JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 27, Number 7, 2020

Pp. 1157–1170
DOI: 10.1089/cmb.2019.0321

Explaining Gene Expression Using Twenty-One MicroRNAs

AMIR ASIAEE,[1,*] ZACHARY B. ABRAMS,[2,*] SAMANTHA NAKAYIZA,[2]
DEEPA SAMPATH,[3] and KEVIN R. COOMBES[2]

ABSTRACT

The transcriptome of a tumor contains detailed information about the disease. Although advances in sequencing technologies have generated larger data sets, there are still many questions about exactly how the transcriptome is regulated. One class of regulatory elements consists of microRNAs (or miRs), many of which are known to be associated with cancer. To better understand the relationships between miRs and cancers, we analyzed ~9000 samples from 32 cancer types studied in The Cancer Genome Atlas. Our feature reduction algorithm found evidence for 21 biologically interpretable clusters of miRs, many of which were statistically associated with a specific type of cancer. Moreover, the clusters contain sufficient information to distinguish between most types of cancer. We then used linear models to measure, genome-wide, how much variation in gene expression could be explained by the 21 average expression values ("scores") of the clusters. Based on the ~20,000 per-gene R^2 values, we found that (1) mean differences between tissues of origin explain about 36% of variation; (2) the 21 miR cluster scores explain about 30% of the variation; and (3) combining tissue type with the miR scores explained about 56% of the total genome-wide variation in gene expression. Our analysis of poorly explained genes shows that they are enriched for olfactory receptor processes, sensory perception, and nervous system processing, which are necessary to receive and interpret signals from *outside* the organism. Therefore, it is reasonable for those genes to be always active and not get downregulated by miRs. In contrast, highly explained genes are characterized by genes translating to proteins necessary for transport, plasma membrane, or metabolic processes that are heavily regulated processes *inside* the cell. Other genetic regulatory elements such as transcription factors and methylation might help explain some of the remaining variation in gene expression.

Keywords: feature extraction, microRNA, mRNA, gene regulation, gene expression prediction.

[1]Mathematical Biosciences Institute, The Ohio State University, Columbus, Ohio, USA.
[2]Department of Biomedical Informatics, The Ohio State University, Columbus, Ohio, USA.
[3]Division of Hematology, Department of Internal Medicine, The Ohio State University, Columbus, Ohio, USA.
*These authors contributed equally to this work.

1. INTRODUCTION

MicroRNAs (miRs) are a class of noncoding RNAs that play a key role in negatively regulating messenger RNA (mRNA) by complementarily binding to the mRNA and inducing degradation or translational repression (He and Hannon, 2004; Rupaimoole and Slack, 2017). This powerful form of gene regulation was evolutionarily developed as a way to protect cells against retroviruses (Houzet and Jeang, 2011). The miR-mRNA interaction is entangled since each miR can regulate hundreds of mRNAs, whereas each mRNA can be controlled by multiple miRs. Furthermore, miRs regulate and are regulated by different types of long noncoding RNAs (Garzon et al., 2010; Peng and Croce, 2016). In cancer, a single miR can act as either an oncogene or a tumor-suppressor in different contexts (Garzon et al., 2010). Two notable examples of miRs relation to cancer are the association of miR-21 with melanoma (Melnik, 2015) and the effect of the tumor suppressor miR-34a in liver cancer (Daige et al., 2014).

Since miRs influence multiple pathways involved in cancer, there has been an ongoing effort to target them to reduce the risk of developing resistance to therapy (Garzon et al., 2010; Iorio and Croce, 2012; Peng and Croce, 2016). Beyond the challenges in developing inhibitors or mimics for miRs and difficulties in delivering those chemicals to the tumor, our limited understanding of the *downstream effect* of miRs is a main reason that prevents drugs targeting miRs from reaching the bedside of patients. Such limited knowledge is speculated to be the root cause of the failure of the first miR-targeted therapy (i.e., MRX34 targeting miR-34 for liver cancer) (Dragomir et al., 2018) due to severe side effects, which underscores the need for in-depth understanding of the role of miRs in cancer and gene regulation.

The complex set of connections between miRs and other cellular molecules makes it extremely difficult to predict and analyze the precise role of a single miR in human cancer. This complexity is amplified by the high false-positive rates of computational tools that try to predict miR-mRNA interactions based on sequence complementarity or evolutionary conservation (Riffo-Campos et al., 2016).

In this article, we take a machine learning approach to further understand miR-cancer and miR-mRNA relationships. We leverage our newly developed feature extraction algorithm, Thresher, to extract 21 biologically meaningful clusters of miRs. Then, we use the means of these 21 clusters (''scores'') as features for clustering cancers and predicting gene expression. We show that using only 21 miR scores, we can distinguish 32 cancer types of The Cancer Genome Atlas (TCGA) (Cancer Genome Atlas Research Network et al., 2013) data set, and we can explain a large portion of the variation in gene expression.

From a biological point of view, we interpret the extracted 21 miR clusters and then examine their association to different cancer types. Our results indicate significant correlations linking miR clusters with a particular cancer or set of cancers. Finally, we perform gene enrichment analysis for sets of genes whose expressions are poorly or well explained by miR scores and interpret our findings.

2. METHODS

In this section, we describe our data set, preprocessing, and analysis methods. Figure 1 summarizes the steps of our analysis.

2.1. Data

The data used in this study were collected from TCGA (Cancer Genome Atlas Research Network et al., 2013). TCGA is a pan-cancer public data repository that holds both clinical and omics data for over 10,000 patient samples. We used the FireBrowse web portal to identify and download the data from patients included in this study. Patients were selected based on the presence of matched mRNA and miR sequencing data. In total, $n = 8895$ patient samples reflecting 32 different cancer types were obtained.

2.2. Processing and filtering

We normalized the sequencing data from individual samples by computing reads per kilobase per million (Mortazavi et al., 2008). Data were then log2 transformed. We filtered the data by removing miRs that had a read count of zero in 75% of patients. After filtering, $p = 470$ miRs remained. The following steps are performed on the 8895×470 data matrix.

