Making sense of DNA methylation data

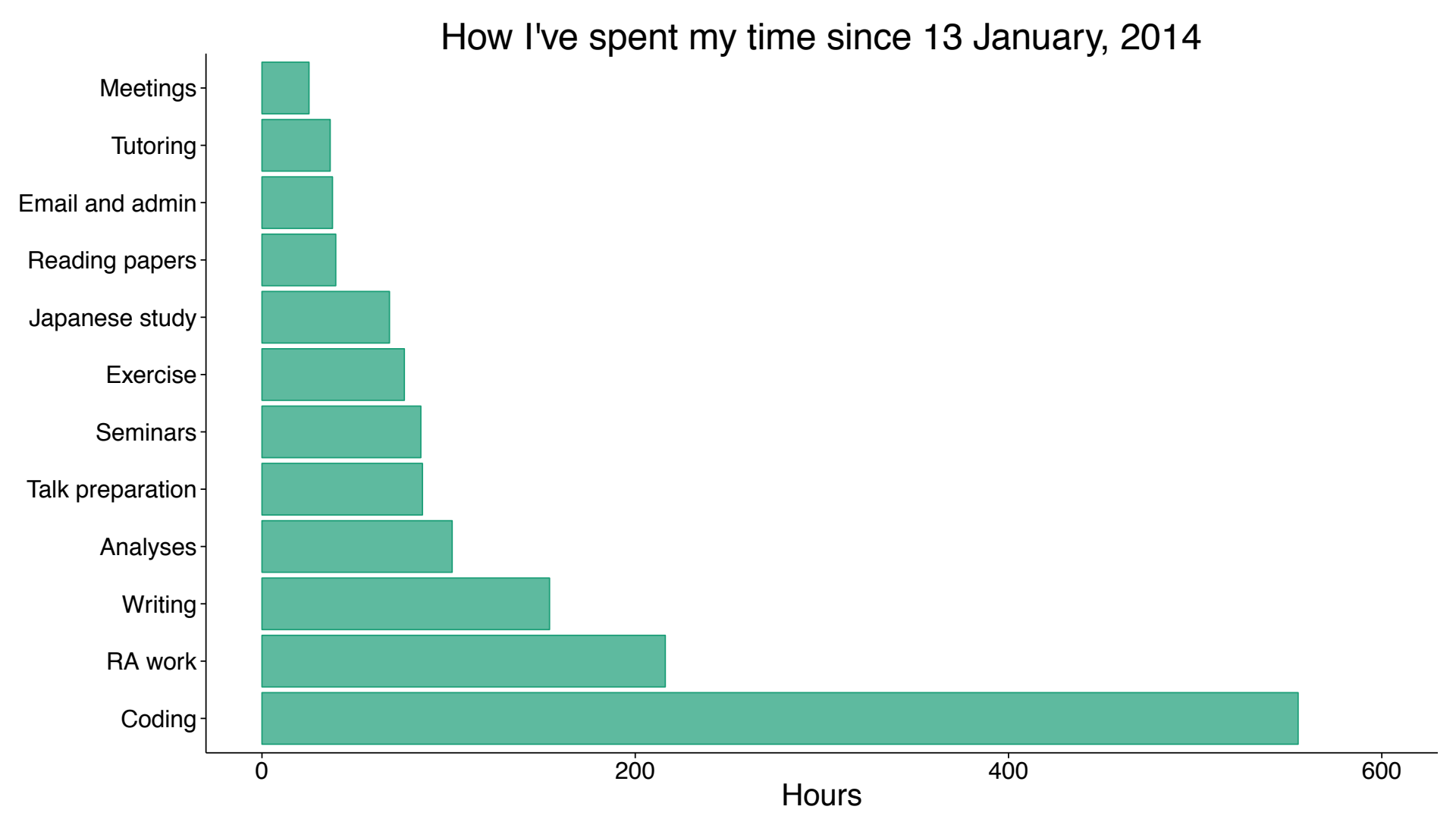
Peter Hickey

@PeteHaitch

15 September 2014

How I spend my time

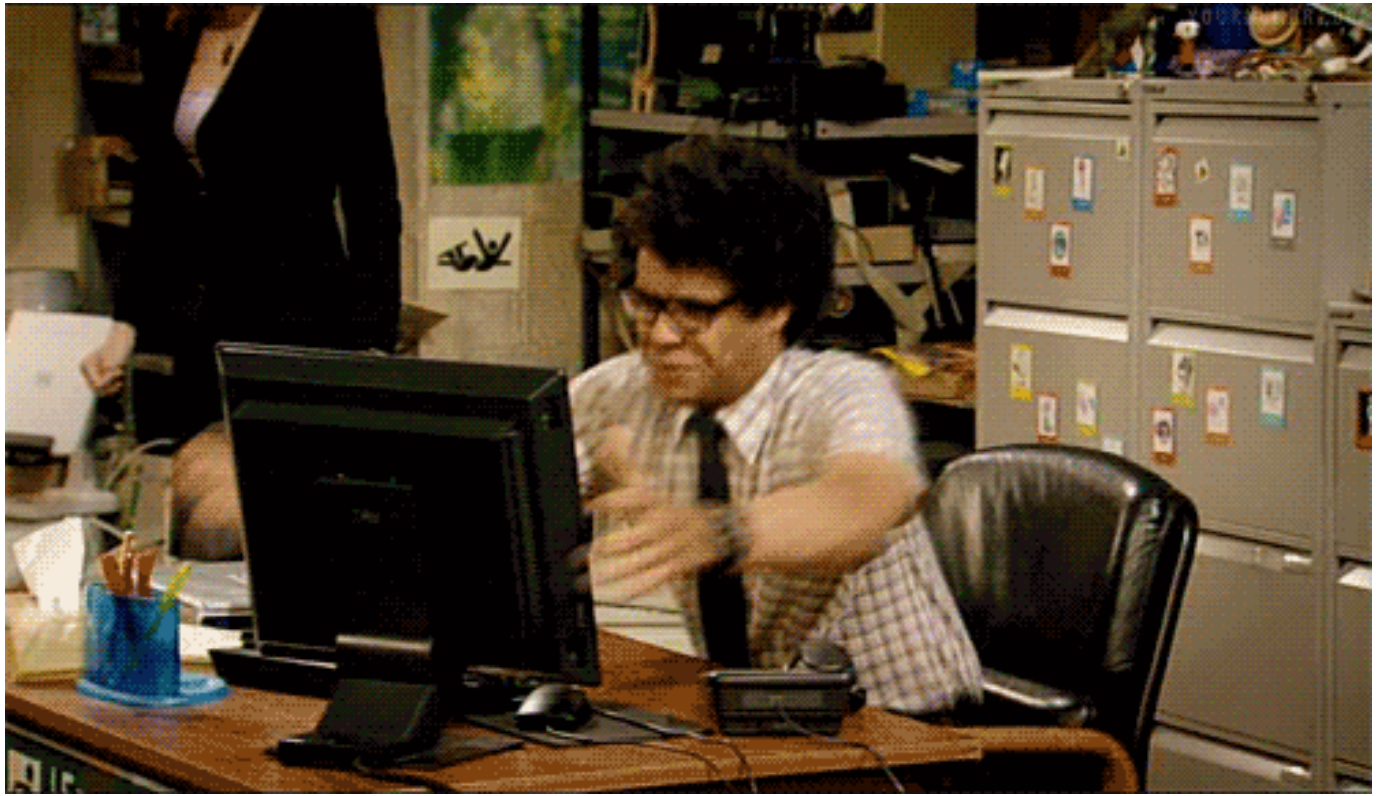
“Oh, you’re a statistician…”


How I've spent my time since 13 January, 2014
Meetings
Tutoring
Email and admin
Reading papers
Japanese study
Exercise
Seminars
Talk preparation
Analyses
Writing
RA work
Coding
0
200
400
600
Hours

How I spend my time

http://gifurl.com/gifs/1145

How I spend my time



http://wifflegif.com/gifs/511132-richard-ayoade-the-it-crowd-gif

Project aim

How to analyse whole-genome bisulfite-sequencing experiments to learn about DNA methylation?

Exploratory data analyses

https://www.flickr.com/photos/theloushe/4640871734

Developing methods

http://tstoaddicts.files.wordpress.com/2013/08/simpsons-questions.gif

Implementing methods

http://www.mhjaso.com/blog/wp-content/uploads/2014/06/Homer-Tries-to-Build-a-Barbeque-Set.jpg

Collaboration

http://cdn.amazinganimalstories.com/wp-content/uploads/2013/11/ATT00038.jpg

5mCs and one assay

ACGCGAAACGTTCTATCGG
TGCGCTTTGCAAGATAGCC

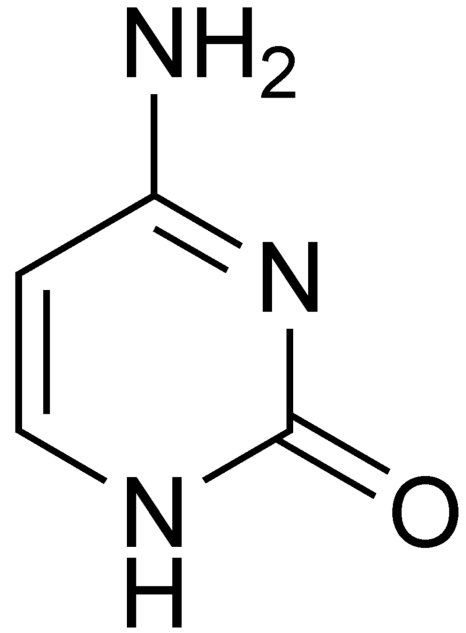
ACGCGAAACGTTCTATCGG
TGCGCTTTGCAAGATAGCC

m m m m

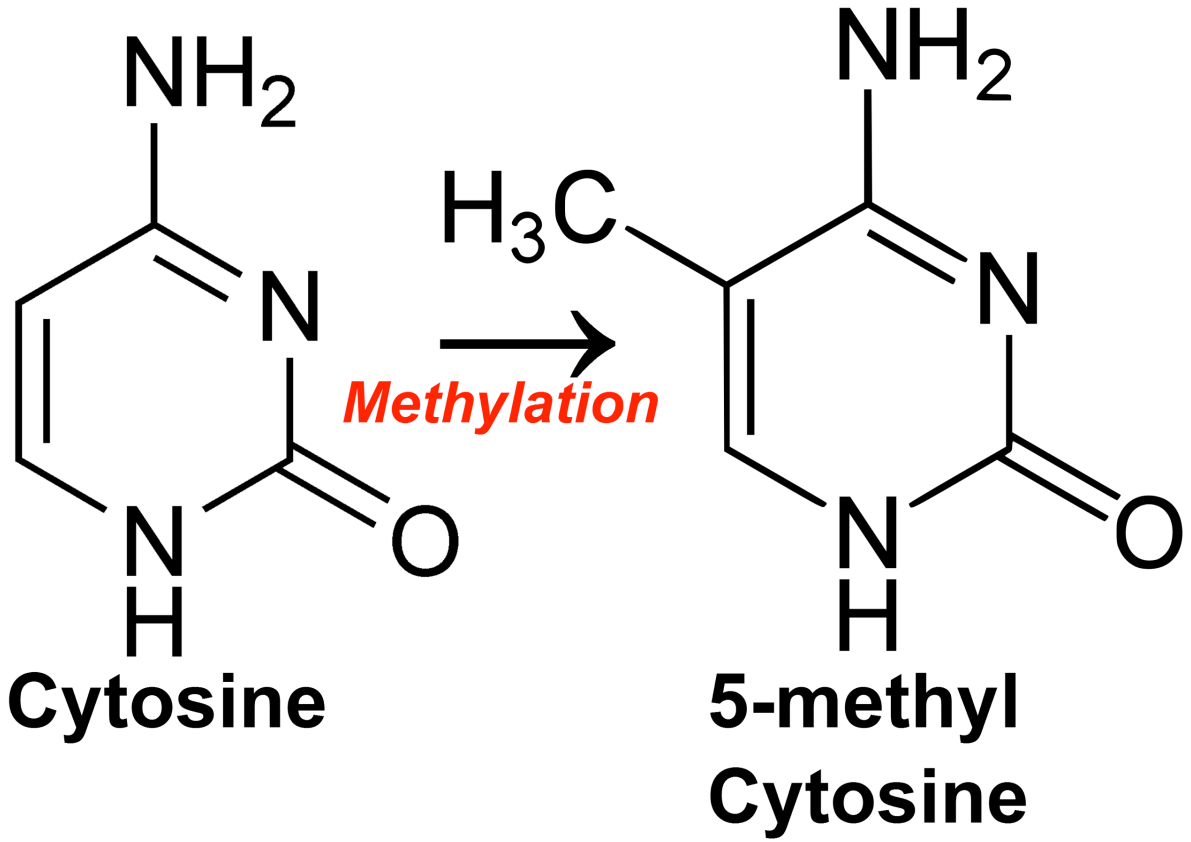
ACGCGAAACGTTCTATCGG

TGCGCTTTGCAAGATAGCC

m m m



Cytosine



ACGCGAAACGTTCTATCGG
TGCGCTTTGCAAGATAGCC

m m m m

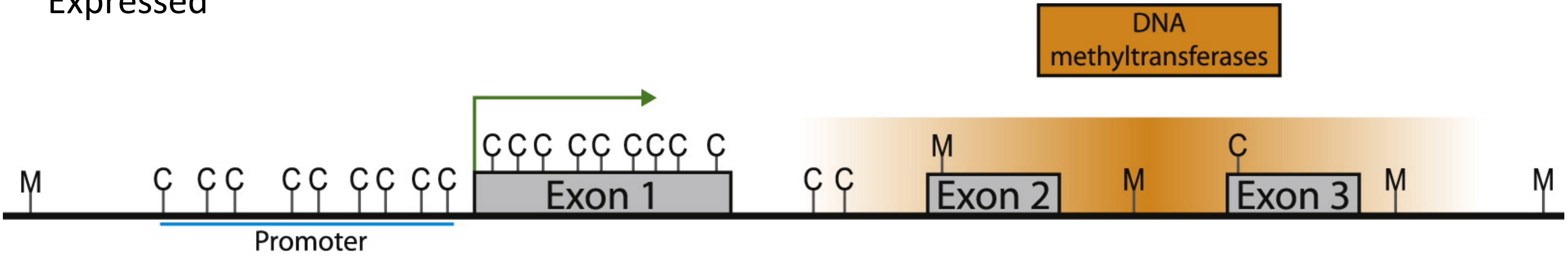
ACGCGAAACGTTCTATCGG

TGCGCTTTGCAAGATAGCC

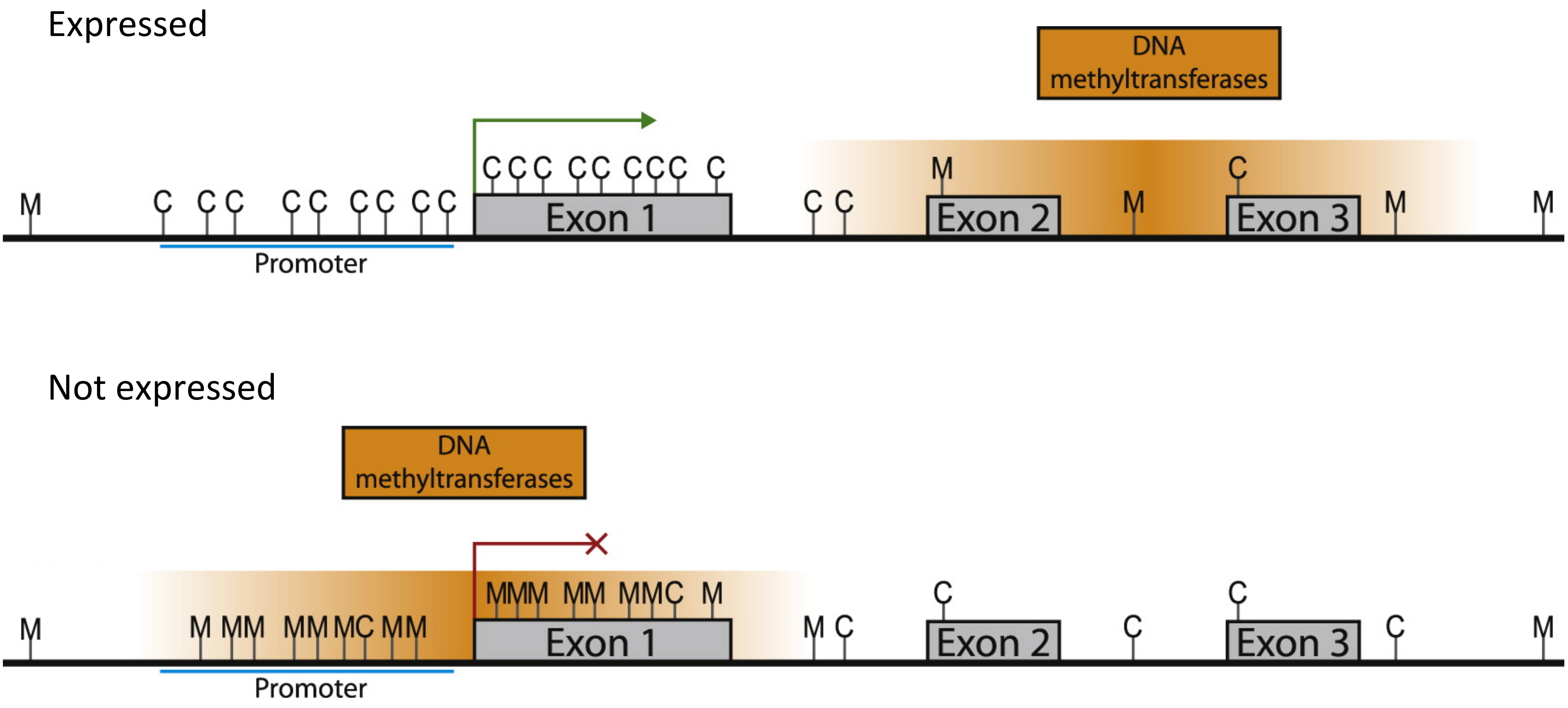
m m m

http://episona.com/wp-content/uploads/2014/01/Agouti-Mouse-1024x802.jpg

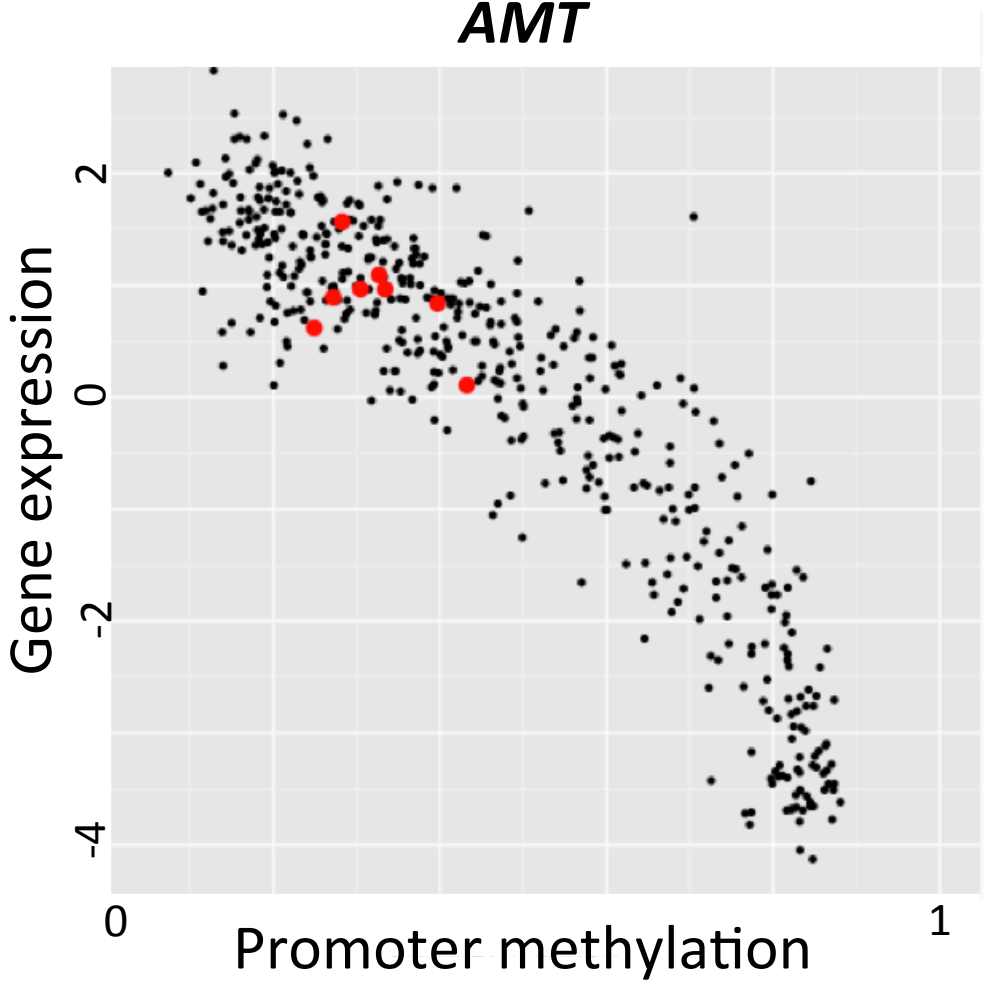
Expressed



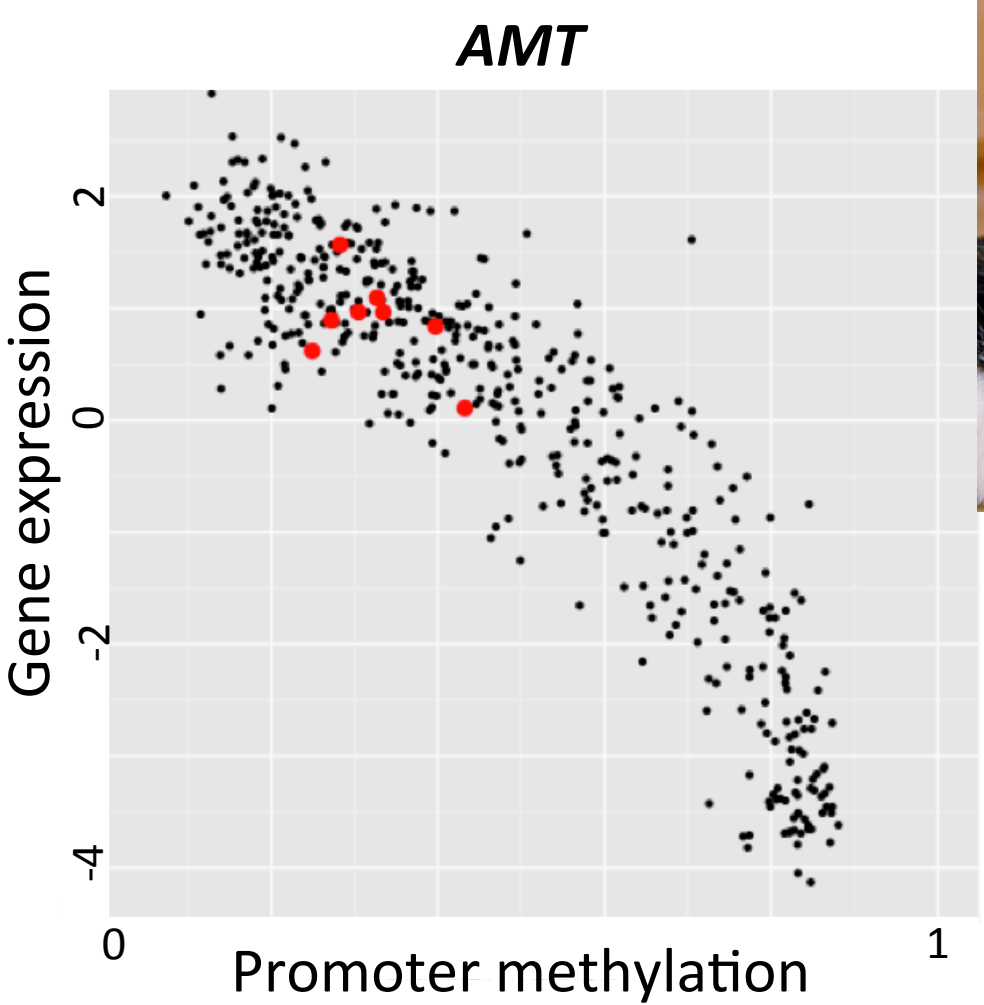
http://www.ous-research.no/kristensen/images/projects/Promoter_methylation.jpg



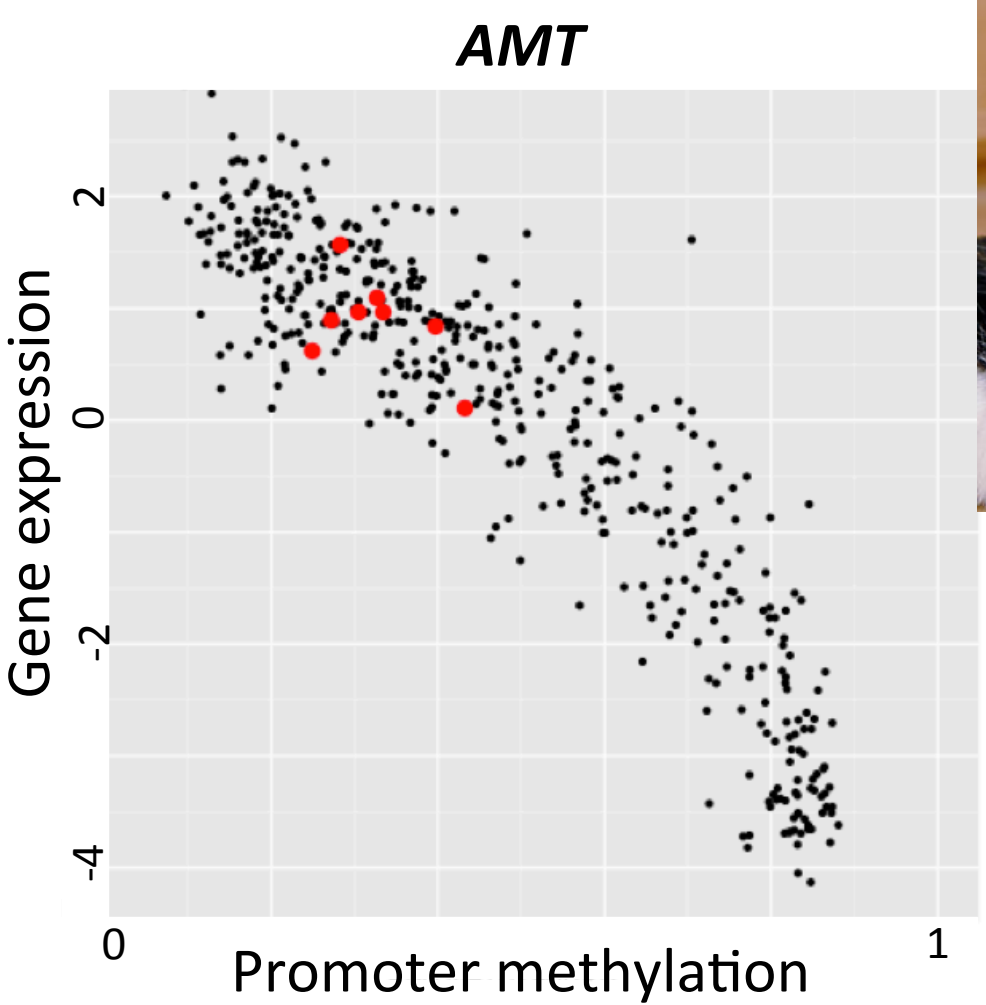
http://www.ous-research.no/kristensen/images/projects/Promoter_methylation.jpg



Cancer Genome Atlas Research Network. "Integrated genomic analyses of ovarian carcinoma." Nature 474.7353 (2011): 609-615.

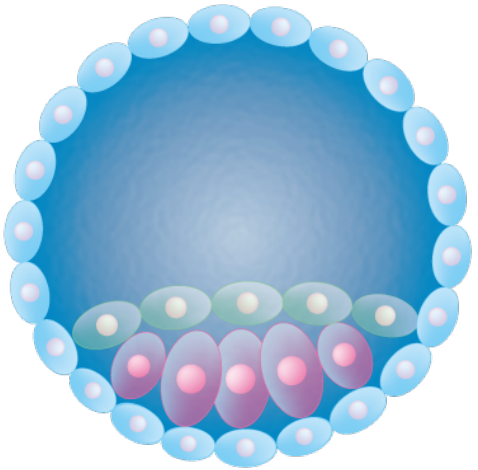


Cancer Genome Atlas Research Network. "Integrated genomic analyses of ovarian carcinoma." Nature 474.7353 (2011): 609-615.



