Introduction to ChIP-seq and ATAC-seq

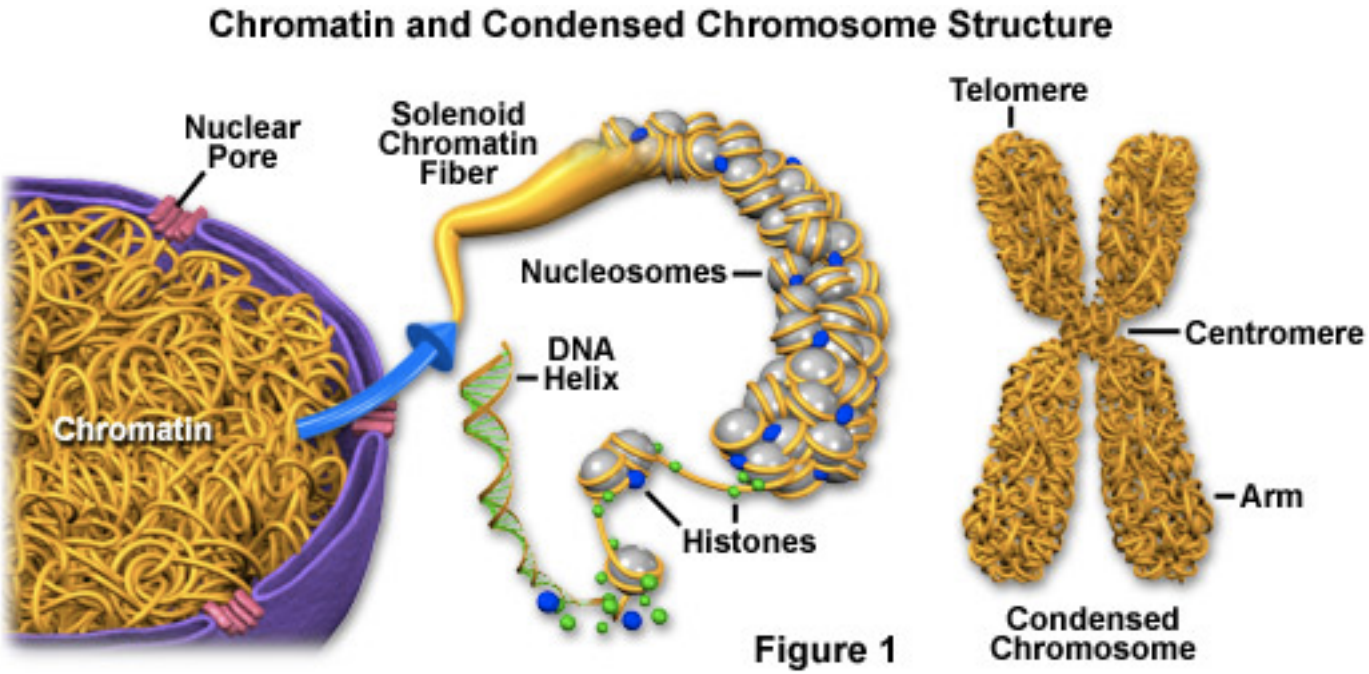
Emily Brown

May 18, 2020

Introduction to ChIP-seq and ATAC-seq

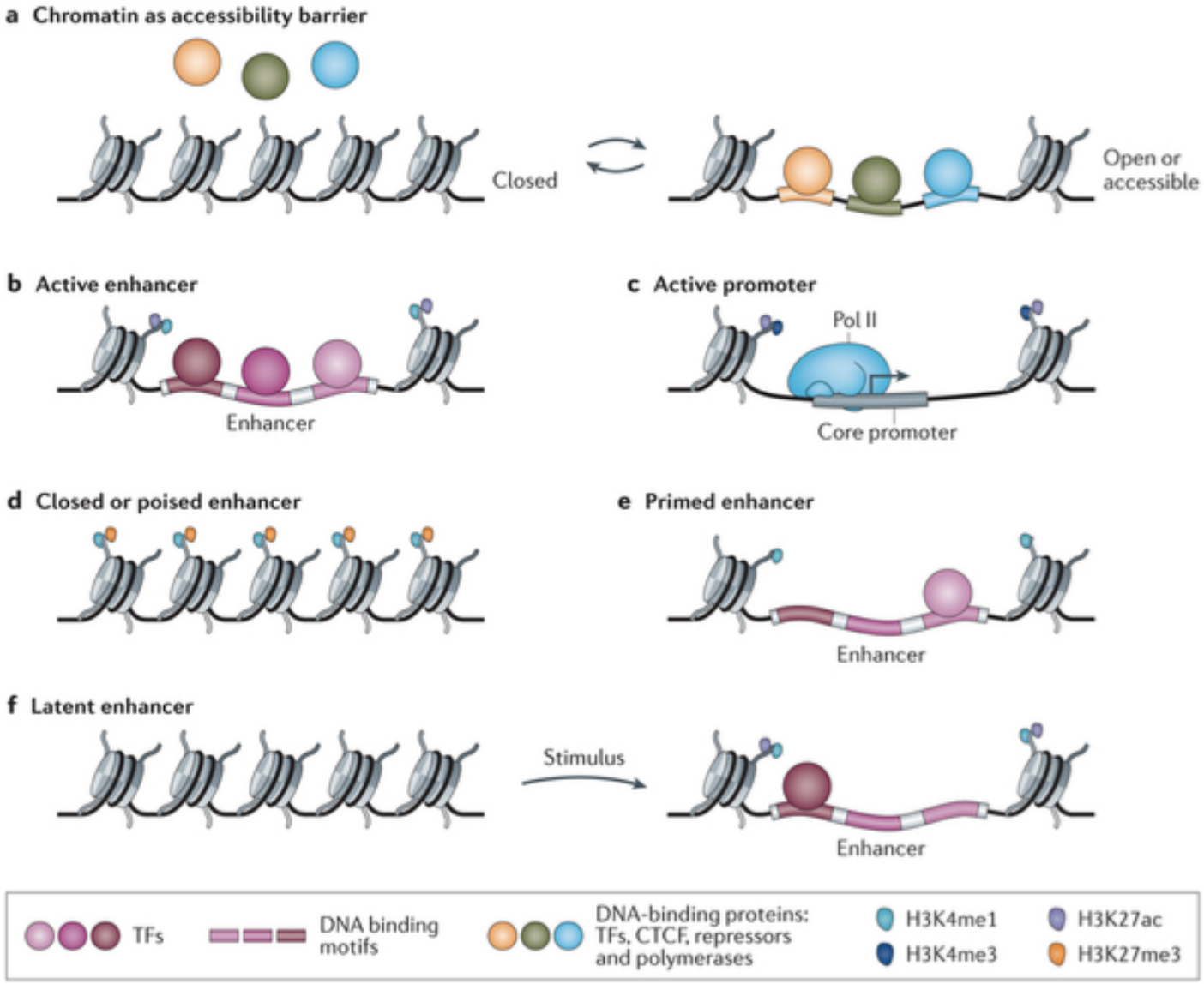
- Why look at chromatin structure on a genome-wide basis?
- How does ChIP-seq work?
  - How is ATAC-seq different?
- When would you choose to do ATAC-seq vs. ChIP-seq?
- How does the analysis work?
  - How does it differ for ChIP-seq and ATAC-seq?
  - Downstream analyses and follow-ups

What can you gain by looking at chromatin structure?



https://hbctraining.github.io/Intro-to-ChIPseq/

Chromatin structure determines if a gene is expressed



Shlyueva, et al (2014)

Nature Reviews | Genetics

What can you gain by looking at chromatin structure?

Characterize a cell type

- what is the chromatin landscape of the cell type?
- what transcription factor (TF) binding sites are open or closed?

What can you gain by looking at chromatin structure?

Characterize a cell type

Characterize a DNA-binding protein

- where does protein bind in given cell type(s)?
- what is the protein's binding motif?

What can you gain by looking at chromatin structure?

Characterize a cell type

Characterize a DNA-binding protein

Follow up differential expression analysis

- identify candidate TFs responsible for differential expression

What can you gain by looking at chromatin structure?

Characterize a cell type

Characterize a DNA-binding protein

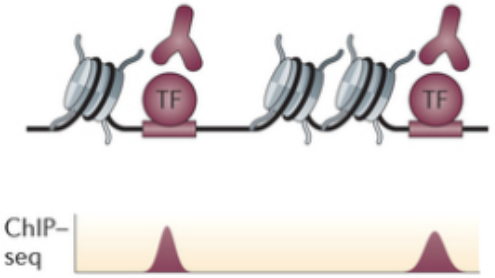
Follow up differential expression analysis

Characterize developmental trajectories

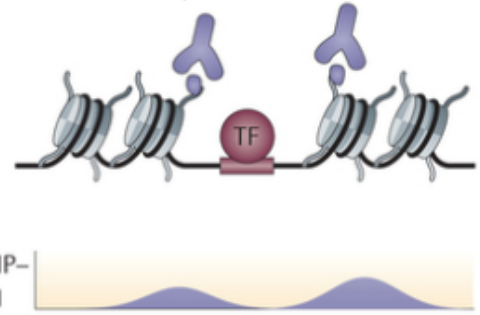
- follow pioneering TFs during development
- track histone modification changes

How to detect chromatin structure genome-wide?

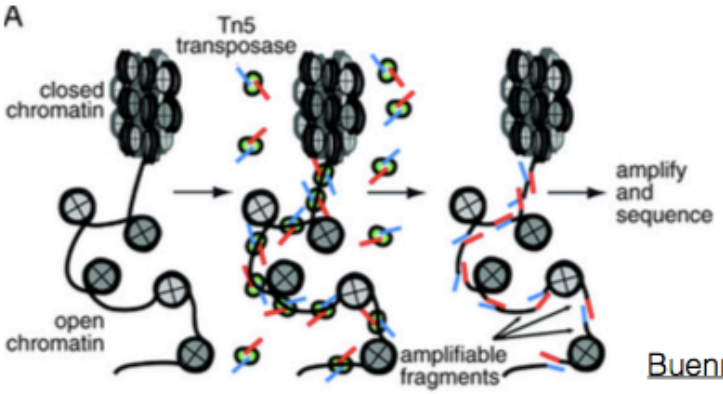
ChIP-seq for a TF



ChIP-seq for chromatin marks

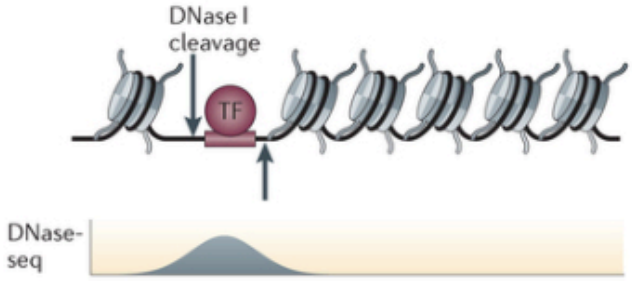


ATAC-seq



Buenrostro et al., 2015

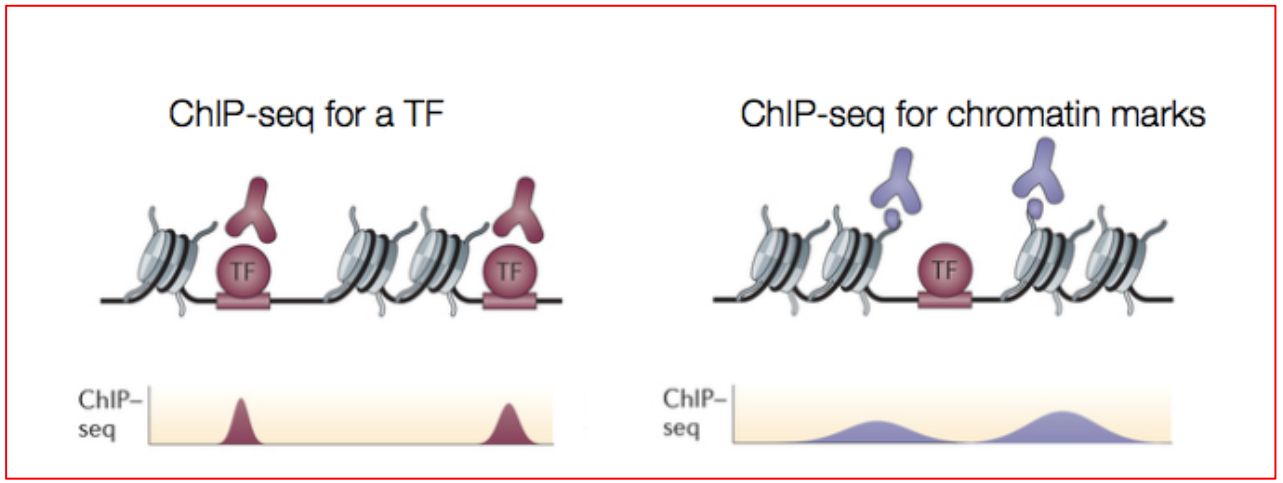
DNase-seq

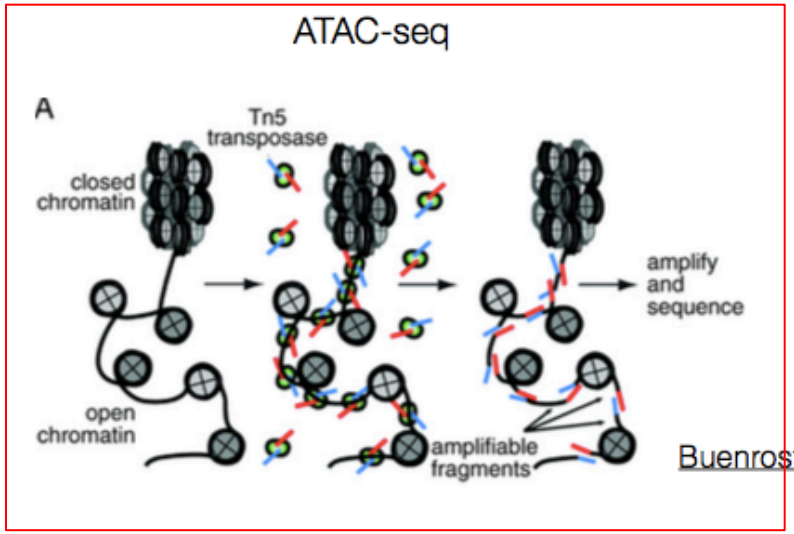


Shlyueva, et al (2014)

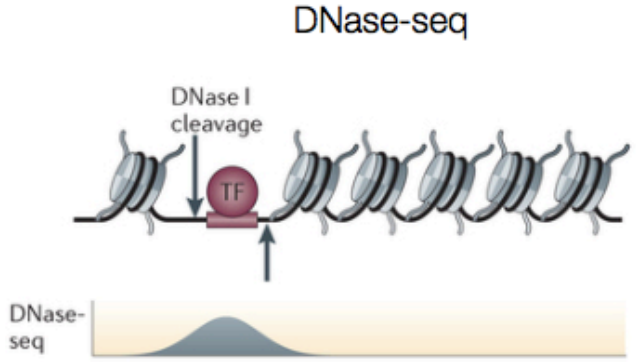
Also ChIA-PET and chromosome-conformation capture (3C+) based methods to interrogate the 3D structure of the genome

How to detect chromatin structure genome-wide?





