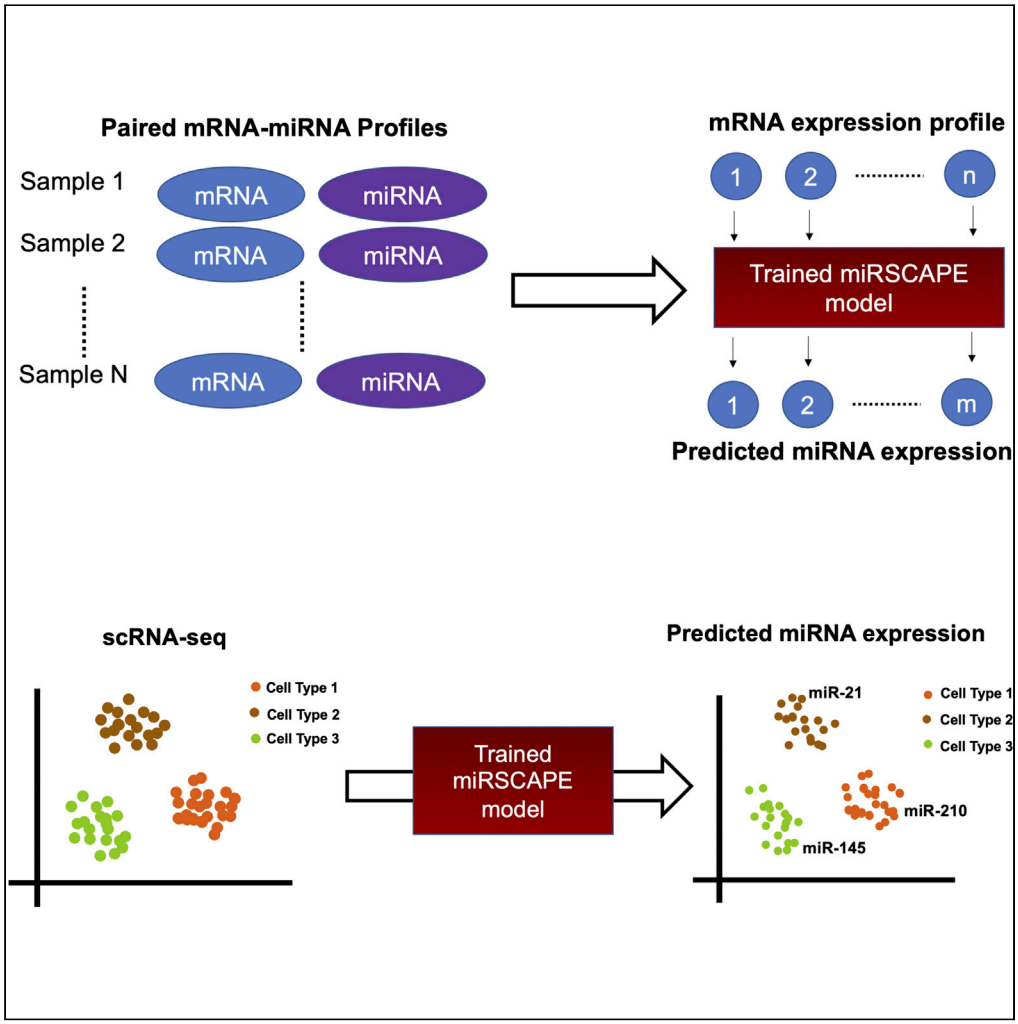
iScience

CellPress
OPEN ACCESS

Article

miRSCAPE - inferring miRNA expression from scRNA-seq data



Gulden Olgun, Vishaka Gopalan, Sridhar Hannenhalli

sridhar.hannenhalli@nih.gov

Highlights

Novel machine learning-based tool to infer miRNA expression at single cell level

Predicts miRNA activity with high accuracy in various contexts

Recaps miRNAs associated with specific cellular states and suggests novel candidates

Provides a general framework to predict other types of molecular data at a single cell

Olgun et al., iScience 25, 104962
September 16, 2022
https://doi.org/10.1016/j.isci.2022.104962

iScience

CellPress
OPEN ACCESS

Article

miRSCAPE - inferring miRNA expression from scRNA-seq data

Gulden Olgun,[1] Vishaka Gopalan,[1] and Sridhar Hannenhalli[1,2,*]

SUMMARY

Our understanding of miRNA activity at cellular resolution is thwarted by the inability of standard scRNA-seq protocols to capture miRNAs. We introduce a novel tool, miRSCAPE, to infer miRNA expression in a sample from its RNA-seq profile. We establish miRSCAPE's accuracy in 10 tumor and normal cohorts demonstrating its superiority over alternatives. miRSCAPE accurately infers cell type-specific miRNA activities (predicted versus observed fold-difference correlation ~0.81) in two independent scRNA-seq datasets. We apply miRSCAPE to infer miRNA activities in scRNA clusters in pancreatic and lung adenocarcinomas, as well as in 56 cell types in the human cell landscape (HCL). In pancreatic and breast cancer scRNA-seq data, miRSCAPE recapitulates miRNAs associated with stemness and epithelial-mesenchymal transition (EMT) cell states, respectively. Overall, miRSCAPE recapitulates and refines miRNA biology at cellular resolution. miRSCAPE is freely available and is easily applicable to scRNA-seq data to infer miRNA activities at cellular resolution.

INTRODUCTION

miRNAs are small non-coding RNAs that typically bind to 3′ untranslated regions (UTR) of their target mRNAs and regulate their expression via diverse mechanisms, including mRNA degradation (Valencia-Sanchez, 2006) and translational inhibition (Bartel, 2004). miRNAs play critical roles in most fundamental cellular processes, from development to homeostasis (Bartel, 2004; Jovanovic and Hengartner, 2006) and consequently are implicated in several diseases, including cancer (Peng and Croce, 2016).

Single-cell RNA sequencing (scRNA-seq) technologies have advanced our understanding of molecular mechanisms at the single-cell level (Peng and Croce, 2016), revealing cellular heterogeneity and identifying previously unknown cellular subpopulations in a variety of contexts (Han et al., 2020; Luecken and Theis, 2019). However, these technologies are yet to benefit the field of miRNAs because the reverse transcription stage of the current scRNA-seq protocol relies on poly(A) capture, which mature miRNAs lack. This limitation has precluded the standard scRNA technologies from profiling miRNAs and, consequently, there are very few published miRNA profiles at single-cell resolution (Faridani et al., 2016; Wang et al., 2019). Even though there are large miRNA atlases that profile genome-wide miRNA expression in specific tissues or purified cell culture (Lorenzi et al., 2021; deRie et al., 2017; McCall et al., 2017), there is a general lack of *in vivo* miRNA expression profiles in single cells, which severely limits our understanding of miRNAs function and dynamics at cellular resolution, and investigation of miRNAs' role in the emergence of cell states.

Previous attempts to infer miRNA expression from the expression of protein-coding genes have relied on the assumption that reduced expression of a miRNA's putative targets, where targets are ascertained based on various sequence-based approaches (Chang et al., 2008; Setty et al., 2012; Agarwal et al., 2015), is indicative of the miRNA activity (Israel et al., 2009; Nielsen and Pedersen, 2021). In particular, a recent tool, miReact (Nielsen and Pedersen, 2021), ascertains the activity of miRNA in a transcriptomic profile based on the cross-gene correlation between the miRNA seed sequence match score in the gene's 3′ UTR and the gene's expression, where a strong negative correlation indicates miRNA activity. Such approaches are similar in spirit to SCENIC (Aibar et al., 2017), a computational tool, which infers the activity of a transcription factor in a cell based on the expression of its putative targets. However, because of (1) highly incomplete and noisy nature of miRNA-target relationships and (2) a variable effect of miRNA induction on its targets' expressions(Rzepiela et al., 2018), owing to diverse mechanisms underlying the effect of a miRNA on its targets, reliance on putative targets alone to infer miRNA activity is far from ideal; we

[1]Cancer Data Science Lab, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA

[2]Lead contact

*Correspondence: sridhar.hannenhalli@nih.gov

https://doi.org/10.1016/j.isci.2022.104962

explicitly demonstrate this assertion. Instead, we hypothesize that the complex direct and downstream indirect regulatory links between miRNAs and the expression of other mRNAs (both protein-coding and non-coding) can be exploited to infer miRNAs at a single-cell level much more effectively than based on putative targets alone. Here, we report a computational tool– miRSCAPE –to infer miRNA expression in single cell clusters from the scRNA-seq data.

First, based on large cohorts of paired miRNA-mRNA profiles in TCGA, GTEx (Lonsdale et al., 2013), and CCLE (Barretina et al., 2012), we establish the cross-validation accuracy of miRSCAPE. We demonstrate miRSCAPE's superior accuracy relative to several alternatives, notably, miReact. We then validate miRSCAPE predictions in multiple independent datasets with experimentally profiled miRNAs in single cells or purified cell types. In two independent datasets where cell type-specific miRNA profiles for multiple cell types are available, miRSCAPE accurately infers cell type-specific miRNAs with an average correlation between predicted and observed inter-cell type fold-difference ~ 0.81. In a single cell miRNA induction experiment data (Rzepiela et al., 2018), miRSCAPE, trained independently, successfully distinguishes the cells with miRNA induction. In scRNA-seq data from pancreas and breast cancer, miRSCAPE identifies miRNAs associated with specific cellular states such as stemness and epithelial-mesenchymal-transition. We demonstrate the general utility of miRSCAPE by applying it to scRNA-seq data from pancreatic and lung cancers and, as a resource, we provide top active miRNAs in 56 cell types in the human cell landscape (HCL) (Han et al., 2020). In each of these applications, miRSCAPE accurately recapitulated the miRNAs previously implicated in each of these contexts and in many cases revealed a cell type specific role of individual miRNAs.

Overall, miRSCAPE is a versatile and robust computational tool to infer miRNA expression from scRNA-seq data. Application of miRSCAPE to the vast collections of available scRNA-seq will substantially expand our understanding of gene regulatory networks at cellular resolution. miRSCAPE is freely available as a standalone tool for the community at https://github.com/hannenhalli-lab/miRSCAPE.

RESULTS

Overview of miRSCAPE

Figure 1A illustrates the overall schematic of miRSCAPE. Details are in the STAR Methods section, but briefly, given a large compendium of paired miRNA-mRNA bulk RNA-seq data (Table S1), for each miRNA independently, we train an eXtreme Gradient Boosting (XGBoost) model to infer the miRNA expression based on the global mRNA profile of the sample (STAR Methods). The model accuracy is defined as the Spearman rank-order correlation between the predicted and the observed miRNA expression in the test samples, either within a sample across miRNAs or for a miRNA across samples. We establish the superiority of miRSCAPE relative to alternative approaches, including miReact. We assess miRSCAPE's accuracy in multiple independent datasets of experimentally profiled miRNAs in cancer cell lines and purified hematopoietic cell types, as well as a miRNA induction experiment. In single-cell validation, where miRNA and mRNA are profiled for multiple cell types, to quantify the model accuracy, we assess the extent to which the cross-cell-type fold-differences of the predicted miRNA expression values are correlated with those of the observed miRNA expression values.

miRNA Prediction in bulk data

We first benchmarked the accuracy of miRSCAPE in cancer bulk paired miRNA-mRNA RNAseq gene expression data from TCGA. We selected 10 cancer types from TCGA having at least a hundred matched pairs of miRNA and mRNA samples, and in each cancer type independently, we estimated the 5-fold cross-validation (CV) prediction accuracy of miRSCAPE. To minimize the inclusion of potential false positive annotations in miRNA databases (Fromm et al., 2021), and to maximize the inclusion of relevant miRNAs, we excluded the miRNAs expressed in fewer than half of the samples within each cancer type. When comparing the predicted and observed expression of a given miRNA across test samples, miRSCAPE achieved an accuracy of 0.45 Spearman correlation (STAR Methods) on average, ranging from 0.39 to 0.51 across the 10 cohorts. On average across cancer types, miRSCAPE achieves an accuracy of 0.4 or higher for 58% of the miRNAs (Figure 1B). We demonstrate that the cross-sample accuracy is sufficient to identify differentially expressed miRNAs between two cohorts (Figure S1). Although we use Spearman correlation between predicted and observed miRNA expression to quantify model accuracy, the predicted values can be interpreted as miRNA expression (Figure S2).

