

# Epigenomics: ChIP-seq and ATAC-seq

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

Some slides adapted from:

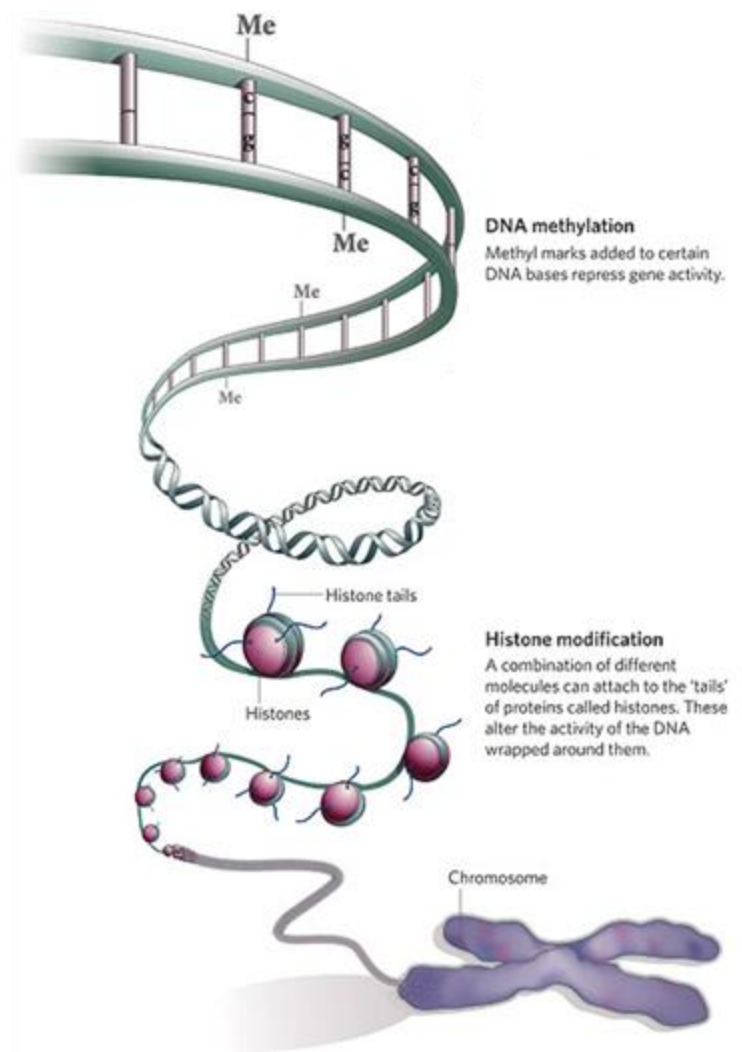
<https://github.com/genome/bfx-workshop>

<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- $\alpha$ -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 NF1 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 Forshaw 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 Transcripts by RNA-Seq," *Nature Methods* 5, no. 7 (July 2008): 621–628, doi:10.1038/nmeth.1225.

**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:** Yohji 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:** Kazuki 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 Rashbash, "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 Hashashony 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/nature09816.

**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/nature09852.

**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/ma.2008810.

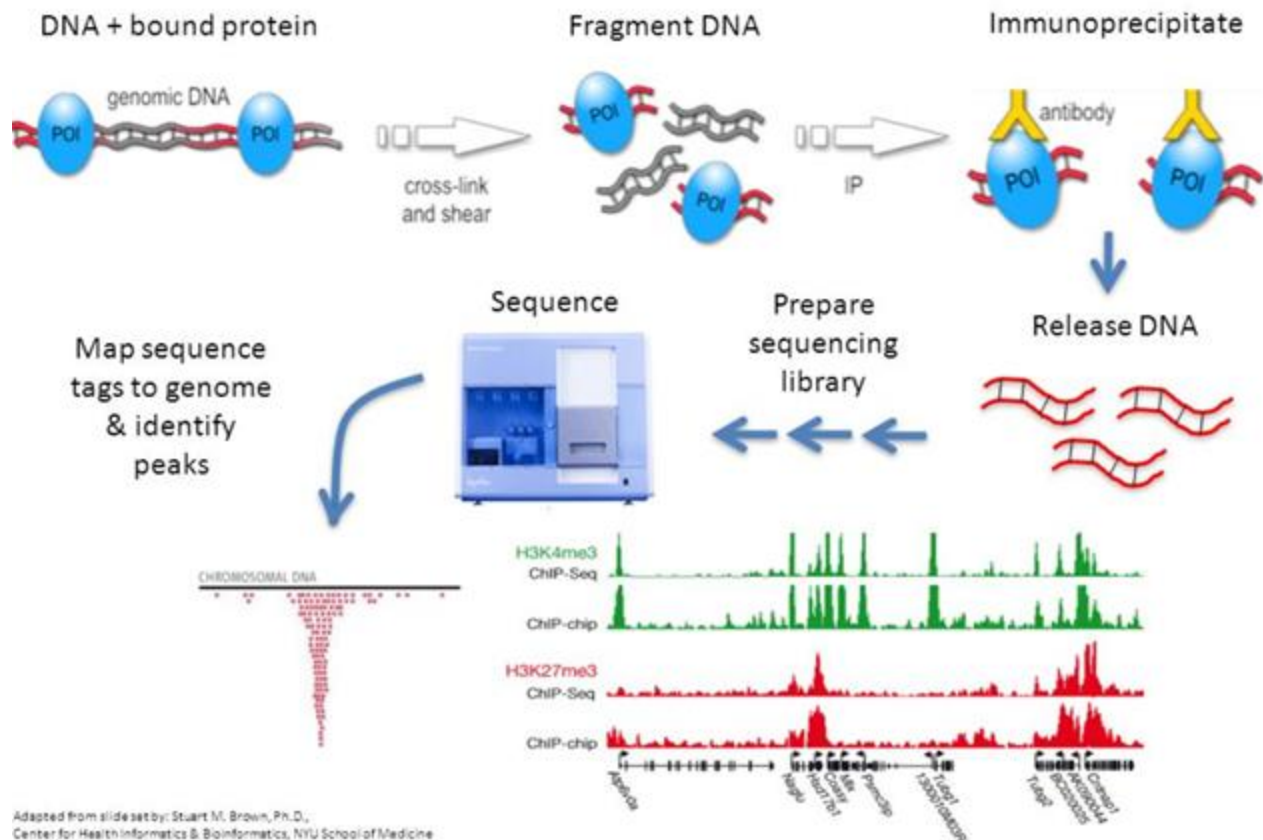
**FRT-Seq:** Lira Mamanova et al., "FRT-Seq: Amplification-free, Strand-specific Transcriptome Sequencing," *Nature Methods* 7, no. 2 (February 2010): 130–132, doi:10.1038/nmeth.1417.

**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.

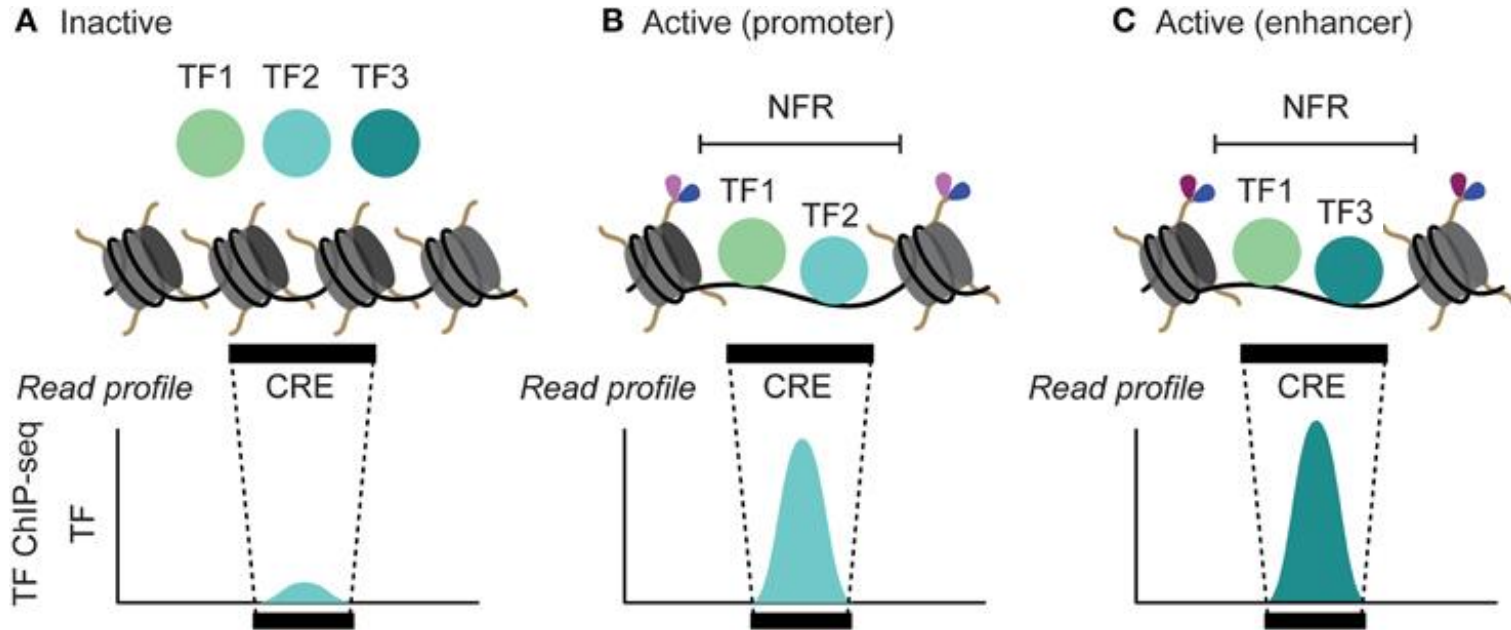
**PRO-Seq:** Hyoung Kwek et al., "Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing," *Science* 339, no. 6122 (February 22, 2013): 950–953, doi:10.1126/science.1229386.

**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

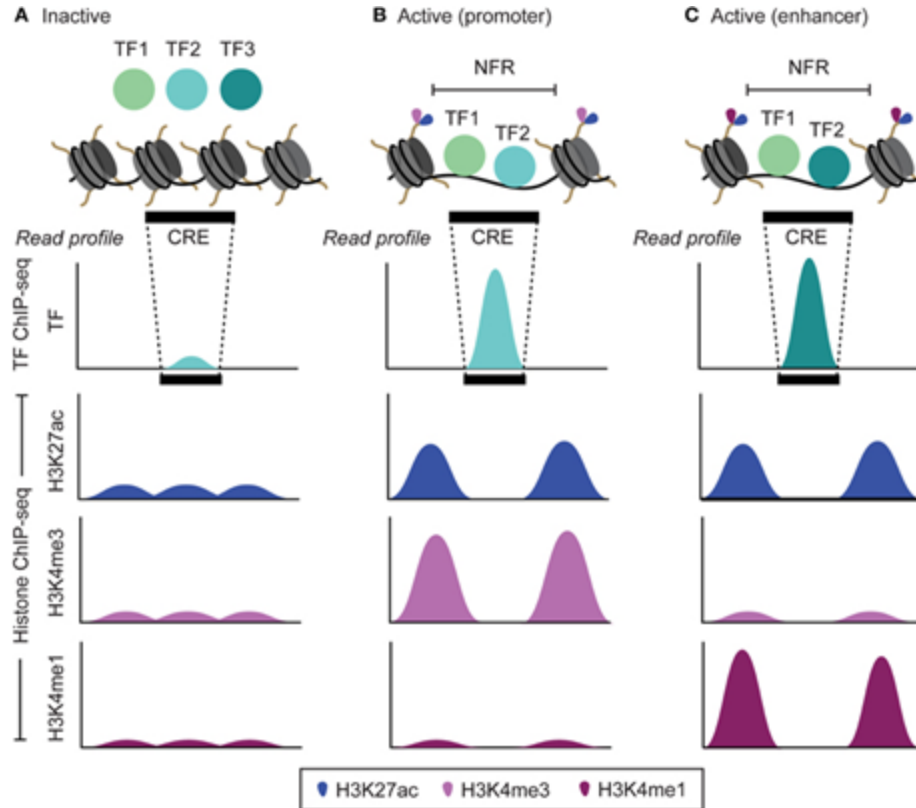


# Mapping transcription-factor binding locations



NFR = nucleosome free region  
CRE = Cis regulatory element

# Mapping histone modifications



H3K4me3:

