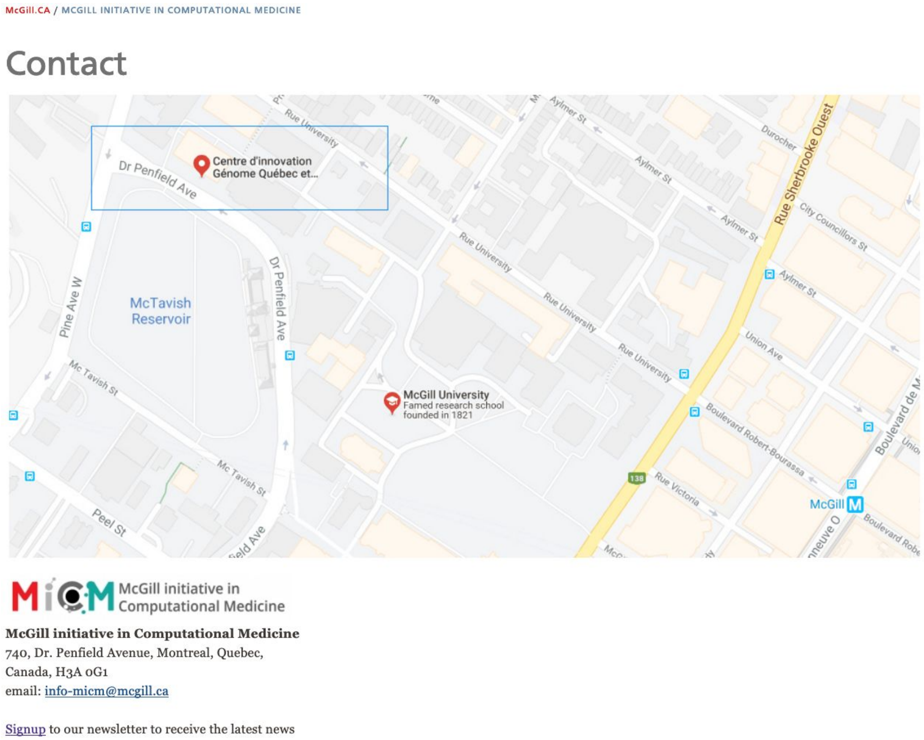
ChIP-seq analysis

Instructor: Ariel Madrigal Aguirre
TA: Adrien Osakwe
November 23, 2022

Mission : aims to deliver inter-disciplinary research programs and empower the use of data in health research and health care delivery



McGill.CA / MCGILL INITIATIVE IN COMPUTATIONAL MEDICINE

Contact

McGill initiative in
Computational Medicine

McGill initiative in Computational Medicine
740, Dr. Penfield Avenue, Montreal, Quebec,
Canada, H3A 0G1
email: info-micm@mcgill.ca

Signup to our newsletter to receive the latest news

https://www.mcgill.ca/micm

Outline:

1 Introduction

2 Alignment and identification of binding sites

3 Quality control

4 Visualization

5 Motif finding and gene set enrichment analysis

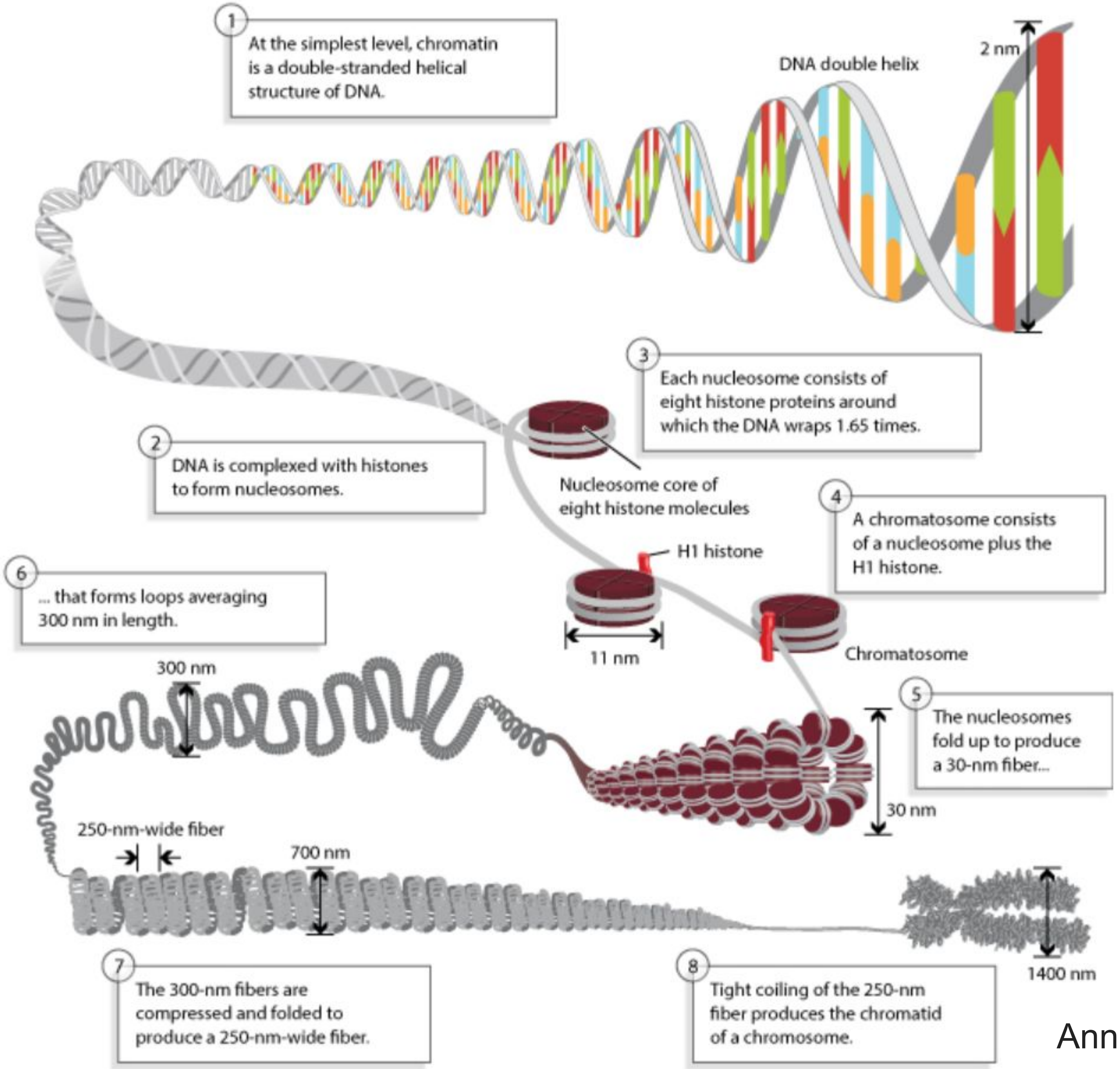
6 Concluding remarks

This is an interactive workshop :)

Feel free to interrupt or raise your hand to ask questions

Part 1: Introduction to ChIP-seq

Chromatin



1 At the simplest level, chromatin is a double-stranded helical structure of DNA.

2 DNA is complexed with histones to form nucleosomes.

3 Each nucleosome consists of eight histone proteins around which the DNA wraps 1.65 times.

4 A chromatosome consists of a nucleosome plus the H1 histone.

5 The nucleosomes fold up to produce a 30-nm fiber...

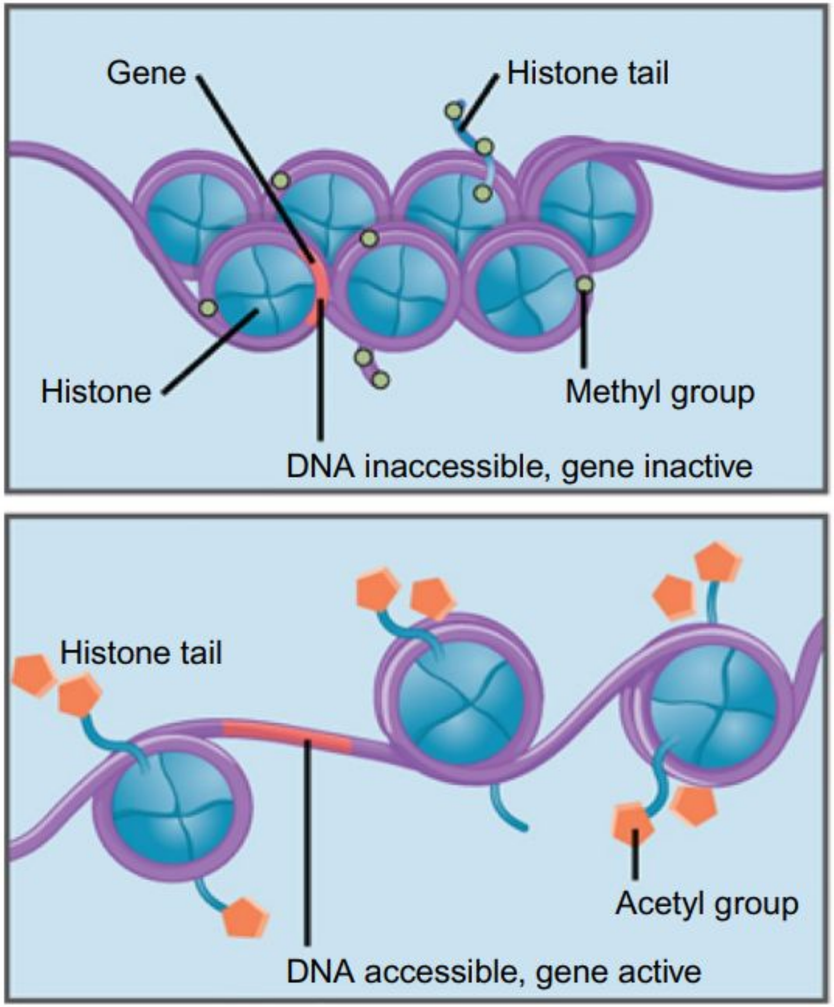
6 ... that forms loops averaging 300 nm in length.

7 The 300-nm fibers are compressed and folded to produce a 250-nm-wide fiber.

8 Tight coiling of the 250-nm fiber produces the chromatid of a chromosome.

Annunziato, A. (2008)

Chromatin

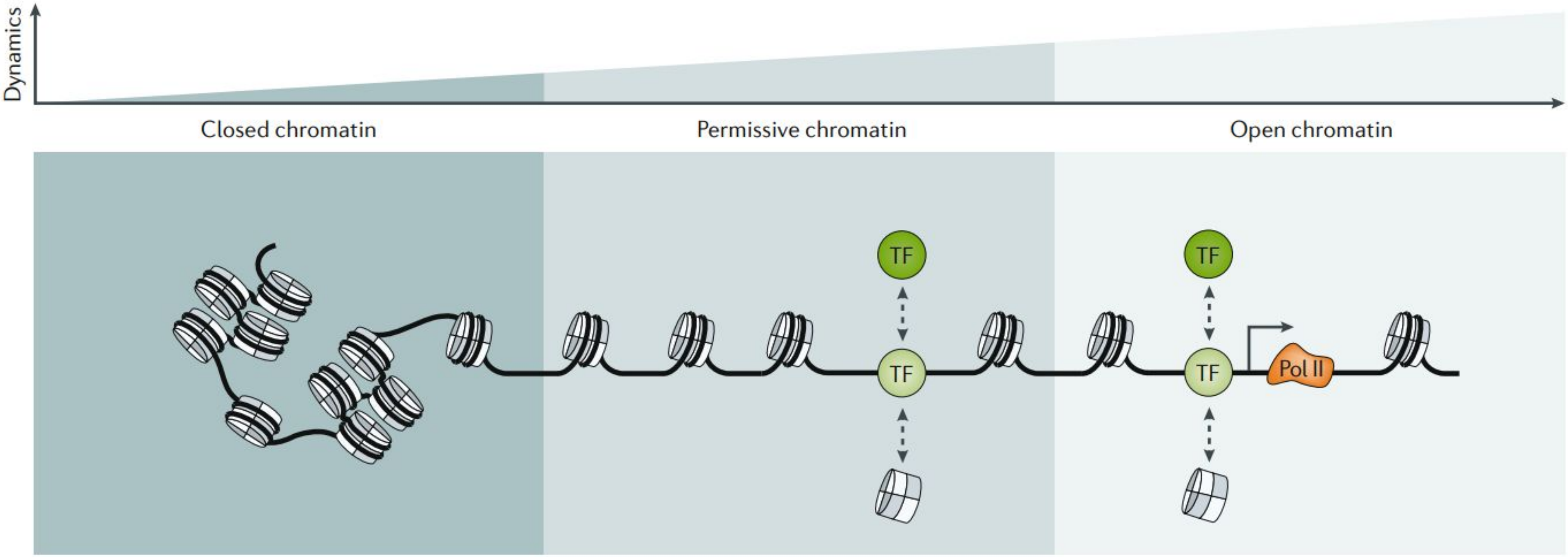


Methylation of DNA and histones causes nucleosomes to pack tightly together. Transcription factors cannot bind the DNA, and genes are not expressed.

Histone acetylation results in loose packing of nucleosomes. Transcription factors can bind the DNA and genes are expressed.

Mobley (2019)

Chromatin

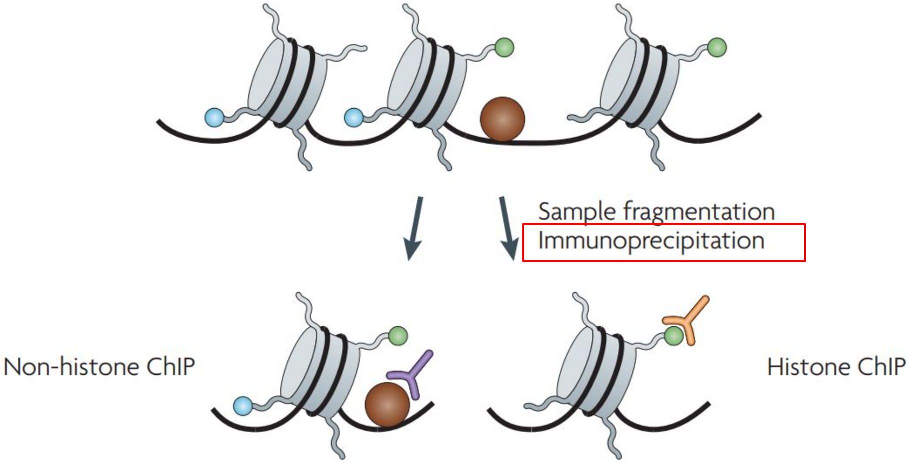


The accessible genome comprises ~2–3% of total DNA sequence yet captures more than 90% of regions bound by Transcription Factors (ENCODE)

Klemm, S. (2019)

What are we looking for?

Interactions between proteins (Histone modifications and DNA binding proteins) and DNA

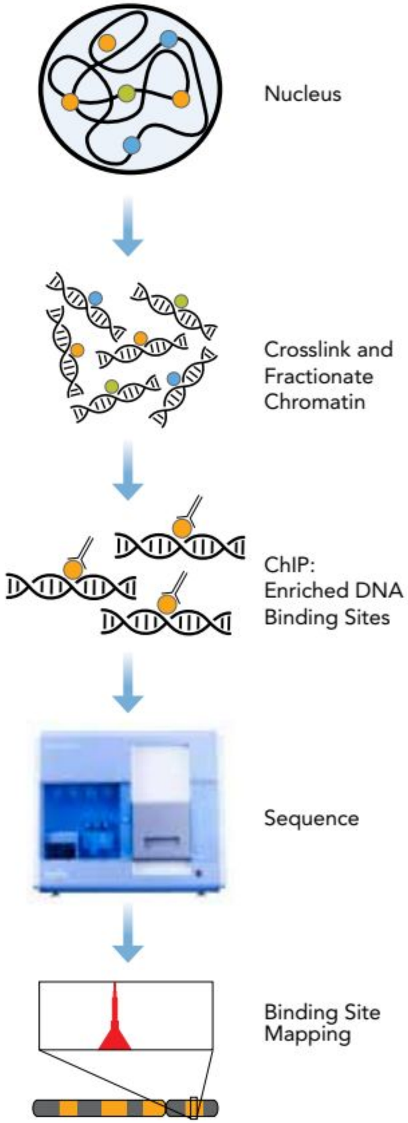


How can we find this?

