Published online DD MM YYYY

Nucleic Acids Research, 2023, Vol. xx, No. xx **1–9**
doi:10.1093/nar/gkn000

Transcriptome Complexity Disentangled: A Regulatory Molecules Approach

Amir Asiaee [1,*], Zachary B. Abrams [2] and Kevin R. Coombes [3]

[1]Department of Biostatistics, Vanderbilt University Medical Center, 2525 West End Avenue, Nashville, TN 37203, USA , [2]Institute for Informatics, Washington University, 4444 Forest Park Ave, St. Louis, MO 63108, USA , and [3]Department of Population Health Science, Medical College of Georgia, 1120 15th St, Augusta, GA 30912, USA

Received YYYY-MM-DD; Revised YYYY-MM-DD; Accepted YYYY-MM-DD

ABSTRACT

Gene regulatory networks play a critical role in understanding cell states, gene expression, and biological processes. Here, we investigated the utility of transcription factors (TFs) and microRNAs (miRNAs) in creating a low-dimensional representation of cell states and predicting gene expression across 31 cancer types. We identified 28 clusters of miRNAs and 28 clusters of TFs, demonstrating that they can differentiate tissue of origin. Using a simple SVM classifier, we achieved an average accuracy of 92.8% in tissue classification. We also predicted the entire transcriptome using Tissue-Agnostic and Tissue-Aware models, with average R^2 values of 0.45 and 0.70, respectively. Our Tissue-Aware model, using 56 selected features, showed comparable predictive power to the widely-used L1000 genes. However, the model's transportability was impacted by covariate shift, particularly inconsistent microRNA expression across datasets.

INTRODUCTION

The transcriptome is a critical element of gene expression regulation, comprising the entire set of messenger RNA molecules within an organism or cell. Due to its ease of measurement in comparison to other gene expression components, the transcriptome plays a significant role in advancing our understanding of differentiation (1), development (2), and disease (3). Essentially, it serves as a proxy signal that captures the cell's state, allowing for a mathematical and concrete representation of the underlying biology (4, 5).

Although RNA-Seq has become the most popular gene expression measurement technology due to the significant drop in its cost since 2008 (6), it may not always be the optimal solution. For instance, when measuring gene expression on a scale of millions of experiments to determine the cell's response to various environments (such as drugs or gene perturbation), the low cost per sample (ranging from 10 to 200, depending on the technology) accumulates to a considerable amount of money (7). Additionally, when high-quality RNA samples are unavailable, alternative sensitive techniques like qPCR may be preferable, although they are not ideal for whole transcriptome analysis due to the need for specific primers, limited throughput, and technical variability (8).

To address these challenges, the NIH LINCS project (7) developed a new, low-cost, high-throughput method for measuring the "reduced representation" of the gene expression profile consisting of 1000 landmark genes (L1000). This method enables the inference of non-measured transcripts with 81% accuracy. The authors employed this approach to scale the Connectivity Map project's samples from 492 original samples (9) to more than 1.3 million samples (7), enabling the discovery of functional connections among diseases, genetic perturbations, and drug actions. In the case of low-quality RNA samples, the same set of predictive models can be used for predicting transcriptome from the (potentially adjusted) qPCR measurement of the L1000 genes.

Notably, the L1000 genes are chosen based on a data-driven approach rather than prior biological knowledge, as highlighted by Subramanian et al. (7). In short, their approach is based on clustering of genes in the Principal Component space and selecting the centroids by an iterative peel-off procedure. Surprisingly, the authors found no enrichment of the landmarks in any specific functional class (such as transcription factors) using hypergeometric overlap statistics (10). The selection method and enrichment results are peculiar from a causal point of view because they ignore the fact that some parts of the transcriptome act as the causes of the others. For instance, certain transcripts produce transcription factor proteins that regulate the transcription of other genes (11). Therefore the abundance of former transcripts can be considered the causes of the latter ones. Taking into account the causal relationships between transcriptome components when making predictions can have two significant implications: firstly, by treating the causes as the main predictors for the entire transcriptome, the size of the "reduced representation" could be substantially reduced; secondly, incorporating causal predictors into the model could potentially result in more robust predictions across various environments with different perturbations. In this study, we

*To whom correspondence should be addressed. Tel: +1 615 3222067; Email: amir.asiaeetaheri@vumc.org

