DNA methylation

study design, bench work, and statistical approaches

Amanda Lea

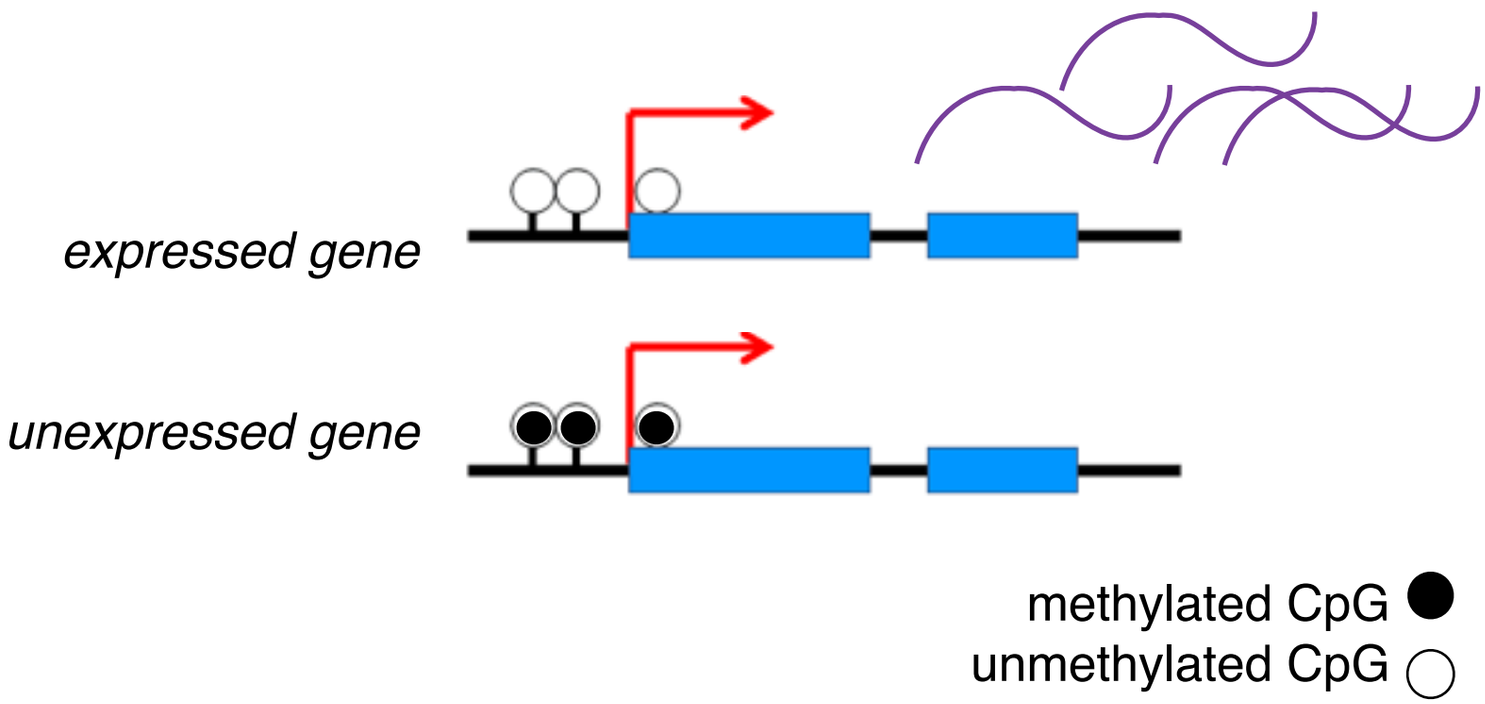
September 22, 2016

Some things we'll talk about:

- What is DNA methylation? Why do we care about it?
- How can we measure genome-wide DNA methylation (in a cost effective way, in non-model organisms)?
- What do (sequencing-based) DNA methylation data look like? What are some particular challenges in working with these data?
- What models have been proposed to analyze differential methylation (and what are their relative shortcomings)?

DNA methylation

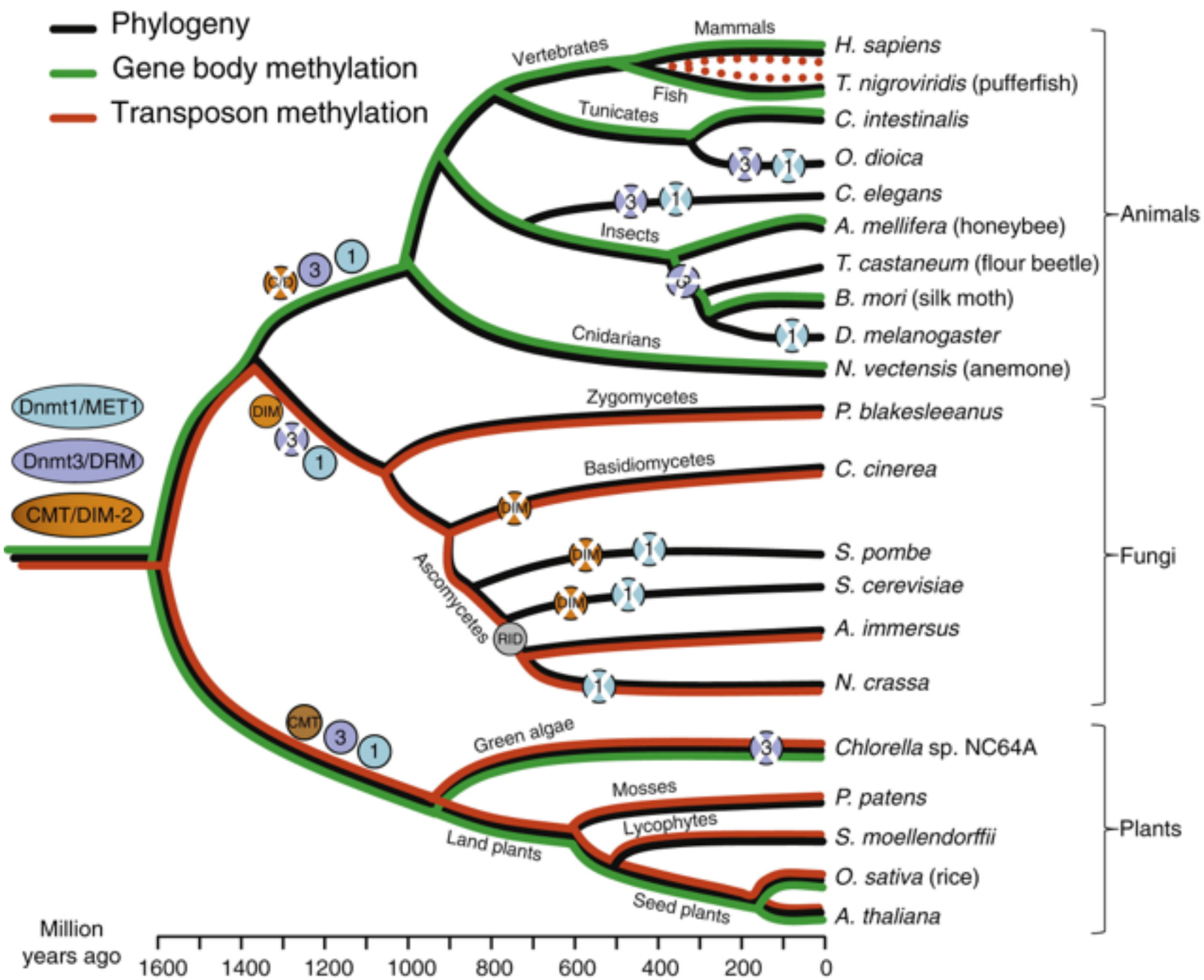
- one of many epigenetic gene regulatory mechanisms
- covalent addition of methyl groups to cytosine bases
- occurs at ‘CpG sites’ in vertebrates
- enriched in/near genes and regulatory regions
- key functions = X inactivation, suppression of transposable elements, imprinted genes



To review, differences in DNA methylation are associated with:

To review, differences in DNA methylation are associated with:

- evolutionary history, species differences



Rematch & Zilberman 2010, Current Biology

To review, differences in DNA methylation are associated with:

- evolutionary history, species differences
- environmental exposures

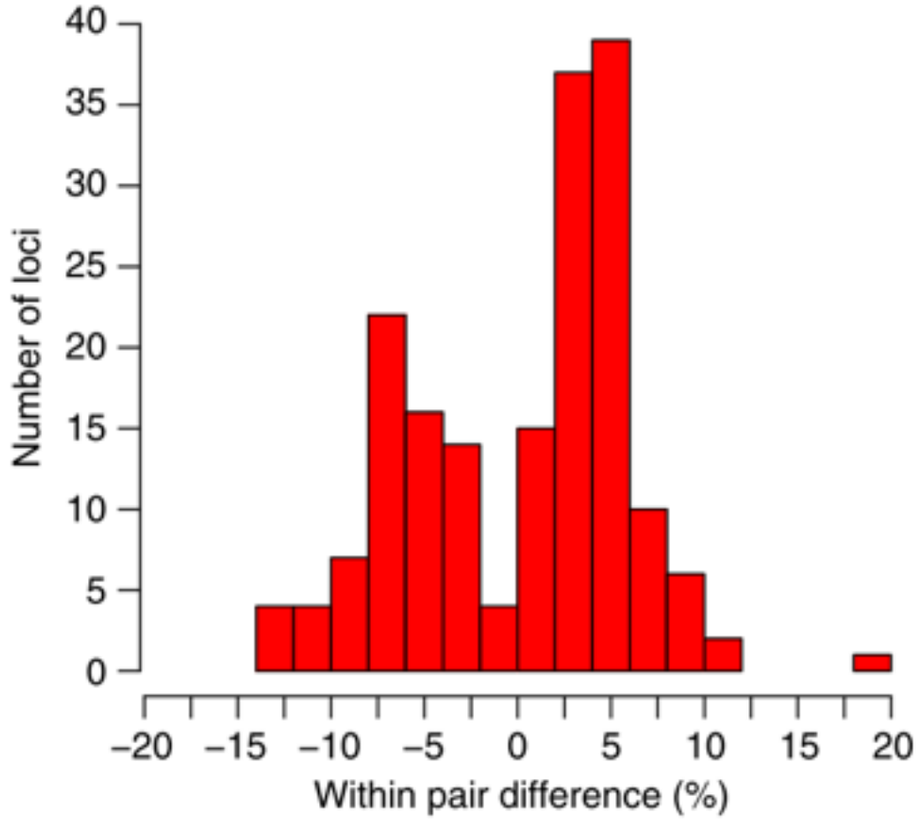
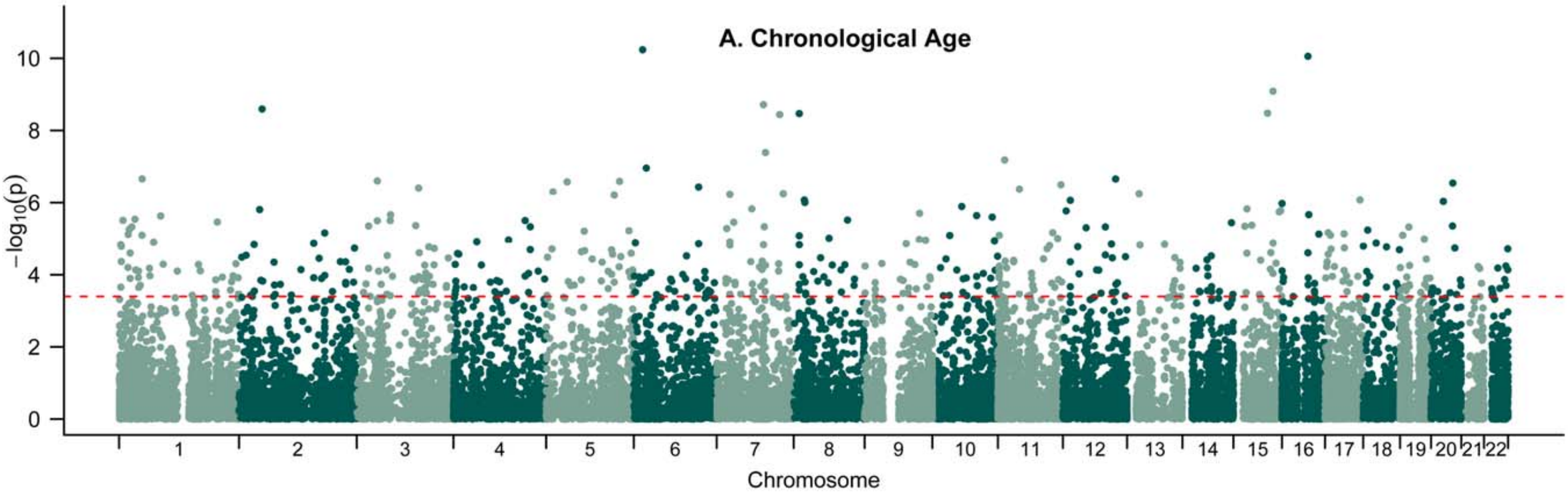


Figure 2 | The average within-pair difference for the 181 regions associated with prenatal famine exposure after correction for multiple testing. A histogram for the average within pair difference (%) between the exposed and unexposed siblings. A positive number reflects relative higher DNA methylation levels in the exposed.

Toni et al. 2014, Nature Communications

To review, differences in DNA methylation are associated with:

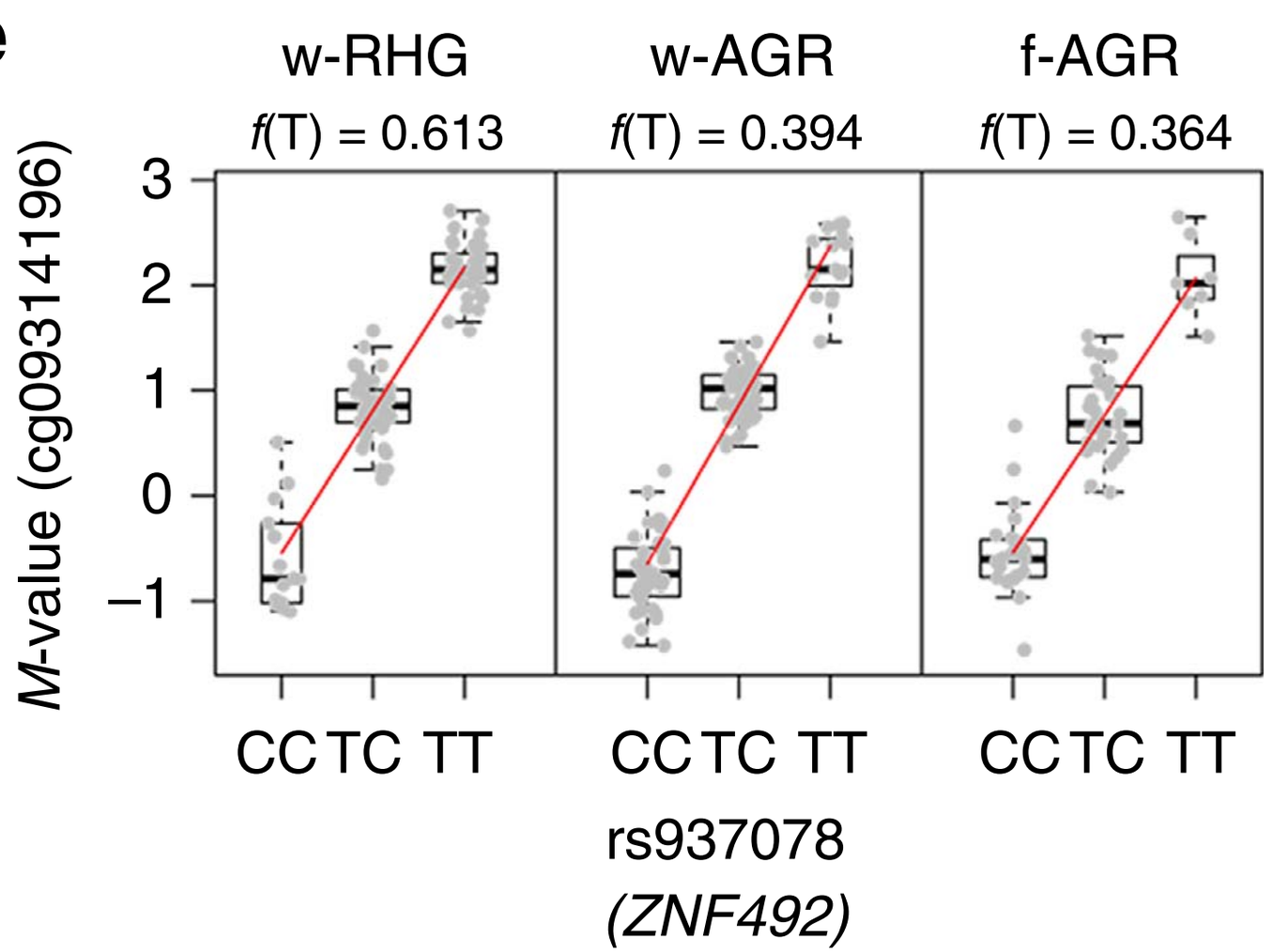
- evolutionary history, species differences
- environmental exposures
- age (also life history or developmental stage)



Bell et al. 2012, PLoS Genetics

To review, differences in DNA methylation are associated with:

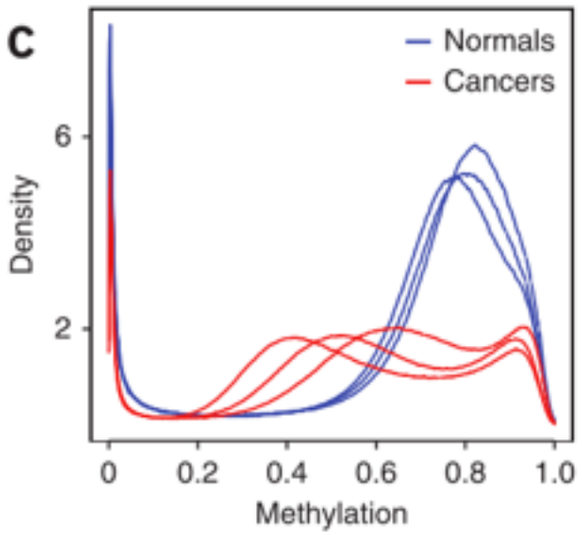
- evolutionary history, species differences
- environmental exposures
- age (also life history or developmental stage)
- genetic variation

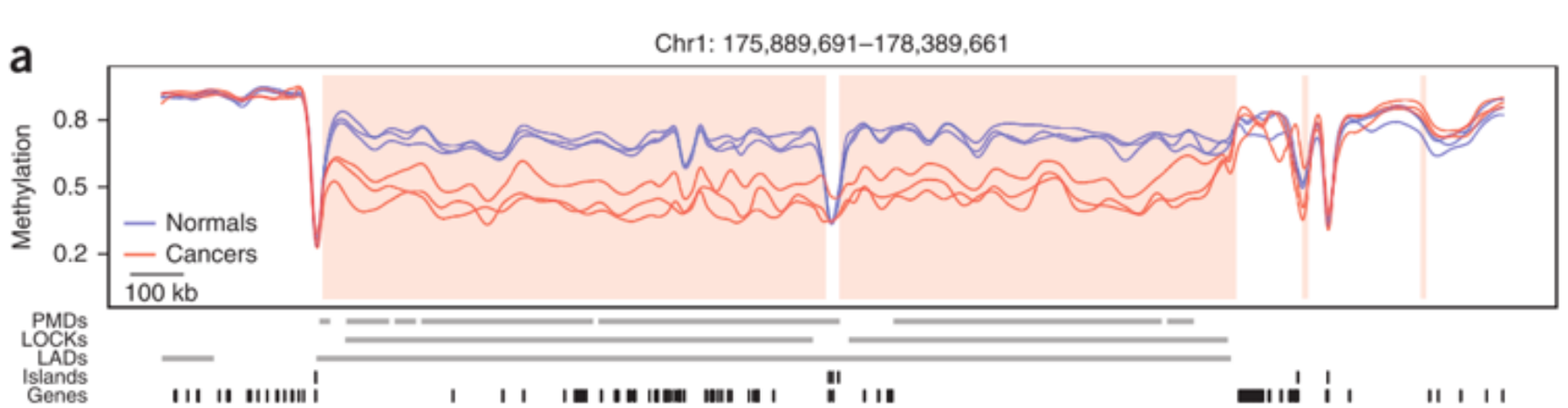


Fagny et al. 2015, Nature Communications

To review, differences in DNA methylation are associated with:

- evolutionary history, species differences
- environmental exposures
- age (also life history or developmental stage)
- genetic variation
- disease status

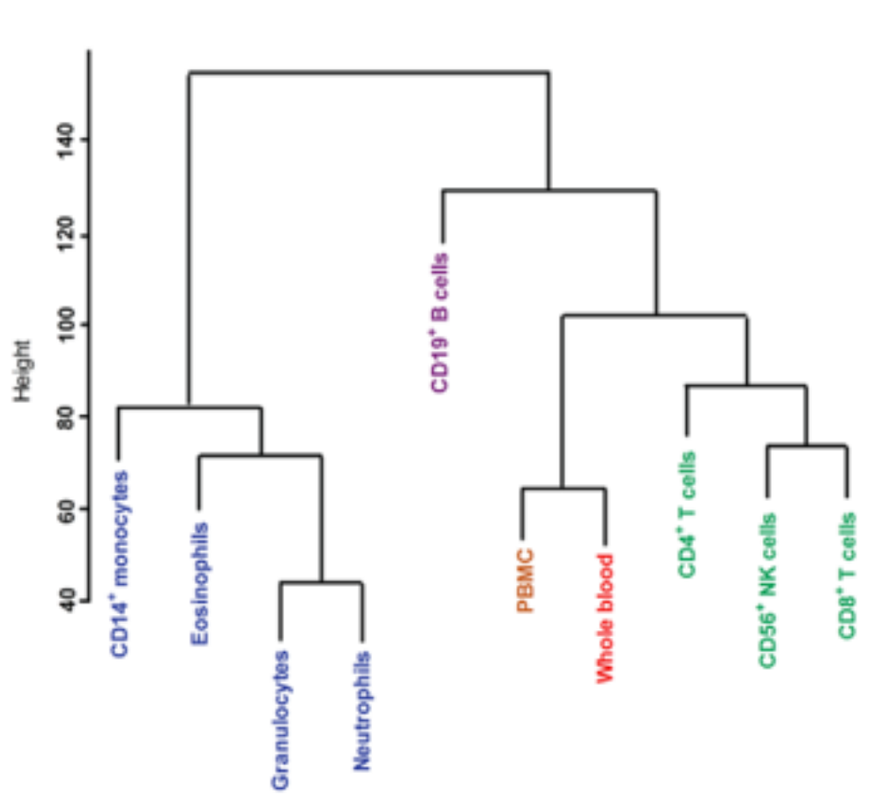


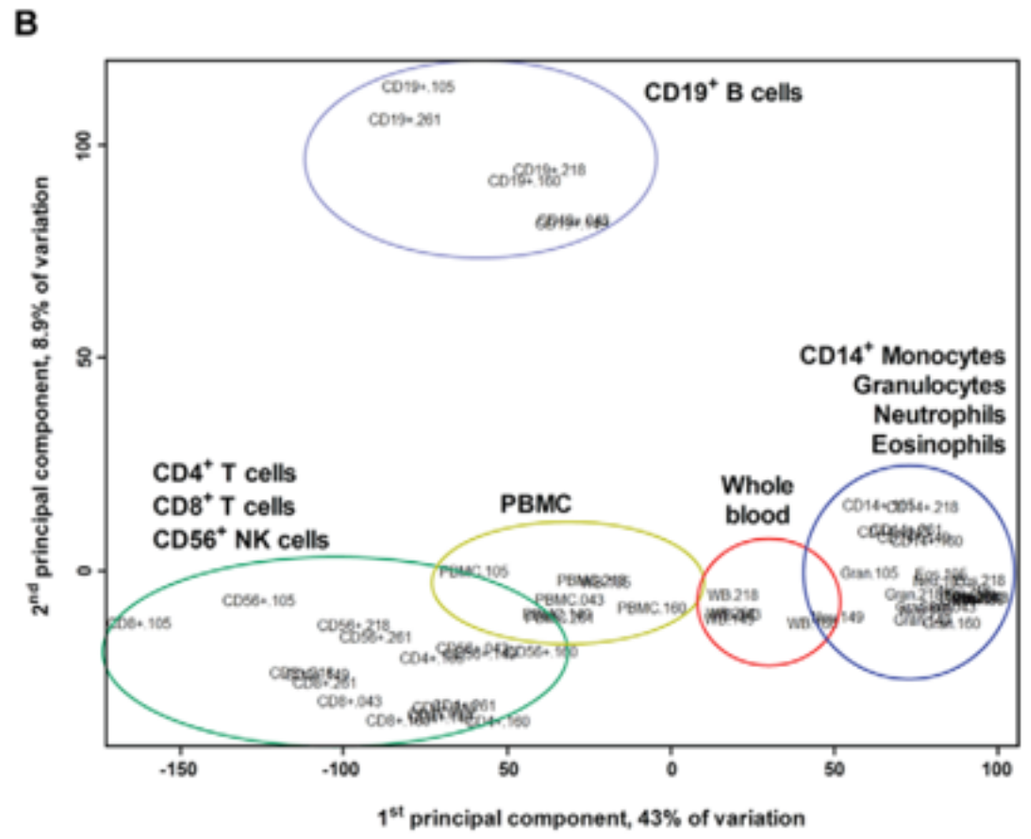


Hansen et al. 2011, Nature Genetics

To review, differences in DNA methylation are associated with:

- evolutionary history, species differences
- environmental exposures
- age (also life history or developmental stage)
- genetic variation
- disease status
- cell type identity/composition





Reinius et al. 2012, PLoS One

To review, differences in DNA methylation are associated with:

- evolutionary history, species differences
- environmental exposures
- age (also life history or developmental stage)
- genetic variation
- disease status
- cell type identity/composition

Interesting!

Also, important to account for these factors in your study!

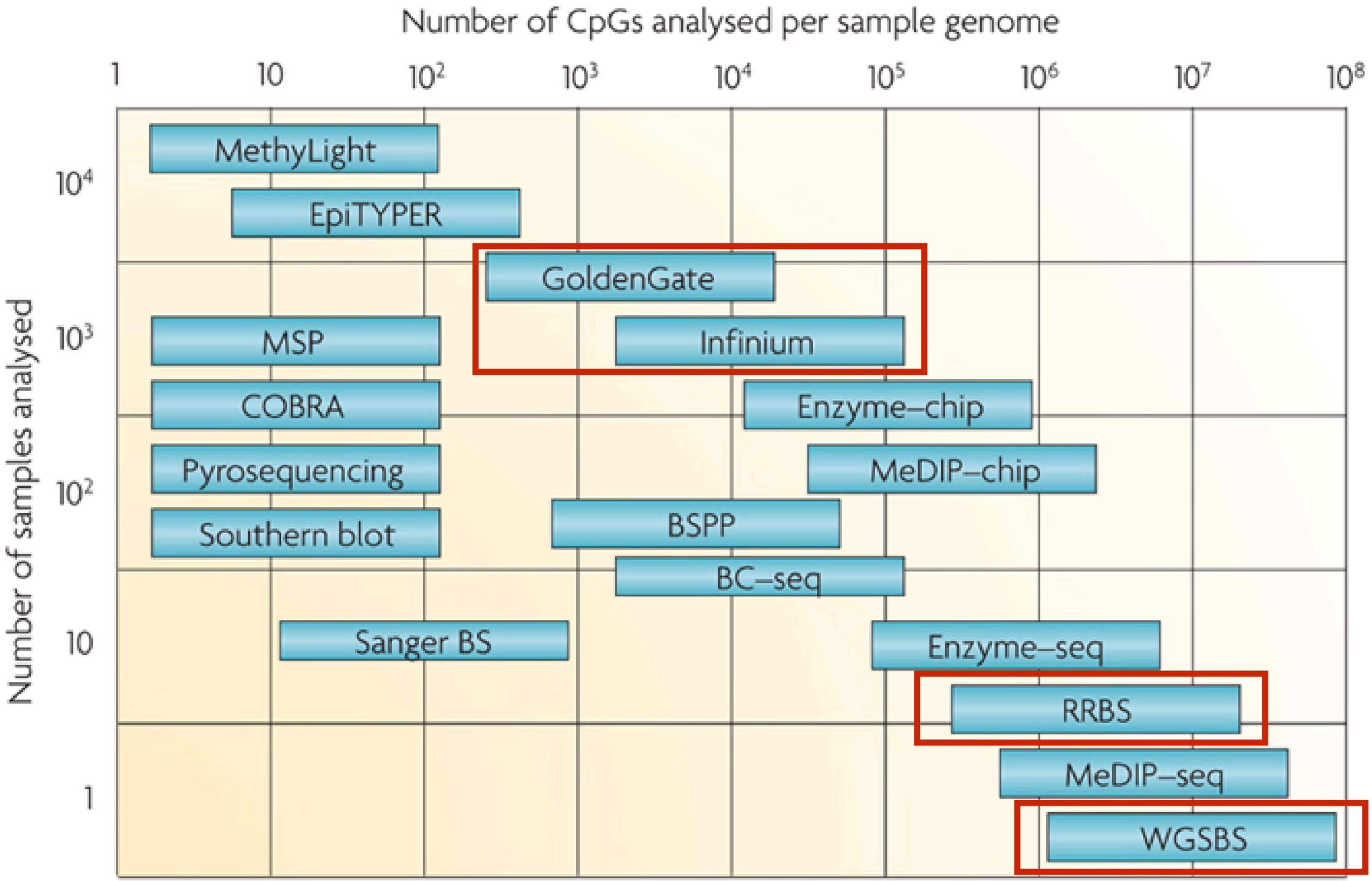
We're talking about this kind of DNA methylation study:

1. Get biological samples
2. Extract DNA
3. Measure DNA methylation levels at many loci using a high-throughput approach
4. Test for associations between predictor of interest and DNA methylation levels at each measured site (controlling for other stuff)
5. Correct for multiple hypothesis testing (e.g., FDR)
6. Make some inference based on which genes/regions exhibit differential methylation

We're talking about this kind of DNA methylation study:

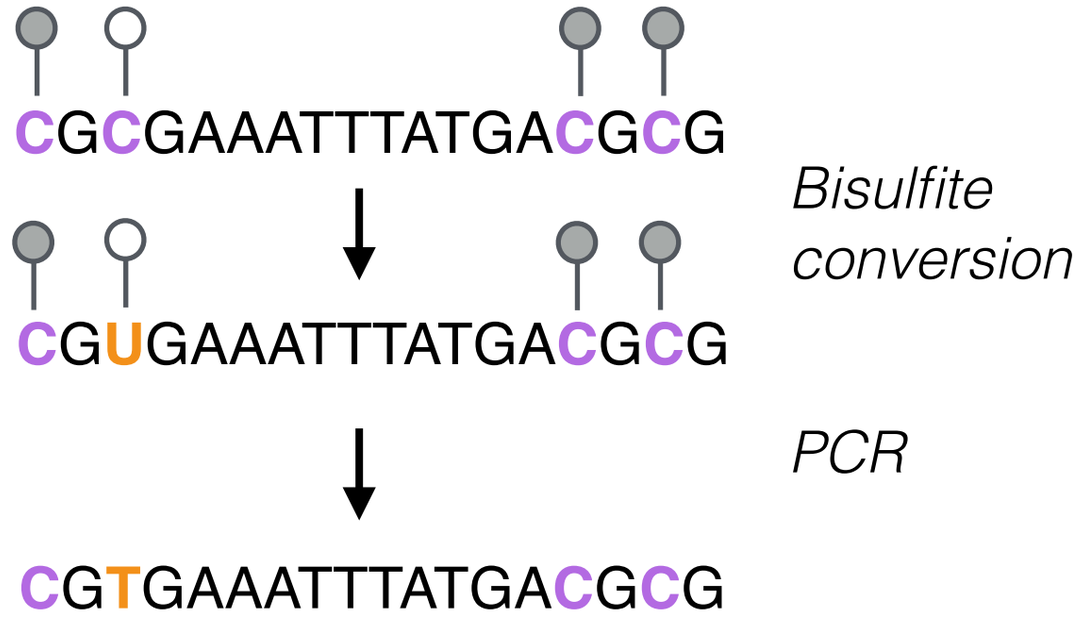
1. Get biological samples
2. Extract DNA
3. Measure DNA methylation levels at many loci using a high-throughput approach
4. Test for associations between predictor of interest and DNA methylation levels at each measured site (controlling for other stuff)
5. Correct for multiple hypothesis testing (e.g., FDR)
6. Make some inference based on which genes/regions exhibit differential methylation

Ways to measure DNA methylation levels

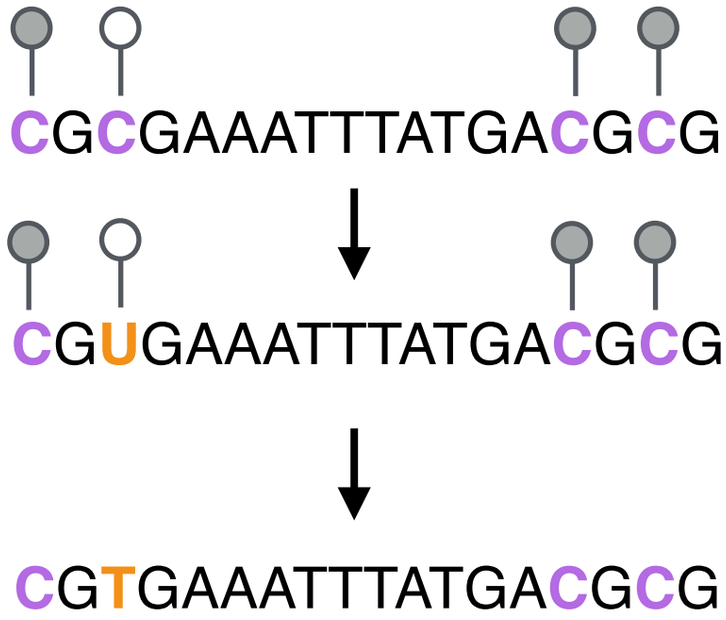


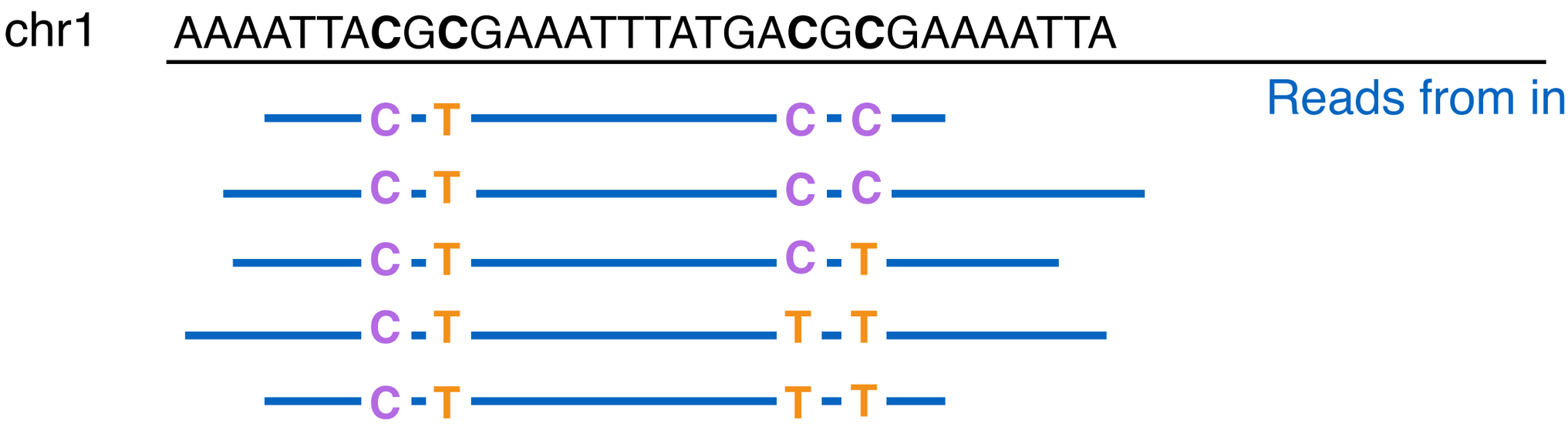
Laird et al. 2010, Nature Reviews Genetics