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

NFR = nucleosome free region  
CRE = Cis regulatory element

# CUT&RUN

Potentially less off-target/background signal

Very similar, from a data perspective

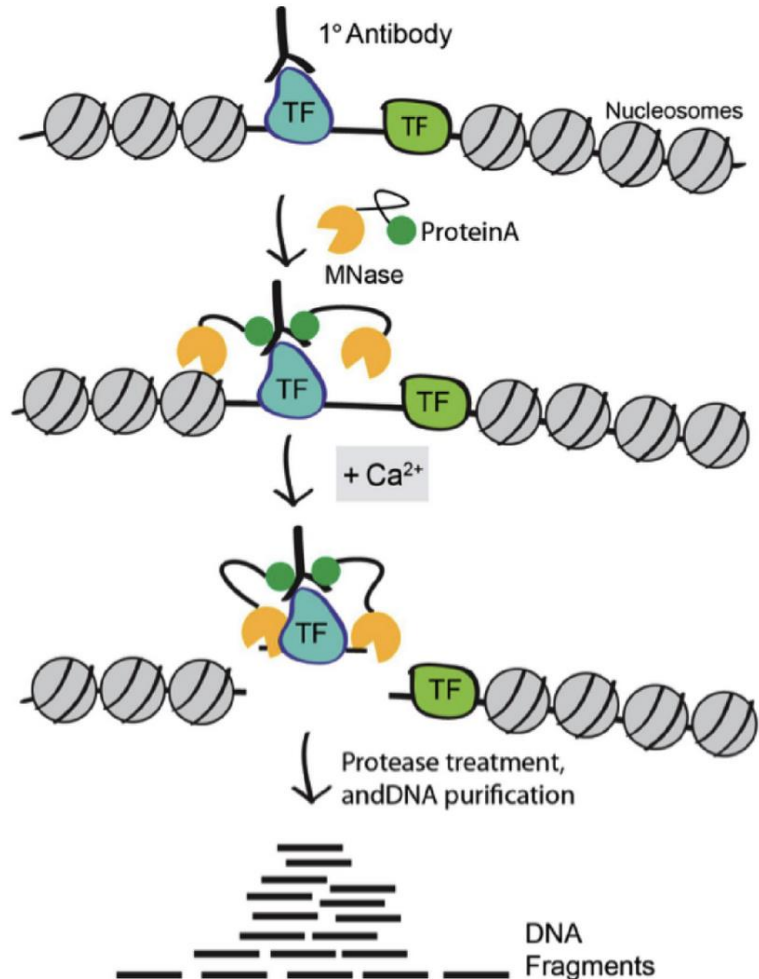
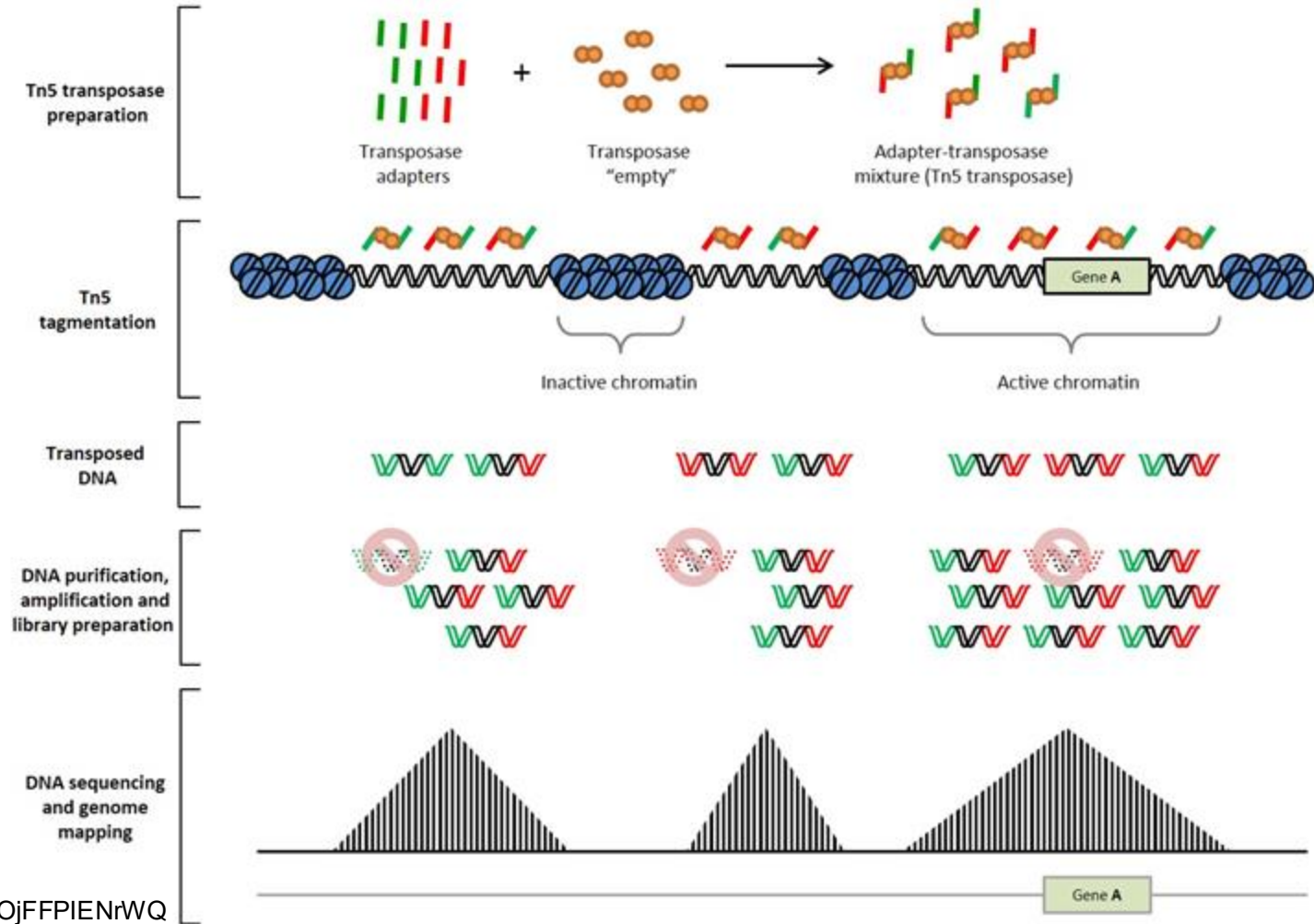


Figure 1. CUT&RUN schematic (see text for details).



# ATAC-seq

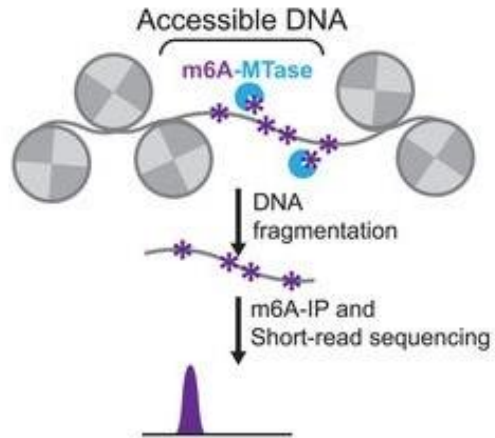
DNA accessibility



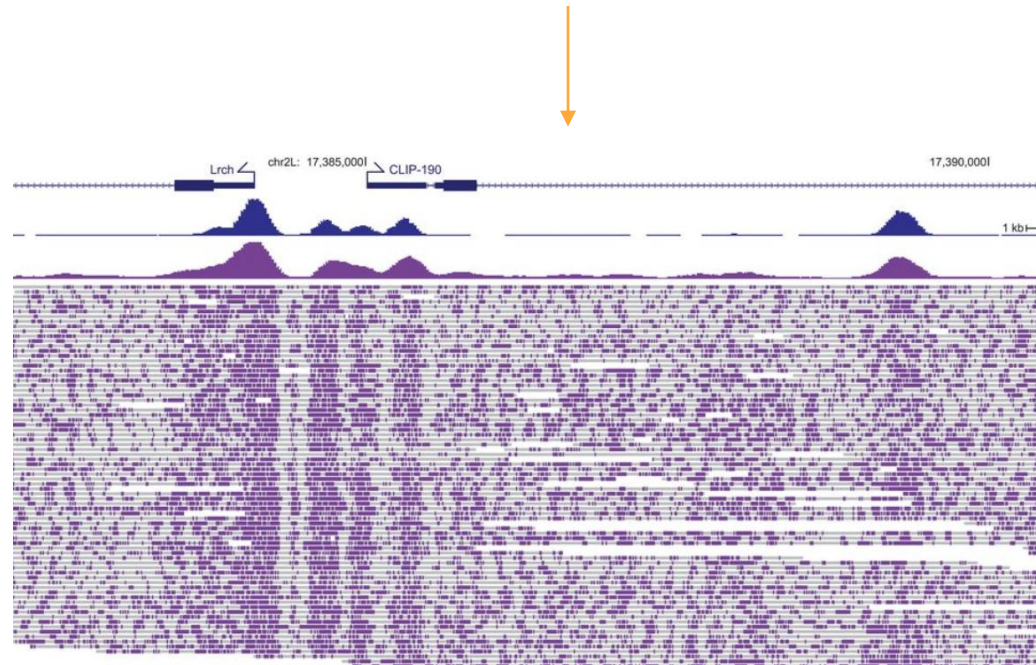


# Fiber-seq

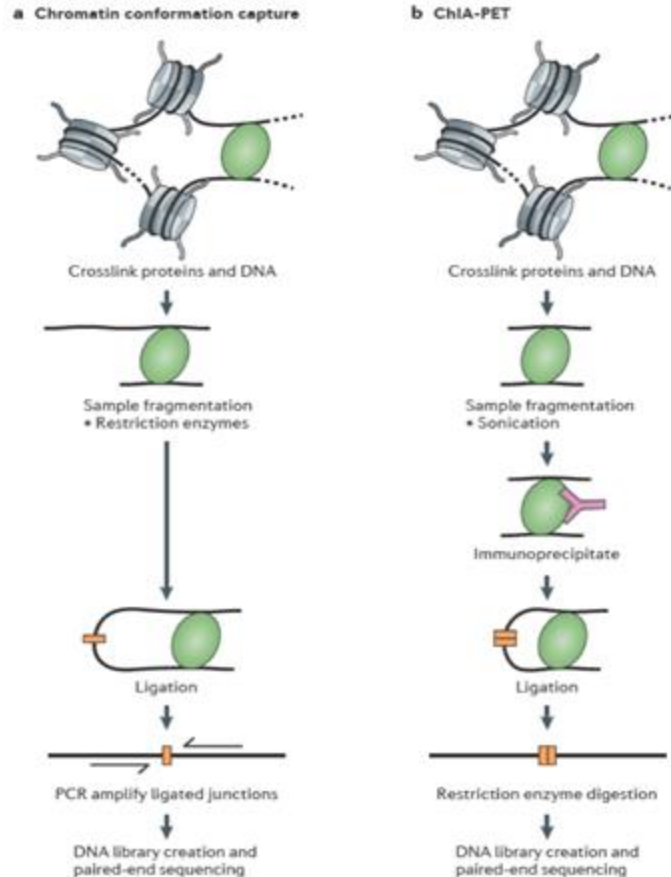
**Methyltransferase-based assay:**



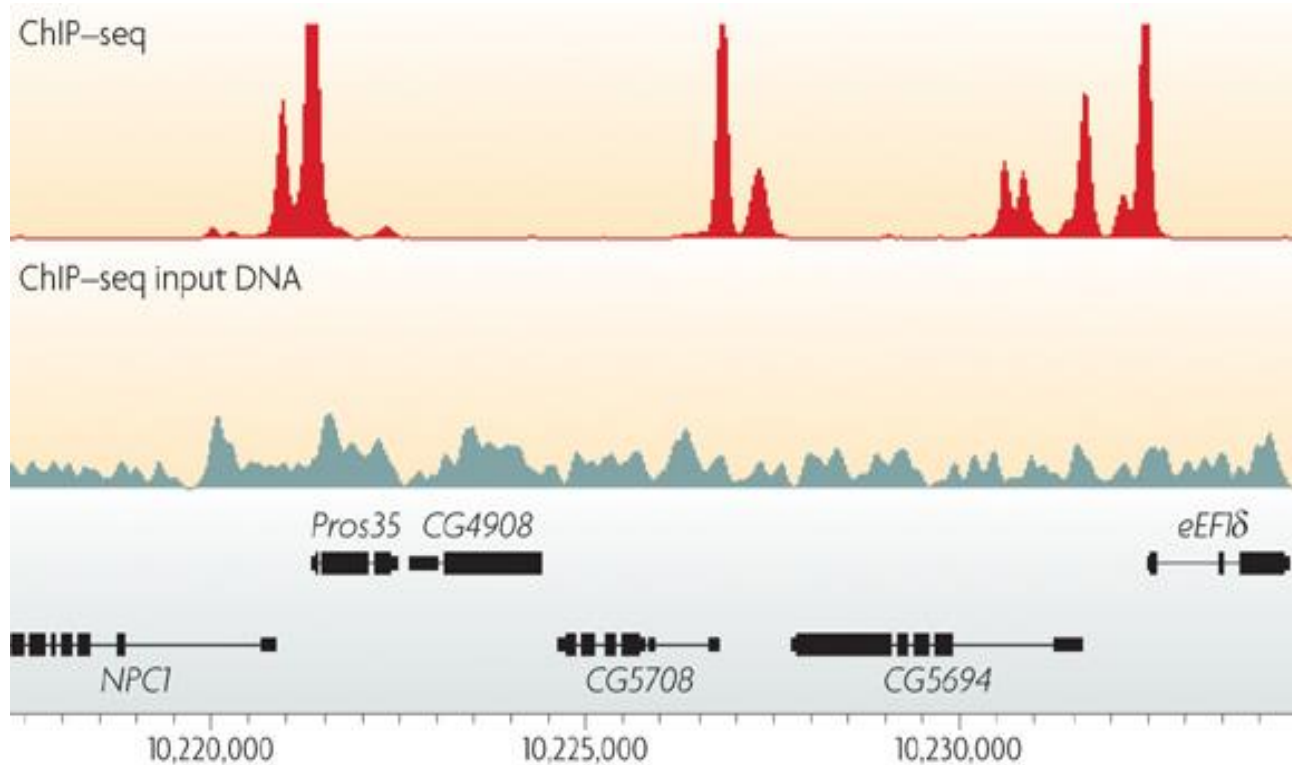
## Long-read sequencing (PacBio/ONT)



