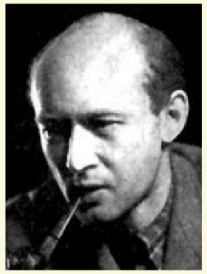
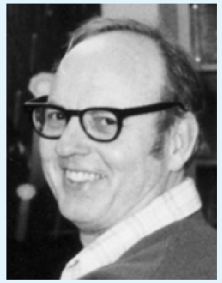
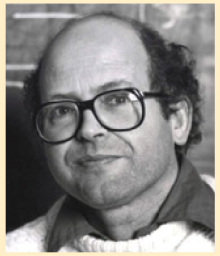
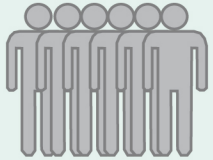
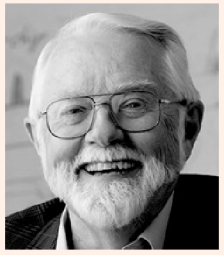
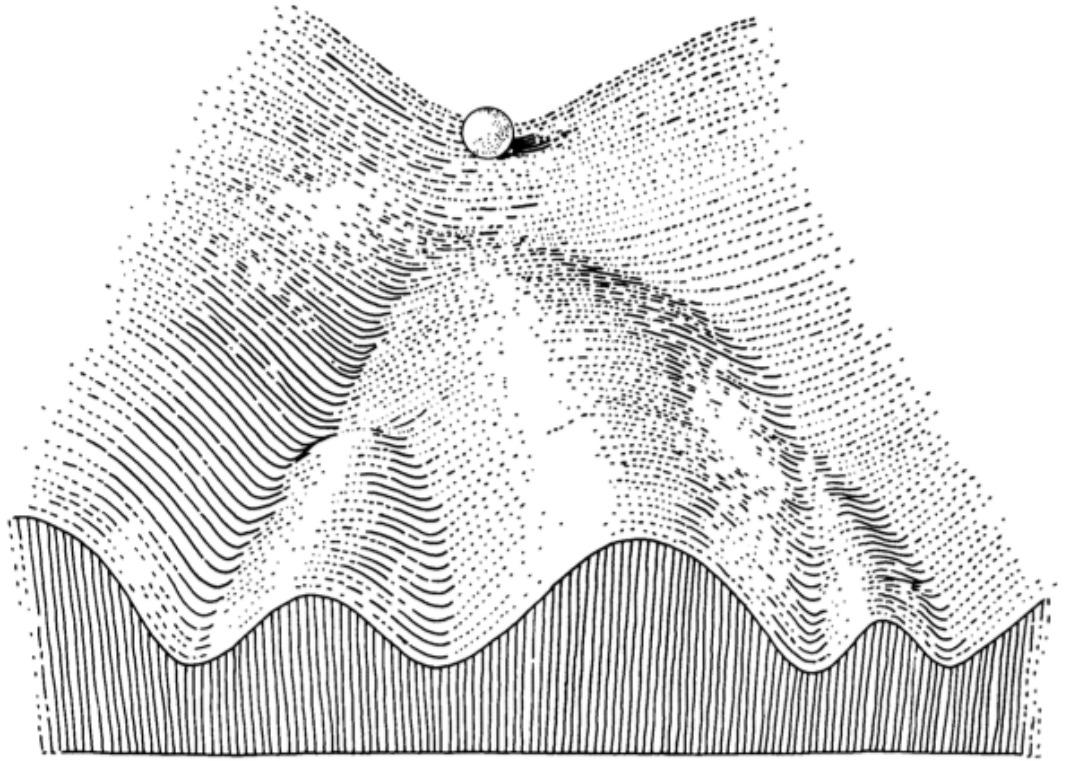
Epigenetics and Chromatin State

What is epigenetics?

		Intended meaning of ‘epigenetic’:	Why this is a problem:
	Conrad Hal Waddington	The **epigenetic landscape** was proposed to resolve how epigenesis could have genetic influences.	All Waddington was trying to do was to stop the embryologists and geneticists from fighting with each other.
	David Nanney	A non-nuclear heritability of long-term cellular memory.	He was trying to explain mating type decisions in isogenic *Tetrahymena,* with evidence that this information was carried in the cytoplasm, not the nucleus. That’s what he meant by epigenetic.
	(John Pugh) Robin Holliday	A mutational mechanism involving changes in DNA methylation, not DNA sequence. Later extended to mean gene regulation.	Changed the definition mid-use from a type of mutation to transcriptional regulation, and equated epigenetic properties with molecular regulators, whereas definitions had previously been based on cell properties.
	Everyone	Back-translated epi- (above, upon) -genetics (DNA sequence) to mean **any** biochemical process regulating the genome.	Such a broad definition encompasses any molecular regulator that can differ from cell to cell, reducing its meaningfulness.
	Arthur Riggs	Attempted to rein in the use of epi- (above) -genetics (DNA sequence) by requiring molecular process to be heritable through cell division	By including both mitotic and meiotic cell division, he generated the **multigenerational** definition of epigenetics

How do we define epigenetics?

- **All cells have the same DNA; how are different cell types determined?**
- For our purposes, epigenetics refers to non-DNA changes to DNA that modify gene expression or regulate the genome.
- These changes are reversible
  - Induced pluripotent stem cells
  - Some evidence that differentiated cells can revert to a stem cell-like state under certain circumstances, for example liver trauma



Images: Autism Daily

Levels of Chromation Organization

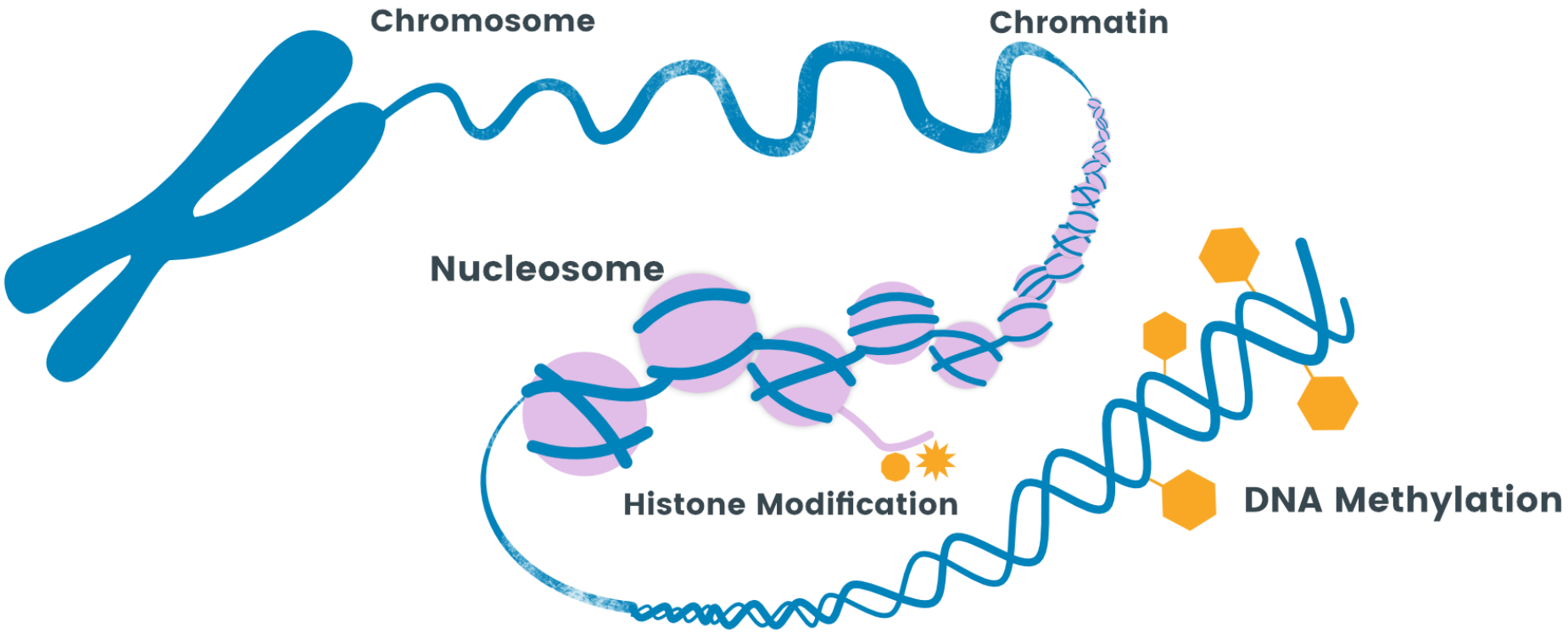


Image: labclinics

DNA Methylation

- Addition of a methyl group to a cytosine
- Majority is on Cs followed by Gs (CpGs) and the individual cytosines are analyzed
- DNA methylation regulates:
  - Transcriptional repression
  - X-chromosome inactivation
  - Imprinting
  - Silencing repetitive elements
- Enriched in genes
  - May be a signal for transcription
  - May be involved in alternative splicing
  - May inhibit alternative
- Mutagenic