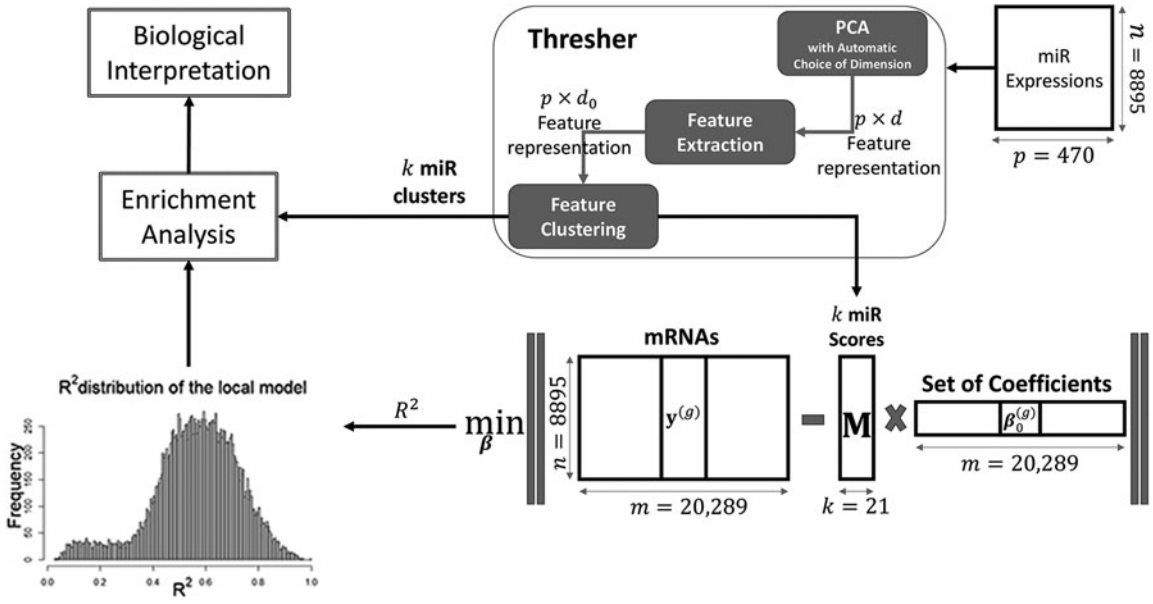

FIG. 1. Workflow of our analysis. Starting from miR expression data (top right), we use the Thresher algorithm to extract biologically meaningful cluster of miRs and perform enrichment analysis to interpret them (top left). We also use the scores (mean expression of miRs in each cluster) of miR clusters to predict the gene expression of all genes (bottom right). From the R^2 results, we select genes that are poorly or highly explained by miRs and perform enrichment analysis to better understand their biological similarities (bottom left). miRs, microRNAs.

2.3. Feature extraction

After data processing and filtering, we analyzed the miR data using version 1.1.1 of the Thresher R package (Wang et al., 2018). Thresher has three main steps: principal components analysis (PCA), outlier filtering, and clustering on hyperspheres using von Mises-Fisher distributions (Banerjee et al., 2005). During PCA, Thresher automatically determines the number d of significant principal components (PCs) by an adaptation of a graphical Bayesian model of Auer and Gervini (2008). For each feature $i \in 1, \ldots, 470$, we have a d-dimensional ''feature representation vector'' $v_i \in \mathbb{R}^d$. Here v_i contains the ''loadings'' of the feature on all d components and represents the total contribution of the feature to the data matrix. For feature selection, Thresher uses $\|v_i\|_2 \geq 0.35$ as a criterion to retain useful features, discarding the less important ones (Wang et al., 2018) and reducing the number of features to $p_0 \leq p$. Finally, it clusters the *directions* of the remaining feature representation vectors on the hypersphere using mixtures of von Mises-Fisher distributions to k clusters where k is determined by Bayesian Information Criterion (BIC) (Wang et al., 2018). When applied to omics data sets, Thresher has shown to be able to recover one-dimensional biologically interpretable clusters of features (Abrams et al., 2018). So, after applying Thresher, the data matrix for each cluster of features should contain only one significant PC. Unlike dimension reduction by PCA, each cluster reflects a natural collection of highly correlated miRs or genes that can be interpreted biologically.

For our TCGA miR data set, $p_0 = p = 470$, which means that no miR was filtered out by the Thresher. Also, the dimension of the feature representation vector (number of significant PCs determined by graphical Bayesian model) is determined as $d = 21$ and the final number of identified miR clusters (determined by BIC) is $k = 21$. We take the mean expression of miRs in each cluster as a new feature, which we call the *miR score* of that cluster. All the following analysis is performed on the resulting 8895×21 miR score matrix, $\mathbf{M}$.

2.4. Data visualization

To show that the 21-dimensional miR score has retained valuable information on the original data set, we visualize the miR score matrix $\mathbf{M}$ using the t-stochastic neighbor embedding (t-SNE) algorithm (van der Maaten and Hinton, 2008) as implemented in version 0.13 of the Rtsne R package.

2.5. Prediction

We use the score matrix $\mathbf{M}$ to predict 20,289 gene expressions of each of $n = 8895$ samples using ordinary least squares. To fit $m = 20{,}289$ linear models efficiently, we used MultiLinearModel function of version 3.1.6 of the ClassComparison R package. We call this set of linear models the *tissue-agnostic model*. Therefore, for each gene $g \in 1, \ldots, 20{,}289$, we are minimizing $\left\|\mathbf{y}^{(g)} - \mathbf{M}\boldsymbol{\beta}_0^{(g)} - b_0^{(g)}\right\|_2$ to estimate the coefficients $\hat{\boldsymbol{\beta}}_0^{(g)}$ and the intercept (bias) $\hat{b}_0^{(g)}$ of the tissue-agnostic model, where $\mathbf{y}^{(g)} \in \mathbb{R}^n$ is the vector of expression of g in all samples.

Cancer type contributes to gene expression through both miRs and other biological pathways. To explore the effect of cancer type on gene expression prediction performance, we included tissue of origin information in regressions in two ways. First, we treated the cancer type as a categorical variable and included it by dummy coding in our model. In the other approach, we fitted cancer-specific linear models to predict the residuals $\mathbf{r}^{(g)} = \mathbf{y}^{(g)} - \mathbf{M}\hat{\boldsymbol{\beta}}_0^{(g)} - \hat{b}_0^{(g)}$ of the tissue-agnostic model. We call this set of linear models the *tissue-aware model*, which minimizes $\left\|\mathbf{r}_t^{(g)} - \mathbf{M}_t\boldsymbol{\beta}_t^{(g)} - b_t^{(g)}\right\|_2$, where $\mathbf{M}_t$ and $\mathbf{r}_t^{(g)}$ contain subset of rows of $\mathbf{M}$ and elements of $\mathbf{r}^{(g)}$ for cancer type t, respectively. Also, $\boldsymbol{\beta}_t^{(g)}$ and $b_t^{(g)}$ are the corresponding cancer-specific parameters.

