Explaining Gene Expression Using Twenty-One MicroRNAs

Amir Asiaee [*1,2] **Zachary B. Abrams** [*1] **Samantha Nakayiza** [1] **Deepa Sampath** [3] **Kevin R. Coombes** [1]

Abstract

The transcriptome, or gene expression profile, of a tumor contains detailed information about the disease. Although advances in sequencing technologies have generated larger and more informative data sets, there are still many questions about 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 microRNAs and different cancers, we analyzed ∼9000 samples from 32 cancer types studies in The Cancer Genome Atlas (TCGA). Using the Thresher R package to perform feature reduction, we found evidence for 21 biologically interpretable clusters of miRs. Many of these clusters 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 ("score") of the clusters. Based on the ∼20,000 per-gene R^2 values, we found that (a) mean differences between cancer types explain about 40% of variation; (b) the 21 miR cluster scores explain about 30% of variation, and (c) combining cancer 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 down-regulated 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. The distribution of R^2 values suggests that other genetic regulatory elements such as transcription factors (TF) and methylation would help explain some of the remaining variation in gene expression. By building a combined microRNA-TF-methylation model, we can potentially predict the majority of human transcriptomic expression.

1. Introduction

MicroRNAs (miRs) are a class of non-coding 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 & Hannon, 2004; Rupaimoole & Slack, 2017). This powerful form of gene regulation was evolutionarily developed as a way to protect cells against retroviruses (Houzet & Jeang, 2011). The miR-mRNA interaction is entangled since each miR can regulate hundreds of mRNAs, while each mRNA can be controlled by multiple miRs. Further, miRs regulate and are regulated by different types of long non-coding RNAs (lncRNAs) (Peng & Croce, 2016; Garzon et al., 2010). In cancer, a single miR can act as either an oncogene or a tumor-suppressor in different contexts (Garzon et al., 2010). Two notable example 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; Peng & Croce, 2016; Iorio & Croce, 2012). 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

[*]Equal contribution [1]Department of Biomedical Informatics, The Ohio State University, Columbus, USA [2]Mathematical Biosciences Institute, The Ohio State University, Columbus, USA [3]Department of Internal Medicine, Division of Hematology, The Ohio State University, Columbus, USA. Correspondence to: Amir Asiaee <asiaeetaheri.1@osu.edu>, Zachary B. Abrams <Zachary.Abrams@osumc.edu>, Kevin R. Coombes <coombes.3@osu.edu>.

Workshop on Computational Biology at the International Conference on Machine Learning, Long Beach, CA, 2019.

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 paper, 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 variation in gene expression.

From 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 microRNA 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, pre-processing, and analysis methods. Figure 1 summarizes the steps of our analysis.

Data. The data used in this study were collected by The Cancer Genome Atlas (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 microRNA sequencing data. In total, $n = 8895$ patient samples reflecting 32 different cancer types were obtained.

Processing and Filtering. We normalized the sequencing data from individual samples by computing reads per kilobase per million (RPKM) (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, $p = 470$ microRNAs remained. The following steps are performed on the 8895×470 data matrix.

Feature Extraction. After data processing and filtering, we analyzed the microRNA 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 (Auer & 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 is determined as $d = 21$ and the final number of identified miR clusters 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 of the following analysis is performed on the resulting 8895×21 miR score matrix, $\mathbf{M}$.

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 (Maaten & Hinton, 2008) as implemented in version 0.13 of the Rtsne R package.

Prediction. We use the score matrix $\mathbf{M}$ to predict 20289 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 *global model*. Therefore, for each gene $g \in 1, \ldots, 20289$, we are minimizing $\|\mathbf{y}^{(g)} - \mathbf{M}\boldsymbol{\beta}_0^{(g)}\|_2$ to find the coefficients $\boldsymbol{\beta}_0^{(g)}$ of the global model, where

