bioRxiv preprint doi: https://doi.org/10.1101/2024.06.11.598484; this version posted June 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Mutual information for detecting multi-class biomarkers when integrating multiple bulk or single-cell transcriptomic studies

Jian Zou[1], Zheqi Li[2,3], Neil Carleton[4,5,6], Steffi Oesterreich[4,5,7], Adrian V. Lee[4,5,7], and George C. Tseng[8,*]

[1]Department of Statistics, School of Public Health, Chongqing Medical University, Chongqing, 400016, Chongqing, China

[2]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, 02215, Massachusetts, USA

[3]Department of Medicine, Harvard Medical School, Boston, 02215, Massachusetts, USA

[4]Women's Cancer Research Center, UPMC Hillman Cancer Center (HCC), Pittsburgh, 15232, Pennsylvania, USA

[5]Magee-Womens Research Institute, Pittsburgh, 15213, Pennsylvania, USA

[6]Medical Scientist Training Program, School of Medicine, University of Pittsburgh, Pittsburgh, 15213, Pennsylvania, USA

[7]Department of Pharmacology & Chemical Biology, University of Pittsburgh, Pittsburgh, 15213, Pennsylvania, USA

[8]Department of Biostatistics, University of Pittsburgh, Pittsburgh, 15213, Pennsylvania, USA

[*]Corresponding author: ctseng@pitt.edu

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.11.598484; this version posted June 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Motivation: Biomarker detection plays a pivotal role in biomedical research. Integrating omics studies from multiple cohorts can enhance statistical power, accuracy and robustness of the detection results. However, existing methods for horizontally combining omics studies are mostly designed for two-class scenarios (e.g., cases versus controls) and are not directly applicable for studies with multi-class design (e.g., samples from multiple disease subtypes, treatments, tissues, or cell types).

Results: We propose a statistical framework, namely Mutual Information Concordance Analysis (MICA), to detect biomarkers with concordant multi-class expression pattern across multiple omics studies from an information theoretic perspective. Our approach first detects biomarkers with concordant multi-class patterns across partial or all of the omics studies using a global test by mutual information. A post hoc analysis is then performed for each detected biomarkers and identify studies with concordant pattern. Extensive simulations demonstrate improved accuracy and successful false discovery rate control of MICA compared to an existing MCC method. The method is then applied to two practical scenarios: four tissues of mouse metabolism-related transcriptomic studies, and three sources of estrogen treatment expression profiles. Detected biomarkers by MICA show intriguing biological insights and functional annotations. Additionally, we implemented MICA for single-cell RNA-Seq data for tumor progression biomarkers, highlighting critical roles of ribosomal function in the tumor microenvironment of triple-negative breast cancer and underscoring the potential of MICA for detecting novel therapeutic targets.

Availability: https://github.com/jianzou75/MICA

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.11.598484; this version posted June 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1. Introduction

Biomarker detection provides information for early disease diagnosis and is a critical element in biomedical research [Liu et al., 2020]. Integration of data from multiple cohorts is a common approach to improve reliability and statistical power of biomarker detection. If a biomarker demonstrate a similar pattern across multiple studies, it provides robustness and high likelihood of success in subsequent translation and clinical applications. In transcriptomic analysis, differential expression (DE) analysis stands as the predominant method for identifying biomarker expression pattern within individual studies [Costa-Silva et al., 2017, Conesa et al., 2016, McDermaid et al., 2019]. However, the majority of DE techniques are tailored for two-class scenarios (e.g., case versus control), faltering in multi-class scenarios. Popular methods such as limma [Ritchie et al., 2015], although capable of handling multiple classes, primarily offer statistical tests for aggregated differential information in a global sense rather than considering the expression patterns. This limitation highlights a paucity of methods adept at delineating multi-class expression patterns.

To address the integration of omics analysis results from multiple cohorts, two popular approaches emerge in the literature: combining p-values and combing effect sizes. The former has been widely discussed. For example, Fisher's method sums up the log-transformed p-values, and each p-value is assumed to follow standard uniform distribution under the null hypothesis. In addition to Fisher's method, Stouffer [Stouffer et al., 1949], minimum p-value [Tippett et al., 1931], higher criticism [Donoho and Jin, 2004], and adaptive Fisher method [Li and Tseng, 2011] have been developed under this category and are widely used in the omics study integration, such as GWAS [Begum et al., 2012], transcriptomics [Tseng et al., 2012], and methylation [Smith et al., 2018]. Random effects models [DerSimonian and Kacker, 2007], an example of the latter approach, decompose each study's observed treatment effects into the actual effect size and the study-specific noise. These methods, however, have limitations to combine multi-class differential information. P-value combination methods focus on significance without considering multi-class patterns, while effect size combination is restricted to two-class scenarios. To our knowledge, the min-MCC method [Lu et al., 2010] is the only established approach for detecting concordant multi-class biomarkers across multiple studies, The method, however, has two major drawbacks on overlooking the situation when only partial studies share the multi-class pattern and not distinguishing between cases where all pairs of studies have a uniformly low concordance and cases where only one pair has a very low concordance.

To address these challenges, we introduced Mutual Information Concordance Analysis (MICA), a novel

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.11.598484; this version posted June 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

two-stage framework for multi-class biomarker detection combining multiple studies from the perspective of information theory. The first stage employs the generalized mutual information with one-sided correction (gMI_+) to overcome the aforementioned drawbacks. The second stage involves a post-hoc pairwise analysis to identify studies sharing the concordant expression pattern. In 2024, where sequencing studies are ubiquitous, having a method like MICA can be a powerful tool for integrating datasets and enhancing the detection of robust biomarkers. We focus on bulk and single-cell transcriptomic applications in this paper but the method are readily applicable to other omics data types.

As a visual demonstration, Fig 1A shows three example genes $Amacr$, $Pole4$ and $Mcrip2$ detected by MICA to have concordant multi-class (WT: wild type mice; LCAD: LCAD mutated mice; VLCAD: VLCAD mutated mice) pattern across all or partial studies (tissues) (enclosed by red rectangles) while $Mrpl51$ is not detected due to heterogeneous patterns in all four tissues. Post-hoc pairwise analysis in the second stage then determines the studies (enclosed by yellow and blue triangles) that contribute to such concordance for the genes identified in the first step. Specifically, all four tissues share the same multi-class expression pattern in $Amacr$. Only brown fat, heart and liver tissues but not skeletal tissue share the same multi-class expression pattern in $Pole4$. Interestingly, in $Mcrip2$ gene, brown fat and liver share one concordant pattern, while Heart and Skeletal share a different concordant pattern.

The paper is structured as follows. In Section 2, we firstly review the existing method multi-class correlation (MCC) and min-MCC [Lu et al., 2010], followed by a reappraisal from an information theoretic perspective, where we demonstrate improved properties of the MICA framework. A simulation study and three real-world bulk and single-cell transcriptomic applications (Section 3) are conducted to compare min-MCC and MICA. Conclusions and discussions of MICA are included in Section 4.

2. Methods

We assume input data to contain K classess ($K \geq 2$) for detecting multi-class patterns in S transcriptomic studies for integration. For simplicity, we skip subscript of genes and denote x_{ski} as the gene expression for one gene in study s ($1 \leq s \leq S$), class k ($1 \leq k \leq K$), and sample i ($1 \leq i \leq n_{sk}$). For clarity, when discussing the two-study scenario (i.e., $S = 2$), we employ x_{ki} to represent the gene expression in study X, and similarly y_{ki} for study Y.

