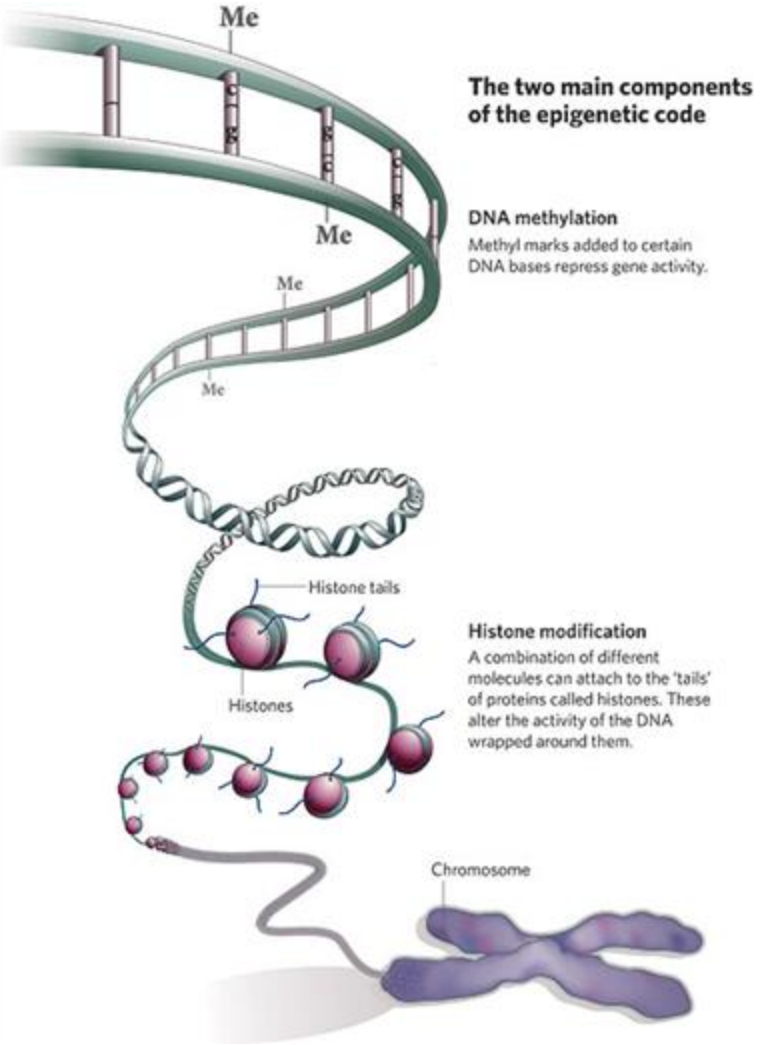
Epigenomics: ChIP-seq and ATAC-seq

Chris Miller, Ph.D.
Washington University in St. Louis

Some slides adapted from:
https://github.com/genome/bfx-workshop
https://github.com/quinlan-lab/applied-computational-genomics

Epigenomics

- Alterations of DNA state or accessibility
- Wrapped around histones
- Bound by transcription factors
- etc



105+ *-seq assays

from Lior Pachter's blog

Nucleo-Seq: Anton Valouev et al., "Determinants of Nucleosome Organization in Primary Human Cells," *Nature* 474, no. 7352 (June 23, 2011): 516–520, doi:10.1038/nature10002.

DNAse-Seq: Gregory E. Crawford et al., "Genome-wide Mapping of DNase Hypersensitive Sites Using Massively Parallel Signature Sequencing (MPSS)," *Genome Research* 16, no. 1 (January 1, 2006): 123–131, doi:10.1101/gr.4074106.

DNAseI-Seq: Jay R. Hesselberth et al., "Global Mapping of protein-DNA Interactions in Vivo by Digital Genomic Footprinting," *Nature Methods* 6, no. 4 (April 2009): 283–289, doi:10.1038/nmeth.1313.

Sono-Seq: Raymond K. Auerbach et al., "Mapping Accessible Chromatin Regions Using Sono-Seq," *Proceedings of the National Academy of Sciences* 106, no. 35 (September 1, 2009): 14926–14931, doi:10.1073/pnas.0905443106.

Hi-C-Seq: Erez Lieberman-Aiden et al., "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome," *Science* 326, no. 5950 (October 9, 2009): 289–293, doi:10.1126/science.1181369.

ChIA-PET-Seq: Melissa J. Fullwood et al., "An Oestrogen-receptor-α-bound Human Chromatin Interactome," *Nature* 462, no. 7269 (November 5, 2009): 58–64, doi:10.1038/nature08497.

FAIRE-Seq: Hironori Waki et al., "Global Mapping of Cell Type–Specific Open Chromatin by FAIRE-seq Reveals the Regulatory Role of the NFI Family in Adipocyte Differentiation," *PLoS Genet* 7, no. 10 (October 20, 2011): e1002311,

NOMe-Seq: Theresa K. Kelly et al., "Genome-wide Mapping of Nucleosome Positioning and DNA Methylation Within Individual DNA Molecules," *Genome Research* 22, no. 12 (December 1, 2012): 2497–2506, doi:10.1101/gr.143008.112.

ATAC-Seq: Jason D. Buenrostro et al., "Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-binding Proteins and Nucleosome Position," *Nature Methods* advance online publication (October 6, 2013), doi:10.1038/nmeth.2688.

Genome variation

RAD-Seq: Nathan A. Baird et al., "Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers," *PLoS ONE* 3, no. 10 (October 13, 2008): e3376, doi:10.1371/journal.pone.0003376.

Freq-Seq: Lon M. Chubiz et al., "FREQ-Seq: A Rapid, Cost-Effective, Sequencing-Based Method to Determine Allele Frequencies Directly from Mixed Populations," *PLoS ONE* 7, no. 10 (October 31, 2012): e47959, doi:10.1371/journal.pone.0047959.

CNV-Seq: Chao Xie and Martti T. Tammi, "CNV-seq, a New Method to Detect Copy Number Variation Using High-throughput Sequencing," *BMC Bioinformatics* 10, no. 1 (March 6, 2009): 80, doi:10.1186/1471-2105-10-80.

Novel-Seq: Iman Hajirasouliha et al., "Detection and Characterization of Novel Sequence Insertions Using Paired-end Next-generation Sequencing," *Bioinformatics* 26, no. 10 (May 15, 2010): 1277–1283, doi:10.1093/bioinformatics/btq152.

TAm-Seq: Tim Forshew et al., "Noninvasive Identification and Monitoring of Cancer Mutations by Targeted Deep Sequencing of Plasma DNA," *Science Translational Medicine* 4, no. 136 (May 30, 2012): 136ra68, doi:10.1126/scitranslmed.3003726.

DNA replication

Repli-Seq: R. Scott Hansen et al., "Sequencing Newly Replicated DNA Reveals Widespread Plasticity in Human Replication Timing," *Proceedings of the National Academy of Sciences* 107, no. 1 (January 5, 2010): 139–144, doi:10.1073/pnas.0912402107

ARS-Seq: Ivan Liachko et al., "High-resolution Mapping, Characterization, and Optimization of Autonomously Replicating Sequences in Yeast," *Genome Research* 23, no. 4 (April 1, 2013): 698–704, doi:10.1101/gr.144659.112.

Sort-Seq: Carolin A. Müller et al., "The Dynamics of Genome Replication Using Deep Sequencing," *Nucleic Acids Research* (October 1, 2013): gkt878, doi:10.1093/nar/gkt878.

Transcription

RNA-Seq: Ali Mortazavi et al., "Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq," *Nature Methods* 5, no. 7 (July 2008): 621–628, doi:10.1038/nmeth.1226.

GRO-Seq: Leighton J. Core, Joshua J. Waterfall, and John T. Lis, "Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters," *Science* 322, no. 5909 (December 19, 2008): 1845–1848, doi:10.1126/science.1162228.

Quartz-Seq: Yohei Sasagawa et al., "Quartz-Seq: a Highly Reproducible and Sensitive Single-cell RNA-Seq Reveals Non-genetic Gene Expression Heterogeneity," *Genome Biology* 14, no. 4 (April 17, 2013): R31, doi:10.1186/gb-2013-14-4-r31.

CAGE-Seq: Hazuki Takahashi et al., "5' End-centered Expression Profiling Using Cap-analysis Gene Expression and Next-generation Sequencing," *Nature Protocols* 7, no. 3 (March 2012): 542–561, doi:10.1038/nprot.2012.005.

Nascent-Seq: Joseph Rodriguez, Jerome S. Menet, and Michael Rosbash, "Nascent-Seq Indicates Widespread Cotranscriptional RNA Editing in Drosophila," *Molecular Cell* 47, no. 1 (July 13, 2012): 27–37, doi:10.1016/j.molcel.2012.05.002.

Precapture RNA-Seq: Tim R. Mercer et al., "Targeted RNA Sequencing Reveals the Deep Complexity of the Human Transcriptome," *Nature Biotechnology* 30, no. 1 (January 2012): 99–104, doi:10.1038/nbt.2024.

Cel-Seq: Tamar Hashimshony et al., "CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification," *Cell Reports* 2, no. 3 (September 27, 2012): 666–673, doi:10.1016/j.celrep.2012.08.003.

3P-Seq: Calvin H. Jan et al., "Formation, Regulation and Evolution of Caenorhabditis Elegans 3'UTRs," *Nature* 469, no. 7328 (January 6, 2011): 97–101, doi:10.1038/nature09616.

NET-Seq: L. Stirling Churchman and Jonathan S. Weissman, "Nascent Transcript Sequencing Visualizes Transcription at Nucleotide Resolution," *Nature* 469, no. 7330 (January 20, 2011): 368–373, doi:10.1038/nature09652.

SS3-Seq: Oh Kyu Yoon and Rachel B. Brem, "Noncanonical Transcript Forms in Yeast and Their Regulation During Environmental Stress," *RNA* 16, no. 6 (June 1, 2010): 1256–1267, doi:10.1261/rna.2038810.

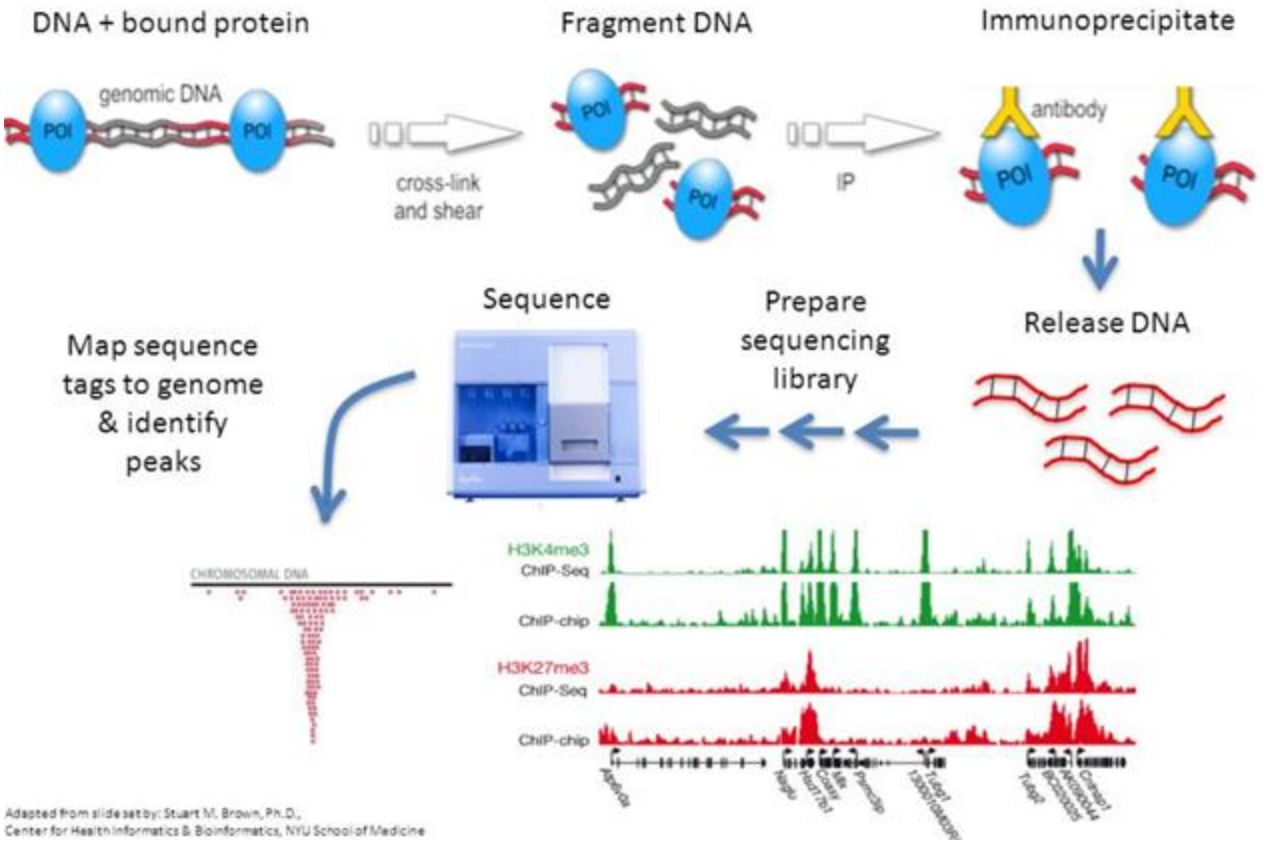
FRT-Seq: Lira Mamanova et al., "FRT-seq: Amplification-free, Strand-specific Transcriptome Sequencing," *Nature Methods* 7, no. 2 (February 2010): 130–132, doi:10.1038/nmeth.1417.

3-Seq: Andrew H. Beck et al., "3'-End Sequencing for Expression Quantification (3SEQ) from Archival Tumor Samples," *PLoS ONE* 5, no. 1 (January 19, 2010): e8768, doi:10.1371/journal.pone.0008768.

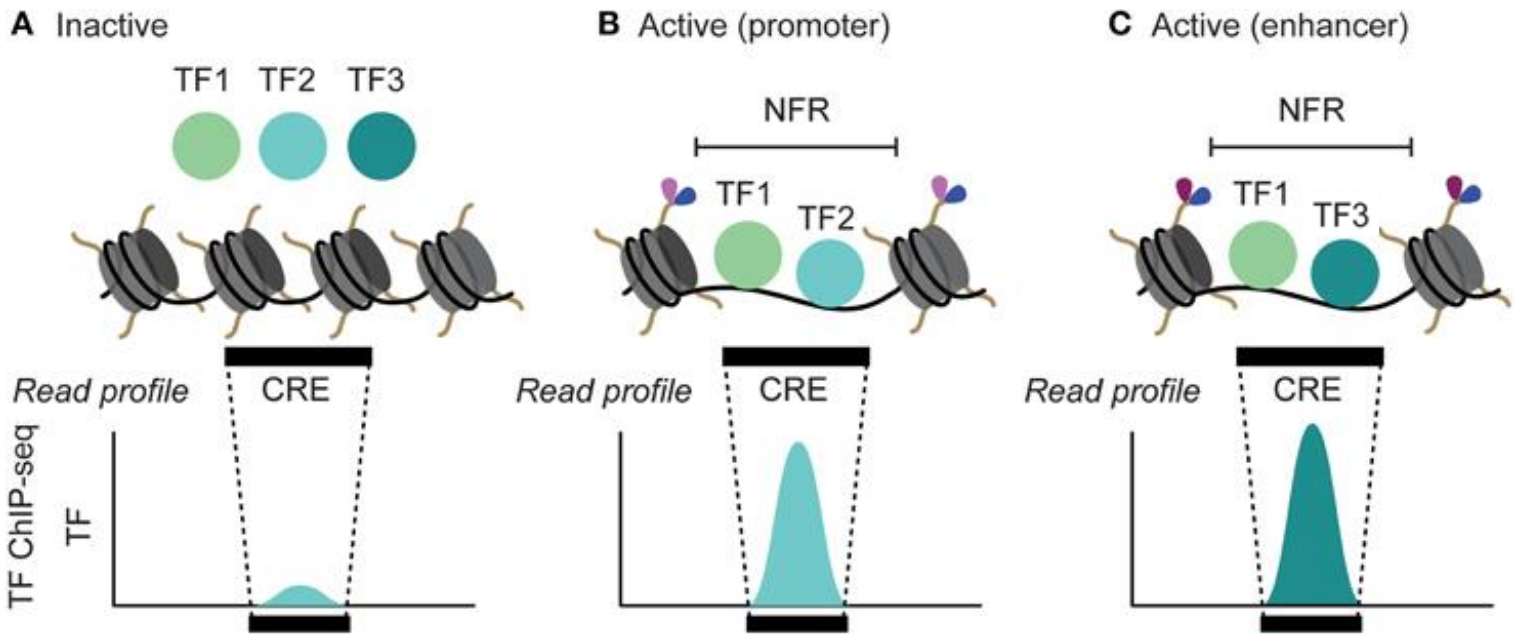
PRO-Seq: Hojoong Kwak et al., "Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing," *Science* 339, no. 6122 (February 22, 2013): 950–953, doi:10.1126/science.1229386.

Bru-Seq: Artur Veloso et al., "Genome-Wide Transcriptional Effects of the Anti-Cancer Agent Camptothecin," *PLoS ONE* 8, no. 10 (October 23, 2013): e78190, doi:10.1371/journal.pone.0078190.