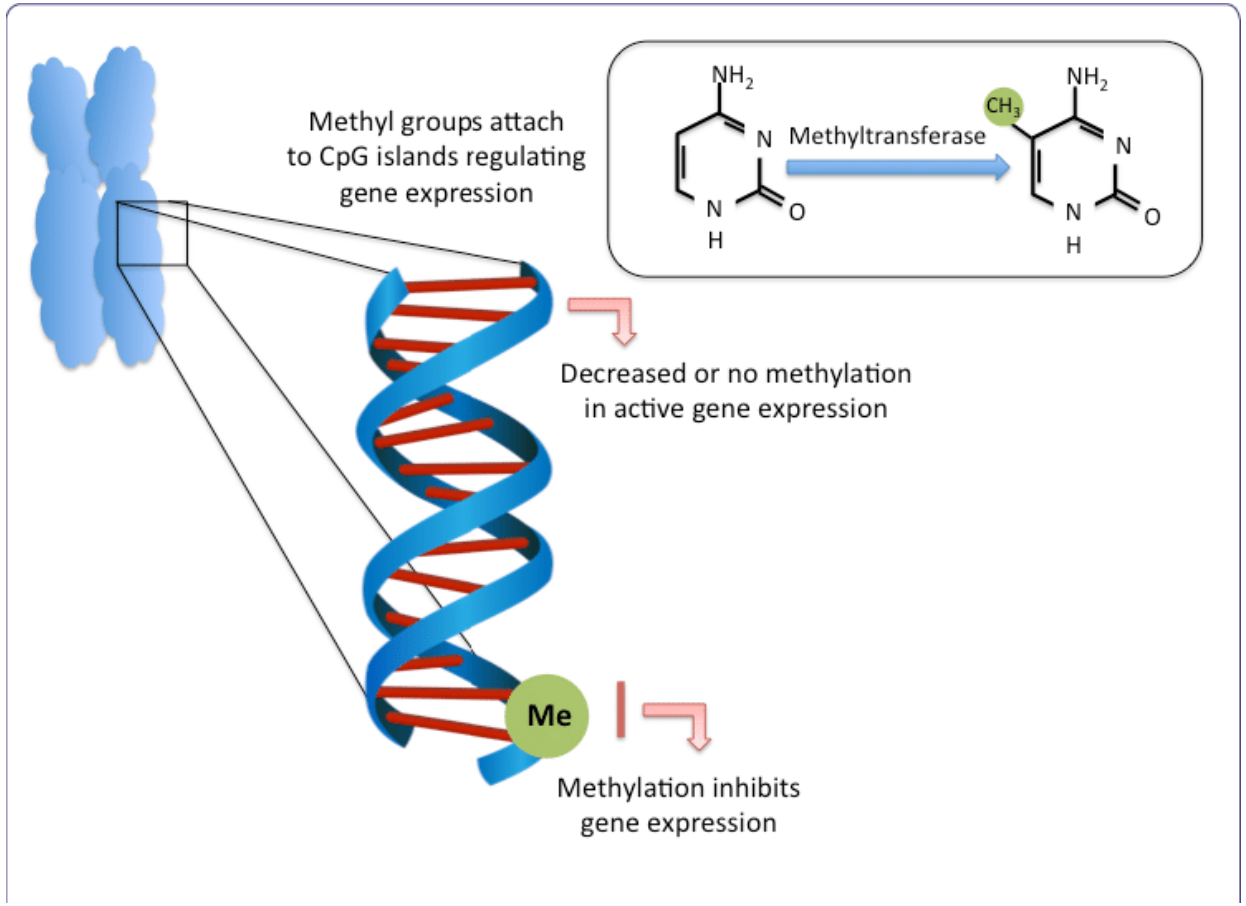


Image: Caroll 2015

Histone Modifications

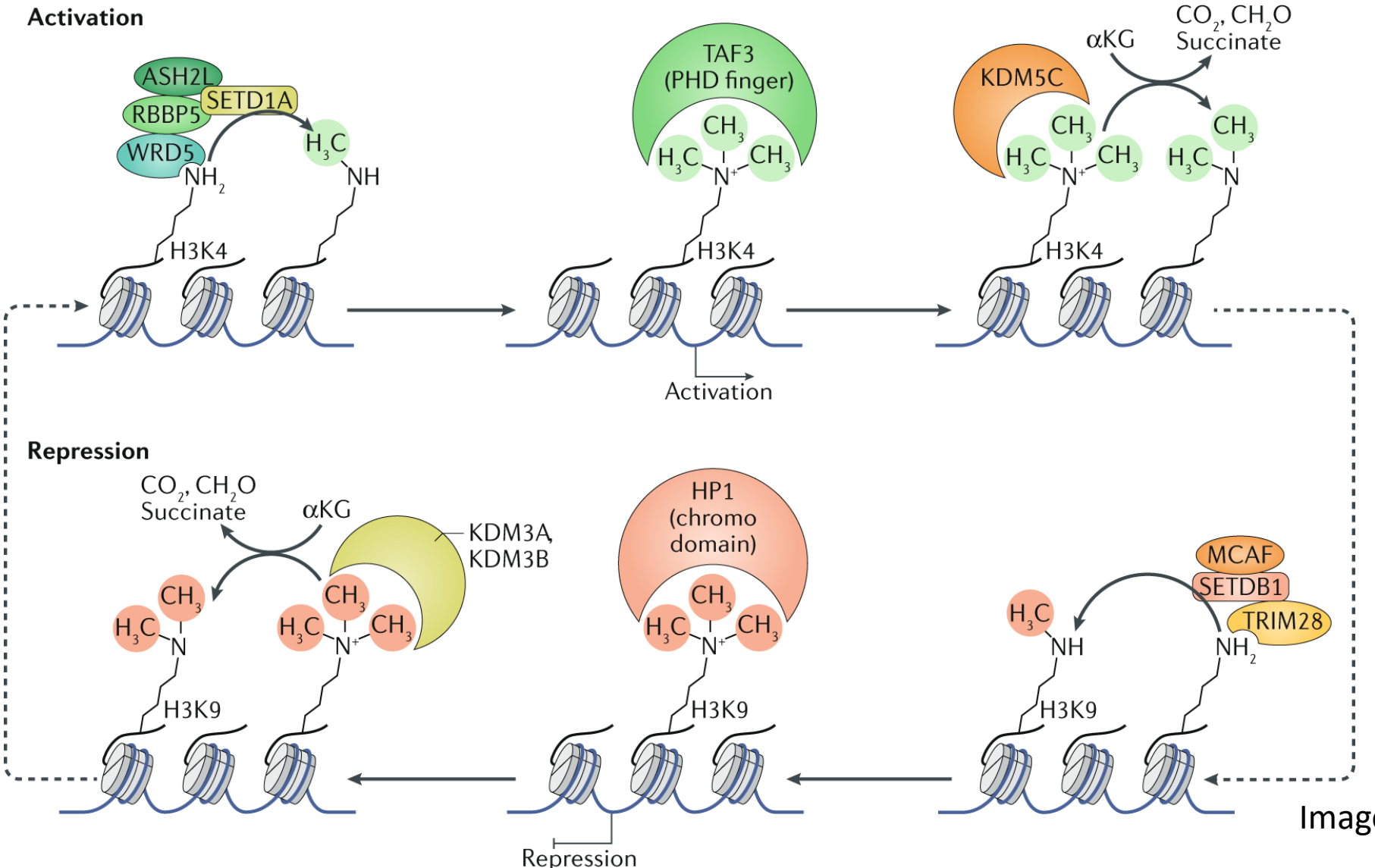


Image: Banner 2011

Sequencing Considerations (Review from Last Week)

- Sequencing Depth
  - Confidence in your measurement increases as sequencing depth increases
  - Need 20-30 million reads per sample for most NGS sequencing applications
  - Cost increases with increased depth
- Read Length
  - The longer the read length the more sequencing you get.
  - Price increases with read length.
  - Depending on the assay, a long read length can mean you'll read into adapters and waste a lot of sequencing.
- Single vs Paired-End Reads
  - Double the information with paired-end reads, but more expensive
  - Need paired-end reads for some applications, like detecting alternative splicing

Bisulfite Sequencing (BS-seq) for DNA Methylation

What is BS-seq?

