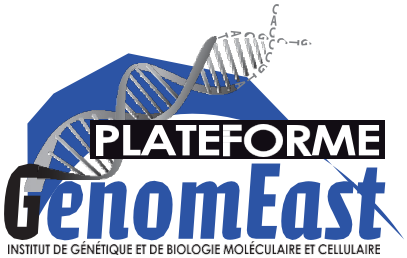
# Epigenetics assays

Analysis of gene expression regulation

Stéphanie Le Gras

GenomEast Platform, Illkirch, France

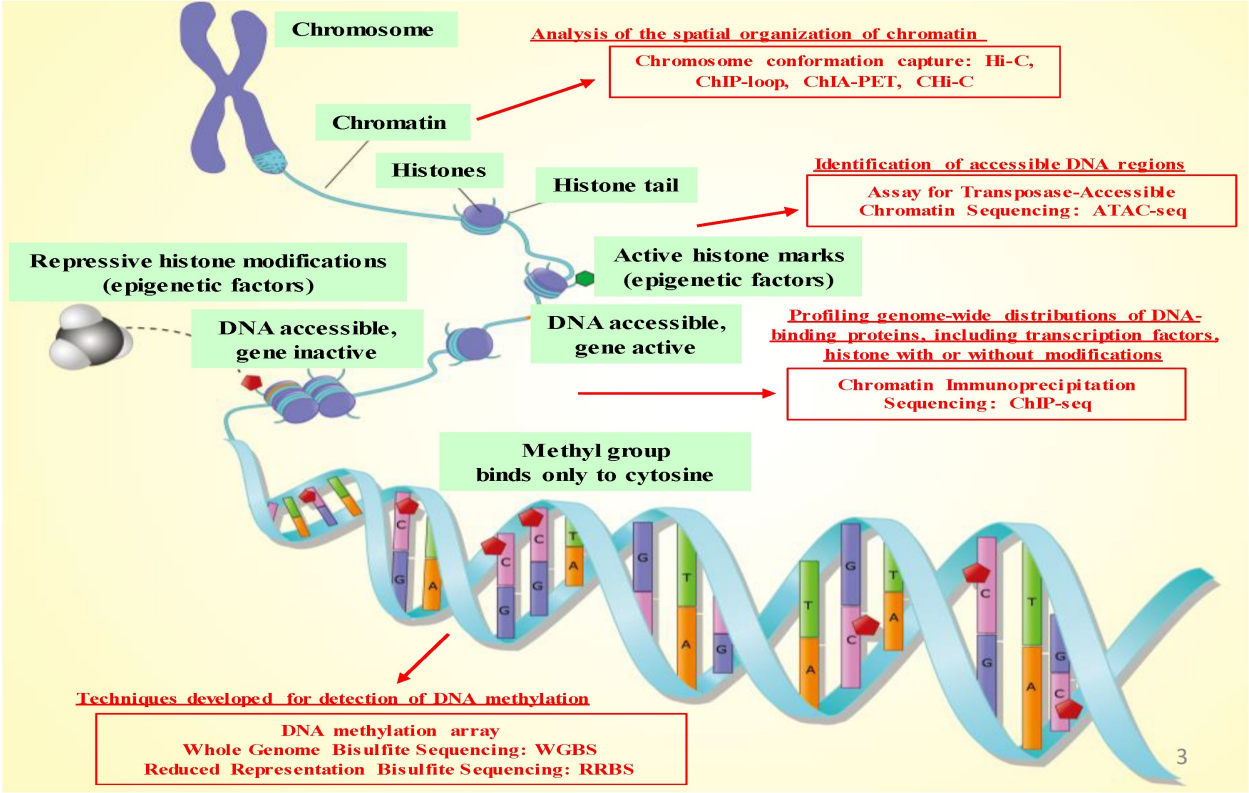


# Epigenetics

- Epigenetics is the study of changes in gene function that are mitotically and/or meiotically heritable and that do not entail a change in DNA sequence. (Wu et al, 2001)
- Epigenetic mechanisms such as DNA methylation, Histone Post-Translational Modification (PTM), Nucleosome positioning are affected by development (in utero, childhood), environmental chemicals, drugs/pharmaceuticals, aging, diet,…

# Technologies for epigenetics analysis

Hamamoto et al, 2020



# ChIP-seq

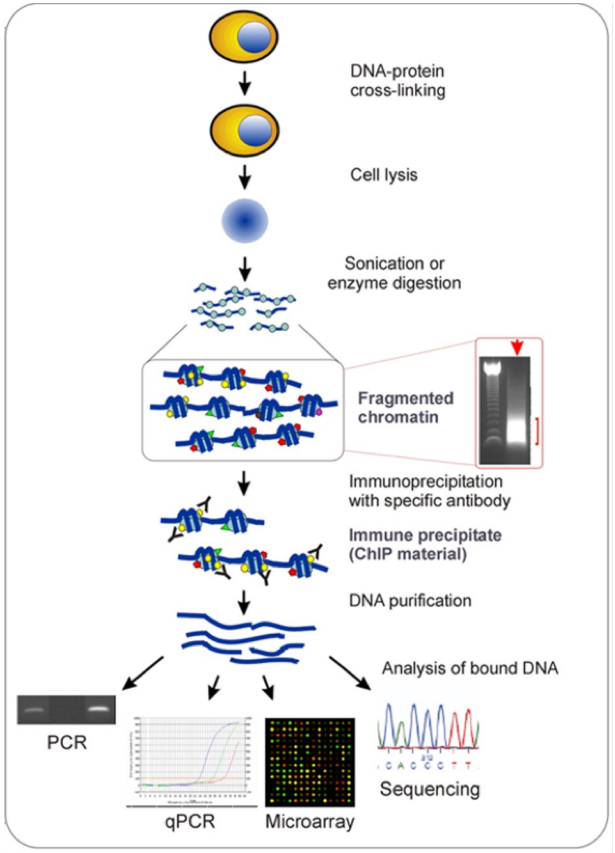
**Chromatin ImmunoPrecipitation followed by sequencing**

# ChIP-seq

**ChIP (=Chromatin Immuno-Precipitation)**

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

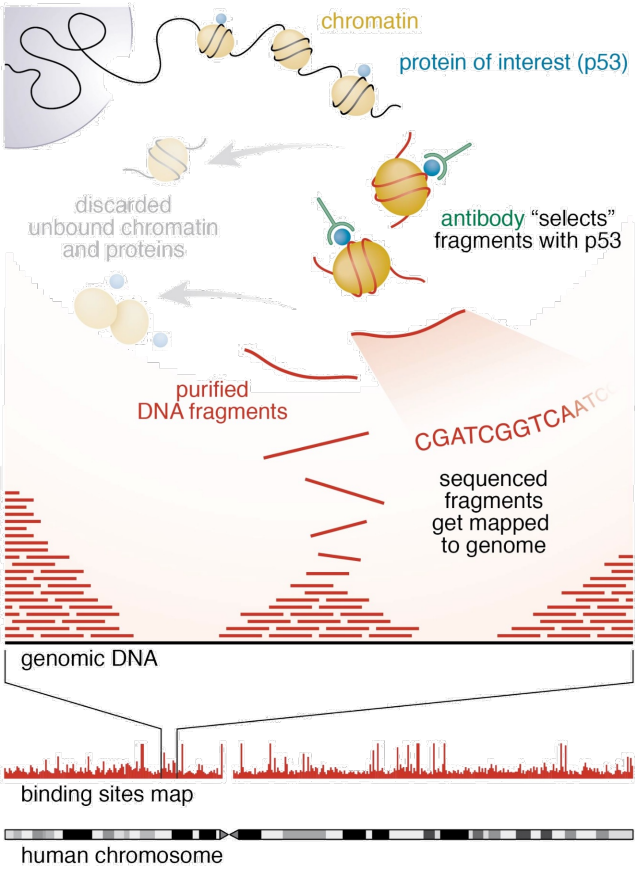
- small-scale: PCR / qPCR
- large-scale:
- microarray = **ChIP-on-chip**
- sequencing = **ChIP-seq**




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

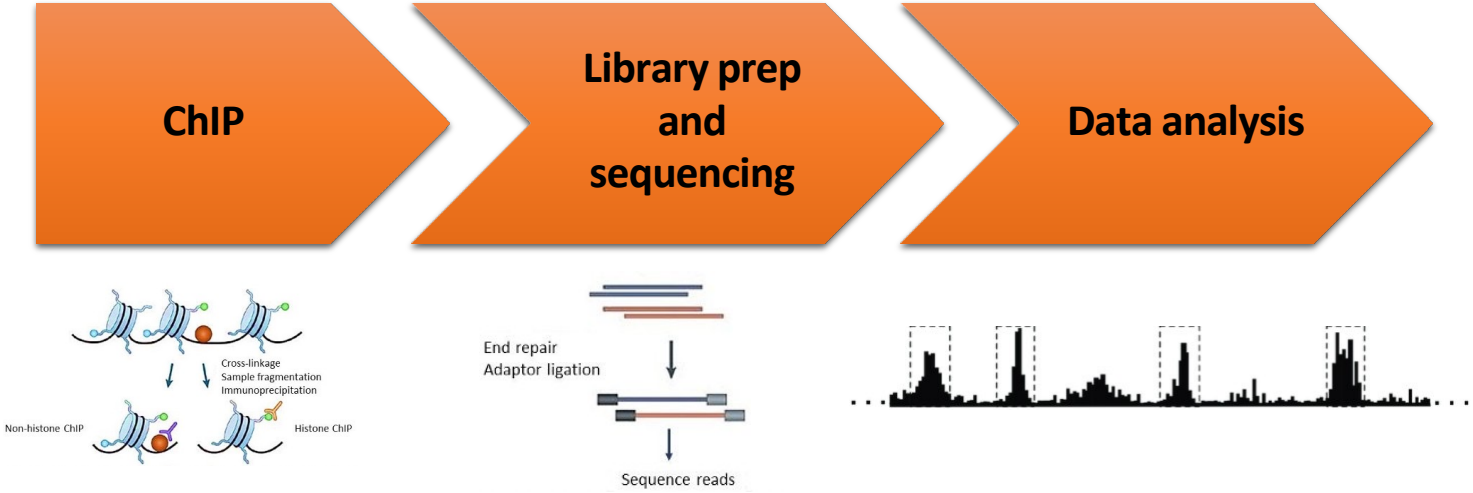
# ChIP-seq

- Johnson et al, 2007
- Alternative to ChIP-on-chip (hybridization technique)
- ChIP-seq combines chromatin immunoprecipitation and sequencing to analyze protein interactions with DNA
- It can be used to detect and analyze
  - Binding sites of various proteins bound to DNA such as transcription factors (TF ChIP-seq)
  - Position of histone post translational modifications (Histone ChIP-seq)
  - Nucleotide modification such methylation (MeDIP-seq)
- Expected results of a ChIP-seq experiment are genomic regions with significant sequenced read enrichments (also called peaks)

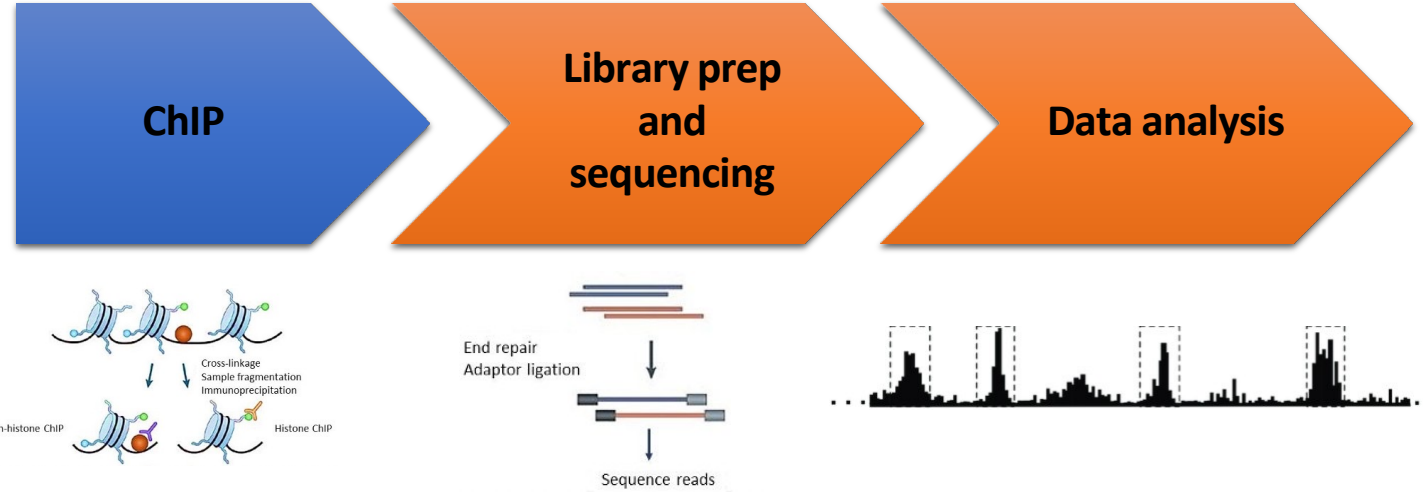



https://hbctraining.github.io/Intro-to-ChIPseq/lessons/01_Intro_chipseq_data_organization.html

# ChIP-seq process

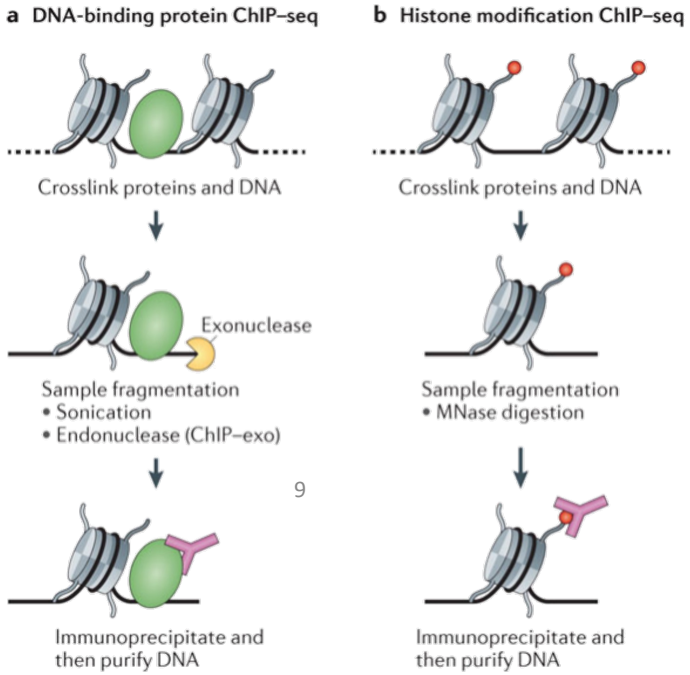


# Chromatin ImmunoPrecipitation



# Chromatin ImmunoPrecipitation

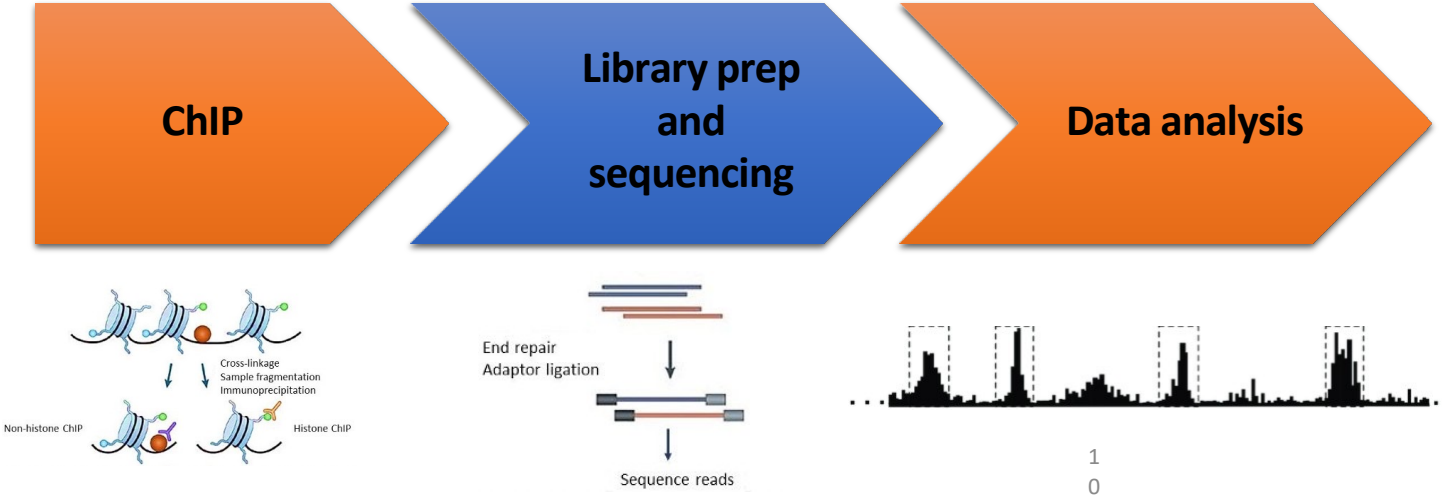
- ChIP is used to isolate DNA fragments bound by proteins or modified histones of interest using an antibody specifically design to target them
- It requires a high number of cells (1-10 million cells)
- It requires highly tested and verified antibody working specifically for ChIP






Furey, 2012

# Library prep and sequencing



# Experimental design

# ENCODE

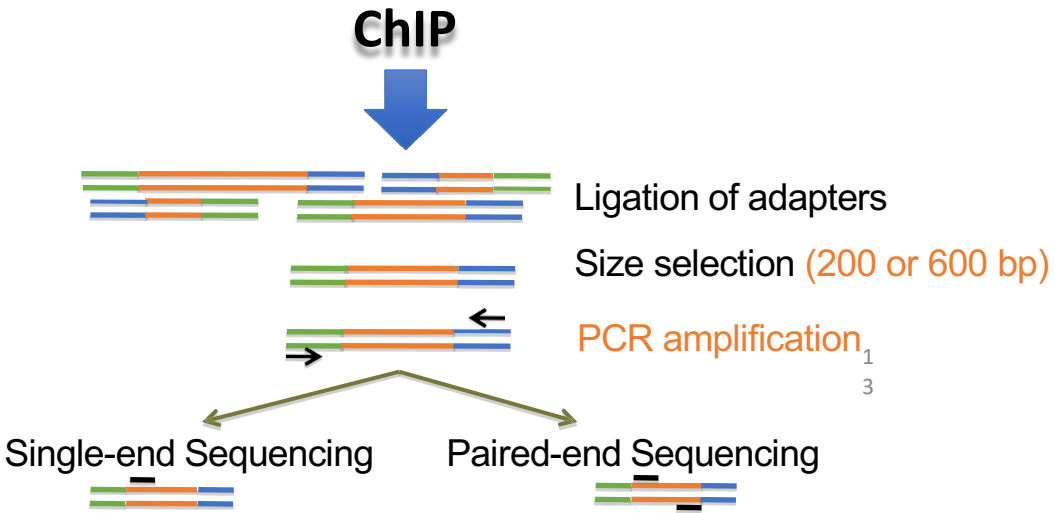
- The Encyclopedia of DNA Elements (ENCODE) Consortium has carried out thousands of ChIP-seq experiments and has used this experience to develop a set of working standards and guidelines


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/

# Library prep

- Starting material: ChIP sample (1-10ng of sheared DNA)

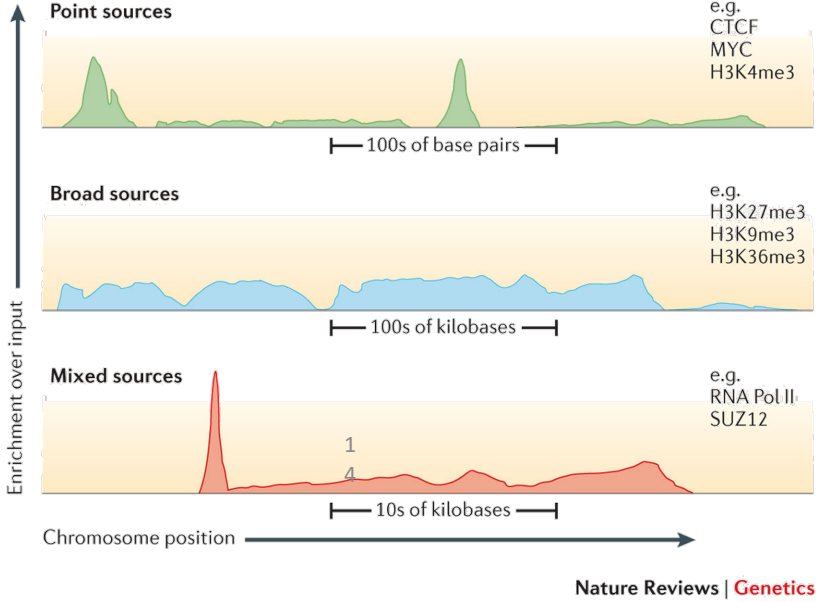


# Sequencing

- Paired-end or single-end
- Short reads (50nt)
- Number of reads needed depends on:
  - chipped protein,
  - type of expected profile,
  - number of expected binding sites,
  - size of the genome of interest.

Ex:

- For human genomes, 20 million uniquely mapped read sequences are suggested for point-source peaks, or 40 million for broad-source peaks.
- For fly genome: 8 million reads
- For worm genome: 10 million reads



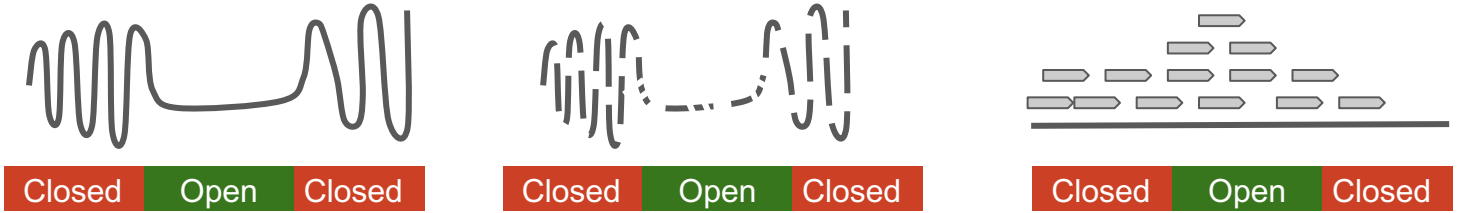

Park, 2009

# Controls

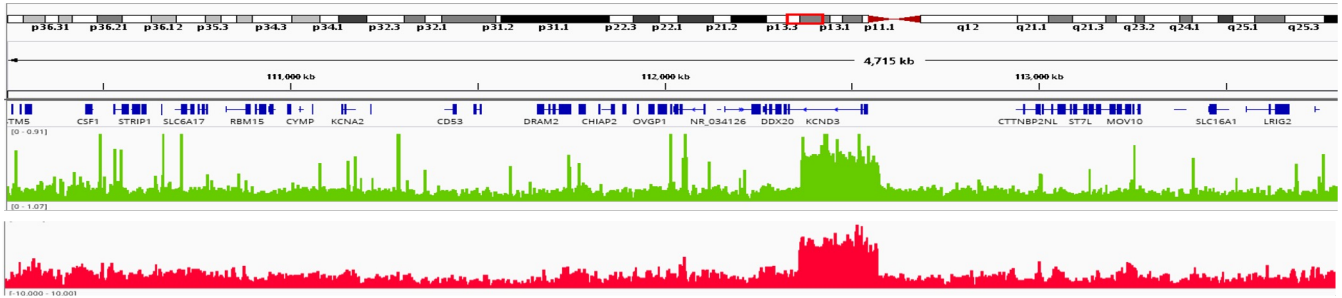
- Used mostly to filter out false positives (high level of noise)
  - Idea: potential false positive will be enriched in both treatment and control.
- A control will fail to filter out false positives if its enrichment profile is very different from the enrichment profile of false positive regions in 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**

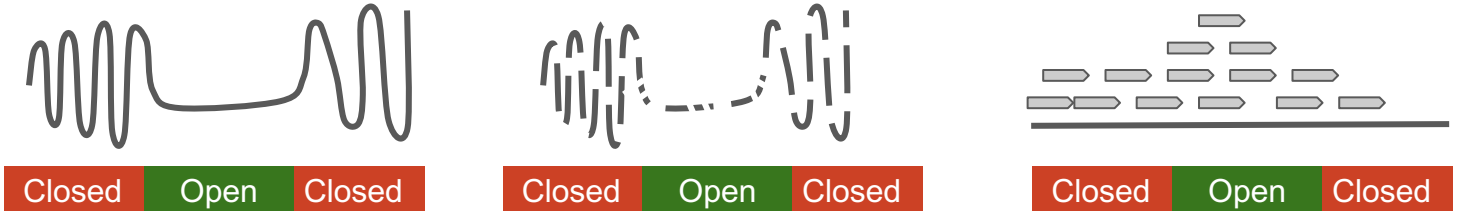



  - **Amplified regions (CNV) are expected to produce more reads**



# 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 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

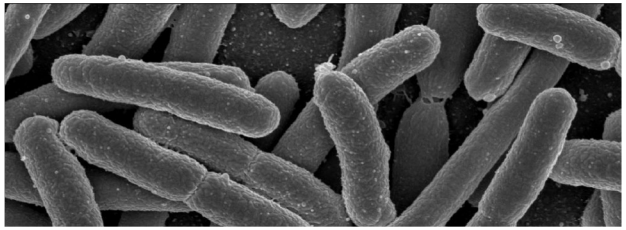
# Dataset used

OPEN ACCESS Freely available online

PLOS | GENETICS

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

Kevin S. Myers[1,2], Huihuang Yan[3¤a], Irene M. Ong[3], Dongjun Chung[4¤b], Kun Liang[4,5¤c], Frances Tran[6¤d], Sündüz Keleş[4,5], Robert Landick[3,6,7*], Patricia J. Kiley[2,3*]

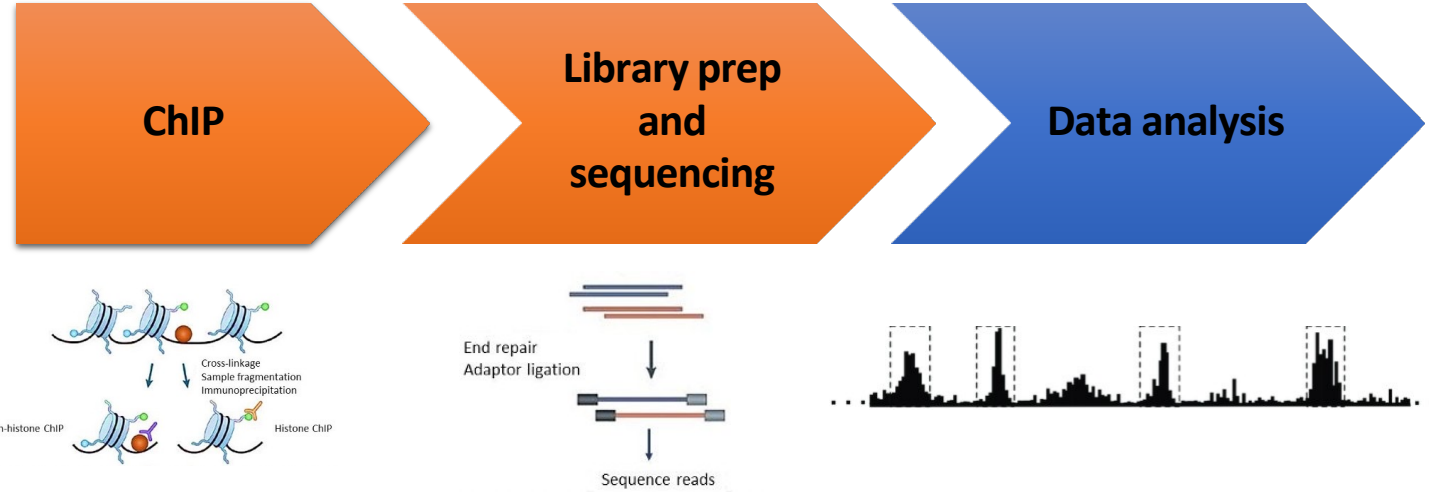


- All experiments (GEO): GSE41187
- Experiment: **FNR** IP ChIP-seq Anaerobic A (SRX189773 - SRR576933)
- Control: anaerobic INPUT DNA (SRX189778 - SRR576938)

