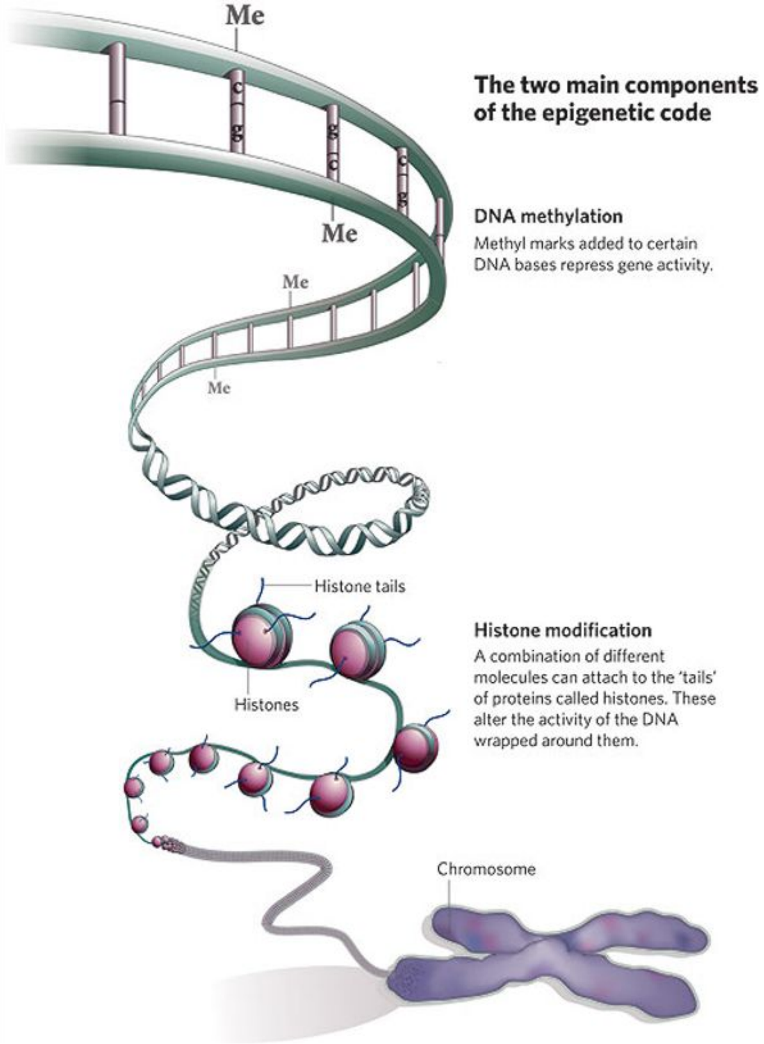
Bioinformatics Workshop

Session #16

Bisulfite Sequencing and Analysis

Chris Miller

Epigenetics



The two main components of the epigenetic code

DNA methylation

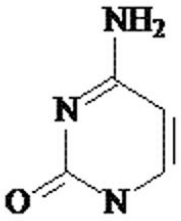
Methyl marks added to certain DNA bases repress gene activity.

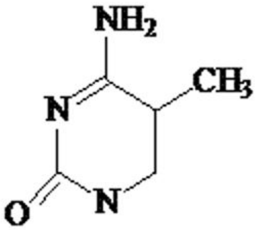
Histone modification

A combination of different molecules can attach to the 'tails' of proteins called histones. These alter the activity of the DNA wrapped around them.

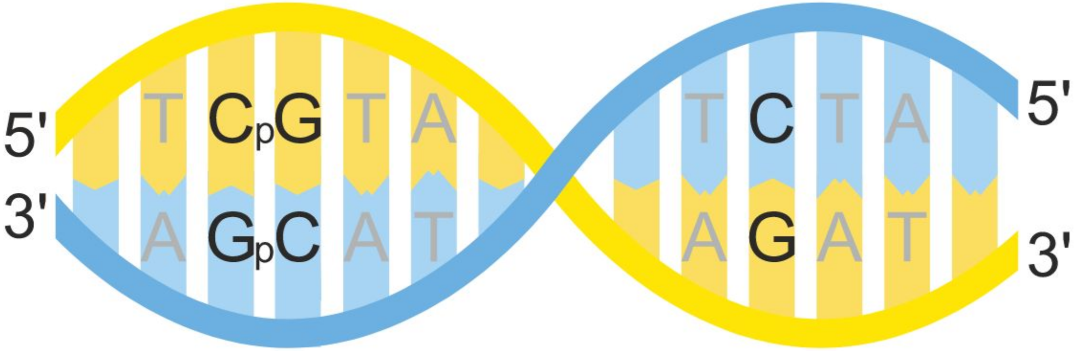
http://www.nature.com/nature/journal/v441/n7090/full/441143a.html

DNA Methylation





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



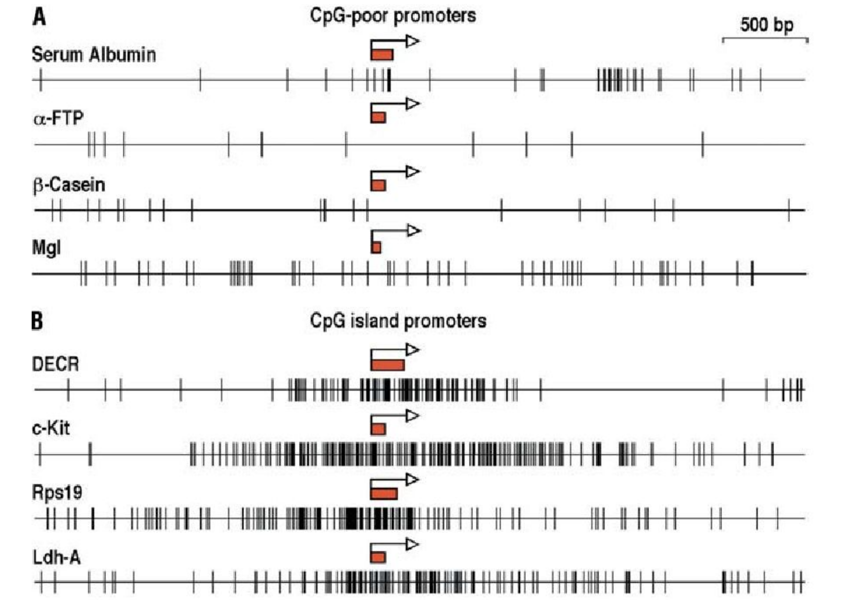
- https://en.wikipedia.org/wiki/CpG_site#/media/File:CpG_vs_C-G_bp.svg

DNA Methylation

- CpG Islands

- Length >= 200 bp
  GC% > 50%
  o/e CpG ratio > 60%

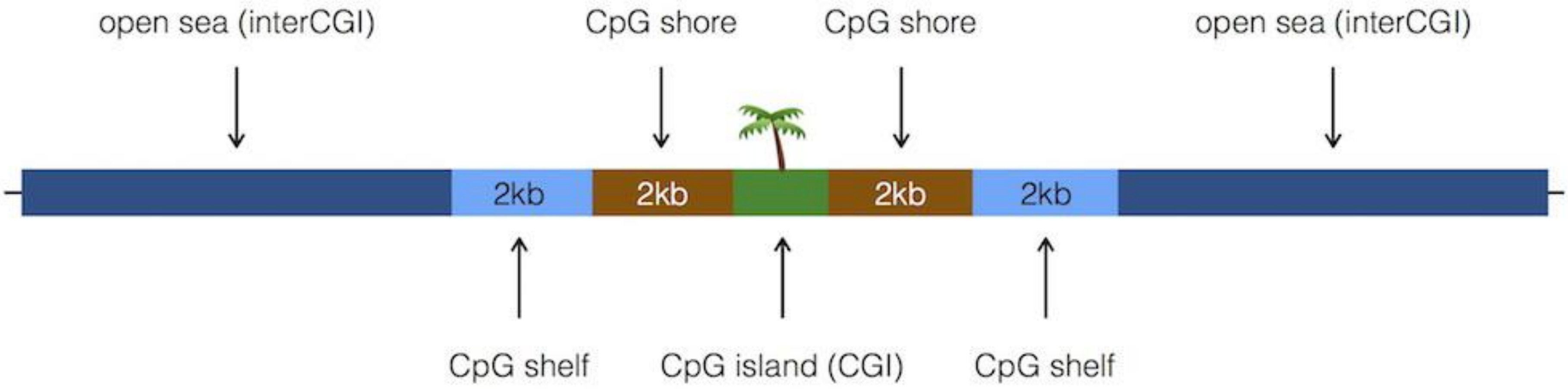
- Selective pressure/
  Evolutionary constraint



ses of promoters in human and mouse. (A) Examples of CpG-poor promoters. Red boxes and arrows