Cancer Genome Atlas Research Network. "Integrated genomic analyses of ovarian carcinoma." Nature 474.7353 (2011): 609-615.

http://commons.wikimedia.org/wiki/File:Calico_cat_-_Phoebe.jpg#mediaviewer/File:Calico_cat_-_Phoebe.jpg



http://commons.wikimedia.org/wiki/File%3ABlastocyst_embryo.png

m m m m

ACGCGAAACGTTCTATCGG

TGCGCTTTGCAAGATAGCC

m m m

m m m m

ACGCGAAACGTTCTATCGG

ACGCGAAACGTTCTATCGG

ACGCGAAACGTTCTATCGG

+

PCR amplification

=

ACGCGAAACGTTCTATCGG

+

PCR amplification

=

ACGCGAAACGTTCTATCGG

ACGCGAAACGTTCTATCGG

ACGCGAAACGTTCTATCGG

+

Sodium bisulfite

=

ACGCGAAACGTTCTATCGG

+

Sodium bisulfite

=

ACGUGAAACGTTCTATCGG

ACGUGAAACGTTCTATCGG

ACGUGAAACGTTCTATCGG

+

PCR amplification

=

ACGUGAAACGTTCTATCGG

+

PCR amplification

=

ACGTGAAACGTTCTATCGG

ACGUGAAACGTTCTATCGG

+

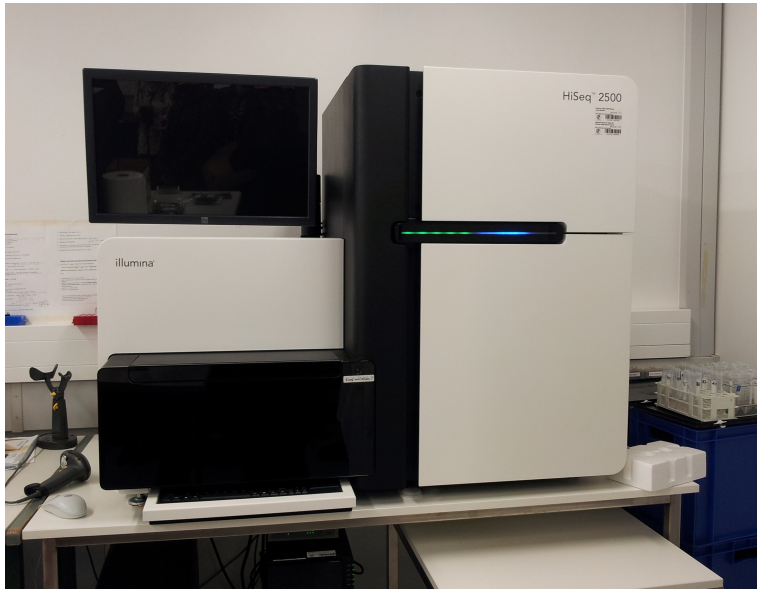
PCR amplification

=

ACGTGAAACGTTCTATCGG

Bisulfite treatment of DNA

+

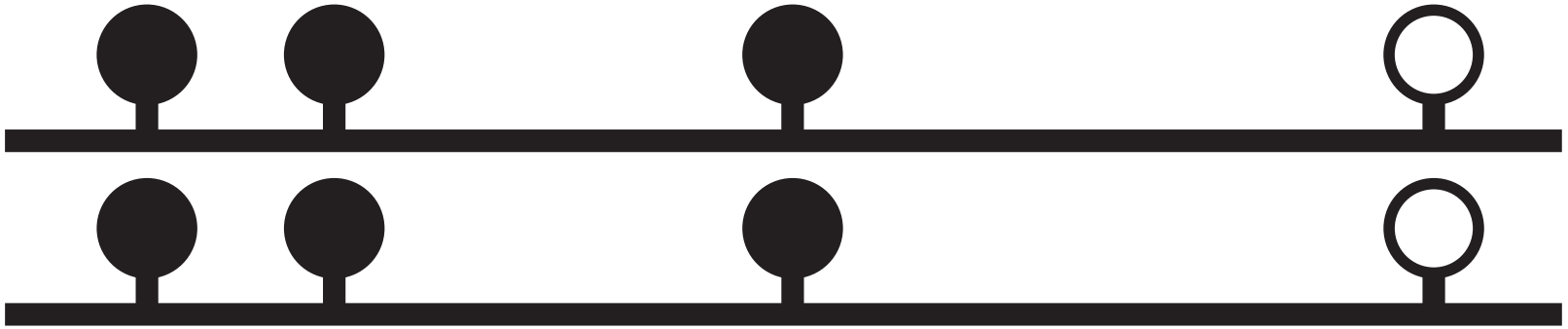


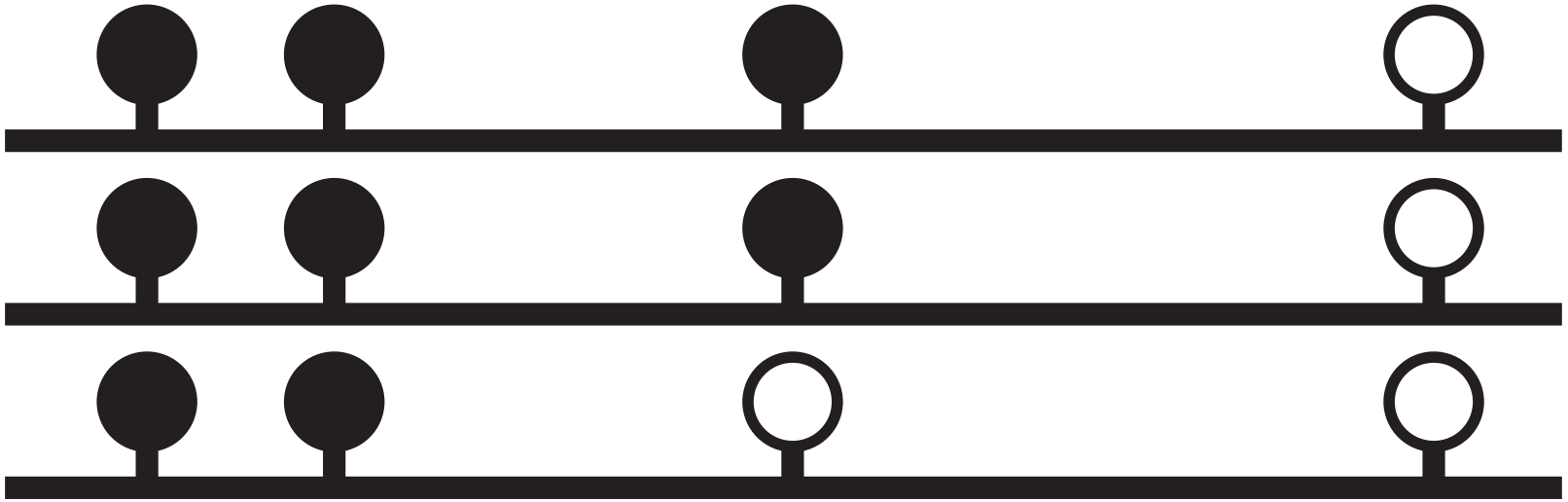
=

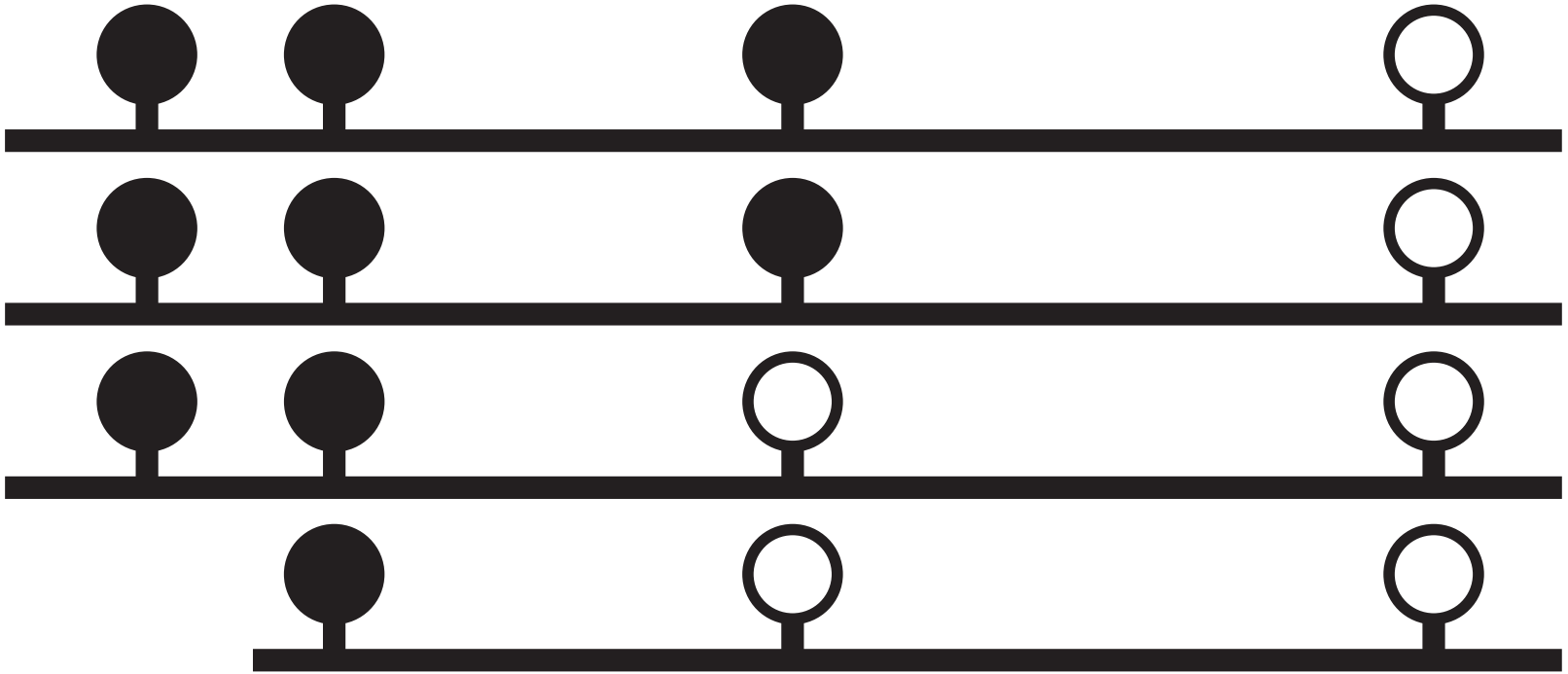
Whole-genome bisulfite-sequencing

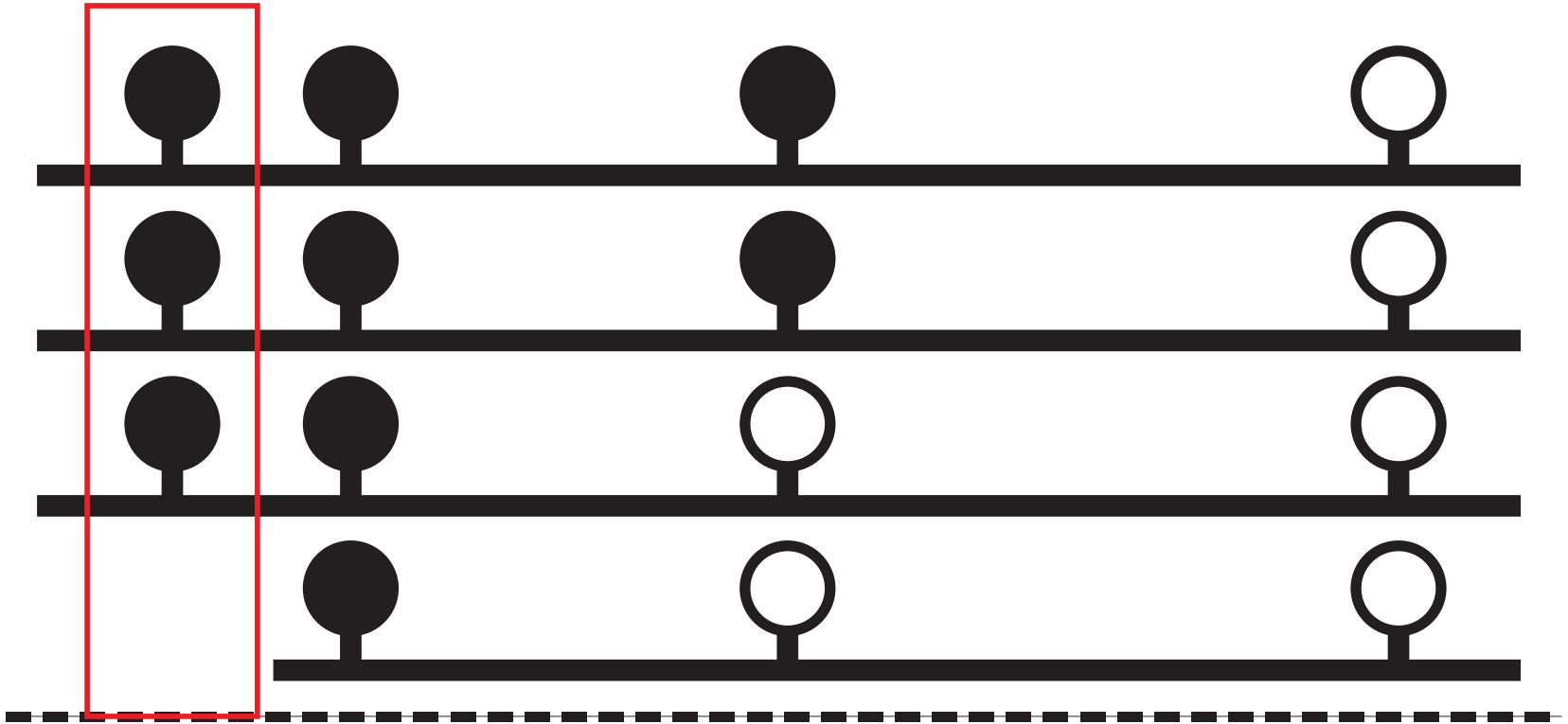
ACGCGAAACGTTCTATCG


ACGCGAAAACGTTCTATCG

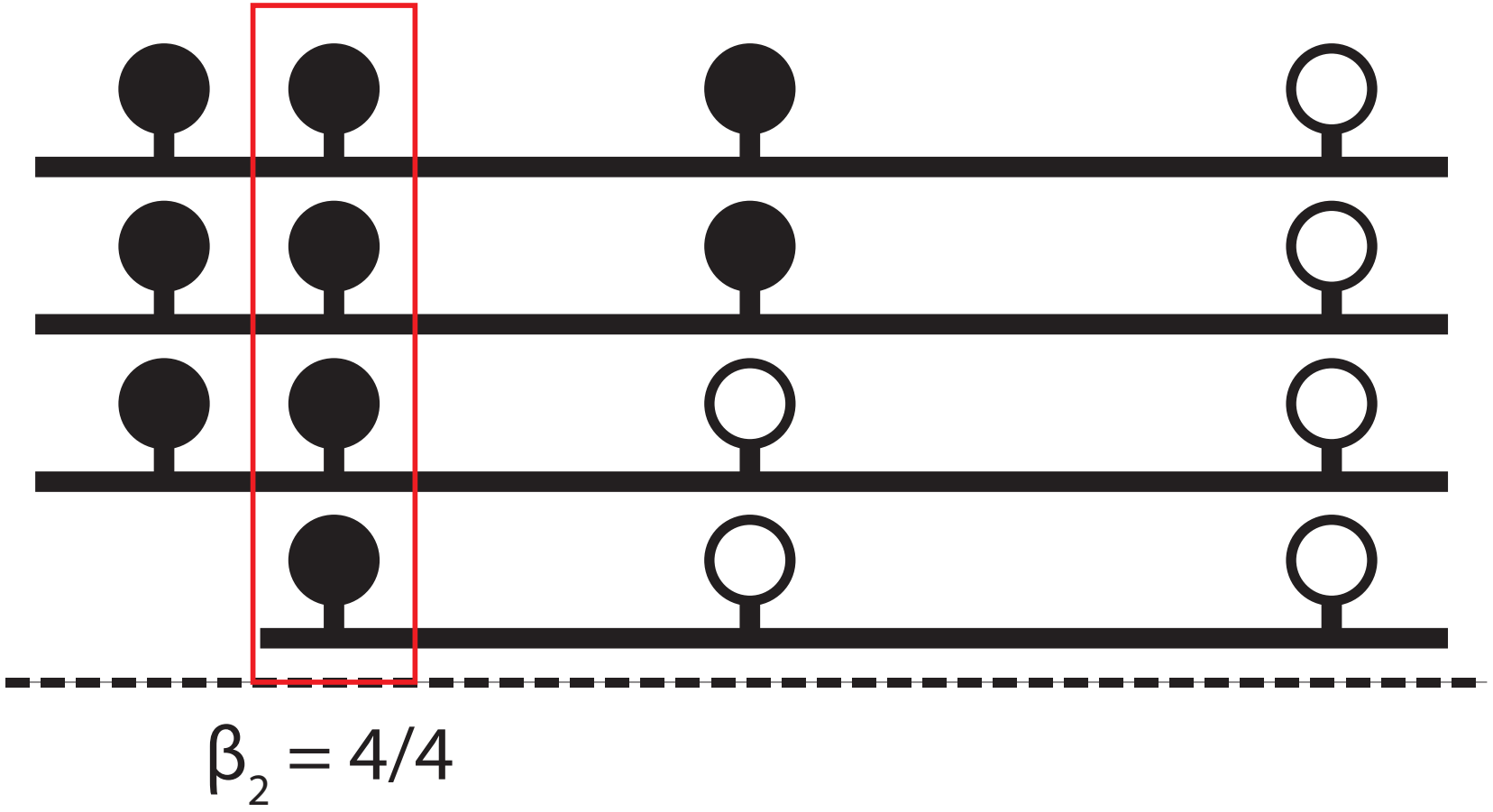


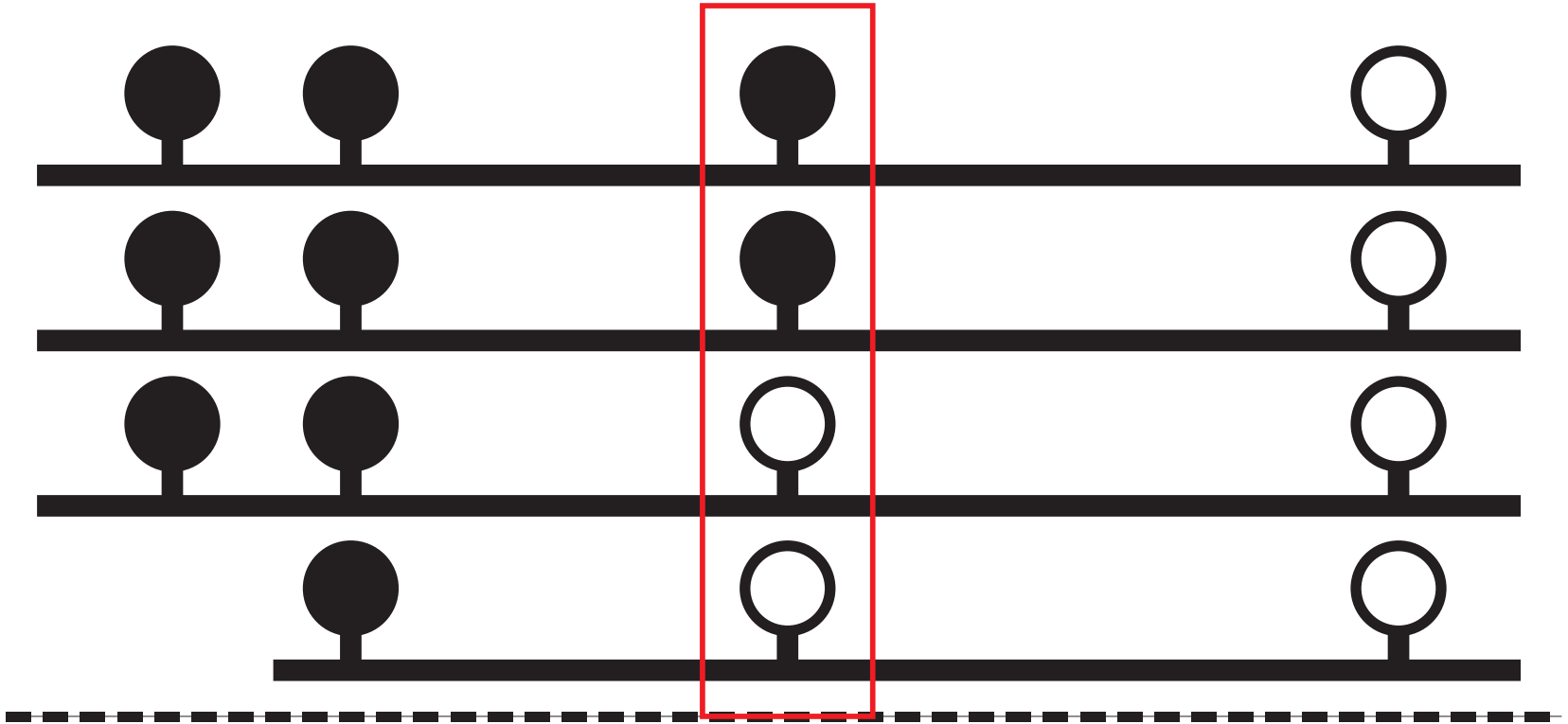