Buenrostro et al., 2015

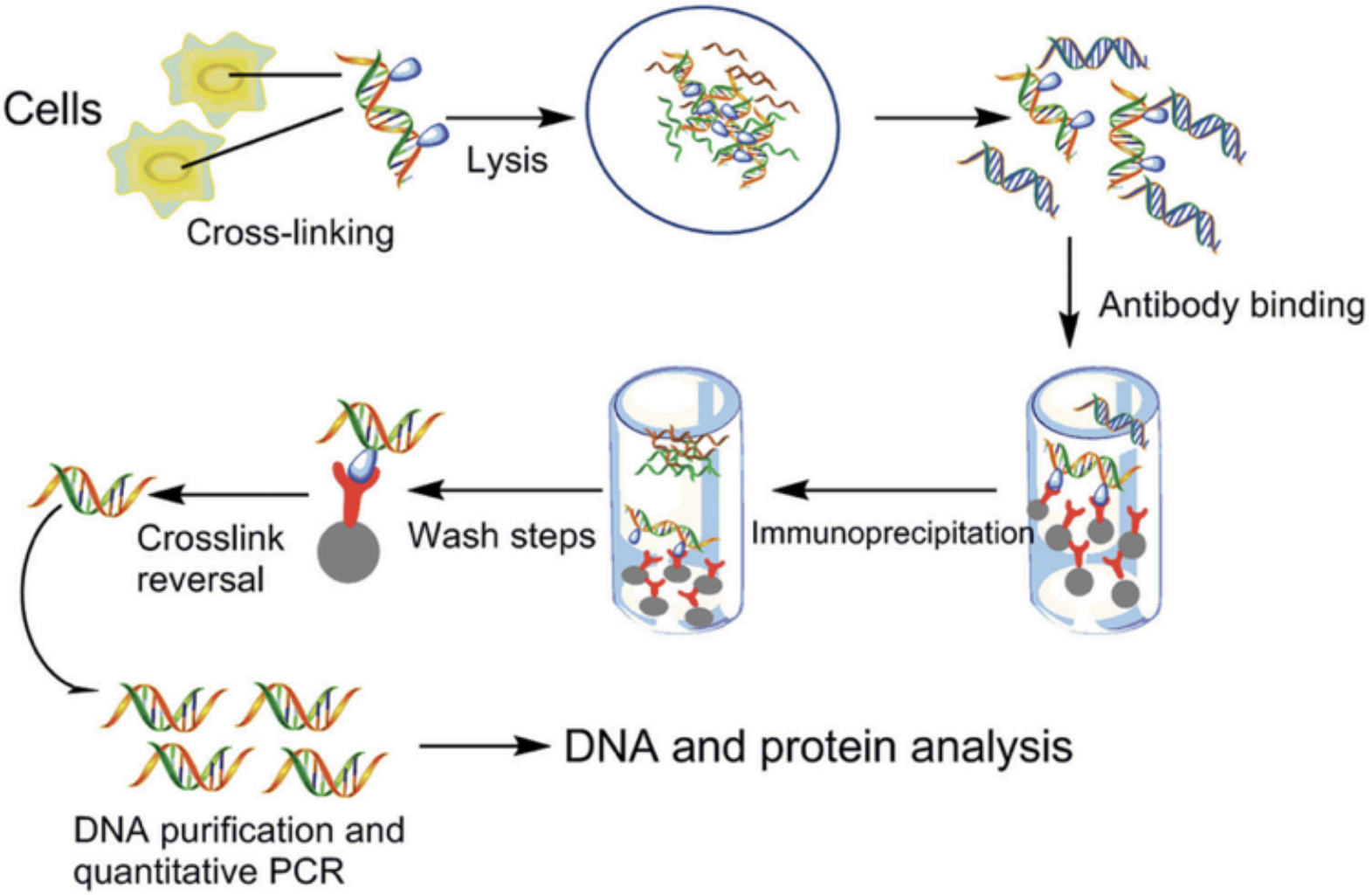


Shlyueva, et al (2014)

Also ChIA-PET and chromosome-conformation capture (3C+) based methods to interrogate the 3D structure of the genome

How does ChIP-seq work?

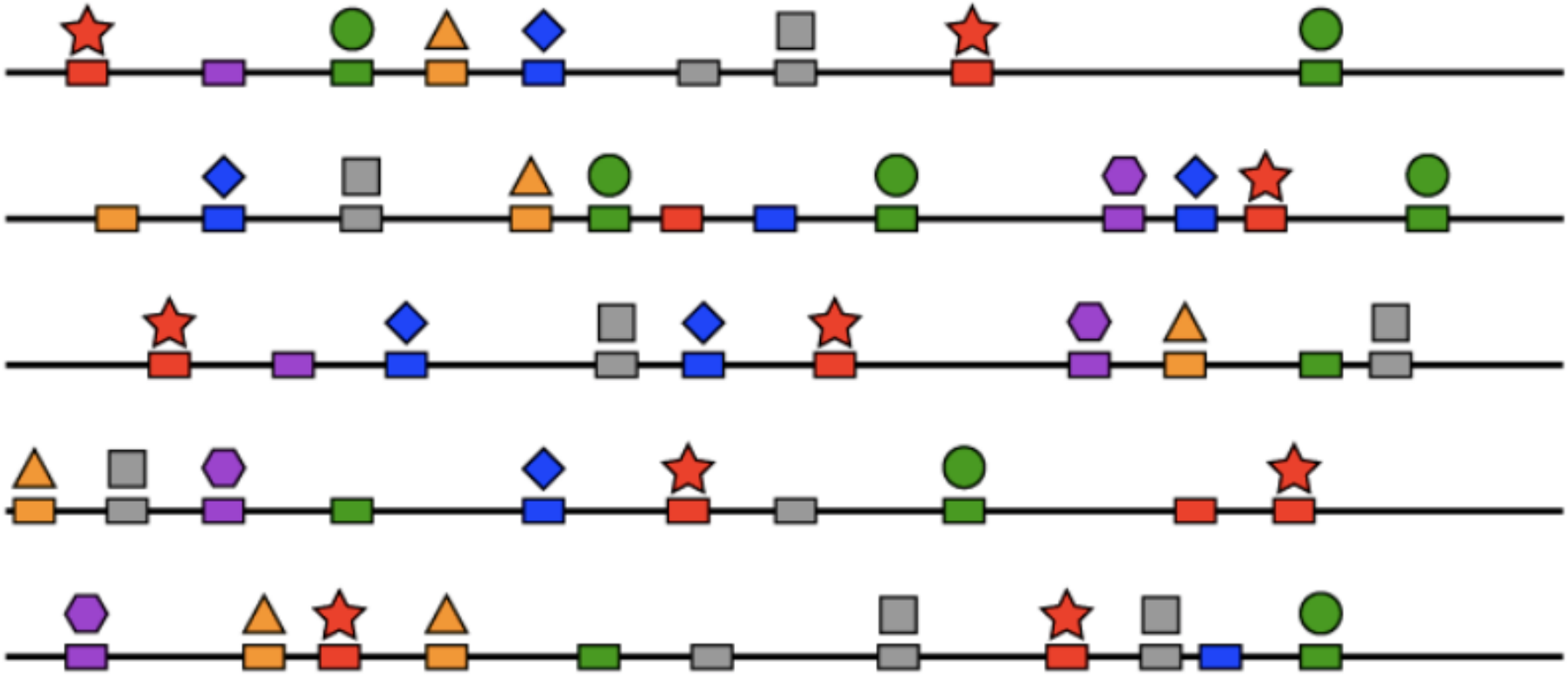
Chromatin **I**mmuno-**P**recipitation



How does ChIP-seq work?

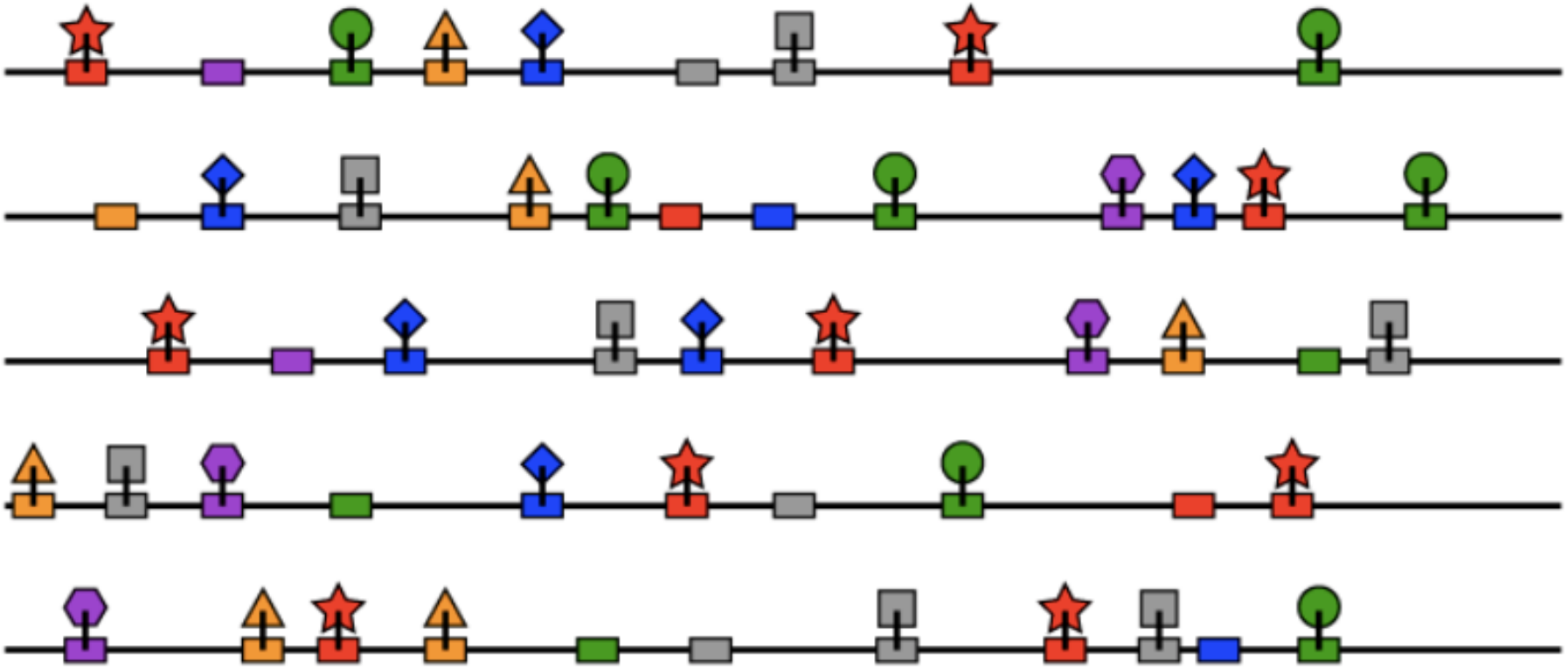
Starting point: isolated chromatin

Note: need a fair amount of starting material because ChIP will enrich for small portion
(some suggest starting point of 10^7 cells)



How does ChIP-seq work?

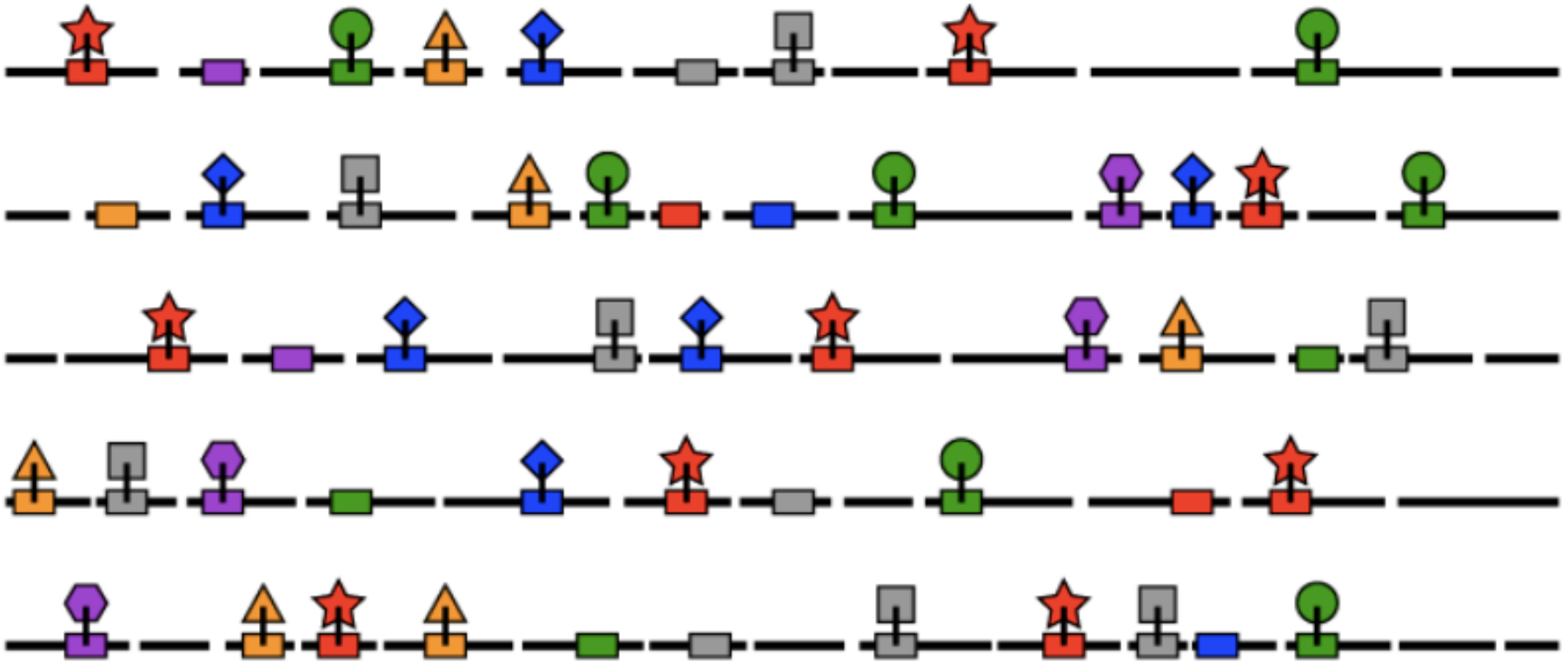
Step 1: cross-link proteins to DNA



How does ChIP-seq work?

