Detection and accurate False Discovery Rate control of differentially methylated regions

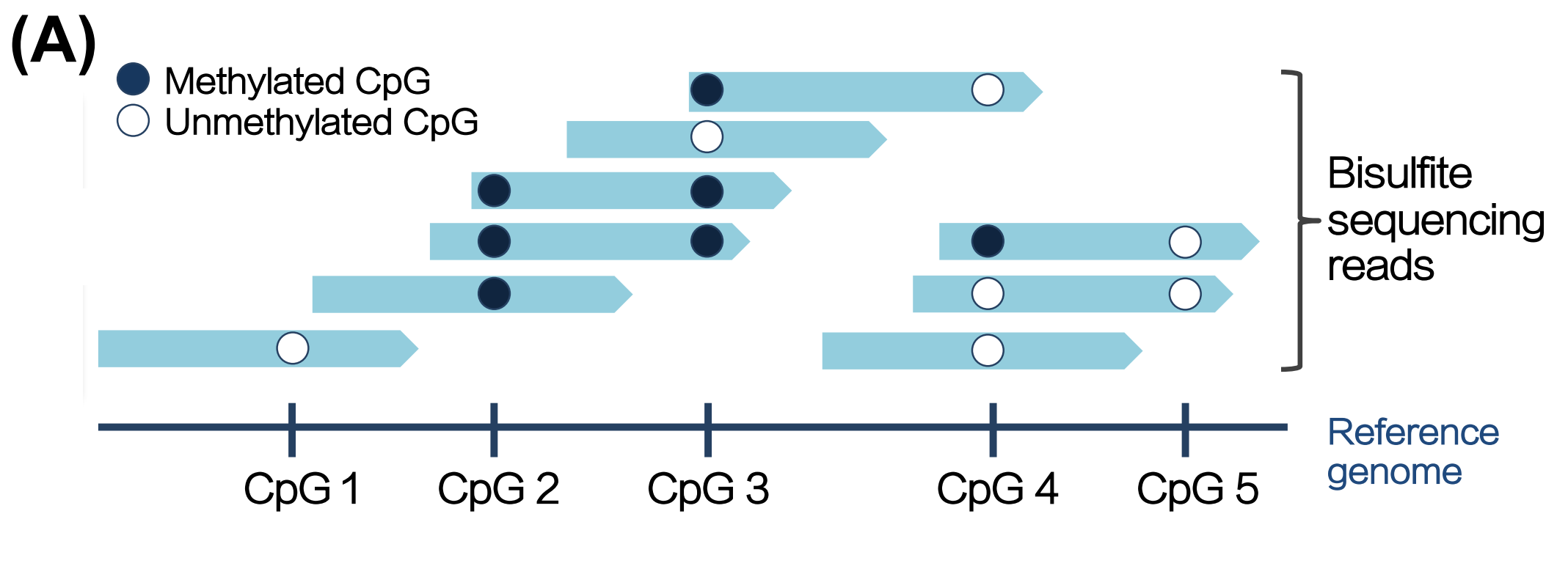
Keegan Korthauer[1,2*], Sutirtha Chakraborty[3], Yuval Benjamini[4], Rafael A. Irizarry[1,2]

[1]Harvard T.H. Chan School of Public Health, [2]Dana-Farber Cancer Institute, [3]Novartis, [4]Hebrew University of Jerusalem

