Integrating Multi-Omic Data with Penalized Cox Models for Improved Cancer Prognostic Signatures

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Andrew Willems
December 2024

Copyright © 2024 by Andrew Willems.
All rights reserved.

DEDICATION

For my parents who have sacrificed much to give me the best opportunities to explore my love of computing in a biological context.

ACKNOWLEDGEMENTS

First and foremost, I want to acknowledge my advisor Dr. Tian Hong who has been instrumental in shaping my view of doing good scientific work. For seeing my potential as a researcher and helping me grow tremendously to consider all the perspectives of a problem especially before jumping into a problem. I want to thank him for all his support and insight over the past half decade.

I also would like to acknowledge Dr. Nicholas Panchy as he also has been instrumental in helping me develop as a scientist. I am thankful for our many conversations on science and life that have made these past five years so enriching.

Next, I would like to acknowledge my closest friend, Ben Regan. He has been a constant source of emotional support over the course of this marathon. I am so thankful for his friendship.

Finally, I would like to acknowledge my family for their constant love and support through this process.

ABSTRACT

Cancer research faces a significant challenge in identifying crucial genes for prediction, treatment, and amelioration. The vast array of potential genes complicates the identification of genomic markers that effectively predict outcomes while remaining interpretable. This thesis introduces a novel miRNA-targeting metric that scores genes based on their interactions with miRNAs. We use this metric in an integrated approach to elucidate a small number of important biomarkers in colorectal cancer (CRC) and small cell lung cancer (SCLC).

We developed a method integrating gene prioritization steps with miRNA targeting information and penalized Cox regression models to identify biomarkers that accurately predict cancer survival. Our approach, which combines multiple types of omics information, significantly outperforms existing methods in biomarker identification from large omics datasets.

For instance, in CRC, our method achieved a concordance index (a measure of predictive accuracy) of 0.75 compared to 0.65 for standard gene expression-based approaches. Similarly, in SCLC, our method showed improved performance over commonly used biomarker discovery tools such as DESeq2 and edgeR. Moreover, our approach demonstrated good performance across various other cancer types, suggesting its broad applicability.

The identified signatures show strong literature support, with many genes previously implicated in various cancers. Furthermore, these biomarkers are enriched in cancer-relevant pathways, including iron metabolism, NOTCH3 signaling, and ALK signaling. Our pipeline's multi-step approach allows for easy substitution of other model types, enhancing its flexibility while maintaining interpretability by using a limited number of genes in the final signatures.

This research contributes to the field by providing a computationally accessible, adaptable, and effective method for identifying interpretable biomarkers. These findings have potential implications for personalized cancer prognosis and treatment strategies, paving the way for more targeted therapeutic approaches.

Table of Contents

LIST OF TABLES **7**

LIST OF FIGURES **7**

LIST OF SUPPLEMENTAL TABLES **12**

LIST OF SUPPLEMENTAL FIGURES **12**

CHAPTER I INTRODUCTION **15**

Difficulty of Building Prognostic Models in a Big Data Context 16
Survey of Prognostic Models Across Cancers 19
Challenges of Prognostic Models in Rare Cancers 23
EMT Background and Use in Prognostic Models 25
Thesis Objectives and Approach 26

REFERENCES **27**

CHAPTER II USING SINGLE-CELL RNA SEQUENCING AND MICRORNA TARGETING DATA TO IMPROVE COLORECTAL CANCER SURVIVAL PREDICTION **30**

Abstract 32
Introduction 32
Results 34
Discussion 41
Conclusions 43
Methods 43
2.1. Model Overview *43*
2.2. miRNA Metric for Ranking Genes *44*
2.3. Pre-processing scRNA-Seq Data *44*
2.4. Inferring EMT-Based Pseudotime *45*
2.5. MAD Metric for Ranking Genes *46*
2.6. Inference of Switch-Like Differential Expression along Single-Cell Trajectories (SDE) Metric for Ranking Genes *46*
2.7. Overall Gene Prioritization Scoring *47*
2.8. LASSO Regularized Cox Model *48*
2.9. Concordance Index (C-index) *48*
2.10. Kaplan-Meier Survival Analysis *49*
2.11. Datasets *50*

2.12. Implementation *50*
REFERENCES 51
APPENDIX 56

CHAPTER III COMPARATIVE SINGLE-CELL TRANSCRIPTOMES OF DOSE AND TIME DEPENDENT EPITHELIAL–MESENCHYMAL SPECTRUMS **63**

ABSTRACT 65
INTRODUCTION 65
METHODS 67
Cell culture *67*
Single-cell RNA-sequencing *67*
Data processing and integration *69*
Projection of single cell data in reduced dimensional space *70*
Signature genes in EMT spectrum *71*
GO Enrichment *72*
Prognostic models *72*
Clustering from shared-nearest-neighbor graphs *73*
RESULTS 74
A single-cell transcriptomic landscape of dose-dependent EMT reveals a continuum-like spectrum *74*
Dose-dependent states show less variation and are more separable than time-dependent states in the E/M space *77*
En route cell clusters unique to time dependent EMT program *84*
Signature genes from both dose- and time-dependent datasets contribute to better prognostic models *86*
DISCUSSION 88
DATA AVAILABILITY 90
REFERENCES 91
APPENDIX 94

CHAPTER IV AN EXPRESSION-BASED PROGNOSTIC MODEL FOR SMALL CELL LUNG CANCER **103**

ABSTRACT 105
INTRODUCTION 105
MATERIALS & METHODS 108
Model *108*
Backwards elimination procedure *108*
Synthetic Dataset Bootstrap procedure *108*
Power calculations *109*
Synthetic Sample Bootstrap procedure *109*
Datasets *110*
Implementation *110*
RESULTS 110
DISCUSSION 117
CONCLUSION 119
DATA AVAILABILITY 120
REFERENCES 120

CONCLUSION **123**

LIST OF TABLES

Table 1. Top Risk Increasing Genes Identified by TCGA-COAD Cox model: List of the top risk increasing genes identified by our Cox model on the TCGA-COAD dataset 38
Table 2. Top Risk Increasing Genes Identified by TCGA-READ Cox model: List of the top risk increasing genes identified by our Cox model on the TCGA-READ dataset..... 39
Table 3. Top Risk Decreasing Genes Identified by TCGA-COAD Cox model: List of the top risk decreasing genes identified by our Cox model on the TCGA-COAD dataset..... 40
Table 4. Top Risk Decreasing Genes Identified by TCGA-READ Cox model: List of the top risk decreasing genes identified by our Cox model on the TCGA-READ dataset. 40
Table 5. Statistically Significantly Enriched COAD Pathways: Most enriched pathways influenced by the genes with the largest hazard ratios associated with increased risk in our TCGA-COAD Cox model. P-value threshold < 0.05 FDR. Hazard ratio threshold > 2 41
Table 6. Statistically Significantly Enriched READ Pathways: Most enriched pathways influenced by the genes with the largest hazard ratios are associated with increased risk in our TCGA-READ Cox model. P-value threshold < 0.05 FDR. Hazard ratio threshold > 2..... 41
Table 7. Common language effective size of separation of dosage groups based on EMT scores 77
Table 8. Common language effect size of separation of time and dosage groups based on EMT scores of integrated data..... 81
Table 9. Whole Genome Sequencing Sample Characteristics. This table presents a modified version of the whole genome sequencing data from George et al., summarizing key clinical and demographic variables for 72 cancer patients. The variables include vital status, treatment modalities (radiation and chemotherapy), tumor staging (TNM), smoking status, sex, and age distribution. For N and M staging, 'X' indicates that the staging could not be determined. The age classification of 'Old' or 'Young' is based on the median age of patients in our dataset; patients above the median age are classified as 'Old', while those below are classified as 'Young'..... 110

LIST OF FIGURES

Figure 1: Example of Underfit, Properly fit, and Overfit Models. Here we show a toy example of what a un underfit model (Degree 1), properly fit (Degree 4), and an overfit model (Degree 15). We use a portion of the cosine function (orange line) to show how fitting a polynomial with varying numbers of degrees (teal line) impacts impact how well the model fits the observed data (blue dots). 17
Figure 2: Overall Design of CCsc. CCsc is composed of 5 steps. (**1**) Obtaining pre-processed scRNA-seq matrix (genes by cells) and performing imputation and denoising via the MAGIC package and then filtering for low abundance samples. (**2**) Using filtered, denoised, and imputed scRNA-seq matrix to perform metric dependent pre-processing steps. (A) For the SDE metric this involves ordering the scRNA-seq matrix by pseudotime based on the expression of the VIM marker. (B) For the miRNA metric, this involves sending the matrix to our novel miRNA metric. (C) We calculate the MAD metric from the scRNA-seq matrix. (**3**) In this step, we generate a

distinct gene list from each of our three metrics. (**4**) We integrated and merged the three separate gene lists into a single master list based on their score. From this master list, we then send N number of genes from the list as inputs to a LASSO penalized Cox model. The metric weights (a1, a2 and a3) are optimized via a grid search that involves the next step. (**5**) This step yields the outputs of our penalized Cox model that uses concordance index to assess our model's accuracy. We also generate Kaplan–Meier survival curves, based on the model output, to further validate the efficacy of the genes selected by our model. Based on the Cox model outputs, we can optimize the gene weights of our linear model to achieve the best performance. 34
Figure 3: Combination of metrics is better than any individual metric.(**A**) We tested each of our models' metrics and compared them to CCsc MS, CCsc MM, and CCsc MMS and a set of randomly selected genes, equal to the number of genes used by our method on the TCGA-COAD dataset. (**B**) We made the same comparison on the TCGA-READ dataset. We show that integrating the lists of ranked genes from each metric provides better performance than any one individually. (**C**) Kaplan–Meier estimate based on our model's top set of predictors for the TCGA-COAD dataset. (**D**) Kaplan–Meier estimate based on our model's top set of predictors for the TCGA-READ dataset. The shaded regions represent 95% confidence intervals of the survival estimates. The p-value threshold for significance is < 0.05. 36
Figure 4: CCsc MMS Outperforms Other Methods. (A) We tested CCsc against several well-established tools on the TCGA-COAD dataset. We compared the mean concordance index performance of CCsc with its ideal weighting to DESeq2, edgeR, scDD, and DEsingle. We gave each method its optimal number of genes and ran each with its recommended settings, according to their respective best practices. (B) Same comparison but on the TCGA-READ dataset. We show that CCsc outperforms single-cell methods (scDD and DEsingle) and that bulk RNA-seq methods can be optimized for single-cell RNA-seq data (DESeq2, edgeR). 37
Figure 5: Top coefficients identified by our model. (**A**) Top risk decreasing (green, left) and top risk increasing (red, right) genes identified by our Cox model in TCGA-COAD. (**B**) Top risk decreasing genes (green, left) and top risk increasing (red, right) genes identified by our Cox model in TCGA-READ. For risk increasing genes the hazard ratio threshold is > 2. For risk decreasing genes the hazard ratio threshold is < 0.5. 38
Figure 6: Analysis overview and progression of dose-dependent EMT at single-cell level. **(A)** A schematic of our analysis in this study. The analyses involve the existing time-course data (top) contain MCF10A cells at different time points following a common TGF-β treatment (about 200 pM; Deshmukh et al. [13]) as well as the dose-dependent data (bottom) containing MCF10A cells treated with different dosage levels of TGF-β after a fixed time period representing near-steady-state. Gene expression of cells from both experiments were measured using single-cell RNA-sequencing and were subsequently used individually and integrated for downstream analyses. **(B)** Projection of dose treatment single-cell expression data using UMAP. The color of individual points indicates the dose of TGF-β treatment from 0 pM (red) to 800 pM (pink). **(C-D)** Contour plots of gene set scores of E (x-axis) and M (y-axis) genes using M PC1 (C) and M PC2 (D) for M-scores. Color indicates the dose of TGF-β as in (B). Circles indicate the mean E- and M-score of samples from each dose point and the associated error bars show the standard deviation (**E-G**) Overlay of nnPCA-derived scores from the first E (E PC1, E), first M (M PC1, F) and second M (M PC2, G) principal components. The color of individuals points indicates the score from low (blue) to high (red). (**H-J**) Overlay of the scaled expression of EMT marker genes CDH1 (an epithelial marker, H), VIM (a mesenchymal marker, I), and FN1 (a highly

expressed mesenchymal gene, J). The color of individual points indicates the Z-score of expression of each gene from low (blue) to high (red). 76
Figure 7: Continuity of integrated single cell dose and time data in low-dimensional projections. (A) Projection of integrated dose and time data using UMAP. Each panel uses the same underlying UMAP for samples, but the coloring of points has different meaning. Left panel: color indicates the origin of the sample from the Days 0, 4 and 8 of the time experiment (red), the Days 0, 1, 2 and 3 of the time experiment (green), or the dose experiment (blue). Middle panel: color indicates the treatment dose of samples from the dose experiment; time samples are masked. Right panel: color indicates the time of treatment for samples from the time experiment, dose samples are masked. **(B-C)** Contour plots of gene set scores of E (x-axis) and M (y-axis) genes using nnPCA for dose (B) and time (C) samples from integrated data. Color indicates the dose of TGF-β treatment from 0 pM (red) to 800 pM (pink) for dose data and time of treatment from 0 days (red) to 8 days (pink) for time data. Circles indicate the mean E- and M-score of samples from each dose point and the associated error bars show the standard deviation. **(D)** Boxplots show the distribution of E (left) and M (right) scores across different dose treatments from integrated data. Color indicates the dose of TGF-β as in (B). **(E)** Boxplots show the distribution of E (left) and M (right) scores across different time treatments from integrated data. Color indicates the time of treatment as in (C). 80
Figure 8: Correlation of non-classical EMT genes with E and M scores in integrated data. (A-B) Correlation of non-EMT gene expression with sample E-scores (x-axis) and M Pc1-scores (y-axis) in dose (A) and time (B) samples from the integrated dataset. Each point represents one of the 15000 most variable genes from the integrated single cell dataset, excluding those genes used to define the E- and M- scores. Blue points indicate genes which are anti-correlated with EMT progression ($R > 0.25$ for E-scores, < -0.25 for M-scores), while red points indicate genes which are correlated with EMT progression ($R < -0.25$ for E-scores, > 0.25 for M-scores). (**C-D**) Correlation of non-EMT gene expression with sample E-scores (x-axis) and M PC2-scores (y-axis) in dose (C) and time (D) samples from the integrated dataset. Labels and colors are defined as in (**A-B**). (**E-F**) The correlation between E-scores with M PC1 (bottom) and M PC2 (top) across different treatments in dose (E) and time (F) data. The color of each cell indicates the strength of anti-correlation from weak (red) to strong (blue). 83
Figure 9: Enrichment of dose and time samples in E- and M-score space. (A) A contour map showing the distributions of time (blue) and dose (red) samples in E- and M-score space. **(B)** Bubble chart showing the enrichment of time samples in different segments of E- and M-score space. Each point represents a segment of E- and M-score space defined by a particular quartile of E-score (x-axis) and M-score (y-axis). The size of the point corresponds to the total number samples in the segment and the color of the point represent the log-odds of time sample enrichment from low (blue) to high (red) with a log-odds of 0 (white) indicating balanced representation. (**C-F**) Bubble charts showing the differential expression of key EMT genes, CDH1 (C), EPCAM (D) VIM (E), and FN1 (F) in different segments of E- and M-score space. Point positions and size are defined as in (B). The color of the point represents the log fold change of expression of the gene from low (blue) to high (red) with a log fold change of 0 (white) indicating no-change relative to other samples. All fold change values are shown regardless of significance. 86
Figure 10: Performance of prognostic models. (A) Bubble chart shows the average performance of Cox hazard models built using the up-regulated genes from different segments of the E- and M-score space. Each point represents a segment defined by a particular quartile of E-

score (x-axis) and M-score (y-axis). The size of the point corresponds to the average number predictor genes across all models built using the up-regulated genes from that segment. Color represents the C-index averaged across all models of cancer datasets, where C-index is a measure of the correspondence between the models predicted risk and survivorship, similar to the Area Under the Receiver Operating Curve (AUC-ROC). Red represents higher C-index (better performance). Black represents lower C-index (poorer) performance. White represents the approximate median of scores (0.645). **(B)** Heatmap of individual model performance across all combinations of segment-gene-derived models (x-axis) and cancer datasets (y-axis). The color of individual cells represents the C-index of the model scaled against all models of that cancer datasets to account for data-set-level differences in model performance. Segments are ordered across the x-axis by the increased average scaled C-index from left to right, while cancer datasets are ordered by increasing number of samples from bottom to top. The shading of segment coordinates along the x-axis distinguishes between the extreme M state (yellow), E-low intermediates (green), M-high intermediates (light blue), and other segments (dark blue). **(C-D)** Bubble charts show the average performance Cox hazard models built using the up-regulated genes identified using only time (C) or dose (D) samples from different segments of E- and M-score space. Point position, size, and color as defined as in (A). 88

Figure 11: Flowchart of the Prognostic Model Development Process. The six-step methodology for constructing our prognostic model is depicted. Step 1: Data Preparation entails the preprocessing of bulk and single-cell RNA sequencing (scRNA-seq) datasets, consistent with methods outlined in our previous publication. Step 2: Gene Prioritization incorporates a tripartite metric approach— Single-cell Differential Expression (SDE), miRNA-target gene interactions, and Median Absolute Deviation (MAD)—to refine an integrated gene set, which is subsequently processed through a LASSO penalized Cox regression model in steps 3 and 4. Step 5: Predictor Optimization involves a backward elimination process tailored to adhere to the event-per-variable ratio of 1:10, thereby minimizing overfitting. This step iteratively excludes genes to maximize the concordance index (c-index), retaining a gene set with optimal predictive power. Step 6: Validation assesses the performance of the identified gene set using 10-fold cross-validation and the Kaplan-Meier risk estimator, ensuring robustness and reliability of the prognostic model. Note: the SDE plot seen next to the UMAP in step 2A is from Campbell et al. 108

Figure 12: Composite Metrics Outperform Individual Measures. We evaluated the performance of individual metrics (M = MAD, S = SDE, miR = miRNA) and compared them to linear combinations of these metrics in (A) a single dataset and (B) 1,000 simulated datasets based on the survival distribution of the dataset used in (A). Combining all three metrics yielded the highest overall concordance index (c-index) in 10-fold cross-validation. To validate the signature identified in (A), we applied it to 1,000 datasets and calculated the mean of the mean 10-fold cross-validation across all 1,000 datasets in (B). We also assessed the performance of the signature in progression-free survival. In this analysis, both SDE and miRNA alone performed as well as the three-metric model. In (D), we demonstrated that the six-gene signature identified by the three-metric model effectively stratifies patients into high-risk and low-risk groups. Significance in (D) was determined using a p-value threshold of < 0.05 using the log-rank test. 114

Figure 13: Validation of miRMS model. In (A) we show the the performance of our signature on the training dataset with the hazard ratios of each of the genes in the model. In (B), we have the top 5 enriched pathways from the Reactome database from the genes in our model. In (C) we

have same but for the enriched pathways from the Wikipathways database. Both pathway enrichment tools reveal that our genes are associated with pathways related to iron metabolism, NOTCH signaling, and various aspects of Gamma signaling. Thes pathways have been shown to play important roles in various cancers. In (D) we performed 10-fold cross validation on the initial dataset and took the mean of this performance for miRMS along with several commenly used diffeential-expression methods. We see that miRMS outperforms all methods. In (E) we perform the same analysis as in (D) but on our 1,000 synthetic datasets to compare the methods to further validate miRMS's performance benefits compared to other methods. Finally, in panel (F) we investigated whether miRMS performance was influenced by sex and saw that it performed well for both sexes but that it has better mean performance for females. This is inline with literature that finds women have better outcomes from SCLC. *** p-value < 0.001. Combined Score is defined as the -log(p-value) * odds ratio. log is the natural logorithm. p-value threshold is < 0.05. n in (F) based on 0.2 effect size at 80% power. 115
Figure 14: Differential Predictive Ability of the miRMS Signature Across Biological & Sociological Variables. Here we employed a four-gene miRMS signature and subjected it to rigorous evaluation using distinct subsets of the initial training dataset. These subsets were generated via simulated sampling based on power calculations. Our primary objective was to scrutinize the signature's predictive performance across the spectrum of the TNM Classification of Malignant Tumors (TNM) standard and other importnat biological or sociological variables. Specifically, the training dataset was meticulously stratified to encompass only those samples that conformed to specific TNM staging criteria. Subsequently, simulated samples were meticulously generated in accordance with the predetermined power calculations. Following this, we conducted a comprehensive assessment of our four-gene signature, leveraging the power of 10,000 iterations of the 10-fold cross-validation technique. (A) When categorizing all patients with stage I or II cancer as "early" and those annotated as stage III or IV as "late," our findings align with expectations, revealing a higher model accuracy for patients with less advanced tumor staging. (B) An exploration of N stages 0, 1, and 2 demonstrated comparable model performance for lymph node stages 0 and 1. However, a noticeable decline in performance was observed when predicting samples characterized by the most advanced lymph node staging. (C) Interestingly, our results exhibit superior predictive performance for samples classified as "metastatic" (1) compared to those categorized as "non-metastatic" (0). This observation may be attributed to the limitations inherent in a binary classification of samples into metastatic or non-metastatic categories. (D) Upon categorizing patients as "old" or "young" based on the median age of the entire cohort, our model displayed marginally enhanced performance for older patients. This phenomenon could be attributed to the typicality of older patients, while younger SCLC patients may present with additional complexities, potentially influencing earlier cancer diagnoses and, consequently, predictive accuracy. (E) Intriguingly, our model exhibited heightened accuracy for patients with no history of smoking. (F) Finally, patients who had not previously undergone chemotherapy were found to be better predicted by our signature. All p-values reported as < 0.05. P-values were calculated employing the Wilcoxon Ranked Sum test (A, C-F) or ANOVA (B). Pairwise p-values were further assessed using Dunn's test (B). The "n" represents the number of samples within each group, as determined by power calculations based on an effect size of 0.2 at 80% power. 116
Figure 15: Comparative Performance of miRMS Against Deep Learning Methods and Across SCLC Subtypes (A) Performance comparison of miRMS with deep learning methods using the top 20 genes identified by our gene prioritization pipeline. miRMS outperforms both

deepSurv (DS) and Random Forest (RF), as well as a set of randomly selected genes (RG). This demonstrates the effectiveness of our gene prioritization approach and the superiority of miRMS even when all methods are optimized. (B) Differential performance of our identified signature across SCLC subtypes, based on the ANPY subtyping system proposed by Rudin et al. The most common subtype (A) shows a predictive accuracy of approximately 0.72. The Y subtype was excluded from analysis due to insufficient sample size (n=2) in our dataset. Statistical significance was assessed using ANOVA followed by Dunn's test for multiple comparisons, with p-values adjusted using the Benjamini-Hochberg method (α = 0.05). 117

LIST OF SUPPLEMENTAL TABLES

Supplemental Table 1. Genes Identified by Model Regulated by miRNA. Genes are either directly or indirectly regulated by miRNAs..... 62

LIST OF SUPPLEMENTAL FIGURES

Supplemental Figure 1. miRNA metric (**A**) Heatmap depicting the interactions between the experimentally validated miRNAs and their targets in READ data. Here we show a small subset of the genes and miRNAs that make up our miRNA metric for the TCGA-COAD dataset. These were acquired by first taking the overlap of the miRNAs from miRMap and dbDEMC 2.0. Then these common miRNAs were submitted to the TargetScan database, and these miRNAs' top N gene targets were returned. The number of times each miRNA interacted with each gene was then calculated by simply increasing the number for a gene by one if it interacted with a particular miRNA. Red is high binding; blue is low binding. Once all miRNA-target pairs were evaluated, the genes were ranked from the greatest number of interactions to the fewest. This score was then normalized to between 0 and 1. (**B**) Example of some of the top gene targets for the COAD dataset. 56
Supplemental Figure 2. Pseudo-temporal ordering analysis. (**A)** UMAP of gene expression of 30 randomly selected epithelial (E) annotated genes (left panel). UMAP of gene expression of 30 randomly selected mesenchymal (M) annotated genes (center panel). UMAP of gene expression of 4 hallmark EMT genes that represent E (CDH1) or M (VIM, VIMP, ZEB1) gene expression states in the Li et al data (right panel). (**B)** UMAP of Li et al data colored by pseudotime values based on the gene expression profile of VIM (left). Histograms of the distributions of the Spearman correlation coefficients between gene expression levels and pseudotime values for all E annotated (center) or M annotated (right) genes. 57
Supplemental Figure 3. LASSO regularization gives the best performance. Heatmaps showing our grid searches at several different values of alpha for (**A)** 0, (**B)** 0.5, and (**C)** 1. We plot the mean 10-fold cv at each combination of miRNA and miRNA target number. 58

