De novo detection and accurate inference of differentially methylated regions

Keegan Korthauer, PhD

Joint Statistical Meetings, Vancouver, CA
29 July 2018

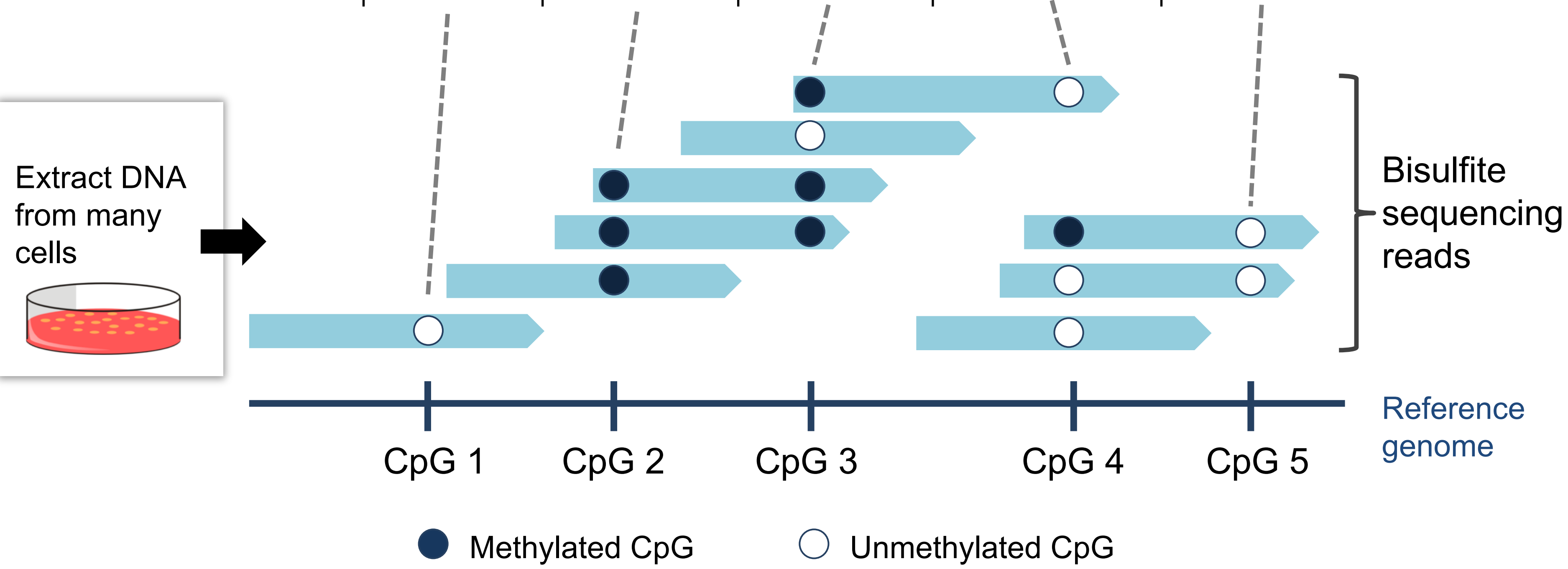
DANA-FARBER CANCER INSTITUTE

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

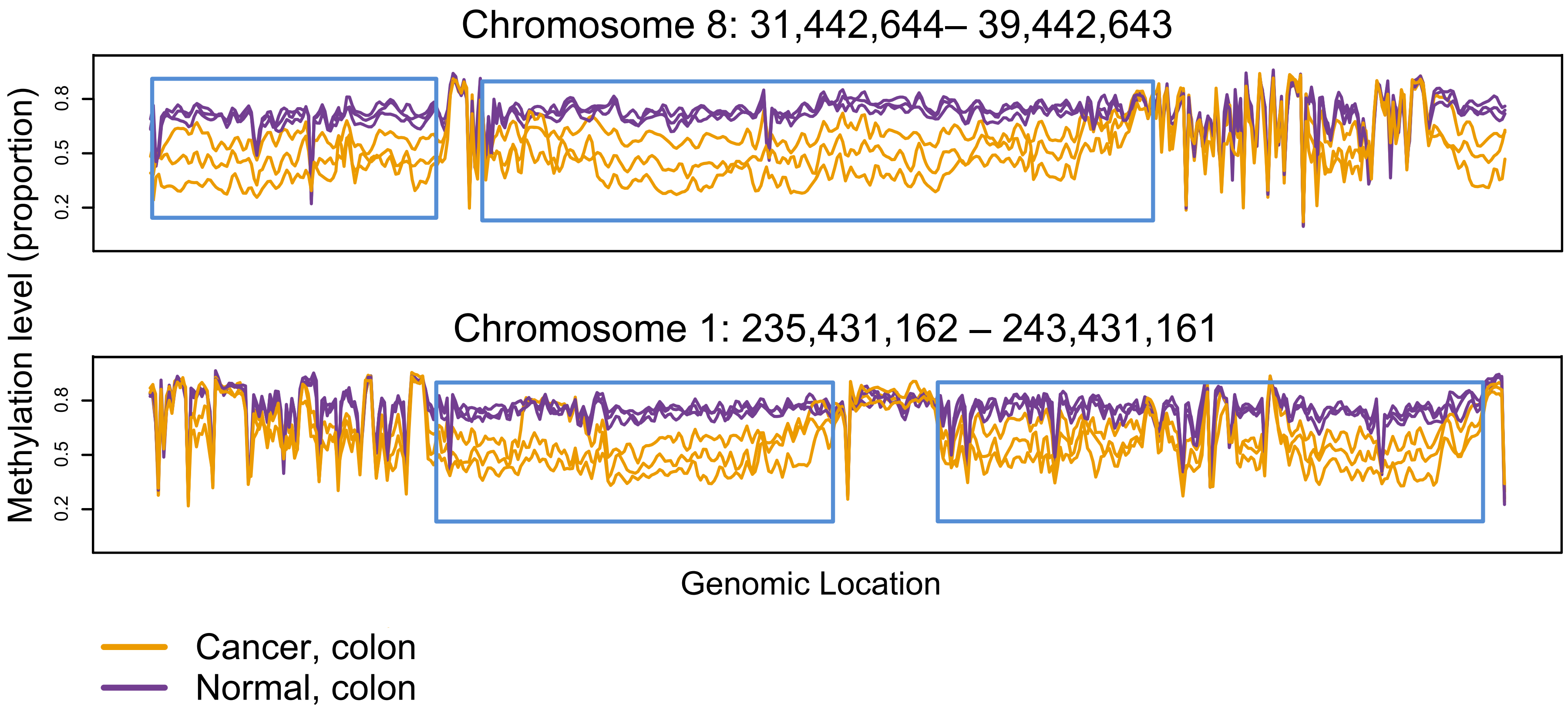
Whole Genome Bisulfite Sequencing (WGBS)

