**Moffitt Cancer Center**
Biostatistics & Bioinformatics

**Morsani College of Medicine**
Oncologic Sciences

# Regulatory Genomic Sequencing

USF Master's Program in
Bioinformatics and Computational Biology

Mingxiang Teng

March 21, 2023

TE

what is regulatory genomic sequencing

**ChatGPT**

Regulatory genomic sequencing is the study of the non-coding regions of the genome that control gene expression. These regions include DNA sequences that act as switches to turn genes on or off, as well as other functional elements such as enhancers, silencers, and insulators.

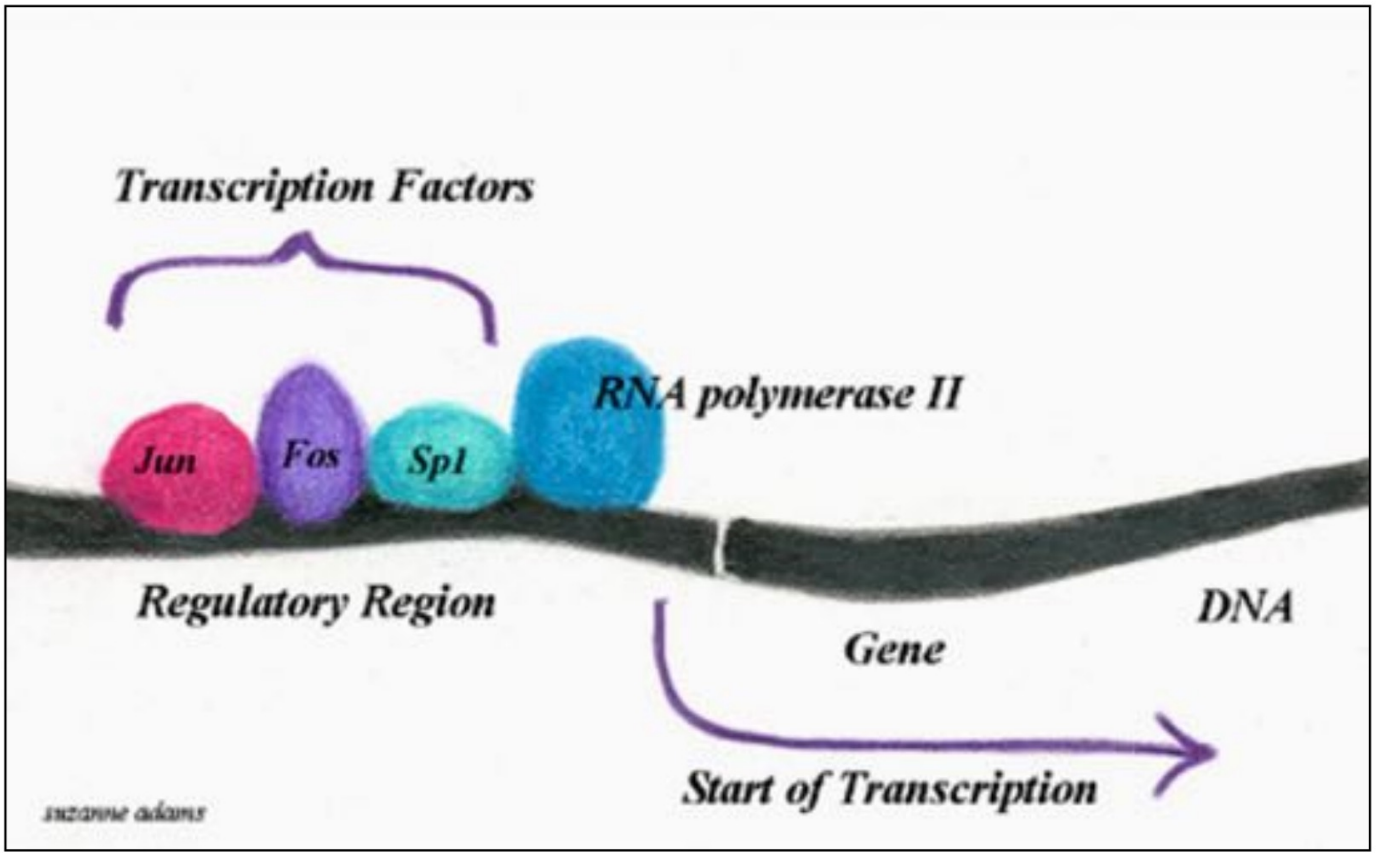
Regulatory genomic sequencing can involve a range of techniques, including chromatin immunoprecipitation (ChIP) to identify DNA-binding proteins that interact with regulatory elements, sequencing of RNA transcripts to identify genes that are actively expressed, and epigenetic profiling to investigate modifications to DNA and chromatin structure that affect gene expression.

The information obtained from regulatory genomic sequencing can provide insight into how genes are regulated during development, in response to environmental stimuli, and in disease states. This can help to identify potential therapeutic targets and inform the development of new treatments for genetic disorders and other diseases.

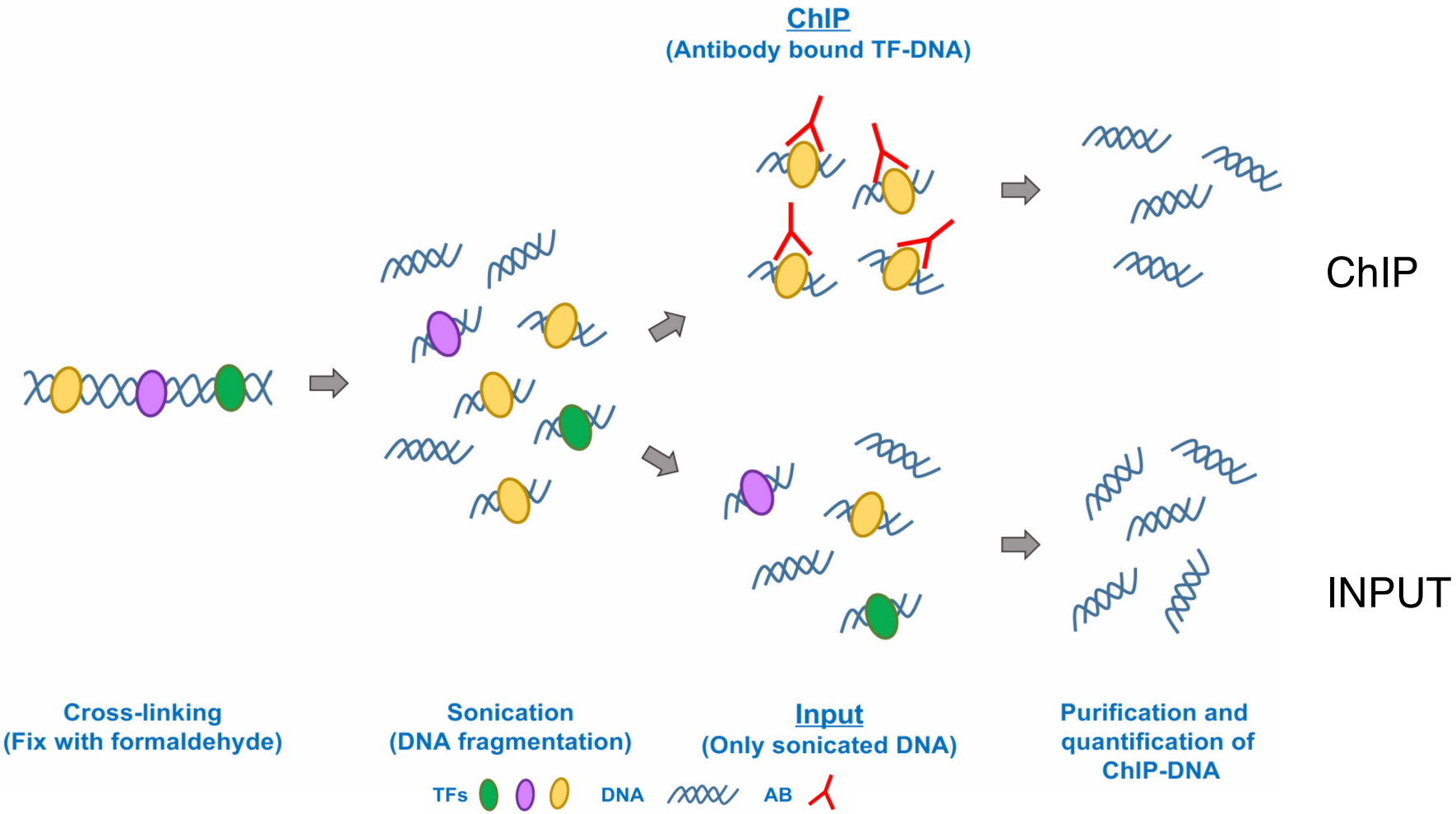
# OUTLINE

- ChIPseq data types
- Public data resources
- **Exploratory data analysis** (QC)
- **Routine analysis**
- Super enhancers
- **Differential binding analysis**
- **Peak annotation**
- Integrative analysis