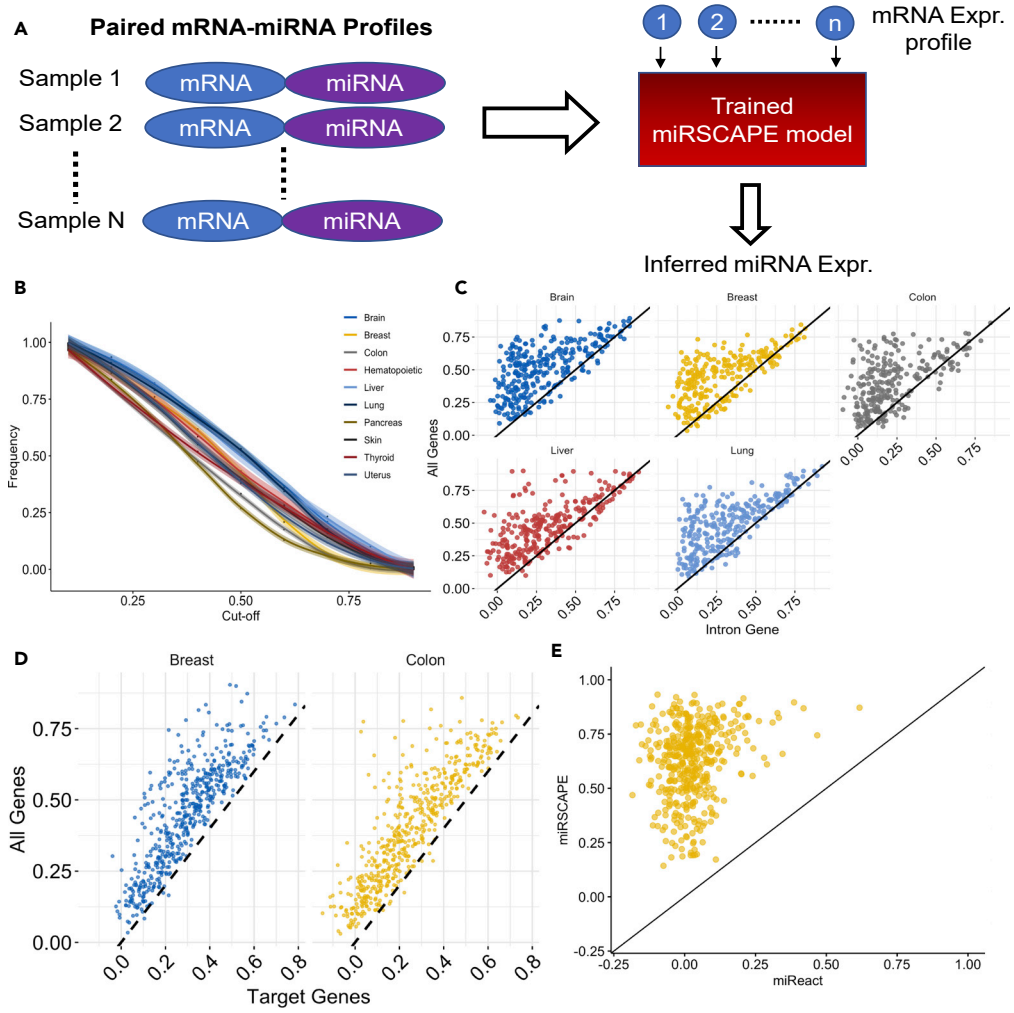


Figure 1. miRNA Prediction in Bulk Data

(A) miRSCAPE workflow. See text for details.

(B) miRSCAPE miRNA-wise cross-sample accuracy in 10 cohorts. The fraction of miRNAs (yaxis) having cross-validation accuracy greater than a given cut-off (xaxis).

(C) miRSCAPE accuracy compared with a model based only on the miRNA-containing gene. Accuracy of predicting intronic miRNAs when the model is trained on their host genes alone (xaxis) is compared with miRSCAPE accuracy based on all genes (yaxis).

(D) miRSCAPE accuracy comparison using all versus target-only genes as features. In the scatterplots, target-only-based accuracy (xaxis) is contrasted with all genes-based accuracy (yaxis) for two tissues; each dot represents a miRNA.

(E) Comparison between the accuracy of miReact and miRSCAPE. On a pooled TCGA cohort of five cancer types, miRSCAPE accuracy (yaxis) is compared with miReact accuracy (xaxis).

Our study is based on the widely used miRBase annotated miRNAs. Recognizing potential false positive miRNA annotations in miRBase, especially for lowly expressed miRNAs, an alternative curated database – miRGeneDB, was proposed (Fromm et al., 2021). Interestingly, we found that miRSCAPE cross-validation accuracies is much greater for the subset of miRNAs annotated in miRGeneDB (Figure S3). Specifically, for instance, on average across all cancer types the fraction of miRGeneDB-annotated miRNAs achieving an accuracy of 0.4 or greater was 0.76, compared to 0.58 when all miRNAs in miRbase were considered. Evolutionary conservation of miRNAs (Bartel, 2004) is utilized in annotating miRNAs to reduce false positives in the miRbase database. Segregating miRNAs by their conservation as provided in TargetScan v8 (Agarwal et al., 2015), we found that broadly conserved and conserved miRNAs are predicted with higher accuracy. For example, respectively in liver and colon, 89% and 80% miRNAs were highly predictable (accuracy $\geq$ 0.4) compared to 58% and 42% for the rest (Figure S4). Thus, our reported accuracy based

on miRbase-annotated miRNAs could be considered an underestimate. However, because it is not easy to ascertain false positives in miRBase, in the following, we include all miRNAs annotated in miRBase for greater coverage in various downstream analysis; as mentioned above, miRNAs expressed in fewer than half of the samples were excluded. Next, we investigated various potential correlates of accuracy and compared the accuracy of miRSCAPE with a number of alternatives.

First, we find a minimal effect of the number of samples on miRSCAPE accuracy (Figure S5). Second, we found the accuracy to be low, as expected, for very lowly expressing miRNAs but little dependence on miRNA expression beyond certain point (Figure S6). Third, half of the known miRNAs reside within an intron of another gene (Steiman-Shimony et al., 2018). Even though such intragenic miRNAs are expected to be co-expressed with the host gene, this is not universally true and for a substantial fraction of miRNAs a previous study found no clear correlation between the miRNA and host gene expression (Tan et al., 2019). Consistent with this, we find that for the intragenic miRNAs, our full miRSCAPE model far outperforms the alternative model based only on the host gene (Figure 1C).

Fourth, we compared miRSCAPE, which utilizes all genes as features, with the alternative model based only on the annotated targets of the miRNA, as was done previously (Israel et al., 2009) (Figure S7). A primary mechanism by which a miRNA affects its target mRNA is via mRNA degradation, which would imply a negative correlation between miRNA and the mRNA. However, this expected relation does not always hold given the variety of mechanisms by which a miRNA affects its target. As shown in Tan et al. (2019), positive miRNA-gene correlations are common across cancer types. Therefore, we assessed the extent to which a miRNA's activity is informed by the expression of its known target genes relative to other genes. To this end, we first compared the fraction of all genes and the fraction of known targets that are highly correlated with a given miRNA. In three sample cohorts (CCLE, colon cancer, breast cancer), we observe that the miRNA exhibits a significantly higher fraction of correlation with all genes compared with the known targets suggesting that non-target genes may contribute significantly to predicting a miRNA's activity (Figure S7). Next, we directly compared the cross-validation predictability of each miRNA using either only the experimentally known targets or using all genes. Again, as expected from the above, in all three cases, broadly for all miRNAs, the prediction accuracy using all genes is substantially greater than that achieved based on known targets alone (Figure S7). We therefore decided to use all genes to predict miRNAs and ascertained (Figure 1D) that miRSCAPE consistently outperforms the alternative model based only on the annotated targets of the miRNA.

Finally, we compared miRSCAPE directly (STAR Methods) with a recently published tool miReact (Nielsen and Pedersen, 2021) that infers a miRNA activity in a sample based on the genome-wide correlation between miRNA seed match in a gene's 3′ UTR and the gene's expression level. As shown in Figure 1E, miRSCAPE substantially outperforms miReact; Figure S8 includes the comparison for individual cancer cohorts. More specifically, in their pan-cancer application, miReact manuscript reports the top 50 most predictable miRNAs, having an average accuracy of 0.24. By comparison, in our pan-cancer application, the average accuracy of the top 50 most predictable miRNAs is 0.86. Regulatory networks are known to be highly context-specific, and consequently, a model trained in one regulatory context is not expected to perform well on data from a highly diverged context.

In many practical instances, one may want to predict miRNAs in a context/cell type/tissue which was not covered by the training cohort. We assessed miRSCAPE's applicability in such a scenario in 5 TCGA cohorts. We trained the model using samples from four of the five cancer types and predicted miRNA expression levels in the fifth left-out cohort. Across the five cancer types, and across miRNAs, miRSCAPE achieved an average accuracy of 0.27 (Figure S9). More specifically, on average 33% of miRNAs achieved an accuracy of 0.4 or higher in the 4 left-out cohort. This fraction was only 0.01 in the breast cancer cohort when trained on the other four cancer types. Relatively lower performance in breast cancer is consistent with the fact that the breast cancer samples are transcriptionally the most divergent from other cancers (Figure S9).

Characterizing broadly predictable miRNAs and important gene features

Next, we assessed the extent to which the highly predictable miRNAs and important features (genes) underlying those predictions are shared across cancer tissues. Figure 2A shows the number of highly predictable miRNAs in each tissue at two accuracy thresholds, and Figure 2B shows that, of the 10 tissues, on

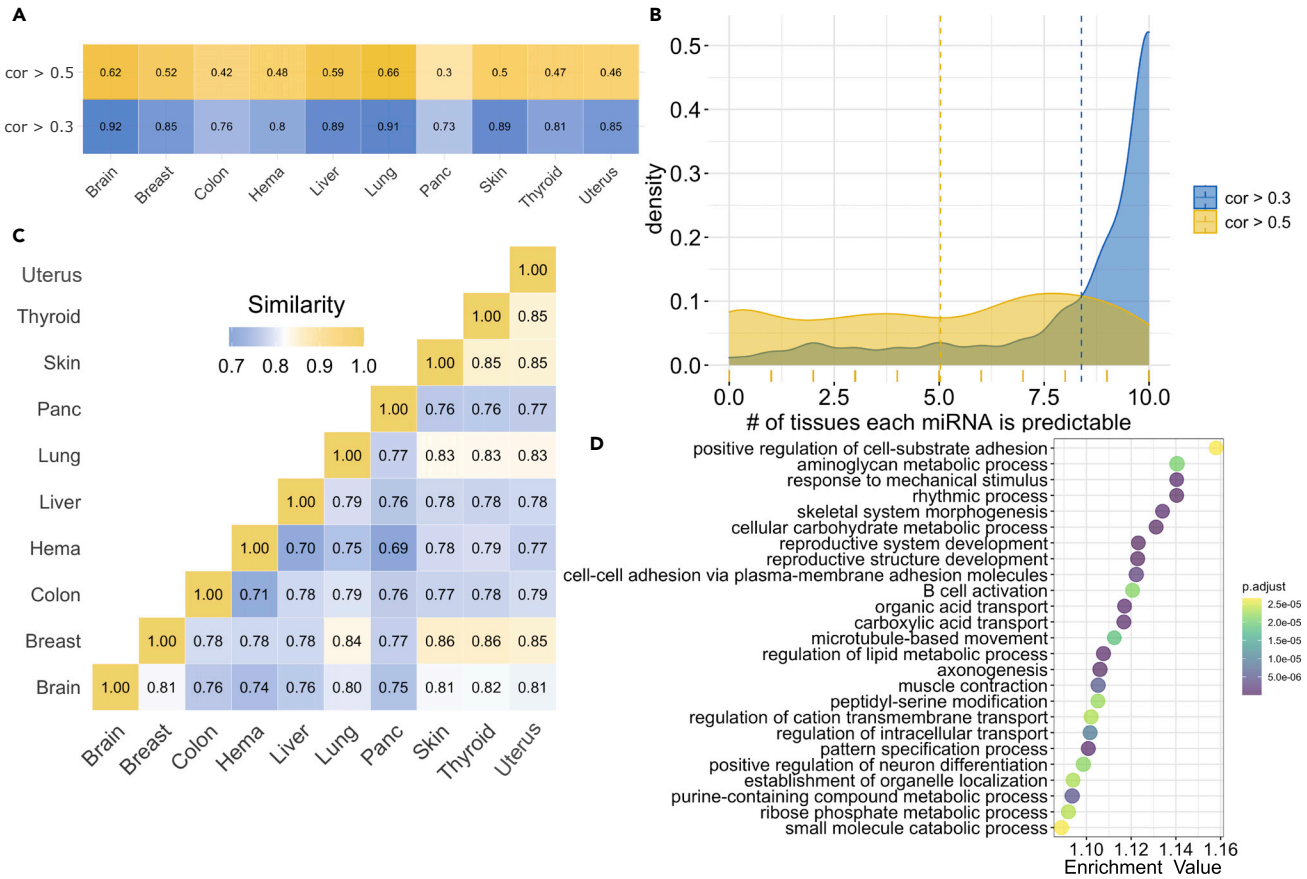


Figure 2. Predictable miRNAs and the contributing gene features

(A) The fraction of miRNAs that are highly predictable for each tissue type at two accuracy cutoffs.

(B) For two different accuracy thresholds to consider a miRNA predictable in a tissue, the plot shows the distribution of the number of tissues (xaxis) in which a miRNA is predictable in.

(C) Cross-tissue similarity (Jaccard index) in important features.

(D) Top 25 enriched biological processes among the globally most important gene features.

average, each miRNA is predictable with rho >0.3 in 8 tissues and with rho >0.5 in 5 tissues. The XGBoost algorithm underlying miRSCAPE additionally reports the most important features underlying the prediction of each miRNA (STAR Methods). First, we assessed whether experimentally known targets of a miRNA are preferentially deemed important by miRSCAPE. Toward this, for a given miRNA, we rank all genes based on the number of tissues in which the gene is deemed an important feature for the miRNA and tested, using the Wilcoxon test, whether known targets rank higher than the rest of the genes. Indeed, we found this to be the case for 67% of the miRNAs. Next, we compared, for each pair of cancers, the most important features in each cancer type (those that are deemed important for at least 20% of the miRNAs in the tissue). As shown in Figure 2C, whereas each cancer has specific important features, there is also a substantive overlap in important features across cancer types (Jaccard similarity ranges from 0.69 to 0.86); even when we only consider the features deemed important for at least 80% of the miRNAs (Figure S10), Jaccard similarity remains substantially high, ranging from 0.34 to 0.72. In addition, we observed that tissue-specific important features for a highly predictable miRNA (rho >0.8) are specifically expressed in the tissue (STAR Methods). Our observations that (1) targets are more likely to be important predictors and (2) more important tissue-specific features are more highly expressed in the tissue are consistent with the fact that targets are generally more highly expressed (Figure S11). Lastly, we defined a set of globally important features as those deemed important for at least 20% of the miRNAs in at least 5 cancer types and performed functional enrichment analysis. As shown in Figure 2D, the globally important gene feature genes are largely related to metabolic processes among others.