Step 2: fragment cross-linked chromatin

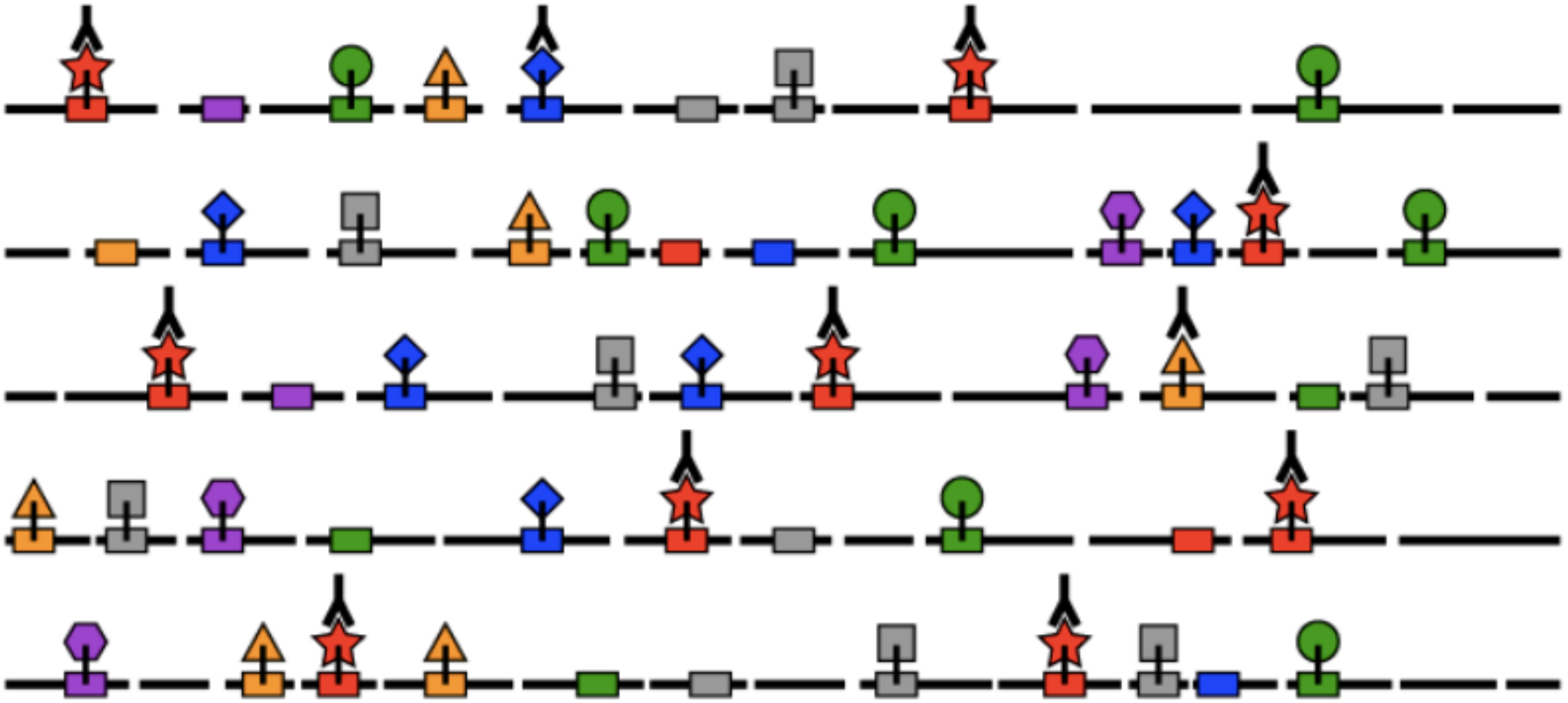
- fragment DNA into 200-500 bp
- need to optimize fixation time and sonication time to get fragments within this range



How does ChIP-seq work?

Step 3: introduce protein-specific antibody

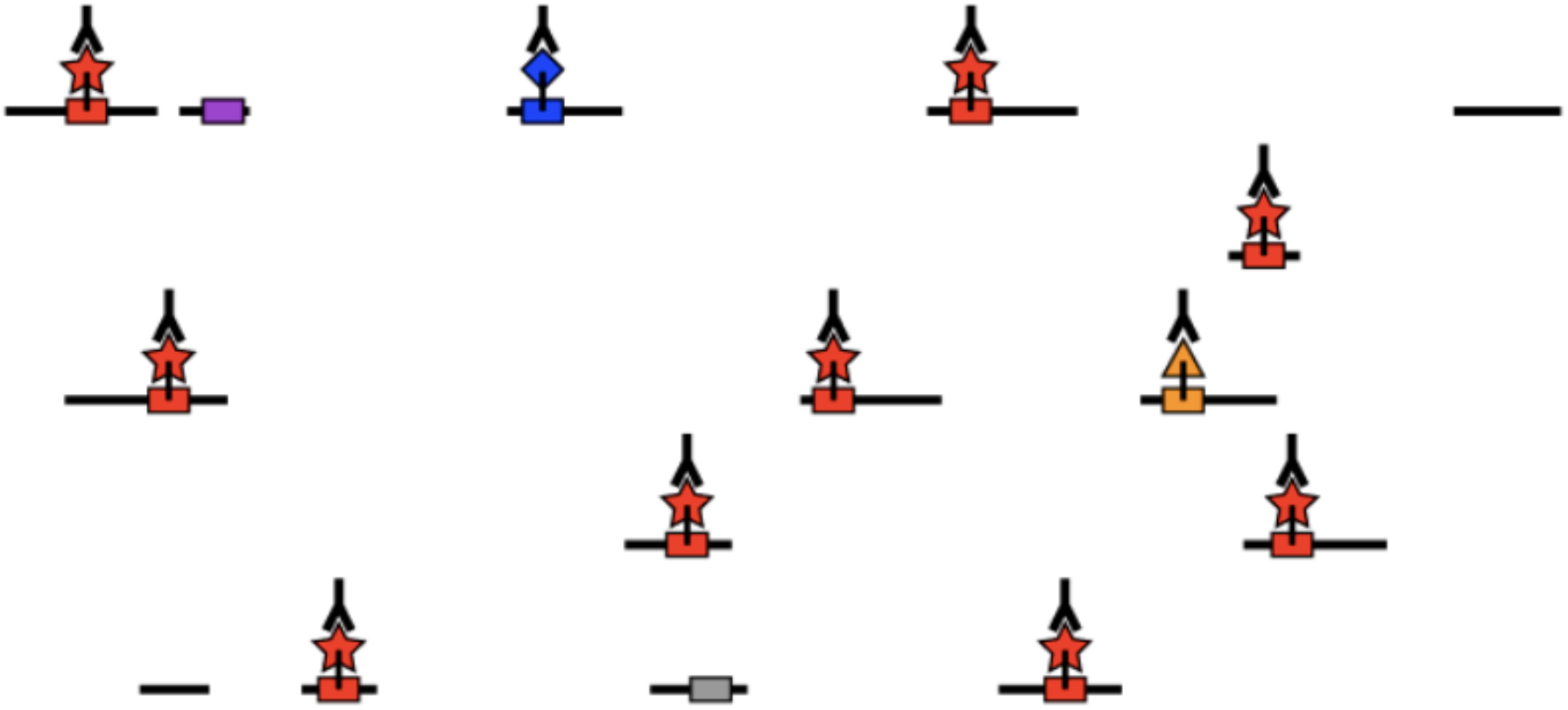
- amount of antibody per IP must be optimized



How does ChIP-seq work?

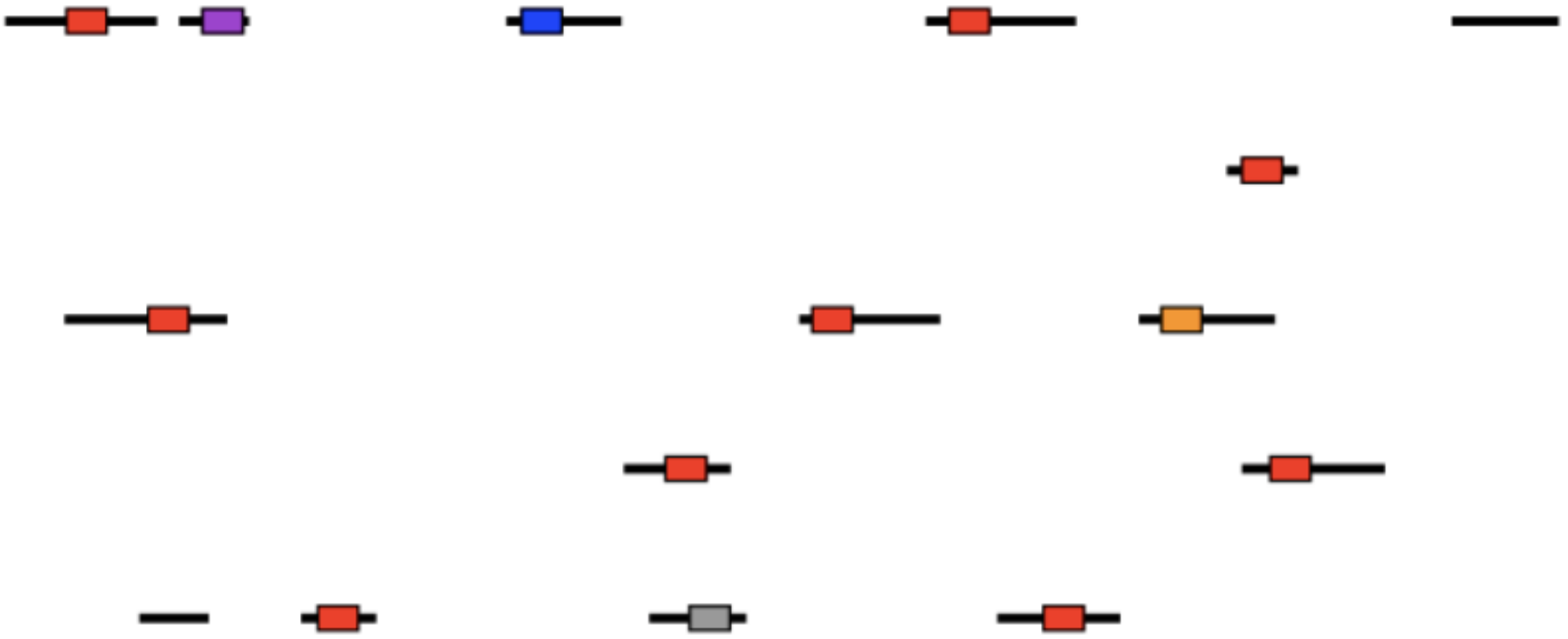
Step 4: Immunoprecipitate

- certain amount of non-specific binding
- some naked DNA will make it through as well

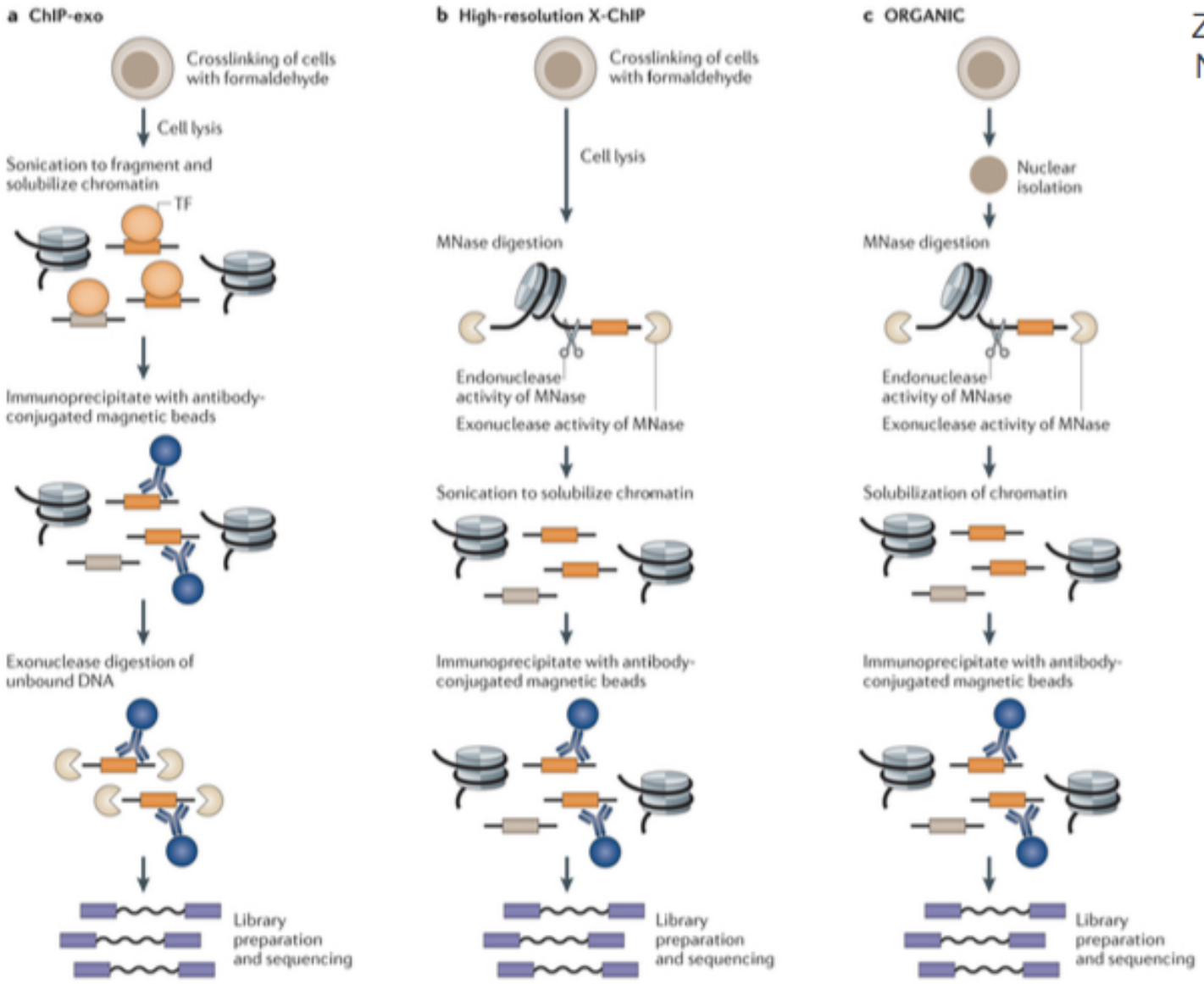


How does ChIP-seq work?

Step 5: reverse cross links and go into library prep



Quick note on high-resolution ChIP methods

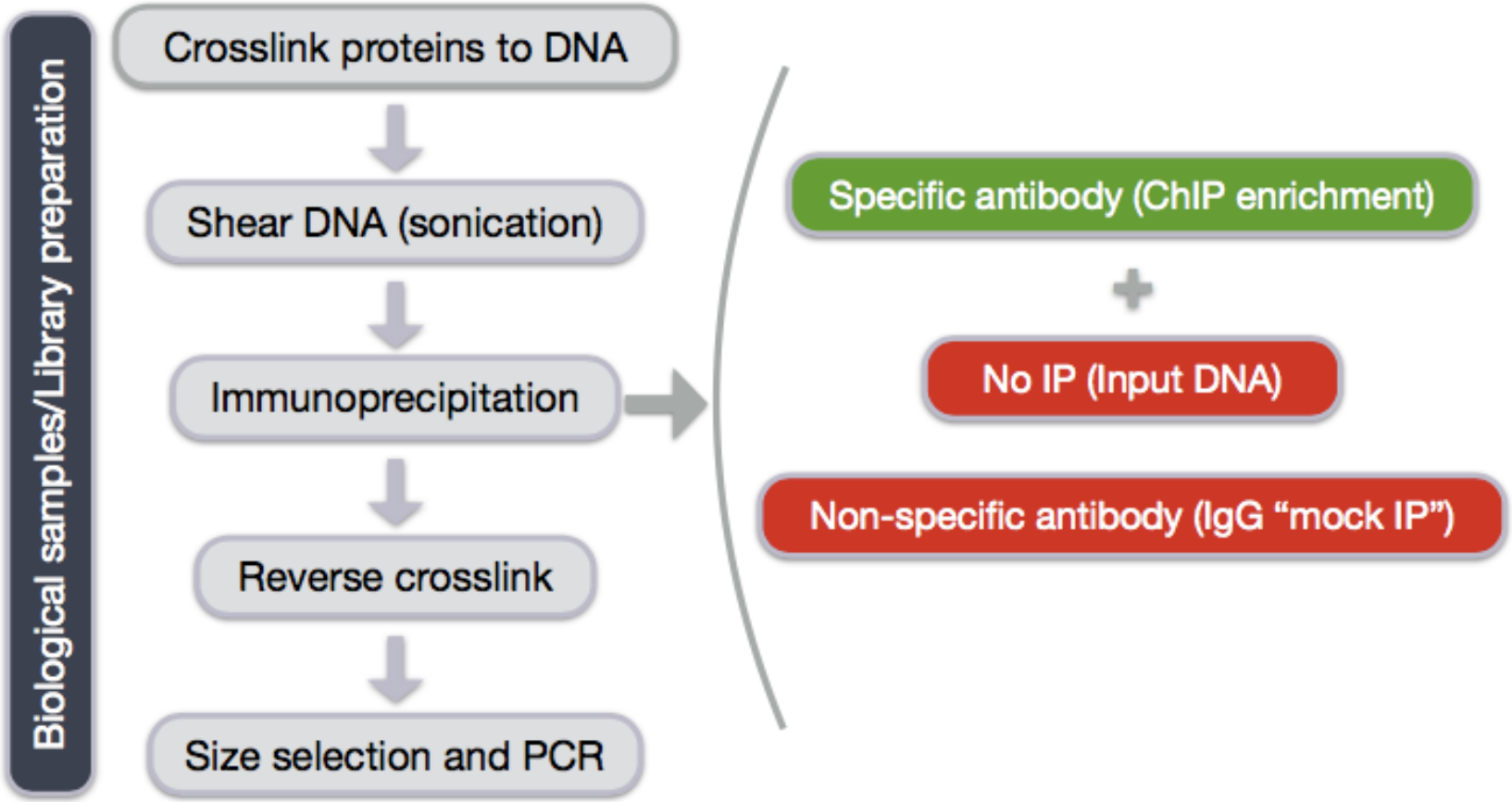


Zentner & Henikoff (2014).
Nature Reviews Genetics.

