Accurate inference of DNA methylation data:
Statistical challenges lead to biological insights

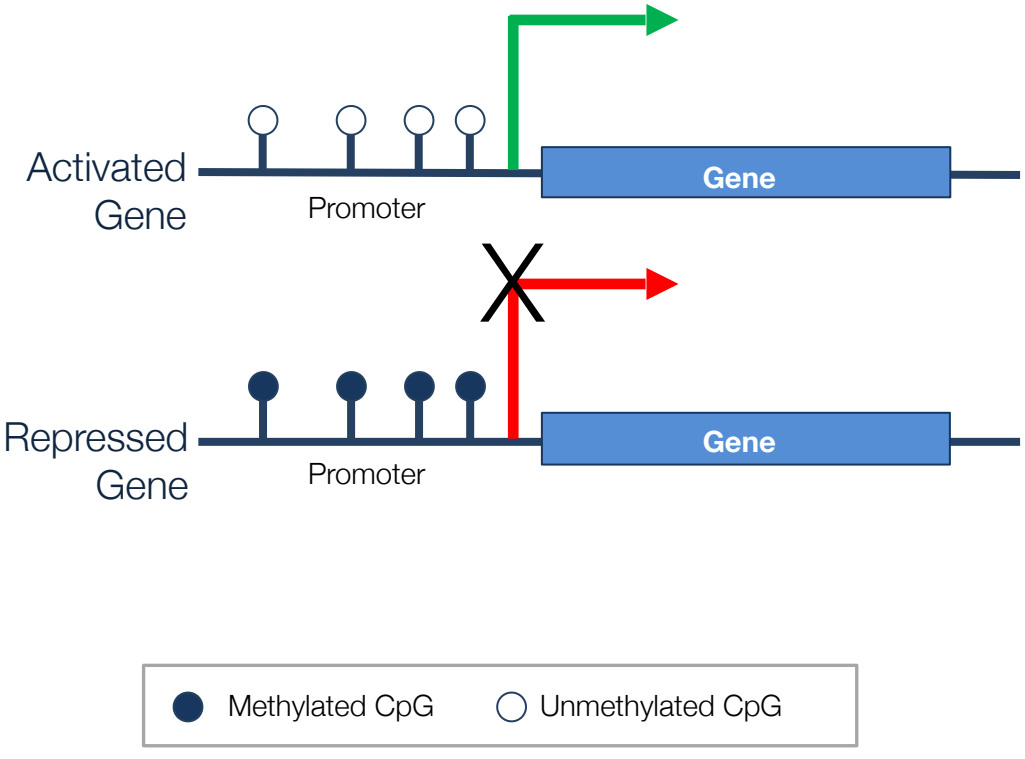
Keegan Korthauer, PhD
Postdoctoral Research Fellow

CFCE Seminar
Dana-Farber Cancer Institute
15 February 2019

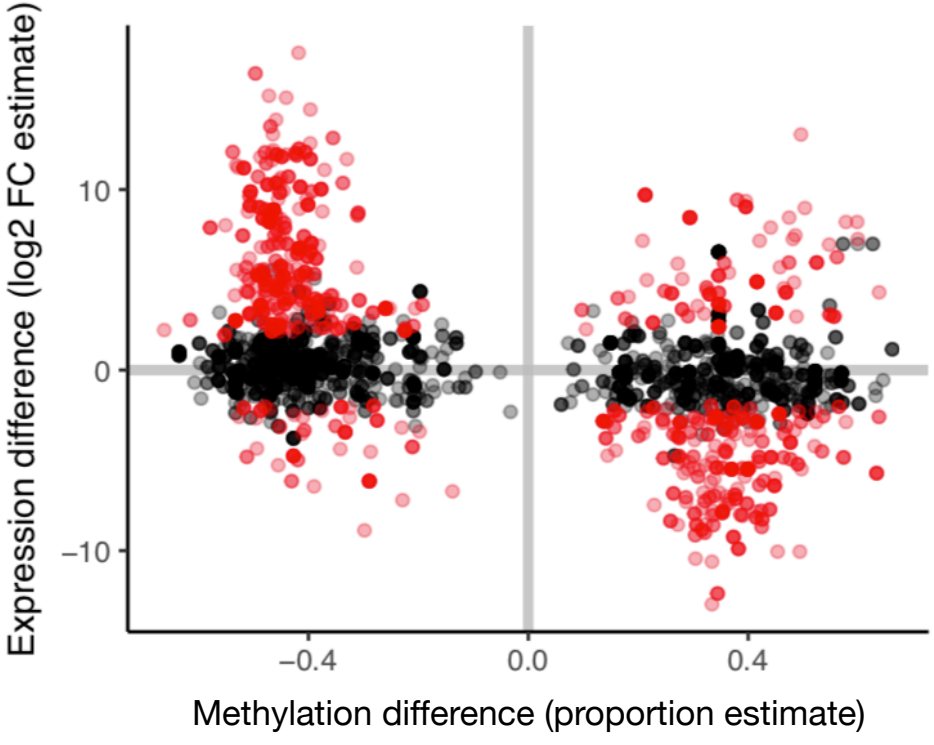
DANA-FARBER CANCER INSTITUTE

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

Role of DNA methylation in transcriptional regulation



Correlation or causation?



Korthauer et al., 2018 (*Biostatistics*)

First genome-wide study of causality

New Results – September 2017

Cold Spring Harbor Laboratory

bioRχiv

THE PREPRINT SERVER FOR BIOLOGY

Frequent lack of repressive capacity of promoter DNA methylation identified through genome-wide epigenomic manipulation

Ethan Edward Ford, Matthew R. Grimmer, Sabine Stolzenburg, Ozren Bogdanovic, Alex de Mendoza, Peggy J. Farnham, Pilar Blancafort, Ryan Lister

doi: https://doi.org/10.1101/170506

First genome-wide study of causality

New Results – September 2017

Frequent lack of repressive capacity of promoter DNA methylation identified through genome-wide epigenomic manipulation

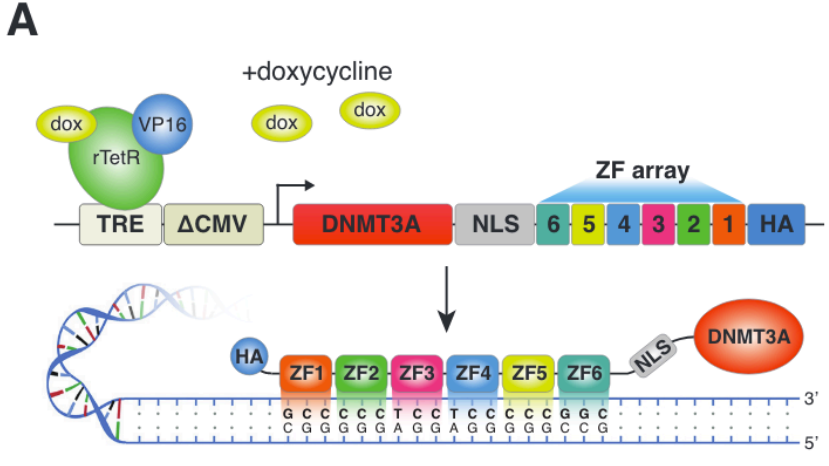
Ethan Edward Ford, Matthew R. Grimmer, Sabine Stolzenburg, Ozren Bogdanovic, Alex de Mendoza, Peggy J. Farnham, Pilar Blancafort, Ryan Lister

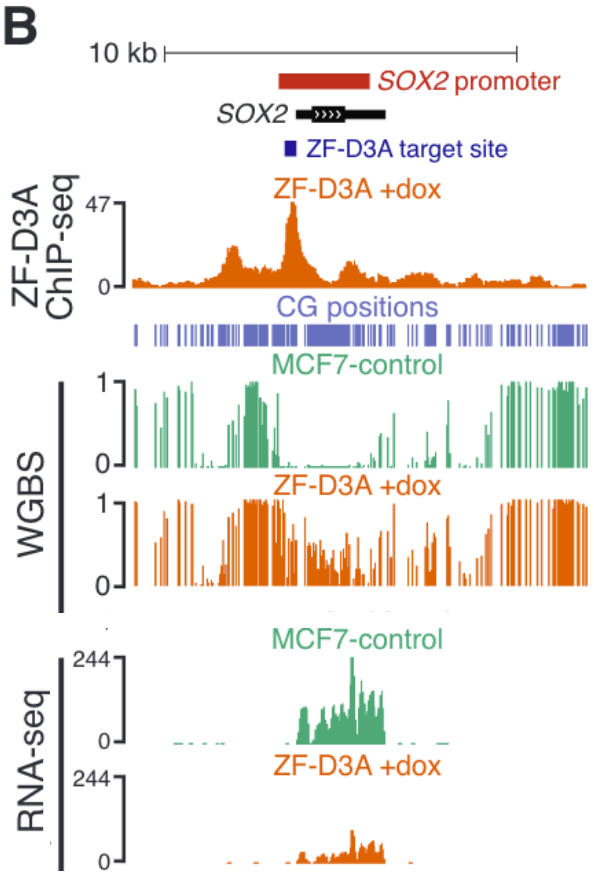
doi: https://doi.org/10.1101/170506

"promoter DNA methylation is **not generally sufficient** for transcriptional inactivation"

Forcible methylation of promoters

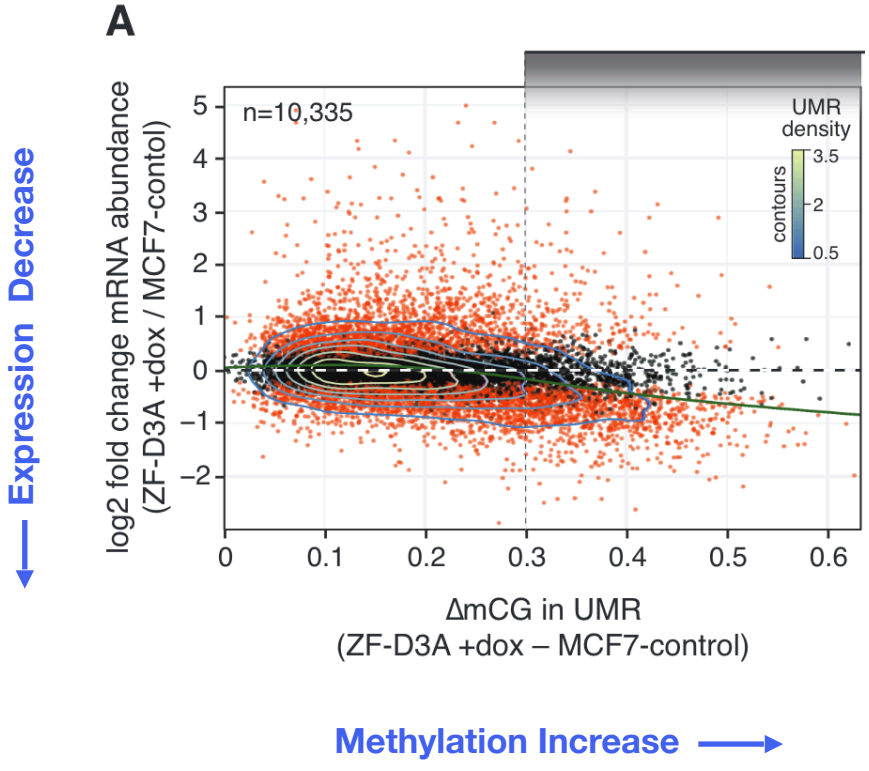
Figure 1 from Ford et al., 2017 (*bioRxiv*)





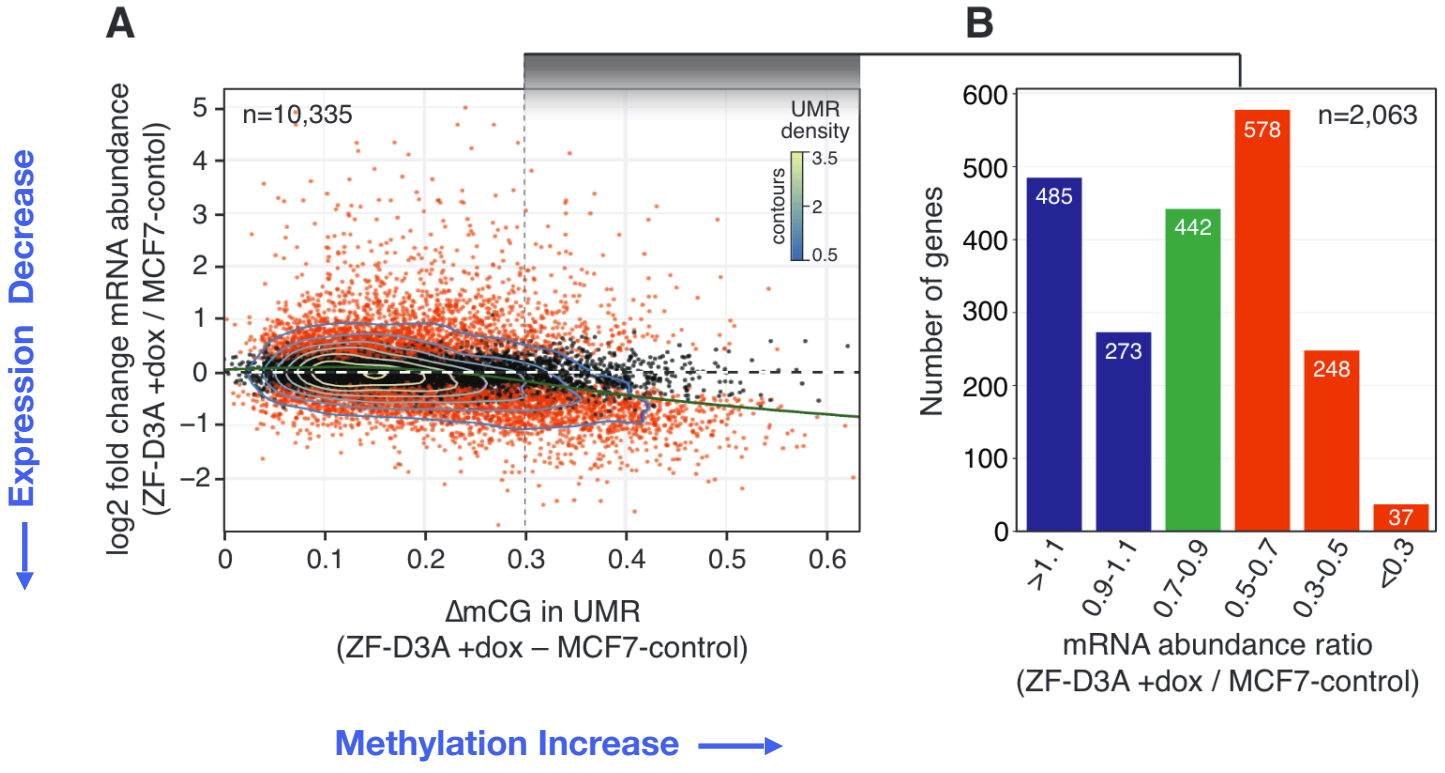
Conclusion: methylation not generally sufficient for gene repression

Figure 5 from Ford et al., 2017 (*bioRxiv*)



Conclusion: methylation not generally sufficient for gene repression

Figure 5 from Ford et al., 2017 (*bioRxiv*)



Statistical challenges

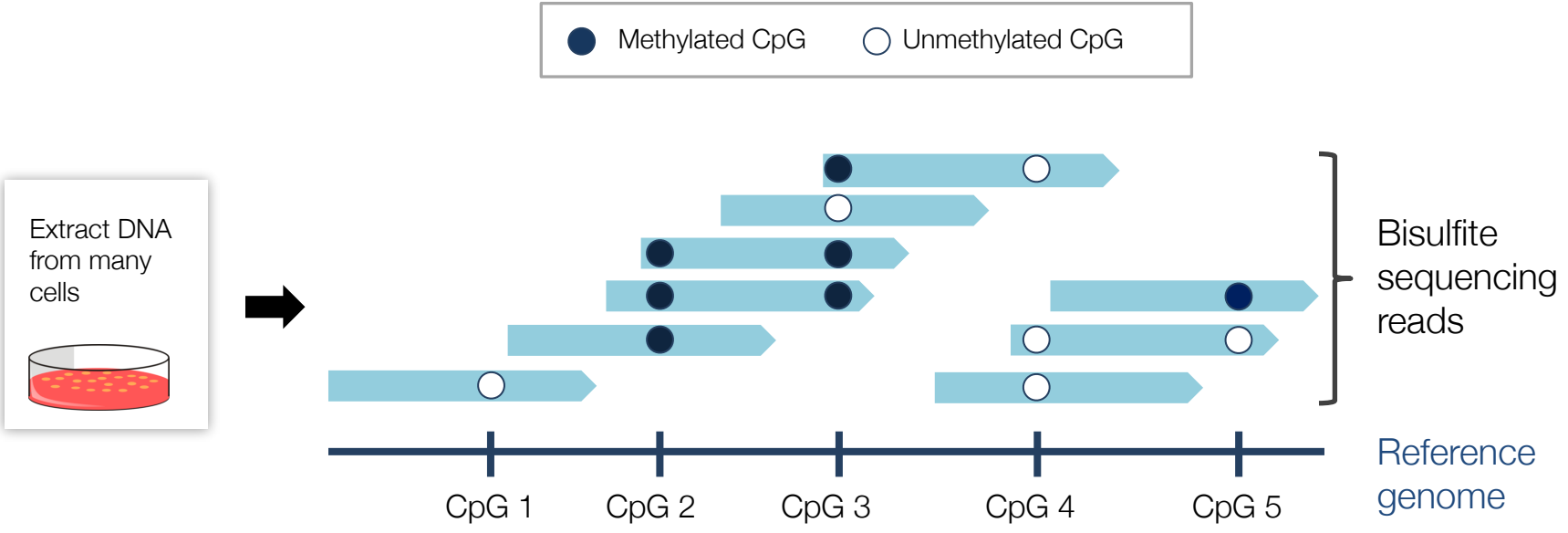
Challenges of methylation sequencing analysis

1. Small sample sizes
2. Region-level inference
3. Biological and spatial variability

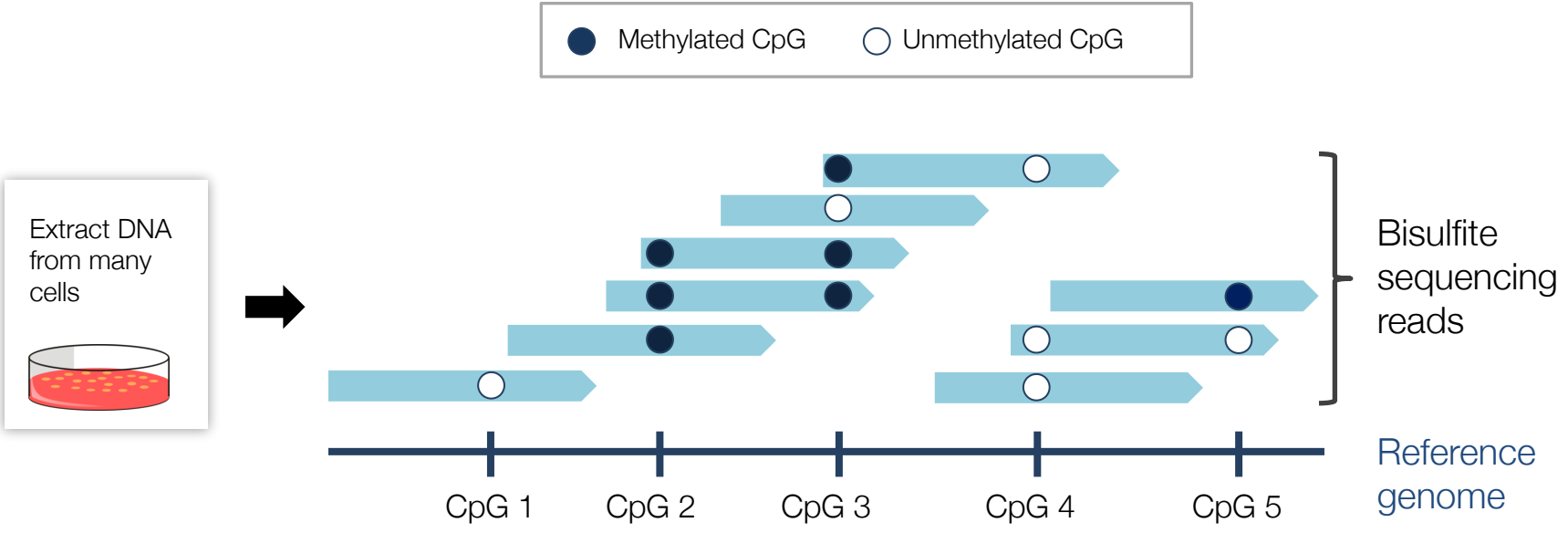
Challenges of methylation sequencing analysis

