Regulatory Genomics

Yongjin Park

Slides credit
(borrowed a lot from):

Carles Boix

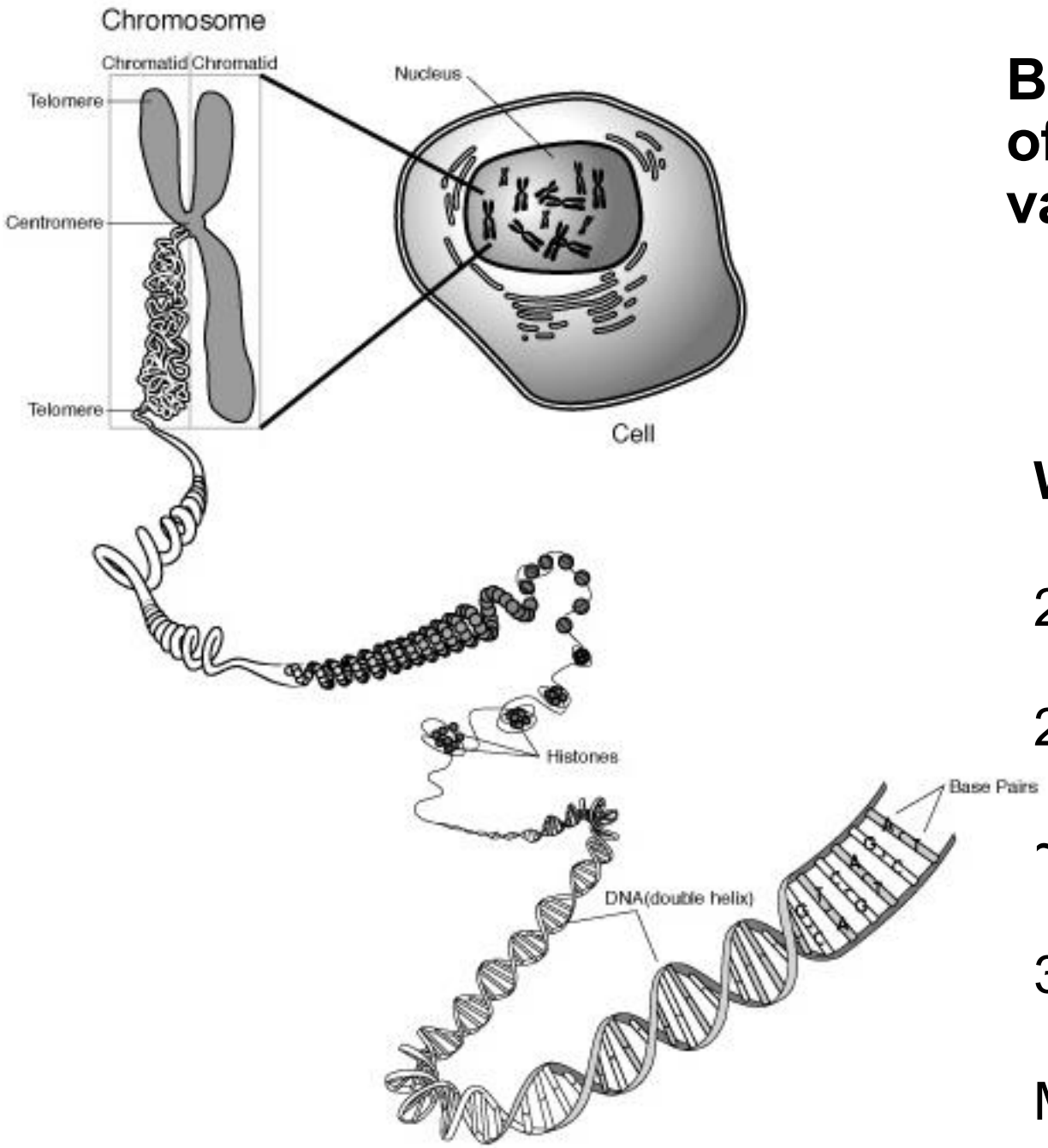
Abhishek Sarkar

Manolis Kellis

Roadmap of STAT/BIOF/GSAT540 2026

data generation followed by *methodology*

3. Formulating biological questions
4. RNA-seq data
5. Experimental design
6. Single-cell genomics
7. Matrix factorization
10. GWAS
11. **Regulatory genomics** (Today)
12. Regression (for reg. genomics)
13. Biological networks
14. Clustering



Building blocks of genetic variation

Within each cell:

2 copies of the genome

23 chromosomes

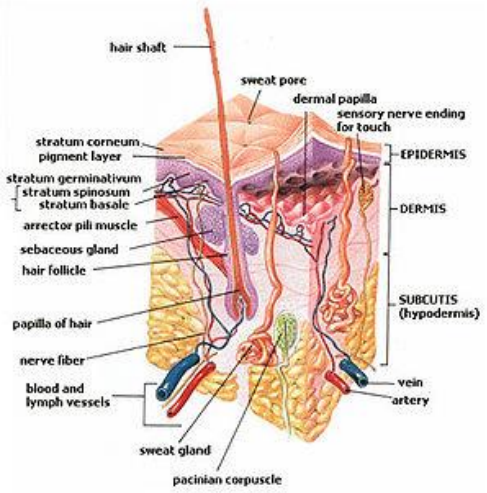
~20,000 genes

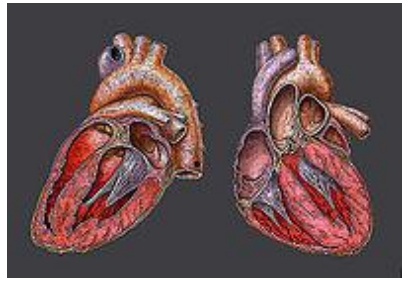
3.2B letters of DNA

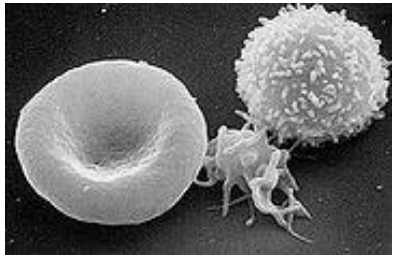
Millions of polymorphic sites

One Genome – Many Cell Types

```
ACCAGTTACGACGGT
CAGGGTACTGATACC
CCAAACCGTTGACCG
CATTTACAGACGGGG
TTTGGGTTTTGCCCC
ACACAGGTACGTTAG
CTACTGGTTTAGCAA
TTTACCGTTACAACG
TTTACAGGGTTACGG
TTGGGATTTGAAAAA
AAGTTTGAGTTGGTT
TTTTCACGGTAGAAC
GTACCGTTACCAGTA
```







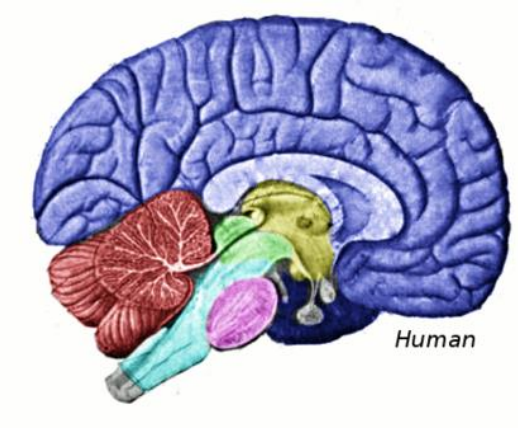
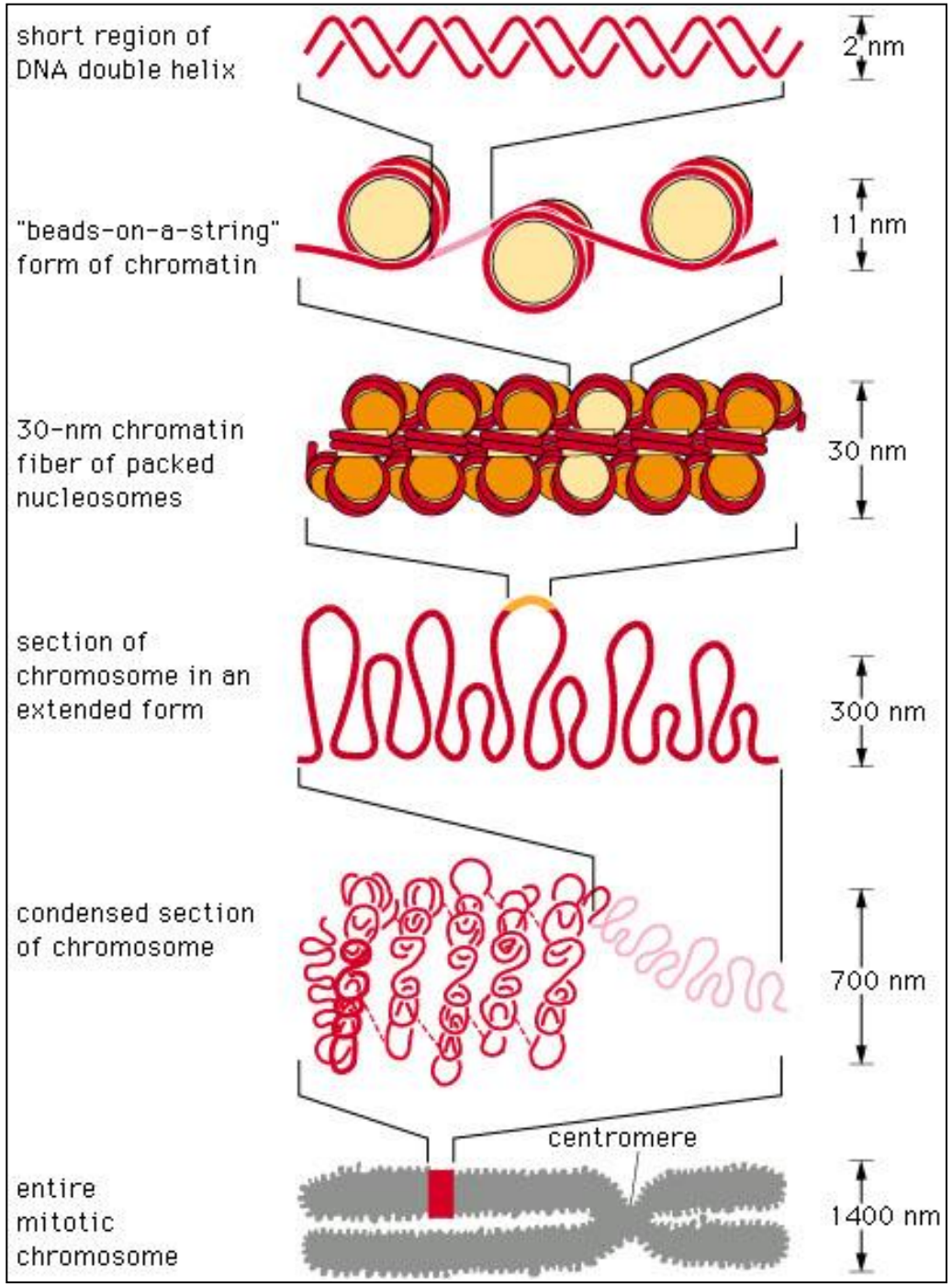


Image Source wikipedia

DNA packaging

- Why packaging
  - DNA is very long
  - Cell is very small
- Compression
  - Chromosome is 50,000 times shorter than extended DNA
- Using the DNA
  - Before a piece of DNA is used for anything, this compact structure must open locally
- Now emerging:
  - Role of accessibility
  - State in chromatin itself
  - Role of 3D interactions



Three types of epigenetic modifications

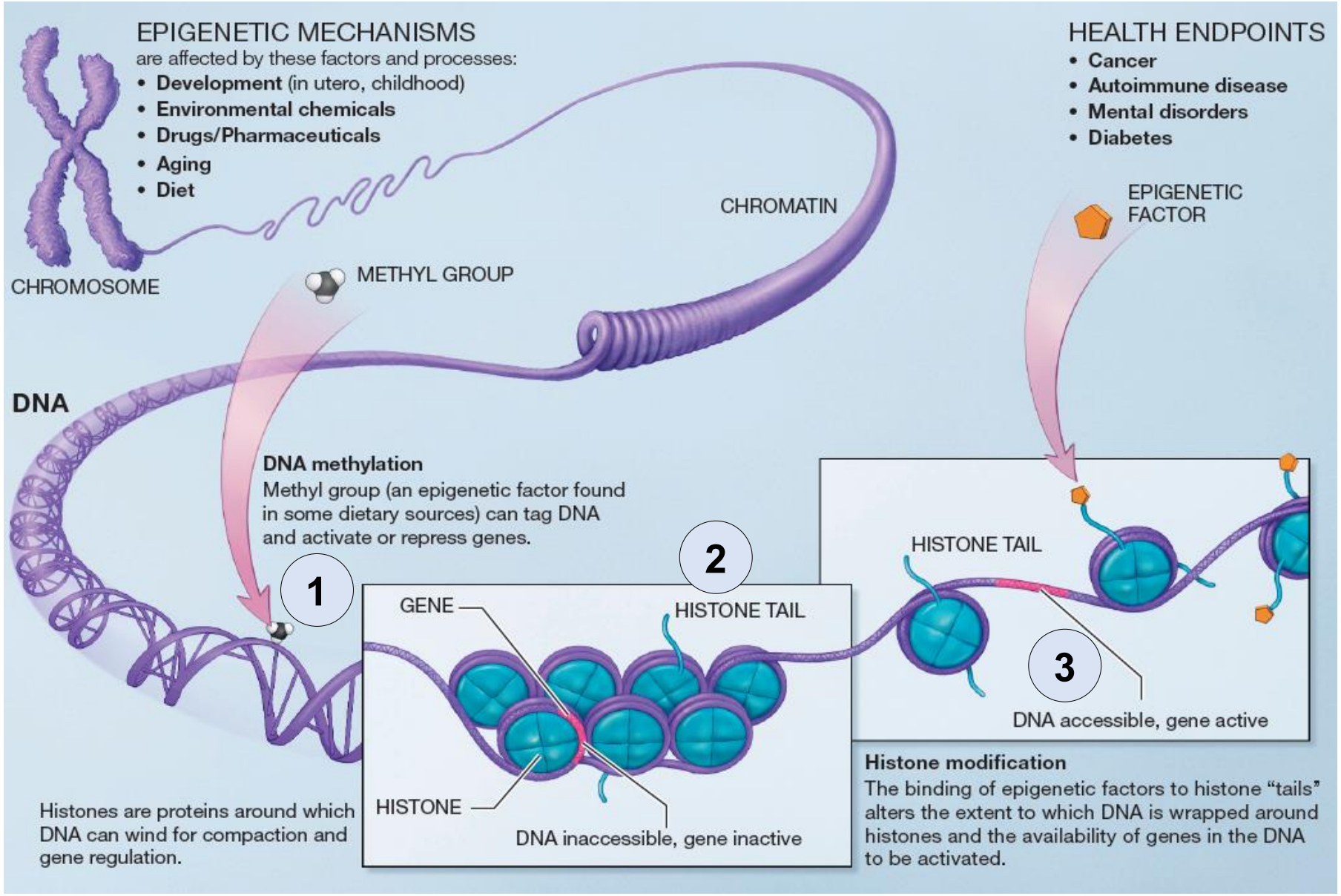
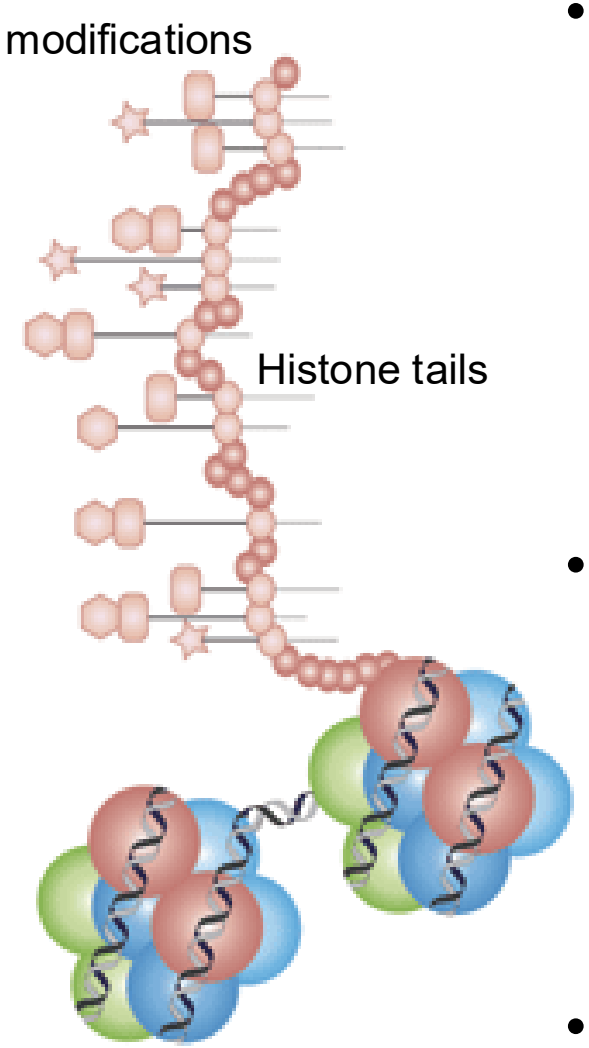


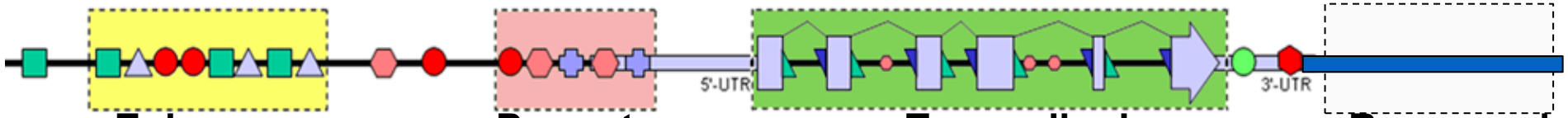
Image source: http://nihroadmap.nih.gov/epigenomics/

100s of histone tail modifications



- 100+ different histone modifications
  - Histone protein → H3/H4/H2A/H2B
  - AA residue → Lysine4(K4)/K36…
  - Chemical modification → Met/Pho/Ubi
  - Number → Me-Me-Me(me3)
  - Shorthand: H3K4me3, H2BK5ac
- In addition:
  - DNA modifications
  - Methyl-C in CpG / Methyl-Adenosine
  - Nucleosome positioning
  - DNA accessibility
- The constant struggle of gene regulation
  - TF/histone/nucleo/GFs/Chrom compete

Combinations of marks encode epigenomic state



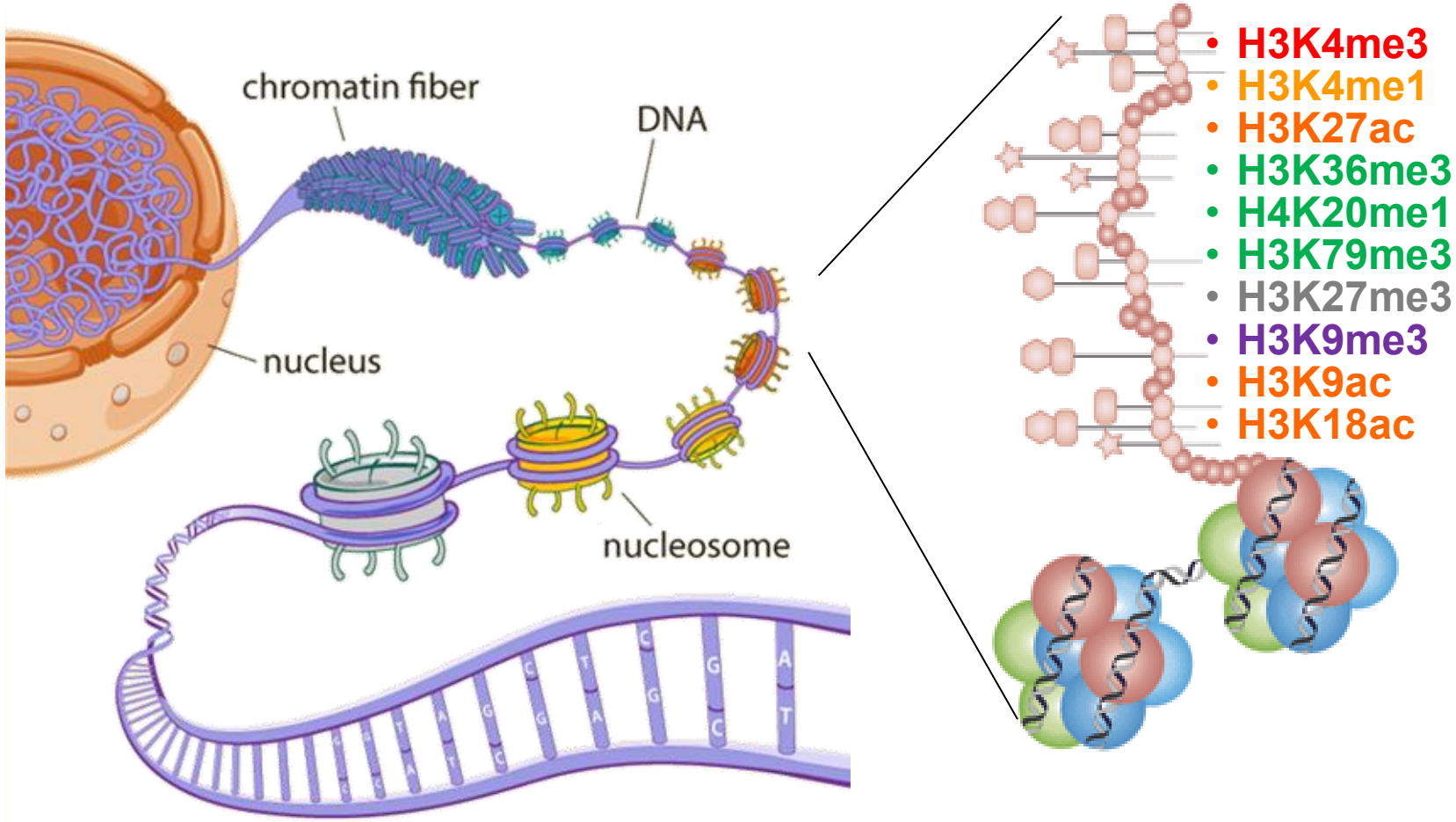
Enhancers
- H3K4me1
- H3K27ac
- DNase

Promoters
- H3K4me3
- H3K9ac
- DNase

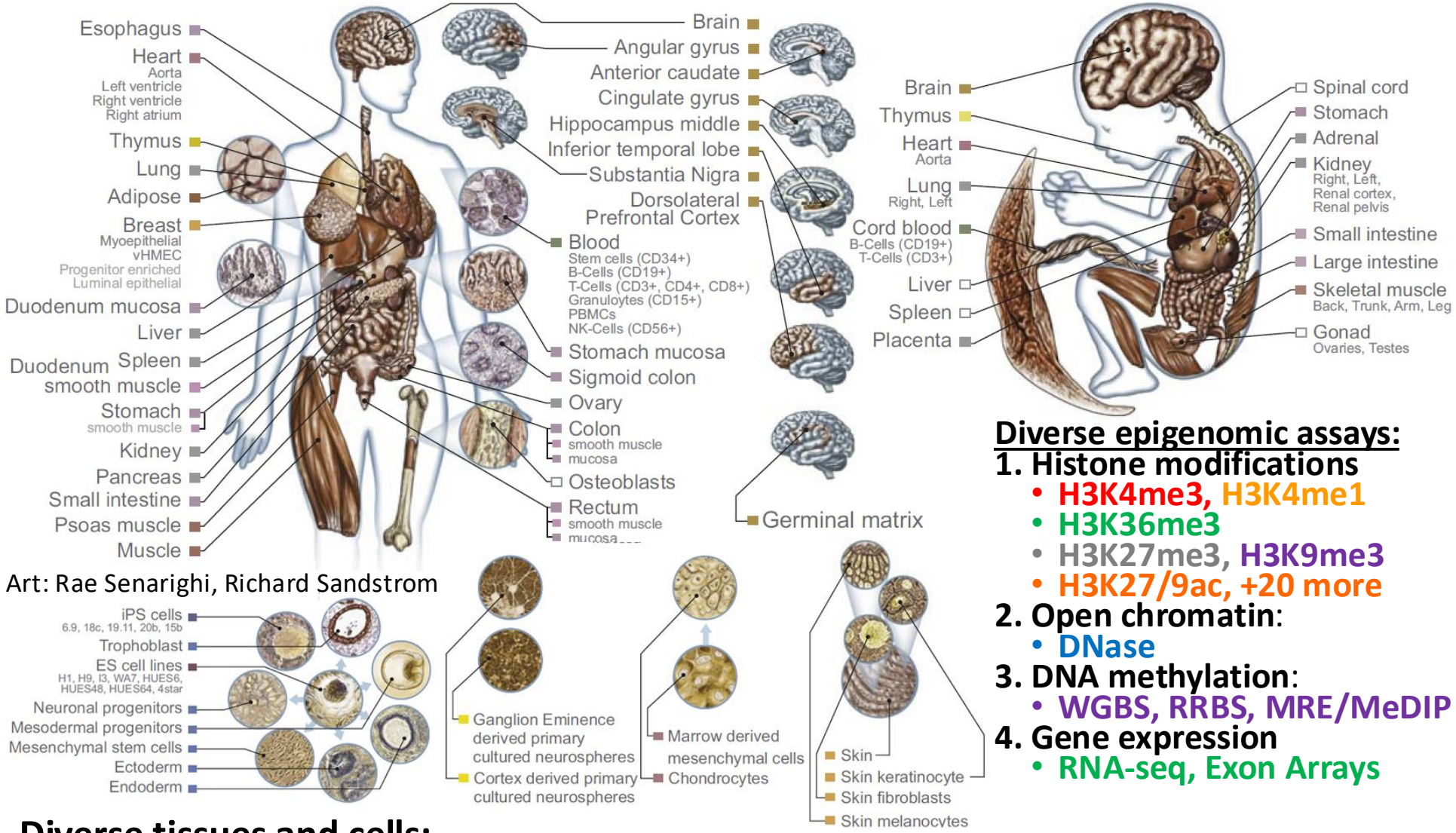
Transcribed
- H3K36me3
- H3K79me2
- H4K20me1

Repressed
- H3K9me3
- H3K27me3
- DNAmethyl

- 100s of known modifications, many new still emerging
- Systematic mapping using ChIP-, Bisulfite-, DNase-Seq



Epigenomics Roadmap across 100+ tissues/cell types



Art: Rae Senarighi, Richard Sandstrom

Diverse epigenomic assays:

1. **Histone modifications**
   - **H3K4me3, H3K4me1**
   - **H3K36me3**
   - **H3K27me3, H3K9me3**
   - **H3K27/9ac, +20 more**
2. **Open chromatin**:
   - **DNase**
3. **DNA methylation**:
   - **WGBS, RRBS, MRE/MeDIP**
4. **Gene expression**
   - **RNA-seq, Exon Arrays**

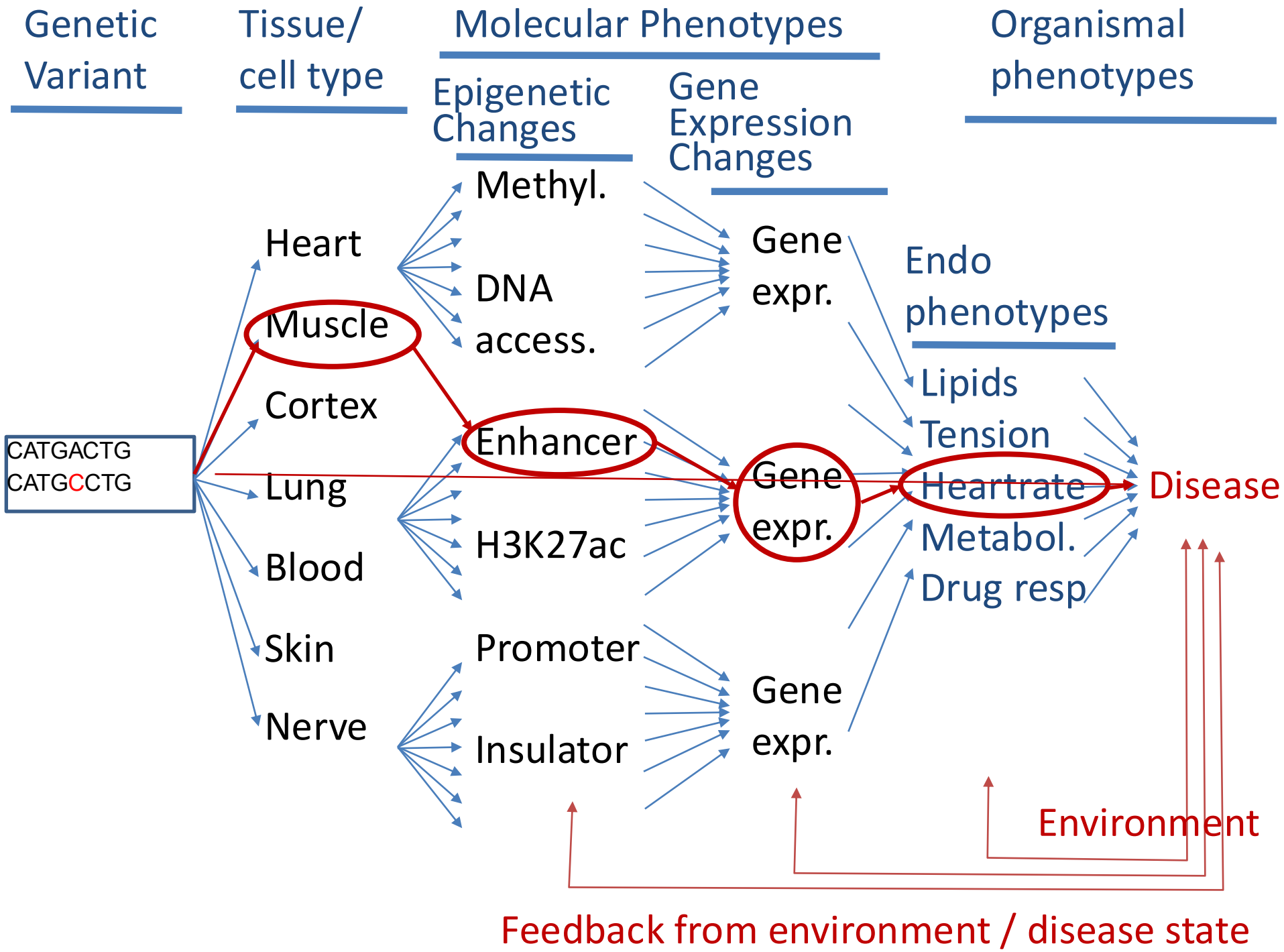
Diverse tissues and cells:

1. **Adult tissues and cells** (brain, muscle, heart, digestive, skin, adipose, lung, blood…)
2. **Fetal tissues** (brain, skeletal muscle, heart, digestive, lung, cord blood…)
3. **ES cells, iPS, differentiated cells** (meso/endo/ectoderm, neural, mesench, trophobl)

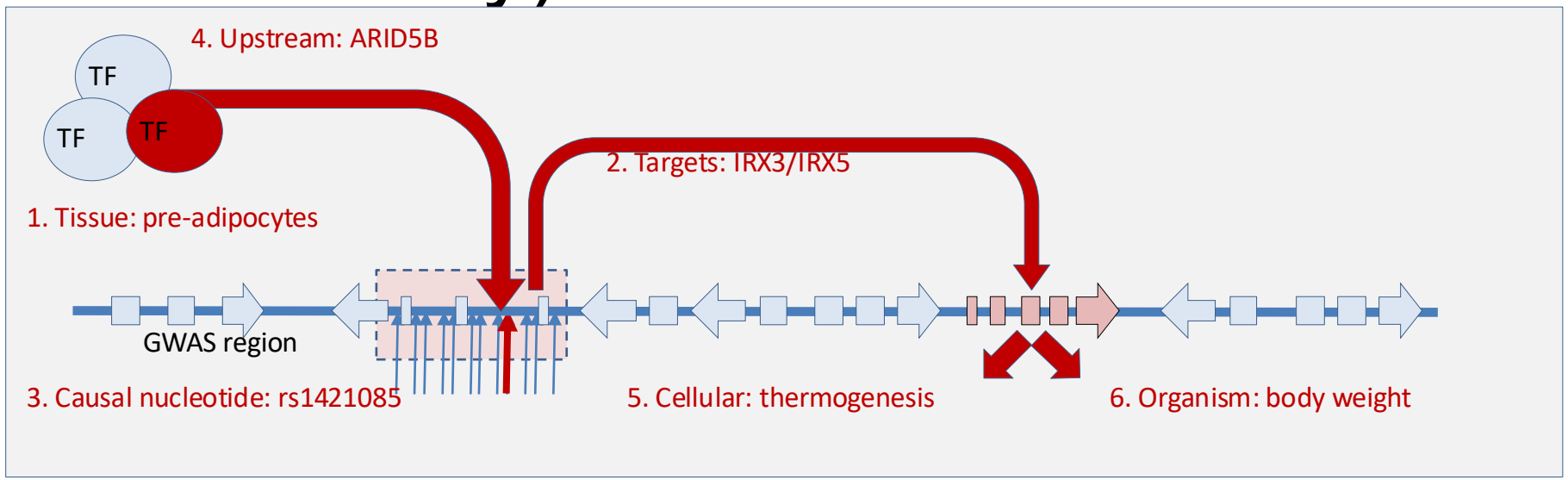
Ongoing epigenomic mapping projects

- Mapping multiple modifications
- In multiple cell types
- In multiple individuals
- In multiple species
- In multiple conditions
- With multiple antibodies
- Across the whole genome

- First waves published
- Lots more in pipeline
- Time for analysis!


