Epigenomics data analysis

11/30/2020 – 12/18/2020

Qi Sun, William Lai, and Jeff Glaubitz

Bioinformatics Facility & Epigenomics Facility

Genome feature annotation & Epigenomics data analysis

Week 1
Genome features

- File formats
- Public databases
- Genome browser

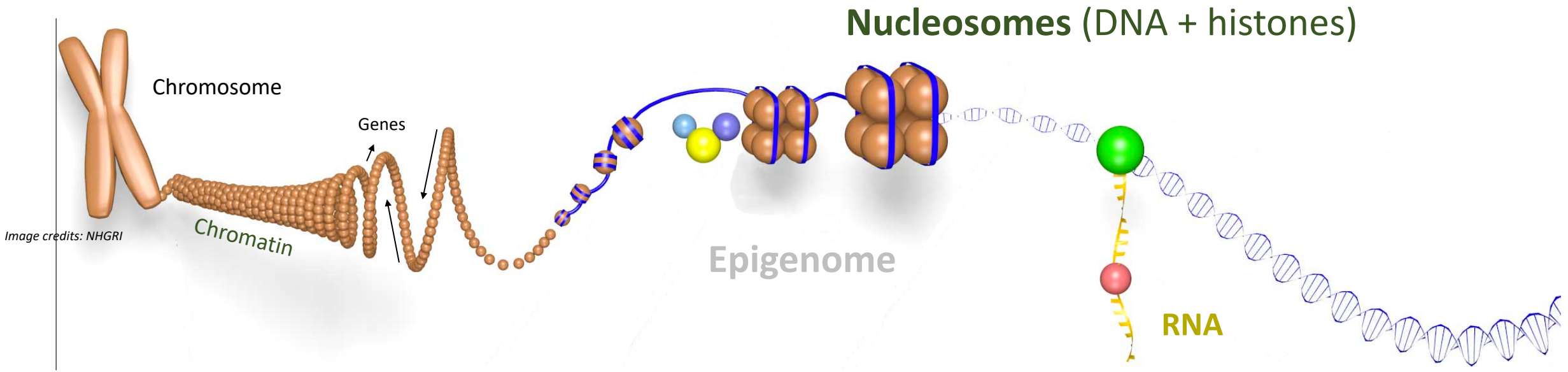
Week 2
Peak calling

- Peak calling
- Data QC

Week 3
Integrated analysis

- Bedtools
- Deeptools
- Homer

What is the Epigenome?



What is the Epigenome?

- We define the epigenome here as ‘epi-’ (on top of) the genome
  - i.e. Histone modifications, transcription factor binding, 3D nuclear architecture

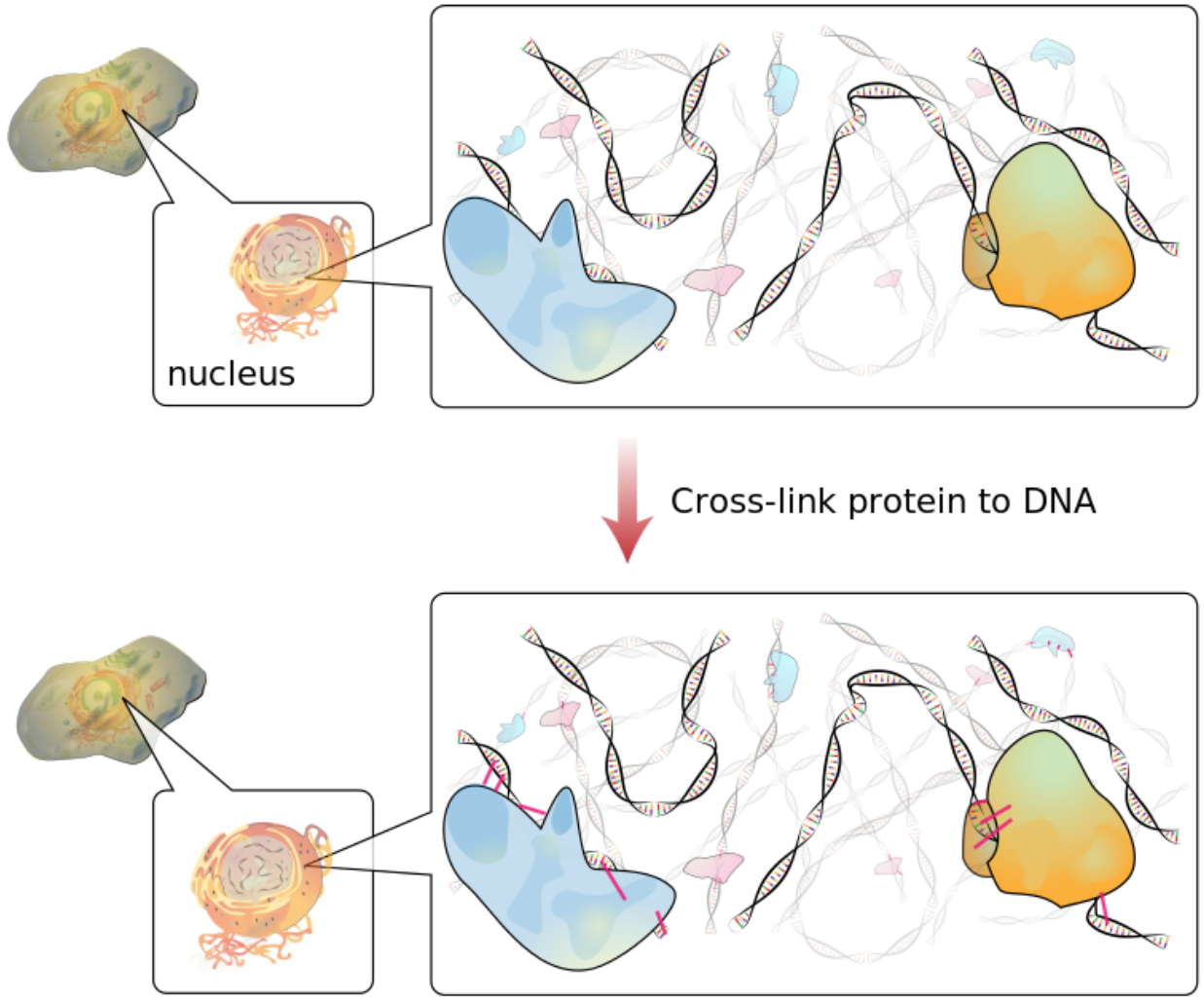
- This is semantically (but not *necessarily* biologically) distinct from epigenetic
  - i.e. DNA methylation