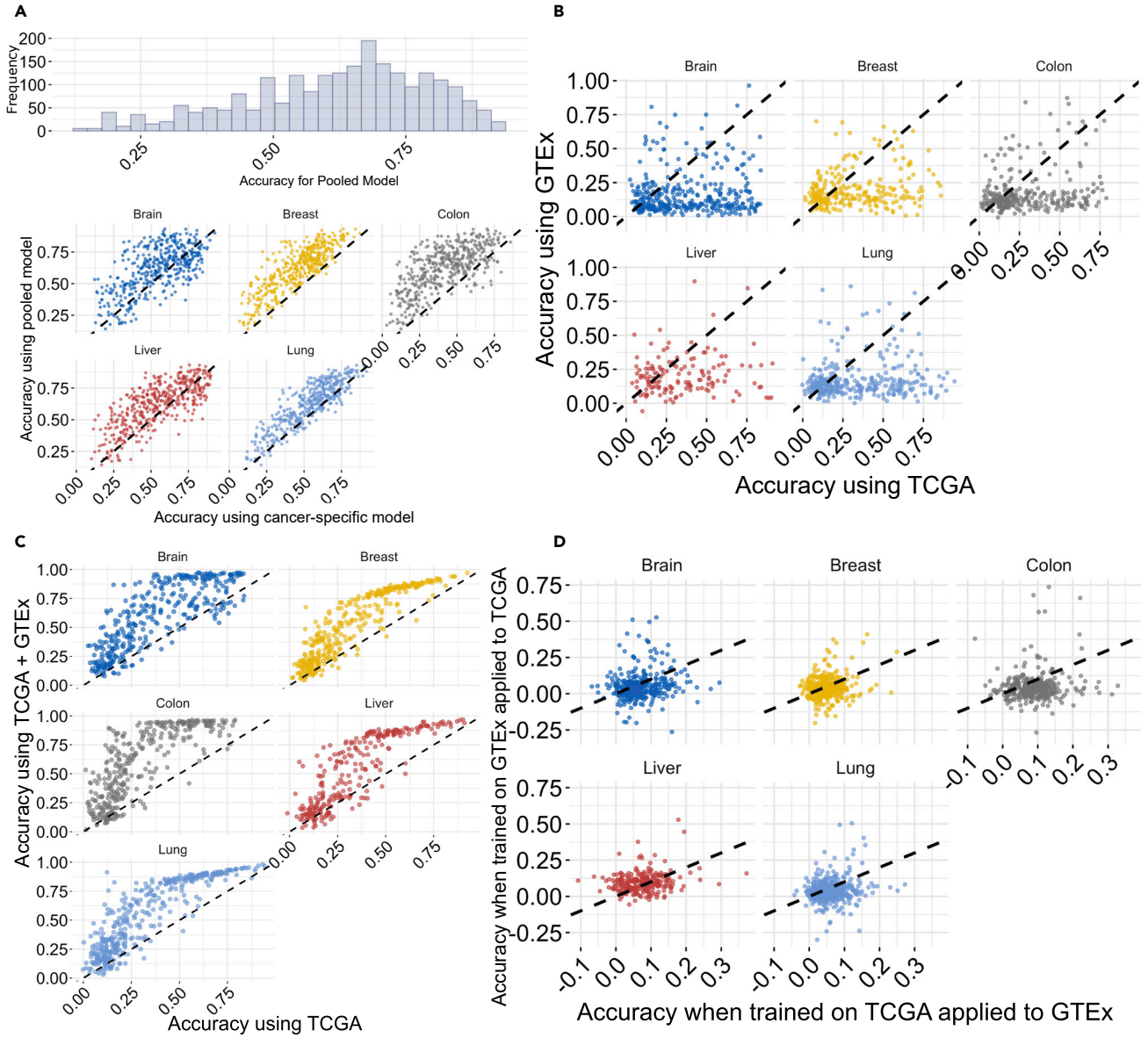


Figure 3. Effects of the inter-sample heterogeneity on model accuracy

(A) Comparison of accuracies based on the individual cancer samples versus samples pooled across five cancer types. The histogram (top) shows the distribution of accuracies for all miRNAs for the pooled samples, and the scatterplot (bottom) compares the accuracy for the pooled model (yaxis) with that for tissue-specific models (xaxis); each dot represents a miRNA.

(B) Comparison of accuracies for the TCGA cancer samples (xaxis) with those in the normal GTEx counterparts (yaxis).

(C) Comparison of accuracies for the TCGA samples (xaxis) versus those for the samples pooled across TCGA and the normal GTEx counterpart (yaxis).

(D) Cancer versus Normal cross-training and testing accuracy. The scatterplot shows the comparison between accuracies when the model is trained on TCGA samples and applied to the corresponding GTEx samples (xaxis) with the accuracies when the model is trained on GTEx samples and applied to the corresponding TCGA samples (yaxis).

Greater sample heterogeneity results in improved model accuracy

To assess the effect of sample heterogeneity on prediction accuracy, we quantified the model performance on 8,089 samples pooled across five cancer types (brain, breast, colon, liver, and lung). Figure 3A shows that, across all miRNAs, the pooled model (average CV accuracy: 0.62) performs substantially better than the cancer type-specific model – the average increase in CV accuracy is 0.2. The improved performance is likely because the model can capture the major differences in miRNA and other gene expression

values across tissue types. This is, however, not explained by the higher sample size in the pooled set (Figure S5).

Given a greater cross-sample heterogeneity in cancer compared with the healthy counterpart, we compared the prediction accuracies in five cancer types from TCGA with the accuracies in their healthy counterparts from GTEx (Lonsdale et al., 2013), as well as with the accuracy on pooled normal and cancer samples. We excluded miRNAs expressed in fewer than 10% of the samples in either the normal or the cancer cohort. As expected, a greater expression variability in cancer results in a more informative model yielding greater accuracy (Figure 3B), with an average accuracy of 0.34 in cancer and 0.18 in normal samples. When we pool the normal and the cancer samples, presumably because of the differential expression between normal and cancer and further increase in variability, the prediction accuracy is substantially increased to 0.54 on average (Figure 3C); this improved accuracy in the pooled data is consistent with the results obtained above when pooling multiple tissue types (Figure 3A). We further observed that the accuracy of a model trained on cancer samples and tested on corresponding normal samples is substantially better than the converse, again implicating greater heterogeneity in cancer on model accuracy (Figure 3D). However, consistent with our leave-one-tissue-out benchmarking above (Figure S9), we observed that the cross-cohort accuracy is lower than within-cohort cross-validation accuracy, suggesting substantive transcriptional and regulatory differences across tissues as well as between normal and cancer samples of the same tissue. Overall, these results suggest the context-specificity of the model as well as a substantive positive impact of sample heterogeneity on the model accuracy.

miRSCAPE accurately predicts cell type-specific miRNA expression

Having established the accuracy of miRSCAPE in the bulk data, next we assessed the extent to which miRSCAPE, trained on bulk data, can infer miRNA expression in particular cell types (Figure 4A and STAR Methods). We tested this in multiple independent datasets where miRNA and mRNA profiles are available for individual cell types based on either single cell profiling or bulk profiling of purified cell types.

Faridani et al. (2016) (GEO: GSE81287) measured single-cell miRNA expression profiles in the brain (166 cells) and kidney (45 cells). We trained miRSCAPE separately on the bulk brain and kidney cancers from TCGA and applied the models respectively to the scRNA-seq profiles of brain and kidney cells obtained from the HCL (Han et al., 2020), after bootstrapping (STAR Methods) yielding inferred cell type-specific miRNA profiles. We then compared the kidney versus brain fold-change (FC) based on the predicted miRNA expression with those based on observed miRNA expression from Faridani et al. (2016) Figure 4B shows that across 262 miRNAs, the predicted and the observed FC are highly correlated (Spearman rho = 0.8). Even when we used a single model trained on the pooled brain and kidney bulk data and applied the same model to predict the miRNA expression for both kidney and brain scRNA data, miRSCAPE achieved a high concordance between the predicted and observed FC (Spearman rho = 0.73; Figure 4B).

Faridani et al. also investigate miRNA expression in naive and primed human ES cells and observe that miR-302 family is more highly expressed in primed cells, whereas the miR-371-3 is more highly expressed in naive cells. We identified a dataset by Han et al. (GEO: GSE107552) that includes naive and primed human stem cells mRNA profiles. We applied miRSCAPE trained on the CCLE cohort and predicted miRNA expression in the naive and primed ES cells. miRSCAPE recapitulates the results of Faridani et al.; mir-302 predicted expression is higher in primed cells (logFC = 0.38, p-adj = 5.9×10^{-273}), whereas miR-371 is more highly expressed in naive cells (logFC = −1.84, p-adj = 8.4×10^{-158}).

Isakova et al. (2020) have applied Smart-seq-total simultaneously to profile within the same cell both miRNAs as well as protein-coding mRNAs in the skin (277 cells), breast (90 cells), and kidney (245 cells). We trained cell-type-specific miRSCAPE models for skin, breast, and kidney from the samples for the corresponding cancer types in TCGA (STAR Methods) and applied those models to the scRNA-seq data, after bootstrapping (STAR Methods) to predict cell-type specific miRNA profiles in the three cell types. For each cell type pair, as above, we compared the FC between the predicted and observed cell type-specific miRNAs. As shown in Figure 4C, the Spearman rank correlation coefficient for each of the three comparisons range from 0.76 to 0.89.

The paired miRNA and mRNA profiling in same cells in the Isakova et al. data provides an opportunity to assess miRSCAPE accuracy when trained and tested directly on scRNA-seq data. For each miRNA that is

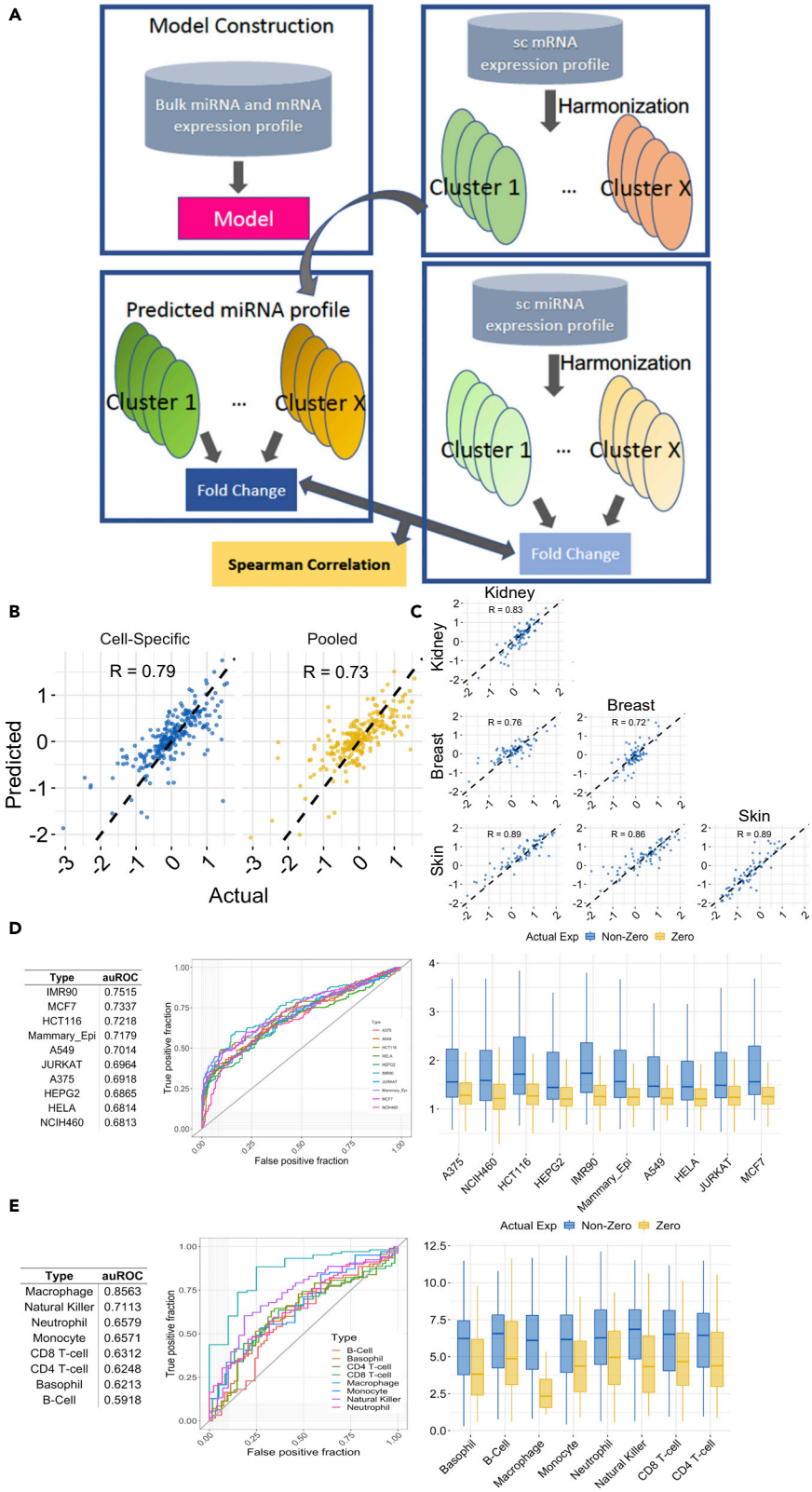


Figure 4. miRSCAPE validation in independent single cell data

(A) Validation pipeline. A model is learned for the matching bulk data and then miRNA expression is inferred from the single-cell mRNA data for individual cell types. To assess miRSCAPE predictability, for each miRNA, the fold-differences

