Controlling for conservation in genome-wide DNA methylation studies

M. Singer and L. Pachter, BMC Genomics(2015)

Saket Choudhary

July 6, 2016

DNA Methylation

- Either cytosine(C) or adenine(A) undergoes methylation
- Typically represses gene expression
- Typically occurs in a `CpG' context (C followed by G)
- Methylated C often deaminates to T

Yule-Simpson Effect: Example from University Admission

	Female	Male
Applicants	550	550
Admitted	28.2%	41.8%

Acceptance percentage of female candidates is far less

Yule-Simpson Effect: Example from University Admission

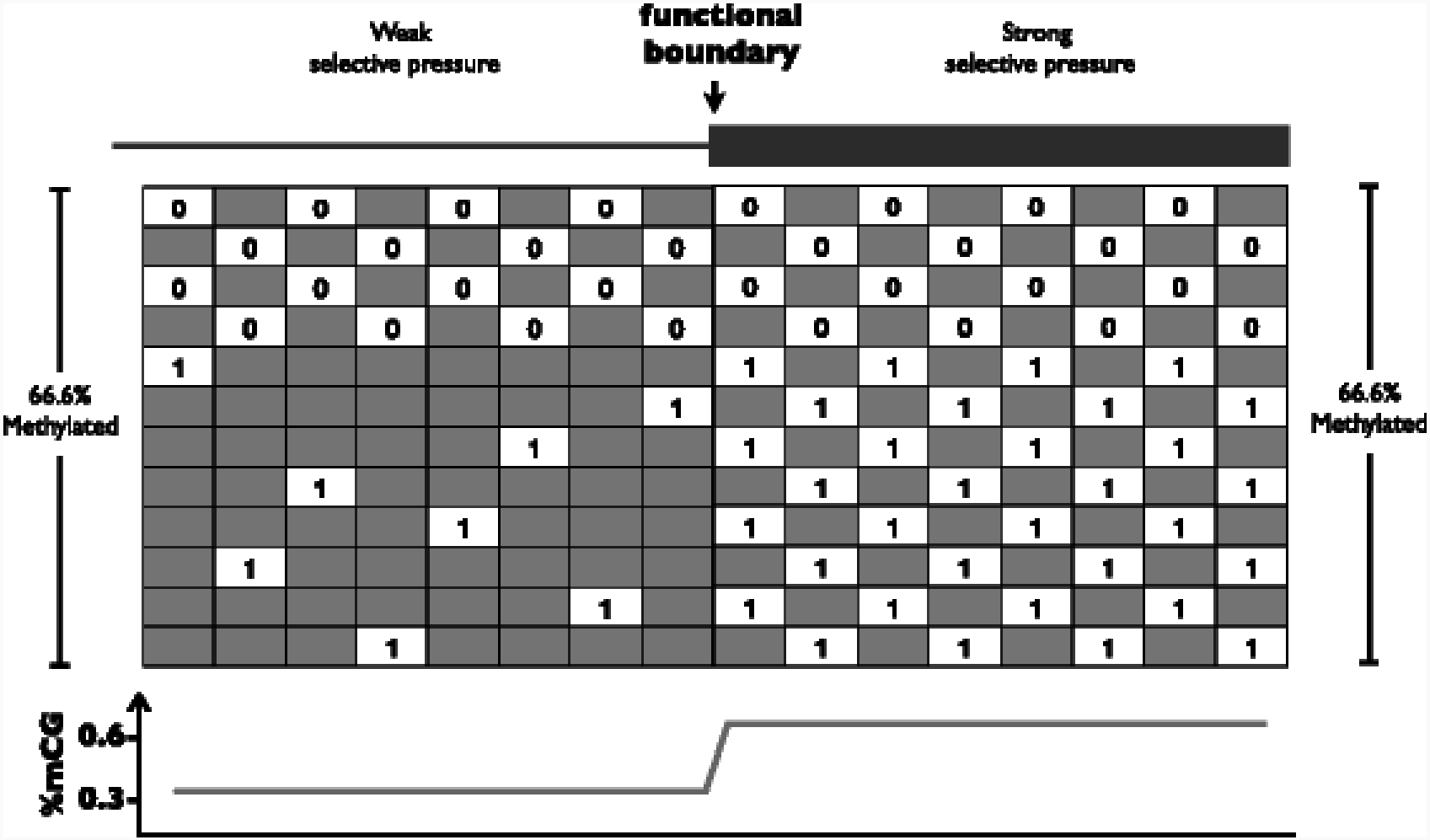
	Female	Male
Applicants	550	550
Admitted	28.2%	41.8%