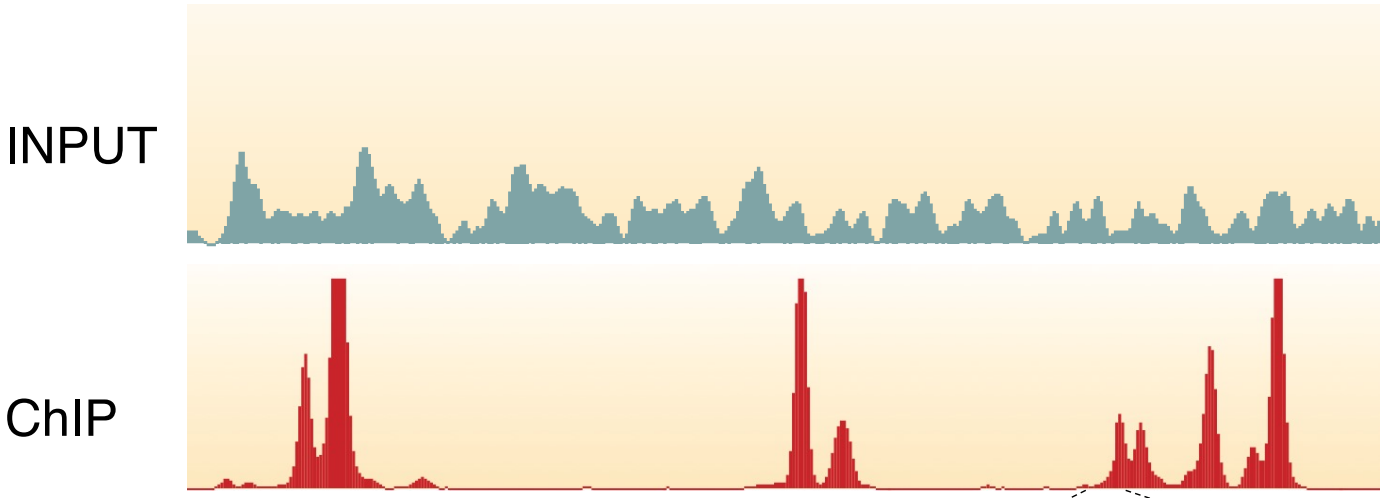
# Gene regulation

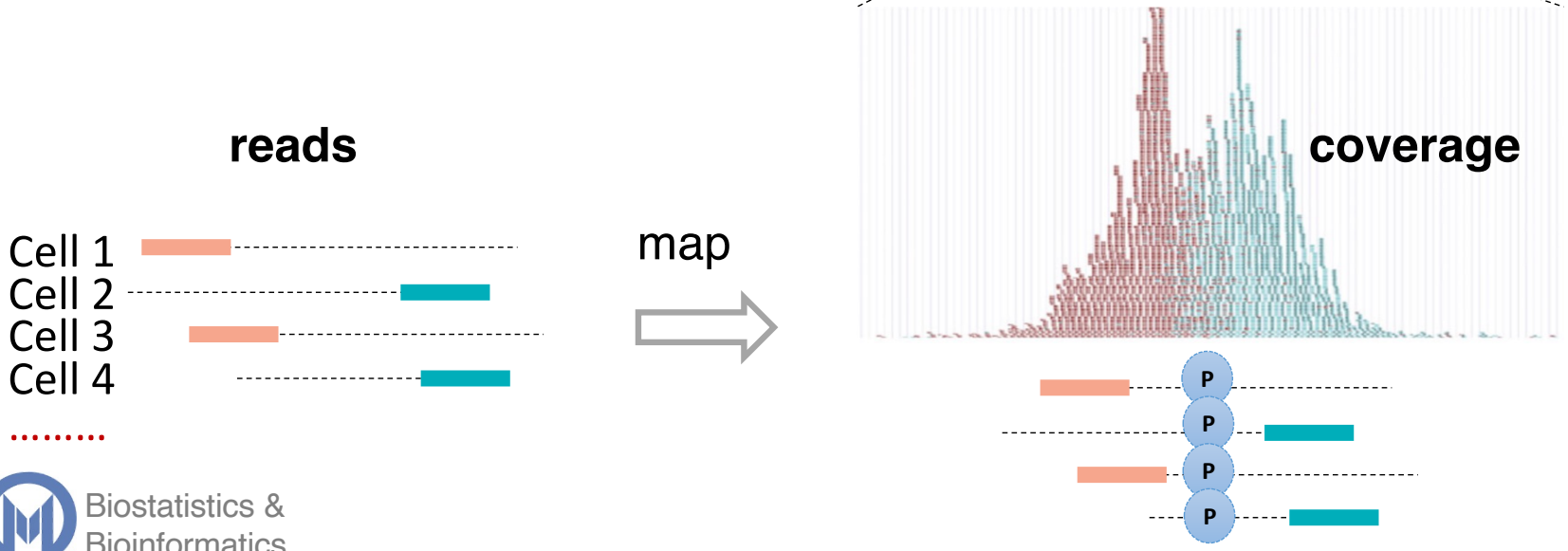


# Experiment protocol



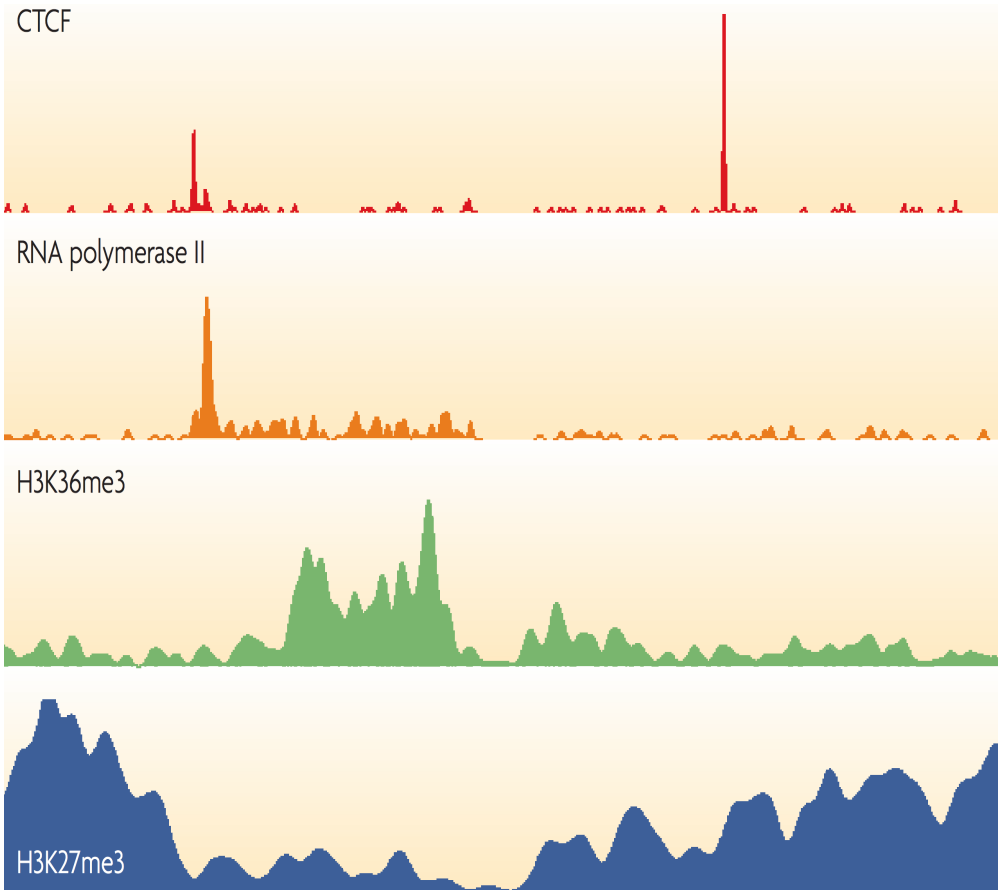
# Genome-wide signal



# Typical data types

- Transcription Factors
  - Point source (CTCF)
  - Mixed signal (Pol2)
- Chromatin Marks
  - Point source (H3K4Me3, H3K27Ac)
  - Broad region (H3K36Me3)
  - Mixed signal (H3K27Me3)



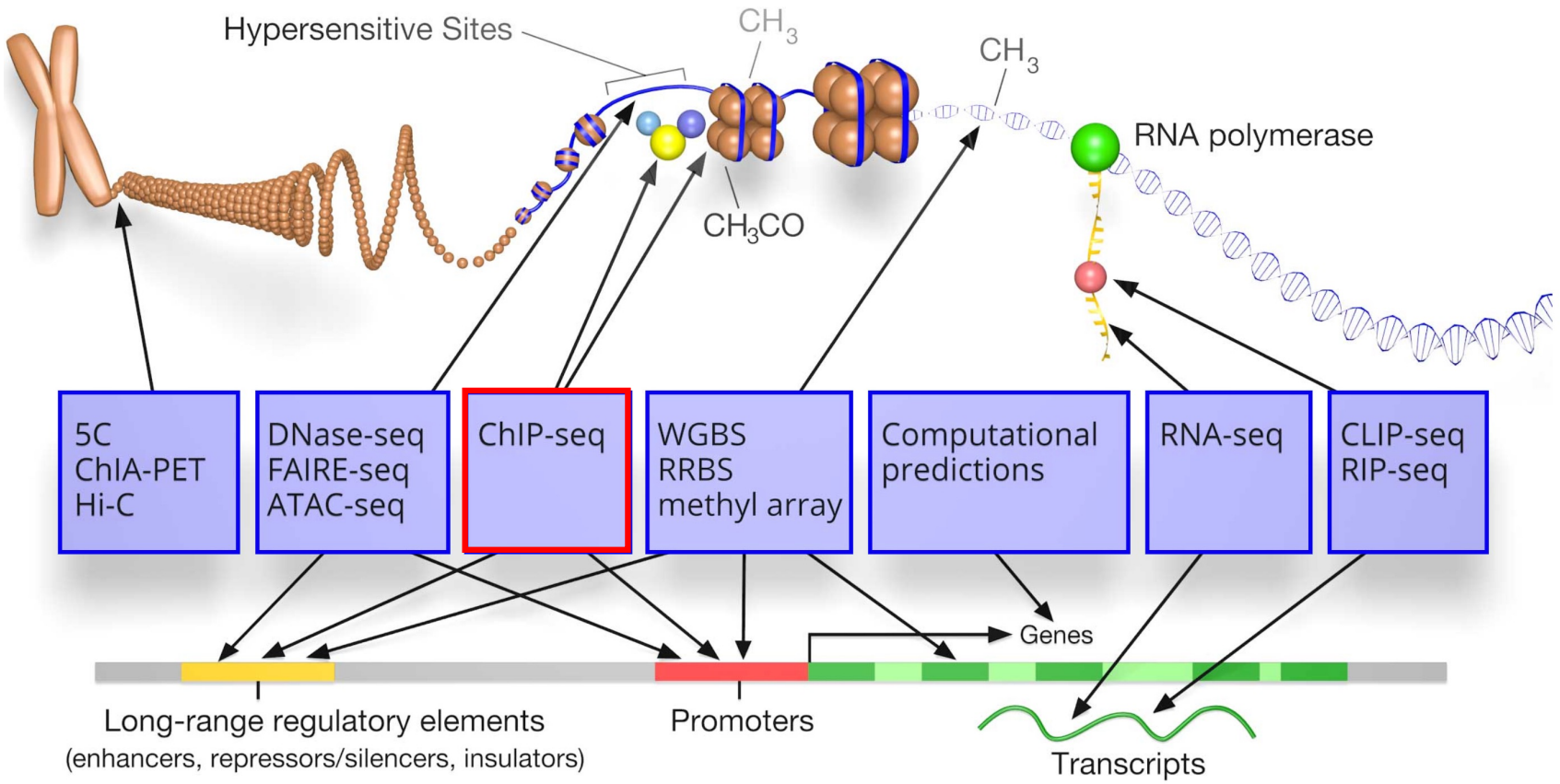
# Typical data types (cont.)

**Histone Marks**

| Broad Marks | Narrow Marks | Exceptions |
|---|---|---|
| • H3F3A<br>• H3K27me3<br>• H3K36me3<br>• H3K4me1<br>• H3K79me2<br>• H3K79me3<br>• H3K9me1<br>• H3K9me2<br>• H4K20me1 | • H2AFZ<br>• H3ac<br>• H3K27ac<br>• H3K4me2<br>• H3K4me3<br>• H3K9ac | • H3K9me3 |
| 45M usable reads | 20M usable reads | 45M usable reads |

**Transcription Factor**: 20M usable reads

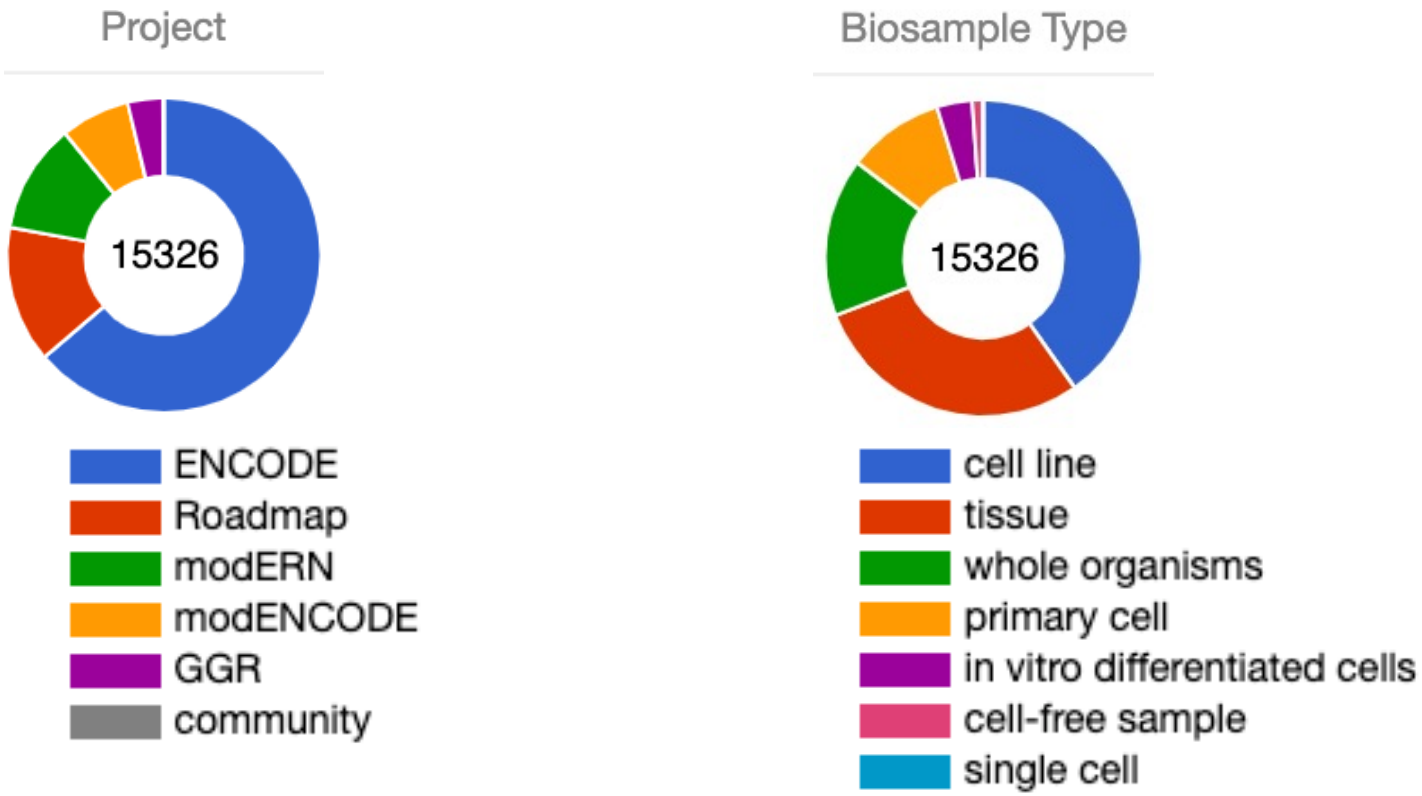
# Public data: Encyclopedia of DNA Elements



