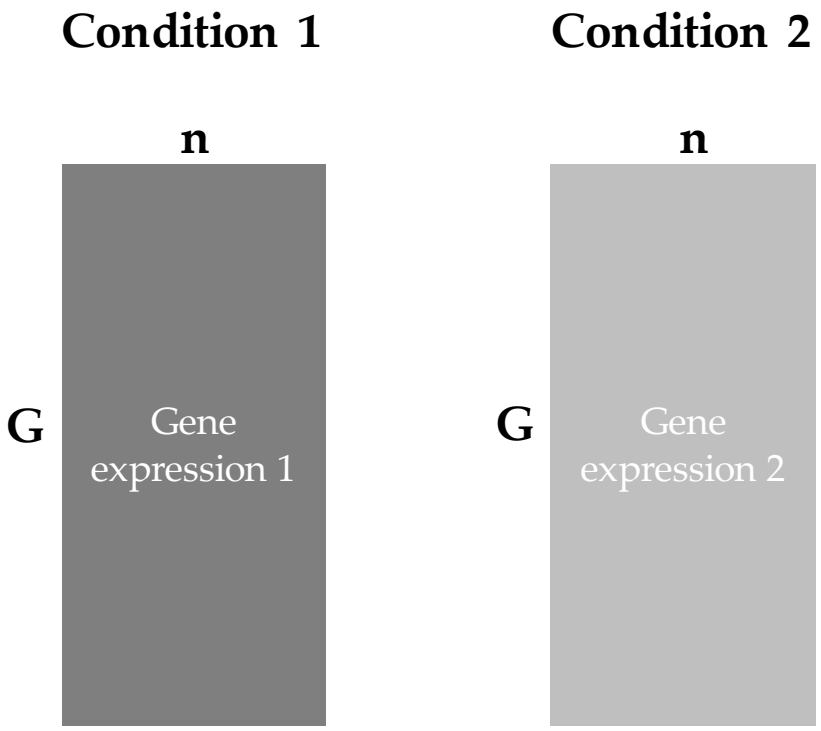
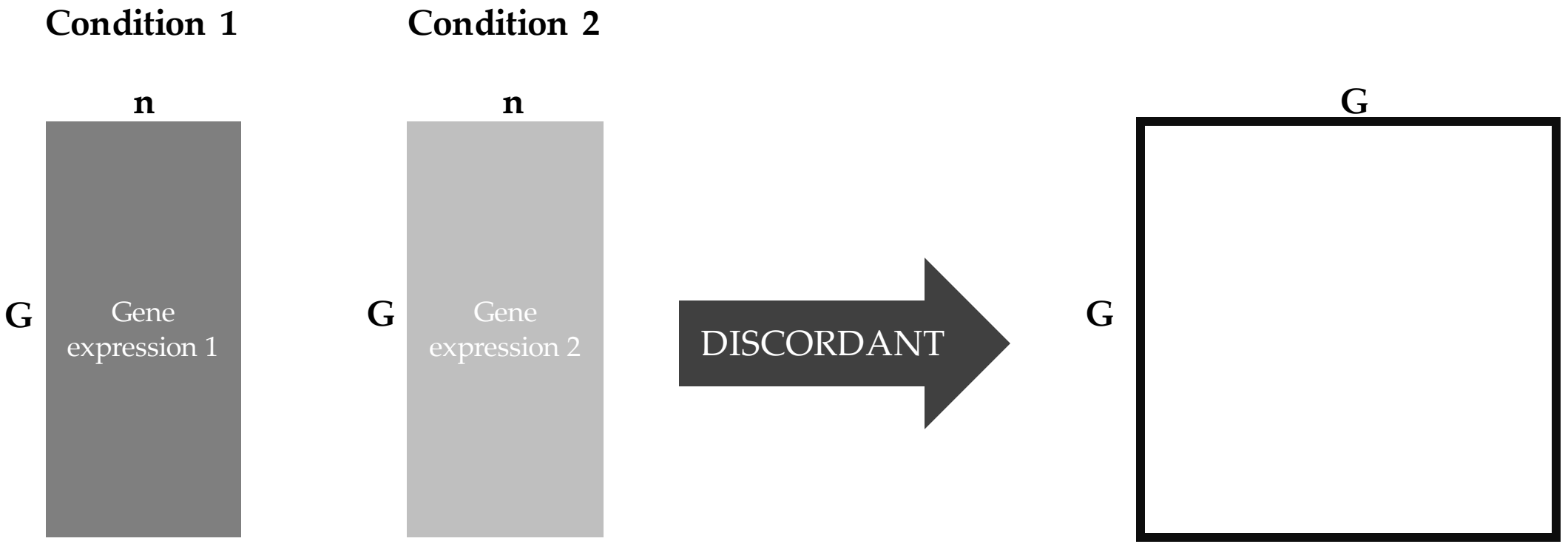
THE DISCORDANT METHOD: A NOVEL APPROACH FOR DIFFERENTIAL CORRELATION

