Journal of Data Science 19 (4), 555–568
October 2021
DOI: 10.6339/21-JDS1019
Statistical Data Science

A Pan-Cancer Network Analysis with Integration of miRNA-Gene Targeting for Multiomics Datasets

Henry Linder[1] and Yuping Zhang[1,*]

[1]*Department of Statistics, University of Connecticut, Storrs, Connecticut, USA*

Abstract

Large-scale genomics studies provide researchers with access to extensive datasets with extensive detail and unprecedented scope that encompasses not only genes, but also more experimental functional units, including non-coding microRNAs (miRNAs). In order to analyze these high-fidelity data while remaining faithful to the underlying biology, statistical methods are necessary that can reflect the full range of understanding in contemporary molecular biology, while remaining flexible enough to analyze a wide range of data and complex phenomena. Leveraging multiple omics datasets, miRNA-gene targets as well as signaling pathway topology, we present an integrative linear model to analyze signaling pathways. Specifically, we use a mixed linear model to characterize tumor and healthy tissue, and execute statistical significance testing to identify pathway disturbances. In this paper, pan-cancer analysis is performed for a wide range of signaling pathways. We discuss specific findings from this analysis, as well as an interactive data visualization available for public consumption that contains the full range of our analytic findings.

Keywords *hypothesis testing; integrative statistical learning; large-scale inference; multi-view data integration*

1 Introduction

Modern large-scale genomic data is unprecedented not only in breadth, but also depth: in addition to omics measurements on thousands of genes across multiple data collection platforms, observations from non-gene cellular functional units are now collected as a matter of routine. These data offer insight into larger processes in systems biology, and the opportunity for structural statistical modeling of molecular processes is particularly promising.

In order to use these novel data types to more fully characterize the human genome, valid statistical methods are critical to harnessing this bioinformatic resource. One important class of molecular entities are microRNAs (miRNAs), small non-coding RNAs that are understood to regulate gene expression in a targeted fashion. miRNAs have gained visibility in recent work that situates genetic activity in a larger, systems context that includes miRNAs, but the sophistication of existing analytic methods varies. A typical clinical approach considers miRNAs individually and the impacts of specific miRNAs on cellular processes. For instance, Hamilton et al. (2013) identified a co-active group of specific miRNAs that contribute to a wide range of cancerous tumors. Likewise, Kim et al. (2017) offered a meta-analysis of the degree to which activity in certain miRNAs can be used as a prognostic tool to differentiate types of osteosarcoma. Sun et al. (2017) performed experiments on a single miRNA in colorectal samples, and compared

*Corresponding author. Email: yuping.zhang@uconn.edu.

Received March 12, 2021; Accepted July 17, 2021

their findings with samples drawn from a sample of colon cancer tumors from a large-scale omics study. miRNAs have also been considered from the perspective of treatment: Gurbuz and Ozpolat (2019) discussed miRNA-targeted therapy for treatment of pancreatic cancer, and performed a meta-analysis of a variety of specific results.

More comprehensive analyses situate miRNAs within a broader biological context, often with an emphasis on integration of multiple data types. Dhawan et al. (2018) performed a pan-cancer analysis of miRNA contributions to gene activity. They integrated gene-level methylation and copy number with miRNA expression and estimated a penalized regression model for gene expression. Falzone et al. (2018) took a different approach to integration, by considering interactions between miRNAs and gene networks. The massive scale of miRNA data, encompassing several thousand distinct entities, has also motivated open-access, interactive databases, often implemented as web applications. These expand the scope of miRNA analysis: Wong et al. (2017) produced a pan-cancer analysis, with exploratory analysis available through a web application interface. Interactive analysis has benefited from these methods, as well. Backes et al. (2016) published a database of miRNAs that play regulatory roles for gene pathways.

Integrating multi-omics data with network topology for network analysis plays an important role in understanding complex diseases and biological systems. For instance, Zhang et al. (2017) integrates gene expression, copy number variations, methylation data with network topology to dissect pathway disturbances. In this paper, we integrate observations of miRNA profiles, gene expression, methylations and copy number variations to analyze pathway disturbance and differential activity between tumor and healthy tissue for 14 cancer types. The paper proceeds as follows. We summarize the datasets used for our analysis in Section 2, and introduce the functional relationships between miRNAs and genes. In Section 3, we introduce the integrative pathway model, with a focus on miRNA-gene integration in a graphical model. Section 4 gives an overview of our pan-cancer pathway analysis, specific results of pathway disturbances, and outlines the interactive data visualization. Finally, we summarize and discuss our findings in Section 5.

2 Data

Our approach to integration and joint modeling of multivariate omics observations is motivated by real-world, large-scale omics studies. We used data collected and published by The Cancer Genome Atlas (TCGA) (Tomczak et al., 2015), which includes omics measurements across more than 30 cancers. We considered two sample populations, one of cancerous tissue samples collected from primary tumor tissues, and the other matched healthy tissue, that is, healthy tissue collected from the corresponding anatomical site as the relevant tumor. We downloaded the data from the NCI Genomic Data Commons (Grossman et al., 2016) using the `TCGA-Assembler` software package for R, version 2.0.0 (Zhu et al., 2014; Wei et al., 2017).

For each tissue sample, we considered observations measured on two distinct functional units, non-coding miRNAs and genes. For the miRNAs, we obtained expression values, collected using the RNASeqV2 platform. Likewise, we obtained RNASeq expression for genes, as well as copy number variation and methylation data. For the expression data, we used normalized values for both miRNAs and genes. We used normalized values provided by TCGA, namely, reads per million (RPM) for miRNAs, and fragments per kilobase per mission reads mapped, upper quartile (FPKM-UQ) for genes. Both methods are conventional normalizations for gene expression values, and the application of gene expression transformations to miRNA expression is

widespread in the literature—see, for example, Stokowy et al. (2014). We verified the empirical similarity between RPM and FPKM-UQ values for both miRNAs and genes, by calculating each ourselves, but we used the TCGA normalized values to ensure maximal data coherence and alignment with the exact data published by TCGA. To the normalized values of miRNA and gene expression, we applied a $\log_2$ transformation. Copy number, with germline copy number variants omitted, was measured using the Affymetrix Genome-Wide Human SNP Array 6.0 platform. To aggregate the copy number observations collected by DNA region, we took the average copy number value across all regions corresponding to a given gene. Methylation was collected using the HumanMethylation450 BeadChip platform, and we aggregated CpG site-level observations using the average methylation beta value across all CpG sites for a given gene.