© 2023 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

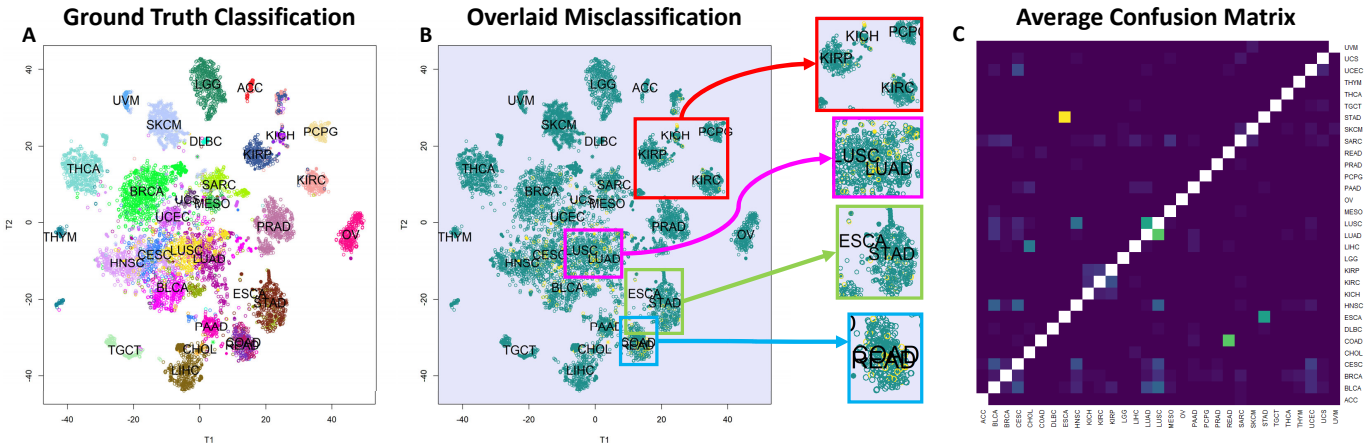

Figure 1. Cancer Classification and Visualization Through Regulatory Molecules. (A) A 2D t-SNE plot of the 8895x56 input matrix, consisting of selected TFs and miRNAs, with colors corresponding to known cancer types, showcasing the ability to differentiate most cancers using 56 regulatory molecules. **(B)** An enlarged t-SNE plot focusing on misclassified samples, marked in yellow, indicating that classification errors are more common in highly similar tissues. **(C)** An average confusion matrix generated from 10-fold cross-validation, exposing the error pattern of the SVM classifier with 92.8% accuracy. The diagonal elements (accurate classifications) have been removed to improve the heatmap's scaling for enhanced error visualization. Lighter matrix entries (excluding diagonals) represent larger errors, and the error pattern predominantly coincides with the t-SNE visualization of errors in panel B. Classifying similar cancer pairs like STAD and ESCA, COAD and READ, LUSC and LUAD, higher-order similar cancers such as KIRP, KIRC, and KICH, as well as squamous cancers (or cancers with a squamous cell subtype) including LUSC, CESC, HNSC, and BLCA, is challenging.

use Transcription Factor (TF) and microRNA (miR), two classes of regulatory molecules known to causally regulate transcription, as predictors to demonstrate that by measuring only 28 of each class, we can explain 70% variation of the transcriptome, i.e., expression of approximately 20,000 genes. Our findings suggest that the transcriptome has a low-dimensional structure that can be captured by measuring the expression of only 56 regulatory molecules. Each of our selected regulatory molecules represents a cluster of regulatory molecules, which we call biological components. Through gene enrichment analysis, we show that biological components consist of genes that are co-expressed and work together to control various biological processes (12, 13), which is in contrast to the selected L1000 landmark genes. Additionally, we show that our approach requires two orders of magnitude fewer measurements than the traditional L1000 approach while achieving comparable whole transcriptome prediction accuracy. This could lead to the development of cheaper transcriptome assays and accommodate low RNA sample regimes. Finally, we provide biological insights into potential reasons for certain genes being easily or hardly predictable using TFs and miRNAs by analyzing the set of genes that are poorly or well-explained by our model.

MATERIALS AND METHODS

Data Description

The data used in this study were collected from The Cancer Genome Atlas (TCGA) (14), a public data repository that contains clinical and omics data for over 10,000 patient samples from 32 different cancer types. To obtain the transcript data, we used the FireBrowse web portal and selected samples with matched mRNA and microRNA sequencing data. We obtained $n=8895$ samples from 31 different cancer types, excluding Glioblastoma due to its small sample size.

We downloaded the mRNA expression level data that were quantified using the RSEM method (15) and normalized using FPKM-UQ (Upper quartile normalized fragments per Kilobase of transcript per Million mapped reads). We obtained the microRNA expression data in the form of counts of mature microRNA, normalized to reads per million mapped reads (RPM). We log2 transformed both the mRNA and miRNA datasets. We filtered the data by removing miRNAs with zero read counts in 75% of the samples, resulting in $p=470$ miRNAs remaining (13). Using the Transcription Factor Catalog (16), we identified 486 human transcription factors and separated them from the 20289 measured gene expressions (12). The resulting dataset consisted of 8895 $\times$ (486 + 470) dimensions, comprising TFs and miRNAs as our original features and the rest of the transcriptome with 19803 gene expressions for the 8895 samples.

Feature Selection

In this study, we aim to understand the computational complexity of transcriptomes by selecting a low-dimensional representation that captures most of the information about the sample's state. Specifically, we focus on regulatory molecules miRNAs and TFs as features and aim to identify a small subset of them that can be used in downstream predictive analyses while maintaining comparable performance to using the whole transcriptome. To test our proposed method, we analyzed $\sim$9000 samples from 31 cancer types studied in The Cancer Genome Atlas (TCGA).

We choose feature selection instead of dimensionality reduction methods like Principal Component Analysis (PCA) because measuring the corresponding low dimensional latent features often requires measuring the original signal, which can be expensive and challenging for low-quality RNA

samples. Additionally, interpreting the biological relevance of latent features that are superpositions of numerous genes can be challenging (17). We chose not to use prediction-guided feature selection methods (18), such as LASSO (19), as our goal was to choose features in a task-agnostic manner, ensuring that the performance of any downstream prediction task remains comparable to the situation where the entire transcriptome was measured.

Our feature selection process employs the Thresher package (20), which automatically selects the optimal number of clusters for clustering TFs and miRNAs into biological components. We then select the medoid of each cluster as the representative gene for the corresponding biological component. The Thresher pipeline begins by utilizing PCA to derive a low-dimensional representation of each feature. To identify the number of significant principal components, we use a variant of the Bayesian model selection method (21). We further refine the feature set by removing outliers using a length threshold for each feature representation. A threshold of 0.35 was chosen based on several simulation studies (20). Finally, we cluster the directions of feature representation vectors corresponding to points on the hypersphere by utilizing mixtures of von Mises-Fisher distributions (22). We select the optimal number of regulatory molecule clusters using the Bayesian Information Criterion (23). To demonstrate that the selected regulatory molecules capture most of the transcriptome's information, we perform various predictive tasks summarized in the subsequent sections.

Assessing Predictive Power of Selected Regulatory Molecules

Tissue Classification Using Selected Regulatory Molecules. To show that the selected 56 regulatory molecules are informative, we use them to predict the tissue of origin for TCGA samples. This is a multi-class classification problem (24) with 31 outcome labels. To train our classifier, we use a Support Vector Machine (SVM) algorithm (25) and report the average generalization error on hold-out data through 10-fold cross-validation.

Transcriptome Prediction Using Selected Features and Tissue Type Information. Our goal is to demonstrate that the 56 selected features provide an accurate representation of the entire transcriptome. In the first experiment, we use linear regression to predict the entire transcriptome using the selected features. We have approximately 20,000 regression tasks, with each task consisting of 56 features and $\sim$9,000 samples. For each outcome gene, we obtain a model with 57 parameters, consisting of 56 coefficients and one intercept.