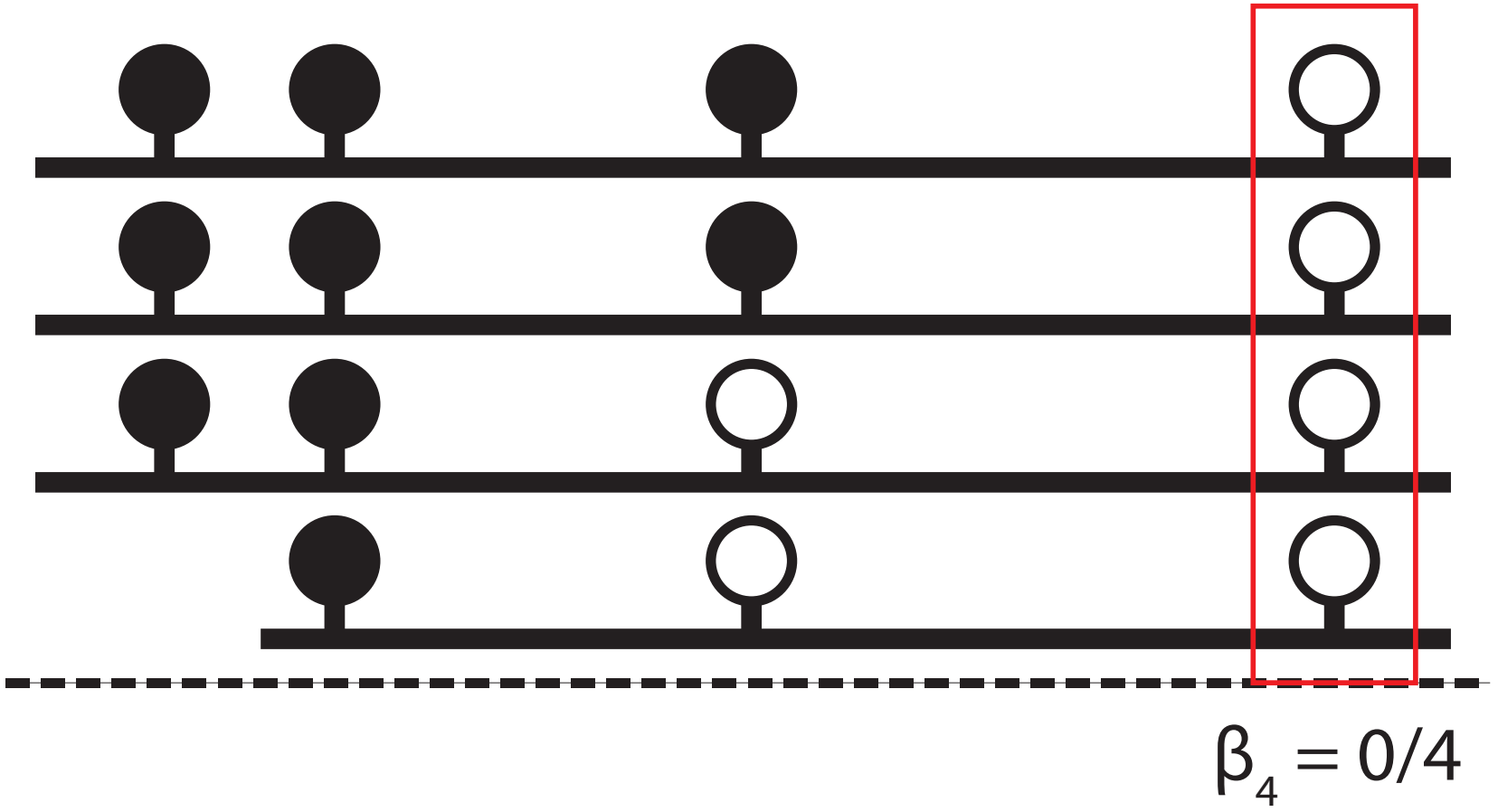


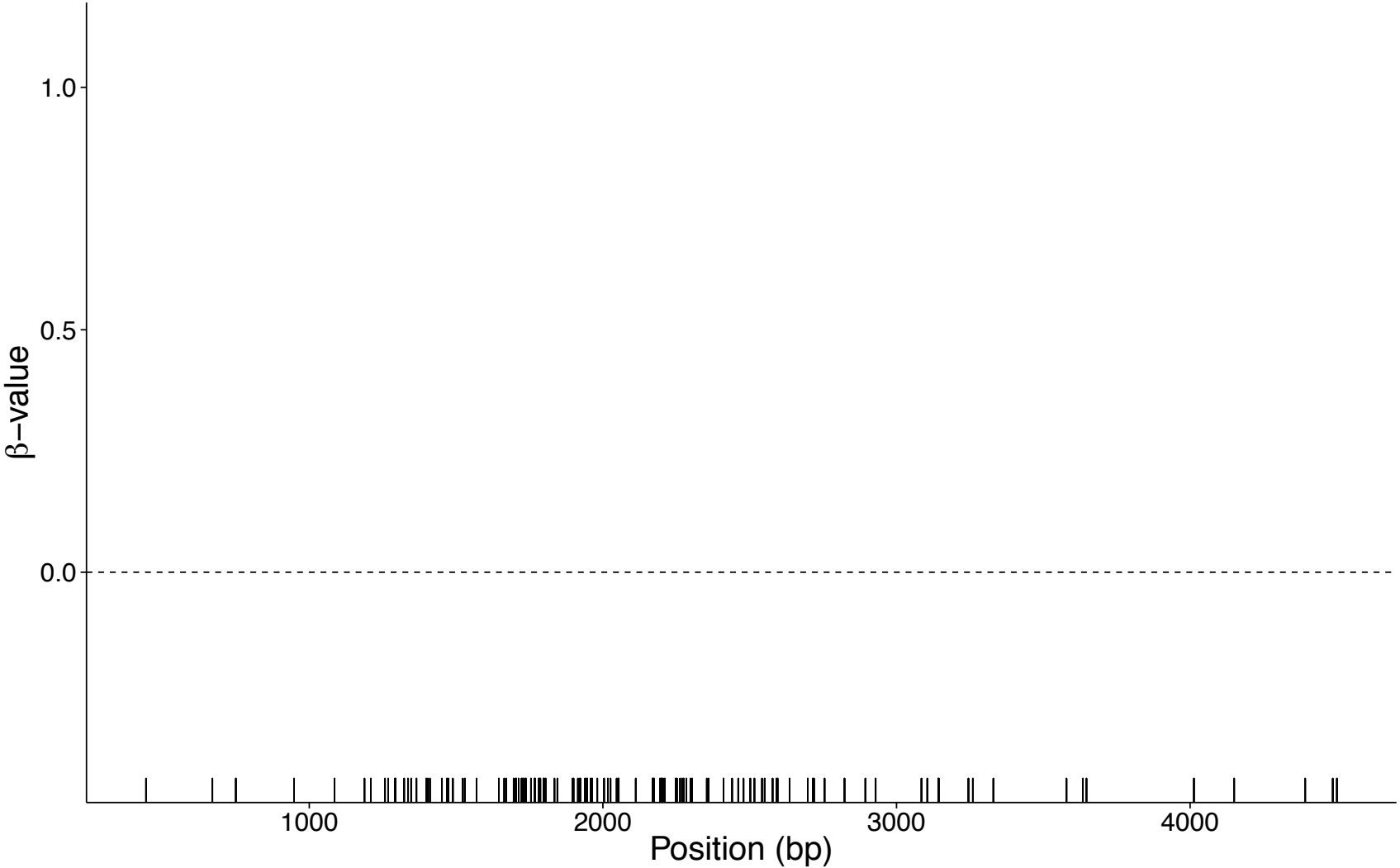
$\beta_1 = 3/3$

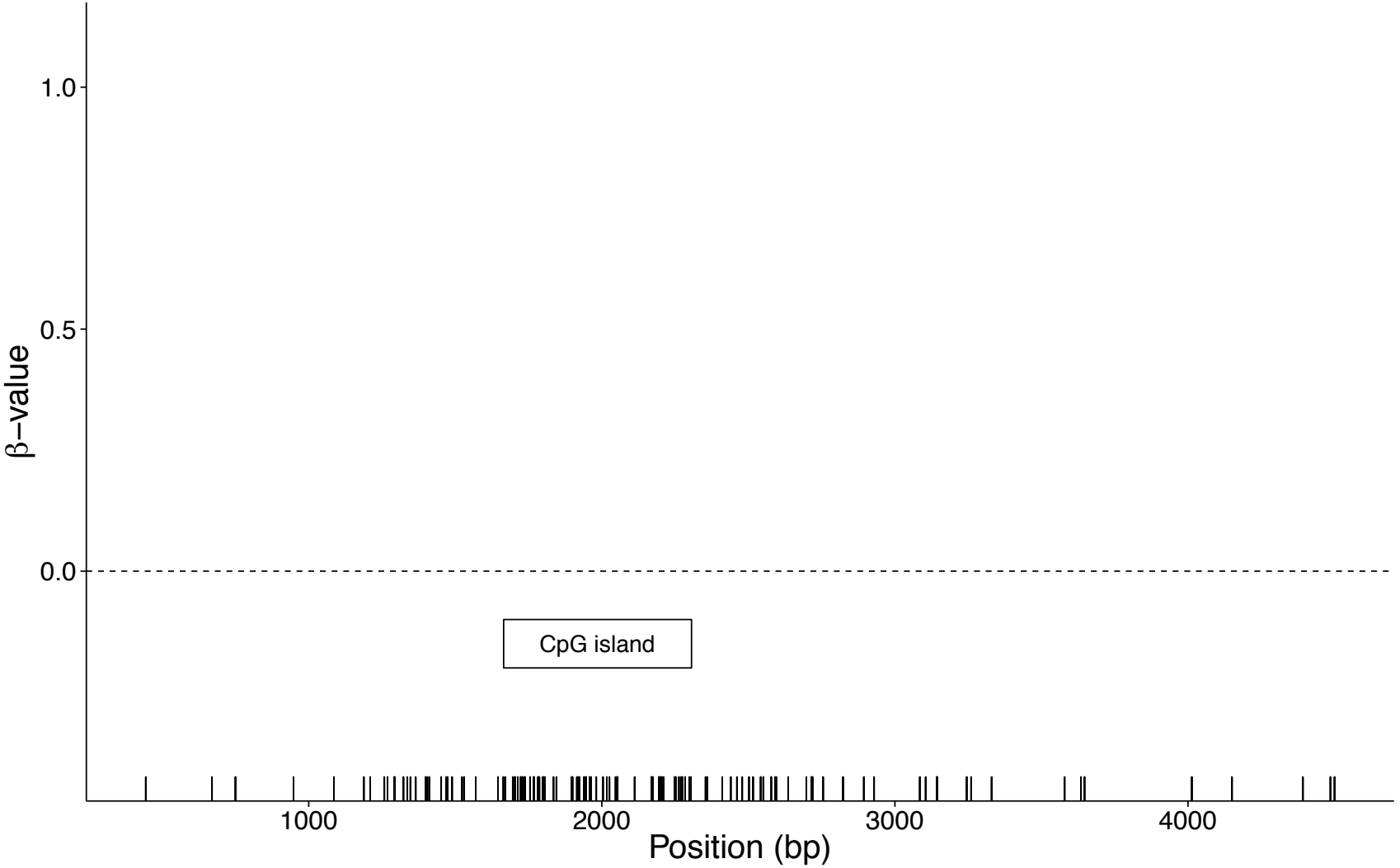

$\beta_2 = 4/4$

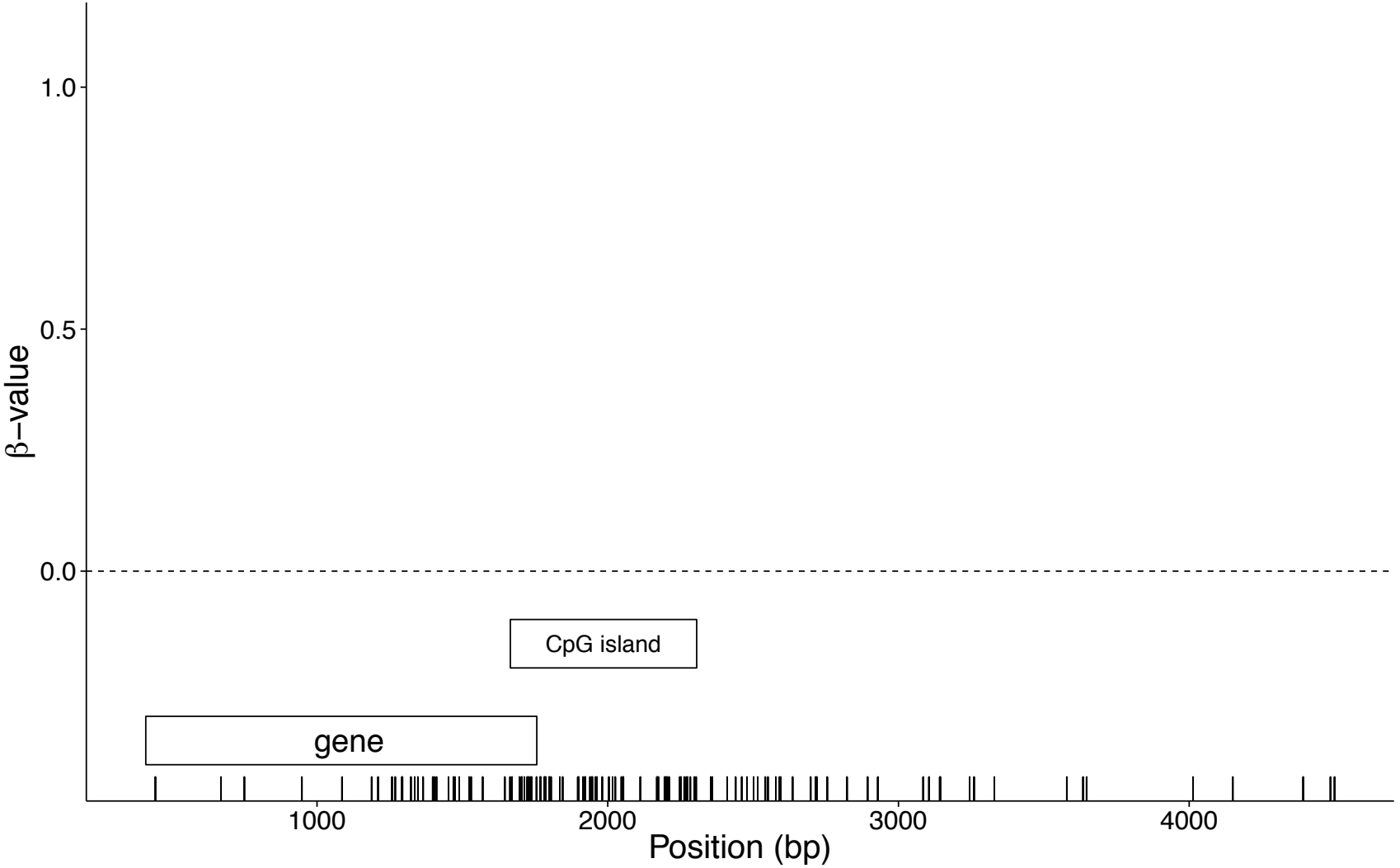


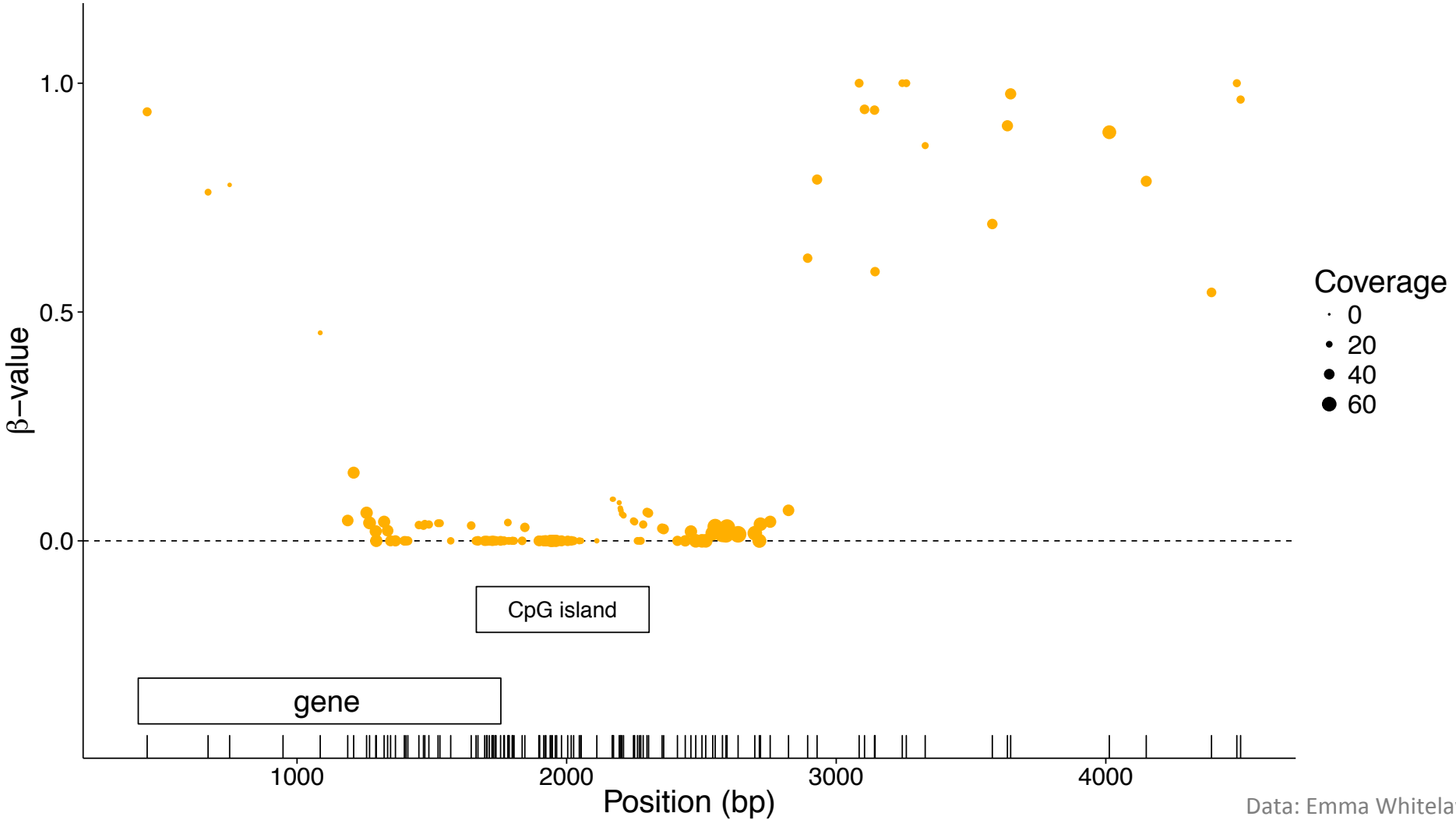
$\beta_3 = 2/4$


$\beta_4 = 0/4$


1.0
0.5
0.0
β−value
1000
2000
3000
4000
Position (bp)


1.0
0.5
0.0
β–value
CpG island
1000
2000
3000
4000
Position (bp)

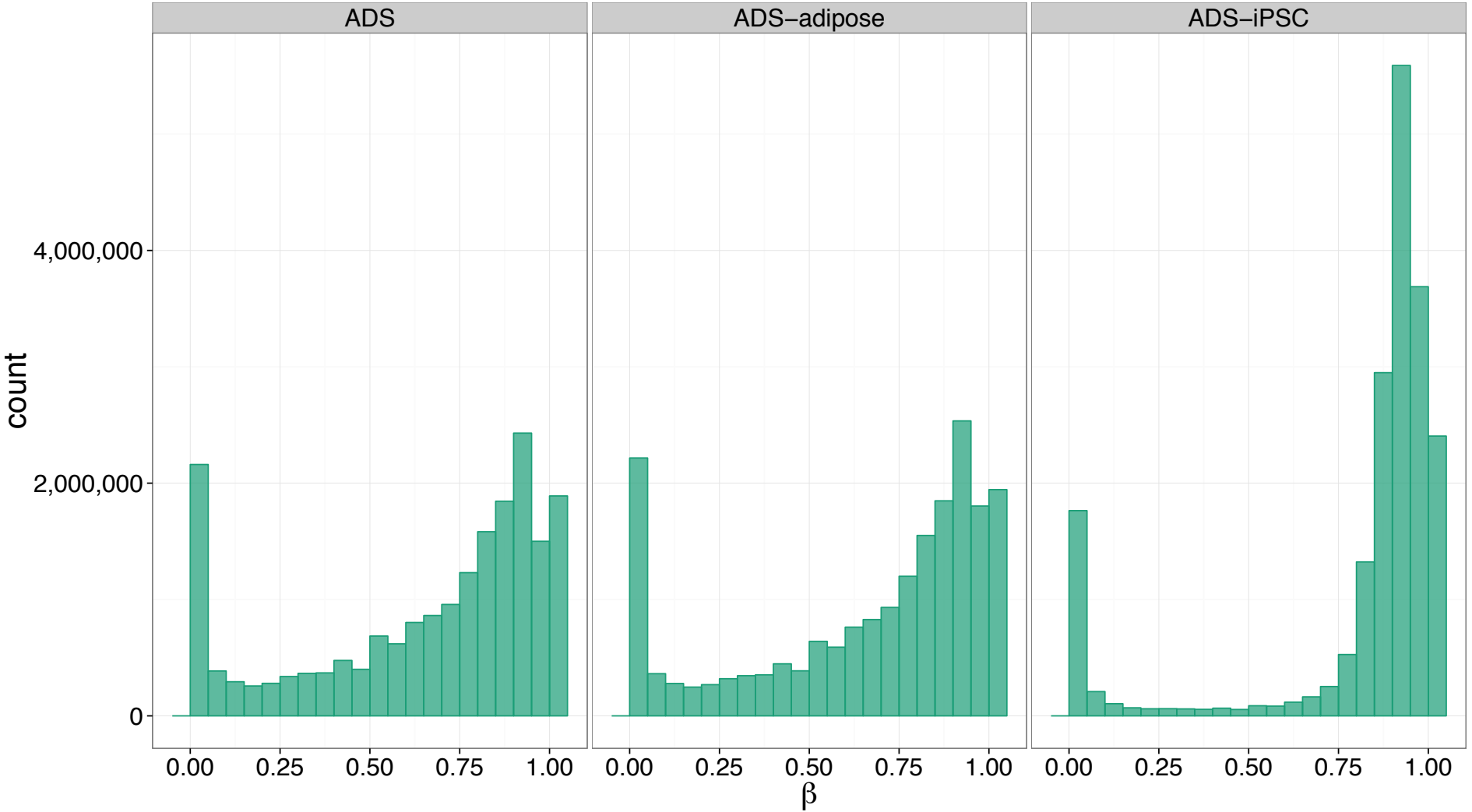

1.0
0.5
0.0
β−value
CpG island
gene
1000
2000
3000
4000
Position (bp)

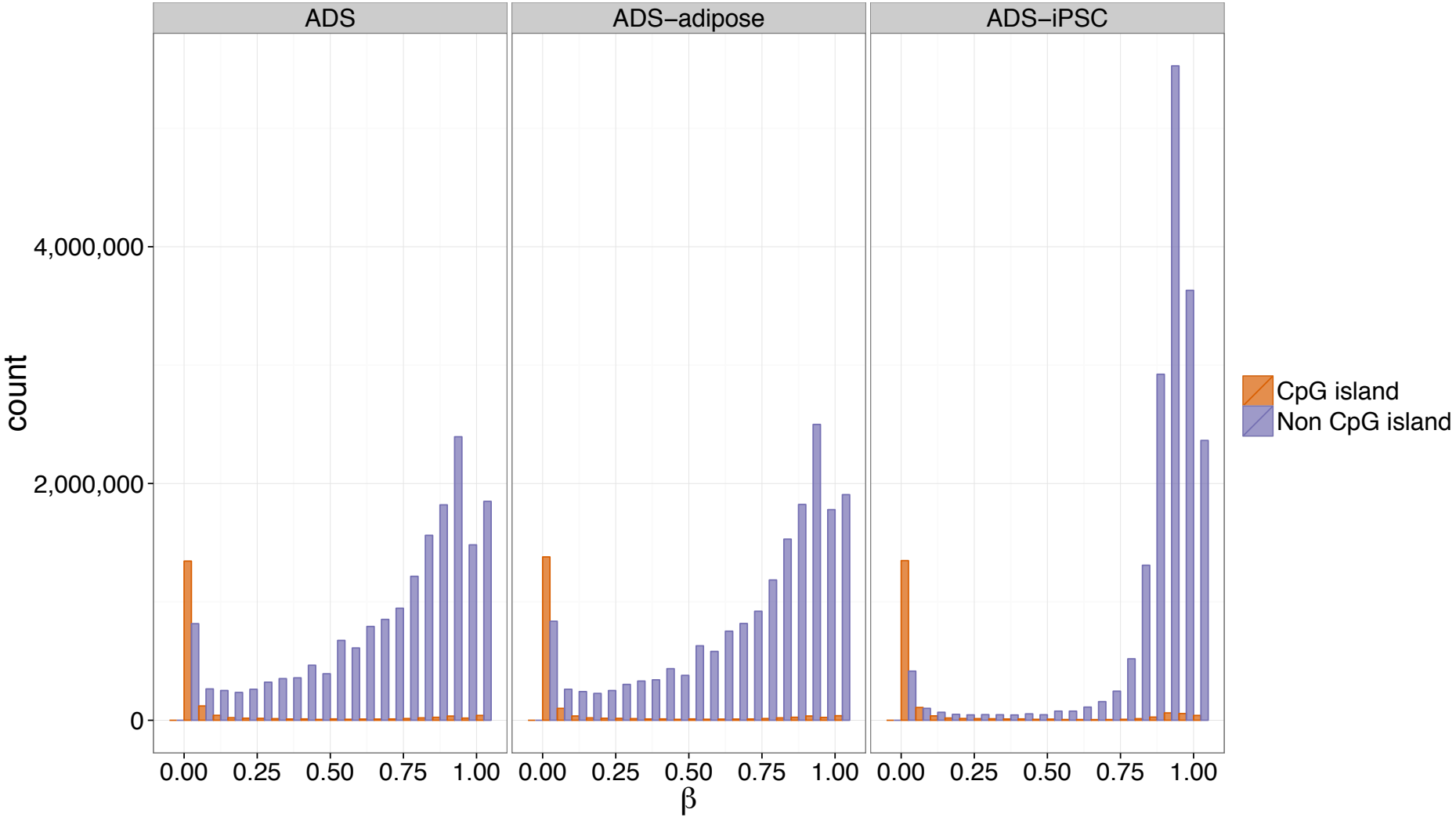

1.0
0.5
0.0
β–value
CpG island
gene
1000
2000
3000
4000
Position (bp)
Coverage
0
20
40
60
Data: Emma Whitelaw

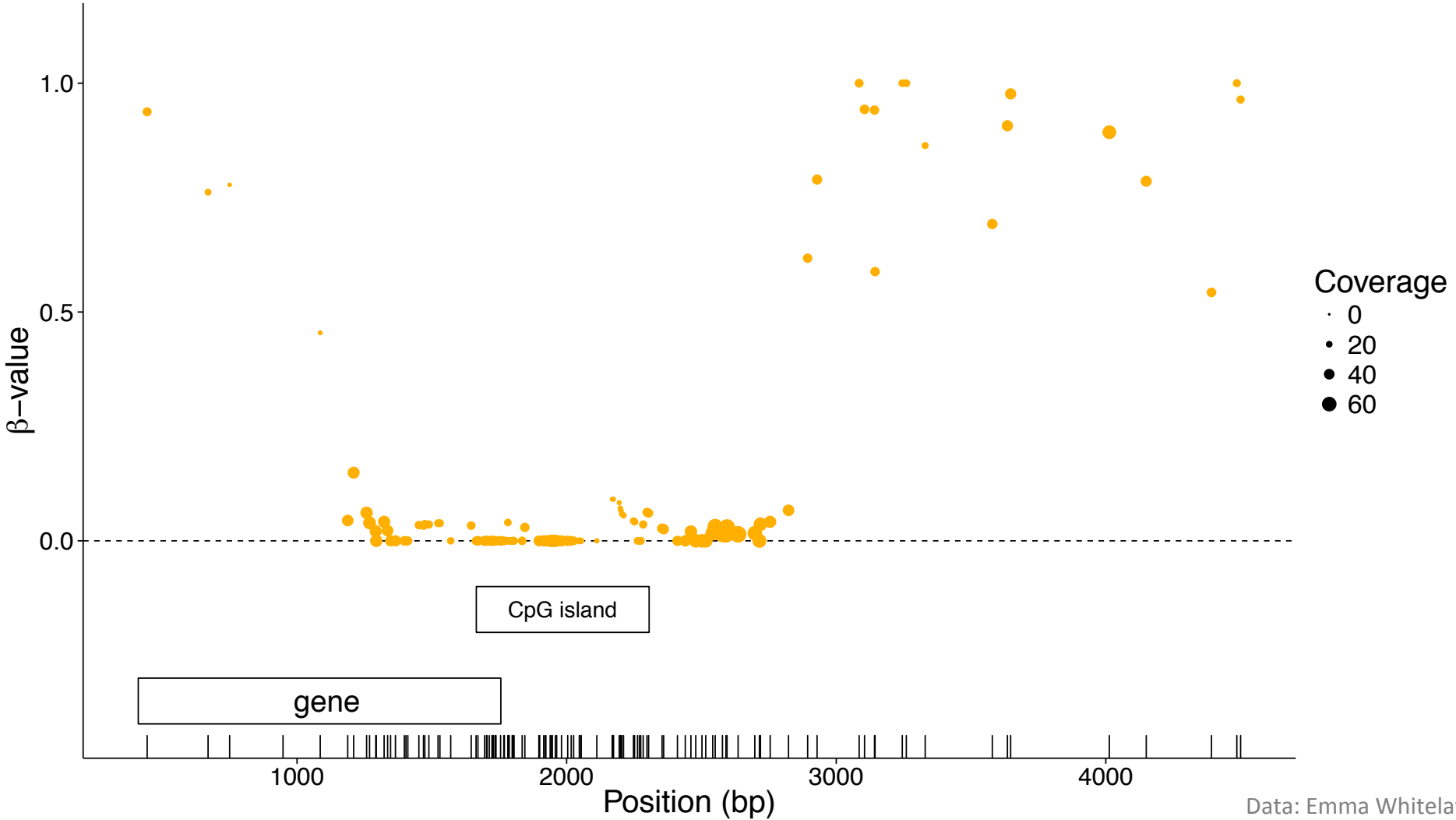
Lister data

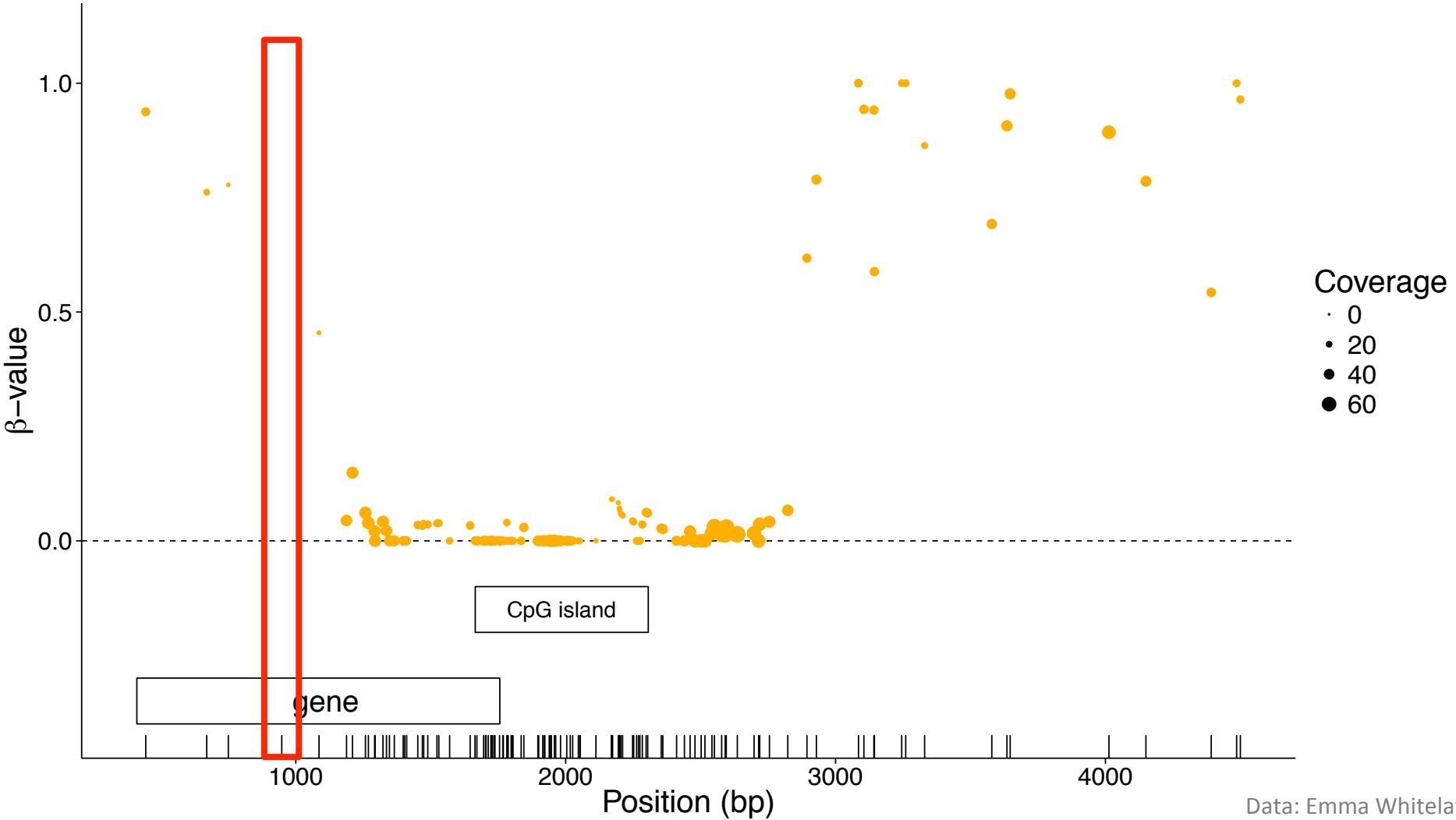
	ADS	*ADS-adipose*	*ADS-iPSC*
Organism	Human (female)	Human (female)	Human (female)
Cell type	Somatic	Somatic	Induced pluripotent stem cell (iPSC)
Description	Adipose	Adipocytes derived from *ADS*	iPSC line derived from *ADS*
Sequencing	75 bp paired-end	75 bp paired-end	75 bp paired-end
Average coverage	23×	24×	26×

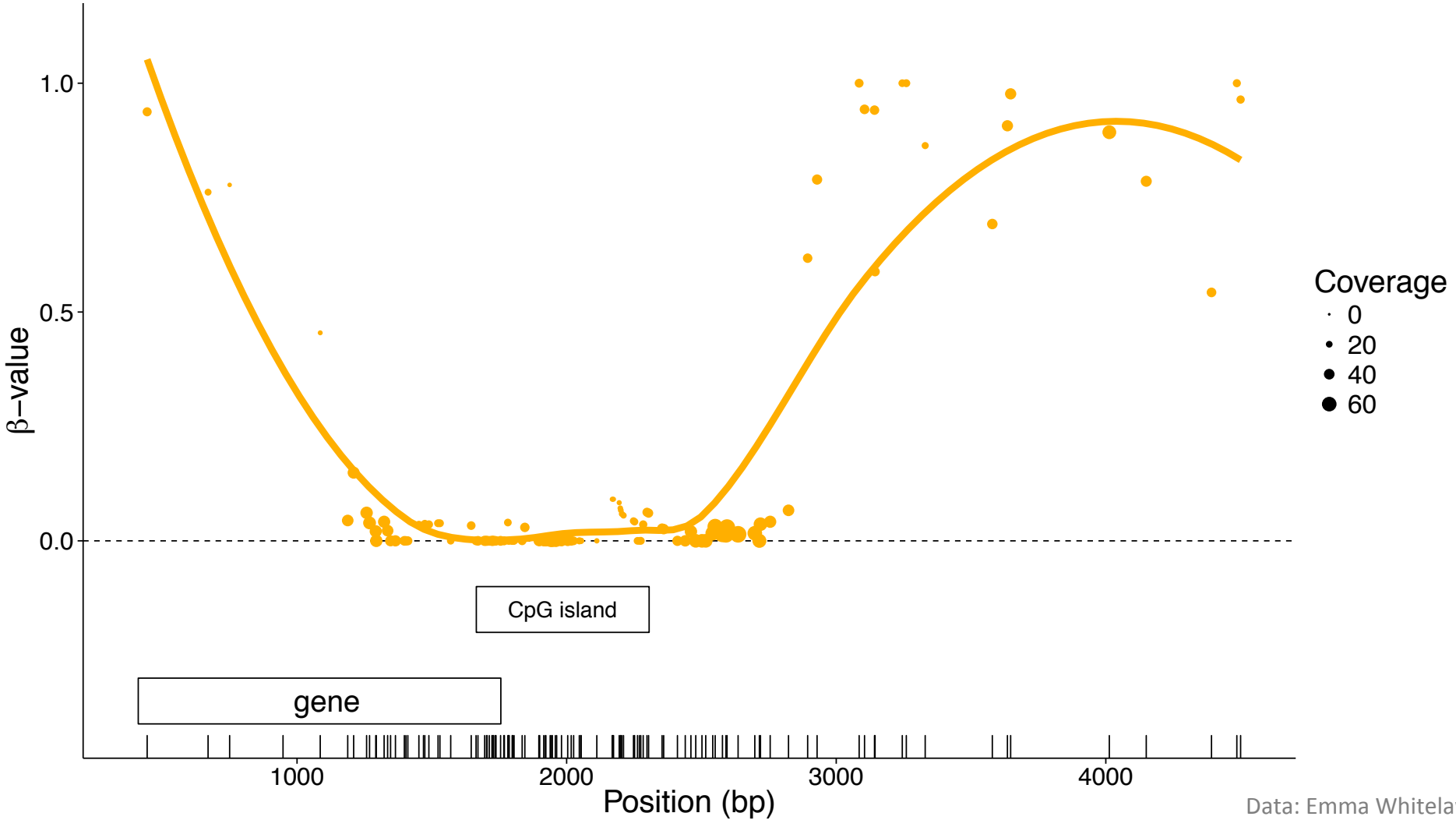
Lister, Ryan, et al. "Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells." Nature 471.7336 (2011): 68-73.