ChIP-seq


DNA + bound protein
genomic DNA
POI
POI
cross-link
and shear
Fragment DNA
POI
POI
IP
Immunoprecipitate
antibody
POI
POI
Release DNA
Prepare
sequencing
library
Sequence
Map sequence
tags to genome
& identify
peaks
CHROMOSOMAL DNA
H3K4me3
ChIP-Seq
ChIP-chip
H3K27me3
ChIP-Seq
ChIP-chip
Adapted from slide set by: Stuart M. Brown, Ph.D.,
Center for Health Informatics & Bioinformatics, NYU School of Medicine

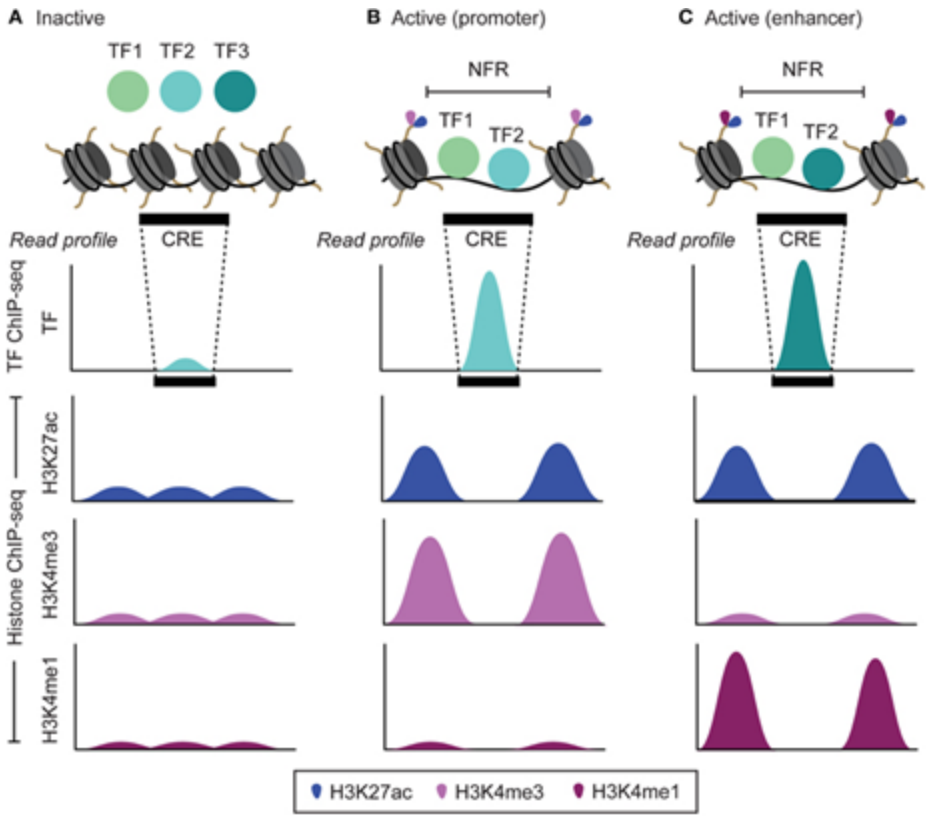
Mapping transcription-factor binding locations



NFR = nucleosome free region
CRE = Cis regulatory element

http://journal.frontiersin.org/article/10.3389/fgene.2015.00188/full

Mapping histone modifications



H3K4me3:

H3 = name of histone
K4 = 4th lysine residue
me3 = tri-methylation

NFR = nucleosome free region
CRE = Cis regulatory element

http://journal.frontiersin.org/article/10.3389/fgene.2015.00188/full

CUT&RUN

Potentially less off-target/background signal

Very similar, from a data perspective

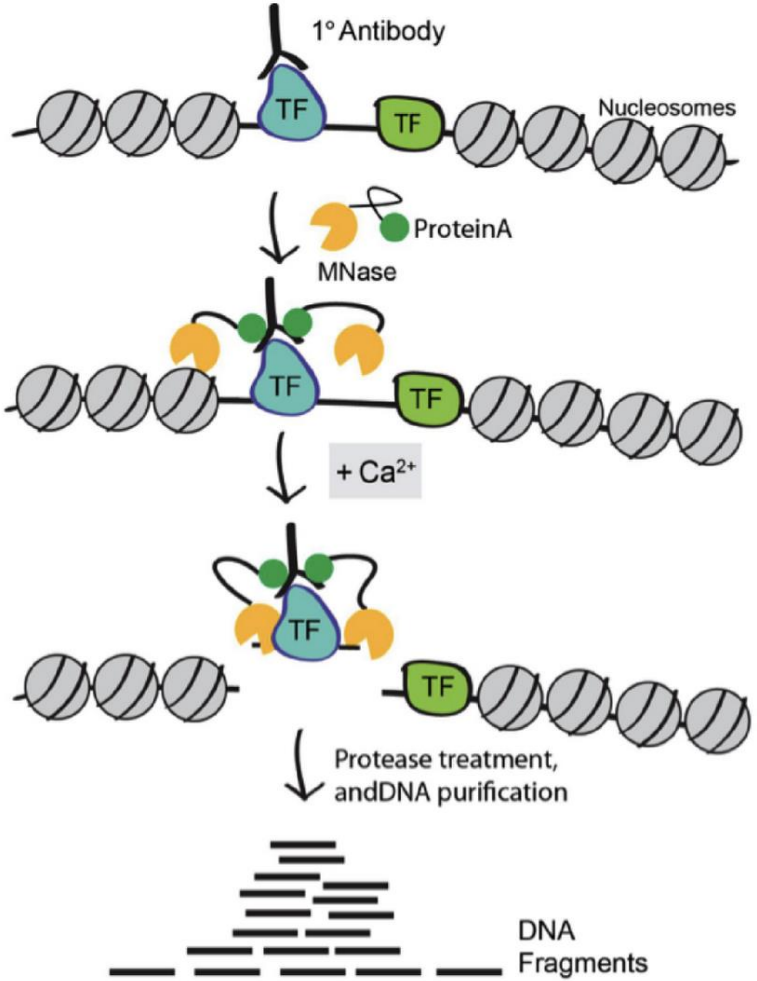
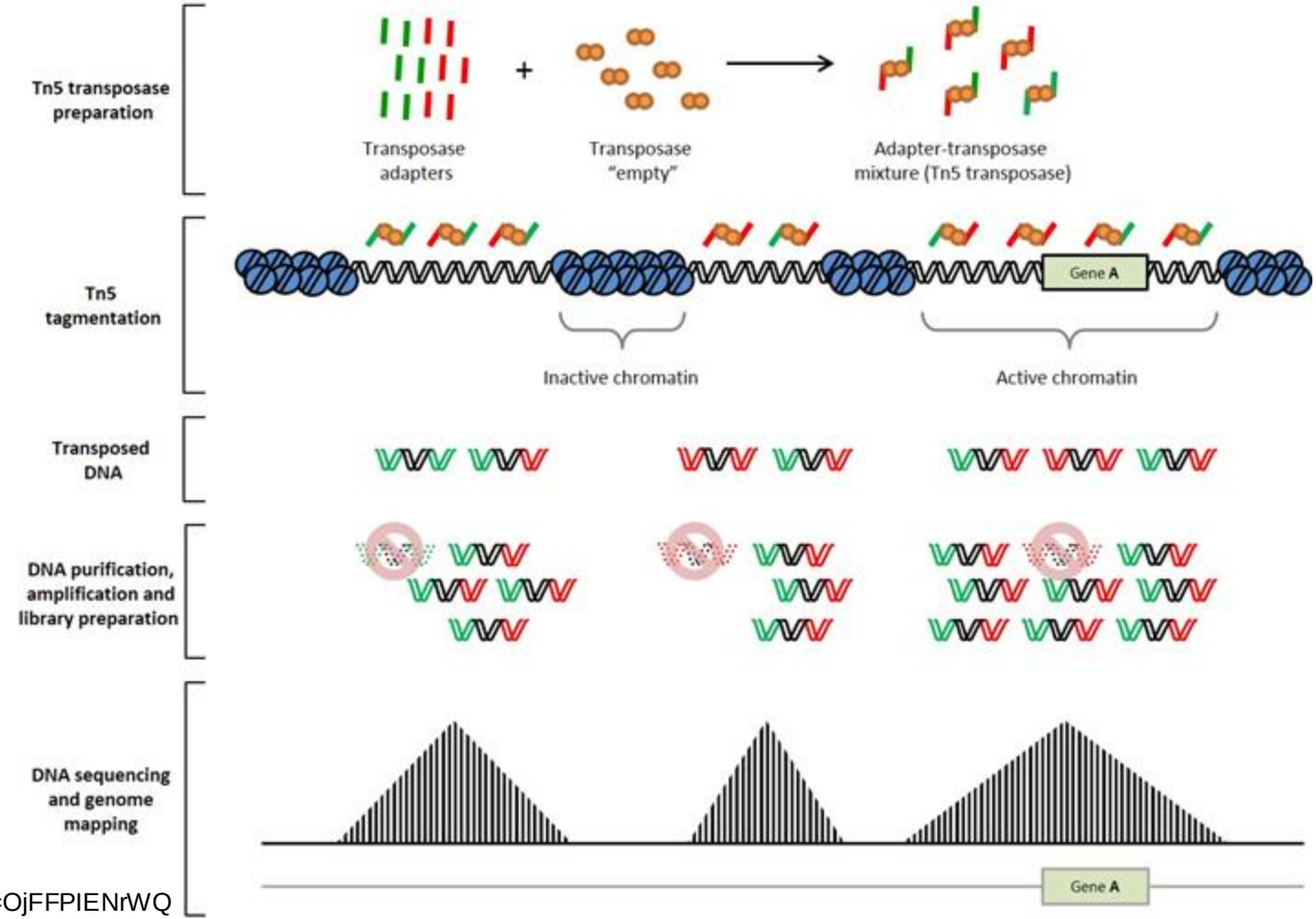


Figure 1. CUT&RUN schematic (see text for details).

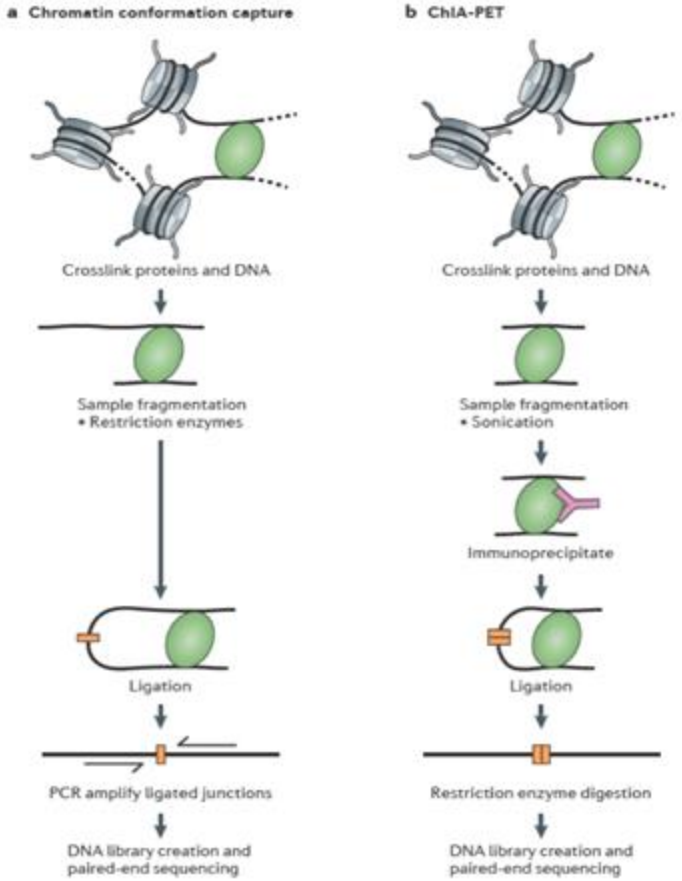
ATAC-seq

DNA accessibility



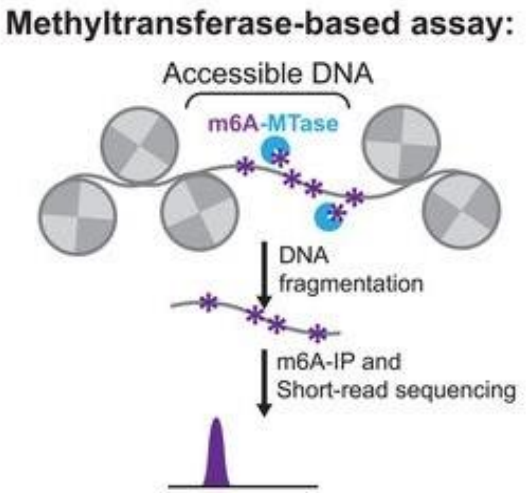
https://www.youtube.com/watch?v=OjFFPIENrWQ

Examining 3-D DNA interactions in the nucleus

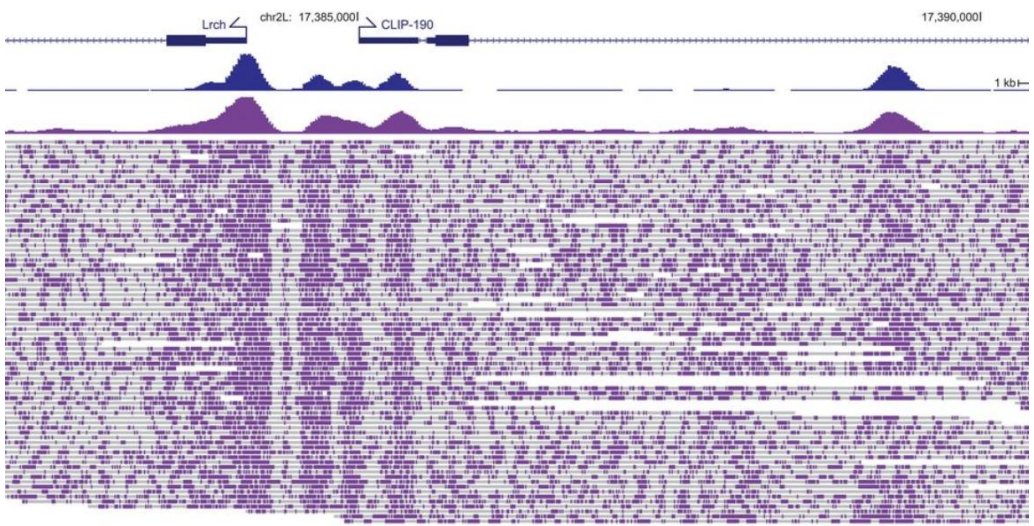


Nature Reviews Genetics 2012 13:840

Fiber-seq

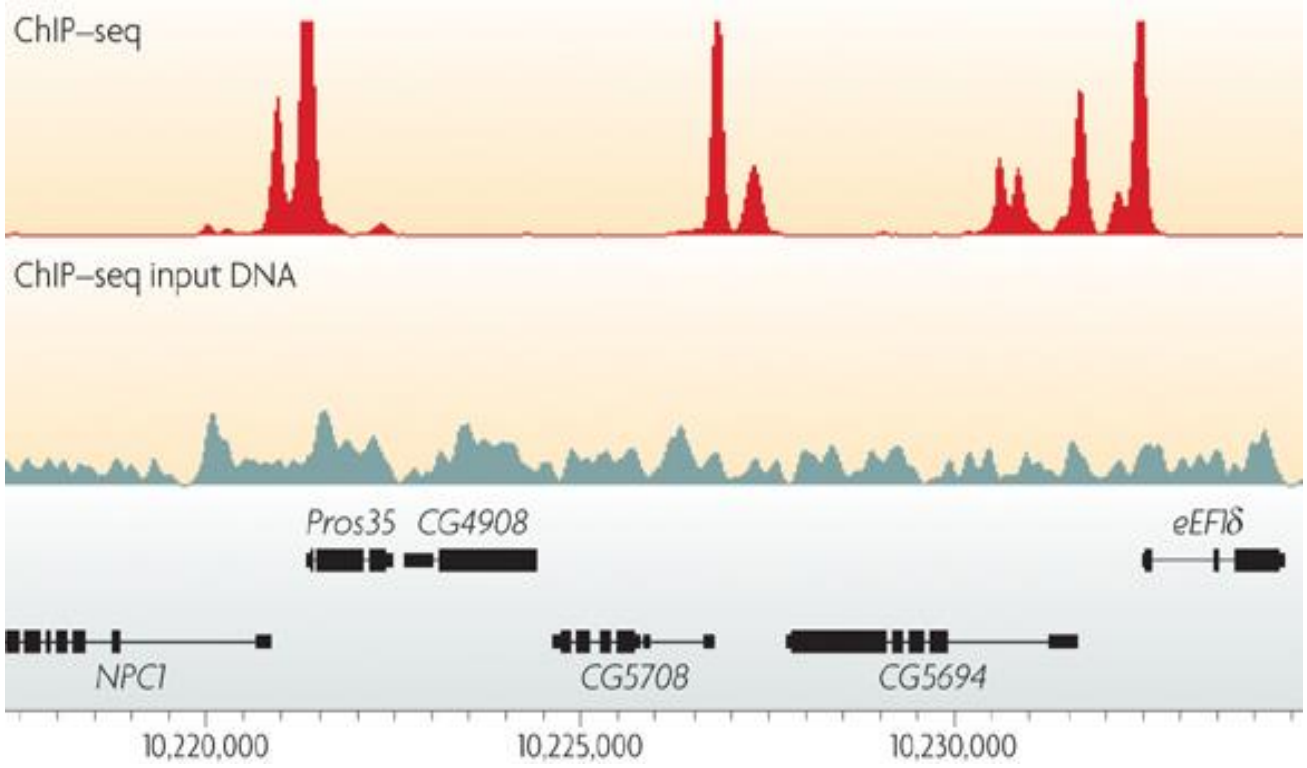


Long-read sequencing (PacBio/ONT)