In the high-dimensional setting where $p \gg n$ minimizing

$$\min_{\forall t:\boldsymbol{\beta}_t} \sum_t \left\|\mathbf{y}_t^{(g)} - \mathbf{M}_t(\boldsymbol{\beta}_0^{(g)} + \boldsymbol{\beta}_t^{(g)})\right\|_2^2 + \lambda_0\|\boldsymbol{\beta}_0\|_1 + \sum_{t=1} \lambda_t\|\boldsymbol{\beta}_t\|_1 \tag{1}$$

has been of recent interest in the statistical machine learning community and is known by multiple names. It is a form of multitask learning (Jalali et al., 2010; Zhang and Yang, 2017) when you consider prediction of expression in each cancer as a task. It is also called data sharing (Gross and Tibshirani, 2016) since information contained in data of different cancer is shared through the common parameter $\boldsymbol{\beta}_0^{(g)}$. Finally, it has been called data enrichment (Chen et al., 2015; Asiaee et al., 2018, 2019) because you enrich your data set with pooling multiple samples from different but related data sources.

In our low-dimensional setting $p = 21 \ll n = 8895$, we do not need the regularization in Equation (1) to induce sparsity on parameters, therefore all $\lambda_i = 0$. Then, the remaining loss $\sum_t \left\|\mathbf{y}_t^{(g)} - \mathbf{M}_t(\boldsymbol{\beta}_0^{(g)} + \boldsymbol{\beta}_t^{(g)})\right\|_2^2$ is overparameterized because of unregularized (unconstrained) $\boldsymbol{\beta}_0^{(g)}$, meaning that one can chose $\boldsymbol{\beta}_0^{(g)}$ arbitrarily because only the summation of two parameters $\widetilde{\boldsymbol{\beta}}_g = \boldsymbol{\beta}_0^{(g)} + \boldsymbol{\beta}_t^{(g)}$ matters. Therefore, our tissue-aware model minimizes

$$\min_{\forall t:\boldsymbol{\beta}_t} \sum_t \left\|\mathbf{y}_t^{(g)} - \mathbf{M}_t\widetilde{\boldsymbol{\beta}}_t^{(g)} - b_t^{(g)} - b_0^{(g)}\right\|_2, \tag{2}$$

which means that all samples across tissues are shared to estimate the tissue-agnostic intercept $b_0^{(g)}$ for each gene g and then each tissue gets to learn its own linear model parameters $(\widetilde{\boldsymbol{\beta}}_g, b_t^{(g)})$. Note that Equation (2) is a special case of data sharing Equation (1) in low dimension and has more degree of freedom than only including tissue as a categorical covariate in regression.

2.6. Performance measure

We need a measure that summarizes the ability of miRs to predict expression over all genes. Mean square error (MSE) or Root MSE (RMSE) are standard measures of prediction performance of a linear regression. But since each response vector $\mathbf{y}^{(g)}$ has different variability, taking the mean of MSE or RMSE over 20,289 linear regressions is not particularly informative. One way to circumvent this issue is to work with normalized RMSE (NRMSE) where RMSE is divided by the mean, range, or interquartile range of $\mathbf{y}^{(g)}$. The problem with NRMSE, however, is that we do not know how to distinguish between good or bad prediction performance.

For these reasons, we use the R^2 statistic to report prediction performance. R^2 for the gth response is defined as

$$R_g^2 = 1 - \frac{\left\|\mathbf{y}^{(g)} - \mathbf{f}^{(g)}\right\|_2^2}{\left\|\mathbf{y}^{(g)} - \bar{\mathbf{y}}^{(g)}\right\|_2^2},$$

where $\mathbf{f}^{(g)}$ is our prediction, that is, $\mathbf{M}\boldsymbol{\beta}_0^{(g)}$ or $\mathbf{M}_t(\boldsymbol{\beta}_0^{(g)} + \boldsymbol{\beta}_t^{(g)})$ in tissue agnostic or aware models, respectively. R^2 can be thought of as a measure of the percentage of variance explained and is $0 \leq R_g^2 \leq 1$, so we can meaningfully compare the performance of regression across different responses and take its average $\bar{R}^2 = \frac{1}{m}\sum_{g=1}^{m} R_g^2$ as the overall power of miRs in predicting the transcriptome. Note that R^2 is related to MSE normalized by variance as

$$R^2 = 1 - \frac{MSE}{Var}.$$

2.7. Gene and miR enrichment analysis

To interpret miR clusters, we performed enrichment analyses using three main approaches:

1. comparing with clinically known miRs (Hydbring and Badalian-Very, 2013);
2. the miRs enrichment and annotation (miEAA) tool (Backes et al., 2016), which performs Fisher's exact tests based on the input set of miRs; and
3. the ToppGene tool (Chen et al., 2009), which performs Fisher's exact tests based on the input set of genes.

Since the input to ToppGene is a list of genes rather than a list of miRs, we calculated the gene list for each miRs cluster by selecting genes that were significantly correlated with the mean expression of miRs in the cluster.

3. RESULTS

3.1. Differentiating 32 cancers with 21 miR scores

After applying the unsupervised, nonlinear, t-SNE projection algorithm to the 8895×21 matrix $\mathbf{M}$ of miR scores, we visualized the result in two dimensions (Fig. 2). We colored the plot using the known cancer types to better understand the patterns. In general, most cancers can be separated purely based on their miR profile. There are a few examples where multiple diseases are overlapping. For instance, colon and rectal cancers almost perfectly overlap each other, which makes biological sense given the similarity of the tissues from which these cancers originate. Another intriguing example is the relationship between the three different forms of kidney cancer. While all three can be clearly distinguished, the matched normal samples form a fourth group, representing the ''same'' normal kidney profile. Overall, this t-SNE plot demonstrates that the majority of cancers can be distinguished purely based on their miR profile and overlaps tend to be based on the similarity of tissue type.

3.2. Viewing miR scores across cancer types

To determine if there were differences in the expression of miRs across the cancer types, we generated bean plots, three of which are shown in Figure 3. The bean plots were generated per cluster by plotting cancer types on the x-axis and the miR score on the y-axis. Expression levels and thus miR scores were plotted on a log scale. The bean plots illustrate the variation in expression across different cancer types for each miR cluster. This helps inform the biological interpretation of the cluster since many clusters are noteworthy for having a set of ''outlier'' diseases compared with the majority of cancer types.