In the second experiment, we use tissue type information available in the TCGA dataset to improve our model. Instead of predicting the outcome for each gene as a single task, we predict the outcome in each tissue as a separate task, resulting in 31 tasks for TCGA. We first fit a global regression model across all tissues and then fit the residuals of the global model in each tissue through a secondary regression. This can be considered a form of multi-task learning (26), where a shared intercept is learned for all tasks along with one linear model per task (13, 27, 28). Our first experiment, which does not incorporate tissue information, is called Tissue-Agnostic, while our second multi-task regression experiment, which uses tissue type as a task, is termed Tissue-Aware.

Biological Insights from Evaluating Transcriptome Prediction Accuracy

To gain insights into the biological implications of transcriptome prediction accuracy, we evaluated the Tissue-Agnostic and Tissue-Aware models and examined how much variance in the transcriptome they accounted for. Our analysis focused on the distribution of R^2 values, which represents the proportion of variation in gene expression that each model explains. We created a histogram of R^2 values for the nearly 20,000 genes predicted and fitted a beta mixture model to identify poorly, moderately, and well-explained genes from the histogram. We applied the expectation-maximization algorithm with a range of mixture components ($k=1,2,3,4$) and used the Bayesian Information Criterion (BIC) score to determine the optimal number of components. The R^2 distribution of both models showed evidence of two beta distribution components, which we labeled as poorly and well-explained components.

To gain insight into which biological pathways are influenced by TFs and miRs, and which pathways are less affected by these regulatory molecules, we performed gene enrichment analysis using the ToppGene platform (29). Specifically, we focused on genes in the bottom 5th percentile of the poorly explained component and in the top 99th percentile of the well-explained component for each model. Our goal was to identify pathways that are heavily controlled by these regulatory molecules, regardless of tissue involvement in the modeling. To accomplish this, we examined the intersection of pathways enriched for poorly-explained genes in each model, as well as for well-explained genes.

Comparison of Performance between Selected Regulatory Molecules and L1000 Genes

The L1000 Gene Expression Signature Profiling Technology and the performance of the 978 landmark genes it uses have been widely studied for identifying new therapeutic targets and drug combinations (7). However, due to the lack of access to the data set used for their selection and prediction, we were unable to test the performance of our 56 selected regulatory molecules on the same data set. To ensure a fair comparison, we instead used the L1000 genes to fit TCGA data from scratch.

We conducted a comparison between our Tissue-Aware model, which employs 56 genes, and the linear model, which uses L1000 genes. To evaluate the performance of the models, we computed their average R^2 scores and correlation scores. The correlation score, introduced in the L1000 landmark paper (7), measures the overall performance of the algorithm in predicting gene expression. For each gene, the predicted expression and ground truth expression are compared using the correlation score R_g. To determine if R_g is significant, a null distribution of correlations between all inferred transcripts and all measured transcripts is used. The inference is deemed accurate if R_g is higher than the 95th percentile of the null distribution. The correlation score of the algorithm represents the overall percentage of genes that are accurately inferred by

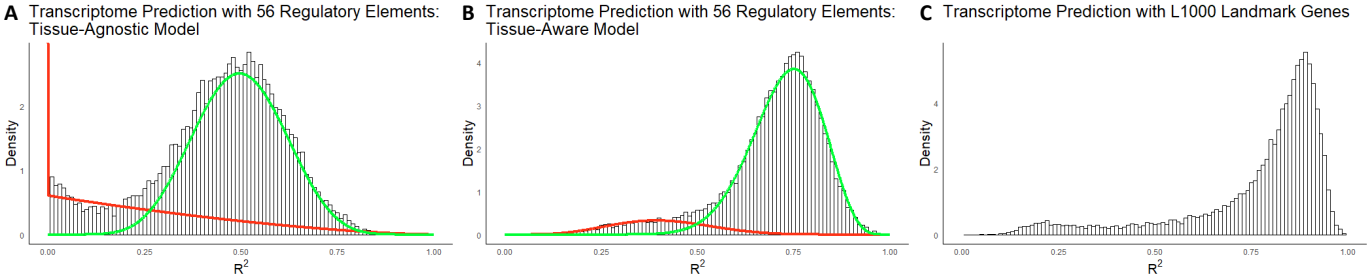

Figure 2. Gene Expression Prediction and Comparison of Predictive Power between Selected Regulatory Molecules and L1000 Genes. (A) and **(B)** R^2 histograms for predicting 20,289 gene expressions using 8,895 samples in TCGA with Tissue-Agnostic and Tissue-Aware models, respectively, illustrating the mixture of two components: well-explained genes (green curve) and poorly-explained genes (red curve). The Tissue-Aware model demonstrates enhanced performance, with an average R^2 value of 0.70 compared to 0.45 for the Tissue-Agnostic model. **(C)** R^2 distribution of the linear models fitted using 1,000 landmark genes(7), displaying an average R^2 value of 0.77. The distribution exhibits a more pronounced bimodality, indicating that compared to the Tissue-Aware model, some genes are better explained, while the poorly-explained tail is also thicker. The Tissue-Aware model achieves comparable predictive power using only 56 predictors according to the R^2 measure and attains a correlation score of 0.9998 compared to 0.9975 for the L1000 genes (not shown), highlighting the effectiveness of the selected 56 regulatory molecules in predicting the transcriptome.

the method. The linear model trained on the data from (7) that uses L1000 landmark genes as features has a reported correlation score of 83%.

Evaluating Cross-Dataset Transportability of Tissue-Aware Transcriptome Prediction Model

To evaluate the transportability capabilities of our Tissue-Aware transcriptome prediction model, we tested it on unseen data from the Genotype-Tissue Expression (GTEx) project (30). To ensure consistency with TCGA normalization, we downloaded gene count data from the GTEx portal and normalized the mRNA and miRNA expressions separately. We then used selected TF and miRNA as predictors for the remaining mRNA expression. However, four miRNA cluster medoids were unavailable in GTEx, so we selected the second closest member to the centroid as the cluster representative. We continued this iterative process until representatives were found for each biological component. Finally, employing the Tissue-Aware model, we used the newly selected regulatory molecules to predict the transcription of 16607 mRNA (that we could match with TCGA) in 17382 samples. However, to use the Tissue-Aware model, a mapping between tissue labels in TCGA and GTEx was required.