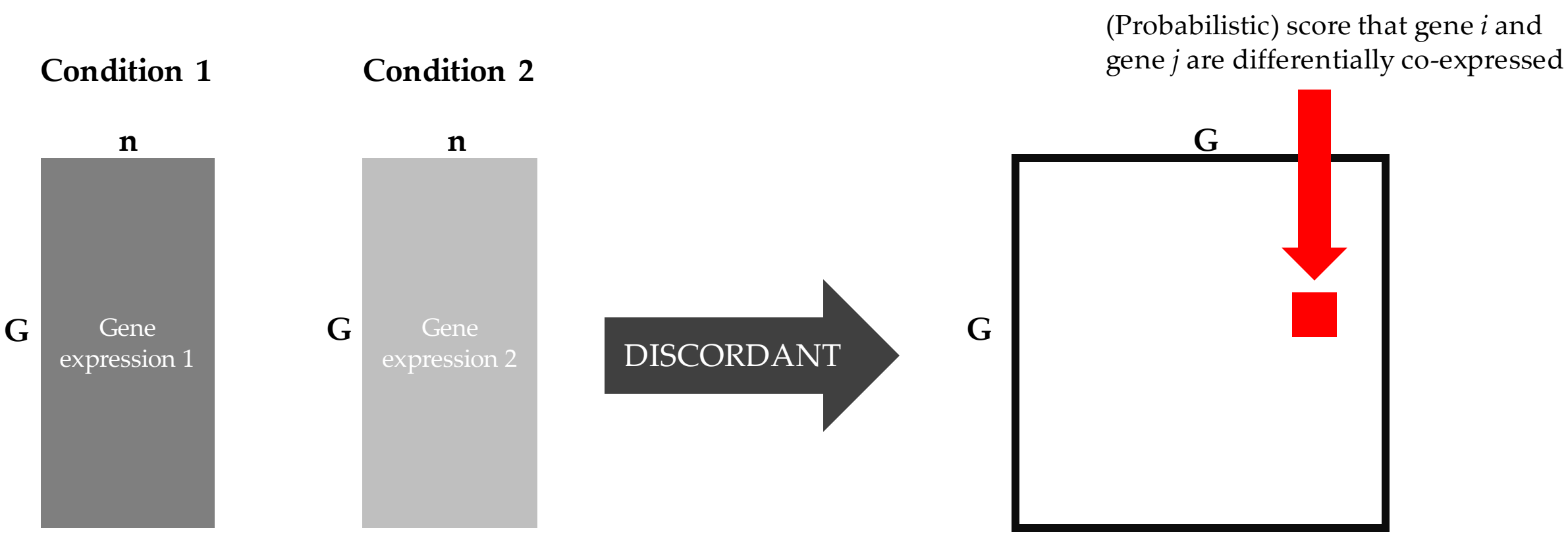
Charlotte Siska, Russell Bowler, Katerina Kechris; 2015

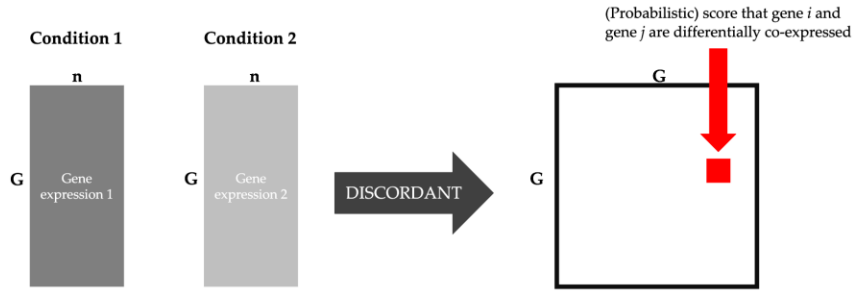
By Soel Micheletti · JQ Lab Journal Club · June 2023

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

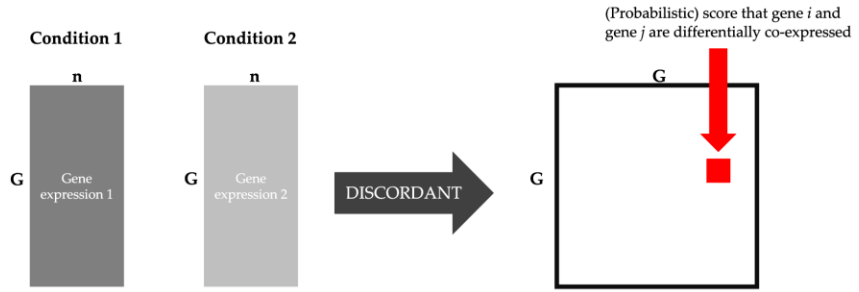

Condition 1
Condition 2
n
n
G
Gene expression 1
G
Gene expression 2


Condition 1
Condition 2
n
n
G
G
G
G
Gene expression 1
Gene expression 2
DISCORDANT

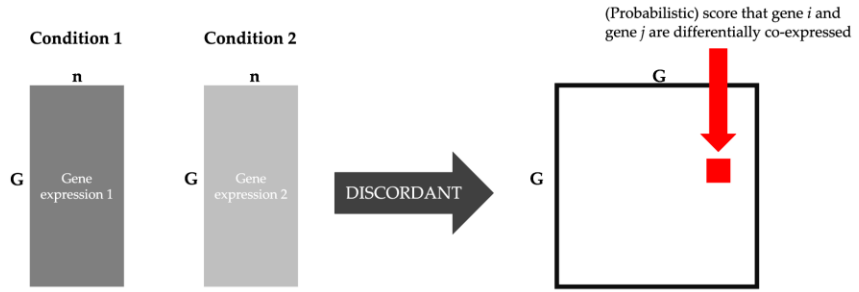

Condition 1
Condition 2
n
n
G
G
Gene expression 1
Gene expression 2
DISCORDANT
(Probabilistic) score that gene *i* and gene *j* are differentially co-expressed
G
G



- Compute co-expression matrix for both conditions



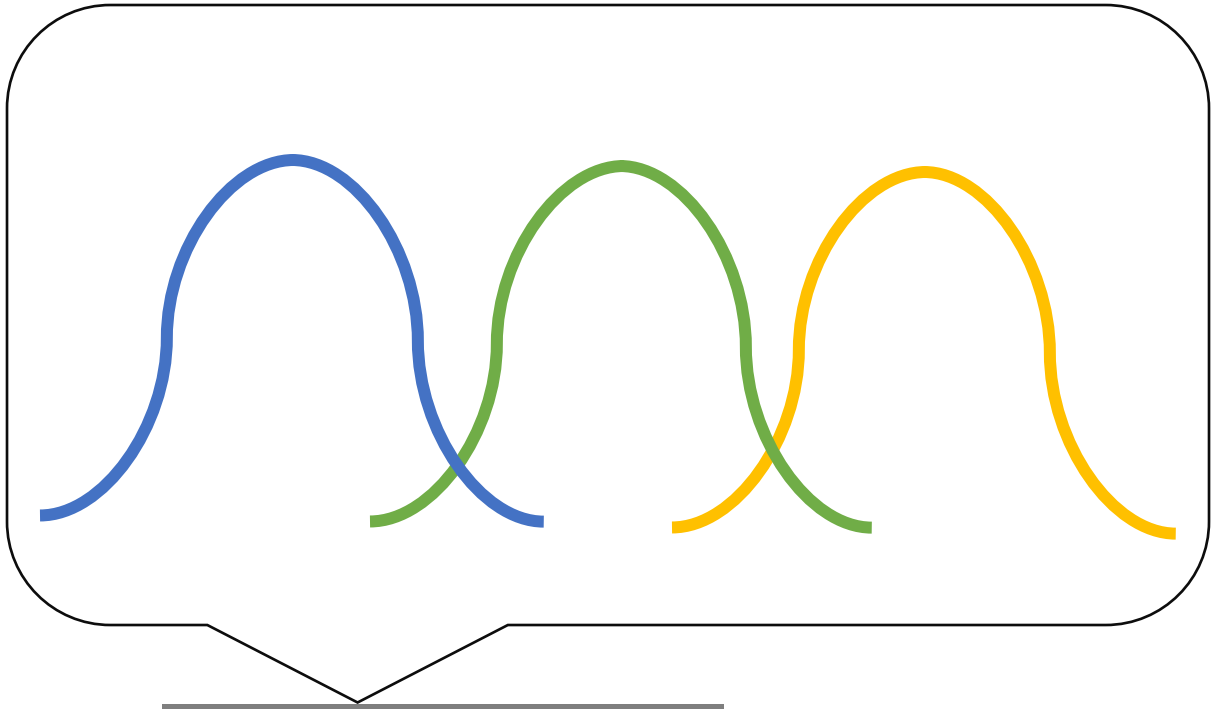
- Compute co-expression matrix for both conditions
- Apply Fisher's transformation to convert Pearson's correlation coefficients into z-scores



