重慶醫科大學
CHONGQING MEDICAL UNIVERSITY

Mutual information for detecting multi-class biomarkers when integrating multiple omics studies

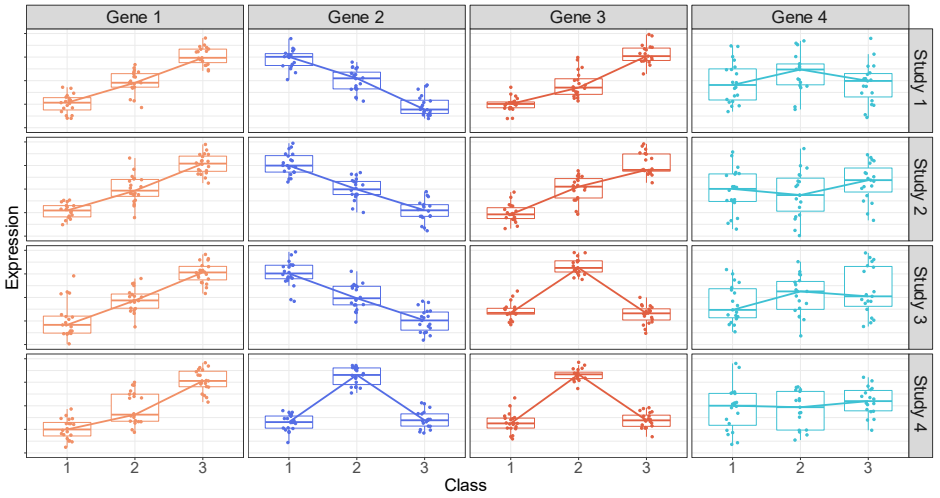
Jian Zou, Ph.D.
Department of Statistics, School of Public Health
Chongqing Medical University
jianzou@cqmu.edu.cn

2024-07-19

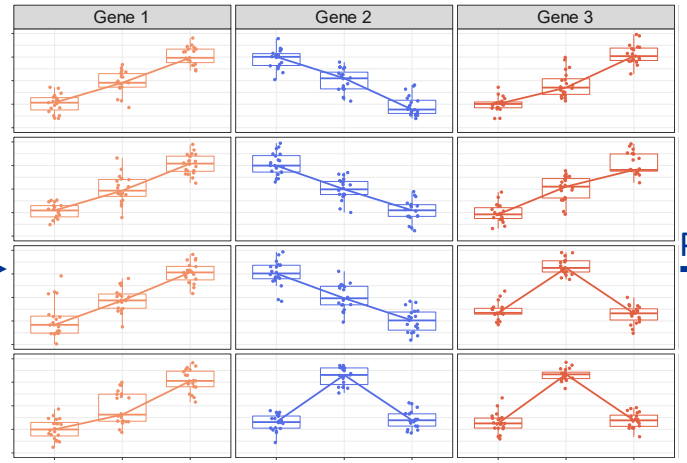
Biomarker detection

- **Biomarker detection**, which provides accurate biological information for early disease diagnosis, is a crucial element in biomedical research.
- **Study integration** is a common approach for improving the reliability and power of biomarker detection. If a biomarker shows similar patterns across multiple studies, we could believe that it is a robust choice for disease indication.
- Combining p-values and combing effect sizes are two leading solutions for study integration.
  - The p-value combination only focuses on the significance level without considering the data pattern
  - The effect size combination is only available in the two-class scenario (usually the disease vs. normal).