Methylation Sequencing Data

	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

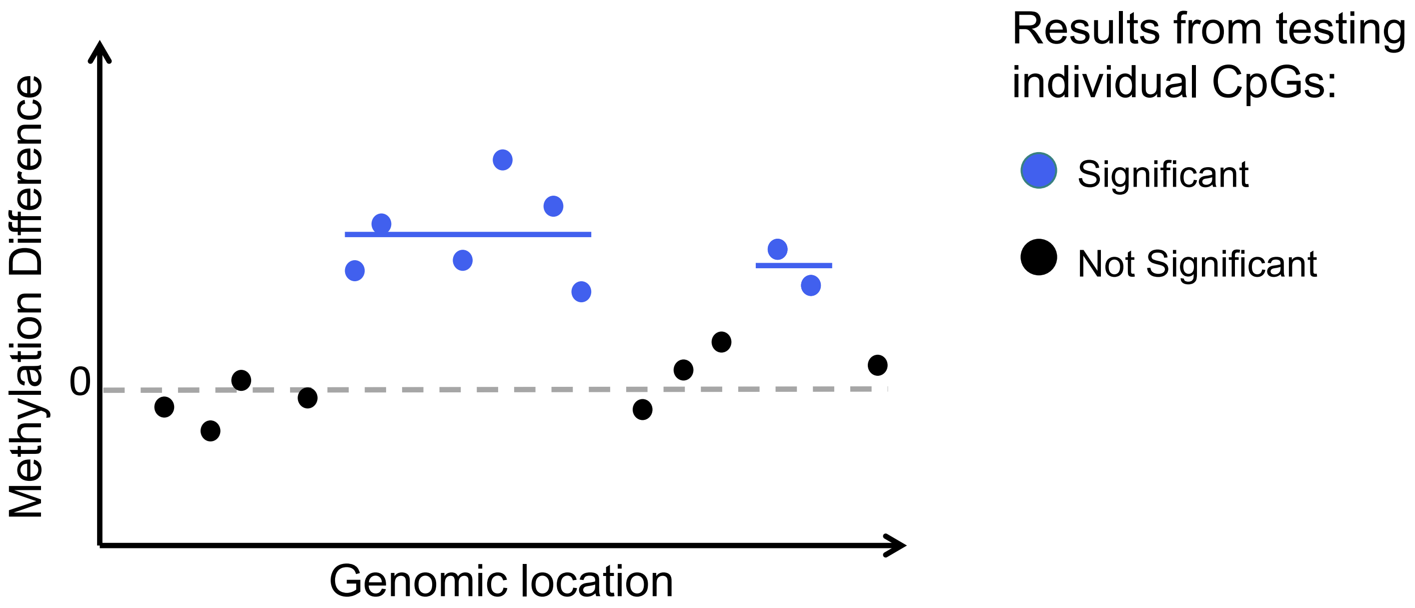


Differentially Methylated Regions (DMRs)



Adapted from Hansen et al. 2014, *Genome Research*

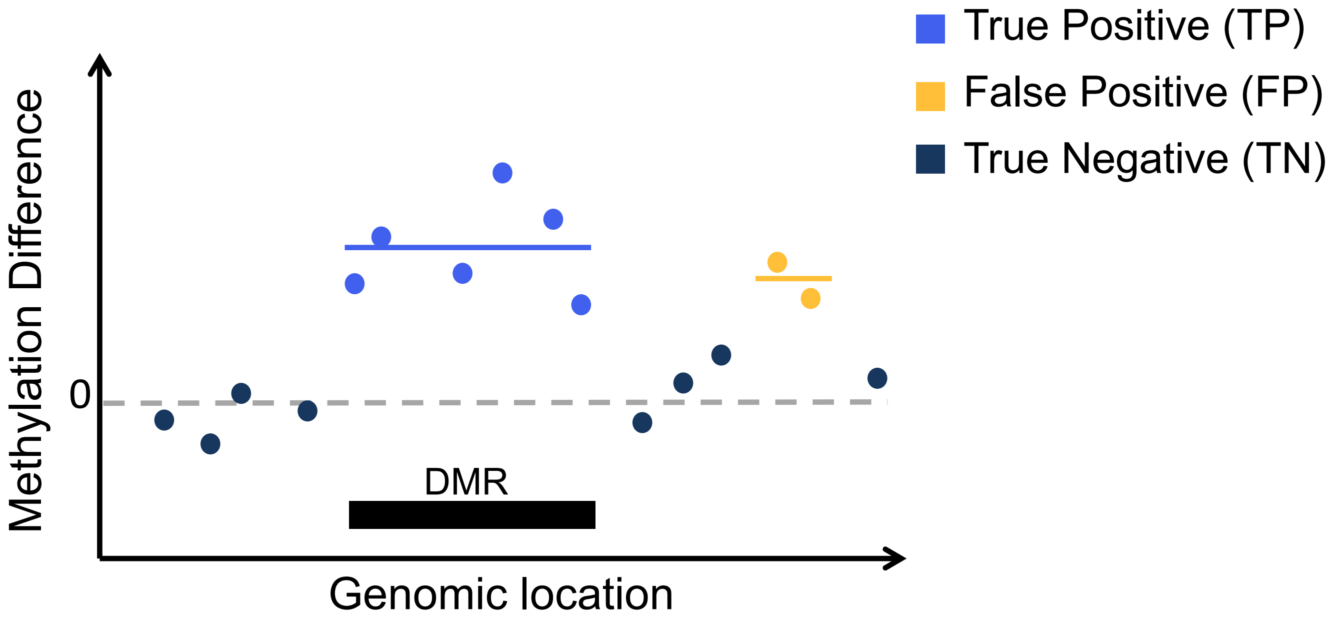
Previous methods: Grouping significant CpGs



Examples:

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

Error rate not controlled at the region level

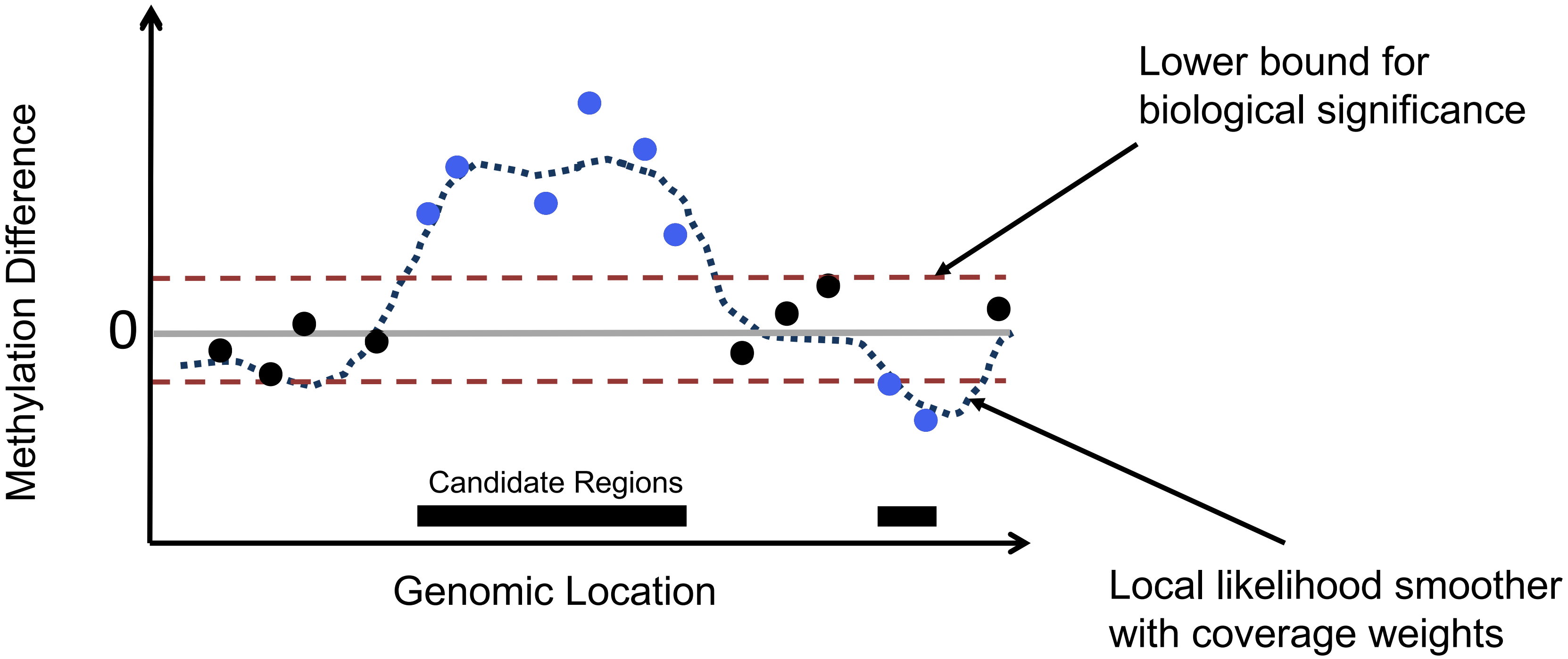


$$\text{False Discovery Rate } (FDR) = E\left[\frac{FP}{TP+FP}\right]$$

$$\widehat{FDR}_{CpG} = \frac{2}{8} = 0.25 \quad vs \quad \widehat{FDR}_{DMR} = \frac{1}{2} = 0.50$$ ⚠

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

Genome-wide scan of CpG methylation difference

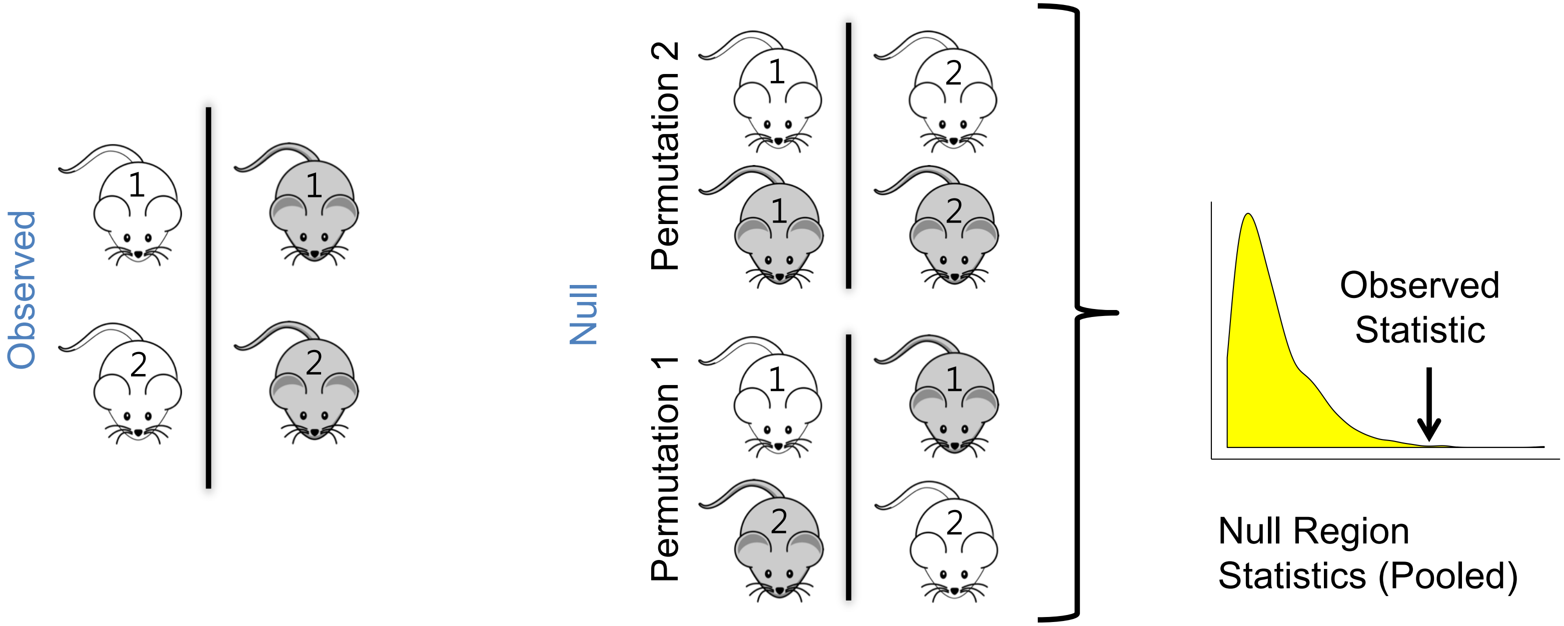


Lower bound for biological significance

Local likelihood smoother with coverage weights

dmrseq: (2) Assess region-level signal

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



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, where $j \in C_s$

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

$$\boxed{H_0: \beta_{1r} = 0}$$

Region-level model fitting

Generalized Least Squares (GLS) with variance stabilizing transformation:

arcsine link transformation (Park & Wu 2016)