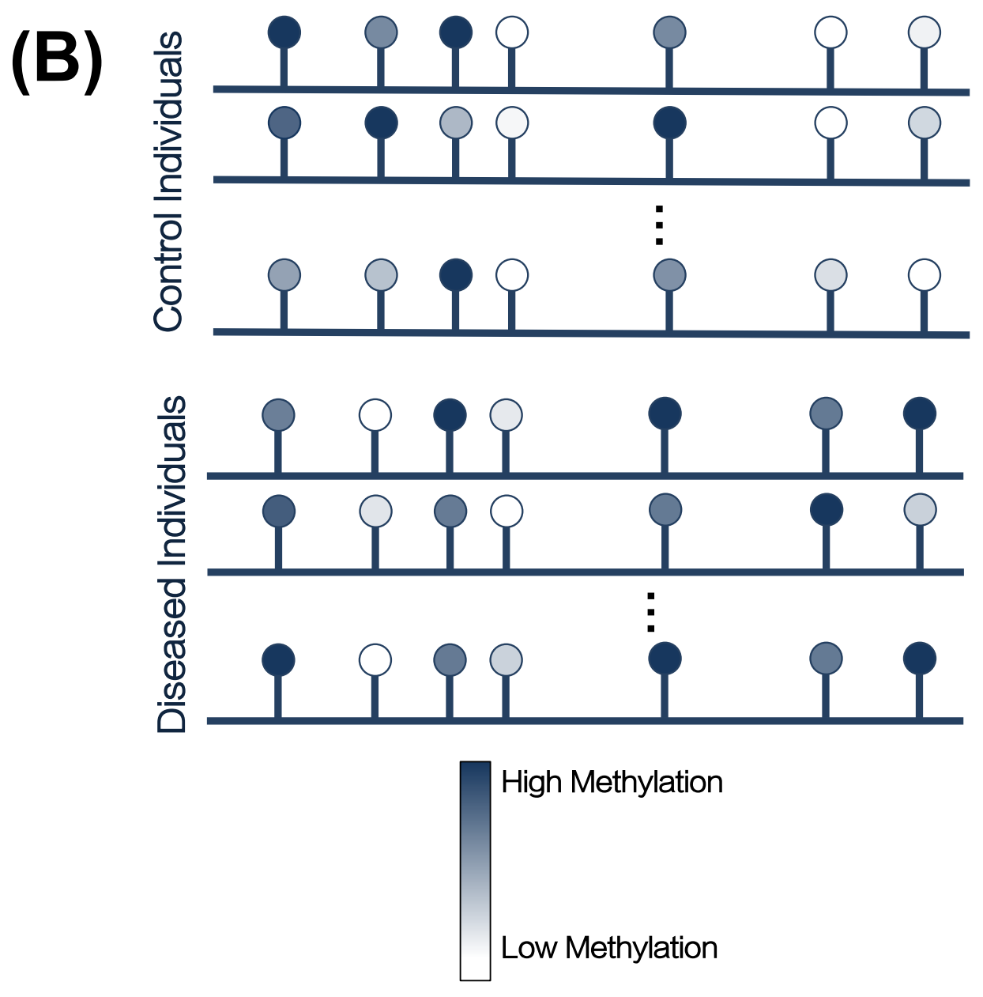
Abstract

- A central question in the analysis of bisulfite sequencing data is to detect regions with systematic differences between conditions
- Current computational approaches for detecting these so-called **Differentially Methylated Regions (DMRs)** do not provide accurate statistical inference
- Major challenges in reporting uncertainty are (1) accounting for a genome wide scan to detect regions, and (2) the limited sample sizes of typical experiments
- The R package ***dmrseq*** overcomes these challenges using a permutation-based approach to **detect and perform accurate inference for differential methylation**
- We find that the new method improves the specificity and sensitivity of lists of regions and **accurately controls the False Discovery Rate (FDR)** in experiments with as few as two samples per group [1]

Measuring methylation with Whole Genome Bisulfite Sequencing (WGBS)