- Use sodium bisulfite to convert unmethylated cytosines to thymines, then sequence
- Two versions
  - Whole Genome Bisulfite Sequencing
  - Reduced Representation Bisulfite Sequencing
    - Use MspI restriction enzyme to cut at CCGG and select for regions that are likely to have cytosines
    - Sequence a subsample that is representative of the entire genome
    - Much cheaper

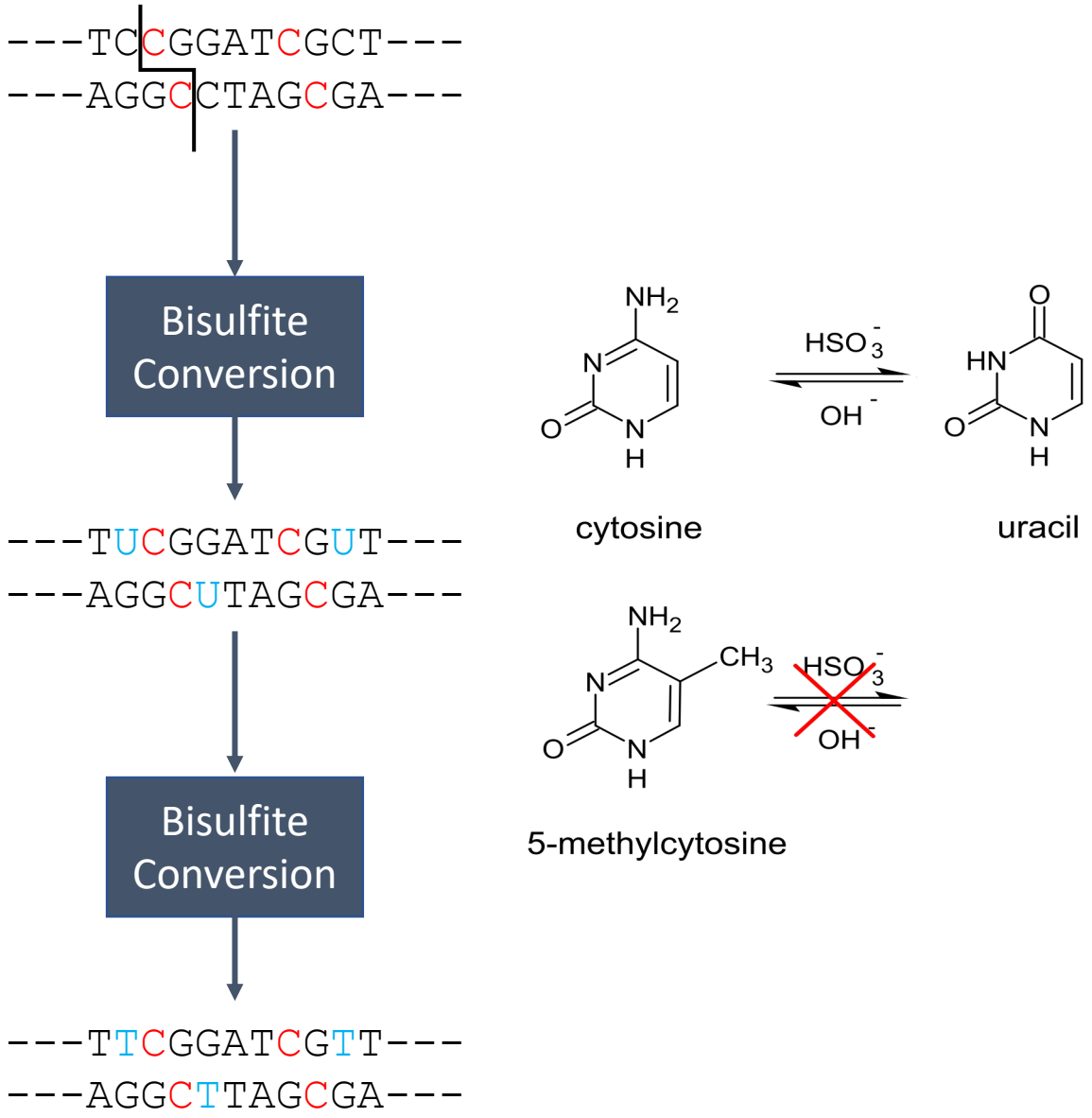


Image: Custom, Wikipedia, Ref: Greenberg 2019

How are BS-seq libraries prepared?

- Special reagents
  - Sodium bisulfite
  - Methylated sequencing adapters
- Which version are you doing?
  - RRBS, all steps
  - WGBS starts at step 3
- RRBS was optimized for a read length of 80 base pairs (bp), so sequence at a shorter read length (if possible).