ADS
ADS−adipose
ADS−iPSC
count
4,000,000
2,000,000
0
0.00
0.25
0.50
0.75
1.00
β

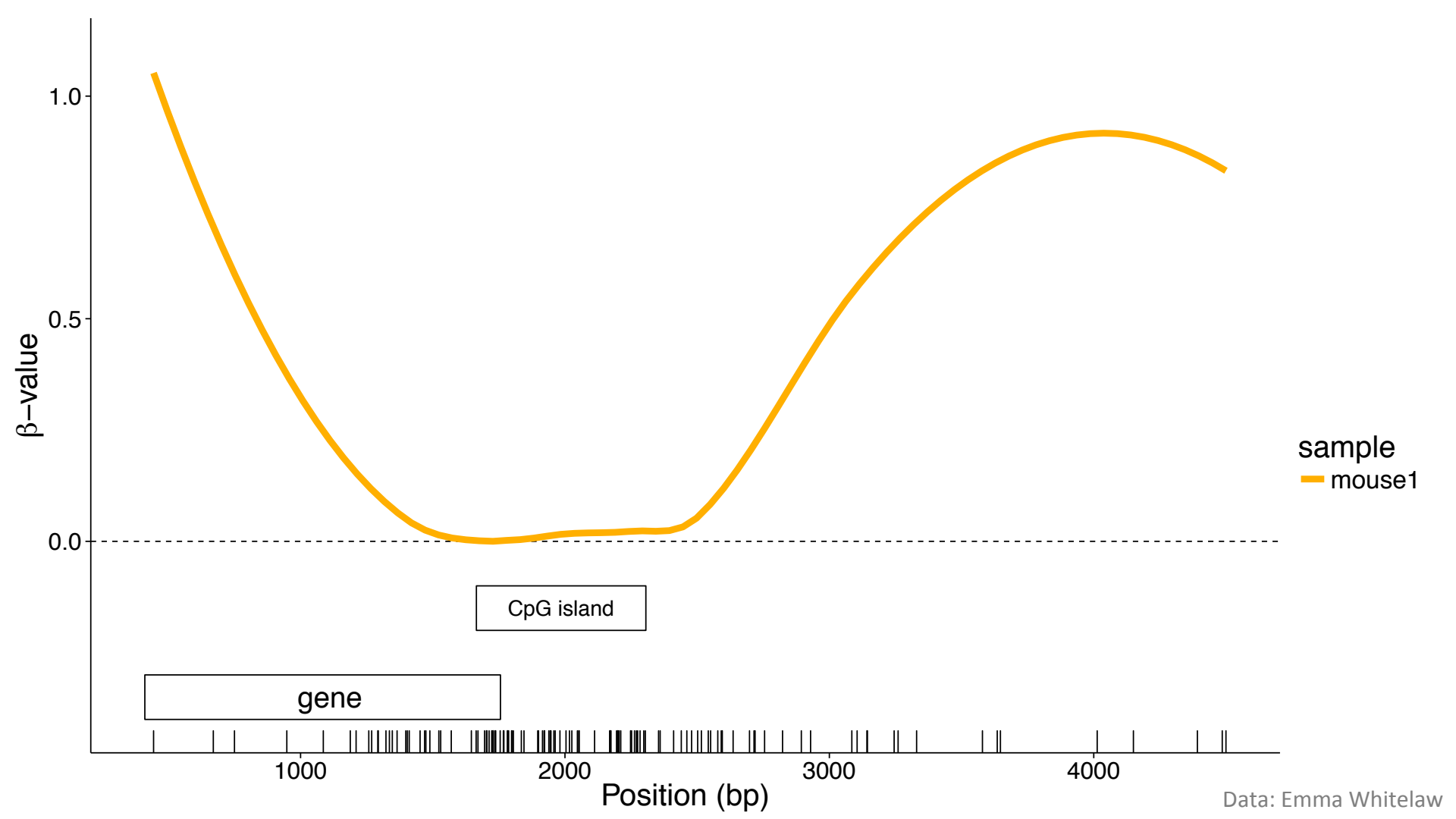

ADS
ADS−adipose
ADS−iPSC
count
4,000,000
2,000,000
0
0.00
0.25
0.50
0.75
1.00
β
CpG island
Non CpG island

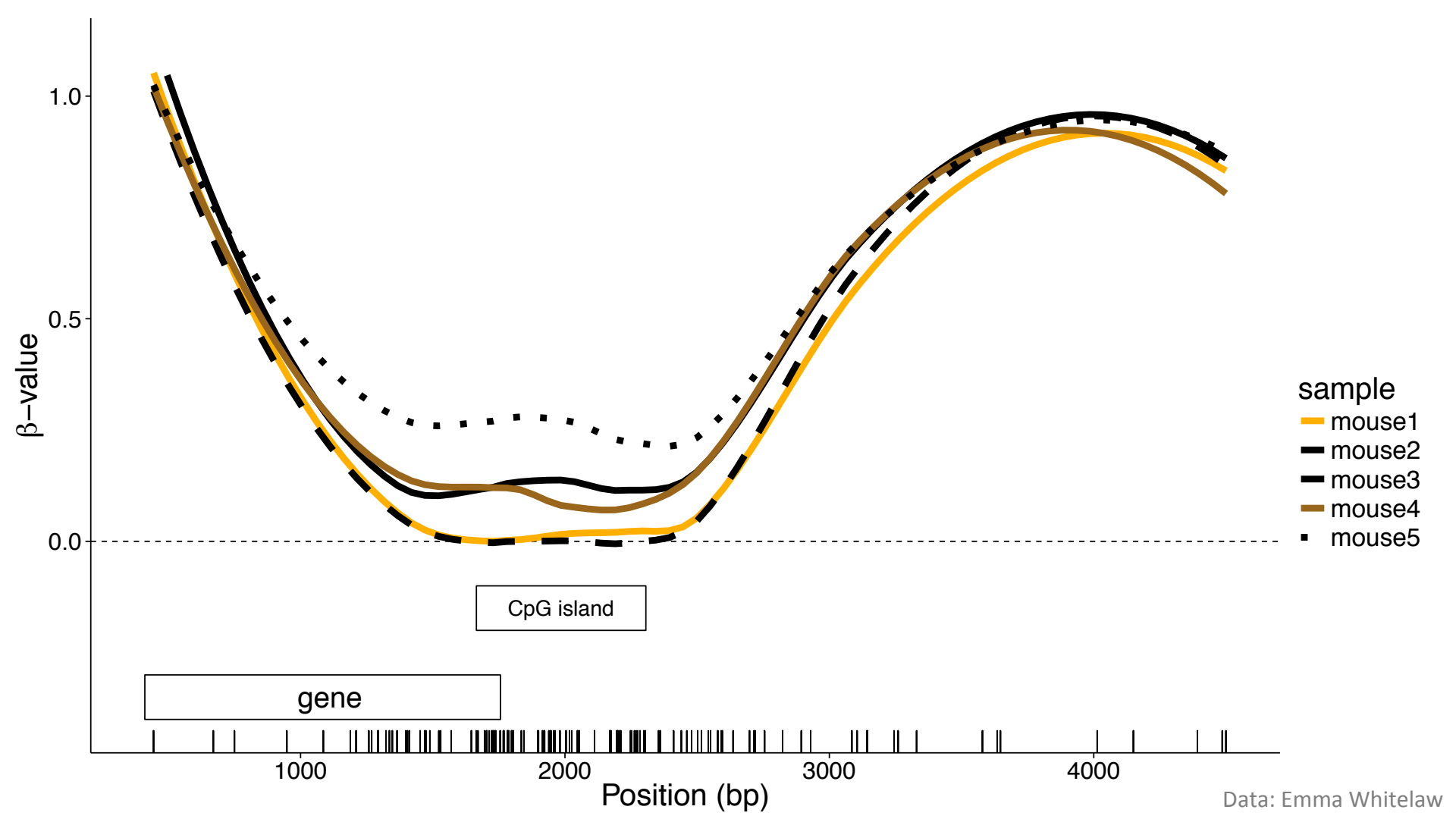

1.0
0.5
0.0
β–value
CpG island
gene
1000
2000
3000
4000
Position (bp)
Coverage
0
20
40
60
Data: Emma Whitelaw

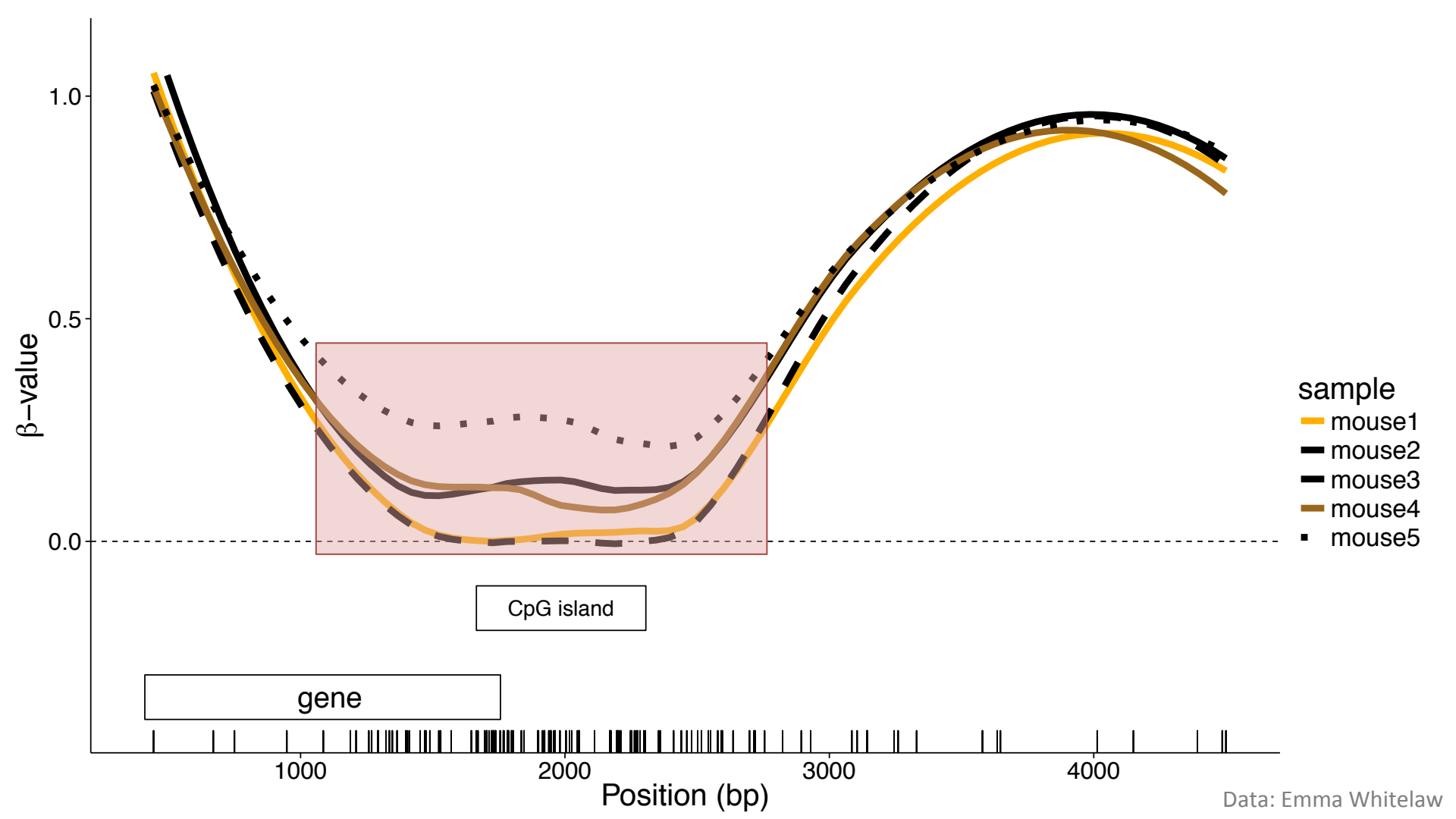

1.0
0.5
0.0
β–value
CpG island
gene
1000
2000
3000
4000
Position (bp)
Coverage
0
20
40
60
Data: Emma Whitelaw



Data: Emma Whitelaw


1.0
0.5
0.0
β−value
CpG island
gene
1000
2000
3000
4000
Position (bp)
sample
mouse1
Data: Emma Whitelaw


1.0
0.5
0.0
β−value
CpG island
gene
1000
2000
3000
4000
Position (bp)
sample
mouse1
mouse2
mouse3
mouse4
mouse5
Data: Emma Whitelaw


1.0
0.5
0.0
β−value
CpG island
gene
1000
2000
3000
4000
Position (bp)
sample
mouse1
mouse2
mouse3
mouse4
mouse5
Data: Emma Whitelaw

Co-methylation

Co-methylation = co-occurence

“The presence of methylation over a stretch of neighboring CpG positions”

Schatz, Philipp, Dimo Dietrich, and Matthias Schuster. "Rapid analysis of CpG methylation patterns using RNase T1 cleavage and MALDI-TOF." Nucleic acids research 32.21 (2004): e167-e167.

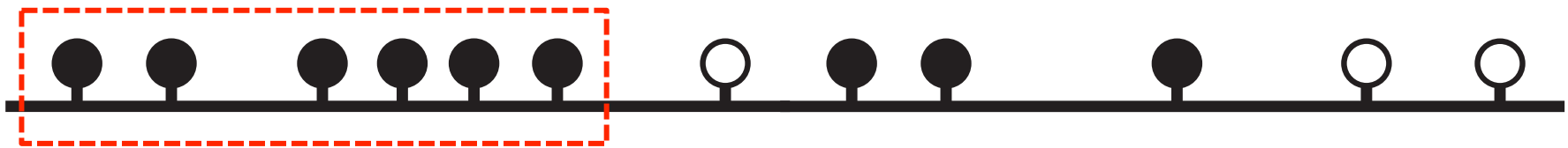
Co-methylation = co-occurence

“The presence of methylation over a stretch of neighboring CpG positions”

Schatz, Philipp, Dimo Dietrich, and Matthias Schuster. "Rapid analysis of CpG methylation patterns using RNase T1 cleavage and MALDI-TOF." Nucleic acids research 32.21 (2004): e167-e167.

Co-methylation = co-occurence

“The presence of methylation over a stretch of neighboring CpG positions”



Schatz, Philipp, Dimo Dietrich, and Matthias Schuster. "Rapid analysis of CpG methylation patterns using RNase T1 cleavage and MALDI-TOF." Nucleic acids research 32.21 (2004): e167-e167.

Co-methylation = correlation

“The relationship between the degree of methylation over distance”

Eckhardt, Florian, et al. "DNA methylation profiling of human chromosomes 6, 20 and 22." Nature genetics 38.12 (2006): 1378-1385.

Co-methylation = correlation

"The relationship between the degree of methylation over distance"

Eckhardt, Florian, et al. "DNA methylation profiling of human chromosomes 6, 20 and 22." Nature genetics 38.12 (2006): 1378-1385.

Co-methylation = correlation

"The relationship between the degree of methylation over distance"

1. Within-fragment co-methylation
2. Correlation of β-values

Eckhardt, Florian, et al. "DNA methylation profiling of human chromosomes 6, 20 and 22." Nature genetics 38.12 (2006): 1378-1385.

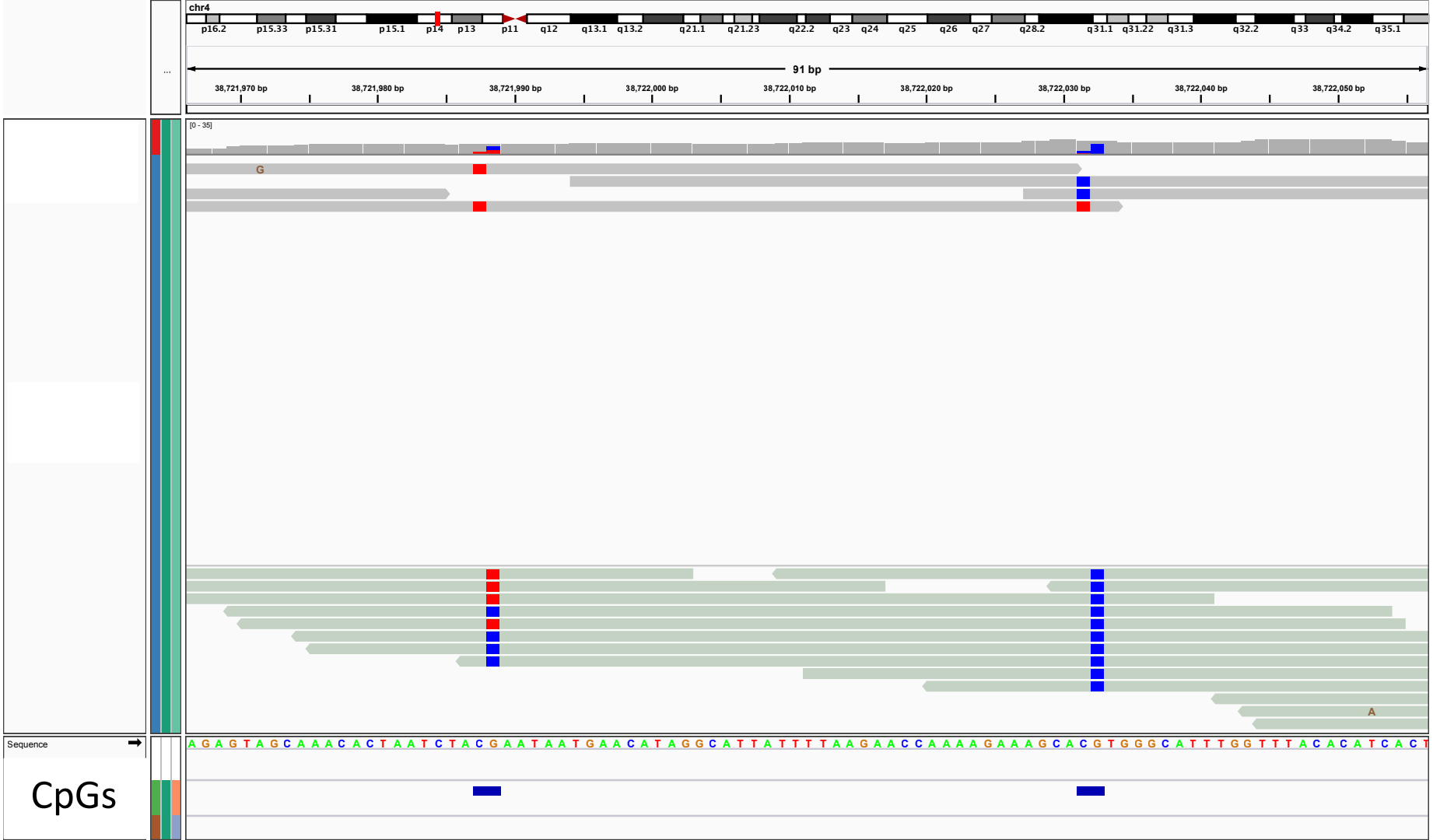
Co-methylation = correlation

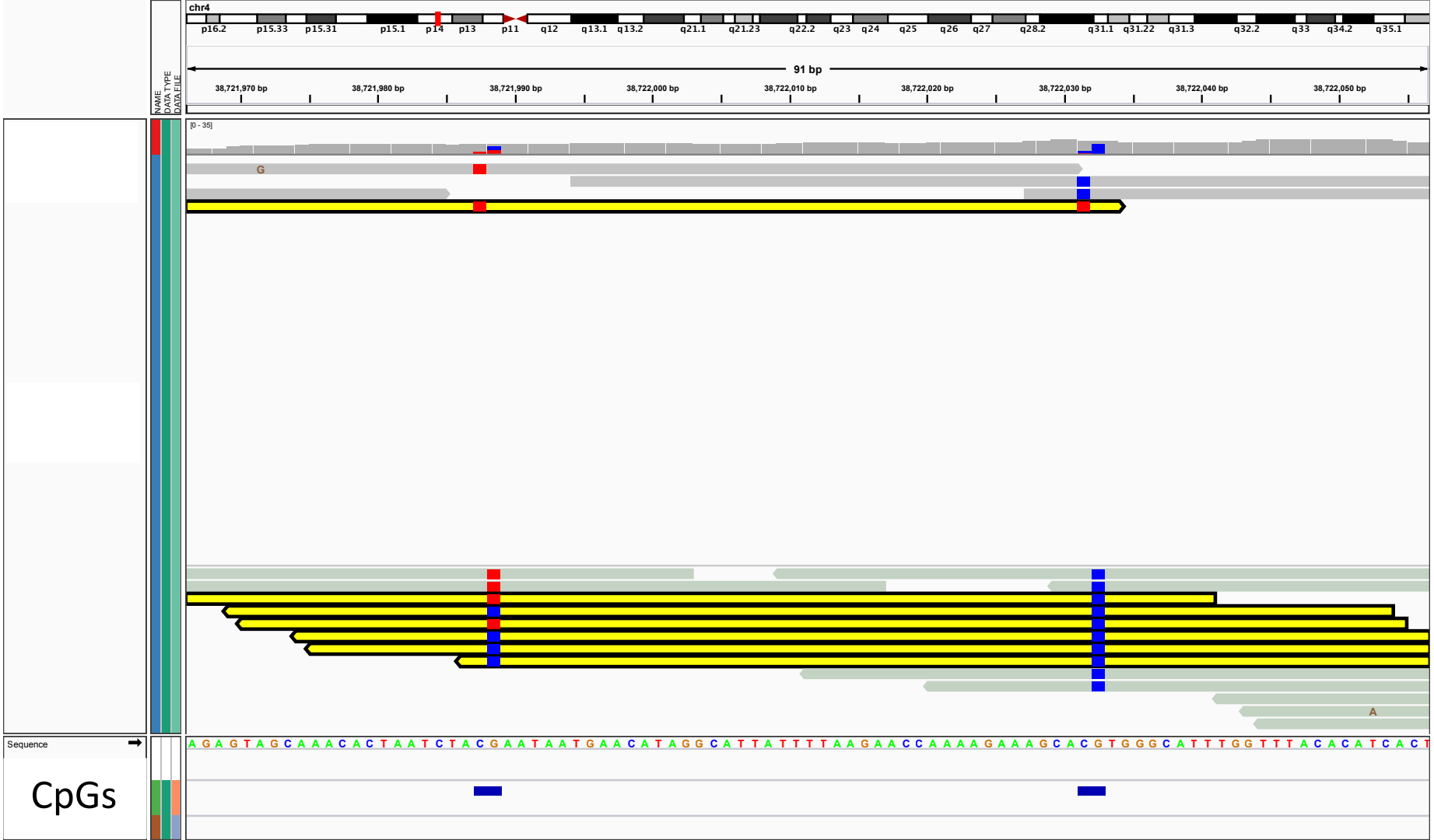
"The relationship between the degree of methylation over distance"

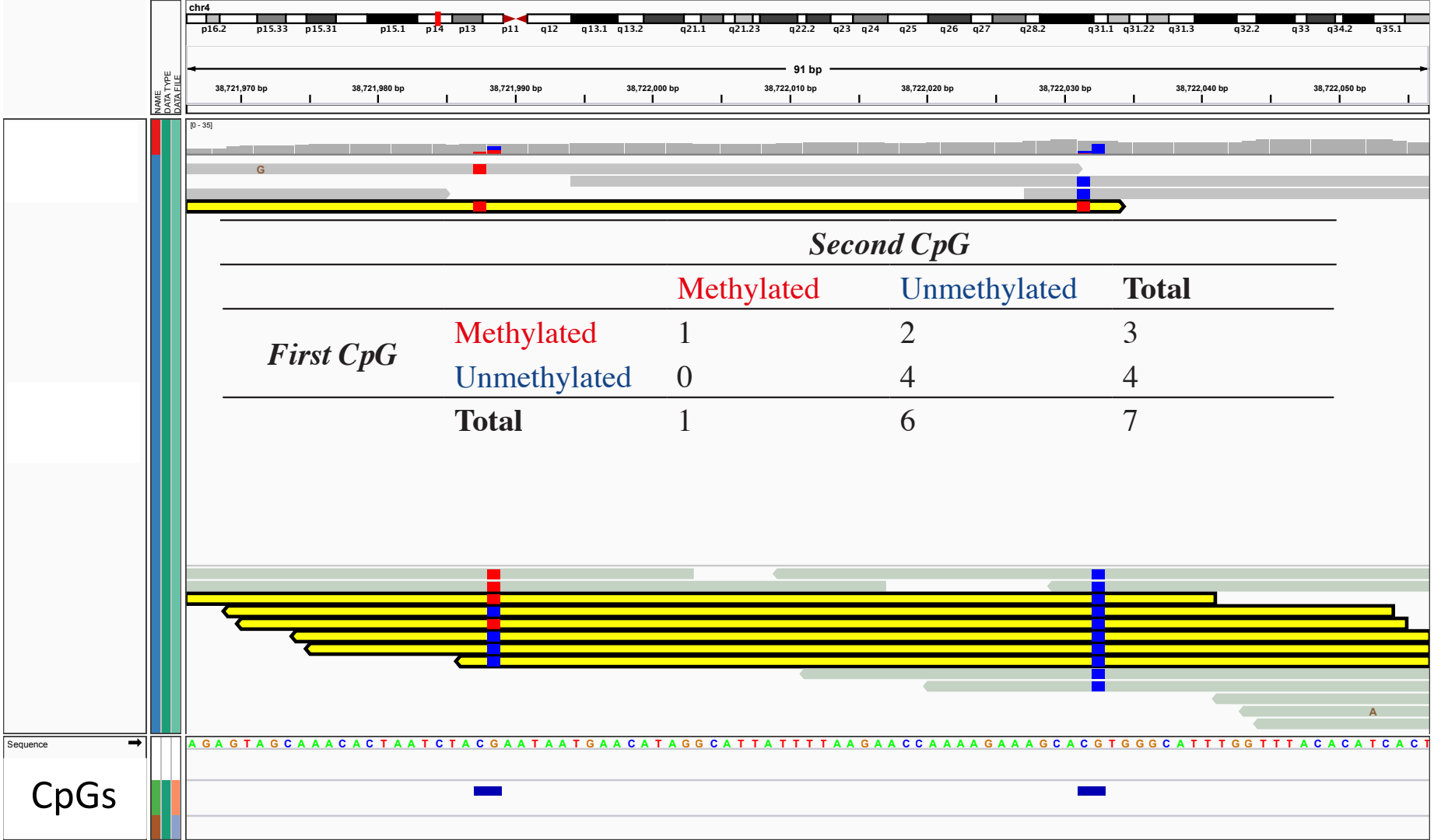
1. Within-fragment co-methylation

2. Correlation of β-values

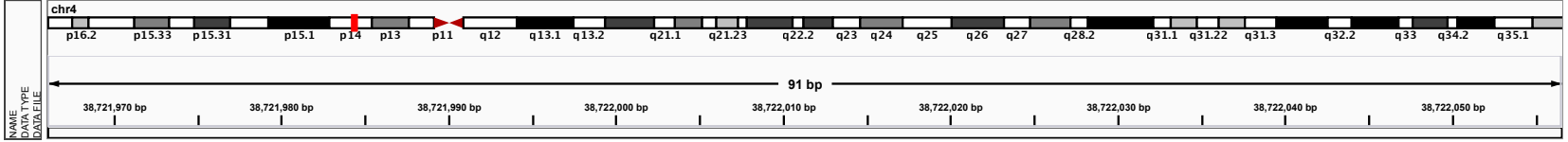
Eckhardt, Florian, et al. "DNA methylation profiling of human chromosomes 6, 20 and 22." Nature genetics 38.12 (2006): 1378-1385.

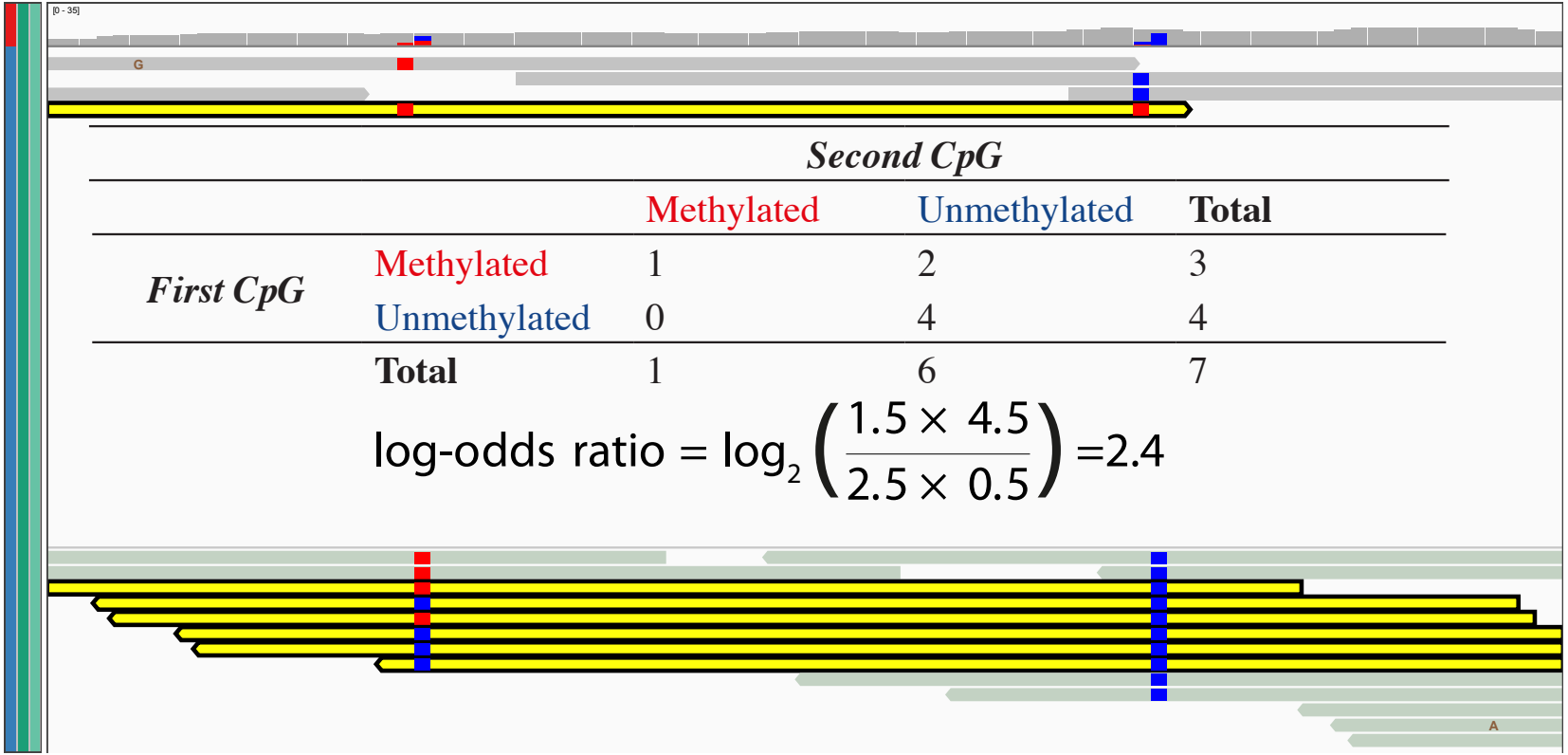

chr4
p16.2 p15.33 p15.31 p15.1 p14 p13 p11 q12 q13.1 q13.2 q21.1 q21.23 q22.2 q23 q24 q25 q26 q27 q28.2 q31.1 q31.22 q31.3 q32.2 q33 q34.2 q35.1
91 bp
38,721,970 bp
38,721,980 bp
38,721,990 bp
38,722,000 bp
38,722,010 bp
38,722,020 bp
38,722,030 bp
38,722,040 bp
38,722,050 bp
[0 - 35]
G
A
Sequence
A G A G T A G C A A A C A C T A A T C T A C G A A T A A T G A A C A T A G G C A T T A T T T T A A G A A C C A A A A G A A A G C A C G T G G G C A T T T G G T T T A C A C A T C A C T
CpGs


chr4
p16.2 p15.33 p15.31 p15.1 p14 p13 p11 q12 q13.1 q13.2 q21.1 q21.23 q22.2 q23 q24 q25 q26 q27 q28.2 q31.1 q31.22 q31.3 q32.2 q33 q34.2 q35.1
91 bp
38,721,970 bp
38,721,980 bp
38,721,990 bp
38,722,000 bp
38,722,010 bp
38,722,020 bp
38,722,030 bp
38,722,040 bp
38,722,050 bp
NAME
DATA TYPE
DATA FILE
[0 - 35]
G
A
Sequence
A G A G T A G C A A A C A C T A A T C T A C G A A T A A T G A A C A T A G G C A T T A T T T T A A G A A C C A A A A G A A A G C A C G T G G G C A T T T G G T T T A C A C A T C A C T
CpGs



		Second CpG		
		Methylated	Unmethylated	**Total**
First CpG	Methylated	1	2	3
	Unmethylated	0	4	4
	Total	1	6	7





Second CpG

		Methylated	Unmethylated	**Total**
First CpG	Methylated	1	2	3
	Unmethylated	0	4	4
	Total	1	6	7

$$\text{log-odds ratio} = \log_2\left(\frac{1.5 \times 4.5}{2.5 \times 0.5}\right) = 2.4$$