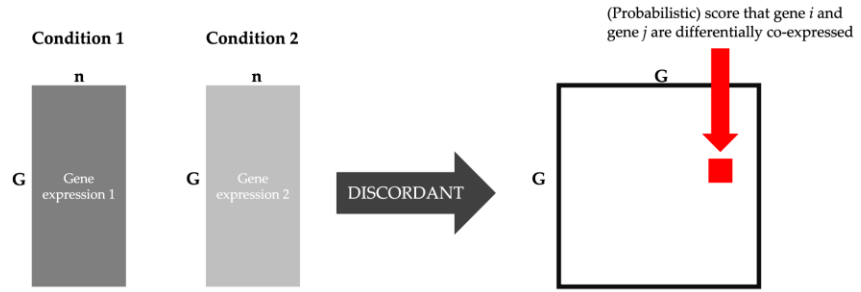
- Compute co-expression matrix for both conditions
- Apply Fisher's transformation to convert Pearson's correlation coefficients into z-scores

Matrix of z-scores for condition 1

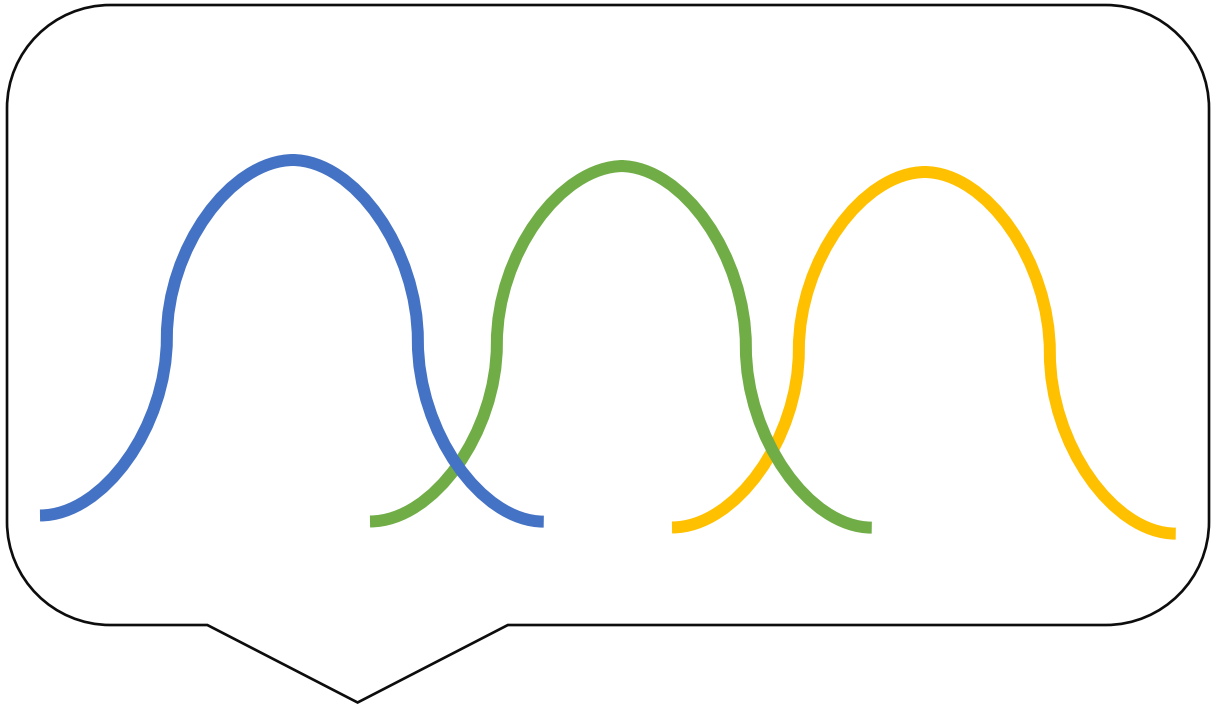
Matrix of z-scores for condition 2

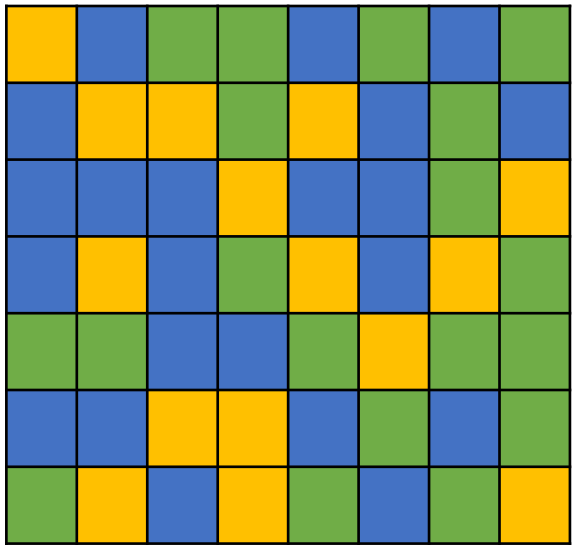