Acceptance percentage of female candidates is far less

	Female		Male	
	Applicants	Admitted	Applicants	Admitted
Department A	150	50%	400	50%
Department B	400	20%	150	20%

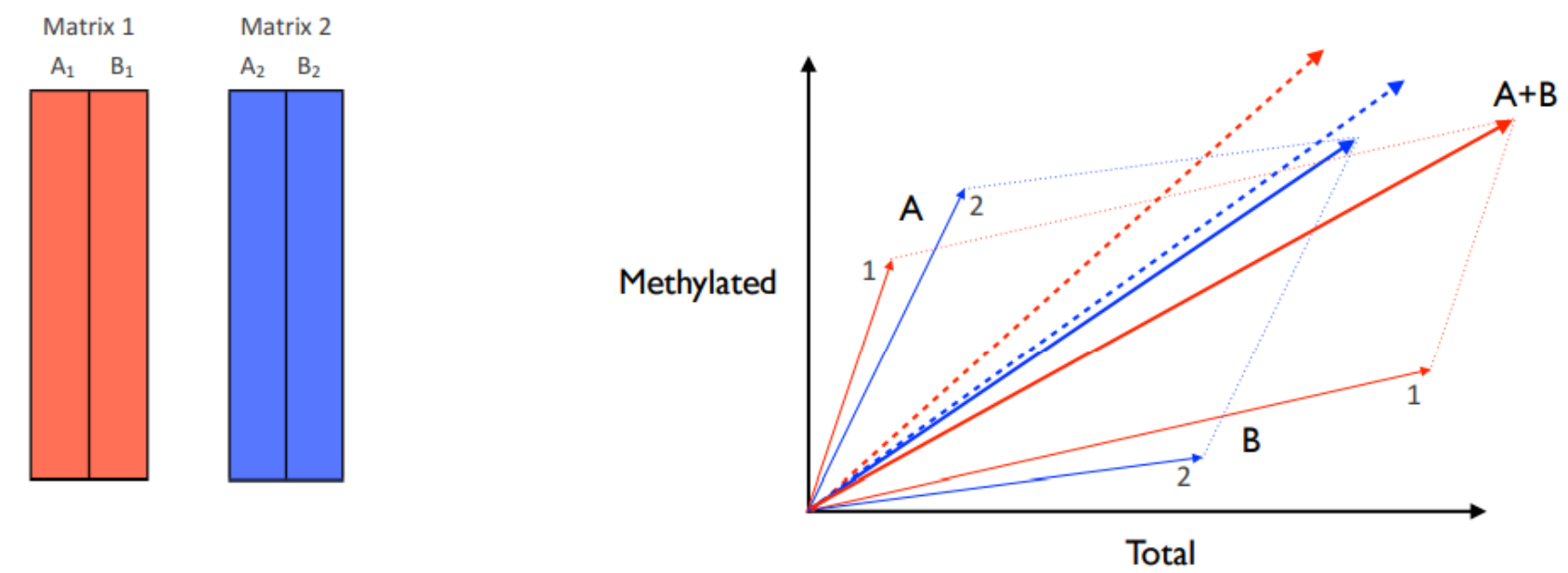
Departments do not display gender specific bias. Lower admission rates in female arises due females applying to department which are `harder' to get into