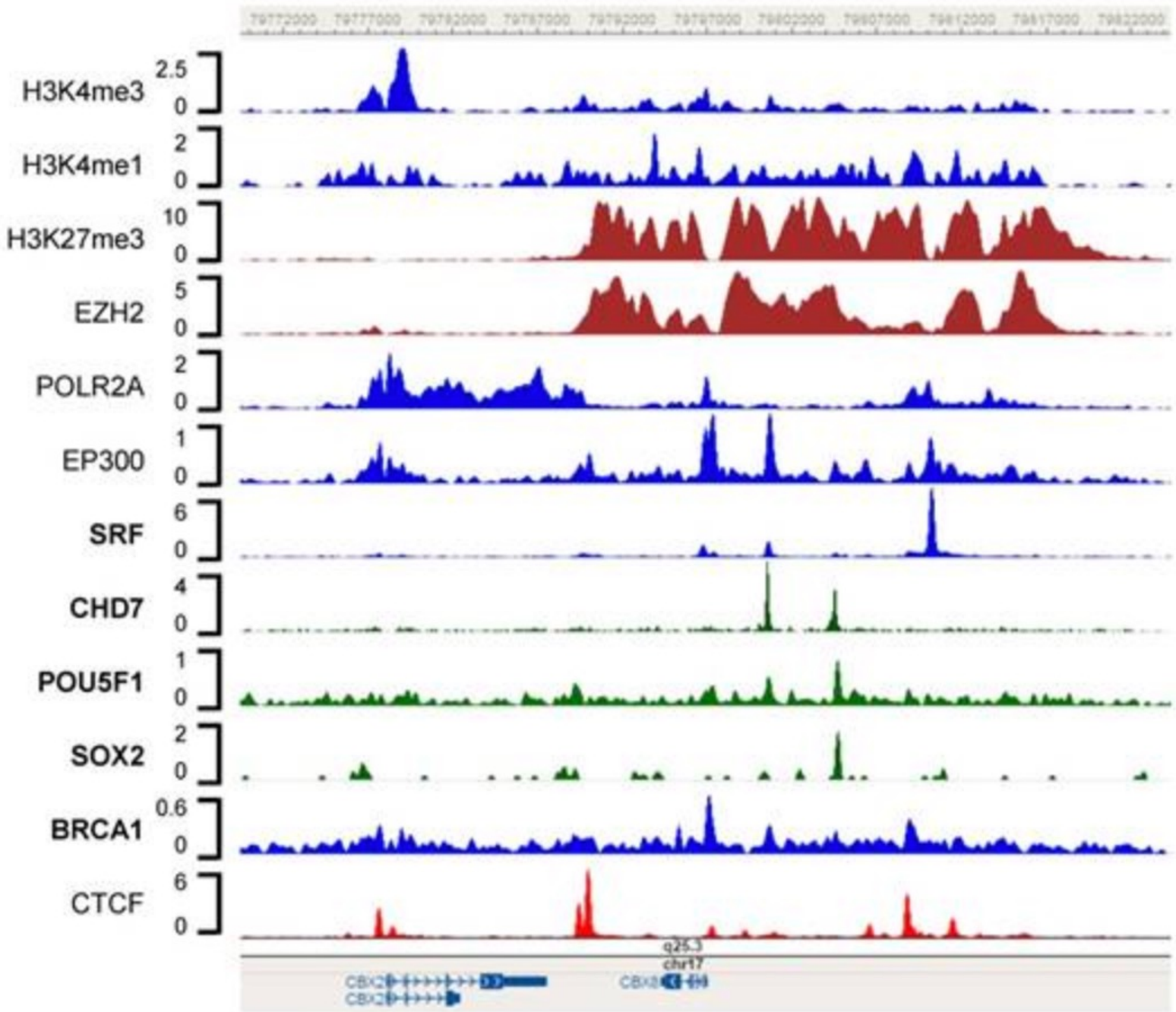

Biostatistics & Bioinformatics

Alternatives: *GEO, dbGaP*

# Public data: Encyclopedia of DNA Elements




Alternatives: *GEO, dbGaP*




Biostatistics & Bioinformatics

hESC-H1 cells

# Exploratory data analysis

# Exploratory data analysis – Quality Control

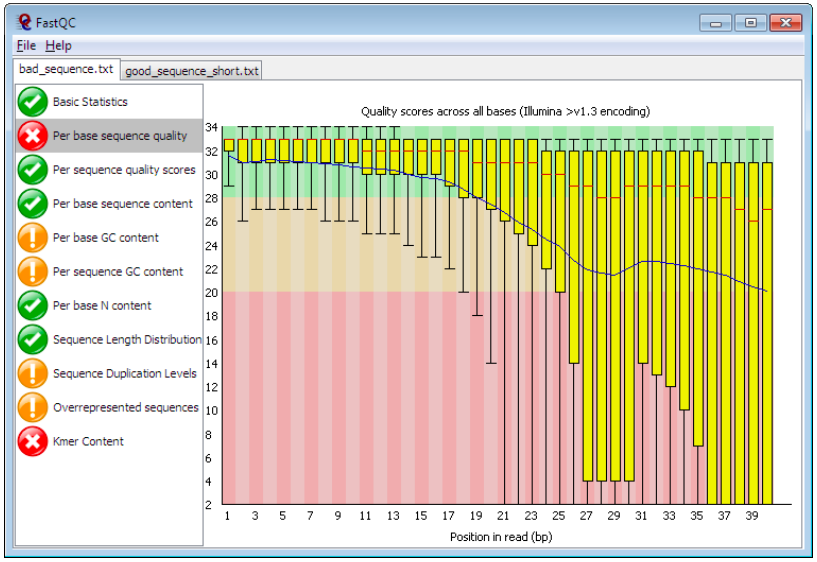
- Sequencing quality score of a given base:

  $Q = -10\log_{10}(e)$

  where e is the probability of base call being wrong

- **Higher Q score** indicate a smaller probability of error

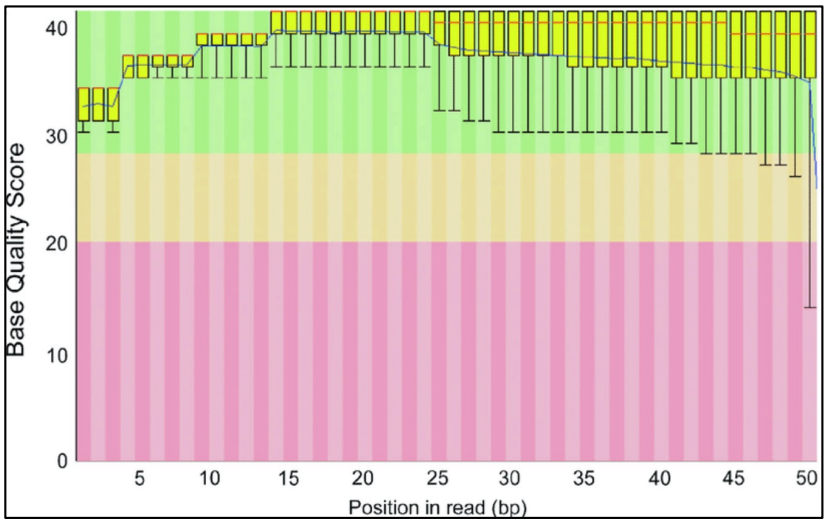
| Quality Score | Probability of Incorrect Base Call | Inferred Base Call Accuracy |
|---|---|---|
| 10 (Q10) | 1 in 10 | 90% |
| 20 (Q20) | 1 in 100 | 99% |
| 30 (Q30) | 1 in 1000 | 99.9% |



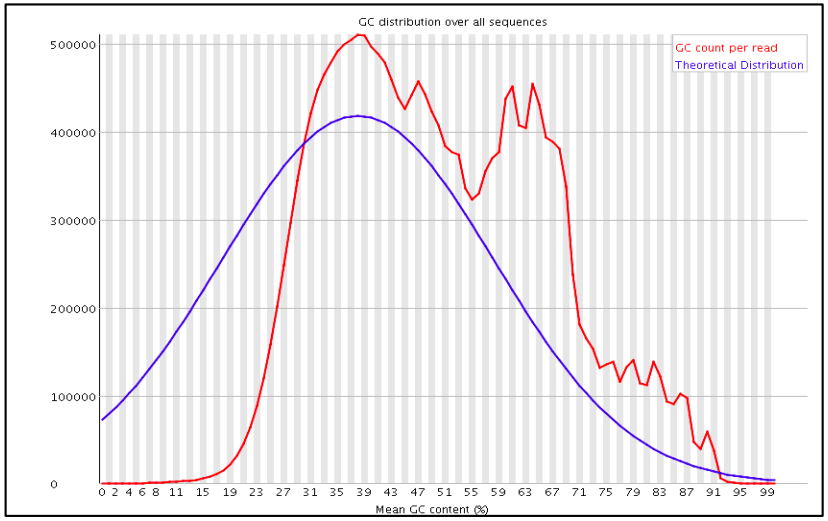

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

# Exploratory data analysis - QC

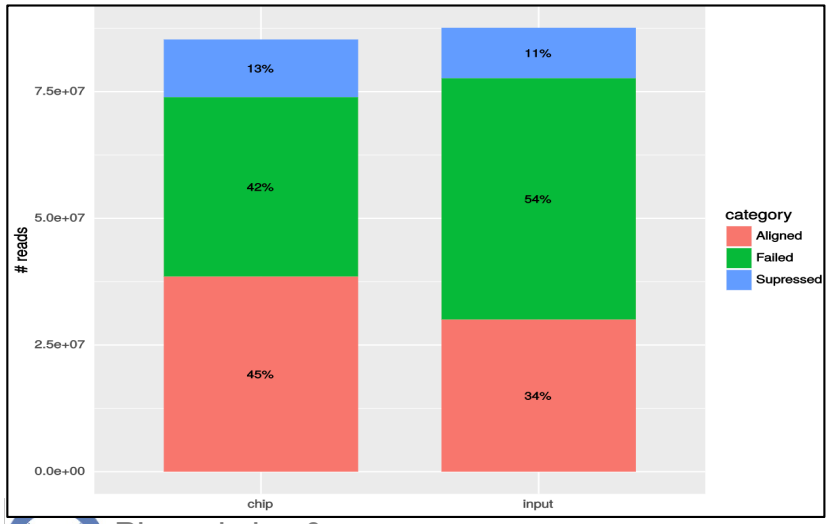
## Base quality




## GC content




## Mapping quality




## Mapping quality – more

Non-Redundant Fraction:

- NRF > 0.9

PCR Bottlenecking Coefficient 1

- PBC1 > 0.9

PCR Bottlenecking Coefficient 2

- PBC2 > 10

Fraction of Reads in Peaks

- FRiP > 1%

# Exploratory data analysis – specific metrics

- Non-Redundant Fraction (NRF) = Number of distinct uniquely mapping reads / Total number of reads.
- PCR Bottlenecking Coefficient 1 (PBC1) = $M_1/M_{DISTINCT}$
- PCR Bottlenecking Coefficient 2 (PBC2) = $M_1/M_2$

Non-Redundant Fraction:
- NRF > 0.9

PCR Bottlenecking Coefficient 1
- PBC1 > 0.9

PCR Bottlenecking Coefficient 2
- PBC2 > 10

Fraction of Reads in Peaks
- FRiP > 1%

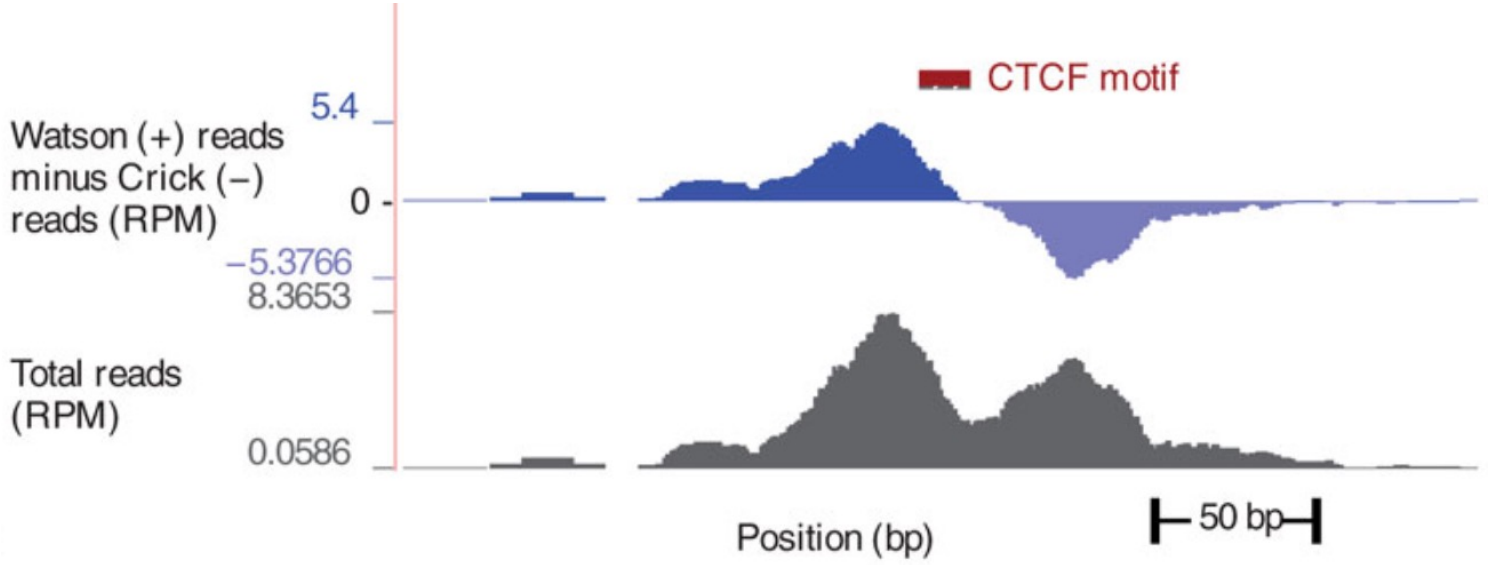
- Fraction of reads in peaks (FRiP) = Fraction of all mapped reads that fall into the called peak regions
- Cross correlation = correlation read counts between forward and reverse strands

$M_1$: # of genomic locations where exactly one read maps uniquely

$M_2$: # of genomic locations where two reads map uniquely

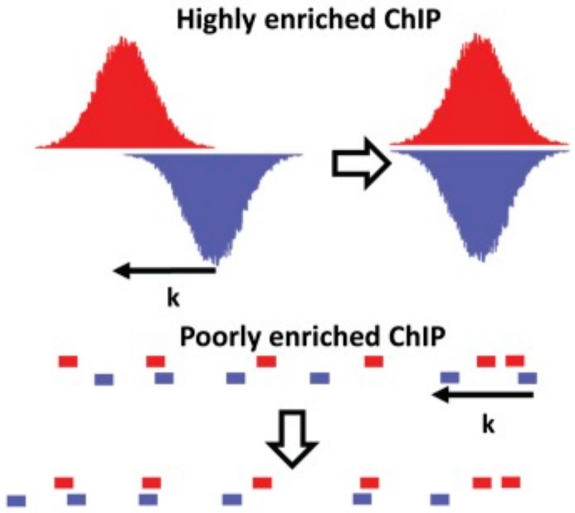
$M_{DISTINCT}$: # of distinct genomic locations to which some read maps uniquely

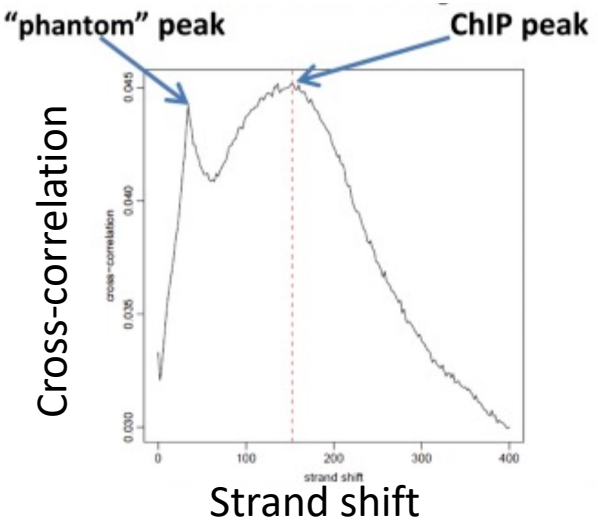
# Exploratory data analysis – cross correlation