CpGs

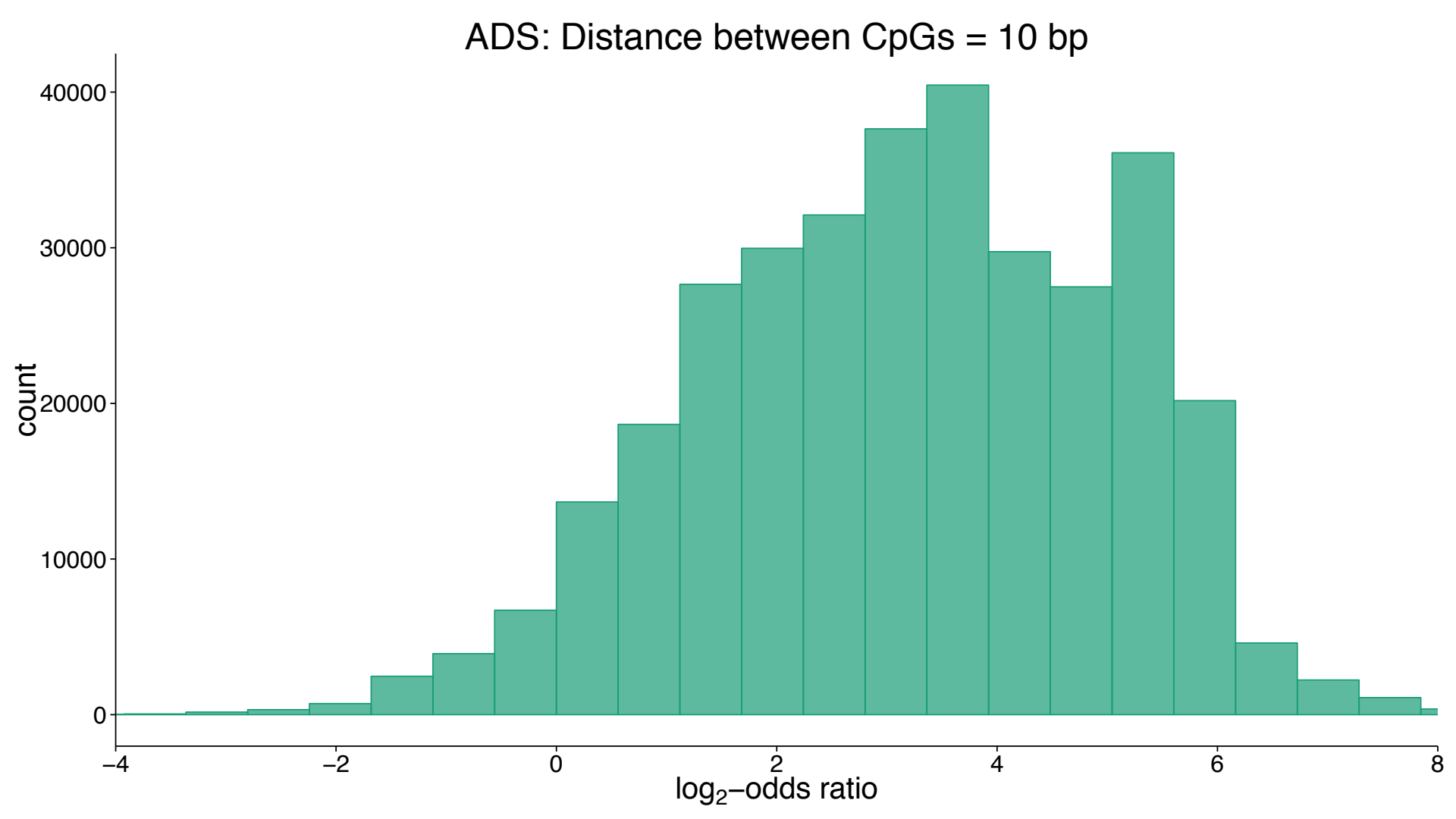
Do this 50 million times per sample

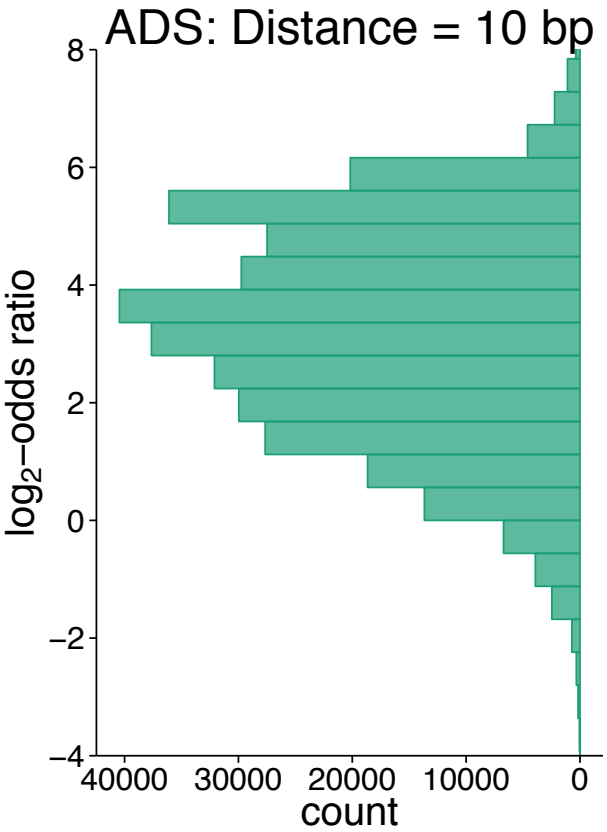
chr	strand	pos1	pos2	MM	MU	UM	UU
chr1	+	469	471	0	5	1	0
chr1	+	471	484	1	0	4	1
chr1	+	484	489	3	2	1	0
chr1	+	489	493	4	2	1	1
chr1	+	493	497	4	1	3	0
chr1	+	497	525	6	0	1	0
chr1	+	525	542	4	0	0	0
chr1	+	525	563	1	0	0	0
...	...	...	...	...	...	...	...

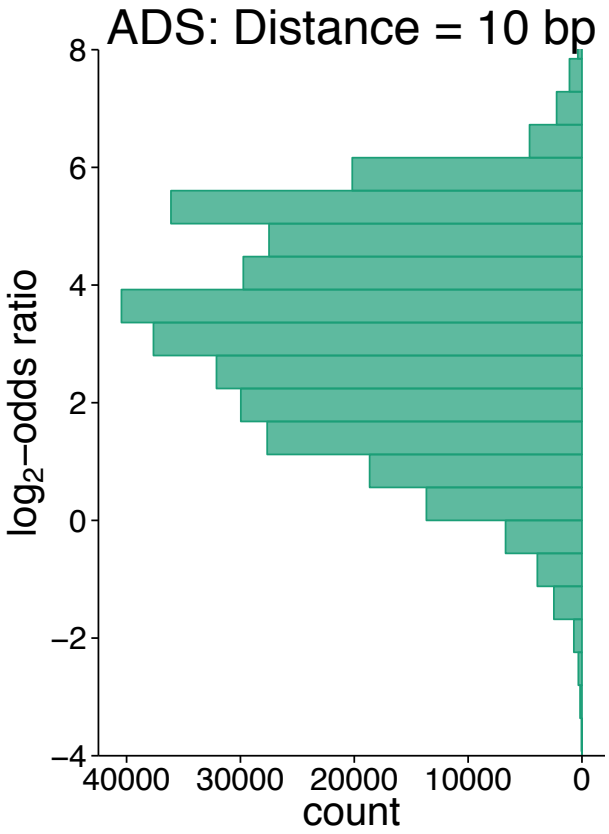
Do this 50 million times per sample

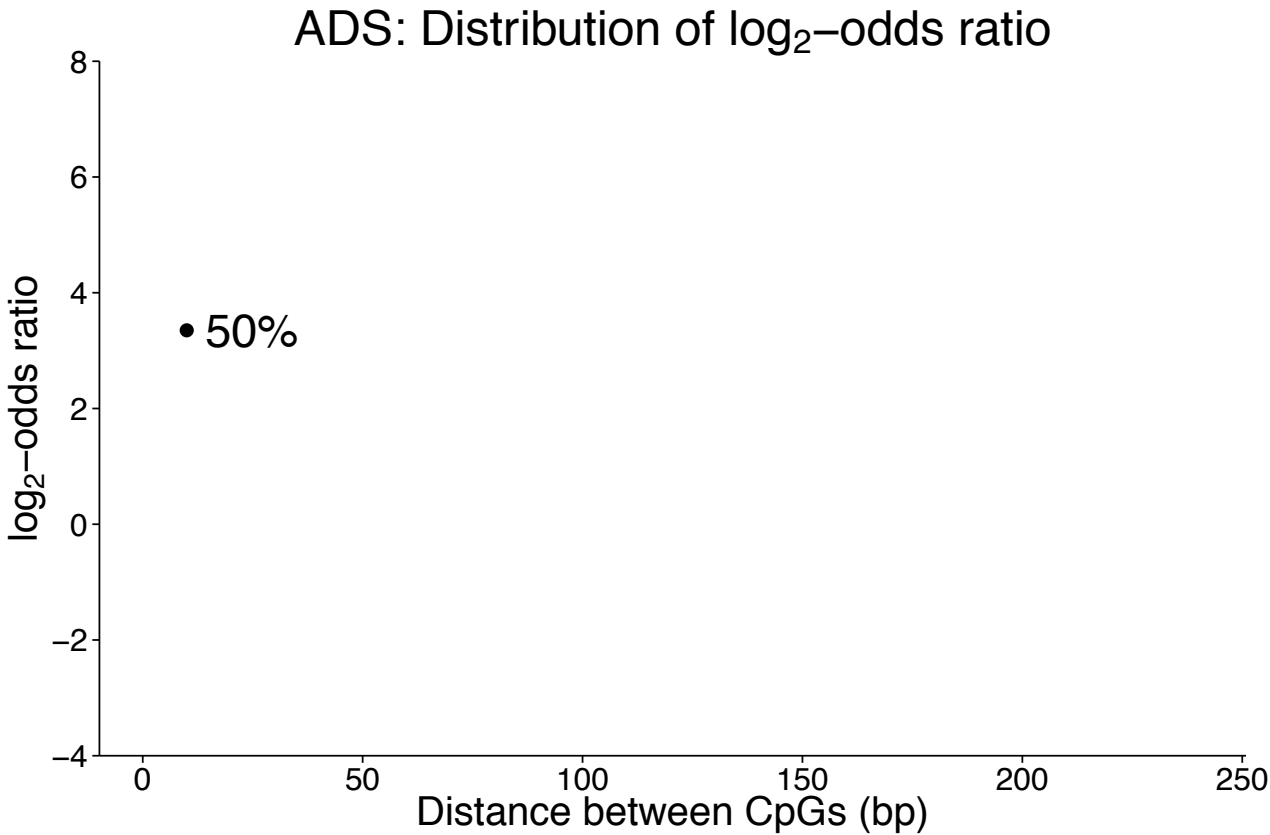
methtuple

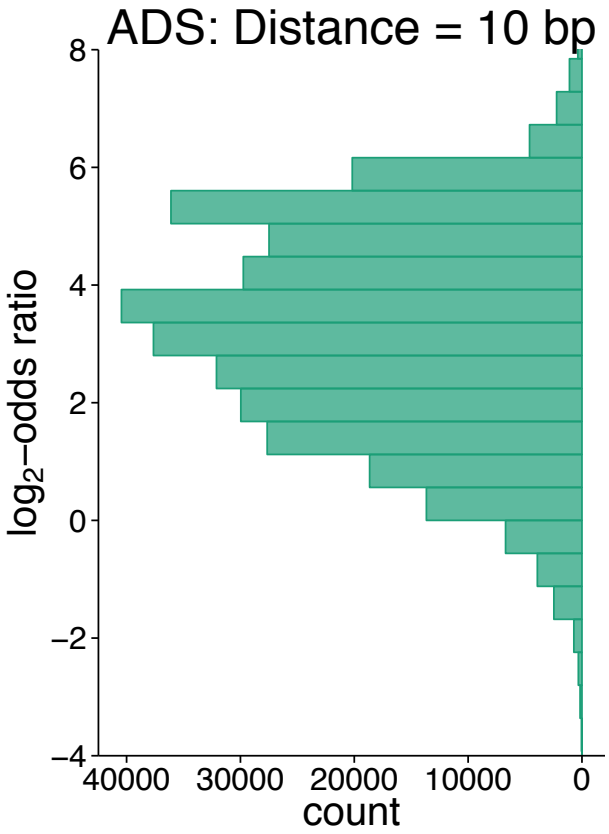
www.github.com/PeteHaitch/methtuple

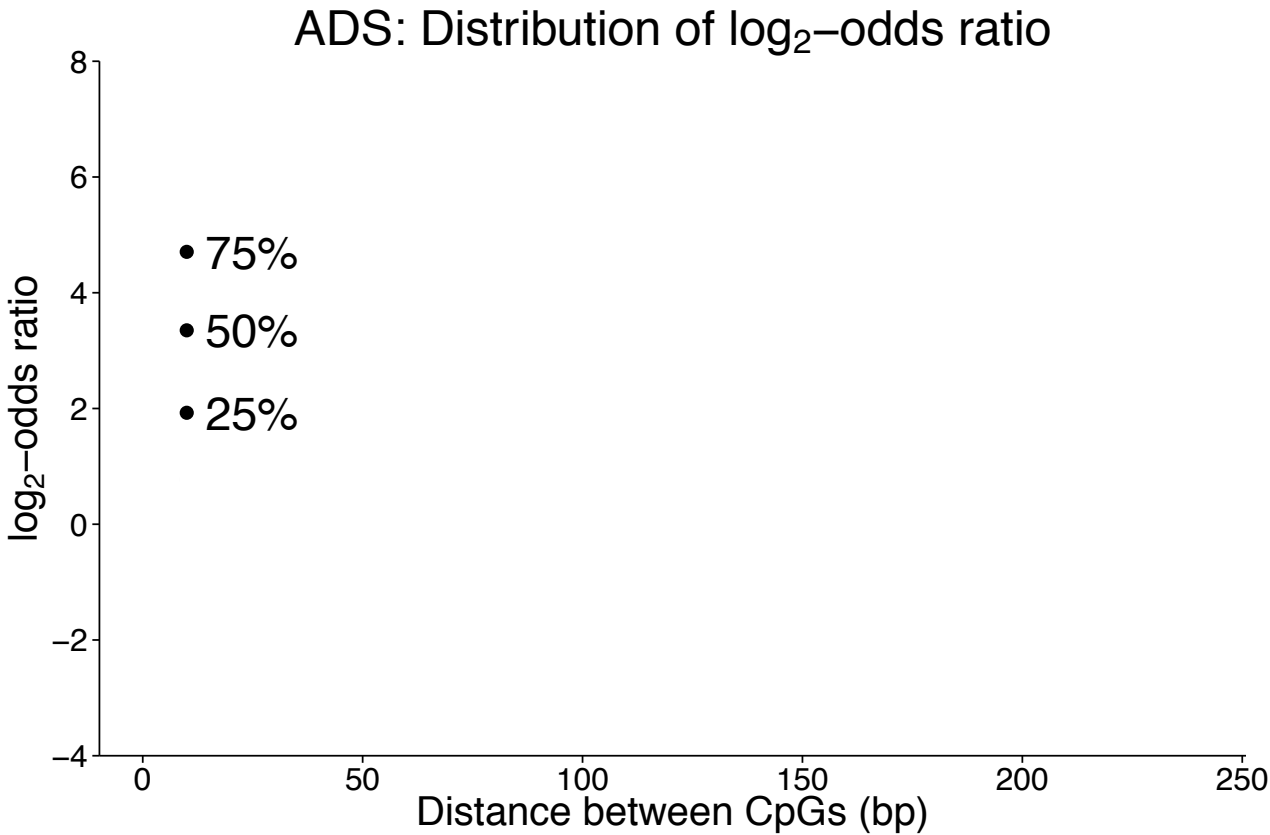

ADS: Distance between CpGs = 10 bp
count
40000
30000
20000
10000
0
−4
−2
0
2
4
6
8
log2−odds ratio

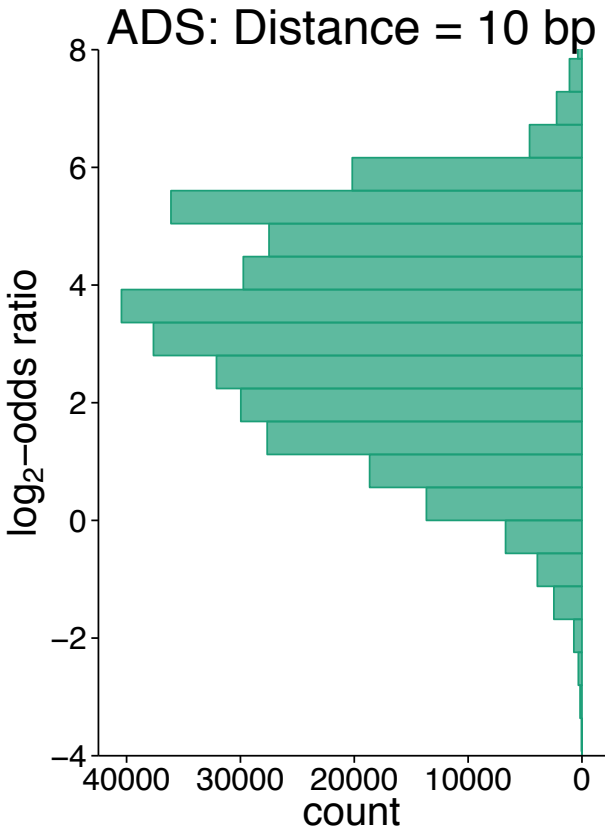

ADS: Distance = 10 bp
log2–odds ratio
8
6
4
2
0
−2
−4
40000
30000
20000
10000
0
count

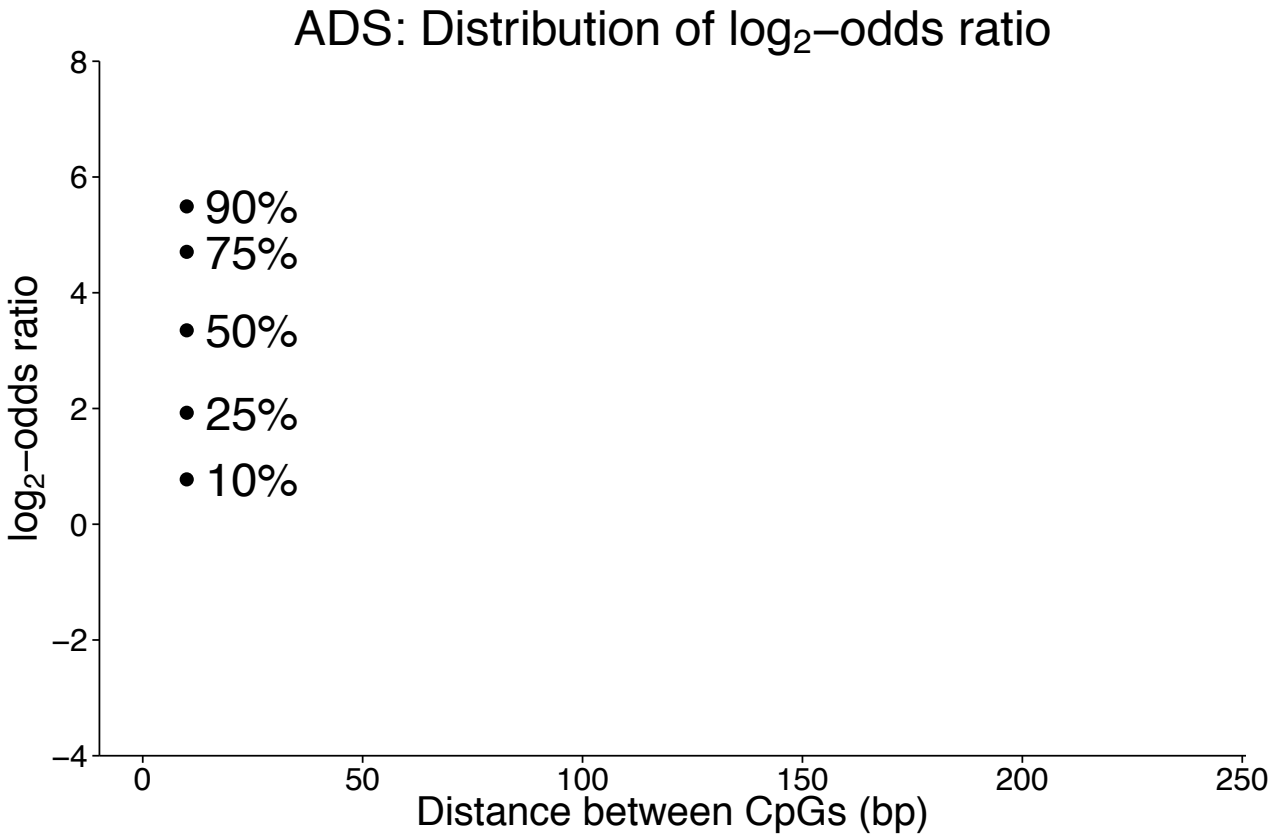

ADS: Distance = 10 bp
log2–odds ratio
8
6
4
2
0
−2
−4
40000
30000
20000
10000
0
count

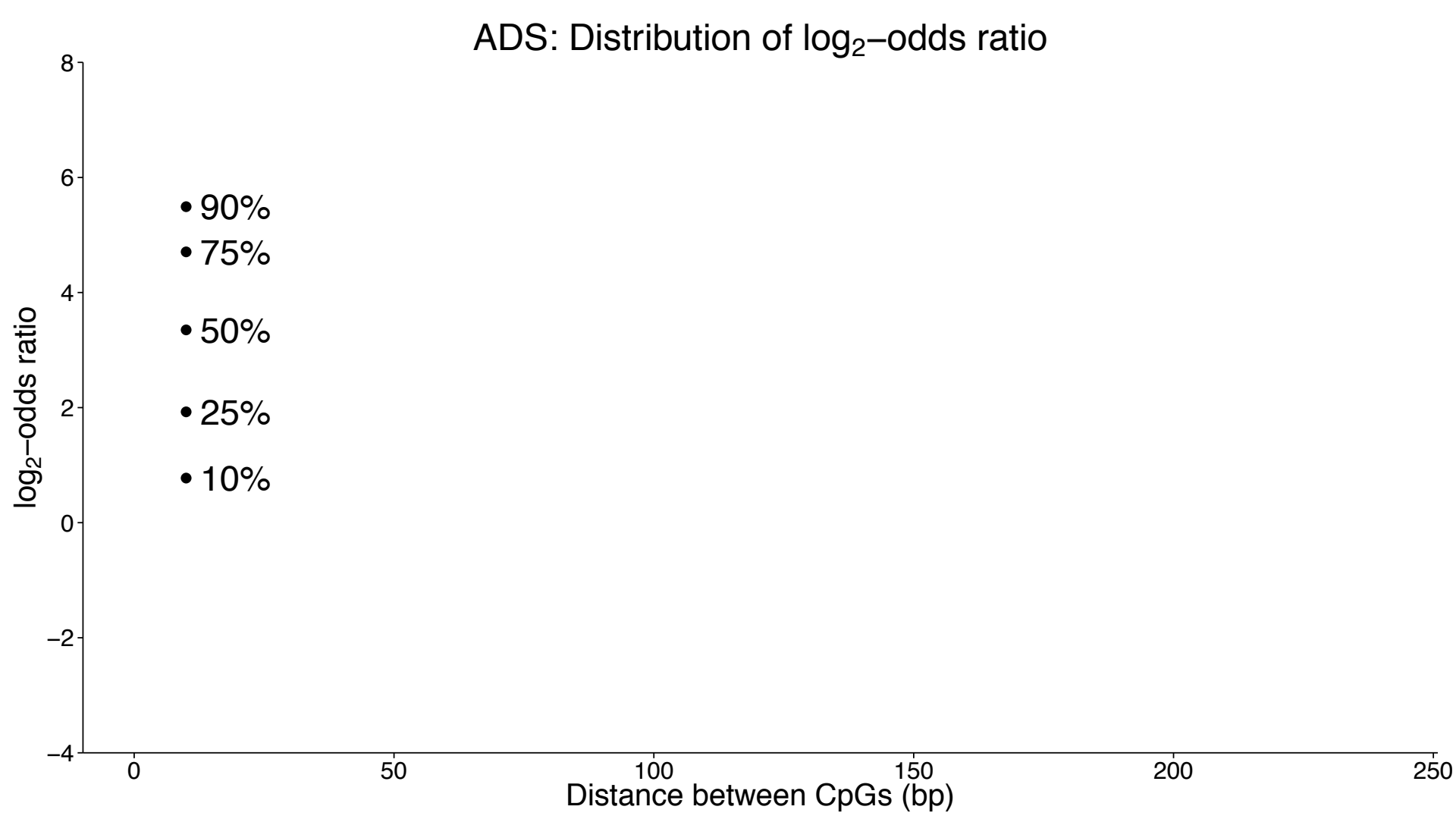

ADS: Distribution of log2–odds ratio
log2–odds ratio
8
6
4
2
0
−2
−4
• 50%
0
50
100
150
200
250
Distance between CpGs (bp)

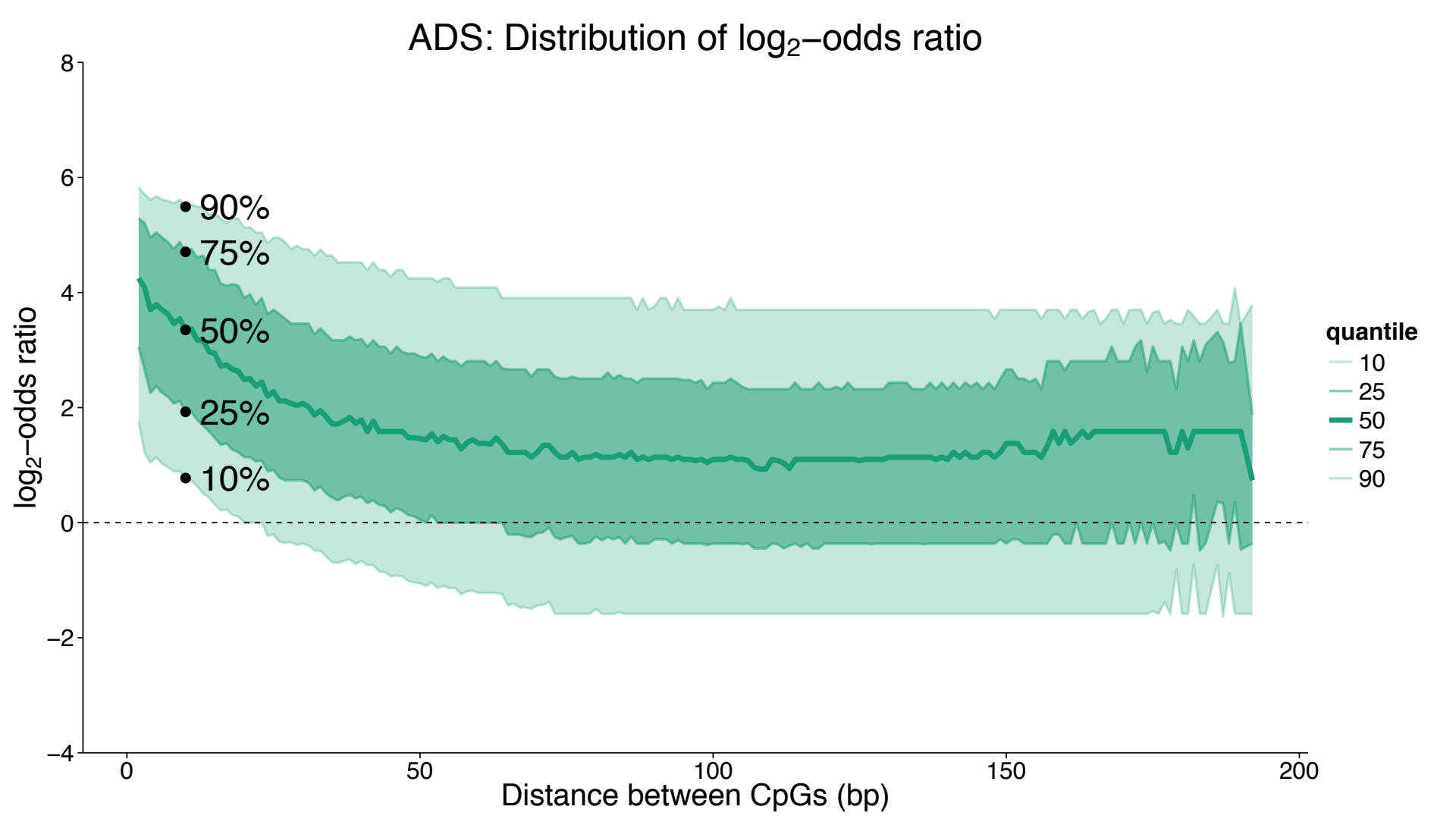

ADS: Distance = 10 bp
log2–odds ratio
8
6
4
2
0
−2
−4
40000
30000
20000
10000
0
count

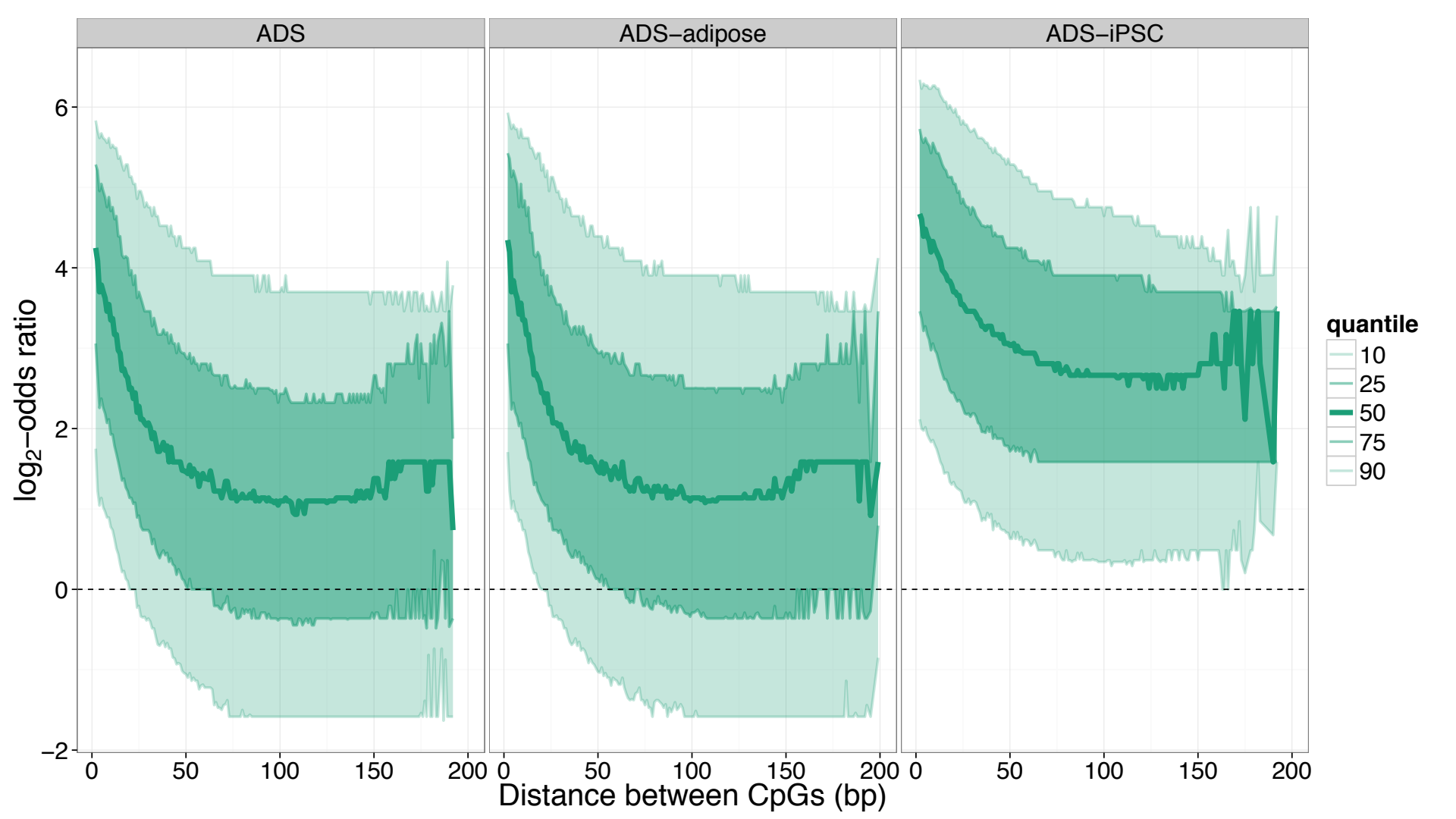

ADS: Distribution of log2–odds ratio
log2–odds ratio
8
6
4
2
0
−2
−4
• 75%
• 50%
• 25%
0
50
100
150
200
250
Distance between CpGs (bp)

