



# Atelier ChIP-Seq

Elodie Darbo, Stéphanie Le Gras, Rachel Legendre,  
Denis Puthier, Morgane Thomas-Chollier, Tao Ye



# Contents

- [Introduction](#)
- [Experimental Design](#)
- [Quality Control of the reads](#)
- [Mapping](#)
- [Quality control on mapped reads](#)
- [Visualisation et normalisation](#)
- [Peak Calling](#)
- [Motifs Analysis](#)
- [Annotation](#)
- [Conclusions](#)



# TIPS

- **Keep track** of all command lines you run. You can for example, create a text file in which you write every commands you run.
- Give **content-explicit names** to the files you're generating.
- Give to files the **right extension**.
- **Create directories with explicit names!!**
- **Compress big files** (with gzip for instance).

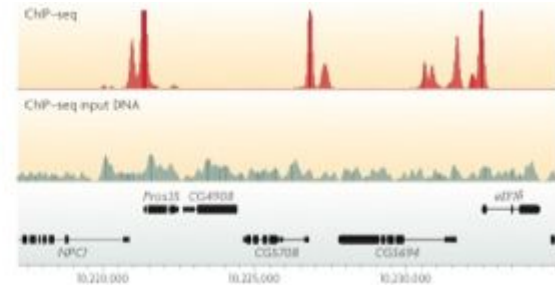


# Introduction

# Chip-Seq analysis

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

```
@SRR002012.1 Oct4:5:1:871:340  
GGCGCACTTACACCCTACATCCATTG  
+  
IIIIIG1?II;IIIIIIIIIII1%.I7I
```



Reads



Peaks

# Chip-Seq analysis

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

Reads



Peaks



Motifs



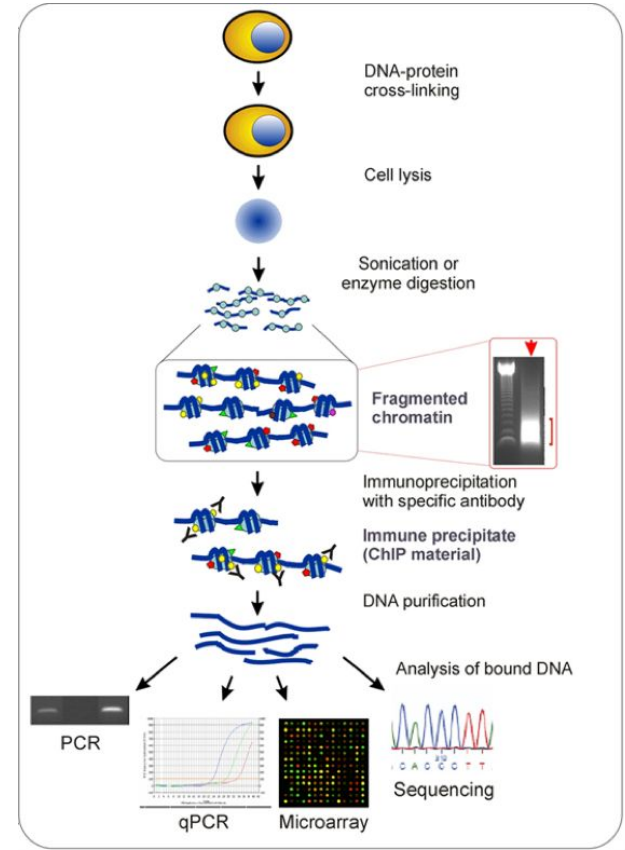
Annotations

# 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**



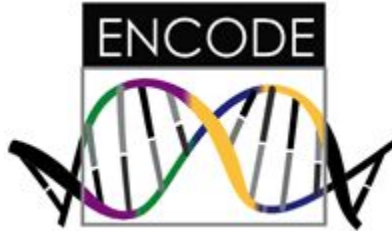


# Experimental design



# 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.

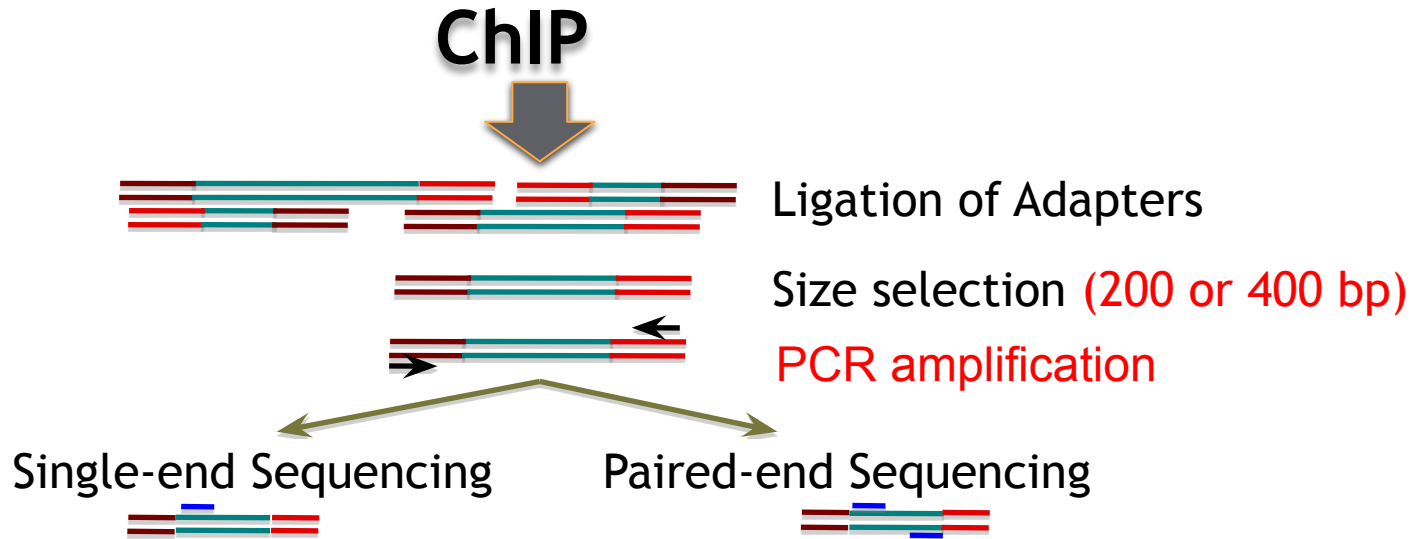
See: <https://www.encodeproject.org/about/experiment-guidelines/>

# 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 NextSeq 2000
- No. of reads per sample: ~40 millions per sample
  - HiSeq 4000 : 8 samples per lane
  - NextSeq 2000, p2 (only PE) : 10 samples per flowcell
  - NextSeq 2000, p3 : 25-30 samples per flowcell
- Length of DNA fragment : ~200bp
- No. of cycle per run : 50

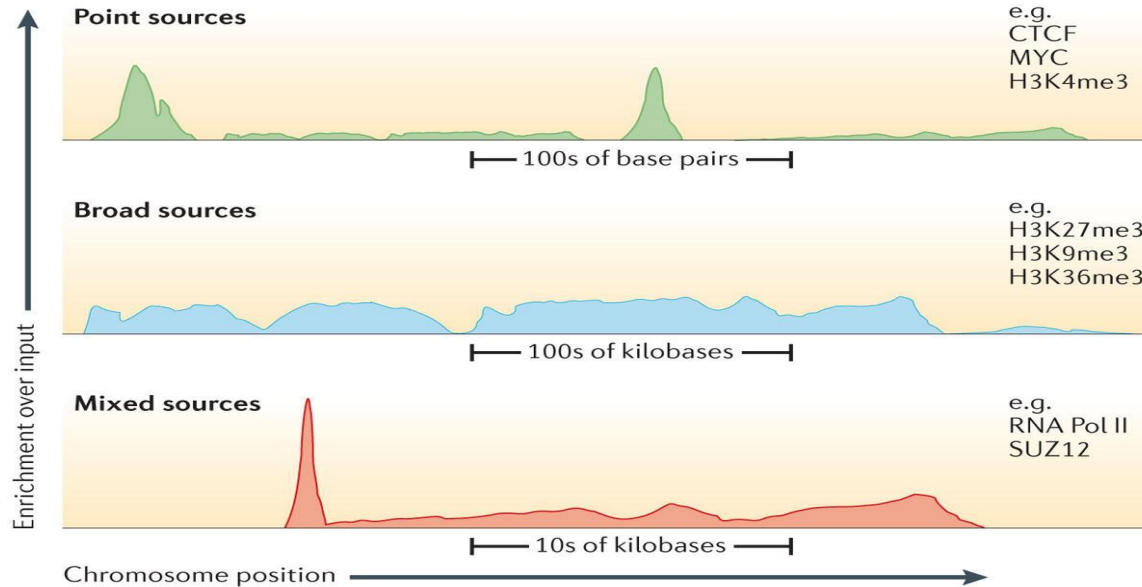


# Single end or paired end ?

- 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 **repeat regions**
  - ☹️ The price! But 2 x 40bp is affordable

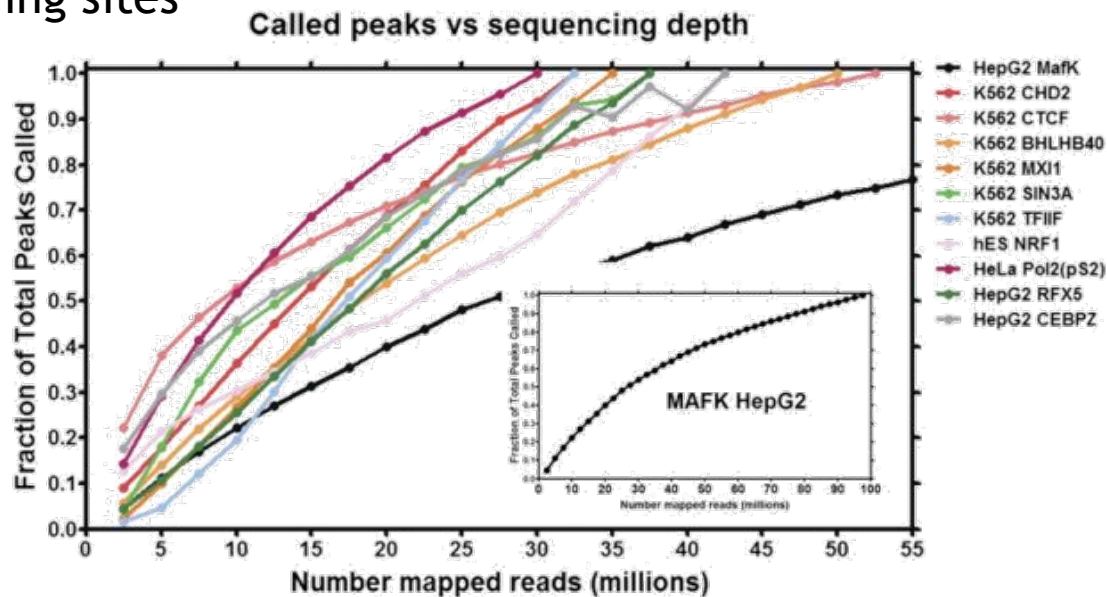
# Sequencing depth

- Consider the depth needed depending on:
  - Chipped protein



# Sequencing depth

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



# 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 ?

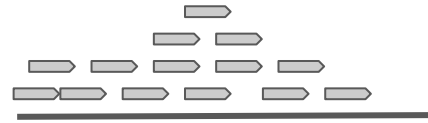
- The input is used to model local noise level
  - Accessible regions are expected to produce more reads



Closed Open Closed

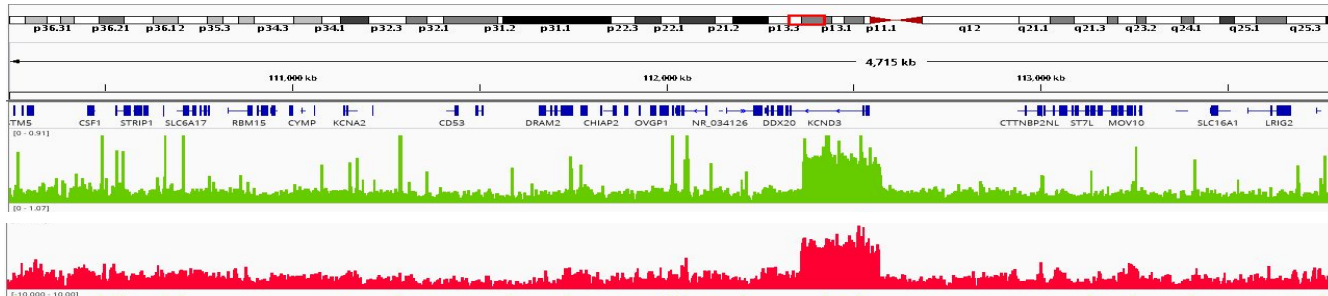


Closed Open Closed



Closed Open Closed

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



# Why an Input is required ?

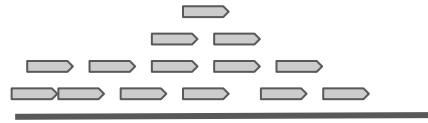
- The input is used to model local noise level
  - Accessible regions are expected to produce more reads



Closed Open Closed



Closed Open Closed



Closed Open Closed

- 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 Récepteur
  - Induced by Dexamethasone (Dex)
  - Control vehicule = Ethanol (EthOH)
- KO of your protein of interest
- Non flagged cell lines
- ...



# 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.

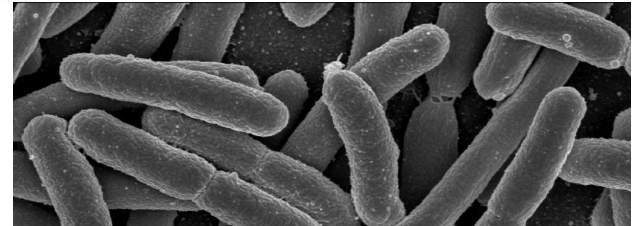


Data analysed in this course

# Dataset used

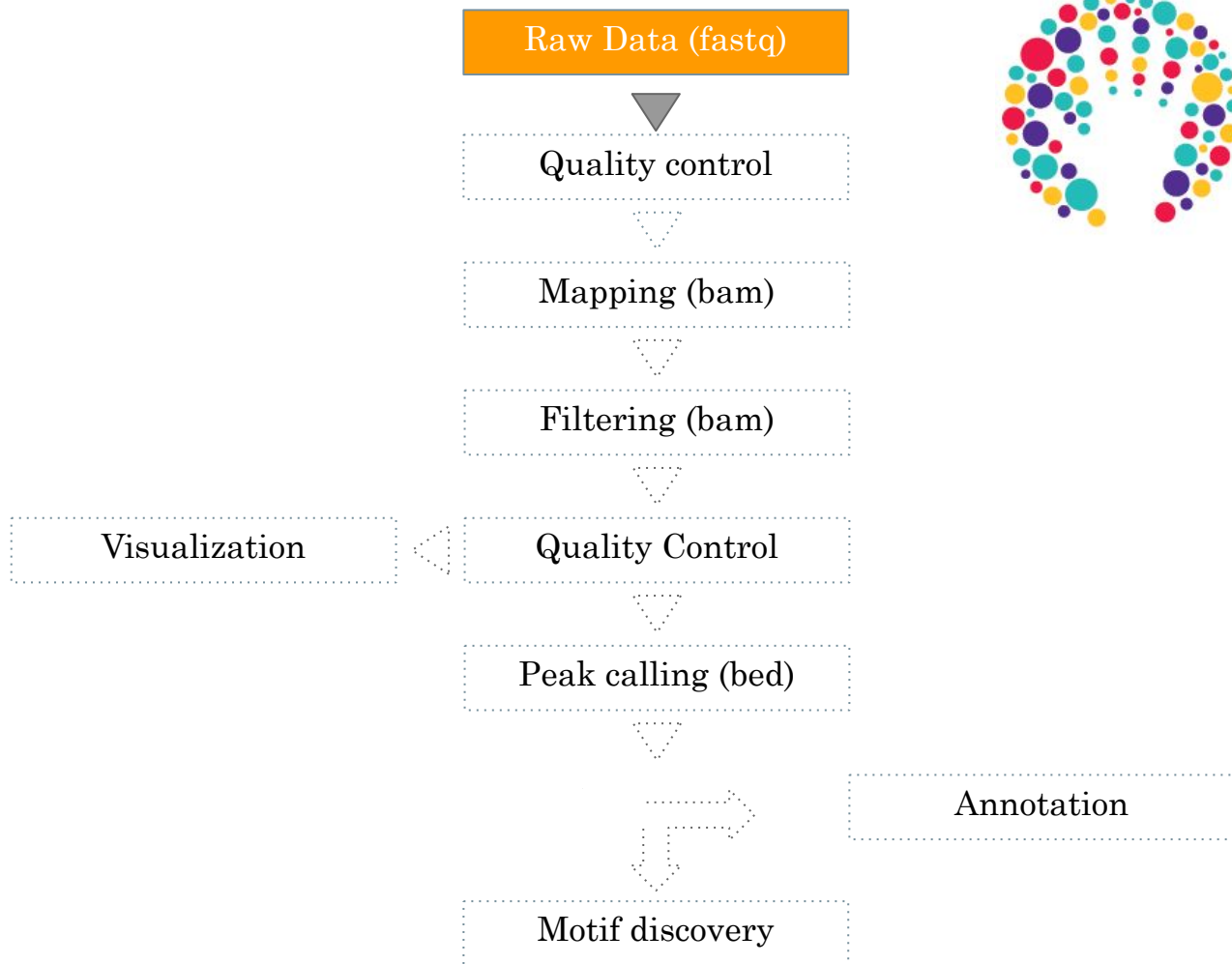
## Genome-scale Analysis of *Escherichia coli* FNR Reveals Complex Features of Transcription Factor Binding

Kevin S. Myers<sup>1,2</sup>, Huihuang Yan<sup>3aa</sup>, Irene M. Ong<sup>3</sup>, Dongjun Chung<sup>4ab</sup>, Kun Liang<sup>4,5ac</sup>, Frances Tran<sup>6ad</sup>, Sündüz Keleş<sup>4,5</sup>, Robert Landick<sup>3,6,7\*</sup>, Patricia J. Kiley<sup>2,3\*</sup>



- All experiments (GEO): GSE41187
- Experiment: FNR IP ChIP-seq Anaerobic A (SRX189773 - SRR576933)
- Control: anaerobic INPUT DNA (SRX189778 - SRR576938)

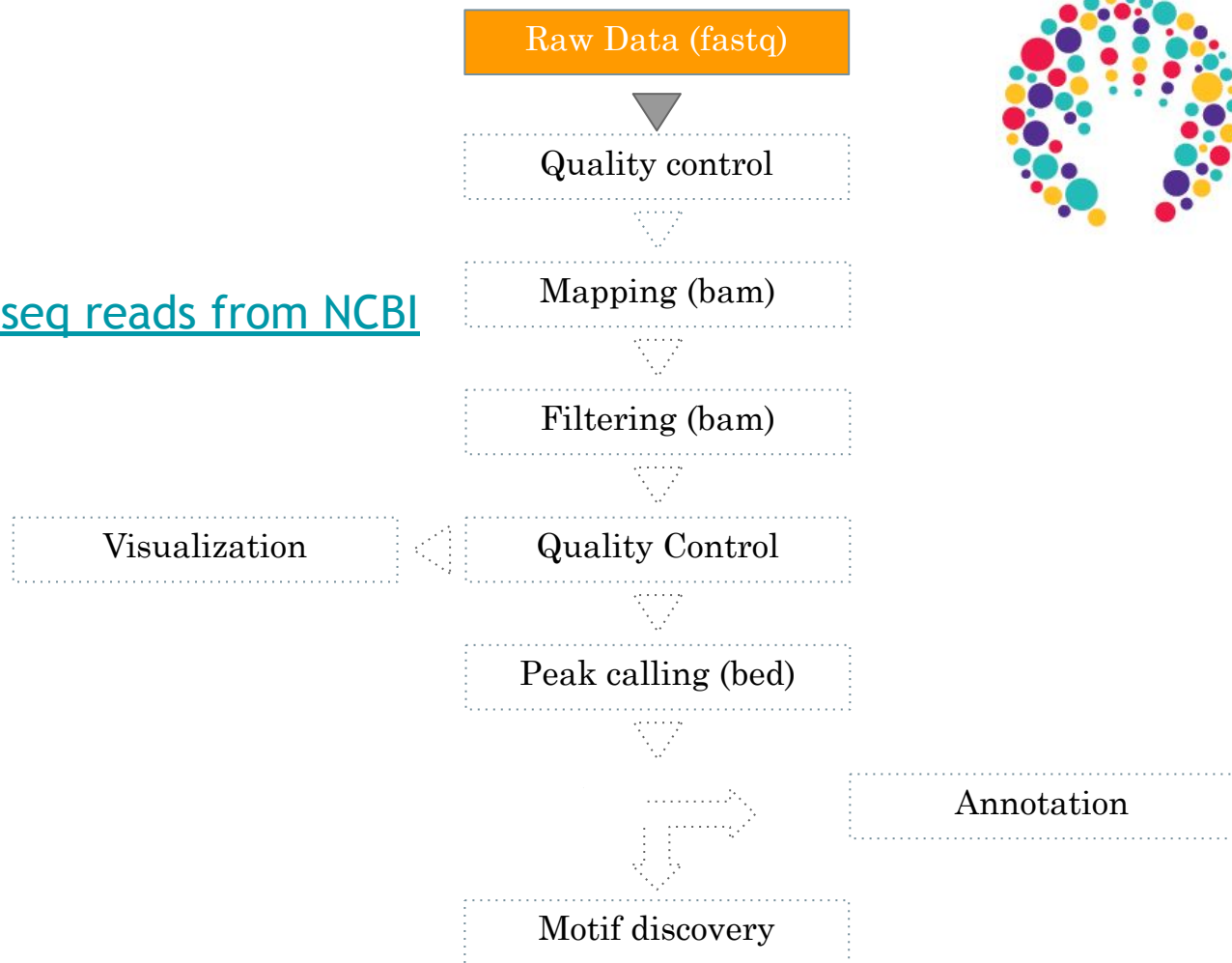
# Protocol





# Protocol

- [Downloading ChIP-seq reads from NCBI](#)



# Protocol



- [Connect to the server and set up your environment](#)



# Quality control of the reads

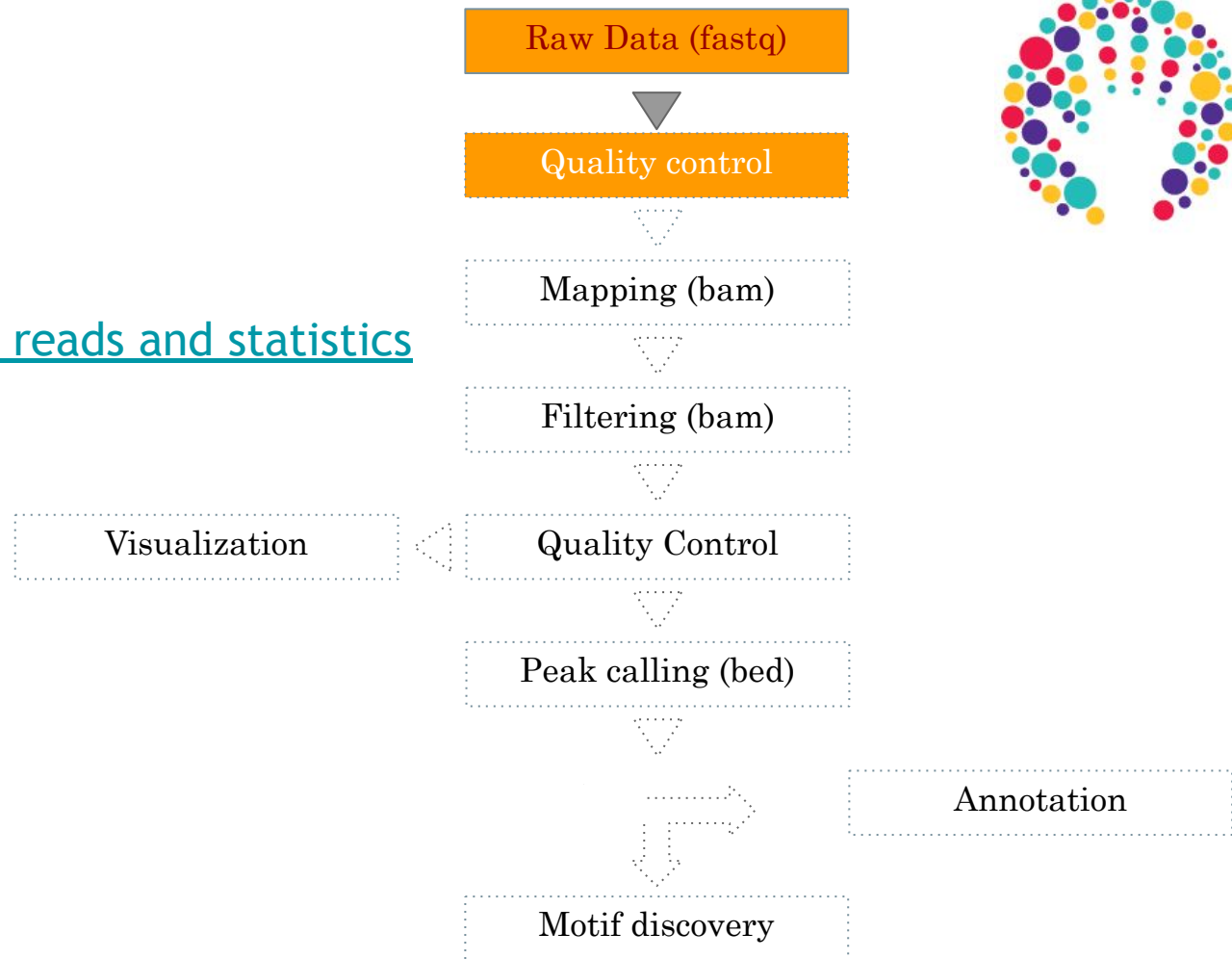


# Quality control of the reads

- As for any NGS datasets
- FastQC program (c.f. Course “preprocessing” Monday afternoon)

# Protocol

- Quality control of the reads and statistics





# Mapping

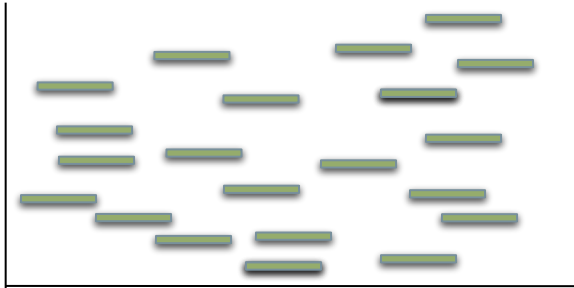
# Mapping

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

Ref. Genome



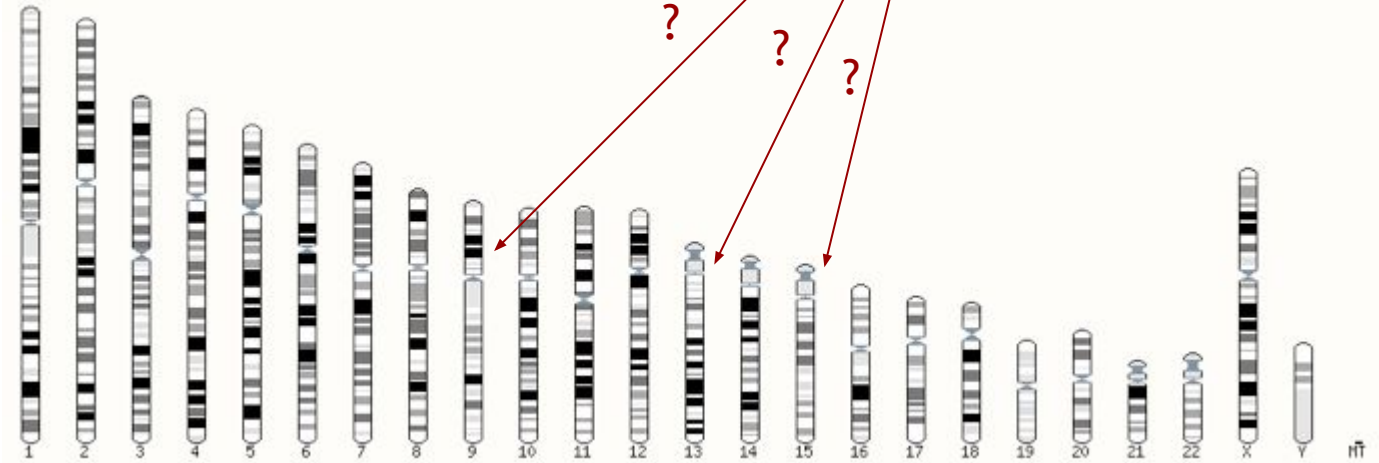
Reads



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

# Mapping example

ATGCGATTA

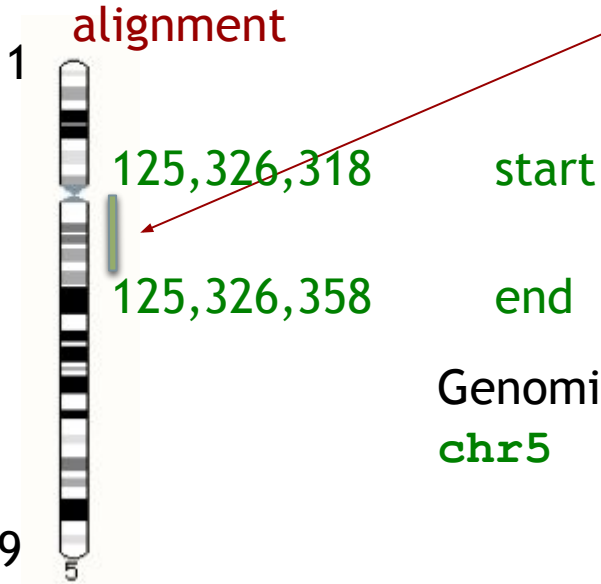


Human chromosomes



# Genomic coordinates

ATGCGATTA



Genomic coordinate of the mapped read :