Figure 4. *Continued*
between two cell types using the observed and the predicted miRNA expression are estimated, and the two sets of fold differences are compared across miRNAs using Spearman correlation.
(B) Validation on Faridani et al.data. Scatterplots comparing fold difference across cell types using observed miRNA expression (Xaxis) and predicted expression (yaxis). The left plot is based on cell type-specific models, and the right plot is based on a single pooled model, along with Spearman correlation coefficient (R).
(C) Validation on Isakova et al. Refer to B for details. Off-diagonal plots compare two cell types, and the diagonal plots compare one cell type (row) with the other two cell types pooled.
(D) Validation on ENCODE and McCall et al. Boxplot for miRSCAPE prediction values and the corresponding ROC curves and the table for auROC values to represent how miRSCPE discriminates the miRNAs with zero and non-zero expression in 10 ENCODE cell lines.
(E) Validation on hematopoietic cell types. Boxplot for miRSCAPE prediction values and the corresponding ROC curves and the table for auROC values show the extent to which miRSCAPE discriminates the miRNAs having experimentally detectable levels of expression from those with no detection in the cell type in 8 major hematopoietic cell types.

expressed in at least 5% of the cells or at most 95% of the cells, using the top 1,000 most variable genes as features, we trained and tested miRSCAPE, at single cell level, in leave-one-out cross-validation. Figure S12 shows that miRSCAPE successfully distinguishes the expressed and unexpressed miRNAs even when trained directly on single-cell data.

McCall et al. (2017) provide the landscape of human cell-specific microRNA expression in 42 cancer or immortalized cell lines (as well as other cell lines and tissues). We identified 10 ENCODE cell lines matching those profiled in McCall et al. We trained a model on the paired miRNA-mRNA data in 942 CCLE cell lines and predicted the expression values of 427 miRNAs in each of the 10 cell lines independently based on expression profiles in ENCODE and compared the predicted miRNA expressions with the measured values in McCall et al. However, in each cell line, only a small subset of miRNAs have detected (non-zero) expression in McCall et al. and therefore our standard approach to quantify accuracy based on cross-sample correlation of predicted and observed values was not feasible. Instead, we assessed miRSCAPE accuracy in two alternative ways. First, we found that in each cell line, the predicted values of miRNAs with non-zero expression are significantly (and substantially) higher than those with zero expression (Figure 4D). We assessed the accuracy with which the predicted values could discriminate the observed miRNAs from the undetected miRNAs in terms of auROC and again observed a reasonable accuracy across cell lines (Wilcoxon pvalue = 9.3×10^{-19}, average auROC = 0.71, Figure 4D). Second, for each miRNA, we grouped the cell lines as positive (non-zero expression of the miRNA) and negative (zero expression of the miRNA) and compared the mean expression of positive and negative groups across miRNAs. miRSCAPE predicted value for positive cell lines was significantly greater than those for negative cell lines (Paired one-side Wilcoxon pvalue = 1.9×10^{-9}).

Similar to the above application, next, we assessed miRSCAPE's ability to infer miRNA expression in hematopoietic cell types. Because we could not obtain both miRNA and mRNA profiles in human hematopoietic cell types, we instead assessed the extent to which miRSCAPE trained on human bulk data can infer miRNA activities in mouse hematopoietic cell types. Toward this, we trained miRSCAPE on 151 acute myeloid leukemia (AML) samples in TCGA. We then obtained transcriptional profiles of >15,000 mouse cells across 8 major immune cell types from the Zilionis et al. study (Zilionis et al., 2019) (GEO: GSE127465), pooled and harmonized the data for each cell type, and applied the miRSCAPE model to infer cell type-specific miRNA activities. Using purified bulk miRNA-seq data for the 8 hematopoietic cell types (Petriv et al., 2010). As for the McCall et al. dataset, very few miRNAs had non-zero expression in any cell type, and we followed the same assessment approach as above. For each cell type, individually, we assessed whether the inferred expression of miRNAs could distinguish the miRNAs having experimentally detectable levels of expression in the cell type from the undetected miRNAs. In all 8 cell types, this was indeed the case (Wilcoxon pvalue < 2.5×10^{-7}, average auROC = 0.67); Figure 4E shows the ROC curves and the auROC values for each cell type, clearly suggesting that a model trained in human bulk data can still identify cell type miRNAs in the mouse. Second, we compared predicted values of miRNAs in their positive (non-zero expression) and negative (zero expression) cell lines and found that the miRSCAPE predicted value for positive cell lines was significantly greater than those for negative cell lines (Paired one-side Wilcoxon pvalue = 0.03).

As final validations, we assessed whether miRSCAPE could recapitulate miRNA activity in miRNA knock-out (Loeb et al., 2012) and miRNA induction (Rzepiela et al., 2018) experiments. Loeb et al. (2012) have provided

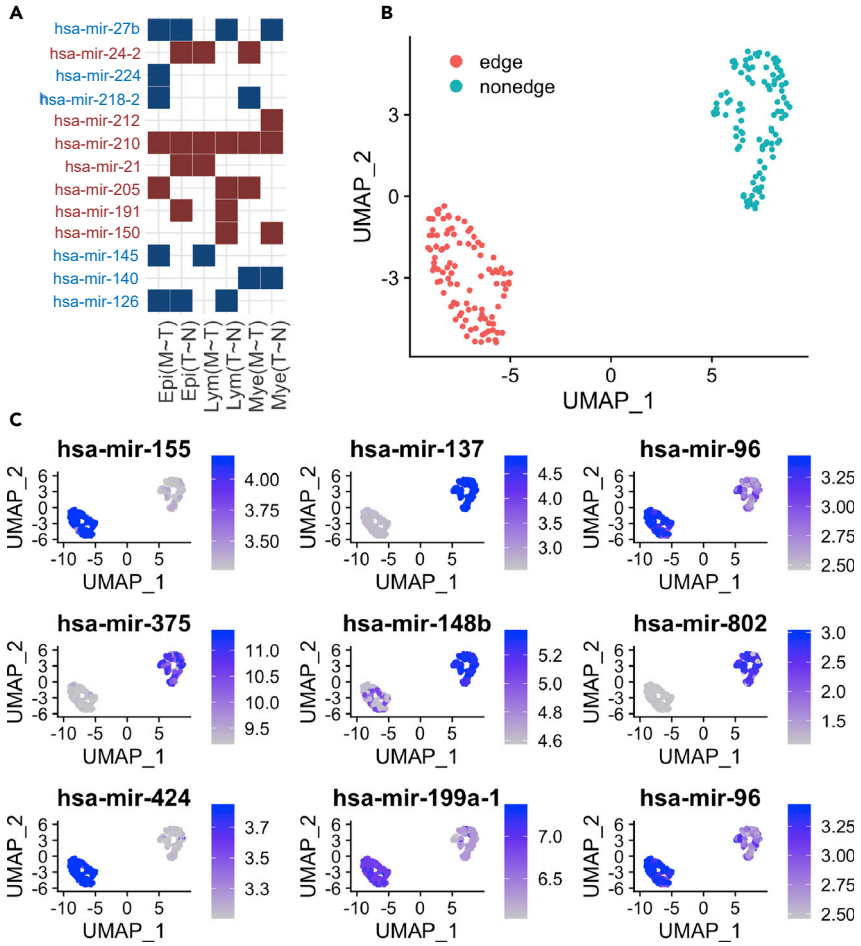


Figure 5. miRSCAPE application to single cell data

(A) Application to lung adenocarcinoma. Heatmap for predicted fold changes of the miRNAs (yaxis) in each of the epithelial (Epi), lymphoid (Lym), and myeloid (Mye) cell types in Primary Tumor ~ Normal (T ~ N) and Metastasis ~ Primary Tumor (M ~ T) (xaxis). Red represents the upregulated and blue represents the downregulated miRNAs.

(B) Application to Oncogenic states in healthy pancreatic acinar cells UMAP plot for pancreatic acinar cell states based on their global predicted miRNA profiles.

(C) Feature Plot for the select miRNAs. For selected miRNAs, the predicted expression among the edge and non-edge acinar cells are depicted.

the mRNA expression data between WildType (WT) and miR-155 deficient (155KO) primary CD4 T cells in the mouse. We applied the miRSCAPE model for miR-155 trained on 151 human AML samples in TCGA and predicted the expression of miR-155 in WT (3 samples) and 155KO (3 samples) mouse samples based on gene orthology mapping (STAR Methods). Despite the fact that a model trained on human data was applied to mouse samples, we observed significantly lower predicted expression of mir-155 in 155KO relative to WT (FC = 0.55, Wilcoxon pvalue = 8.54×10^{-18}).

Rzepiela et al. (2018) have reported scRNA-seq data before and after induction of miR-199a-5p and miR-199a-3p in two cell lines. We applied miRSCAPE models for two miRNAs, both trained on CCLE cell lines, and predicted their expression in the induced and uninduced cells. First, based on filtered, bootstrapped and pooled single cell data (STAR Methods), in all 4 cases (2 miRNAs in 2 cell types), we observed significantly higher predicted values in induced cells (log FC = 6.2, 6.0, 5.9, 4.9; all Wilcoxon pvalues < E−72). As a control, we also predicted the other available miRNAs in the CCLE. We observed that miR-199a-5p and miR-199a-3p ranked as the top miRNAs in the miR199 cell line and was ranked among the top 5 miRNAs in the KTN cell line. Next, when we applied the CCLE trained model to single cell transcriptomes, without pooling multiple cells (STAR Methods), despite data sparsity, we observed significantly higher predicted

values in induced cells (logFC = 2.6, 2.6, 1.9, 1.5; Wilcoxon pvalues < E−36) consistent with fold changes observed by Rzepiela et al. In this single cell application, we compared miRSCAPE performance with that of miReact (STAR Methods). We found that miReact could not consistently infer the upregulation of miR199 in induced cells (log FC = 0.33, 0.15, −1.00, −1.13).

Overall, in all independent validations in single-cell or bulk purified cell type datasets, our model trained on human bulk data, including cross-species application, achieved a high concordance between the predicted and the observed miRNA expression, firmly establishing the generalizability and robustness of miRSCAPE.

Application of miRSCAPE to scRNA-seq data across a variety of contexts

Having established the accuracy of miRSCAPE via cross-validation and in multiple independent datasets, we set out to demonstrate miRSCAPE's application to scRNA-seq data. Although XGBoost is designed to cope with sparse data (Chen and Guestrin, 2016), the extreme sparsity (missing feature values) of typical scRNA-seq data is a major challenge for XGBoost, and indeed for any machine learning approach. To this end, we assessed the effects of data sparsity on miRSCAPE's accuracy. To mimic the scRNA data at fixed read depth (10K, 100K, 1M) per sample/cell, we randomly down-sampled (STAR Methods) sequencing reads in each test sample and assessed the accuracy as above. We performed the cross-validation analysis in pooled 5 TCGA cohorts (brain, breast, colon, liver, and lung). As expected, we observed performance deterioration at lower depth (Figure S13), as was also observed previously (Nielsen and Pedersen, 2021). However, for high-depth scRNA-seq data with ~100K reads per cell, on average 53% miRNAs could be predicted with accuracy $\geq$ 0.4, compared to 87% of miRNAs for the original bulk data. This deterioration in performance is a limitation of the data and not necessarily of the model. We have discussed this further below where we showcase a variety of single cell applications of miRSCAPE.

Below, in various single cell applications we follow one of the three strategies. In most cases, to address the sparsity, we pool transcriptomes of randomly selected subset of cells (of certain type or cell cluster). In one instance, where our goal was to assess miRSCAPE's ability to identify miRNAs associated with EMT state, we pooled the k-nearest neighbors of each cell in the Principal Component space and pooled their transcriptome. Finally, to validate miRSCAPE on a single cell miRNA induction experiment, we apply miRSCAPE trained on bulk data directly to single cell transcriptomes. The specifics of these three modalities are provided in STAR Methods and in context below.

Pancreatic ductal adenocarcinoma (PDAC)

We applied miRSCAPE, trained on TCGA PDAC cohort, to scRNA-seq data in PDAC (Peng et al., 2019) comprising 57,730 cells from 35 donors. The cell annotations were obtained from Peng et al. (2019) and here, we focused on three cell types – acinar, normal ductal type 1, and the malignant (Ductal type II) cells. We applied miRSCAPE to each cell type separately after bootstrapping (STAR Methods) to predict miRNA expression.

The cell type that gives rise to PDAC is not entirely resolved and both acinar as well as ductal cells represent likely candidates for the cell of origin for PDAC (Ferreira et al., 2017). We, therefore, compared the differential miRNA expression in the malignant cells relative to acinar and ductal (Type I) cells individually as well as relative to pooled acinar and ductal cells. Table S2 shows the significantly differential miRNAs in the malignant cells in all three comparisons.

