Detection and inference of differentially methylated regions from bisulfite sequencing

Keegan Korthauer

@keegankorthauer

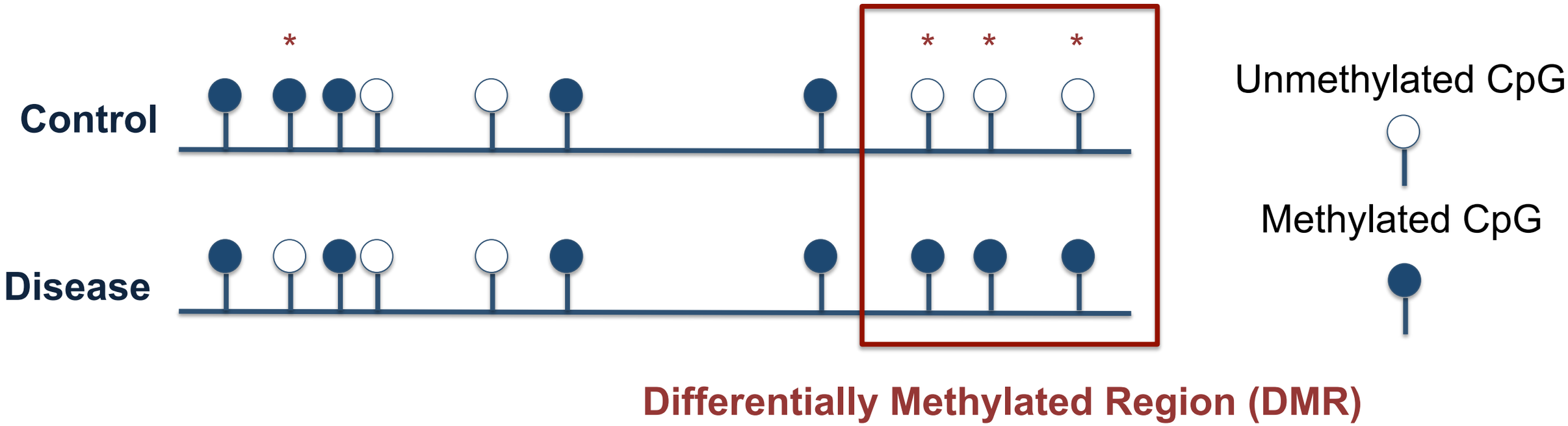
Bioconductor Meeting 2017

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

Differential methylation

Differential methylation commonly studied in

- Cancer
- Developmental stages / Aging
- Tissue types



Whole Genome Bisulfite Sequencing

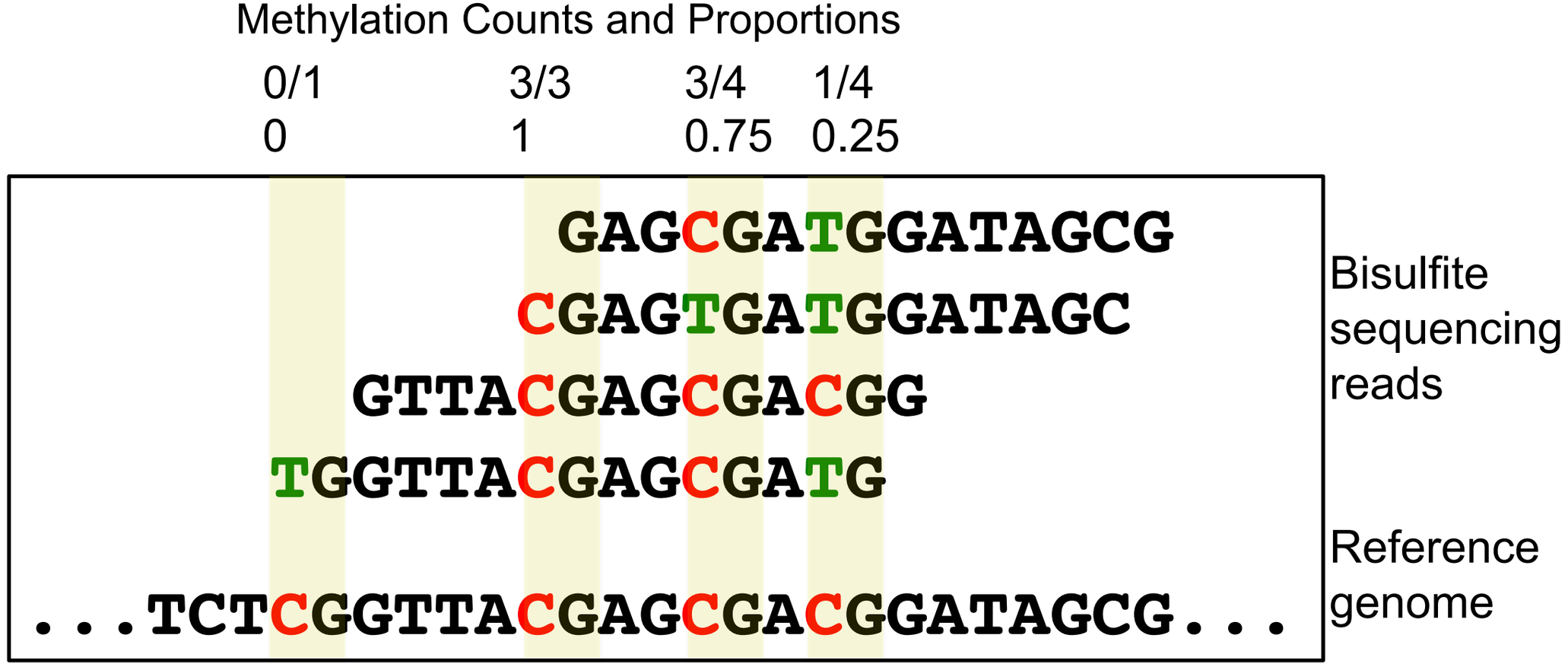
Treat DNA with bisulfite before sequencing