Yule-Simpson Effect: Averaging methylation states can be misleading



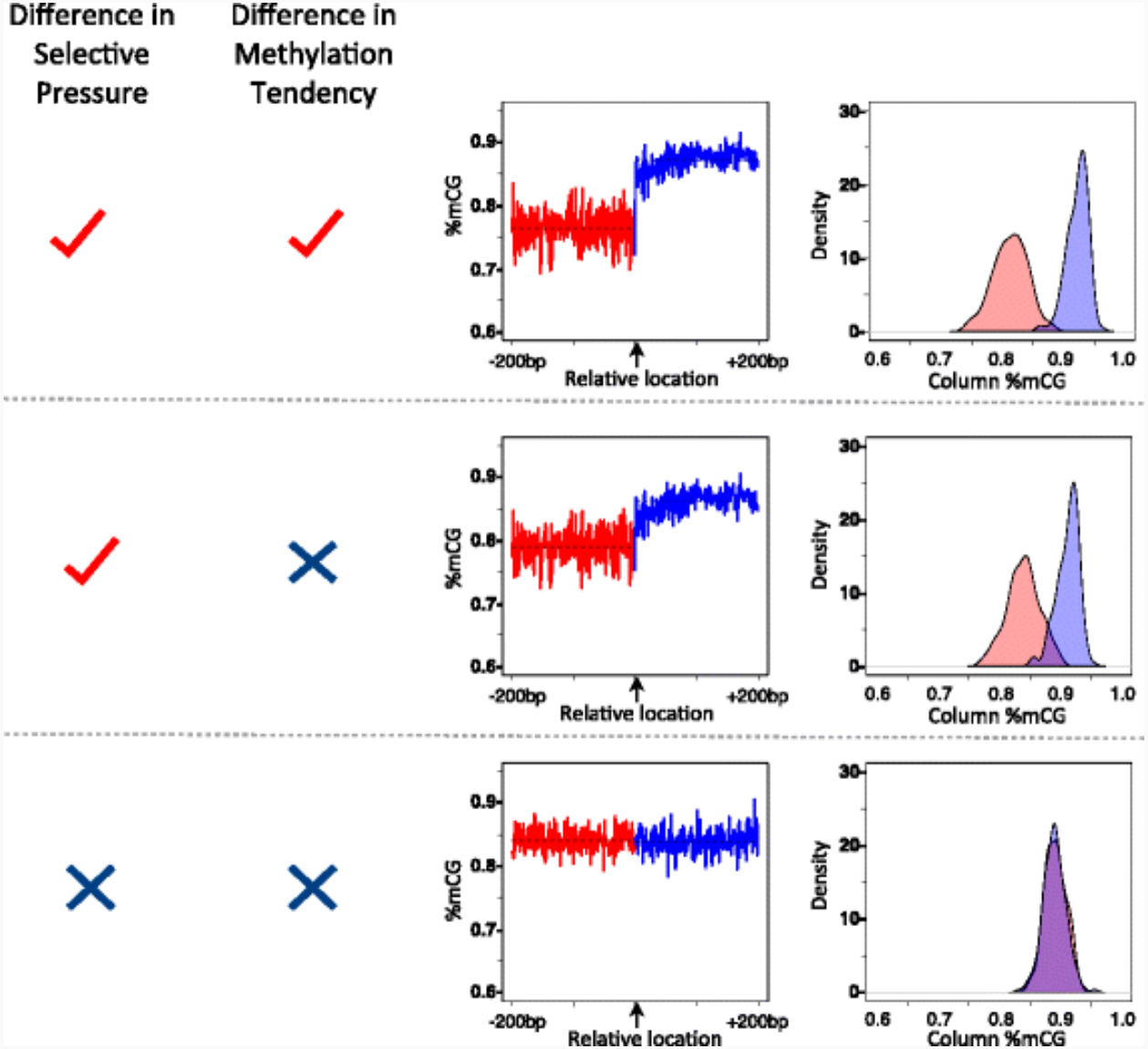
There is NO difference in methylation levels. YS Effect: Low frequency of methylated CGs in the left matrix

Yule-Simpson Effect: Geometric Interpretation



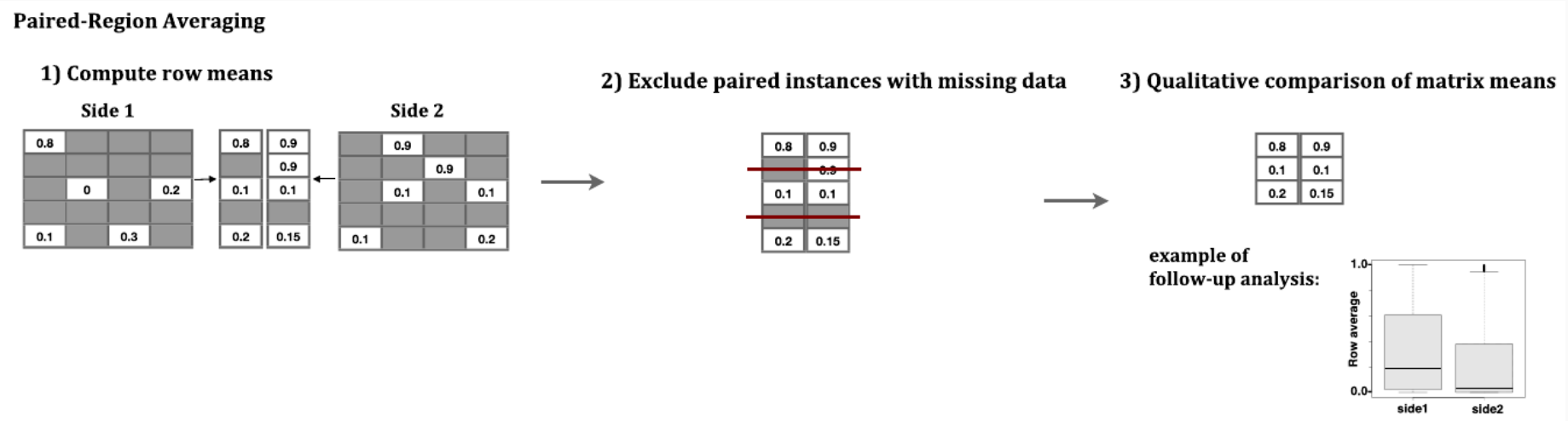
Average of the slopes is reverse of the slopes of average

