A Brief Overview of Epigenomics

BIOTRAC Lecture

Sean, Soonweng, Cho, PhD

Kennedy Krieger Institute

and

Johns Hopkins University School of Medicine

 sean.cho@jhmi.edu

 github.com/sean-cho

Outline

What is epigenomics?

Why epigenomics?

How

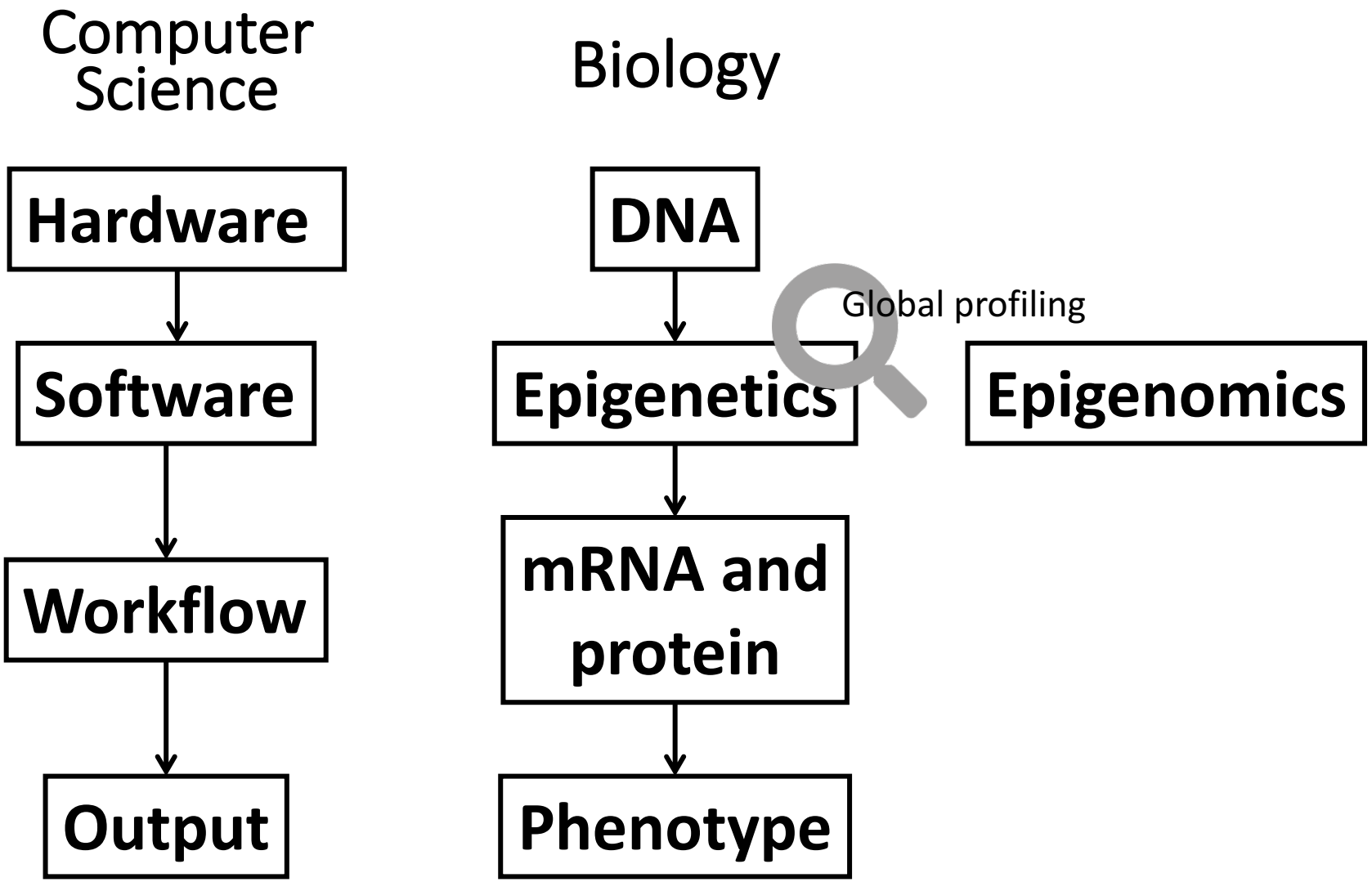
- DNA methylation
- Histone modifications
- Chromatin features
  - Accessibility
  - HiC
- Integrative epigenomics
  - Imputation
  - Chromatin state
- Integrative analyses
- Study Design
  - Limitations and considerations
  - Balancing and Blocking

The future of epigenomics

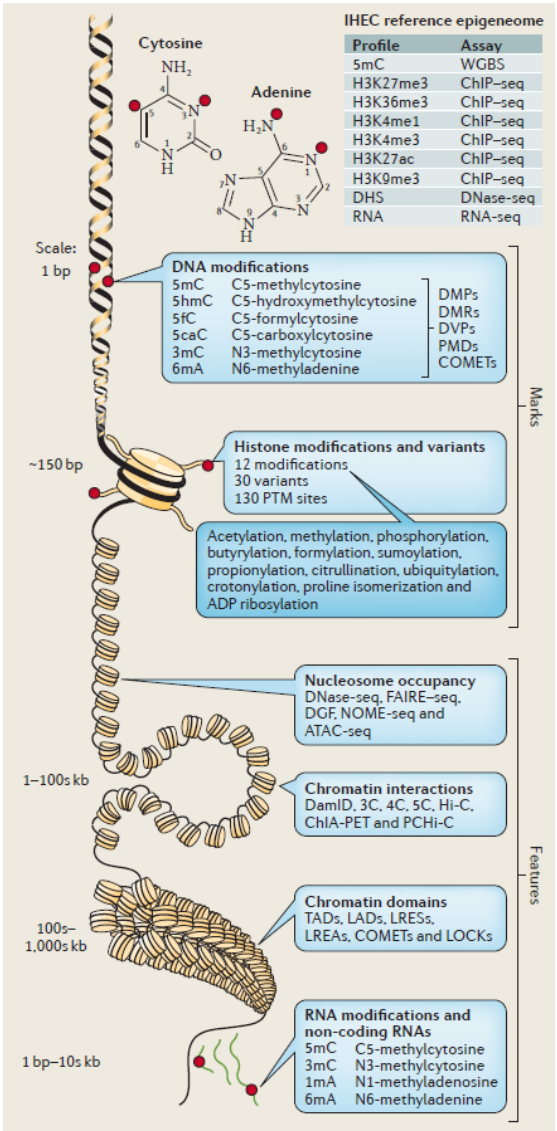
- Nanopore/MinION sequencing
- Single cell epigenomics

What is epigenetics?

Epigenetics



Epigenetics ➡ epigenomics



WGBS

DNA methylation

H3K4me3

Histone modifications

DNase

Chromatin features

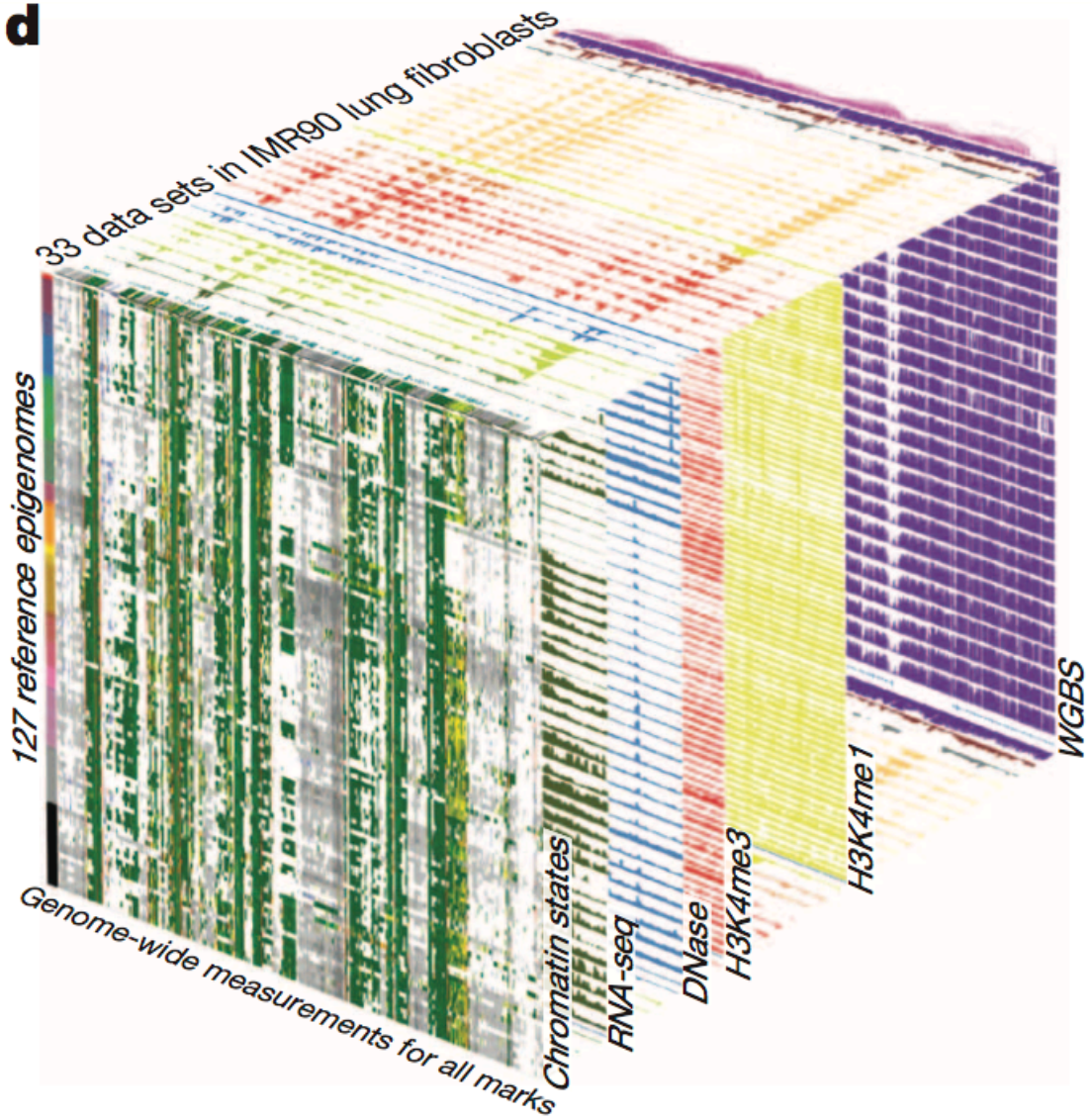
RNA-seq

Non-coding RNA

Stricker, S. H., Köferle, A., & Beck, S. (2017). From profiles to function in epigenomics. *Nature Reviews Genetics*, *18*(1), 51.

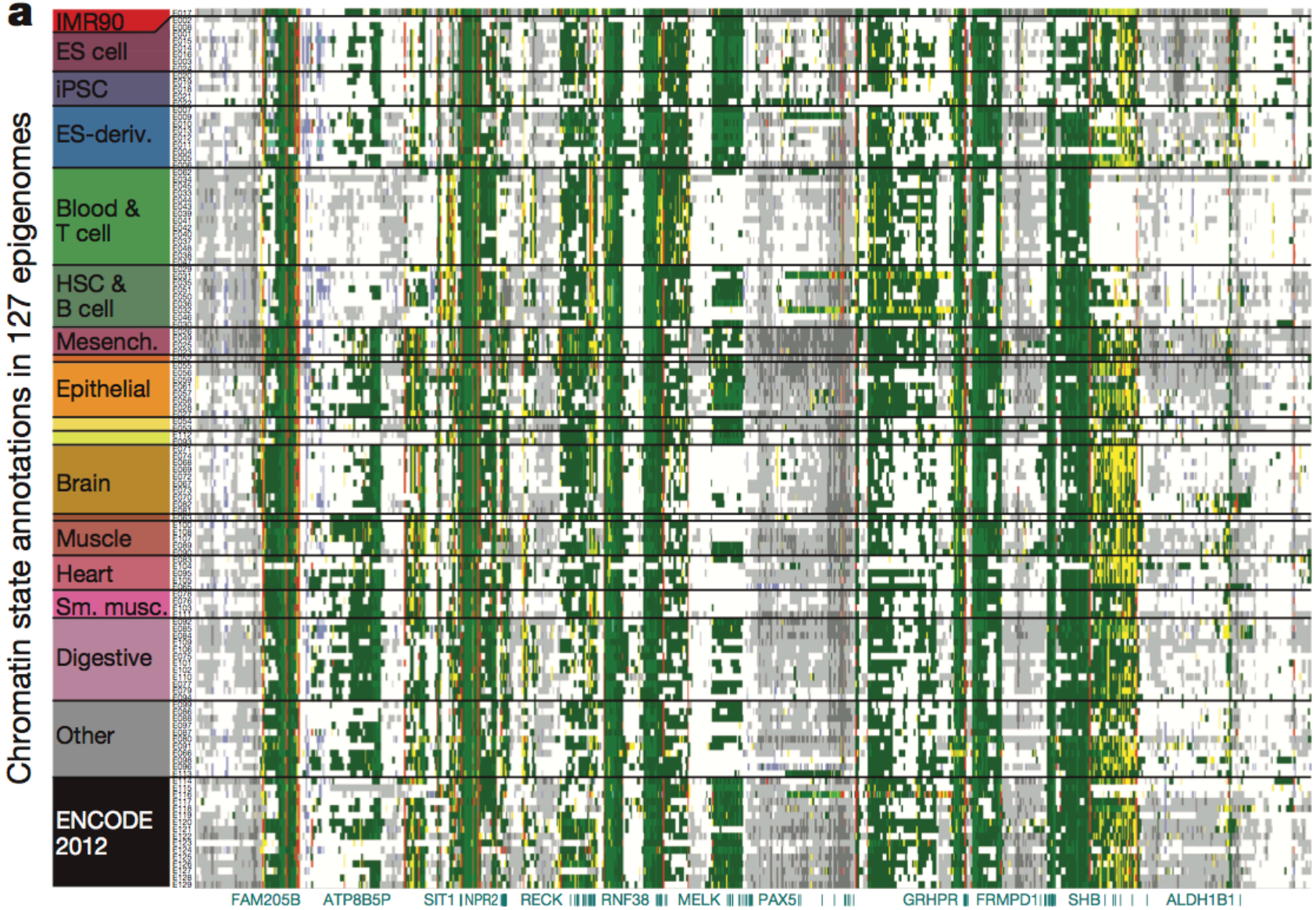
Roadmap, E. C., ... & Ziller, M. J. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, *518*, 317-330

Integrated into cell specific epigenomes that govern the function, state, and fate of cells



Roadmap, E. C., ... & Ziller, M. J. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, *518*, 317-330

Roadmap Epigenomics Consortium

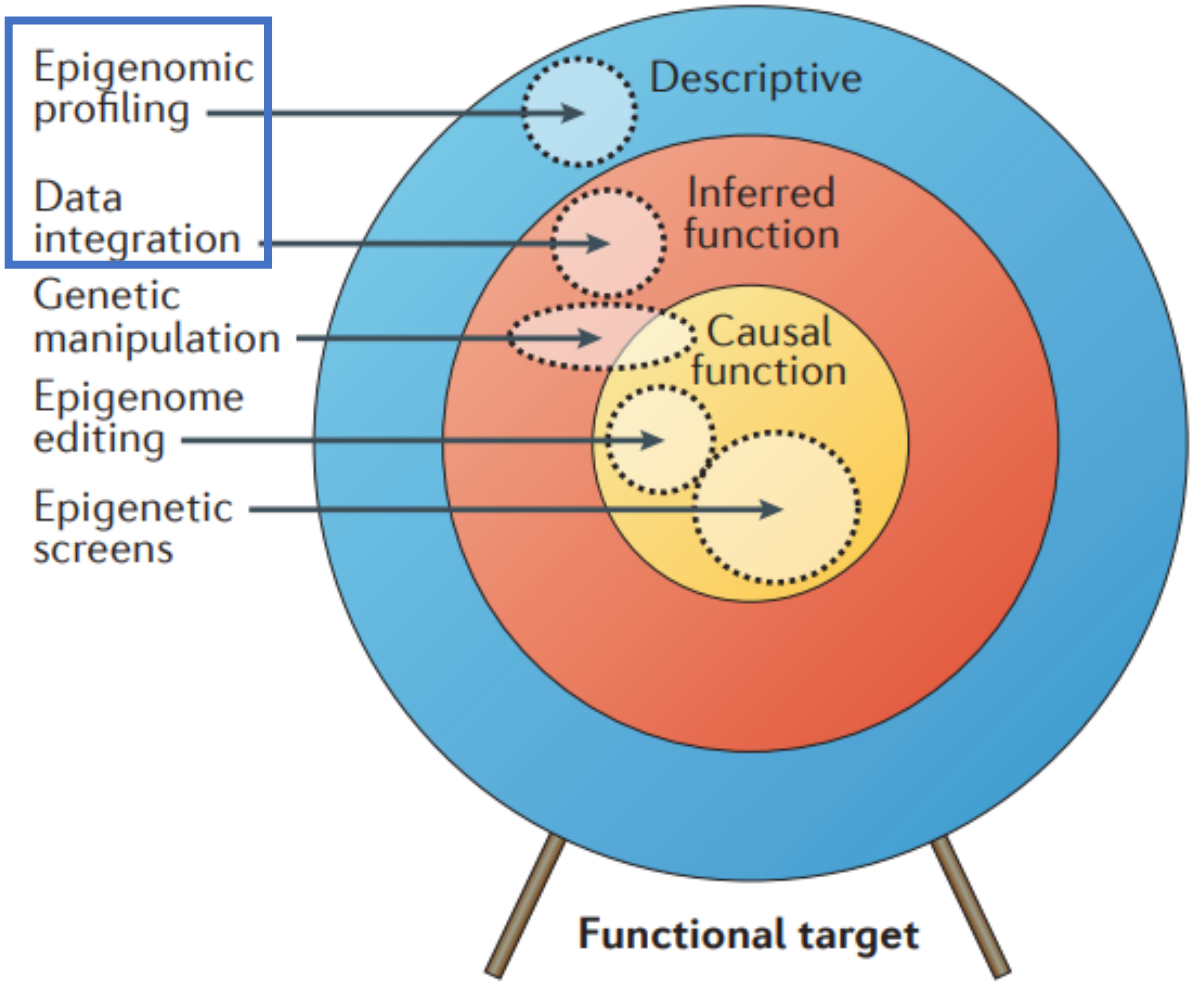


Roadmap, E. C., ... & Ziller, M. J. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, *518*, 317-330

Why epigenomics?

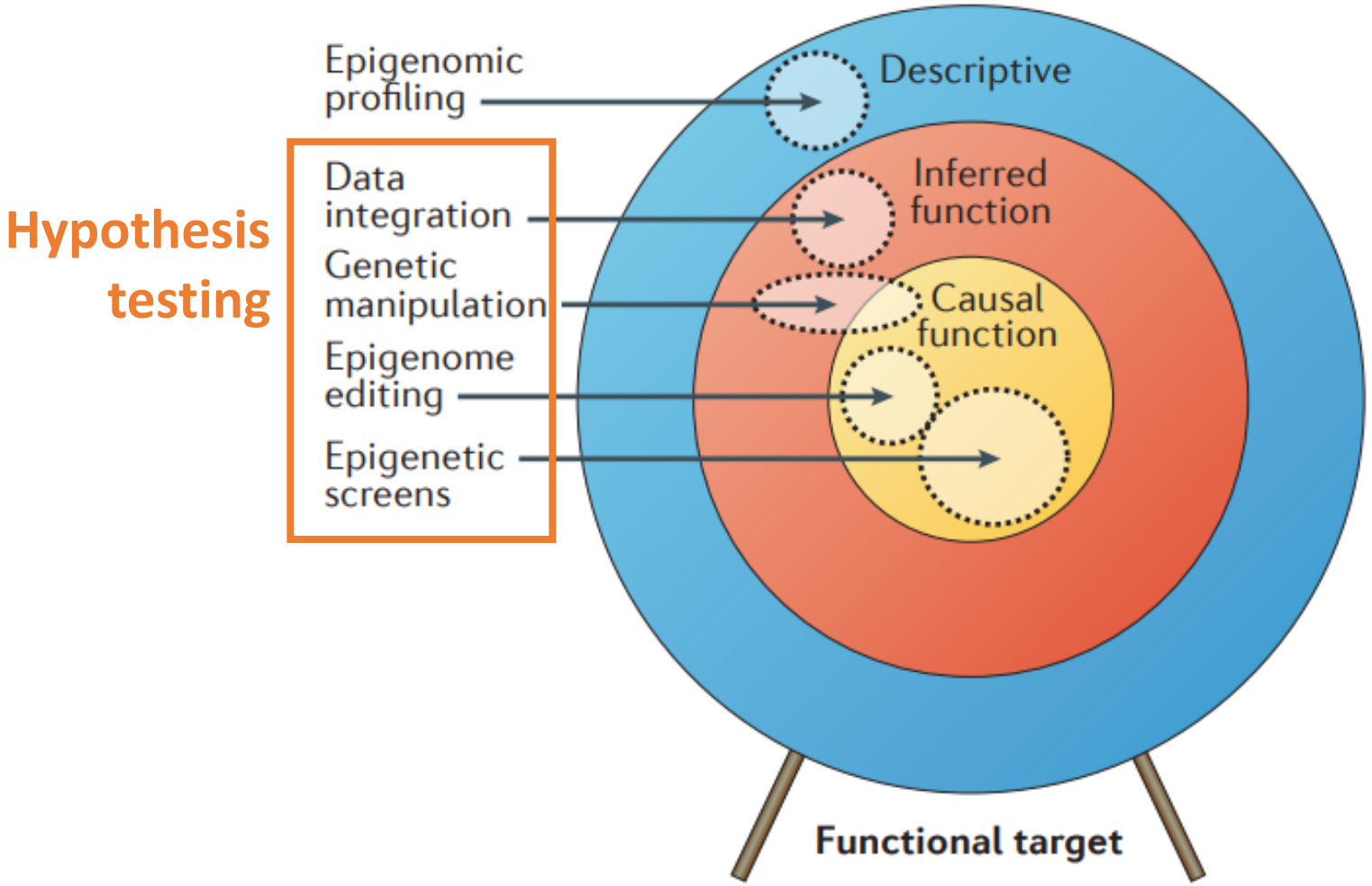
Why epigenomics?

Hypothesis generating



Stricker, S. H., Köferle, A., & Beck, S. (2017). From profiles to function in epigenomics. *Nature Reviews Genetics*, *18*(1), 51.

Why epigenomics?



Stricker, S. H., Köferle, A., & Beck, S. (2017). From profiles to function in epigenomics. *Nature Reviews Genetics*, *18*(1), 51.