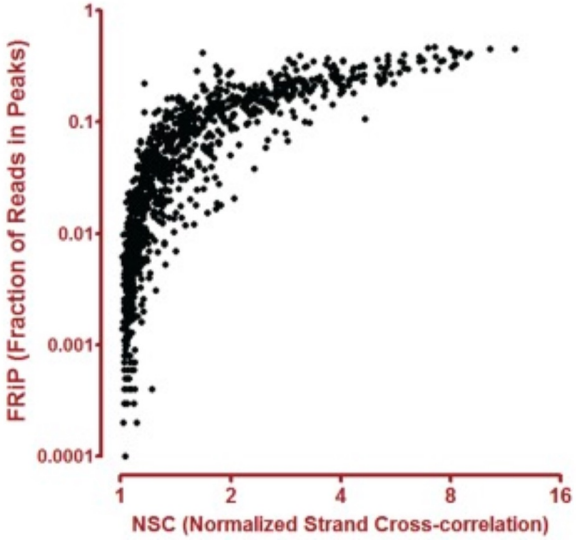


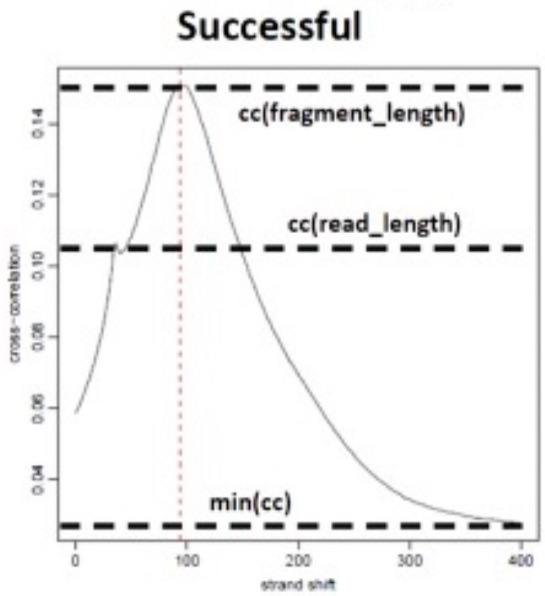

- Random fragmentation with sonication
- Single-end reads randomly sequenced on fragments
- Coverage symmetry between two strands

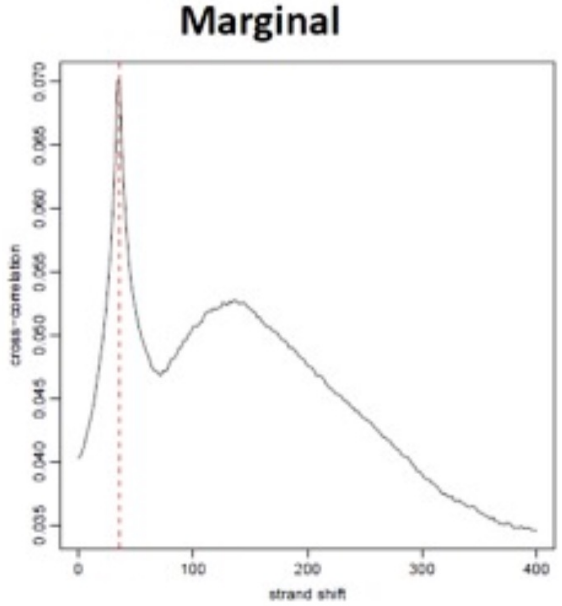
# Exploratory data analysis – cross correlation

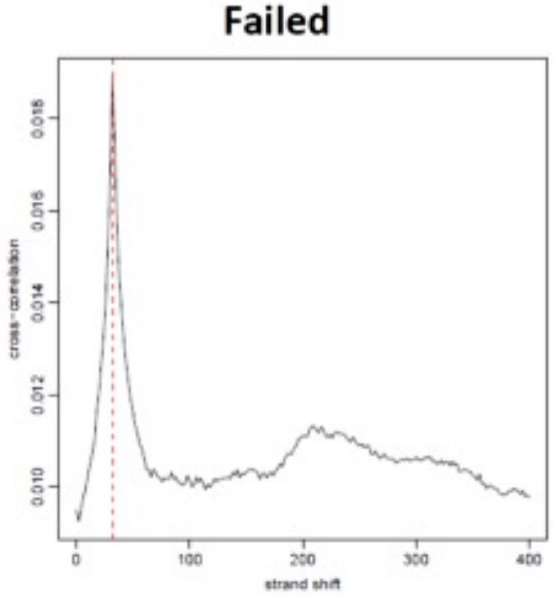


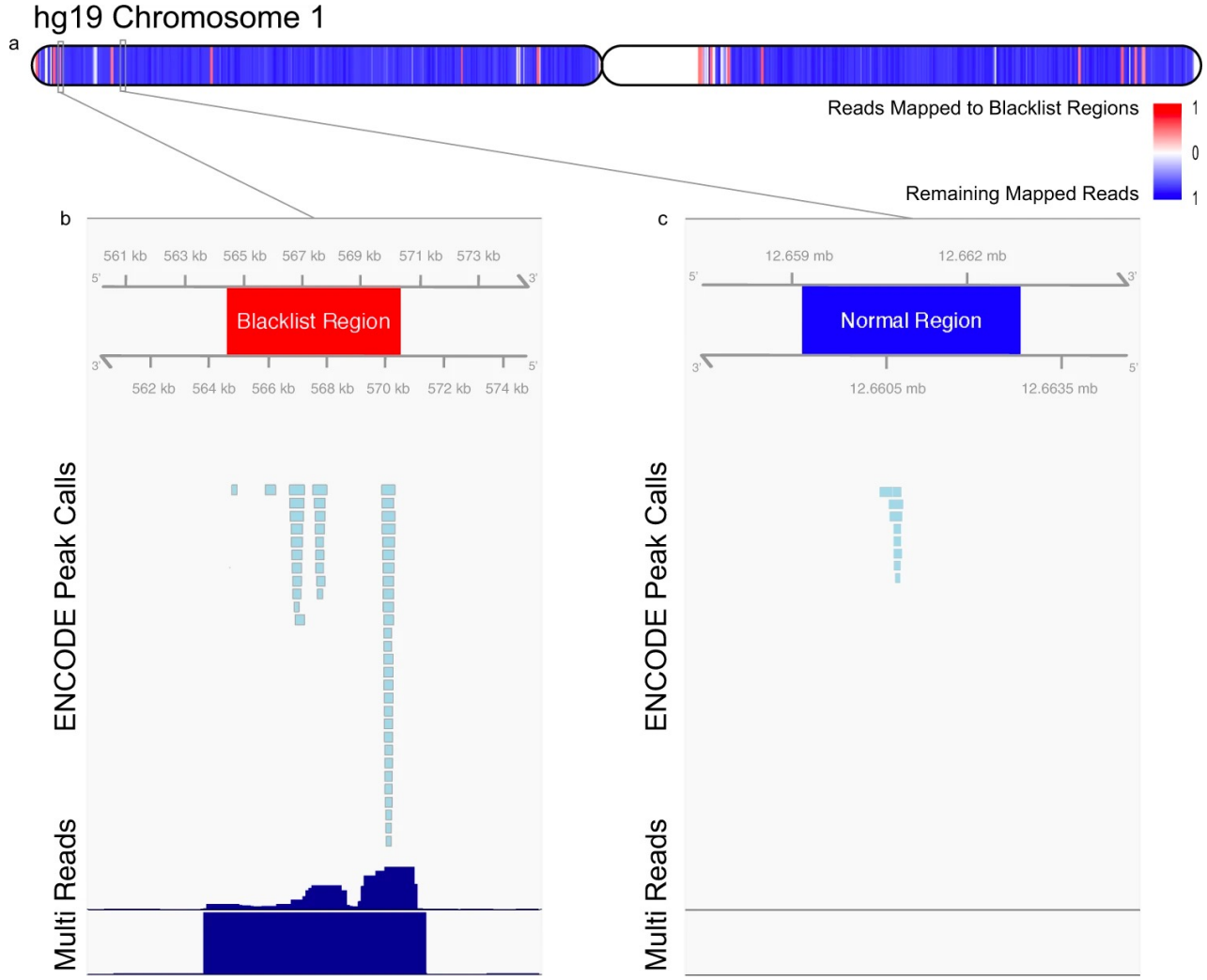



# Exploratory data analysis – blacklist regions




- Happening when low mappability or other
- Changing from genome build to genome build

Biostatistics & Bioinformatics

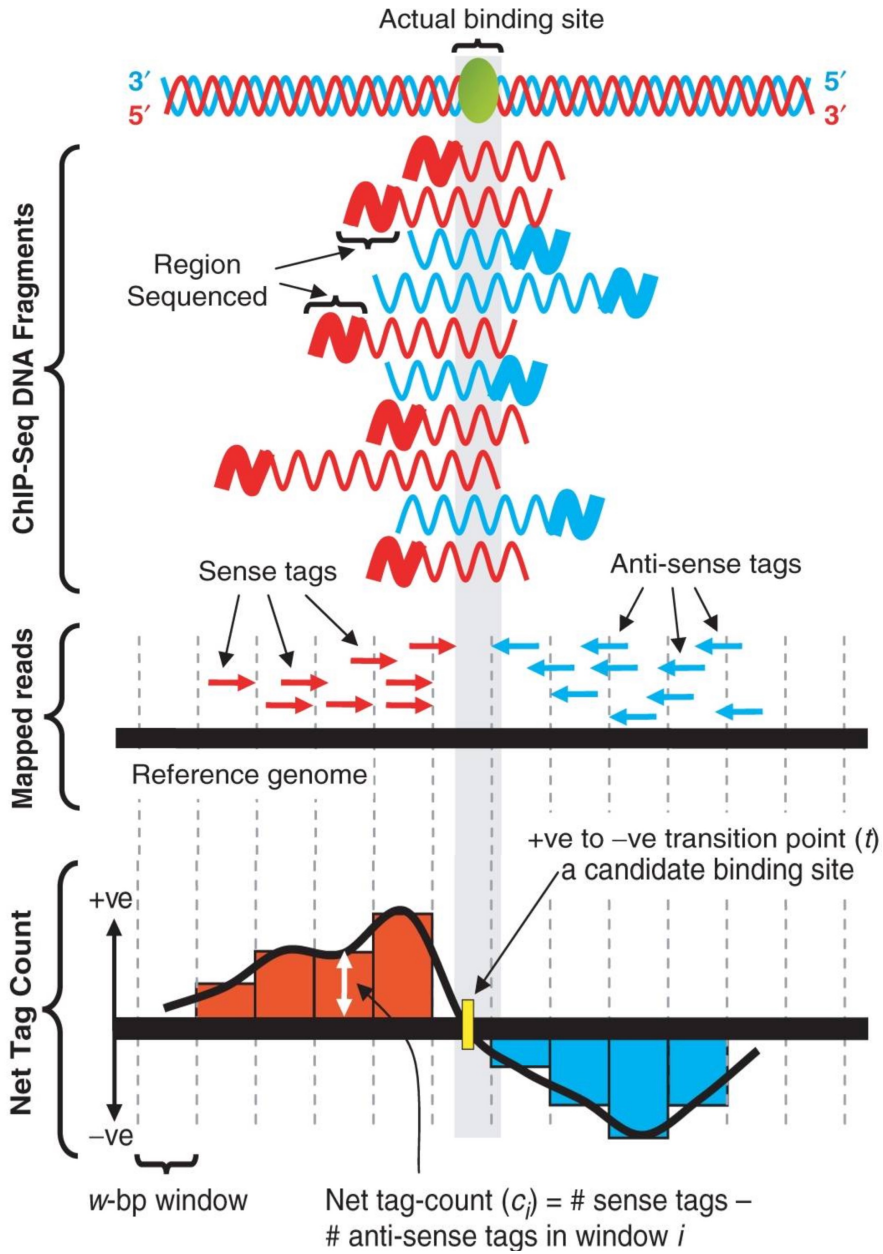
# Routine analysis

# Routine analysis – preprocessing

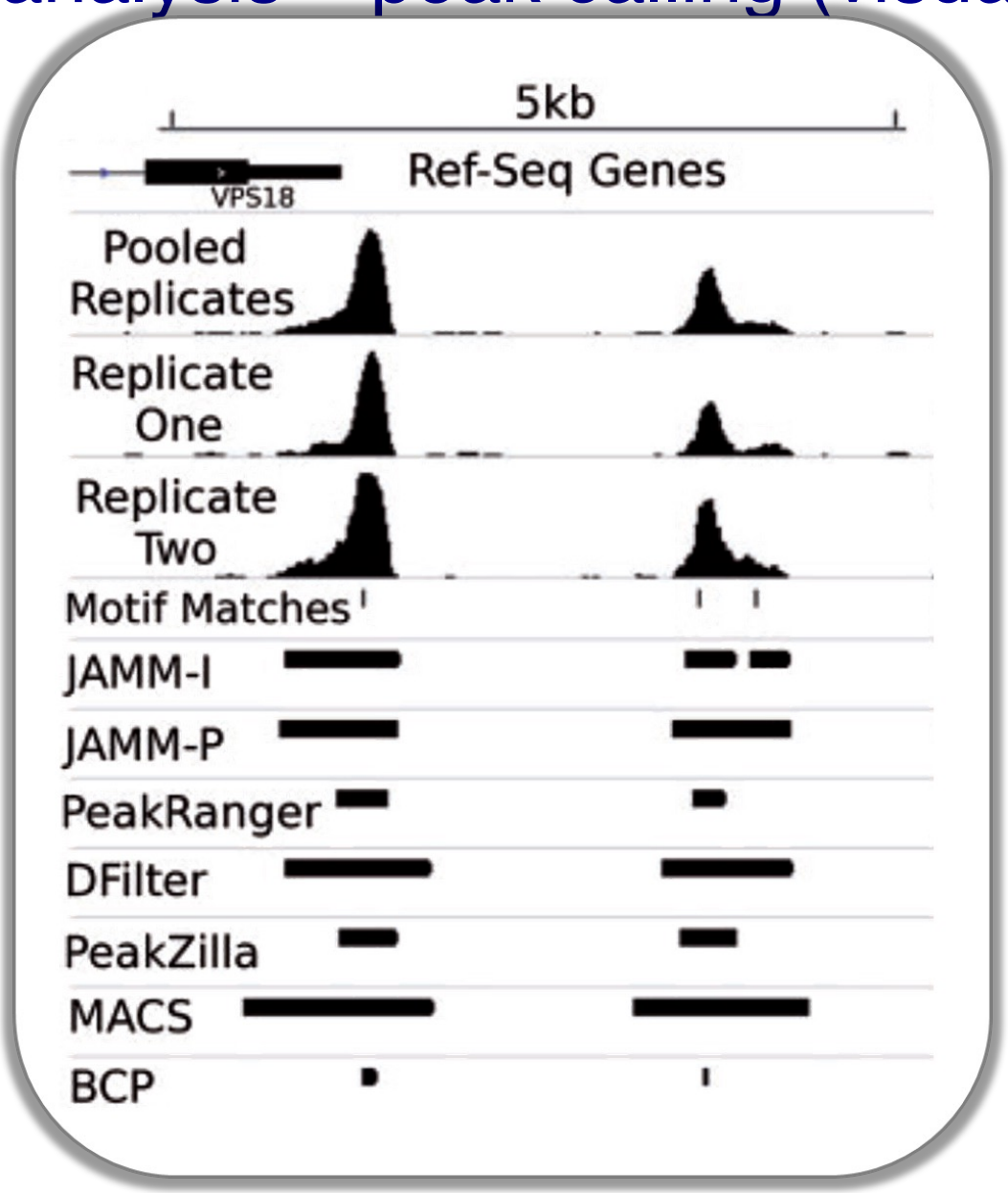
- Quality control (FastQC, Picard, …)
- Mapping (BWA, Bowtie2, …)
- Duplicate removal (SAMtools, Galaxy, …)
- Blacklist removal (in-house code)
- GC bias removal (gcapc,…)
- In-house metrics