Genetic Variant
Tissue/ cell type
Molecular Phenotypes
Epigenetic Changes
Gene Expression Changes
Organismal phenotypes
CATGACTG
CATGCCTG
Heart
Muscle
Cortex
Lung
Blood
Skin
Nerve
Methyl.
DNA access.
Enhancer
H3K27ac
Promoter
Insulator
Gene expr.
Gene expr.
Gene expr.
Endo phenotypes
Lipids
Tension
Heartrate
Metabol.
Drug resp
Disease
Environment
Feedback from environment / disease state

Dissection of disease mechanism (e.g. FTO/obesity)



1. Establish relevant **tissue/cell type: pre-adipocytes**
2. Establish downstream **target** gene(s): **IRX3 and IRX5**
3. Establishing **causal** nucleotide variant: **rs1421085**
4. Establish upstream **regulator** causality: **ARID5B**
5. Establish **cellular** phenotypic consequences: **thermogenesis**
6. Establish **organismal** phenotypic consequences: **body weight**

Today's lecture

1. Technologies for measuring regulatory activities
2. DNA Methylation (credit: Keegan Korthauer)
3. Multiomic Data Integration with Regulatory Genomics

Chromatin Immunoprecipitation + Sequencing

Goal: Identify where proteins (e.g., TFs) bind to DNA genome-wide

Experimental Protocols:

1. Cross-link proteins to DNA
2. Fragment chromatin
3. Immunoprecipitate with antibody
4. Sequence DNA fragments
5. Map reads to genome
6. **Call peaks**

ChIP-chip and ChIP-Seq technology overview

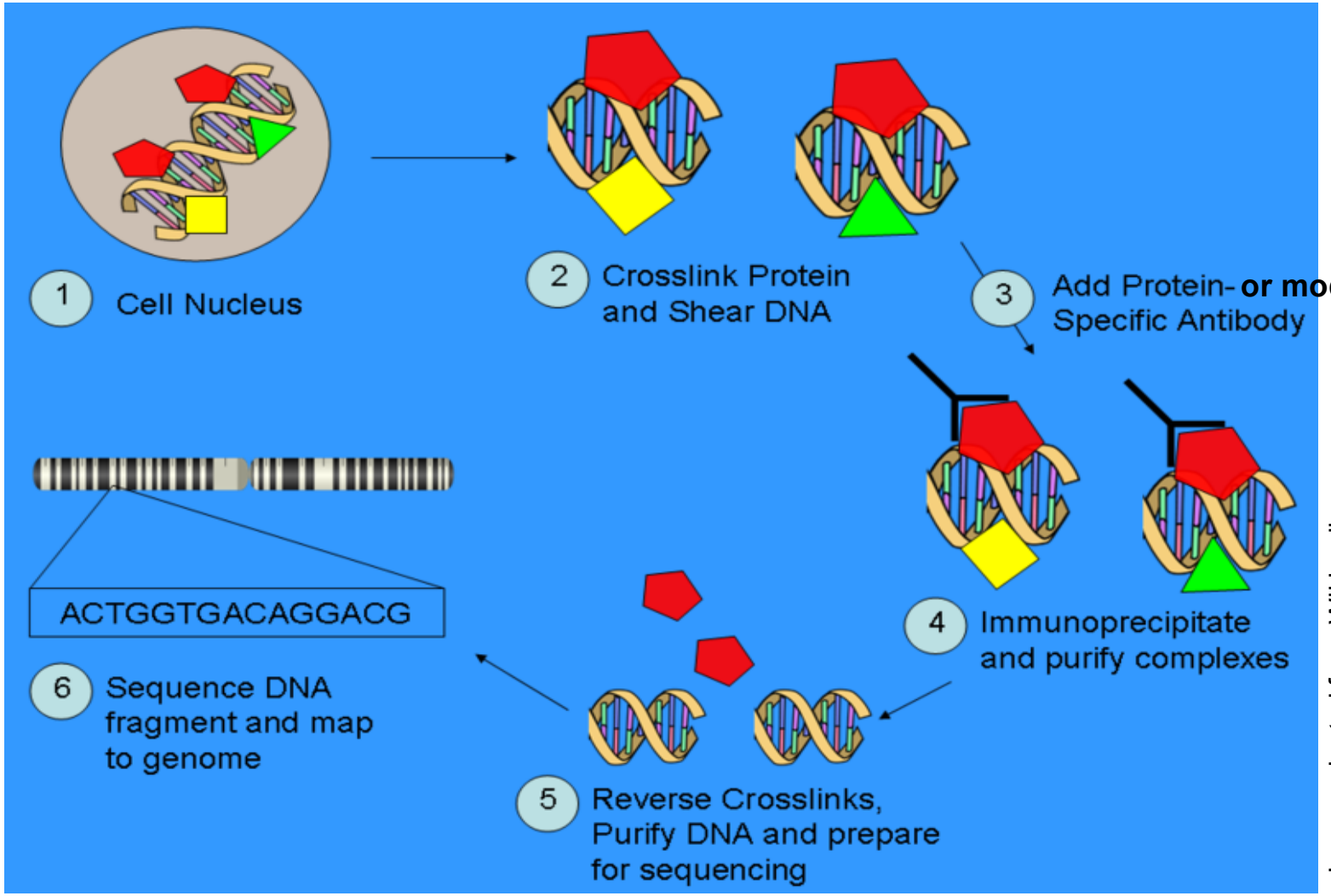


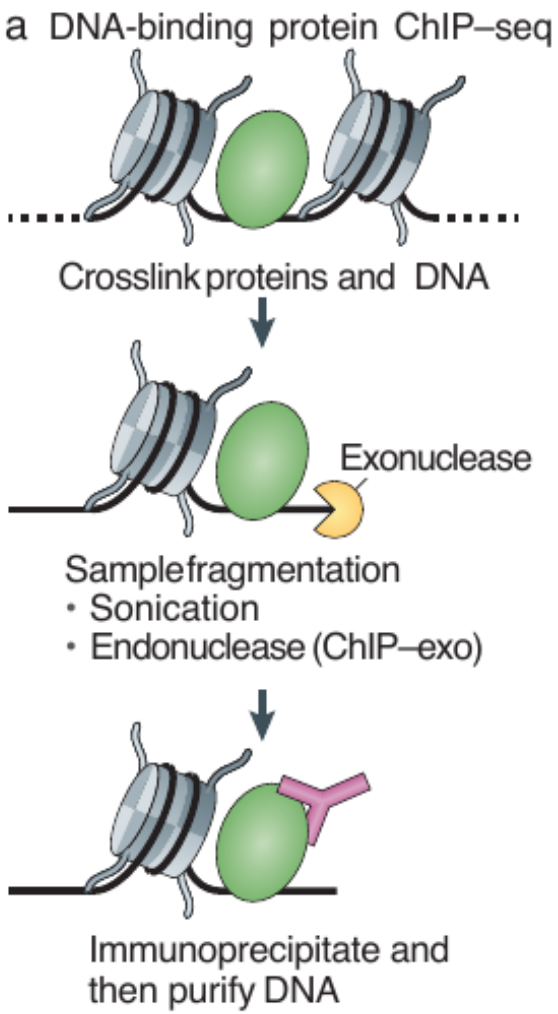
Image adapted from Wikipedia

Modification-specific antibodies → Chromatin Immuno-Precipitation

followed by: ChIP-chip: array hybridization

ChIP-Seq: Massively Parallel Next-gen Sequencing

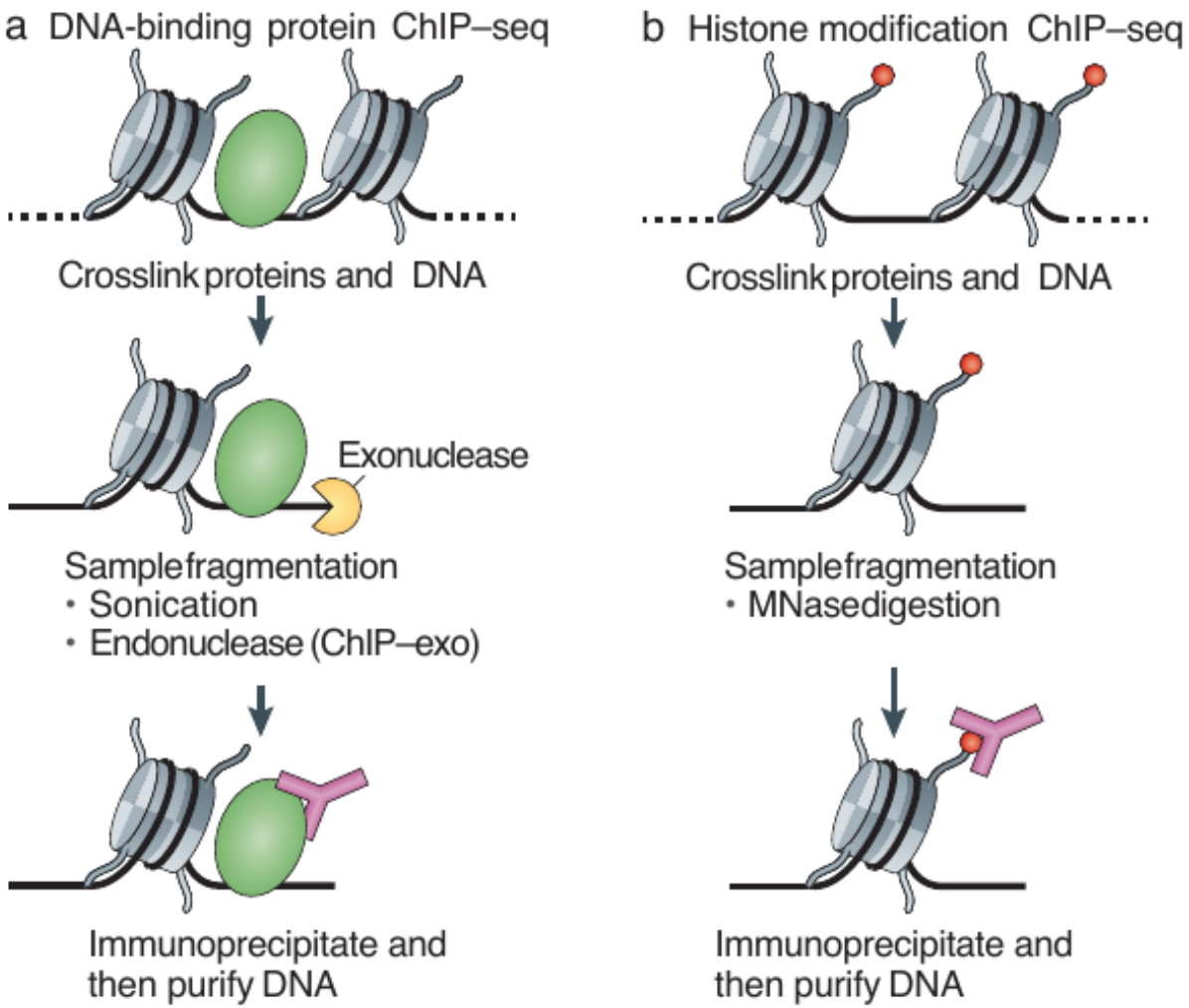
How ChIP-seq works



- Binding
- Sequencing

Furey (2012)

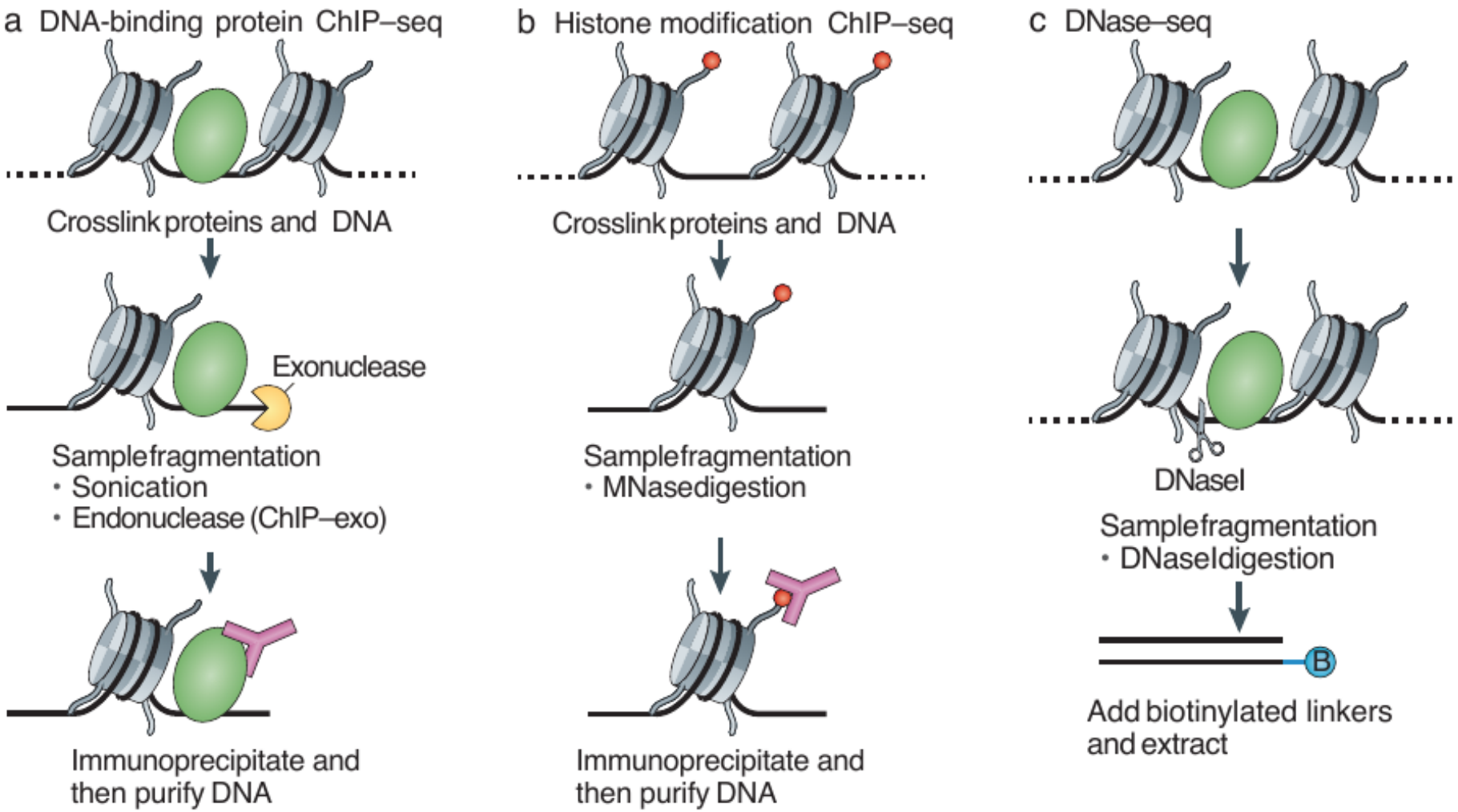
How ChIP-seq works



- Binding
- Sequencing

Furey (2012)

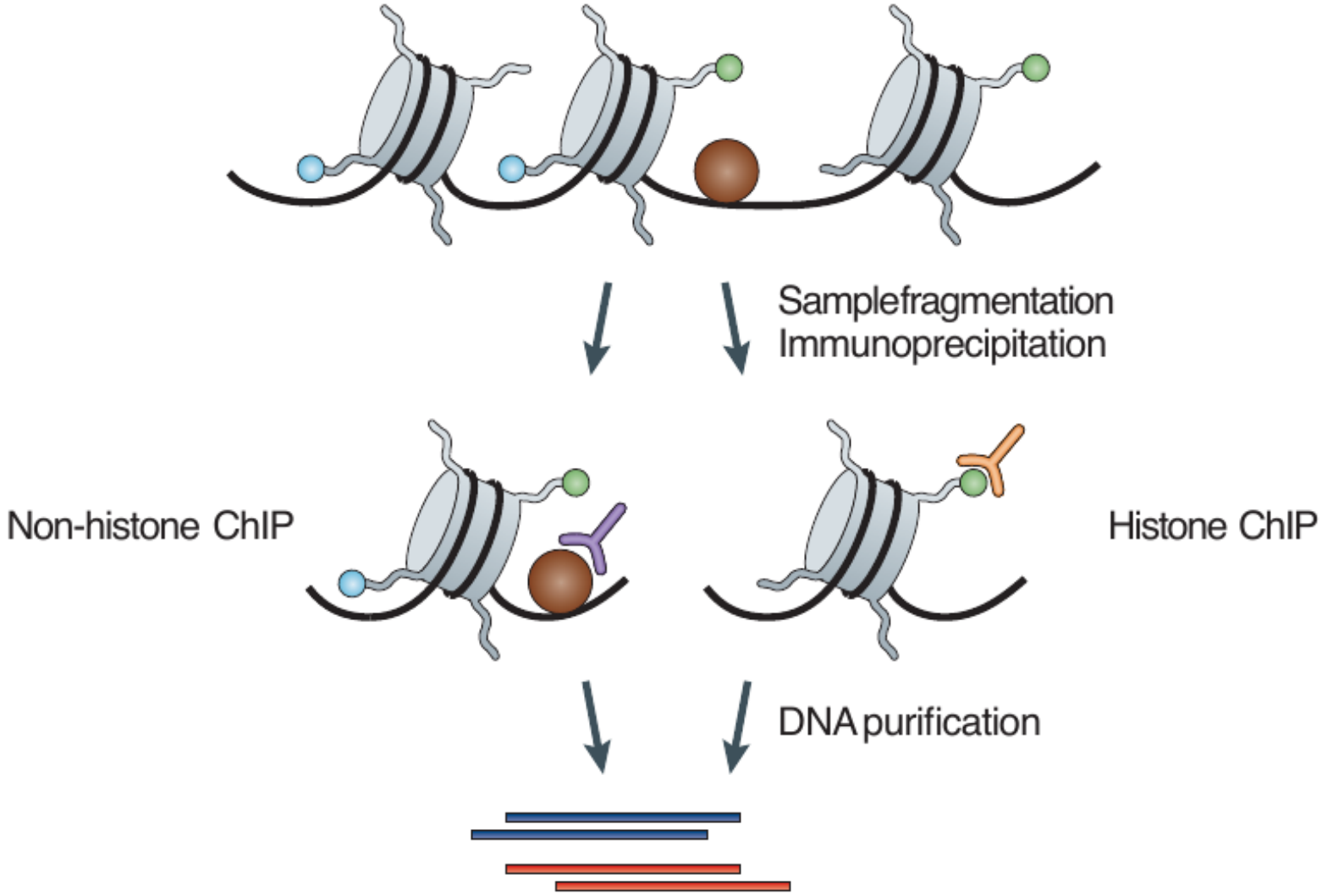
How ChIP-seq works



- Binding
- Sequencing

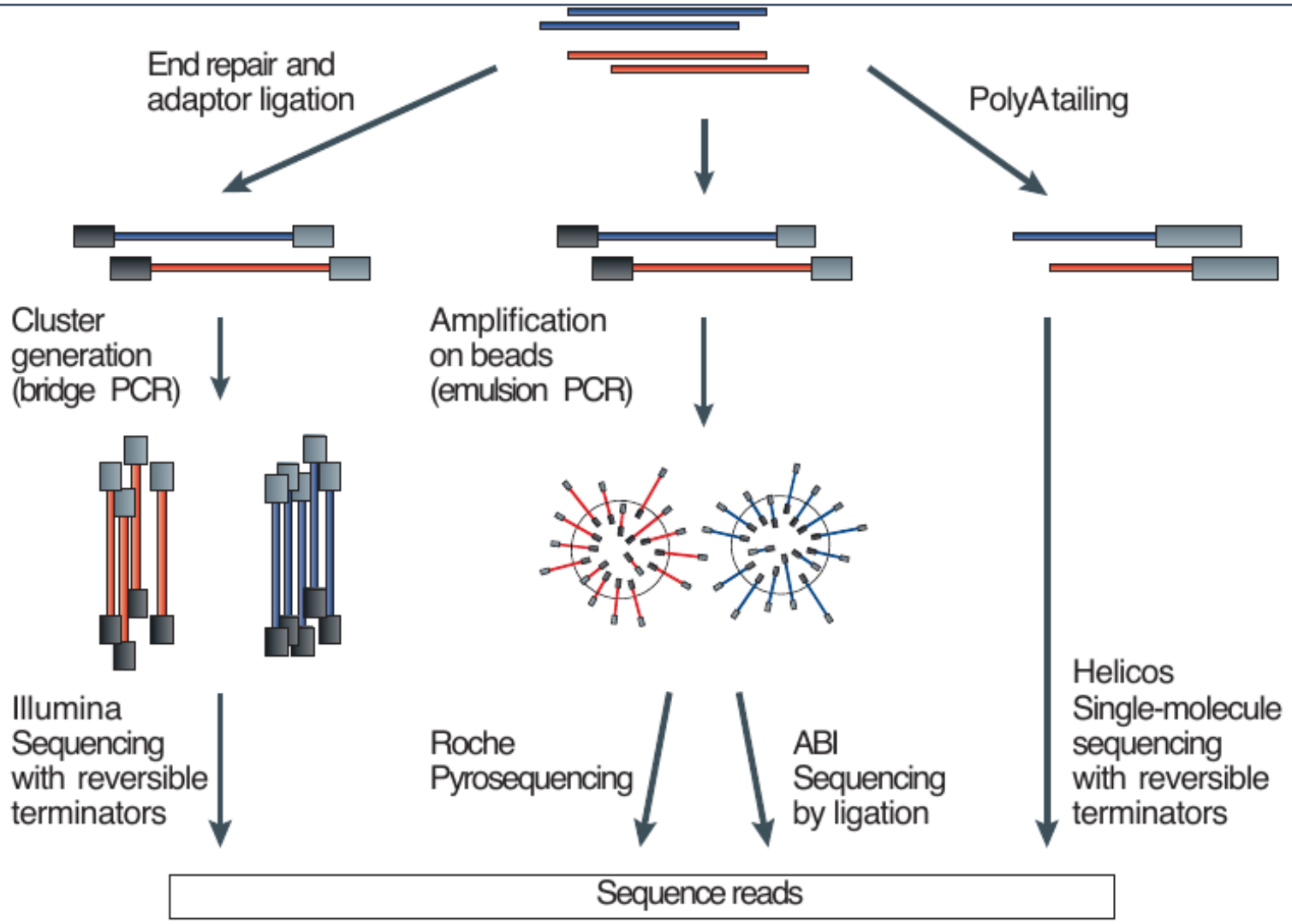
Furey (2012)

How ChIP-seq works



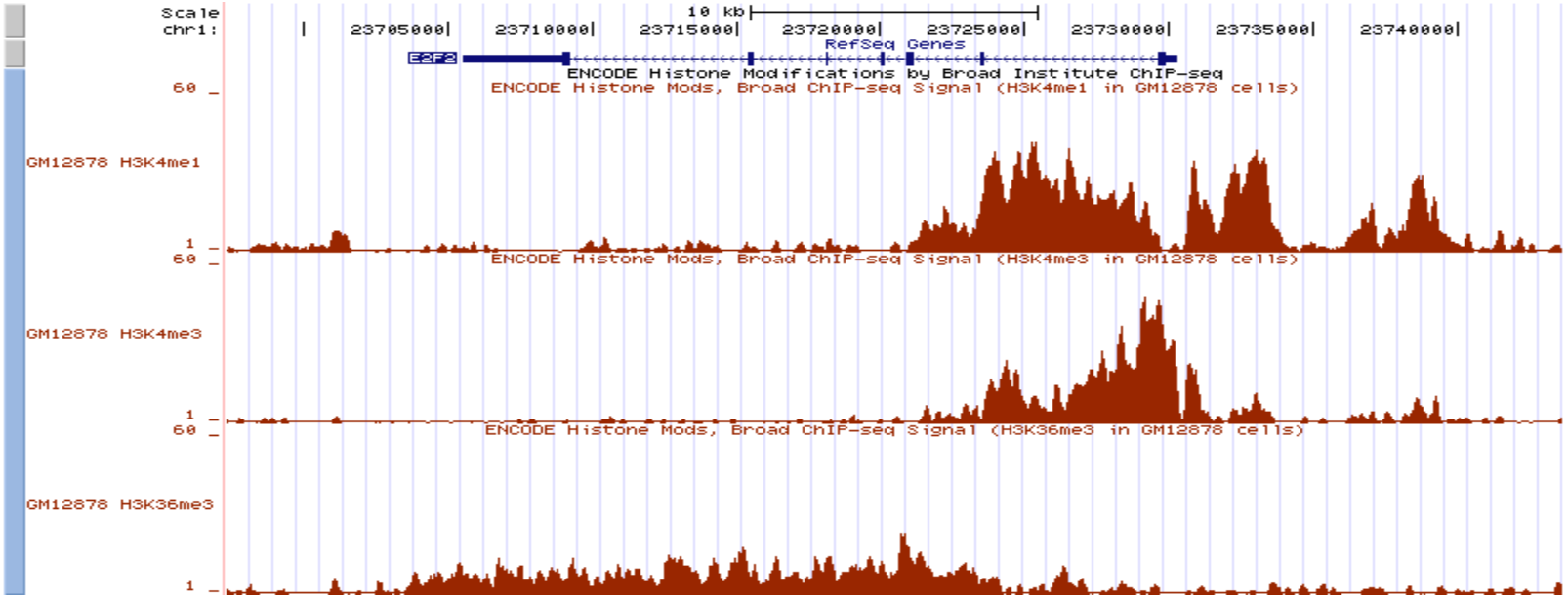
(Park 2009)

How ChIP-seq works



(Park 2009)

ChIP-Seq Histone Modifications: What the raw data looks like



- Each sequence tag is 30 base pairs long
- Tags are mapped to unique positions in the ~3 billion base reference genome
- Number of reads depends on sequencing depth. Typically on the order of 10 million mapped reads.

Quality control metrics

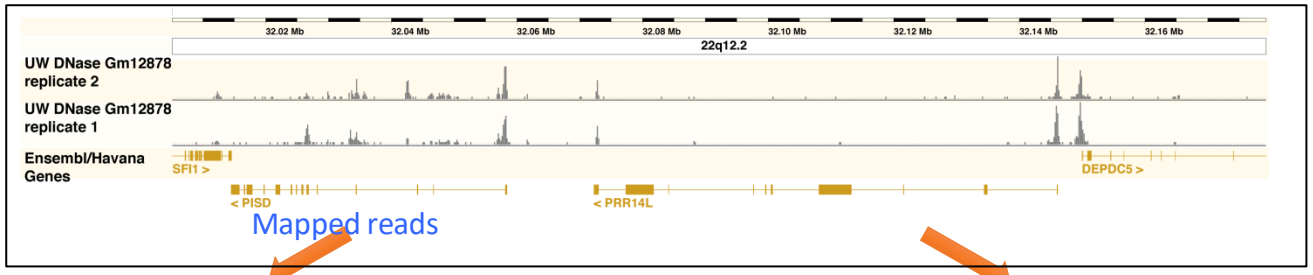
ChIP vs. Input DNA

Read quality

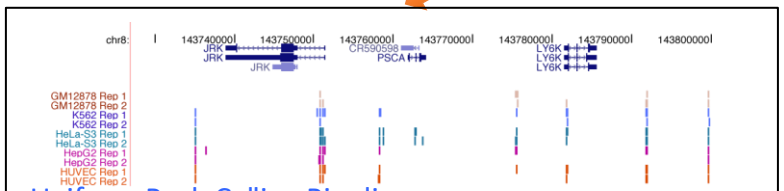
Mappability

Library complexity

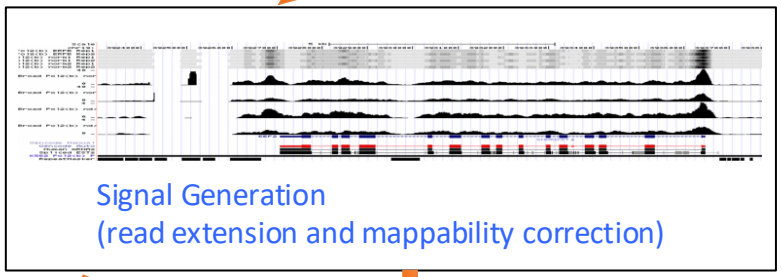
ENCODE uniform processing pipeline



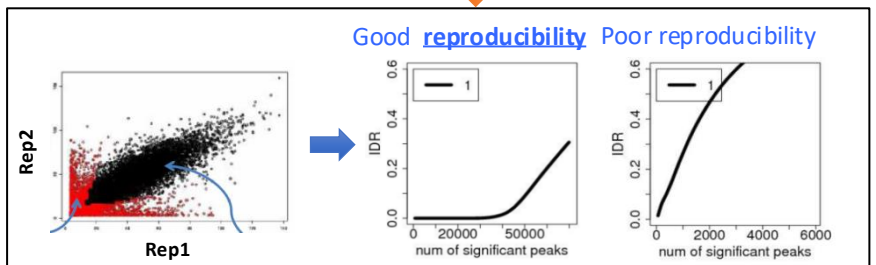
Mapped reads



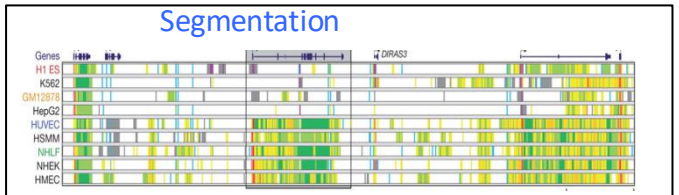
Uniform Peak Calling Pipeline



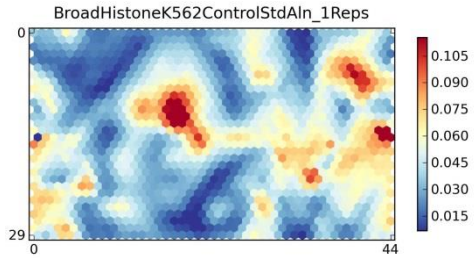
Signal Generation
(read extension and mappability correction)



IDR Processing, Quality control and Blacklist Filtering

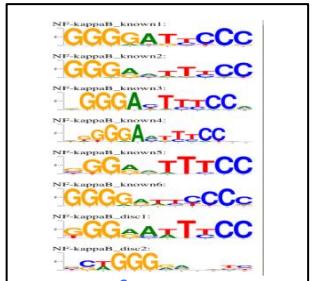


Segmentation

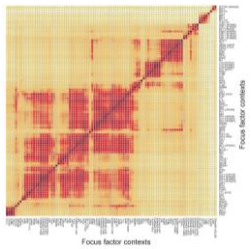


ChromHMM/Segway

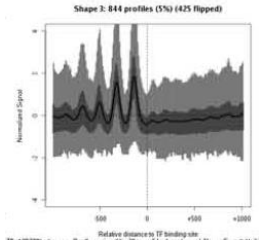
Self Organising Maps



Motif Discovery

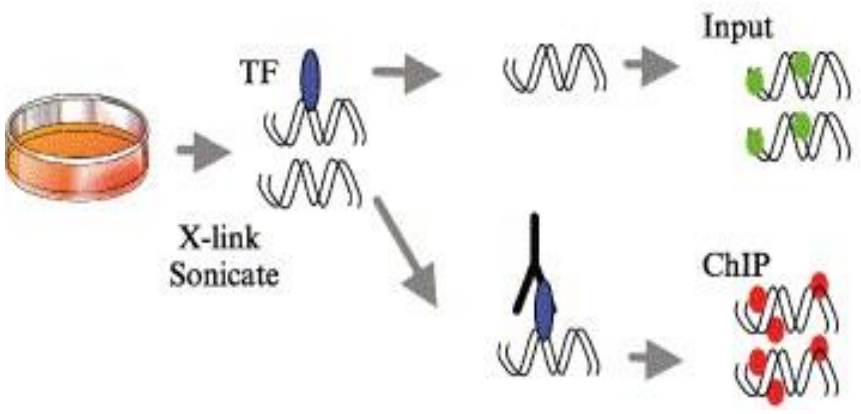


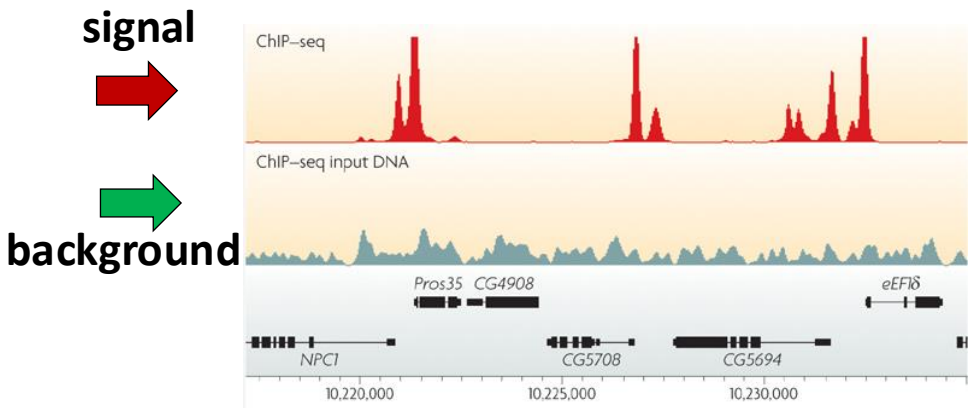
Co-association analysis



Signal Aggregation over elements

QC1: Use of input DNA as control dataset



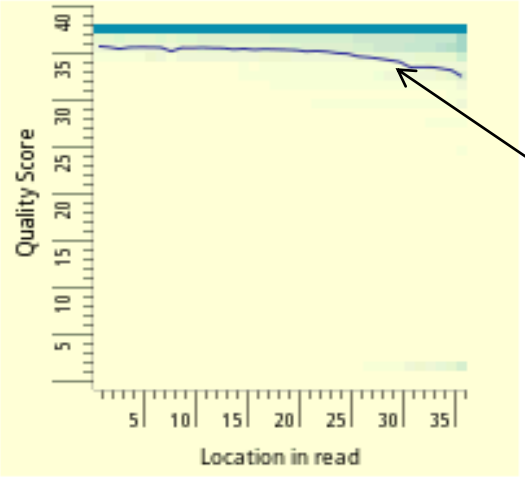


- **Challenge:**
  - Even without antibody: Reads are **<u>not</u>** uniformly scattered
- **Sources of bias in input dataset scatter:**
  - Non-uniform fragmentation of the genome
  - Open chromatin fragmented more easily than closed regions
  - Repetitive sequences over-collapsed in the assembled genome.
- **How to control for these biases:**
  - Remove portion of DNA sample before ChIP step
  - Carry out control experiment without an antibody (input DNA)
  - Fragment input DNA, sequence reads, map, use as background

QC2: Read-level sequencing quality score Q>10

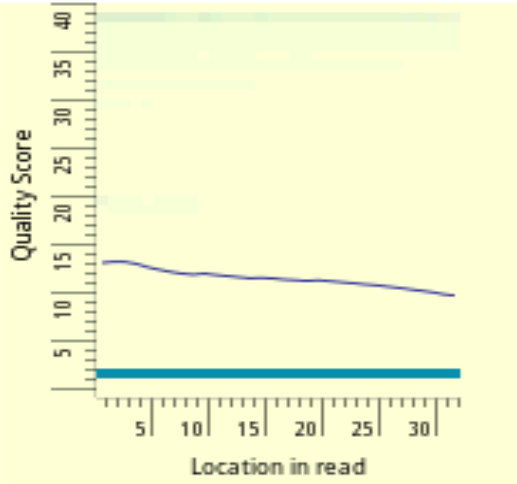
Read quality histograms

High quality reads



average base score per position

Low quality reads



- Each column is a color-coded histogram
- Encodes fraction of all mapped reads that have base score Q (y-axis) at each position (x-axis)
- Darker blue = higher density

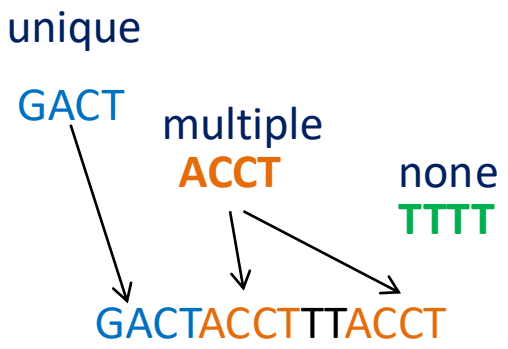
- Read quality tends to drop towards the ends of reads
- Low average per base score implies greater probabilty of mismappings.

- Typically, reject reads whose average score Q < 10

QC3: Fraction of short reads mapped >50%

Reads can map to:

- exactly one location (uniquely mapping)
- multiple locations (repetitive or multi-mapping)
- no locations (unmappable)



Dealing with multiply-mapping reads:

- Conservative approach: do not assign to any location
- Probabilistic approach: assign fractionally to all locations
- Pair-end approach: use paired end read to resolve ambiguities in repeat reads

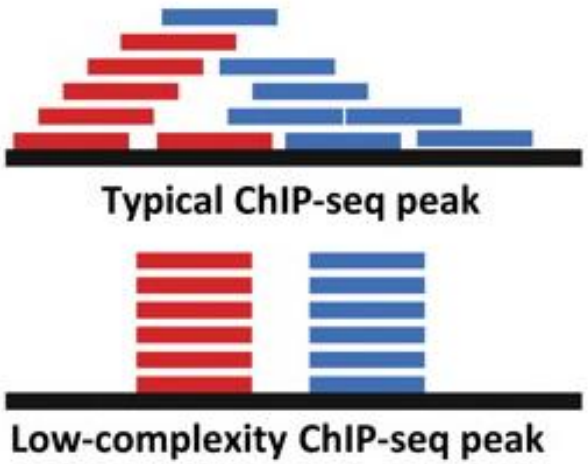
Absence of reads in a region could be due to:

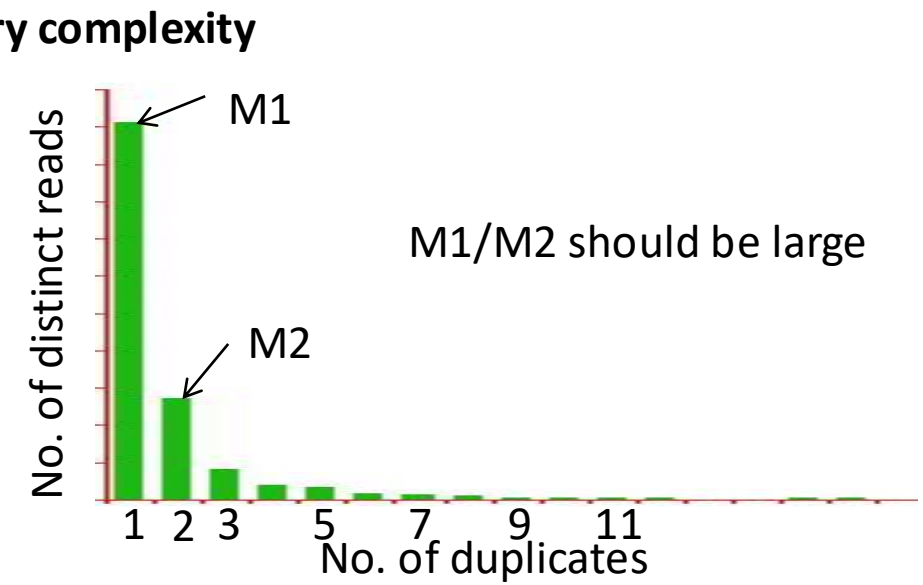
- No assembly coverage in that region (e.g. peri-centromeric region)
- Too many reads mapping to this location (e.g. repetitive element)
- No activity observed in this location (e.g. inactive / quiescent / dead regions)

Dealing with mappability biases:

- ‘Black-listed’ regions, promiscuous across many datasets
- ‘White-listed’ regions, for which at least some dataset has unique reads

QC4: Library complexity: non-redundant fraction





How many distinct uniquely mapping read? How many duplicates?

If your sample does not contain sufficient DNA and/or you over-sequence, you will simply be repeatedly sequencing PCR duplicates of a restricted pool of distinct DNA fragments. This is known a **low-complexity library** and is not desirable.