# TIPS

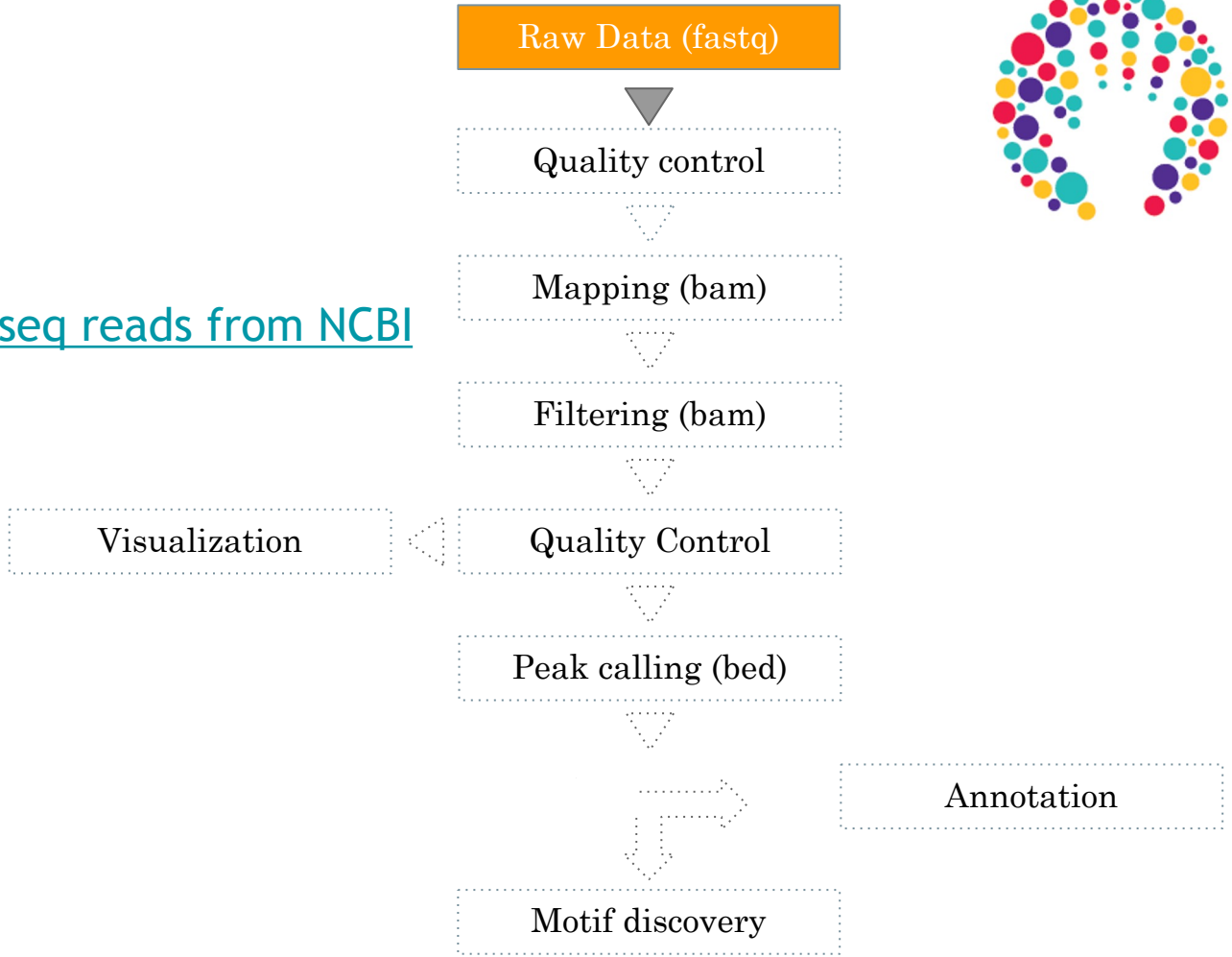
- **Keep track** of all command lines you run. You can for example, create a text file in which you write every commands you run.
- Give **content-explicit names** to the files you're generating.
- Give to files the **right extension**.
- **Create directories** with **explicit names**!!
- **Compress**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

# Analysis of chIP-seq data



# Protocol

- Downloading ChIP-seq reads from NCBI



# Protocol

- Connect to the server and set up your environment

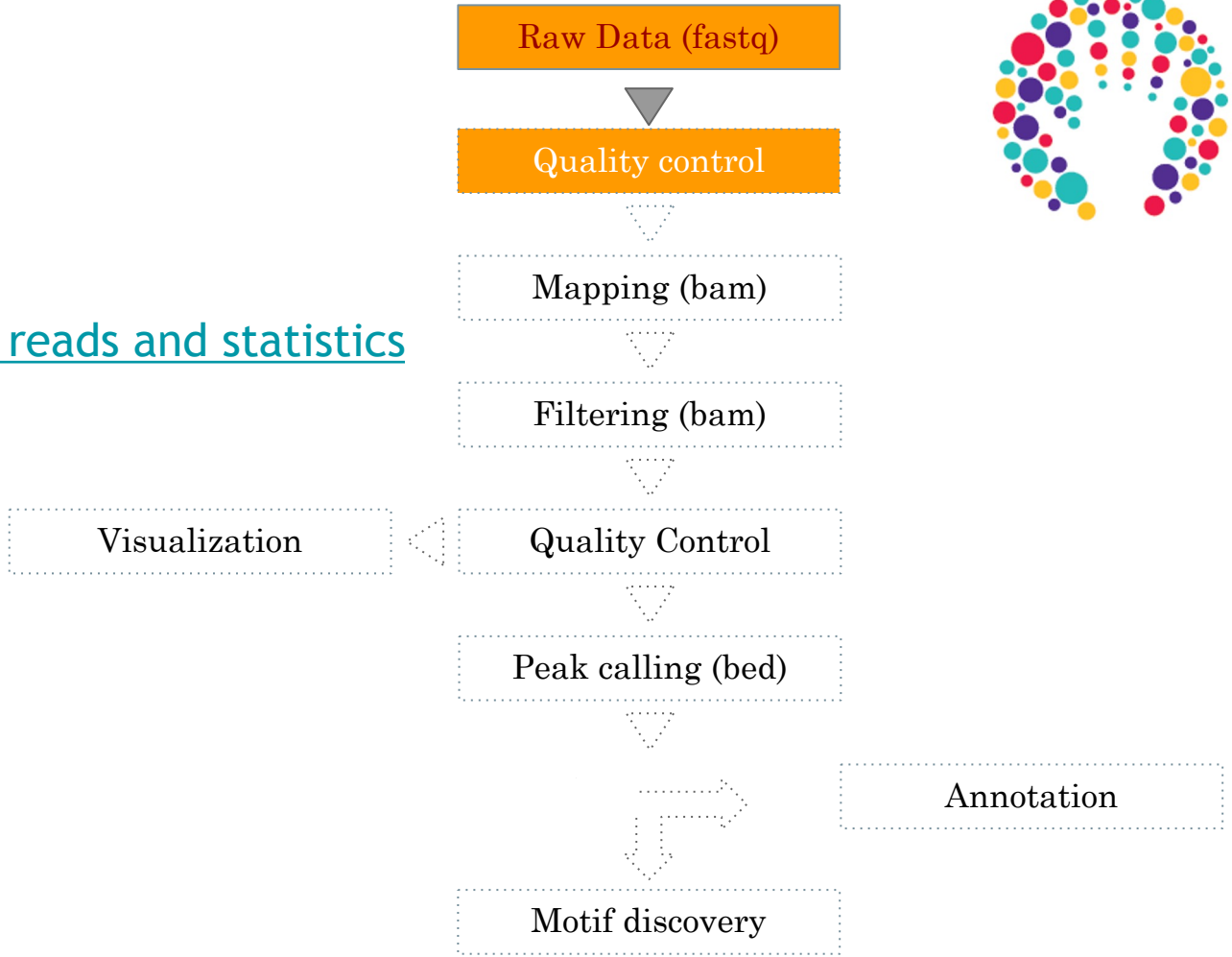
# Quality control of the reads

# Quality control of the reads

- As for any NGS datasets
- FastQC program

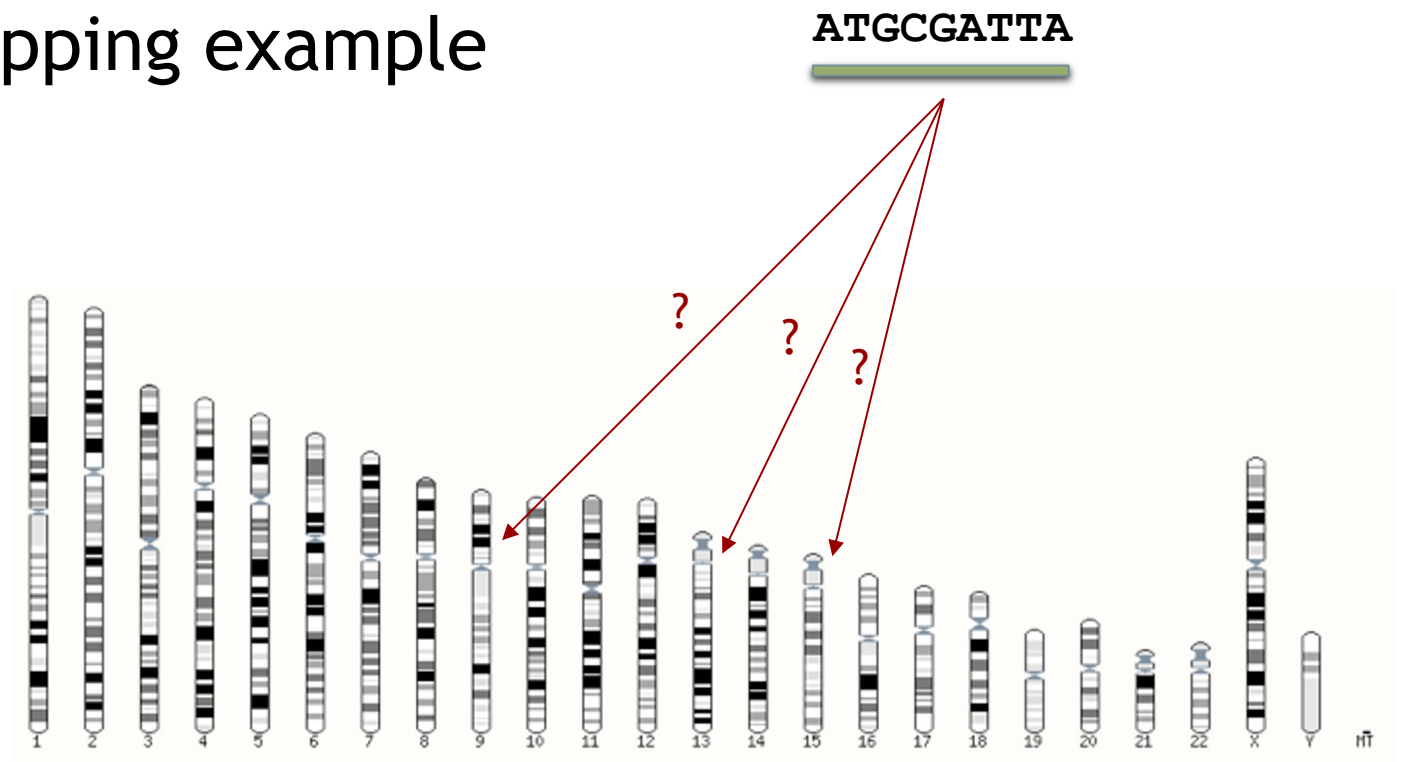
# Protocol

- Quality control of the reads and statistics



# Mapping

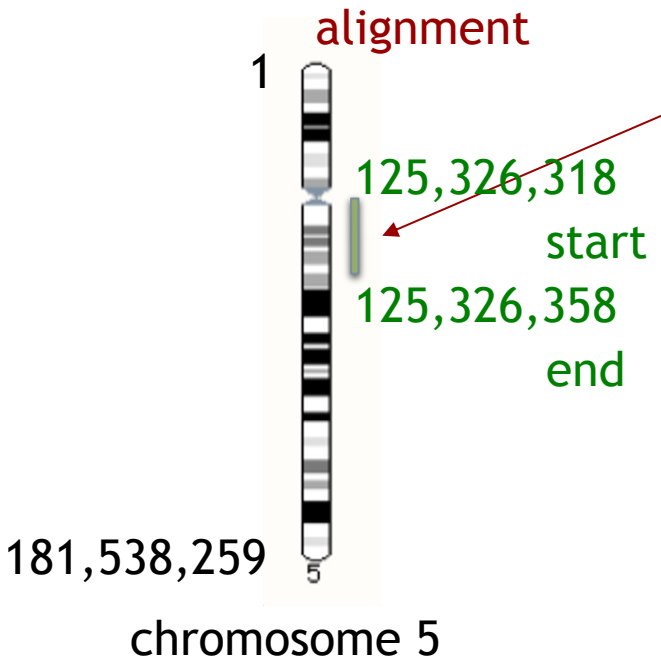
# Mapping example




Human chromosomes

# Genomic coordinates

ATGCGATTA





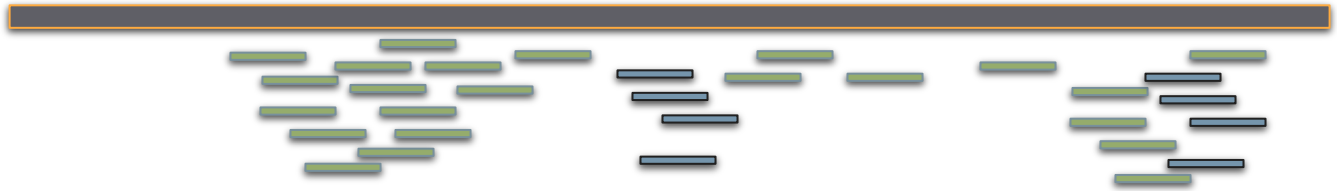
Genomic coordinate of the mapped read :

```
chr5   125326318 125326358 +
```

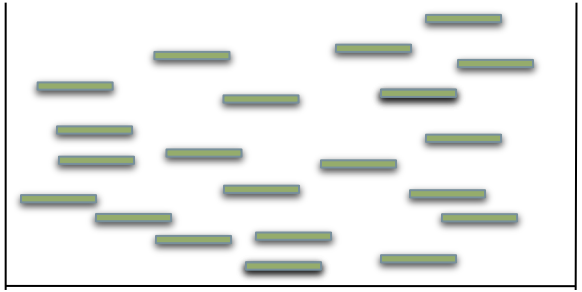
# Mapping

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

Ref. Genome



Reads



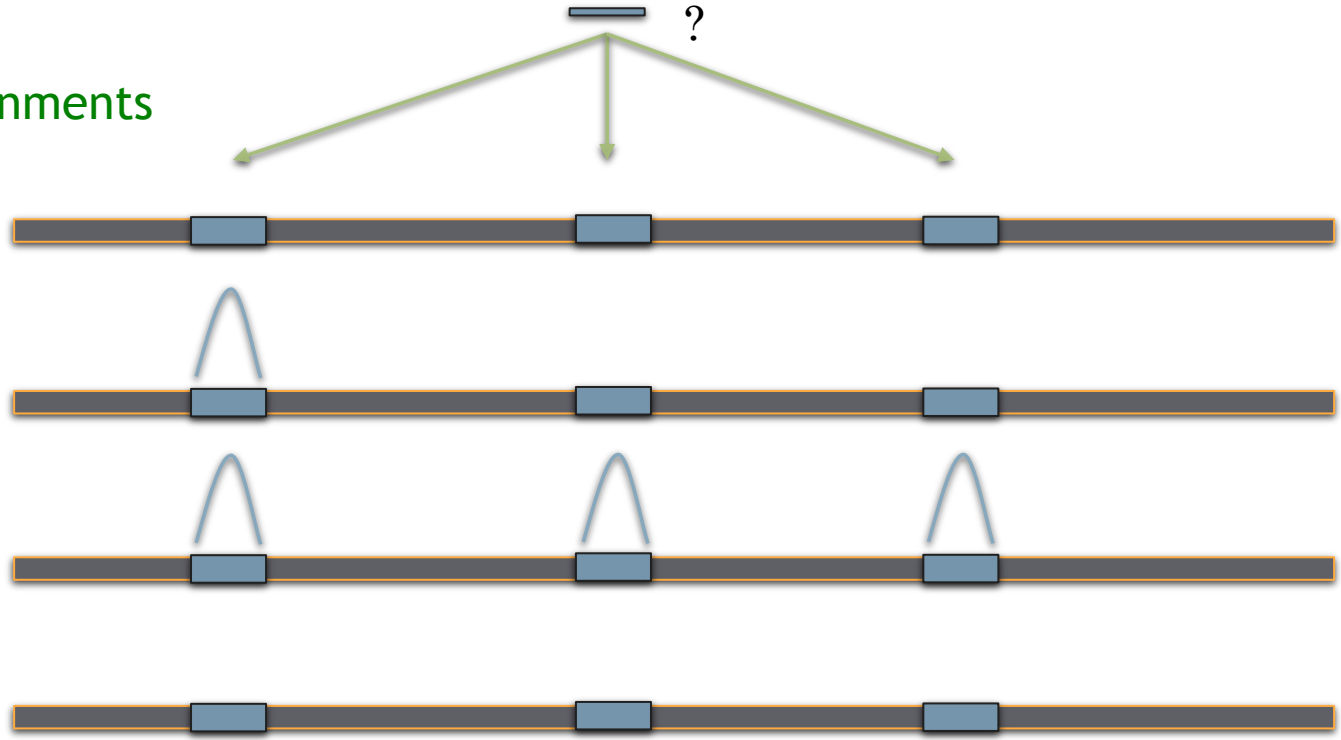
- One position in the genome
- Many possible positions (Repeat regions, duplicate regions, pseudogenes…)

# Mapping tool used: Bowtie

- (c.f. Course “mapping” Denis Puthier Monday afternoon)
- 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



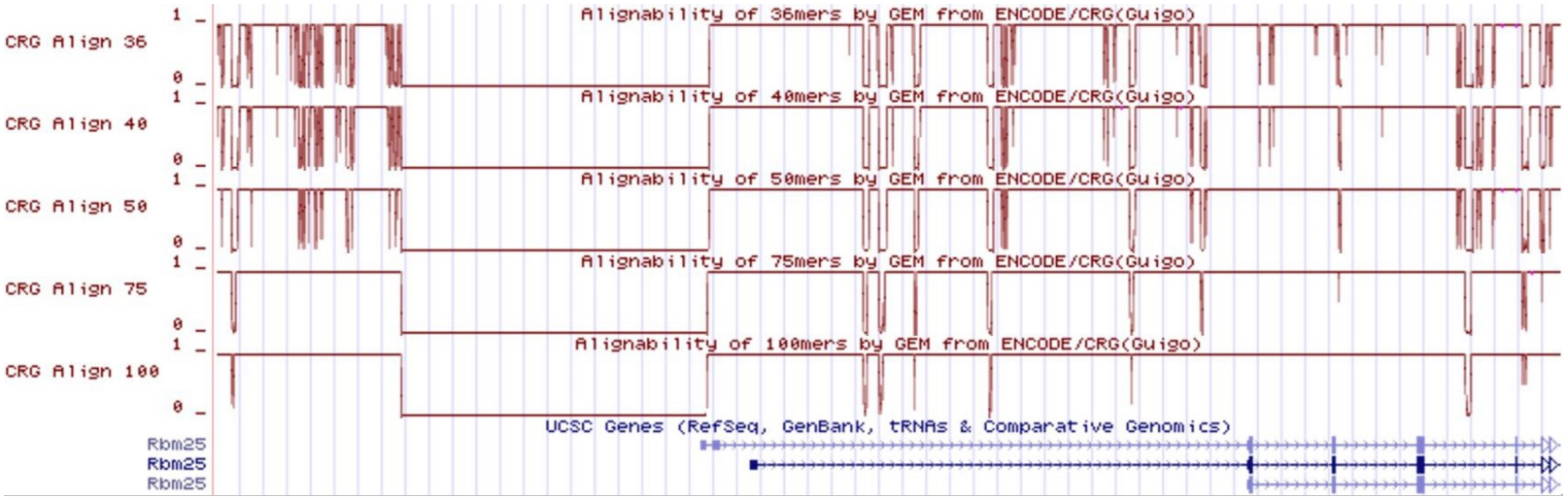

Keep 1 position randomly

Keep all possible position

Keep none

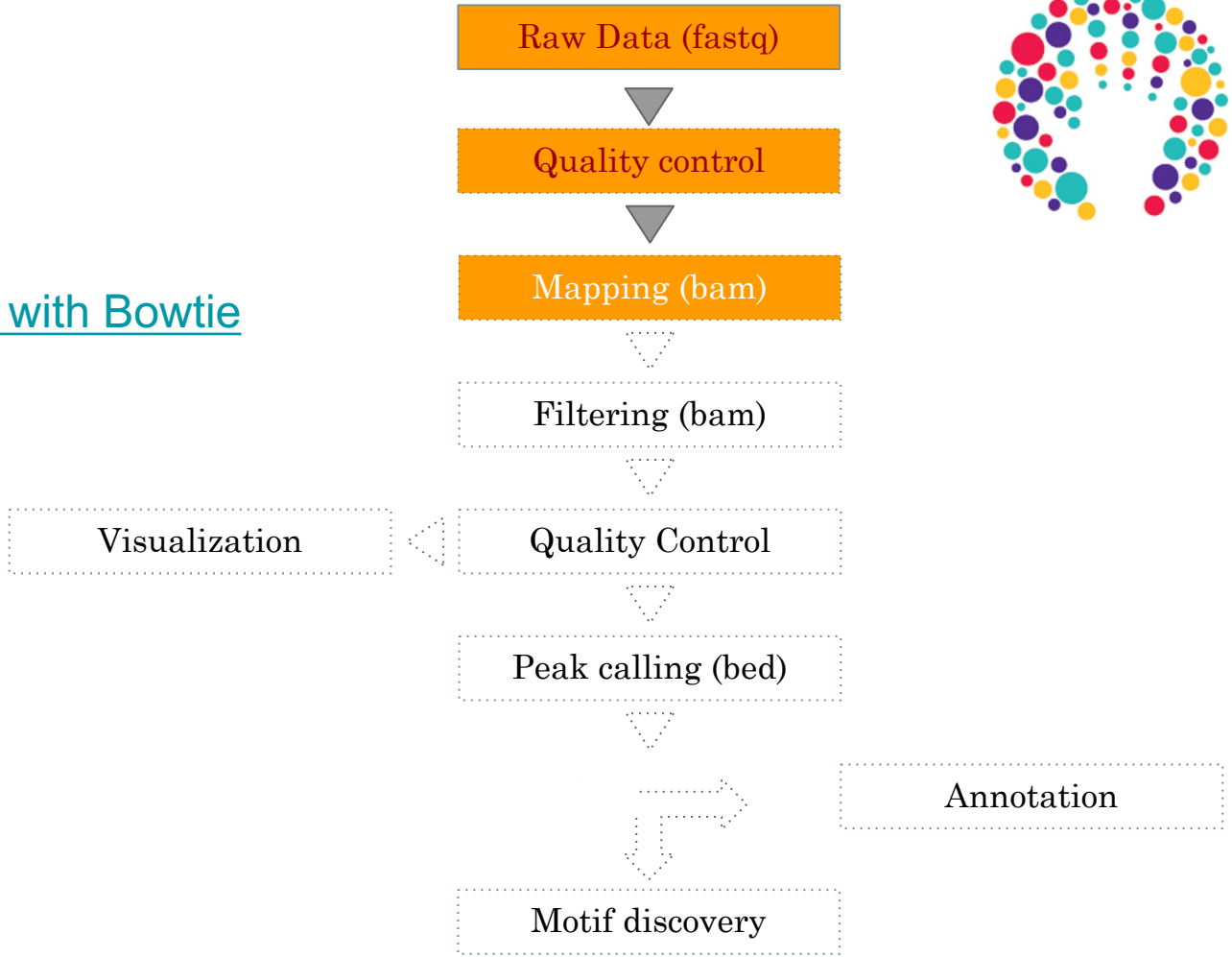
# Mappability

- 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)