Mutual Information Concordance Analysis (MICA)

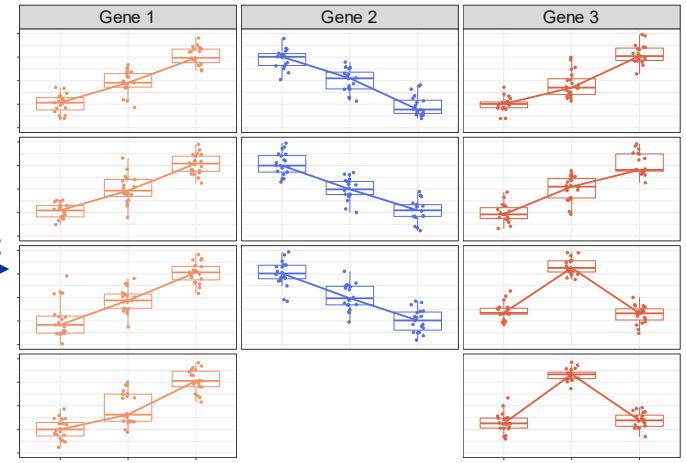


Overall →



Identify the genes with concordant expression patterns

Post-hoc →



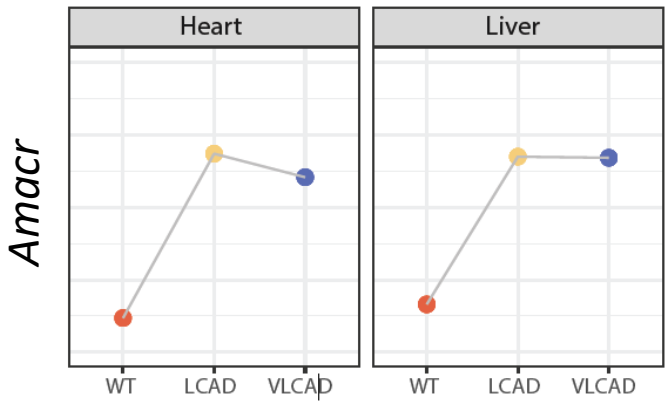
Identify the studies which contribute to the concordance

Problem Setting

- Assume *K* classes ($K \geq 2$) and *S* transcriptomic studies.
- Annotate x_{ski} as the gene expression for one gene in study *s*, class *k,* sample *i.*

- If there is only 2 studies (*X* and *Y*) and no replicates, we could simply use the Pearson correlation

$$Cor_{X,Y} = \rho_{X,Y} = \frac{\sum_{k=1}^{K}(x_{k1} - \bar{x})(y_{k1} - \bar{y})}{\sqrt{\sum_{k=1}^{K}(x_{k1} - \bar{x})^2 \sum_{k=1}^{K}(y_{k1} - \bar{y})^2}}$$



Multi-class correlation (MCC) and min-MCC

- When there are multiple replicates within each class:
- For study *X*, the observed gene expression x_{kj} from sample *j* class *k* is assumed to be obtained from $X_k \sim N(\mu_{X_k},\ \sigma^2_{X_k})$, where $X_k \perp\!\!\!\perp X_{k'}$, $(\forall\ k \neq k')$.
- *X* can be naturally defined as a mixture distribution of X_k $(k = 1{:}\,K)$, where w_k is the class weight.

$$f_X(x) = \sum_{k=1}^{K} w_k f_{X_k}(x)$$

$$E(X) = \mu_X = \sum_{k=1}^{K} w_k \mu_{X_k}, \qquad Var(X) = \sigma_X^2 = \sum_{k=1}^{K} w_k\left(\sigma^2_{X_k} + \mu^2_{X_k}\right) - \mu_X^2$$

- Study *Y* is similarly defined, and Y_k is independent with X_k.
- The above-mentioned parameters can all be directly estimated from the data.

Multi-class correlation (MCC) and min-MCC

- MCC and min-MCC (Lu *et al*., 2010) are the only available statistics to detect such biomarkers for now.
- Multi-class correlation (MCC) is therefore defined as

$$MCC = \rho = \frac{E(XY) - EX \cdot EY}{\sqrt{Var(X) \cdot Var(Y)}} = \frac{(\sum_{k=1}^{K} w_k \, \mu_{X\,k} \mu_{Y_k}) - \mu_X \cdot \mu_Y}{\sigma_X \cdot \sigma_Y}$$

- For multiple *S* studies, min-MCC is then defined as minimum value of pair-wise MCC

$$min - MCC = \min_{1 \le u < v \le S} (MCC_{(u),(v)})$$

- However, the hypothesis test $HS_{min-MCC}$ for min-MCC is

$$\{H_0 \colon \exists\ \rho_{ij} \le 0\ vs. H_A \colon \forall\ \rho_{ij} > 0\}$$

All the studies should contain a consistent pattern simultaneously.