Enrich for these interactions and find the DNA sequences that are over-represented and represent binding

Park (2009)

ChIP-seq experimental workflow

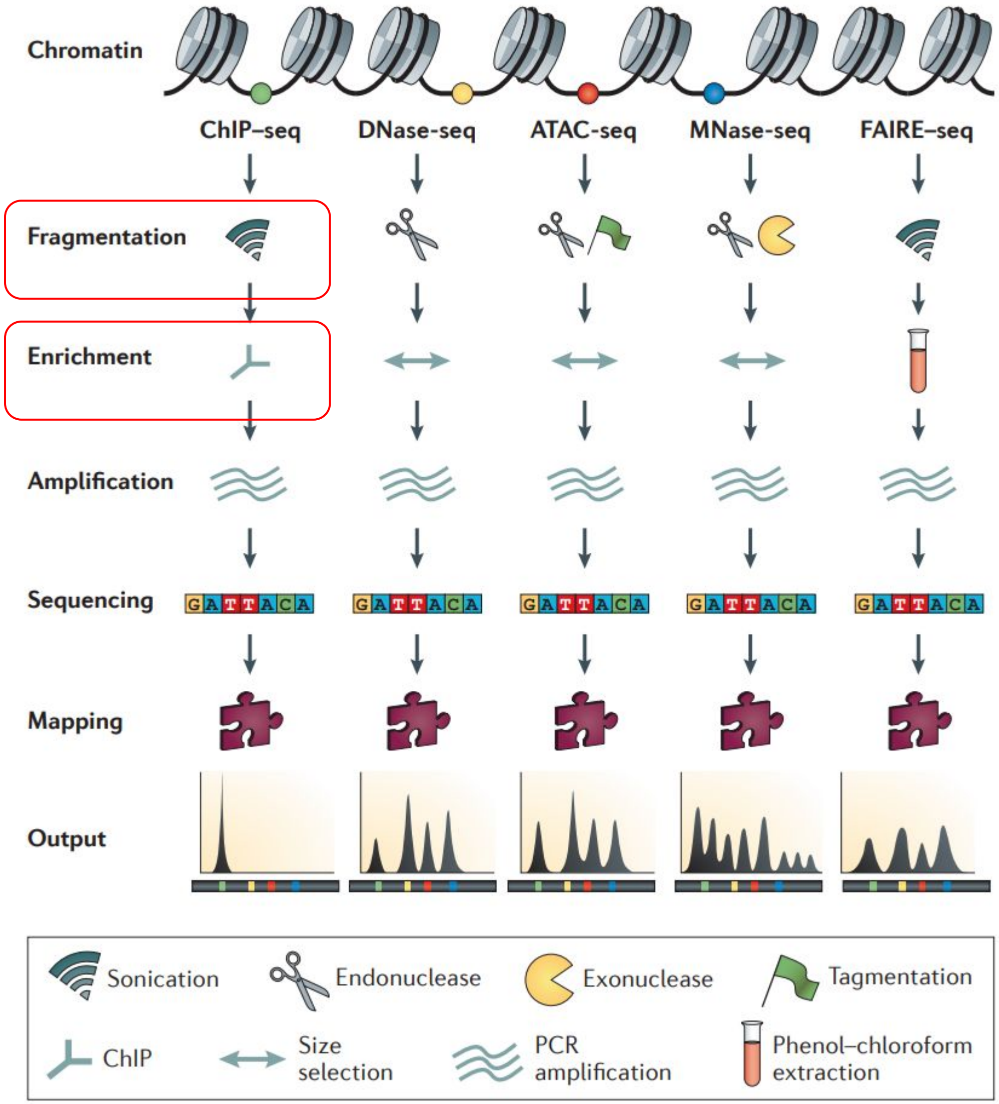

Nucleus
Crosslink and Fractionate Chromatin
ChIP: Enriched DNA Binding Sites
Sequence
Binding Site Mapping

Usually using sonication

Immunoprecipitation

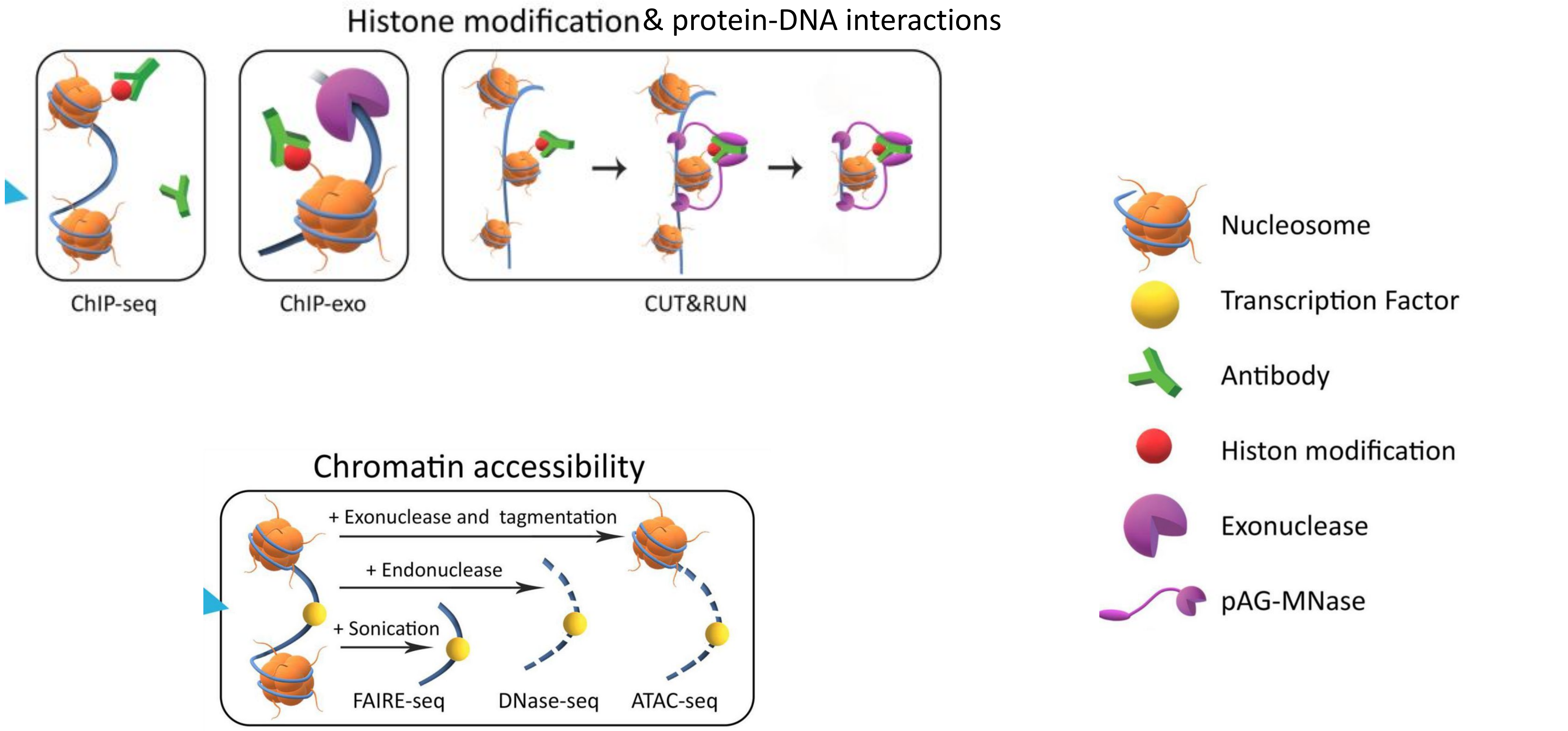
Illumina datasheet

Comparison of ChIP-seq to other techniques



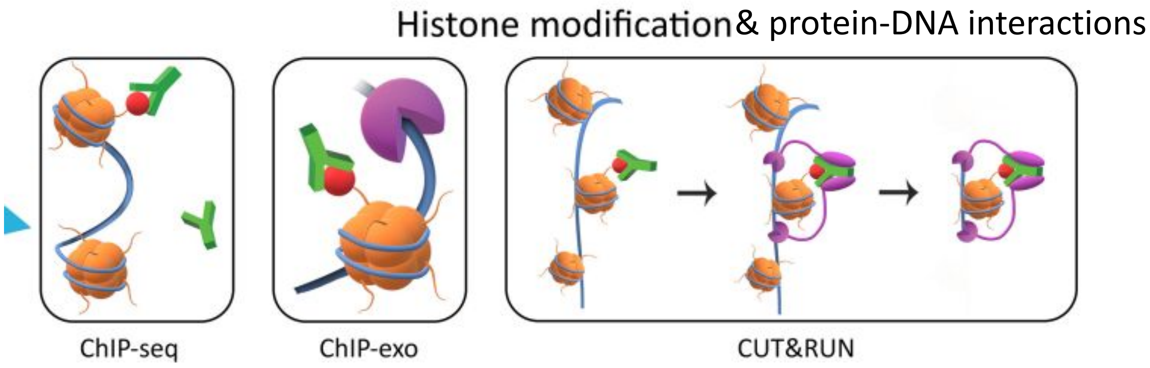
Meyer & Liu (2014)

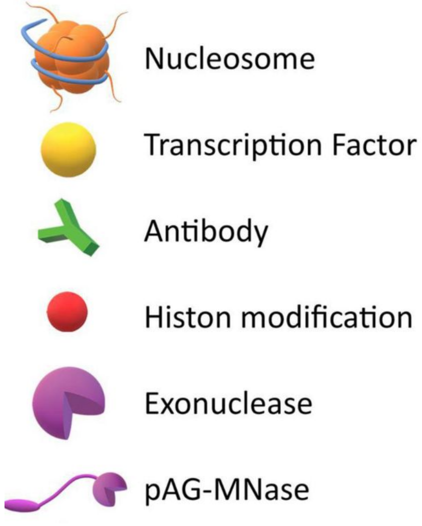
Comparison of ChIP-seq to other techniques



Mehrmohamadi (2021)

Comparison of ChIP-seq to other techniques





ChIP-exo
ChIP- nexus

Higher resolution of binding sites, from hundreds of base pairs in ChIP-seq to a single base resolution

CUT & RUN
CUT & TAG

Lower input requirements
Higher resolution

Mehrmohamadi (2021)

Sources of bias in ChIP-seq: sonication

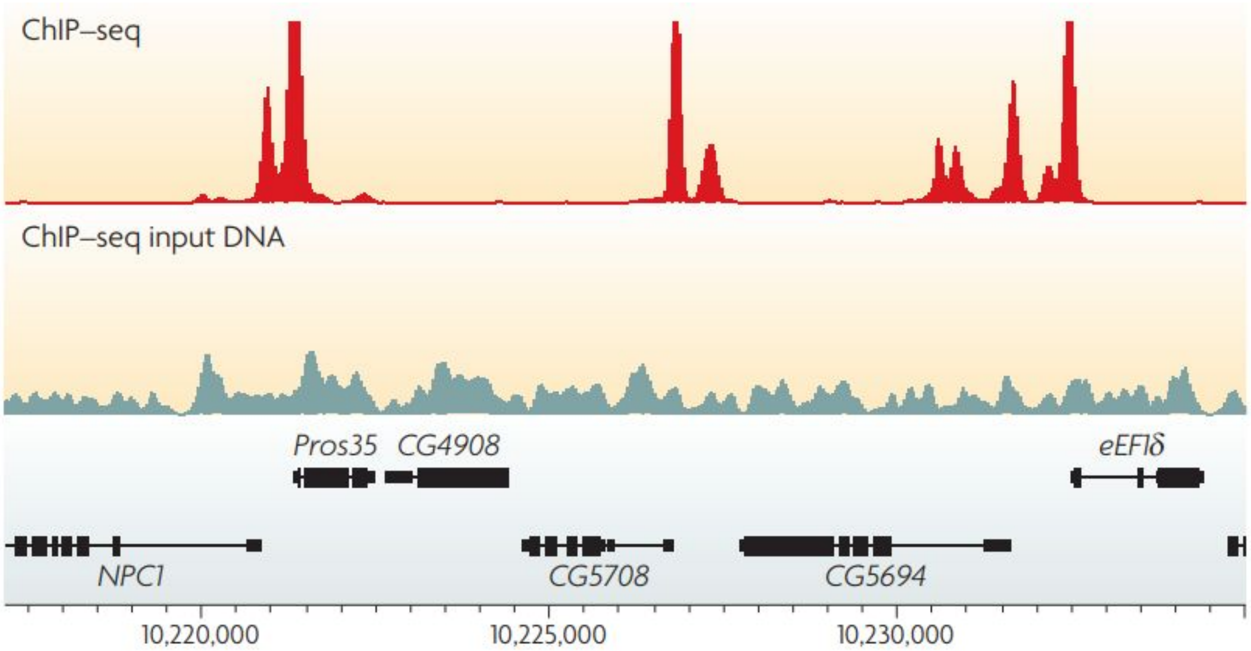
The problem

Shearing of DNA(usually by sonication), does not result in uniform fragmentation of the genome

- open chromatin regions tend to be fragmented more easily than closed regions, which creates an uneven distribution of sequence tags across the genome

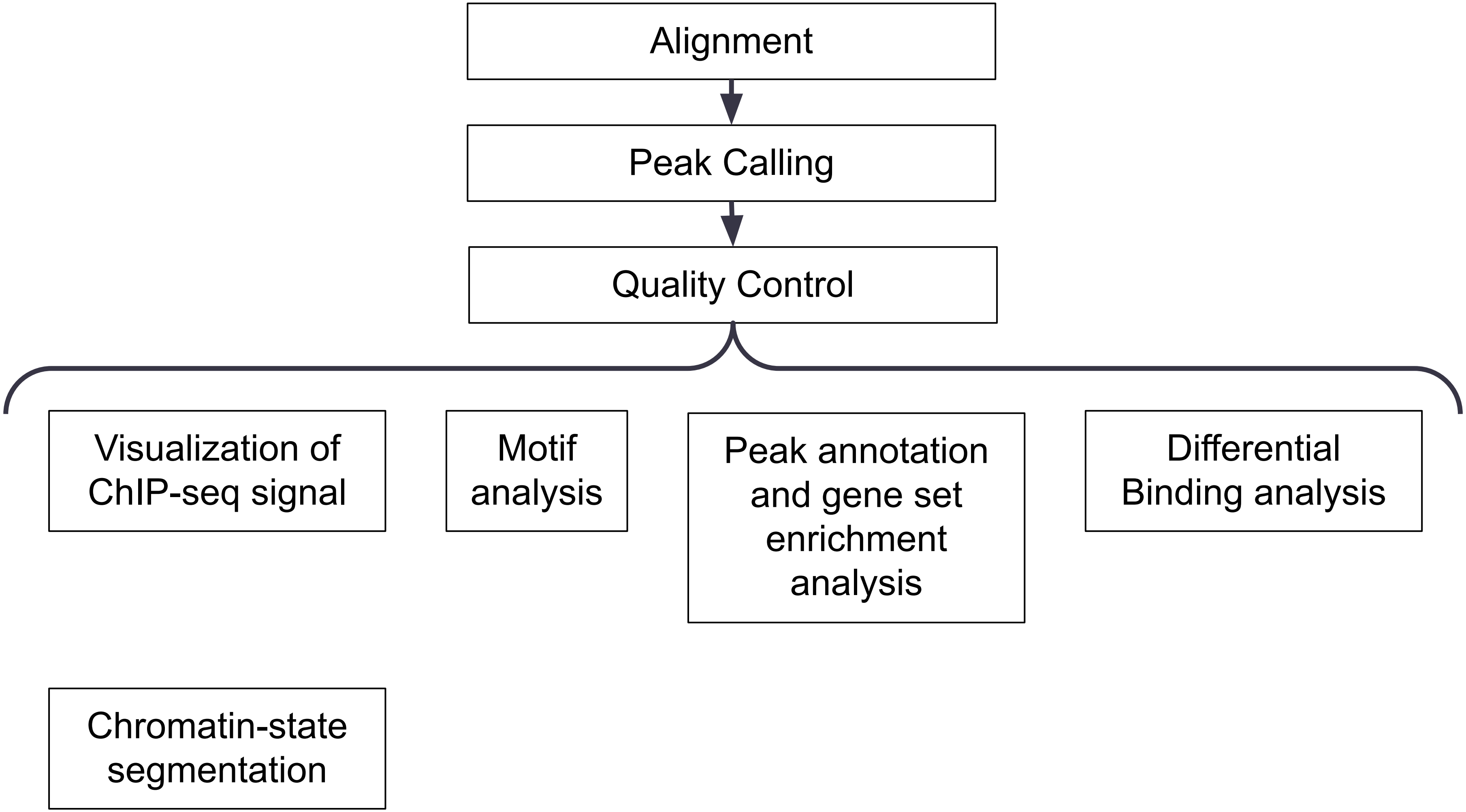
Input DNA Control:

- The ChIP experiment without the ‘immunoprecipitation’ step (no antibody)
- Corrects for bias related to the shearing of DNA and amplification



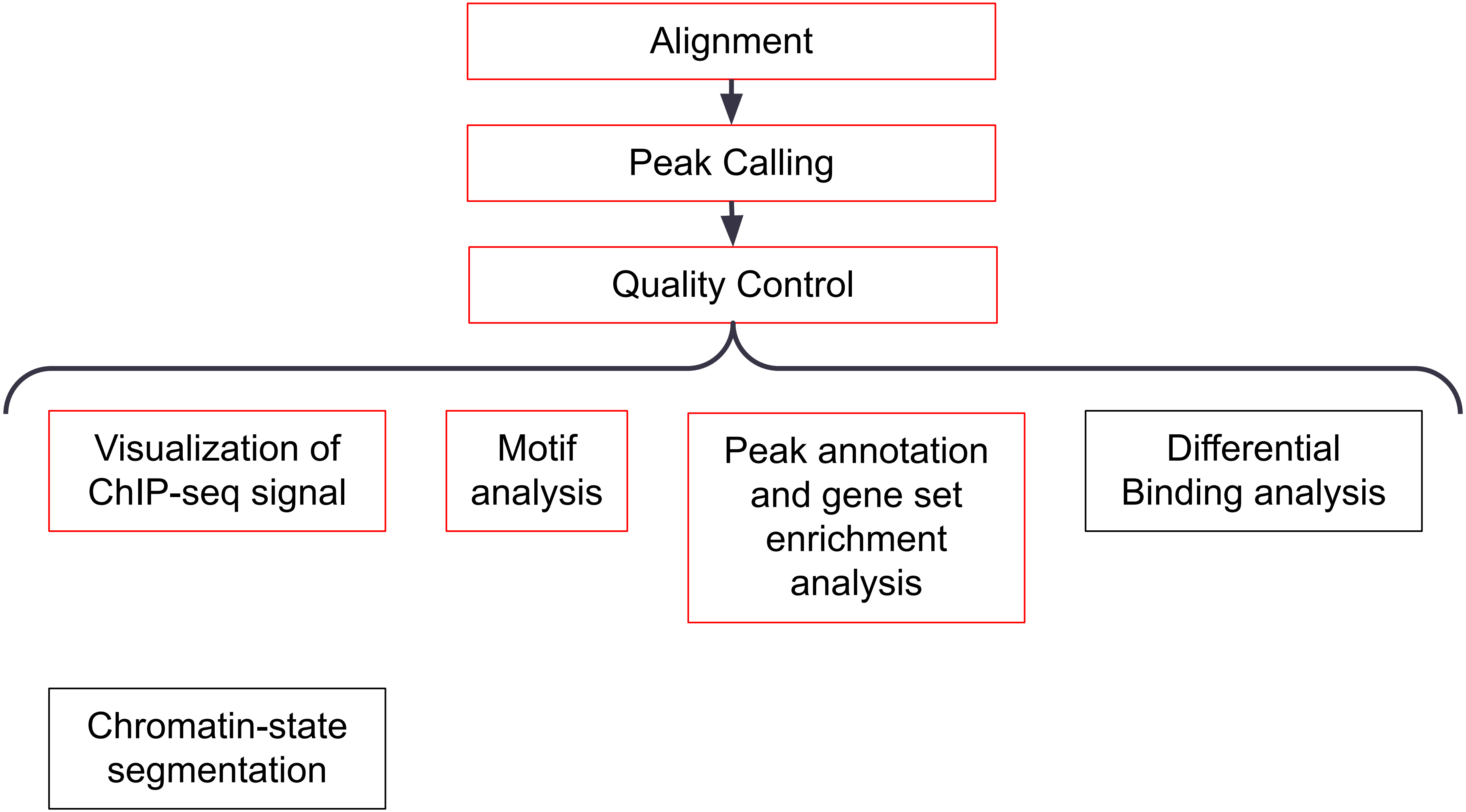
Park (2009)

ChIP-seq analysis


Alignment
Peak Calling
Quality Control
Visualization of ChIP-seq signal
Motif analysis
Peak annotation and gene set enrichment analysis
Differential Binding analysis
Chromatin-state segmentation

Based on Santiago et al (2018)

ChIP-seq analysis



Alignment

Peak Calling

Quality Control

Visualization of ChIP-seq signal

Motif analysis

Peak annotation and gene set enrichment analysis

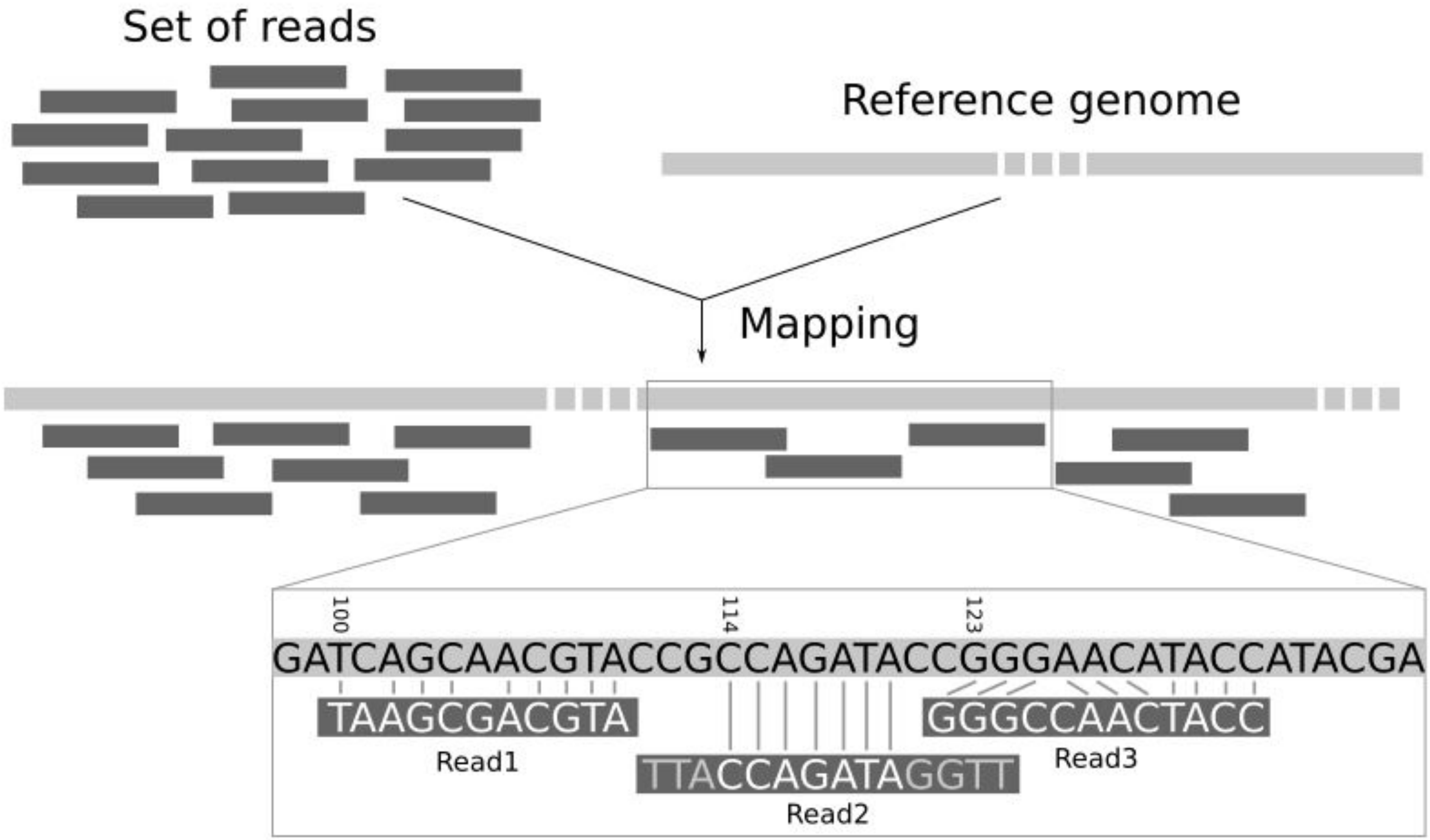
Differential Binding analysis

Chromatin-state segmentation

Based on Santiago et al (2018)

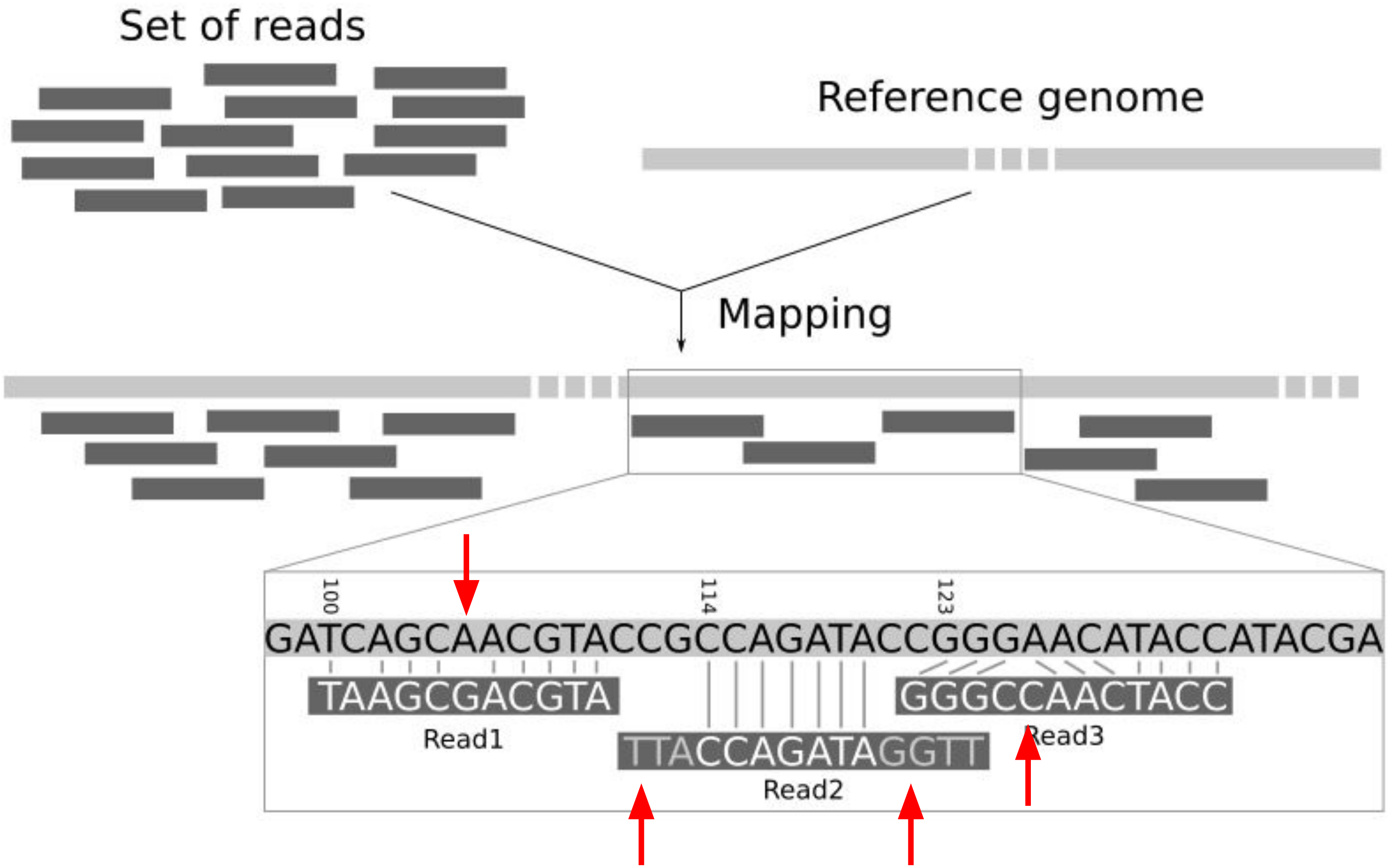
Part 2: Alignment and identification of binding sites

Mapping of short reads



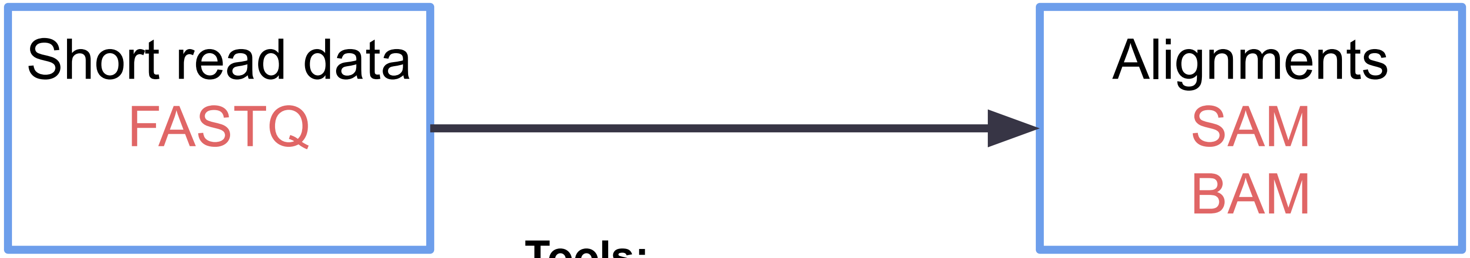
Galaxy training

Mapping of short reads



Galaxy training

Mapping of short reads



Short read data
FASTQ

Alignments
SAM
BAM

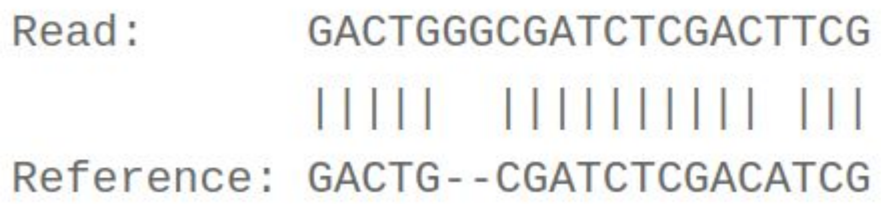
Tools:

-Bowtie
-Bowtie2
-BWA

Bowtie :
-Short reads (< 50) and no gapped-alignments
Bowtie2:
- Supports gapped alignment.
-For reads longer than about 50 bp Bowtie 2 is generally faster, more sensitive, and uses less memory than Bowtie 1
BWA:
-Very similar to Bowtie2 although slower

Bowtie2

Supports gapped alignment



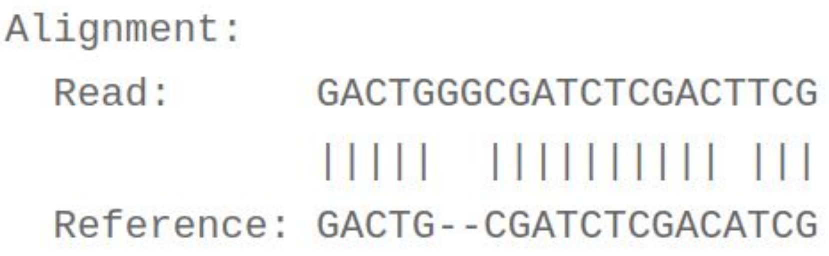
```
Read:      GACTGGGCGATCTCGACTTCG
           |||||  |||||||||| |||
Reference: GACTG--CGATCTCGACATCG
```

- Dash symbol represents a gap (insertion/deletion)
- Vertical bars represent matches

Bowtie2

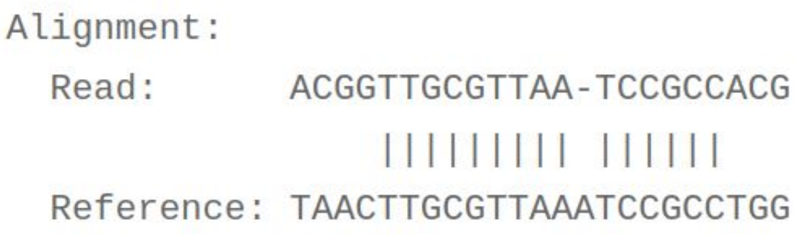
2 modes:

- End-to-end alignment (default mode): it searches for alignments involving all of the read characters.