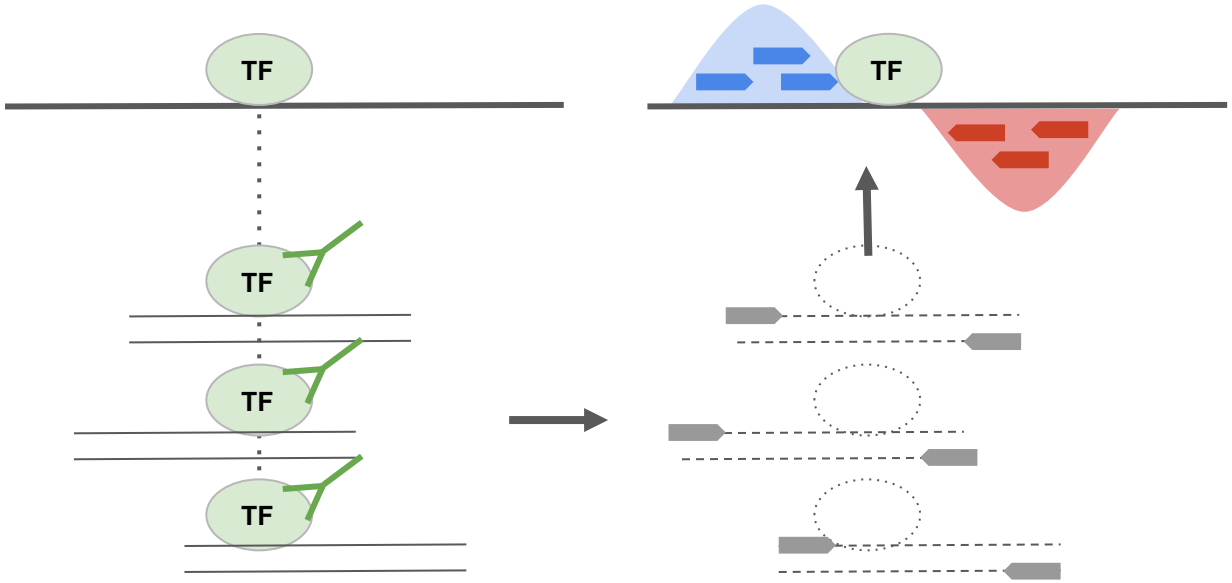
# Protocol

- Mapping the reads with Bowtie



# Mapping: expected signal

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

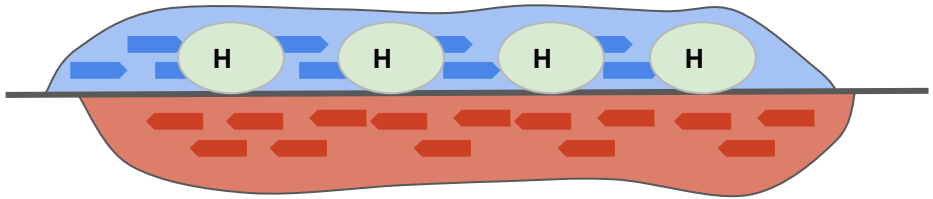



**The binding site itself is generally not sequenced !**

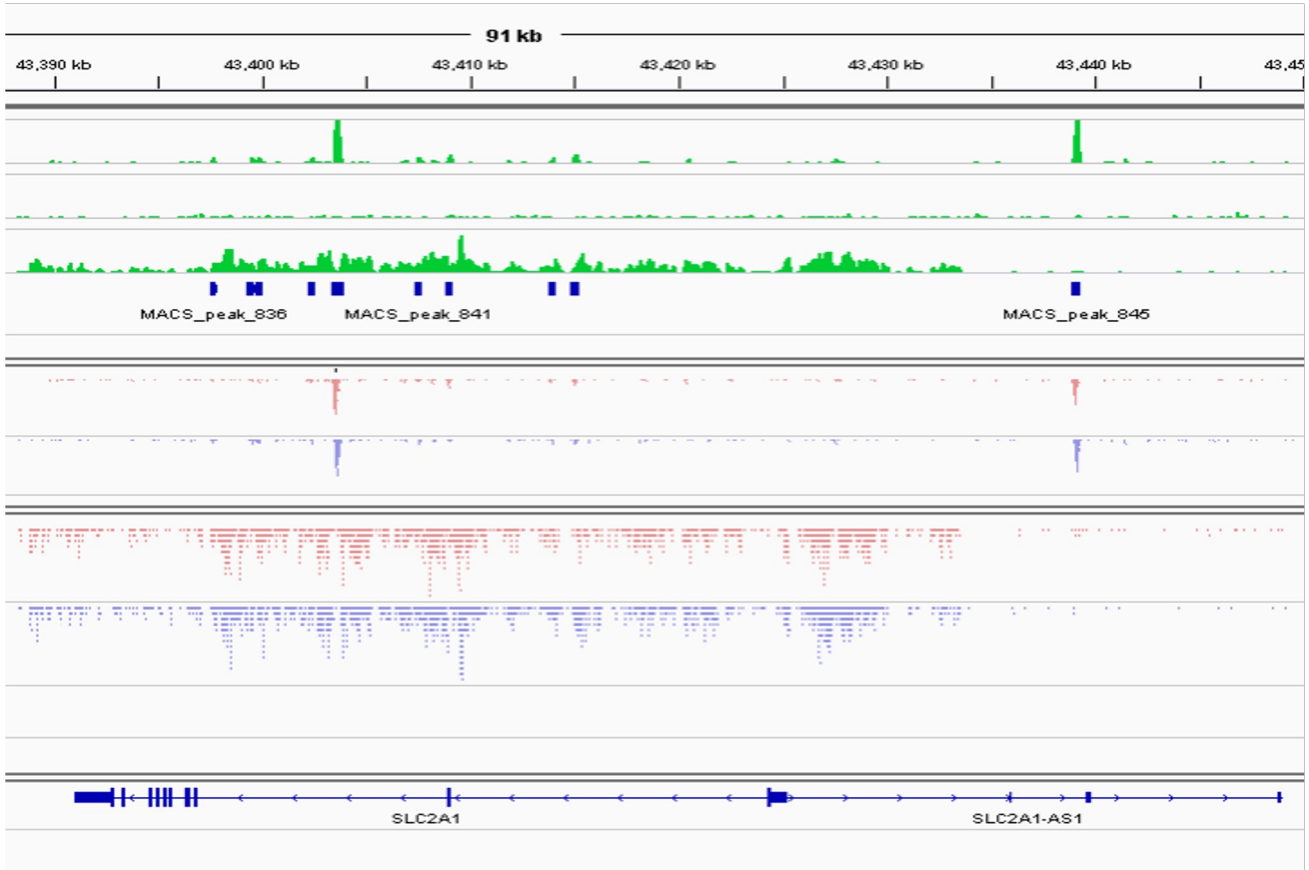
Sens alignments
Rev/comp alignments

# 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)

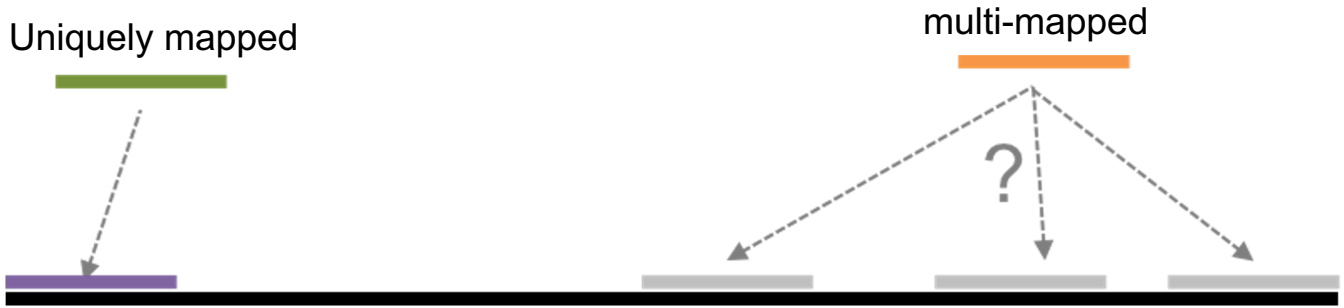
# Filtering mapped reads

# Which reads to filter ?

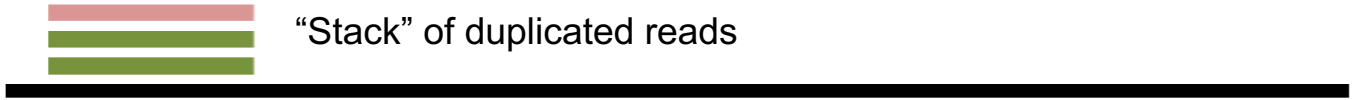
- Low-quality read alignments
  - Tool : samtools
- Multi-mapped reads (unless removed during the mapping step)
  - Tool : samtools
- Duplicated reads (PCR duplicates)
  - Tool : Picard MarkDuplicates

# 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)

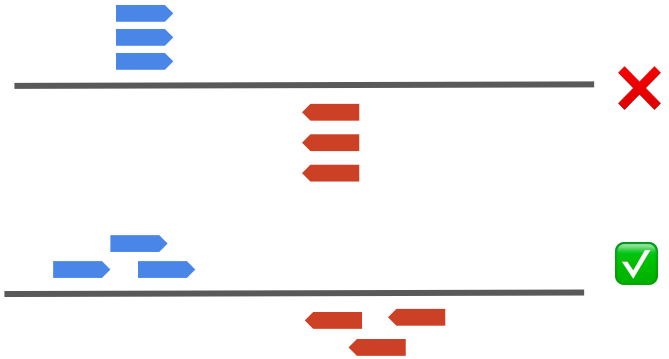


# PCR duplicates

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

# QC : PBC (PCR Bottleneck Coefficient)

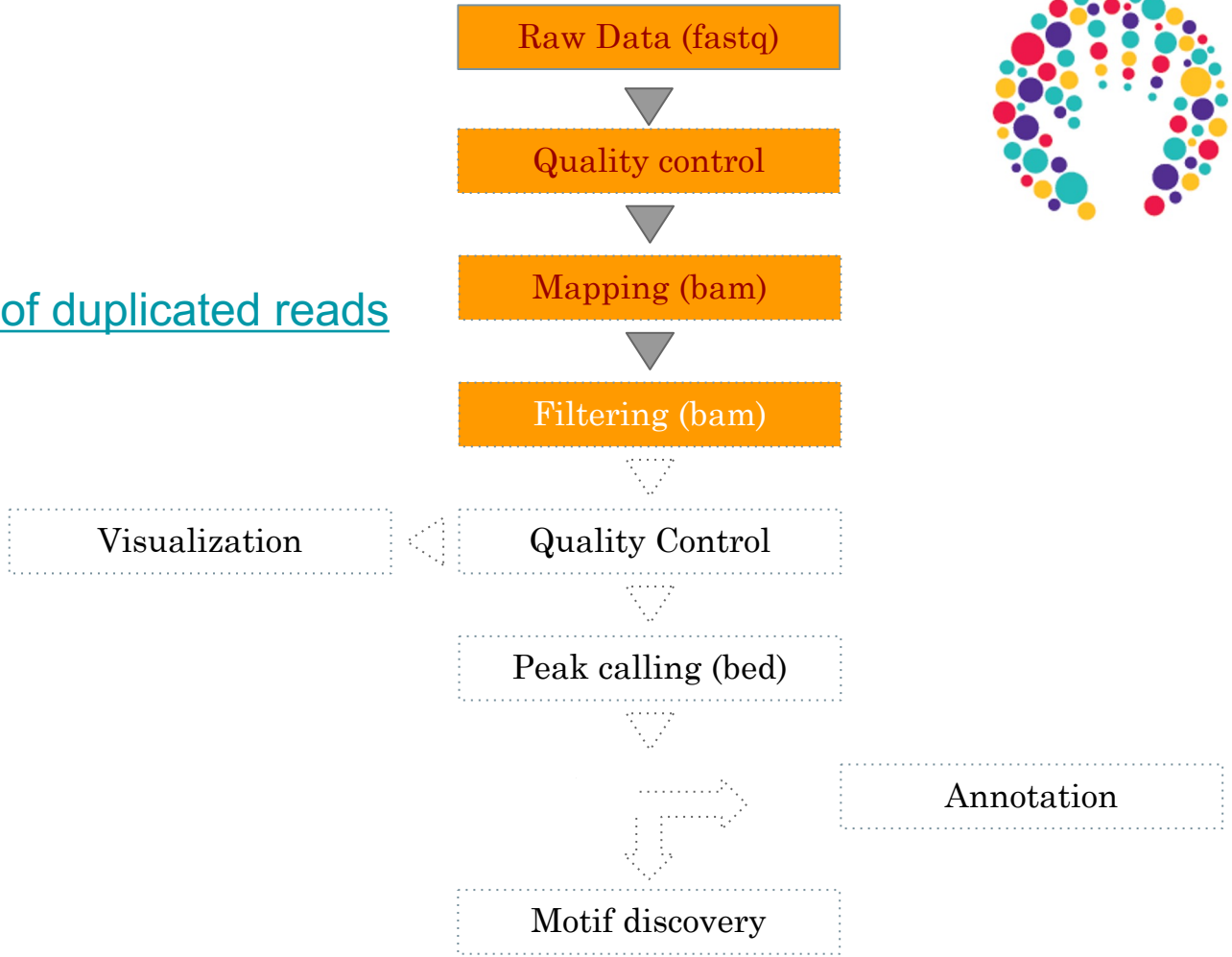
- An approximate measure of library complexity
- PBC = $N_1/N_d$
  - $N_1$= Genomic position with 1 read aligned
  - $N_d$ = Genomic position with ≧ 1 read aligned
- Value :
  - 0-0.5: severe bottlenecking
  - 0.5-0.8: moderate bottlenecking
  - 0.8-0.9: mild bottlenecking
  - 0.9-1.0: no bottlenecking



https://genome.ucsc.edu/ENCODE/qualityMetrics.html

# Protocol

- Estimating the number of duplicated reads



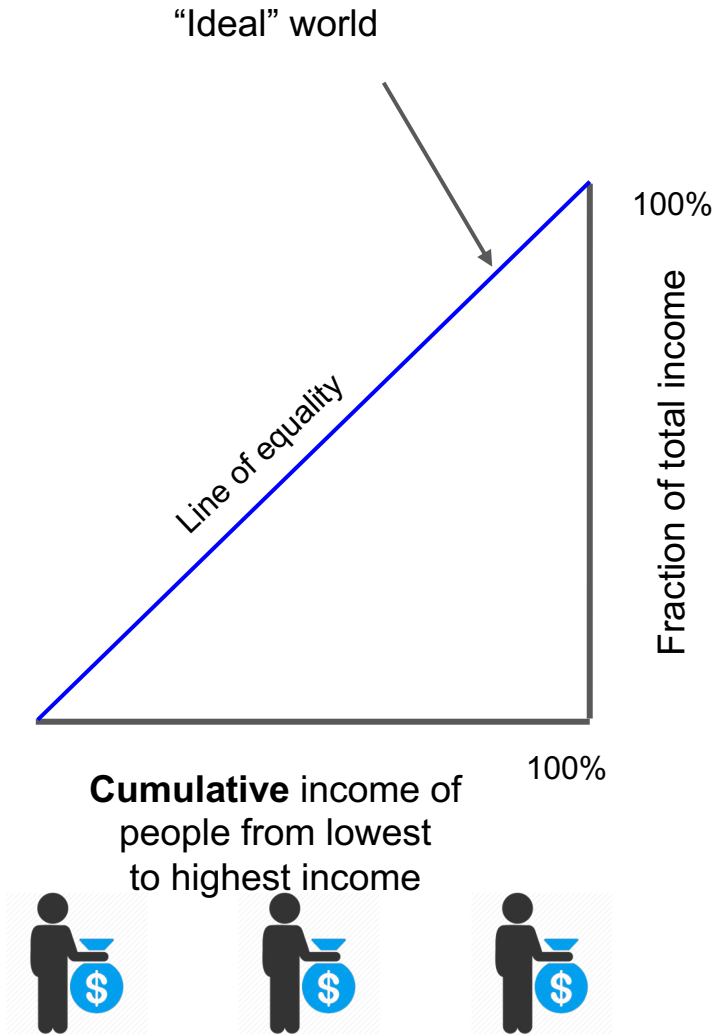
# Quality Control on mapped reads

# 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.

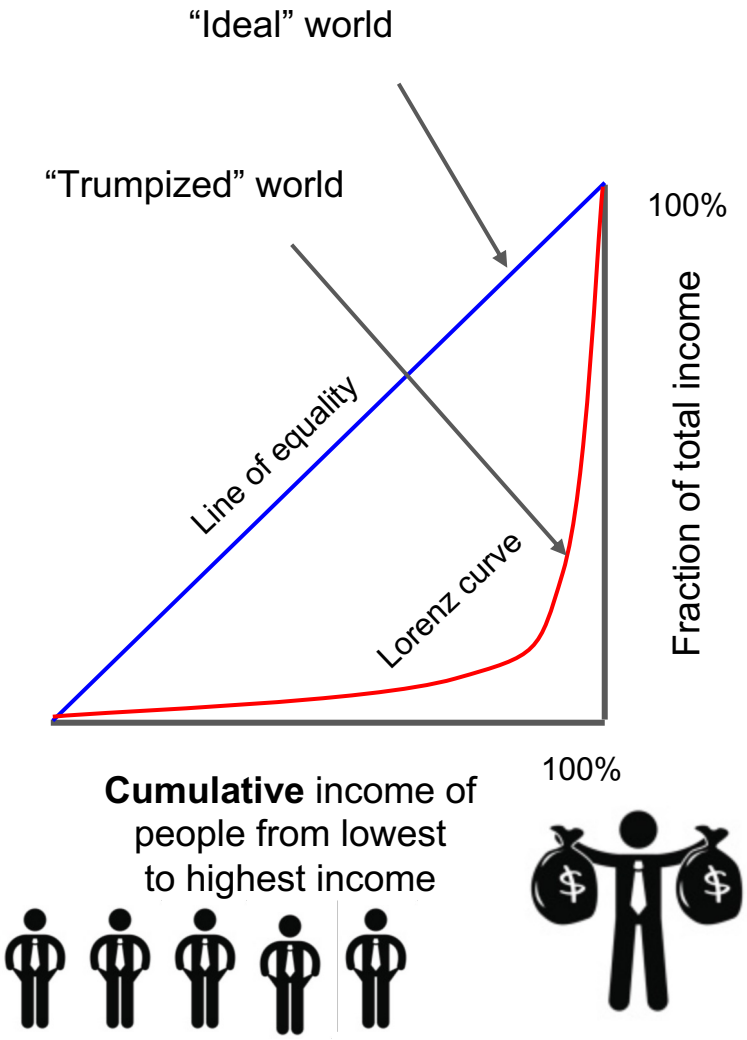
# Lorenz curve

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



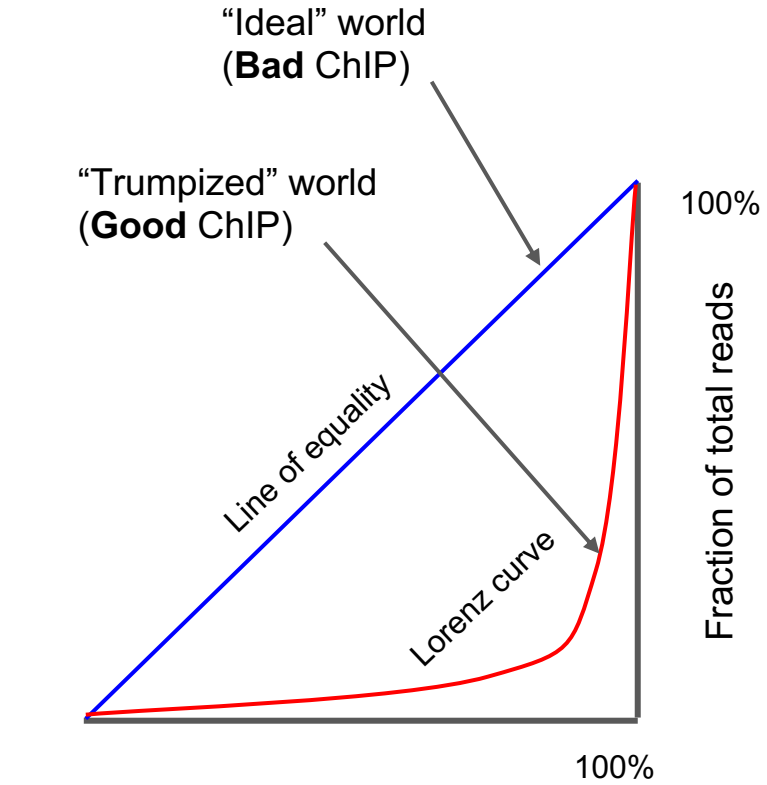
# Lorenz curve

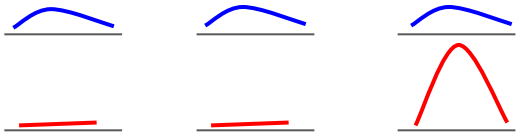
- Analyze income among workers by computing cumulative sum.
  - If uniform income distribution :
    - Straight line
  - If they were trumpized
    - Lorenz curve



# Lorenz curve

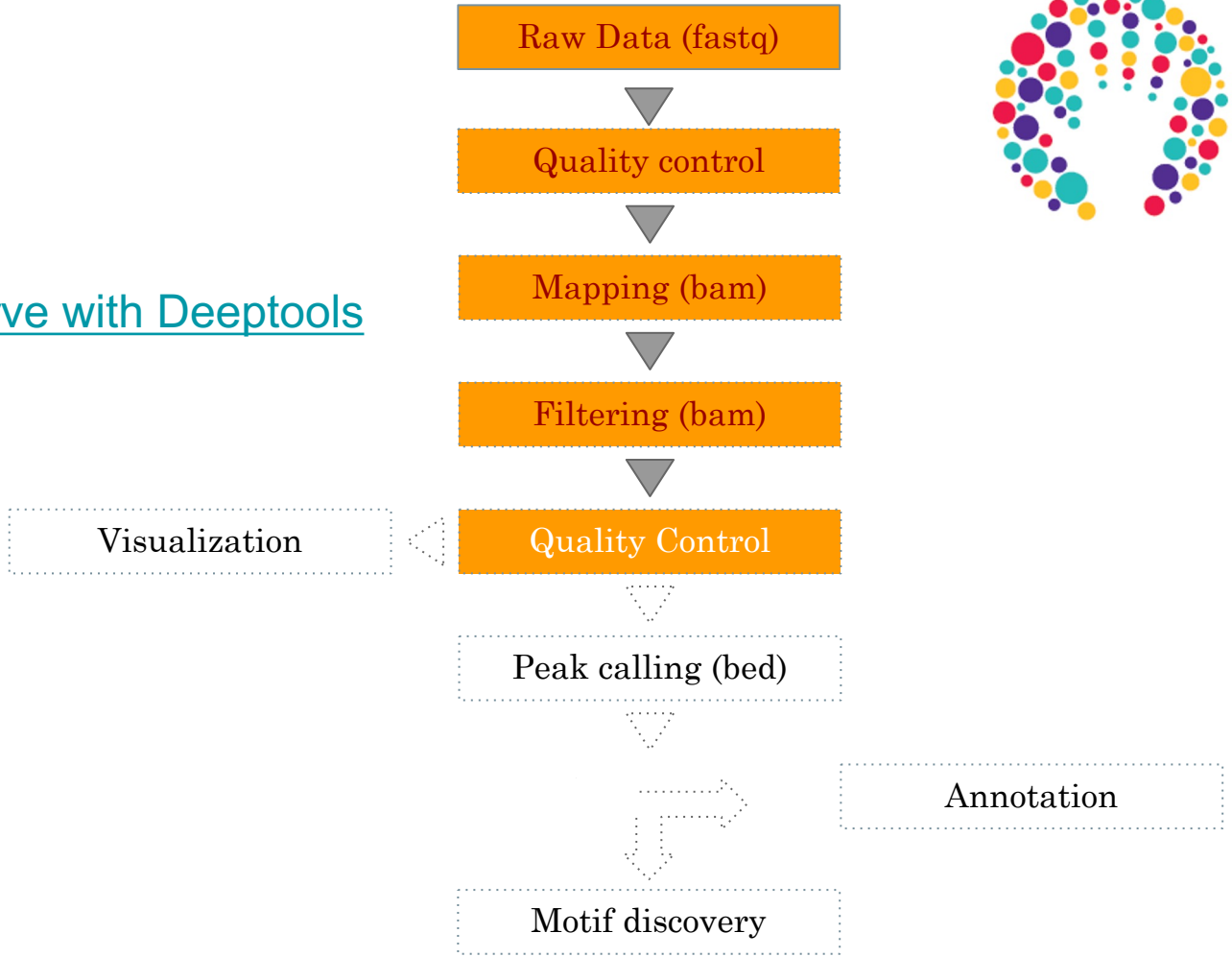
- Analyze income among workers by computing cumulative sum.
  - If uniform income distribution :
    - Straight line
  - If they were trumpized
    - Lorenz curve
- Here the workers are the genome windows and incomes are reads