# Routine analysis – peak calling (location & width)

- Peak for sense tags will be 1/2 the fragment length upstream
- Binding site position = mid-way between sense tag peak & antisense tag peak.
- To get binding site peak, shift sense downstream by ½ fragsize & antisense upstream by ½ fragsize.
- INPUT consideration

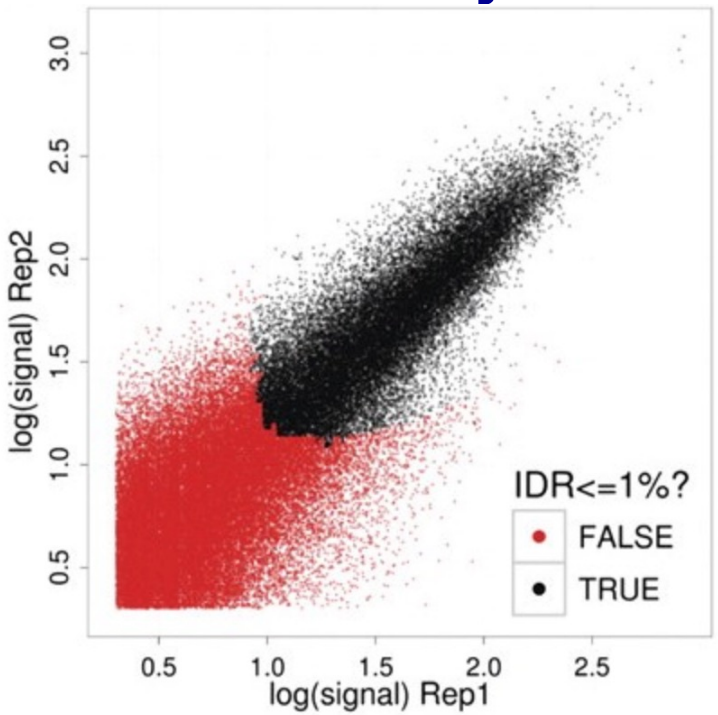


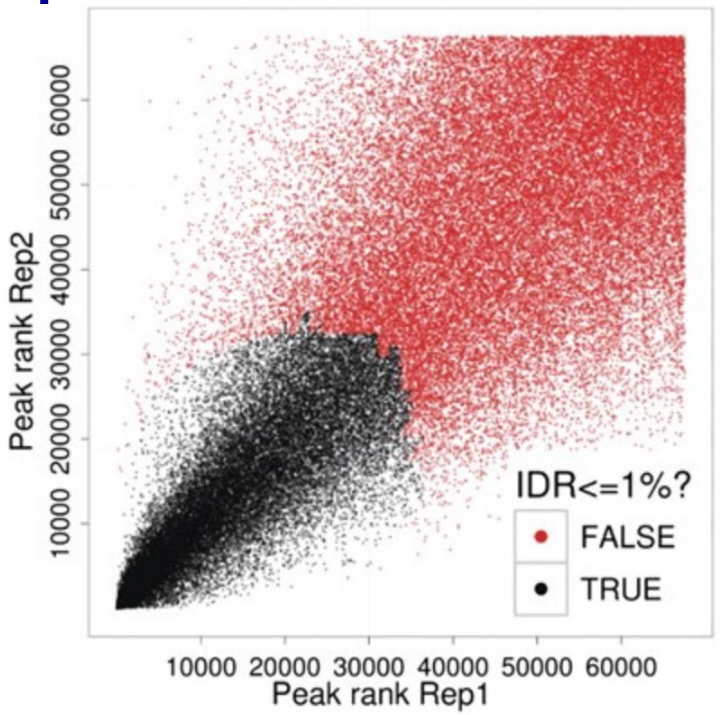
# Routine analysis – peak calling (visualization)




Other tools:
SICER, SPP, …

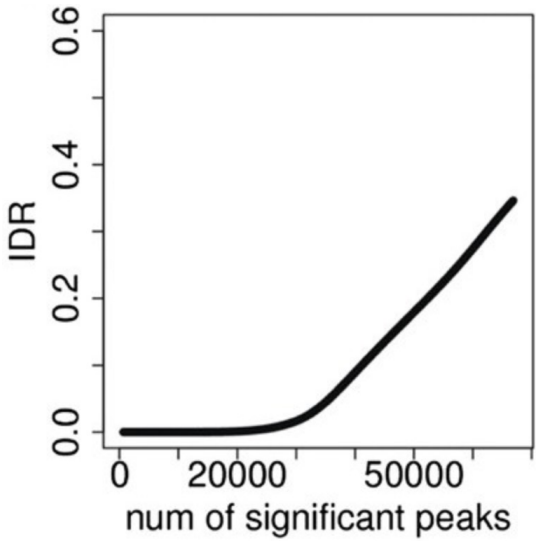
# Routine analysis – replicate concordance




Two replicates or pseudoreps

**IDR**: irreproducible discovery rate




Biostatistics & Bioinformatics

# Routine analysis – peak calling (QC)

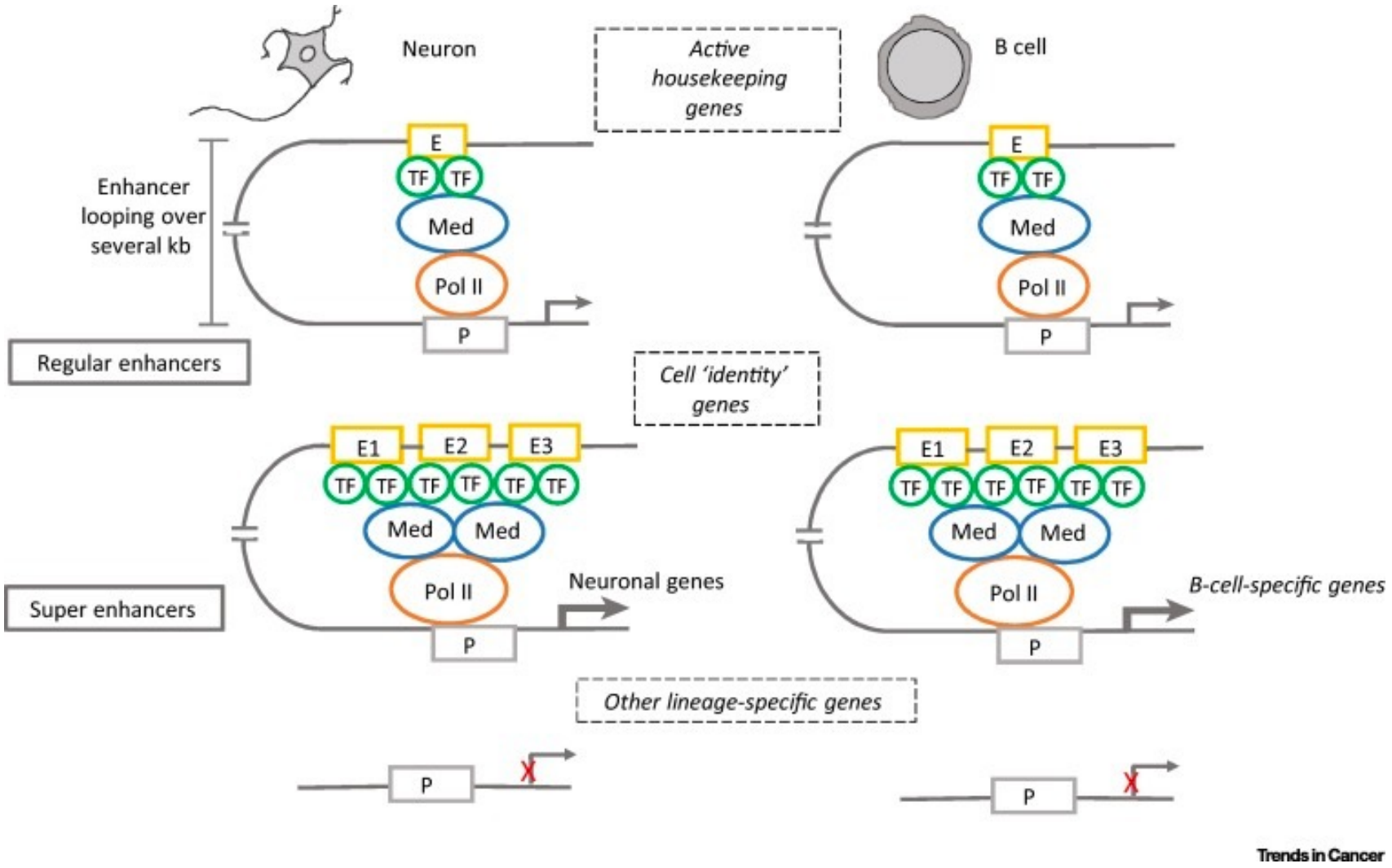
How many peaks ?

Where the peaks locate ?

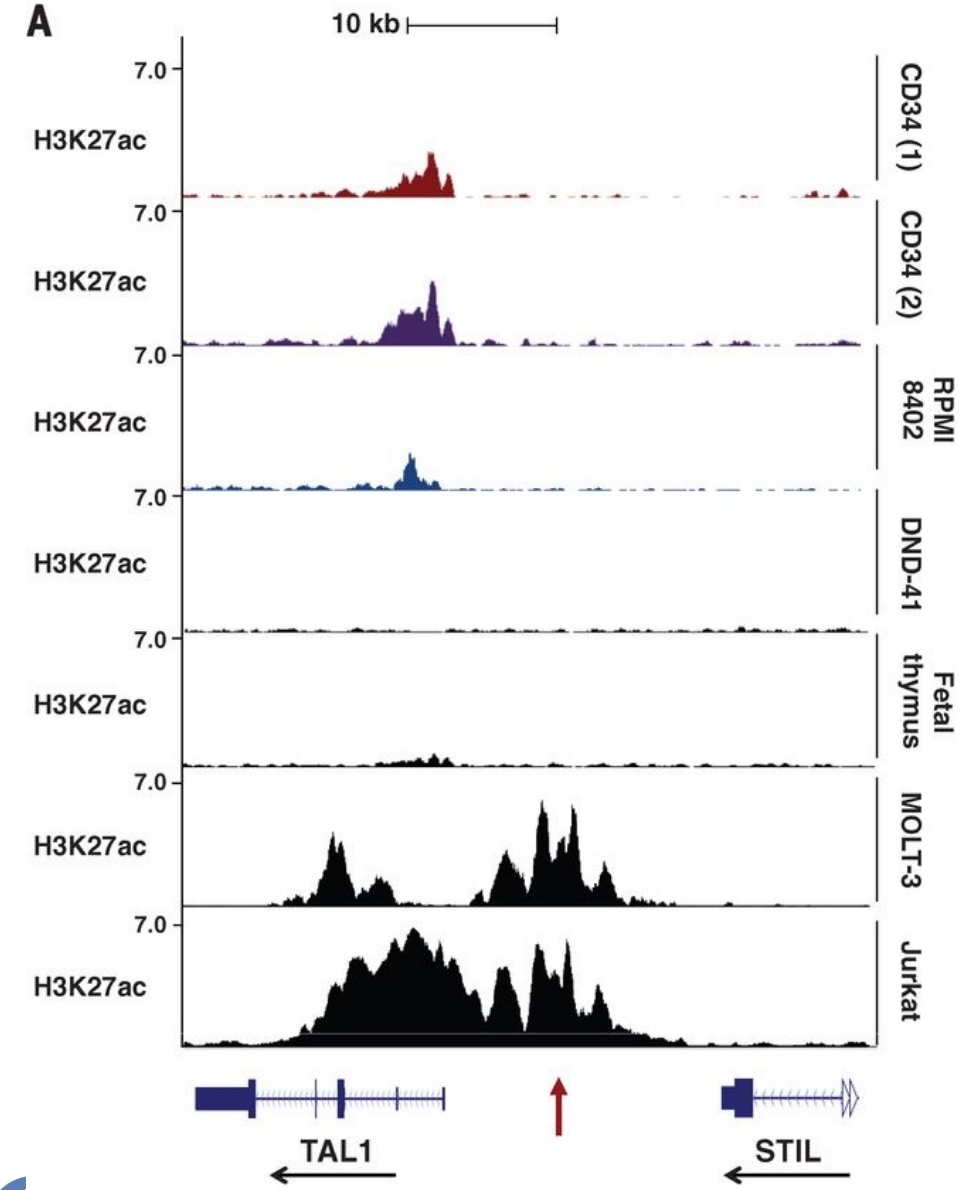
What the peak width distribute ?