- Methylated C -> **unaffected**
- Unmethylated C -> **appear as T**

Whole Genome Bisulfite Sequencing

Treat DNA with bisulfite before sequencing

- Methylated C -> **unaffected**
- Unmethylated C -> **appear as T**



Methods for DMR detection

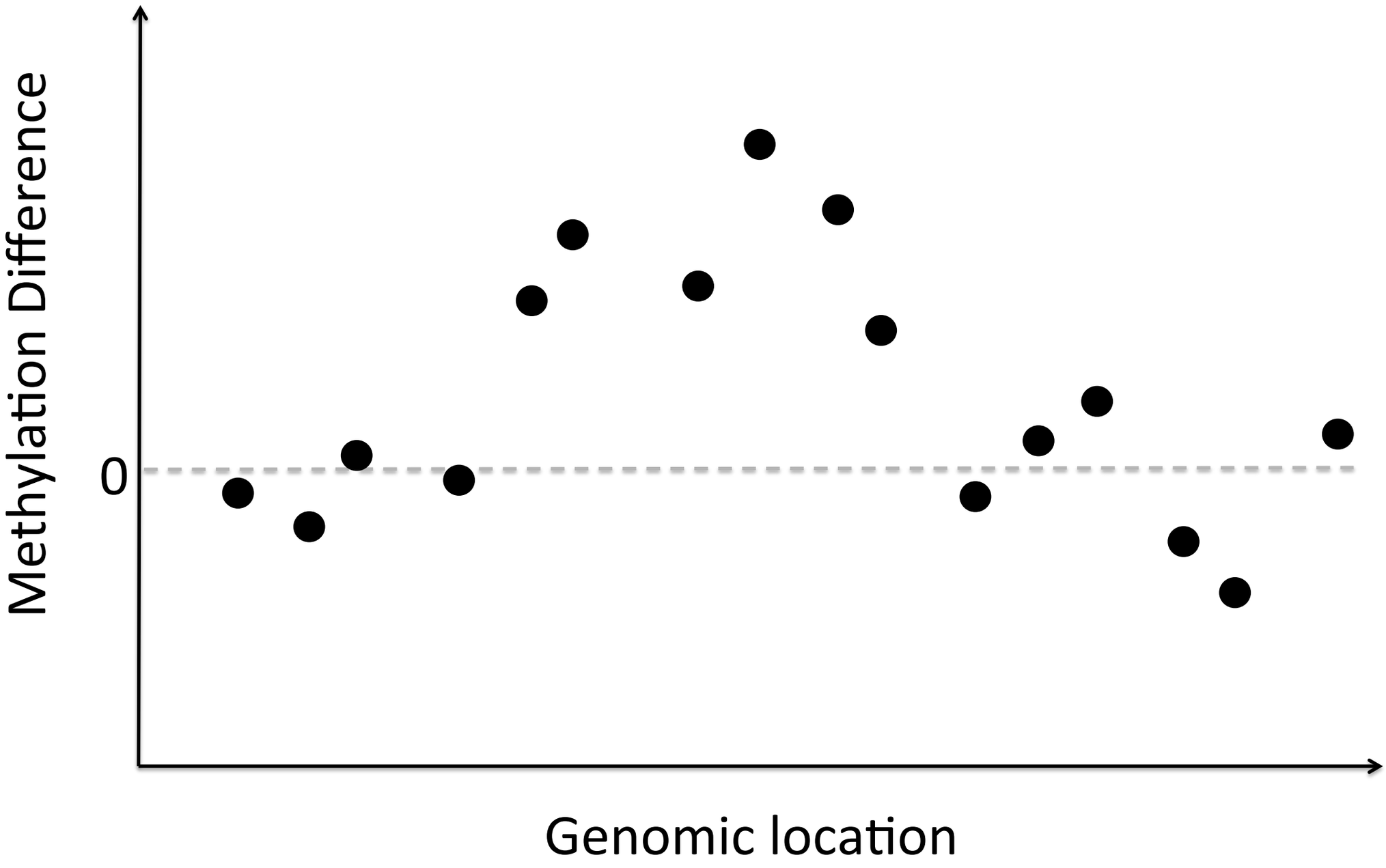
Challenges

- Have to handle small sample sizes !
- Accommodate known sources of variability
- Detect region boundaries