- Drawbacks:
  - Overlook the situation when only partial studies share the multi-class pattern
  - Cases where all pairs of studies have a uniformly low concordance *vs.* only one pair has a very low concordance

Mutual Information Concordance Analysis (MICA)

- We assumed *X* and *Y* to be jointly bivariate normal and annotate *Z* and $Z^{\perp\!\!\!\perp}$ as the bivariate random variables when *X* and *Y* are correlated or not respectively.

$$Z \sim N\left(\begin{pmatrix}\mu_X\\ \mu_Y\end{pmatrix}, \begin{bmatrix}\sigma_X^2 & \rho\sigma_X\sigma_Y\\ \rho\sigma_X\sigma_Y & \sigma_Y^2\end{bmatrix}\right), \qquad Z^{\perp\!\!\!\perp} \sim N\left(\begin{pmatrix}\mu_X\\ \mu_Y\end{pmatrix}, \begin{bmatrix}\sigma_X^2 & 0\\ 0 & \sigma_Y^2\end{bmatrix}\right)$$

- The mutual information between Z and $Z^{\perp\!\!\!\perp}$ is

$$MI = D_{KL}(Z||Z^{\perp\!\!\!\perp}) = -\frac{1}{2}\log(1-\rho^2)$$

D_{KL} means the Kullback-Leibler divergence, and ρ is exactly the MCC between *X* and *Y*.

- We define MICA in two-study scenario as

$$MICA = -\frac{1}{2}log(1-\rho_+^2)$$

where $\rho_+ = \rho \cdot \mathbb{1}(\rho > 0) + 0 \cdot \mathbb{1}(\rho \leq 0)$, and $\mathbb{1}$ is the indication function.

Mutual Information Concordance AnalysIs (MICA)

- For *S* studies, we have $Z \sim N(\boldsymbol{\mu}, \Sigma)$ and $Z^{\perp\!\!\!\perp} \sim N(\boldsymbol{\mu}, \Sigma^{\perp\!\!\!\perp})$, where

$$\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_S)^T$$

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \cdots & \rho_{1,S_+}\sigma_1\sigma_S \\ \vdots & \ddots & \vdots \\ \rho_{S,1_+}\sigma_S\sigma_1 & \cdots & \sigma_S^2 \end{bmatrix}, \qquad \Sigma^{\perp\!\!\!\perp} = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_S^2 \end{bmatrix}$$

- We can define the MICA in the multiple study cases (Watanabe, 1960)

$$MICA = D_{KL}(Z||Z^{\perp\!\!\!\perp}) = -\frac{1}{2}log\left(\frac{|\Sigma|}{|\Sigma^{\perp\!\!\!\perp}|}\right) = -\frac{1}{2}\left(log|\Sigma| - \sum_{s=1}^{S} log\,\sigma_s^2\right)$$

- The hypothesis test HS_{MICA} for MICA is

$$\{H_0\colon \forall\, \rho_{ij} \le 0\ vs.\, H_A\colon \exists\, \rho_{ij} > 0\}$$

All or part of the studies contain a consistent pattern simultaneously.