Supplemental Figure 4. Determination of Ideal miRNA and miRNA target values. Heatmaps showing our grid searches of many combinations of miRNAs and miRNA targets for (**A)** TCGA-COAD and (**B)** TCGA-READ. We send each combination of miRNAs and miRNA targets through our pipeline, and then we send the ideal number of genes in each integrated list of genes from our model into the LASSO Cox model for TCGA-COAD and TCGA-READ datasets. The gene size for TCGA-COAD is 2,500, and the gene size for TCGA-READ is 600. We then plot the concordance index of each combination. 59
Supplemental Figure 5. CCsc MMS Uses Comparable Gene Set Size to Existing Methods.(A) We ran each method at optimal gene set size and then extracted the non-zero betas from the LASSO penalized Cox model (i.e., Active genes). We see that, on average, CCsc has a much higher concordance index than other methods while using approximately the same number of genes. We also see this in TCGA-READ (**B)** 59
Supplemental Figure 6. CCsc MMS Outperforms other Methods in Non-TCGA CRC Data. We performed the same analysis on a non-TCGA CRC dataset. We found that CCsc MMS continues to outperform its individual components (**A**). Additionally, it outperformed all other methods on this new dataset (**B**). 60
Supplemental Figure 7. Highly Ranked Gene Targets Show Reduced Expression Level. Each of the top performing genes in our signatures showed much reduced expression compared to the mean of all genes present in the (**A**) TCGA-COAD and (**B**) TCGA-READ datasets respectively. This was true for both the top genes that were found to increase patient risk (left panels) and to decrease patient risk (right panels). Horizontal red line denotes the mean expression of each dataset. 61
Supplemental Figure 8. Positioning of unlabeled dose samples in UMAP space.(A) Projection of dose-dependent single cell expression data using UMAP. The color of individual points indicates whether the sample has a treatment level label (black) or is missing a label (red). **(B)** Scatter plots of gene set scores of E (x-axis) and M (y-axis) genes using nnPCA for dose data. Circles indicate the mean E- and M-score of samples from each shared nearest neighbor cluster, with the red dots representing labeled samples only and the blue points representing unlabeled samples. The associated error bars show the standard deviation. 94
Supplemental Figure 9. Location of putative homotypic doublets in UMAP space.(A-B) Projection of labeled dose samples (A) and time samples (B) from integrated expression data using UMAP. The color of individual points indicates the nearest neighbor clusters to which is belong except for black points which indicate putative homotypic doublets detected by DoubletFinder. 94
Supplemental Figure 10. The impact of cell-cycle correction of projection and clustering.(A-B) Projection of dose-dependent single cell expression data using UMAP without (A) and with (B) correction based on cell cycle scores. The color of individual point indicates the inferred cell-cycle state: red = G1, green = G2M, blue = S. 95
Supplemental Figure 11. Clustering based on shared-nearest-neighbor graphs in dose UMAP and E- and M-score space. (A) Projection of dose treatment single cell expression data using UMAP. The color of individual points indicates the clusters based on shared-nearest-neighbor graphs to which the sample was assigned **(B)** Contour plots of gene set scores of E (x-axis) and M (y-axis) genes using nnPCA for dose data. Color indicates cluster of samples in UMAP space, with individual colors corresponding directly to clusters in (A). Circles indicate the mean E- and M-score of samples from each dose point and the associated error bars show the standard deviation. **(C)** Bar charts show the frequency (y-axis) of different treatment dosages (x-

axis) in Cluster 0 (left) and Cluster 2 (right). The frequency of each dosage is broken down by the inferred cell-cycle state: S (blue), G2/M (green) and G1 (red). 96
Supplemental Figure 12. Positioning of samples from different experiments in unintegrated UMAP Space. Projection of unintegrated dose and time data using UMAP. Color indicates the origin of the sample from the Days 0, 4 and 8 of the time experiment (red), the Days 0, 1, 2 and 3 of the time experiment (green), or the dose experiment (blue). 97
Supplemental Figure 13. Overlays of principal component scores on integrated data. (A) Projection of dose treated samples from integrated single-cell samples using UMAP which have been overlaid with scores derived from the principal components (PCs) of integrated data: E PC1 (top), M PC1 (middle), M PC2 (bottom). The color of each point indicates the value of the score from high (red) to low (blue). **(B)** Projection of integrated single-cell samples using UMAP which have been overlaid with scores derived from the components (PCs) of the unintegrated dose data: E PC1 (top), M PC2 (middle), M PC1 (bottom). Colors are as in (A). 98
Supplemental Figure 14. Overlays of principal component scores on dose-dependent data.(A) Projection of dose-dependent single-cell samples using UMAP which have been overlaid with scores derived from the principal components (PCs) of dose-dependent data: E PC1 (top), M PC1 (middle), M PC2 (bottom). The color of each point indicates the value of the score from high (red) to low (blue). **(B)** Projection of dose-dependent single-cell samples using UMAP which have been overlaid with scores derived from the components (PCs) of the integrated data: E PC1 (top), M PC2 (middle), M PC1 (bottom). Colors are as in (A). 99
Supplemental Figure 15. Similarity in nearest neighbor clustering between dose-dependent and integrated data.(A) Projection of dose-dependent single-cell samples using UMAP which have been overlaid with nearest neighbor clusters from dose-dependent (top) and integrated (bottom) data. The color of each point represents the cluster of the sample respective to the legend beneath each figure. **(B)** Projection of dose treated samples from integrated single-cell samples using UMAP which have been overlaid with nearest neighbor clusters from integrated data (top) and dose-dependent data (bottom). The color of each point represents the cluster of the sample respective to the legend beneath each figure. 100
Supplemental Figure 16. Overlap of unintegrated E- and M-scores across time and dose data. Scatter plot of E-scores (x-axis) and M-scores (y-axis) of samples from the Days 0,4,8 of time experiment (red), Days 0,1,2,3 of time experiment (green) and dose experiment (blue) prior to full integration. Note that time samples have been integrated relative to one another as in Deshmukh et al.[13], but have not been integrated with the samples from the dose experiment. 101
Supplemental Figure 17. Clustering based on shared-nearest-neighbor graphs in integrated UMAP and E- and M-score space. (A) (above) Projection of time treatment single cell expression data using UMAP on integrated data. The color of individual points indicates the cluster based on shared-nearest-neighbor graphs to which the sample was assigned. The dotted circles highlight Cluster 1. (below) Bubble charts showing the fraction of samples from Cluster 1 in different segments of E- and M-score space. Each point represents a segment of E- and M-score space defined by a particular quartile of E-score (x-axis) and M-score (y-axis). The size of the point corresponds to the total number samples in the segment and the color of the point represents the fraction of samples from the corresponding cluster from 0 (black) to 0.5 (red). **(B)** (above) Projection of dose treatment single cell expression data using UMAP on integrated data. The color of individual points indicates the cluster based on shared-nearest-neighbor graphs to which the sample was assigned. The dotted circle highlights Cluster 8. (below). Bubble charts

showing the fraction of samples from Cluster 8 in different segments of E- and M-score space. Point positions, size and color as are defined as in (B). 102

CHAPTER I INTRODUCTION

Difficulty of Building Prognostic Models in a Big Data Context

A fundamental challenge in computational biology, particularly in modeling, is the curse of dimensionality. This phenomenon refers to the analytical difficulties that arise when dealing with datasets containing many more predictors (p) than subjects (n) [1]. The big data era has exacerbated this issue, with some experimental designs collecting thousands of predictors for even a single sample. This is especially prevalent in various biological data contexts, such as analyses involving single nucleotide polymorphisms (SNPs), gene expression, or transcription factors. Consequently, identifying which predictors are biologically significant for outcomes of interest - such as disease severity, patient survival, or response to novel therapeutics - has become increasingly complex.

The curse of dimensionality manifests in several specific ways, primarily through multicollinearity, multiple testing issues, overfitting, and data sparsity as articulated by Altman et al. [1-4]. Multicollinearity occurs when many predictors in a model are highly correlated, making it difficult to discern which variables truly drive the prediction of an outcome. This is particularly critical in biological contexts, where predictions can have significant consequences, such as estimating a patient's disease severity and potential lifespan.

Multiple testing problems arise due to running many statistical tests on a small number of samples. By running many tests mathematically, the likelihood that spurious statistically significant results will be found. For example, in differential gene expression analysis comparing thousands of genes between conditions, the likelihood of falsely identifying significant differences increases substantially at standard significance thresholds (e.g., $\alpha = 0.05$). To mitigate this, various correction methods are employed. The Bonferroni correction divides the α threshold by the total number of tests (e.g., 0.05/20,000 genes $\approx 2.5 \times 10^{-6}$ for gene expression studies). Alternatively, the Benjamini-Hochberg procedure controls the false discovery rate by ranking p-values and comparing them to critical values based on the desired FDR level [5].

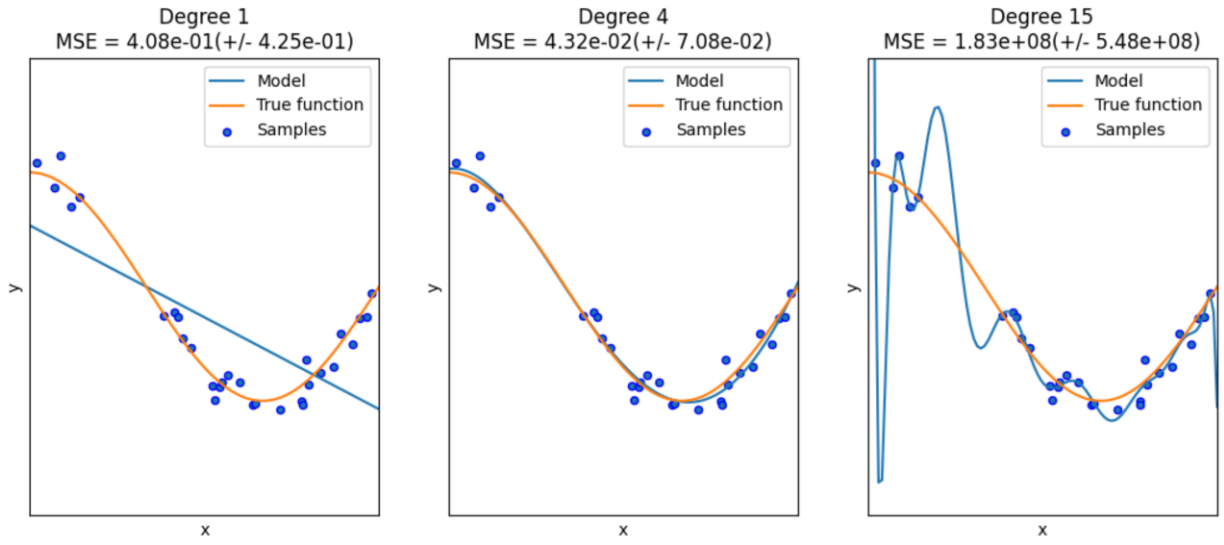


Figure 1: Example of Underfit, Properly fit, and Overfit Models. Here we show a toy example of what a un underfit model (Degree 1), properly fit (Degree 4), and an overfit model (Degree 15). We use a portion of the cosine function (orange line) to show how fitting a polynomial with varying numbers of degrees (teal line) impacts impact how well the model fits the observed data (blue dots).

Overfitting occurs when a model captures not only the true relationships between variables but also the noise inherent in the dataset (Figure 1). This leads to optimistic performance estimates on training data but poor generalization to unseen data. To address this, researchers often partition their data into training and testing sets or employ k-fold cross-validation for more robust model evaluation.

Data sparsity is particularly problematic in single-cell RNA-seq experiments, where high dropout rates result in many zero values. This sparsity complicates the distinction between true biological variation and technical noise [6-8], reduces statistical power, and challenges many statistical tests that rely on distributional assumptions violated by sparse data.

To address these challenges, researchers have developed various prognostic models, including classical statistical methods, machine learning approaches, and deep learning techniques [8-12]. Each type of model offers unique strengths and weaknesses in tackling the curse of dimensionality.

Classical statistical methods provide interpretable results with well-established theoretical foundations. They often require fewer samples to achieve reliable results and offer clear measures of statistical significance and effect sizes [9]. However, these methods may struggle with highly non-linear relationships and can be sensitive to outliers and violations of distributional assumptions [10]. Additionally, they often require manual feature selection, which can be challenging in high-dimensional spaces [11].

Machine learning approaches, on the other hand, can capture complex, non-linear relationships in data and often perform well with high-dimensional datasets [12]. Many of these methods, such as random forests, inherently perform feature selection [13]. Nevertheless, machine learning techniques may require larger sample sizes for optimal performance [14]. Some methods, like support vector machines, can be computationally intensive, and their results can be less interpretable than those of classical statistical methods [15].

Deep learning techniques excel at capturing highly complex patterns in data and can automatically learn features from raw data [16]. They often outperform other methods on very large datasets

[17]. However, these approaches typically require very large sample sizes to achieve optimal performance [18]. Moreover, the results from deep learning models can be difficult to interpret, leading to the so-called black box problem [19]. Deep learning methods are also computationally intensive and may require specialized hardware, such as GPUs, for efficient training and deployment [20].

The choice of method often depends on the specific characteristics of the dataset, the research questions at hand, and the trade-offs between model performance, interpretability, and computational requirements. In practice, researchers may employ a combination of these approaches to leverage their complementary strengths and mitigate their individual weaknesses.

As we transition from discussing the general challenges in prognostic modeling to examining specific approaches across various cancer types, it's important to note how these fundamental issues manifest in real-world applications. The following section provides a comprehensive survey of prognostic models in different cancers, illustrating how researchers have attempted to address the curse of dimensionality and related challenges in diverse oncological contexts.

Survey of Prognostic Models Across Cancers

Prognostic models in cancer have evolved significantly over the past few decades, reflecting advancements in both our understanding of cancer biology and computational methodologies. This evolution has been marked by a shift from reliance on purely clinical factors to the incorporation of molecular and genetic data, highlighting the increasing complexity and precision of cancer prognosis.

Phung et al. conducted an extensive review of 58 prognostic models for breast cancer developed between 1982 and 2016 [21]. Their analysis revealed a predominance of Cox proportional hazard models (32 out of 58), likely due to their ability to handle censored data common in survival analysis. These models primarily utilized established clinical markers such as tumor grade, tumor size, and nodal status. With the growing recognition of HER2's importance, some newer models incorporated HER2 status as a predictor.

The most common performance metric across these studies was the concordance index (c-index), which measures the model's ability to correctly rank patient outcomes [22]. These models, primarily utilizing established clinical markers such as tumor grade, tumor size, and nodal status, face significant limitations. The focus on clinical prognostic markers, while valuable, does not fully capture the potential contributions of genetic biomarkers, which were not widely used in clinical practice at the time of these studies.

In the realm of colorectal cancer (CRC), Smith et al. reviewed 16 prognostic models from 8 publications [23]. These models, focused on predicting 5- and 10-year survival, primarily used clinical and lifestyle factors such as alcohol use, dietary preferences, and body mass index. The models showed good calibration and discriminatory performance (c-index of 0.70-0.71) in large validation cohorts, suggesting their potential utility in population-based screening programs.

However, the landscape of CRC prognostics is evolving with the introduction of molecular markers. Koncina et al. examined common biomarkers like *KRAS*, *BRAF*, and microsatellite instability (MSI) in CRC [24]. Interestingly, they found that *KRAS* and *BRAF* had limited prognostic value in non-metastatic CRC, highlighting the complexity of translating molecular data into clinically relevant prognostic tools.

Ahluwalia et al. advocated for a more comprehensive approach, suggesting that truly powerful prognostic models should integrate multiple types of biomarkers [25]. They proposed that combining multi-gene signatures with diverse data sources, including immune-based information, could significantly improve prognostic accuracy. This multi-omic approach could potentially provide more accurate prognoses beyond traditional staging information, especially for specific patient subgroups and varying tumor severities [26].

The potential of molecular markers is further exemplified in endometrial cancer (EC) research. Coll-de la Rubia et al. identified 255 proteins with prognostic value in EC through a meta-analysis of 398 studies [27]. They highlighted the Cancer Genome Atlas (TCGA) subtyping system for EC, based on *TP53* and *KRAS* mutations and overall mutation load, as a significant advancement. However, their analysis revealed that only a small subset of these potential biomarkers showed consistent association with EC across multiple studies, underscoring the challenges in identifying robust molecular prognostic factors.

In head and neck cancer (HNC), Russo et al. reviewed six studies utilizing Cox models [28]. They emphasized the importance of employing shrinkage techniques like LASSO (Least Absolute Shrinkage and Selection Operator) to enhance model stability and interpretability. LASSO is a regression analysis method that performs both variable selection and regularization to improve the prediction accuracy and interpretability of the resulting statistical model. This recommendation reflects a growing recognition of the need for more sophisticated statistical approaches in handling high-dimensional data typical in molecular studies.

Prostate cancer research has also seen advancements in prognostic modeling. Wang et al. developed a five-gene signature using univariate Cox models, achieving moderate performance

(c-index 0.60-0.71) across multiple cohorts [29]. Zhang et al. took a more comprehensive approach, integrating genomic markers, SNPs, and CNVs into a multi-gene Cox model [30]. Their model showed improved performance, achieving an Area Under the Curve (AUC) of 0.78, and could stratify patients into distinct prognostic subtypes. The AUC refers to the area under the Receiver Operating Characteristic (ROC) curve, which is a graphical plot illustrating the diagnostic ability of a binary classifier system as its discrimination threshold is varied. An AUC of 0.78 indicates good discriminatory power, with the model correctly classifying a randomly chosen positive instance higher than a randomly chosen negative one 78% of the time.

Similar trends are observed in small cell lung cancer (SCLC) research, where prognostic models are increasingly incorporating molecular data alongside clinical factors [31, 32]. Despite these advancements, a fundamental limitation persists across many of these models: they often focus on a single type of predictive information, be it gene expression, clinical factors, or histological features [33-39]. This narrow focus inherently limits the models' predictive power and fails to capitalize on the wealth of diverse data now available in cancer research.

As we move from examining prognostic models across various cancer types to focusing on the unique challenges presented by rare cancers, it becomes evident that the limitations of single data type approaches are particularly pronounced in these contexts. The next section delves into the specific difficulties encountered when developing prognostic models for rare cancers, using small cell lung cancer (SCLC) as a prime example.

Challenges of Prognostic Models in Rare Cancers

While abundant genomic data exists for common cancers like colorectal cancer (CRC), rare cancers such as small cell lung cancer (SCLC) present unique challenges in prognostic model development due to limited available datasets. Despite receiving relatively less attention, SCLC is a significant health concern, comprising approximately 15% of all new lung cancer cases in the United States, with an alarming 80% of cases already metastasized at the time of diagnosis [40]. This high rate of metastasis underscores the critical need for accurate prognostic models in SCLC, despite the inherent difficulties in their development.

The scarcity of genomic data with associated survival information for SCLC significantly complicates the construction of robust prognostic models. This limitation constrains researchers to select fewer predictors to avoid overfitting, a particular concern given the high-dimensional nature of genomic data. The impact of data paucity is especially pronounced for deep learning models, which typically require substantially larger datasets compared to classical statistical models. This disparity in data requirements often forces researchers to rely on simpler models for rare cancers, potentially sacrificing the predictive power that more complex models might offer with larger datasets.

To address these challenges, researchers have turned to advanced statistical methods, particularly synthetic data generation techniques. Bootstrapping has emerged as a robust approach to augment limited datasets while maintaining the underlying distributions of survival times and genomic profiles [41]. This method involves randomly sampling the existing data with replacement, effectively scaling up small datasets. When combined with k-fold cross-validation, bootstrapping provides a solid foundation for assessing model performance, even in the context of limited data.

However, while synthetic data generation offers a partial solution, it is not without limitations. Synthetic data may not fully capture the complexity and nuances of real patient data, particularly in the context of rare and aggressive cancers like SCLC. Therefore, there remains a pressing need for innovative approaches that can maximize the utility of limited datasets without compromising the validity of prognostic predictions.

The challenges in developing prognostic models for rare cancers like SCLC extend beyond mere data limitations. They also encompass issues of model interpretability, generalizability across diverse patient populations, and the integration of multiple data types (genomic, clinical, and environmental factors). Addressing these challenges requires a multifaceted approach, combining advanced statistical techniques with domain-specific biological knowledge.

As the field progresses, there is growing interest in developing hybrid models that can leverage the strengths of both classical statistical approaches and modern machine learning techniques. Such models aim to balance the need for robust performance on limited datasets with the potential for capturing complex, non-linear relationships in cancer progression.

Having explored the challenges specific to rare cancers, we now turn our attention to a crucial biological process that plays a significant role in cancer progression and has implications for prognostic modeling: Epithelial-Mesenchymal Transition (EMT). The following section examines the background of EMT and its relevance in developing more accurate and comprehensive prognostic models.

Epithelial-mesenchymal transition and Its Use in Prognostic Models

Epithelial-mesenchymal transition (EMT) is a dynamic process underlying the progression of many cancers. This biological phenomenon involves the transformation of epithelial cells into mesenchymal-like cells along a continuum of states [42]. During EMT, epithelial cells lose their characteristic features, including tight junctions, adherens junctions, and apical-basal polarity, and acquire mesenchymal traits such as elongated morphology and enhanced motility. This transition enables cells from primary tumors to metastasize to distant locations, making EMT a critical target for cancer therapeutics.

The EMT process is initiated by well-defined molecular triggers, with transforming growth factor beta (TGF-β) being a primary inducer. It is regulated by a complex network of genes and transcription factors, including *ZEB1* and *SNAIL1*, which orchestrate the phenotypic changes. Beyond its role in cancer, EMT plays crucial functions in embryonic development, wound healing, and fibrosis [43, 44].

Extensive research has focused on linking the spectrum of EMT states to cancer severity and progression. Numerous studies have correlated the expression of specific EMT states and associated genes with poor outcomes across various cancer types [45-47]. This has led to the incorporation of EMT-related information into prognostic scores, which have shown promise in predicting various cancer outcomes.

However, significant challenges remain in fully elucidating the relationship between the intricate partial states of EMT and patient survival outcomes. Current prognostic models often rely on a single data type as predictors, limiting their ability to capture the complexity of EMT and its impact

on cancer progression. This limitation is particularly pronounced in cancers such as small cell lung cancer (SCLC), where high-throughput data with associated survival outcomes are scarce.

The development of more sophisticated prognostic models that can integrate multiple data types and accurately represent the EMT spectrum is crucial. Such models could potentially offer improved discriminatory power in predicting patient outcomes and guide more personalized treatment strategies. Addressing these challenges requires innovative approaches that can effectively synthesize diverse data sources and overcome the limitations of current prognostic modeling techniques in cancer research.

Thesis Objectives and Approach

This thesis addresses critical challenges in cancer prognostics by developing an innovative approach to integrate multiple types of omics data for constructing relevant prognostic signatures across various cancer types. The primary objectives of this research are:

1. To demonstrate that multi-omic modeling approaches are superior to single data type approaches in developing prognostic models for cancer.
2. To establish that gene prioritization and ranking methods are superior to random gene selection and are crucial for identifying interpretable signatures.
3. To establish that gene prioritization and ranking methods are superior to commonly used methods of biomarker discovery.

Central to this work is the development of methods to prioritize and rank genes, effectively tackling the curse of dimensionality inherent in large-scale genomic datasets. By leveraging diverse data

sources, including those related to EMT states, this research aims to generate interpretable and powerful signatures capable of predicting patient survival outcomes with high accuracy.

The approach is designed to be particularly valuable in contexts where high-throughput data with associated survival outcomes are scarce, such as in rare cancers like small cell lung cancer (SCLC). Through this integrated, multi-omics strategy, this thesis seeks to advance the field of cancer prognostics, offering a more comprehensive and nuanced understanding of the molecular factors influencing patient survival.

References

1. Altman, N. and M. Krzywinski, *The curse (s) of dimensionality.* Nat Methods, 2018. **15**(6): p. 399-400.
2. Daoud, J.I. *Multicollinearity and regression analysis*. in *Journal of Physics: Conference Series*. 2017. IOP Publishing.
3. Noble, W.S., *How does multiple testing correction work?* Nature biotechnology, 2009. **27**(12): p. 1135-1137.
4. Lever, J., M. Krzywinski, and N. Altman, *Points of significance: model selection and overfitting.* Nature methods, 2016. **13**(9): p. 703-705.
5. Benjamini, Y. and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* Journal of the Royal statistical society: series B (Methodological), 1995. **57**(1): p. 289-300.
6. Stegle, O., S.A. Teichmann, and J.C. Marioni, *Computational and analytical challenges in single-cell transcriptomics.* Nature Reviews Genetics, 2015. **16**(3): p. 133-145.
7. Grün, D., L. Kester, and A. Van Oudenaarden, *Validation of noise models for single-cell transcriptomics.* Nature methods, 2014. **11**(6): p. 637-640.
8. Van Dijk, D., et al., *Recovering gene interactions from single-cell data using data diffusion.* Cell, 2018. **174**(3): p. 716-729. e27.
9. Harrell, F.E., *Regression modeling strategies.* R package version, 2012: p. 6.2-0.
10. Hastie, T., et al., *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. 2009: Springer.
11. Guyon, I. and A. Elisseeff, *An introduction to variable and feature selection.* Journal of machine learning research, 2003. **3**(Mar): p. 1157-1182.
12. Breiman, L., *Statistical modeling: The two cultures (with comments and a rejoinder by the author).* Statistical science, 2001. **16**(3): p. 199-231.
13. Díaz-Uriarte, R. and S. Alvarez de Andrés, *Gene selection and classification of microarray data using random forest.* BMC bioinformatics, 2006. **7**: p. 1-13.
14. Figueroa, R.L., et al., *Predicting sample size required for classification performance.* BMC medical informatics and decision making, 2012. **12**: p. 1-10.

15. Burges, C.J., *A tutorial on support vector machines for pattern recognition.* Data mining and knowledge discovery, 1998. **2**(2): p. 121-167.
16. LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning.* nature, 2015. **521**(7553): p. 436-444.
17. Esteva, A., et al., *Dermatologist-level classification of skin cancer with deep neural networks.* nature, 2017. **542**(7639): p. 115-118.
18. Sun, C., et al. *Revisiting unreasonable effectiveness of data in deep learning era*. in *Proceedings of the IEEE international conference on computer vision*. 2017.
19. Lipton, Z.C., *The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.* Queue, 2018. **16**(3): p. 31-57.
20. Raina, R., A. Madhavan, and A.Y. Ng. *Large-scale deep unsupervised learning using graphics processors*. in *Proceedings of the 26th annual international conference on machine learning*. 2009.
21. Phung, M.T., S. Tin Tin, and J.M. Elwood, *Prognostic models for breast cancer: a systematic review.* BMC cancer, 2019. **19**: p. 1-18.
22. Uno, H., et al., *On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data.* Statistics in medicine, 2011. **30**(10): p. 1105-1117.
23. Smith, T., et al., *Comparison of prognostic models to predict the occurrence of colorectal cancer in asymptomatic individuals: a systematic literature review and external validation in the EPIC and UK Biobank prospective cohort studies.* Gut, 2019. **68**(4): p. 672-683.
24. Koncina, E., et al., *Prognostic and predictive molecular biomarkers for colorectal cancer: updates and challenges.* Cancers, 2020. **12**(2): p. 319.
25. Ahluwalia, P., R. Kolhe, and G.K. Gahlay, *The clinical relevance of gene expression based prognostic signatures in colorectal cancer.* Biochimica et Biophysica Acta (BBA)-Reviews on Cancer, 2021. **1875**(2): p. 188513.
26. Kuipers, E.J., et al., *Colorectal cancer.* Nature reviews. Disease primers, 2015. **1**: p. 15065.
27. Coll-de la Rubia, E., et al., *Prognostic biomarkers in endometrial cancer: a systematic review and meta-analysis.* Journal of clinical medicine, 2020. **9**(6): p. 1900.
28. Russo, D., et al., *Development and validation of prognostic models for oral squamous cell carcinoma: a systematic review and appraisal of the literature.* Cancers, 2021. **13**(22): p. 5755.
29. Wang, Y., et al., *Identification of a robust five-gene risk model in prostate cancer: a robust likelihood-based survival analysis.* International journal of genomics, 2020. **2020**(1): p. 1097602.
30. Zhang, P., et al., *Identification and validation of a novel anoikis-related prognostic model for prostate cancer.* Molecular Genetics & Genomic Medicine, 2024. **12**(4): p. e2419.
31. Nagy, Á., G. Munkácsy, and B. Győrffy, *Pancancer survival analysis of cancer hallmark genes.* Scientific reports, 2021. **11**(1): p. 6047.
32. Jones, G.S., et al., *A systematic review of survival following anti-cancer treatment for small cell lung cancer.* Lung Cancer, 2020. **141**: p. 44-55.
33. Sardo, E., et al., *Multi-omic approaches in colorectal cancer beyond genomic data.* Journal of Personalized Medicine, 2022. **12**(2): p. 128.
34. Lu, C., et al., *A prognostic model for overall survival of patients with early-stage non-small cell lung cancer: a multicentre, retrospective study.* The Lancet. Digital health, 2020. **2**(11): p. e594-e606.
35. Li, J., et al., *A novel six-gene-based prognostic model predicts survival and clinical risk score for gastric cancer.* Frontiers in Genetics, 2021. **12**: p. 615834.

36. Zhou, W., et al., *Development of a Ten-lncRNA Signature Prognostic Model for Breast Cancer Survival: A Study with the TCGA Database.* Analytical Cellular Pathology, 2020. **2020**(1): p. 6827057.
37. Hu, D., et al., *Development of an autophagy-related gene expression signature for prognosis prediction in prostate cancer patients.* Journal of Translational Medicine, 2020. **18**: p. 1-12.
38. Bruni, D., H.K. Angell, and J. Galon, *The immune contexture and Immunoscore in cancer prognosis and therapeutic efficacy.* Nature Reviews Cancer, 2020. **20**(11): p. 662-680.
39. Jiang, L.-h., et al., *Predictive risk-scoring model for central lymph node metastasis and predictors of recurrence in papillary thyroid carcinoma.* Scientific Reports, 2020. **10**(1): p. 710.
40. Jain, A.S., et al., *Everything old is new again: Drug repurposing approach for non-small cell lung cancer targeting MAPK signaling pathway.* Frontiers in Oncology, 2021. **11**: p. 741326.
41. Iba, K., et al., *Re-evaluation of the comparative effectiveness of bootstrap-based optimism correction methods in the development of multivariable clinical prediction models.* BMC Medical Research Methodology, 2021. **21**: p. 1-14.
42. Pastushenko, I. and C. Blanpain, *EMT Transition States during Tumor Progression and Metastasis.* Trends in Cell Biology, 2019. **29**(3): p. 212-226.
43. Lamouille, S., J. Xu, and R. Derynck, *Molecular mechanisms of epithelial–mesenchymal transition.* Nature reviews Molecular cell biology, 2014. **15**(3): p. 178-196.
44. Hao, Y., D. Baker, and P. Ten Dijke, *TGF-β-mediated epithelial-mesenchymal transition and cancer metastasis.* International journal of molecular sciences, 2019. **20**(11): p. 2767.
45. Gibbons, D.L. and C.J. Creighton, *Pan-cancer survey of epithelial–mesenchymal transition markers across the Cancer Genome Atlas.* Developmental Dynamics, 2018. **247**(3): p. 555-564.
46. Grosse-Wilde, A., et al., *Stemness of the hybrid epithelial/mesenchymal state in breast cancer and its association with poor survival.* PLoS ONE, 2015. **10**(5).
47. Karacosta, L.G., et al., *Mapping lung cancer epithelial-mesenchymal transition states and trajectories with single-cell resolution.* Nature Communications, 2019. **10**(1).