Visualization ! (bigwig)

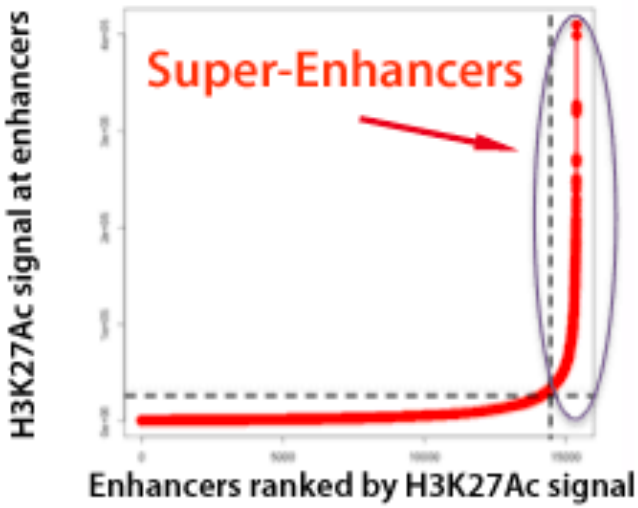
# Super enhancers



# Super enhancers – H3K27ac

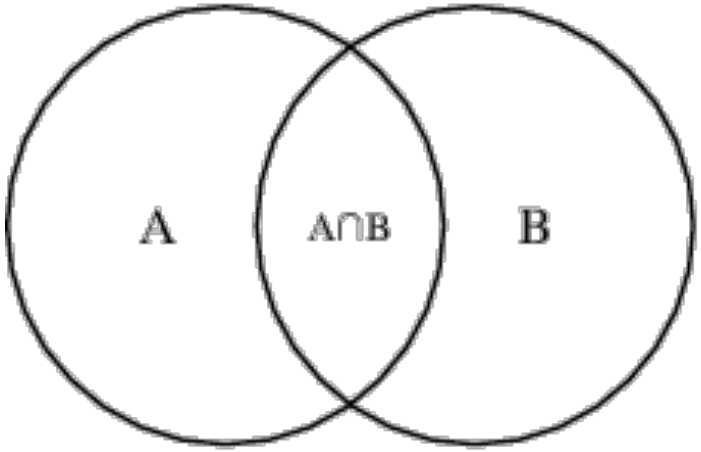


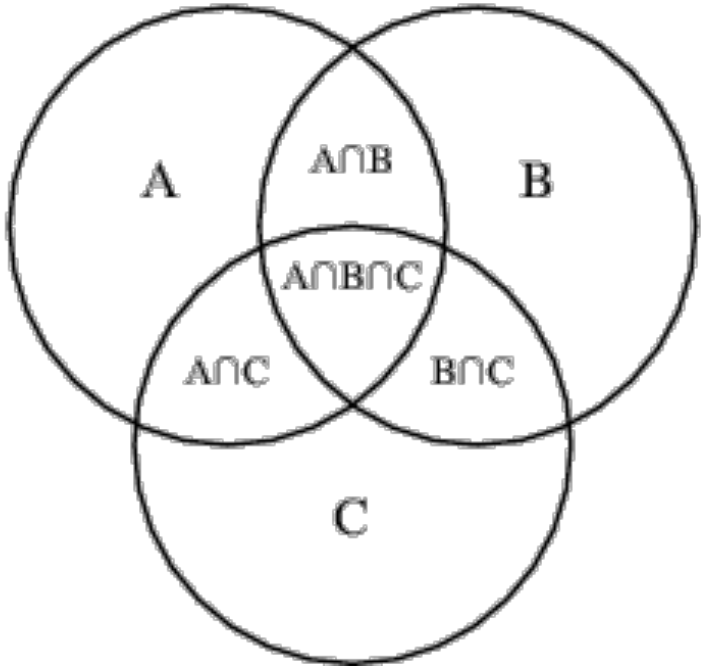

**ROSE**: stitch neighboring enhancers



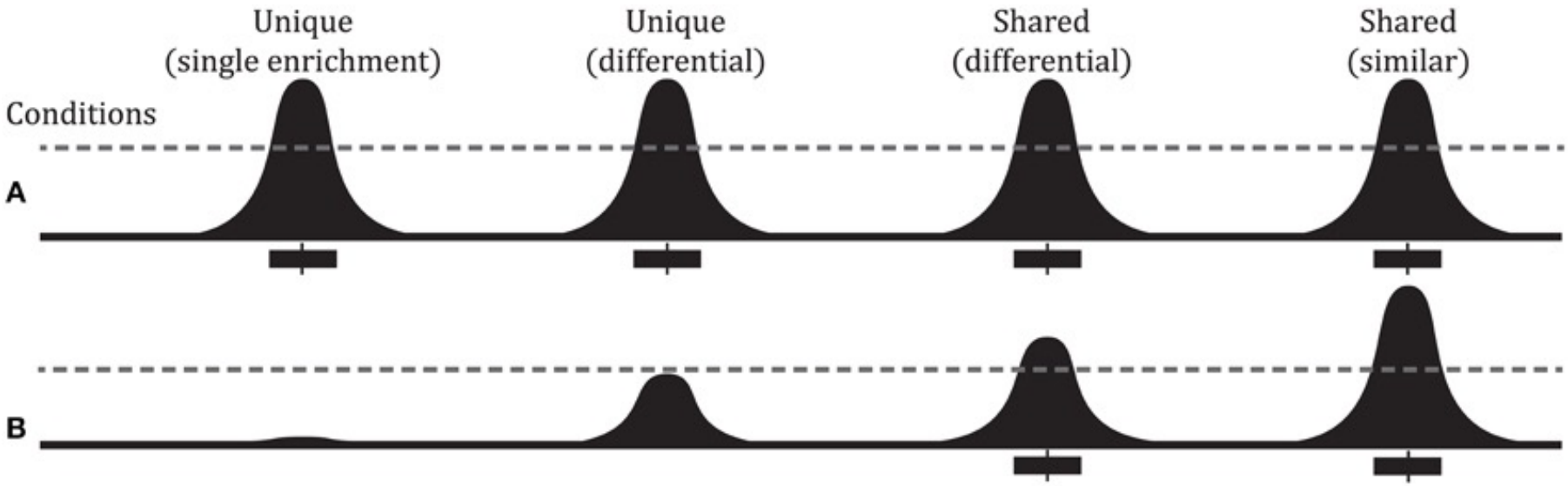
# Differential binding analysis

# Differential binding analysis – the easiest way



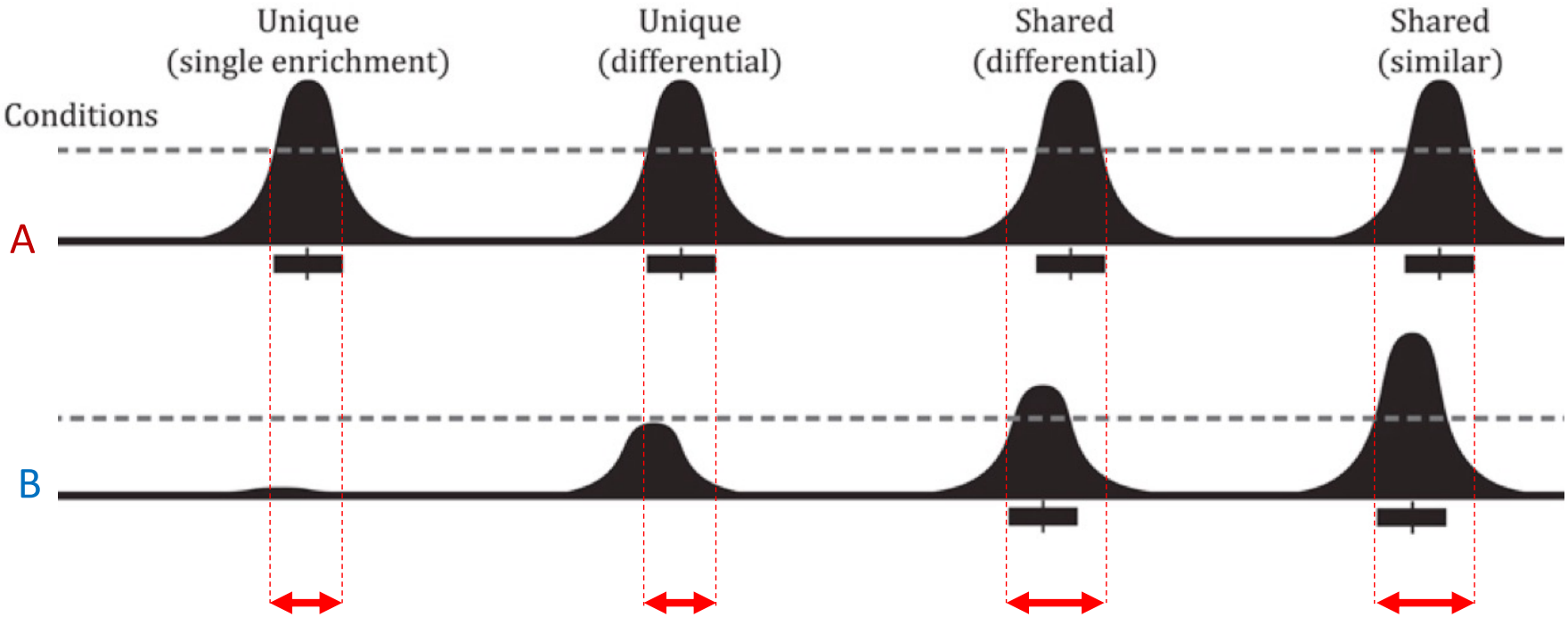



# Differential binding analysis – the targets




- Similar to differential gene expression testing
- Normalization is critical
- Alternative targets: genome-wide bins
- Less worry about INPUT

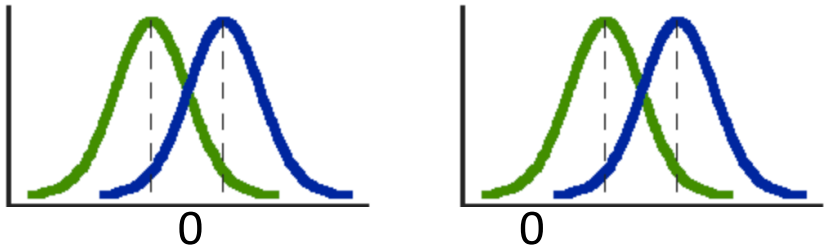
# Differential binding analysis – the targets




| | A-rep1 | A-rep2 | B-rep1 | B-rep2 |
|---|---|---|---|---|
| peak1 | $RC_{A1-1}$ | $RC_{A2-1}$ | $RC_{B1-1}$ | $RC_{B2-1}$ |
| peak2 | $RC_{A1-2}$ | $RC_{A2-2}$ | $RC_{B1-2}$ | $RC_{B2-2}$ |
| | … | … | … | … |
| peakj | $RC_{A1-j}$ | $RC_{A2-j}$ | $RC_{B1-j}$ | $RC_{B2-j}$ |

# Differential binding analysis – testing

- Statistical testing assumed on normal distribution won't work



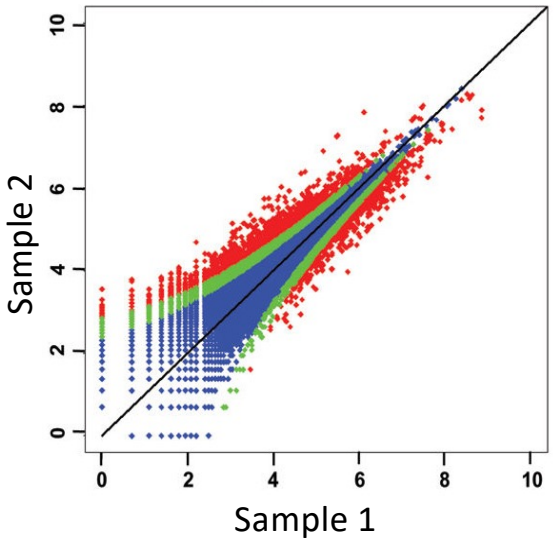

- ChIP-seq counts

| | Condition A | Condition B |
|---|---|---|
| Peak 1 | 20 | 2 |
| Peak 2 | 200 | 20 |

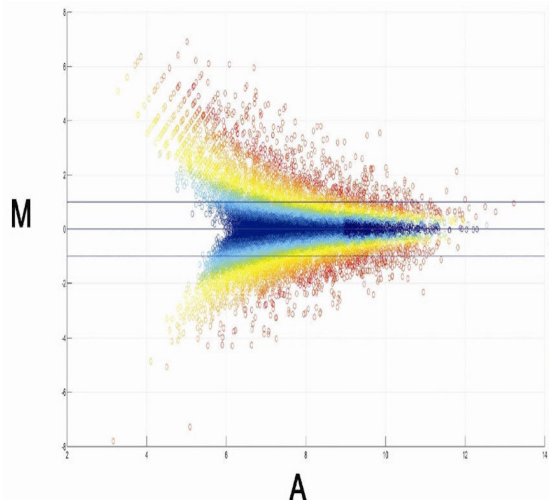
**DiffBind**: based on DESeq2, edgeR, limma-voom
**csaw**: based on genomic windows