After formatting and transformation, the TCGA datasets contained 47548 omics measurements across the four data platforms of miRNA expression, gene expression, gene copy number, and gene methylation. In particular, this included 20502 genes and 1870 miRNAs. Because our data analysis is oriented towards signaling pathways, we also downloaded the NCI Pathway Interaction Database (PID), published by Schaefer et al. (2008). The database consists of 212 signaling pathways, specified as directed graphs giving functional relationships between genes. Subsequently, we considered only those genes in the TCGA dataset that occurred in at least one of the PID pathways. This produced 2363 genes for analysis, as well as the corresponding copy number and methylation observations for each gene.

The final components of our integrative pathway analysis are miRNAs. These non-coding functional units are widely understood to target genes (Lewis et al., 2005), possibly serving a regulatory role for gene activity. However, because of the large numbers of both miRNAs and genes, exhaustive experimental validation of miRNA-gene target relationships is computationally costly to the point of prohibitive (Agarwal et al., 2015). To partially address this combinatoric hurdle, databases have been assembled of experimentally-validated miRNA-gene targets, as well as computational predictions of miRNA-gene interactions. To complete our dataset, we downloaded the mirDIP database published by Tokar et al. (2017). The database aggregates more than 30 sources of miRNA-gene targets, and assigns to each pairing a confidence score intended to capture the cumulative support across the literature for the presence of a target. The mirDIP authors also stratified these scores into "confidence classes," the strongest of which is relationships for which the evidence is "very confident" of the target interaction. Therefore, we used as the basis for our omics miRNA dataset the set of all miRNAs with a "very confident" target among the 2363 genes in both the TCGA dataset and the PID pathways. This yielded a total of 510 miRNAs in our integrative dataset.

3 Methods

We build the integrative pathway model in three stages: first, we specify a directed graph to represent a signaling pathway in terms of genes; second, we extend the graph to include miRNA-gene targeting; and third, we again augment the graph, to integrate the secondary data on copy number and methylation for genes.

3.1 Network Construction

We start with a signaling pathway that consists of p genes. Denote a set of graph vertices by $\mathcal{V}_{\mathrm{E}}$, where each vertex $v \in \mathcal{V}_{\mathrm{E}}$ corresponds to a gene, and the subscript "E" emphasizes that each vertex corresponds to an observation of gene expression. We assume the topology of the

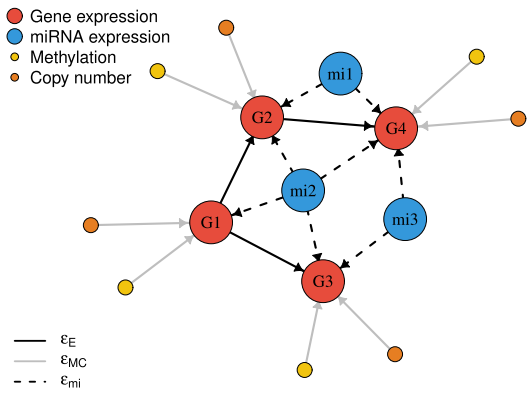

Figure 1: A toy example. The graph consists of a root node (G1) with two children (G2, G3), one of which (G2) has a child of its own (G4). Associated with each gene are copy number and methylation vertices. The toy graph includes three miRNAs, two of which (mi1, mi3) each target two genes, and one (mi2) which targets all four genes. In total, the graph contains $g = 15$ vertices.

signaling pathway is known, as is the case with the PID pathways introduced in Section 2. This topology is specified as a set of directed edges $\mathcal{E}_{\mathrm{E}}$ between the elements of $\mathcal{V}_{\mathrm{E}}$. We then define the signaling pathway as the graph $\mathcal{G}_{\mathrm{E}} = \{\mathcal{V}_{\mathrm{E}}, \mathcal{E}_{\mathrm{E}}\}$.

Next, we introduce a set of vertices that correspond to observations of miRNA expression. Denote the set of miRNA vertices by $\mathcal{V}_{\mathrm{mi}}$, and without loss of generality, we suppose that each element in $\mathcal{V}_{\mathrm{mi}}$ targets at least one gene in $\mathcal{V}_{\mathrm{E}}$. Denote the number of miRNAs by m, i.e., $|\mathcal{V}_{\mathrm{mi}}| = m$. We consider known the miRNA-gene target relationships between elements of $\mathcal{V}_{\mathrm{E}}$ and $\mathcal{V}_{\mathrm{mi}}$, as is the case with the mirDIP database introduced in Section 2. We encode the miRNA-gene targeting as a set of directed edges $\mathcal{E}_{\mathrm{mi}}$, where each edge $e \in \mathcal{E}_{\mathrm{mi}}$ leads from a miRNA vertex in $\mathcal{V}_{\mathrm{mi}}$ to a gene vertex in $\mathcal{V}_{\mathrm{E}}$. This yields a graph $\mathcal{G}_{\mathrm{miE}} = \{\mathcal{V}_{\mathrm{miE}}, \mathcal{E}_{\mathrm{miE}}\}$ where $\mathcal{V}_{\mathrm{miE}} = \mathcal{V}_{\mathrm{E}} \cup \mathcal{V}_{\mathrm{mi}}$ and $\mathcal{E}_{\mathrm{miE}} = \mathcal{E}_{\mathrm{E}} \cup \mathcal{E}_{\mathrm{mi}}$.