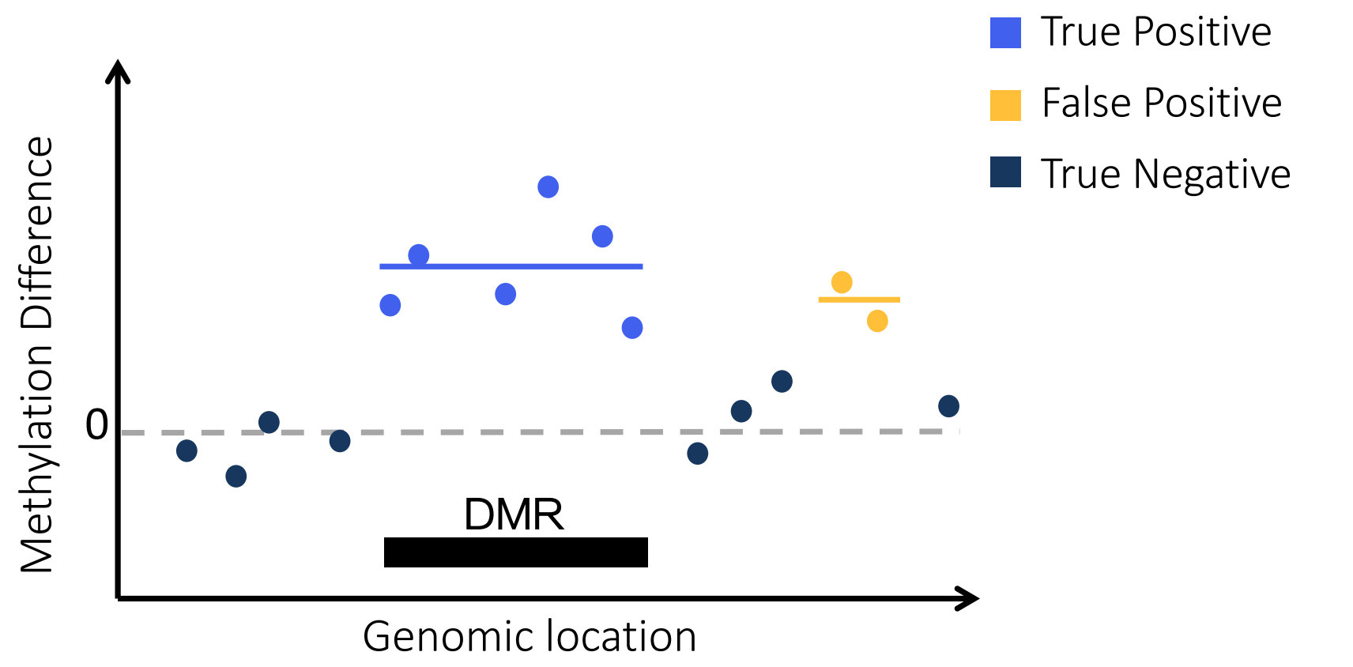


	CpG 1	CpG 2	CpG 3	CpG 4	CpG 5
Methylated Count (M)	0	3	3	1	0
Coverage (N)	1	3	4	4	2
Proportion (M/N)	0	1	0.75	0.25	0



WGBS data and differential methylation (A) Although CpGs on a single DNA strand have binary methylation status, WGBS measures the proportion of methylated reads covering each CpG. Total coverage of each CpG varies due to sampling error. **(B)** The task is to identify regions (groups of CpGs) that are significantly associated with some phenotype of interest, such as disease status. Spatial, individual, and coverage variability need to be accounted for.

Grouping significant CpGs does not control error rate of DMRs



False Discovery Rate (FDR) =
# False Discoveries / Total # Discoveries

$\text{FDR}_{\text{CpG}} = 2/8 = 0.25$

vs

$\text{FDR}_{\text{DMR}} = 1/2 = 0.50$

Controlling FDR at the CpG level does **not** guarantee FDR control at the region level

Region-Level Modeling

CpG level:

$$M_{ijr} | N_{ijr}, p_{ijr} \sim Bin(N_{ijr}, p_{ijr})$$
$$p_{ijr} \sim Beta(a_{irs}, b_{irs})$$
$$\pi_{irs} = \frac{a_{irs}}{(a_{irs}+b_{irs})}$$

M_{ijr} = methylated read count
N_{ijr} = total coverage
p_{ijr} = methylation proportion
π_{irs} = methylation proportion for condition s
i indexes CpGs, j indexes samples, $j \in C_s$
s indicates biological condition

Region level:

$$g(\boldsymbol{\pi}_r) = \boldsymbol{X\beta}_r = \underbrace{\sum_{l=1}^{L_r} \beta_{0lr} 1_{[i=l]}}_{\text{loci-specific intercept}} + X_j\beta_{1r}$$

condition effect

Fit with Generalized Least Squares (GLS) and variance-stabilizing arcsine transformation [3]:

$$Z_{ijr} = arcsin(2\, M_{ijr}/N_{ijr} - 1)$$

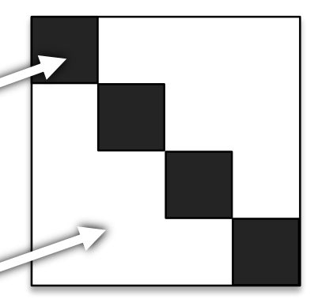
$$\boldsymbol{Z}_r = \boldsymbol{X\beta}_r + \boldsymbol{\epsilon}_r$$

where $E[\boldsymbol{\epsilon}_r] = \boldsymbol{0}$ and $Var[\boldsymbol{\epsilon}_r] = \boldsymbol{V}_r$