To avoid issues arising from treating each tissue as a separate task, such as misaligned tissue labels across databases and suboptimal tissue label resolution, we opted to cluster TCGA data and determine the optimal number of tasks in a data-driven manner. Specifically, we clustered TCGA samples into Pseudo-Tissues using K-means clustering (23) with a range of values $K = 1, 5, 10, 15, 20, 25, 30, 35$, and performed 5-fold cross-validation to monitor the (Pseudo)Tissue-Aware model's test performance in the whole transcriptome prediction task for TCGA. We selected the number of clusters (tasks) that resulted in the minimum test error, and for each cluster of samples, we used the centroid to represent the Pseudo-Tissue. To evaluate the model's performance over GTEx, we assigned each GTEx sample to the Pseudo-Tissue whose representative had the smallest Euclidean distance to the sample. We then used the trained model for that Pseudo-Tissue to predict the transcriptome.

RESULTS

Biological Relevance of Chosen Features

Our methodology identifies 28 miRNA clusters and 28 TF clusters, using their medoids as the selected features for our analysis. Prior studies have shown that the TFs (12) and miRNAs (13) clusters recognized by Thresher are associated with specific biological processes and exhibit only one principal direction of variation. Consequently, these clusters can be viewed as biological components representing cell states, as opposed to principal components. We employed the unsupervised, nonlinear t-SNE projection technique (31) to the 8895×56 input matrix consisting of selected TFs and miRNAs, and visualized the outcome in a two-dimensional plot shown in Figure 1A. To enhance our comprehension of the patterns, we color-coded the plot based on known cancer types. Overall, the majority of cancer types can be differentiated using the 56 chosen regulatory molecules.

High Accuracy Tissue Classification with 56 Regulatory Molecules

Our analysis uses 56 regulatory molecules to perform multi-class tissue classification using a vanilla SVM classifier. Figure 1C illustrates the average confusion matrix obtained from 10-fold cross-validation, where the average accuracy of the classifier is 92.8%. However, the accuracy over tissue pairs is bimodal, indicating that the model is able to distinguish most pairs with 100% accuracy but struggles to classify certain pairs. Notably, challenging tissue pairs include COAD (Colon Adenocarcinoma) and READ (Rectal Adenocarcinoma), LUSC (Lung Squamous Cell Carcinoma) and LUAD (Lung Adenocarcinoma), and ESCA (Esophageal Carcinoma) and STAD (Stomach Adenocarcinoma). Additionally, higher-order classification confusion is observed for KIRP (Kidney Renal Papillary Cell Carcinoma), KIRC (Kidney Renal Clear Cell Carcinoma), and KICH (Kidney Chromophobe), which are all subtypes of renal cell carcinoma as well as squamous cancers including LUSC, CESC (Cervical Squamous Cell Carcinoma), HNSC (Head and Neck Squamous Cell Carcinomas), and BLCA (Bladder Carcinoma, with a basal-squamous subtype). The t-SNE plot of the clusters is shown in

Figure 1B, with misclassified samples highlighted in yellow. The plot is zoomed in to emphasize that errors are more frequent in tissues that are highly similar.

Highly Accurate Prediction of Gene Expression using Selected 56 Features

Our analysis employed the selected 56 features to predict the expression of approximately 20,000 genes using Tissue-Agnostic and Tissue-Aware models. Figure 2A and B illustrate the R^2 distribution histogram for both models, revealing that both histograms are mixtures of two components. The tail analysis procedure used Beta mixture distributions fitted to the histograms, with the green curve representing the component covering the right tail (i.e., well-explained genes) and the red curve representing the left tail (i.e., poorly explained genes). Notably, the distribution of poorly-explained genes is thicker and skewed toward zero in the Tissue-Agnostic model in comparison to the Tissue-Aware model. Additionally, the variance explained for well-explained genes is significantly improved in the Tissue-Aware model. The average R^2 values for the Tissue-Agnostic and Tissue-Aware models are 0.45 and 0.70, respectively.

Identifying Pathways Regulated by TFs, miRs, and Alternative Mechanisms

We obtained four sets of genes by using the 5th and 99th percentile cutoffs for poorly and well-explained genes in the tails of the R^2 distributions of the Tissue-Agnostic and Tissue-Aware models. In the Tissue-Agnostic model, we identified 256 poorly explained genes, while in the Tissue-Aware model, there were 106 poorly explained genes. Following gene set pathway enrichment, we found a high degree of overlap between enriched pathways that were considered significant. For both models, the number of significant pathways was limited to ten, with nine of them shared between the two models, as shown in the left column of Table 1. In the Tissue-Agnostic model, we identified 280 well-explained genes, while in the Tissue-Aware model, we identified 228 well-explained genes. After performing pathway enrichment analysis, we found 123 significantly enriched pathways in the Tissue-Agnostic model and 90 significant pathways in the Tissue-Aware model. Of these pathways, only 12 were significant in both models and are listed in the right column of Table 1.

Comparison of Predictive Power of Selected Regulatory Molecules and L1000 Genes for Transcriptome

In this experiment, we compare the predictive power of our selected 56 regulatory molecules to that of the L1000 genes. Figure 2C shows the R^2 distribution of the linear model fitted using L1000 genes, with an average R^2 value of 0.77 over the entire transcriptome. In contrast, our Tissue-Aware model, using only 56 predictors, has an average R^2 value of 0.70, indicating comparable predictive power. The R^2 distribution of the L1000 genes also exhibits a similar bimodal shape as observed in the Tissue-Aware model. Furthermore, while the L1000 genes reach a correlation score of 0.9975, our Tissue-Aware model achieves a score of 0.9998. These results demonstrate that our selected 56 regulatory molecules are almost as good predictors of the transcriptome as the widely-used L1000 genes and, in some cases, even outperform them.