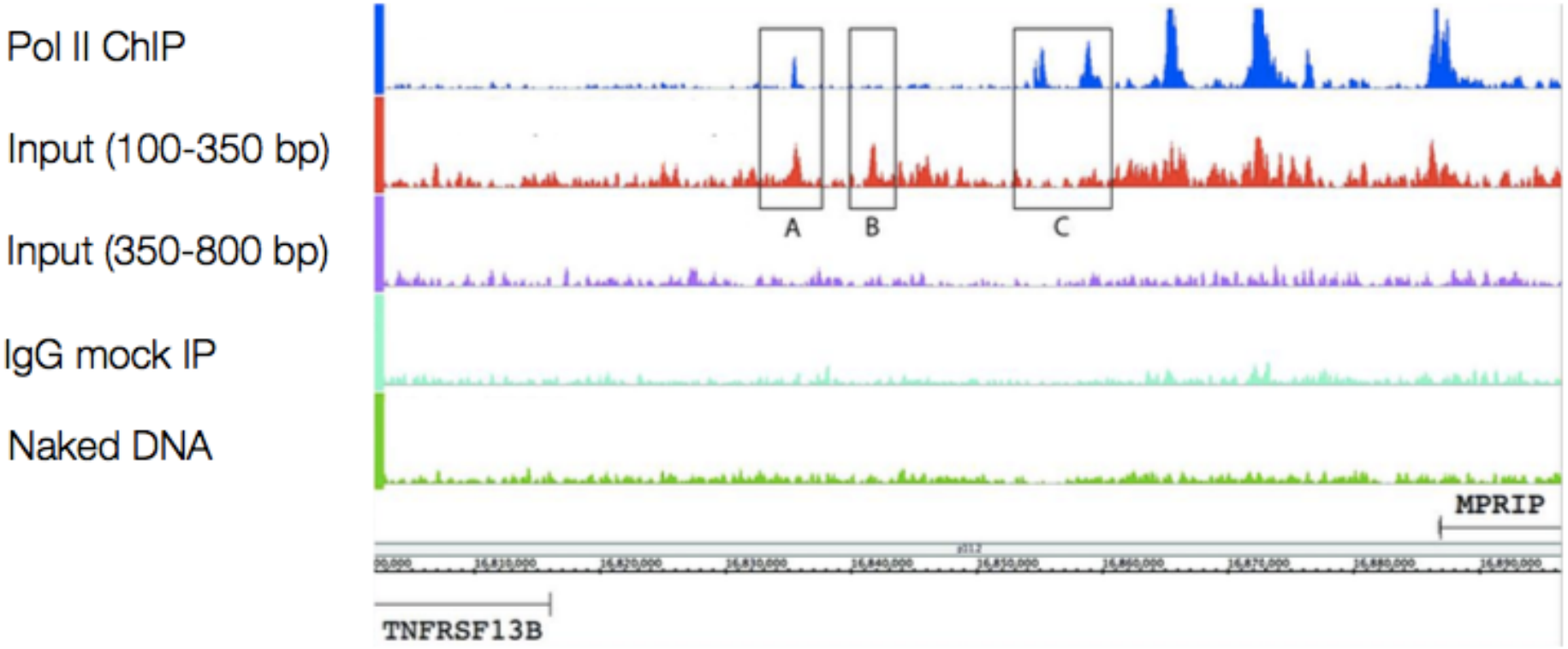
Why are controls necessary for ChIP-seq?

- Open chromatin regions fragment more easily than closed regions
- Repetitive sequences may seem to be enriched
- Uneven distribution of sequence tags across the genome
- Hyper-ChIPable regions
- ENCODE provides a “Black List” of troublesome genomic regions

Why are controls necessary for ChIP-seq?



Why are controls necessary for ChIP-seq?

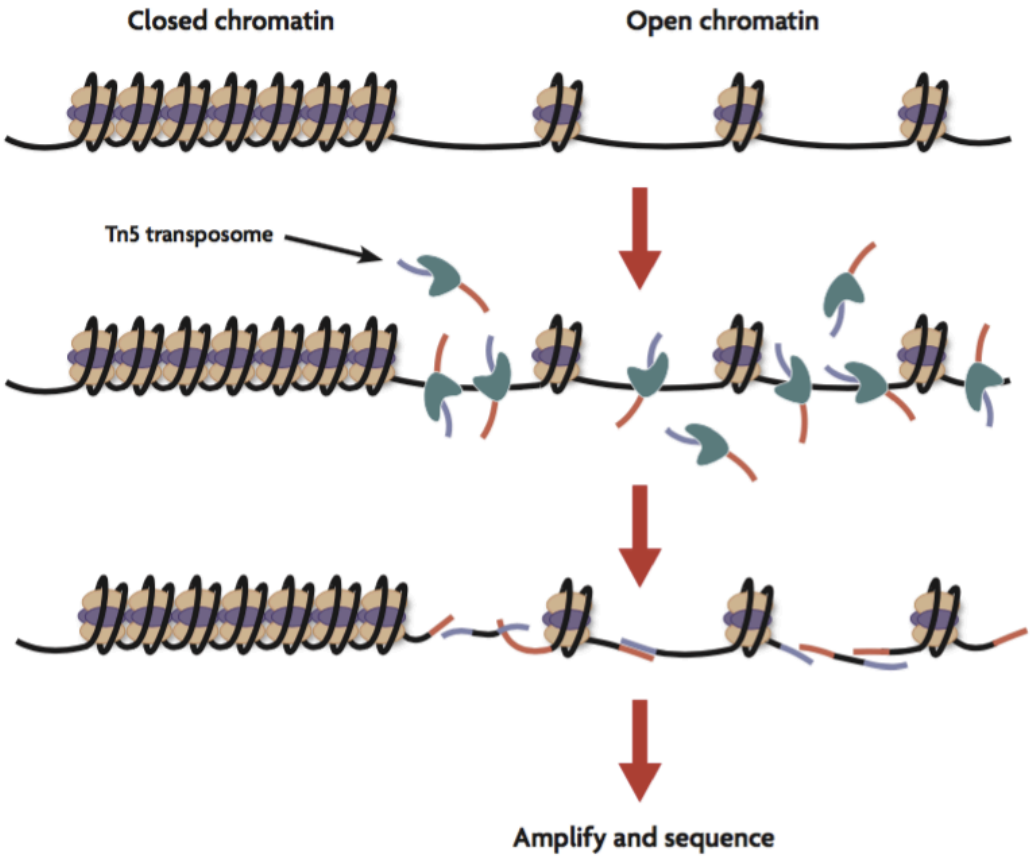


Parameters for a successful ChIP

- Efficient and specific antibody
- Sufficient amount of starting material
- ChIP DNA yield depends on multiple factors
  - cell type in question
  - abundance of the mark or protein (histones have higher binding "area" than TFs, so will pull down more DNA)
  - Antibody quality
- Chromatin fragmentation
  - Size matters!
    - Too big → reduced signal to noise ratio
    - Too small → strong bias to open/ promoter regions in both IP AND control
  - Fragmentation parameters can vary between cell types
- Stringency of washes

How does ATAC-seq work?

Assay of **T**ransposase-**A**ccessible **C**hromatin



How does ATAC-seq work?

Basic steps:

- harvest cells
  - must be intact, homogenous, and counted
- transposition reaction
  - simultaneously fragment and add barcodes (“tagmentation”)
- amplify and construct sequencing library

How does ATAC-seq work?

Basic steps:

- harvest cells
- transposition reaction
- amplify and construct sequencing library

Notes:

- Key factor for quality is ratio of healthy cells to transposase
  - over- or under-digestion lowers quality of experiment
  - fixed or mechanically sheared cells also tend to reduce data quality

How does ATAC-seq work?

Basic steps:

- harvest cells
- transposition reaction
- amplify and construct sequencing library

Notes:

- Key factor for quality is ratio of healthy cells to transposase
- No antibodies required! No fixation! No sonication!

How does ATAC-seq work?

Basic steps:

- harvest cells
- transposition reaction
- amplify and construct sequencing library

Notes:

- Key factor for quality is ratio of healthy cells to transposase
- No antibodies required! No fixation! No sonication!
- Note that there is not an input control here
  - If goal is characterizing the chromatin landscape of a given cell type, etc, naked DNA (as opposed to chromatin) as control for tagmentation

When does it make more sense to do ATAC-seq vs. ChIP-seq?

When to do ChIP-seq:

- if you are interested in a specific DNA-binding protein
- if you are interested in repressed (closed) domains (heterochromatin or polycomb)
- if you are looking at longer-term changes – ATAC tends to pick up more dynamic changes

When to do ATAC-seq:

- if you don't have an expectation of what TFs/ DNA-binding proteins/ histone modifications are changing
- if you're more interested in a signal of changes in chromatin instead of the specific factor responsible

How does the data analysis work?

Broadly, it's similar to RNA-seq...

... because it's count data!

The main difference is we don't know ***where*** we're going to count reads...

... so the first step is defining the regions of interest.

How does the data analysis work?

Peak calling

Starting point: fastq file

unique read ID,
always starts with “@”

sequence

dummy line,
always starts with “+”

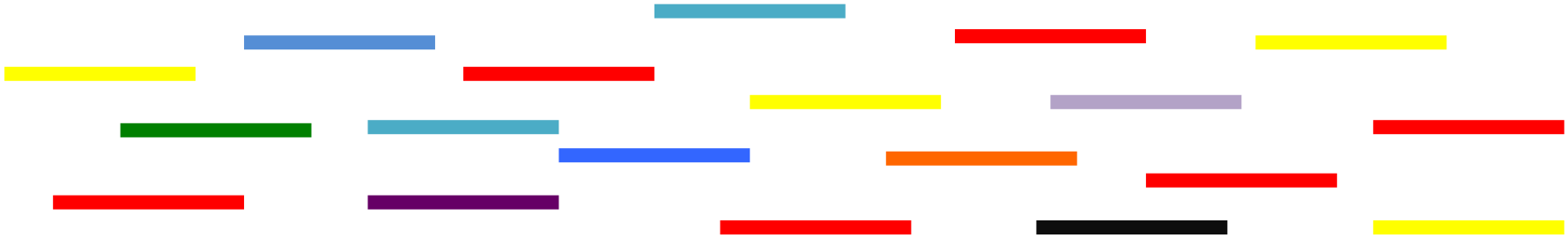
quality score for each base

Very large files! For each of 10s of millions of reads, there are 4 lines

How does the data analysis work?

Peak calling

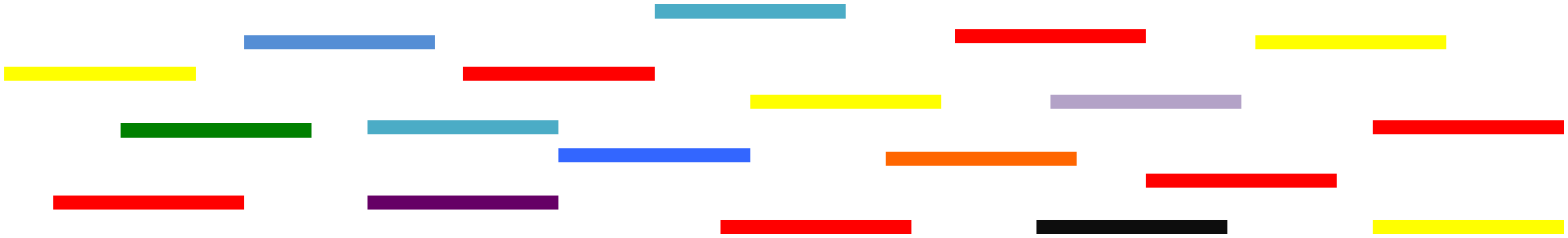
Starting point: fastq file
Just a bunch of bits of DNA sequence



How does the data analysis work?

Peak calling

Starting point: fastq file
Just a bunch of bits of DNA sequence



Step 1: align reads

Where in the genome do these short reads come from?

[BWA, Bowtie2]

How does the data analysis work?