https://www.science.org/doi/10.1126/science.aaz1646

Peak-calling



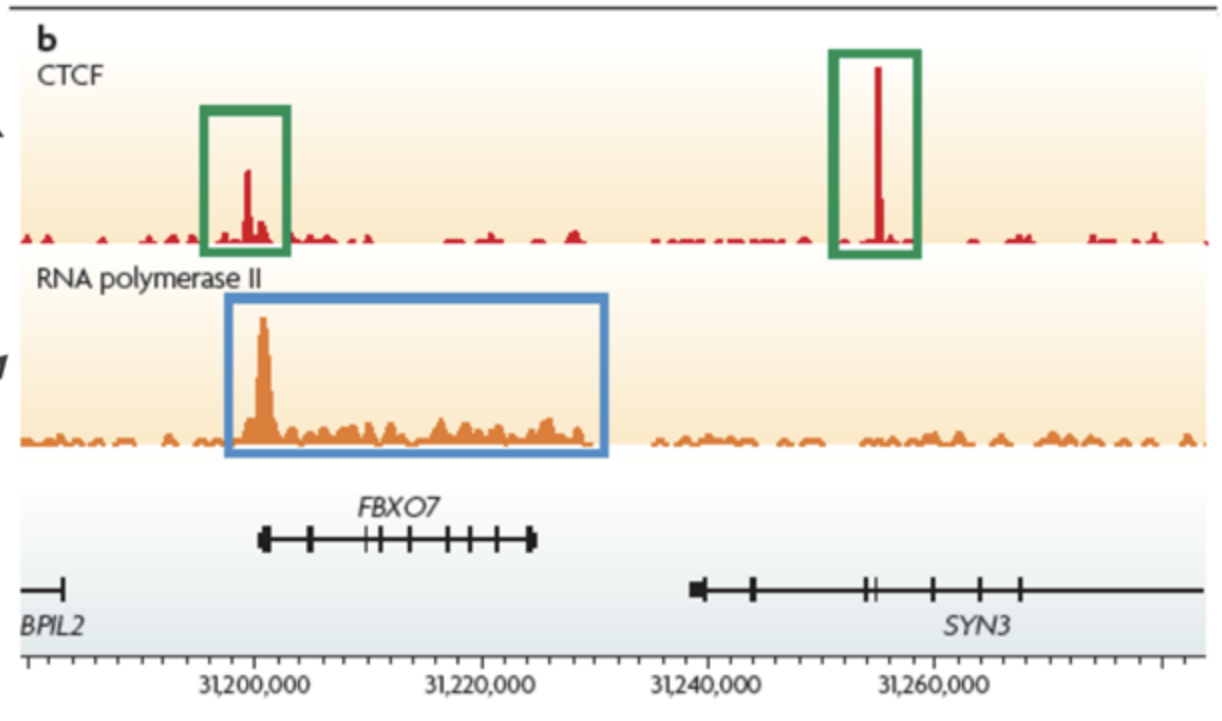
Fundamentally a signal vs noise problem

Nature Reviews Genetics 10, 669-680 (October 2009)

Proteins bind in different ways

Transcription factor – tight, highly-peaked binding region

RNA PolII – enriched at TSS but bound throughout gene body



https://www.ebi.ac.uk/sites/ebi.ac.uk/files/content.ebi.ac.uk/materials/2013/131021_HTS/introduction_to_chip_seq_data_and_analysis-b.gerle_.pdf

How much sequence coverage do we need?

Transcription factor – tight, highly-peaked binding region

For mammalian TFs, other chromatin mods like enhancer-associated histone marks:

- have on the order of thousands of binding sites,
- 20 million reads may be adequate
- (4 million reads for worm and fly TFs).

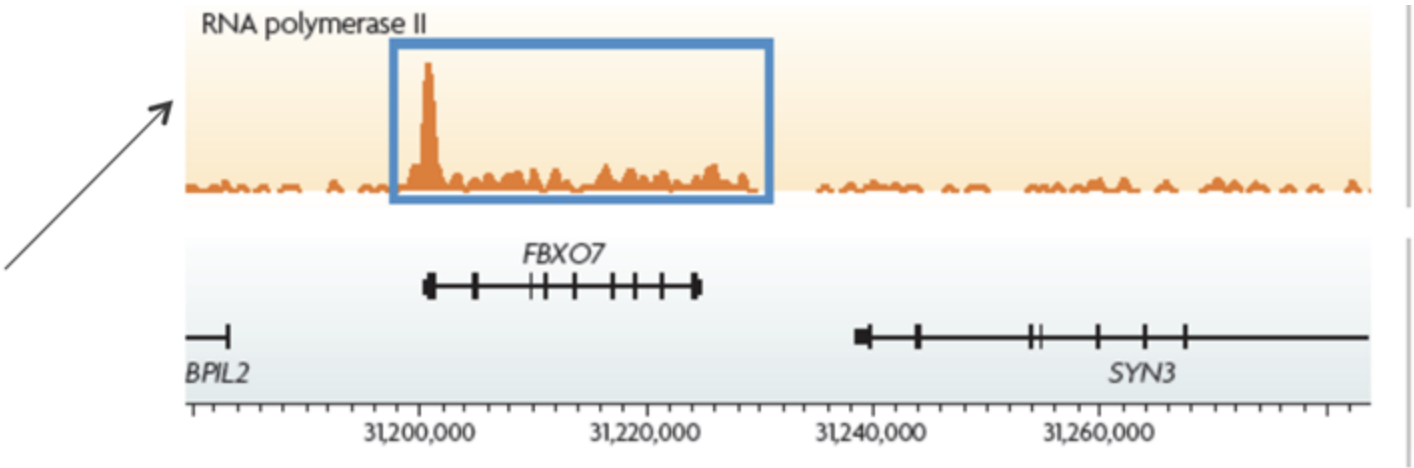
https://www.ebi.ac.uk/sites/ebi.ac.uk/files/content.ebi.ac.uk/materials/2013/131021_HTS/introduction_to_chip_seq_data_and_analysis-b.gerle_.pdf

How much sequence coverage do we need?

More binding sites (e.g., RNA Pol II) or broader factors, including most histone marks,

- require more reads, up to 60 million for mammalian ChIP-seq.

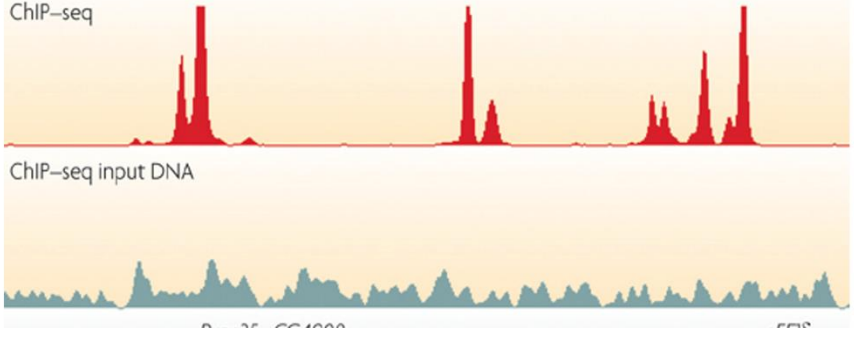
RNA PolII – enriched at TSS but bound throughout gene body



https://www.ebi.ac.uk/sites/ebi.ac.uk/files/content.ebi.ac.uk/materials/2013/131021_HTS/introduction_to_chip_seq_data_and_analysis-b.gerle_.pdf

How much sequence coverage do we need?

- In TF experiments, Control samples should be sequenced **more deeply** than the ChIP-seq samples
    - True for assays involving diffused broad-domain chromatin data.
- This is to ensure sufficient coverage of a substantial portion of the genome and non-repetitive autosomal DNA regions.



ChIP-seq statistics

Continuous variables:

- Your exact height
- Your dog's exact weight
- The winning time in a race
- Exact distance between stars
- Your exact age
- Time it takes a computer to complete a task.

Discrete variables:

- The number of lightbulbs that burn out in a warehouse in a given week.
- The number of heads when flipping a coin 50 times.
- The number of students in a class
- The number of times you forget the attachment to an email on Fridays.
- The number of green M&Ms in a bag
- **The number of times a given base is sequenced.**

The Poisson distribution:

discrete distribution to model coverage

$$P(k \text{ discrete events}) = \lambda^k e^k / k!$$

Where e is Euler's constant (2.718), and ƛ is the average number of occurrences of an event

Example: the "hundred year flood". Thus ƛ=1 (1 catastrophic flood every 100 years

Example: the "hundred year flood".

Thus ƛ=1 (1 catastrophic flood every 100 years

$$P(k \text{ overflow floods in 100 years}) = \frac{\lambda^k e^{-\lambda}}{k!} = \frac{1^k e^{-1}}{k!}$$

$$P(k = 0 \text{ overflow floods in 100 years}) = \frac{1^0 e^{-1}}{0!} = \frac{e^{-1}}{1} = 0.368$$

$$P(k = 1 \text{ overflow flood in 100 years}) = \frac{1^1 e^{-1}}{1!} = \frac{e^{-1}}{1} = 0.368$$

$$P(k = 2 \text{ overflow floods in 100 years}) = \frac{1^2 e^{-1}}{2!} = \frac{e^{-1}}{2} = 0.184$$

k	P(*k* overflow floods in 100 years)
0	0.368
1	0.368
2	0.184
3	0.061
4	0.015
5	0.003
6	0.0005

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

Example: expected number of goals in a World Cup game.

Average number of goals is 2.5

$$P(k \text{ goals in a match}) = \frac{2.5^k e^{-2.5}}{k!}$$

$$P(k = 0 \text{ goals in a match}) = \frac{2.5^0 e^{-2.5}}{0!} = \frac{e^{-2.5}}{1} = 0.082$$

$$P(k = 1 \text{ goal in a match}) = \frac{2.5^1 e^{-2.5}}{1!} = \frac{2.5 e^{-2.5}}{1} = 0.205$$

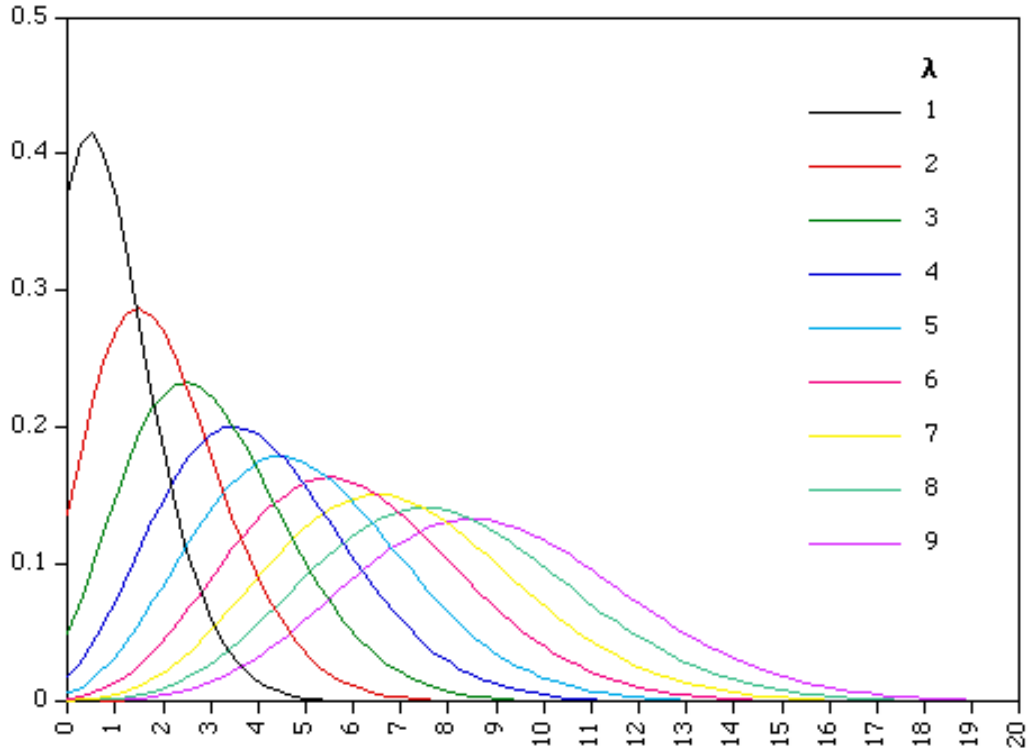
$$P(k = 2 \text{ goals in a match}) = \frac{2.5^2 e^{-2.5}}{2!} = \frac{6.25 e^{-2.5}}{2} = 0.257$$

Note that λ does not have to be a countable integer.

k	P(*k* goals in a World Cup soccer match)
0	0.082
1	0.205
2	0.257
3	0.213
4	0.133
5	0.067
6	0.028

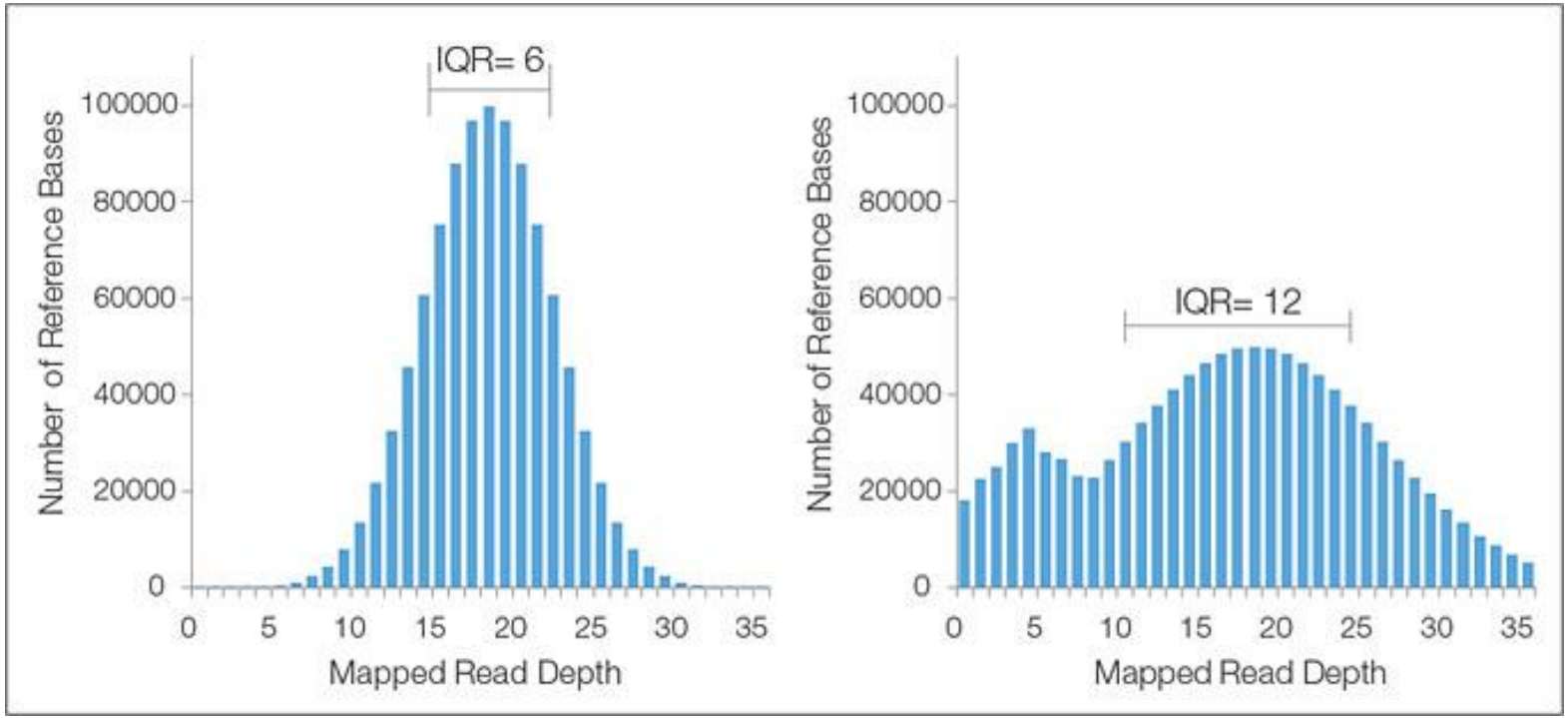
https://en.wikipedia.org/wiki/Poisson_distribution

Poisson distribution with different values of ƛ



https://www.umass.edu/wsp/resources/poisson/

Ideally, sequencing coverage will follow a Poisson distribution. But...