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

$Var(M_{ijr}/N_{ijr}) \propto \pi_{ijr}(1 - \pi_{ijr})$ but $Var(Z_{ijr}) \approx \frac{1+(N_{ijr}-1)\gamma_{irs}}{N_{ijr}}$

↓ ↓

Variance depends on mean Variance independent of mean

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

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

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{r}} = \left(\boldsymbol{X}^{t}\boldsymbol{V}_{\boldsymbol{r}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{V}_{\boldsymbol{r}}^{-1}\boldsymbol{X}^{t}\boldsymbol{V}_{\boldsymbol{r}}^{-1}\boldsymbol{Z}_{\boldsymbol{r}}$$

Account for variability across samples and locations

(1) Correlation: Continuous Autoregressive (CAR) model

$$\rho(Z_{ijr}, Z_{kjr}) = e^{-\phi_r |t_{ir} - t_{kr}|}$$

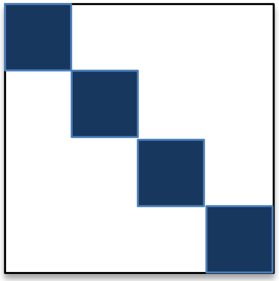
t_{ir} = genomic location of CpG i

(2) Variability dependent on coverage

$$Var(Z_{ijr}) \propto \frac{1}{N_{i \cdot r}}$$

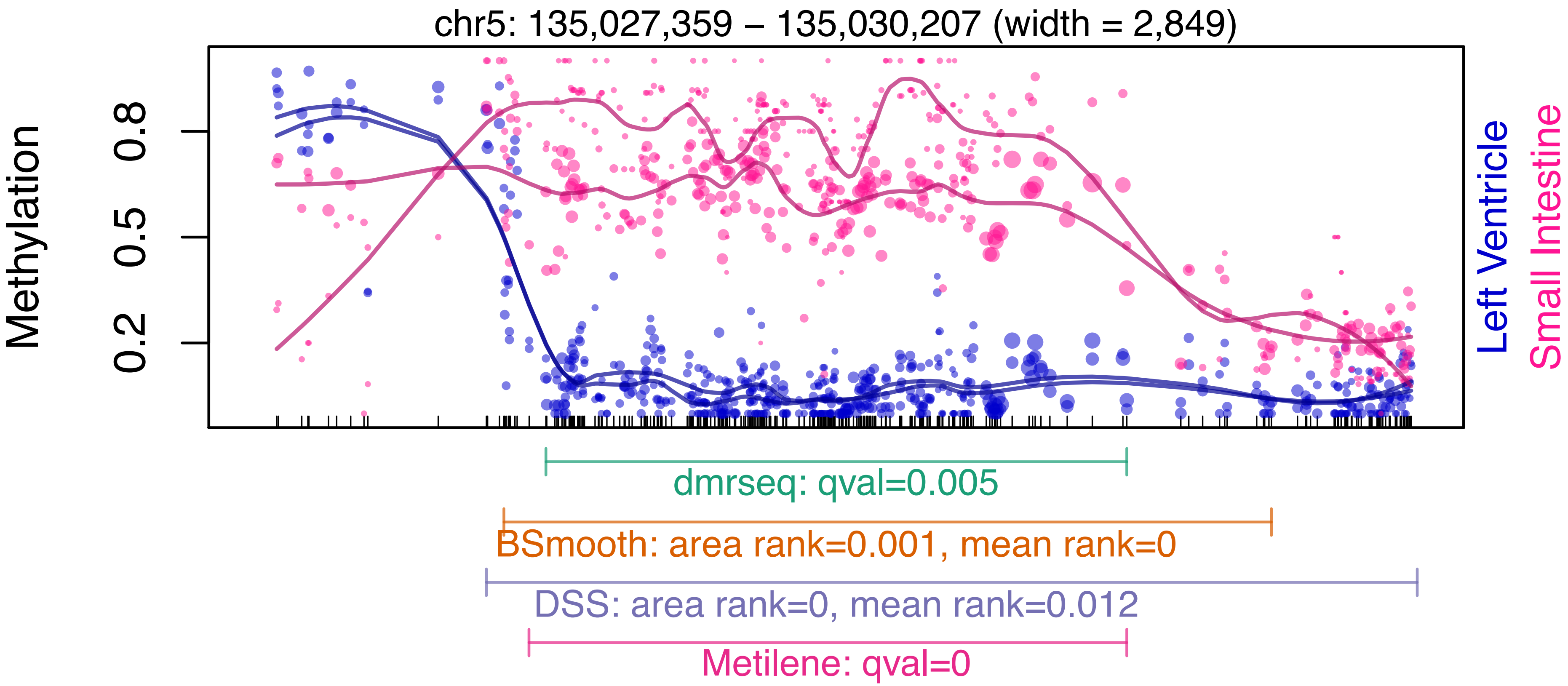
(3) Within sample correlation

Independent samples



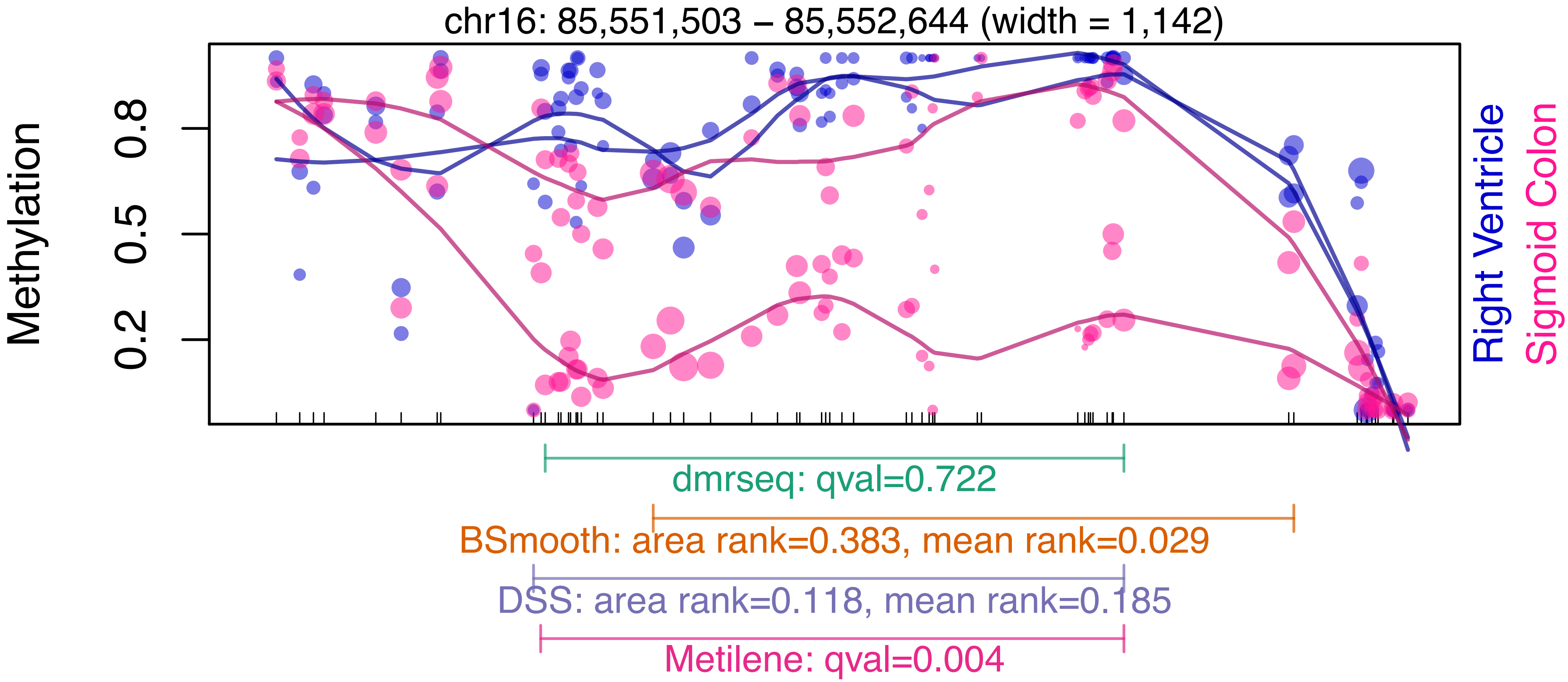
$$Cov(Z_{ijr}, Z_{ij^*r}) = 0$$

Example: highly ranked DMR across all methods



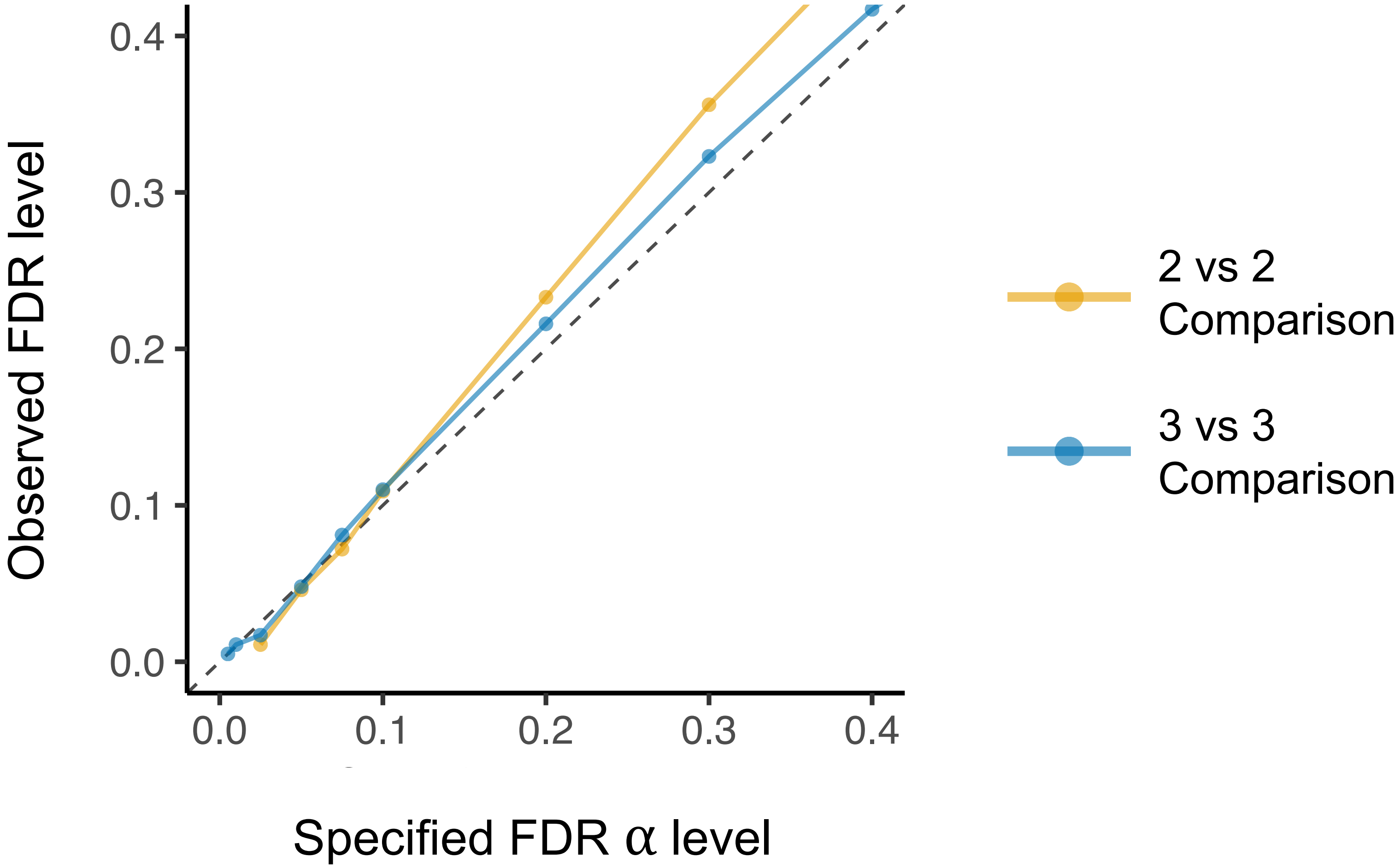
Korthauer et al., 2018