Case 01:
Identification of differentially methylated loci and biologically relevant clusters in breast cancer

EMBO
Molecular Medicine

Research Article
Epigenetic portraits of human breast cancers

DNA methylation profiling reveals a predominant immune component in breast cancers

*Sarah Dedeurwaerder[1†], Christine Desmedt[2†], Emilie Calonne[1], Sandeep K. Singhal[2], Benjamin Haibe-Kains[2,3], Matthieu Defrance[1], Stefan Michiels[2], Michael Volkmar[1], Rachel Deplus[1], Judith Luciani[1], Françoise Lallemand[2], Denis Larsimont[4], Jérôme Toussaint[2], Sandy Haussy[2], Françoise Rothé[2], Ghizlane Rouas[2], Otto Metzger[2], Samira Majjaj[2], Kamal Saini[2], Pascale Putmans[1], Gérald Hames[5], Nicolas van Baren[6], Pierre G. Coulie[5], Martine Piccart[7], Christos Sotiriou[2**,†], François Fuks[1*,†]*

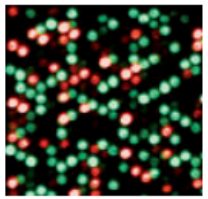
Overview

MAIN SET OF PATIENTS:
123 breast tissues

Discovery cohort

VALIDATION SET OF PATIENTS:
125 breast tissues

Independent validation cohort



14, 475 genes
(27,578 CpGs)
investigated per sample

Important, especially for classification and biomarkers studies to validate that discovery is applicable to other general datasets.

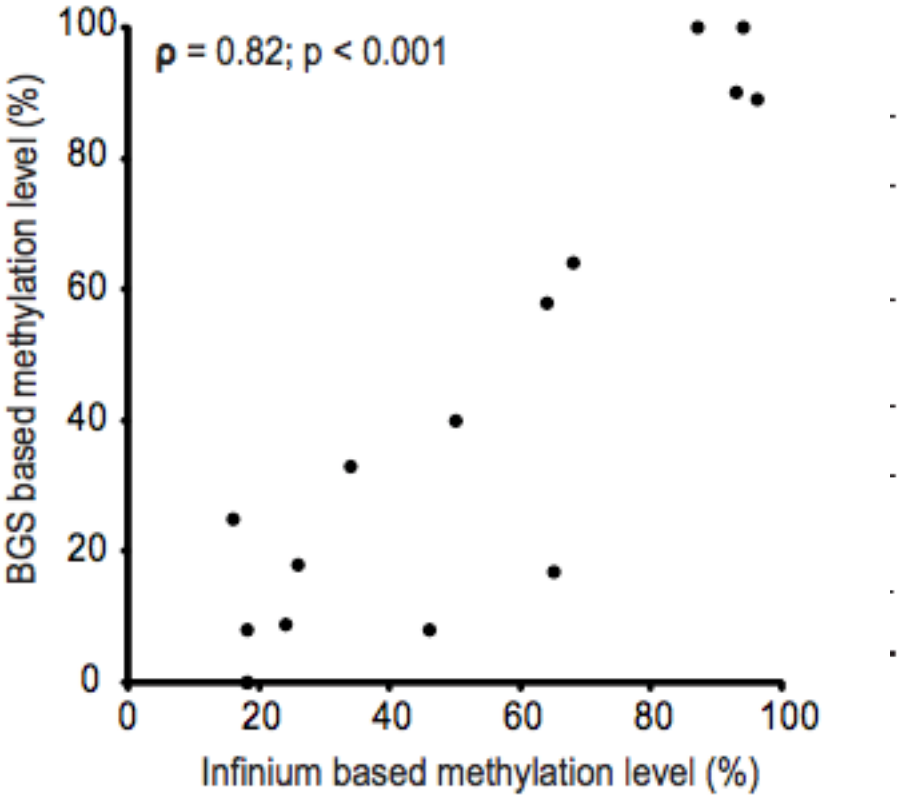
Identification of differentially methylated loci

Gene	Illumina ID	Infinium methylation (frequency, %)	Reported methylation data (frequency, %; technique)*	Correlation Infinium vs. reported methylation data
RASSF1A	cg00777121	71	70; MSP / 56; MSP / 56; MPS	++
	cg08047457	72	65; MSP	++
	cg21554552	70	65; MSP	++
CCND2	cg25425078	9	46; MSP / 28; MSP / 55; MSP	+
APC	cg16970232	39	45; MSP / 28; MSP / 39; MSP / 49; MSP	++
	cg20311501	35	45; MSP / 28; MSP / 39; MSP / 49; MSP	++
RARß2	cg27486427	12	17; PS / 0; PS	++
	cg26124016	4	23; MSP	+
CDH13	cg08747377	17	33; MSP	++
SDHB	cg24305835	0	0; MS-HRM	++
	cg03861428	0	0; MS-HRM	++
FH	cg06806184	0	0; MS-HRM	++

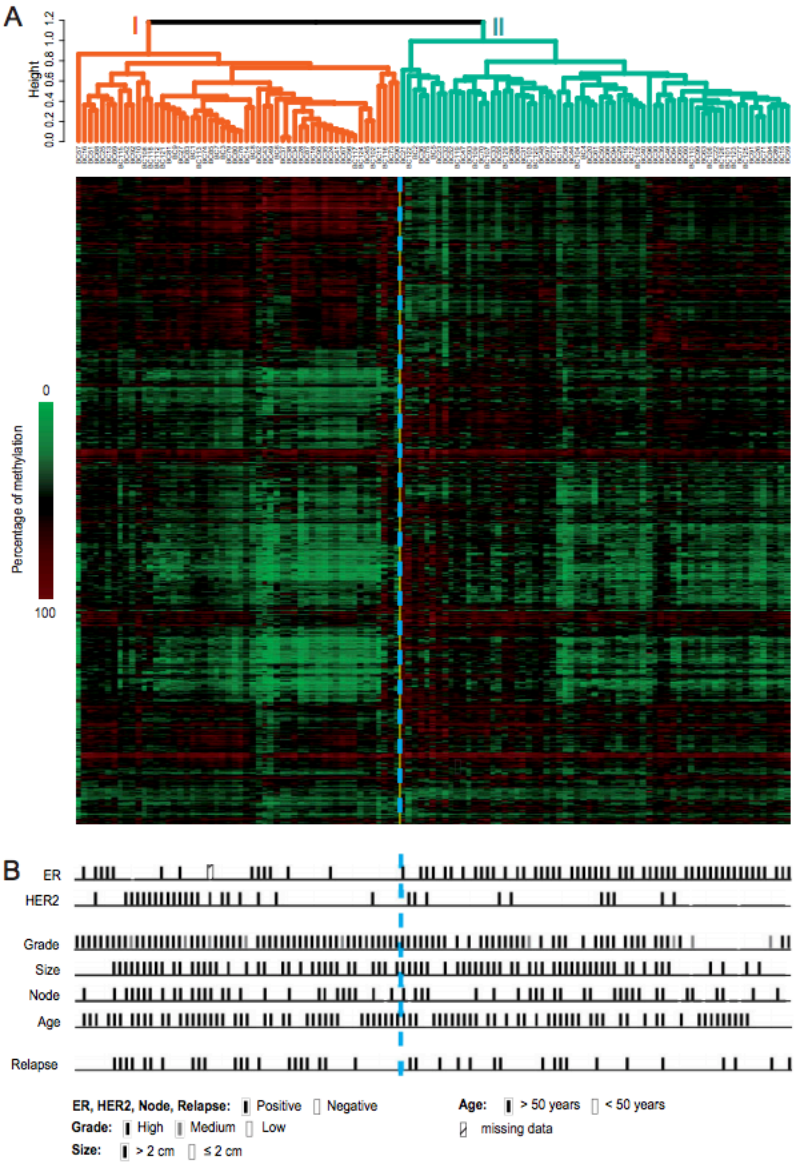
Identification of differentially methylated loci

Gene	Illumina ID	Infinium methylation (frequency, %)	Reported methylation data (frequency, %; technique)*
RASSF1A	cg00777121	71	70; MS
	cg08047457	72	65; MS
	cg21554552	70	65; MS
CCND2	cg25425078	9	46; MS
APC	cg16970232	39	45; MS
	cg20311501	35	45; MS
RARß2	cg27486427	12	17; PS
	cg26124016	4	23; MS
CDH13	cg08747377	17	33; MS
SDHB	cg24305835	0	0; MS-
	cg03861428	0	0; MS-
FH	cg06806184	0	0; MS-

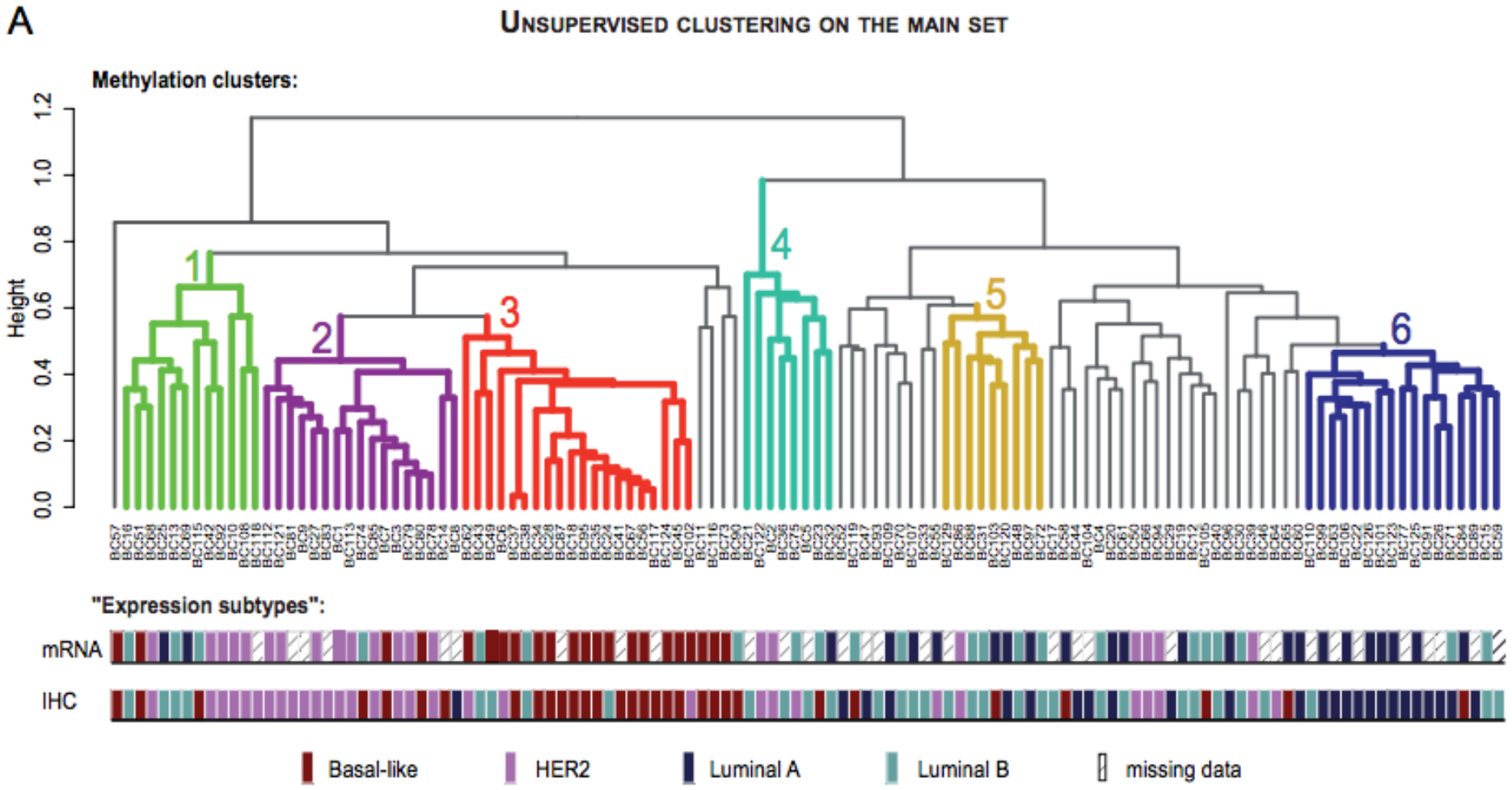
Correlation Infinium vs. reported methylation data



Unsupervised clustering identifies primary clusters by ER status

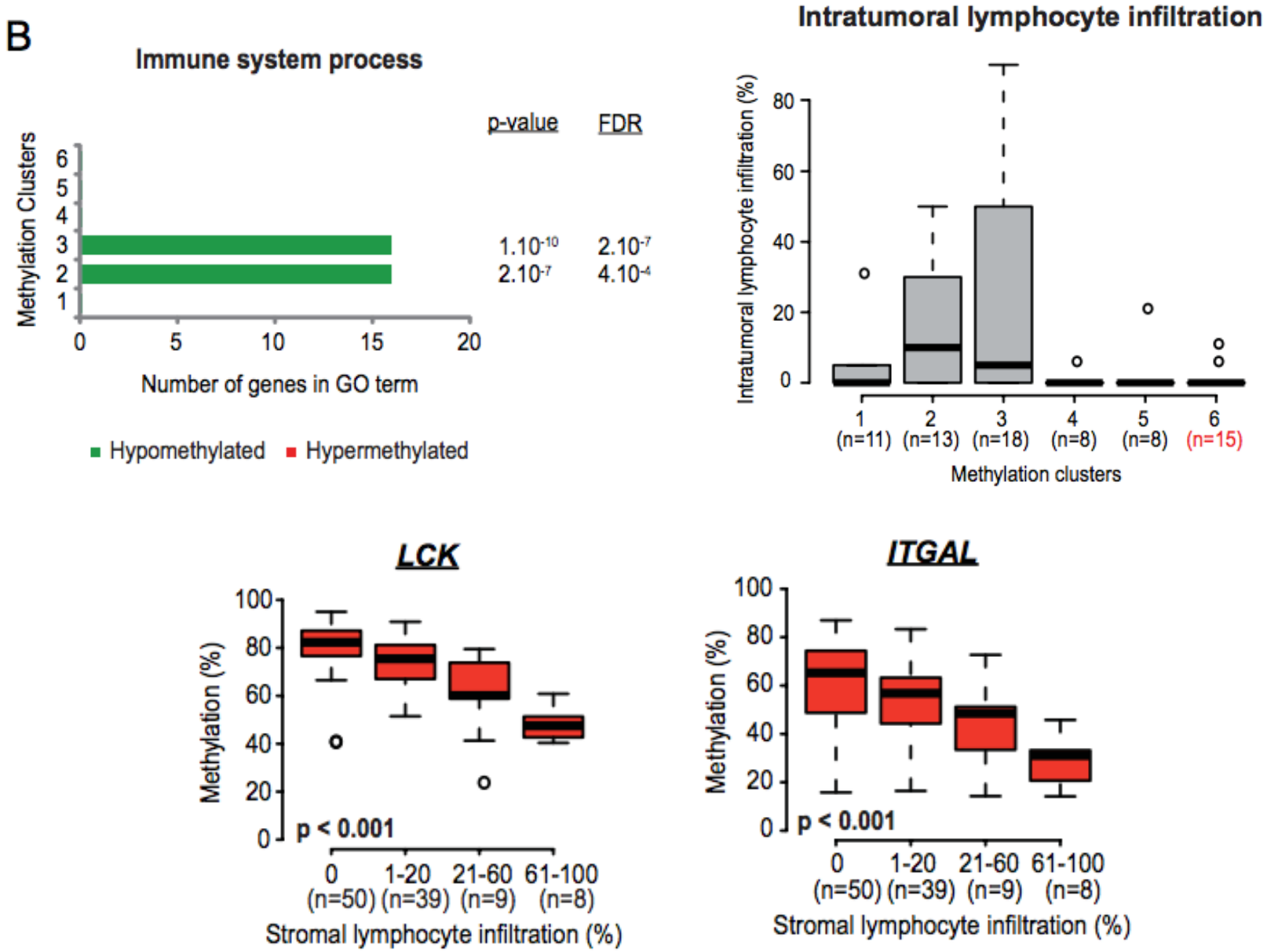


Six stable, biologically relevant clusters



Validated the presence of these clusters in an independent data set

Identification of immune-related clusters

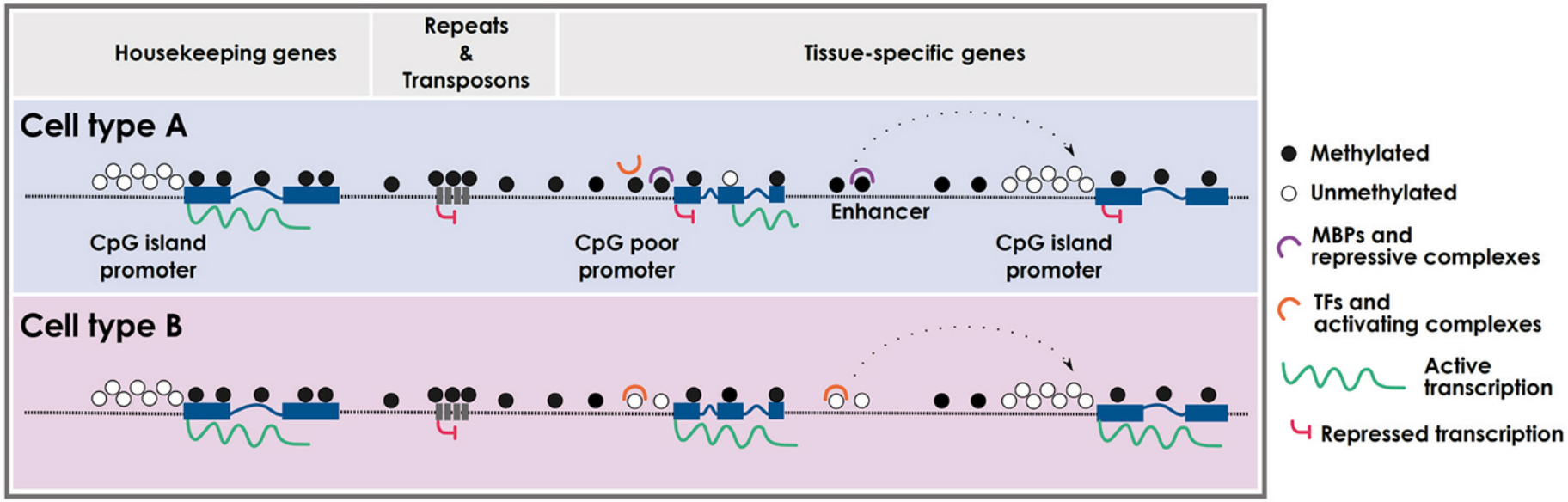


How?

Epigenomics methods

DNA Methylation

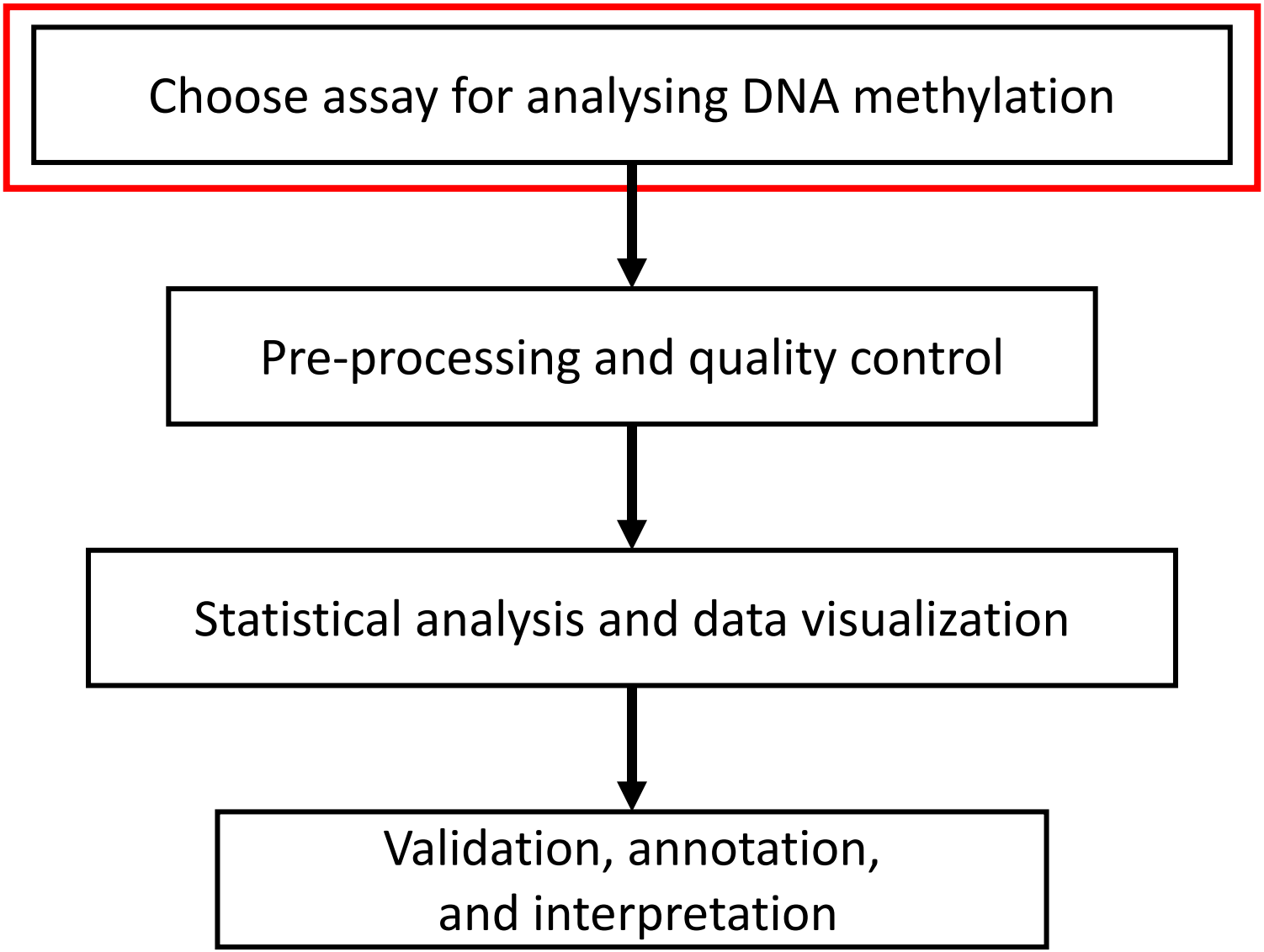
“Typical” 5mC DNA methylation in mammalian genome



- CpG island promoter and enhancer methylation is inversely proportional to gene expression
- Gene body methylation is correlated with gene expression of highly transcribed genes
- CpG methylation is typical in repeats and transposons
- Changes in these patterns, both globally and loci-specific, have been observed in disease progression