Two main strategies

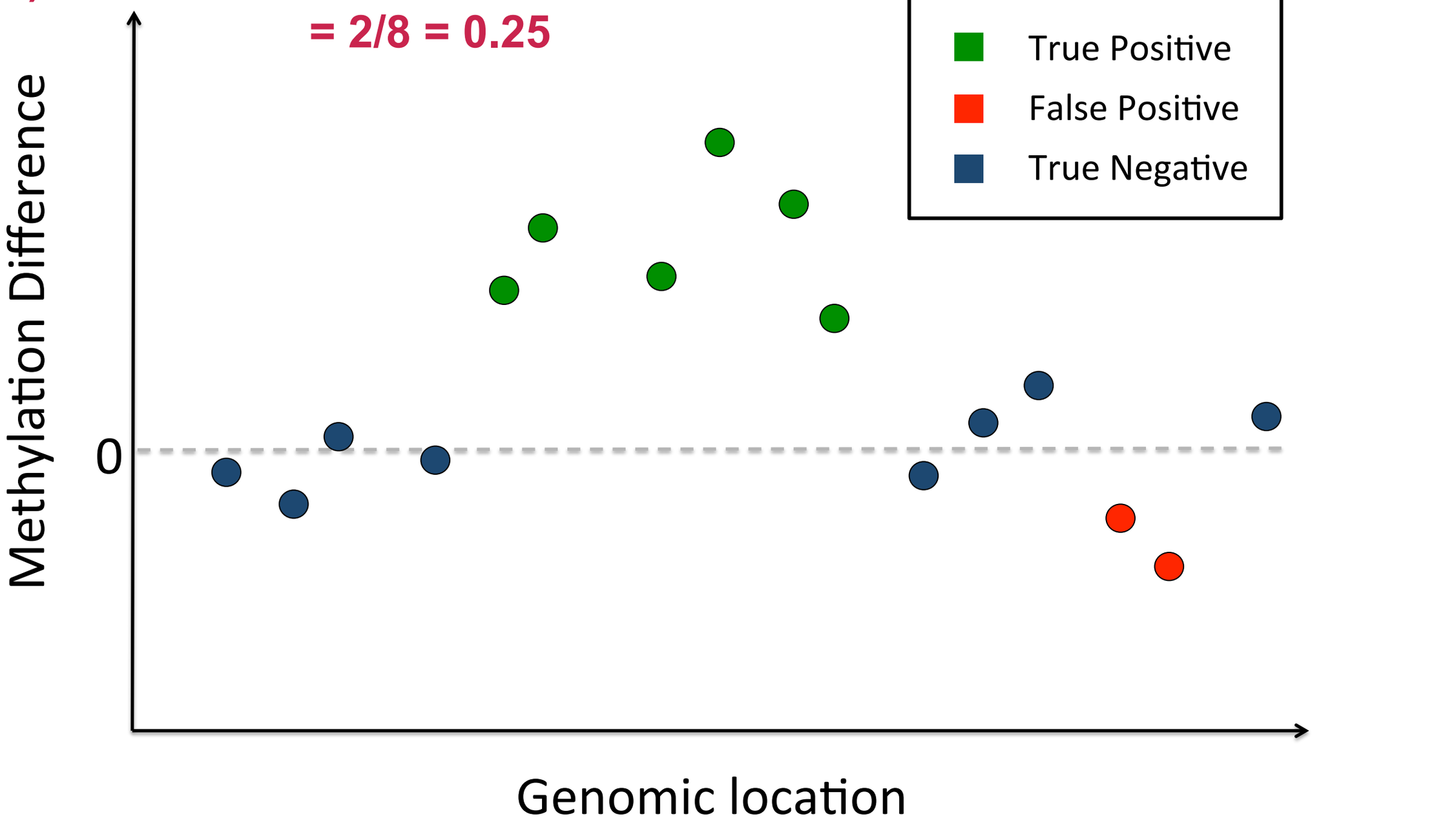
- Detect significant CpGs, then group together neighbors to form regions (DSS, BSmooth)
- Targeted regions / sliding windows specified in advance (BiSeq, MOABS)

Danger of grouping significant CpGs



Danger of grouping significant CpGs

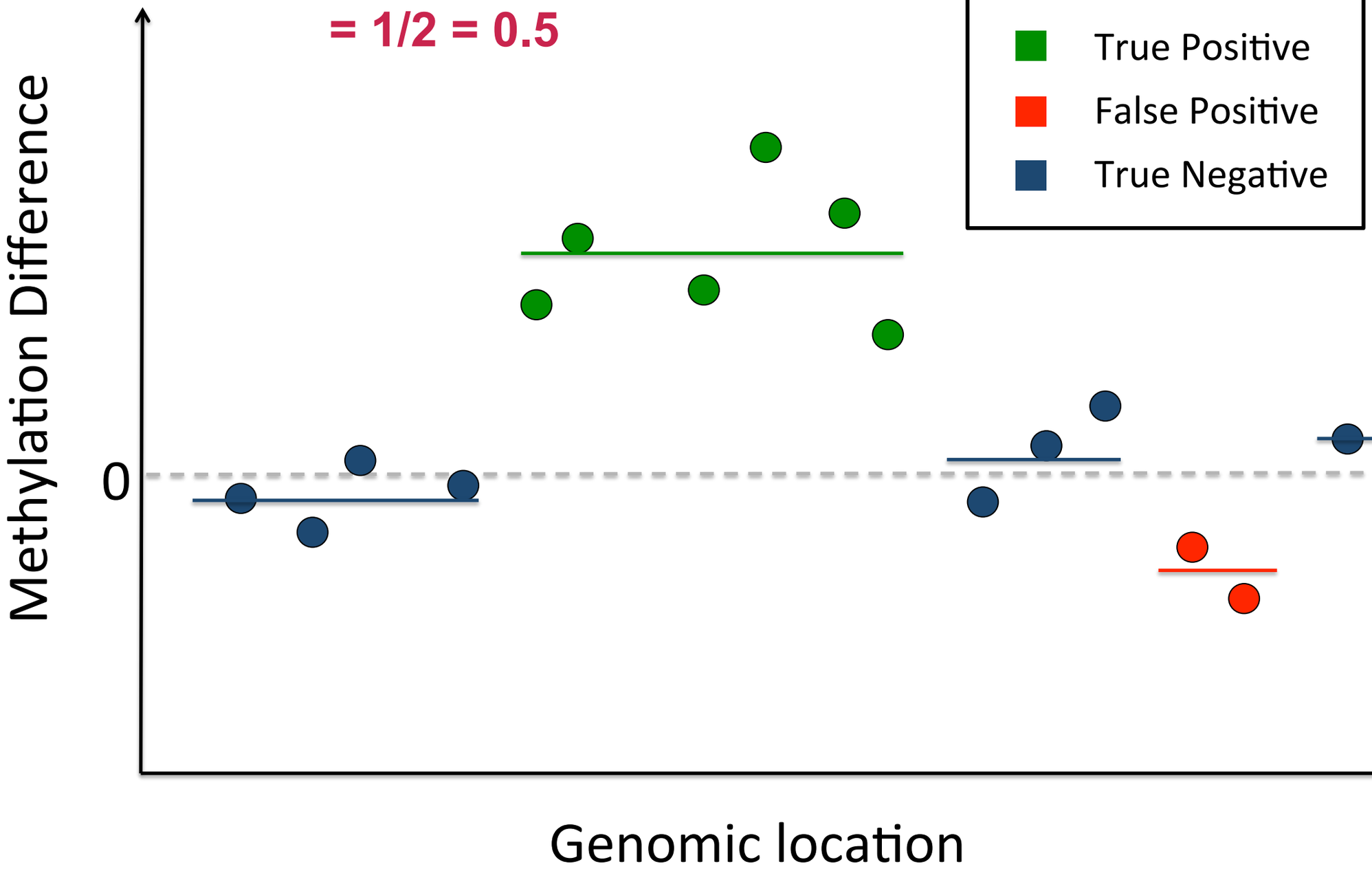
False Discovery Rate = # False Discoveries / Total # Discoveries
(FDR)
= 2/8 = 0.25