CHAPTER II
USING SINGLE-CELL RNA SEQUENCING AND MICRORNA TARGETING DATA TO IMPROVE COLORECTAL CANCER SURVIVAL PREDICTION

A version of this chapter was originally published by Andrew Willems, Nicholas Panchy, and Tian Hong:

Willems A, Panchy N, Hong T. Using Single-Cell RNA Sequencing and MicroRNA Targeting Data to Improve Colorectal Cancer Survival Prediction. *Cells*. 2023; 12(2):228. https://doi.org/10.3390/cells12020228

My contributions to this work include: (1) conceptualization, (2) methodology, (3) software, (4) validation, (5) investigation, and (6) writing and editing the manuscript.

Abstract

Colorectal cancer has proven to be difficult to treat as it is the second leading cause of cancer death for both men and women worldwide. Recent work has shown the importance of microRNA (miRNA) in the progression and metastasis of colorectal cancer. Here, we develop a metric based on miRNA-gene target interactions, previously validated to be associated with colorectal cancer. We use this metric with a regularized Cox model to produce a small set of top-performing genes related to colon cancer. We show that using the miRNA metric and a Cox model led to a meaningful improvement in colon cancer survival prediction and correct patient risk stratification. We show that our approach outperforms existing methods and that the top genes identified by our process are implicated in NOTCH3 signaling and general metabolism pathways, which are essential to colon cancer progression.

Introduction

Colorectal cancer (CRC) is estimated to develop in 945,000 patients worldwide yearly, with approximately 492,000 dying from the disease. Its complexity leads to challenges with interpreting the relationship between the many inputs of classical statistical models and their outputs related to large datasets. Current work in predicting the prognosis of CRC, using gene expression data, has focused on identifying novel biomarkers to predict survival and treatment outcomes through differential gene expression [1, 2]. These approaches have been developed to work with next-generation bulk RNA sequencing (RNA-seq) [3-5], a technique that only generates the averages of gene expression across cells. Models that prioritize marker genes, based on bulk RNA-seq, are, therefore, unable to describe and utilize widely observed cancer cell heterogeneity for prognostic predictions [6-8]. In addition to the limitation in data, many prognostic models for colorectal

cancer use unregularized Cox models to perform survival analysis. This can lead to issues with overfitting to training data and challenges with including additional inputs to a model.

Controlling gene expression at the post-transcriptional level is not only crucial for cancer progression, but also potentially useful for developing anticancer therapeutics. MicroRNAs (miRNAs) are small (22nts), post-transcriptional regulators of messenger RNA. They have been studied in various systems and have recently been found to play essential roles in cancer regulation and progression [9-12]. While there are a variety of approaches to identifying biomarkers across different cancers, it remains challenging to integrate miRNA target information with RNA-sequencing data, primarily containing mRNA transcripts for prognostic predictions of CRC.

In this work, we use advances in single-cell RNA sequencing (scRNA-seq) and miRNA targeting to improve the prediction of survival outcomes for CRC patients. We develop an integrated gene prioritization method that combines miRNA–mRNA binding and target expression data (Figure 1). We show that using miRNAs and scRNA-seq provides better a predictive performance than other methods. Additionally, our method determined markers that have previously been found to be associated with other types of cancer. Finally, we show that our method can be used on other large cancer datasets to potentially find novel biomarkers and improve survival prediction accuracy.

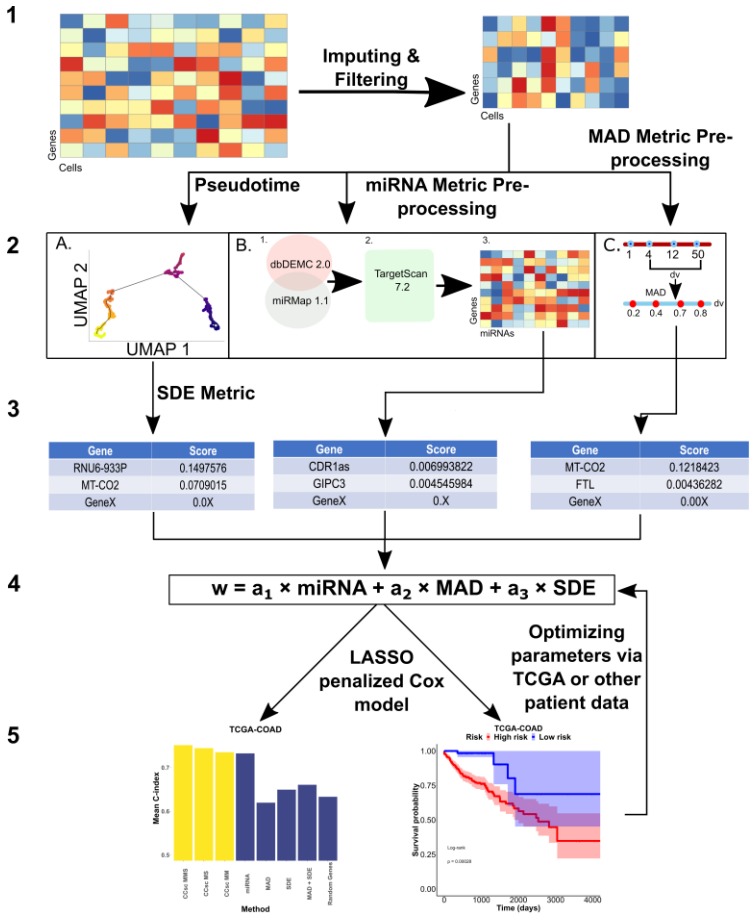


Figure 2: Overall Design of CCsc. CCsc is composed of 5 steps. (**1**) Obtaining pre-processed scRNA-seq matrix (genes by cells) and performing imputation and denoising via the MAGIC package and then filtering for low abundance samples. (**2**) Using filtered, denoised, and imputed scRNA-seq matrix to perform metric dependent pre-processing steps. (A) For the SDE metric this involves ordering the scRNA-seq matrix by pseudotime based on the expression of the VIM marker. (B) For the miRNA metric, this involves sending the matrix to our novel miRNA metric. (C) We calculate the MAD metric from the scRNA-seq matrix. (**3**) In this step, we generate a distinct gene list from each of our three metrics. (**4**) We integrated and merged the three separate gene lists into a single master list based on their score. From this master list, we then send N number of genes from the list as inputs to a LASSO penalized Cox model. The metric weights (a1, a2 and a3) are optimized via a grid search that involves the next step. (**5**) This step yields the outputs of our penalized Cox model that uses concordance index to assess our model's accuracy. We also generate Kaplan–Meier survival curves, based on the model output, to further validate the efficacy of the genes selected by our model. Based on the Cox model outputs, we can optimize the gene weights of our linear model to achieve the best performance.

Results

CCsc has three metrics to prioritize genes for prognostic predictions: miRNA (based on disease-related miRNAs), MAD (based on variability of gene expression), and SDE (based on switch-like behaviors of gene expression) (see Materials and Methods for details). To evaluate these three metrics, we tested multiple versions of CCsc. They include CCsc miRNA + MAD (CCsc

MM), CCsc miRNA + SDE (CCsc MS), and CCsc miRNA + MAD + SDE (CCsc MMS) (Figure 2A, B). To prioritize genes, we used a recent scRNA-seq dataset for colorectal cancer cells [13], and we used TCGA-COAD (Colon Adenocarcinoma) and -READ (Rectum adenocarcinoma) datasets for prognostic performance evaluations (see Methods for details). We compared each combination's mean 10-fold cross-validated C-index, while holding the number of genes constant. We did this for both TCGA-COAD and TCGA-READ. Based on our test results (Figure 7 and Figure 8 Appendix), we concluded that combining all three metrics gave us the best mean concordance index performance across both datasets. In addition, we found that CCsc MMS could separate high-risk from low-risk patients for both the TCGA-COAD and TCGA-READ datasets (Figure 2C, D). Taken together, the combinations of all three metrics (i.e., CCsc MMS) perform better than other choices, suggesting the importance of prioritizing genes based on both miRNA-targeting information and the summary statistics of expression. We therefore used CCsc MMS for our subsequent analyses.

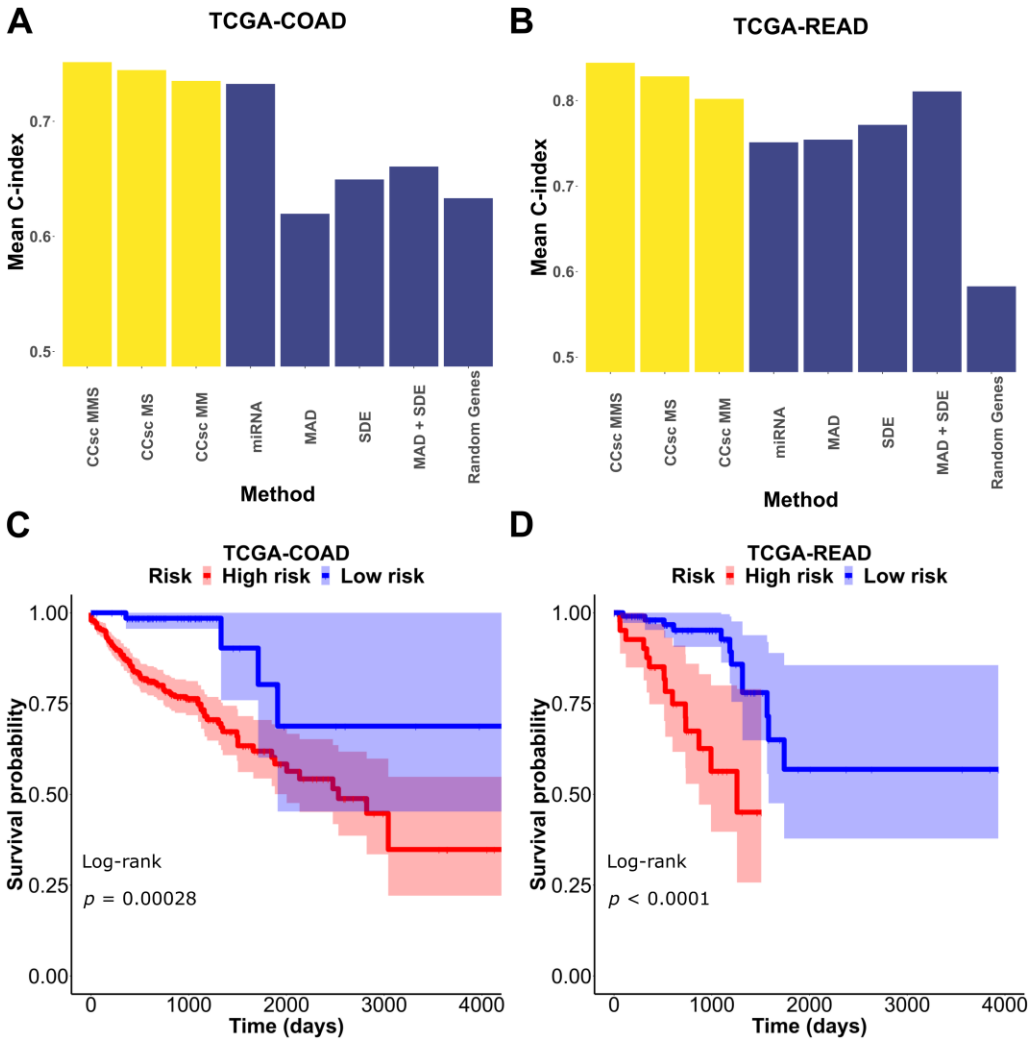


Figure 3: Combination of metrics is better than any individual metric.(**A**) We tested each of our models' metrics and compared them to CCsc MS, CCsc MM, and CCsc MMS and a set of randomly selected genes, equal to the number of genes used by our method on the TCGA-COAD dataset. (**B**) We made the same comparison on the TCGA-READ dataset. We show that integrating the lists of ranked genes from each metric provides better performance than any one individually. (**C**) Kaplan–Meier estimate based on our model's top set of predictors for the TCGA-COAD dataset. (**D**) Kaplan–Meier estimate based on our model's top set of predictors for the TCGA-READ dataset. The shaded regions represent 95% confidence intervals of the survival estimates. The p-value threshold for significance is < 0.05.

To further evaluate the importance of miRNA-targeting in gene prioritization, we compared the mean 10-fold cross-validation performance of CCsc MMS to that of several other methods that select genes based on differential expression. For this comparison, we chose two methods initially designed for bulk RNA-seq analysis, but that have been updated to use single-cell data (DESeq2 and edgeR), and two methods designed specifically for scRNA-seq data (scDD and DEsingle). For both TCGA-COAD and TCGA-READ, CCsc MMS has the best mean 10-fold performance

compared to all other methods. In the case of COAD, CCsc has a markedly better performance (0.7514 vs. ~0.65 for all other methods) and a slightly higher performance with READ (0.8442 vs. 0.8023 for scDD) (Figure 3).

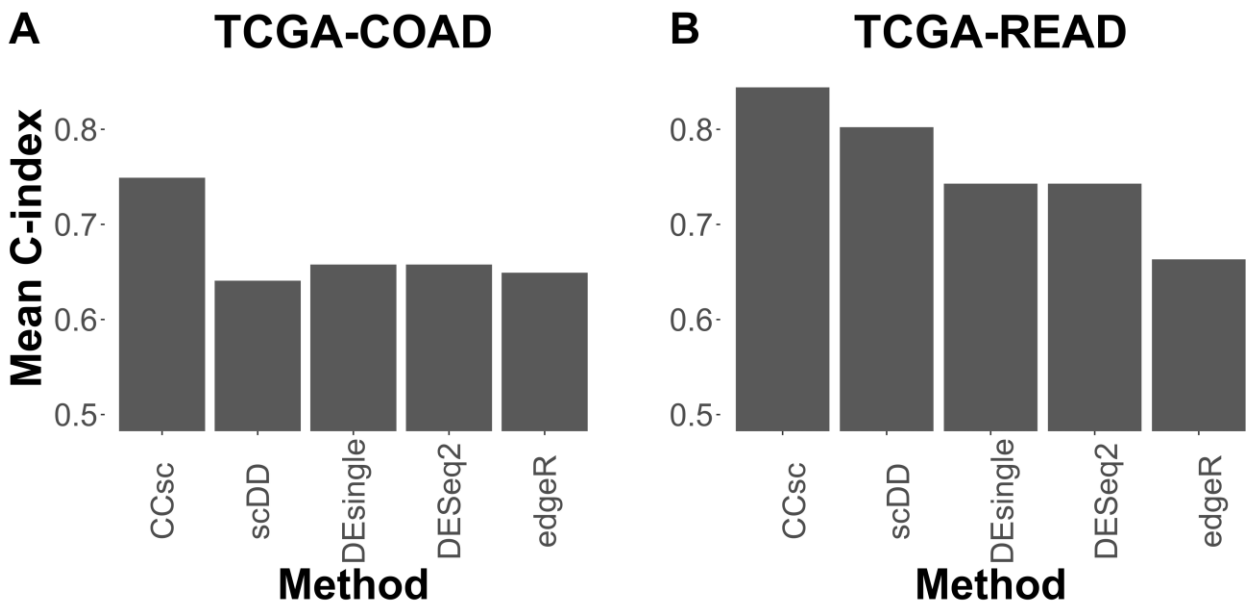


Figure 4: CCsc MMS Outperforms Other Methods. (A) We tested CCsc against several well-established tools on the TCGA-COAD dataset. We compared the mean concordance index performance of CCsc with its ideal weighting to DESeq2, edgeR, scDD, and DEsingle. We gave each method its optimal number of genes and ran each with its recommended settings, according to their respective best practices. (B) Same comparison but on the TCGA-READ dataset. We show that CCsc outperforms single-cell methods (scDD and DEsingle) and that bulk RNA-seq methods can be optimized for single-cell RNA-seq data (DESeq2, edgeR).

We then asked what the top predictors in the model were. We found 13 genes that had a hazard ratio > 2 for COAD, and 16 such genes for READ. Next, we found 12 genes that had a hazard ratio < 0.5 for COAD, and 19 genes for READ (Figure 4A, B). Many of the genes that increase patient risk (12/13 COAD and 11/16 READ) were previously implicated in cancer progression (Tables 1 and 2).

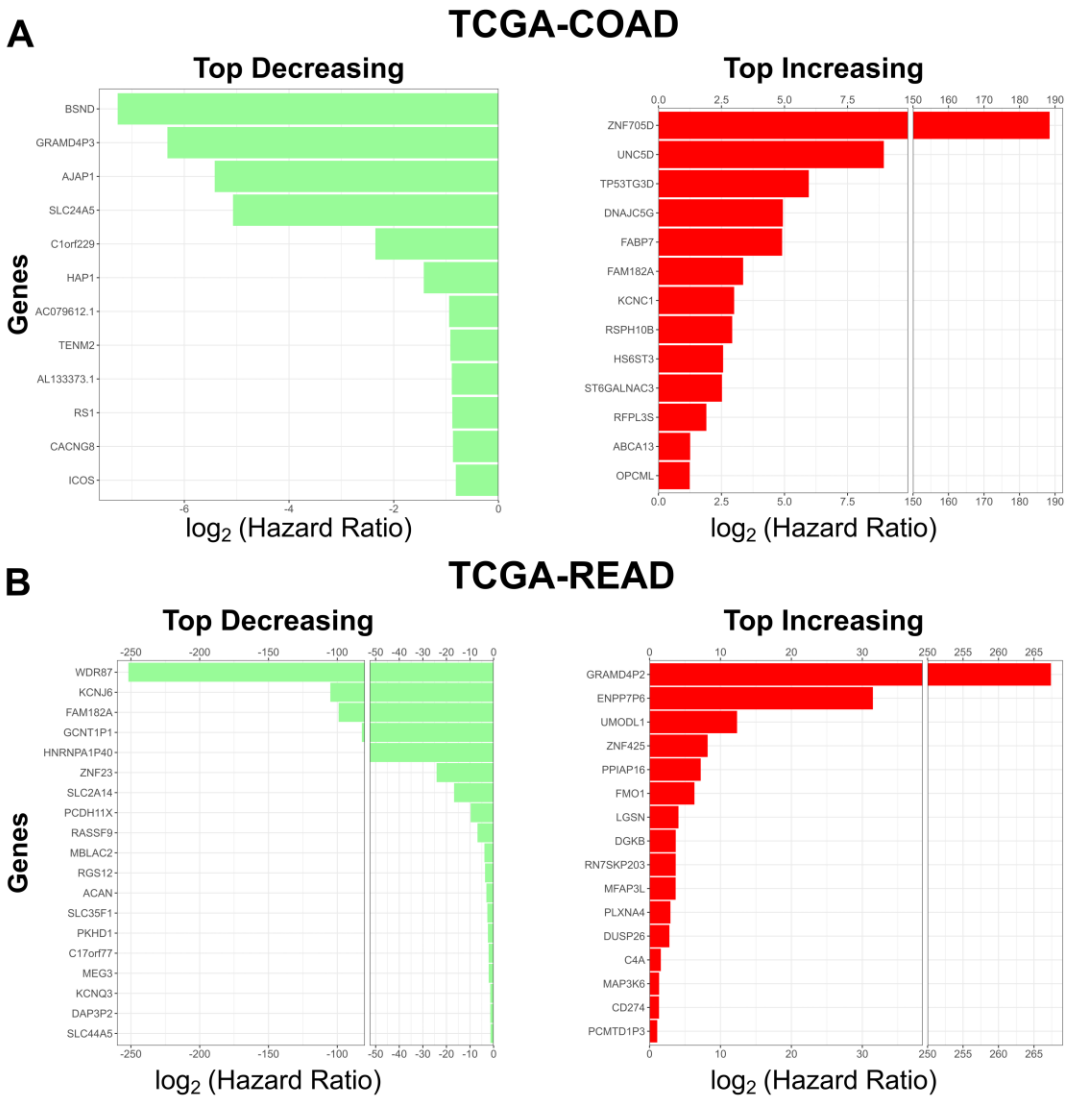


Figure 5: Top coefficients identified by our model. (**A**) Top risk decreasing (green, left) and top risk increasing (red, right) genes identified by our Cox model in TCGA-COAD. (**B**) Top risk decreasing genes (green, left) and top risk increasing (red, right) genes identified by our Cox model in TCGA-READ. For risk increasing genes the hazard ratio threshold is > 2. For risk decreasing genes the hazard ratio threshold is < 0.5.

Table 1. Top Risk Increasing Genes Identified by TCGA-COAD Cox model: List of the top risk increasing genes identified by our Cox model on the TCGA-COAD dataset

Gene	Literature Support
ZNF705D	[14]
UNC5D	[15]
TP53TG3D	[16, 17]
ST6GALNAC3	[18, 19]
RSPH10B	[20]
KCNC1	[21, 22]
HS6ST3	[23, 24]
FAM182A	None
FABP7	[25, 26]

DNAJC5G	[27]
ABCA13	[28]
OPCML	[29]
RFPL3S	[30]

Table 2. Top Risk Increasing Genes Identified by TCGA-READ Cox model: List of the top risk increasing genes identified by our Cox model on the TCGA-READ dataset.

Gene	**Literature Support**
ZNF425	[31]
UMODL1	[32]
RN7SKP203	Pseudogene
PPIAP16	Pseudogene
PLXNA4	[33, 34]
MFAP3L	[35, 36]
LGSN	[37]
GRAMD4P2	Pseudogene
FMO1	[38]
ENPP7P6	Pseudogene
DUSP26	[39, 40]
DGKB	[41, 42]
MAP3K6	[43]
CD274	[44]
C4A	[45]
PCMTD1P3	Pseudogene

Genes such as *OPCML* have been found to be silenced in tumors, and when reactivated, they lead to cancer tumor inhibition. Several of the genes found to increase patient risk in CRC by our model have been found to be silenced in other types of cancers. Additionally, we examined the genes with the smallest hazard ratios to see which of our model's genes might be indictive of better clinical outcomes. We found several genes that were associated with a substantial reduction in patient risk and have been associated with a decreased patient risk in various cancers (Tables 3 and 4). Of the top predictors in READ, without the support of the literature, were all annotated as pseudogenes. However, given the large hazard ratio of these genes in our model, our results suggest that it either plays a role in the pathological process or serves as a signal for cellular changes that lead to cancer progression.

Table 3. Top Risk Decreasing Genes Identified by TCGA-COAD Cox model: List of the top risk decreasing genes identified by our Cox model on the TCGA-COAD dataset.

Gene	Literature Support
AJAP1	[46]
SLC24A5	None (potassium-dependent sodium/calcium exchanger)
CACNG8	[47]
C1orf229	None
GRAMD4P3	Pseudogene
ICOS	[48]
HAP1	[49]
TENM2	[50]
AC079612.1	[51]
AL133373.1	None
BSND	[52]
RS1	[53]

Table 4. Top Risk Decreasing Genes Identified by TCGA-READ Cox model: List of the top risk decreasing genes identified by our Cox model on the TCGA-READ dataset.

Gene	Literature Support
FAM182A	None
SLC35F1	[54]
RGS12	[55]
PKHD1	[56]
GCNT1P1	Pseudogene
KCNJ6	None (potassium channel)
RASSF9	[57]
DAP3P2	Pseudogene
WDR87	None
KCNQ3	[58]
PCDH11X	[59]
MEG3	[60]
MBLAC2	[61]
SLC44A5	None (Predicted to enable transmembrane transporter activity)
HNRNPA1P40	Pseudogene
ZNF23	[62]
ACAN	[63]
SLC2A14	[64]

Next, we sought to see what pathways and cellular processes were influenced by these genes. We submitted the genes from Figure 4 (10 for COAD and 13 for READ) separately to Reactome version 3.7 (https://reactome.org) (Accessed 1 December 2022), and identified pathways related

to *NOTCH3* signaling and Flavin-containing monooxygenases (FMO) oxidizing nucleophiles (Table 5 and 6). The pathways impacted by the top genes have been implicated in CRC progression and metastasis [65-70].

Table 5. Statistically Significantly Enriched COAD Pathways: Most enriched pathways influenced by the genes with the largest hazard ratios associated with increased risk in our TCGA-COAD Cox model. P-value threshold < 0.05 FDR. Hazard ratio threshold > 2

Pathway	Literature Support	FDR P-value
NOTCH3 Intracellular Domain Regulates Transcription	[65]	1.98×10^{-2}
Voltage gated Potassium channels	[66]	1.98×10^{-2}
Signaling by NOTCH3	[67]	2.62×10^{-2}

Table 6. Statistically Significantly Enriched READ Pathways: Most enriched pathways influenced by the genes with the largest hazard ratios are associated with increased risk in our TCGA-READ Cox model. P-value threshold < 0.05 FDR. Hazard ratio threshold > 2.

Pathway	Literature Support	FDR P-value
Activation of C3 and C5	[71]	1.48×10^{-3}
STAT3 nuclear events downstream of ALK signaling	[72]	4.74×10^{-3}
Signaling by ALK	[73]	1.72×10^{-2}
FOXO-mediated transcription of oxidative stress, metabolic and neuronal genes	[69]	1.72×10^{-2}

Discussion

CRC has proven to be a complex disease that, despite the marked research focus and improvements in biomarker detection, still has many open questions. Based on the important roles that miRNAs play in the regulation of many cellular processes, including processes related to CRC, we developed a novel metric that uses the prevalence of miRNA-target interactions to prioritize genes for prognostic models. In conjunction with the switchde and MAD methods, we create an integrated model, CCsc MMS, to improve the ability to accurately predict colorectal patient

survival. We show this performance improvement across multiple large datasets related to colon and rectal cancer. We show that our model outperforms multiple existing methods, including DESeq2, DEsingle, scDD, and edgeR, which have been developed for identifying differentially expressed genes with both bulk and single-cell RNA seq data. The improvement is facilitated by the incorporation of the LASSO regularization. This regularization allows the simplification of the model and avoids overfitting. When we examined the weight of these genes, we found that many of our model's top performers are implicated in various types of cancers, including CRC. We then performed pathway analysis with Reactome and found that this handful of genes is enriched in pathways related to NOTCH3 signaling and potassium channels, which are important in CRC. We attempted pathway enrichment analysis for both the top genes associated with higher patient risk and those that were associated with lower patient risk. The top markers in the lower risk analysis for both TCGA-COAD and TCGA-READ did not meet our significance criteria, and hence no pathways were identified for the lower patient risk genes. In addition, we found that our method uses comparable numbers of active genes to those from these existing methods, while giving better performance (Figure S5). We also found that our approach had a satisfactory performance for non-TCGA CRC datasets (Figure S6). Finally, we asked if any of the top predictors were known to be regulated by miRNAs. We found that many of the genes most impactful on patient survival are directly or indirectly regulated by miRNAs in disease settings (Table S1). In addition, we quantified the mean expression of all the genes in each of our COAD and READ signatures and compared them to the overall mean across genes for each of the datasets. For both COAD and READ, we observed that all genes in our signature sets were well below the mean expression of all genes in the datasets (Figure S7).

Conclusions

In conclusion, we developed a novel metric based on miRNA-gene target interactions that improved an integrated model's predictive performance in CRC. We demonstrated that our method, CCsc MMS, outperforms existing methods and that we have a more interpretable model by using a LASSO regularization. We show that CCsc MMS is a valuable method for predicting the survival of CRC patients and offering an interpretable and insightful way to examine the most important genes in a large data context. We show that many of these largest coefficients are enriched for various aspects of NOTCH3 signaling, potassium, and overall metabolism, which has been shown to play an important role in CRC.