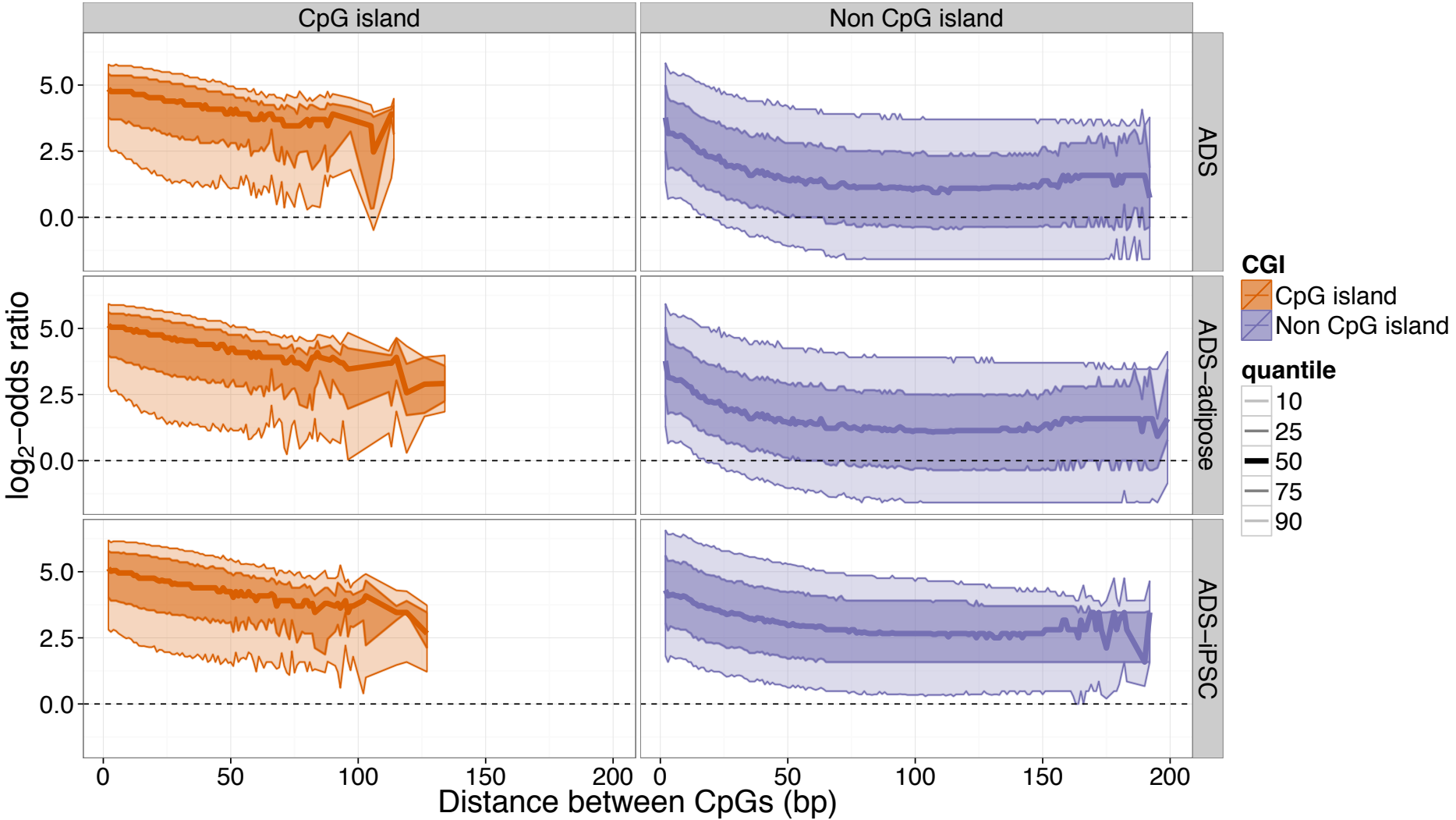

ADS: Distance = 10 bp
log2–odds ratio
8
6
4
2
0
−2
−4
40000
30000
20000
10000
0
count


ADS: Distribution of log2–odds ratio
log2–odds ratio
8
6
4
2
0
−2
−4
• 90%
• 75%
• 50%
• 25%
• 10%
0
50
100
150
200
250
Distance between CpGs (bp)


ADS: Distribution of log2−odds ratio
log2−odds ratio
8
6
4
2
0
−2
−4
• 90%
• 75%
• 50%
• 25%
• 10%
0
50
100
150
200
250
Distance between CpGs (bp)


ADS: Distribution of log2−odds ratio
8
6
4
2
0
−2
−4
log2−odds ratio
90%
75%
50%
25%
10%
0
50
100
150
200
Distance between CpGs (bp)
quantile
10
25
50
75
90


ADS
ADS−adipose
ADS−iPSC
log2−odds ratio
6
4
2
0
−2
0
50
100
150
200
Distance between CpGs (bp)
quantile
10
25
50
75
90


CpG island
Non CpG island
ADS
ADS−adipose
ADS−iPSC
log2−odds ratio
5.0
2.5
0.0
0
50
100
150
200
Distance between CpGs (bp)
CGI
CpG island
Non CpG island
quantile
10
25
50
75
90

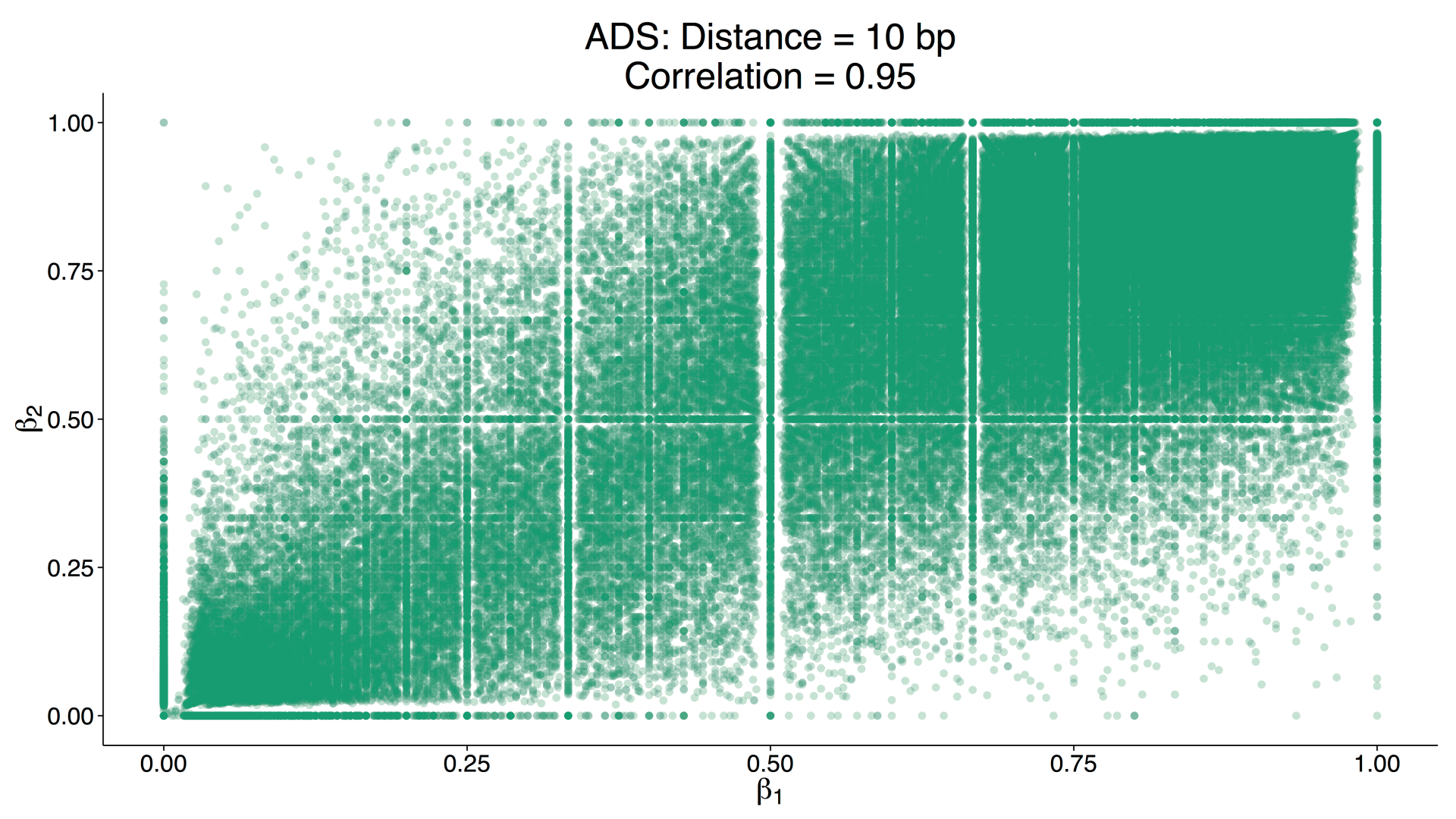
Co-methylation = correlation

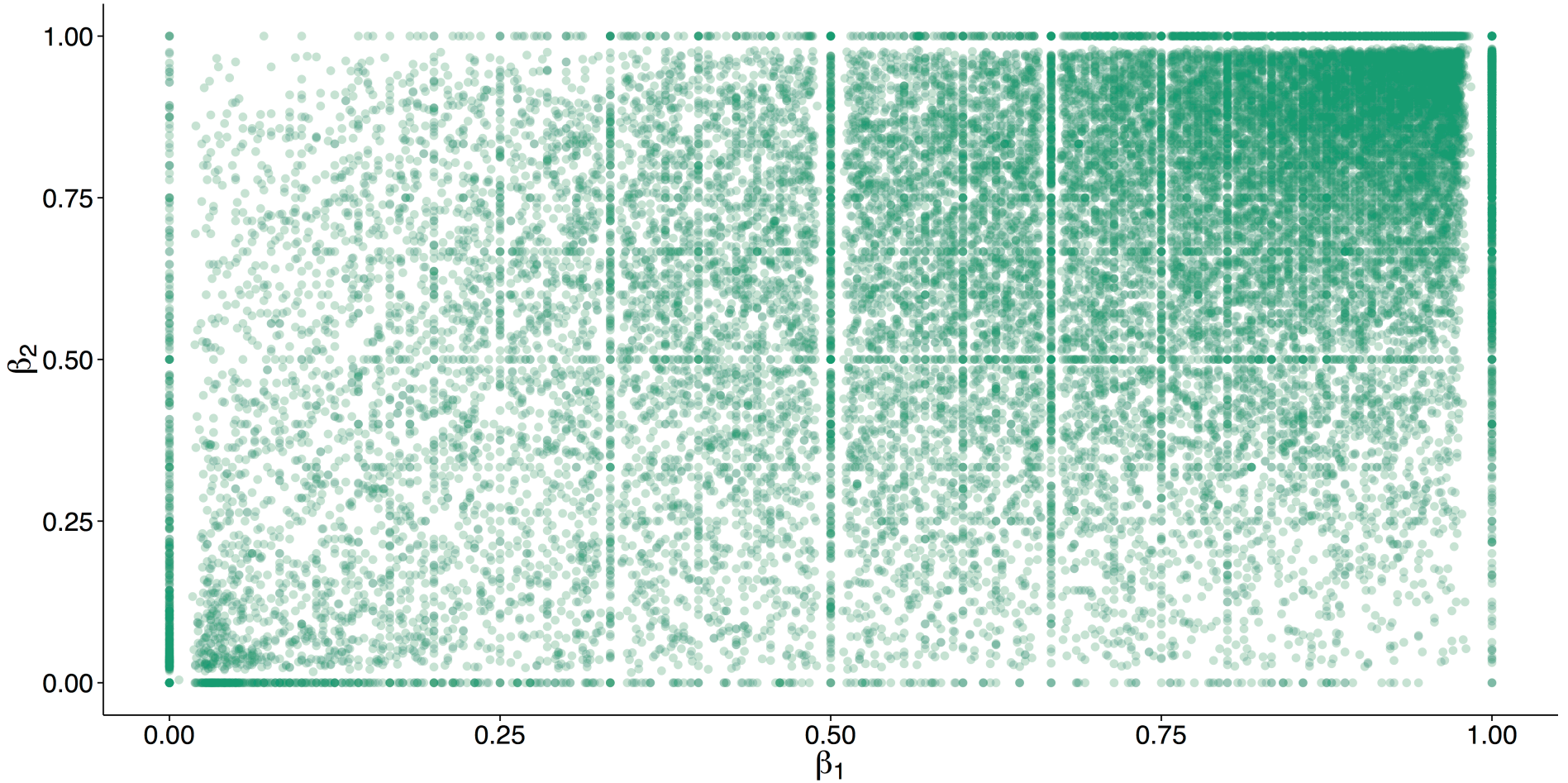
"The relationship between the degree of methylation over distance"

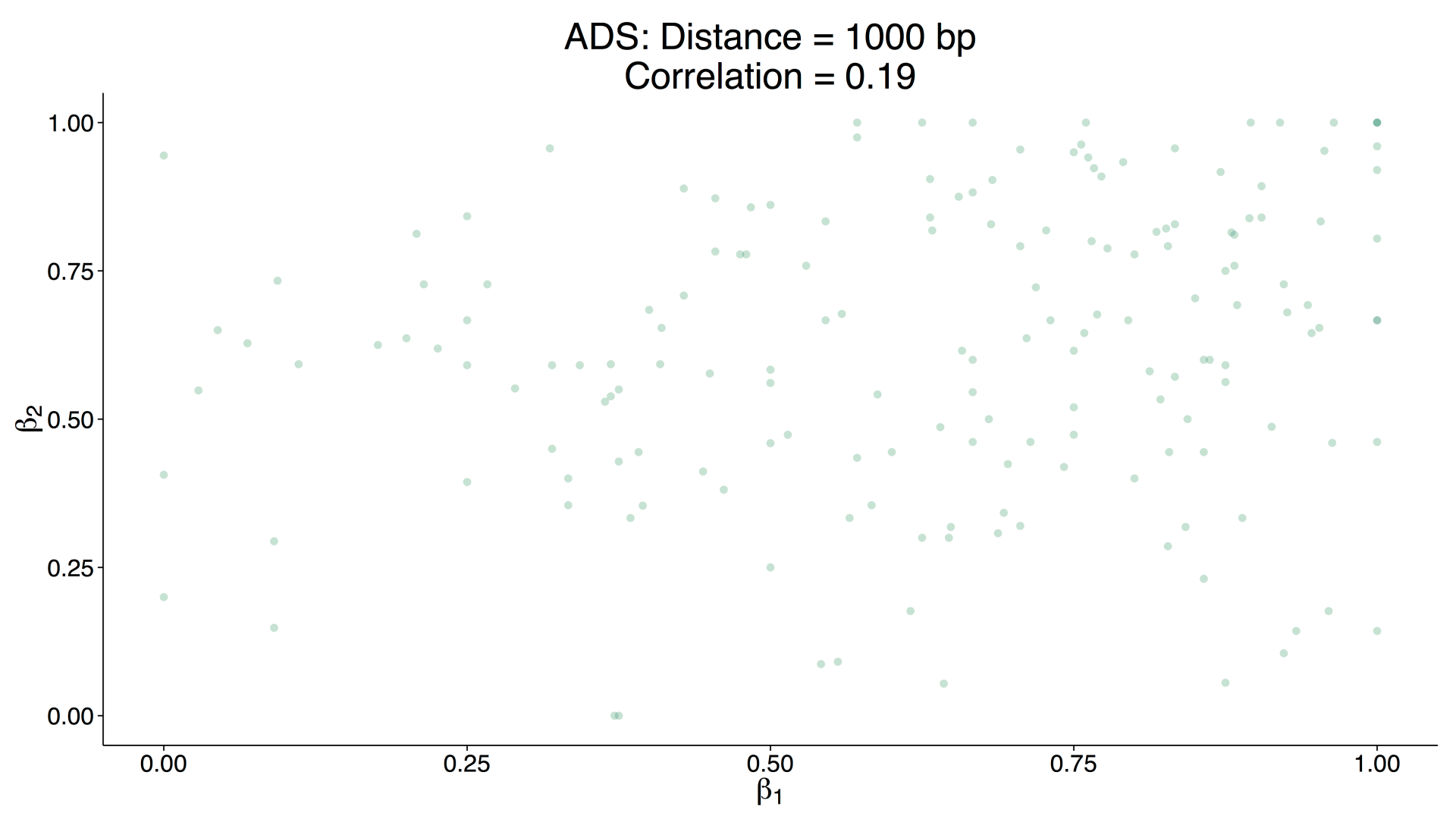
1. Within-fragment co-methylation

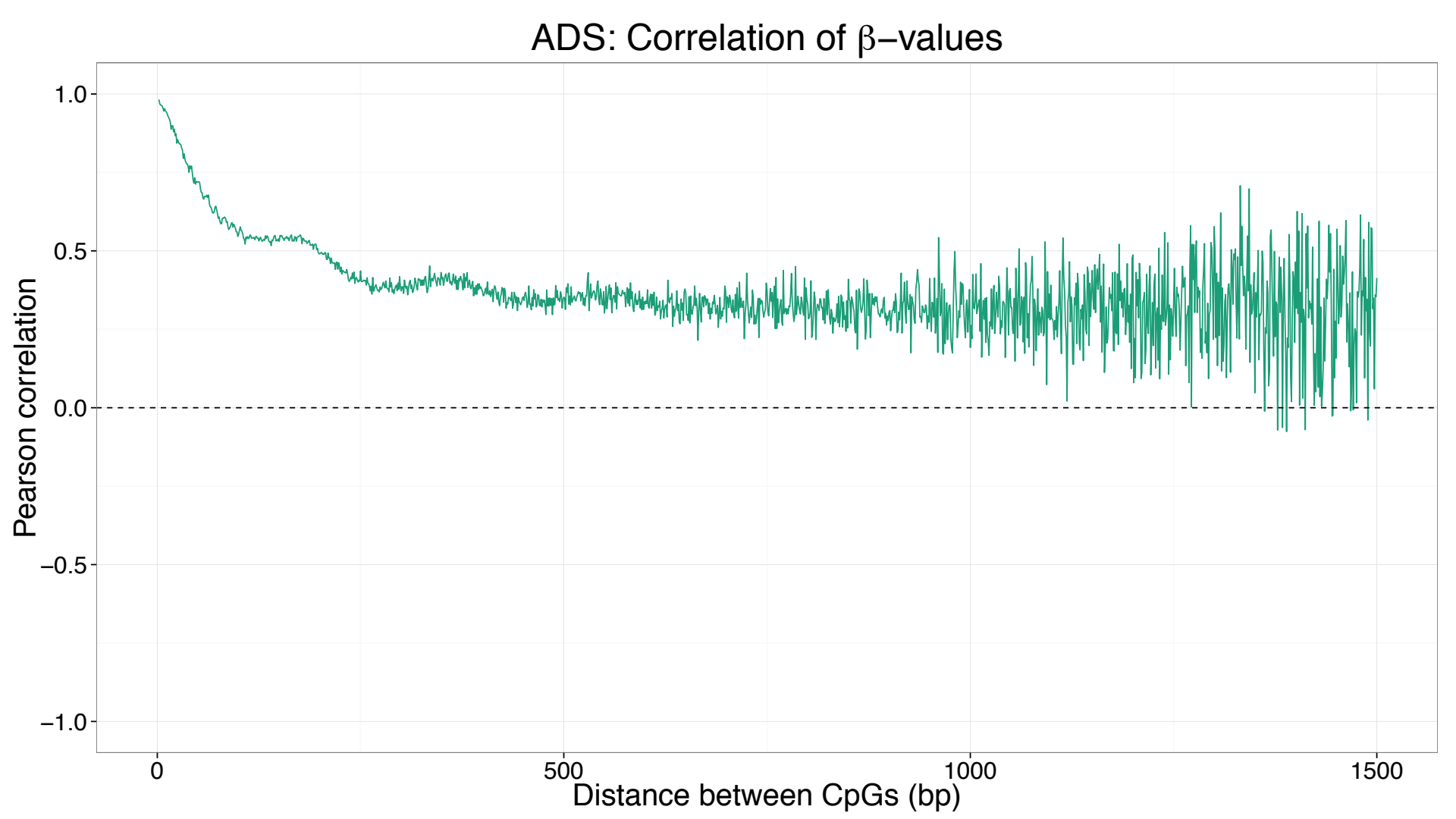
2. Correlation of β-values

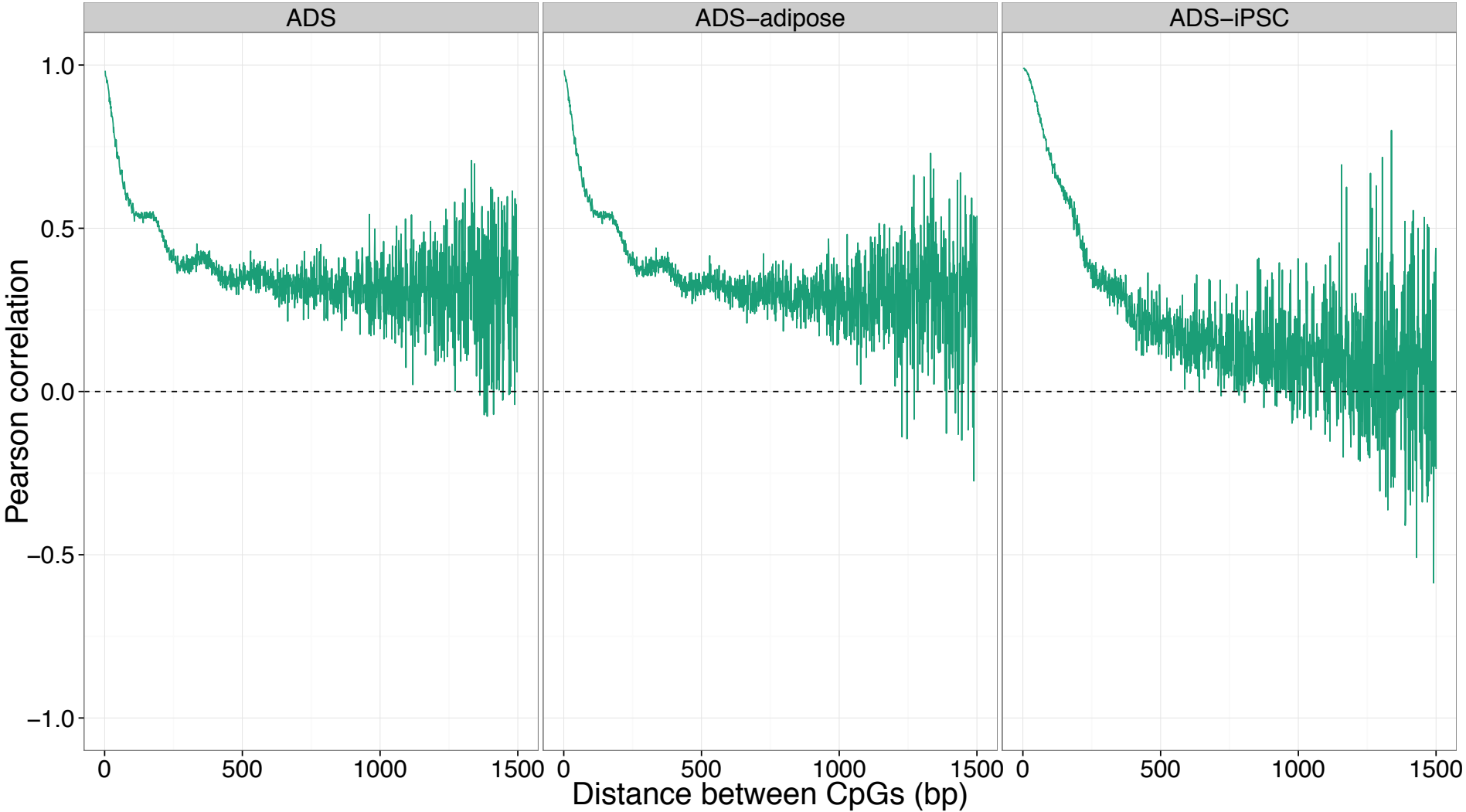
Eckhardt, Florian, et al. "DNA methylation profiling of human chromosomes 6, 20 and 22." Nature genetics 38.12 (2006): 1378-1385.

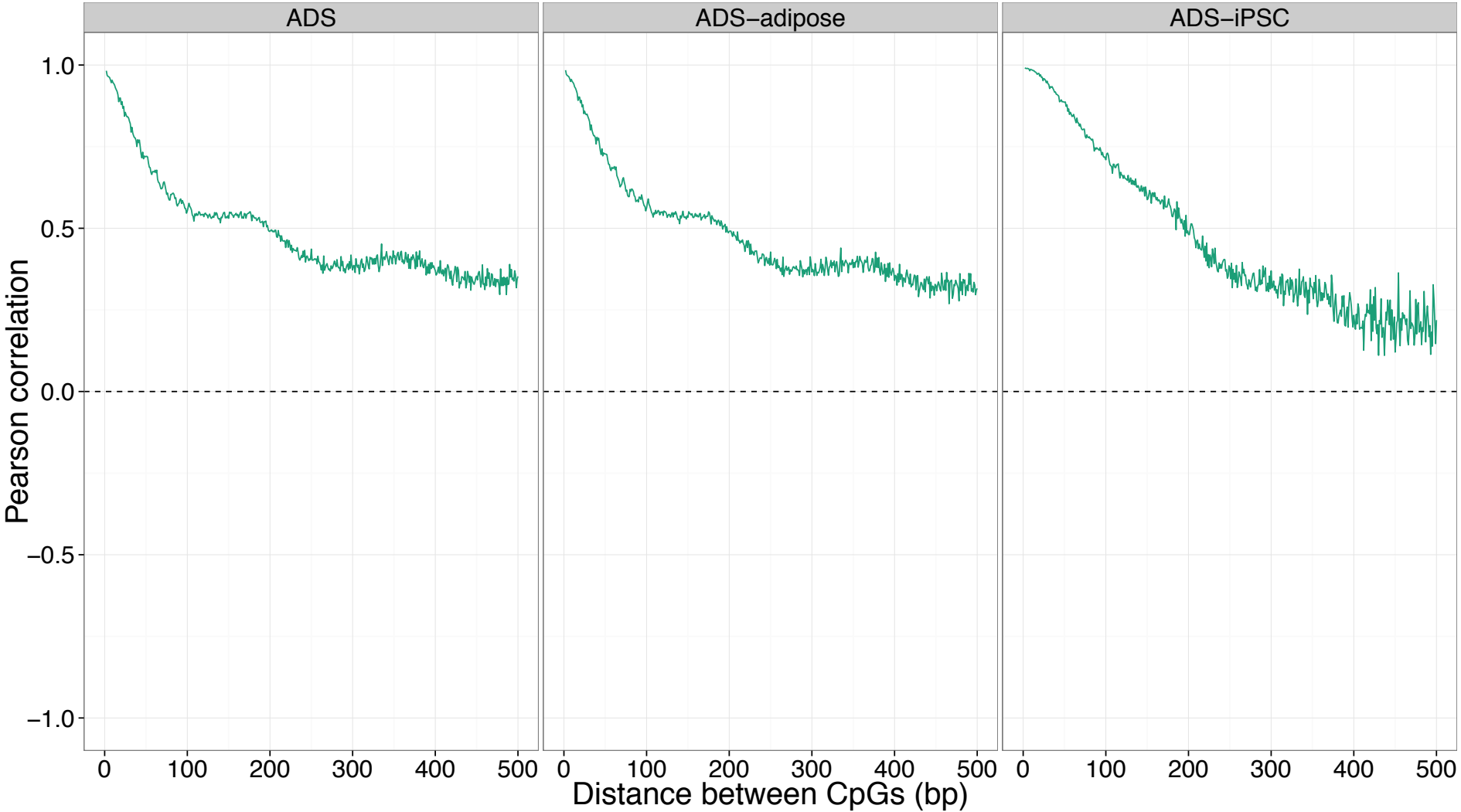

ADS: Distance = 10 bp
Correlation = 0.95
1.00
0.75
0.50
0.25
0.00
β2
0.00
0.25
0.50
0.75
1.00
β1