1. Small sample sizes
2. Region-level inference
3. Biological and spatial variability

Whole genome bisulfite sequencing (WGBS)



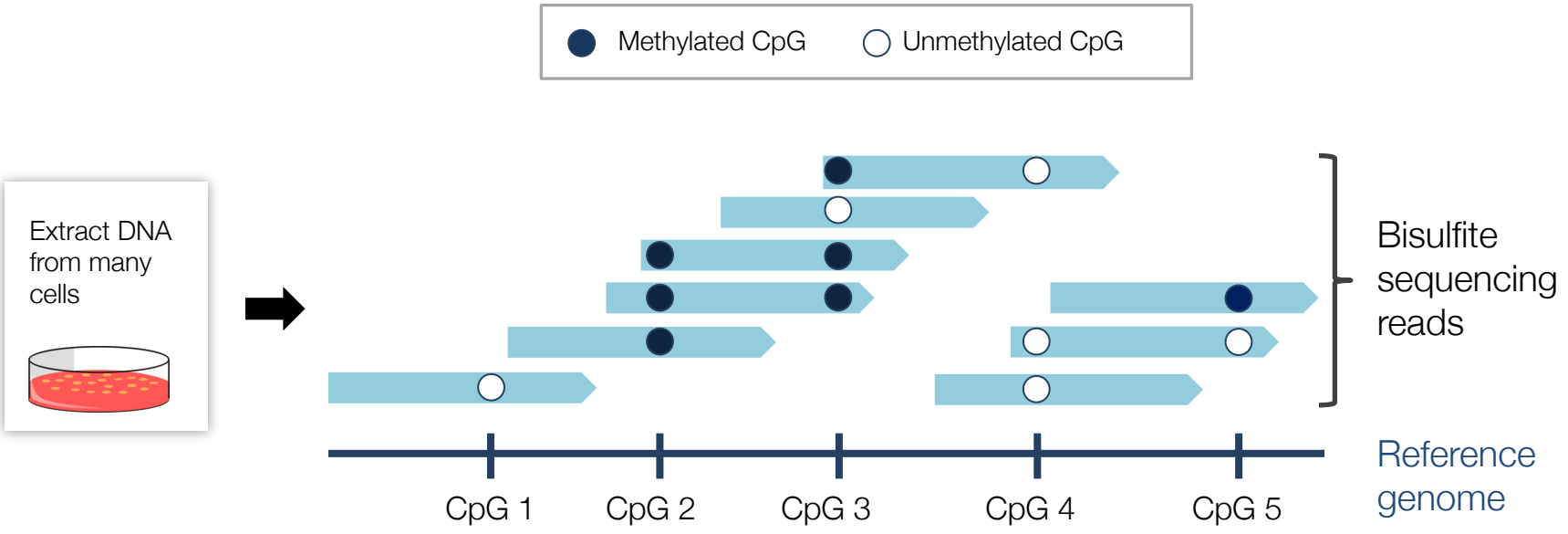
Whole genome bisulfite sequencing (WGBS)



Methylated Count (M)	0	3	3	0	1
Coverage (N)	1	3	4	3	2
Proportion (M/N)	0	1	0.75	0	0.50

Methylation sequencing data

Whole genome bisulfite sequencing (WGBS)

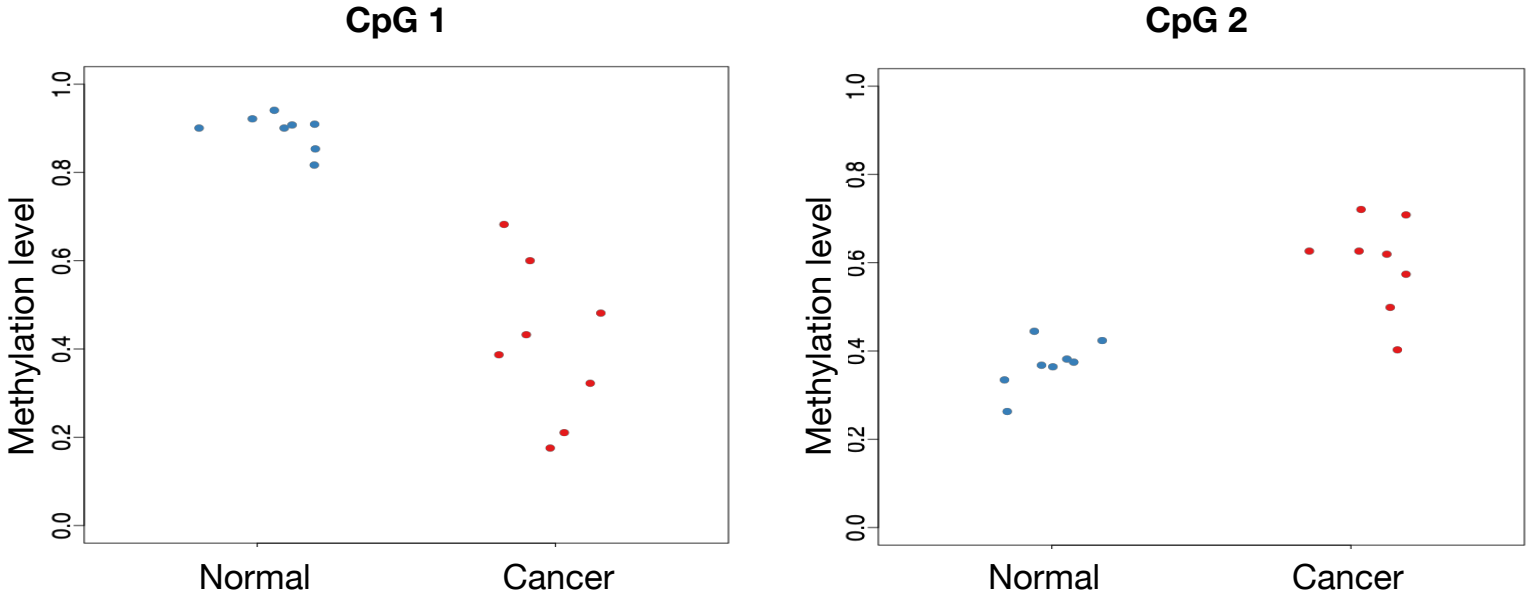


Methylated Count (M)	0	3	3	0	1
Coverage (N)	1	3	4	3	2
Proportion (M/N)	0	1	0.75	0	0.50

Methylation sequencing data

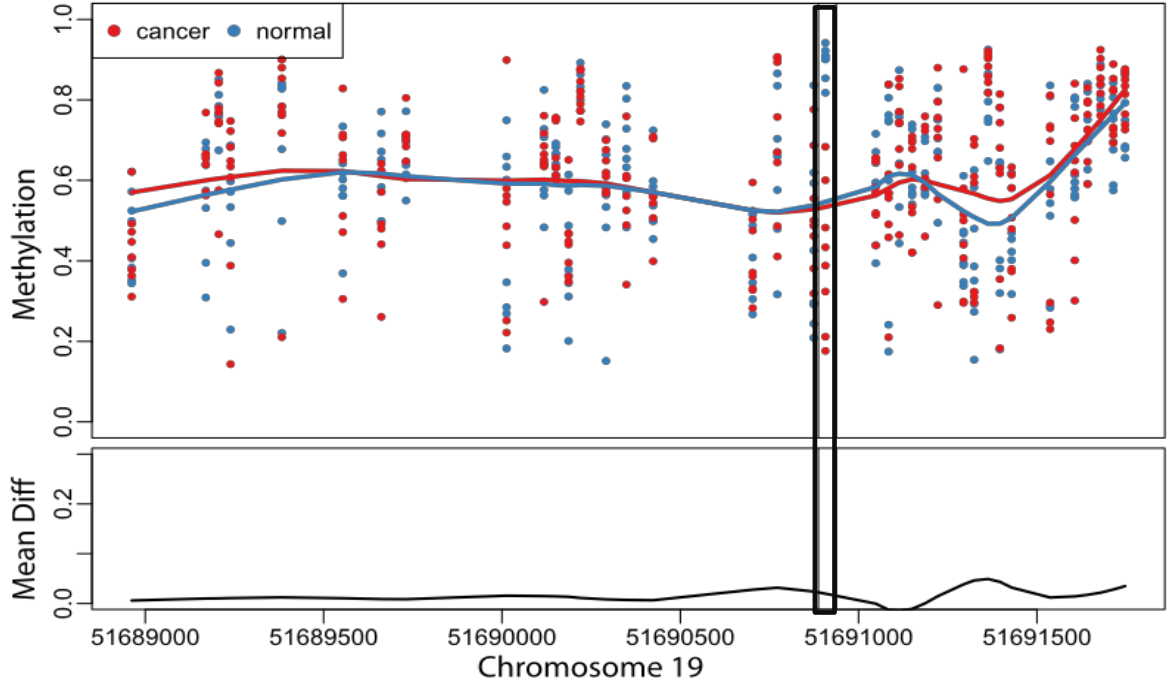
⚠ WGBS cost ≈ WGS cost

Differential methylation of individual CpGs



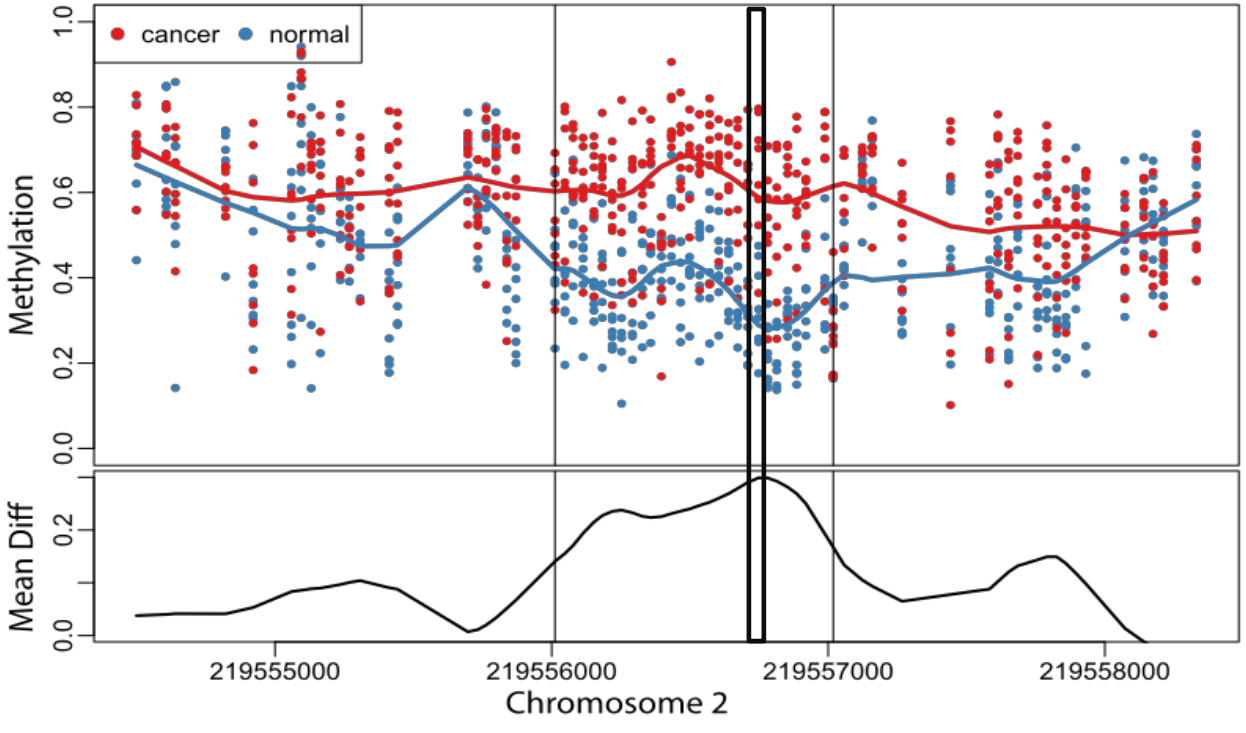
Adapted from Jaffe et al., 2012 (*Int J Epidemiol*)

CpG 1



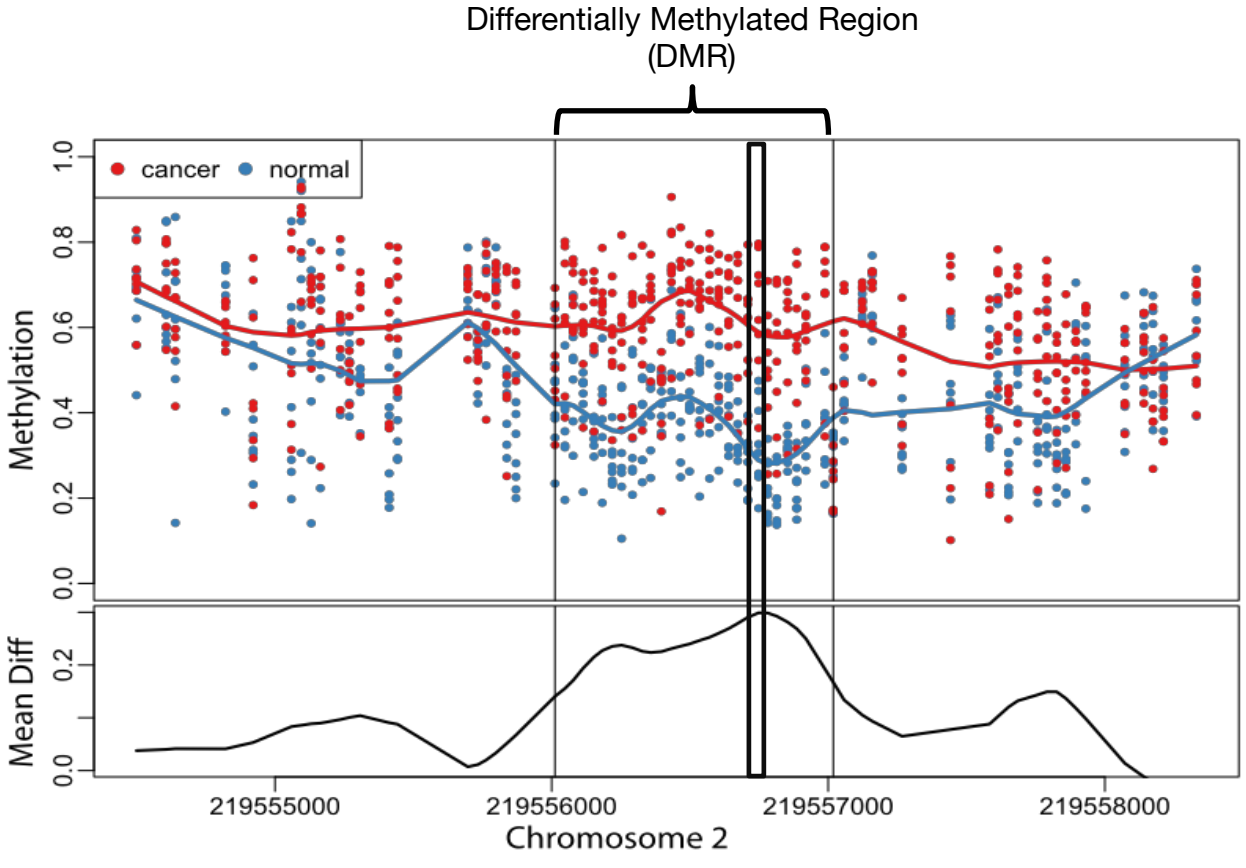
Adapted from Jaffe et al., 2012 (*Int J Epidemiol*)

CpG 2



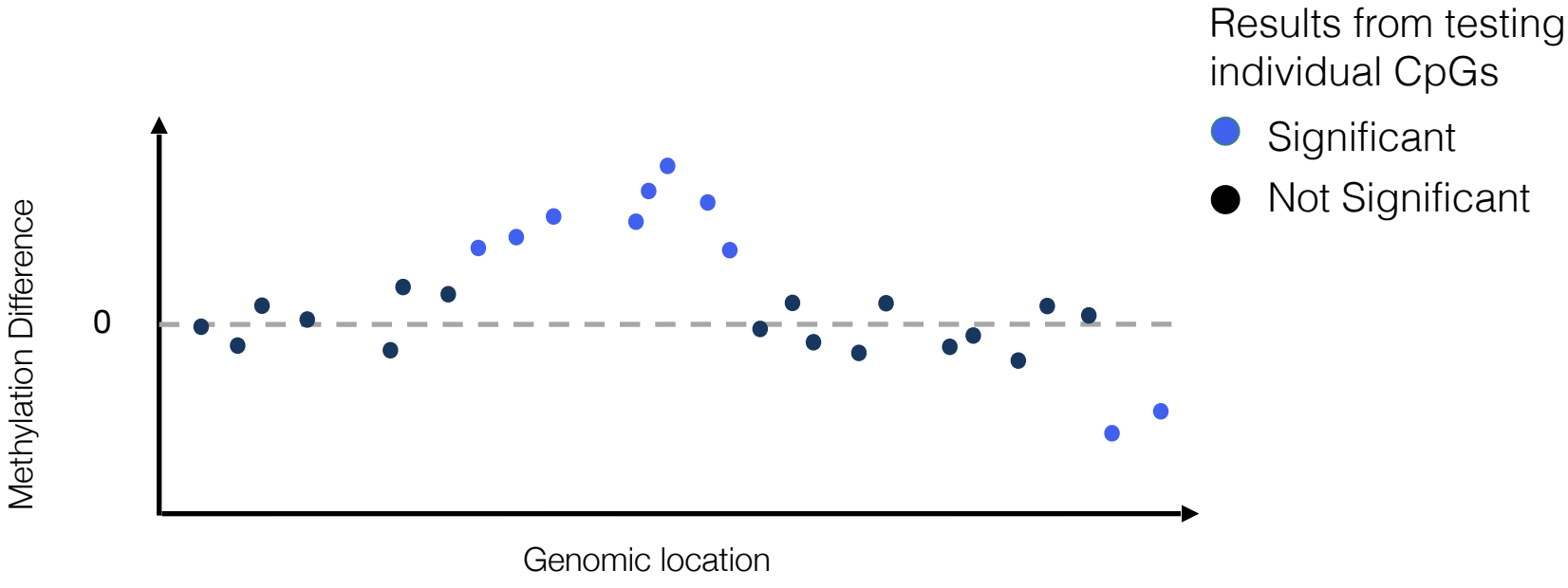
Adapted from Jaffe et al., 2012 (*Int J Epidemiol*)