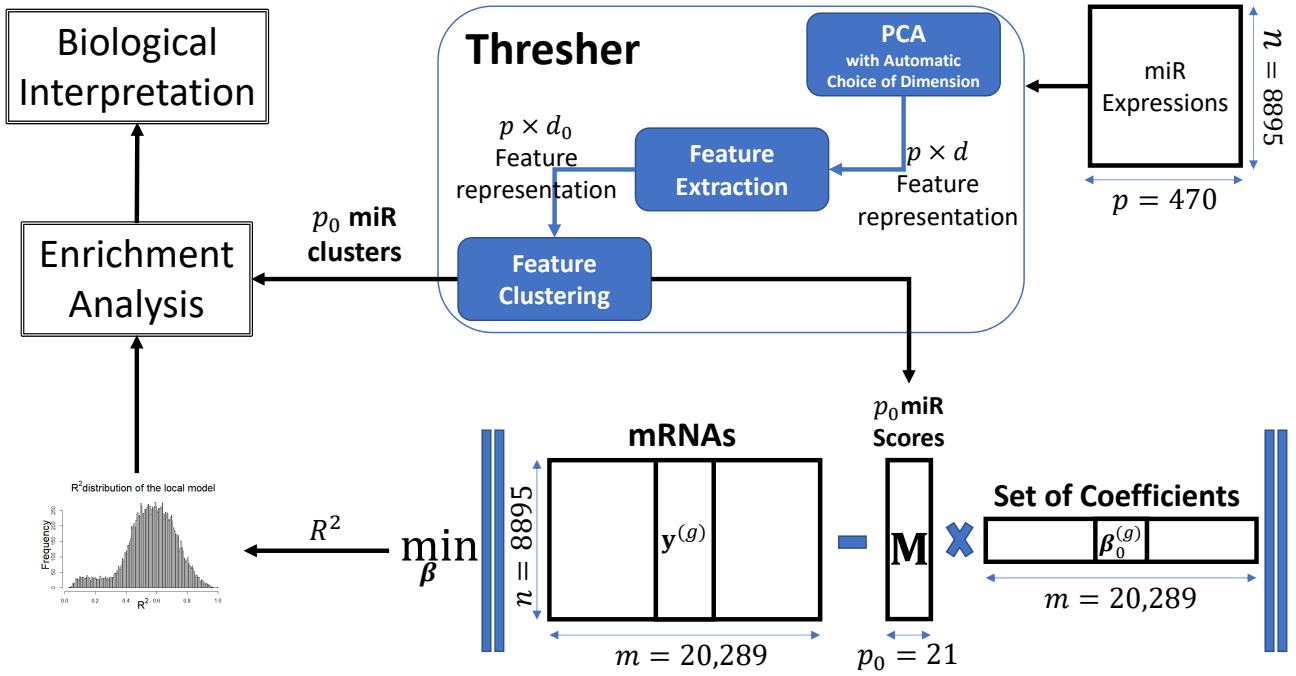


Figure 1. Workflow of our analysis. Starting from miR expression data (top right), we use the Thresher algorithm to extract biologically meaningful cluster of miRs and perform enrichment analysis to interpret them (top left). We also use the scores (mean expression of miRs in each cluster) of miR clusters 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).

$\mathbf{y}^{(g)} \in \mathbb{R}^n$ 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 fitted cancer-specific linear models to predict the residuals $\mathbf{r}^{(g)} = \mathbf{y}^{(g)} - \mathbf{M}\boldsymbol{\beta}_0^{(g)}$ of the global model. We call this set of linear models the *local model*, which is equivalent to minimizing $\|\mathbf{y}_t^{(g)} - \mathbf{M}_t(\boldsymbol{\beta}_0^{(g)} + \boldsymbol{\beta}_t^{(g)})\|_2$ where $\mathbf{M}_t$ contains rows of $\mathbf{M}$ for cancer type t and $\boldsymbol{\beta}_t^{(g)}$ is the corresponding cancer-specific parameter.