Bisulfite sequencing

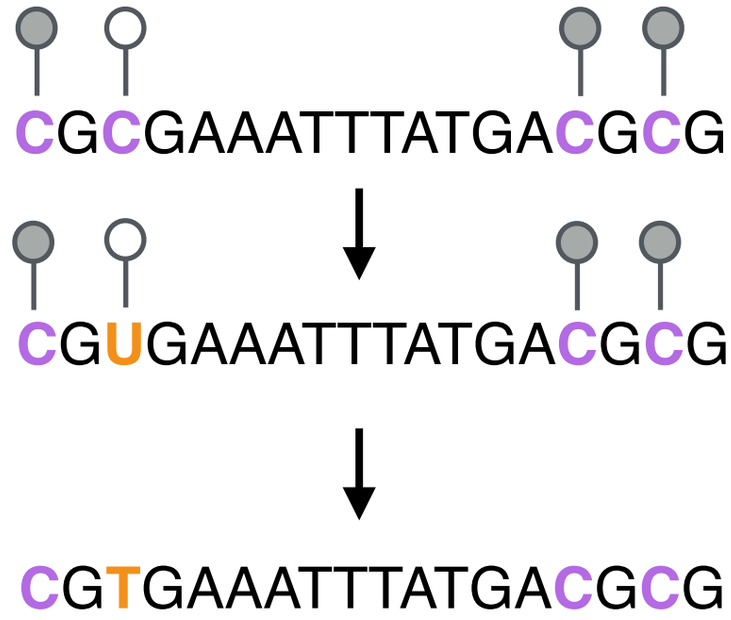


Bisulfite sequencing





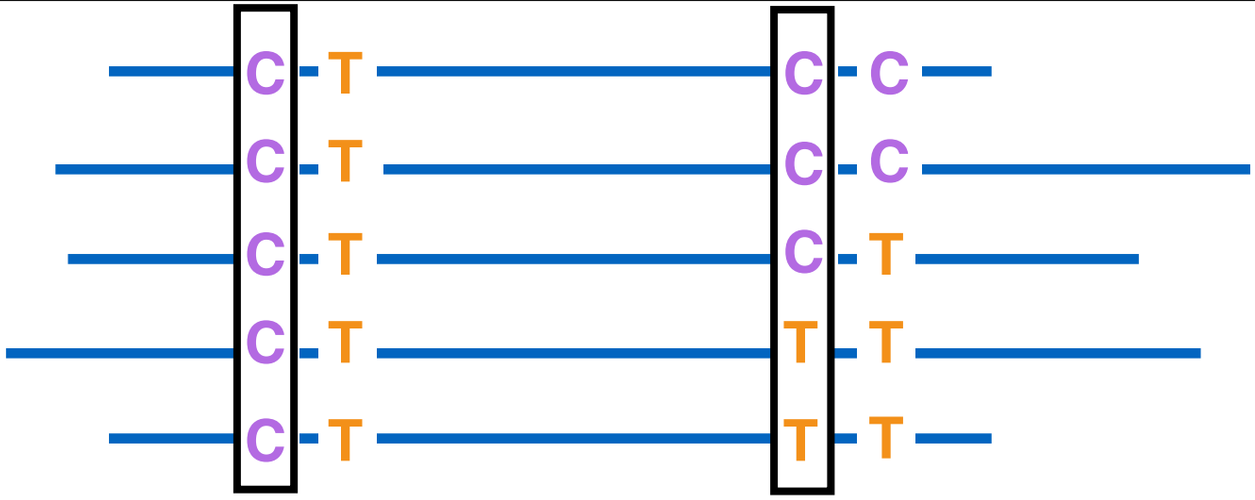
Bisulfite sequencing



Bisulfite conversion

PCR

chr1 AAAATTA**C**G**C**GAAATTTATGA**C**G**C**GAAAATTA



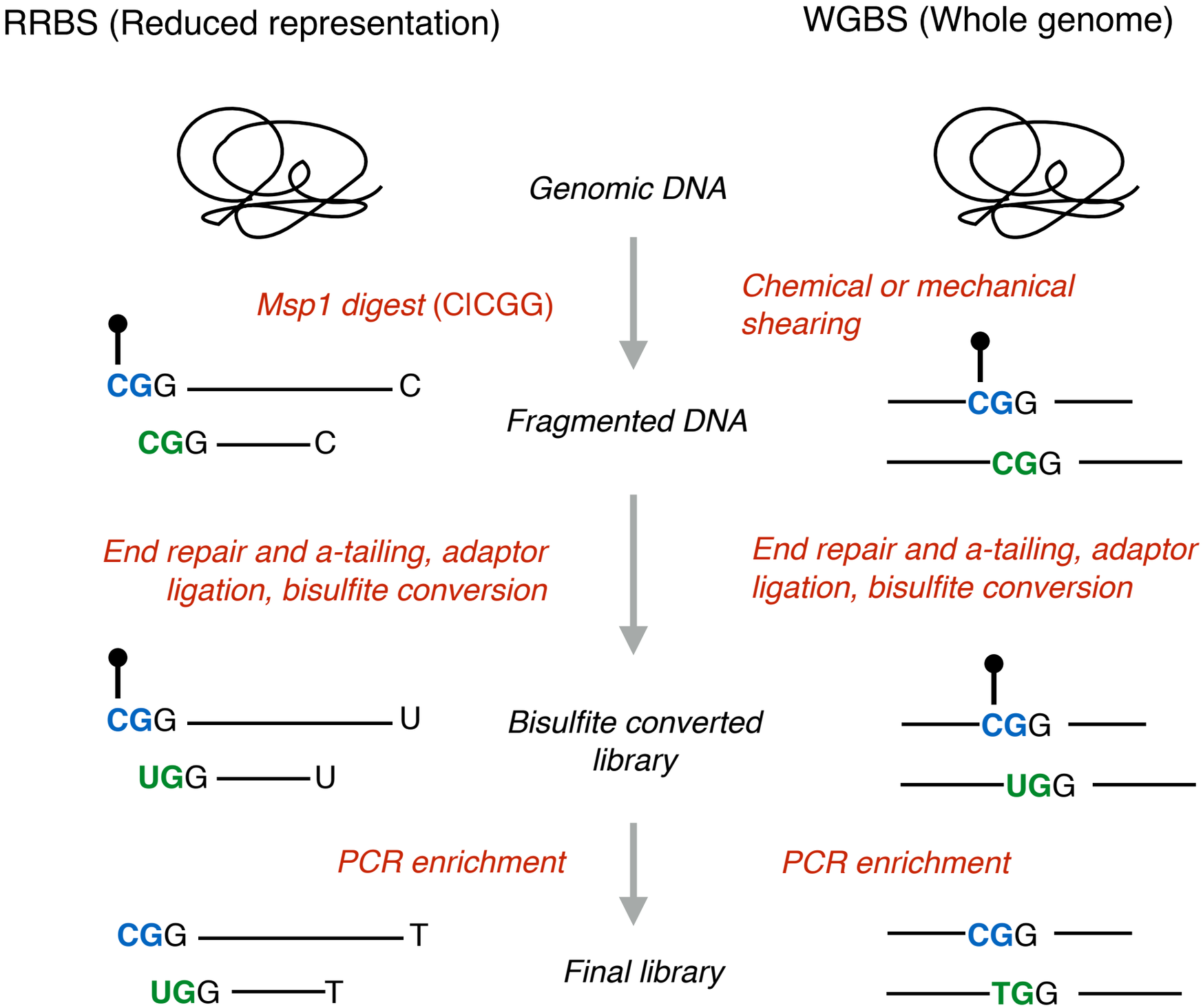
Reads from individual 1

5/5 = 100% methylated

3/5 = 60% methylated

DNA methylation level = C / (C+T)

Bisulfite sequencing



Bisulfite sequencing

RRBS (Reduced representation)

a few mil CpG sites

for a primate-sized genome, cost <$150

need <200ng of DNA

WGBS (Whole genome)

~28 mil CpG sites

Genomic DNA

Msp1 digest (CICGG)

Chemical or mechanical shearing

CGG C

CGG C

Fragmented DNA

CGG C

CGG C

End repair and a-tailing, adaptor ligation, bisulfite conversion

End repair and a-tailing, adaptor ligation, bisulfite conversion

CGG U

UGG U

Bisulfite converted library

CGG U

UGG U

PCR enrichment

PCR enrichment

CGG T

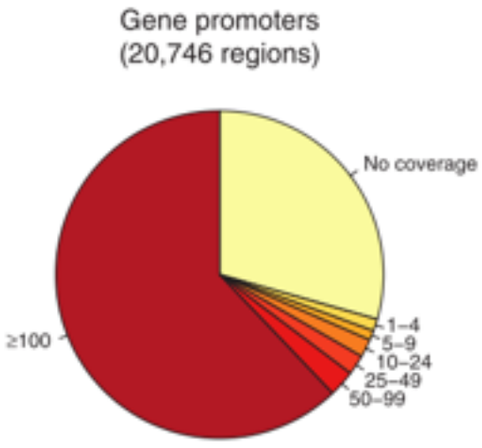
UGG T

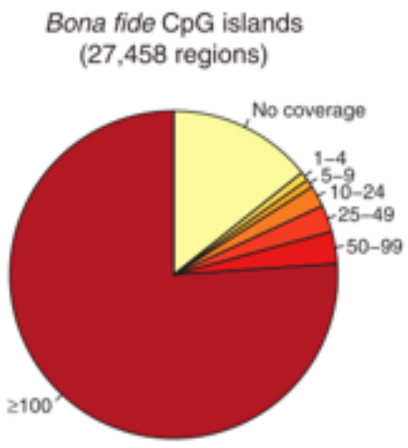
Final library

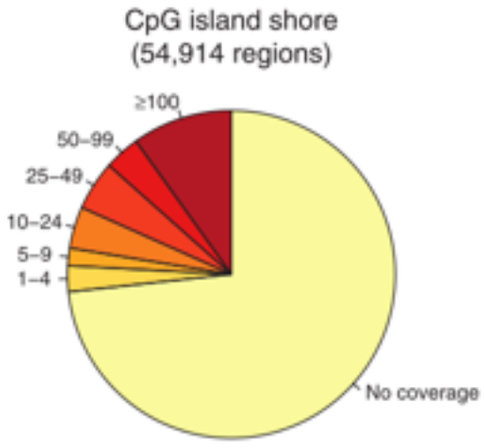
CGG T

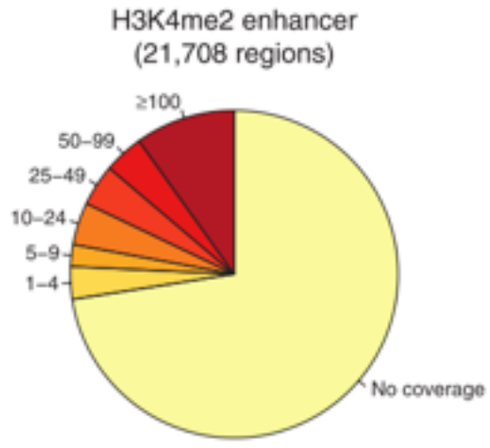
TGG T

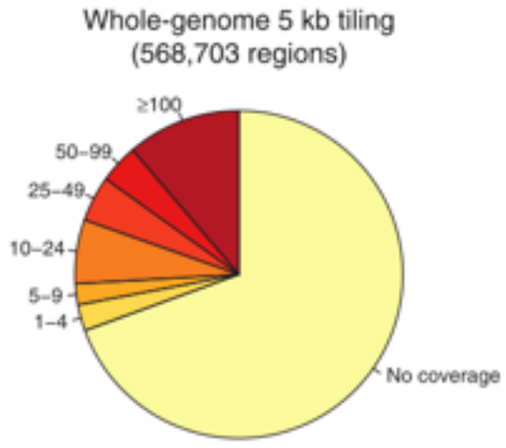
RRBS enriches for functionally important regions of the genome