**chr5 125326318 125326358 +**

181,538,259

chromosome 5

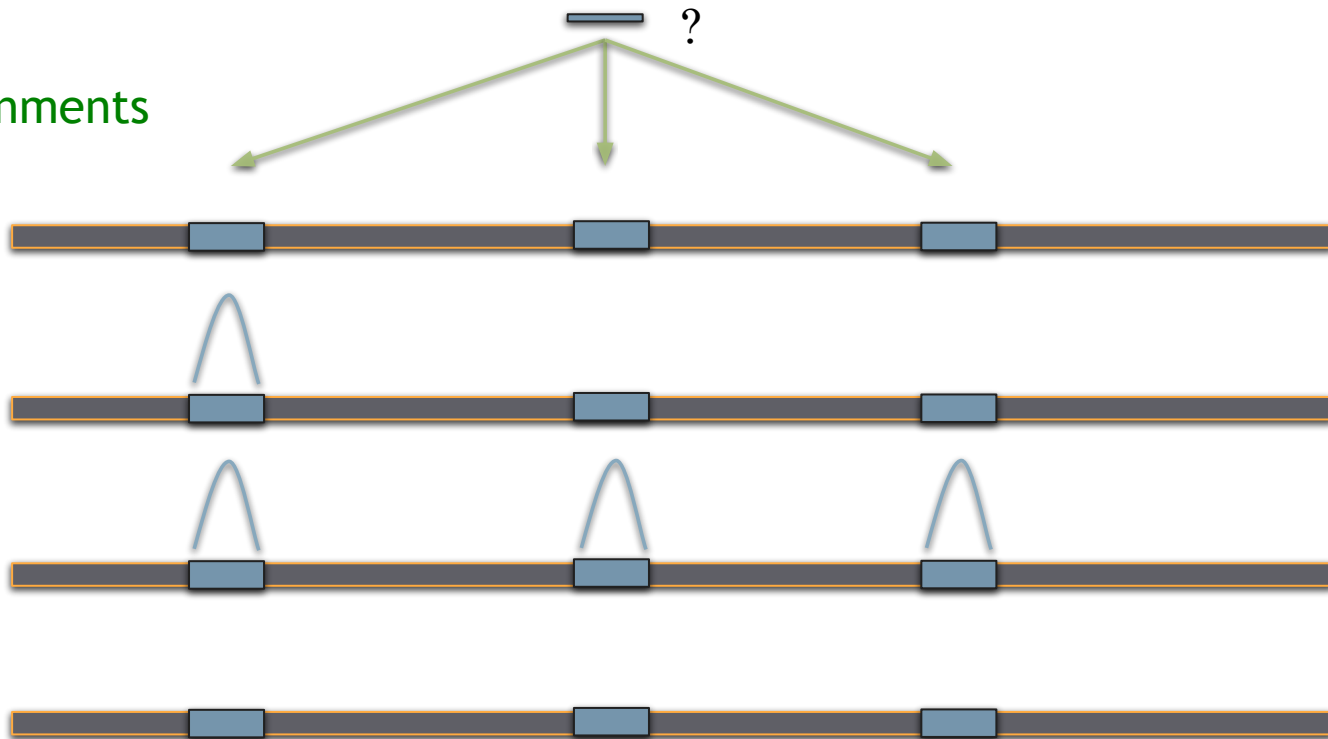


# Mapping tool used: Bowtie

- 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



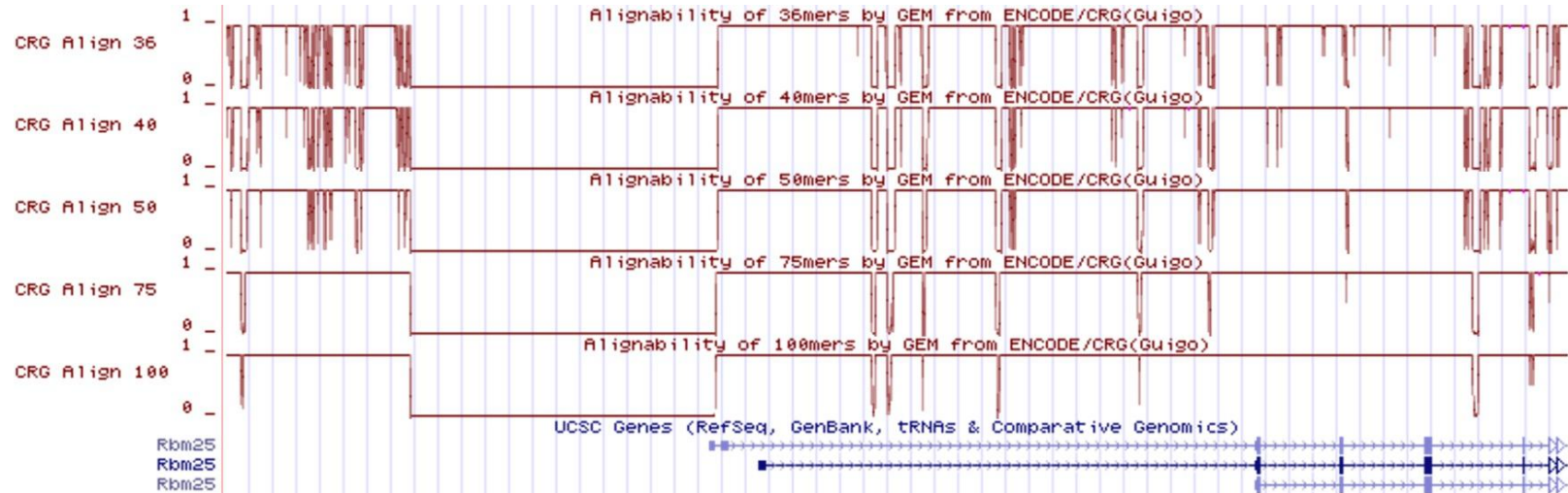
Keep 1 position  
randomly

Keep all  
possible position

Keep none

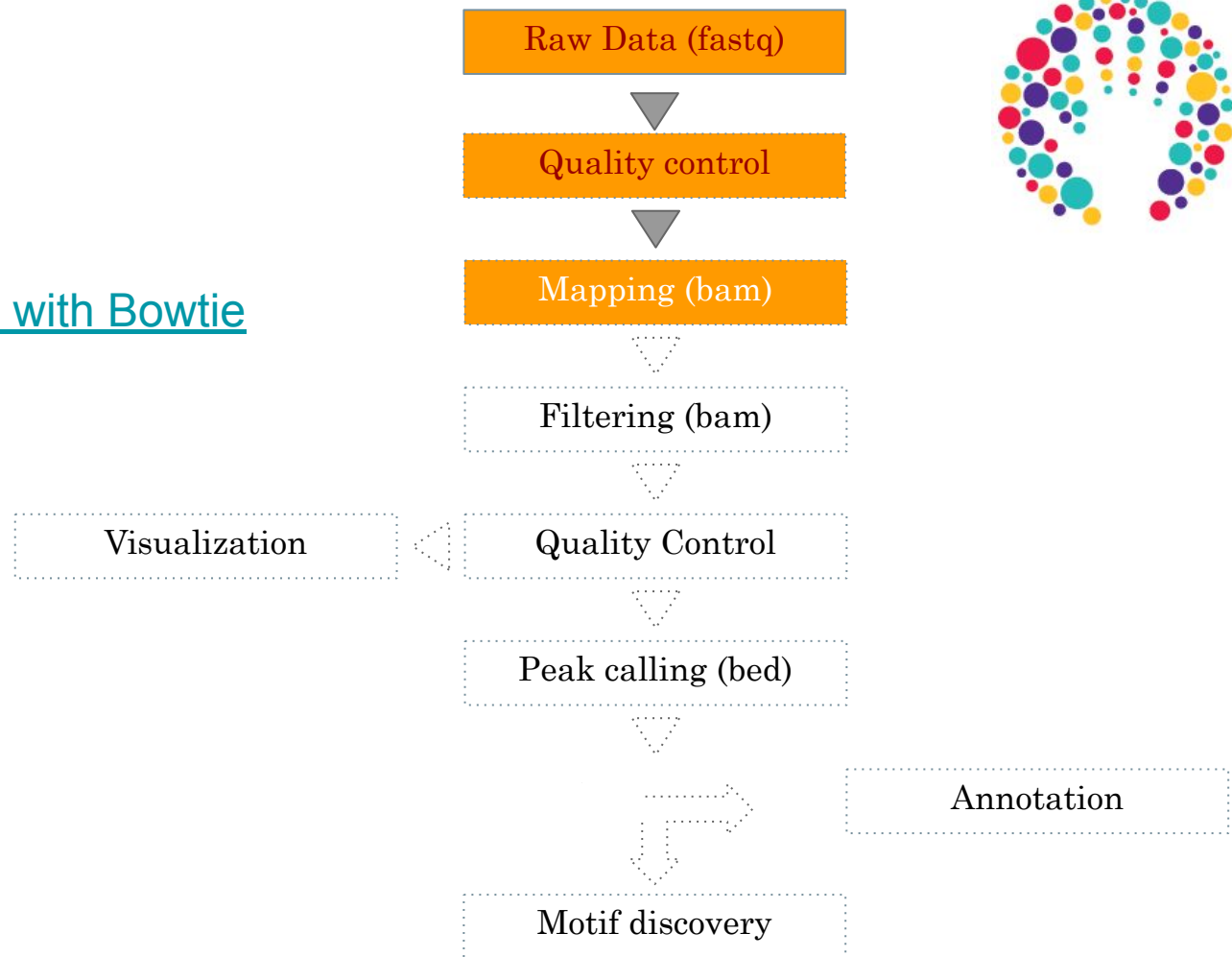
# Mappability

- 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)



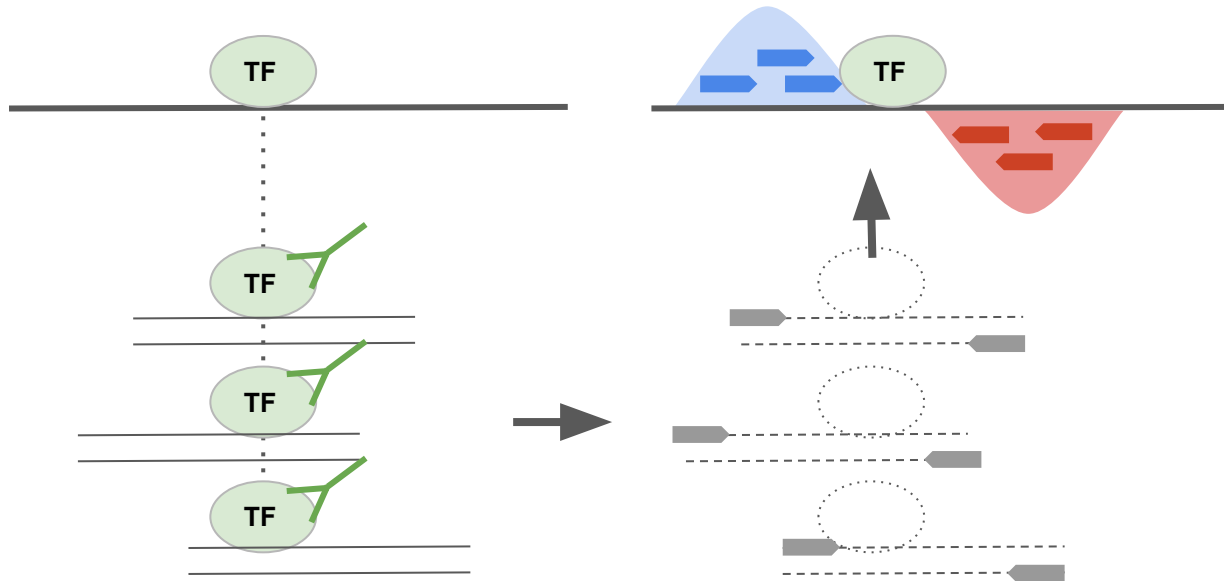
# Protocol

- Mapping the reads with Bowtie



# Mapping: expected signal

- For a transcription factor signal is expected to be sharp

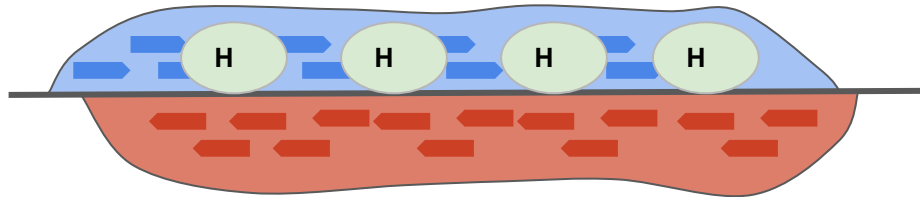


The binding site itself is generally not sequenced !

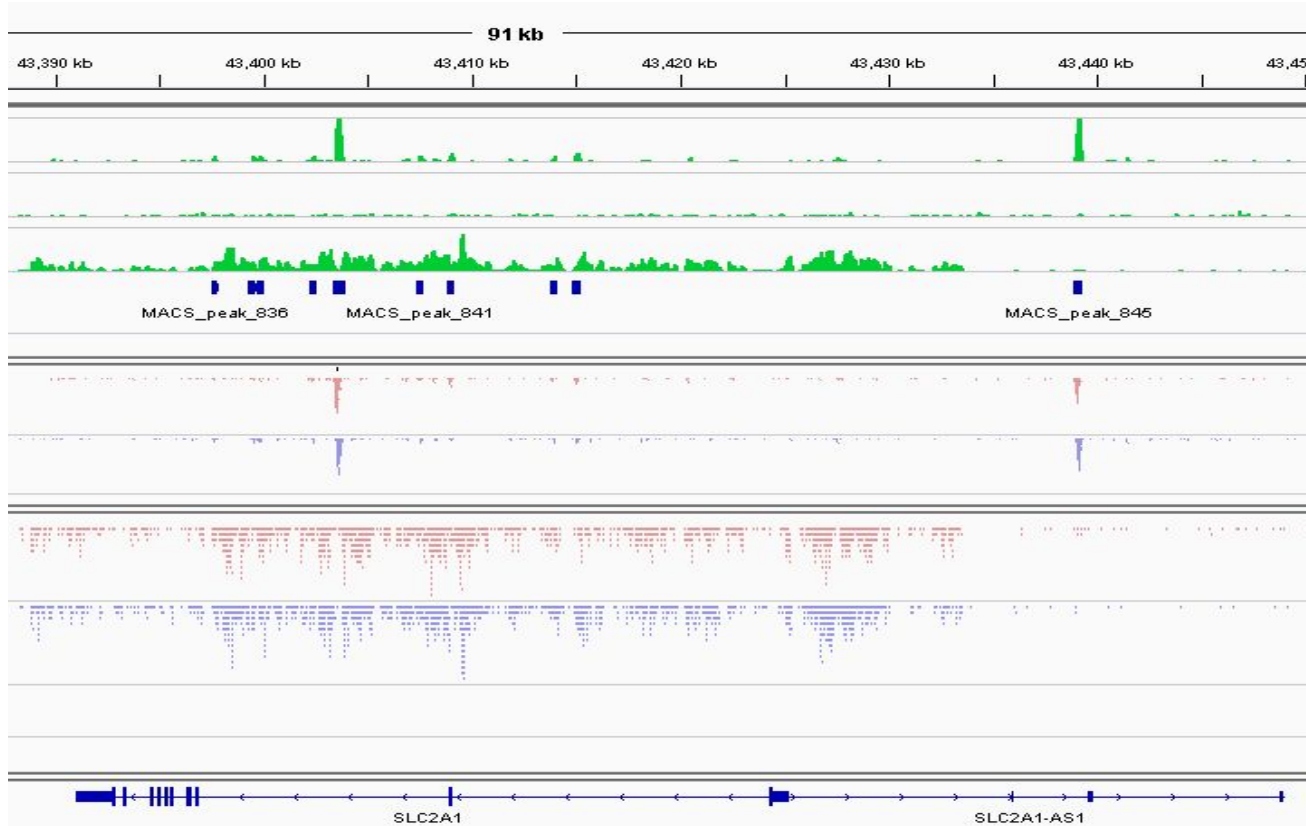
— Sens alignments  
— Rev/comp alignments

# 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)





# Filtering mapped reads

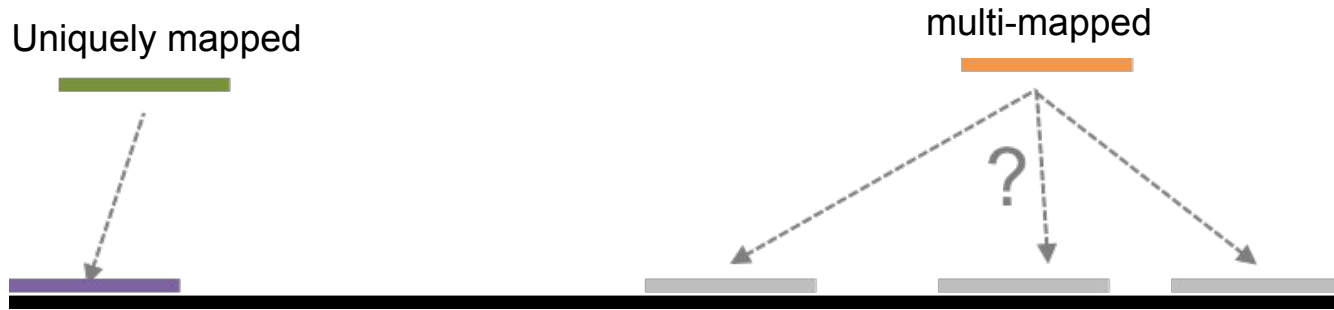


# Which reads to filter ?

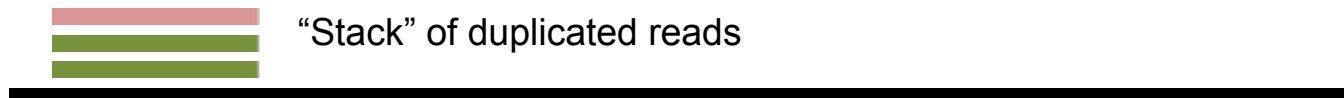
- Low-quality read alignments
  - Tool : samtools
- Multi-mapped reads (unless removed during the mapping step)
  - Tool : samtools
- Duplicated reads (PCR duplicates)
  - Tool : Picard MarkDuplicates

# 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)

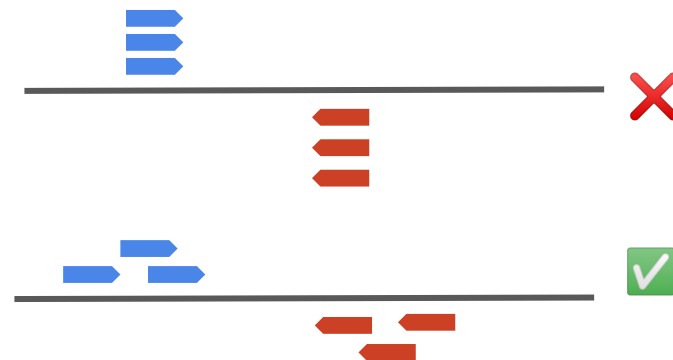


# PCR duplicates

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

# QC : PBC (PCR Bottleneck Coefficient)

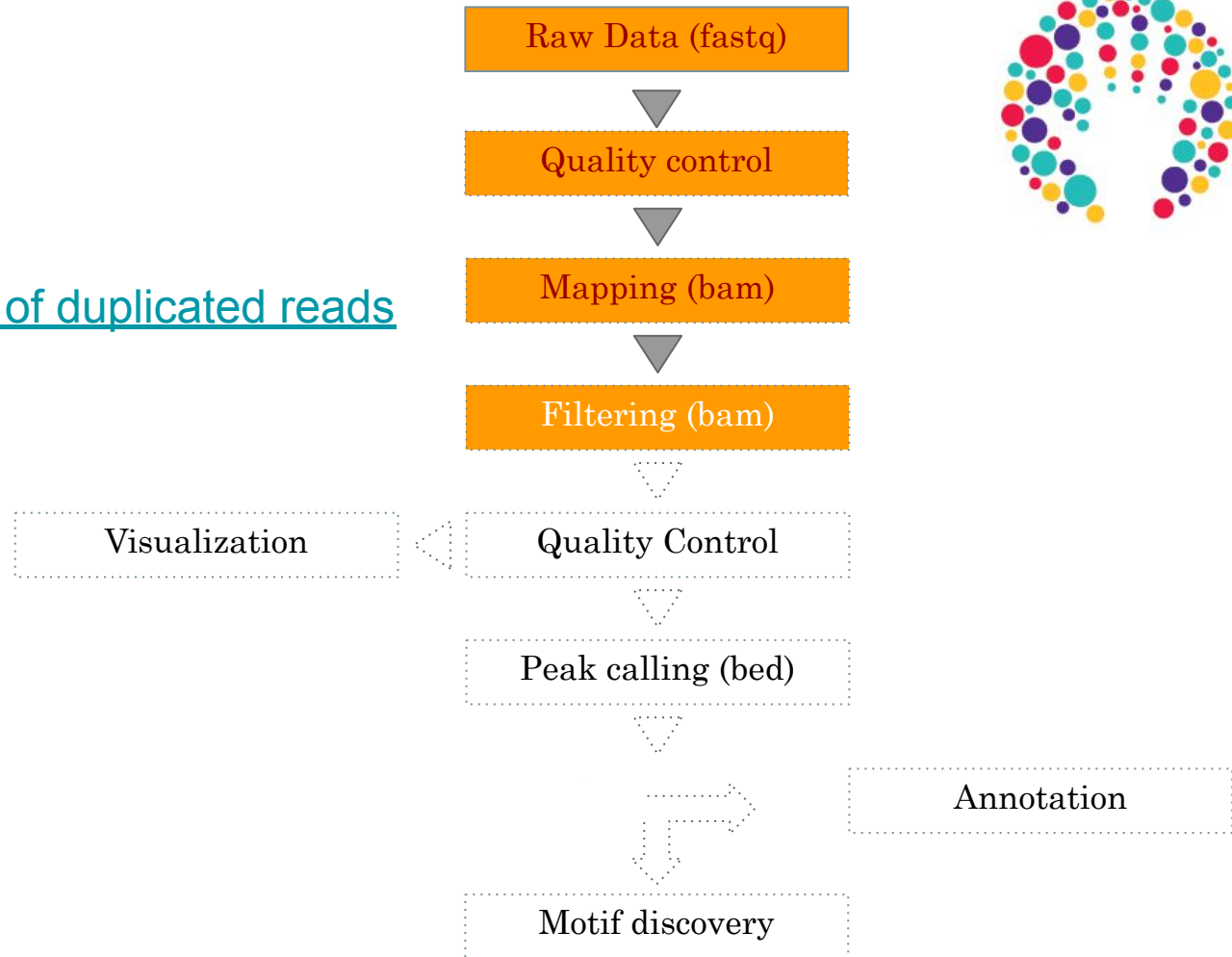
- An approximate measure of library complexity
- $PBC = N_1 / N_d$ 
  - $N_1$  = Genomic position with 1 read aligned
  - $N_d$  = Genomic position with  $\geq 1$  read aligned
- Value :
  - 0-0.5: severe bottlenecking
  - 0.5-0.8: moderate bottlenecking
  - 0.8-0.9: mild bottlenecking
  - 0.9-1.0: no bottlenecking



# Protocol



- Estimating the number of duplicated reads





# Quality Control on mapped reads

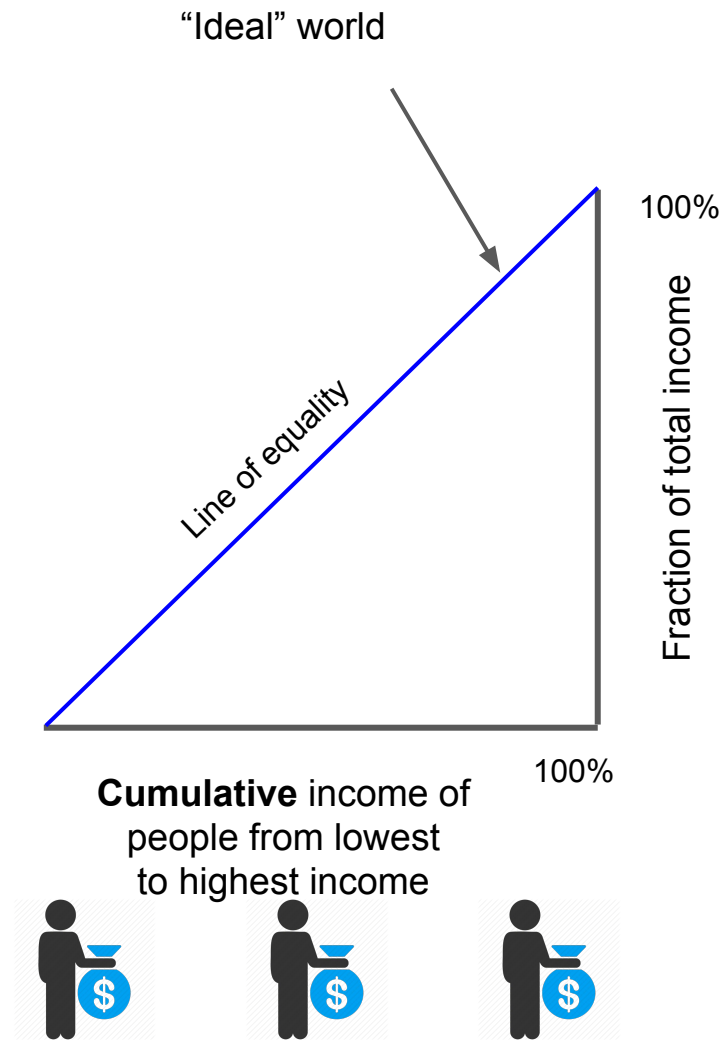
# Assessing ChIP quality

- Guidelines from ENCODE
- Various metrics
  - Check **duplicate** rate (see previous Filtering section)
  - Use a **Lorenz Curve** (implemented in **Deeptools fingerprint**)
  - Look at **strand cross-correlation** (implemented in **SPP BioC package** and **phantompeakqualtools**)
  - Fraction of reads in peaks (**FRiP**, as proposed by ENCODE), but requires to find peaks.