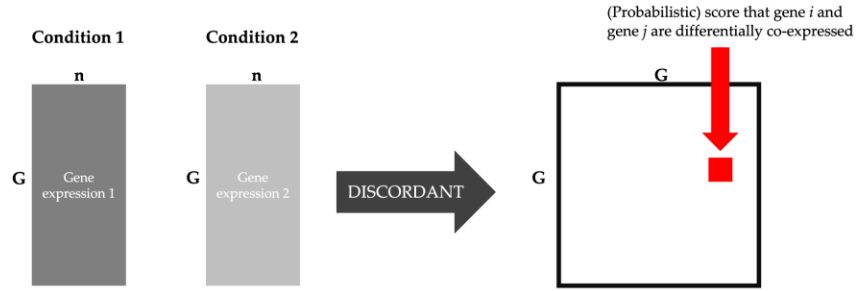
Matrix of z-scores for condition 1



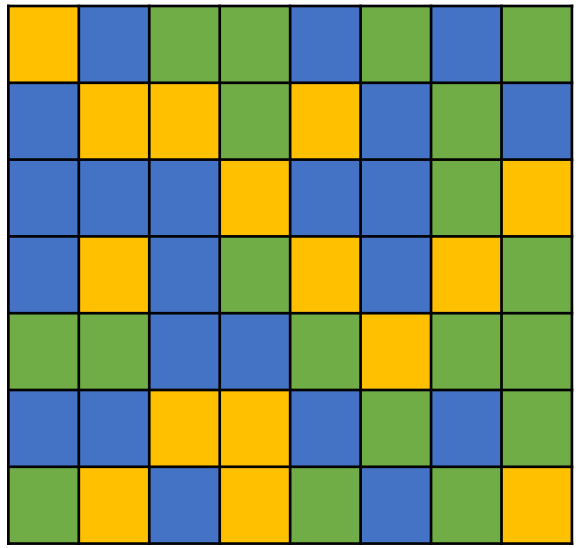
Matrix of z-scores for condition 2

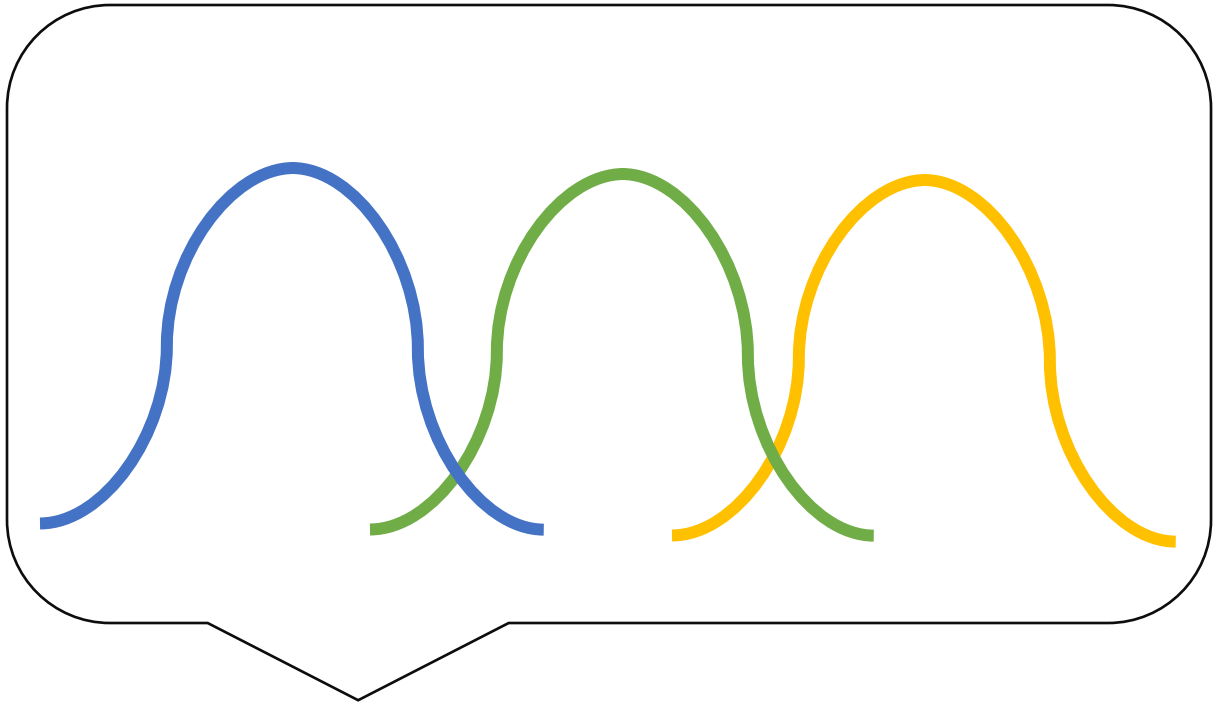


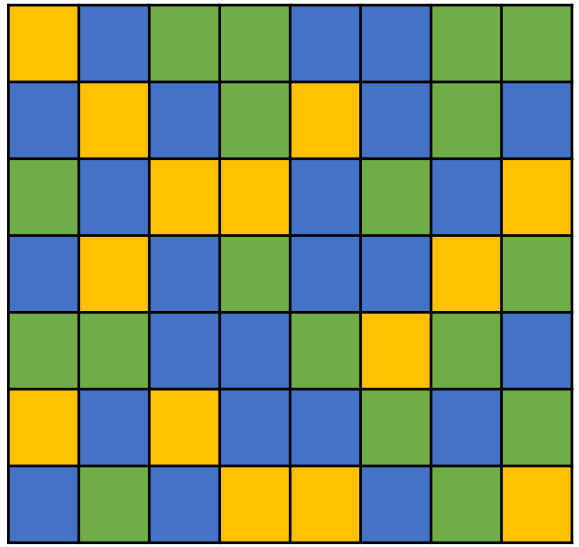




Matrix of z-scores for condition 2

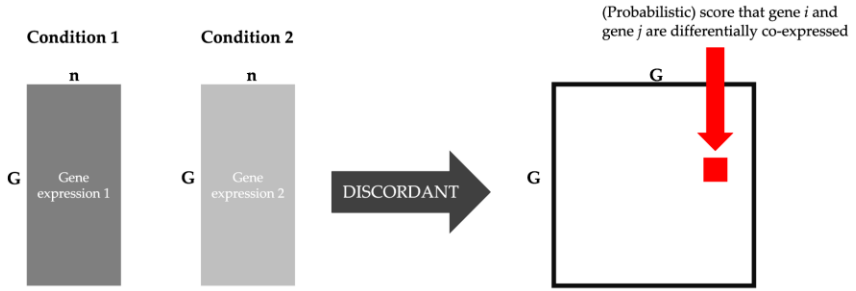


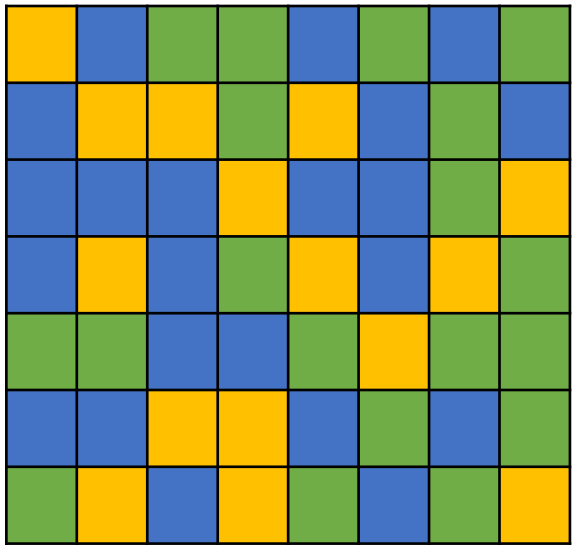