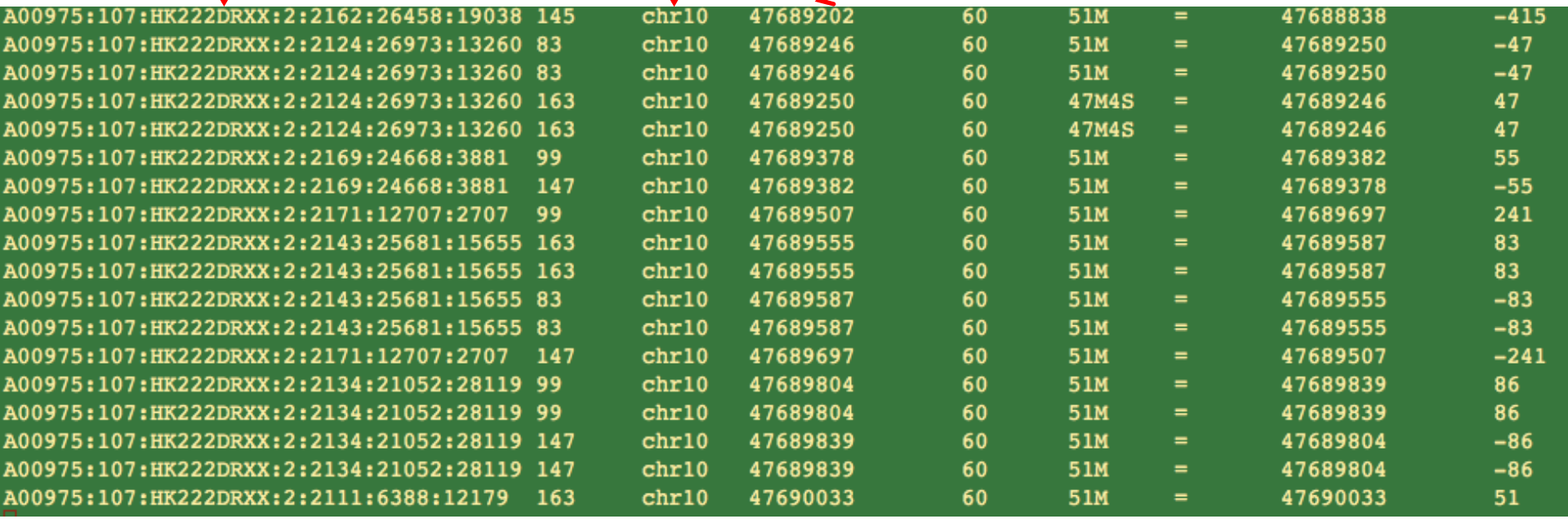
Peak calling

Intermediate file: sam/ bam file
(they're the same thing!)

unique read ID, now without the "@"

chromosome where read aligned

position where read aligned

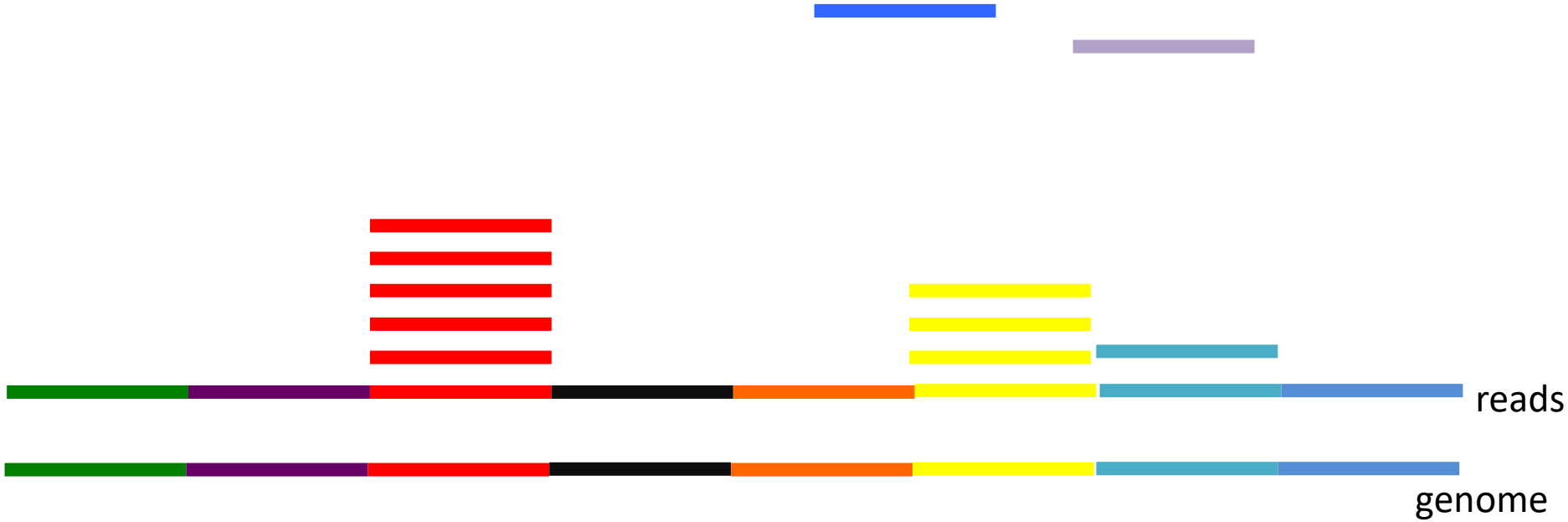


```
A00975:107:HK222DRXX:2:2162:26458:19038 145   chr10  47689202  60  51M  =  47688838  -415
A00975:107:HK222DRXX:2:2124:26973:13260 83    chr10  47689246  60  51M  =  47689250  -47
A00975:107:HK222DRXX:2:2124:26973:13260 83    chr10  47689246  60  51M  =  47689250  -47
A00975:107:HK222DRXX:2:2124:26973:13260 163   chr10  47689250  60  47M4S  =  47689246  47
A00975:107:HK222DRXX:2:2124:26973:13260 163   chr10  47689250  60  47M4S  =  47689246  47
A00975:107:HK222DRXX:2:2169:24668:3881  99    chr10  47689378  60  51M  =  47689382  55
A00975:107:HK222DRXX:2:2169:24668:3881  147   chr10  47689382  60  51M  =  47689378  -55
A00975:107:HK222DRXX:2:2171:12707:2707  99    chr10  47689507  60  51M  =  47689697  241
A00975:107:HK222DRXX:2:2143:25681:15655 163   chr10  47689555  60  51M  =  47689587  83
A00975:107:HK222DRXX:2:2143:25681:15655 163   chr10  47689555  60  51M  =  47689587  83
A00975:107:HK222DRXX:2:2143:25681:15655 83    chr10  47689587  60  51M  =  47689555  -83
A00975:107:HK222DRXX:2:2143:25681:15655 83    chr10  47689587  60  51M  =  47689555  -83
A00975:107:HK222DRXX:2:2171:12707:2707  147   chr10  47689697  60  51M  =  47689507  -241
A00975:107:HK222DRXX:2:2134:21052:28119 99    chr10  47689804  60  51M  =  47689839  86
A00975:107:HK222DRXX:2:2134:21052:28119 99    chr10  47689804  60  51M  =  47689839  86
A00975:107:HK222DRXX:2:2134:21052:28119 147   chr10  47689839  60  51M  =  47689804  -86
A00975:107:HK222DRXX:2:2134:21052:28119 147   chr10  47689839  60  51M  =  47689804  -86
A00975:107:HK222DRXX:2:2111:6388:12179  163   chr10  47690033  60  51M  =  47690033  51
```

Very large files! For each of 10s of millions of reads, there is lots of info

How does the data analysis work?

Peak calling

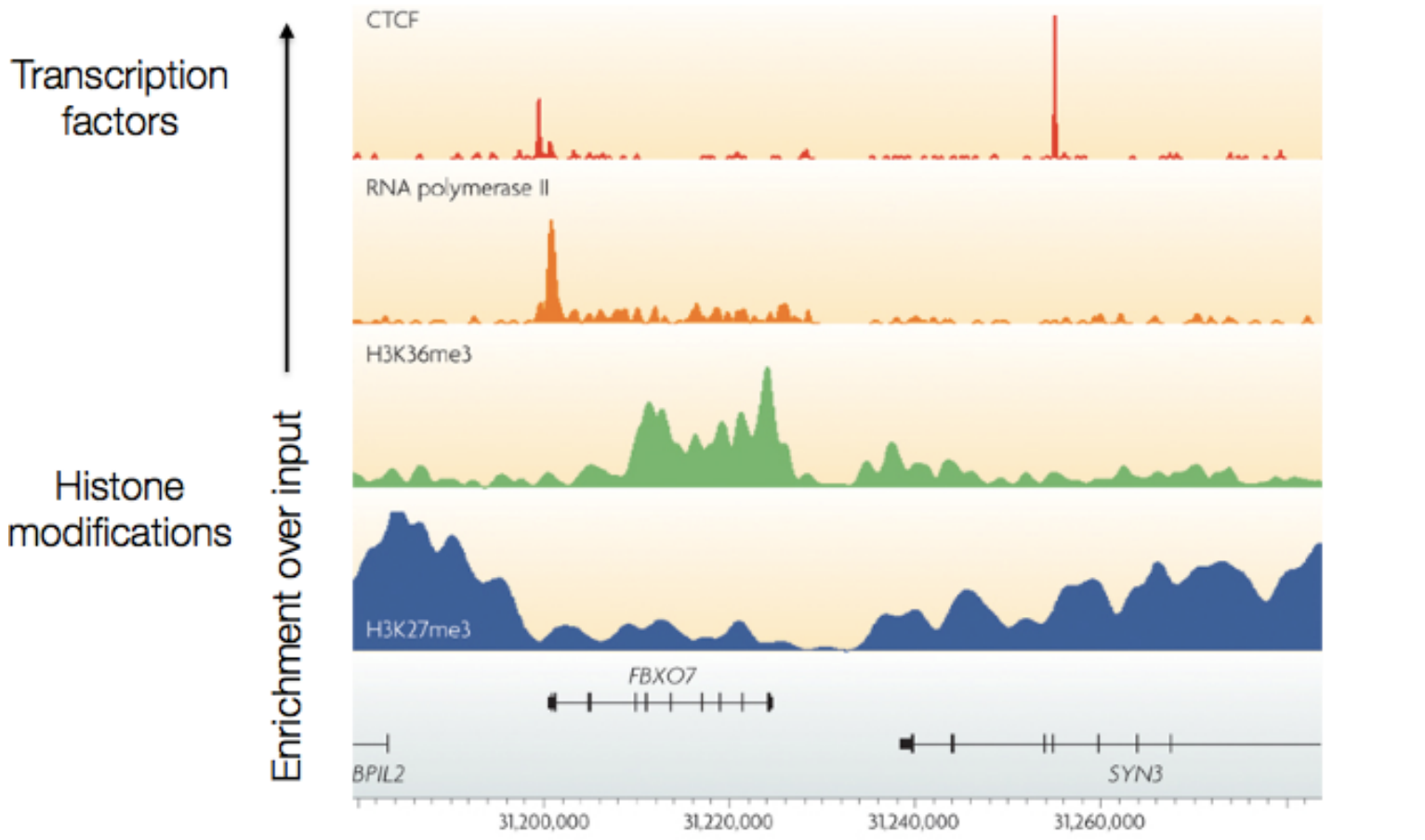


Step 2: call peaks

What regions of the genome show an enrichment for mapped reads?

How does the data analysis work?

Peak calling

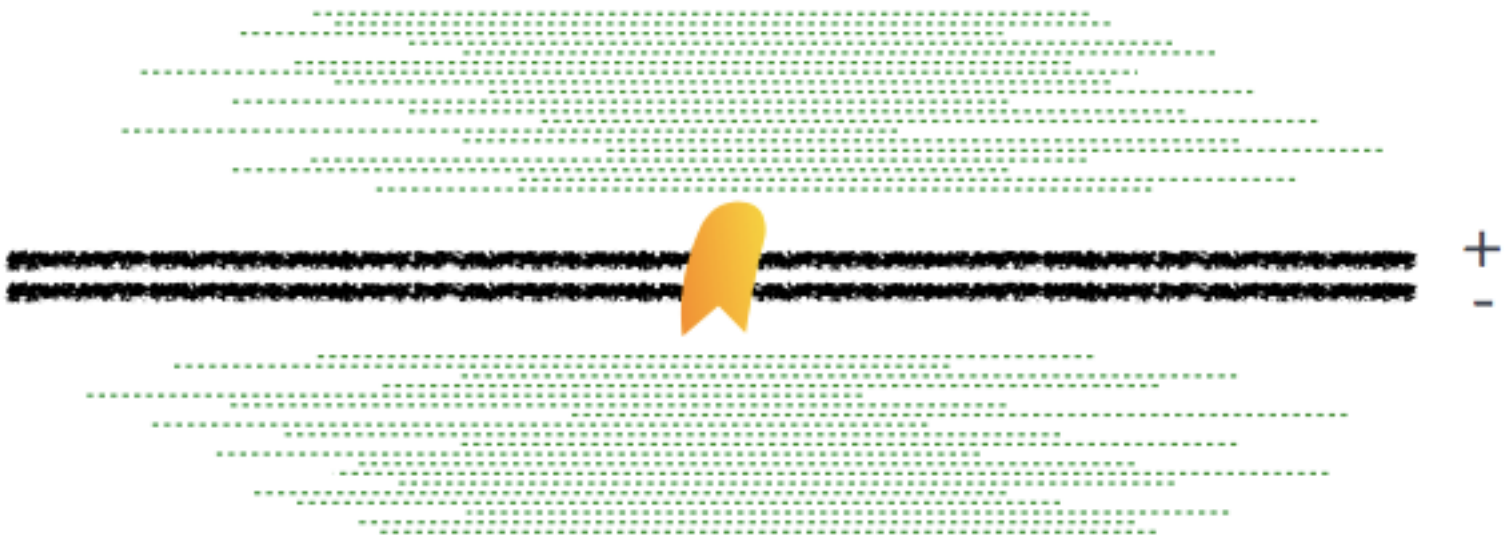


Adapted from Park (2009). Nature Reviews Genetics.

How does the data analysis work?

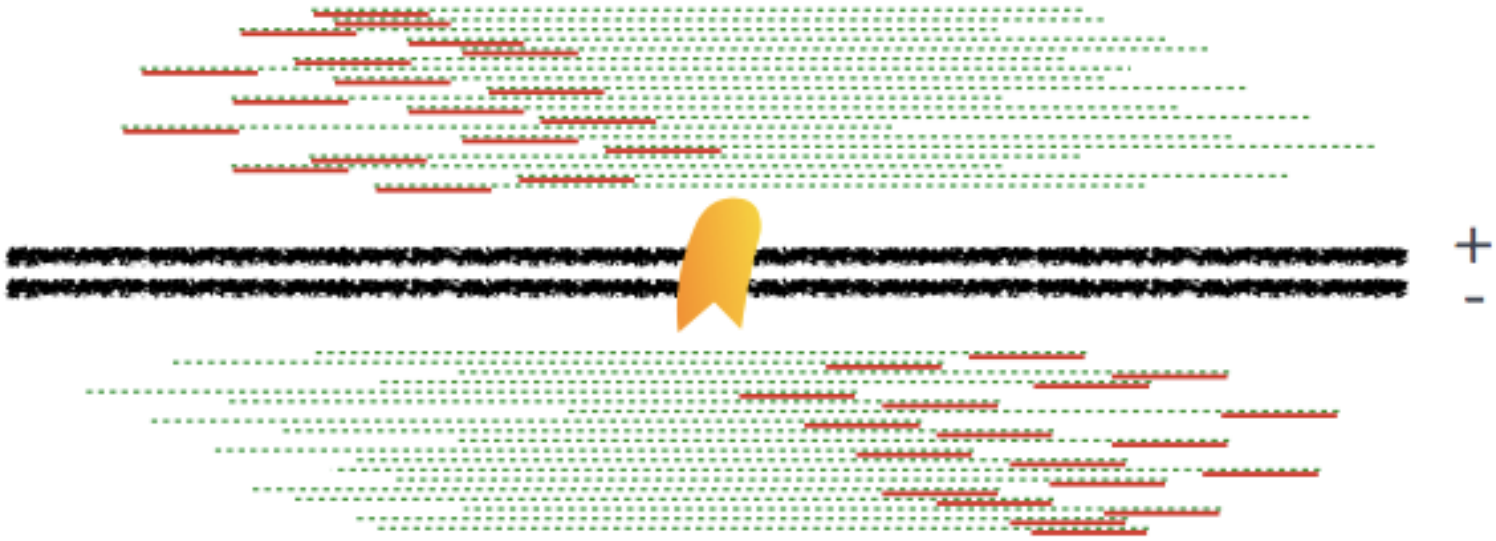
Simplest case: relatively narrow peaks (< 200bp or so)