Permutation test

We use θ to denote four statistics (MCC, min-MCC, MICA)

1. Compute statistics θ_g for gene *g.*
2. Permutate the group label *B* times and calculate the permutated statistics $\theta_g^{(b)}$, where $1 \leq b \leq B$.
3. Calculate the p-value of θ_g

$$p(\theta_g) = \frac{\sum_{b=1}^{B} \sum_{g'=1}^{G} I(\theta_{g'}^{(b)} \geq \theta_g)}{G \cdot B}$$

4. Obtain the p-values $p(\theta_g)$ for each gene where $1 \leq g \leq G$ and estimate q-values for *G* genes using Benjamin-Hochberg (FDR) procedure. ($p_{(i)}$ is ordered *i*-th p-value)

$$q_{(i)} = min\{\min_{\{j \geq i\}} \left\{ \frac{G \cdot p_{(j)}}{j} \right\}, 1\}$$

Simulation

- We conduct the same simulation with the MCC study (Lu *et al.*, 2010) to identify the genes showing concordant patterns for three classes among three studies.
- 2,000 genes from four expression patterns are simulated for each study.
  - 300 genes (category I) have concordant expression across three studies
  - 100 genes (category II) have discordant expression across three studies
  - 100 genes (category III) have concordant expression in study 1 and 2 only
  - 1,500 genes (category Null) do not include any signals
- One gene with a q-value < 0.05 is seen as informative.

Simulation

- MICA tends to detect more genes comparing to min-MCC.
- MICA can detect the genes with partially shared concordant expression.

Effect size	Methods	I (300)	II (100)	III (100)	Null (1500)
0.5	min-MCC	208.74	0.15	12.84	8.91
	MICA	236.61	6.65	37.22	11.20
0.6	min-MCC	266.01	0.07	17.94	11.37
	MICA	284.45	10.45	61.88	14.18
0.7	min-MCC	290.15	0.01	22.06	12.47
	MICA	297.51	13.26	81.90	15.62

The average number of detected genes which show the concordant expression pattern.

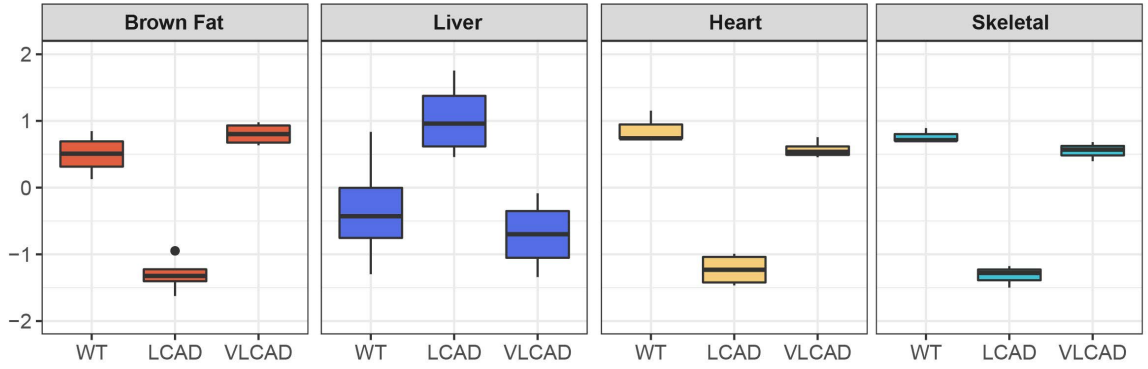
Study types	Min-MCC	MICA
↗ ↗ ↗	✓	✓
↗ ↘ ↕	X	X
↗ ↗ –	X	✓
– – –	X	X