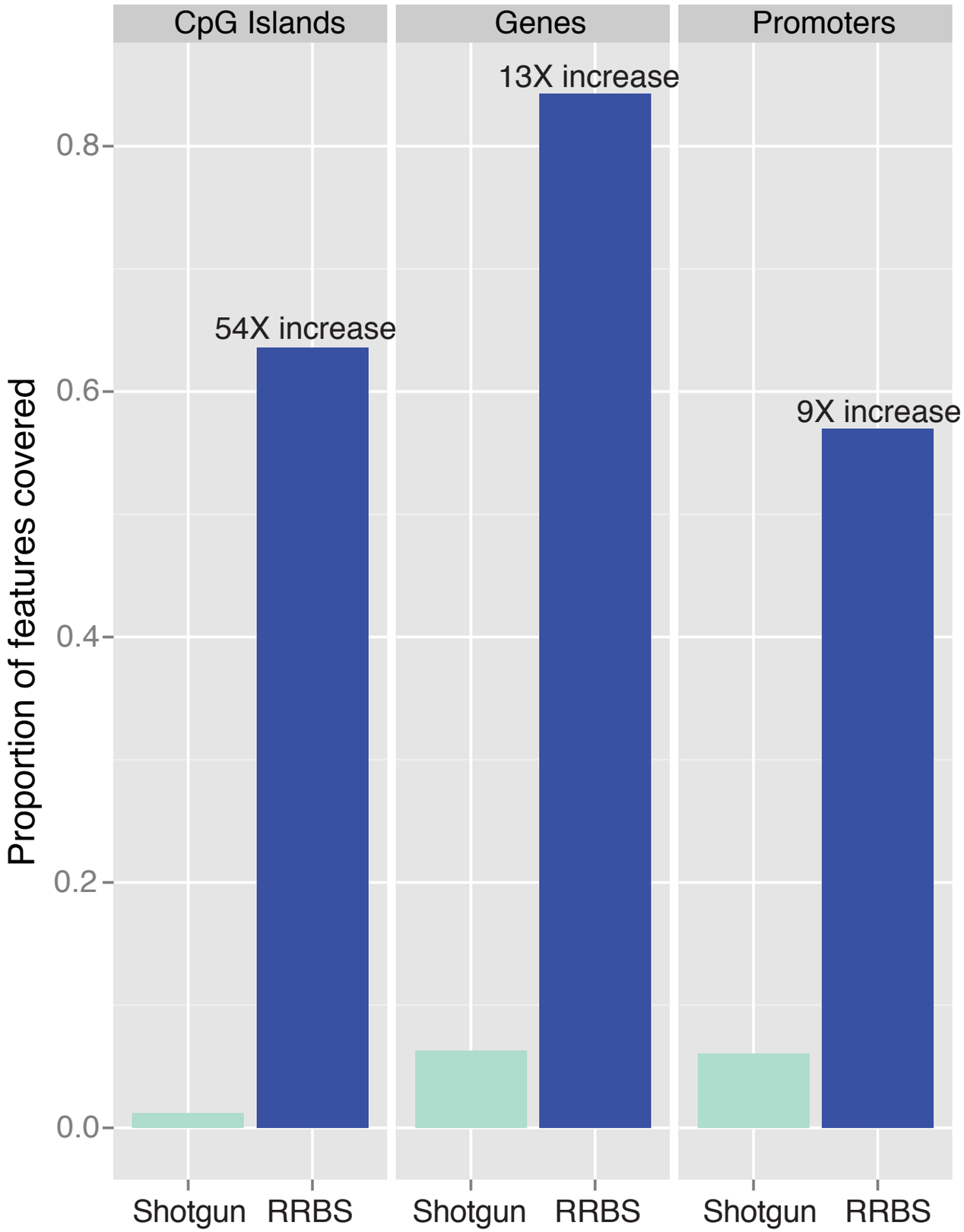






Gu et al. 2011, Nature Protocols

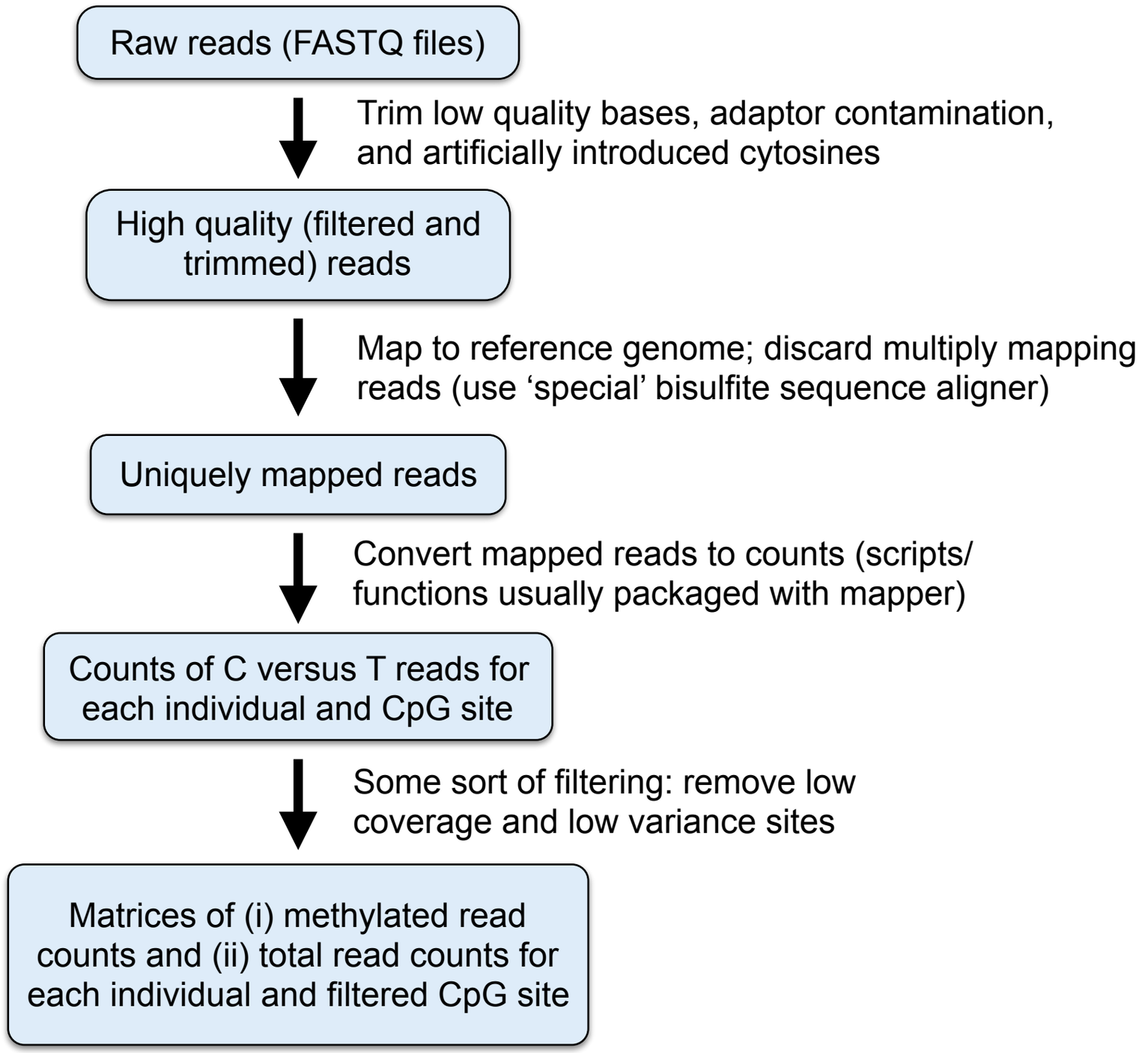
RRBS enriches for functionally important regions of the genome



Bisulfite sequencing data

Bisulfite sequencing data

- First - process, map, and filter your data



Note, there are special mappers for bisulfite sequencing data

BMC Bioinformatics

BioMed Central

Methodology article

Open Access

BSMAP: whole genome bisulfite sequence MAPping program

Yuanxin Xi and Wei Li*

BIOINFORMATICS **APPLICATIONS NOTE**

Vol. 27 no. 11 2011, pages 1571–1572
doi:10.1093/bioinformatics/btr167

Sequence analysis

Advance Access publication April 14, 2011

Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications

Felix Krueger* and Simon R. Andrews

Cell Reports

Resource

Differential DNA Methylation Analysis without a Reference Genome

Johanna Klughammer,[1] Paul Datlinger,[1] Dieter Printz,[2] Nathan C. Sheffield,[1] Matthias Farlik,[1] Johanna Hadler,[1] Gerhard Fritsch,[2] and Christoph Bock[1,3,4]

SOFTWARE

Open Access

BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data

Weilong Guo[1,2], Petko Fiziev[3], Weihong Yan[4], Shawn Cokus[2], Xueguang Sun[5], Michael Q Zhang[1,6], Pao-Yang Chen[7*] and Matteo Pellegrini[2,8*]

Bisulfite sequencing data

- Now you have a count matrix!

Data matrix for *n* individuals measured at *q* loci
(methylated reads / total reads)

	ID_1	ID_2	…	ID_n
loc_1	2/7	20/70	…	40/140
loc_2	2/3	40/60	…	40/140
loc_3	5/5	50/50		100/100
…	…	…	…	…
loc_q	0/0	1/3	…	4/6

Bisulfite sequencing data

Data matrix for *n* individuals measured at *q* loci
(methylated reads / total reads)

	ID_1	ID_2	…	ID_n
loc_1	2/7	20/70	…	40/140
loc_2	2/3	40/60	…	40/140
loc_3	5/5	50/50		100/100
…	…	…	…	…
loc_q	0/0	1/3	…	4/6

Biological variable = a proportion or level (ranging from 0-1)

Data = counts

Bisulfite sequencing data

Data matrix for *n* individuals measured at *q* loci
(methylated reads / total reads)

	ID_1	ID_2	…	ID_n
loc_1	2/7	20/70	…	40/140
loc_2	2/3	40/60	…	40/140
loc_3	5/5	50/50		100/100
…	…	…	…	…
loc_q	0/0	1/3	…	4/6

Missing data

Bisulfite sequencing data

Data matrix for *n* individuals measured at *q* loci
(methylated reads / total reads)

	ID_1	ID_2	…	ID_n
loc_1	2/7	20/70	…	40/140
loc_2	2/3	40/60	…	40/140
loc_3	5/5	50/50		100/100
…	…	…	…	…
loc_q	0/0	1/3	…	4/6

Sites with little variance in methylation levels

Bisulfite sequencing data

Data matrix for *n* individuals measured at *q* loci
(methylated reads / total reads)

	ID_1	ID_2	…	ID_n
loc_1	2/7	20/70	…	40/140
loc_2	2/3	40/60	…	40/140
loc_3	5/5	50/50		100/100
…	…	…	…	…
loc_q	0/0	1/3	…	4/6

Variation in coverage across **individuals** (by orders of magnitude)

Bisulfite sequencing data

Data matrix for *n* individuals measured at *q* loci
(methylated reads / total reads)

	ID_1	ID_2	…	ID_n
loc_1	2/7	20/70	…	40/140
loc_2	2/3	40/60	…	40/140
loc_3	5/5	50/50		100/100
…	…	…	…	…
loc_q	0/0	1/3	…	4/6

Variation in coverage across **sites** (by orders of magnitude)

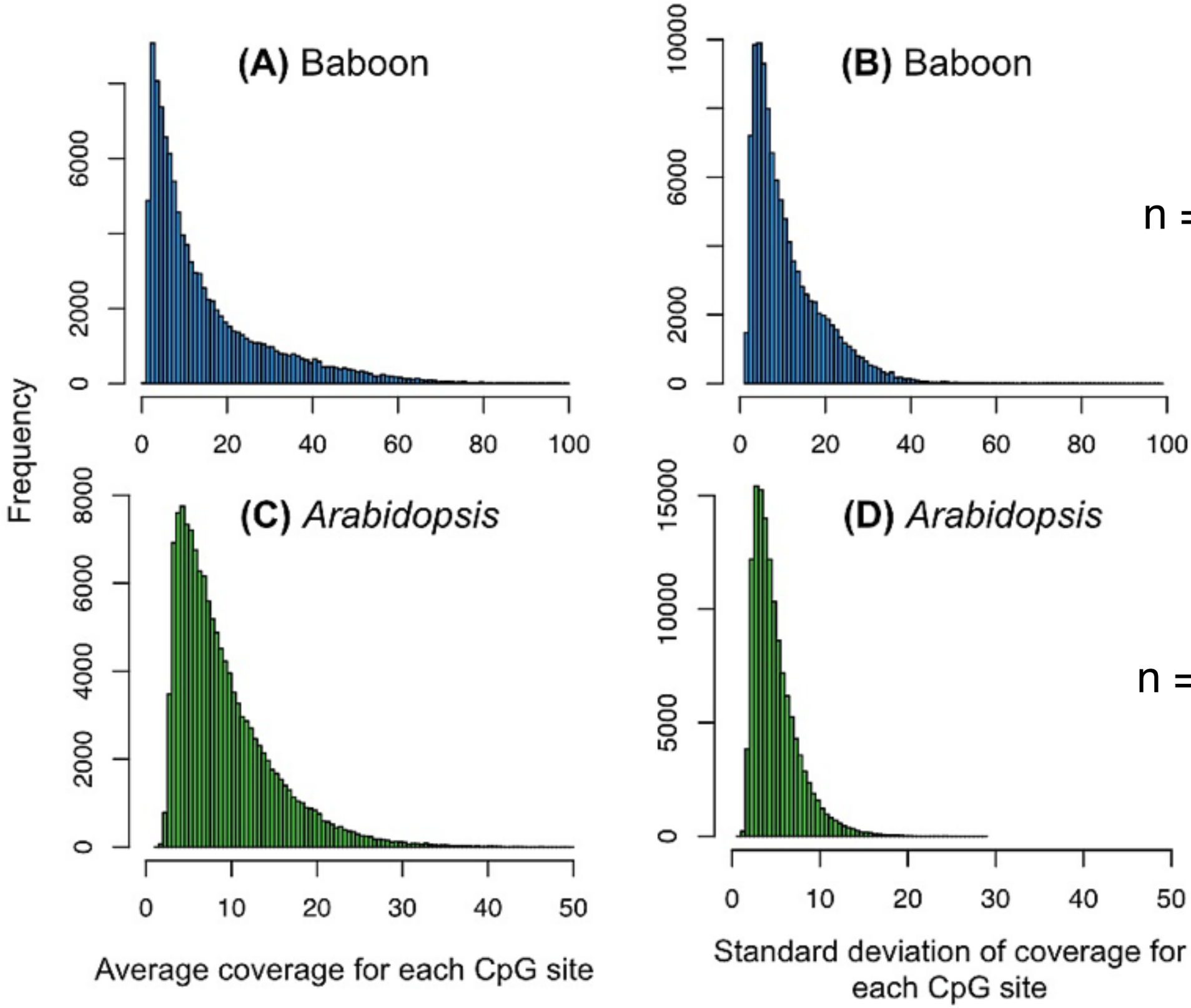
Bisulfite sequencing data

Data matrix for *n* individuals measured at *q* loci
(methylated reads / total reads)

	ID_1	ID_2	…	ID_n
loc_1	2/7	20/70	…	40/140
loc_2	2/3	40/60	…	40/140
loc_3	5/5	50/50		100/100
…	…	…	…	…
loc_q	0/0	1/3	…	4/6

Unlike RNA-seq, all of this total read count variation is technical

Bisulfite sequencing data

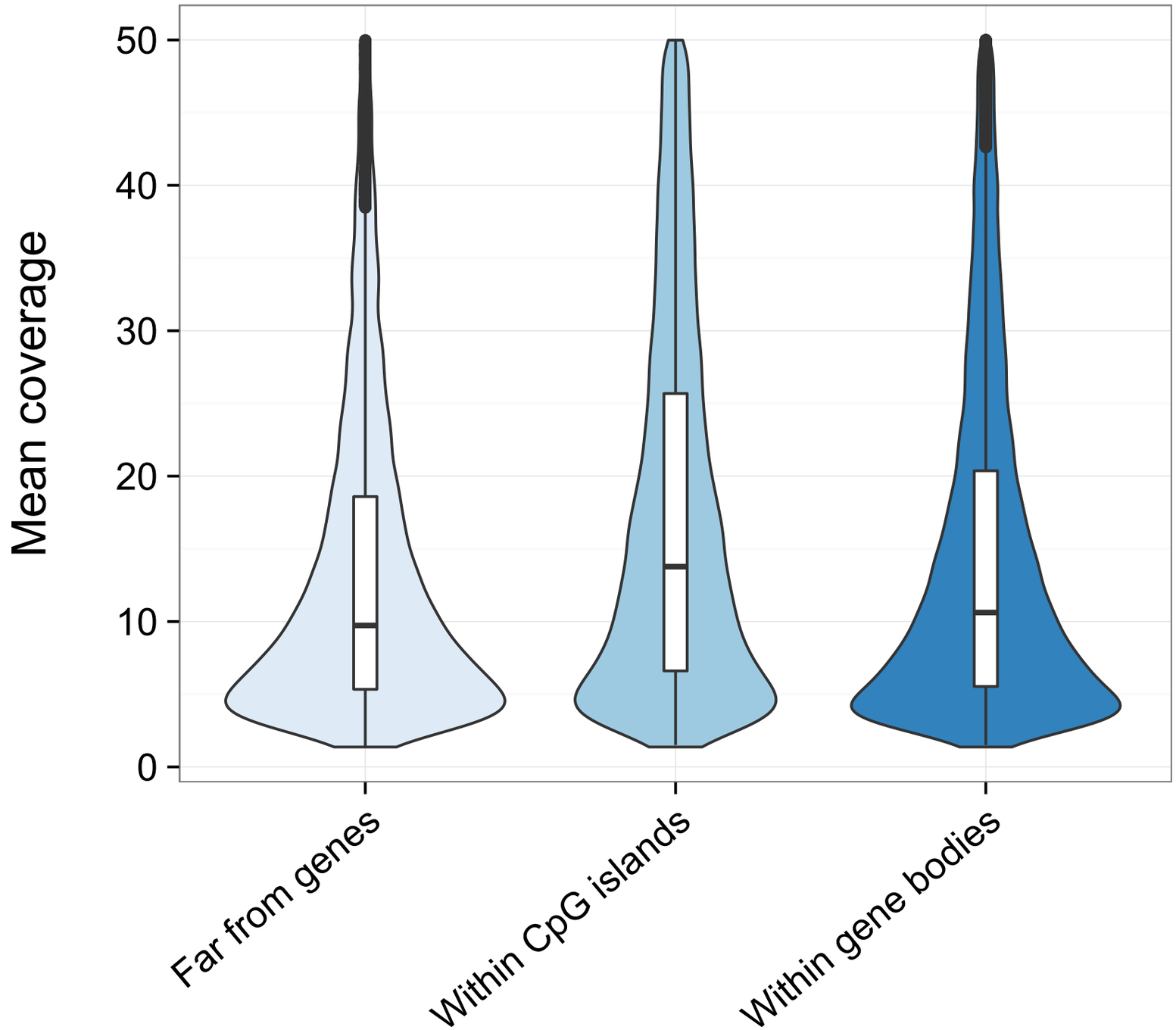


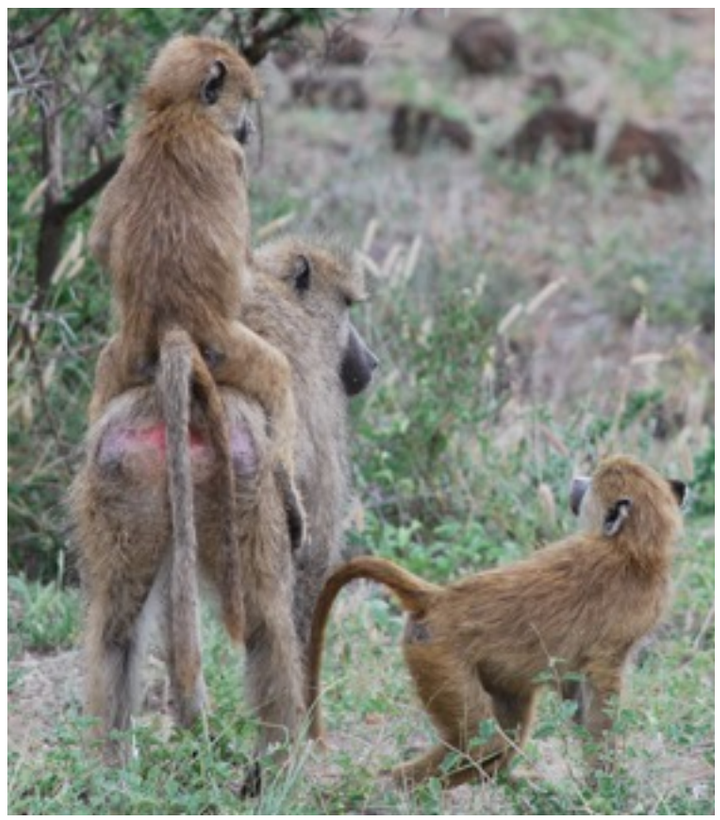
n = 50 (RRBS)

n = 24 (WGBS)

Schmitz et al. 2013, Nature; Lea et al. 2015, PLoS Genetics

Bisulfite sequencing data





Bisulfite sequencing data

	ID_1	ID_2	…	ID_n
loc_1	2/7	20/70	…	40/140
loc_2	2/3	40/60	…	40/140
loc_3	5/5	50/50		100/100
…	…	…	…	…
loc_q	0/0	1/3	…	4/6

transform?

