Epigenomics assays: ChIP-seq and ATAC-seq

Analysis of Next-Generation Sequencing Data

Friederike Dündar

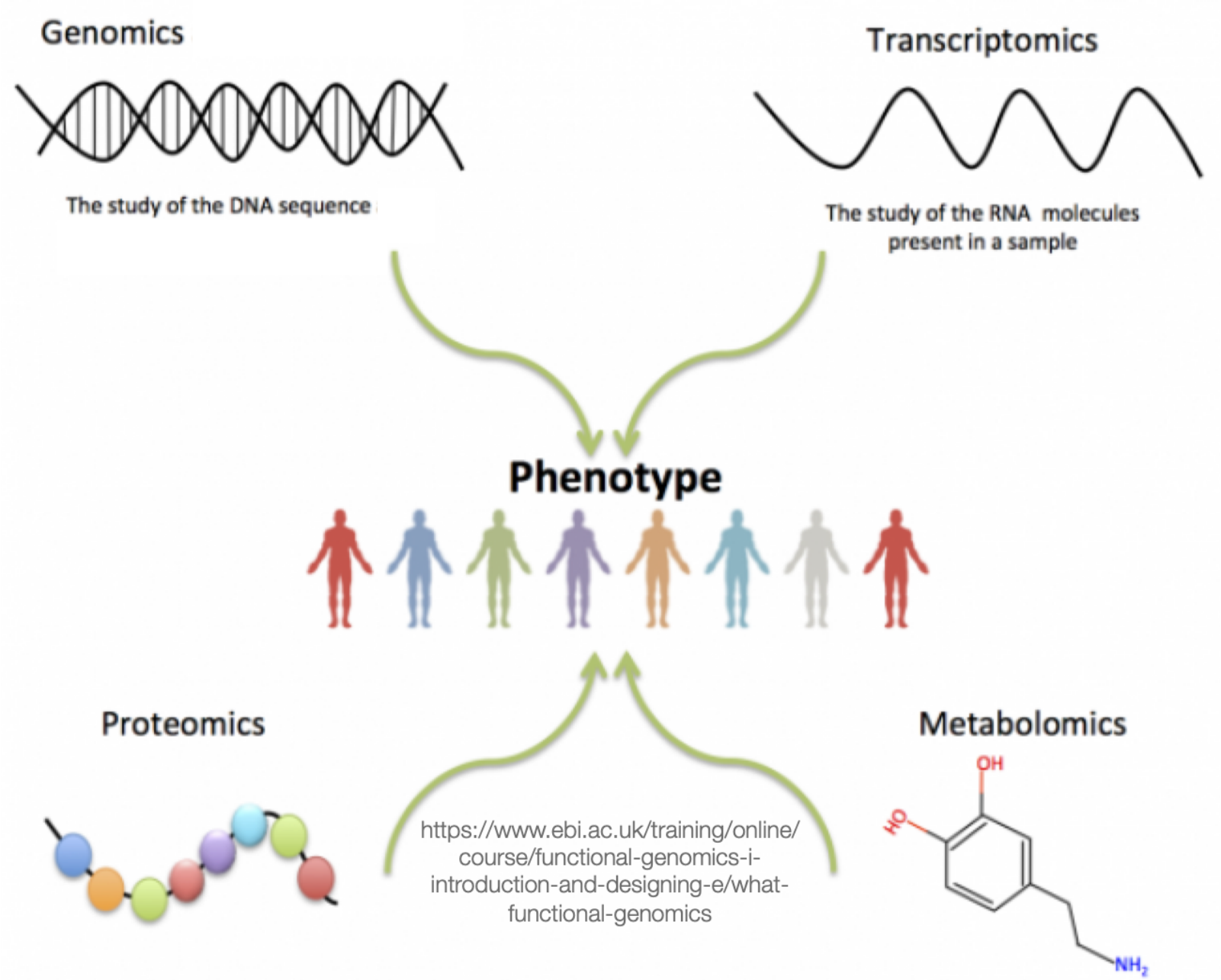
Applied Bioinformatics Core

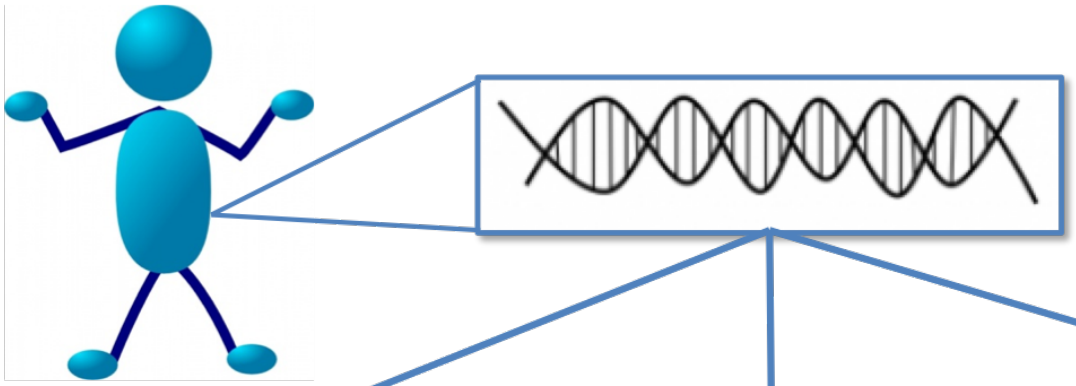
Slides at https://bit.ly/2T3sjRg[1]

March 31, 2020

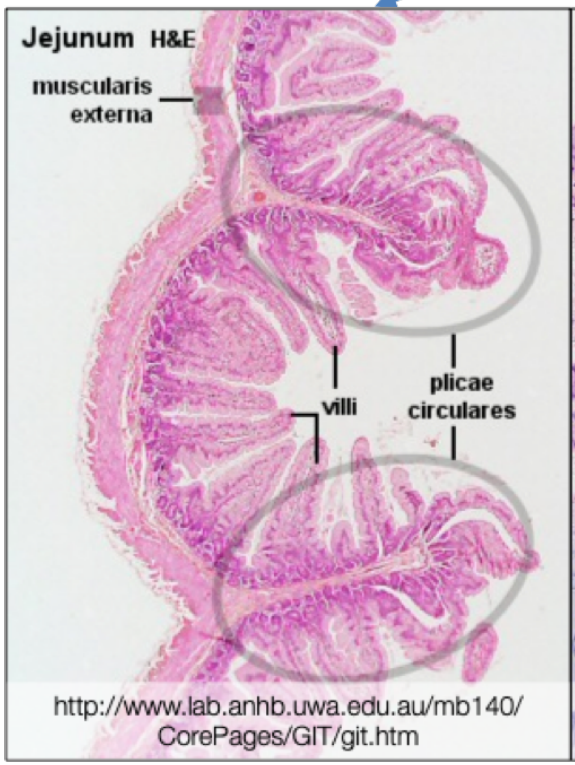
[1] https://physiology.med.cornell.edu/faculty/skrabanek/lab/angsd/schedule_2020/

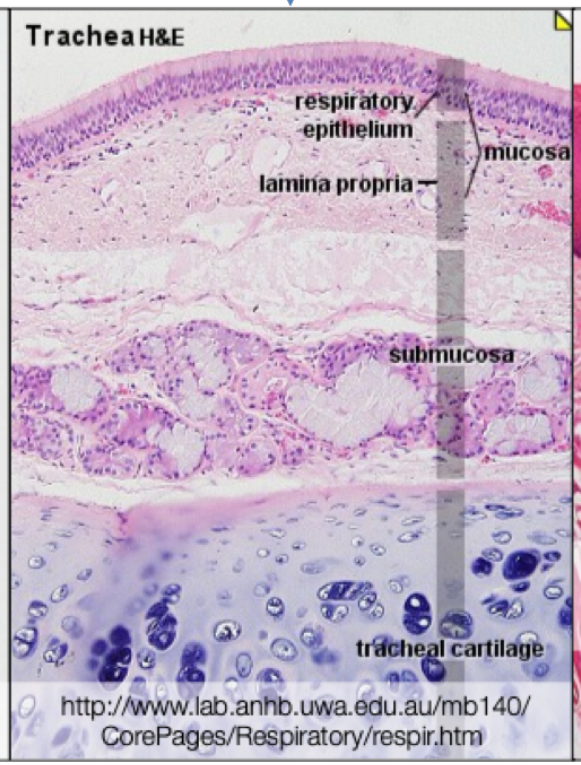
From DNA to phenotype


Genomics
The study of the DNA sequence
Transcriptomics
The study of the RNA molecules present in a sample
Phenotype
Proteomics
Metabolomics
OH
HO
NH2
https://www.ebi.ac.uk/training/online/course/functional-genomics-i-introduction-and-designing-e/what-functional-genomics



The **same DNA** code gives rise to vastly **different cell types**.

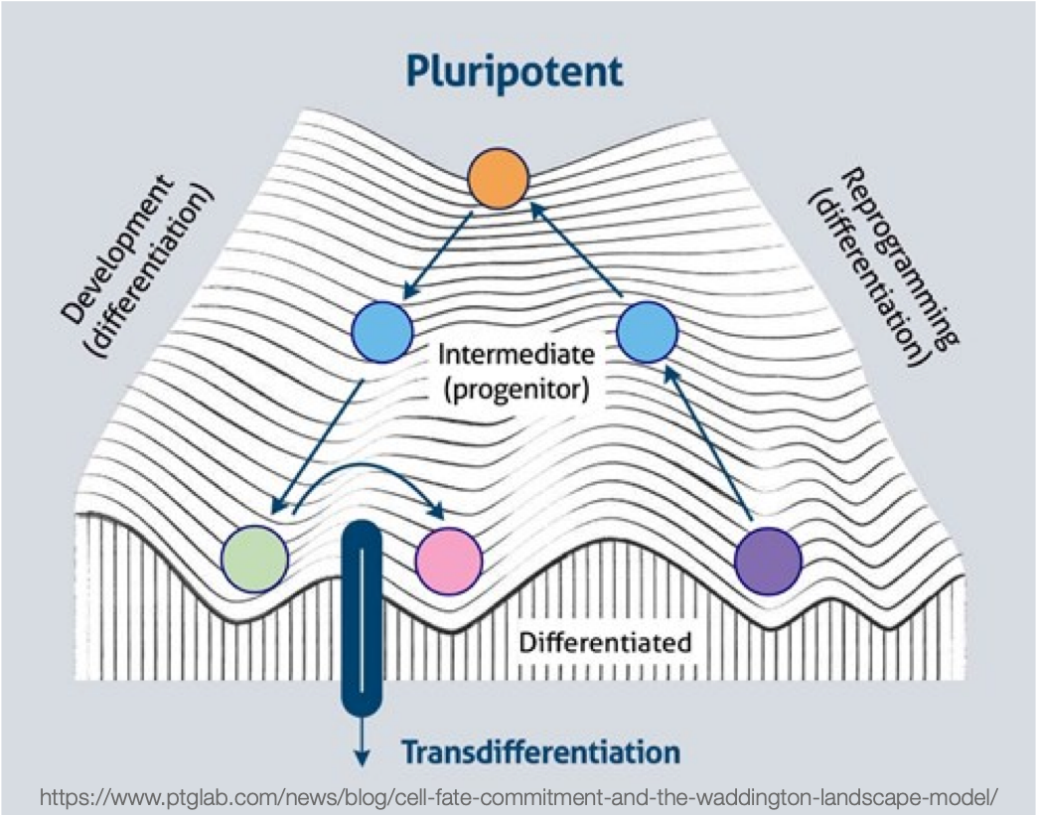




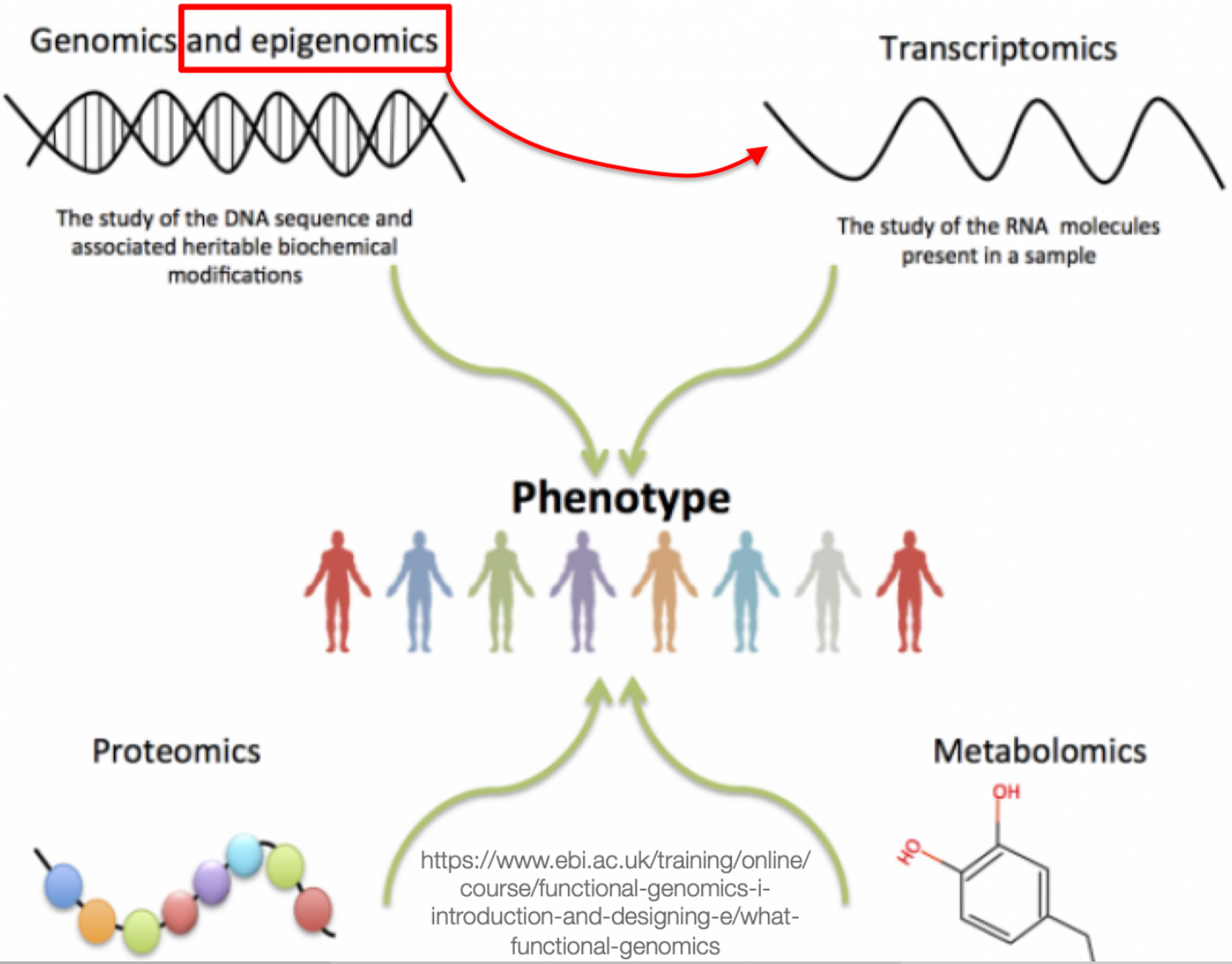
Epigenetics

Waddington's definition of epigenetics

Epigenetics encompasses the molecular mechanisms by which the genes of the genotype bring about phenotypic changes [Waddington, 1942].

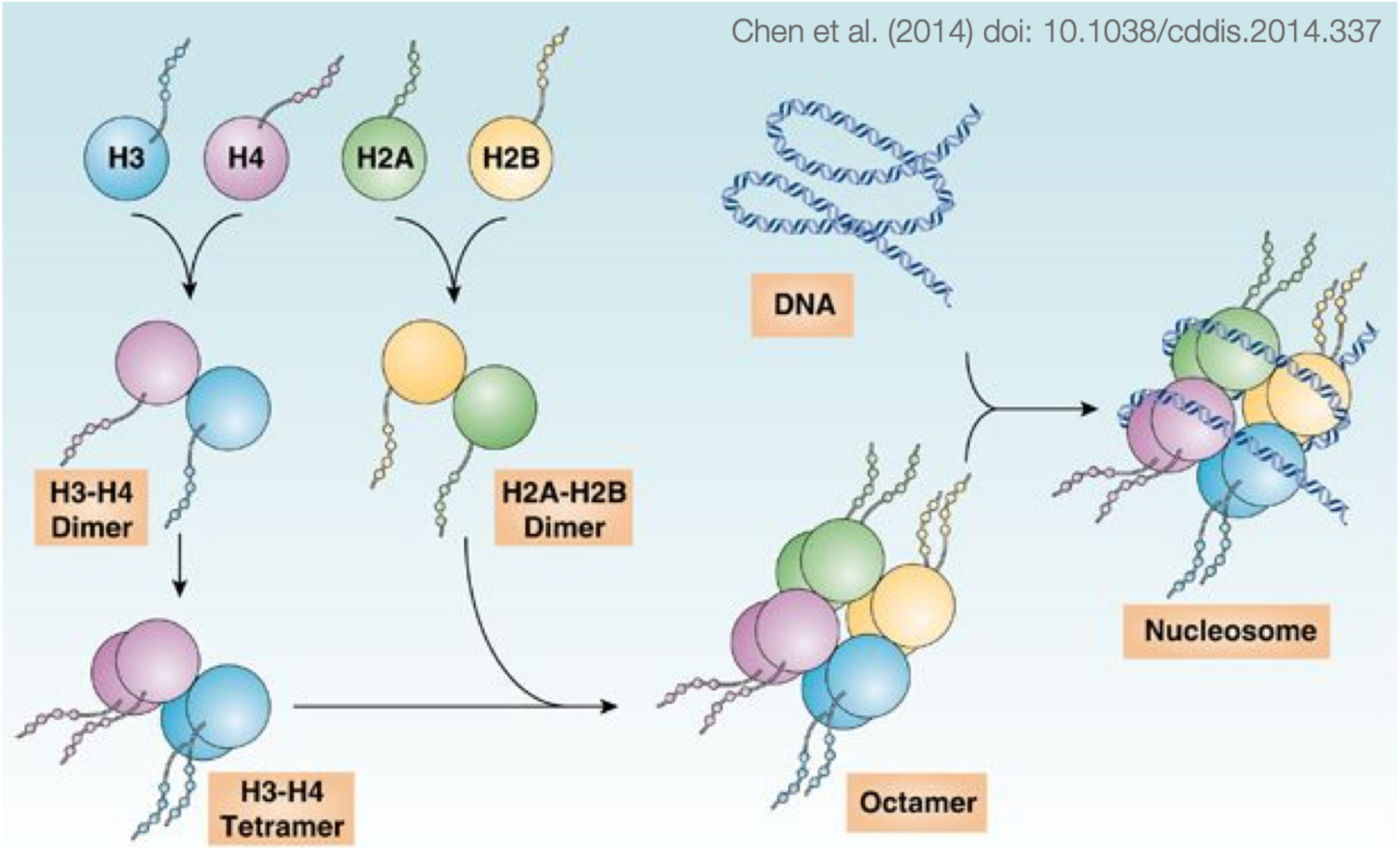


Epigenetics: understanding how the genetic code is interpreted (~ gene expression)

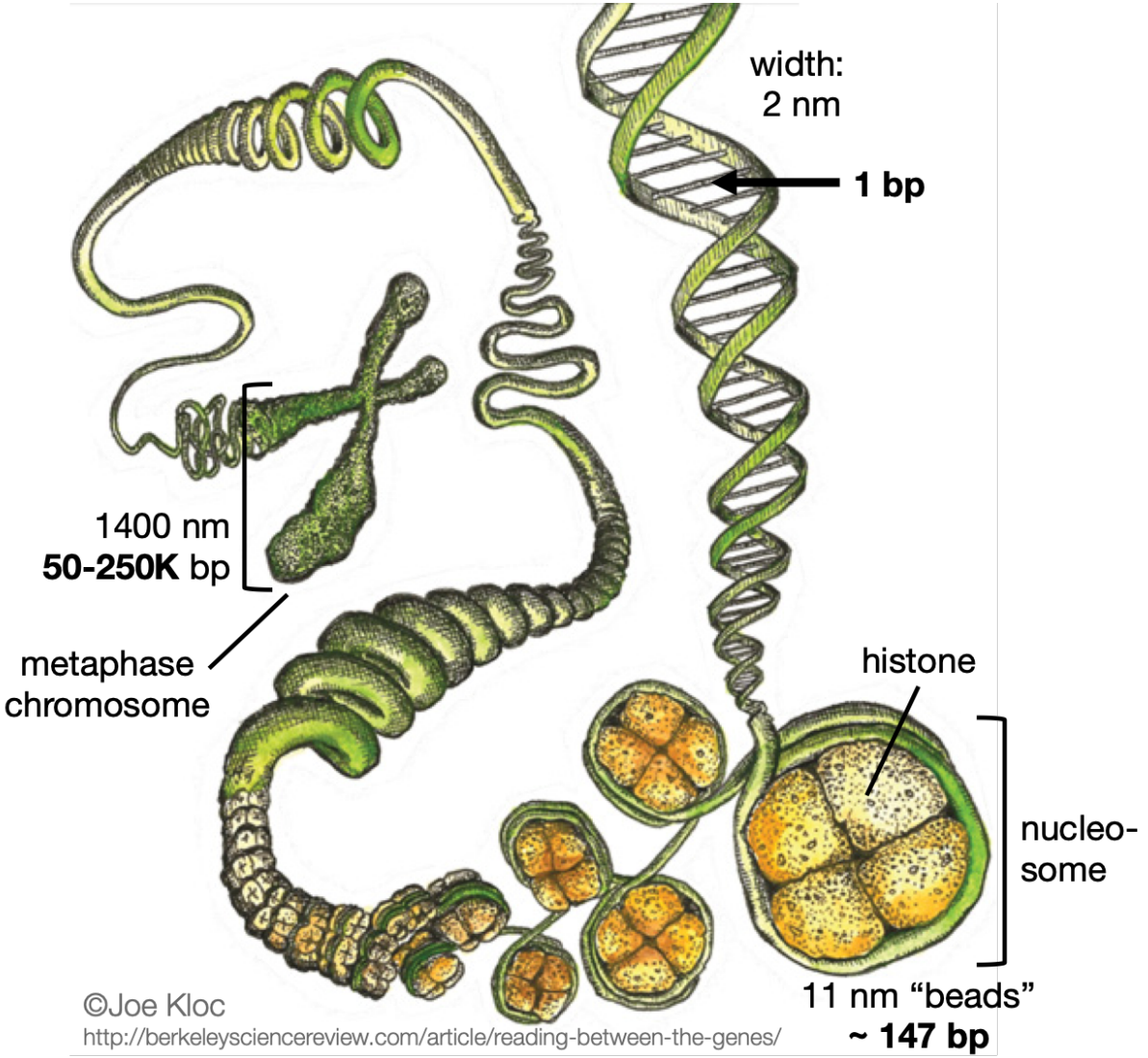


DNA does not occur naked in eukaryotic cells

Histone proteins are small alkaline proteins around which the DN**A** molecule is wrapped.



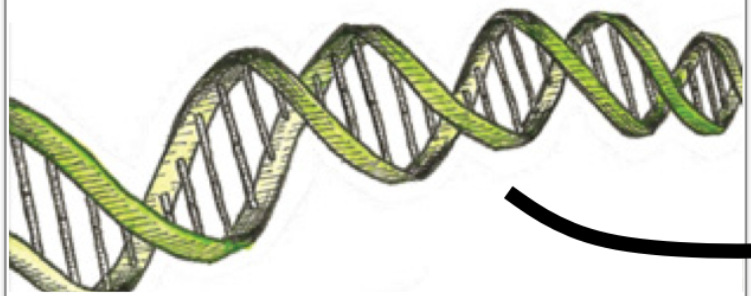
Chromatin = DNA + proteins + ncRNA



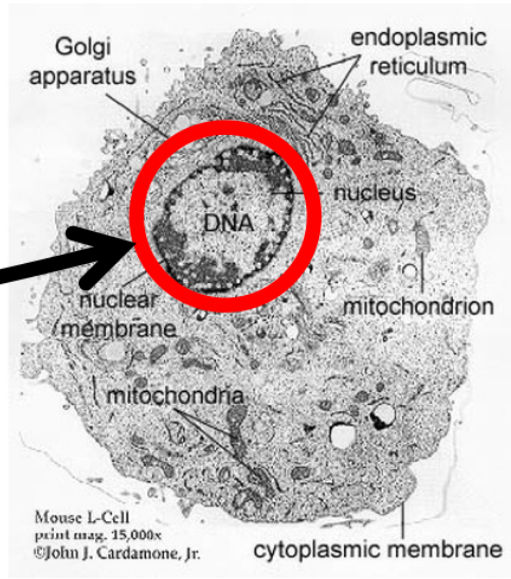
The most obvious function of chromatin is **DNA compaction**.