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

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

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

where $\mathbf{f}^{(g)}$ is our prediction, i.e., $\mathbf{M}\boldsymbol{\beta}_0^{(g)}$ or $\mathbf{M}_t(\boldsymbol{\beta}_0^{(g)} + \boldsymbol{\beta}_t^{(g)})$ in global or local 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

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

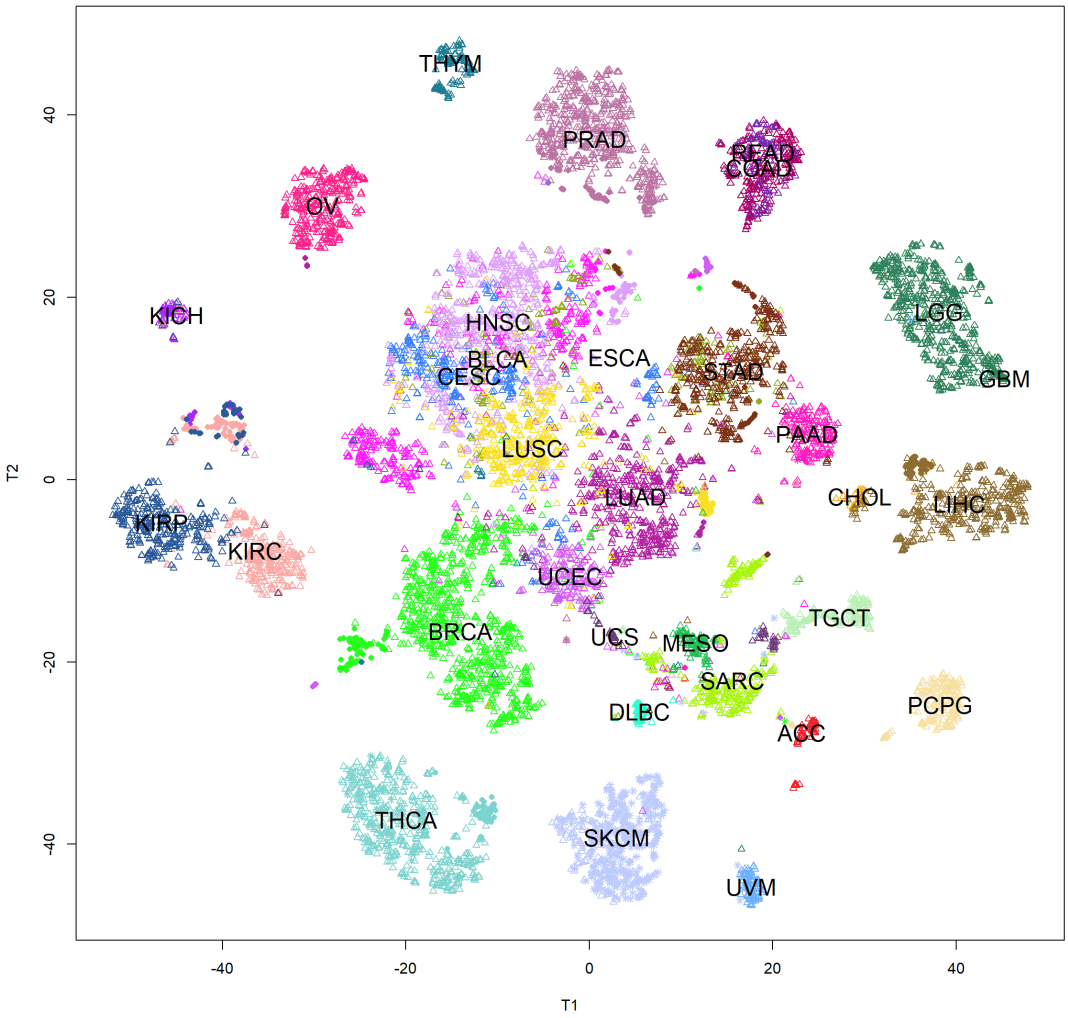


Figure 2. Plot of the non-linear t-SNE map of samples from 21-dimensional miR score space into two dimensions. Primary tumors are plotted with an open circle, metastases with a hollow triangle, and normal samples with an asterisk. Importantly most of the 32 different cancer types are distinguishable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

Gene and MicroRNA Enrichment Analysis. To interpret miR clusters, we performed enrichment analyses using three main approaches:

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

ples 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 Fig 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 to the majority of cancer types.