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.11.598484; this version posted June 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

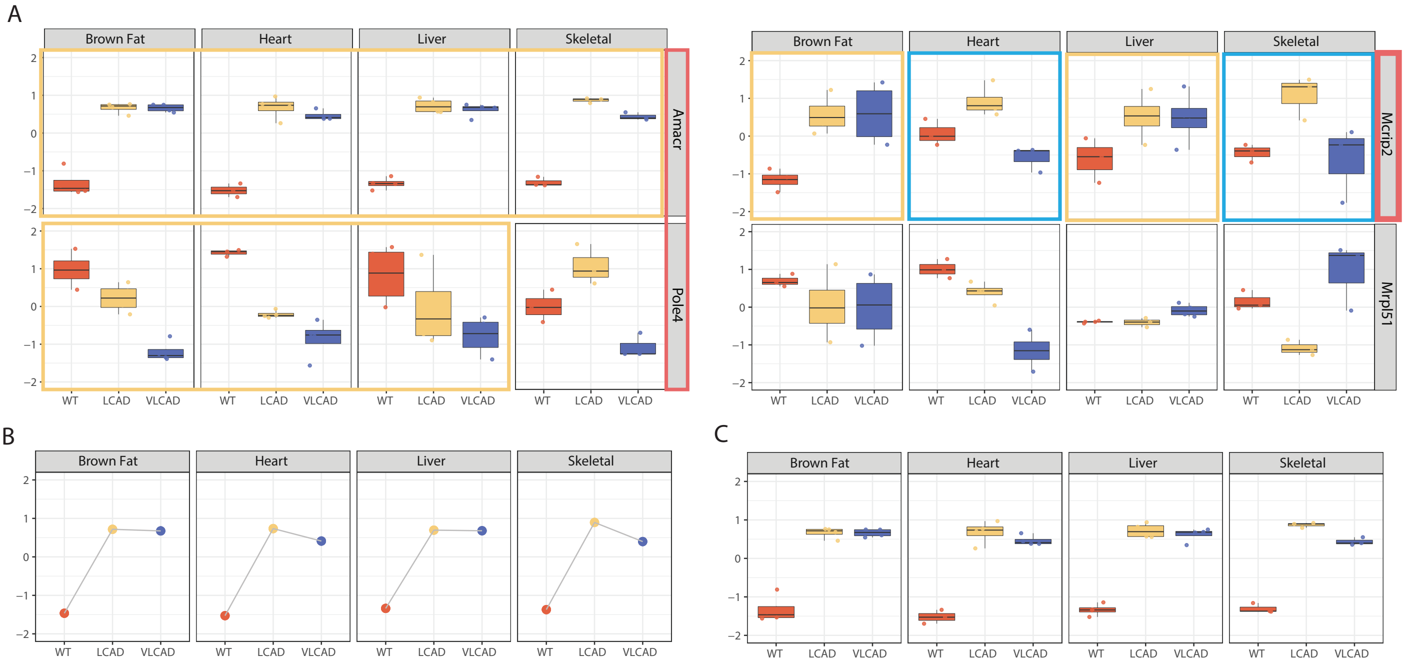

Figure 1: **Depiction of the MICA Framework.** The MICA framework was illustrated using the mouse metabolism dataset. (A) The application of the MICA framework to different gene types. Genes *Amacr*, *Pole*4, and *Mcrip*2 that exhibited consistent patterns across studies were initially identified by a global test using gMI_+ (highlighted by the red triangle). Subsequent post-hoc tests using MI_+ (highlighted by yellow and blue triangles) detected the studies sharing this consistency for each gene. (B) The scenario without replicates for each class within each study, displaying the median expression of the *Amacr* gene within each class and tissue. (C) The scenario with multiple replicates for each class within each study, displaying the expression of all samples for the *Amacr* gene.

2.1 Multi-class correlation (MCC)

We start from the case of two studies ($S = 2$) with expression vectors X and Y. We first consider the simplest case wherein $n_{sk} = 1$ for all the studies s ($1 \leq s \leq S$) and the classes k ($1 \leq k \leq K$) (Fig 1B). Under this circumstance, the intuitive strategy for calculating the concordance (correlation) between study X and Y utilizing Pearson correlation is as follows:

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

where $\bar{x}$ and $\bar{y}$ respectively denote the means of x_{k1} and y_{k1} for all $1 \leq k \leq K$.

When there are replicates within each class from each study ($n_{sk} > 1$, Fig 1C), Pearson correlation is no longer viable. For study X, the observed gene expression x_{ki} is assumed to be obtained from $X_k \sim N(\mu_{X_k}, \sigma^2_{X_k})$, where $X_k \perp\!\!\!\perp X_{k'}$ ($\forall\ k \neq k'$). Therefore, study X can be naturally defined as a mixture

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.11.598484; this version posted June 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

distribution of X_k $(k = 1 : K)$, where each class is assumed to be equally weighted.

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

$$E(X) = \mu_X = \frac{1}{K} \sum_{k=1}^{K} \mu_{X_k}$$

$$Var(X) = \sigma_X^2 = \frac{1}{K} \sum_{k=1}^{K} (\sigma_{X_k}^2 + \mu_{X_k}^2) - \mu_X^2$$

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.

$$\hat{\mu}_{X_k} = \sum_{j=1}^{n_{X_k}} x_{kj}/n_{X_k}$$

$$\hat{\sigma}_{X_k}^2 = \sum_{j=1}^{n_{X_k}} (x_{kj} - \hat{\mu}_{X_k})^2/n_{X_k}$$

Multi-class correlation (MCC) is therefore defined as

$$\begin{aligned} MCC_{(X,Y)} = \rho_{(X,Y)} &= \frac{E(XY) - EX \cdot EY}{\sqrt{Var(X) \cdot Var(Y)}} \\ &= \frac{(\frac{1}{K} \cdot \sum_{k=1}^{K} \mu_{X_k}\mu_{Y_k} - \mu_X \cdot \mu_Y)}{\sigma_X \cdot \sigma_Y} \end{aligned}$$

For multiple S studies $(S > 2)$, min-MCC [Lu et al., 2010] is then defined as the minimum value of MCC statistics across all the pair-wise study combinations:

$$\textit{min-MCC} = min_{U \neq V}(MCC_{(U,V)})$$

The hypothesis test HS_A for min-MCC to detect concordant expression pattern across all S studies is H_0: $\exists\ \rho_{ij} \leq 0$ vs. H_A: $\forall\ \rho_{ij} > 0$, where ρ_{ij} represents the measurement of concordance in the multi-class pattern between study i and j. In addition to computational burden when S is large, min-MCC has two drawbacks. First, it neglects the situation when the concordant multi-class pattern only exists in partial studies due to its stringent requirement for consistency across all studies. Second, it cannot differentiate between scenarios where all study pairs have uniformly low concordance and scenarios where only one pair has very low concordance, which can lead to misinterpretations.

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.11.598484; this version posted June 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2.2 Mutual information concordance analysis (MICA) 105

To overcome the issues above, we revisit this problem from the aspect of information theory. We assumed X and Y to be jointly bivariate normal and denote Z and $Z^{\perp\!\!\!\perp}$ as the bivariate random variables when X and Y are correlated ($\rho \neq 0$) or no correlation respectively.

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

Therefore, we can define the mutual information between Z and $Z^{\perp\!\!\!\perp}$ as

$$
\begin{aligned}
MI_{(X,Y)} &= D_{KL}(Z||Z^{\perp\!\!\!\perp}) \\
&= \frac{1}{2}\Big(\mathrm{tr}(\Sigma_{Z^{\perp\!\!\!\perp}}^{-1}\Sigma_Z) + (\mu_{Z^{\perp\!\!\!\perp}} - \mu_Z)^T\Sigma_{Z^{\perp\!\!\!\perp}}^{-1}(\mu_{Z^{\perp\!\!\!\perp}} - \mu_Z) \\
&\quad -k - \log\left(\frac{|\Sigma_Z|}{|\Sigma_{Z^{\perp\!\!\!\perp}}|}\right)\Big) \\
&= \frac{1}{2}\left(\mathrm{tr}\left(\begin{bmatrix}\frac{1}{\sigma_X^2} & 0 \\ 0 & \frac{1}{\sigma_Y^2}\end{bmatrix}\begin{bmatrix}\sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2\end{bmatrix}\right)\right. \\
&\quad \left. -2 - \log\left(\frac{\sigma_X^2\sigma_Y^2 - (\rho\sigma_X\sigma_Y)^2}{\sigma_X^2\sigma_Y^2}\right)\right) \\
&= -\frac{1}{2}\log(1-\rho^2)
\end{aligned}
$$

D_{KL} means the Kullback-Leibler divergence, and ρ is exactly the MCC between X and Y. To be consistent with MCC and limits to the positive correlation, we define the one-sided corrected mutual information (MI_+) as

$$
MI_{(X,Y)+} = -\frac{1}{2}\log(1-(\rho_+)^2)
$$

where $\rho_+ = \rho \cdot \mathbb{1}_{\rho>0}$. 106

In the two-study scenario, we can find that MI_+ is equivalent to MCC, but it is more straightforward to generalize to more than two studies. For S studies, we have $Z \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $Z^+ \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}^+)$, and

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.11.598484; this version posted June 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$Z^{\perp\!\!\!\perp} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{\perp\!\!\!\perp})$, where

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

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \cdots & \rho_{1,S}\sigma_1\sigma_S \\ \vdots & \ddots & \vdots \\ \rho_{1,S}\sigma_1\sigma_S & \cdots & \sigma_S^2 \end{bmatrix}$$

$$\boldsymbol{\Sigma}^{+} = \begin{bmatrix} \sigma_1^2 & \cdots & \rho_{1,S_+}\sigma_1\sigma_S \\ \vdots & \ddots & \vdots \\ \rho_{1,S_+}\sigma_1\sigma_S & \cdots & \sigma_S^2 \end{bmatrix}$$

$$\boldsymbol{\Sigma}^{\perp\!\!\!\perp} = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_S^2 \end{bmatrix}$$

