De novo detection and accurate inference of differentially methylated regions

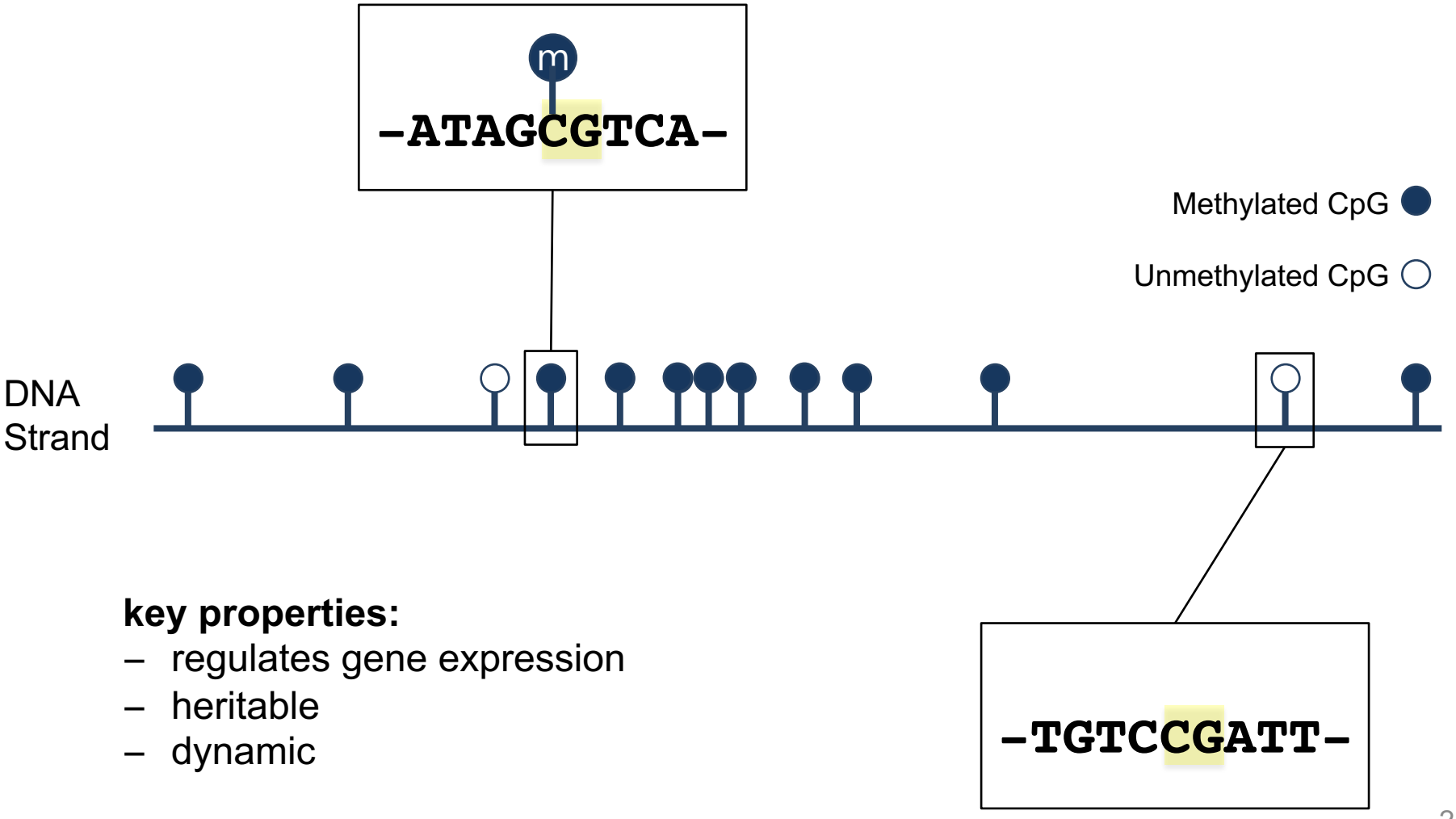
Keegan Korthauer, PhD

AISC, Greensboro, NC
6 October 2018

DANA-FARBER CANCER INSTITUTE

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

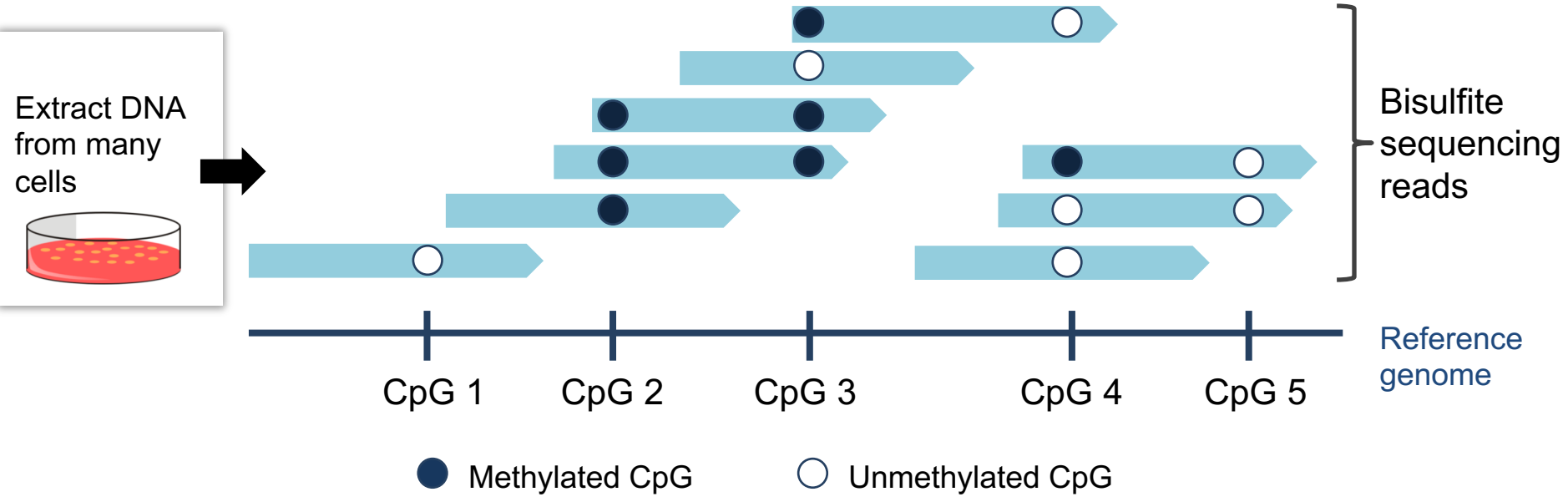
DNA Methylation: The fifth base?



key properties:

- regulates gene expression
- heritable
- dynamic

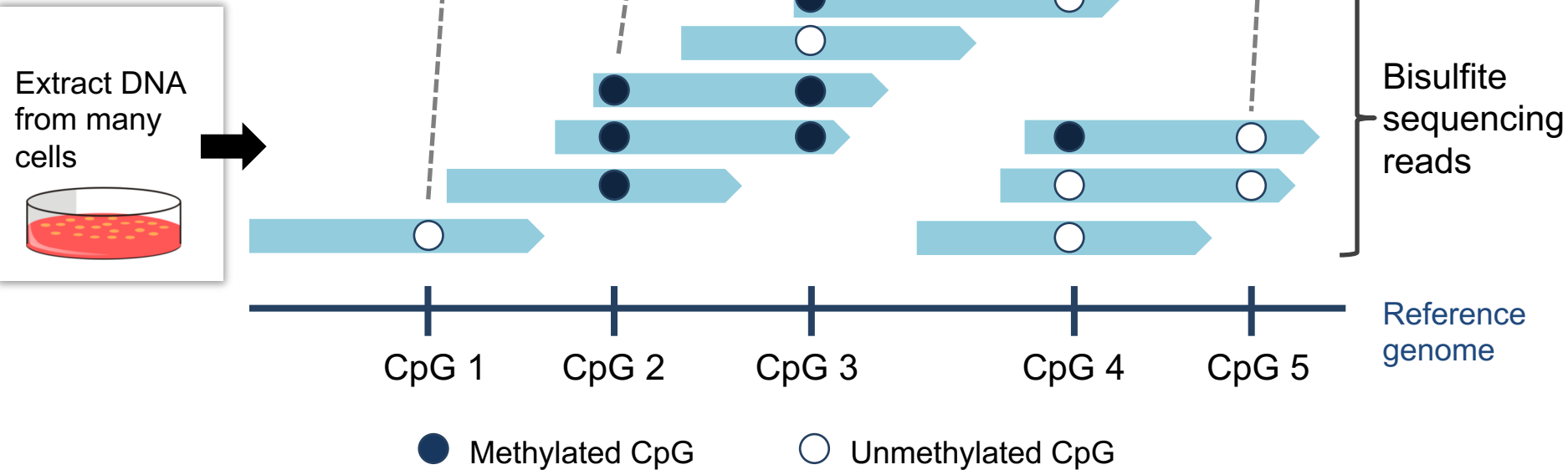
Whole Genome Bisulfite Sequencing (WGBS)



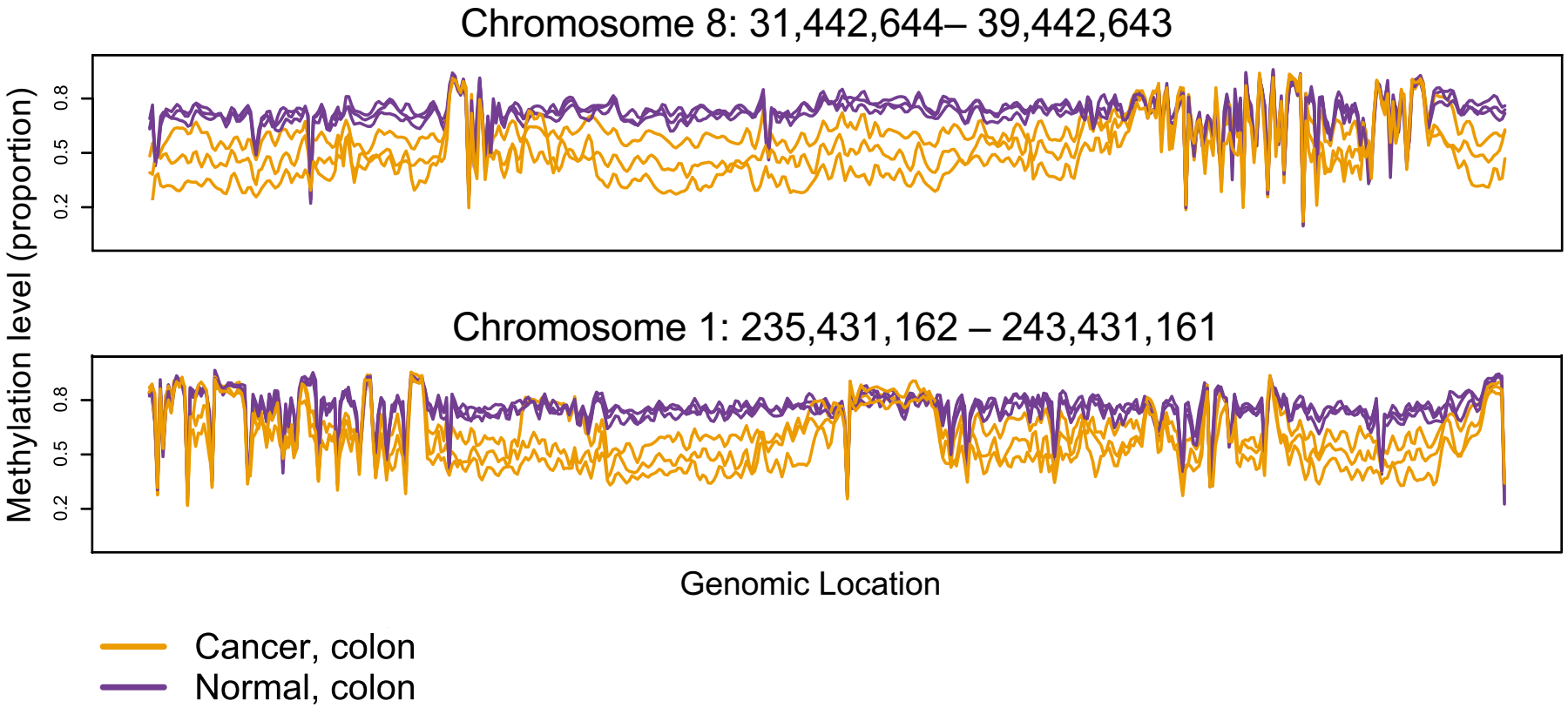
Whole Genome Bisulfite Sequencing (WGBS)