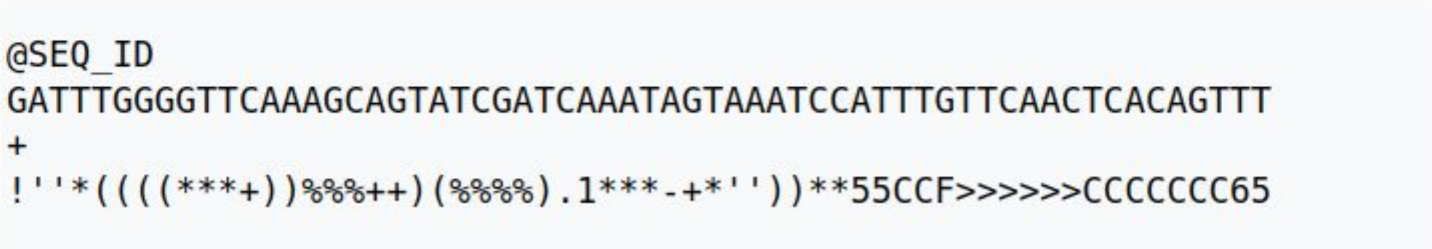
```
Alignment:
  Read:      GACTGGGCGATCTCGACTTCG
             |||||  |||||||||| |||
  Reference: GACTG--CGATCTCGACATCG
```

- Local alignment: some of the characters at the ends of the read do not participate (also known as "soft-trimming" or "soft-clipped")



```
Alignment:
  Read:      ACGGTTGCGTTAA-TCCGCCACG
                 ||||||||| ||||||
  Reference: TAACTTGCGTTAAATCCGCCTGG
```

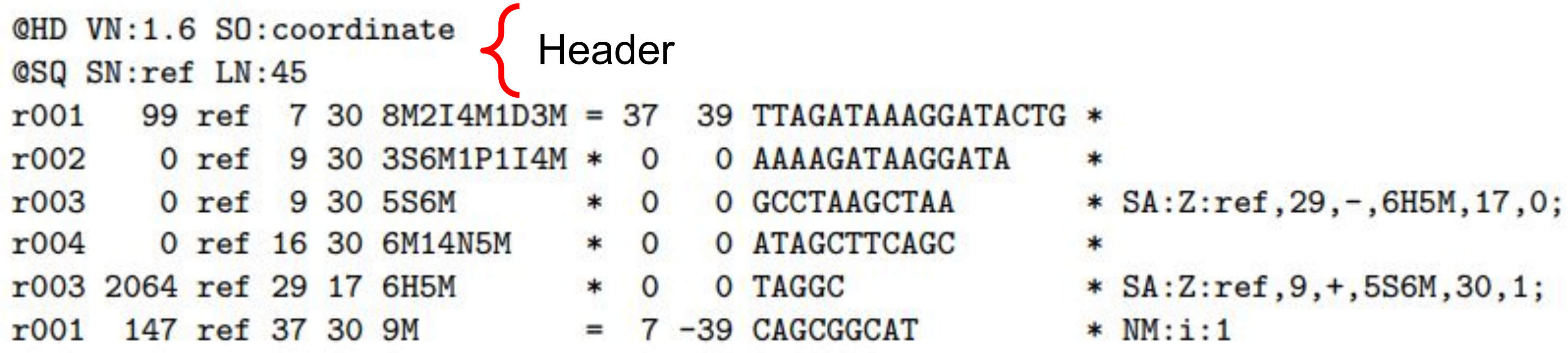
Quick review of the formats: FASTQ



```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

1. Sequence ID
2. Raw sequence
3. Begins with a ‘+’ character; optionally followed by sequence ID and/or other description
4. Quality values of the sequence

Quick review of the formats: SAM



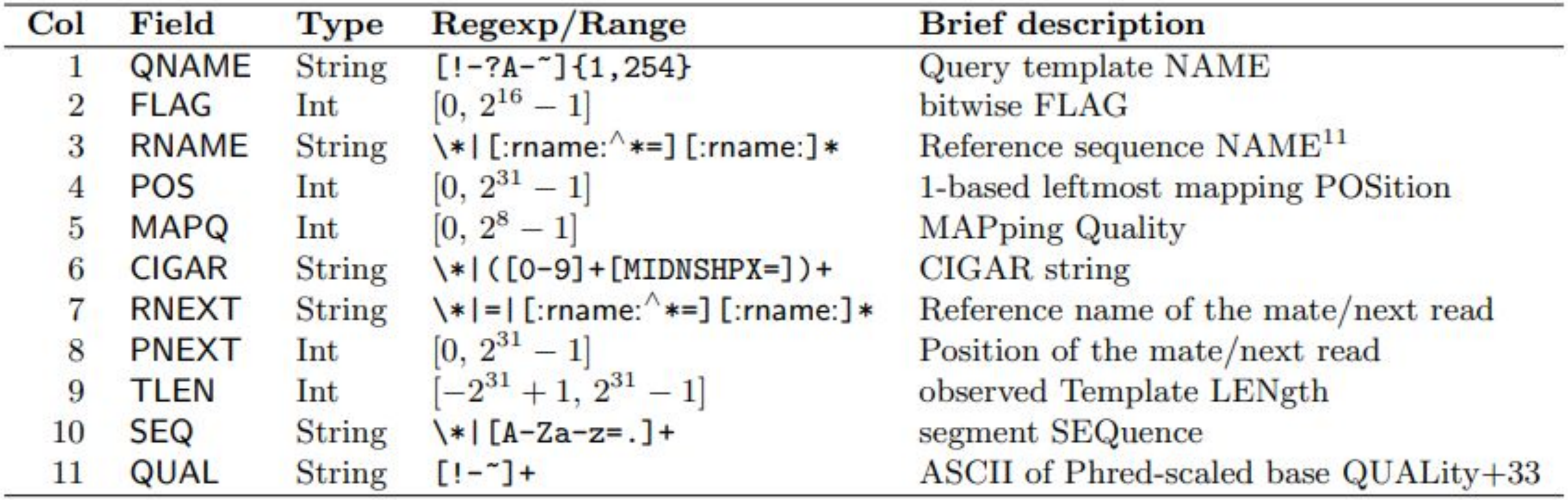
```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M * 0    0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       * 0    0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    * 0    0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       * 0    0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         = 7  -39 CAGCGGCAT         * NM:i:1
```

Header

SAM: Sequence Alignment Map
BAM: Binary (compressed) SAM

A great tool to work with SAM/BAM : **Samtools**

Quick review of the formats: SAM

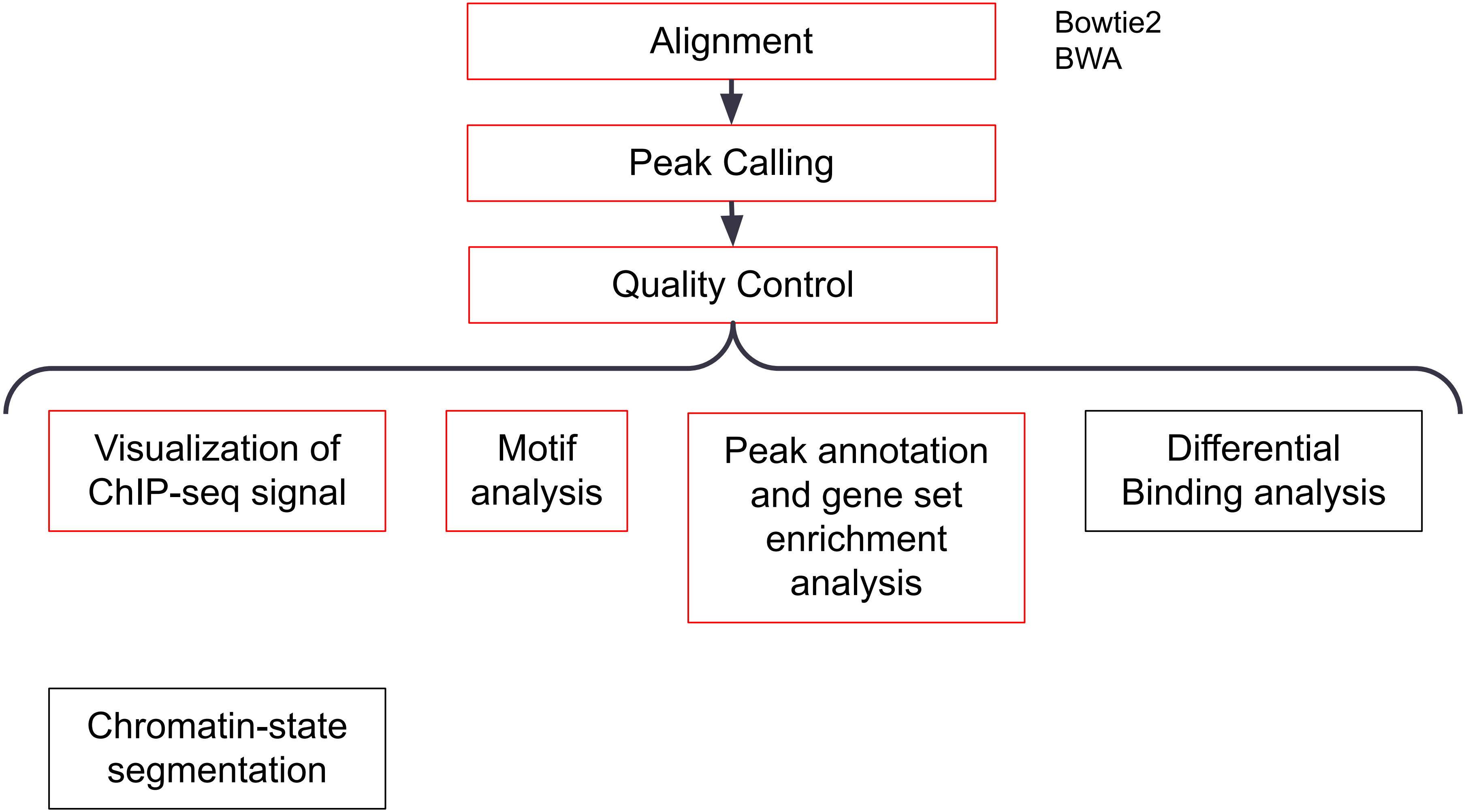


Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!-?A-~]{1,254}	Query template NAME
2	FLAG	Int	$[0, 2^{16} - 1]$	bitwise FLAG
3	RNAME	String	*|[:rname:^*=][:rname:]*	Reference sequence NAME[11]
4	POS	Int	$[0, 2^{31} - 1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0, 2^{8} - 1]$	MAPping Quality
6	CIGAR	String	*|([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	*|=|[:rname:^*=][:rname:]*	Reference name of the mate/next read
8	PNEXT	Int	$[0, 2^{31} - 1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31} + 1, 2^{31} - 1]$	observed Template LENgth
10	SEQ	String	*|[A-Za-z=.]+	segment SEQuence