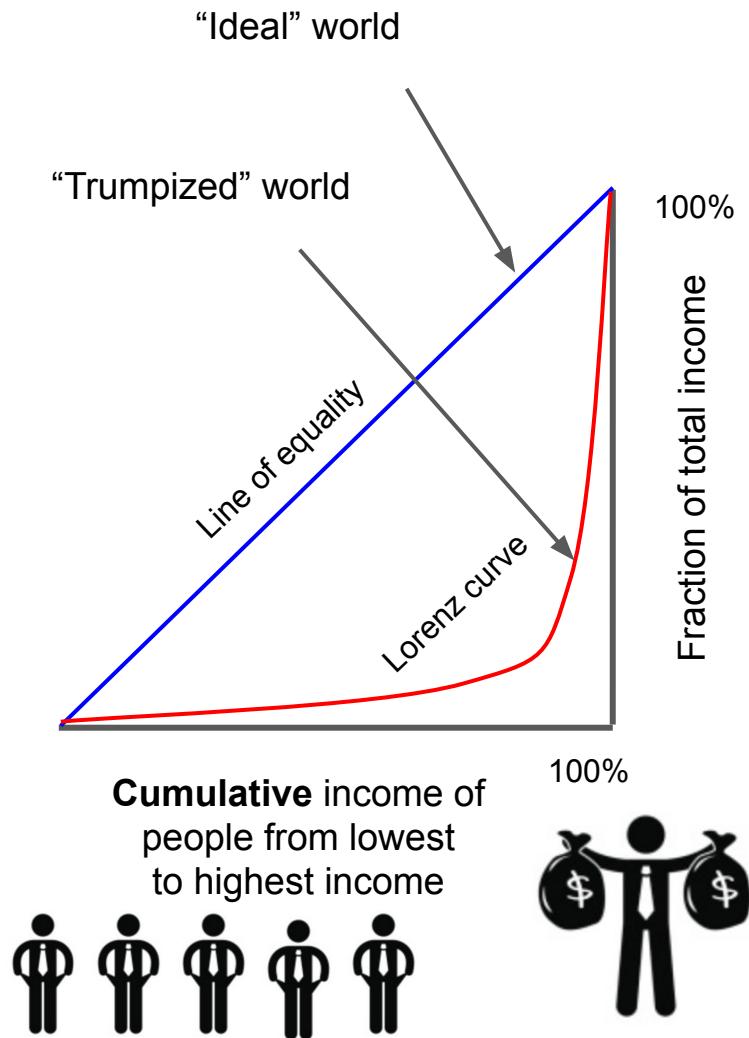
# Lorenz curve

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



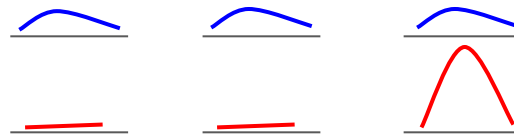
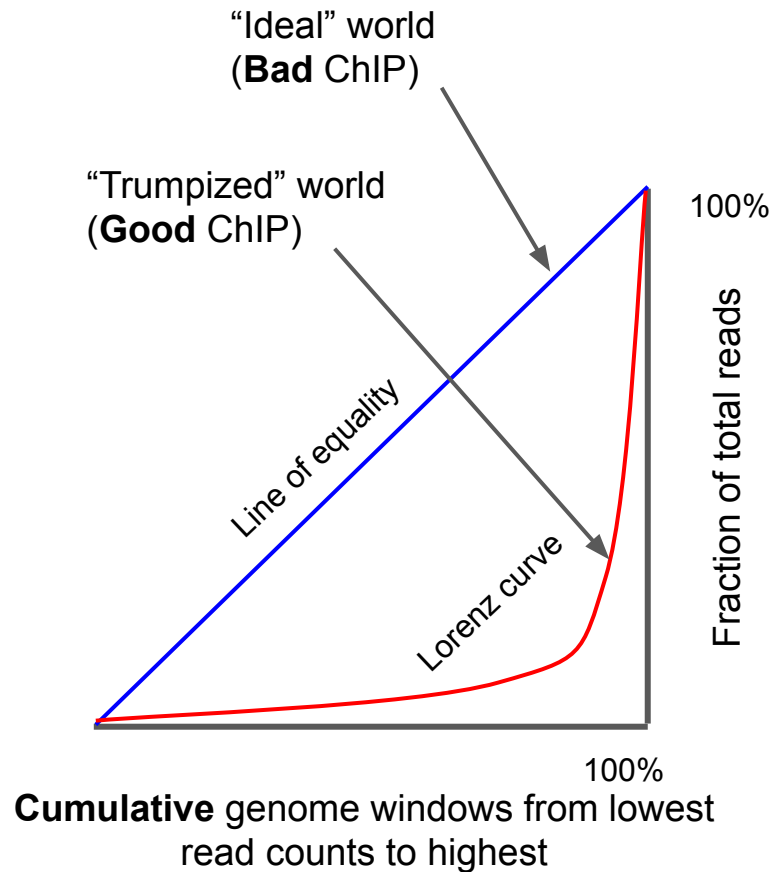
# Lorenz curve

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



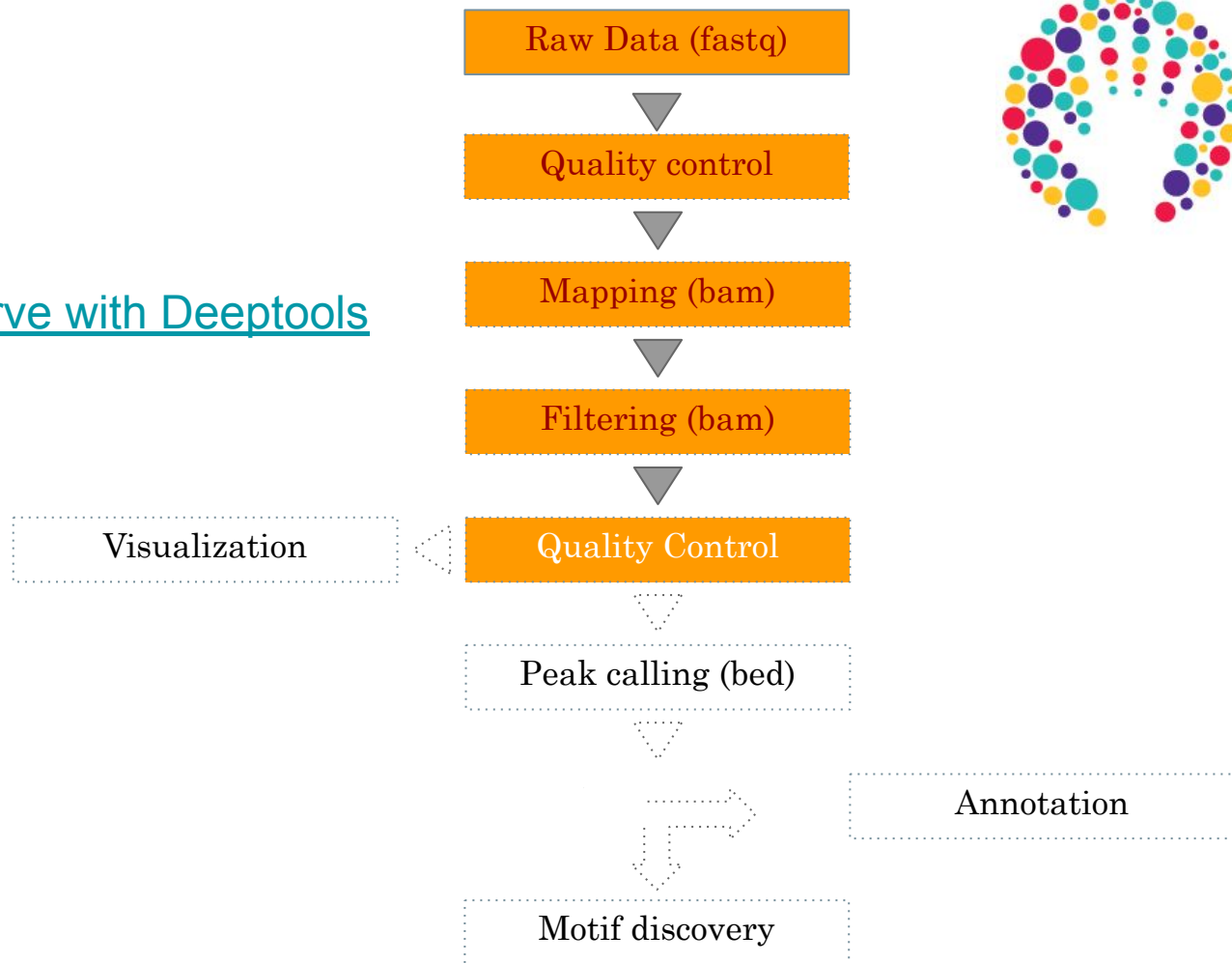
# Lorenz curve

- 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



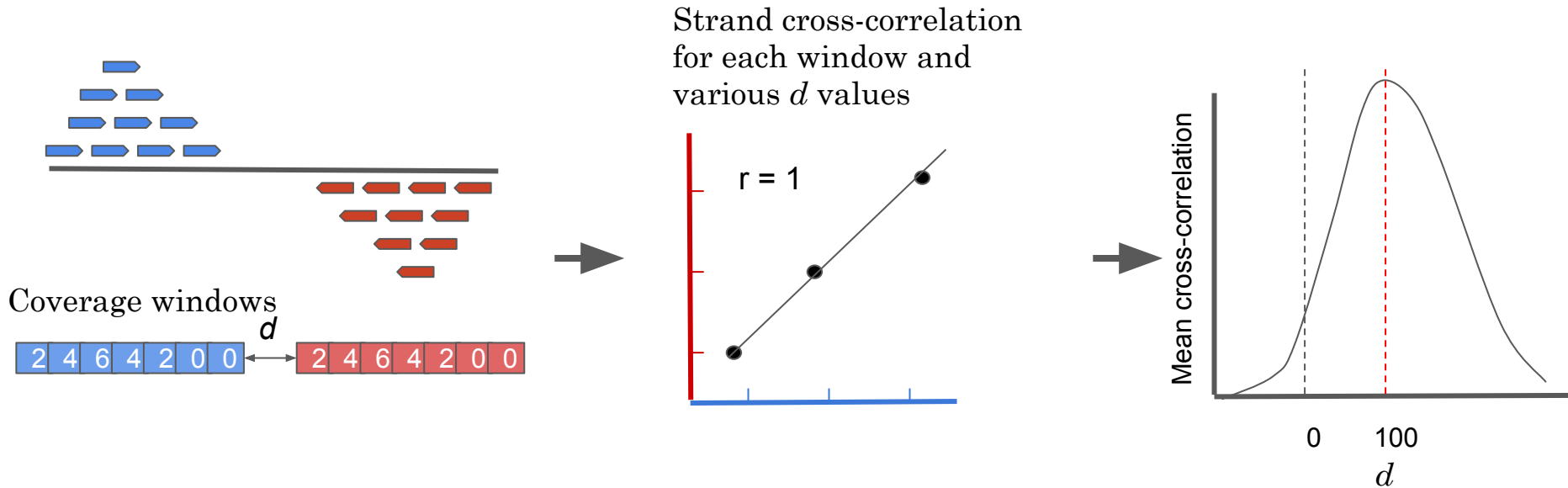
# Protocol

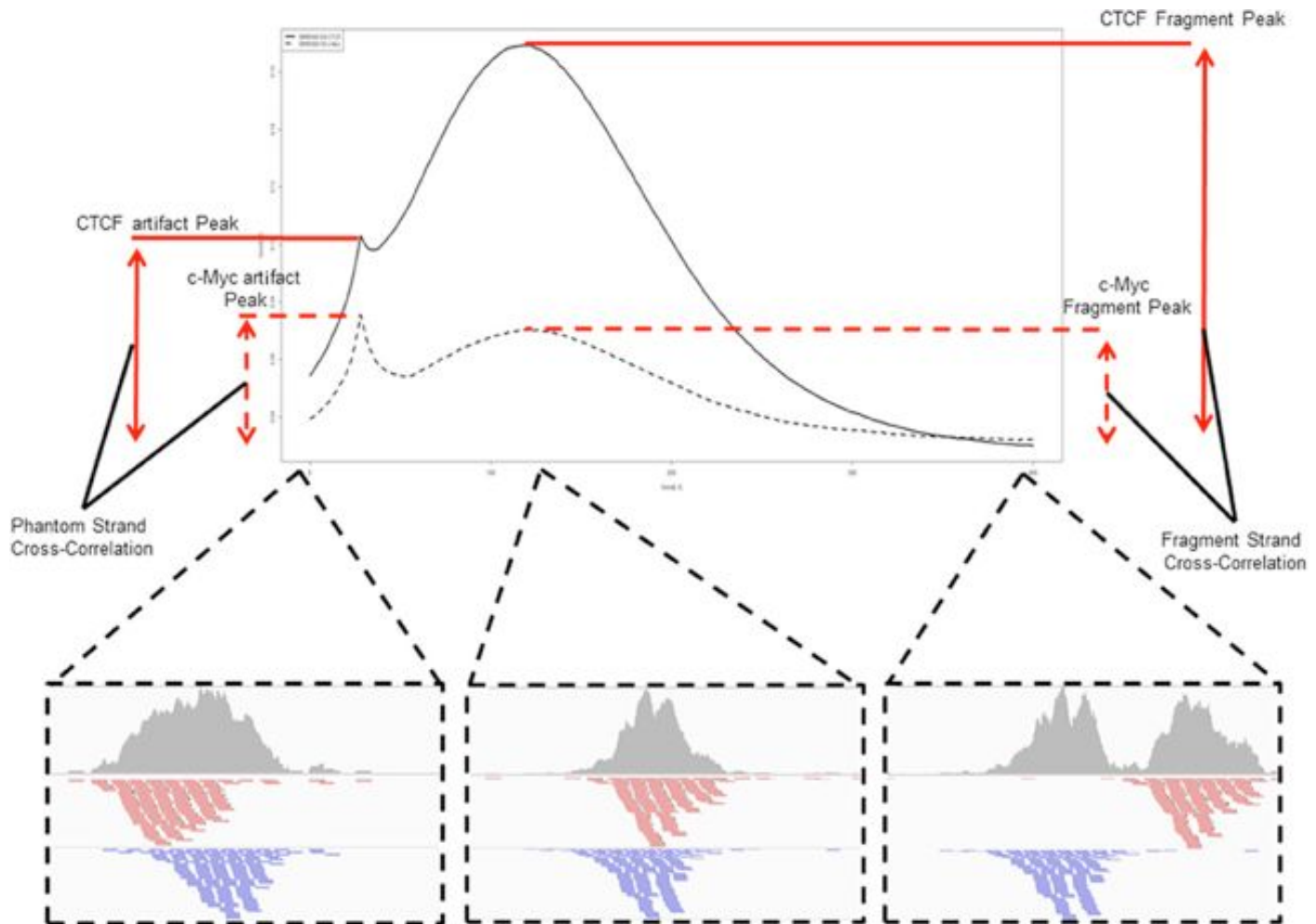
- [Plot the Lorenz curve with Deeptools](#)



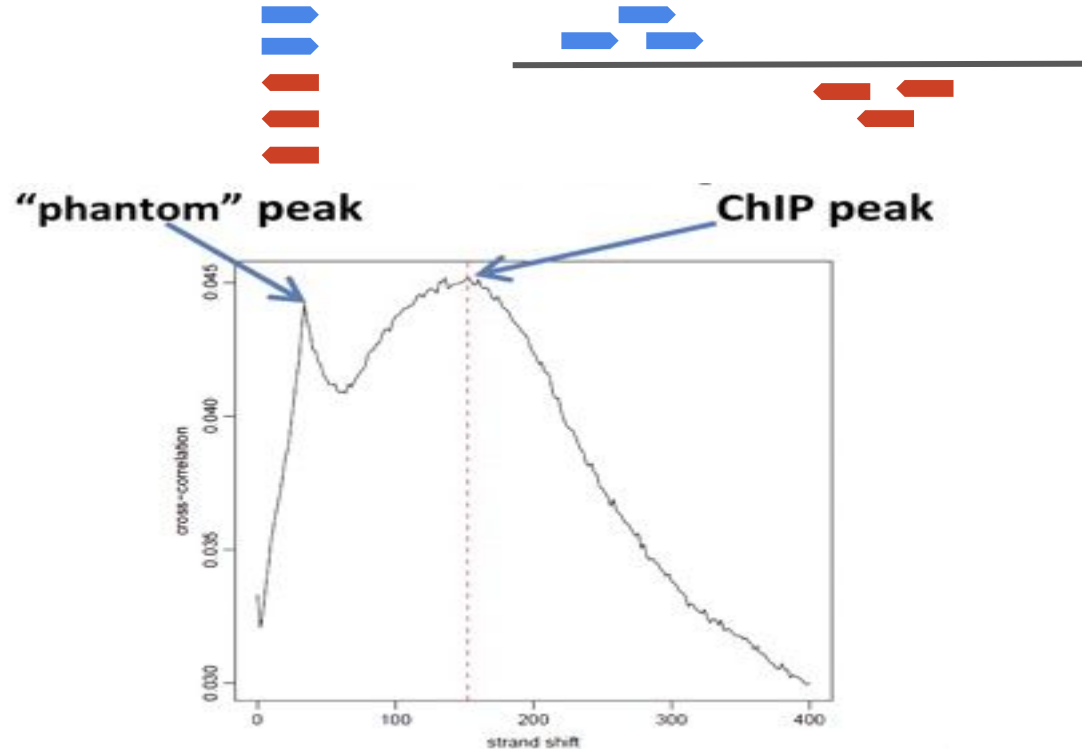
# Strand cross-correlation

- Compute strand cross correlation for each window  $w$  across the genome.
- Use various distance  $d$  and compute the mean cross-correlation observed

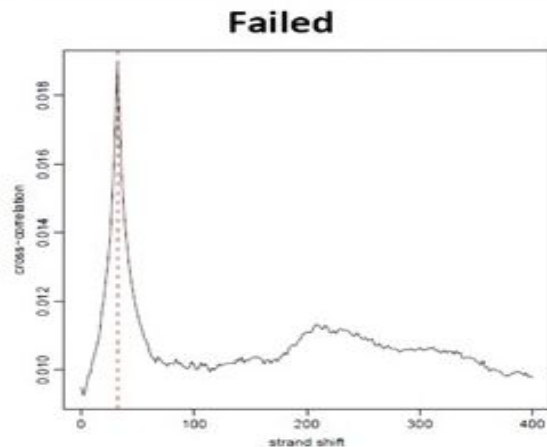
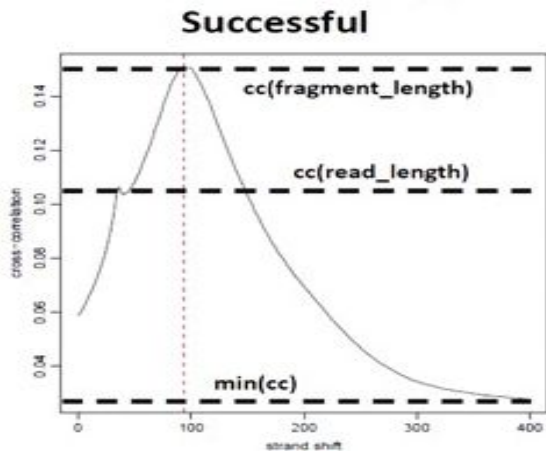




# Strand cross-correlation



# Strand cross-correlation



NSC: normalized strand coefficient

$$NSC = \frac{cc(\text{fragment length})}{\min(cc)}$$


NSC  $\geq$  1.05 is recommended

Relative strand correlation (RSC)

$$RSC = \frac{cc(\text{fragment length}) - \min(cc)}{cc(\text{read length}) - \min(cc)}$$

RSC  $\geq$  0.8 is recommended

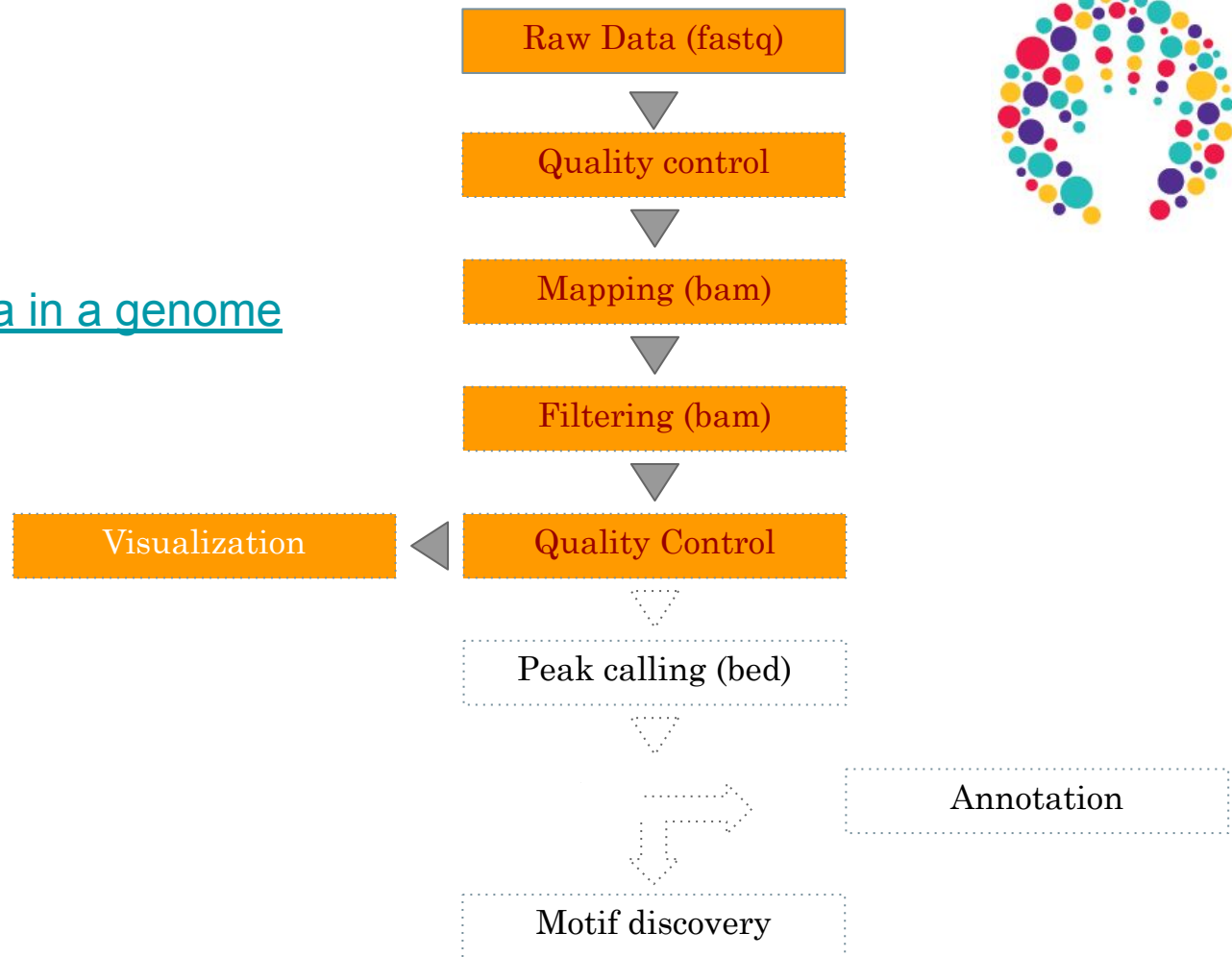







Visualization: computing a genomic coverage file

# Protocol

- [Visualizing the data in a genome browser](#)

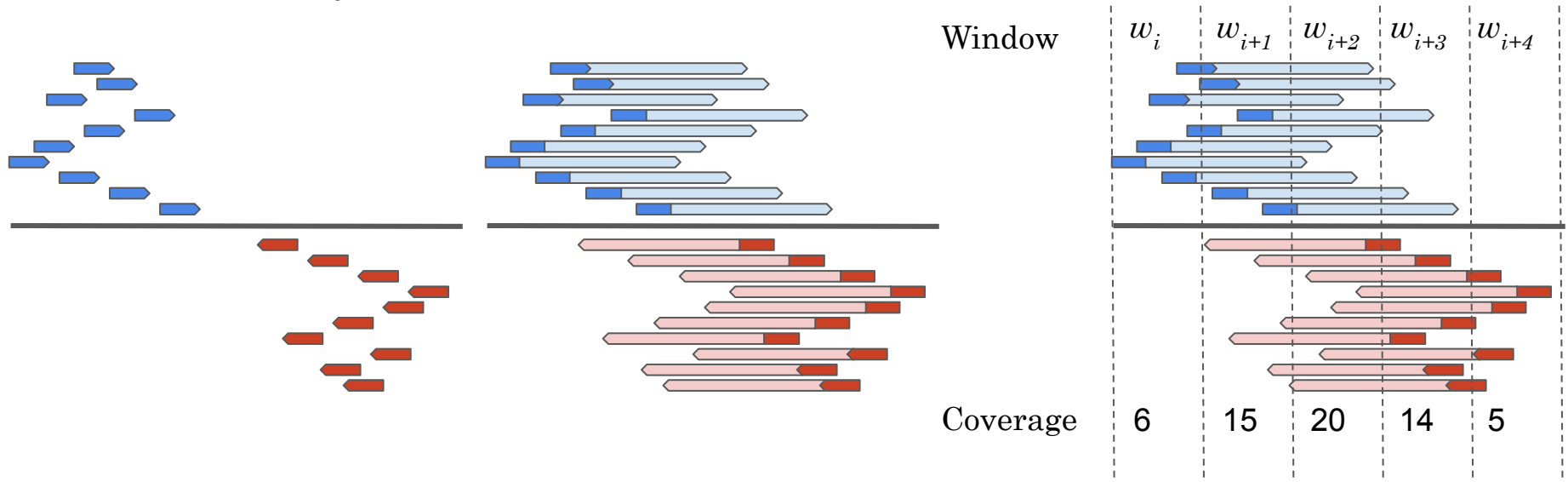


# Bam files are fat

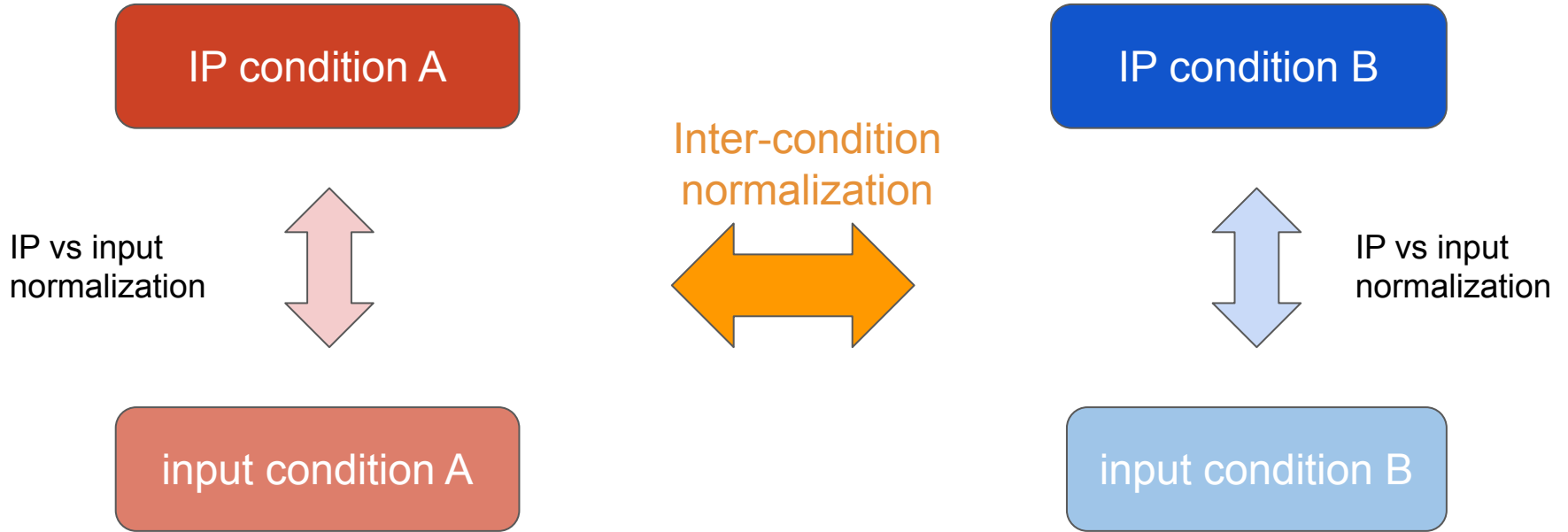
- **BAM files are fat** as they do contain exhaustive information about read alignments
  - Memory issues (can only visualize fraction of the BAM)
- Need a more **lightweight** file format containing **only genomic coverage** information:
  -  **Wig** (not compressed, not indexed)
  -  **TDF** (compressed, indexed)
  -  **BigWig** (compressed, indexed)