Example: dmrseq accounts for sample variability

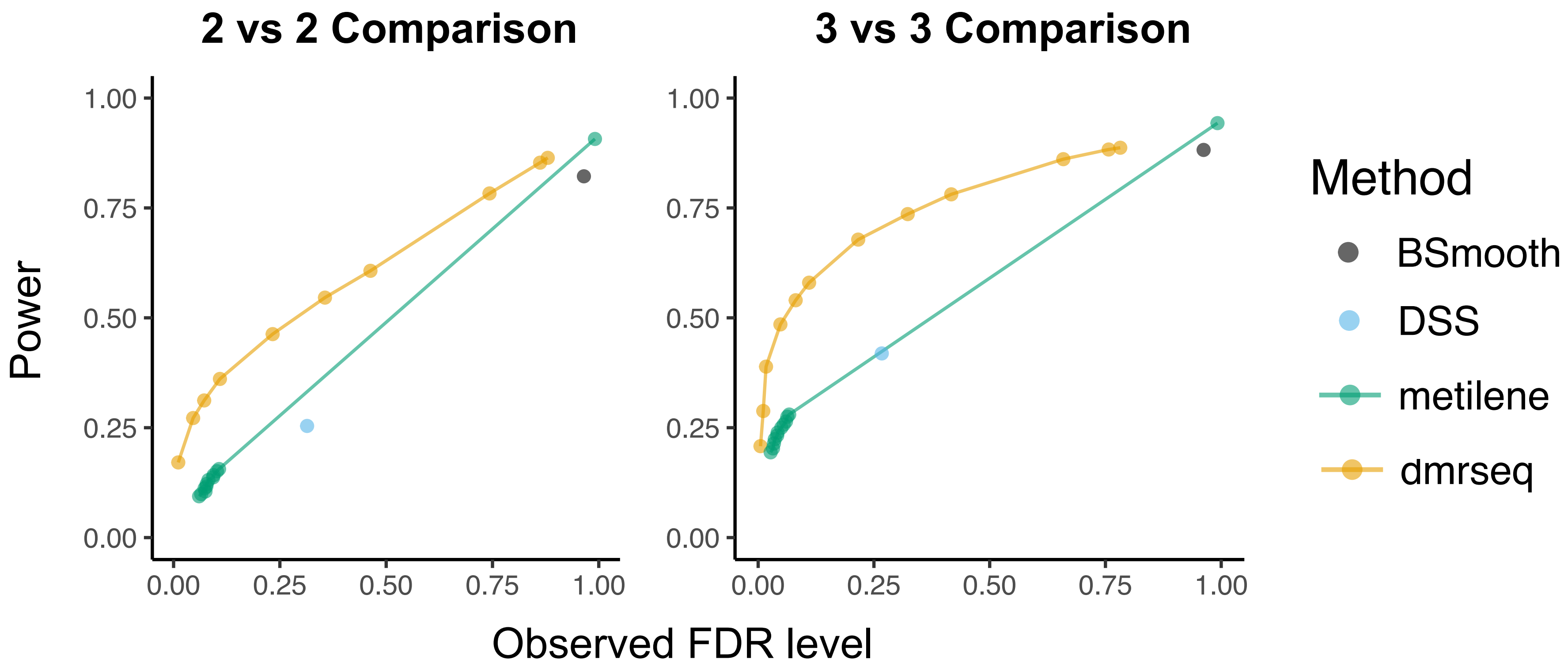


Korthauer et al., 2018

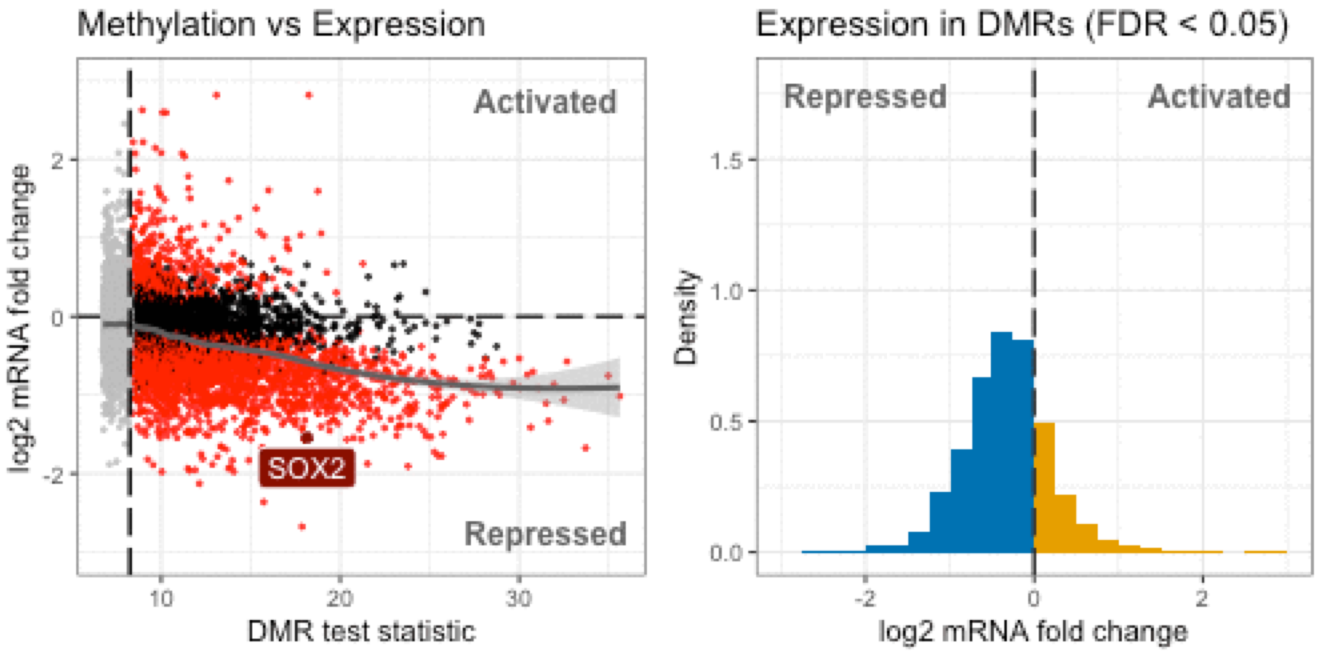
Accurate FDR control in simulation



High sensitivity and specificity in simulation



Significant DMRs enriched for biological signal



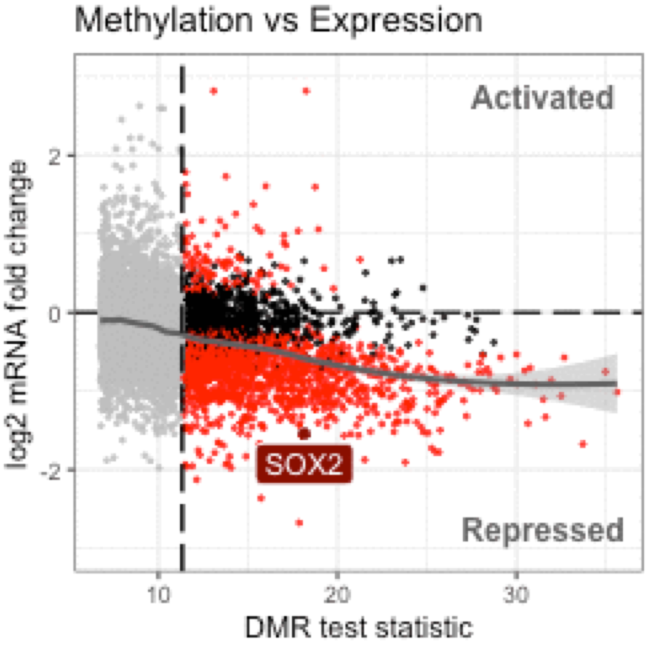
← Expression Decrease

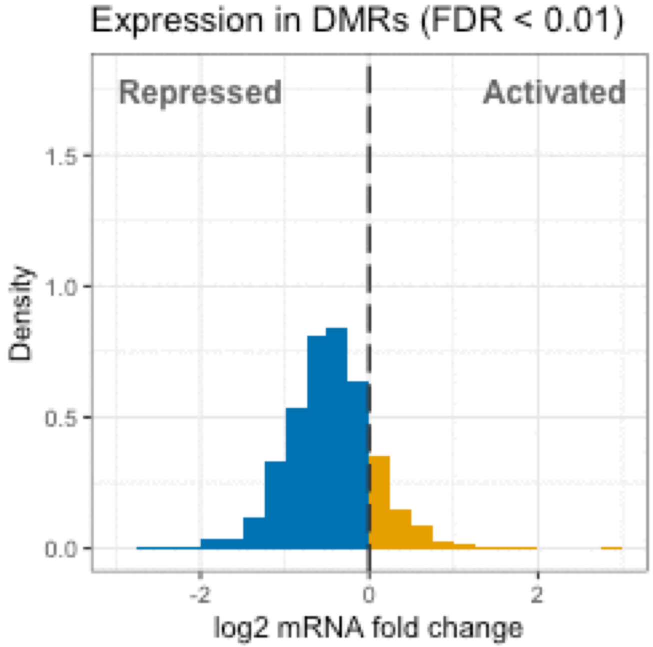
Methylation Increase →

Data from Ford et al., 2017, *bioRxiv*

Significant DMRs enriched for biological signal

↓ Expression Decrease

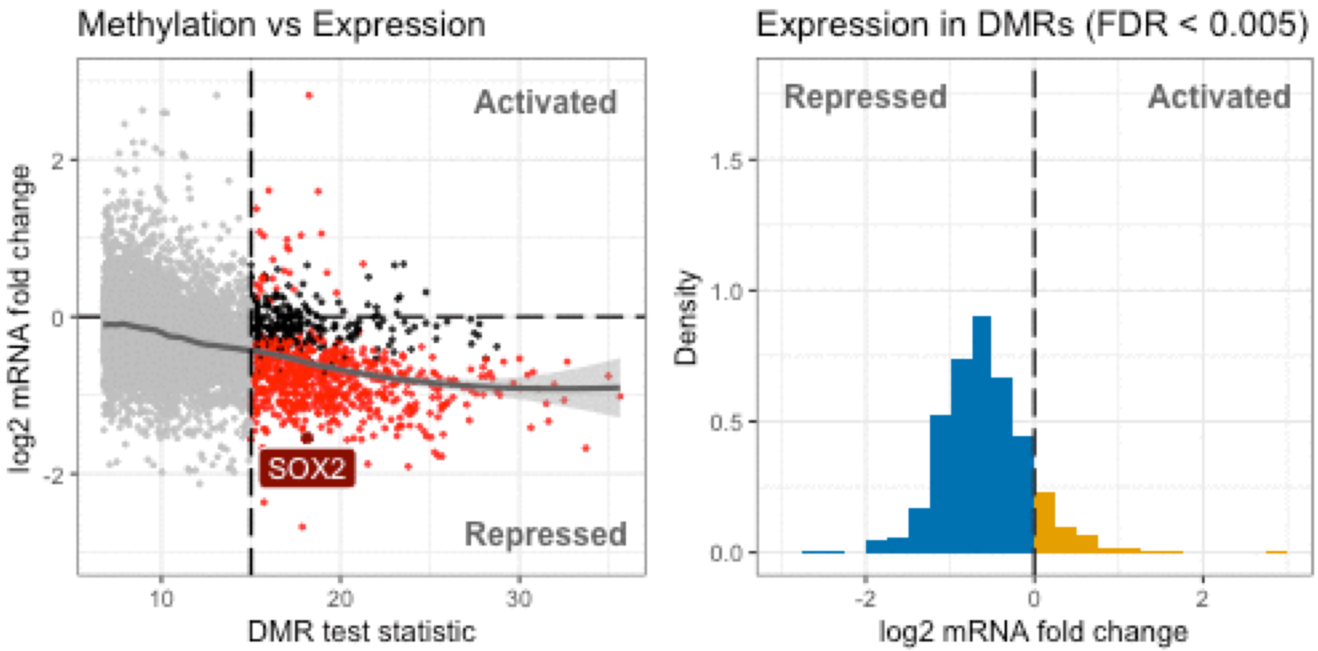




Methylation Increase →

Data from Ford et al., 2017, *bioRxiv*