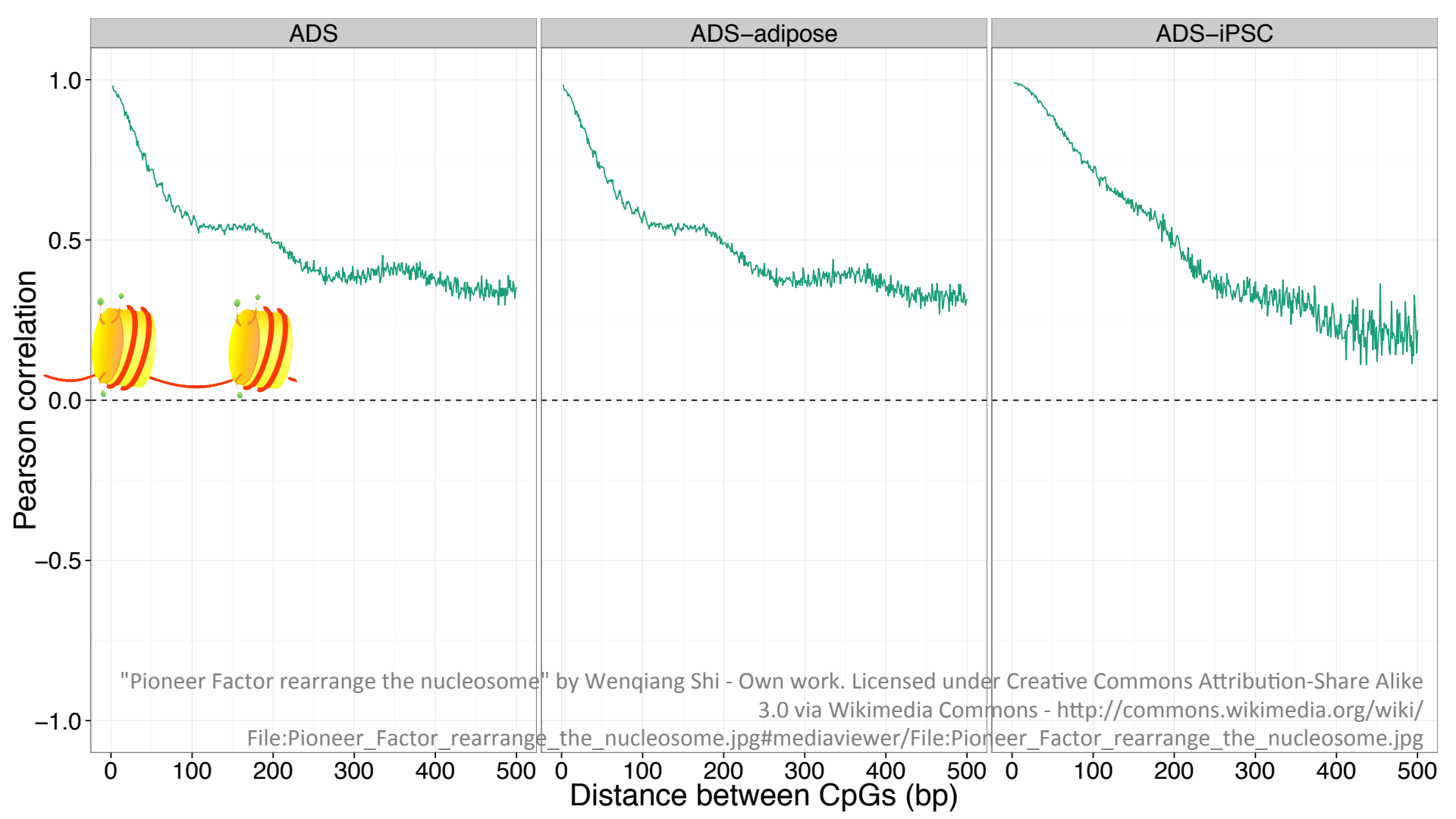

ADS: Distance = 100 bp
Correlation = 0.55
1.00
0.75
0.50
0.25
0.00
β2
0.00
0.25
0.50
0.75
1.00
β1

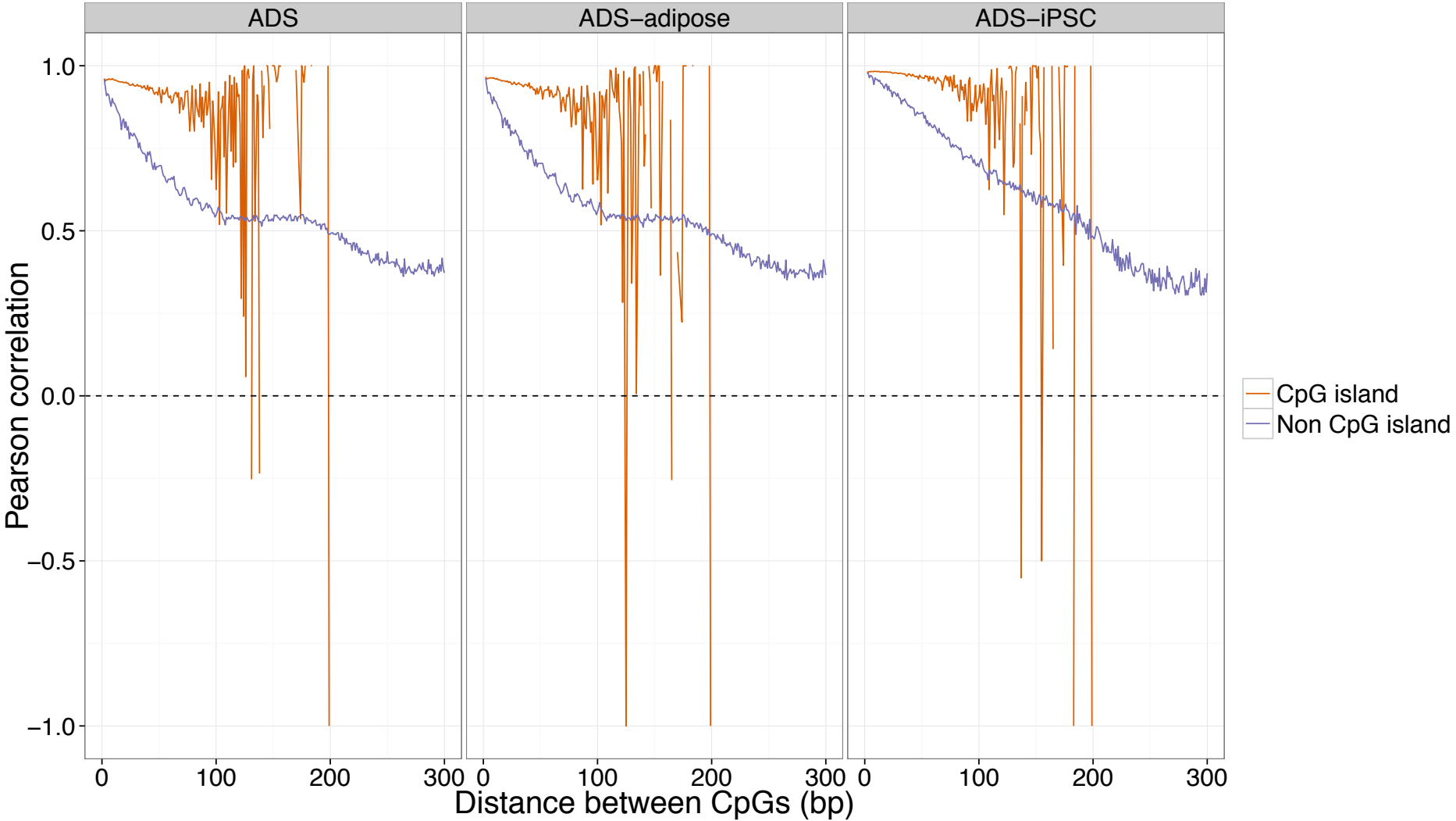

ADS: Distance = 1000 bp
Correlation = 0.19
1.00
0.75
0.50
0.25
0.00
β2
0.00
0.25
0.50
0.75
1.00
β1


ADS: Correlation of β−values
1.0
0.5
0.0
−0.5
−1.0
Pearson correlation
0
500
1000
1500
Distance between CpGs (bp)


ADS
ADS−adipose
ADS−iPSC
Pearson correlation
1.0
0.5
0.0
−0.5
−1.0
0
500
1000
1500
Distance between CpGs (bp)

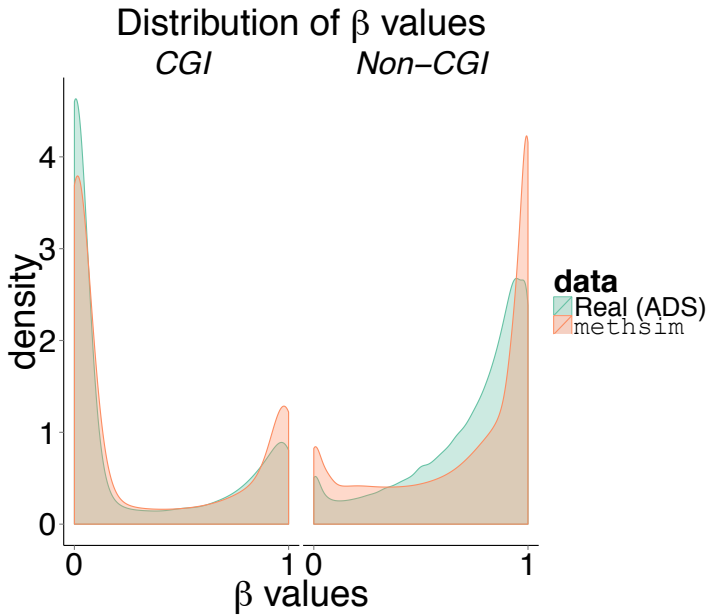

ADS
ADS−adipose
ADS−iPSC
Pearson correlation
1.0
0.5
0.0
−0.5
−1.0
0
100
200
300
400
500
Distance between CpGs (bp)

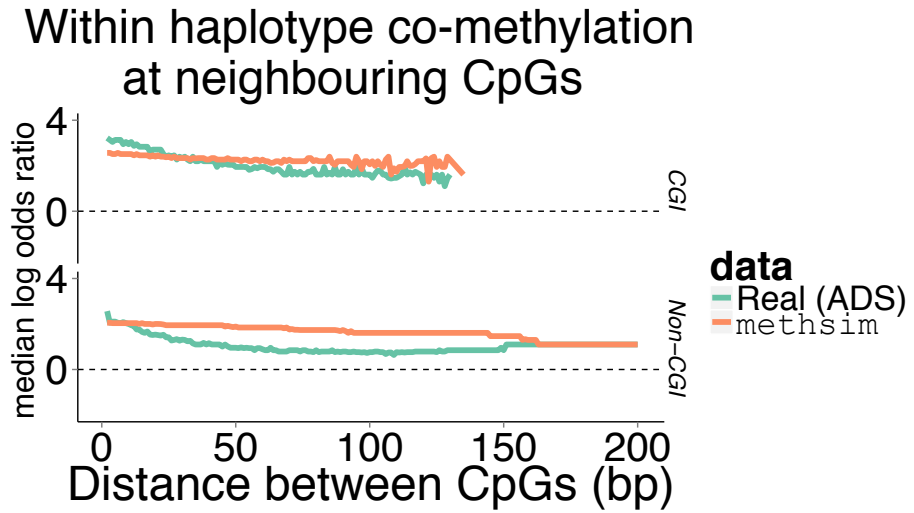


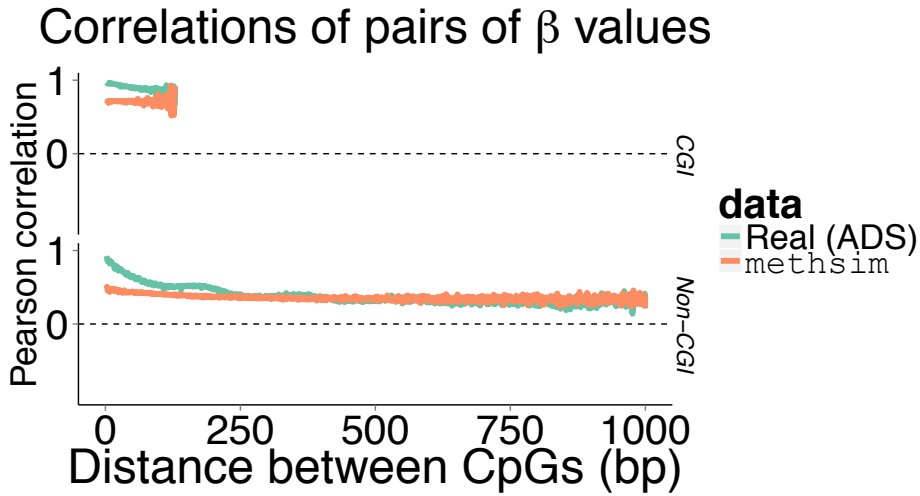

ADS
ADS−adipose
ADS−iPSC
1.0
0.5
0.0
−0.5
−1.0
Pearson correlation
0
100
200
300
Distance between CpGs (bp)
CpG island
Non CpG island

methsim

www.github.com/PeteHaitch/methsim

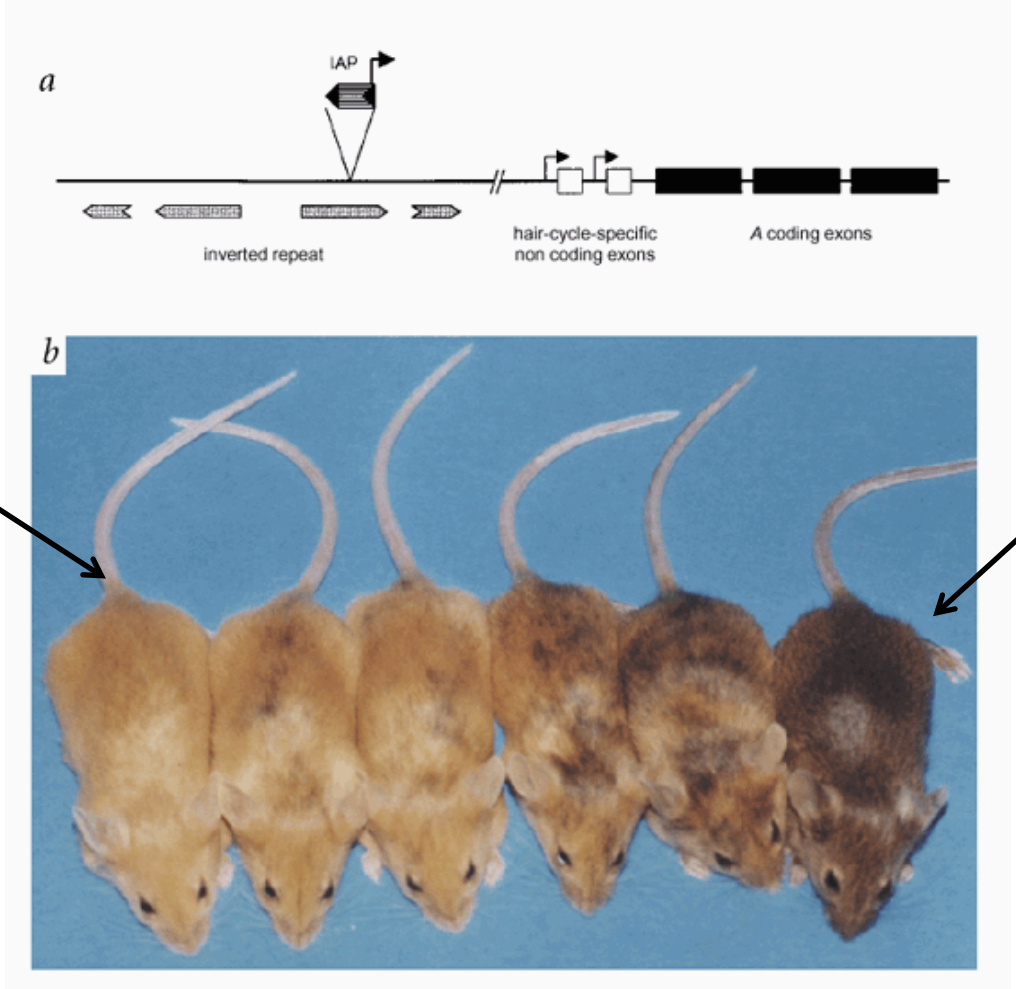






Some of these mice are not like the others (we hope…)

Methylome of the agouti viable yellow mouse (A^{vy})

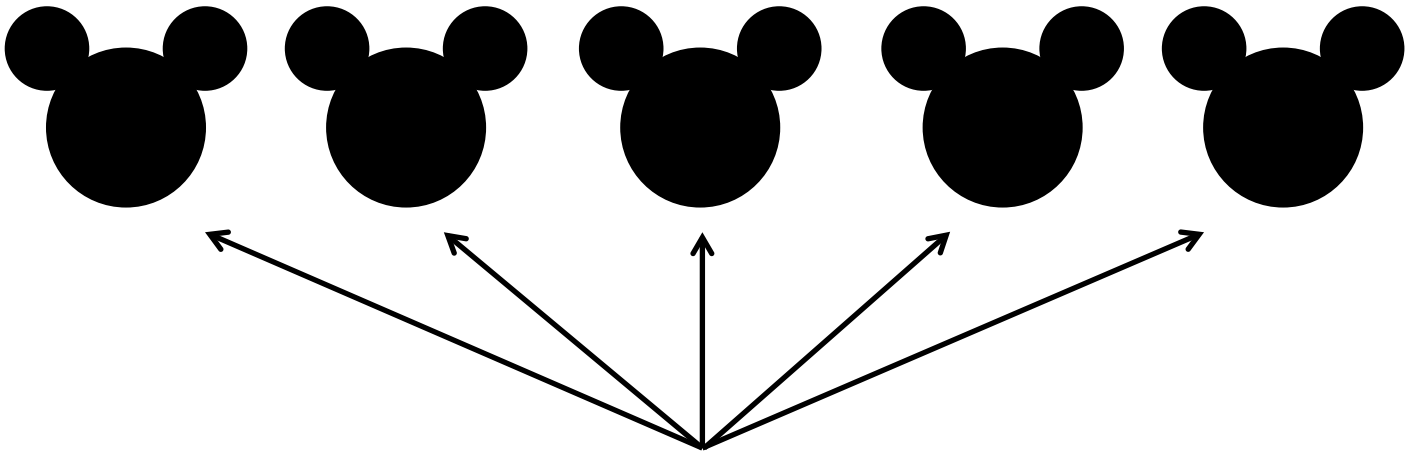


Unmethylated A^{vy}

Methylated A^{vy}

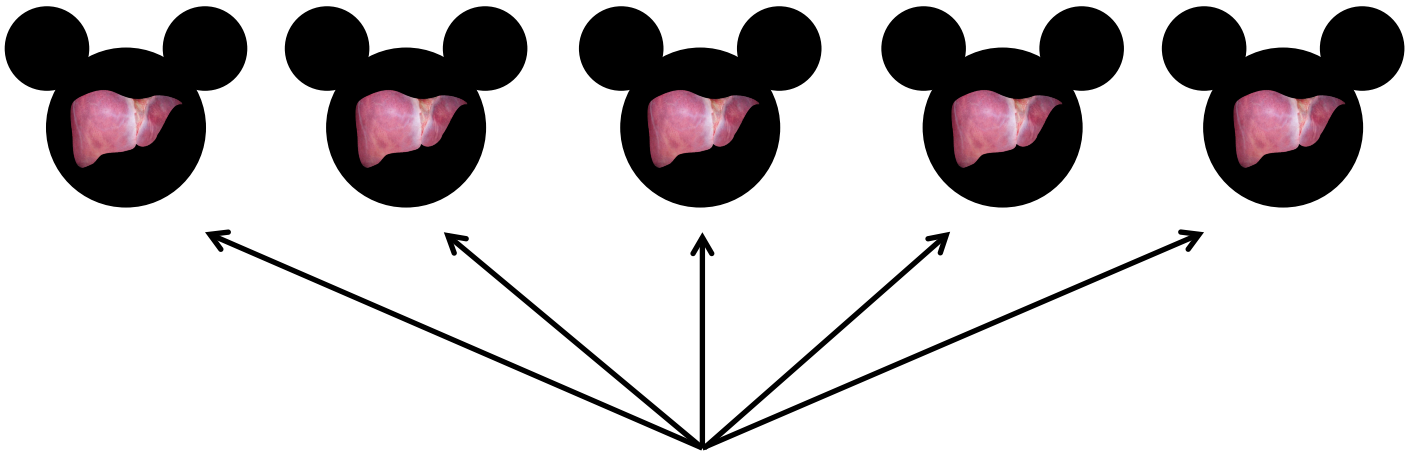
Morgan, Hugh D., et al. "Epigenetic inheritance at the agouti locus in the mouse." Nature genetics 23.3 (1999): 314-318.