Shared Pathways Enriched for	
Poorly-Explained Genes	**Well-Explained Genes**
Antimicrobial Peptides	Cell Adhesion Molecules (CAMs)
Beta Defensins	Hematopoietic Cell Lineage
Defensins	Tight Junction
Developmental Biology	Catecholamine Biosynthesis, Tyrosine =>Dopamine =>Noradrenaline => Adrenaline
GPCR Downstream Signaling	Hemostasis
Keratinization	MET Receptor Activation
Olfactory Signaling Pathway	Network map of SARS-CoV-2 Signaling J—pathway
Olfactory Transduction	Oligodendrocyte Specification and Differentiation, Leading to Myelin components for CNS Signaling by MSTL
Signaling by GPCR	Signaling by MSTL
	T Cytotoxic Cell Surface Molecules
	THelper Cell Surface Molecules
	Tight Junction Interactions

Table 1. Pathway Enrichment Analysis for Poorly and Well-Explained Genes. The table presents the results of gene set pathway enrichment analysis for genes poorly explained (left column) and well-explained (right column) common between Tissue-Agnostic and Tissue-Aware models. Out of 10 pathways enriched for genes poorly explained by both models, nine are shared, demonstrating that these pathways are not heavily regulated by TFs and miRNAs regardless of tissue type. Conversely, 123 and 90 pathways were enriched for well-explained genes in Tissue-Agnostic and Aware models, respectively, among which only the listed 12 pathways (right column) were shared between the models. This indicates that tissue-specific regulatory mechanisms play a crucial role in governing the expression of well-explained genes. Poorly explained genes are potentially regulated by alternative mechanisms such as signaling pathways, post-translational modifications, protein-protein interactions, and epigenetic regulation, while well-explained genes are likely heavily regulated by TFs and miRNAs, participating in various cellular functions.

Covariate Shift Affects Transportability of Tissue-Aware Model

To predict gene expression across different datasets, first, we investigate the generalization error (32) of our Tissue-Aware model as a function of the number of Pseudo-Tissues. Figures 3A and B depict the cross-validation curves to tune the number of Pseudo-Tissues (clusters) for RMSE and correlation score, respectively, using the TCGA dataset. The optimal number of Pseudo-Tissues is found to be 15, based on both criteria. However, when testing the transportability (out-of-distribution generalization (33)) of the model on the GTEx dataset, the performance is poorer than expected due to the covariate shift (34), where the distribution of predictors across datasets differs. In particular, many microRNAs have zero expression in GTEx, making prediction challenging, Figure 4C and D. In light of this observation, we re-did the experiment using only TFs, as they showed more consistent expression distributions across datasets, Figure 4A and B. Training a Tissue-Aware model with 28 TFs on TCGA data and testing on GTEx, we achieved improved results with an average RMSE of 1.3 and

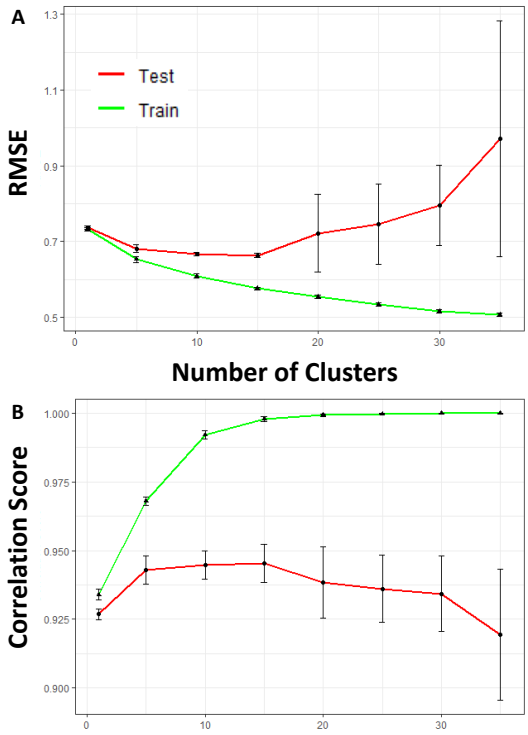

Figure 3. Optimizing Pseudo-Tissues for Tissue-Aware Model Generalization. (A) and **(B)** Cross-validation curves for tuning the number of Pseudo-Tissues (sample clusters) based on RMSE and correlation score, respectively, using the TCGA dataset. The optimal number of Pseudo-Tissues is determined to be 15 for both criteria. However, cross-validation assumes that the train and test data share the same distributions, which is not the case when distribution shifts are present, such as when selected miRNA expressions are almost zero in GTEx, thus hindering the transportability of the predictive model.

a correlation score of 0.55 during testing compared to when miRNAs were included in the model.

DISCUSSION

The Role of Causal Predictors in Robust Transcriptome Representation

Understanding gene regulatory networks, capturing cell states through gene expression profiles, and obtaining low-dimensional representations of cell states are three essential and interconnected biological questions. In this study, we demonstrated that by using our prior knowledge of the critical roles of transcription factors and microRNAs in gene regulation, we can establish a low-dimensional representation of cell states and infer the entire transcriptome from a limited number of regulatory molecules. The value of a reduced cell state representation lies in its ability to capture the gene expression distribution not only under normal conditions but also under various perturbations, such as drugs, mutations, or gene knockouts. These perturbations fundamentally change the joint distributions of gene expressions, generating interventional distributions (35) over expressions that can differ significantly from the normal state. To obtain the minimum number of genes needed to explain most transcriptome variations, it is necessary to understand or infer the gene regulatory network and select master regulators to fully predict the outcomes of perturbations. We hypothesized that using TFs and miRNAs as a shortcut can work in place of full gene regulation inference because of their significant role in the regulation of various biological processes. Consequently, under any perturbation or environment, their values serve as robust predictors of the entire transcriptome. Our hypothesis is supported by two key findings. Firstly, our feature selection pipeline identified regulatory molecule clusters that have biological significance, as reported in (12, 13). Secondly, the selected features demonstrate strong predictive capability in supervised learning tasks across different environments, such as different tissues, and under different perturbations, such as DNA aberrations in cancer cells. Together, these results provide further evidence in support of our hypothesis.