Methods

2.1. Model Overview

The general pipeline that we develop in this study is called Colon Cancer Single-cell (CCsc). CCsc aims to prioritize genes for the input of a LASSO penalized Cox model. Gene prioritization integrates the rank of genes targeted by CRC-specific miRNAs and the ranks revealed by the scRNA-seq data of relevant tumors. We use one miRNA-based metric and two scRNA-seq-based metrics to construct a linear model with metric weights that are further optimized in the subsequent step, using bulk RNA-seq data from patients with survival data.

CCsc combines the outputs of three different metrics to get a final ranked list of genes. The metrics used are the interaction scores from the different miRNAs and the genes that they bind to (miRNA), the Median Absolute Deviation (MAD) of gene expression profiles, and the inference of switch-

like differential expression along single-cell trajectories of genes in different stages of EMT (SDE) (R Core Team, Austria, 4.2.2) (University of Oxford, England, 1.22.0) [74, 75].

2.2. miRNA Metric for Ranking Genes

The miRNA metric involves first obtaining the overlap of two databases (miRmap and dbDEMC); dbDEMC contains many experimentally validated microRNAs for the cancer of interest (Computational Genetics Group, China, 2.0) (Zdobnov Group, Switzerland, 1.2.0) [76, 77]. We took the intersection of the cancer-specific miRNAs from dbDEMC with all of the miRNAs in miRMap. dbDEMC has collated 2224 differentially expressed miRNAs from 36 different cancer types. The miRNAs common to the dbDEMC cancer of interest and present in miRMap were then submitted to TargetScan to acquire the top-ranked genes that interact with each miRNA (Whitehead Institute for Biomedical Research, Cambridge, MA, USA, 7.2.0). The optimal number of miRNAs and target genes were determined later. We created a $g \times m$ matrix, where g is the number of genes, and m is the number of microRNAs based on the interaction data from TargetScan. Each matrix entry is either 0 for a non-interacting pair or 1 for an interacting pair. We summed the score of each gene and ranked them from highest to lowest, and returned this list of ranked genes to be used in our combined linear model (Figure 5), i.e., genes were ranked based on their total numbers of interacting microRNAs in this metric. We did this for all shared miRNAs for the cancer of interest.

2.3. Pre-processing scRNA-Seq Data

Gene ranking in CC Single-cell (CCsc) has three metrics, two of which depend on processed scRNA-seq data. The single-cell dataset comes from 11 primary colorectal tumors in [13]. Briefly,

the study used the 11 tumors and matched normal mucosal tissue to test an algorithm (reference component analysis) to improve clustering accuracy and elucidate new colorectal cancer subtypes [13]. This single-cell data is the foundation for both the median absolute deviation and the inference of the switch-like differential expression along single-cell trajectories (MAD and SDE) metrics. We filtered the single-cell data set to include only those genes expressed in at least 10 cells. Next, we normalized the library size with the library.size.normalize command from the phateR package (Krishnaswamy Lab, New Haven, CT, USA, 1.0.7). This command performs normalization on input data, so that the sum of the expression values for each cell sums to 1, then returns the normalized matrix to the metric space using the median UMI count per cell, effectively scaling all cells as if they were sampled evenly. We then took the square root of the dataset to transform the data, but avoid the instabilities and pseudo counts needed when taking the log of the dataset, following best practices from [78]. Next, we used the MAGIC package to denoise scRNA-seq data and imputed the missing gene expression profiles (Krishnaswamy Lab, New Haven, CT, USA, 2.0.3) [79]. Briefly, it calculates a cell–cell distance matrix in reduced dimensions. An adaptive Gaussian kernel converts the distance matrix to a cell–cell affinity (similarity) matrix. Additional steps are used to create a Markov transition matrix. The denoised scRNA-seq matrix is created by multiplying the exponentiated Markov transition matrix by the gene expression matrix.

2.4. Inferring EMT-Based Pseudotime

The SDE metric requires an ordering of cells that represents an activation or deactivation process. Despite the absence of time-series information, we inferred the pseudotime ordering of cells in the scRNA-seq data [80]. We used a scRNA-seq dataset for colorectal tumors that primarily contain epithelial cells [13]. Epithelial-mesenchymal transition is a known driver for epithelial plasticity

and tumor progression for colorectal cancer and several other cancer types [13, 81-84]. There, we used the expression levels of a specific mesenchymal signature gene, *VIM*, as an approximation for cells progressing through EMT [84-86]. In the scRNA-seq dataset, we found that the expression levels of most epithelial (E) genes were negatively correlated with pseudotime, whereas those of mesenchymal (M) genes were positively correlated with pseudotime (Figure 6) [83, 84]. This suggests a reasonable pseudo-temporal ordering of epithelial tumor cells. It should be noted that, here, we do not use pseudo-temporal ordering to infer trajectories of cell state transitions during development. We, therefore, do not analyze the connectivity of cell state attractors (cell types). Instead, the ordering allows us to analyze the variations in expression for all genes in a relatively uniform framework of cell states (see the description of the SDE method below).

2.5. MAD Metric for Ranking Genes

MAD is calculated with

$$m_g = median(e_1, e_2, \ldots, e_c) \quad (1)$$

$$MAD(g) = median(|e_1 - m_g|, |e_2 - m_g|, \ldots, |e_c - m_g|) \quad (2)$$

where e_i represents the expression level of a gene g in cell i, and c is the number of cells.

2.6. Inference of Switch-Like Differential Expression along Single-Cell Trajectories (SDE) Metric for Ranking Genes

To calculate SDE, we used the R package switchde, which estimates the differentiation of switch-like genes in different stages of EMT. It defines a sigmoid function

$$f(t_c; \mu_g^0;\ k_g;\ t_g^0) = \frac{2\mu_g^0}{1 + \exp\left(-\kappa_g\left(t_c - t_g^0\right)\right)} \quad (3)$$

to fit the profile of a gene g concerning a pseudotime. In Eq 2, μ_g^0 corresponds to the average peak expression; κ_g is the activation nonlinearity; t_c is the active pseudotime of g in cell c; and t_g^0 is

the offset of the activation. Intuitively, κ_g represents how quickly the gene g is up or downregulated along the pseudotime and is used as the metric score. Note that κ_g may reflect how genes are switched on or off dynamically due to transcriptional bursting [87], but it can also include other sources of variability of gene expression, such as mutations and post-transcriptional regulations [88, 89]. Therefore, the parameter should be viewed as a characteristic variational pattern of gene expression across a cell population in the pseudo-temporal trajectory, common to all genes.

2.7. Overall Gene Prioritization Scoring

To make ranges of gene scores consistent across the three metrics (MAD, SDE and miRNA), each gene's final score for each metric was normalized to be between 0 and 1, by dividing its raw score by the sum of the score of all genes for the metric. Next, we combined the scores from the three metrics using the function

$$\boldsymbol{w = a_1 w_{SDE} + a_2 w_{MAD} + a_3 w_{miRNA}} \quad (4)$$

where w_{SDE}, w_{MAD} and w_{miRNA} are the normalized metric scores. Parameters a_1, a_2 and a_3 are metric weights for combining the metric scores into the overall score w. If a gene does not have a score for miRNA metric due to the lack of miRNA targeting information, its score is assumed to be zero. We ranked the genes based on overall scores, and we selected the genes with high scores (see details later) for the subsequent analysis. The values of a_1, a_2 and a_3 were determined by a grid search with an interval of 0.1, and the constraint of the sum of these parameters is 1. The performance of the grid search is based on the Cox model and the concordance index described below.

2.8. LASSO Regularized Cox Model

To build prognostic models and to further select important genes from the prioritized lists, we used bulk RNA-seq data and the associated patient survival data from the Cancer Genome Atlas (TCGA) or cBioPortal for Cancer Genomics. We used a LASSO regularized Cox model to determine the concordance index and top-performing coefficients. This regularization uses the L1 lasso penalty. This allows the number of coefficients to be constrained based on the value of the penalty weight parameter λ that we use. The regularization was used to avoid the overfitting of our models. Regularizations help to address overfitting and make Cox models more interpretable. The LASSO regularization, which involves finding a subset of predictors which give a model's best overall performance, is used here to improve the interpretability of our model [90-93]. We used the glmnet package in R to perform all the LASSO regularized Cox-model fittings (Hastie Lab, Stanford, CA, USA, 4.1-6) [94, 95].

2.9. Concordance Index (C-index)

The concordance index (C-index) is the primary metric for assessing our method's effectiveness. This metric is analogous to the area under the curve–receiver–operator characteristic, but is applied specifically for survival analysis situations. It is calculated by

$$\boldsymbol{C-index} = \frac{\sum\sum_{i<j}\left[\boldsymbol{I(t_i < t_j)I(r_i > r_j)I(\delta_i \equiv 1) + I(t_i > t_j)I(r_i < r_j)I(\delta_j \equiv 1)}\right]}{\sum\sum_{i<j}\left[\boldsymbol{I(t_i < t_j)I(\delta_i \equiv 1) + I(t_i > t_j)I(\delta_j \equiv 1)}\right]}. \quad (5)$$

The C-index is equal to the concordance probability P ($r_i > r_j | z_i < z_j$) for a randomly selected pair of patients I and j. Unfortunately, we cannot observe the potential survival time for some patients who are lost to follow-up or are event free at the end of a study (right censored). Given this, the observed survival time $t_i = min$ (z_i, c_i), where c_i is the potential right censoring time; δ_i is the

censoring status. An event (e.g., death) is when $\delta_i = 1$. *I ()* is the indicator function, and *r* is equal to the risk score for patients *i* and *j,* respectively.

Essentially, this metric assesses the ability of a set of input predictors to accurately judge whether a patient with a particular risk score will get cancer in a specific period. Specifically, the C-index judges whether a model has discriminatory power and accurately ranks the patient's survival time when considering their calculated risk scores. In an ideal case, a model would ideally separate all patients based on their risk score into their correct group and would have a performance of 1. A model with a C-index of 0.5 is classified as a random predictor [96].

2.10. Kaplan-Meier Survival Analysis

The other primary output of our model is a Kaplan–Meier risk estimator (Terry Therneau, Rochester, MN, 3.4-0) [97]. This metric attempts to determine how well a set of inputs in a model can correctly stratify patients in our datasets as high risk or low risk, based on their gene expression profiles. The survival probability s at time t is given by

$$s_t = \frac{\text{Number of subjects living at the start} - \text{Number of subjects who died}}{\text{Number of subjects living at the start}} \quad (6)$$

The Kaplan–Meier survival analysis uses the log-rank test to assess whether the high and low-risk groups' survival times differ statistically. The statistic is given by

$$\text{Log} - \text{rank test statistic} = \frac{(O - E)^2}{E} \quad (7)$$

$\boldsymbol{O}$ is the total of the observed events and $\boldsymbol{E}$ is the total of the expected events. For each event of interest (death), we calculate the number of deaths observed and the number of deaths expected, if there was really no difference between our groups. This calculation is performed for both the high and low-risk groups in our data. We then sum the number of observed and expected events to get

$\boldsymbol{O}$ and $\boldsymbol{E}$ in (5). If a survival time is censored, that particular individual is considered to be at risk of dying in the week of the censoring, but not in subsequent weeks [98].

A p-value < 0.05 is statistically significant for our Kaplan–Meier survival analysis.

2.11. Datasets

The datasets in this study include several single-cell and bulk RNA-seq datasets, and we used several scRNA-seq based and bulk RNA-seq based methods either for constructing our model or benchmarking. The scRNA-seq data was obtained from a previous study on combined tumor samples of 11 colorectal cancer patients [13]. The bulk RNA-seq data of colon and rectal cancer were obtained from the Cancer Genome Atlas (TCGA), which were used to train the Cox models. The colon cancer dataset includes 461 cases, and the rectal cancer dataset includes 172 cases. The TCGA datasets can be accessed through the Genome Data Commons web portal or the TCGAbiolinks R package (https://portal.gdc.cancer.gov) (Accessed on 1 December 2022) [99]. The additional bulk dataset from the cBioPortal for Cancer Genomics contains 79 cases and can be accessed through the cBioPortal web interface (https://www.cbioportal.org) (Accessed on 1 December 2022) [100]. The scRNA-seq data can be found at GEO under accession GSE81861.

2.12. Implementation

We have implemented our metric and model in R on GitHub (https://github.com/compbiolover/CC-Singlecell) (Accessed 1 December 2022). All code and datasets used in this manuscript are available there.

References

1. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome biology, 2014. **15**(12): p. 1-21.
2. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-140.
3. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* 2001.
4. Venter, J.C., et al., *The sequence of the human genome.* science, 2001. **291**(5507): p. 1304-1351.
5. Nagalakshmi, U., et al., *The transcriptional landscape of the yeast genome defined by RNA sequencing.* Science, 2008. **320**(5881): p. 1344-1349.
6. Qian, M., et al. *Detection of single cell heterogeneity in cancer*. in *Seminars in Cell & Developmental Biology*. 2017. Elsevier.
7. Tellez-Gabriel, M., et al., *Tumour heterogeneity: the key advantages of single-cell analysis.* International journal of molecular sciences, 2016. **17**(12): p. 2142.
8. Zhao, N. and J.M. Rosen. *Breast cancer heterogeneity through the lens of single-cell analysis and spatial pathologies*. in *Seminars in Cancer Biology*. 2021. Elsevier.
9. Li, X., et al., *A novel single-cell based method for breast cancer prognosis.* PLoS computational biology, 2020. **16**(8): p. e1008133.
10. Bartel, D.P., *MicroRNAs: genomics, biogenesis, mechanism, and function.* cell, 2004. **116**(2): p. 281-297.
11. Di Leva, G., M. Garofalo, and C.M. Croce, *MicroRNAs in cancer.* Annual Review of Pathology: Mechanisms of Disease, 2014. **9**: p. 287-314.
12. Lu, T.X. and M.E. Rothenberg, *MicroRNA.* Journal of Allergy and Clinical Immunology, 2018. **141**(4): p. 1202-1207.
13. Li, H., et al., *Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors.* Nature genetics, 2017. **49**(5): p. 708-718.
14. Pratama, R., et al., *Authentication of differential gene expression in oral squamous cell carcinoma using machine learning applications.* BMC oral health, 2021. **21**(1): p. 1-8.
15. Zhu, Y., Y. Li, and A. Nakagawara, *UNC5 dependence receptor family in human cancer: a controllable double-edged sword.* Cancer Letters, 2021. **516**: p. 28-35.
16. Callen, D.F., *Revisiting the identification of breast cancer tumour suppressor genes defined by copy number loss of the long arm of chromosome 16.* bioRxiv, 2021.
17. Lu, Q., et al., *LncRNA TP53TG1 Promotes the Growth and Migration of Hepatocellular Carcinoma Cells via Activation of ERK Signaling.* Non-coding RNA, 2021. **7**(3): p. 52.
18. Dong, Y., et al., *Comprehensive analysis of coexpressed long noncoding RNAs and genes in breast cancer.* Journal of Obstetrics and Gynaecology Research, 2019. **45**(2): p. 428-437.
19. Kasper, B.T., S. Koppolu, and L.K. Mahal, *Insights into miRNA regulation of the human glycome.* Biochemical and biophysical research communications, 2014. **445**(4): p. 774-779.

20. Brim, H., et al., *Streptococcus sp. VT_162 infection of colon cancer cell lines induces mRNAs that associate with poor prognosis.* Cancer Research, 2020. **80**(16_Supplement): p. 6102-6102.
21. Chen, S., et al., *Methylation gene KCNC1 is associated with overall survival in patients with seminoma.* Oncology Reports, 2021. **45**(5): p. 1-10.
22. Gu, X.-Y., et al., *MicroRNA is a potential target for therapies to improve the physiological function of skeletal muscle after trauma.* Neural Regeneration Research, 2022. **17**(7): p. 1617.
23. Iravani, O., B.-H. Bay, and G.W.-C. Yip, *Silencing HS6ST3 inhibits growth and progression of breast cancer cells through suppressing IGF1R and inducing XAF1.* Experimental Cell Research, 2017. **350**(2): p. 380-389.
24. Guo, Y., et al., *Downregulation of HS6ST2 by miR-23b-3p enhances matrix degradation through p38 MAPK pathway in osteoarthritis.* Cell death & disease, 2018. **9**(6): p. 1-15.
25. Ma, R., et al., *FABP7 promotes cell proliferation and survival in colon cancer through MEK/ERK signaling pathway.* Biomedicine & Pharmacotherapy, 2018. **108**: p. 119-129.
26. Tian, X., et al., *Circ_ZFR affects FABP7 expression to regulate breast cancer progression by acting as a sponge for miR-223-3p.* Thoracic Cancer, 2022. **13**(9): p. 1369-1380.
27. Misawa, K., et al., *Identification of novel methylation markers in HPV-associated oropharyngeal cancer: Genome-wide discovery, tissue verification and validation testing in ctDNA.* Oncogene, 2020. **39**(24): p. 4741-4755.
28. Araújo, T., et al., *Recurrent amplification of RTEL1 and ABCA13 and its synergistic effect associated with clinicopathological data of gastric adenocarcinoma.* Molecular Cytogenetics, 2016. **9**(1): p. 1-7.
29. Li, C., et al., *OPCML is frequently methylated in human colorectal cancer and its restored expression reverses EMT via downregulation of smad signaling.* American journal of cancer research, 2015. **5**(5): p. 1635.
30. Guo, J., et al., *Long Non-Coding RNA RFPL3S Functions as a Biomarker of Prognostic and Immunotherapeutic Prediction in Testicular Germ Cell Tumor.* Frontiers in Immunology, 2022. **13**.
31. Wang, Y., et al., *A novel human KRAB-related zinc finger gene ZNF425 inhibits mitogen-activated protein kinase signaling pathway.* BMB reports, 2011. **44**(1): p. 58-63.
32. Wang, W., et al., *Overexpression of Uromodulin-like1 accelerates follicle depletion and subsequent ovarian degeneration.* Cell death & disease, 2012. **3**(11): p. e433-e433.
33. Kigel, B., et al., *Plexin-A4 promotes tumor progression and tumor angiogenesis by enhancement of VEGF and bFGF signaling.* Blood, The Journal of the American Society of Hematology, 2011. **118**(15): p. 4285-4296.
34. Mawaribuchi, S., et al., *mRNA and miRNA expression profiles in an ectoderm-biased substate of human pluripotent stem cells.* Scientific reports, 2019. **9**(1): p. 1-13.
35. Lou, X., et al., *MFAP3L activation promotes colorectal cancer cell invasion and metastasis.* Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease, 2014. **1842**(9): p. 1423-1432.
36. Ye, J., et al., *MicroRNA-671-5p inhibits cell proliferation, migration and invasion in non-small cell lung cancer by targeting MFAP3L.* Molecular Medicine Reports, 2022. **25**(1): p. 1-8.
37. Nakatsugawa, M., et al., *Novel spliced form of a lens protein as a novel lung cancer antigen, Lengsin splicing variant 4.* Cancer science, 2009. **100**(8): p. 1485-1493.

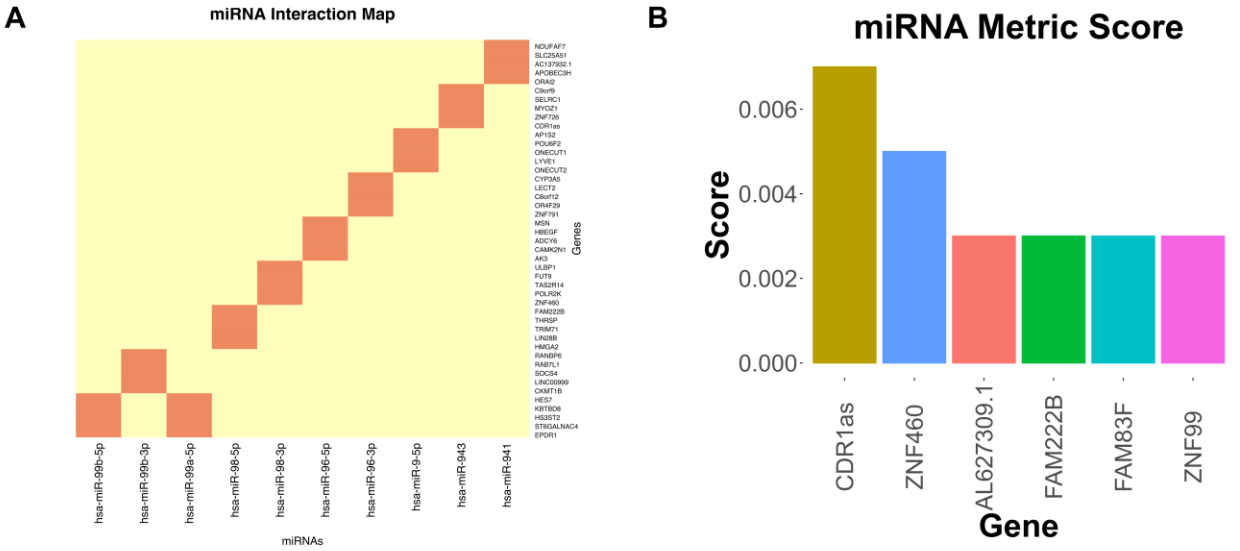
38. Zhang, T., et al., *Overexpression of flavin-containing monooxygenase 5 predicts poor prognosis in patients with colorectal cancer.* Oncology letters, 2018. **15**(3): p. 3923-3927.
39. Yu, W., et al., *A novel amplification target, DUSP26, promotes anaplastic thyroid cancer cell growth by inhibiting p38 MAPK activity.* Oncogene, 2007. **26**(8): p. 1178-1187.
40. Thompson, E.M. and A.W. Stoker, *A review of DUSP26: structure, regulation and relevance in human disease.* International Journal of Molecular Sciences, 2021. **22**(2): p. 776.
41. Zhang, Z.Y., et al., *Metabolic reprogramming-associated genes predict overall survival for rectal cancer.* Journal of cellular and molecular medicine, 2020. **24**(10): p. 5842-5849.
42. Kefas, B., et al., *A miR-297/hypoxia/DGK-α axis regulating glioblastoma survival.* Neuro-oncology, 2013. **15**(12): p. 1652-1663.
43. Gaston, D., et al., *Germline mutations in MAP3K6 are associated with familial gastric cancer.* PLoS genetics, 2014. **10**(10): p. e1004669.
44. Kula, A., et al., *The role of genetic polymorphism within PD-L1 gene in cancer. Review.* Experimental and Molecular Pathology, 2020. **116**: p. 104494.
45. Wang, Y., et al., *A Comprehensive Prognostic Analysis of Tumor-Related Blood Group Antigens in Pan-Cancers Suggests That SEMA7A as a Novel Biomarker in Kidney Renal Clear Cell Carcinoma.* International journal of molecular sciences, 2022. **23**(15): p. 8799.
46. Han, J., et al., *Deregulated AJAP1/β-catenin/ZEB1 signaling promotes hepatocellular carcinoma carcinogenesis and metastasis.* Cell death & disease, 2017. **8**(4): p. e2736-e2736.
47. He, Z. and B. Li, *Recent progress in genetic and epigenetic profile of diffuse gastric cancer.* Cancer Translational Medicine, 2015. **1**(3).
48. Shamsdin, S.A., et al., *Associations of ICOS and PD. 1 gene variants with colon cancer risk in the Iranian population.* Asian Pacific Journal of Cancer Prevention: APJCP, 2018. **19**(3): p. 693.
49. Peng, L., et al., *APEX1 regulates alternative splicing of key tumorigenesis genes in non-small-cell lung cancer.* BMC Medical Genomics, 2022. **15**(1): p. 1-10.
50. Peppino, G., et al., *Teneurins: Role in Cancer and Potential Role as Diagnostic Biomarkers and Targets for Therapy.* International Journal of Molecular Sciences, 2021. **22**(5): p. 2321.
51. Sun, Y., et al., *Identification of lnc RNAs related to prognosis of patients with colorectal cancer.* Technology in cancer research & treatment, 2020. **19**: p. 1533033820962120.
52. Shinmura, K., et al., *BSND and ATP6V1G3: novel immunohistochemical markers for chromophobe renal cell carcinoma.* Medicine, 2015. **94**(24).
53. Zhang, T., et al., *RS1 gene is a novel prognostic biomarker for lung adenocarcinoma.* Thoracic Cancer, 2022.
54. Winter, G.E., et al., *The solute carrier SLC35F2 enables YM155-mediated DNA damage toxicity.* Nature chemical biology, 2014. **10**(9): p. 768-773.
55. Wang, Y., et al., *RGS12 Is a Novel Tumor-Suppressor Gene in African American Prostate Cancer That Represses AKT and MNX1 ExpressionRGS12 in African American Prostate Cancer.* Cancer research, 2017. **77**(16): p. 4247-4257.
56. Ward, C.J., et al., *Germline PKHD1 mutations are protective against colorectal cancer.* Human genetics, 2011. **129**(3): p. 345-349.
57. Xu, Q., et al., *Clinical importance of PLA2R1 and RASSF9 in thyroid cancer and their inhibitory roles on the Wnt/β-catenin pathway and thyroid cancer cell malignant behaviors.* Pathology-Research and Practice, 2022. **238**: p. 154092.