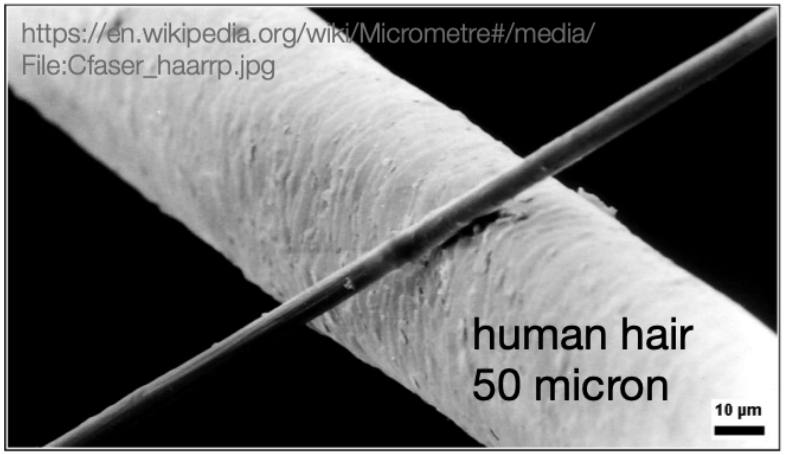
DNA compaction

DNA length:
2 m (6'6")



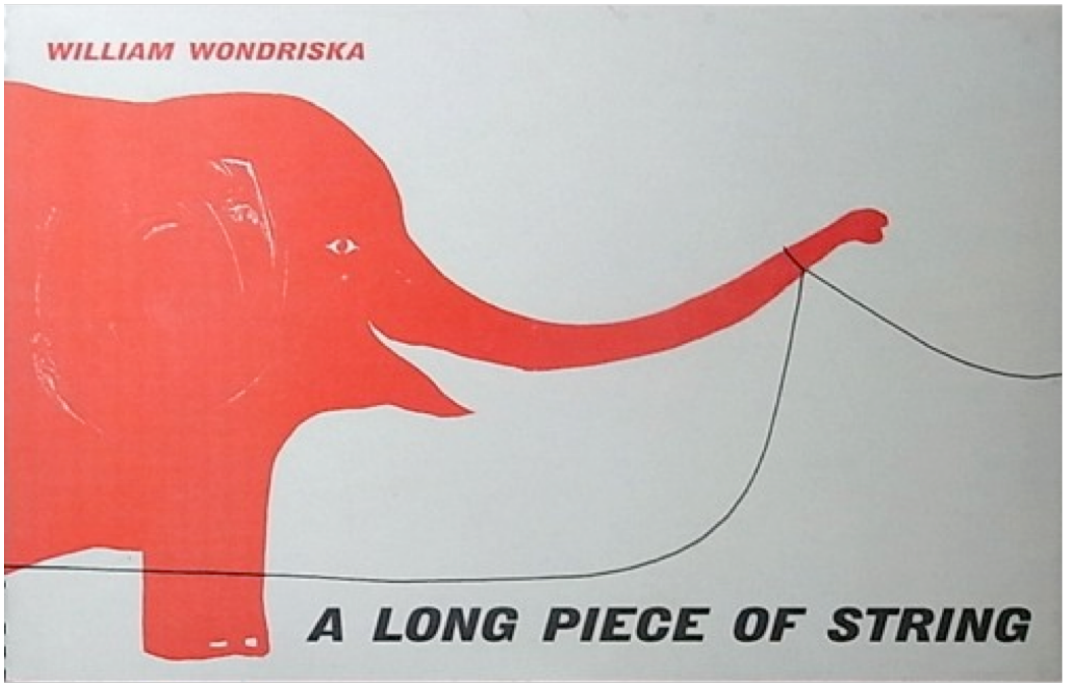
eukaryotic nucleus
diameter: ~ **2 micron** (10e-6 m)

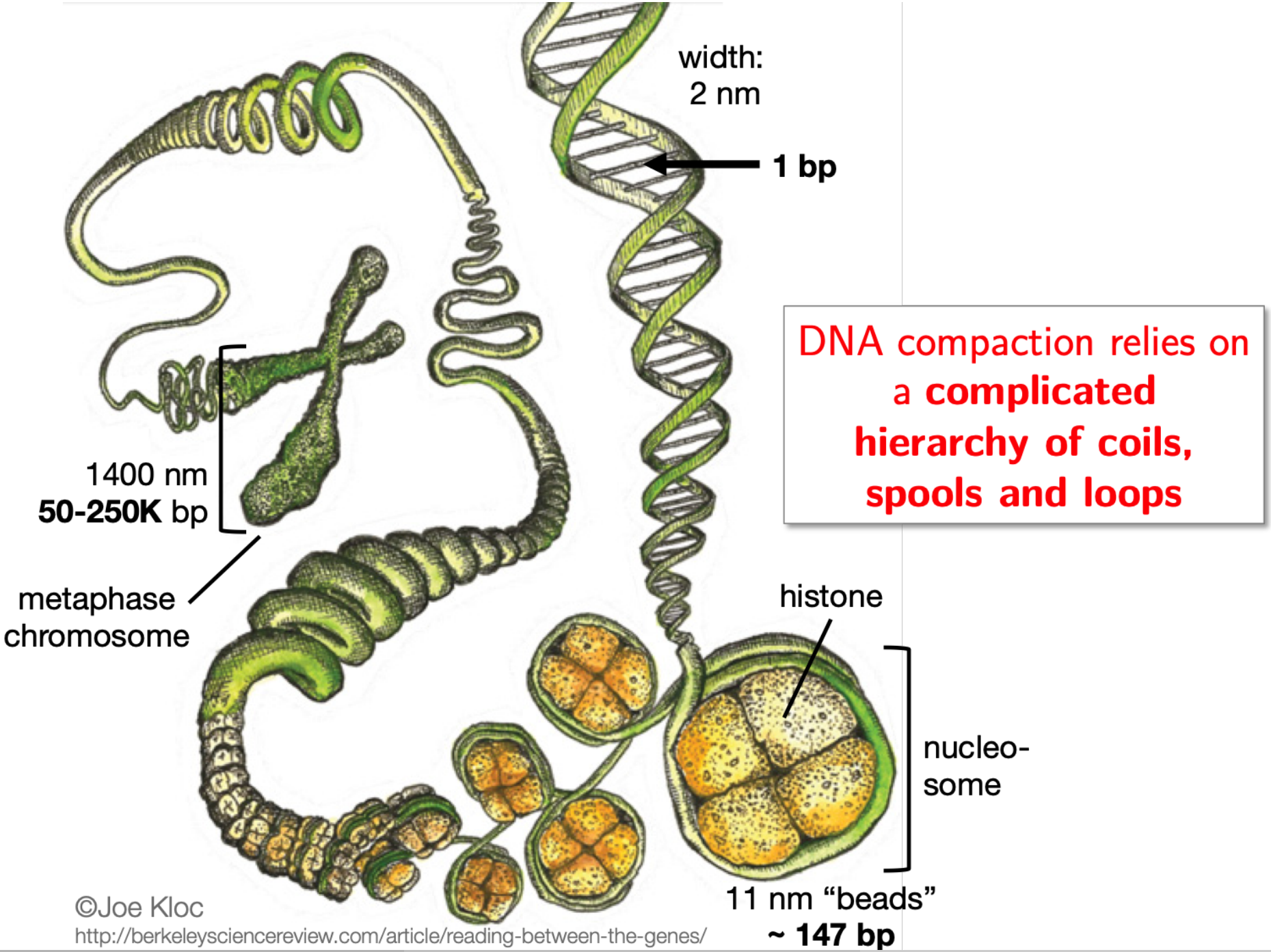




DNA compaction

Example for relatively trivial compaction:
375 m (~1230 ft) of yarn packed into a ball of about 10 cm x 4 cm (4"x1.6") using **simple coils**



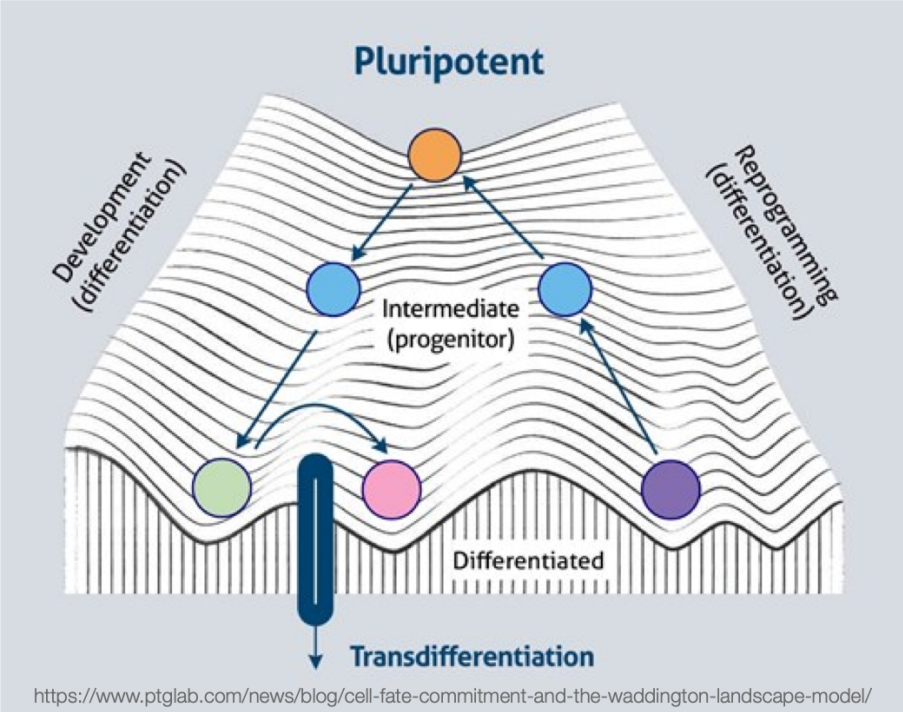


DNA compaction relies on a **complicated hierarchy of coils, spools and loops**

Studying Chromatin

From DNA to phenotype: epigenetics

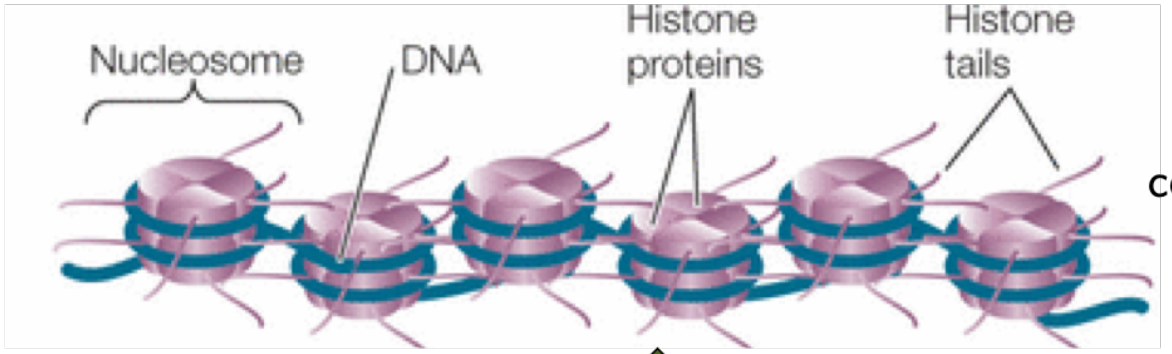
The current assumption is that the **chromatin structure** is an essential part of defining an individual cell's fate, i.e. by interacting tightly with DNA and regulating access to it, chromatin has a key role in how transcription is achieved in a highly **time- and tissue-dependent** manner.



"Understanding the chromatin structure can give a perspective of how a certain mRNA expression state was reached and how a cell might advance."
[Winter et al., 2015]

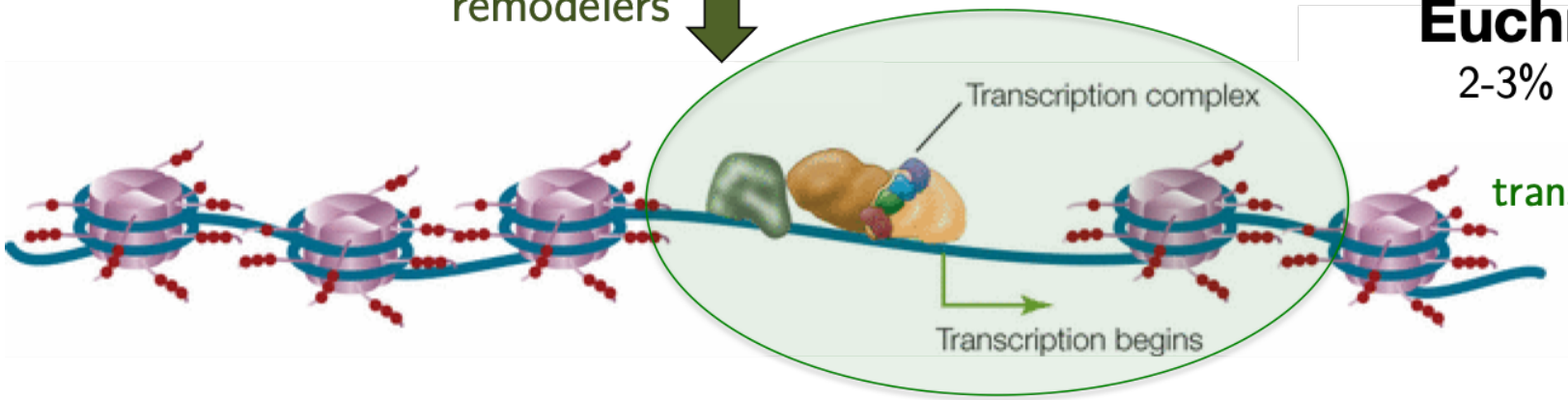
2 basic chromatin states based on nucleosome occupancy

For transcription to occur, the RNA Pol II machinery needs to access the **naked** DNA strand, i.e. the chromatin needs to be made **locally accessible**.



Heterochromatin
> 95% of the DNA
compacted and transcriptionally repressed

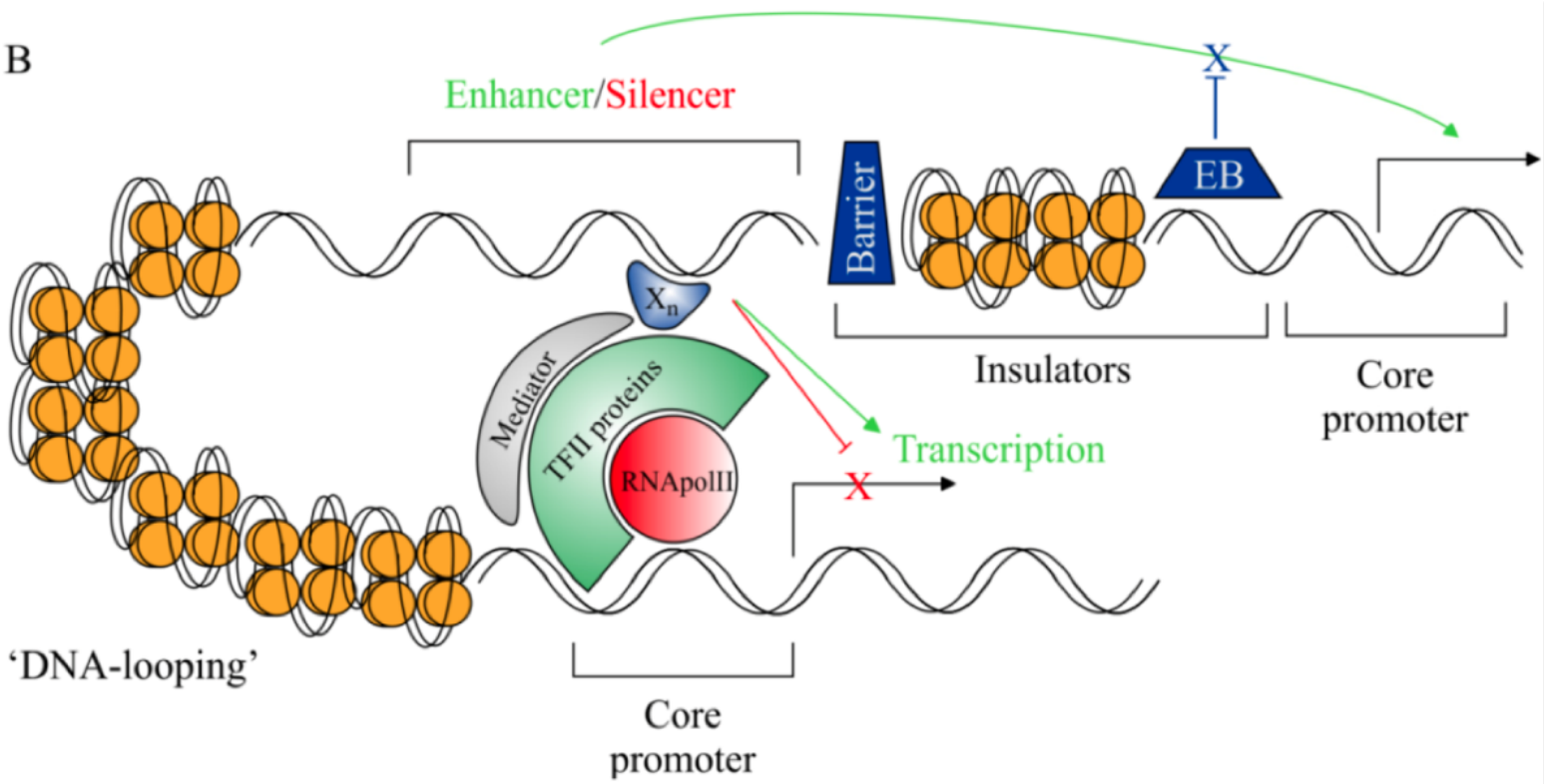
nucleosome remodelers



Euchromatin
2-3% of the DNA
"open" and transcriptionally permissive

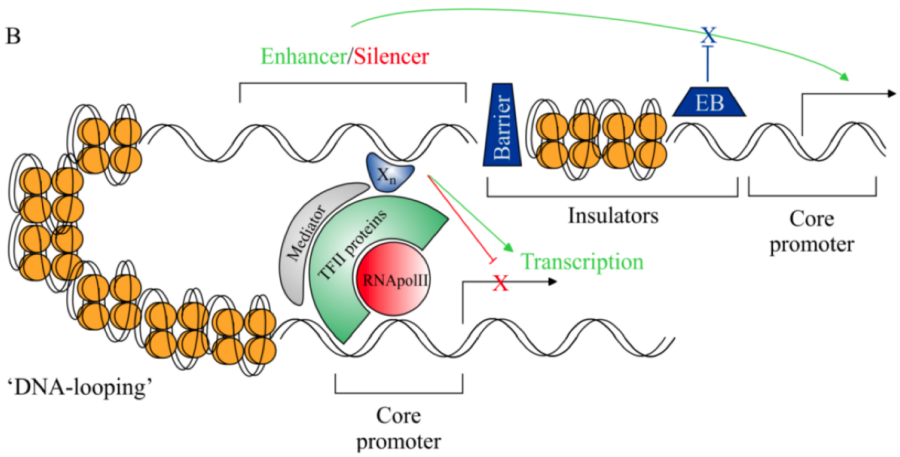
Open chromatin harbors numerous regulatory elements

- **Trans-regulatory elements** = DNA encoding transcription factors ⇒ the actual effectors are *proteins* (e.g. RNA Pol II, Mediator, TF
- **Cis-regulatory elements (CRE)** = non-protein-coding DNA that regulates transcription of neighboring genes ⇒ the effectors are thought to be (at least partially) the *DNA sequences* (and sometimes their corresponding *transcripts*)



Open chromatin harbors numerous regulatory elements

Cis-regulatory elements (CRE) = non-protein-coding DNA that influences gene transcription



promoter:

- "beginning" of a gene: region where the Pol II and its co-factors (100s!) assemble
- between 500-3,000 bp

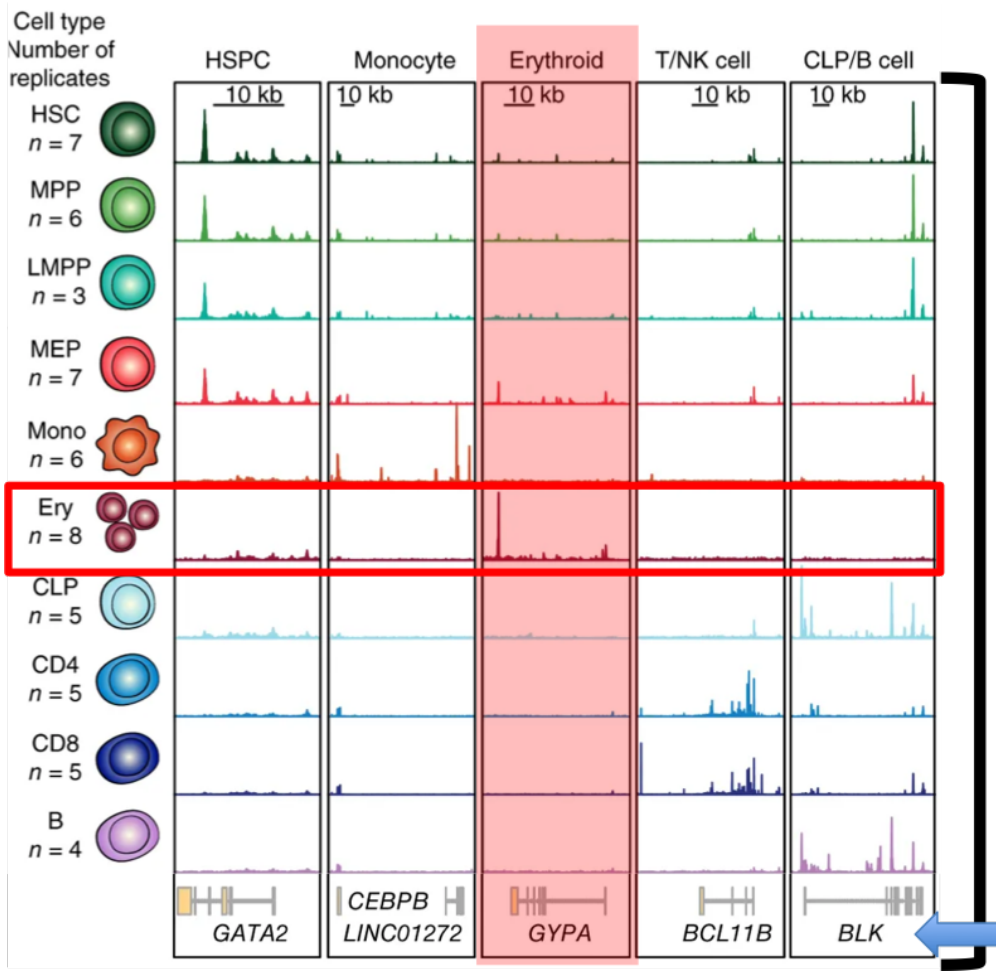
Distant CREs:

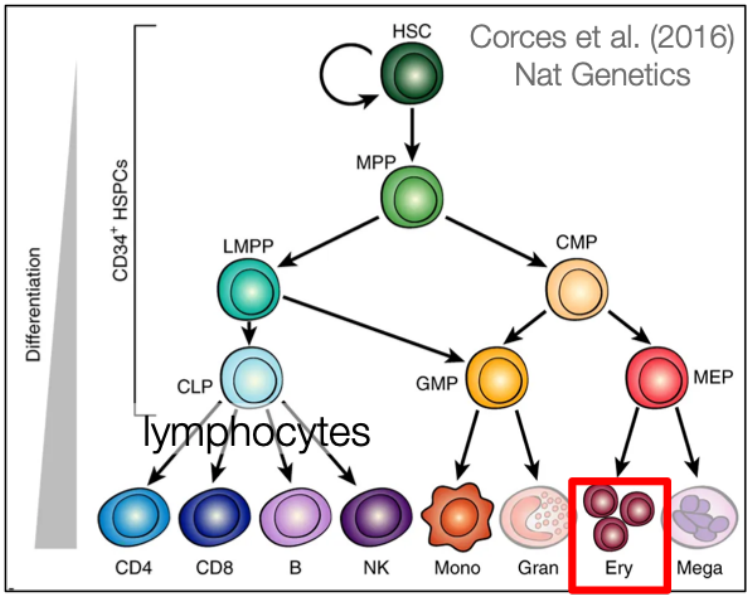
- **enhancers**, **silencers**; fairly small (ca. 50-100 bp)
- regions where additional TF or inhibitor proteins bind
- often form indirect interactions with the target promoter

insulators:

- e.g. prevent chromatin condensation of active regions
- some insulators maintain enhancers' specificities by blocking them from impinging on other genes

Understanding cell-type specific chromatin accessibility patterns helps dissect different cell type lineages





chromatin accessibility profiles for these genes in different cell types

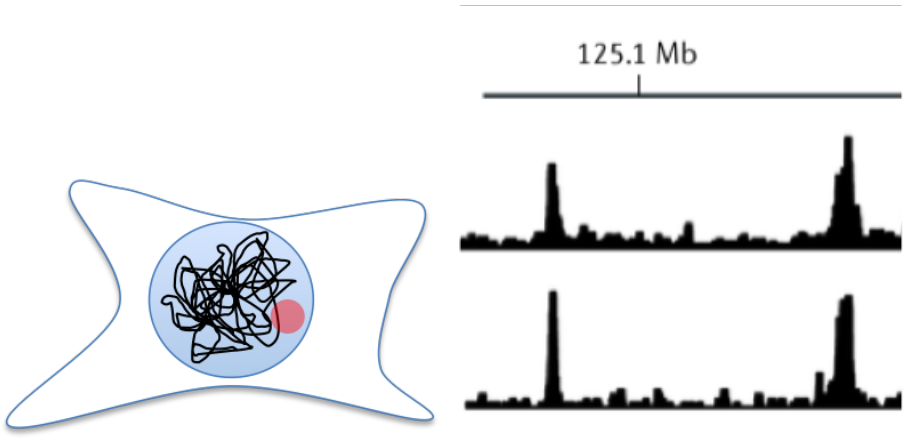
genes encoding lineage-specifying TF

NGS techniques for studying chromatin and DNA modifications

Basic concept

***Enriching* for DNA regions of interest** and inferring their location via NGS-based quantification.

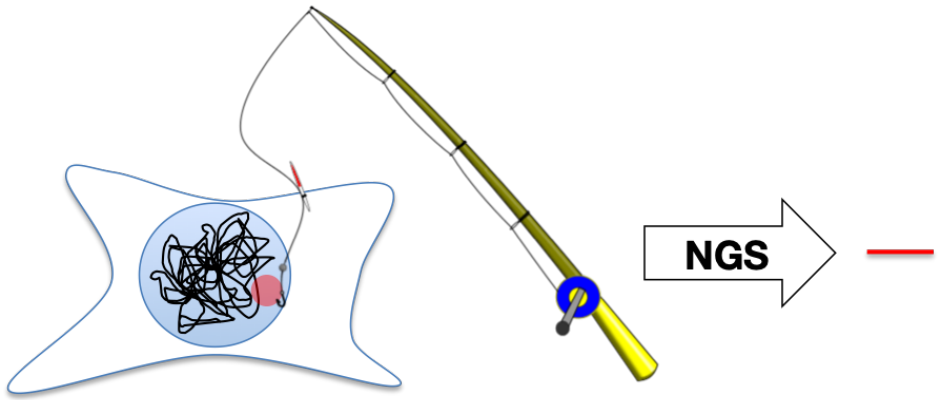
Similar to RNA-seq, we're trying to **quantify** regions of interest. In contrast to RNA-seq, however, we're quantifying **DNA** regions with specific properties – such as being accessible – that make them amenable to biochemical enrichment strategies that exploit these properties.