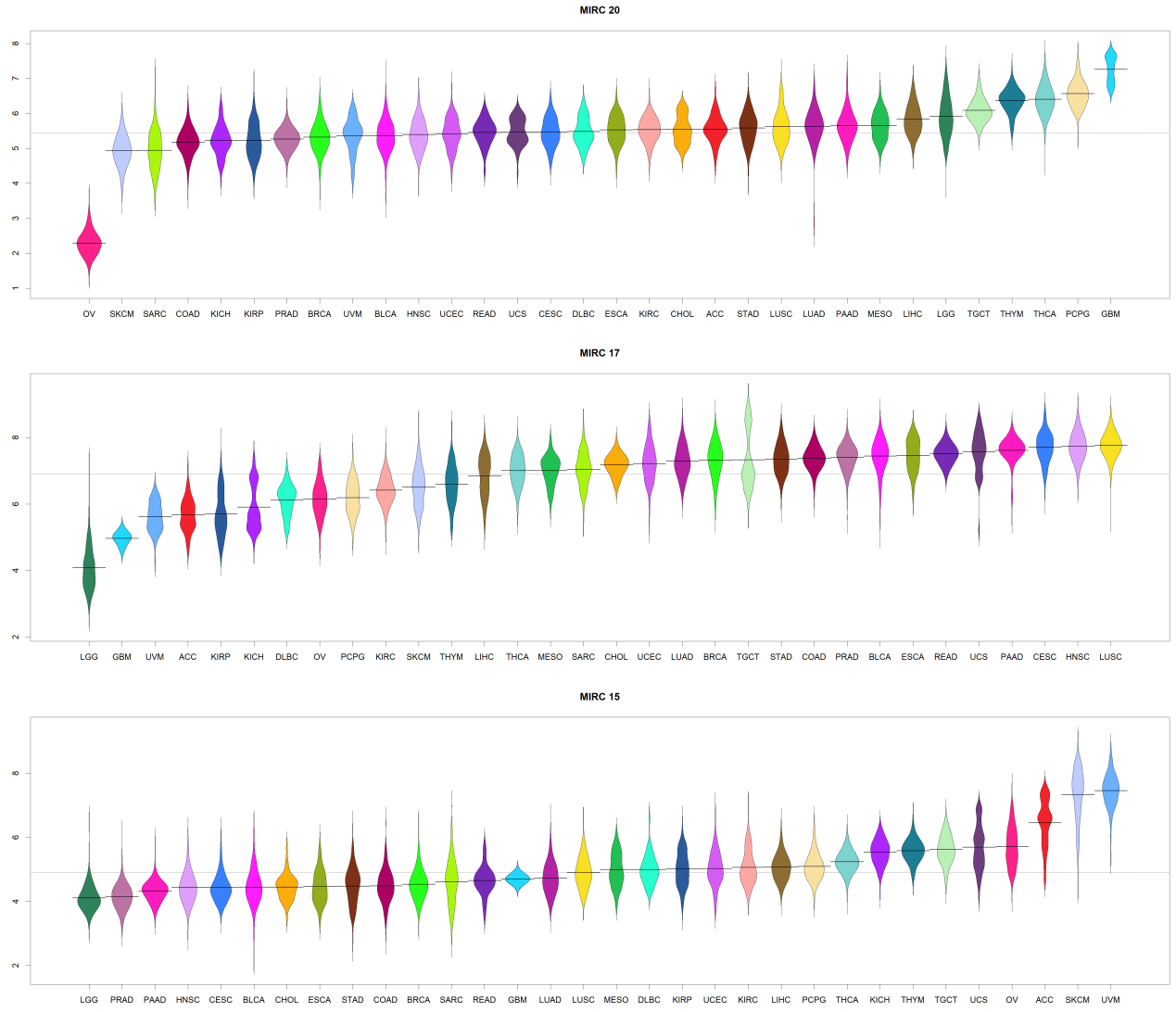


Figure 3. Bean plots of the log of miR score associated with cancer type. Horizontal line is the mean of the log of the score across all cancers. Important to note cancer type per sub-figure are: a) Ovarian b) Squamous cell c) Melanomas.