# Examining 3-D DNA interactions in the nucleus



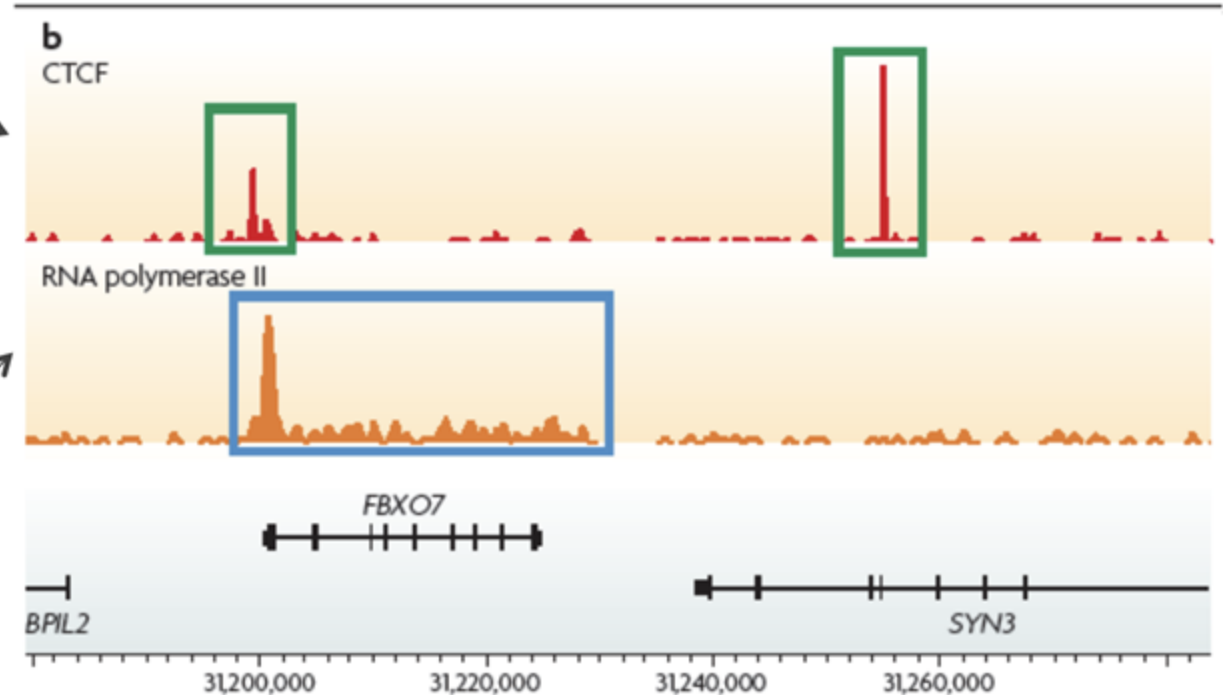
# Peak-calling



Fundamentally a  
signal vs noise problem

# Proteins bind in different ways

Transcription factor – tight, highly-peaked binding region



RNA PolII – enriched at TSS but bound throughout gene body



# 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:

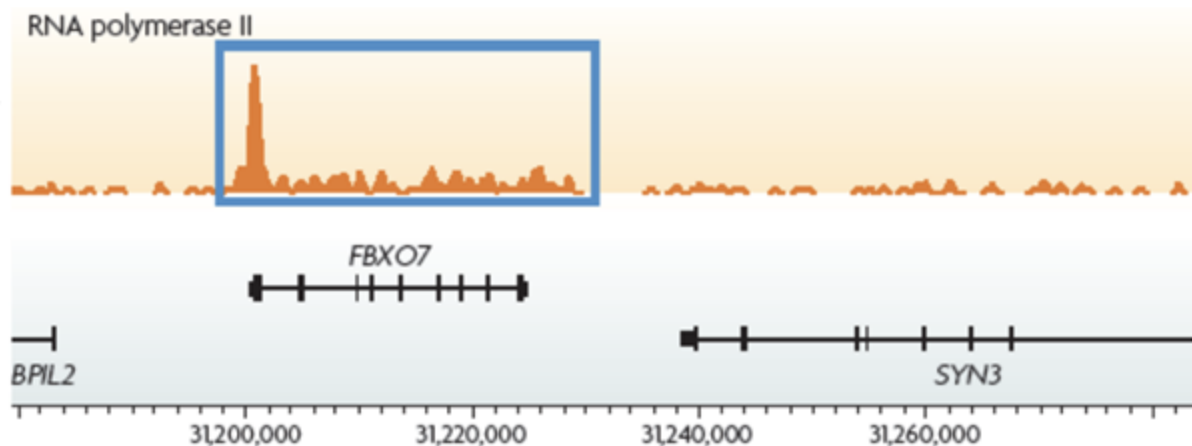
- have on the order of thousands of binding sites,
- 20 million reads may be adequate
- (4 million reads for worm and fly TFs).

# How much sequence coverage do we need?

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

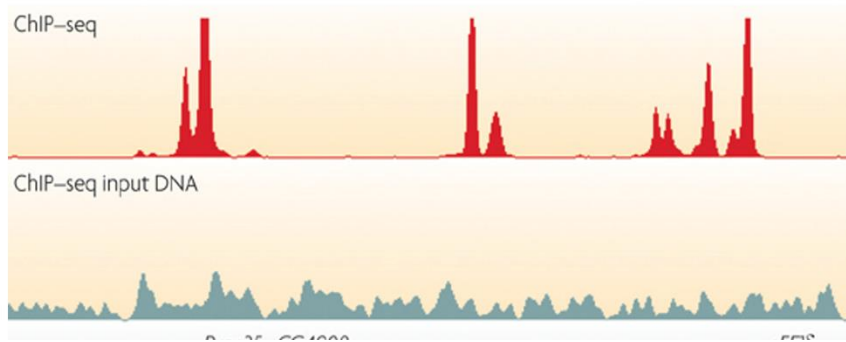
- require more reads, up to 60 million for mammalian ChIP-seq.

RNA PolII – enriched  
at TSS but bound  
throughout gene  
body



# 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

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

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

Example: the "hundred year flood". Thus  $\lambda=1$  (1 catastrophic flood every 100 years)

# Example: the "hundred year flood".

Thus  $\lambda=1$  (1 catastrophic flood every 100 years)

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

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

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

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

$k$	$P(k \text{ overflow floods in 100 years})$
0	0.368
1	0.368
2	0.184
3	0.061
4	0.015
5	0.003
6	0.0005

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

Average number of goals is 2.5

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

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

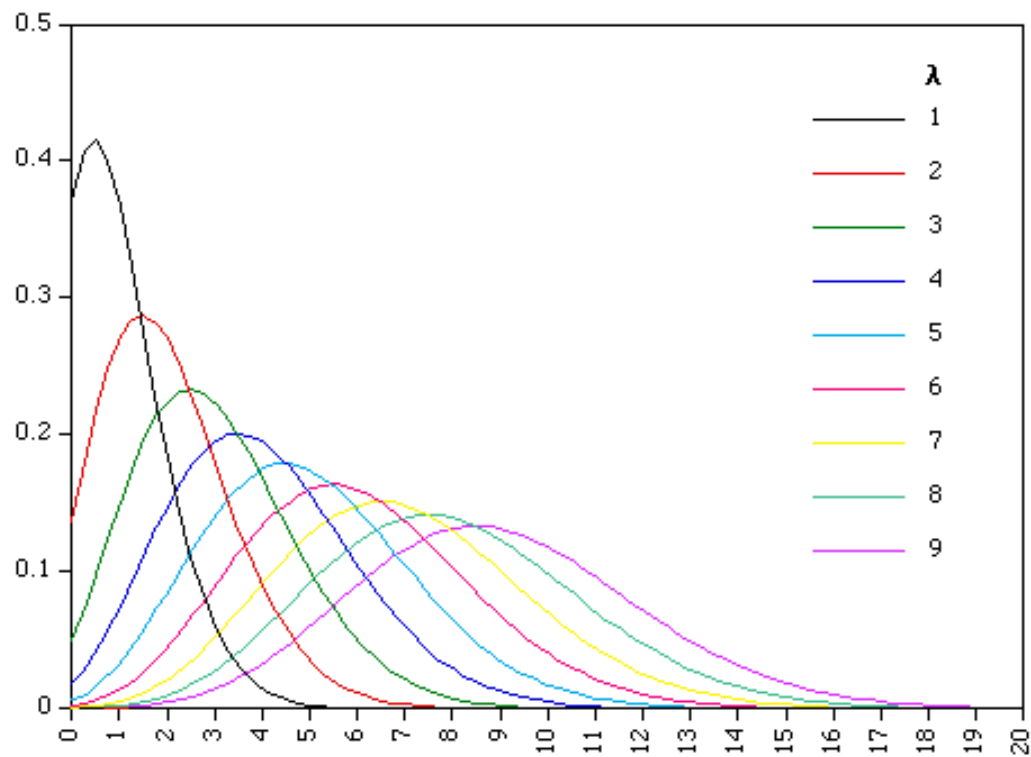
$$P(k = 1 \text{ goal in a match}) = \frac{2.5^1 e^{-2.5}}{1!} = \frac{2.5 e^{-2.5}}{1} = 0.205$$

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

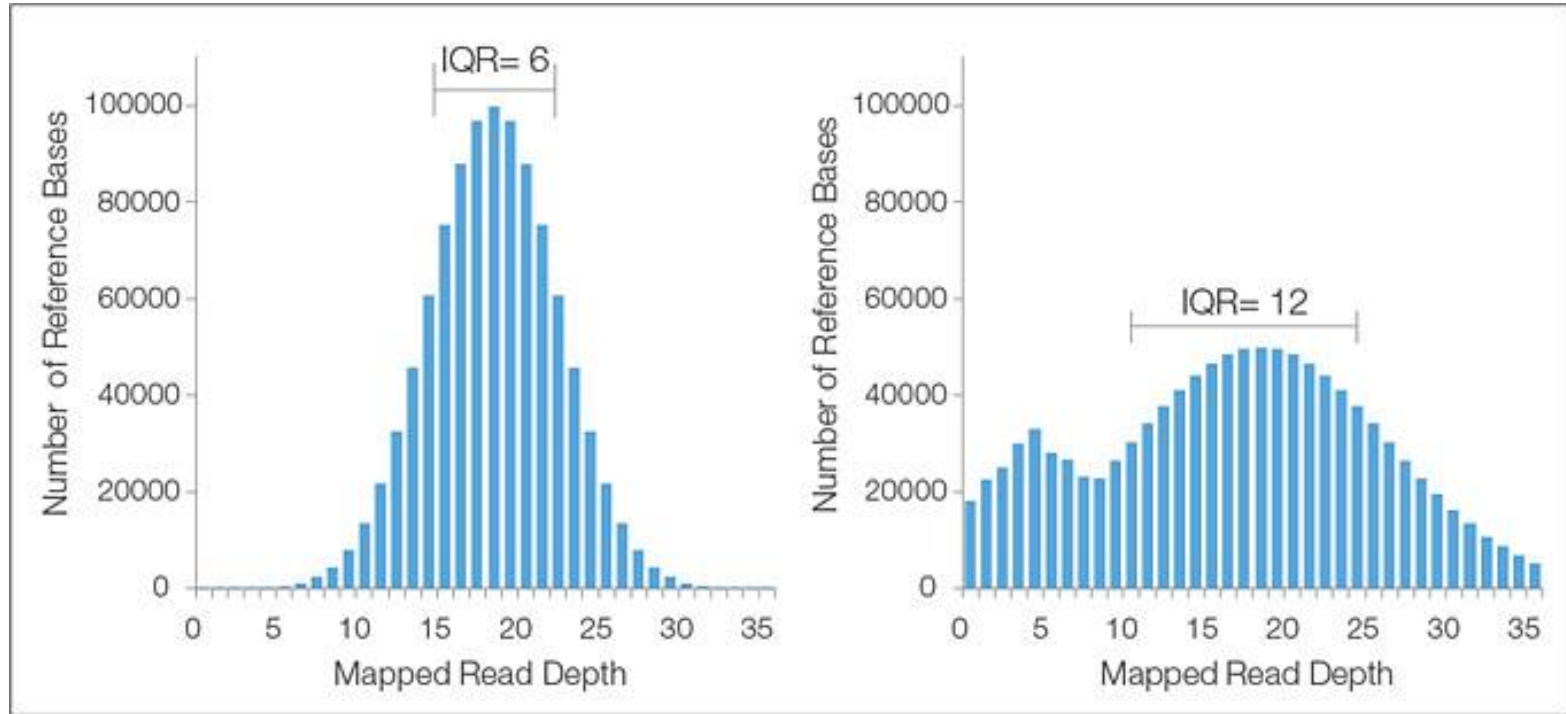
Note that  $\lambda$  does not have to be a countable integer.

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

# Poisson distribution with different values of $\lambda$



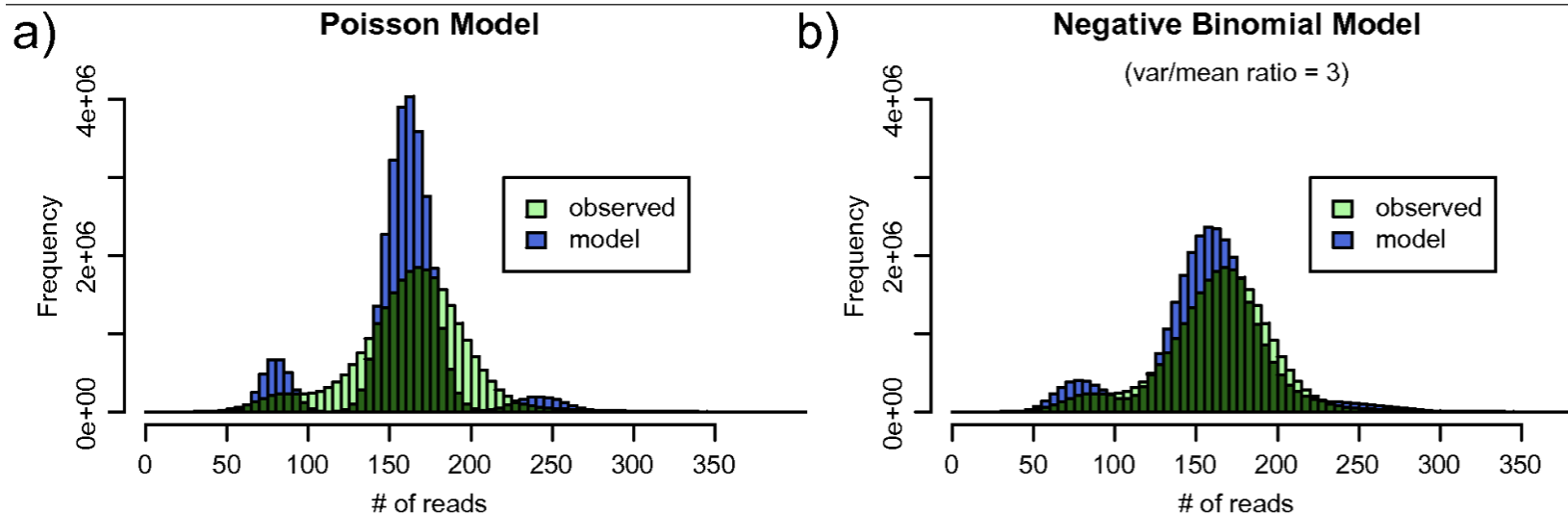
Ideally, sequencing coverage will follow a Poisson distribution. But...