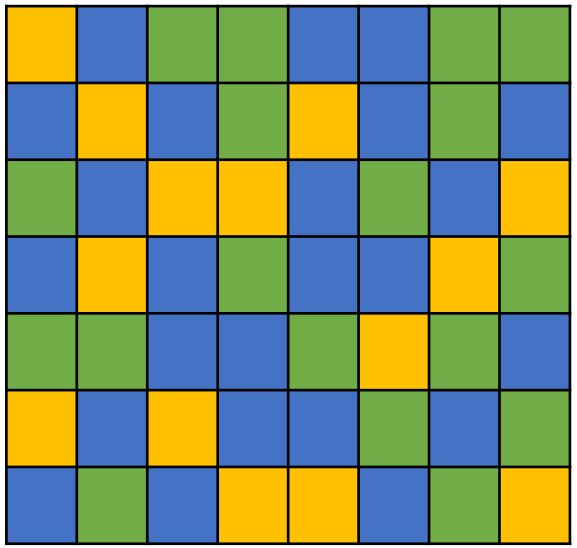


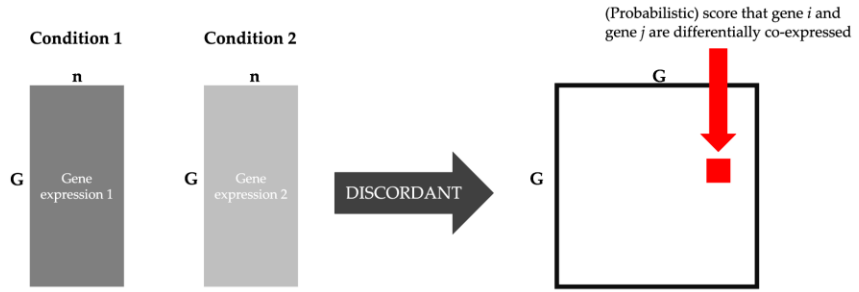
Main idea:

- Correlation smaller than average
- Average correlation
- Correlation larger than average





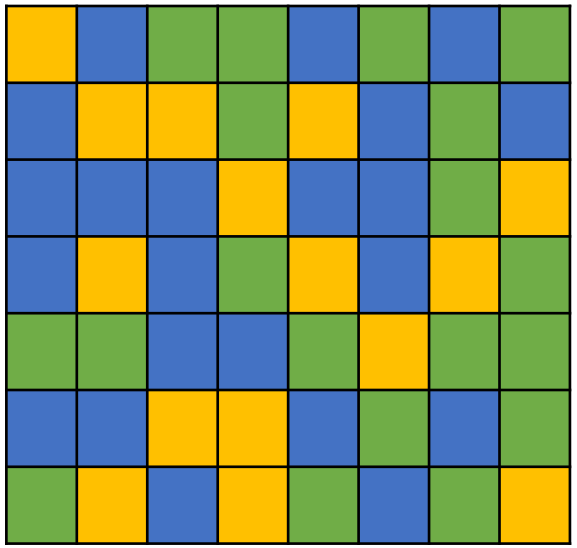


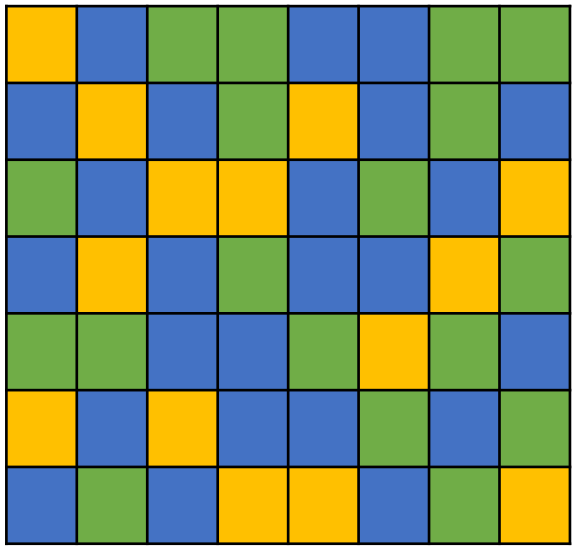


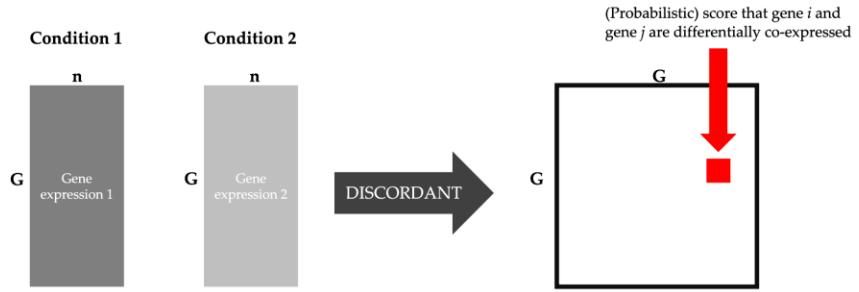
Main idea:

- Correlation smaller than average
- Average correlation
- Correlation larger than average

A pair is differentially coexpressed if it belongs to different categories





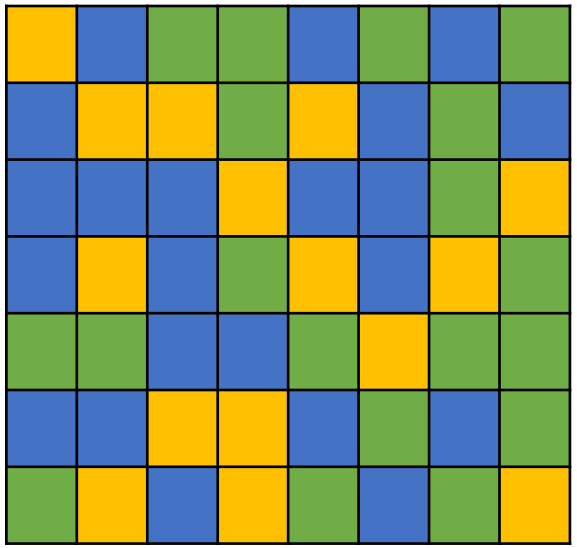


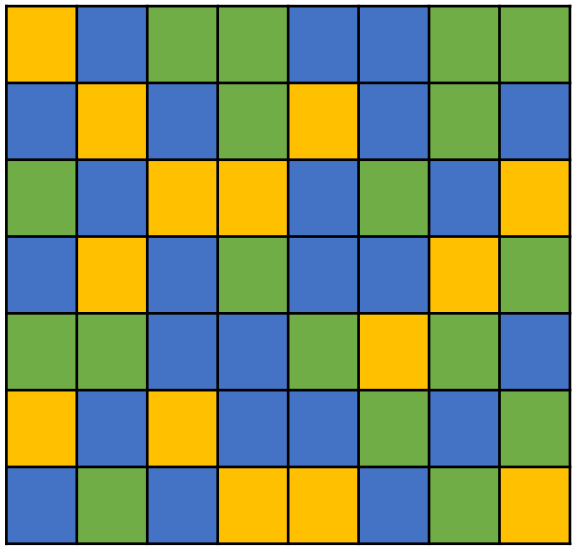
Main idea:

- Correlation smaller than average
- Average correlation
- Correlation larger than average

A pair is differentially coexpressed if it belongs to different categories

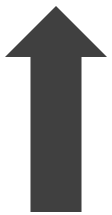
For each gene pair, compute Pr[blue], Pr[green], and Pr[orange]





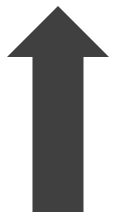
$$\underset{\Theta=[\Theta_1,\Theta_2]}{\arg\max} \prod_{k} Pr\left[z_1(k), z_2(k) \middle| \Theta\right]$$

$$\underset{\Theta=[\Theta_1,\Theta_2]}{\arg\max} \prod_{k} Pr\left[z_1(k), z_2(k) \mid \Theta\right]$$



Gene pairs

$$\underset{\Theta=[\Theta_1,\Theta_2]}{\arg\max} \prod_k Pr\left[z_1(k), z_2(k) \mid \Theta\right]$$



Mean, covariance, and weight for each normal distribution

$$\underset{\Theta=[\Theta_1,\Theta_2]}{\arg\max} \prod_k Pr\left[z_1(k), z_2(k) \middle| \Theta\right] = \prod_k Pr\left[z_1(k) \middle| \Theta_1\right] Pr\left[z_2(k) \middle| \Theta_2\right]$$

$$\underset{\Theta=[\Theta_1,\Theta_2]}{\arg\max} \prod_k Pr\left[z_1(k), z_2(k) \middle| \Theta\right] = \prod_k Pr\left[z_1(k) \middle| \Theta_1\right] Pr\left[z_2(k) \middle| \Theta_2\right]$$

(Simplyfing assumption: independence)

$$\underset{\Theta=[\Theta_1,\Theta_2]}{\arg\max} \prod_k Pr\left[z_1(k), z_2(k) \middle| \Theta\right] = \prod_k Pr\left[z_1(k) \middle| \Theta_1\right] Pr\left[z_2(k) \middle| \Theta_2\right]$$

(Simplyfing assumption: independence)

$$= \prod_k \sum_{i=0}^{2} \sum_{j=0}^{2} w_i^{(1)} \mathcal{N}(z_1(k) | \mu_i^{(1)}, \Sigma_i^{(1)}) w_j^{(2)} \mathcal{N}(z_2(k) | \mu_j^{(2)}, \Sigma_j^{(2)})$$

$$\underset{\Theta=[\Theta_1,\Theta_2]}{\arg\max} \prod_k Pr\left[z_1(k), z_2(k) \middle| \Theta\right] = \prod_k Pr\left[z_1(k) \middle| \Theta_1\right] Pr\left[z_2(k) \middle| \Theta_2\right]$$

(Simplyfing assumption: independence)

$$= \prod_k \sum_{i=0}^{2} \sum_{j=0}^{2} w_i^{(1)} \mathcal{N}(z_1(k) | \mu_i^{(1)}, \Sigma_i^{(1)}) w_j^{(2)} \mathcal{N}(z_2(k) | \mu_j^{(2)}, \Sigma_j^{(2)})$$

Difficult optimization: non-convex with constraints

$$\underset{\Theta=[\Theta_1,\Theta_2]}{\arg\max} \prod_k Pr\left[z_1(k), z_2(k) \middle| \Theta\right] = \prod_k Pr\left[z_1(k) \middle| \Theta_1\right] Pr\left[z_2(k) \middle| \Theta_2\right]$$

(Simplyfing assumption: independence)

$$= \prod_k \sum_{i=0}^{2} \sum_{j=0}^{2} w_i^{(1)} \mathcal{N}(z_1(k) | \mu_i^{(1)}, \Sigma_i^{(1)}) w_j^{(2)} \mathcal{N}(z_2(k) | \mu_j^{(2)}, \Sigma_j^{(2)})$$

Difficult optimization: non-convex with constraints ==> **EM algorithm**

$$\underset{\Theta=[\Theta_1,\Theta_2]}{\arg\max} \prod_k Pr\left[z_1(k), z_2(k) \middle| \Theta\right] = \prod_k Pr\left[z_1(k) \middle| \Theta_1\right] Pr\left[z_2(k) \middle| \Theta_2\right]$$