- The nature of inheritance is beyond the scope of this workshop!

Common epigenomic assays

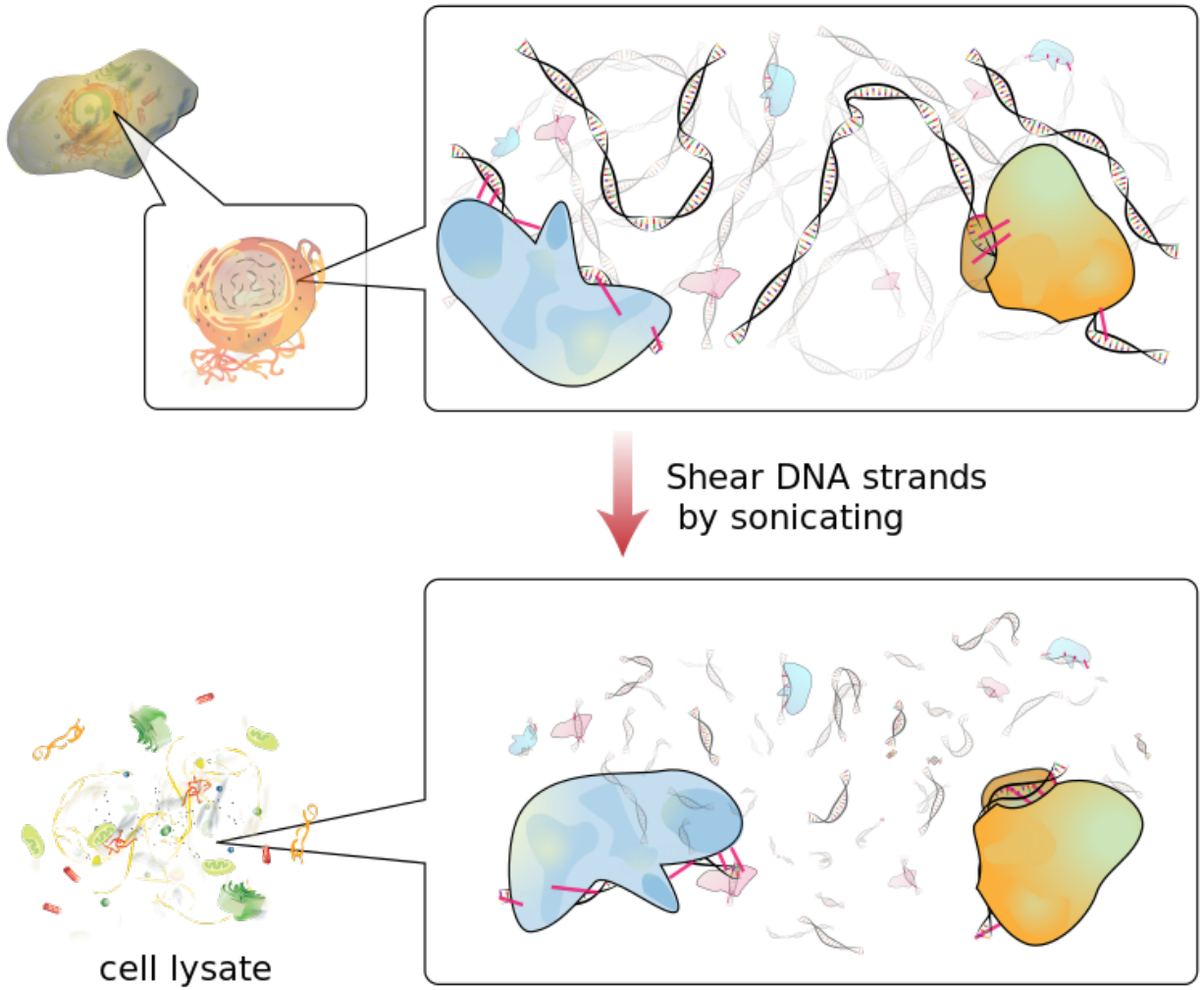
- Measure the expression of genes/enhancers
  - RNA-seq, PRO-seq, NET-seq, START-seq, etc…
- Measure the binding of proteins
  - ChIP-chip, ChIP-seq, ChIP-exo, CUT&TAG, etc…
- Measure accessible regions of the genome
  - MNase-seq, DNase-seq, ATAC-seq, CUT&RUN, etc…
- Measure the 3-D architecture of the genome
  - 4C, 5C, HI-C, SPRITE, GAM, etc…
- Specialized assays abound!
  - Single-cell variants of much of the above
  - STARR-seq (enhancer characterization)
  - Repli-seq (mapping replication)