Poisson

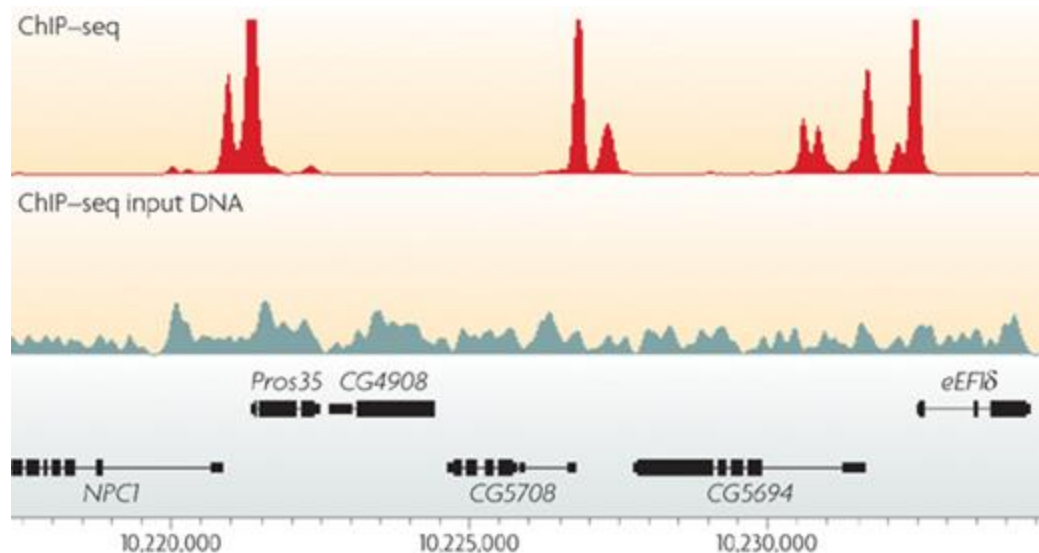
Not Poisson.  
Overly "dispersed"

# Negative binomial fits sequencing coverage data much better





# Comparative ChIP-seq: scaling and normalizing

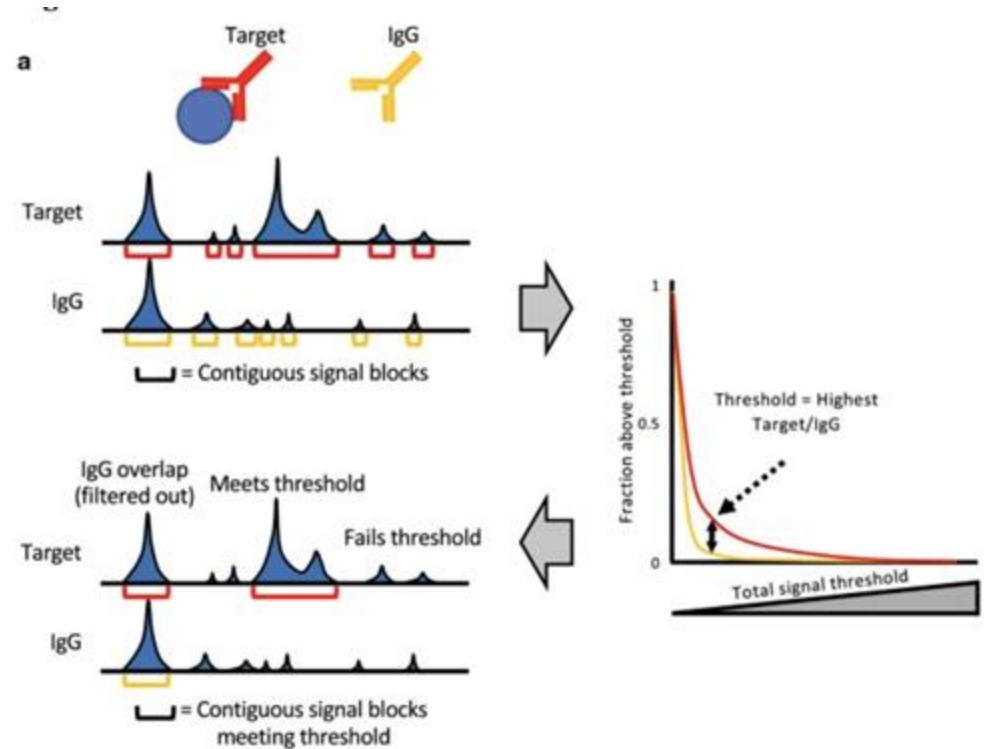


Sequencing depths from TF ChIP should be compared to control (input DNA) to get a sense of the noise.

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.**

# Peak-calling

MACS2, HOMER, SEACR, etc



# ChIP-seq peak callers

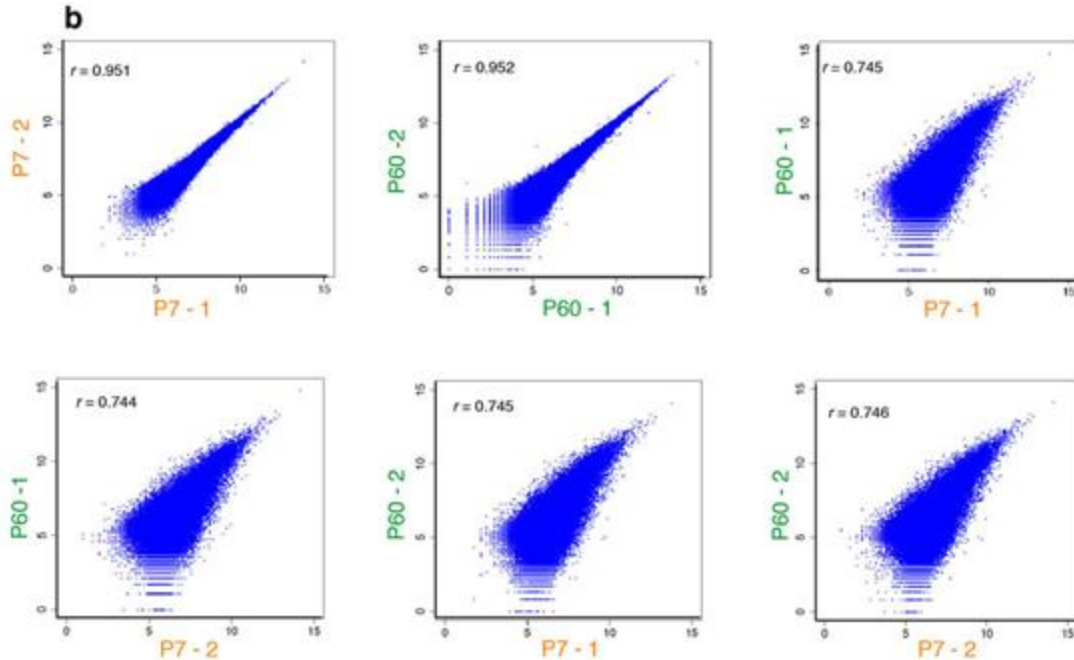
Table S1. Examples of peak callers employed in ChIP-seq.

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

<sup>†</sup> Only for post-processing.

MACS is probably the most widely used

# Replicates, replicates, replicates

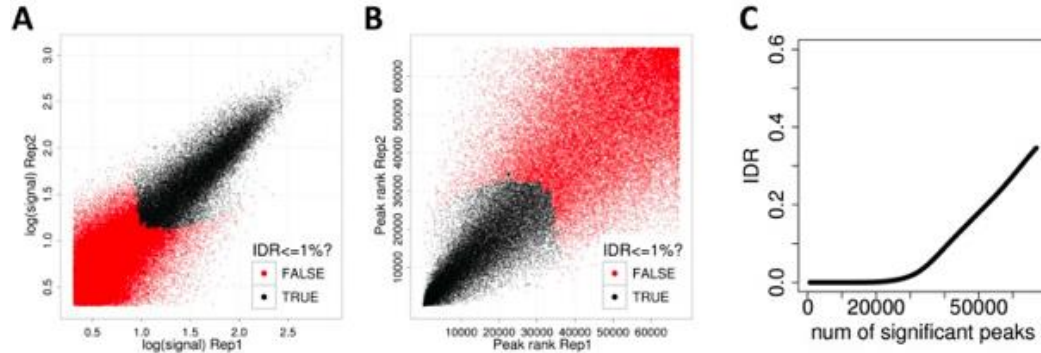


Pearson Correlation is one metric indicative of overall reproducibility

(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.

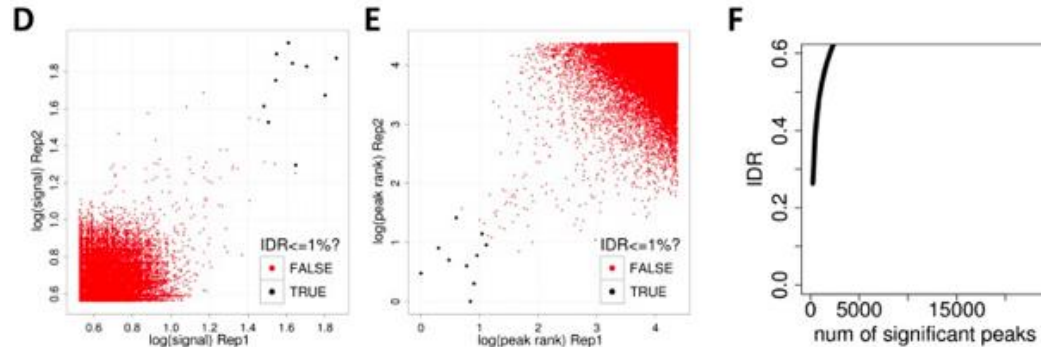
# Irreproducibility Discovery Rate (IDR)

**RAD21 Replicates (high reproducibility)**



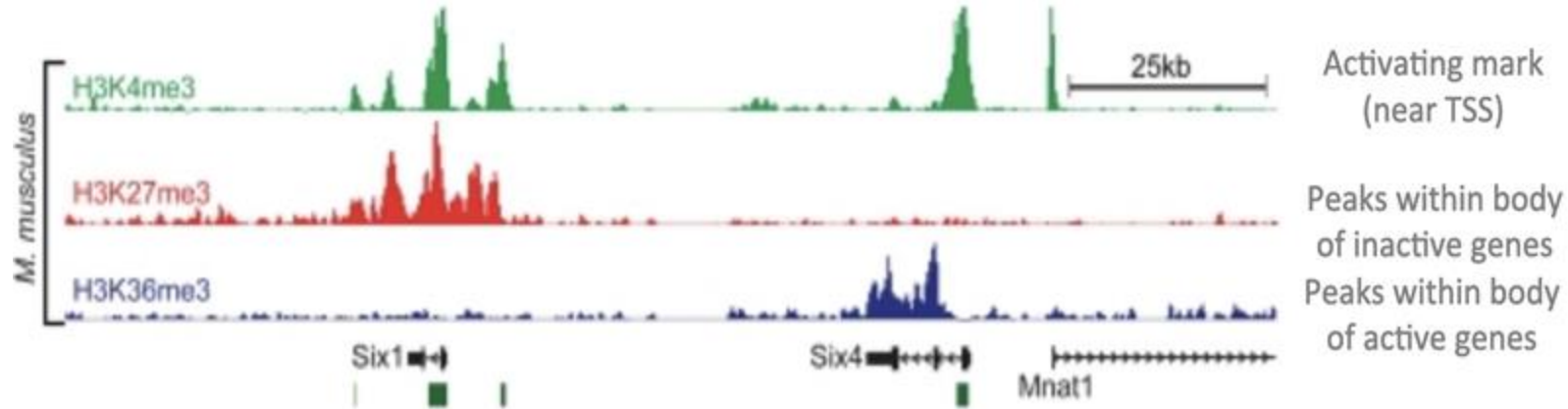
Peaks that show up consistently among replicates are more likely to be real!

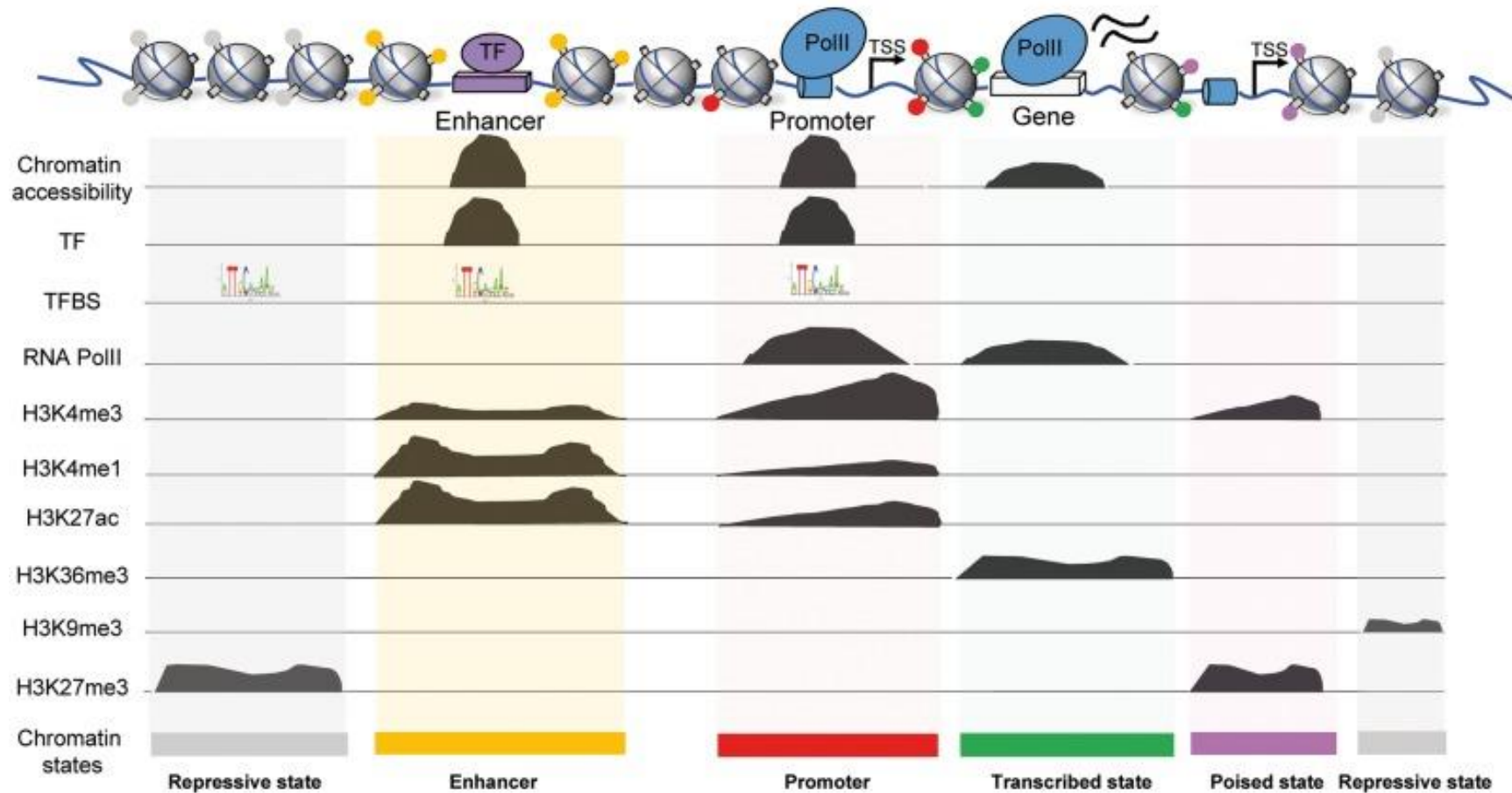
**SPT20 Replicates (low reproducibility)**