(Simplyfing assumption: independence)

$$= \prod_k \sum_{i=0}^{2} \sum_{j=0}^{2} w_i^{(1)} \mathcal{N}(z_1(k) | \mu_i^{(1)}, \Sigma_i^{(1)}) w_j^{(2)} \mathcal{N}(z_2(k) | \mu_j^{(2)}, \Sigma_j^{(2)})$$

- Initialize parameters $\Theta_i = \{w_0^{(i)}, w_1^{(i)}, w_2^{(i)}, \mu_0^{(i)}, \mu_1^{(i)}, \mu_2^{(i)}, \Sigma_0^{(i)}, \Sigma_1^{(i)}, \Sigma_2^{(i)}\}$ for both conditions i

$$\underset{\Theta=[\Theta_1,\Theta_2]}{\arg\max} \prod_k Pr\left[z_1(k), z_2(k) \middle| \Theta\right] = \prod_k Pr\left[z_1(k) \middle| \Theta_1\right] Pr\left[z_2(k) \middle| \Theta_2\right]$$

(Simplyfing assumption: independence)

$$= \prod_k \sum_{i=0}^{2} \sum_{j=0}^{2} w_i^{(1)} \mathcal{N}(z_1(k) | \mu_i^{(1)}, \Sigma_i^{(1)}) w_j^{(2)} \mathcal{N}(z_2(k) | \mu_j^{(2)}, \Sigma_j^{(2)})$$

- Initialize parameters $\Theta_i = \{w_0^{(i)}, w_1^{(i)}, w_2^{(i)}, \mu_0^{(i)}, \mu_1^{(i)}, \mu_2^{(i)}, \Sigma_0^{(i)}, \Sigma_1^{(i)}, \Sigma_2^{(i)}\}$ for both conditions i
- Iterate until convergence
  - **E-step**

    Assign to each sample a probability of being in 'color i' in the first condition and 'color j' in the second condition
  - **M-step**

    Solve the MLE using the soft labels

$$\underset{\Theta=[\Theta_1,\Theta_2]}{\arg\max} \prod_k Pr\left[z_1(k), z_2(k) \middle| \Theta\right] = \prod_k Pr\left[z_1(k) \middle| \Theta_1\right] Pr\left[z_2(k) \middle| \Theta_2\right]$$

(Simplyfing assumption: independence)

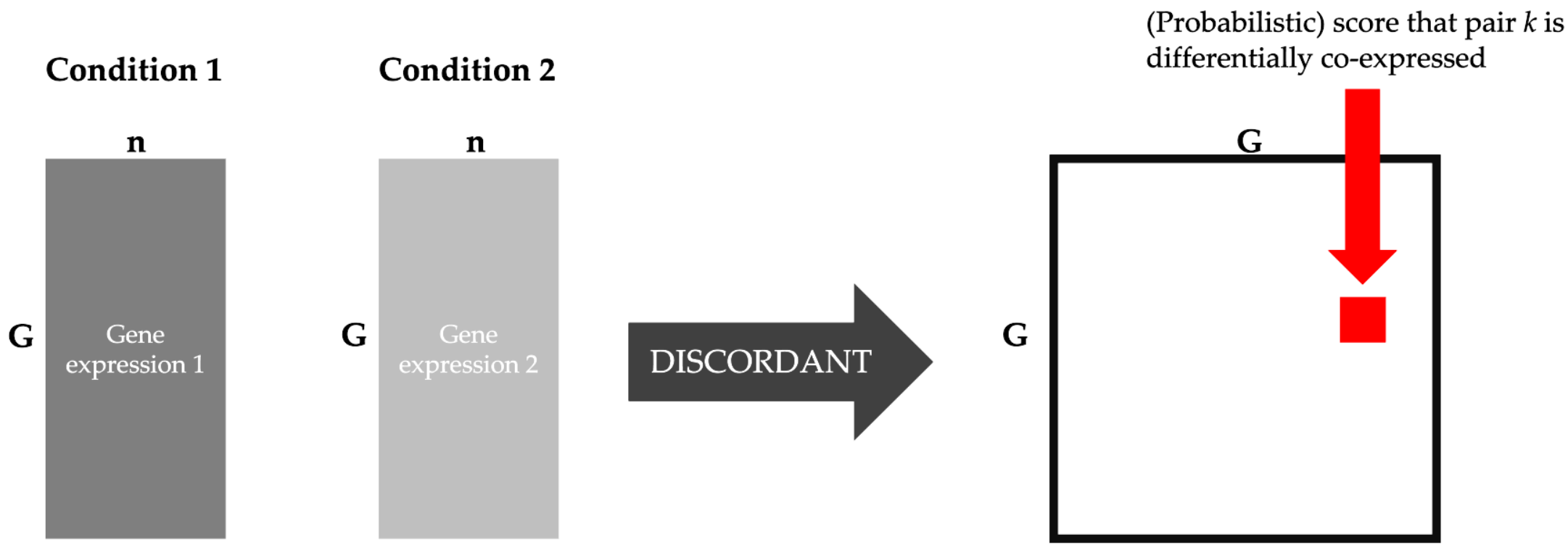
$$= \prod_k \sum_{i=0}^{2} \sum_{j=0}^{2} w_i^{(1)} \mathcal{N}(z_1(k) | \mu_i^{(1)}, \Sigma_i^{(1)}) w_j^{(2)} \mathcal{N}(z_2(k) | \mu_j^{(2)}, \Sigma_j^{(2)})$$