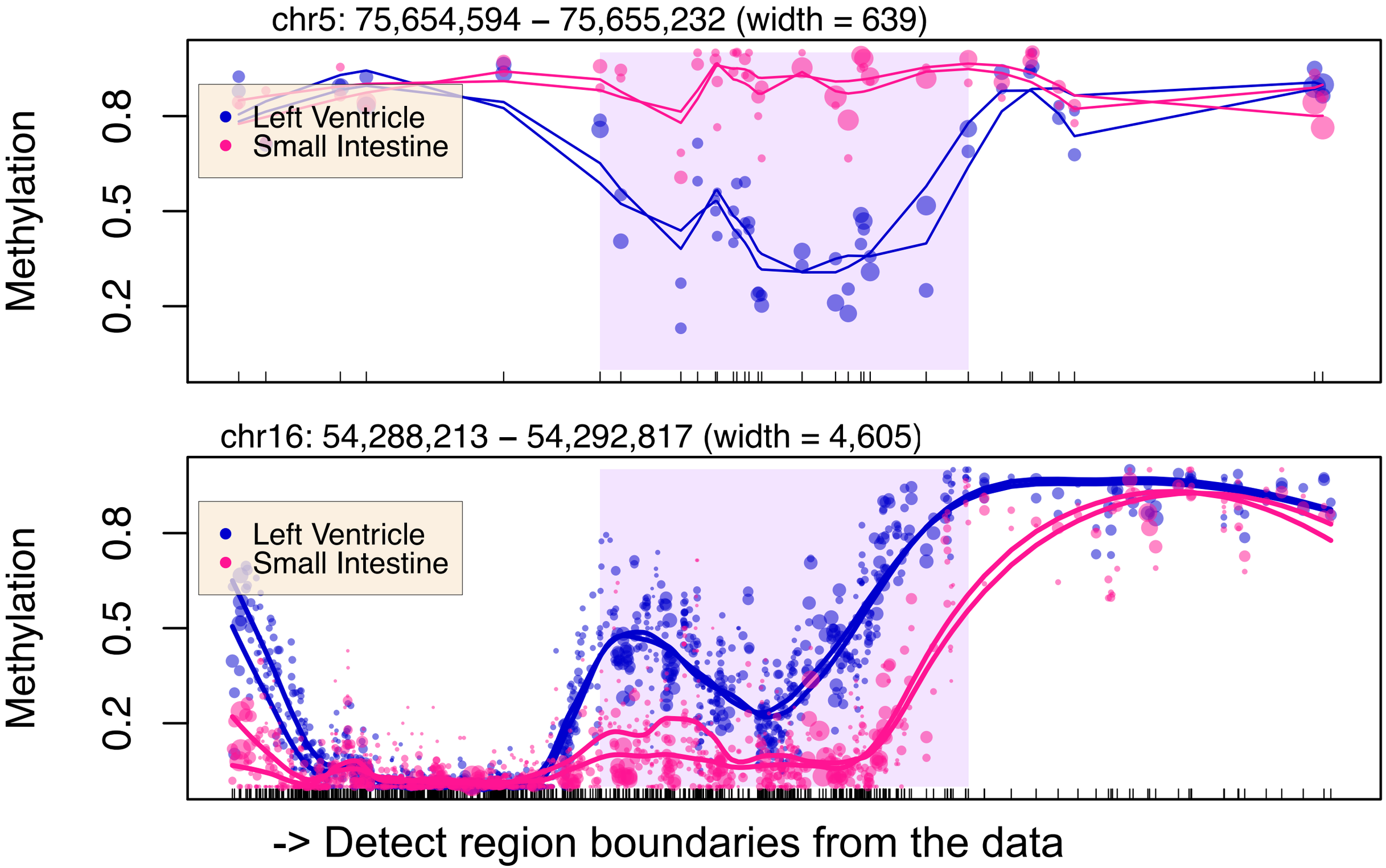
Danger of grouping significant CpGs

False Discovery Rate (FDR) = $\frac{\text{\# False Discoveries}}{\text{Total \# Discoveries}}$