Ovarian (OV) cancer in cluster 20 (Fig. 3a) is a great example, with a mean value of 2.4 compared with the next lowest cancer type with a mean value of 5.0. This demonstrates that miRs in cluster 20 are highly underexpressed in ovarian cancer compared with all other cancer types. The bean plots can also be used to find similarities across different cancers. Cluster 15 (Fig. 3b) distinguishes melanomas from other forms of cancer. This can be seen because both skin melanoma (SKCM) and uveal melanoma (UVM) have much higher mean expression than any other cancers. Again, this makes biological sense given the underlying

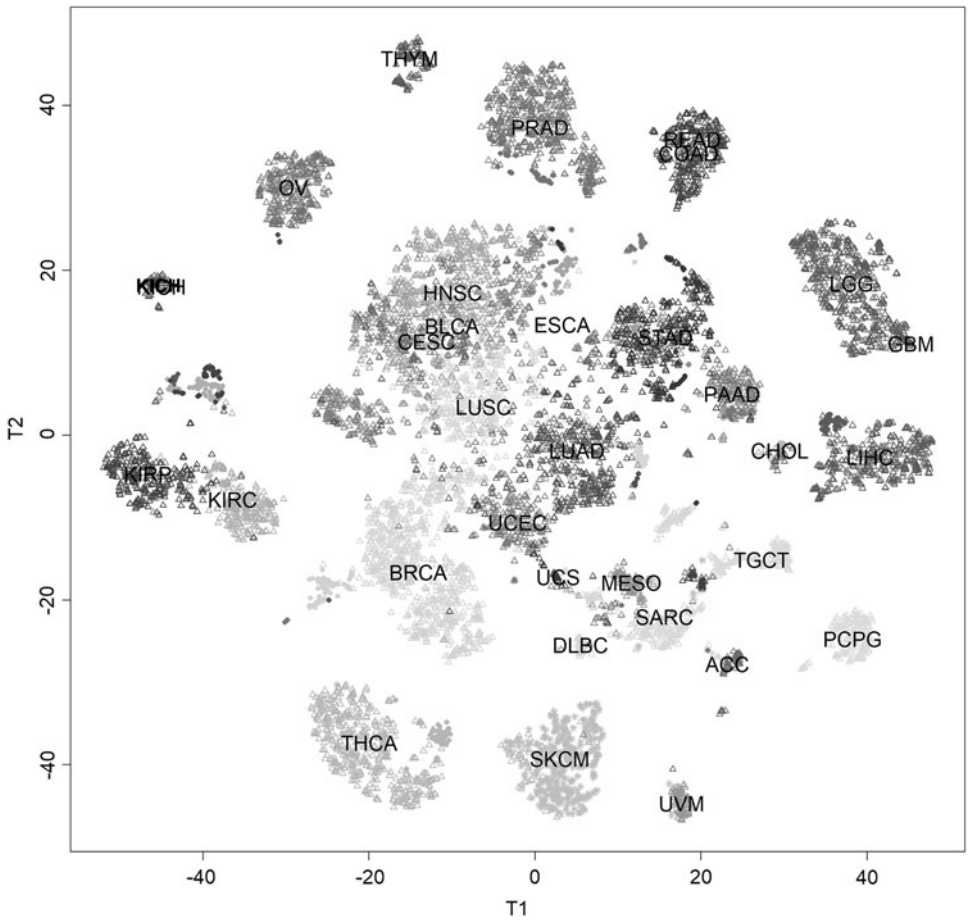

FIG. 2. Plot of the nonlinear t-SNE map of samples from 21-dimensional miR score space into two dimensions. Primary tumors are plotted with an open circle, metastases with a hollow triangle, and normal samples with an asterisk. Importantly, most of the 32 different cancer types are distinguishable using only the 21 miR clusters. This indicates that miR expression can be a distinguishing factor of many types of cancer. t-SNE, t-stochastic neighbor embedding.

similarity of the origins of these two diseases. Finally, cluster 17 (Fig. 3c) shows a similarity across different types of squamous cell cancers, namely lung squamous cell (LUSC), head and neck squamous cell (HNSC), and cervical squamous cell (CESC) carcinomas. Although these cancers arise in different organ systems, the underlying biology of these cancers appears to be similar since they come from the same progenitor cells.

3.3. Enrichment analysis of 21 miR clusters

To dig deeper into the potential interpretations of miR clusters, we performed a set of enrichment analyses (Table 1). First, a list of clinically known miRs was taken from the study ''Clinical applications of microRNAs'' (Hydbring and Badalian-Very, 2013). We cross-referenced these clinically known miRs and noted in which of the 21 clusters they were found; these are listed as ''Known Important miRs'' in Table 1. Second, we ran each cluster of miRs through the miEAA online analysis tool, which enables users to enter a list of miRs to perform a set of enrichment analyses based on Fisher's exact tests over a variety of categories, such as diseases, chromosomes, and pathways. Top hits from this analysis are presented in the ''miEAA'' column of Table 1. Finally, we performed a ToppGene enrichment analysis. Since ToppGene only uses gene lists as input and not miRs, we calculated the gene list per cluster by selecting genes that were significantly correlated with the miR score of each cluster ($|r| \geq 0.4$). We ran each of these 21 lists of genes through ToppGene and report the top results in the ''ToppGene'' column of Table 1.