Ovarian (OV) cancer in cluster 20 (Fig 3A) is a great example, with a mean value of 2.4 compared to the next lowest cancer type with a mean value of 5.0. This demonstrates that miRs in cluster 20 are highly under expressed in ovarian cancer compared to all other cancer types. The bean plots can also be used to find similarities across different cancers. Cluster 17 (Fig 3B) 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. Finally, cluster 15 (Fig 3C) 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 cancer. Again, this makes biological sense given the underlying similarity of the origins of these two diseases.

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 & 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 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 ($|r| \geq .4$). We ran each of these 21 lists of genes through ToppGene, and report the top results in the "ToppGene" column of Table 1.

3.4. Predicting gene expression across cancers with 21 miR scores

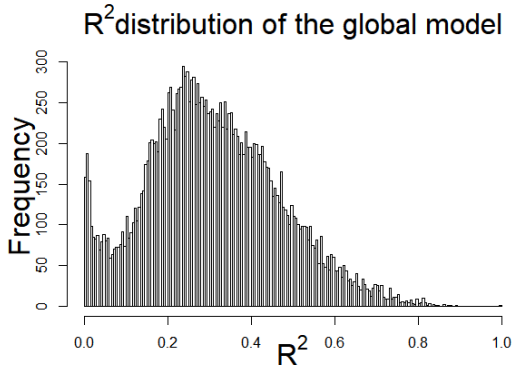
Our main goal was to understand how much of the variability of transcriptome can be explained by miRs. To this end, we fitted the global and local models explained in Section 2. Intuitively, the global model should capture the variability

ID	#of miRs	Known Important miRs	miEAA	ToppGene
1	9	miR-196a, miR-10b	NA	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	NA	Lymphoma, Interleukin-2 binding
4	4	miR-181a, miR-130a	NA	Common carcinoma
5	46	NA	NA	Autonomic nervous system development
6	35	miR-92b, let-7e, miR-181b	Melanoma, Lung Neoplasms, Chromosome 22	NA
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	NA	Chromosome X	NA
10	29	miR-106b-3p, miR-345	NA	Cervix carcinoma, Malignant neoplasm of ovary
11	2	NA	NA	NA
12	15	miR-148b	NA	NA
13	26	miR-19b, miR-106b-5p	Chromosome 13	RNA binding, RNA splicing
14	3	miR-193b	NA	NA
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	NA	Liver neoplasms
19	13	miR-200c, miR-141	NA	Adenocarcinoma, Squamous cell carcinoma
20	29	NA	NA	NA
21	50	miR-187, miR-193a-5p, miR-92a	Melanoma, Alzheimers Disease, RenalCcancer	Cell Cycle, chromosomal part, Chromosome Breakage

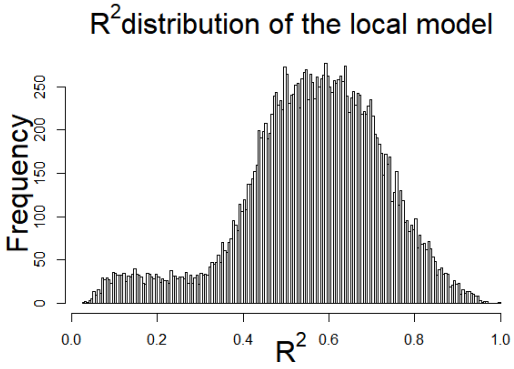
Table 1. Results of enrichment analysis of 21 miR clusters with three different methods.

due to the underlying similarity between different tissues or cancer types, and the local model should fit the intra-tissue variability of gene expression. Figure 4 illustrates the distribution of R^2 for the global (Fig 4a) and local (Fig 4b) models. The global model on average could explain 31% of the variation across the transcriptome, but still had a large group of genes that are poorly explained, i.e., the spike near $R^2 = 0$ in Fig 4a. When the tissue information was included in the analysis, the local model could explain 56% of the variability of gene expression, a large improvement over the global model. Also, the number of very poorly explained genes was substantially reduced, Fig. 4b.