- **Histogram of no. of duplicates**
- Non-redundant fraction (NRF) = $\frac{\text{No. of 'distinct' unique-mapping reads}}{\text{No. of unique-mapping reads}}$
- NRF should be > 0.8 when 10M < #reads < 80M unique-mapping reads

Cross-correlation analysis

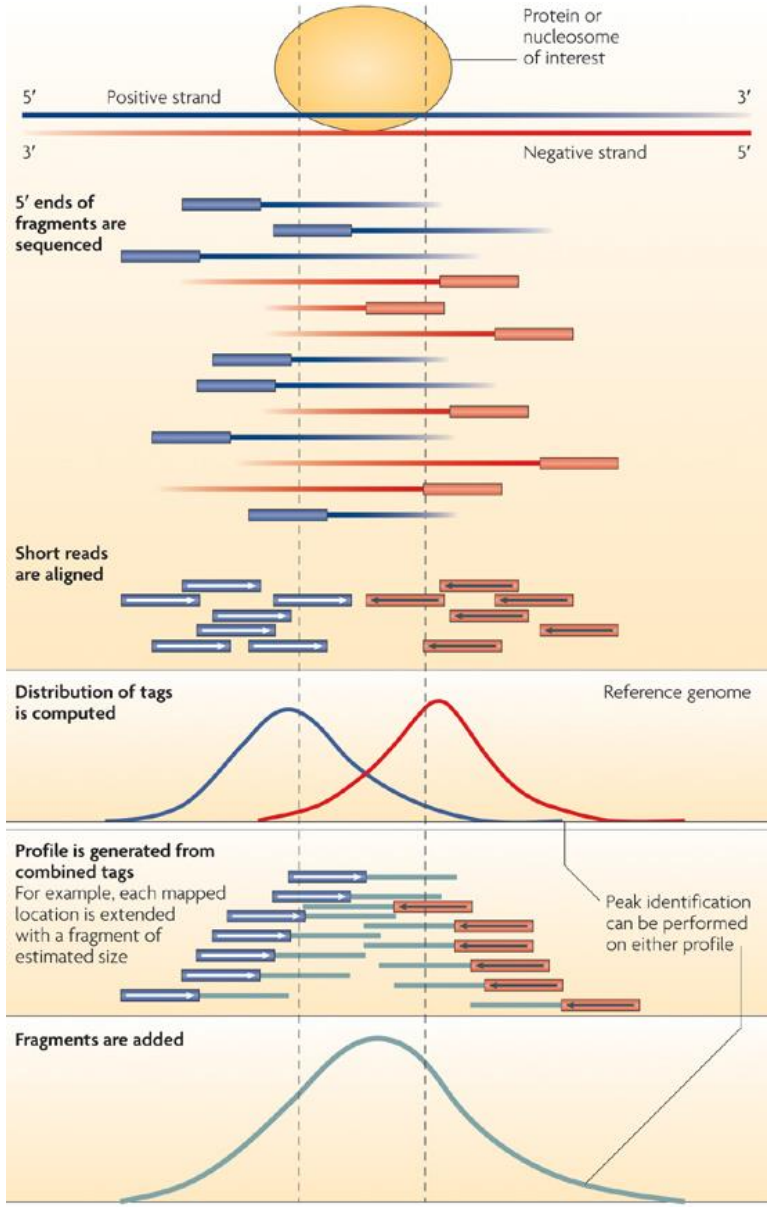
Exploiting forward and reverse reads

Fragment-length peak

Phantom read-length peak

ChIP-seq: exploiting forward and reverse reads

(Chromatin immunoprecipitation followed by sequencing)



Nature Reviews | Genetics

Multiple IP fragments are obtained corresponding to each binding event

Ends of the fragments are sequenced i.e. "Short-reads/tags"

- Typically ~36 bp, 50 bp, 76 bp or 101 bp

Single-end (SE) sequencing

- Randomly sequence one of the ends of each fragment

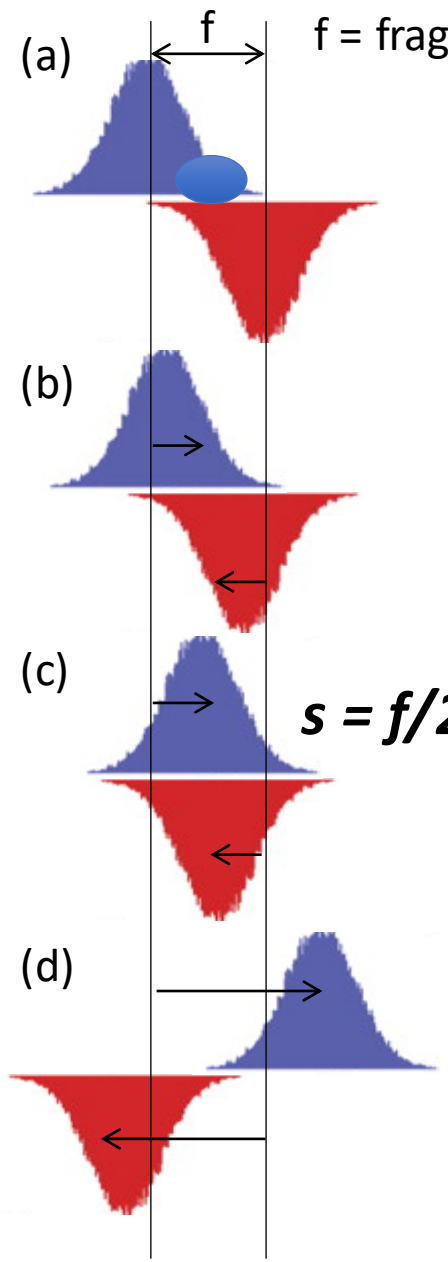
Paired-end (PE) sequencing

- sequence both ends of each fragment

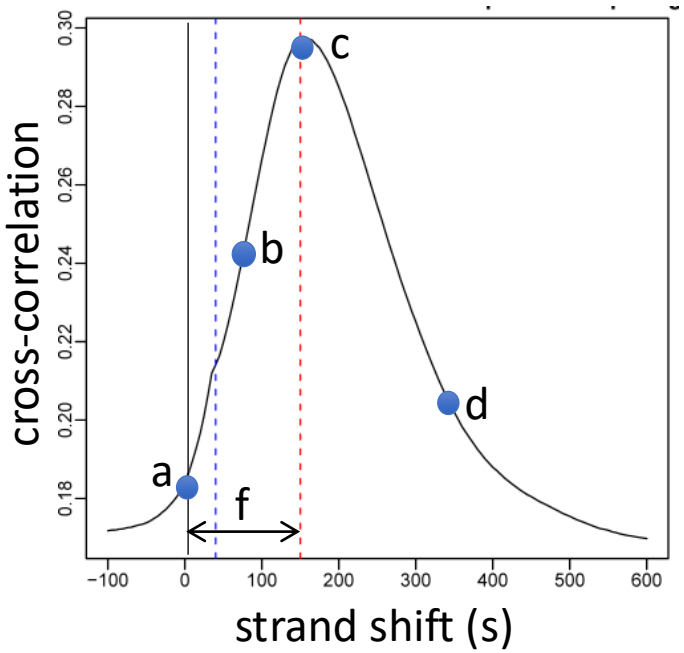
Canonical "stranded mirror distribution of short-reads" after mapping reads to genome

- Heaps of reads on the **+ strand** and **– strand** separated by a distance ~= fragment length

Strand cross-correlation (CC) analysis



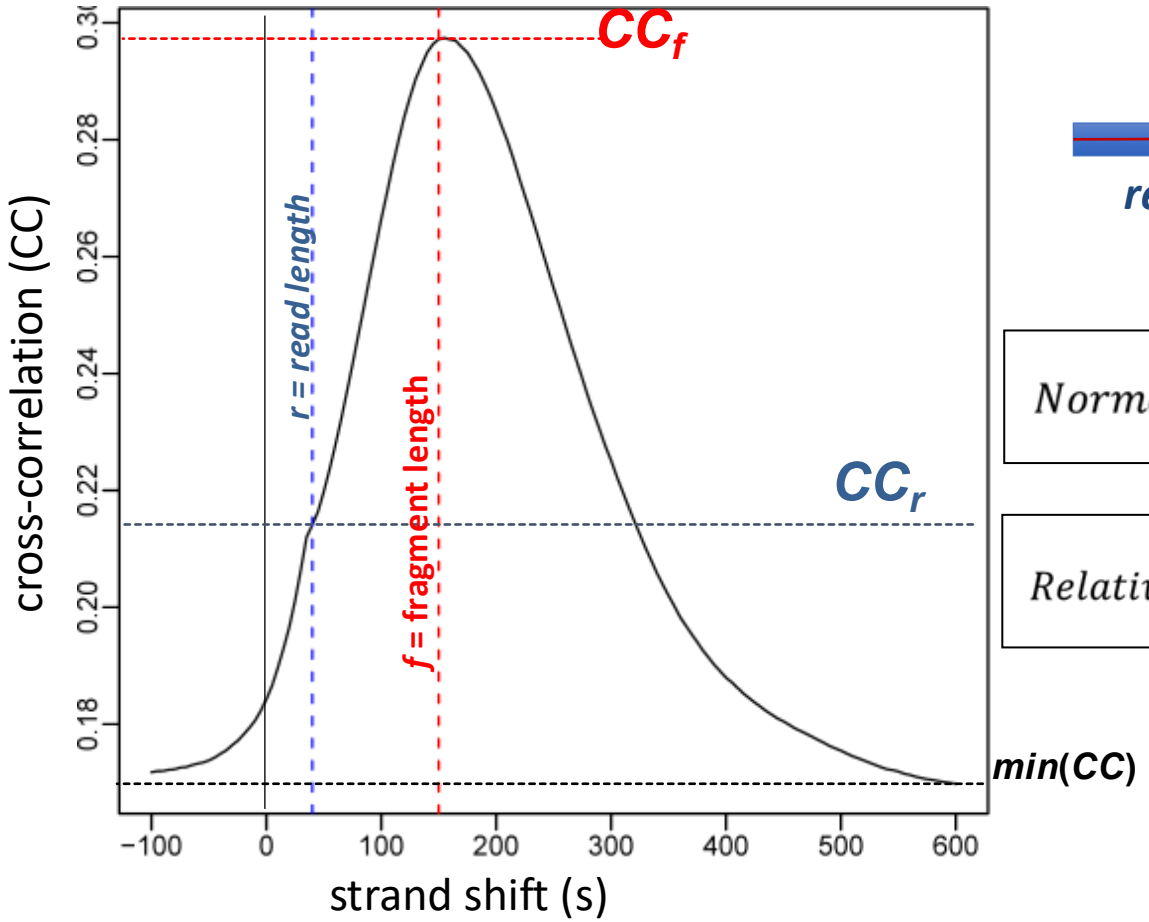
$s = f/2 + f/2$



1. Calculate forward and reverse strand signals
2. Shift both by specified offset towards each other
3. Calculate correlation of two signals at that shift
4. Correlation peaks at **fragment length** offset ***f***

f is the length at which ChIP DNA is fragmented

Cross-correlation at *read* vs. *fragment* length

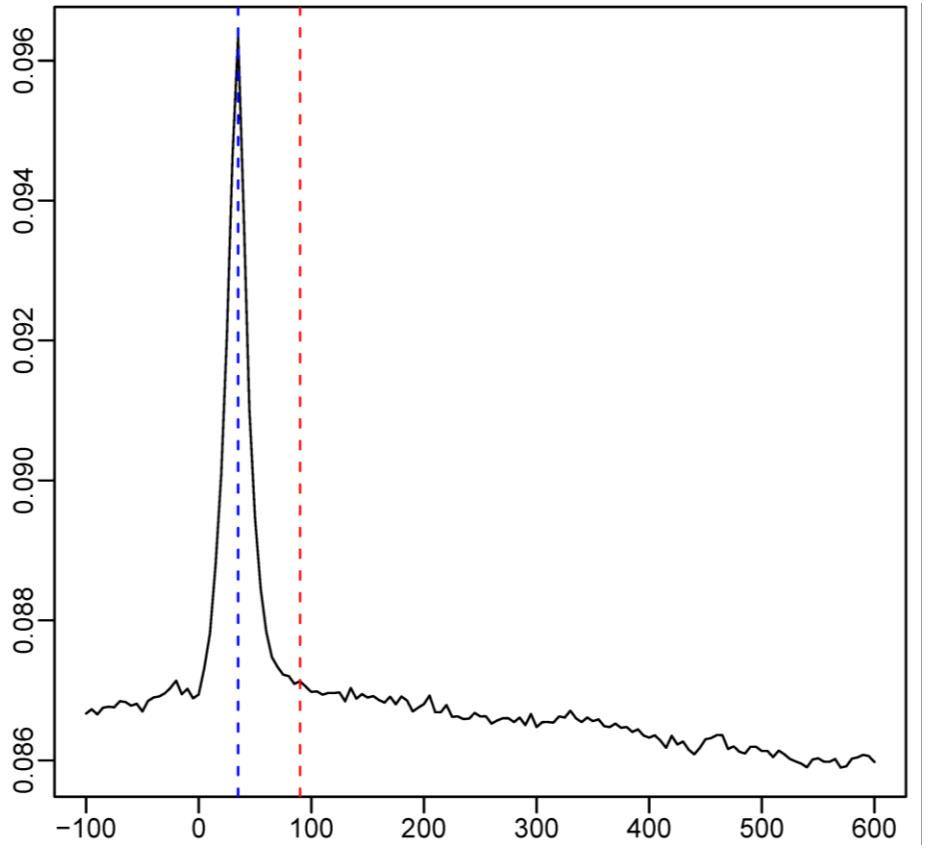


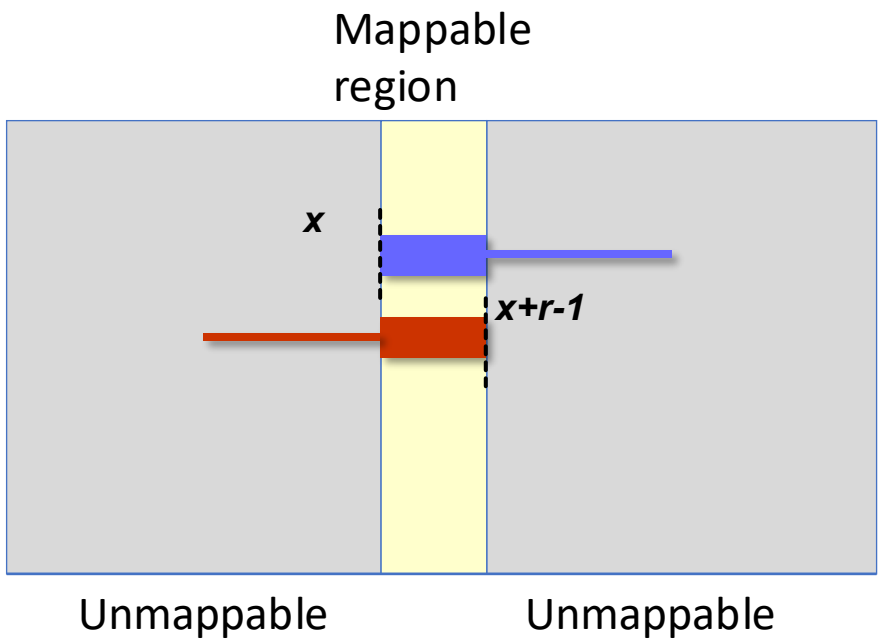
$$Normalized\ strand\ CC(NSC) = \frac{CC_f}{\min(CC)}$$

$$Relative\ strand\ CC(RSC) = \frac{(CC_f - \min(CC))}{(CC_r - \min(CC))}$$

- Sign of a good dataset:
  - High absolute cross-correlation at *fragment* length (NSC)
  - High *fragment* length CC relative to *read* length CC (RSC)

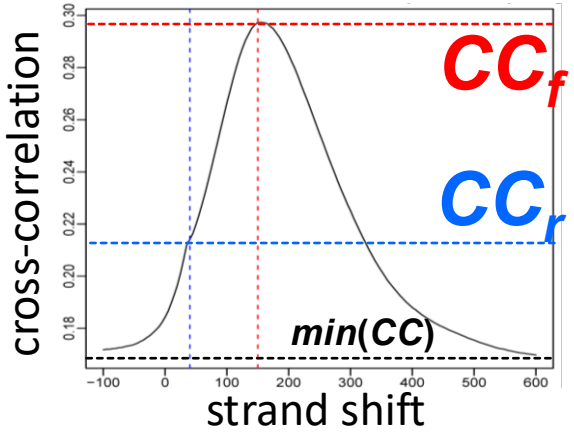
Where does *read* cross-correlation come from?



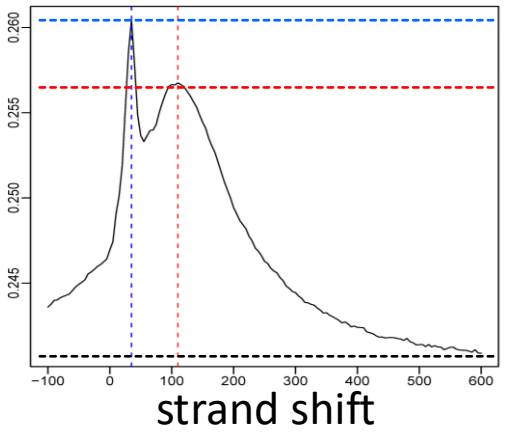


- Input dataset (no ChIP) shows 'phantom' peak at ***read*** length only
- Due to read mappability:
  - If position 'x' is uniquely mappable on + strand
  - Then position 'x+r-1' is uniquely mappable on – strand
- *Fragment*-length peak should always dominate the read-length peak

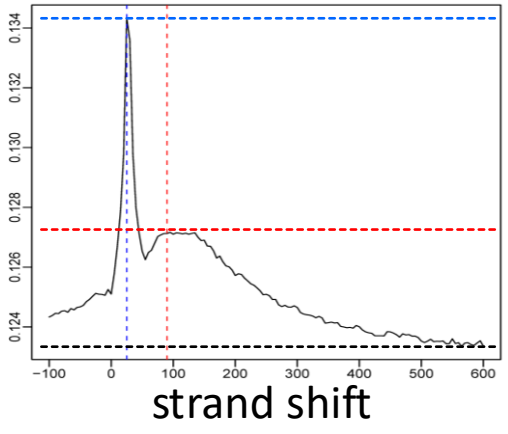
Example of good, medium, bad CC datasets



Highly quality



Medium quality



Low quality

$$Normalized\ strand\ CC(NSC) = \frac{CC_f}{\min(CC)}$$

$$Relative\ strand\ CC(RSC) = \frac{(CC_f - \min(CC))}{(CC_r - \min(CC))}$$

For highly enriched datasets, fragment length cross-correlation peak should be able to beat read-length phantom peak

RSC should be > 1

Peak calling: signal vs. background

Key idea: Find genomic regions with more reads than expected by chance

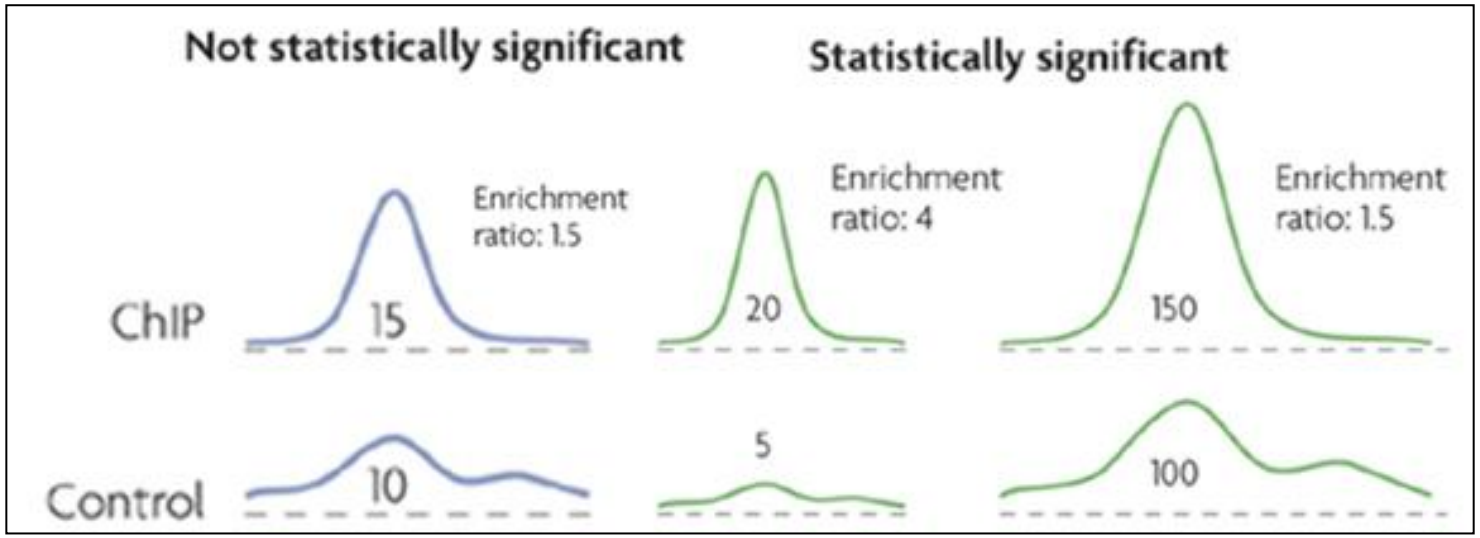
- Divide genome into bins (e.g., 200 bp)
- Count reads in each bin
- Compare ChIP vs Input (control)
- Statistical test: Is enrichment significant?

Common approaches:

- **MACS2**: Poisson model, local background
- **normR**: Binomial mixture model
- **DiffBind**: Differential binding analysis

Multiple testing: FDR correction across all bins!

Peak calling: detect regions of enrichment



Goal: Transform read counts into **normalized intensity signal**

Steps:

1. Estimate fragment-length *f* using strand cross-correlation analysis
2. Extend each read from 5' to 3' direction to fragment length *f*
3. Sum intensity for each base in 'extended reads' from both strands
4. Perform same operation on input-DNA control data (correct for sequencing depth differences)
5. Calculate enrichment ratio value for every position in the genome

Result: Enrichment fold difference for ChIP / control signal

Peak calling thresholds

Poisson p-value thresholds

- Read count model: Locally-adjusted' Poisson distribution

$$P(count = x) = \frac{\lambda_{local}^{x} \exp(-\lambda_{local})}{x!}$$

- $\lambda_{local} = \max(\lambda_{BG}, [\lambda_{1k},] \lambda_{5k}, \lambda_{10k})$ estimated from control data
- Poisson *p*-value = $P(count \geq x)$
- *q*-value : Multiple hypothesis correction

Peaks: Genomic locations that pass a user-defined *p*-value (e.g. 1e-5) or *q*-value (e.g. 0.01) threshold

Empirical False discovery rates

- Swap ChIP and input-DNA tracks
- Recompute *p*-values
- At each *p*-value, eFDR = Number of control peaks / Number of ChIPpeaks
- Use an FDR threshold to call peaks

Issues with peak calling thresholds

Cannot set a universal threshold for empirical FDRs and p-values

- Depends on ChIP and input sequencing depth
- Depends on binding ubiquity of factor
- Stronger antibodies get an advantage

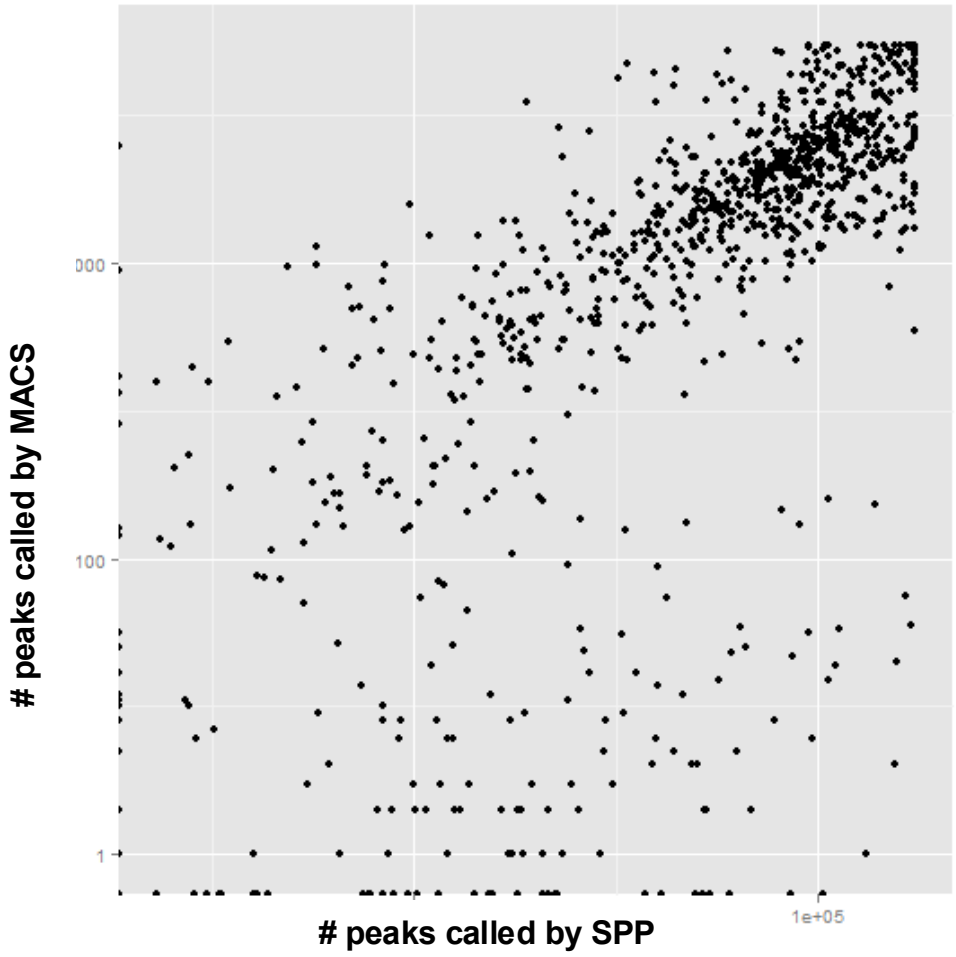
FDRs quite unstable

- Small changes in threshold => massive changes in peak numbers

Difficult to compare results across peak callers with a fixed threshold

- Different methods to compute eFDR or q-values

(at FDR = 1% cutoff)



ChIP-seq Peak Calling Exercise

```r
library(AnnotationHub)
library(rtracklayer)

data_dir <- "../data/chipseq/"
dir.create(data_dir, showWarnings = FALSE, recursive = TRUE)

chr <- "chr21" # Focus on chr21 to keep data manageable

# Download ENCODE CTCF ChIP peaks
peaks_file <- paste0(data_dir, "ctcf_gm12878_peaks.rds")
if.needed(peaks_file, {
    cat("Downloading ENCODE CTCF peaks...\n")
    ah <- AnnotationHub()
    query_peaks <- query(ah, c("CTCF", "GM12878", "narrowPeak", "ENCODE"))
    chip_peaks_all <- query_peaks[[1]]
    chip_peaks <- chip_peaks_all[seqnames(chip_peaks_all) == chr]
    saveRDS(chip_peaks, file = peaks_file)
})
chip_peaks <- readRDS(peaks_file)
```

ChIP-seq Peak Calling Exercise

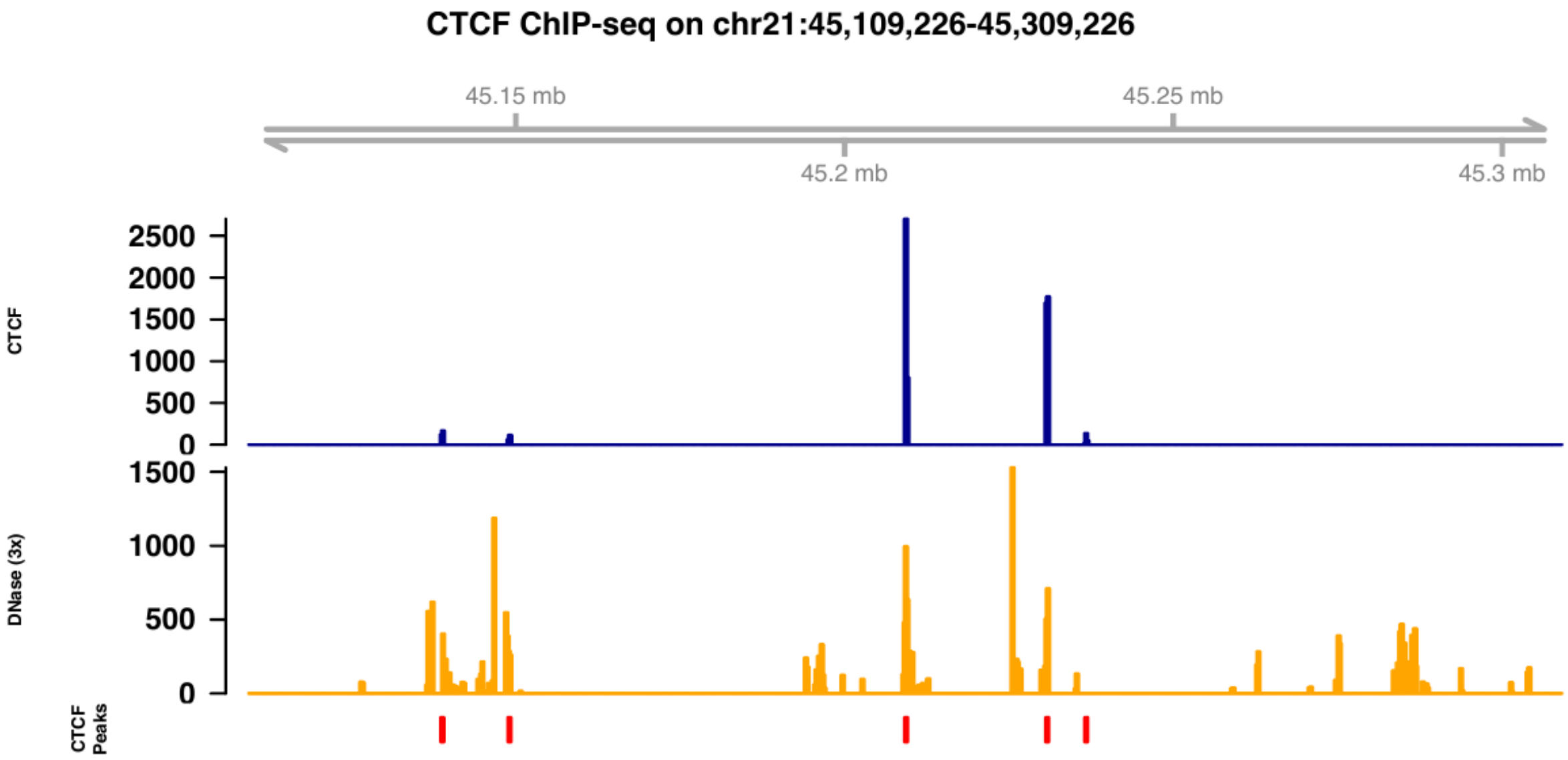
```r
# Download ENCODE DNase-seq peaks (open chromatin for comparison)
dnase_file <- paste0(data_dir, "dnase_gm12878_peaks.rds")
if (!file.exists(dnase_file)) {
    cat("Downloading ENCODE DNase-seq data...\n")
    ah <- AnnotationHub()
    query_dnase <- query(ah, c("DNase", "GM12878", "narrowPeak", "ENCODE"))
    if (length(query_dnase) > 0) {
        dnase_peaks_all <- query_dnase[[1]]
        dnase_peaks <- dnase_peaks_all[seqnames(dnase_peaks_all) == chr]
        saveRDS(dnase_peaks, file = dnase_file)
        cat(sprintf("Saved %d DNase peaks\n", length(dnase_peaks)))
    } else {
        cat("No DNase data found\n")
        saveRDS(GRanges(), file = dnase_file)
    }
}
dnase_peaks <- readRDS(dnase_file)
```

Peak Calling with normr

```r
## Create genomic bins
binsize <- 1000
bins <- GRanges(chr, IRanges(start = seq(1, 48e6, binsize),
                             width = binsize))

## Count reads in bins
chip_counts <- countOverlaps(bins, chip_reads)
dnase_counts <- countOverlaps(bins, dnase_reads)

## Call peaks using normR
fit <- normr::enrichR(treatment = chip_counts,
                      control = dnase_counts,
                      genome = bins,
                      verbose = FALSE)

## Get significant peaks (FDR < 0.05)
peaks_called <- normr::getRanges(fit)
peaks_sig <- peaks_called[peaks_called$qvalue < 0.05]
```

- Foreground (CTCF) vs. background (control, DNase)
- Enrichment-based Binomial Tests
- Input: piled signals
- Output: genomic regions (BED)
- Narrow vs. broad peaks

Visualizing CTCF Binding Sites (using `Gviz`)



Selecting meaningful peaks using reproducibility

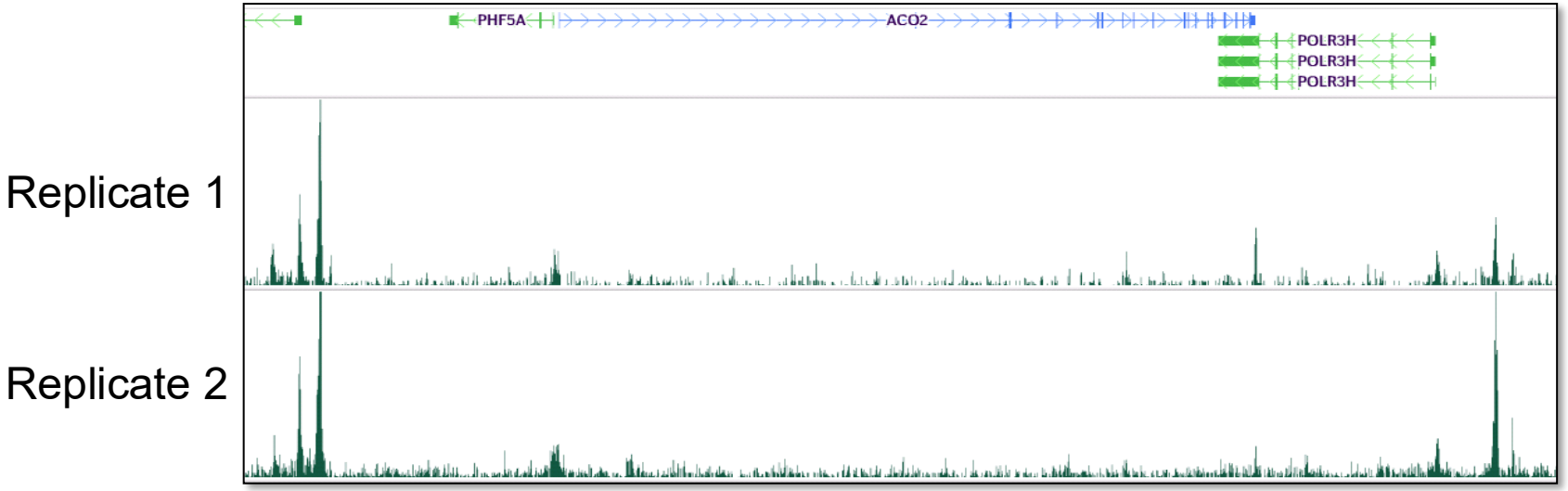
Use peak ranks in replicate experiments

IDR: Irreproducible Discovery Rate

http://anshul.kundaje.net/projects/idr

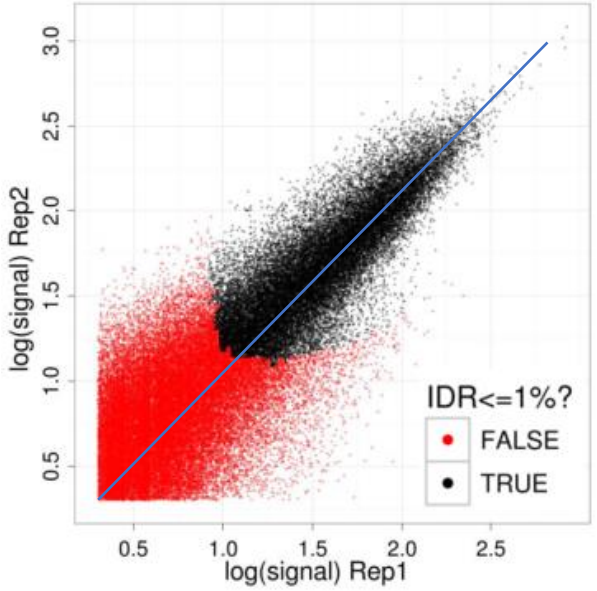
A. Kundaje, Q. Li , B. Brown, J. Rozowsky, S. Wilder, M. Gerstein, I. Dunham, E. Birney, P. Bickel

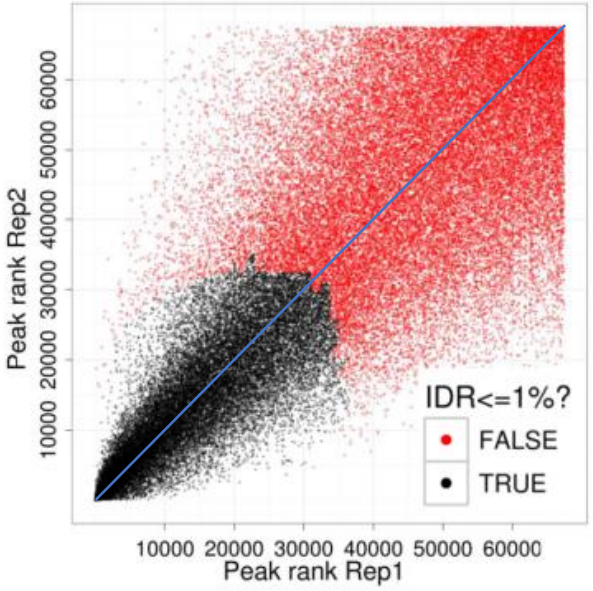
How to combine two replicates

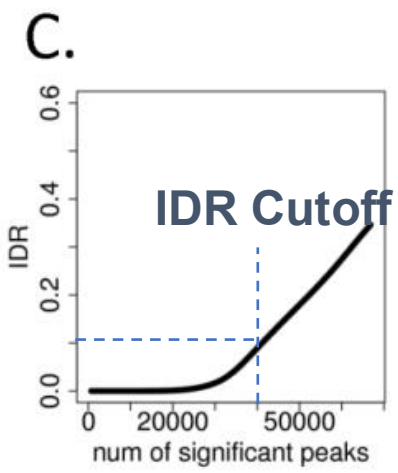


- Challenge:
  - Replicates show small differences in peak heights
  - Many peaks in common, but many are unique
- Problem with simple solutions:
  - Union: too lenient, keeps garbage from both
  - Intersection: too stringent, throws away good peaks
  - Sum: does not exploit independence of two datasets

IDR idea: Exploit peak rank similarity in replicates







- Key idea: True peaks will be highly ranked in both replicates
  - Keep going down rank list, until ranks are no longer correlated
  - This cutoff could be different for the two replicates
  - The actual peaks included may differ between replicates
- Adaptively learn optimal peak calling threshold
  - **F**DR threshold of 10% ➔ 10% of peaks are false (widely used)
  - **I**DR threshold of 10% ➔ 10% of peaks are not reproducible

The IDR model: A two component mixture

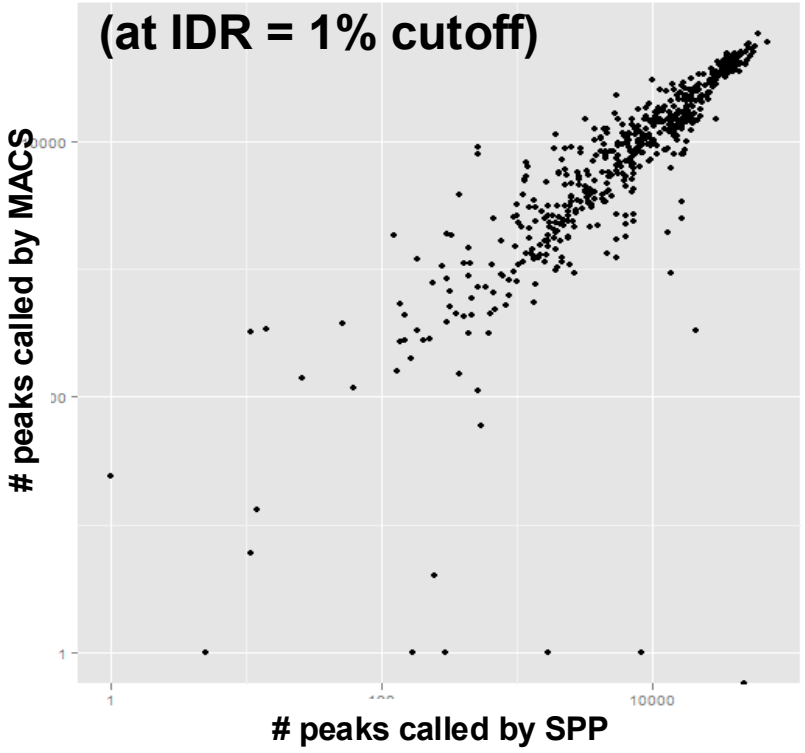
- Looking only at ranks means that the marginals are uniform, so all the information is encoded in the joint distribution.
- Model the joint distribution of ranks as though it came from a two component Gaussian mixture model:

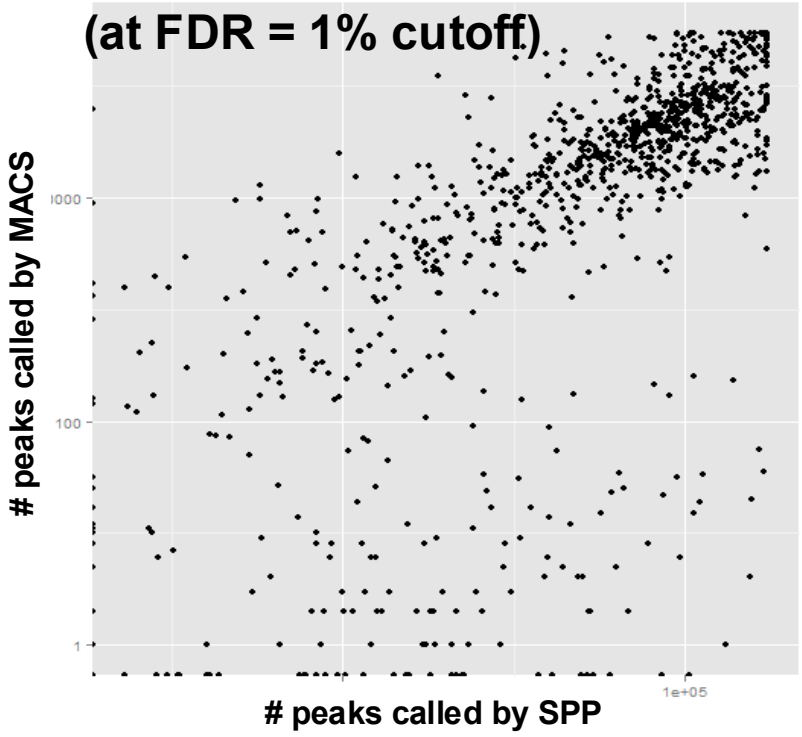
$$(x, y) \sim pN(\mu, \mu, \sigma, \sigma, \rho) + (1 - p)N(0,0,1,1,0)$$

- This can be fit via an EM-like algorithm.