Challenges in Cancer Tissue Classification

Our findings demonstrate that these selected features can effectively distinguish between 31 cancer types, as illustrated by the t-SNE plot (Figure 1A). Moreover, the classification accuracy obtained by our vanilla SVM classifier is promising at 92.8%, although the model faces some challenges with specific tissue pairs and subtypes. Biologically, the misclassification of these tissues can be rationalized. COAD and READ are both very similar subtypes of colorectal cancer (36), while LUSC and LUAD are subtypes of lung cancer, and ESCA and STAD are both types of gastrointestinal cancers. A similar situation arises for KIRP, KIRC, and KICH, where all are subtypes of renal cell carcinoma that originate from different cell types in close proximity to the kidney.

Enhanced Gene Expression Prediction Using Selected Features and Tissue-Aware Models

The prediction of gene expression using the selected 56 features was highly accurate, with an average R^2 value of 0.45 for the Tissue-Agnostic model and 0.70 for the Tissue-Aware model. The Tissue-Aware model showed significant improvement over the Tissue-Agnostic model, suggesting that incorporating tissue-specific information into our multi-task learning framework can enhance gene expression prediction. There are two potential biological explanations for the effectiveness of the Tissue-Aware model. First, several other types of regulatory molecules, such as epigenetic modifiers (37) and chromatin remodeling complexes (38), play essential roles in gene regulation besides TFs and miRs. Some of these, like DNA methylation and histone modification, are crucial in determining cell identity and ultimately tissue type (39, 40). Therefore, using tissue type can serve as a surrogate for these omitted regulators. Furthermore, from a causal perspective, the interaction between regulatory molecules can cause the causal effect of TFs and miRNAs to differ in each environment. Consequently, our multi-task learning framework, which fits a model per environment with shared parameters across environments, is a more biologically plausible model.

We hypothesized that the causal relevance of TFs and miRs, along with our flexible multi-task learning framework, would enable a succinct representation of the transcriptome. We

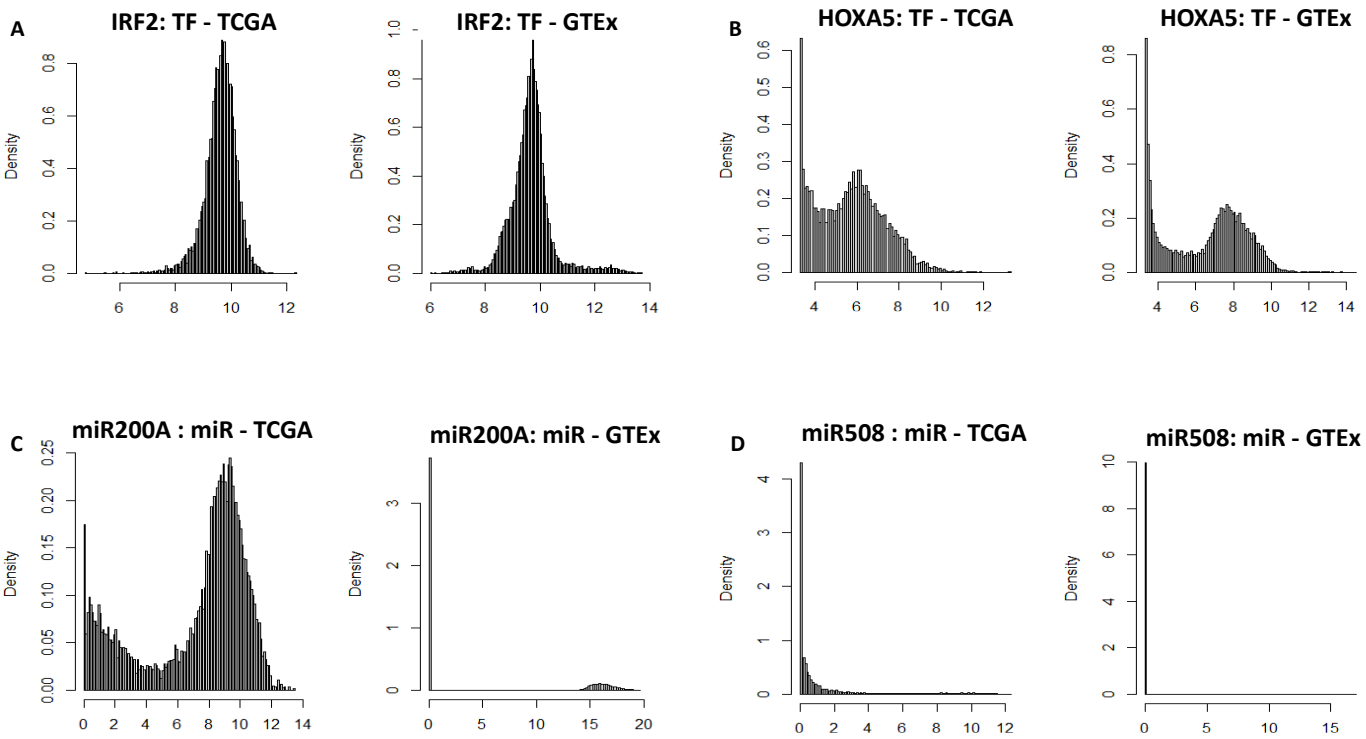

Figure 4. Shift in Distributions of Regulatory Molecules across TCGA and GTEx Affecting Model Transportability. (A) and **(B)** Display distributions of two selected transcription factors (IRF2 and HOXA5) across both datasets. Although differences exist between TF distributions, they generally share the same support and shape, allowing predictive model adjustments using techniques like domain adaptation. **(C)** and **(D)** Contrastingly, illustrate the vastly different distributions of two selected miRNAs (miR200A and miR508) in TCGA and GTEx. Notably, many selected miRNAs exhibit zero expression in almost all GTEx samples, making model transportation challenging. Differences in microRNA expression between datasets may stem from factors such as experimental techniques, data preprocessing, biological variability, and post-mortem interval between tissue collection and RNA extraction for GTEx samples.

confirmed this hypothesis by demonstrating that our selected 56 features are as predictive as the widely-used L1000 genes. The predictive power of our 56 regulatory molecules was nearly as good as the L1000 genes, according to both R^2 and correlation score measures. In some cases, our Tissue-Aware model even outperformed the L1000 genes, highlighting the effectiveness of our feature selection method, which is guided by biological knowledge about the causal role of regulatory molecules in transcription.