Chromatin immunoprecipitation (ChIP-seq)



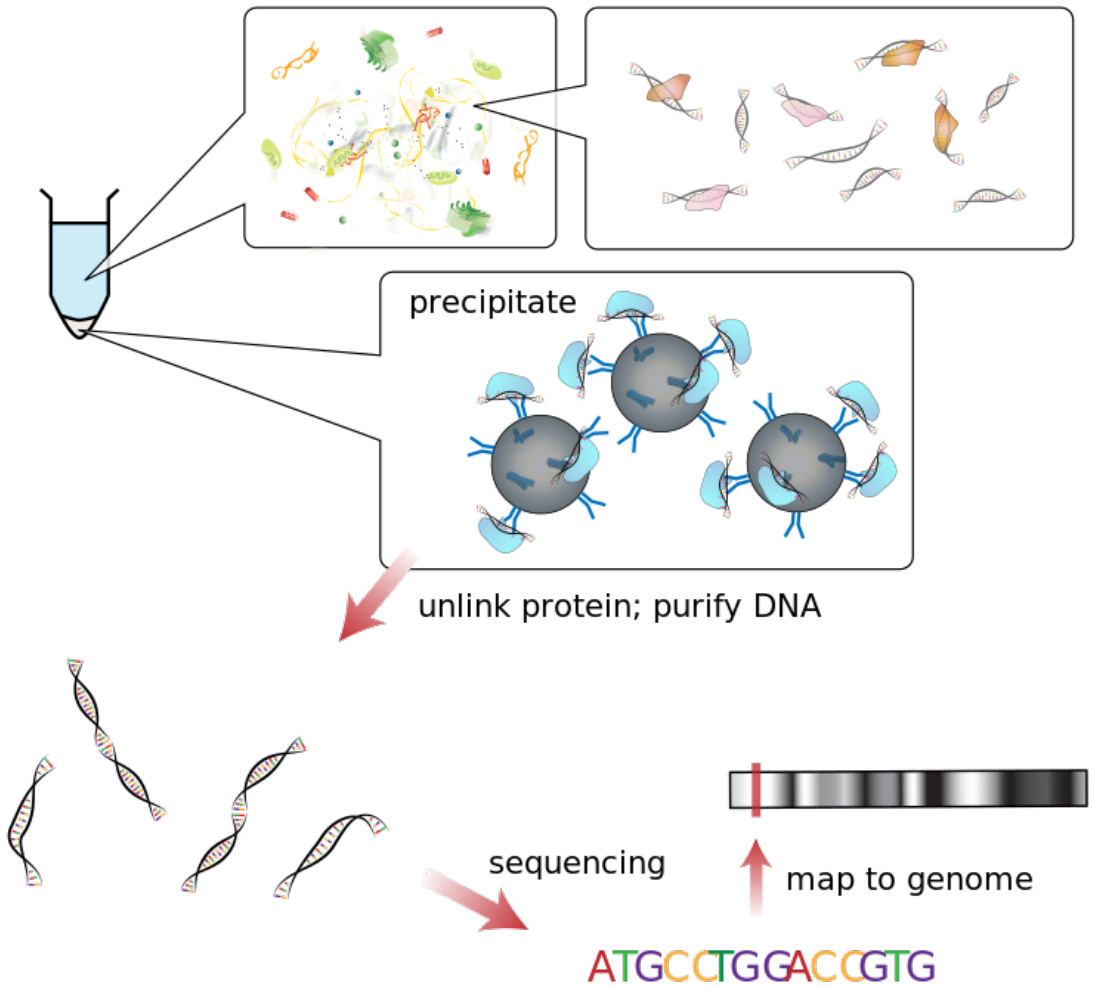
https://en.wikipedia.org/wiki/Chromatin_immunoprecipitation