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

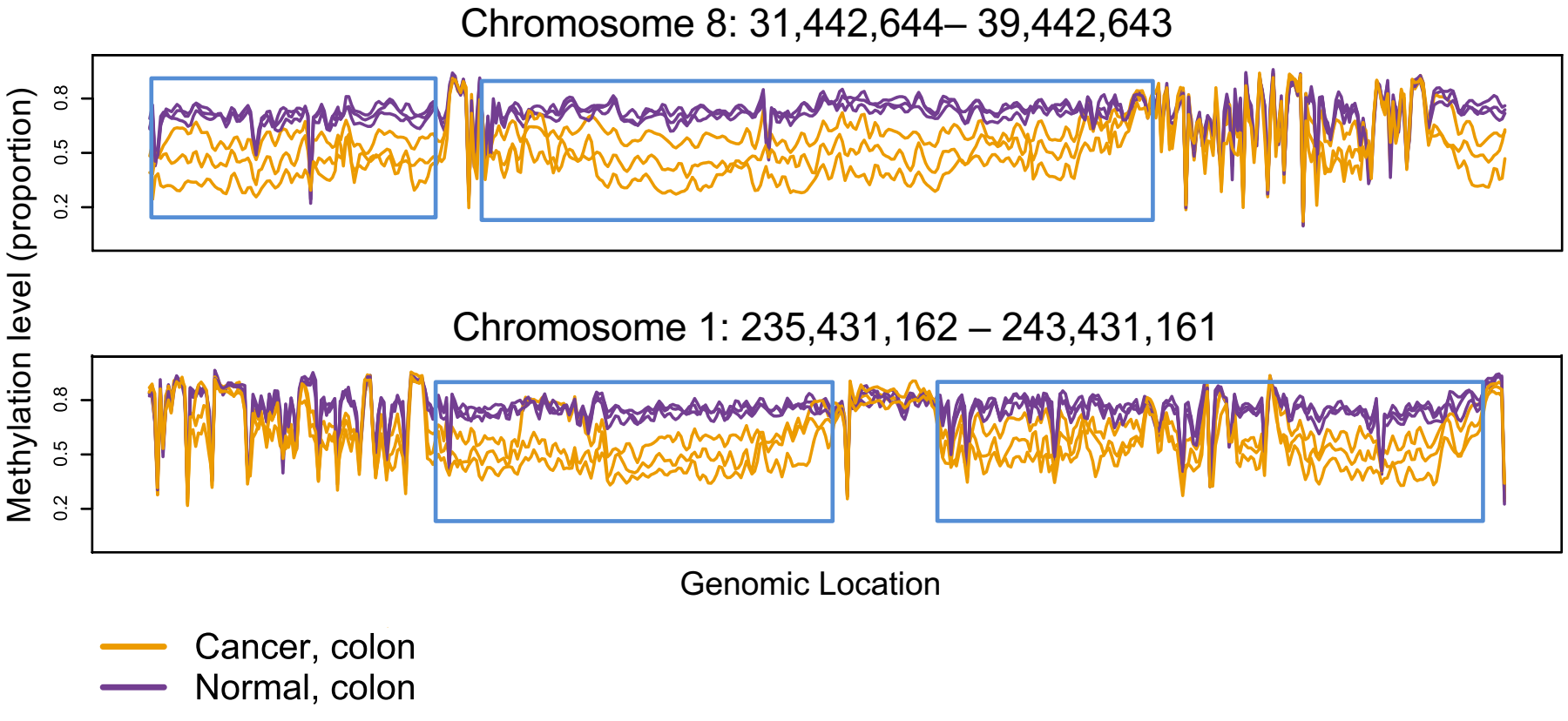


Differentially Methylated Regions (DMRs)



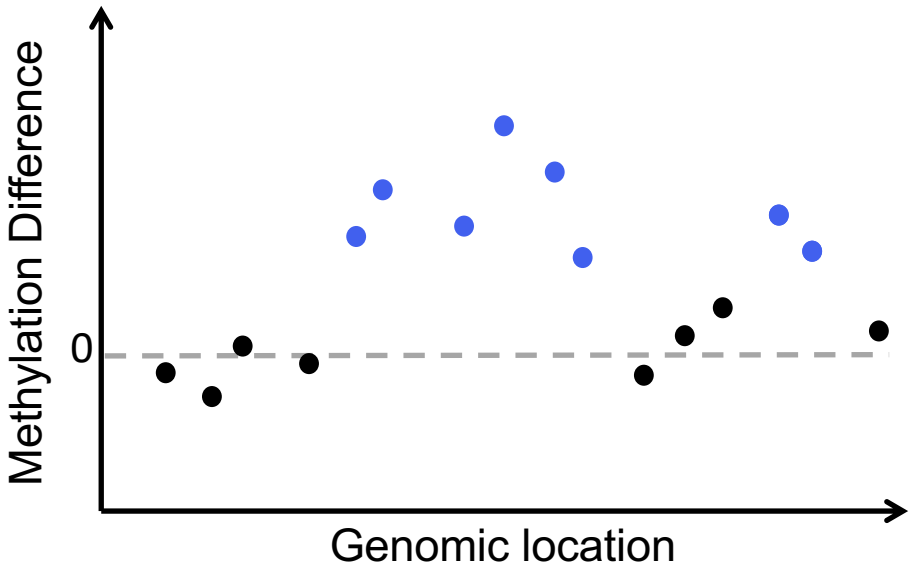
Adapted from Hansen et al. 2014, *Genome Research*

Differentially Methylated Regions (DMRs)



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

Previous methods: Grouping significant CpGs



Results from testing individual CpGs:

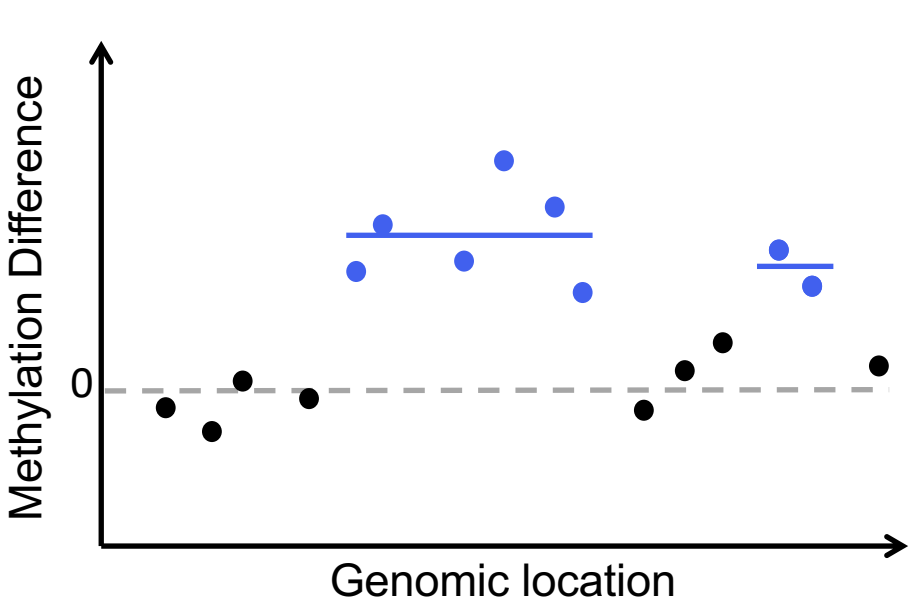
Significant

Not Significant

Examples:

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

Previous methods: Grouping significant CpGs



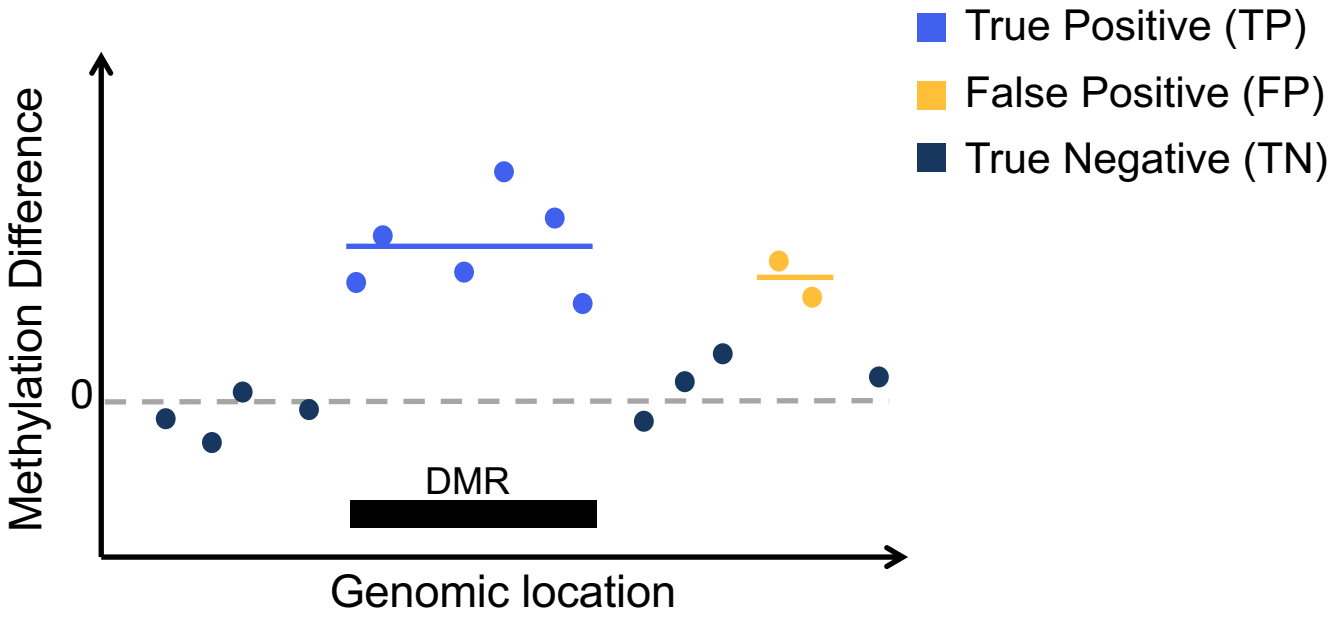
Results from testing individual CpGs:

- Significant
- Not Significant

Examples:

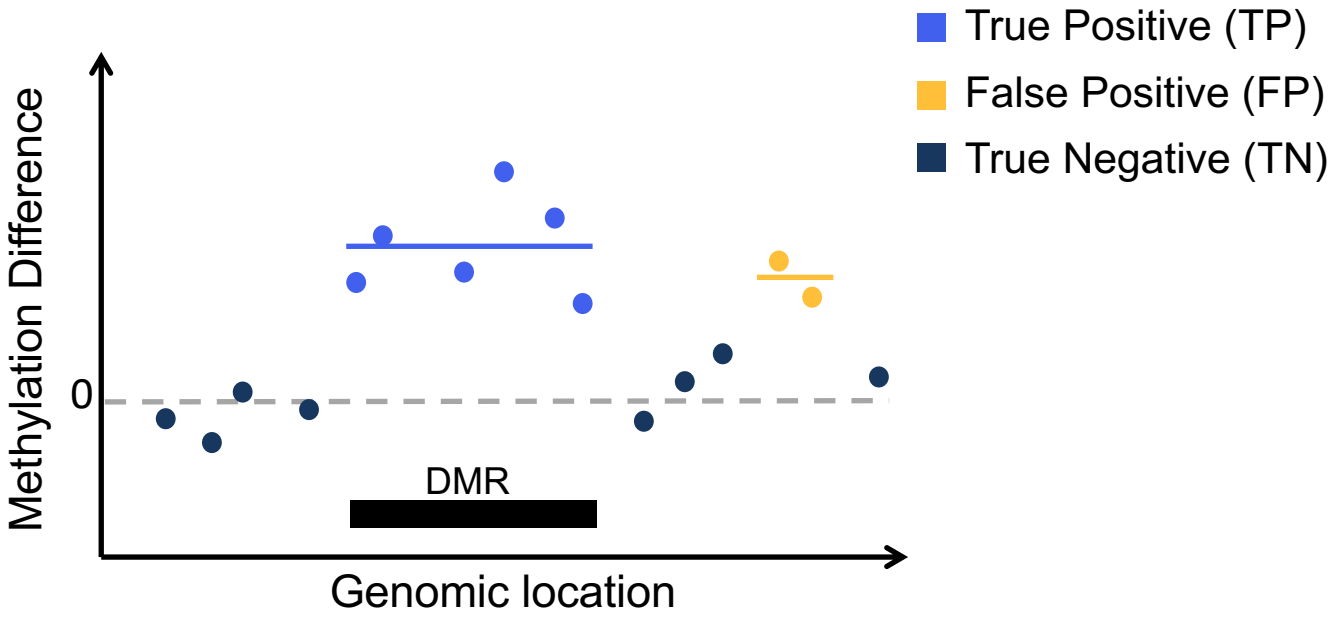
- Bsmooth (Hansen et al., 2012)
- DSS (Feng et al., 2014; Wu et al., 2015)

Error rate not controlled at the region level



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

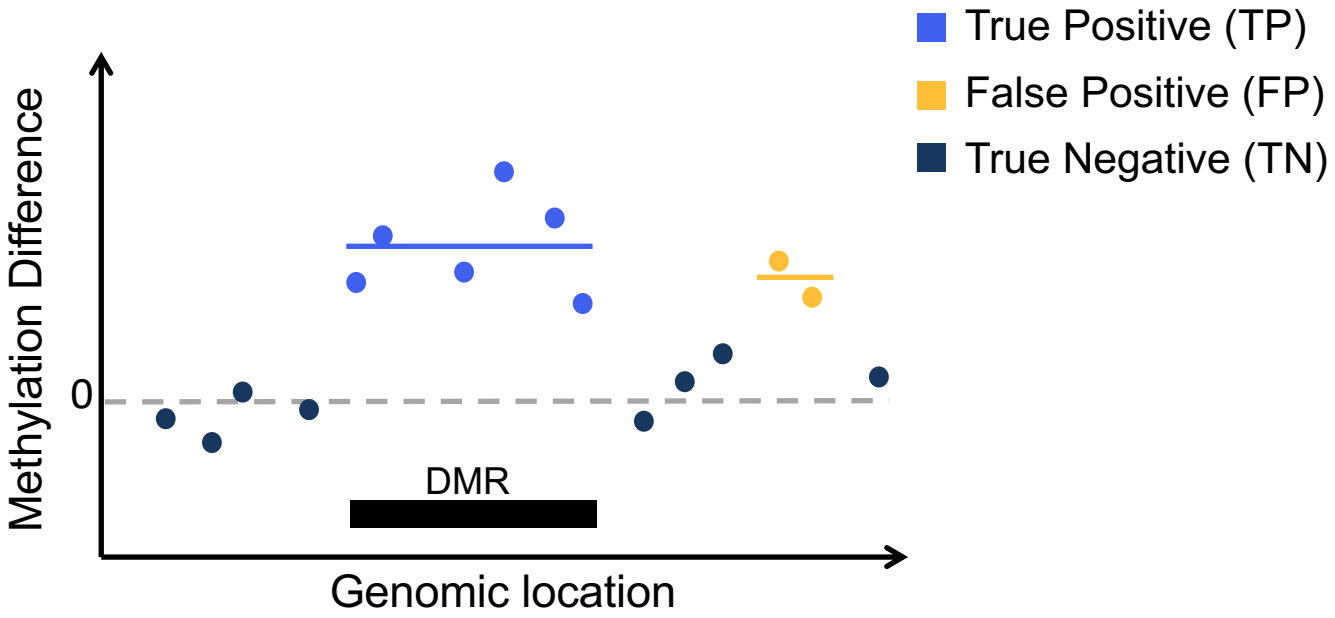
Error rate not controlled at the region level