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

Key fields:

- FLAG: Information about the alignment
- MAPQ: Mapping quality is related to "uniqueness” Higher == “more unique”

ChIP-seq analysis



Alignment

Peak Calling

Quality Control

Visualization of ChIP-seq signal

Motif analysis

Peak annotation and gene set enrichment analysis

Differential Binding analysis

Chromatin-state segmentation

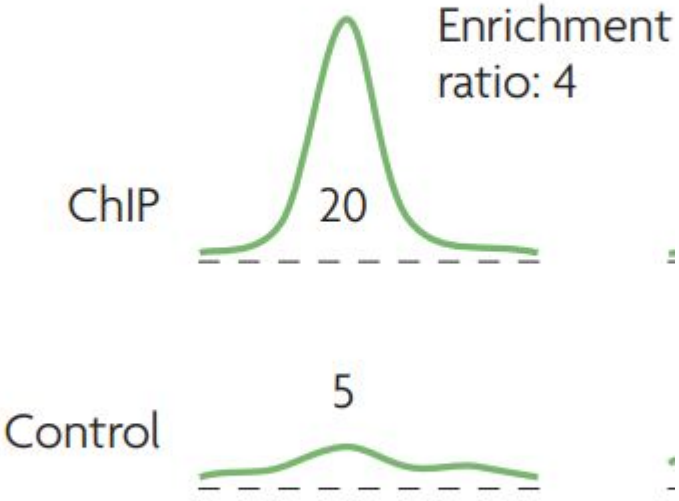
Bowtie2
BWA

Based on Santiago et al (2018)

Peak Calling

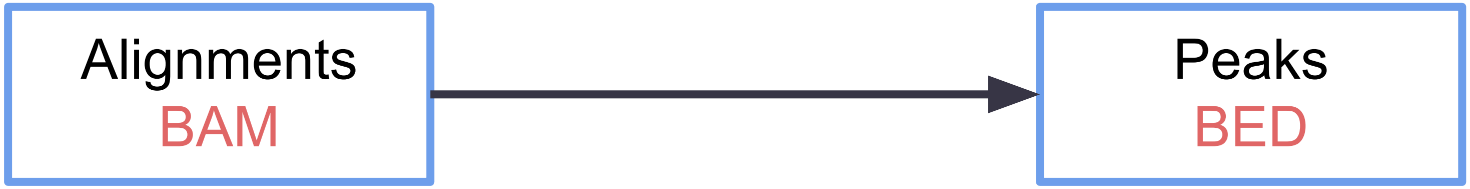
What is our goal?

Identify the regions of the genome where the ChIPed protein is bound by finding regions with significant numbers of mapped reads (compared to input control)



Park (2009)

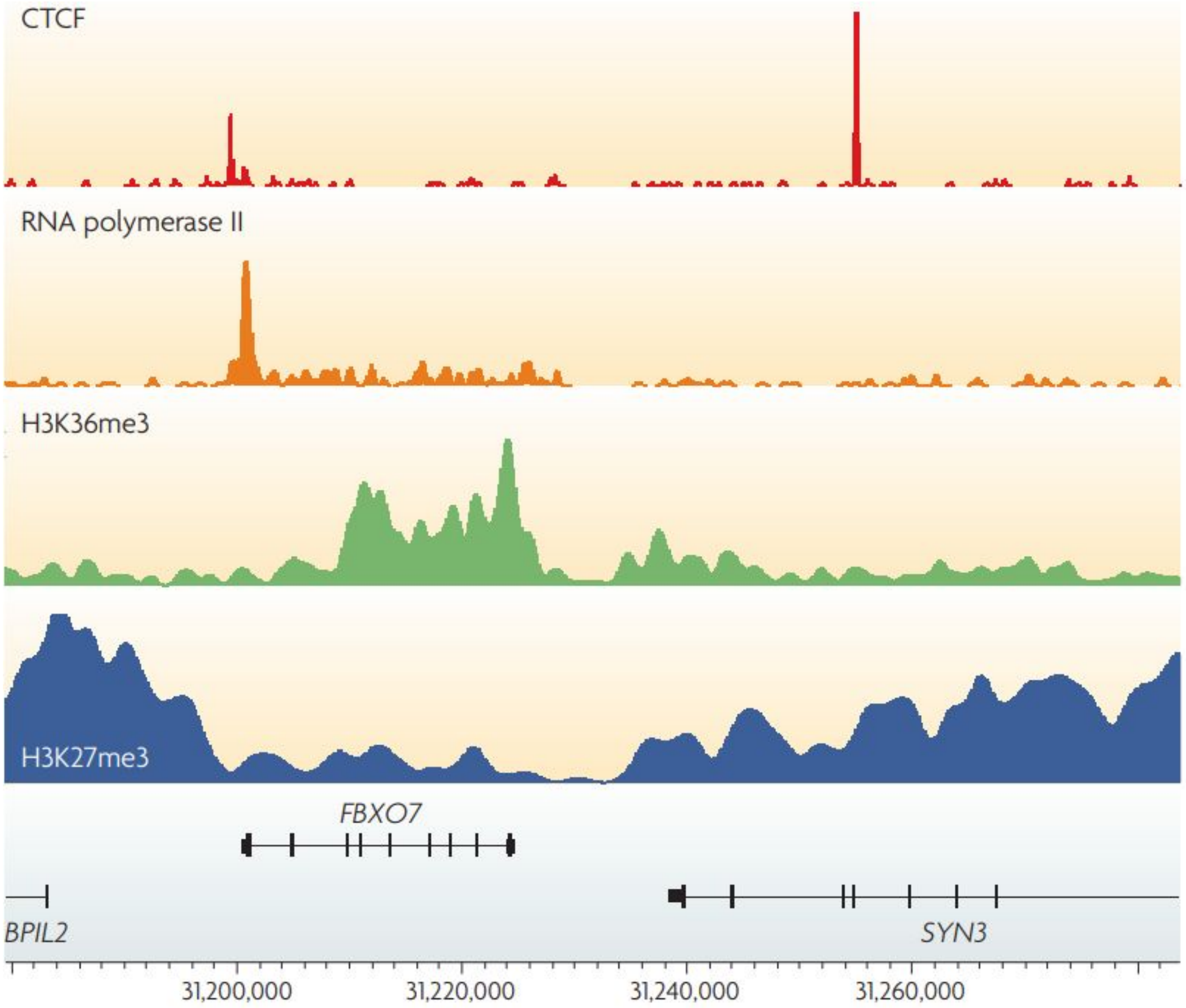
Peak Calling



Tools:

- -MACS2
- -SICER
- -SPP
- -HOMER
- -BroadPeak

Variability in ChIP-seq signals



Sharp

Mixed: Sharp + Broad

Broad signal enriched in transcribed region

Broad signal enriched in non-transcribed region

Park (2009)

McGill initiative in Computational Medicine

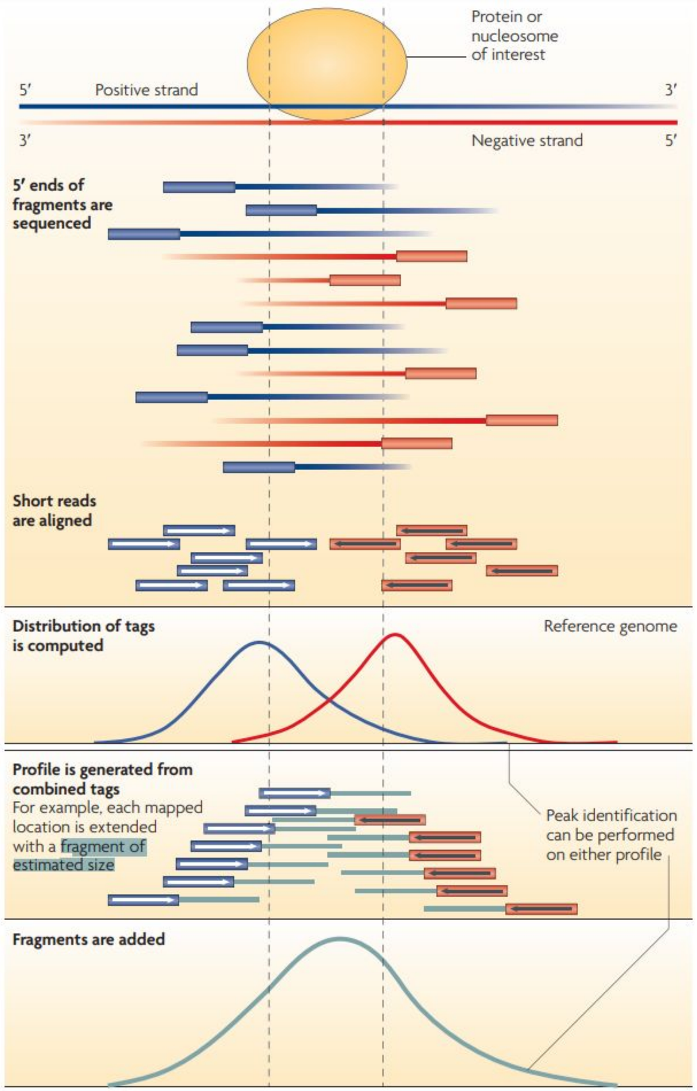
MACS2

Model-based analysis of ChIP-seq

1. Estimate fragment length
2. Compare coverage against input control

Fragment size is estimated in single end data with MACS2

What happens in paired-end data?



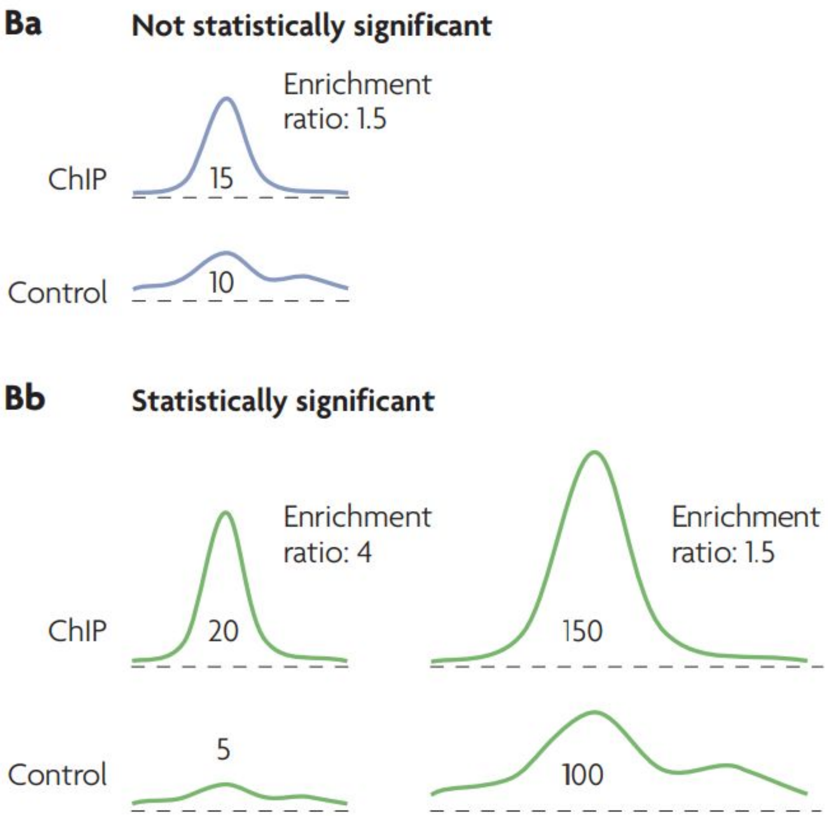
Protein or nucleosome of interest
5′
Positive strand
3′
3′
Negative strand
5′
5′ ends of fragments are sequenced
Short reads are aligned
Distribution of tags is computed
Reference genome
Profile is generated from combined tags
For example, each mapped location is extended with a fragment of estimated size
Peak identification can be performed on either profile
Fragments are added

Park (2009)

MACS2

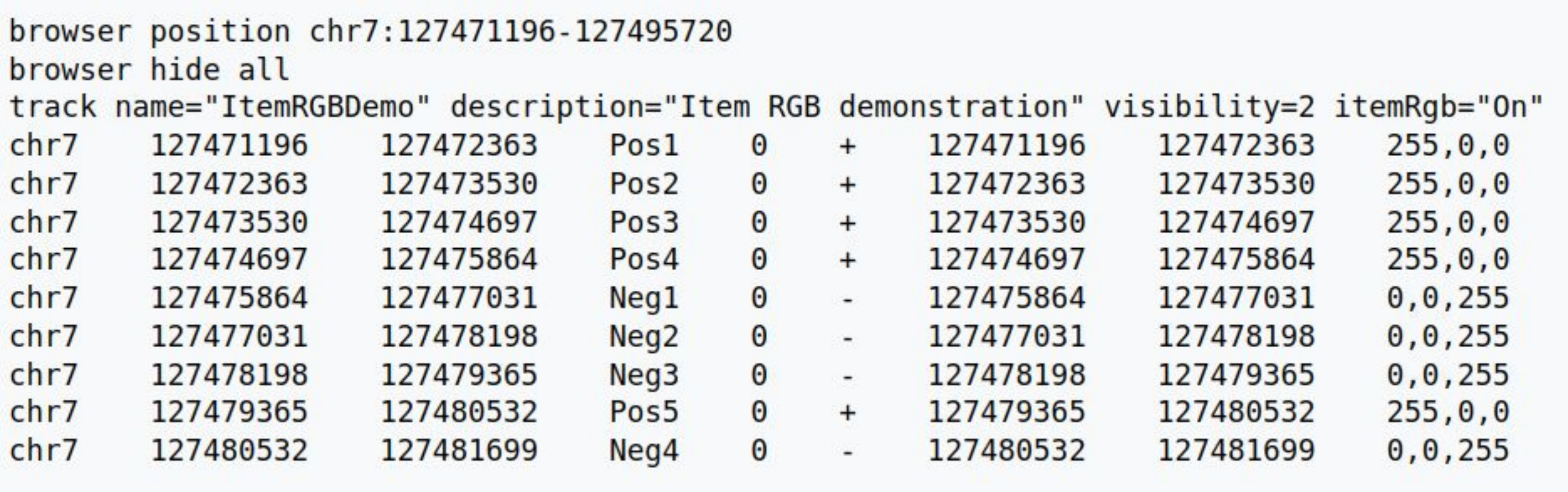
Model-based analysis of ChIP-seq

MACS2 models the tag distribution using a Poisson Model



Accounts for the ratio as well as the absolute tag numbers

Quick review of the formats: BED