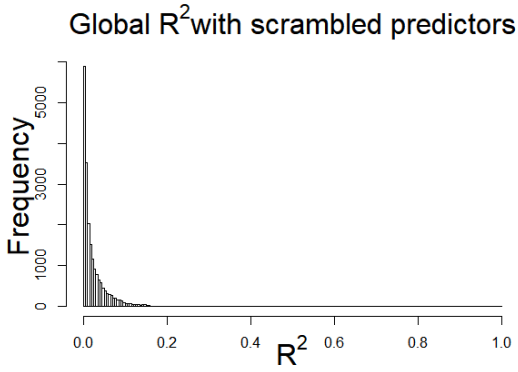
To test whether the performance of either model is due to random chance (and also to isolate the predictive power of tissue type), we supplemented our analysis with *scrambled* versions of our original prediction experiment. We permuted features of each sample independently to break any relation between the features and the outcome. In other words, we replaced each row i of the score matrix $\mathbf{M}$ with a random permutation π_i specific to that row, i.e., $\mathbf{m}_i = \pi_i(\mathbf{m}_i)$. As we expected in the global model (Fig. 4c) the mean R^2 substantially dropped and became almost zero. To our initial surprise, the R^2 mean for the scrambled local model was 0.43 which is higher than the un-scrambled global model but lower than the local one (Fig. 4d). Since we have broken the relation between features and outcome the only part of the model that can meaningfully contribute to this performance is the **intercept** of the linear model. And we know that the intercept in the linear model is $\bar{\mathbf{y}}^{(g)} - \boldsymbol{\beta}_t^{(g)} \bar{\mathbf{x}}^{(g)}$. Since we have scrambled each row separately, $\bar{\mathbf{x}}^{(g)} \in \mathbb{R}^{21}$ (the average vector of all features) should not contribute to the prediction power. Therefore, all of the transcriptome variability that we can explain in the scrambled local model is coming from the average per-tissue expression, $\bar{\mathbf{y}}^{(g)}$. The fact that the average gene expressions per tissue is a strong predictor of the gene expression biologically makes sense but the amount of that and its overlap with the predictive power of miRs have not previously been established. Going back to the poorly explained genes in the un-scrambled global model (Fig. 4a), we observed that most of these genes are only expressed in specific tissues and therefore, with



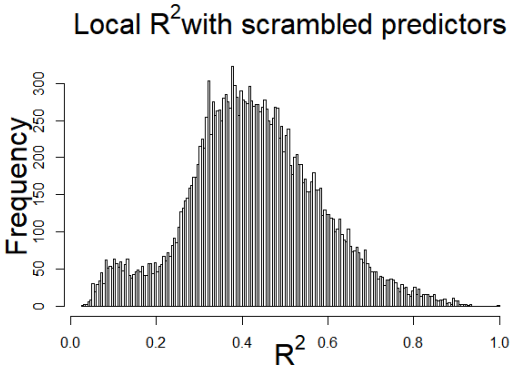
(a) $\bar{R}^2 = 0.31$



(b) $\bar{R}^2 = 0.56$



(c) $\bar{R}^2 = 0.02$



(d) $\bar{R}^2 = 0.43$

Figure 4. Distribution of $R^2$and its mean for predicting $20,289$ gene expressions using 21 miR scores. a) Global model. b) Local model. c) Scrambled global model. d) Scrambled local model.

tissue type information, it is easier to predict them. This fact can be deduced from the difference between the R^2 distribution of un-scrambled global and local models where the the spike near $R^2 = 0$ in Fig. 4a has been smoothed out in Fig. 4b.

Remark 1. Note that since the number of samples $n = 8895$ is much larger than the dimension $p = 21$, we can solve the ordinary least square for both global and local regressions without any model selection. Therefore, we are not tuning any hyper-parameter (like the regularization parameter of Ridge and LASSO regressions), and cross-validation is not necessary to prevent over-fitting. To check the generalization error, we have split samples of each cancer type 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 papers that explored the causal role of miR in mRNA regulation (Chen & Lu, 2018; Zhang et al., 2014; Nalluri et al., 2017). 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.