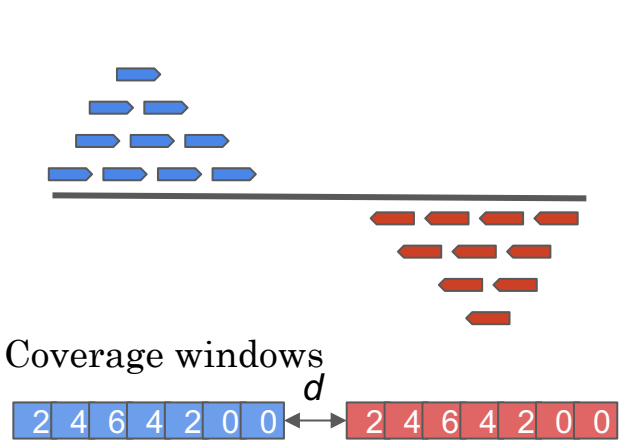
# Protocol

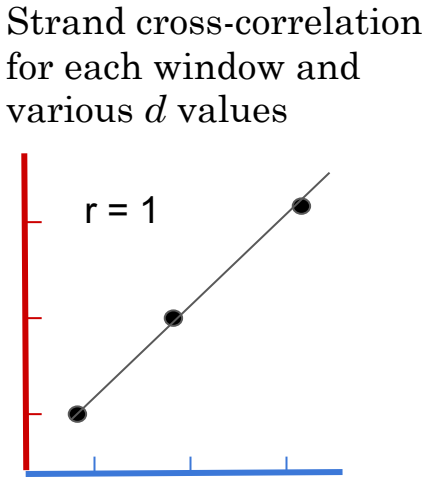
- Plot the Lorenz curve with Deeptools

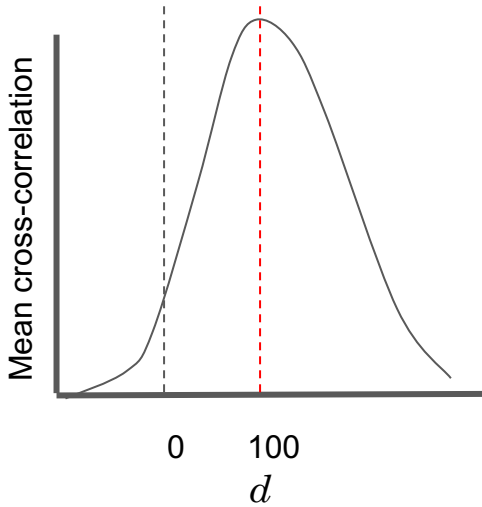


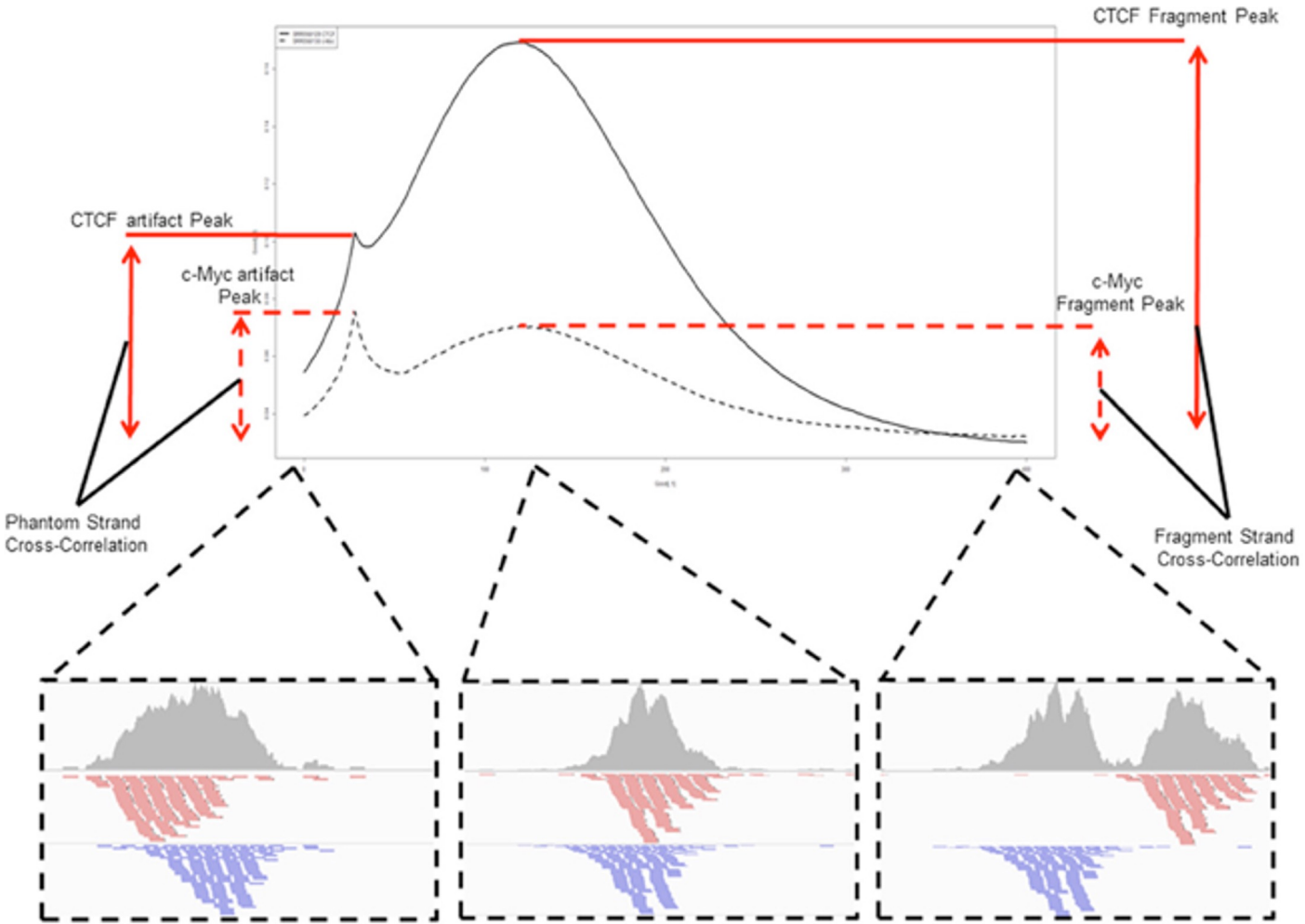
# Strand cross-correlation

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



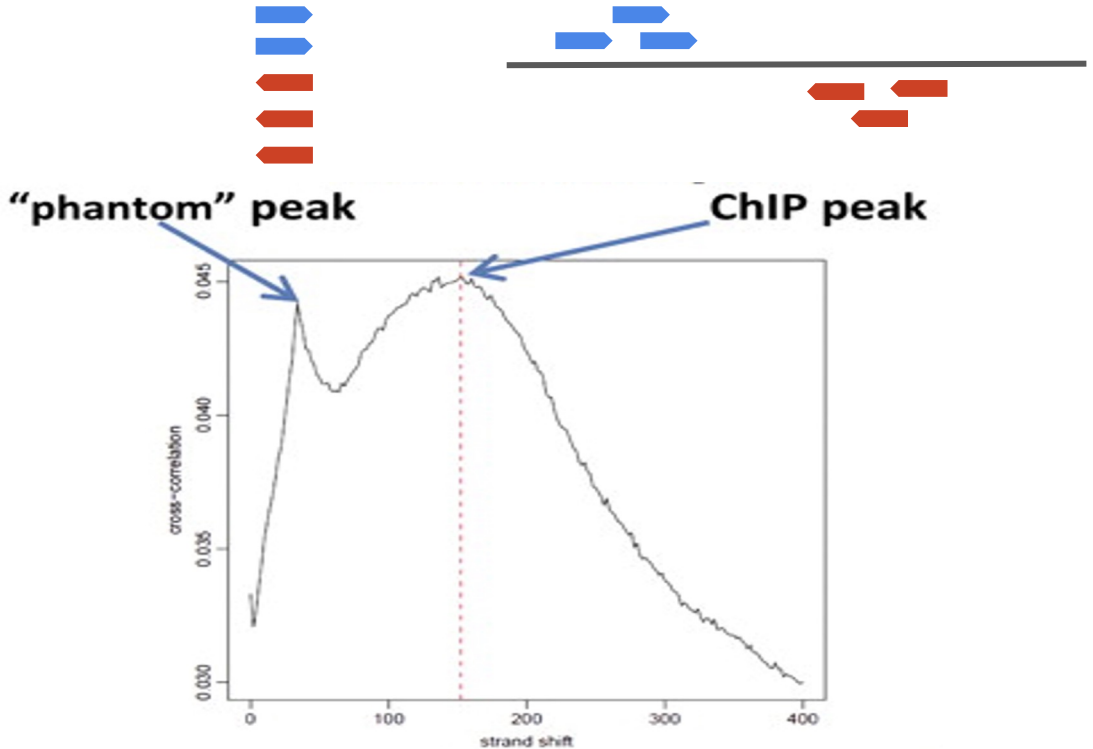



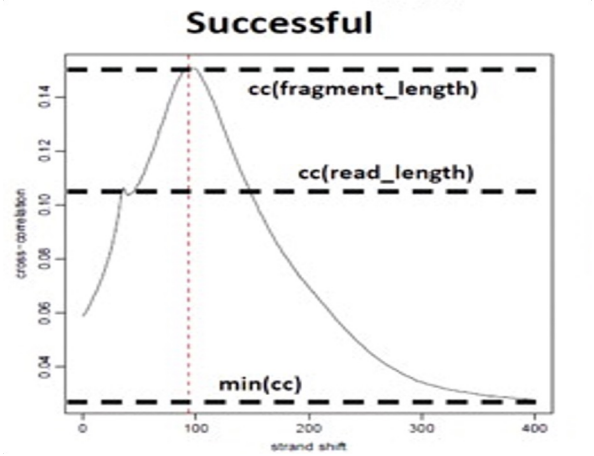


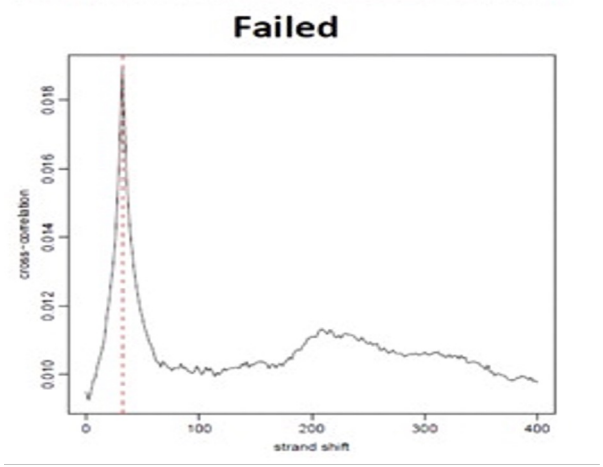

Carroll et al, *Frontiers in genetics*, 2014

# Strand cross-correlation



# Strand cross-correlation




NSC: normalized strand coefficient

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

NSC ≥ 1.05 is recommended

Relative strand correlation (RSC)

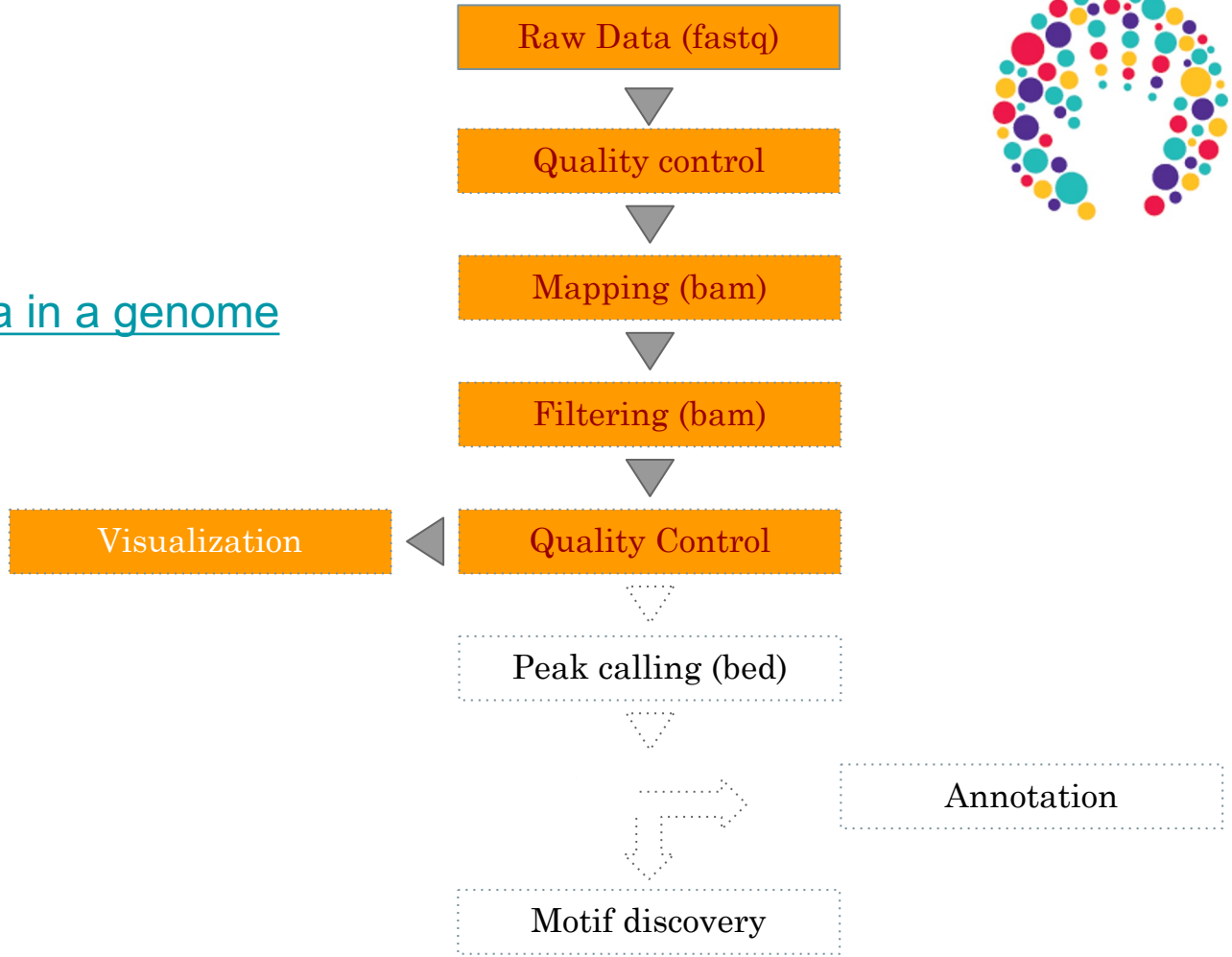
$$RSC = \frac{cc(fragment\ length) - min(cc)}{cc(read\ length) - min(cc)}$$

RSC ≥ 0.8 is recommended

# Visualization: computing a genomic coverage file

# Protocol

- Visualizing the data in a genome browser



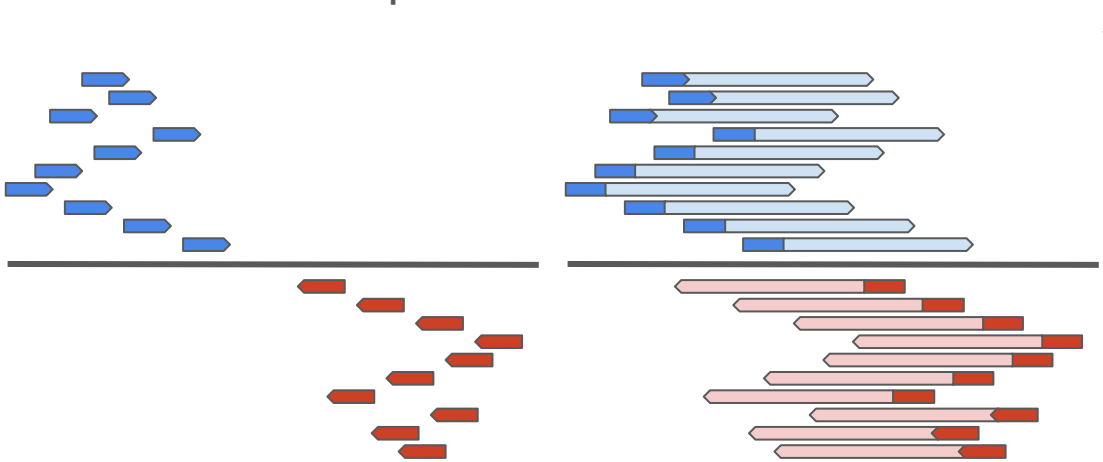
# Bam files are fat

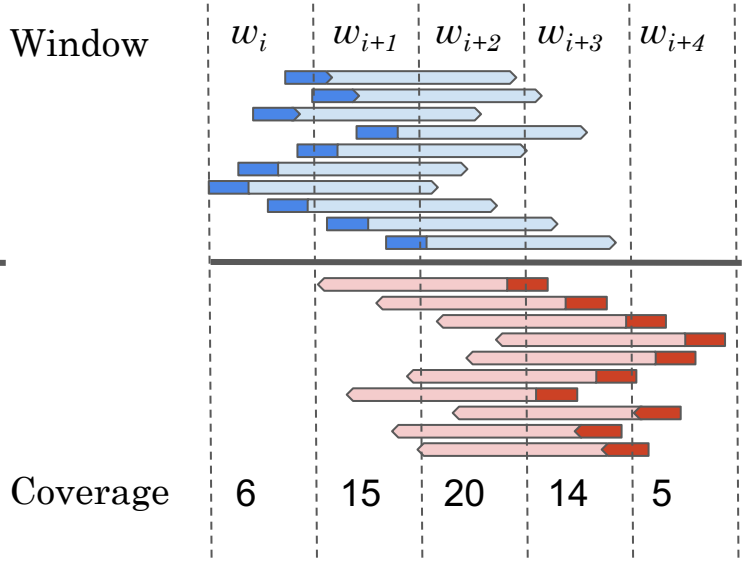
- **BAM files are fat** as they do contain exhaustive information about read alignments
  - Memory issues (can only visualize fraction of the BAM)

- 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

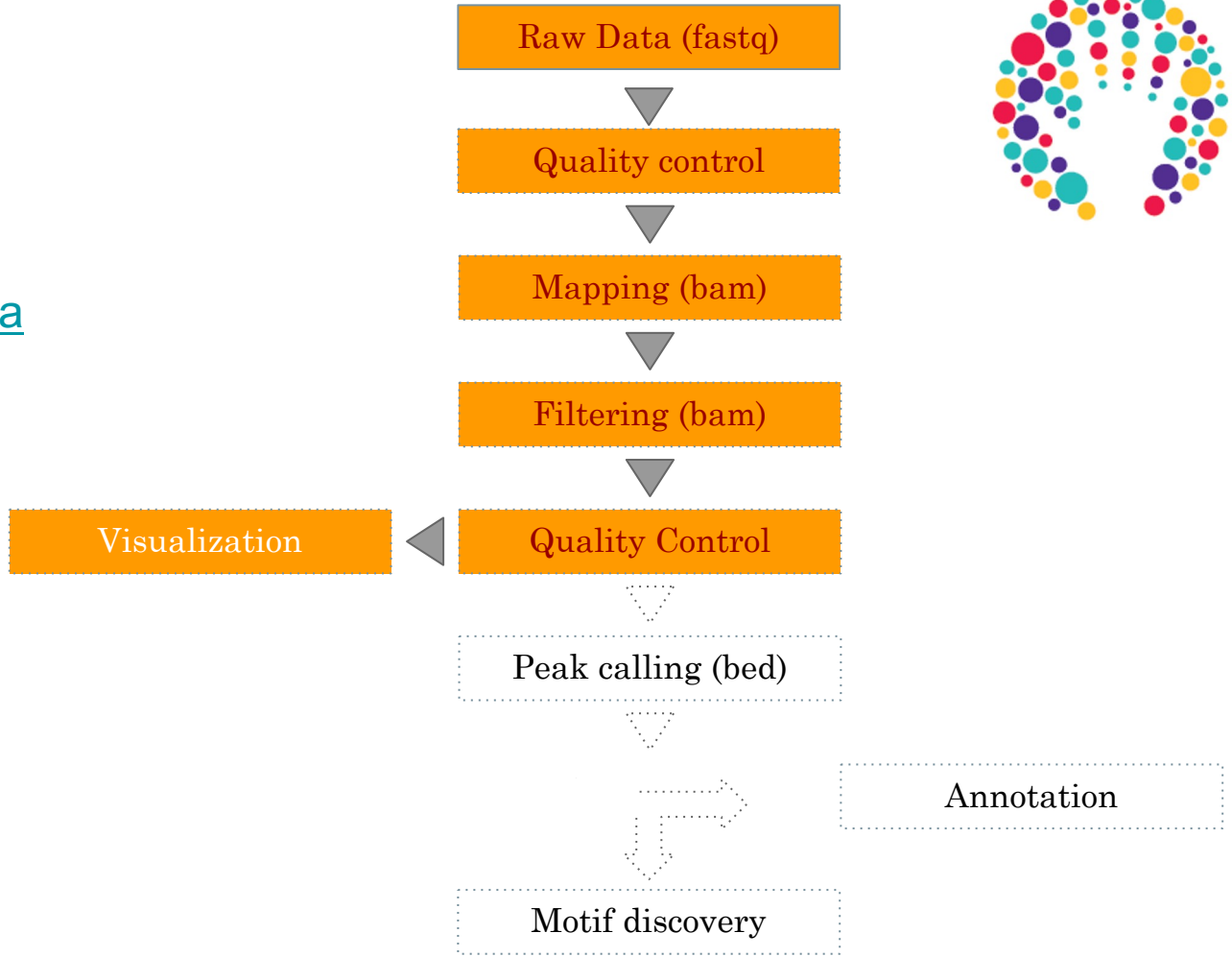
- BAM files **do not contain fragment location** but read location
- We need to extend reads to compute fragments coordinates before coverage analysis
- Not required for PE





# Protocol

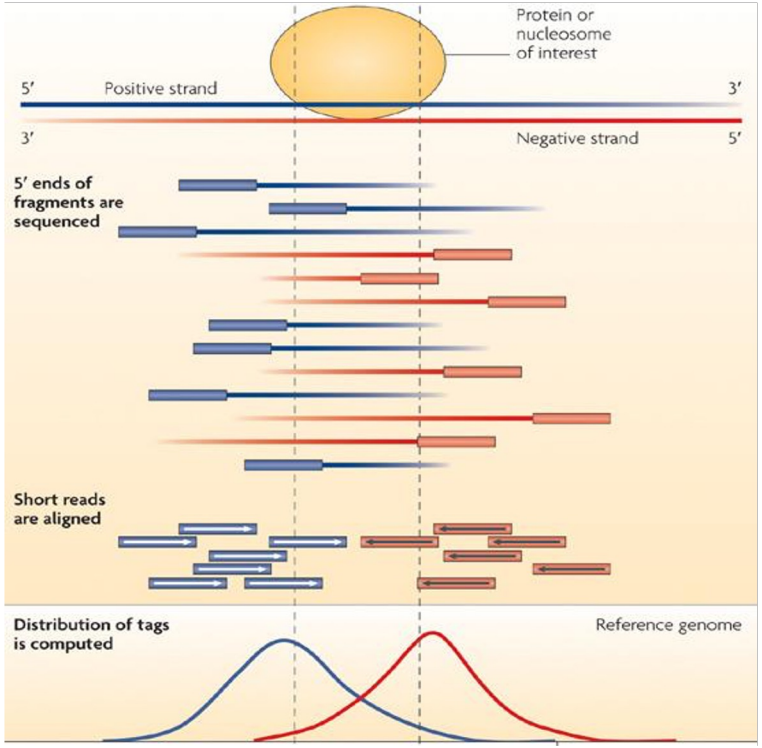
- Viewing scaled data



# Peak Calling

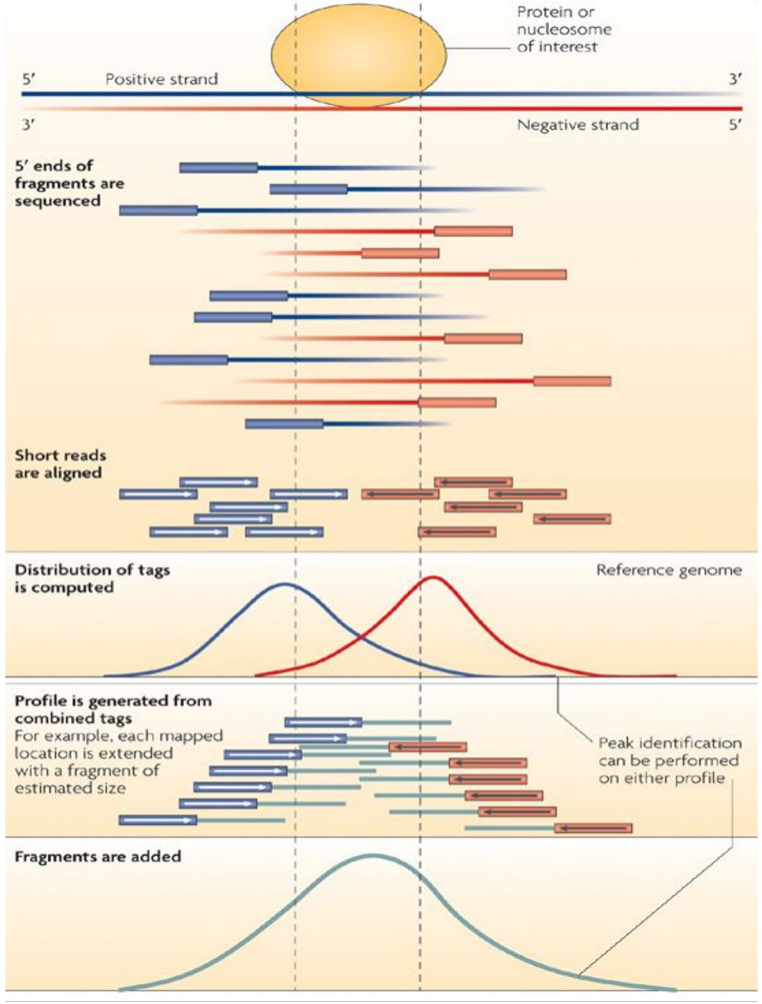
# From reads to peaks

- 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



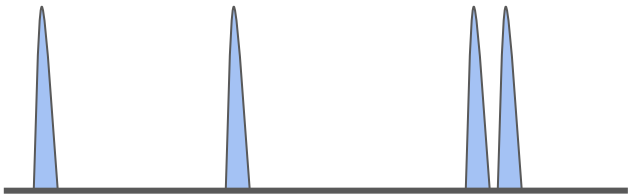
# From reads to peaks

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

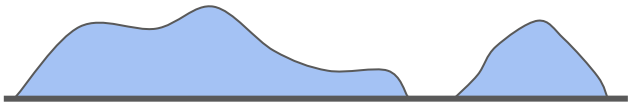


# Peak callers

- 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

# Peak callers

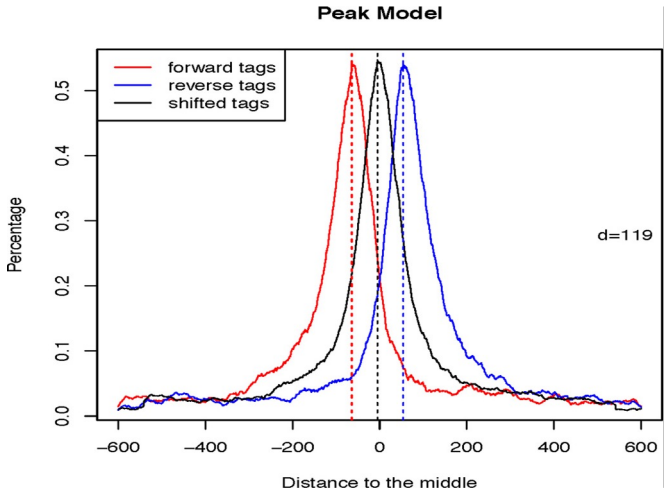
| | Profile | Peak criteria[a] | Tag shift | Control data[b] | Rank by | FDR[c] | User input parameters[d] | Artifact filtering: strand-based/ duplicate[e] | Refs. |
|---|---|---|---|---|---|---|---|---|---|
| CisGenome v1.1 | Strand-specific window scan | 1: Number of reads in window 2: Number of ChIP reads minus control reads in window | Average for highest ranking peak pairs | Conditional binomial used to estimate FDR | Number of reads under peak | 1: Negative binomial 2: conditional binomial | Target FDR, optional window width, window interval | Yes / Yes | 10 |
| ERANGE v3.1 | Tag aggregation | 1: Height cutoff High quality peak estimate, per-region estimate, or input | High quality peak estimate, per-region estimate, or input | Used to calculate fold enrichment and optionally *P* values | *P* value | 1: None 2: $\frac{\#\ control}{\#\ ChIP}$ | Optional peak height, ratio to background | Yes / No | 4,18 |
| FindPeaks v3.1.9.2 | Aggregation of overlapped tags | Height threshold | Input or estimated | NA | Number of reads under peak | 1: Monte Carlo simulation 2: NA | Minimum peak height, subpeak valley depth | Yes / Yes | 19 |
| F-Seq v1.82 | Kernel density estimation (KDE) | *s* s.d. above KDE for 1: random background, 2: control | Input or estimated | KDE for local background | Peak height | 1: None 2: None | Threshold s.d. value, KDE bandwidth | No / No | 14 |
| GLITR | Aggregation of overlapped tags | Classification by height and relative enrichment | User input tag extension | Multiply sampled to estimate background class values | Peak height and fold enrichment | 2: $\frac{\#\ control}{\#\ ChIP}$ | Target FDR, number nearest neighbors for clustering | No / No | 17 |
| MACS v1.3.5 | Tags shifted then window scan | Local region Poisson *P* value | Estimate from high quality peak pairs | Used for Poisson fit when available | *P* value | 1: None 2: $\frac{\#\ control}{\#\ ChIP}$ | *P*-value threshold, tag length, mfold for shift estimate | No / Yes | 13 |
| PeakSeq | Extended tag aggregation | Local region binomial *P* value | Input tag extension length | Used for significance of sample enrichment with binomial distribution | *q* value | 1: Poisson background assumption 2: From binomial for sample plus control | Target FDR | No / No | 5 |
| QuEST v2.3 | Kernel density estimation | 2: Height threshold, background ratio | Mode of local shifts that maximize strand cross-correlation | KDE for enrichment and empirical FDR estimation | *q* value | 1: NA 2: $\frac{\#\ control}{\#\ ChIP}$ as a function of profile threshold | KDE bandwidth, peak height, subpeak valley depth, ratio to background | Yes / Yes | 9 |
| SICER v1.02 | Window scan with gaps allowed | *P* value from random background model, enrichment relative to control | Input | Linearly rescaled for candidate peak rejection and *P* values | *q* value | 1: None 2: From Poisson *P* values | Window length, gap size, FDR (with control) or *E*-value (no control) | No / Yes | 15 |
| SiSSRs v1.4 | Window scan | $N_- - N_+$ sign change, $N_+ +$ $N_-$ threshold in | Average nearest paired tag distance | Used to compute fold-enrichment distribution | *P* value | 1: Poisson 2: control distribution | 1: FDR 1,2: $N_+ + N_-$ threshold | Yes / Yes | 11 |

Pepke et al, 2009

# MACS [Zhang et al, 2008]

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

- **slides *2*bandwidth* windows** across the genome to find regions with tags **more than *mfold* 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