# Coverage file and read extension

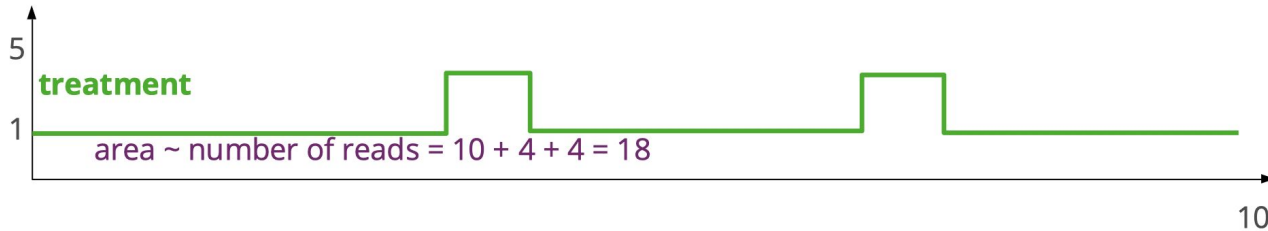
- BAM files do not contain fragment location but read location
- We need to extend reads to compute fragments coordinates before coverage analysis
- Not required for PE



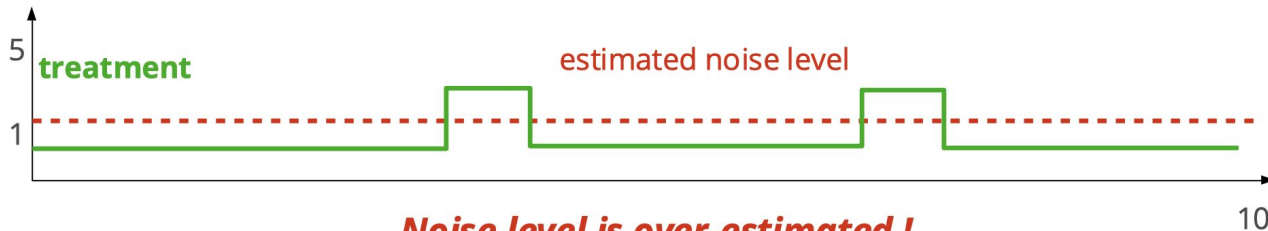
# Library size normalization



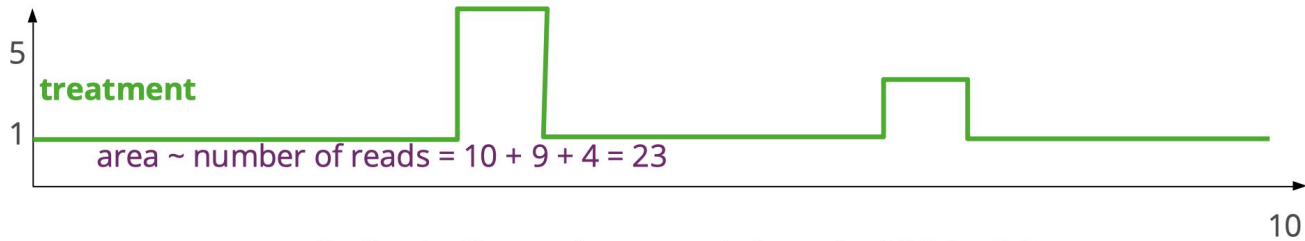
# Library size normalization (input vs IP)



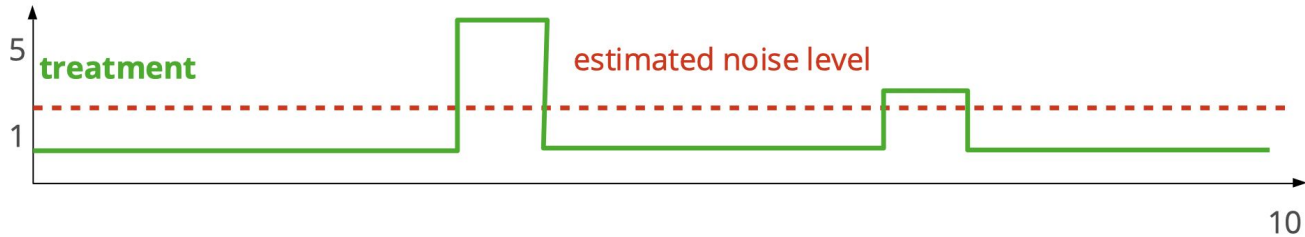
Scaling by library size : upscale input by  $18/10 = 1.8$



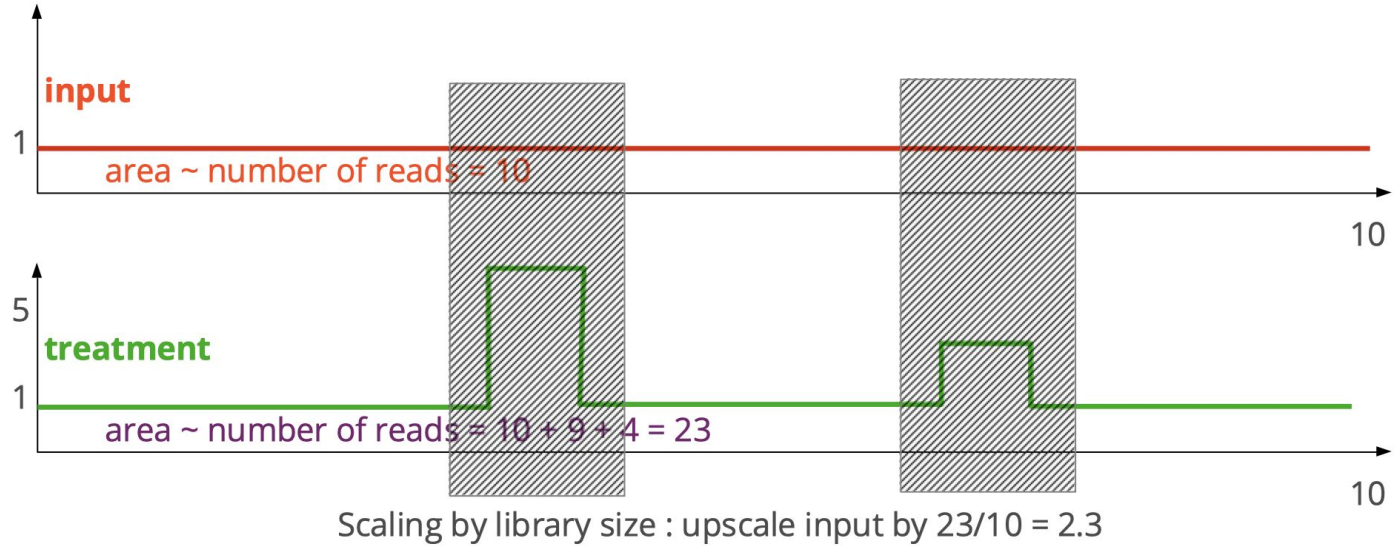
# Library size normalization (input vs IP)



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



# Library size normalization (input vs IP)



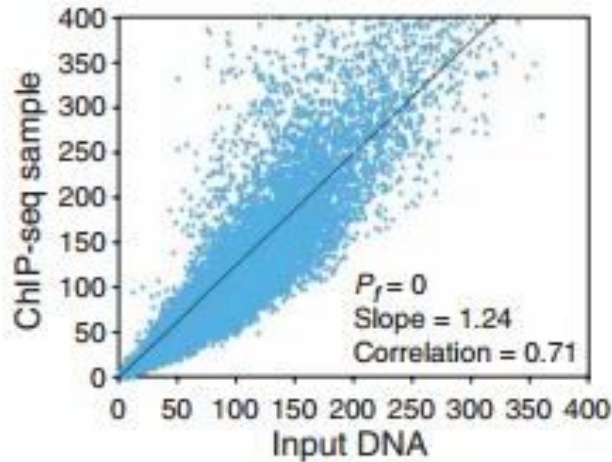


# Library size normalization

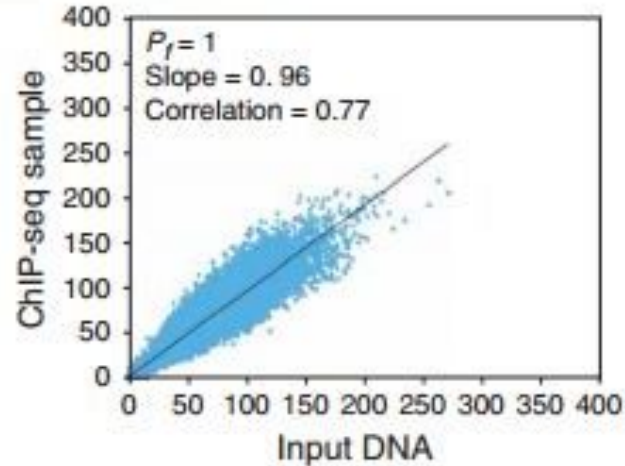
PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls

Example of PeakSeq

Joel Rozowsky , Ghia Euskirchen, Raymond K Auerbach, Zhengdong D Zhang, Theodore Gibson, Robert Bjornson, Nicholas Carriero, Michael Snyder & Mark B Gerstein 



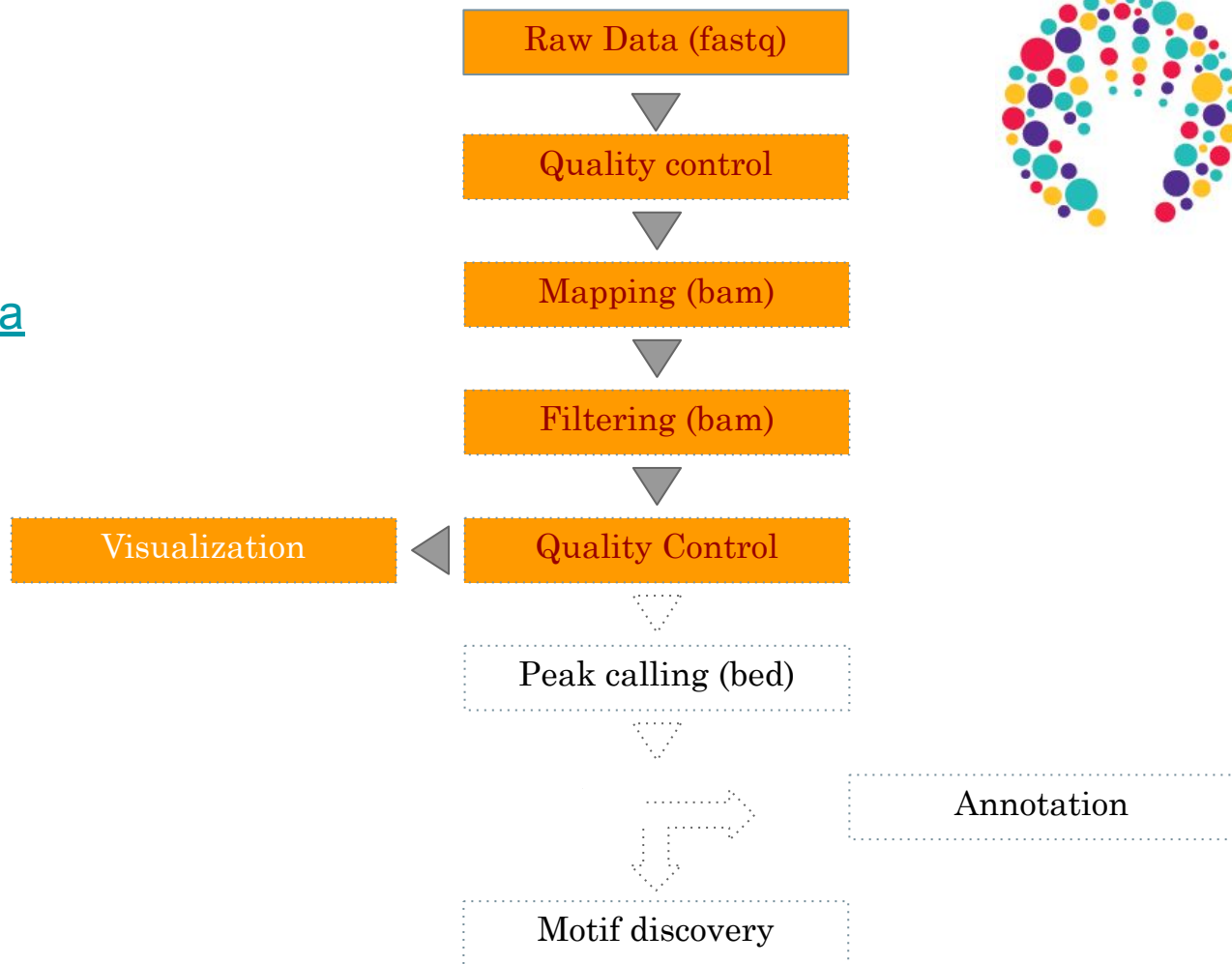
All peaks



Signal peaks removed

# Protocol

- [Viewing scaled data](#)





# Peak Calling

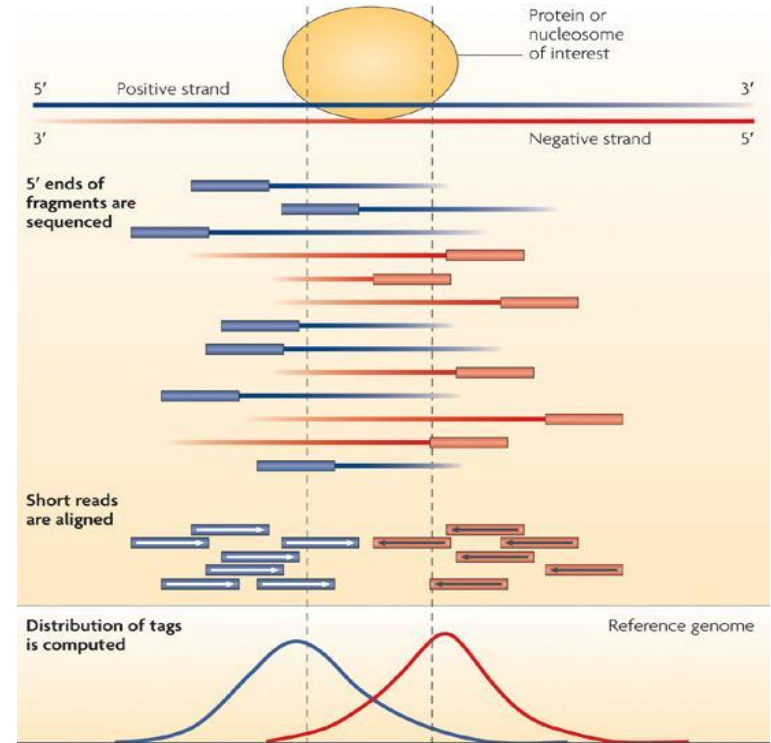
Reads



Peaks

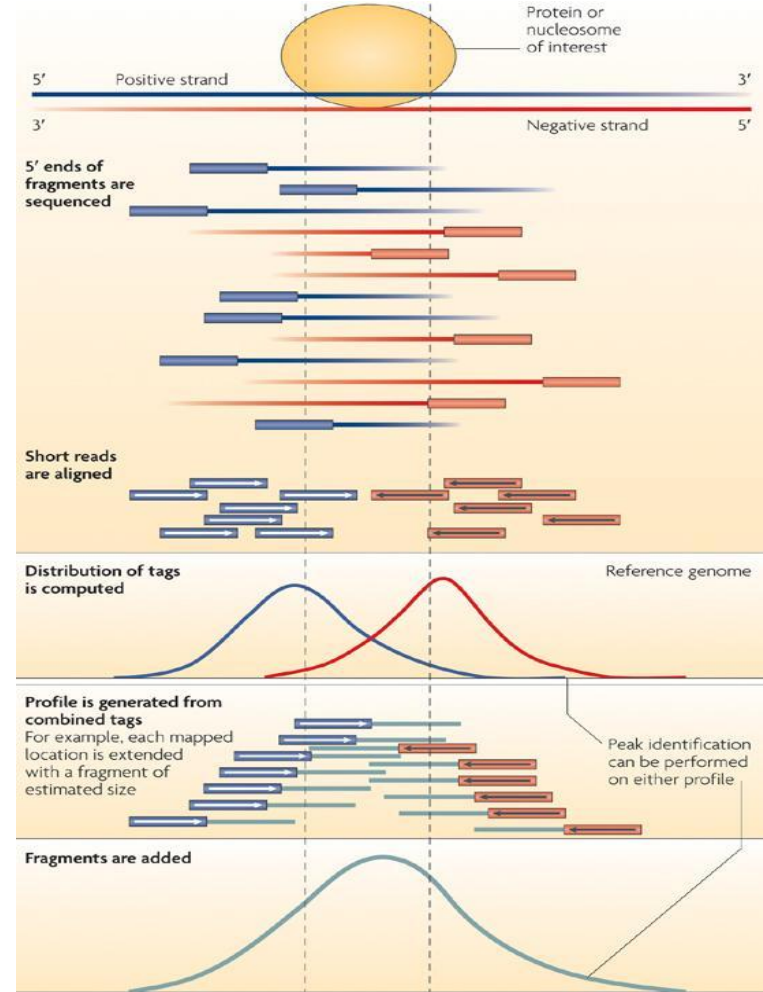
# From reads to peaks

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



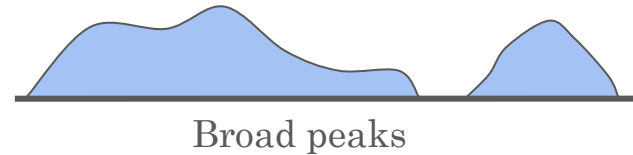
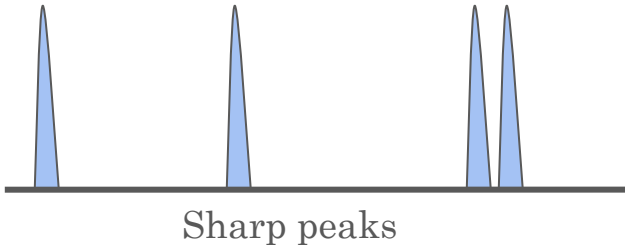
# From reads to peaks

- 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,...



# A variety of peak callers

- 60 programs listed on OMICTOOLS
- Most support a control

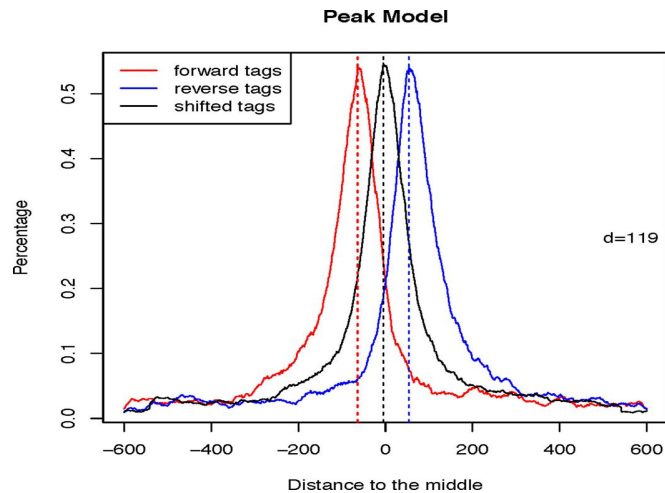
The screenshot displays the OMICTOOLS website interface. At the top, the OMICTOOLS logo is on the left, and a search bar with the text "Find the best of bioinformatics" is on the right. Below the header, a navigation breadcrumb shows "HIGH-THROUGHPUT SEQUENCING > CHIP-SEQ ANALYSIS > PEAK CALLING". The main content area is titled "PEAK CALLING SOFTWARE TOOLS | CHIP SEQUENCING DATA ANALYSIS". A sub-header reads "Identification of genomic regions of interest in ChIP-seq data, commonly referred to as peak-calling, aims to find the locations of transcription factor binding sites, modified histones or nucleosomes. Source text: (Cairns et al., 2011) BayesPeak-an... Read more". Below this is a "FILTERS" section. The main list features five tool cards, each with a gear icon, a "Desktop" label, and a brief description:

- MACS**: Model-based Analysis for ChIP-Seq. 5 stars (1 review), 1 discussion, 4 favorites. Description: "A software to analyze data generated by short read sequencers. MACS empirically models the shift size of ChIP-Seq tags, and uses it to improve the spatial resolution of predicted binding sites. It..."
- HOMER**: Hypergeometric Optimization of Motif EnRichment. 5 stars (5 reviews), 0 discussions, 4 favorites. Description: "A suite of tools for Motif Discovery and next-gen sequencing analysis. HOMER contains many useful tools for analyzing ChIP-Seq, GRO-Seq, RNA-Seq, DNase-Seq, Hi-C and numerous other types of..."
- SICER**: A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. 5 stars (1 review), 0 discussions, 1 favorite.
- SPP**: An R package for analysis of ChIP-seq and other functional sequencing data. SPP has been designed to detect protein binding positions with high accuracy. SPP can also examine the saturation level of...
- Scripture**: A method for transcriptome reconstruction that relies solely on RNA-Seq reads and an assembled genome to build a transcriptome ab initio. The statistical methods to estimate read coverage...

# MACS [Zhang et al, 2008]

## 1. Modeling the shift size of CHIP-Seq tags

- slides  $2 \times \text{bandwidth}$  windows across the genome to find regions with tags more than  $m\text{fold}$  enriched relative to a random tag genome distribution
- randomly samples 1,000 of these highly enriched regions
- separates their + and - reads, and aligns them by the midpoint between their + and - read centers
- define  $d$  as the distance in bp between the summit of the two distribution



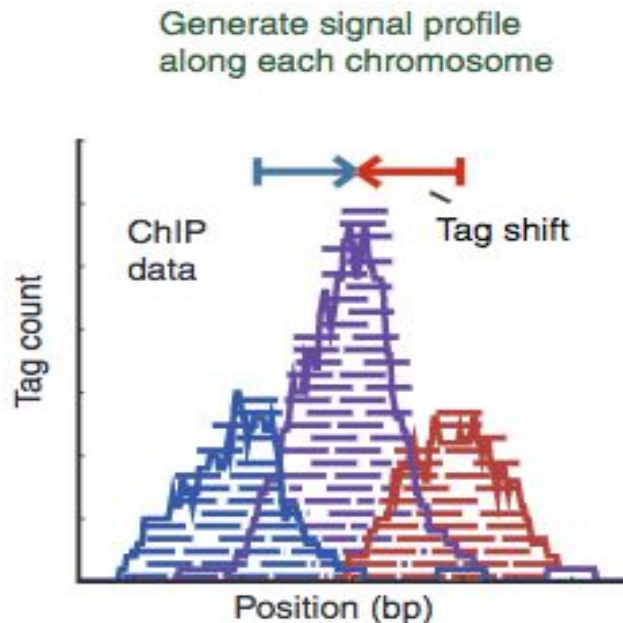


# MACS [Zhang et al, 2008]

## 2. Peak detection

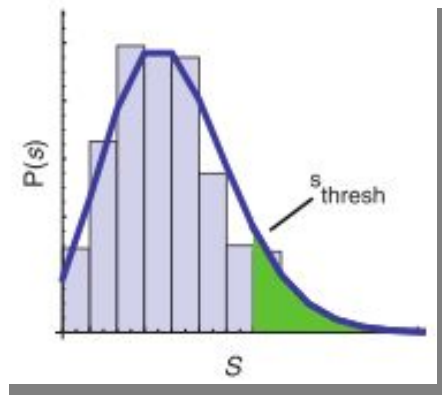
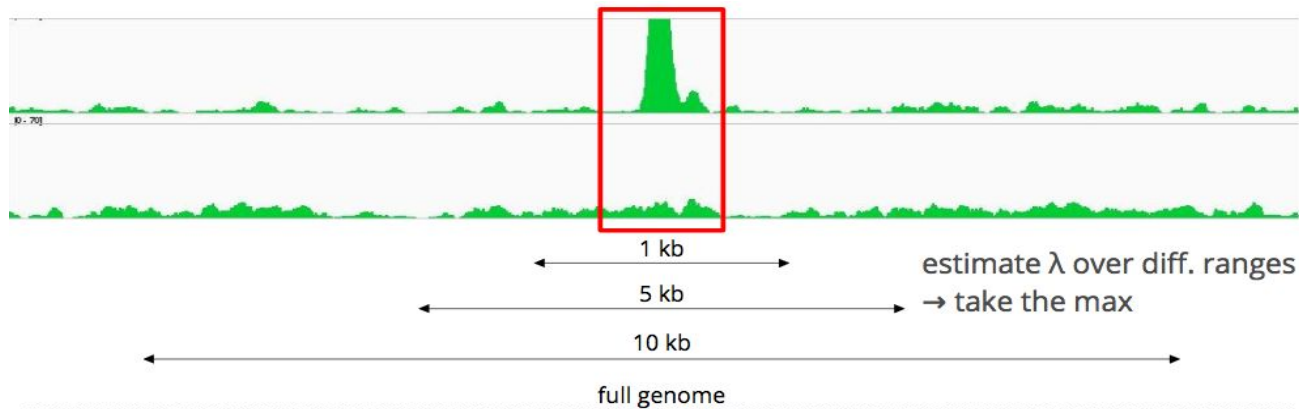
- Scales the total Input read count to be the same as the total ChIP read count
- Duplicate read removal
- Reads are shifted by  $d/2$

( $d$  value is the model obtained in step 1)



# MACS [Zhang et al, 2008]

- Slides  $2d$  windows across the genome to find candidate peaks with a significant read enrichment (Poisson distribution  $p$ -value based on  $\lambda_{BG}$ , default  $10^{-5}$ )
- Estimate parameter  $\lambda_{local}$  of Poisson distribution
- Keep peaks significant under  $\lambda_{BG}$  and  $\lambda_{local}$  and with  $p$ -value  $<$  threshold

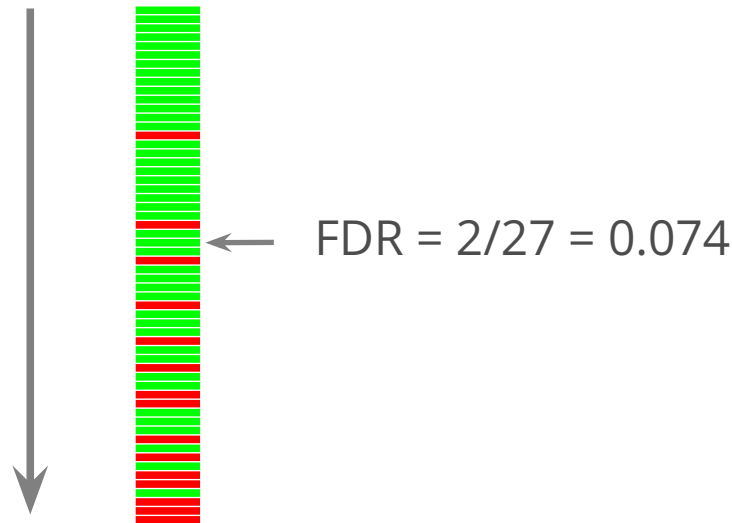


# MACS [Zhang et al, 2008]

## 3. Multiple testing correction (FDR)

- Swap treatment and input and call negative peaks
- Take all the peaks (neg + pos) and sort them by increasing p-values

$$\text{FDR}(p) = \frac{\# \text{ Negative peaks with p-value} < p}{\# \text{ Selected peaks}}$$