Therefore, we can define the concordance measurement for multiple studies, which is the generalized mutual information (gMI), also known as total correlation [Watanabe, 1960].

$$\begin{aligned} gMI_{(X_1,X_2,...,X_S)} = D_{KL}(Z||Z^{\perp\!\!\!\perp}) &= -\frac{1}{2}\log\left(\frac{|\boldsymbol{\Sigma}|}{|\boldsymbol{\Sigma}^{\perp\!\!\!\perp}|}\right) \\ &= -\frac{1}{2}\left(\log|\boldsymbol{\Sigma}| - \sum_{s=1}^{S}\log\sigma_s^2\right) \end{aligned}$$

Similarly, to only consider the positive concordance, we define the generalized one-sided corrected mutual information (gMI_+) as

$$\begin{aligned} gMI_{(X_1,X_2,...,X_S)+} &= D_{KL}(Z^{+}||Z^{\perp\!\!\!\perp}) \\ &= -\frac{1}{2}\left(\log|\boldsymbol{\Sigma}^{+}| - \sum_{s=1}^{S}\log\sigma_s^2\right) \end{aligned}$$

2.3 Procedure of concordant biomarker detection

Based on the generalized mutual information above, the mutual information concordance analysis (MICA) is developed in two steps.

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.11.598484; this version posted June 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2.3.1 Global test for concordant biomarker detection

In the first step, we first deploy the generalized one-sided corrected mutual information (gMI_+) to ascertain if a gene exhibits concordant multi-class pattern across multiple studies. This determination hinges on the hypothesis test, namely HS_B, H_0: $\forall\ \rho_{ij} \leq 0$ vs. H_A: $\exists\ \rho_{ij} > 0$. The permutation test (see Section 2.3.3) is employed for assessing p-values and q-values of this global test for each gene.

2.3.2 Post-hoc test to detect subset of studies with concordant multi-class pattern

If the null hypothesis in the global test is rejected, we proceed to identify the subest of studies with concordant multi-class pattern. Specifically, we select the largest subset of studies where every pair of studies in it shows a significant p-value, indicating concordance. For this purpose, we employ the one-sided corrected mutual information (MI_+) to examine all feasible pairs of studies (i, j). This analysis is conducted under the hypothesis setting HS_C for the study pair i and j by H_0: $\rho_{ij} \leq 0$ vs. H_A: $\rho_{ij} > 0$, with p-values inferred by permutation test (see Section 2.3.3).

2.3.3 Permutation test for the four statistics

Permutation test is designed to obtain the significance levels for MI_+ and gMI_+ since an analytical solution is not achievable. We use θ to denote them for using permutation test to evaluate p-values and q-values. To compare with existing methods, we use the same permutation analysis for MCC and min-MCC.

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,

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

4. (If multiple genes are screened simultaneously) Obtain the p-values $p(\theta_g)$ for each gene where $1 \leq g \leq G$, and estimate q-values for gene i using Benjamini-Hochberg procedure. ($p_{(j)}$ is ordered j-th p-value)

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

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.11.598484; this version posted June 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3. Results

In this section, we first applied MICA for simulations to evaluate the type I error and power of multi-class biomarker detection. The method is then applied to two bulk transcriptome applications: mouse metabolism-related studies [Lu et al., 2010], and estrogen treatment expression profiles [Li et al., 2023]. In the third application, we investigate the capability of MICA in single cell RNA-Seq data for tumor progression biomarkers detection. [Tokura et al., 2022, Wu et al., 2021a, Xu et al., 2021].

3.1 Simulation

We devised simulations involving five distinct types of genes from four studies (details in Supplement Table S1). Gene Type I represents perfect concordance with all four studies. In Gene Type II, studies 1, 2 and 3 show concordant expression. Gene Type III highlights pairwise concordance, showing agreement between studies 1 and 2 and a separate concordance between studies 3 and 4. Finally, Gene Type IV contains noises across all four studies, without any discernible pattern. There are 10 biological replicates within each class from each study, and the simulation is repeated for 500 times for evaluation. We then compare the performance of min-MCC and MICA in terms of type I error control and power.

MICA outperformed min-MCC in terms of signal detection power. For Gene Type I, where all studies were concordant, MICA achieved a detection rate of 0.836 against 0.638 for min-MCC at the p-value threshold of 0.05. In the more complex scenarios of Gene Types II and III, where only part of the studies were concordant, MICA maintained performance (0.748 in Gene Type II, 0.936 in Gene Type III), while min-MCC faltered (0.184 in Gene Type II, 0.174 in Gene Type III). Figure 2A-C provide a direct comparison of the respective powers of MICA and min-MCC at varying p-value thresholds across the three gene types. For the negative control, Gene Type IV, MICA exhibited an error rate of 0.058, slightly higher than the 0.048 error rate observed in min-MCC.

Following the assessment of individual genes, we expanded the simulation to encompass gene expression matrices for a genome-wide power comparison. We prepared 2000 genes expression for each dataset, distributed evenly across four gene types with 500 genes each. A total of 200 datasets was simulated for this analysis. After preparing the receiver operating characteristic (ROC) curves for each simulated dataset, the averaged area under the curve (AUC) of MICA was 0.97 (sd = 0.004), in contrast to 0.59 (sd = 0.02) for min-MCC. Figure 2D shows an ROC curve of the data aggregated across 200 simulated

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.11.598484; this version posted June 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