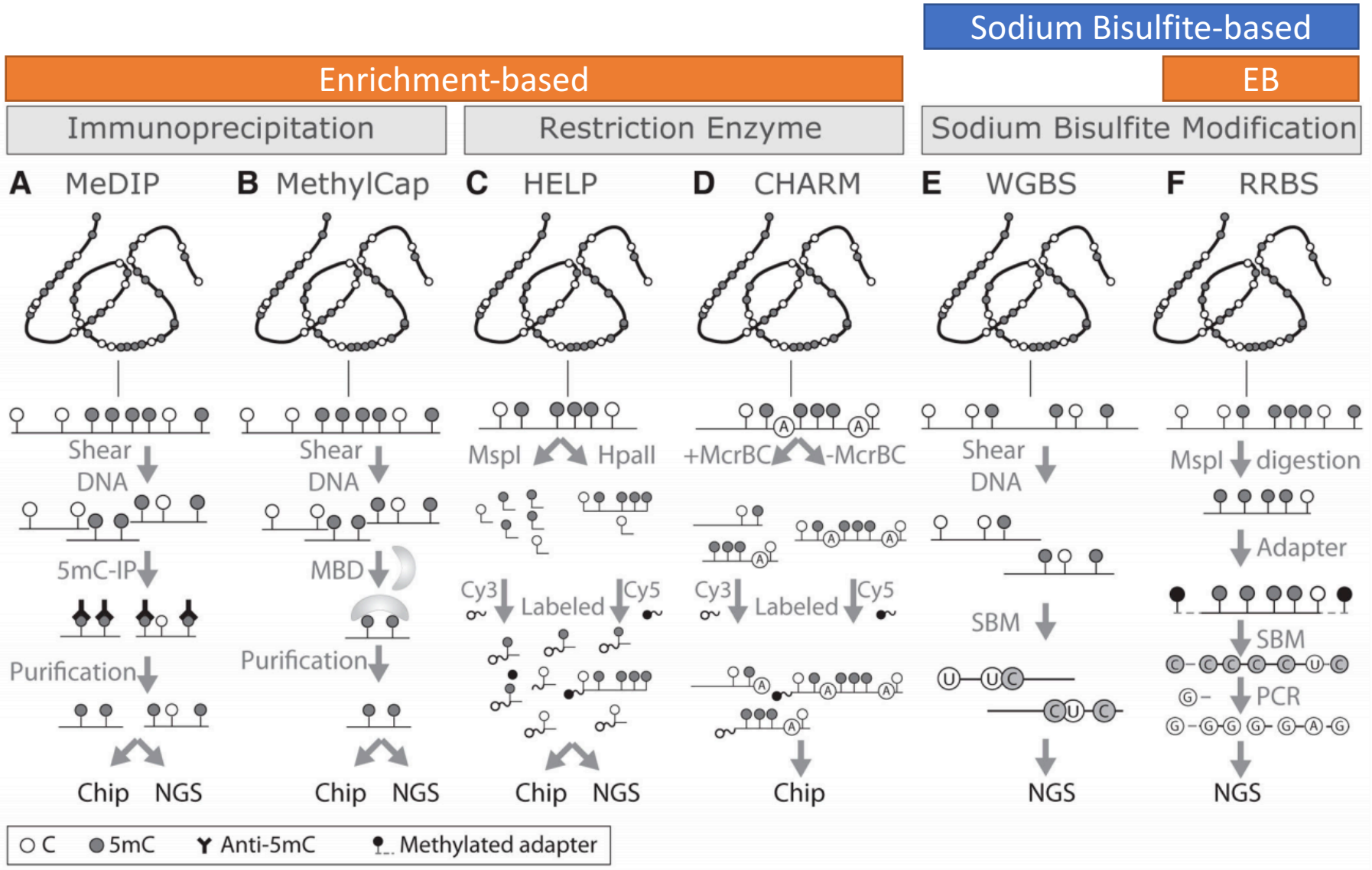
Carrió, E., & Suelves, M. (2015). DNA methylation dynamics in muscle development and disease. *Frontiers in aging neuroscience*, *7*, 19

Anatomy of a DNA methylome project

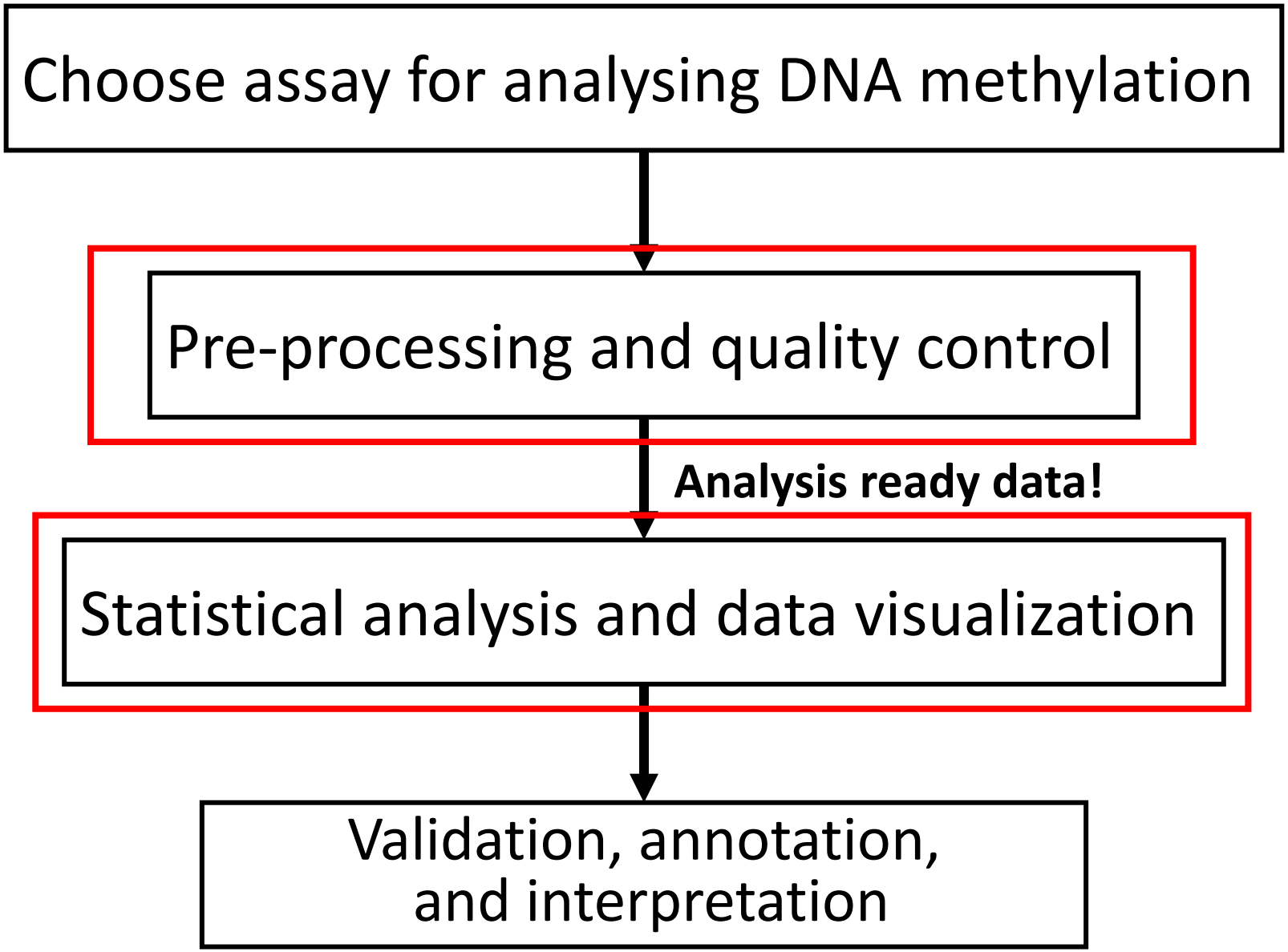


Bock, C. (2012). Analysing and interpreting DNA methylation data. *Nature Reviews Genetics*, *13*(10), 705.

Epigenomics methods for studying 5-mC DNA methylation



Marzese, D. M., & Hoon, D. S. (2015). Emerging technologies for studying DNA methylation for the molecular diagnosis of cancer. *Expert review of molecular diagnostics*, *15*(5), 647-664.



Bock, C. (2012). Analysing and interpreting DNA methylation data. *Nature Reviews Genetics*, *13*(10), 705.

Goal: Identify differentially methylated probe/site

Linear model

Observed signal for probe p

Coefficients for probe p

$$E(Y_p) = X\beta_p$$

Design matrix

$$\sim \beta_D + \beta_1 + \cdots + \beta_n + \varepsilon$$

Identifier	Disease	Gender	…	Age
Case01	1	1	…	21
Case02	1	0	…	25
Ctrl01	0	1	…	22
Ctrl02	0	0	…	27

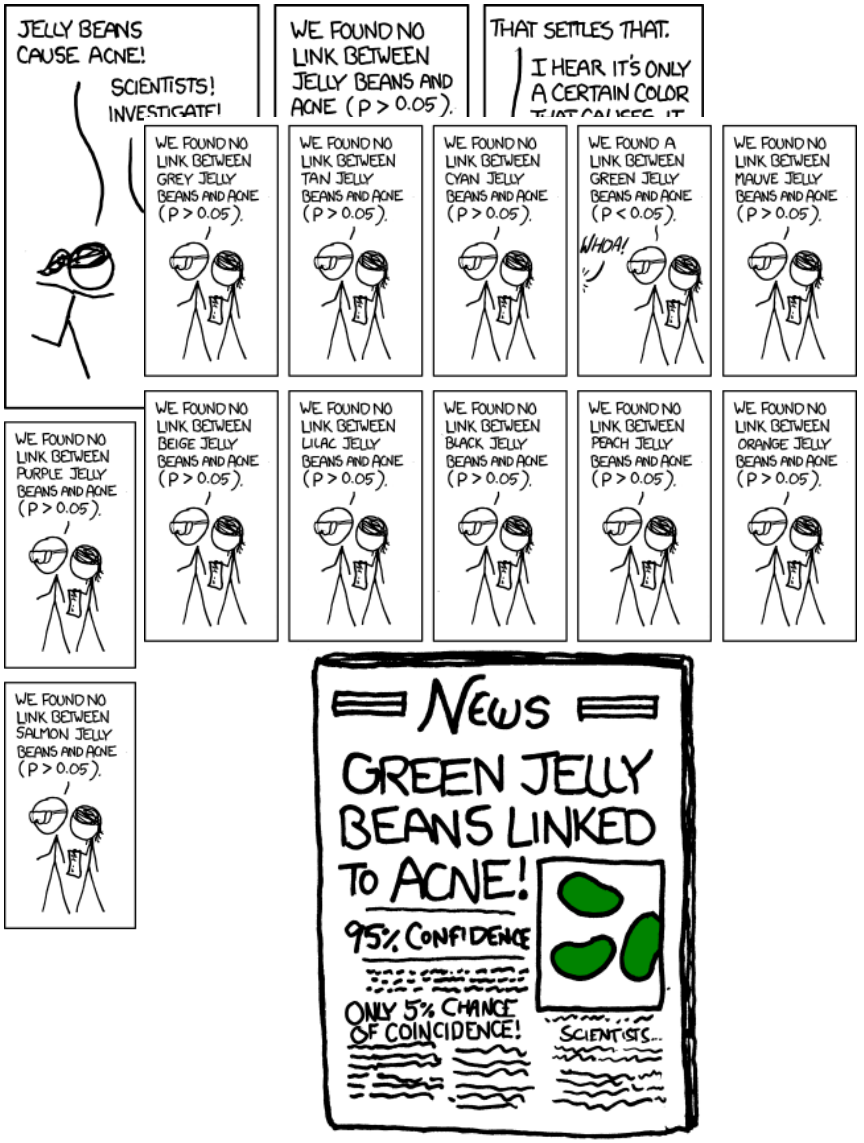
Multiple hypothesis correction

$$E(Y_1) = X\beta_1$$

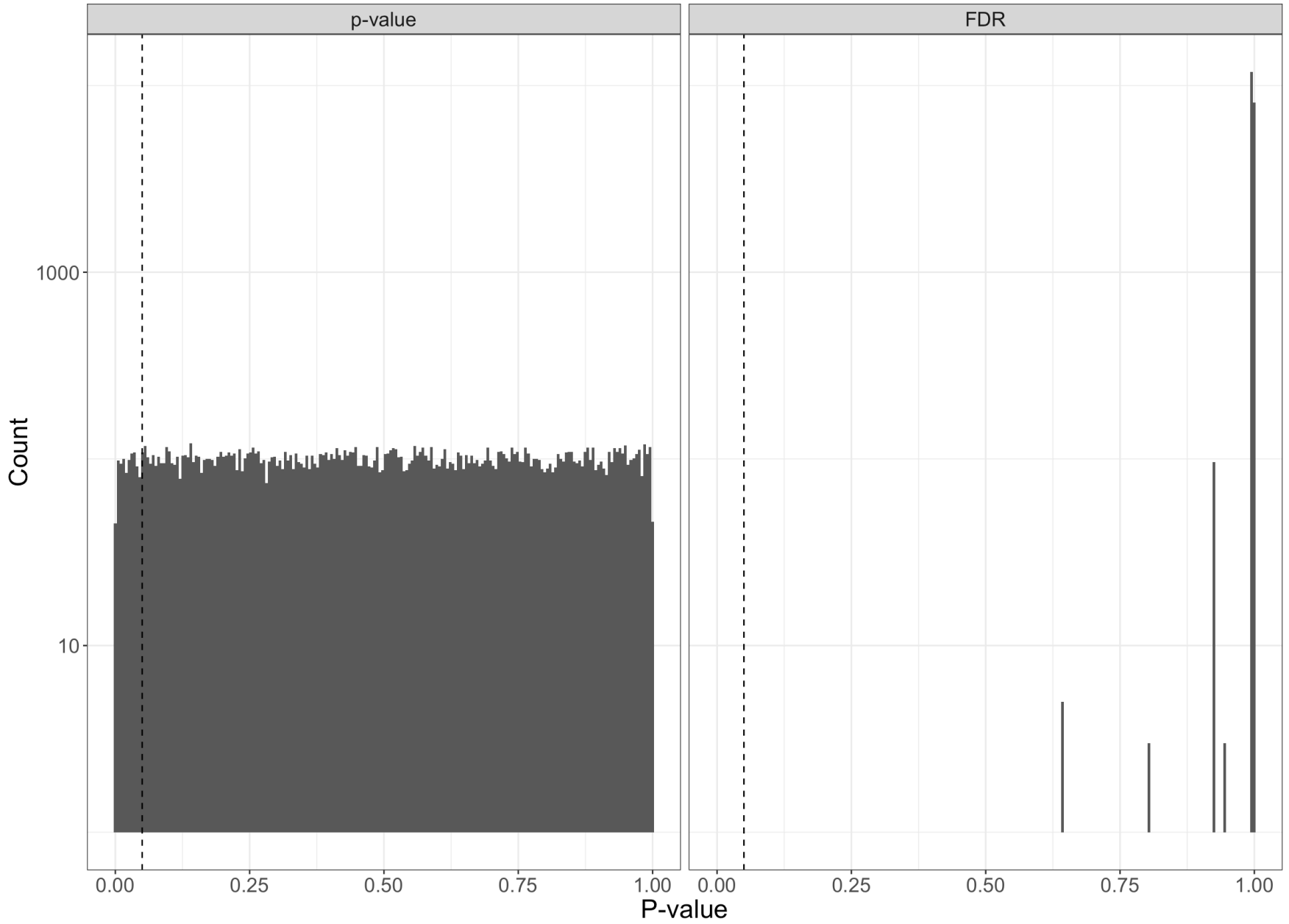
…

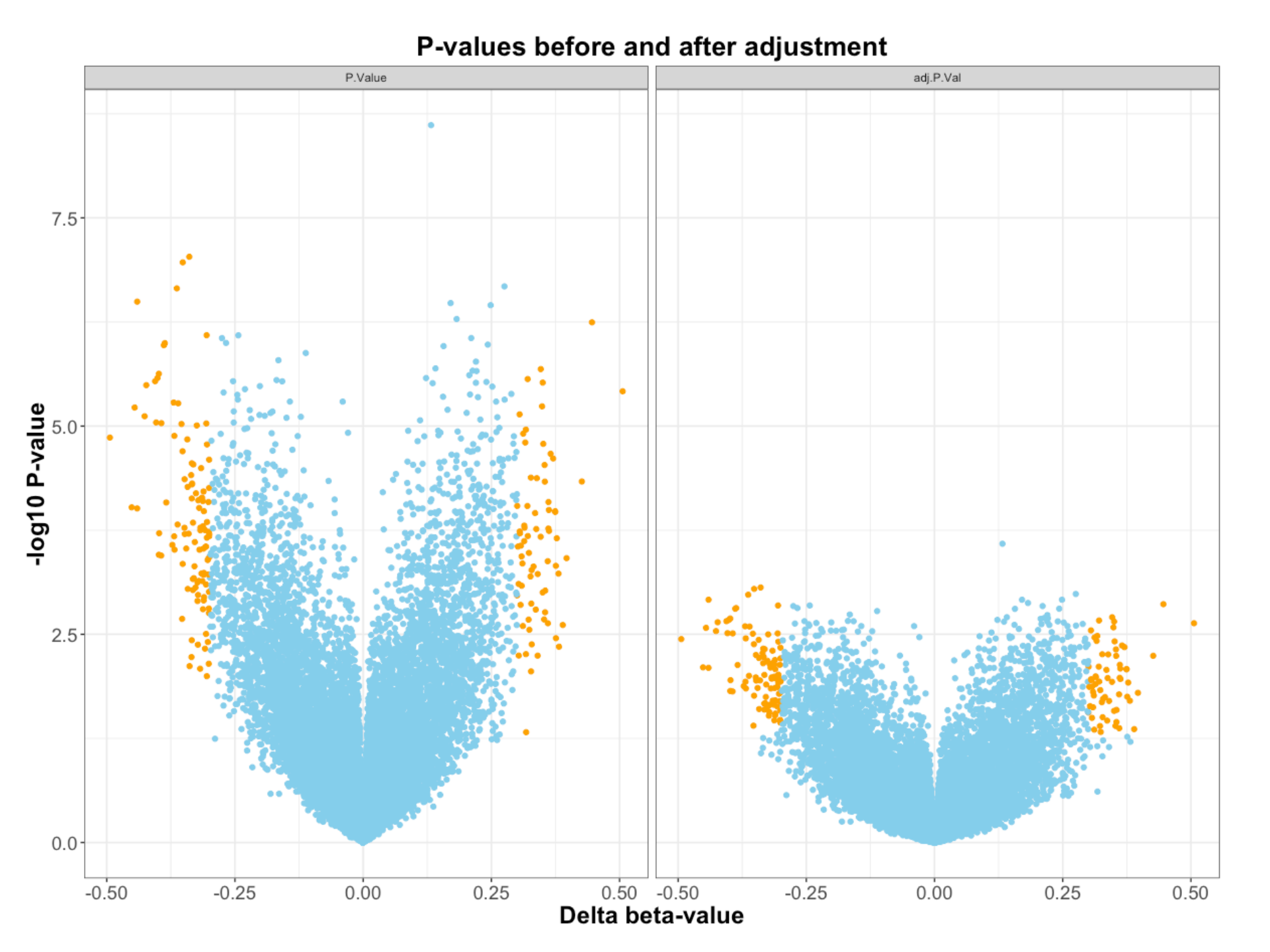
$$E(Y_n) = X\beta_n$$

n = 450,000



https://xkcd.com/882/


20,000 permutations of H0 = True
p-value
FDR
1000
10
Count
0.00
0.25
0.50
0.75
1.00
P-value

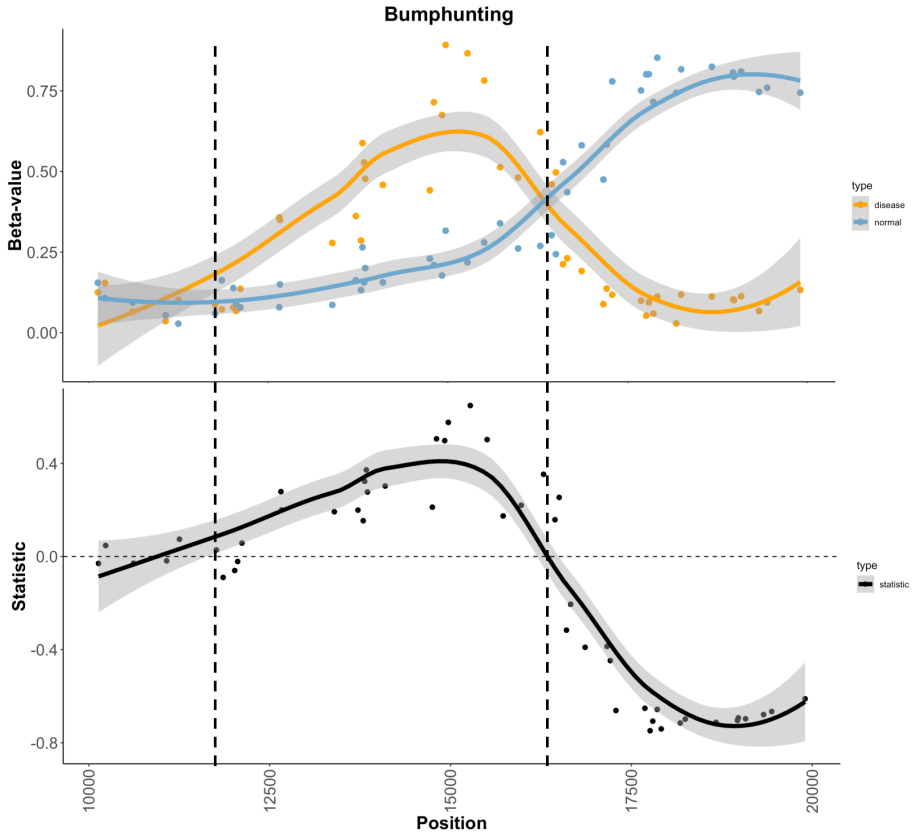

P-values before and after adjustment
P.Value
adj.P.Val
-log10 P-value
7.5
5.0
2.5
0.0
-0.50
-0.25
0.00
0.25
0.50
Delta beta-value