IDR leads to higher consistence between peak callers

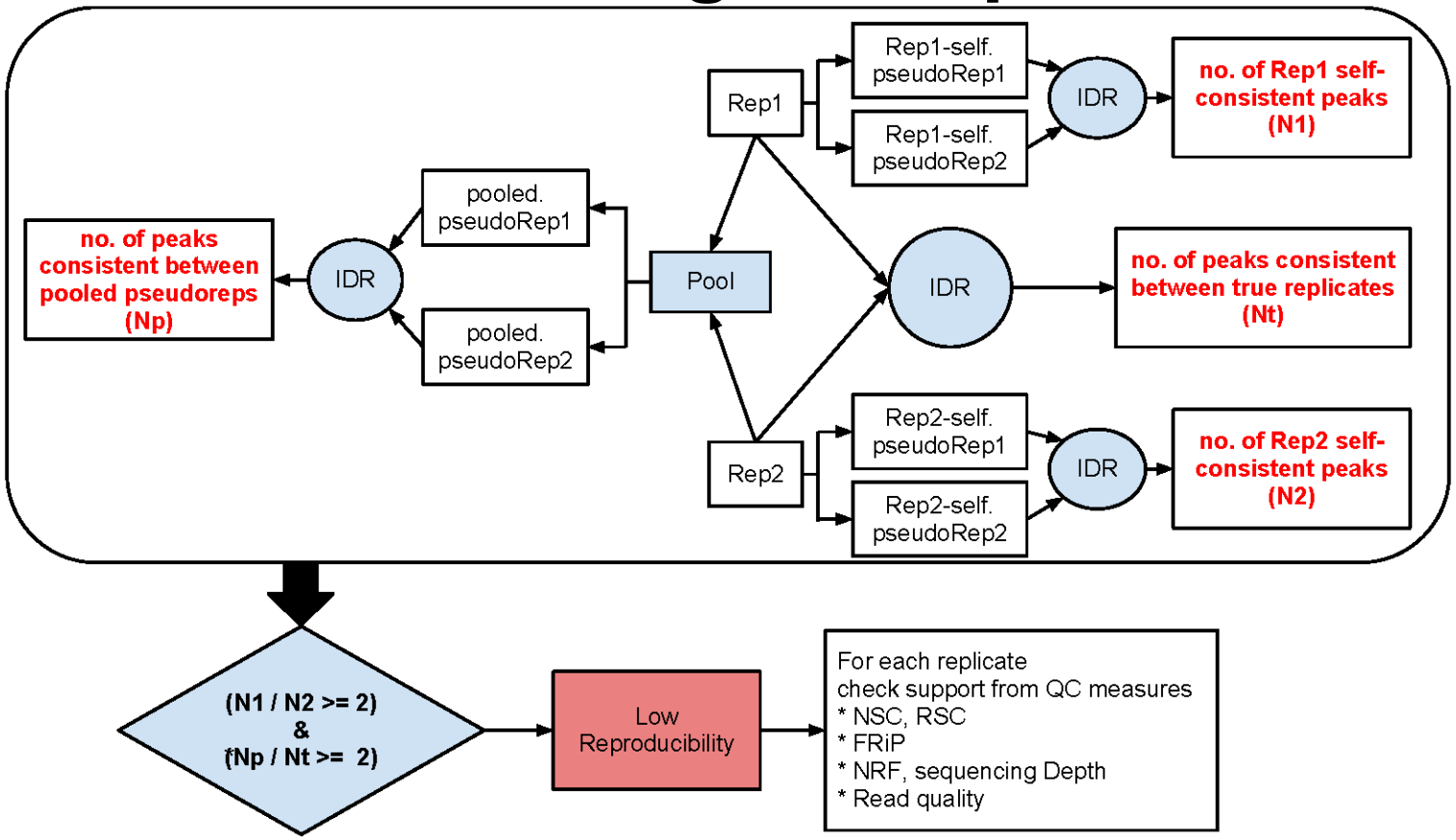
IDR = Irreproducible Discovery Rate FDR = False Discovery Rate





- Compare number of peaks found by two different peak callers
- IDR thresholds are far more robust and comparable than FDR
- FDR only relies on enrichment over input, IDR exploits replicates

What if we don't have good replicates?

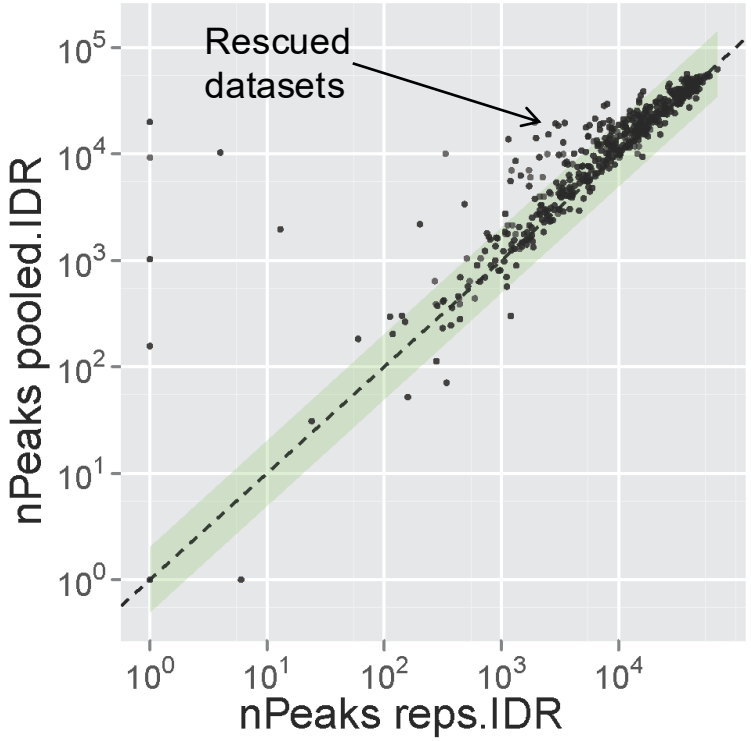


- IDR pipeline uses replicates when they are available
- IDR pipeline also evaluates each replicate individually
  - Pooling strategy to generate pseudo-replicates

➔ Can pin-point 'bad' replicates that may lead to low reproducibility

➔ Can estimate IDR thresholds when replicates are not available

Only one good replicate: Pseudo-replicates



- IDR pipeline can be used to rescue datasets with only one good replicate (using pseudo-replicates)
- IDR pipeline can also be used to call optimal thresholds on a dataset with a single replicate (e.g. when there isn't enough material to perform multiple reps)

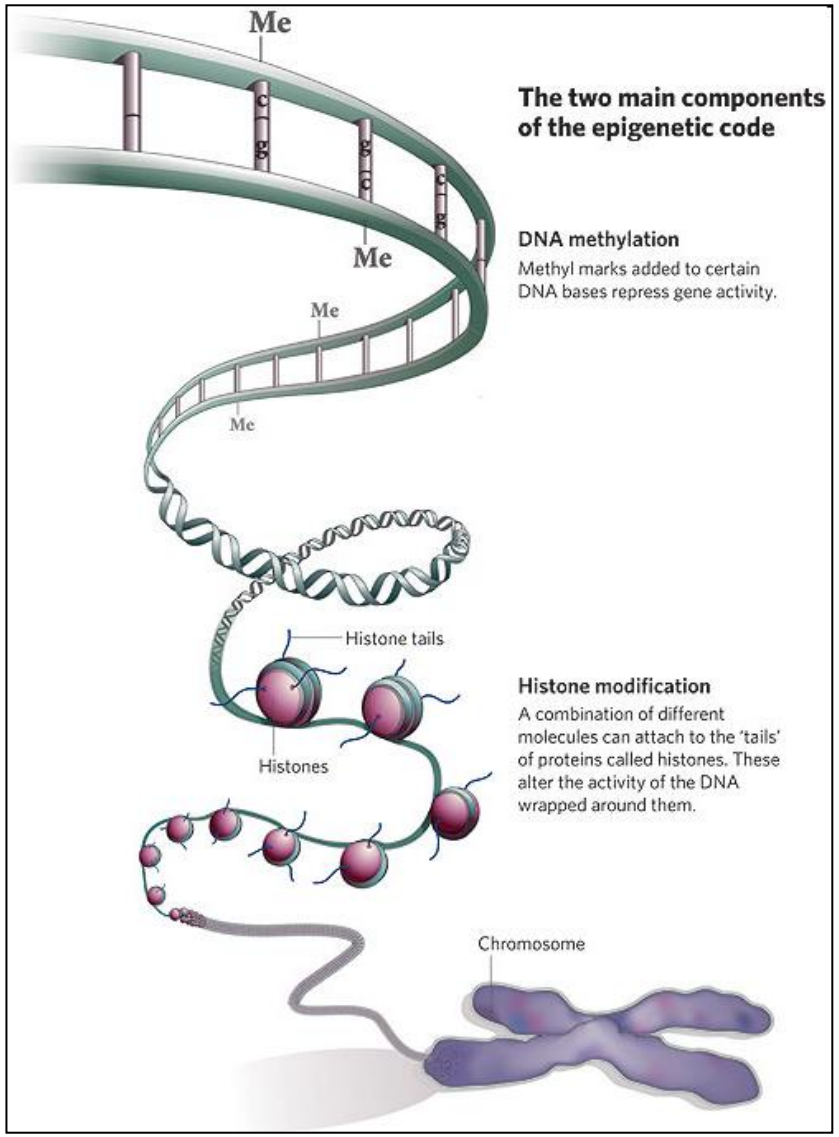
Key Concepts from ChIP-seq Peak Calling

- **Biological**: ChIP-seq identifies protein-DNA binding sites
- **Statistical** models:
  - Poisson/Binomial models for count data
  - Local background estimation
  - Multiple testing correction (FDR)
- **Computational** pipeline:
  - Millions of bins $\rightarrow$ massive multiple testing problem
  - Need efficient algorithms (`MACS2`, `normr`)
- **Quality control**:
  - Peak width distribution (sharp vs broad)
  - Fraction of reads in peaks (FRiP score)
  - Replicate concordance

Today's lecture

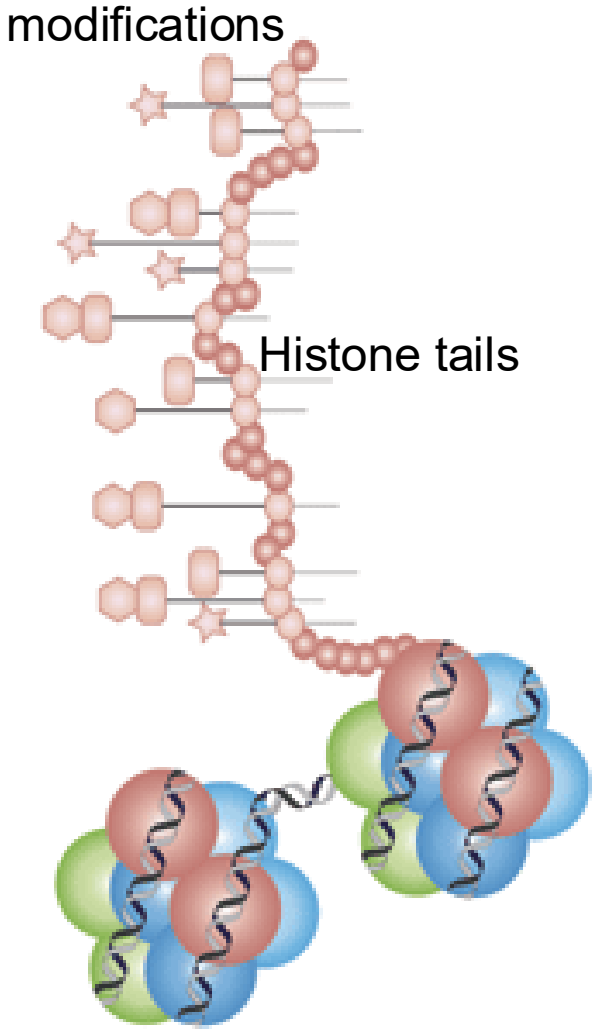
1. Technologies for measuring regulatory activities
2. DNA Methylation (credit: Keegan Korthauer)
3. Multiomic Data Integration with Regulatory Genomics

Chromatin signatures for genome annotation



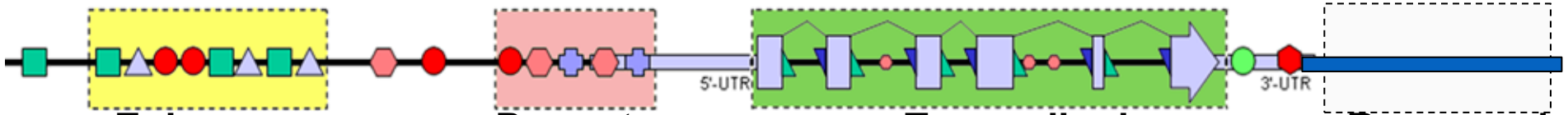
- **Challenges**
  - Dozens of marks
  - Complex combinatorics
  - Diversity and dynamics
- **Histone code hypothesis**
  - Distinct function for distinct combinations of marks?
  - Both additive and combinatorial effects
- **How do we find biologically relevant ones?**
  - Unsupervised approach
  - Probabilistic model
  - Explicit combinatorics

100s of histone tail modifications



- 100+ different histone modifications
  - Histone protein → H3/H4/H2A/H2B
  - AA residue → Lysine4(K4)/K36…
  - Chemical modification → Met/Pho/Ubi
  - Number → Me-Me-Me(me3)
  - Shorthand: H3K4me3, H2BK5ac
- In addition:
  - DNA modifications
  - Methyl-C in CpG / Methyl-Adenosine
  - Nucleosome positioning
  - DNA accessibility
- The constant struggle of gene regulation
  - TF/histone/nucleo/GFs/Chrom compete

Combinations of marks encode epigenomic state

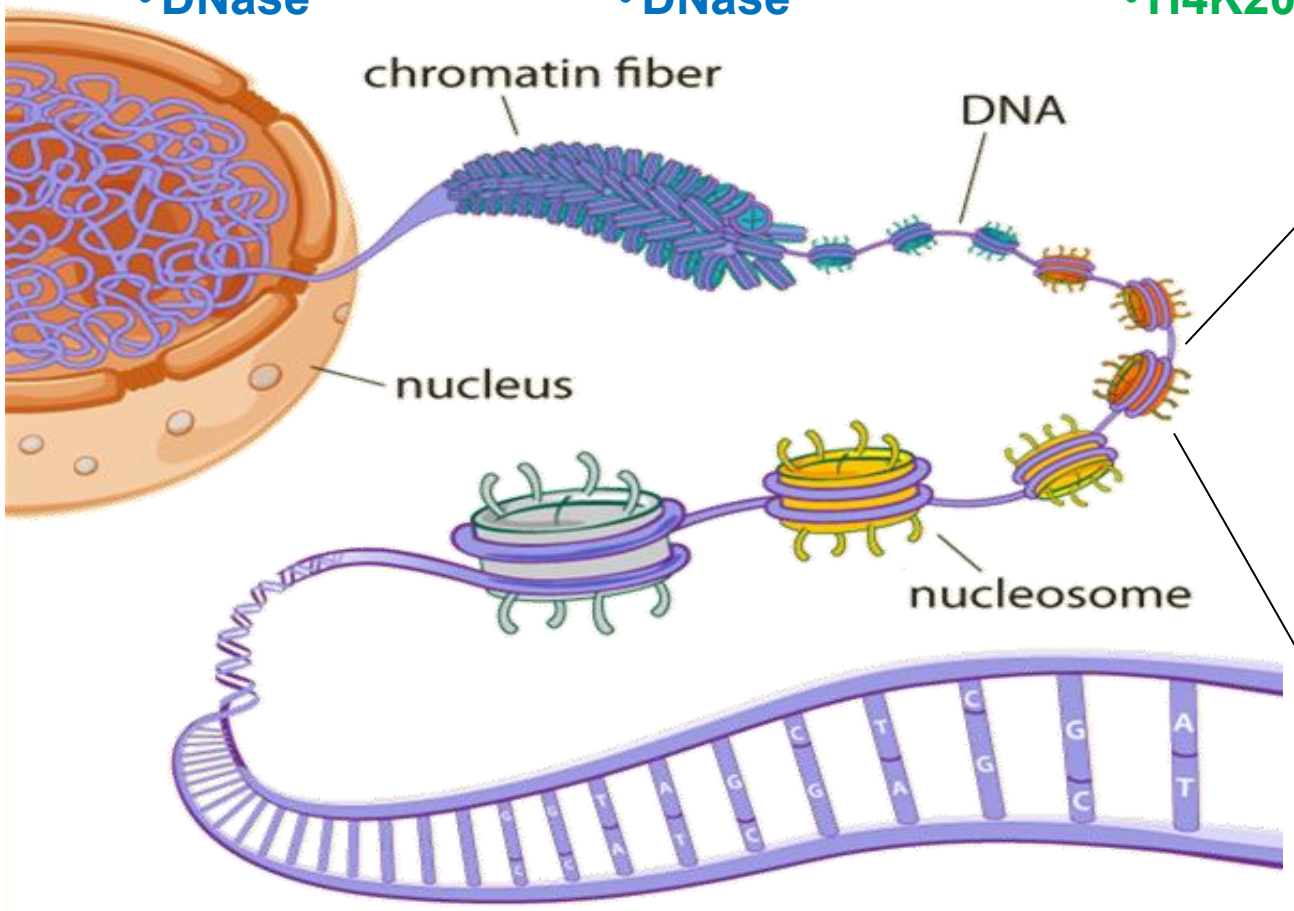


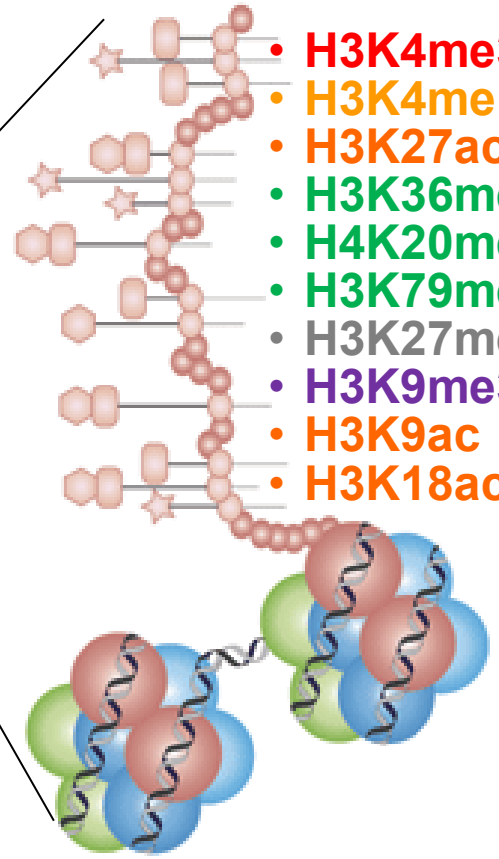
Enhancers
- H3K4me1
- H3K27ac
- DNase

Promoters
- H3K4me3
- H3K9ac
- DNase

Transcribed
- H3K36me3
- H3K79me2
- H4K20me1

Repressed
- H3K9me3
- H3K27me3
- DNAmethyl

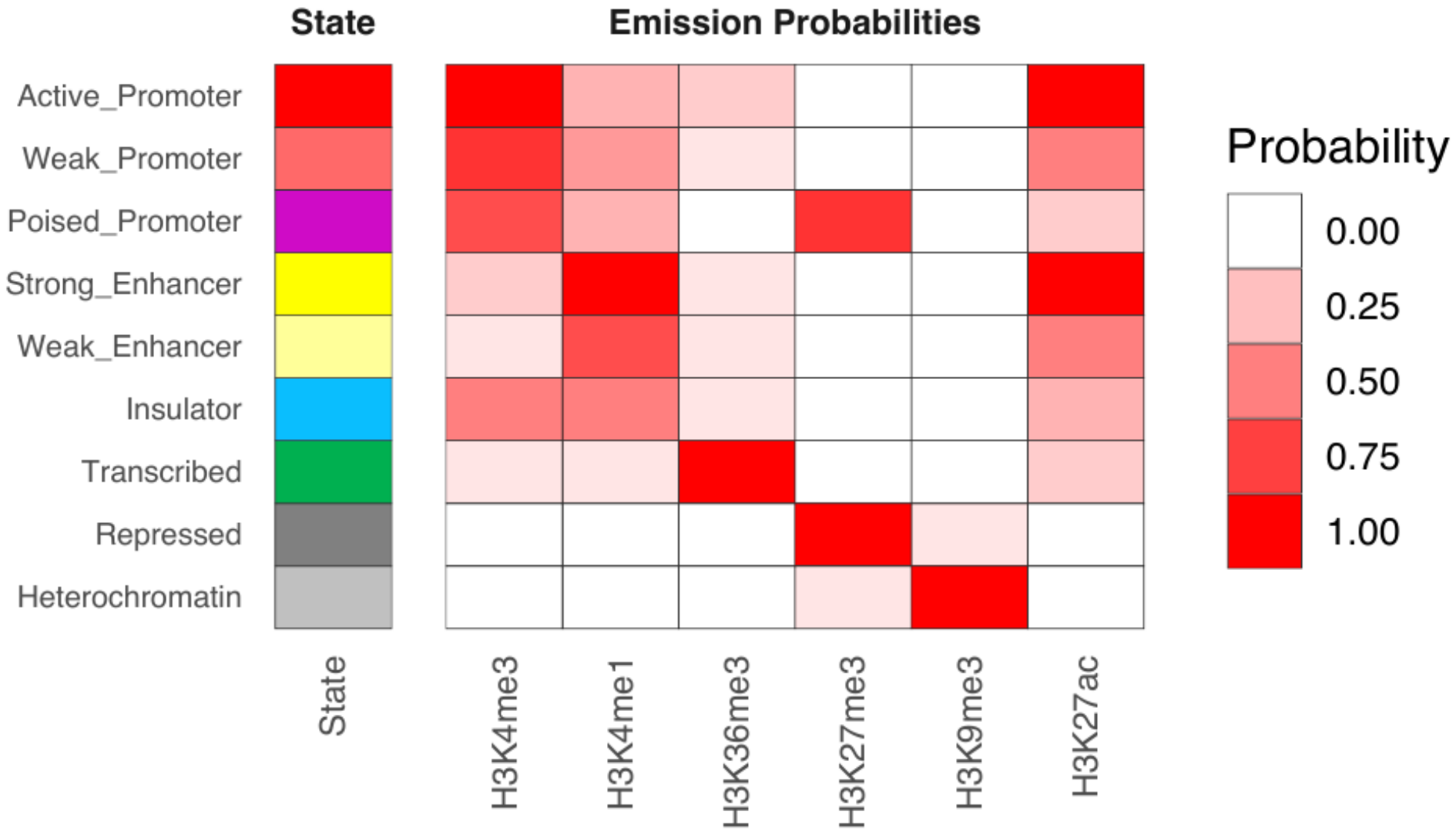




- H3K4me3
- H3K4me1
- H3K27ac
- H3K36me3
- H4K20me1
- H3K79me3
- H3K27me3
- H3K9me3
- H3K9ac
- H3K18ac

- 100s of known modifications, many new still emerging
- Systematic mapping using ChIP-, Bisulfite-, DNase-Seq

ChromHMM: Emission Probabilities



ChromHMM: Integrating Multiple Histone Marks

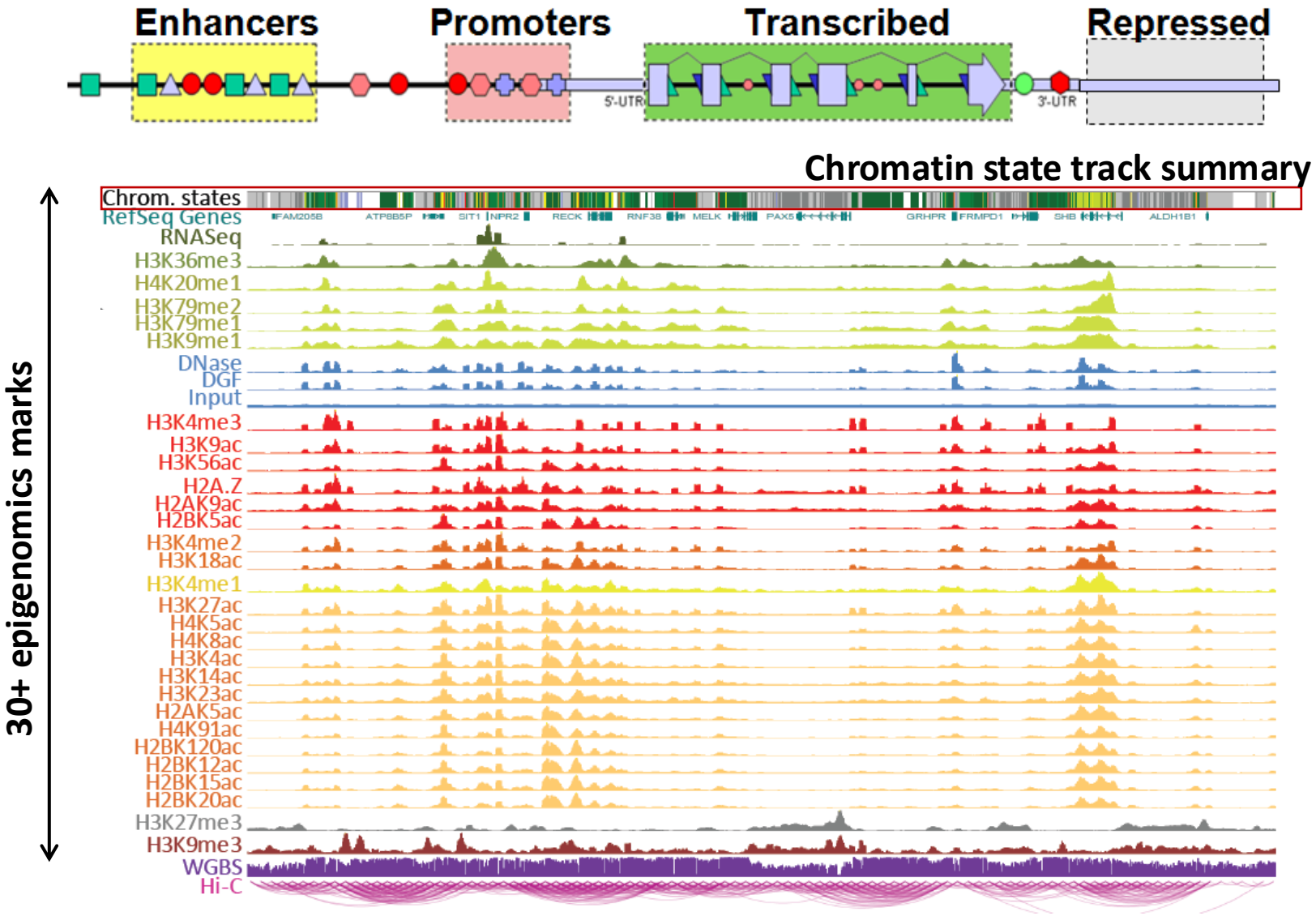
How it works:

- Input: Multiple histone modification ChIP-seq signals
- Model: Hidden Markov Model learns chromatin states
- Each state defined by emission probabilities (previous slide)
- Output: Genome segmented into functional states

In our genome browser:

- ChromHMM track shows where each state occurs
- CTCF peaks often coincide with specific states (insulators, enhancers)

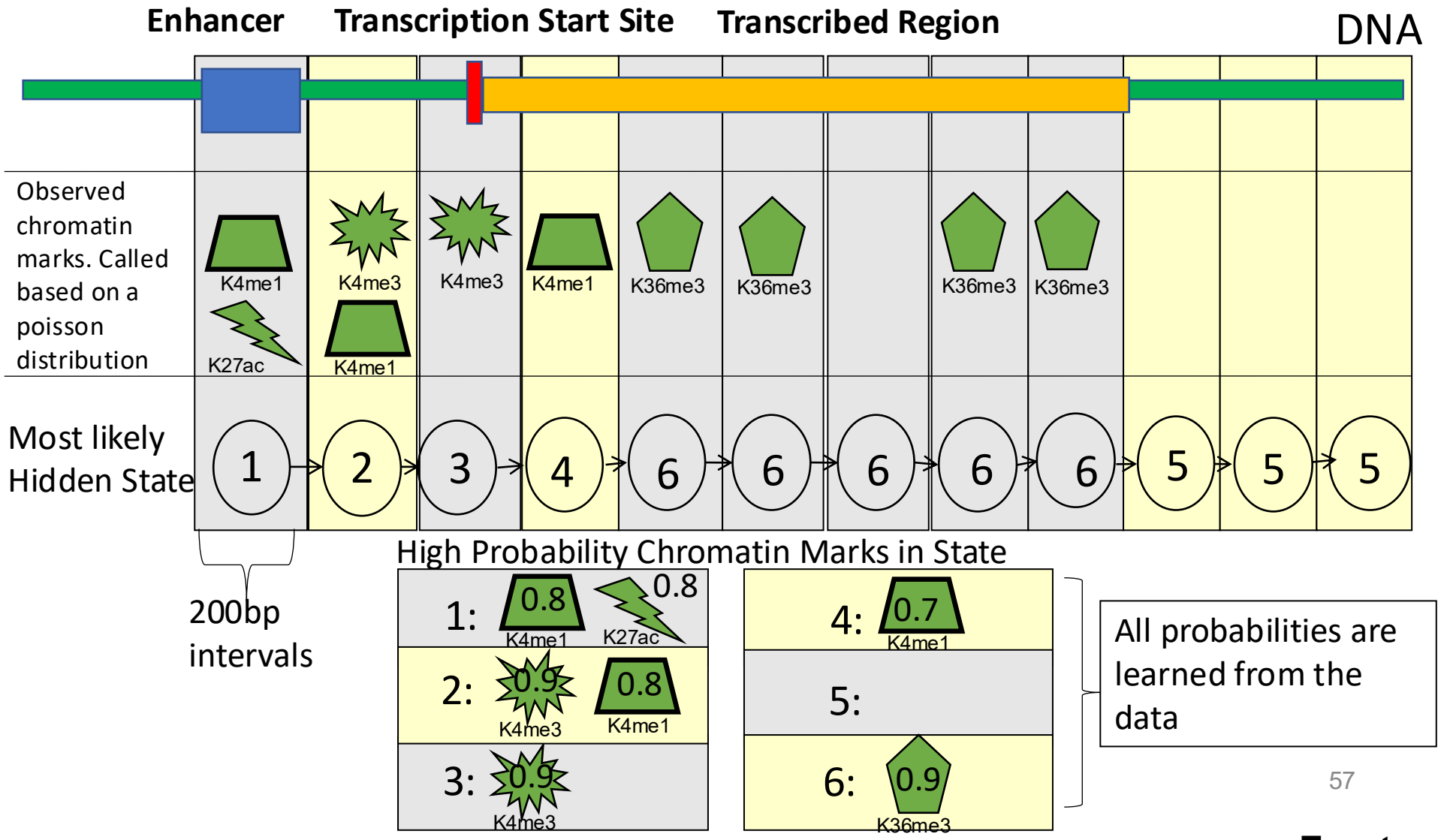
Summarize multiple marks into chromatin states



WashU Epigenome Browser

ChromHMM: multi-variate hidden Markov model

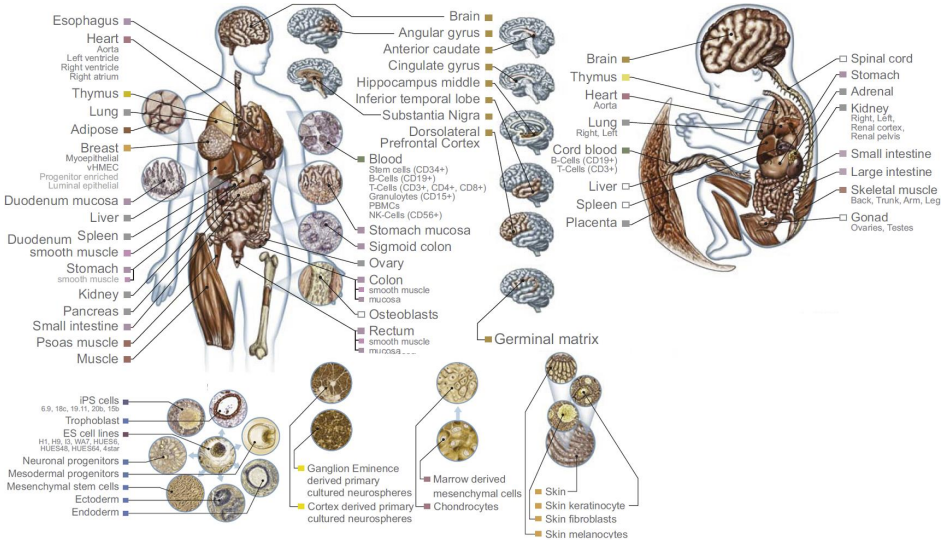
Multivariate HMM for Chromatin States



Ernst and Kellis
Nature Biotech 2010

Epigenomic mapping across 100+ tissues/cell types

Diverse tissues and cells



Adult tissues and cells (brain, muscle, heart, digestive, skin, adipose, lung, blood…)

Fetal tissues (brain, skeletal muscle, heart, digestive, lung, cord blood…)

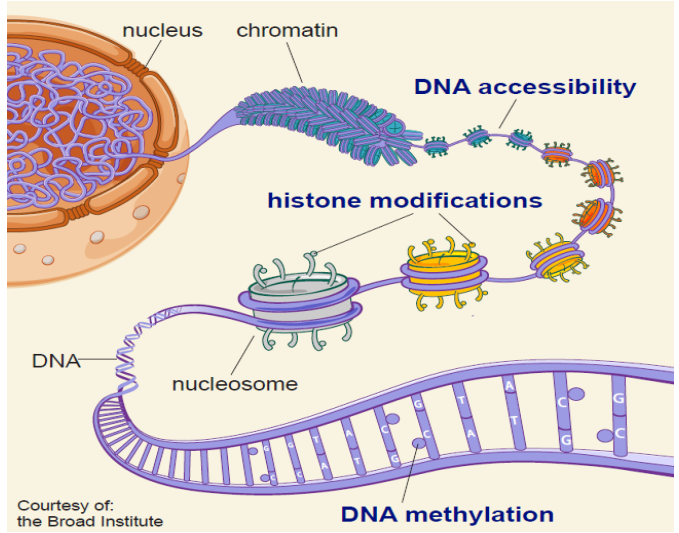
ES cells, iPS, differentiated cells (meso/endo/ectoderm, neural, mesench…)

ROADMAP epigenomics PROJECT

ENCODE

X

Diverse epigenomic assays



Histone modifications
- **H3K4me3, H3K4me1, H3K36me3**
- **H3K27me3, H3K9me3, H3K27/9ac**
- **+20 more**

Open chromatin:
- **DNA accessibility**

DNA methylation:
- **WGBS, RRBS, MRE/MeDIP**

Gene expression
- **RNA-seq, Exon Arrays**

ENCODE: Study nine marks in nine human cell lines

9 marks

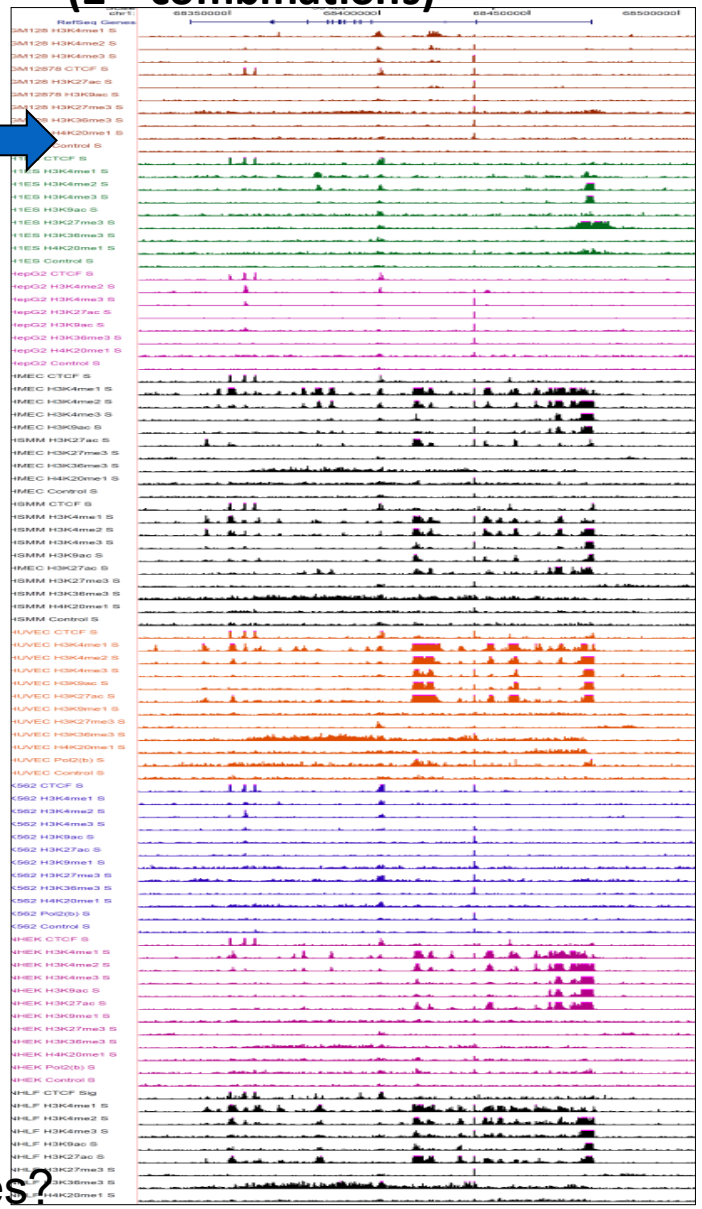
H3K4me1
H3K4me2
H3K4me3
H3K27ac
H3K9ac
H3K27me3
H4K20me1
H3K36me3
CTCF
+WCE
+RNA

X

9 human cell types

HUVEC	Umbilical vein endothelial
NHEK	Keratinocytes
GM12878	Lymphoblastoid
K562	Myelogenous leukemia
HepG2	Liver carcinoma
NHLF	Normal human lung fibroblast
HMEC	Mammary epithelial cell
HSMM	Skeletal muscle myoblasts
H1	Embryonic

81 Chromatin Mark Tracks (2^{81} combinations)



Brad Bernstein ENCODE Chromatin Group

b.