Remark 3. Finally, although we are reporting prediction results when using mean of each miR cluster as predictors (miR score), the same analysis using the closest to the mean miRs as features produced the same result and therefore omitted from presentation. This fact justifies our paper title.

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 cut offs for high ($R^2 > 0.8$) and low ($R^2 < 0.05$) information based on the R^2 distributions from Figure 4a. These cut offs resulted in 35 high and 1089 low influenced genes, respectively. Both gene lists were independently run through ToppGene to perform 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 which interpret and process signals from outside the organism, so it makes sense that you would not want to turn off or down regulate 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 local model (Fig. 4b) 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 global model, so we only presented the result of the global 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 1 to 3 cancer types from the remaining cancer types based on differential expression of microRNAs. 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.

	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

Table 2. Results of enrichment analysis of genes highly or poorly explained by miRs. (GO= Gene Ontology; BP = Biological Process; CC = Cellular Component; MF = Molecular Function)

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 local and global models help explain how miRs and tissue-specificity affect gene expression. In the global 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 global model is R^2 = 0.31, through extrapolation approximately 30% of all transcriptomic variation is due to miR expression patterns. The related global scrambled model emphasizes this point since the mean R^2 value is almost zero. This scrambled model shows that the only variability being taken into account is the miR clusters themselves. This indicates that the clusters Thresher generated have clear biological meaning.

The local model differed from the global model by incorporating cancer type into the linear model. Thus, the local 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 local model. However, the local scrambled model also had an average R^2 value of approximately 0.43, dramatically higher than the global scrambled model. The local scrambled model does not take into consideration the microRNA clusters but does consider the underlying tissue-specific transcriptomic differences among cancers. Thus, approximately 40% of all transcriptomic variation is tissue-specific. Taking both tissue and miR patterns into account yields a non-additive average R^2 value of 0.56. This indicates that miR cluster expression and tissue-specific gene expression patterns account for approximately 56% of all transcriptomic variation. 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 non-zero features to systematically deal with the zero-inflation problem which is present in many other computation biology application (Van 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 & 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 multi-response models where the goal is to simultaneously fit regression models for each task and learn the covariance structure between the outcomes (Kim & Xing, 2009; Chen & Banerjee, 2017).

6. Conclusion

In this paper we determined the amount of variability that miRs play in influencing transcriptomic expression patterns. Using data from The Cancer Genome Atlas (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. 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.

Acknowledgements

The authors thank the Mathematical Biosciences Institute (MBI) and the Ohio State University Comprehensive Cancer Center – James for their supports.

References

Abrams, Z. B., Zucker, M., Wang, M., Asiaee Taheri, A., Abruzzo, L. V., and Coombes, K. R. Thirty biologically interpretable clusters of transcription factors distinguish cancer type. *BMC genomics*, 19(1):738, 2018.

Asiaee, A., Oymak, S., Coombes, K. R., and Banerjee, A. High dimensional data enrichment: Interpretable, fast, and Data-Efficient. 2018.

Auer, P. and Gervini, D. Choosing principal components: A new graphical method based on bayesian model selection. *Communications in Statistics - Simulation and Computation*, 37(5):962–977, 2008.

Backes, C., Khaleeq, Q. T., Meese, E., and Keller, A. miEAA: microRNA enrichment analysis and annotation. *Nucleic acids research*, 44(W1):W110–6, 2016.

Banerjee, A., Dhillon, I. S., Ghosh, J., and Sra, S. Clustering on the unit hypersphere using von Mises-Fisher distributions. *Journal of machine learning research: JMLR*, 6: 1345–1382, 2005.

Cancer Genome Atlas Research Network, Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M. The cancer genome atlas Pan-Cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

Chen, A., Owen, A. B., and Shi, M. Data enriched linear regression. *Electronic journal of statistics*, 9(1):1078–1112, 2015.

Chen, J., Bardes, E. E., Aronow, B. J., and Jegga, A. G. ToppGene suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic acids research*, 37: W305–11, 2009.

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

Chen, S. and Banerjee, A. Alternating estimation for structured High-Dimensional Multi-Response models. In *Advances in Neural Information Processing Systems 30*, pp. 2838–2848. 2017.