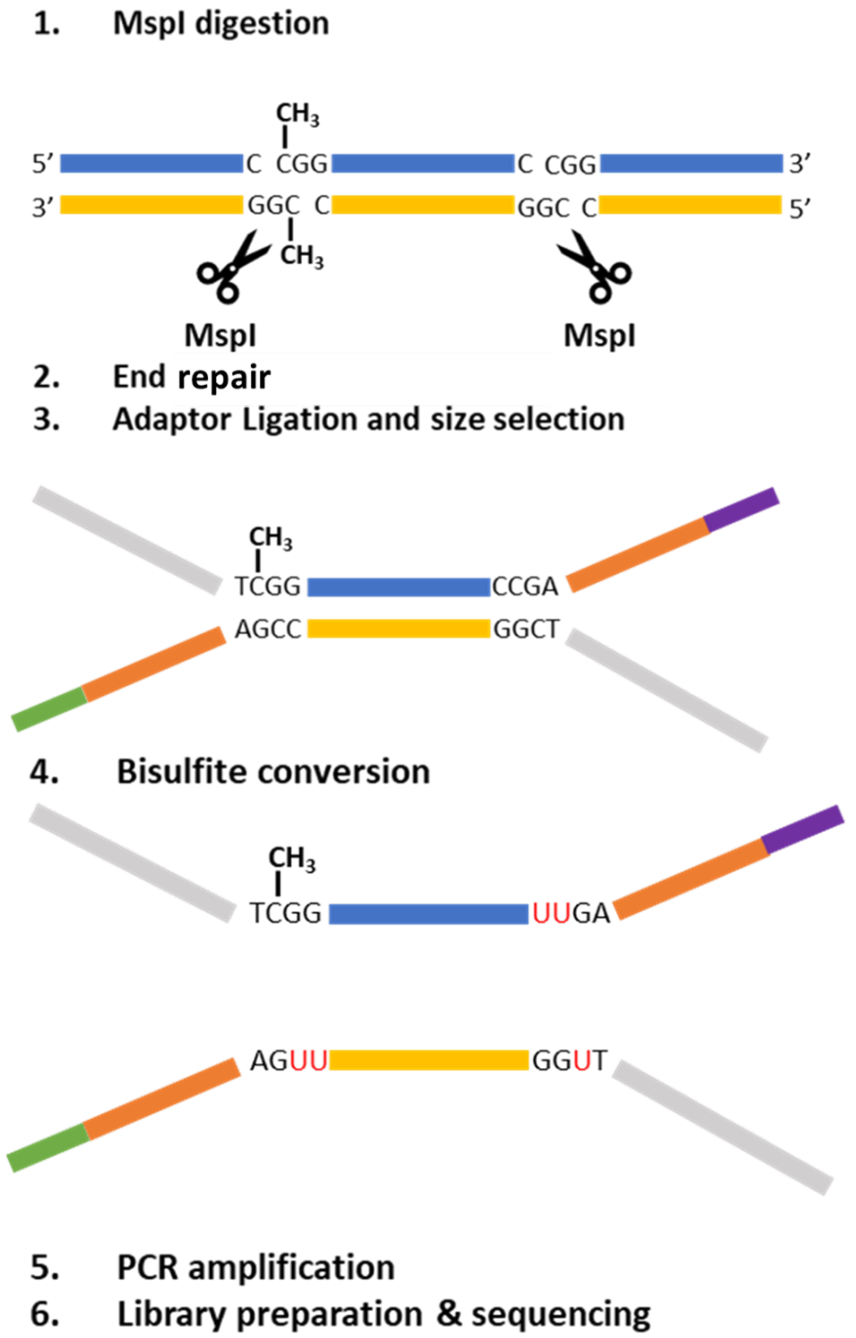


Image: Courtesy of Dr. Jaroslav Jelinek, Ref: Greenberg 2019

Process BS-seq Data

Standard Processing

1. Remove unwanted sequences
2. Align to reference genome
3. Count feature of interest

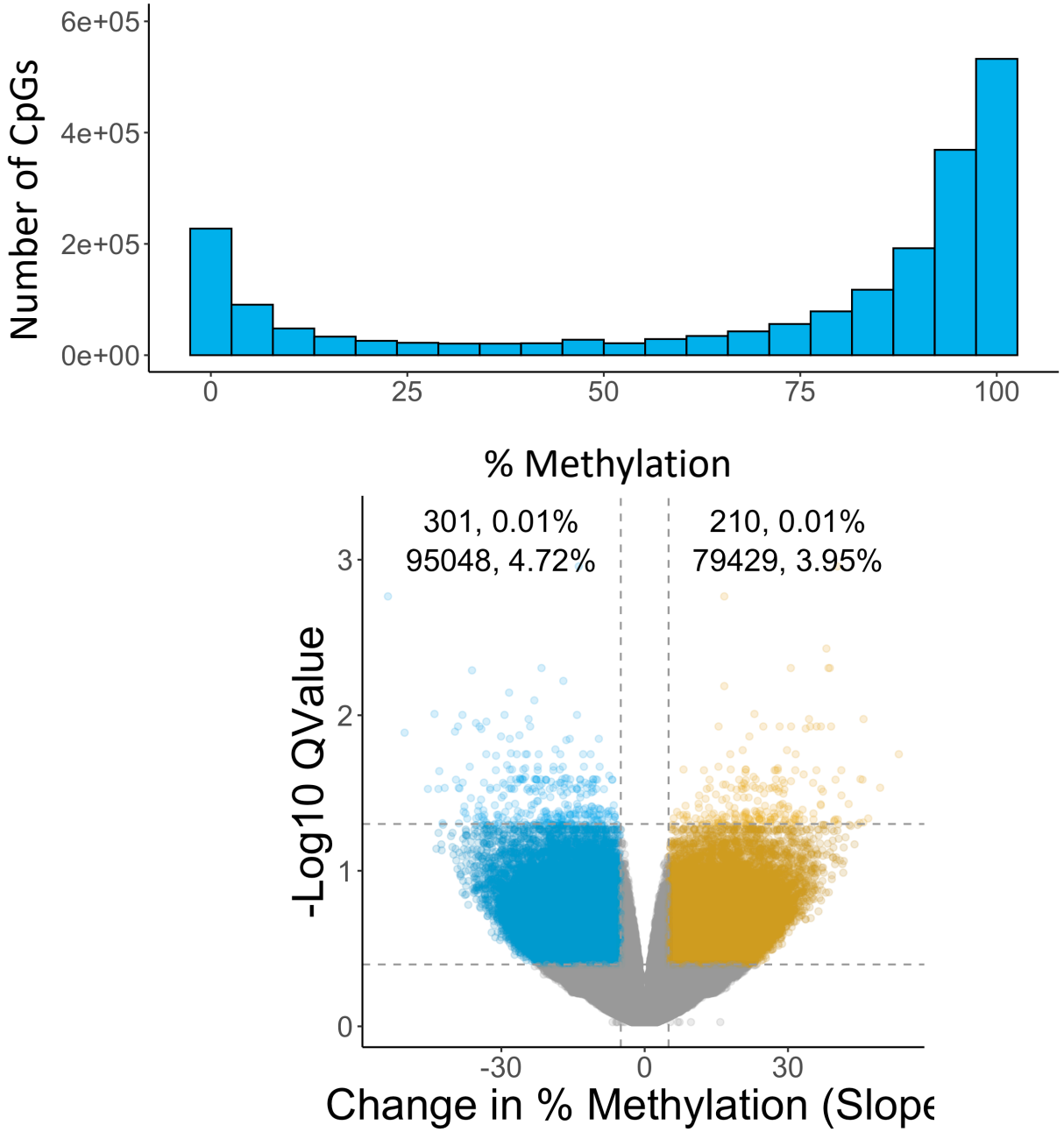
Bisulfite Sequencing

1. Remove unwanted sequences
2. Align to 2 different versions of the reference genome
    1. Normal reference genome
    2. Bisulfite converted reference genome (All Cs changed to Ts)
3. Count number of methylated Cs and number of Ts where Cs should be

Ref: Greenberg 2019

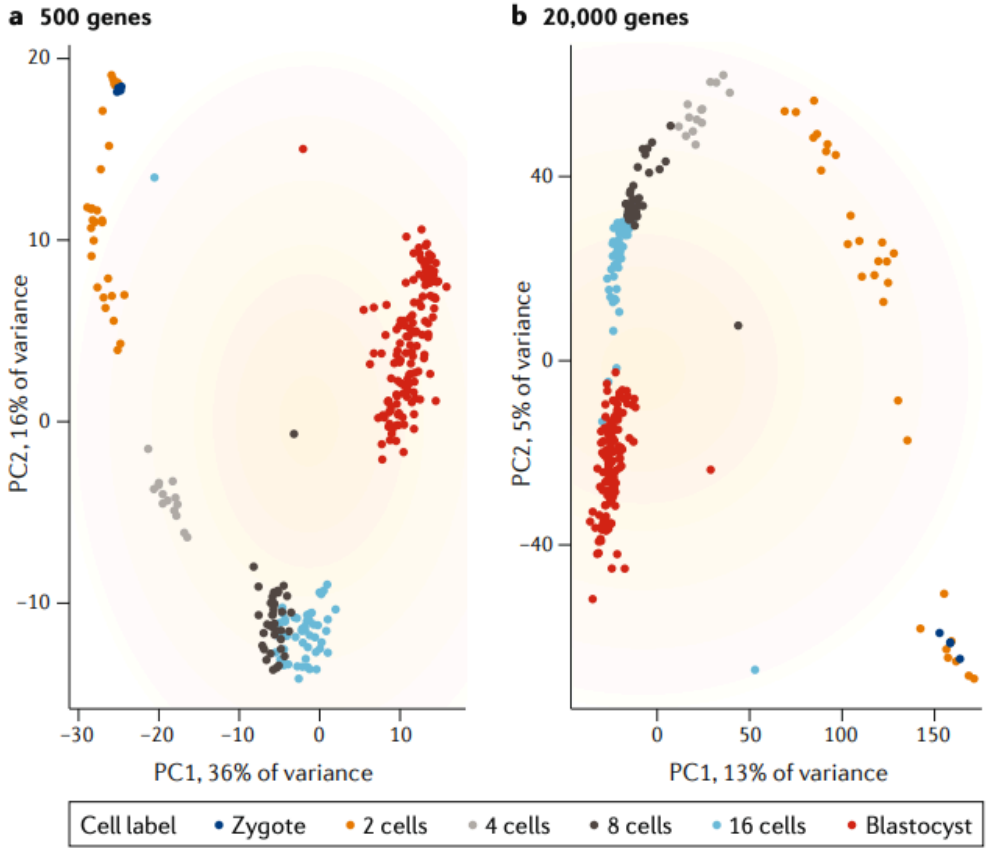
Analyzing BS-seq Data

- Reported as percent methylation, i.e. “This CpG is 80% methylated”
  - Individual Cs are either methylated (1) or unmethylated (0)
  - But because what we detect is an average of all the cells in the population sequenced, it’s reported as a percent.
- Differential Methylation
  - Test association of each CpG with the phenotype of interest
  - Correct for multiple testing



Images: KK (unpublished), Ref: Greenberg 2019

Analyzing single cell BS-seq Data



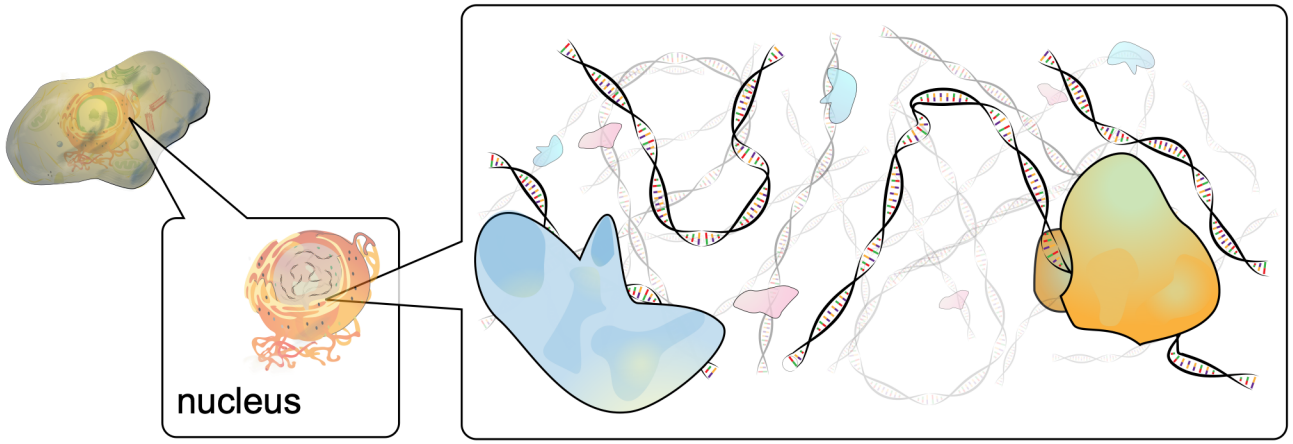
- Library prep is the same, except you need to separate out single cells somehow first
- Cluster cells and identify cell type first
  - Guess on patterns of methylation or gene expression
  - Label with information from a single cell atlas (doesn't exist for methylation yet though)
- Test for differential methylation between clusters rather than or in addition to between conditions

Image: Kiselev 2019, Ref: Karemaker 2018

Chromatin Immunoprecipitation Sequencing (ChIP-seq) for DNA-associated Proteins and Histone Modifications

What is ChIP-seq?

- ChIP-seq = **Ch**romation **I**mmuno**P**recipitation sequencing
- Analyzes features associated with DNA by targeting with an antibody against the feature of interest
  - DNA-associated proteins
  - Histone modifications



Images: Wikipedia, labclinics

How are ChIP-seq libraries prepared?