CpG 2



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

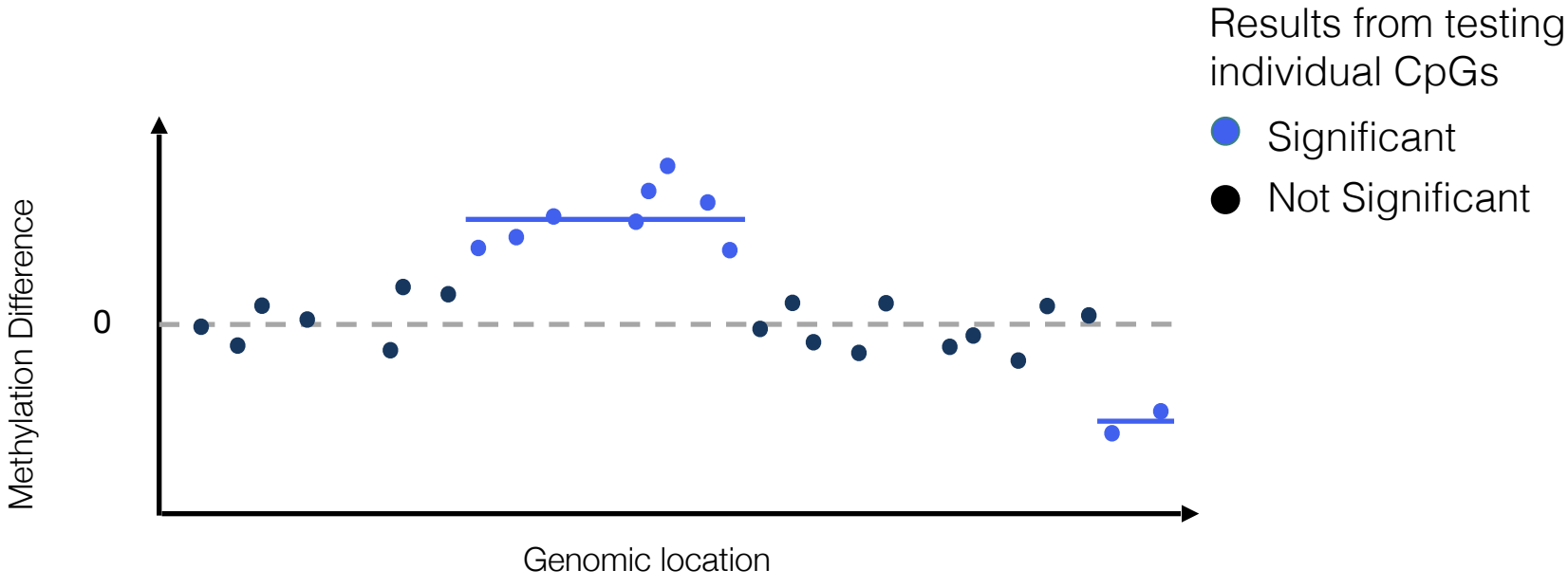
Previous methods: Grouping significant CpGs



Examples:

- Bsmooth (Hansen et al., 2012)
- DSS (Feng et al., 2014; Wu et al., 2015) – used by Ford et al.

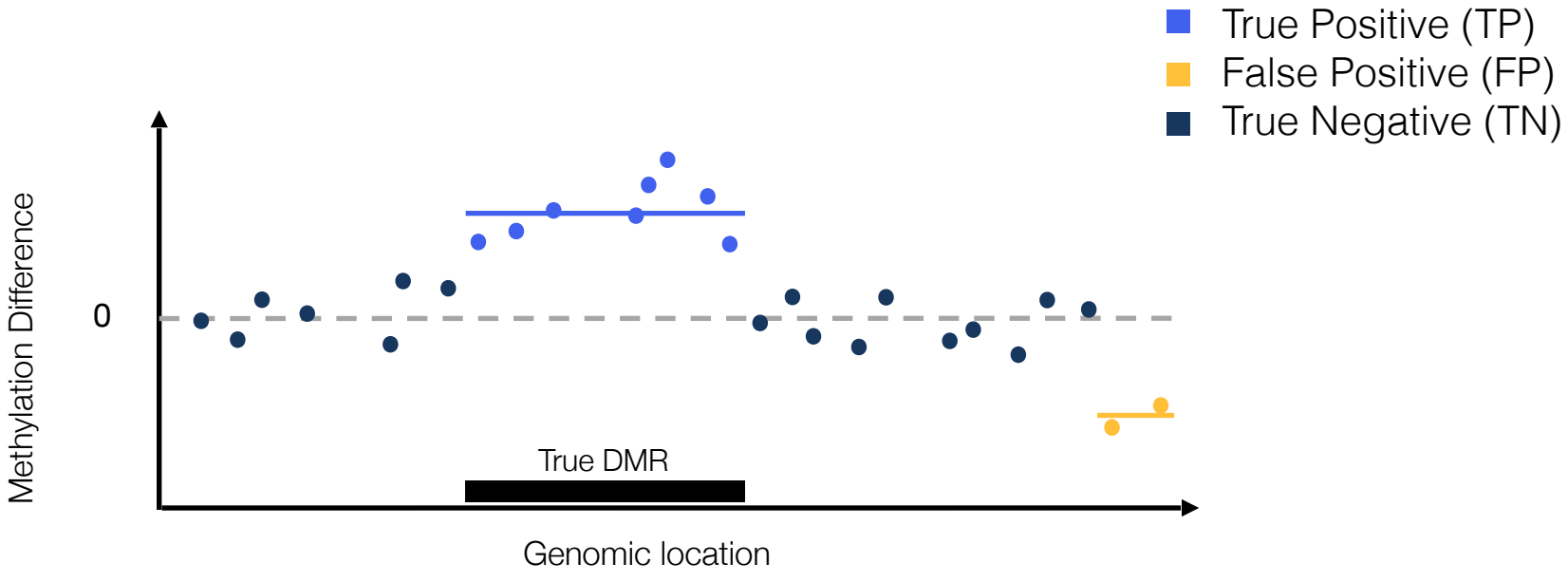
Previous methods: Grouping significant CpGs



Examples:

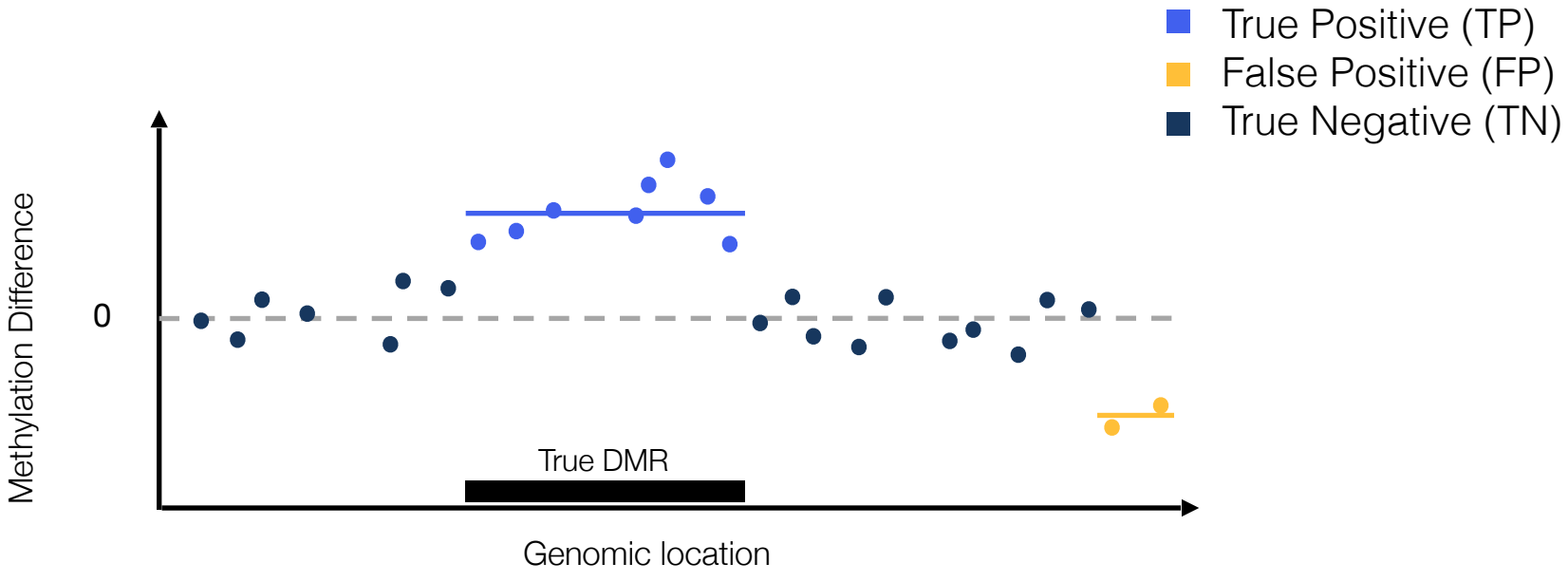
- Bsmooth (Hansen et al., 2012)
- DSS (Feng et al., 2014; Wu et al., 2015) – used by Ford et al.

Error rate not controlled at the region level



$$\text{False Discovery Rate (FDR)} = \mathrm{E}\left[\frac{\mathrm{FP}}{\mathrm{TP}+\mathrm{FP}}\right]$$

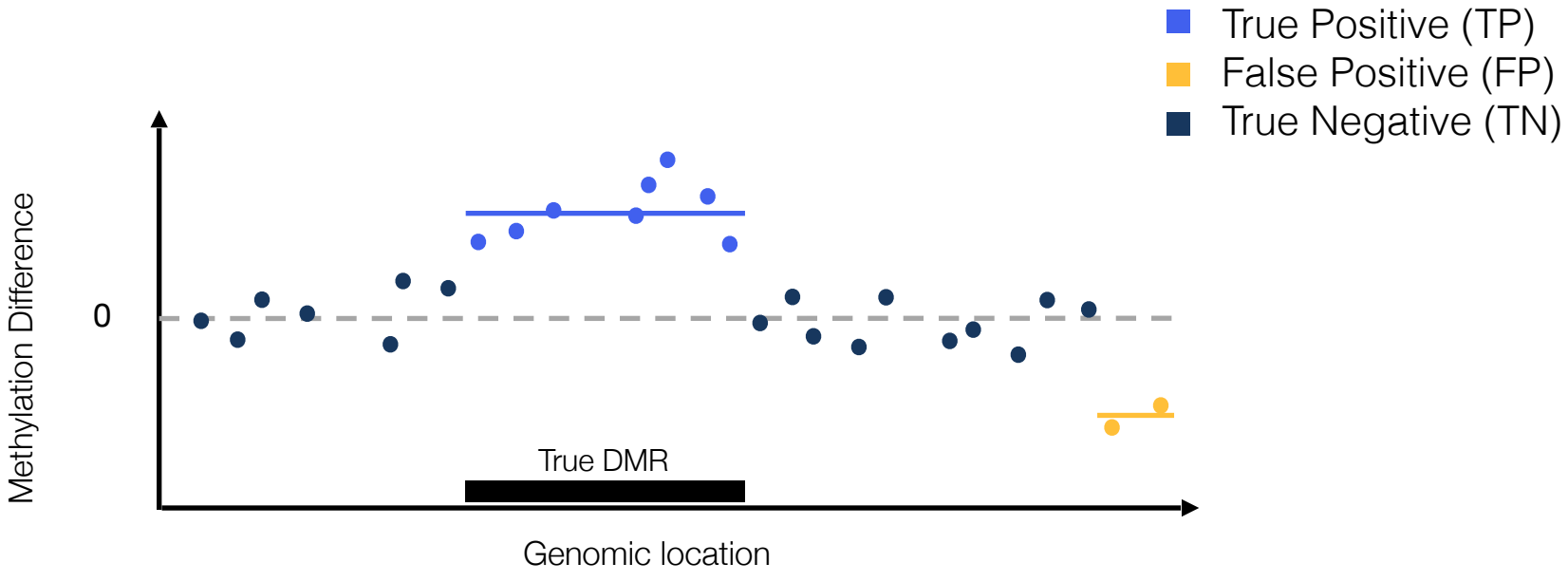
Error rate not controlled at the region level



$$\text{False Discovery Rate (FDR)} = \mathrm{E}\left[\frac{\mathrm{FP}}{\mathrm{TP}+\mathrm{FP}}\right]$$

$$\hat{FDR}_{CpG} = \frac{2}{10} = 0.2$$

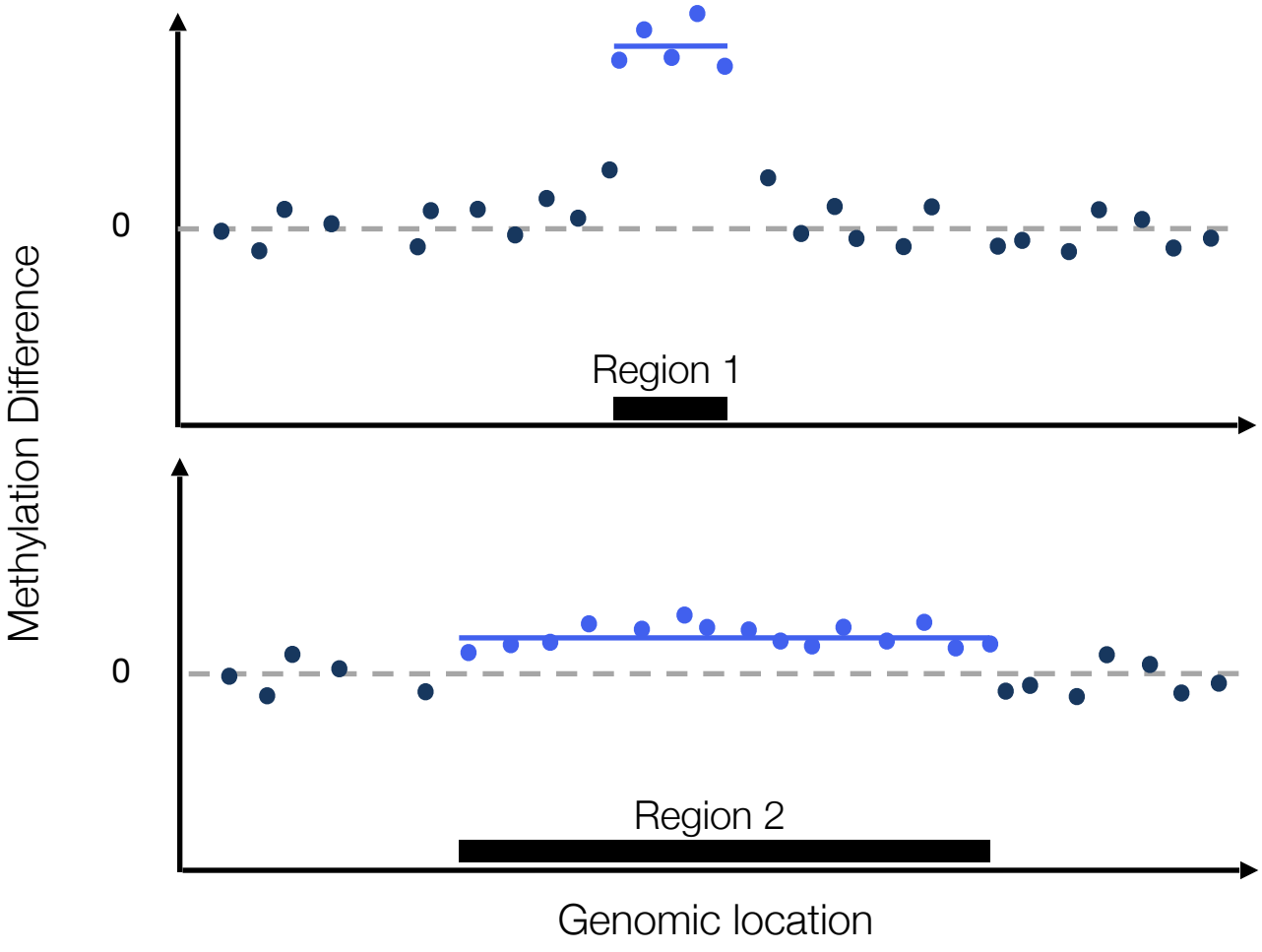
Error rate not controlled at the region level



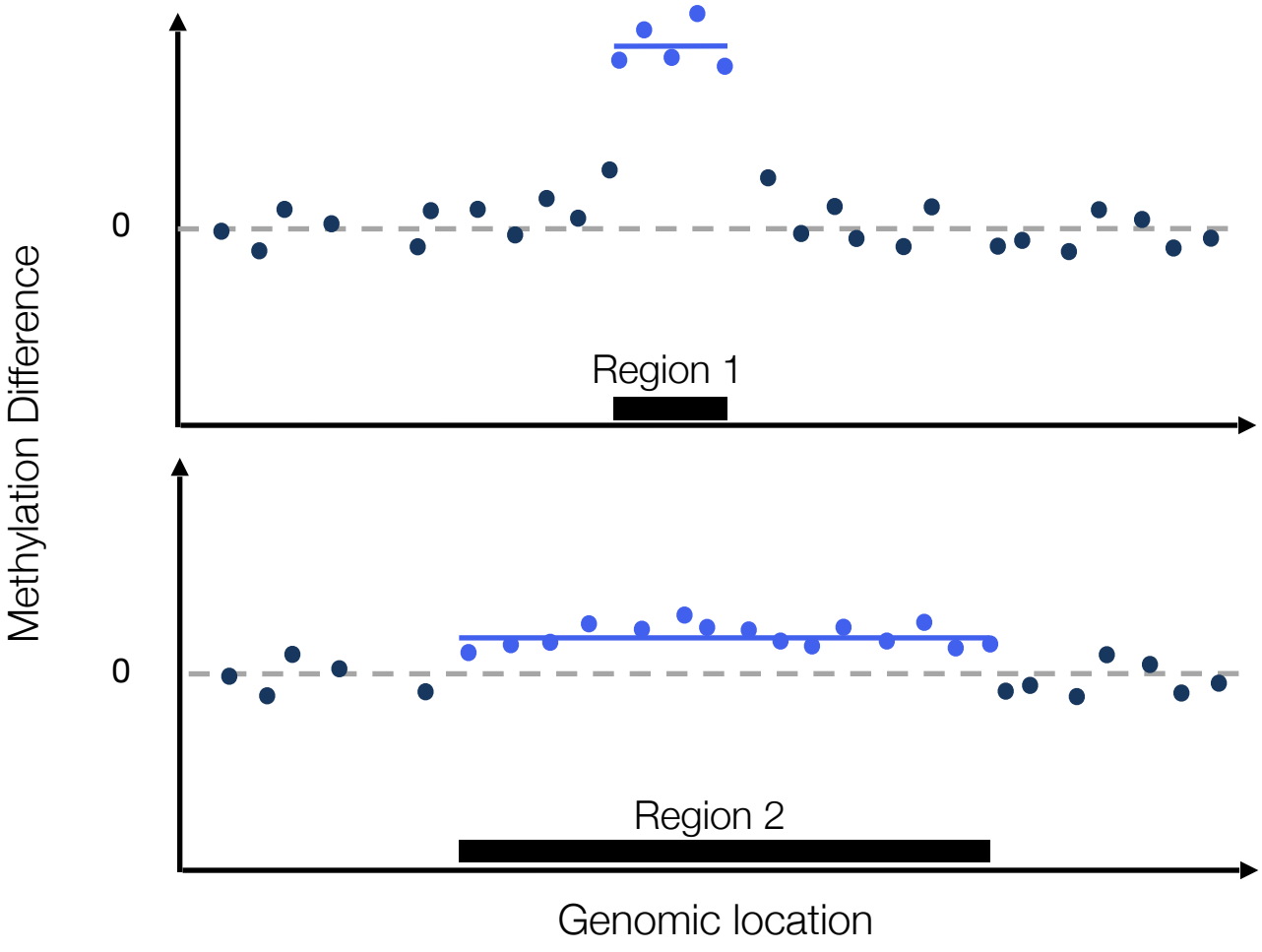
$$\text{False Discovery Rate (FDR)} = \mathrm{E}\left[\frac{\mathrm{FP}}{\mathrm{TP}+\mathrm{FP}}\right]$$

$$\hat{FDR}_{CpG} = \frac{2}{10} = 0.2 \quad vs \quad \hat{FDR}_{DMR} = \frac{1}{2} = 0.5$$