= 1/2 = 0.5

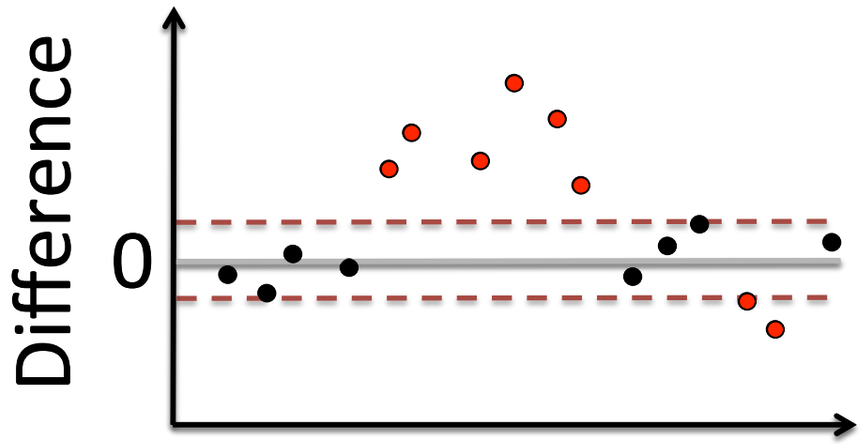


DMRs come in all shapes and sizes

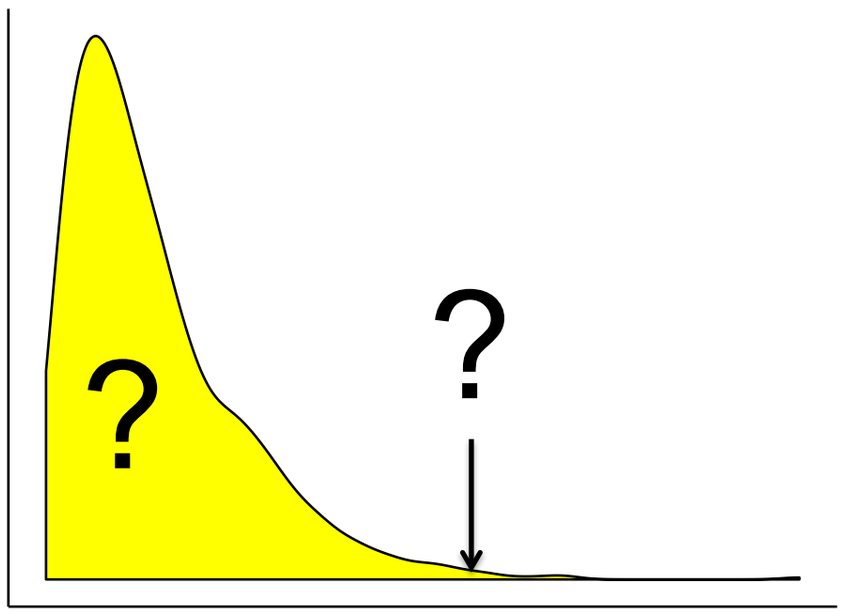


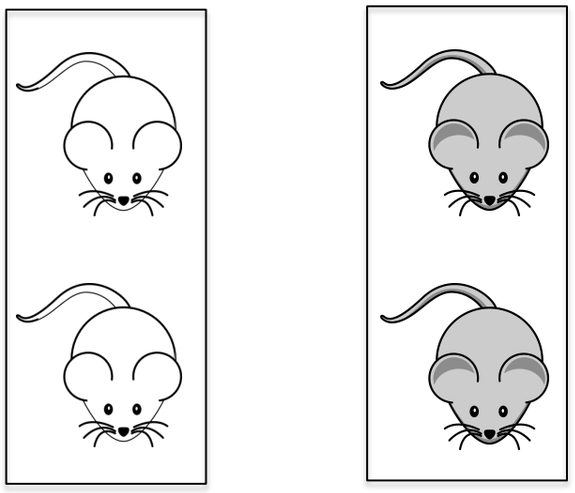
Our approach: **dmrseq**

1. Detect regions by scanning the genome for candidate regions



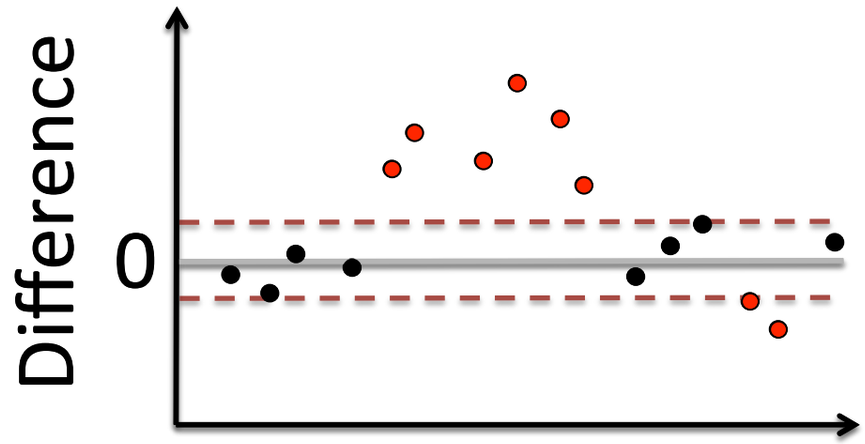
2. Region-level summary scores compared to a null to evaluate significance