58. Shorthouse, D., et al., *THE ROLE OF POTASSIUM CHANNELS IN THE PATHOGENESIS OF GASTROINTESTINAL CANCERS AND THERAPEUTIC POTENTIAL.* bioRxiv, 2022: p. 2020.03. 10.984039.
59. Pancho, A., et al., *Protocadherins at the crossroad of signaling pathways.* Frontiers in Molecular Neuroscience, 2020. **13**: p. 117.
60. Ghafouri-Fard, S. and M. Taheri, *Maternally expressed gene 3 (MEG3): A tumor suppressor long non coding RNA.* Biomedicine & Pharmacotherapy, 2019. **118**: p. 109129.
61. Lechner, S., et al., *Target deconvolution of HDAC pharmacopoeia reveals MBLAC2 as common off-target.* Nature Chemical Biology, 2022: p. 1-9.
62. Zhang, X., et al., *ZNF23 suppresses cutaneous melanoma cell malignancy via mitochondria-dependent pathway.* Cellular Physiology and Biochemistry, 2017. **43**(1): p. 147-157.
63. Vafaeie, F., et al., *ACAN, MDFI, and CHST1 as Candidate Genes in Gastric Cancer: A Comprehensive Insilco Analysis.* Asian Pacific Journal of Cancer Prevention, 2022. **23**(2): p. 683-694.
64. Amir Shaghaghi, M., et al., *The SLC2A14 gene, encoding the novel glucose/dehydroascorbate transporter GLUT14, is associated with inflammatory bowel disease.* The American Journal of Clinical Nutrition, 2017. **106**(6): p. 1508-1513.
65. Serafin, V., et al., *Notch3 signalling promotes tumour growth in colorectal cancer.* The Journal of pathology, 2011. **224**(4): p. 448-460.
66. Abdul, M. and N. Hoosein, *Voltage-gated potassium ion channels in colon cancer.* Oncology reports, 2002. **9**(5): p. 961-964.
67. Varga, J., et al., *AKT-dependent NOTCH3 activation drives tumor progression in a model of mesenchymal colorectal cancer.* Journal of Experimental Medicine, 2020. **217**(10).
68. Pardo, L.A. and W. Stühmer, *The roles of K+ channels in cancer.* Nature Reviews Cancer, 2014. **14**(1): p. 39-48.
69. Farhan, M., et al., *Role of FOXO transcription factors in cancer metabolism and angiogenesis.* Cells, 2020. **9**(7): p. 1586.
70. Ma, J., et al. *FOXO family in regulating cancer and metabolism*. in *Seminars in cancer biology*. 2018. Elsevier.
71. Rutkowski, M.J., et al., *Cancer and the Complement CascadeCancer and the Complement Cascade.* Molecular Cancer Research, 2010. **8**(11): p. 1453-1465.
72. Gu, Y., I.S. Mohammad, and Z. Liu, *Overview of the STAT-3 signaling pathway in cancer and the development of specific inhibitors.* Oncology letters, 2020. **19**(4): p. 2585-2594.
73. Pilling, A.B., et al., *ALK is a critical regulator of the MYC-signaling axis in ALK positive lung cancer.* Oncotarget, 2018. **9**(10): p. 8823.
74. Campbell, K.R. and C. Yau, *switchde: inference of switch-like differential expression along single-cell trajectories.* Bioinformatics, 2017. **33**(8): p. 1241-1242.
75. Howell, D.C., *Median absolute deviation.* Encyclopedia of statistics in behavioral science, 2005.
76. Yang, Z., et al., *dbDEMC 2.0: updated database of differentially expressed miRNAs in human cancers.* Nucleic acids research, 2017. **45**(D1): p. D812-D818.
77. Vejnar, C.E. and E.M. Zdobnov, *MiRmap: comprehensive prediction of microRNA target repression strength.* Nucleic acids research, 2012. **40**(22): p. 11673-11683.
78. lab, K. *Rmagic EMT Tutorial*. 2018 [cited 2020 December 1 2020]; Tutorial on MAGIC best practices]. Available from:

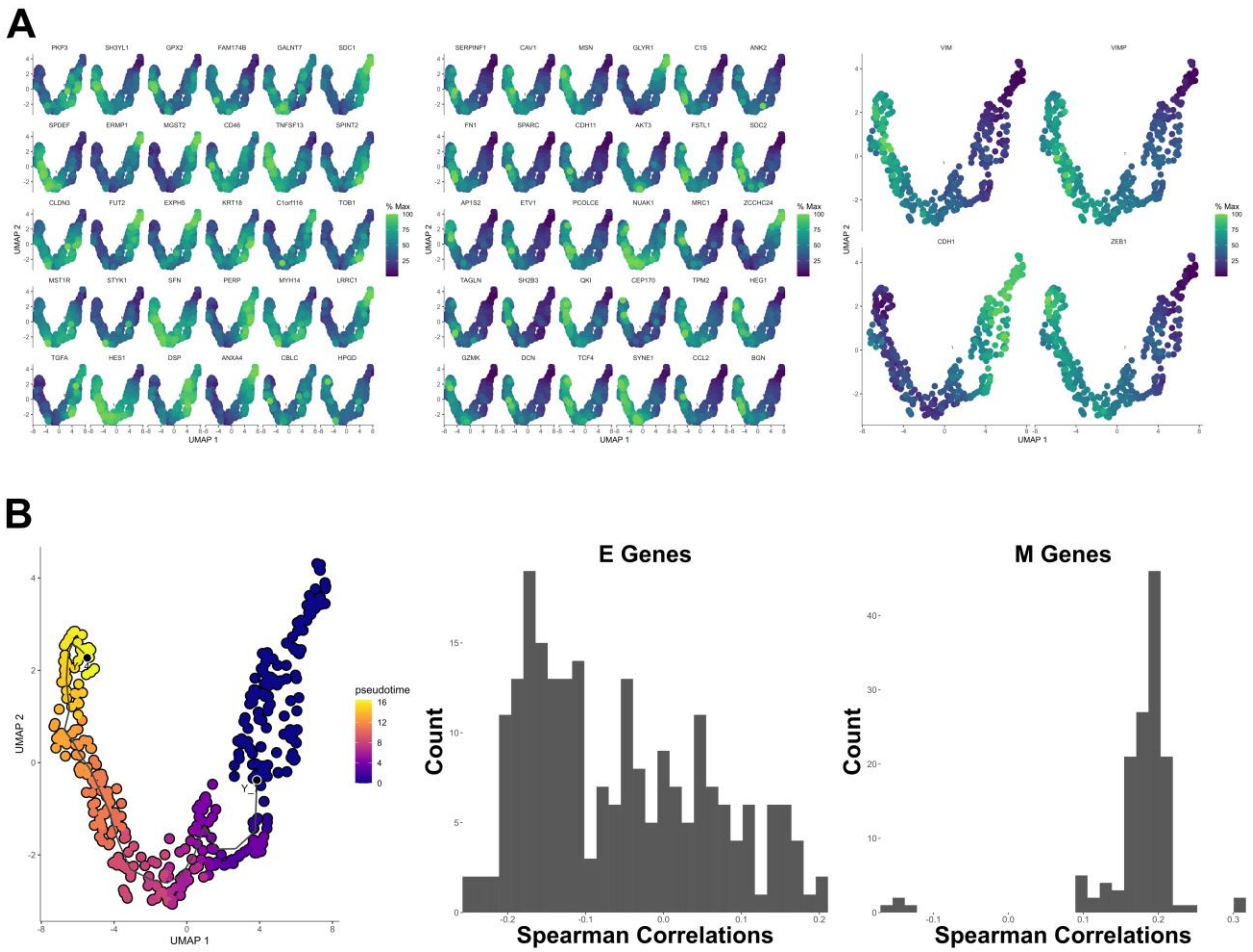
http://htmlpreview.github.io/?https://github.com/KrishnaswamyLab/MAGIC/blob/master/Rmagic/inst/examples/emt_tutorial.html.
79. Van Dijk, D., et al., *Recovering gene interactions from single-cell data using data diffusion.* Cell, 2018. **174**(3): p. 716-729. e27.
80. Cao, J., et al., *The single-cell transcriptional landscape of mammalian organogenesis.* Nature, 2019. **566**(7745): p. 496-502.
81. Rokavec, M., et al., *IL-6R/STAT3/miR-34a feedback loop promotes EMT-mediated colorectal cancer invasion and metastasis.* The Journal of clinical investigation, 2014. **124**(4): p. 1853-1867.
82. Vu, T. and P.K. Datta, *Regulation of EMT in colorectal cancer: a culprit in metastasis.* Cancers, 2017. **9**(12): p. 171.
83. Panchy, N., et al., *Integrative transcriptomic analysis reveals a multiphasic epithelial–mesenchymal spectrum in cancer and non-tumorigenic cells.* Frontiers in oncology, 2020. **9**: p. 1479.
84. Panchy, N., et al., *Comparative single-cell transcriptomes of dose and time dependent epithelial–mesenchymal spectrums.* NAR Genomics and Bioinformatics, 2022. **4**(3).
85. Wang, Z., et al., *Vimentin expression is required for the development of EMT-related renal fibrosis following unilateral ureteral obstruction in mice.* American Journal of Physiology-Renal Physiology, 2018. **315**(4): p. F769-F780.
86. Loboda, A., et al., *EMT is the dominant program in human colon cancer.* BMC medical genomics, 2011. **4**(1): p. 1-10.
87. Zhang, J., Q. Nie, and T. Zhou, *Revealing dynamic mechanisms of cell fate decisions from single-cell transcriptomic data.* Frontiers in genetics, 2019. **10**: p. 1280.
88. Hu, X., et al., *Genetic alterations and oncogenic pathways associated with breast cancer subtypes.* Molecular cancer research, 2009. **7**(4): p. 511-522.
89. Nordick, B., et al., *Nonmodular oscillator and switch based on RNA decay drive regeneration of multimodal gene expression.* Nucleic Acids Research, 2022. **50**(7): p. 3693-3708.
90. Zhao, P. and B. Yu, *On model selection consistency of Lasso.* The Journal of Machine Learning Research, 2006. **7**: p. 2541-2563.
91. Tibshirani, R., *Regression shrinkage and selection via the lasso.* Journal of the Royal Statistical Society: Series B (Methodological), 1996. **58**(1): p. 267-288.
92. Yamada, M., et al., *High-dimensional feature selection by feature-wise kernelized lasso.* Neural computation, 2014. **26**(1): p. 185-207.
93. Muthukrishnan, R. and R. Rohini. *LASSO: A feature selection technique in predictive modeling for machine learning*. in *2016 IEEE international conference on advances in computer applications (ICACA)*. 2016. IEEE.
94. Simon, N., et al., *Regularization paths for Cox's proportional hazards model via coordinate descent.* Journal of statistical software, 2011. **39**(5): p. 1.
95. Zou, H. and T. Hastie, *Regularization and variable selection via the elastic net.* Journal of the royal statistical society: series B (statistical methodology), 2005. **67**(2): p. 301-320.
96. Uno, H., et al., *On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data.* Statistics in medicine, 2011. **30**(10): p. 1105-1117.
97. Goel, M.K., P. Khanna, and J. Kishore, *Understanding survival analysis: Kaplan-Meier estimate.* International journal of Ayurveda research, 2010. **1**(4): p. 274.
98. Bland, J.M. and D.G. Altman, *The logrank test.* Bmj, 2004. **328**(7447): p. 1073.

99. Colaprico, A., et al., *TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data.* Nucleic acids research, 2016. **44**(8): p. e71-e71.
100. Chatila, W.K., et al., *Genomic and transcriptomic determinants of response to neoadjuvant therapy in rectal cancer.* Nature Medicine, 2022. **28**(8): p. 1646-1655.

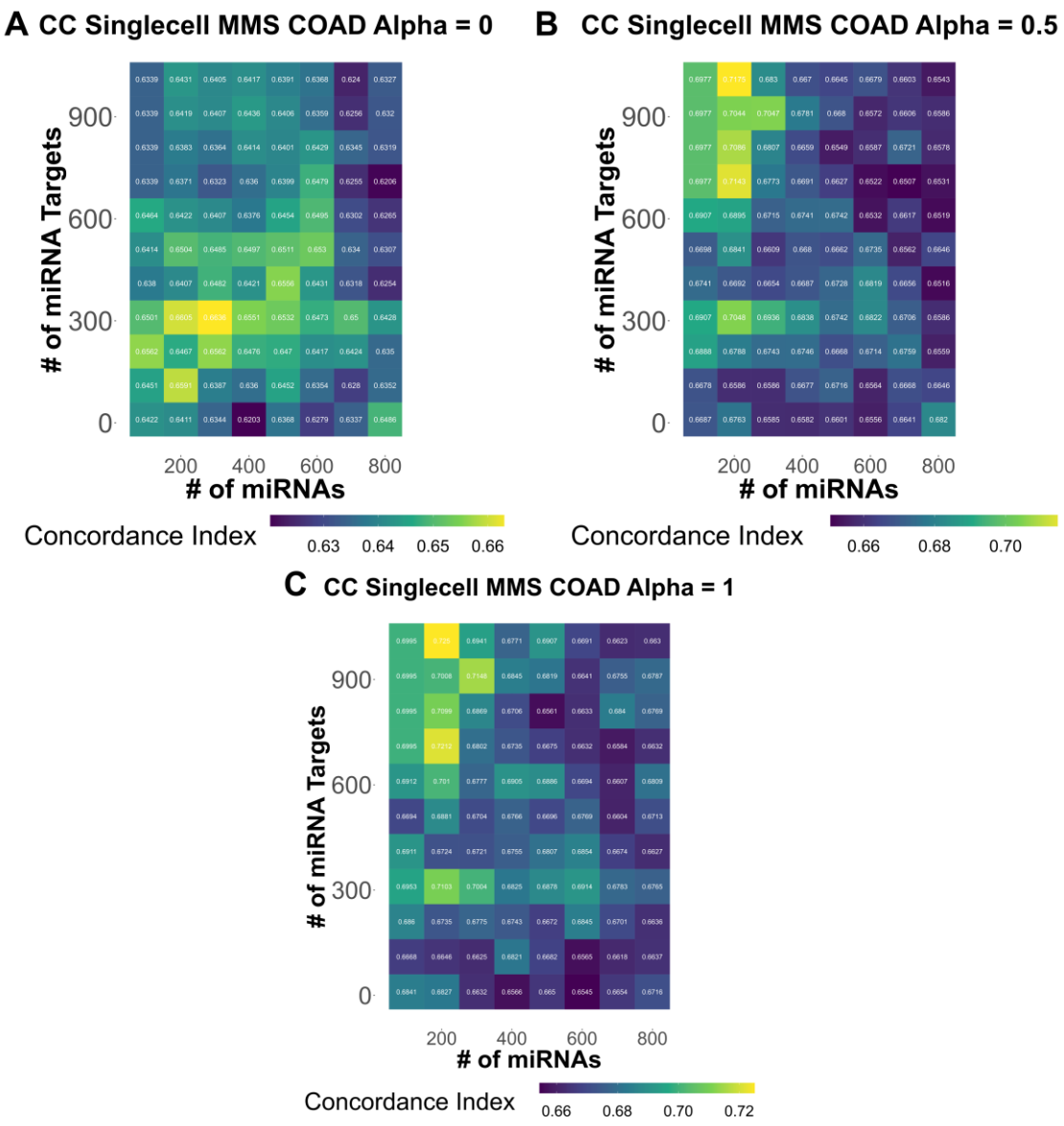
Appendix



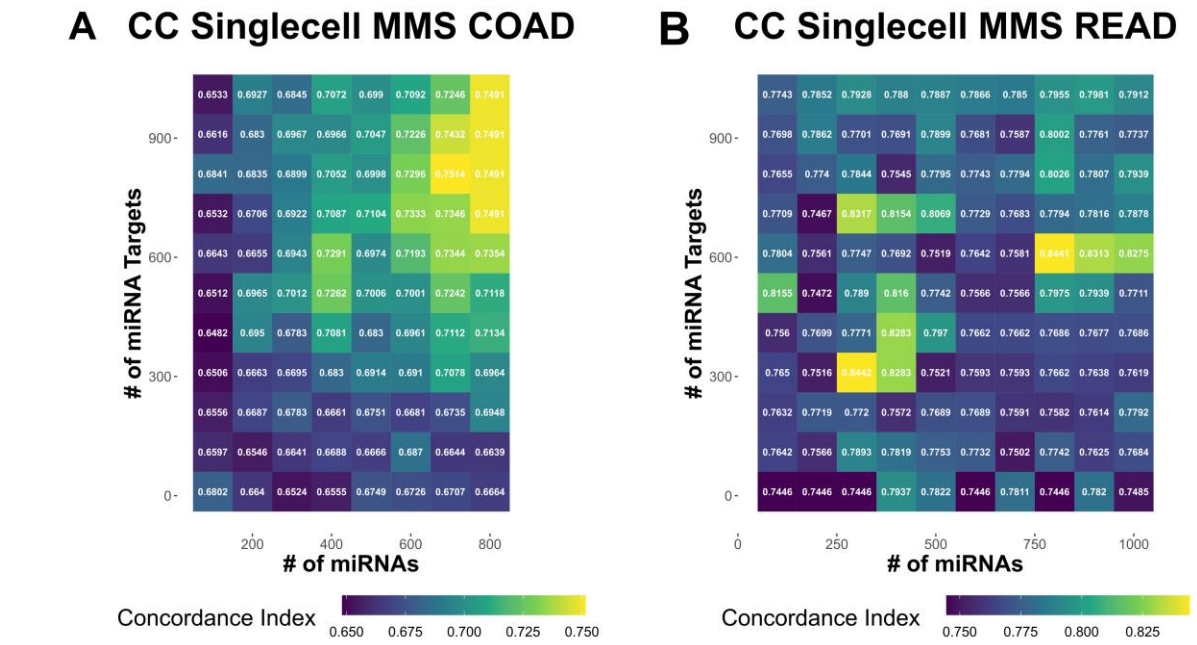
Supplemental Figure 1. miRNA metric (**A**) Heatmap depicting the interactions between the experimentally validated miRNAs and their targets in READ data. Here we show a small subset of the genes and miRNAs that make up our miRNA metric for the TCGA-COAD dataset. These were acquired by first taking the overlap of the miRNAs from miRMap and dbDEMC 2.0. Then these common miRNAs were submitted to the TargetScan database, and these miRNAs' top N gene targets were returned. The number of times each miRNA interacted with each gene was then calculated by simply increasing the number for a gene by one if it interacted with a particular miRNA. Red is high binding; blue is low binding. Once all miRNA-target pairs were evaluated, the genes were ranked from the greatest number of interactions to the fewest. This score was then normalized to between 0 and 1. (**B**) Example of some of the top gene targets for the COAD dataset.



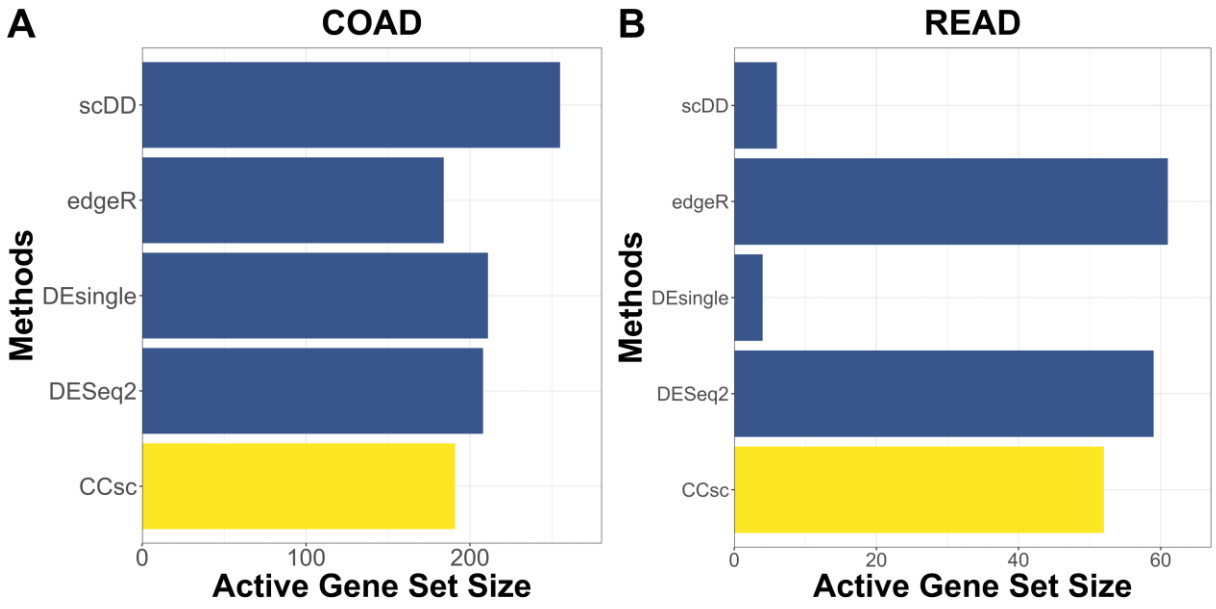
Supplemental Figure 2. Pseudo-temporal ordering analysis. (**A)** UMAP of gene expression of 30 randomly selected epithelial (E) annotated genes (left panel). UMAP of gene expression of 30 randomly selected mesenchymal (M) annotated genes (center panel). UMAP of gene expression of 4 hallmark EMT genes that represent E (CDH1) or M (VIM, VIMP, ZEB1) gene expression states in the Li et al data (right panel). (**B)** UMAP of Li et al data colored by pseudotime values based on the gene expression profile of VIM (left). Histograms of the distributions of the Spearman correlation coefficients between gene expression levels and pseudotime values for all E annotated (center) or M annotated (right) genes.



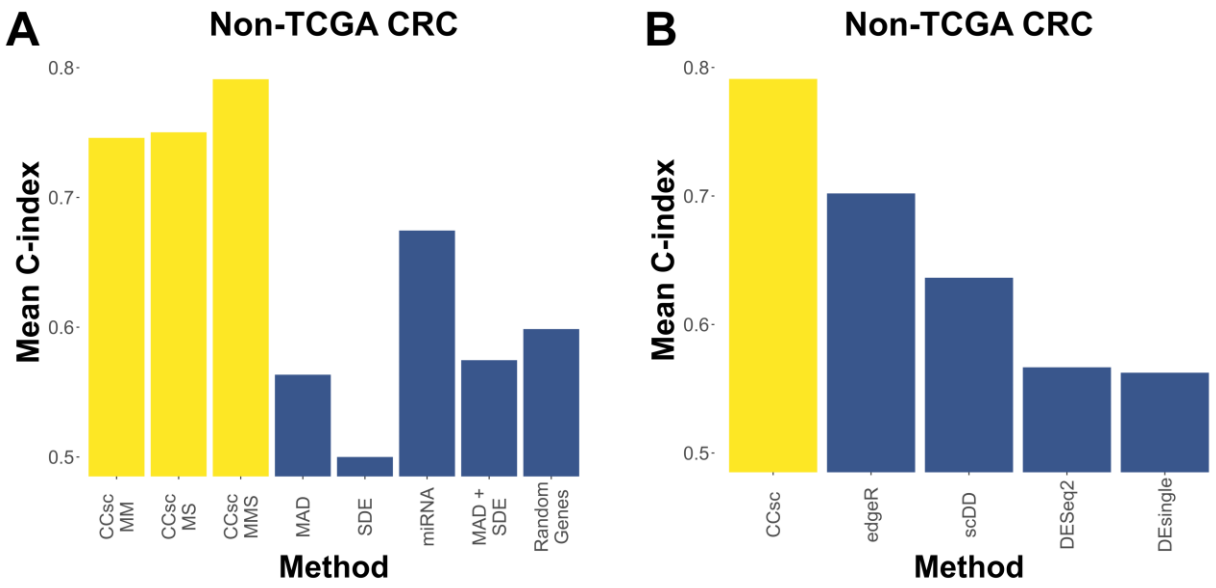
Supplemental Figure 3. LASSO regularization gives the best performance. Heatmaps showing our grid searches at several different values of alpha for **(A)** 0, **(B)** 0.5, and **(C)** 1. We plot the mean 10-fold cv at each combination of miRNA and miRNA target number.



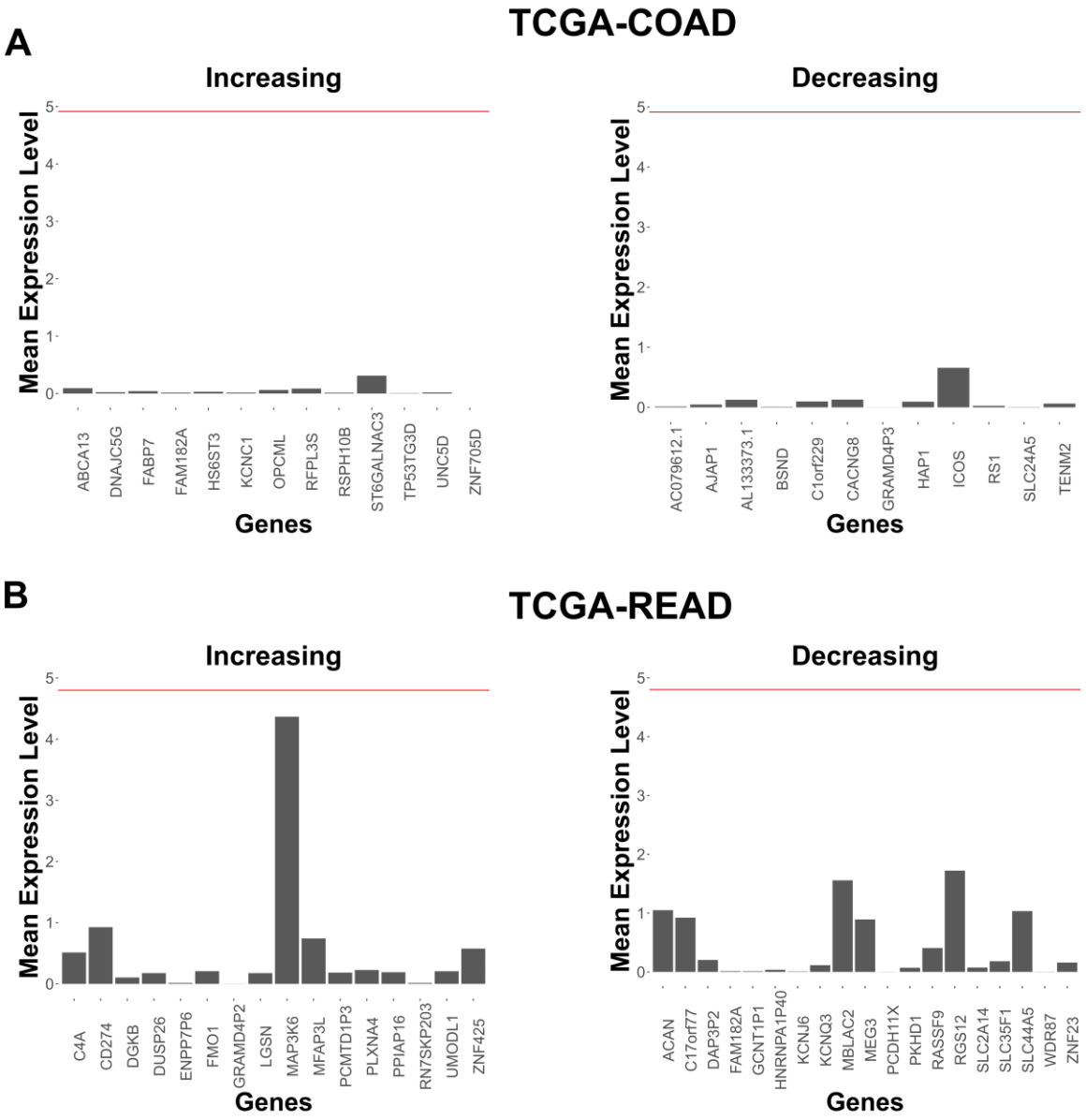
Supplemental Figure 4. Determination of Ideal miRNA and miRNA target values. Heatmaps showing our grid searches of many combinations of miRNAs and miRNA targets for (**A)** TCGA-COAD and (**B)** TCGA-READ. We send each combination of miRNAs and miRNA targets through our pipeline, and then we send the ideal number of genes in each integrated list of genes from our model into the LASSO Cox model for TCGA-COAD and TCGA-READ datasets. The gene size for TCGA-COAD is 2,500, and the gene size for TCGA-READ is 600. We then plot the concordance index of each combination.



Supplemental Figure 5. CCsc MMS Uses Comparable Gene Set Size to Existing Methods. (A) We ran each method at optimal gene set size and then extracted the non-zero betas from the LASSO penalized Cox model (i.e., Active genes). We see that, on average, CCsc has a much higher concordance index than other methods while using approximately the same number of genes. We also see this in TCGA-READ (**B)**



Supplemental Figure 6. CCsc MMS Outperforms other Methods in Non-TCGA CRC Data. We performed the same analysis on a non-TCGA CRC dataset. We found that CCsc MMS continues to outperform its individual components (**A**). Additionally, it outperformed all other methods on this new dataset (**B**).



Supplemental Figure 7. Highly Ranked Gene Targets Show Reduced Expression Level. Each of the top performing genes in our signatures showed much reduced expression compared to the mean of all genes present in the (**A**) TCGA-COAD and (**B**) TCGA-READ datasets respectively. This was true for both the top genes that were found to increase patient risk (left panels) and to decrease patient risk (right panels). Horizontal red line denotes the mean expression of each dataset.

Gene	Directly Regulated	Indirectly Regulated	Citation
ZNF705D	N	N	N/A
UNC5D	Y	N	[15]
TP53TG3D	Y	Y	[17]
ST6GALNAC3	Y	N	[19]
RSPH10B	N	N	N/A
KCNC1	Y	N	[22]
HS6ST3	Y	N	[24]

FAM182A	N	N	N/A
FABP7	N	Y	[26]
DNAJC5G	N	N	N/A
ZNF425	N	N	N/A
UMODL1	N	N	N/A
RN7SKP203	N	N	N/A
PPIAP16	N	N	N/A
PLXNA4	Y	N	[34]
MFAP3L	Y	N	[36]
LGSN	N	N	N/A
GRAMD4P2	N	N	N/A
FMO1	N	N	N/A
ENPP7P6	N	N	N/A
DUSP26	N	Y	[40]
DGKB	Y	**N**	[42]

Supplemental Table 1. Genes Identified by Model Regulated by miRNA. Genes are either directly or indirectly regulated by miRNAs.

CHAPTER III
COMPARATIVE SINGLE-CELL TRANSCRIPTOMES OF DOSE AND TIME DEPENDENT EPITHELIAL–MESENCHYMAL SPECTRUMS

A version of this chapter was originally published by Nicholas Panchy, Kazuhide Watanabe, Masataka Takahashi, Andrew Willems and Tian Hong:
Panchy N, Watanabe K, Takahashi M, Willems A, Hong T Comparative single-cell transcriptomes of dose and time dependent epithelial–mesenchymal spectrums. *NAR Genom Bioinform*. 2022, 4(3): lqac072

My contributions to this work include: (1) prognostic model, (2) data analysis, (3) reading of manuscript.