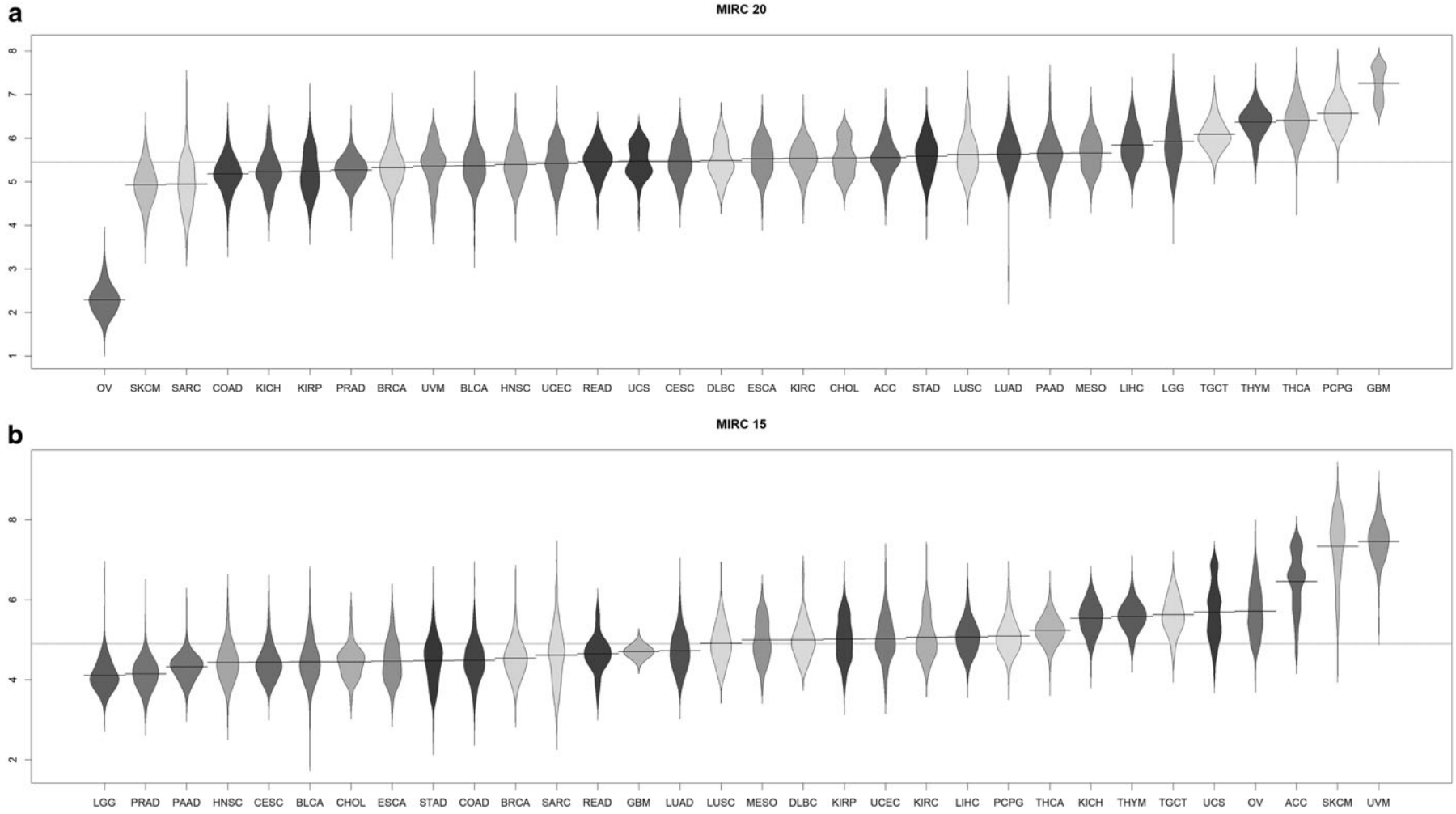

FIG. 3. Bean plots of the log of miR score associated with cancer type. **(a)** Ovarian. **(b)** Melanomas. **(c)** Squamous cell. The horizontal line is the mean of the log of the score across all cancers.

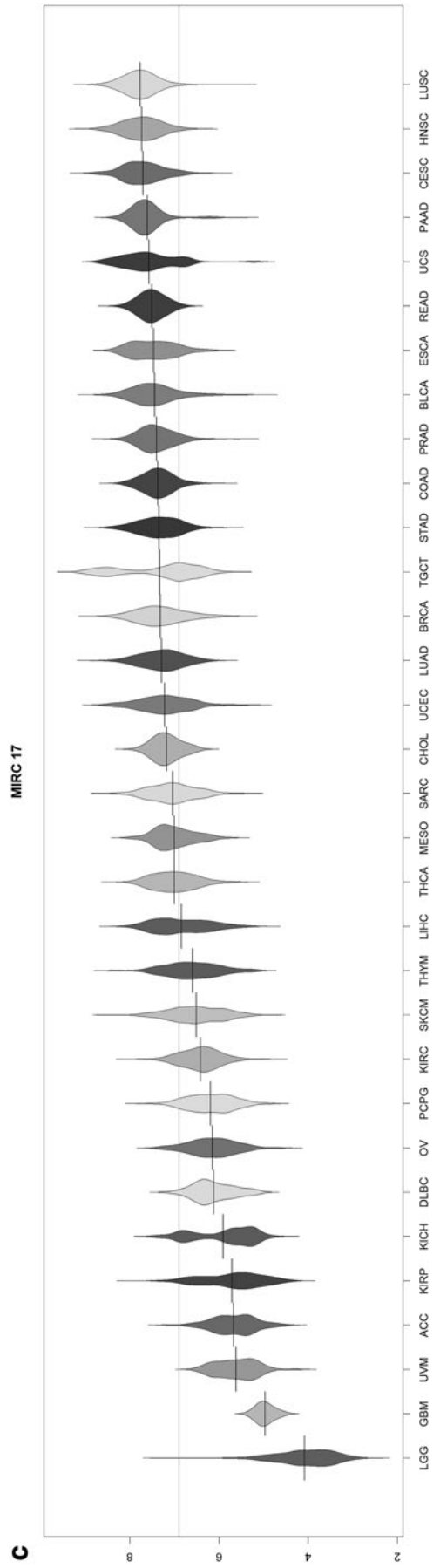

FIG. 3. (*Continued*).

TABLE 1. RESULTS OF ENRICHMENT ANALYSIS OF 21 MICRORNA CLUSTERS WITH THREE DIFFERENT METHODS

ID	*No. of miRs*	*Known important miRs*	*miEAA*	*ToppGene*
1	9	miR-196a, miR-10b	NS	Sequence-specific DNA binding
2	26	miR-142-3p, miR-21-5p, miR-31-3P, miR-34a	Chromosome 17	Enzyme inhibitor activity
3	19	let-7i, miR-29a, miR-31-5P	NS	Lymphoma, interleukin-2 binding
4	4	miR-181a, miR-130a	NS	Common carcinoma
5	46	NS	NS	Autonomic nervous system development
6	35	miR-92b, let-7e, miR-181b	Melanoma, lung neoplasms, chromosome 22	NS
7	24	miR-99a	Chromosome 11	Cell cycle
8	30	miR-363, miR-138, miR-9	Melanoma, lung cancer, pancreatic cancer	Schizophrenia, Alzheimer's disease
9	17	NS	Chromosome X	NS
10	29	miR-106b-3p, miR-345	NS	Cervix carcinoma, malignant neoplasm of ovary
11	2	NS	NS	NS
12	15	miR-148b	NS	NS
13	26	miR-19b, miR-106b-5p	Chromosome 13	RNA binding, RNA splicing
14	3	miR-193b	NS	NS
15	22	miR-509	Chromosome 8 and X	Metastatic melanoma, melanosome membrane
16	30	miR-146a, miR-210	Melanoma pathways, neoplasms	Chromosome breakage, cell cycle process
17	36	miR-152, miR-205, miR-21-3p, miR-145, miR-214, miR-193a-3p, miR-27b, miR-375	Chromosome 5	Squamous cell carcinoma, cell junction
18	5	miR-192, miR-194	NS	Liver neoplasms
19	13	miR-200c, miR-141	NS	Adenocarcinoma, squamous cell carcinoma
20	29	NS	NS	NS
21	50	miR-187, miR-193a-5p, miR-92a	Melanoma, Alzheimer's disease, renal cancer	Cell cycle, chromosomal part, chromosome breakage

miEAA, miRs enrichment and annotation; miRs, microRNAs; NS, non-significant.

3.4. Predicting gene expression across cancers with 21 miR scores

Our main goal was to understand how much of the variability of transcriptome can be explained by miRs and tissue type. To this end, we fitted

1. the *tissue-agnostic* model where features are 21 miR scores,
2. the *tissue-only* model where tissue type is a single dummy coded feature.

Then, to estimate the joint effect of the tissue type and miRs scores, we fitted

3. the *combination* of above 1 and 2 models where features are combined,
4. the *tissue-aware* model explained in Section 2.