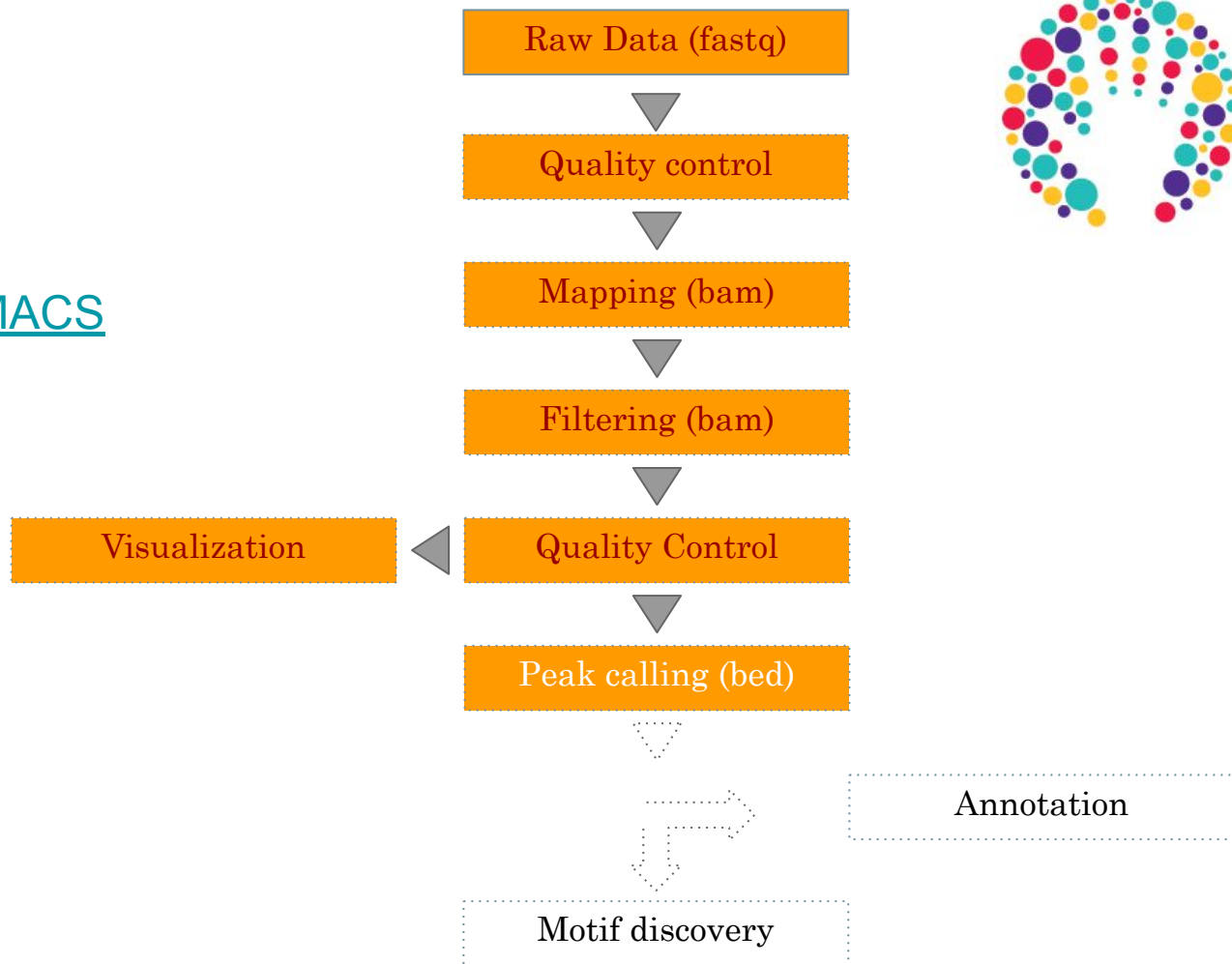


# MACS in summary

- 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](#)  
(stop after step 3)

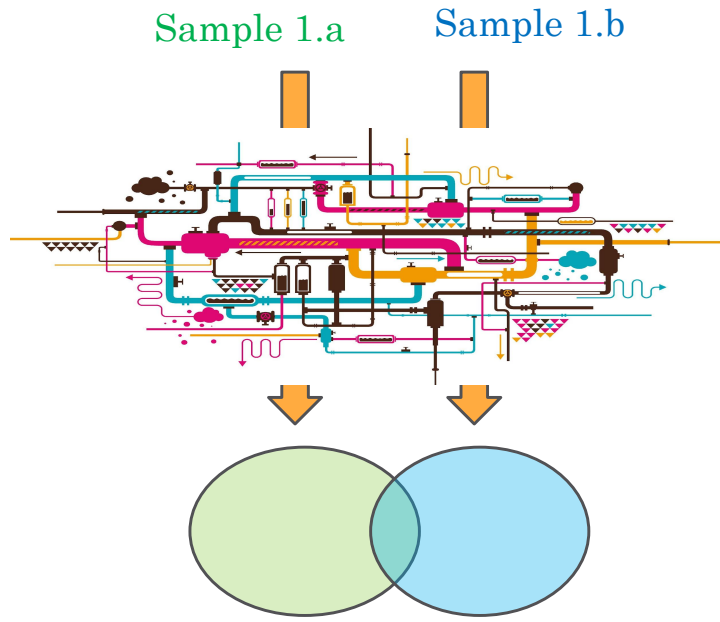




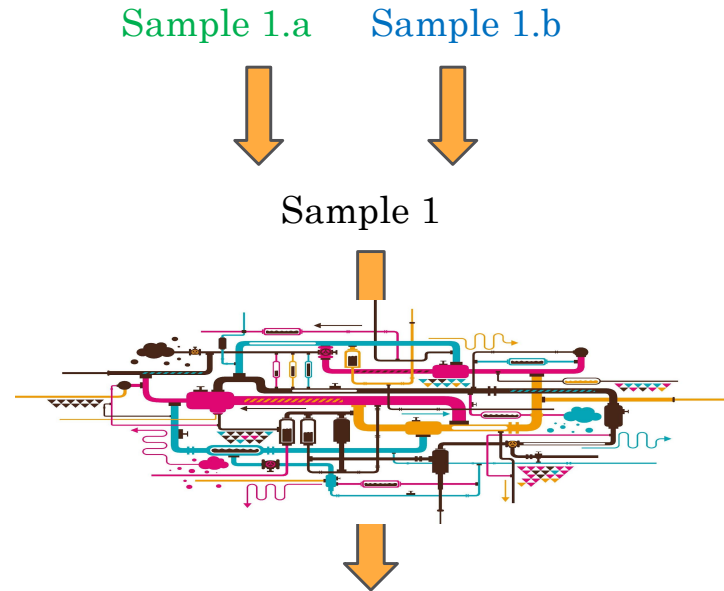
# How to deal with replicates

# How to deal with replicates

Analyze samples separately and takes union or intersection of resulting peaks



Merge samples prior to the peak calling (e.g recommended by MACS) => “pooling”





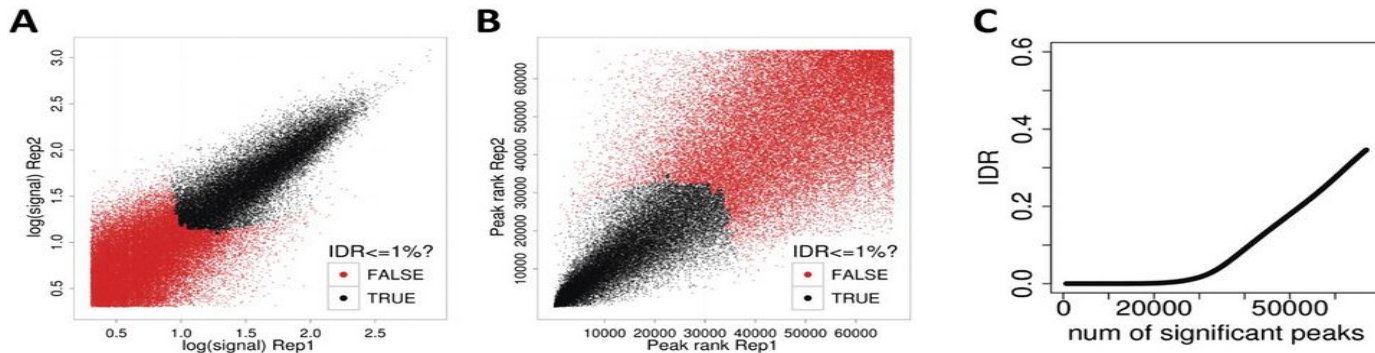
# IDR - Irreproducible Discovery Rate (ENCODE)

- Measures consistency between replicates
- Uses reproducibility in score rankings between peaks in each replicate to determine an optimal cutoff for significance.
- Idea:
  - The most significant peaks are expected to have high consistency between replicates
  - The peaks with low significance are expected to have low consistency

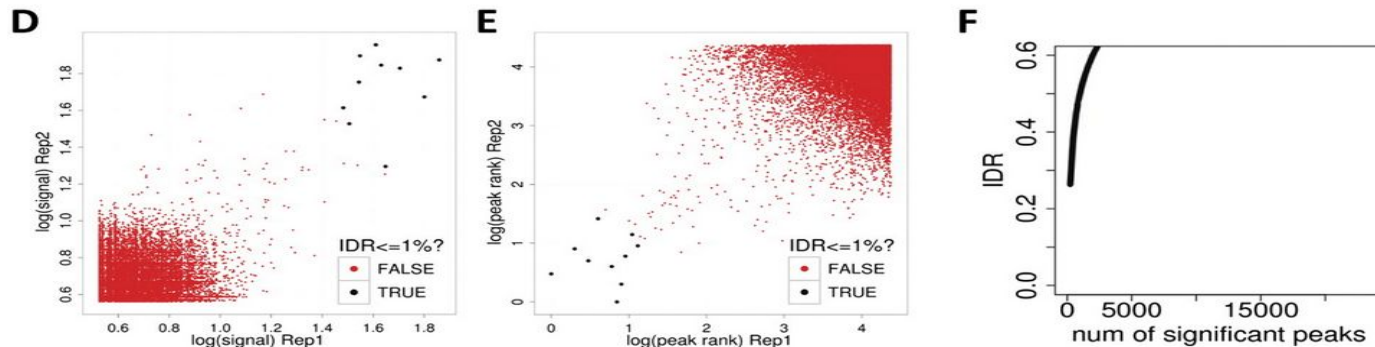


# IDR

## RAD21 Replicates (high reproducibility)



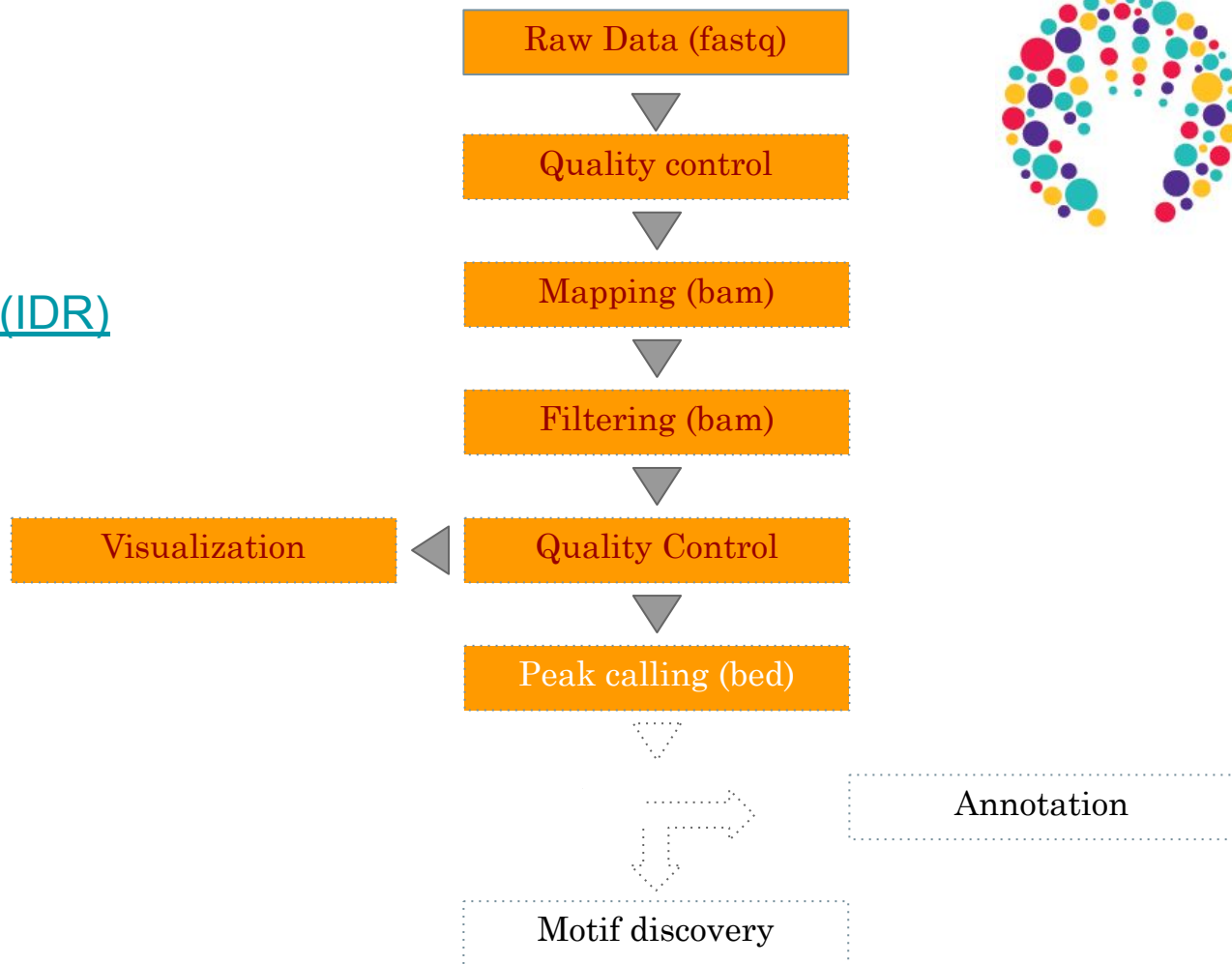
## SPT20 Replicates (low reproducibility)



(!) IDR doesn't work on broad source data!

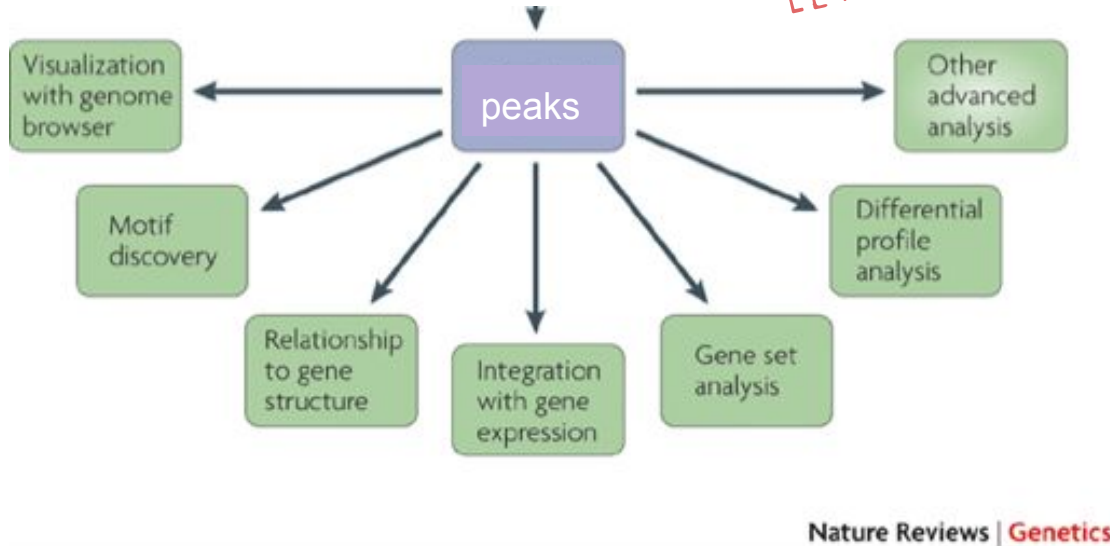
# Protocol

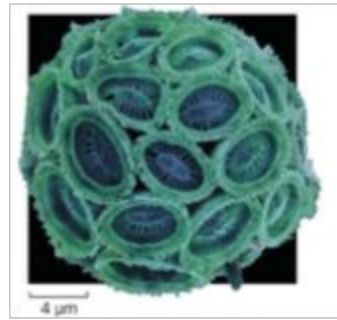
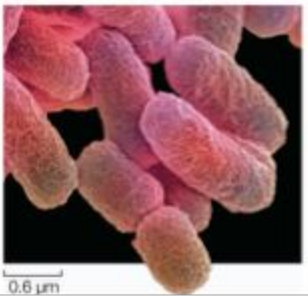
- Combine replicate (IDR)



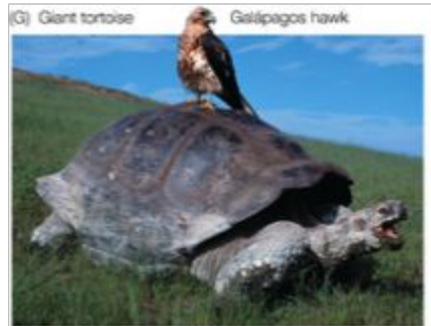
# Processing steps are over !

LET'S DO BIOLOGY !!!





**What is the biological question ?**





**What is the biological question ?**

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



**What is the biological question ?**

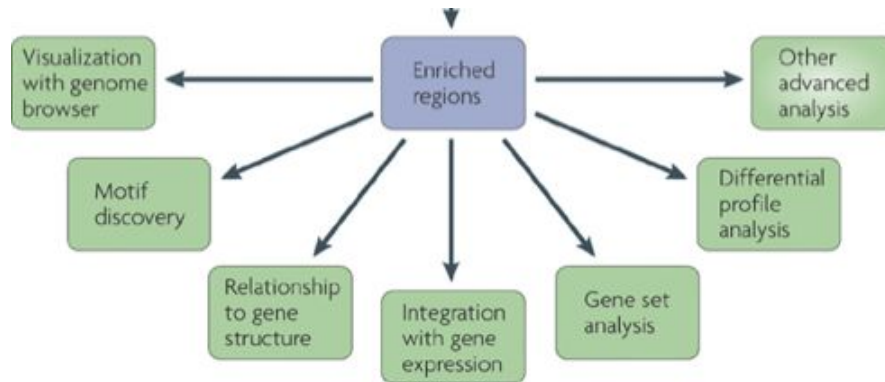
~~« see if you can find something in the data »~~

# 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) ?
- Proteomic (eg: mass spectrometry) ?
- Transgenics ?
- Will depend on
  - the organism
  - available infrastructure



# Discovering motifs in peaks

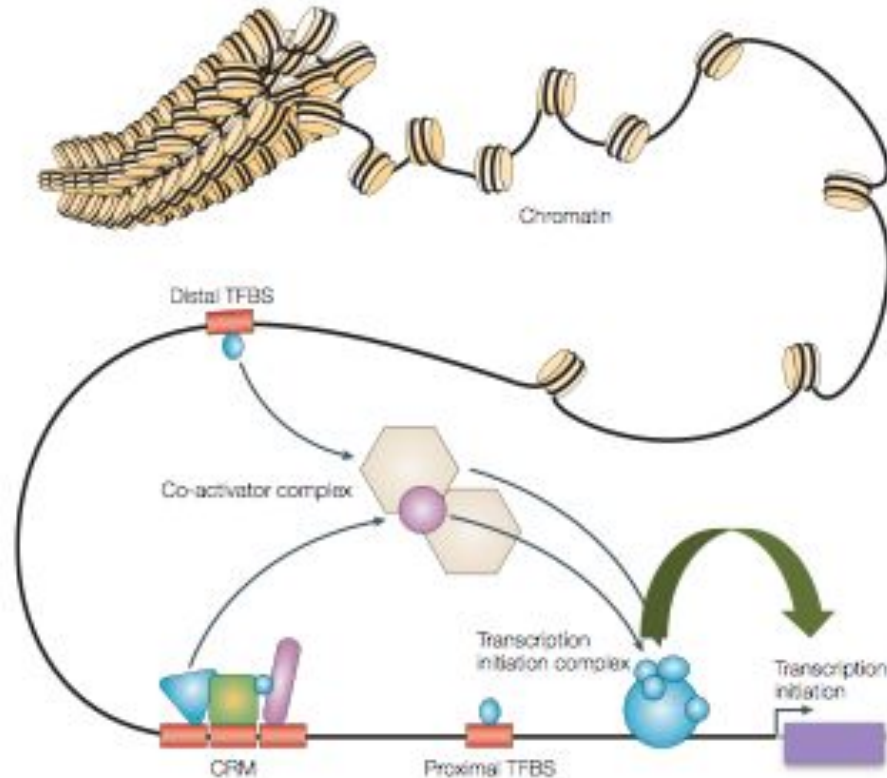


Motifs

Annotations

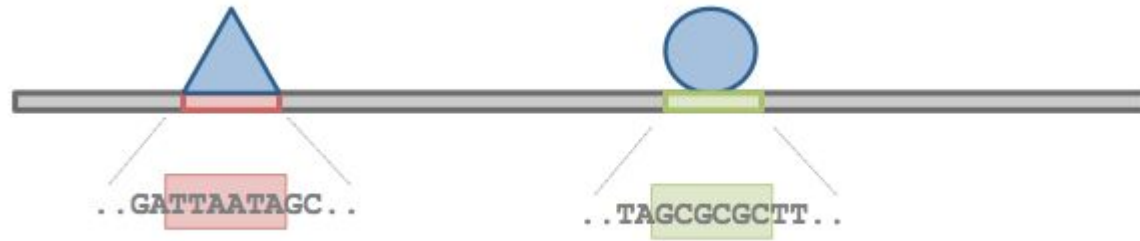
# Biological concepts of transcriptional regulation

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

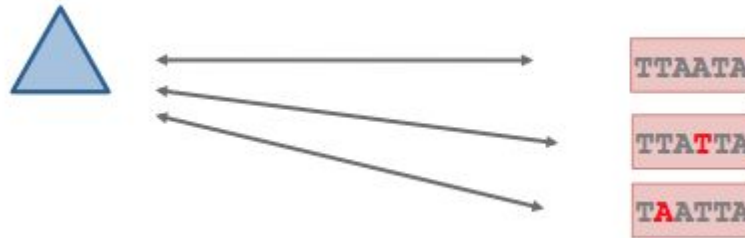


# Transcription factor specificity

*How do TF « know » where to bind DNA ?*



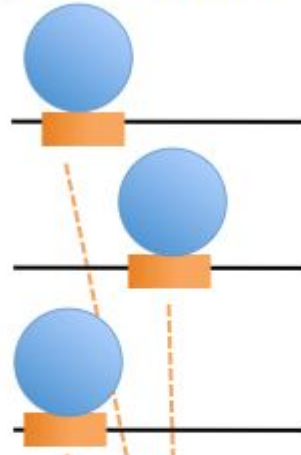
TF recognize TFBS with specific DNA sequences



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

# Binding specificity

transcription factor



cis-regulatory elements



binding motif

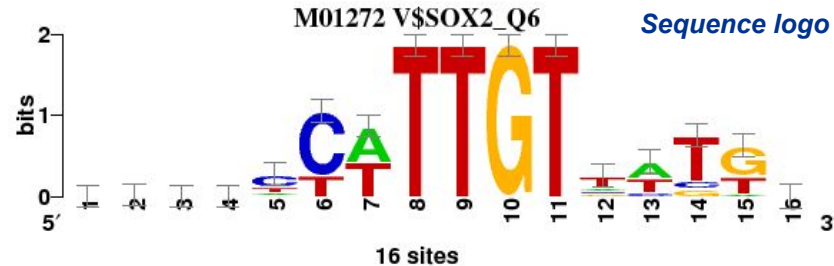
# From binding sites to binding motif

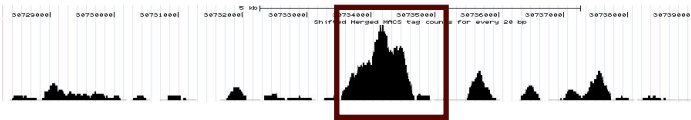
*Collection of binding sites  
used to build the Sox2 matrix  
(TRANSFAC M01272)*

R15133 GCCCTCATTGTTATGC  
 R15201 AAACCTCTTTGTTTGGA  
 R15231 TTCACCATTGTTCTAG  
 R15267 GACTCTATTGTCTCTG  
 R16367 GATATCTTTGTTTCTT  
 R17099 TGCACCTTTGTTATGC  
 R19276 AATTCACATTGTTATGA  
 R19367 AAACCTCTTTGTTTGGA  
 R19510 ATGGACATTGTAATGC  
 R22342 AGGCCTTTTGTCCCTGG  
 R22344 TGTGCTTTTGTNNNNN  
 R22359 C'TCAACTTTGTAATTT  
 R22961 GCAGCCATTGTGATGC  
 R23679 CACCCCTTTGTTATGC  
 R25928 TTTTCTATTGTTTTTA  
 R27428 AAAGGCATTGTGTTTC

*Position-specific scoring matrix (PSSM)*

<b>A</b>	6	7	4	4	2	0	8	0	0	0	0	2	7	0	1	4
<b>C</b>	2	2	6	5	9	12	0	0	0	0	0	2	2	2	0	6
<b>G</b>	4	3	2	4	1	0	0	0	0	16	0	2	0	2	9	3
<b>T</b>	4	4	4	3	4	4	8	16	16	0	16	9	6	11	5	2





ChIP-seq peaks

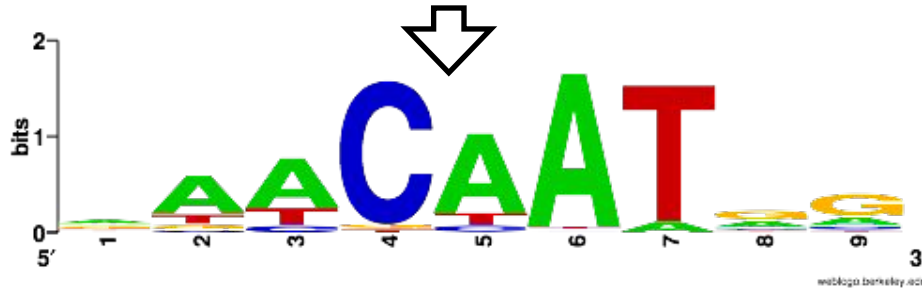
```

>mm9_chr1_39249116_39251316_+
gagaggaagggggagaaagagggagggggagGGTGATAGGTAGCCAGGAG
CCAATGGGGGCGTTTTCTTGTCCAGGCCACTGCTGGAATGTGAGATGT
AGAATGACCCAAAGAGAGCTGCCAAGACAGAGCTCTGCCCCAGGAATTGA
ACTCAAAGGTGTCAGAAAGCAGGTGGCCTTTGTGCACCTGGCGCGGGGA
CGTGGCTCCCTCTTCCGGCTGGTCTAGCCAGGtgccctgccctgccctgcc
gccGTGATCTCTGGACGCCAGTAGAGGGTTGTTGTGGGTTTGGGTGAAAC
ACGCCACCCCTGAGCTCTTCCGCGGGGCTAGCAATCTCCCCATCACCCCA
TTCGCGCTCAGAACCCCTCAGCGGATAACAGCAGGCCTGGTTCCCCG
  
```

DNA sequence

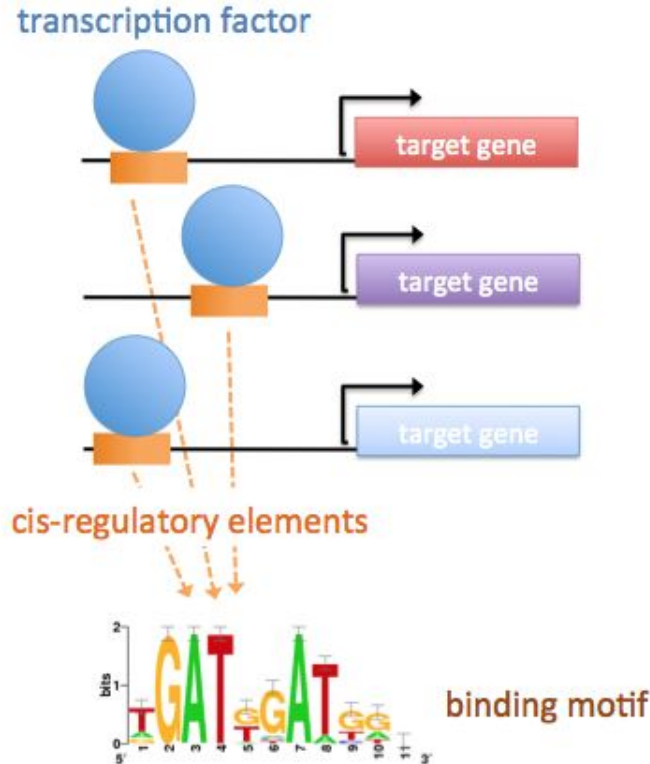
A	[24	54	59	0	65	71	4	24	9]
C	[7	6	4	72	4	2	0	6	9]
G	[31	7	0	2	0	1	1	38	55]
T	[14	9	13	2	7	2	71	8	3]