Finally, we integrate copy number and methylation into the graph, following the EMC integration method of Zhang et al. (2017). We consider two sets of p vertices $\mathcal{V}_{\mathrm{C}}$ and $\mathcal{V}_{\mathrm{M}}$, corresponding to observations of copy number and methylation for each of the p genes in the signaling pathway, respectively. Define sets of edges $\mathcal{E}_{\mathrm{C}}$ and $\mathcal{E}_{\mathrm{M}}$, each containing p directed edges, one leading from each vertex to the corresponding gene expression vertex in $\mathcal{E}_{\mathrm{E}}$ for the gene on which copy number or methylation is observed. Then, we define the fully integrated graph as $\mathcal{G} \equiv \mathcal{G}_{\mathrm{miEMC}} = \{\mathcal{V}_{\mathrm{miEMC}}, \mathcal{E}_{\mathrm{miEMC}}\}$, where $\mathcal{V}_{\mathrm{miEMC}} = \mathcal{V}_{\mathrm{E}} \cup \mathcal{V}_{\mathrm{mi}} \cup \mathcal{V}_{\mathrm{M}} \cup \mathcal{V}_{\mathrm{C}}$ and $\mathcal{E}_{\mathrm{miEMC}} = \mathcal{E}_{\mathrm{E}} \cup \mathcal{E}_{\mathrm{mi}} \cup \mathcal{E}_{\mathrm{M}} \cup \mathcal{E}_{\mathrm{C}}$.

Without loss of generality, we suppose that $\mathcal{V}_{\mathrm{miEMC}}$ contains $g = 3p + m$ vertices. In real-world datasets like the TCGA data we used in our analysis, it is often the case that copy number or methylation observations are unavailable for some genes in $\mathcal{V}_{\mathrm{E}}$. Moreover, because of the recency and experimental nature of identifying miRNA-gene targets, it is possible that $\mathcal{V}_{\mathrm{mi}}$ is incomplete. We remedy these potential issues by dropping the relevant copy number or methylation vertices, and observing that the directed nature of the integration scheme means that missing vertices in $\mathcal{V}_{\mathrm{mi}}$, $\mathcal{V}_{\mathrm{C}}$, or $\mathcal{V}_{\mathrm{M}}$ may be removed without altering the topology of $\mathcal{G}$ with respect to the signaling pathway itself.

As an illustrative example, we show in Figure 1 a toy network with the integration scheme described above. The signaling pathway consists of four genes: a root node with two child nodes,

one of which also has its own child node. We include three miRNAs, two that target two genes each, and one that targets all four genes. We also include separate graph vertices for copy number and methylation. The fully-integrated graph consists of $g = 15$ vertices.

Having defined an integrative graph for the four omics data types, we pivot to the task of representing the graph in a mathematically accessible fashion, which can be conveniently used as the basis for a linear statistical model. We start with the unweighted graph adjacency matrix $\mathbf{A}^\star$, a $g \times g$ matrix with elements $\alpha_{jk} \in \{0, 1\}$, $j, k = 1, \ldots, g$. The elements of the unweighted adjacency matrix are defined as an indicator function:

$$\alpha_{jk} = \begin{cases} 1 & \text{if } \mathcal{E}_{\text{miEMC}} \text{ contains a directed edge from graph vertex } k \text{ to vertex } j \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

We can decompose $\mathbf{A}^\star$ into the three integrative pieces described above, for (1) the signaling pathway, (2) miRNA-gene targets, and (3) gene-level copy number and methylation:

$$\mathbf{A}^\star = \begin{pmatrix} \mathbf{A}^\star_{\text{E}} & \mathbf{A}^\star_{\text{mi}} & \mathbf{A}^\star_{\text{M}} & \mathbf{A}^\star_{\text{C}} \\ \mathbf{O}_{(2p+m)\times p} & \mathbf{O}_{(2p+m)\times m} & \mathbf{O}_{(2p+m)\times p} & \mathbf{O}_{(2p+m)\times p} \end{pmatrix} \tag{2}$$

Here, $\mathbf{A}^\star_{\text{E}}$ is an indicator matrix for the signaling pathway; $\mathbf{A}^\star_{\text{mi}}$ indicates the miRNA-gene targets; $\mathbf{A}^\star_{\text{C}} = \mathbf{A}^\star_{\text{M}} = \mathbf{I}_p$ indicate the copy number and methylation integration, respectively; and $\mathbf{O}_{q\times s}$ is a $q \times s$ matrix of zeros. The elements of each block submatrix are all defined as in Equation 1, and the zero matrices reflect that we do not include any interaction between the secondary omics features (miRNAs, gene copy number, gene methylation), nor do we include any feedback from genes to omics features.

Within the context of Gaussian graphical models, graph dependence is parameterized as conditional dependence. From this perspective, we may characterize α_{jk} as an indicator for conditional dependence of vertex j on vertex k, given the remaining $(g - 2)$ vertices in the fully-integrated graph $\mathcal{G}$. Moreover, the strength of graph association is formalized as partial correlation, and corresponds to the inverse of the covariance matrix of a multivariate Gaussian distribution (Loh and Wainwright, 2012). The partial correlation between two variables X and Y is defined conditionally with respect to a set of auxiliary random variables $\mathcal{Z}$ which may also associate with X and Y. The partial correlation coefficient is calculated by first regressing X and Y on all elements of $\mathcal{Z}$, and obtaining the residuals $X_{\setminus\mathcal{Z}}$, $Y_{\setminus\mathcal{Z}}$. Here, we use the notation of the orthogonal complement, defined as $X_{\setminus\mathcal{Z}} = X - \mathcal{P}_{\mathcal{Z}}X$, where $\mathcal{P}_{\mathcal{Z}}$ is a linear projection onto the elements of $\mathcal{Z}$—that is, a linear regression Krämer et al. (2009). Then, the partial correlation r_{XY} is simply the Pearson correlation coefficient between the two residuals, i.e., $\text{corr}(X_{\setminus\mathcal{Z}}, Y_{\setminus\mathcal{Z}})$ (Kim, 2015).