= binding site
---- = size selected DNA fragment



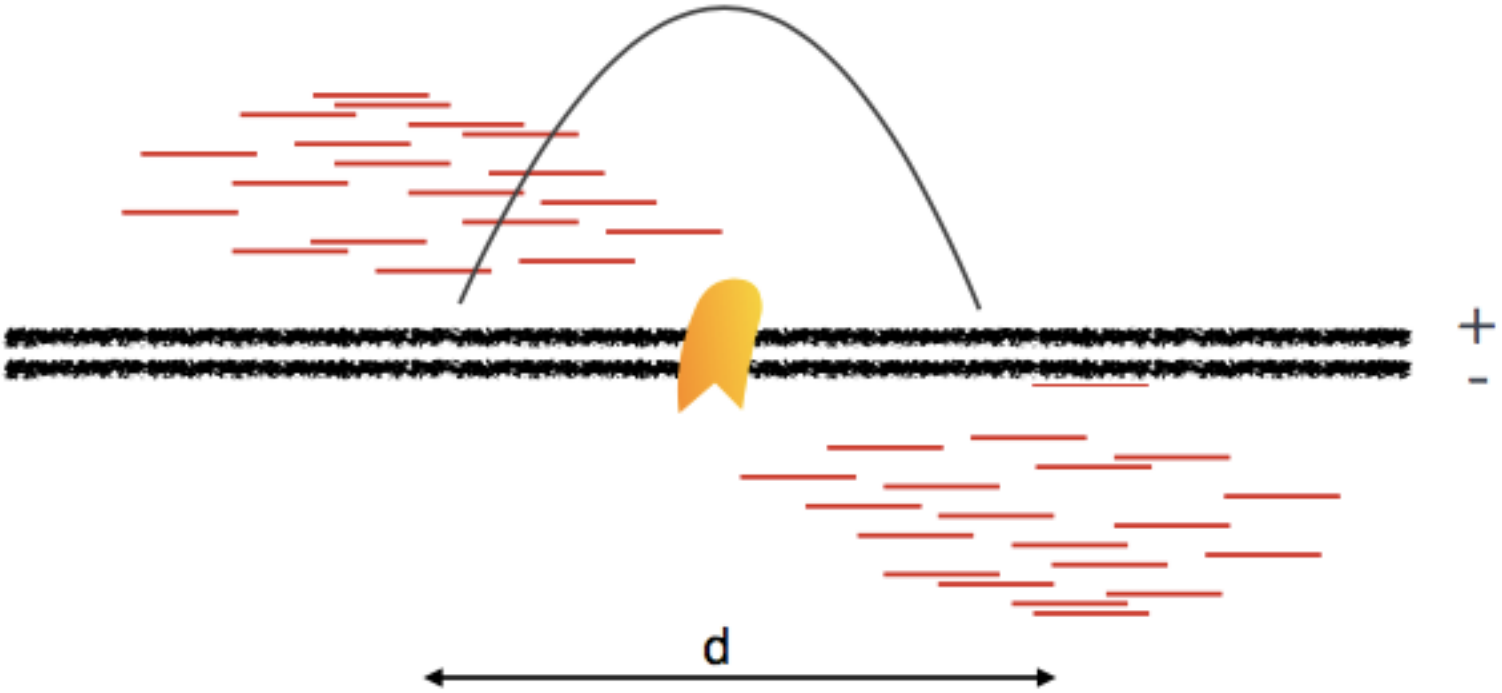
How does the data analysis work?

Simplest case: relatively narrow peaks (< 200bp or so)



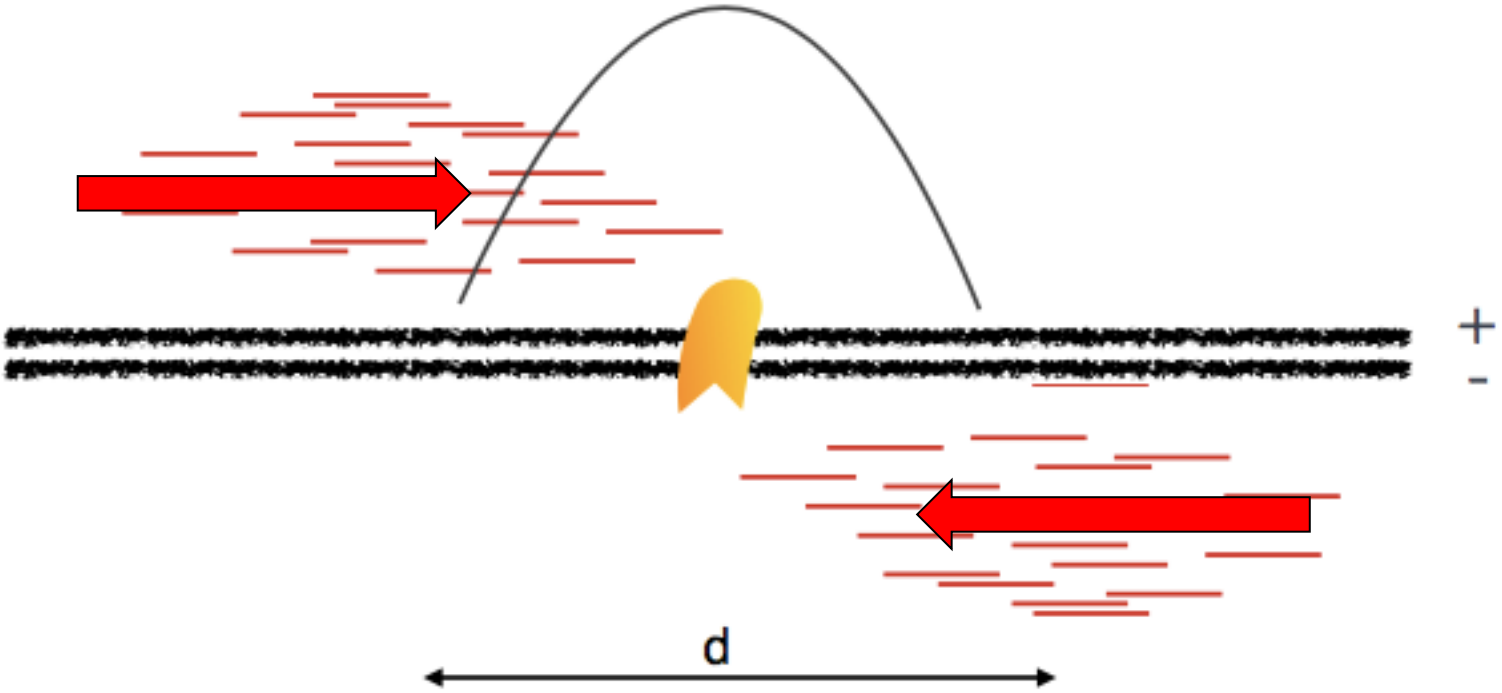
How does the data analysis work?

Simplest case: relatively narrow peaks (< 200bp or so)



How does the data analysis work?

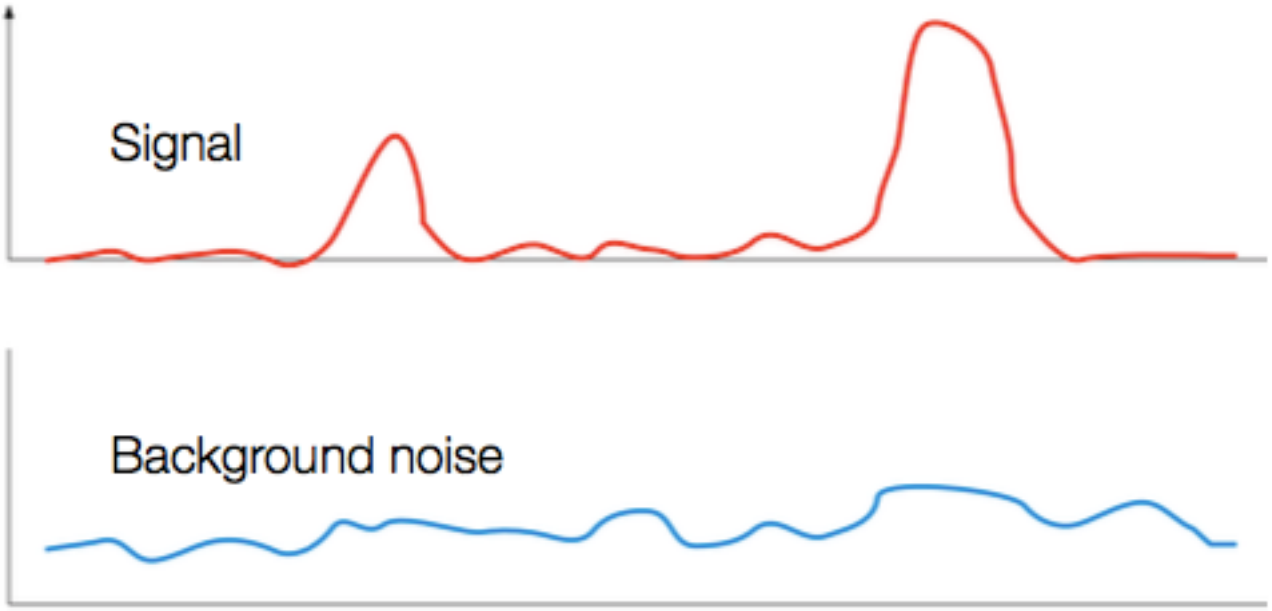
Simplest case: relatively narrow peaks (< 200bp or so)



First part of peak calling: look for regions with reads pointing towards each other

How does the data analysis work?

Simplest case: relatively narrow peaks (< 200bp or so)



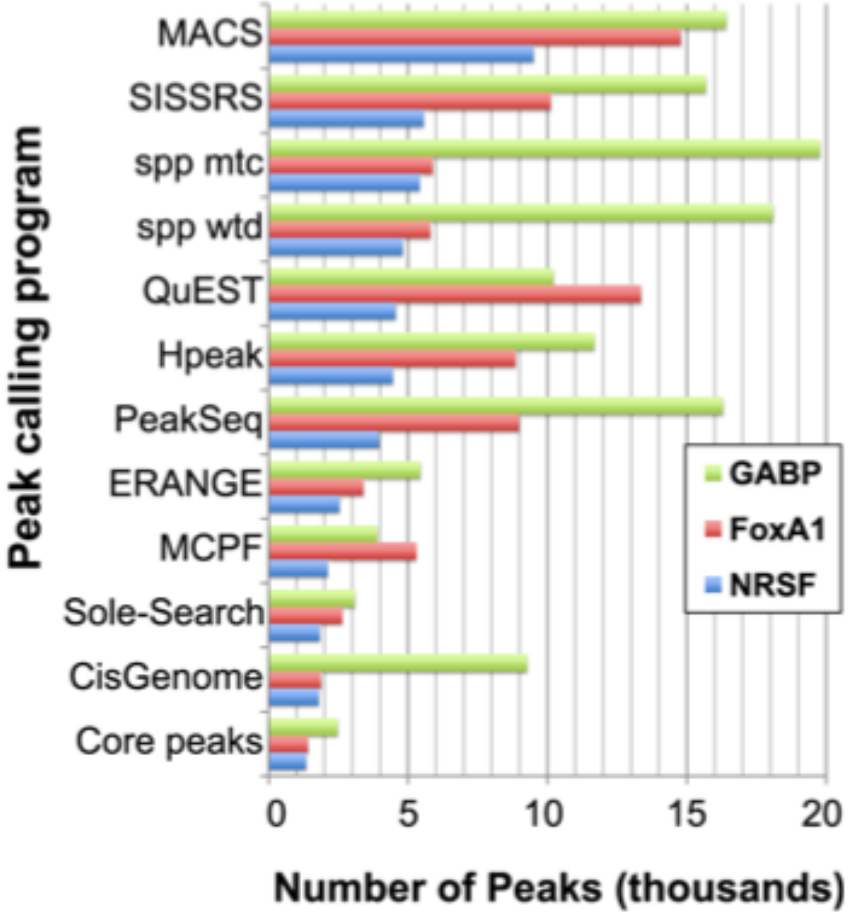
Second component of peak calling: how do you decide when signal is high enough to merit calling a peak, especially when background noise varies by genomic location?

How does the data analysis work?

- Most peak-calling algorithms model the expected number of reads for a genomic region/ window using a Poisson distribution
- Often there is more variance in real data than assumed by the Poisson (overdispersion)
- MACS (model-based analysis of ChIP-Seq) uses multiple Poisson distributions to model the local background noise within each region using the input data

What peak caller to use?

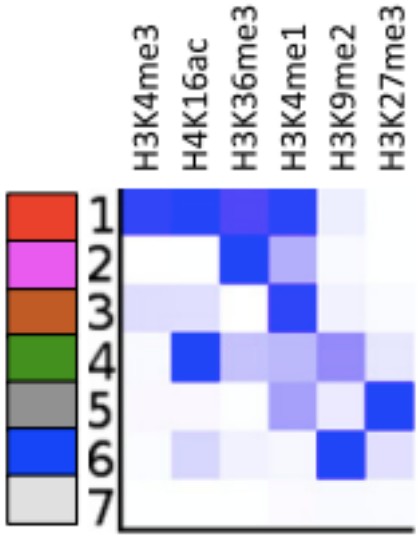
- Most peak callers agree on strongest peaks, but there is variability in number of peaks detected
- Pick one that's widely used
- Pick one that's maintained
- Always perform some visual inspection of peak calling using IGV or similar

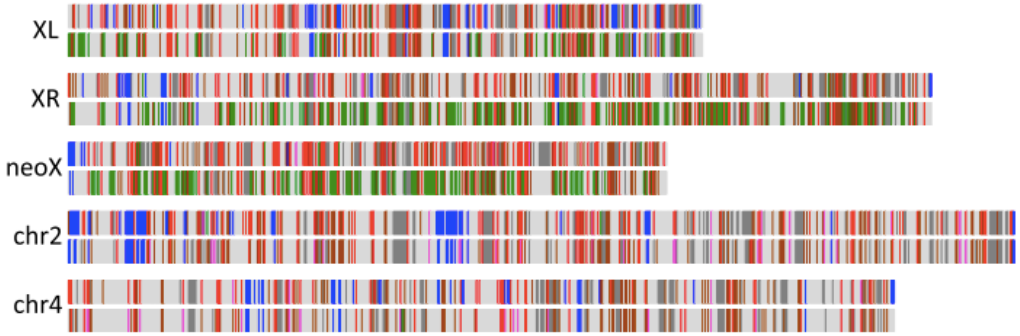


Wilbanks & Facciotti (2010). PLoS ONE.

A note on very broad peaks and multiple datasets

- Some histone modifications cover very large stretches of the genome (can be megabases long), making them recalcitrant to peak callers
- A window-based approach may be more useful here, looking at enrichment for all windows of a given size (5kb, for example)
- If multiple histone modifications are evaluated, there are tools available to identify dominant combinations of marks and annotate the genome **[ChromHMM, ChromClust, segway, others]**





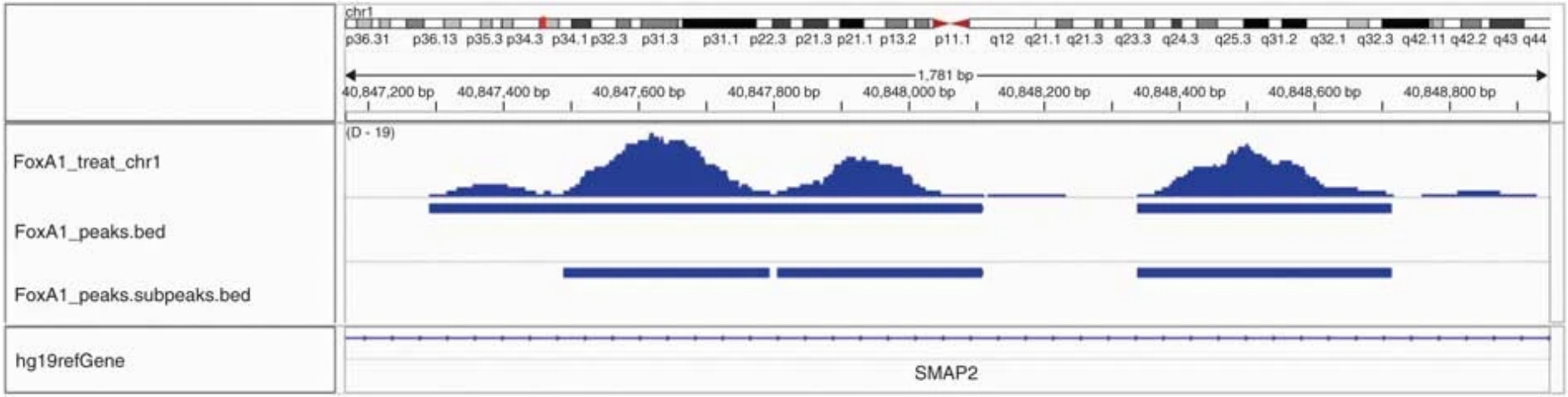
Brown & Bachtrog (2014). Genome Research

Some notes on analysis for ATAC-seq

- There is no input sample for ATAC
  - Some peak callers work just fine with no input (MACS2)
  - In practice, often need a higher signal to confidently call a peak
  - Fortunately, ATAC detects **sharp** peaks, which tend to have higher signal : noise ratio
- Need to remove mitochondrial reads
  - Mito reads often comprise large portion of ATAC-seq reads

How does the data analysis work?