# Useful technique in sequencing comparison

Scatter plot



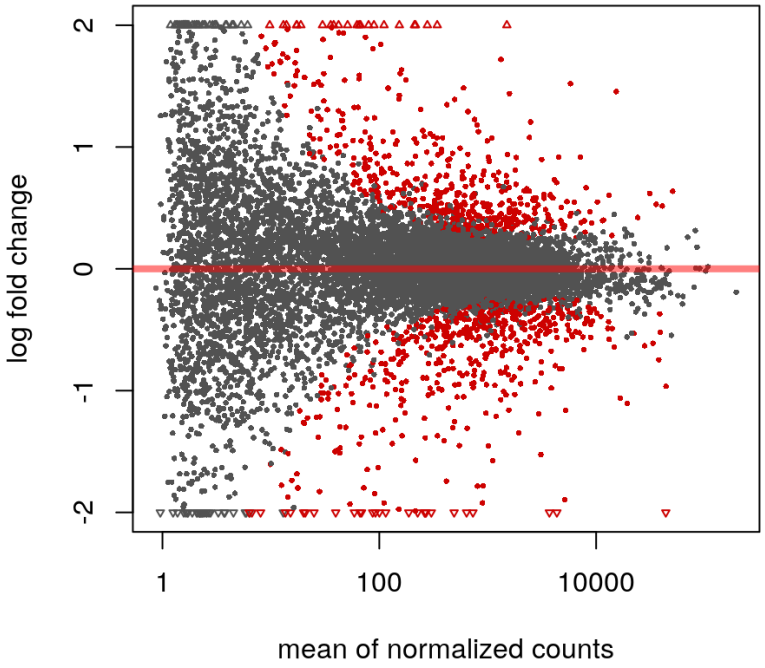

M-A plot



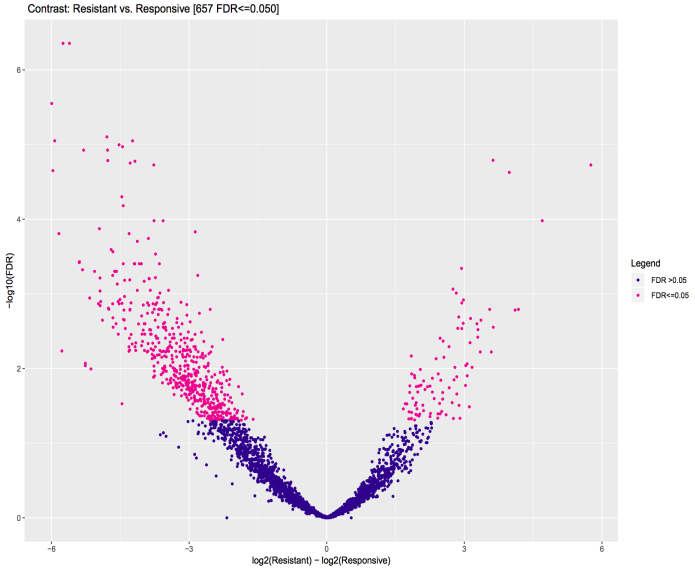

Biostatistics & Bioinformatics

# Differential binding analysis – significance

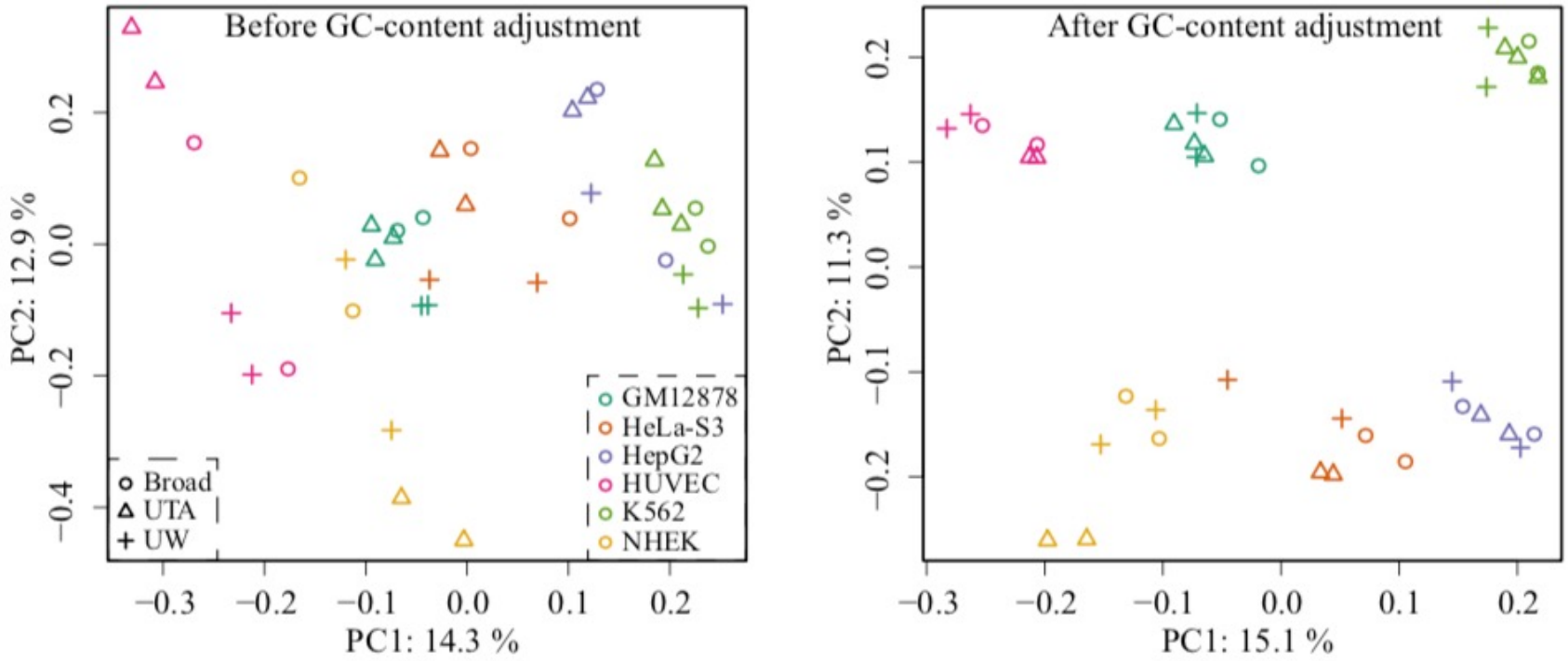
M-A plot




Volcano plot

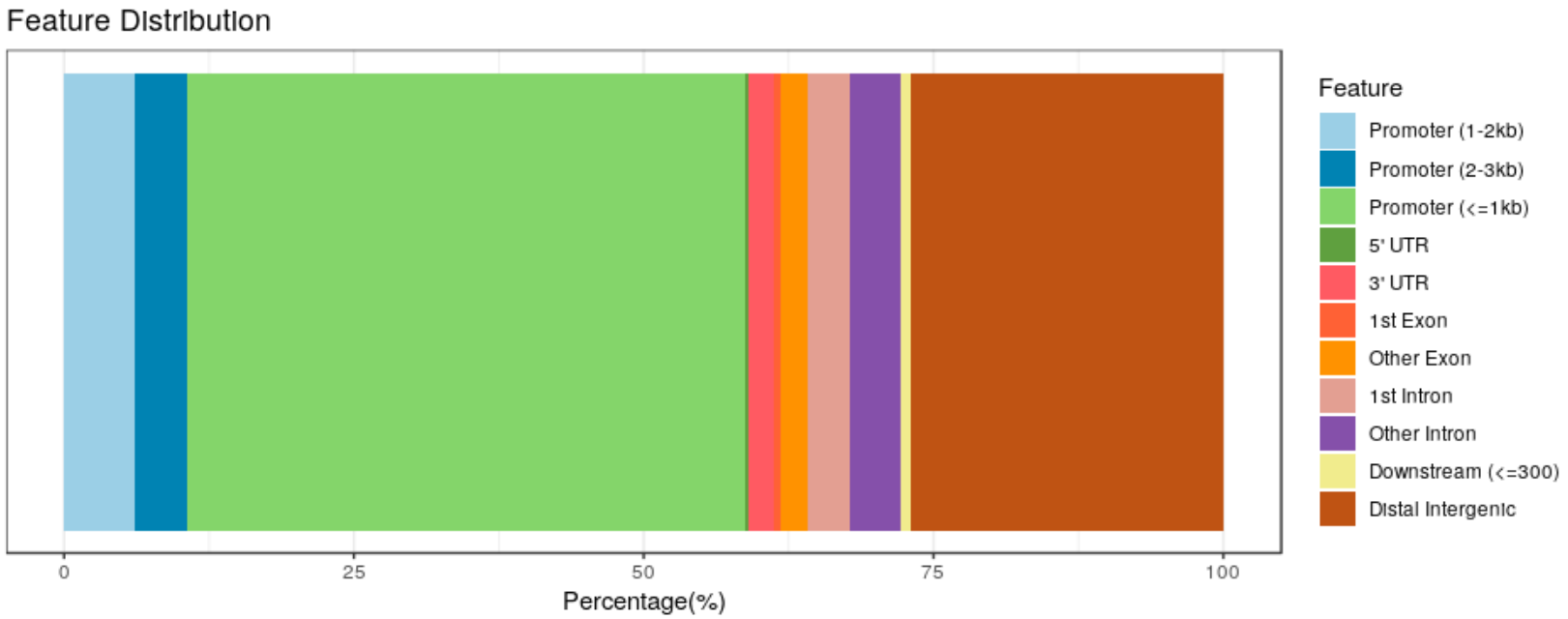


# Differential binding analysis – PCA



# Peak annotation

# Peak annotation – peak or differential peak

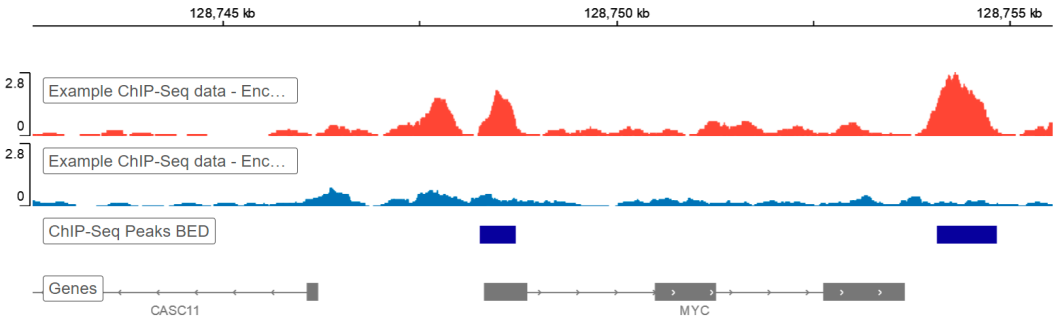



- Connecting peaks to mutations
- Connecting peaks to RNA-seq signals
- Connecting peaks to other ChIP-seq

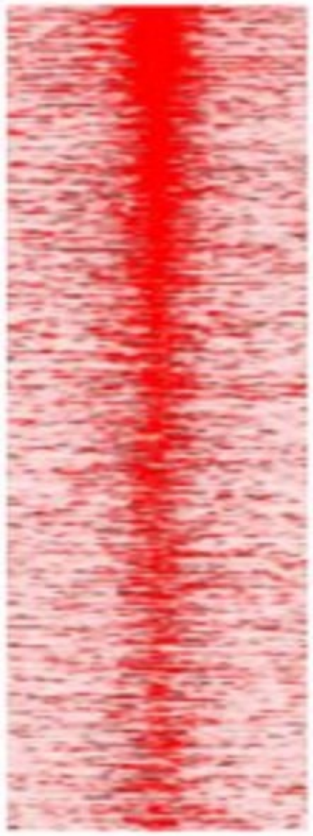
Tools: **ChIPseeker, PAVIS, Homer, GenomicRanges**