For our integrated pathway graph $\mathcal{G}$, we consider observations $\mathbf{y}_1, \ldots, \mathbf{y}_N$, where $\mathbf{y}_i \in \mathbb{R}^g$ contains elements for each graph vertex, that is, gene and miRNA expression, gene copy number, and gene methylation. The graph $\mathcal{G}$ and the unweighted adjacency matrix $\mathbf{A}^\star_{\text{miEMC}}$ then specify functional relationships between the elements of $\mathbf{y}_i$, and we may use these observation vectors to estimate the partial correlation coefficients r_{jk} between graph vertices j and k. Then, we define the weighted adjacency matrix to be $\mathbf{A} \equiv \mathbf{A}_{\text{miEMC}}$, which has elements $a_{jk} = r_{jk}\alpha_{jk}$, where $\alpha_{jk} \in \{0, 1\}$ is the corresponding element of $\mathbf{A}^\star_{\text{miEMC}}$.

3.2 Network Differential Expression Analysis

In order to build a linear statistical model, we extend the NetGSA (Shojaie and Michailidis, 2009) and EMC-NetGSA (Zhang et al., 2017), which defined the influence matrix $\mathbf{\Lambda}$ to express the

mutual, cumulative network effects of the elements of $\mathbf{y}_i$ induced by the graph structure of $\mathcal{G}$. In particular, they found that $\mathbf{\Lambda} = (\mathbf{I}_g - \mathbf{A})^{-1}$ in the special case of a directed acyclic graph (DAG). This condition was relaxed by Shojaie and Michailidis (2010), who found the relationship holds for all adjacency matrices with eigenvalues smaller than 1 in magnitude. Moreover, they offered an approximation to the influence matrix that induces the spectral constraint on $\mathbf{A}$ for any directed graph, which permits application of the linear model to arbitrary signaling pathways. Without loss of generality, we consider the base case of the DAG influence matrix, with the understanding that when the adjacency matrix does not satisfy the constraints, we use instead the approximation of Shojaie and Michailidis (2010).

Given the correspondence between the graph topology of $\mathcal{G}$ and the covariance of a multivariate Gaussian random variable noted above, the NetGSA model uses the influence matrix as a design matrix in a linear regression model, that is, $\mathbb{E}\mathbf{y}_i = \mathbf{\Lambda}\boldsymbol{\beta}$. Here, $\boldsymbol{\beta}$ is an unknown regression parameter that we may interpret as the network-adjusted mean activity for the elements of $\mathbf{y}_i$. Moreover, to account for correlated deviations from the mean within each subject, the NetGSA model is a mixed model, with the form

$$\mathbf{y}_i = \mathbf{\Lambda}\boldsymbol{\beta} + \mathbf{\Lambda}\boldsymbol{\gamma}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, N \tag{3}$$

$$\boldsymbol{\gamma}_i \sim N_g(\mathbf{0}, \sigma_\gamma^2 \mathbf{I}_g) \tag{4}$$

$$\boldsymbol{\epsilon}_i \sim N_g(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_g) \tag{5}$$

Here, $\boldsymbol{\gamma}_i$ is a sample-level random effect, with the consequence that the covariance of $\mathbf{y}_i$ is given by $\sigma_\gamma^2 \mathbf{\Lambda}\mathbf{\Lambda}' + \sigma_\epsilon^2 \mathbf{I}_g$.

In addition to a linear model, the NetGSA framework also provides for hypothesis testing between two populations, generically labeled as "treatment" and "control" which correspond in our analysis to "tumor" and "healthy" tissue, respectively. For each sample $\mathbf{y}_i$, a class label $c_i \in \{\mathrm{T}, \mathrm{C}\}$ is available, indicating the sample population as treatment or control. The sample sizes N_C and N_T are such that $N_\mathrm{C} + N_\mathrm{T} = N$, i.e., every sample can be categorized as either control or treatment. Then, the model in Equation (3) is parameterized with separate adjacency and influence matrices $\mathbf{A}^c$ and $\mathbf{\Lambda}^c$, as well as population-specific activity parameters $\boldsymbol{\beta}^c$, $c = \mathrm{T}, \mathrm{C}$.

The model is easily estimated using restricted maximum likelihood, using results for the classical mixed model. The final component to permit significance testing for differences in subsets of the activity parameters $\boldsymbol{\beta}^\mathrm{T}$, $\boldsymbol{\beta}^\mathrm{C}$ is the network contrast $\boldsymbol{\ell} = (-\mathbf{b} \cdot \mathbf{b}\mathbf{\Lambda}_\mathrm{C}, \mathbf{b} \cdot \mathbf{b}\mathbf{\Lambda}_\mathrm{T})$, where $\cdot$ denotes an element-wise product. It accounts for the cumulative effect of nodes in the signaling pathway graph. Here, $\mathbf{b}$ is a vector of length g with elements in $\{0, 1\}$ that indicate which features in the elements of $\mathbf{y}_i$ should be included in a test for inequality of network-adjusted means. The test statistic $T \propto \boldsymbol{\ell}\boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\boldsymbol{\beta}^{\mathrm{C}'}, \boldsymbol{\beta}^{\mathrm{T}'})'$, follows a Student's t distribution, with degrees of freedom estimated using the Satterthwaite approximation.