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

$$\widehat{FDR}_{CpG} = \frac{2}{8} = 0.25$$

Error rate not controlled at the region level

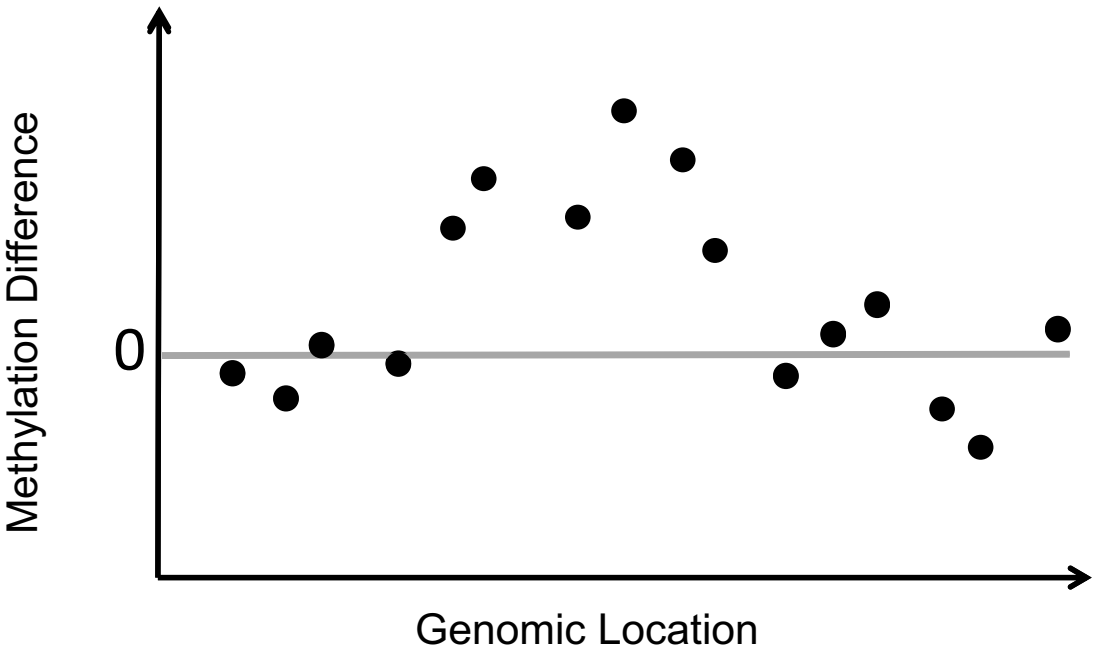


$$\text{False Discovery Rate } (FDR) = E\left[\frac{FP}{TP + FP}\right]$$

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

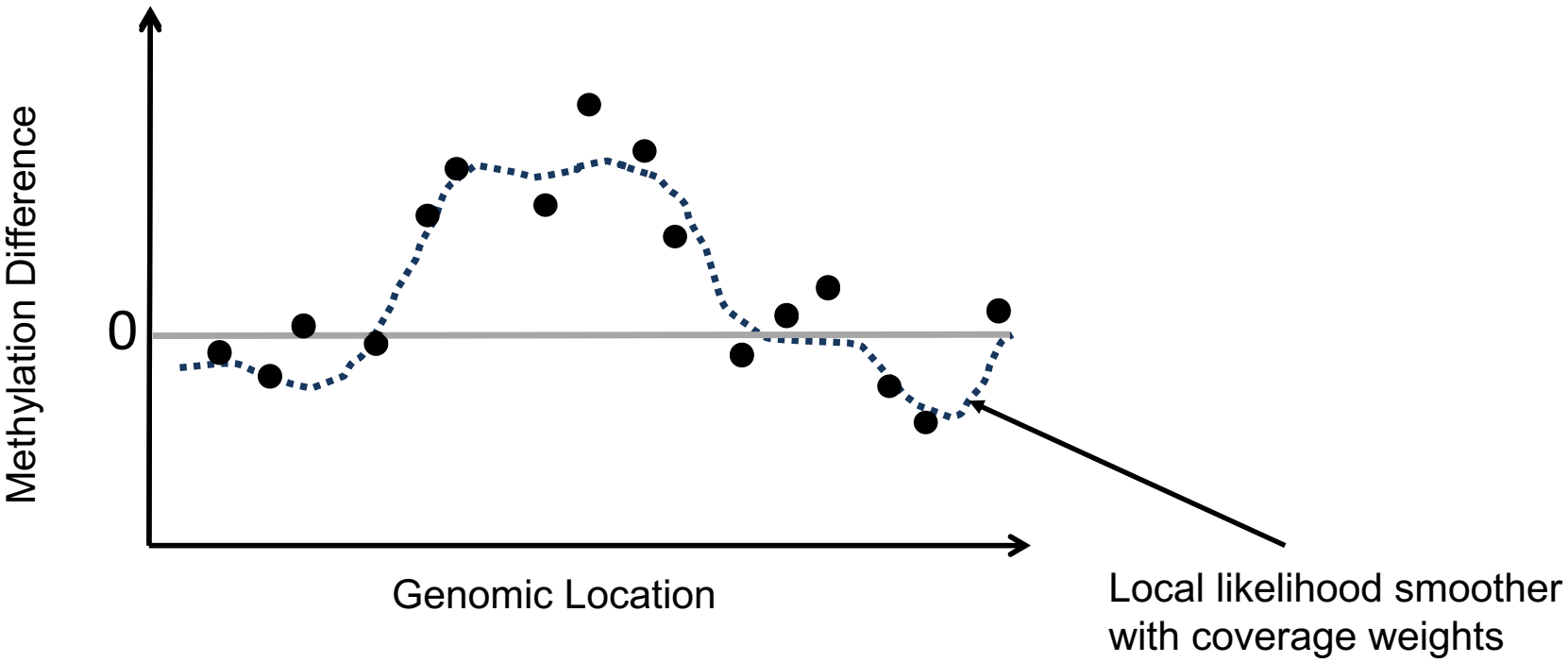
dmrseq: (1) Detect *de novo* candidate regions

Genome-wide scan of CpG methylation difference



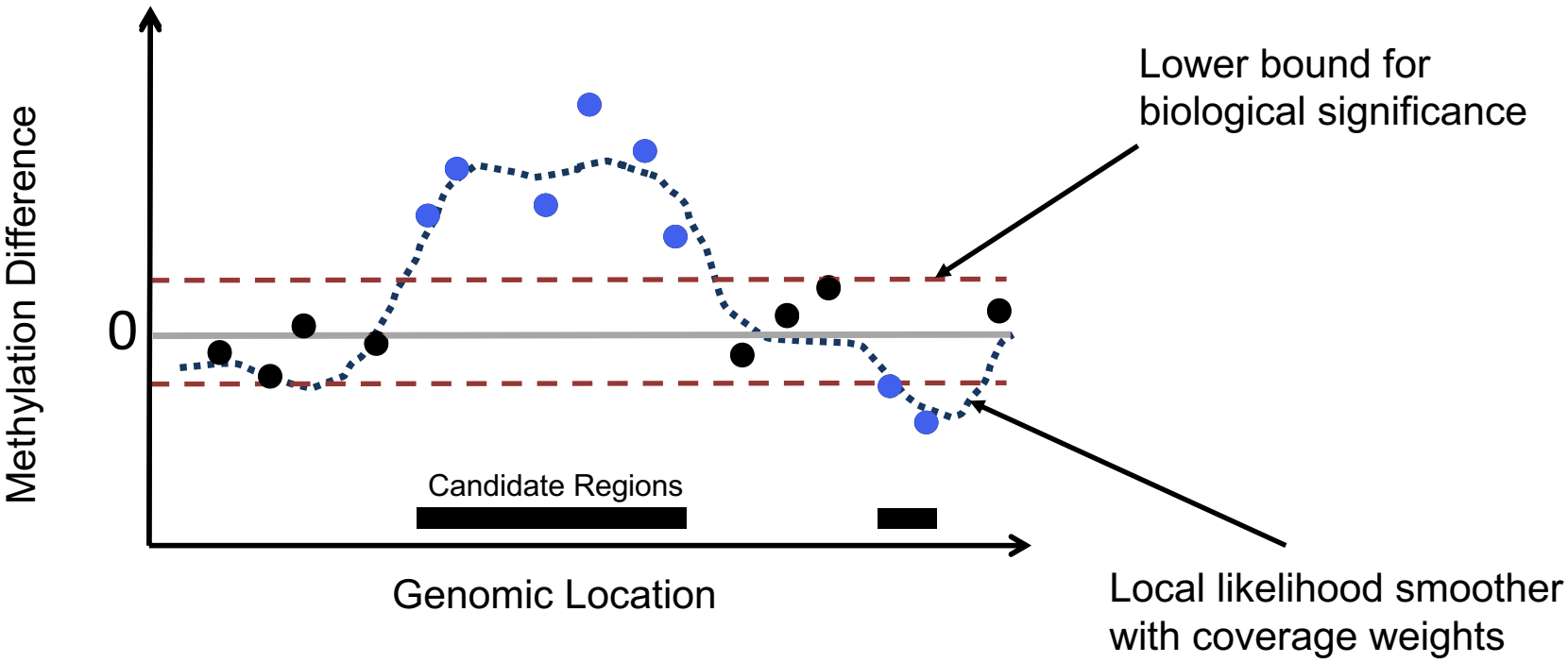
dmrseq: (1) Detect *de novo* candidate regions

Genome-wide scan of CpG methylation difference



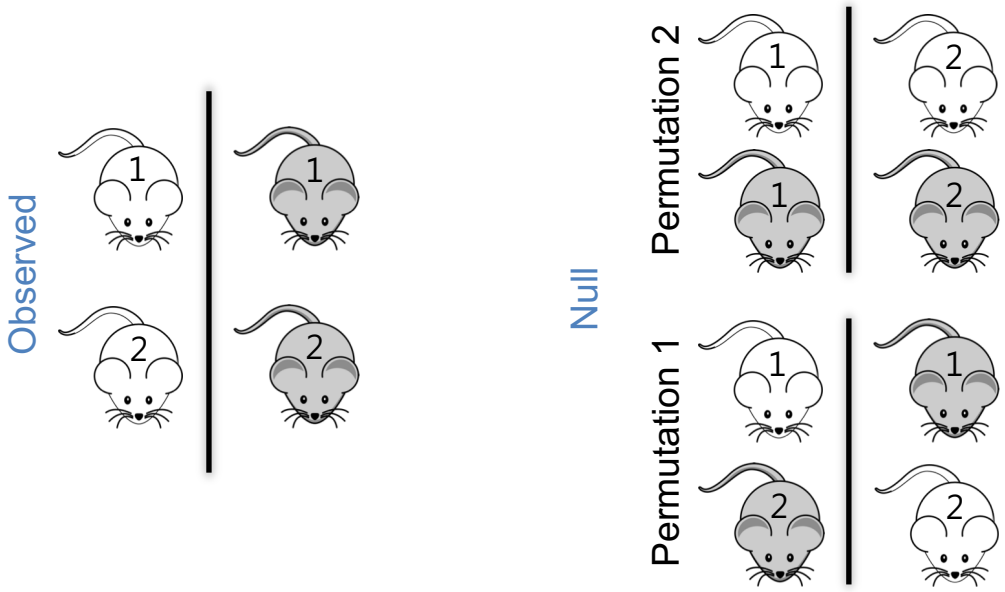
dmrseq: (1) Detect *de novo* candidate regions

Genome-wide scan of CpG methylation difference



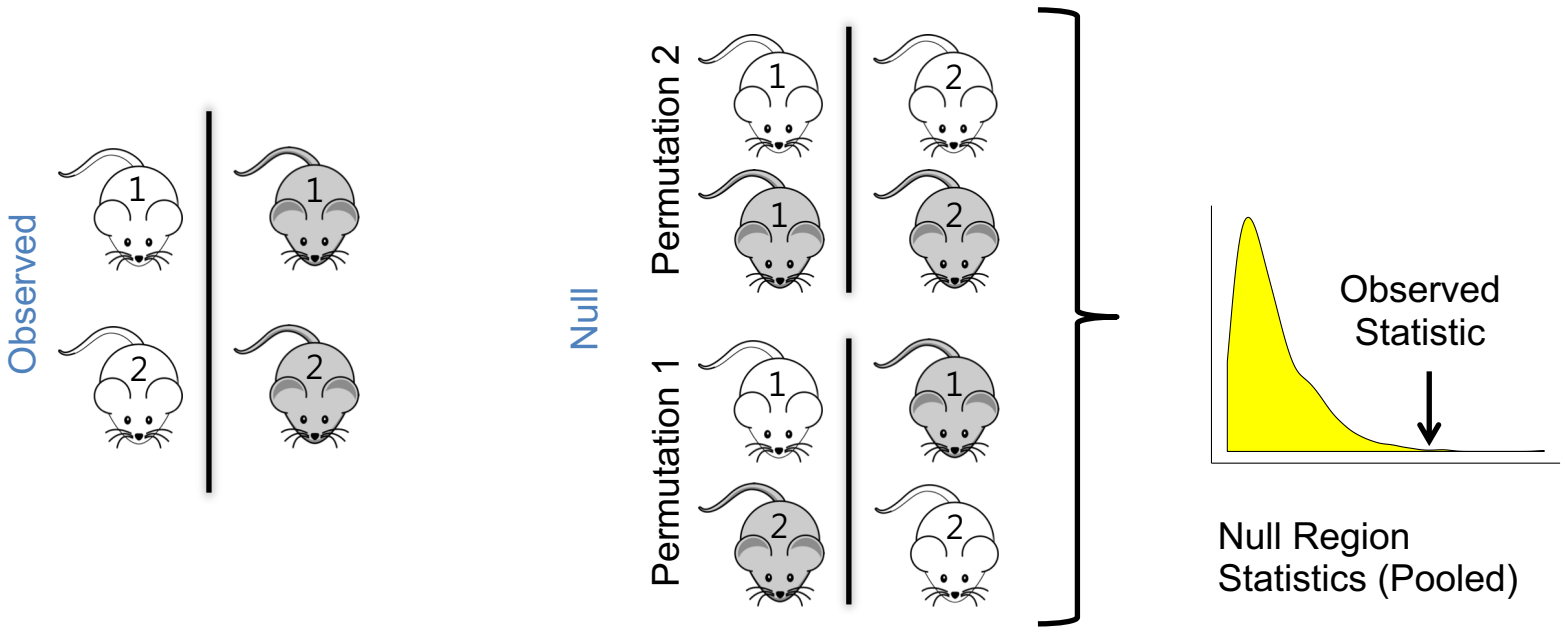
dmrseq: (2) Assess region-level signal

- 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 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 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:

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

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:

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

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:

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

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

Region-level model fitting

Generalized Least Squares (GLS) with variance stabilizing transformation:

arcsine link transformation (Park & Wu 2016)