NGS techniques for studying chromatin and DNA modifications

Basic concept

***Enriching* for DNA regions of interest** and inferring their location via NGS-based quantification.

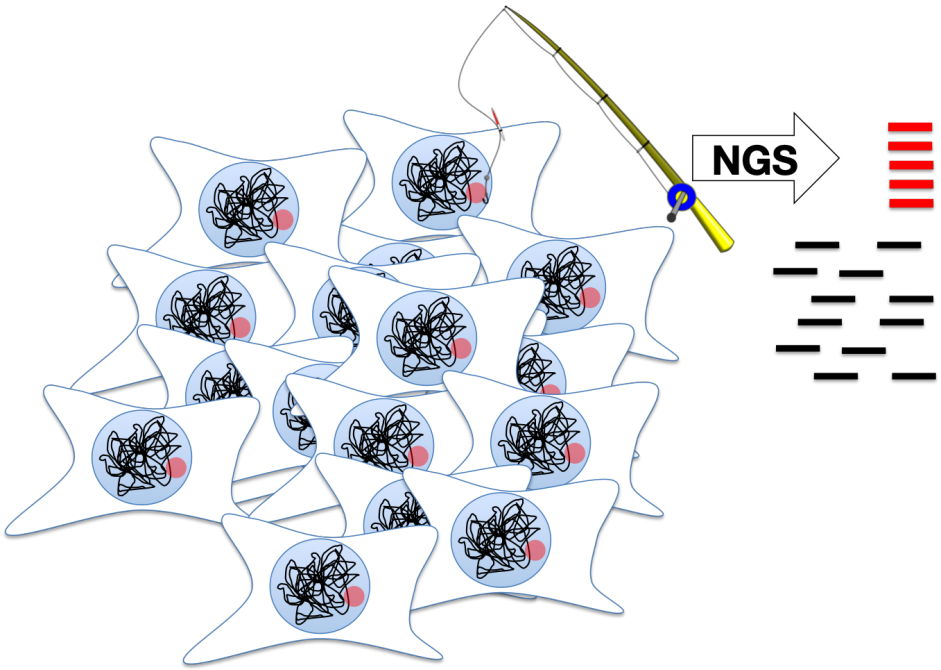


- red dot = region of interest, e.g. transcription factor binding site

NGS techniques for studying chromatin and DNA modifications

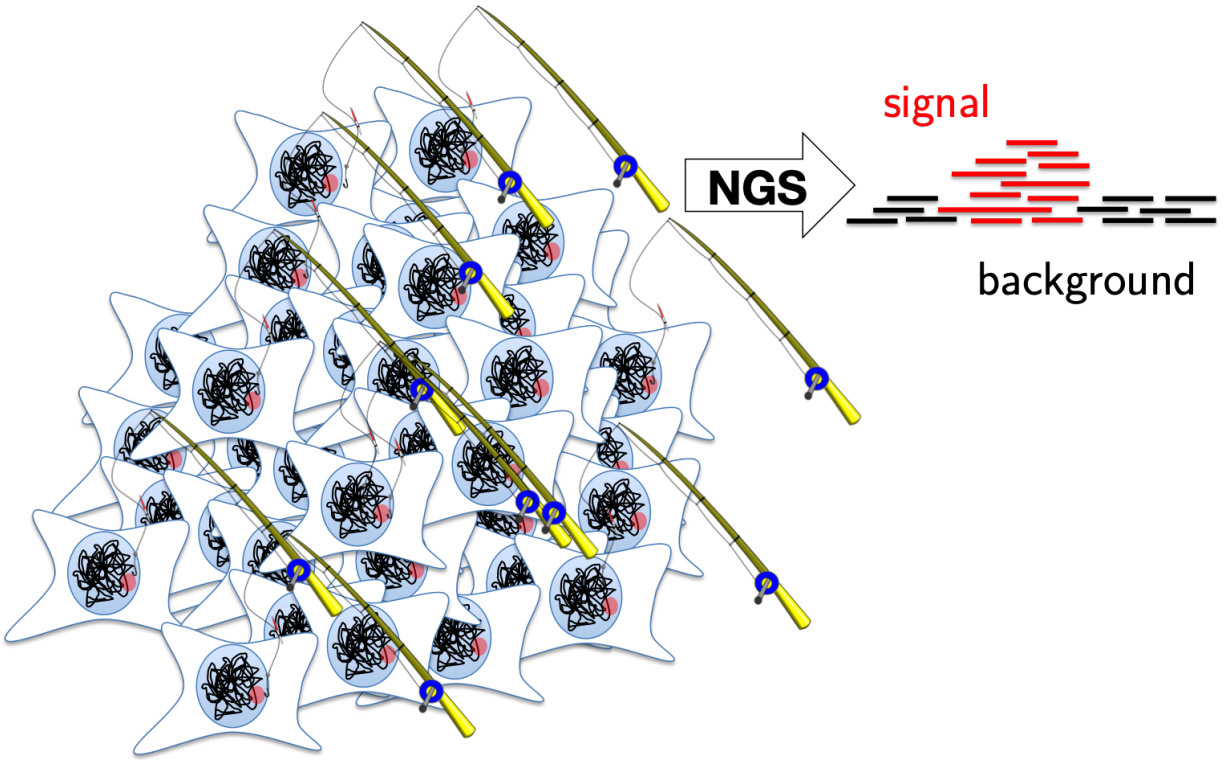
Basic concept

***Enriching* for DNA regions of interest** and inferring their location via NGS-based quantification.



NGS techniques for studying chromatin and DNA modifications

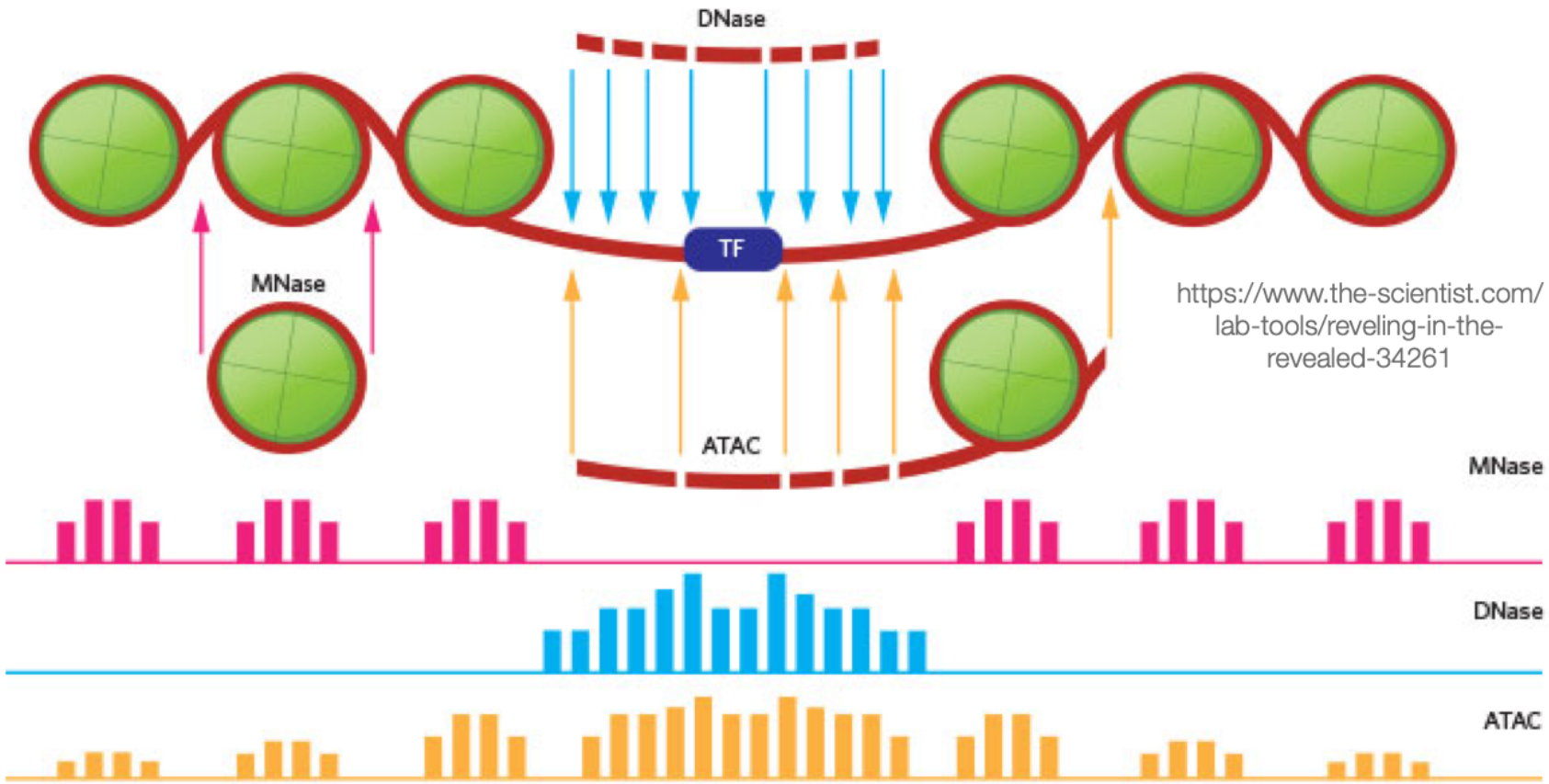
***Enriching* for DNA regions of interest** and inferring their location via NGS-based quantification.



ATAC-seq principles

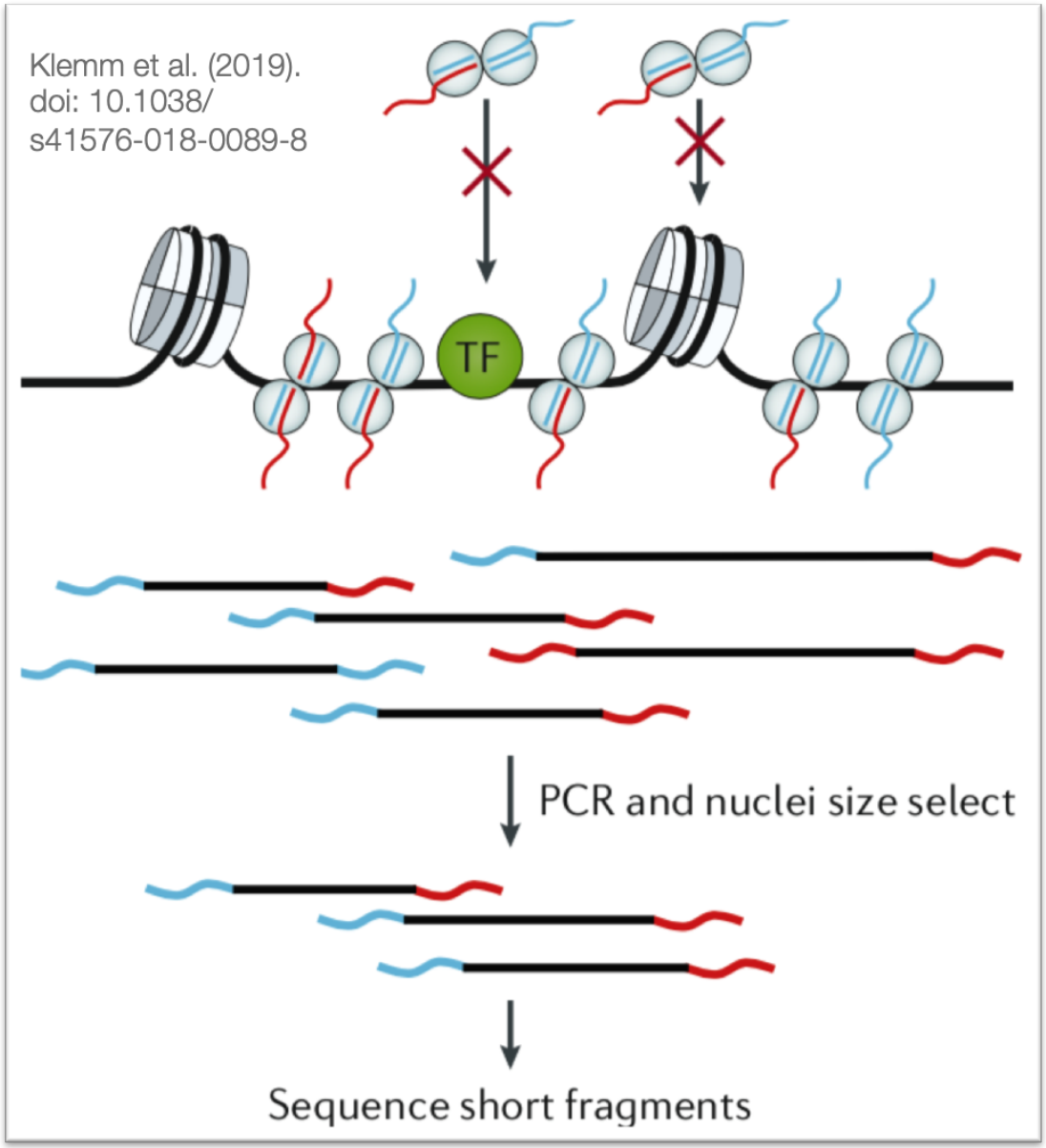
Identifying accessible chromatin regions

Active CRE (promoters, gene bodies, enhancers, TFBS) are expected to be accessible.



Open chromatin is identified via **ATAC**-, DNase-, MNase-seq (and more).

Assay for transposase-accessible chromatin (ATAC)



Tn5 transposase with sequencing adapters

↓

"attacks" **nucleosome-free DNA** regions

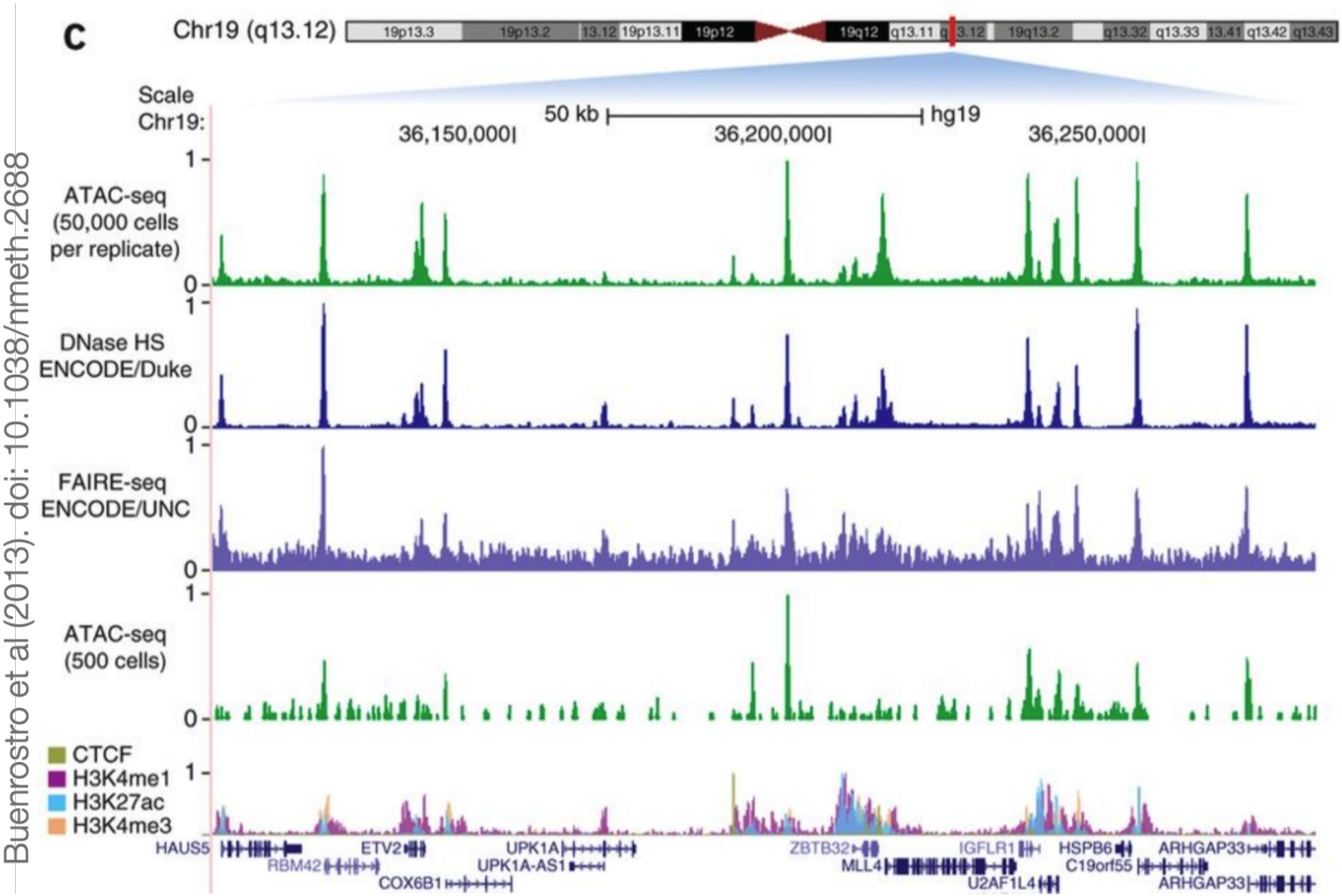
↓

TAGMENTATION
= DNA **fragmentation** + adaptor-**tagging**

↓

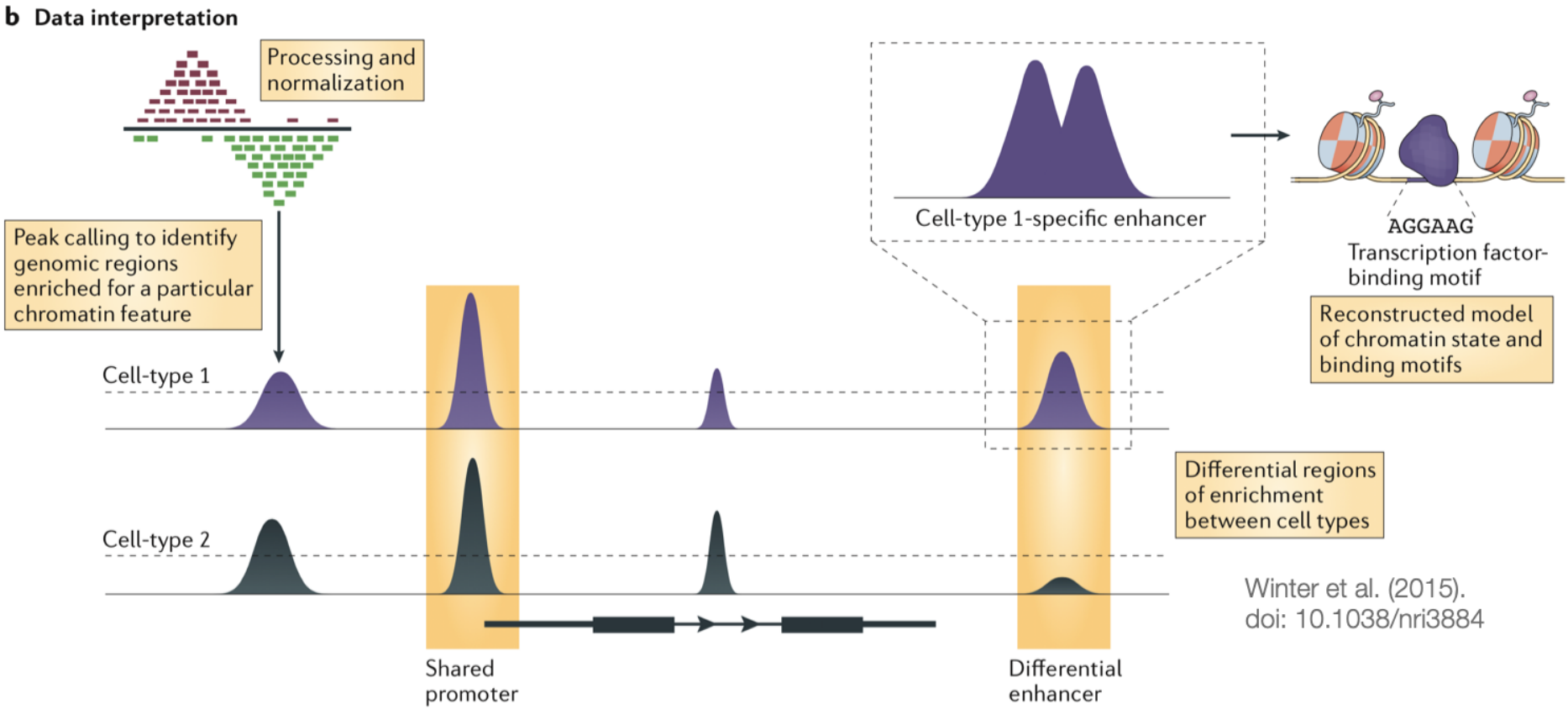
fragments that will be sequenced represent nucleosome-free DNA

ATAC-seq profiles

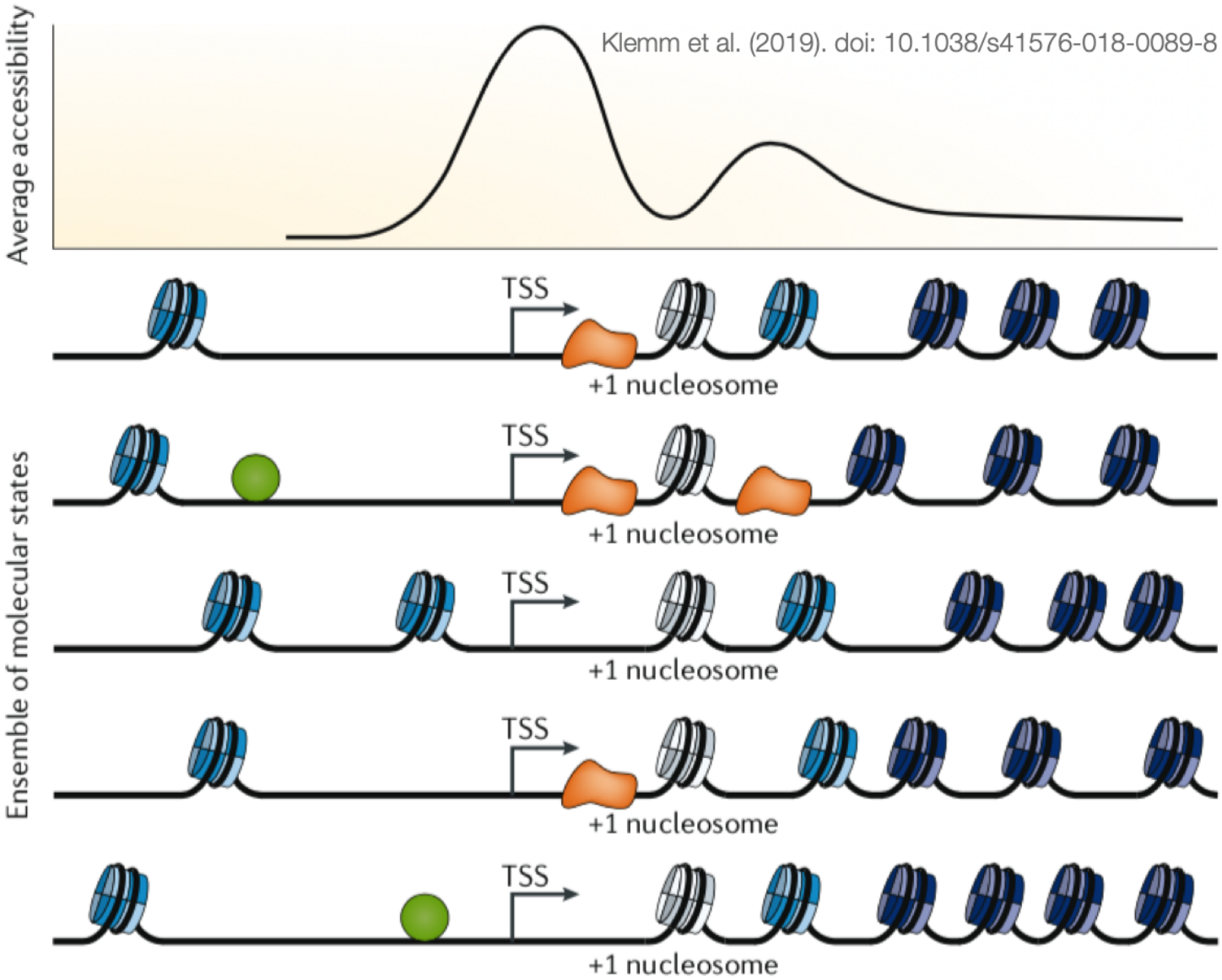


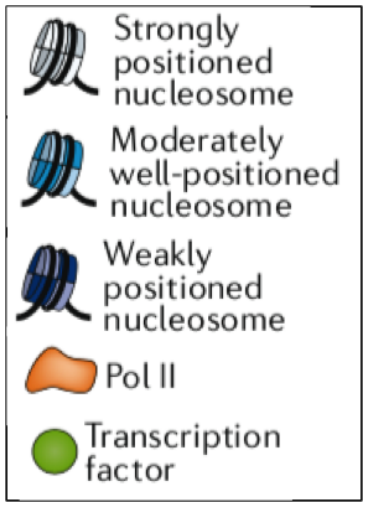
Buenrostro et al (2013). doi: 10.1038/nmeth.2688

Interpretation of ATAC-seq data



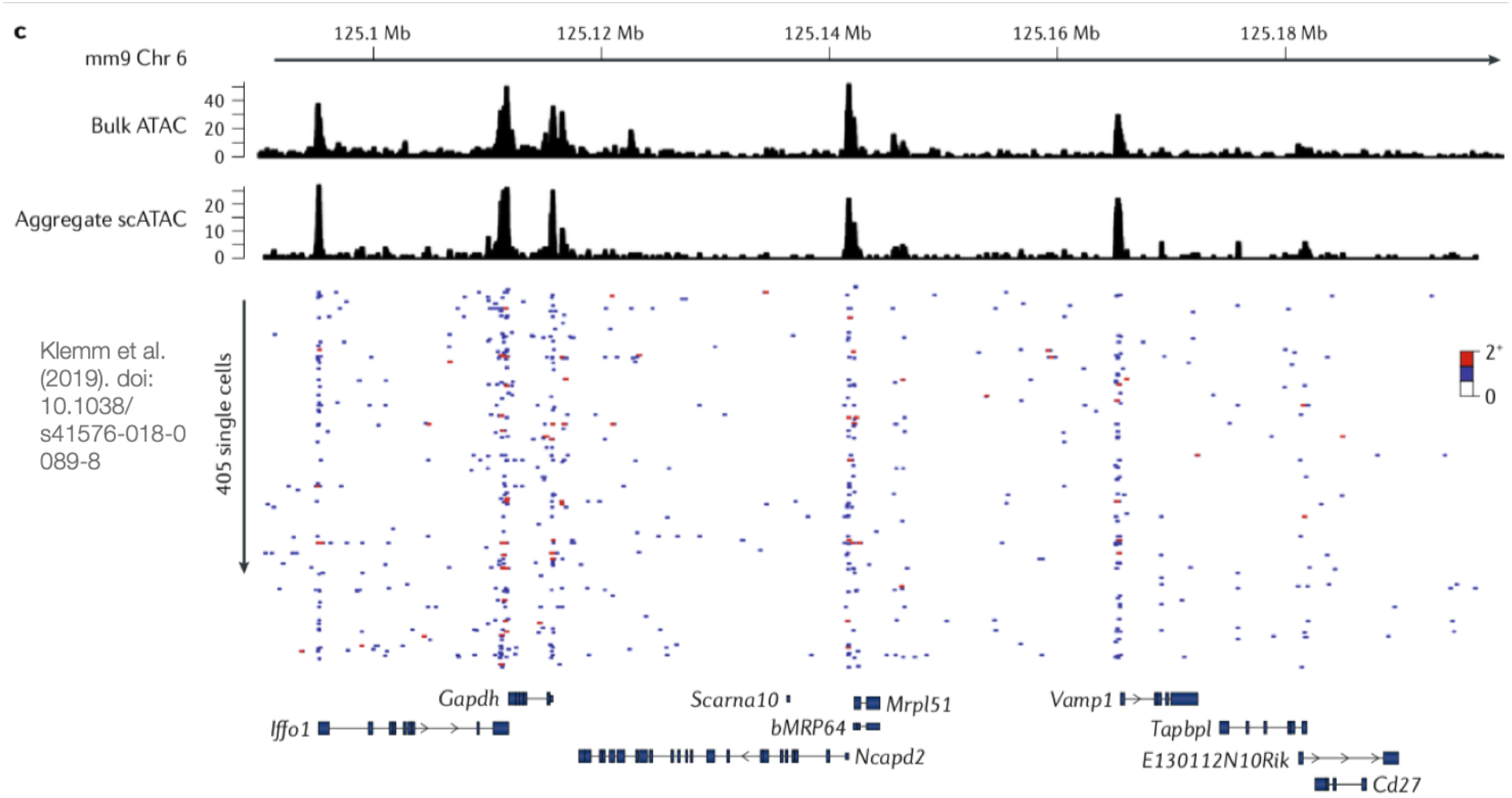
ATAC-seq profiles are typically population snapshots