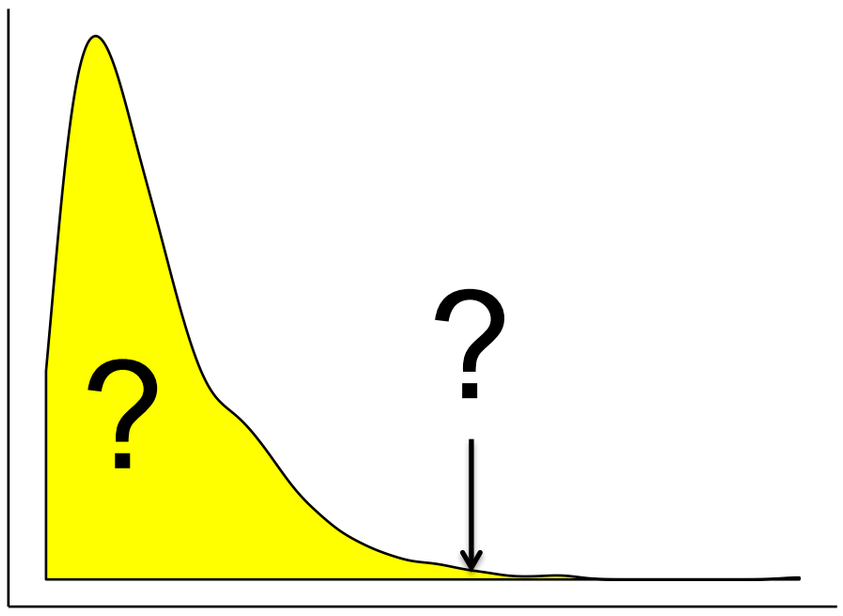


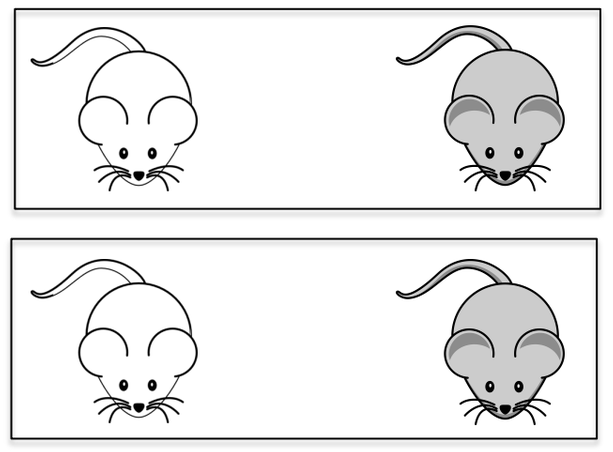
Our approach: **dmrseq**

1. Detect regions by scanning the genome for candidate regions



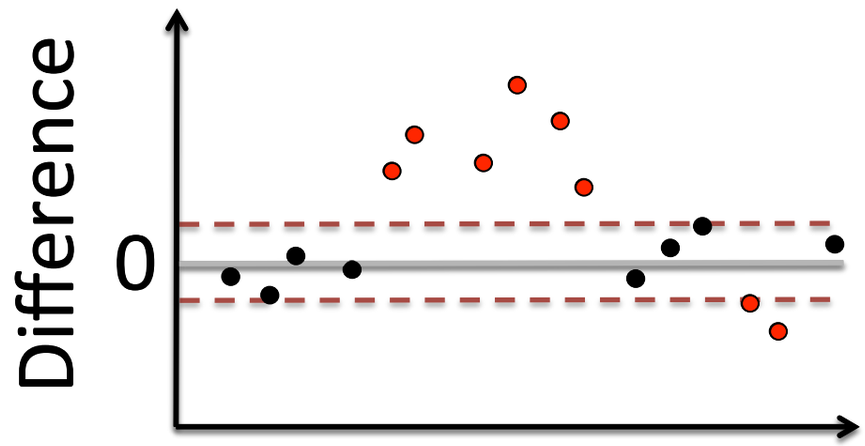
2. Region-level summary scores compared to a null to evaluate significance



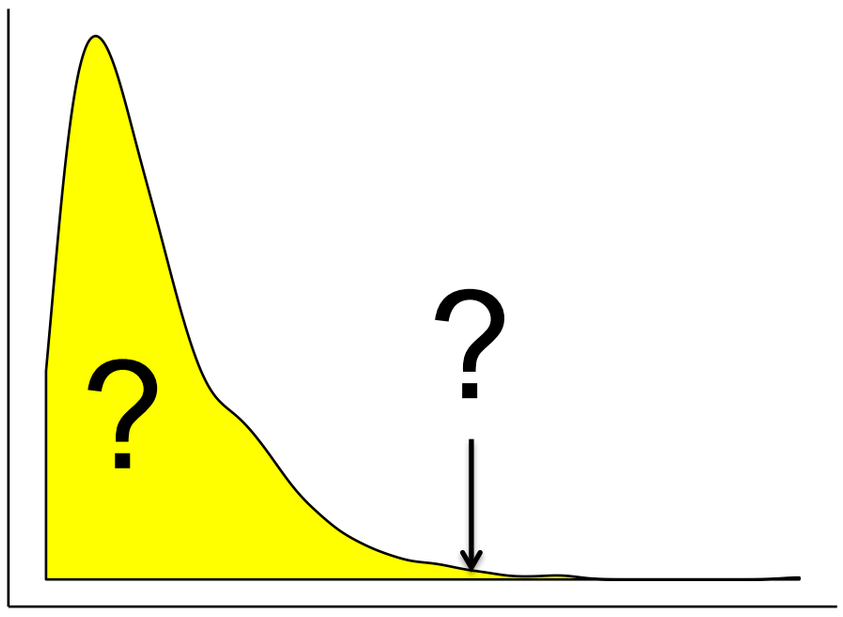


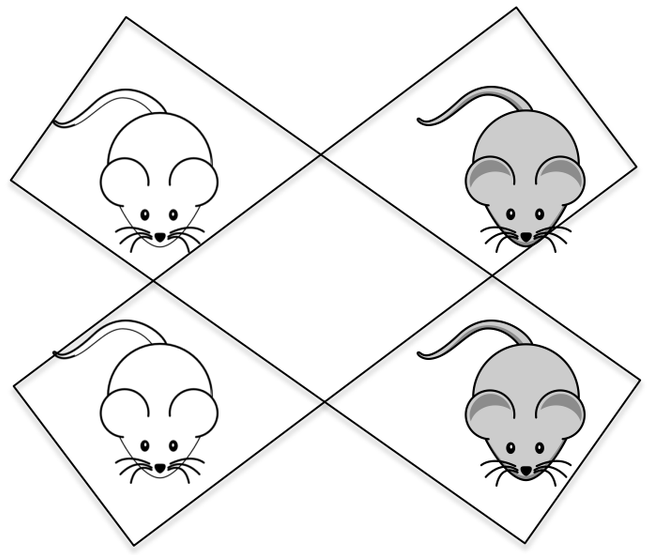
Our approach: **dmrseq**

1. Detect regions by scanning the genome for candidate regions



2. Region-level summary scores compared to a null to evaluate significance





Region level Summary Scores

Summary statistic that is approximately exchangeable across the genome so we can generate a **pooled null**

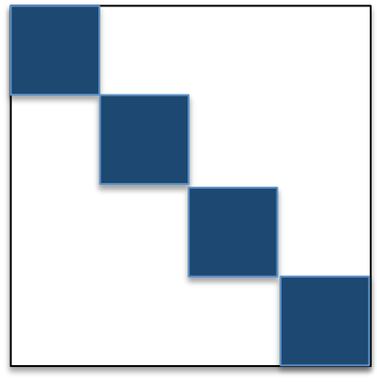
- Biological variability among samples
- Correlation among nearby CpGs
- Higher variability in CpGs with lower coverage