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

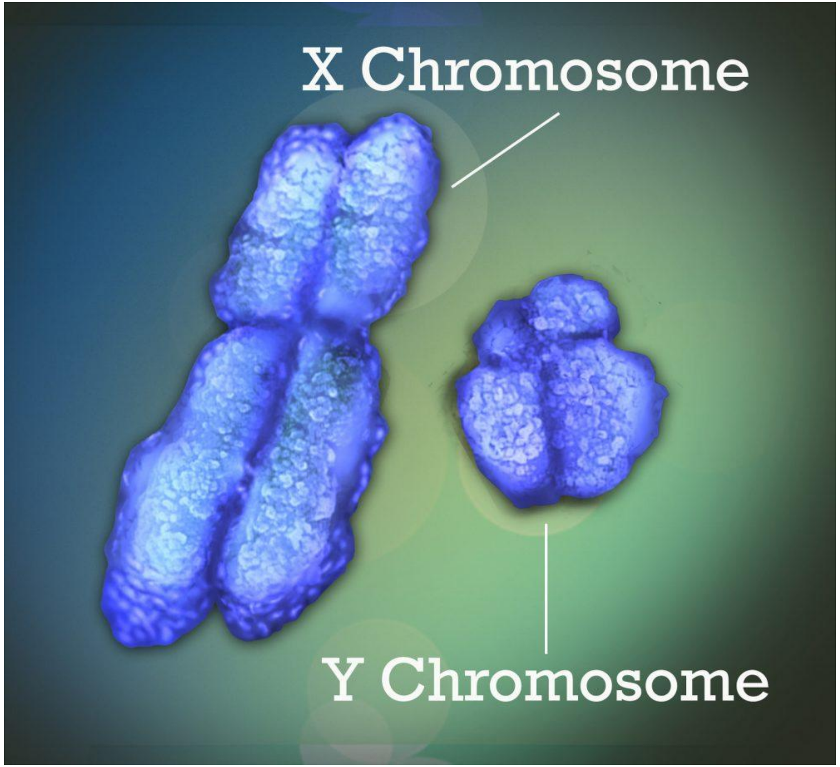
Islands, shores, and shelves



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

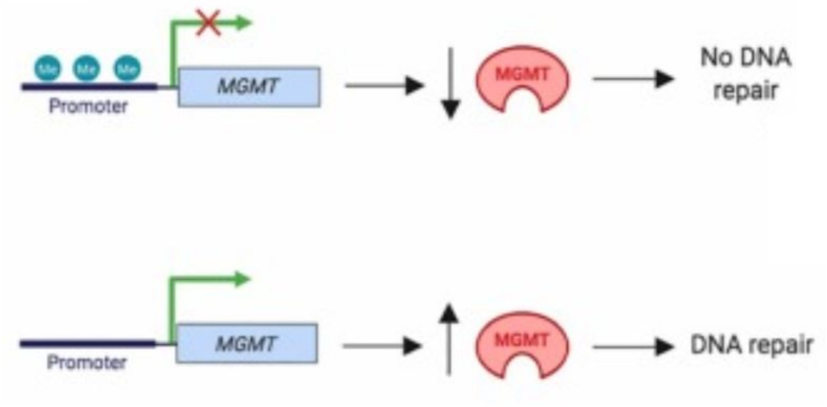


MGMT and Temozolomide

- TMZ is an alkylating agent - damages DNA, causes cell death

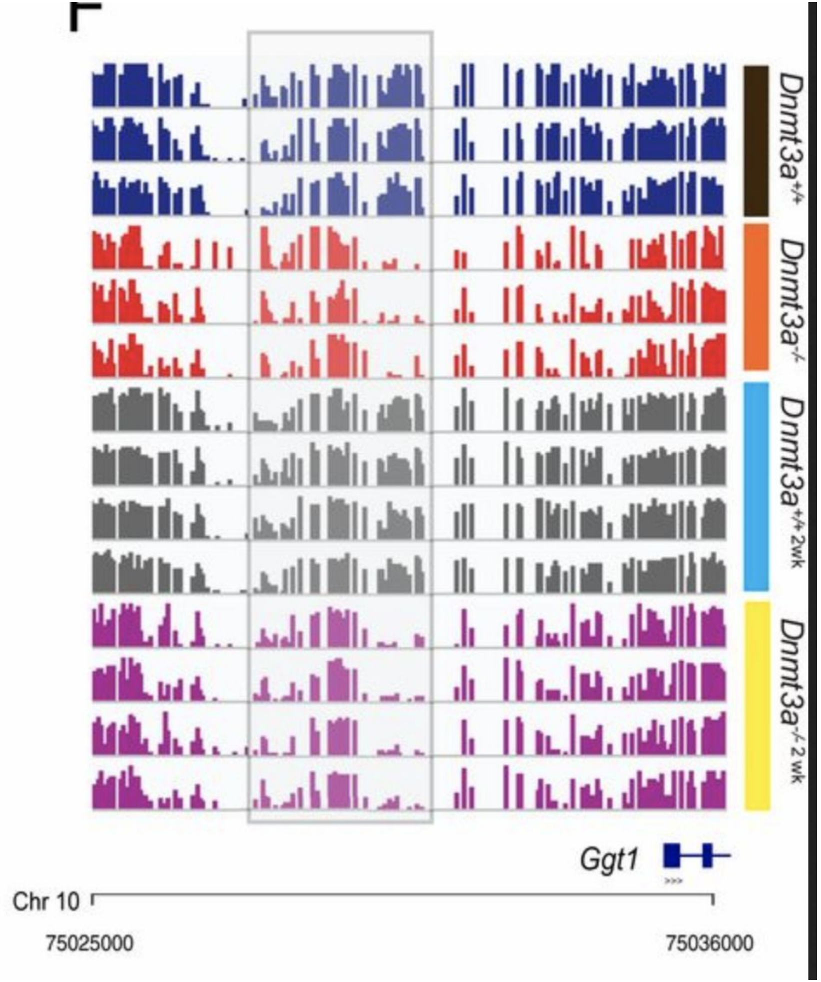
- MGMT “cleans up” the damage

- Methylation of the MGMT promoter is linked to better outcomes!