Real application: mouse metabolism data analysis

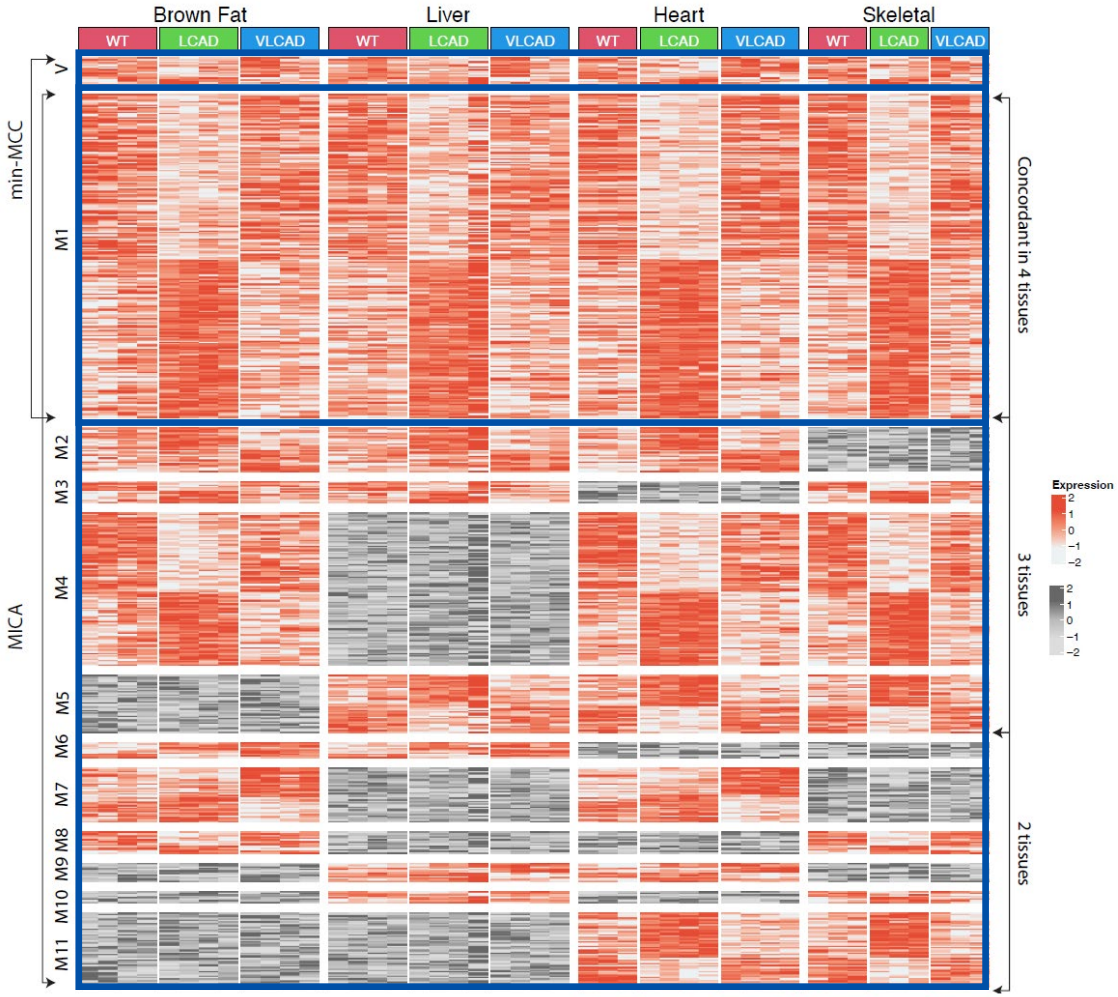
- Microarray data of 3 genotypes of mice: wild-type, LCAD knock-out, and VLCAD knock-out.
  - VLCAD deficiency: common energy metabolism disorder in children
  - LCAD deficiency: impaired fatty acid oxidation and develop a disease similar to other mitochondrial disorders.
- 4 types of tissues (brown fat, skeletal, liver, and heart) from each of the 12 mice.
- Genes with little information content were filtered out, leaving 4288 genes for analysis.

Real application: mouse metabolism data analysis

- A total of 1,394 concordant genes were identified through MICA analysis (q-value < 0.05), and they were further classified according to the post-hoc pattern.



- *Blvrb*, which is related to metabolism and functions in liver, showed the largest MICA statistic, while min-MCC failed to detect it.
- It exhibits the highest gene expression in the liver among multiple tissues (GTEx), suggesting unique liver-specific metabolic functions.