Within Sample: $\widehat{Cov}(\boldsymbol{Z}_{jr}) = \widehat{\boldsymbol{V}}_{jr} = \hat{\sigma}_r^2 \widehat{\boldsymbol{R}}_{jr}$

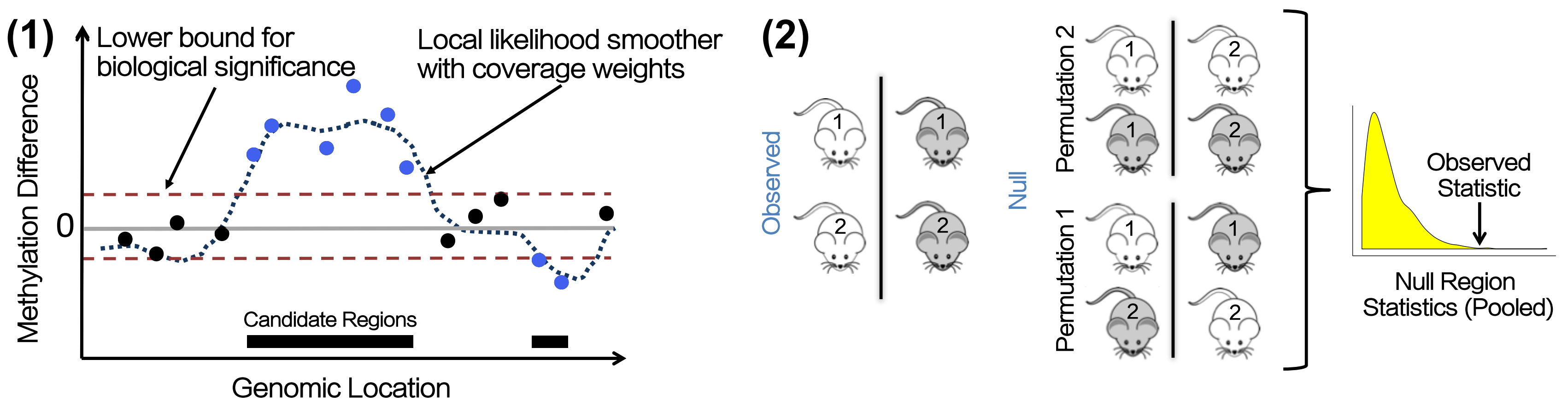
with ik^{th}elelement of $\boldsymbol{R}_{jr}$: $\{\widehat{\boldsymbol{R}}_{jr}\}_{\boldsymbol{ik}} = \frac{e^{-\hat{\Phi}_r |t_{ir} - t_{kr}|}}{\sqrt{N_{i\cdot r}\, N_{k\cdot r}}}$