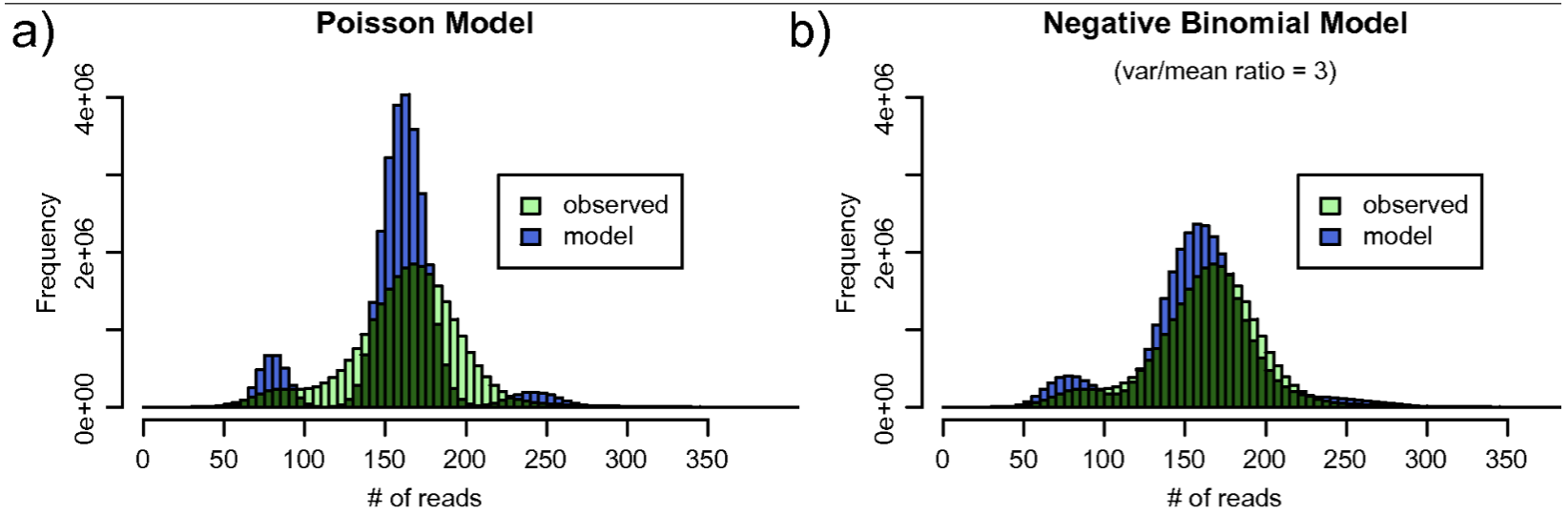


Poisson

Not Poisson.
Overly "dispersed"

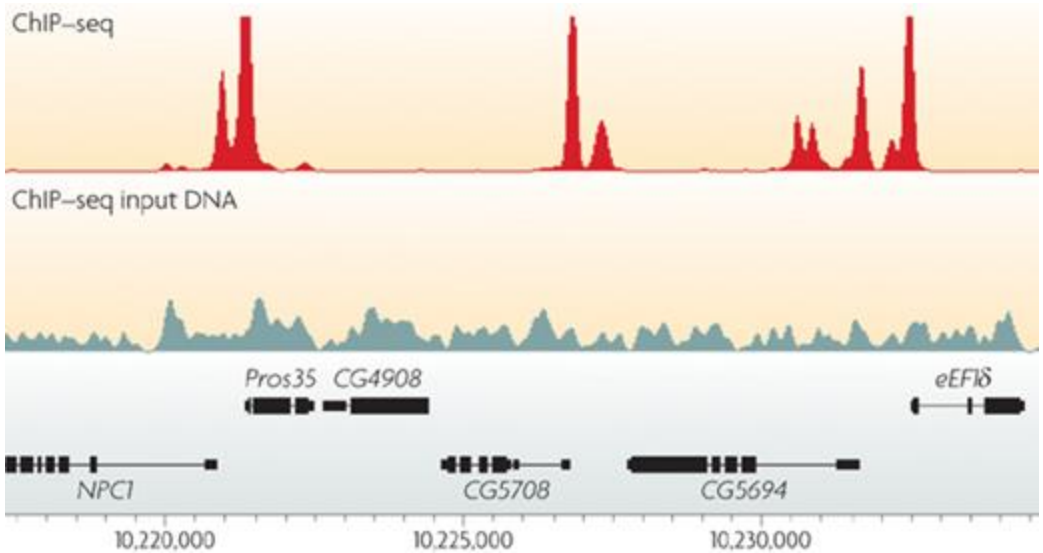
https://www.umass.edu/wsp/resources/poisson/

Negative binomial fits sequencing coverage data much better



https://www.umass.edu/wsp/resources/poisson/

Comparative ChIP-seq: scaling and normalizing



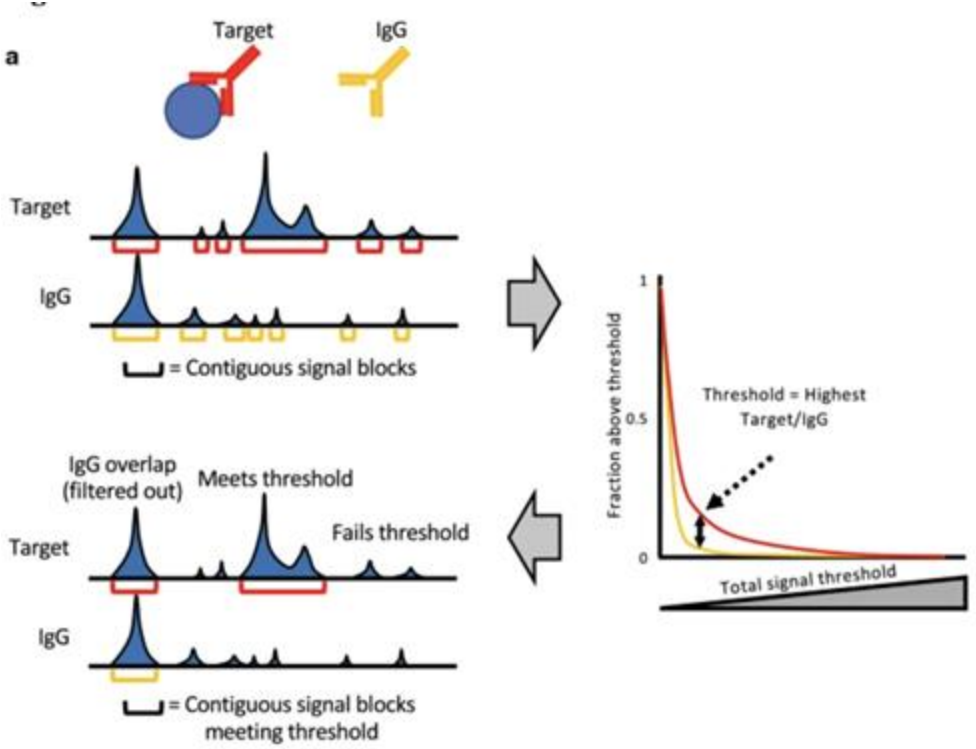
Sequencing depths from TF ChIP should be compared to control (input DNA) to get a sense of the noise.

When comparing two ChIP-seq experiments, you need to normalize the counts / peaks before doing so. **E.g., comparing experiments where one had 10 million reads and the other had 100 million reads.**

Nature Reviews Genetics 10, 669-680 (October 2009)

Peak-calling

MACS2, HOMER, SEACR, etc



ChIP-seq peak callers

Table S1. Examples of peak callers employed in ChIP-seq.

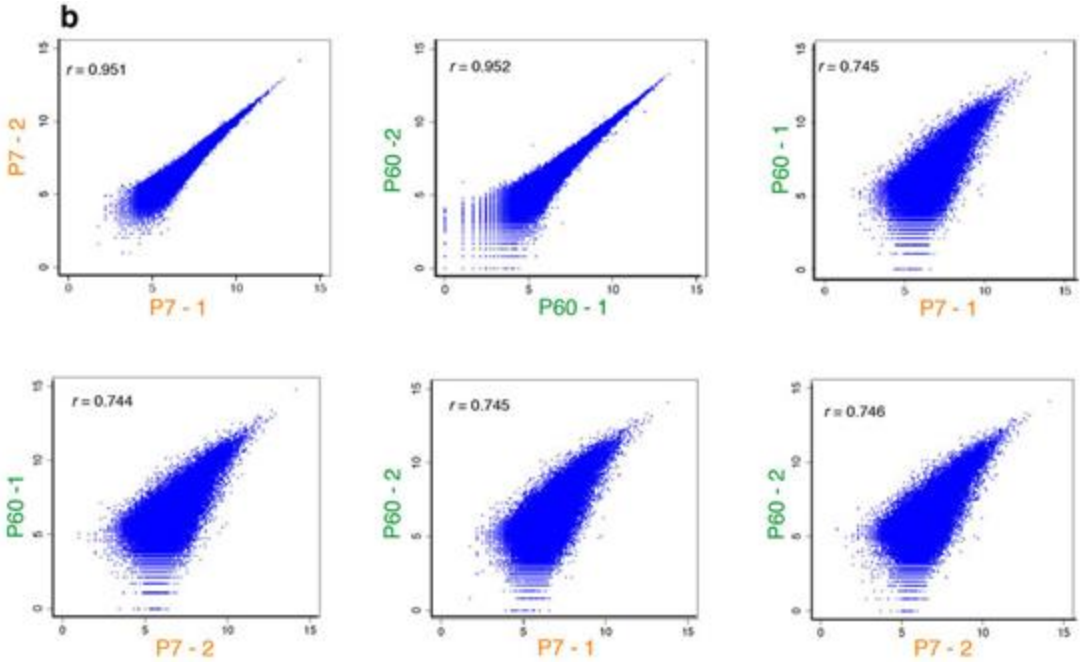
Software tool	Version	Availability	Point-source (peaks)	Broad regions (domains)
BayesPeak [88]	1.10.0	http://bioconductor.org/packages/release/bioc/html/BayesPeak.html	Yes	
BEADS[‡] [84]	1.1	http://beads.sourceforge.net/	Yes	Yes
CCAT [91]	3.0	http://cmb.gis.a-star.edu.sg/ChIPSeq/paperCCAT.htm		Yes
CisGenome [56]	2.0	http://www.biostat.jhsph.edu/~hji/cisgenome/	Yes	
CSAR [85]	1.10.0	http://bioconductor.org/packages/release/bioc/html/CSAR.html	Yes	
dPeak	0.9.9	http://www.stat.wisc.edu/~chungdon/dpeak/	Yes	
GPS/GEM [67,18]	1.3	http://cgs.csail.mit.edu/gps/	Yes	
HPeak [87]	2.1	http://www.sph.umich.edu/csg/qin/HPeak/	Yes	
MACS [17]	2.0.10	https://github.com/taoliu/MACS/	Yes	Yes
NarrowPeaks[‡]	1.4.0	http://bioconductor.org/packages/release/bioc/html/NarrowPeaks.html	Yes	
PeakAnalyzer/ PeakSplitter[‡] [89]	1.4	http://www.bioinformatics.org/peakanalyzer	Yes	
PeakRanger [93]	1.16	http://ranger.sourceforge.net/	Yes	Yes
PeakSeq [24]	1.1	http://info.gersteinlab.org/PeakSeq	Yes	
polyaPeak[‡]	0.1	http://web1.sph.emory.edu/users/hwu30/polyaPeak.html	Yes	
RSEG [92]	0.6	http://smithlab.usc.edu/histone/rseg/		Yes
SICER [90]	1.1	http://home.gwu.edu/~wpeng/Software.htm		Yes
SIPeS [21]	2.0	http://gmdd.shgmo.org/Computational-Biology/ChIP-Seq/download/SIPeS	Yes	
SISSRs [19]	1.4	http://sissrs.rajajothi.com/	Yes	
SPP [9]	1.1	http://compbio.med.harvard.edu/Supplements/ChIP-seq/	Yes	Yes
USeq [97]	8.5.1	http://sourceforge.net/projects/useq/	Yes	
ZINBA [86]	2.02.03	http://code.google.com/p/zinba/	Yes	Yes

[‡] Only for post-processing.

MACS is probably the most widely used

http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003326

Replicates, replicates, replicates

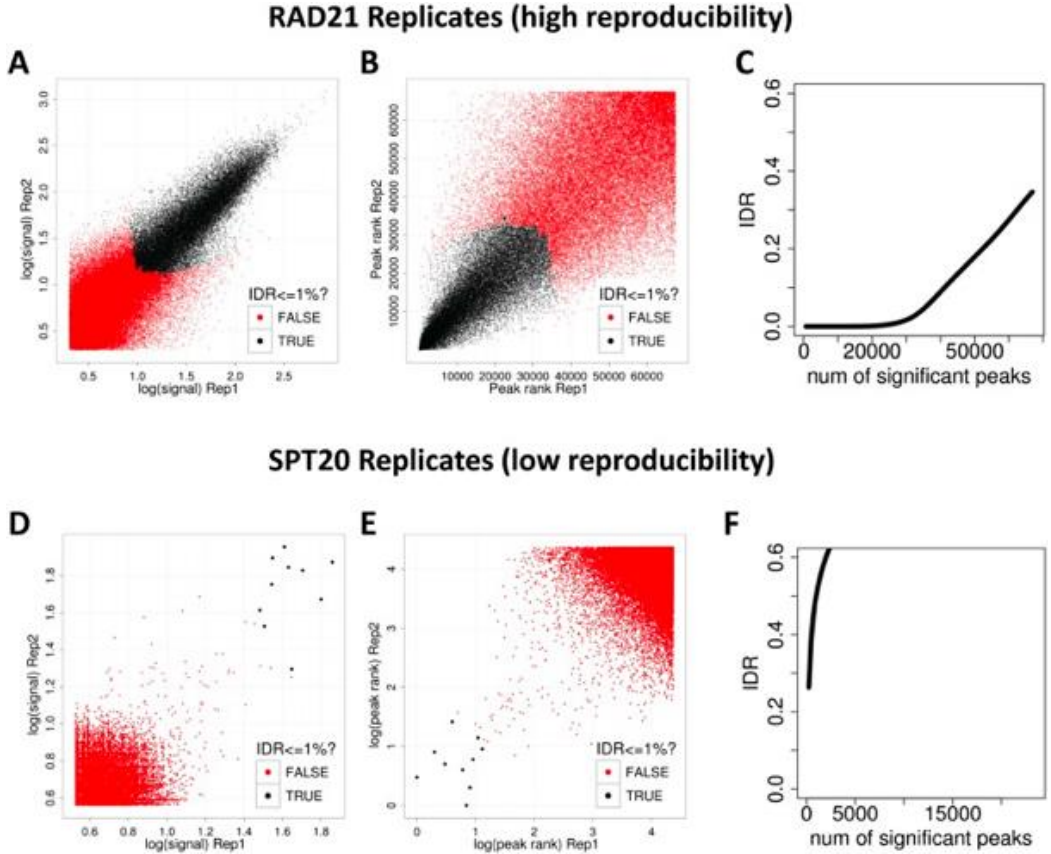


(**b**) Scatterplots of pairwise Zic ChIP-seq replicates with Pearson correlation (*r*) displayed. **Note the correlations are much higher between biological replicates of the same developmental stage than between P7 and P60 cerebellum**.

Pearson Correlation is one metric indicative of overall reproducibility

http://www.nature.com/neuro/journal/v18/n5/full/nn.3995.html

Irreproducibility Discovery Rate (IDR)

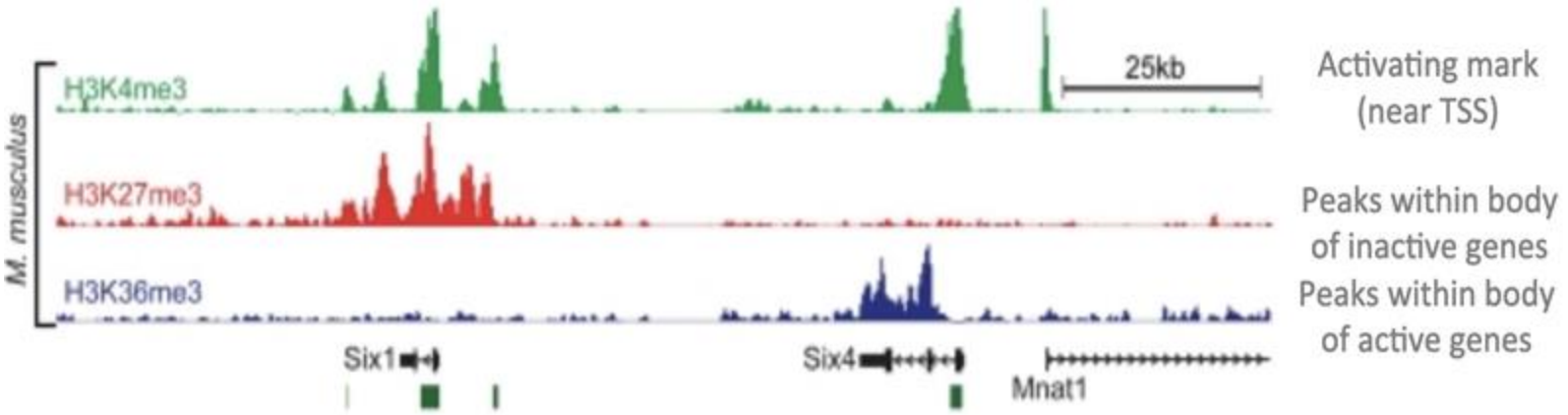


Peaks that show up consistently among replicates are more likely to be real!