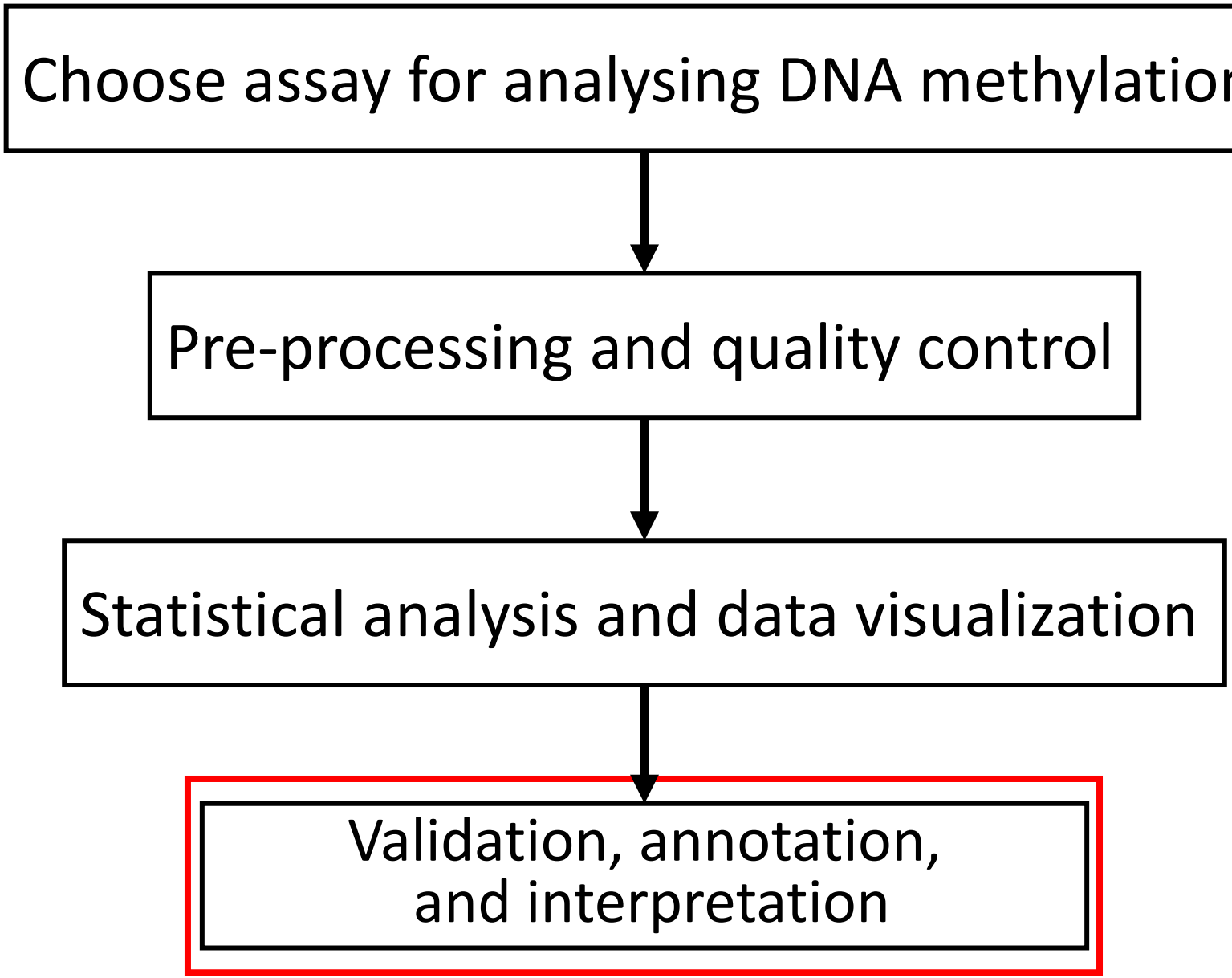
Differentially methylated regions (DMRs)

- Identify regions of differential methylation rather than a CpG site
  - Change across a region is biologically more relevant
  - Less features being tested
- Calculate differential methylation probabilities across continuous sites over a region

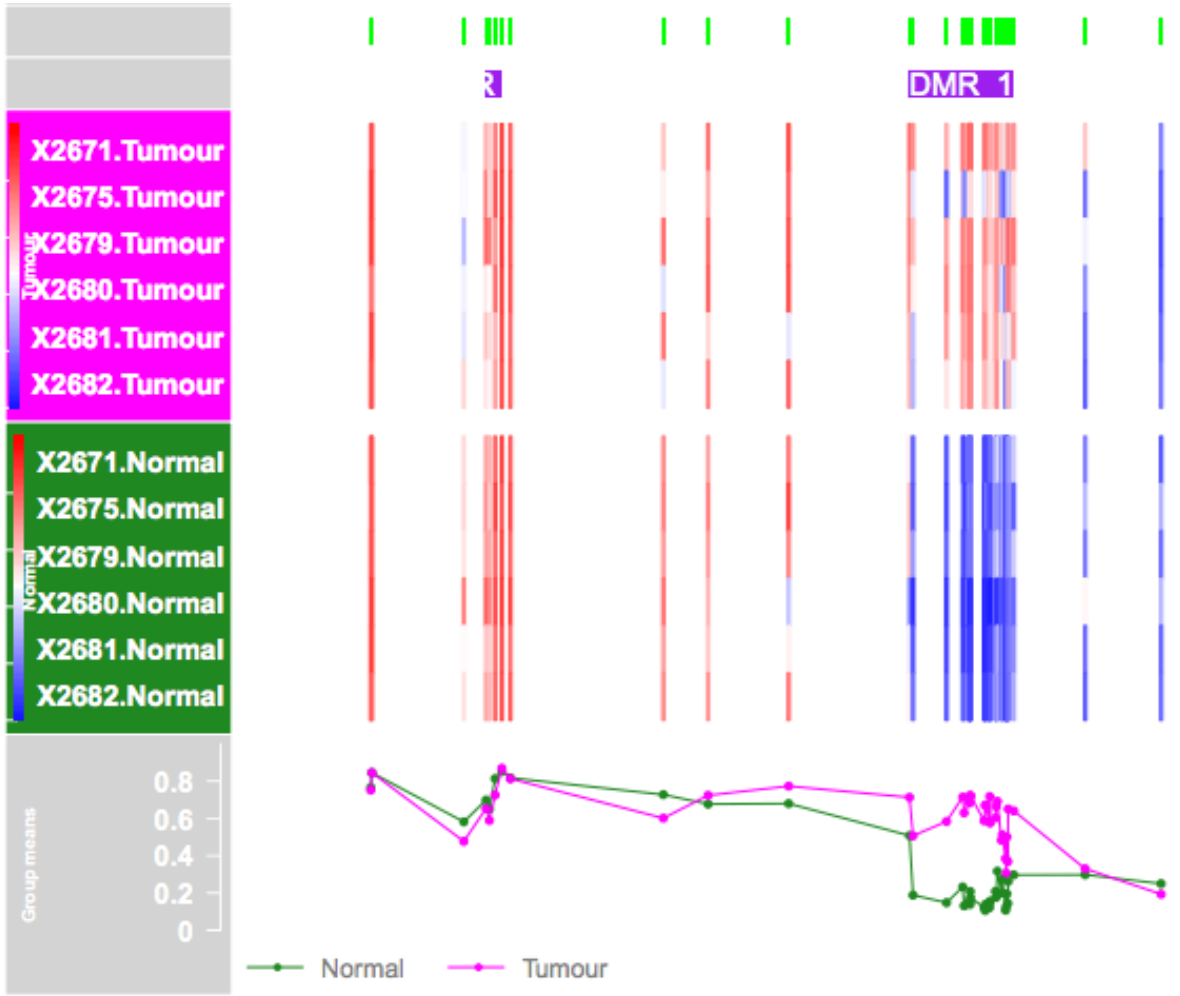
Software:

- minfi::bumphunter
- DMRcate
- CHAMP::probe lasso
- comb-p
- BS-seq





Bock, C. (2012). Analysing and interpreting DNA methylation data. *Nature Reviews Genetics*, *13*(10), 705.

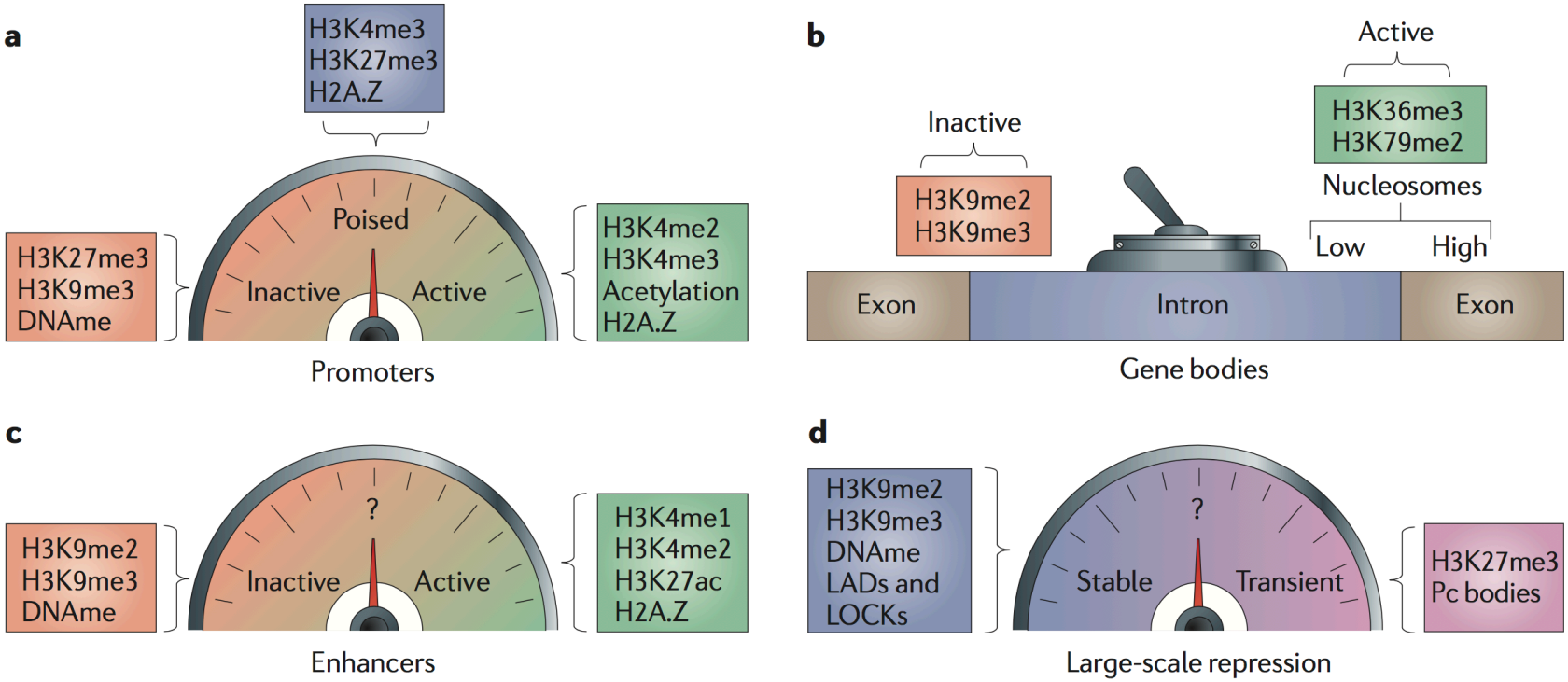


https://bioconductor.org/packages/release/bioc/vignettes/DMRcate/inst/doc/DMRcate.pdf

Histone modifications

The histone code hypothesis

- Combinatorial histone post-translational modifications (PTMs) encrypt the recruitment of different histone effectors
- These combinations define specific chromatin states
- These states are heritable

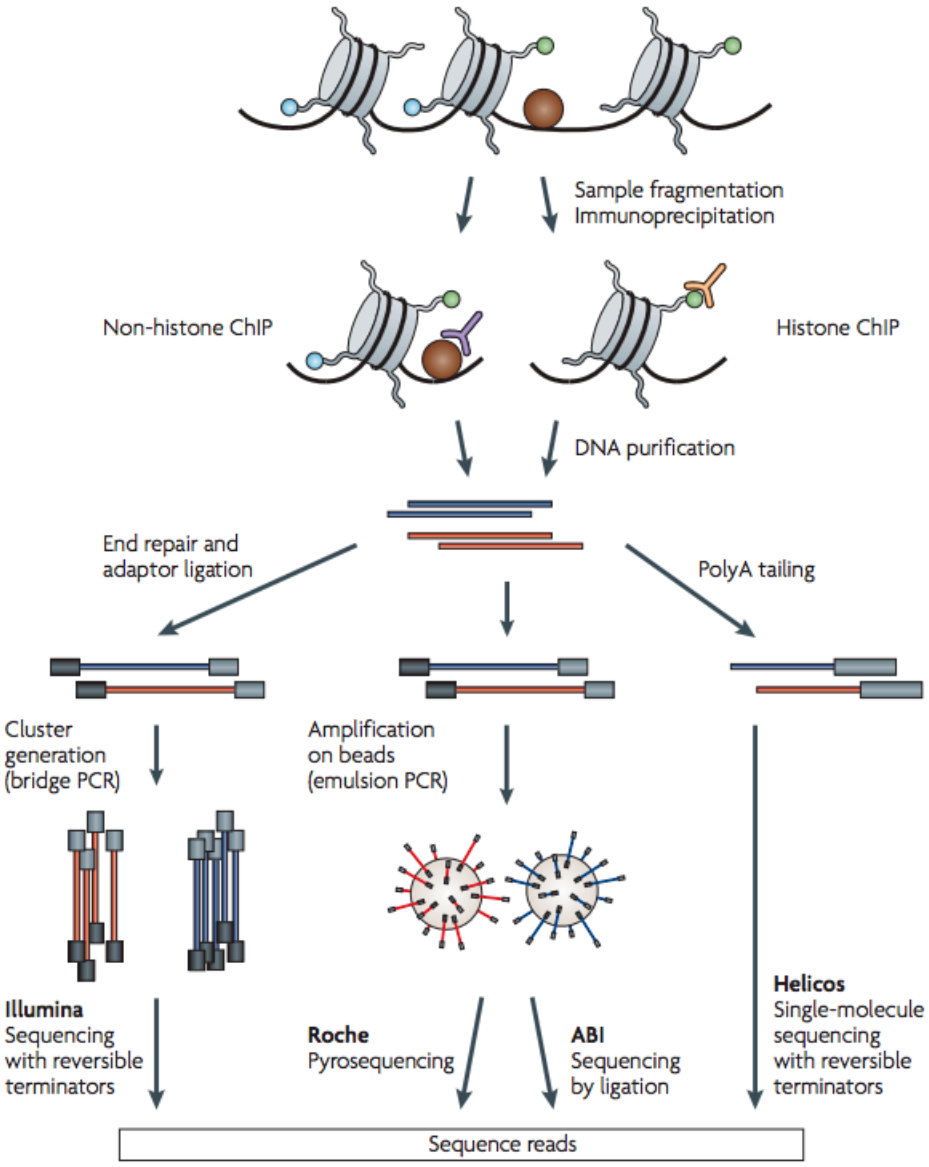


Strahl, B. D., & Allis, C. D. (2000). The language of covalent histone modifications. *Nature*, *403*(6765), 41

Turner, B. M. (2000). Histone acetylation and an epigenetic code. *Bioessays*, *22*(9), 836-845

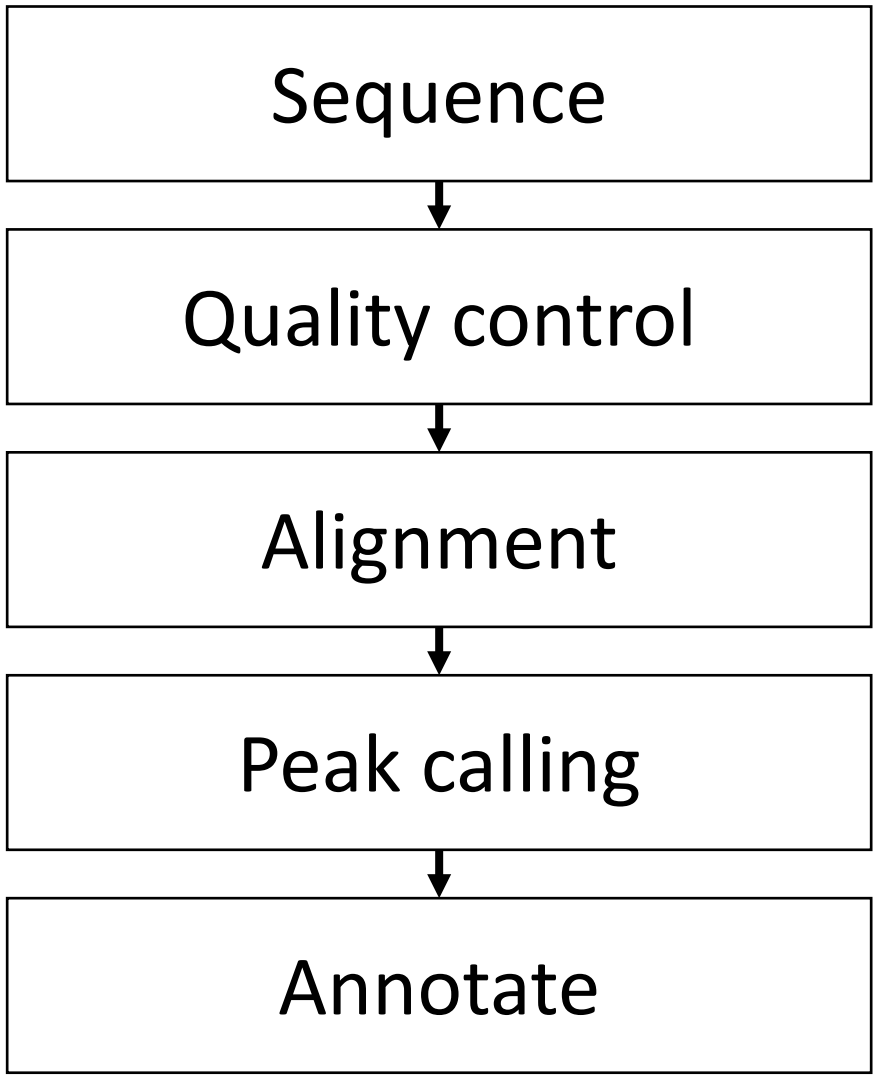
Zhou, V. W., Goren, A., & Bernstein, B. E. (2011). Charting histone modifications and the functional organization of mammalian genomes. *Nature Reviews Genetics*, *12*(1), 7

ChIP-Seq: Assay

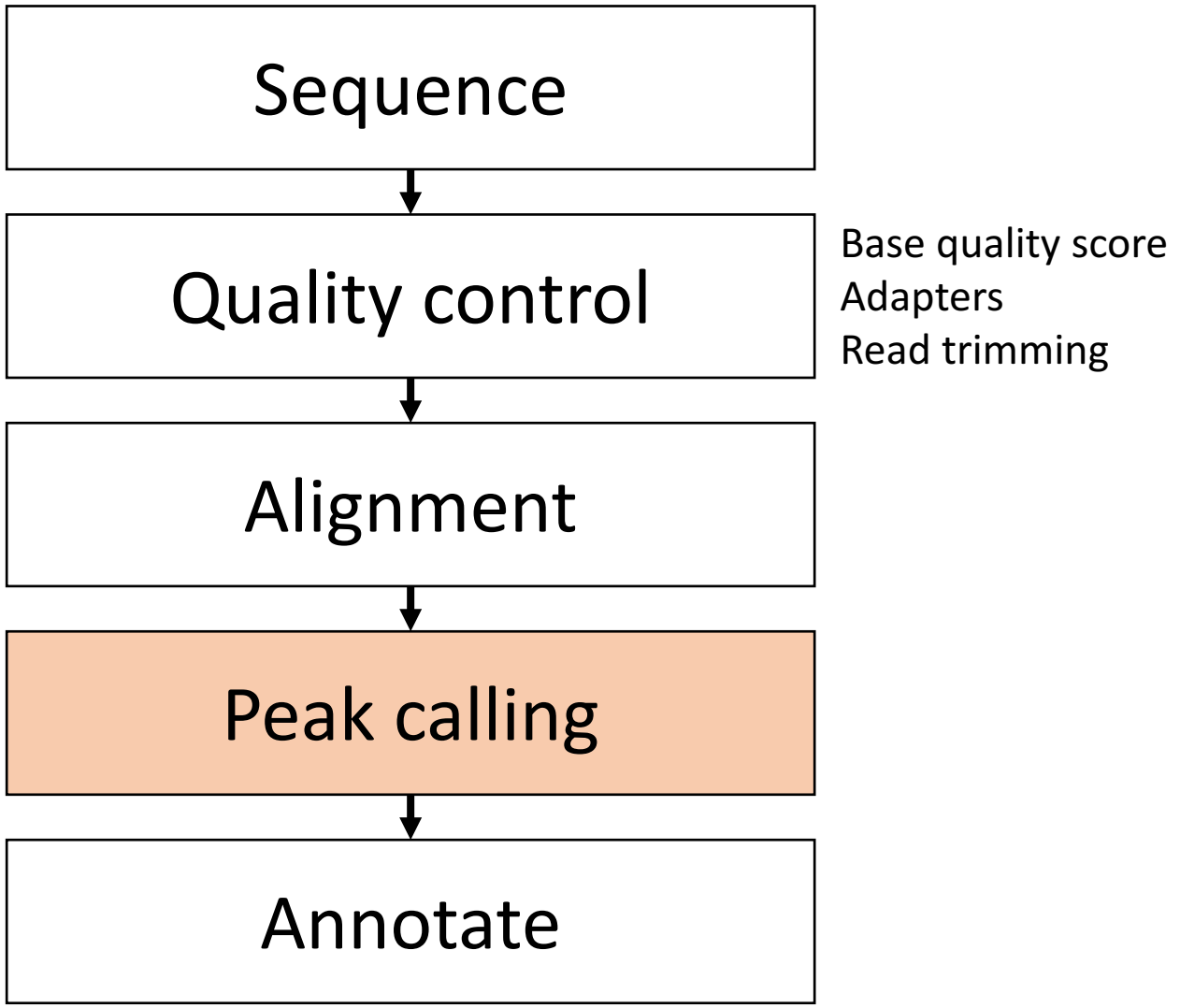


Park, P. J. (2009). ChIP–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, *10*(10), 669