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

Region-level model fitting

Generalized Least Squares (GLS) with variance stabilizing transformation:

arcsine link transformation (Park & Wu 2016)

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

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

Variance depends on mean

Variance independent of mean

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

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

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

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

Account for variability across samples and locations

(1) Correlation: Continuous Autoregressive (CAR) model

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

t_{ir} = genomic location of CpG i

Account for variability across samples and locations

(1) Correlation: Continuous Autoregressive (CAR) model

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

$t_{ir} =$ genomic location of CpG i

(2) Variability dependent on coverage

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

Account for variability across samples and locations

(1) Correlation: Continuous Autoregressive (CAR) model

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

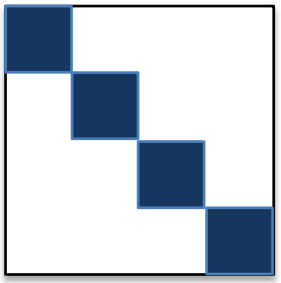
t_{ir} = genomic location of CpG i

(2) Variability dependent on coverage

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

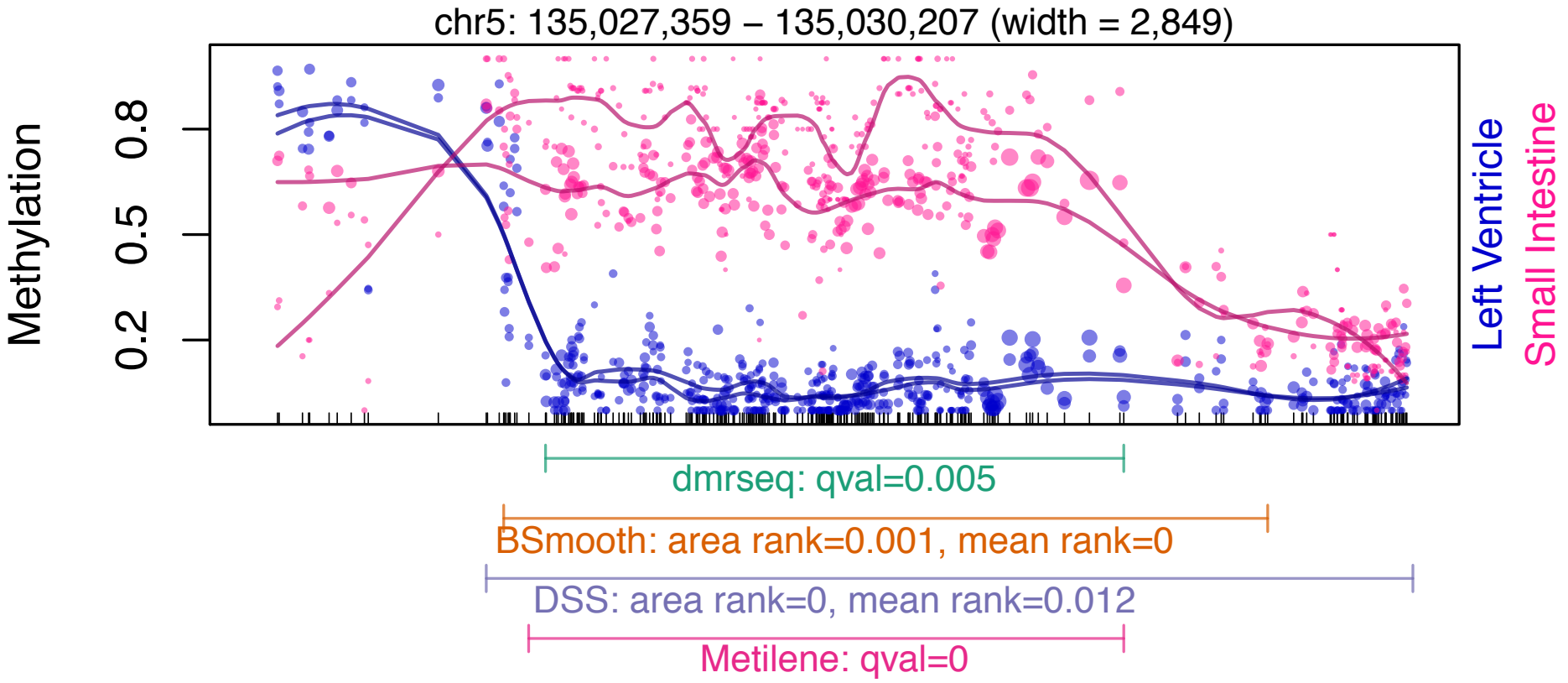
(3) Within sample correlation

Independent samples



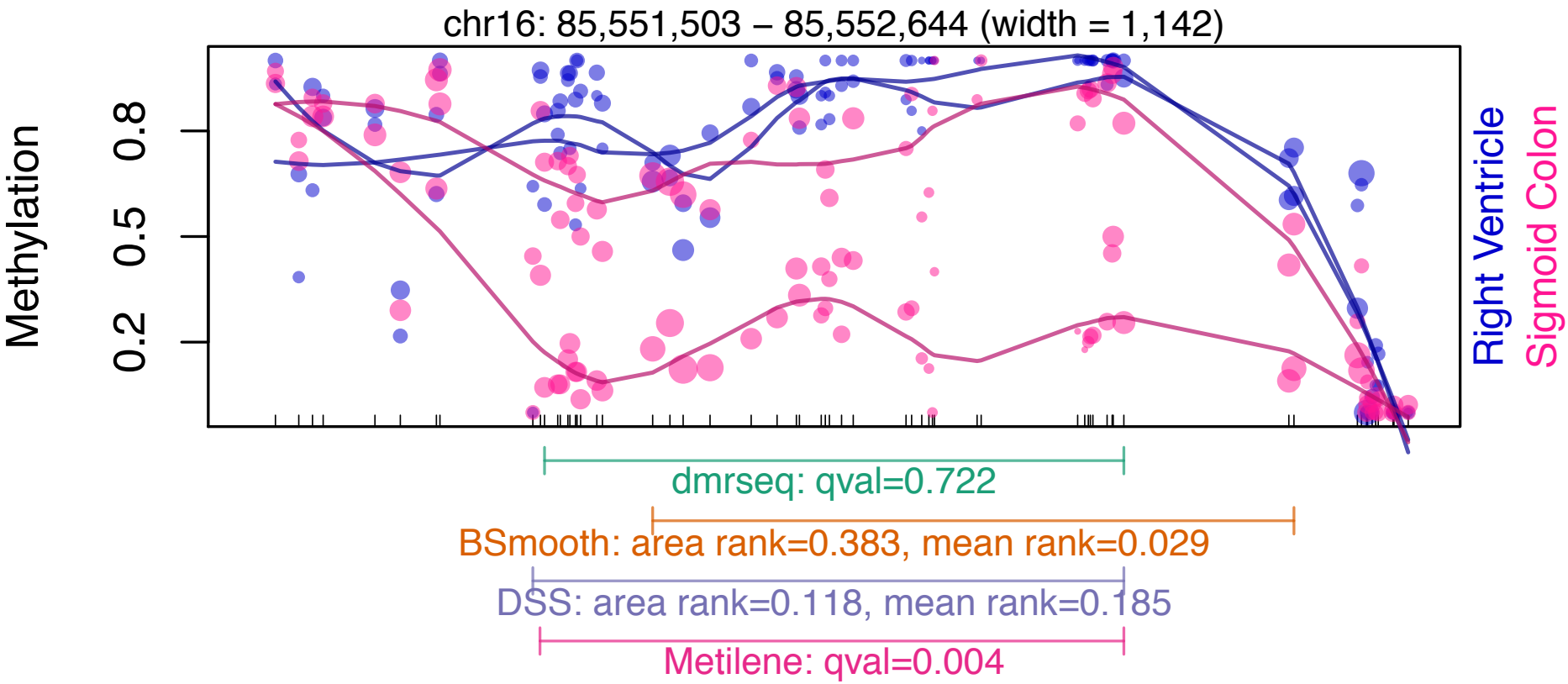
$$Cov\left(Z_{ijr}, Z_{ij^* r}\right) = 0$$

Example: highly ranked DMR across all methods



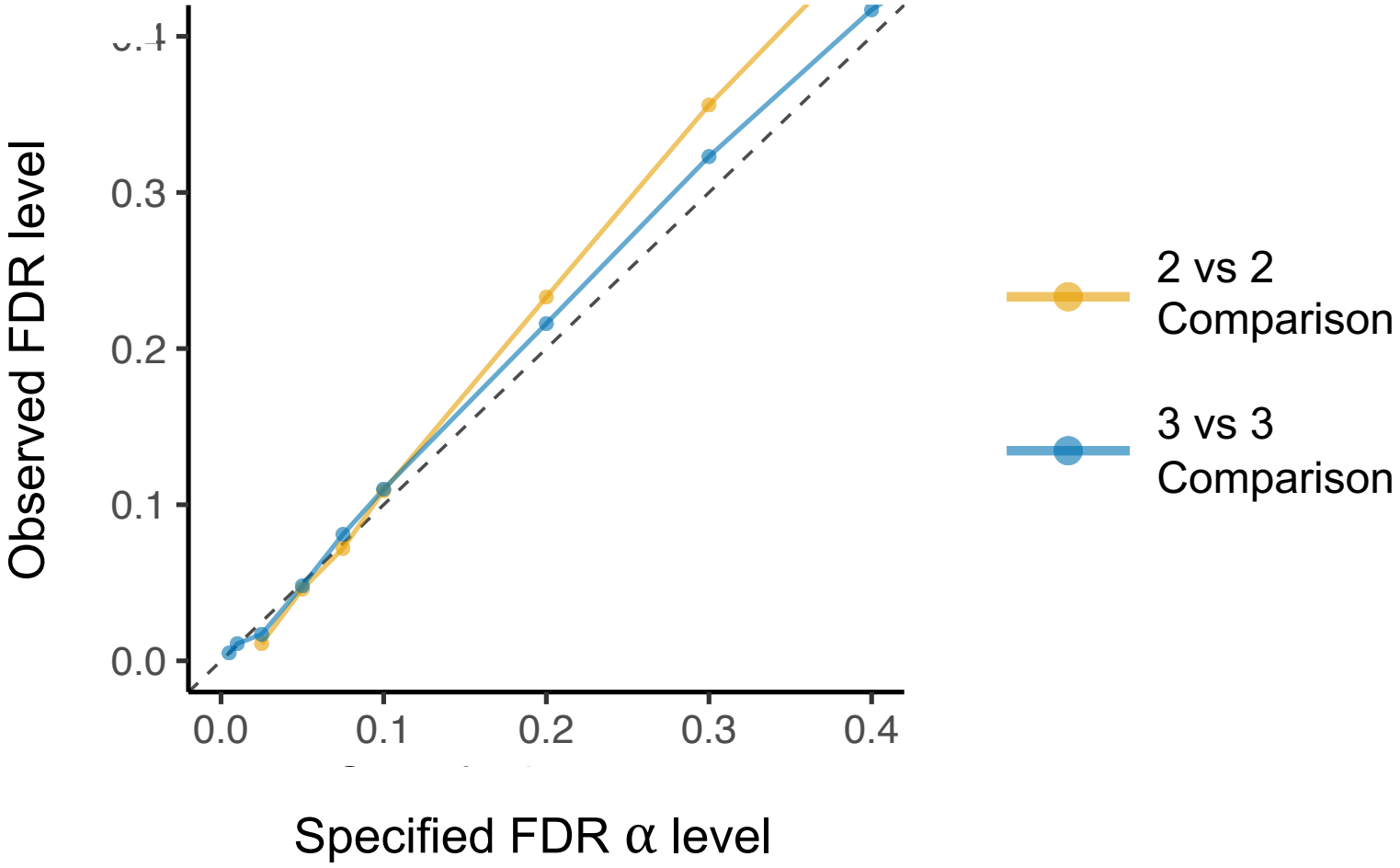
Korthauer et al., 2018 (*Biostatistics*)