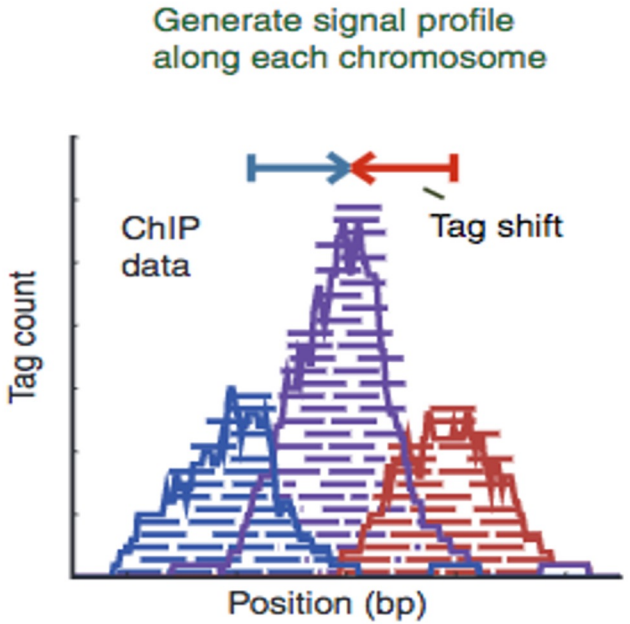


# MACS [Zhang et al, 2008]

**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

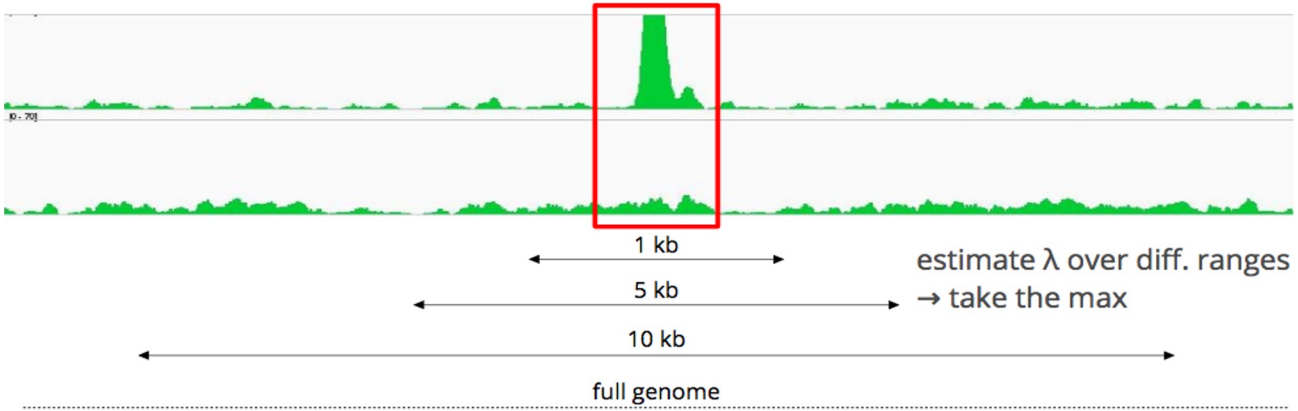
(*d* value is the model obtained in step 1)

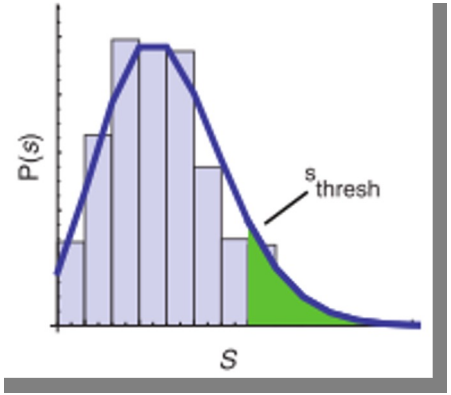



Pepke et al, 2009

# 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

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

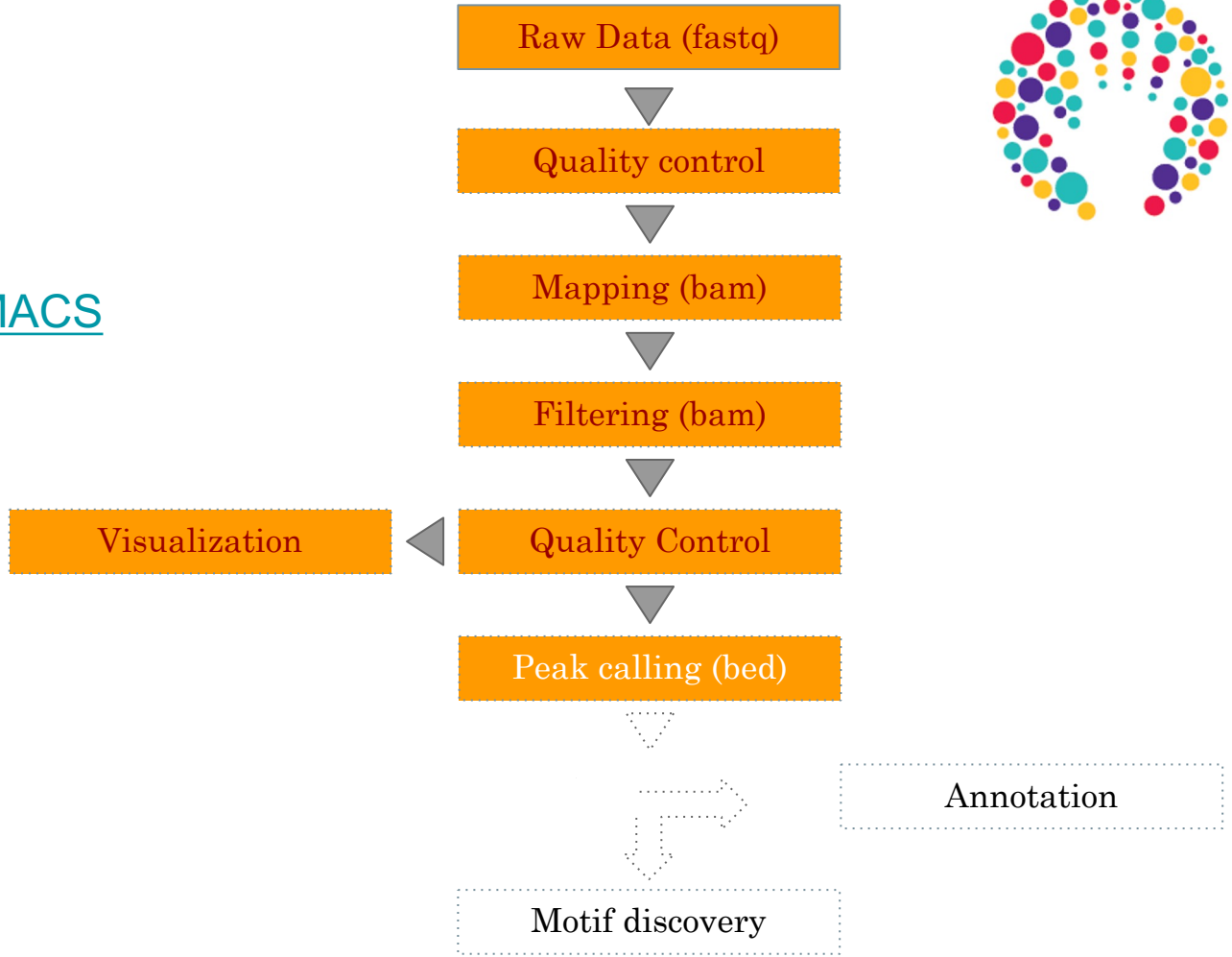
# MACS in summary

- 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

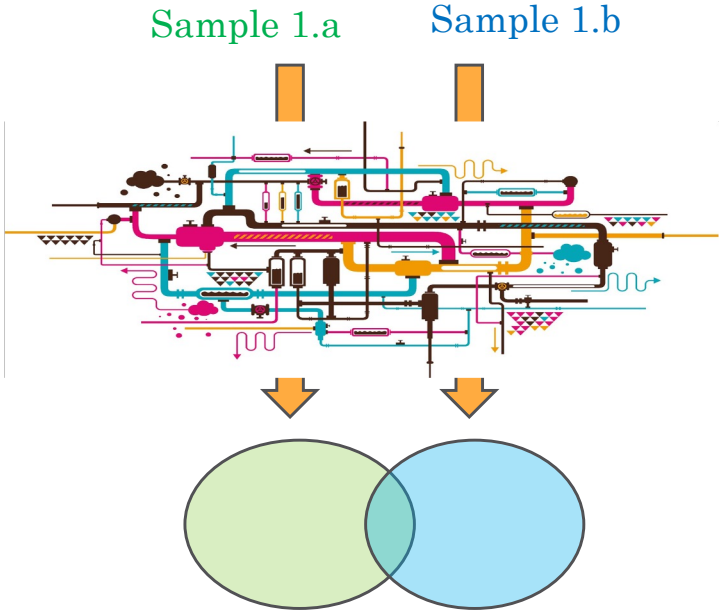
  (stop after step 5)



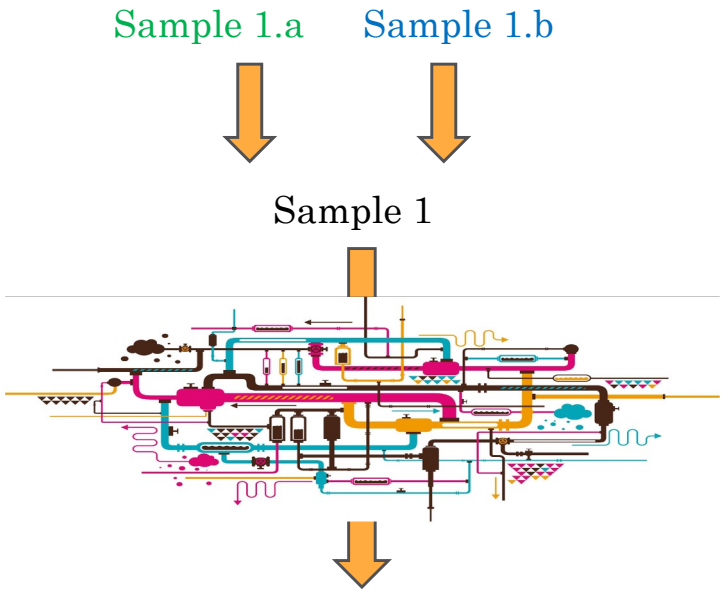
# How to deal with replicates

# 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”

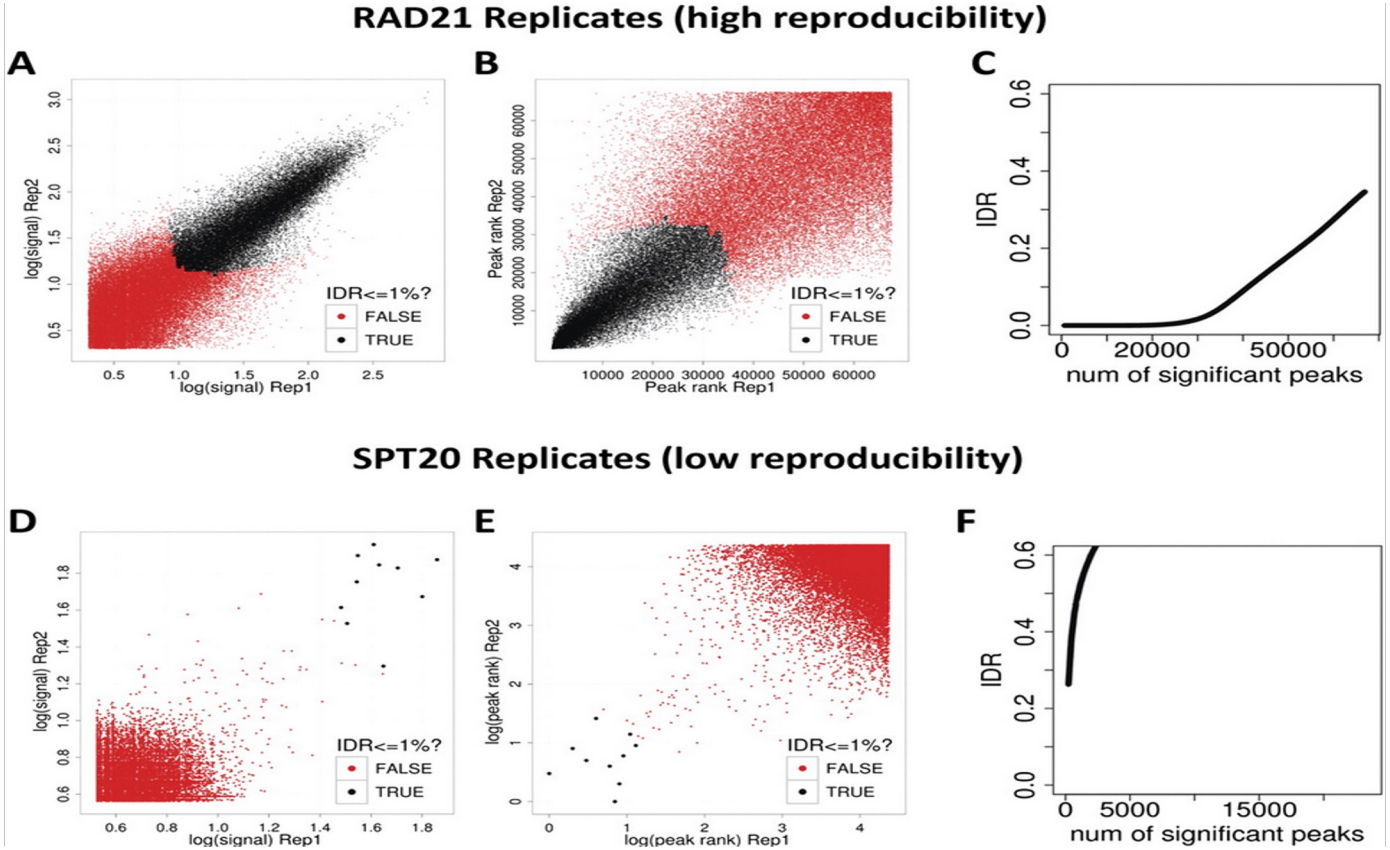


# IDR - Irreproducible Discovery Rate (ENCODE)

- 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

https://sites.google.com/site/anshulkundaje/projects/idr

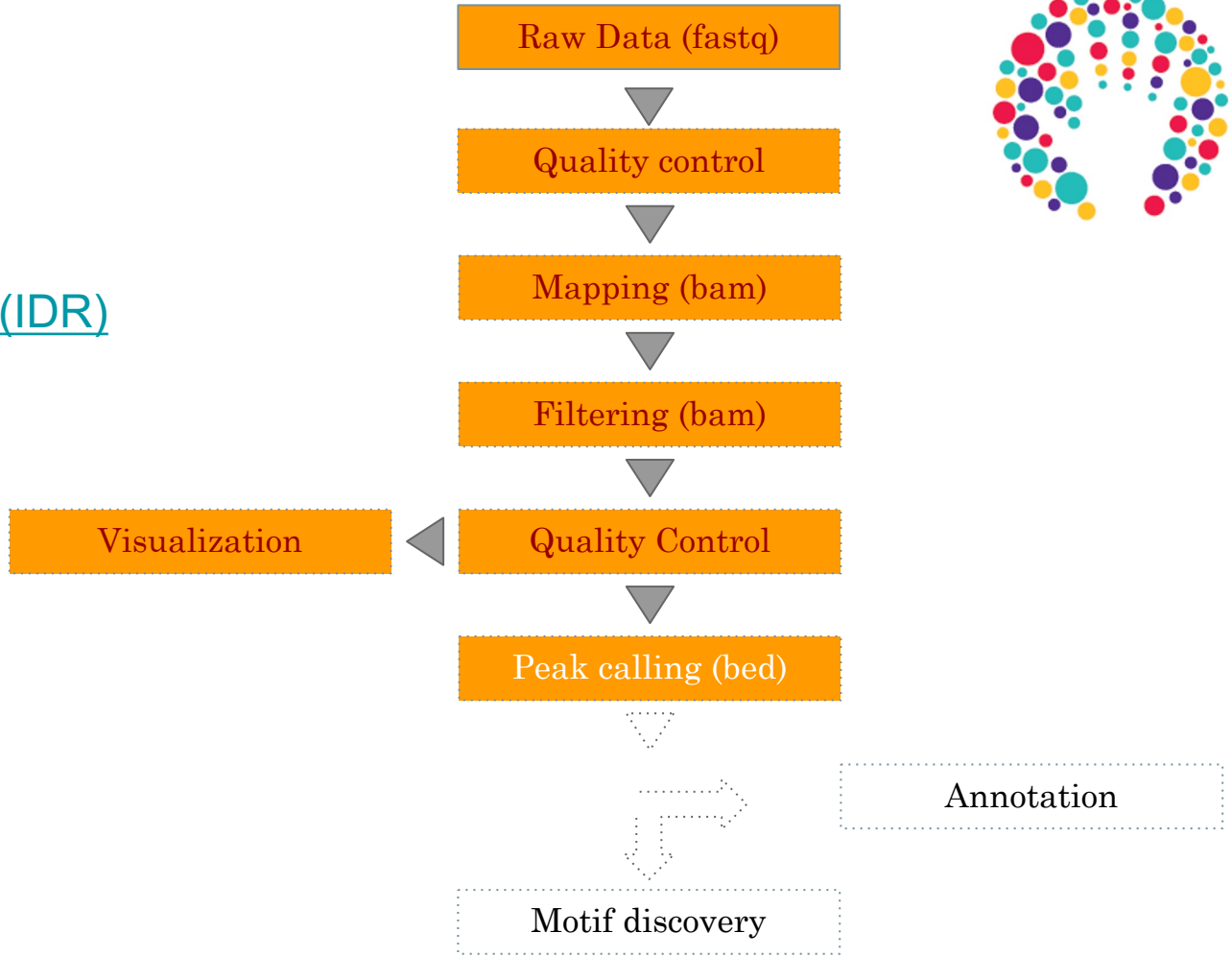
# IDR




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

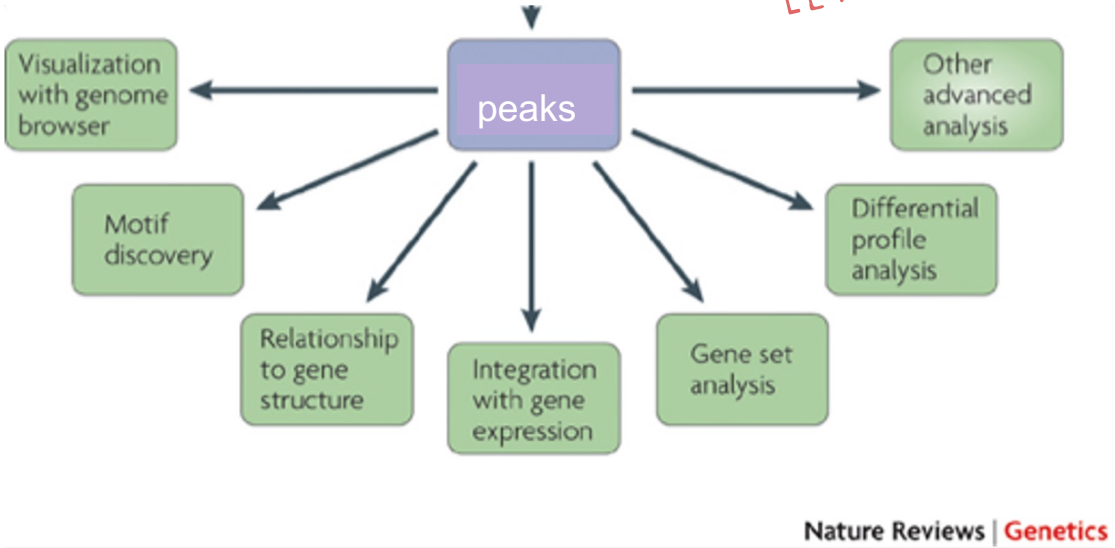
# Protocol

- Combine replicate (IDR)

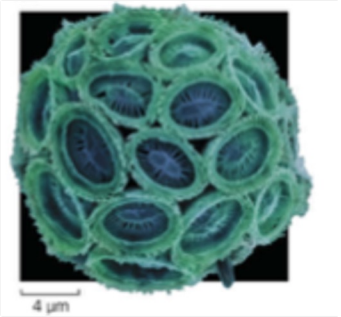


# Processing steps are over !

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

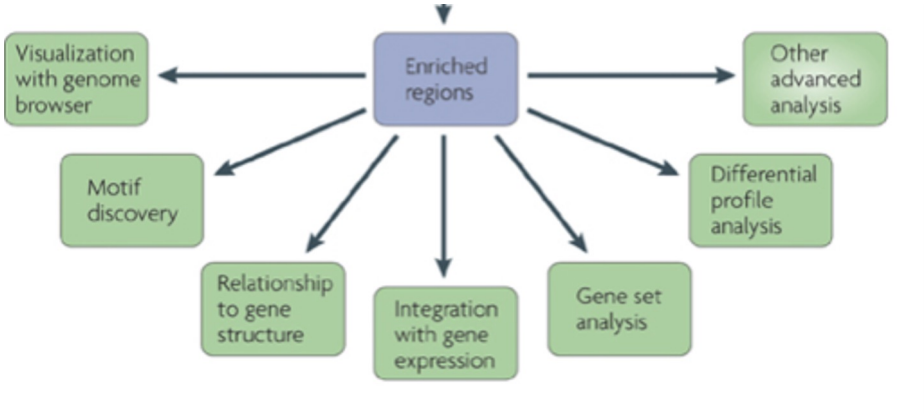
- Which **regulated genes** are directly regulated by a given TF ?

- 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 ?

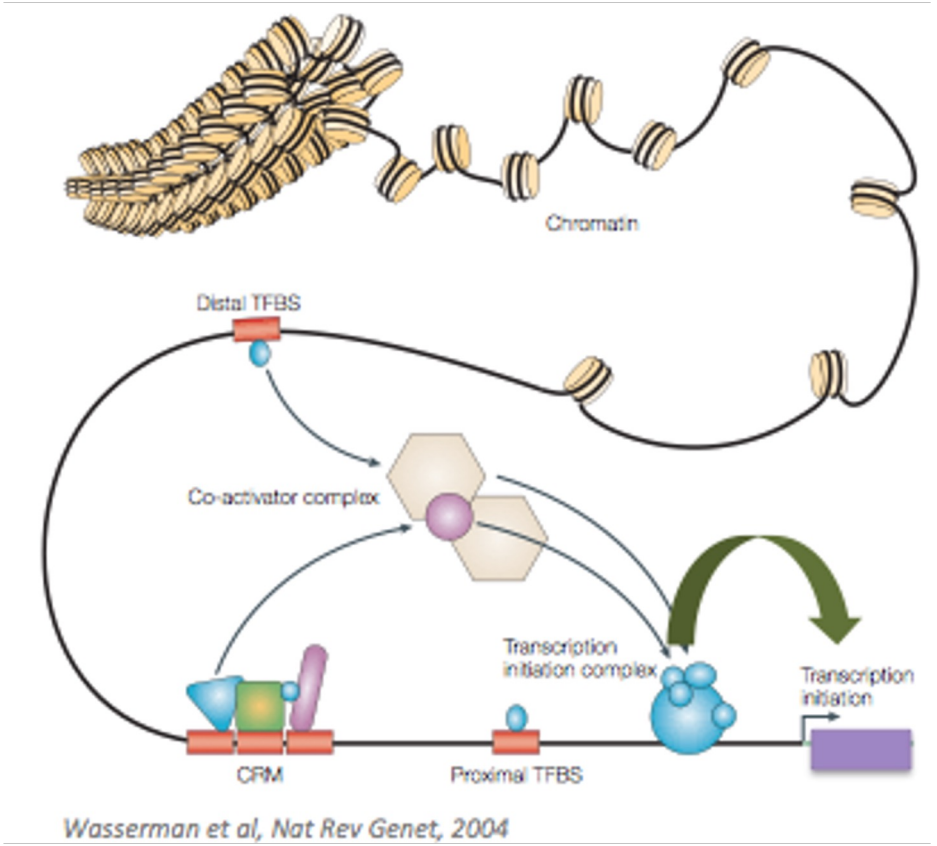
## What can be the following experimental work ?