Methylation Patterns

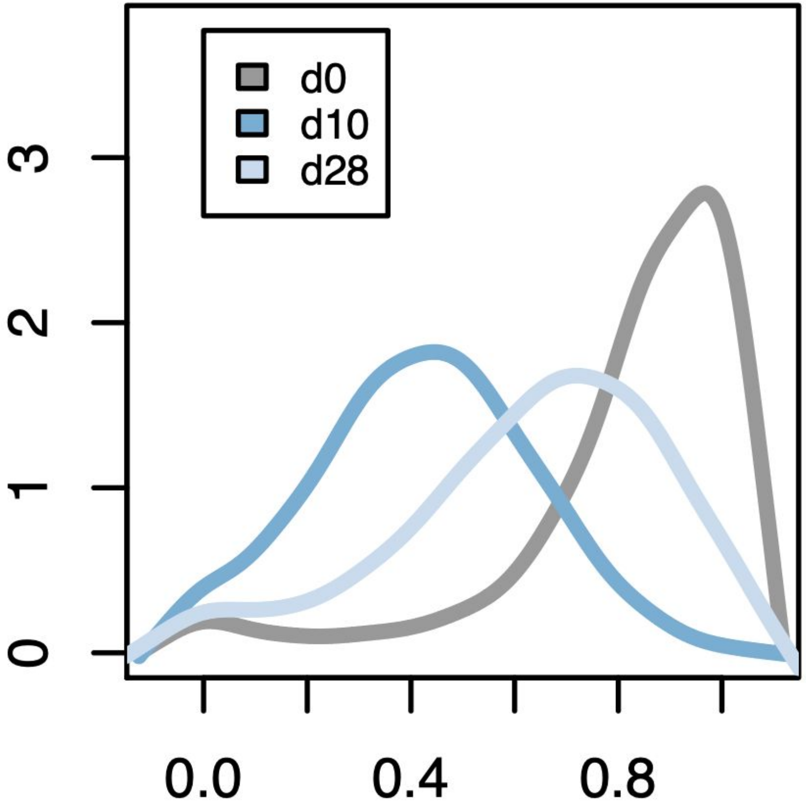
- Methyltransferases that act locally



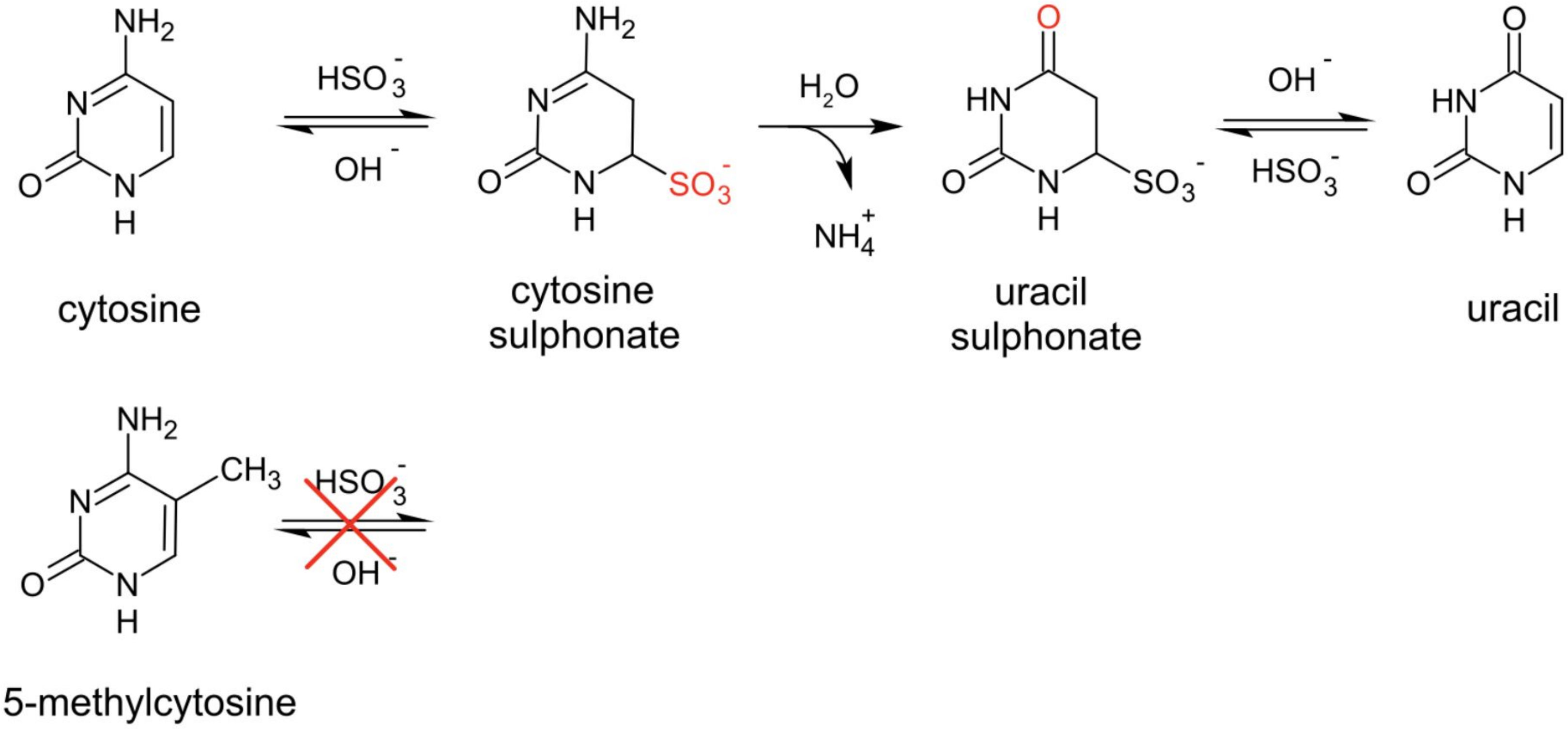
Methylation Patterns

- Methyltransferases that act locally

- Other alterations (or treatments) that act globally

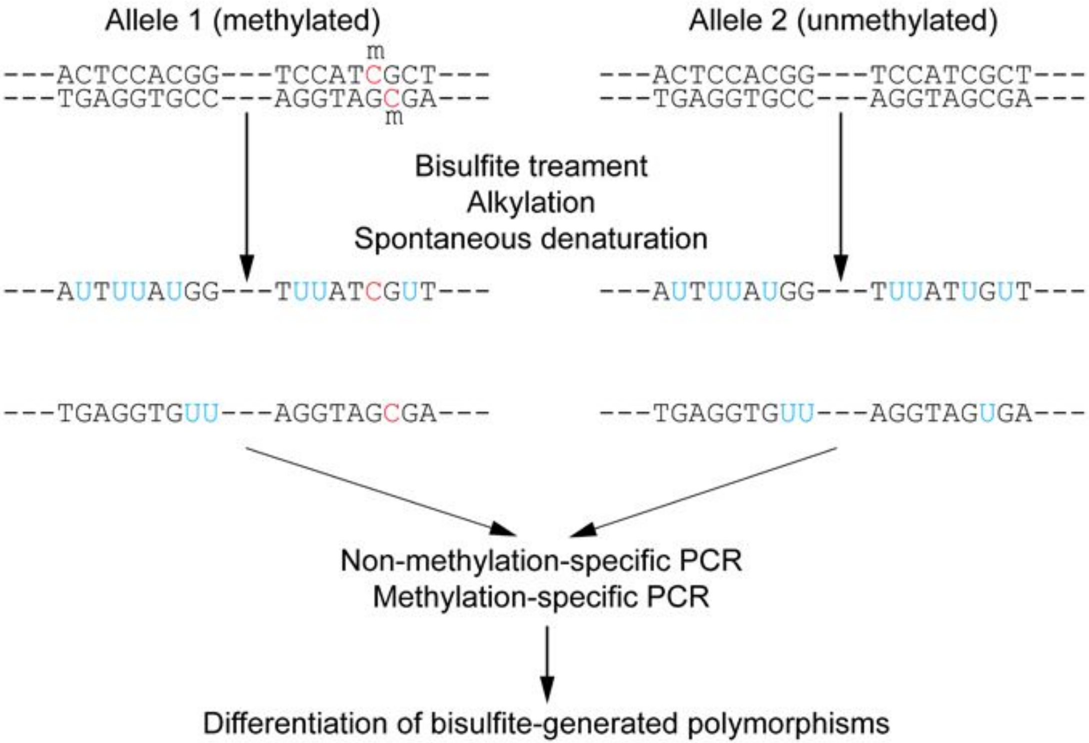


Bisulfite sequencing



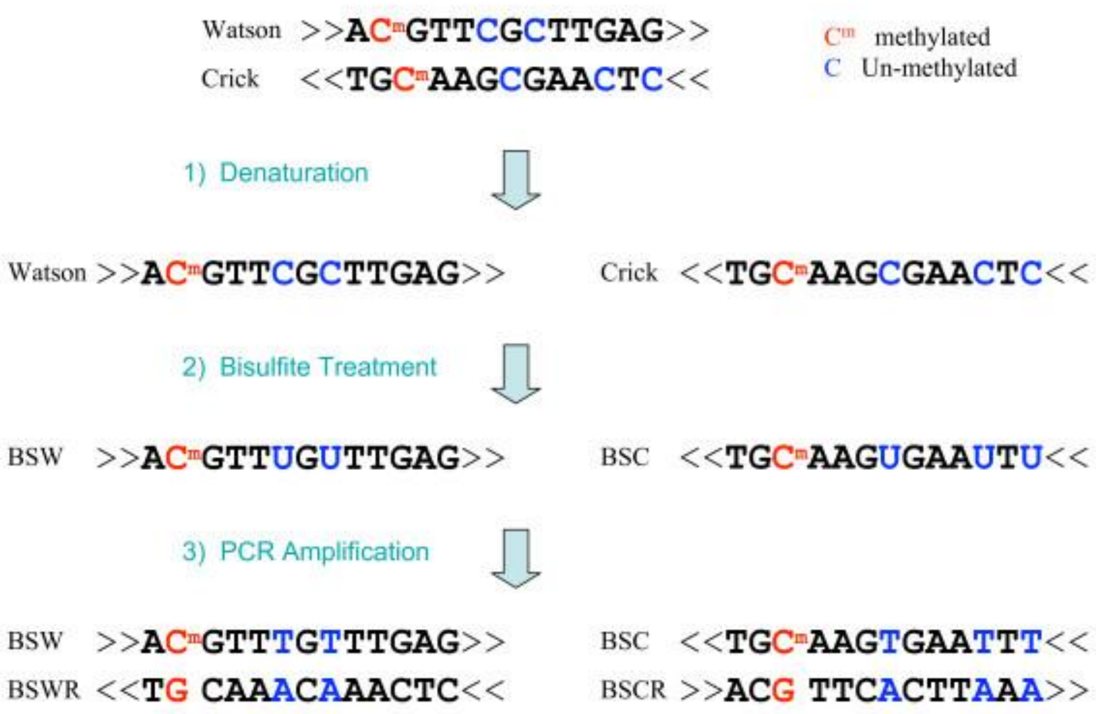
https://en.wikipedia.org/wiki/File:Bisulfite_conversion.svg

Bisulfite sequencing


Allele 1 (methylated)
Allele 2 (unmethylated)
m
---ACTCCACGG---TCCATCGCT---
---TGAGGTGCC---AGGTAGCGA---
m
---ACTCCACGG---TCCATCGCT---
---TGAGGTGCC---AGGTAGCGA---
Bisulfite treament
Alkylation
Spontaneous denaturation
---AUTUUAUGG---TUUATCGUT---
---AUTUUAUGG---TUUATUGUT---
---TGAGGTGUU---AGGTAGCGA---
---TGAGGTGUU---AGGTAGUGA---
Non-methylation-specific PCR
Methylation-specific PCR
Differentiation of bisulfite-generated polymorphisms

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

Bisulfite sequencing



Watson >>AC^{m}GTTCGCTTGAG>>

Crick <<TGC^{m}AAGCGAACTC<<

C^m methylated

C Un-methylated

1) Denaturation

Watson >>AC^{m}GTTCGCTTGAG>>

Crick <<TGC^{m}AAGCGAACTC<<

2) Bisulfite Treatment

BSW >>AC^{m}GTTUGUTTGAG>>

BSC <<TGC^{m}AAGUGAAUTU<<

3) PCR Amplification

BSW >>AC^{m}GTTTGTTTGAG>>

BSC <<TGC^{m}AAGTGAATTT<<

BSWR <<TG CAAACAAACTC<<

BSCR >>ACG TTCACTTAAA>>

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

Methylation calling

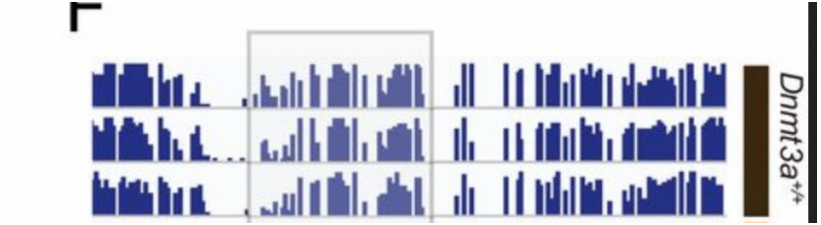
- Determine methylation fraction at each site in the genome
    - Count the Cs and Ts, taking strandedness into account
    - Some tools account for SNPs while doing this

-

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

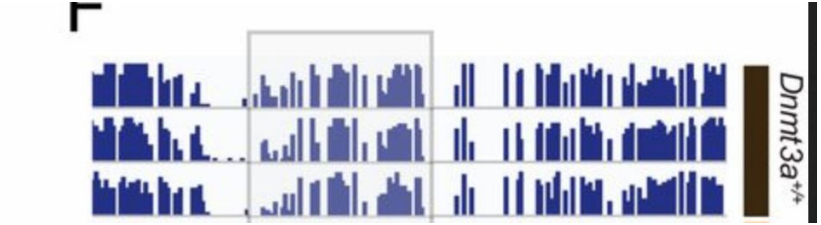
- Why isn't every position 0%, 50% or 100%?