Discovered motif



Motif logo

# De novo motif discovery



*Problem :*

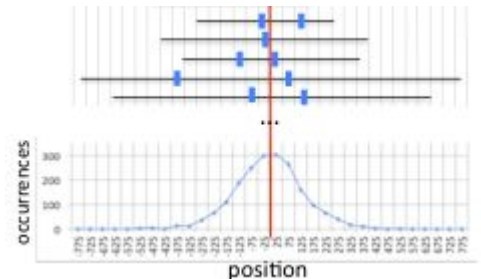
*How can we model/describe  
the binding specificity of  
a given TF ?*

*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

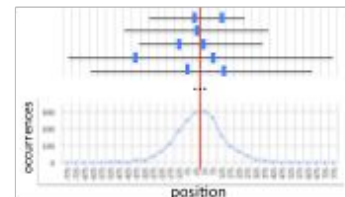
*Why do we need new approaches for genome-wide datasets ?*

# New approaches for ChIP-seq datasets

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



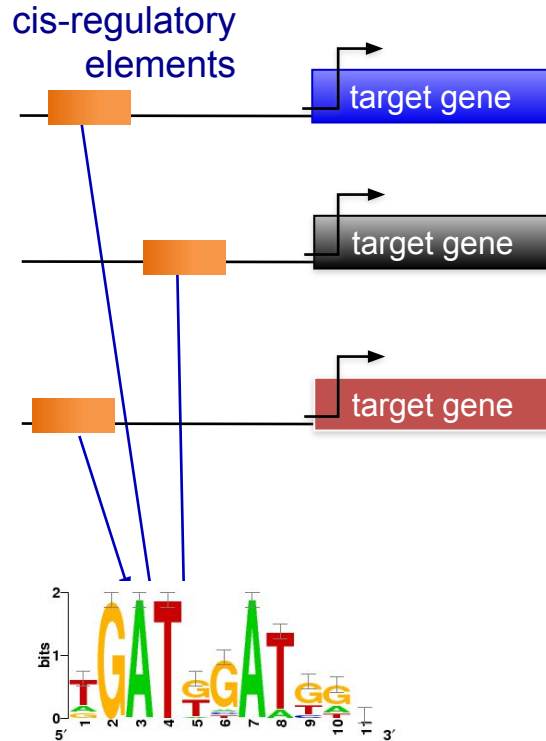
- **the problem is slightly different**
    - promoters: 200-2000bp from co-regulated genes
- ↓
- peaks: 300bp, positional bias



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

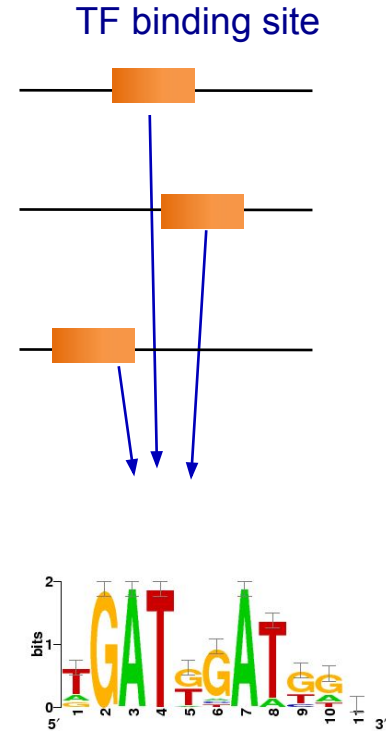
# De novo motif discovery

Case 1: promoters of co-expressed genes



binding motif  
(represented as a  
sequence logo)

Case 2: ChIP-seq peaks



# Regulatory sequence Analysis Tools (rsat.eu)

## Regulatory Sequence Analysis Tools

Welcome to **Regulatory Sequence Analysis Tools (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.



maintained by TAGC - Université Aix Marseilles, France



maintained by RegulonDB - UNAM, Cuernavaca, Mexico



maintained by plateforme ABIMS Roscoff, France



maintained by Ecole Normale Supérieure Paris, France



maintained by Bruno Contreras Moreira, Spain



maintained by SLU Global Bioinformatics Center, Uppsala, Sweden

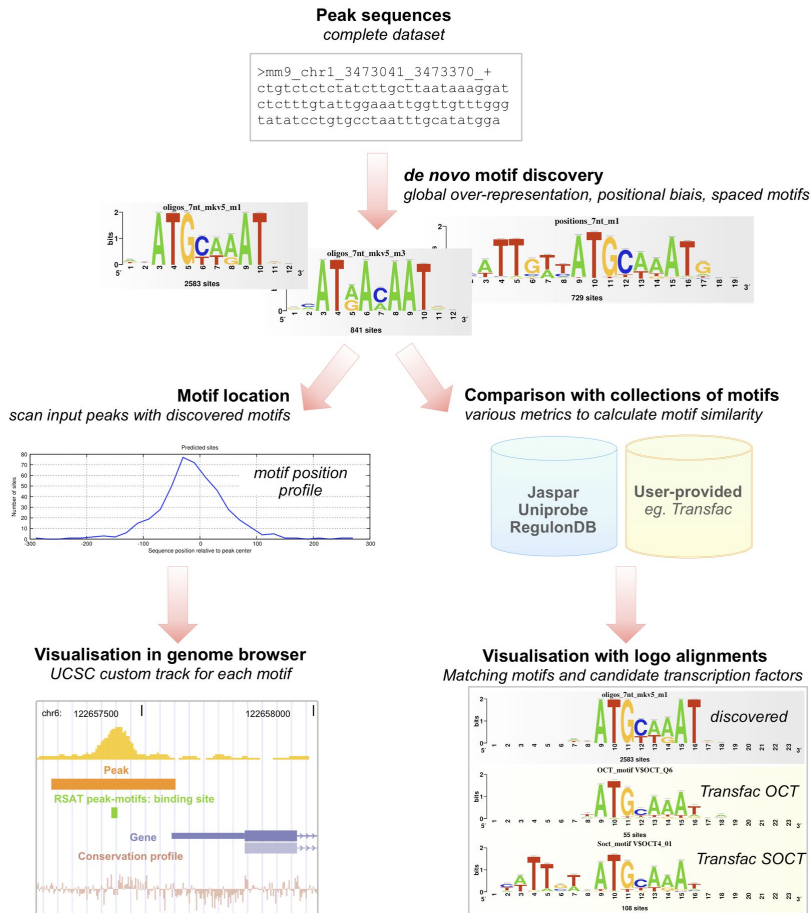
### Citing RSAT complete suite of tools:

- 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.

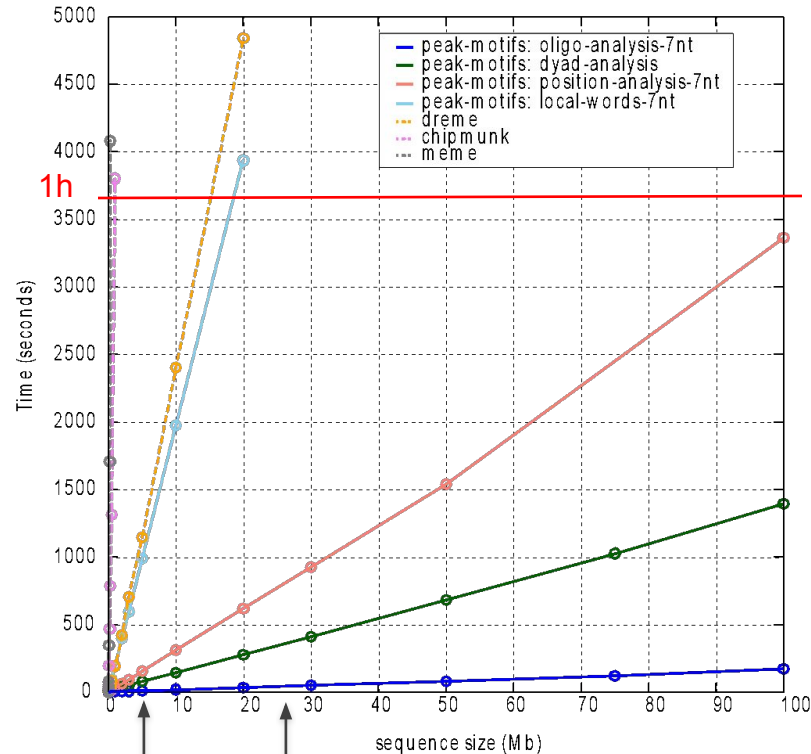
# Peak-motifs

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



# Peak-motifs: why providing yet another tool?

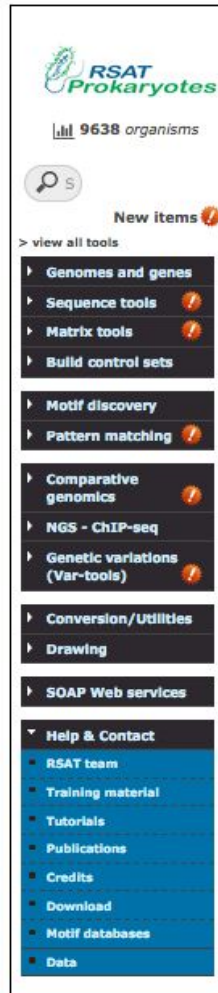
- fast and scalable
- treat full-size datasets
- using 4 complementary algorithms
  - Global over-representation
    - oligo-analysis
    - dyad-analysis (spaced motifs)
  - Positional bias
    - position-analysis
    - local-words



size limit of other websites

typical ChIP-seq dataset

# RSAT menu



→ 1. Get sequences

→ 2. Run the analysis

→ 3. Visualization

→ Help: tutorials,



# RSAT Web forms

**RSA-tools - retrieve sequence**

Tool name

Returns upstream, downstream or ORF sequences for a list of genes

Tool description

Remark: If you want to retrieve sequences from an organism that is in the [Ensembl](#) database, we recommend to use the [retrieve-ensembl-seq](#) program instead

Single organism Organism

Multiple organisms

Genes  all  selection

Upload gene list from file

Query contains only IDs (no synonyms)

Feature type  CDS  mRNA  tRNA  rRNA  scRNA

Sequence type  From  To

Prevent overlap with neighbour genes (noorf)

Mask repeats (only valid for organisms with annotated repeats)

Admit imprecise positions

Sequence format

Sequence label

Tool parameters

Output

Go button (launches the analysis)

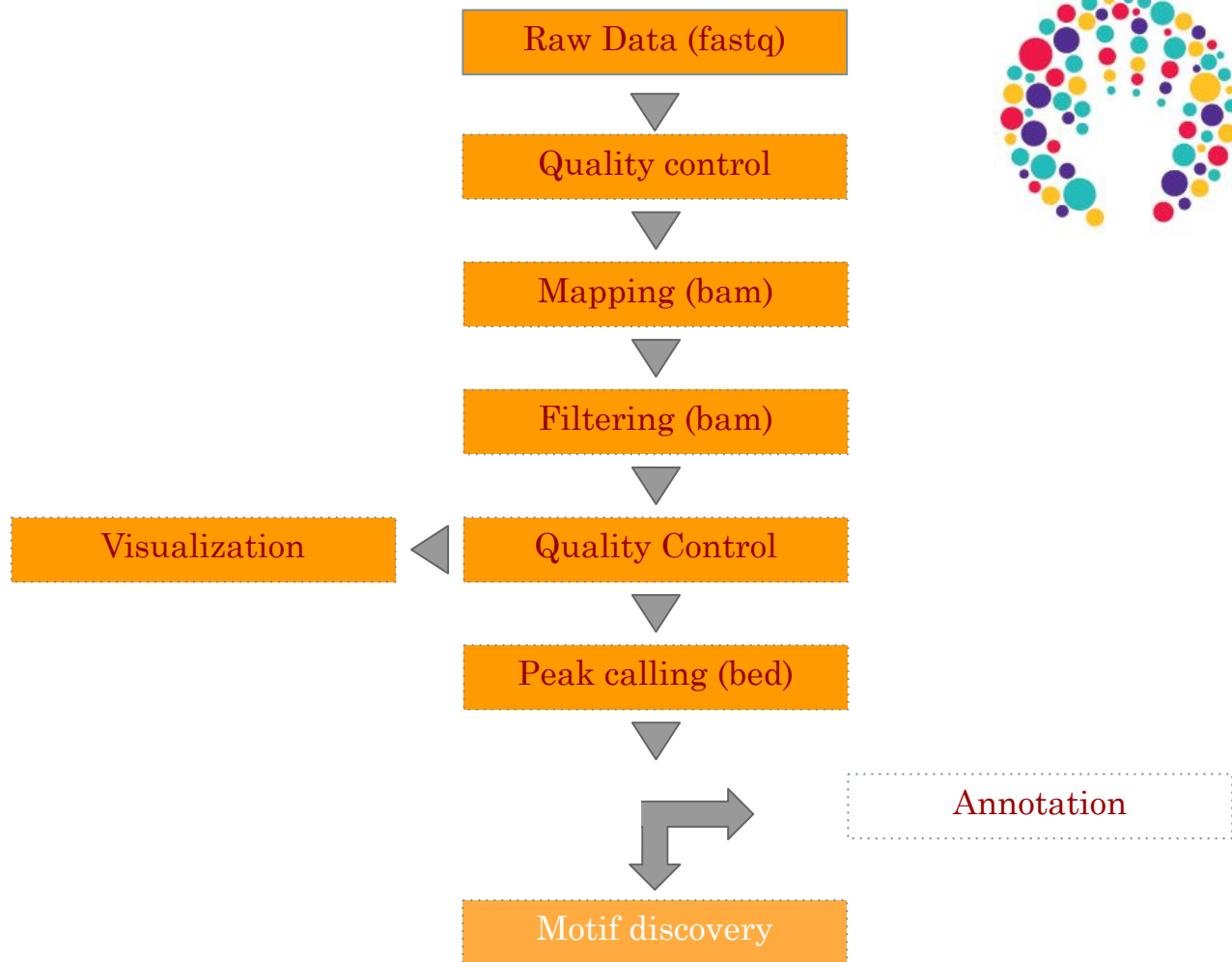
Demo button (fill in the form for test purposes)

[MANUAL TUTORIAL](#) [MAIL](#)

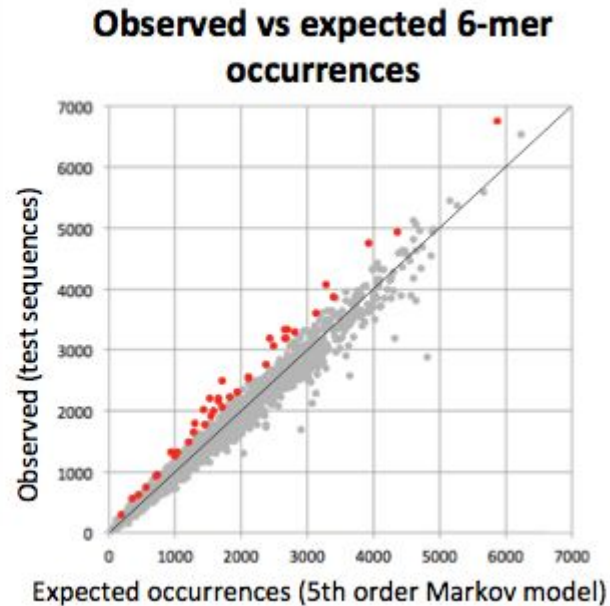
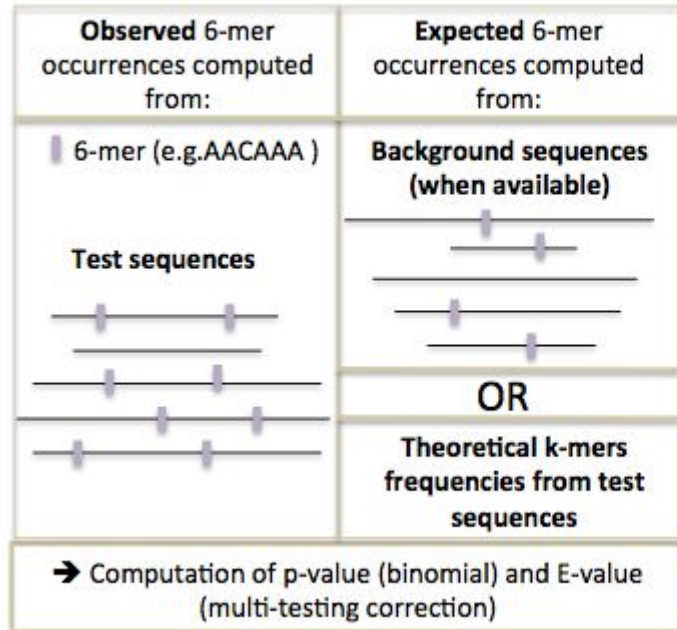
Help

# Protocol

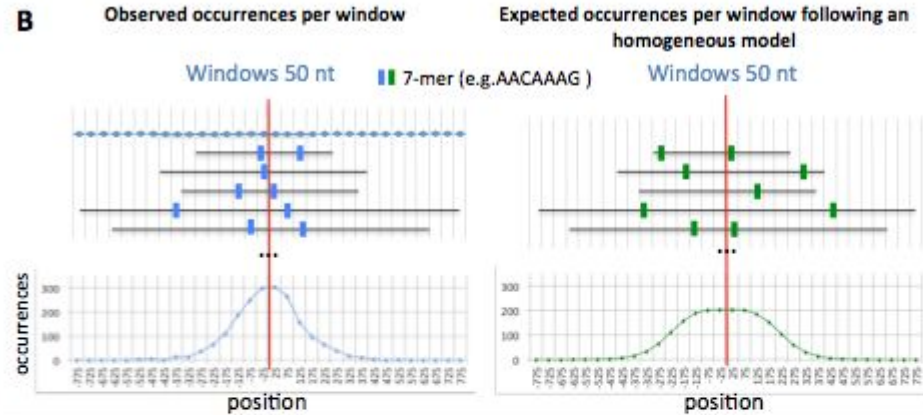
- Motif analysis



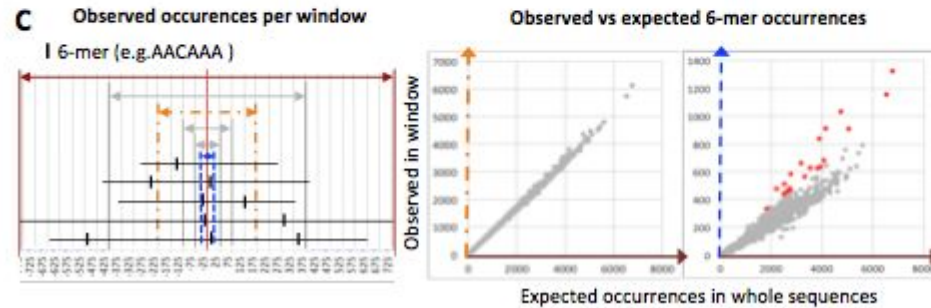
# Motif discovery: frequency



# Motif discovery: positional bias



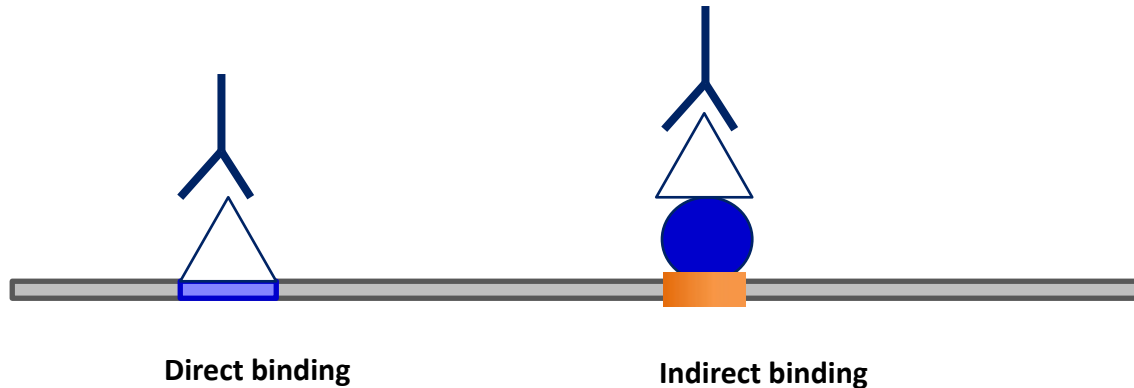
position-analysis



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!



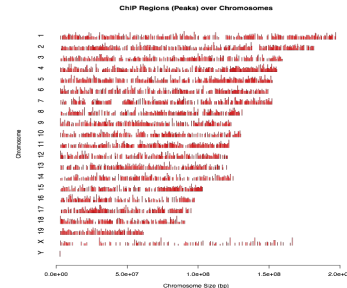
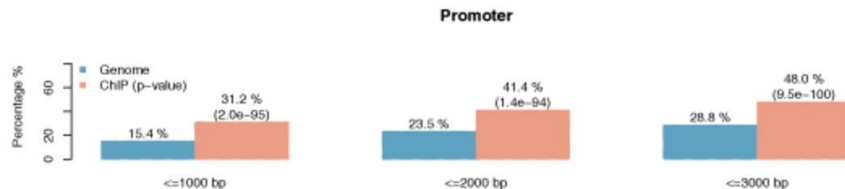


# Annotating peaks

**Annotations**

# 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?





# Various tools available

- CHIPseeker (Bioconductor) <https://goo.gl/BemEsw>
- bedtools annotate : <http://bedtools.readthedocs.io>
- HOMER annotatePeaks.pl

Warning : rely on the organism annotation and assembly version

=> not all organisms supported by all programs !



# Which are the closest genes?



## **HOMER**

Software for motif discovery and ChIP-Seq analysis

---

HOMER is a well-maintained suite of tools for functional genomics sequencing data sets. It can perform peak-calling and motif analysis, but we will use it for annotation of the peaks only.

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

ChIP-seq peaks

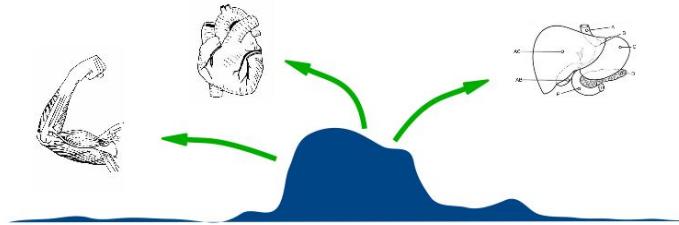


Genes



Ontology terms

GO Molecular Function  
GO Biological Process  
Disease Ontology  
Pathways  
...



# Various tools available

## These tools work with gene lists

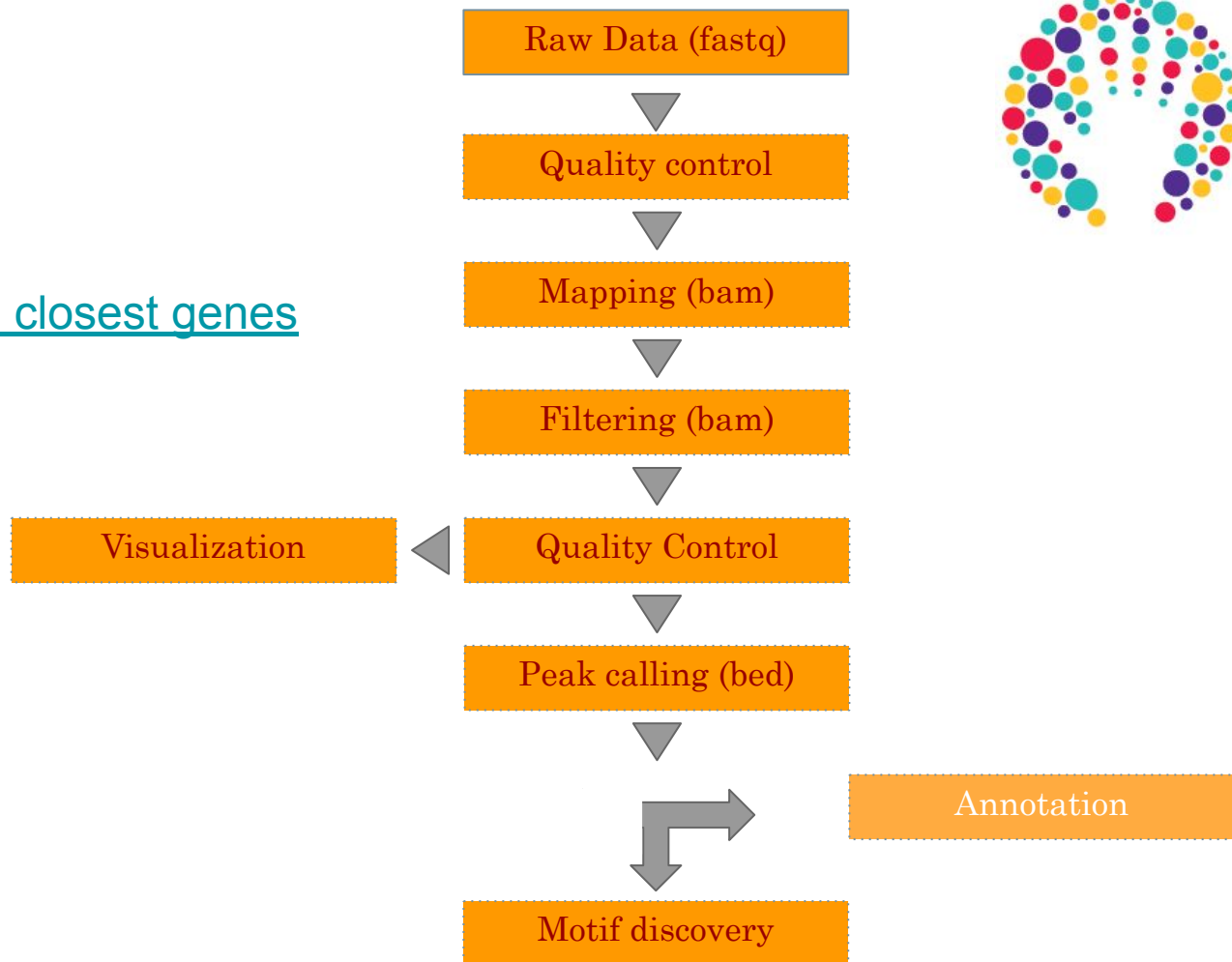
- **GSEA:** <http://www.broadinstitute.org/gsea>
- **gProfiler:** <http://biit.cs.ut.ee/gprofiler/gost>
- **GSA:** <https://gsan.labri.fr/>
- **HOMER:** <http://homer.salk.edu/homer>
- **DAVID:** <http://david.abcc.ncifcrf.gov>

## These tools work with regions (BED files)

- **EnrichR:** <http://amp.pharm.mssm.edu/Enrichr/enrich>
- **LOLA** (Bioconductor) <https://goo.gl/pWDZEs>
- **GREAT:** <http://great.stanford.edu/public/html/>

# Protocol

- Associate peaks to closest genes



Now that we have the genes,

Are there some functional categories over-represented ?

ChIP-seq peaks



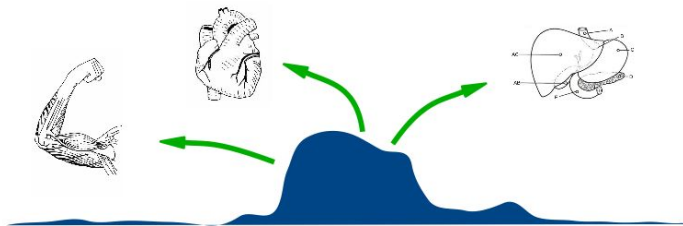
Genes



Ontology terms

GO Molecular Function  
GO Biological Process  
Disease Ontology  
Pathways

...