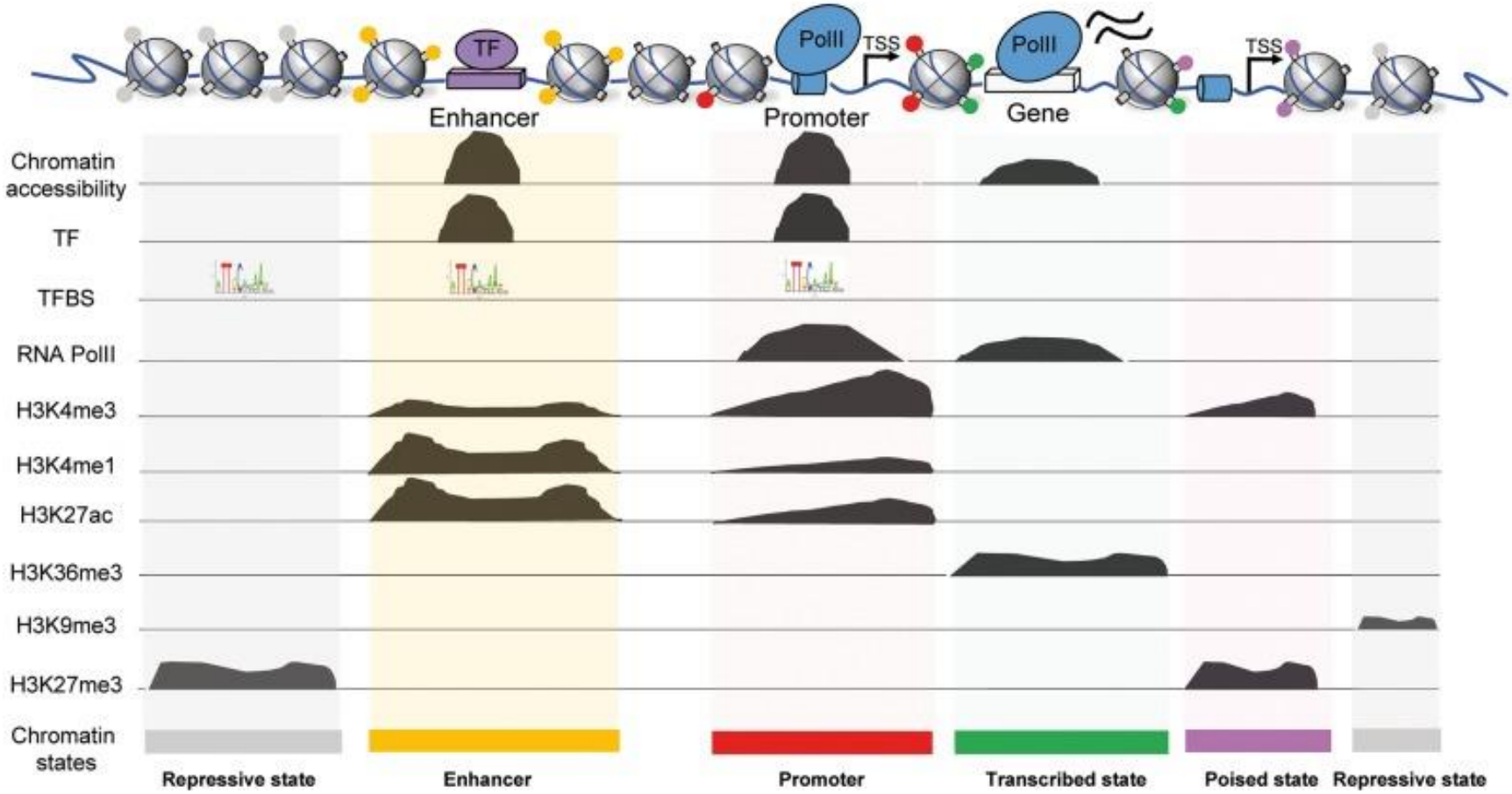
Fantastic resource for learning ChIP-seq analysis

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

Interpretation

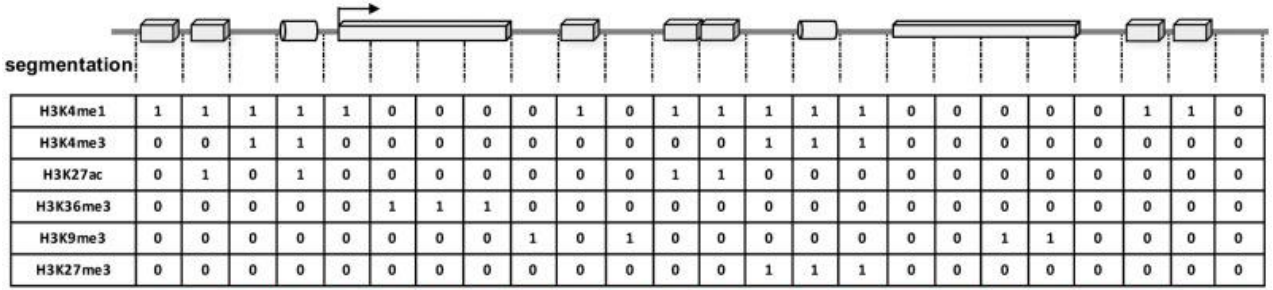


https://www.ebi.ac.uk/sites/ebi.ac.uk/files/content.ebi.ac.uk/materials/2013/131021_HTS/introduction_to_chip_seq_data_and_analysis-b.gerle_.pdf

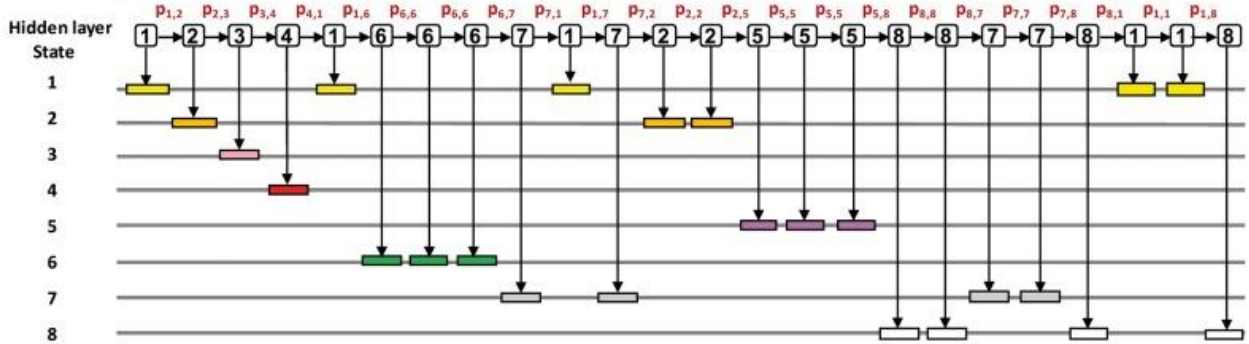


https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5888983/

Genome Segmentation with ChromHMM



H3K4me1	1	1	1	1	1	0	0	0	0	1	0	1	1	1	1	1	0	0	0	0	0	1	1	0
H3K4me3	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0
H3K27ac	0	1	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
H3K36me3	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H3K9me3	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0
H3K27me3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0

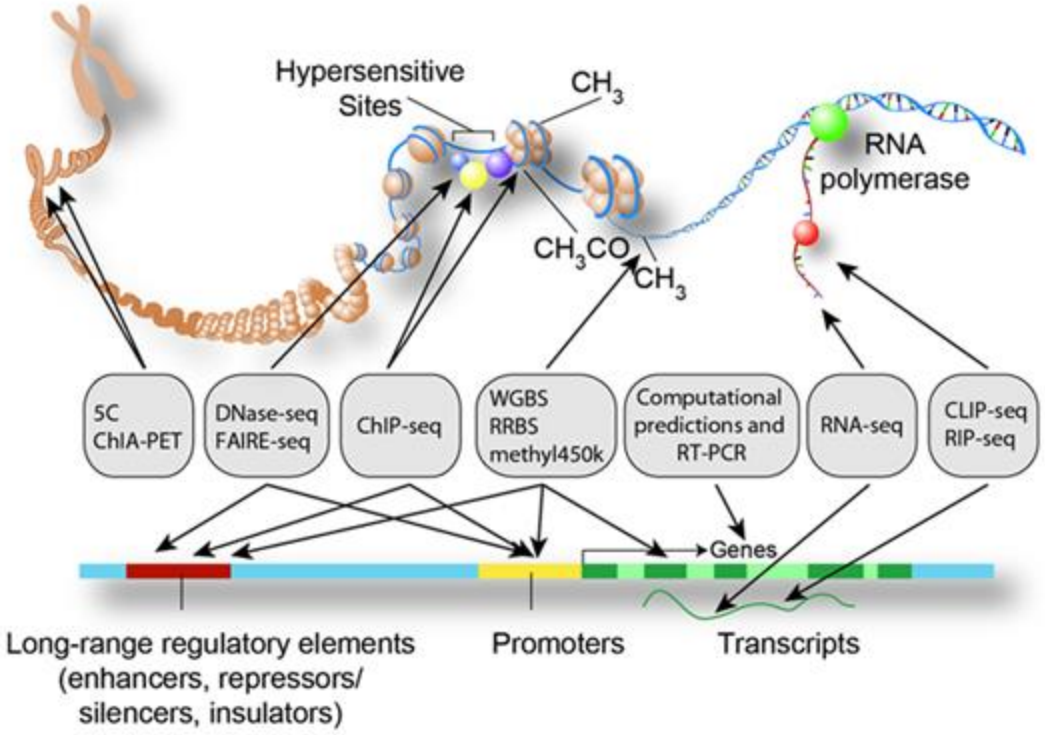


Chromatin state	State annotation
1	Poised enhancer
2	Active enhancer
3	Weak promoter
4	Active promoter
5	Poised promoter
6	Transcriptional elongation
7	Repressive state
8	Low signal state

Be aware that segmentation may be tissue or cell-type specific!

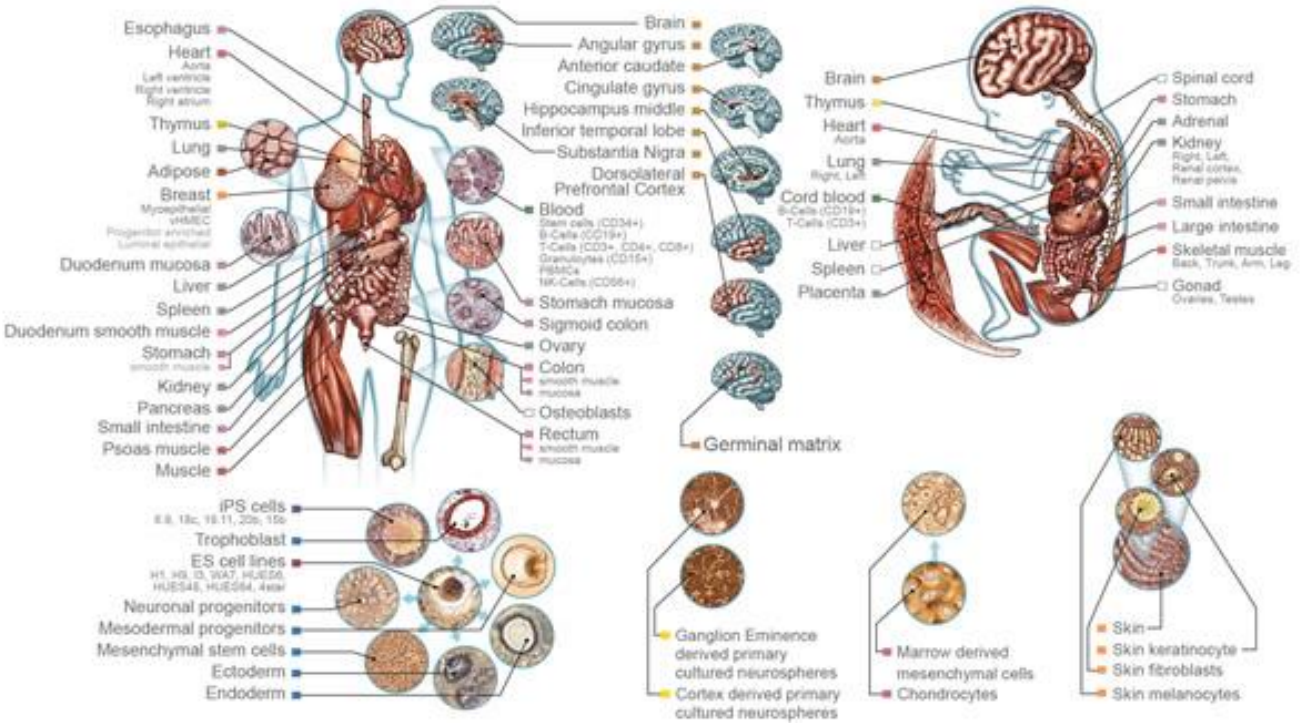
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5888983/

The Encyclopedia of DNA Elements Project



https://www.encodeproject.org/

Roadmap Epigenomics Project



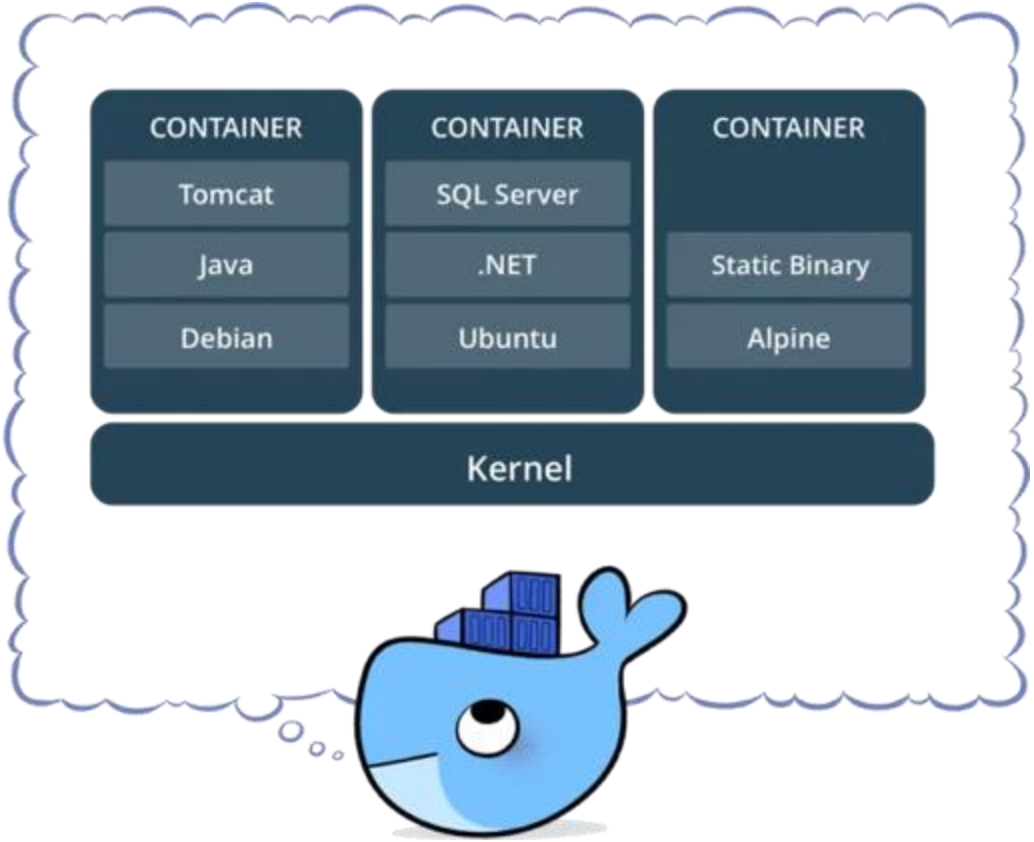
http://commonfund.nih.gov/epigenomics/index

ChIP-seq Exercise

Computing Environments

- Laptop
  - You administer
  - You control completely
- Shared compute cluster
  - A sysadmin or group administers it
  - You control very little
- Docker (containers)
  - Sysadmins handle the hardware
  - You control the software almost completely

Docker containers



Docker Exercise

https://gist.github.com/chrisamiller/fab79aa56b259e01744330e1d245085e

Anatomy of a docker command

```
docker run -v /workspace:/data -it mgibio/gatk-cwl:3.6.0 /bin/bash
```

Anatomy of a docker command

```
docker run -v /workspace:/data -it mgibio/gatk-cwl:3.6.0 /bin/bash
```

tells the **docker** program that you want to **run** an image

Anatomy of a docker command

```
docker run -v /workspace:/data -it mgibio/gatk-cwl:3.6.0 /bin/bash
```

mounting directories:
the `/workspace` directory on your actual computer will be linked to
the `/data` directory inside your image

Any part of your computer that isn't linked explicitly
will be invisible inside the docker container!

Sometimes we'll use bash substitution to link the current directory:

```
-v $(pwd -P):/data
```

Anatomy of a docker command

```
docker run -v /workspace:/data -it mgibio/gatk-cwl:3.6.0 /bin/bash
```

run docker interactively, that is, drops you into a terminal

Anatomy of a docker command

```
docker run -v /workspace:/data -it mgibio/gatk-cwl:3.6.0 /bin/bash
```

the docker image that you want to run

this one is the `gatk-cwl` image,
hosted in the `mgibio` organization on dockerhub
and we're pulling the image that is tagged `3.6.0`

If you don't provide a tag, it looks for a tag called `latest`

(Why might using tags be important for your analyses?)

Anatomy of a docker command

```
docker run -v /workspace:/data -it mgibio/gatk-cwl:3.6.0 /bin/bash
```

The command that you want to run inside of docker.