Intuitively, the tissue-agnostic model should capture the variability due to the underlying similarity between different tissues or cancer types, and the tissue-only model should only predict the mean expression level of a gene in a tissue type, that is, $\overline{\mathbf{y}}_t^{(g)}$. Figure 4 illustrates the distribution of R^2 for the tissue-agnostic (Fig. 4a) and tissue-only (Fig. 4b) models. The tissue-agnostic model on average could explain 31% of the variation across the transcriptome but still had a large group of genes that are poorly explained,

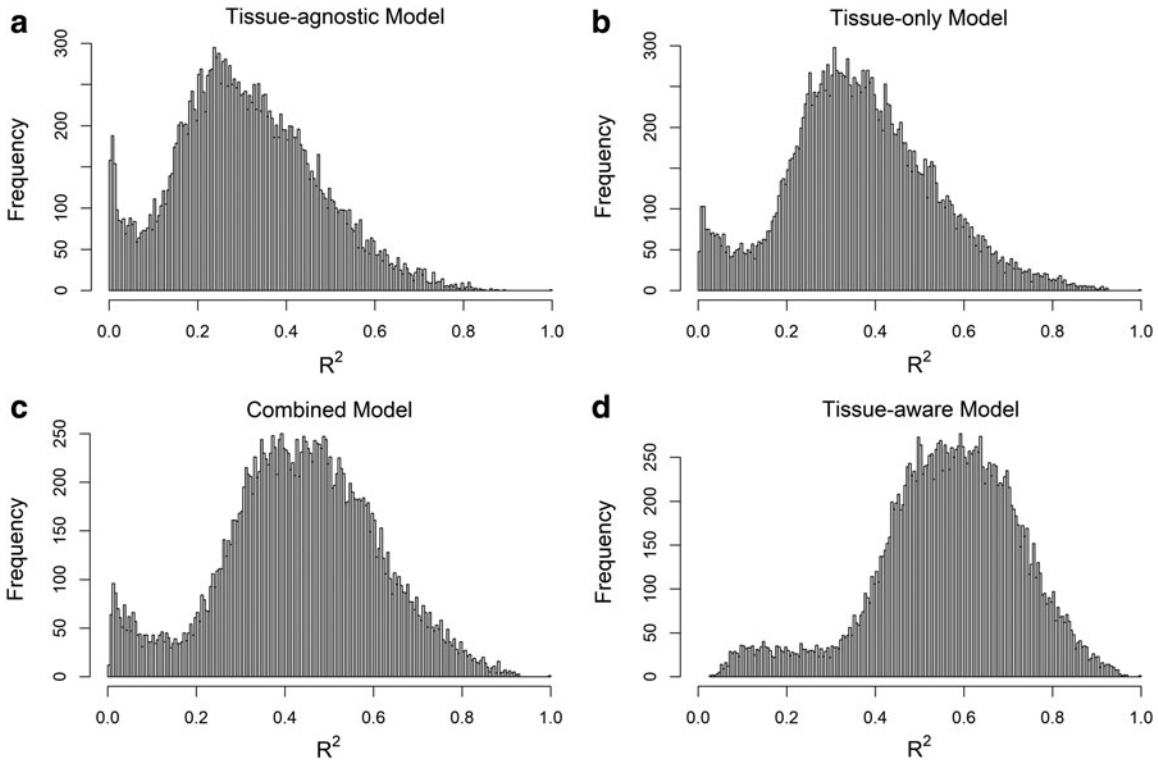

FIG. 4. Distribution of R^2 and its mean for predicting 20,289 gene expressions using 21 miR scores and tissue type: **(a)** Tissue-agnostic model: average $R^2 = 0.31$. **(b)** Tissue-only model: average $R^2 = 0.36$. **(c)** Combined model: average $R^2 = 0.43$. **(d)** Tissue-aware model: average $R^2 = 0.56$.

that is, the spike near $R^2 = 0$ in Figure 4a. The tissue-only model on average could explain 36% of the variation over all genes, which interestingly is more than the predictive power of miRs.

Combined and tissue-aware models are two separate ways of incorporating tissue information in prediction with miRs. We expected that the tissue-aware model would fit the intra-tissue variability of gene expression better because it has more degrees of freedom compared with the combined model. Figure 4c and d fulfills this expectation as the tissue-aware model could explain 56% of the variability of gene expression while 43% is the mean R^2 for the combined model. Note the large improvement (25%) of the tissue-aware model over the tissue-agnostic model. Also, the number of very poorly explained genes was substantially reduced (Fig. 4c).

Remark 1. Since the number of samples $n = 8895$ is much larger than the dimension $p = 21$, we can solve the ordinary least square regressions without any model selection. Therefore, we are not tuning any hyperparameter (like the regularization parameter in Ridge and LASSO regressions), and cross-validation is not necessary to prevent overfitting. To check the generalization error, we have split samples of each cancer type to two equal-size train and test parts. Then, we computed the parameters from training samples and calculated the R^2 from test samples. The distribution of R^2 and its mean was very similar to what we have presented above, and therefore, we are omitting that from the presentation.

Remark 2. We want to emphasize that our proposed model is a predictive model, not a *causal* one. There are a few articles that explored the causal role of miR in mRNA regulation (Zhang et al., 2014; Nalluri et al., 2017; Chen and Lu, 2018). Also, there is a new body of work in statistical machine learning literature that links causality to *invariant prediction* across heterogeneous data source (Peters et al., 2016, 2017). Using invariant prediction principles and different data sources such as TCGA and Gene-Tissue Expression (GTEx) (GTEx Consortium, 2017), we may be able to go beyond prediction and recover the causal effects of miRs in gene regulation.

Table 2. Results of Enrichment Analysis of Genes Highly or Poorly Explained by microRNAs

	Highly predictable genes	*Poorly predictable genes*
GO:MF terms	Transporter activity	Olfactory receptor activity
GO:BP terms	Organic hydroxy compound metabolic process	Sensory perception of smell, nervous system process
GO:CC terms	Plasma membrane region	Intermediate filament
Pathways	Complement and coagulation cascades	Olfactory transduction
Gene families	Solute carriers, apolipoproteins	Keratin-associated proteins, olfactory receptors

BP, biological process; CC, cellular component; GO, Gene Ontology; MF, molecular function.

Remark 3. Finally, although we are reporting prediction results when using the mean of each miR cluster as predictors (miR score), the same analysis using individual miRs that are closest to the mean as features produced the same result and therefore omitted from the presentation. This fact justifies our article title, where we announce that we use 21 miRs to explain gene expression.