1. Cross-link proteins to DNA
2. Sonicate to fragment DNA
3. Add biotinylated antibodies against feature of interest
   - Protein
   - Histone mark
4. Pull down antibody
5. Ligate sequencing adapters
6. PCR amplification

You MUST have an IgG control

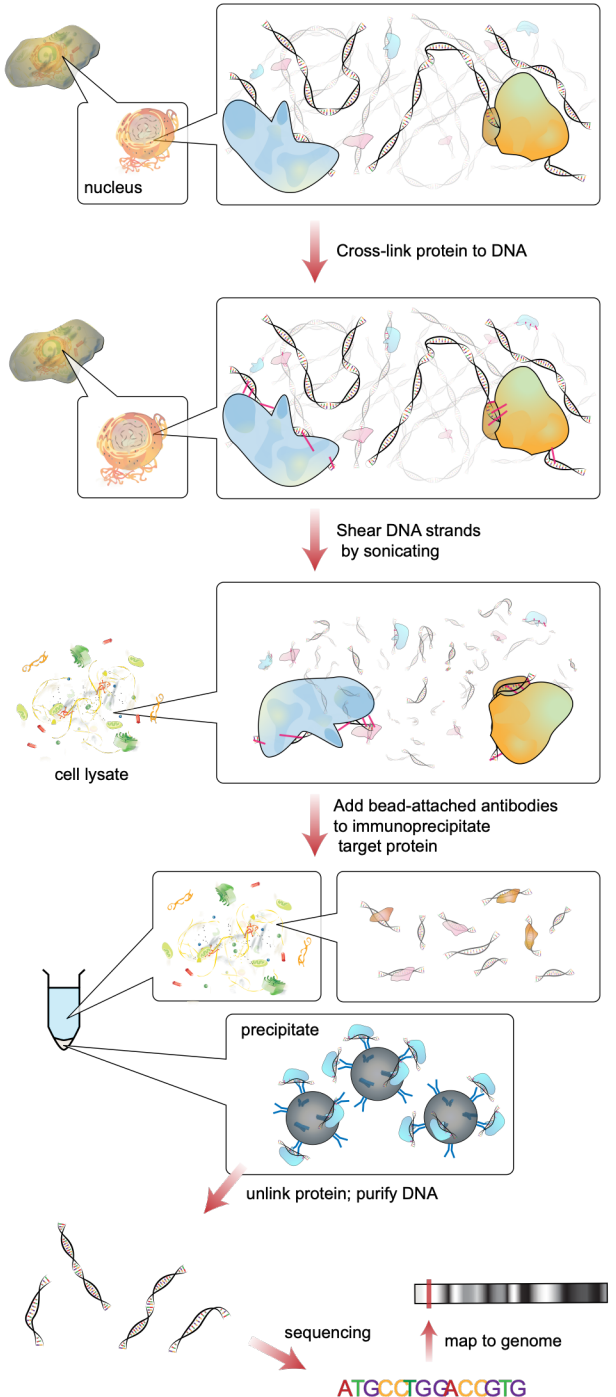


Image: Wikipedia,
Ref: Nakato 2017

Process ChIP-seq Data

Standard Processing

1. Remove unwanted sequences
2. Align to reference genome
3. Count feature of interest

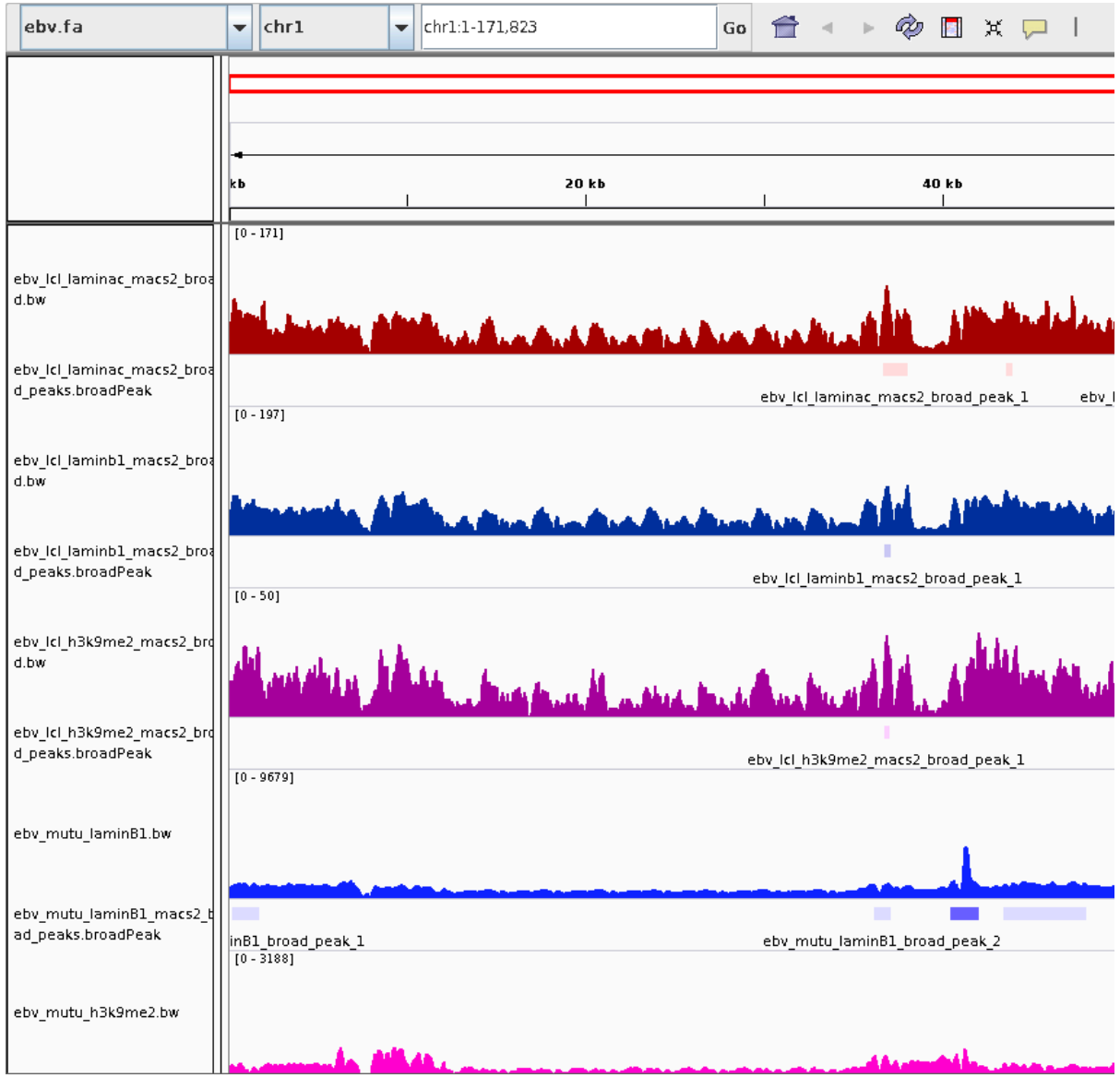
ChIP-seq

1. Remove unwanted sequences
2. Align to the reference genome
3. Call peaks

Ref: Nakato 2017

Calling Peaks

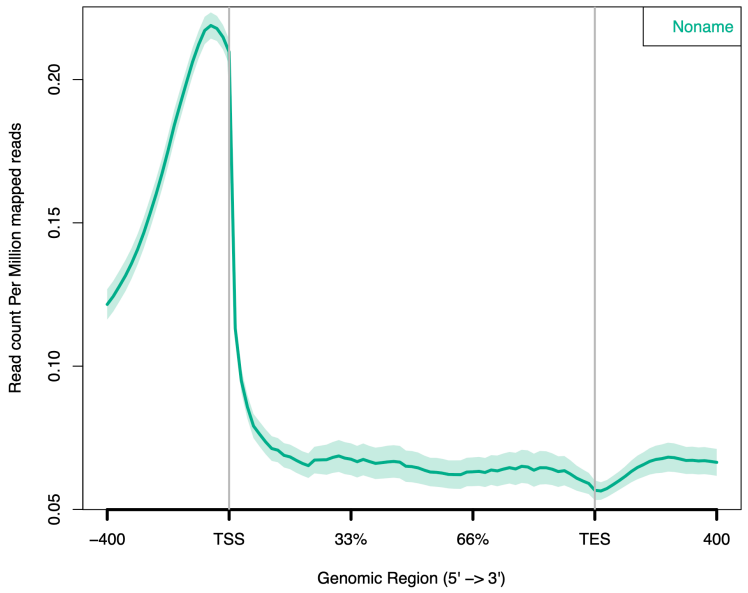
- Efficiency of selecting only the crosslinked DNA isn't great
- Off-target antibody binding
- Call peaks:
  - Finds regions where there is more DNA than background = peak
  - Compares peaks in the sample to control IgG peaks to make sure the observed peak isn't due to off-target IgG binding
- Statistical significance for the peak that must be corrected for multiple testing