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

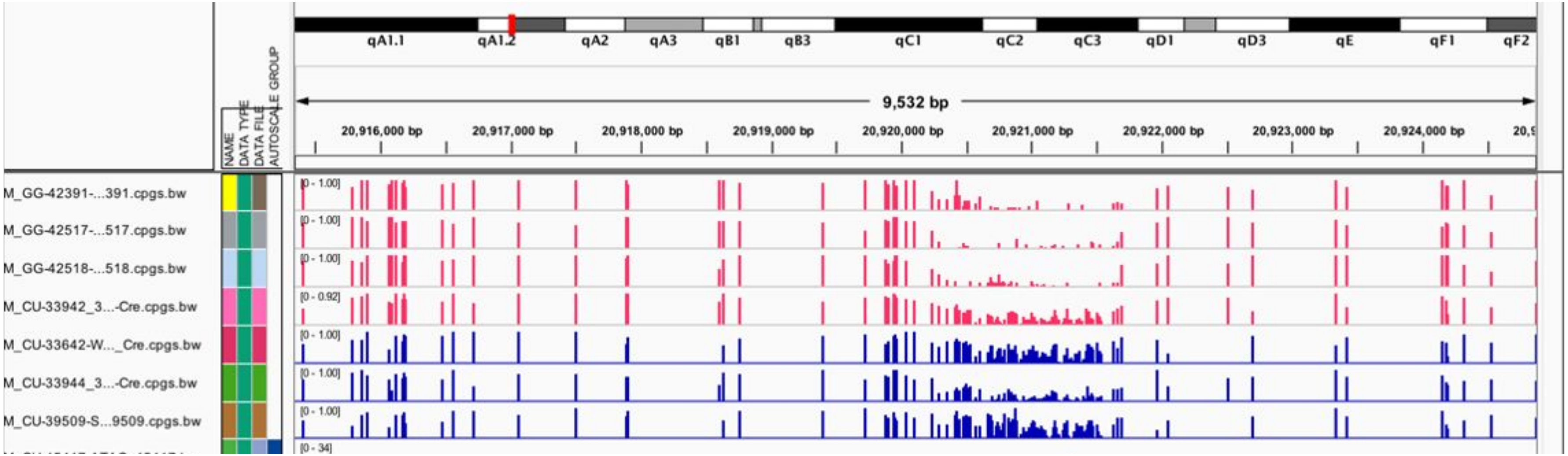
- Pileup: BAM/CRAM > VCF
  - (entries for every site, allele frequencies)

- VCF > bedgraph
  - chr, start, stop, beta_value (methylation fraction)

- bedgraph > bigwig (for visualization in IGV)

- We have a workflow for this!

IGV visualization

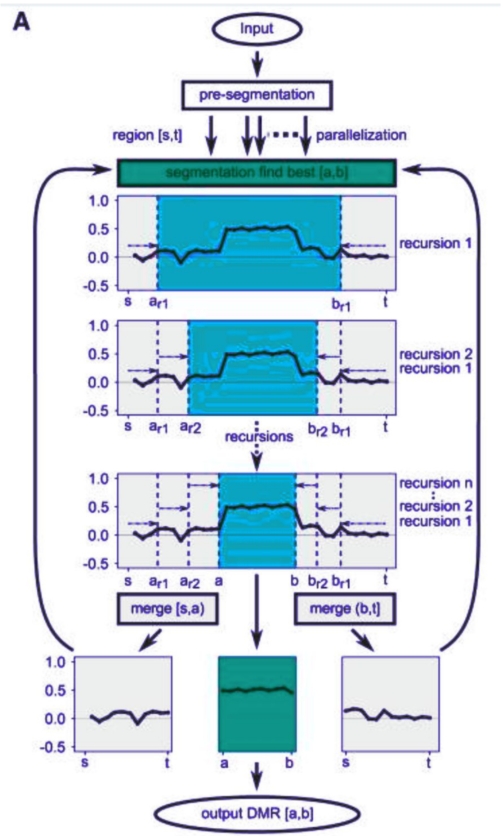


Differentially methylated regions

- Comparing two groups to find changes

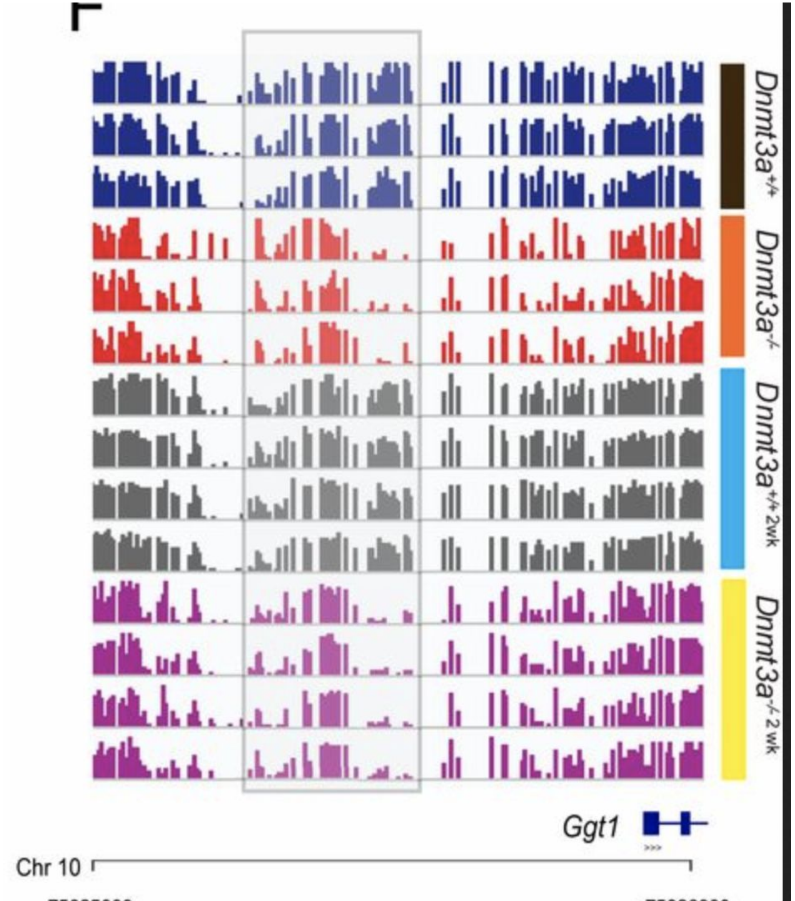
- Finding DMRs is a segmentation problem