Example: dmrseq accounts for sample variability



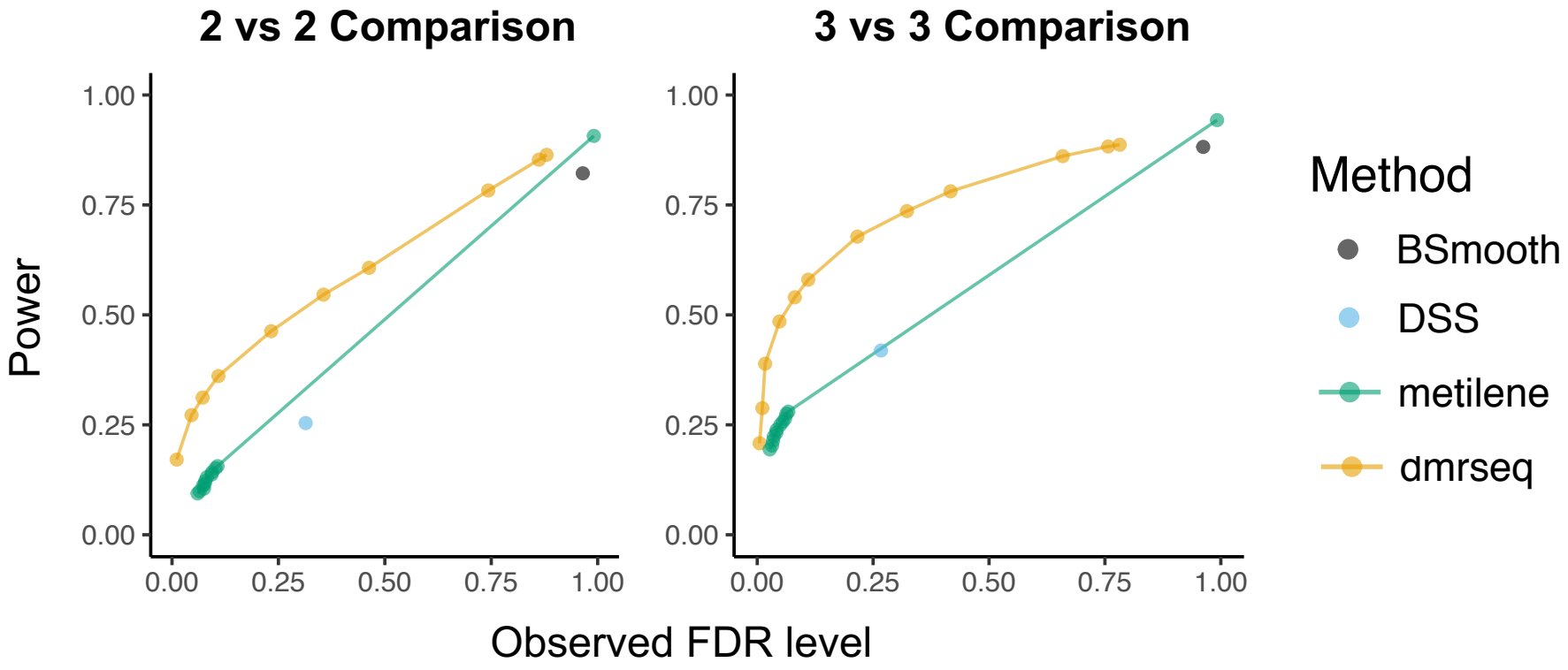
Korthauer et al., 2018 (*Biostatistics*)

Accurate FDR control in simulation



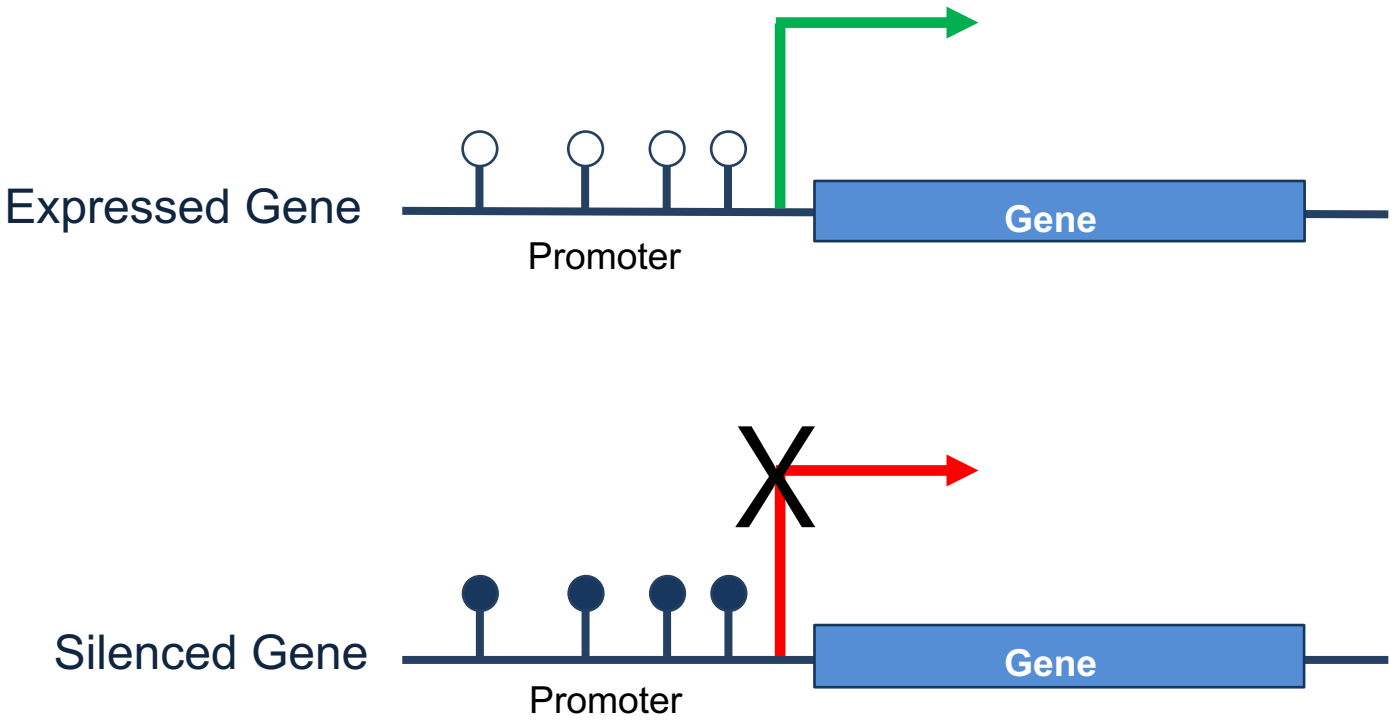
Korthauer et al., 2018 (*Biostatistics*)

High sensitivity and specificity in simulation



Korthauer et al., 2018 (*Biostatistics*)