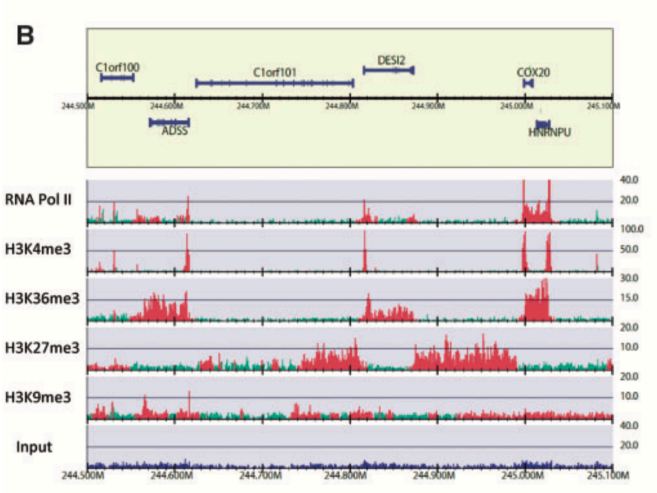


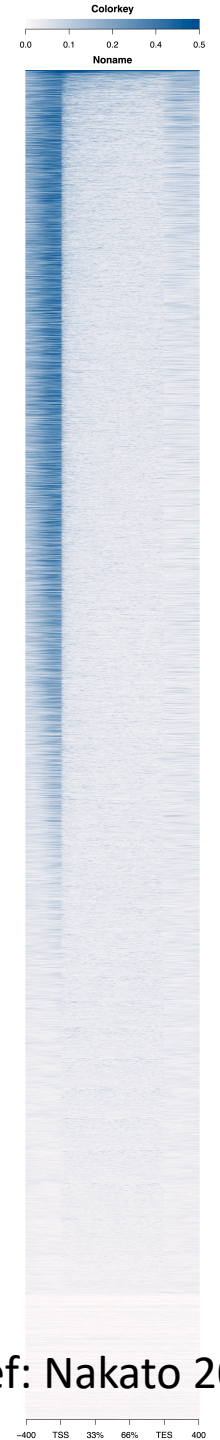
Images: Courtesy of Jozef Madzo (unpublished), Ref: Nakato 2017

Analyzing ChIP-seq Data

- How is the data represented?
  - Traces
  - Heatmaps of the reads
  - Representative example
- Look at peaks
- Look at regions of interest like along the genebody
- Mostly people just say something is present in one condition and absent/reduced/enhanced in another, but can do a permutation test to put a p-value to it







Images: KK (unpublished), Nakato 2017, Ref: Ref: Nakato 2017

ATAC-seq for Chromatin Accessibility

What is ATAC-seq?

- ATAC-seq = **A**ssay for **T**ransposase **A**ccessible **C**hromatin **seq**uencing
- Use a hyperactive mutant Tn5 Transposase to cut regions of open chromatin
- Where is chromatin open?

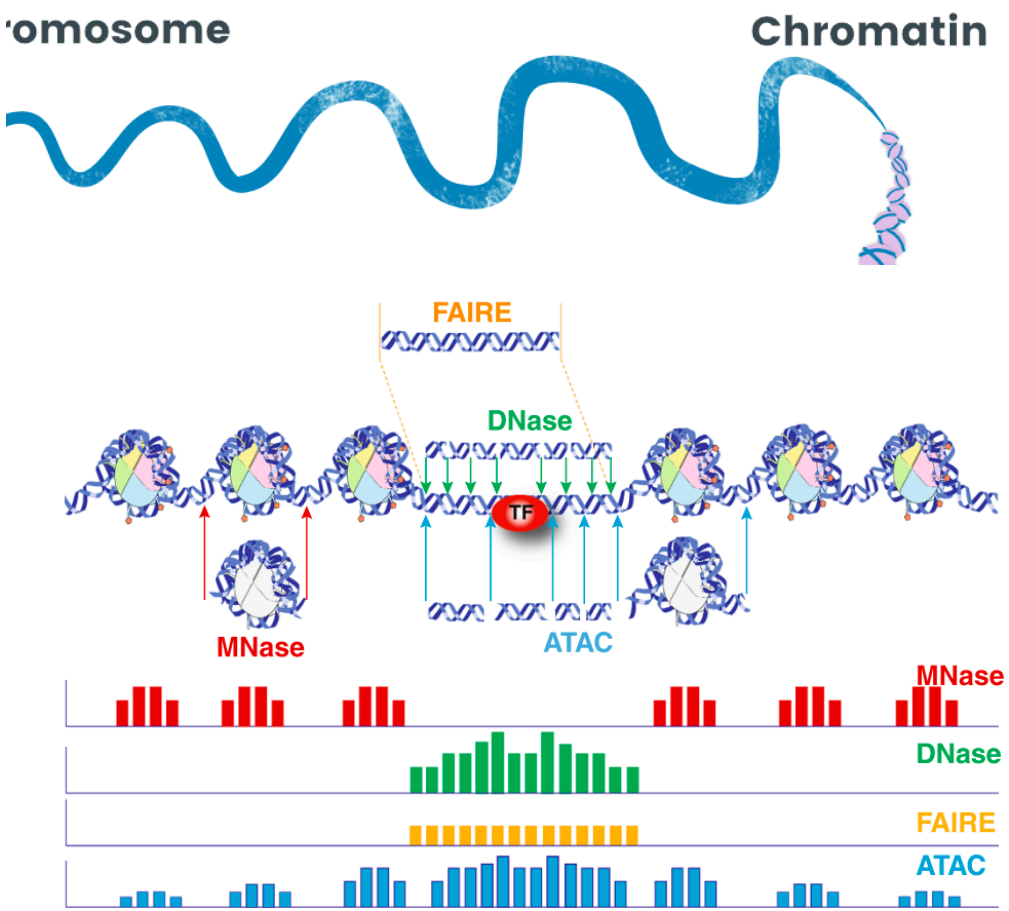


Figure 1 Schematic diagram of current chromatin accessibility assays performed with typical experimental conditions. Representative DNA fragments generated by each assay are shown, with end locations within chromatin defined by colored arrows. Bar diagrams represent data signal obtained from each assay across the entire region. The footprint created by a transcription factor (TF) is shown for ATAC-seq and DNase-seq experiments.

Images: labclinics, Yan 2020, Ref: Yan 2020

How are ATAC-seq libraries prepared?

1. Extract nuclei from cells
2. Treat nuclei with Tn5 transposes which will cut open chromatin and ligate sequencing adapters in a single step (tagmentation)
3. PCR amplification

- Note:
  - Paired-end sequencing preferred because the distance between the pairs shows you where nucleosomes are
  - Unlike ChIP-seq no normalization control (like IgG) required