Spatial Variability



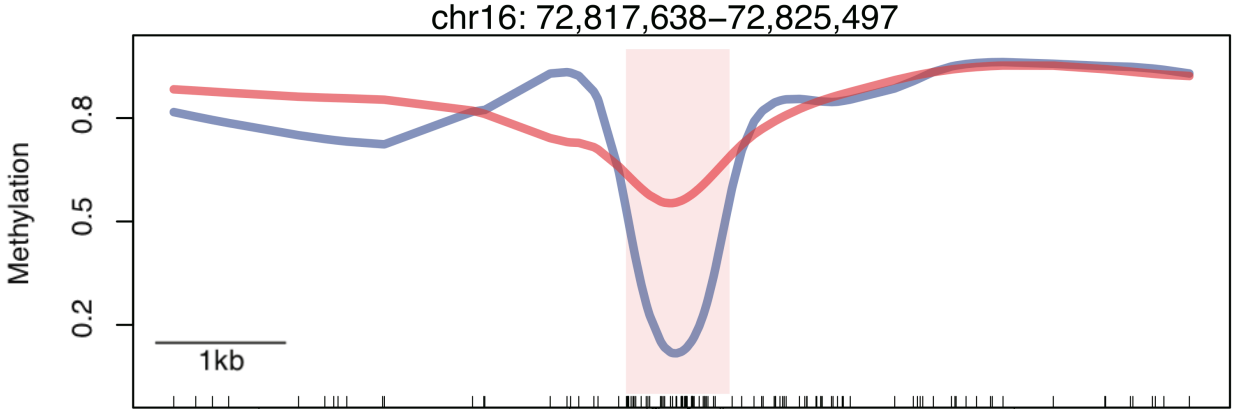
Spatial Variability



Prioritized by mean difference statistics

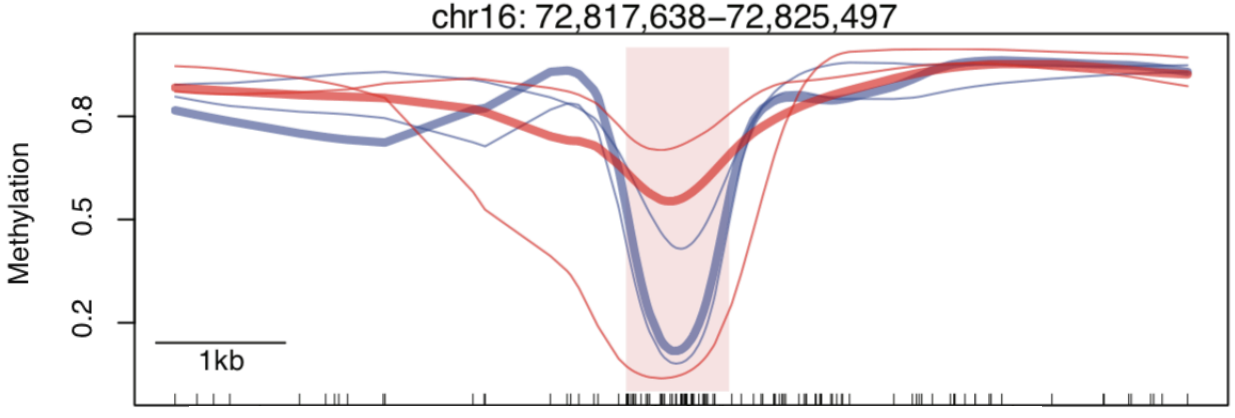
Prioritized by area (sum) statistics

Biological variability



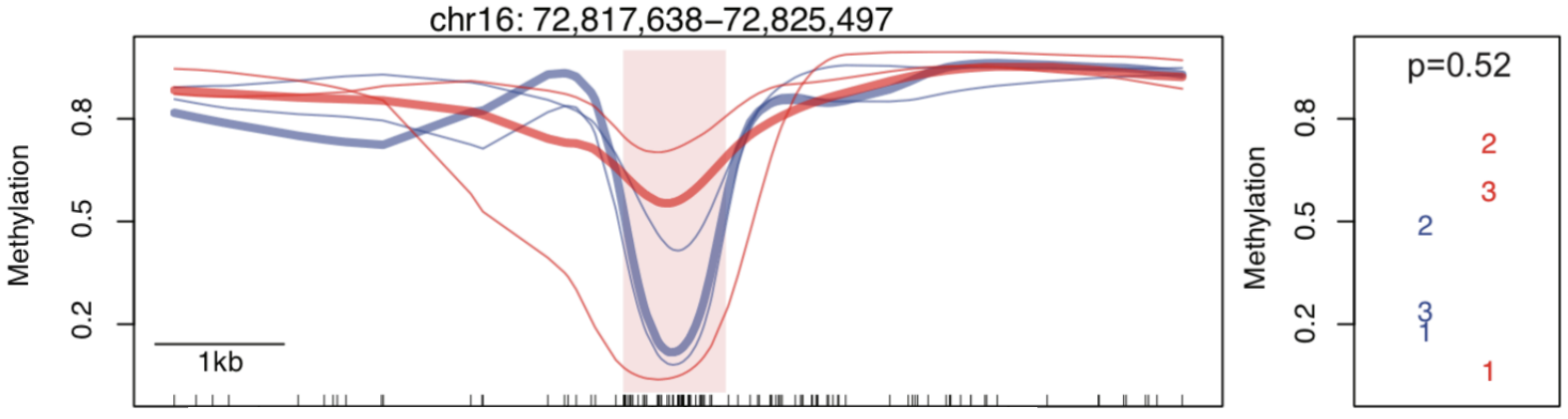
Adapted from Hansen et al., 2011 (*Nature Genetics*)

Biological variability



Adapted from Hansen et al., 2011 (*Nature Genetics*)

Biological variability



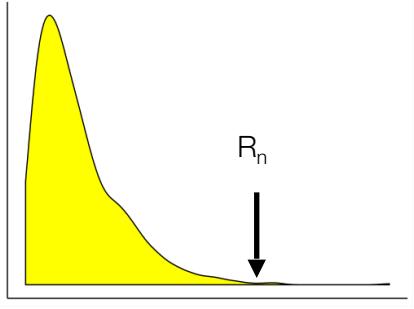
Adapted from Hansen et al., 2011 (*Nature Genetics*)

Methodology

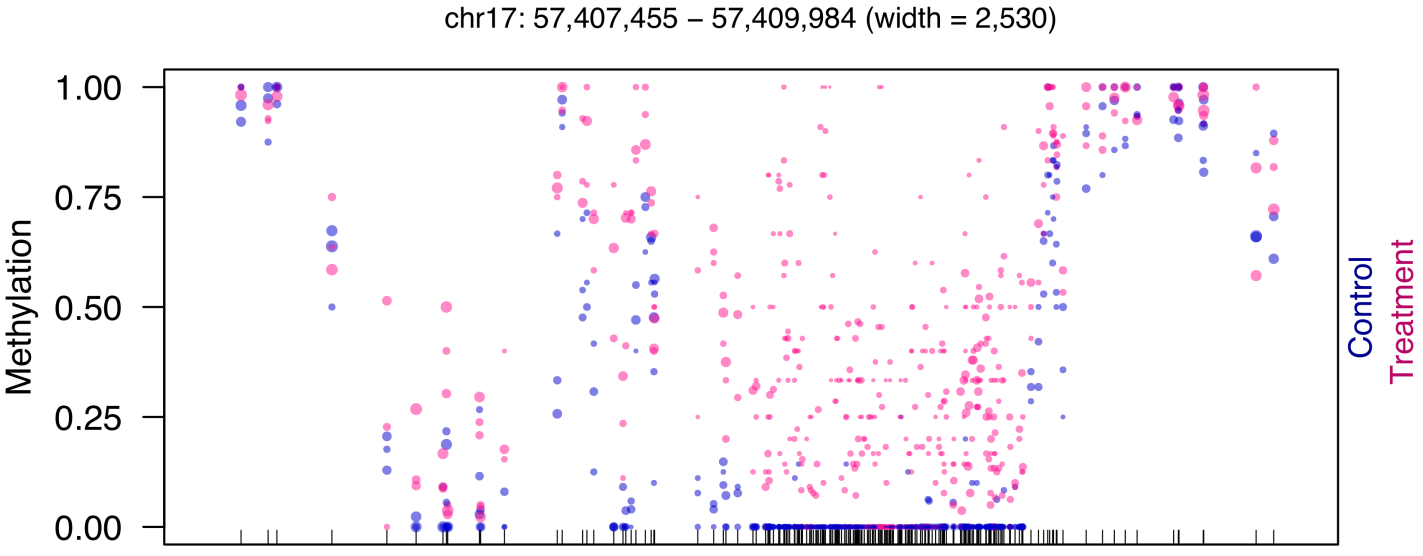
dmrseq: two-stage approach

1. Detect *de novo* candidate regions

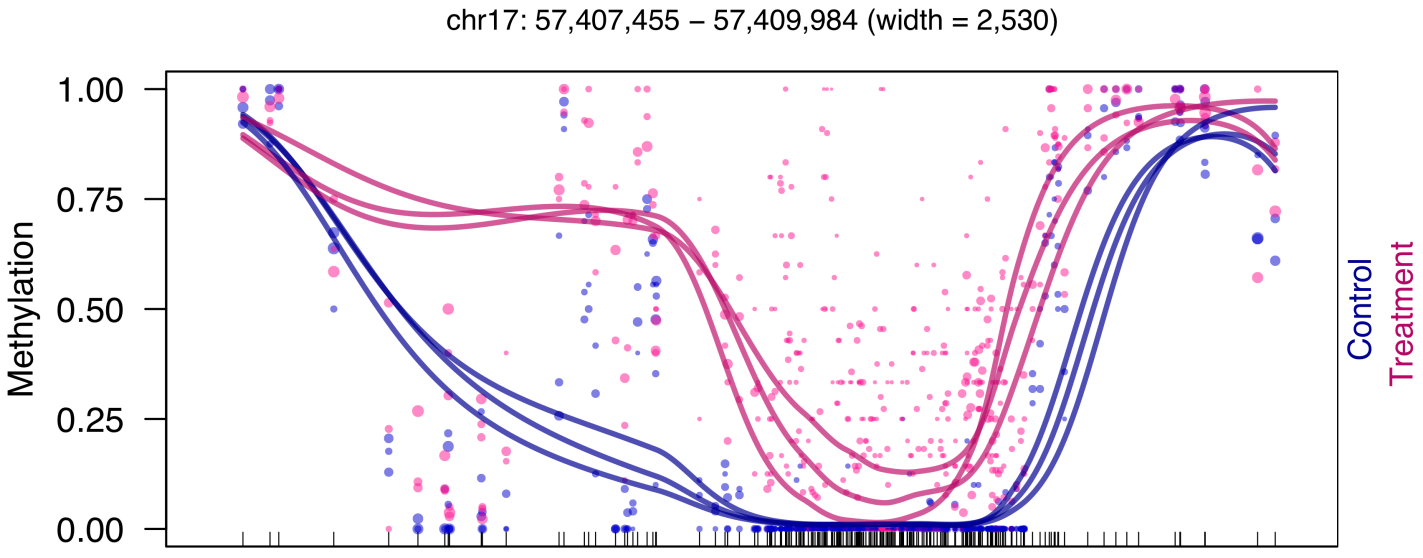
2. Evaluate statistical significance



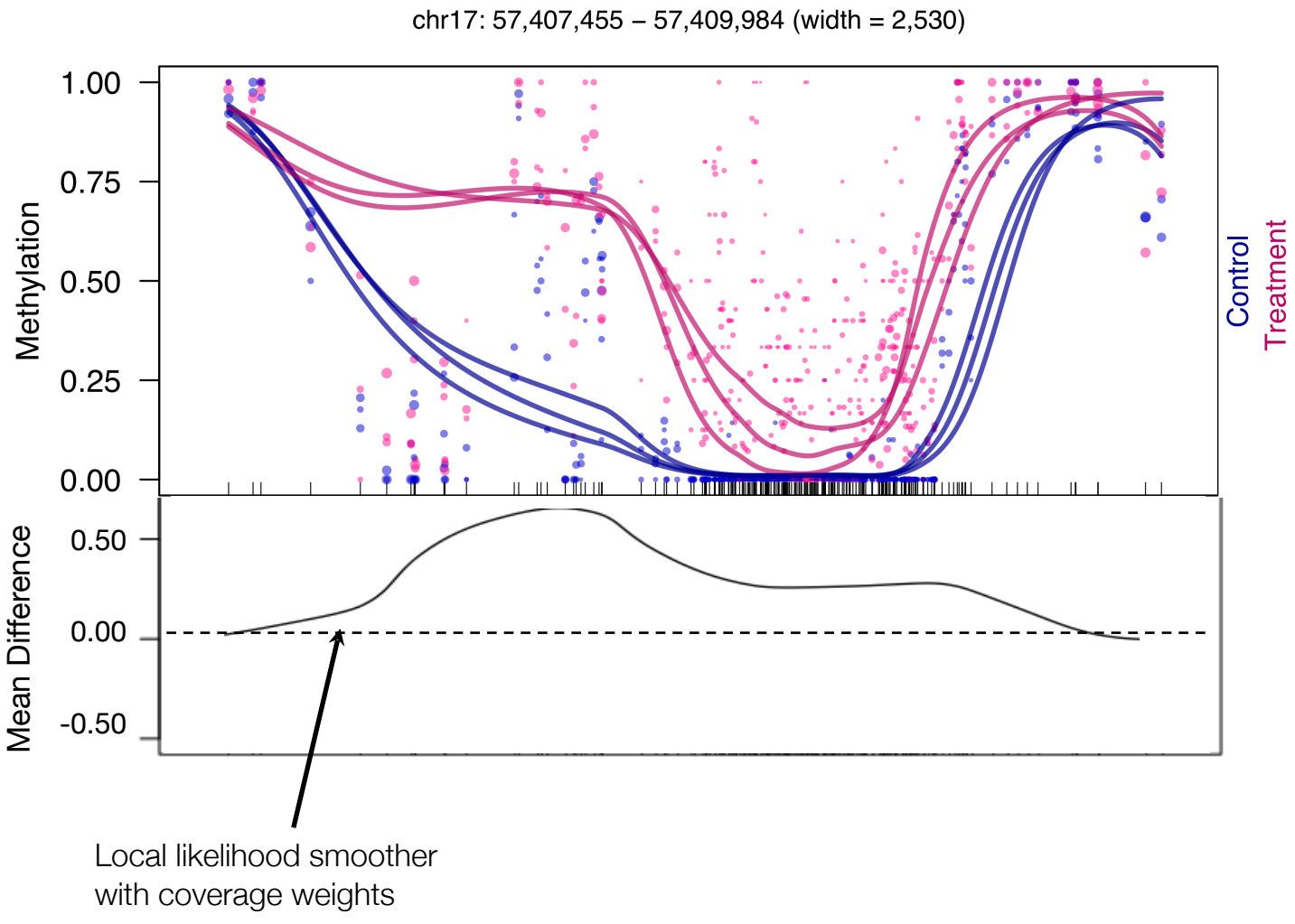
dmrseq: (1) Detect *de novo* candidate regions



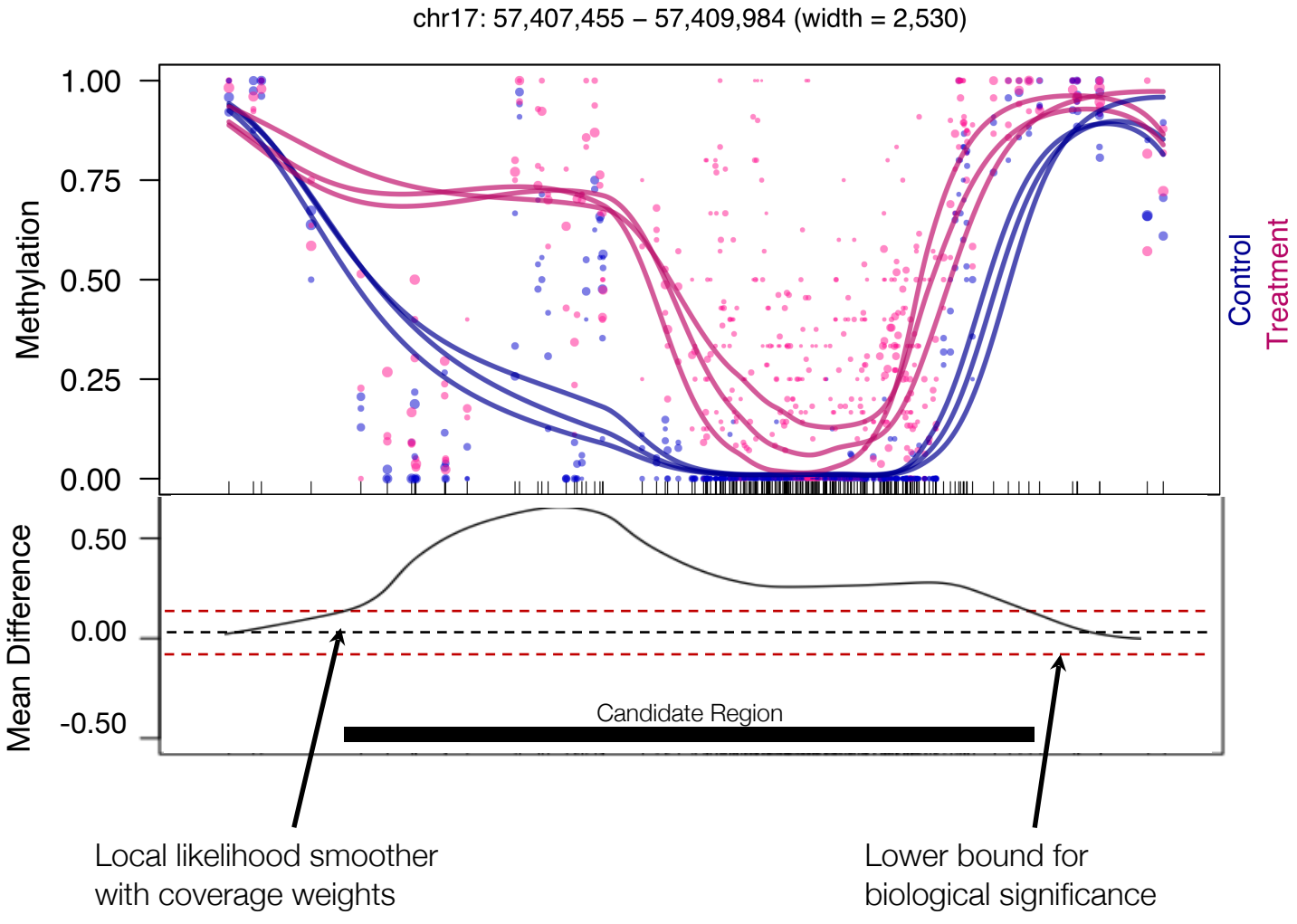
dmrseq: (1) Detect *de novo* candidate regions