Experimental design



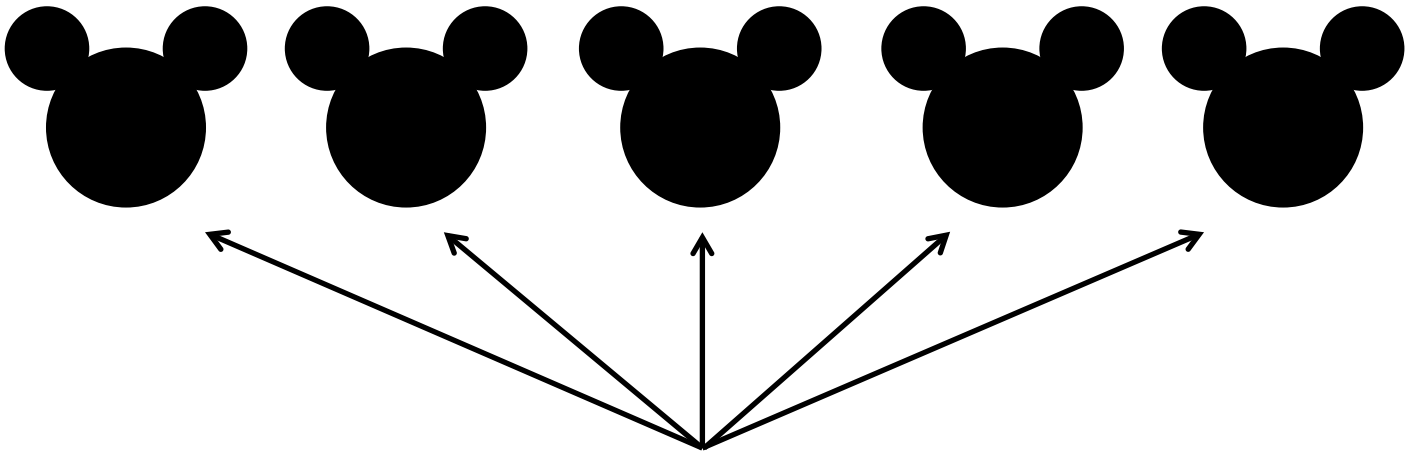
C57BL/6

Experimental design



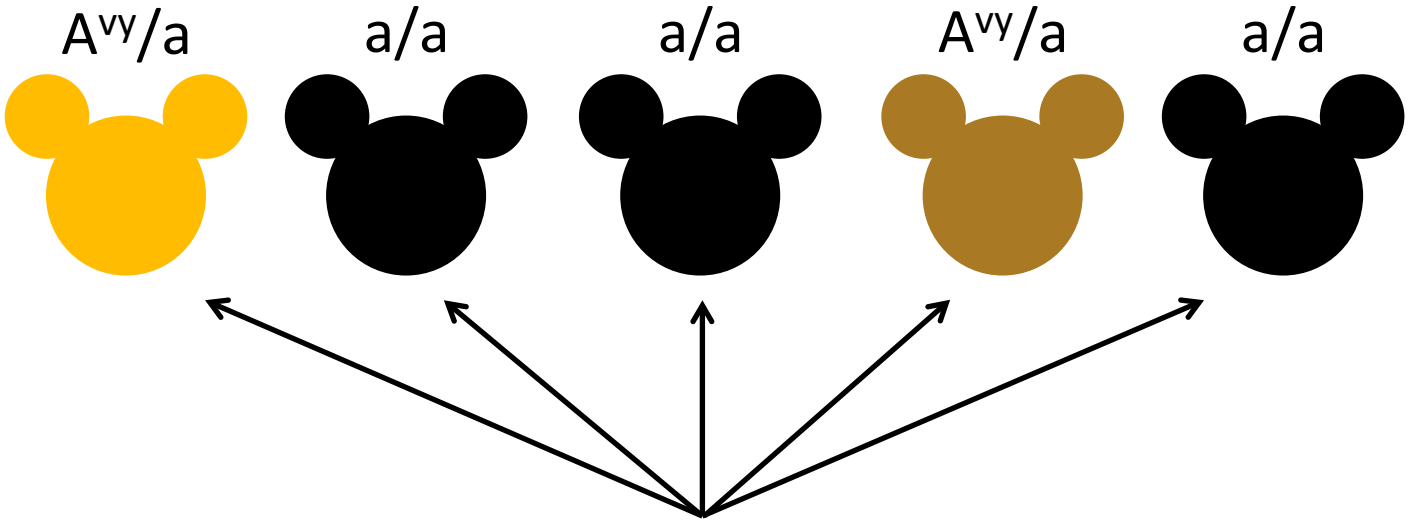
C57BL/6

Experimental design



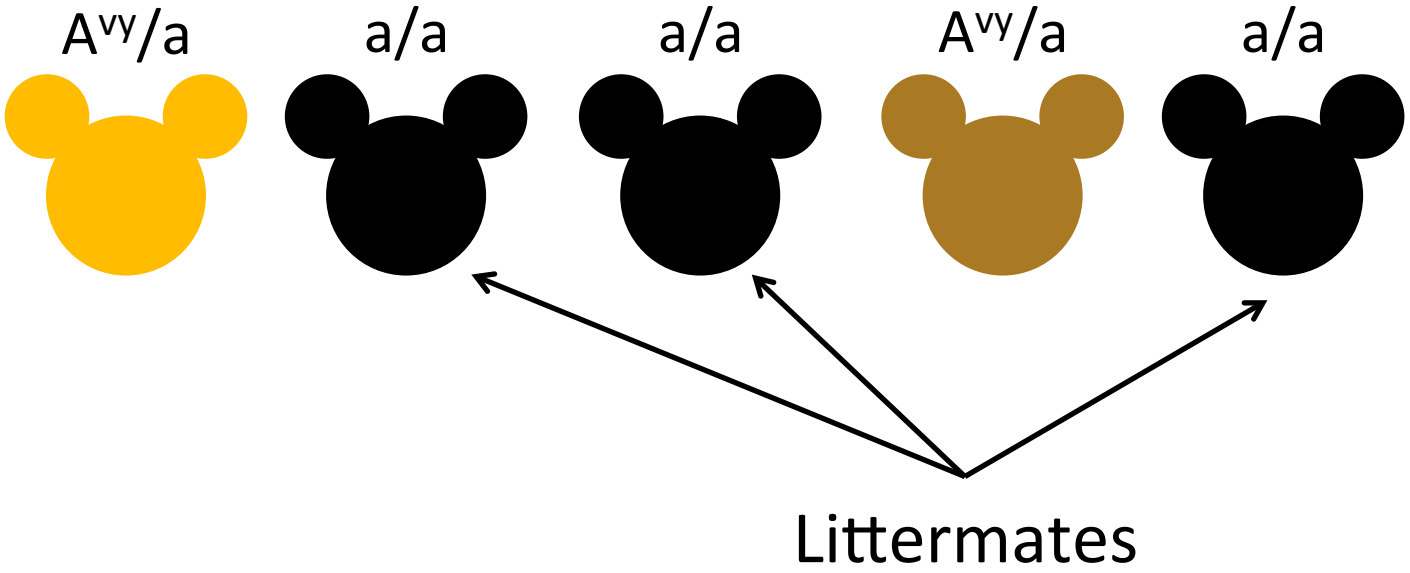
C57BL/6
Isogenic*

Experimental design

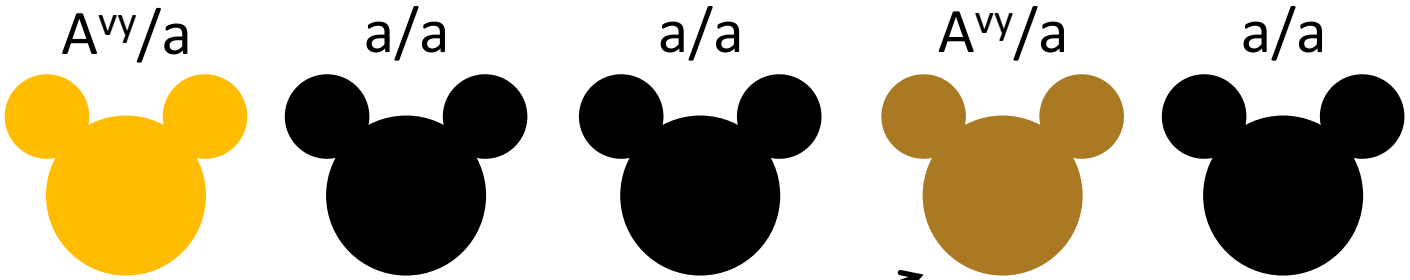


C57BL/6
Isogenic*

Experimental design

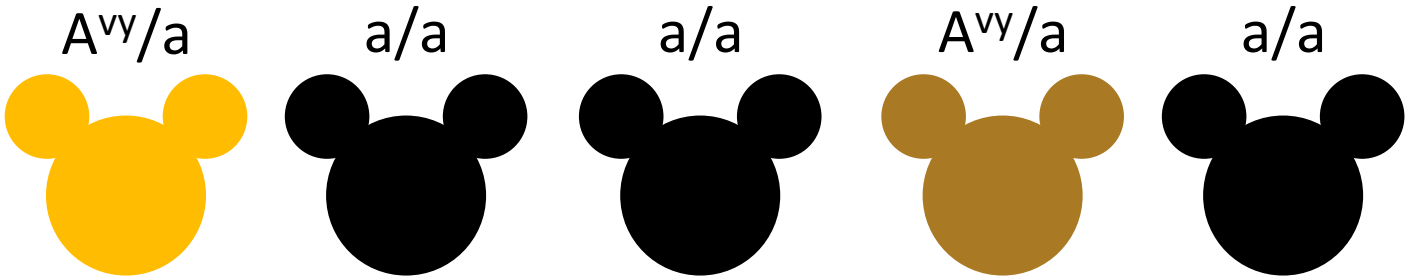


Experimental design

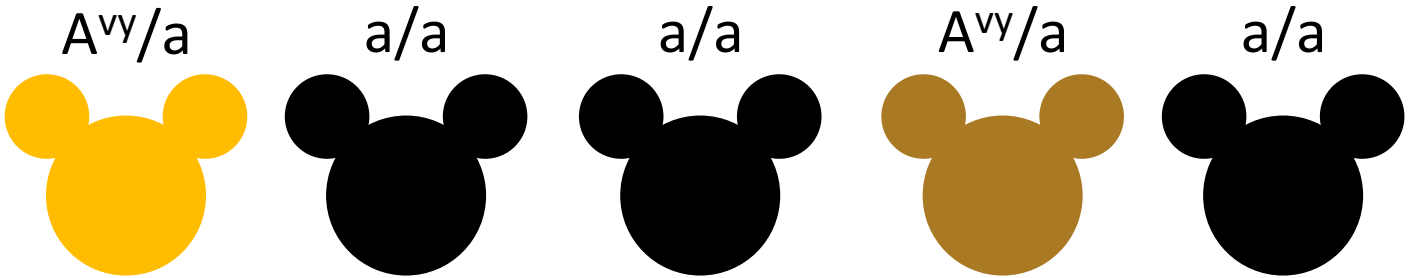


A^{vy} + 3 Mb of C3H

Experimental design



Experimental design

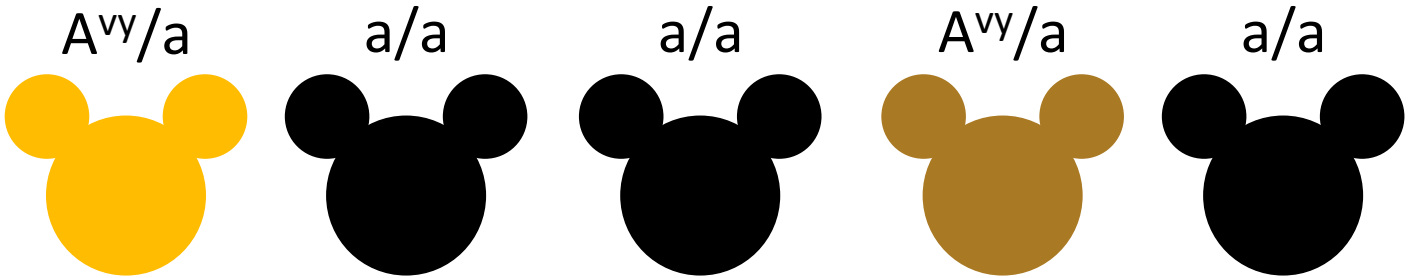


+

30× whole-genome bisulfite-sequencing

=

Experimental design



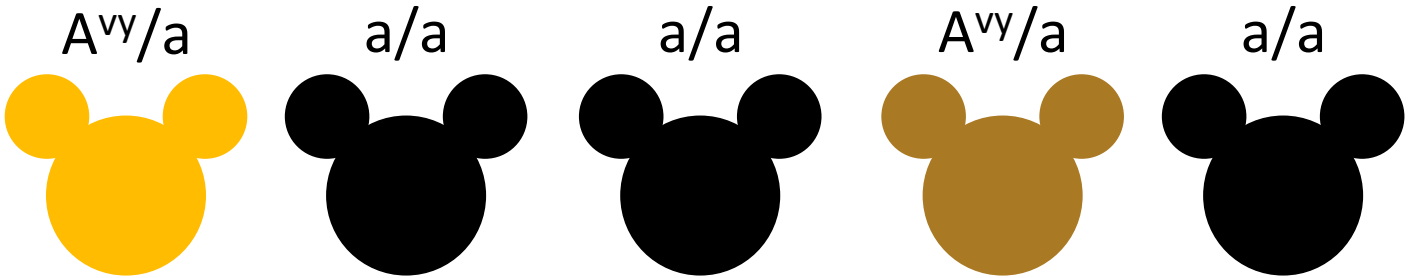
+

30× whole-genome bisulfite-sequencing

=

$$$

Experimental design



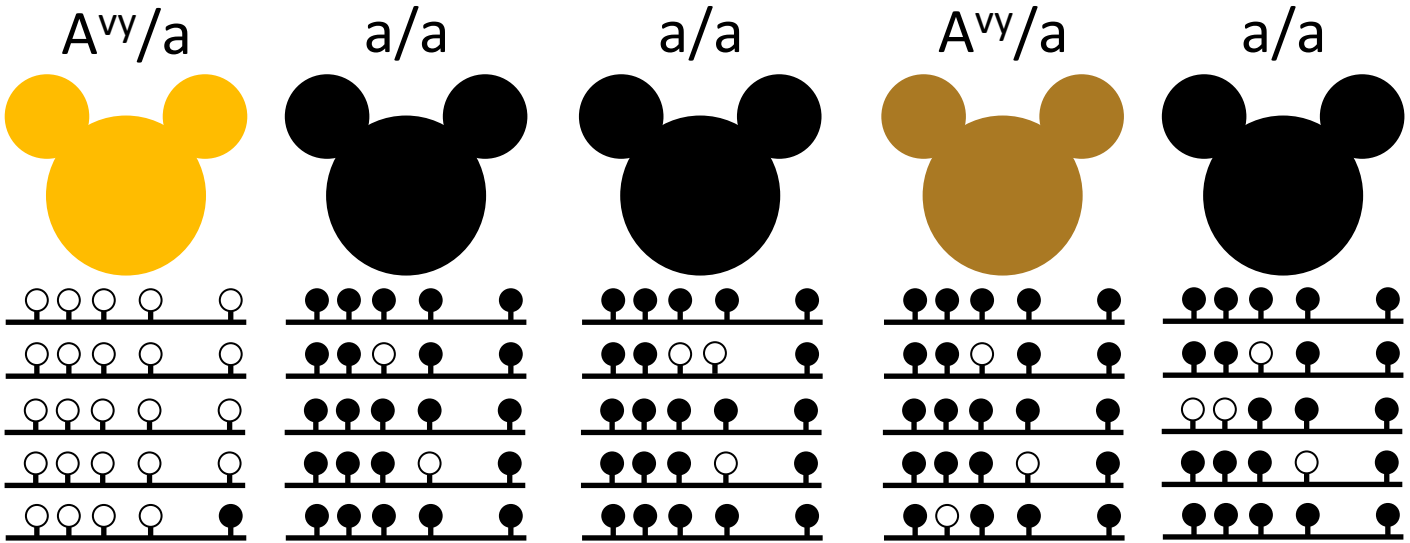
+

30× whole-genome bisulfite-sequencing

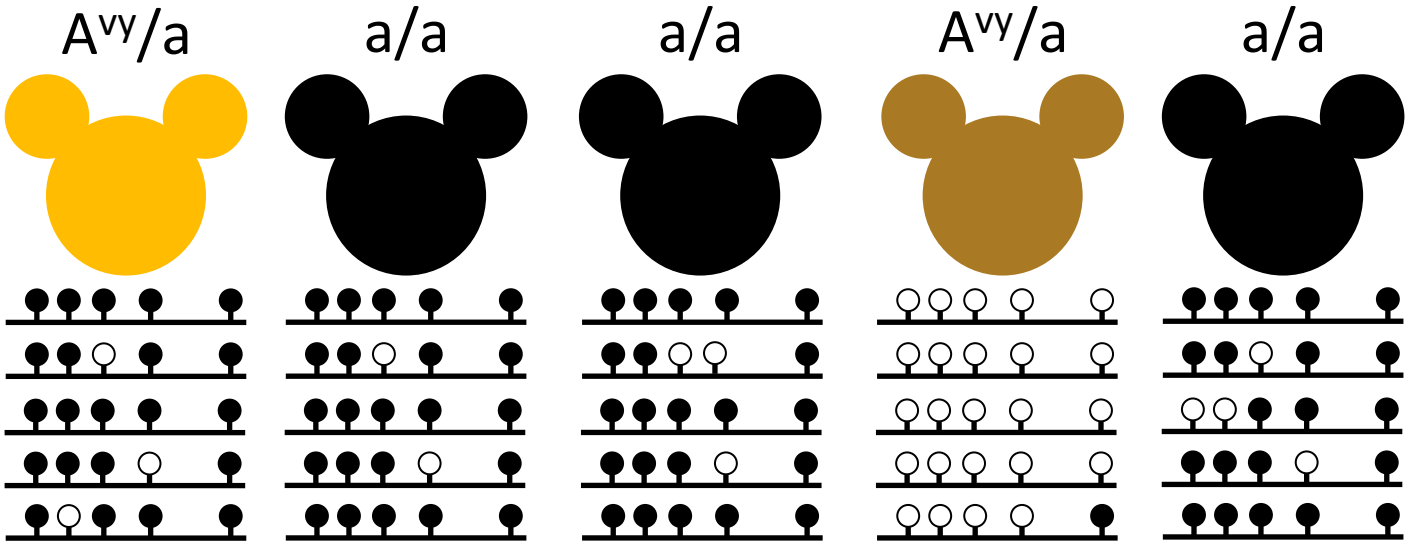
=

epialleles

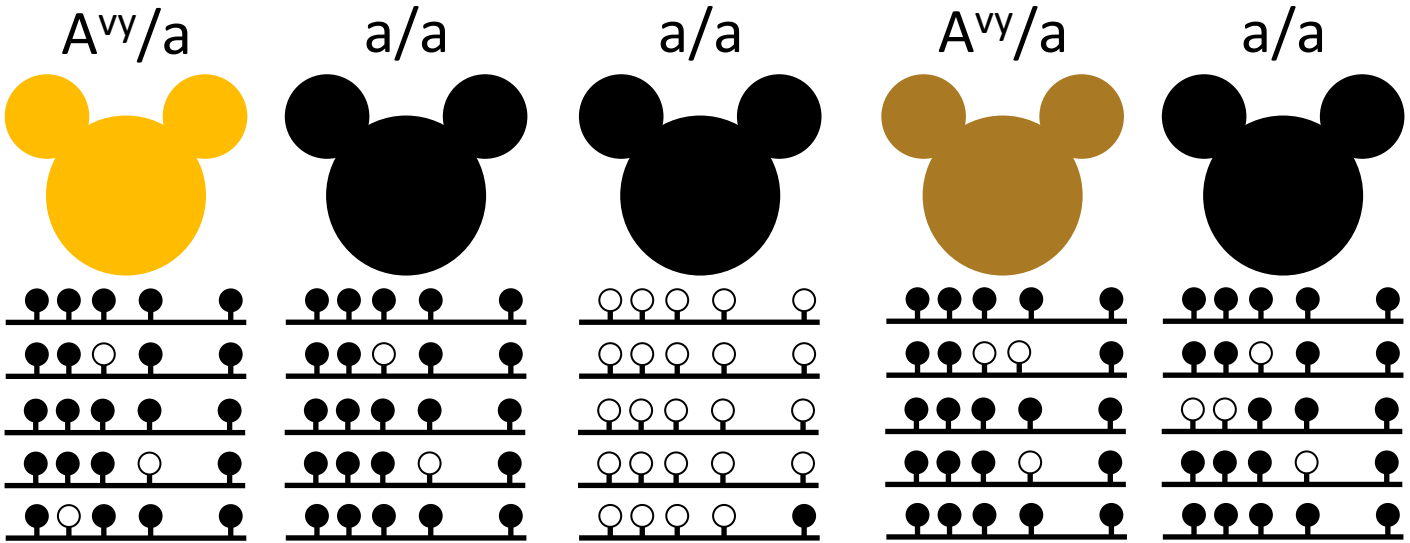
Epialleles



Epialleles



Epialleles



Method

1. Some of these mice are not like the others?
2. Is my neighbour also different?
3. Is my neighbour different in the same way as me?

Some of these mice are not like the others?

	Mouse1	Mouse2	Mouse3	Mouse4	Mouse5
Methylated	17	31	15	23	9
Unmethylated	1	3	0	1	1

P-value = 0.76

Some of these mice are not like the others?

	Mouse1	Mouse2	Mouse3	Mouse4	Mouse5
Methylated	38	79	59	69	44
Unmethylated	1	2	1	2	46

P-value = 2×10^{-25}

Some of these mice are not like the others?

	Mouse1	Mouse2	Mouse3	Mouse4	Mouse5
Methylated	38	79	59	69	44
Unmethylated	1	2	1	2	46

P-value = 2×10^{-25}

P-value < *threshold*

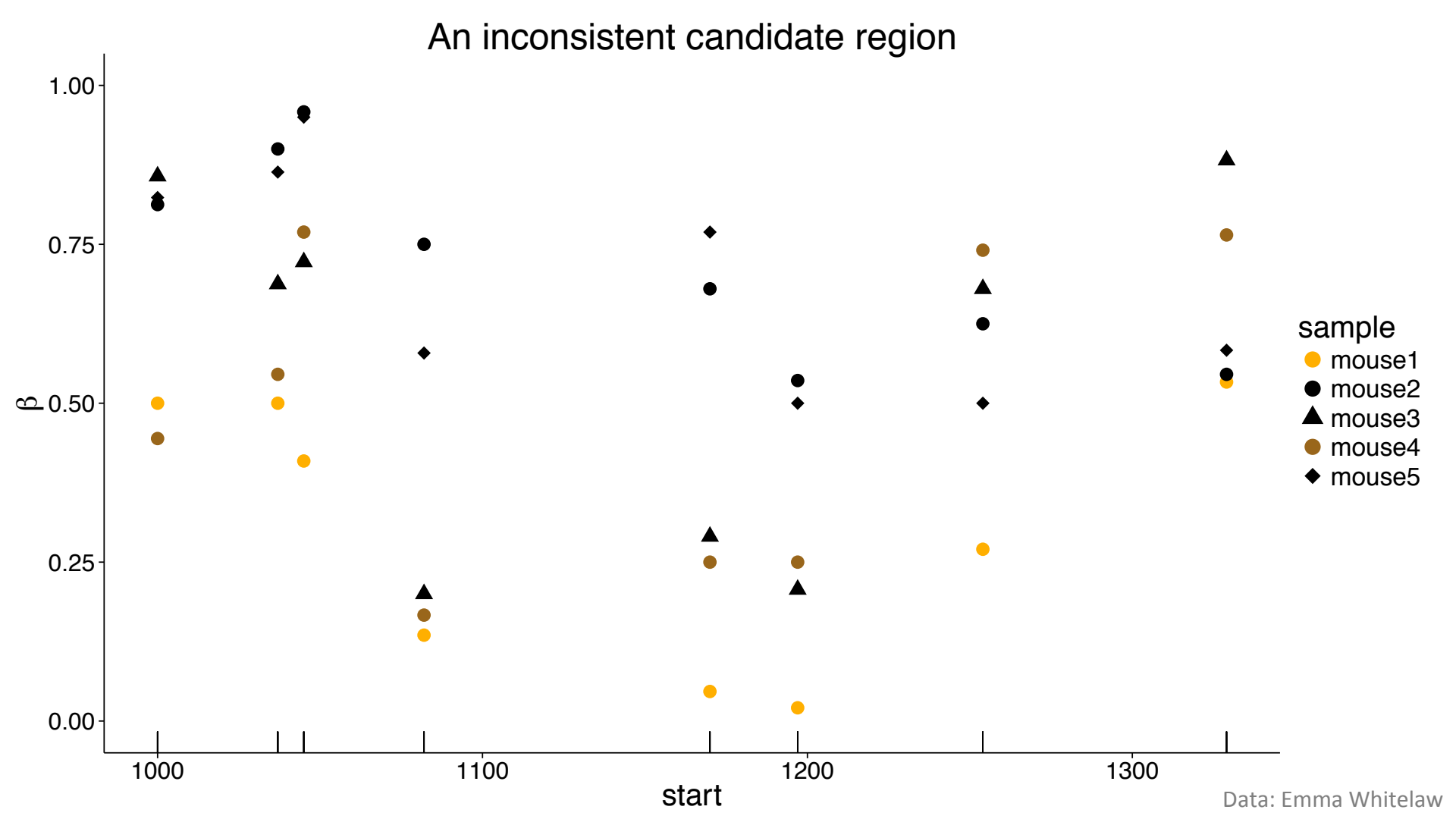
→ (candidate) differentially methylated CpG (**DMC**)

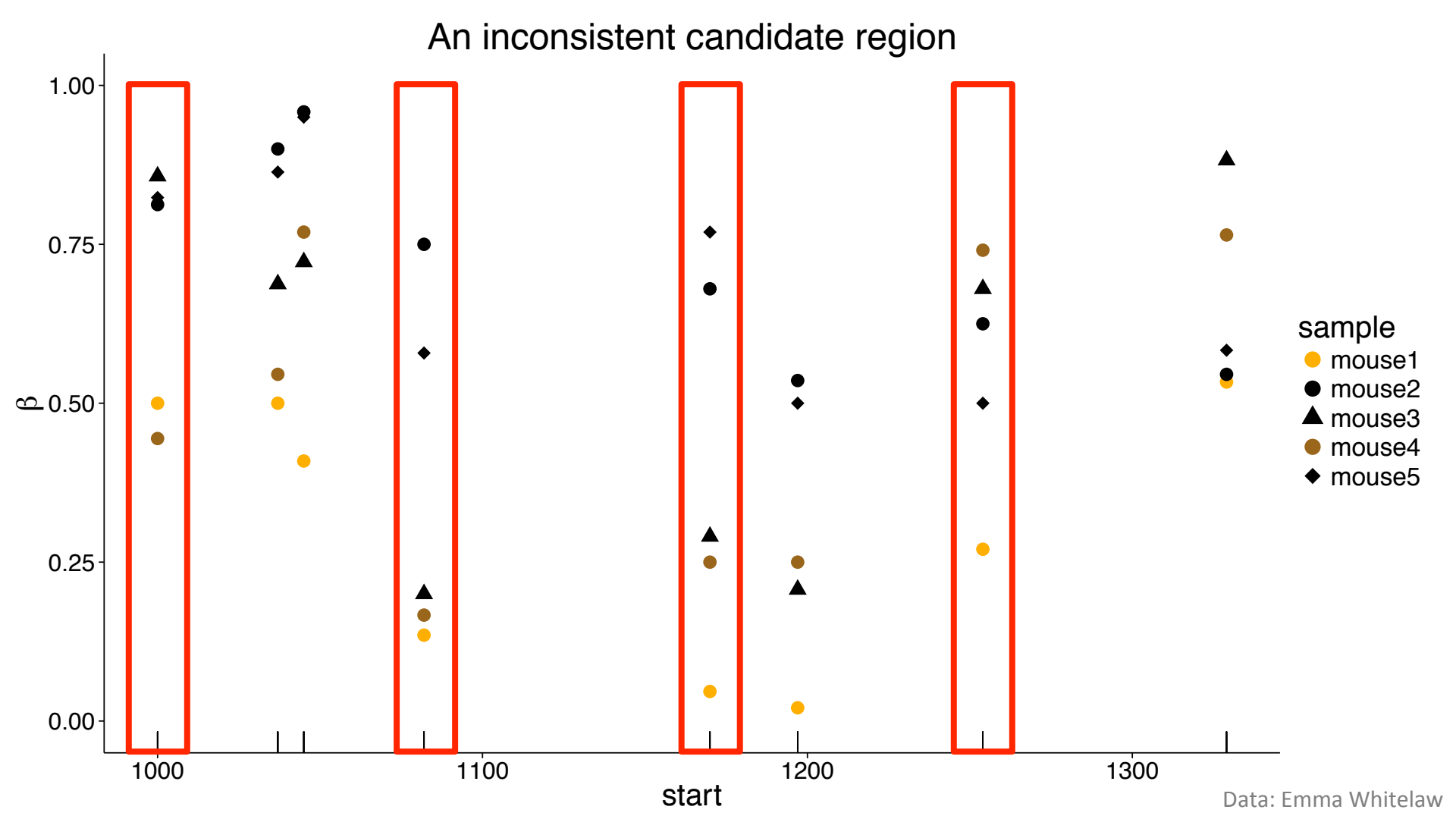
Is my neighbour also different?

Is my neighbour also different?

1. Find runs of CpGs
   - P-value < *threshold*
   - Within *distance* of next CpG
   - Some allowance for missing or “insignificant” CpGs
2. Filter candidate runs
   - Run contains enough CpGs

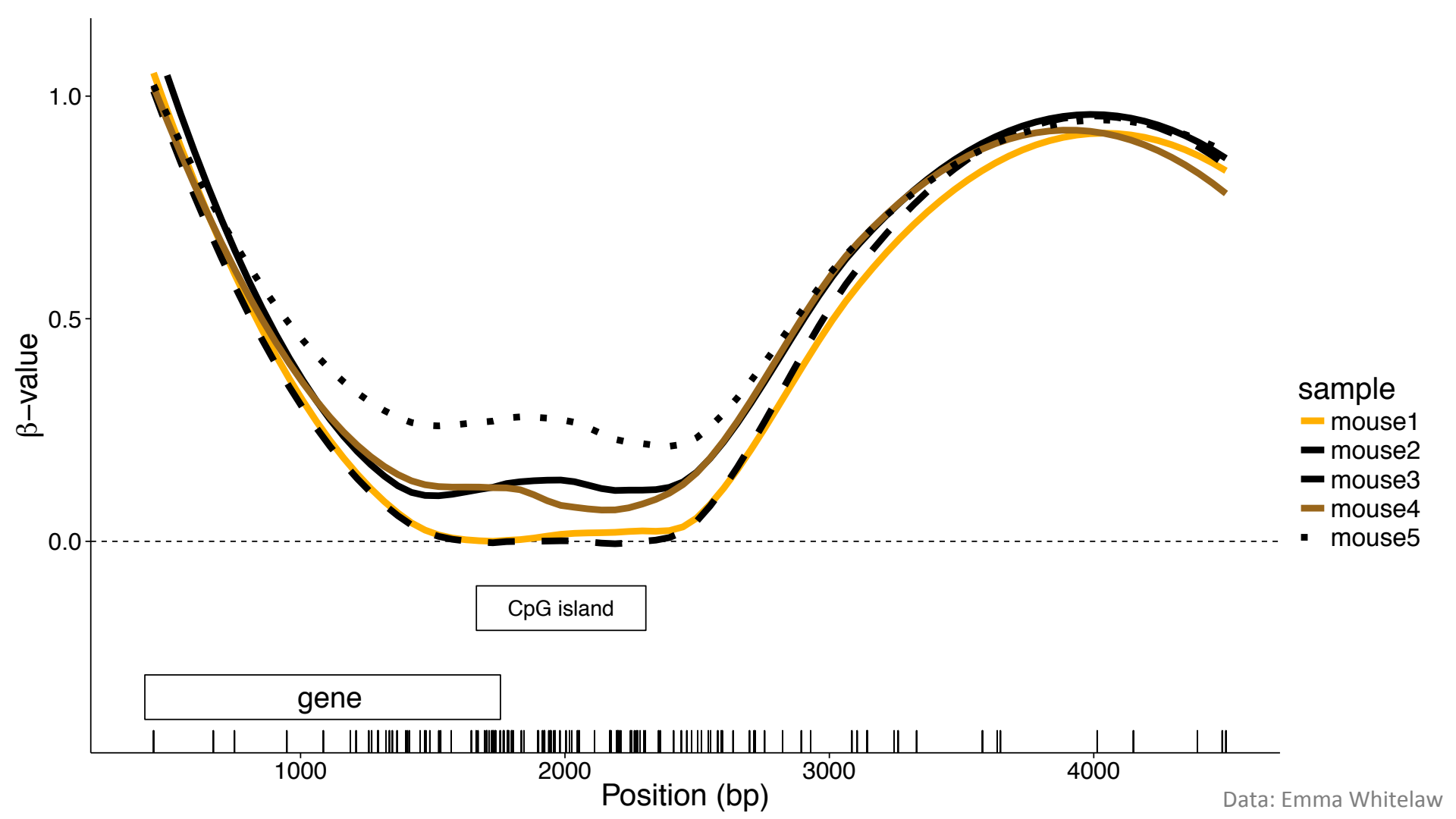
Is my neighbour different in the same way as me?


An inconsistent candidate region
1.00
0.75
0.50
0.25
0.00
β
1000
1100
1200
1300
start
sample
mouse1
mouse2
mouse3
mouse4
mouse5
Data: Emma Whitelaw


An inconsistent candidate region
1.00
0.75
0.50
0.25
0.00
β
1000
1100
1200
1300
start
sample
mouse1
mouse2
mouse3
mouse4
mouse5
Data: Emma Whitelaw

Is my neighbour different in the same way as me?

- Flag regions with 3-way interaction between sample × methylation level × position
  - Not quite what we want
  - So plot, plot, plot


1.0
0.5
0.0
β−value
CpG island
gene
1000
2000
3000
4000
Position (bp)
sample
mouse1
mouse2
mouse3
mouse4
mouse5
Data: Emma Whitelaw

How I work, what I found, and what I’m proud of

Summary

“Doesn’t the gardener lavish more care on the thorns than on the flowers”

- *Hartman* in Metamorphisis by S.Y. Agnon
Via @erichlya

“You can observe a lot by watching”

- Yogi Berra

- Robert Van Winkle

http://4.bp.blogspot.com/-N-caq8YoBaQ/UpPZs5EChnI/AAAAAAAABIw/Pb8g7jjJJGA/s1600/Stop+Collaborate.jpg

- Robert Van Winkle

a.k.a. Vanilla Ice

http://4.bp.blogspot.com/-N-caq8YoBaQ/UpPZs5EChnI/AAAAAAAABIw/Pb8g7jjJJGA/s1600/Stop+Collaborate.jpg

What I found

- Estimated strong spatial dependence of DNA methylation

What I found

- Estimated strong spatial dependence of DNA methylation
- Cell-type differences in dependence structure

What I found

- Estimated strong spatial dependence of DNA methylation
- Cell-type differences in dependence structure
- Evidence of higher order chromatin structure in spatial dependence data

What I’m proud of

www.github.com/PeteHaitch

Acknowledgements

Terry Speed

Peter Hall

Acknowledgements

Data

- Ryan Lister et al. (UWA, Salk Institute)
- Sue Clark, Aaron Statham (Garvan Institute)
- Emma Whitelaw, Harry Oey (La Trobe)
- Kasper Hansen, Rafael Irizarry (Johns Hopkins, Harvard)
- **Everyone who makes their data publicly available**

Acknowledgements

Data

- Ryan Lister et al. (UWA, Salk Institute)
- Sue Clark, Aaron Statham (Garvan Institute)
- Emma Whitelaw, Harry Oey (La Trobe)
- Kasper Hansen, Rafael Irizarry (Johns Hopkins, Harvard)
- **Everyone who makes their data publicly available**

Methodology & technology

- Kasper Hansen, Rafael Irizarry (Johns Hopkins, Harvard)
- Felix Krueger (Babraham Institute)
- Toby Sargeant (WEHI)
- Keith Satterley (WEHI)
- Bioconductor developers
- WEHI Bioinformatics
- **Everyone who makes their software open source**

Acknowledgements

Data

- Ryan Lister et al. (UWA, Salk Institute)
- Sue Clark, Aaron Statham (Garvan Institute)
- Emma Whitelaw, Harry Oey (La Trobe)
- Kasper Hansen, Rafael Irizarry (Johns Hopkins, Harvard)
- **Everyone who makes their data publicly available**

Methodology & technology

- Kasper Hansen, Rafael Irizarry (Johns Hopkins, Harvard)
- Felix Krueger (Babraham Institute)
- Toby Sargeant (WEHI)
- Keith Satterley (WEHI)
- Bioconductor developers
- WEHI Bioinformatics
- **Everyone who makes their software open source**

Funding: APA and VLSCI

Acknowledgements

Data

- Ryan Lister et al. (UWA, Salk Institute)
- Sue Clark, Aaron Statham (Garvan Institute)
- Emma Whitelaw, Harry Oey (La Trobe)
- Kasper Hansen, Rafael Irizarry (Johns Hopkins, Harvard)
- **Everyone who makes their data publicly available**

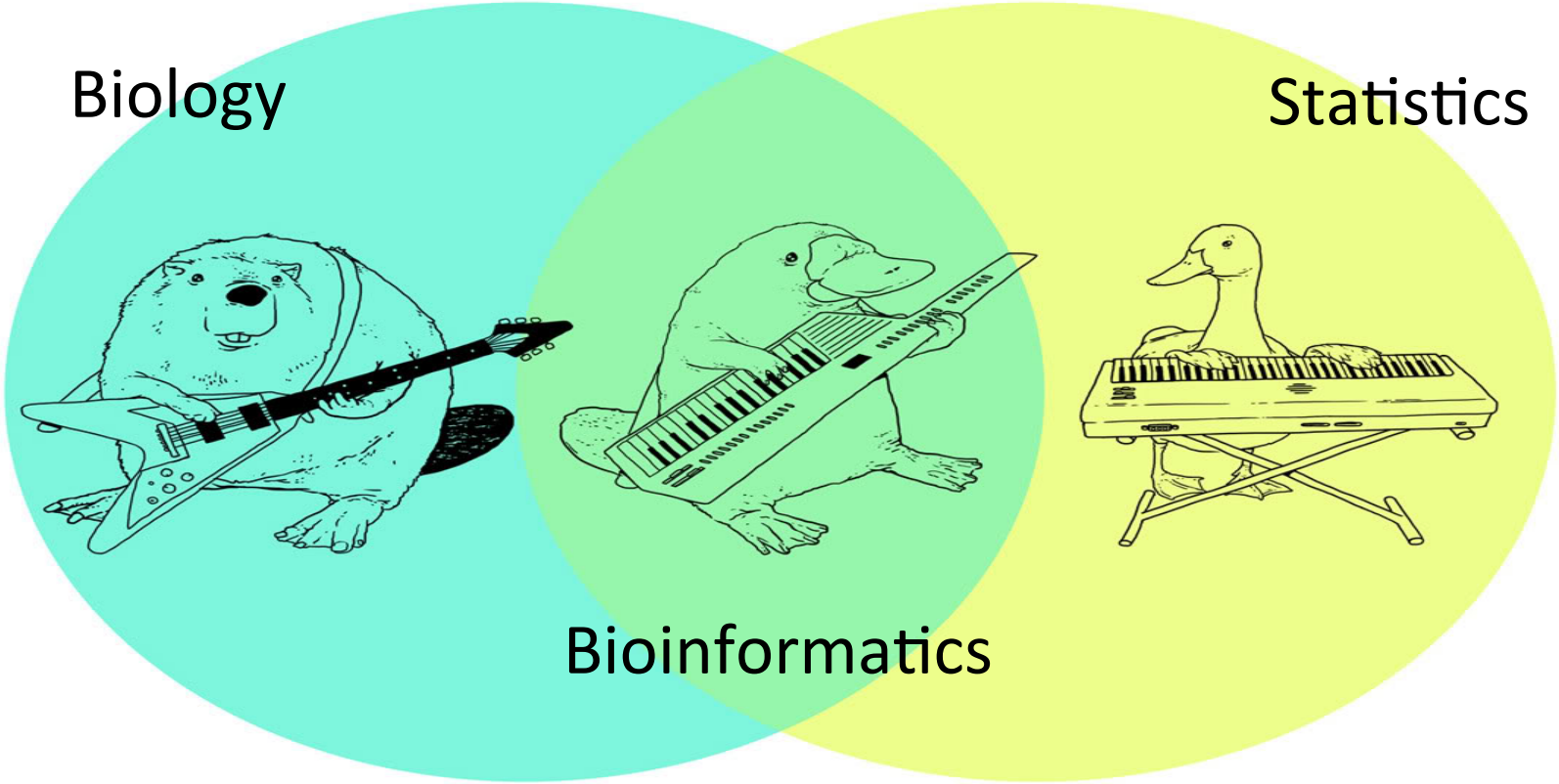
Methodology & technology

- Kasper Hansen, Rafael Irizarry (Johns Hopkins, Harvard)
- Felix Krueger (Babraham Institute)
- Toby Sargeant (WEHI)
- Keith Satterley (WEHI)
- Bioconductor developers
- WEHI Bioinformatics
- **Everyone who makes their software open source**

Funding: APA and VLSCI

Sanity: Family and friends

http://s18.photobucket.com/user/endriquelimones/media/tensographics_math.jpg.html



www.peterhickey.org/presentations