4 Results

In preparation for pan-cancer pathway analysis, we obtained observations of gene expression, miRNA expression, methylation, and copy number from TCGA, as described in Section 2. As additional pre-processing, we imputed missing values in the matrix formed by all available samples across all four data modalities, using the full data matrix of all gene expression, miRNA expression, copy number, and methylation features for any gene in the 212 pathways of the NCI Pathway Interaction Database (Schaefer et al., 2008) (PID), or miRNAs present in the

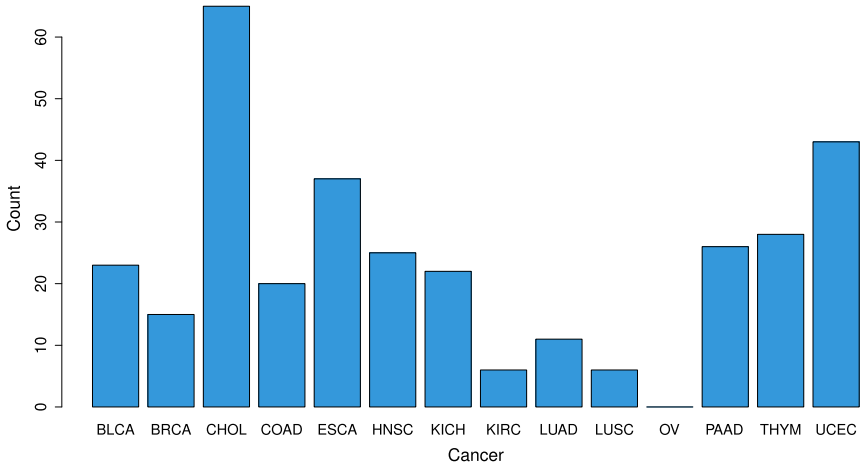

Figure 2: Number of pathways that change significance at the $\alpha = 0.05$ level between the fully-integrated miEMC-NetGSA analysis and the baseline NetGSA performed on expression data only. p-values are Benjamini-Hochberg adjusted to control the false discovery rate. Cancers are labeled according to the TCGA abbreviation codes, which are listed with the full cancer name in Table 1.

mirDIP database that target those genes. For imputation, we applied the iterative integrated imputation (I3) method proposed by Linder and Zhang (2019), a low-rank procedure adapted from the structured matrix completion method introduced by Cai et al. (2016).

After separate imputation on the data matrix for each of 29 TCGA cancer studies, we considered the subset of cancers for which all 10 BRAF pathway genes were observed, and for which more than 10 samples were available in both the tumor and normal tissue sample populations. This yielded 14 cancers for integrative pathway analysis. Of the 212 pathways included in the PID, we analyzed 173 pathways, based on the availability of gene expression observations for all genes in the pathway.

Within each cancer, we performed four pathway analyses: a fully-integrated analysis including miRNAs, methylation, and copy number with gene expression ("miEMC-NetGSA"); integration of miRNA and gene expression only ("miE-NetGSA"); integration of methylation, copy number, and gene expression only ("EMC-NetGSA"); and an expression-only baseline analysis ("E-NetGSA", or simply "NetGSA"). Within each combination of cancer and integration scheme, we applied the p-value adjustment method of Benjamini and Hochberg (1995) (BH) to control the false discovery rate.

Figure 2 gives counts for each of the 14 cancers of the number of pathways that were insignificant at the $\alpha = 0.05$ level for the baseline expression-only pathway analysis, but were significant in the miEMC-NetGSA analysis. The cancers are labeled using the TCGA abbreviation codes, which are listed along with the corresponding, full cancer name in Table 1. As the figure makes clear, miRNA integration consistently leads to increased statistical significance in the pathway analyses. A risk is that this systematically elevates the significance of the same pathways across cancers, not due to a true biological pathway disturbance, but instead deterministically because of an increased number of genomic features. However, this is apparently not the case: Figure 3 shows an indicator matrix for changes in significance by pathway across all cancers. Despite some overlap in the pathways that change significance, the pathways that develop significance

Table 1: Cancer names and abbreviation codes for the 14 cancers included in the pan-cancer analysis.

Name	Code
Bladder urothelial carcinoma	BLCA
Breast invasive carcinoma	BRCA
Cholangiocarcinoma	CHOL
Colon adenocarcinoma	COAD
Esophageal carcinoma	ESCA
Head and neck squamous cell carcinoma	HNSC
Kidney chromophobe	KICH
Kidney renal clear cell carcinoma	KIRC
Lung adenocarcinoma	LUAD
Lung squamous cell carcinoma	LUSC
Ovarian serous cystadenocarcinoma	OV
Pancreatic adenocarcinoma	PAAD
Thymoma	THYM
Uterine corpus endometrial carcinoma	UCEC

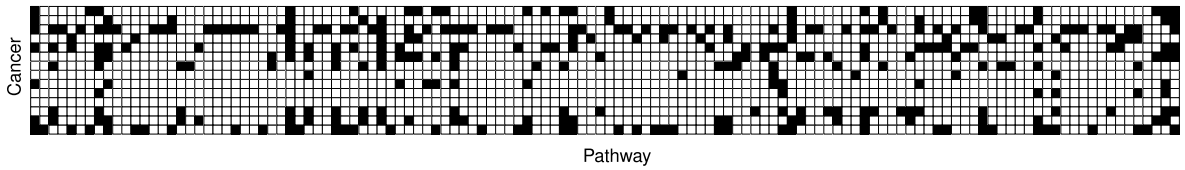

Figure 3: Indicator matrix for the pathways that change significance between expression-only NetGSA and the fully-integrated miEMC-NetGSA, counted in Figure 2. Rows correspond to cancers, columns correspond to pathways. A black cell indicates the pathway changed significance in the cancer.

in concert with the miRNA integration are not consistent across cancers. This suggests the significance changes are due to differential contributions of the various miRNA and omics features across cancers, with varying effects that depend on the individual cancer in question.

Figure 4 shows the results of significance tests for colon adenocarcinoma (COAD) for the pathways that change significance between the expression-only and fully-integrated analyses. The figure shows the $-\log_{10}$ transformation of the BH-adjusted p-values for each full pathway under the four integration schemes for the pathway analysis. The analyses exhibit several different patterns.