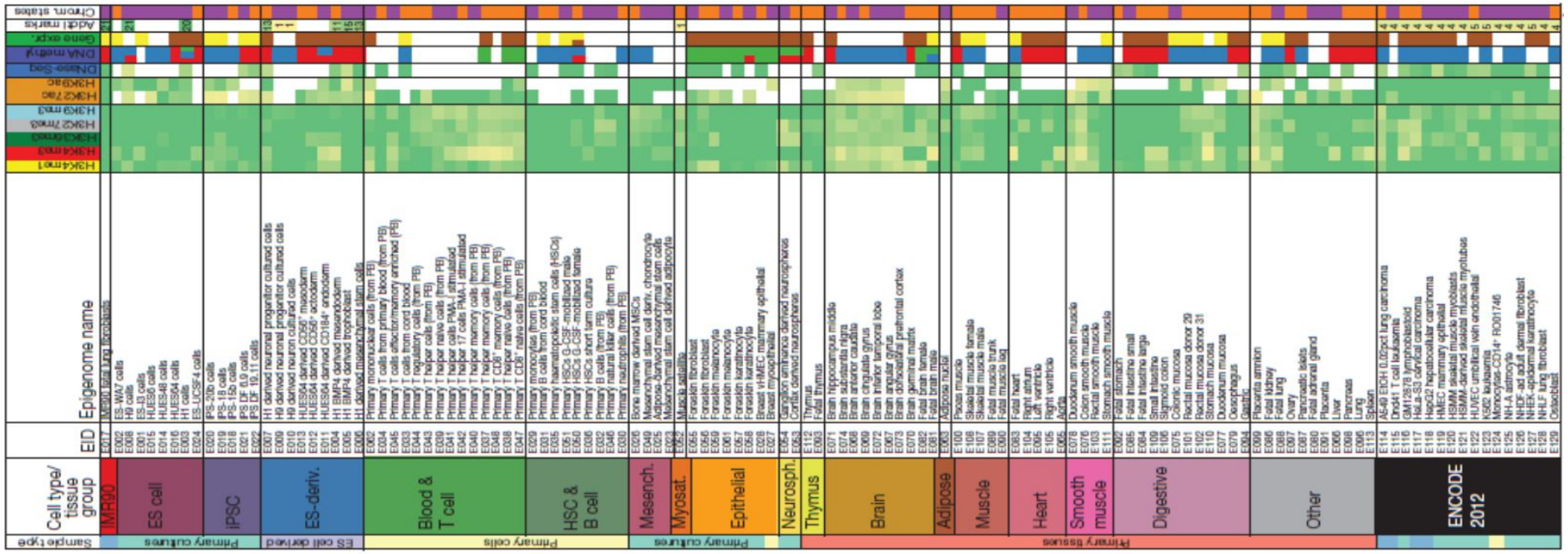
Chromatin States: State	CTCF	H3K27me3	H3K36me3	H4K20me1	H3K4me1	H3K4me2	H3K4me3	H3K27ac	H3K9ac	WCE
1	16	2	2	6	17	93	99	96	98	2
2	12	2	6	9	53	94	95	14	44	1
3	13	72	0	9	48	78	49	1	10	1
4	11	1	15	11	96	99	75	97	86	4
5	5	0	10	3	88	57	5	84	25	1
6	7	1	1	3	58	75	8	6	5	1
7	2	1	2	1	56	3	0	6	2	1
8	92	2	1	3	6	3	0	0	1	1
9	5	0	43	43	37	11	2	9	4	1
10	1	0	47	3	0	0	0	0	0	1
11	0	0	3	2	0	0	0	0	0	0
12	1	27	0	2	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0
14	22	28	19	41	6	5	26	5	13	37
15	85	85	91	88	76	77	91	73	85	78

Chromatin Mark Observation Frequency (%)

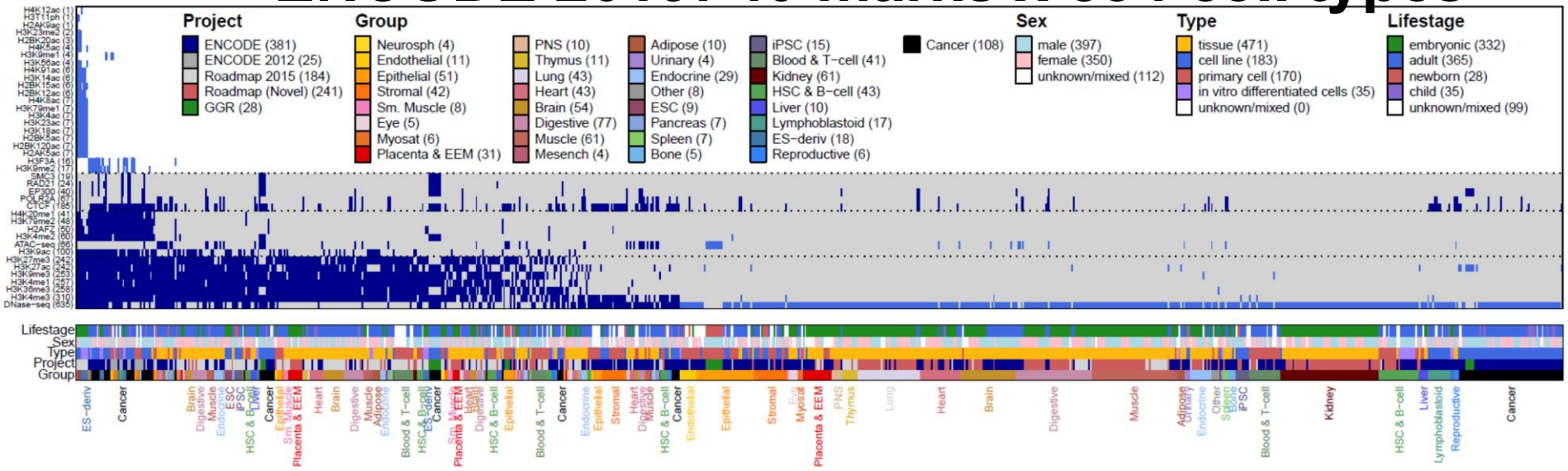
How to learn single set of chromatin states?

Ernst et al, *Nature* 2011

Roadmap 2015: 12 marks x 127 cell types

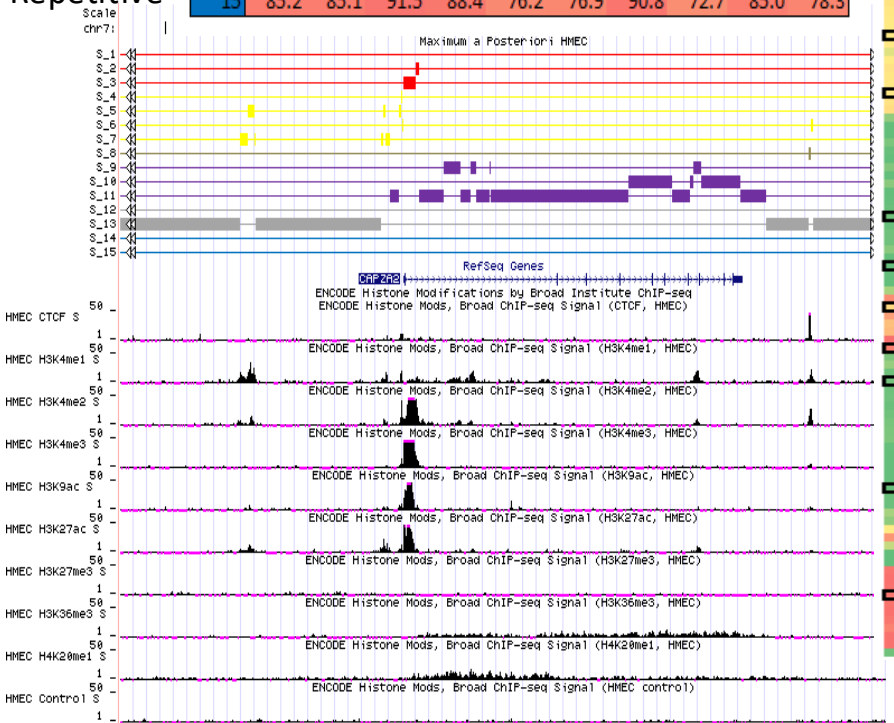


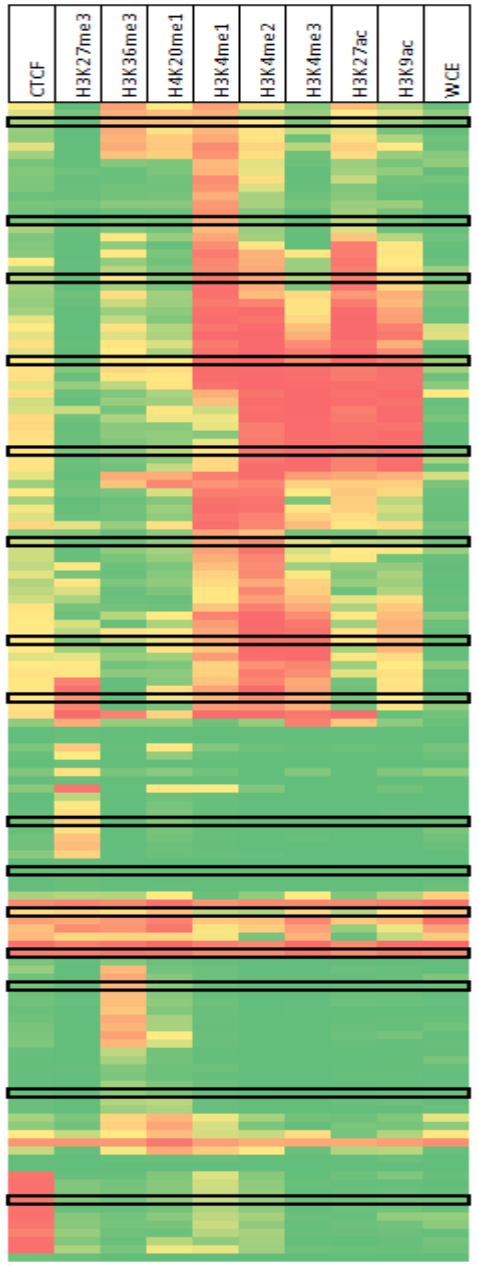
ENCODE 2019: 40 marks x 834 cell types



Solution 1: Learn independent models and cluster

	state	CTCF	H3K27me3	H3K36me3	H4K20me1	H3K4me1	H3K4me2	H3K4me3	H3K27ac	H3K9ac	WCE
Promoter	1	13.2	72.0	0.2	9.1	47.9	77.8	49.5	1.3	10.2	0.7
	2	11.9	1.9	6.1	9.0	52.7	93.7	95.0	14.1	44.1	0.9
	3	16.4	1.5	2.4	5.5	17.0	92.6	99.0	95.7	98.1	1.9
Candidate enhancer	4	11.4	0.6	14.5	11.3	96.3	99.3	75.1	97.2	85.7	3.7
	5	5.3	0.2	9.5	2.6	88.1	56.8	5.3	84.4	24.9	1.5
	6	6.7	0.9	1.0	3.2	58.3	74.7	8.4	5.8	5.4	0.8
	7	1.6	0.6	1.6	1.3	56.5	2.7	0.4	5.9	1.6	0.6
Insulator	8	91.5	1.8	0.9	2.8	6.3	3.3	0.4	0.5	1.0	0.8
Transcribed	9	4.6	0.3	43.2	43.1	36.5	11.5	1.9	9.1	3.9	1.3
	10	1.2	0.1	47.2	2.7	0.4	0.0	0.1	0.3	0.3	0.5
	11	0.4	0.1	2.7	1.7	0.2	0.1	0.1	0.2	0.3	0.4
Repressive	12	0.9	26.8	0.0	2.1	0.4	0.1	0.1	0.1	0.1	0.4
	13	0.2	0.4	0.0	0.1	0.1	0.0	0.0	0.0	0.1	0.1
Repetitive	14	21.9	27.9	19.1	41.0	5.7	4.8	25.9	5.3	13.1	37.5
	15	85.2	85.1	91.5	88.4	76.2	76.9	90.8	72.7	85.0	78.3





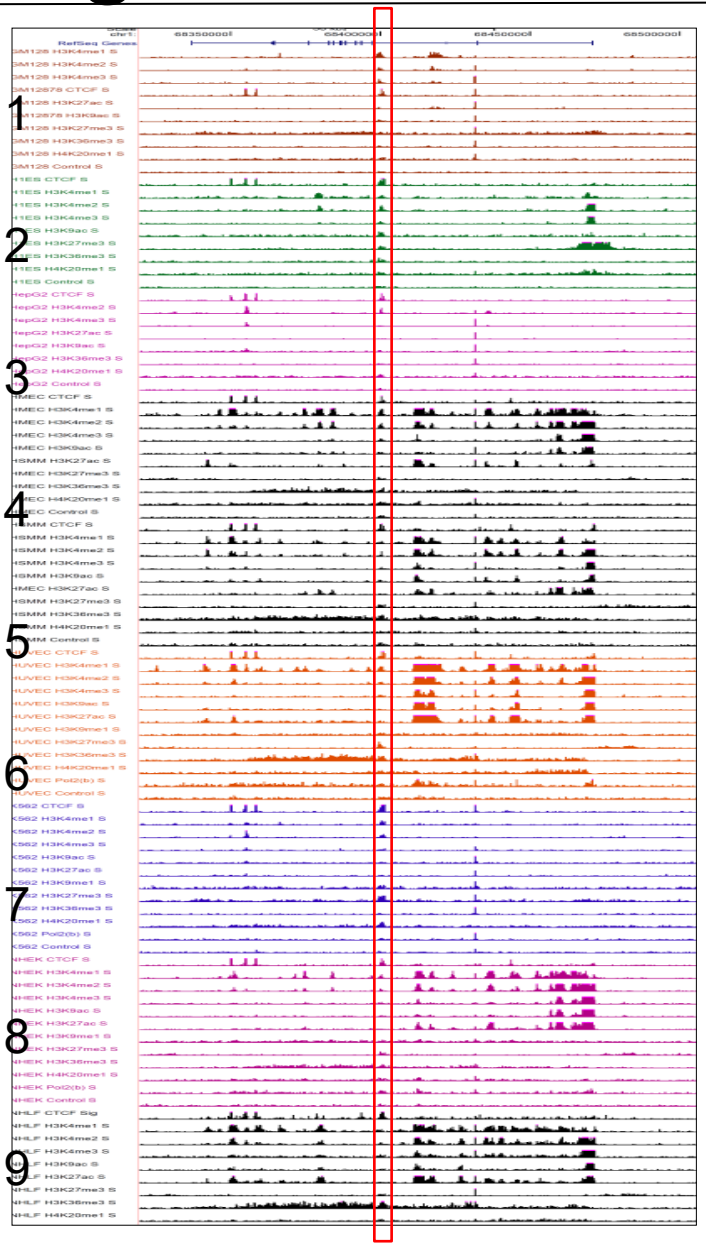
Basic approach:

a) Train a k-state model in each cell type independently
b) Cluster models learned independently
c) Merge clusters and re-apply to each cell type

How to cluster

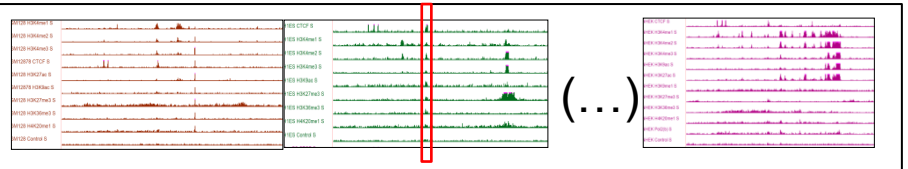
a) Using emission probability matrix: most similar definitions
b) Using genome annotation: posterior probability decoding

Joint learning of states across multiple cell types



Solution 2: Stacking

- Learns each combination of activity as a separate state
- Ex: ES-specific enhancers: enhancer marks in ES, no marks in other cell types

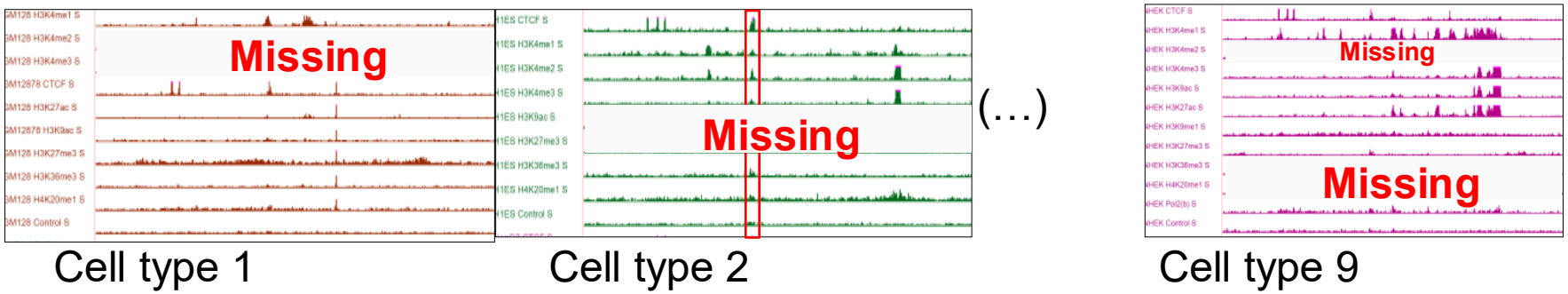


Cell type 1 Cell type 2 Cell type 9

Solution 3: Concatenation

- Requires that profiled marks are the same (or treat as missing data)
- Ensures common state definitions across cell types

Joint learning with <u>different subsets</u> of marks (Solution 3)



Option (a) Treat missing tracks as missing data

- EM framework allows for unspecified data points
- As long as pairwise relationship observed in some cell type

Option (b) Chromatin mark imputation

- Explicitly predict max-likelihood chromatin track for missing data
- Less powerful if ultimate goal is chromatin state learning

ENCODE: Study 9 marks in nine human cell lines

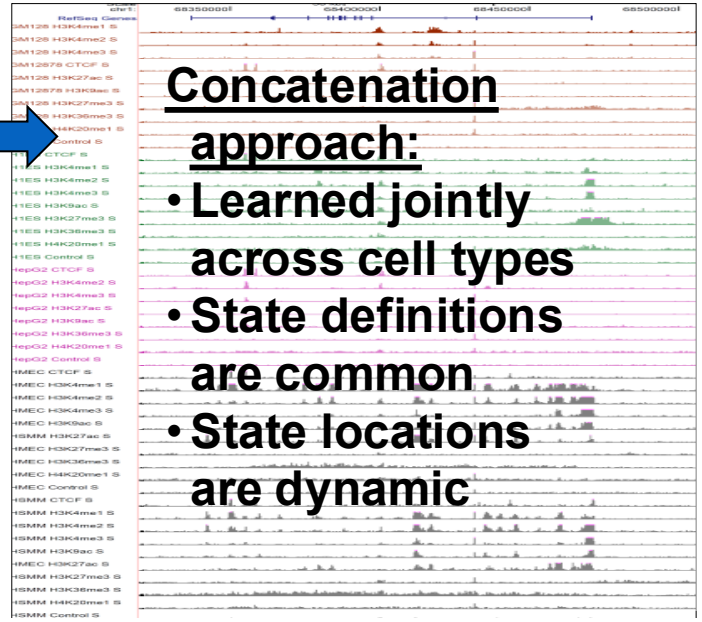
9 marks

H3K4me1
H3K4me2
H3K4me3
H3K27ac
H3K9ac
H3K27me3
H4K20me1
H3K36me3
CTCF
+WCE
+RNA

X

HUVEC	Umbilical vein endothelial
NHEK	Keratinocytes
GM12878	Lymphoblastoid
K562	Myelogenous leukemia
HepG2	Liver carcinoma
NHLF	Normal human lung fibroblast
HMEC	Mammary epithelial cell
HSMM	Skeletal muscle myoblasts
H1	Embryonic

81 Chromatin Mark Tracks



Concatenation approach:

- **Learned jointly across cell types**
- **State definitions are common**
- **State locations are dynamic**

Brad Bernstein ENCODE Chromatin Group

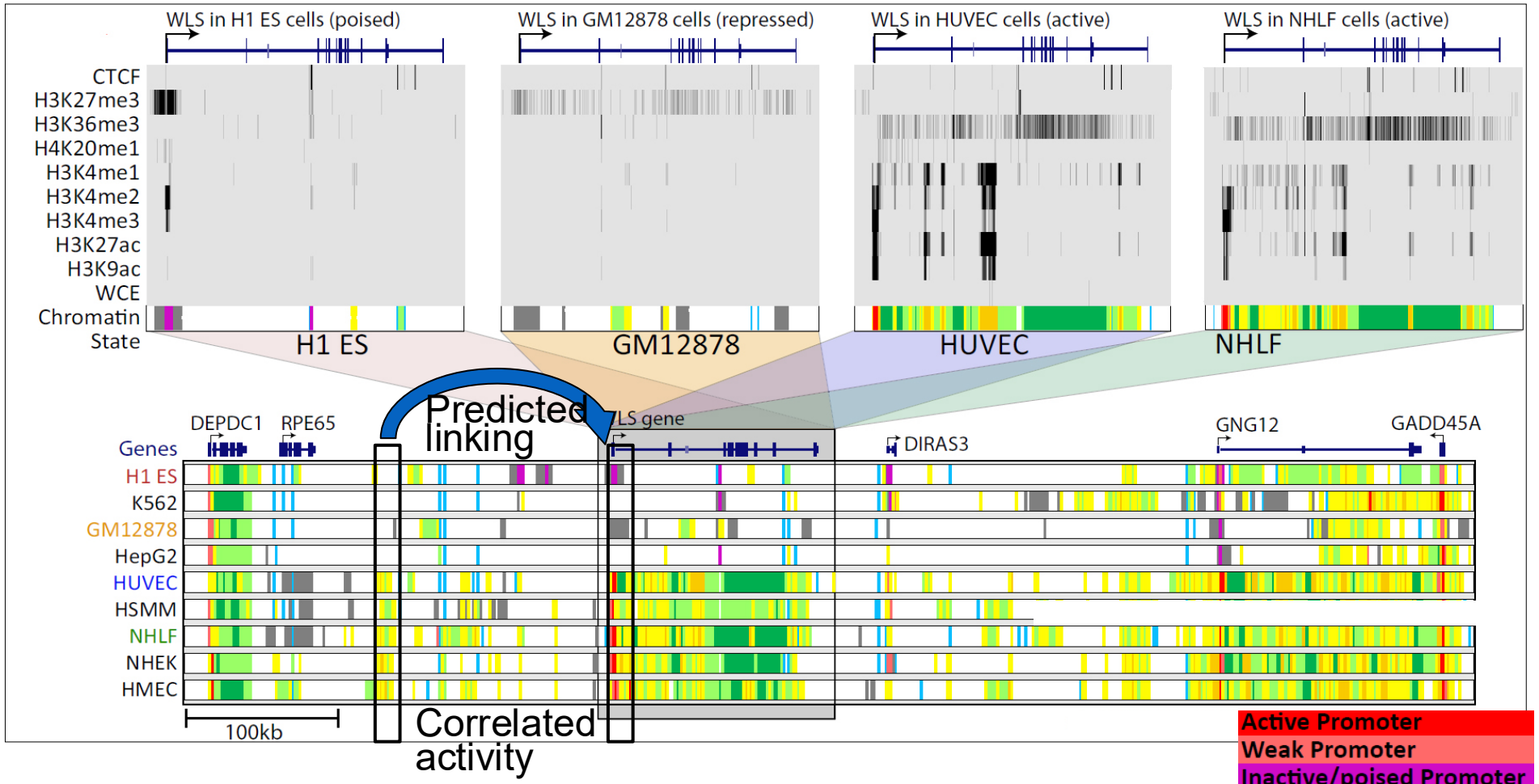
b.

Chromatin States

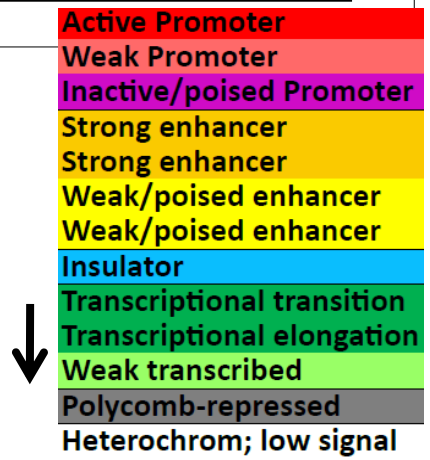
State	CTCF	H3K27me3	H3K36me3	H4K20me1	H3K4me1	H3K4me2	H3K4me3	H3K27ac	H3K9ac	WCE	Coverage Median	Coverage H1 ES	Coverage GM	Median Length	+/-2kb TSS	Conserved non-exon	DNase (K562)	C-Myc (K562)	NF-κB (GM12878)	Transcript	Nuclear Lamina (NHLF)	Candidate state annotation
1	16	2	2	6	17	93	99	96	98	2	0.6	0.5	1.2	1.0	83	3.8	23.3	82.0	40.7	0.2	0.15	Active Promoter
2	12	2	6	9	53	94	95	14	44	1	0.5	1.2	1.3	0.4	58	2.8	15.3	12.6	5.8	0.6	0.30	Weak Promoter
3	13	72	0	9	48	78	49	1	10	1	0.2	4.0	1.0	0.6	49	4.3	10.8	3.1	1.0	0.4	0.68	Inactive/poised Promoter
4	11	1	15	11	96	99	75	97	86	4	0.7	0.1	1.1	0.6	23	2.7	23.1	31.8	49.0	1.3	0.05	Strong enhancer
5	5	0	10	3	88	57	5	84	25	1	1.2	0.2	0.7	0.6	3	1.8	13.6	6.3	15.8	1.4	0.10	Strong enhancer
6	7	1	1	3	58	75	8	6	5	1	0.9	1.3	1.0	0.2	17	2.4	11.9	5.7	7.0	1.1	0.31	Weak/poised enhancer
7	2	1	2	1	56	3	0	6	2	1	1.9	1.2	1.1	0.4	4	1.5	5.1	0.6	2.4	1.3	0.20	Weak/poised enhancer
8	92	2	1	3	6	3	0	0	1	1	0.5	1.4	1.0	0.4	3	1.5	12.8	2.5	1.2	1.1	0.61	Insulator
9	5	0	43	43	37	11	2	9	4	1	0.7	1.3	1.0	0.8	4	1.1	4.5	0.7	0.8	2.4	0.02	Transcriptional transition
10	1	0	47	3	0	0	0	0	0	1	4.3	0.6	1.2	3.0	1	0.9	0.3	0.0	0.0	2.5	0.11	Transcriptional elongation
11	0	0	3	2	0	0	0	0	0	0	12.5	1.3	0.8	2.6	2	0.9	0.3	0.0	0.1	1.9	0.24	Weak transcribed
12	1	27	0	2	0	0	0	0	0	0	4.1	0.3	0.7	2.8	5	1.4	0.3	0.0	0.1	0.8	0.63	Polycomb-repressed
13	0	0	0	0	0	0	0	0	0	0	71.4	1.0	1.0	10.0	1	0.9	0.1	0.0	0.0	0.7	1.30	Heterochrom; low signal
14	22	28	19	41	6	5	26	5	13	37	0.1	0.9	1.2	0.6	3	0.4	1.9	0.3	0.2	0.4	1.44	Repetitive/CNV
15	85	85	91	88	76	77	91	73	85	78	0.1	0.9	1.0	0.2	1	0.2	5.9	9.5	7.4	0.4	1.30	Repetitive/CNV

Chromatin Mark Observation Frequency (%) | (%) (fold) | (kb) | (%) | Functional enrichments (fold)

Chromatin states dynamics across nine cell types

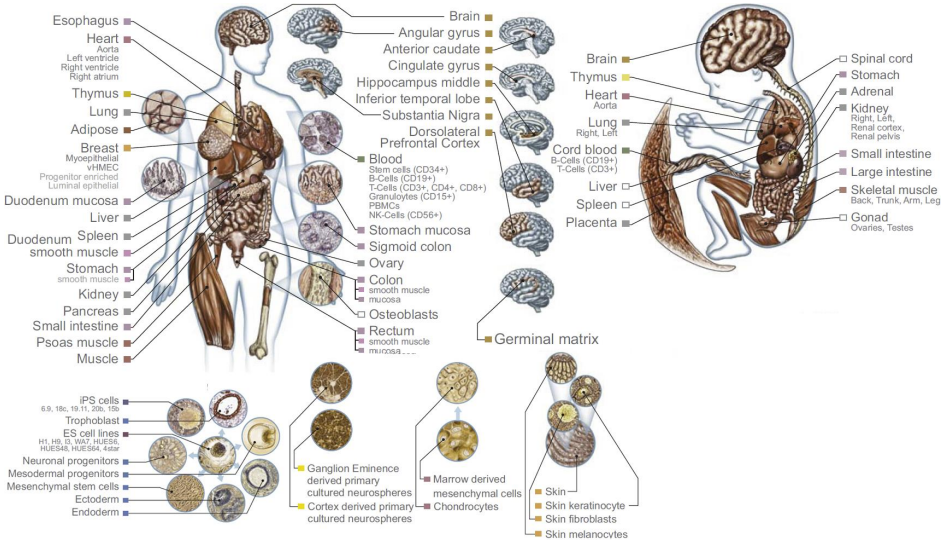


- Single annotation track for each cell type
- Summarize cell-type activity at a glance
- Can study 9-cell activity pattern across



Epigenomic mapping across 100+ tissues/cell types

Diverse tissues and cells



Adult tissues and cells (brain, muscle, heart, digestive, skin, adipose, lung, blood…)

Fetal tissues (brain, skeletal muscle, heart, digestive, lung, cord blood…)

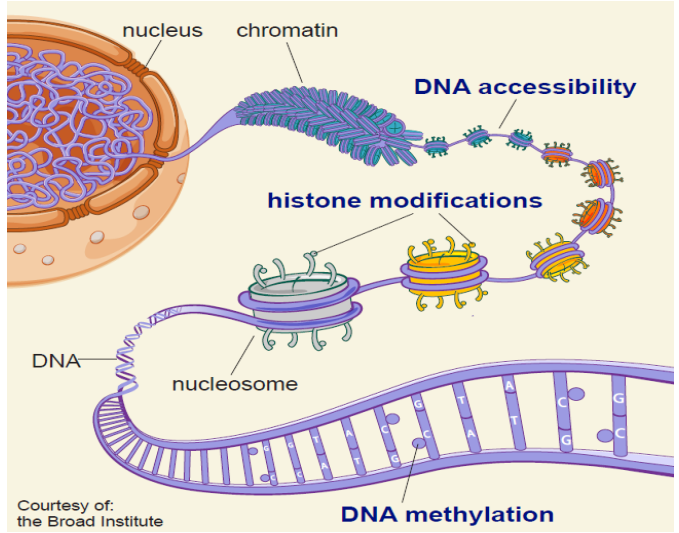
ES cells, iPS, differentiated cells (meso/endo/ectoderm, neural, mesench…)

ROADMAP epigenomics PROJECT

ENCODE

X

Diverse epigenomic assays



Histone modifications
- **H3K4me3, H3K4me1, H3K36me3**
- **H3K27me3, H3K9me3, H3K27/9ac**
- **+20 more**

Open chromatin:
- **DNA accessibility**

DNA methylation:
- **WGBS, RRBS, MRE/MeDIP**

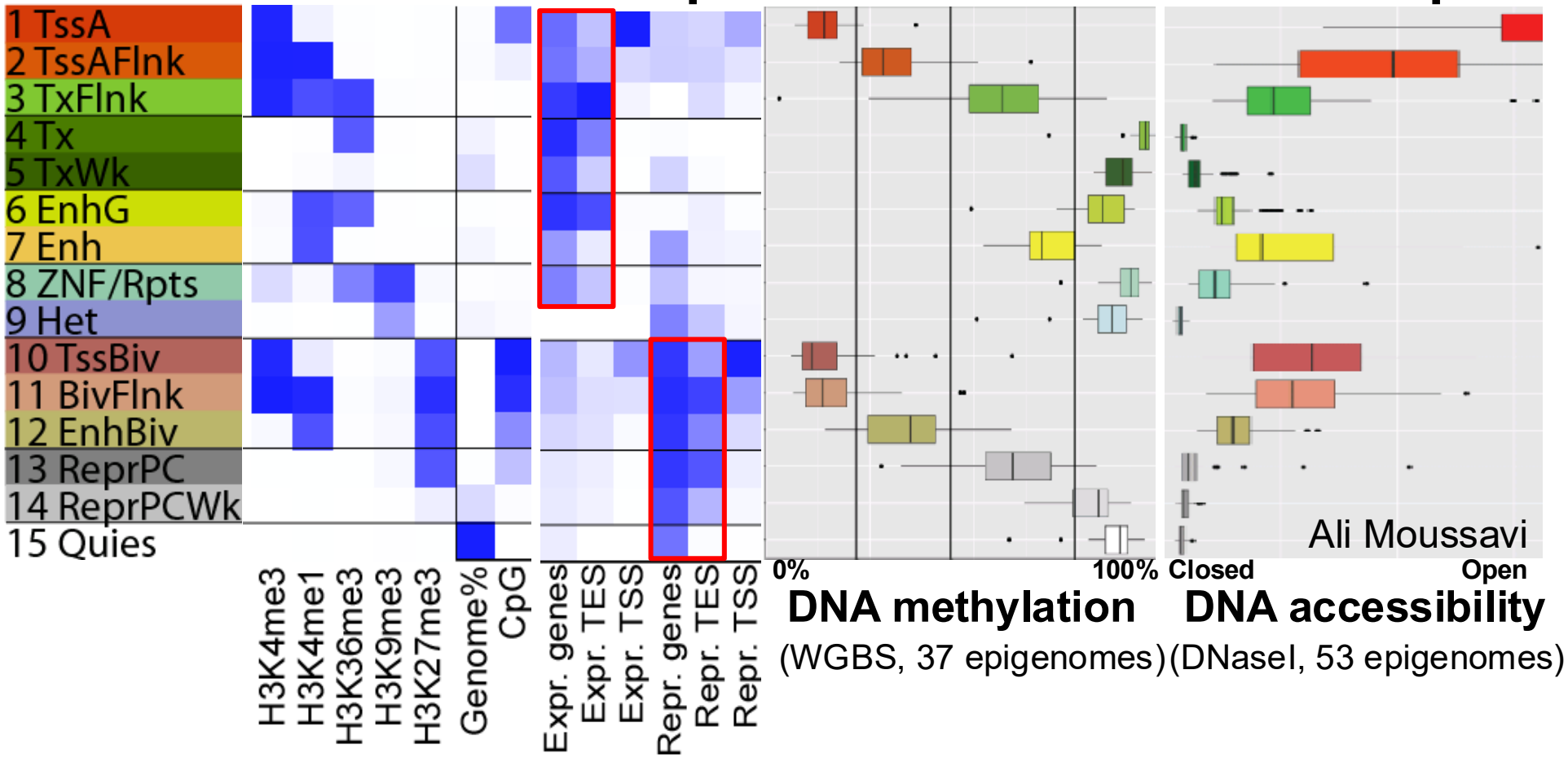
Gene expression
- **RNA-seq, Exon Arrays**

Chromatin states across 127 epigenomes

Reveal epigenomic variability: enh/prom/tx/repr/het

Anshul Kundaje

States show distinct mCpG, DNase, Tx, Ac profiles



TssA vs. TssBiv: diff. activity, both open, both unmethylated!