YS effect arises due to different rates of selection



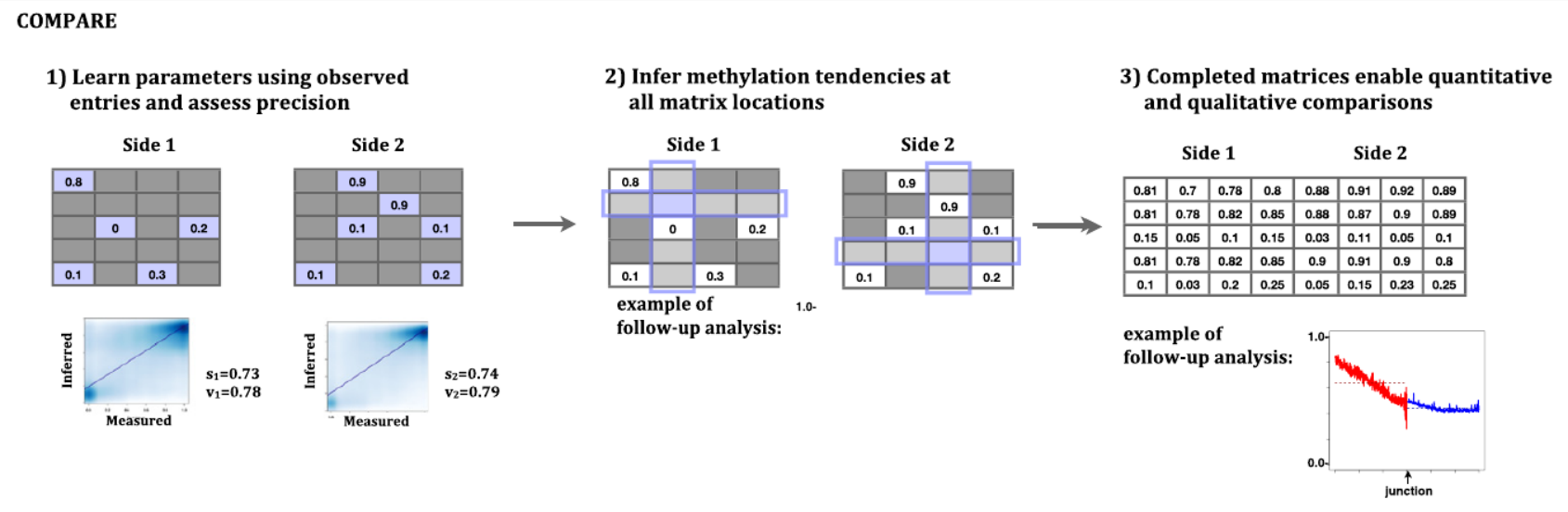
CpG = Missing data due to evolutionary constraint