Significant DMRs enriched for biological signal

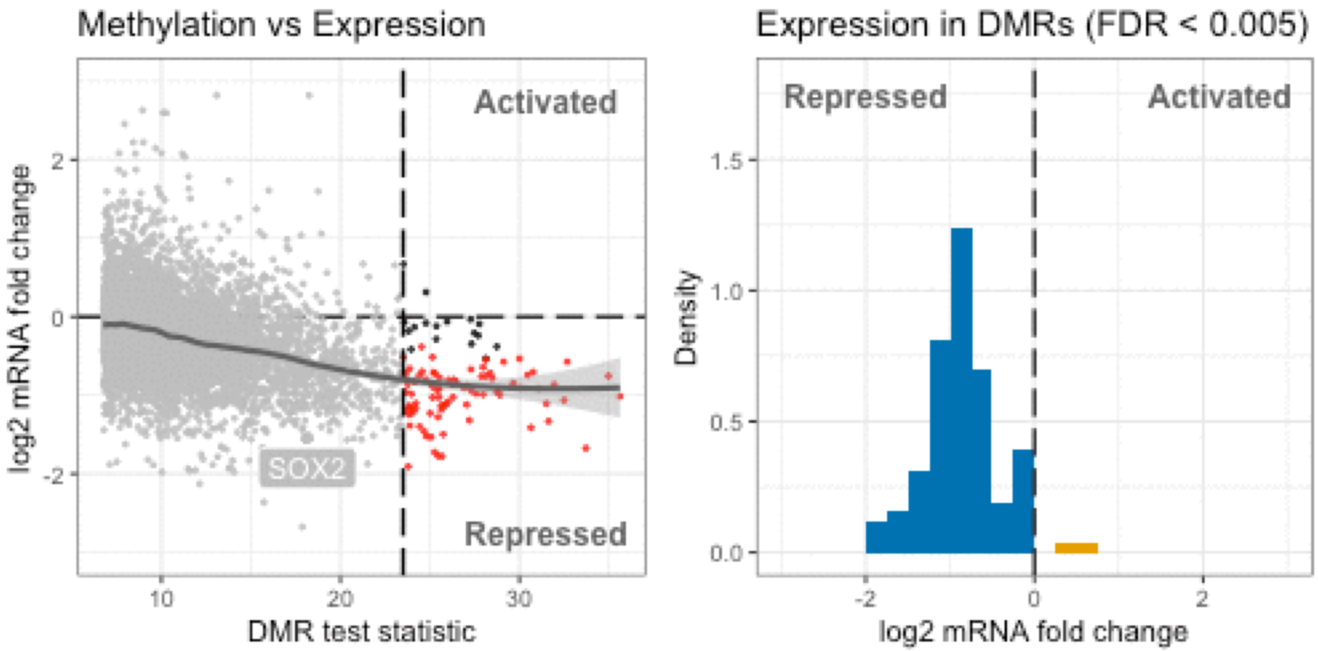


← Expression Decrease

Methylation Increase →

Data from Ford et al., 2017, *bioRxiv*

Significant DMRs enriched for biological signal

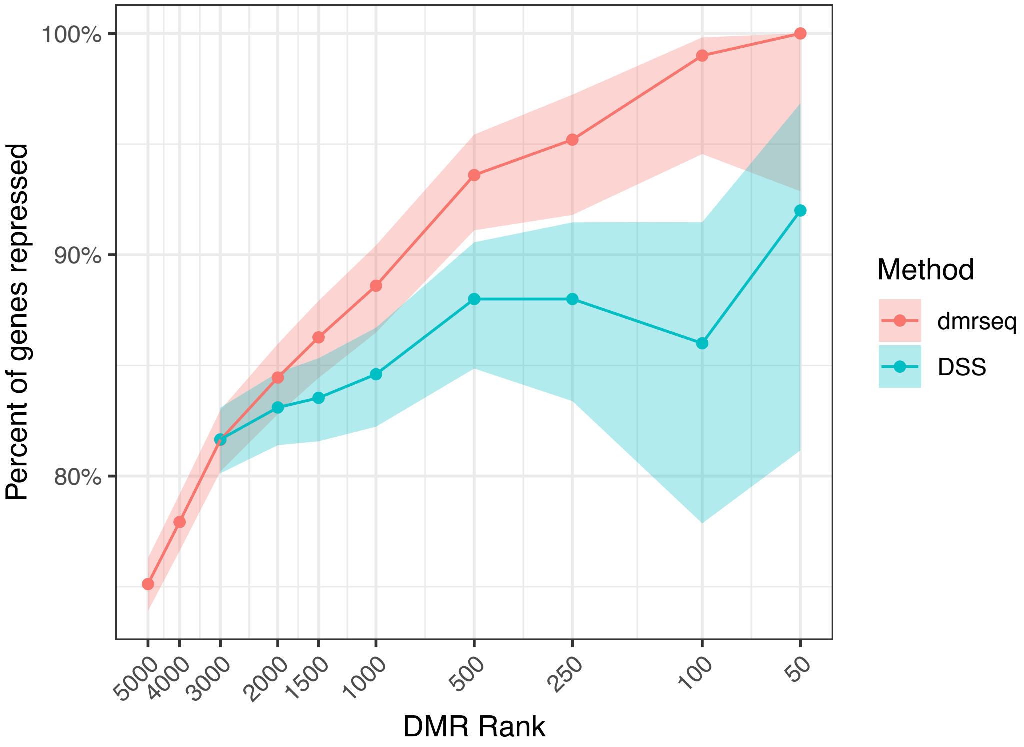


← Expression Decrease

Methylation Increase →

Data from Ford et al., 2017, *bioRxiv*

Increased biological signal in dmrseq DMRs



Data from Ford et al., 2017, *bioRxiv*

Summary

- dmrseq **identifies and prioritizes DMRs** from bisulfite sequencing experiments
- **Models signal at the region level** in order to account for sample and spatial variability