Based on bulk transcriptomes, Mazza et al. (2017) have reported differentially expressed miRNAs in PDAC relative to the normal pancreas, of which 39 are included in our study. Table S3, shows all miRNAs identified as up- or down-regulated in PDAC. Unlike Mazza et al. which compares bulk data, our approach enables us to compare specifically the malignant cells against normal acinar or the normal ductal (DC1), the two potential precursors of malignant cells. We consider our result consistent if the Mazza et al. result is recapitulated in at least one of the two comparisons. Overall, 77% (30/39) of the miRNAs were consistently recapitulated in our analysis. As a concordant example, miR-221, identified by Mazza et al. and miRSCAPE, is known to be upregulated in pancreatic cancer cell lines (Xu et al., 2015; Wu et al., 2020; Papaconstantinou et al., 2013) and plays a role in invasion, drug resistance, and apoptosis in PDAC (Wu et al., 2020). As a discordant example, miR-29a has a tumor-suppressive role in PDAC (Dey et al., 2020), and whereas Mazza et al. failed to identify miR-29A, miRSCAPE predicted miR-29a to be downregulated in malignant cells relative to acinar cells but surprisingly, upregulated relative to ductal type I cells. This suggests a pleiotropic

and potentially context-specific role, and also suggests acinar cells as potential PDAC cells of origin (Kopp et al., 2012). It may also explain why it may have been missed in bulk data comparison.

Lung cancer

Next, we applied miRSCAPE to scRNA-seq of lung adenocarcinoma (Kim et al., 2020) consisting of 208,506 cells derived from 44 individuals including 11 biopsies from adjacent-normal tissues, 14 primary tumors, and 9 metastatic tumors (Kim et al., 2020). Cell cluster annotations were obtained from the original publication, and here we focus on three cell types - epithelial, lymphoid, and myeloid cells. For each cell type, separately in normal, primary tumor, and metastatic tumor samples, as above, we pooled randomly selected cells to predict miRNA expression distributions in each cell type and each condition.

We compared the primary tumors with normal samples and also the metastatic tumors with the primary tumors separately in epithelial, lymphoid, and myeloid cells and identified the top 20 upregulated and downregulated miRNAs among 402 miRNAs (Table S4) for each cell type in each comparison. Yanaihara et al. (2006) have reported a small set of differentially expressed miRNAs in lung tumors relative to normal controls, of which 6 downregulated and 7 miRNAs were included in our study. Based on our predicted miRNA expression in the three cell types, we assessed the differential regulation of the 13 miRNAs in primary tumor versus normal and metastasis versus primary tumor. As shown in Figure 5A, miRSCAPE consistently identifies 6 up and 3 downregulated miRNAs in lymphoid cells, 5 up and 3 downregulated miRNAs in myeloid cells, 5 up and 5 downregulated miRNAs in epithelial cells.

More generally, miRSCAPE not only recapitulates many of the miRNAs associated with lung cancer, it in fact reveals their cell type specific roles (e.g., miR-4664 is upregulated specifically in epithelial cells of the tumor), which in some cases is even opposing in different cell types (e.g., miR-328 is downregulated in primary tumor myeloid cells but upregulated in the metastatic lymphoid cells). Many of our detected tumor-associated miRNAs in immune cells are known to be associated with leukemias, e.g., miR-21.

In lung epithelial cells, miRSCAPE recapitulates the results of Yamada et al. (2013), showing upregulation of miR-21 in lung tumor epithelial cells. However, in contrast to the epithelial cells, miRSCAPE predicts miR-21 as down-regulated in the myeloid cells of the tumor. This may represent a normal immune response consistent with a previous finding that miR-21 inhibition reduces the proportion of myeloid-derived suppressor cells in lung cancer (Meng et al., 2020). Of interest, miR-21 is also known to promote proliferation in AML (Li et al., 2019). Furthermore, miRSCAPE identified miR-100 among the top upregulated miRNAs in tumor myeloid cells, and like miR-21, miR-100 is a known oncomiR for AML (Bai et al., 2012), suggesting a link between miRNA functions in the lymphoid cells across contexts.

Comparing metastatic and non-metastatic lung adenocarcinomas, Sun et al. (2019) found five miRNAs to be upregulated in the metastatic tumors, among which miR-210 was included in our dataset. miRSCAPE successfully identifies miR-210 among the top 20 upregulated miRNAs in the brain metastasis compared to the primary tumor in epithelial cells. Similar to epithelial cells, miR-210 is also upregulated in the myeloid and the lymphoid cells of the primary tumor relative to normal lung samples. In addition, miR-145 is known to be downregulated in LUAD patients with brain metastasis (Sun et al., 2019). Consistently, miRSCAPE identifies miR-145 to be downregulated in LUAD patients with brain metastasis, specifically in the epithelial cells.

MiR-328 represents another interesting case. It promotes myeloid differentiation (Beitzinger and Meister, 2010). Eiring et al. (2010) have shown loss of miR-328 in CML. We observed the downregulation of miR-328 in the lung tumor myeloid cells. However, in contrast, miR-328 is upregulated in the lymphoid cells of the brain metastatic tumors, consistent with Arora et al. (2011) again suggesting a complex, context-specific role of miRNAs revealed by miRSCAPE.

Analyzing the peripheral blood lymphocytes of pulmonary sarcoidosis (linked with lung cancer), Kiszałkiewicz et al. (2016) found significant upregulation of miR-222 and significant downregulation of let-7f in patients compared to controls. miRSCAPE recapitulates these findings in lung cancer lymphoid cells versus normal lung lymphoid cells.

Overall, our results suggest that miRSCAPE not only identifies key miRNAs involved in lung cancer it also provides an opportunity to investigate cell type-specific roles of these miRNAs in the context of lung cancer.

Epithelial-mesenchymal transition (EMT) cell states in breast cancer

Cook and Vanderhyden (Cook and Vanderhyden, 2020) provide scRNA-seq data for ~5k cells representing the temporal epithelial-mesenchymal transition (EMT) response in breast cancer. Cell annotations were derived from the original publication, and we only considered the cells in 0 and 7 days after EMT induction time points. We pooled nearest neighboring cells as a representative for each cell (STAR Methods) and applied the miRSCAPE model trained on the TCGA BRCA cohort on these pooled single cell transcriptomes. For each miRNA, we estimated the cross-cell Spearman correlation between the given miRNA's predicted expression and the EMT score using an established EMT signature (Cook and Vanderhyden, 2020) (STAR Methods). Table S5 includes the top 20 positively and negatively correlated miRNAs.

miR-577 and miR-200a are known to suppress EMT (Mongroo and Rustgi, 2010; Yin et al., 2018). miRSCAPE predicts these miRNAs as inversely correlated with EMT (Spearman Correlation −0.495 and −0.492, respectively (Mongroo and Rustgi, 2010; Yin et al., 2018)). Zhang et al. (Zhang and Ma, 2012) list 14 miRNAs that either promote (8 miRNAs) or suppress (6 miRNAs) EMT in breast cancer. Among the 6 promoter and 5 suppressor miRNAs included in our study, miRSCAPE recapitulated all 6 promoter miRNAs, and 2 suppressor miRNAs (correlation values for each miRNA: miR-10b = 0.82, miR-9 = 0.64, miR-155 = 0.62, miR-335 = 0.58, miR-21 = 0.49, miR-126 = 0.48, miR-31 = −0.06, miR-200 = −0.49). Koleckova et al. (2021) provide 3 downregulated and 9 upregulated miRNAs associated with EMT. Of the 8 miRNAs included in our study, miRSCAPE correctly predicted 5 (correlation values for miR-182 = 0.67, miR-574 = 0.63, miR-885 = 0.61, miR-22 = 0.60, miR-185 = 0.26).

Oncogenic states in healthy pancreatic acinar cells

In our previous work (Gopalan et al., 2021), we have identified a subpopulation of healthy pancreatic acinar cells (termed edge cells) that express markers of various oncogenic programs, including stemness and pancreatic progenitor program. We obtained the edge and non-edge cell (as annotated in (Gopalan et al., 2021)) transcriptomes from Peng et al. (2019), obtained bootstrapped representation of the single cell transcriptome (STAR Methods), and applied miRSCAPE (trained on TCGA pancreatic cancer data) to predict miRNA expression in the two cell types. Figure 5B shows the two cell populations based on the predicted miRNA profiles. Importantly, miRSCAPE identified several miRNAs that are differentially expressed between the edge and non-edge cells (Figure 5C), that have previously been implicated with proliferation and oncogenesis in pancreatic ductal adenocarcinomas (Ma et al., 2017; Neault et al., 2016; Tang et al., 2013; LaConti et al., 2011; Ge et al., 2022). A few representative examples are illustrated in Figure 5C.

Predicting miRNAs in 56 human cell types

Finally, we set out to chart a comprehensive global landscape of miRNA expression across all human cell types profiled via scRNA-seq in the HCL (Han et al., 2020). Our goal was to train a single global model that captures the co-variation among miRNAs and mRNAs and is uniformly applicable to all cell types. Toward this, we collected tumor and normal samples across ten diverse tissues from TCGA and GTEx (Table S1), comprising 13,764 samples. To reduce the sample space without compromising the captured variance, for each tissue separately, we clustered the tumor and the normal samples using k-means clustering into 100 clusters. We then selected the medoid of each cluster as its representative, yielding 100 samples for each tissue, with a total of 1,000 samples representing the global variation in the human body. These 1,000 samples were used to build a single global miRSCAPE model.

We obtained scRNA-seq data from HCL (Han et al., 2020), encompassing >700,000 cells across 56 different cell types comprising 36 adult, 18 fetal, 1 neonatal, and 1 placental sample. We pooled the scRNA-seq profiles after bootstrapping (STAR Methods) for each cell type separately and applied the global miRSCAPE model to each cell type yielding inferred activity distributions of 523 miRNAs across the 56 cell types. The top 10 upregulated and downregulated miRNAs in each cell type relative to all other cell types are provided in Table S6. Figure S14 shows the UMAP plot of the 56 cell types based on the predicted miRNA expression profiles and miRSCAPE recapitulates previous finding to a reasonable degree. Rie et al. (deRie et al., 2017) have reported an atlas of miRNAs in a large number of tissue/cell types and specifically provided, for each miRNA, the tissue/cell type where it has the highest expression. For a handful of tissues, we checked whether miRSCAPE was able to recapitulate the results in Rie et al. (1) Rie et al. reported that miR-9 is most expressed in Neural stem cells, miRSCAPE also found that both miR-9-3 (logFC:1.68, rank:2) and miR-9-1 (logFC:1.42, rank: 6) are enriched in the fetal brain compared to other tissues. (2) Rie et al. found that miR-23b has the highest expression in the pulmonary artery. Consistently, miRSCAPE

detected that miR-23b is enriched (logFC:4.77, rank: 48) in the adult heart compared to other tissues. (3) Rie et al. found that miR-204 has the highest expression in retinal pigment epithelial cells, and miRSCAPE detected that miR-204 (logFC:0.64, rank: 8) is enriched in fetal eyes compared to other tissues. (4) According to Rie et al. miR-23a, miR-24, and miR-27a are highly expressed in amniotic membrane cells and miRSCAPE predicted mir-23a (logFC: 1.63, rank:14), miR-24 (logFC:1.42, rank: 26), miR-27a (logFC: 1.08, rank:47) to be highly and preferentially expressed in the placenta.

Table S7 lists the top 20 differential miRNAs between the fetal and the adult tissues of each type. These results are meant as a public resource for follow up analyses. As such, knowledge of developmentally regulated miRNAs in individual tissues is relatively limited for us to corroborate our findings. Nevertheless, in the few tissues where such data are available, miRSCAPE recapitulates previous findings to a reasonable degree.

Thum et al. (Thum et al., 2007) report 52 upregulated and 40 downregulated miRNAs between human fetal and adult heart tissue. Of these 46 (30 upregulated and 16 downregulated) were included in our dataset of human cell atlas (Table S8). Among those 46 genes, miRSCAPE recapitulated 28 upregulated and 10 downregulated miRNAs corresponding to a concordance rate of 82.61%.

Tang et al. (Tang et al., 2011) performed a PCR-based quantification of a small set of miRNAs in matched human fetal and adult organs (including heart, kidney, liver, lung), and reported a handful of miRNAs highly expressed in the fetal stage in particular organs. For instance, they report that the let-7 family (7a, 7b, 7c, 7days, 7f and 7g) are highly expressed in the fetal stage compared to the adult stage in all organs. Consistently, miRSCAPE predicts upregulation in all let-7 families in fetal heart (logFC = 7.46), kidney (Max logFC = 0.6) and liver (logFC = 2.2). They also observe an overexpression for miR-26a in the fetal lung and kidney. Consistently, miRSCAPE predicted miR-26a to be upregulated in fetal lung (logFC = 2.04) and kidney (logFC = 0.5).

Burgess et al. (Burgess et al., 2015) identified 114 upregulated and 72 downregulated miRNAs from fetal to pediatric stage in liver. Among these, miRSCAPE includes 81 upregulated, 51 downregulated miRNAs. miRSCAPE correctly recapitulates 65 upregulated, 20 downregulated miRNAs (Table S9).

Williams et al. (Williams et al., 2007) report 9 upregulated and 13 downregulated miRNAs during the development of human lung. Of these 4 upregulated and 7 downregulated included in our study, miRSCAPE successfully recapitulates 6 of them. miRNAs (miR-26b = 1.52, miR-370 = −0.14, miR-154 = −0.22, miR-337 = −0.35, miR-134 = −0.74, miR-199b = −1.48).