Correction Method 1: Paired region averaging

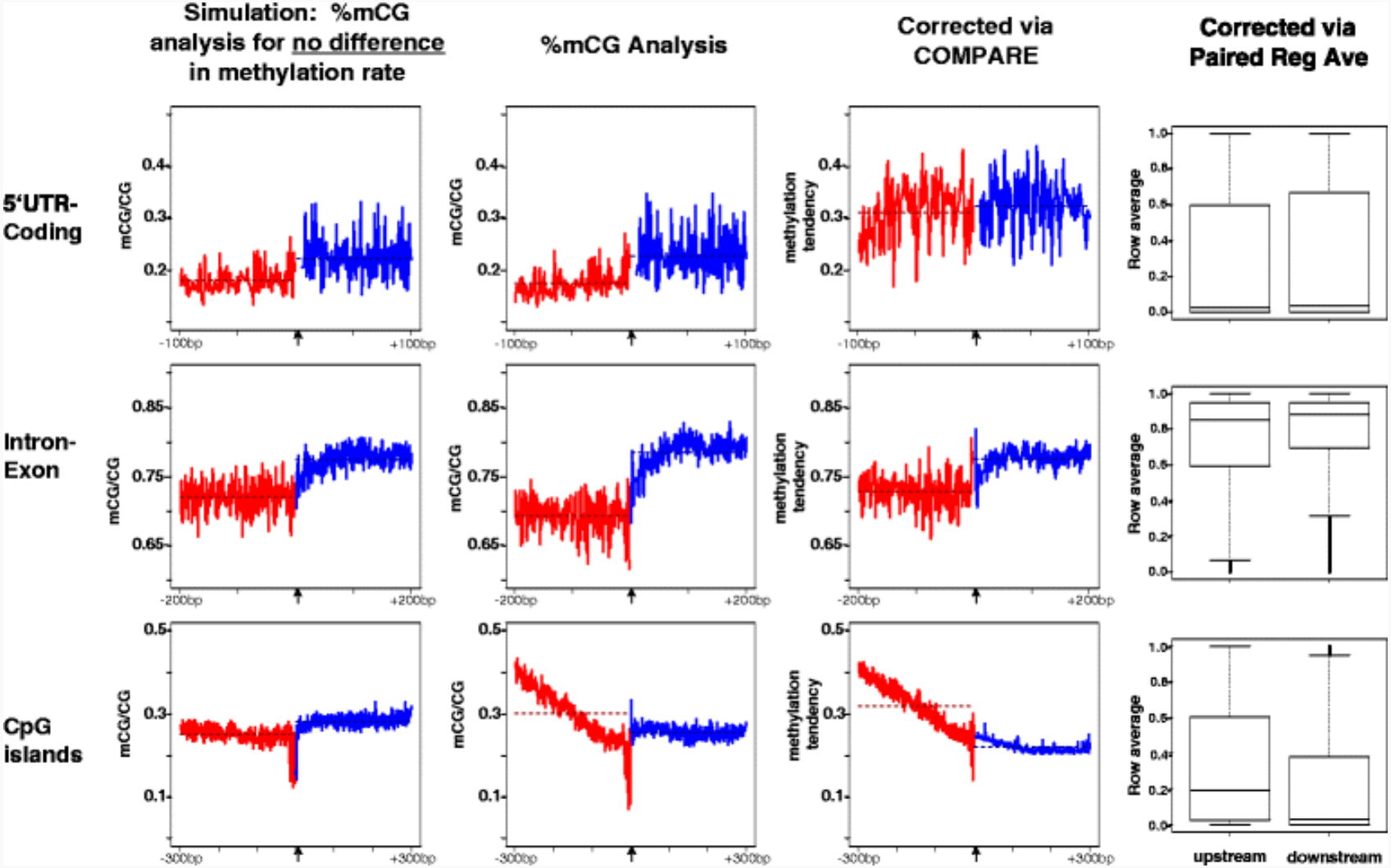


Discarding data overcomes YS-effect. Only qualitative comparisons permitted

Correction Method 2: COMPARE (COMparison of Phenotypes Averaged by REgion)



Key findings



Key findings

- Reanalysis of 5'UTR-coding boundaries revealed no significant difference in methylation tendencies
- Intro-exon junctions in both human and Arabidopsis revealed difference in methylation levels

Correction Method 2: COMPARE (COMparison of Phenotypes Averaged by REgion)

$$M_{i,j} = \frac{1}{1 + e^{-(b_j B_{i,j} + x_j X_{i,j} + y_j Y_{i,j} + z_j)}}$$

$$X_{i,j} = \text{Mean(row i) excluding } (i,j) \text{ entry}$$

$$Y_{i,j} = \text{Proportion of sites in row i with missing data}$$

$$B_{i,j} = \text{Indicator variable for methylability}$$

$$b_j, x_j, y_j, z_j = \text{Learned parameters}$$

Summary

- Näive averaging approaches can be heavily biased due to non-uniformity of underlying distribution
- Paired region averaging is a non-parametric approach that accounts for this non-uniformity for quanlitative comparison
- COMPARE is a parametric approach allowing quantitative comparison between regions