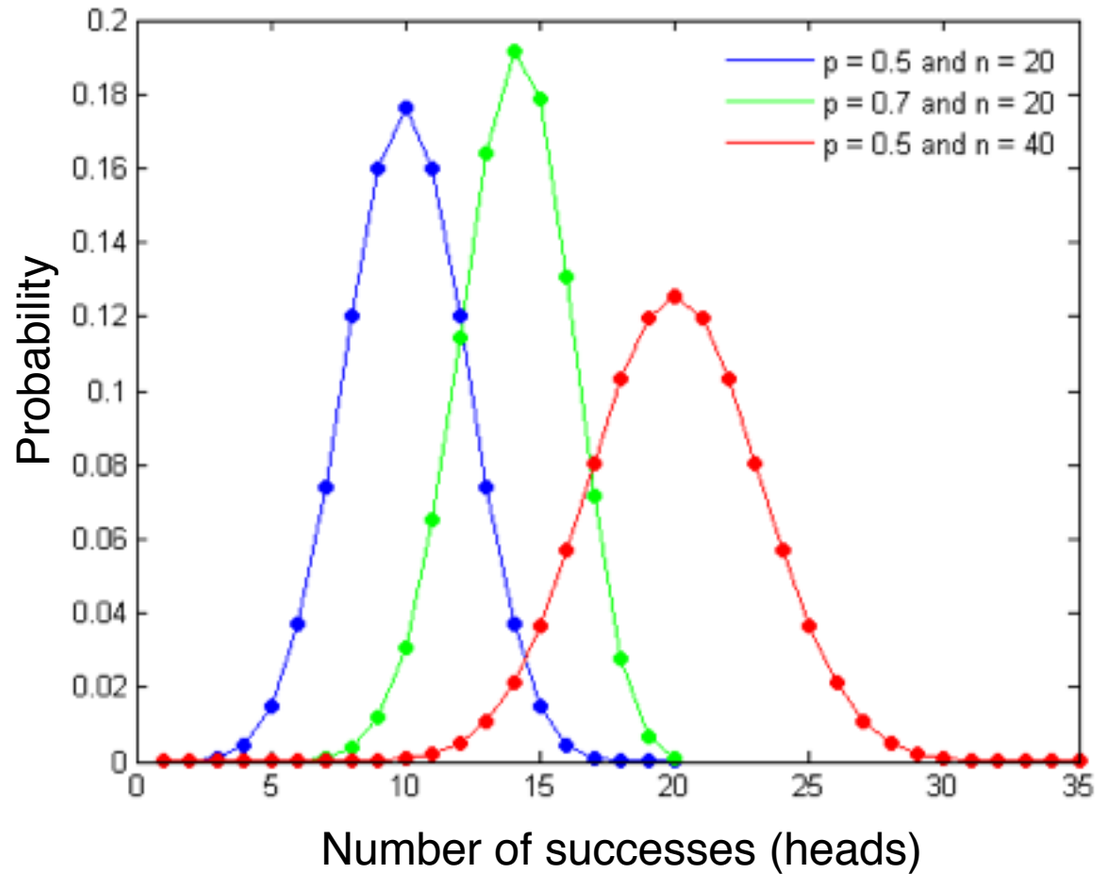
	ID_1	ID_2	…	ID_n
loc_1	0.28	0.28	…	0.29
loc_2	0.67	0.67	…	0.29
loc_3	1	1		1
…	…	…	…	…
loc_q	NA	0.33	…	0.67

Methods for analyzing bisulfite sequencing data

Programs	Method	Model counts?
Many	t-test or Wilcoxon rank-sum test	No (model proportion)
Many	Fisher's exact test	Yes
Many	Linear regression	No (model normalized proportion)
Many	Logistic/Binomial regression	Yes
DSS (Feng et al. 2014), **MOABS** (Sun et al. 2014), **RADMeth** (Dolzhenko & Smith 2014)	Based on a beta binomial regression	Yes

Beta binomial models

- $m_i \sim \text{Bin}(n_i, p_i)$, where p_i is the unknown methylation level of the site (m_i and n_i are the observed counts)

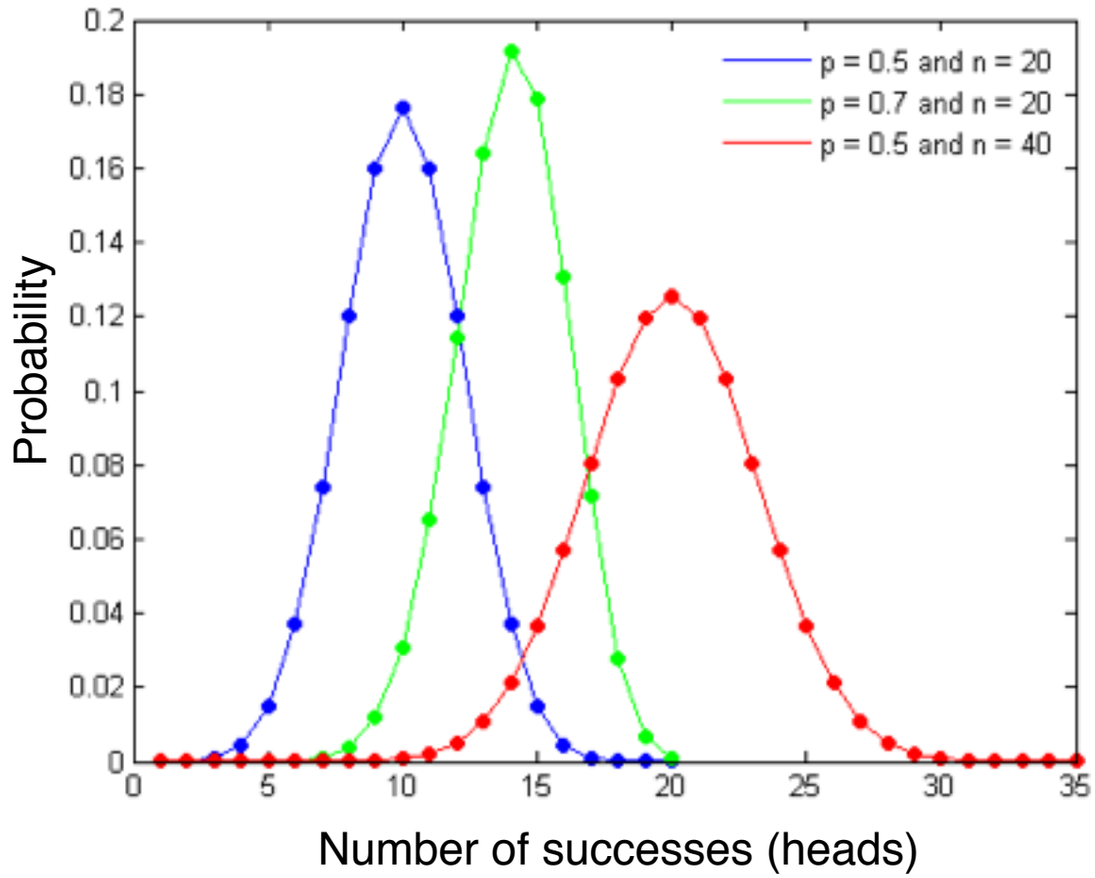


Beta binomial models

- $m_i \sim \text{Bin}(n_i, p_i)$, where p_i is the unknown methylation level of the site (m_i and n_i are the observed counts)

- The *binomial* assumes:

  $E(m_i) = n_i * p_i$
  $\text{Var}(m_i) = n_i * p_i * (1 - p_i)$



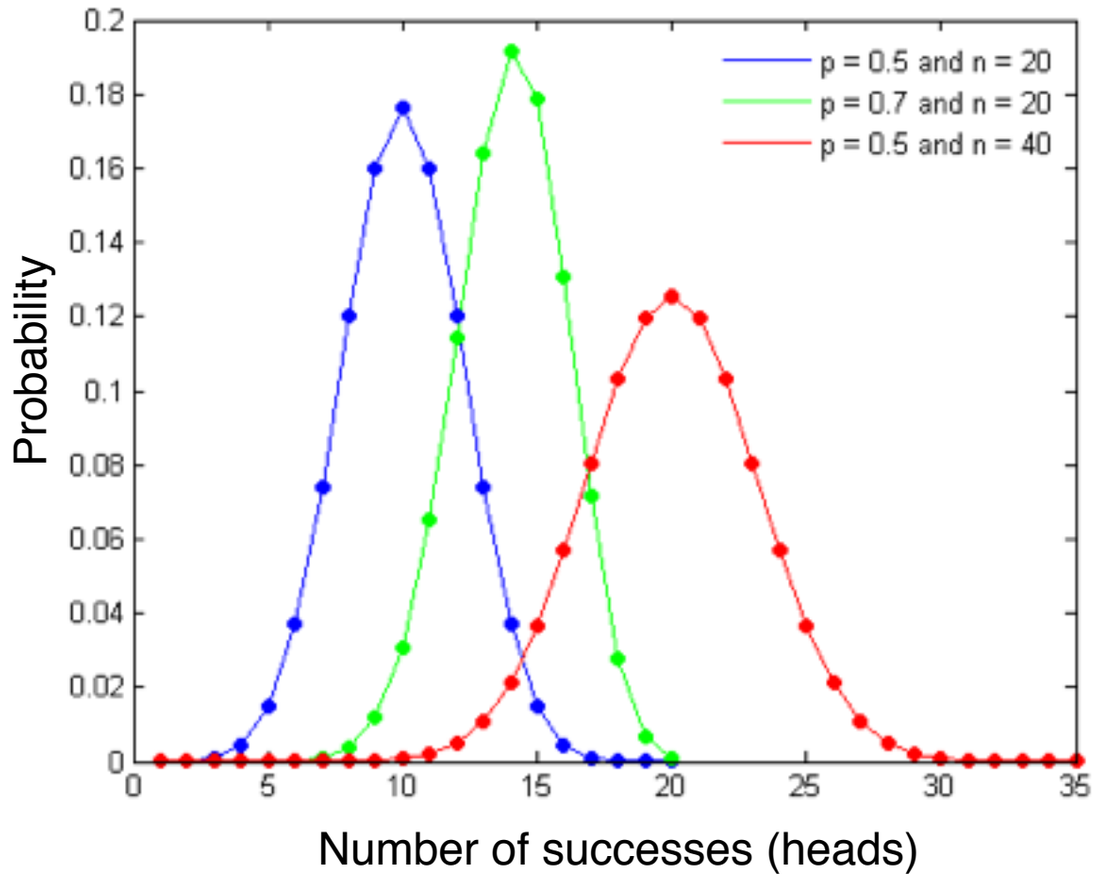
- But counts from bisulfite-sequencing data are over dispersed

Beta binomial models

- $m_i \sim \text{Bin}(n_i, p_i)$, where p_i is the unknown methylation level of the site (m_i and n_i are the observed counts)

- The *binomial* assumes:

  $E(m_i) = n_i * p_i$
  $\text{Var}(m_i) = n_i * p_i * (1 - p_i)$



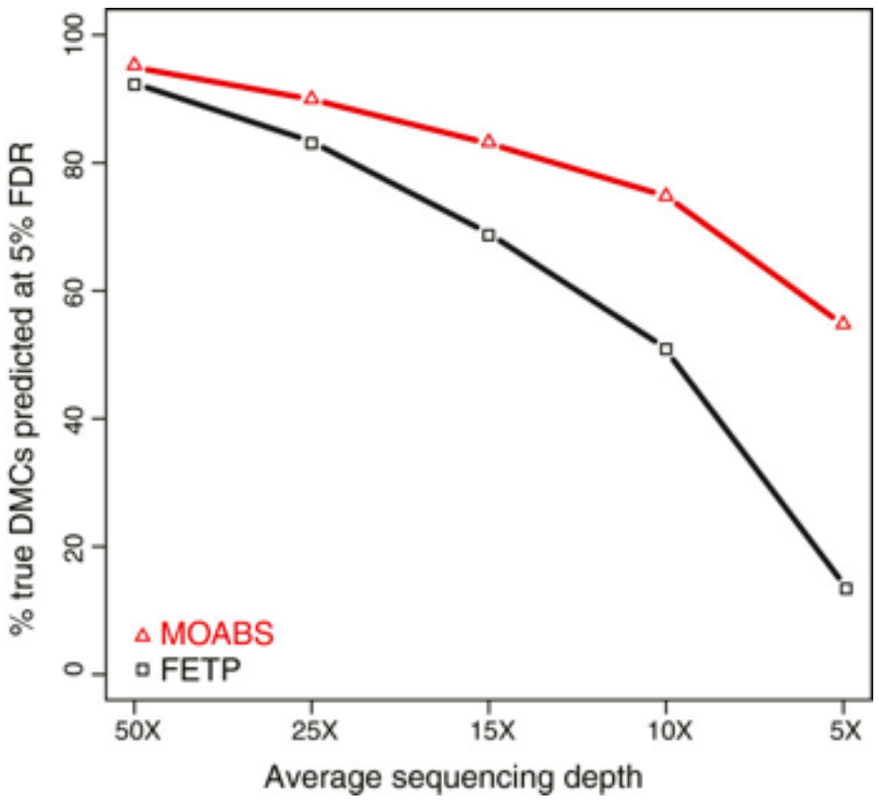
- But counts from bisulfite-sequencing data are over dispersed

- The *beta-binomial* assumes p_i is a random variable and follows a beta distribution (accounts for overdispersion)

- Not surprisingly, *beta-binomial* models outperform non count-based methods, as well as binomial models for analyzing DNA methylation datasets