For some pathways, integration of any variety increases the significance substantially from the expression-only analysis. One such pathway is IL23-mediated signaling events, which increases in significance somewhat after integration of methylation and copy number, and increases more substantially after integration of miRNAs. For other pathways, such as signaling events mediated by PRL, integration of miRNAs alone increases the significance substantially, whereas full integration with methylation and copy number dampens the increase in significance. In general, this pattern is widespread: other pathways with smaller significance in the miEMC-NetGSA analysis compared with the miE-NetGSA analysis include EGFR-dependent

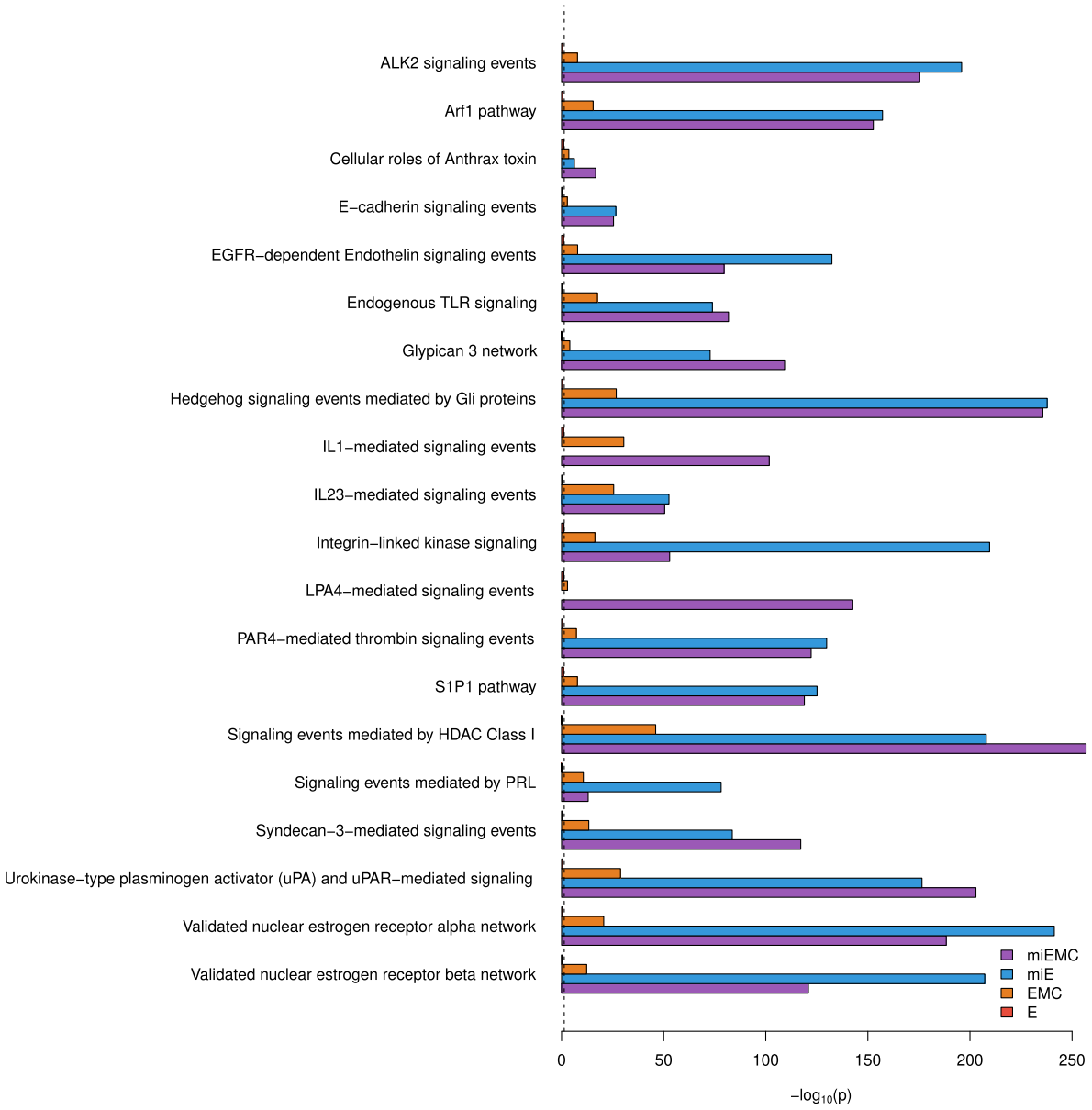

Figure 4: Significance test results for the 20 pathways that changed significance between expression-only and fully-integrated miEMC-NetGSA analyes in colon adenocarcinoma (COAD). Plotted are $-\log_{10}(p)$-values for the four pathway analyses: miEMC-NetGSA integrated miRNA expression, gene expression, methylation, and copy number; miE-NetGSA integrates miRNA and gene expression; EMC-NetGSA integrates gene expression, methylation, and copy number; and E-NetGSA is expression-only NetGSA. p-values were adjusted using the Benjamini-Hochberg method to control the false discovery rate.

endothelin signaling events, integrin-linked kinase signaling, and the validated nuclear estrogen receptor beta network.

On the other hand, there is sometimes a synergistic effect to integration of miRNA with the methylation and copy number. One such pathway is cellular roles of anthrax toxin, which

has substantially stronger significance after integration of both miRNAs with methylation and copy number, compared with either the miRNA or gene-level omics features alone. In other cases, such as signaling events mediated by HDAC Class I and syndecan-3-mediated signaling events, the integration of miRNAs increases the significance substantially, and that significance increases further after integration of the gene-level features. Finally, in other pathways, including ALK2 signaling events, the Arf1 pathway, and the S1P1 pathway, any integration with miRNAs markedly increases the significance, but integration of methylation and copy number does not have a strong effect.

To illustrate the analytic advantage of the fully-integrated analysis, we considered the Notch signaling pathway in cholangiocarcinoma. Kwon et al. (2017) found that DNA methylation inhibited expression of the miR-34a, which in turn targets components of the Notch pathway, which contributes to tumorigenesis.

In our analysis, the p-value for the Notch pathway was significant in expression-only and EMC analyses, at the $\alpha = 0.05$ level, but the miE-NetGSA analysis was not significant. However, joint integration of miRNA, gene expression, methylation, and copy number observations resulted in highly significant pathway disturbance in cholangiocarcinoma tumors.