Real application: EstroGene

- The EstroGene project focuses on improving the understanding of the estrogen receptor and its role in the development of breast cancer and aims to document and integrate the publicly available estrogen-related sequencing datasets.
- We considered studies that included gene expression data (microarray and RNA-seq) and limited our analysis to the samples with estrogen receptor positive (ER+) treated with estradiol (E2) for varying durations.
- We combined the samples by cell line and sequencing technology.
- Treatment durations categories:
  - Short (< 6 hours)
  - Medium (>= 6 hours and <= 24 hours)
  - Long (> 24 hours)

Real application: EstroGene

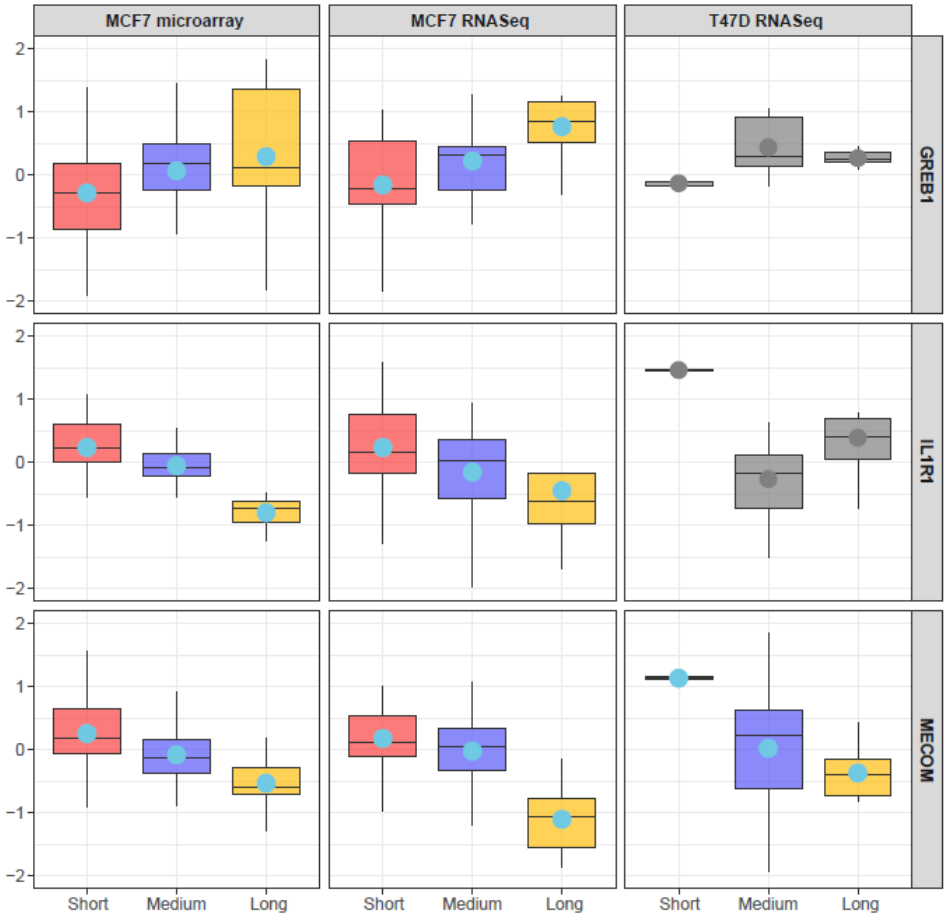
- *GREB1* and *IL1R1* were widely reported as E2 activated and repressed genes.
- *MECOM* was the only gene identified by MICA and min-MCC simultaneously which was not recognized as a biomarker for E2 treatment.

LISA is used to determine the transcription factors (TF) and chromatin regulators related to concordant genes.