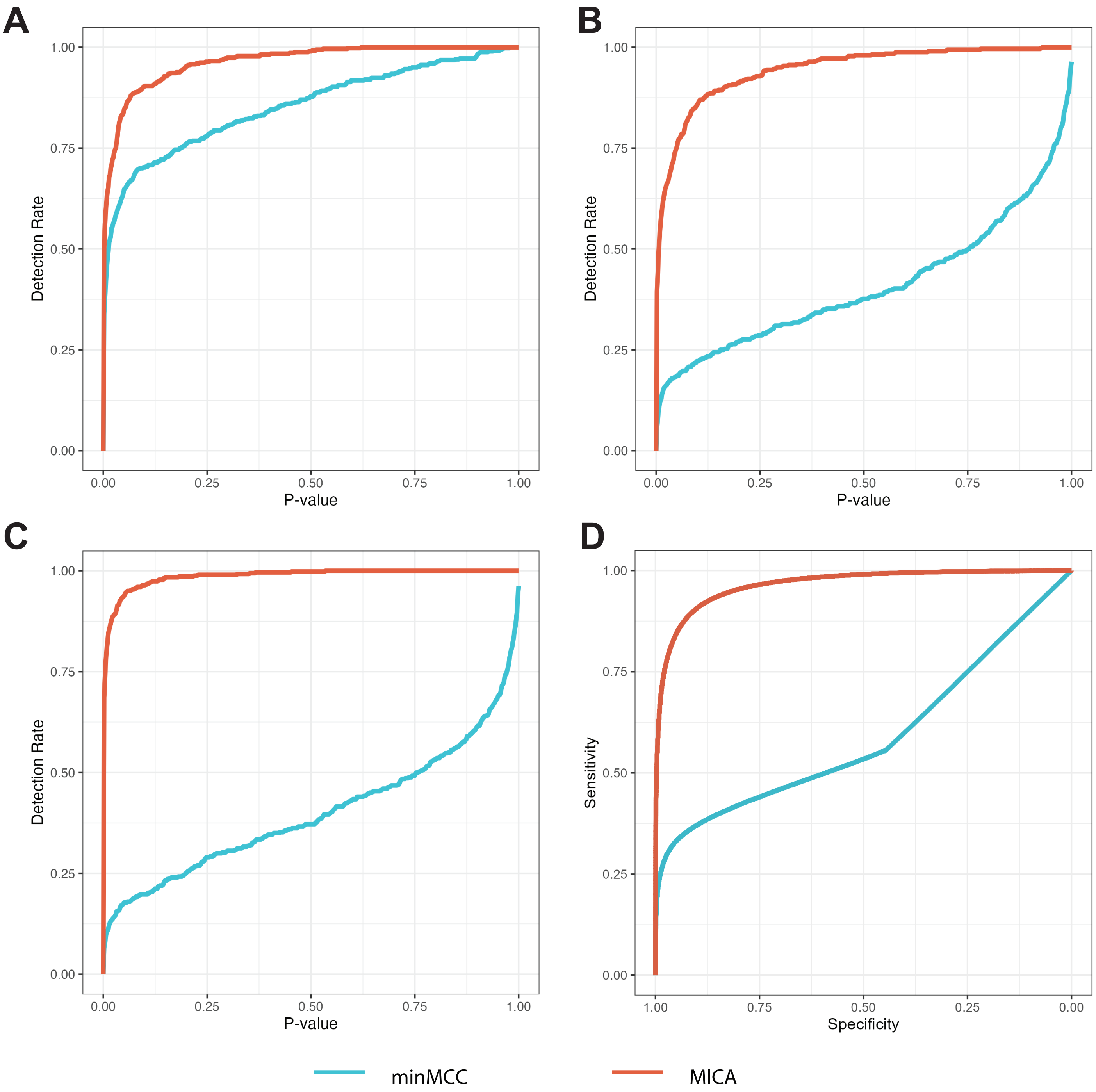

Figure 2: **Comparative performance assessment of MICA and min-MCC via simulation.** MICA consistently outperforms min-MCC in signal detection across Gene Types I-III. Genome-wide analysis further corroborates superior efficacy of MICA in biomarker identification. (A) Statistical power analysis in Gene Type I. (B) Statistical power analysis in Gene Type II. (C) Statistical power analysis in Gene Type III. (D) Aggregated ROC curves from 200 simulated datasets.

datasets, substantiating the superior performance of MICA. Employing a q-value threshold of 0.05, MICA achieved the sensitivity of 0.79 and the specificity of 0.97, whereas min-MCC has sensitivity and specificity at 0.23 and 0.99, respectively. 156 157 158

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.11.598484; this version posted June 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3.2 Application 1: mouse metabolism bulk transcriptomic studies

In this section, we applied MICA to the study analyzed in the min-MCC paper [Lu et al., 2010]. Bulk expression profiles are measured in mice with three genotypes (wild-type, LCAD knock-out, and VLCAD knock-out). LCAD deficiency is associated with impaired fatty acid oxidation, and VLCAD deficiency is associated with energy metabolism disorders in children. Microarray experiments were conducted on tissues from 12 mice (four mice per genotype) including brown fat, liver, heart, and skeletal. The expression changes across genotypes were studied, and genes with little information content were filtered out to have 4,288 genes remained for downstream analysis. Four samples were identified with quality defects and were excluded from further analysis.

A total of 730 concordant genes were identified through MICA analysis, while min-MCC only detected 245 concordant genes (q-value < 0.01), suggesting tissue heterogeneity. To evaluate the necessity of MICA, we classified the detected genes into three subsets: genes identified by min-MCC only (V), genes detected by min-MCC and MICA simultaneously (M1), and genes identified only by MICA (M2-M11). In the third subset, we classified genes into 10 modules based on post-hoc MICA results and clustered genes within the same module using K-means. The number of clusters was determined using the NbClust R package [Charrad et al., 2014].

Figure 3 and Supplement Figure S1 display the expression patterns for each gene module. Genes in Module V exhibited ambiguous expression patterns. Meanwhile, genes in Module M1, which were partitioned into two clusters, exhibited high concordance across all four tissues. We performed a QIAGEN Ingenuity Pathway Analysis (IPA) [Krämer et al., 2014] on genes in M1. Apart from the pathways known to be associated with LCAD and VLCAD [Nsiah-Sefaa and McKenzie, 2016], such as oxidative phosphorylation and acyl-CoA hydrolysis, metabolism and mitochondria-related pathways like arsenate detoxification, tetrapyrrole synthesis, and heme biosynthesis were also detected (Supplement Table S2). Additionally, genes in M1 showed more similar expression patterns in wild-type and VLCAD knock-out mice compared to LCAD knock-out mice, supporting previous findings that LCAD knock-out mice exhibit a more severe phenotype than VLCAD knock-out mice [Maher et al., 2010].

Modules M2-M11 demonstrated concordance pattern in a subset of tissues that were not detected by the min-MCC method. Among modules concordant in three tissues (M2-M5), Module M4 contained the largest number of genes (107 genes), showing impacts of LCAD and VLCAD knockouts in all tissues except for liver. *Blvrb* in Module M4 displayed the highest MICA statistic (MICA $= 2.31$, p-value $= 0$), although

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.11.598484; this version posted June 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

it was not identified by the min-MCC method (min-MCC = -0.71, p-value = 1) (Supplement Figure S2). *Blvrb* demonstrated lower expression in LCAD knockout samples in brown fat, heart, and skeletal tissues, but its expression was higher in the liver. Though *Blvrb* has no reported direct relation with LCAD and VLCAD, it is involved in metabolism, converting biliverdin to bilirubin in the liver [Consortium et al., 2017]. According to the Human Protein Atlas (proteinatlas.org) and the GTEx database [Lonsdale et al., 2013, Uhlén et al., 2015], *Blvrb* showed the highest gene expression in the liver among multiple tissues, indicating liver-specific functions not seen in the other three tissues.

The IPA analysis on genes in M4 (Supplement Table S2) emphasizes the distinct role of liver and the necessity to identify concordant pattern genes in a subset of tissues/studies. Specifically, 9 of the top 15 pathways, such as superpathway of methionine degradation and guanosine nucleotides degradation III, identified are related to metabolism, highlighting the role of liver.

In summary, MICA significantly outperforms min-MCC by identifying more concordant genes and uncovering tissue-specific gene expression patterns that min-MCC misses. This underscores the necessity of MICA for capturing the complexity of the partially concordant gene expression.

3.3 Application 2: bulk transcriptomic data in the EstroGene project

The EstroGene project [Li et al., 2023] focuses on improving the understanding of the estrogen receptor and its role in the development of breast cancer. It aims to document and integrate the publicly available estrogen-related datasets, including RNA-Seq, microarray, ChIP-Seq, ATAC-Seq, DNase-Seq, ChIA-PET, Hi-C, GRO-Seq and others, to establish a comprehensive database that allows for customized data search and visualization. Specifically, in this case, MICA can help identify genes that are consistently regulated by estradiol (E2) over different time points across multiple studies, which is critical for understanding the dynamics of estrogen receptor signaling in breast cancer.

In this subsection, we only 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) doses greater than 1nM for varying duration. We first combined the samples by cell line and sequencing technology. To further analyze the data, we then classified the treatment duration into three categories: short ($<$ 6 hours), medium ($\geq$ 6 hours and $\leq$ 24 hours), and long ($>$ 24 hours). Finally, we normalized the data for the newly pooled studies using trimmed mean of M values (TMM) [Bullard et al., 2010] followed by ComBat [Johnson et al., 2007] with the study indication as a batch covariate. These

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.11.598484; this version posted June 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