Beta binomial models

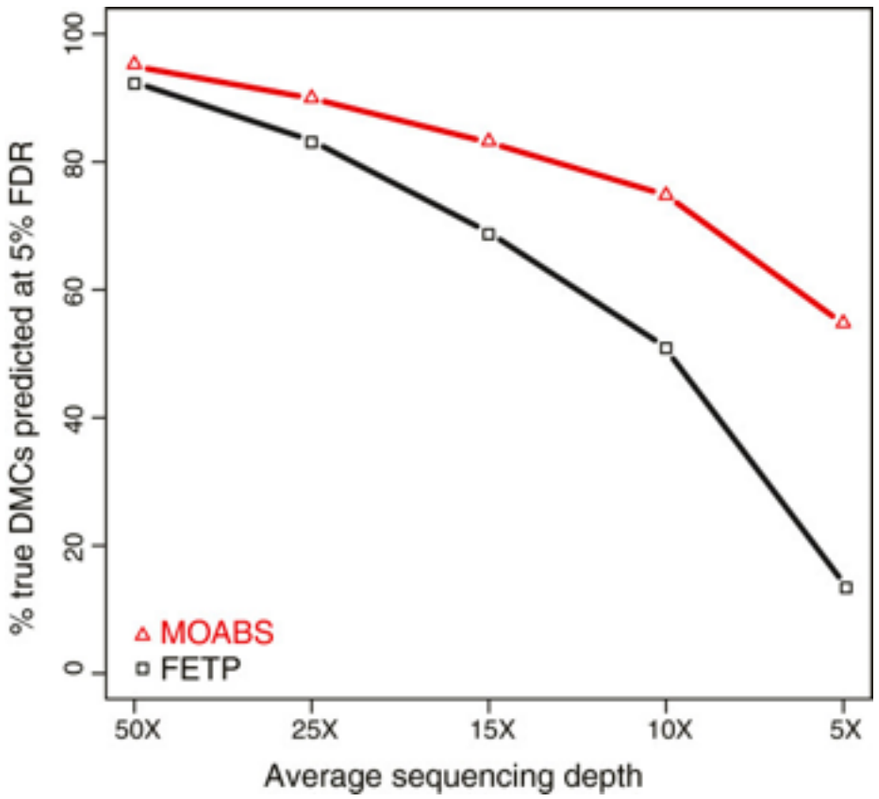
Comparison between 2 samples



MOABS: Sun et al. 2014, Genome Biology

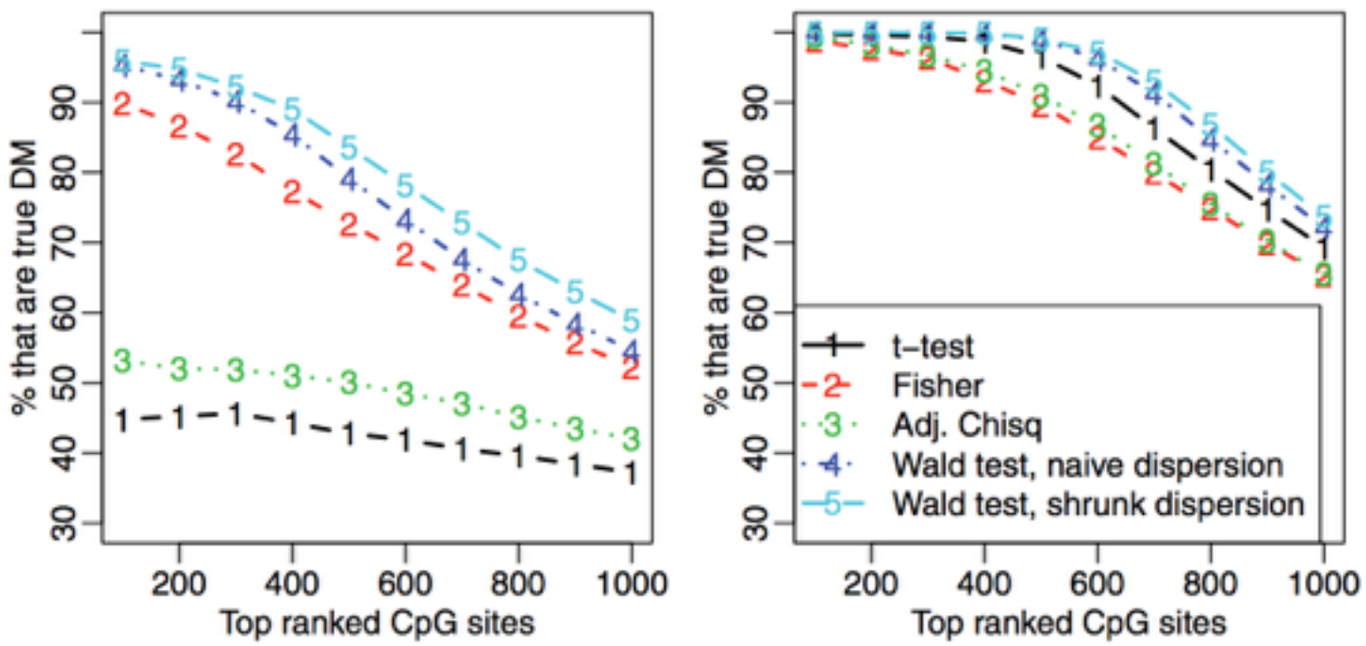
Beta binomial models

Comparison between 2 samples



MOABS: Sun et al. 2014, Genome Biology

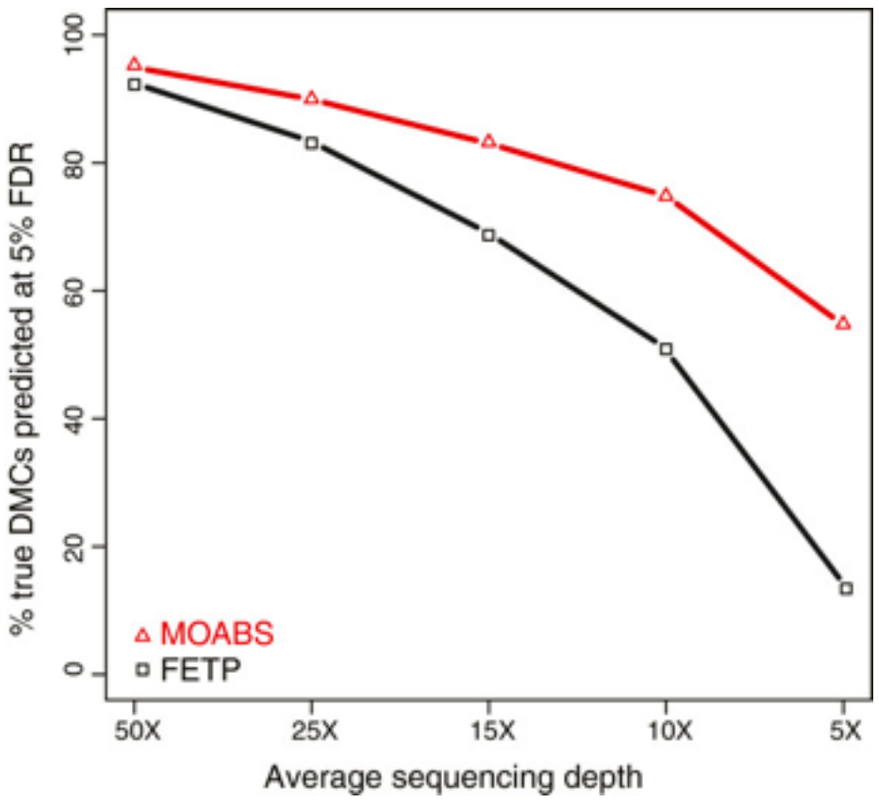
Comparison between 2 groups



DSS: Feng et al. 2014, Nucleic Acids Research

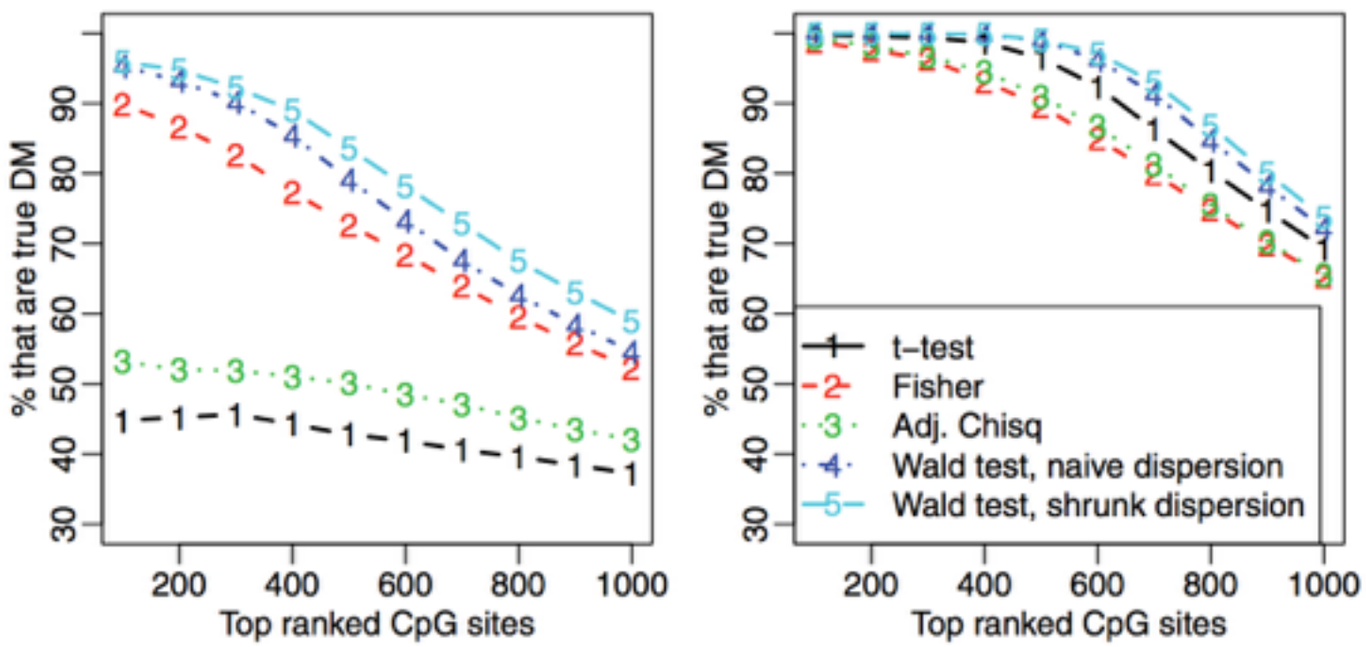
Beta binomial models

Comparison between 2 samples



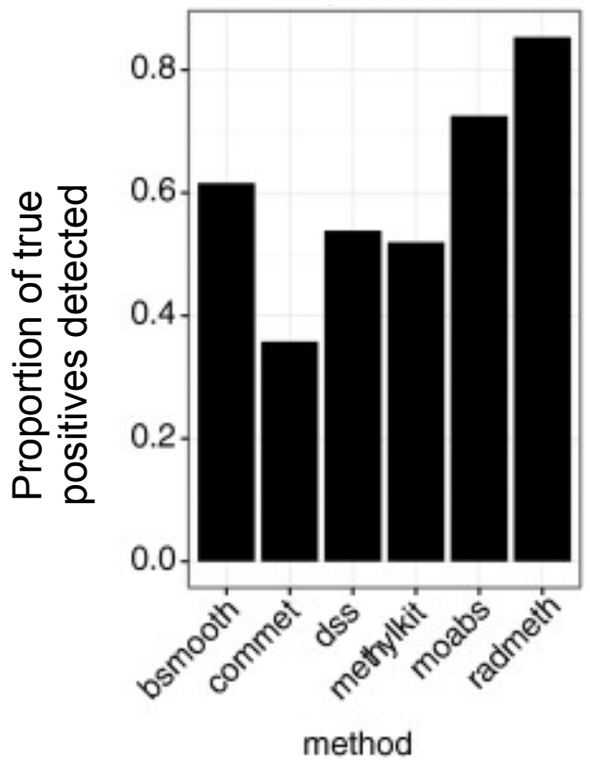
MOABS: Sun et al. 2014, Genome Biology

Comparison between 2 groups



DSS: Feng et al. 2014, Nucleic Acids Research

Multiple biological replicates, continous predictor



RadMeth: Dolzhenko and Smith 2014, BMC Bioinformatics

Beta binomial models

These tools are an improvement, but are relatively inflexible and few allow researchers to model additional covariates (e.g.,):

- technical/batch effects effects
- age
- sex
- cell type heterogeneity

Beta binomial models

These tools are an improvement, but are relatively inflexible and few allow researchers to model additional covariates (e.g.,):

- technical/batch effects effects
- age
- sex
- cell type heterogeneity

- also, genetic effects!

Table 1 Average correlation across all probes of normalised methylation measurements between relative pairs

Relationship	Pairs (n)	Correlation	Expected[a]
MZ twins	67	0.200	h^2
DZ twins	111	0.109	$h^2/2$
Siblings	262[b]	0.090	$h^2/2$
Parent-Offspring	362[b]	0.089	$h^2/2$
Mother-Offspring	190	0.097	$h^2/2$
Father-Offspring	172	0.085	$h^2/2$
Parent-Parent	58	0.023	0
Unrelated	187,331[b]	-0.002	0

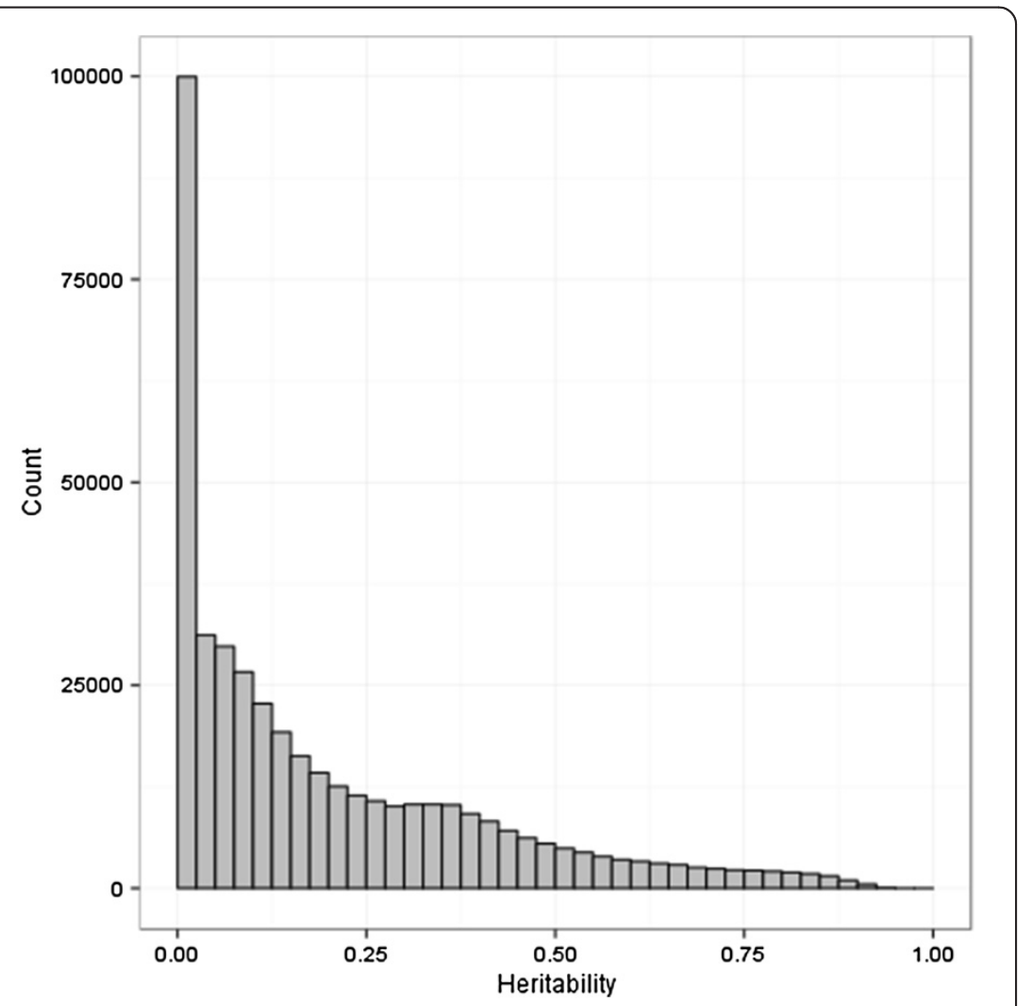


Figure 1 Distribution of heritability estimates for DNA methylation levels. The average genetic heritability estimate is 0.199. A zero estimate for genetic heritability was observed in 17.1% of cases indicating that genetic heritability results in transgenerational inheritance of DNA methylation for at least 65.8% of probes.

McRae et al. 2014, Genome Biology

Methods for analyzing bisulfite sequencing data

Programs	Method	Model counts?	Control for relatedness/ structure?	Control for covariates?
Many	t-test or Wilcoxon rank-sum test	No	No	No
Many	Fisher's exact test	Yes	No	No
Many	Linear regression	No	No	Yes
Many	Logistic/Binomial regression	Yes	No	Yes
DSS (Feng et al. 2014), **MOABS** (Sun et al. 2014), **RADMeth** (Dolzhenko & Smith 2014)	Based on a beta binomial model	Yes	No	No/Yes

Methods for analyzing bisulfite sequencing data

Programs	Method	Model counts?	Control for relatedness/ structure?	Control for covariates?
Many	t-test or Wilcoxon rank-sum test	No	No	No
Many	Fisher's exact test	Yes	No	No
Many	Linear regression	No	No	Yes
Many	Logistic/Binomial regression	Yes	No	Yes
DSS (Feng et al. 2014), **MOABS** (Sun et al. 2014), **RADMeth** (Dolzhenko & Smith 2014)	Based on a beta binomial regression	Yes	No	No/Yes
EMMA/EMMAX (Kang et al. 2010), **GEMMA** (Zhou & Stephens 2014)	Based on a linear mixed effects model	No	**Yes**	Yes

The Linear Mixed Model

$$\mathbf{y} = \mu + \mathbf{x}\beta + \mathbf{u} + \boldsymbol{\epsilon},$$
$$\mathbf{u} \sim \text{MVN}_n(0, \sigma_g^2 \mathbf{K}), \quad \boldsymbol{\epsilon} \sim \text{MVN}_n(0, \sigma_e^2 \mathbf{I}_{n \times n}),$$

- $\mathbf{y}$ is a n-vector of phenotypes,
- $\mathbf{x}$ is a n-vector of genotypes for the SNP,
- $\mathbf{u}$ is a n-vector of random effects,
- $\boldsymbol{\epsilon}$ is a n-vector of residual errors,
- $\mathbf{K}$ is a known $n \times n$ relatedness matrix, estimated either from a pedigree or genotypes.

$$\sigma^2_g = \sigma^2 h^2 \text{ and } \sigma^2_e = \sigma^2 (1\text{-}h)^2$$

Methods for analyzing bisulfite sequencing data

Programs	Method	Model counts?	Control for relatedness/ structure?	Control for covariates?
Many	t-test or Wilcoxon rank-sum test	No	No	No
Many	Fisher's exact test	Yes	No	No
Many	Linear regression	No	No	Yes
Many	Logistic/Binomial regression	Yes	No	Yes
DSS (Feng et al. 2014)**, MOABS** (Sun et al. 2014)**, RADMeth** (Dolzhenko & Smith 2014)	Based on a beta binomial regression	Yes	No	No/Yes
EMMA/EMMAX (Kang et al. 2010)**, GEMMA** (Zhou & Stephens 2014)	Based on a linear mixed effects model	**No**	Yes	Yes

Methods for analyzing bisulfite sequencing data

Programs	Method	Model counts?	Control for relatedness/ structure?	Control for covariates?
Many	t-test or Wilcoxon rank-sum test	No	No	No
Many	Fisher's exact test	Yes	No	No
Many	Linear regression	No	No	Yes
Many	Logistic/Binomial regression	Yes	No	Yes
DSS (Feng et al. 2014)**, MOABS** (Sun et al. 2014)**, RADMeth** (Dolzhenko & Smith 2014)	Based on a beta binomial regression	Yes	No	No/Yes
EMMA/EMMAX (Kang et al. 2010)**, GEMMA** (Zhou & Stephens 2014)	Based on a linear mixed effects model	No	Yes	Yes
MACAU (Lea et al. 2015)	Based on a binomial mixed effects model	Yes	Yes	Yes

The Linear Mixed Model

$$\mathbf{y} = \mu + \mathbf{x}\beta + \mathbf{u} + \boldsymbol{\epsilon},$$
$$\mathbf{u} \sim \text{MVN}_n(0, \sigma_g^2 \mathbf{K}), \quad \boldsymbol{\epsilon} \sim \text{MVN}_n(0, \sigma_e^2 \mathbf{I}_{n \times n}),$$

- $\mathbf{y}$ is a n-vector of phenotypes,
- $\mathbf{x}$ is a n-vector of genotypes for the SNP,
- $\mathbf{u}$ is a n-vector of random effects,
- $\boldsymbol{\epsilon}$ is a n-vector of residual errors,
- $\mathbf{K}$ is a known $n \times n$ relatedness matrix, estimated either from a pedigree or genotypes.

$$\sigma^2_g = \sigma^2 h^2 \text{ and } \sigma^2_e = \sigma^2 (1\text{-}h)^2$$

A binomial mixed effects model

methylated counts ~ Bin(total counts, p)

$$\text{log(p/(1-p))} = \mu + \mathbf{x}\beta + \mathbf{u} + \boldsymbol{\epsilon},$$

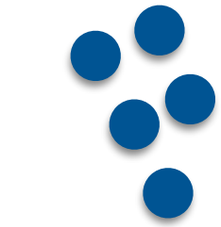
$$\mathbf{u} \sim \text{MVN}_n(0, \sigma_g^2 \mathbf{K}), \quad \boldsymbol{\epsilon} \sim \text{MVN}_n(0, \sigma_e^2 \mathbf{I}_{n \times n}),$$

- $\mathbf{y}$ is a n-vector of phenotypes,
- $\mathbf{x}$ is a n-vector of genotypes for the SNP,
- $\mathbf{u}$ is a n-vector of random effects,
- $\boldsymbol{\epsilon}$ is a n-vector of residual errors,
- $\mathbf{K}$ is a known $n \times n$ relatedness matrix, estimated either from a pedigree or genotypes.