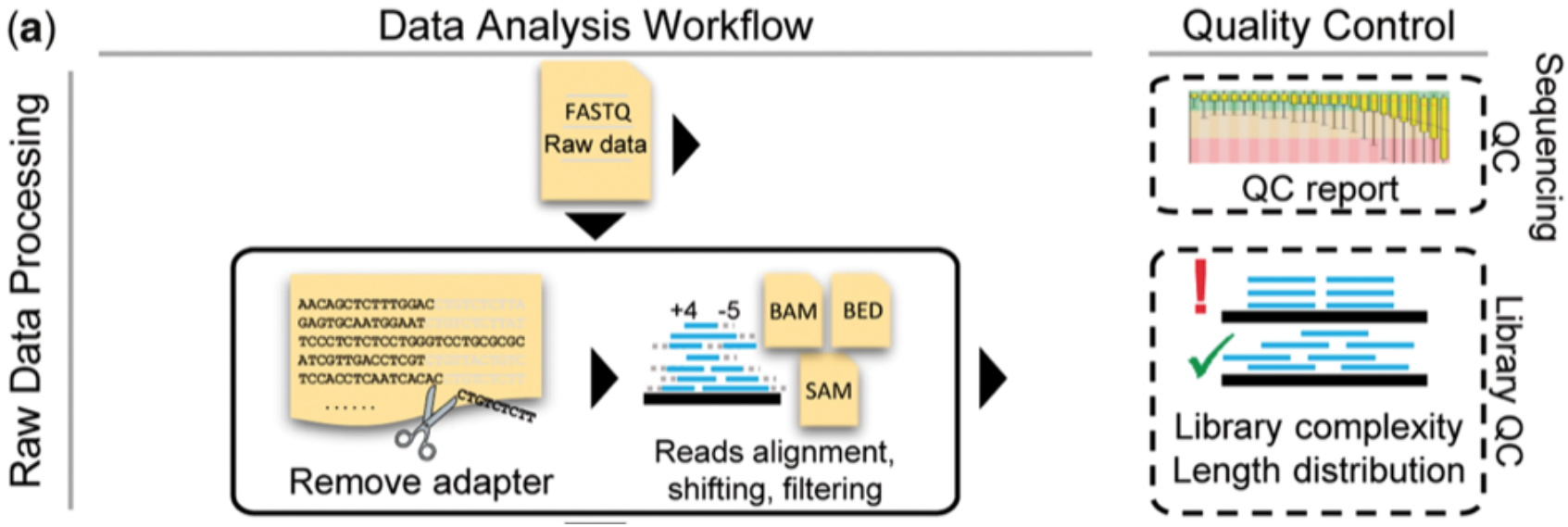


ATAC-seq profiles usually represent the **average accessibility** of a heterogeneous collection of single molecules.

ATAC-seq profiles are typically population snapshots, but scATAC-seq is possible



Processing ATAC-seq data



- the usual QC of FASTQ and alignment apply
- alignment should be performed with an aligner tailored for **genome** sequencing, i.e. not `STAR`, but rather `BWA`

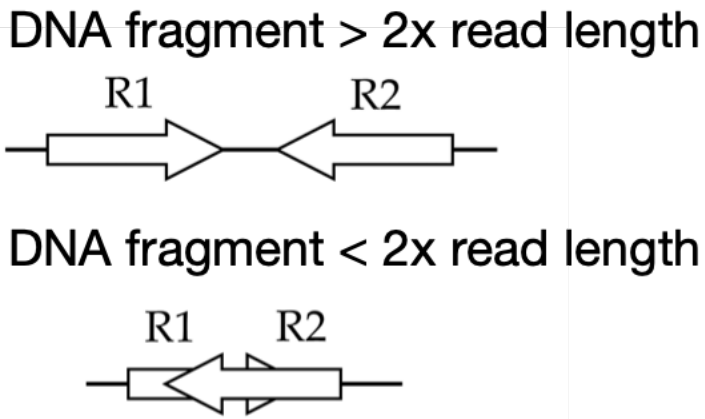
Established ATAC-seq pipelines

- **ENCODE**
  - lots of QC scores and guidelines for identfying samples that worked/failed
  - somewhat cumbersome implementation
- **Tom Carroll**'s R-based workflow
  - mostly follows ENCODE's guidelines
  - every command is shown including some explanations about important parameters
  - R is not the best-suited environment for some of the steps (e.g. bigWig generation)
- **Harvard FAS**
  - some steps of the ENCODE pipeline are re-worked/re-thought
  - alternative peak caller (not yet peer-reviewed, but more versatile/ATAC-seq-oriented than MACS2)

See Yan et al. [2020] for a detailed processing guide.

Raw data processing: FASTQ to BAM

- **FastQC** – the usual suspects: sequencing quality, duplications, contaminations
- **adapter removal** may be warranted
  - PE sequencing will often lead to frequent adapter sequences for ATAC-seq data because many *fragments* are shorter than 2x50bp

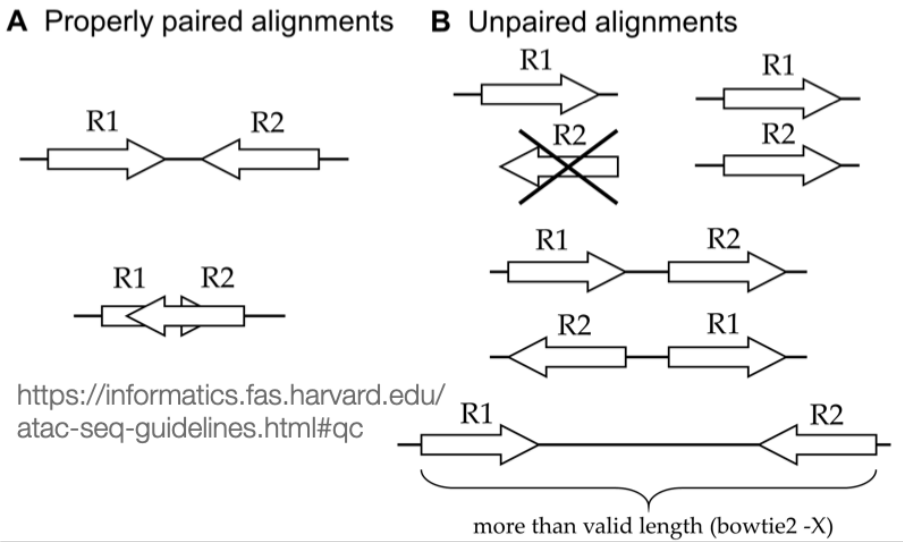


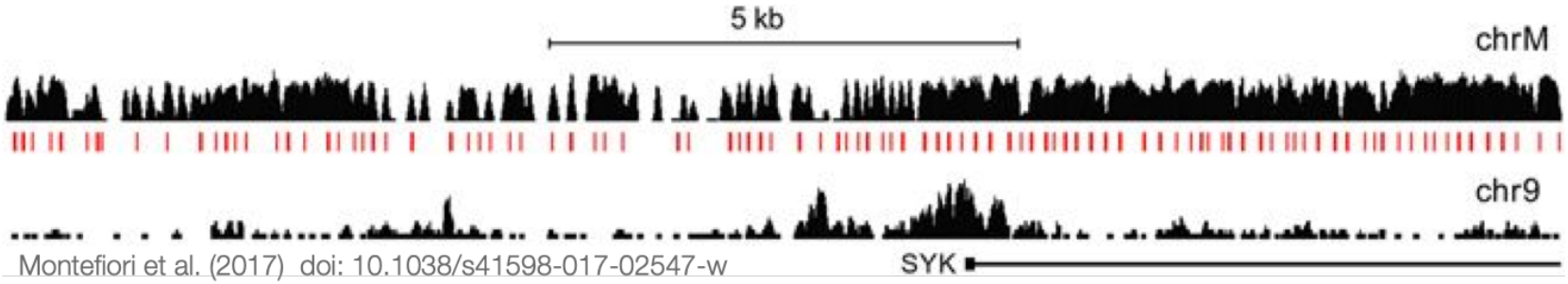
- genome **aligners** for short reads, e.g. Bowtie2 or **BWA**

Raw data QC: filtering the BAM files

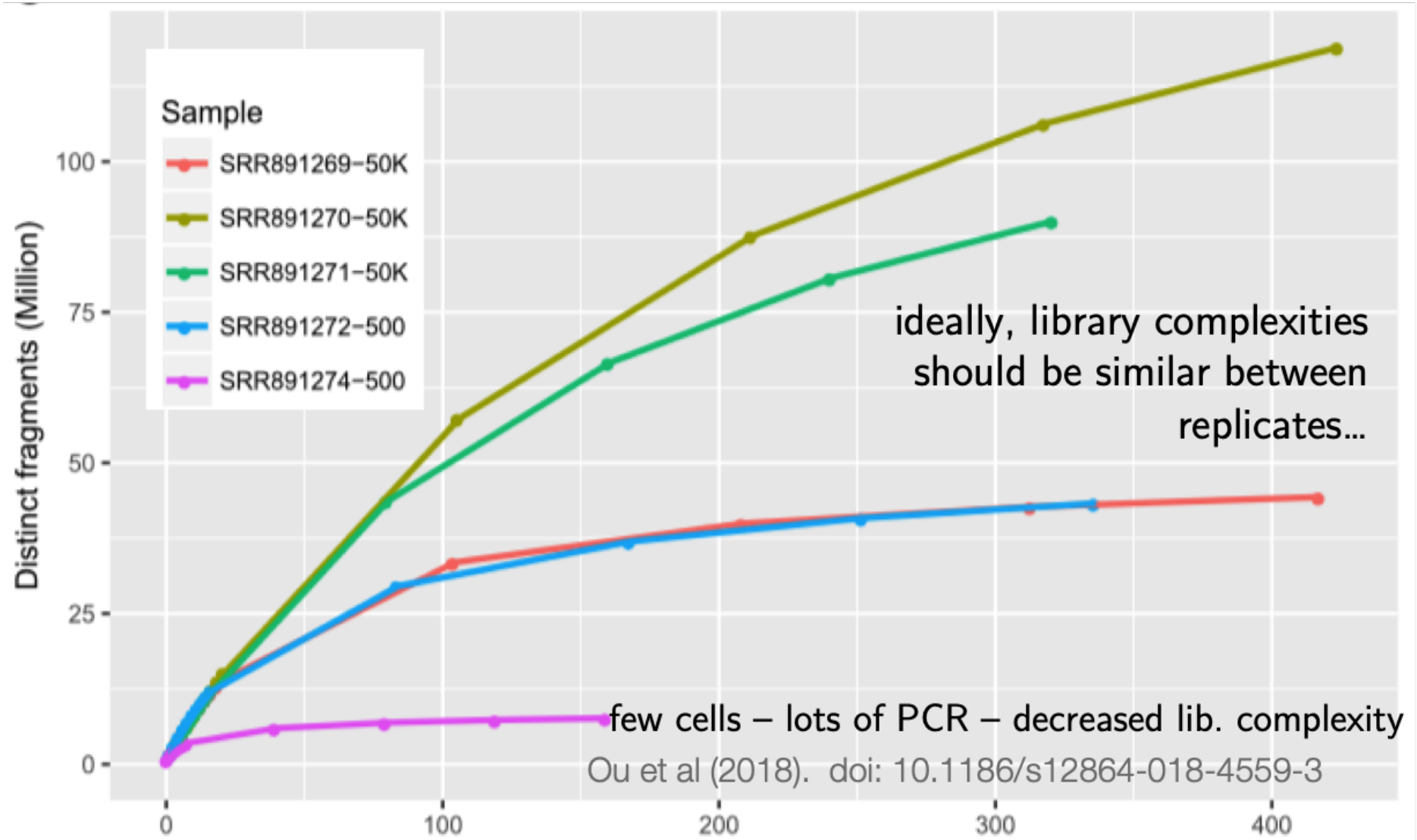
The following reads are removed:

- mitochondrial reads
- discordantly "paired" reads
- non-uniquely aligned reads
- PCR duplicates
- reads corresponding to fragments < 40 bp (see slides about fragment size distributions)
- reads overlapping with blacklisted regions

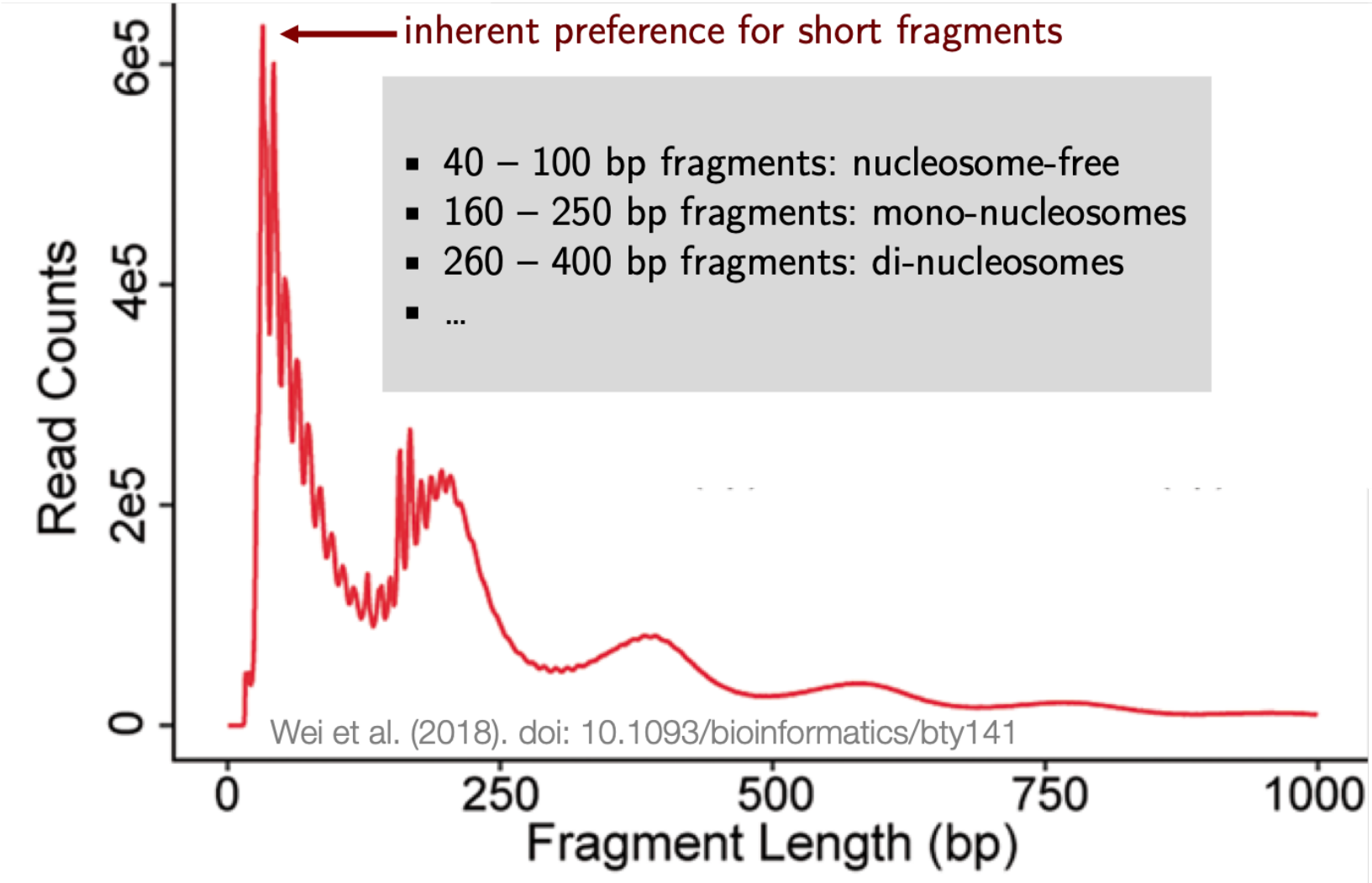




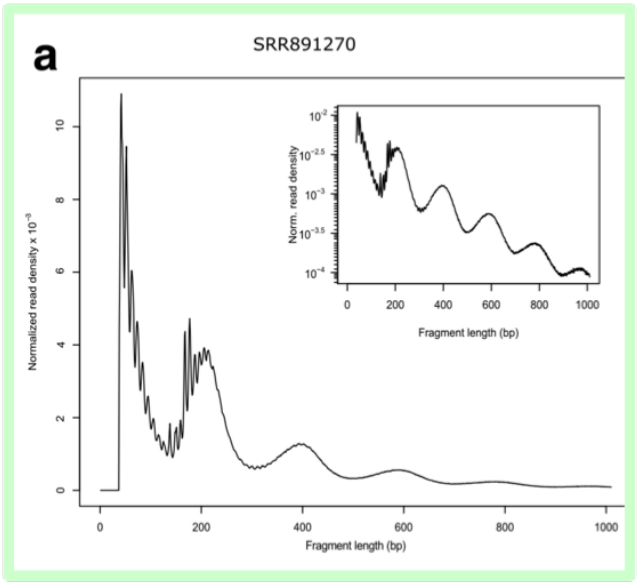
PCR duplicates are frequent – more so for low cell numbers!



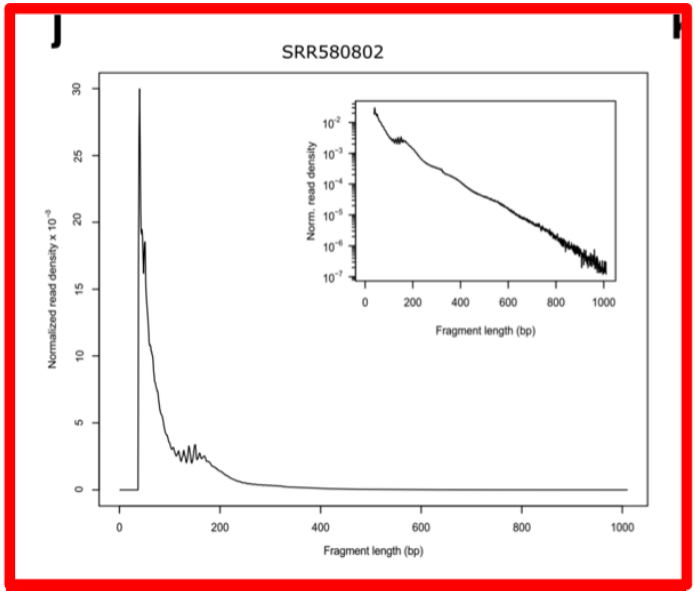
The dominant fragment size distribution signal in ATAC-seq should reflect the nucleosome pattern

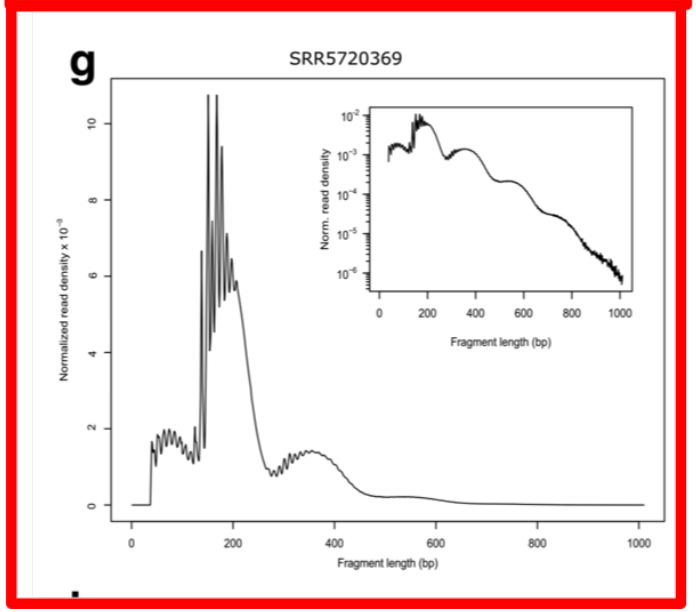


Examples of ATAC-seq frag. size distributions



Ou et al (2018). doi: 10.1186/s12864-018-4559-3



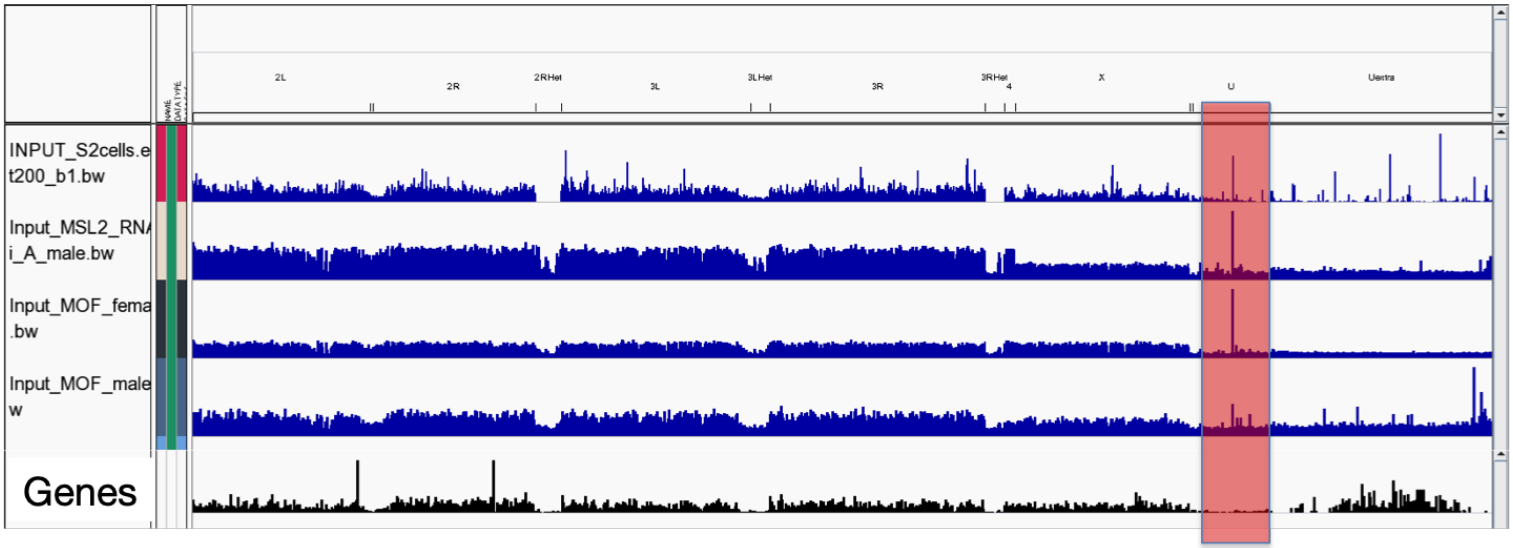


- typical problems seen here:
  - overdigestion/too much Tn5
  - too little Tn5/incomplete digestion
  - flawed size selection

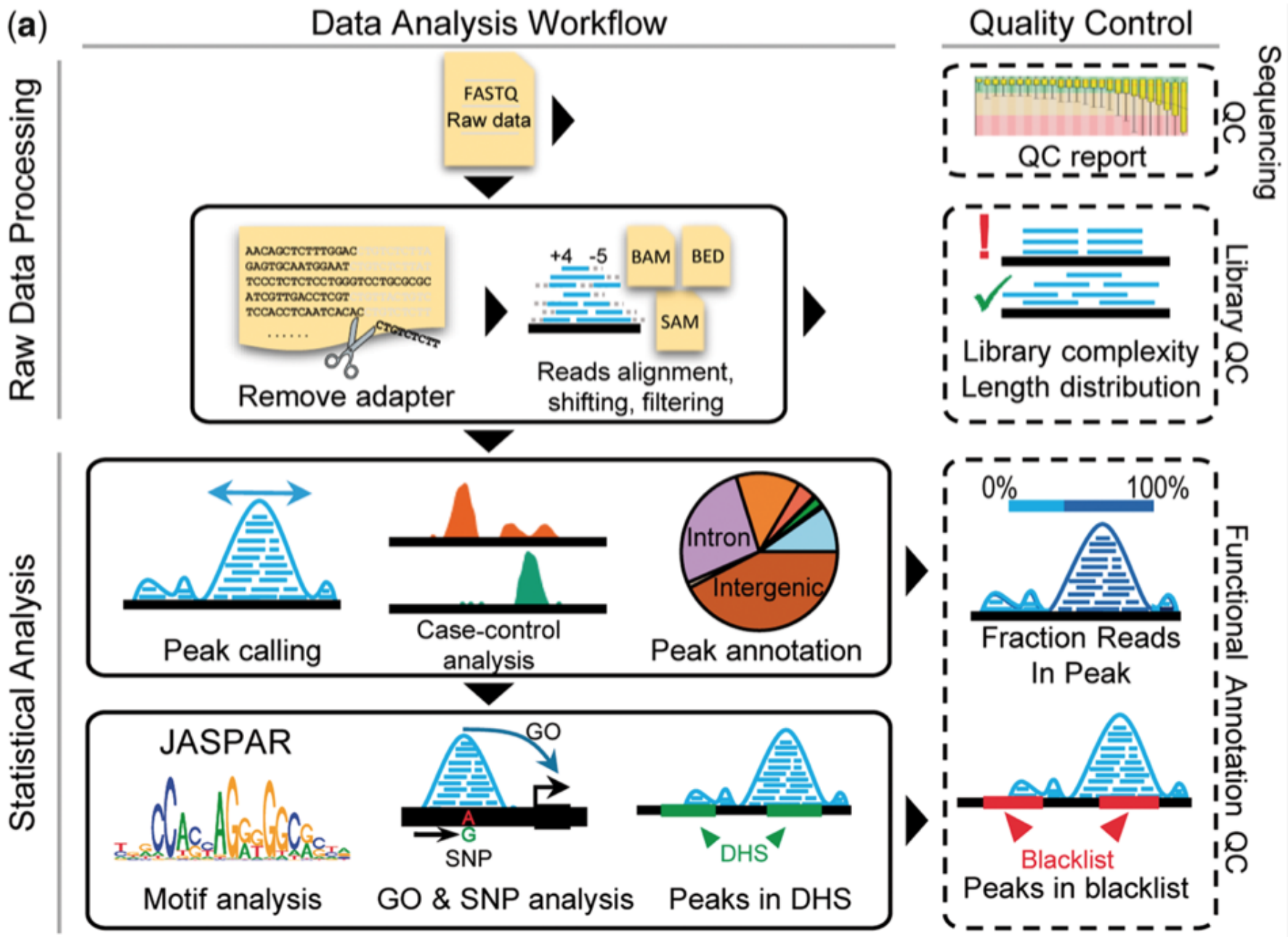
Blacklisted regions: regions with spurious signals