Fantastic resource for learning ChIP-seq analysis  
<https://github.com/hbctraining/Intro-to-ChIPseq>

# Interpretation

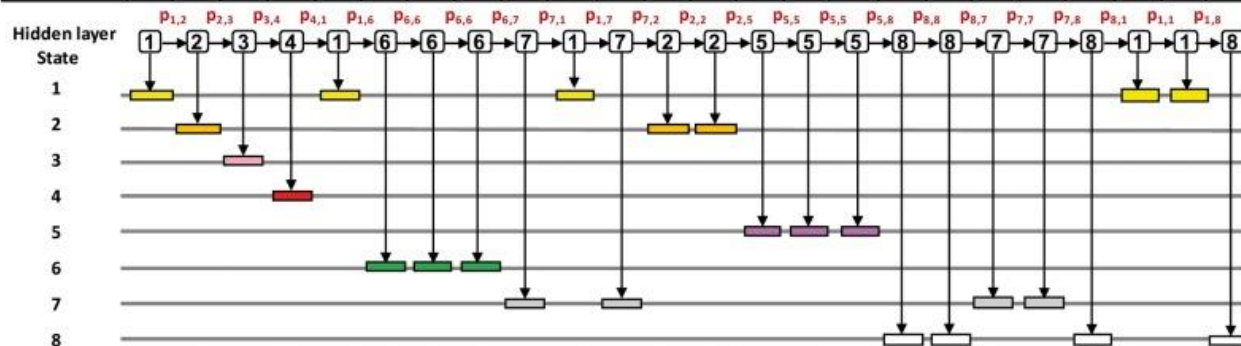








H3K4me1	1	1	1	1	1	0	0	0	0	1	0	1	1	1	1	1	0	0	0	0	0	1	1	0
H3K4me3	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0
H3K27ac	0	1	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
H3K36me3	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H3K9me3	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0
H3K27me3	0	0	0	0	0	0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

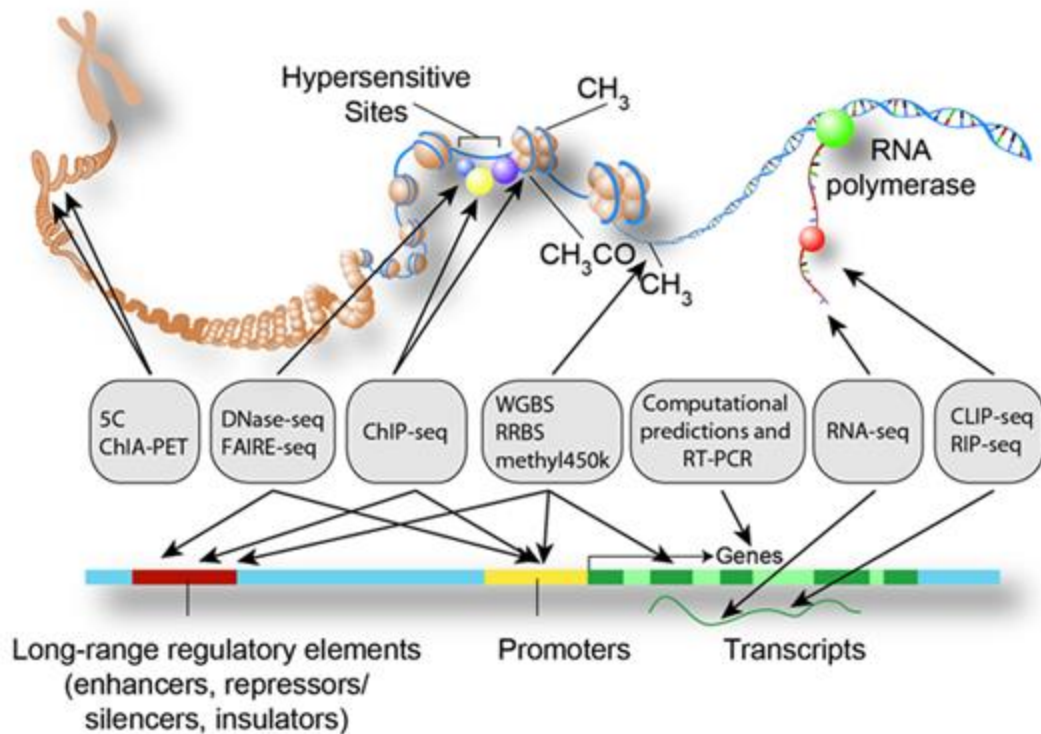


Chromatin state	State annotation
1	Poised enhancer
2	Active enhancer
3	Weak promoter
4	Active promoter
5	Poised promoter
6	Transcriptional elongation
7	Repressive state
8	Low signal state

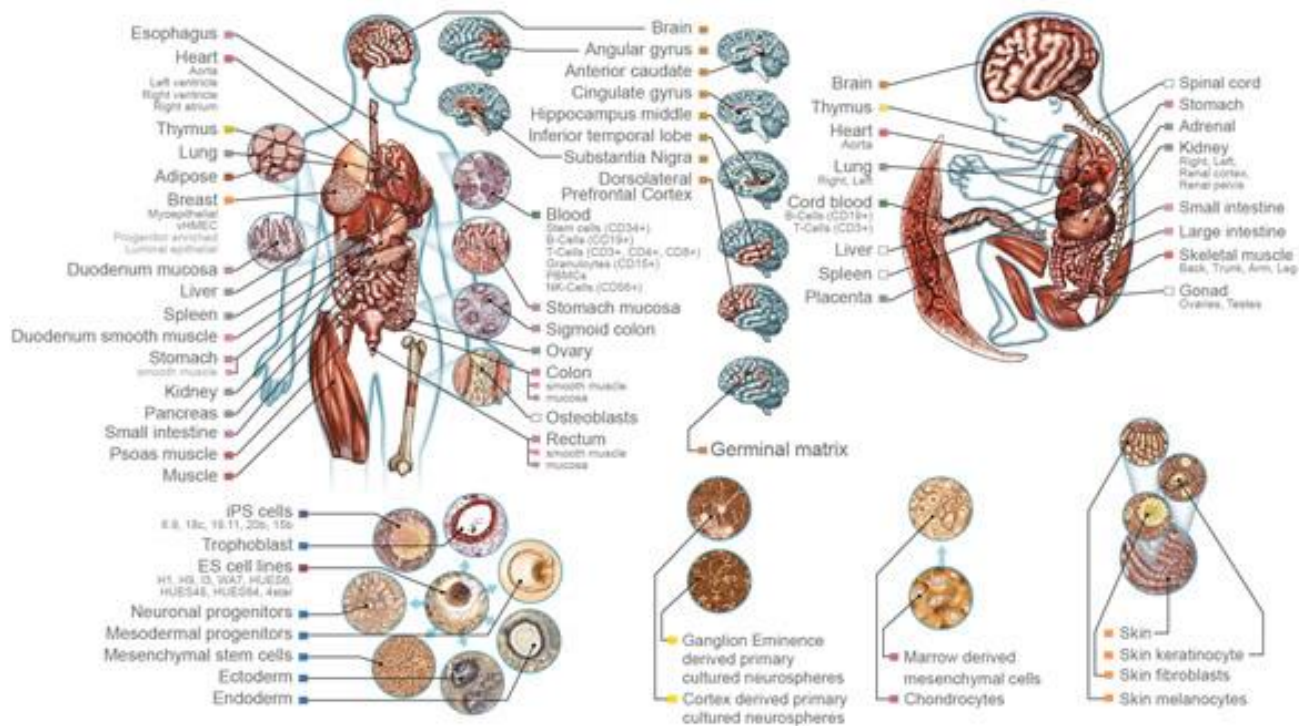
## Genome Segmentation with ChromHMM

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

# The Encyclopedia of DNA Elements Project



# Roadmap Epigenomics Project



# ChIP-seq Exercise

# Epigenomics: Bisulfite sequencing

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

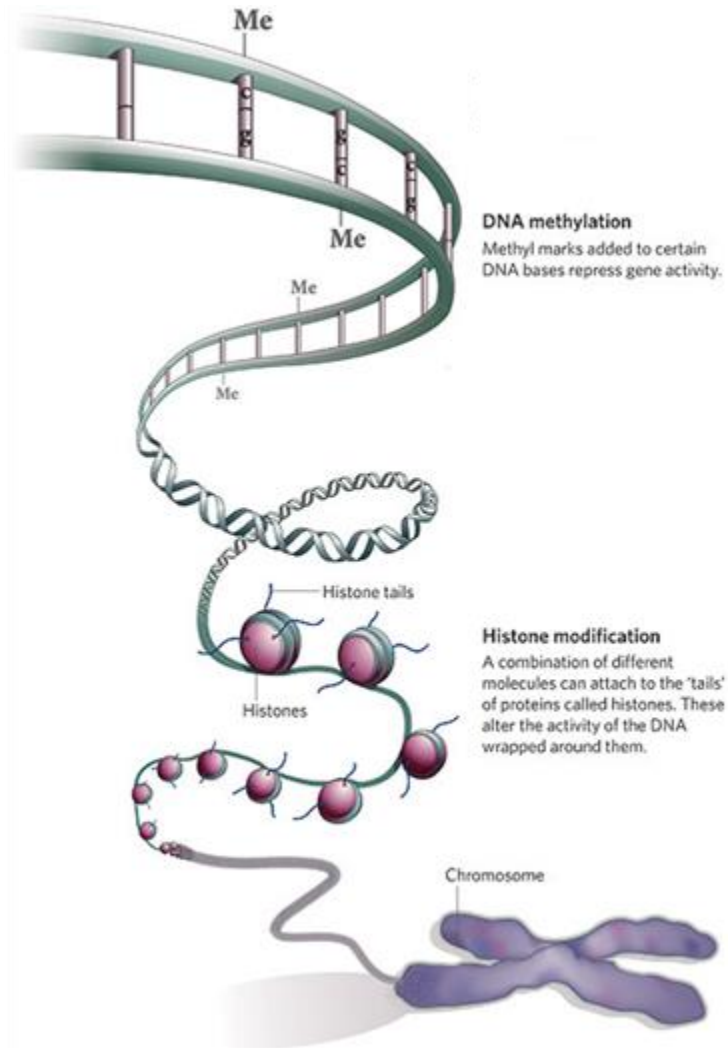
Some slides adapted from:

<https://github.com/genome/bfx-workshop>

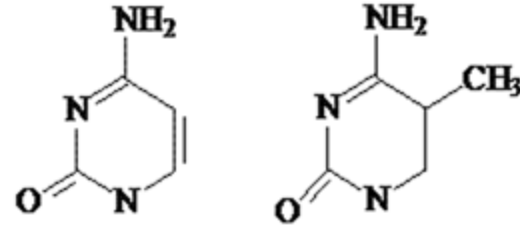
<https://github.com/quinlan-lab/applied-computational-genomics>



# Epigenetics



# DNA Methylation



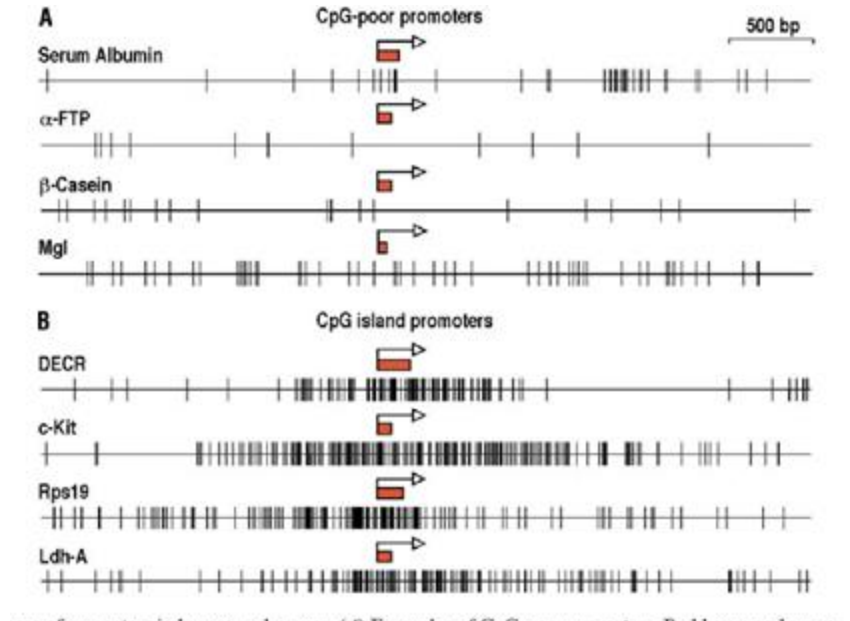
- Mostly happens at CpGs
- About 25 million CpGs in human genome