Figure 5 shows parameter estimates for the healthy and tumor tissue in the TCGA dataset. Although many of the parameter estimates remain the same across the integration schemes—for instance, WNT7A in the tumor population, and FZD2 in the healthy population—some parameters exhibit substantial changes in magnitude or sign. These include LRP6 in the tumor population, IGFBP4 in the healthy population, WNT5A in the tumor population, and DKK1 in the healthy population. The variability and sign changes observed in these parameters indicate strong structural changes implied by the various integration schemes, and reinforce the importance of biologically-founded statistical analysis.

Taken together, these results suggest real effects due to miRNA integration that are not uniform in character: sometimes, miRNA integration clarifies weak significance, other times it supplements information in a gene-level analysis, and in still others, it exaggerates significance and is regulated by methylation and copy number integration. Moreover, we confirm known pathway disturbance that relates to a specific miRNA, miR-34A, in cholangiocarcinoma, and identify substantial effects on the parameter estimates due to the different integration schemes.

The results discussed above represent only a small subset of our overall analysis across the full sets of cancers and pathways. In order to promote access to and exploration of the results of our analysis, we also built an interactive data visualization, which we published online for public access at:

https://zhang-lab.shinyapps.io/pathway-analysis-mirna/

Figure 6 gives screenshots from the data visualization, which permits dynamic interaction with the graph topology of each signaling pathway, as well as plots of the significance and test statistics for each full pathway as well as individual features, and coefficient estimates for the integrative pathway model.

5 Discussion

In this paper, we constructed an integrative pan-cancer network analysis with observations of gene expression, copy number, and methylation, as well as miRNA expression across 14 cancers and more than 170 signaling pathways. We found that integration of miRNAs has a differential

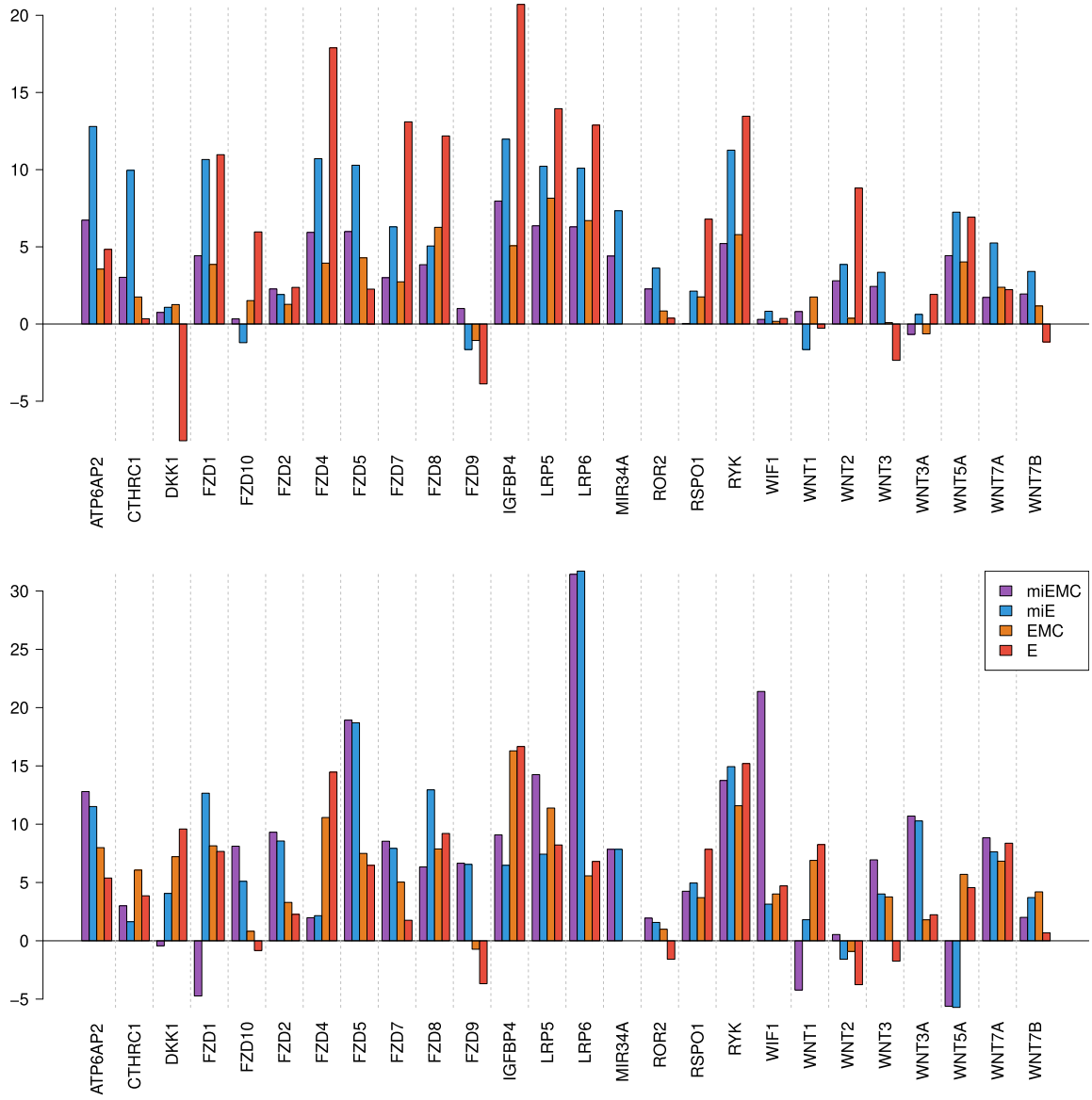

Figure 5: Network-adjusted activity parameters for genes in the Notch pathway and miR-34a, in cholangiocarcinoma, healthy tissue (top) and tumorous tissue (bottom).

impact depending on the pathway under consideration. This indicates real biological changes in interpretation when estimating models using different sets of omics features. We also gave details of a comprehensive, interactive data visualization for the full set of testing results.