Regulatory Mechanisms in Gene Expression

Alternative Regulation in Poorly Explained Genes. Genes poorly explained by both Tissue-Agnostic and Aware models may not be heavily regulated by TFs and miRs. Gene enrichment analysis (left column of Table 1) indicates that these genes might be regulated by other mechanisms, such as signaling pathways (41), post-translational modifications (42), protein-protein interactions (43), and epigenetic regulation (39). These alternative mechanisms allow rapid and dynamic control of cellular responses, immune defense, and maintenance of cellular integrity. Examples of such pathways include antimicrobial peptides, keratinization, olfactory signaling, and GPCR downstream signaling.

TF and miRNA Regulation in Well Explained Genes. Conversely, genes well explained by both models are likely heavily regulated by TFs and miRNAs. Gene enrichment analysis (right column of Table 1) supports this hypothesis, showing that these genes are involved in various cellular functions, including cell adhesion, differentiation, neurotransmission, immune response, and barrier integrity. Tight regulation by miRNAs and TFs allows cells to respond appropriately to stimuli (44) and maintain tissue homeostasis (45, 46), development (47), and function. Examples of such pathways include cell adhesion molecules (CAMs), hematopoietic cell lineage, tight junctions, catecholamine biosynthesis, and hemostasis.

Impact of Covariate Shift on Transportability

The Tissue-Aware model's transportability was evaluated by testing its performance on the GTEx dataset. Although the model performed well on the TCGA dataset, the covariate shift, particularly the inconsistent microRNA expression across datasets, hindered its performance on the GTEx dataset. Figure 4 illustrates the distributions of some selected TFs and miRNAs in TCGA and GTEx. While differences exist between TF distributions, they generally share the same support and shape, whereas some selected miRNAs are entirely absent in GTEx. By re-evaluating the model using only TFs, we observed improved results, suggesting that including miRNAs with low expression in most samples can negatively impact prediction performance. This finding underscores the importance of considering the robustness of selected features across datasets to enhance the

model's transportability. Potential reasons for the significant differences in microRNA expression between the two datasets include differences in experimental techniques, data preprocessing, biological variability, and post-mortem interval between tissue collection and RNA extraction for GTEx samples.

CONCLUSION

In this study, we demonstrated that using a small subset of transcription factors and microRNAs as selected features can effectively represent cell states and predict gene expression across 31 cancer types. Our results support the hypothesis that these regulatory molecules, due to their causal roles in gene regulation, can provide robust predictors of the transcriptome. The Tissue-Aware model, which incorporates tissue-specific information, outperformed the Tissue-Agnostic model, suggesting the importance of context for gene regulation. While our selected features were nearly as predictive as the L1000 genes, the model's transportability was affected by the covariate shift, particularly in microRNA expression. By using only TFs in the model, we improved our performance on the GTEx dataset, emphasizing the need for robust feature selection across different datasets. Future studies should explore additional regulatory molecules, such as epigenetic modifiers and chromatin remodeling complexes, as well as their interactions, to enhance the understanding of gene regulation and improve model transportability.

DATA AVAILABILITY

All the code and data used for analysis can be found on GitHub at -`https://github.com/aalabmaster/transcriptome_disentangled`. The paper's data is also publicly available and accessible via our GitHub repository.

ACKNOWLEDGEMENTS

Author contributions: Conceptualization, K.R.C.; methodology and investigation, A.A., K.R.C, Z.B.A.; writing, A.A.

ACKNOWLEDGEMENTS

NIH [R00HG011367 to A.A.].

Conflict of interest statement. None declared.