3.5. Enrichment analysis of genes highly or poorly predictable by 21 miR scores

After performing the R^2 analyses, we were interested in determining if there was a pattern in the genes at either end of the R^2 distribution. In other words, is there some similarity between the genes that are highly influenced by miRs (large R^2) or that are poorly influenced by miRs (small R^2). So, we selected cutoffs for high ($R^2 > 0.8$) and low ($R^2 < 0.05$) information based on the R^2 distributions from Figure 4a. These cutoffs resulted in 35 high and 1089 low influenced genes, respectively. Both gene lists were independently run through ToppGene to perform gene enrichment analyses (Table 2).

In general, low information genes are involved in olfactory receptor processes, sensory perception, and nervous system processing. These are all systems that interpret and process signals from outside the organism, so it makes sense that you would not want to turn off or downregulate any of these external sensory systems. In contrast, the high information genes are characterized by being transporters, plasma membrane proteins, or involved in metabolic process. All these processes require the transportation or processing of internal components rather than involving outside influence. This makes it more important to regulate on a smaller scale to maintain the health of the cell. These are also processes that are more associated with the cells response to viruses, linking back to the original biological development of miRs as an antiviral factor. We also have performed the same analysis using the R^2 distribution of the tissue-aware model, and although the number of genes resulted from the thresholding the R^2 was different, the enrichment analysis results were similar to that of the tissue-agnostic model, so we only presented the result of the tissue-agnostic model.

4. DISCUSSION

The fact that Thresher was able to reduce the set of 470 miRs into 21 one-dimensional clusters illustrates the potential complexity of the role of miRs in human cancer. The majority of the 21 clusters distinguish one to three cancer types from the remaining cancer types based on differential expression of miRs. Thus, it seems likely that these separations are largely determined by tissue type as opposed to a global mechanism of cancer. However, we also found that the miR profiles in cancer are different from the profiles of their corresponding normal tissues. This can be seen in kidney, lung, and breast cancer in the t-SNE plot (Fig. 2). In all cases, the samples from normal tissue can be clearly seen as a separate entity somewhat removed from the cancer samples. In the case of kidney cancers, the plot also shows that different forms of cancer from the same organ can develop different distinct miR expression profiles.

The enrichment analyses of the 21 miR clusters show some interesting results (Table 1). Overall, there was an uneven distribution of prior information across the clusters. Some clusters, such as 16, contain a large number of clinically known miRs and significant enrichment results for both the miRs and gene lists. However, other clusters, such as 20, had no prior information or significant enrichment findings. This may indicate that some of these miRs have been associated with each other in the literature, whereas others have no such literature associations. Another interesting finding is the high number of chromosomes pulled out

of the miEAA analysis. Seven of the 21 clusters had a significant association with an individual chromosome. This indicates that there may be a regulatory connection between miRs that are physically located near each other on the same chromosome.

The R^2 results for both the tissue-aware and tissue-agnostic models help explain how miRs and tissue specificity affect gene expression. In the tissue-agnostic model, only the miRs are taken into account when calculating the linear models to generate R^2 values. Thus, only the miR expression affects each genes' R^2 value. Since the average value in the tissue-agnostic model is $R^2 = 0.31$, through extrapolation, $\sim 30\%$ of all transcriptomic variations are due to miR expression patterns. The tissue-aware model differed from the tissue-agnostic model by incorporating cancer type into the linear model. Thus, the tissue-aware model used tissue-specific gene expression patterns as part of the calculation of R^2 values. This explains the substantial increase in the average R^2 value, increasing from 0.31 to 0.56 in the tissue-aware model. However, the tissue-only model also had an average R^2 value of ~ 0.36. The tissue-only model does not take into consideration the miR clusters but does consider the underlying tissue-specific transcriptomic differences among cancers. Thus, $\sim 36\%$ of all transcriptomic variations are tissue-specific. Taking both tissue and miR patterns into account yields a nonadditive average R^2 value of 0.56. This indicates that miR cluster expression and tissue-specific gene expression patterns account for $\sim 56\%$ of all transcriptomic variations. This is a major step forward in understanding how the human transcriptome is influenced and regulated.

5. FUTURE DIRECTIONS

Based on our findings, there are multiple future directions to explore. First of all, we think that by adding other regulatory elements such as transcription factors and methylation as features to our analysis, we should be able to explain much more of the transcriptome's variability. Exploring biological interpretation of genes that are highly or poorly explained by each regulatory element separately and also jointly may shed light on important underlying biological pathways that regulate gene expression across tissue types. In addition, performing the same analysis on healthy samples such as those in the GTEx (GTEx Consortium, 2017) and comparing our results on TCGA cancer samples should provide us with new insights into the role of regulatory elements in cancer.

From a methodological point of view, there are several interesting directions to explore. First, our preliminary analysis shows that although the infrequent miRs that we discard in preprocessing are zero-inflated, they are also tissue-specific and may contain valuable information. Therefore, we are looking for methods to go beyond simple thresholding nonzero features to systematically deal with the zero-inflation problem, which is present in many other computation biology application (den Berge et al., 2018).

Second, we want to compare Thresher feature extraction findings with that of other automatic feature selection methods based on regularization, such as Ridge regression and LASSO (Hastie et al., 2009). An interesting avenue for exploration is combining Thresher's extracted feature clusters with more complicated penalized regression method like group LASSO (Yuan and Lin, 2006) and sparse group LASSO (Simon et al., 2013).

Finally, we are treating the prediction of expression of each gene as a separate task. In reality, many gene expressions are correlated and modeling these relations explicitly may boost the prediction performance. A suitable machine learning tool for predicting the related outcomes is multiresponse models where the goal is to simultaneously fit regression models for each task and learn the covariance structure between the outcomes (Kim and Xing, 2009; Chen and Banerjee, 2017).

6. CONCLUSION

In this article, we determined the amount of variability that miRs play in influencing transcriptomic expression patterns. Using data from TCGA, we were able to break down all miRs into 21 one-dimensional clusters. These 21 clusters explained 31% of the total variability found in human transcriptomic expression within the TCGA cohort. When combined with tissue information, miRs could explain 56% of the transcriptome. This result helps explain the amount of regulatory power that miR exert across the human transcriptome and how different miRs are differentially associated with specific tissue and cancer types.

ACKNOWLEDGMENTS

The authors thank the Mathematical Biosciences Institute (MBI) and the Ohio State University Comprehensive Cancer Center—James for their supports. We also thank Jared Huling for helpful comments about the data sharing model.

AUTHOR DISCLOSURE STATEMENT

The authors declare they have no competing financial interests.

FUNDING INFORMATION

The authors would like to thank the Mathematical Biosciences Institute (MBI) at Ohio State University for partially supporting this research. MBI receives its funding through the National Science Foundation grant DMS 1440386. Also, support was provided by U.S. NIH grant P30 CA016058 and by startup funds (KRC) from the Ohio State University College of Medicine. Finally, DS acknowledges support from the Leukemia & Lymphoma Society's grant 6538–18.