DISCUSSION

Single-cell RNA-seq technologies have matured and are routinely used to generate large scRNA-seq data, effectively capturing protein-coding genes, in a wide variety of contexts. However, analogous technology specifically for small non-coding RNA sequencing, specifically miRNAs, is significantly lagging and has only been demonstrated in a handful of cell types (Faridani et al., 2016; Isakova et al., 2020). This has left a substantial gap in our understanding of miRNA transcriptional dynamics at cellular resolution. To overcome this limitation, here we report a machine learning tool, miRSCAPE, to predict the miRNA expression in single-cell clusters from their genome-wide mRNA profiles.

First, based on paired miRNA-mRNA profiles in ~5,000 samples in TCGA spanning 10 cancer types and ~4,500 samples in the corresponding healthy tissues in GTEx (Lonsdale et al., 2013), we establish the cross-validation accuracy of miRSCAPE in multiple scenarios. Within a test sample, miRSCAPE can accurately rank miRNAs by expression level (average cross-miRNA correlation in predicted and actual ranks within a sample ~0.93). For a given miRNA, the correlation between the predicted and the observed expression across the test samples is 0.45 on average across 10 cohorts which, as we show both in bulk and in single cell application, is sufficient for an accurate identification of miRNAs that are differentially expressed between two sets of samples (Figures S1, 4B, and 4C). We demonstrate miRSCAPE's superior accuracy relative to several alternatives, notably, the only comparable tool — miReact. We validate miRSCAPE predictions in multiple independent datasets of experimentally profiled miRNAs in cancer cell lines (McCall et al., 2017) and purified hematopoietic cell types (Petriv et al., 2010), as well as a miRNA induction experiment (Rzepiela et al., 2018). Next, in two independent datasets where cell type-specific miRNA

profiles are available (HEK-GBM (Faridani et al., 2016), naive and primed human stem cell (Faridani et al., 2016), kidney-breast-skin (Isakova et al., 2020)), along with the mRNA profiles of those cell types, we show that miRSCAPE, trained on TCGA data, can accurately infer cell type-specific miRNA activities with an average correlation between predicted and observed inter-cell type fold-difference $\sim$ 0.81. We also demonstrate that when miRSCAPE is trained and tested on the single cell data directly, without pooling, miRSCAPE can distinguishes the expressed and unexpressed miRNAs (Figure S12). We further demonstrate that in scRNA-seq data from pancreas and breast cancer, miRSCAPE can identify miRNAs associated with specific cellular states such as stemness (e.g., miR-155) and epithelial-mesenchymal-transition (e.g., miR-200a), respectively. Finally, we demonstrate the general utility of miRSCAPE by applying it to scRNA-seq data from pancreatic ductal adenocarcinoma (PDAC), Lung cancers, as well as in 56 cell types in the HCL (Han et al., 2020). In each of these applications, miRSCAPE accurately recapitulated the miRNAs previously implicated in each of these contexts and, in many cases, revealed a cell type specific role of individual miRNAs. Our tool and the associated freely available software open up the possibility to leverage a vast compendium of scRNA-seq datasets to understand miRNA activities at cellular resolution.

The mechanistic premise of miRSCAPE is that the miRNA activity is reflected, directly or indirectly, in the global transcriptomic profiles. However, the global transcriptomic profile reflects not only miRNA activity but several other cellular features, such as protein activities, metabolite levels, enhancer activity, DNA methylation, etc., and in principle, machine learning can predict these other cellular features from the global transcriptomic profiles. Thus, the miRSCAPE framework can be extended to estimate these other features at a cellular resolution if appropriate paired data are available for training the model. It is also worth noting that using protein expression levels instead of mRNA levels may result in a better model, however, large cohorts of paired miRNA-protein expression data are currently not available to train and assess such a model.

One potential limitation in applying miRSCAPE to scRNA-seq data is sparsity of scRNA-seq data. Indeed, at sequencing read counts of 50–100K read per cell, higher than in typical 10X scRNA-seq data, the performance of miRSCAPE deteriorates, as we show above. This is, however, not necessarily a limitation of the approach, but the data. Even in actual scRNA-seq data, any meaningful biological interpretation is made at the level of cell clusters and not for individual cells. For instance, although the previous tool miReact was applied to predict miRNA expression in single cells, the authors nevertheless interpreted the predicted miRNA at the level of cell clusters. The concept of pooling similar cells into a 'pseudo-cell' in order to increase the read depth for downstream analysis was recently proposed (Tosches et al., 2018). In fact, a recent paper (Squair et al., 2021) compared various single cell differential expression methods and found that the methods that pooled cells in a cluster first had the best performance. All this evidence suggests that single cell data is best interpreted in terms of groups of cells, and not individual cells. As a practical matter, this is the approach we have chosen to take. Hence, as a practical matter, we pool cells within a single cell cluster and apply miRSCAPE to the pooled (and therefore not sparse) transcriptomic profile to infer miRNA activity at the level of single cell clusters. However, we have shown the efficacy of miRSCAPE in correctly recapitulating miRNA induction in single cells in Rzepiela et al. data. Furthermore, scRNA-seq data sparsity is not a major limitation because miRSCAPE can still identify miRNA activity for distinct cell types or cell states as long as those types or states are discernable based on the scRNA-seq profile by the standard tool (Hao et al., 2020). To add, the notion of a 'cluster' is relative, and in a typical scRNA-seq application, one can either define clusters at a very high resolution or even pool nearest neighbors to estimate smoothed miRNA expression values.

miRSCAPE is expected to perform the best when it is trained on a large training set that captures the transcriptional diversity of the specific context it is applied to. However, when precise cell or tissue type data is not available, one can consider a closely matching context for training. In a situation where multiple cells or tissue types are involved, a global model spanning all such contexts can be useful (as in our HCL application), with a small loss in performance compared to context-specific models, as demonstrated in our HEK-GBM validation (Figure 4B). When the training context is very different from the application context, a more substantive loss in performance is expected, as quantified in Figure S9. However, when using a global model, loss in performance may be compensated by the ability to uniformly apply a single model, making the inferences in specific sub-contexts directly comparable.

Overall, miRSCAPE enables studying the miRNA transcriptional dynamics in a vast variety of contexts where scRNA-seq data has been profiled, from development to homeostasis to diseases including cancer. We

have comprehensively benchmarked miRSCAPE and demonstrated its utility in multiple contexts and made the tool freely available. miRSCAPE thus represents an impactful advance toward leveraging the scRNA-seq data to expand our understanding of transcriptional dynamics at cellular resolution. miRSCAPE is available at https://github.com/hannenhalli-lab/miRSCAPE.

Limitations of the study

miRSCAPE has a few limitations. The first one is that unlike previous approaches that are based on presumed targets of the miRNA and do not require specific training, miRSCAPE needs to be trained on a large cohort of paired miRNA-mRNA data. The second potential limitation is context-specificity of the model. miRSCAPE is expected to have slightly reduced accuracy when the cellular/tissue context of the application sample does not perfectly match those of the training cohort. In such case we recommend a pan-context model that we have provided. Lastly, when applying to scRNA-seq data, given the large dropout rate, a pre-trained model may not work and miRSCAPE needs to be retrained using only the genes that are detected at a reasonable level in the specific scRNA-seq data.

DISCLAIMER

The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Data collection
  - Model construction and performance evaluation
  - Application of the model to scRNA-seq data
  - Comparing miRSCAPE with miReact in bulk data
  - Validation of miRSCAPE in miRNA induction experiment
  - Data downsampling
  - Analysis of tissue-specific gene features
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.104962.

ACKNOWLEDGMENTS

This work was supported by funding from the Intramural Research Program, National Cancer Institute, National Institutes of Health. The results here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. We thank Stefan Muljo, Shuo Gu, Thomas Gonatopoulos, Shan Li, Piyush Agrawal, and Sarthak Sahoo for their valuable feedback. We thank the authors of miReact for providing us the TCGA correlation values of miReact and Dr. Mihaela Zavolan for providing us their miRNA induction single cell transcriptomic data.

AUTHOR CONTRIBUTIONS

G.O. processed the data, analyzed the results, and wrote the manuscript. V.G. helped scRNA data transform, gave feedback and ran the miReact. S.H. supervised the project and wrote the manuscripts.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 2, 2022
Revised: May 9, 2022
Accepted: August 12, 2022
Published: September 16, 2022

REFERENCES

Agarwal, V., Bell, G.W., Nam, J.-W., and Bartel, D.P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. Elife *4*.

Aibar, S., González-Blas, C.B., Moerman, T., Huynh-Thu, V.A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., et al. (2017). SCENIC: single-cell regulatory network inference and clustering. Nat. Methods *14*, 1083–1086.

Arora, S., Ranade, A.R., Tran, N.L., Nasser, S., Sridhar, S., Korn, R.L., Ross, J.T.D., Dhruv, H., Foss, K.M., Sibenaller, Z., et al. (2011). MicroRNA-328 is associated with (non-small) cell lung cancer (NSCLC) brain metastasis and mediates NSCLC migration. Int. J. Cancer *129*, 2621–2631.

Bai, J., Guo, A., Hong, Z., and Kuai, W. (2012). Upregulation of microRNA-100 predicts poor prognosis in patients with pediatric acute myeloid leukemia. OncoTargets Ther. *5*, 213–219.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, J., Kryukov, G.V., Sonkin, D., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature *483*, 603–607.

Bartel, D.P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell *116*, 281–297.

Beitzinger, M., and Meister, G. (2010). Preview. MicroRNAs: from decay to decoy. Cell *140*, 612–614.

Burgess, K.S., Philips, S., Benson, E.A., Desta, Z., Gaedigk, A., Gaedigk, R., Segar, M.W., Liu, Y., and Skaar, T.C. (2015). Age-related changes in MicroRNA expression and pharmacogenes in human liver. Clin. Pharmacol. Ther. *98*, 205–215.

Chang, Y.-M., Juan, H.-F., Lee, T.-Y., Chang, Y.-Y., Yeh, Y.-M., Li, W.-H., and Shih, A.C.-C. (2008). Prediction of human miRNAs using tissue-selective motifs in 3′ UTRs. Proc. Natl. Acad. Sci. USA *105*, 17061–17066.

Chen, T., and Guestrin, C. (2016). XGBoost. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM).

Cook, D.P., and Vanderhyden, B.C. (2020). Context specificity of the EMT transcriptional response. Nat. Commun. *11*, 2142.

De Rie, D., Abugessaisa, I., Alam, T., Arner, E., Arner, P., Ashoor, H., Åström, G., Babina, M., Bertin, N., Burroughs, A.M., et al. (2017). An integrated expression atlas of miRNAs and their promoters in human and mouse. Nat. Biotechnol. *35*, 872–878.

Dey, S., Kwon, J.J., Liu, S., Hodge, G.A., Taleb, S., Zimmers, T.A., Wan, J., and Kota, J. (2020). miR-29a is Repressed by MYC in pancreatic cancer and its Restoration drives tumor-suppressive effects via downregulation of LOXL2. Mol. Cancer Res. *18*, 311–323.

Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc. *4*, 1184–1191.

Eiring, A.M., Harb, J.G., Neviani, P., Garton, C., Oaks, J.J., Spizzo, R., Liu, S., Schwind, S., Santhanam, R., Hickey, C.J., et al. (2010). miR-328 functions as an RNA decoy to modulate hnRNP E2 regulation of mRNA translation in leukemic blasts. Cell *140*, 652–665.

Faridani, O.R., Abdullayev, I., Hagemann-Jensen, M., Schell, J.P., Lanner, F., and Sandberg, R. (2016). Single-cell sequencing of the small-RNA transcriptome. Nat. Biotechnol. *34*, 1264–1266.

Ferreira, R.M.M., Sancho, R., Messal, H.A., Nye, E., Spencer-Dene, B., Stone, R.K., Stamp, G., Rosewell, I., Quaglia, A., and Behrens, A. (2017). Duct- and acinar-derived pancreatic ductal adenocarcinomas show distinct tumor progression and marker expression. Cell Rep. *21*, 966–978.

Fromm, B., Høye, E., Domanska, D., Zhong, X., Aparicio-Puerta, E., Ovchinnikov, V., Umu, S.U., Chabot, P.J., Kang, W., Aslanzadeh, M., et al. (2021). MirGeneDB 2.1: toward a complete sampling of all major animal phyla. Nucleic Acids Res. *50*, D204–D210.

Ge, W., Goga, A., He, Y., Silva, P.N., Hirt, C.K., Herrmanns, K., Guccini, I., Godbersen, S., Schwank, G., and Stoffel, M. (2022). miR-802 suppresses acinar-to-ductal reprogramming during early pancreatitis and pancreatic carcinogenesis. Gastroenterology *162*, 269–284.

Gopalan, V., Singh, A., Rashidi Mehrabadi, F., Wang, L., Ruppin, E., Arda, H.E., and Hannenhalli, S. (2021). A transcriptionally distinct subpopulation of healthy acinar cells exhibit features of pancreatic progenitors and PDAC. Cancer Res. *81*, 3958–3970.

Han, X., Chen, H., Huang, D., Chen, H., Fei, L., Cheng, C., Huang, H., Yuan, G.-C., and Guo, G. (2018). Mapping human pluripotent stem cell differentiation pathways using high throughput single-cell RNA-sequencing. Genome Biol. *19*, 47.

Han, X., Zhou, Z., Fei, L., Sun, H., Wang, R., Chen, Y., Chen, H., Wang, J., Tang, H., Ge, W., et al. (2020). Construction of a human cell landscape at single-cell level. Nature *581*, 303–309.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Iii, Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zagar, M., et al. (2020). Integrated analysis of multimodal single-cell data. Cell *184*, 3573–3587.e29.

