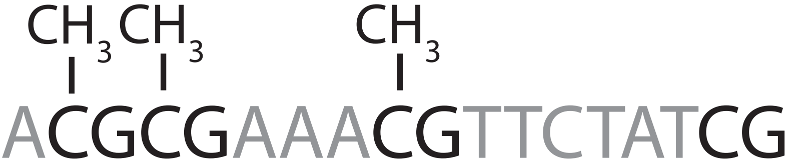
DNA methylation

ACGCGAAACGTTCTATCG

DNA methylation

ACGCGAAACGTTCTATCG

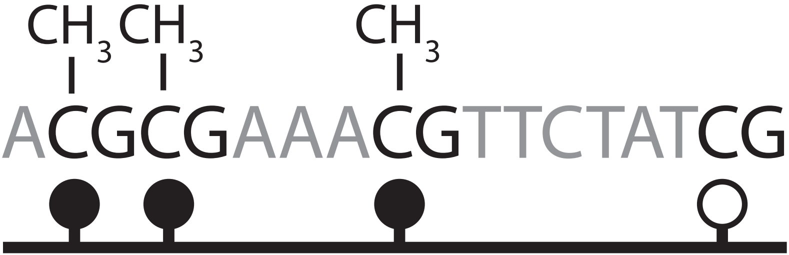
DNA methylation



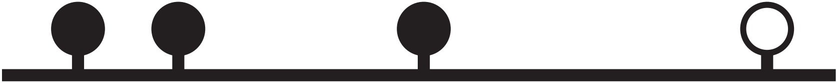
CH_3 CH_3 CH_3

ACGCGAAACGTTCTATCG

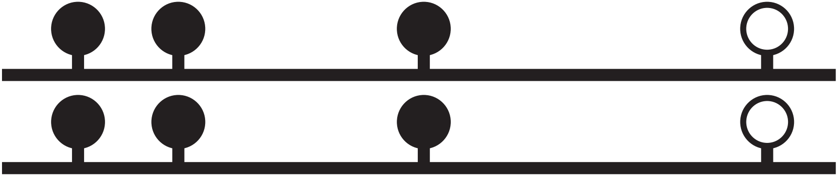
DNA methylation



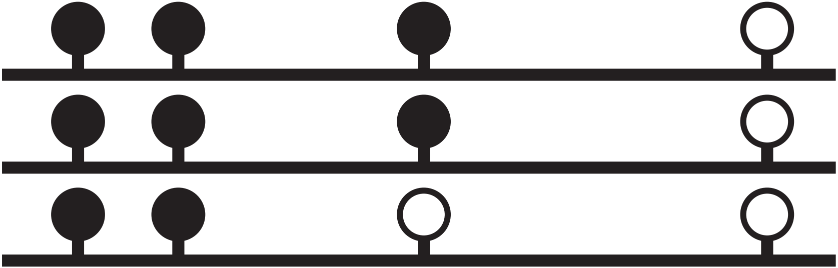
Measuring DNA methylation



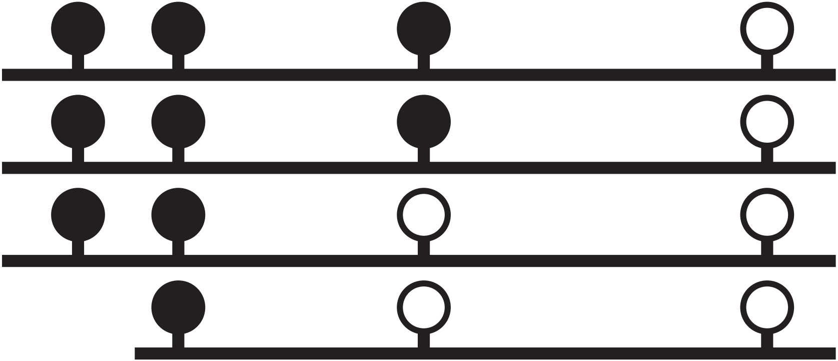
Measuring DNA methylation



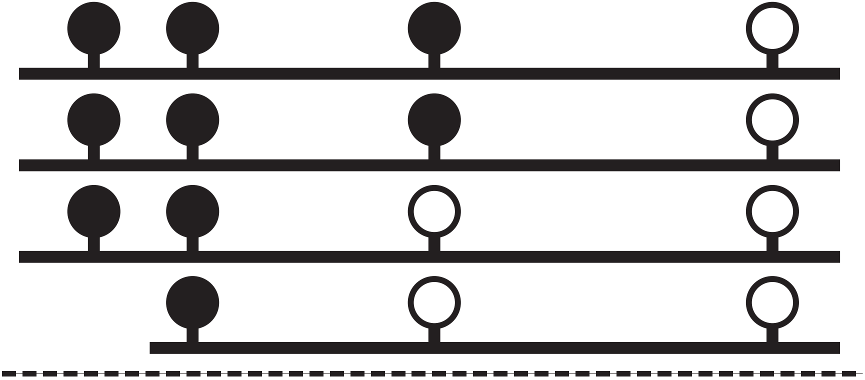
Measuring DNA methylation



Measuring DNA methylation



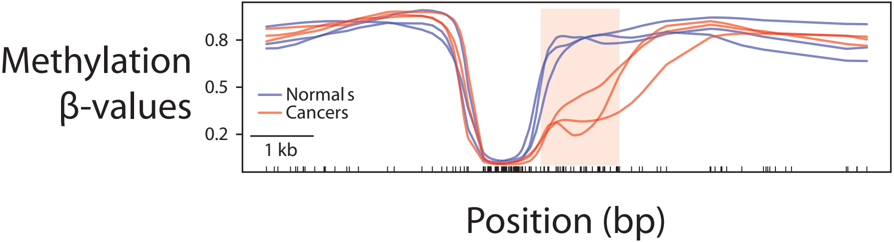
Measuring DNA methylation



$\beta_i = 3/3$ $\beta_{i+2} = 2/4$ $\beta_{i+3} = 0/4$

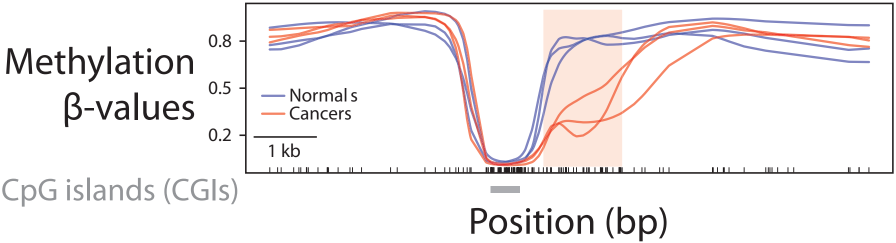
$\beta_{i+1} = 4/4$

Differentially methylated regions (DMRs)[1]



[1] Hansen, K. D. et al. Nat Genet 43, 768–775 (2011)

Differentially methylated regions (DMRs)[1]



[1] Hansen, K. D. et al. Nat Genet 43, 768–775 (2011)

Why I care about simulating DNA methylation data

Methods development and validation

- Do methods designed to find DMRs actually work?
- What method reigns supreme?

Why I care about simulating DNA methylation data

Methods development and validation

- Do methods designed to find DMRs actually work?
- What method reigns supreme?

How to decide?

- No “gold standard” data ⇒ simulate

Why I care about simulating DNA methylation data

Methods development and validation

- Do methods designed to find DMRs actually work?
- What method reigns supreme?

How to decide?

- No "gold standard" data ⇒ simulate
- No simulation software ⇒ I'm writing `methsim`.

Simulation approaches

Simulate β-values

- Simulate independent $\beta_i \overset{d}{=} Beta(\mu_i, \nu_i)$ + induce correlation via variogram model.

Simulation approaches

Simulate β-values

- Simulate independent $\beta_i \overset{d}{=} Beta(\mu_i, \nu_i)$ + induce correlation via variogram model.
- Re-sample real data in a way that tries to preserve correlation structure.