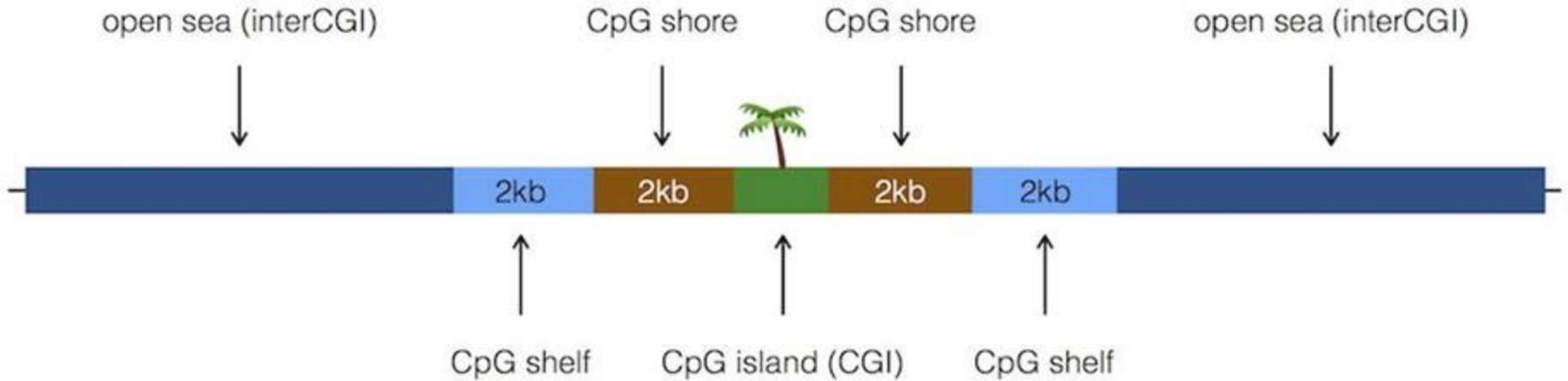


# DNA Methylation

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

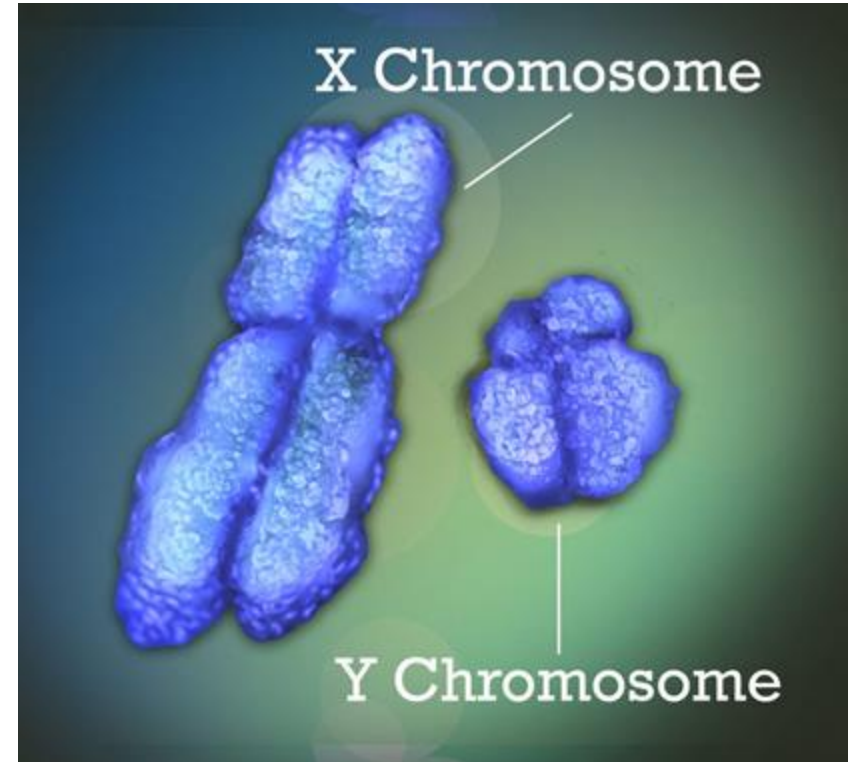


# Islands, shores, and shelves



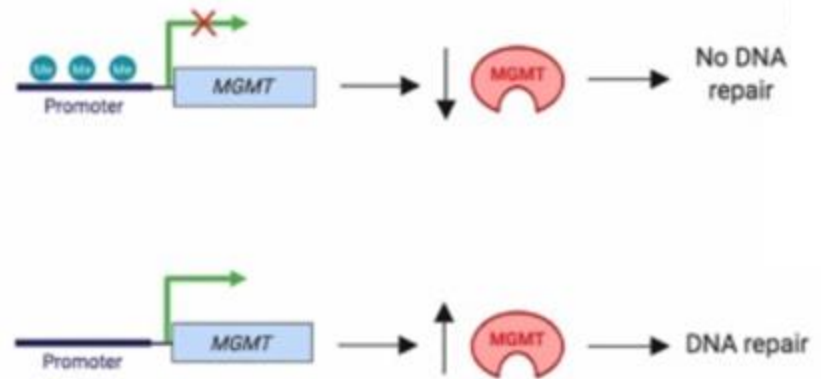
# What does DNA methylation do?

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



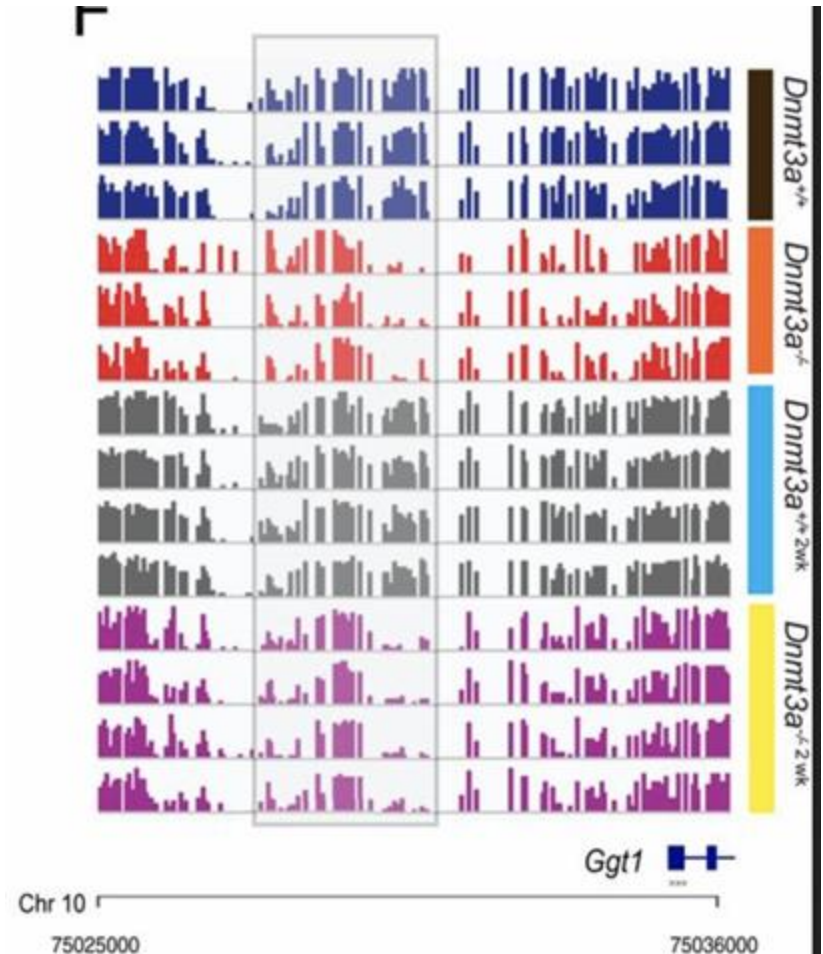
# MGMT and Temozolomide

- 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!



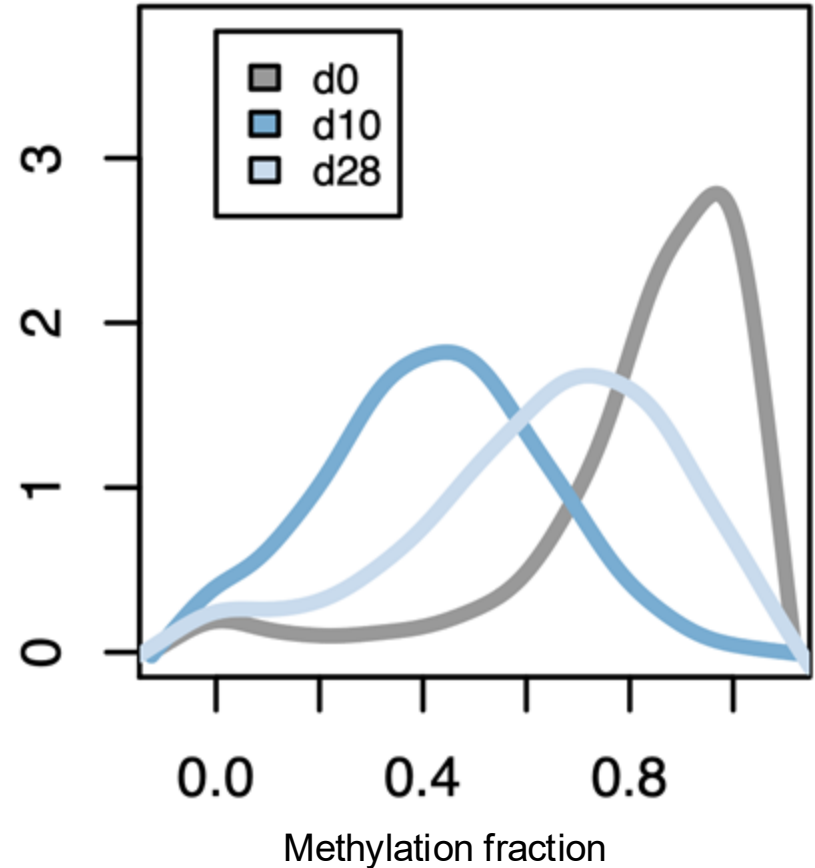
# Methylation Patterns

- Methyltransferases that act locally

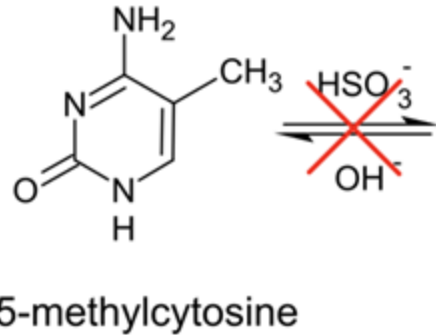
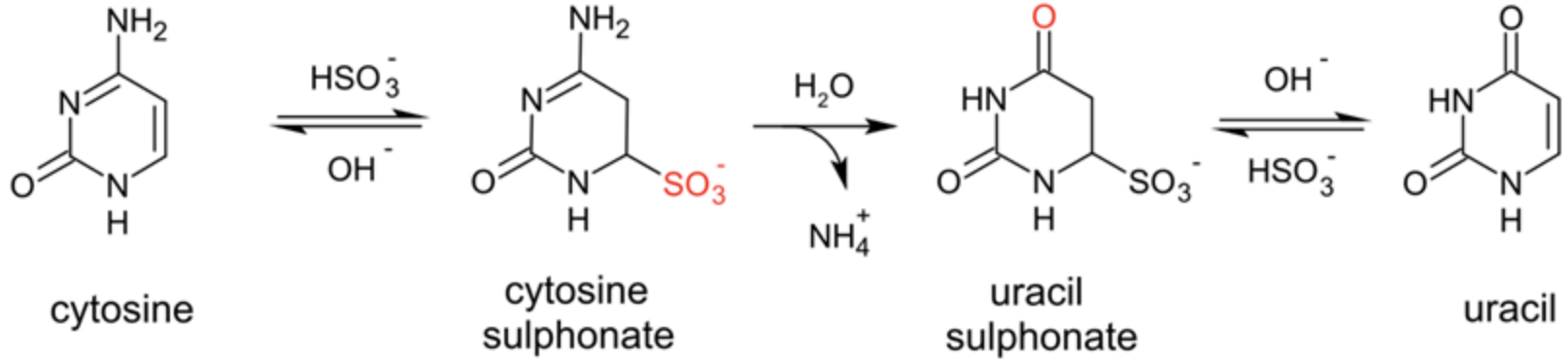


# Methylation Patterns

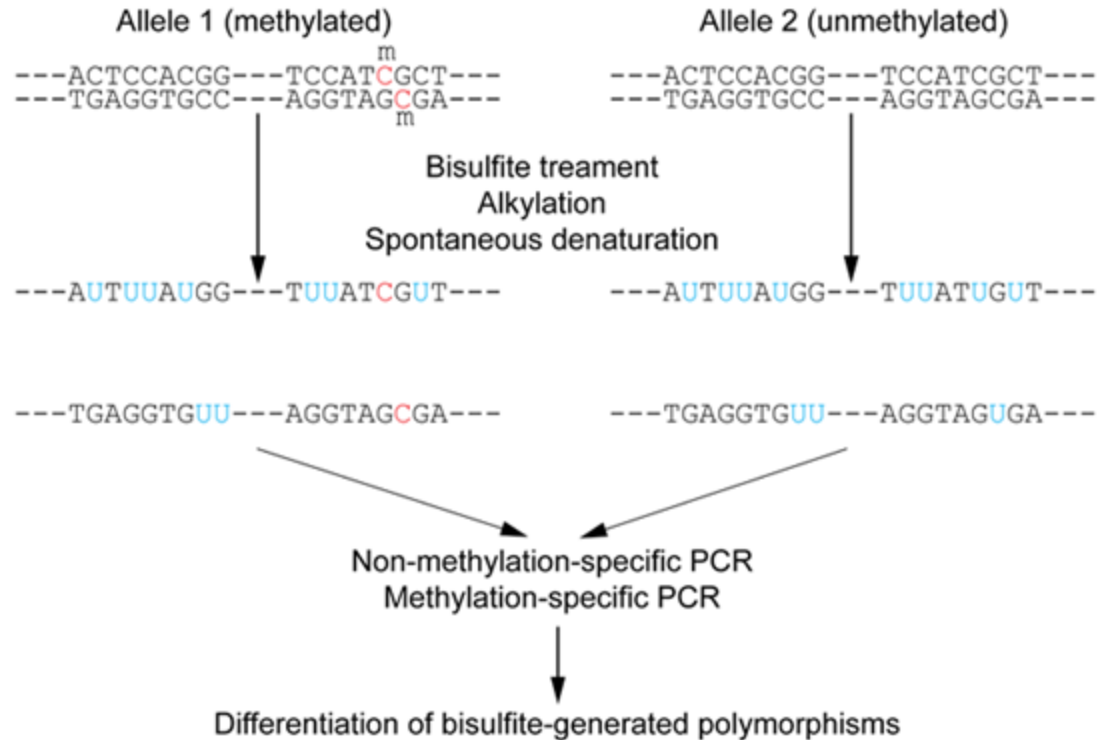
- Methyltransferases that act locally
- Other alterations (or treatments) that act globally