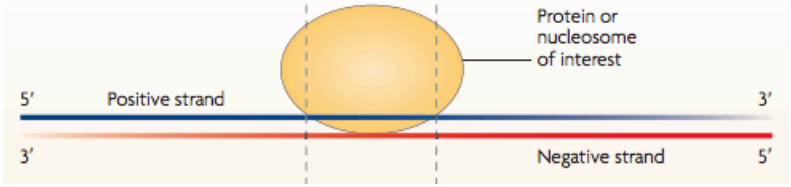
ChIP-Seq: Analysis pipeline

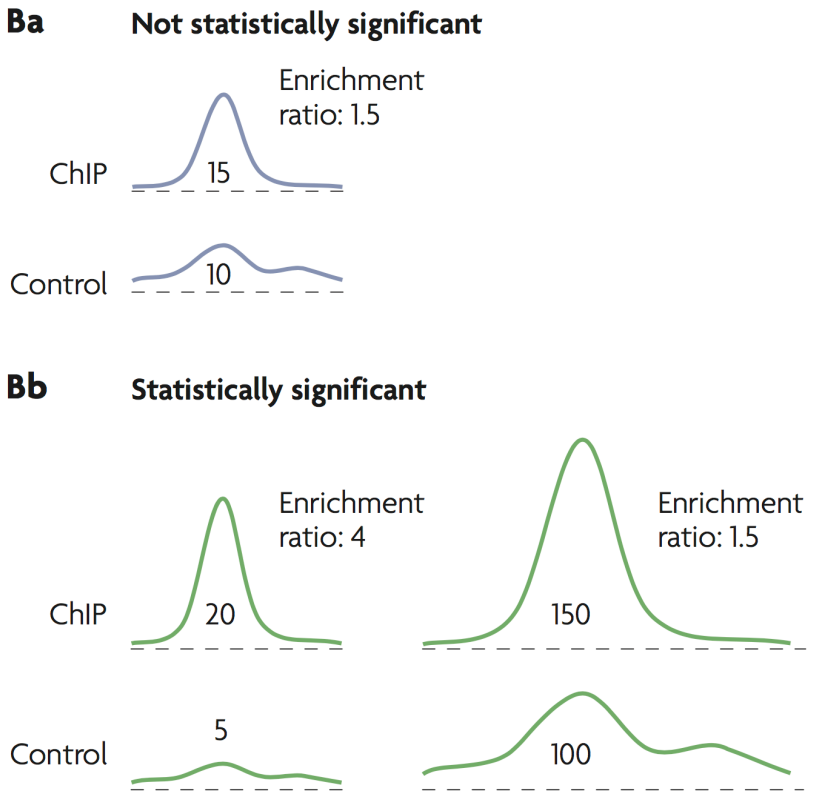


ChIP-Seq: Analysis pipeline



Goal: Identify enrichment peaks

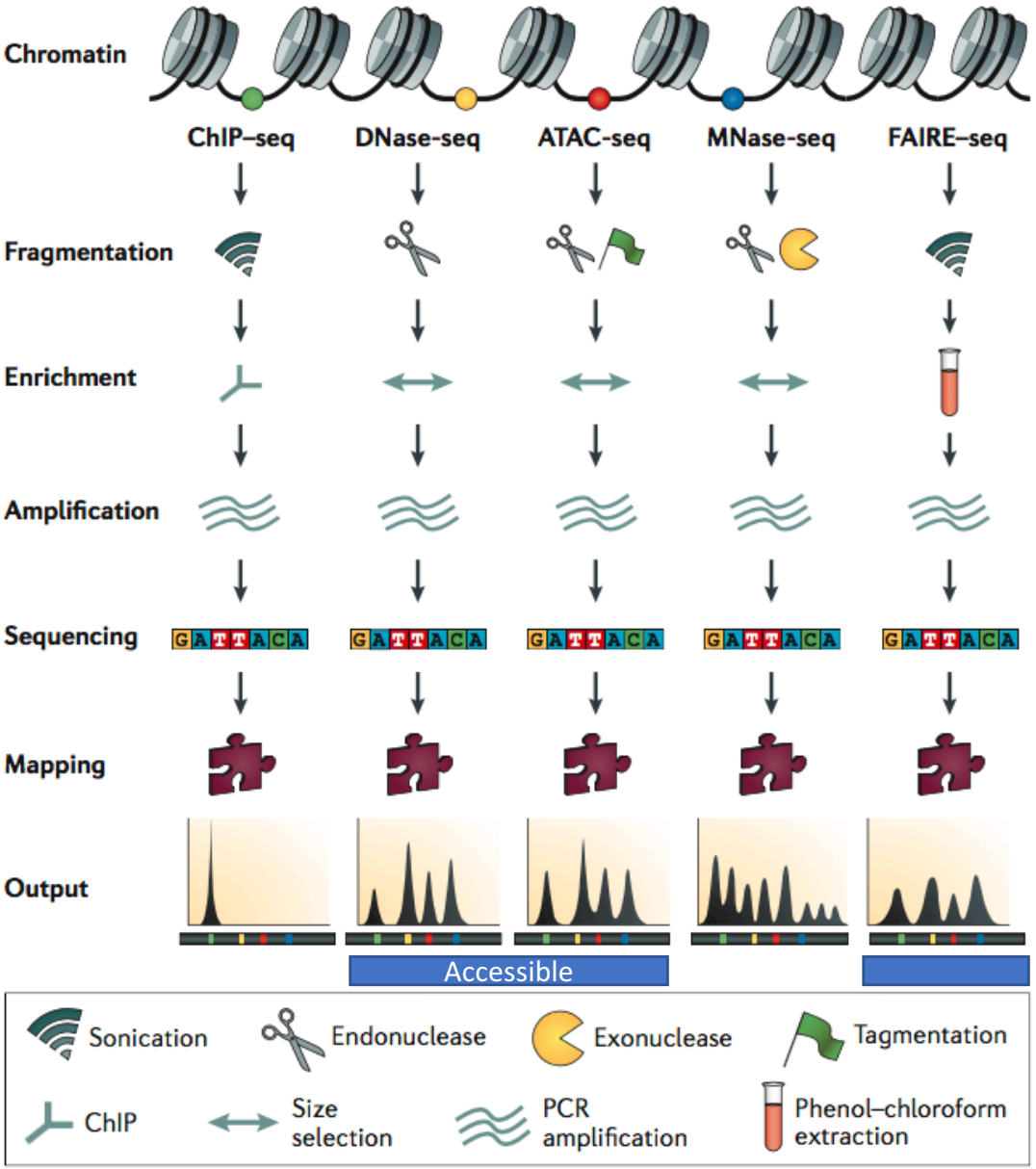




Park, P. J. (2009). ChIP–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, *10*(10), 669

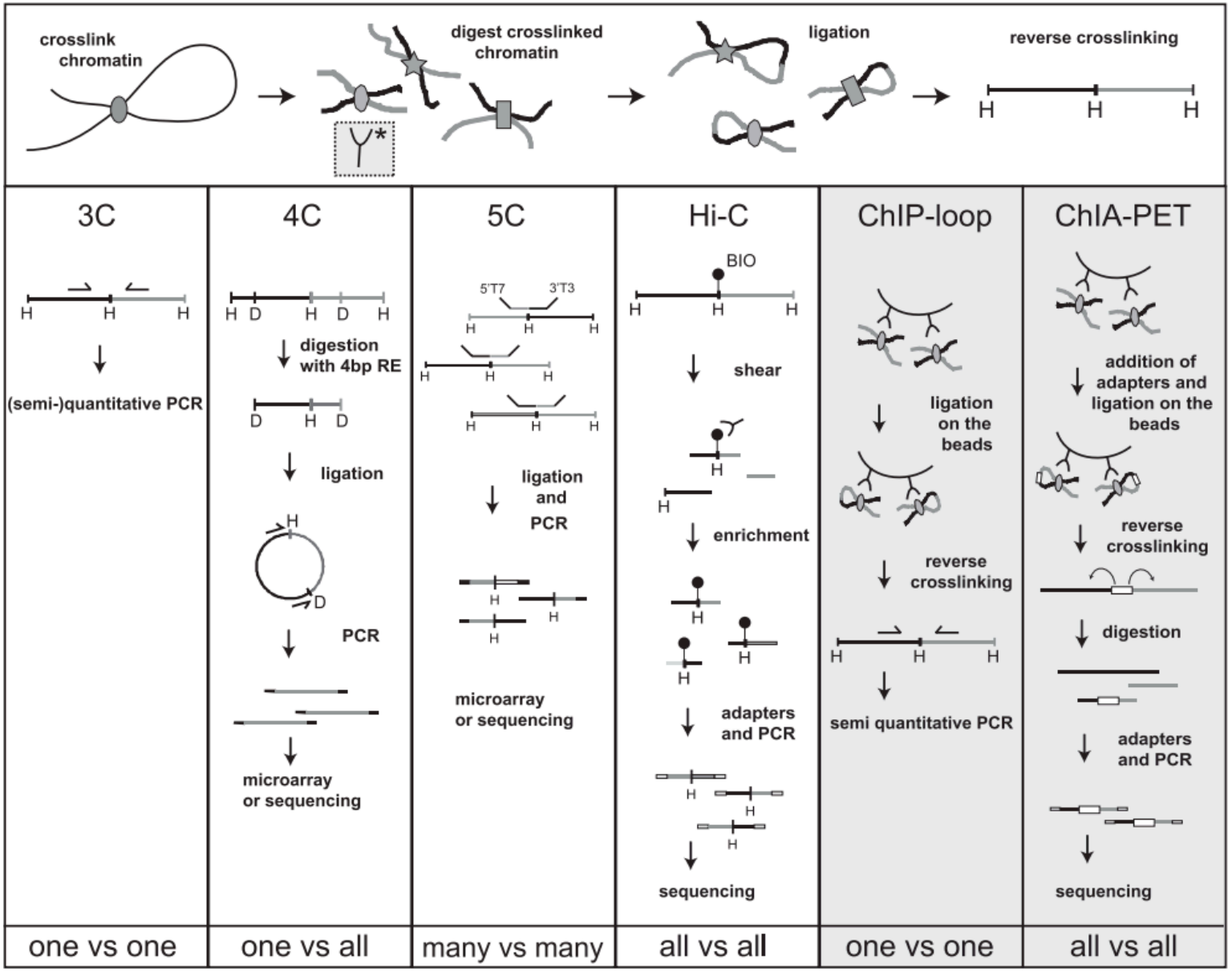
Chromatin features

Chromatin features: accessibility



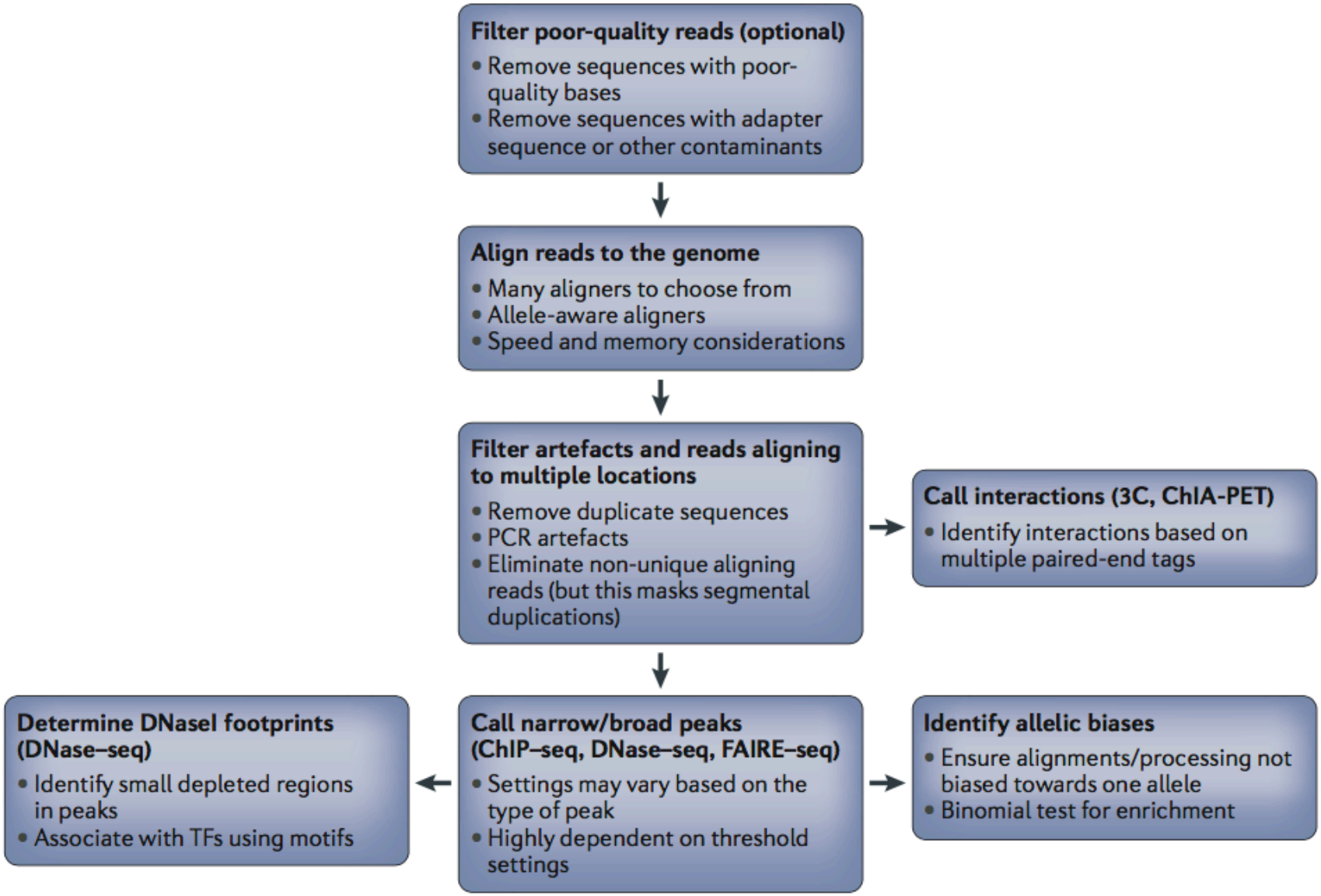
Meyer, C. A., & Liu, X. S. (2014). Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat Rev Genetics*, *15*(11), 709

Chromatin features: long range interactions



de Wit, E., & de Laat, W. (2012). A decade of 3C technologies: insights into nuclear organization. *Genes & development*, *26*(1), 11-24

Post-sequencing bioinformatics pipeline

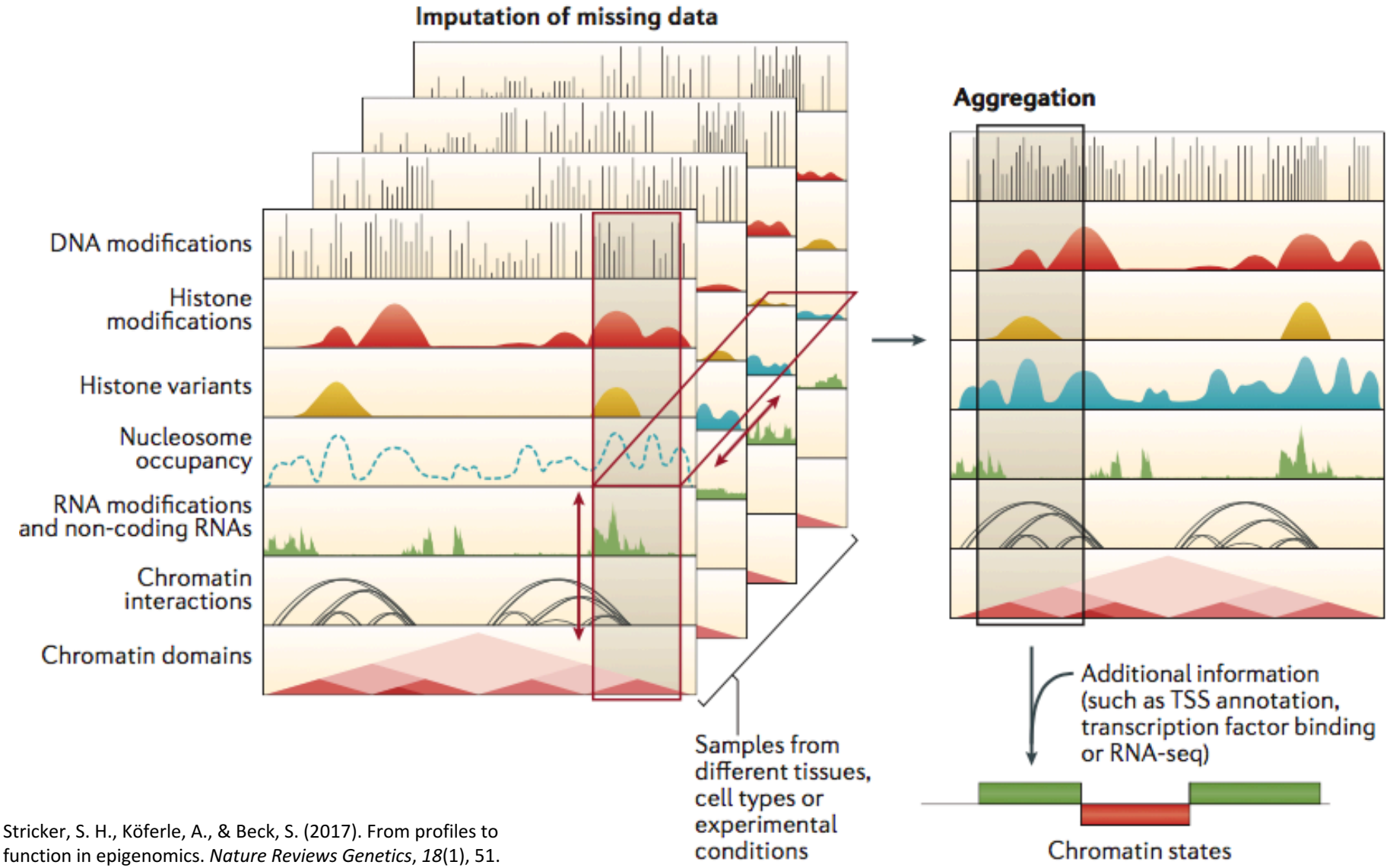


Furey, T. S. (2012). ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. *Nature Reviews Genetics*, *13*(12), 840

Integration

Putting it all together

Imputation and integration to infer functional states



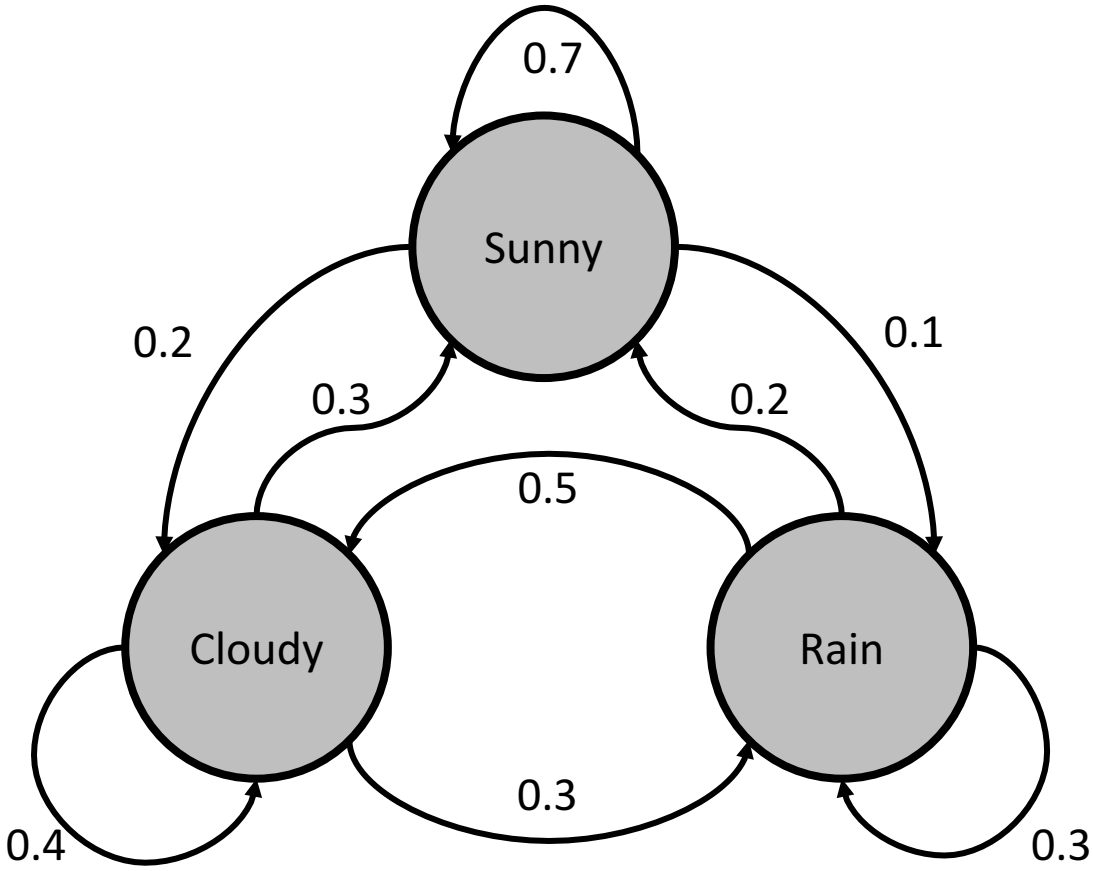
Stricker, S. H., Köferle, A., & Beck, S. (2017). From profiles to function in epigenomics. *Nature Reviews Genetics*, *18*(1), 51.