- Achieves **accurate False Discovery Rate control** by generating a null distribution that pools information across the genome
- Detailed in “Detection and accurate False Discovery Rate control of differentially methylated regions from Whole Genome Bisulfite Sequencing” (*Biostatistics,* 2018)
- dmrseq R package available on Bioconductor

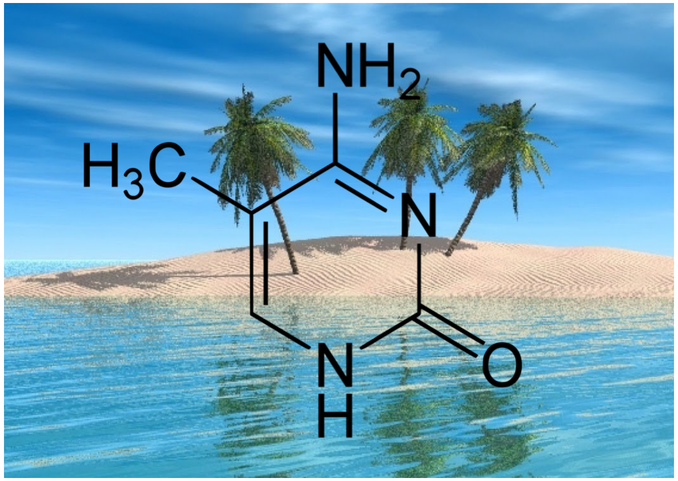
Acknowledgements

<ins>Dana-Farber/Harvard Chan</ins>

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

<ins>Collaborators</ins>

Sutirtha Chakraborty
Yuval Benjamini



Contact

keegan@jimmy.harvard.edu
keegankorthauer
kkorthauer.org