$\sigma^2_g = \sigma^2 h^2$ and $\sigma^2_e = \sigma^2 (1\text{-}h)^2$

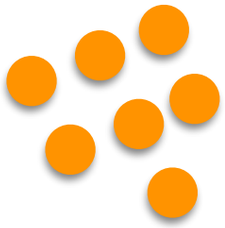
Xiang Zhou
University of Michigan

Correlation (R^2) between population structure and predictor variable



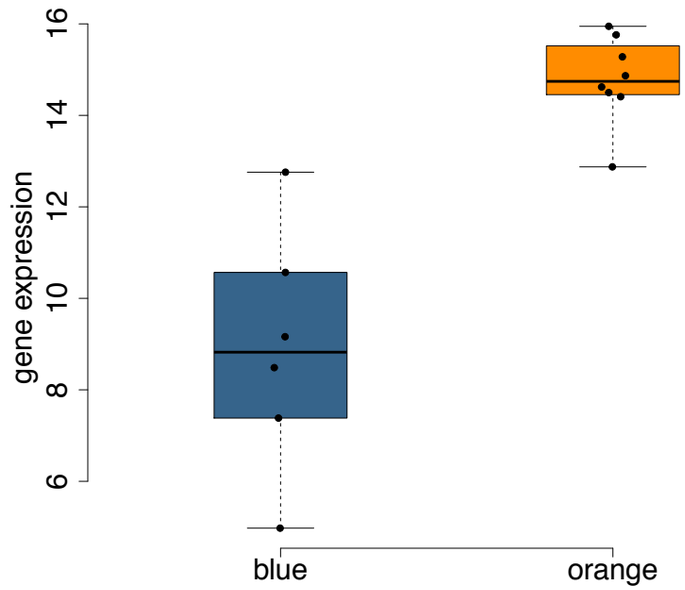
high status

high status family



low status

low status family



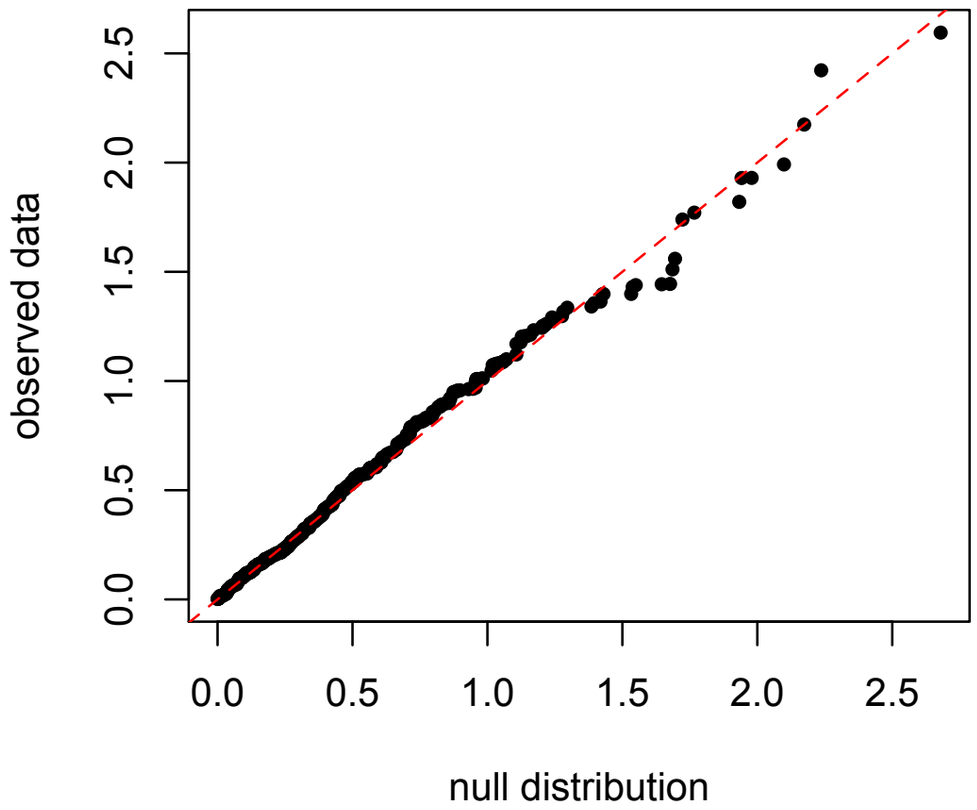
Correlation (R^2) between population structure and predictor variable

Genomic control factor (λ)

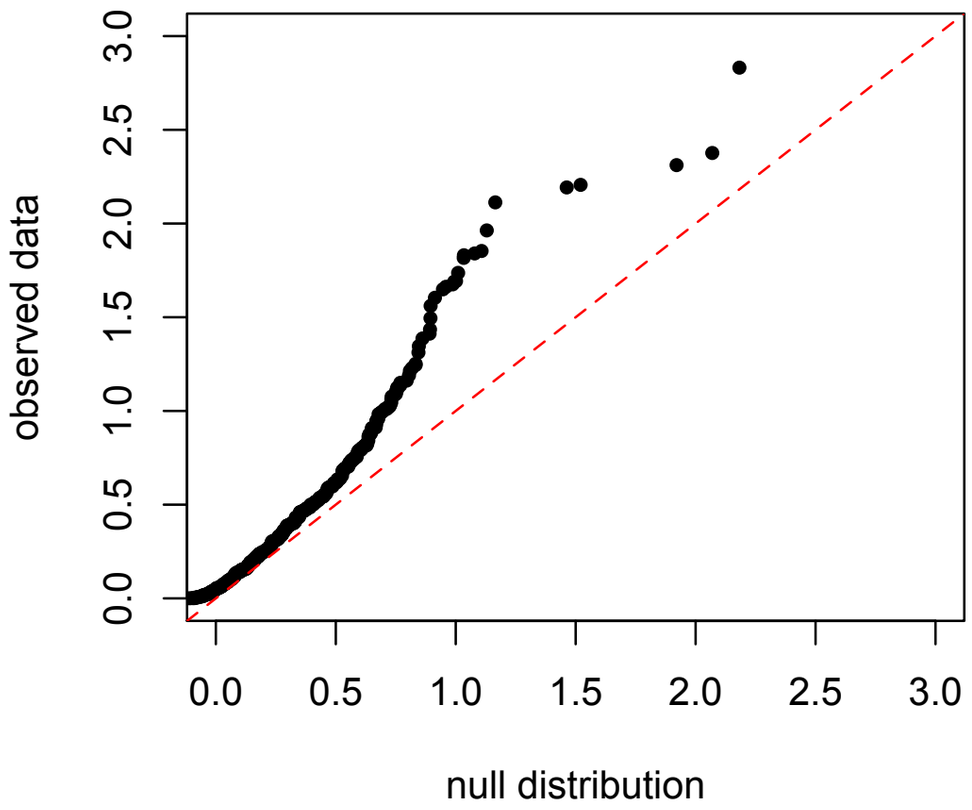
Correlation (R^2) between population structure and predictor variable

Genomic control factor (λ)

genomic control factor ~ 0

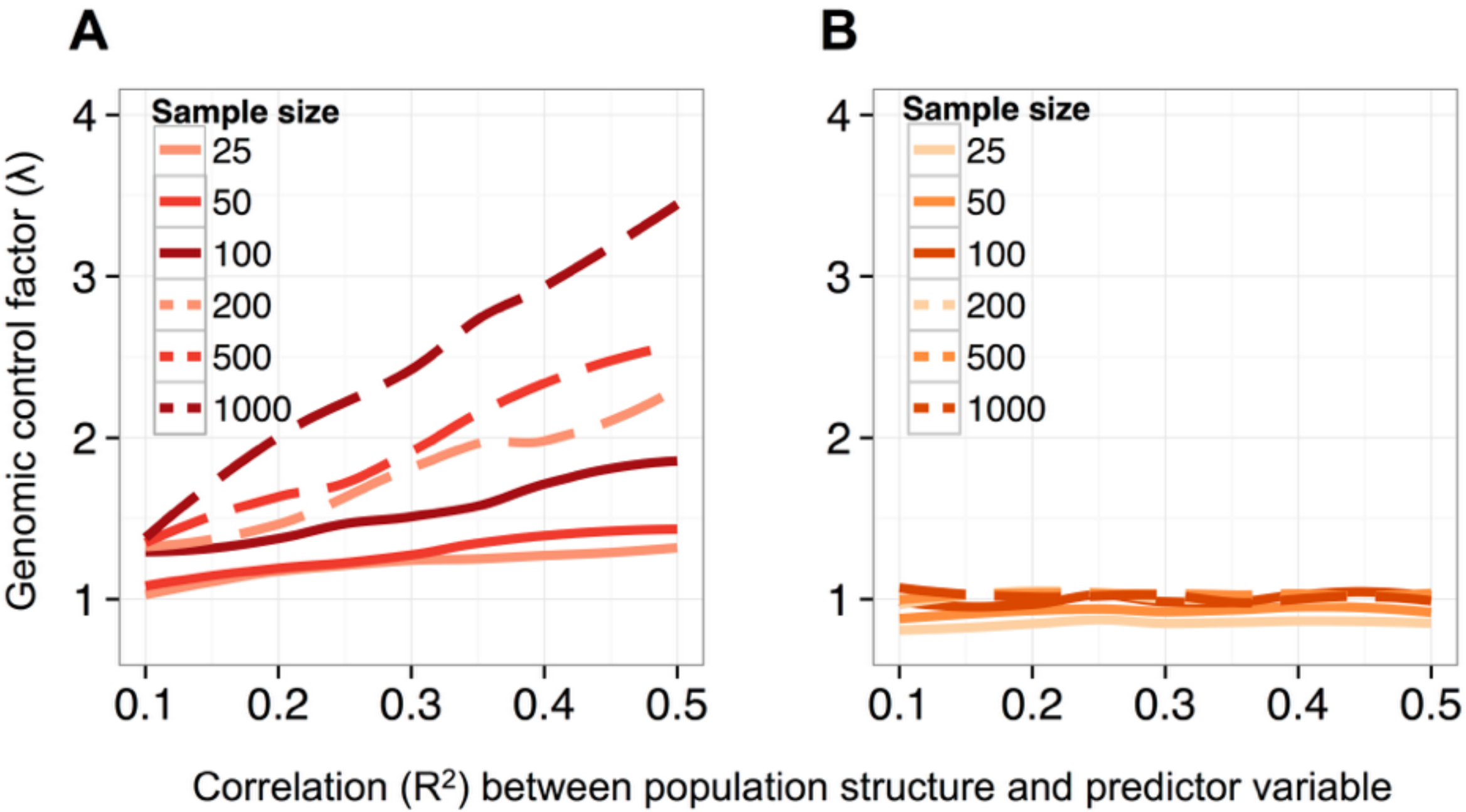


genomic control factor >> 0



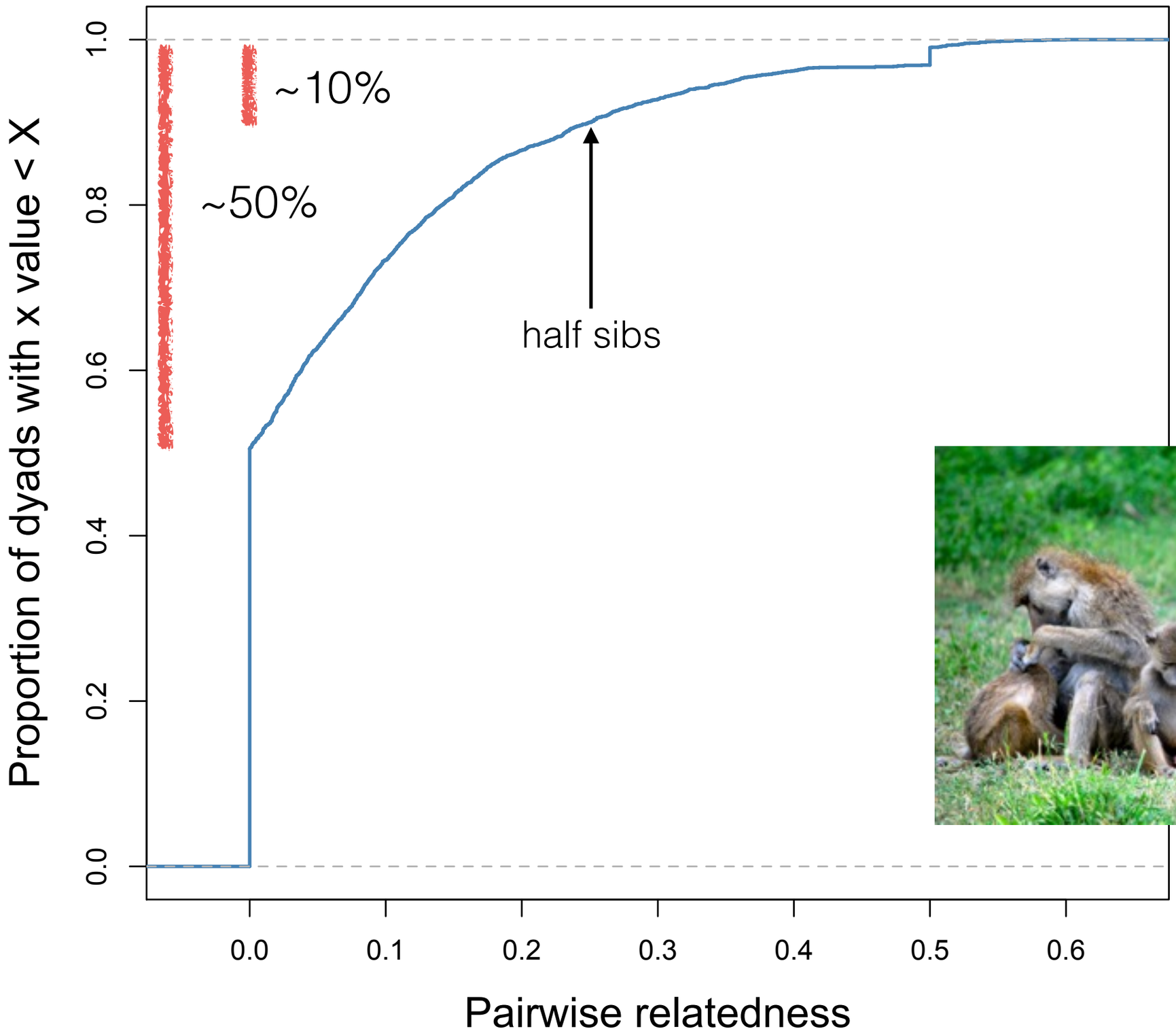
Correlation (R^2) between population structure and predictor variable

Binomial mixed effects models minimize false positives when genetic structure is confounded with the predictor of interest (beta binomial models do not)



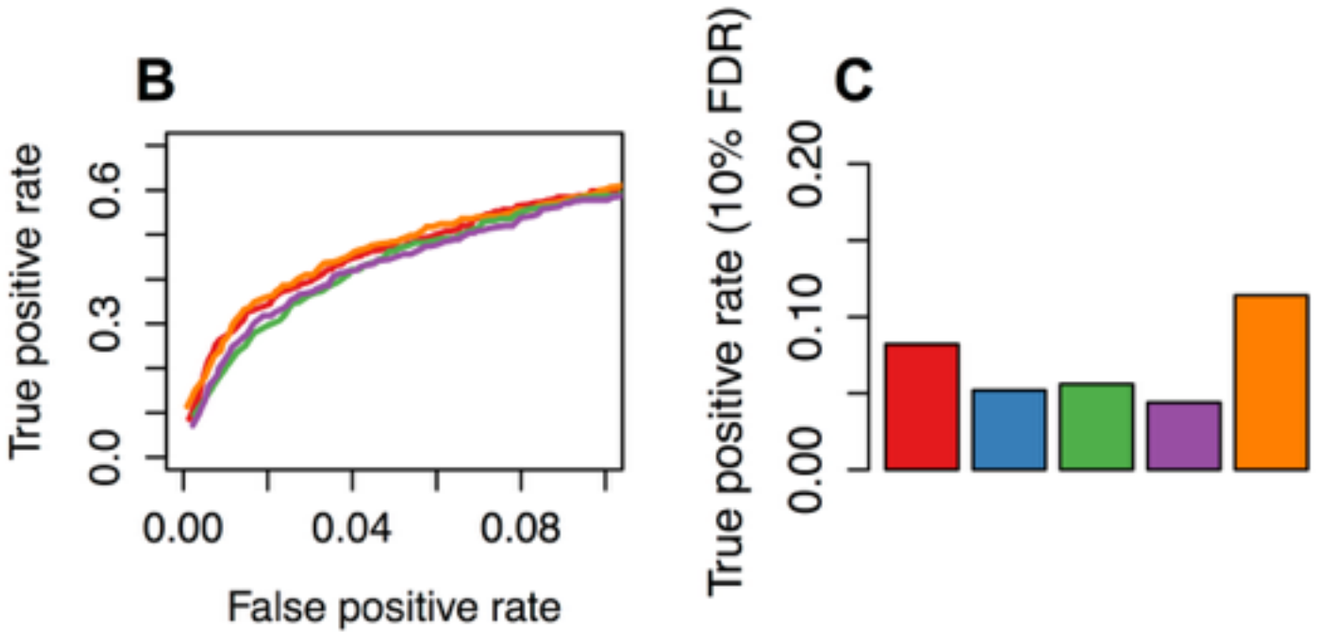
Lea et al. 2015, PLoS Genetics

What about samples that contain relatives (but there's no cofound)?