ChromHMM: Inferring chromatin states

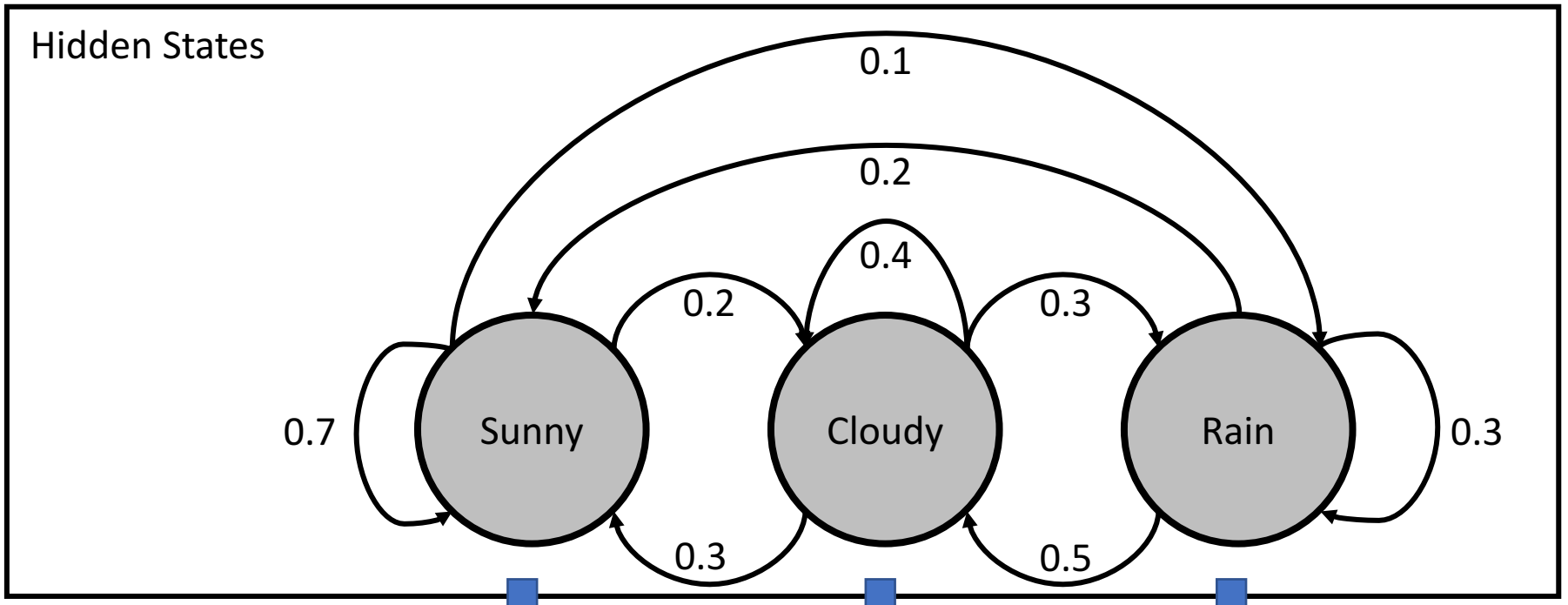
- Purpose: identify segments of the genome as biologically relevant functional elements
- ChromHMM uses a multivariate hidden Markov model (HMM) to infer chromatin states
- Chromatin states are derived from combinations of epigenetics marks (histone modifications, chromatin features)
- Chromatin states include:
  - Active/Weak
  - Promoter, enhancer, insulator
  - Polycomb

Simple Hidden Markov Model

Markov Chain for weather



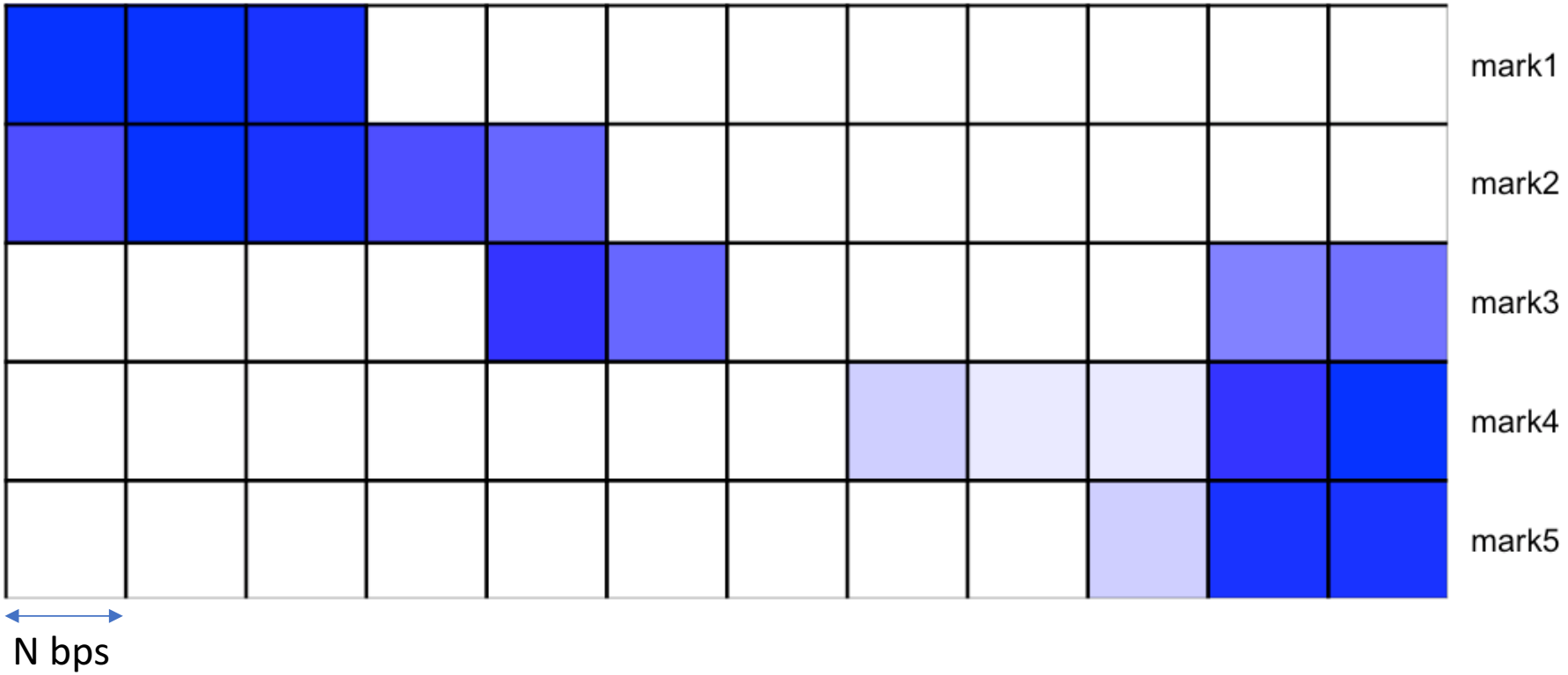
Simple Hidden Markov Model



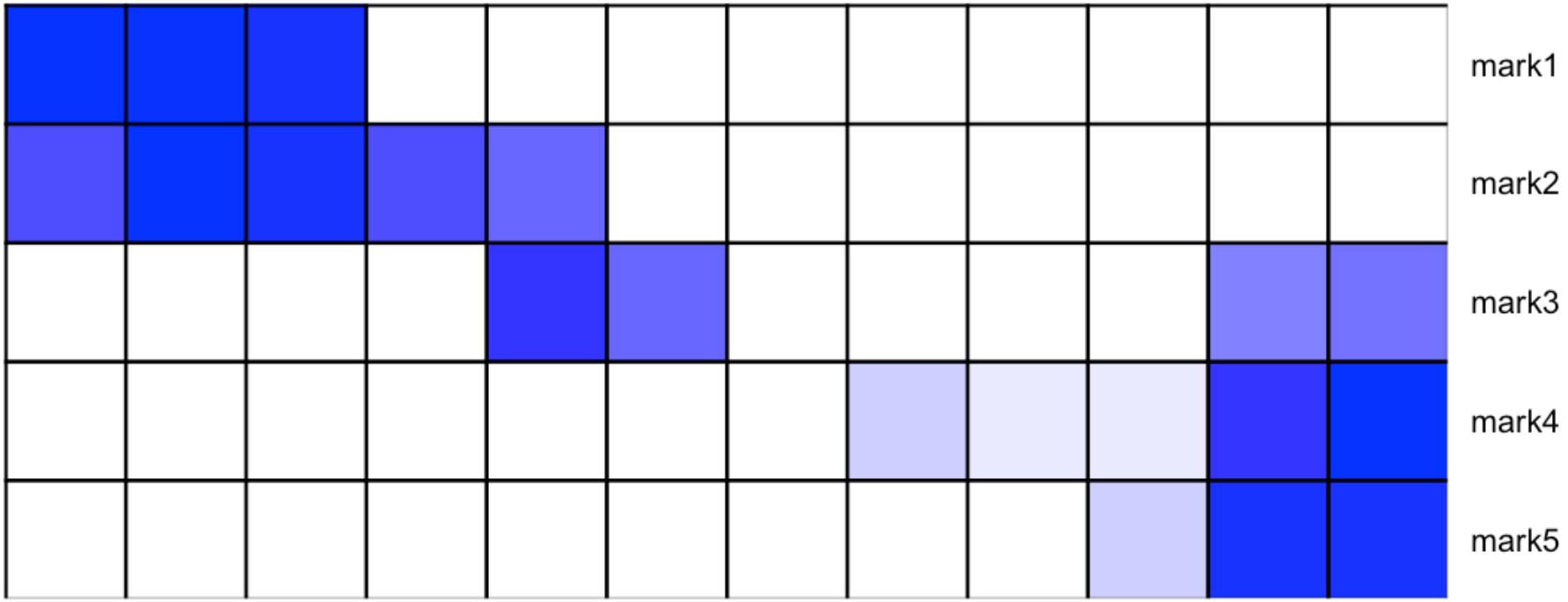
Emission probabilities

	Sunny	Cloudy	Rain
Raincoat	0.1	0.6	0.9
No raincoat	0.9	0.4	0.1

Observation

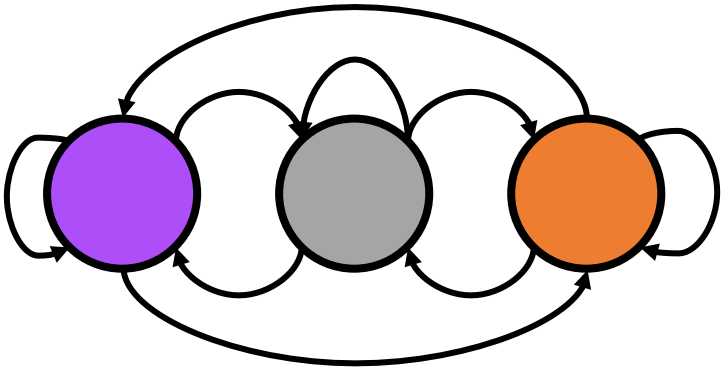


Observation

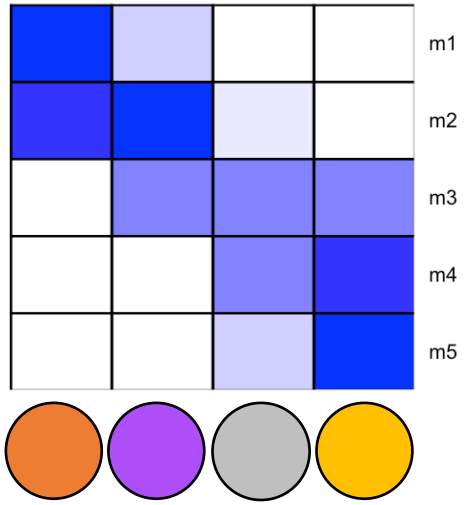


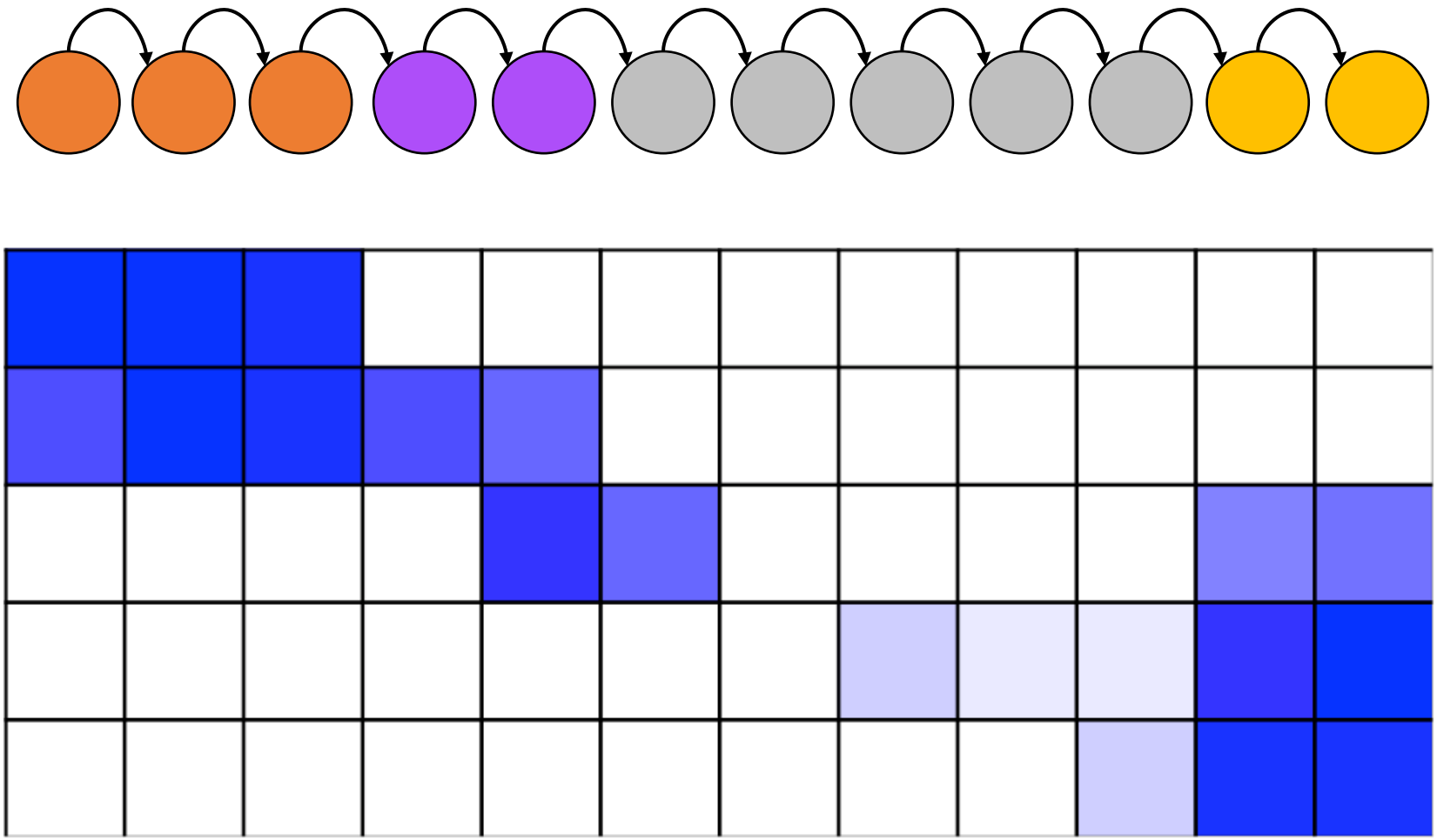
Learn from the data

1. Transition probabilities

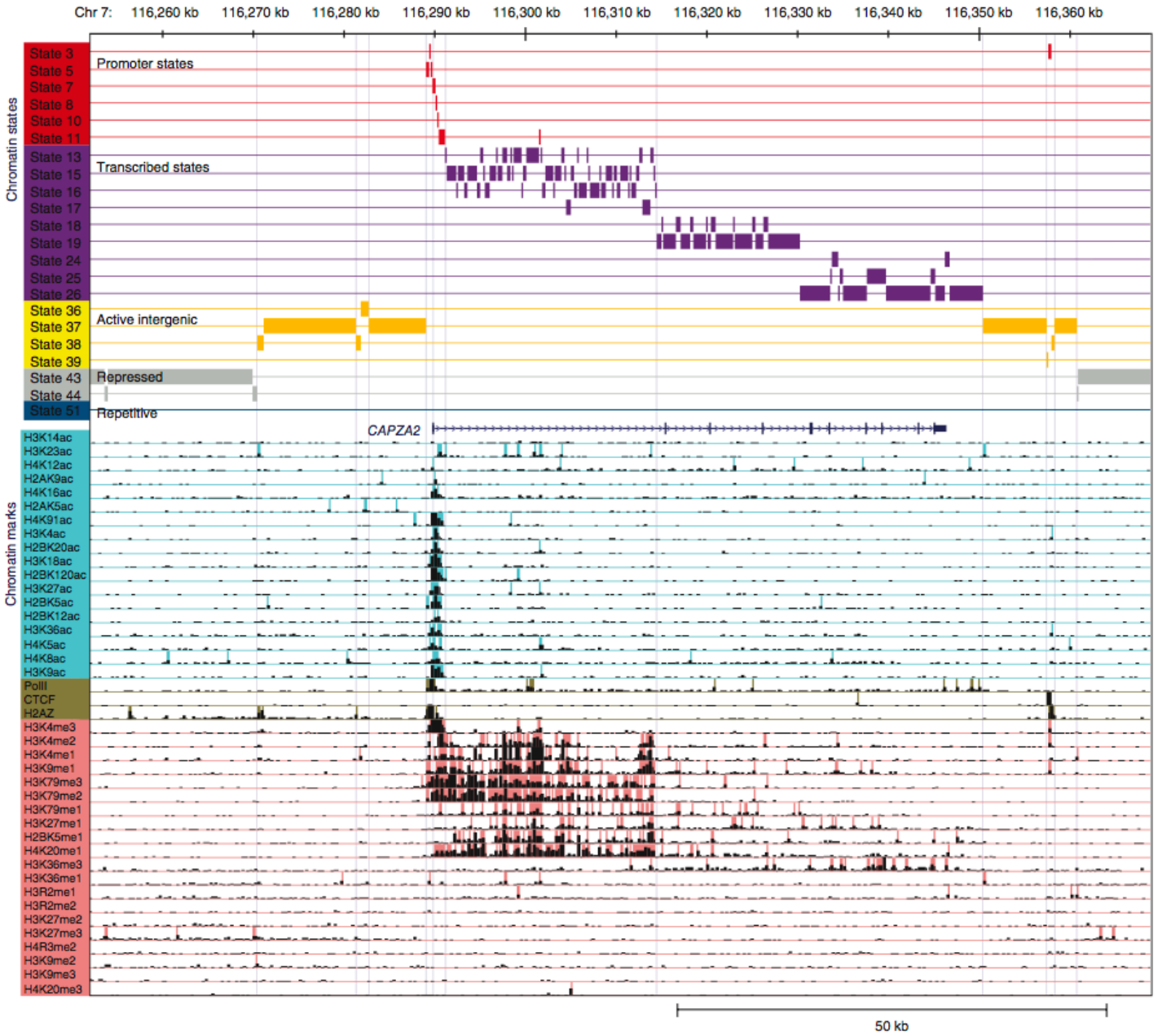


2. Emission probabilities




Chromatin states
mark1
mark2
mark3
mark4
mark5
Genomic Annotations
TSS
Gene body
Enhancer

Chromatin states



Ernst, J., & Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature biotechnology*, *28*(8), 817

Integrative genomics

Case 02:
Integrating genomics, epigenomics, and experimental biology to identify causal allele and regulatory mechanism in obesity

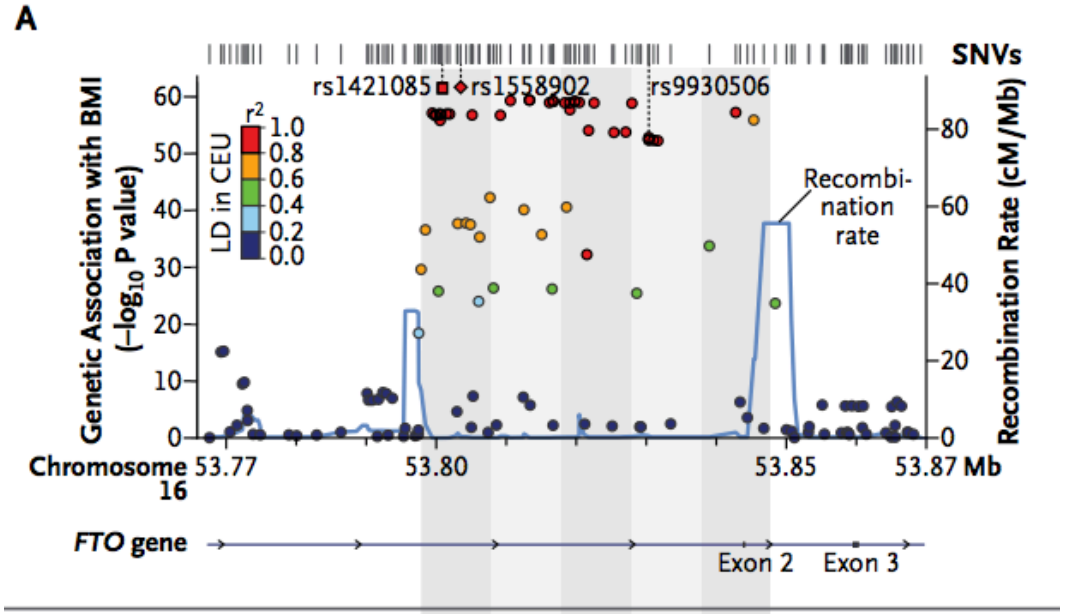
The NEW ENGLAND JOURNAL of MEDICINE

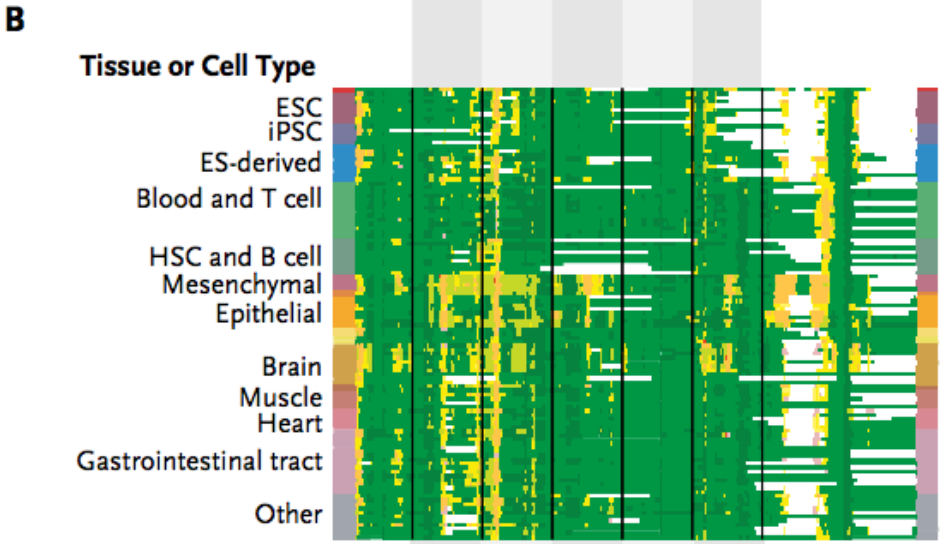
ESTABLISHED IN 1812 SEPTEMBER 3, 2015 VOL. 373 NO. 10

FTO Obesity Variant Circuitry and Adipocyte Browning in Humans

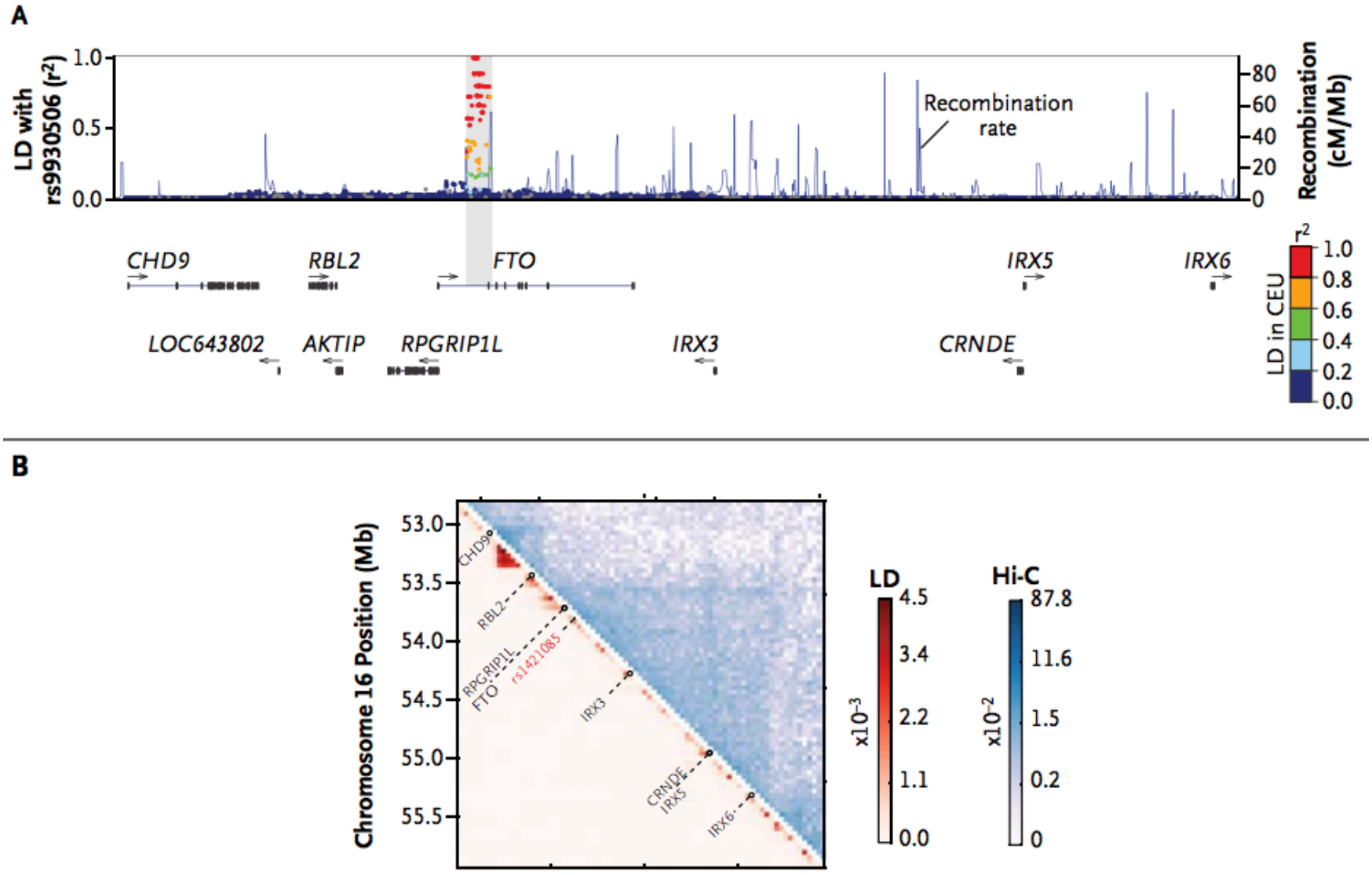
Melina Claussnitzer, Ph.D., Simon N. Dankel, Ph.D., Kyoung-Han Kim, Ph.D., Gerald Quon, Ph.D., Wouter Meuleman, Ph.D., Christine Haugen, M.Sc., Viktoria Glunk, M.Sc., Isabel S. Sousa, M.Sc., Jacqueline L. Beaudry, Ph.D., Vijitha Puviindran, B.Sc., Nezar A. Abdennur, M.Sc., Jannel Liu, B.Sc., Per-Arne Svensson, Ph.D., Yi-Hsiang Hsu, Ph.D., Daniel J. Drucker, M.D., Gunnar Mellgren, M.D., Ph.D., Chi-Chung Hui, Ph.D., Hans Hauner, M.D., and Manolis Kellis, Ph.D.