- typically appear uniquely mappable
- often found at specific types of repeats such as centromeres, telomeres and satellite repeats
- especially important to remove these regions *before* computing measures of similarity

Blacklists were generated empirically by the (mod)ENCODE consortium: http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/




```
bedtools intersect -abam reads.bam -b blacklisted.bed > filtered_reads.bam
```

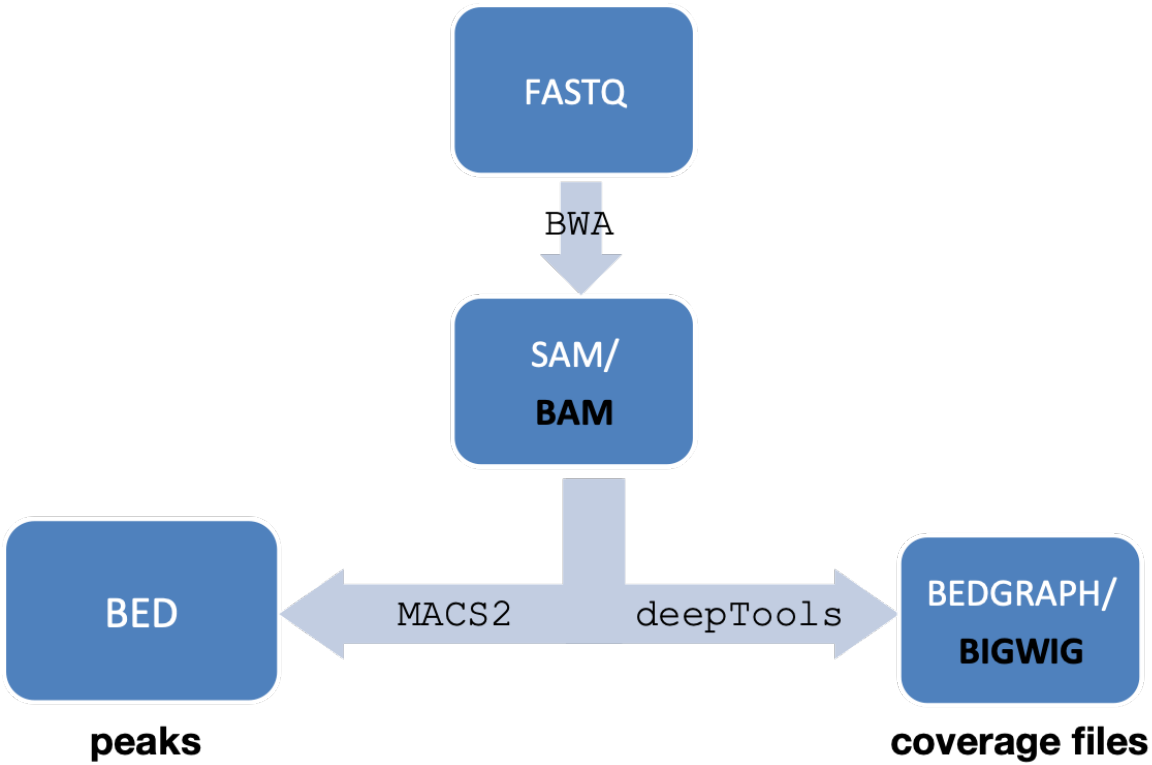


Wei et al. (2018). doi: 10.1093/bioinformatics/bty141

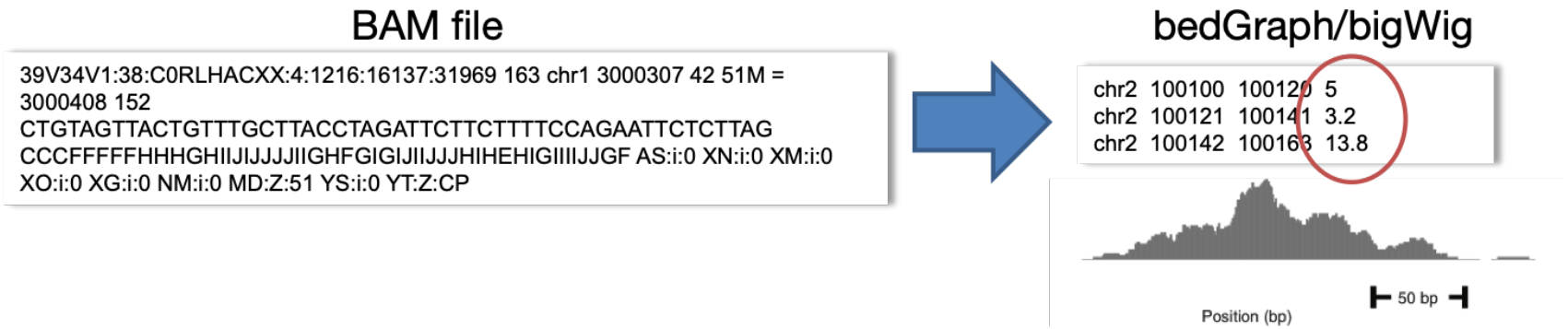
Checking the signal enrichment for ATAC-seq

Following the filtering of the BAM files, the next QC steps include:

- fraction of reads in **peaks** (**FRiP**)
- enrichments around active **TSS**
- visual inspection (genome browser!)



Checking the signal enrichment: generating coverage files

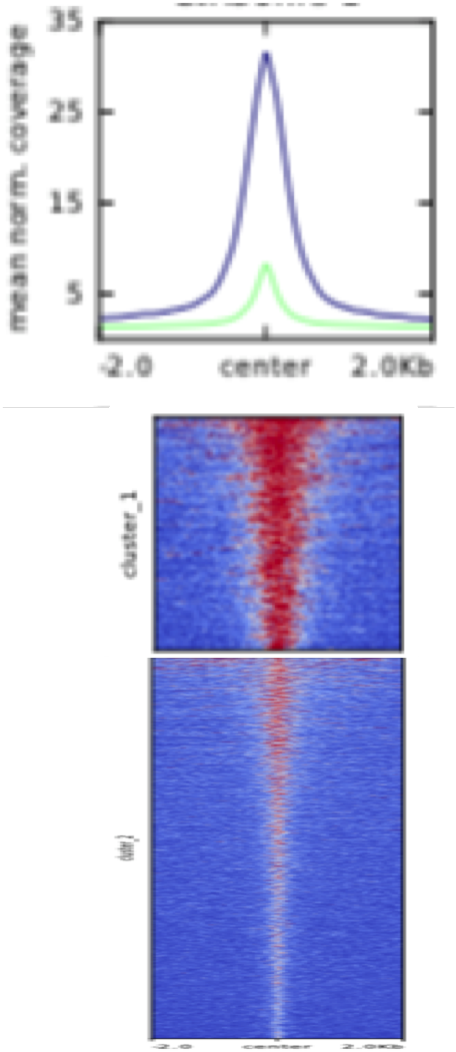


deepTools [Ramírez et al., 2016] offers the `bamCoverage` function that is fairly versatile and flexible

- check out the documentation!
- several types of normalization to account for sequencing depth differences
  - **RPGC** (reads per gen. content): $\frac{\textit{reads per bin}}{(\textit{all reads} * \textit{fragment length} / \textit{effective genome size})}$ recommended
  - **RPKM**: division by total number of reads

```
bamCoverage --bam a.bam -o a.SeqDepthNorm.bw --binSize 10 \
    --normalizeUsing RPGC --effectiveGenomeSize 2150570000 \
    --ignoreForNormalization chrX -minFragmentLength 40
```

Checking the signal enrichment: TSS focus



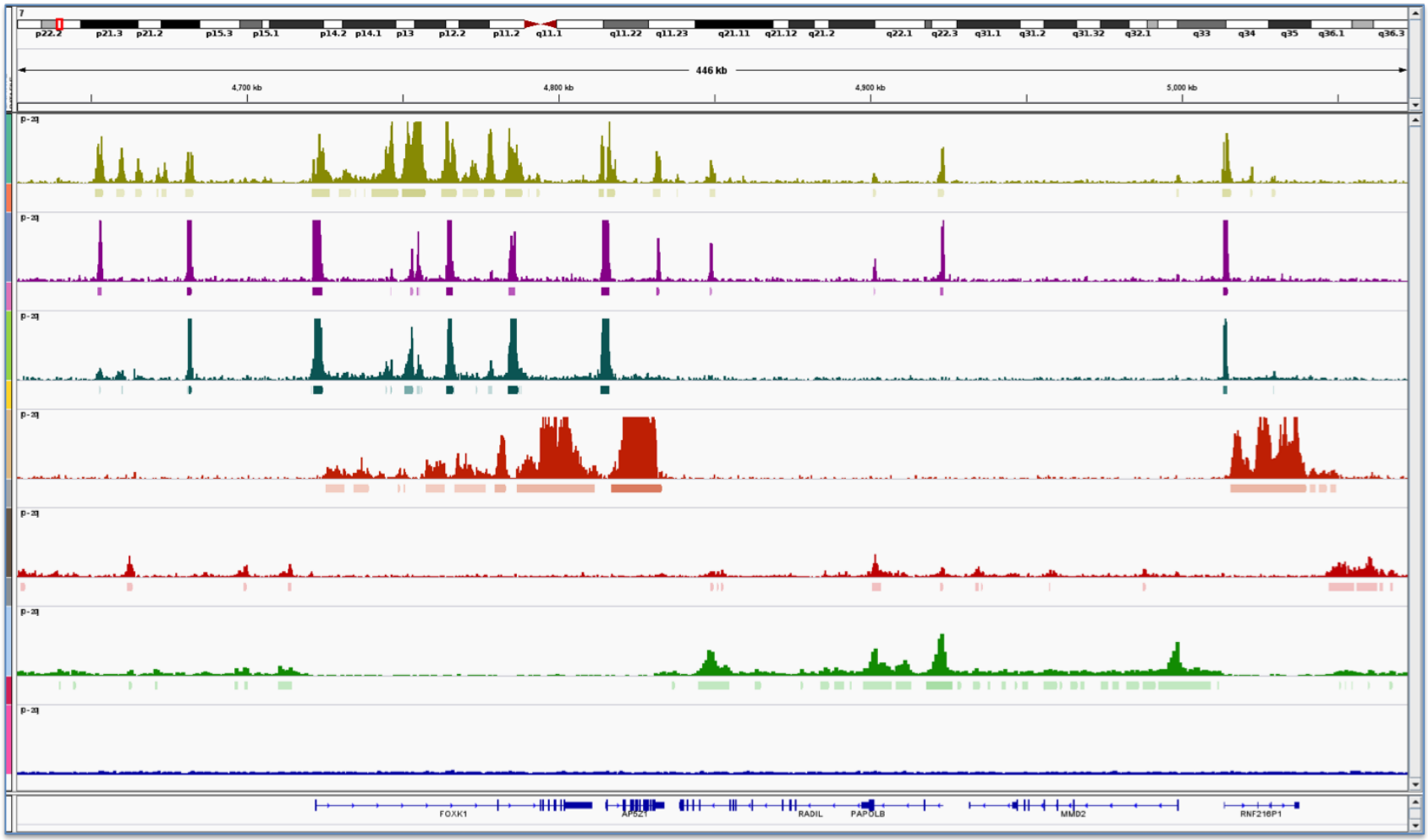
deepTools offers functions for visualizations of the bigWig files

```
$ computeMatrix reference-point \
 -S ATACseq.bigwig -R genes.bed \
 --referencePoint TSS \
 -a 2000 -b 2000 \ ## bp before and
                   # after refPoint
 -out ATAC_TSS.tab.gz

$ plotHeatmap -m ATAC_TSS.tab.gz \
  -out hm_ATAC.png \
 --heatmapHeight 15 \
 --refPointLabel center
```

Checking the signal enrichment: peak calling

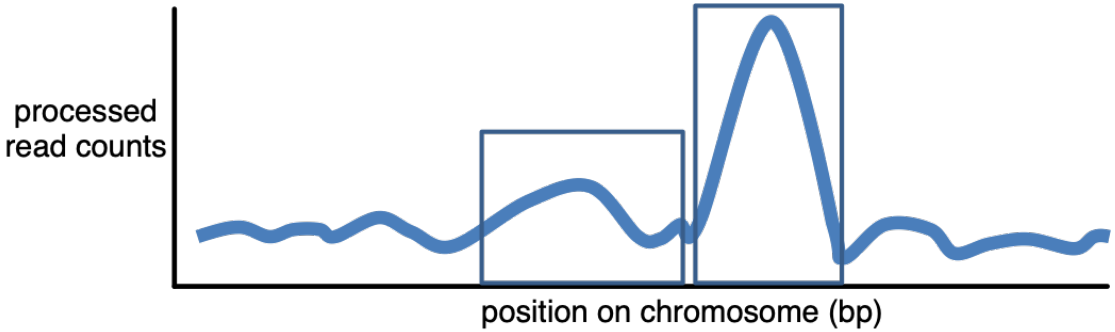
= identifying regions with higher read coverage than expected based on the background



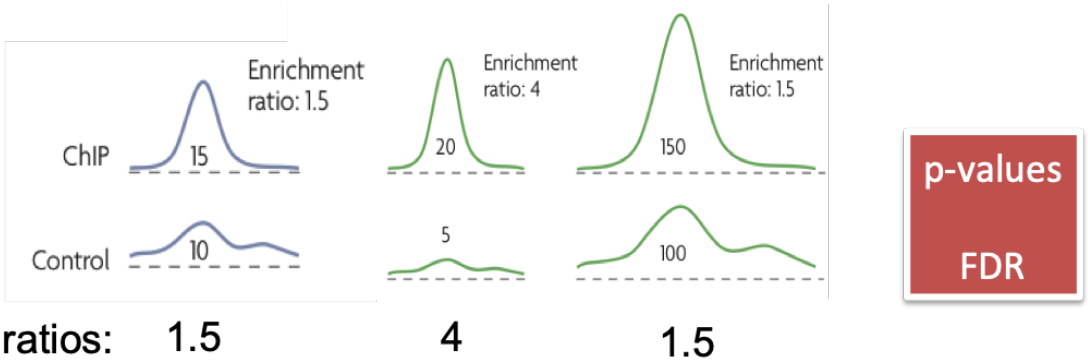
Checking the signal enrichment: peak calling

Starting from the BAM file:

1. generate a signal of *fragment* counts along the genome
2. identify regions of *enrichment*

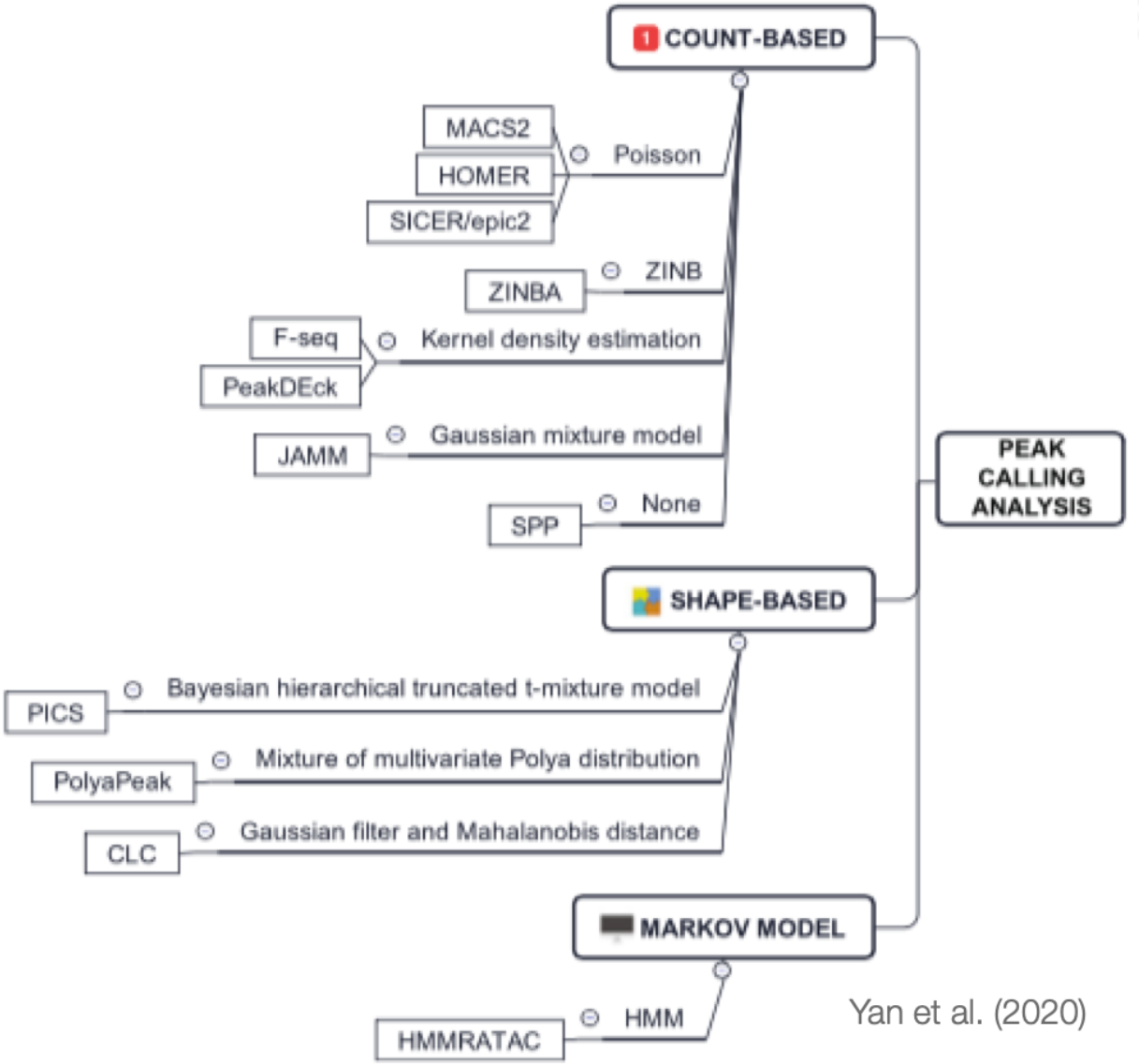


3. assess *significance* of enrichment



We usually use MACS [Zhang et al., 2008]; mostly because it's part of most pipelines, not because it's such a great tool (but it has proven itself to be fairly robust and useful).

Peak calling



Yan et al. (2020)

Peak calling

Identifying and assessing regions of **enrichment** with MACS

1. Sliding a window of length 2 x bandwidth (= half of estimated sonication size) across genome and determine read counts
2. Retain windows with counts > MFOLD (fold-enrichment of treatment/back-ground)
3. PEAKS: probability of an enrichment being stronger than expected
   - H0: reads are randomly distributed throughout the genome following a Poisson distribution
   - Determine the background distribution (λ) by sliding a window of size 2 x fragment size across the background to estimate the local coverage

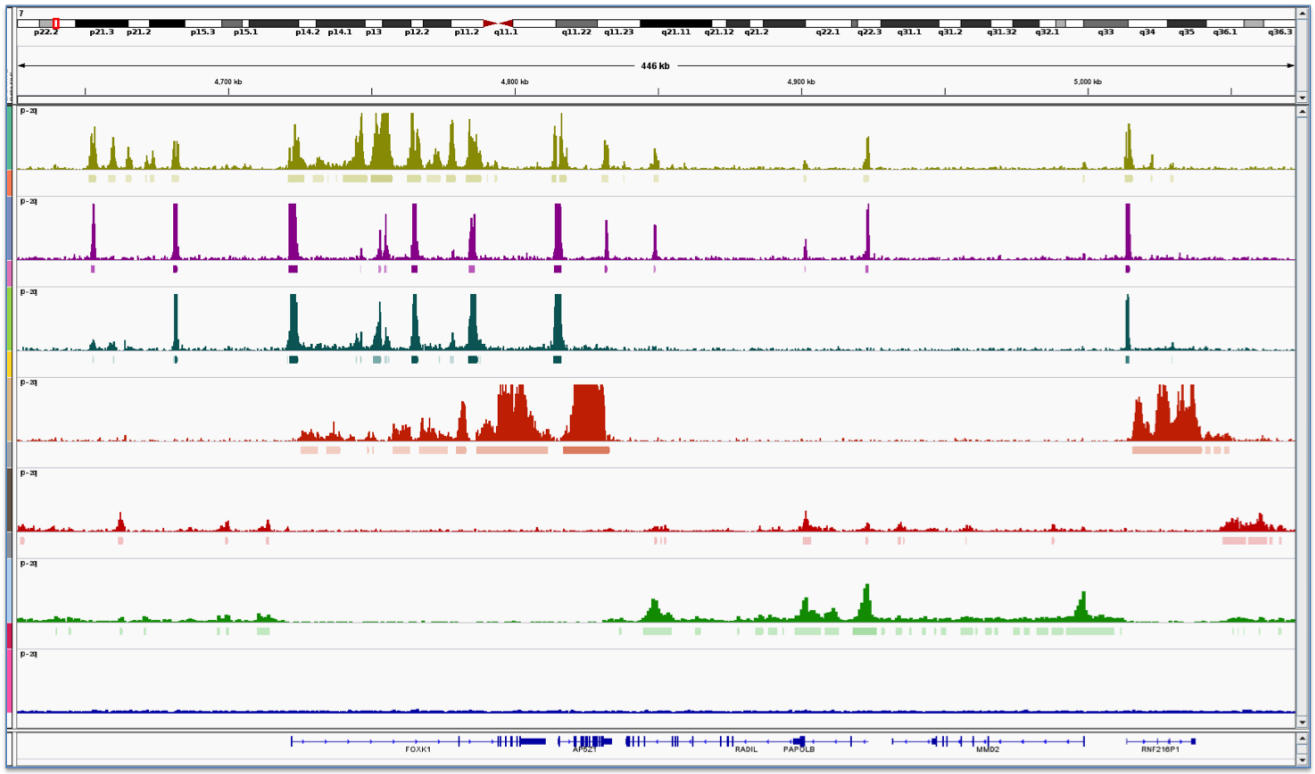
```
MACS2 callpeak -t pairedEnd.bam -f BAMPE --outdir path/to/output/ \
      --name pairedEndPeakName -g 2.7e9
```

See Tom Carroll's pipeline for detailed MACS2 commands.

The result of MACS is a BED file of regions with sign. enrichments, i.e. peaks.

Checking signal enrichments: FRiP

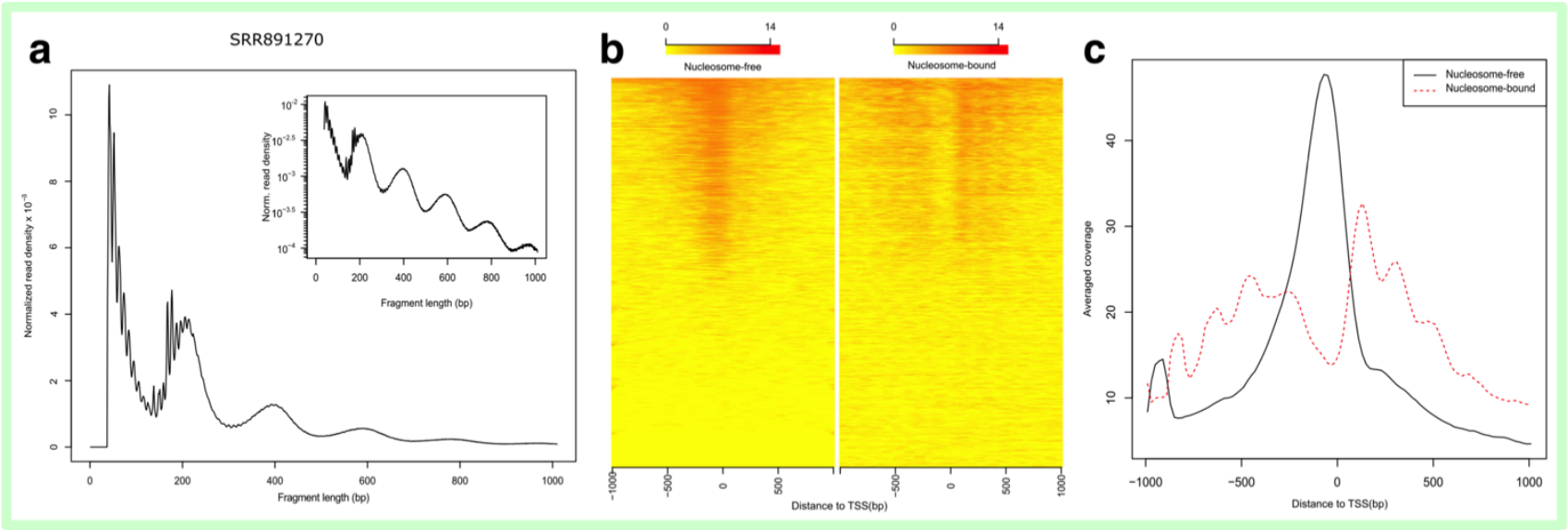
$$FRiP = \frac{reads\ in\ peaks}{total\ reads}$$

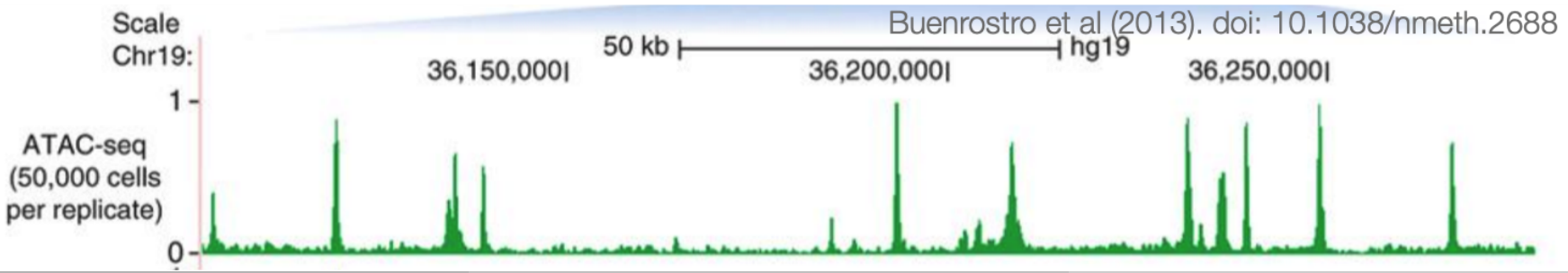


FRiP $>$ 0.3 is optimal; FRiP $>$ 0.2 acceptable by ENCODE standards.