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

# *De novo* motif discovery

# 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**

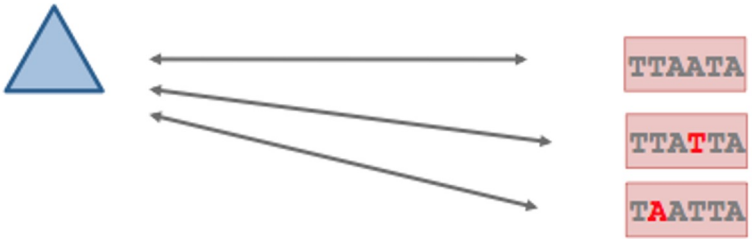



*Wasserman et al, Nat Rev Genet, 2004*

# Transcription factor specificity

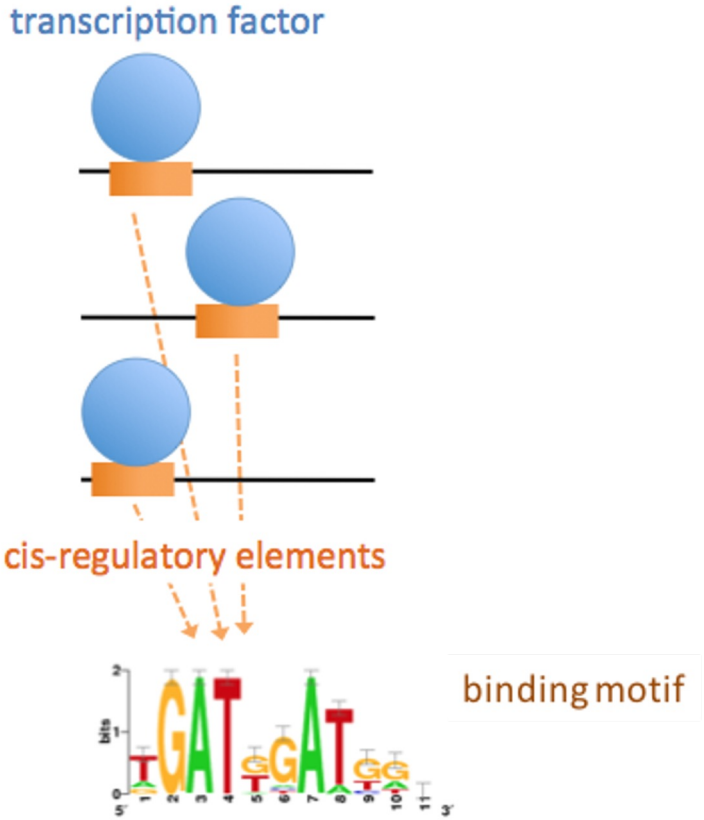
*How do TF « know » where to bind DNA ?*


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

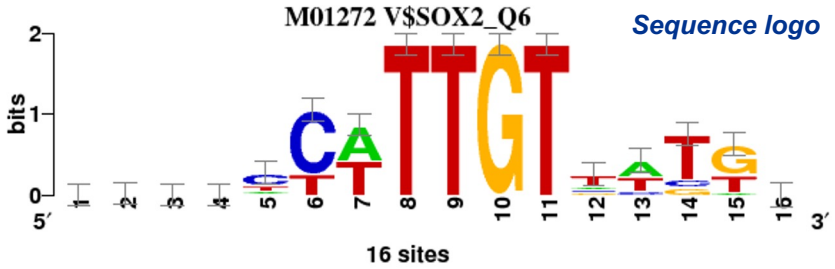
***Collection of binding sites***
*used to build the Sox2 matrix*
*(TRANSFAC M01272 )*

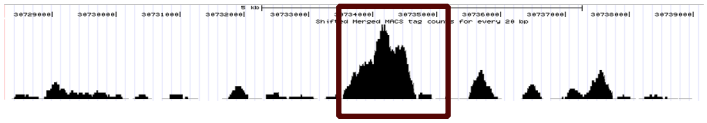
```
R15133   GCCCTCATTGTTATGC
R15201   AAACTCTTTGTTTGGA
R15231   TTCACCATTGTTCTAG
R15267   GACTCTATTGTCTCTG
R16367   GATATCTTTGTTTCTT
R17099   TGCACCTTTGTTATGC
R19276   AATTCCATTGTTATGA
R19367   AAACTCTTTGTTTGGA
R19510   ATGGACATTGTAATGC
R22342   AGGCCTTTTGTCCTGG
R22344   TGTGCTTTTGTNNNNN
R22359   CTCAACTTTGTAATTT
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 |

***Sequence logo***





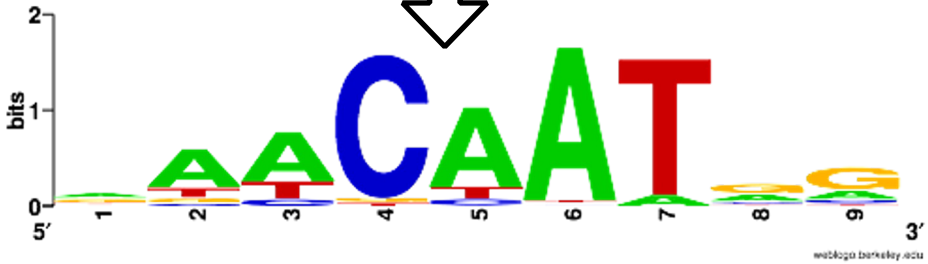
ChIP-seq peaks