dmrseq: (1) Detect *de novo* candidate regions

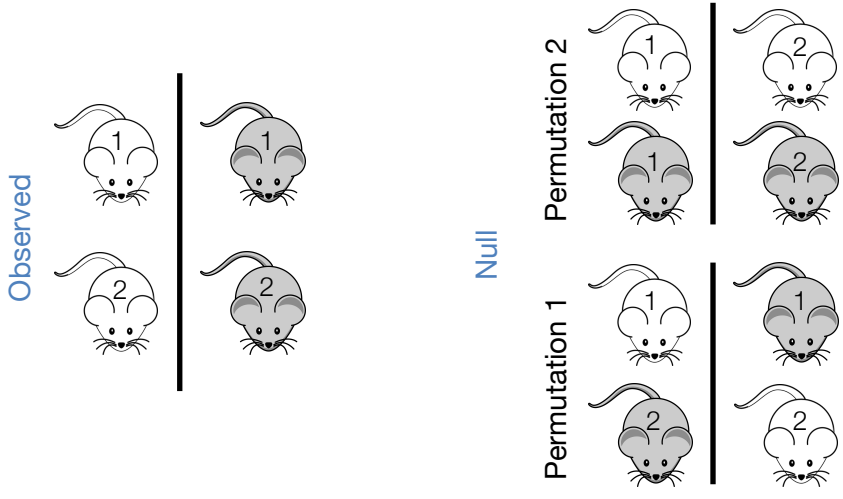


dmrseq: (1) Detect *de novo* candidate regions



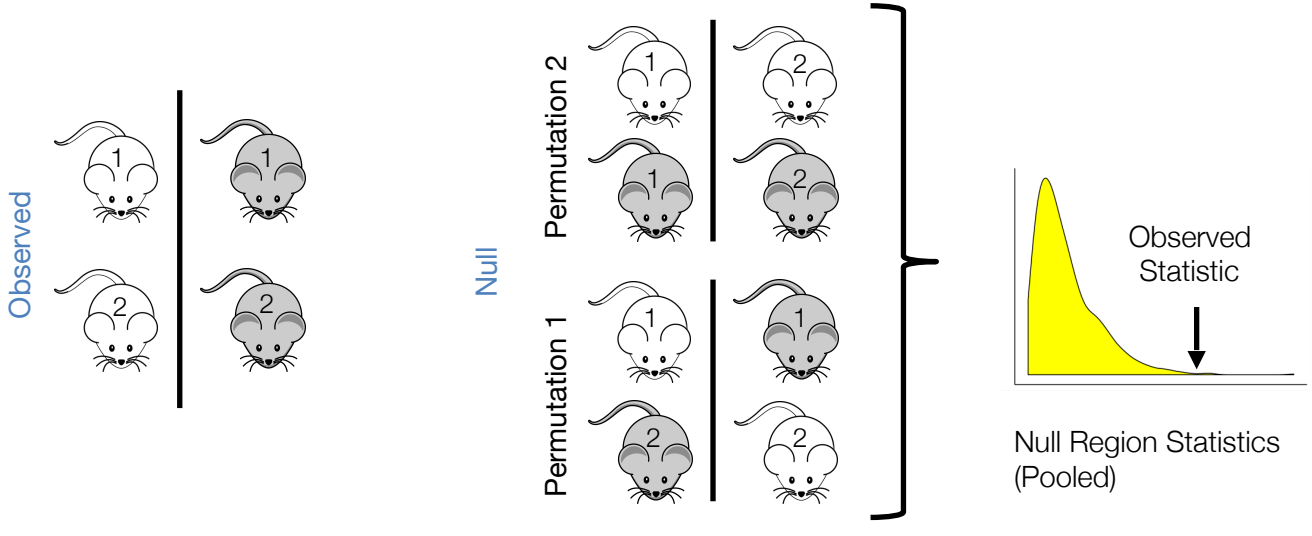
dmrseq: (2) Assess region-level signal

- Formulate region-level summary statistic
- Compare region statistics against null permutation distribution to evaluate significance

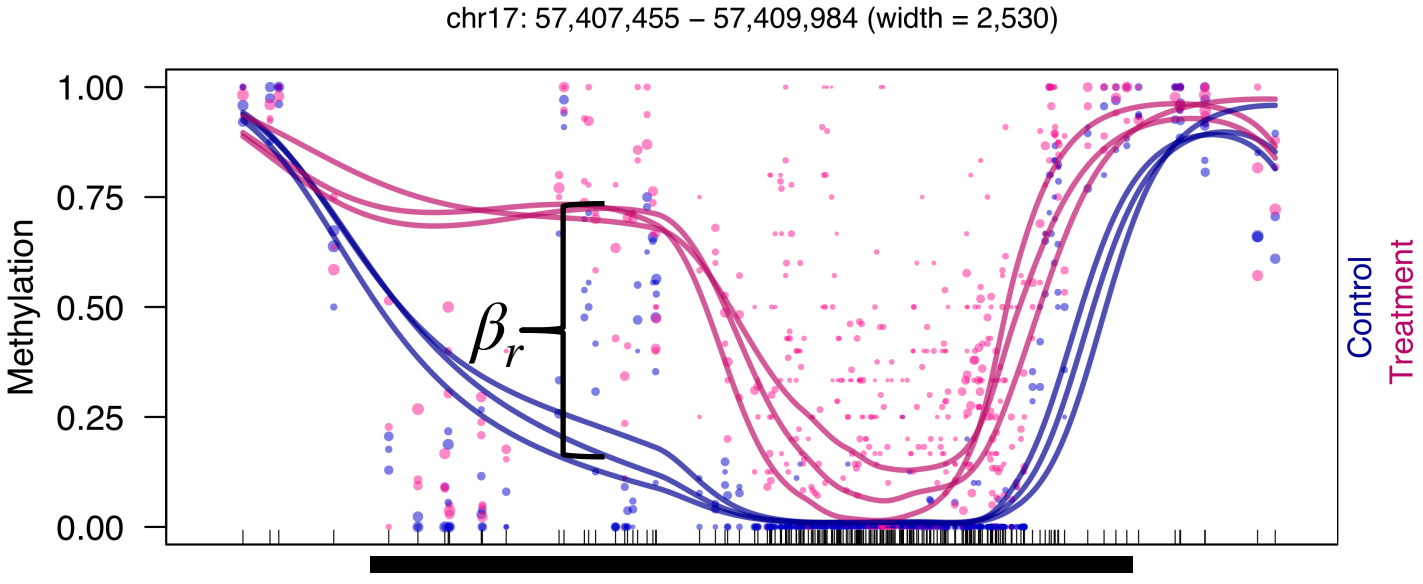


dmrseq: (2) Assess region-level signal

- Formulate region-level summary statistic
- Compare region statistics against null permutation distribution to evaluate significance



Region-level summary statistic



Captures signal across the entire region, accounting for:

- spatial correlation
- variability within biological condition
- coverage

Evaluation

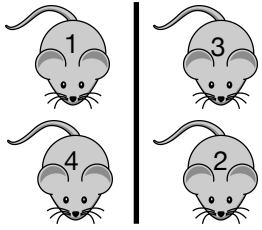
Simulation to assess FDR and power

Control samples split into two artificial conditions

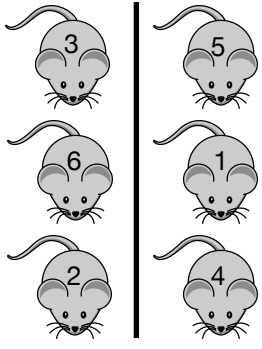
+

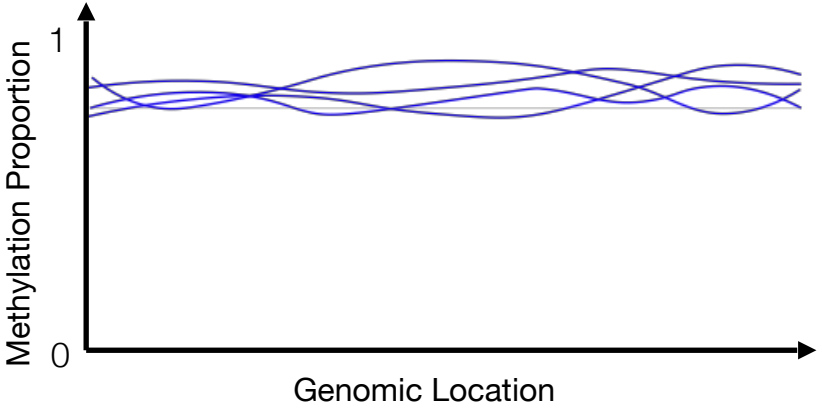
In silico DMRs added at random locations

2 vs 2 Comparison



3 vs 3 Comparison





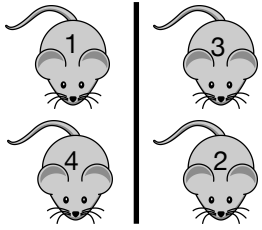
Simulation to assess FDR and power

Control samples split into two artificial conditions

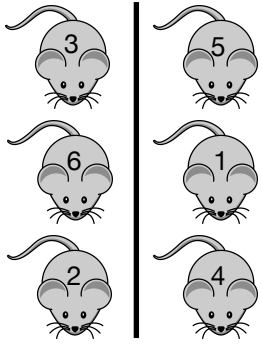
+

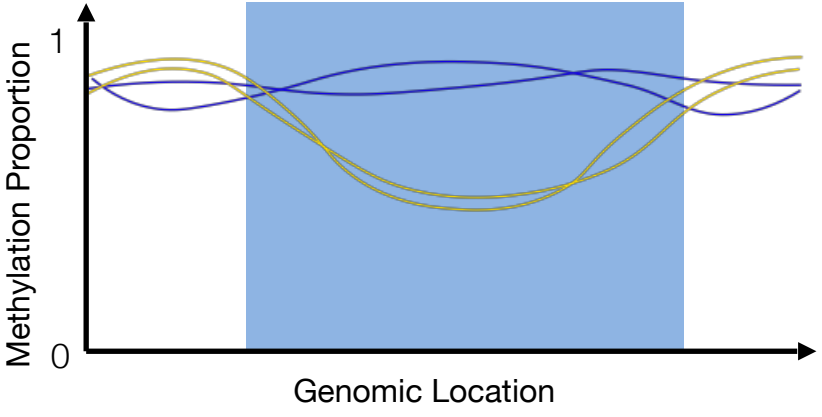
In silico DMRs added at random locations

2 vs 2 Comparison

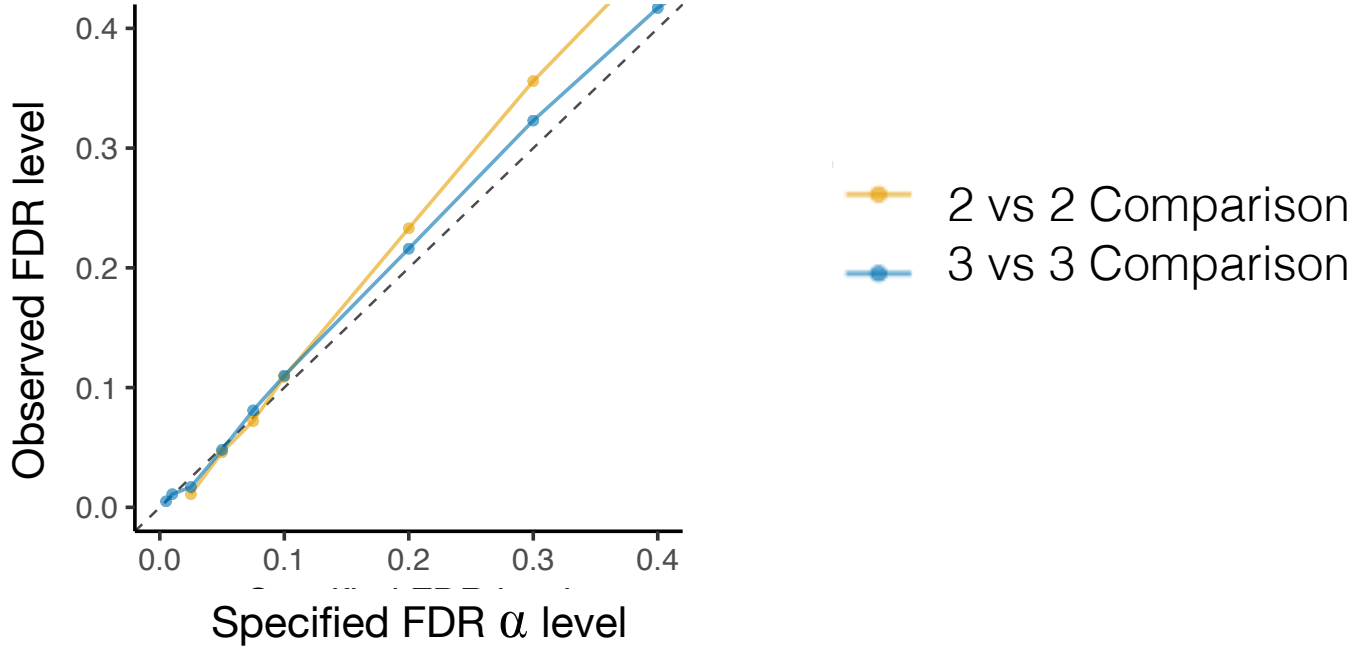


3 vs 3 Comparison



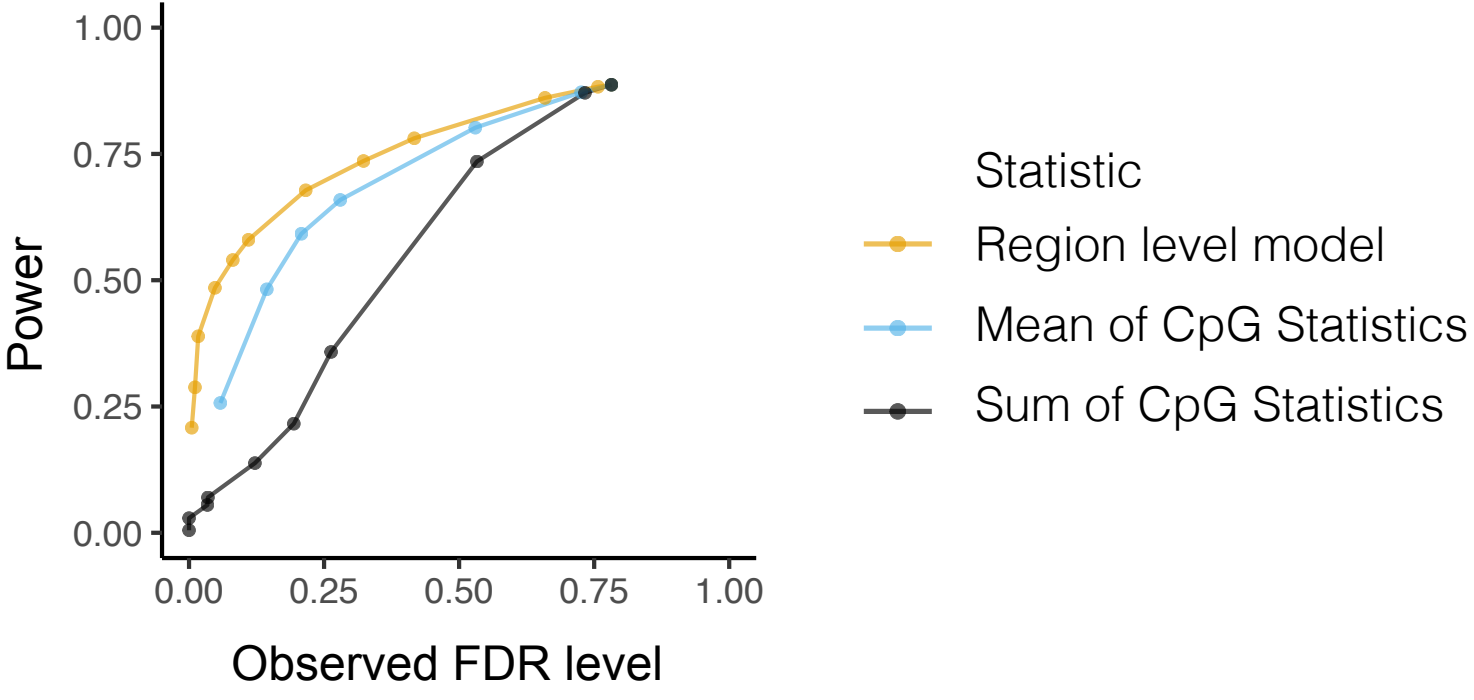


Accurate FDR control in simulation

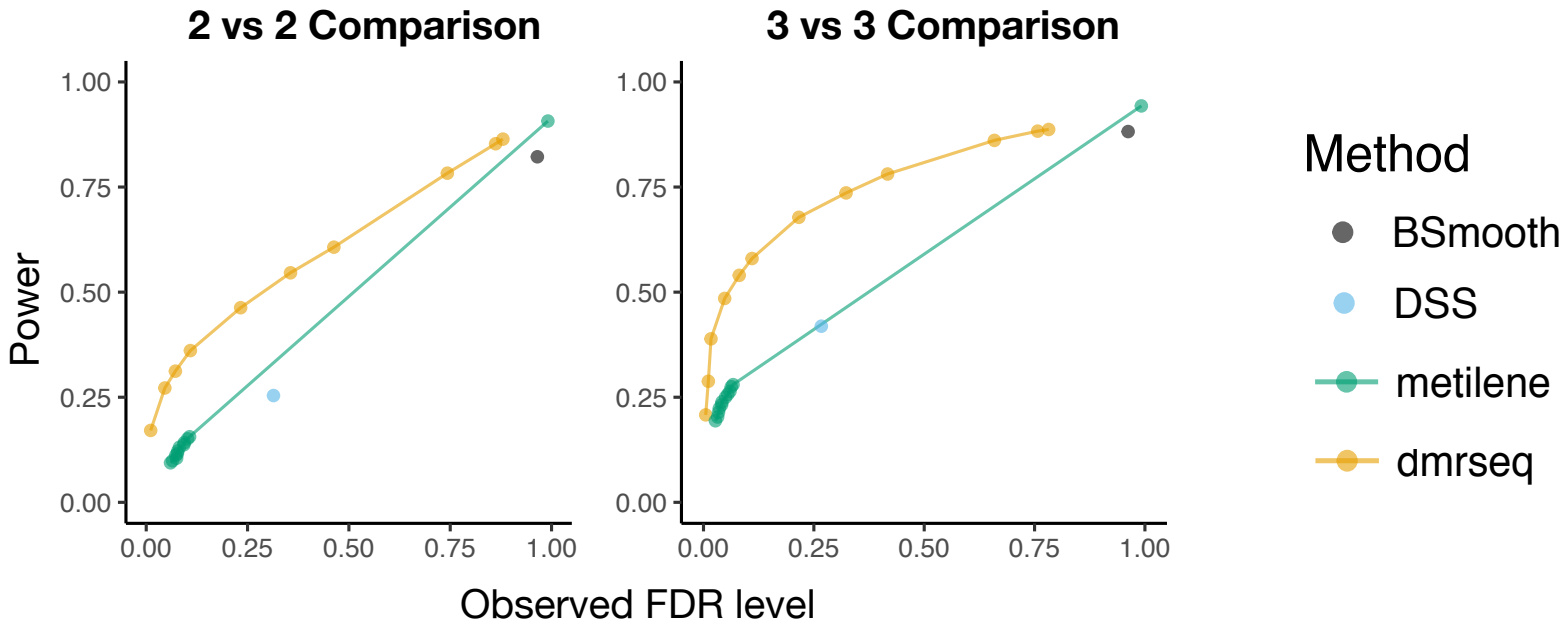


Korthauer et al., 2018 (*Biostatistics*)