In this case, we're starting a shell so that we can do command line stuff

Anatomy of a docker command

```
docker run -v /workspace:/data mgibio/gatk-cwl:3.6.0 \
gatk --java-options '-Xmx7g' HaplotypeCaller \
-R /workspace/inputs/references/genome/ref_genome.fa \
-I align/WGS_Norm_merged_sorted_mrkdup_bqsr.bam \
-O /workspace/germline/WGS_Norm_HC_calls.vcf \
--bam-output /workspace/germline/WGS_Norm_HC_out.bam
```

In this example, we removed -it and just added the command we want to run. Docker will automatically exit after running this command.

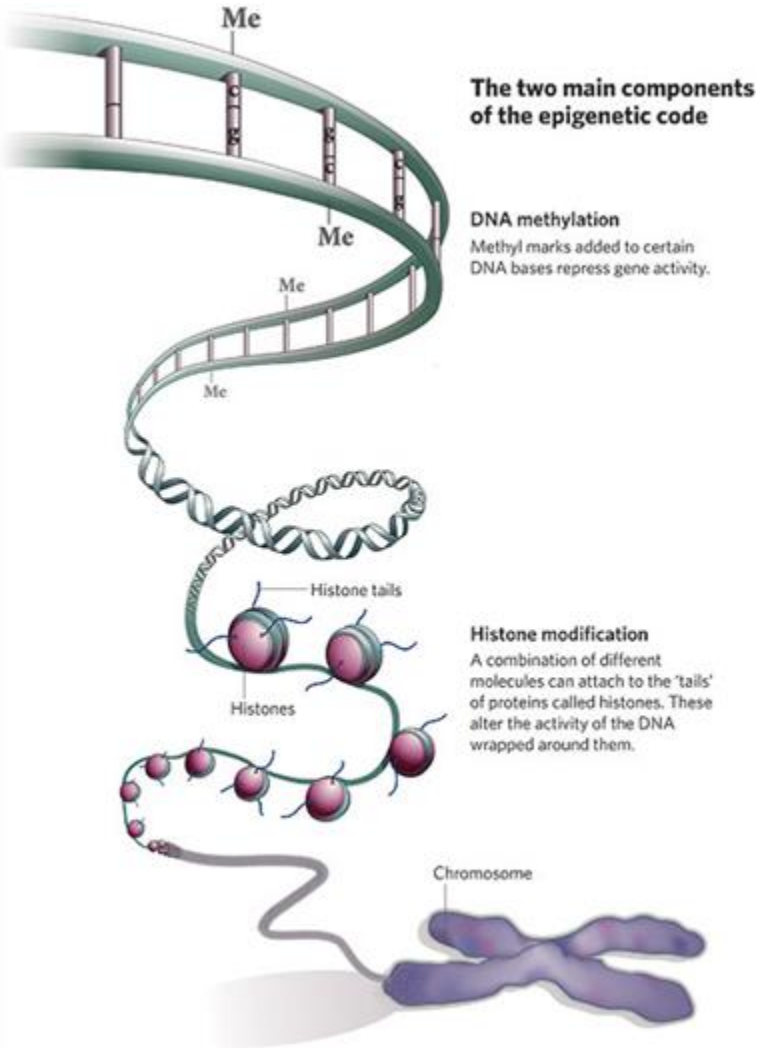
ChIP-seq Exercise

Epigenomics: Bisulfite sequencing

Chris Miller, Ph.D.
Washington University in St. Louis

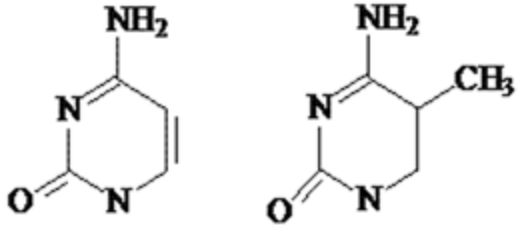
Some slides adapted from:
https://github.com/genome/bfx-workshop
https://github.com/quinlan-lab/applied-computational-genomics

Epigenetics



http://www.nature.com/nature/journal/v441/n7090/full/441143a.html

DNA Methylation

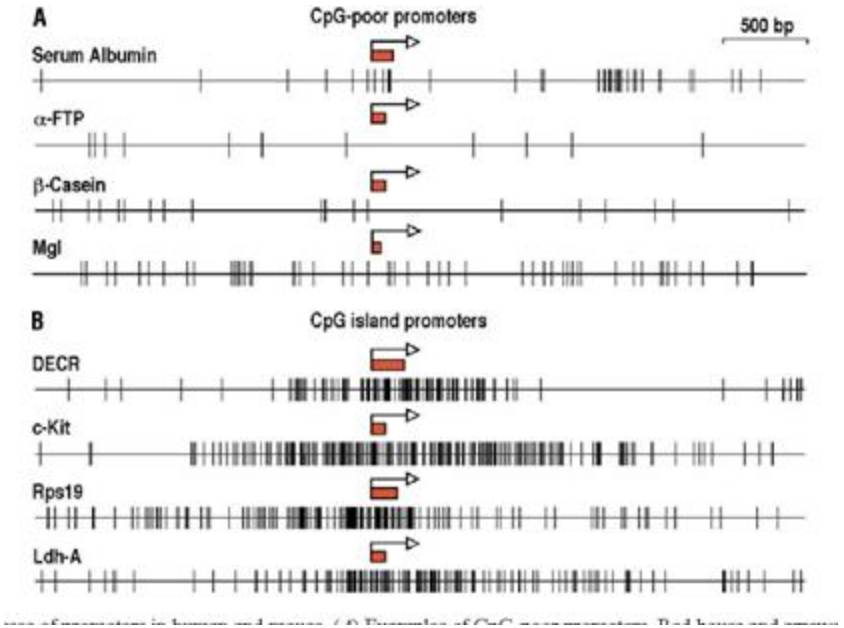


- Mostly happens at CpGs
- About 25 million CpGs in human genome

- https://en.wikipedia.org/wiki/CpG_site#/media/File:CpG_vs_C-G_bp.svg

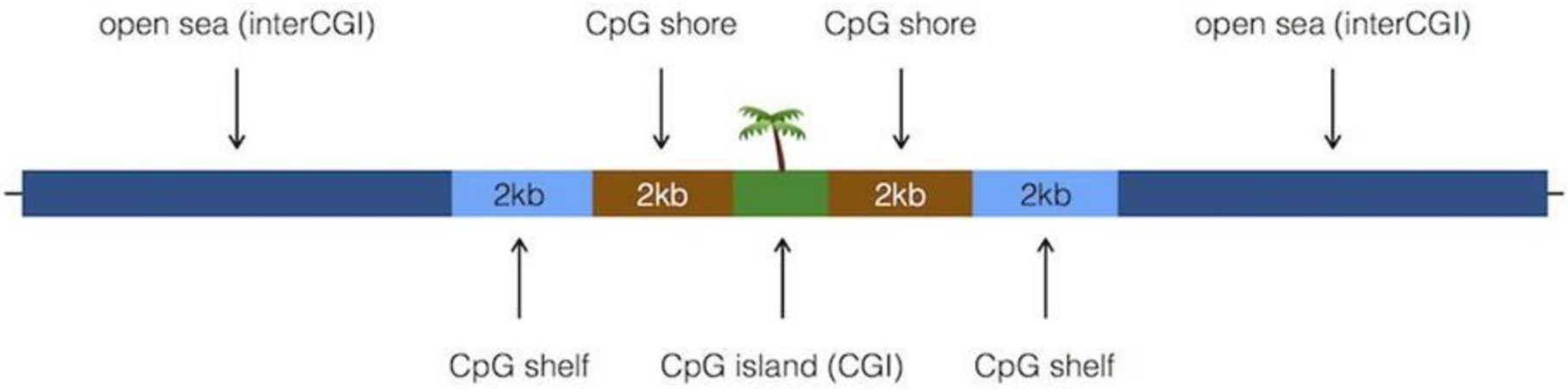
DNA Methylation

- CpG Islands
- Length >= 200 bp
  GC% > 50%
  o/e CpG ratio > 60%
- Selective pressure/
  Evolutionary constraint



- DOI:10.1007/s00018-003-3088-6

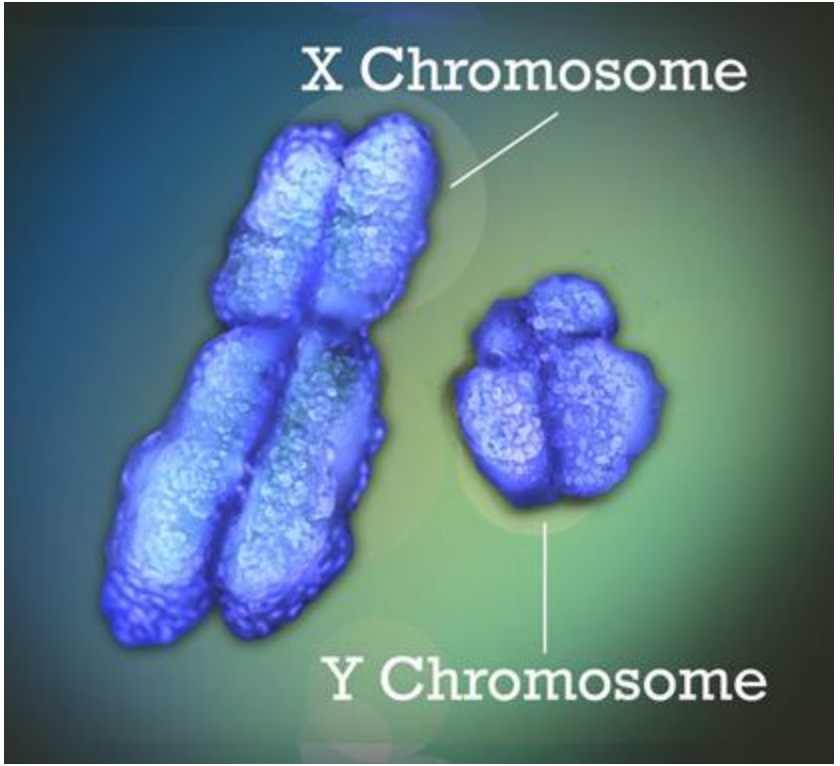
Islands, shores, and shelves



https://bioconductor.org/packages/release/bioc/vignettes/annotatr/inst/doc/annotatr-vignette.html

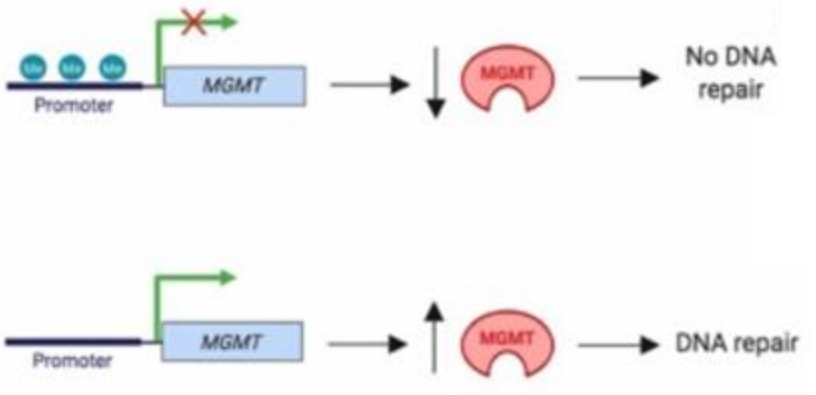
What does DNA methylation do?

- The short answer: It depends!
- X-chromosome inactivation
- Silencing of transposable elements
- Cellular differentiation
- Cancer - hypo/hypermethylation



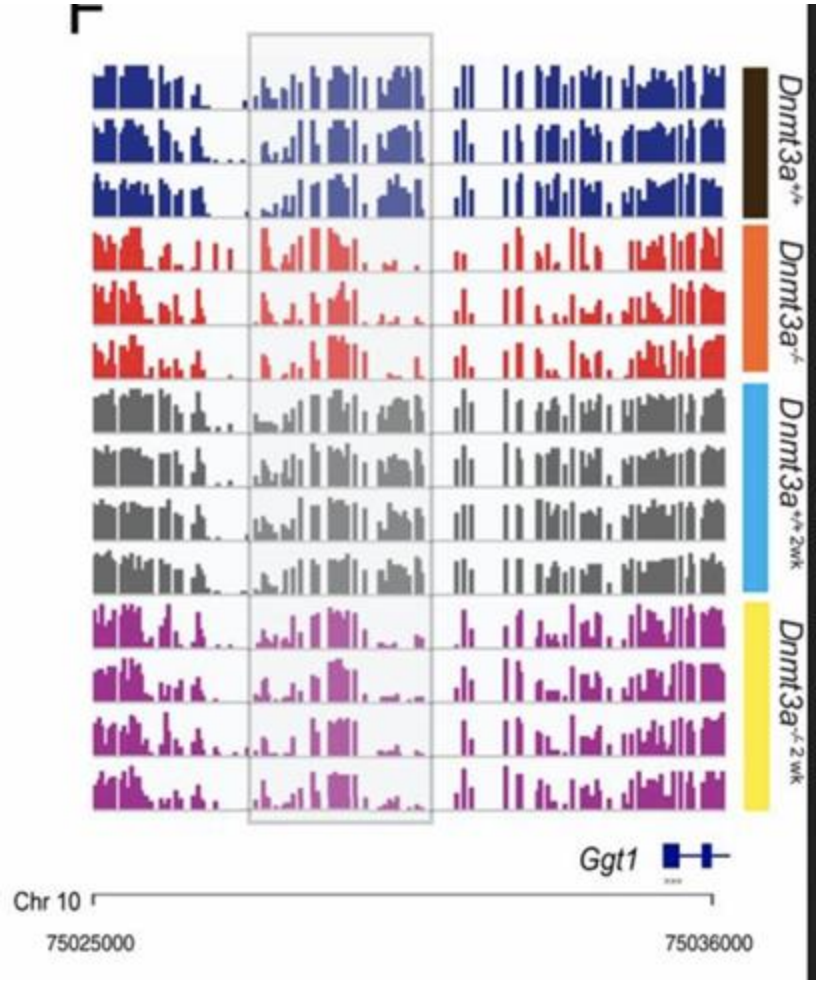
MGMT and Temozolomide

- TMZ is an alkylating agent - damages DNA, causes cell death
- MGMT “cleans up” the damage
- Methylation of the MGMT promoter is linked to better outcomes!