Generalized Least Squares (GLS) regression

$$\arcsin(2\pi_{ijr} - 1) = \sum_{l} \alpha_{lr} 1_{[i=1]} + \beta_r X_j + \varepsilon_{ijr}$$

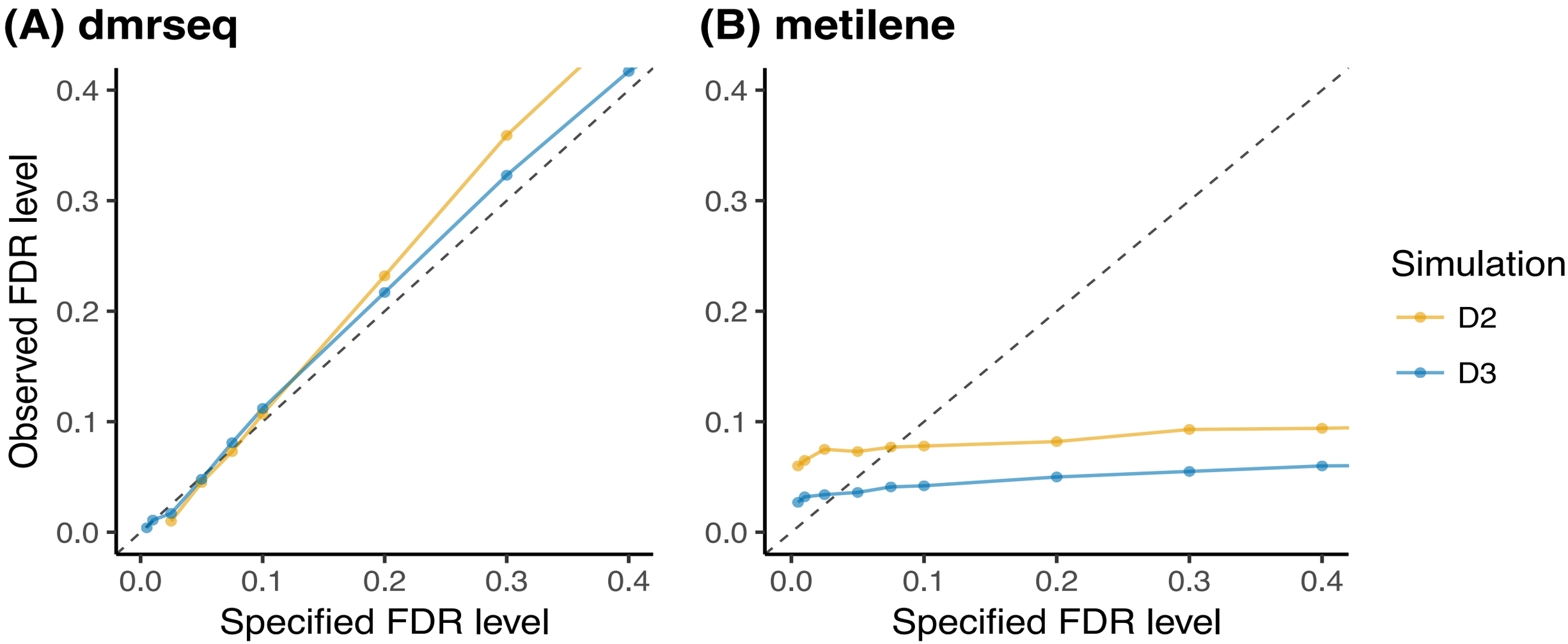
- Nested autoregressive correlated error structure