Abstract

Epithelial-mesenchymal transition (EMT) is a cellular process involved in development and disease progression. Intermediate EMT states were observed in tumors and fibrotic tissues, but previous *in vitro* studies focused on time-dependent responses with single doses of signals; it was unclear whether single-cell transcriptomes support stable intermediates observed in diseases. Here, we performed single-cell RNA-sequencing with human mammary epithelial cells treated with multiple doses of TGF-β. We found that dose-dependent EMT harbors multiple intermediate states at nearly steady state. Comparisons of dose- and time-dependent EMT transcriptomes revealed that the dose-dependent data enable higher sensitivity to detect genes associated with EMT. We identified cell clusters unique to time-dependent EMT, reflecting cells *en route* to stable states. Combining dose- and time-dependent cell clusters gave rise to accurate prognosis for cancer patients. Our transcriptomic data and analyses uncover a stable EMT continuum at the single-cell resolution, and complementary information of two types of single-cell experiments.

Introduction

Epithelial-mesenchymal transition (EMT) is a cellular process in which epithelial (E) cells undergo fate switches towards mesenchymal (M) types. This process renders the loss of apical-basal polarity and the gain of migratory properties. EMT plays crucial roles in development and disease progressions such as metastasis and fibrosis [1, 2]. EMT is not a binary process. In tumor cells, for example, intermediate (partial) EMT states were observed [3-5], and it was suggested that there is an association between intermediate EMT states and metastatic potentials [6, 7]. Interestingly, intermediate EMT states can also be observed *in vitro* with epithelial cell lines treated with EMT signals, such as TGF-β [3, 8], and these *in vitro* experiments provide useful insights into molecular

programs underlying partial EMT [9]. For example, experiments with genetically perturbed cells have suggested that interconnected feedback loops in gene regulatory networks can generate multiple intermediate EMT states [10]. Additionally, mathematical models postulated stability of these states arising from intricate gene regulatory networks [10-12].

At the fundamental level, intermediate EMT states can be understood as either cell states *en route* to M-like states, or those stable states induced by weak (low-dose) EMT signals in the microenvironment. Recent single-cell transcriptomic studies showed that the time-dependent EMT programs contain intermediate states that delineate a continuum-like EMT spectrum [13-15]. However, it is unclear whether stable cell states in EMT program induced by multiple levels of signals support a continuum or a discrete EMT spectrum. While previous dose-dependent single-cell experiments with two EMT markers (E-cadherin for E, Vimentin for M) support the existence of intermediate EMT states [8, 10], much less is known about the transcriptomic profiles of the dose-dependent EMT spectrum.

In this work, we performed single-cell RNA-sequencing (scRNA-seq) using human mammary epithelial (MCF10A) cells treated with multiple concentrations of TGF-β. We found that the dose-dependent EMT program is a continuum containing multiple intermediate states that are stable after two-week treatment of TGF-β. We performed integrated analyses with our dataset and a recent time-dependent scRNA-seq dataset for the same cell line and EMT inducer [13] (**Figure 12A**). We found that the dose-dependent EMT spectrum has a stronger anti-correlation of E and M transcriptional programs than the time-dependent spectrum. While both spectrums show strong cell-to-cell variability and continuum-like patterns, the dose-dependent dataset has higher separability in terms of the groups of cells with neighboring labels (similar doses vs. similar time points). These differences enable higher sensitivity for the dose-dependent model to detect non-

canonical EMT genes that are associated with the core EMT programs in terms of the expression pattern. Furthermore, the time-dependent dataset contains unique cell clusters at E-low region in the transcriptomic space, which correspond to *en route* cell states that do not appear at steady state. We found that signature genes in both dose- and time-enriched clusters are useful for prognostic predictions of cancer patients. Our analyses revealed key differences between dose- and time-dependent EMT programs in terms of the underlying dynamical processes, and showed the widespread existence of stable EMT continuum under multiple assumptions that may be relevant to physiological and pathological conditions.

Methods

Cell culture

MCF10A cells were obtained from ATCC and grown in DMEM/F12(1:1) medium with 5% horse serum, epidermal growth factor (10 ng/mL), cholera toxin (100 ng/mL), and insulin (0.023 IU/mL). For TGF-β treatment, cells were incubated with indicated concentrations (**Figure 12B**) of human TGF-β1 protein (R&D systems) in the complete culture medium. The culture medium was replaced daily, and cells were passaged right before reaching full confluency.

Single-cell RNA-sequencing

MCF10A cells were first labelled with Perturb-seq vectors without sgRNA expression using guide barcodes (GBCs) that were originally used to identify sgRNAs [16]. Barcoded MCF10A cells were then treated with different dosages of TGF-β for 14 days and single cells were prepared and mixed at a concentration of approximately 1,000 cell/μL. Transcriptome library generation was performed following the Chromium Single Cell 3′ Reagents Kits v2 (following the CG00052 Rev

B. user guide) where we target 10,000 cells per sample for capture. GBC library was generated from a fraction (5ng) of amplified whole transcriptome by dial-out PCR method according to a previous publication [16]. Both libraries were mixed at 9:1 ratio and sequenced by paired-end sequencing (26bp Read 1 and 98bp Read 2) with a single sample index (8bp) on the Illumina HiSeq 2500. Generated FASTQ files were aligned utilizing 10× Genomics Cell Ranger 2.1.0. Each library was aligned to an indexed hg38 genome using Cell Ranger Count. The cell barcode (CBC)-GBC table was generated from the GBC library and used to identify the treatment groups. We first performed doublet exclusion with standard 10X pipeline based on RNA amounts, and then excluded cell barcodes that were assigned with multiple GBCs (12.7% of all cells assigned with GBCs) [16]. Nonetheless, we examined the 2402 unlabeled dose samples in the dose data set, which are roughly evenly distributed across UMAP and E- and M-score space (**Figure S1A**) (see the method for computing the scores below). We also compared the average E and M-scores of labeled and unbaled samples in each nearest neighbor cluster and found them to be similar (**Figure S1B**) and excluding them from calculating E and M-scores showed little affect ($R > 0.999$ with original scores). These results suggest that the unlabeled cells are merely missing annotation rather than contamination or doublets, which are consistent with prior observations of GBC labeled data [16] and thus were retained for processing and scoreing. Additionally, following the identification of multiple putative end states in both dose and integrate data, we also looked at potential homotypic doublets with individual treatment samples. To do this, we used DoubletFinder [17, 18]. Using 2.4% for time data which targeted 3000 cells per treatment, and 8.0% for dose data which targeted 10,000 cells adjusted for the proportion of each labeled dose treatment in the dataset to estimate the homotypic rate, we found 90 putative homotypic doublets in dose data, which are broadly distributed in integrated data (**Figure S2A),** and 332 in the time data, which showed some

clustering around the edge of one cluster **(Figure S2B).** However, excluding these doublets did not affect the pattern of separation of treatment samples, anti-correlation of E- and M-scores, or enrichment of time and dose samples in quartile clusters. Based on these observations, we chose not to filter these samples.

Data processing and integration

Sequencing data for time-course single-cell data from Deshmukh et al. [13] was obtained from National Center for Biotechnology Information Sequence Read Archive (BioProject ID: PRJNA698642) and mapped using to the same human genome assembly as our dose (CRXh38.84) data using Cell Ranger [19]. Aligned sequences we processed using the Seurat package (version 4.0.2) in R [20]. Gene names between experiments were correlated using the HGNChelper package [21], using the suggested gene symbol for each gene except when it would create a duplicate reference. Genes were filtered from individual runs if they did not appear in three or more cells. We then filtered each dataset for cells with fewer than 500 features or more than three median absolute deviations beyond the median number of features of the sample set (i.e., the long time, short time, and dose datasets). We additionally eliminated any sample with a fraction of mitochondrial reads that was greater than 0.2. We then integrated all time and dose data following the procedure used in Deshmukh et al. to best preserve the relationship between time samples observed in the study [13]. Briefly, the integration involves identifying the most variable genes in each dataset, defining ‘anchor’ samples between datasets using canonical correlation analysis, correcting the expression of related anchor samples and finally propagating this correction to other samples based on the similarity to the anchors. We normalized and calculated cell cycles scores for each dataset independently prior to integration and used the 15000 most variable genes to identify anchors. We applied the same top 15000 variable genes filter to unintegrated dose data for

calculating UMAP, nnPCA scores and clustering. Finally, each dataset was scaled across genes and the expression was corrected for the total read count, percent of mitochondrial reads, and cell-cycle phases using S and G2M scores. The importance of correction for cell-cycle phase can be seen by contrasting the UMAP and nearest neighbor clustering of dose data with and without cell-cycle score correction (**Figure S3**).

Projection of single cell data in reduced dimensional space

Projection of single cell data was done using the scaled expression values for both unintegrated and integrated data. For this process we employed three main approaches, Uniform Manifold Approximation and Projection (UMAP), average Z-scores, and non-negative Principal Component Analysis (nnPCA). For UMAP, we performed principal component analysis (PCA). We then used the first 15 principal components to construct the map using the RunUMAP function from Seurat with defaults parameters beyond specifying the PCA input and size. UMAP was primarily used to assess the continuity of both unintegrated and integrated datasets. Average Z-score was determined using the gsva package [22]. nnPCA-based scores were computed by performing non-negative principal component analysis on subsets of the scaled dataset defined using epithelial (E) and mesenchymal (M) genes identified by Tan et al. [23]. We used the nsprcomp function from the R package of the same name with the option nneg=TRUE, ncomp=25, em_maxiter = 10000, em_tol = 0.00001 and a consistent seed (set.seed(5)) to identify the top twenty five components for each subset [24]. While the first twenty-five components were calculated, as the function greedily optimizes the variance explained by each component in order, we initially considered only the top E and M components and added the second M component when clustering suggested the possibility of multiple paths of progression. Further details about our nnPCA scoring approach can be found in Panchy et al. [25]. In addition to E and M axes, we analyzed the relationship between

EMT progression and PD-1 ligation. The latter was represented by the gene set M4534 from the Molecular Signature Database [26].

Signature genes in EMT spectrum

To examine the distribution of time and dose samples, we divided E and M-score space into a 4×4 grid based on the 25th, 50th and 75th percentile of E and M-scores. To obtain results comparable to progression plots in **Figure 2**, enrichment only considered labeled time dosage samples. Odds of the enrichment of time and dose samples in each segment of the 4×4 grid of E and M-score space were calculated using the Fisher's exact test implemented in R (fisher.test). Differentially expressed marker genes for each segment were identified using the FindAllMarkers function from Seurat. As recommended, for FindAllMarkers we used unintegrated, normalized counts as integration introduces non-independence between samples. In FindAllMarkers, we used the logistic regression framework ('LR') so we could use the number reads per cell, percent mitochondrial genes and cell cycles scores as latent variables as well as the experiment of origin (dose, time long, and time short) to account for batch effect. We also manually corrected *p*-values to account for comparing all possible genes (i.e. those with non-zero expression in at least one sample including retesting EMT genes discussed below) across all clusters, as FindAllMarkers does not account for testing multiple clusters. We used the Bonferroni Correction, which is the method used by FindAllMarkers, with a cutoff of 0.05 along with requiring absolute value average Log2 fold-change > 0.25 (the standard setting of FindAllMarkers). Because EMT genes were used to construct cell clusters, raising a possible concern of a bias due to feature reuse, we iteratively removed each EMT gene from the list, recalculated EMT scores, and looked at differential expression for the removed gene in a clustering context agnostic to its influence. Overall, E and M scores were robust to the loss of any individual genes ($R > 0.99$), but cluster membership vary for

marginal samples around the borders of clusters, so we filtered our EMT markers genes for significant differential expression in clusters agnostic to it. This led to the removal of 12, 14 and 33 EMT marker genes from all, time and dose clusters respectively, representing 1.7%, 1.8% and 6.7% of EMT marker genes. Notably we did not apply any average Log fold-change threshold to gene agnostic results as a primary concern was biasing *p*-values due to feature reuse. Finally, the same function was also used to compare VIM, FN1, EPCAM, and FN1 across all samples, and between time and dose samples in the middle 50% of E and M scores, except we use the options "logfc.threshold=0, min.pt=0, return.thresh=1" to return all results for plotting fold change FC values. Additionally, batch correction was excluded for comparison of time and dose data.

GO Enrichment

GO enrichment of identified genes set was performed using the clusterProfiler package in R [27], using a background of all expressed genes in the unintegrated data, and significance was assessed using the Benjamini-Hochberg adjusted *p*-value with a cutoff of 0.05. Otherwise, we used default parameters which filters overly narrow (<10 genes) and overly board (>500 genes) sets as well as terms which do not overlap with the gene set of interest. Each combination of gene set and ontology (Biological Process, Cellular Component, and Molecular Function) were tested independently.

Prognostic models

TCGA bulk RNA-seq data and associated meta-data were obtained from TCGAbiolinks [28]. Of the 33 available datasets (cancer types), we selected those with at least ten patients in both the survivor and non-survivor groups, which gave rise to the 27 cancer types listed in **Figure 5B**. For each dataset, FPKM expression values were extracted from all non-normal tissue samples, filtered for duplicate samples from the same aliquot, and normalized to Log plus 1 TPM values. We then

calculated overall survival as our survival metric, which is defined as the time from date of diagnosis to the date of death (for non-survivors) or date of last follow up (for survivors), the latter being right censored in our model. Samples with missing or non-positive values for survival time were dropped as the modelling approach does not tolerate these values. For features, we used the set of up-regulated marker genes (a fold change in expression > the 75th percentile of all significant marker genes across all segments at an adjusted p-value < 0.05) in each segment. Each combination of cancer expression data and gene set was then used to construct Cox proportional hazard models using glmnet implemented in R. A glmnet function employed Lasso regularization (alpha = 1). The function optimizes the lambda parameter which influences the strength of regularization. For each lambda value, a 10-fold cross validation was performed, and the performance was measured by C-index, which is the proportion of concordant pairs to total pairs in the dataset, i.e. the proportion of all possible samples pairs where increased model hazard corresponds to reduced survival. For each model, we used the default range and interval of the tested lambda values determined by glmnet. In theory, the stronger Lasso regularization should reduce the number of active predictors (genes) in the model, but we do not actively seek the minimized predictors. Instead, we report the model which maximizes the average cross-fold C-index. This measure is roughly analogous to AUC-ROC in a censored data context such as survival.

Clustering from shared-nearest-neighbor graphs

To confirm that the associations we found in E- and M-score space reflect the full dataset, we examined sample associations across the 15000 most variable expressed genes in the integrated and dose only dataset using the clustering based on shared-nearest-neighbor graphs via Seurat. After scaling and correcting single cell data, principal component analysis was done using RunPCA with 15 components (npcs=15) and then passed to FindNeighbors with options

reduction='pca', dims=1:15. Finally, FindCluster was run with resolution = 0.5 and default parameters and seeding otherwise. Comparisons between clusters we done using the Jaccard package in R which implements tests from Ref [29].

Results

A single-cell transcriptomic landscape of dose-dependent EMT reveals a continuum-like spectrum

To characterize the transcriptomic spectrum with multiple levels of EMT signals, we performed dose-dependent induction of EMT with MCF10A cells and analyzed cells at a near-steady-state time point (14 days after TGF-β treatment) using scRNA-seq (**Figure 12A**, **red box**). Transcriptomic profiles of 8876 cells with dosage annotation were identified after a standard filtering process and each condition of TGF-β concentration (dose) yielded more than 800 cells. We found that cells treated with various concentrations of TGF-β showed a continuous spectrum when visualized in the low-dimensional Uniform Manifold Approximation and Projection (UMAP) space (**Figure 12B**). To visualize the transcriptomic variability with interpretable, functional space, we used a recently developed projection method based on nonnegative principal component analysis (nnPCA) [25]. Previously identified epithelial-associated genes (E-genes) and mesenchymal-associated genes (M-genes), of which 203 E- and 137 M-genes were present in the processed dataset, were used to construct low-dimensional space using the first principal component of E scores and either the first (M PC1, **Figure 12C**) or the second (M PC2, **Figure 12D**) principal component of M-scores [23]. We observed a progression of MCF10A cells from E-high-M-low state to E-low-M-high state with increasing concentrations of TGF-β. The effect of the progression was saturated as higher concentration of TGF-β treatment overlap in both E- and

M-scores (**Figure 12C-D**), such that our scores and significantly separated treatments of 100 pM and below, but not above **(Table 1).** Furthermore, we observed the continuous progression of key E-genes (e.g. CDH1) and M-genes (e.g. VIM and FIN1) expression (**Figure 12E-G**). The continuity of the transition and the saturation of the progression were similar to time-dependent EMT in MCF10A cells reported recently [13]. Additionally, we found that the expression of VIM and FN1 followed the M PC1 and M PC2 scores, respectively, when mapped to UMAP space, and both regions are distinct from the high E-score/EPCAM region on the opposite end of UMAP space (**Figure 12H-J**). Furthermore, these regions are reflected in the nearest neighbor clusters of dose samples (**Figure S4A**), which also showed a progression from E-high M-low to E-low M-high in the E- and M-score space (**Figure S4B**). The divergence of the mesenchymal-like cell states represented by these two clusters appears to be driven by a difference in high (≥100pM) dose cells (Cluster 0) and predominantly lower dose cells in G1 (Cluster 2), which may have undergone cell cycle arrest [30] (**Figure S4C-D**). Notably, neither dimensionality reductions nor nearest neighbor clusters depend on dose labels, yet all approaches showed a progression of the expression of EMT genes with multiple possible end states. Overall, our results show that the near-steady-state EMT program of MCF10A cells has a continuum-like spectrum that is independent of projection methods and sample labels.

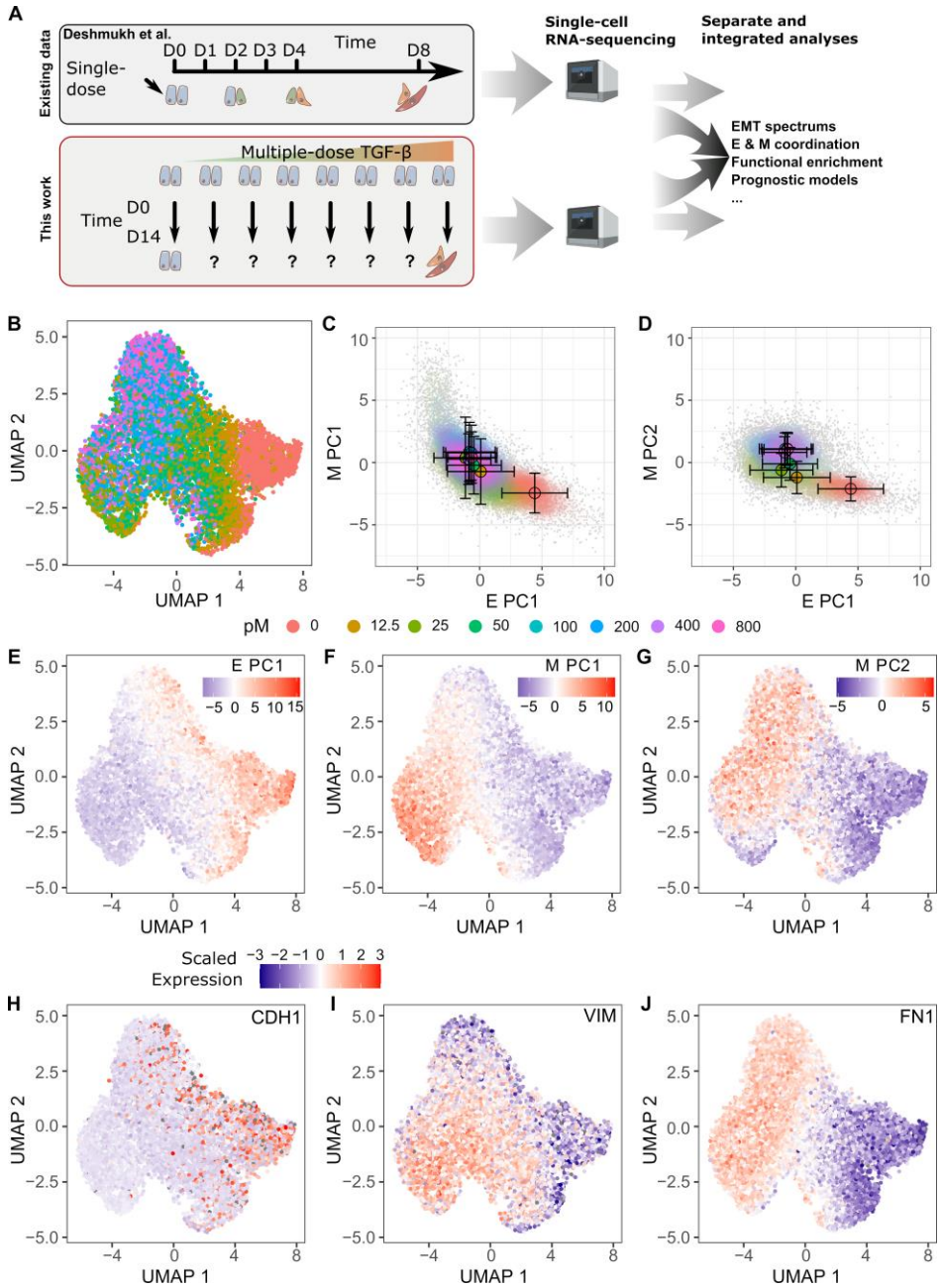


Figure 6: Analysis overview and progression of dose-dependent EMT at single-cell level. (A) A schematic of our analysis in this study. The analyses involve the existing time-course data (top) contain MCF10A cells at different time points following a common TGF-β treatment (about 200 pM; Deshmukh et al. [13]) as well as the dose-dependent data (bottom) containing MCF10A cells treated with different dosage levels of TGF-β after a fixed time period representing near-steady-state. Gene expression of cells from both experiments were measured using single-cell RNA-sequencing and were subsequently used individually and integrated for downstream analyses. **(B)** Projection of dose treatment single-cell expression data using UMAP. The color of individual points indicates the dose of TGF-β treatment from 0 pM (red) to 800 pM (pink). **(C-D)** Contour plots of gene set scores of E (x-axis) and M (y-axis) genes using M PC1 (C) and M PC2 (D) for M-scores. Color indicates the dose of TGF-β as in (B). Circles indicate the mean E- and M-score of samples from each dose point and the associated error bars show the standard deviation (**E-G**) Overlay of nnPCA-derived scores from the first E (E PC1, E), first M (M PC1, F) and second M (M PC2, G) principal components. The color of individuals points indicates the score from low (blue) to high (red). (**H-J**) Overlay of the scaled expression of EMT marker genes CDH1 (an epithelial marker, H), VIM (a mesenchymal marker, I), and FN1 (a highly expressed mesenchymal gene, J). The color of individual points indicates the Z-score of expression of each gene from low (blue) to high (red).

Table 7. Common language effective size of separation of dosage groups based on EMT scores

Comparison	E PC1	M PC1	M PC2
0 vs. 12.5 pM	0.879*	0.700*	0.706*
12.5 vs. 25 pM	0.638*	0.601*	0.624*
25 vs. 50 pM	0.403	0.474	0.609*
50 vs. 100 pM	0.551*	0.641*	0.743*
100 vs. 200 pM	0.474	0.495	0.498
200 vs. 400 pM	0.528	0.425	0.438
400 vs. 800 pM	0.461	0.522	0.554*

* Indicates statistically significant separation based on Mann-Whitney U-test ($p < 0.05$ after Bonferroni correction)

Dose-dependent states show less variation and are more separable than time-dependent states in the E/M space

We next compared the dose-dependent EMT and time-dependent EMT at the single-cell resolution. We first integrated our dose-dependent dataset (abbreviated as dose dataset) with the previously published time-dependent dataset (abbreviated as time dataset) using the same approach as in Deshmukh et al. [13] (see **Methods)**. Because the time-dependent dataset contains two sets of experiments (short-time and long-time conditions), the integration involved three experiments. While comparison can be performed with unintegrated datasets, the integration is helpful as it reduces the effect of experimental batches not associated with the meaningful biological differences and increases the similarity between common cell states. As expected, we found that the integrated data contains cells from all experiments distributed in one region more continuously (**Figure 2A**) compared to the same analysis of unintegrated data (**Figure S5**). Furthermore, the untreated control in the dose-dependent dataset and the Time The 0 control in the time-dependent

datasets were located in similar regions in the expression space (**Figure 2A**, middle and right panels). These results suggest that the datasets can be compared in a reasonably uniform framework.

The data integration resulted in some alteration of the gene expression values and coverage (207 E and 144 M genes), but the progression of EMT in the dose-dependent manner was preserved. In fact, comparing dose samples between dose only and integrated data, the first E and M components in each data set were highly correlated (R = 0.84 and 0.81 respectively). We also found a strong correlation between dose M PC1 and integrated M PC2 (R = 0.78), as well as between dose M PC2 and integrated M PC1 (R = 0.79), but only moderate correlation between dose M PC2 and integrated M PC2 (R=0.21). As such, it is unsurprising that when we applied dose PCs to the dose samples in integrated data, we saw a similar distribution of scores (**Figure S6**) and vice versa for integrated PCs applied to dose data (**Figure S7**), particularly between dose M PC1 and integrated M PC2 as well as integrated M PC1 and dose M PC2. Likewise, we observed a similar correspondence between nearest neighbor clusters defined using each data set (**Figure S8**); in particular, dose Cluster 0 had a significant overlap with dose samples in the union of integrated Clusters 0 and 3 (Jaccard index = 0.74, *p*-value < 1e-5 based on bootstrap testing), while dose Cluster 2 had significant overlap with dose samples from the union of Clusters 1 and 10 (Jaccard index = 0.76, *p*-value < 1e-5 based on bootstrap testing). Notably, Cluster 1 is predominantly time samples, suggesting that the region identified by integrated M PC2 scores is not simply an artifact of the dose data.

For subsequent definition of E- and M-score space, we chose to use M PC1 as it represents the greatest variation in M-genes across our data and is more correlated with time (R = 0.48) and dose (R = 0.42) progression, than M2 (R = 0.16 for time and R = 0.06 for dose). In this E- and M-score

space, the time-dependent data points were distributed more broadly compared to the dose-dependent data (**Figure 2A-C**). Nonetheless, the continuous E-to-M progression was observed (**Figure 2C and E**). Note that the broader distribution of cells in the time-dependent dataset was also observed with unintegrated data (**Figure S9**). This broader distribution of time-dependent data likely reflects distinct transient trajectories during the response to the EMT signal, and the difference in these trajectories may be caused by different initial conditions at unperturbed state of MCF10A cells as well as temporal fluctuations. Dosage data maintained a similar pattern of separability as the unintegrated data with significant differentiation between conditions up to the putative saturation point between 100, though we could no longer separate 25 and 50 pM treatments as we could with dose data. Time samples, however, show greater separability between the earliest (0 to 2 days) and latest (4 vs 8 days) time points, but not interior points, which includes 3 vs 4 days, the breakpoint between the two time data batches (**Table 2**).