Region-level modeling improves power to detect DMRs

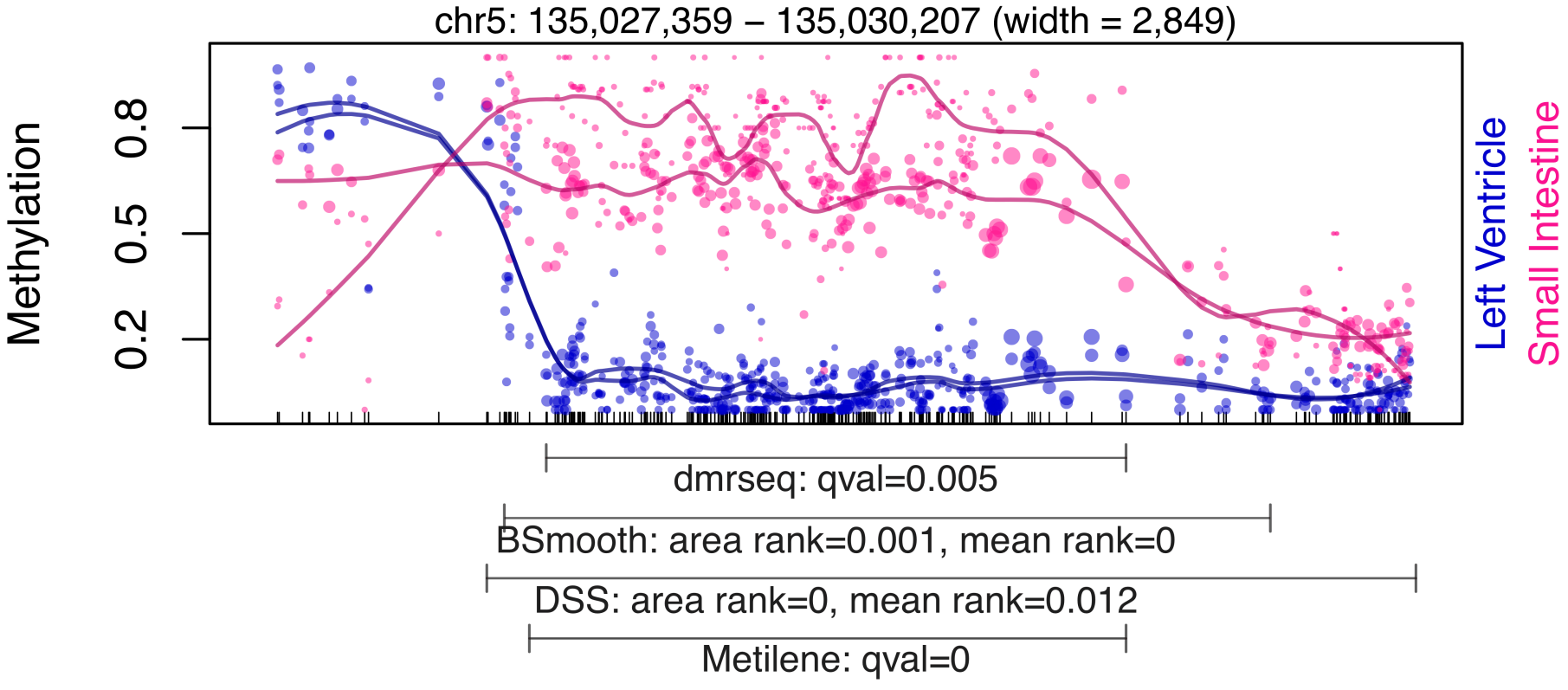


High sensitivity and specificity in simulation



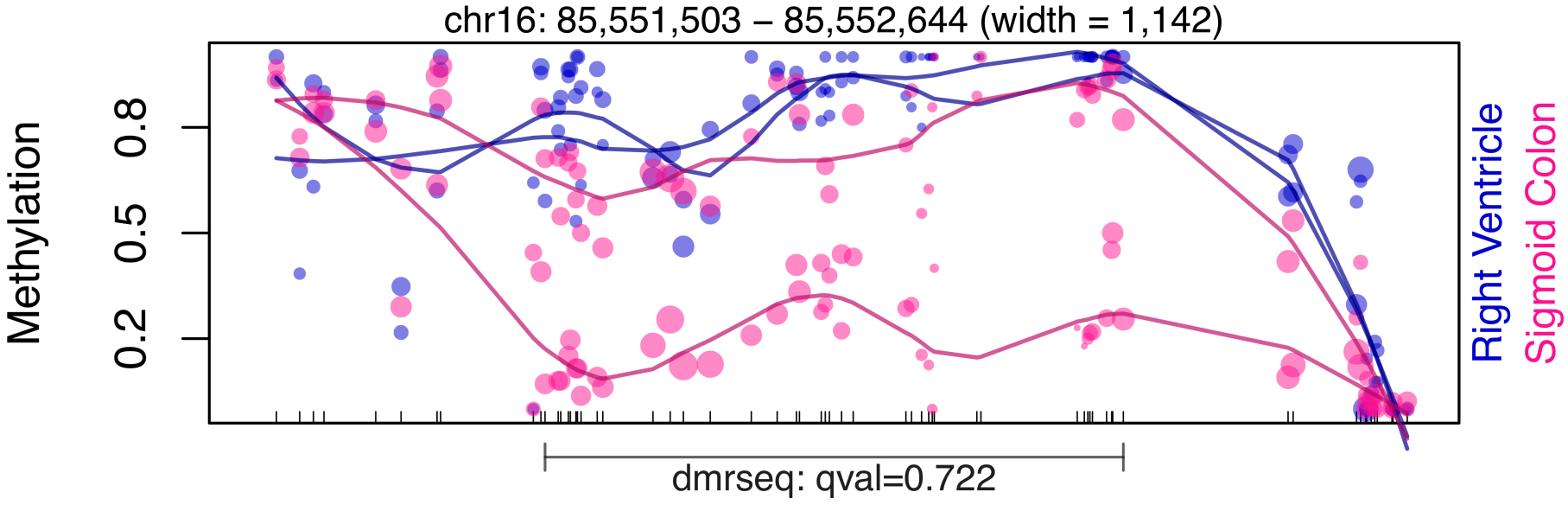
Korthauer et al., 2018 (*Biostatistics*)

Example: highly ranked DMR across all methods



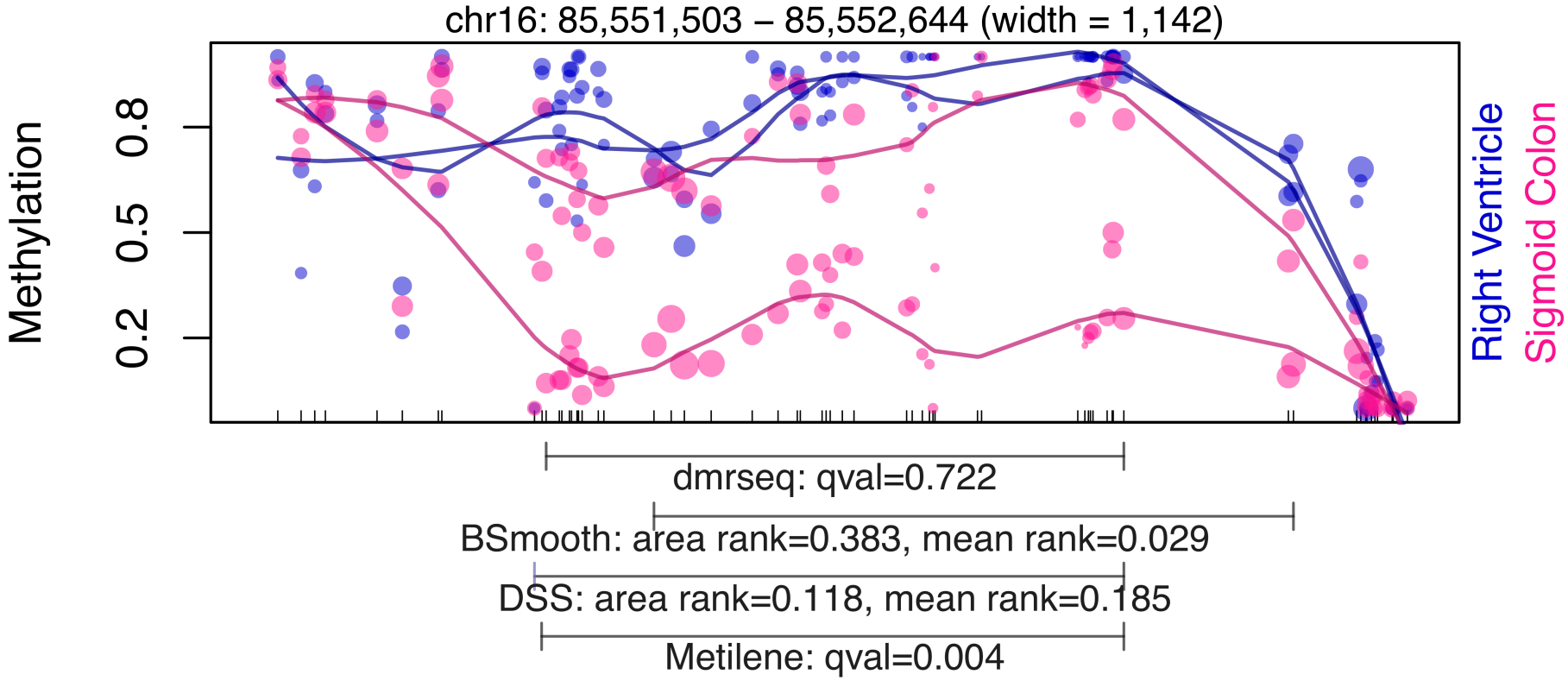
Korthauer et al., 2018 (*Biostatistics*)

Example: dmrseq accounts for sample variability



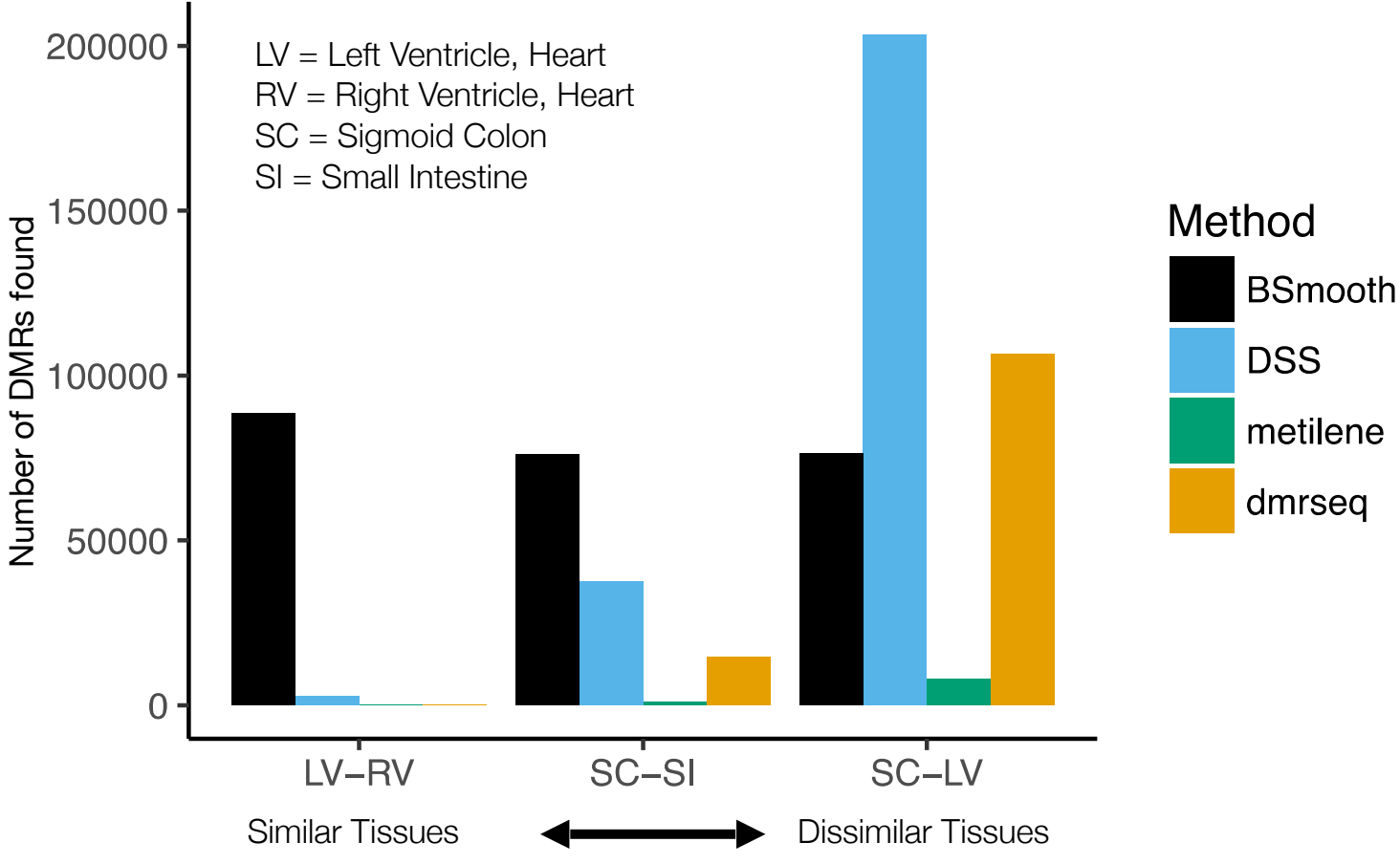
Korthauer et al., 2018 (*Biostatistics*)

Example: dmrseq accounts for sample variability

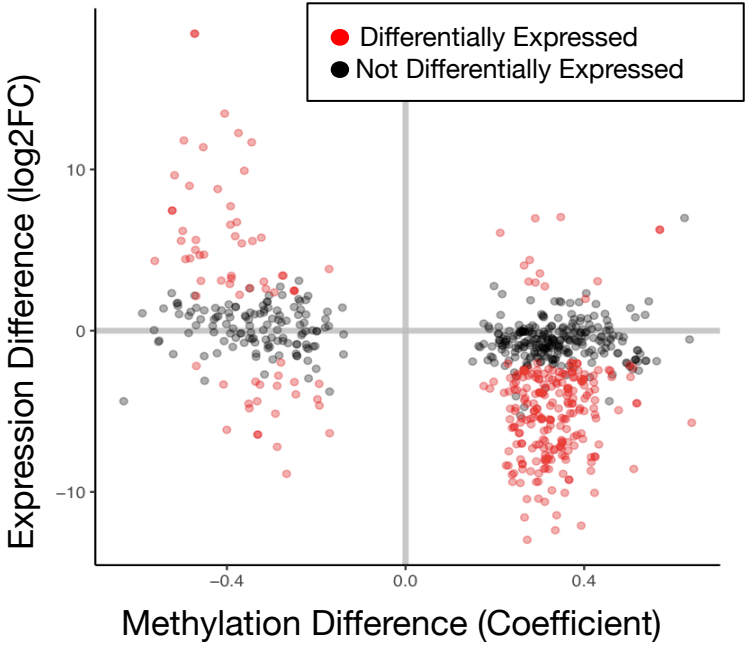


Korthauer et al., 2018 (*Biostatistics*)

Roadmap case study: Tissue-specific DMRs

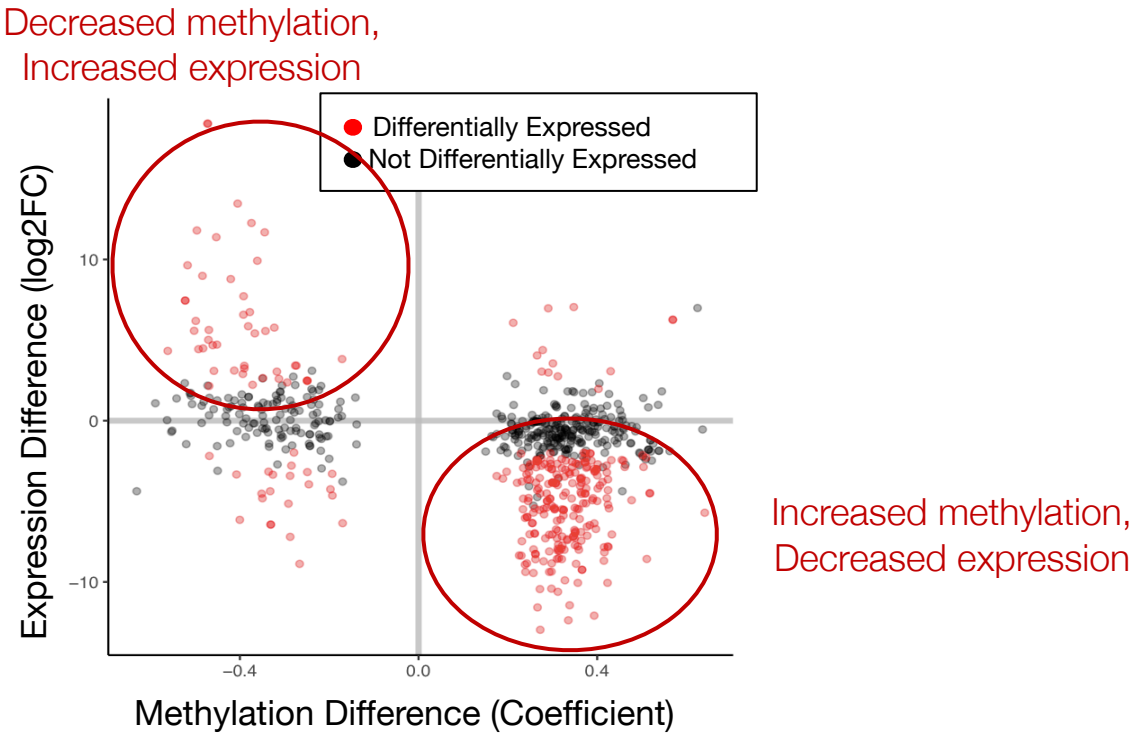


Validation of DMRs in promoter regions



Korthauer et al., 2018 (*Biostatistics*)

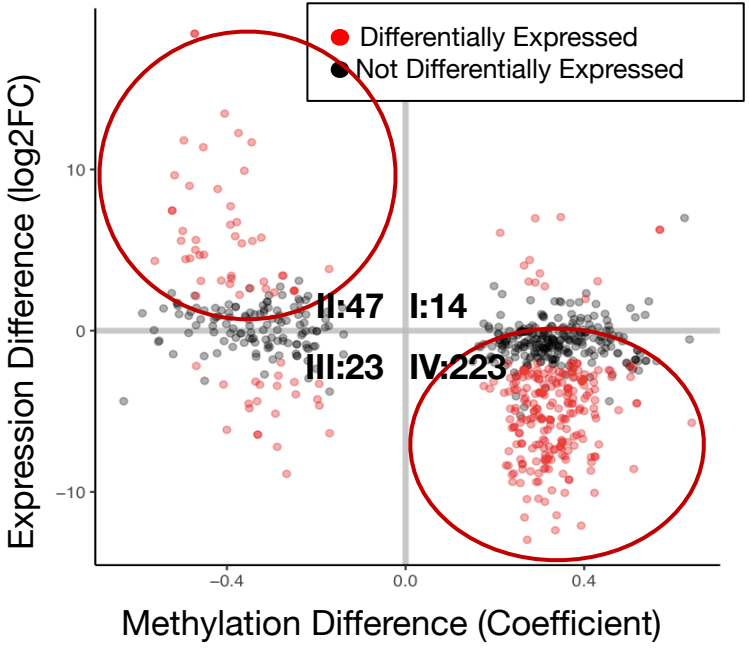
Validation of DMRs in promoter regions



Korthauer et al., 2018 (*Biostatistics*)