# HOMER [Heinz et al., Mol Cell 2010]

## Gene Ontology Analysis of Associated Genes: annotatePeaks go option

P-value	LogP	Term	GO Tree	GO ID	# of Genes in Term	# of Target Genes in Term	# of Total Genes	# of Target Genes	Common Genes
2.912e-26	-5.880e+01	immune response	biological process	GO:0006955	349	35	18091	168	Il10,Cd14,Malt1,Ccl2,Ccl7,Ifih:
3.912e-26	-5.850e+01	immune system process	biological process	GO:0002376	679	45	18091	168	S100a9,Egr1,Il10,Cd14,Malt1,C
1.823e-25	-5.696e+01	cytokine activity	molecular function	GO:0005125	178	27	18371	167	Gdf15,Il10,Csf2,Ccl9,Ccl2,Ccl7
3.372e-23	-5.174e+01	defense response	biological process	GO:0006952	430	35	18091	168	Il10,Cd14,Malt1,Nupr1,Ccl2,Cc

## Genome Ontology: Looking for Enriched Genomic Annotations: annotatePeaks genomeOntology option

Total Input Regions (0.936444051404582.pos): 25961, 33798473 bp

P-value	Log P-value	Annotation	Ann Group	#features	Coverage(bp)	AvgFeatureSize[ref=1301]	Overlap(#peaks)	Overlap(bp)	Expected Overlap(bp, gsize=2.00e+09)	Log Ratio Enrichment	Log P-value(+ underrepresented)	P-value
1e-3660	-8428.50	cpgIsland	basic	28691	21842742	761	9426	5545349	369125	2.71	-8428.50	0.00e+00
1e-2673	-6155.47	promoters	basic	44477	30002652	674	8579	5141062	507021	2.32	-6155.47	0.00e+00
1e-1027	-2363.56	utr5	basic	57703	5423448	93	7002	1516346	91652	2.81	-2363.56	0.00e+00
1e-381	-876.84	exons	basic	503529	73292986	145	9092	3171853	1238595	0.94	-876.84	0.00e+00
1e-341	-783.19	protein-coding	basic	483461	66805131	138	8609	2876255	1128955	0.94	-783.19	0.00e+00
1e-105	-239.83	coding	basic	407555	43508461	106	6592	1482965	735259	0.70	-239.83	6.94e-105
1e-71	-162.42	GC_rich Low_complexity Low_complexity	repeats	13724	552081	40	2716	120042	9329	2.55	-162.42	2.89e-71
1e-60	-136.15	miscRNA	basic	11332	4544003	400	592	287477	76790	1.32	-136.15	7.46e-60
3.14e-20	-44.91	tts	basic	44477	28239519	634	1124	718919	477226	0.41	-44.91	3.14e-20
1.21e-10	-22.84	CG Simple_repeat Simple_repeat	repeats	1241	71601	57	313	15761	1210	2.57	-22.84	1.21e-10
1.06e-09	-20.66	C-rich Low_complexity Low_complexity	repeats	9534	1007297	105	673	53335	17022	1.14	-20.66	1.06e-09

# 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:
- 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:
- 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,hg38), 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.

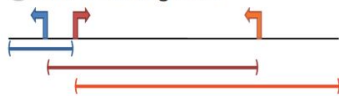
### Basal plus extension



Proximal:  kb upstream,  kb downstream, plus Distal: up to  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  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  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.

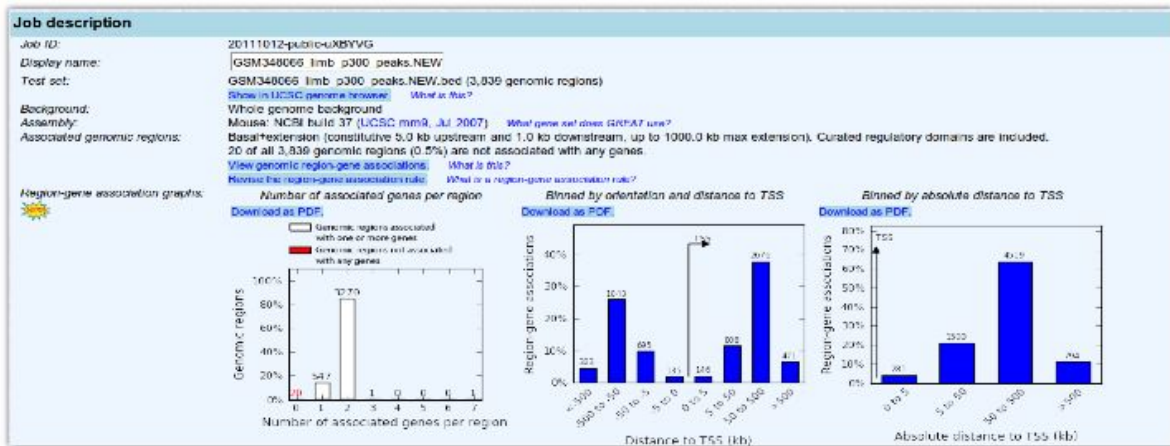
 Gene Transcription Start Site (TSS)

Note: Only human (hg19, hg38), mouse (mm9, mm10) and zebrafish (danRer7) genomes are supported



# GREAT

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



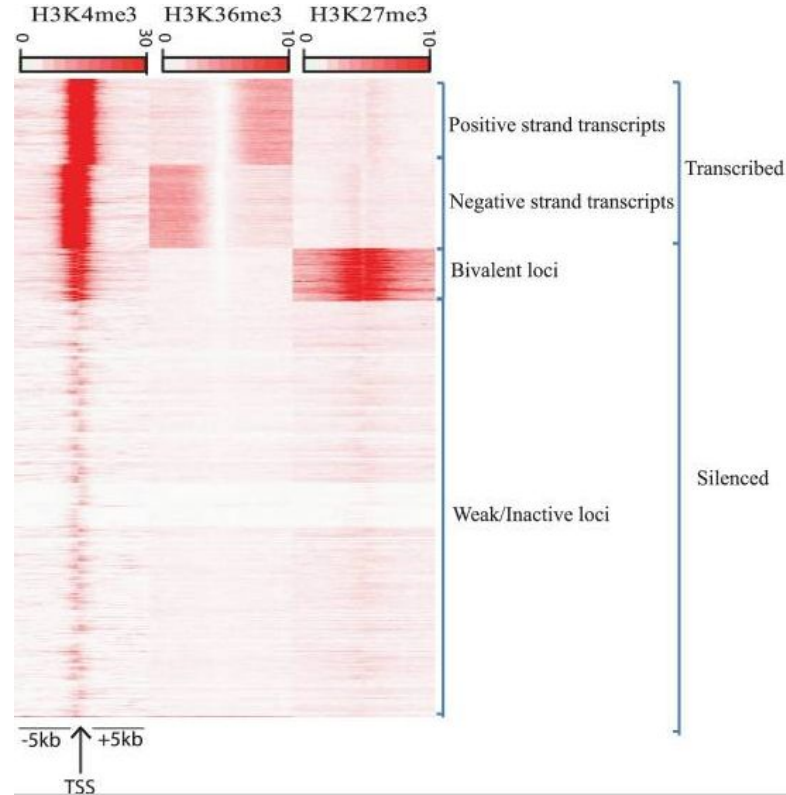
**X Mouse Phenotype** Global Controls

Table controls:  Shown top rows in this table:   Term annotation count: Min:  Max:

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
<a href="#">abnormal limbs/digits/tail morphology</a>	2	<a href="#">2.0559e-91</a>	<a href="#">6.6837e-88</a>	2.1465	780	20.32%	6	<a href="#">2.5295e-40</a>	2.2020	278	661	8.31%
<a href="#">abnormal craniofacial morphology</a>	3	<a href="#">9.3822e-91</a>	<a href="#">2.0334e-87</a>	2.0082	887	23.10%	10	<a href="#">8.9231e-36</a>	2.0382	297	786	8.88%
<a href="#">abnormal limb morphology</a>	5	<a href="#">2.4990e-80</a>	<a href="#">3.2497e-77</a>	2.3077	604	15.73%	9	<a href="#">7.4787e-37</a>	2.4541	202	444	6.04%
<a href="#">abnormal appendicular skeleton morphology</a>	10	<a href="#">3.0255e-70</a>	<a href="#">1.9672e-67</a>	2.3450	517	13.47%	17	<a href="#">3.9549e-30</a>	2.4098	172	385	5.14%
<a href="#">abnormal skeleton extremities morphology</a>	12	<a href="#">3.2687e-69</a>	<a href="#">1.7711e-66</a>	2.3724	488	13.00%	21	<a href="#">7.0557e-29</a>	2.4222	163	363	4.87%
<a href="#">abnormal paw/hand/foot morphology</a>	13	<a href="#">4.0300e-69</a>	<a href="#">2.0156e-66</a>	2.6813	404	10.52%	23	<a href="#">5.4818e-28</a>	2.7186	126	250	3.77%
<a href="#">abnormal head morphology</a>	14	<a href="#">6.4657e-67</a>	<a href="#">3.0029e-64</a>	2.0134	672	17.50%	25	<a href="#">2.9042e-27</a>	2.0982	223	585	6.67%
<a href="#">abnormal digit morphology</a>	18	<a href="#">1.0543e-61</a>	<a href="#">3.8064e-59</a>	2.6982	358	9.33%	36	<a href="#">1.2033e-25</a>	2.7998	109	210	3.26%
<a href="#">abnormal cartilage morphology</a>	23	<a href="#">7.3728e-58</a>	<a href="#">2.0843e-55</a>	2.3432	430	11.20%	29	<a href="#">1.1337e-26</a>	2.5089	140	301	4.19%
<a href="#">abnormal skeleton development</a>	24	<a href="#">3.5769e-56</a>	<a href="#">9.6904e-54</a>	2.0833	530	13.81%	38	<a href="#">5.2377e-25</a>	2.1414	185	466	5.53%
<a href="#">abnormal long bone morphology</a>	25	<a href="#">4.6593e-56</a>	<a href="#">1.2118e-53</a>	2.3374	419	10.91%	43	<a href="#">4.9983e-24</a>	2.3923	140	317	4.19%

# Other analyses

- Clustering peaks  
(Deeptools, HOMER, seqMINER)

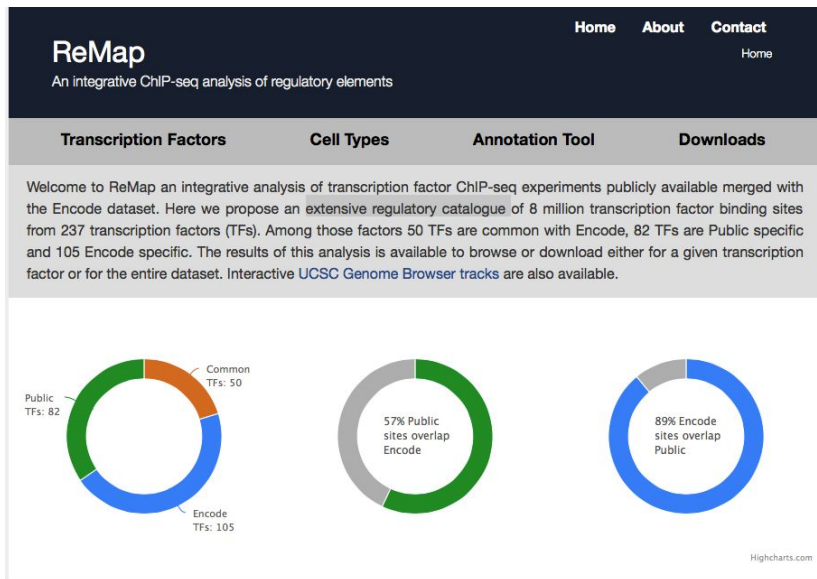


Ye et al, 2011

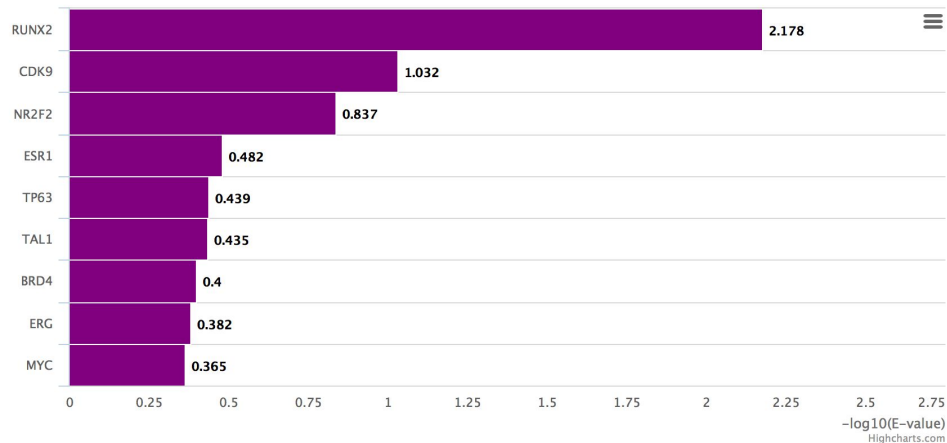
The darker the red the higher the read enrichment

# Other analyses

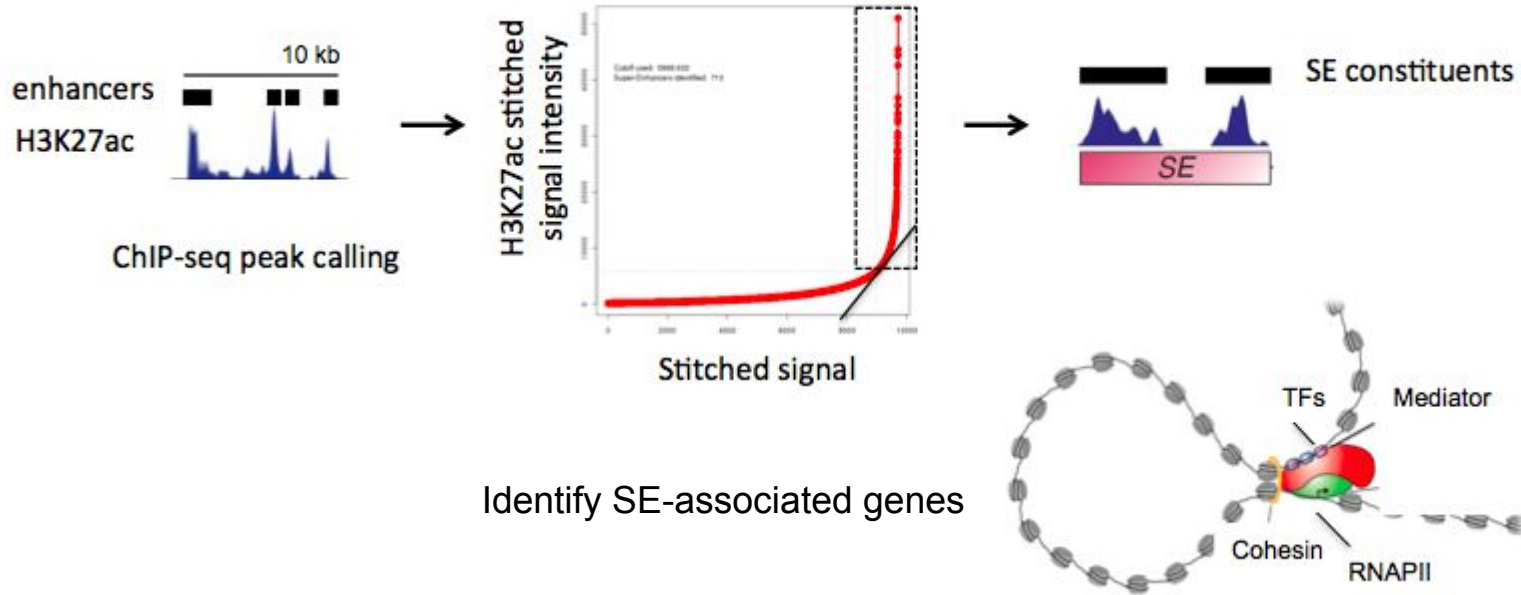
- ReMAP (<http://tagc.univ-mrs.fr/remap/>)
  - Is my peak dataset enriched for known TF peaks?



Enriched TFs in intersection

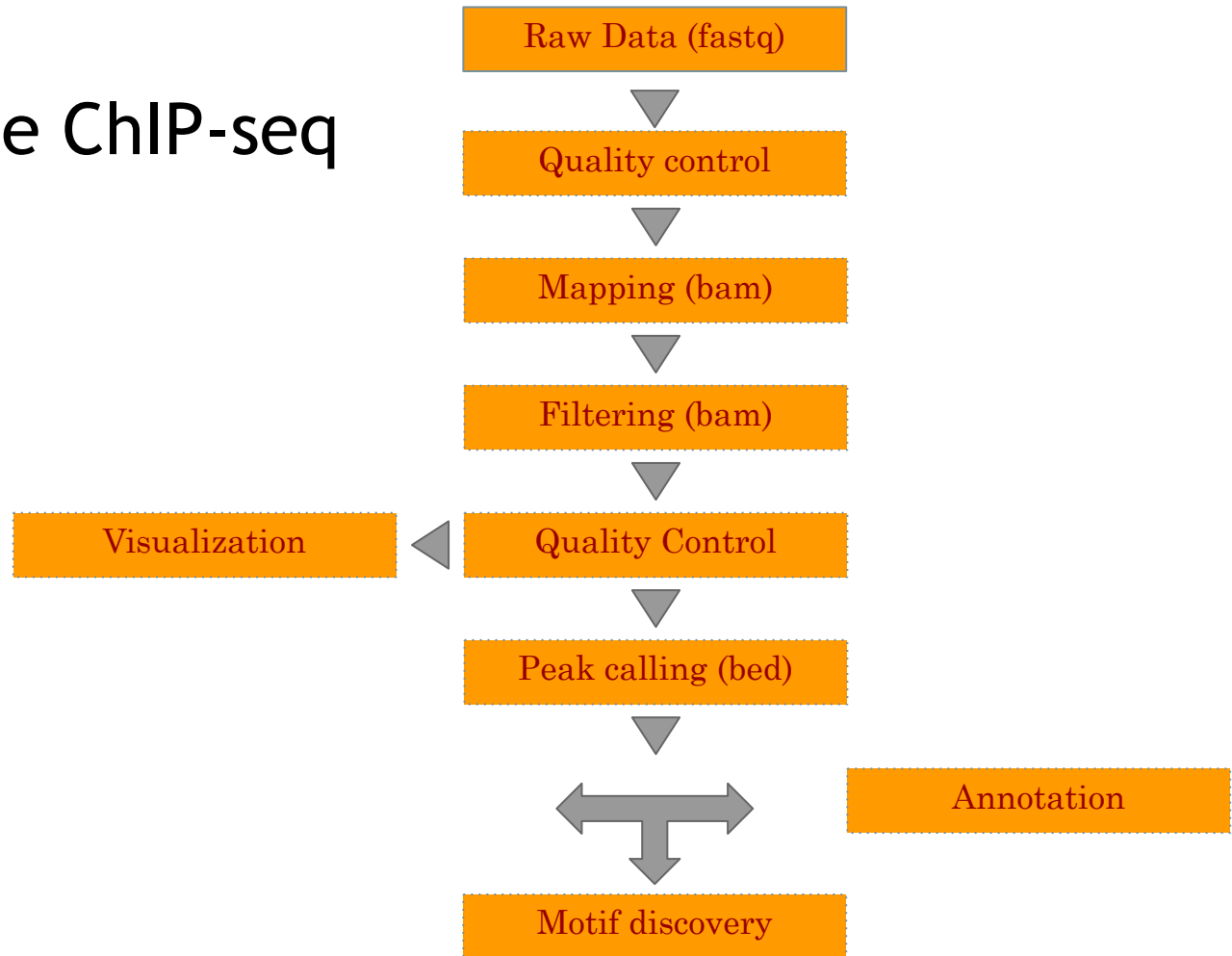


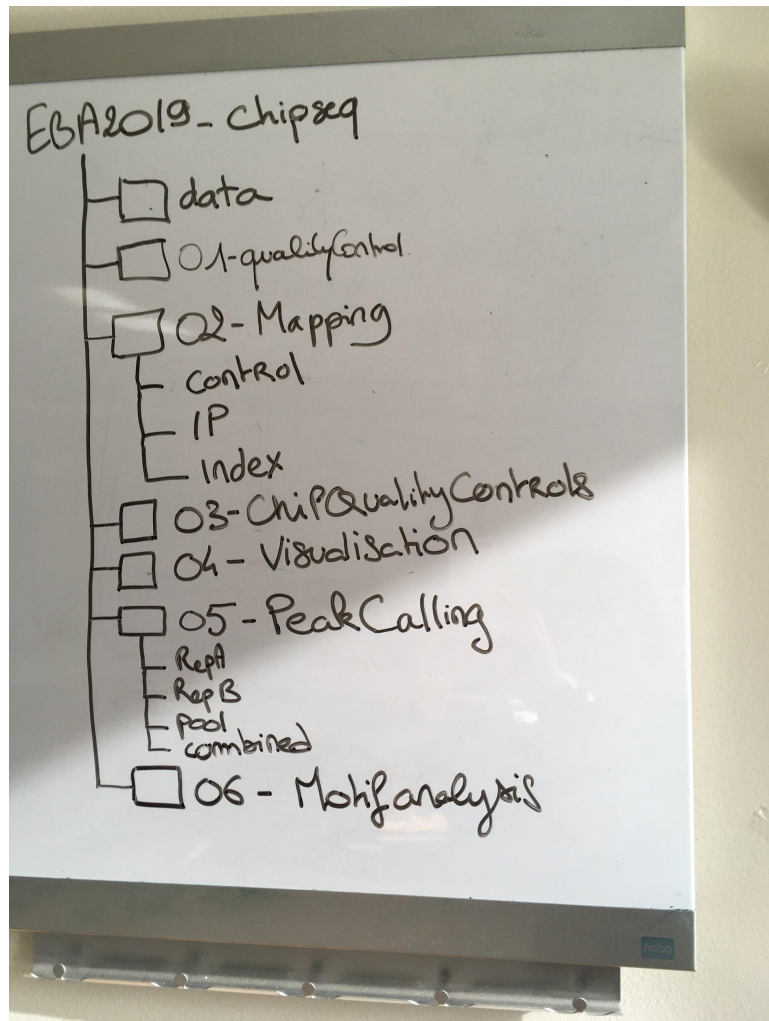
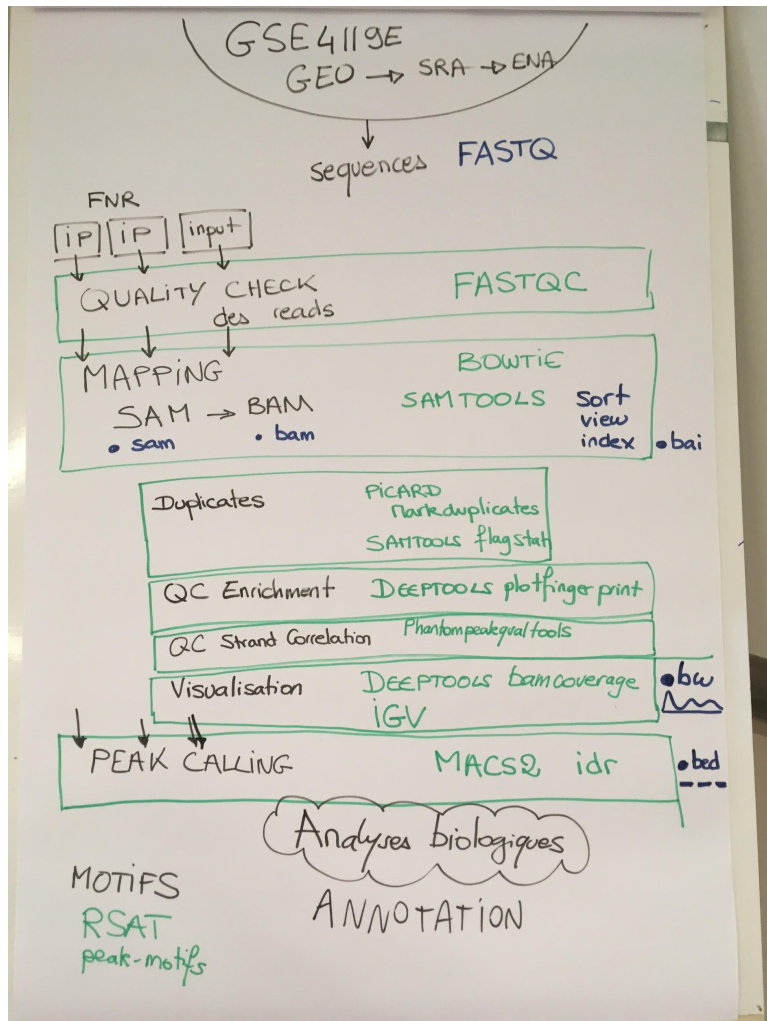
# Identification of Super-Enhancers from ChIP-seq peaks with ROSE [Loven et al., Cell 2013]



# Conclusions atelier Chip-Seq

# Bilan du pipeline ChIP-seq







# Beyond ChIP-seq : ChIP-exo



crosslink



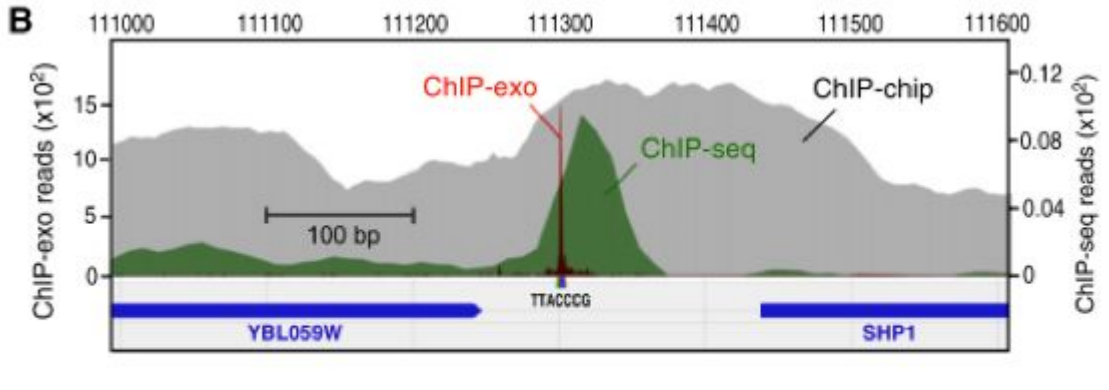
sonication



antibody



exonuclease





# Beyond ChIP-seq

## Experimental techniques



crosslink



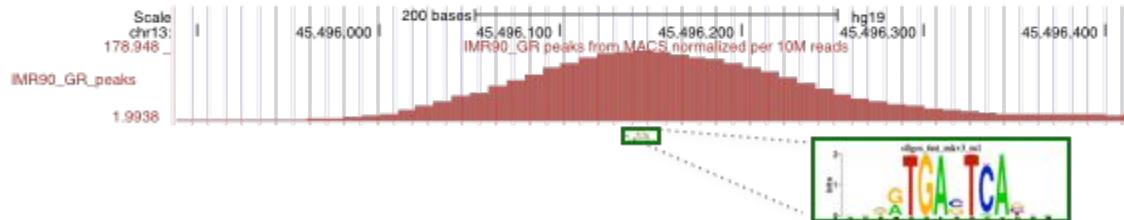
sonication



antibody

## Improvement aimed

higher resolution => 300bp to 1bp



# Beyond ChIP-seq : ChIP-nexus

## Experimental techniques



crosslink



sonication



antibody



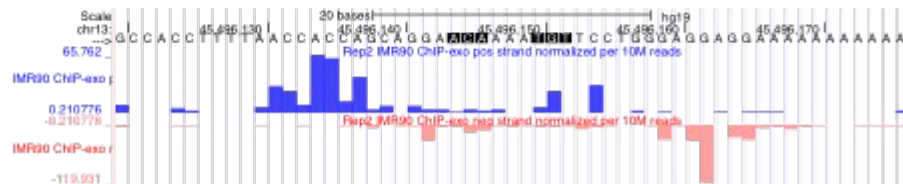
exonuclease



barcode

## Improvement aimed

Get rid of PCR artifacts



# Beyond ChIP-seq : native ChIP

## Experimental techniques



~~crosslink~~



~~sonication~~



antibody



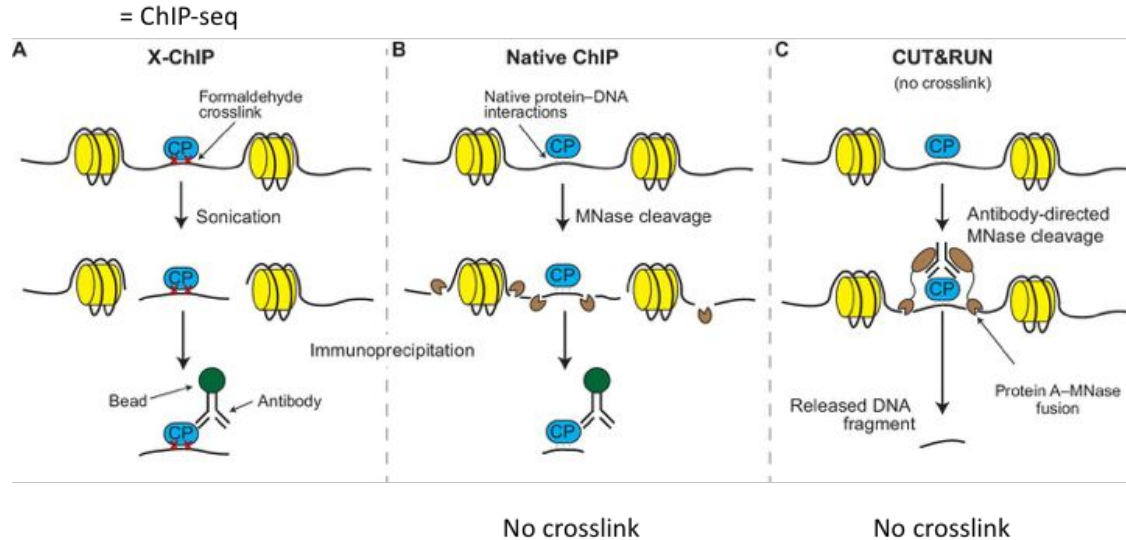
endonuclease

## Improvement aimed

Avoid formaldehyde 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



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

# 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

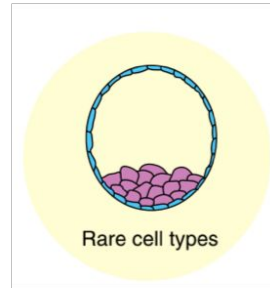
*Dahl & Gilfillan, Briefings in Functional Genomics, 2017*