GWAS locus prioritization using epigenome map of human cells

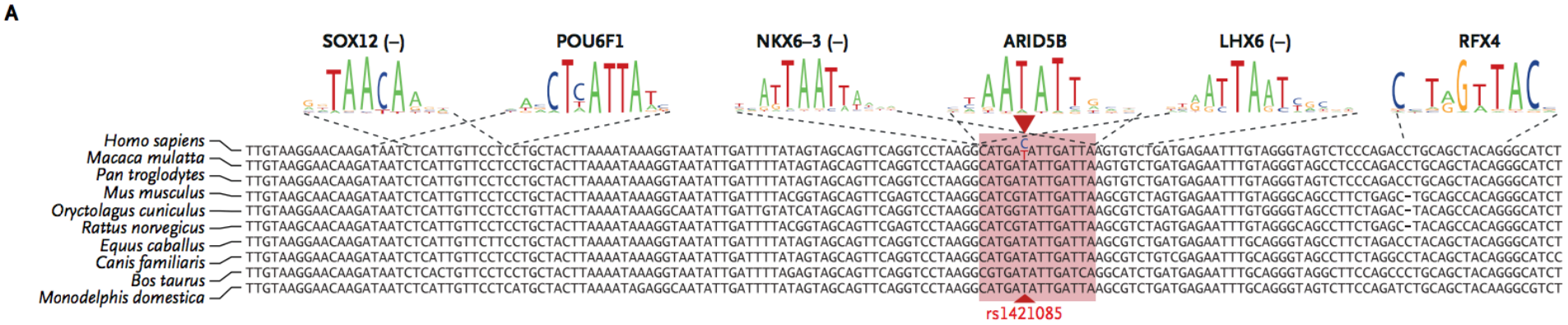




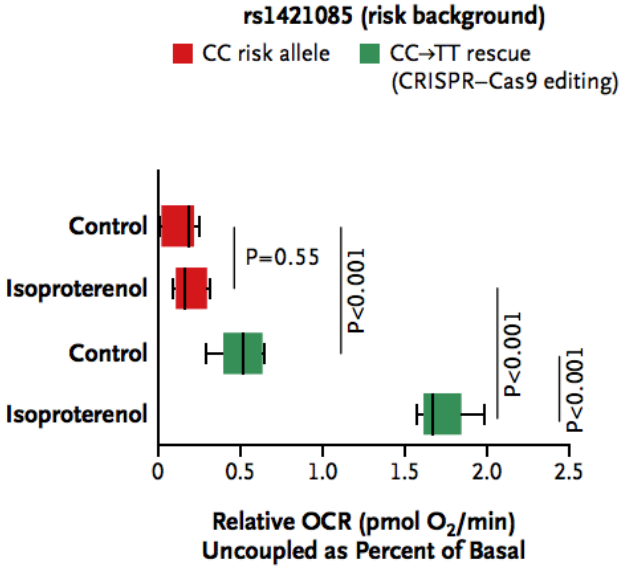
Hi-C data reveals 2-Mb topologically associated domain (TAD)



Identification of an SNV in linkage disequilibrium with candidate risk allele which disrupts an ARID5B repressor motif



CRISPR-CAS9 rescue of homozygous risk allele restores oxygen consumption in adipocytes from carriers



Limitations and considerations

Enrichment methods

Bisulfite-based methods

- Incomplete bisulfite conversion
  - Leads to increased methylation signal
  - Identification
    - Control probes in microarray
    - Methylation of non-CpG sites
    - Methylation-specific PCR of non-CpG sites

Affinity-based methods

- Highly dependent on antibody used
  - Include input control and other appropriate controls

Measurement

Array

- Sequence specific bias within reference genome
- Neighboring SNPs may interfere with binding

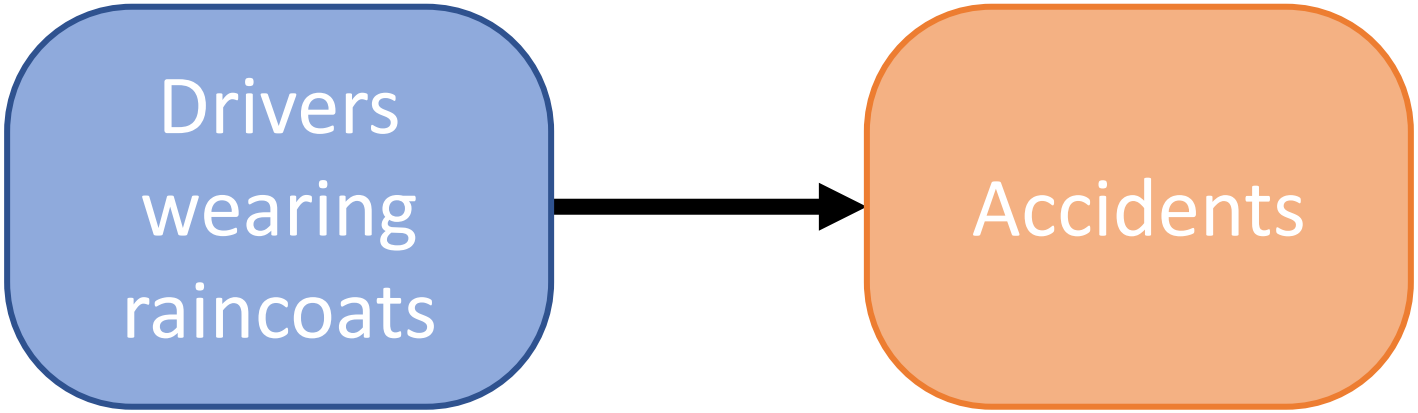
Sequencing

- Alignment
  - Repetitive regions are difficult to align
  - WGBS aligners have different biases depending on reference genome
    - Operator can also change parameters such as mismatch penalty
- Uneven depth across genome, may bias peak calling
  - Include input DNA for capture-based methods

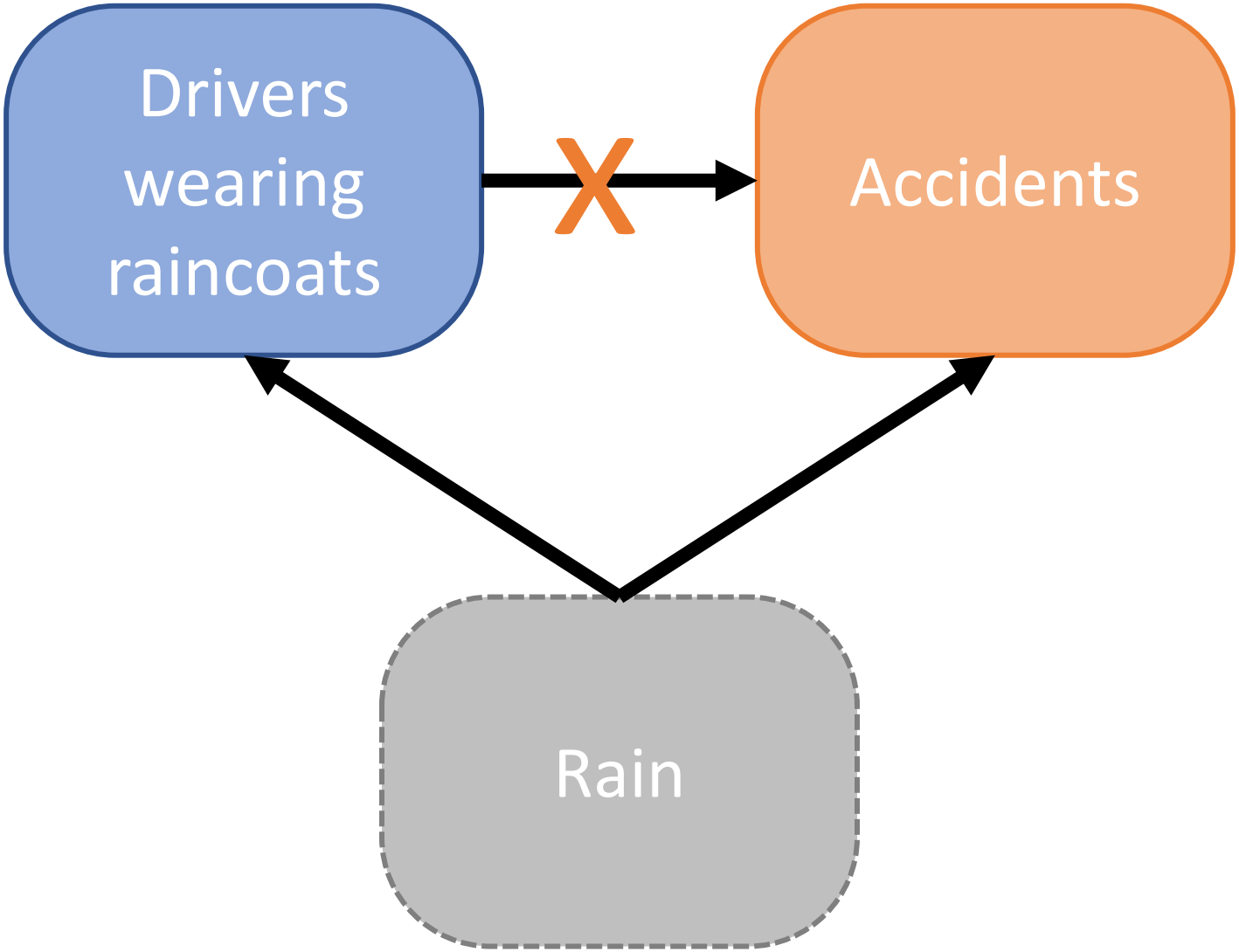
General statistical & design considerations

- Correlation does not mean causation
  - Experiments are required to establish causality
- Balanced study
- Replication using an independent cohort is important
  - Especially for biomarker or clustering studies
  - Training and validation cohorts should be representative of each other
- Orthogonal validation of sequencing or microarray results
- Accounting for potential sources of variation

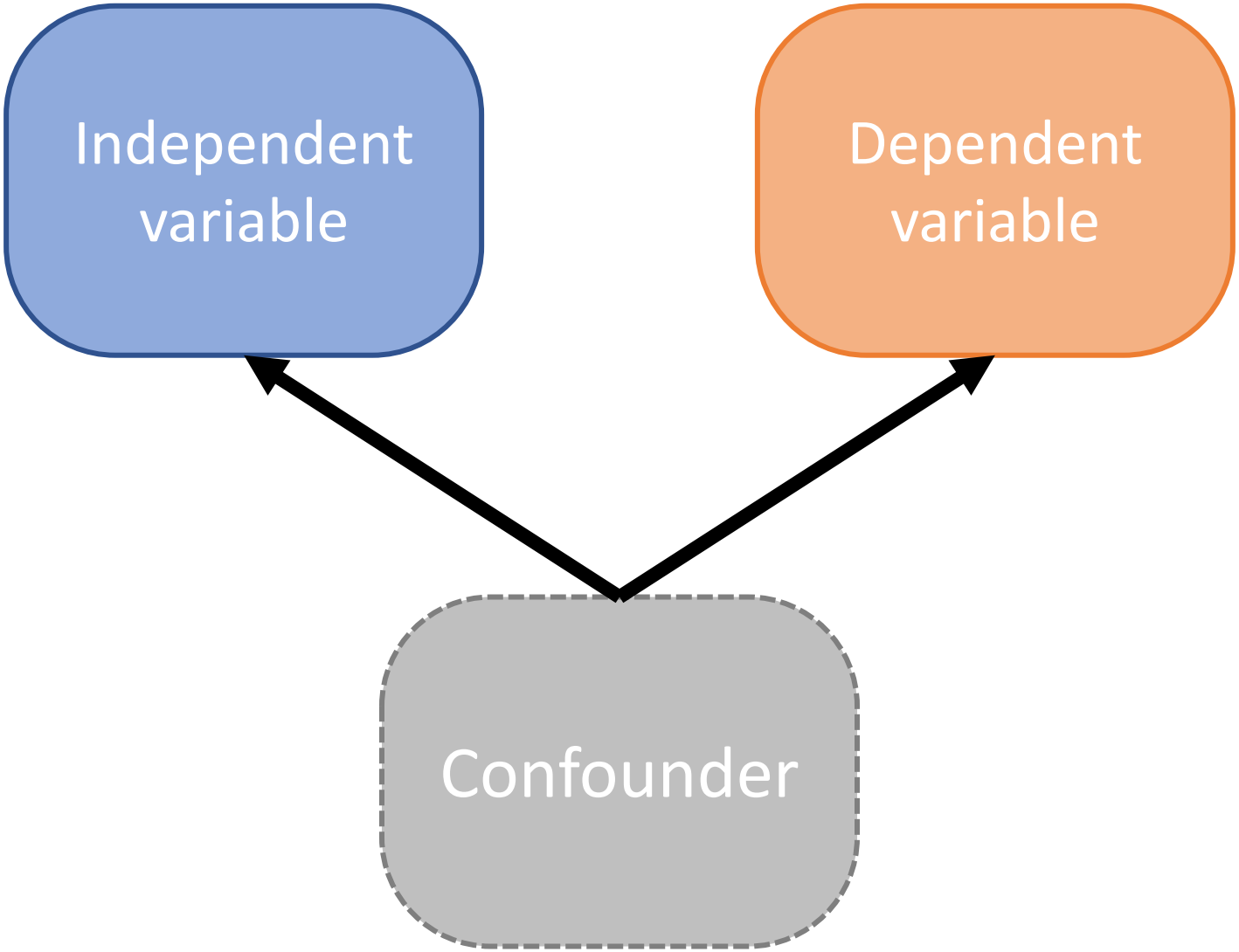
Confounding



Confounding



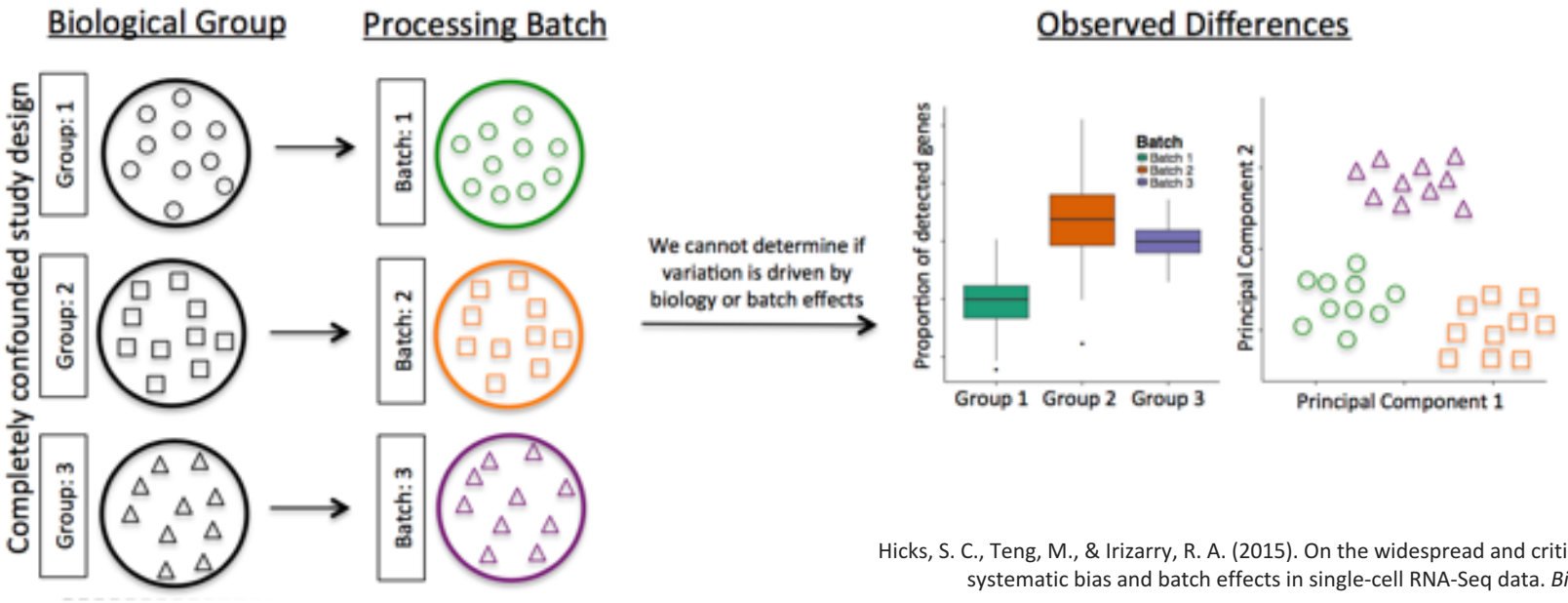
Confounding



Batch as a confounder

Disease
Treatment
Handedness

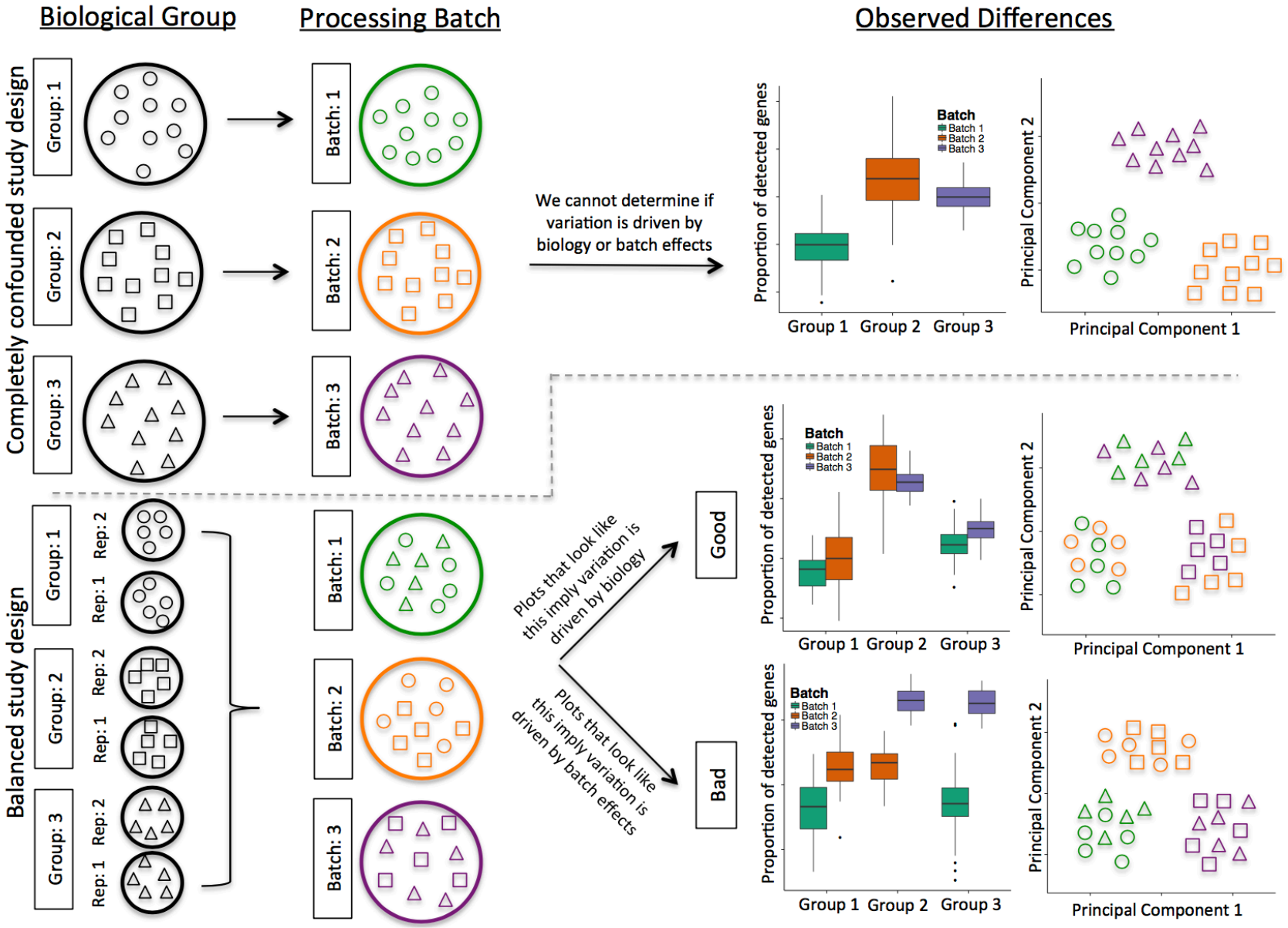
Date
Operator
Plate



Hicks, S. C., Teng, M., & Irizarry, R. A. (2015). On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data. *BioRxiv*, *25528*