Validation of DMRs in promoter regions

Decreased methylation,
Increased expression



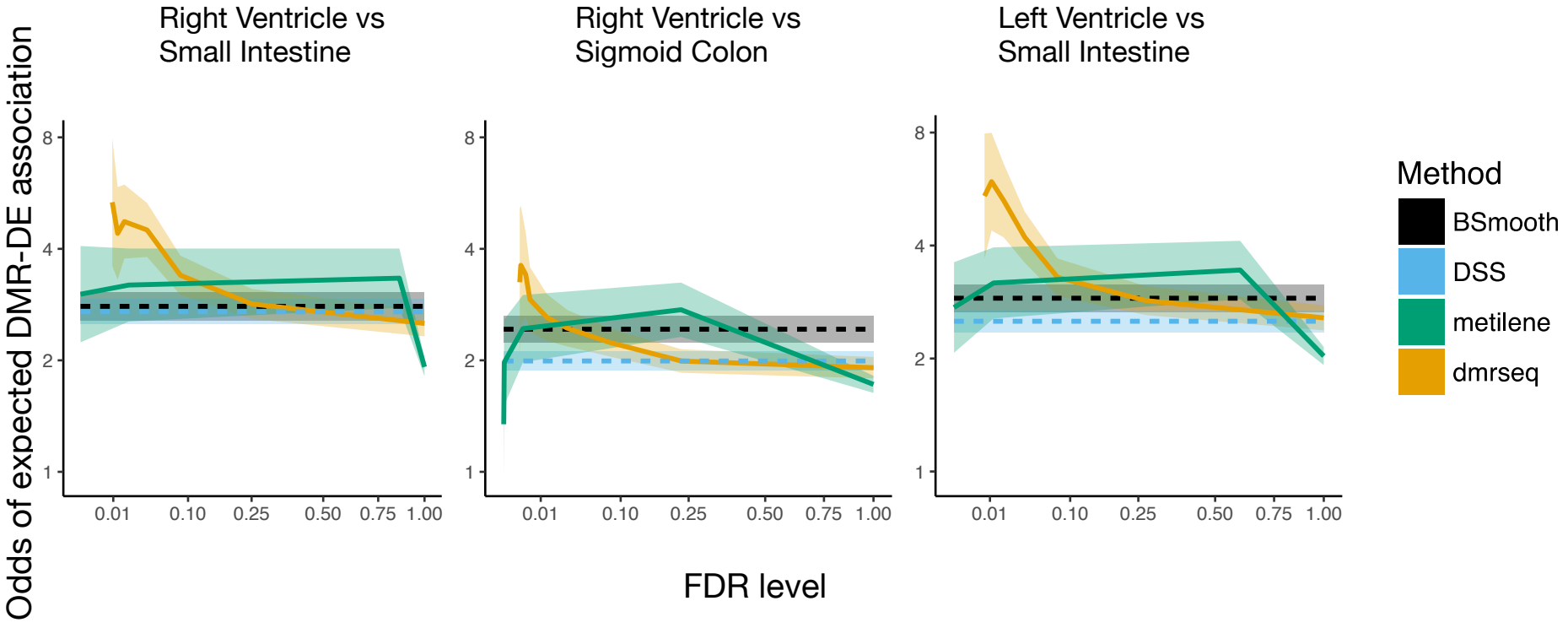
Increased methylation,
Decreased expression

Odds Statistic:

$$\frac{\text{Expected direction}}{\text{Unexpected Direction}} = \frac{\text{II} + \text{IV}}{\text{I} + \text{III}} = \frac{47 + 223}{14 + 23} = 7.30$$

Korthauer et al., 2018 (*Biostatistics*)

Validation of DMRs in promoter regions

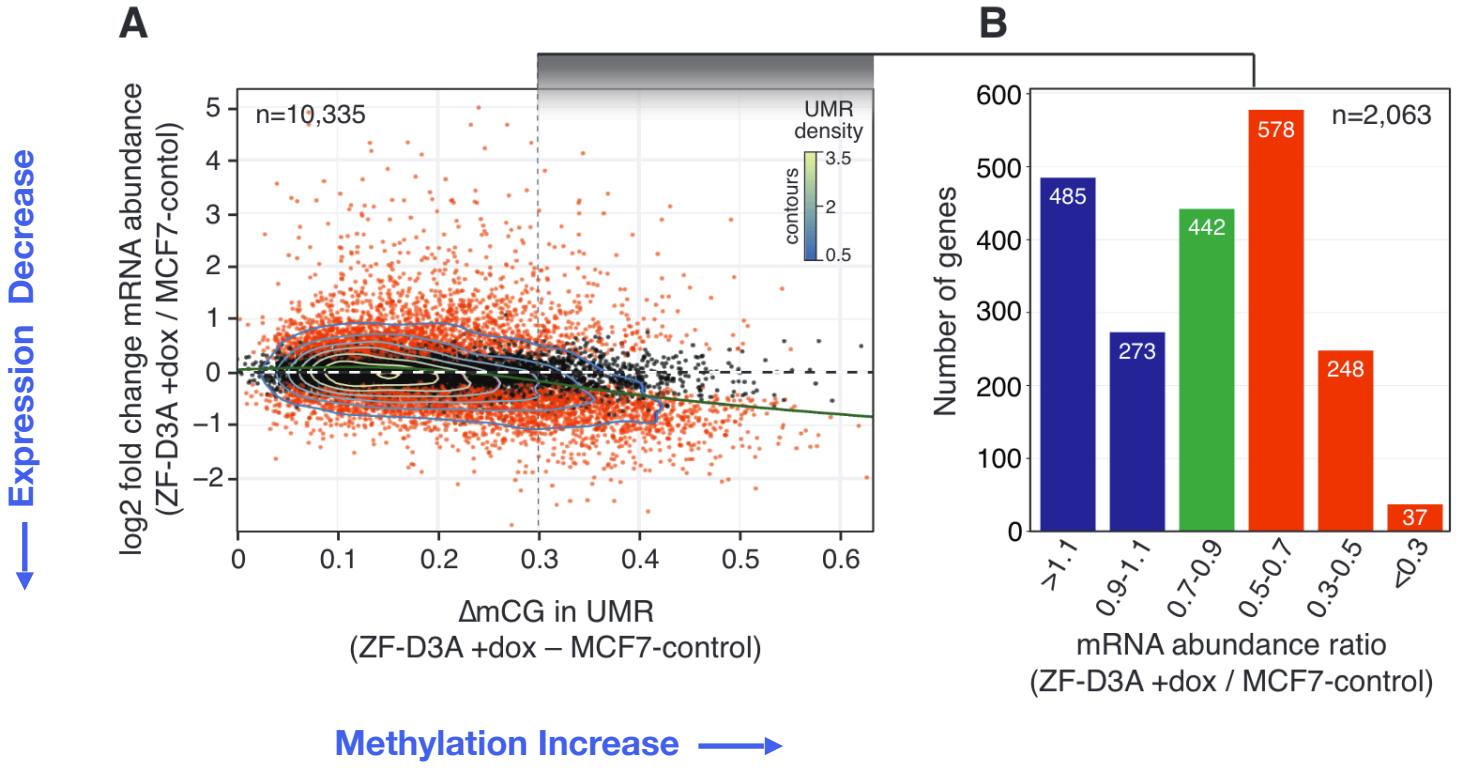


Korthauer et al., 2018 (*Biostatistics*)

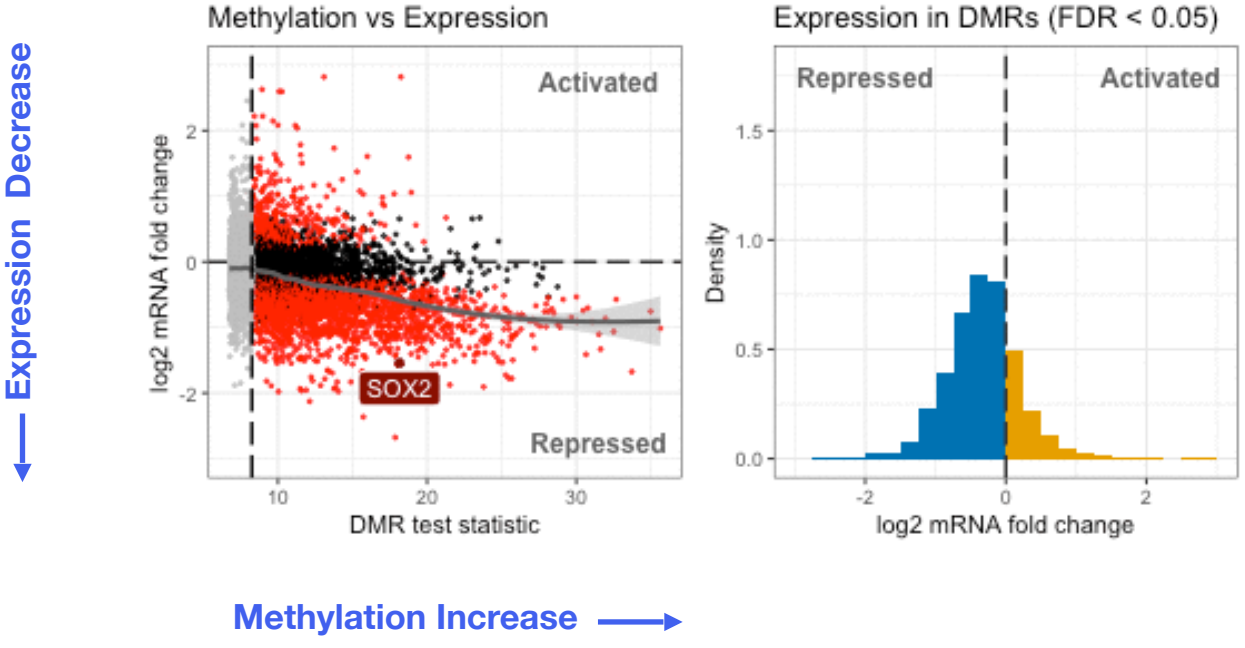
Biological insights

Landmark study finds methylation not generally sufficient to repress gene expression

Figure 5 from Ford et al., 2017 (*bioRxiv*)

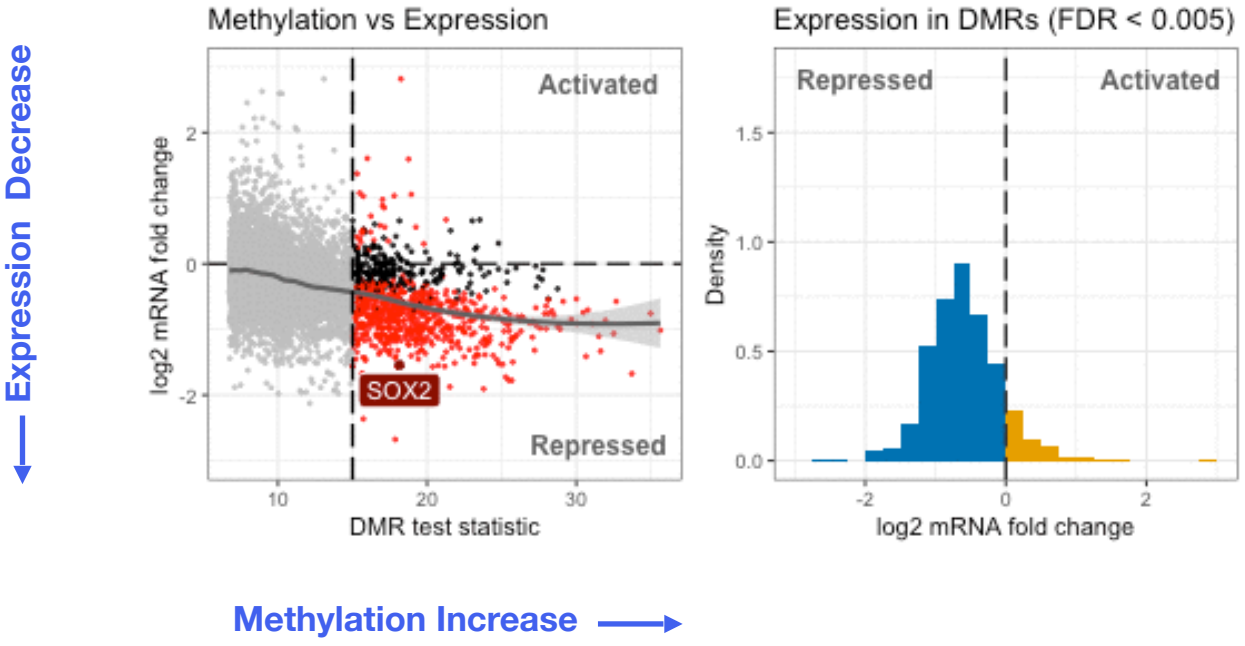


Methylation of promoters overwhelmingly represses gene expression



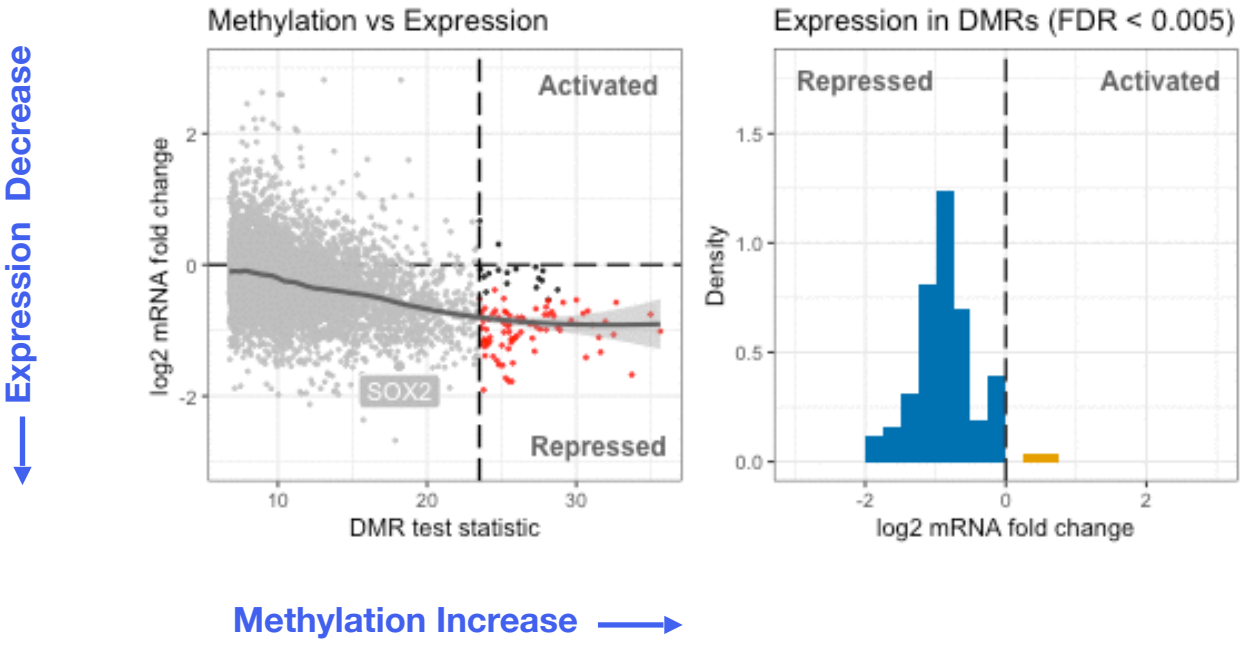
Korthauer & Irizarry, 2018 (*bioRxiv*)

Methylation of promoters overwhelmingly represses gene expression



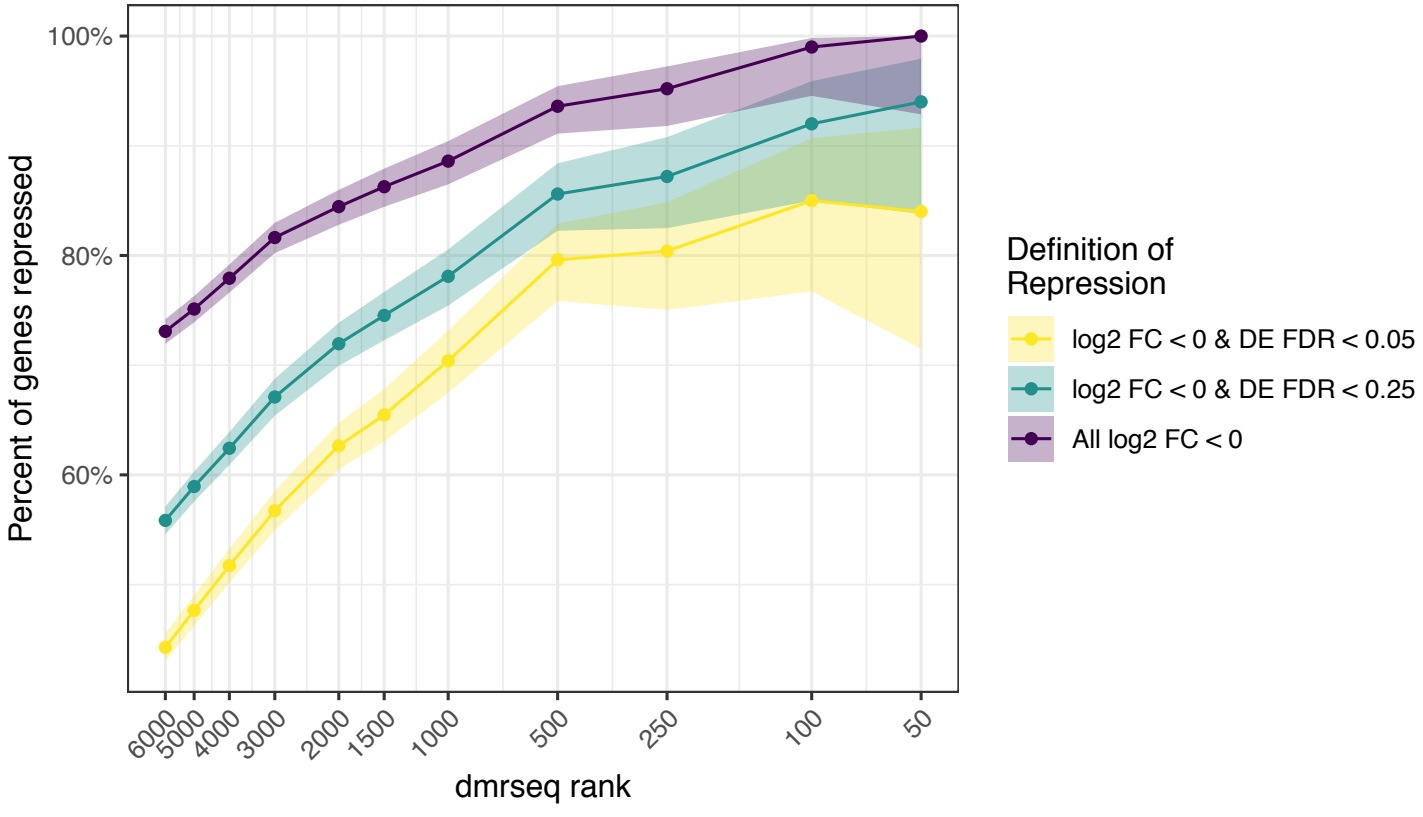
Korthauer & Irizarry, 2018 (*bioRxiv*)