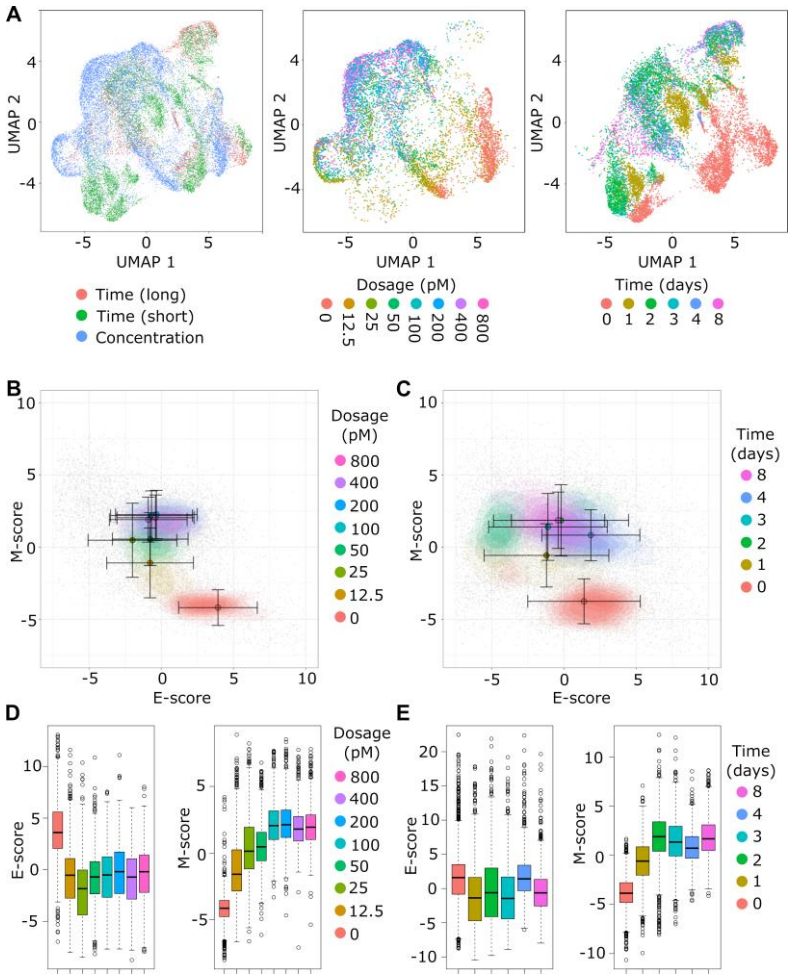


Figure 7: Continuity of integrated single cell dose and time data in low-dimensional projections. (A) Projection of integrated dose and time data using UMAP. Each panel uses the same underlying UMAP for samples, but the coloring of points has different meaning. Left panel: color indicates the origin of the sample from the Days 0, 4 and 8 of the time experiment (red), the Days 0, 1, 2 and 3 of the time experiment (green), or the dose experiment (blue). Middle panel: color indicates the treatment dose of samples from the dose experiment; time samples are masked. Right panel: color indicates the time of treatment for samples from the time experiment, dose samples are masked. **(B-C)** Contour plots of gene set scores of E (x-axis) and M (y-axis) genes using nnPCA for dose (B) and time (C) samples from integrated data. Color indicates the dose of TGF-β treatment from 0 pM (red) to 800 pM (pink) for dose data and time of treatment from 0 days (red) to 8 days (pink) for time data. Circles indicate the mean E- and M-score of samples from each dose point and the associated error bars show the standard deviation. **(D)** Boxplots show the distribution of E (left) and M (right) scores across different dose treatments from integrated data. Color indicates the dose of TGF-β as in (B). **(E)** Boxplots show the distribution of E (left) and M (right) scores across different time treatments from integrated data. Color indicates the time of treatment as in (C).

Table 8. Common language effect size of separation of time and dosage groups based on EMT scores of integrated data

Comparison	E PC1	M PC1	M PC2
Time			
0 vs. 1 days	0.605*	0.885*	0.524*
1 vs. 2 days	0.530*	0.797*	0.683*
2 vs. 3 days	0.502	0.418	0.445
3 vs. 4 days	0.469	0.371	0.478
4 vs. 8 days	0.593*	0.605*	0.569*
Dosage			
0 vs. 12.5 pM	0.887*	0.892*	0.833*
12.5 vs. 25 pM	0.616*	0.685*	0.603*
25 vs. 50 pM	0.363	0.528	0.420
50 vs. 100 pM	0.472	0.753*	0.516
100 vs. 200 pM	0.468	0.524	0.478
200 vs. 400 pM	0.546*	0.434	0.513
400 vs. 800 pM	0.455	0.533*	0.474

* Indicates statistically significant separation based on Mann-Whitney U-test ($p < 0.05$ after Bonferroni correction)

In addition to the difference in the width of the distribution, the dose-dependent dataset had stronger anti-correlation between E and M scores (R = -0.495) than the time dataset (R = -0.307) **(Figure 2B and C**). We hypothesized that the stronger coordination between E and M transcriptional programs can facilitate the discovery of EMT-associated genes within our integrated dataset that were not classically considered EMT genes in previous studies [23, 31]. Indeed, with the same threshold of Pearson correlation coefficient (0.25), the dose-dependent dataset revealed greater numbers of genes that had association (correlation with M-scores and anti-correlation with E-scores or vice-versa) with the overall expression of E and M genes (167) compared to the time-dependent dataset (41) (**Figure 3A and B**). The genes that showed association with E-program in the dose-dependent dataset but not in the time-dependent dataset were enriched with GO terms such as keratinization, keratinocyte differentiation and epidermis development, while M-correlated genes included those associated with integrin, chemokine binding, and neuronal components like the dendrite and protosynapse (**Supplemental Tables S1**

and S2) while there were too few (0 E and 2 M) found only in the time data to test. Notably, TGF-β has been shown to be involved in keratinocyte growth arrest [32], which is consistent with keratinocyte differentiation being associated with the E-program. We then focused on specific genes there were specifically correlated with EMT progression in dose data, but not time data. FARP1 was among the genes most highly correlated M-scores scores (R = 0.60) in dose data, but not in time data (R = 0.18). FARP1 was recently shown to be important for cancer cell motility and associated with poor prognosis [33]. Similarly, ESM1 was highly correlated with M-scores in dose (R = 0.47) but not time (R = 0.10), and it contributes to the metastasis in colorectal cancer via NK-κB activation [34]. In contrast, we saw comparable anti-correlation between E-scores and M PC2 in dose (R = -0.764) and time (R = -0.666), with 212 anti-correlated genes in dose and 102 in time using the 0.25 threshold. Among these genes, those unique to dose data showed similar, but fewer significantly overrepresented GO terms (**Supplemental Tables S3 and S4**). M-program genes from time data were enriched in a variety of metabolic processes, including those related to fatty acids, alcohol, and steroids, while E-program genes from time data were found in actin filament and cell-cell junction terms, but this represents only 2 or 3 of 7 genes **(Supplemental Tables S5 and S6).** We found two genes correlated with M PC2 in time data, but not dose, SLC20A1 and HPCAL1, which are involved in Wnt signaling in tumor cells, though in this context SLC20A1 showed antagonism to mesenchymal markers without affect epithelial markers, suggesting a possible atypical/intermediate EMT state [35, 36]. Finally, we considered anti-correlation between E and M-scores at individual treatment points in dose **(Figure 3E**) and time (**Figure 3F**). For M PC2, values were relatively similar across treatments and datasets, consistent with the higher overall anti-correlation for this principal component. Dose and time data show more distinctions for M PC1: dose data have the strongest anti-correlation for low dose (<50 pM)

after which it saturates, while anti-correlations in time samples are strong at the end points (0 and 8 days), but weak for the intermediate samples. This would suggest dose treatment achieves the greatest anti-correlation prior to saturation, while time treatments are most anti-correlated at end points of no or protracted treatment. However, care must be taken with interpreting these results as the presence of multiple path or end points of EMT suggested by our data, indicates these results may be influenced by Simpson's paradox compared to the overall correlation scores which intentionally consider the variation across multiple treatment populations. Note that the correlation analyses (**Figure 3A and B**) were performed in the absence of dose and time treatment labels, so the correlations were primarily driven by the intrinsic EMT continuum. Our results suggest that compared to the time-dependent data, the dose-dependent scRNA-seq data may provide higher sensitivity to detect non-classical EMT genes that are coordinated by the core EMT module.

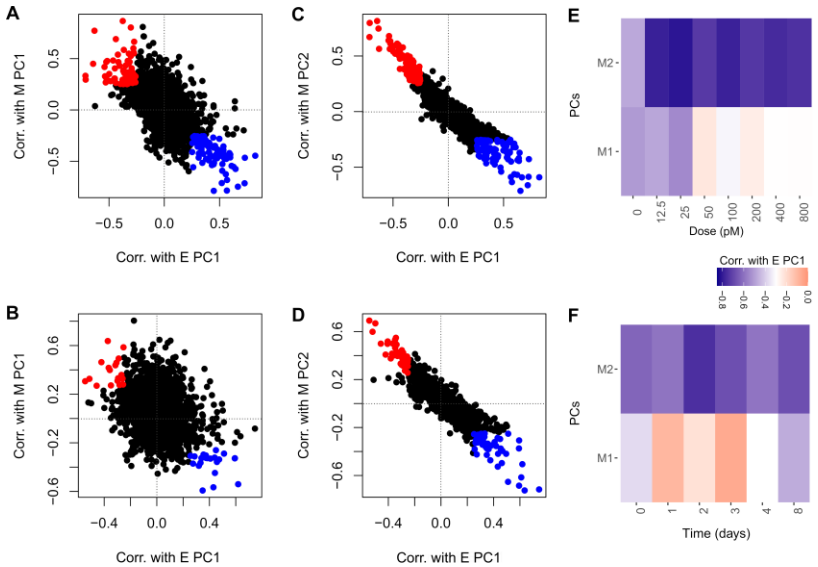


Figure 8: Correlation of non-classical EMT genes with E and M scores in integrated data. (A-B) Correlation of non-EMT gene expression with sample E-scores (x-axis) and M Pc1-scores (y-axis) in dose (A) and time (B) samples from the integrated dataset. Each point represents one of the 15000 most variable genes from the integrated single cell dataset, excluding those genes used to define the E- and M- scores. Blue points indicate genes which are anti-correlated with EMT progression ($R > 0.25$ for E-scores, < -0.25 for M-scores), while red points indicate genes which are correlated with EMT progression ($R < -0.25$ for E-scores, > 0.25 for M-scores). (**C-D**) Correlation of non-EMT gene expression with sample E-scores (x-axis) and M PC2-scores (y-axis) in dose (C) and time (D) samples from the integrated dataset. Labels and colors are defined as in (**A-B**). (**E-F**) The correlation between E-scores with M PC1 (bottom) and M PC2 (top) across different treatments in dose (E) and time (F) data. The color of each cell indicates the strength of anti-correlation from weak (red) to strong (blue).

En route cell clusters unique to time dependent EMT program

Based on the distinct distributions of cells between time- and dose-dependent EMT programs (**Figure 4A**), we wondered whether our comparative study can reveal expression regions containing cells that are *en route* to the M state rather than (partially) stabilized at intermediate EMT attractors. To simplify the representation of the expression profiles in the EMT spectrum, we focused on a 4×4 grid in the E- and M-score space. We found that a region with low scores of one (E or M) program, and low-to-medium scores of the other program (M or E) is enriched with cells in the time-dependent EMT program (**Figure 4B, lower left**). This result reveals a transient EMT path that may primarily involve the relatively low expression of both E and M genes. Nonetheless, the enrichment of time data points in the regions of high E-gene activity and M-gene activity suggested a possible alternative transient EMT path (**Figure 4B, right**) [37, 38]. We next considered the relationship between these transient paths, particularly the E-low and M-low region, with our alternative M-score. To do this, we considered the nearest neighbor clusters identified in the integrated data which coincide with high expression in M PC2 and found that Cluster 1 was broadly distributed across the lowest E-quartile and is predominantly time samples (**Figure S10A**). Cluster 10, which was mostly dose samples, is also in the lowest E quartile (**Figure S10B**), but was more concentrated in the highest M cluster (Fisher's Exact Test, odds = 1.73, *p*-value = 6.9e-12*)*, consistent with transient progression through the region. Notably, both clusters have high M PC2 expression in the integrated data (compare **Figure S6** and **Figure S8**), suggesting that while there is transience in the highest variance of component of M-genes, there is a signature of expression for these sample that has been identified by the second M principal component. Finally, we looked at differentially expressed genes across all sixteen segments of the 4×4 grid and focused first on the known EMT marker genes CDH1, EPCAM, VIM, and FN1. Interestingly,

while the transient path crossing the E-low-M-low region is generally consistent with the profiles of the E marker CDH1 and EPCAM as well as M markers VIM and FN1 (**Figure 4C-F**), a distinct sequence of M gene activations was observed in the hypothetical E-to-M path: for example, VIM was activated before FN1 in this path (**Figure 4E and F**). This observation is consistent with our earlier transcriptomic data showing significant diversity of M-genes in response to EMT signals [31]. We also considered the differential expression of these genes between time and dose samples in the middle 50% of E and M scores. We found a significant difference (adjusted p-value < 0.05) in the expression of VIM, FN1, and EPCAM, with FN1 favoring dose samples (average Log2 fold change = 2.14) while EPCAM (average Log2 fold change = 0.31) and VIM (average Log2 fold change = 1.29) favor time samples. However, the fold change of these genes in the central clusters was low compared to clusters around the extrema (**Figure 4B**). As such, while this result is consistent with the expected expression differences associated with transient EMT, it is difficult to infer a specific biological meaning to this difference as the underlying expression values are moderate.

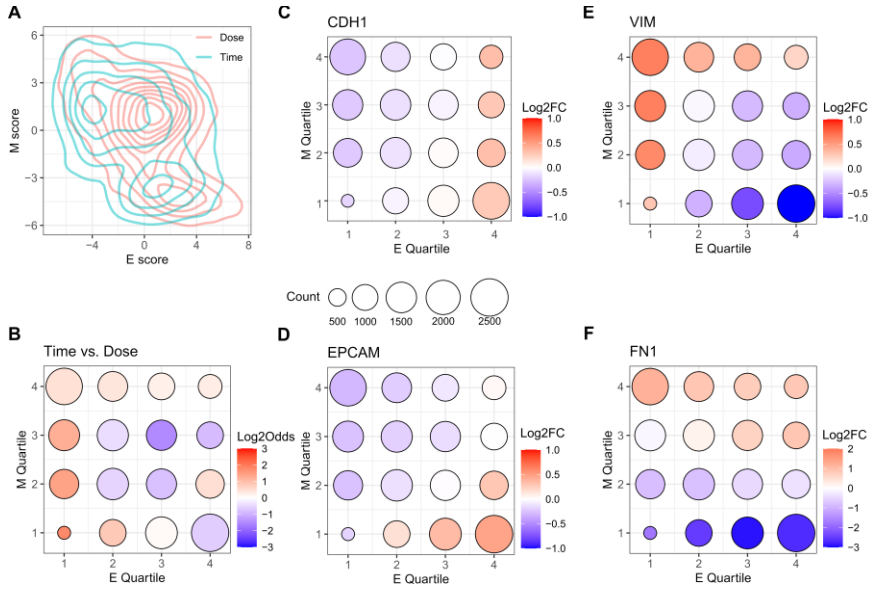


Figure 9: Enrichment of dose and time samples in E- and M-score space. (A) A contour map showing the distributions of time (blue) and dose (red) samples in E- and M-score space. **(B)** Bubble chart showing the enrichment of time samples in different segments of E- and M-score space. Each point represents a segment of E- and M-score space defined by a particular quartile of E-score (x-axis) and M-score (y-axis). The size of the point corresponds to the total number samples in the segment and the color of the point represent the log-odds of time sample enrichment from low (blue) to high (red) with a log-odds of 0 (white) indicating balanced representation. (**C-F**) Bubble charts showing the differential expression of key EMT genes, CDH1 (C), EPCAM (D) VIM (E), and FN1 (F) in different segments of E- and M-score space. Point positions and size are defined as in (B). The color of the point represents the log fold change of expression of the gene from low (blue) to high (red) with a log fold change of 0 (white) indicating no-change relative to other samples. All fold change values are shown regardless of significance.

Signature genes from both dose- and time-dependent datasets contribute to better prognostic models

To examine the roles of the signature genes at various locations of the EMT spectrum in prognosis, we again focused on the sixteen segments of the 4 × 4 grid of the EMT space containing both time and dose data. First, we expanded our search for differentially expressed genes to include all possible (non-zero expression) genes and selected those in the top 75th percentile of fold change values after filtering for significance (adjusted p-value < 0.05) and minimum average Log2 fold-change threshold (absolute value of 0.25). We then obtained 27 diverse cancer datasets with tumor RNA-seq and patient data from The Cancer Genome Atlas (TCGA). We used the signature genes of each segment to construct a Lasso penalized Cox model for predicting the survival outcomes of cancer patients for each of the 27 cancer types. Interestingly, segments with high average prognostic performance measured in C-index were primarily found in the lowest E-score quartile (**Figure 5A**, **left edge**) and highest M-score quartile (**Figure 5A, top edge**)**.** We next scaled performance index across cancer types to account for dataset level variance in model construction, (**Figure 5B**). We found that, in general, E-low intermediate states (green) and M-high intermediates (light blue) both outperformed other (blue) segments, apart from (3,4) and (1,1), with similar average performance to the extreme M state (yellow), although there was a large degree of variance across cancers and on the high end of the performance spectrum. The superior prognostic performance of the extreme M-state and certain intermediate states was also observed

in the recent study based on time-dependent EMT alone [13], and our analysis revealed the characteristics of these high-performing groups in the 2-dimensional EMT spectrum. We also considered the predictive power of signature genes defined using only time and only dose samples. We found that signature genes from time samples alone had the highest C-index in the lowest E-quartile (**Figure 5C**), while signature genes from dose samples alone had high C-index only in the high M end of the lowest E-quartile (**Figure 5D**), reflecting difference in distribution of Clusters 1 and 10 (see **Figure S10**). This suggests that gene differentially enriched in time and dose samples contribute to the predictive power of lower E-quartile sample. However, time and dose samples also showed unique prognostic values, as high E and high M signatures segments had high C-index only in time data, but high E low M signature genes had high C-index only in dose data. This suggests that there may be further differences in expression between transient and steady state EMT that were not highlighted by our M-scores because they do not represent broad variance across both data sets. Comparably, the moderate E- and M-score clusters are weak predictors in all cases, likely related to the few signature genes found there. As such, further study of the extrema of EMT expression in both transient and steady state conditions may be important for identifying prognostic predictors.

We next asked whether the comparable contributions of dose- and time- dependent EMT to the prognostic models can be partially explained by their involvement in immune evasion activity [39, 40]. We used a list for genes previously known to be upregulated by PD-1 ligation [26], as a gene set representative of immune evasion, and we found that progressions of both dose- and time-dependent EMT had moderately positive correlations with the expression of the gene set ($R \approx 0.3$). This result was expected because some M-genes such as VIM and LGALS1 are in this PD-1 upregulation set.

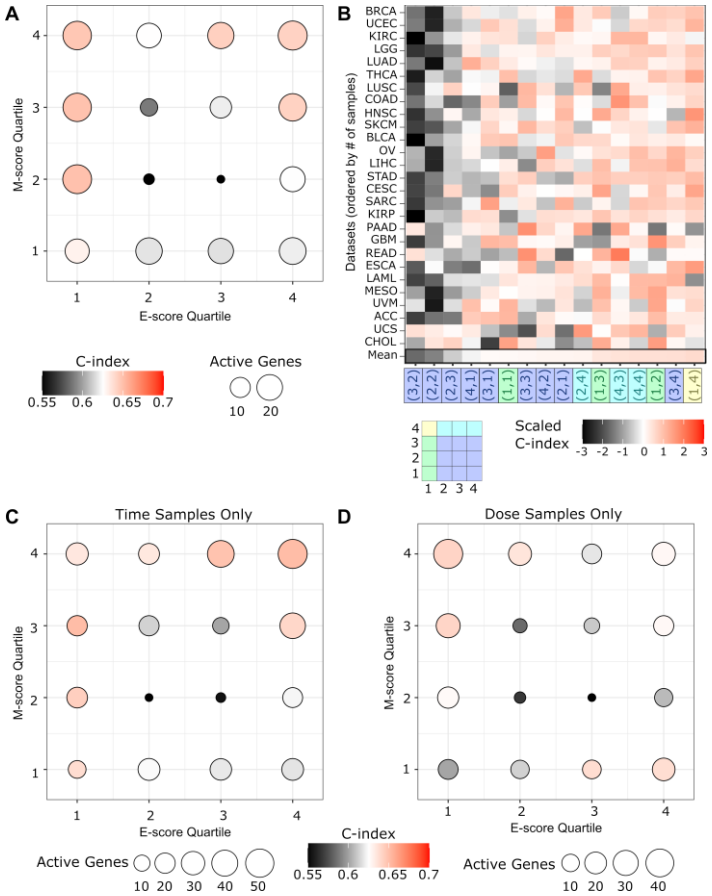


Figure 10: Performance of prognostic models. (A) Bubble chart shows the average performance of Cox hazard models built using the up-regulated genes from different segments of the E- and M-score space. Each point represents a segment defined by a particular quartile of E-score (x-axis) and M-score (y-axis). The size of the point corresponds to the average number predictor genes across all models built using the up-regulated genes from that segment. Color represents the C-index averaged across all models of cancer datasets, where C-index is a measure of the correspondence between the models predicted risk and survivorship, similar to the Area Under the Receiver Operating Curve (AUC-ROC). Red represents higher C-index (better performance). Black represents lower C-index (poorer) performance. White represents the approximate median of scores (0.645). **(B)** Heatmap of individual model performance across all combinations of segment-gene-derived models (x-axis) and cancer datasets (y-axis). The color of individual cells represents the C-index of the model scaled against all models of that cancer datasets to account for data-set-level differences in model performance. Segments are ordered across the x-axis by the increased average scaled C-index from left to right, while cancer datasets are ordered by increasing number of samples from bottom to top. The shading of segment coordinates along the x-axis distinguishes between the extreme M state (yellow), E-low intermediates (green), M-high intermediates (light blue), and other segments (dark blue). **(C-D)** Bubble charts show the average performance Cox hazard models built using the up-regulated genes identified using only time (C) or dose (D) samples from different segments of E- and M-score space. Point position, size, and color as defined as in (A).

Discussion

Time dependent EMT processes have been extensively studied with single-cell transcriptomics in recent years [3, 13, 14]. While these studies provided substantial information about EMT progression in multiple contexts, the connection between the intermediate cell states observed in

these experiments and cell attractors was elusive. Using near-steady-state single-cell transcriptomic profiling, we showed that the EMT spectrum can be described as a stable continuum under multiple levels of EMT signals. This information complements earlier studies with single-dose time course data, and it provides stronger evidence supporting the existence of multiple intermediate EMT states that widely exist in tumors and metastatic cells [5]. Furthermore, by comparing the time- and dose- dependent single-cell data, we identified groups of cells that are exclusively *en route* to M state, which shows the possibility that some cells can transiently deactivate the epithelial program before activating the mesenchymal program. Nonetheless, it is likely that many other *en route* cells are also close to the stable intermediate EMT states, and these states have intermediate levels of both E and M genes. Furthermore, given the heterogeneous microenvironment of cells under physiological conditions, the cells states may be determined by both the multi-level EMT signals and the time-dependent stages of the EMT process, so both dose dependent and time dependent *in vitro* data can contribute to the understanding of the EMT program *in vivo*.

Previous mathematical models have provided mechanistic insights into the gene network structures supporting multiple intermediate EMT states [10-12, 41-44]. While interconnected positive feedback loops involving a few genes can govern discrete intermediate states, the continuum-like, wide distributions of cells in the EMT spectrum even under the near-steady-state condition suggest that the discrete cells states may only partially explain the stable phenotypical heterogeneity. At least two mechanisms may explain the gap between the existing theories the observed continuum: realistic gene regulatory networks may contain much more positive feedback loops than those described by existing models [45], and these loops can support many intermediate states; dynamical and reversible cell-state transitions at stationary phase, such as those driven by

transcriptional noise and slow-timescale oscillations [46], can give rise to cells that are far away from point attractors in the gene expression space. While it is possible that many unknown transcriptional and post-transcriptional circuits may contribute to the numbers and dynamical properties of the EMT attractors, we cannot exclude the possibility that the technical noise in scRNA-seq experiments may mask the effects of discrete attractors. Nonetheless, existing models have provided a strong theoretical foundation for stable intermediate EMT states, and future development of the theories and quantitative experiments will be important to the understanding of EMT continuum.

EMT involves dramatic phenotypic changes of cells. It is therefore expected that transcriptome-wide alteration is induced during the multi-stage transition. This suggests that classically defined core EMT genes may not be sufficient to provide a holistic view of the EMT program. We showed that the dose-dependent single-cell transcriptomes can be useful to identify genes that show expression patterns highly correlated with the core EMT genes at the single-cell level. The dose-dependent scRNA-seq experiment can therefore provide crucial links of the core EMT networks to the rest of the transcriptomes, and some of these connections may not be revealed by time-course scRNA-seq experiments due to the large numbers of cells that transiently activate parts of the transcriptional program. Nonetheless, the two types of scRNA-seq experiments contain complementary information, and we suggest that both can be used in future studies to reveal cellular programming that determines cell-to-cell variabilities in cell populations.

Data Availability

All code is available at GitHub repository: https://github. com/panchyni/Time and Dose.

Sequencing data for dose-dependent EMT in MCF10A cells are deposited in the GEO database (GSE213753).

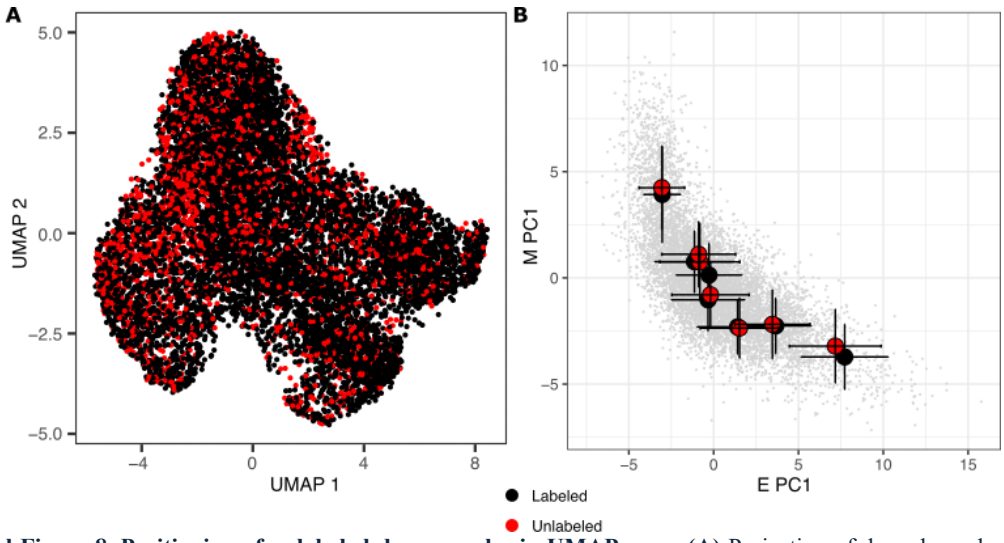
References

1. Mani, S.a., et al., *The epithelial-mesenchymal transition generates cells with properties of stem cells.* Cell, 2008. **133**(4): p. 704-15.
2. Thiery, J.P., et al., *Epithelial-mesenchymal transitions in development and disease.* Cell, 2009. **139**(5): p. 871-90.
3. Karacosta, L.G., et al., *Mapping lung cancer epithelial-mesenchymal transition states and trajectories with single-cell resolution.* Nature Communications, 2019. **10**(1): p. 5587.
4. Pastushenko, I. and C. Blanpain, *EMT Transition States during Tumor Progression and Metastasis.* Trends in cell biology, 2018.
5. Pastushenko, I., et al., *Identification of the tumour transition states occurring during EMT.* Nature, 2018. **556**: p. 463-468.
6. Kröger, C., et al., *Acquisition of a hybrid E/M state is essential for tumorigenicity of basal breast cancer cells.* Proceedings of the National Academy of Sciences, 2019: p. 201812876.
7. Grosse-Wilde, A., et al., *Stemness of the hybrid Epithelial/Mesenchymal State in Breast Cancer and Its Association with Poor Survival.* PLoS One, 2015. **10**(5): p. e0126522.
8. Zhang, J., et al., *TGF-β -induced epithelial-to-mesenchymal transition proceeds through stepwise activation of multiple feedback loops.* Science Signaling, 2014. **7**(345): p. ra91-ra91.
9. Wang, X. and J.P. Thiery, *Harnessing Carcinoma Cell Plasticity Mediated by TGF-β Signaling.* Cancers, 2021. **13**(14): p. 3397.
10. Hong, T., et al., *An Ovol2-Zeb1 Mutual Inhibitory Circuit Governs Bidirectional and Multi-step Transition between Epithelial and Mesenchymal States.* PLoS Comput Biol, 2015. **11**(11): p. e1004569.
11. Huang, B., et al., *Interrogating the topological robustness of gene regulatory circuits by randomization.* PLoS computational biology, 2017. **13**(3): p. e1005456.
12. Subbalakshmi, A.R., et al., *A computational systems biology approach identifies SLUG as a mediator of partial Epithelial-Mesenchymal Transition (EMT).* Cells Tissues Organs, 2022. **211**(6): p. 105-118.
13. Deshmukh, A.P., et al., *Identification of EMT signaling cross-talk and gene regulatory networks by single-cell RNA sequencing.* Proceedings of the National Academy of Sciences, 2021. **118**(19).
14. Cook, D.P. and B.C. Vanderhyden, *Context specificity of the EMT transcriptional response.* Nature Communications, 2020. **11**(1): p. 1-9.
15. Ramirez, D., V. Kohar, and M. Lu, *Toward modeling context-specific EMT regulatory networks using temporal single cell RNA-Seq data.* Frontiers in molecular biosciences, 2020. **7**: p. 54.
16. Dixit, A., et al., *Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens.* Cell, 2016. **167**(7): p. 1853-1866.