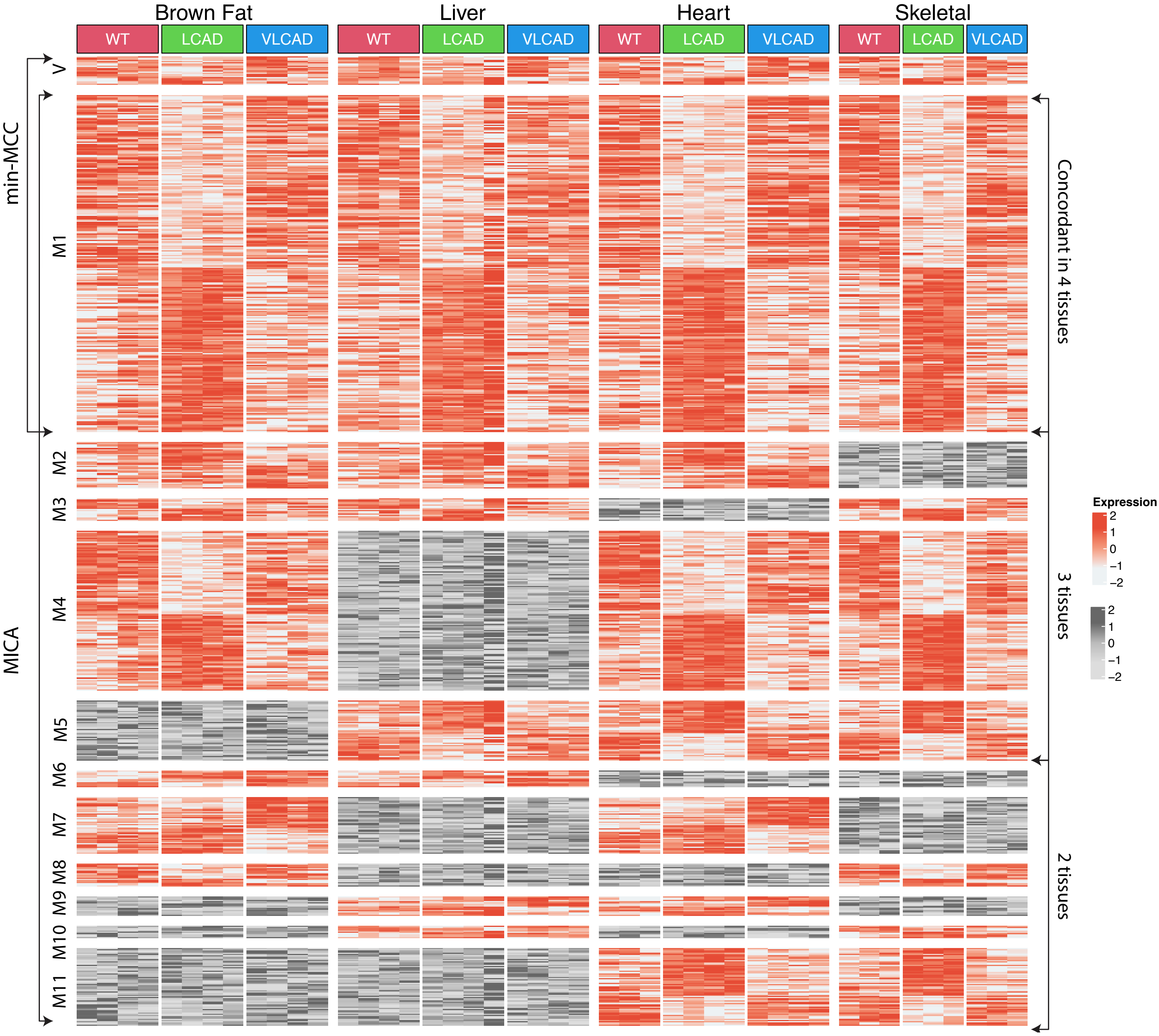

Figure 3: **The heatmap of the gene expression patterns of different gene modules across four tissues in mouse metabolism data analysis.** The rows represent the genes, and the columns represent the samples. V includes genes detected by min-MCC only, while M1 includes genes detected by min-MCC and MICA at the same time. The genes in M2-M11 were identified by MICA alone and categorized by the contributing studies using MICA post-hoc analysis. Studies that contribut-ed to the concordance are shown in red panel, while those that did not are shown in gray.

steps resulted in three pooled studies: MCF7 microarray (25 samples in short treatment, 34 in medium treatment, and 7 in long treatment), MCF7 RNA-Seq (49 in short treatment, 62 in medium treatment, and 10 in long treatment), and T47D RNA-Seq (3 in short treatment, 22 in medium treatment, and 11 in long treatment). 1,983 genes were intersected across multiple platforms for downstream analysis.

We first validated the two well-established benchmark genes, *GREB*1 and *IL*1*R*1, which have been

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.11.598484; this version posted June 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

widely reported as E2 activated and repressed genes [Cheng et al., 2018, Rae et al., 2005, Schaefer et al., 2005, Lavigne et al., 2008]. Figure 4 revealed the up- and down-regulation of $GREB1$ and $IL1R1$ in MCF7 microarray and RNA-Seq studies. However, these trends were not observed in the T47D RNA-Seq study. Specifically, while T47D cells exhibited a decreasing trend in $IL1R1$ gene regulation from short to the combined medium and long durations ($p < 0.05$ from t-test), the trend reversed, showing an increase between medium and long durations ($p < 0.05$ from t-test), and no trend was observed in $GREB1$ gene ($p = 0.85$ from ANOVA test). This inconsistency is likely due to the inherent heterogeneity of breast cancer. Despite the inconsistency across all three studies, MICA evaluated the partial trend as concordant. As a result, MICA identified both genes as concordant with q-values of 0.01 and 0, while the min-MCC detected them with larger q-values of 0.03 and 0.06, respectively.

In addition to validating known markers, we are also able to detect novel biomarkers. For example, $MECOM$ was the only gene identified by MICA and min-MCC with q-values = 0 simultaneously (Figure 4). Prior to our study, $MECOM$ was not recognized as a biomarker for E2 treatment although it is known as a transcriptional regulator and oncogene. Indeed, when we analyzed 1,459 ER+ breast cancer patients in the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) database [Curtis et al., 2012], we observed that higher $MECOM$ gene expression was associated with worse hazard ratio (HR) in terms of overall survival (HR = 2.27, p-value = 0.048) and relapse-free survival (HR = 3.34, p-value = 0.015).

To determine if this association is specific to HR+ tumors, we also performed survival analysis in other subtypes, including triple-negative breast cancer (TNBC) and HER2+ cohorts. In the TNBC cohort (n = 299), we did not observe a significant association (p-value = 0.18 for OS and p-value = 0.49 for RFS). Similarly, in the HER2+ cohort (n = 236), there was no significant association (p-value = 0.44 for OS and p-value = 0.90 for RFS). These findings suggest that the association of $MECOM$ with survival outcomes is specific to HR+ tumors, which could strengthen the link between $MECOM$ and endocrine response.

The potential mechanism of the clinical prognosis could partially be explained by the regulation of estrogen receptor, as we observed several consistent ER binding sites at transcription start sites (TSS) proximity from ChIP-seq data in the EstroGene website. The mechanistic link of $MECOM$ to estrogen receptor and E2 treatment, however, needs further investigation.

In total, MICA identified 403 concordant genes (q-value < 0.05). To gain a deeper understanding of the upstream transcription factors associated with these genes, we applied LISA, an algorithm that uses

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.11.598484; this version posted June 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

chromatin profile and H3K27ac ChIP-seq data to determine the transcription factors (TF) and chromatin regulators related to a given gene set [Qin et al., 2020]. Among the top-ranked TFs (Supplement Table S3), $ESR1$ and $FOXA1$ are the TFs that have previously been reported to be associated with E2 [Chaudhary et al., 2017, Theodorou et al., 2013]. In addition, $SMC1A$ and $CTCF$, the first two candidates, suggests a potential role of topologically associating domain (TAD) in the regulation of these gene [Rinzema et al., 2022, DeMare et al., 2013]. These findings revealed that the E2 response may involve gene regulation through chromatin looping mechanisms. Further experimental studies are needed to fully elucidate the underlying mechanisms.

3.4 Application 3: tumor progression biomarker detection in scRNA-seq breast cancer studies

In this subsection, we apply MICA to a scRNA-Seq dataset to compare three stages ($K = 3$) of triple-negative breast cancer (TNBC) progression using treatment-naive tissues: ductal carcinoma in situ (DCIS) ($N = 5$), primary tumor ($N = 5$), and lymph node metastasis ($N = 2$). Understanding the progression from DCIS, a precursor of invasive breast cancer, to primary tumors and eventually to metastatic disease is crucial for identifying biomarkers of tumor progression. The application of MICA in this case provides valuable insights into the molecular changes driving cancer metastasis, which is essential for developing targeted therapies and improving patient outcomes.

Data were obtained from three publications [Tokura et al., 2022, Wu et al., 2021a, Xu et al., 2021]. We implemented the scATOMIC [Nofech-Mozes et al., 2023] to annotate single cells to five cell types (B cell, CD4 T cell, CD8 T cell, macrophage and tumor cells) for downstream analysis. The distribution of cell types can be found in Supplement Table S4. Within each study, total count normalization was applied [Hao et al., 2023]. We treat the five cell types as independent studies ($S = 5$) and apply MICA to 6,644 genes after preprocessing. 2,703 genes exhibited concordant expression patterns across two or more cell types (q-value < 0.001). Notably, of the 86 genes associated with ribosomal functions, 82 exhibited concordance, which underscores the substantial role of protein synthesis in tumor progression.