Methylation is a transcriptional silencing mark



Landmark study finds little influence of methylation on expression

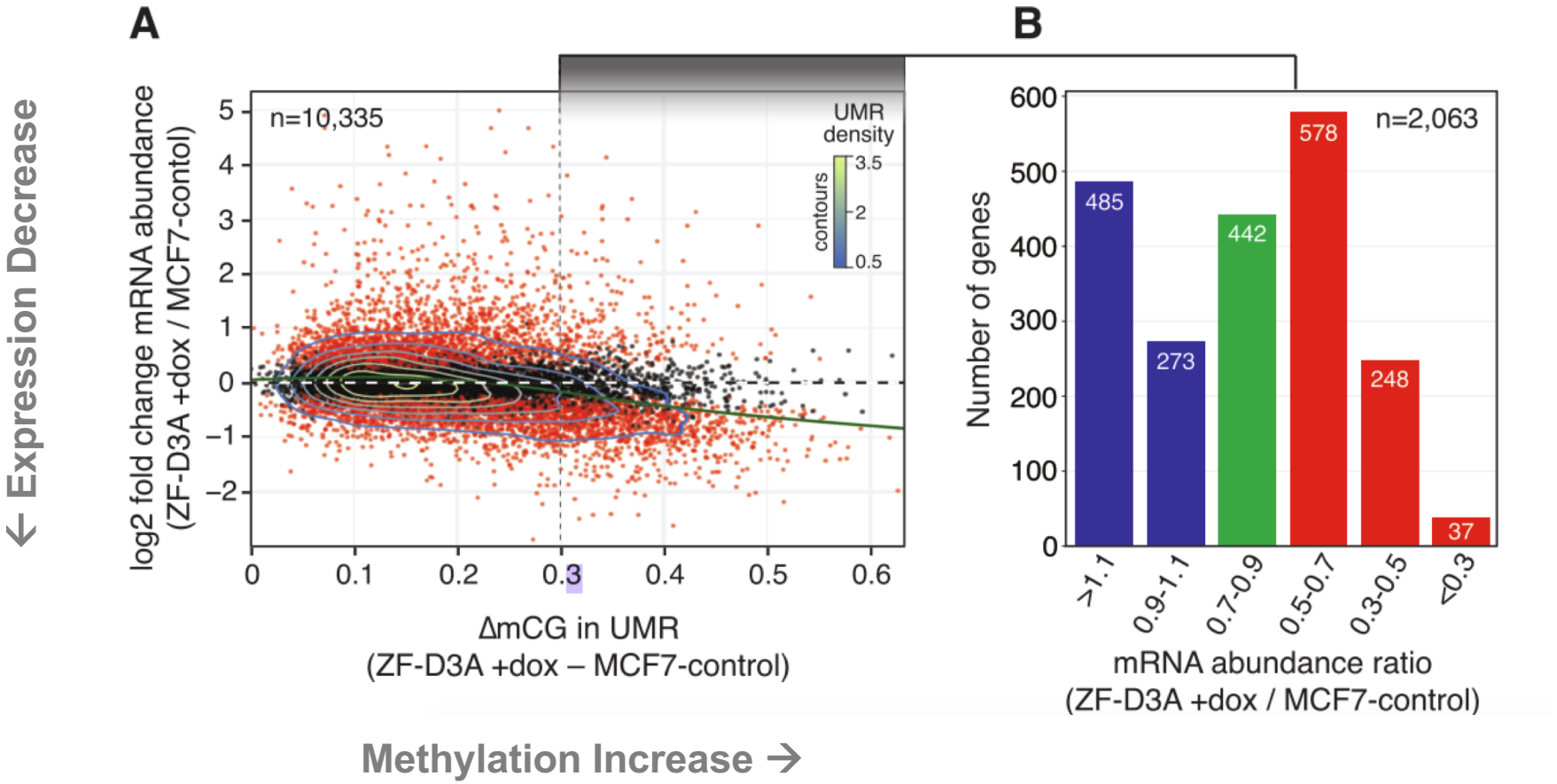
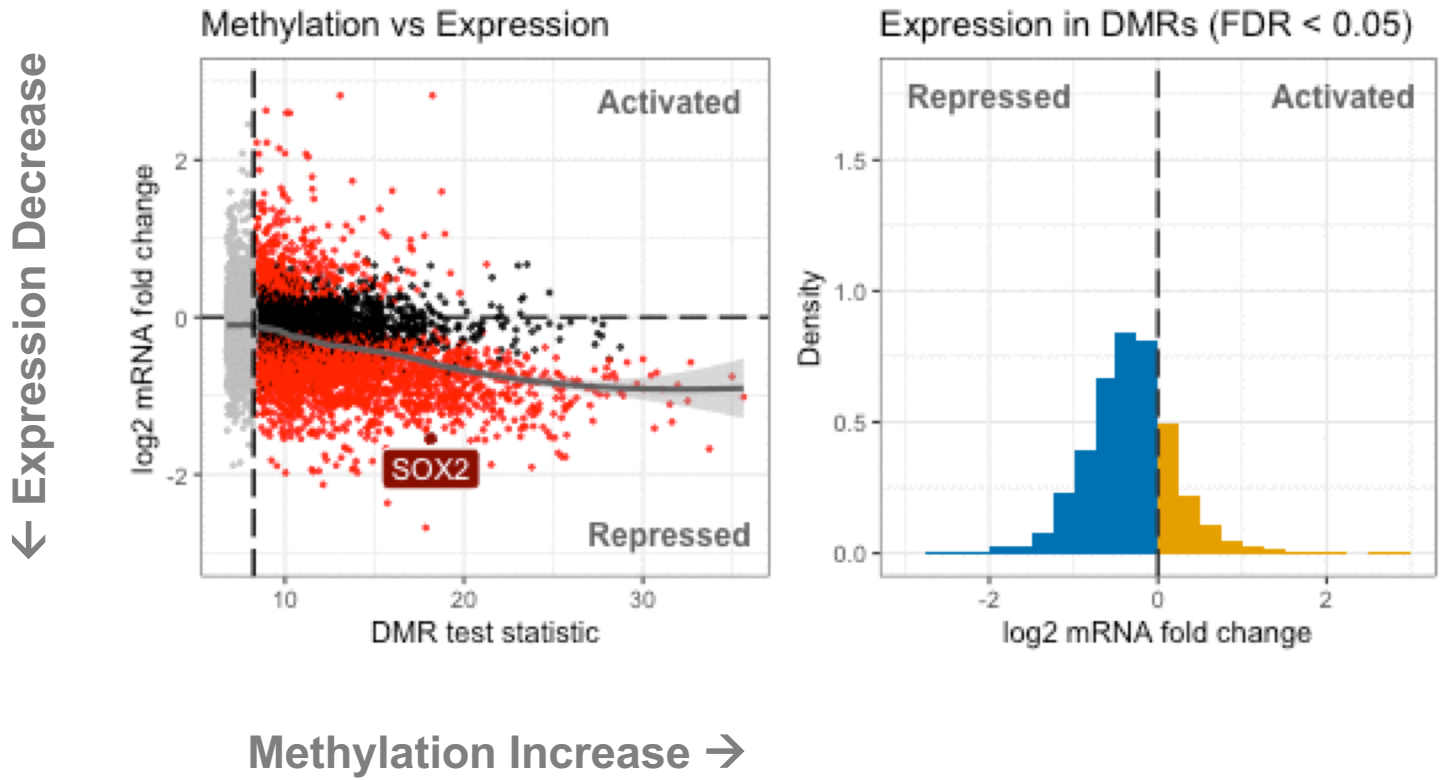


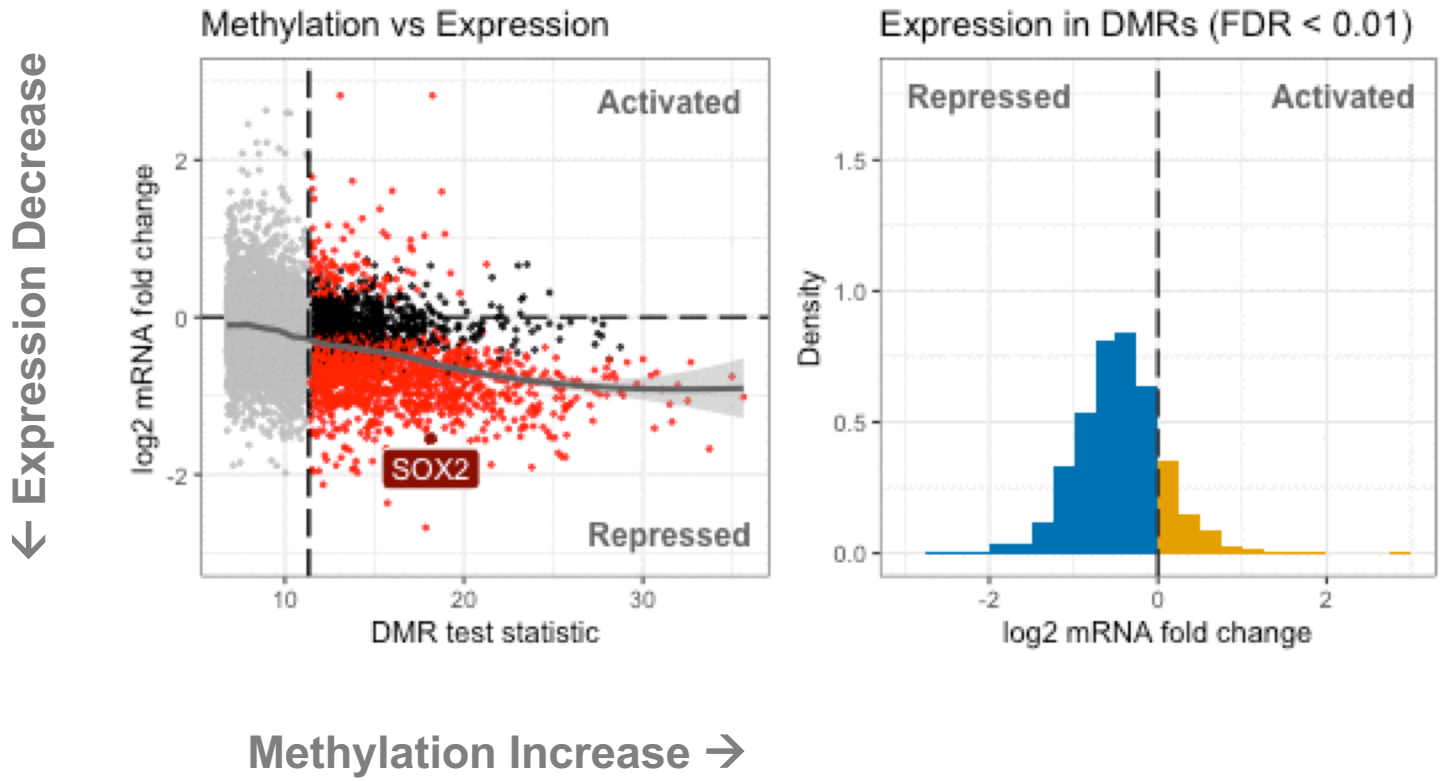
Figure 5, Ford et al., 2017 (*bioRxiv*)

Reanalysis reveals DMRs enriched for biological signal



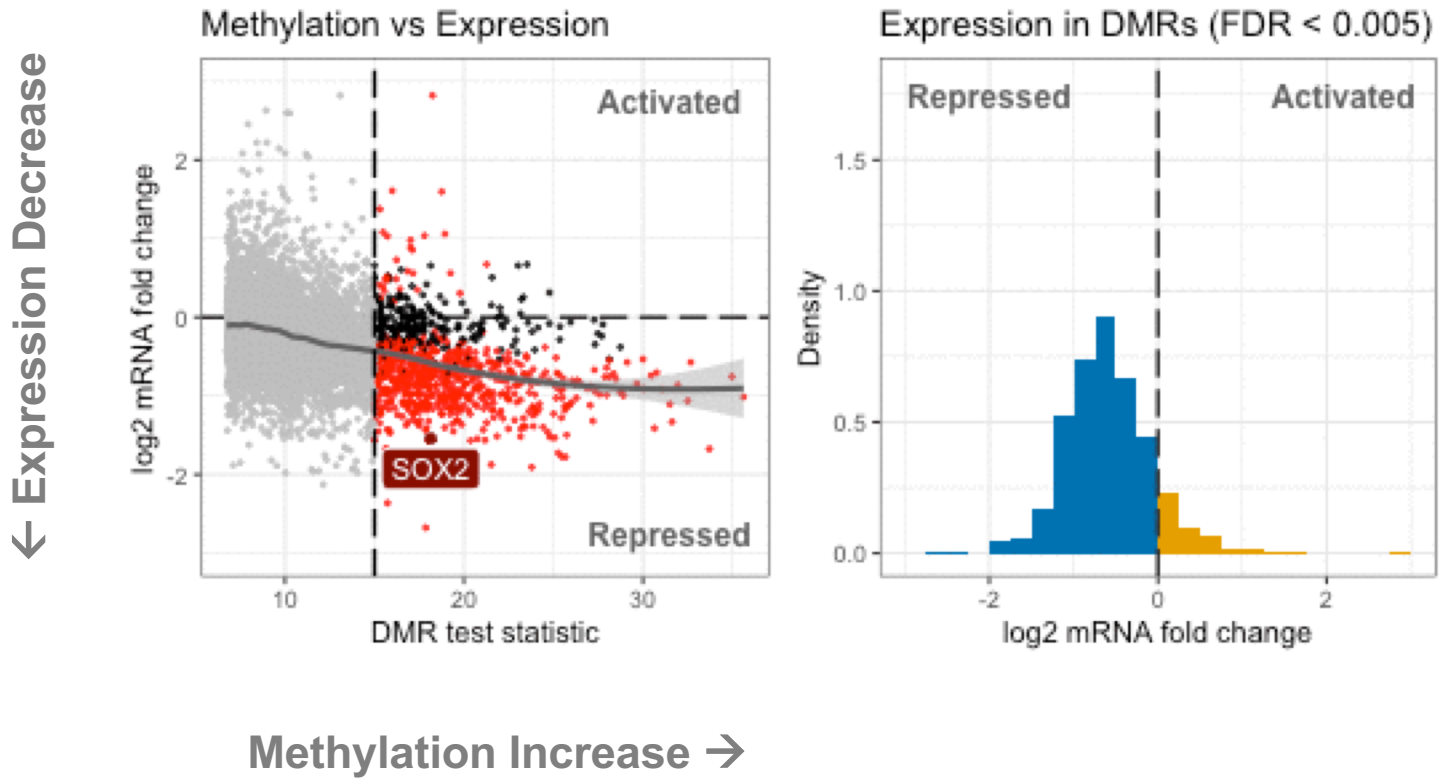
Korthauer & Irizarry, 2018 (*bioRxiv*)