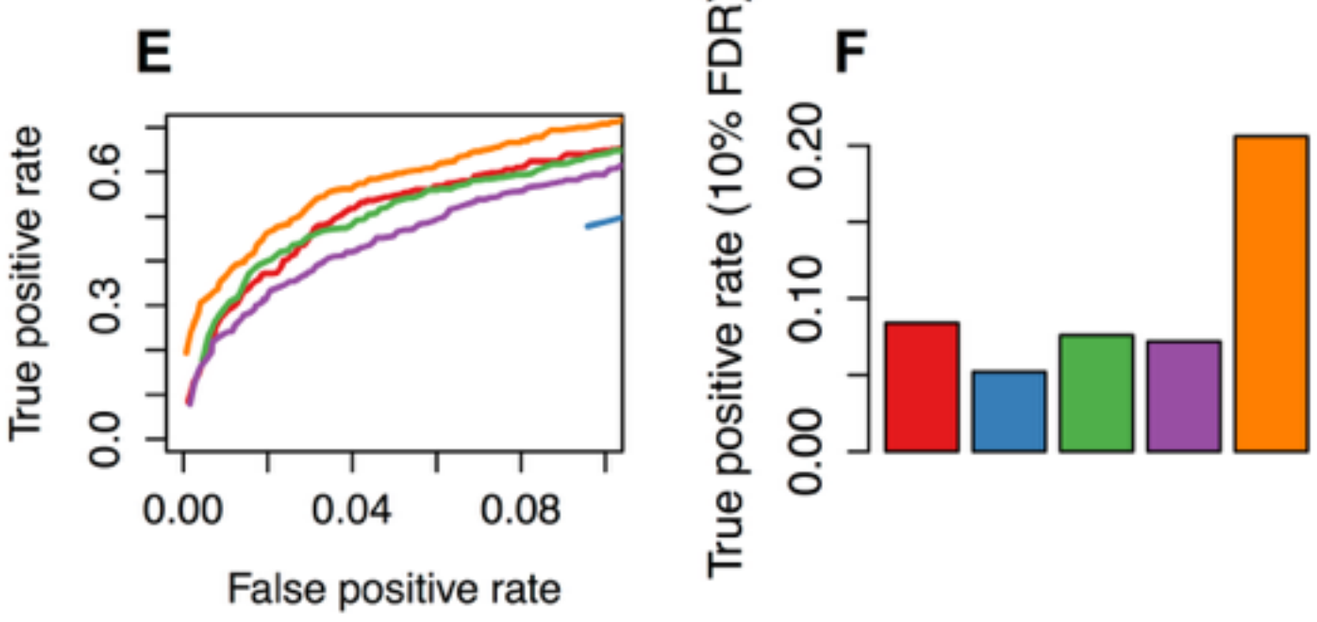


Binomial mixed effects models improve power when the sample contains related individuals (relative to beta binomial model)

h^2=0.3



h^2=0.6



Beta-binomial Binomial GEMMA Linear MACAU

Lea et al. 2015, PLoS Genetics

To wrap up, some key points:

- DNA methylation is affected by many factors (age, cell type composition, genetic variation, environmental conditions) that we want to study but should also account for

- Bisulfite sequencing (RRBS or WGBS) is the method of choice for those that work on non-model organisms

- Bisulfite sequencing data is count data, with a lot of variation/unique coverage properties

- Count-based models generally provide better power (than non count-based models)

- Controlling for relatedness is important and can (i) reduce false positive associations and (ii) increase power to detect true effects

Some important things we didn’t talk about:

Some important things we didn’t talk about:

- Batch effects (they exist in DNA methylation data too)!
  - Bisulfite conversion rate, sequencing depth, library prep, etc.
  - Same approaches to identify them apply, e.g., PCA

Some important things we didn’t talk about:

- Batch effects (they exist in DNA methylation data too)!
  - Bisulfite conversion rate, sequencing depth, library prep, etc.
  - Some approaches to identify them apply, e.g., PCA

- False discovery rates and empirical nulls (they apply here too)!

Some important things we didn't talk about:

- Batch effects (they exist in DNA methylation data too)!
  - Bisulfite conversion rate, sequencing depth, library prep, etc.
  - Some approaches to identify them apply, e.g., PCA

- False discovery rates and empirical nulls (they apply here too)!

- Cell type heterogeneity
  - Measure it!
  - Deconvolution
  - Understand magnitude of bias using methylation data from pure cell types

Some important things we didn't talk about:

- Batch effects (they exist in DNA methylation data too)!
  - Bisulfite conversion rate, sequencing depth, library prep, etc.
  - Some approaches to identify them apply, e.g., PCA

- False discovery rates and empirical nulls (they apply here too)!

- Cell type heterogeneity
  - Measure it!
  - Deconvolution
  - Understand magnitude of bias using methylation data from pure cell types

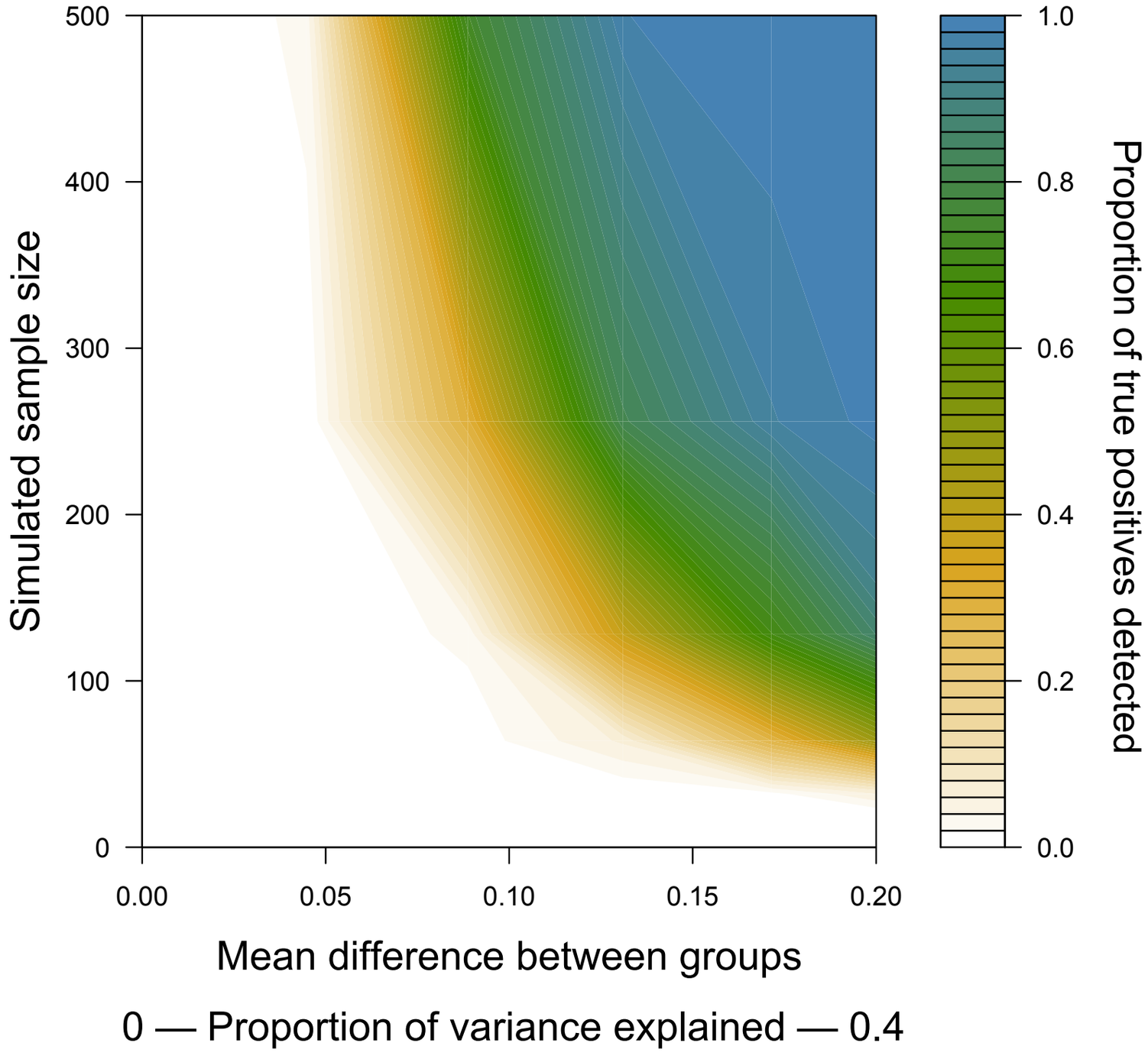
- Other kinds of analyses, e.g.,
  - Genotyping from bisulfite-sequencing data
  - meQTL
  - Identifying differentially methylated regions

Some important things we didn’t talk about:

- Power and sample size
- More biological replicates is always good (if you’re on a budget, think about using RRBS and/or reducing your sequencing depth)

Some important things we didn't talk about:

- Power and sample size
- More biological replicates is always good (if you're on a budget, think about using RRBS and/or reducing your sequencing depth)



DNA methylation

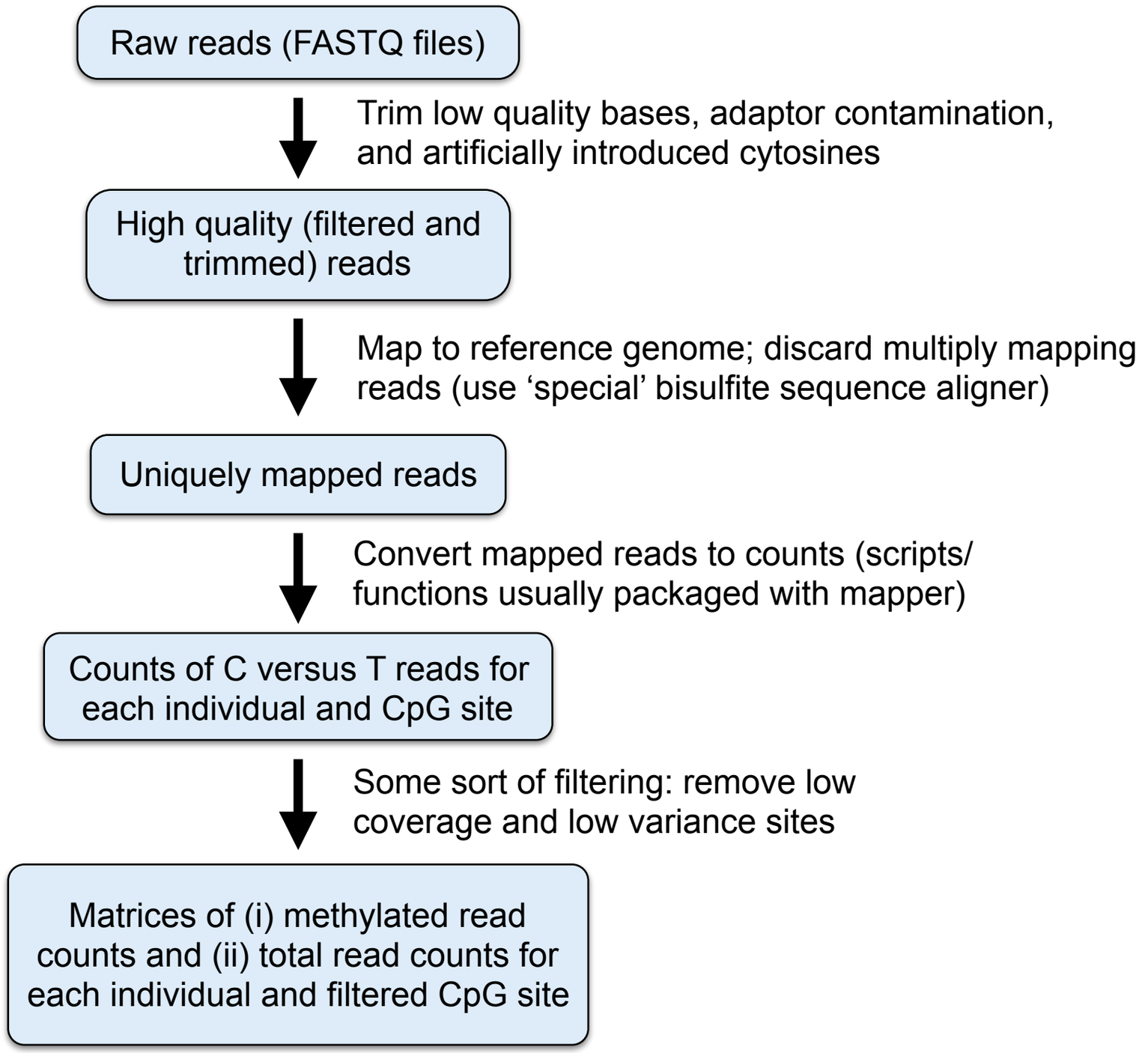
hands on data analysis

Amanda Lea

September 22, 2016

Bisulfite sequencing data

- First - process, map, and filter your data



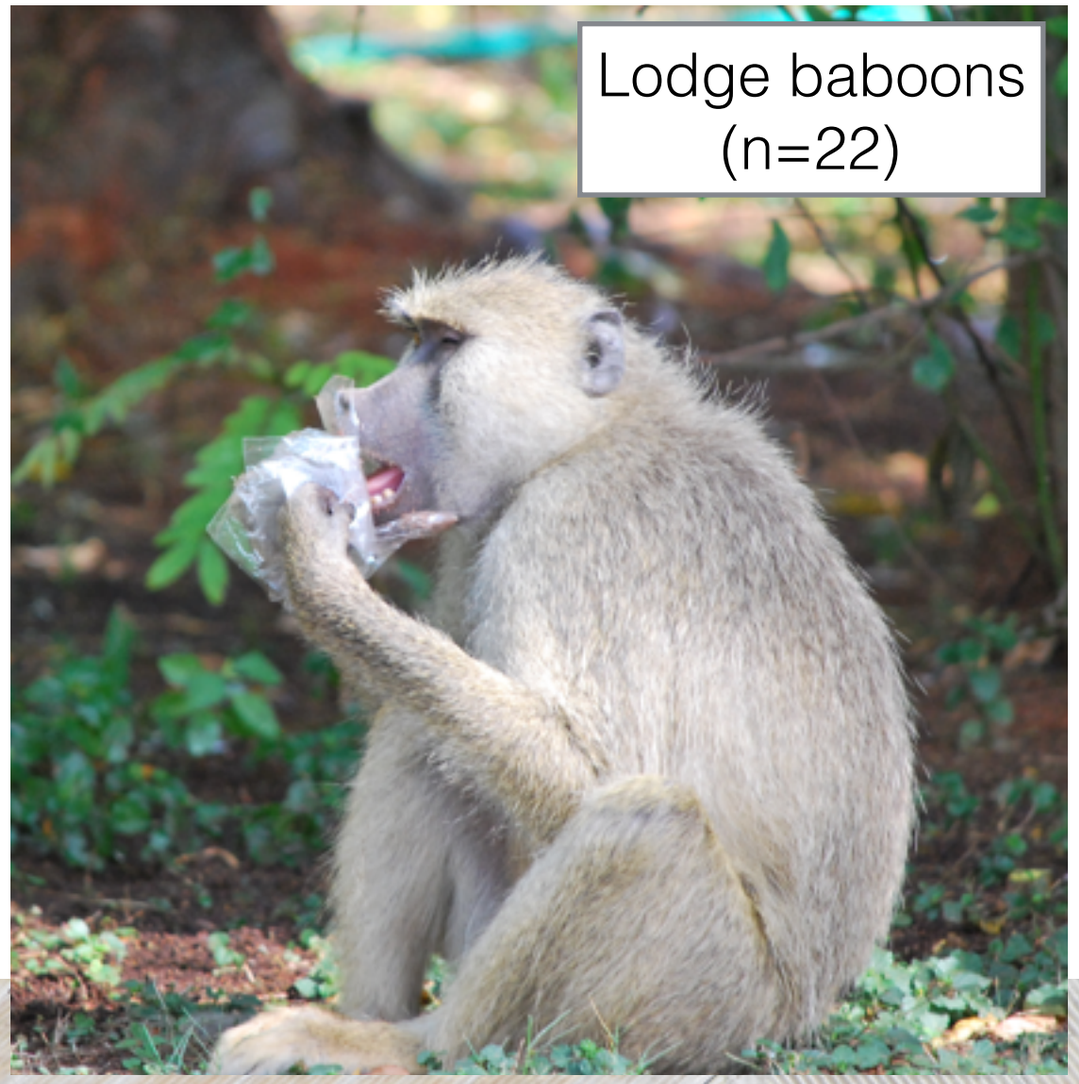
Bisulfite sequencing data

- Now you have a count matrix!

Data matrix for *n* individuals measured at *q* loci
(methylated reads / total reads)

	ID_1	ID_2	…	ID_n
loc_1	2/7	20/70	…	40/140
loc_2	2/3	40/60	…	40/140
loc_3	5/5	50/50		100/100
…	…	…	…	…
loc_q	0/0	1/3	…	4/6

An RRBS dataset



Lea et al. 2016, Molecular Ecology

FYI - you can use these simulations to do power analyses

- Is your sample size sufficient to detect your expected effect size?