In a further analysis, we aimed to identify the immune-tumor discordant genes, which exhibit concordant expression patterns across the first four tumor microenvironment cell types yet discordant patterns in tumor cells, as they progress from ductal carcinoma *in situ* (DCIS) to primary and subsequently to metastatic stages. To achieve this goal, we select from the 2,703 genes using criteria of any post-hoc pairwise

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.11.598484; this version posted June 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

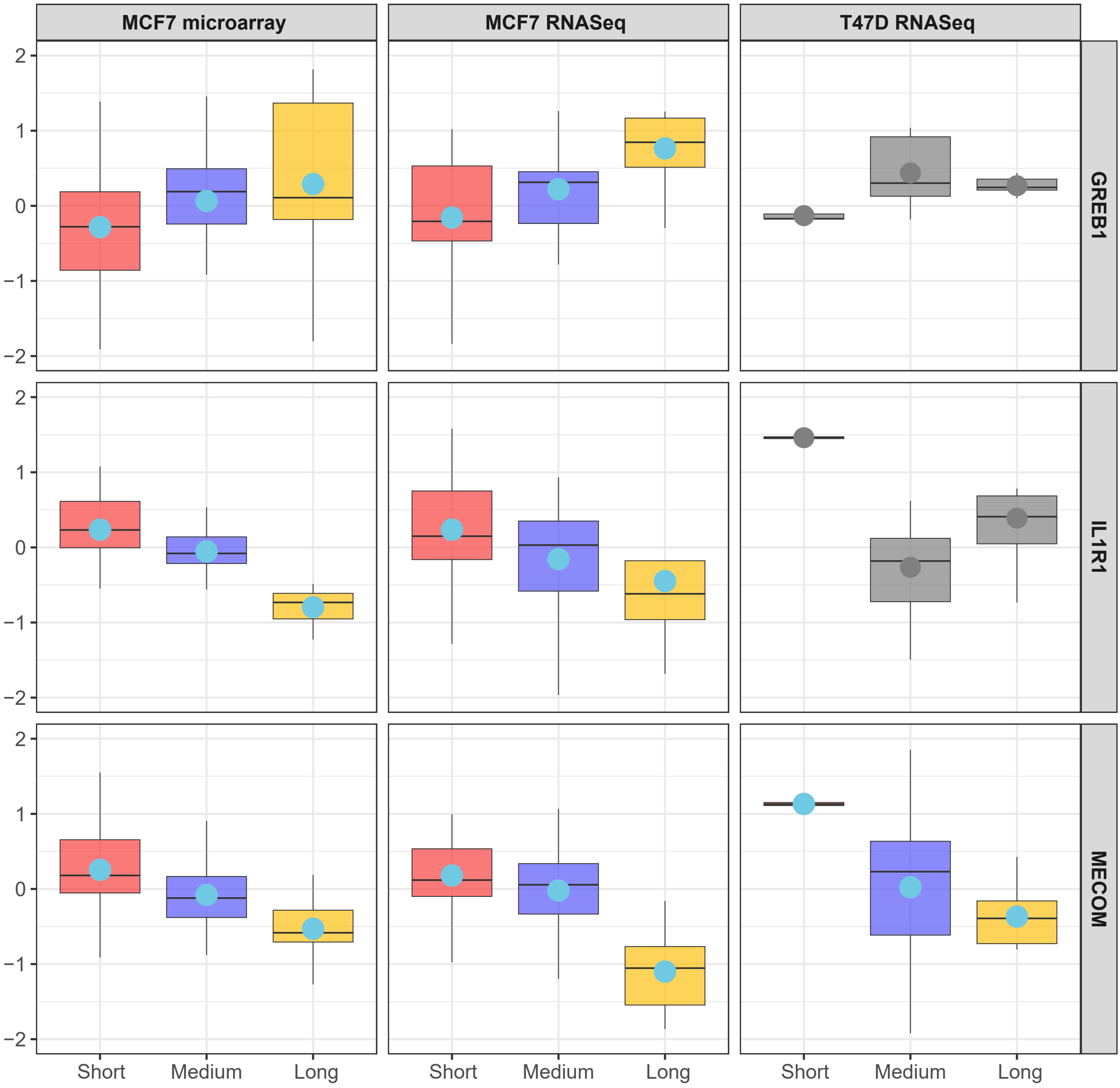

Figure 4: **The expression patterns of *GREB1*, *IL1R1*, and *MECOM* across three data sources.** $GREB1$ and $IL1R1$ are widely reported as E2 activated and repressed genes and were detected by MICA while failed to be identified by min-MCC. $MECOM$ was the only gene detected by MICA and min-MCC simultaneously. The averaged expression is shown as a blue circle.

p-values among immune cell types being less than 0.001 (i.e., all four immune cell types have concordant pattern to each other), and all pairwise p-values between a immune cell type and tumor exceeding 0.5 (i.e. all four immune cell types have discordant pattern to tumor). This analysis detected 198 genes (Supplement Table S5). Figure 5 illustrates the expression patterns of $RPS15A$ and $RPS25$, the first two genes

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.11.598484; this version posted June 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

with the highest gMI_+ statistics. The numbers of cells for each tumor type and each cell type are shown in the x-axis labels. Both genes are related to ribosomal functions, suggesting a hypothesis that protein synthesis is downregulated in immune cells as the tumor progresses, while it is upregulated in tumor cells.

Gene set enrichment analysis [Wu et al., 2021b] of these immune-tumor discordant genes, conducted using the Gene Ontology (GO) knowledge base [Aleksander et al., 2023], revealed significant enrichment in two functional groups: cell junction and ribosome-related pathways (Figure 6). The former is closely associated with the tumor progression, such as altered cell adhesion and migration, while the latter confirms our earlier findings and underscores the importance of protein synthesis in the tumor microenvironment during the progression.

4. Discussion and conclusions

Horizontal integration of multiple transcriptomic studies to identify disease biomarkers is an effective tool for accurate and reproducible detection [Cohn and Becker, 2003, Trikalinos et al., 2008]. To date, min-MCC is the only available method to detect the biomarkers with concordant multi-study multi-class expression patterns [Lu et al., 2010]. However, since min-MCC cannot identify the partially concordant biomarkers and is insensitive to the pairwise high concordance, we revisited this problem from the aspect of information theory and proposed a two-step framework MICA (Mutual Information Concordance Analysis). Both the simulation and real application results demonstrate the superiority of the MICA framework in selecting more informative biomarkers and elucidating underlying disease mechanisms towards translational research.

The three real applications contain a variety of biological and clinical scenarios and demonstrate wide applicability of MICA. In the mouse metabolism example, the biological objective is to detect biomarkers changed in wild type, LCAD or VLCAD mutation ($K = 3$) across four tissues ($S = 4$). Since biomarker pattern may different across different tissues, categorization of detected biomarkers in the heatmap of Figure 3 allows structured biological investigation. In the second EstroGene Project example, we investigate biomarkers with differential expression changes in short, medium and long treatment duration ($K = 3$) in three cell line-platform studies ($S = 3$). Ideally we expect similar expression pattern across studies but we indeed observe different multi-class pattern in different cell lines (e.g., GREB1 and IL1R1 in Figure 4). The third TNBC scRNA-seq example demonstrates an intriguing finding of tumor progression marker detection (DCIS, primary and metastatic tumor; $K = 3$) in five cell types of single cells ($S = 5$). The

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.11.598484; this version posted June 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