REFERENCES

Abrams, Z.B., Zucker, M., Wang, M., et al. 2018. Thirty biologically interpretable clusters of transcription factors distinguish cancer type. *BMC Genomics* 19, 738.

Asiaee, A., Oymak, S., Coombes, K.R., et al. 2018. High dimensional data enrichment: Interpretable, fast, and data-efficient. *arXiv* 1806.04047.

Asiaee, A., Oymak, S., Coombes, K.R., et al. 2019. Data enrichment: Multi-task learning in high dimension with theoretical guarantees. In Gretton, A., and Lee, H. (eds.): *Adaptive and Multitask Learning Workshop at the ICML.* IMLS, Long Beach, CA.

Auer, P., and Gervini, D. 2008. Choosing principal components: A new graphical method based on Bayesian model selection. *Commun. Stat. Simul. Comput.* 37, 962–977.

Backes, C., Khaleeq, Q.T., Meese, E., et al. 2016. miEAA: microRNA enrichment analysis and annotation. *Nucleic Acids Res.* 44(W1), W110–W116.

Banerjee, A., Dhillon, I.S., Ghosh, J., et al. 2005. Clustering on the unit hypersphere using von Mises-Fisher distributions. *J. Mach. Learn. Res.* 6, 1345–1382.

Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., et al. 2013. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* 45, 1113–1120.

Chen, A., Owen, A.B., and Shi, M. 2015. Data enriched linear regression. *Electron. J. Stat.* 9, 1078–1112.

Chen, J., Bardes, E.E., Aronow, B.J., et al. 2009. ToppGene suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* 37, W305–W311.

Chen, L., and Lu, X. 2018. Discovering functional impacts of miRNAs in cancers using a causal deep learning model. *BMC Med. Genomics* 11(Suppl 6), 116.

Chen, S., and Banerjee, A. 2017. Alternating estimation for structured high-dimensional multi-response models. *Adv. Neural Inf. Process. Syst.* 30, 2838–2848.

Daige, C.L., Wiggins, J.F., Priddy, L., et al. 2014. Systemic delivery of a miR34a mimic as a potential therapeutic for liver cancer. *Mol. Cancer Ther.* 13, 2352–2360.

den Berge, K., Perraudeau, F., Soneson, C., et al. 2018. Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biol.* 19, 24.

Dragomir, M., Mafra, A.C.P., Dias, S.M.G., et al. 2018. Using microRNA networks to understand cancer. *Int. J. Mol. Sci.* 19.

Garzon, R., Marcucci, G., and Croce, C.M. 2010. Targeting microRNAs in cancer: Rationale, strategies and challenges. *Nat. Rev. Drug Discov.* 9, 775–789.

Gross, S.M., and Tibshirani, R. 2016. Data shared Lasso: A novel tool to discover uplift. *Comput. Stat. Data Anal.* 101, 226–235.

GTEx Consortium. 2017. Genetic effects on gene expression across human tissues. *Nature* 550, 204–213.

Hastie, T., Tibshirani, R., and Friedman, J. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)*. Springer, New York, NY.

He, L., and Hannon, G.J. 2004. MicroRNAs: Small RNAs with a big role in gene regulation. *Nat. Rev. Genet.* 5, 522–531.

Houzet, L., and Jeang, K.-T. 2011. MicroRNAs and human retroviruses. *Biochim. Biophys. Acta* 1809, 686–693.

Hydbring, P., and Badalian-Very, G. 2013. Clinical applications of microRNAs. *F1000Research* 2, 136.

Iorio, M.V., and Croce, C.M. 2012. MicroRNA dysregulation in cancer: Diagnostics, monitoring and therapeutics. A comprehensive review. *EMBO Mol. Med.* 4, 143–159.

Jalali, A., Ravikumar, P., Sanghavi, S., et al. 2010. A dirty model for multi-task learning. *Adv. Neural Inf. Process. Syst.* 23, 964–972.

Kim, S., and Xing, E.P. 2009. Tree-guided group Lasso for multi-response regression with structured sparsity, with an application to eQTL mapping. *arXiv* 0909.1373.

Melnik, B.C. 2015. MiR-21: An environmental driver of malignant melanoma? *J. Transl. Med.* 13, 202.

Mortazavi, A., Williams, B.A., McCue, K., et al. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628.

Nalluri, J.J., Rana, P., Barh, D., et al. 2017. Determining causal miRNAs and their signaling cascade in diseases using an influence diffusion model. *Sci. Rep.* 7, 8133.

Peng, Y., and Croce, C.M. 2016. The role of MicroRNAs in human cancer. *Signal Transduct. Target. Ther.* 1, 15004.

Peters, J., Bühlmann, P., and Meinshausen, N. 2016. Causal inference by using invariant prediction: Identification and confidence intervals. *J. R. Stat. Soc. Series B. Stat. Methodol.* 78, 947–1012.

Peters, J., Janzing, D., and Schölkopf, B. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA.

Riffo-Campos, Á. L., Riquelme, I., and Brebi-Mieville, P. 2016. Tools for sequence-based miRNA target prediction: What to choose? *Int. J. Mol. Sci.* 17, 1987.

Rupaimoole, R., and Slack, F.J. 2017. MicroRNA therapeutics: Towards a new era for the management of cancer and other diseases. *Nat. Rev. Drug Discov.* 16, 203–222.

Simon, N., Friedman, J., Hastie, T., et al. 2013. A sparse-group Lasso. *J. Comput. Graphical Stat.* 22, 231–245.

van der Maaten, L., and Hinton, G. 2008. Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.

Wang, M., Abrams, Z.B., Kornblau, S.M., et al. 2018. Thresher: Determining the number of clusters while removing outliers. *BMC Bioinformatics* 19, 9.

Yuan, M., and Lin, Y. 2006. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Series B. Stat. Methodol.* 68, 49–67.

Zhang, J., Le, T.D., Liu, L., et al. 2014. Identifying direct miRNA-mRNA causal regulatory relationships in heterogeneous data. *J. Biomed. Inform.* 52, 438–447.

Zhang, Y., and Yang, Q. 2017. A survey on multi-task learning. *arXiv* 1707.08114.

Address correspondence to:
Dr. Amir Asiaee
Mathematical Biosciences Institute
The Ohio State University
379 Jennings Hall, 3rd Floor
1735 Neil Ave.
Columbus, OH 43210
USA

E-mail: asiaeetaheri.1@gmail.com

Dr. Zachary B. Abrams
Department of Biomedical Informatics
The Ohio State University
310D Lincoln Tower
1800 Cannon Drive
Columbus, OH 43210
USA

E-mail: zachary.abrams@osumc.edu