Methylation Patterns

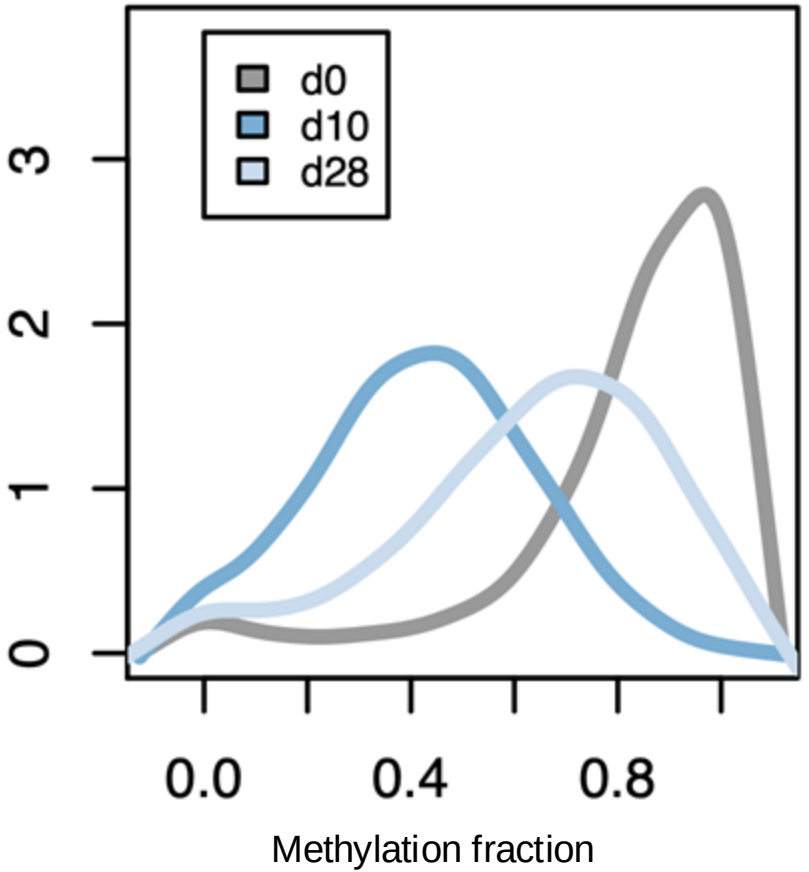
- Methyltransferases that act locally



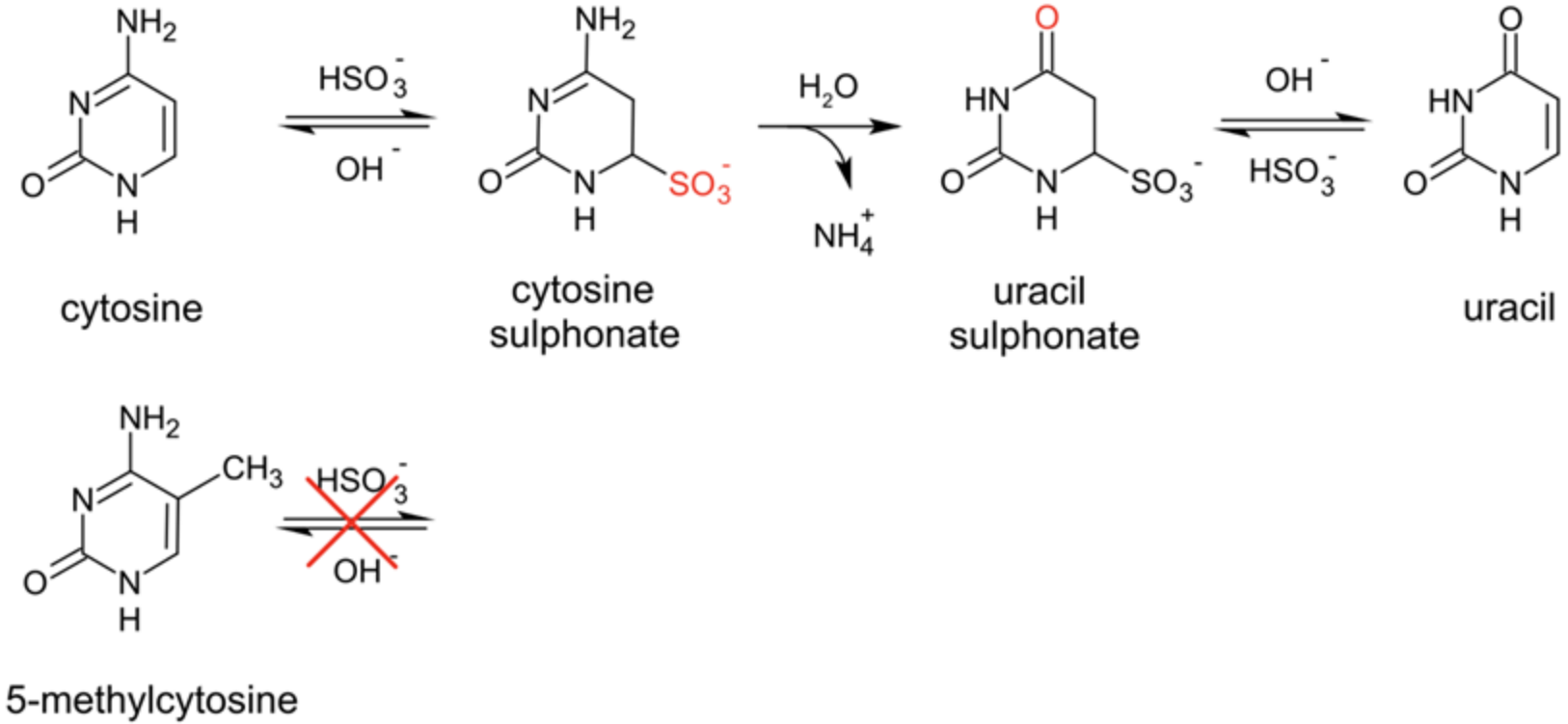
Methylation Patterns

- Methyltransferases that act locally

- Other alterations (or treatments) that act globally

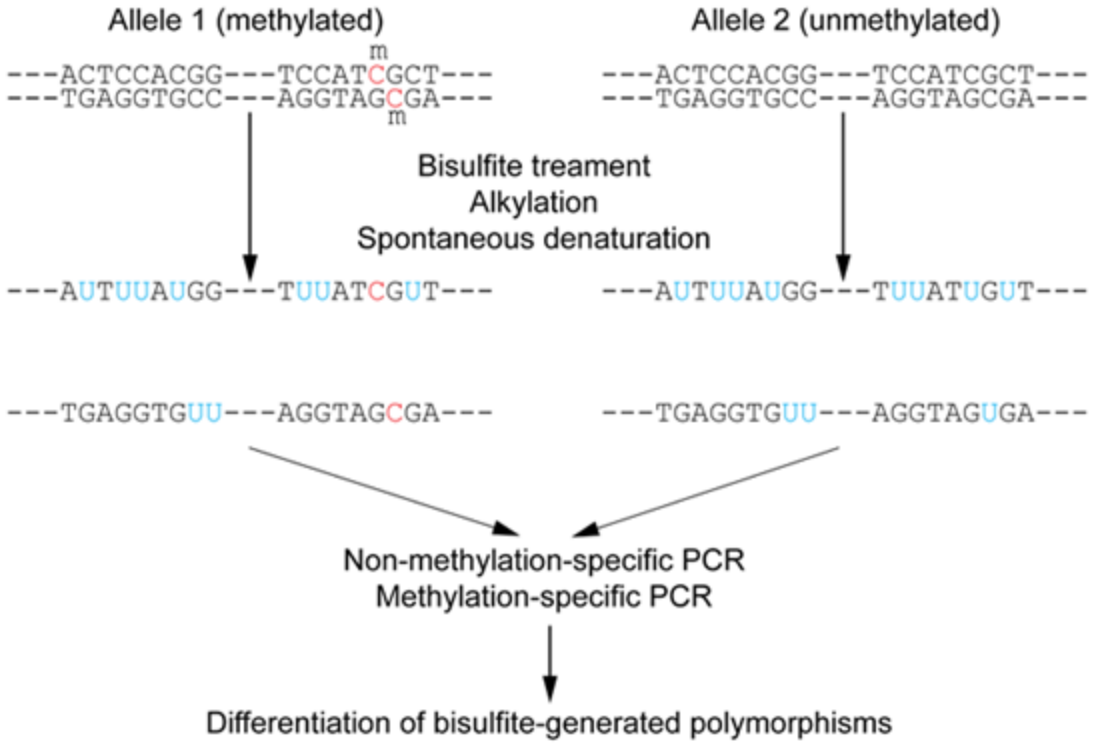


Bisulfite sequencing



https://en.wikipedia.org/wiki/File:Bisulfite_conversion.svg

Bisulfite sequencing



https://en.wikipedia.org/wiki/File:Wiki_Bisulfite_sequencing_Figure_1_small.png

Bisulfite sequencing

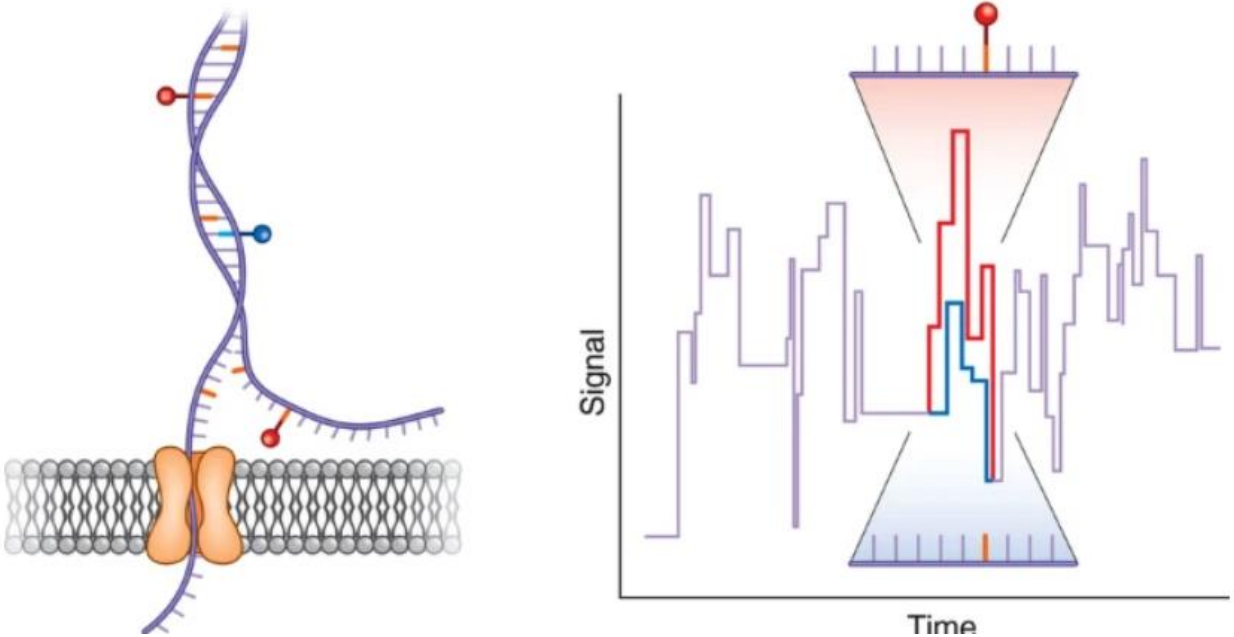
BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics 10:232

Whole-genome Bisulfite Sequencing (WGBS)

- Need a special aligner - has to expect many C > T mismatches!

- BSMAP
- bismark
- BWA-meth
- biscuit

Direct detection with long read sequencing



Can be used for 5mC as well as m6A in direct RNAseq

Methylation calling

- Determine methylation fraction at each site in the genome
    - Count the Cs and Ts, taking strandedness into account
    - Some tools account for SNPs while doing this
-

Methylation calling

- Determine methylation fraction at each site in the genome
  - Count the Cs and Ts, taking strandedness into account
  - Some tools account for SNPs while doing this

- Why isn’t every position 0%, 50% or 100%?

Methylation calling

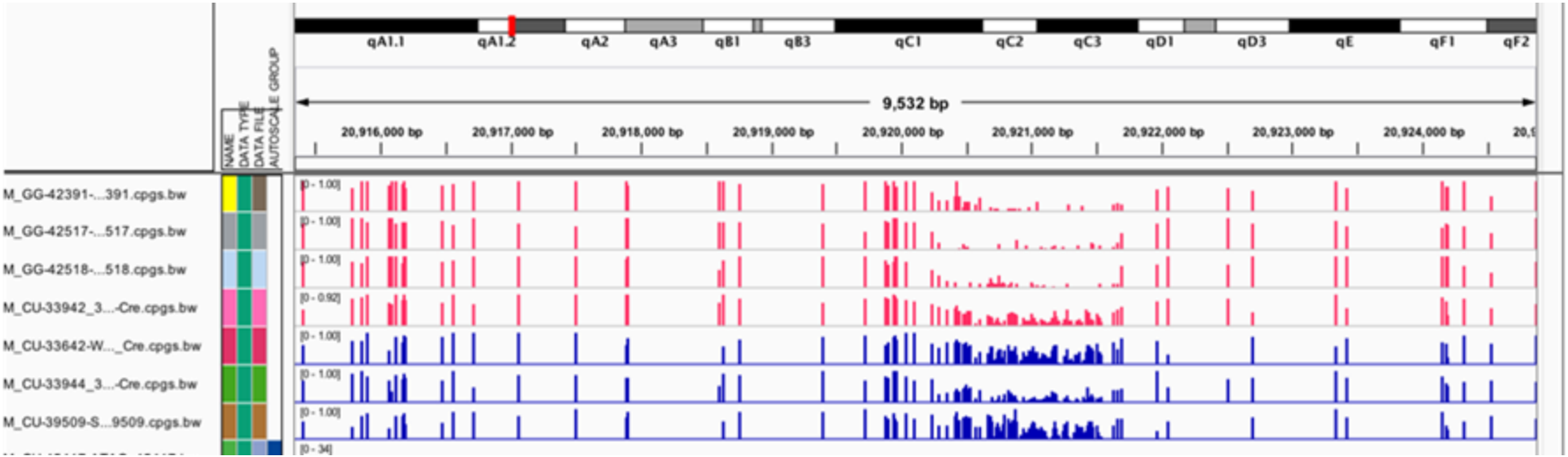
- Determine methylation fraction at each site in the genome
  - Count the Cs and Ts, taking strandedness into account
  - Some tools account for SNPs while doing this

- Why isn’t every position 0%, 50% or 100%?
  - we’re sequencing a population of cells!

Workflow/File formats

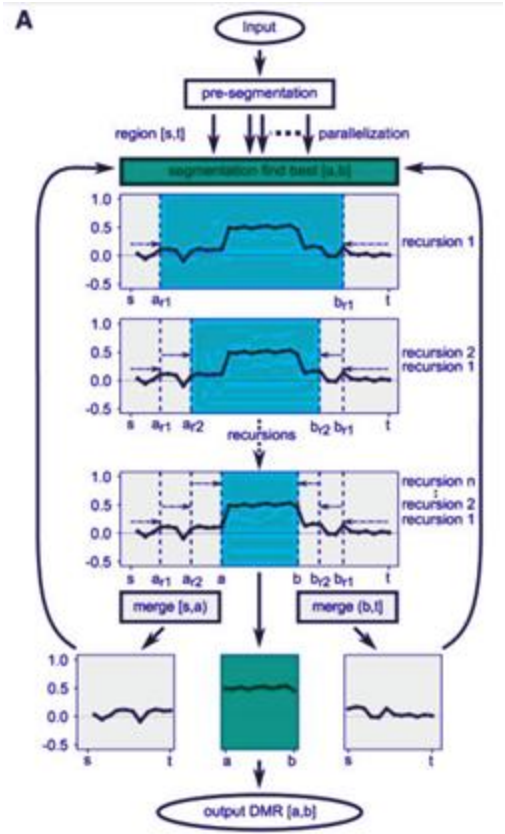
- Aligning: FASTQ > BAM/CRAM
- Pileup: BAM/CRAM > VCF
  - (entries for every site, allele frequencies)
- VCF > bedgraph
  - chr, start, stop, beta_value (methylation fraction)
- bedgraph > bigwig (for visualization in IGV)
- There are workflows for this!

IGV visualization



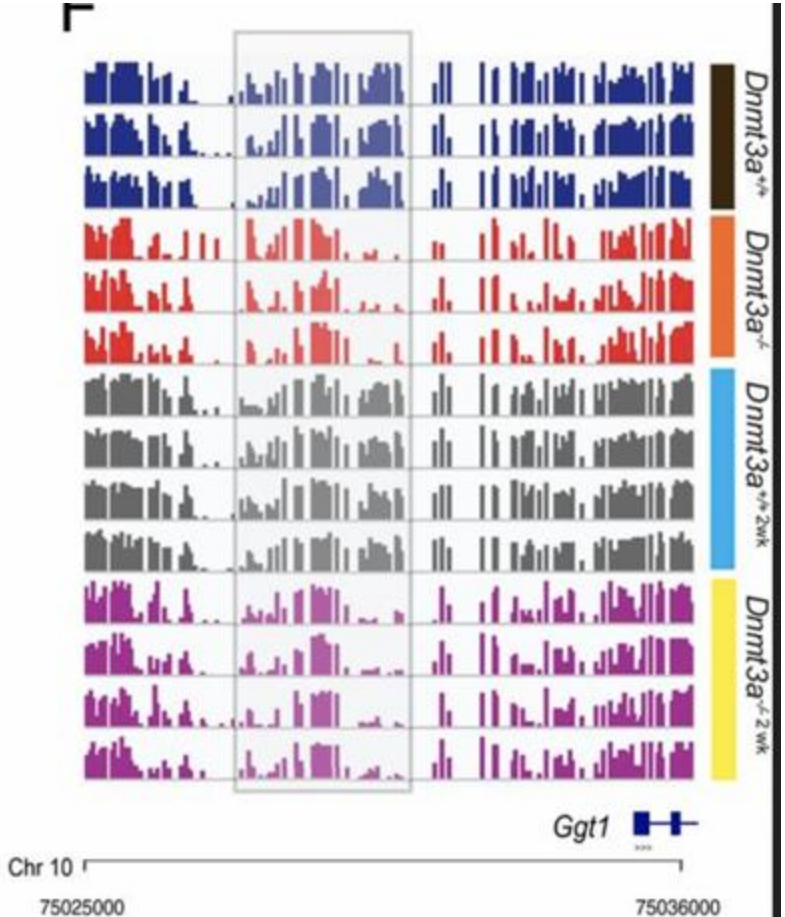
Differentially methylated regions

- Comparing two groups to find changes
- Finding DMRs is a segmentation problem
- We use a tool called metilene