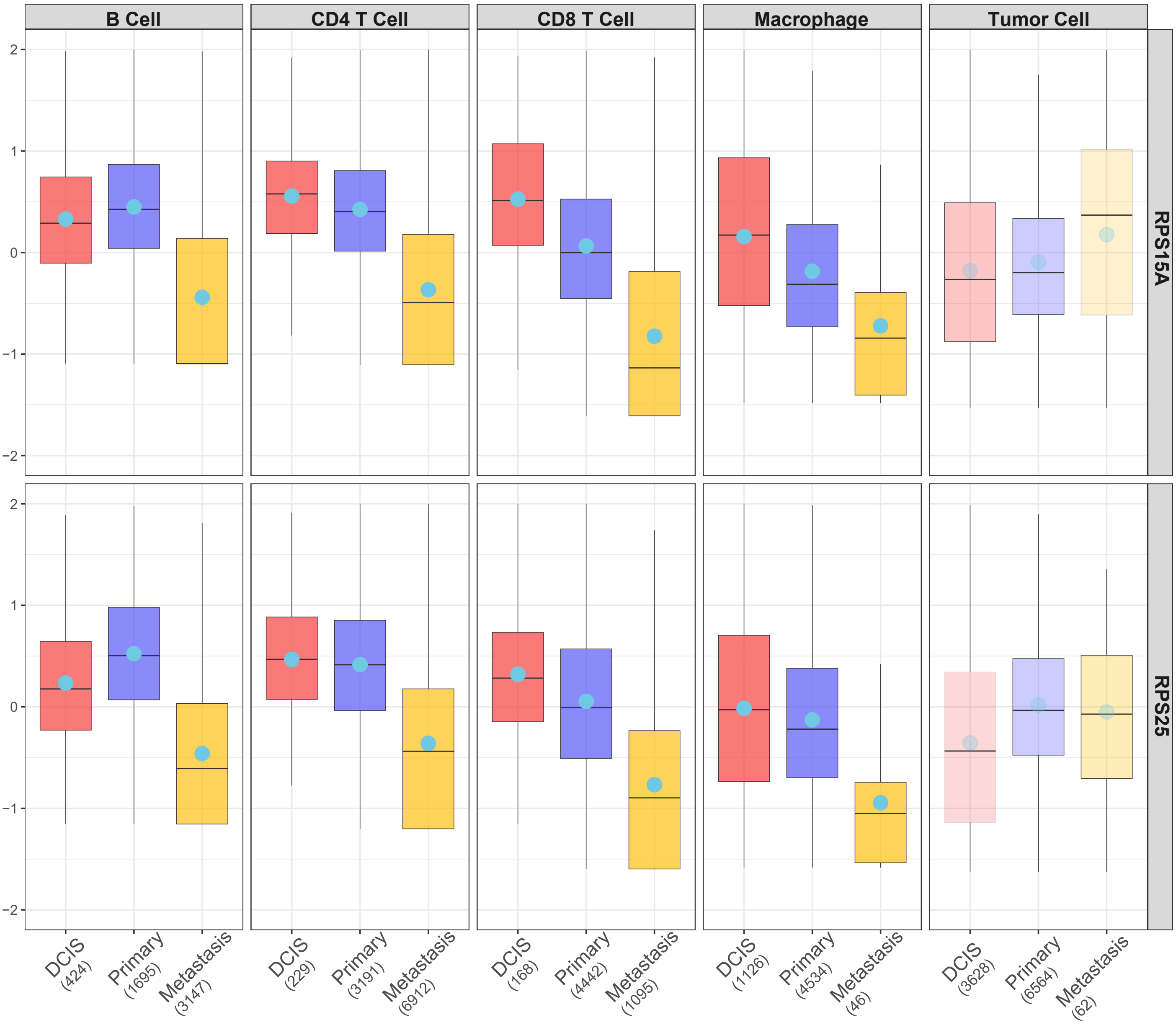

Figure 5: **The expression pattern of** $RPS15A$ **and** $RPS25$**.** $RPS15A$ and $RPS25$ are the top 2 immune-tumor discordant genes with largest gMI_+ statistics. In immune cells, a decreasing expression trend is observed as the tumor progresses, whereas an opposing pattern is evident in tumor cells. The averaged expression is shown as a blue circle. The numbers of cells for each tumor type and each cell type are shown in the parentheses of x-axis labels

result identifies a set of biomarkers with concordant tumor progression pattern in the four immune-related cell types (i.e., B cell, CD4 T cell, CD8 T cell and Macrophage) while almost opposite pattern in tumor cells. We believe the wide range of applications not only demonstrate wide applicability of MICA but also will inspire its novel applications by other researchers.

One advantage of MICA is its scalable computing when K, S and biological replicate sample sizes increase. In the simulation of $K = 3$, $S = 4$ and 30 samples, the computing time for 2000 genes and 500 permutations takes 18.9 minutes using the high performance computing (HPC) with 50 threads parallel

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.11.598484; this version posted June 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

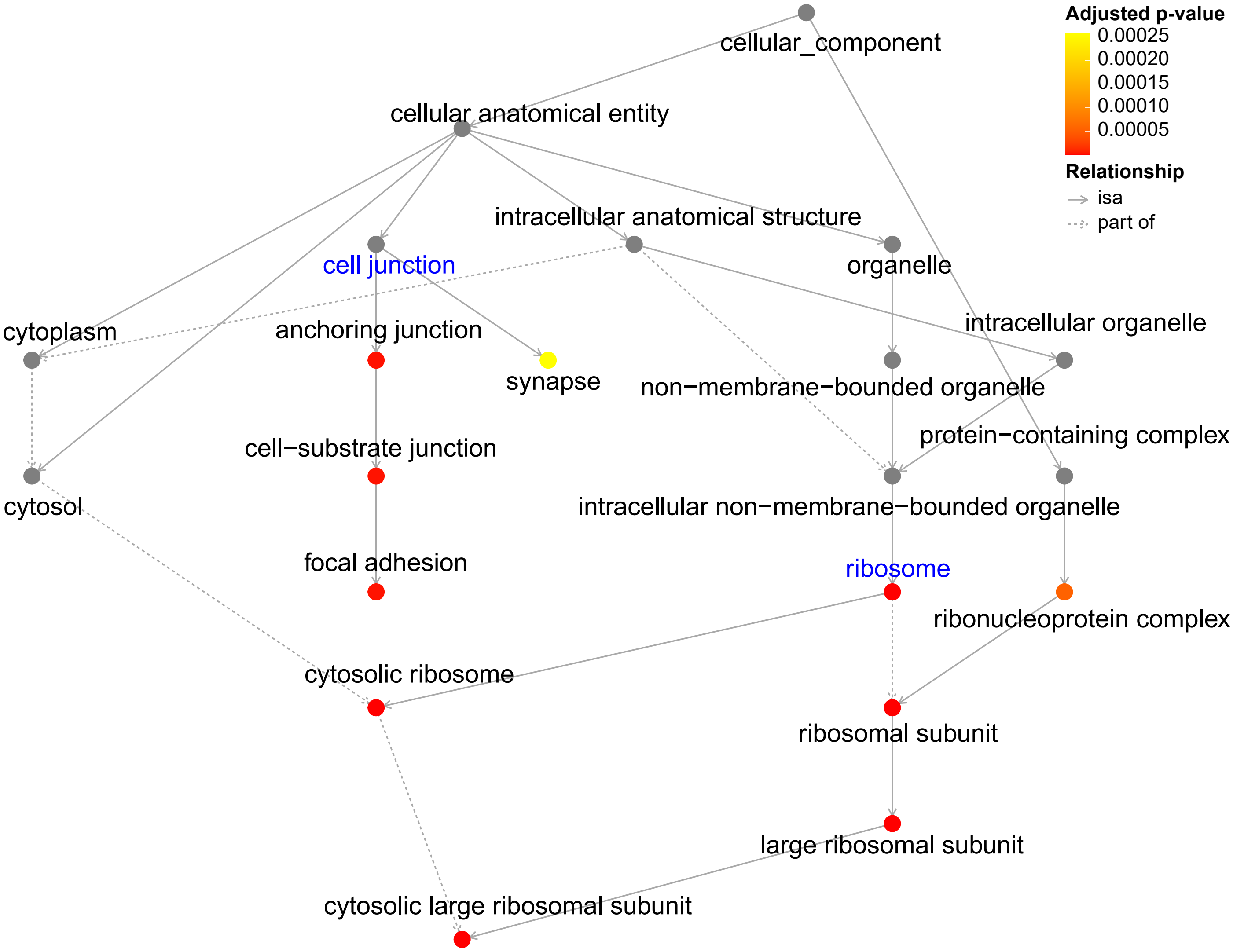

Figure 6: **GO directed acyclic graph for enrichment analysis of immune-tumor discordant genes.** Solid arrows (“isa”) denote that the term at the arrowhead is a subtype of the term at the tail, establishing a hierarchical relationship. Dashed arrows (“part of”) indicate that the term at the arrowhead is an essential part of the term at the tail, signifying structural inclusion. Dot colors represent the adjusted p-values from the enrichment analysis. Notably, cell junction and ribosome-related pathways are significantly enriched among the immune-tumor discordant genes.

design. The current method performs analysis for each gene independently although the permutation scheme keeps gene dependence structure by permuting class labels when generating the null distribution for p-value assessment. An R package, namely MICA, and all programming code are available on GitHub for reproducing figures and results in this paper.

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.11.598484; this version posted June 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Funding

JZ and GCT are partially funded by NIH R01LM014142. This research was supported in part by the University of Pittsburgh Center for Research Computing, RRID:SCR 022735, through the resources provided. Specifically, this work used the HTC cluster, which is supported by NIH award number S10OD028483. This work was supported by the Breast Cancer Research Foundation (AVL and SO]; Susan G. Komen Scholar awards (SAC110021 to AVL and SAC160073 to SO]; the National Cancer Institute (R01CA221303 to SO, R01256161 to AVL; 5F30CA264963-02 to NC).

References

S. A. Aleksander, J. Balhoff, S. Carbon, J. M. Cherry, H. J. Drabkin, D. Ebert, M. Feuermann, P. Gaudet, N. L. Harris, et al. The gene ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031, 2023.

F. Begum, D. Ghosh, G. C. Tseng, and E. Feingold. Comprehensive literature review and statistical considerations for gwas meta-analysis. *Nucleic Acids Research*, 40(9):3777–3784, 2012.

J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit. Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC Bioinformatics*, 11(1):1–13, 2010.