Finally, we have peaks called! Now we know what regions we're going to assess for differential binding



How does the data analysis work?

Finally, we have peaks called! Now we know what regions we're going to look at for downstream analysis

Peak file in bed or xls format

Location of peak

Measures of peak "confidence"

chr	start	end	length	abs_summit	pileup	-log10(pvalue)	fold_enrichment	-log10(qvalue)	name
chr1	3137337	3137639	303	3137365	20.00	15.35908	7.65395	12.62340	../macs/11-27-14_S13_L0
chr1	3139076	3139338	263	3139269	10.00	5.21834	3.75127	2.90311	../macs/11-27-14_S13_L002.macs_peak_2
chr1	3275393	3275646	254	3275422	10.00	6.63368	4.63728	4.22607	../macs/11-27-14_S13_L002.macs_peak_3
chr1	3307970	3308111	142	3308016	13.00	9.74147	6.00700	7.18350	../macs/11-27-14_S13_L002.macs_peak_4
chr1	3308341	3308458	118	3308412	16.00	12.98686	7.29422	10.31853	../macs/11-27-14_S13_L0
chr1	3445931	3446070	140	3446050	8.00	4.96005	3.86165	2.66221	../macs/11-27-14_S13_L002.macs_peak_6
chr1	3501589	3501712	124	3501642	9.00	5.80498	4.26713	3.45147	../macs/11-27-14_S13_L002.macs_peak_7
chr1	3671086	3671714	629	3671176	12.00	7.67499	4.93348	5.21088	../macs/11-27-14_S13_L002.macs_peak_8
chr1	3749453	3749567	115	3749456	8.00	4.28342	3.42291	2.05093	../macs/11-27-14_S13_L002.macs_peak_9
chr1	3749864	3750392	529	3749923	24.00	20.46337	9.44646	17.60649	../macs/11-27-14_S13_L0
chr1	3847111	3847317	207	3847201	22.00	18.89270	9.14512	16.07044	../macs/11-27-14_S13_L0
chr1	3896484	3896708	225	3896548	18.00	15.27619	8.15236	12.54271	../macs/11-27-14_S13_L0
chr1	4020898	4021014	117	4020919	7.00	4.03185	3.37257	1.84224	../macs/11-27-14_S13_L002.macs_peak_13
chr1	4021916	4022125	210	4022075	13.00	9.68809	5.97399	7.13403	../macs/11-27-14_S13_L002.macs_peak_14
chr1	4022290	4022483	194	4022420	9.00	5.83763	4.28806	3.47887	../macs/11-27-14_S13_L002.macs_peak_15

Small(ish) files – one line per peak, so maybe just 100k's of lines

Common QC metrics for ChIP-seq

https://www.encodeproject.org/data-standards/

Specific metrics vary for TF vs. histone modification

Broadly, need to have biological replicates, input for each replicate, and well-characterized antibody

Sequencing depth guidelines vary depending on expected “peaky-ness” of mark

Common QC metrics for ATAC-seq

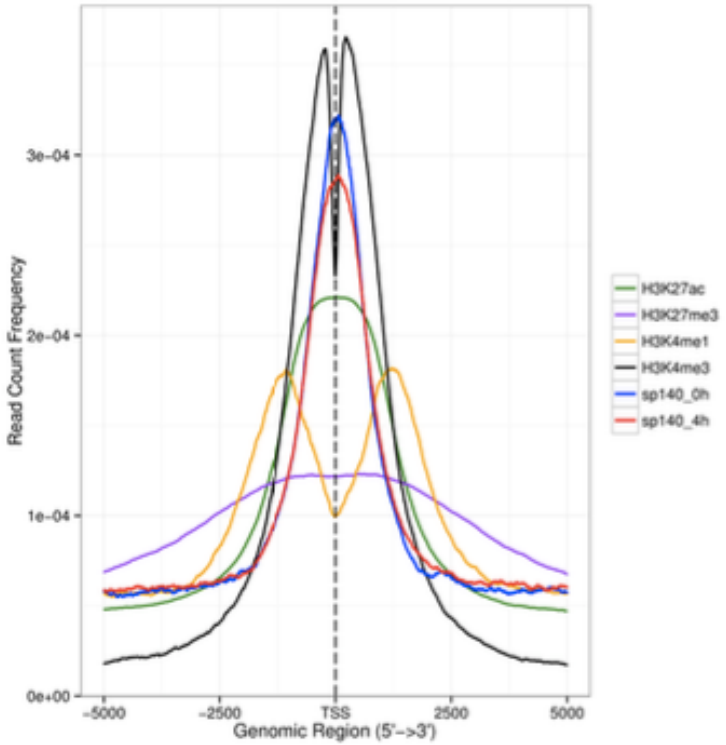
https://www.encodeproject.org/atac-seq/

FRiP score: **F**raction **R**eads **i**n **P**eaks
→ of all the reads sequenced, what proportion map to regions called as peaks?
(ENCODE suggests 0.2 or higher for high quality experiment)

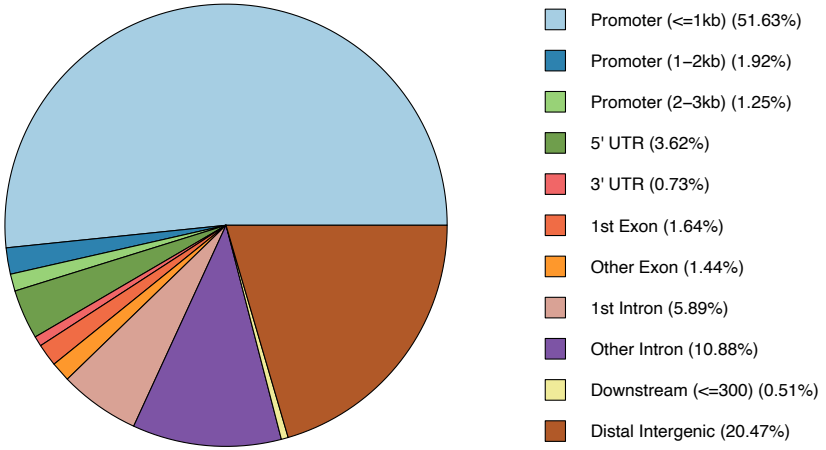
TSS Enrichment Score: (specific for ATAC)
→ How much higher is signal within 1kb of TSS than flanking 100bp?
(ENCODE suggests score of 5 or higher for high quality experiment)

Downstream analyses

Annotation of peaks– where are they relative to TSS or other genomic features?



ATAC peak set

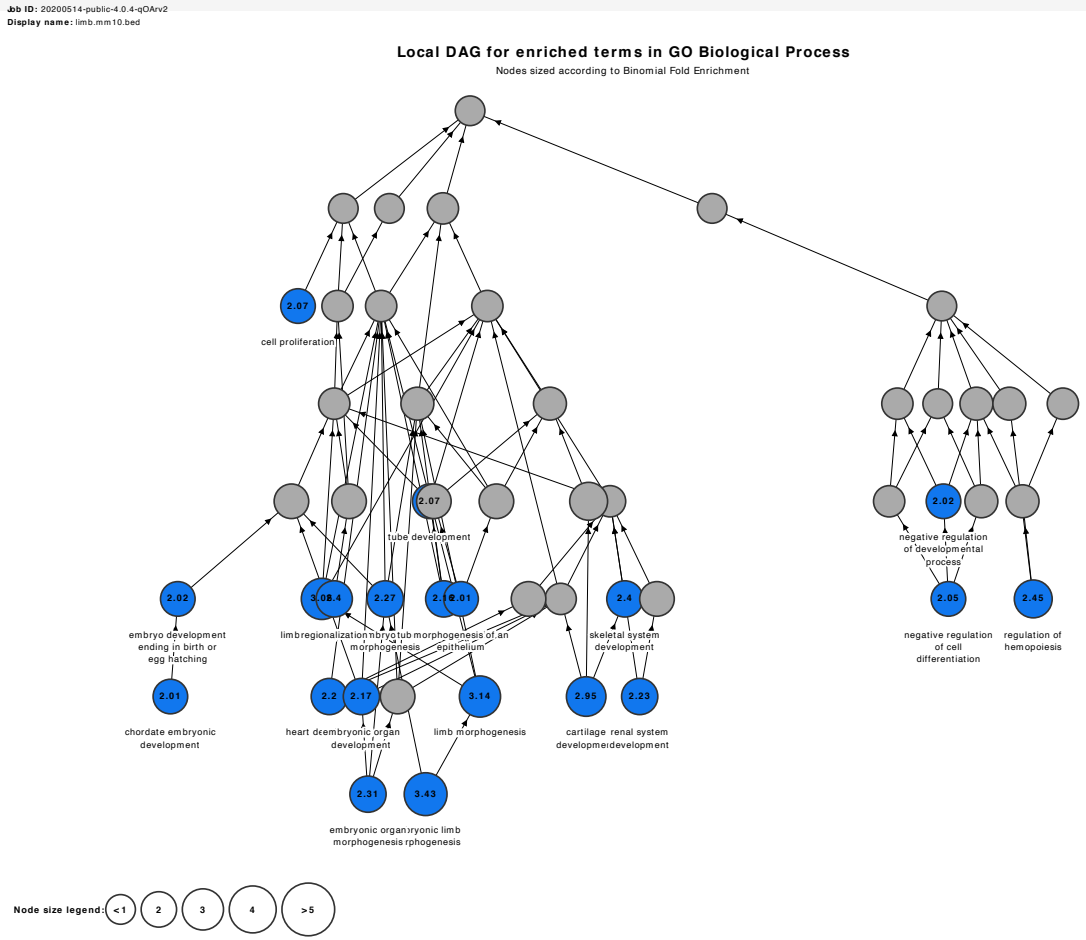


Downstream analyses

Functional enrichment analysis http://great.stanford.edu/public/html/index.php

GREAT: Genomic Regions Enrichment of Annotations Tool

GREAT predicts functions of *cis*-regulatory regions.



Downstream analyses

Motif discovery

[MEME suite, Homer, ChiLin]

Homer Known Motif Enrichment Results

Homer *de novo* Motif Results
Gene Ontology Enrichment Results
Known Motif Enrichment Results (txt file)
Total Target Sequences = 942, Total Background Sequences = 47744

Rank	Motif	Name	P-value	log P-pvalue	q-value (Benjamini)	# Target Sequences with Motif	% of Targets Sequences with Motif	# Background Sequences with Motif	% of Background Sequences with Motif
1		Fli1(ETS)/CD8-FLI-ChIP-Seq(GSE20898)/Homer	1e-37	-8.733e+01	0.0000	572.0	60.72%	18964.1	39.72%
2		PU.1(ETS)/ThioMac-PU.1-ChIP-Seq(GSE21512)/Homer	1e-37	-8.561e+01	0.0000	289.0	30.68%	6811.9	14.27%
3		Elf4(ETS)/BMDM-Elf4-ChIP-Seq(GSE88699)/Homer	1e-36	-8.420e+01	0.0000	505.0	53.61%	15934.4	33.38%

Downstream analyses

Differential binding

In principal, this is very similar to differential expression analysis

- It's count data, but instead of counting reads from RNA in genes, counting reads in peaks (or windows)

- Many differential binding tools actually use differential expression tools (DESeq2, EdgeR) in the background

- MACS2 also has a differential binding tool with looser replicate requirements

Downstream analyses

Differential binding

Quick refresher on how tools like DESeq2 work

- normalize counts between samples to account for differences in library size
- estimate dispersion for each condition
- fit a negative binomial model and test for deviation using Wald test or Likelihood Ratio Test

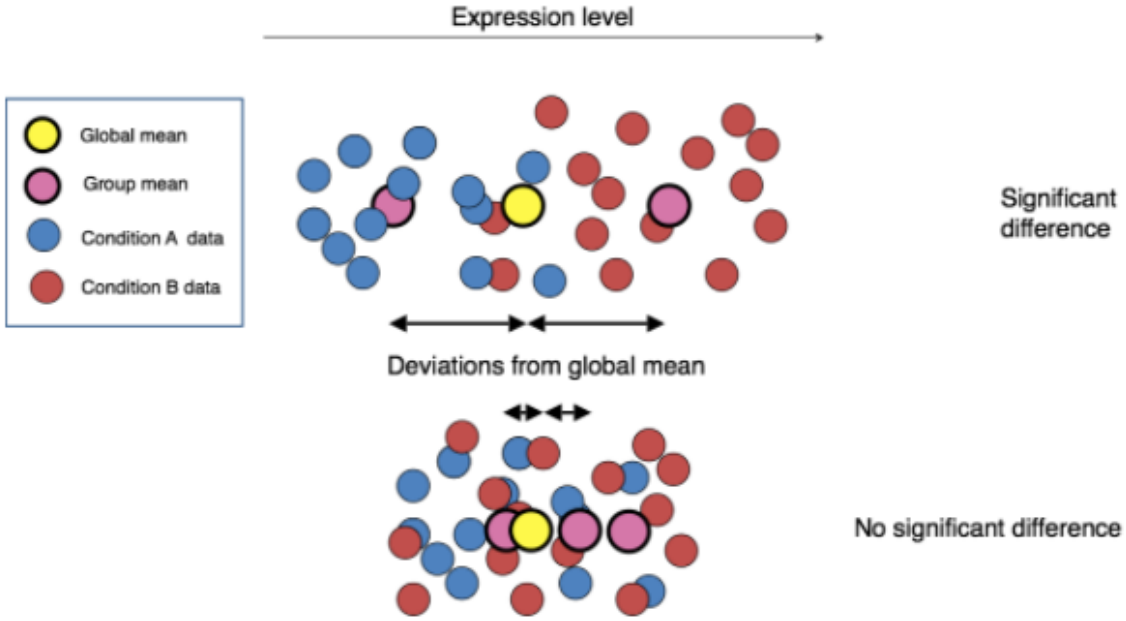
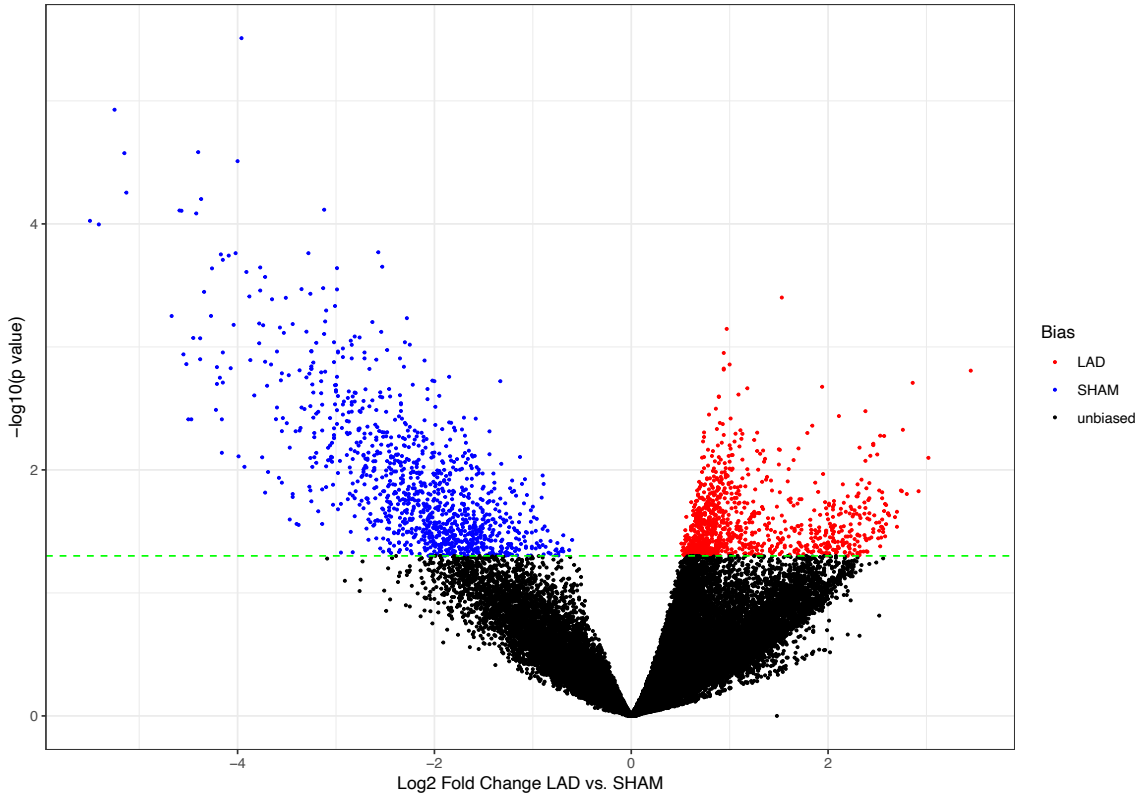