Chromatin immunoprecipitation (ChIP-seq)



https://en.wikipedia.org/wiki/Chromatin_immunoprecipitation

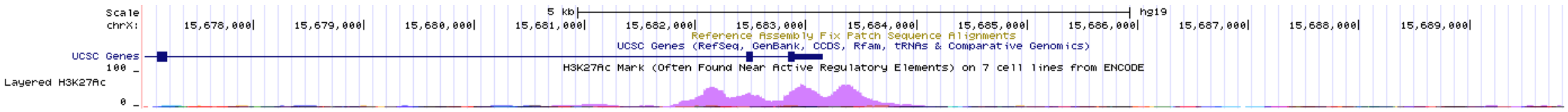
Chromatin immunoprecipitation (ChIP-seq)



https://en.wikipedia.org/wiki/Chromatin_immunoprecipitation

What does the data look like?

Gene browser level



High Occupancy

chr5:546658-546666(-)

chr3:289396-289404(+)

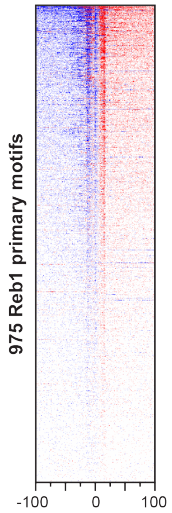
chr14:652001-652009(-)

chr4:74266-74274(-)

chr13:135333-135341(+)

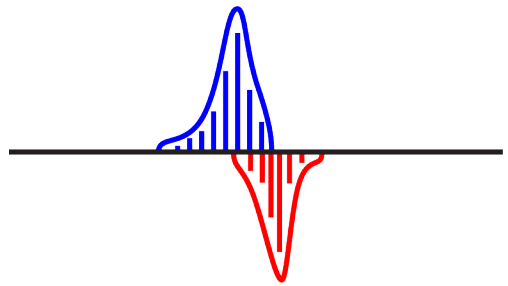
Low Occupancy

Heatmap



Average

Composite plots



Principles of peak calling

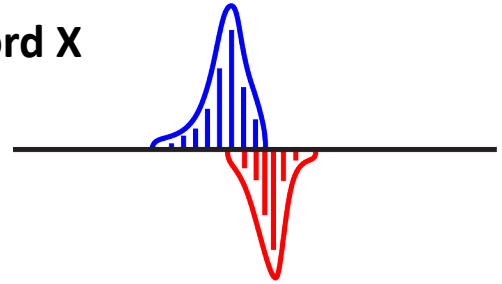
Peak!
Protein is bound here

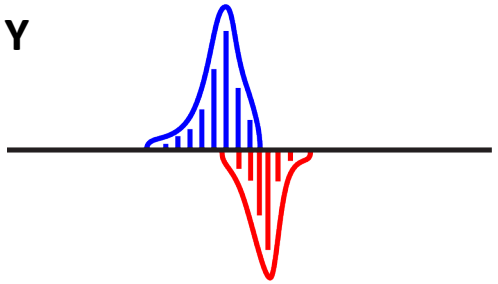
Background is high here!
Unknown if protein is bound here

Genome coord X

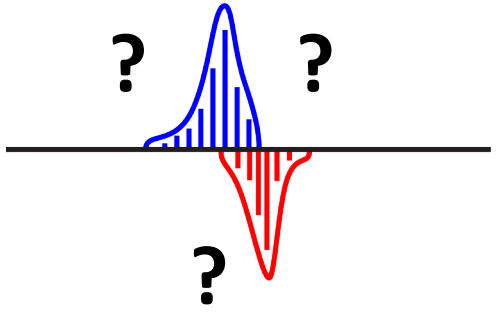
Genome coord Y

Sample





Control
i.e. Mock, Input, Both



Blacklist regions

- Specific regions of the genome that produce artefactual data often irrespective of assay or target
- What causes this?
  - Lots of different and confounding reasons…
  - https://www.nature.com/articles/s41598-019-45839-z
- What can I do about it?
  - Generally recommended to filter them out
  - BUT… Many of these regions are interesting in their own right and you might miss potentially novel biological function by ignoring them completely.
  - Your use case will weigh heavily in how you choose to handle them