M. Charrad, N. Ghazzali, V. Boiteau, and A. Niknafs. Nbclust: an r package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, 61:1–36, 2014.

S. Chaudhary, B. M. Krishna, and S. K. Mishra. A novel foxa1/esr1 interacting pathway: A study of oncomine™ breast cancer microarrays. *Oncology Letters*, 14(2):1247–1264, 2017.

M. Cheng, S. Michalski, and R. Kommagani. Role for growth regulation by estrogen in breast cancer 1 (greb1) in hormone-dependent cancers. *International Journal of Molecular Sciences*, 19(9):2543, 2018.

L. D. Cohn and B. J. Becker. How meta-analysis increases statistical power. *Psychological Methods*, 8(3): 243, 2003.

A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, et al. A survey of best practices for rna-seq data analysis. *Genome Biology*, 17(1):1–19, 2016.

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.11.598484; this version posted June 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

U. Consortium et al. Uniprot: the universal protein knowledgebase. *Nucleic Acids Research*, 45(Database issue):D158, 2017.

J. Costa-Silva, D. Domingues, and F. M. Lopes. Rna-seq differential expression analysis: An extended review and a software tool. *PLoS ONE*, 12(12):e0190152, 2017.

C. Curtis, S. P. Shah, S.-F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 2012.

L. E. DeMare, J. Leng, J. Cotney, S. K. Reilly, J. Yin, R. Sarro, and J. P. Noonan. The genomic landscape of cohesin-associated chromatin interactions. *Genome Research*, 23(8):1224–1234, 2013.

R. DerSimonian and R. Kacker. Random-effects model for meta-analysis of clinical trials: an update. *Contemporary Clinical Trials*, 28(2):105–114, 2007.

D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994, 2004.

Y. Hao, T. Stuart, M. H. Kowalski, S. Choudhary, P. Hoffman, A. Hartman, A. Srivastava, G. Molla, S. Madad, C. Fernandez-Granda, and R. Satija. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology*, 2023. doi: 10.1038/s41587-023-01767-y. URL https://doi.org/10.1038/s41587-023-01767-y.

W. E. Johnson, C. Li, and A. Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

A. Krämer, J. Green, J. Pollard Jr, and S. Tugendreich. Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics*, 30(4):523–530, 2014.

J. A. Lavigne, Y. Takahashi, G. V. Chandramouli, H. Liu, S. N. Perkins, S. D. Hursting, and T. T. Wang. Concentration-dependent effects of genistein on global gene expression in mcf-7 breast cancer cells: an oligo microarray study. *Breast Cancer Research and Treatment*, 110:85–98, 2008.

J. Li and G. C. Tseng. An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *Annals of Applied Statistics*, 5:994–1019, 2011.

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.11.598484; this version posted June 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Z. Li, T. Li, M. E. Yates, Y. Wu, A. Ferber, L. Chen, D. D. Brown, J. S. Carroll, M. J. Sikora, G. C. Tseng, et al. The estrogene database reveals diverse temporal, context-dependent, and bidirectional estrogen receptor regulomes in breast cancer. *Cancer Research*, pages CAN–23, 2023.

R. Liu, X. Ye, and T. Cui. Recent progress of biomarker detection sensors. *Research*, 2020, 2020.

J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, et al. The Genotype-Tissue expression (GTEx) project. *Nature Genetics*, 45(6):580–585, 2013.

S. Lu, J. Li, C. Song, K. Shen, and G. C. Tseng. Biomarker detection in the integration of multiple multi-class genomic studies. *Bioinformatics*, 26(3):333–340, 2010.

A. C. Maher, A.-W. Mohsen, J. Vockley, and M. A. Tarnopolsky. Low expression of long-chain acyl-coa dehydrogenase in human skeletal muscle. *Molecular genetics and metabolism*, 100(2):163–167, 2010.

A. McDermaid, B. Monier, J. Zhao, B. Liu, and Q. Ma. Interpretation of differential gene expression results of rna-seq data: review and integration. *Briefings in Bioinformatics*, 20(6):2044–2054, 2019.

I. Nofech-Mozes, D. Soave, P. Awadalla, and S. Abelson. Pan-cancer classification of single cells in the tumour microenvironment. *Nature Communications*, 14(1):1615, 2023.

A. Nsiah-Sefaa and M. McKenzie. Combined defects in oxidative phosphorylation and fatty acid β-oxidation in mitochondrial disease. *Bioscience Reports*, 36(2), 2016.

Q. Qin, J. Fan, R. Zheng, C. Wan, S. Mei, Q. Wu, H. Sun, M. Brown, J. Zhang, C. A. Meyer, et al. Lisa: inferring transcriptional regulators through integrative modeling of public chromatin accessibility and chip-seq data. *Genome Biology*, 21(1):1–14, 2020.

J. M. Rae, M. D. Johnson, J. O. Scheys, K. E. Cordero, J. M. Larios, and M. E. Lippman. Greb1 is a critical regulator of hormone dependent breast cancer growth. *Breast Cancer Research and Treatment*, 92:141–149, 2005.

N. J. Rinzema, K. Sofiadis, S. J. Tjalsma, M. J. Verstegen, Y. Oz, C. Valdes-Quezada, A.-K. Felder, T. Filipovska, S. van der Elst, Z. de Andrade dos Ramos, et al. Building regulatory landscapes reveals that an enhancer can recruit cohesin to create contact domains, engage ctcf sites and activate distant genes. *Nature Structural & Molecular Biology*, 29(6):563–574, 2022.

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.11.598484; this version posted June 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, 2015.

T. M. Schaefer, J. A. Wright, P. A. Pioli, and C. R. Wira. Il-1β-mediated proinflammatory responses are inhibited by estradiol via down-regulation of il-1 receptor type i in uterine epithelial cells. *The Journal of Immunology*, 175(10):6509–6516, 2005.

R. G. Smith, E. Hannon, P. L. De Jager, L. Chibnik, S. J. Lott, D. Condliffe, A. R. Smith, V. Haroutunian, C. Troakes, S. Al-Sarraj, et al. Elevated dna methylation across a 48-kb region spanning the hoxa gene cluster is associated with alzheimer's disease neuropathology. *Alzheimer's & Dementia*, 14(12):1580–1588, 2018.

S. A. Stouffer, E. A. Suchman, L. C. DeVinney, S. A. Star, and R. M. Williams Jr. *The american soldier: Adjustment during army life.(studies in social psychology in world war ii), vol. 1*. Princeton Univ. Press, 1949.

V. Theodorou, R. Stark, S. Menon, and J. S. Carroll. Gata3 acts upstream of foxa1 in mediating esr1 binding by shaping enhancer accessibility. *Genome Research*, 23(1):12–22, 2013.

L. H. C. Tippett et al. The methods of statistics. *The Methods of Statistics*, 1931.

M. Tokura, J. Nakayama, M. Prieto-Vila, S. Shiino, M. Yoshida, T. Yamamoto, N. Watanabe, S. Takayama, Y. Suzuki, K. Okamoto, et al. Single-cell transcriptome profiling reveals intratumoral heterogeneity and molecular features of ductal carcinoma in situ. *Cancer Research*, 82(18):3236–3248, 2022.

T. A. Trikalinos, G. Salanti, E. Zintzaras, and J. P. Ioannidis. Meta-analysis methods. *Advances in Genetics*, 60:311–334, 2008.

G. C. Tseng, D. Ghosh, and E. Feingold. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Research*, 40(9):3785–3799, 2012.

M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, et al. Tissue-based map of the human proteome. *Science*, 347 (6220):1260419, 2015.

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.11.598484; this version posted June 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

S. Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of Research and Development*, 4(1):66–82, 1960.

S. Z. Wu, G. Al-Eryani, D. L. Roden, S. Junankar, K. Harvey, A. Andersson, A. Thennavan, C. Wang, J. R. Torpy, N. Bartonicek, et al. A single-cell and spatially resolved atlas of human breast cancers. *Nature genetics*, 53(9):1334–1347, 2021a.

T. Wu, E. Hu, S. Xu, M. Chen, P. Guo, Z. Dai, T. Feng, L. Zhou, W. Tang, L. Zhan, et al. clusterprofiler 4.0: A universal enrichment tool for interpreting omics data. *The innovation*, 2(3), 2021b.

K. Xu, R. Wang, H. Xie, L. Hu, C. Wang, J. Xu, C. Zhu, Y. Liu, F. Gao, X. Li, et al. Single-cell rna sequencing reveals cell heterogeneity and transcriptome profile of breast cancer lymph node metastasis. *Oncogenesis*, 10(10):66, 2021.