```
>mm9_chr1_39249116_39251316_+
gagaggaaggggagaaagagggaggggagGGTGATAGGTAGCCAGGAG
CCAATGGGGGCGTTTTCCTTGTCCAGGCCACTTGCTGGAATGTGAGATGT
AGAATGACCCAAAGAGAGCTGCCAAGACAGAGCTCTGCCCCAGGAATTGA
ACTCAAAGGGTGTCAGAAAGCAGGTGGCCTTTGTGCACCTGGCGCGGGGA
CGTGGCTCCCCTCTTCCGGCTGGTCTAGCCAGGtgcctgcctgcctgcct
gccGTGATCTCTGGACGCCAGTAGAGGGTTGTTGTGGGTTTGGGTGAAAC
ACGCCACCCCTGAGCTCTTCCGCGGGGCTAGCAATCTCCCCATCACCCCA
TTCGCGCTCAGAACCCCCTCAGCGA TAACAGCAGGCCTGGTTCCCCG
```

DNA sequence

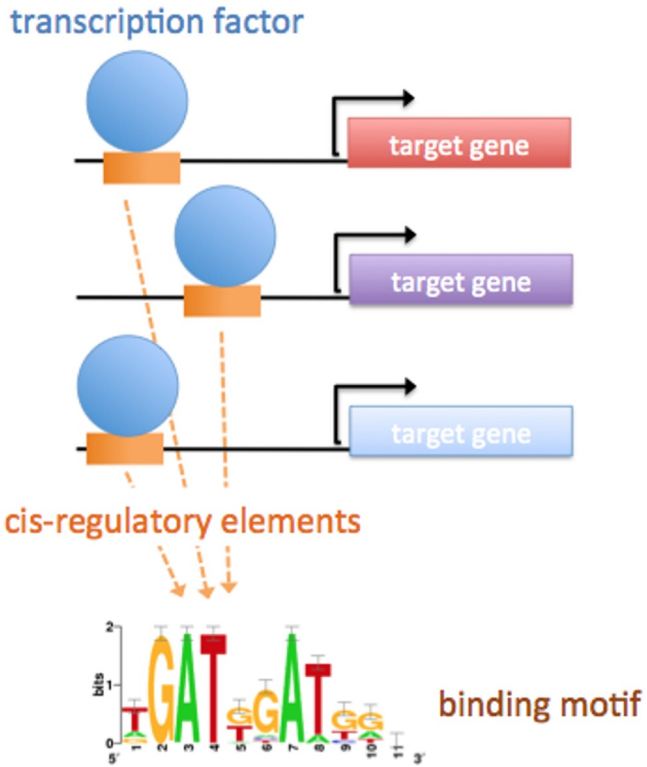
```
A  [24  54  59   0  65  71   4  24   9 ]
C  [ 7   6   4  72   4   2   0   6   9 ]
G  [31   7   0   2   0   1   1  38  55 ]
T  [14   9  13   2   7   2  71   8   3 ]
```

Discovered motif




Motif logo

# *De novo* motif discovery




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

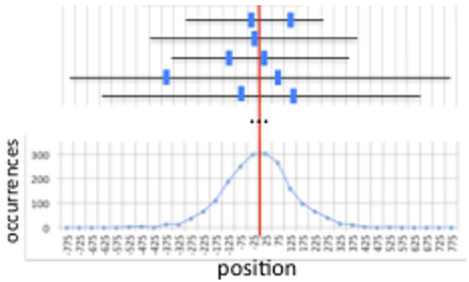
*If there is a common regulating factor, can we discover its motif only using these sequences ?*

# *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

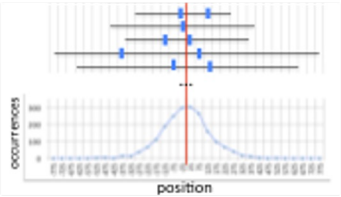
- MEME (Bailey et al., 1994)
- **RSAT oligo-analysis (van Helden et al., 1998)**
- AlignACE (Roth et al. 1998)
- **RSAT position-analysis (van Helden et al., 2000)**
- Weeder (Pavesi et al. 2001)
- MotifSampler (Thijs et al., 2001)
- … many others

*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 !!!!!!

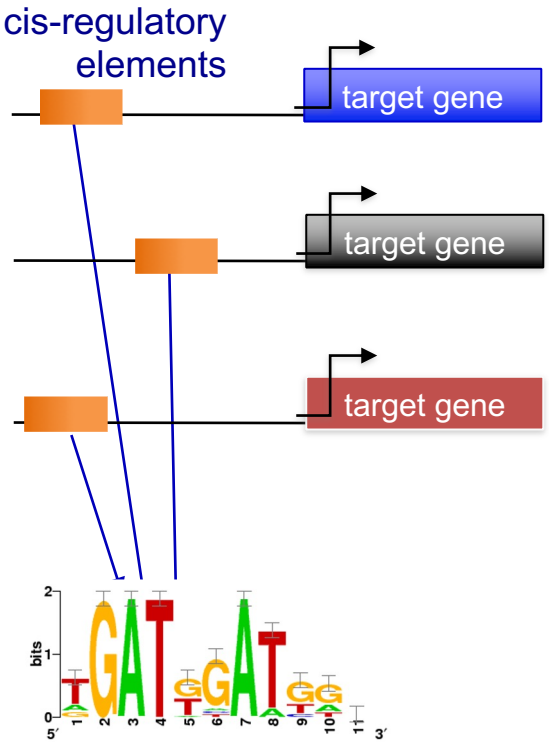

- **the problem is slightly different**
  - promoters: 200-2000bp from co-regulated genes
  - peaks: 300bp, positional bias



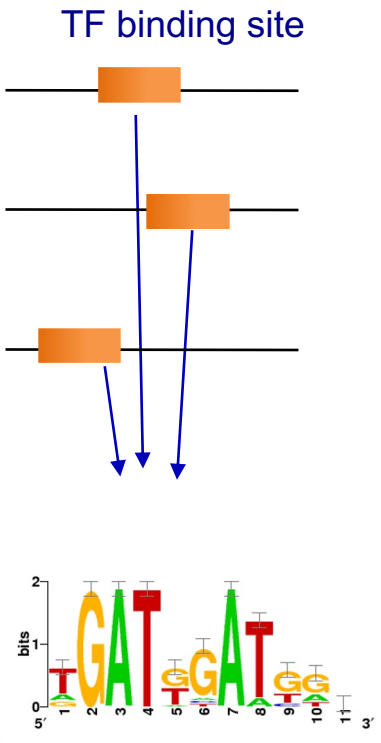

- **motif analysis: not just for specialists anymore !**
  - complete user-friendly workflows

# De novo motif discovery

## Case 1: promoters of co-expressed genes

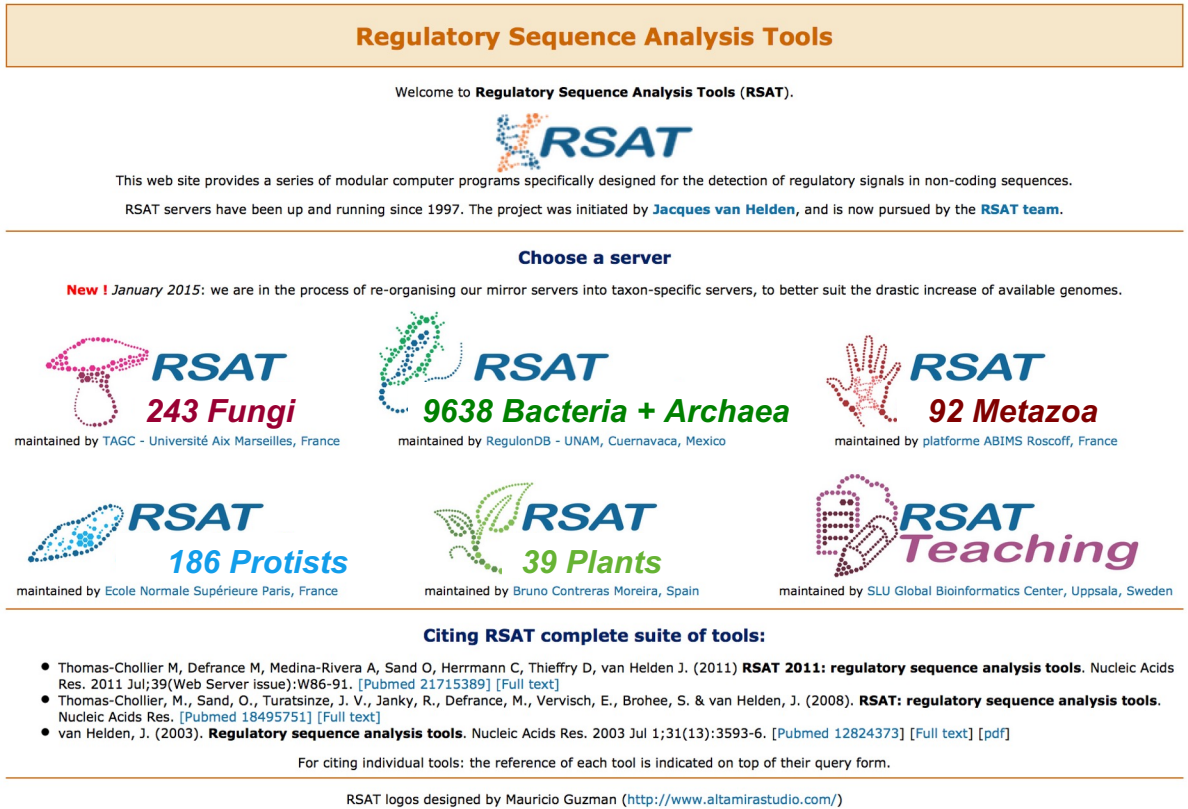



## Case 2: ChIP-seq peaks



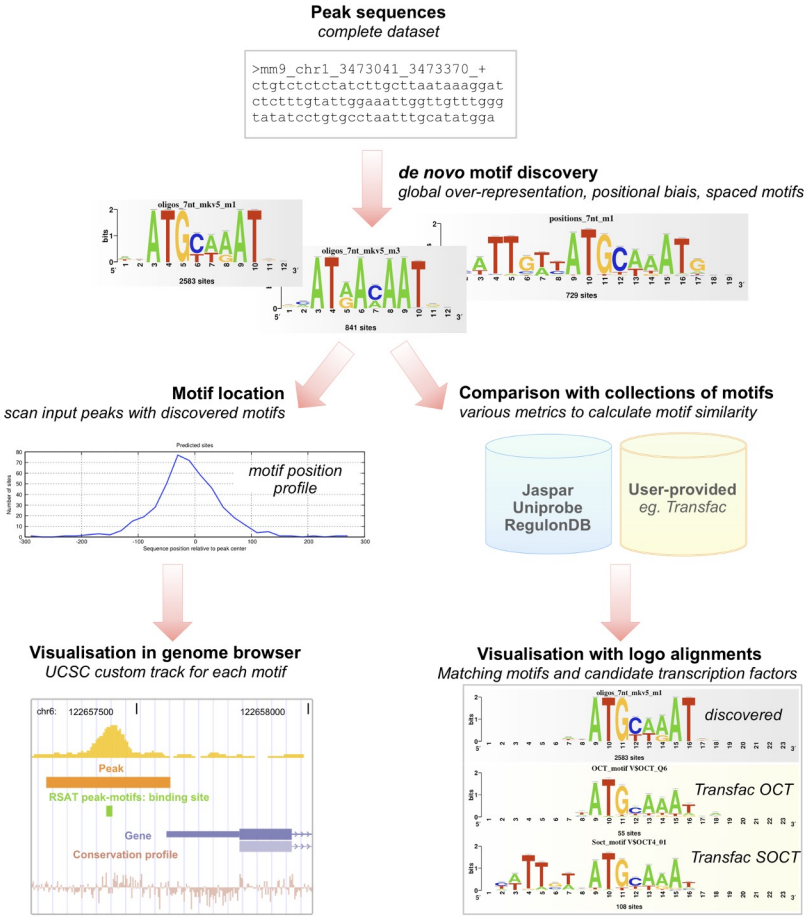

binding motif
(represented as a sequence logo)

# Regulatory sequence Analysis Tools (rsat.eu)



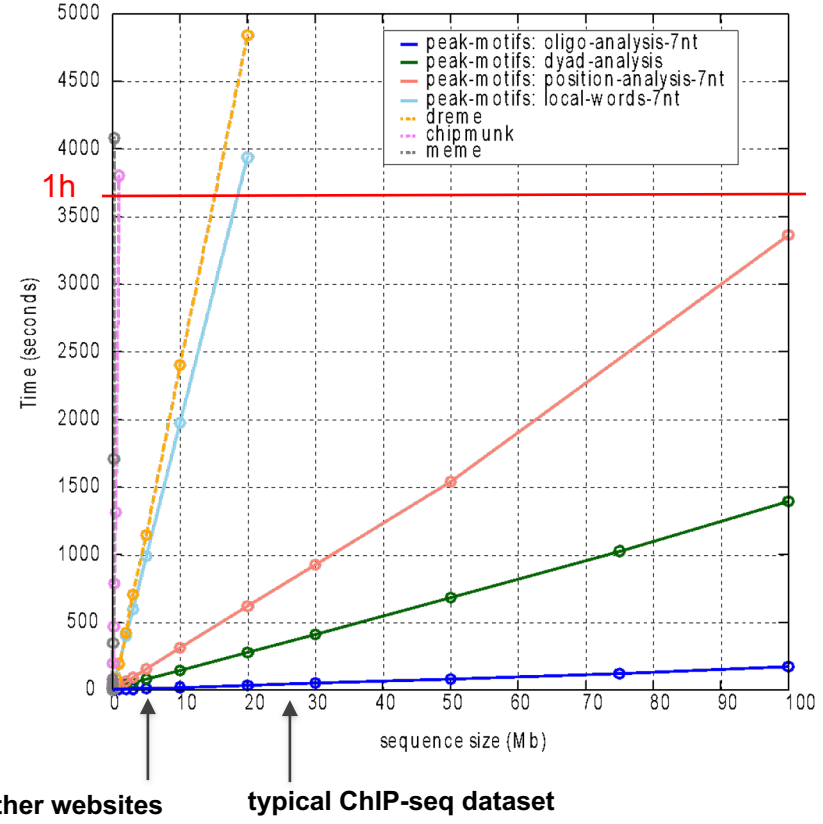
# Peak-motifs

- 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

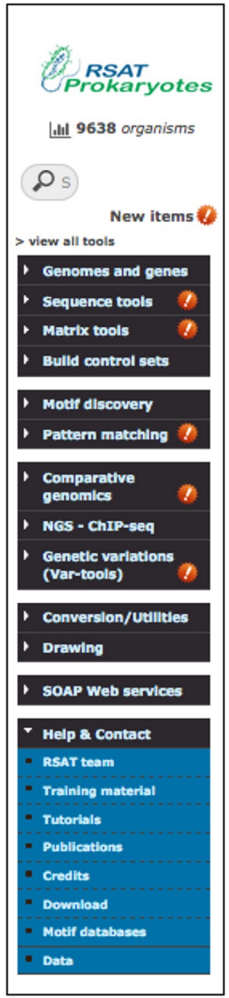



**size limit of other websites**

**typical ChIP-seq dataset**

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

# RSAT menu



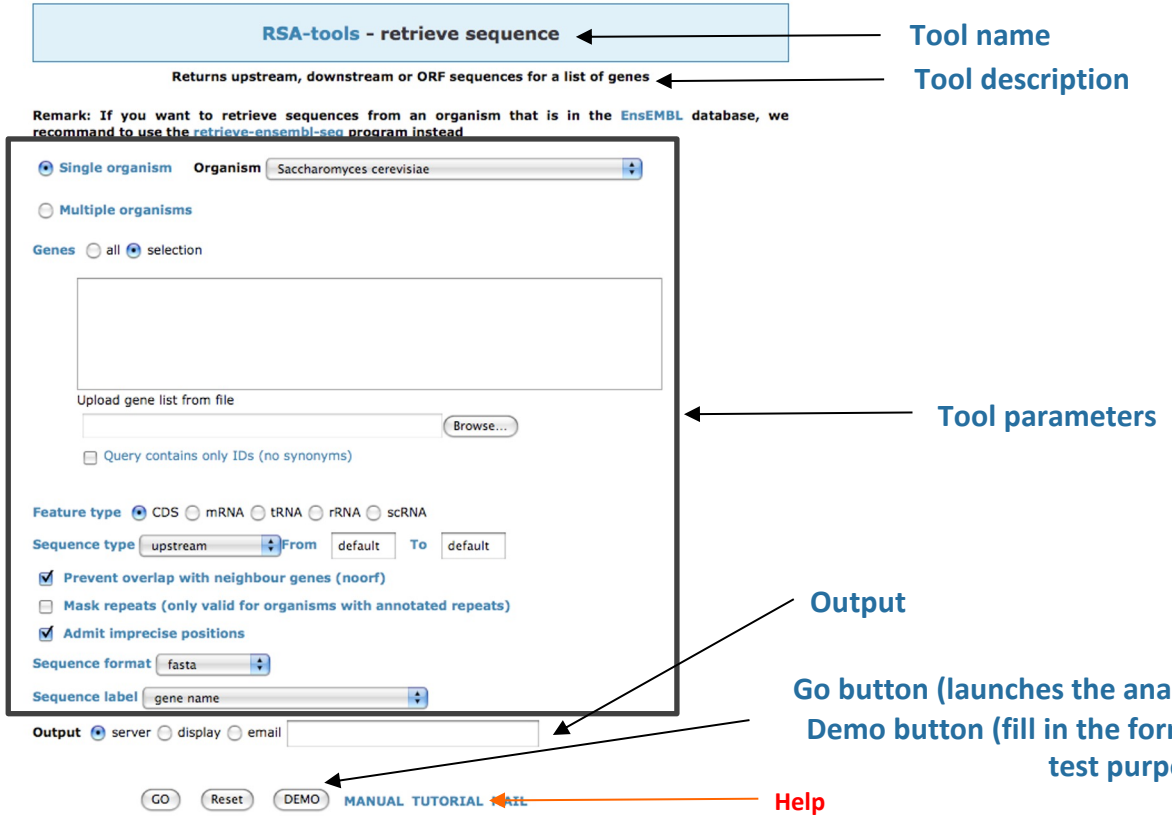

1. Get sequences

2. Run the analysis

3. Visualization

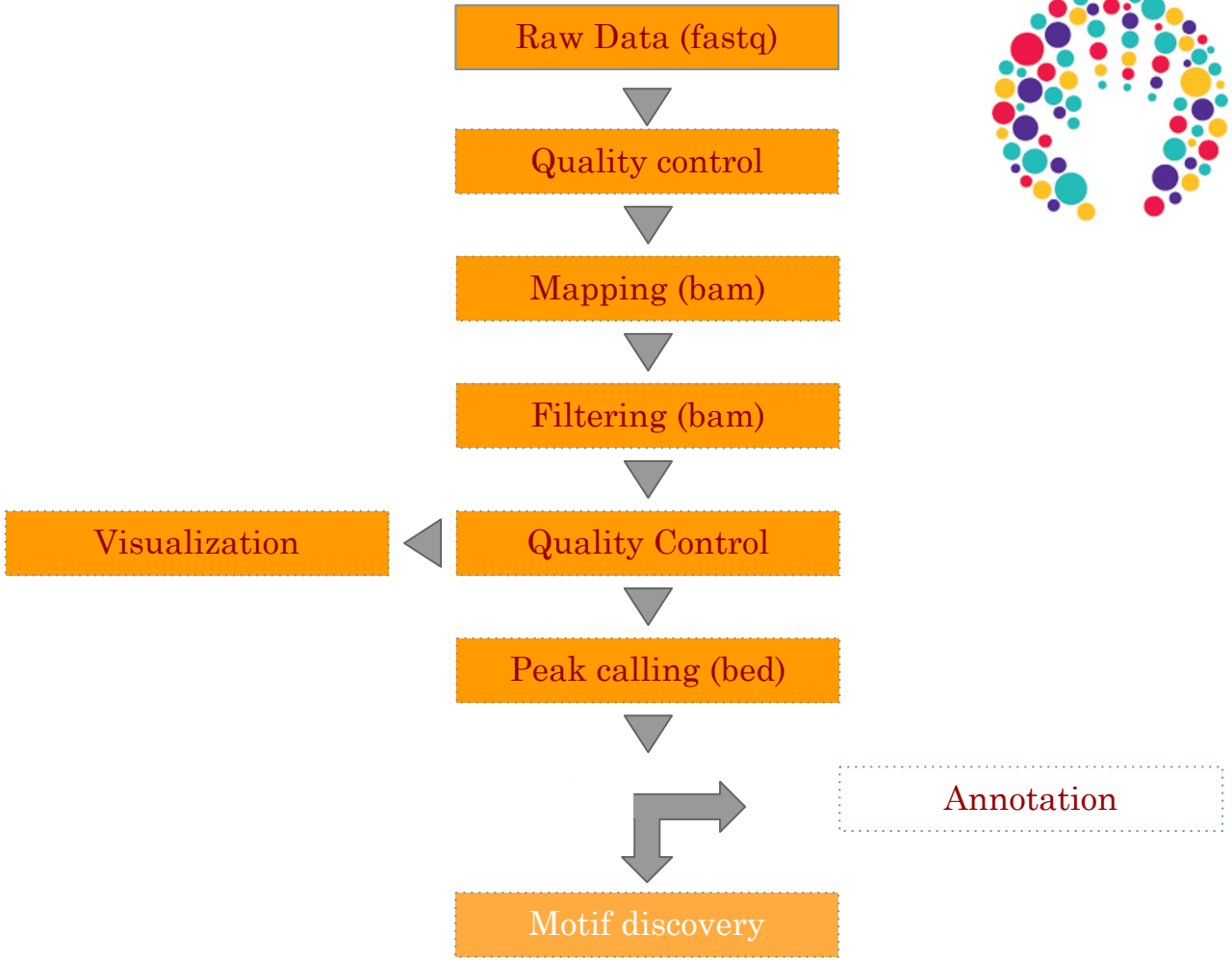
Help: tutorials,

# RSAT Web forms

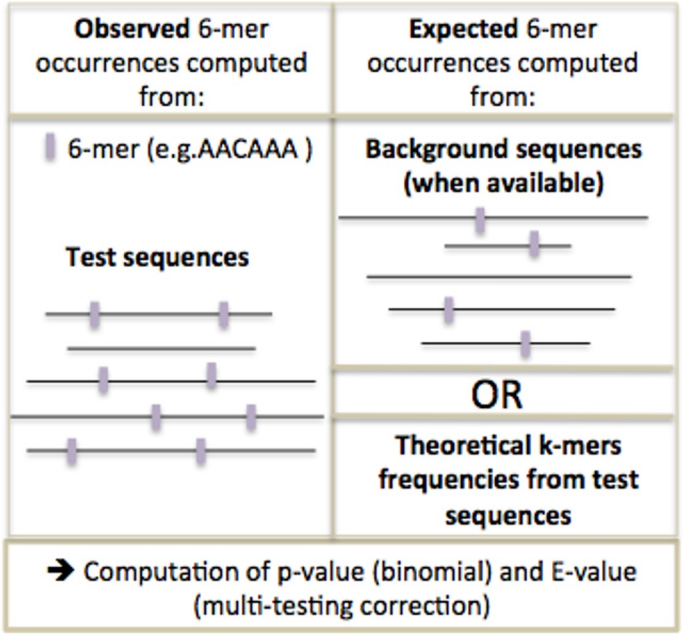


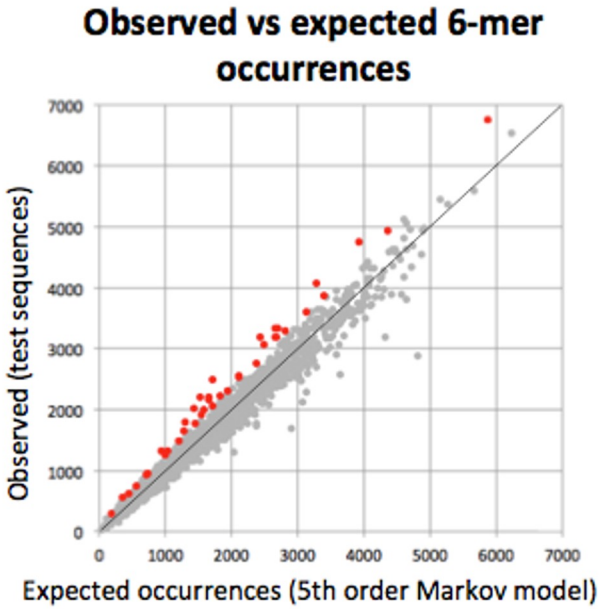
# Protocol

- Motif analysis

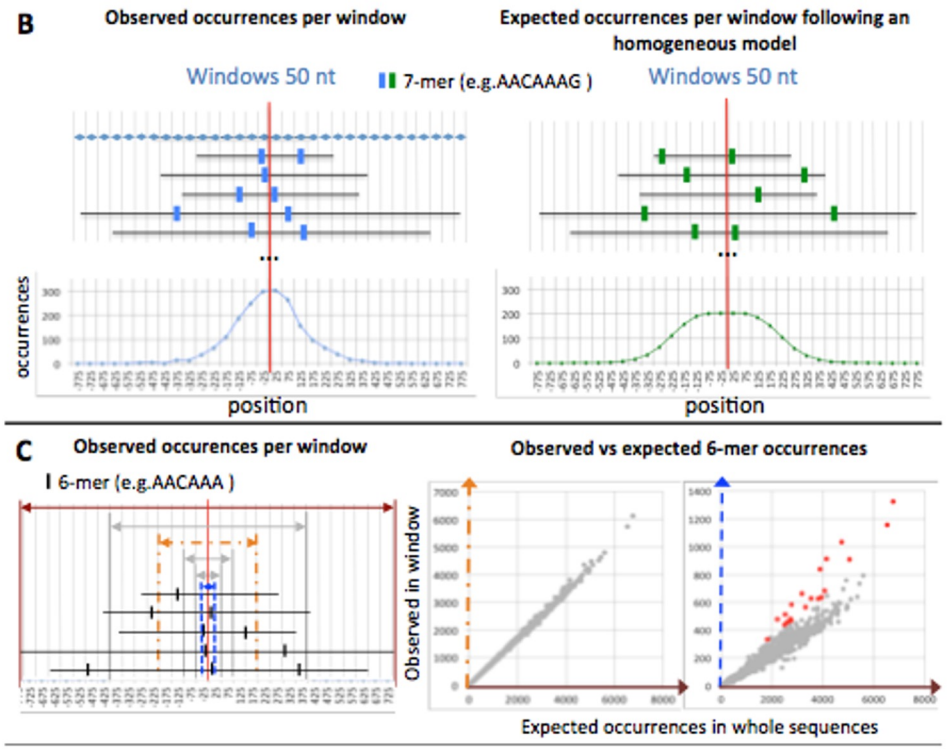


# Motif discovery: frequency



# Motif discovery: positional bias

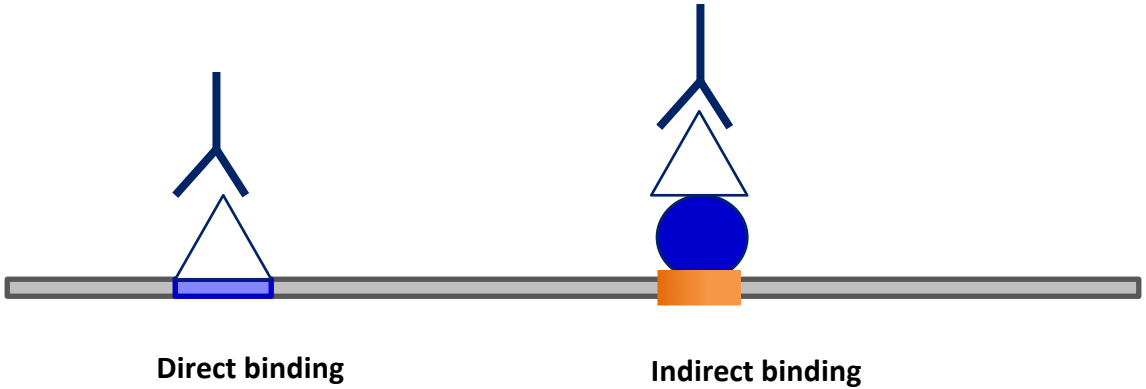



**position-analysis**

**local-words**

# Direct versus indirect binding

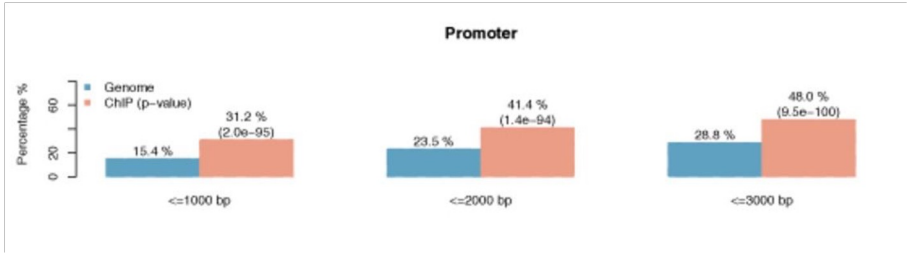
ChIP-seq does not necessarily reveal **direct binding:** The motif of the targeted TF is not always found in peaks!

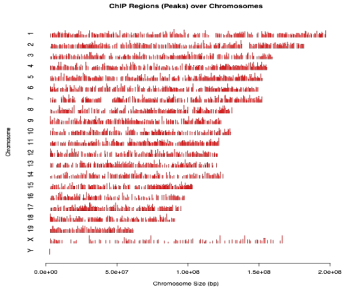


# Peak annotation

# 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?

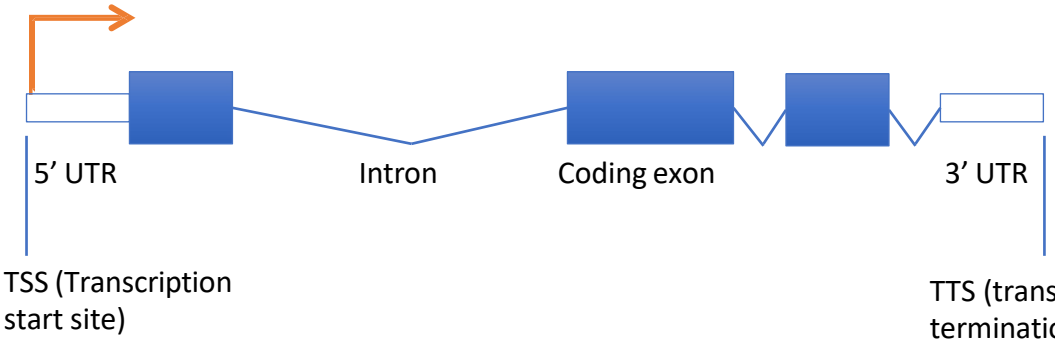
**HOMER**

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.

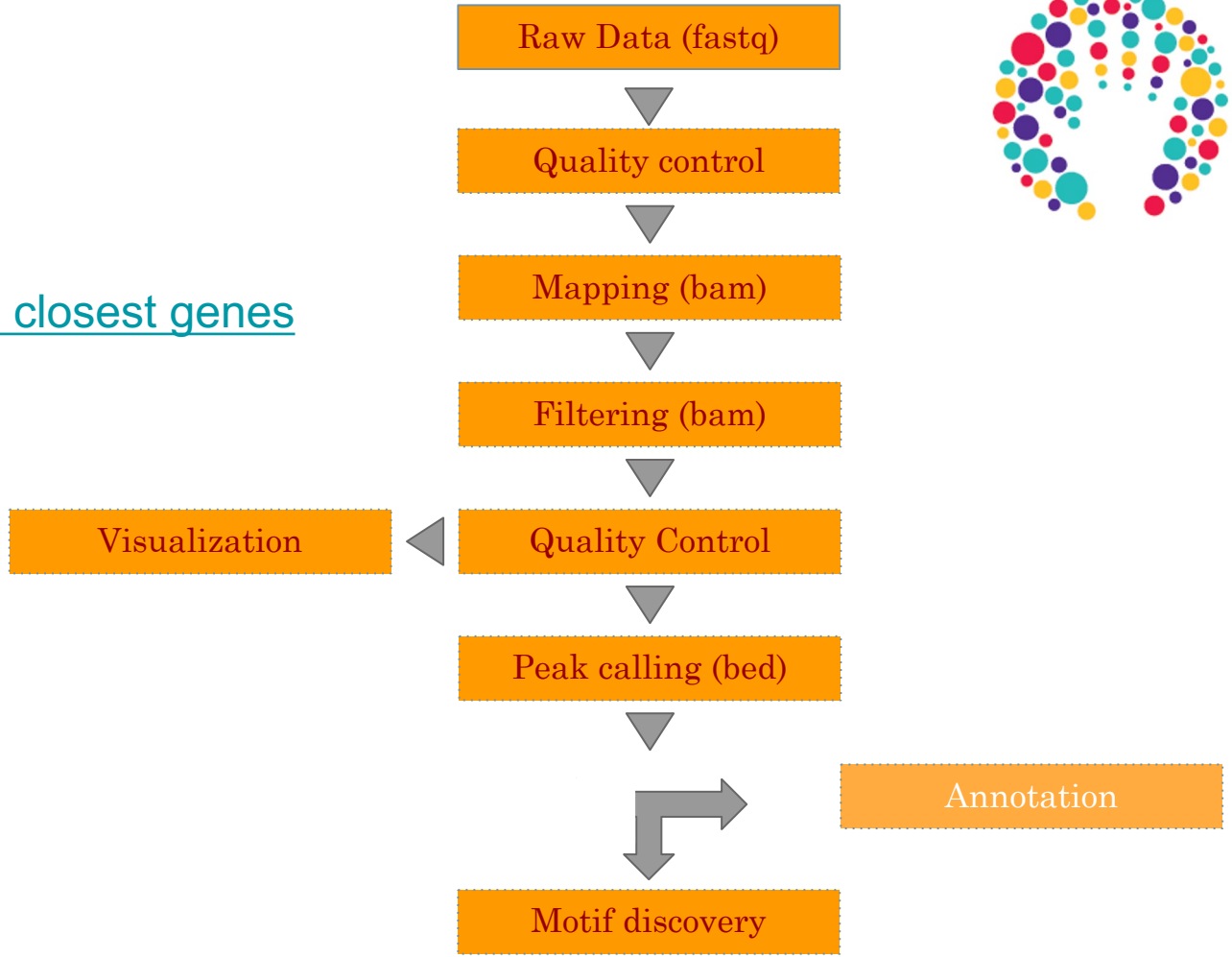
# Peak annotation

- Goal: assigning a peak to one or many genome features (genes/transcripts) to understand which genes are possibly regulated by the protein of interest binding
- The name of the gene is important as well as the genic region where the peak is located
- Example of Homer tools:
  - Determines the distance to the nearest Transcription Start Site (TSS) and assigns the peak to that gene
  - Determines the genomic annotation of the region **occupied by the center** of the peak/region. Possible genomic annotation:



# Protocol

- Associate peaks to closest genes

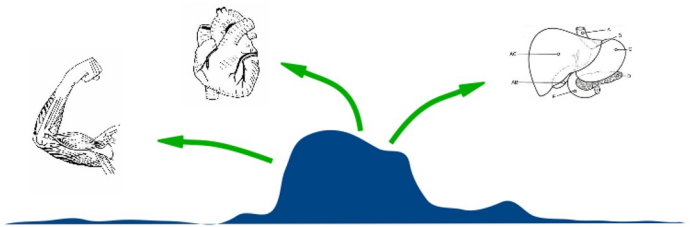


Now that we have the genes,

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

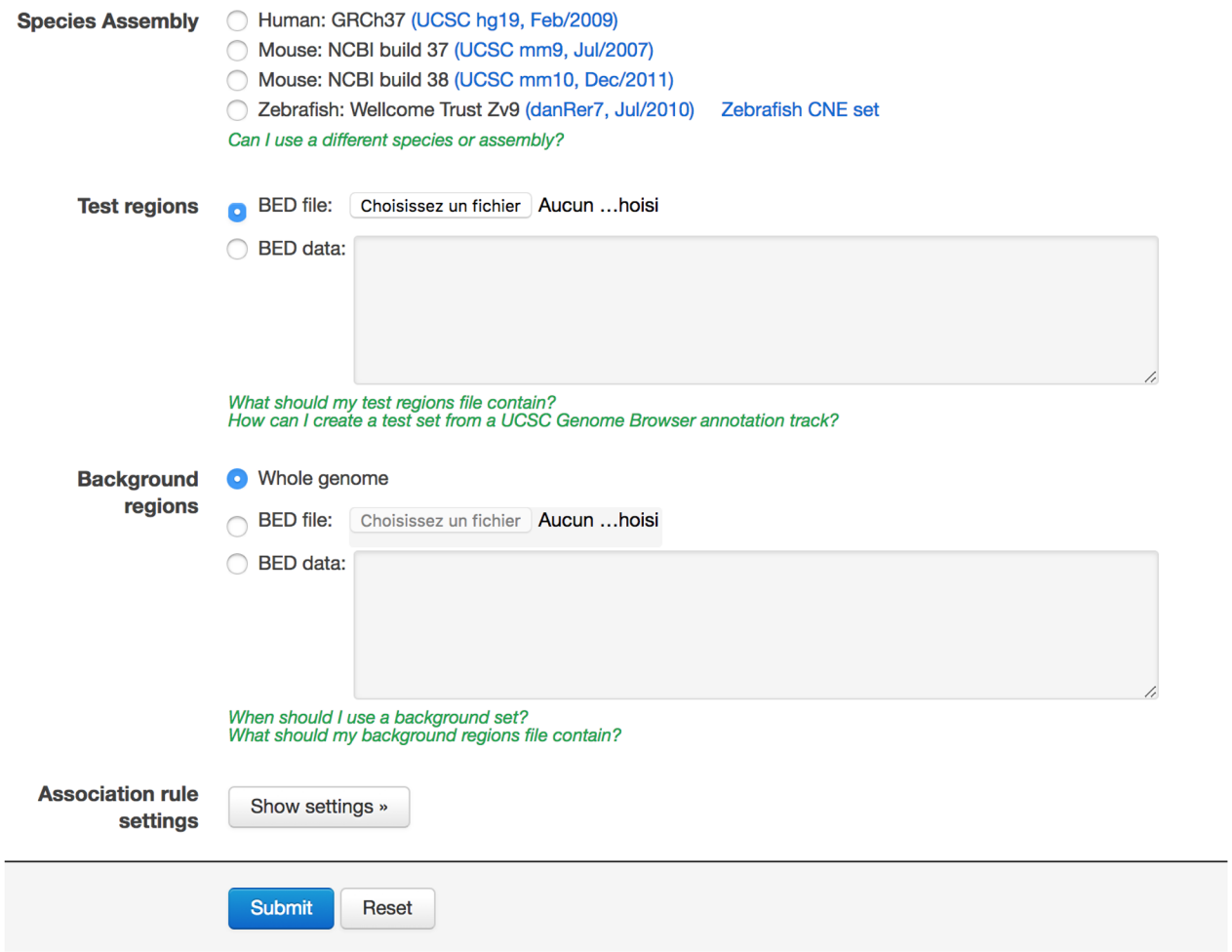
| **P-value** | LogP | Term | GO Tree | GO ID | # of Genes in Term | # of Target Genes in Term | # of Total Genes | # of Target Genes | Common Genes |
|---|---|---|---|---|---|---|---|---|---|
| 2.912e-26 | -5.880e+01 | immune response | biological process | GO:0006955 | 349 | 35 | 18091 | 168 | Il10,Cd14,Malt1,Ccl2,Ccl7,Ifih1 |
| 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 | CGGn\|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



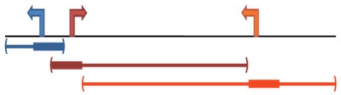

Note: Only human (hg19), mouse (mm9, mm10) and zebrafish (danRer7) genomes are supported

# 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.

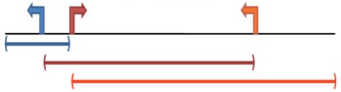
(•) Basal plus extension



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.

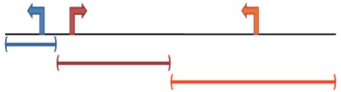
( ) Two nearest genes



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

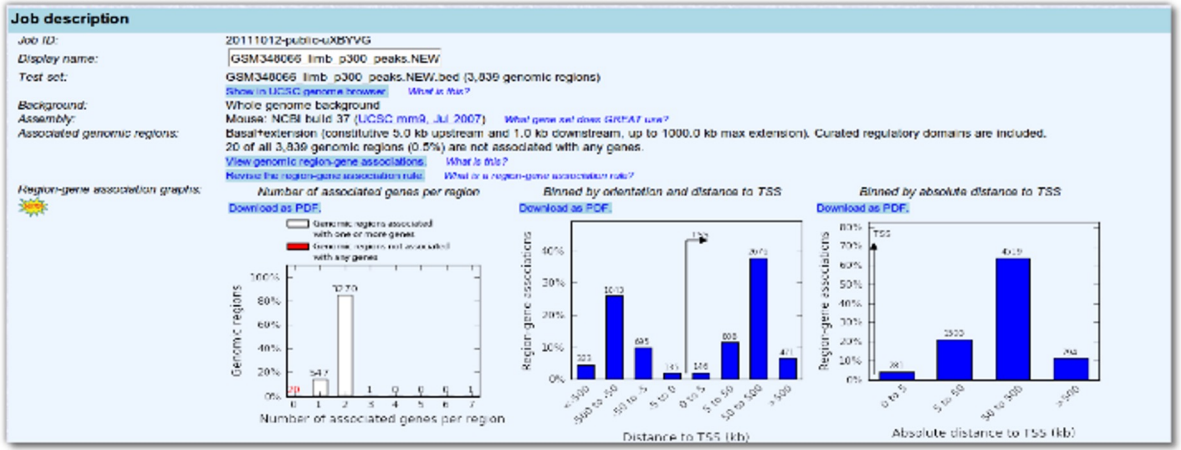
**Gene regulatory domain definition:** Each gene is assigned a regulatory domain that extends in both directions to the midpoint between the gene's TSS and the nearest gene's TSS but no more than the maximum extension in one direction.

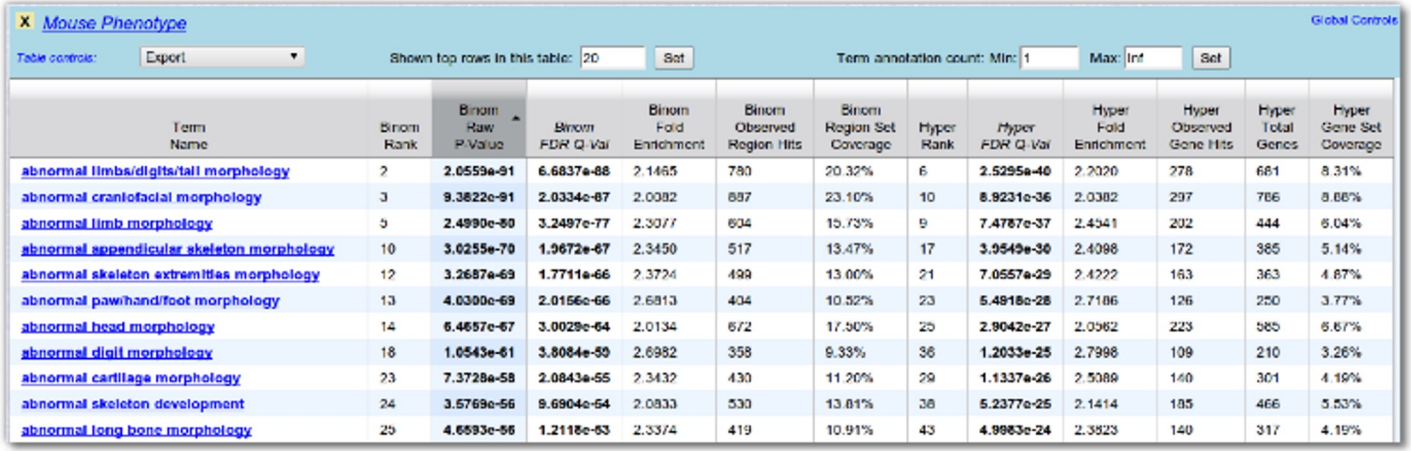
Gene Transcription Start Site (TSS)