Our analysis reflects current trends in molecular biology and functional genomics, and leverages the large volumes of omics data to provide insight into the biological processes and signaling events that contribute to a range of different cancer types. Our model incorporates known biological functions, and permits principled analysis of these omics data. Our findings support the value of integrative analysis, and provide a novel tool for comparisons of populations to obtain better understanding of the mechanisms that give rise to complex diseases. For future work, it would be worth considering integrative models with more types of biological data.

Figure 6: Screenshots of interactive data visualization of the pan-cancer pathway analysis. The interface includes dynamic visualization of signaling pathway graph topology (top), results of inference for the full pathway and individual features (middle), and coefficient estimates (bottom).

Supplementary Material

Supplementary Materials include descriptions for data and software.

References

Agarwal V, Bell GW, Nam JW, Bartel DP (2015). Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, 4: e05005.

Backes C, Kehl T, Stöckel D, Fehlmann T, Schneider L, Meese E, et al. (2016). miRPathDB: A new dictionary on microRNAs and target pathways. *Nucleic Acids Research*, 45(D1): D90–D96. gkw926.

Benjamini Y, Hochberg Y (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, Methodological*, 57(1): 289–300.

Cai T, Cai TT, Zhang A (2016). Structured matrix completion with applications to genomic data integration. *Journal of the American Statistical Association*, 111(514): 621–633.

Dhawan A, Scott JG, Harris AL, Buffa FM (2018). Pan-cancer characterisation of microrna across cancer hallmarks reveals microrna-mediated downregulation of tumour suppressors. *Nature Communications*, 9(1): 5228.

Falzone L, Scola L, Zanghì A, Biondi A, Di Cataldo A, Libra M, et al. (2018). Integrated analysis of colorectal cancer microRNA datasets: Identification of microRNAs associated with tumor development. *Aging (Albany NY)*, 10(5): 1000.

Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, et al. (2016). Toward a shared vision for cancer genomic data. *The New England Journal of Medicine*, 375(12): 1109–1112.

Gurbuz N, Ozpolat B (2019). MicroRNA-based targeted therapeutics in pancreatic cancer. *Anticancer Research*, 39(2): 529–532.

Hamilton MP, Rajapakshe K, Hartig SM, Reva B, McLellan MD, Kandoth C, et al. (2013). Identification of a pan-cancer oncogenic microRNA superfamily anchored by a central core seed motif. *Nature Communications*, 4: 2730.

Kim S (2015). ppcor: An R package for a fast calculation to semi-partial correlation coefficients. *Communications for Statistical Applications and Methods*, 22(6): 665.

Kim YH, Goh TS, Lee CS, Oh SO, Kim JI, Jeung SH, et al. (2017). Prognostic value of microRNAs in osteosarcoma: A meta-analysis. *Oncotarget*, 8(5): 8726.

Krämer N, Schäfer J, Boulesteix AL (2009). Regularized estimation of large-scale gene association networks using graphical Gaussian models. *BMC Bioinformatics*, 10(1): 384.

Kwon H, Song K, Han C, Zhang J, Lu L, Chen W, et al. (2017). Epigenetic silencing of miRNA-34a in human cholangiocarcinoma via EZH2 and DNA methylation: Impact on regulation of Notch pathway. *The American Journal of Pathology*, 187(10): 2288–2299.

Lewis BP, Burge CB, Bartel DP (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1): 15–20.

Linder H, Zhang Y (2019). Iterative integrated imputation for missing data and pathway models with applications to breast cancer subtypes. *Communications for Statistical Applications and Methods*, 26(4): 411–430.

Loh PL, Wainwright MJ (2012). Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. In: *Advances in Neural Information Processing Systems*,

2087–2095. Curran Associates, Inc.

Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, et al. (2008). PID: The pathway interaction database. *Nucleic Acids Research*, 37(suppl_1): D674–D679.

Shojaie A, Michailidis G (2009). Analysis of gene sets based on the underlying regulatory network. *Journal of Computational Biology*, 16(3): 407–426.

Shojaie A, Michailidis G (2010). Network enrichment analysis in complex experiments. *Statistical Applications in Genetics and Molecular Biology*, 9(1): Article 22, 34 pages.

Stokowy T, Eszlinger M, Świerniak M, Fujarewicz K, Jarząb B, Paschke R, et al. (2014). Analysis options for high-throughput sequencing in miRNA expression profiling. *BMC Research Notes*, 7(1): 144.

Sun M, Song H, Wang S, Zhang C, Zheng L, Chen F, et al. (2017). Integrated analysis identifies microRNA-195 as a suppressor of Hippo-YAP pathway in colorectal cancer. *Journal of Hematology & Oncology*, 10(1): 79.

Tokar T, Pastrello C, Rossos AE, Abovsky M, Hauschild AC, Tsay M, et al. (2017). mirDIP 4.1-integrative database of human microRNA target predictions. *Nucleic Acids Research*, 46(D1): D360–D370.

Tomczak K, Czerwińska P, Wiznerowicz M (2015). The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Contemporary Oncology*, 19(1A): A68.

Wei L, Jin Z, Yang S, Xu Y, Zhu Y, Ji Y (2017). TCGA-assembler 2: Software pipeline for retrieval and processing of TCGA/CPTAC data. *Bioinformatics*, 34(9): 1615–1617.

Wong NW, Chen Y, Chen S, Wang X (2017). OncomiR: An online resource for exploring pan-cancer microRNA dysregulation. *Bioinformatics*, 34(4): 713–715.

Zhang Y, Linder MH, Shojaie A, Ouyang Z, Shen R, Baggerly KA, et al. (2018). Dissecting pathway disturbances using network topology and multi-platform genomics data. *Statistics in Biosciences*, 10: 86–106.

Zhu Y, Qiu P, Ji Y (2014). TCGA-assembler: Open-source software for retrieving and processing TCGA data. *Nature Methods*, 11(6): 599.