Hinske, L.C., França, G.S., Torres, H.A.M., Ohara, D.T., Lopes-Ramos, C.M., Heyn, J., Reis, L.F.L., Ohno-Machado, L., Kreth, S., and Galante, P.A.F. (2014). miRIAD-integrating microRNA inter- and intragenic data. Database *2014*, bau099.

Huang, H.-Y., Lin, Y.-C.-D., Li, J., Huang, K.-Y., Shrestha, S., Hong, H.-C., Tang, Y., Chen, Y.-G., Jin, C.-N., Yu, Y., et al. (2020). miRTarBase 2020: updates to the experimentally validated microRNA-target interaction database. Nucleic Acids Res. *48*, D148–D154.

Isakova, A., Neff, N., and Quake, S.R. (2020). Single cell profiling of total RNA using Smart-seq-total. Preprint at bioRxiv. https://doi.org/10.1101/2020.06.02.131060.

Israel, A., Sharan, R., Ruppin, E., and Galun, E. (2009). Increased microRNA activity in human cancers. PLoS One *4*, e6045.

Jovanovic, M., and Hengartner, M.O. (2006). miRNAs and apoptosis: RNAs to die for. Oncogene *25*, 6176–6187.

Kim, N., Kim, H.K., Lee, K., Hong, Y., Cho, J.H., Choi, J.W., Lee, J.-I., Suh, Y.-L., Ku, B.M., Eum, H.H., et al. (2020). Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. Nat. Commun. *11*, 2285.

Kiszałkiewicz, J., Piotrowski, W.J., Pastuszak-Lewandoska, D., Górski, P., Antczak, A., Górski, W., Domańska-Senderowska, D., Migdalska-Sęk, M., Czarnecka, K.H., Nawrot, E., and Brzeziańska-Lasota, E. (2016). Altered miRNA expression in pulmonary sarcoidosis. BMC Med. Genet. *17*, 2.

Koleckova, M., Ehrmann, J., Bouchal, J., Janikova, M., Brisudova, A., Srovnal, J., Staffova, K., Svoboda, M., Slaby, O., Radova, L., et al. (2021). Epithelial to mesenchymal transition and microRNA expression are associated with spindle and apocrine cell morphology in triple-negative breast cancer. Sci. Rep. *11*, 5145.

Kopp, J.L., Von Figura, G., Mayes, E., Liu, F.-F., Dubois, C.L., Morris, J.P., 4th, Pan, F.C., Akiyama, H., Wright, C.V.E., Jensen, K., et al. (2012). Identification of Sox9-dependent acinar-to-ductal reprogramming as the principal mechanism for initiation of pancreatic ductal adenocarcinoma. Cancer Cell *22*, 737–750.

Laconti, J.J., Shivapurkar, N., Preet, A., Deslattes Mays, A., Peran, I., Kim, S.E., Marshall, J.L., Riegel, A.T., and Wellstein, A. (2011). Tissue and serum microRNAs in the Kras(G12D) transgenic animal model and in patients with pancreatic cancer. PLoS One *6*, e20687.

Li, C., Yan, H., Yin, J., Ma, J., Liao, A., Yang, S., Wang, L., Huang, Y., Lin, C., Dong, Z., et al. (2019). MicroRNA-21 promotes proliferation in acute myeloid leukemia by targeting Krüppel-like factor 5. Oncol. Lett. *18*, 3367–3372.

Loeb, G.B., Khan, A.A., Canner, D., Hiatt, J.B., Shendure, J., Darnell, R.B., Leslie, C.S., and

Rudensky, A.Y. (2012). Transcriptome-wide miR-155 binding map reveals widespread noncanonical microRNA targeting. Mol. Cell *48*, 760–770.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The genotype-tissue expression (GTEx) project. Nat. Genet. *45*, 580–585.

Lorenzi, L., Chiu, H.-S., Avila Cobos, F., Gross, S., Volders, P.-J., Cannoodt, R., Nuytens, J., Vanderheyden, K., Anckaert, J., Lefever, S., et al. (2021). The RNA Atlas expands the catalog of human non-coding RNAs. Nat. Biotechnol. *39*, 1453–1465.

Luecken, M.D., and Theis, F.J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. Mol. Syst. Biol. *15*, e8746.

Ma, D., Tang, S., Song, J., Wu, Q., Zhang, F., Xing, Y., Pan, Y., Zhang, Y., Jiang, J., Zhang, Y., and Jin, L. (2017). Culturing and transcriptome profiling of progenitor-like colonies derived from adult mouse pancreas. Stem Cell Res. Ther. *8*, 172.

Mazza, T., Copetti, M., Capocefalo, D., Fusilli, C., Biagini, T., Carella, M., De Bonis, A., Mastrodonato, N., Piepoli, A., Pazienza, V., et al. (2017). MicroRNA co-expression networks exhibit increased complexity in pancreatic ductal compared to Vater's papilla adenocarcinoma. Oncotarget *8*, 105320–105339.

Mccall, M.N., Kim, M.-S., Adil, M., Patil, A.H., Lu, Y., Mitchell, C.J., Leal-Rojas, P., Xu, J., Kumar, M., Dawson, V.L., et al. (2017). Toward the human cellular microRNAome. Genome Res. *27*, 1769–1781.

Meng, G., Wei, J., Wang, Y., Qu, D., and Zhang, J. (2020). miR-21 regulates immunosuppression mediated by myeloid-derived suppressor cells by impairing RUNX1-YAP interaction in lung cancer. Cancer Cell Int. *20*, 495.

Mongroo, P.S., and Rustgi, A.K. (2010). The role of the miR-200 family in epithelial-mesenchymal transition. Cancer Biol. Ther. *10*, 219–222.

Neault, M., Mallette, F.A., and Richard, S. (2016). miR-137 modulates a tumor suppressor network-inducing senescence in pancreatic cancer cells. Cell Rep. *14*, 1966–1978.

Nielsen, M.M., and Pedersen, J.S. (2021). miRNA activity inferred from single cell mRNA expression. Sci. Rep. *11*, 9170.

Papaconstantinou, I.G., Manta, A., Gazouli, M., Lyberopoulou, A., Lykoudis, P.M., Polymeneas, G., and Voros, D. (2013). Expression of microRNAs in patients with pancreatic cancer and its prognostic significance. Pancreas *42*, 67–71.

Peng, J., Sun, B.-F., Chen, C.-Y., Zhou, J.-Y., Chen, Y.-S., Chen, H., Liu, L., Huang, D., Jiang, J., Cui, G.-S., et al. (2019). Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. Cell Res. *29*, 725–738.

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

Petriv, O.I., Kuchenbauer, F., Delaney, A.D., Lecault, V., White, A., Kent, D., Marmolejo, L., Heuser, M., Berg, T., Copley, M., et al. (2010). Comprehensive microRNA expression profiling of the hematopoietic hierarchy. Proc. Natl. Acad. Sci. USA *107*, 15443–15448.

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

Robinson, D.G., and Storey, J.D. (2014). subSeq: determining appropriate sequencing depth through efficient read subsampling. Bioinformatics *30*, 3424–3426.

Rzepiela, A.J., Ghosh, S., Breda, J., Vina-Vilaseca, A., Syed, A.P., Gruber, A.J., Eschbach, K., Beisel, C., Van Nimwegen, E., and Zavolan, M. (2018). Single-cell mRNA profiling reveals the hierarchical response of miRNA targets to miRNA induction. Mol. Syst. Biol. *14*, e8266.

Sachs, M.C. (2017). plotROC: a tool for plotting ROC curves. J. Stat. Softw. *79*, 2.

Setty, M., Helmy, K., Khan, A.A., Silber, J., Arvey, A., Neezen, F., Agius, P., Huse, J.T., Holland, E.C., and Leslie, C.S. (2012). Inferring transcriptional and microRNA-mediated regulatory programs in glioblastoma. Mol. Syst. Biol. *8*, 605.

Squair, J.W., Gautier, M., Kathe, C., Anderson, M.A., James, N.D., Hutson, T.H., Hudelle, R., Qaiser, T., Matson, K.J.E., Barraud, Q., et al. (2021). Confronting false discoveries in single-cell differential expression. Nat. Commun. *12*, 5692.

Steiman-Shimony, A., Shtrikman, O., and Margalit, H. (2018). Assessing the functional association of intronic miRNAs with their host genes. RNA *24*, 991–1004.

Sun, G., Ding, X., Bi, N., Wang, Z., Wu, L., Zhou, W., Zhao, Z., Wang, J., Zhang, W., Fan, J., et al. (2019). Molecular predictors of brain metastasis-related microRNAs in lung adenocarcinoma. PLoS Genet. *15*, e1007888.

Tan, H., Huang, S., Zhang, Z., Qian, X., Sun, P., and Zhou, X. (2019). Pan-cancer analysis on microRNA-associated gene activation. EBioMedicine *43*, 82–97.

Tang, S., Bonaroti, J., Unlu, S., Liang, X., Tang, D., Zeh, H.J., and Lotze, M.T. (2013). Sweating the small stuff: microRNAs and genetic changes define pancreatic cancer. Pancreas *42*, 740–759.

Tang, Y., Liu, D., Zhang, L., Ingvarsson, S., and Chen, H. (2011). Quantitative analysis of miRNA expression in seven human foetal and adult organs. PLoS One *6*, e28730.

Thum, T., Galuppo, P., Wolf, C., Fiedler, J., Kneitz, S., Van Laake, Doevendans, P.A., Mummery, C.L., Borlak, J., et al. (2007). MicroRNAs in the human heart: a clue to fetal gene reprogramming in heart failure. Circulation *116*, 258–267.

Tosches, M.A., Yamawaki, T.M., Naumann, R.K., Jacobi, A.A., Tushev, G., and Laurent, G. (2018). Evolution of pallium, hippocampus, and cortical cell types revealed by single-cell transcriptomics in reptiles. Science *360*, 881–888.

Wang, N., Zheng, J., Chen, Z., Liu, Y., Dura, B., Kwak, M., Xavier-Ferrucio, J., Lu, Y.-C., Zhang, M., Roden, C., et al. (2019). Single-cell microRNA-mRNA co-sequencing reveals non-genetic heterogeneity and mechanisms of microRNA regulation. Nat. Commun. *10*, 95.

Wilkinson, L. (2011). ggplot2: elegant graphics for data analysis by WICKHAM, H. Biometrics *67*, 678–679.

Williams, A.E., Moschos, S.A., Barnes, P.J., and Lindsay, M.A. (2007). Maternally imprinted microRNAs are differentially expressed during mouse and human lung development. Dev. Dyn. *236*, 572–580.

Wu, X., Huang, J., Yang, Z., Zhu, Y., Zhang, Y., Wang, J., and Yao, W. (2020). MicroRNA-221-3p is related to survival and promotes tumour progression in pancreatic cancer: a comprehensive study on functions and clinicopathological value. Cancer Cell Int. *20*, 443.

Xu, Q., Li, P., Chen, X., Zong, L., Jiang, Z., Nan, L., Lei, J., Duan, W., Zhang, D., Li, X., et al. (2015). miR-221/222 induces pancreatic cancer progression through the regulation of matrix metalloproteinases. Oncotarget *6*, 14153–14164.

Xu, T., Su, N., Liu, L., Zhang, J., Wang, H., Zhang, W., Gui, J., Yu, K., Li, J., and Le, T.D. (2018). miRBaseConverter: an R/Bioconductor package for converting and retrieving miRNA name, accession, sequence and family information in different versions of miRBase. BMC Bioinf. *19*, 514.

Yamada, M., Kubo, H., Ota, C., Takahashi, T., Tando, Y., Suzuki, T., Fujino, N., Makiguchi, T., Takagi, K., Suzuki, T., and Ichinose, M. (2013). The increase of microRNA-21 during lung fibrosis and its contribution to epithelial-mesenchymal transition in pulmonary epithelial cells. Respir. Res. *14*, 95.

Yanaihara, N., Caplen, N., Bowman, E., Seike, M., Kumamoto, K., Yi, M., Stephens, R.M., Okamoto, A., Yokota, J., Tanaka, T., et al. (2006). Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. Cancer Cell *9*, 189–198.

Yin, C., Mou, Q., Pan, X., Zhang, G., Li, H., and Sun, Y. (2018). MiR-577 suppresses epithelial-mesenchymal transition and metastasis of breast cancer by targeting Rab25. Thorac. Cancer *9*, 472–479.

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS *16*, 284–287.

Zhang, Z.J., and Ma, S.L. (2012). miRNAs in breast cancer tumorigenesis (Review). Oncol. Rep. *27*, 903–910.