Between Sample: $Cov(Z_{ijr}, Z_{ij^*r}) = 0$



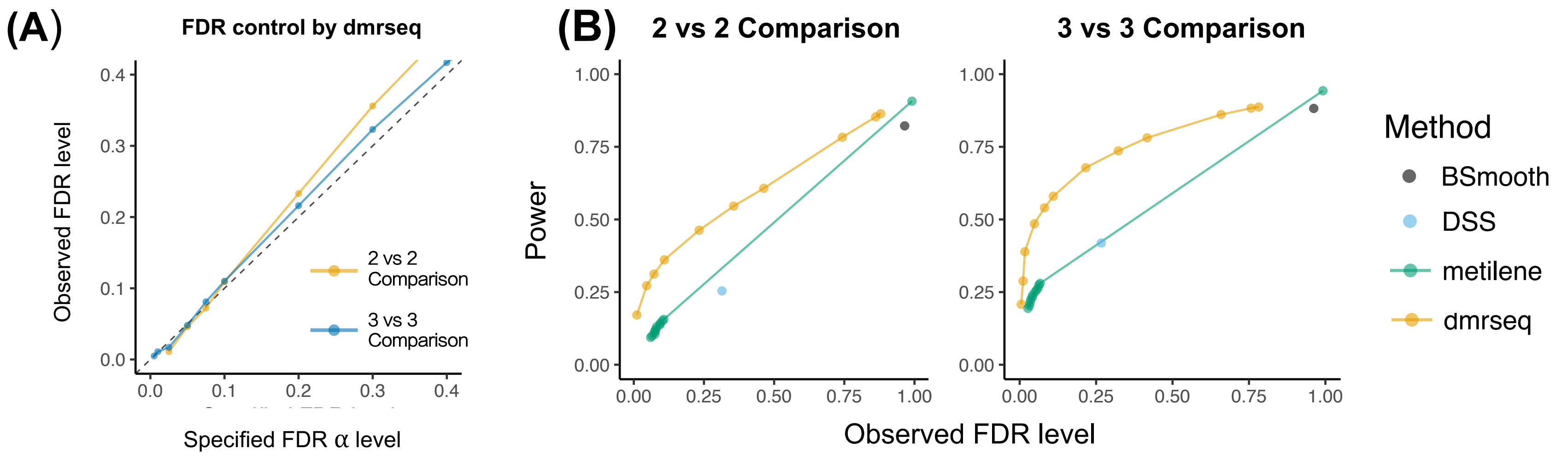
$$H_0: \beta_{1r} = 0, \text{ where } \widehat{\boldsymbol{\beta}}_r = \left(\boldsymbol{X}^t \boldsymbol{V}_r^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{V}_r^{-1} \boldsymbol{X}^t \boldsymbol{V}_r^{-1} \boldsymbol{Z}_r$$

dmrseq: two step approach



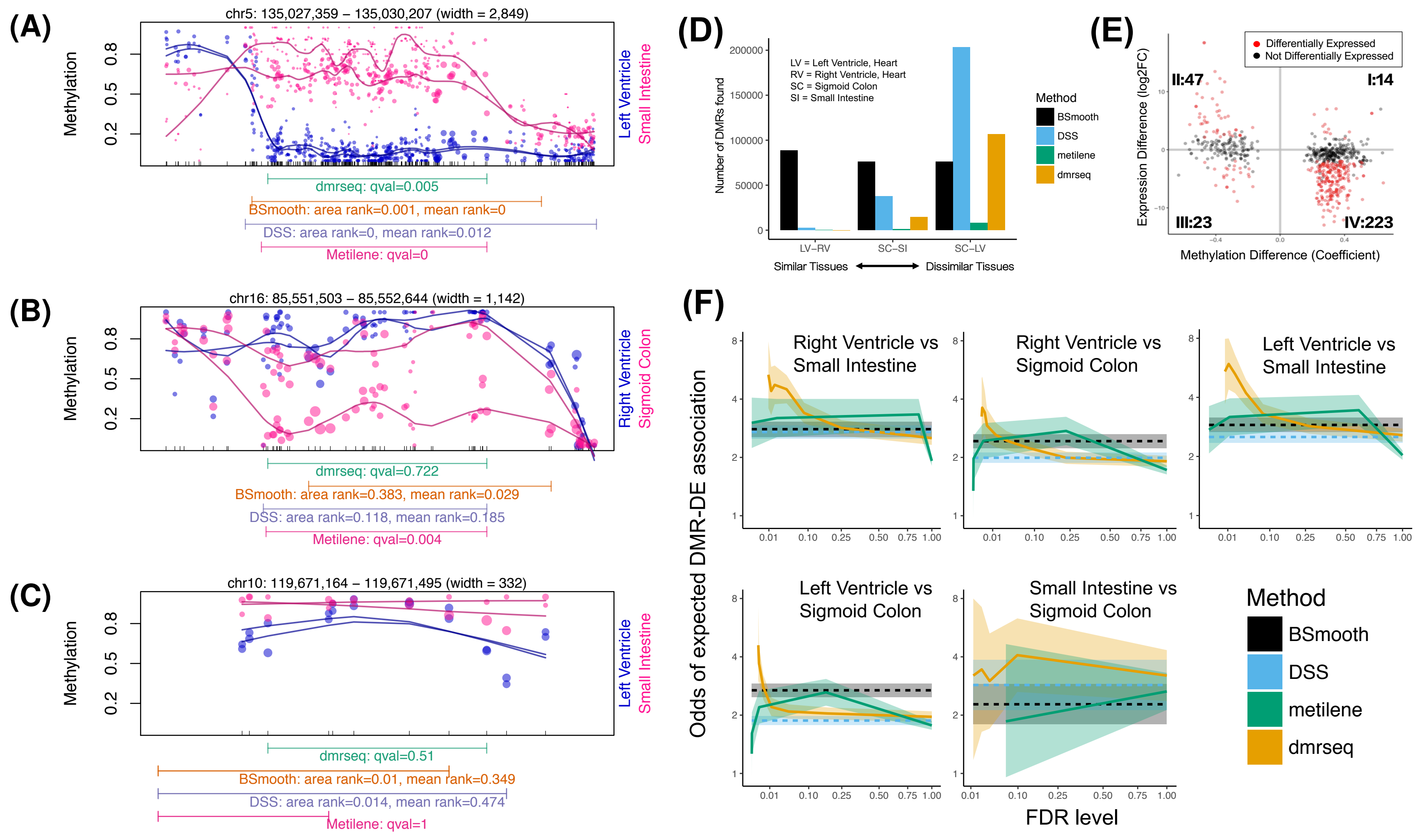
(1) Candidate region detection: Genome wide scan of CpG methylation difference
(2) Evaluate statistical significance: Compare observed region-level summary statistics against pooled null permutation distribution