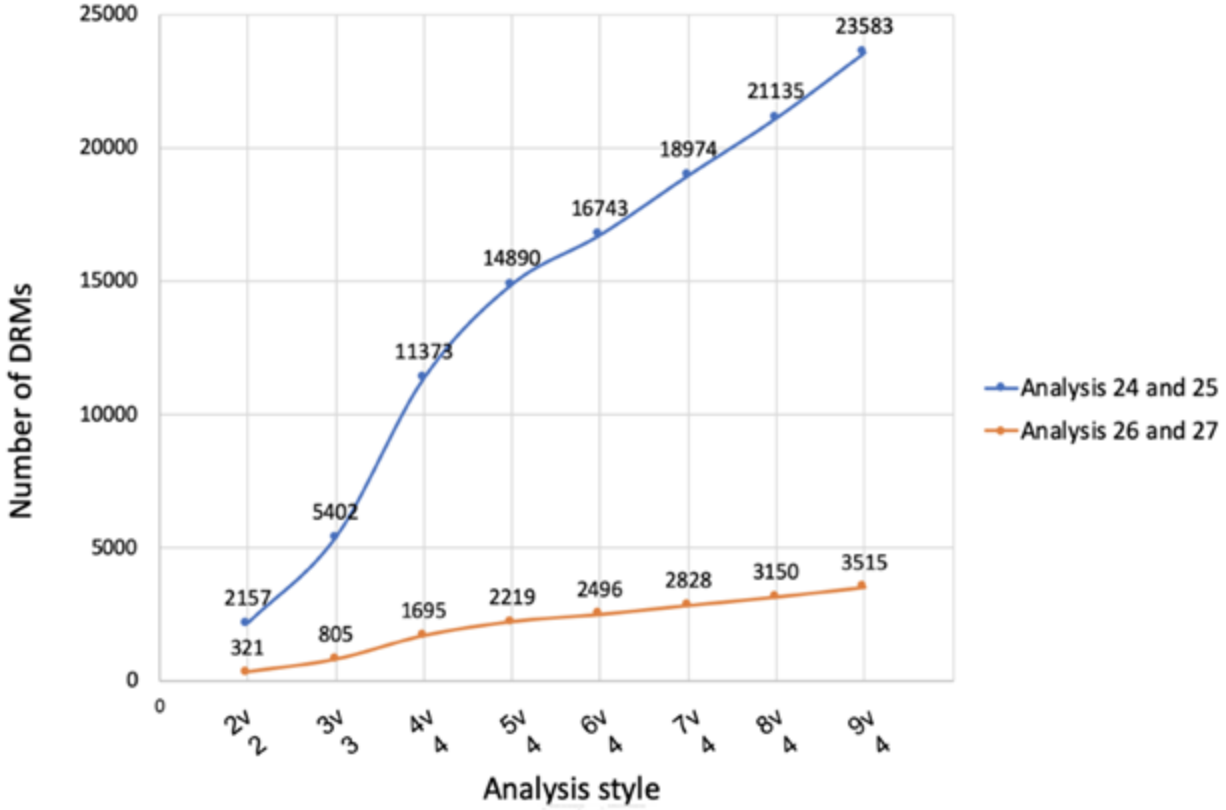


https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4728377/

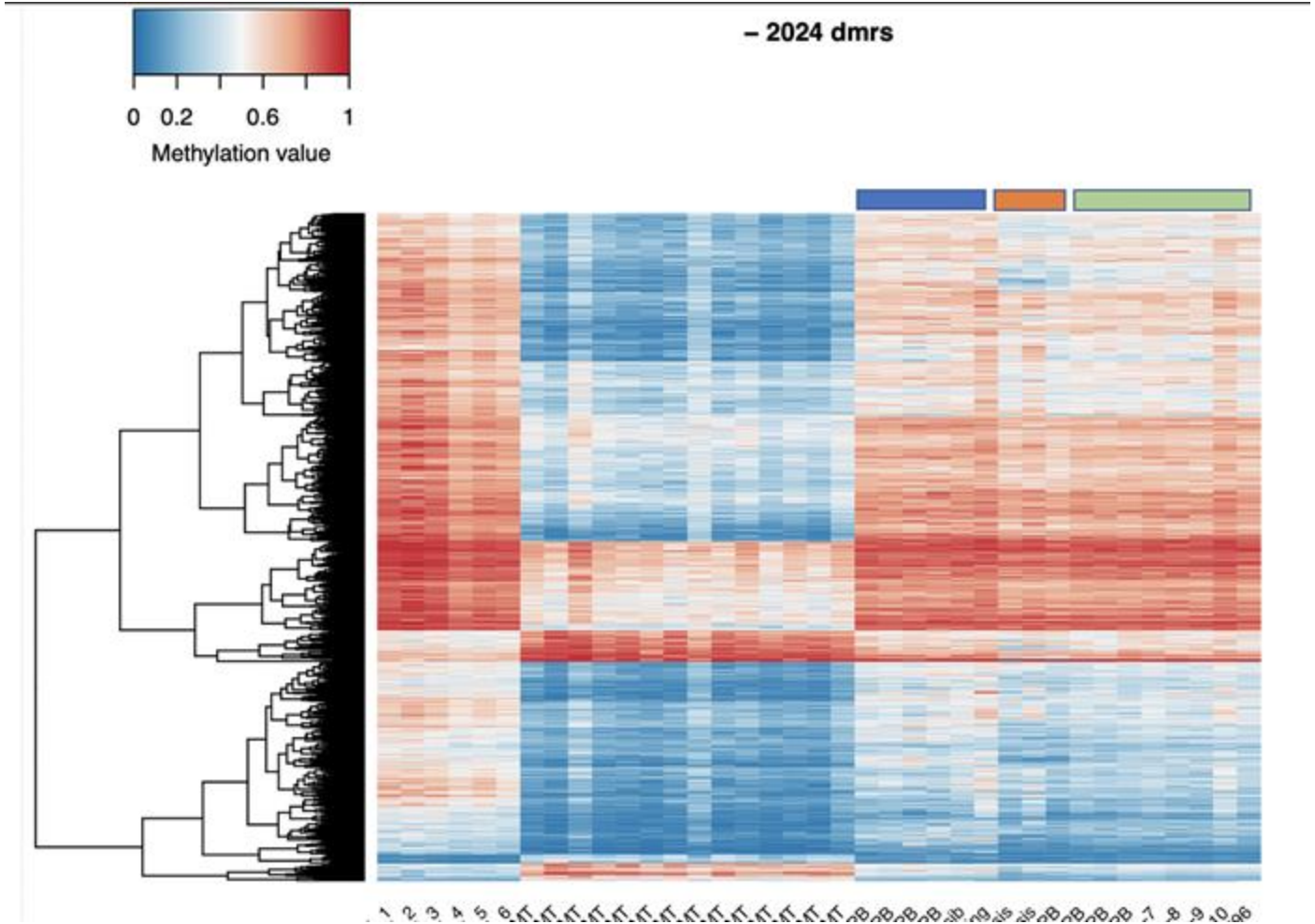
Differentially methylated regions



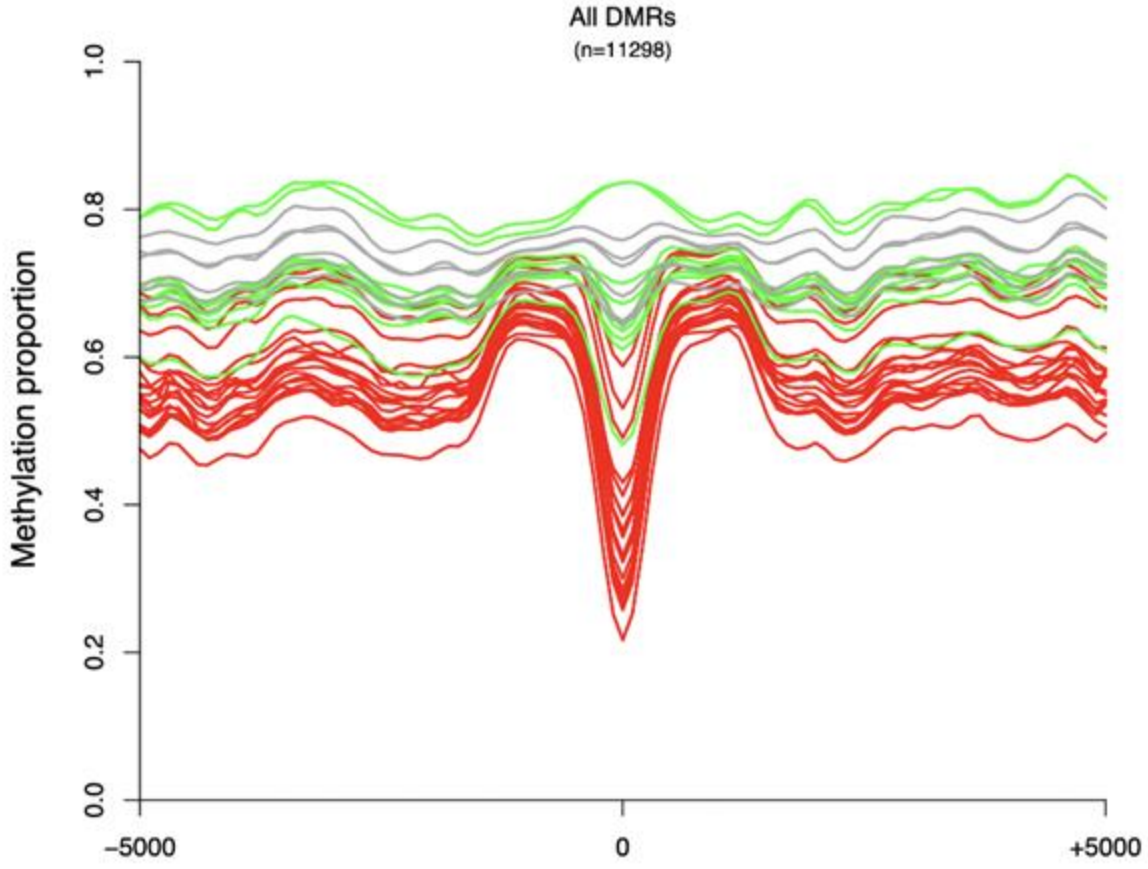
Number of samples matters!



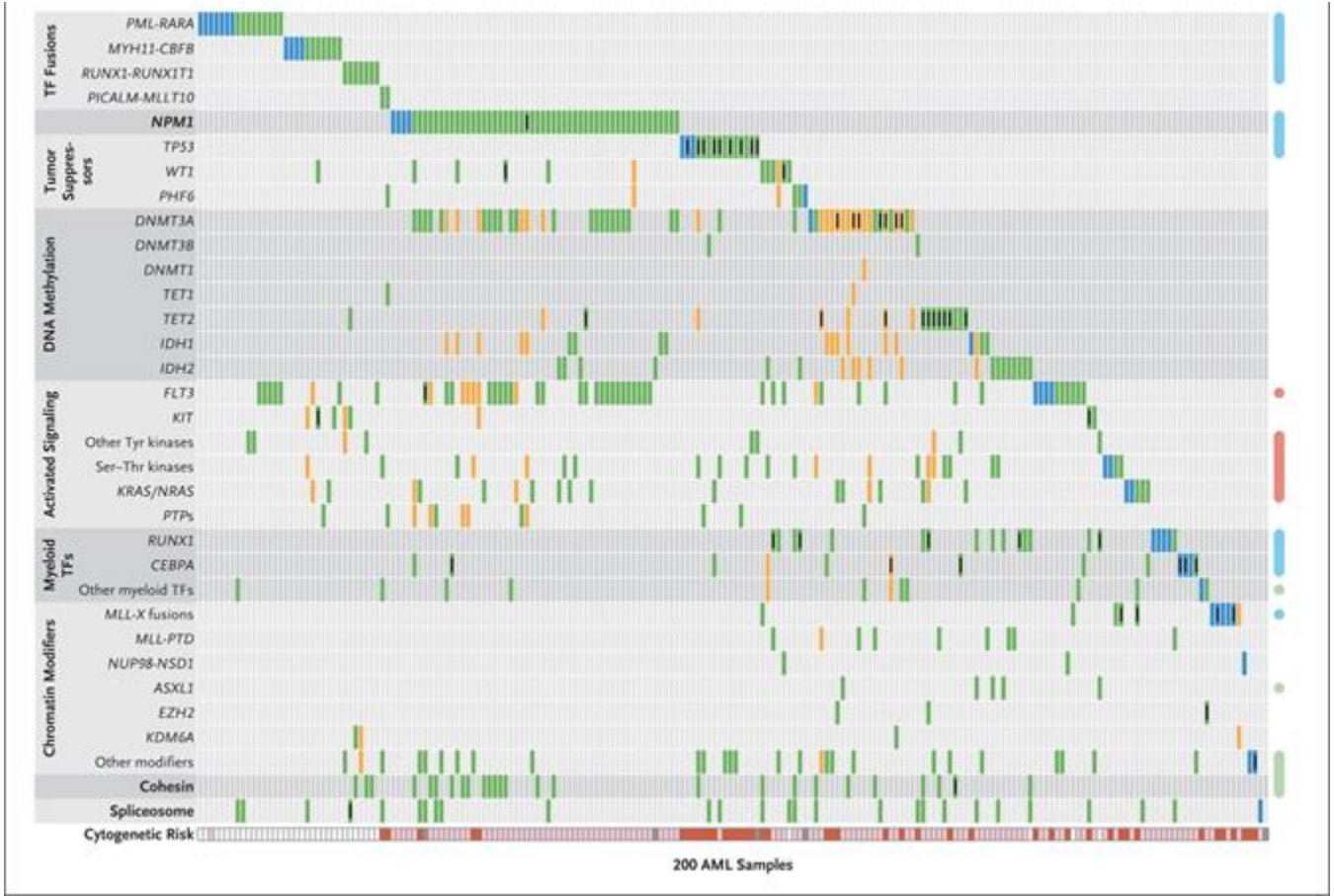
Heatmaps



Canyon Plots

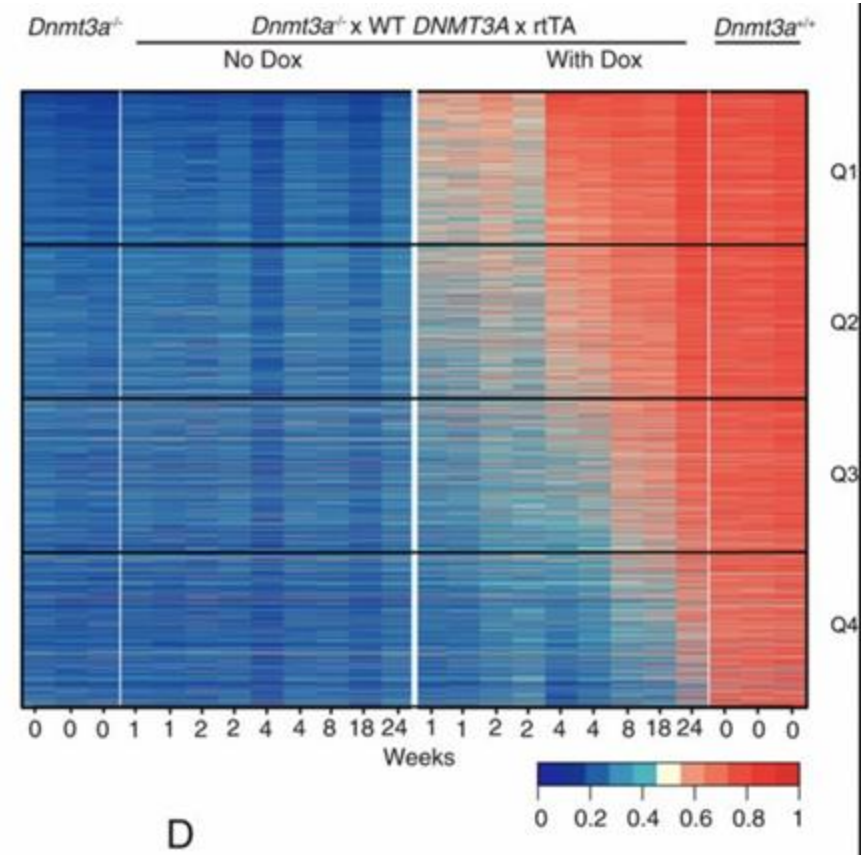


DNMT3A deficiency



DNMT3A deficiency

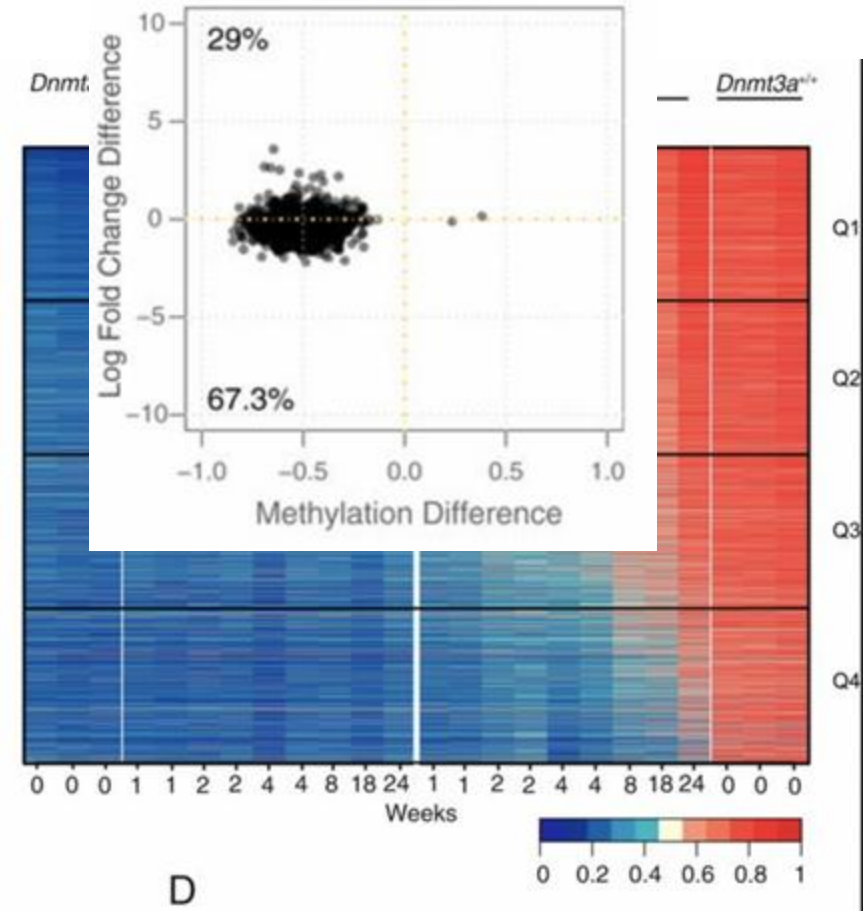
- Mouse models (and human data)
- Looking at context, effects, and reversibility



https://www.pnas.org/content/117/6/3123.full

DNMT3A deficiency

- Mouse models (and human data)
- Looking at context, effects, and reversibility



https://www.pnas.org/content/117/6/3123.full