Simulation approaches

Simulate β-values

- Simulate independent $\beta_i \overset{d}{=} Beta(\mu_i, \nu_i)$ + induce correlation via variogram model.
- Re-sample real data in a way that tries to preserve correlation structure.
- **β-values are summarised measurements.**

Simulation approaches

Simulate β-values

- Simulate independent $\beta_i \overset{d}{=} Beta(\mu_i, \nu_i)$ + induce correlation via variogram model.
- Re-sample real data in a way that tries to preserve correlation structure.
- **β-values are summarised measurements.**
- **Correlations of β-values are spurious.**

Simulation approaches

Simulate β-values

- Simulate independent $\beta_i \overset{d}{=} Beta(\mu_i, \nu_i)$ + induce correlation via variogram model.
- Re-sample real data in a way that tries to preserve correlation structure.
- **β-values are summarised measurements.**
- **Correlations of β-values are spurious.**

Simulation approaches

Simulate β-values

- Simulate independent $\beta_i \overset{d}{=} Beta(\mu_i, \nu_i)$ + induce correlation via variogram model.
- Re-sample real data in a way that tries to preserve correlation structure.
- **β-values are summarised measurements.**
- **Correlations of β-values are spurious.**

Simulate individual methylation events

- Higher resolution.

Simulation approaches

Simulate β-values

- Simulate independent $\beta_i \overset{d}{=} Beta(\mu_i, \nu_i)$ + induce correlation via variogram model.
- Re-sample real data in a way that tries to preserve correlation structure.
- **β-values are summarised measurements.**
- **Correlations of β-values are spurious.**

Simulate individual methylation events

- Higher resolution.
- Contains the mechanistic dependence structure.

Simulation approaches

Simulate β-values

- Simulate independent $\beta_i \overset{d}{=} Beta(\mu_i, \nu_i)$ + induce correlation via variogram model.
- Re-sample real data in a way that tries to preserve correlation structure.
- **β-values are summarised measurements.**
- **Correlations of β-values are spurious.**

Simulate individual methylation events

- Higher resolution.
- Contains the mechanistic dependence structure.
- **Difficult given current data.**

My solution

`methsim`: An R package for simulating whole genome DNA methylation data.

- Parameter distributions estimated from input data.
- Parts written in C++ (via `Rcpp`).
- Results today from a preliminary version of `methsim`.

My solution

methsim: An R package for simulating whole genome DNA methylation data.

- Parameter distributions estimated from input data.
- Parts written in C++ (via Rcpp).
- Results today from a preliminary version of methsim.

Outline of methsim

1. Segment genome into "region of similarity" (MethylSeekR[1])
2. Simulate "meta-haplotypes" within each region using Markov model.
3. Simulate sequencing of reads.

[a] Burger, L., Gaidatzis, D., Schübeler, D. & Stadler, M. B. Nucleic Acids Res (2013). doi:10.1093/nar/gkt599

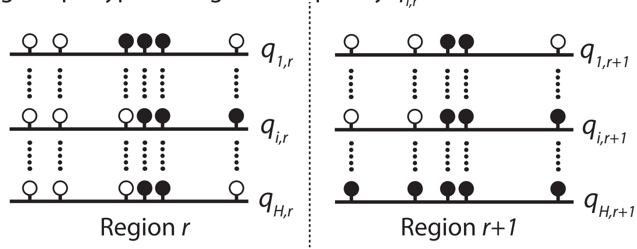
Simulating *meta-haplotypes*

(2) For each region:

Simulate each meta-haplotype using a Markov model

Transition matrices depend on distance between CGs and the type of region

Assign haplotype *i* in region *r* frequency $q_{i,r}$



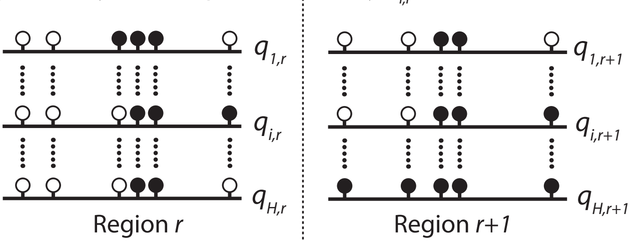
Simulating *meta-haplotypes*

(2) For each region:

Simulate each meta-haplotype using a Markov model

Transition matrices depend on distance between CGs and the type of region

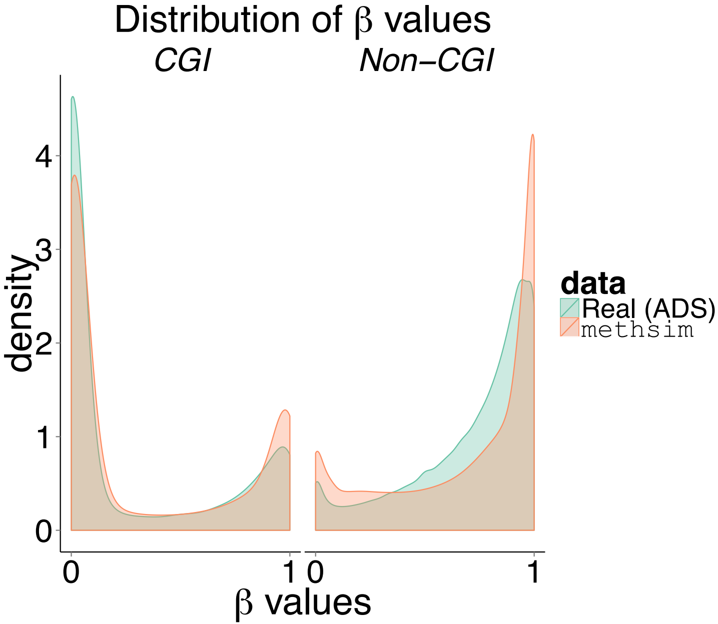
Assign haplotype *i* in region *r* frequency $q_{i,r}$

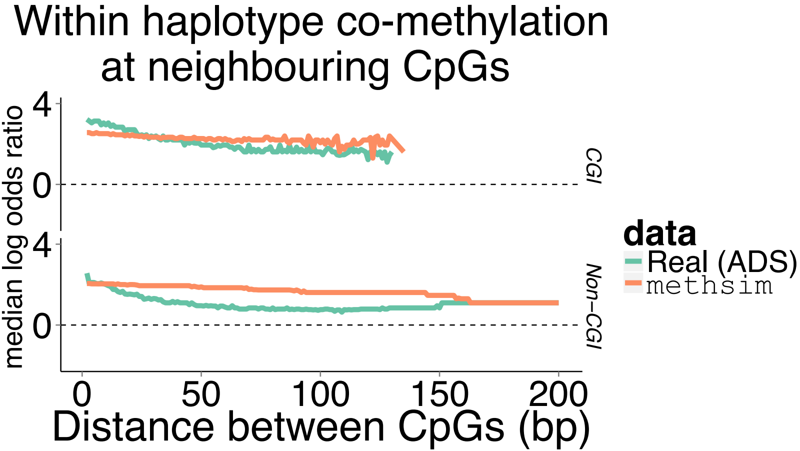


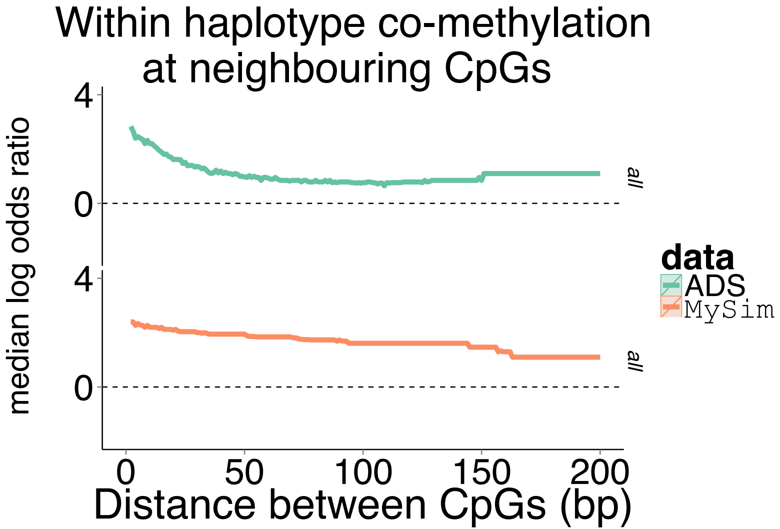
(3) Simulate read positions

Simulate reads for region *r* by sampling from i^{th} haplotype with probability $q_{i,r}$