```
browser position chr7:127471196-127495720
browser hide all
track name="ItemRGBDemo" description="Item RGB demonstration" visibility=2 itemRgb="On"
chr7    127471196    127472363    Pos1    0    +    127471196    127472363    255,0,0
chr7    127472363    127473530    Pos2    0    +    127472363    127473530    255,0,0
chr7    127473530    127474697    Pos3    0    +    127473530    127474697    255,0,0
chr7    127474697    127475864    Pos4    0    +    127474697    127475864    255,0,0
chr7    127475864    127477031    Neg1    0    -    127475864    127477031    0,0,255
chr7    127477031    127478198    Neg2    0    -    127477031    127478198    0,0,255
chr7    127478198    127479365    Neg3    0    -    127478198    127479365    0,0,255
chr7    127479365    127480532    Pos5    0    +    127479365    127480532    255,0,0
chr7    127480532    127481699    Neg4    0    -    127480532    127481699    0,0,255
```

3 required fields:

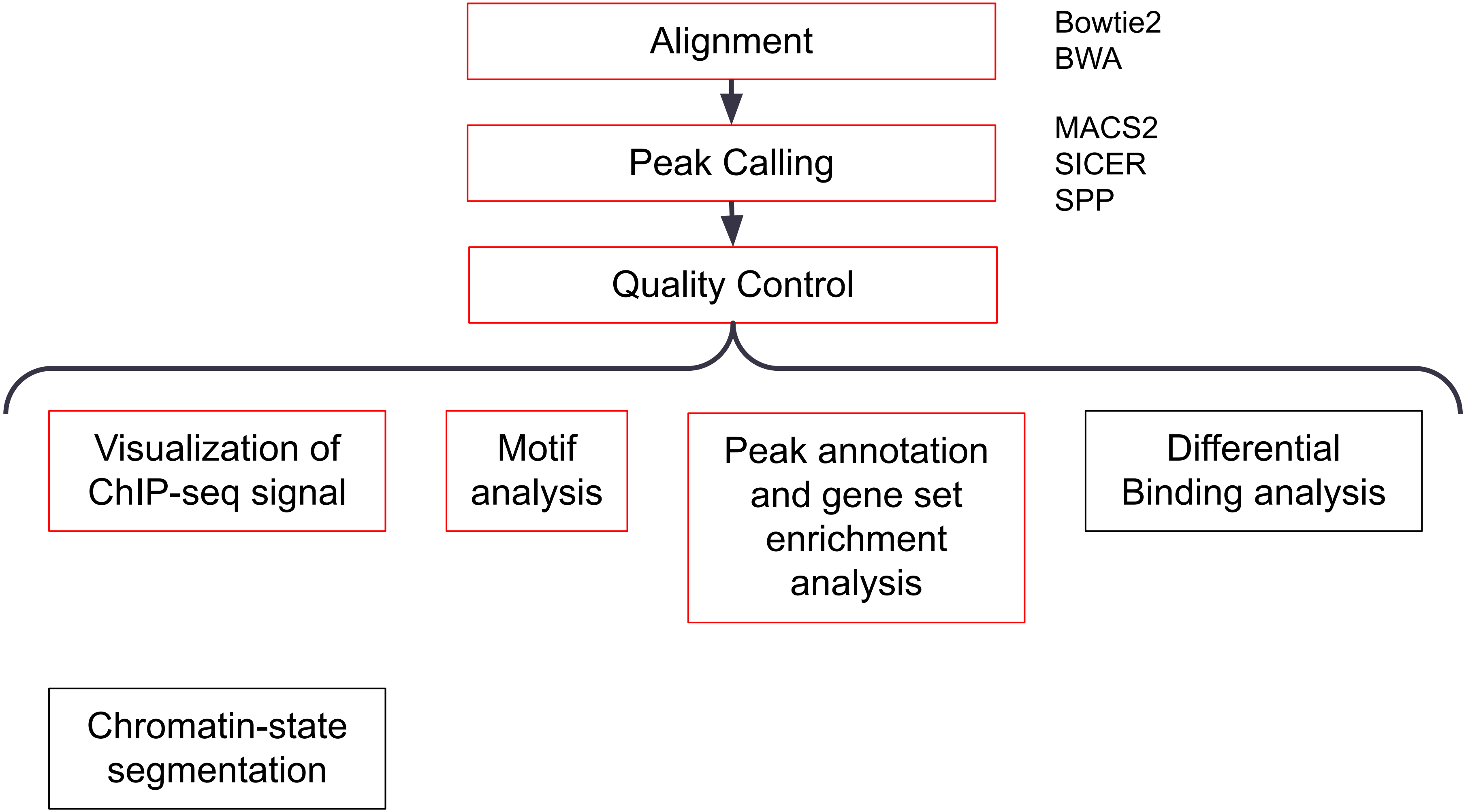
- Chromosome
- Start
- End

9 optional fields:

- Name
- Score
- Strand
- ThickStart
- ThickEnd
- itemRGB
- blockCount
- blockSizes
- blockStarts

A great tool to work with BED: **Bedtools**

ChIP-seq analysis

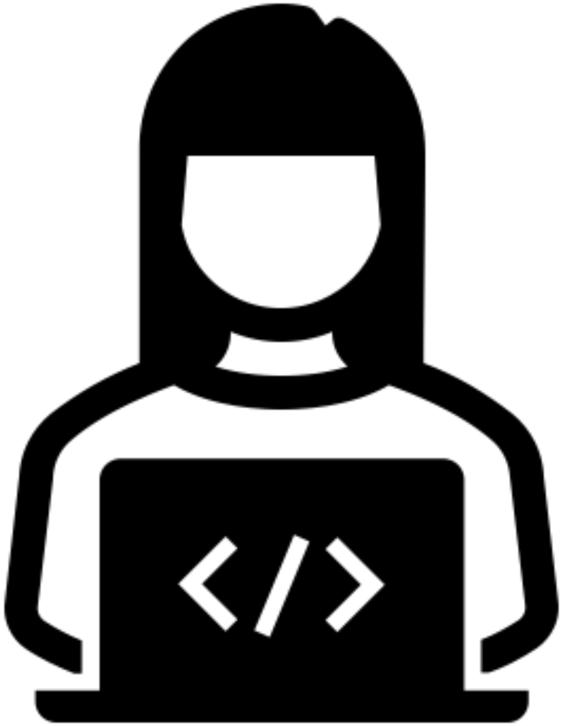

Alignment
Peak Calling
Quality Control
Visualization of ChIP-seq signal
Motif analysis
Peak annotation and gene set enrichment analysis
Differential Binding analysis
Chromatin-state segmentation

Bowtie2
BWA

MACS2
SICER
SPP

Based on Santiago et al (2018)

Hands-on 1



Part 3: Quality control

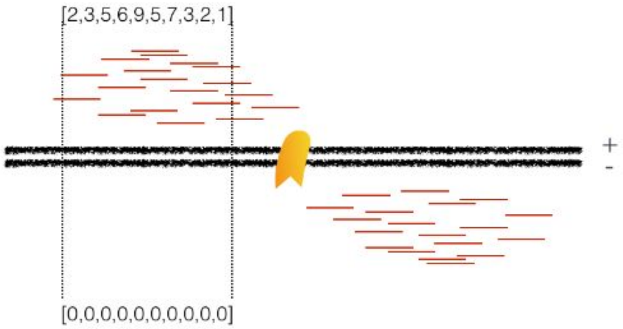
ChIP-seq analysis

Various QC metrics exist:

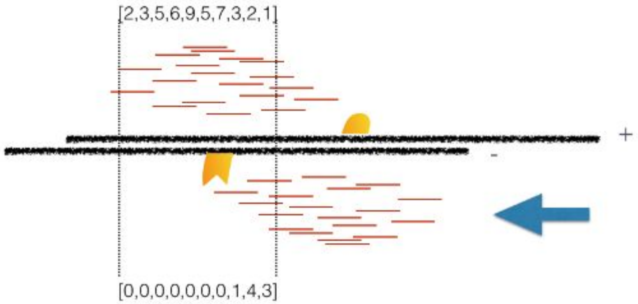
- Cross-correlation
- FRiP (Fraction of reads in peaks)
- Non redundant Fraction (NRF)
- IDR (Irreproducibility Discovery Rate)
- Fingerprint plots
- PBC1 and PBC2

Cross-correlation

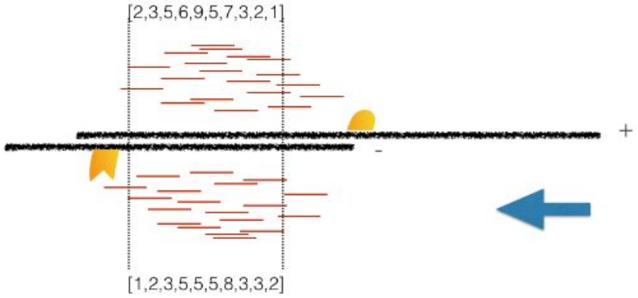
***Plot 1:** At strand shift of zero, the Pearson correlation between the two vectors is 0.*



***Plot 2:** At strand shift of 100bp, the Pearson correlation between the two vectors is 0.389.*

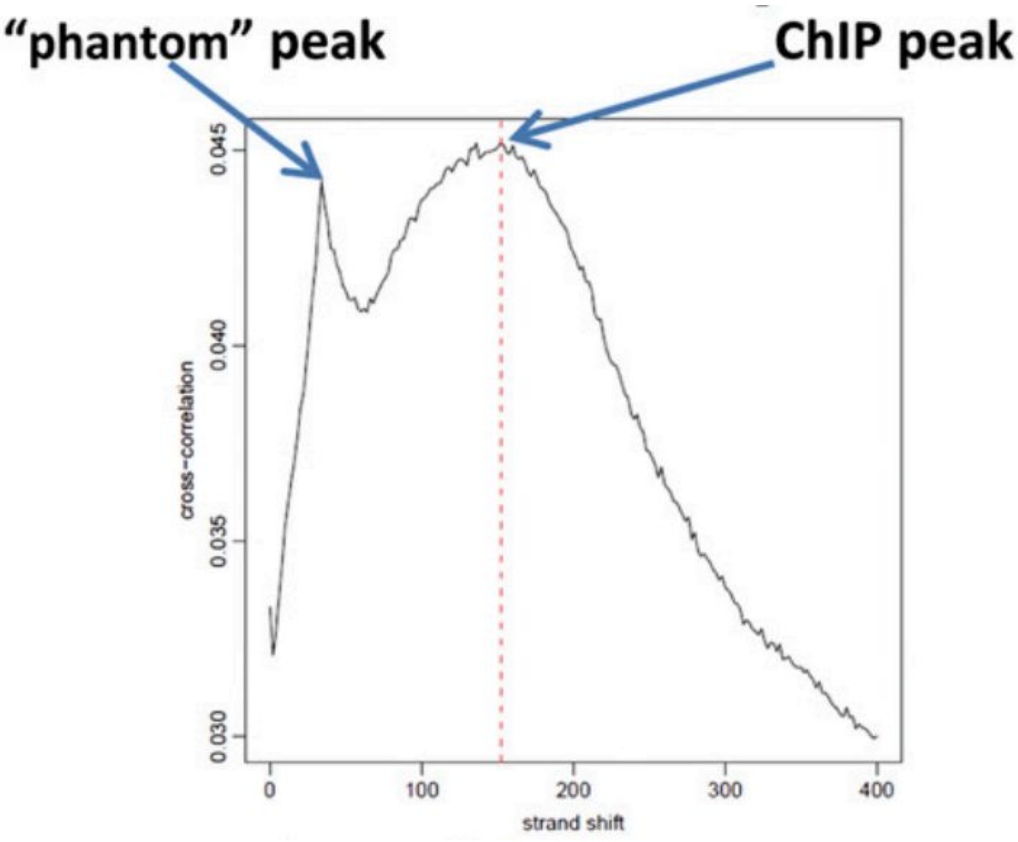


***Plot 3:** At strand shift of 175bp, the Pearson correlation between the two vectors is 0.831.*



HBC training (Online)

Cross-correlation

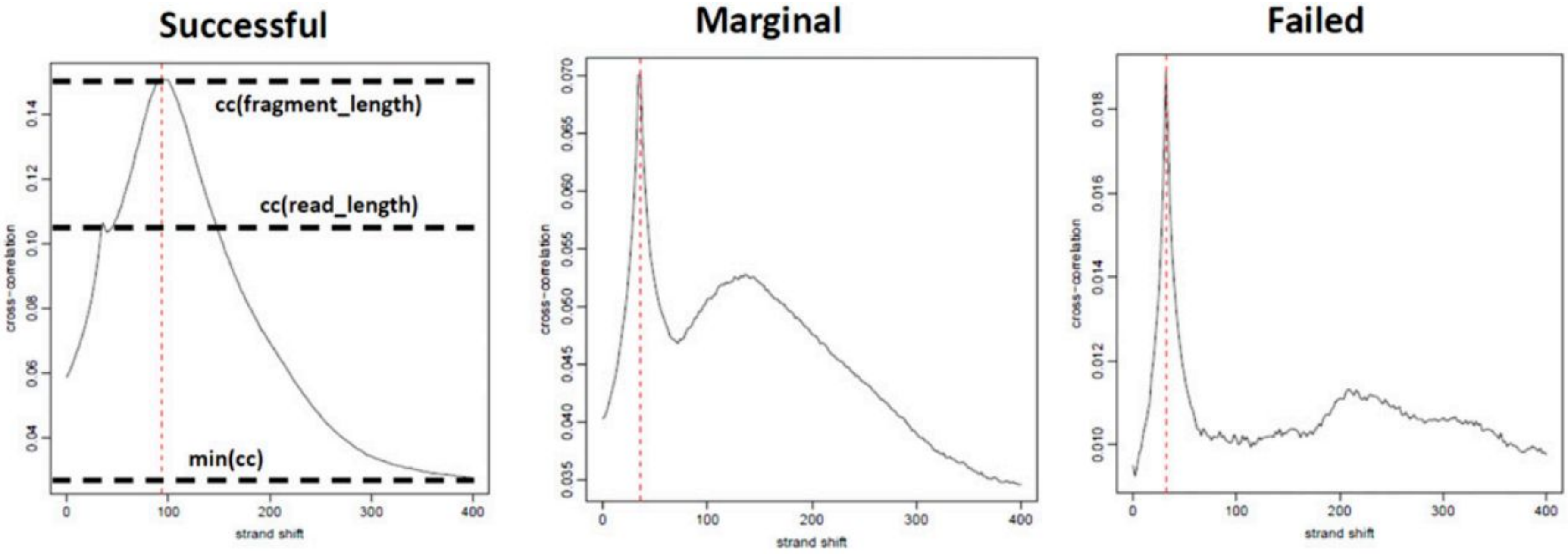


Phantom peak corresponds to the read length

ChIP peak corresponds to the predominant fragment length

Landt *et al* (2012)

Cross-correlation



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

RSC > 1 represents high quality

Landt *et al* (2012)

FRiP

Fraction of all reads mapped that fall in peaks

In general, samples with a FRiP higher than 1% represent good quality, however…

Some limitations:

- Some DNA binding proteins have very few true binding sites (ZNF274 & RNA pol III)
- Dependent on antibody

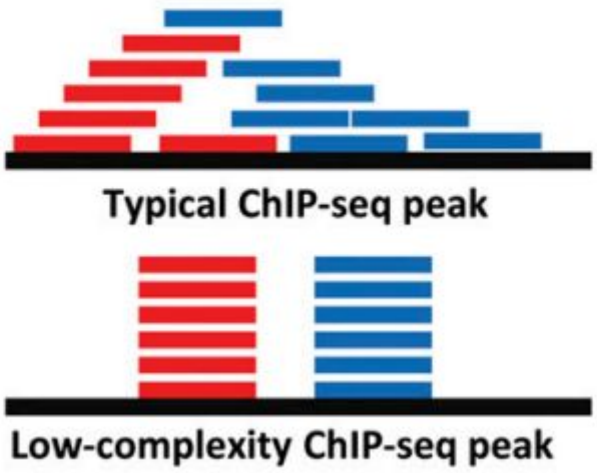
It is still a useful metric to:

- Compare results obtained with the same antibody across cell lines
- Compare different antibodies against the same factor

NRF (Non-redundant Fraction)

Number of distinct uniquely mapping reads (i.e. after removing duplicates) / Total number of mapped reads

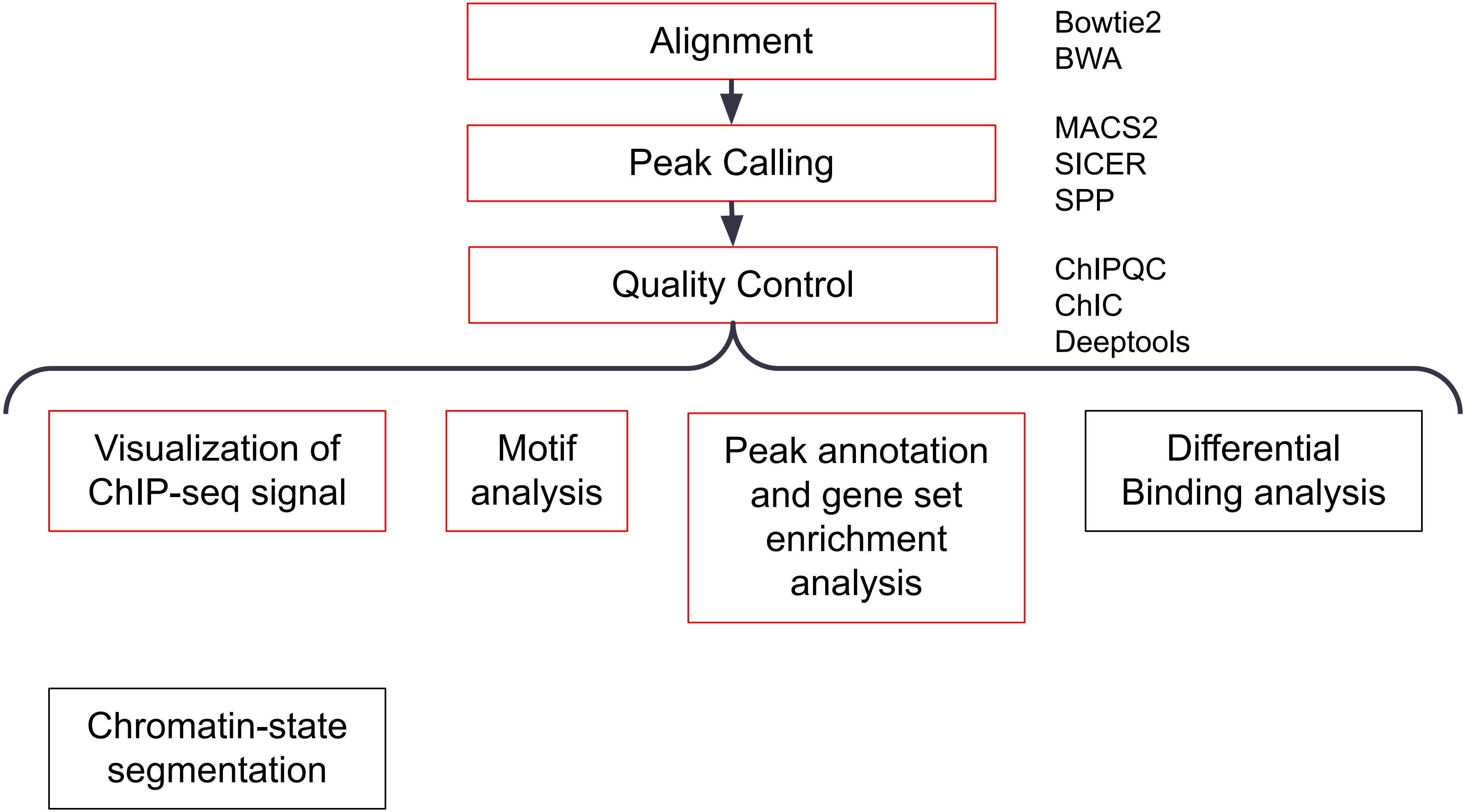
Typically good values are NRF > 0.9 according to the ENCODE standards



Landt *et al* (2012)

McGill initiative in Computational Medicine

ChIP-seq analysis



Alignment