Image credit: Paul Pavlidis, UBC

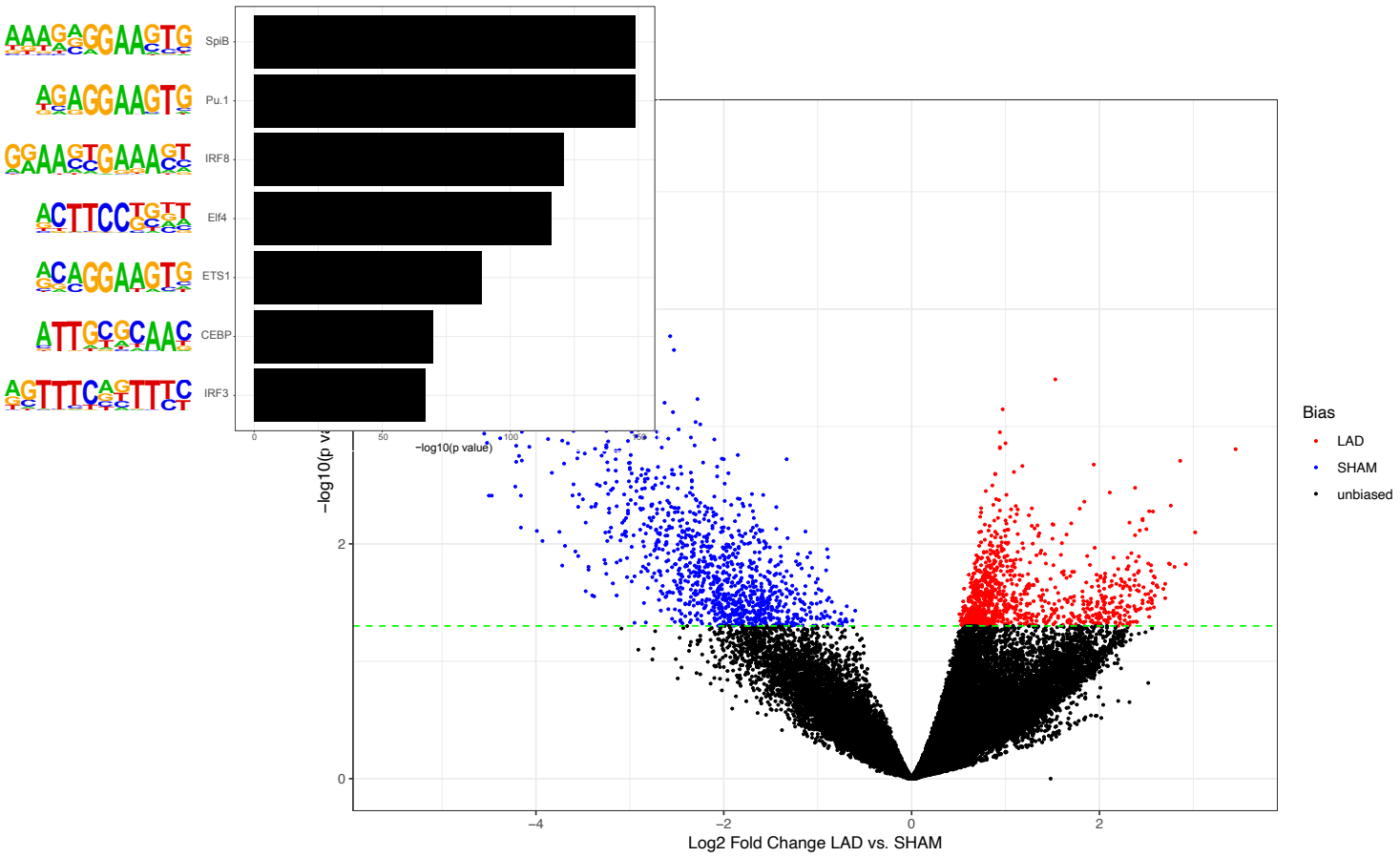
Downstream analyses

Differential binding



Downstream analyses

Differential binding



Downstream analyses

Differential binding

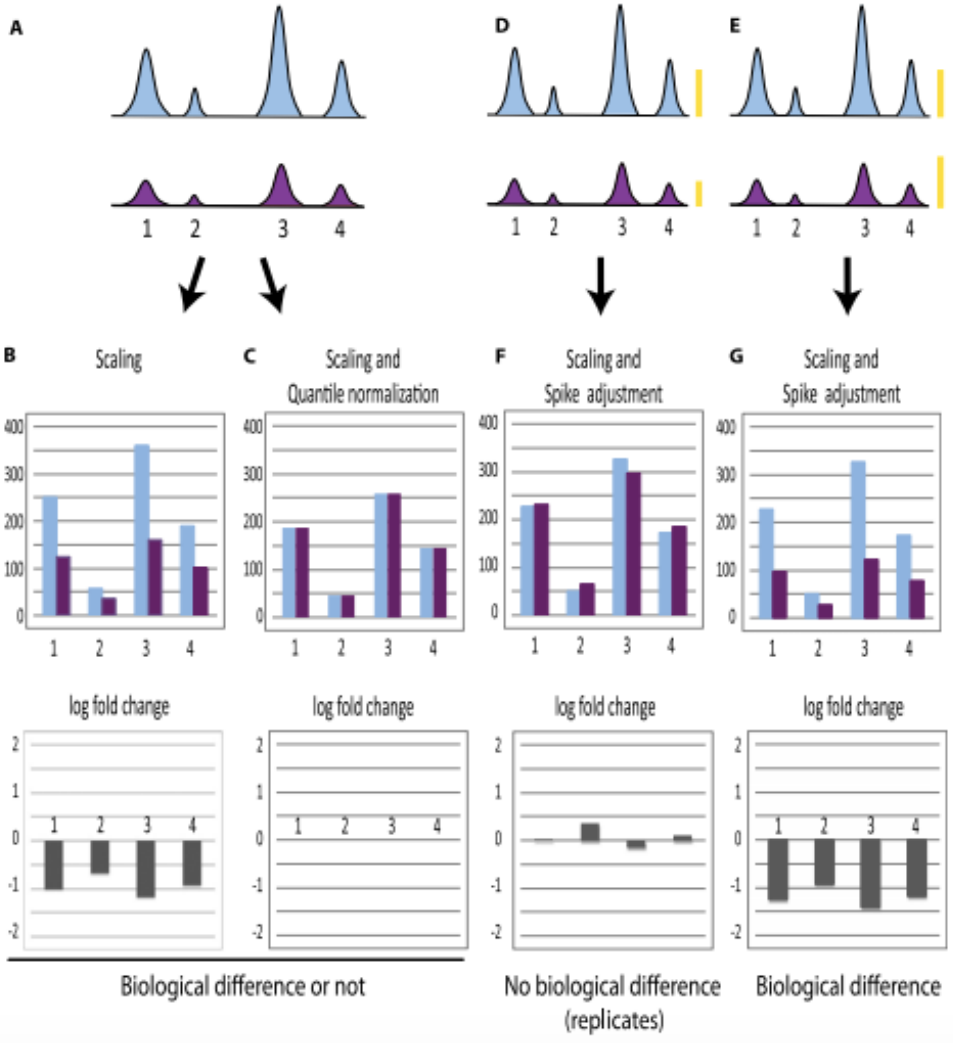
Addition of a “spike-in” can aid in discovery of differential binding

For ChIP:

- use a spike of another species with antibody cross-reactivity

For ATAC:

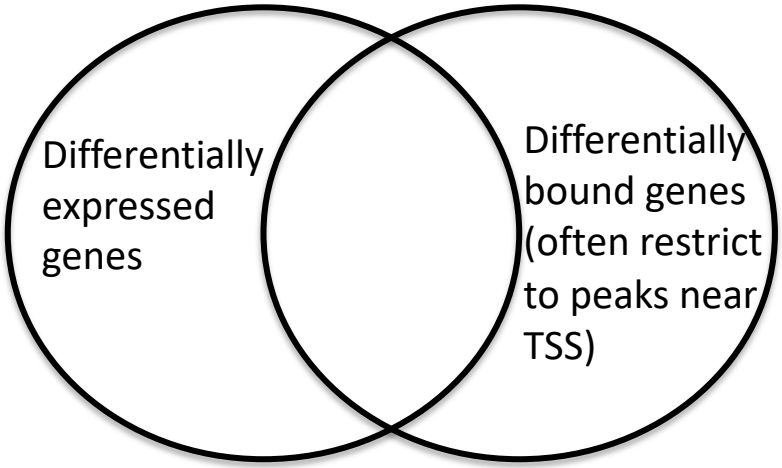
- spike in live cells of another species (Stewart-Morgan et al (2019). Molecular Cell)



Bonhoure et al (2014). Genome Research.

Downstream analyses

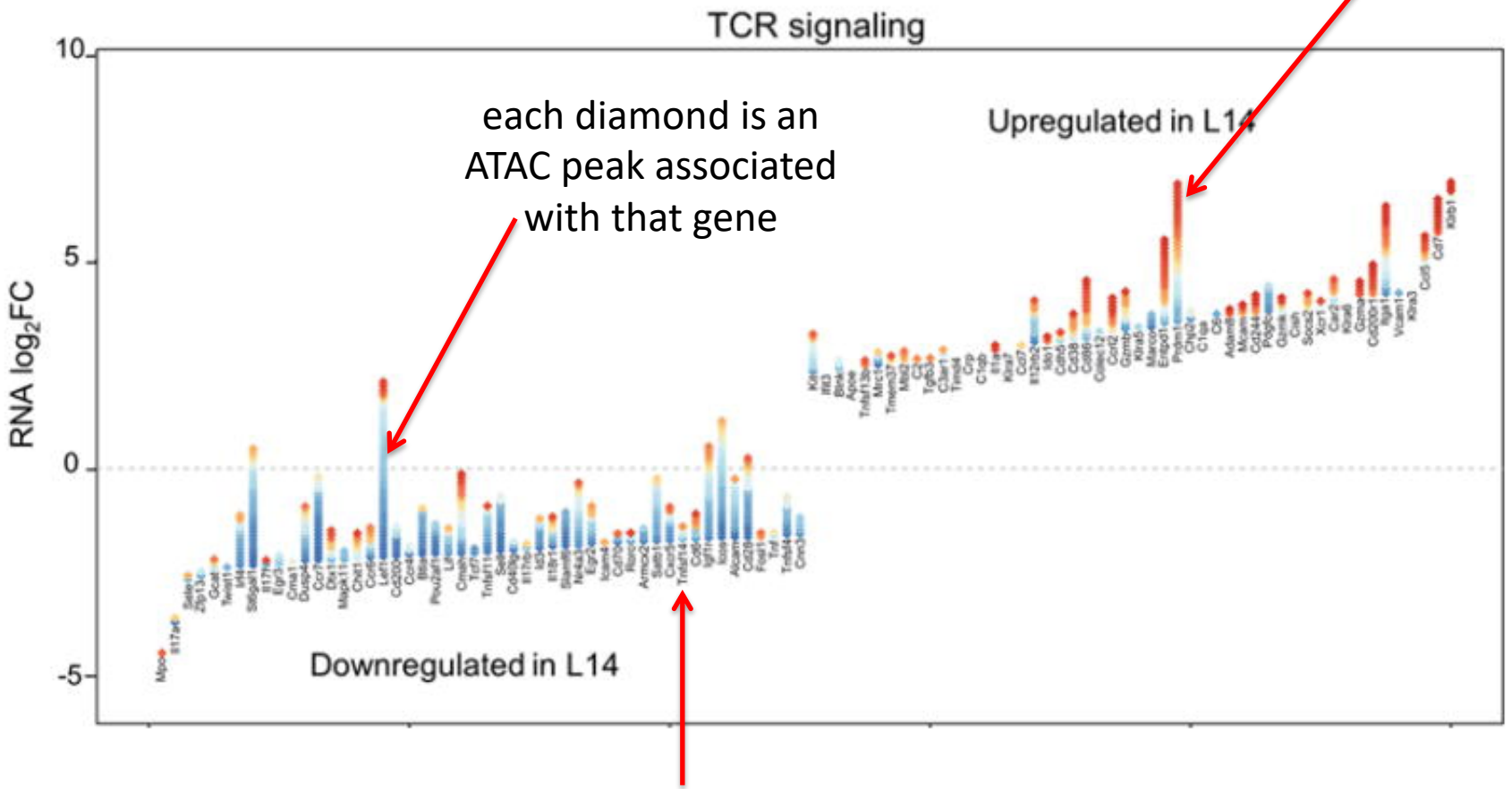
Differential binding and integrating with RNA-seq



Downstream analyses

Differential binding and integrating with RNA-seq

color of diamond indicates gain of peak (red) or loss of peak (blue)



Philip et al (2017). Nature.

Wrapping up

- Basics of ChIP and ATAC-seq protocol
  - When is each appropriate? Depends on research questions
- Basics of data analysis workflow
  - Focus on why and how peaks are called
  - How data analysis differs for ChIP-seq and ATAC-seq datasets
  - Considerations based on “peaky-ness” of data
  - Some ideas for downstream analyses

Questions?

Thank you for listening!