Enh vs. ReprPC: diff. activity, both intermediate DNase/Methyl

Tx: Methylated, closed, actively transcribed

➔ Distinct modes of repression: H3K27me3 vs. DNAme vs. Het

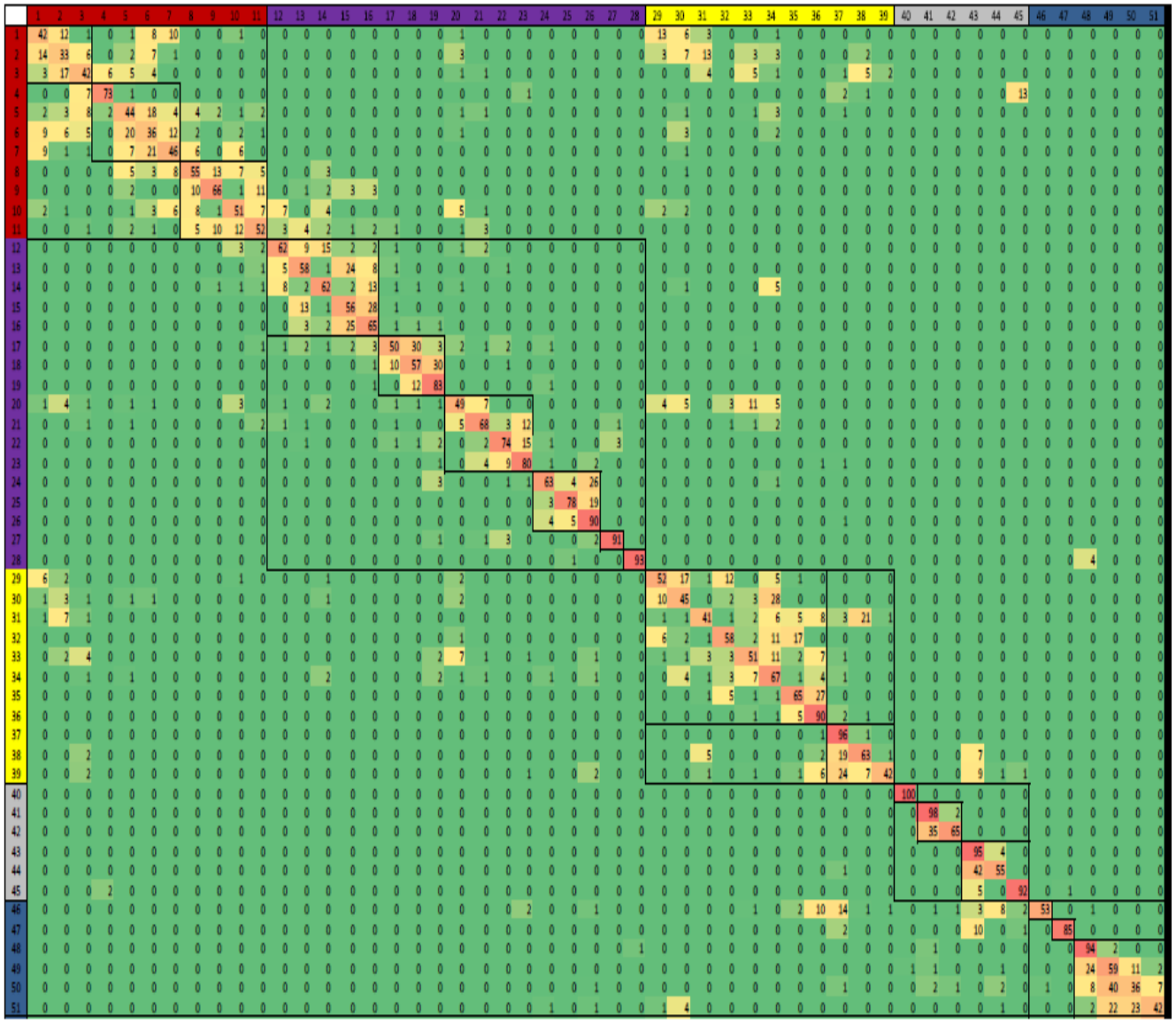
Emission Parameter Matrix $e_k(\boldsymbol{x}_i)$ to learn biology

state	H3K14ac	H3K23ac	H4K12ac	H2AK9ac	H4K16ac	H2AK5ac	H4K91ac	H3K4ac	H2BK20ac	H3K18ac	H2BK120ac	H3K27ac	H2BK5ac	H2BK12ac	H3K36ac	H4K5ac	H4K8ac	H3K9ac	PolII	CTCF	H2AZ	H3K4me3	H3K4me2	H3K4me1	H3K9me1	H3K79me3	H3K79me2	H3K79me1	H3K27me1	H2BK5me1	H4K20me1	H3K36me3	H3K36me1	H3R2me1	H3R2me2	H3K27me2	H3K27me3	H4R3me2	H3K9me2	H3K9me3	H4K20me3
1	3.8	23.6	24.2	18.0	37.7	25.5	95.2	94.8	94.3	99.2	99.6	99.7	98.9	79.1	88.6	93.6	86.9	83.6	51.6	15.7	87.5	94.2	93.8	64.2	87.0	3.8	3.3	12.0	19.4	11.6	3.8	0.5	2.6	1.9	2.1	0.2	0.1	0.2	0.5	0.1	1.8
2	2.5	17.5	9.2	3.2	5.9	6.3	44.6	44.4	47.0	73.2	74.1	85.9	71.2	22.1	33.5	61.9	63.3	35.4	18.1	10.9	91.2	86.7	90.4	66.9	78.3	2.4	2.2	7.9	17.6	8.7	2.3	0.6	2.2	1.7	1.5	0.4	0.5	0.2	0.4	0.1	1.4
3	0.5	5.8	1.8	1.0	0.9	1.2	12.3	9.5	8.8	22.6	21.3	22.8	12.1	2.2	4.2	8.4	12.8	7.1	11.2	16.3	77.1	93.9	80.3	45.6	74.2	1.5	1.3	4.6	4.3	7.0	8.8	0.2	1.4	2.1	1.5	0.2	2.1	0.1	0.1	0.1	1.2
4	0.1	0.8	0.1	0.4	0.3	0.2	1.5	0.9	0.7	2.1	2.1	0.5	0.2	0.1	0.1	0.1	0.3	0.3	1.9	6.2	19.0	77.9	20.8	21.1	26.4	0.3	0.0	0.1	0.0	1.3	14.9	0.1	0.2	1.4	0.9	0.0	10.2	0.1	0.1	0.4	1.3
5	0.0	0.2	0.8	1.3	2.0	0.4	26.6	12.6	6.7	15.8	23.1	26.8	24.3	1.5	4.3	0.9	1.4	7.7	53.5	20.6	21.8	87.0	11.2	2.7	15.7	6.3	3.8	2.5	0.0	5.5	14.2	0.0	0.3	0.2	0.6	0.0	0.0	0.0	0.1	0.0	1.5
6	0.1	1.8	3.6	6.9	6.1	1.9	74.5	63.5	53.0	75.7	84.3	89.4	86.7	20.8	41.5	20.5	21.6	62.7	69.2	25.5	61.2	98.3	37.4	7.1	40.3	5.3	2.7	5.6	0.6	6.0	11.6	0.0	0.5	0.4	0.9	0.0	0.0	0.0	0.0	0.1	1.6
7	1.2	8.7	20.6	43.0	53.7	9.8	98.7	98.6	95.7	99.4	99.9	99.9	99.9	76.5	93.3	81.8	76.6	99.2	88.0	26.9	77.1	99.7	38.3	2.2	37.9	32.1	24.9	14.0	2.6	6.5	16.2	0.1	1.0	0.5	1.2	0.0	0.0	0.0	0.1	0.0	1.9
8	1.2	12.7	5.2	11.9	5.6	6.8	56.9	56.1	37.5	52.4	69.8	89.1	85.5	21.3	24.8	16.7	10.3	60.8	62.1	12.0	31.4	96.7	51.7	14.9	45.3	86.3	80.1	23.8	1.7	6.7	42.2	4.5	0.4	1.1	0.9	0.0	0.0	0.0	0.0	0.0	0.5
9	0.5	7.2	3.0	1.1	0.5	2.1	4.0	7.4	2.4	2.5	11.6	35.3	28.0	2.7	2.8	2.0	1.8	8.6	34.7	4.2	4.5	79.2	41.5	23.8	36.1	86.0	82.6	12.0	1.9	6.7	43.5	7.4	0.2	0.6	0.7	0.0	0.0	0.0	0.0	0.2	0.4
10	4.1	24.8	13.1	17.5	24.4	37.0	90.4	88.6	82.0	89.8	95.5	97.0	95.1	54.0	56.4	67.2	45.7	55.7	46.6	10.2	40.8	84.6	92.3	91.4	92.8	67.1	67.2	63.4	29.2	53.8	65.2	4.5	6.8	5.7	3.4	0.1	0.0	0.3	0.1	0.0	1.0
11	1.6	21.0	3.9	3.8	3.2	8.4	28.0	26.1	14.5	22.6	37.6	56.8	47.4	6.0	6.4	13.5	8.8	18.4	30.9	6.9	20.1	92.4	93.5	94.3	94.5	73.8	74.2	55.1	24.0	57.2	79.9	8.9	6.6	4.4	2.5	0.1	0.0	0.2	0.1	0.1	0.7
12	3.6	17.0	8.9	2.2	14.1	34.9	60.3	51.0	38.8	35.6	56.6	53.3	55.3	11.9	11.5	30.0	15.4	1.7	18.0	2.7	0.7	5.4	58.2	96.0	77.8	87.4	87.0	76.3	41.6	79.8	82.2	13.4	3.1	6.4	3.6	0.3	0.0	0.7	0.2	0.1	0.4
13	1.2	10.8	3.6	0.7	2.5	7.4	9.1	6.5	2.6	2.5	6.5	7.7	5.5	0.5	1.0	5.5	3.3	0.3	10.2	1.5	0.0	2.4	56.2	83.7	82.9	92.7	92.6	64.4	38.2	80.0	89.8	12.0	2.4	3.3	2.1	0.3	0.0	0.4	0.2	0.1	0.3
14	0.7	5.3	7.9	1.0	2.4	18.0	19.8	20.7	14.6	7.9	24.5	20.1	21.8	6.7	6.6	11.3	8.6	0.3	6.9	1.7	2.1	1.3	8.0	33.0	16.3	61.8	62.1	37.9	9.7	14.3	18.3	9.7	0.2	1.7	1.2	0.2	0.0	0.1	0.3	0.4	1.0
15	0.2	1.9	2.9	0.3	0.3	1.5	0.8	1.3	0.3	0.2	1.2	1.5	1.2	0.2	0.4	0.7	1.2	0.1	5.0	0.7	0.0	0.2	11.0	17.3	29.3	84.7	82.2	33.1	8.0	26.4	56.2	5.2	0.2	0.7	0.9	0.1	0.0	0.1	0.1	0.6	0.2
16	0.0	0.3	0.4	0.1	0.1	0.5	0.4	0.6	0.2	0.1	0.5	0.4	0.3	0.1	0.2	0.2	0.3	0.0	0.6	0.3	0.1	0.1	1.2	2.8	3.9	29.0	25.2	8.5	0.7	1.3	7.9	1.2	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.4	0.1
17	1.2	9.8	2.8	0.9	2.4	7.8	6.8	6.1	2.3	4.0	3.5	8.4	3.5	0.3	1.0	9.3	6.6	0.5	3.5	1.1	0.4	1.0	52.3	68.9	83.7	22.7	23.5	61.2	48.3	64.7	57.3	21.8	2.8	4.9	2.0	1.0	0.1	0.5	0.5	0.1	0.5
18	0.3	2.6	2.1	0.4	0.5	2.4	1.4	1.6	0.5	0.6	0.9	1.6	0.7	0.2	0.5	1.9	1.9	0.1	1.6	0.7	0.1	0.1	10.4	9.7	29.1	15.0	13.5	34.0	14.9	19.9	21.4	10.6	0.5	1.2	0.9	0.5	0.1	0.1	0.3	0.2	0.2
19	0.1	0.3	0.5	0.2	0.1	0.5	0.1	0.3	0.0	0.0	0.1	0.2	0.1	0.1	0.1	0.1	0.3	0.0	0.4	0.3	0.0	0.0	0.5	0.2	2.0	4.4	2.9	7.3	1.0	1.0	2.3	1.4	0.0	0.2	0.3	0.1	0.0	0.0	0.1	0.4	0.1
20	2.5	10.7	5.4	3.1	9.9	26.2	58.2	48.8	41.7	49.3	54.8	57.1	51.5	13.0	14.1	31.6	21.7	4.0	14.5	6.7	15.6	20.9	56.8	97.1	70.5	5.4	5.6	33.8	31.1	52.6	38.4	7.4	3.4	6.9	3.8	0.5	0.1	0.7	0.3	0.0	1.0
21	0.2	0.8	2.0	1.3	7.2	11.3	32.3	15.4	11.5	5.7	18.6	8.4	12.4	2.1	1.2	3.2	2.3	0.5	15.1	6.5	0.6	4.7	17.0	68.7	33.3	8.7	6.4	37.2	9.6	65.4	87.7	9.2	1.3	7.1	5.3	0.1	0.2	1.4	0.0	0.0	0.8
22	0.1	0.1	1.1	0.6	6.2	2.4	7.8	1.8	0.6	0.1	1.5	0.8	1.5	0.1	0.1	0.7	0.7	0.1	8.5	1.0	0.0	0.0	5.3	8.4	14.6	15.5	9.1	50.0	9.6	77.5	94.1	22.9	0.5	5.6	4.6	0.1	0.0	1.4	0.0	0.1	0.7
23	0.0	0.1	0.1	0.4	2.0	1.6	4.9	1.2	0.5	0.2	0.9	0.2	0.3	0.0	0.1	0.1	0.1	0.0	1.4	1.1	0.0	0.0	0.5	2.6	1.4	0.5	0.1	5.4	1.3	19.4	36.8	2.5	0.1	1.4	1.5	0.1	0.2	0.3	0.0	0.0	0.2
24	0.3	1.8	2.1	0.9	3.2	3.8	4.0	2.3	0.9	0.6	1.1	3.6	1.9	0.1	0.4	3.7	3.9	0.3	2.2	1.0	0.1	0.1	6.0	4.5	17.2	1.3	0.2	15.6	29.8	29.3	7.1	49.5	1.3	4.7	2.2	1.0	0.0	0.6	0.3	0.1	0.3
25	0.1	0.3	0.8	0.5	0.5	0.6	0.3	0.3	0.1	0.0	0.1	0.5	0.3	0.1	0.1	0.4	0.7	0.1	0.8	0.4	0.0	0.1	0.4	0.1	1.4	0.8	0.1	2.0	8.8	2.8	0.4	60.1	0.4	1.1	0.9	0.6	0.1	0.2	0.3	1.7	0.3
26	0.1	0.2	0.6	0.2	0.2	0.2	0.1	0.2	0.0	0.0	0.1	0.3	0.2	0.0	0.1	0.2	0.4	0.0	0.6	0.3	0.0	0.0	0.3	0.1	0.8	0.2	0.0	0.8	2.3	0.9	0.2	4.2	0.1	0.2	0.4	0.1	0.0	0.0	0.1	0.1	0.0
27	0.0	0.5	4.4	0.4	1.3	1.2	1.3	0.7	0.3	0.1	0.7	2.1	2.4	0.1	0.1	1.6	2.7	0.1	21.7	1.4	0.0	0.0	1.1	1.1	3.5	4.6	1.2	9.9	3.0	7.1	31.7	34.0	0.2	0.7	1.1	0.0	0.0	0.1	0.0	0.3	0.1
28	0.0	0.0	0.2	0.3	0.0	0.4	0.1	0.2	0.2	0.0	0.3	0.1	0.1	0.1	0.0	0.1	0.1	0.0	1.3	0.3	0.0	3.3	0.2	0.5	0.5	13.8	3.4	5.7	1.3	3.0	1.5	68.8	0.1	2.7	2.7	0.3	0.2	0.8	0.4	43.0	74.9
29	4.6	8.4	11.1	6.6	20.4	54.7	88.5	88.1	89.6	86.6	95.3	86.3	86.9	68.1	60.2	67.6	42.6	10.4	13.6	3.8	24.2	7.6	24.7	84.4	25.8	4.9	5.6	14.9	17.6	21.7	5.0	2.9	0.9	4.8	3.1	0.4	0.1	0.4	0.8	0.2	4.4
30	1.2	3.6	8.4	2.4	2.6	13.5	24.9	34.5	34.4	24.6	52.1	60.1	64.6	27.4	23.8	20.7	16.7	3.2	9.1	2.9	12.0	6.9	8.8	35.8	6.5	2.8	2.9	4.4	3.7	3.0	1.0	2.8	0.1	0.9	1.0	0.0	0.1	0.1	0.4	0.6	3.4
31	1.7	7.6	4.7	1.7	2.4	6.8	17.7	18.4	21.5	37.9	31.6	35.4	20.1	9.3	13.0	48.3	57.7	3.3	1.2	5.9	69.0	10.5	16.1	41.8	11.0	0.6	0.5	1.2	10.7	2.2	0.0	1.1	1.4	2.0	1.3	1.1	0.5	0.2	0.8	0.2	1.1
32	1.4	1.6	1.0	1.9	8.1	50.1	72.4	57.2	60.9	42.5	57.6	12.1	14.8	23.6	19.5	16.0	5.4	0.3	1.6	1.7	3.6	0.1	2.1	41.4	5.0	1.8	1.7	12.2	10.9	17.1	4.1	4.1	0.9	6.5	3.0	1.3	0.3	0.6	0.5	0.3	3.8
33	1.2	4.6	0.9	0.9	1.2	9.8	12.9	10.3	7.0	12.7	10.5	9.8	5.7	1.3	2.1	6.3	4.1	0.4	2.6	4.3	8.6	1.5	16.3	77.1	23.5	0.7	0.5	5.1	10.3	10.8	3.9	1.8	1.1	2.9	1.6	0.7	0.3	0.2	0.4	0.1	0.4
34	0.2	0.7	2.0	0.5	0.7	5.4	7.2	6.2	4.9	1.2	8.9	5.7	6.2	2.9	2.2	2.3	2.3	0.1	1.5	0.9	1.0	0.2	0.6	6.3	1.0	1.9	1.7	5.2	2.6	2.2	0.6	1.4	0.0	0.5	0.6	0.0	0.0	0.0	0.1	0.2	1.0
35	0.6	1.1	0.1	1.1	1.3	32.1	29.9	20.0	19.6	16.4	11.8	1.1	0.5	3.2	5.4	4.0	1.5	0.2	0.3	1.0	2.0	0.1	0.3	2.6	0.6	0.4	0.2	1.9	2.6	2.2	0.5	1.7	0.9	4.2	2.0	2.6	0.8	0.3	0.5	0.2	0.6
36	0.2	0.3	0.0	0.5	0.2	5.7	2.5	2.1	1.0	1.6	0.8	0.1	0.1	0.3	0.6	0.3	0.1	0.1	0.1	0.3	0.3	0.0	0.1	0.3	0.1	0.2	0.0	0.2	0.4	0.4	0.1	0.4	0.3	1.3	0.8	0.8	0.5	0.1	0.3	0.3	0.1
37	0.1	0.2	0.0	0.2	0.1	0.7	0.1	0.2	0.0	0.1	0.1	0.1	0.0	0.0	0.1	0.1	0.1	0.0	0.1	0.2	0.2	0.0	0.0	0.1	0.1	0.1	0.1	0.0	0.1	0.1	0.0	0.3	0.1	0.5	0.4	0.2	0.2	0.1	0.1	0.2	0.1
38	0.4	2.5	0.6	0.8	0.2	1.9	1.0	1.9	1.4	3.8	2.2	2.4	0.9	0.8	1.4	4.8	9.2	0.4	0.1	0.7	33.0	1.3	1.5	5.9	0.9	0.2	0.1	0.2	1.7	0.3	0.0	0.3	0.8	1.0	0.9	1.0	1.1	0.2	0.6	0.2	0.3
39	0.1	0.4	0.5	0.6	0.1	1.2	0.7	0.8	1.1	0.9	1.0	0.4	0.2	0.3	0.4	1.2	2.0	0.1	4.1	86.2	12.3	1.4	1.6	5.0	1.2	0.2	0.0	0.3	0.6	0.7	0.7	0.2	0.8	2.5	1.2	0.2	0.8	0.1	0.0	0.2	0.6
40	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0
41	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.1	0.5	0.3	0.1	0.1	1.0	2.0	0.4
42	0.1	0.1	0.0	0.8	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.2	0.1	0.6	0.6	0.1	0.0	0.1	0.1	0.1	0.0	0.0	1.8	0.5	0.0	0.5	1.3	2.0	2.3	8.7	1.2	0.8	5.6	2.3	4.4
43	0.0	0.2	0.0	0.3	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.3	0.4	1.0	0.1	0.5	0.5	0.1
44	0.6	1.1	0.0	1.2	0.2	0.7	0.1	0.4	0.0	0.4	0.1	0.1	0.0	0.0	0.2	0.1	0.1	0.3	0.1	0.6	2.0	0.2	0.2	0.3	0.3	0.1	0.0	0.0	0.3	1.1	0.0	1.6	3.3	5.7	3.1	9.0	9.9	0.9	2.5	0.4	1.0
45	0.1	0.7	0.0	0.5	0.0	0.4	0.1	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.6	0.6	0.4	0.1	0.4	0.1	0.1	0.0	0.0	0.0	0.1	0.2	0.3	0.5	1.0	0.9	0.4	16.4	0.2	0.3	1.1	0.4
46	1.6	0.7	0.4	1.4	4.2	6.1	5.2	3.6	0.8	4.5	2.4	0.7	0.6	0.4	0.5	0.7	0.4	0.5	0.6	2.0	1.1	2.7	2.1	28.4	5.5	0.4	0.1	1.1	8.4	63.8	16.5	5.9	17.8	48.9	25.0	9.8	7.8	11.9	3.7	1.6	17.0
47	0.0	0.1	0.0	0.2	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.3	0.0	0.6	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	1.9	0.1	0.4	0.7	0.3	3.1	0.1	0.7	32.3	1.8
48	0.0	0.0	0.0	0.2	0.0	0.1	0.0	0.1	0.1	0.0	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.2	0.1	0.0	0.1	0.0	0.3	0.0	0.1	0.2	0.1	0.0	1.0	0.1	0.5	0.9	0.6	0.3	0.7	0.7	11.9	37.8
49	2.3	0.5	0.6	4.3	9.5	0.2	0.3	0.3	0.2	0.2	0.3	0.1	0.2	0.2	1.4	0.4	0.7	2.4	0.9	3.0	4.3	3.0	0.3	0.5	0.7	1.2	0.1	0.1	10.4	4.4	0.3	16.1	9.0	12.5	38.0	15.3	5.6	49.1	24.5	49.3	85.1
50	21.0	4.2	7.8	16.1	42.7	1.8	1.1	2.0	0.9	0.8	1.1	0.5	0.7	1.5	6.2	2.4	2.9	12.7	14.1	25.3	16.8	36.0	8.4	17.5	14.1	8.3	1.5	1.3	60.0	40.0	9.2	80.0	68.2	70.4	84.9	64.8	28.1	87.5	76.1	79.8	97.1
51	78.3	38.3	65.7	68.2	92.6	20.6	21.4	30.6	27.7	19.7	20.9	12.5	18.3	25.3	54.7	34.7	43.4	65.5	69.6	79.8	72.7	95.1	60.9	73.4	75.8	48.8	24.9	25.3	95.2	85.3	59.0	96.0	96.5	98.4	98.6	90.5	88.5	97.2	96.4	92.5	99.6

- Multi-variate HMM emits vector of values, not just one value
- Can emit real values (SegWay) or binary presence /absence values (ChromHMM)
- Use to learn mark combinations

Ernst and Kellis
Nature Biotech 2010

Transition matrix a_{kl}



- Learns spatial relationships between neighboring states
- Reveals distinct sub-groups of states
- Reveals transitions between different groups

Assay for Transposase-Accessible Chromatin (ATAC)

Goal: Map open chromatin regions genome-wide

Key insight: No antibody needed!

Protocol:

1. Tn5 transposase + DNA
2. Cuts accessible chromatin preferentially
3. Adds sequencing adapters
4. Sequence DNA fragments
5. Map reads to genome
6. **Call peaks**

Advantages:

- Fast protocol (~3 hours)
- Small cell numbers (500-50K cells)
- No antibodies required
- Genome-wide view

ATAC-seq vs. ChIP-seq

ATAC-seq:

- Measures chromatin accessibility
- All open regions (enhancers, promoters, etc.)
- No protein specificity
- No antibody needed
- Cheaper, faster

Use cases:

- Map all regulatory elements
- Compare cell states
- Find active enhancers
- TF footprinting

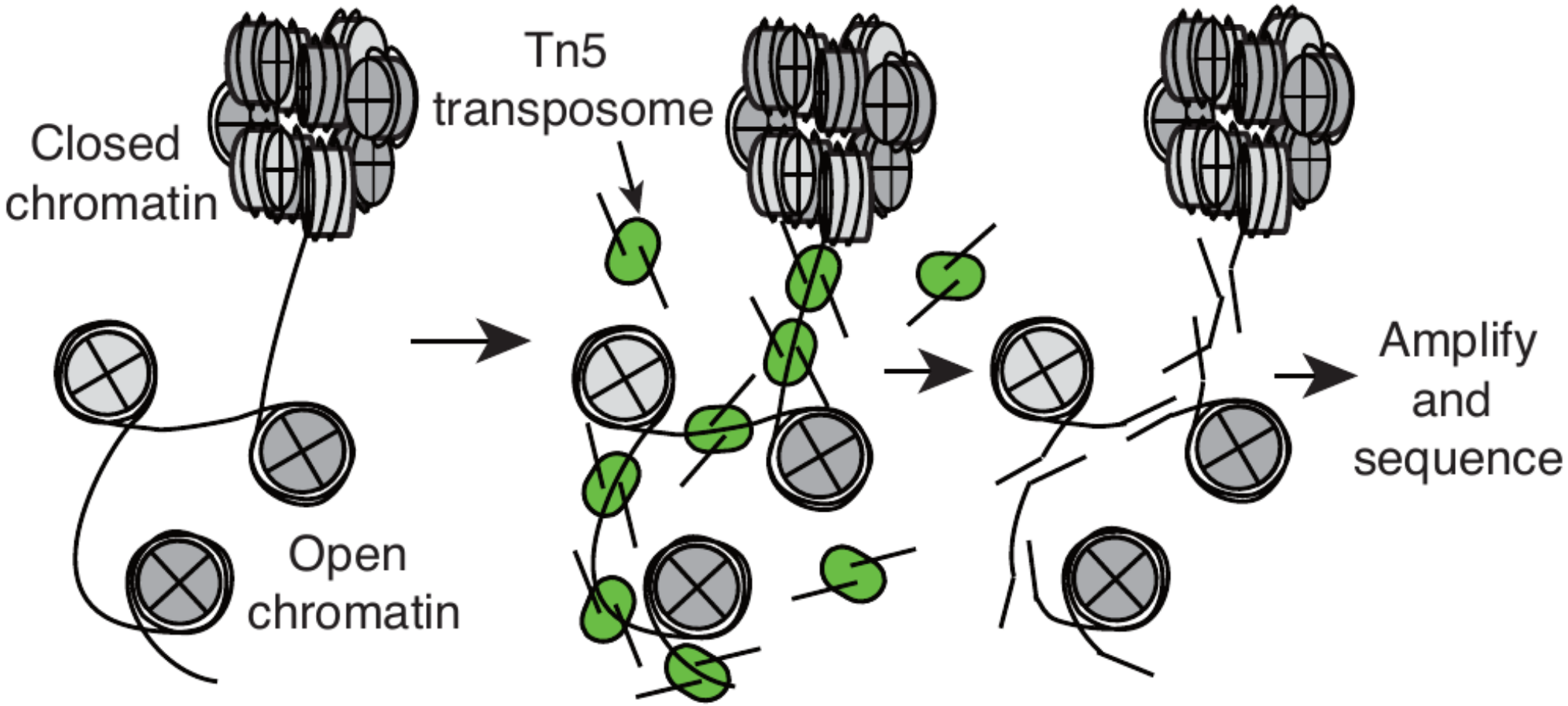
ChIP-seq:

- Measures protein-DNA binding
- Specific protein of interest
- Requires good antibody
- More expensive

Use cases:

- Find TF binding sites
- Map histone modifications
- Validate ATAC peaks
- Mechanistic studies

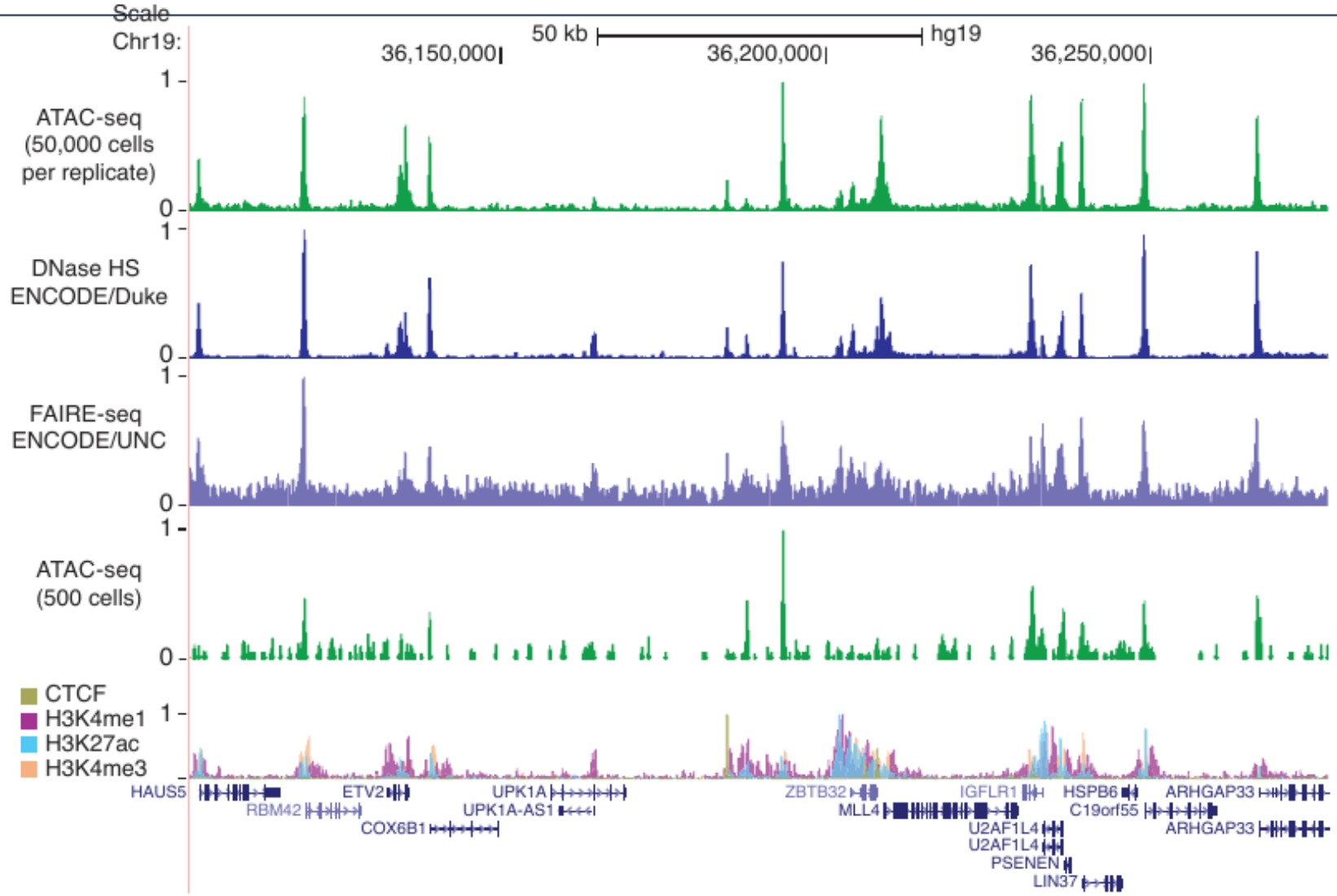
ATAC-seq produces regular fragment sizes



Tn5: **cuts** open DNA and **insert** adapter sequence.

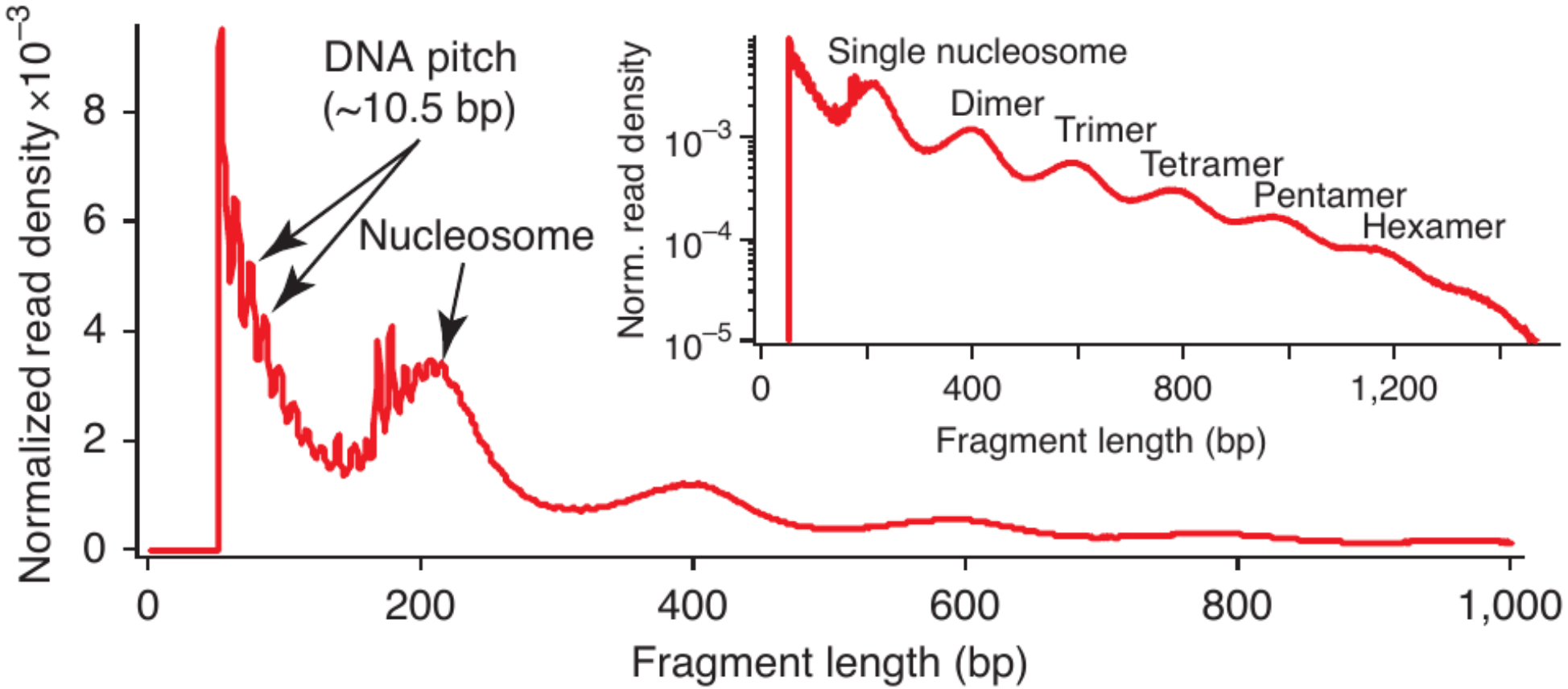
(Buenrostro et al., 2013)

ATAC-seq produces regular fragment sizes



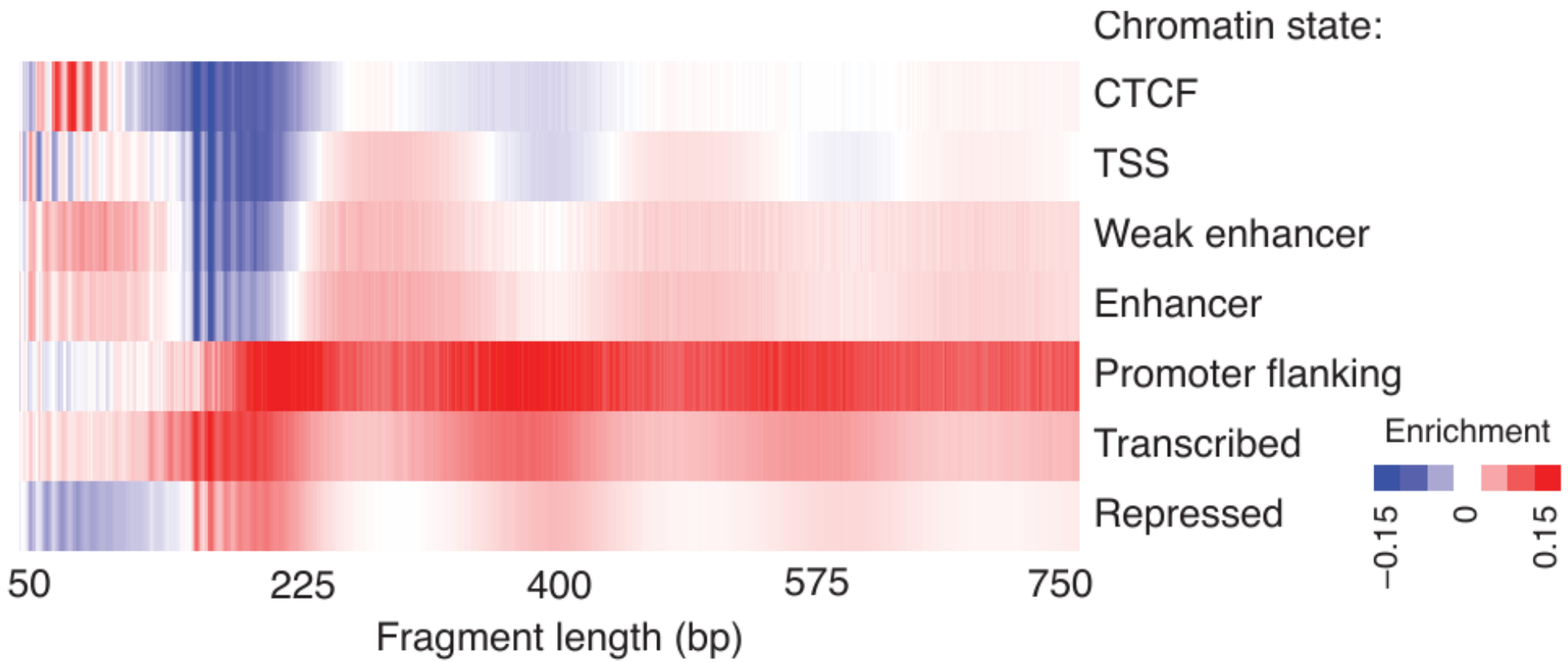
(Buenrostro et al. 2013)

ATAC-seq produces regular fragment sizes



(Buenrostro et al. 2013)

ATAC-seq produces regular fragment sizes



ATAC-seq enriches known gene regulatory elements in genome

(Buenrostro et al. 2013)

Differential Accessibility Analysis

```r
# After peak calling, create count matrix
# Rows = genomic bins, Columns = samples
count_matrix <- matrix(...)  # Read counts per bin per sample

colData <- data.frame( # Set up experimental design
    condition = factor(c(rep("Control", 3), rep("Treatment", 3)))
)

dds <- DESeqDataSetFromMatrix(countData = count_matrix, # Run differential accessibility
                              colData = colData,        #  (same as RNA-seq!)
                              design = ~ condition)
dds <- DESeq(dds)
results <- results(dds, contrast = c("condition", "Treatment", "Control"))

sig_regions <- results[results$padj < 0.05, ] # Get significant regions (FDR < 0.05)
```

Key point: Same statistical framework as RNA-seq (DESeq2)!