Daige, C. L., Wiggins, J. F., Priddy, L., Nelligan-Davis, T., Zhao, J., and Brown, D. Systemic delivery of a mir34a mimic as a potential therapeutic for liver cancer. *Molecular cancer therapeutics*, 13(10):2352–2360, 2014.

Dragomir, M., Mafra, A. C. P., Dias, S. M. G., Vasilescu, C., and Calin, G. A. Using microRNA networks to understand cancer. *International journal of molecular sciences*, 19(7), 2018.

Garzon, R., Marcucci, G., and Croce, C. M. Targeting microRNAs in cancer: rationale, strategies and challenges. *Nature reviews. Drug discovery*, 9(10):775–789, 2010.

Gross, S. M. and Tibshirani, R. Data shared lasso: A novel tool to discover uplift. *Computational statistics & data analysis*, 101:226–235, 2016.

GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, 2017.

Gu, Q. and Banerjee, A. High dimensional structured superposition models. In *Advances in Neural Information Processing Systems 29*, pp. 3691–3699. 2016.

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

He, L. and Hannon, G. J. MicroRNAs: small RNAs with a big role in gene regulation. *Nature reviews. Genetics*, 5 (7):522–531, 2004.

Houzet, L. and Jeang, K.-T. MicroRNAs and human retroviruses. *Biochimica et biophysica acta*, 1809(11-12): 686–693, 2011.

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

Iorio, M. V. and Croce, C. M. MicroRNA dysregulation in cancer: diagnostics, monitoring and therapeutics. a comprehensive review. *EMBO molecular medicine*, 4(3): 143–159, 2012.

Jalali, A., Sanghavi, S., Ruan, C., and Ravikumar, P. K. A dirty model for multi-task learning. In *Advances in Neural Information Processing Systems 23*, pp. 964–972. 2010.

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

Maaten, L. v. d. and Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research: JMLR*, 9: 2579–2605, 2008.

Melnik, B. C. MiR-21: an environmental driver of malignant melanoma? *Journal of translational medicine*, 13: 202, 2015.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5(7):621–628, 2008.

Nalluri, J. J., Rana, P., Barh, D., Azevedo, V., Dinh, T. N., Vladimirov, V., and Ghosh, P. Determining causal miRNAs and their signaling cascade in diseases using an influence diffusion model. *Scientific reports*, 7(1):8133, 2017.

Peng, Y. and Croce, C. M. The role of MicroRNAs in human cancer. *Signal transduction and targeted therapy*, 1:15004, 2016.

Peters, J., Bühlmann, P., and Meinshausen, N. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 78(5):947–1012, 2016.

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

Riffo-Campos, Á. L., Riquelme, I., and Brebi-Mieville, P. Tools for Sequence-Based miRNA target prediction: What to choose? *International journal of molecular sciences*, 17(12), 2016.

Rupaimoole, R. and Slack, F. J. MicroRNA therapeutics: towards a new era for the management of cancer and other diseases. *Nature reviews. Drug discovery*, 16(3): 203–222, 2017.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. A Sparse-Group lasso. *Journal of computational and graphical statistics: a joint publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America*, 22(2):231–245, 2013.

Van den Berge, K., Perraudeau, F., Soneson, C., Love, M. I., Risso, D., Vert, J.-P., Robinson, M. D., Dudoit, S., and Clement, L. Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome biology*, 19(1):24, 2018.

Wang, M., Abrams, Z. B., Kornblau, S. M., and Coombes, K. R. Thresher: determining the number of clusters while removing outliers. *BMC bioinformatics*, 19(1):9, 2018.

Yuan, M. and Lin, Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 68 (1):49–67, 2006.

Zhang, J., Le, T. D., Liu, L., Liu, B., He, J., Goodall, G. J., and Li, J. Identifying direct miRNA-mRNA causal regulatory relationships in heterogeneous data. *Journal of biomedical informatics*, 52:438–447, 2014.

Zhang, Y. and Yang, Q. A survey on Multi-Task learning. 2017.