# Bisulfite sequencing

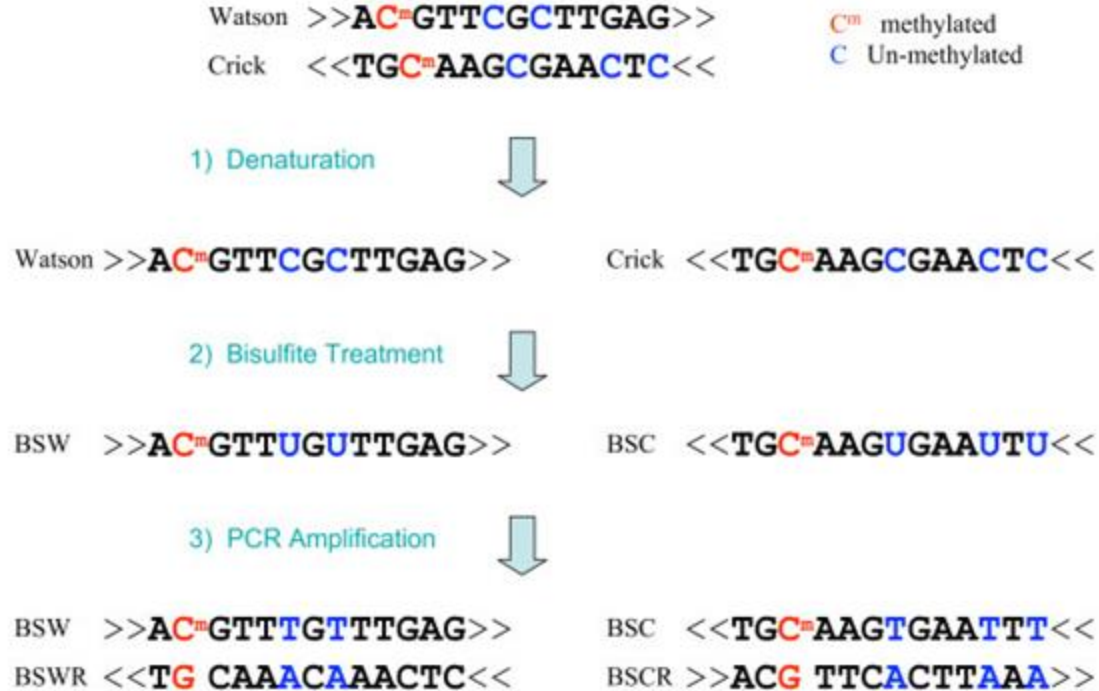


# Bisulfite sequencing





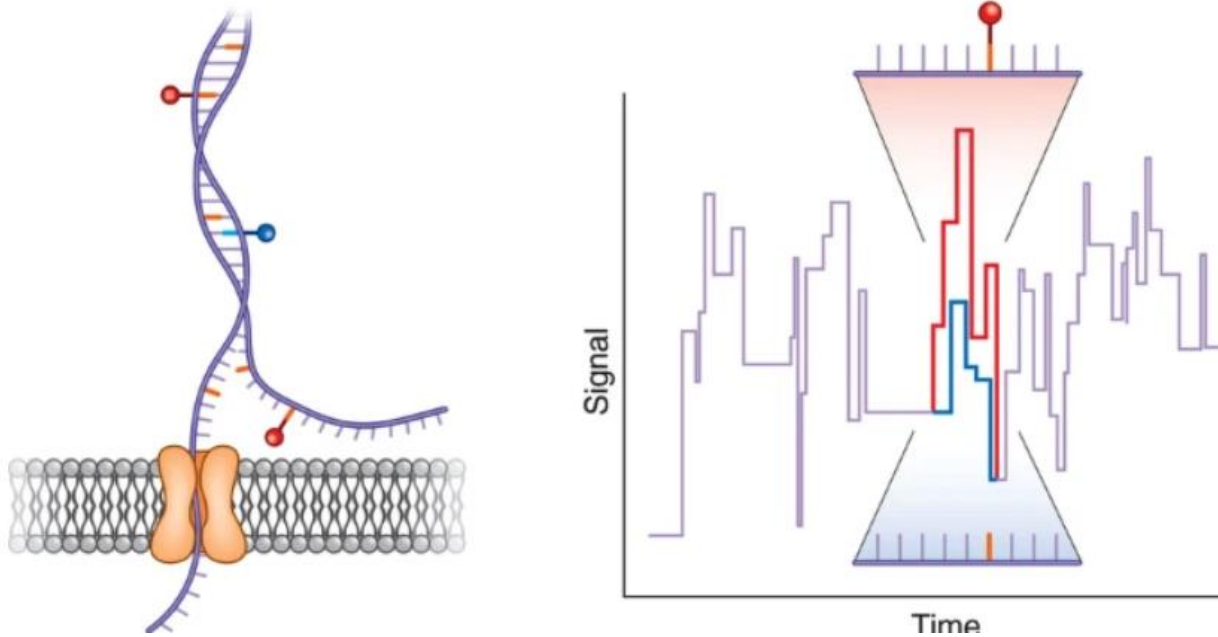
# Bisulfite sequencing



# 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

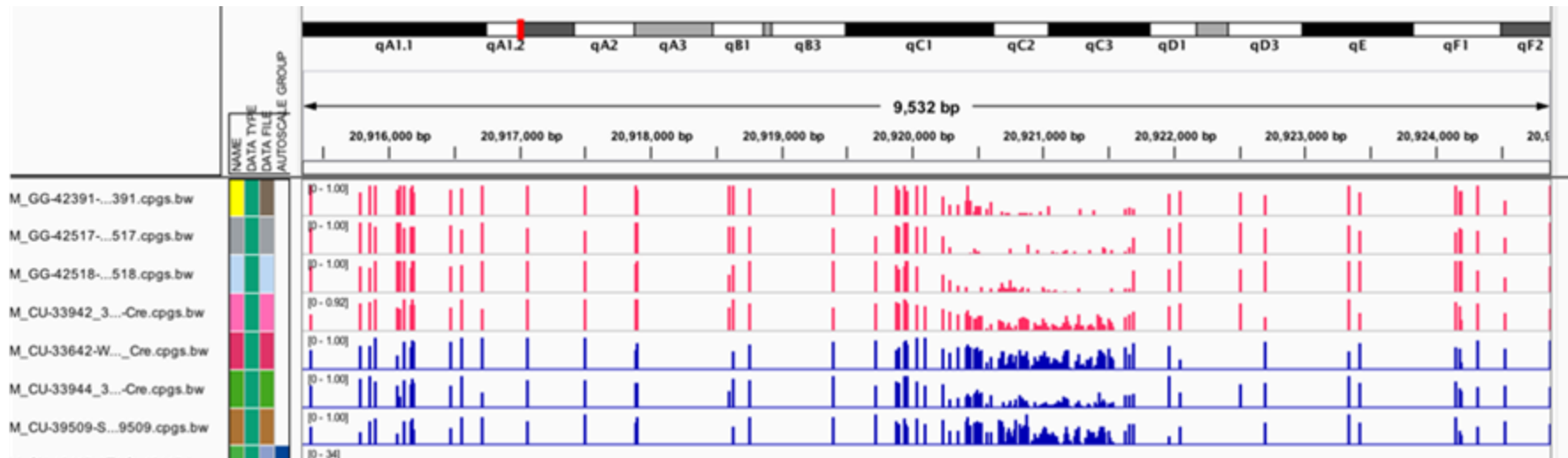
- Determine methylation fraction at each site in the genome
  - Count the Cs and Ts, taking strandedness into account
  - Some tools account for SNPs while doing this
- Why isn't every position 0%, 50% or 100%?
  - we're sequencing a population of cells!



# 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!

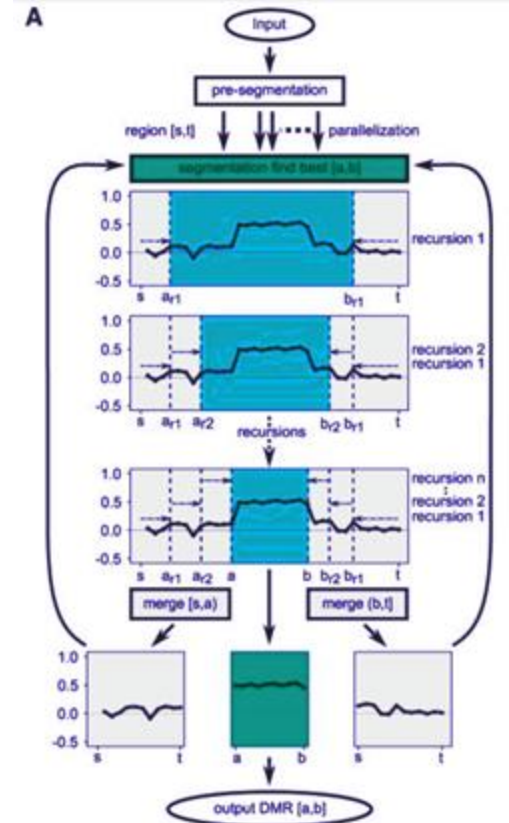
# IGV visualization



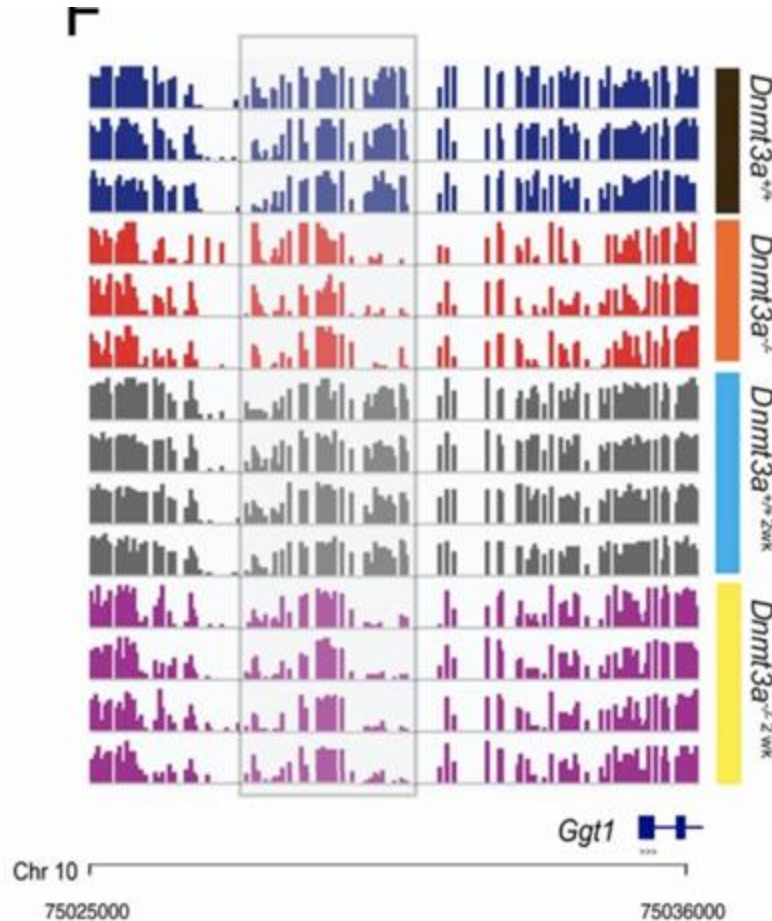


# Differentially methylated regions

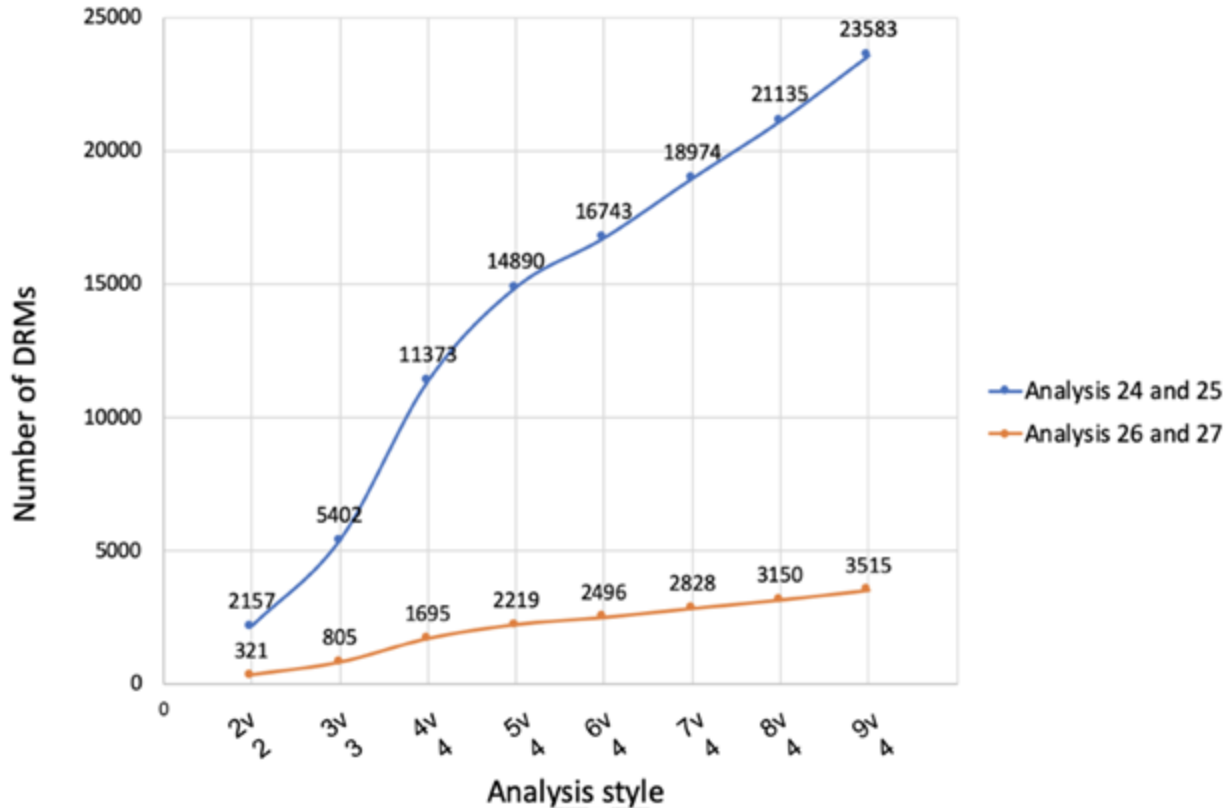
- Comparing two groups to find changes
- Finding DMRs is a segmentation problem
- We use a tool called metilene