Yeast (sacCer3) Blacklist region

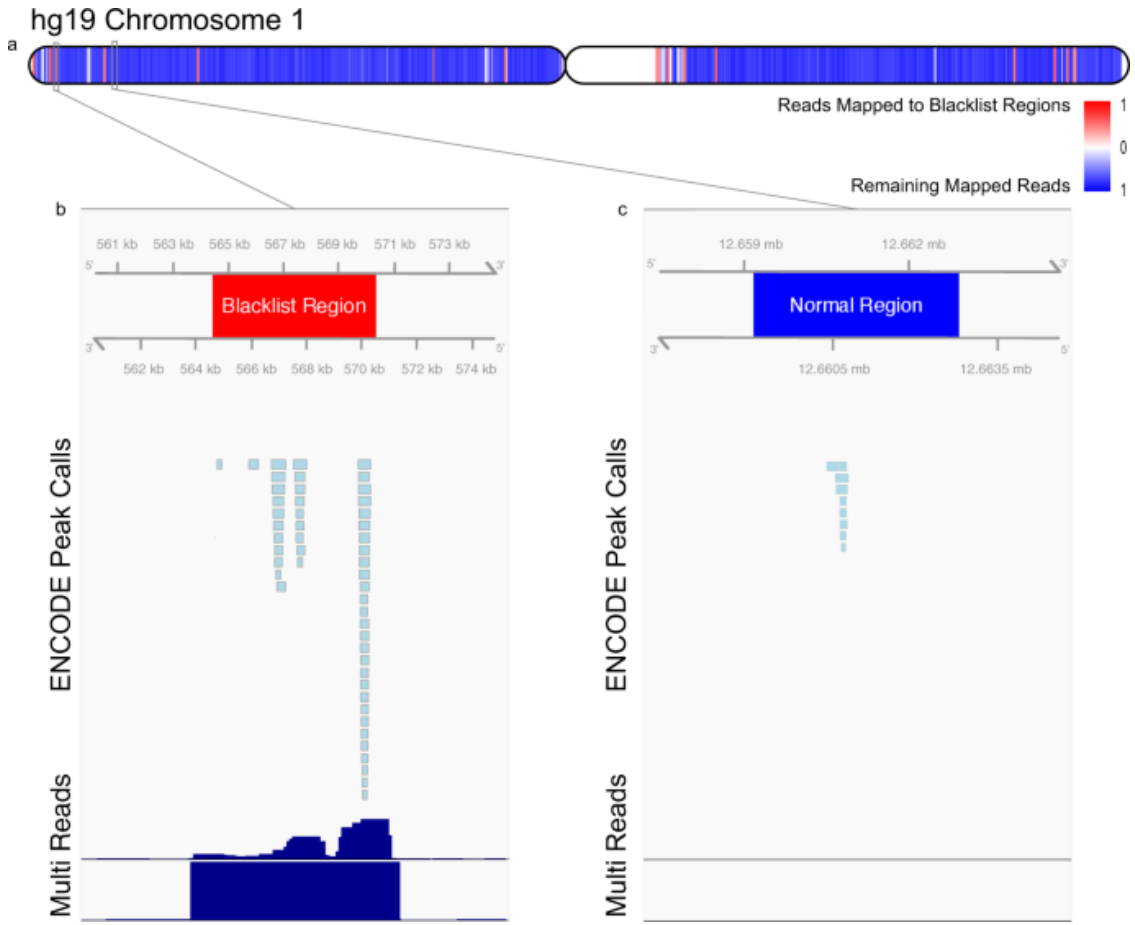
BED format

```
chrXII  451787  468932  rDNA_Locus              .       +
chrM    1       85779   mitochondrial_genome    .       +
```

Blacklist region references

- hg38: https://www.encodeproject.org/files/ENCFF356LFX/
- hg19: https://www.encodeproject.org/files/ENCFF001TDO/
- mm10: https://www.encodeproject.org/annotations/ENCSR636HFF/
- mm9: http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/mm9-mouse/mm9-blacklist.bed.gz
- ce10: http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/ce10-C.elegans/ce10-blacklist.bed.gz
- dm3: http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/dm3-D.melanogaster/dm3-blacklist.bed.gz
- Your particular organism/strain may require its own list depending on your needs/use-case

Blacklist mapping as a Quality Control metric



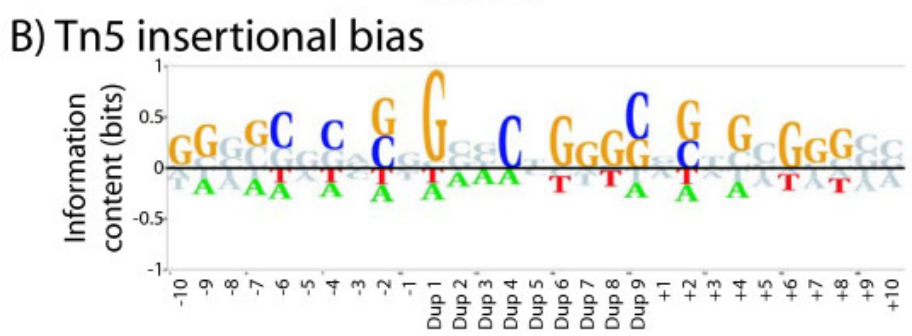
A large number of sequence reads mapping to blacklist regions is typically a sign of a poor or failed ChIP.