QC checklist ATAC-seq

- fragments of 40 - 100 bp size should be over-represented
- 1/3 of the reads should fall into peaks (FRiP)
- very sharp and not too broad enrichments around TSS of active genes
- IGV snapshots: the signal should look sharp and high

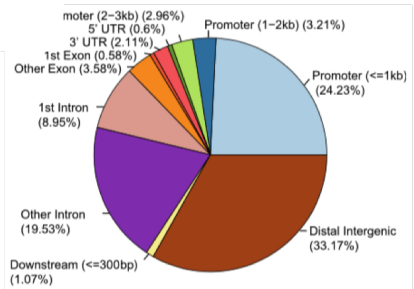


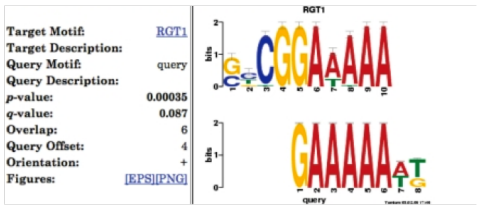


Typical downstream analyses following ATAC-seq peak identification

Peaks = regions of open chromatin

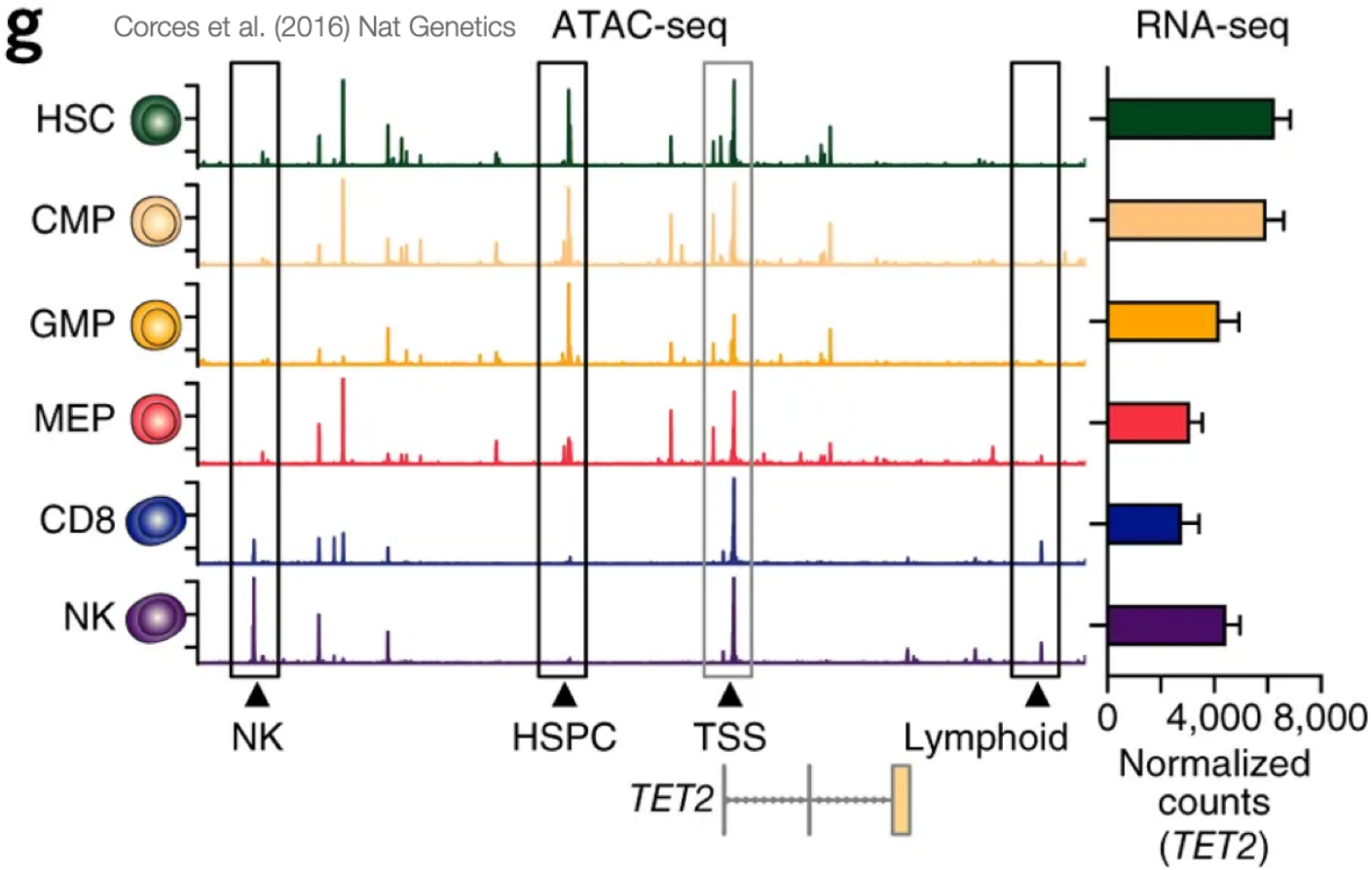
- annotation with **known genes**, i.e. do the peaks overlap with TSS/exons/introns?
  - `bedtools suite` [Quinlan, 2014], `ChIPpeakAnno` [Zhu et al., 2010], `ChIPseeker` [Yu et al., 2015]
- overlap with known **enhancers**, e.g. via `GREAT` McLean et al. [2010]
- **motif** analysis – difficult without additional information b/c TFBS motifs are often very short and exceedingly frequent throughout the genome
  - `MEME suite`: de novo motif detection & motif enrichment analysis





Open chromatin != expression

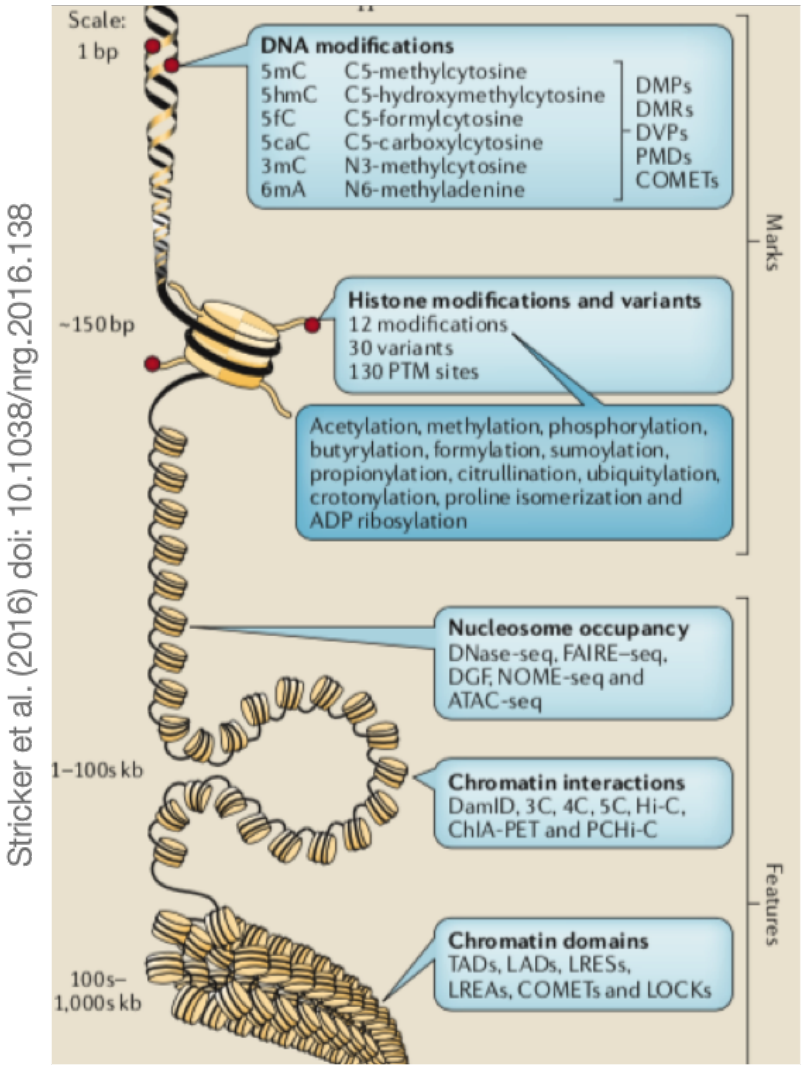
Correlating open chromatin regions with specific gene expression is not straight-forward (except for the TSS, perhaps).



Despite heterogenous chromatin accessibility across the different cell types, the TET gene is constitutively expressed throughout.

ChIP-seq principles

NGS techniques for studying chromatin and DNA modifications



The majority of epigenomics data entails profiles of **nucleosome occupancy**, specific **histone marks** and **transcription factor** binding.

These information are all inferred based on **which DNA sequences** we find **over-represented** in our data set.

NGS techniques for studying chromatin and DNA modifications

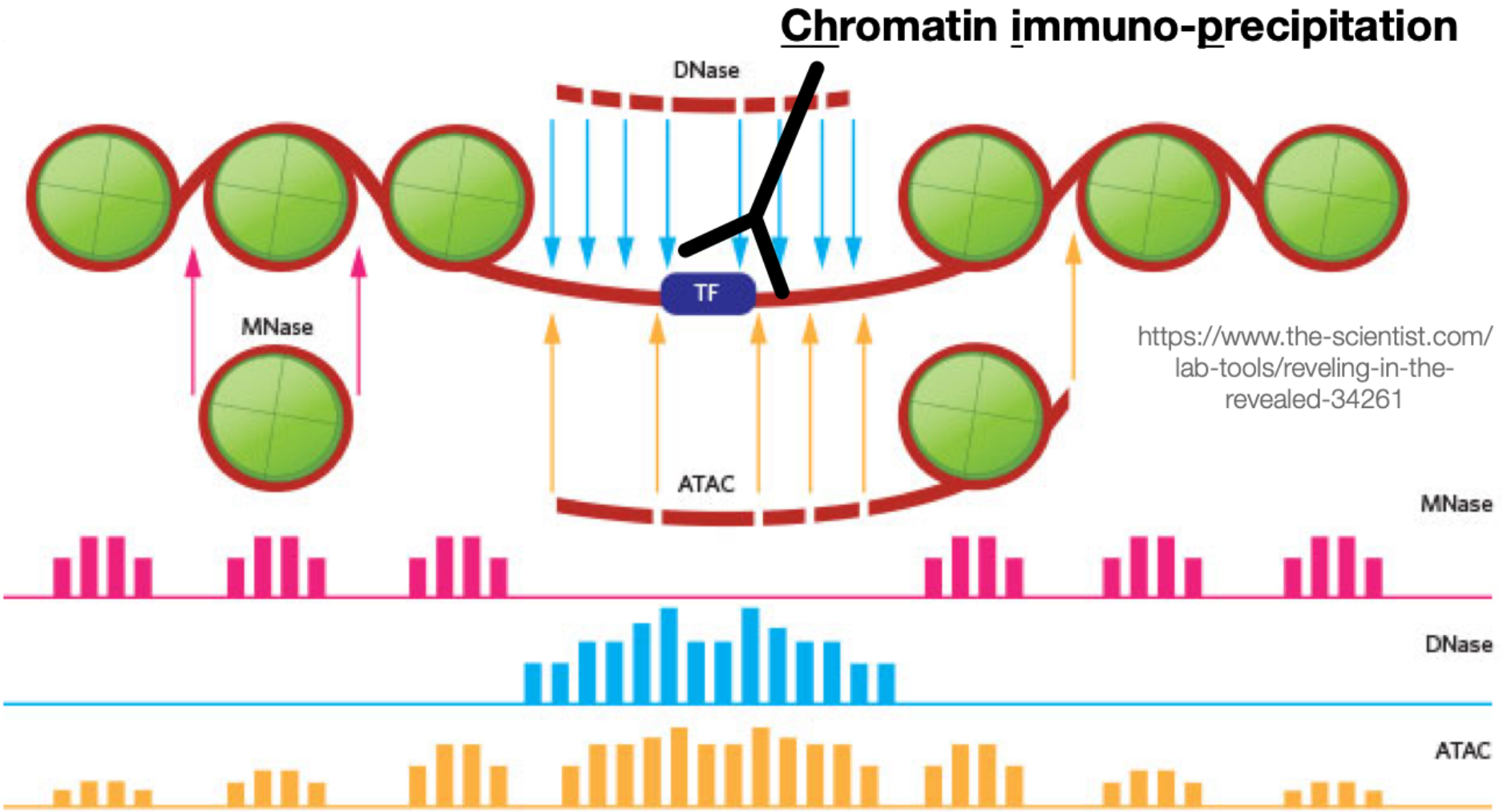
Depending on the type of insights you're interested in, there are different ways of *enrichment*.

How to enrich for the NA	Biological insights	Example technique
Nuclease susceptibility	nucleosome packaging	DNase-seq, MNase-seq
	regulatory regions	ATAC-seq
Affinity-based enrichments	protein-DNA interactions	ChIP-seq
	histone modifications	
	protein-RNA interactions	CLIP-seq
	chromatin-chromatin interactions	ChIA-PET
	RNA modifications	m6A-seq, MeRIP-Seq,
Proximity ligation	chromatin-chromatin interactions	3C, Hi-C, ChIA-PET, ...
Chemical susceptibility	DNA modifications	WGBS, RRBS

Table based on Friedman and Rando [2015]

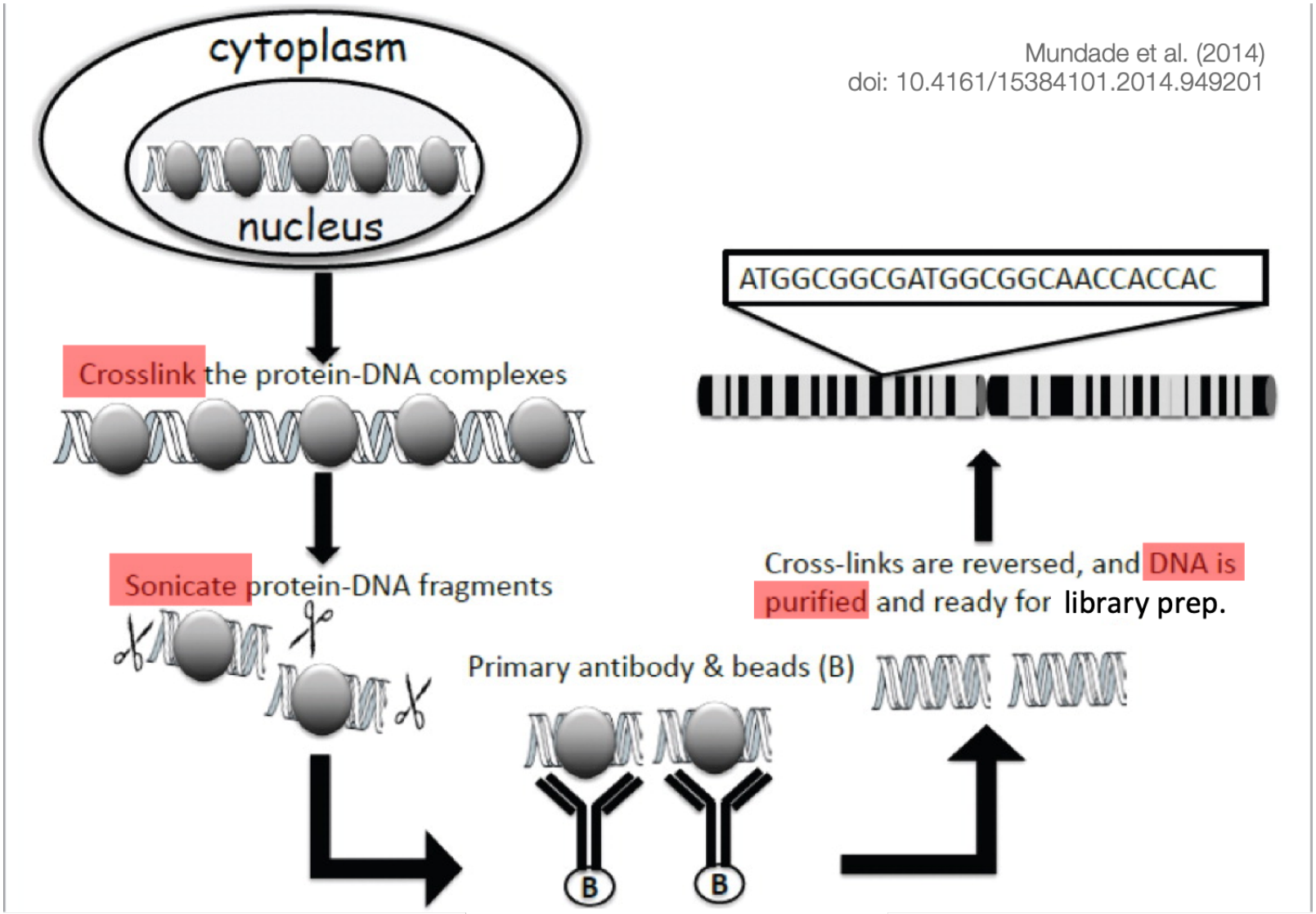
Identifying transcription factor binding sites with ChIP

(The "chromatin" in ChIP just means "any protein interacting with DNA")



The vast majority of TFBS has been found in regions of open chromatin.

Extracting DNA sites bound by a TF of interest

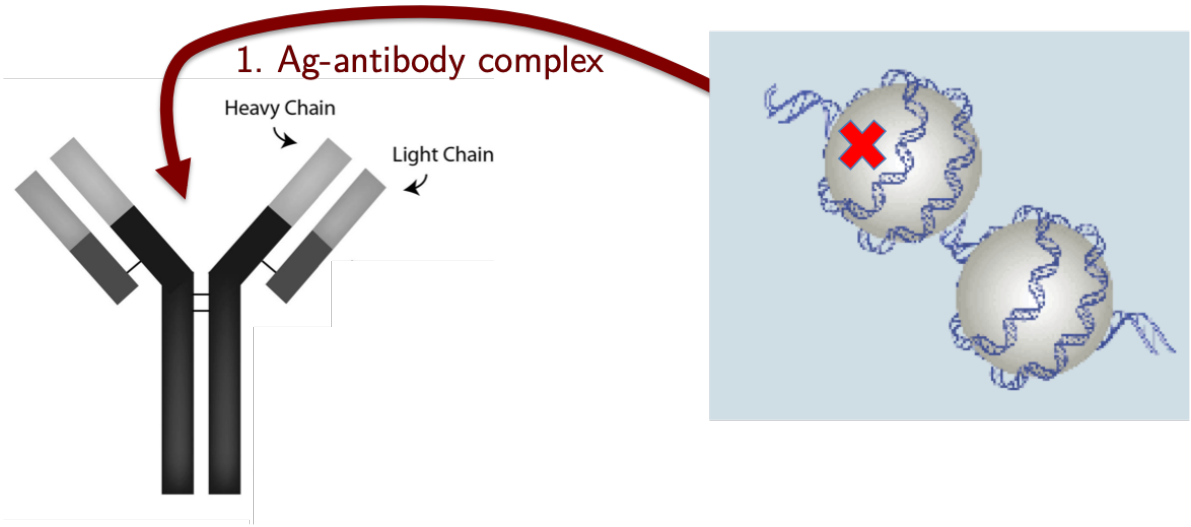


Extracting DNA sites bound by a TF of interest

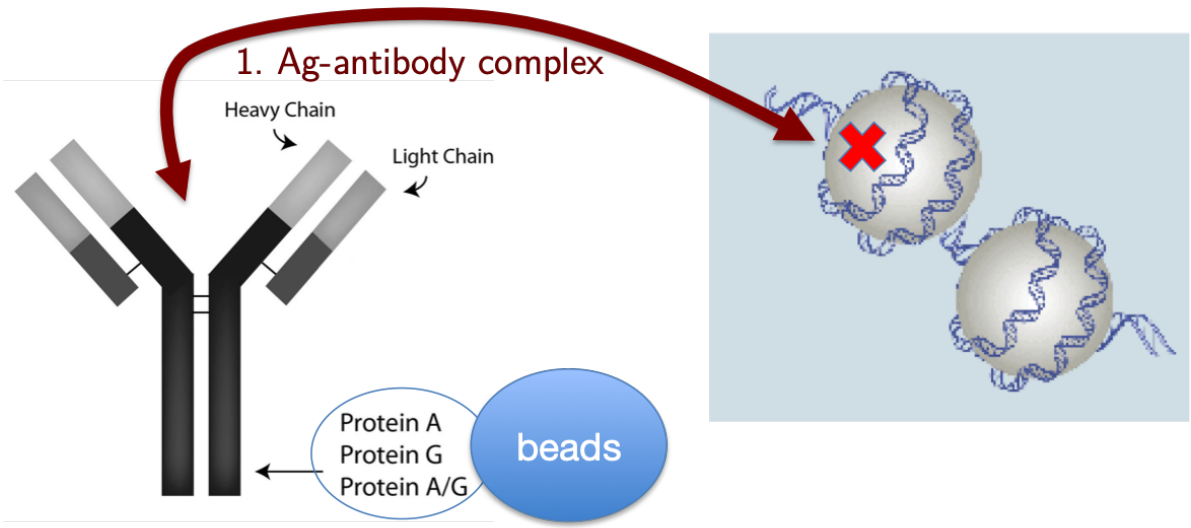
Principles of immunoprecipitation

- based on the principle of **antibody-antigen** interaction: antibody is incubated with cell lysates that contain the target protein bound to DNA
- the DNA-protein-antibody complex is then captured by **antibody-binding proteins** that are attached to magnetic beads
- the DNA bound to the initial target protein can then be eluted from the beads for further analysis.