Accurate FDR control and high power in simulation



Simulation results based on control sample comparison with added DMRs (A) dmrseq achieves accurate FDR control in simulation. **(B)** Compared to alternative approaches Bsmooth [2], DSS [3], and metilene [4], dmrseq achieves greater sensitivity and specificity.

Roadmap tissue-specific DMRs enriched for associations with expression



Results from case study of tissue-specific DMR identification in NIH Epigenome Roadmap [5] (A-C) dmrseq q-values accurately rank regions by statistical significance. Methods that do not account for sample and spatial variability make spurious calls, as in B and C. **(D)** dmrseq finds fewer tissue-specific DMRs when comparing similar tissues. **(E)** DMRs can be validated by correlating expression values of nearby genes. **(F)** Compared to alternative approaches BSmooth [2], DSS [3], and metilene [4], DMRs found by dmrseq have stronger odds of expected association with expression of nearby genes.

Summary

- *dmrseq* **identifies and prioritizes DMRs** from bisulfite sequencing experiments
- Accounts for sample & spatial variability by **modeling signal at the region level**
- Achieves **accurate False Discovery Rate control** by generating a null distribution that pools information across the genome

Contact

Keegan Korthauer, PhD
Postdoctoral Fellow
Harvard T.H. Chan School of Public Health
Dana-Farber Cancer Institute

keegan@jimmy.harvard.edu
@keegankorthauer
kkorthauer.org
github.com/kdkorthauer/dmrseq

References

[1] Korthauer, K., Chakraborty, S., Benjamini, Y., and Irizarry, R. A. (2018). Detection and accurate False Discovery Rate control of differentially methylated regions from Whole Genome Bisulfite Sequencing. *Biostatistics*, to appear.

[2] Hansen, K. D., Langmead, B., and Irizarry, R. A. (2012). BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology*, *13*(10), R83.

[3] Park, Y. and Wu, H. (2016). Differential methylation analysis for BS-seq data under general experimental design. *Bioinformatics*, *32*(10), 1446-1453.

[4] Jühling, F., Kretzmer, H., Bernhart, S. H., et al. (2016). metilene: Fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. *Genome Research*, *26*(2), 256-262.

[5] Schultz, M. D., He, Y., Whitaker, J. W., et al. (2015). Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature*, *523*(7559), 212.