- We use a tool called metilene

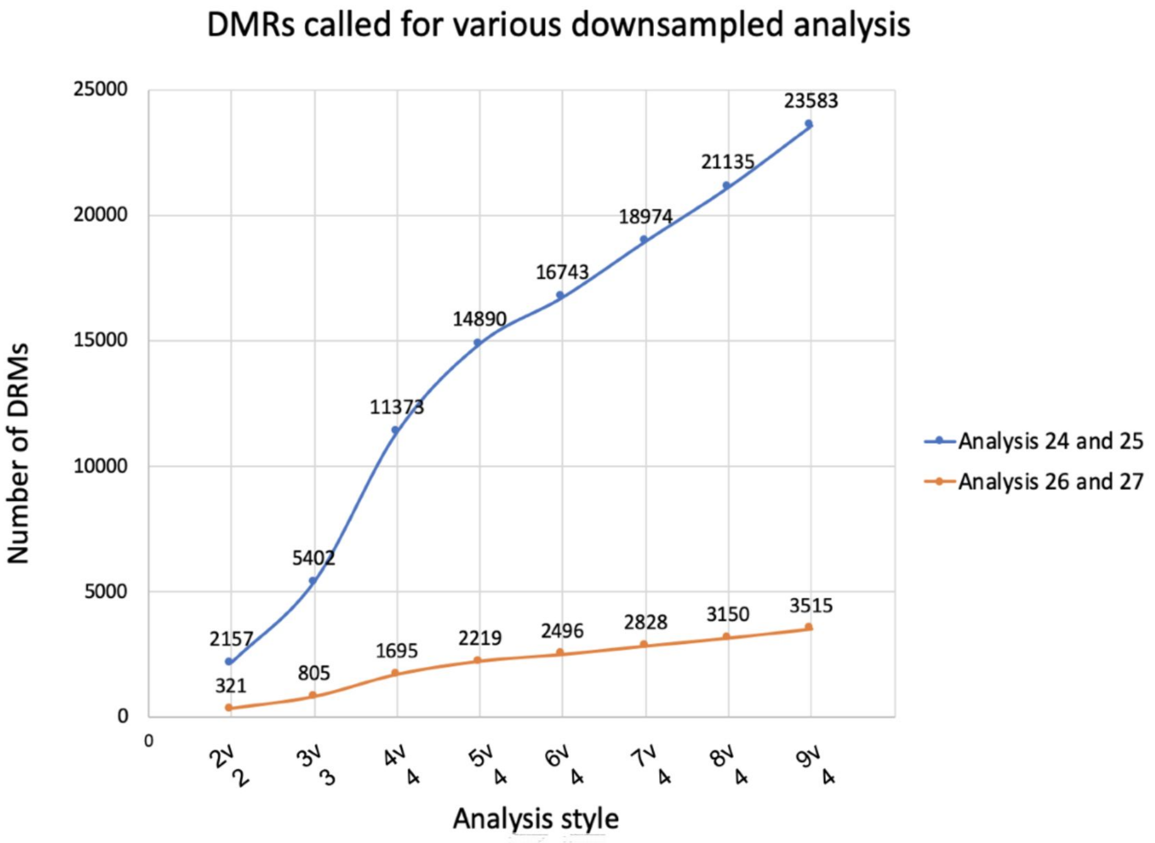


https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4728377/

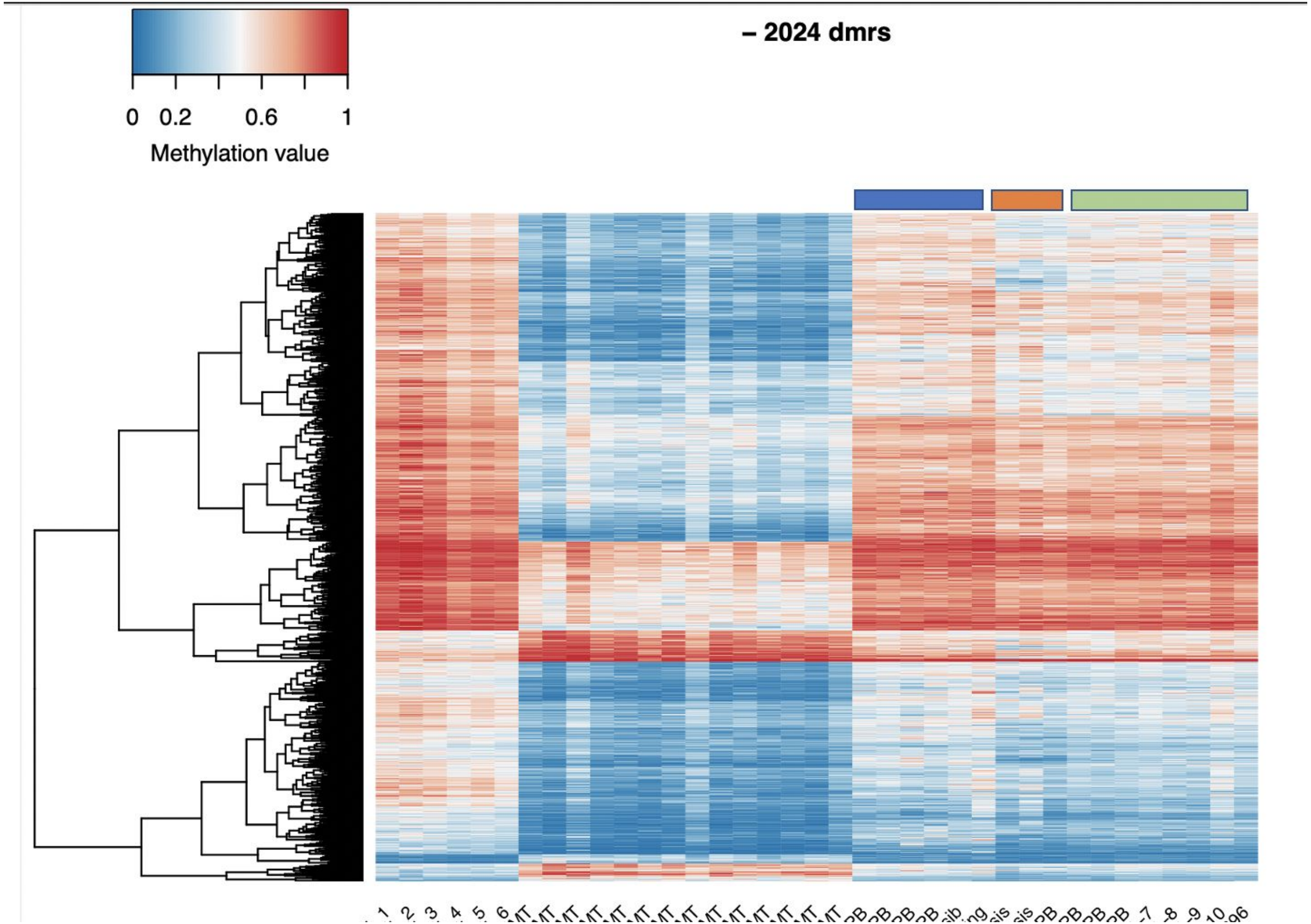
Differentially methylated regions



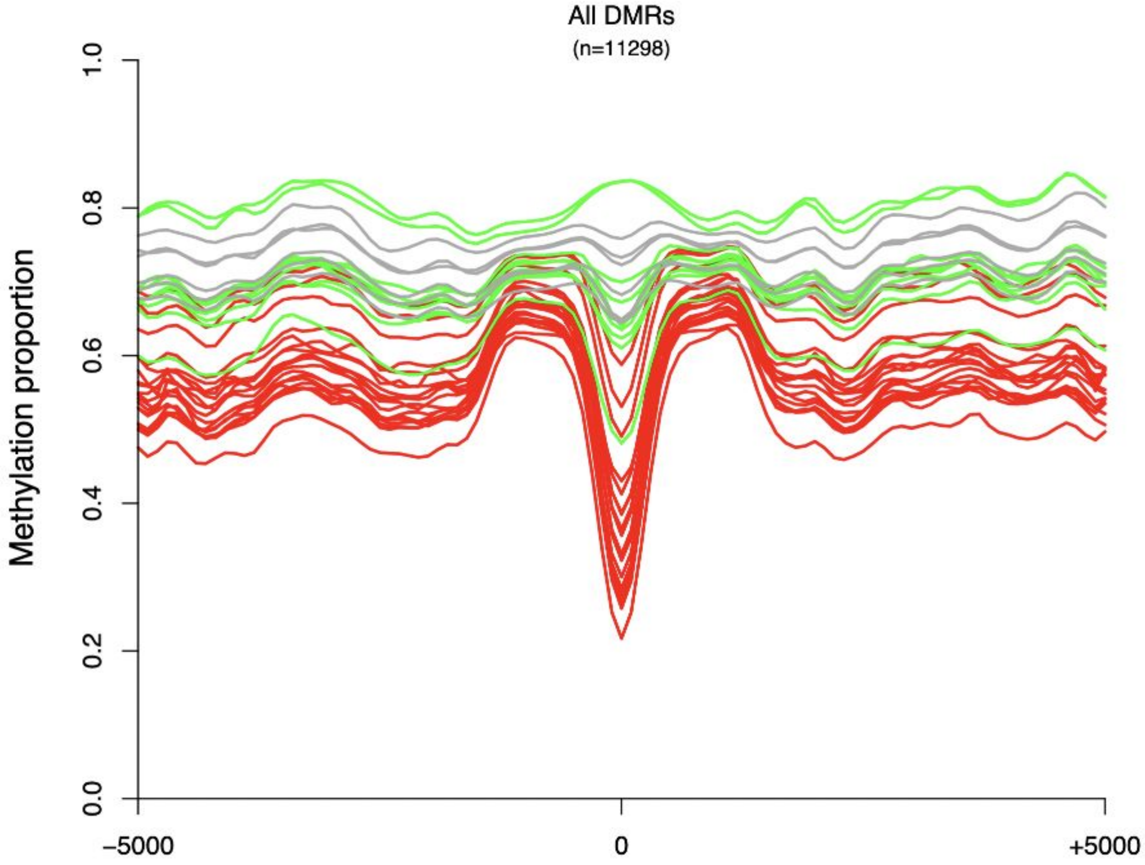
Number of samples matters!