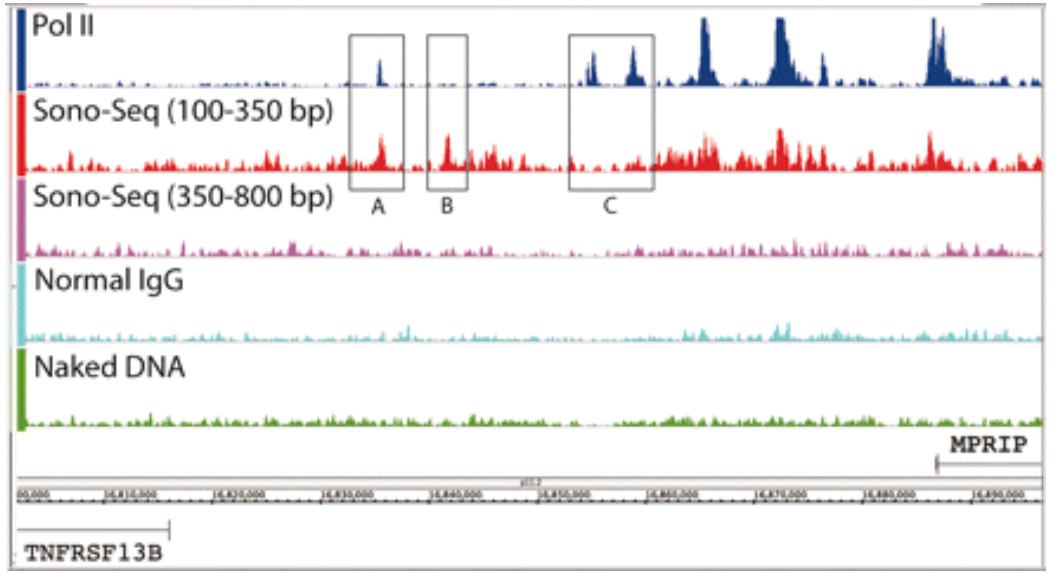
Extreme example: Up to 87% of reads map to blacklist in low quality ENCODE ChIP-seq

https://www.nature.com/articles/s41598-019-45839-z

Bias abounds!

- Bias exists intrinsically within EVERY genomic experiment
- Many attempts have been made to correct it, but it is NOT a solved problem

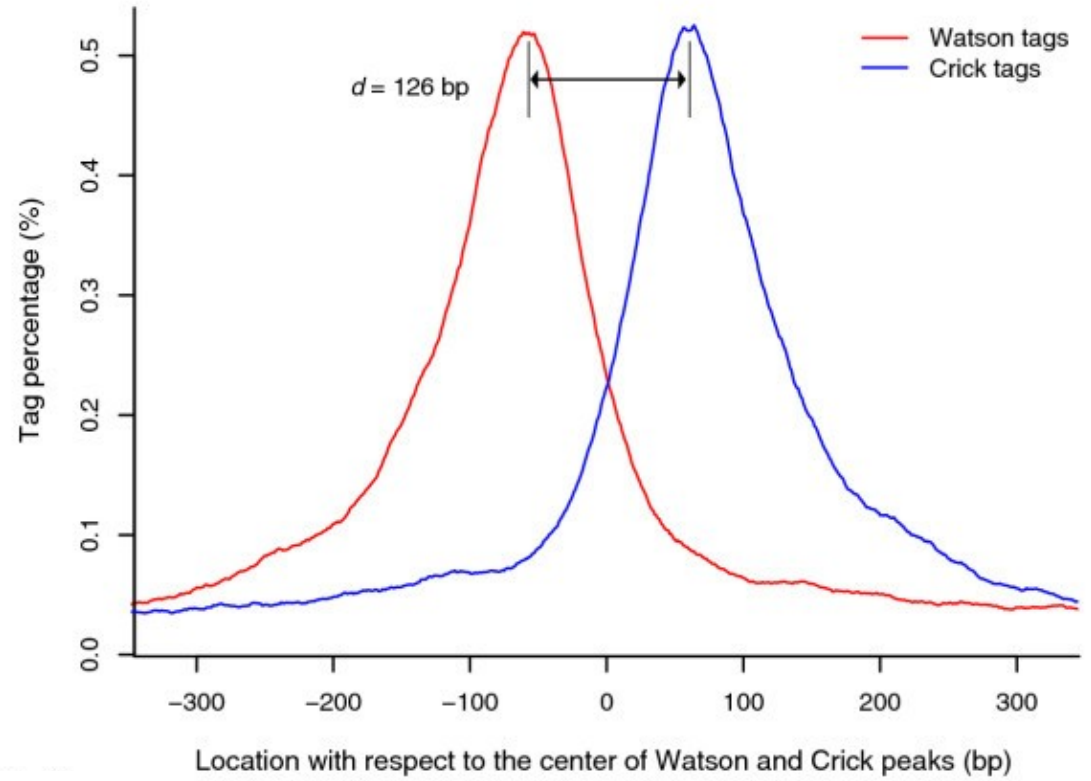




- https://pubmed.ncbi.nlm.nih.gov/22313799/
- https://www.pnas.org/content/106/35/14926
- https://academic.oup.com/nar/article/46/2/e9/4602870