Extracting DNA sites bound by a TF of interest: principles of immunoprecipitation

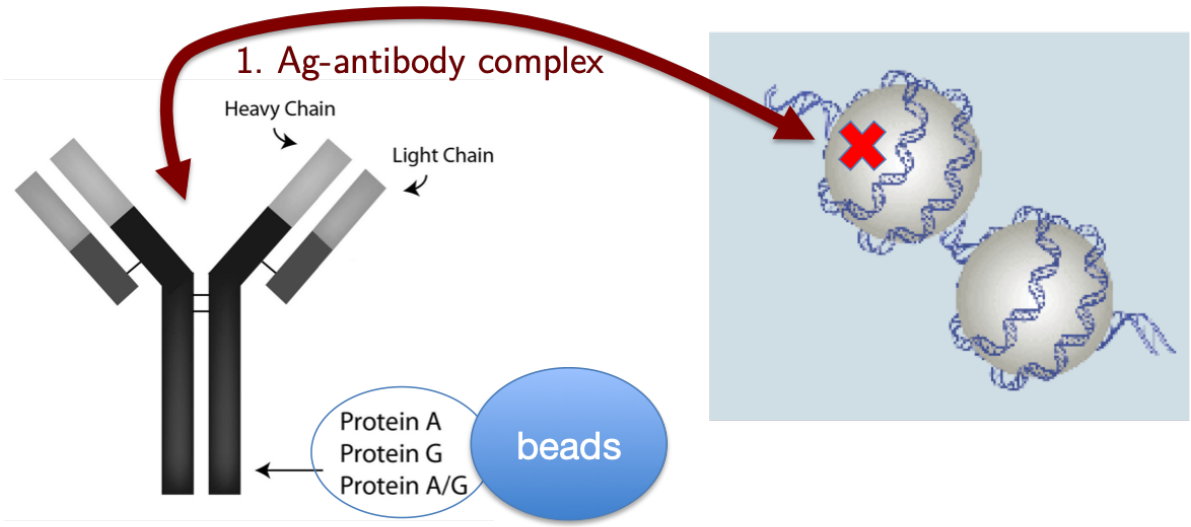


Extracting DNA sites bound by a TF of interest: principles of immunoprecipitation

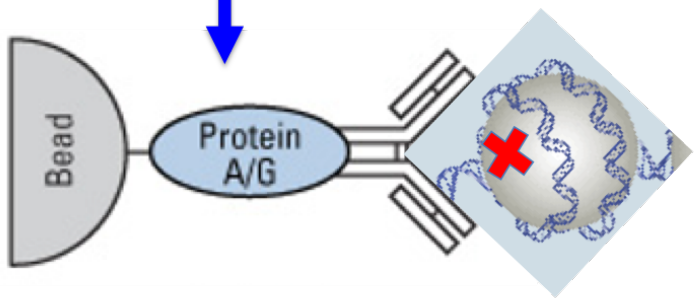


2. Ag-antibody complex bound by an antibody-binding protein (e.g. protein A) + beads

Extracting DNA sites bound by a TF of interest: principles of immunoprecipitation

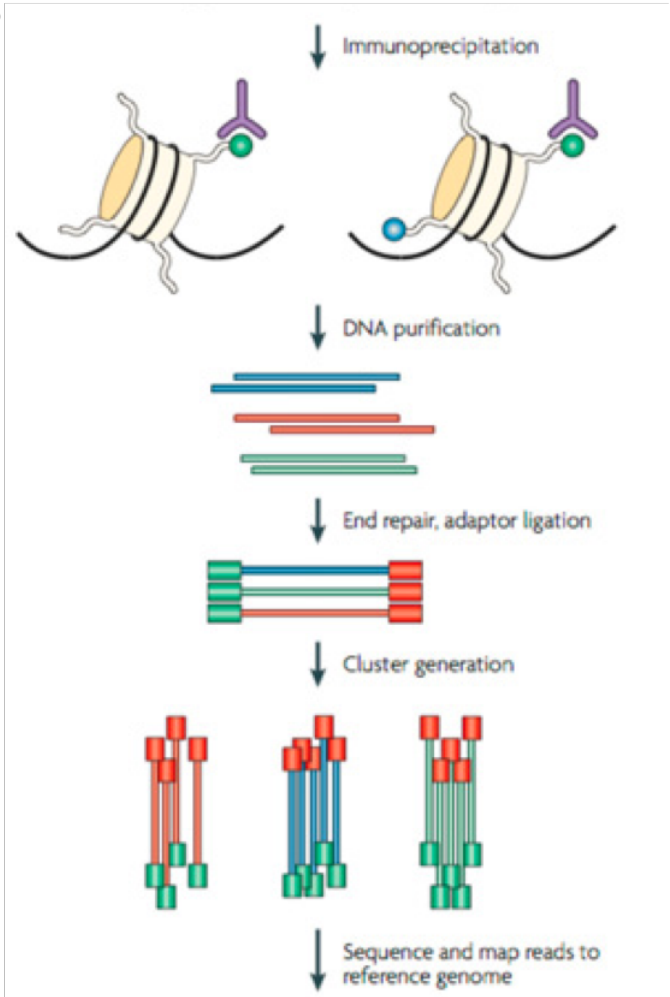


2. Ag-antibody complex bound by an antibody-binding protein (e.g. protein A) + beads



3. Ag-ab-protA complex is purified & DNA is subsequently released

ChIP + NGS = ChIP-seq



http://compbio.pbworks.com/f/1210361096/Histone%20Mod%205-1.jpg

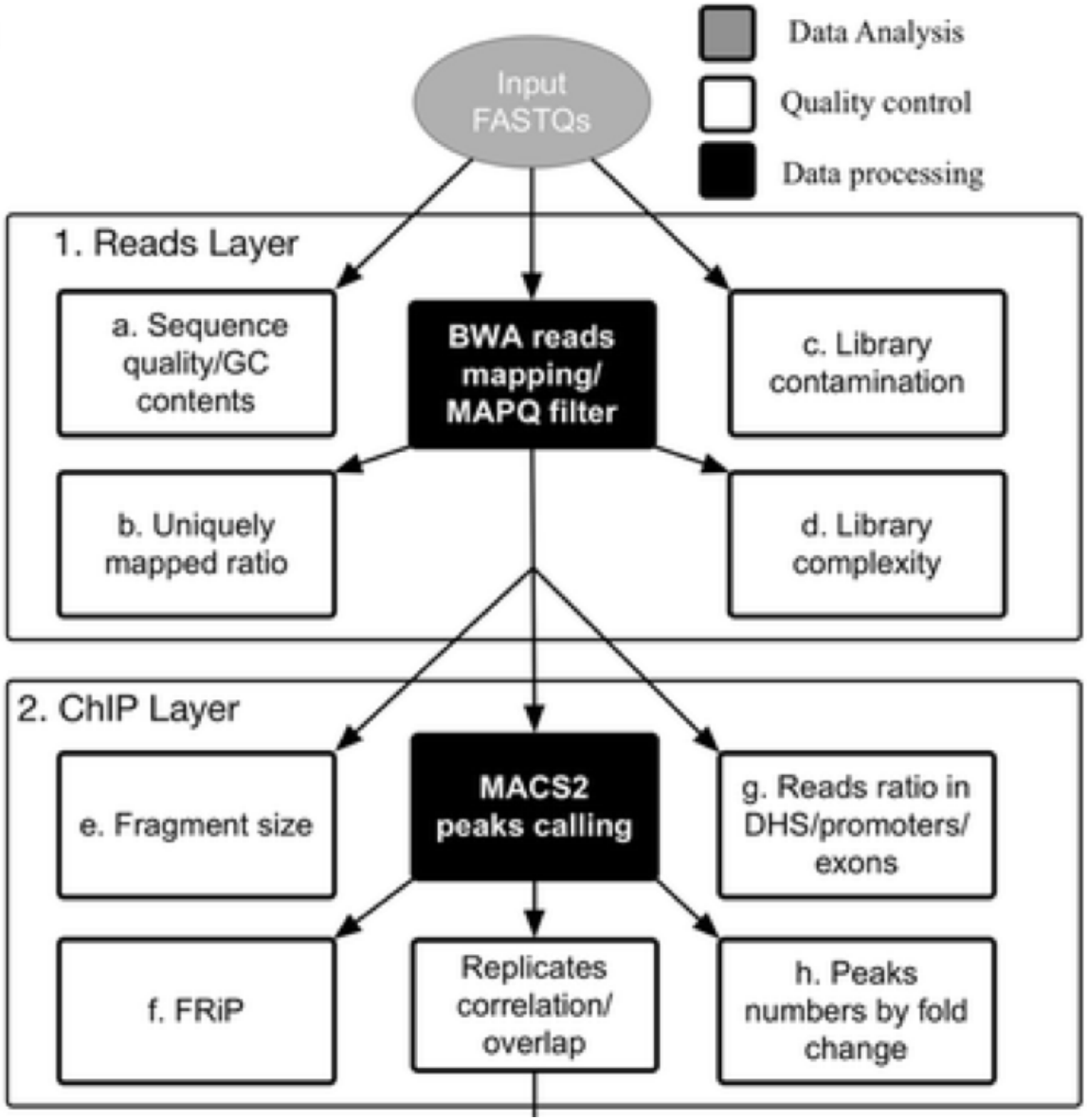
Immunoprecipitation (= *enrichment of DNA bound to the protein of interest*) is followed by high-throughput sequencing of the recovered DNA fragments.

In contrast to ATAC-seq, nobody would say ChIP-seq was "easy"

- **cross-linking** is a frequent source of bias
  - too short $\rightarrow$ proteins will be lost during the sonication
  - the longer the fixation, the more proteins are artificially linked with DNA ("non-specific capturing of reactive soluble proteins" [Baranello et al., 2016])
- **sonication** can be fickle and inherently favors open chromatin regions
- ChIP depends on **antibodies**
  - expensive! (typically 1 vial of antibody per experiment)
  - cross-reactivity: the antibody may bind to more than just the protein of interest
  - successful binding needs incredibly optimized conditions
  - signal-to-noise ratio will depend on how abundantly the protein of interest binds to DNA
- entire protocol takes 3-4 days to complete (before sequencing!)
- requires lots of cells (1-10 mio)

See, for example, Jordán-Pla and Visa [2018] for how to optimize ChIP experiments.

Processing of ChIP-seq data



Qin et al. (2016). doi: 10.1186/s12859-016-1274-4

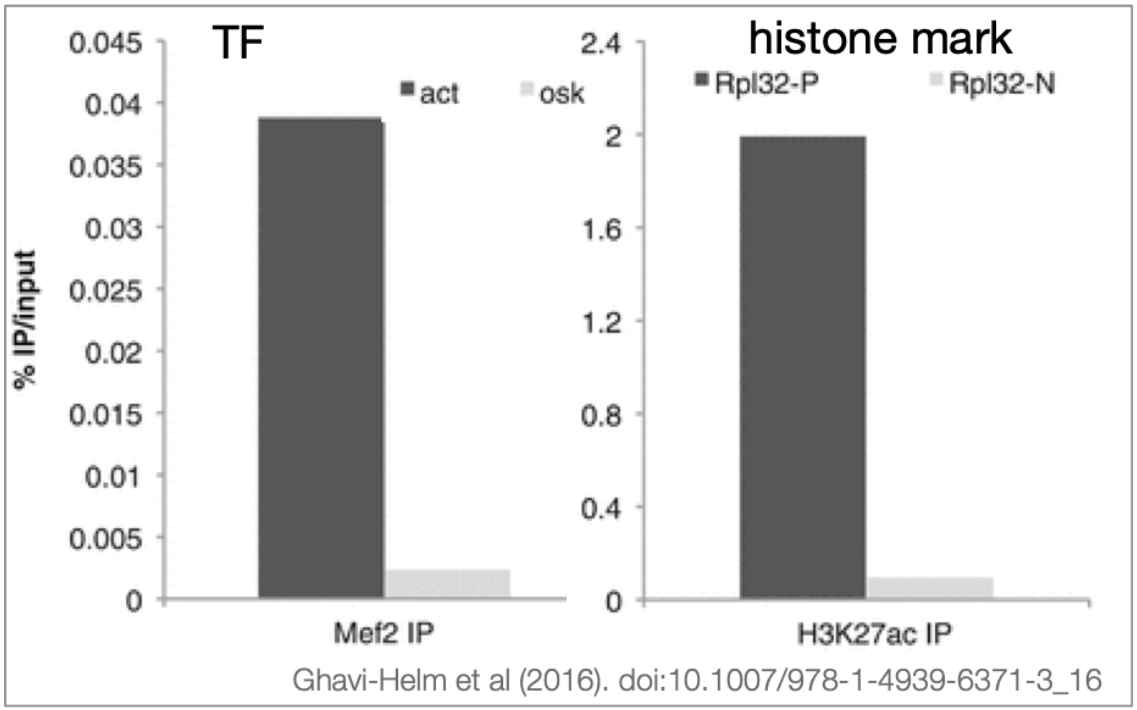
many basic processing steps are the same for ATAC- and ChIP-seq data, but some QC scores differ

Reads within FASTQ files correspond to the captured **DNA**, i.e. pieces captured by the antibody (e.g. against a TF) as well as all the background DNA. In fact, the vast majority will be representative of the entire genome (>95%).

Alignment is necessary prior to the identification of regions where more reads than expected by chance are found (**quantification** and **statistical assessment** = **peak calling**).

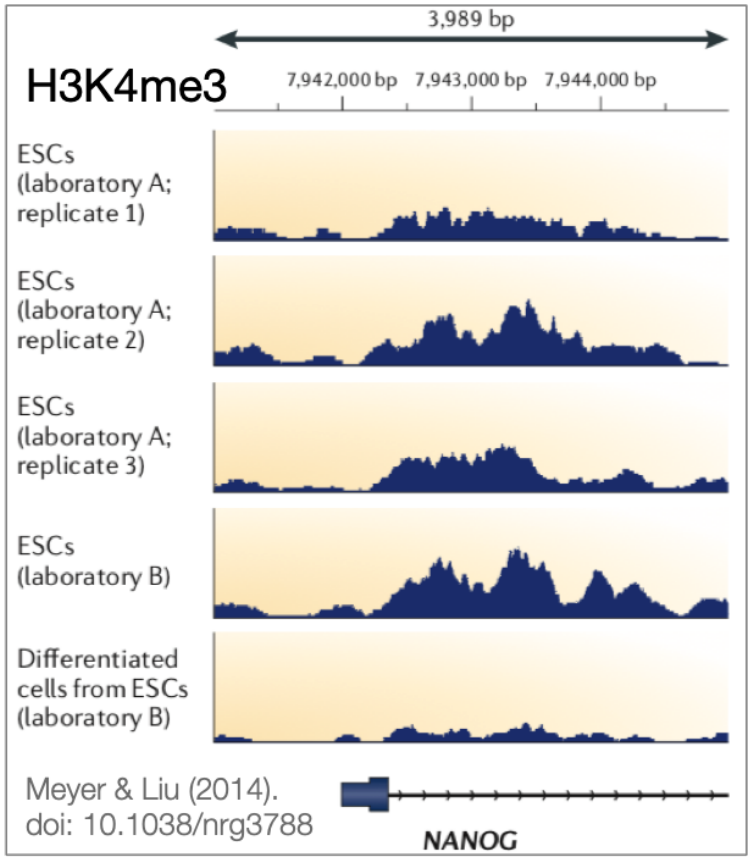
ChIP enrichments are often marginal and variable across experiments

TF often yield well below (!) 1% enrichment, histone marks usually below 10% (check the y-axis here!)

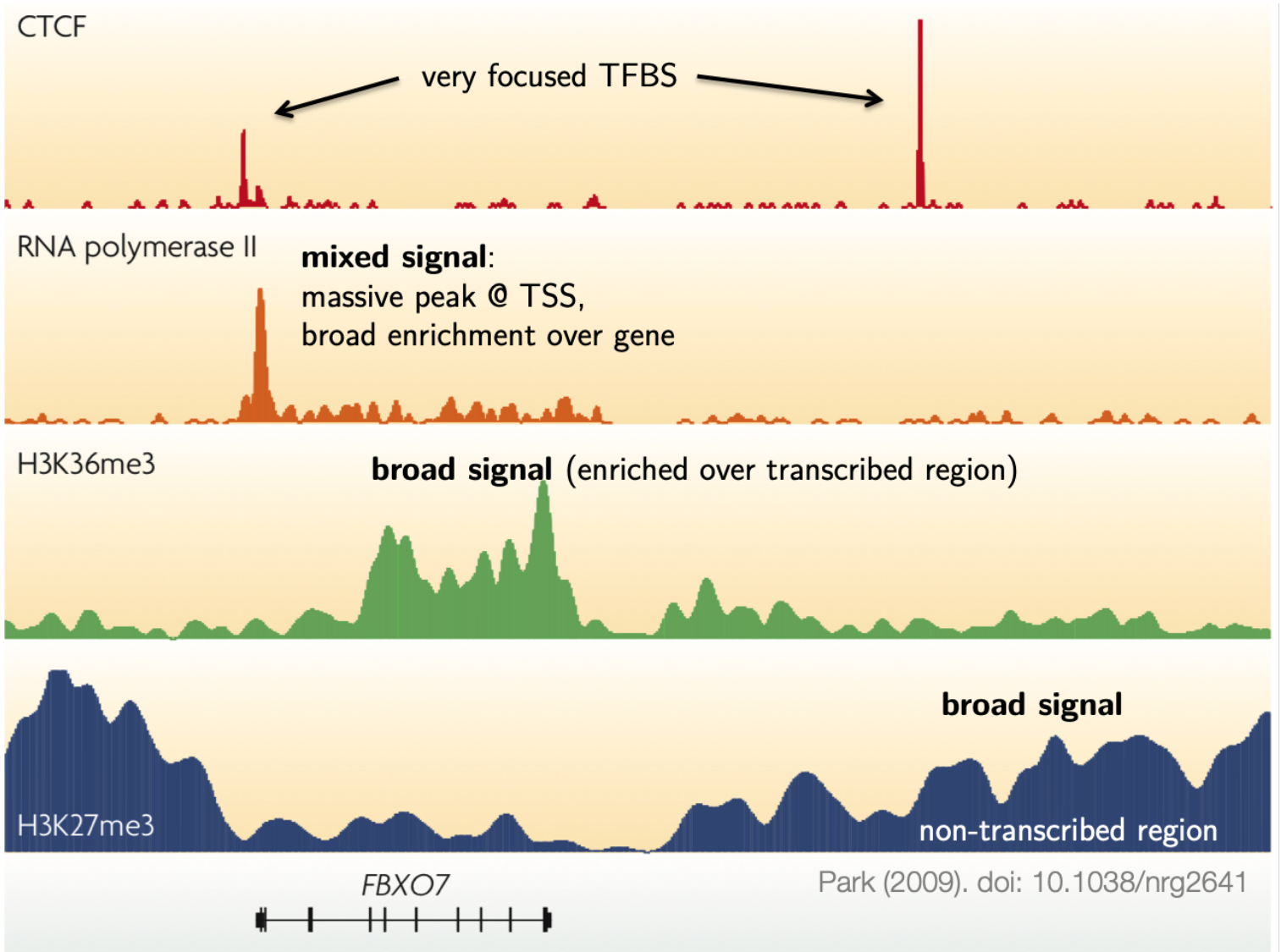


Ghavi-Helm et al (2016). doi:10.1007/978-1-4939-6371-3_16

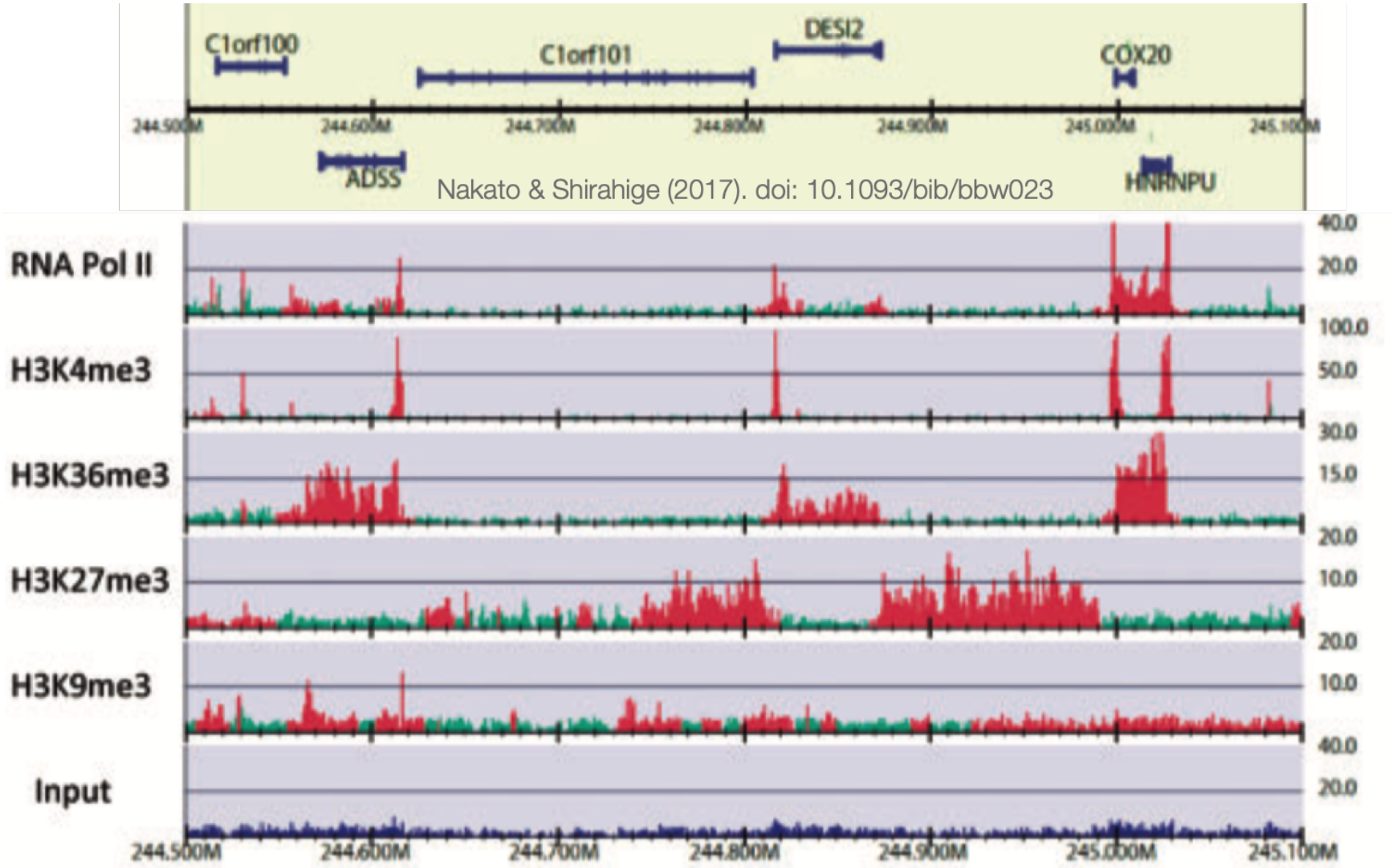
the same histone mark (same antibody) done in different labs



Different types of ChIP'ed factors will yield different types of signals (idealized version)



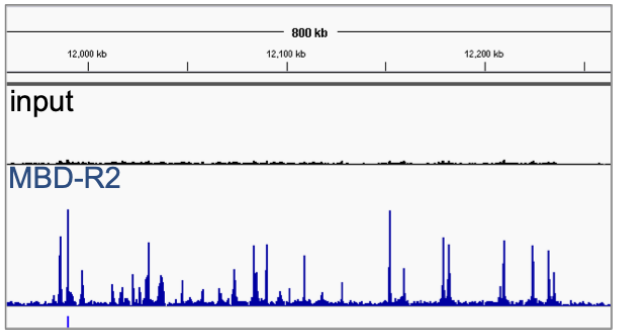
Different types of ChIP'ed factors will yield different types of signals (real-life example)



Peak calling: different ChIP'ed factors require different peak callers

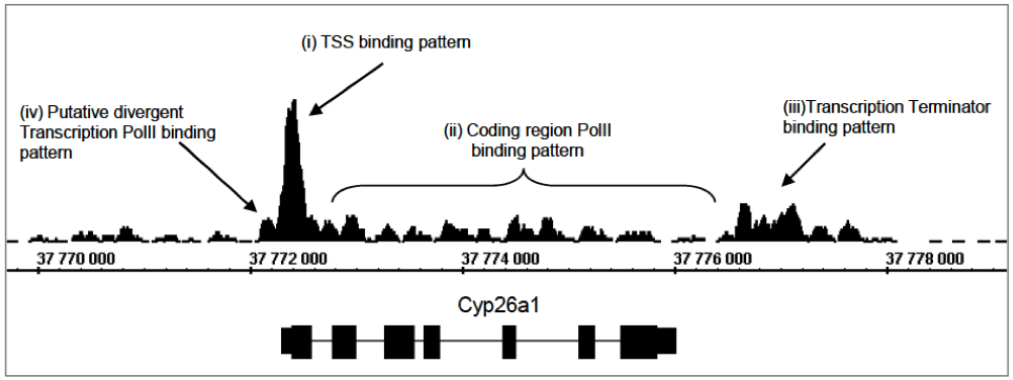
Identifying peaks for sharp, narrow, high enrichments is easy (⇒ MACS).
Assigning stats to broad enrichment is still an unsolved issue.

narrow & strong e.g. MACS



broad signal

mixed signal



See Wilbanks et al. [2010] and Thomas et al. [2017] for evaluations of peak callers.

Peak calling: different ChIP'ed factors require different peak callers

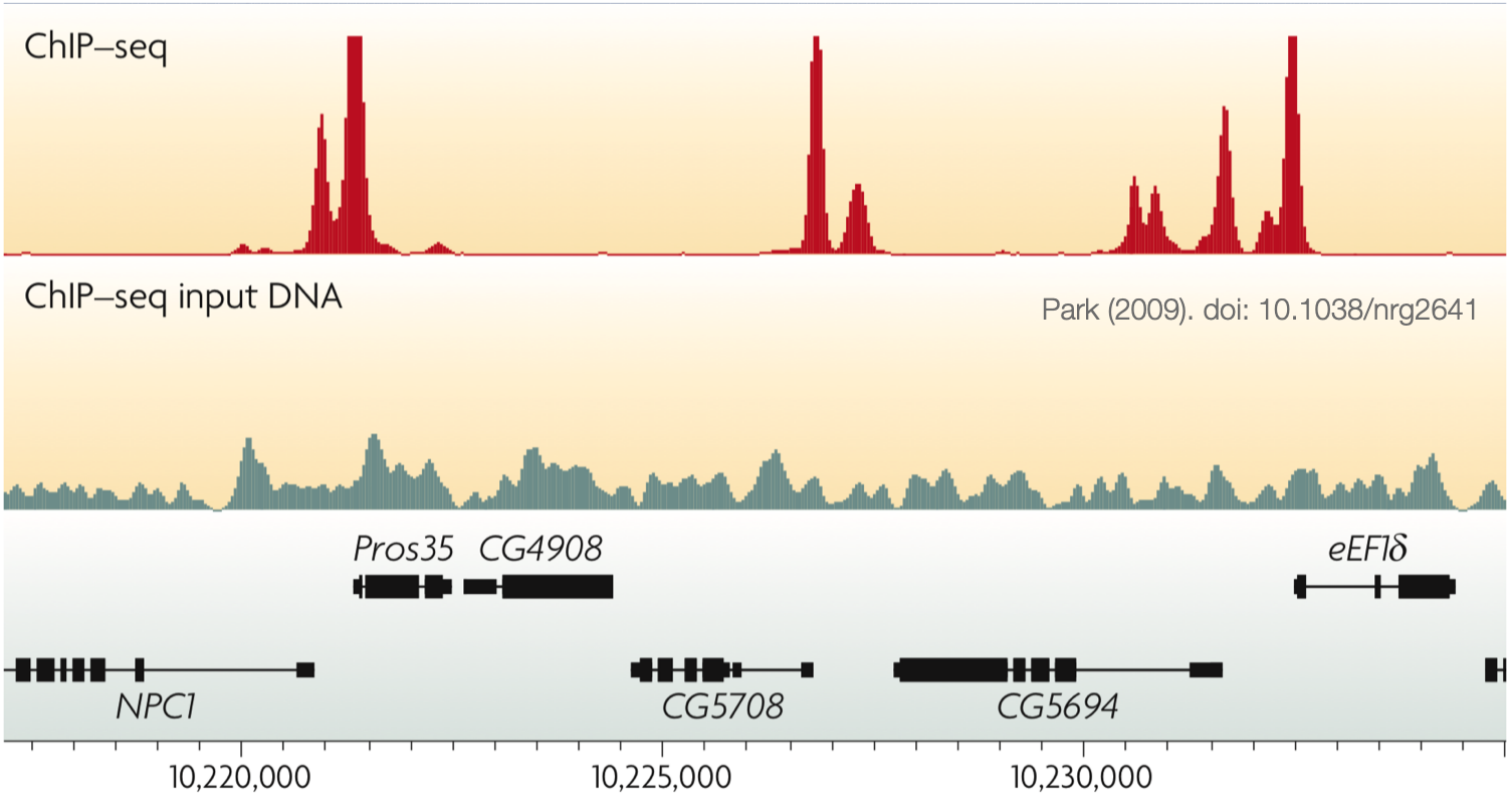
Identifying peaks for sharp, narrow, high enrichments is easy ($\Rightarrow$ MACS).
Assigning stats to broad enrichment is still an unsolved issue.

★ Comprehensive list is at: https://omictools.com/peak-calling-category

MACS2 *(MACS1.4)*	Most widely used peak caller. Can detect narrow and broad peaks.
Epic *(SICER)*	Specialised for broad peaks
BayesPeak	R/Bioconductor
Jmosaics	Detects enriched regions jointly from replicates
T-PIC	Shape based
EDD	Detects megabase domain enrichment
GEM	Peak calling and motif discovery for ChIP-seq and ChIP-exo
SPP	Fragment length computation and saturation analysis to determine if read depth is adequate.