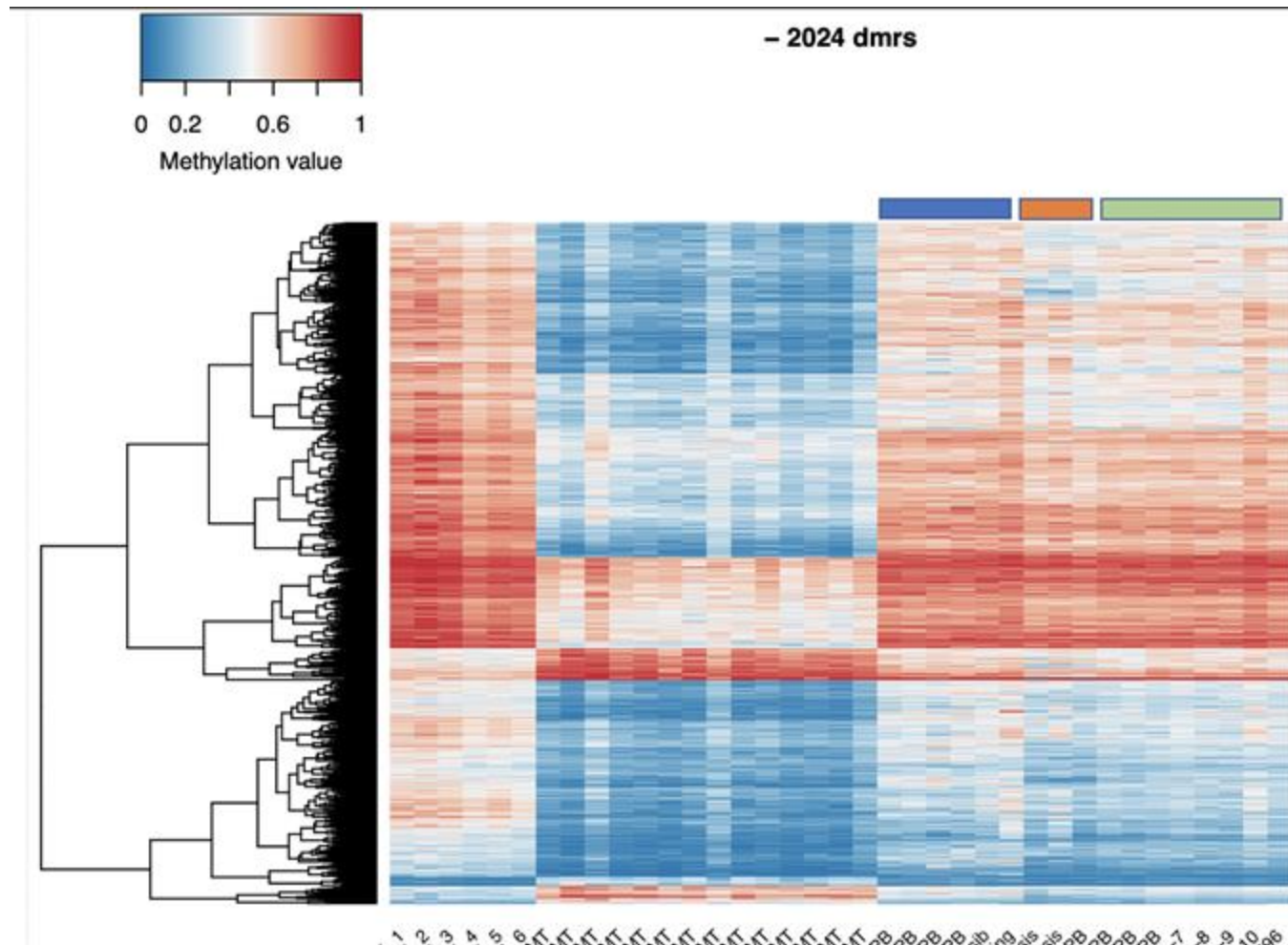
# Differentially methylated regions



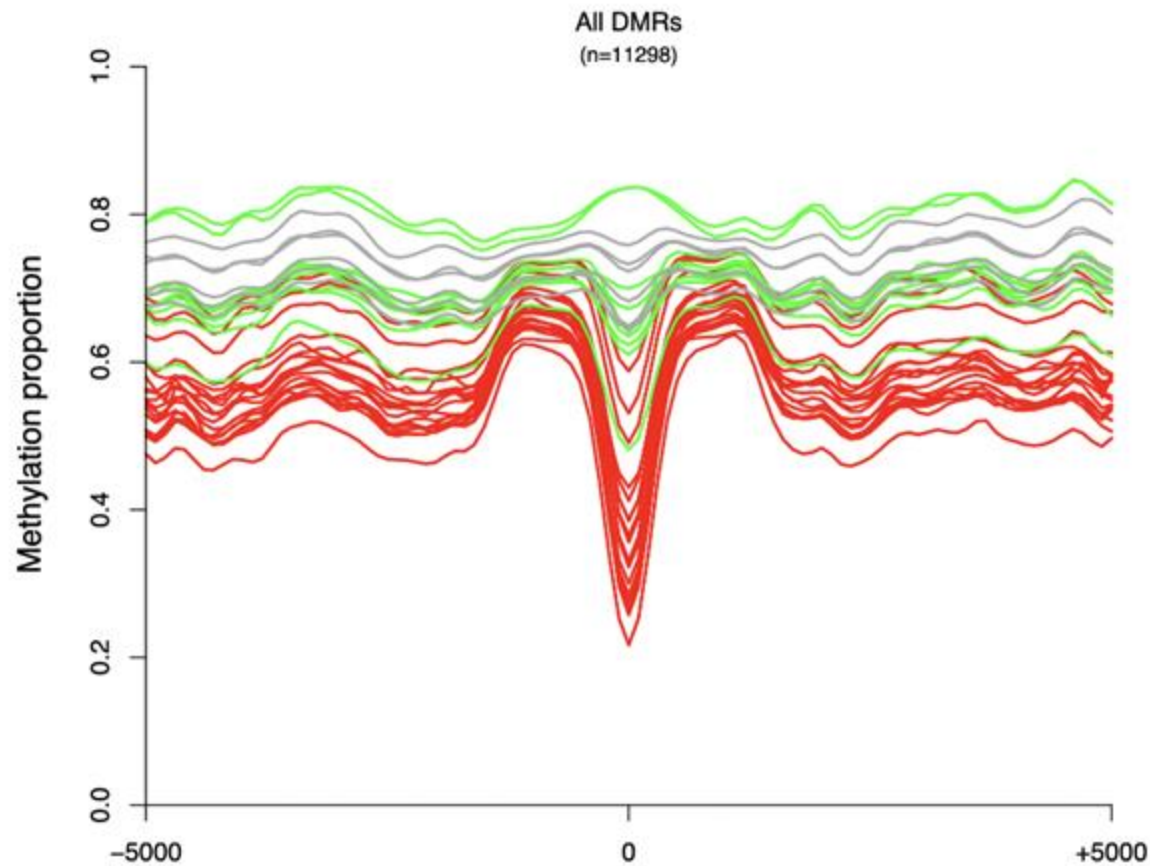
# Number of samples matters!



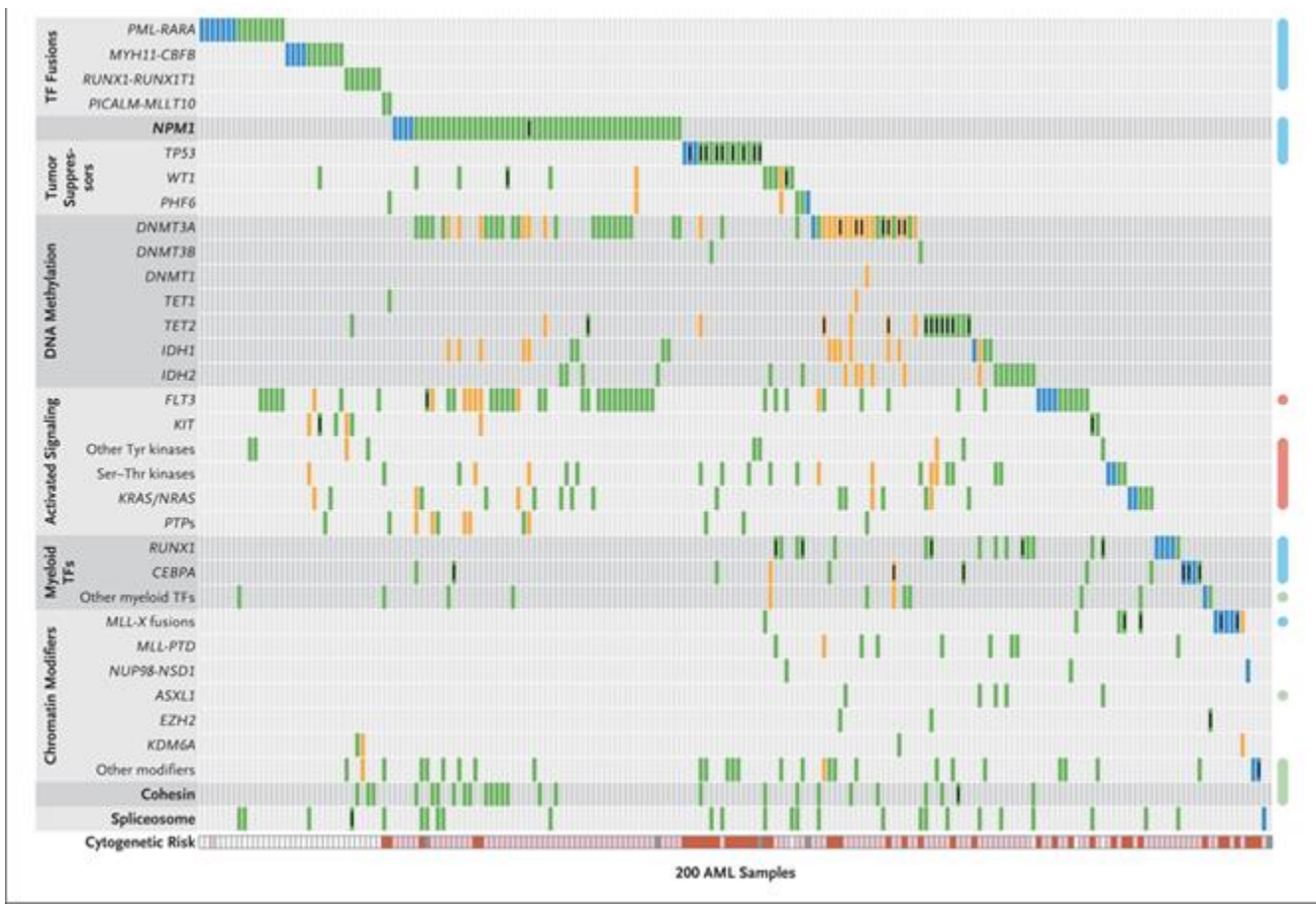
# Heatmaps



# Canyon Plots

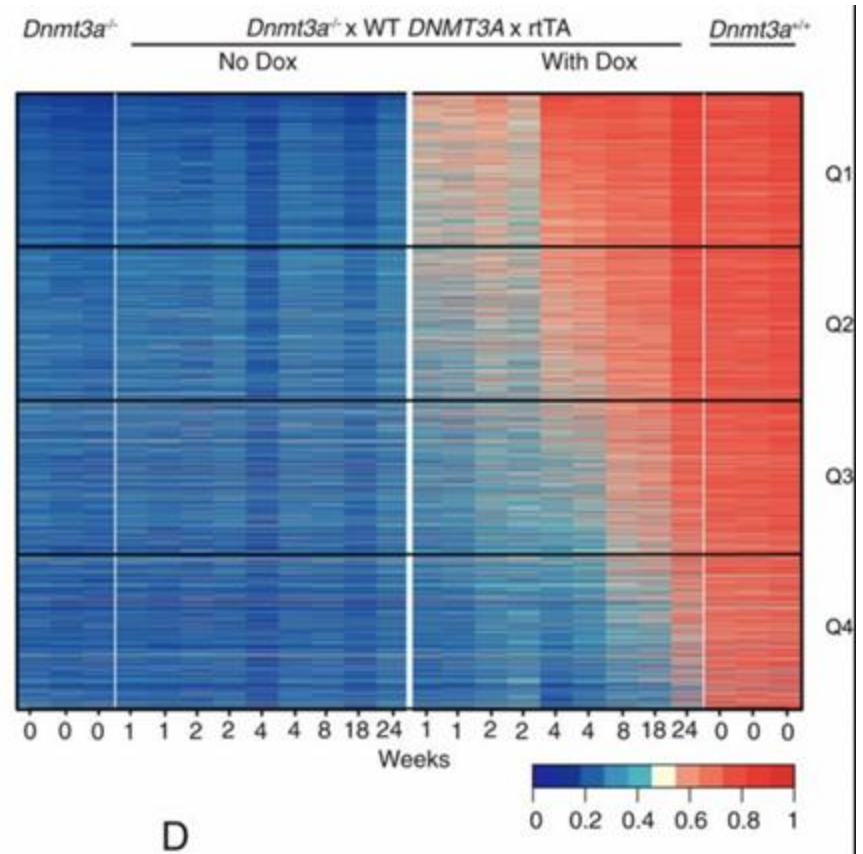


# DNMT3A deficiency



# DNMT3A deficiency

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





# DNMT3A deficiency

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