ATAC-seq Applications

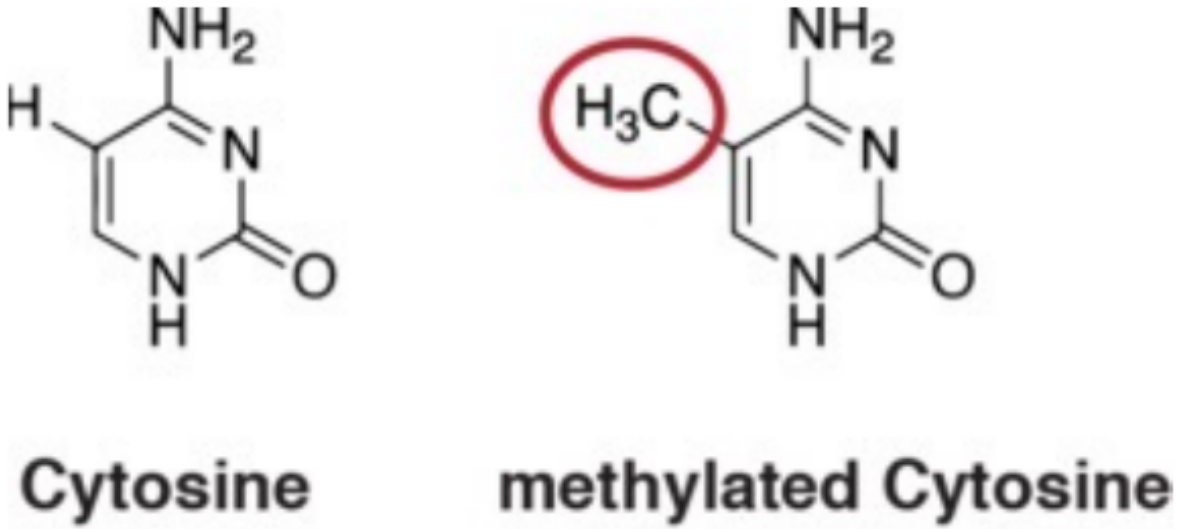
1. **Identify regulatory elements**
   - Promoters (at TSS)
   - Enhancers (distal peaks)
   - Silencers
   - Insulators
2. **Compare cell types/states**
   - Development
   - Differentiation
   - Disease vs healthy
3. **TF footprinting**
   - Bound TF $\rightarrow$ protected DNA
   - Dip in signal at TF motif
   - Infer active TFs
4. **Integration**
   - ATAC + RNA-seq $\rightarrow$ **link enhancers to genes**
   - ATAC + ChIP-seq $\rightarrow$ TF binding
   - ATAC + GWAS $\rightarrow$ **functional variants**

Today's lecture

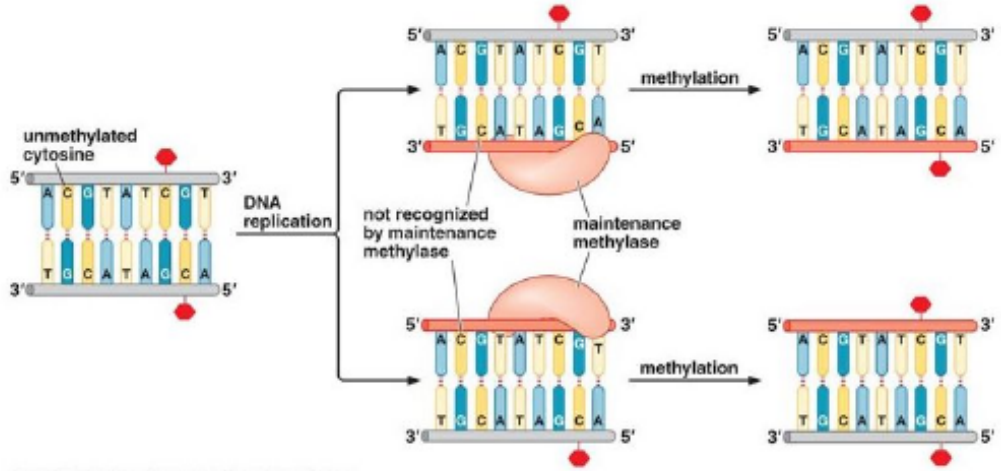
1. Technologies for measuring regulatory activities
2. DNA Methylation (credit: Keegan Korthauer)
3. Multiomic Data Integration with Regulatory Genomics

What is DNA Methylation?

- **5-methylcytosine (5mC)**: a methyl group added to cytosine at CpG dinucleotides
- **Heritable**: copied faithfully during DNA replication (by DNMT1)
- **Regulatory role**:
  - CpG islands at promoters → gene silencing when methylated
  - Enhancer activity correlates with hypomethylation
  - Imprinting, X-inactivation

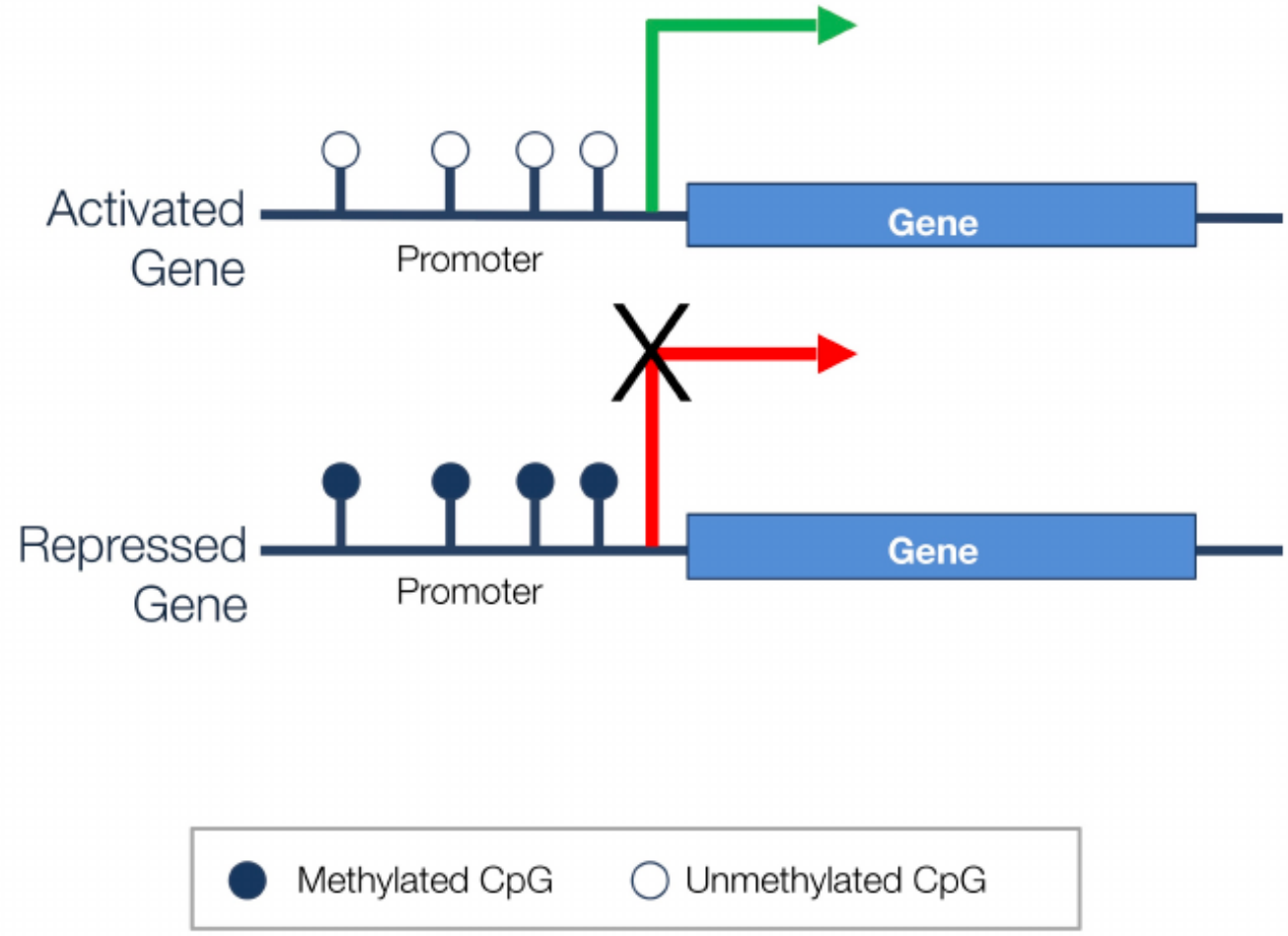


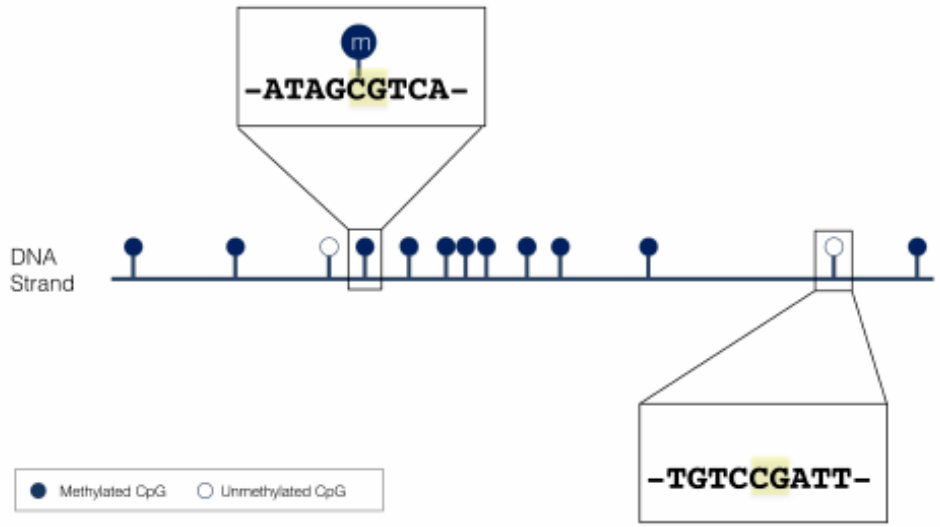
epigentek.com



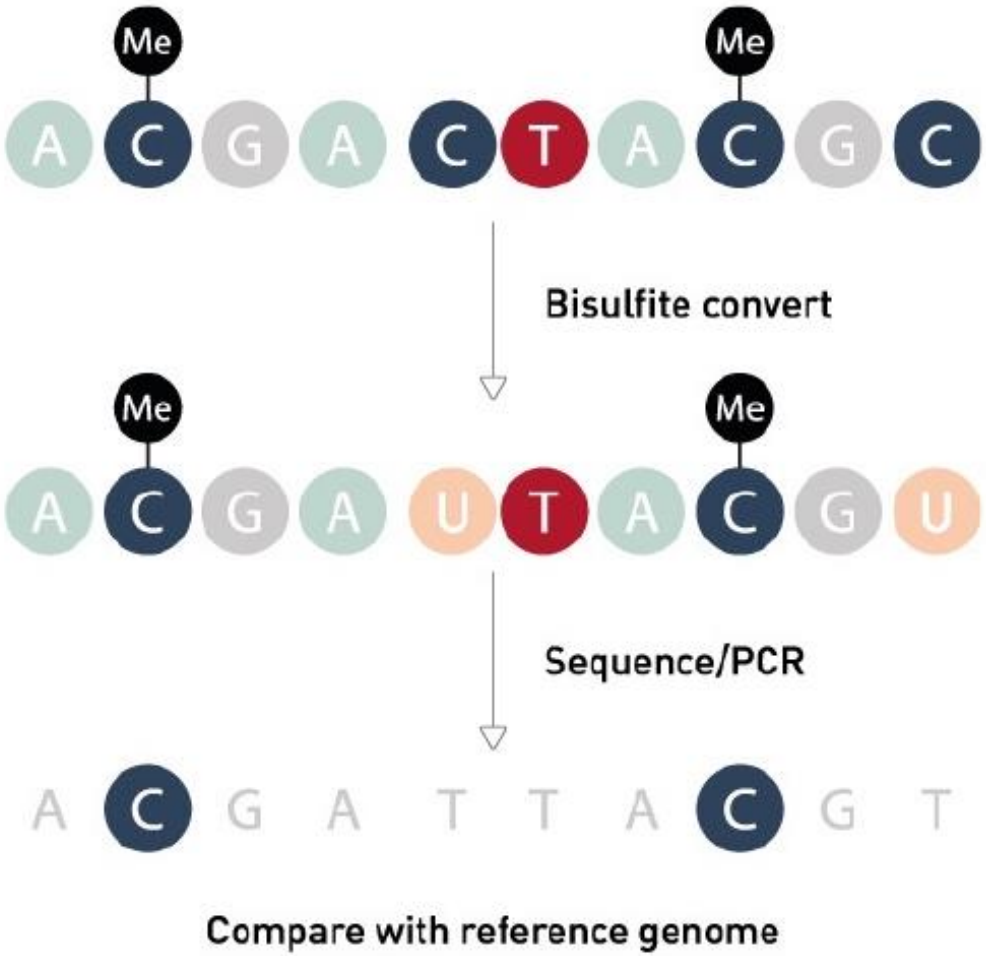
slideplayer.com

DNA Methylation and Gene Regulation

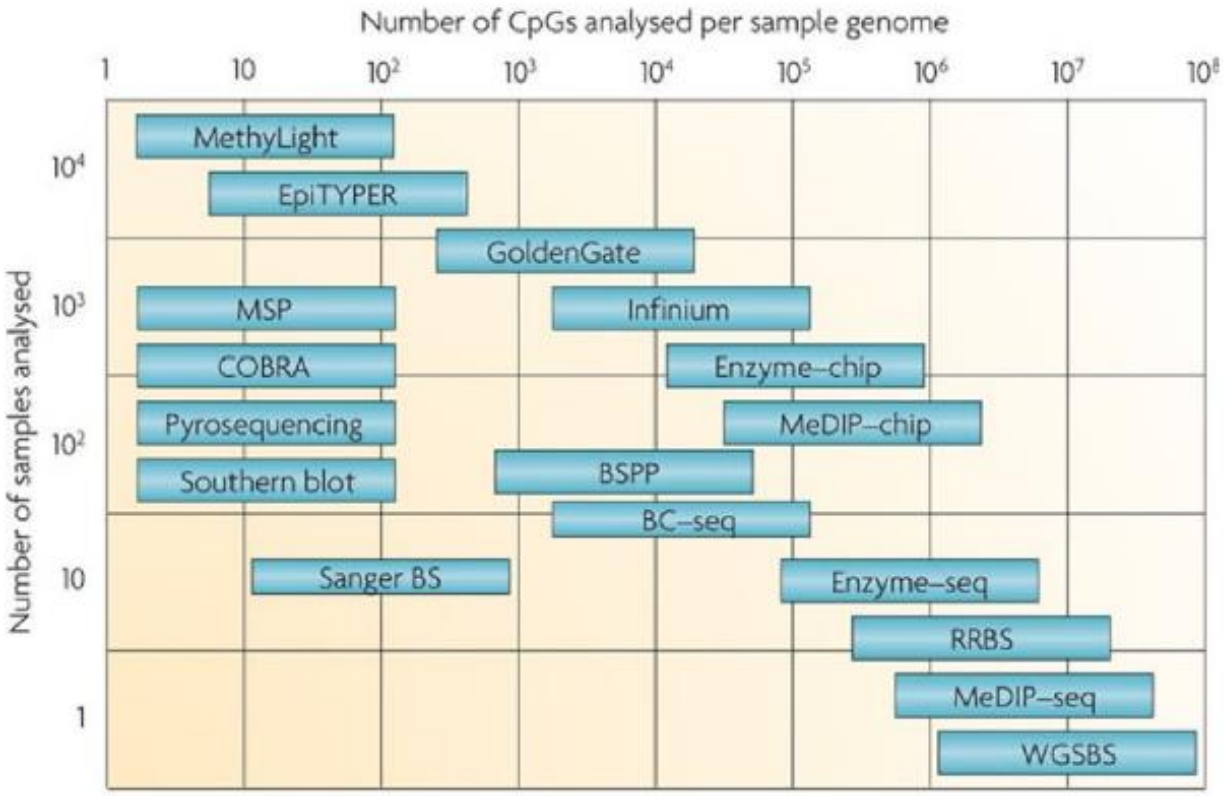




Measuring DNA Methylation

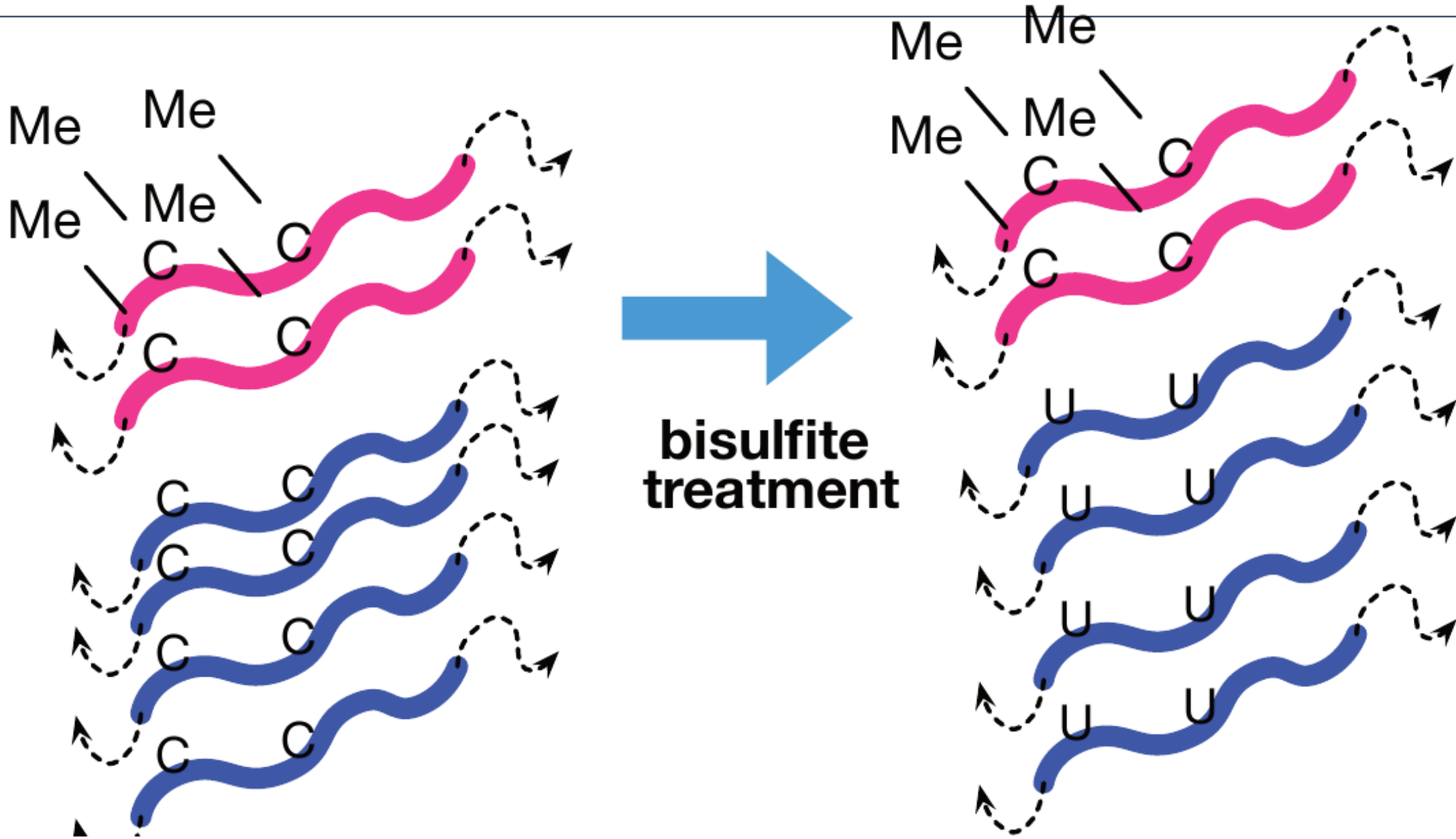


Bisulfite conversion (*diagenode.com*)



(Laird 2010)

Bisulfite Sequencing



Beta (β) values

- Combine two measurements for each CpG probe, measure level of methylation:

$$\beta = \frac{\text{methylated intensity}}{(\text{methylated intensity} + \text{unmethylated intensity})}$$

- Range from 0 to 1
- Easily interpretable as a **methylation proportion**
- Best used for **visualization**

M-values

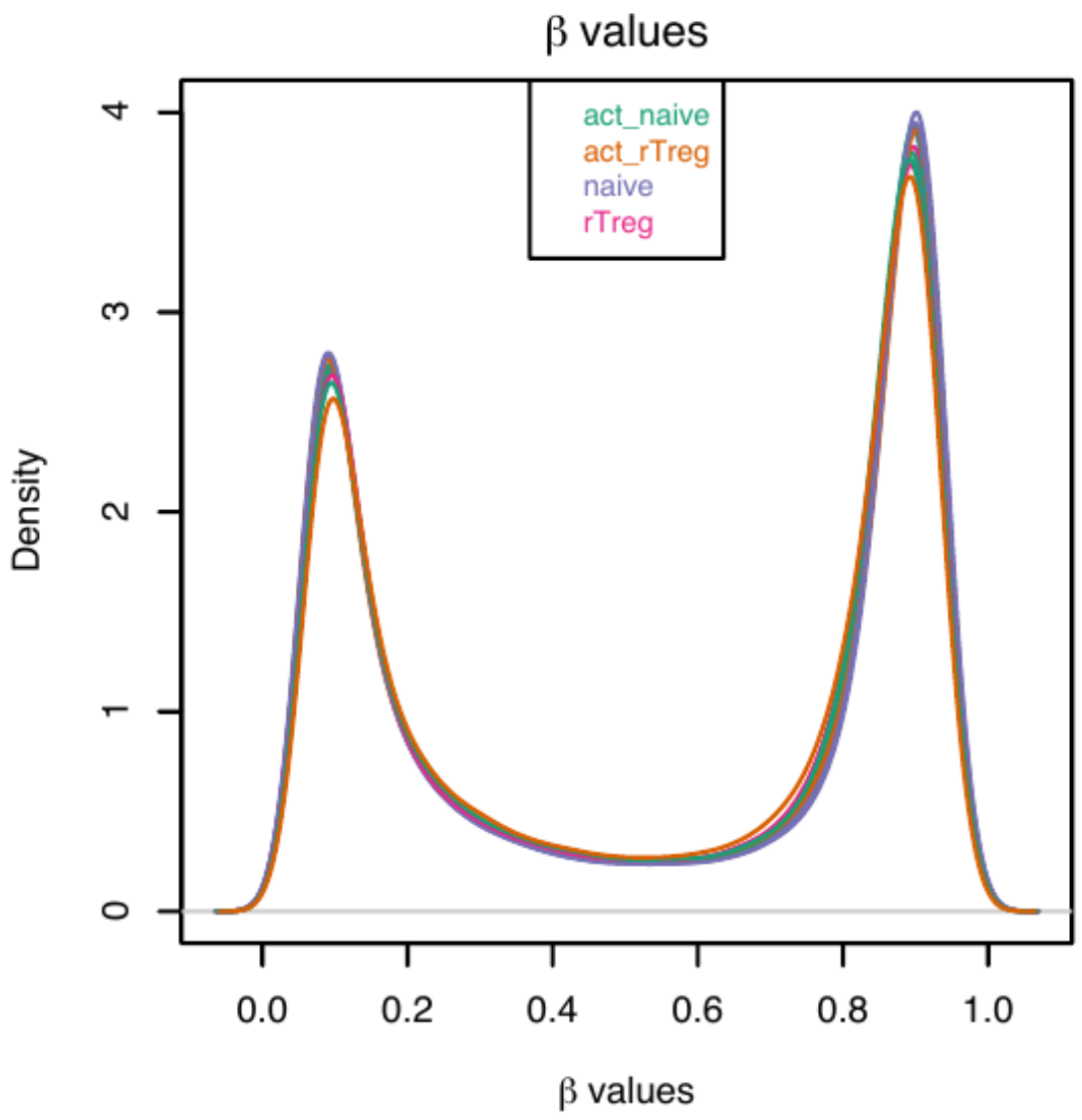
- Transformed β values:

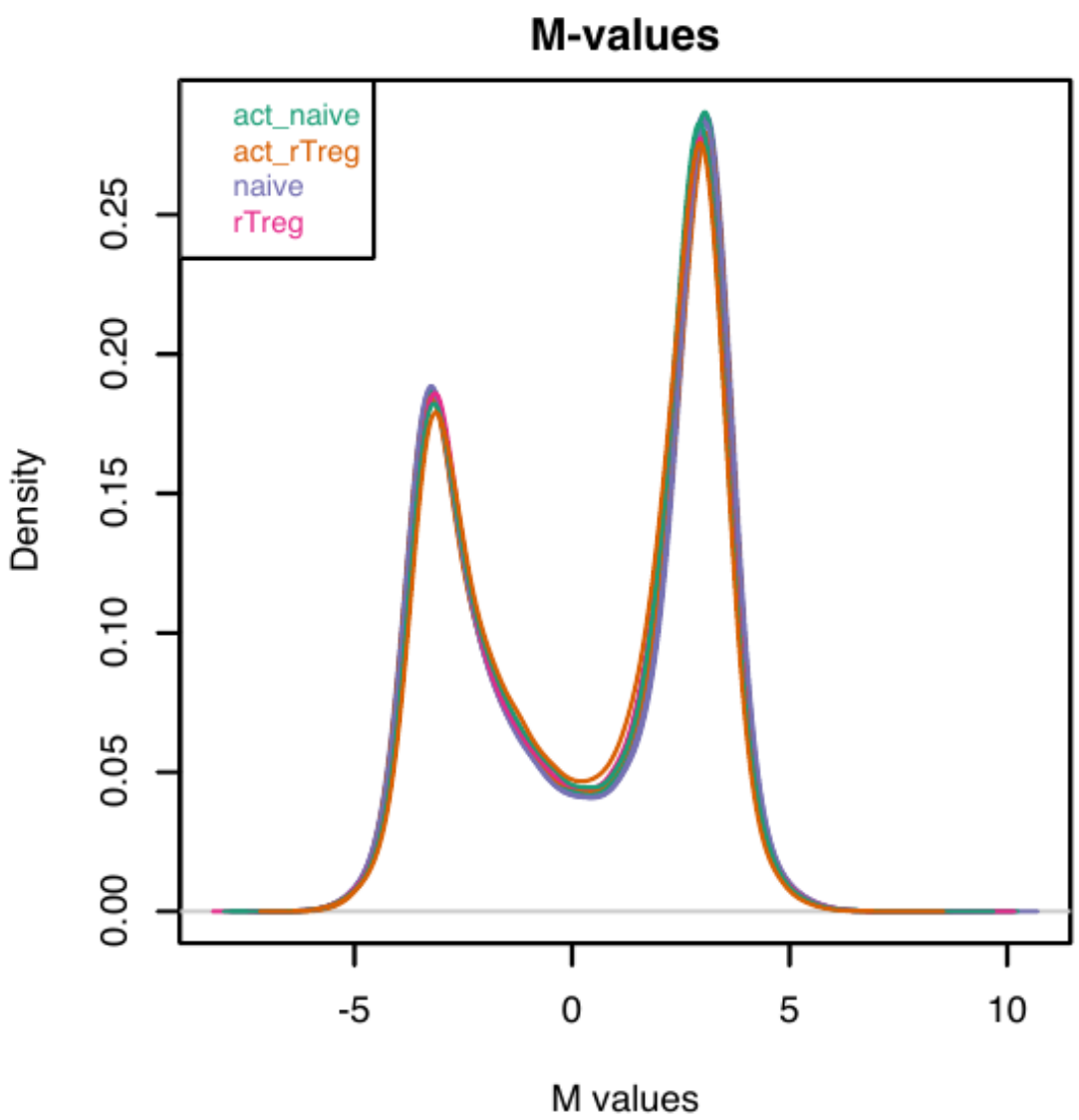
$$M = \log_2\left(\frac{\text{methylated intensity}}{\text{unmethylated intensity}}\right) = \log_2\left(\frac{\beta}{1-\beta}\right)$$

- The *logit* transformation — maps a proportion onto the real line
- Alleviates heteroscedasticity of β values (recall: Binomial(n, p) variance $= np(1-p)$)
- Best used for **statistical testing**

(Du et al. 2010)

Comparing β and M values





Methylation Array Analysis Workflow

- Analysis strategy: **linear regression on M-values** (limma)
- Key R packages:
  - `minfi` — preprocessing and normalization of array data
  - `limma` — differentially methylated cytosines (DMC)
  - `dmrseq` / `bumphunter` — differentially methylated **regions** (DMR)
- Example dataset: sorted T cells (Zhang et al. 2013)
  - Naive vs. activated naive T cells
  - GEO: GSE49667

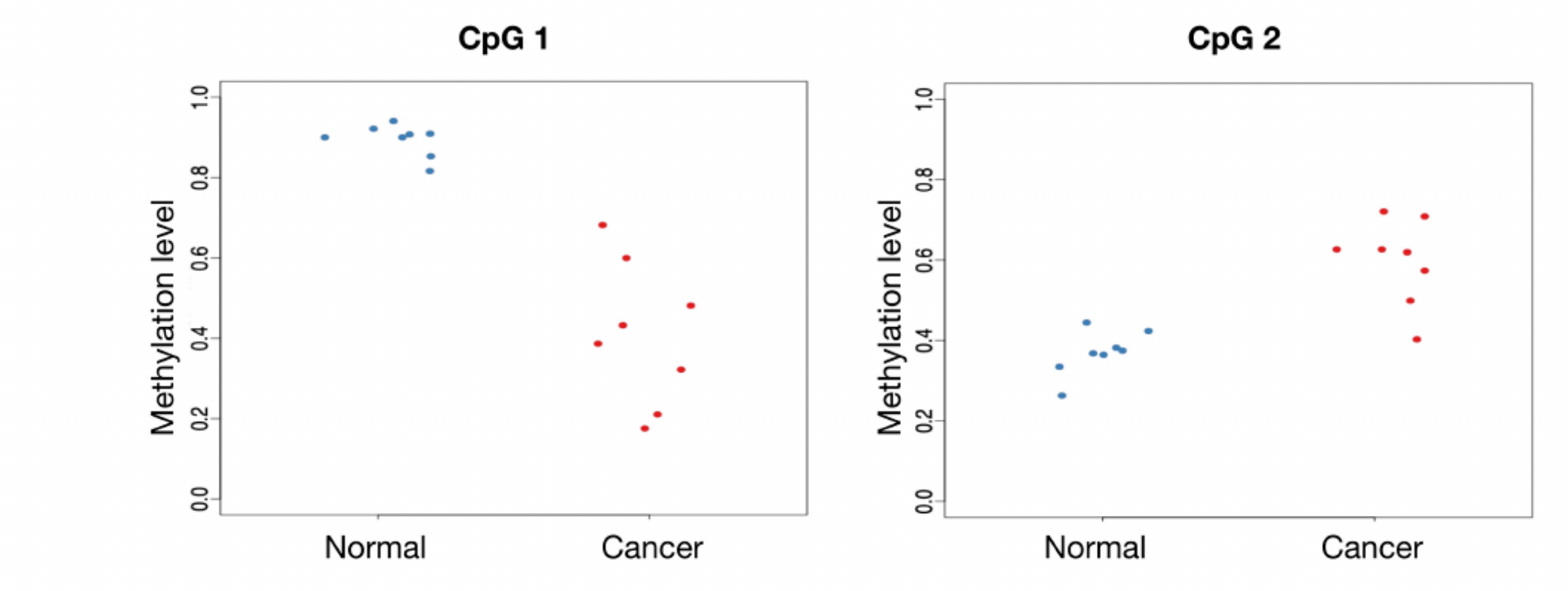
Differentially Methylated Cytosines (DMC)

```r
library(limma)
cellType  <- factor(targets$Sample_Group)
individual <- factor(targets$Sample_Source)
design <- model.matrix(~ cellType + individual, data = targets)

## Fit one linear model per CpG probe
mfit <- lmFit(mVals, design)
cfit <- eBayes(mfit)

## Top differentially methylated CpGs (naive vs. act_naive)
topTable(cfit, coef = "cellTypenaive", number = 5)
```

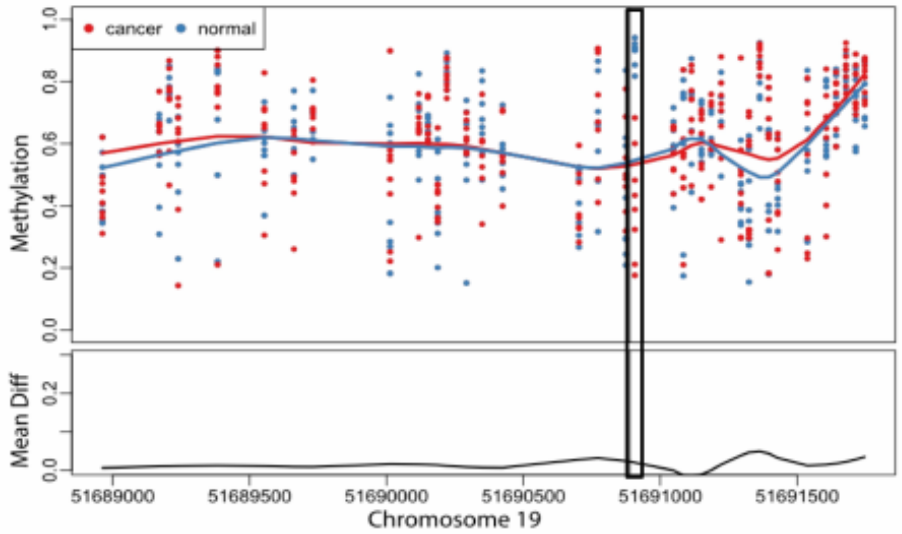
```
##                logFC     AveExpr         t      P.Value   adj.P.Val        B
## cg15459165 -3.892074 -1.47895212 -22.09541 1.993958e-08 0.002793543 6.810420
## cg19704755 -4.139885  0.09100138 -20.35779 3.787335e-08 0.002793543 6.594795
## cg13531460  3.484448 -0.63132888  20.23020 3.978221e-08 0.002793543 6.577256
## cg17137500 -3.007052 -2.47635924 -20.21569 4.000601e-08 0.002793543 6.575245
## cg17048073 -3.650550 -0.87247048 -20.04737 4.270889e-08 0.002793543 6.551674
```

DMCs vs DMRs

- Individual CpG tests → **differentially methylated cytosines (DMCs)**
- Methylation of nearby CpGs is **correlated** and regulated in tandem
- Grouping nearby significant CpGs → **differentially methylated regions (DMRs)**

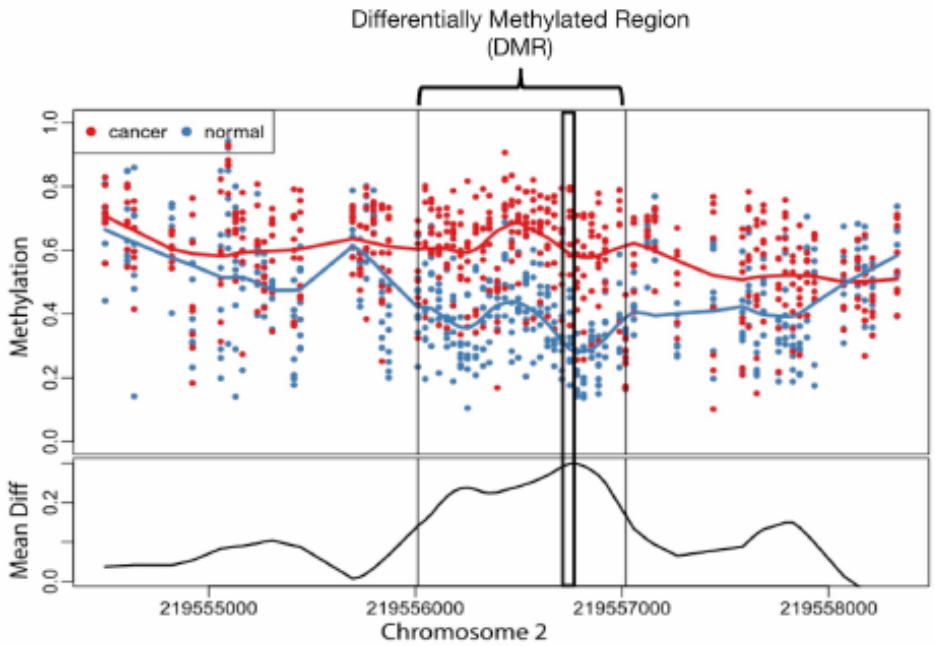


DMCs vs DMRs



Adapted from Jaffe et al., 2012 (*Int J Epidemiol*)

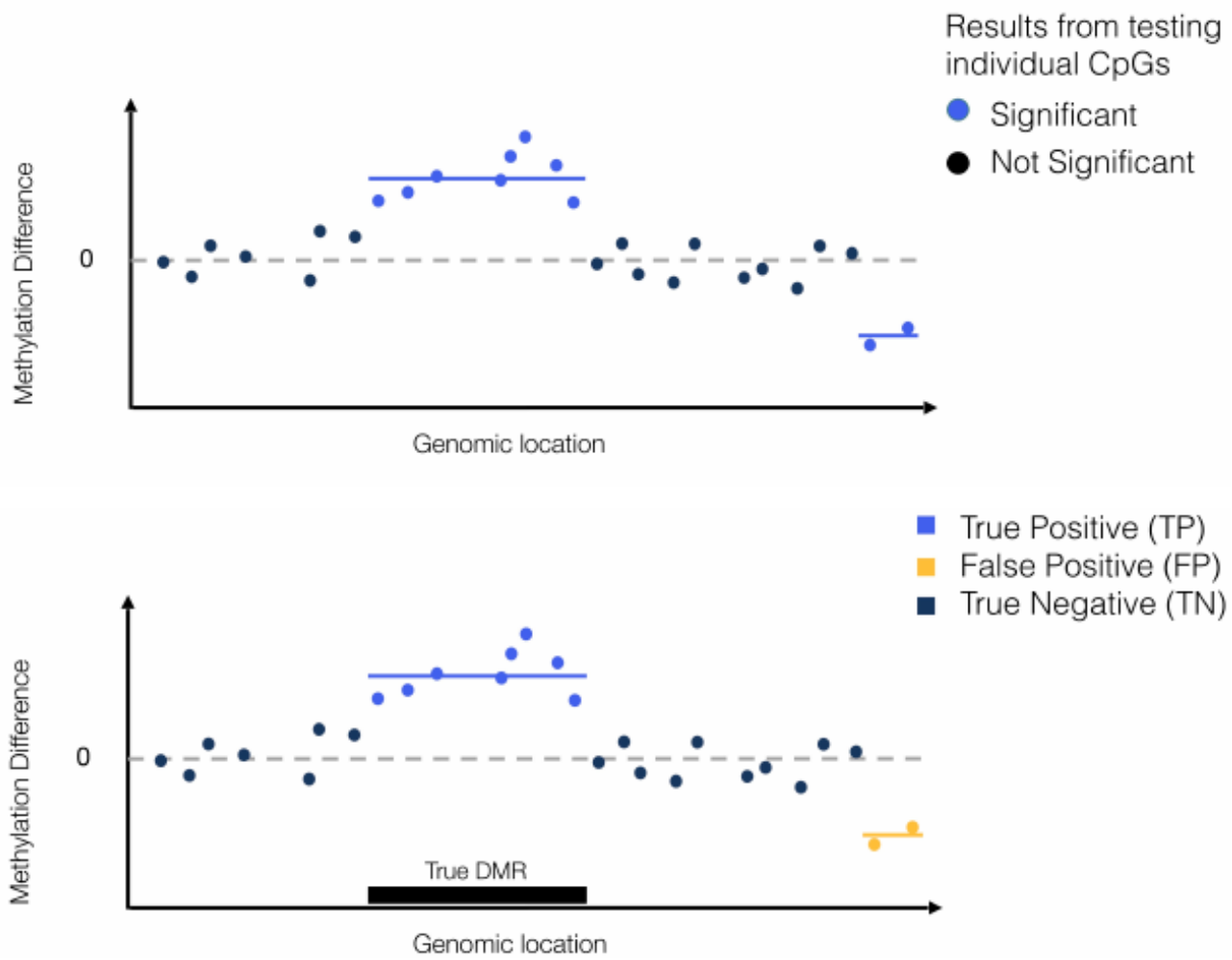
CpG 1



Adapted from Jaffe et al., 2012 (*Int J Epidemiol*)