Peak calling algorithms

Model-based Analysis of ChIP-Seq (MACS2)



- https://genomebiology.biomedcentral.com/articles/10.1186/gb-2008-9-9-r137
- https://github.com/macs3-project/MACS

Step 1: Model fragment size as a function of the 1,000 most enriched regions in the genome

Step 2: Shift reads 5'->3' by that fragment size to make a single vector of data

Step 2.5 (Optional): Scale Sample and Control libraries relative to each other

Step 3: Calculate p-val and call candidate peaks based on fragment size window

Step 4: Call final peaks (FDR using B-H correction relative to control)

Genome wide Event finding and Motif discovery (GEM)

- GEM have been adopted by ENCODE as an additional recommended ChIP-seq peak-caller
  - https://groups.csail.mit.edu/cgs/gem/
  - https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002638
- Algorithmic history
  - GPS -> GEM -> multiGPS -> ChExMix

- Core idea of GEM is that the underlying DNA sequence of a potential peak is considered during peak-calling

GEM Algorithm

1. Predict broad (200-300bp wide) ChIP-seq peaks with GPS
2. Calculated enriched k-mers at broad peaks
   - ATCGA, TACGA, ACGAT, etc…
3. Cluster k-mers into groups (Motif-finding)
   - AGCATNNAGCA
4. Calculate position of motif relative to broad peaks
   - Where in a broad peak is a motif likely to be
5. Predict precise location of protein-binding as a function of where the motif is in the broad peak
6. Repeat motif discovery (Steps 2–3) from the Phase 5 improved event locations.
7. Repeat until convergence (the data doesn’t change after an iteration)

GEM vs MACS2

- Why choose?
- Run both!
- The underlying statistical models of both algorithms (and really every peak-caller) can and DO change some the results of peak-calling
  - Mostly around the lower end of enrichment but that may matter to you depending on what you're studying
  - For reference: EdgeR and DESeq2 will give related but slightly different results in RNA-seq analysis

Replicates

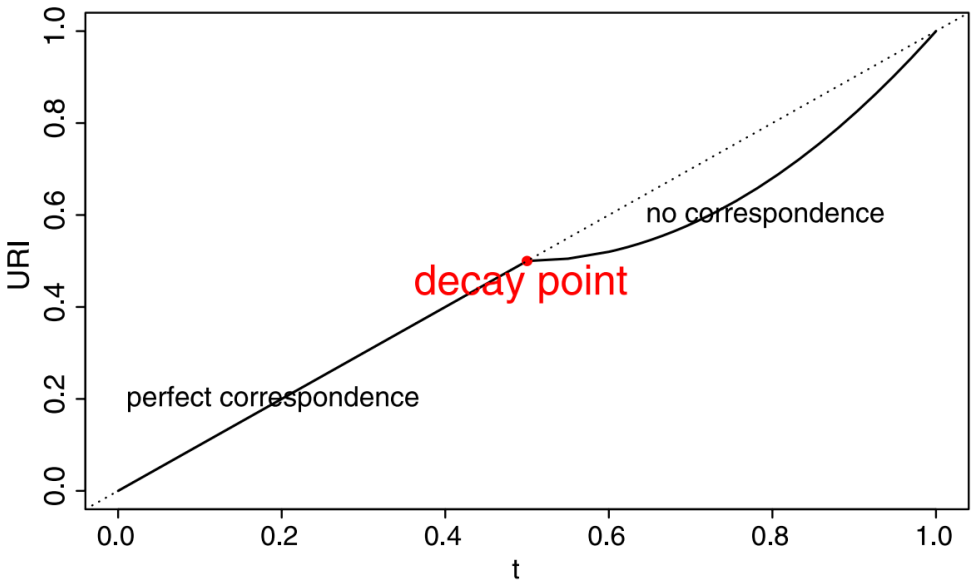
- The level to which biological/technical replicates match each other is also a good metric of quality control

- GEM can handle replicate calling natively but MACS2 does not

Irreproducibility Discovery Rate (IDR)

- For two sets of peaks, rank order each list by occupancy
- True replicates should appear in both lists in approximately the same order
- The point at which peaks no longer match each other (decay point) represents the end of reproducible peaks