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



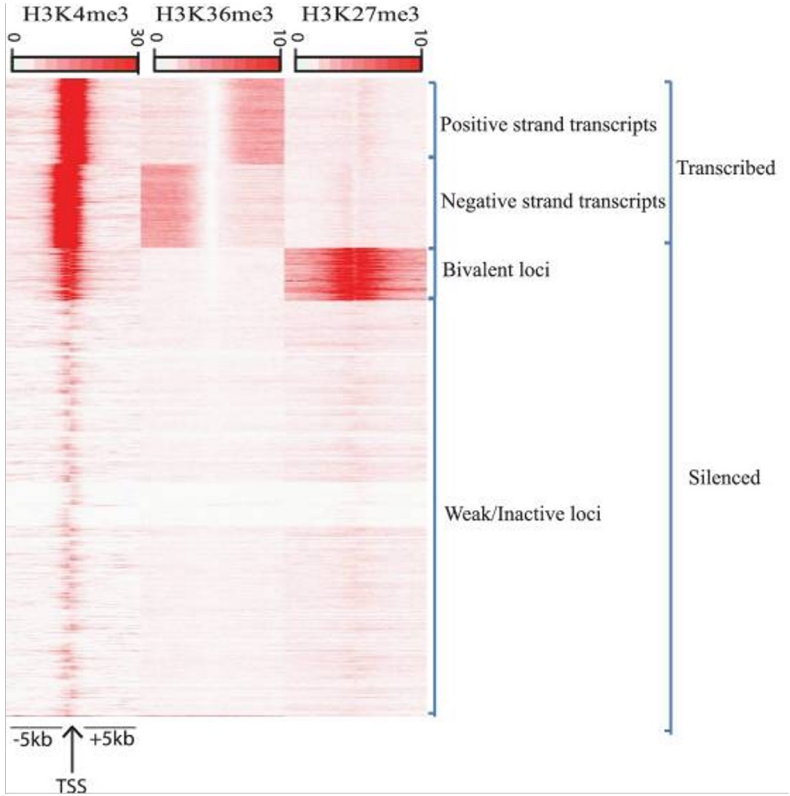




| Term Name | Binom Rank | Binom Raw P-Value | Binom FDR Q-Val | Binom Fold Enrichment | Binom Observed Region Hits | Binom Region Set Coverage | Hyper Rank | Hyper FDR Q-Val | Hyper Fold Enrichment | Hyper Observed Gene Hits | Hyper Total Genes | Hyper Gene Set Coverage |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 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 | 5.4818e-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 | 4.9983e-24 | 2.3823 | 140 | 317 | 4.19% |

# Other analyses

- Clustering peaks

(**Deeptools, HOMER, seqMINER**)



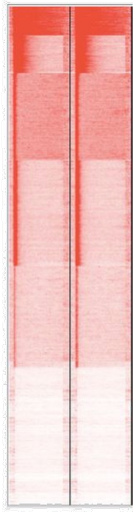

Ye et al, 2011

The darker the red the higher the read enrichment

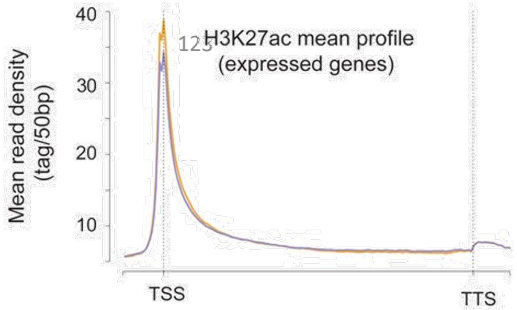
# Meta-profiles / Clustering`

- Global visualization of the data
- Need:
  - Regions of interest
    - Regions around a reference point e.g all TSS +/- 1Kb,…
    - Scaled regions e.g peaks, gene bodies,…
  - Signal data (mapped reads)
- Data can be grouped together with clustering methods such as k-means.

Heatmap

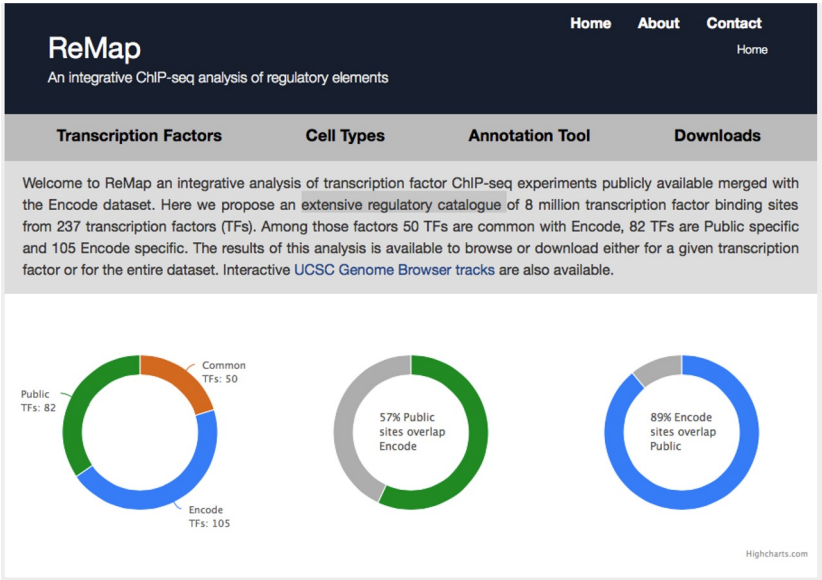


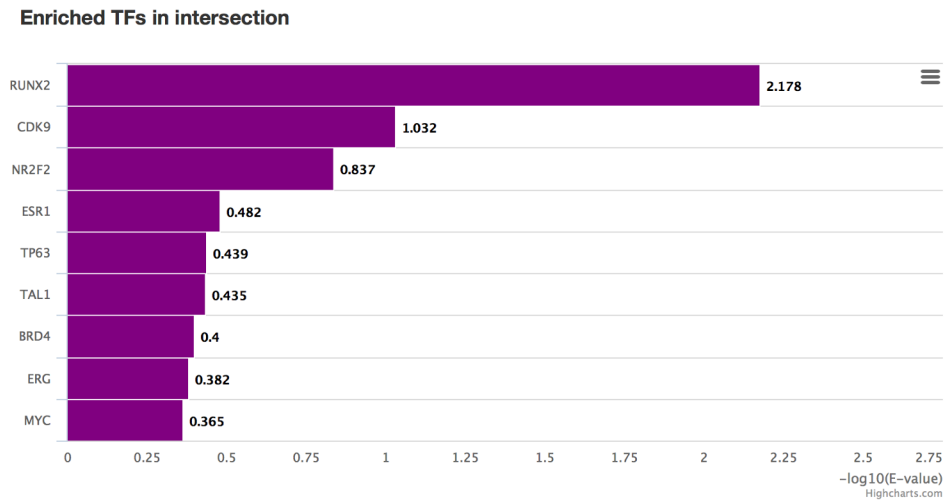
Mean profile



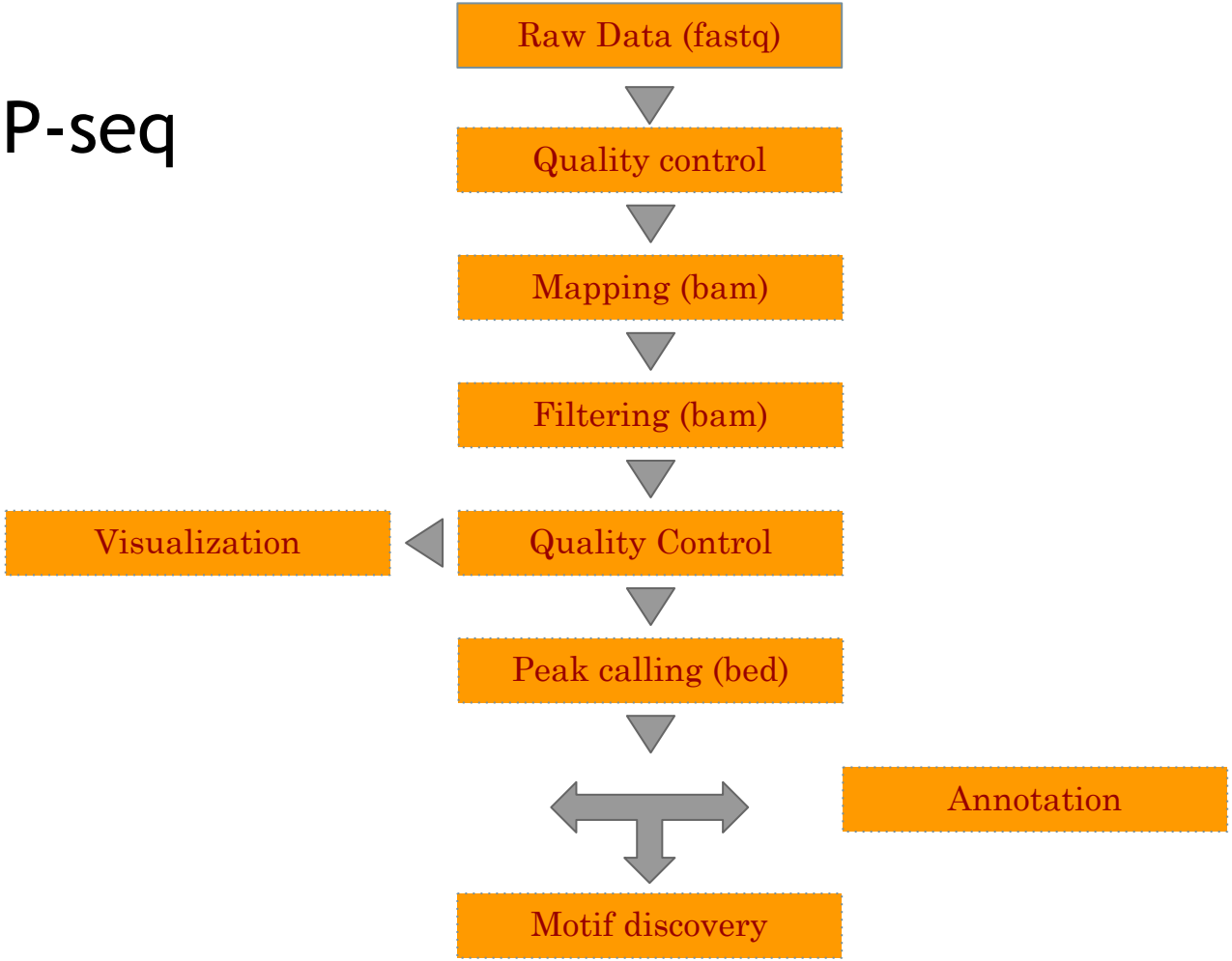
# Other analyses

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



# Overview of ChIP-seq Pipeline

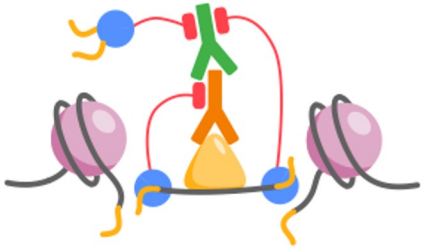


# Beyond ChIP-seq : Cut&TAG (2019)

**CUT&RUN**
Cleavage under targets and release using nuclease

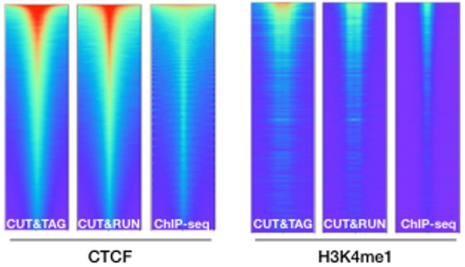
- Cleave adjacent DNA by MNase
- No crosslinking
- Low background

**CUT&TAG**
Cleavage under targets and tagmentation



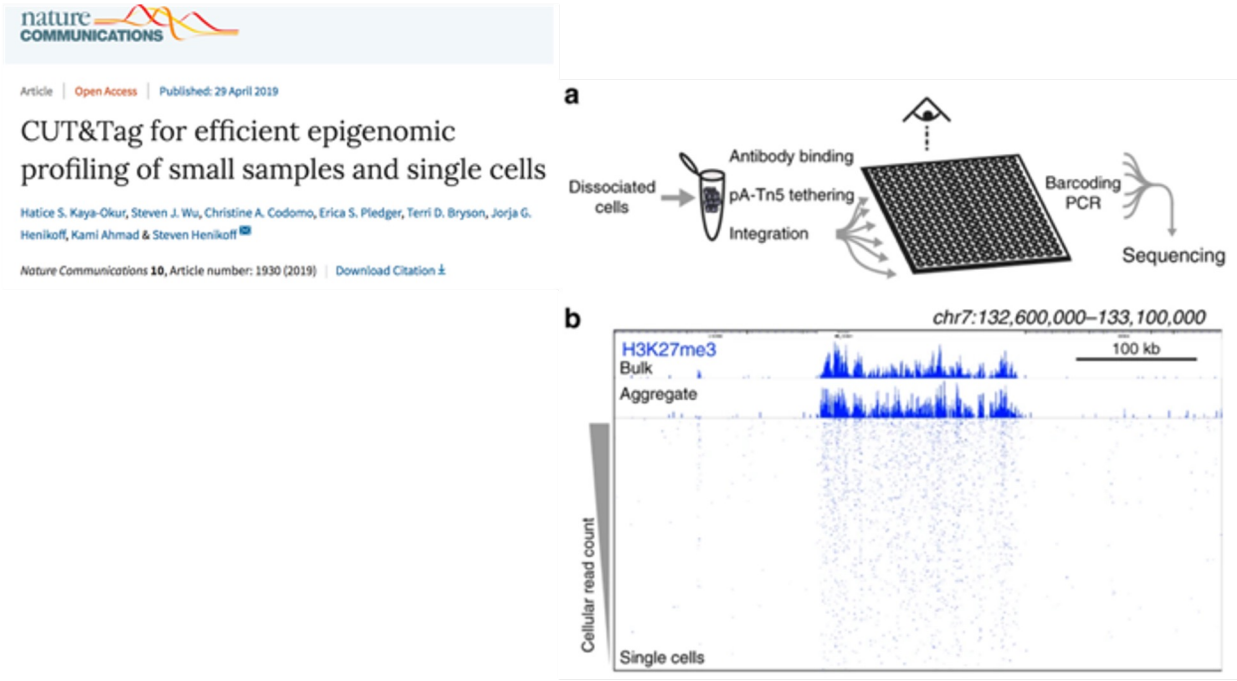
- Cleave near antibody site by Tn5
- No crosslinking
- Low background
- Include adaptor ligation
- Adapted for single-cell

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)




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

# ATAC-seq

***A**ssay for **T**ransposase-**A**ccessible **C**hromatin with highthroughput **seq**uencing*

# 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 ~2–3% 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

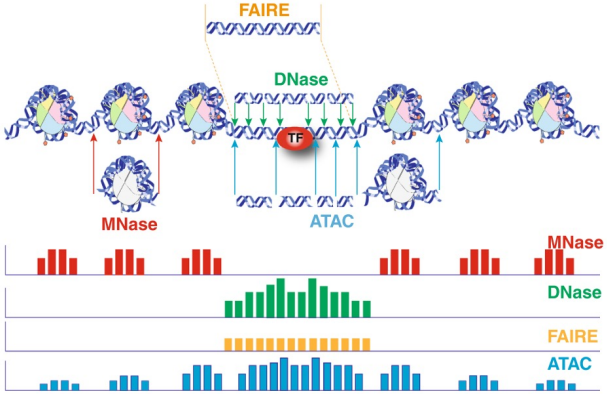
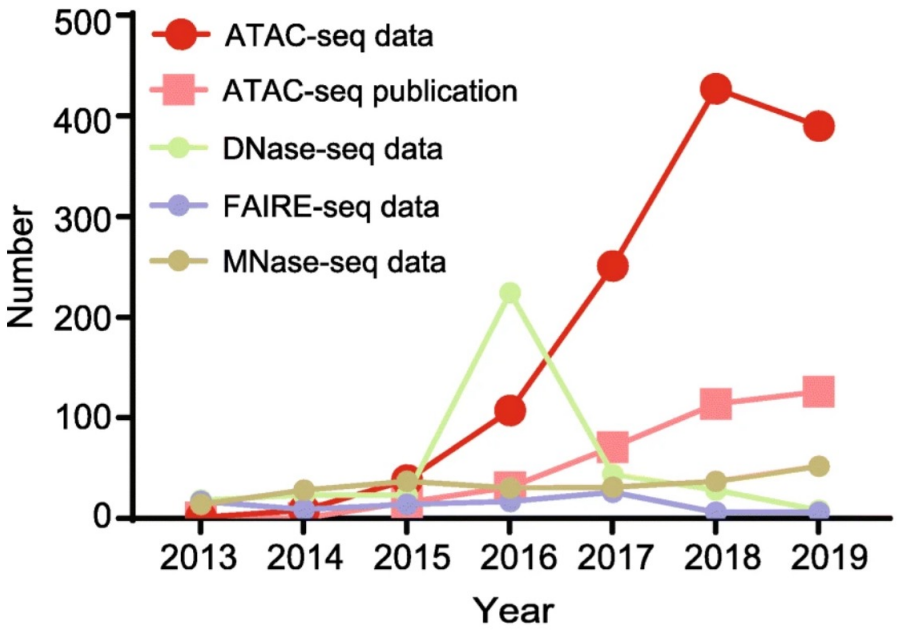



**Figure 1 Schematic diagram of current chromatin accessibility assays performed with typical experimental conditions.** Representative DNA fragments generated by each assay are shown, with end locations within chromatin defined by colored arrows. Bar diagrams represent data signal obtained from each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

(Tsompana and Buck, 2014)

# Chromatin accessibility assays

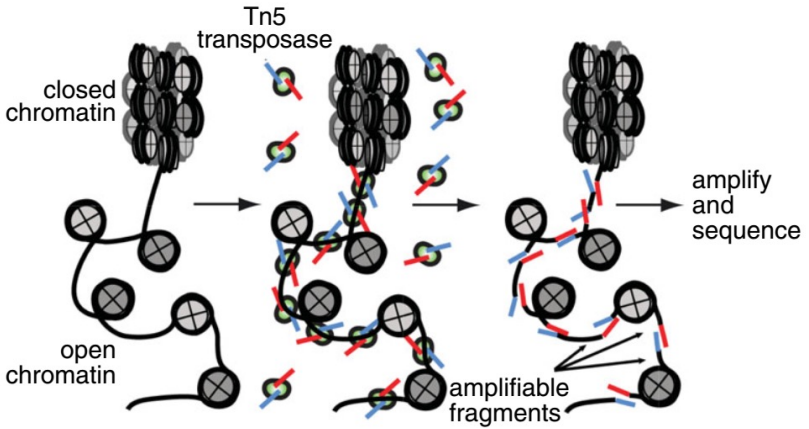
- ATAC-seq has become the most widely used method to detect and analyze open chromatin regions




Yan et al, 2020

# ATAC-seq

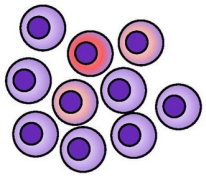
- Buenrostro et al, 2013
- ATAC-seq is a method for determining chromatin accessibility across the genome
- Transcription factor binding sites and positions of nucleosomes can be identified from the analysis of ATAC-Seq
- Advantages of ATAC-seq over other chromatin accessibility assays:
  - The sensitivity 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 cells)
  - single-cell ATAC-seq (scATAC-seq) protocol (Cusanovich et al, 2015)

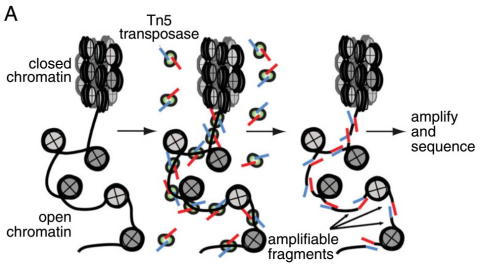


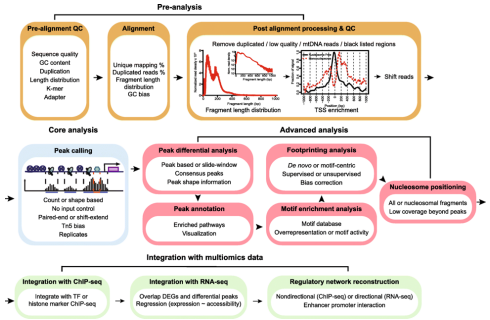

(Buenrostro et al., 2015).

# ATAC-seq process

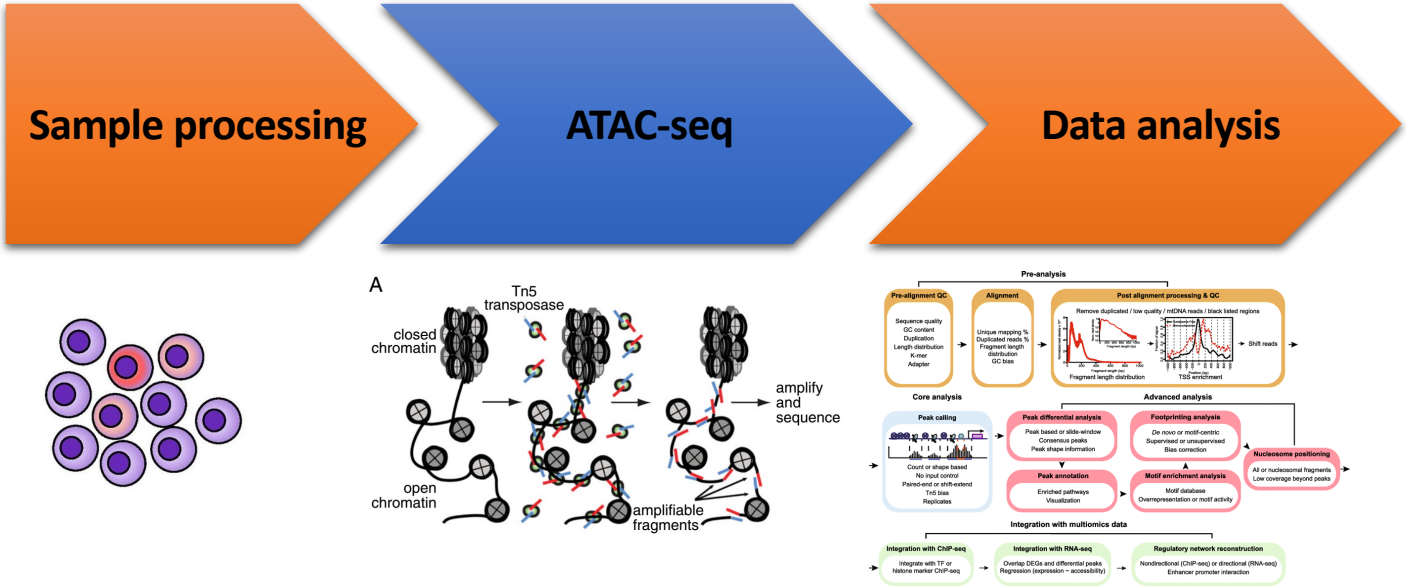
**Sample processing** → **ATAC-seq** → **Data analysis**





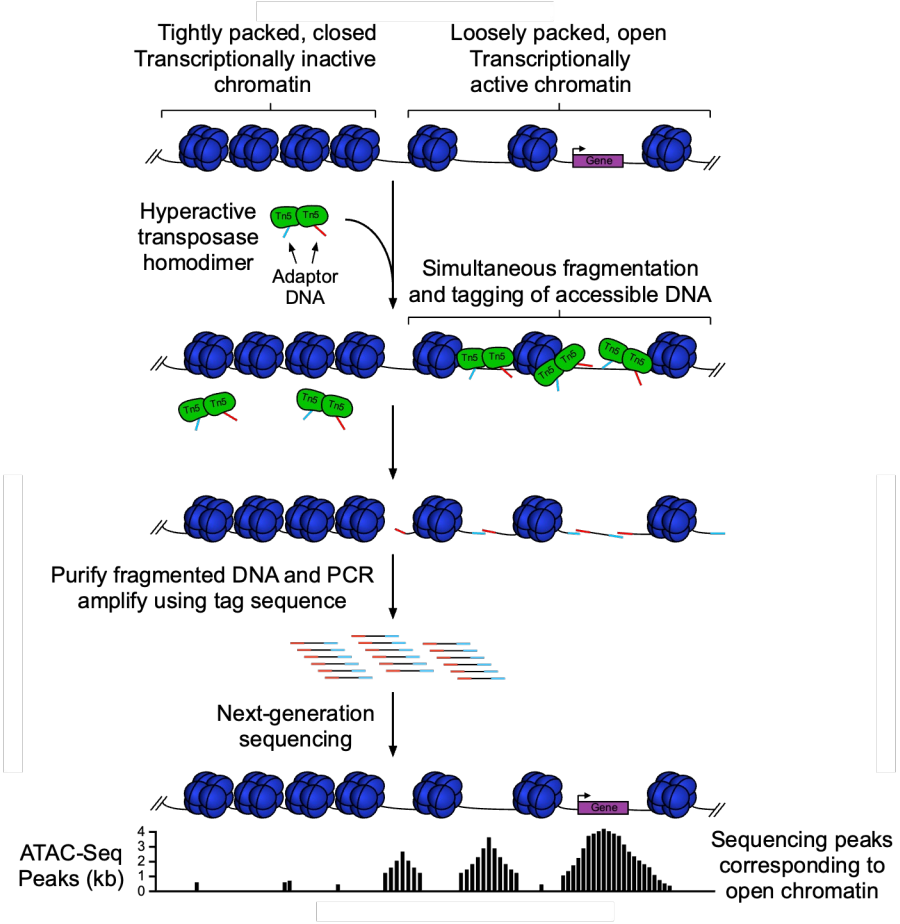



# ATAC-seq


Sample processing
ATAC-seq
Data analysis
A
Tn5
transposase
closed
chromatin
open
chromatin
amplify
and
sequence
amplifiable
fragments
Pre-analysis
Pre-alignment QC
Alignment
Post alignment processing & QC
Core analysis
Advanced analysis
Peak calling
Peak differential analysis
Footprinting analysis
Peak annotation
Motif enrichment analysis
Nucleosome positioning
Integration with multiomics data
Integration with ChIP-seq
Integration with RNA-seq
Regulatory network reconstruction

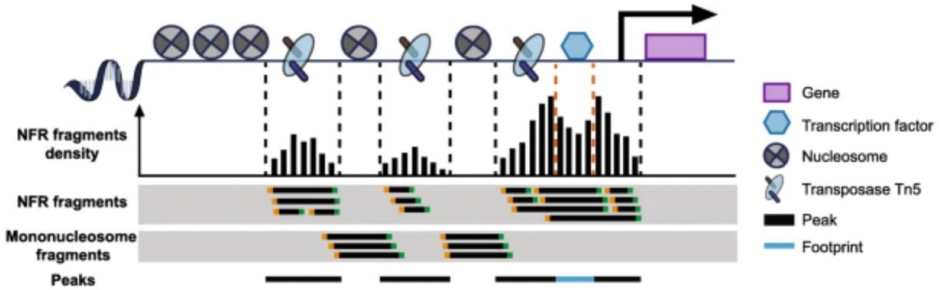
# ATAC-seq

- 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.

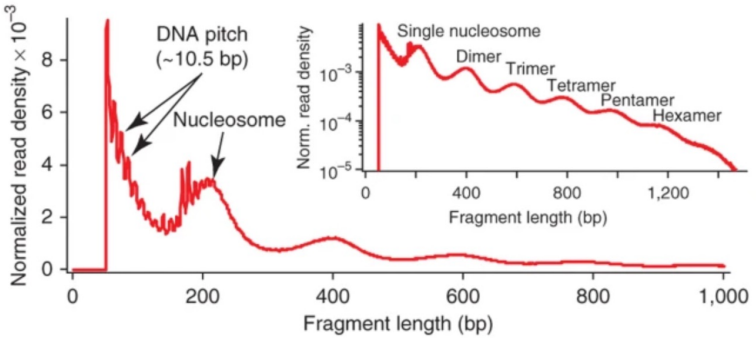


# ATAC-seq

- **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.

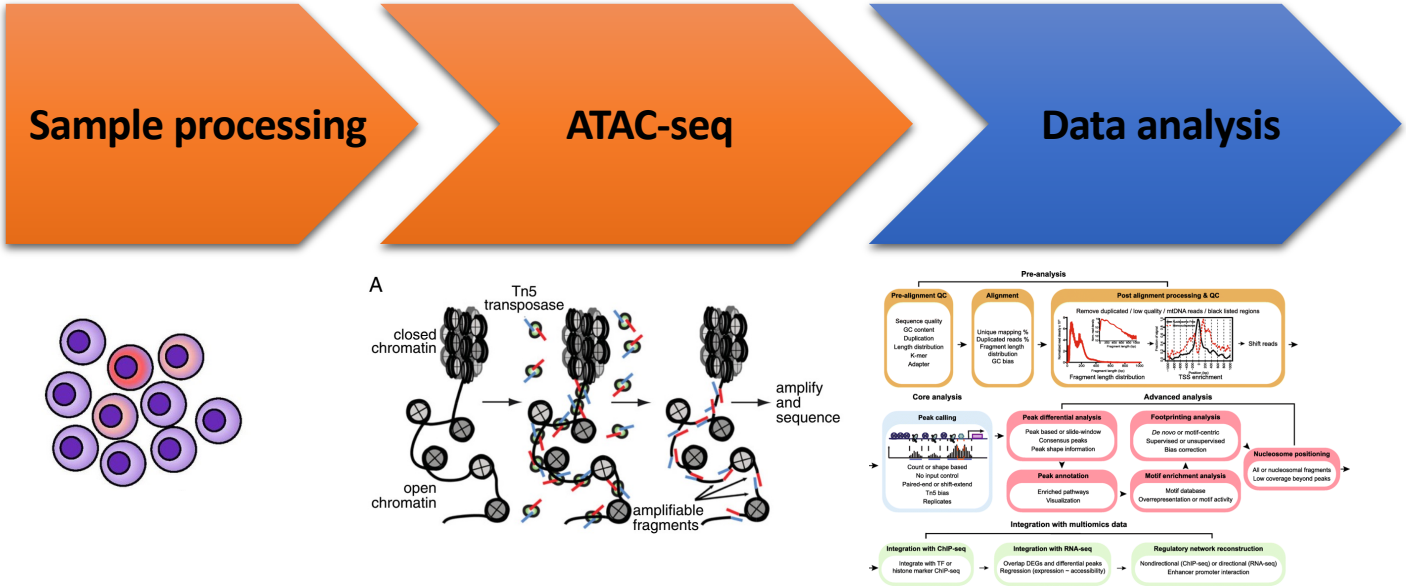



Yan et tal, 2020

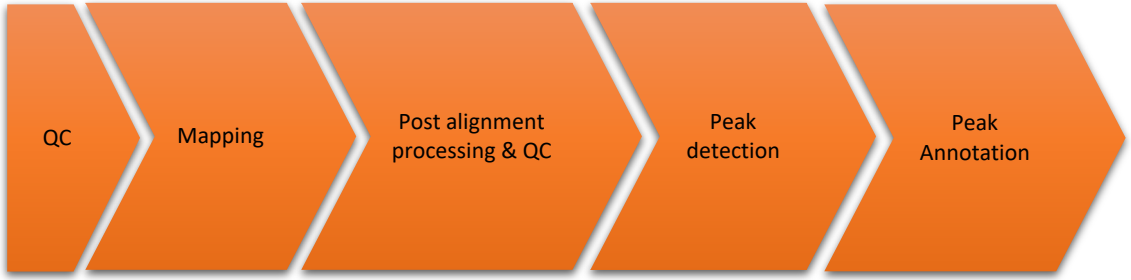



Buenrostro et al, 2023

# Analysis of ATAC-seq data



# Analysis of ATAC-seq data

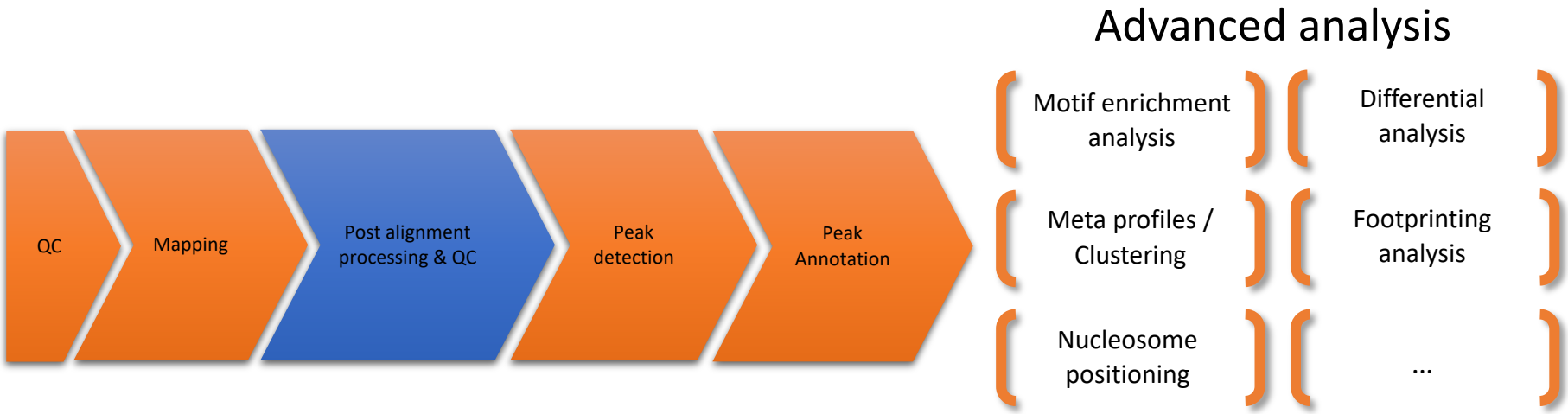



Advanced analysis

- Motif enrichment analysis
- Differential analysis
- Meta profiles / Clustering
- Footprinting analysis
- Nucleosome positioning
- ...

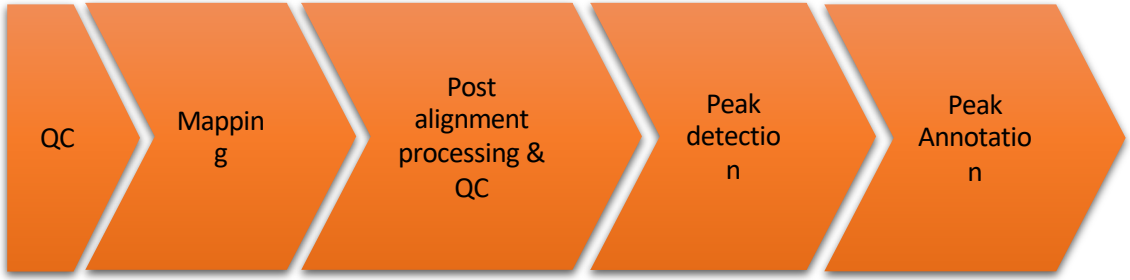
- Overall analysis resemble ChIP-seq data analysis
- Description of particularities of ATAC-seq data analysis

# 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)

# Analysis of ATAC-seq data

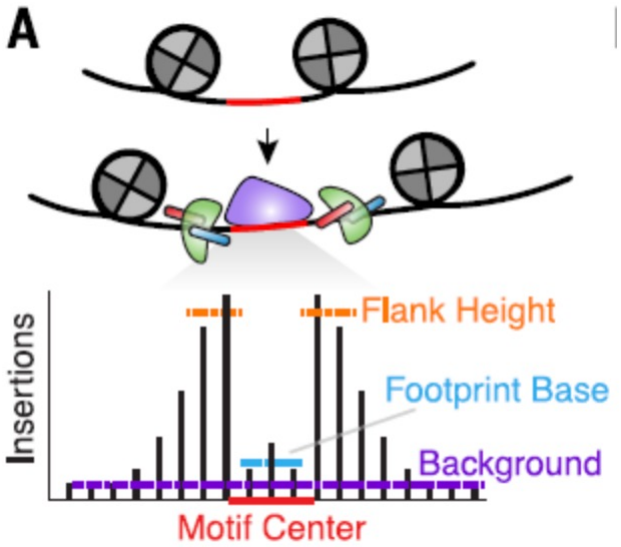



## Advanced analysis

- 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




Corces et al, 2019

# 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)