Bowtie2
BWA

Peak Calling

MACS2
SICER
SPP

Quality Control

ChIPQC
ChIC
Deeptools

Visualization of ChIP-seq signal

Motif analysis

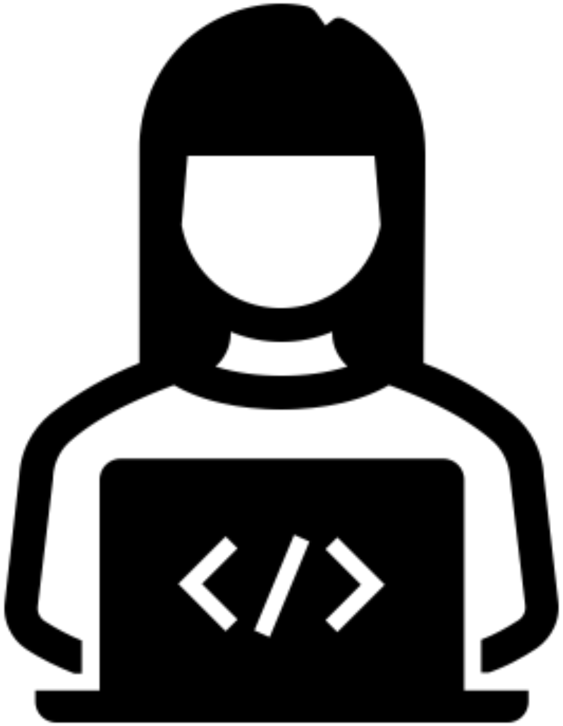
Peak annotation and gene set enrichment analysis

Differential Binding analysis

Chromatin-state segmentation

Based on Santiago et al (2018)

Hands-on 2



Part 4: Visualization

Options for Visualization

Coverage visualization:

- UCSC genome browser
- WashU Epigenome browser
- IGV

Heatmaps/Density plots:

- Deeptools

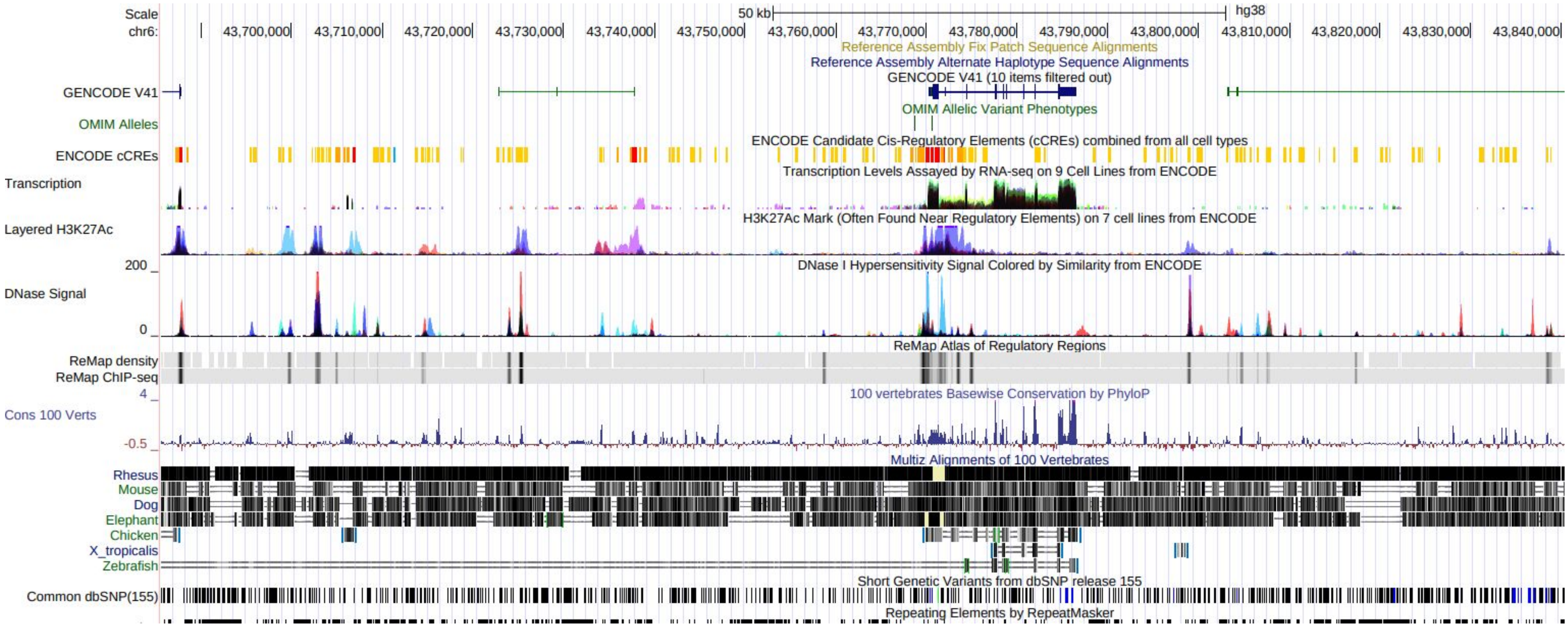
Most of the times, we use .bigwig files as input for visualization

BAM —---> bigWig

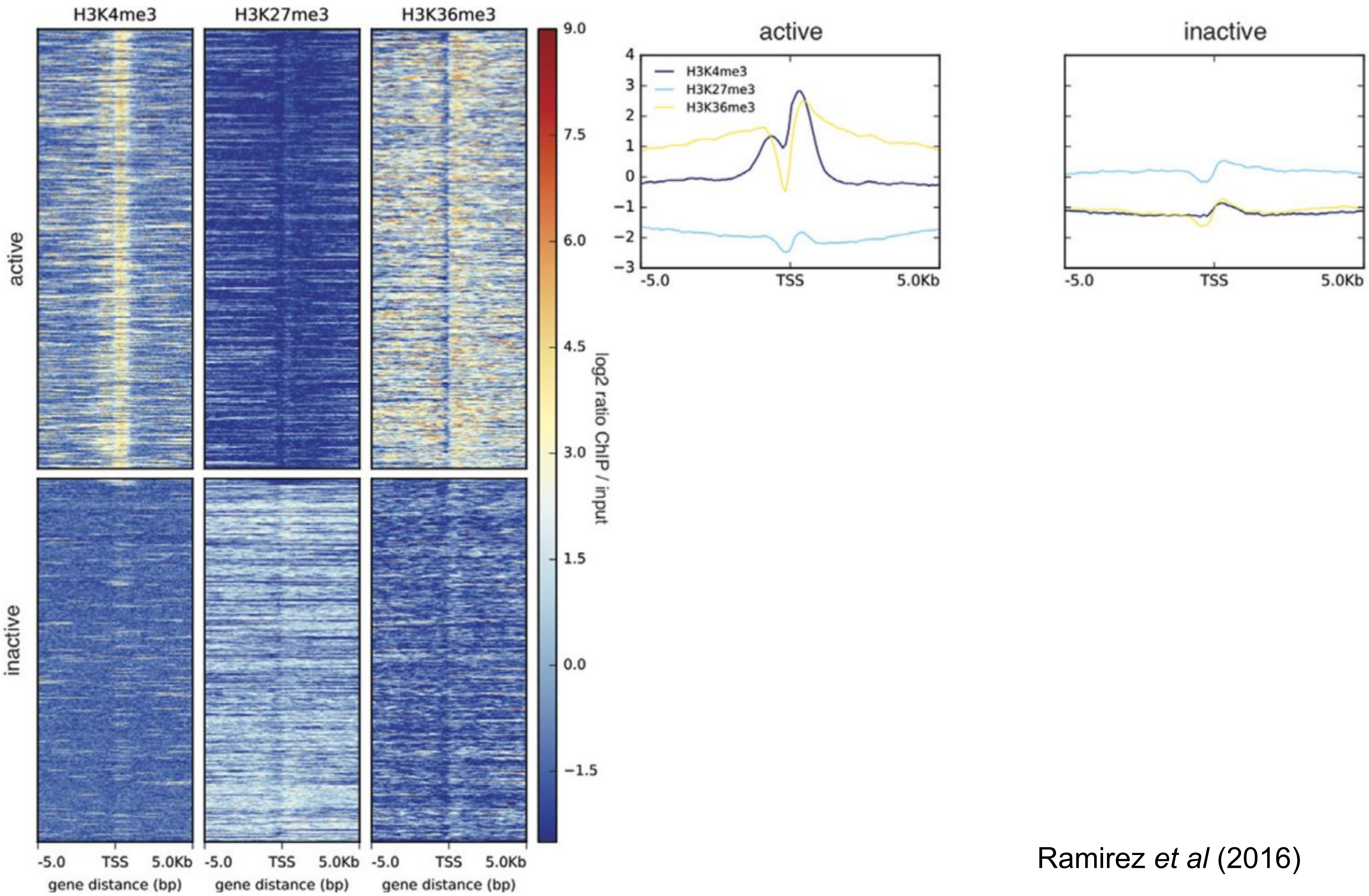
The bigWig format is for display of dense, continuous data that will be displayed as a graph.

UCSC genome browser

https://genome.ucsc.edu/

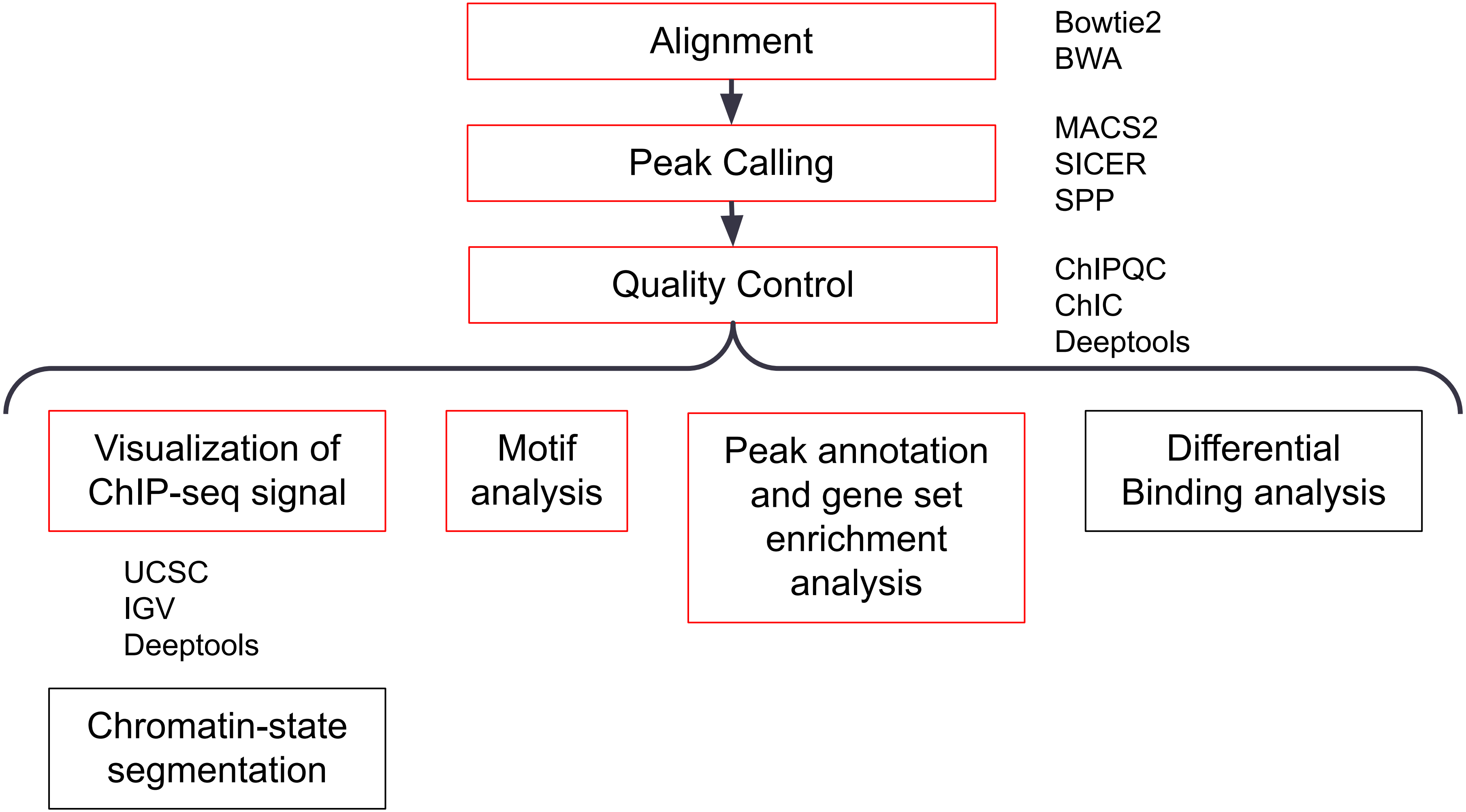

Scale
chr6:
50 kb
hg38
43,700,000
43,710,000
43,720,000
43,730,000
43,740,000
43,750,000
43,760,000
43,770,000
43,780,000
43,790,000
43,800,000
43,810,000
43,820,000
43,830,000
43,840,000
Reference Assembly Fix Patch Sequence Alignments
Reference Assembly Alternate Haplotype Sequence Alignments
GENCODE V41 (10 items filtered out)
GENCODE V41
OMIM Allelic Variant Phenotypes
OMIM Alleles
ENCODE Candidate Cis-Regulatory Elements (cCREs) combined from all cell types
ENCODE cCREs
Transcription Levels Assayed by RNA-seq on 9 Cell Lines from ENCODE
Transcription
H3K27Ac Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
200
DNase I Hypersensitivity Signal Colored by Similarity from ENCODE
DNase Signal
0
ReMap Atlas of Regulatory Regions
ReMap density
ReMap ChIP-seq
4
100 vertebrates Basewise Conservation by PhyloP
Cons 100 Verts
-0.5
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Short Genetic Variants from dbSNP release 155
Common dbSNP(155)
Repeating Elements by RepeatMasker

Deeptools



Ramirez *et al* (2016)

ChIP-seq analysis



Alignment

Bowtie2
BWA

Peak Calling

MACS2
SICER
SPP

Quality Control

ChIPQC
ChIC
Deeptools

Visualization of ChIP-seq signal

UCSC
IGV
Deeptools

Chromatin-state segmentation

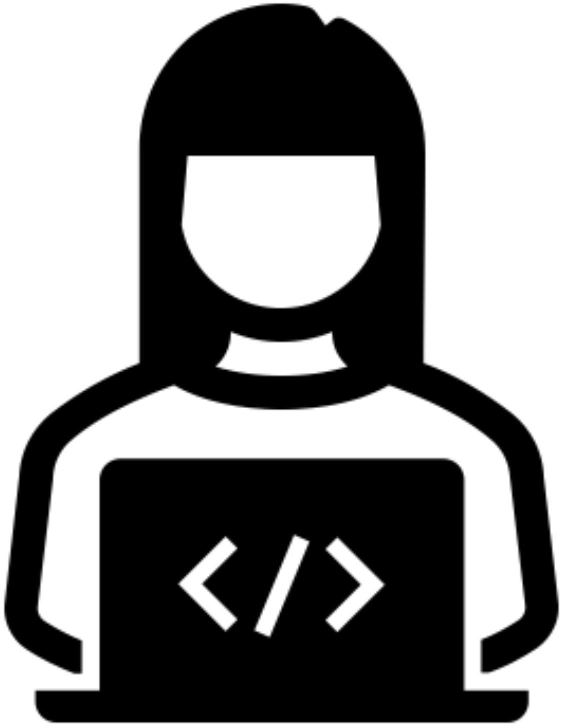
Motif analysis

Peak annotation and gene set enrichment analysis

Differential Binding analysis

Based on Santiago et al (2018)

Hands-on 3



Part 5: Motif finding and gene set enrichment analysis

Available tools

Motif analysis:

- MEME
- HOMER
- JASPAR
- Pscan-ChIP
- RSAT

Gene set enrichment analysis

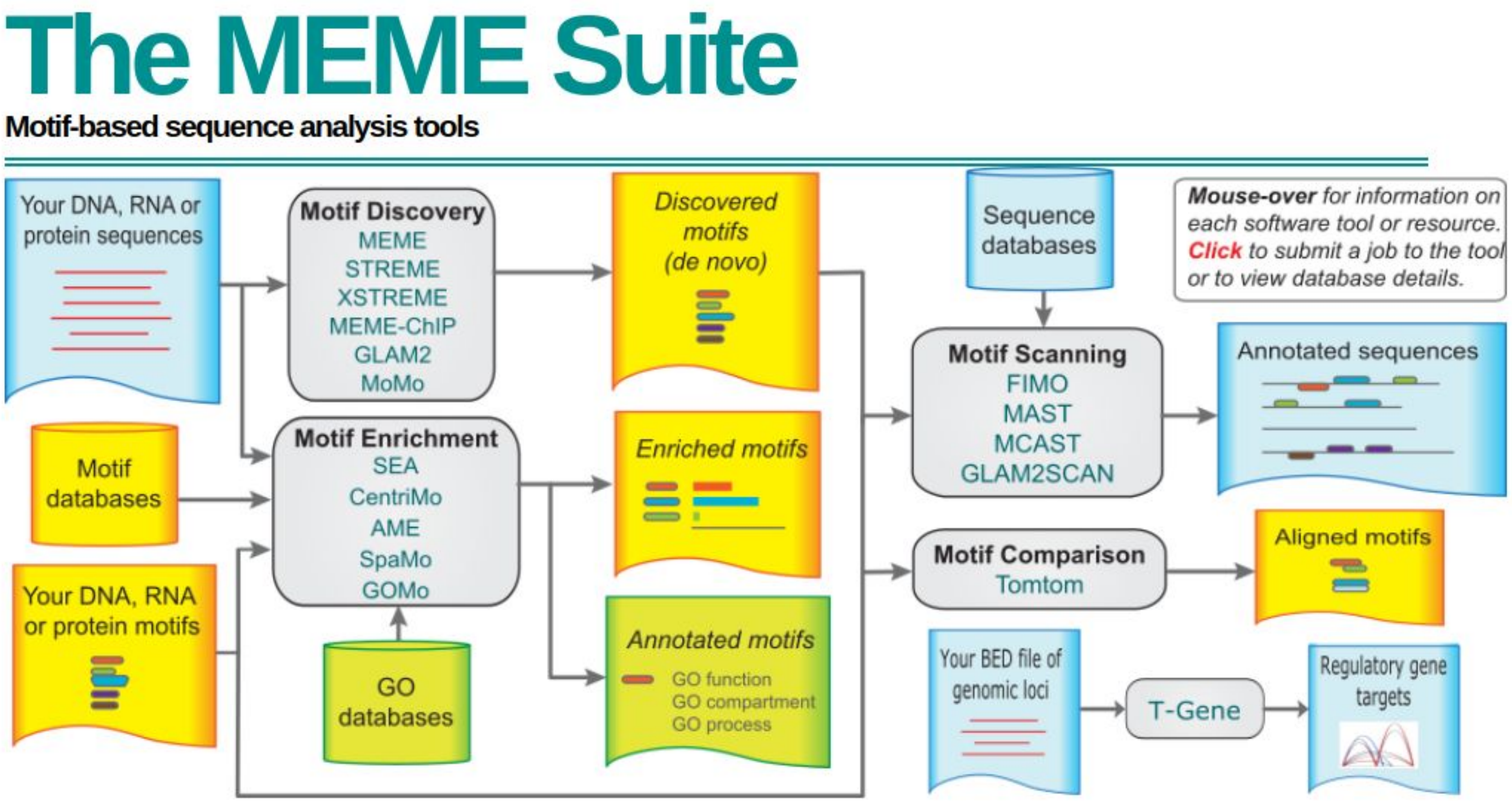
- GREAT
- ChIP Enrich
- Broad Enrich

And many more…

Ramirez *et al* (2016)

MEME

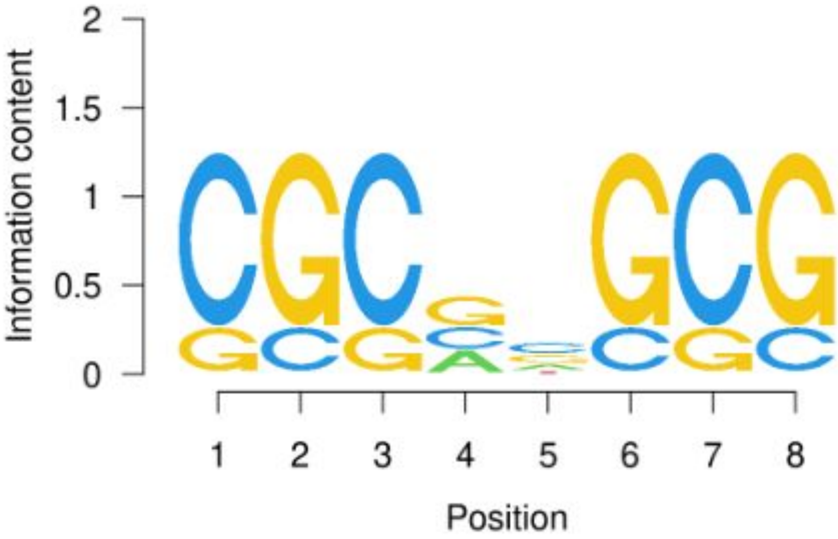
Online: https://meme-suite.org/meme/
Terminal and as a R package: `BiocManager::install("memes")`



MEME

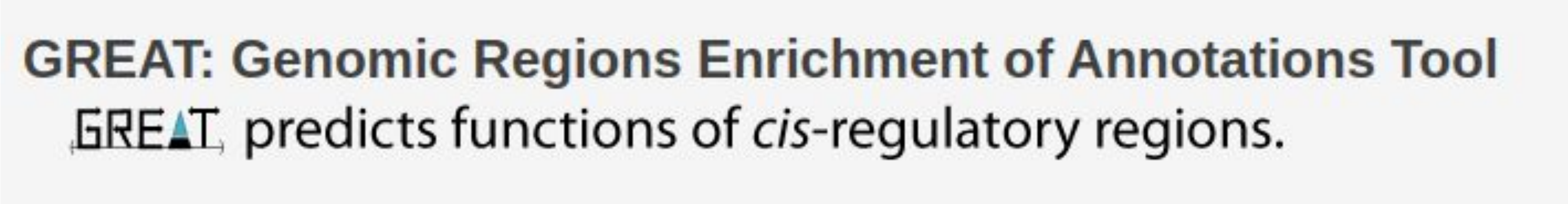
DNA logo

a **sequence logo** is a graphical representation of the sequence conservation of nucleotides



The overall height of the stack is proportional to the information content at that position, while the height of symbols within the stack indicates the relative frequency of each amino or nucleic acid at that position.

GREAT



GREAT: Genomic Regions Enrichment of Annotations Tool

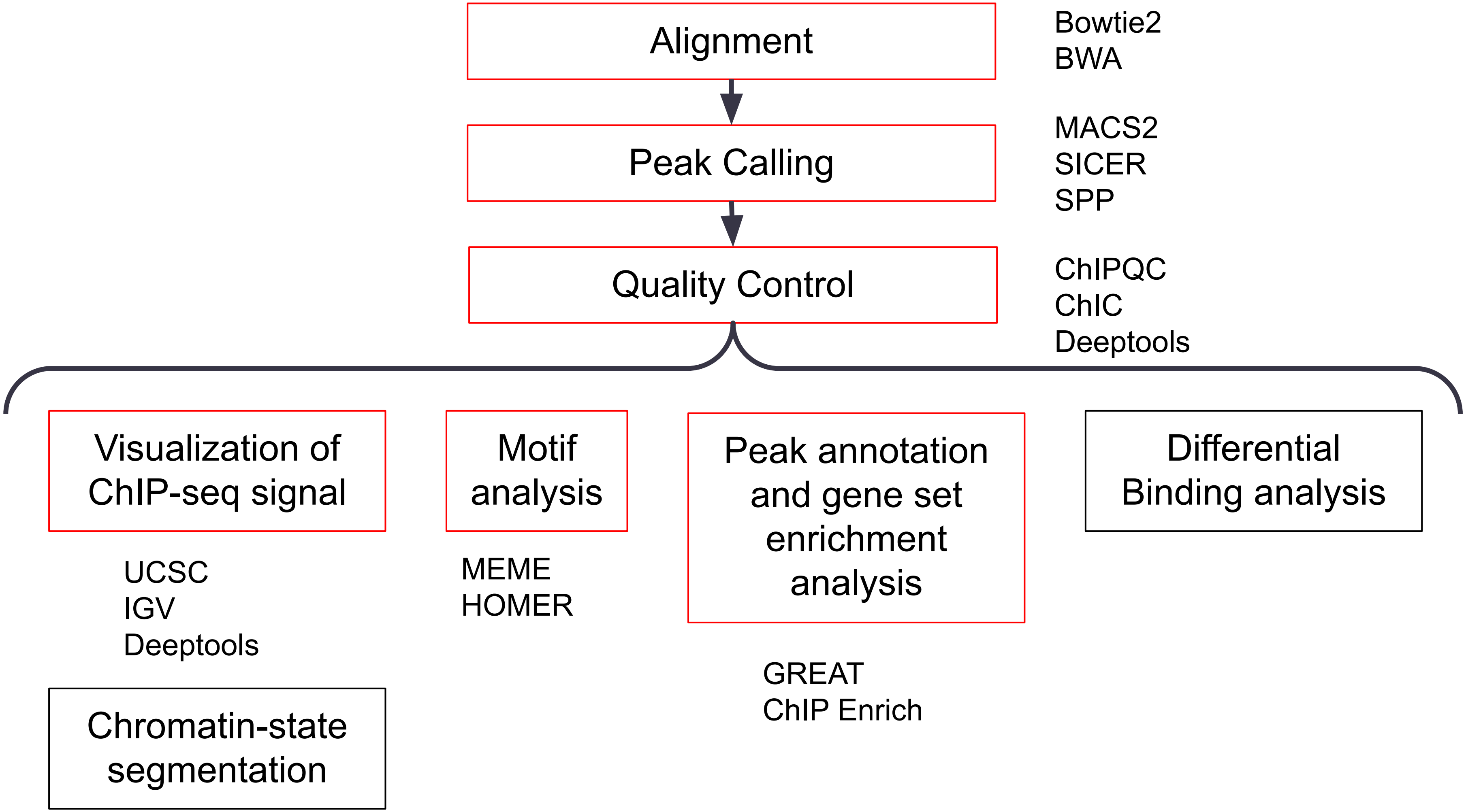
GREAT predicts functions of *cis*-regulatory regions.

Predicts biological functions of cis-regulatory regions:

- Connect your ChIP-seq peaks to genes
- Pathway/GO analysis (accounts for the fraction of the genome involved for a given pathway)