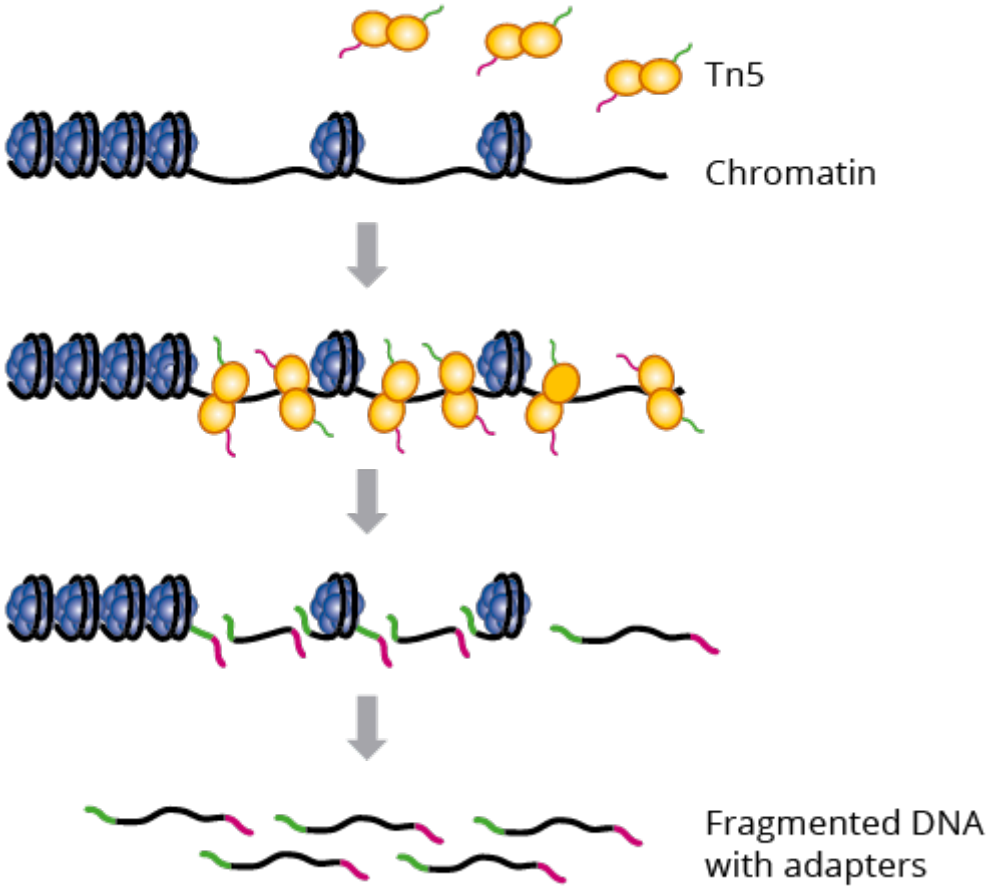


Image: Genewiz, Ref: Yan 2020

Process ATAC-seq Data

Standard Processing

1. Remove unwanted sequences
2. Align to reference genome
3. Count feature of interest

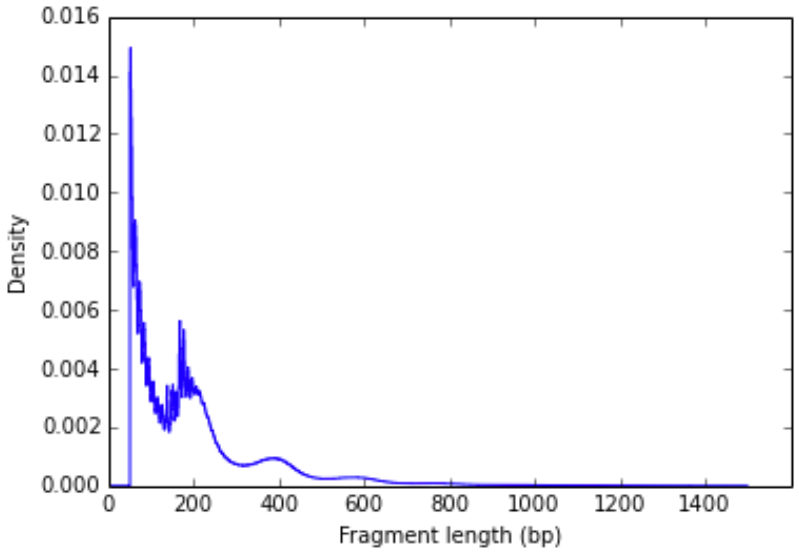
ATAC-seq

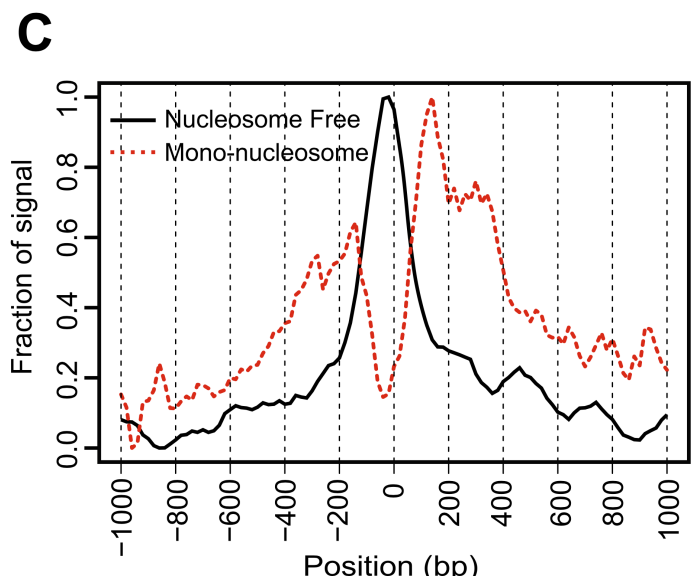
1. Remove unwanted sequences
2. Align to the reference genome
3. Remove duplicates and mitochondrial reads; extremely high proportion of both in ATAC-seq compared to other NGS assays.
4. Call peaks

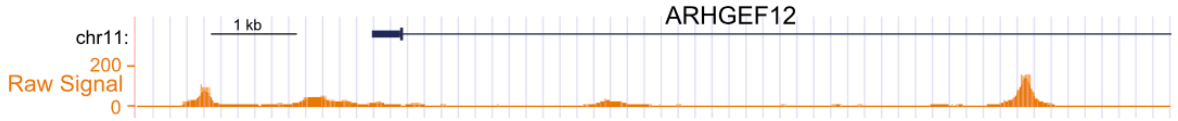
Ref: Yan 2020

Analyzing ATAC-seq Data

- Get clear, crisp peaks at mainly promoters and enhancers.
- Does the location change or the signal change with the experimental condition?
- Visualize
  - Example pileup
  - Traces
  - Heatmaps