CpG 2

FDR at the Region Level



- Grouping significant CpGs post-hoc **inflates FDR**

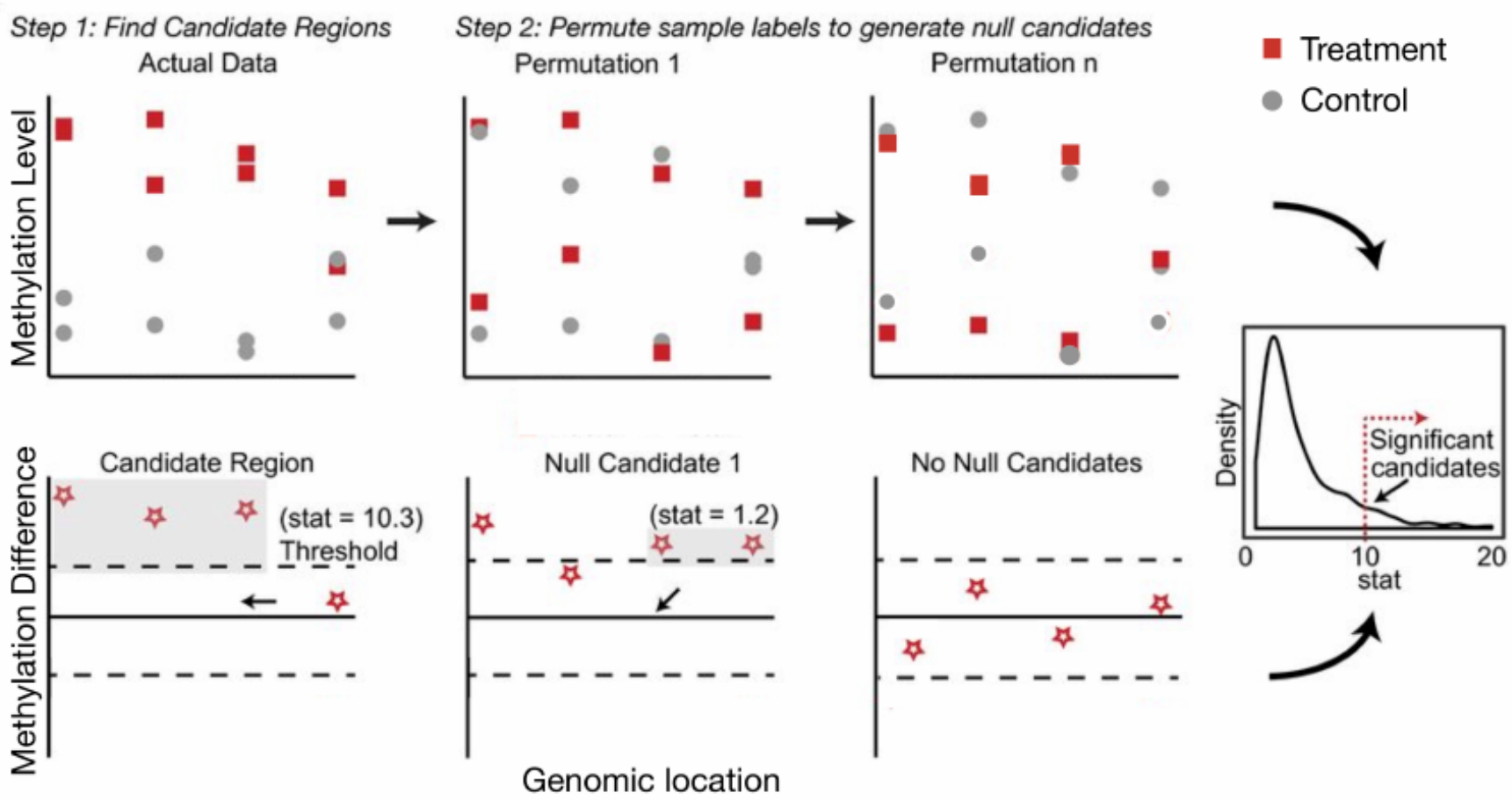
$$FDR_{CpG} = \frac{2}{10} = 0.20$$

$$FDR_{DMR} = \frac{1}{2} = 0.50$$

- Need to model region-level signal directly

`BSmooth` (Hansen et al. 2012), `DSS` (Park and Wu 2016)

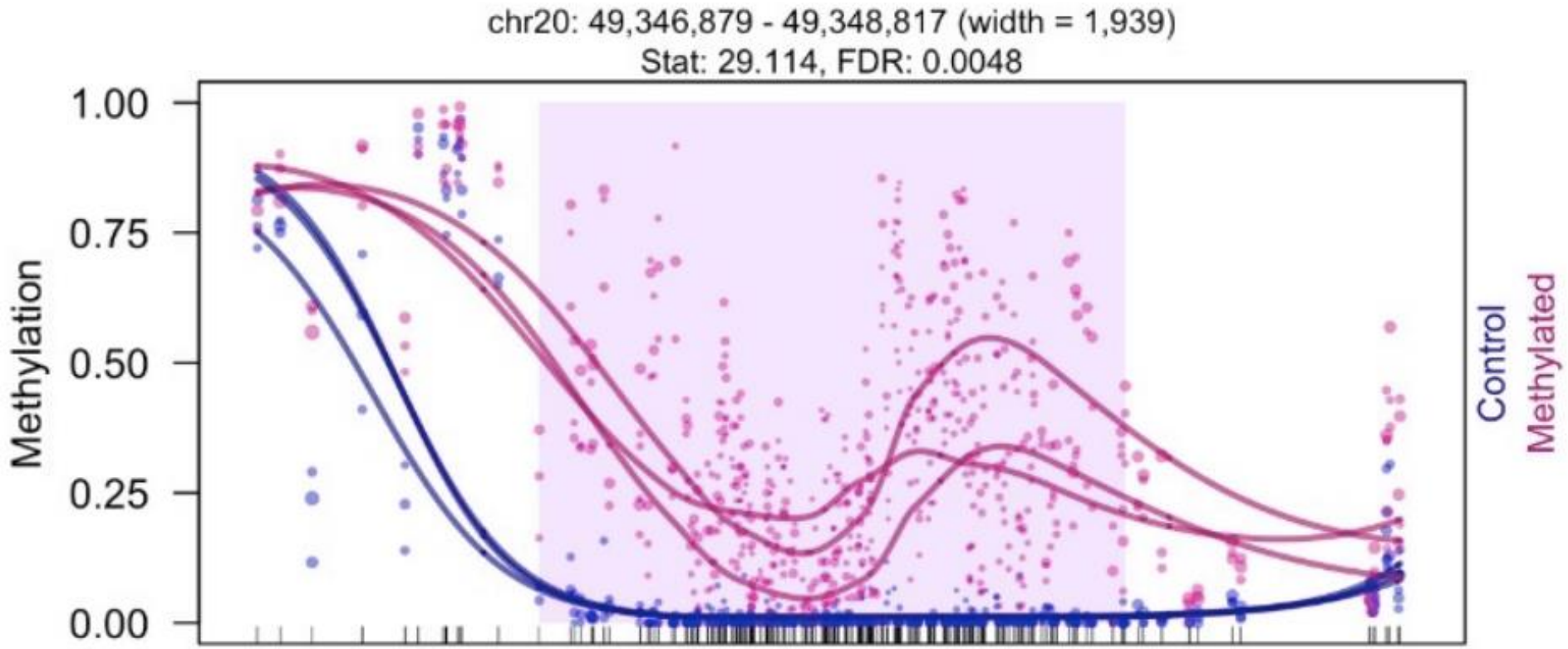
Accurate DMR Inference: `dmrseq`



- Model correlated methylation signal over region
- Permutation for accurate FDR control

(Korthauer et al. 2019)

dmrseq Output



- *De novo* regions with FDR estimate
- Ranked by effect size of methylation difference
- Can adjust for covariates

Summary: DNA Methylation Analysis

- **Data**: Illumina arrays $\rightarrow$ per-CpG β values (proportions)
- **Model**: linear regression on logit-transformed M-values
  - Same `limma` framework as gene expression microarrays
  - Additive models, interactions, paired designs all apply
- **Output**:
  - **DMC** analysis: per-probe test (limma)
  - **DMR** analysis: region-level modelling (dmrseq/bumphunter)
- **Back-transform** estimates to β scale for biological interpretation:

$$\hat{\beta} = \frac{2^{\hat{M}}}{1 + 2^{\hat{M}}}$$

Today's lecture

1. Technologies for measuring regulatory activities
2. DNA Methylation (credit: Keegan Korthauer)
3. Multiomic Data Integration with Regulatory Genomics

From DMRs to Biological Meaning

We found differentially methylated regions — now what?

Key questions:

- Which **genes** are near these regions?
- Which **biological processes** are affected?

Workflow:

$$\text{DMRs} \xrightarrow{\text{proximity}} \text{genes} \xrightarrow{\text{enrichment}} \text{pathways / GO terms}$$

This generalises to **any** set of genomic regions: ATAC peaks, ChIP peaks, GWAS loci

Standard Enrichment: A Problem

Naive approach: map each region to its nearest gene, collect gene list, run hypergeometric test

Why this fails:

- Gene-dense regions contribute many genes $\rightarrow$ **inflated hits**
- Distal enhancers (>5 kb) are silently ignored
- Background is "all genes", not "all genomic space"

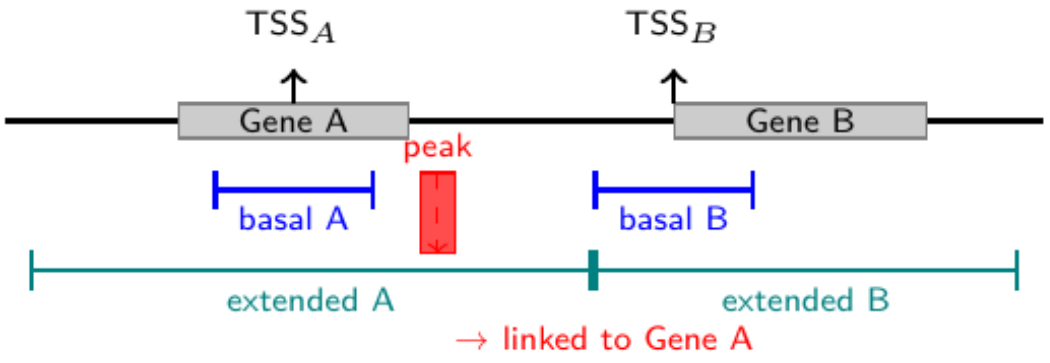
Example: a peak between a 50-gene cluster and an intergenic desert $\rightarrow$ the cluster dominates

Regions → Enrichment Directly

`GREAT` (McLean et al. 2010). Each gene gets a **regulatory domain**:

Basal domain: $\pm$ 5 kb around TSS (always kept)

Extended domain: expands up to 1 Mb, until the next gene's basal boundary



A region can be linked to **multiple** genes; background = all genomic space

GREAT: Why It Works Better

	Standard	GREAT
Input	gene list	genomic regions
Background	all genes	all genome
Long-range	ignored	up to 1 Mb
Gene density bias	yes	corrected

Useful for: ATAC peaks, ChIP peaks, DMRs, GWAS loci

GREAT Step 1: Prepare Input Regions

```r
library(rGREAT)
library(TxDb.Hsapiens.UCSC.hg19.knownGene)

## chip_peaks loaded earlier - keep standard chromosomes
std_chrs <- paste0("chr", c(1:22, "X", "Y"))
ctcf_gr  <- GRanges(as.character(seqnames(chip_peaks)),
                    IRanges(start(chip_peaks), end(chip_peaks)))
ctcf_gr  <- ctcf_gr[as.character(seqnames(ctcf_gr)) %in% std_chrs]
cat(length(ctcf_gr), "CTCF peaks ready\n")
```

```
## 446 CTCF peaks ready
```

- Regions = CTCF ChIP-seq peaks (GM12878, whole genome)
- Same syntax for DMRs, ATAC peaks, or GWAS loci

GREAT Step 2: Run Enrichment

```r
great_file <- "../data/great_ctcf_gobp.rds"
if.needed(great_file, {
    great_res <- great(ctcf_gr,
                       gene_sets  = "GO:BP",
                       tss_source = "TxDb.Hsapiens.UCSC.hg19.knownGene",
                       verbose    = FALSE)
    saveRDS(great_res, great_file)
})
great_res <- readRDS(great_file)
```

- `gene_sets`: `"GO:BP"`, `"GO:MF"`, `"KEGG"`, `"MSigDB:H"`, …
- `tss_source`: genome-specific gene model
- Returns a `GreatObject` (region–gene links + enrichment)

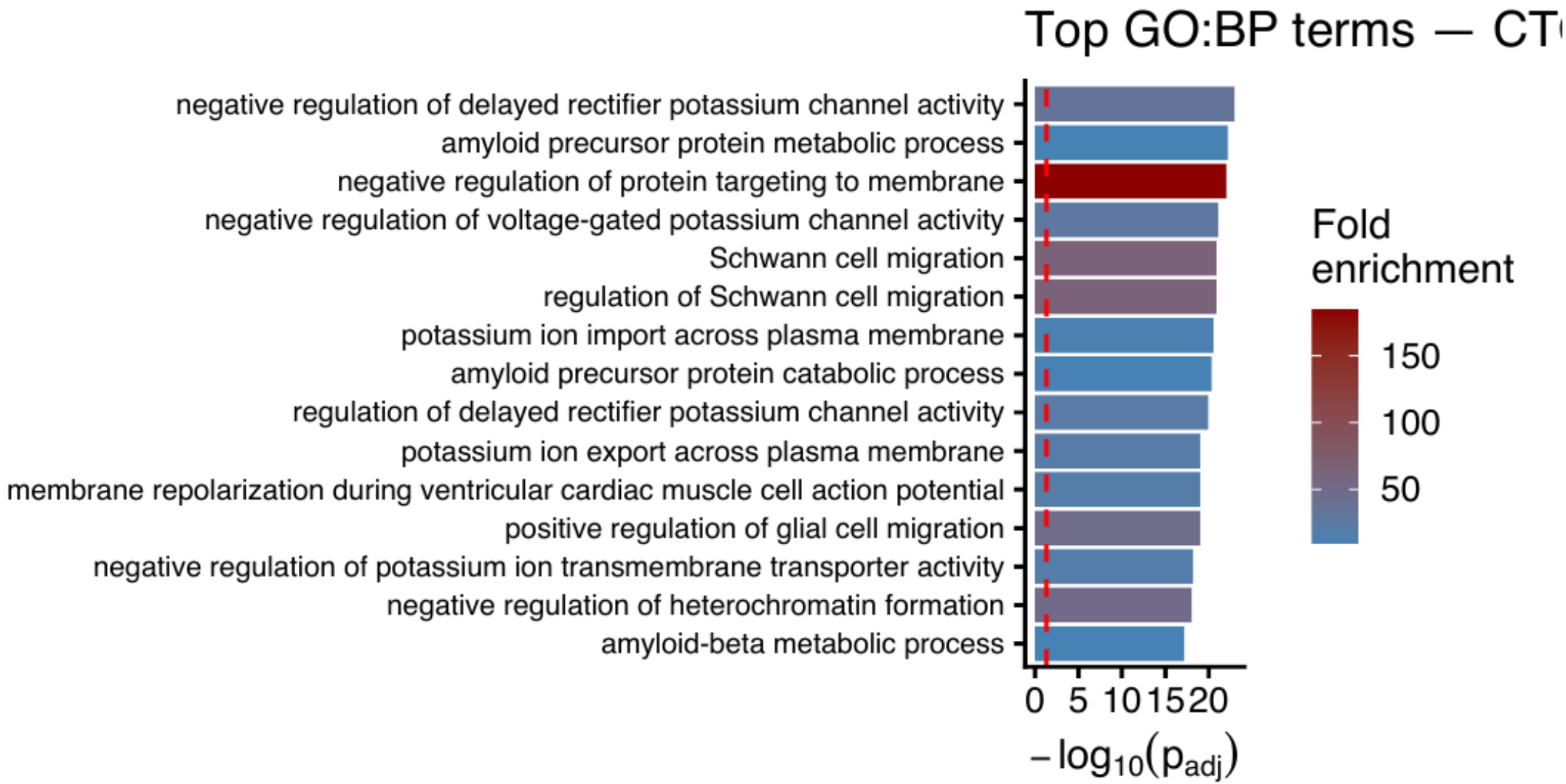
GREAT Step 3: Inspect Results

```r
tb <- getEnrichmentTable(great_res)
cols <- c("description", "observed_region_hits",
          "fold_enrichment", "p_adjust")
head(tb[order(tb$p_adjust), cols], 6)
```

```
##                                                    description
## 1 negative regulation of delayed rectifier potassium channel activity
## 2                          amyloid precursor protein metabolic process
## 3                negative regulation of protein targeting to membrane
## 4     negative regulation of voltage-gated potassium channel activity
## 5                                                Schwann cell migration
## 6                                regulation of Schwann cell migration
##   observed_region_hits fold_enrichment     p_adjust
## 1                   22        36.12004 1.143964e-23
## 2                   38        10.23917 6.198902e-23
## 3                   13       183.74047 9.691236e-23
## 4                   22        27.65168 8.614990e-22
## 5                   16        66.46148 1.283780e-21
## 6                   16        66.46148 1.283780e-21
```

- `fold_enrichment`: observed / expected overlap, `p_adjust`: BH-corrected p-value

GREAT Step 4: Visualize Top Terms



GWAS Variants Are Mostly Non-Coding

Disease GWAS identifies thousands of associated variants

The puzzle: >90% of hits fall outside protein-coding exons

Key questions:

- Where do these variants **act**?
- Which **cell types** are disease-relevant?

Strategy: overlap GWAS loci with cell-type-specific epigenetic annotations

GWAS → Epigenetic States: The Approach

1. Collect GWAS summary statistics (SNPs with $P < 5 \times 10^{-8}$)
2. Download cell-type-specific chromatin annotations (e.g., Roadmap Epigenomics: 127 cell/tissue types)
3. For each cell type: test whether GWAS SNPs are **enriched** in active enhancers
4. The most enriched cell type = likely disease-relevant tissue

Statistical test: Fisher's exact test (or LD-score regression)

GWAS Step 1: Fetch RA Loci from GWAS Catalog

```r
library(gwasrapidd)
ra_file <- "../data/ra_gwas_hits.rds"
if (!file.exists(ra_file)) {
    assoc    <- get_associations(study_id = "GCST90132222")
    variants <- get_variants(study_id    = "GCST90132222")
    ra <- merge(assoc@risk_alleles[, c("association_id","variant_id")],
                variants@variants[, c("variant_id","chromosome_name",
                                      "chromosome_position")],
                by = "variant_id")
    ra$CHR <- paste0("chr", ra$chromosome_name)
    saveRDS(ra, ra_file)
}
ra <- readRDS(ra_file)
cat(nrow(ra), "genome-wide significant RA loci (GCST90132222)\n")
```

```
## 144 genome-wide significant RA loci (GCST90132222)
```

- Study: 22,350 European ancestry samples (Okada et al. 2014)
- Coordinates: hg19 / GRCh37

GWAS Step 2: Build Locus Windows

```r
## Expand each index SNP ±50 kb (covers typical LD block)
sig_gr <- GRanges(ra$CHR,
                  IRanges(ra$chromosome_position - 50000,
                          ra$chromosome_position + 50000))
sig_gr <- sig_gr[start(sig_gr) > 0]

## Chromosome-proportional random background (hg19)
set.seed(42)
chr_sizes <- seqlengths(BSgenome.Hsapiens.UCSC.hg19::
                          BSgenome.Hsapiens.UCSC.hg19)[paste0("chr",1:22)]
chr_s <- sample(names(chr_sizes), 5e4, replace = TRUE,
                prob = chr_sizes / sum(chr_sizes))
pos_s <- mapply(function(ch) sample(chr_sizes[ch] - 1e5, 1) + 50000, chr_s)
bg_gr <- GRanges(chr_s, IRanges(pos_s - 50000, pos_s + 50000))
cat(length(sig_gr), "locus windows,", length(bg_gr), "background windows\n")
```

```
## 144 locus windows, 50000 background windows
```

GWAS Step 3: Load Roadmap H3K27ac Enhancers

```r
enh_dir    <- "../data/roadmap/"
enh_labels <- c(CD4_Naive  = "AH43615", CD4_Memory = "AH43684",
                CD19_Bcell = "AH44900", CD14_Mono  = "AH44890",
                Adult_Liver= "AH43458", Brain      = "AH43501")

## Load from locally cached RDS files (downloaded once via AnnotationHub)
enhancers <- lapply(names(enh_labels), function(ct)
    readRDS(paste0(enh_dir, ct, "_H3K27ac.rds")))
names(enhancers) <- names(enh_labels)
immune_cells <- c("CD4_Naive","CD4_Memory","CD19_Bcell","CD14_Mono")
sapply(enhancers, length)   # peaks per cell type
```

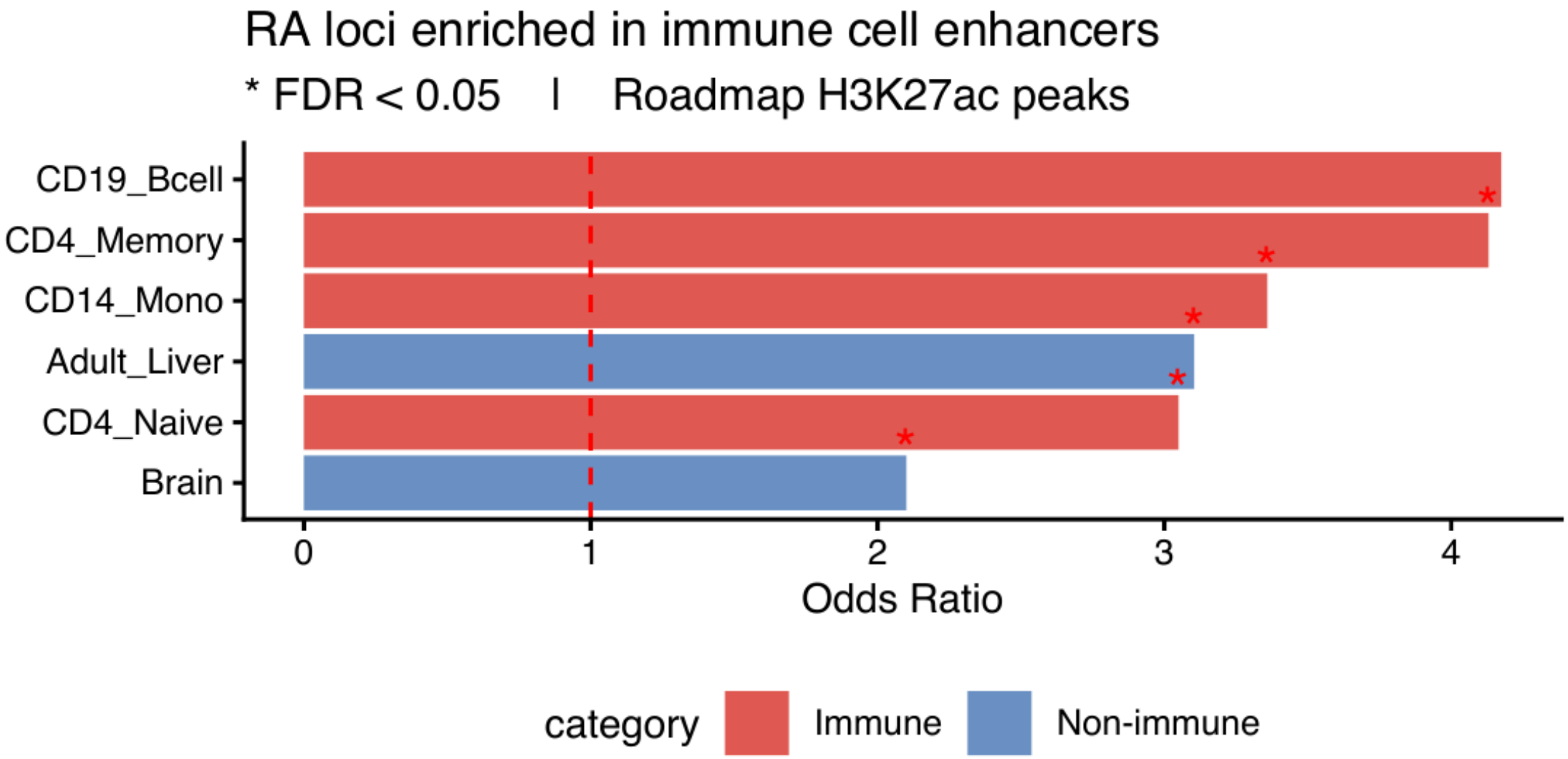
```
##   CD4_Naive  CD4_Memory  CD19_Bcell   CD14_Mono Adult_Liver       Brain
##       66232       92684       65749       61058      121020      188662
```

- H3K27ac = active enhancer / promoter mark
- Roadmap Epigenomics: 127 reference human epigenomes (Kundaje et al. 2015)

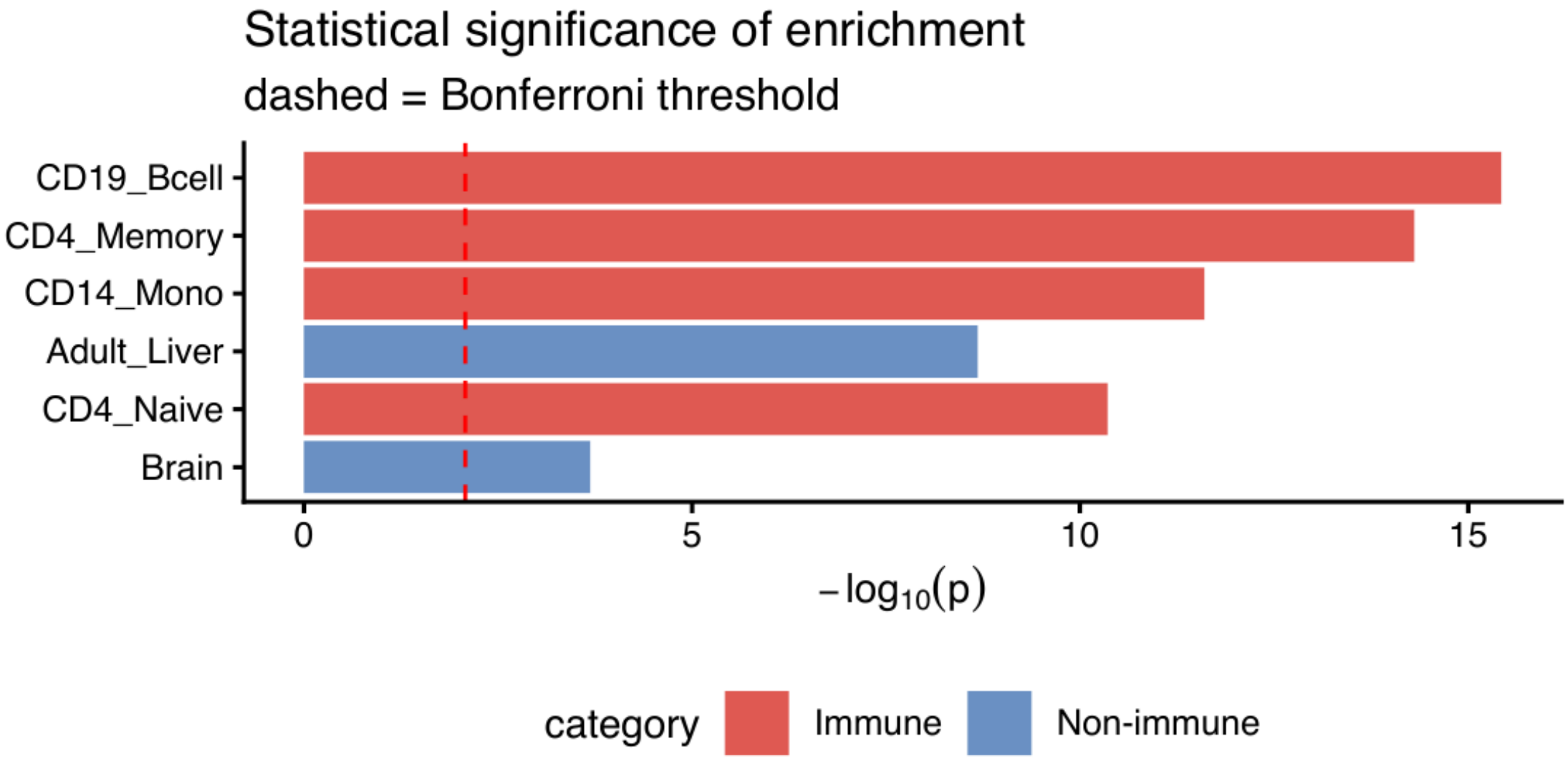
GWAS Step 4: Fisher's Exact Test per Cell Type

```r
enr_file <- "../data/gwas_ra_enrichment.rds"
if.needed(enr_file, {
    enr <- do.call(rbind, lapply(names(enhancers), function(ct) {
        enh <- enhancers[[ct]]
        a <- sum(countOverlaps(sig_gr, enh) > 0)
        b <- length(sig_gr) - a
        c <- sum(countOverlaps(bg_gr,  enh) > 0)
        d <- length(bg_gr)  - c
        ft <- fisher.test(matrix(c(a, b, c, d), 2))
        data.frame(cell_type = ct, OR = ft$estimate,
                   pval = ft$p.value, frac = a / length(sig_gr))
    }))
    enr$fdr      <- p.adjust(enr$pval, "BH")
    enr$category <- ifelse(enr$cell_type %in% immune_cells, "Immune", "Non-immune")
    saveRDS(enr, enr_file)
})
enr <- readRDS(enr_file)
enr$cell_type <- factor(enr$cell_type, levels = enr$cell_type[order(enr$OR)])
```

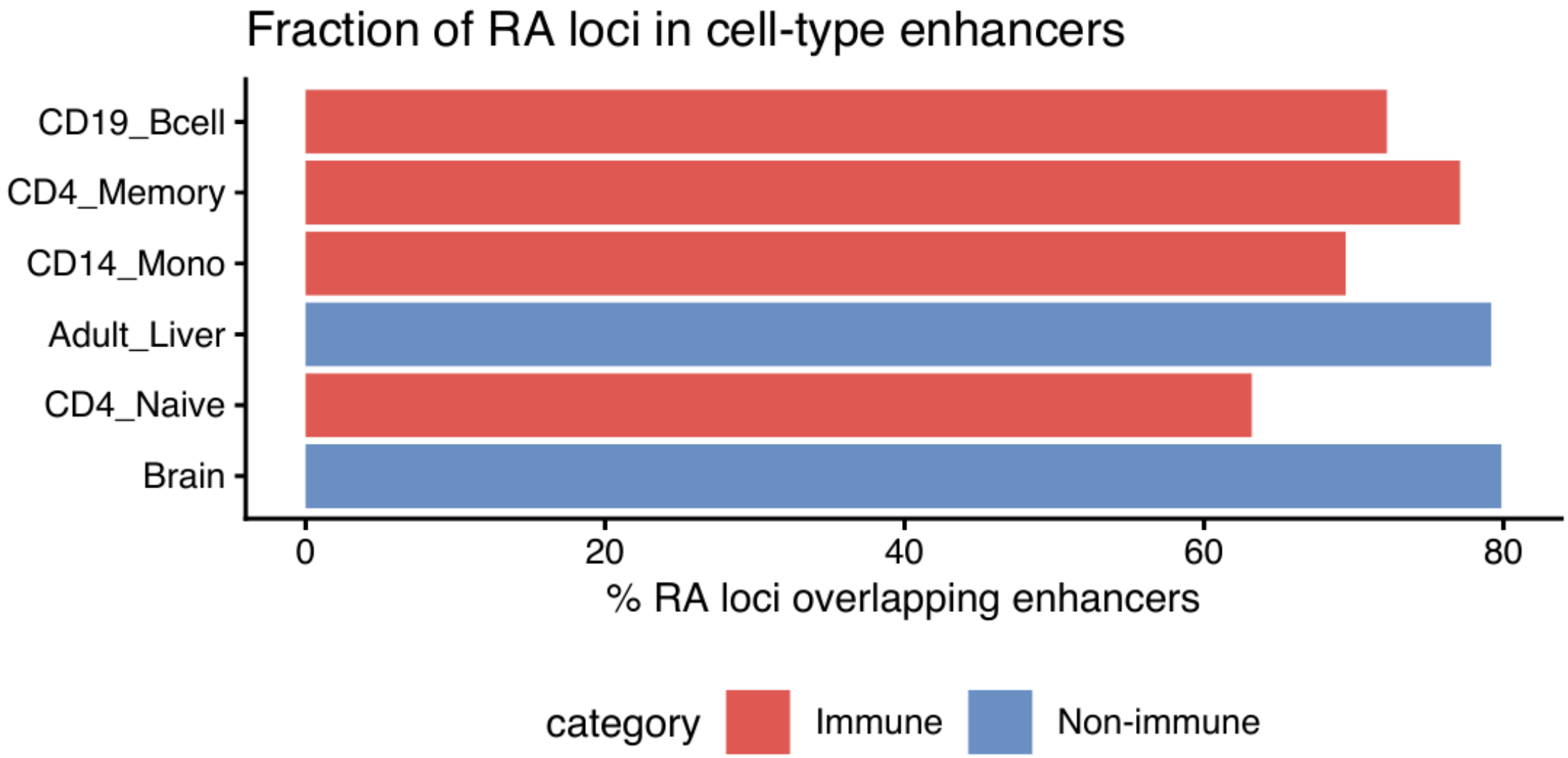
GWAS Enrichment: Odds Ratio



GWAS Enrichment: Statistical Significance



GWAS Enrichment: Overlap Fraction



Limitation: Fisher's Test Ignores LD

The problem: nearby SNPs are correlated through **linkage disequilibrium (LD)**

Two significant SNPs 20 kb apart may represent one causal signal, not two

$\Rightarrow$ overlapping the same enhancer window gets double-counted

Consequence: Fisher's test inflates enrichment in gene-dense, LD-rich regions

Better approaches:

- **Stratified LD score regression** (S-LDSC) — partitions heritability by annotation, models LD (Finucane et al. 2015)
- **GARFIELD** — permutation matching on MAF, distance to TSS, LD
- **GoShifter** — circular permutation within LD blocks

Real-World GWAS–Epigenome Findings (examples)

Rheumatoid arthritis $\rightarrow$ CD4$^+$ T cells	*adaptive immunity*
IBD $\rightarrow$ gut immune cells	*mucosal inflammation*
Type 2 diabetes $\rightarrow$ pancreatic islets	*β-cell dysfunction*
Schizophrenia $\rightarrow$ neurons	*brain circuit defects*
Height $\rightarrow$ chondrocytes	*cartilage growth plates*

(Kundaje et al. 2015)

Why GWAS–Epigenome Integration Matters

Identify disease-relevant cell types

- Guides which cells to study experimentally
- 90% of hits are non-coding — context is everything

Mechanistic insight

- Variant disrupts an enhancer $\rightarrow$ changes TF binding $\rightarrow$ alters gene expression

Therapeutic direction

- Target the right cell type, not just the right gene

Links to scRNA-seq

- Fine-map to specific cell states within a tissue

Roadmap of STAT/BIOF/GSAT540 2026

data generation followed by *methodology*

3. Formulating biological questions
4. RNA-seq data
5. Experimental design
6. Single-cell genomics
7. Matrix factorization
10. GWAS
11. **Regulatory genomics** (Today)
12. Regression (for reg. genomics)
13. Biological networks
14. Clustering

Reference

Buenrostro, Jason D, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. 2013. "Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position." *Nat. Methods* 10 (12): 1213–18.

Du, Pan, Xiao Zhang, Chiang-Ching Huang, et al. 2010. "Comparison of Beta-Value and m-Value Methods for Quantifying Methylation Levels by Microarray Analysis." *BMC Bioinformatics* 11: 587.

Finucane, Hilary K, Brendan Bulik-Sullivan, Alexander Gusev, et al. 2015. "Partitioning heritability by functional annotation using genome-wide association summary statistics." *Nature Genetics* 47: 1228–35.

Reference

Furey, Terrence S. 2012. "ChIP-Seq and Beyond: New and Improved Methodologies to Detect and Characterize Protein-DNA Interactions." *Nat. Rev. Genet.* 13 (12): 840–52.

Hansen, Kasper D, Benjamin Langmead, and Rafael A Irizarry. 2012. "BSmooth: From Whole Genome Bisulfite Sequencing Reads to Differentially Methylated Regions." *Genome Biology* 13 (10): R83.

Korthauer, Keegan, Sutirtha Chakraborty, Yuval Benjamini, and Rafael A Irizarry. 2019. "Detection and Accurate False Discovery Rate Control of Differentially Methylated Regions from Whole Genome Bisulfite Sequencing." *Biostatistics* 20 (3): 367–83.

Reference

Kundaje, Anshul, Wouter Meuleman, Jason Ernst, et al. 2015. "Integrative analysis of 111 reference human epigenomes." *Nature* 518 (7539): 317–30.

Laird, Peter W. 2010. "Principles and Challenges of Genome-Wide DNA Methylation Analysis." *Nature Reviews Genetics* 11 (3): 191–203.

McLean, Cory Y, Dave Bristor, Michael Hiller, et al. 2010. "GREAT Improves Functional Interpretation of Cis-Regulatory Regions." *Nat. Biotechnol.* 28 (5): 495–501.

Okada, Yukinori, Di Wu, Gosia Trynka, et al. 2014. "Genetics of rheumatoid arthritis contributes to biology and drug discovery." *Nature* 506: 376–81.

Reference

Park, Peter J. 2009. "ChIP-Seq: Advantages and Challenges of a Maturing Technology." *Nat. Rev. Genet.* 10 (10): 669–80.

Park, Yun, and Hao Wu. 2016. "DSS: A R/Bioconductor Package for Analyzing DMS and DMR from Bisulfite Sequencing Data." *Bioinformatics* 32 (16): 2446–47.

Zhang, Yue, Jovana Maksimovic, Gaetano Naselli, et al. 2013. "Genome-Wide DNA Methylation Analysis Identifies Hypomethylated Genes Regulated by FOXP3 in Human Regulatory T Cells." *Epigenetics & Chromatin* 6 (1): 34.