- 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

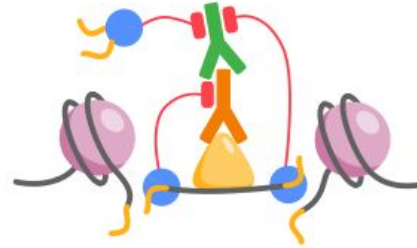
Cleavage under targets and release using nuclease



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

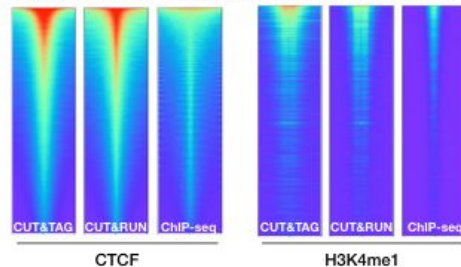
## CUT&TAG

Cleavage near targets and tagmentation



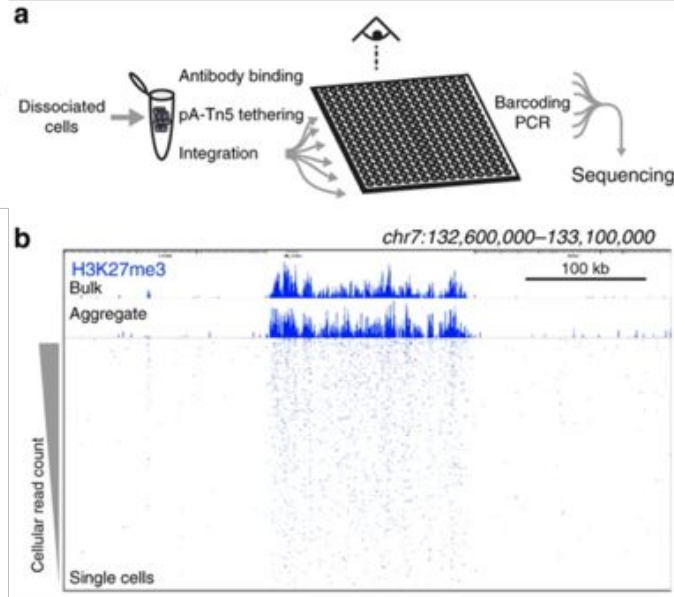
- ✦ Cleave near antibody site by Tn5
- ✦ No crosslinking
- ✦ Low background
- ✦ Include adaptor ligation
- ✦ Adapted for single-cell

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)



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

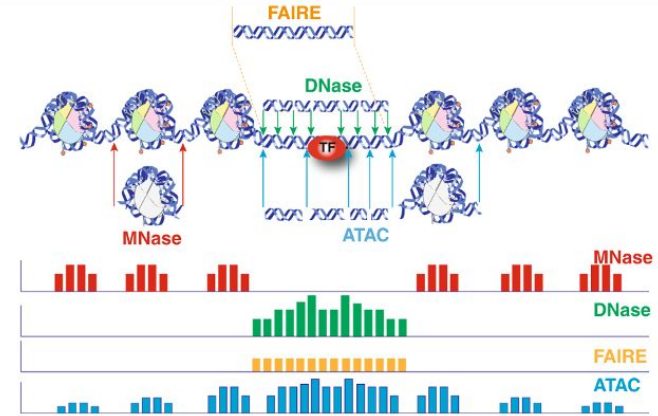
# ATAC-seq

*Assay for Transposase-Accessible Chromatin with highthroughput sequencing*



# Chromatin accessibility assays

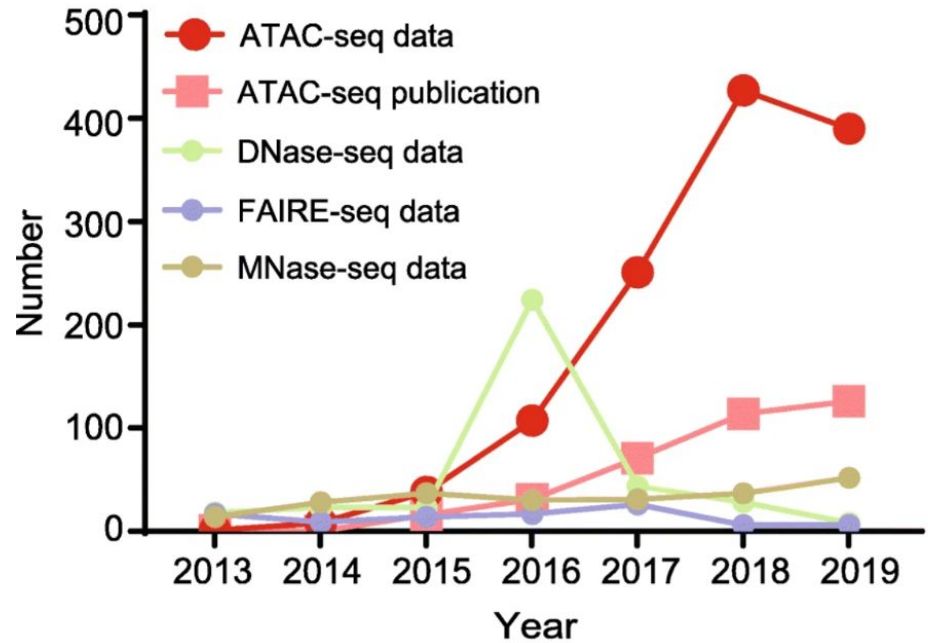
- Chromatin accessibility is the degree to which nuclear macromolecules are able to physically contact chromatinized DNA and is determined by the occupancy and topological organization of nucleosomes as well as other chromatin-binding factors that occlude access to DNA (Klemm et al, 2019)
- Open chromatin regions contains generally transcriptionally active genes
- The accessible genome comprises  $\sim 2\text{--}3\%$  of total DNA sequence yet captures more than 90% of regions bound by TFs (Thurman et al, 2012)
- Chromatin accessibility is measured by quantifying the susceptibility of chromatin to either enzymatic methylation or cleavage of its constituent DNA
- Chromatin accessibility assays (non exhaustive list)  
FAIRE-seq, DNase-seq, MNase-seq, ATAC-seq



**Figure 1** Schematic diagram of current chromatin accessibility assays performed with typical experimental conditions. Representative DNA fragments generated by each assay are shown, with end locations within chromatin defined by colored arrows. Bar diagrams represent data signal obtained from each assay across the entire region. The footprint created by a transcription factor (TF) is shown for ATAC-seq and DNase-seq experiments.

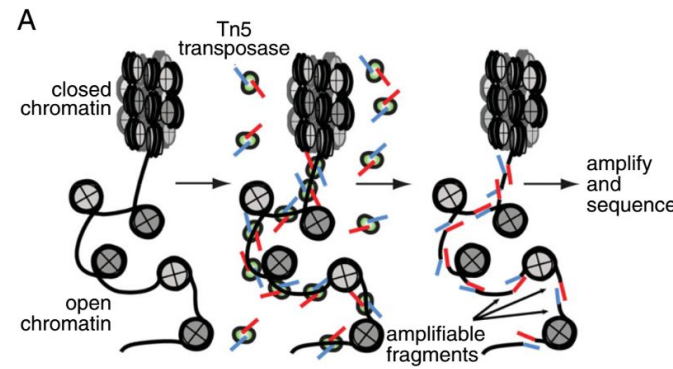
# Chromatin accessibility assays

ATAC-seq has become the most widely used method to detect and analyze open chromatin regions



# ATAC-seq

- Buenrostro et al, 2013
- ATAC-seq is a method for determining chromatin accessibility across the genome
- Transcription factor binding sites and positions of nucleosomes can be identified from the analysis of ATAC-Seq
- Advantages of ATAC-seq over other chromatin accessibility assays:
  - The sensitivity and specificity are comparable to DNase-seq but superior to FAIRE-seq
  - Straightforward and rapidly implemented protocol. ATAC-seq libraries can be generated in less than 3 hours
  - Low number of cells required (500 - 50,000 cells)
  - single-cell ATAC-seq (scATAC-seq) protocol (Cusanovich et al, 2015)

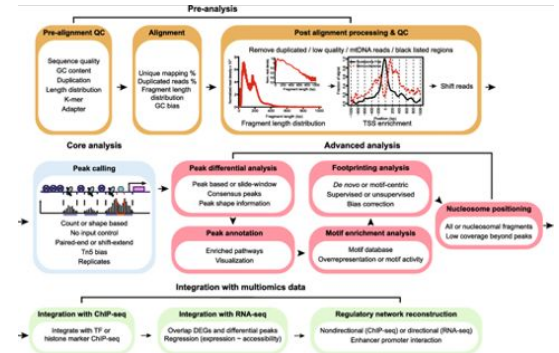
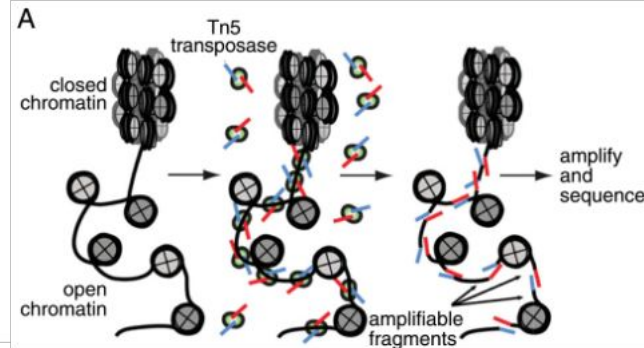
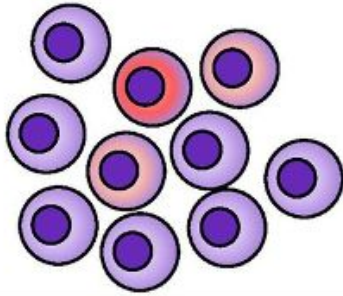


# ATAC-seq process

Sample processing

ATAC-seq

Data analysis

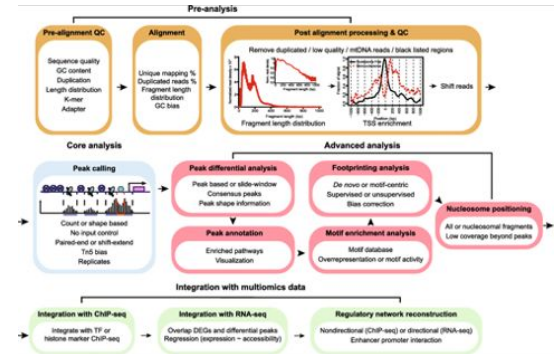
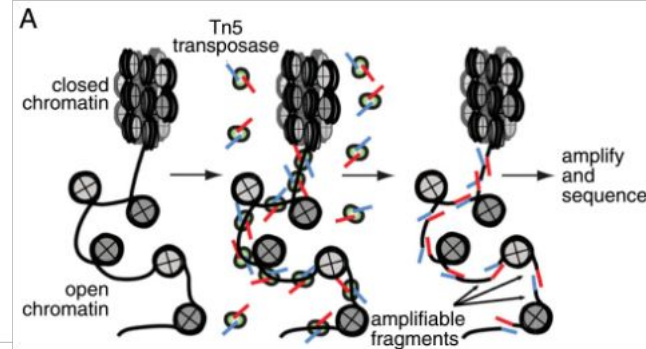
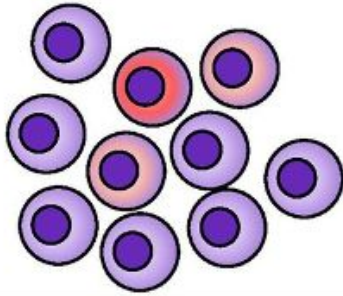


# ATAC-seq

Sample processing

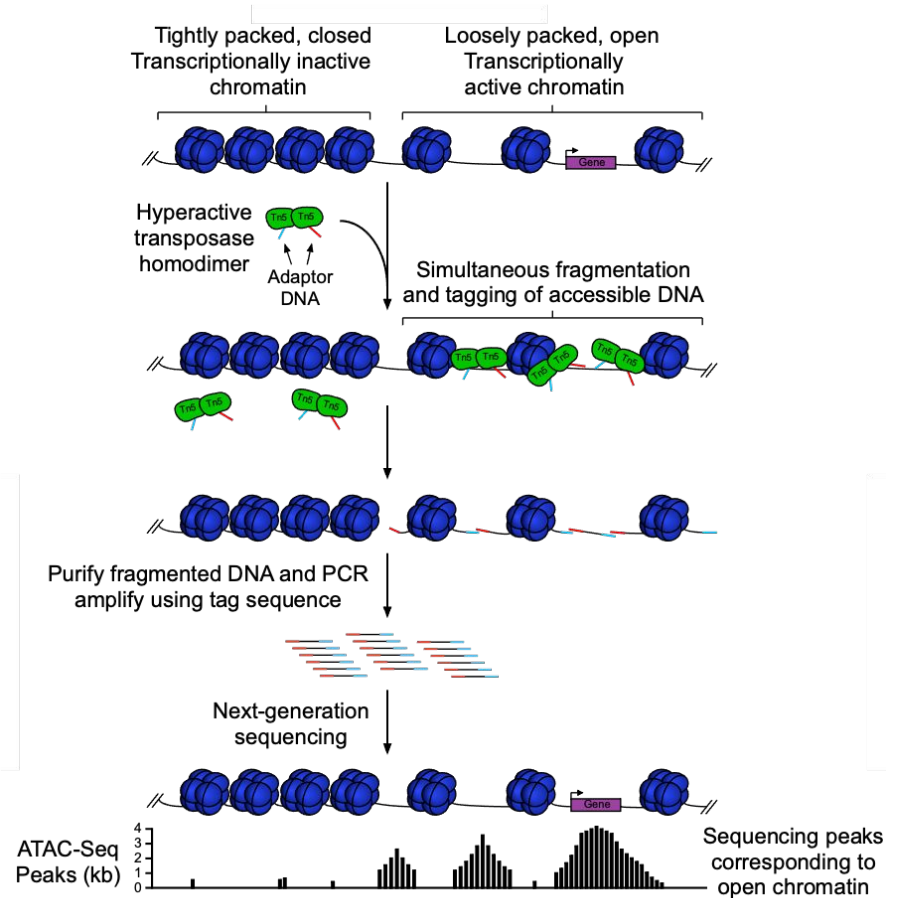
ATAC-seq

Data analysis



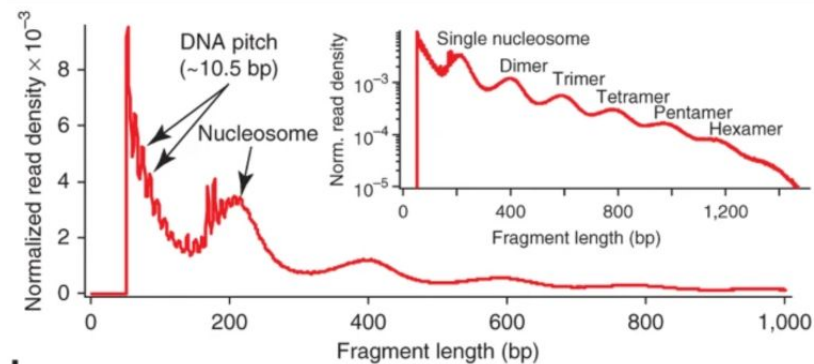
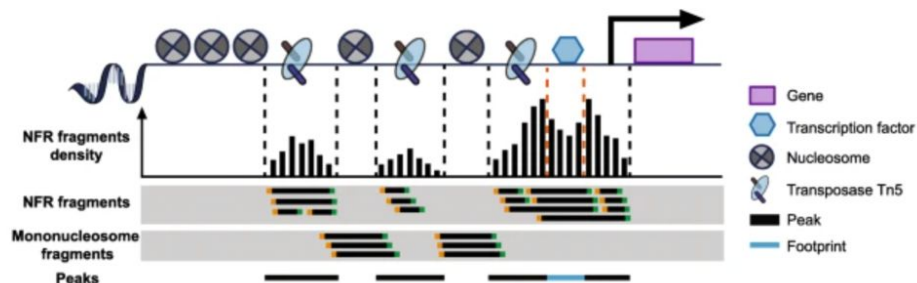
# ATAC-seq

- ATAC-seq protocol utilizes a hyperactive Tn5 transposase to insert sequencing adapters into open chromatin regions
- In a process called "tagmentation", Tn5 transposase cleaves and tags double-stranded DNA with sequencing adaptors
- No additional library prep is needed
- Expected results are enrichments of sequenced reads in open chromatin regions as closed chromatin regions, DNA regions bound by TFs or wrapped around nucleosomes should be protected from Tn5 cleavage activity.



# ATAC-seq

- **Paired-end sequencing** so that by looking at the distance between the two reads of a pair, we know in which the chromatin environment (Nucleosome Free Region (NFR), around a mono, di,-nucleosome, around a TF) of the DNA fragment.

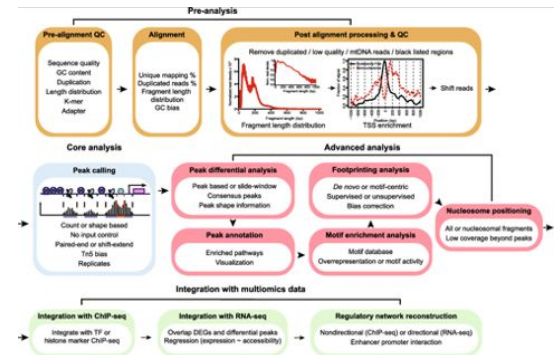
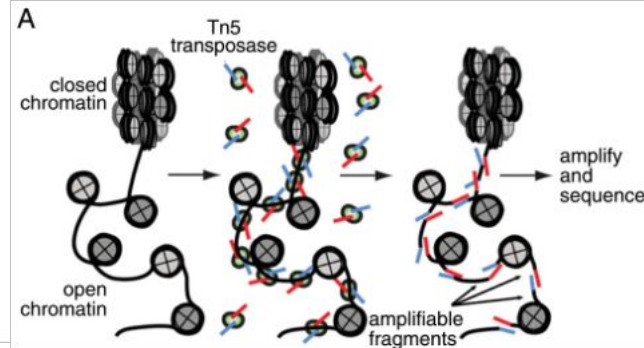
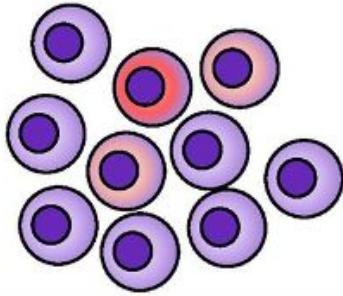


# ATAC-seq process

Sample processing

ATAC-seq

Data analysis





# Analysis of ATAC-seq data

QC

mapping

Post alignment  
processing &  
QC

Peak  
detection

Peak  
annotation

## Advanced analysis

Motif  
enrichment  
analysis

Differential  
analysis

Meta  
profiles /  
Clustering

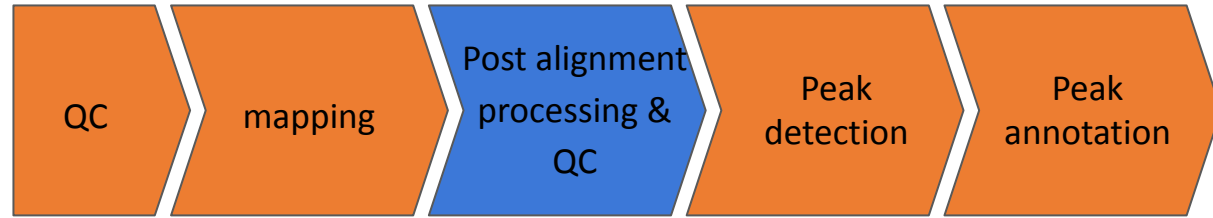
Footprinting  
analysis

Nucleosome  
positioning

...

- Overall analysis resemble ChIP-seq data analysis
- Description of particularities of ATAC-seq data analysis

# Analysis of ATAC-seq data

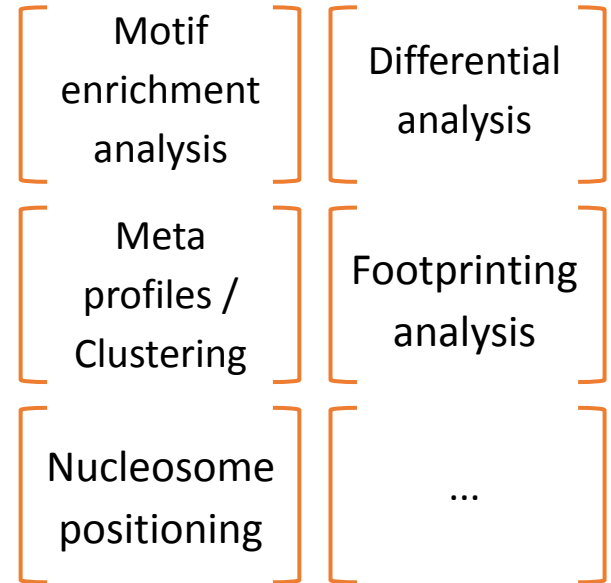


- Some cleaning steps are required for ATAC-seq. For example:

- > A large percentage of reads are derived from mitochondrial DNA. These reads are removed as mitochondrial genome is generally not of interest.

- > Omni-ATAC (Corces et al, 2017)

## Advanced analysis



# Analysis of ATAC-seq data

QC mapping Post alignment processing & QC Peak detection Peak annotation

## Advanced analysis

Motif enrichment analysis

Differential analysis

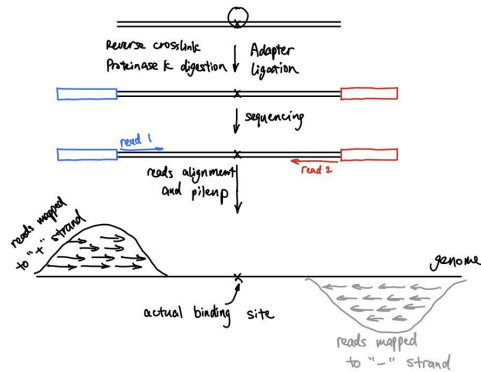
Meta profiles / Clustering

Footprinting analysis

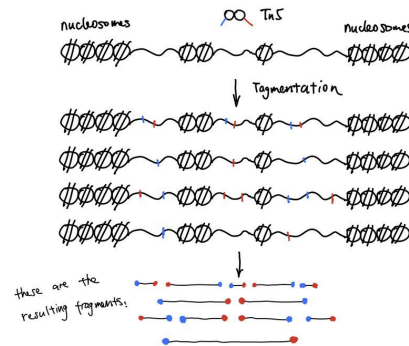
Nucleosome positioning

...

### ChIP-seq



### ATAC-seq



Adapted parameters for peak calling (MACS2) : `--shift 75 --extsize 150 --nomodel -B --SPMR --keep-dup all --call-summits`

# Analysis of ATAC-seq data

QC

mapping

Post alignment  
processing &  
QC

Peak  
detection

Peak  
annotation

## Advanced analysis

Motif  
enrichment  
analysis

Differential  
analysis

Meta  
profiles /  
Clustering

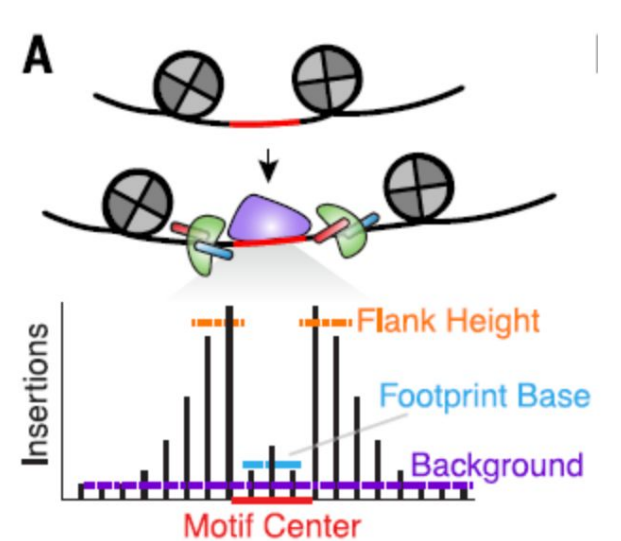
Footprinting  
analysis

Nucleosome  
positioning

...

# Footprinting analysis

- Tn5 cuts in open chromatin regions
- DNA is protected from cleavage at position of TF binding creating a “notch” in ATAC-seq signal
- Footprinting analysis identifies TF activities
  - > Height of the notch reflects TF activity
  - > Compare TF activity between different conditions



# Footprinting analysis

- Volcano plots showing differential TF binding activity as predicted by TOBIAS footprinting analysis in ATAC-seq data of NKp, iNK and mNK from Shin et al. (c) iNK vs NKp; (d) mNK vs NKp; (e) mNK vs iNK.
- Each dot represents a TF
- TFs which activity is changing between the two compared developmental stages are colored (see color legend below volcano plots)

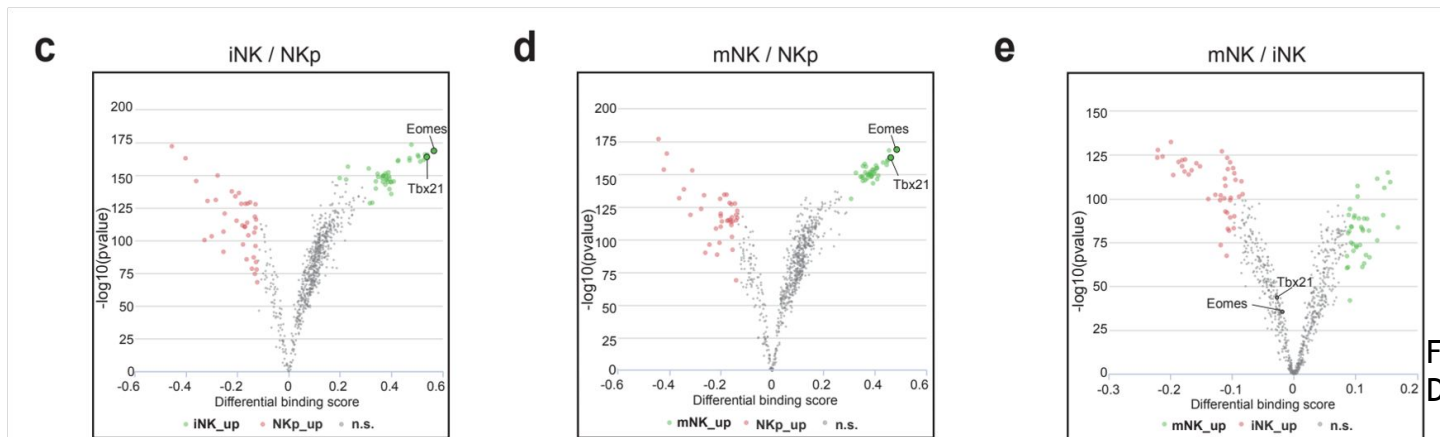


Figure: Zhang et al, 2021  
Data: Shih et al, 2016

# Atelier ChIP-seq: tour de table des données

Les questions qui pourraient moduler le pipeline d'analyse

Narrow peak ou broad peak ?

Paired-end ou single-end ?

Disponibilité du génome de référence (partie annotation) ?

Utilisation de spike-ins

Qualité de l'assemblage du génome ?