- https://projecteuclid.org/euclid.aoas/1318514284



(a) Ψ_n

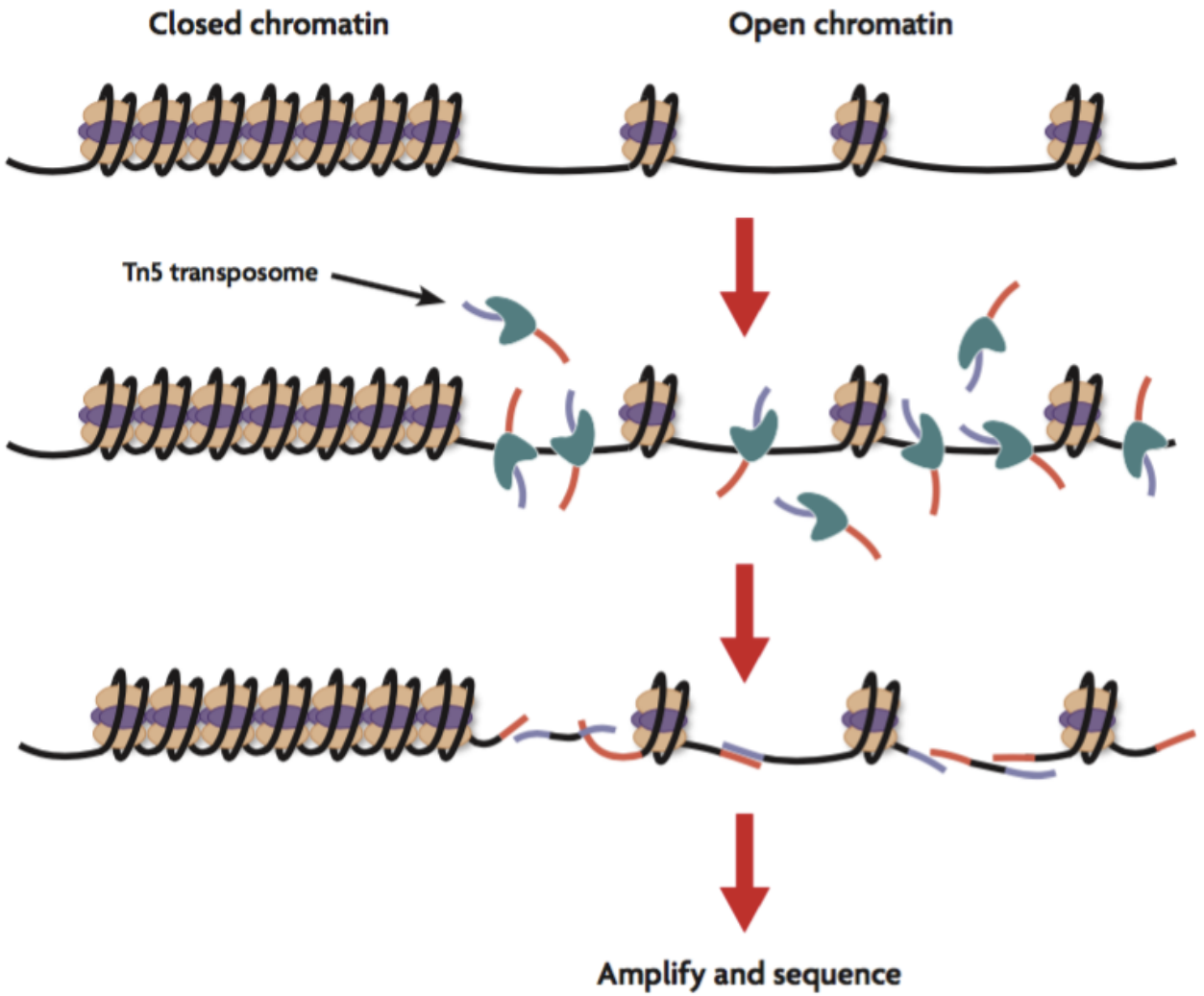
Did my experiment work? (Quality control)

- My factor a sequence-specific transcription factor
  - Is the right motif enriched at peaks?
- My factor bind chromatin (chromatin remodeller)
  - Is it enriched at particular nucleosomes?
- My factor binds in promoters/enhancers/insulators/etc…
  - Does it?

- If I performed replicates, do the replicates match each other?
  - Do the peaks overlap with each other significantly?
  - Do the aligned tags enrich at the same regions?

- General questions:
  - How is my mapping quality?
  - Do my reads map to blacklist more than I expect?
  - What is my PCR duplication rate? (Need UMI and/or paired end sequencing for this)

Assay for Transposase-Accessible Chromatin (ATAC-seq)



https://www.activemotif.com/blog-atac-seq

Peak-calling

- MACS2 remains the default recommended peak-caller for ATAC-seq data as per ENCODE recommendations

- HMMRATAC recently published by creator of MACS2 for ATAC-seq peak calling

MACS2 – modified parameters

- The key difference in MACS2 between ChIP-seq and ATAC-seq is a the assumption of insert size
  - Insert size is assumed to be larger in ATAC-seq since the entire promoter region is expected to be larger than the binding of a given ChIP'ed factor footprint

Replicates

- IDR is again recommended for calculating concordance between replicates

Did my experiment work? (Quality control)

- Are we enriched in expected open chromatin regions?
  - Promoters, enhancers, insulators, etc…
- If I performed replicates, do the replicates match each other?
  - Do the peaks overlap with each other significantly?
  - Do the aligned tags enrich at the same regions?
- General questions:
  - How is my mapping quality?
  - Do my reads map to blacklist more than I expect?
  - What is my PCR duplication rate? (Need UMI and/or paired end sequencing for this)