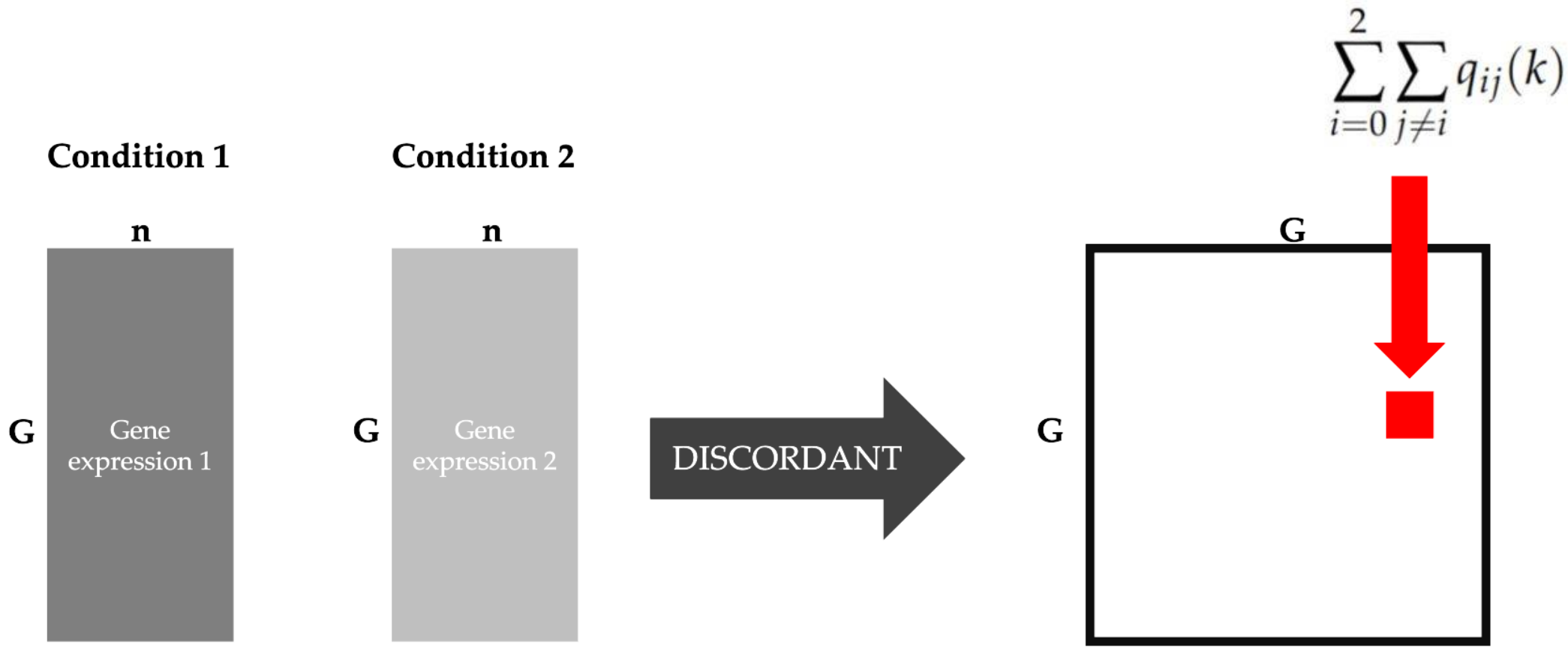
- Initialize parameters $\Theta_i = \{w_0^{(i)}, w_1^{(i)}, w_2^{(i)}, \mu_0^{(i)}, \mu_1^{(i)}, \mu_2^{(i)}, \Sigma_0^{(i)}, \Sigma_1^{(i)}, \Sigma_2^{(i)}\}$ for both conditions i
- Iterate until convergence
  - **E-step**

$$q_{ij}(k) = \frac{w_i^{(1)} \mathcal{N}(z_1(k) | \mu_i^{(1)}, \Sigma_i^{(1)})}{\sum_{\tau=0}^{2} w_\tau^{(1)} \mathcal{N}(z_1(k) | \mu_\tau^{(1)}, \Sigma_\tau^{(1)})} \cdot \frac{w_j^{(2)} \mathcal{N}(z_2(k) | \mu_j^{(2)}, \Sigma_j^{(2)})}{\sum_{\tau=0}^{2} w_\tau^{(2)} \mathcal{N}(z_1(k) | \mu_\tau^{(2)}, \Sigma_\tau^{(2)})}$$

  - **M-step**

    Solve the MLE using the soft labels


Condition 1
Condition 2
(Probabilistic) score that pair k is differentially co-expressed
n
n
G
G
G
Gene expression 1
Gene expression 2
DISCORDANT


Condition 1
n
G
Gene expression 1
Condition 2
n
G
Gene expression 2
DISCORDANT
$\sum_{i=0}^{2} \sum_{j \neq i} q_{ij}(k)$
G
G

Biological validation

- Dataset containing miRNA + mRNA expression for 10 healthy subjects and 21 patients with glioblastoma multiforme
- Interested in the differential correlation between miRNA and mRNA pairs

Biological validation

- Dataset containing miRNA + mRNA expression for 10 healthy subjects and 21 patients with glioblastoma multiforme
- Interested in the differential correlation between miRNA and mRNA pairs

Data	Method	Mean	Median
GBM	Discordant	464.75	347.5
	EBcoexpress	815	607
	Fisher	781	801
	Linear (miRNA-independent)	1095	532.5
	Linear (transcript-independent)	2596.5	787.5

My take-aways

- The probability of differential co-expression can be conveniently used in GSEA

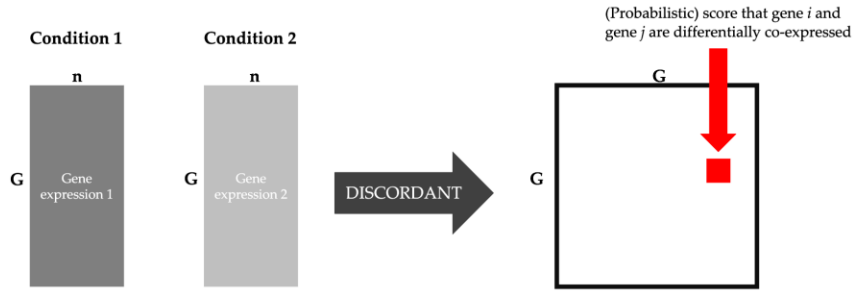
- Methodological paper, no thorough validation with biological data

- Idea of a data-driven binning is elegant:

  - It improves interpretability
  - It enables to determine more differentially correlated pairs than other methods, potentially imrpoving power of detecting disrupted interactions

- Assumption that the z-scores form a tri-modal normal distribution is not discussed

My take-aways



- The probability of differential co-expression can be conveniently used in GSEA
- Methodological paper, no thorough validation with biological data
- Idea of a data-driven binning is elegant:
  - It improves interpretability
  - It enables to determine more differentially correlated pairs than other methods, potentially imrpoving power of detecting disrupted interactions
- Assumption that the z-scores form a tri-modal normal distribution is not discussed