Methylation of promoters overwhelmingly represses gene expression



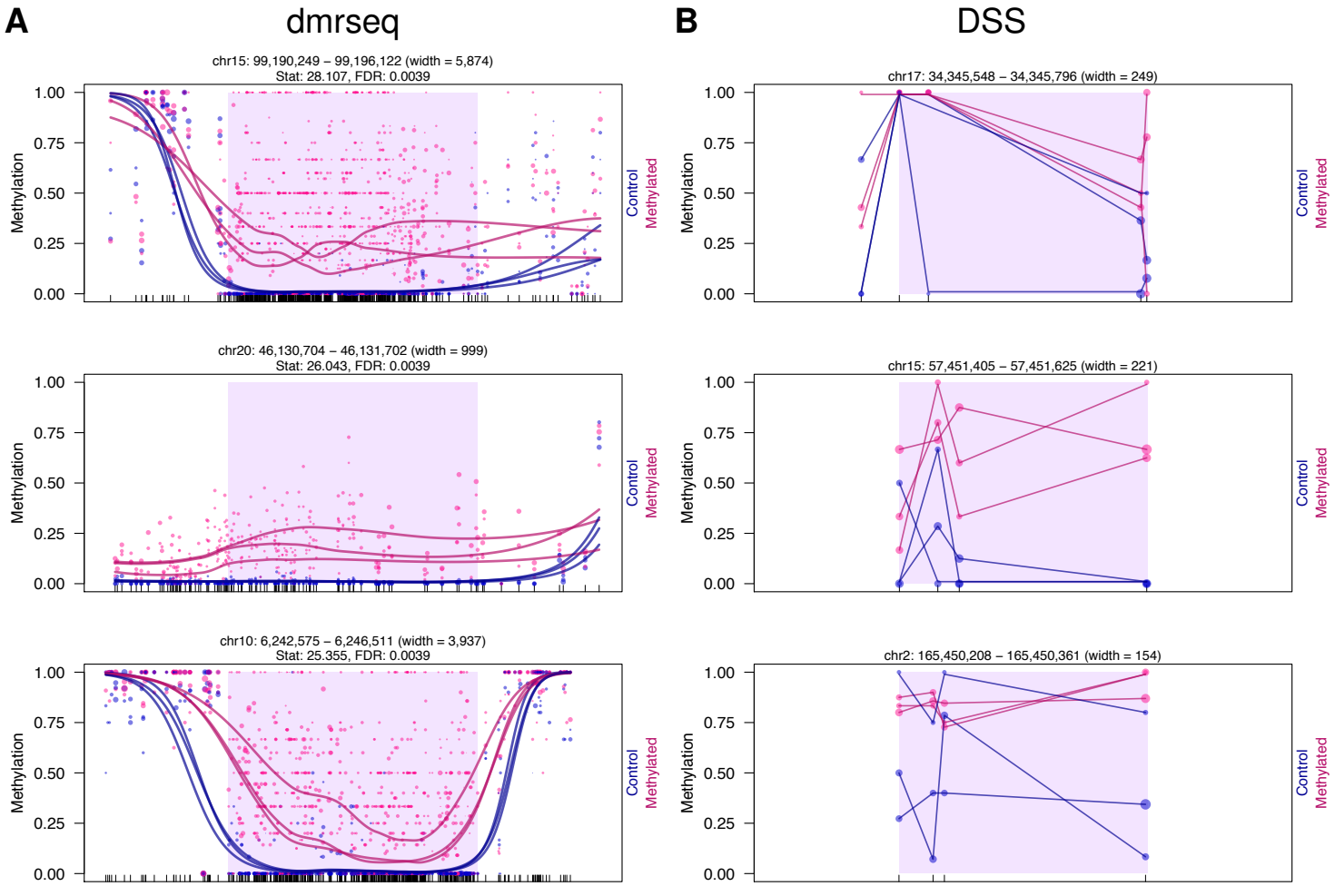
Korthauer & Irizarry, 2018 (*bioRxiv*)

Enrichment increases with significance level



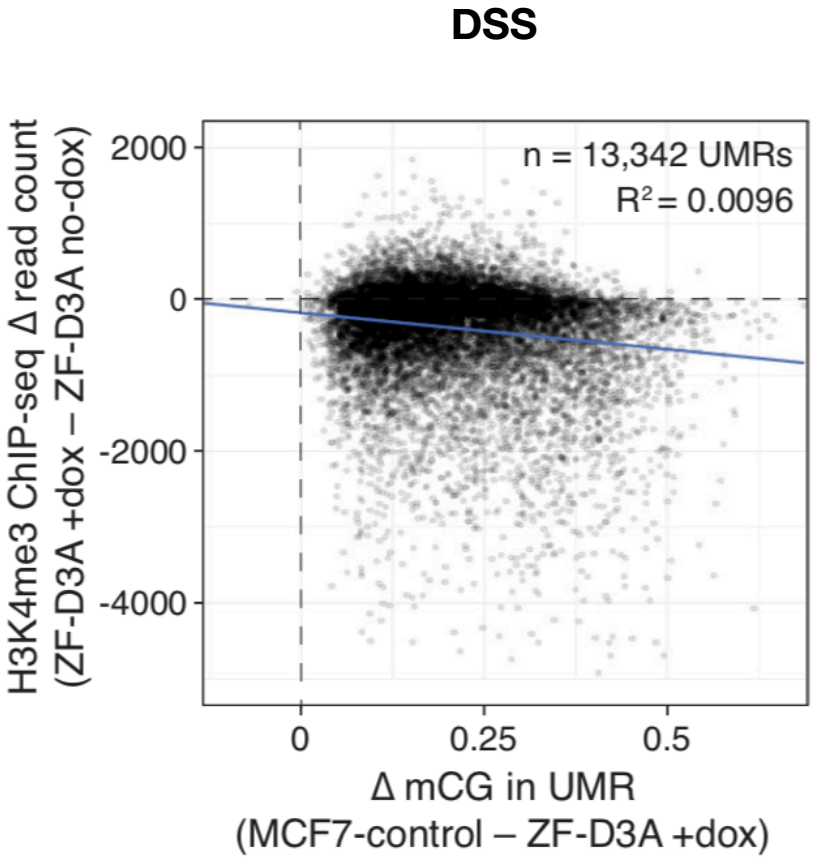
Korthauer & Irizarry, 2018 (*bioRxiv*)

Top-ranked regions found exclusively by each method



Korthauer & Irizarry, 2018 (*bioRxiv*)

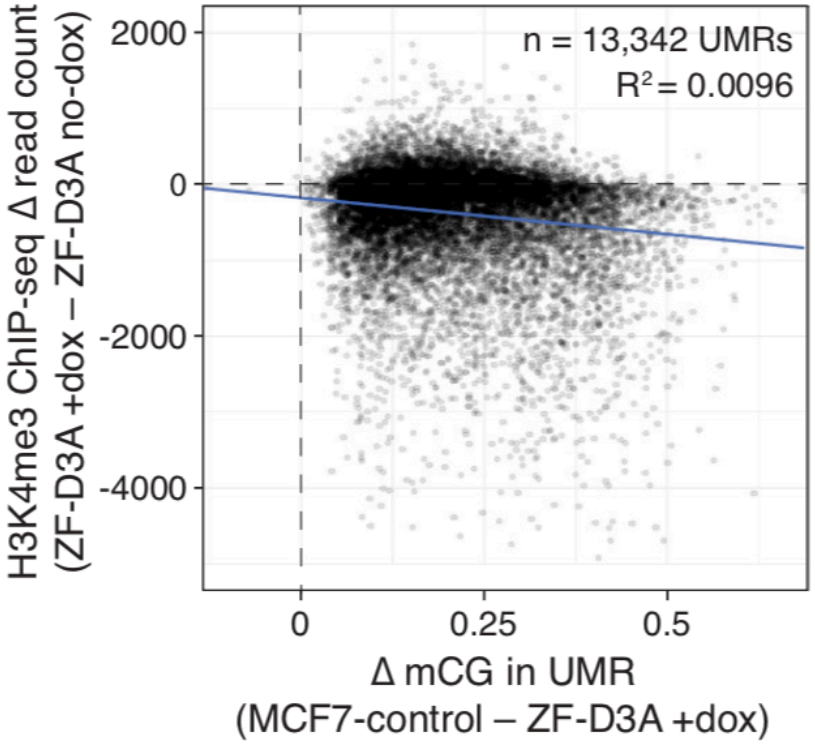
dmrseq shows DNA methylation reduces H3K4 trimethylation



Ford et al., 2017 (*bioRxiv*)

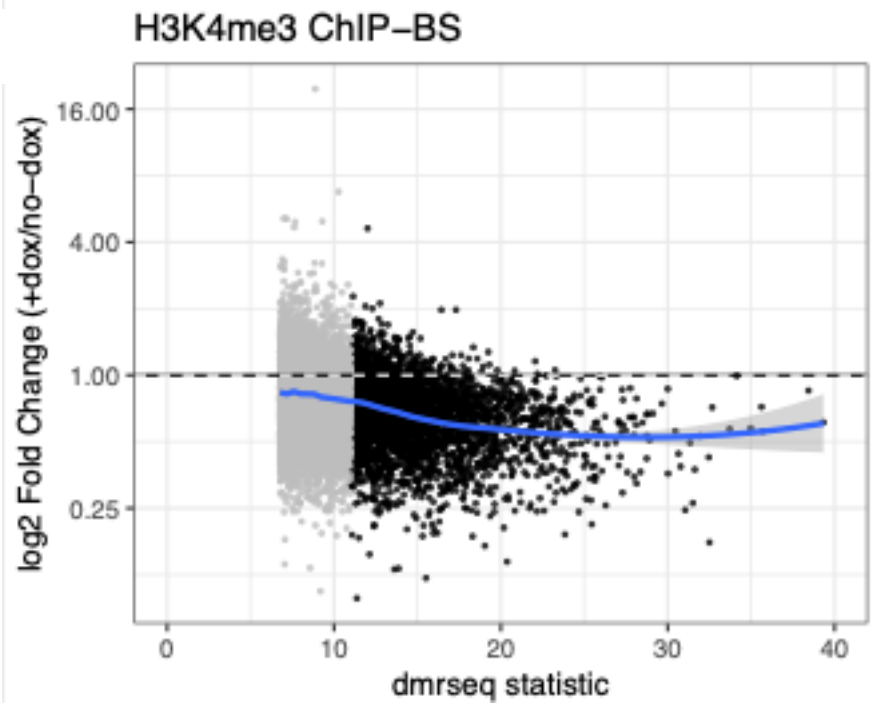
dmrseq shows DNA methylation reduces H3K4 trimethylation

DSS



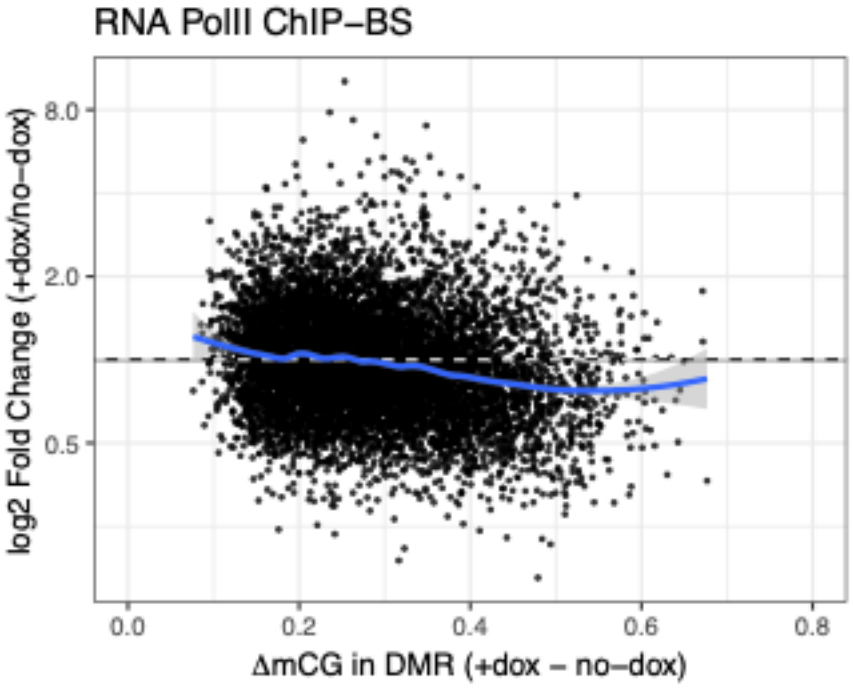
Ford et al., 2017 (*bioRxiv*)

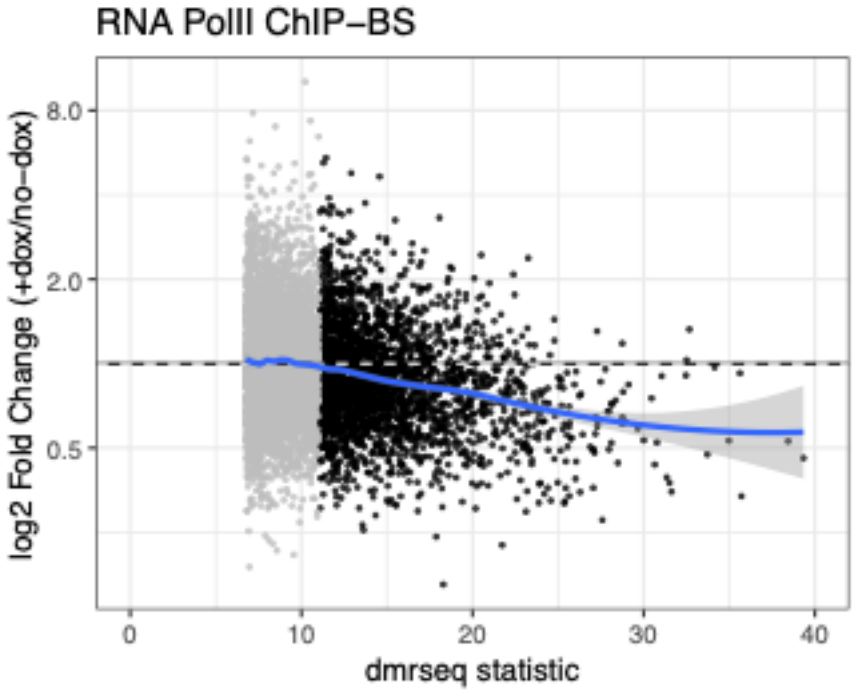
dmrseq



Korthauer & Irizarry, 2018 (*bioRxiv*)

dmrseq shows DNA methylation reduces RNA Pol II activity





Korthauer & Irizarry, 2018 (*bioRxiv*)

dmrseq R package

dmrseq

platforms all | rank 1031 / 1649 | posts 3 / 1 / 7 / 2 | in Bioc 1 year
build ok | updated < 3 months

DOI: 10.18129/B9.bioc.dmrseq

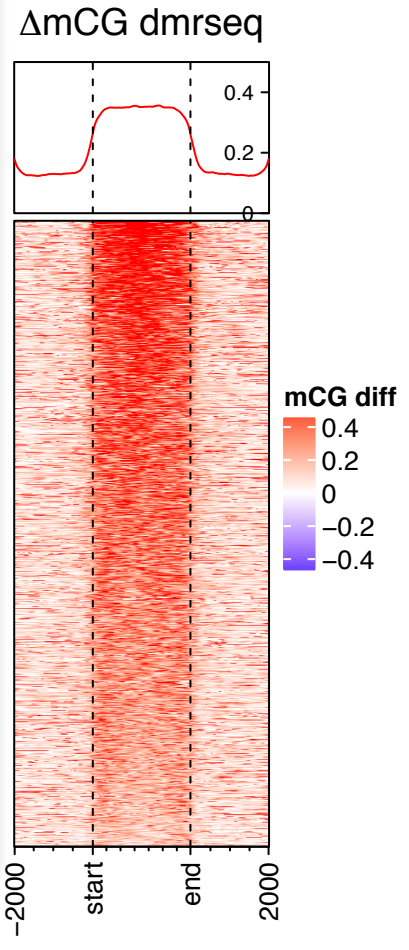
Detection and inference of differentially methylated regions from Whole Genome Bisulfite Sequencing

Bioconductor version: Release (3.8)

This package implements an approach for scanning the genome to detect and perform accurate inference on differentially methylated regions from Whole Genome Bisulfite Sequencing data. The method is based on comparing detected regions to a pooled null distribution, that can be implemented even when as few as two samples per population are available. Region-level statistics are obtained by fitting a generalized least squares (GLS) regression model with a nested autoregressive correlated error structure for the effect of interest on transformed methylation proportions.

Author: Keegan Korthauer <keegan at jimmy.harvard.edu>, Sutirtha Chakraborty <statistuta at gmail.com>, Yuval Benjamini <yuvalbenj at gmail.com>, Rafael Irizarry <rafa at jimmy.harvard.edu>

Maintainer: Keegan Korthauer <keegan at jimmy.harvard.edu>



dmrseq R package

dmrseq

platforms all rank 103

build ok updated

DOI: 10.18129/B9.bioc.dm

Detection and infer
Genome Bisulfite S

Bioconductor version: Rele

This package implements a
on differentially methylatec
on comparing detected reg
as two samples per popula
least squares (GLS) regres
effect of interest on transfo

Author: Keegan Korthauer
gmail.com>, Yuval Benjam

Maintainer: Keegan Kortha

5 Exploring and exporting results

5.1 Explore how many regions were significant

How many regions were significant at the FDR (q-value) cutoff of 0.05? We can find this by counting how many values in the `qval` column of the results data.frame were less than 0.05. You can also subset the regions by an FDR cutoff.

```
sum(regions$qval < 0.05)
```

```
## [1] 144
```

```
# select just the regions below FDR 0.05 and place in a new data.frame
sigRegions <- regions[regions$qval < 0.05,]
```

5.2 Hypo- or Hyper- methylation?

You can determine the proportion of regions with hyper-methylation by counting how many had a positive direction of effect (positive statistic).

```
sum(sigRegions$stat > 0) / length(sigRegions)
```

```
## [1] 0.25
```

To interpret the direction of effect, note that for a two-group comparison **dmrseq** uses alphabetical order of the covariate of interest. The condition with a higher alphabetical rank will become the reference category. For example, if the two conditions are "A" and "B", the "A" group will be the reference category, so a positive direction of effect means that "B" is hyper-methylated relative to "A". Conversely, a negative direction of effect means that "B" is hypo-methylated relative to "A".

5.3 Plot DMRs

Summary

- dmrseq **identifies and prioritizes DMRs** from bisulfite sequencing experiments
  - **Models region level methylation differences** in order to account for sample and spatial variability
  - Quantifies uncertainty using permutation in order to achieve **accurate false discovery rate control**
  - **Reveals the expected link between DNA methylation and gene expression** in the reanalysis of a landmark study
- R package implementation:

- Reproducible analyses from Korthauer et al. (2018, *Biostatistics*) and Korthauer & Irizarry (2018, *bioRxiv*):

DANA-FARBER CANCER INSTITUTE

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

Acknowledgements

Harvard Biostatistics & DFCI Data Sciences

Rafael Irizarry
Claire Duvallet
Stephanie Hicks
Patrick Kimes
Yered Pita-Juarez
Alejandro Reyes
Chinmay Shukla
Mingxiang Teng

Collaborators

Sutirtha Chakraborty
Yuval Benjamini

Data

Ryan Lister
Ethan Ford

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

bioRχiv
THE PREPRINT SERVER FOR BIOLOGY

ENCODE

dmrseq

keegan@jimmy.harvard.edu
@keegankorthauer
kkorthauer.org