Reanalysis reveals DMRs enriched for biological signal



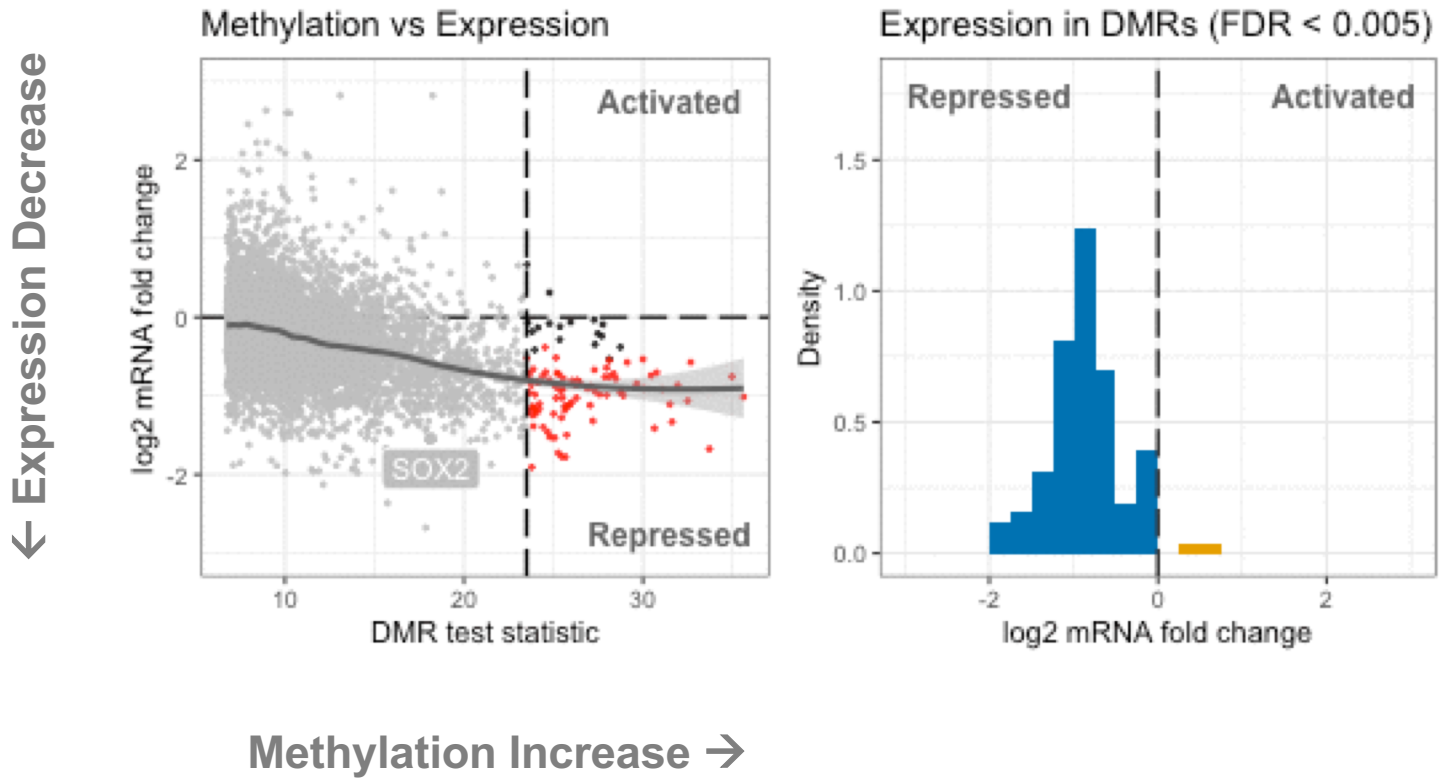
Korthauer & Irizarry, 2018 (*bioRxiv*)

Reanalysis reveals DMRs enriched for biological signal



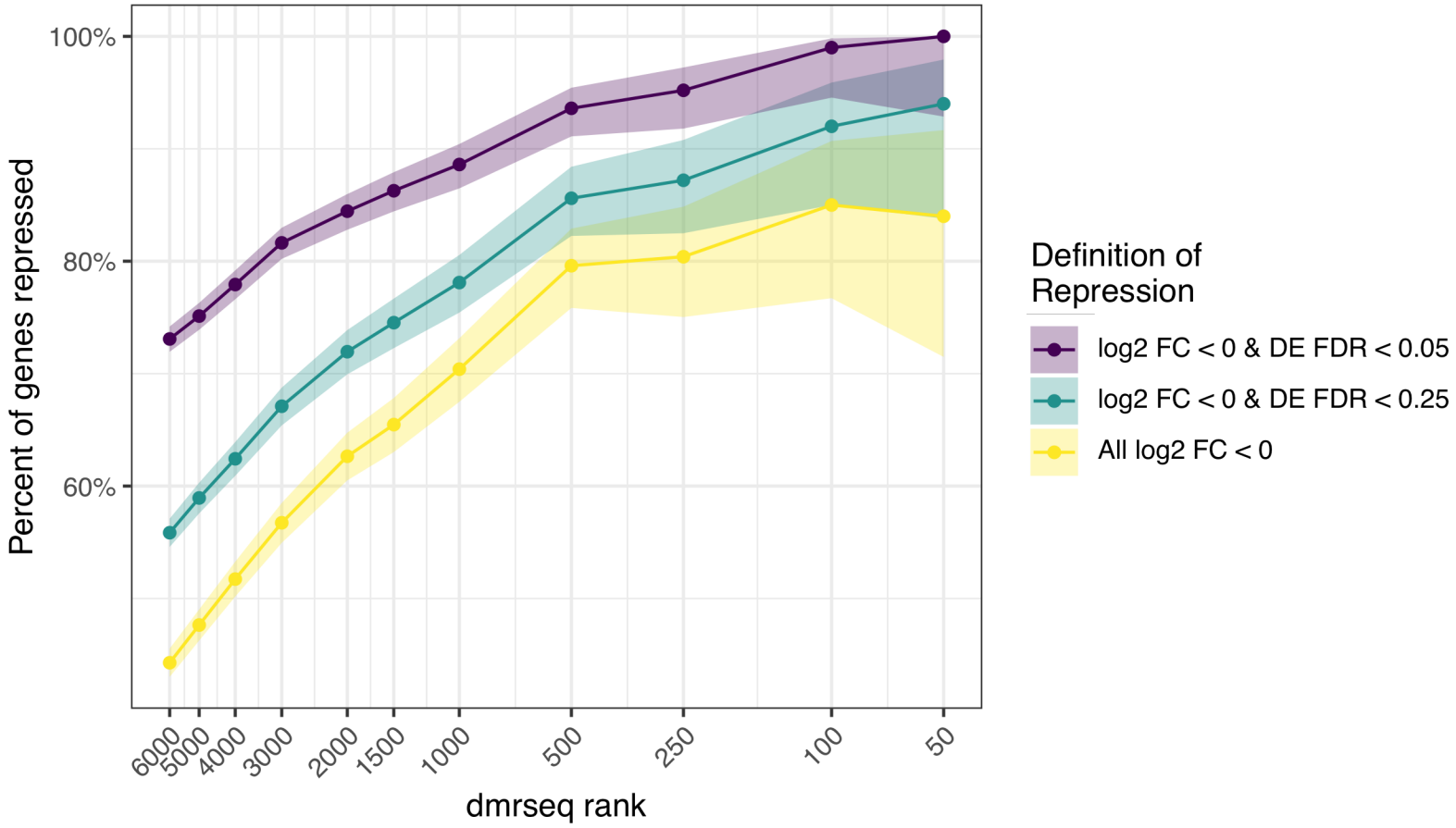
Korthauer & Irizarry, 2018 (*bioRxiv*)

Reanalysis reveals DMRs enriched for biological signal



Korthauer & Irizarry, 2018 (*bioRxiv*)

Enrichment increases with significance level



Korthauer & Irizarry, 2018 (*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
  - Reveals the expected link between DNA methylation and gene expression in the reanalysis of a landmark study
- Learn more:
  - Methodology detailed in Korthauer et al., 2018 (*Biostatistics*)
  - Reanalysis of Ford study detailed in Korthauer & Irizarry, 2018 (*bioRxiv*)
  - R package **dmrseq** available on Bioconductor

Acknowledgements

Dana-Farber/Harvard Chan

Rafael Irizarry

Claire Duvallet

Stephanie Hicks

Patrick Kimes

Yered Pita-Juarez

Alejandro Reyes

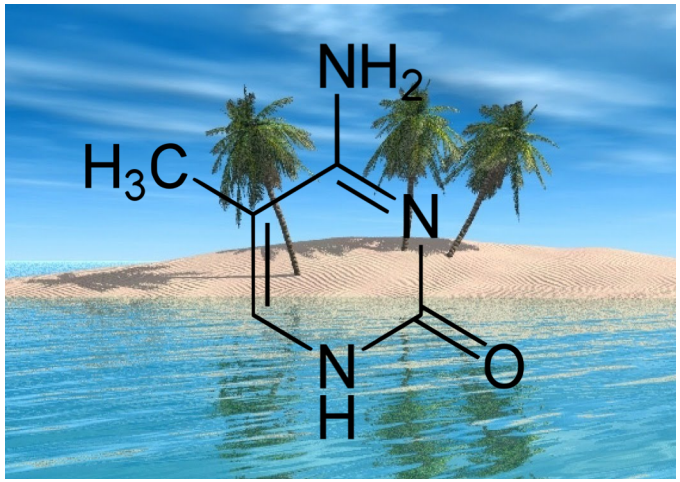
Chinmay Shukla

Mingxiang Teng

Collaborators

Sutirtha Chakraborty

Yuval Benjamini



Contact

keegan@jimmy.harvard.edu

keegankorthauer

kkorthauer.org