- Study design: randomization or blocked design
- Batch correction: estimate signals contributed from known batches and account in model

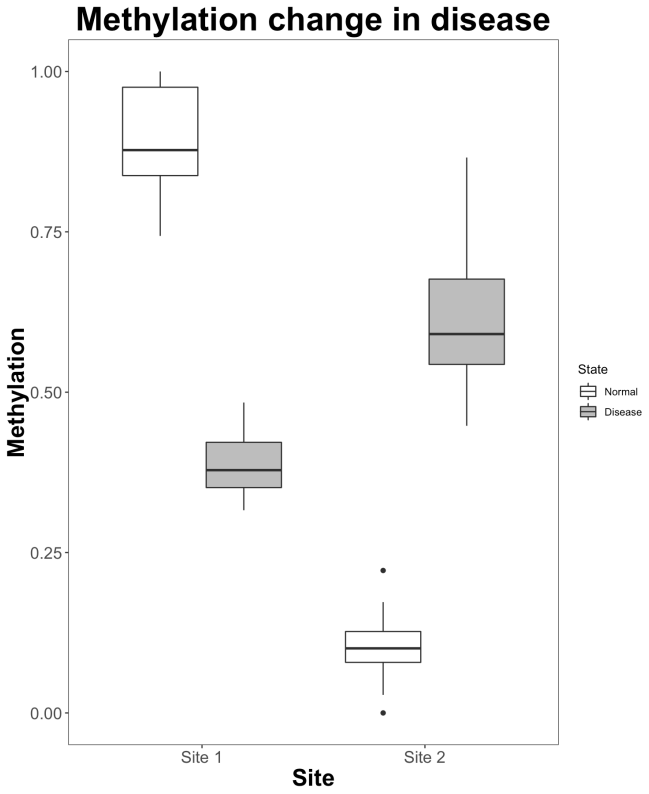
Batch as a confounder

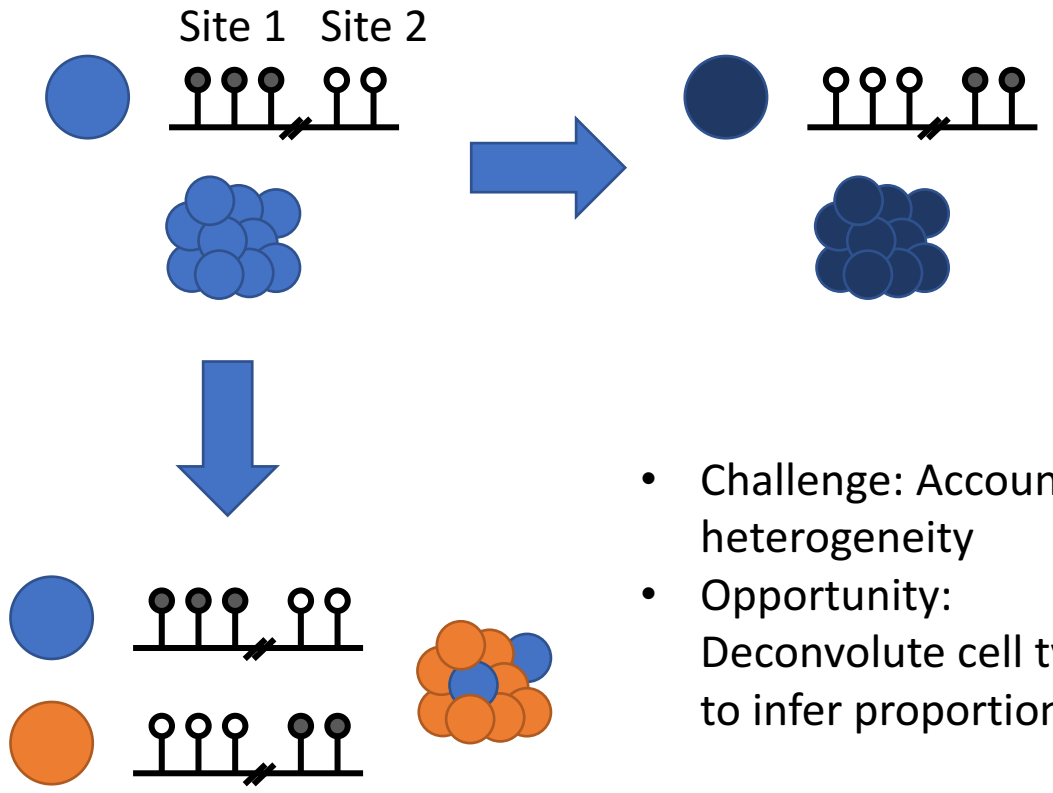


Hicks, S. C., Teng, M., & Irizarry, R. A. (2015). On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data. *BioRxiv*, *25528*

Heterogeneity

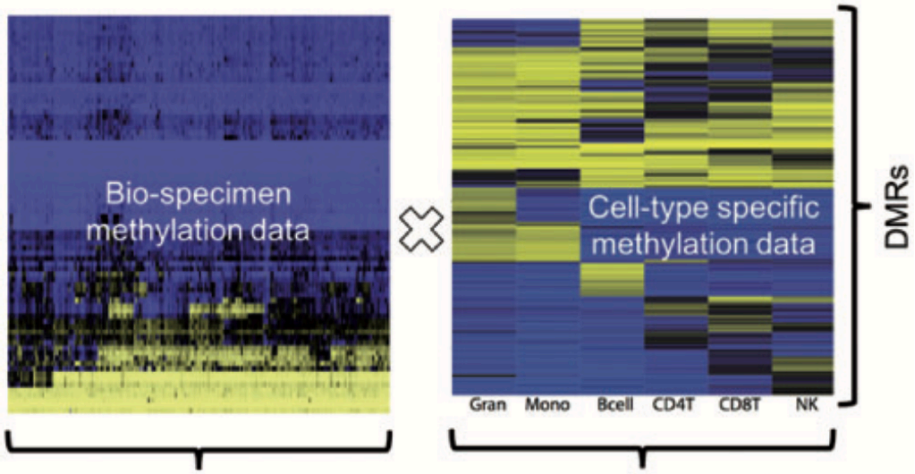
- Cell methylation exists in 3 states:
  - Unmethylated
  - Hemi-methylated
  - Methylated
  - NOTE: Measurement includes a 4th state; missing data
- How are we measuring methylation as a continuous value from 0 – 100%?
- We are measuring a heterogeneous admixture of cells





- Challenge: Account for heterogeneity
- Opportunity: Deconvolute cell types to infer proportion

1. **Reference-based cell type deconvolution using immune cell DMRs (e.g. Houseman method)**



Result: a matrix of samples with estimated immune cell type proportions

Samples 1…n

Gran	Mono	B-cell	CD4T	CD8T	NK
$Gran_1$	$Mono_1$	$B\text{-}cell_1$	$CD4T_1$	$CD8T_1$	NK_1
$Gran_2$	$Mono_2$	$B\text{-}cell_2$	$CD4T_2$	$CD8T_2$	NK_2
…	…	…	…	…	…
$Gran_n$	$Mono_n$	$B\text{-}cell_n$	$CD4T_n$	$CD8T_n$	NK_n

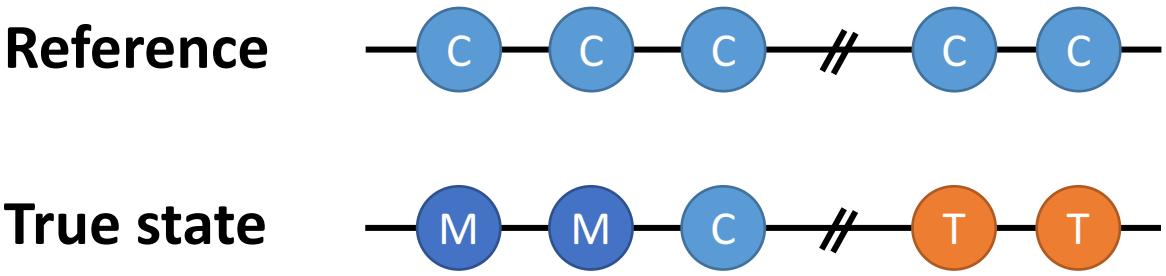
Immune proportion estimates for samples 1…n

Titus, Alexander J., et al. "Cell-type deconvolution from DNA methylation: a review of recent applications." *Human molecular genetics* 26.R2 (2017): R216-R224

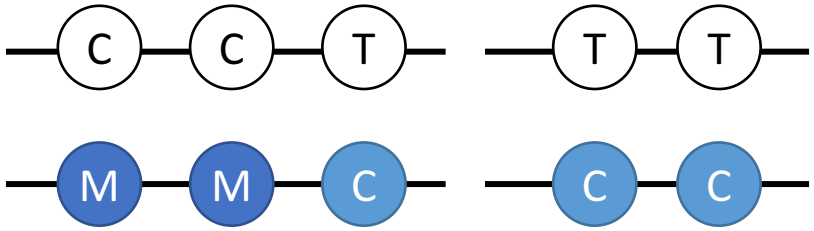
Emerging technologies in epigenomics.

Nanopore sequencing

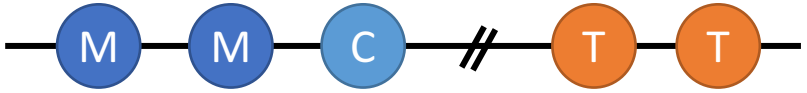
Allows for the detection of 5mC and 5hmC without bisulfite treatment and distinguishes T's from unmethylated C's.

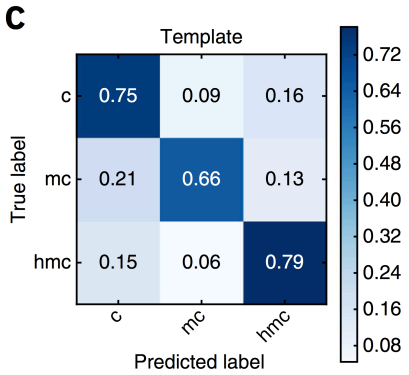


Bisulfite sequencing



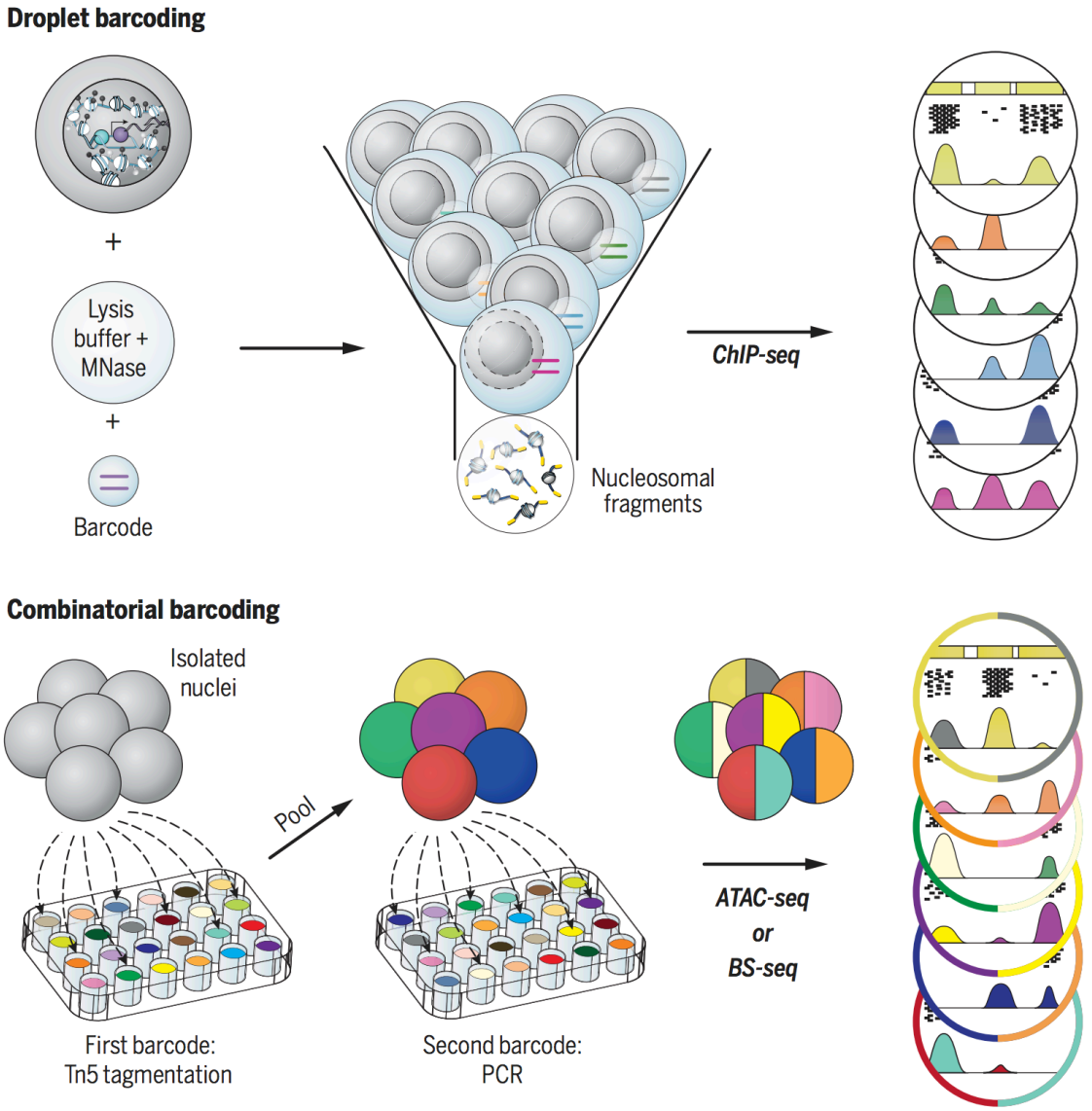
Nanopore sequencing



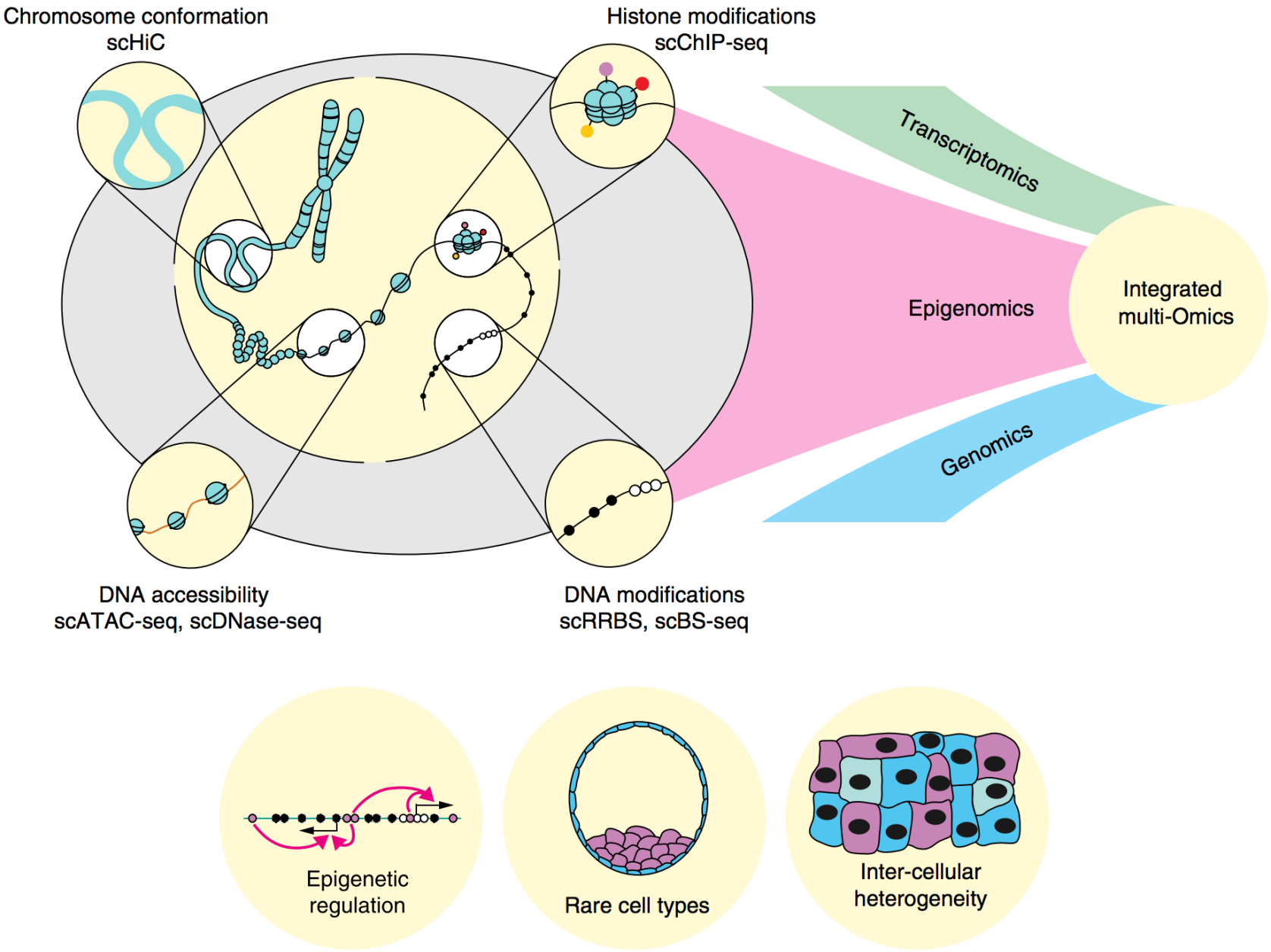


Rand, Arthur C., et al. "Mapping DNA methylation with high-throughput nanopore sequencing." *Nature methods* 14.4 (2017): 411.

Single cell epigenomics



Kelsey, G., Stegle, O., & Reik, W. (2017). Single-cell epigenomics: Recording the past and predicting the future. *Science*, *358*(6359), 69-75

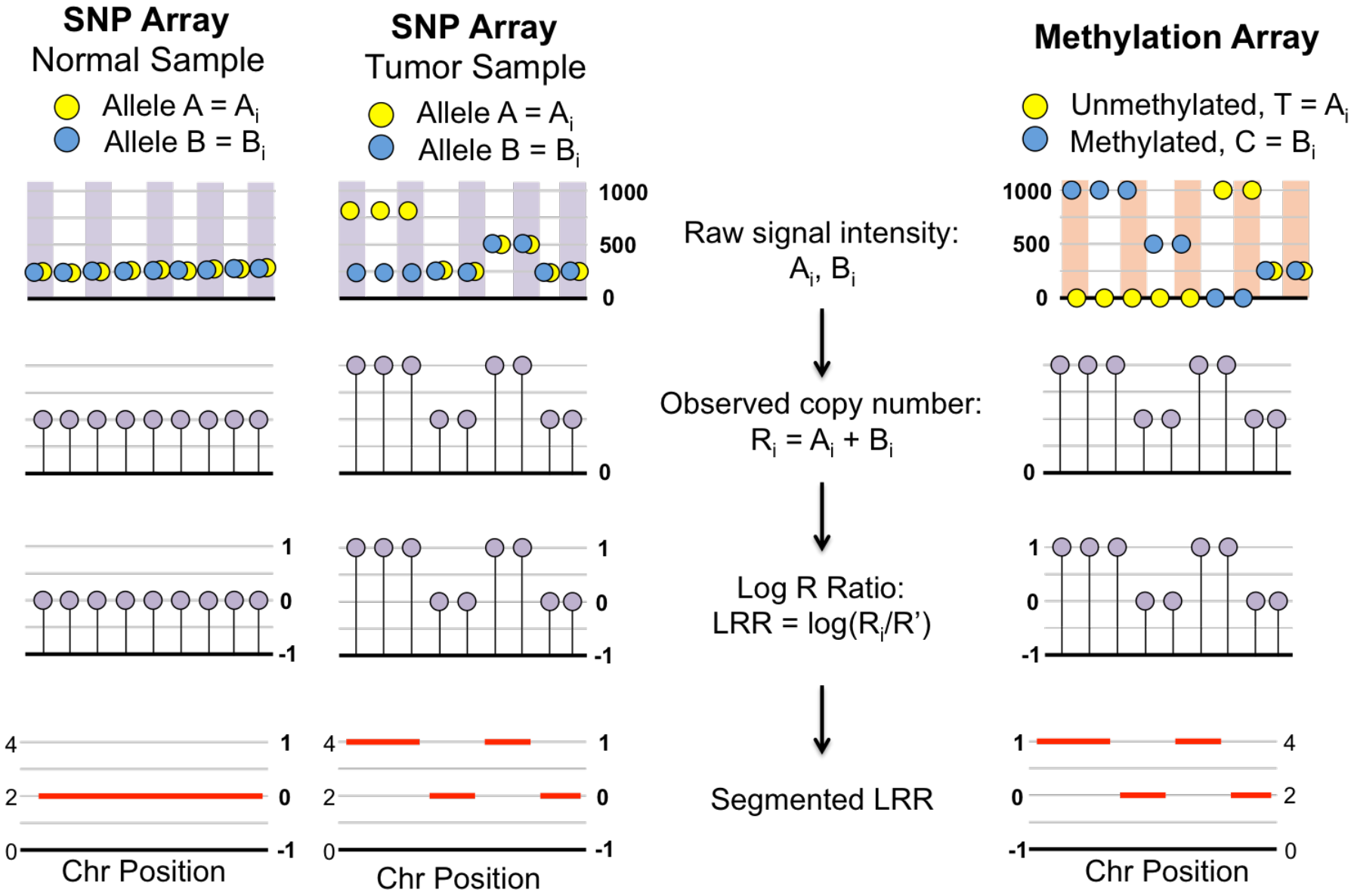


Clark, Stephen J., et al. "Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity." *Genome biology* 17.1 (2016): 72

Thank you.

QUESTIONS?

APPENDIX



Cho et al. *Under review.*