Simulate sequencing error

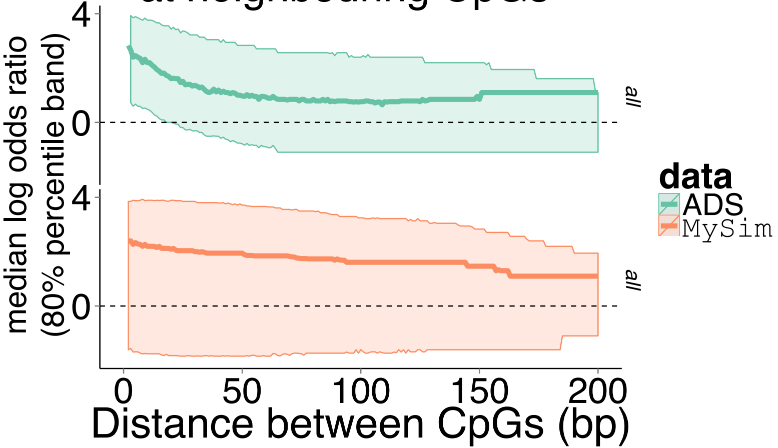

Distribution of β values
CGI
Non–CGI
density
0
1
2
3
4
0
1
0
1
β values
data
Real (ADS)
methsim


Within haplotype co-methylation
at neighbouring CpGs
median log odds ratio
4
0
4
0
CGI
Non–CGI
0
50
100
150
200
Distance between CpGs (bp)
data
Real (ADS)
methsim

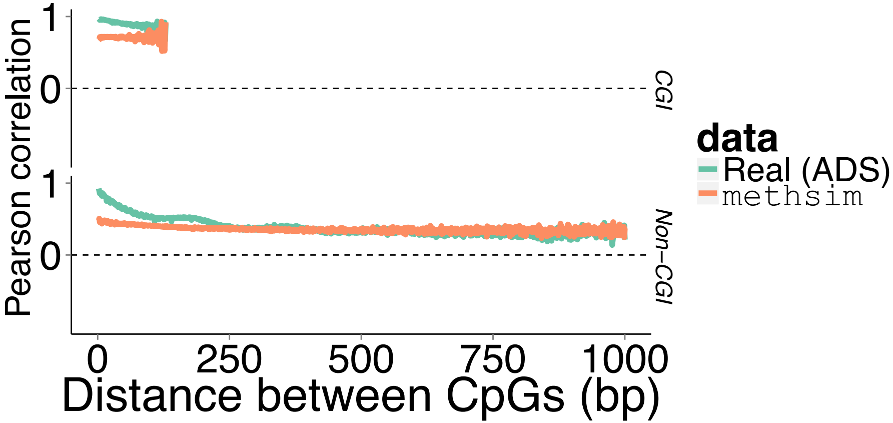


Within haplotype co-methylation
at neighbouring CpGs
median log odds ratio
4
0
4
0
all
all
0
50
100
150
200
Distance between CpGs (bp)
data
ADS
MySim

Within haplotype co-methylation at neighbouring CpGs



Correlations of pairs of β values



Summary

- `methsim` models the mechanistic dependence structure of DNA methylation data.

Summary

- methsim models the mechanistic dependence structure of DNA methylation data.
- Will be using methsim to simulate data with inserted DMRs and compare DMR-detection methods.

Summary

- `methsim` models the mechanistic dependence structure of DNA methylation data.
- Will be using `methsim` to simulate data with inserted DMRs and compare DMR-detection methods.
- `methsim` is open source and developed on GitHub.

Thanks

For advice and supervision

- Terry Speed (WEHI) and Peter Hall (University of Melbourne).

For data

- Ryan Lister (UWA).

For R and C++ help

- Bioconductor and Rcpp mailing lists, especially Martin Morgan.

For funding

- Australian Postgraduate Award, Victorian Life Sciences Computing Initiative.

For sanity

- Friends and family.

To find out more

www.peterhickey.org/ASC2014
GitHub/Twitter: @PeteHaitch