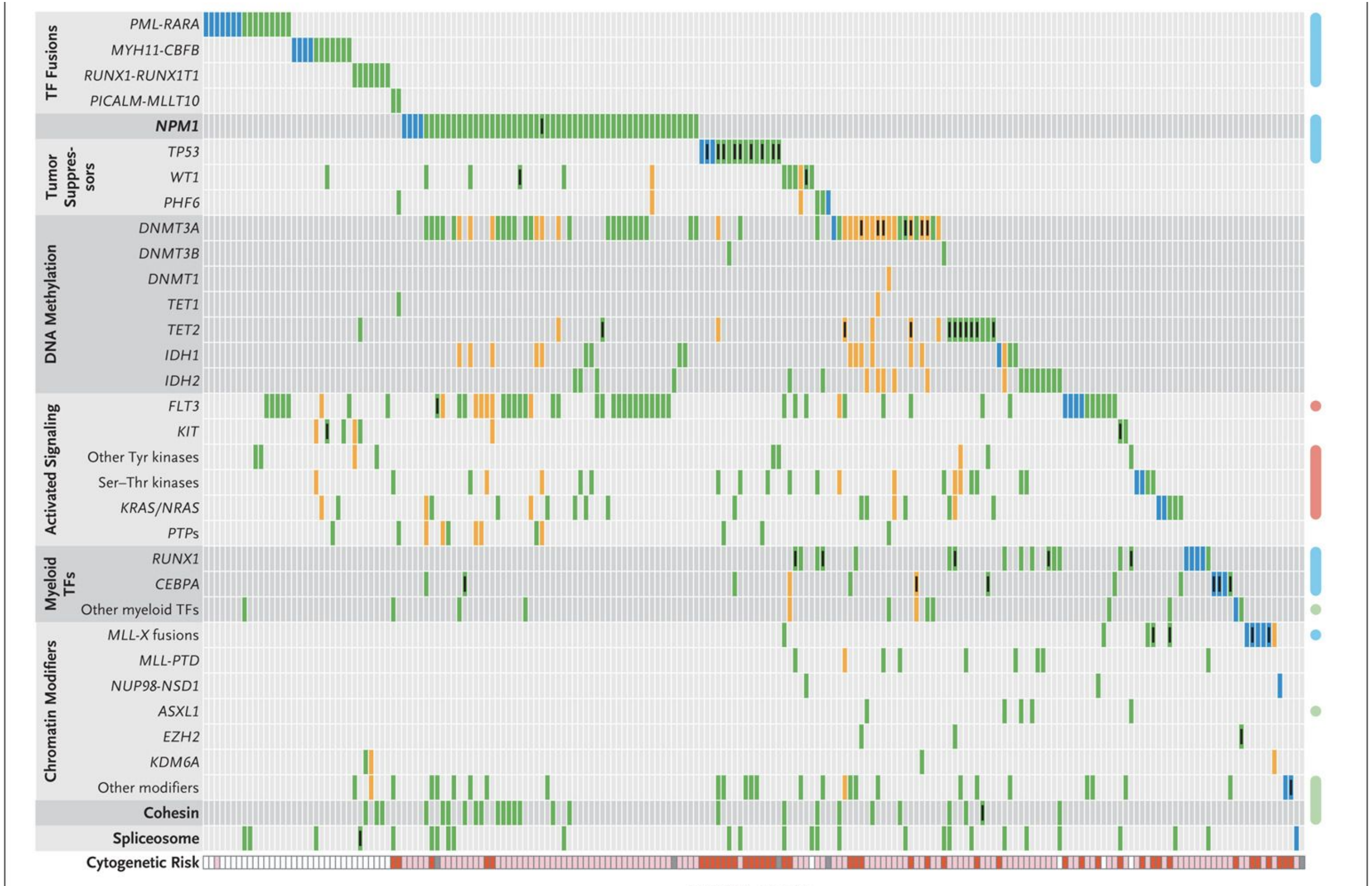
Heatmaps



Canyon Plots



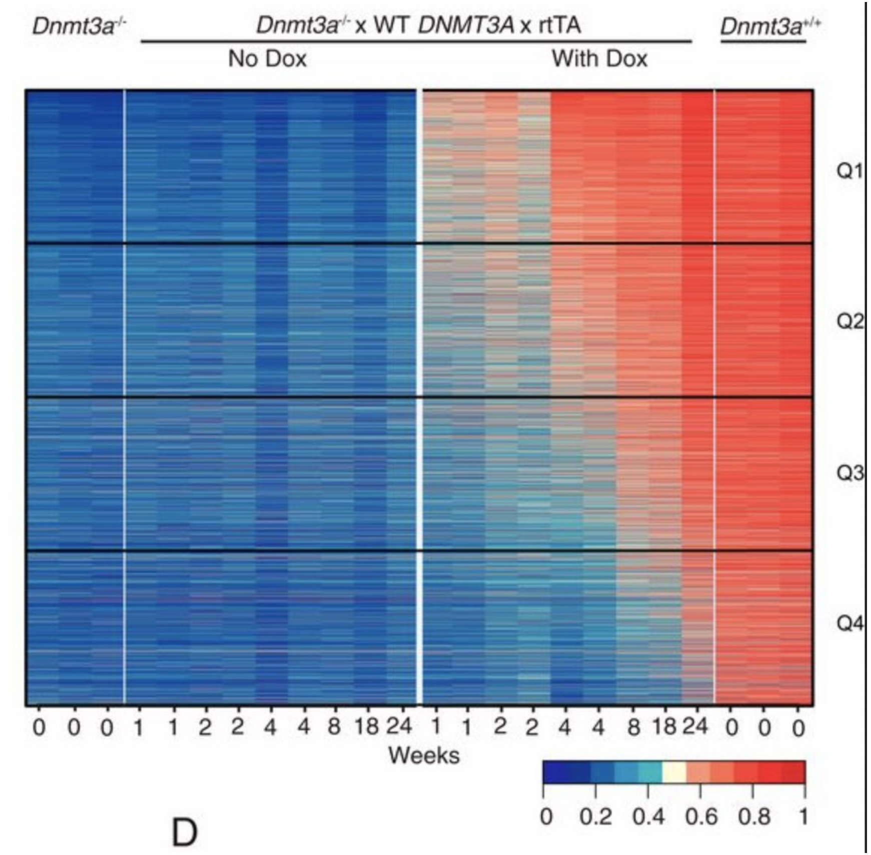
DNMT3A deficiency



DNMT3A deficiency

- Mouse models (and human data)

- Looking at context, effects, and reversibility

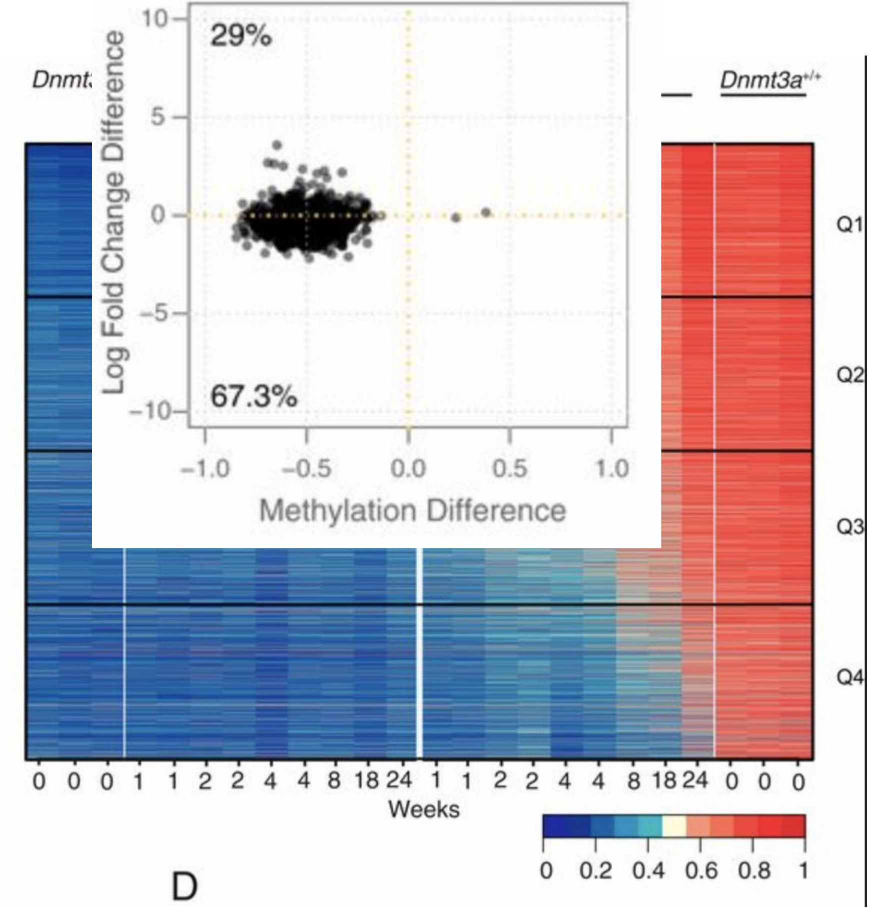


https://www.pnas.org/content/117/6/3123.full

DNMT3A deficiency

- Mouse models (and human data)