Transcription Factor	p-value
SMC1A	3.25E-67
DPF1	7.35E-64
CTCF	3.90E-56
ZMYM3	4.50E-54
NFIA	1.01E-51
ESR1	1.27E-48
BATF3	6.89E-44
MED1	2.46E-43
T	1.80E-31
FOXA1	5.99E-23

Associated with E2

Associated with topologically associating domain

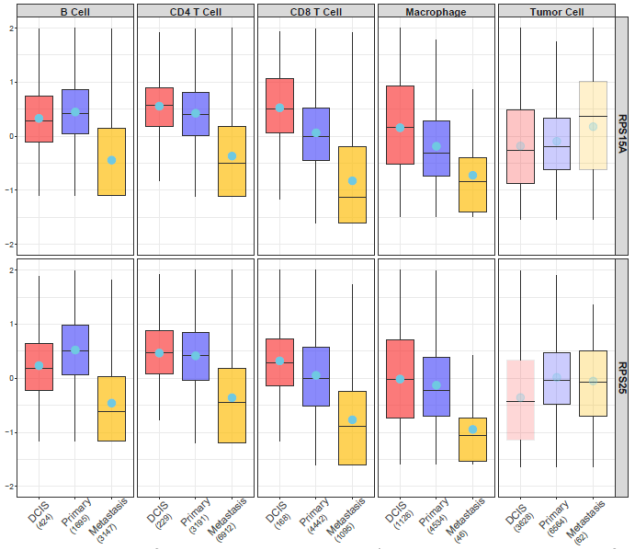


Higher *MECOM* gene expression was associated with worse hazard ratio (HR) in terms of OS (HR = 2.27, p-value = 0.048) and RFS (HR = 3.34, p-value = 0.015).

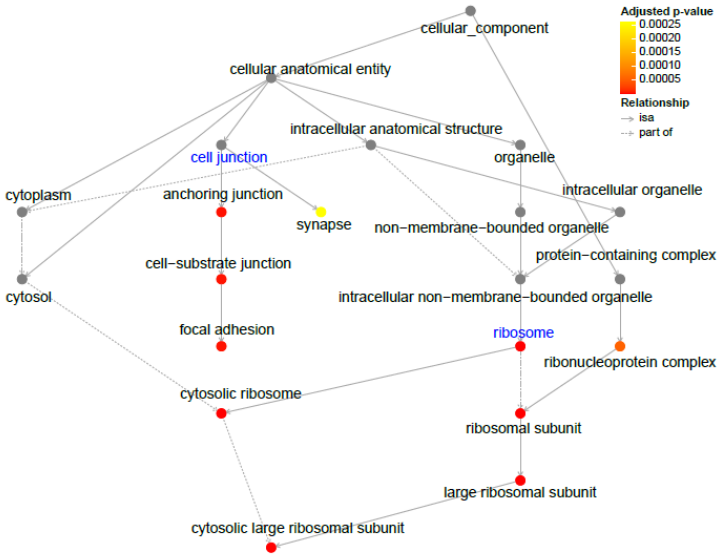
Real application: tumor progression biomarker detection

- We focused on scRNA-seq breast cancer studies.
- **Stages**: ductal carcinoma in situ (DCIS) (N = 5), primary tumor (N = 5), lymph node metastasis (N = 2).
- **Cell types**: B cell, CD4 T cell, CD8 T cell, Macrophage and tumor cell.

- **Tasks**: Identify immune-tumor discordant genes as they progress from DCIS to primary and metastatic stages.
- 198 genes detected.



Expression patterns of RPS15A and RPS25 (related to ribosomal functions)



Enrichment analysis of the immune-tumor discordant genes

Conclusion

- A two-step framework MICA, including an overall and post-hoc pair-wise analysis, is a novel algorithm to detect multi-class concordant biomarker when integrating multiple omics studies.
- Available in bioRxiv https://doi.org/10.1101/2024.06.11.598484 and GitHub https://github.com/jianzou75/MICA

- Possible extensions:
  - It can be generalized to the dependence structure considering the possible relationship among different genes.

Acknowledgement

Advisors:

- Dr. George C. Tseng, University of Pittsburgh
- Dr. Steffi Oesterreich (co-advisor), University of Pittsburgh Medical Center
- Dr. Adrian V. Lee (co-advisor), University of Pittsburgh Medical Center

Collaborators:

- Dr. Zheqi Li, Dana-Farber Cancer Institute
- Dr. Neil Carleton, University of Pittsburgh Medical Center