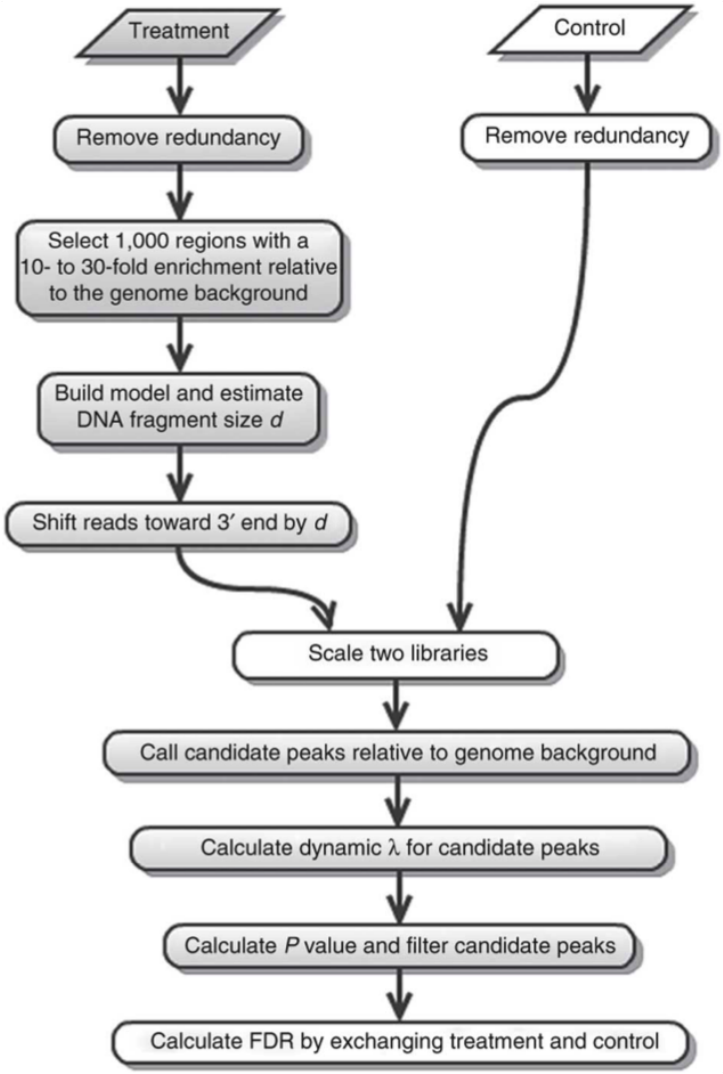
ChIP experiment absolutely require an "input" control

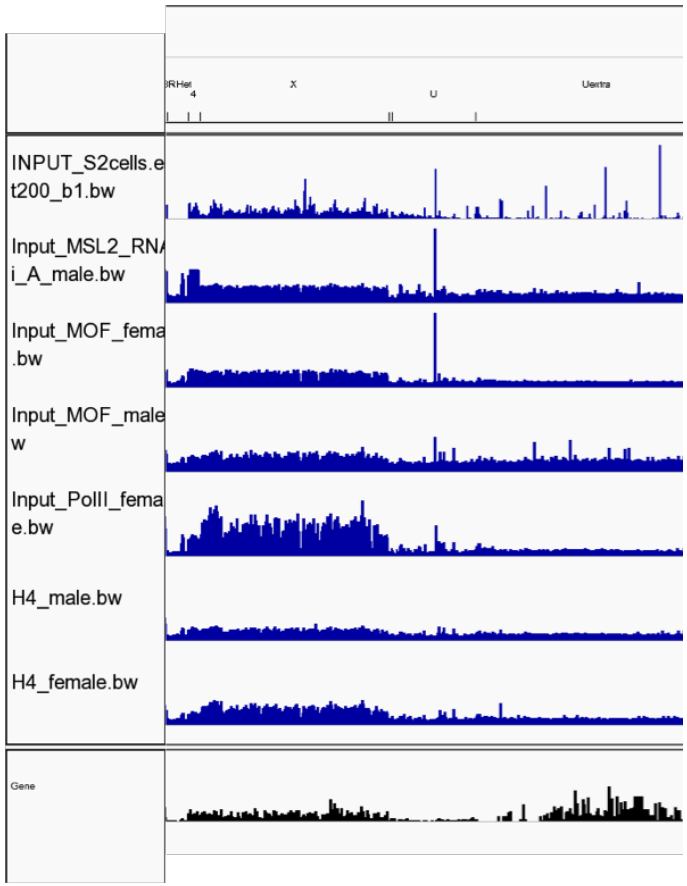
"Input" = the ChIP experiment without the antibody addition



Ideally, input samples should be done in parallel with the ChIP experiments; they should also be sequenced at least as deeply or **more deeply sequenced** than the ChIP samples.

Peak calling: take input samples into consideration!



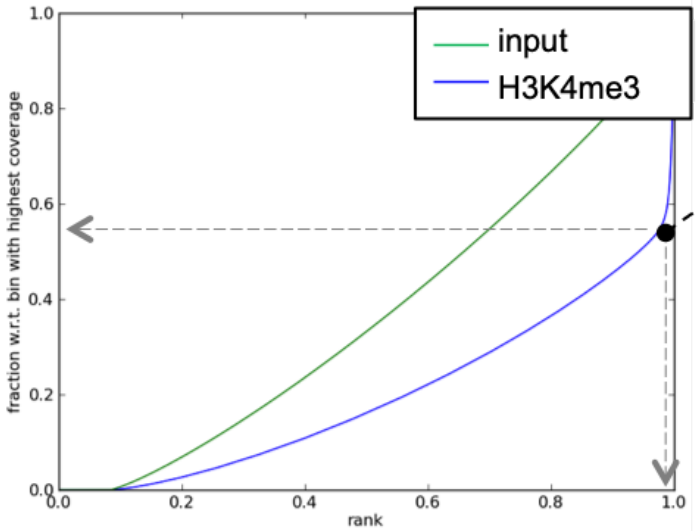


Consider the bioconductor package `GreyListChIP` to define cell-type-specific regions of input biases.

Signal check: fingerprints instead of FRiP

How well can signal & background be separated?

A very specific and strong ChIP enrichment will be indicated by a prominent and steep rise of the cumulative sum towards the highest rank. This means that a big chunk of reads from the ChIP sample is located in few bins which corresponds to high, narrow enrichments typically seen for transcription factors.

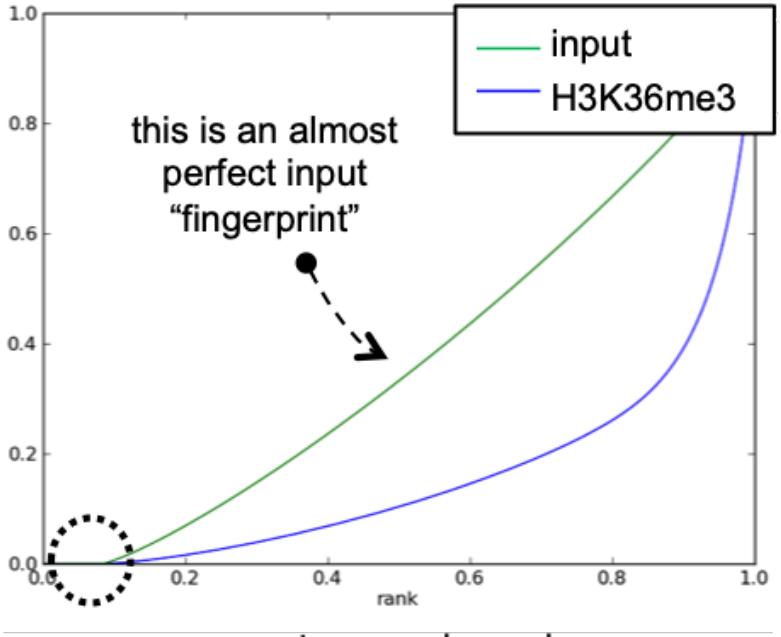


when counting the reads contained in 97% of all genomic bins, only 55% of the maximum number of reads are reached, i.e. 3% of the genome contain a very large fraction of reads!

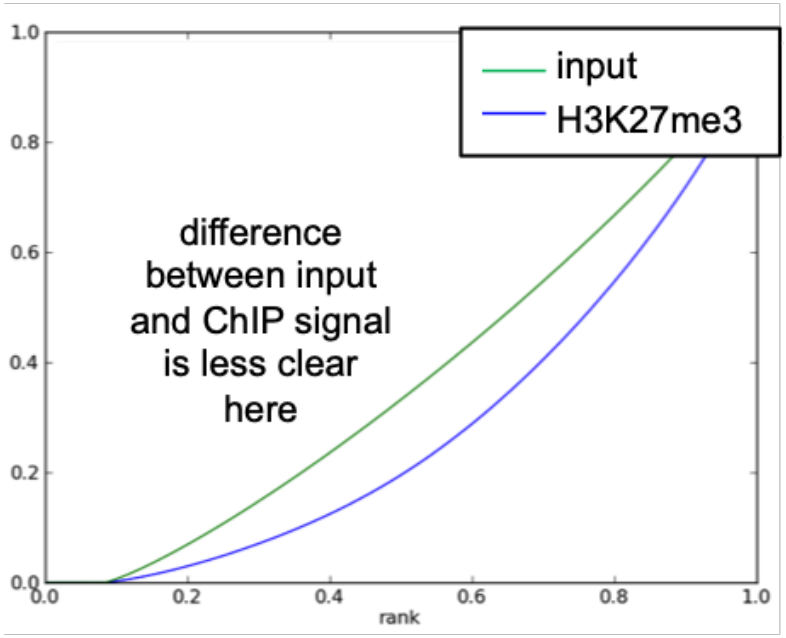
this indicates very **localized**, very **strong** enrichments! (as every biologist hopes for in a ChIP for H3K4me3)

```
## another deepTools function
$ plotFingerprint -b testFiles/*bam --labels H3K4me3 H3K4me1 H3K27me3 \
  --plotFile fingerprints.png --outRawCounts fingerprints.tab
```

Signal check: fingerprints instead of FRiP



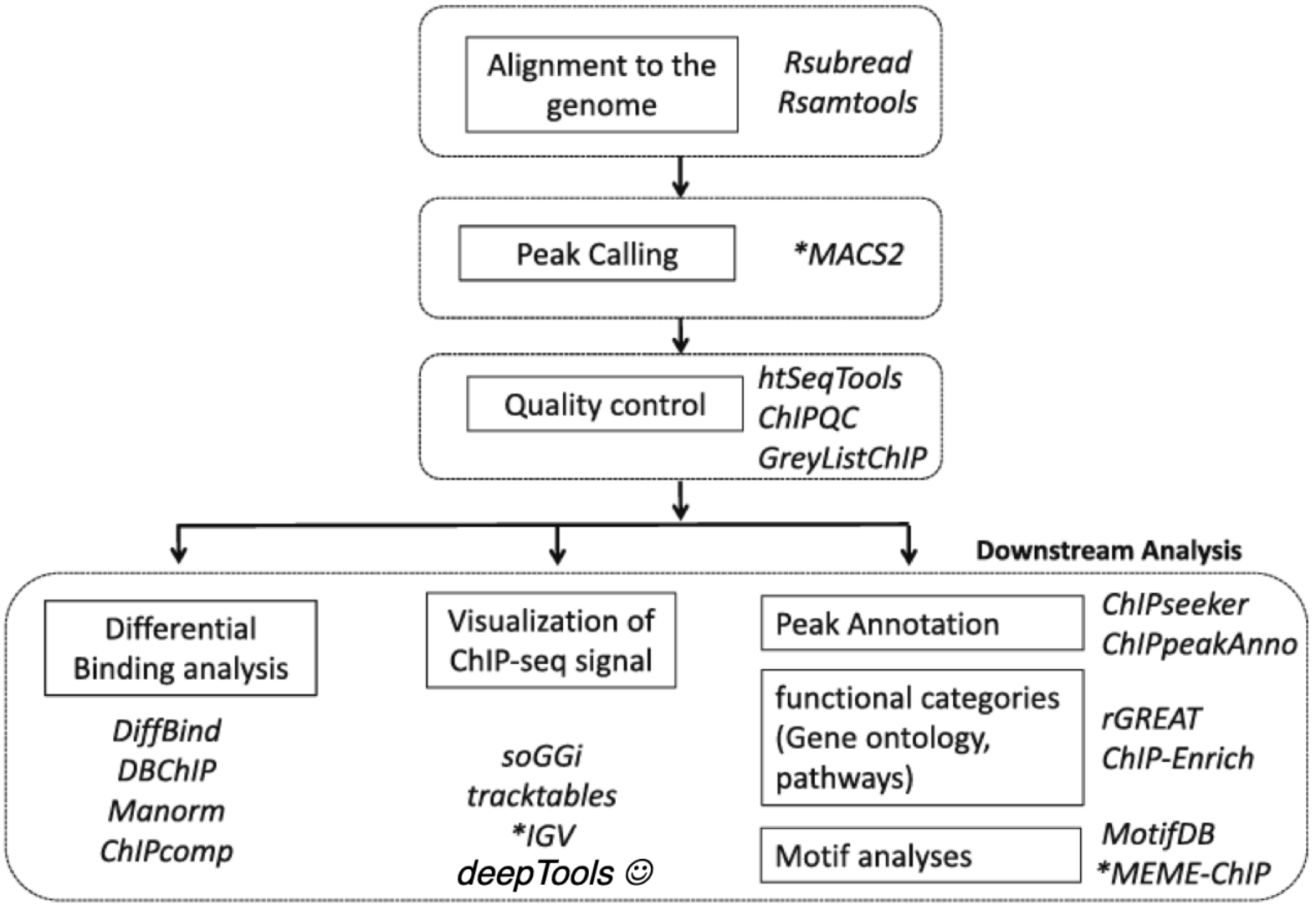
pay attention to where the curves start to rise – this already gives you an assessment of how much of the genome you have not sequenced at all (i.e. bins containing zero reads)



H3K27me3 is a mark that yields **broad** domains instead of narrow peaks

more difficult to distinguish input and ChIP, it does not mean, however, that this particular ChIP experiment failed

Overview of typical ChIP-seq-based analyses



de Santiago, I., & Carroll, T. (2018). *Analysis of ChIP-seq data in R/Bioconductor.* Methods in Molecular Biology

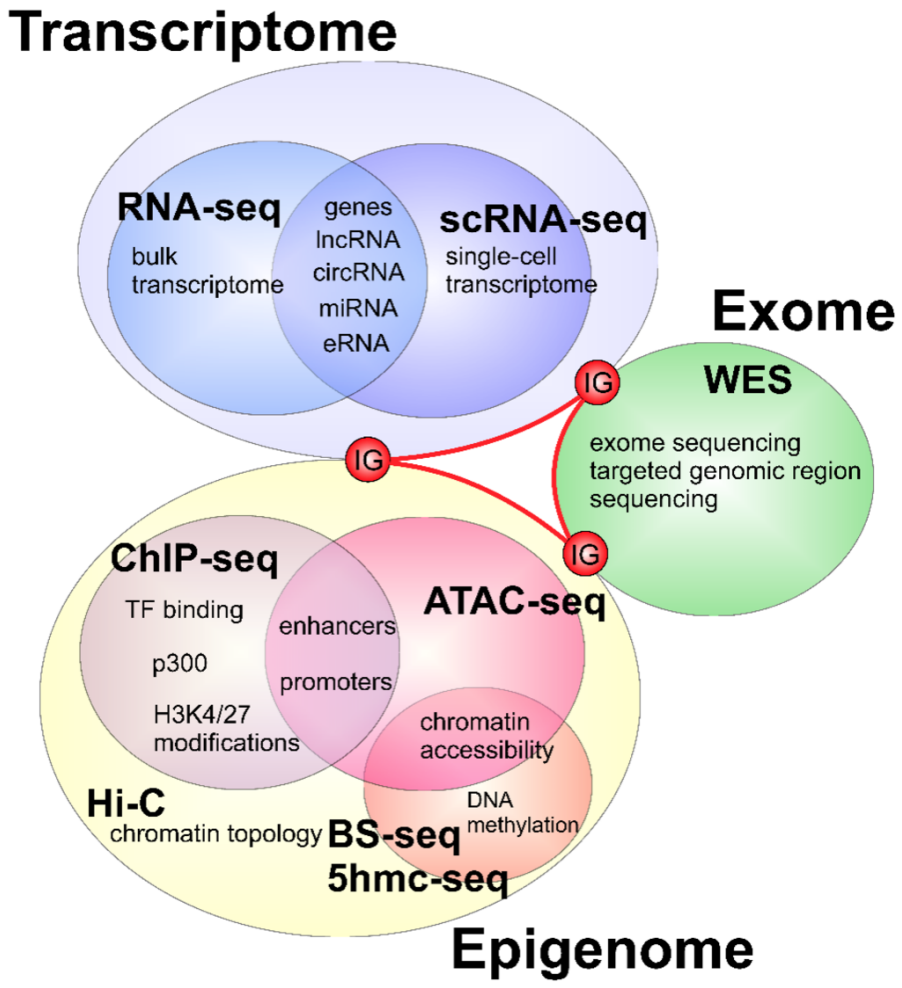
Comparing different ChIP-seq experiments

- comparing the levels of ChIP (and ATAC)-seq enrichments across different conditions is more difficult than one would have hoped for [Guertin et al., 2018]
  - Steinhauser et al. [2016] did a comparison of differential ChIP-seq tools
  - the winner tends to be the bioconductor package `DiffBind`, which is basically a sophisticated wrapper around `DESeq`
- relatively few efforts have been made towards understanding ChIP-seq/ATAC-seq-specific data properties, but the general consensus is that particularly ChIP-seq is awfully noisy and dependent on too many experimental parameters

"Although we would ideally want to study the absolute levels of binding, we have to accept the limitations of ChIP-seq [and ATAC-seq] and adapt by designing experiments in such a way that meaningful conclusions can be drawn from relative levels." [Meyer and Liu, 2014]

Summary

NGS approaches for epigenomics



- DNA = more or less immutable code
- RNA = the code's local read-out
- "epigenome" = additional molecules or chemical DNA modifications that govern the process of DNA-to-RNA transcription
- technically, epigenetics only refers to *heritable* marks that influence transcription [Ptashne, 2013]
- in practice, epigenomics is often used to describe all kinds of aspects of transcription regulation, including highly dynamic ones!

References

Figures taken from the following publications:

[Buenrostro et al., 2013, Chen et al., 2014, Qin et al., 2016, Corces et al., 2016, Cowie et al., 2013, Ramírez et al., 2016, Ernst and Kellis, 2017, Friedman and Rando, 2015, Gaffney et al., 2012, Klemm et al., 2019, Zhang et al., 2008, Meyer and Liu, 2014, Montefiori et al., 2017, Mundade et al., 2014, Nakato and Shirahige, 2017, Ou et al., 2018, Park, 2009, Ptashne, 2013, Splinter and De Laat, 2011, Stricker et al., 2016, Thomas et al., 2017, Waddington, 1942, Wei et al., 2018, Wilbanks and Facciotti, 2010, Winter et al., 2015]

Laura Baranello, Fedor Kouzine, Suzanne Sanford, and David Levens. ChIP bias as a function of cross-linking time. *Chromosome Research*, 2016. doi: 10.1007/s10577-015-9509-1.

Jason D. Buenrostro, Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 2013. doi: 10.1038/nmeth.2688.

R. Chen, R. Kang, X. G. Fan, and D. Tang. Release and activity of histone in diseases. *Cell Death and Disease*, 2014. doi: 10.1038/cddis.2014.337.

M. Ryan Corces, Jason D. Buenrostro, Beijing Wu, Peyton G. Greenside, Steven M. Chan, Julie L. Koenig, Michael P. Snyder, Jonathan K. Pritchard, Anshul Kundaje, William J. Greenleaf, Ravindra Majeti, and Howard Y. Chang. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature Genetics*, 2016. doi: 10.1038/ng.3646.

Philip Cowie, Ruth Ross, and Alasdair MacKenzie. Understanding the Dynamics of Gene Regulatory Systems; Characterisation and Clinical Relevance of cis-Regulatory Polymorphisms. *Biology*, 2013. doi: 10.3390/biology2010064.

Timothy Daley and Andrew D. Smith. Predicting the molecular complexity of sequencing libraries. *Nature Methods*, 2013. ISSN 15487091. doi: 10.1038/nmeth.2375.

Jason Ernst and Manolis Kellis. Chromatin-state discovery and genome annotation with ChromHMM. *Nature Protocols*, 2017. doi: 10.1038/nprot.2017.124.

Nir Friedman and Oliver J. Rando. Epigenomics and the structure of the living genome. *Genome Research*, 2015. doi: 10.1101/gr.190165.115.

Daniel J. Gaffney, Graham McVicker, Athma A. Pai, Yvonne N. Fondufe-Mittendorf, Noah Lewellen, Katelyn Michelini, Jonathan Widom, Yoav Gilad, and Jonathan K. Pritchard. Controls of Nucleosome Positioning in the Human Genome. *PLoS Genetics*, 2012. doi: 10.1371/journal.pgen.1003036.

Michael J. Guertin, Amy E. Cullen, Florian Markowetz, and Andrew N. Holding. Parallel factor ChIP provides essential internal control for quantitative differential ChIP-seq. *Nucleic Acids Research*, 2018. doi: 10.1093/nar/gky252.

Antonio Jordán-Pla and Neus Visa. Considerations on experimental design and data analysis of chromatin immunoprecipitation experiments. In *Methods in Molecular Biology*. 2018. doi: 10.1007/978-1-4939-7380-4_2.

Sandy L. Klemm, Zohar Shipony, and William J. Greenleaf. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, 2019. doi: 10.1038/s41576-018-0089-8.

Cory Y. McLean, Dave Bristor, Michael Hiller, Shoa L. Clarke, Bruce T. Schaar, Craig B. Lowe, Aaron M. Wenger, and Gill Bejerano. GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology*, 2010. doi: 10.1038/nbt.1630.

Clifford A Meyer and X Shirley Liu. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nature Reviews Genetics*, 2014. doi: 10.1038/nrg3788.

Lindsey Montefiori, Liana Hernandez, Zijie Zhang, Yoav Gilad, Carole Ober, Gregory Crawford, Marcelo Nobrega, and Noboru Jo Sakabe. Reducing mitochondrial reads in ATAC-seq using CRISPR/Cas9. *Scientific Reports*, 2017. doi: 10.1038/s41598-017-02547-w.

Rasika Mundade, Hatice Gulcin Ozer, Han Wei, Lakshmi Prabhu, and Tao Lu. Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond. *Cell Cycle*, 2014. doi: 10.4161/15384101.2014.949201.

Ryuichiro Nakato and Katsuhiko Shirahige. Recent advances in ChIP-seq analysis: From quality management to whole-genome annotation. *Briefings in Bioinformatics*, 2017. doi: 10.1093/bib/bbw023.

Jianhong Ou, Haibo Liu, Jun Yu, Michelle A. Kelliher, Lucio H. Castilla, Nathan D. Lawson, and Lihua Julie Zhu. ATACseqQC: A Bioconductor package for post-alignment quality assessment of ATAC-seq data. *BMC Genomics*, 2018. doi: 10.1186/s12864-018-4559-3.

Peter J Park. ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–80, Oct 2009. doi: 10.1038/nrg2641.

M. Ptashne. Epigenetics: Core misconcept. *Proceedings of the National Academy of Sciences*, 2013. doi: 10.1073/pnas.1305399110.

Qian Qin, Shenglin Mei, Qiu Wu, Hanfei Sun, Lewyn Li, Len Taing, Sujun Chen, Fugen Li, Tao Liu, Chongzhi Zang, Han Xu, Yiwen Chen, Clifford A. Meyer, Yong Zhang, Myles Brown, Henry W. Long, and X. Shirley Liu. ChiLin: A comprehensive ChIP-seq and DNase-seq quality control and analysis pipeline. *BMC Bioinformatics*, 2016. doi: 10.1186/s12859-016-1274-4.

Aaron R. Quinlan. BEDTools: The Swiss-Army tool for genome feature analysis. *Current Protocols in Bioinformatics*, 2014. doi: 10.1002/0471250953.bi1112s47.

Fidel Ramírez, Devon P. Ryan, Björn Grüning, Vivek Bhardwaj, Fabian Kilpert, Andreas S. Richter, Steffen Heyne, Friederike Dündar, and Thomas Manke. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic acids research*, 2016. ISSN 13624962. doi: 10.1093/nar/gkw257.

Erik Splinter and Wouter De Laat. The complex transcription regulatory landscape of our genome: Control in three dimensions. *EMBO Journal*, 2011. doi: 10.1038/emboj.2011.344.

Sebastian Steinhauser, Nils Kurzawa, Roland Eils, and Carl Herrmann. A comprehensive comparison of tools for differential ChIP-seq analysis. *Briefings in Bioinformatics*, 2016. doi: 10.1093/bib/bbv110.

Stefan H. Stricker, Anna Köferle, and Stephan Beck. From profiles to function in epigenomics. *Nature Reviews Genetics*, 2016. doi: 10.1038/nrg.2016.138.

Reuben Thomas, Sean Thomas, Alisha K. Holloway, and Katherine S. Pollard. Features that define the best ChIP-seq peak calling algorithms. *Briefings in Bioinformatics*, 2017. doi: 10.1093/bib/bbw035.

C. H. Waddington. The epigenotype. . *International Journal of Epidemiology (2012)*, 1942. doi: 10.1093/ije/dyr184.

Zheng Wei, Wei Zhang, Huan Fang, Yanda Li, and Xiaowo Wang. esATAC: an easy-to-use systematic pipeline for ATAC-seq data analysis. *Bioinformatics*, 2018. doi: 10.1093/bioinformatics/bty141.

Elizabeth G. Wilbanks and Marc T. Facciotti. Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS ONE*, 2010. doi: 10.1371/journal.pone.0011471.

Deborah R. Winter, Steffen Jung, and Ido Amit. Making the case for chromatin profiling: A new tool to investigate the immune-regulatory landscape. *Nature Reviews Immunology*, 2015. doi: 10.1038/nri3884.

Feng Yan, David R. Powell, David J. Curtis, and Nicholas C. Wong. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biology*, 2020. doi: 10.1186/s13059-020-1929-3.

Guangchuang Yu, Li Gen Wang, and Qing Yu He. ChIP seeker: An R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*, 2015. doi: 10.1093/bioinformatics/btv145.

Yong Zhang, Tao Liu, Clifford A. Meyer, Jérôme Eeckhoute, David S. Johnson, Bradley E. Bernstein, Chad Nussbaum, Richard M. Myers, Myles Brown, Wei Li, and X. Shirley Shirley. Model-based analysis of ChIP-Seq (MACS). *Genome Biology*, 2008. doi: 10.1186/gb-2008-9-9-r137.

Lihua J Zhu, Claude Gazin, Nathan D Lawson, Hervé Pagès, Simon M Lin, David S Lapointe, and Michael R Green. ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics*, 11:237, 2010. doi: 10.1186/1471-2105-11-237. URL http://dx.doi.org/10.1186/1471-2105-11-237{%}5Cnhttp://www.ncbi.nlm.nih.gov/pmc/articles/PMC3098059/pdf/1471-2105-11-237.pdf.