$$Var(\varepsilon_{ijr}) =$$

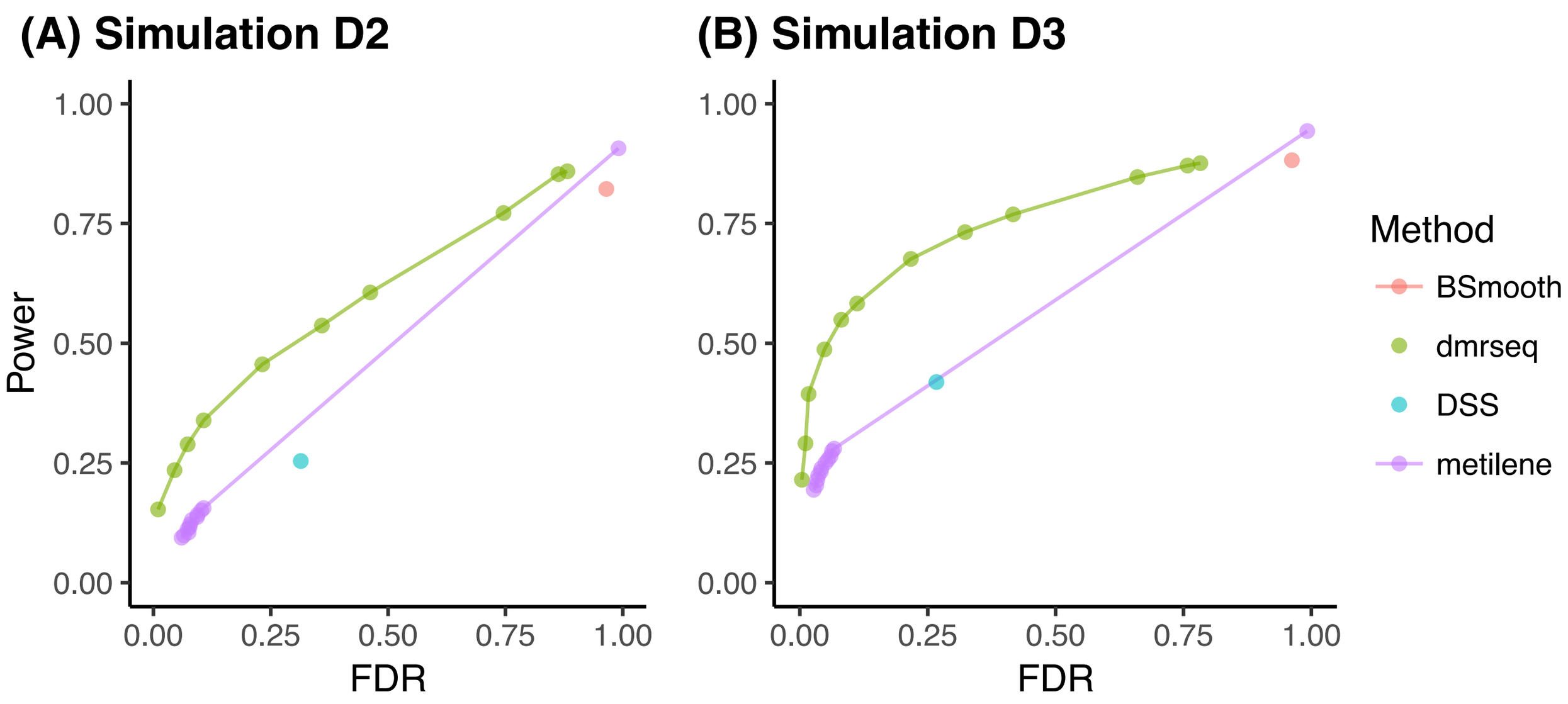


- Coverage-weighted variance

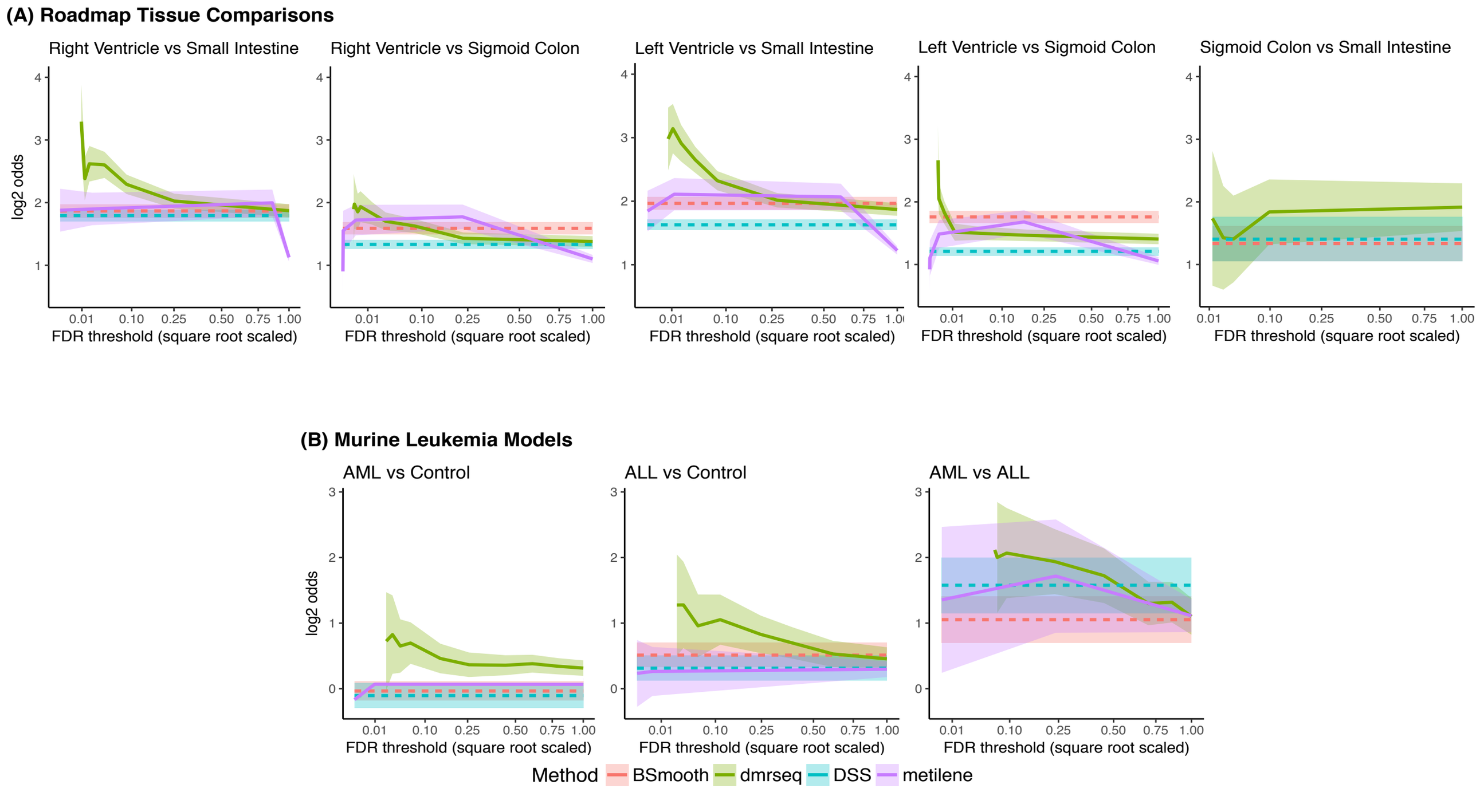
dmrseq accurately controls FDR



dmrseq has high sensitivity



DMRs enriched for associations with differential expression



dmrseq Summary

- dmrseq **identifies and prioritizes DMRs** from bisulfite sequencing experiments

- Computes region summary statistics that account for known sources of variability across the genome

- Achieves **accurate False Discovery Rate control** by generating a null distribution that pools information across the genome

Learn More

Slides
goo.gl/MwQz5f

dmrseq available on GitHub
https://github.com/kdkorthauer/dmrseq

Contact

keegan@jimmy.harvard.edu

@keegsdur

Acknowledgements

Rafa Lab
Rafael Irizarry
Chinmay Shukla

Collaborators
Sutirtha Chakraborty
Yuval Benjamini