McLean *et al* (2010)

ChIP-seq analysis



Alignment

Bowtie2
BWA

Peak Calling

MACS2
SICER
SPP

Quality Control

ChIPQC
ChIC
Deeptools

Visualization of ChIP-seq signal

UCSC
IGV
Deeptools

Motif analysis

MEME
HOMER

Peak annotation and gene set enrichment analysis

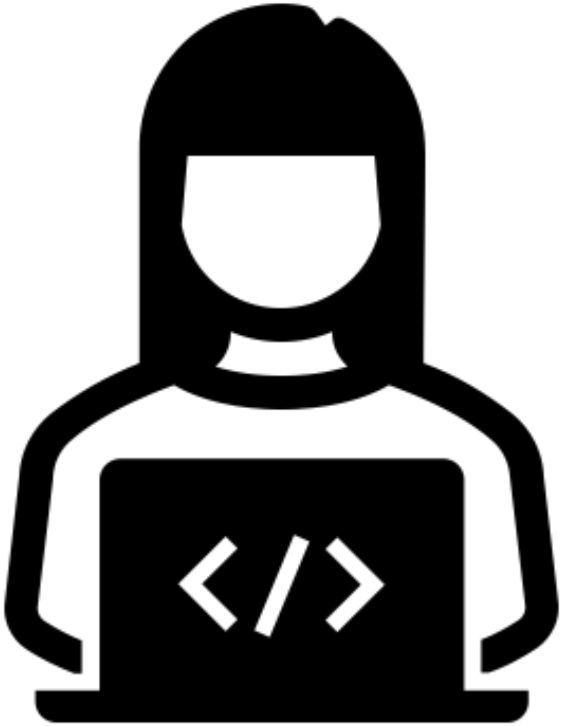
GREAT
ChIP Enrich

Differential Binding analysis

Chromatin-state segmentation

Based on Santiago et al (2018)

Hands-on 4



Part 6: Concluding remarks

ChIP-seq resources

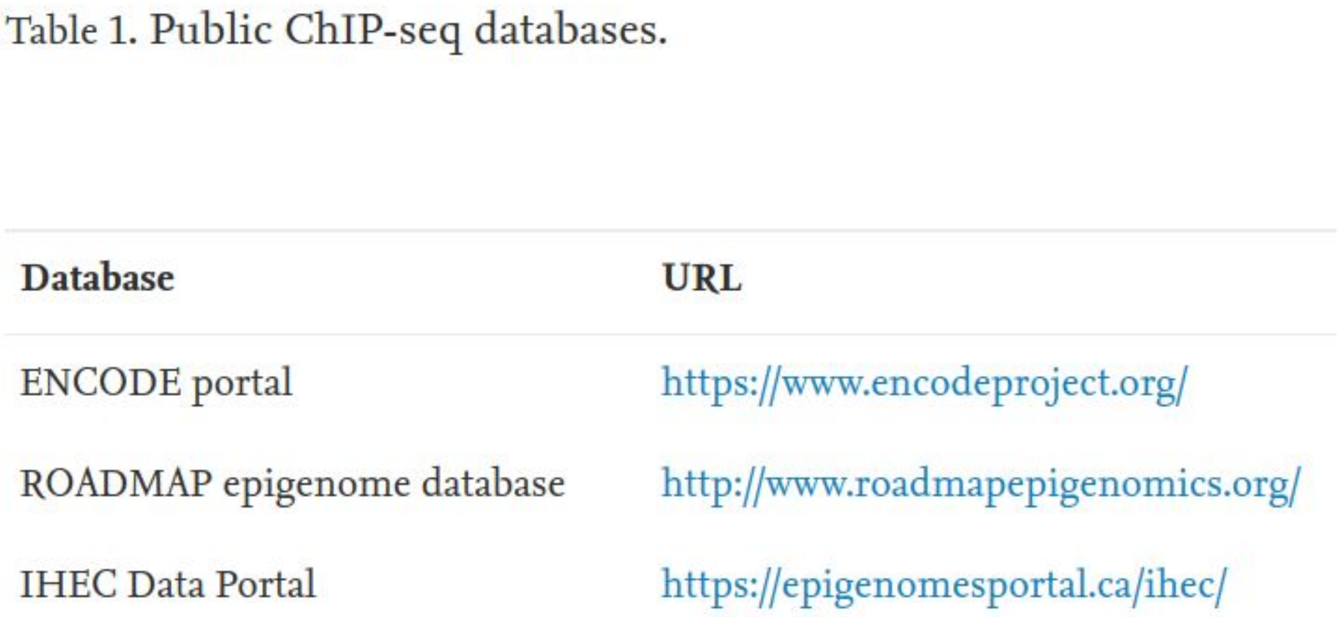


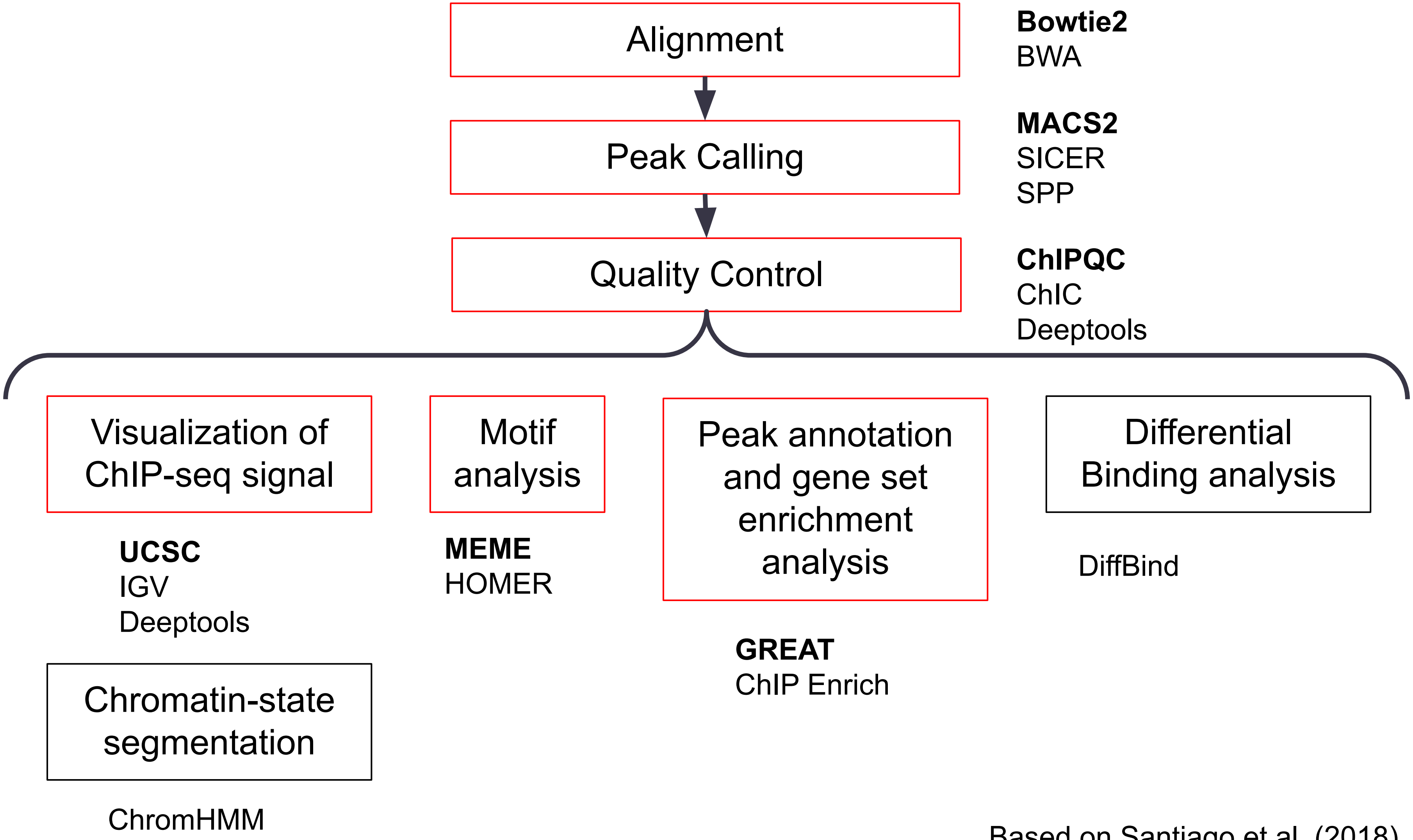
Table 1. Public ChIP-seq databases.

Database	URL
ENCODE portal	https://www.encodeproject.org/
ROADMAP epigenome database	http://www.roadmapepigenomics.org/
IHEC Data Portal	https://epigenomesportal.ca/ihec/

A lot of data is available!

Nakato (2021)

What have we learned?



Alignment

Bowtie2
BWA

Peak Calling

MACS2
SICER
SPP

Quality Control

ChIPQC
ChIC
Deeptools

Visualization of ChIP-seq signal

UCSC
IGV
Deeptools

Motif analysis

MEME
HOMER

Peak annotation and gene set enrichment analysis

GREAT
ChIP Enrich

Differential Binding analysis

DiffBind

Chromatin-state segmentation

ChromHMM

Based on Santiago et al (2018)

Thanks for your attention!

MiCM team:

- MiCM Student Society
- Prof. Guillaume Bourque
- Prof. Celia Greenwood

McGill UNIVERSITY

Keep an eye for the workshops offered by the MiCM!

info-micm@mcgill.ca
https://www.mcgill.ca/micm/

References:

Annunziato, A. (2008). DNA packaging: nucleosomes and chromatin. *Nature Education*, *1*(1), 26.

Klemm, S. L., Shipony, Z., & Greenleaf, W. J. (2019). Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, *20*(4), 207-220.

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., ... & Snyder, M. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome research*, *22*(9), 1813-1831.

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., ... & Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology*, *28*(5), 495-501.

Mehrmohamadi, Mahya, et al. "A comparative overview of epigenomic profiling methods." *Frontiers in Cell and Developmental Biology* (2021): 1990.

Meyer, C. A., & Liu, X. S. (2014). Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nature Reviews Genetics*, *15*(11), 709-721.

Mobley, A. S. (2019). *Neural stem cells and adult neurogenesis*. Academic Press.

Nakato, R., & Sakata, T. (2021). Methods for ChIP-seq analysis: a practical workflow and advanced applications. *Methods*, *187*, 44-53.

Park, P. J. (2009). ChIP–seq: advantages and challenges of a maturing technology. *Nature reviews genetics*, *10*(10), 669-680.

Santiago, I. D., & Carroll, T. (2018). Analysis of ChIP-seq data in R/Bioconductor. In *Chromatin Immunoprecipitation* (pp. 195-226). Humana Press, New York, NY.

Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., ... & Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic acids research*, *44*(W1), W160-W165.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., ... & Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome biology*, *9*(9), 1-9.

Online resources:

https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/mapping/tutorial.html
https://hbctraining.github.io/Intro-to-ChIPseq/lessons/06_combine_chipQC_and_metrics.html
https://physiology.med.cornell.edu/faculty/skrabanek/lab/angsd/schedule_2020/