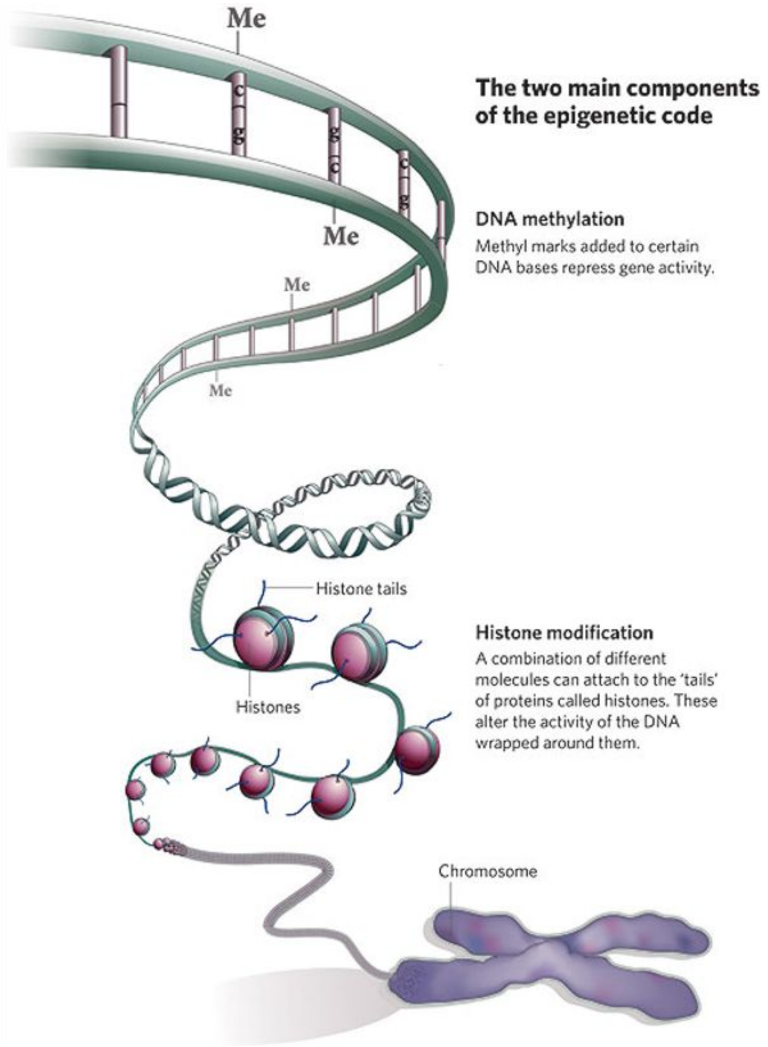
- Looking at context, effects, and reversibility



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

ChIP-seq/ATAC-seq

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



The two main components of the epigenetic code

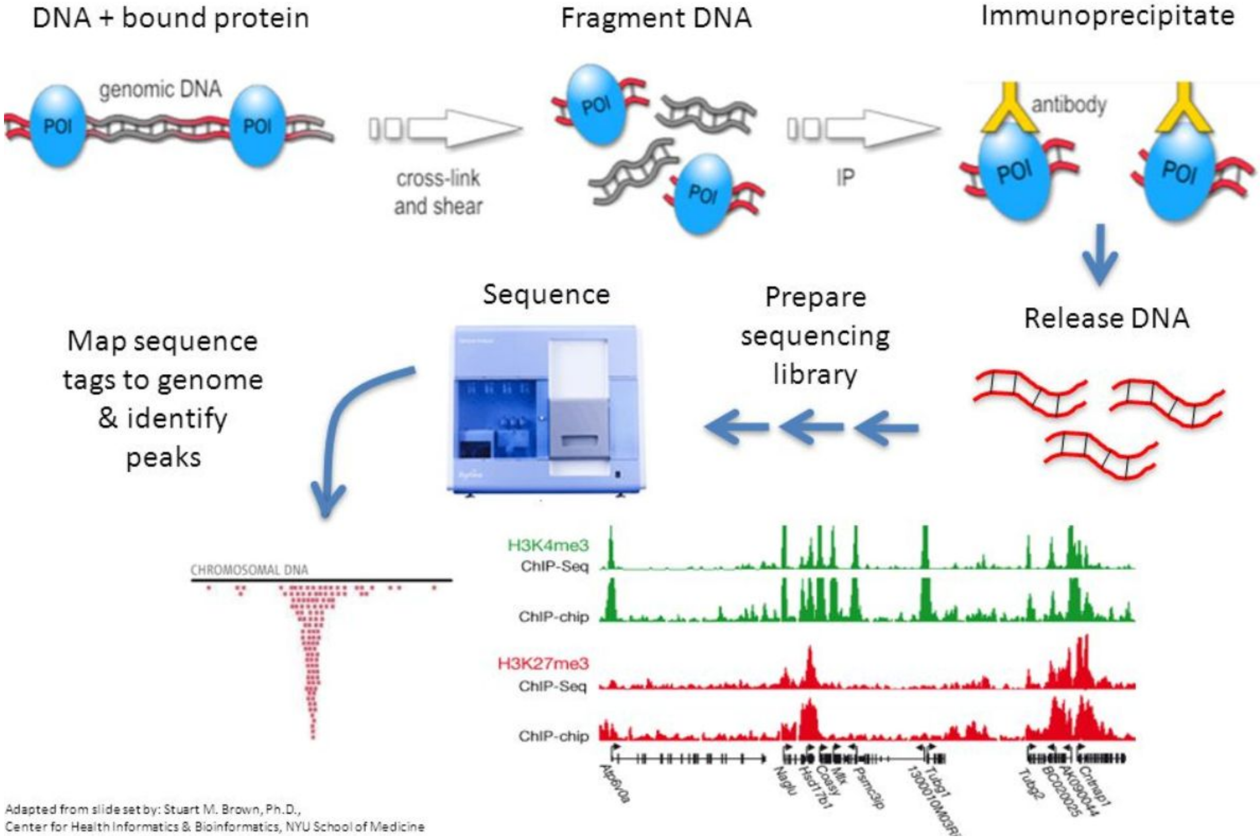
DNA methylation

Methyl marks added to certain DNA bases repress gene activity.

Histone modification

A combination of different molecules can attach to the 'tails' of proteins called histones. These alter the activity of the DNA wrapped around them.