Zilionis, R., Engblom, C., Pfirschke, C., Savova, V., Zemmour, D., Saatcioglu, H.D., Krishnan, I., Maroni, G., Meyerovitz, C.V., Kerwin, C.M., et al. (2019). Single-cell transcriptomics of human and mouse lung cancers reveals conserved myeloid populations across individuals and species. Immunity *50*, 1317–1334.e10.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Cancer bulk RNA-seq	TCGA	https://portal.gdc.cancer.gov
Cancer bulk miRNASeq	TCGA	https://portal.gdc.cancer.gov
Cell line bulk RNASeq	CCLE	https://depmap.org/portal/download/
Cell line bulk miRNASeq	CCLE	https://depmap.org/portal/download/
Normal tissue RNASeq	GTEx v8	https://gtexportal.org/home/datasets
Normal tissue miRNASeq	GTEx v8	https://gtexportal.org/home/datasets
Small seq scRNA brain, kidney, naive and primed human ES cells	Faridani et al. (2016)	GEO: GSE81287
Smart-seq-total skin, breast, kidney cells	Isakova et al. (2020)	GEO: GSE151334
Naive and primed human stem single cell mRNA	Han et al. (2018)	GEO: GSE107552
Mouse hematopoietic single cell mRNA	Zilionis et al. (2019)	GEO: GSE127465
Bulk-purified miRNA	Petriv et al. (2010)	N/A
Single cell mRNA PDAC	Peng et al.,2019	N/A
ScRNA-seq of lung adenocarcinoma	Kim et al. (2020)	GEO: GSE131907
The edge and non-edge acinar cell state annotation	Gopalan et al. (2021)	N/A
Single cell RNA-seq for adult and fetal	Han et al. (2020)	https://db.cngb.org/HCL/index.html.
scRNA-seq data for EMT response in BRCA	Cook and Vanderhyden (2020)	GEO: GSE147405
Mouse mRNA data	Loeb et al. (2012)	GEO: GSE41241
scRNA-seq for miRNA induction experiment	Rzepiela et al. (2018)	N/A
Software and algorithms		
miRSCAPE	This paper	https://doi.org/10.5281/zenodo.6873151
miReact	Nielsen and Pedersen (2021)	https://github.com/muhligs/miReact
XGboost	CRAN	https://cran.r-project.org/web/packages/xgboost/index.html
Limma	Bioconductor	https://bioconductor.org/packages/release/bioc/html/limma.html
Seurat	CRAN	https://cran.r-project.org/web/packages/Seurat/index.html
miRBaseConverter	Bioconductor	https://www.bioconductor.org/packages/release/bioc/html/miRBaseConverter.html
sub-Seq	Bioconductor	https://www.bioconductor.org/packages/release/bioc/html/subSeq.html
clusterProfiler	Bioconductor	https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html
Ggpubr	CRAN	https://cran.r-project.org/web/packages/ggpubr/index.html
ggplot2	CRAN	https://cran.r-project.org/web/packages/ggplot2/index.html
plotROC	CRAN	https://cran.r-project.org/web/packages/plotROC/index.html
biomaRt	Bioconductor	https://bioconductor.org/packages/release/bioc/html/biomaRt.html

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Sridhar Hannenhalli (sridhar.hannenhalli@nih.gov).

Materials availability

This study did not generate new unique reagents.

Data and code availability

This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table.

All original code has been deposited at Github and is publicly available as of the date of publication. DOIs are listed in the key resources table.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Data collection

We obtained matched bulk RNA-Seq and miRNA-Seq gene expression data for cell lines (Cancer Cell Line Encyclopedia, CCLE (Barretina et al., 2012)), normal tissues (GTEx v8 (Lonsdale et al., 2013)), and cancer (TCGA), respectively from URLs https://depmap.org/portal/download/, https://gtexportal.org/home/datasets, and https://portal.gdc.cancer.gov/. Fragments Per Kilobase of transcript per million mapped reads upper quartile (FPKM-UQ) normalized RNA-Seq and reads per million mapped reads (RPM) miRNA-Seq data is obtained from the TCGA. TPM (Transcripts Per Kilobase Million) normalized RNA-seq data is utilized from CCLE and GTEx. We selected ten tissues having more than a hundred matched mRNA and miRNA samples in TCGA and GTEx (Table S1). We used miRNAs that are annotated in miRBase v21. Experimentally known miRNA target information was gathered from the miRTarBase (Huang et al., 2020) and intragenic miRNAs and their host genes are obtained from myriad (Hinske et al., 2014). We used miRGeneDB (Fromm et al., 2021) v2.1 from https://mirgenedb.org/. Conservation information is obtained from TargetScan v8 (Agarwal et al., 2015).

We made use of various single cell mRNASeq and miRNAseq expression (read count) profiles. Small seq single cell miRNA expressions were obtained from Faridani et al. (2016) (GEO: GSE81287) and Isakova et al. (2020) (GEO: GSE151334) studies. Naive and primed human stem single cell mRNA profiles were downloaded from Han et al. (2018) (GEO: GSE107552). Mouse hematopoietic single cell mRNA data and bulk-purified miRNA were collected from Zilionis et al. (2019) study (GEO: GSE127465) and Petriv et al. (2010) study, respectively. Single cell mRNA PDAC data were obtained from Peng et al. (2019) and lung cancer data was obtained from Kim et al. (2020) (GEO: GSE131907). The edge and non-edge acinar cell state annotations for the Peng et al. dataset was obtained from (Gopalan et al., 2021). Single cell RNA-Seq with cell type annotation in the human cell landscape (Han et al., 2020) (HCL) was obtained from https://db.cngb.org/HCL/index.html. We used the scRNA-seq data and cell annotations from Cook et al. (Cook and Vanderhyden, 2020) for the epithelial–mesenchymal transition (EMT) responses in BRCA (GEO: GSE147405). Mouse mRNA data on miR-155 knockout in CD4 T cells are obtained from the Loeb et al. (2012) (GEO: GSE41241).

For miRSCAPE validation of miRNA induction experiment, the scRNA-seq data in two cell lines with induced and uninduced annotation were obtained directly from the authors of Rzepiela et al. (2018).

To compare the performance of miRSCAPE with that of miReact, we obtained the prediction accuracy of each miRNAs (correlation between predicted and observed expression across samples) directly from the authors of Nielsen and Petersen (Nielsen and Pedersen, 2021).

Model construction and performance evaluation

Figure 1A illustrates the overall miRSCAPE pipeline. We chose to use Extreme Gradient Boost (XGBoost) as it designed to cope with sparse data, to reduce overfitting, and for faster training (Chen and Guestrin,

2016). Given a cohort of paired miRNA and mRNA profiles, we train an Extreme Gradient Boost (XGBoost) (Chen and Guestrin, 2016) machine learning model that uses the mRNA expression values of all genes as the features to predict individual miRNA's expression. XGBoost is a decision tree-based ensemble algorithm where new models are iteratively learned to minimize the errors in the previous model using gradient boosting in the function space. In our implementation, we used the XGBoost library in R language (Chen and Guestrin, 2016) to predict the miRNA expressions using all genes' expression values as features. We applied Grid Search, which is an exhaustive search to find optimal hyperparameters from the predefined subset of hyperparameters. Our hyperparameter space is provided in Table S10. For the parameters other than stated in Table S10, we used default parameter settings. To deal with the technological differences between different datasets, gene expression was (0–1)-normalized within each sample before being used as features.

Given a large cohort of paired log10-transformed bulk mRNA and miRNA RNAseq data, we evaluated the model performance based on 5-fold cross-validation which is a resampling method that randomly divides the data into five equal parts and in each iteration, four parts are used for learning/training whereas the fifth left-out part is used for test/validation. Model accuracy was evaluated in two ways: (1) within each test sample, we quantified the Spearman correlation between the predicted and observed expression across all miRNAs (2) for each individual miRNA, we quantified the Spearman correlation between the observed and the predicted expression values across the test samples. Note that the first approach simply tests whether the miRNA rank order based on predicted expression values match the rank order based on the observed expression in a sample, and admittedly is a simpler task. For validation on single cell data, we relied on cases where both sc-mRNA and sc-miRNA profiles are available for the same cells or, in most cases, same cell types. We then predicted cell type-specific miRNA profiles (using the given the cell type-specific scRNA profiles) and then estimated the fold-difference in expression values for each miRNA across pairs of cell types; we did this using both the predicted and the observed miRNA expression values; fold-differences are estimated using the limma package in R (Ritchie et al., 2015). Finally, we quantified the model accuracy as the cross-miRNA Spearman correlation between the observed and predicted fold-differences. For some of the validations - hematopoietic cell types and cancer cell lines, we assessed the prediction accuracy by comparing the predicted expression values of the detected and undetected miRNAs in a given cell type and quantified the accuracy using either the Wilcoxon test or auROC.

For each miRNA, different features make different relative contributions to the model accuracy. We identified those features for follow-up analyses using the *xgb.importance* function of the XGBoost library in R language (Chen and Guestrin, 2016).

Application of the model to scRNA-seq data

To apply a bulk-trained model to scRNA-seq data, we first applied the standard Seurat (Hao et al., 2020) global scaling normalization method. In all single cell applications, except for the EMT analysis, within a single cell type, or Seurat cluster, we generated 50 bootstrapped 'pseudo-bulk' data by sampling without replacement 80% of the cells and pooling their expression values. For EMT analysis, we carried out the PCA based on the 1,000 most variable genes. The first 10 PCs were used to compute the neighborhood graph of cells in PC space. For each (seed) cell, we first identify its 100 nearest neighbors using Euclidean distance and pooled the expression of those 100 cells as a representative of the seed cell. The EMT scores of each pooled cell was computed using the *AddModuleScore* function in Seurat package with genes sourced from the Hallmark EMT gene set (v7.2) from MsigDB. This is done in order to deal with dropout noise prevalent in scRNA-seq data as discussed in detail in Discussion. To be able to apply the model trained on bulk data to the pseudo-bulk data, we ensured that the feature values are (0–1)-normalized, as mentioned above, in each sample both when training and when applying the model to pooled scRNA-seq data.

When comparing the predicted miRNAs (which, by design, have the same sample-wise expression distribution as bulk miRNA) with observed cell type miRNA (which may follow a different distribution), we ensured that the observed cell type miRNA expression profile follows the same sample-wise distribution, by doing a Gaussian transformation to match the bulk miRNA distribution.

Comparing miRSCAPE with miReact in bulk data

Nielsen and Petersen (Nielsen and Pedersen, 2021) provide accuracy of each miRNAs (correlation between predicted and observed expression across samples) for TCGA. The correlation values of each miRNA were

obtained directly from the authors. miReact distinguishes the mature miRNAs based on the arm information (-3p and -5p) for the targets and there can be more than one miRNA expression for a given miRNA in TCGA. Therefore, one precursor miRNA can have multiple correlation values associated with it. To deal with these cases, we converted the mature miRNA names to precursor names using miRBaseConverter (Xu et al., 2018) R package. If there were more than one matching for the same precursor miRNA, we considered the average correlation across the mature miRNAs mapping a precursor.

Validation of miRSCAPE in miRNA induction experiment

Rzepiela et al. (2018) have reported scRNA-seq data before and after induction of miR-199a-5p and miR-199a-3p in two cell lines. We obtained scRNA-seq transcriptomic profiles directly from the authors. The data was provided as a Gene x Cell expression matrix. One of the "genes" corresponded to the GFP tag indicative of miRNA induction; the higher the expression of GFP, the greater the induction. For each of the two cell types (i1999 and i199-KTN1) independently, we divided up the cells into 'uninduced' (no detectable expression of GFP) and 'induced' based on top 20% expression of GFP. We then applied miRSCAPE models for two miRNAs - miR-199a-5p and miR-199a-3p, both trained on CCLE cell lines, and predicted their expression in the induced and uninduced cells, using two strategies. In the first bootstrapping strategy, we generated pooled expression data using bootstrapping (STAR Methods) and predicted the miRNA expressions in these pooled transcriptomes using the CCLE-trained model and compared the predicted expressions of each miRNA between the induced and uninduced cells using Wilcoxon one-sided test. In the second strategy, we applied the CCLE-trained model directly to single cell transcriptomes (without any pooling of cells). Toward this, we retained only the cells having number of detected genes among the top 25% of all cells. Based on the genes that were detected in at least 25% of these cells, we retrained miRSCAPE models for the two miRNAs in CCLE and predicted the miRNA expression in single cell and compared them between the induced and uninduced cells using Wilcoxon one-sided test.

Data downsampling

To assess miRSCAPE performance in typical scRNA-seq data, characterized by small library size, in each sample we randomly selected a prespecified number of reads (library size) and re-quantified the gene expression using sub-Seq R package (Robinson and Storey, 2014). This approach captures reads for each gene proportional to their expression in the bulk data.

Analysis of tissue-specific gene features

Here we compared the tissue-specific models for a given miRNA that is highly predictable among 10 different tissues. For a given miRNA μ, and a pair of tissues T1 and T2 where μ is highly predictable (rho >0.8), we identified the important features for μ, as reported by XGBoost, unique to each tissue. We then checked whether the tissue-specific important features for μ had a higher expression in the tissue where they were deemed important relative to the other tissue where they were not. Across 1,968,990 (μ,T1,T2) triplets we tested, we found that on average tissue-specific important features had 1.6-fold greater expression in the tissue where they were deemed important relative to the other tissue.

QUANTIFICATION AND STATISTICAL ANALYSIS

All biological functional analyses were performed with *enrichGO* function in clusterProfiler (Yu et al., 2012) R package. Figures are generated using the *gghistogram, ggscatter, ggboxplot* functions of ggpubr, and ggplot2 (Wilkinson, 2011) R packages, and ROC plots are created by ggplot2 extension for ROC curves plotROC (Sachs, 2017) R package with *geom_roc* function. To convert the miRNA gene names, we utilized miRBaseConverter (Xu et al., 2018) R package's *miRNA_MatureToPrecursor* or *miRNA_PrecursorToMature* functions where it is appropriate for the miRNA set and *getBM* function in biomaRt (Durinck et al., 2009) R package to convert the gene names. ScRNA analyses, generation of UMAPs, and Feature plots are performed using standard pipeline of the Seurat (Hao et al., 2020) R Package. We used Jaccard index to measure cross-tissue similarity in important features and one-sided Wilcoxon test to assess the prediction accuracy by comparing the predicted expression values of the detected and undetected miRNAs in a given cell type for some of the validations. Specific test used is mentioned in the context.