

# 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

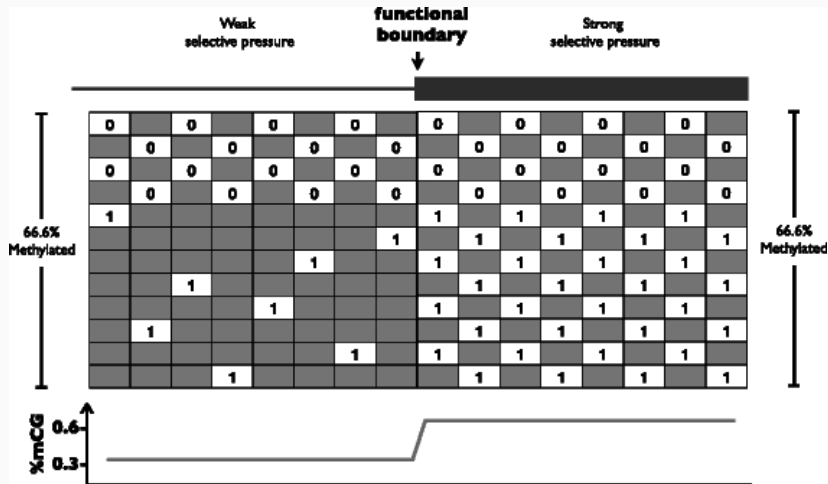
	Female	Male
Applicants	550	550
Admitted	28.2%	41.8%

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%

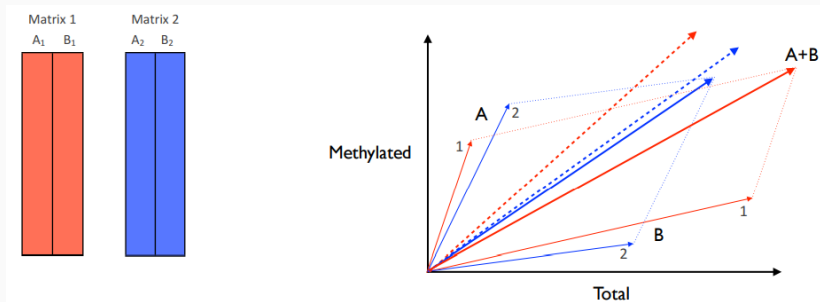
Departments do not display gender specific bias. Lower admission rates in female arises due females applying to department which are 'harder' to get into

# Yule-Simpson Effect: Averaging methylation states can be misleading



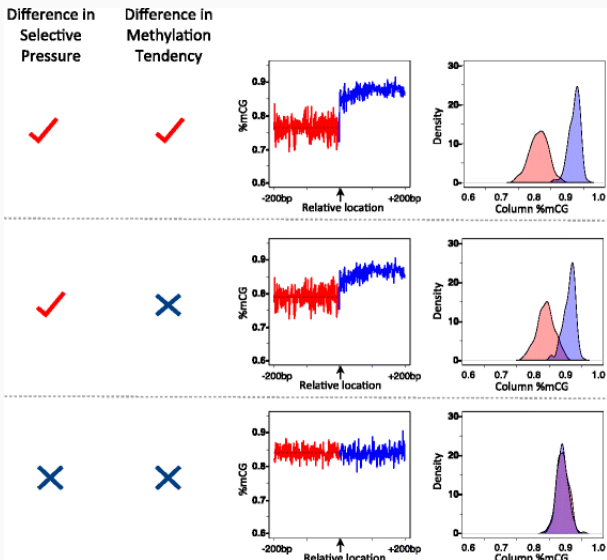
There is NO difference in methylation levels. YS Effect: Low frequency of methylated CGs in the left matrix

# 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

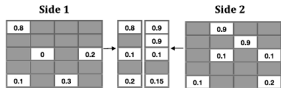


CpG = Missing data due to evolutionary constraint

# Correction Method 1: Paired region averaging

## Paired-Region Averaging

### 1) Compute row means



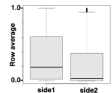
### 2) Exclude paired instances with missing data

0.8	0.9
<del>0.1</del>	<del>0.1</del>
0.2	0.15

### 3) Qualitative comparison of matrix means

0.8	0.9
0.1	0.1
0.2	0.15

example of  
follow-up analysis:



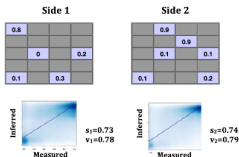
Discarding data overcomes YS-effect. Only qualitative comparisons permitted



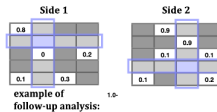
# Correction Method 2: COMPARE (COMparison of Phenotypes Averaged by REgion)

## COMPARE

- 1) Learn parameters using observed entries and assess precision



- 2) Infer methylation tendencies at all matrix locations



- 3) Completed matrices enable quantitative and qualitative comparisons

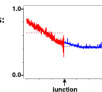
Side 1

0.81	0.7	0.78	0.8	0.88	0.91	0.92	0.89
0.81	0.78	0.82	0.85	0.88	0.87	0.9	0.89
0.15	0.05	0.1	0.15	0.03	0.11	0.05	0.1
0.81	0.78	0.82	0.85	0.9	0.91	0.9	0.8
0.1	0.03	0.2	0.25	0.05	0.15	0.23	0.25

Side 2

0.81	0.7	0.78	0.8	0.88	0.91	0.92	0.89
0.81	0.78	0.82	0.85	0.88	0.87	0.9	0.89
0.15	0.05	0.1	0.15	0.03	0.11	0.05	0.1
0.81	0.78	0.82	0.85	0.9	0.91	0.9	0.8
0.1	0.03	0.2	0.25	0.05	0.15	0.23	0.25

example of follow-up analysis:



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

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

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

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

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

$b_j, x_j, y_j, z_j$  = Learned parameters