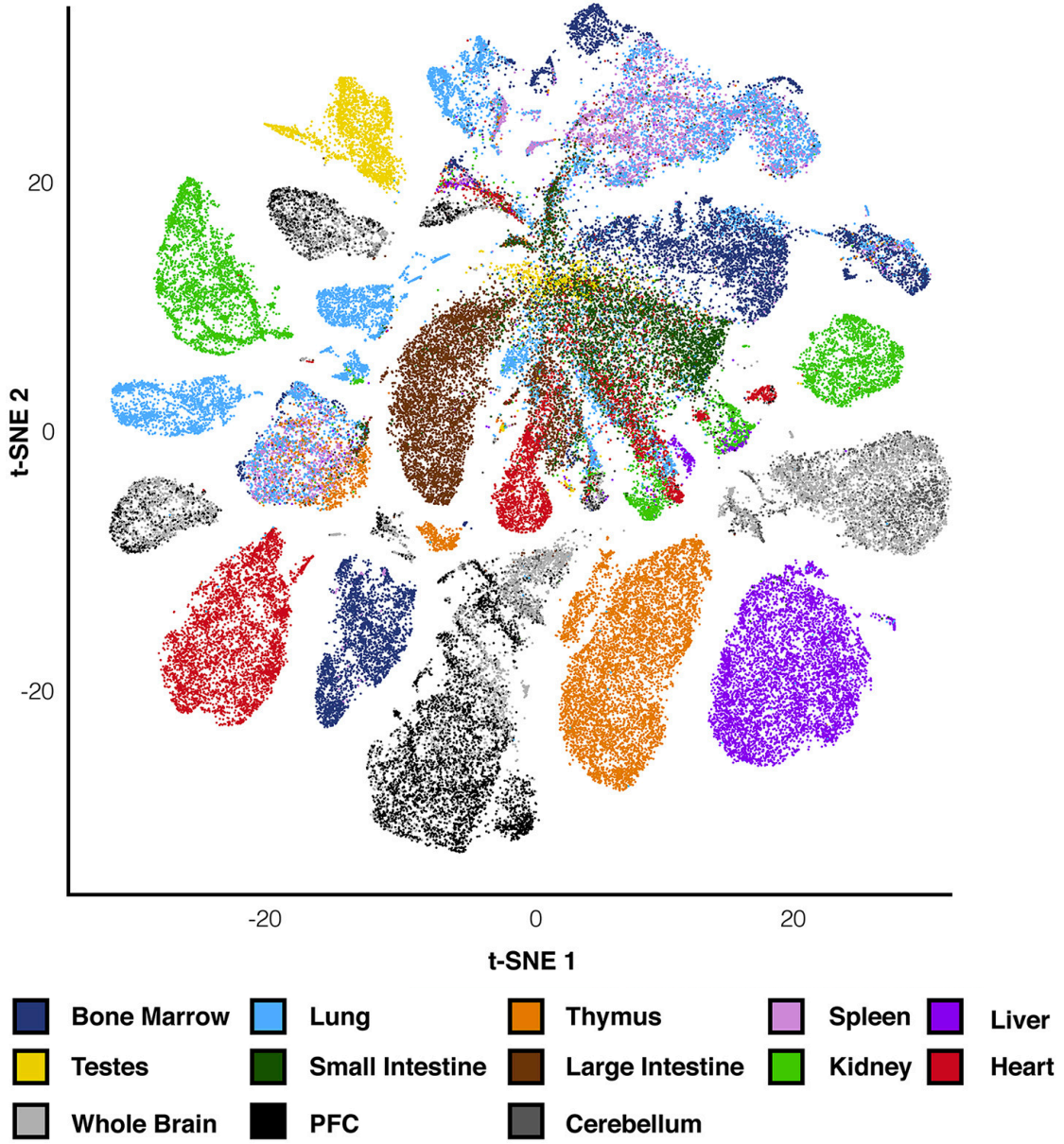




Images: Wikipedia, Yan 2020, Ref: Yan 2020

Analyzing single cell ATAC-seq Data

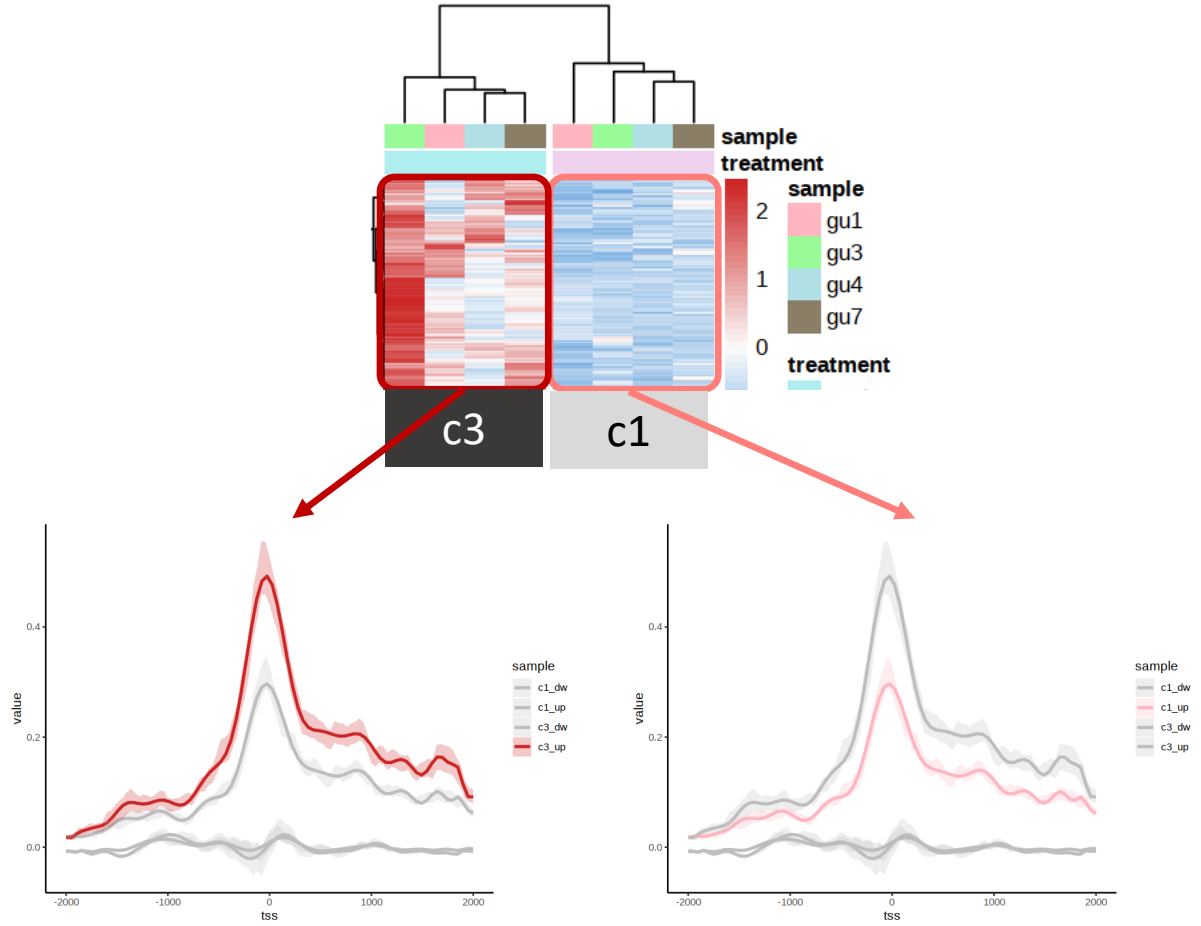
- Additional step before library prep to analyze and barcode single nuclei
- Extra analysis steps
    1. Computationally identify and separate single cells
    2. Cluster and assign cell identity
- Can be easier to identify cell type with scATAC-seq than scRNA-seq because chromatin arrangement is more stable
- scATAC-seq works better for frozen samples than scRNA-seq



Images: Cusanovich 2018, Ref: Cusanovich 2018, Yan 2020

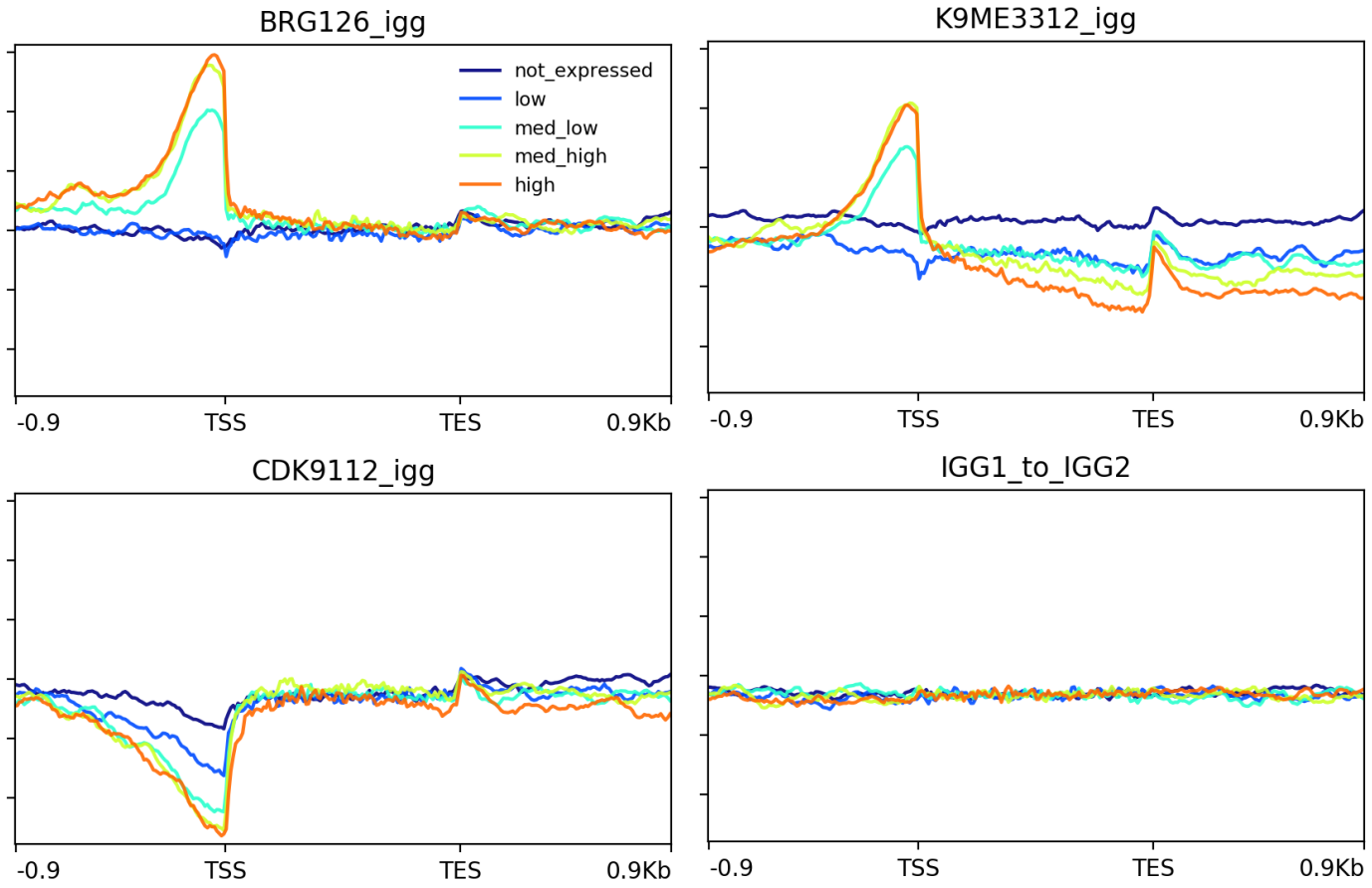
Combing NGS Data by Overlapping Features

- How open is the chromatin at expressed genes?
- What is the methylation status at the promoter of expressed genes?
- Does my histone modification correlate with expressed or unexpressed genes? Is chromatin open where this modification is found?



Combing NGS Data by Overlapping Features

- How open is the chromatin at expressed genes?
- What is the methylation status at the promoter of expressed genes?
- Does my histone modification correlate with expressed or unexpressed genes? Is chromatin open where this modification is found?



References

1. Cusanovich, D. A. et al. A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. Cell 174, 1309–1324.e18 (2018). https://doi.org/10.1016/j.cell.2018.06.052
2. Greenberg, M. V. C. & Bourc'his, D. The diverse roles of DNA methylation in mammalian development and disease. Nature Reviews Molecular Cell Biology 20, 590–607 (2019). https://doi.org/10.1038/s41580-019-0159-6
3. Karemaker, I. D. & Vermeulen, M. Single-Cell DNA Methylation Profiling: Technologies and Biological Applications. Trends in Biotechnology 36, 952–965 (2018). https://doi.org/10.1016/j.tibtech.2018.04.002
4. Nakato, R. & Shirahige, K. Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. Briefings in Bioinformatics bbw023 (2016). https://doi.org/ 10.1093/bib/bbw023
5. Yan, F., Powell, D. R., Curtis, D. J. & Wong, N. C. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. Genome Biology 21, (2020). https://doi.org/10.1186/s13059-020-1929-3

How to can I get started in bioinformatics on my own? (A collection of free online resources)

1. Learn a scripting language
   - R https://r4ds.had.co.nz/
   - Python https://jakevdp.github.io/PythonDataScienceHandbook/
2. Take a free online genomics, biostatistics, bioinformatics classes
   - Biomedical Data Science http://genomicsclass.github.io/book/
   - Modern Statistics for Modern Biology http://web.stanford.edu/class/bios221/book/index.html
   - Bioinformatics Data Skills https://vincebuffalo.com/book/ (book not free sorry!)
3. Go to a MeetUp
   - RLadies Philly https://www.meetup.com/rladies-philly/
   - Philadelphia Python Users Group https://vincebuffalo.com/book/