REFERENCES

1. Mark Richards, Siew-Peng Tan, Jee-Hian Tan, Woon-Khiong Chan, and Ariff Bongso. The transcriptome profile of human embryonic stem cells as defined by sage. *Stem cells*, 22(1):51–64, 2004.
2. Margarida Cardoso-Moreira, Jean Halbert, Delphine Valloton, Britta Velten, Chunyan Chen, Yi Shao, Angélica Liechti, Kelly Ascenção, Coralie Rummel, Svetlana Ovchinnikova, et al. Gene expression across mammalian organ development. *Nature*, 571(7766):505–509, 2019.
3. Glenn A Maston, Sara K Evans, and Michael R Green. Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.*, 7:29–59, 2006.
4. Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63, 2009.
5. Marta Melé, Pedro G Ferreira, Ferran Reverter, David S DeLuca, Jean Monlong, Michael Sammeth, Taylor R Young, Jakob M Goldmann, Dmitri D Pervouchine, Timothy J Sullivan, et al. The human transcriptome across tissues and individuals. *Science*, 348(6235):660–665, 2015.
6. John C Marioni, Christopher E Mason, Shrikant M Mane, Matthew Stephens, and Yoav Gilad. Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, 18(9):1509–1517, 2008.
7. Aravind Subramanian, Rajiv Narayan, Steven M Corsello, David D Peck, Ted E Natoli, Xiaodong Lu, Joshua Gould, John F Davis, Andrew A Tubelli, Jacob K Asiedu, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452, 2017.
8. Stephan Pabinger, Stefan Rödiger, Albert Kriegner, Klemens Vierlinger, and Andreas Weinhäusel. A survey of tools for the analysis of quantitative pcr (qpcr) data. *Biomolecular Detection and Quantification*, 1(1):23–33, 2014.
9. Justin Lamb, Emily D Crawford, David Peck, Joshua W Modell, Irene C Blat, Matthew J Wrobel, Jim Lerner, Jean-Philippe Brunet, Aravind Subramanian, Kenneth N Ross, et al. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *science*, 313(5795):1929–1935, 2006.
10. Wen Fury, Franak Batliwalla, Peter K Gregersen, and Wentian Li. Overlapping probabilities of top ranking gene lists, hypergeometric distribution, and stringency of gene selection criterion. In *2006 international conference of the IEEE engineering in medicine and biology society*, pages 5531–5534. IEEE, 2006.
11. Samuel A Lambert, Arttu Jolma, Laura F Campitelli, Pratyush K Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R Hughes, and Matthew T Weirauch. The human transcription factors. *Cell*, 172(4):650–665, 2018.
12. Zachary B Abrams, Mark Zucker, Min Wang, Amir Asiaee Taheri, Lynne V Abruzzo, and Kevin R Coombes. Thirty biologically interpretable clusters of transcription factors distinguish cancer type. *BMC genomics*, 19:1–14, 2018.
13. Amir Asiaee, Zachary B Abrams, Samantha Nakayiza, Deepa Sampath, and Kevin R Coombes. Explaining gene expression using twenty-one micrornas. *Journal of Computational Biology*, 27(7):1157–1170, 2020.
14. Cancer Genome Atlas Research Network, John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas Pan-Cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.
15. Bo Li and Colin N Dewey. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, 12:1–16, 2011.
16. Debra L Fulton, Saravanan Sundararajan, Gwenael Badis, Timothy R Hughes, Wyeth W Wasserman, Jared C Roach, and Rob Sladek. Tfcat: the curated catalog of mouse and human transcription factors. *Genome biology*, 10:1–14, 2009.
17. Michael Lenz, Franz-Josef Müller, Martin Zenke, and Andreas Schuppert. Principal components analysis and the reported low intrinsic dimensionality of gene expression microarray data. *Scientific reports*, 6(1):1–11, 2016.
18. Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.
19. Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.
20. Min Wang, Zachary B Abrams, Steven M Kornblau, and Kevin R Coombes. Thresher: determining the number of clusters while removing outliers. *BMC bioinformatics*, 19:1–15, 2018.
21. Philipp Auer and Daniel Gervini. Choosing principal components: a new graphical method based on bayesian model selection. *Communications in Statistics—Simulation and Computation®*, 37(5):962–977, 2008.
22. Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, Suvrit Sra, and Greg Ridgeway. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(9), 2005.
23. Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

24. Margherita Grandini, Enrico Bagli, and Giorgio Visani. Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*, 2020.
25. William S Noble. What is a support vector machine? *Nature biotechnology*, 24(12):1565–1567, 2006.
26. Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, 2021.
27. Amir Asiaee, Samet Oymak, Kevin R Coombes, and Arindam Banerjee. High dimensional data enrichment: Interpretable, fast, and data-efficient. *arXiv preprint arXiv:1806.04047*, 2018.
28. Amir Asiaee, Samet Oymak, Kevin R Coombes, and Arindam Banerjee. Data enrichment: Multi-task learning in high dimension with theoretical guarantees. In *Adaptive and Multitask Learning Workshop at the ICML. IMLS, Long Beach, CA*, 2019.
29. Jing Chen, Eric E Bardes, Bruce J Aronow, and Anil G Jegga. Toppgene suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic acids research*, 37(suppl_2):W305–W311, 2009.
30. GTEx Consortium, Kristin G Ardlie, David S Deluca, Ayellet V Segrè, Timothy J Sullivan, Taylor R Young, Ellen T Gelfand, Casandra A Trowbridge, Julian B Maller, Taru Tukiainen, et al. The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.
31. Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.
32. Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.
33. Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.
34. Pirmin Lemberger and Ivan Panico. A primer on domain adaptation. *arXiv preprint arXiv:2001.09994*, 2020.
35. Judea Pearl. *Causality*. Cambridge university press, 2009.
36. Cancer Genome Atlas Network et al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330, 2012.
37. Yuan Cheng, Cai He, Manni Wang, Xuelei Ma, Fei Mo, Shengyong Yang, Junhong Han, and Xiawei Wei. Targeting epigenetic regulators for cancer therapy: mechanisms and advances in clinical trials. *Signal transduction and targeted therapy*, 4(1):62, 2019.
38. Chun Zhang, Jie Lu, and Pingyu Zhang. The roles of chromatin remodeling proteins in cancer. *Current Protein and Peptide Science*, 17(5):446–454, 2016.
39. Bradley E Bernstein, Alexander Meissner, and Eric S Lander. The mammalian epigenome. *Cell*, 128(4):669–681, 2007.
40. Howard Cedar and Yehudit Bergman. Linking dna methylation and histone modification: patterns and paradigms. *Nature Reviews Genetics*, 10(5):295–304, 2009.
41. Jignesh R Parikh, Bertram Klinger, Yu Xia, Jarrod A Marto, and Nils Blï¿ 1/2thgen. Discovering causal signaling pathways through gene-expression patterns. *Nucleic acids research*, 38(suppl_2):W109–W117, 2010.
42. Boxuan Simen Zhao, Ian A Roundtree, and Chuan He. Post-transcriptional gene regulation by mrna modifications. *Nature reviews Molecular cell biology*, 18(1):31–42, 2017.
43. Safir Khan et al. Review on gene regulation: Dna-protein and protein-protein interactions and their regulatory elements. *Journal of Chemistry and Nutritional Biochemistry*, 2(2):35–45, 2021.
44. Masafumi Inui, Graziano Martello, and Stefano Piccolo. Microrna control of signal transduction. *Nature reviews Molecular cell biology*, 11(4):252–263, 2010.
45. Astrid Eijkelenboom and Boudewijn MT Burgering. Foxos: signalling integrators for homeostasis maintenance. *Nature reviews Molecular cell biology*, 14(2):83–97, 2013.
46. Karthikeyan Chandrasekaran, Dwi S Karolina, Sugunavathi Sepramaniam, Arunmozhiarasi Armugam, E Marelyn Wintour, John F Bertram, and Kandiah Jeyaseelan. Role of micrornas in kidney homeostasis and disease. *Kidney international*, 81(7):617–627, 2012.
47. François Spitz and Eileen EM Furlong. Transcription factors: from enhancer binding to developmental control. *Nature reviews genetics*, 13(9):613–626, 2012.