ChIP-seq



DNA + bound protein

Fragment DNA

Immunoprecipitate

Release DNA

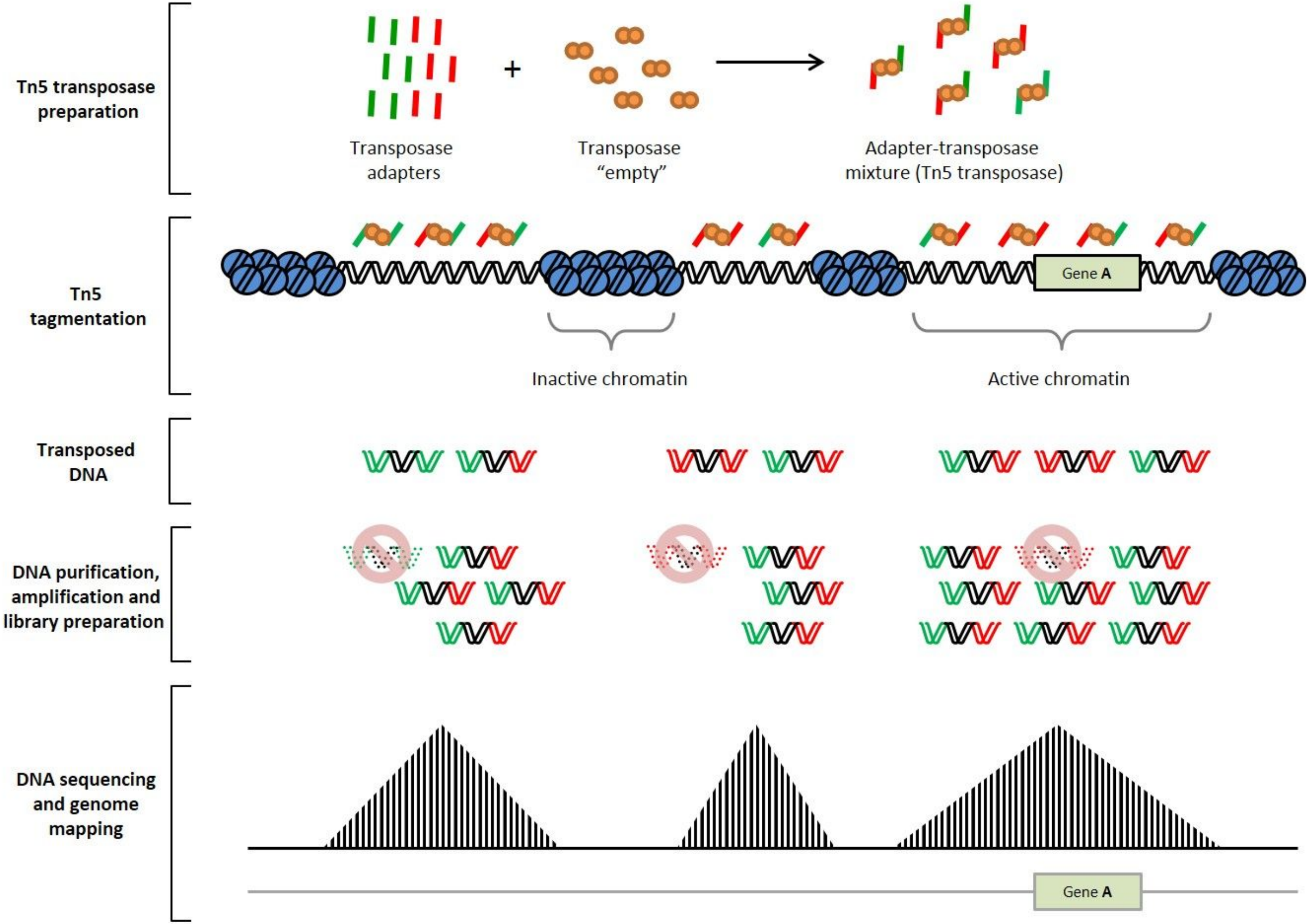
Prepare sequencing library

Sequence

Map sequence tags to genome & identify peaks

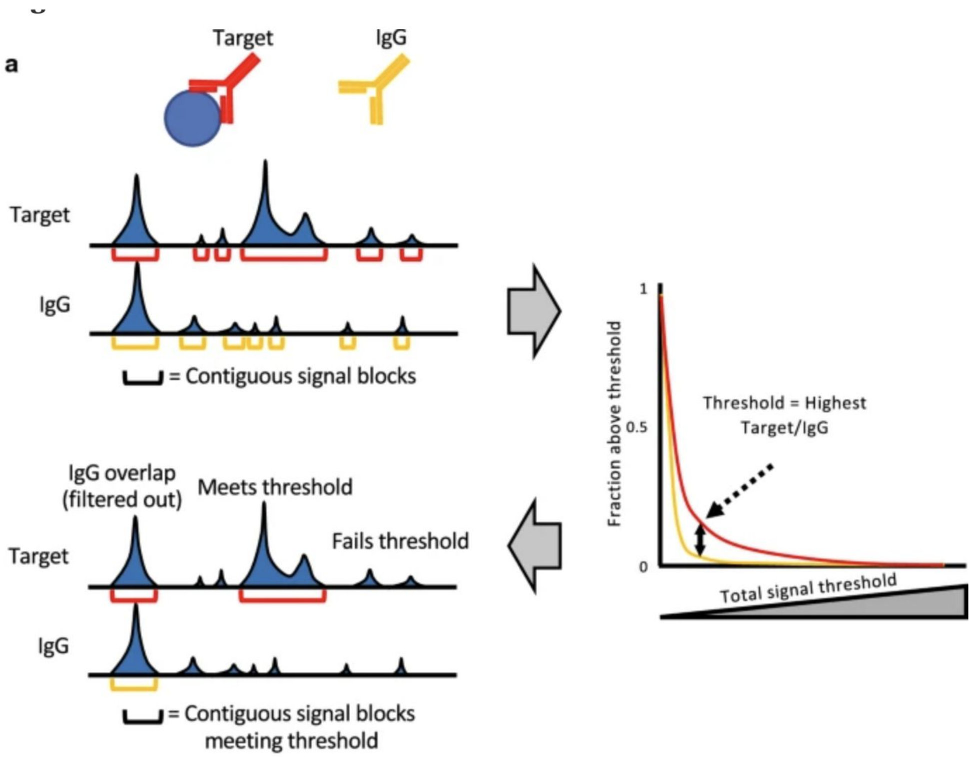
Adapted from slide set by: Stuart M. Brown, Ph.D.,
Center for Health Informatics & Bioinformatics, NYU School of Medicine

ATAC-seq


Tn5 transposase preparation
Transposase adapters
+
Transposase "empty"
Adapter-transposase mixture (Tn5 transposase)
Tn5 tagmentation
Gene A
Inactive chromatin
Active chromatin
Transposed DNA
DNA purification, amplification and library preparation
DNA sequencing and genome mapping
Gene A

Peak-calling

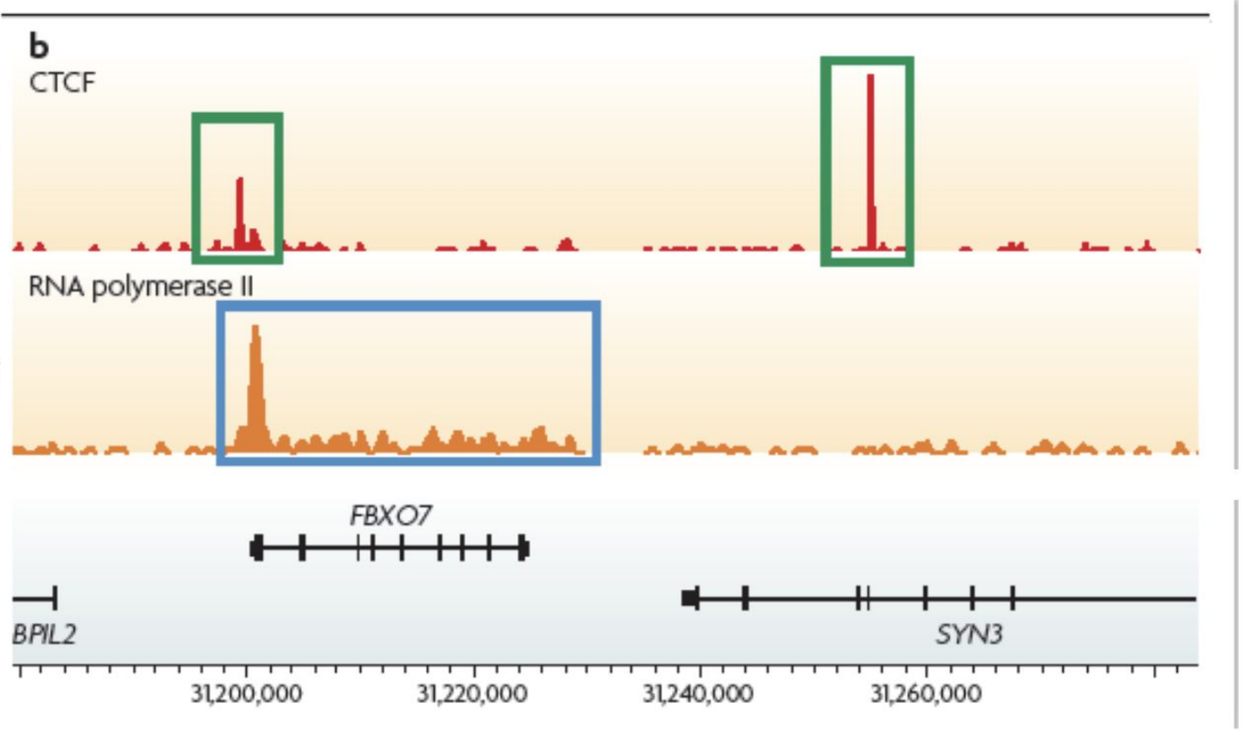
MACS2, HOMER, SEACR, etc


a
Target
IgG
Target
IgG
= Contiguous signal blocks
1
0.5
0
Fraction above threshold
Threshold = Highest Target/IgG
Total signal threshold
IgG overlap (filtered out)
Meets threshold
Fails threshold
Target
IgG
= Contiguous signal blocks meeting threshold

Proteins bind in different ways

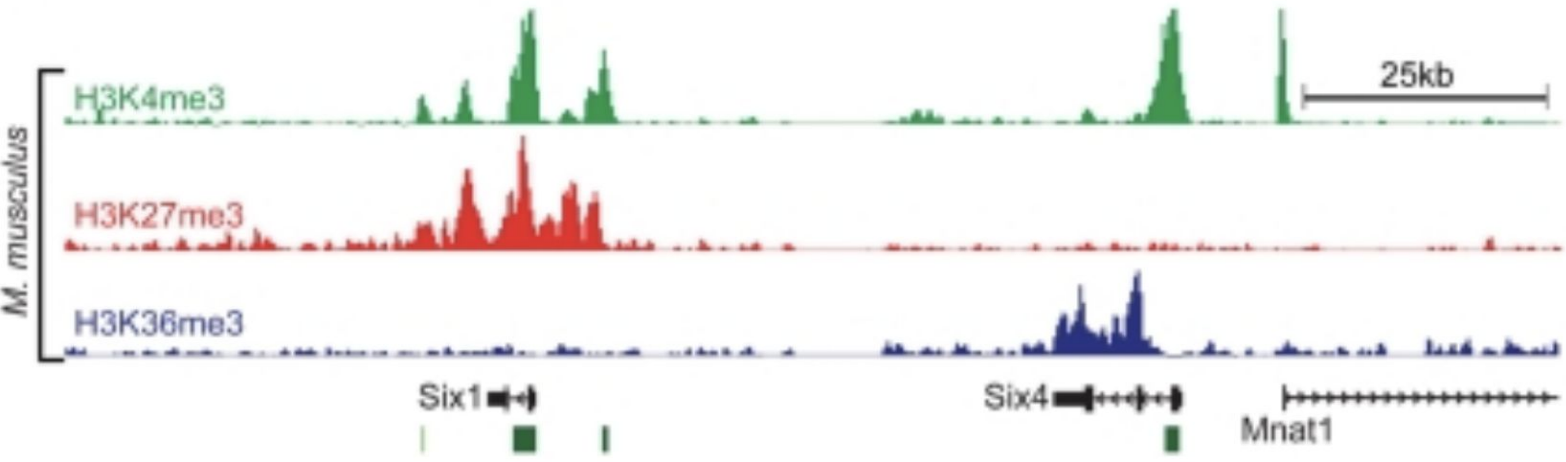
Transcription factor – tight, highly-peaked binding region

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

Interpretation



Activating mark (near TSS)

Peaks within body of inactive genes

Peaks within body of active genes

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