# Peak annotation – Motif

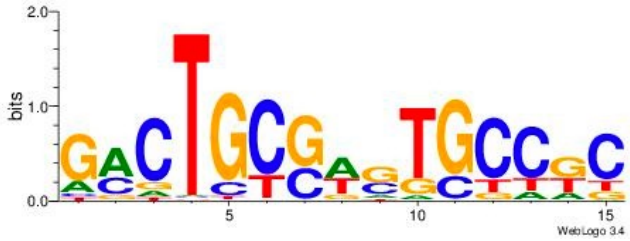
Individual peaks




Genome-wide peaks

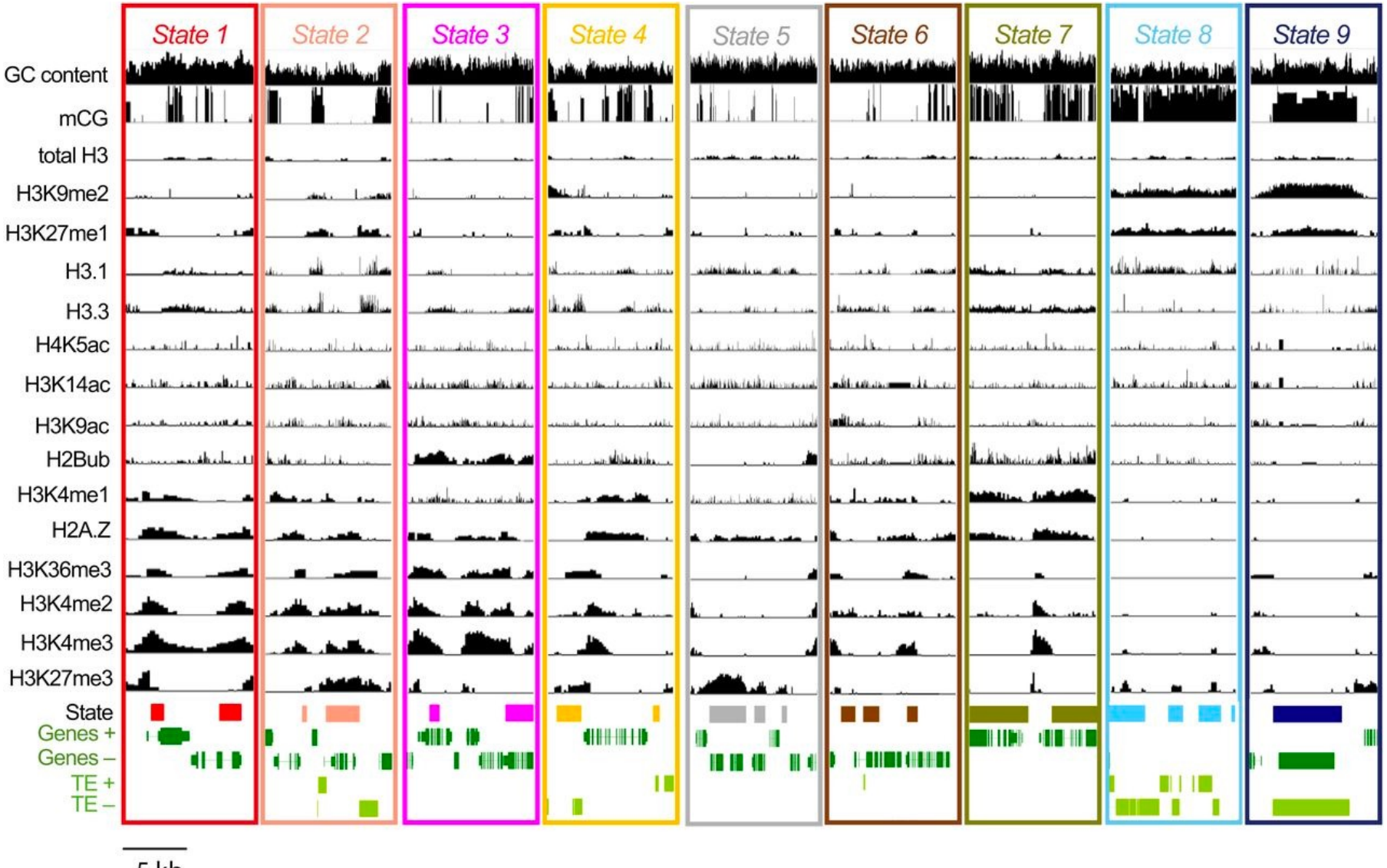


Summarized motif




Tools: **MEME, Homer, PWMEnrich, TFBSTools, deepTools**

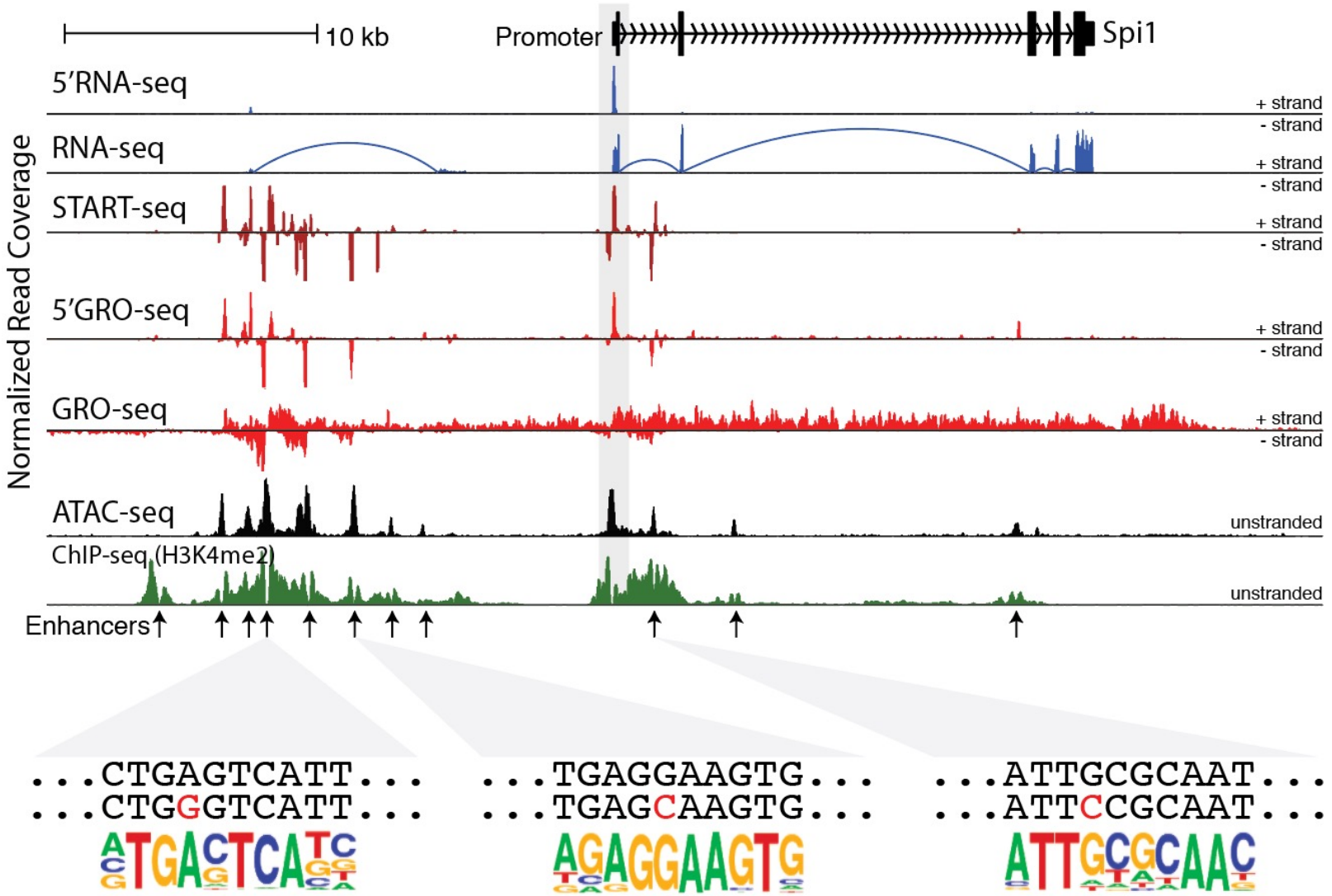
# Peak annotation – same cell multiple ChIP-seq



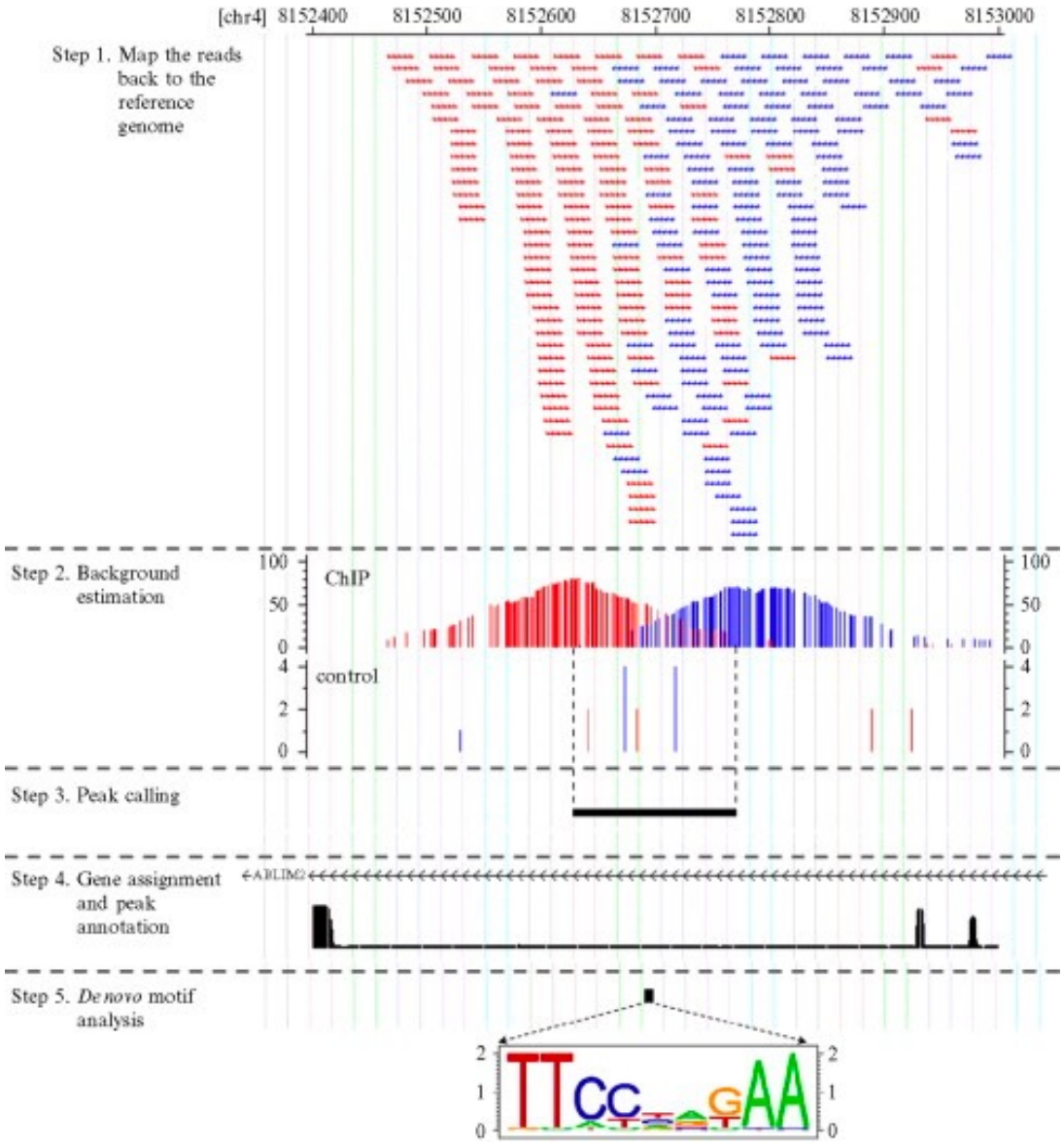

Biostatistics & Bioinformatics

Tool: **ChromHMM**

# Integrative analysis



# Basic pipeline summarization

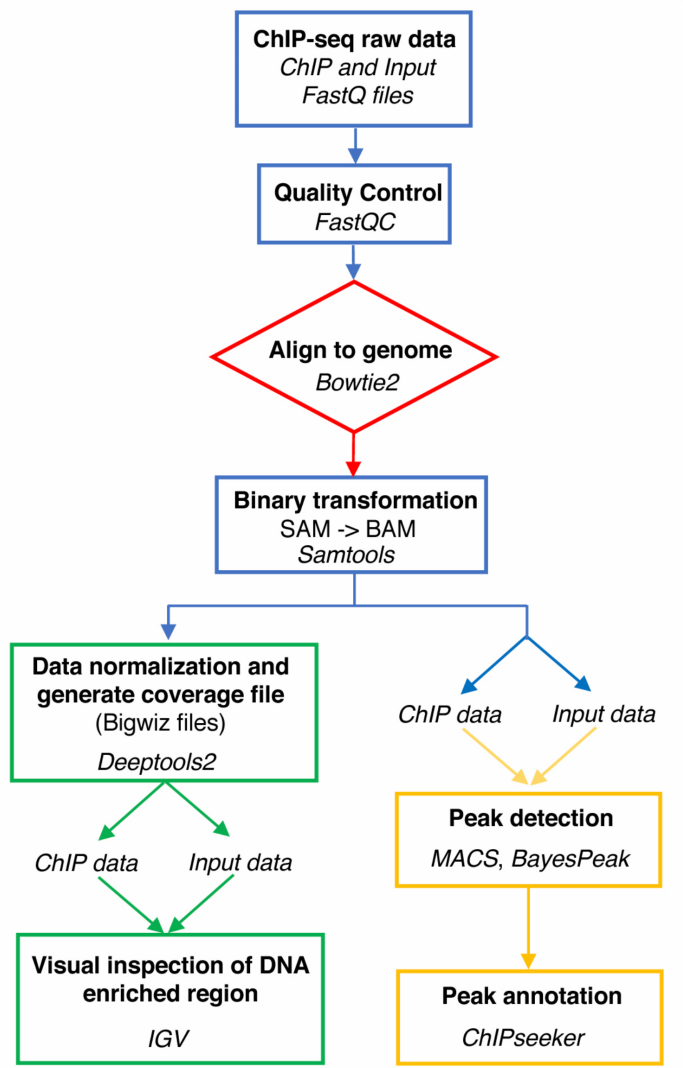



Biostatistics & Bioinformatics

# ChIP-seq Pipeline

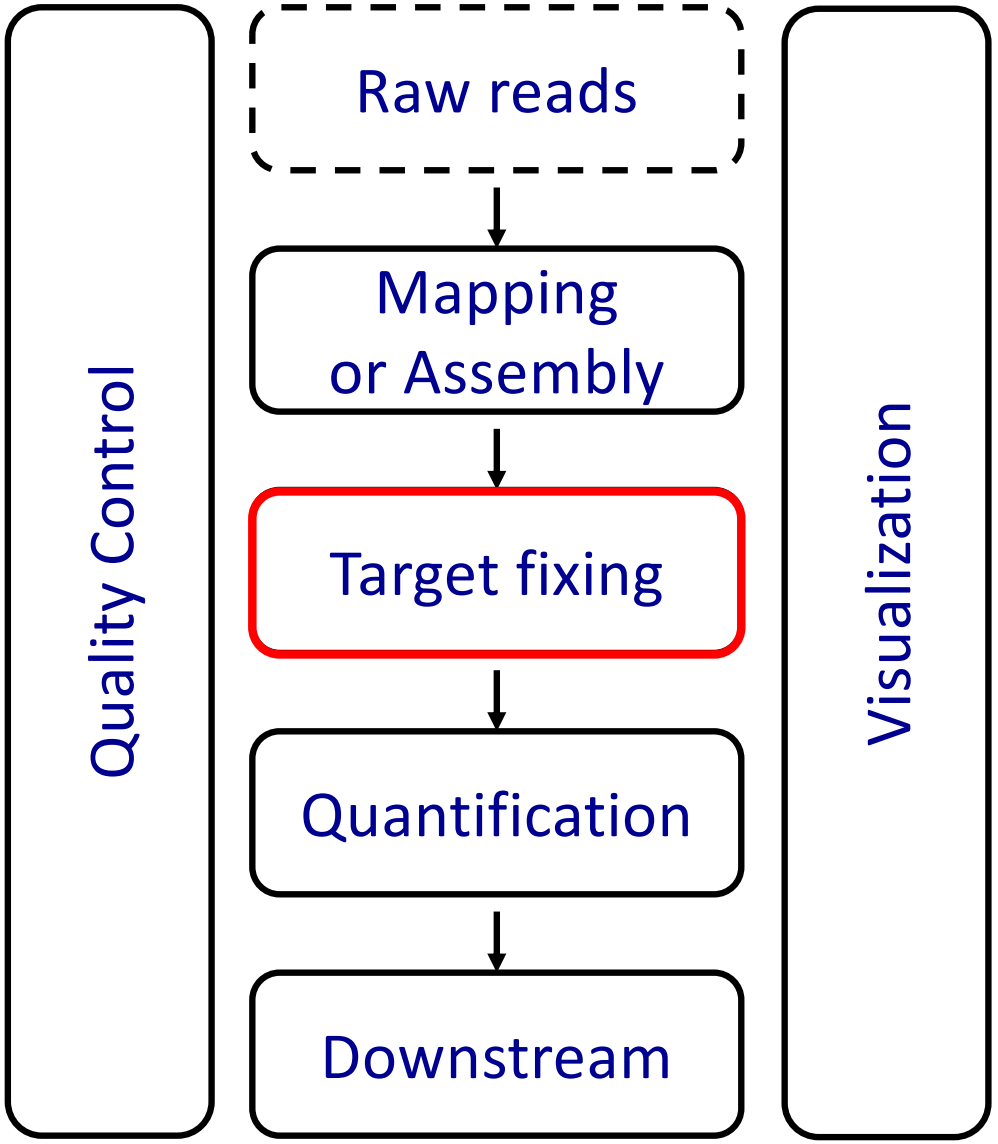
*QC not covered in the pipeline:*

- Before alignment *vs.* after alignment
- Duplicate reads
- Blacklist regions
- Signal to noise ratio
- Mapping rate: ~80% (bowtie2, bwa)
- Peaks: dozens of thousands
- Sophisticated QC from ENCODE




Giner-Lamia *et al.* 2018

# Paradigm in sequencing data analysis



# Next lecture

*R*

*Rstudio*