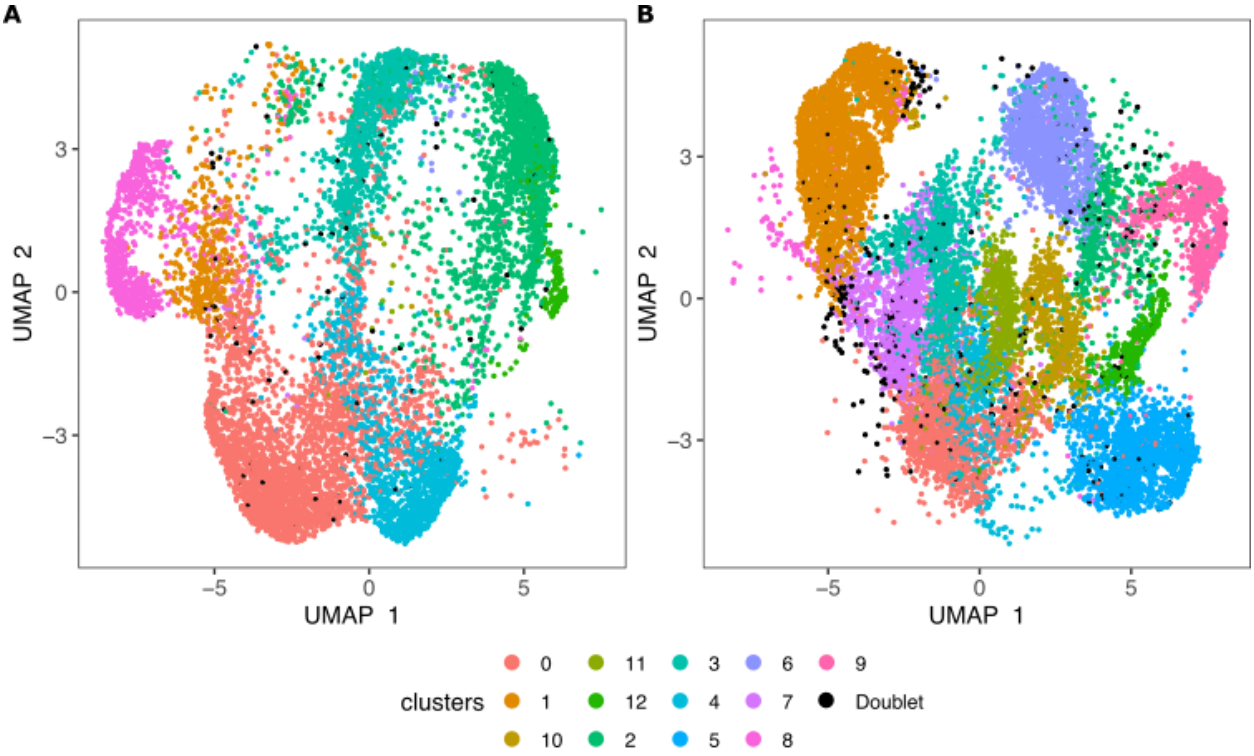
17. Xi, N.M. and J.J. Li, *Benchmarking computational doublet-detection methods for single-cell RNA sequencing data.* Cell systems, 2021. **12**(2): p. 176-194.
18. McGinnis, C.S., L.M. Murrow, and Z.J. Gartner, *DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors.* Cell systems, 2019. **8**(4): p. 329-337.
19. Zheng, G.X.Y., et al., *Massively parallel digital transcriptional profiling of single cells.* Nature communications, 2017. **8**(1): p. 1-12.
20. Hao, Y., et al., *Integrated analysis of multimodal single-cell data.* Cell, 2021. **184**(13): p. 3573-3587.
21. Oh, S., et al., *HGNChelper: identification and correction of invalid gene symbols for human and mouse.* F1000Research, 2020. **9**.
22. Hänzelmann, S., R. Castelo, and J. Guinney, *GSVA: gene set variation analysis for microarray and RNA-Seq data.* BMC Bioinformatics, 2013. **14**(1): p. 7.
23. Tan, T.Z., et al., *Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients.* EMBO molecular medicine, 2014. **6**(10): p. 1279-1293.
24. Sigg, C.D. and J.M. Buhmann. *Expectation-maximization for sparse and non-negative PCA*. in *Proceedings of the 25th international conference on Machine learning*. 2008.
25. Panchy, N., K. Watanabe, and T. Hong, *Interpretable, Scalable, and Transferrable Functional Projection of Large-Scale Transcriptome Data Using Constrained Matrix Decomposition.* Frontiers in Genetics, 2021. **12**: p. 1555.
26. Quigley, M., et al., *Transcriptional analysis of HIV-specific CD8+ T cells shows that PD-1 inhibits T cell function by upregulating BATF.* Nature medicine, 2010. **16**(10): p. 1147-1151.
27. Wu, T., et al., *clusterProfiler 4.0: A universal enrichment tool for interpreting omics data.* The Innovation, 2021. **2**(3): p. 100141.
28. Colaprico, A., et al., *TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data.* Nucleic acids research, 2016. **44**(8): p. e71-e71.
29. Chung, N.C., et al., *Jaccard/Tanimoto similarity test and estimation methods for biological presence-absence data.* BMC bioinformatics, 2019. **20**(15): p. 1-11.
30. Lee, K.-Y. and S.-C. Bae, *TGF-β-dependent cell growth arrest and apoptosis.* BMB Reports, 2002. **35**(1): p. 47-53.
31. Watanabe, K., et al., *Combinatorial perturbation analysis reveals divergent regulations of mesenchymal genes during epithelial-to-mesenchymal transition.* npj Syst Biol Appl, 2019. **5**(1): p. 21.
32. Freedberg, I.M., et al., *Keratins and the keratinocyte activation cycle.* Journal of Investigative Dermatology, 2001. **116**(5): p. 633-640.
33. Hirano, T., et al., *FARP1 boosts CDC42 activity from integrin αvβ5 signaling and correlates with poor prognosis of advanced gastric cancer.* Oncogenesis, 2020. **9**(2): p. 1-14.
34. Kang, Y.H., et al., *ESM-1 regulates cell growth and metastatic process through activation of NF-κB in colorectal cancer.* Cellular signalling, 2012. **24**(10): p. 1940-1949.
35. Zhang, D., et al., *HPCAL 1 promotes glioblastoma proliferation via activation of Wnt/β-catenin signalling pathway.* Journal of Cellular and Molecular Medicine, 2019. **23**(5): p. 3108-3117.

36. Li, J., et al., *Impact of SLC20A1 on the Wnt/β-catenin signaling pathway in somatotroph adenomas.* Molecular medicine reports, 2019. **20**(4): p. 3276-3284.
37. Wang, W., et al., *Epithelial-to-mesenchymal transition proceeds through directional destabilization of multidimensional attractor.* Elife, 2022. **11**: p. e74866.
38. Panchy, N., et al., *Integrative transcriptomic analysis reveals a multiphasic epithelial–mesenchymal spectrum in cancer and non-tumorigenic cells.* Frontiers in Oncology, 2020. **9**: p. 1479.
39. Sahoo, S., et al., *Immunosuppressive traits of the hybrid epithelial/mesenchymal phenotype.* Frontiers in immunology, 2021. **12**.
40. Dongre, A., et al., *Direct and Indirect Regulators of Epithelial–Mesenchymal Transition–Mediated Immunosuppression in Breast CarcinomasEMT and Resistance to Checkpoint Blockade Immunotherapy.* Cancer discovery, 2021. **11**(5): p. 1286-1305.
41. Font-Clos, F., S. Zapperi, and C.A.M. La Porta, *Topography of epithelial–mesenchymal plasticity.* Proceedings of the National Academy of Sciences, 2018. **115**(23): p. 5902.
42. Lu, M., et al., *MicroRNA-based regulation of epithelial-hybrid-mesenchymal fate determination.* Proceedings of the National Academy of Sciences of the United States of America, 2013. **110**(45): p. 18144-18149.
43. Jolly, M.K., et al., *Interconnected feedback loops among ESRP1, HAS2, and CD44 regulate epithelial-mesenchymal plasticity in cancer.* APL Bioengineering, 2018. **2**(3): p. 031908.
44. Tian, X.-J., H. Zhang, and J. Xing, *Coupled reversible and irreversible bistable switches underlying TGFβ-induced epithelial to mesenchymal transition.* Biophysical journal, 2013. **105**(4): p. 1079-89.
45. Nordick, B. and T. Hong, *Identification, visualization, statistical analysis and mathematical modeling of high-feedback loops in gene regulatory networks.* BMC bioinformatics, 2021. **22**(1): p. 1-21.
46. Nordick, B., et al., *Nonmodular oscillator and switch based on RNA decay drive regeneration of multimodal gene expression.* Nucleic Acids Res, 2022. **50**(7): p. 3693.

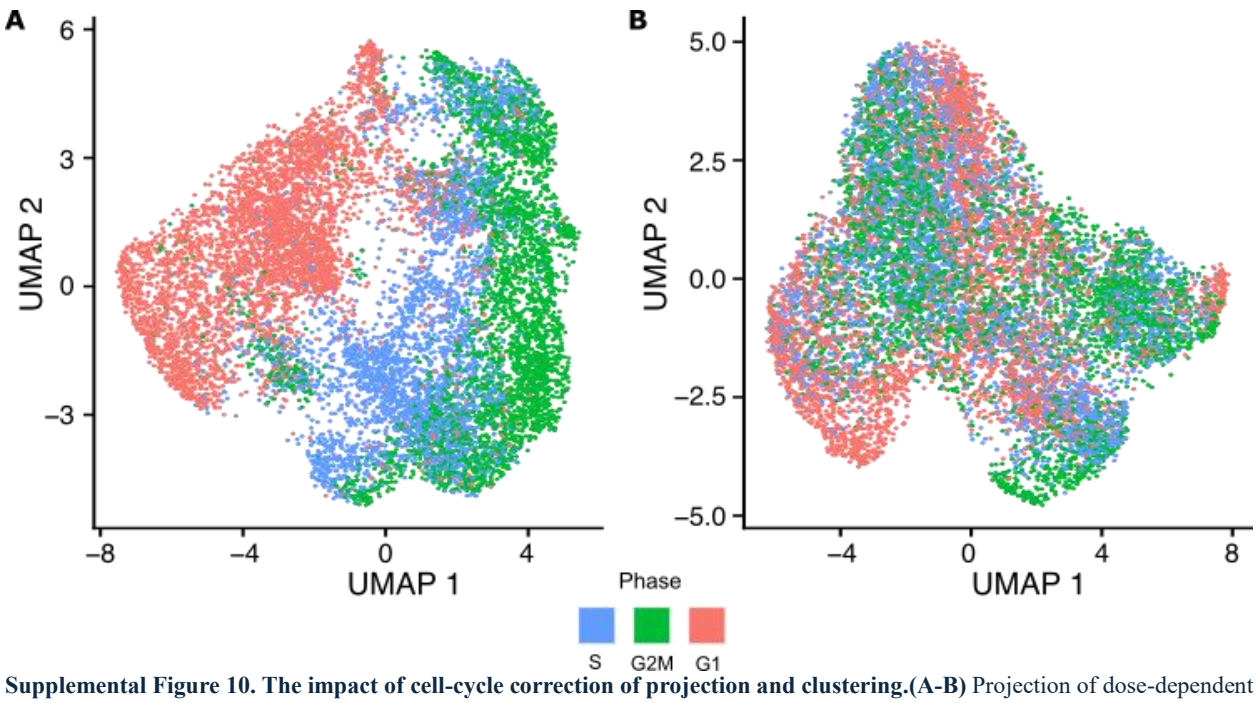
Appendix



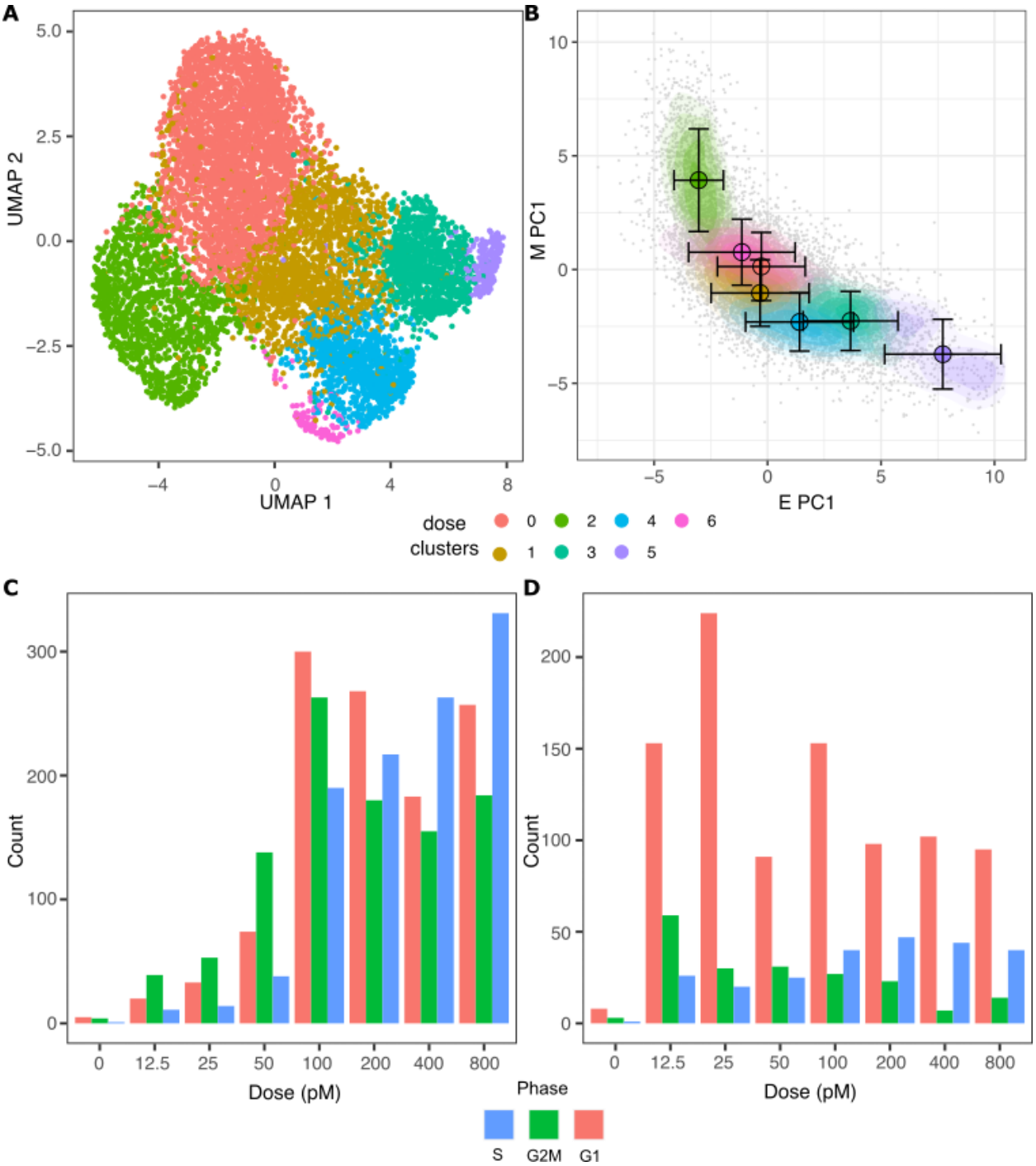
Supplemental Figure 8. Positioning of unlabeled dose samples in UMAP space.(A) Projection of dose-dependent single cell expression data using UMAP. The color of individual points indicates whether the sample has a treatment level label (black) or is missing a label (red). **(B)** Scatter plots of gene set scores of E (x-axis) and M (y-axis) genes using nnPCA for dose data. Circles indicate the mean E- and M-score of samples from each shared nearest neighbor cluster, with the red dots representing labeled samples only and the blue points representing unlabeled samples. The associated error bars show the standard deviation.



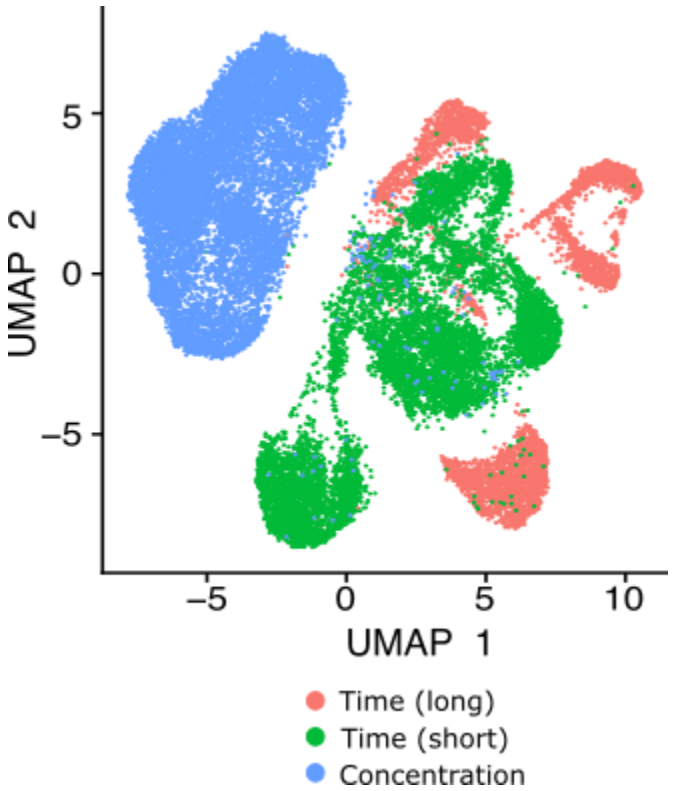
Supplemental Figure 9. Location of putative homotypic doublets in UMAP space.(A-B) Projection of labeled dose samples (A) and time samples (B) from integrated expression data using UMAP. The color of individual points indicates the nearest neighbor clusters to which is belong except for black points which indicate putative homotypic doublets detected by DoubletFinder.



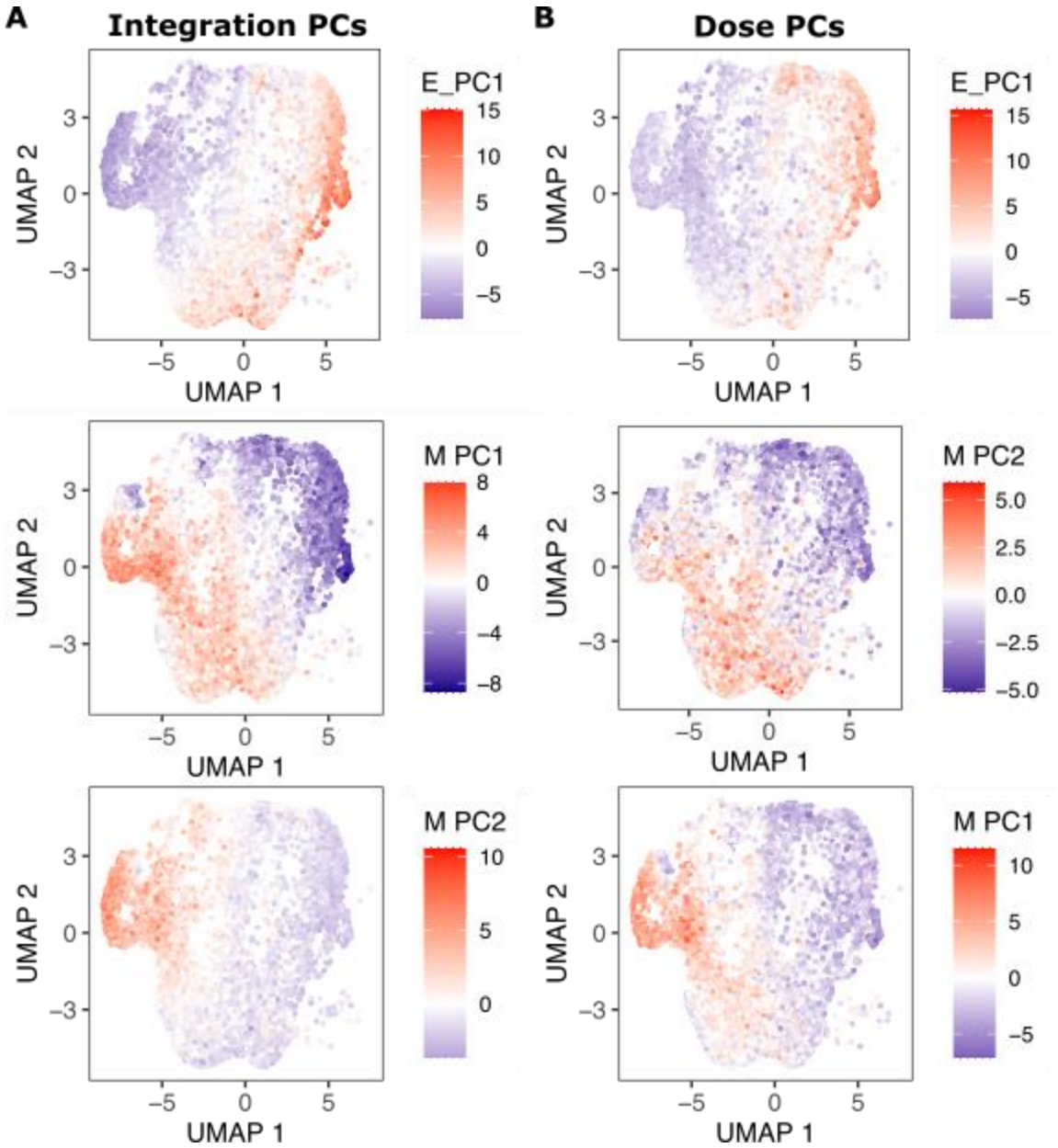
Supplemental Figure 10. The impact of cell-cycle correction of projection and clustering.(A-B) Projection of dose-dependent single cell expression data using UMAP without (A) and with (B) correction based on cell cycle scores. The color of individual point indicates the inferred cell-cycle state: red = G1, green = G2M, blue = S.



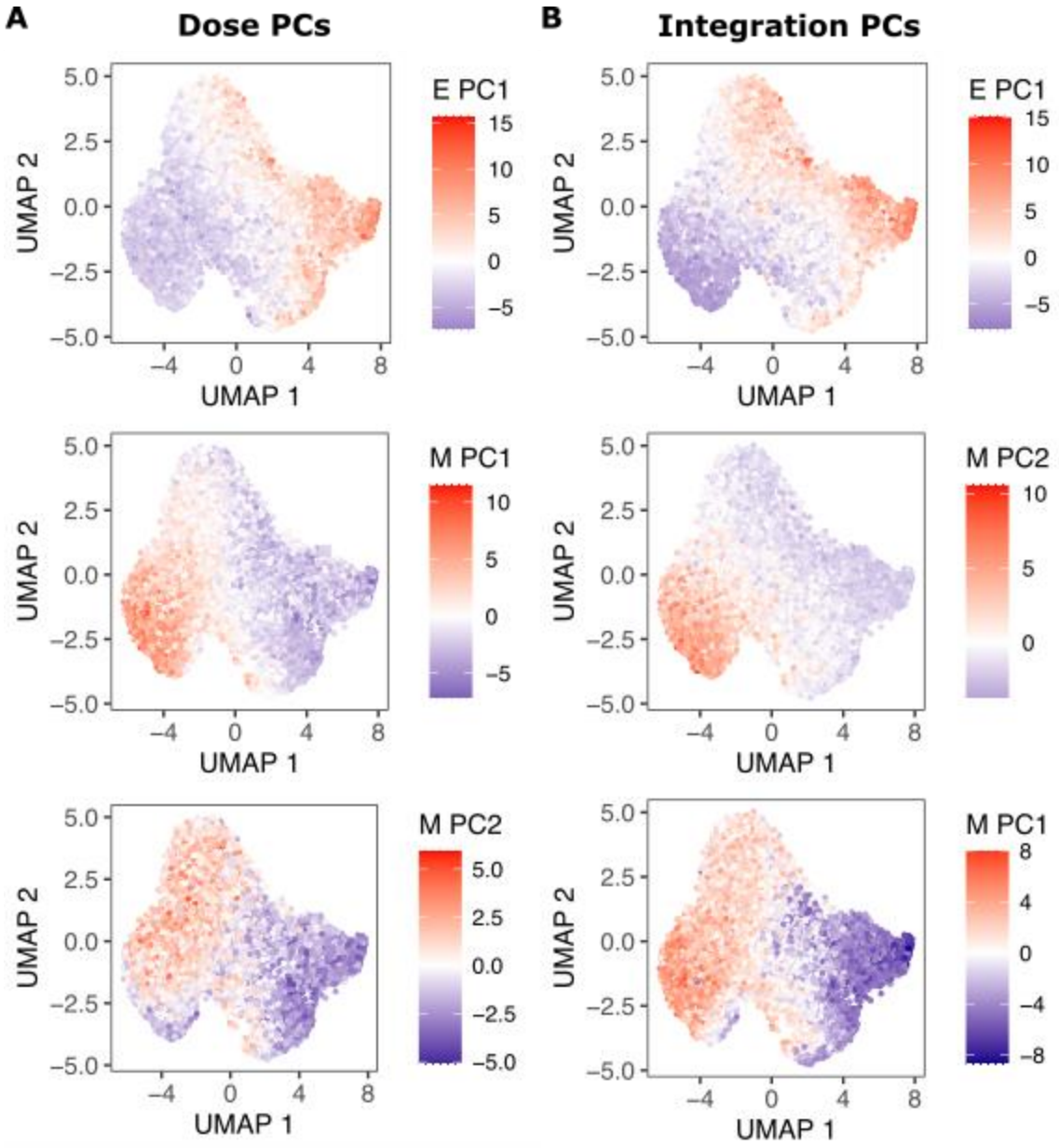
Supplemental Figure 11. Clustering based on shared-nearest-neighbor graphs in dose UMAP and E- and M-score space. **(A)** Projection of dose treatment single cell expression data using UMAP. The color of individual points indicates the clusters based on shared-nearest-neighbor graphs to which the sample was assigned **(B)** Contour plots of gene set scores of E (x-axis) and M (y-axis) genes using nnPCA for dose data. Color indicates cluster of samples in UMAP space, with individual colors corresponding directly to clusters in (A). Circles indicate the mean E- and M-score of samples from each dose point and the associated error bars show the standard deviation. **(C)** Bar charts show the frequency (y-axis) of different treatment dosages (x-axis) in Cluster 0 (left) and Cluster 2 (right). The frequency of each dosage is broken down by the inferred cell-cycle state: S (blue), G2/M (green) and G1 (red).



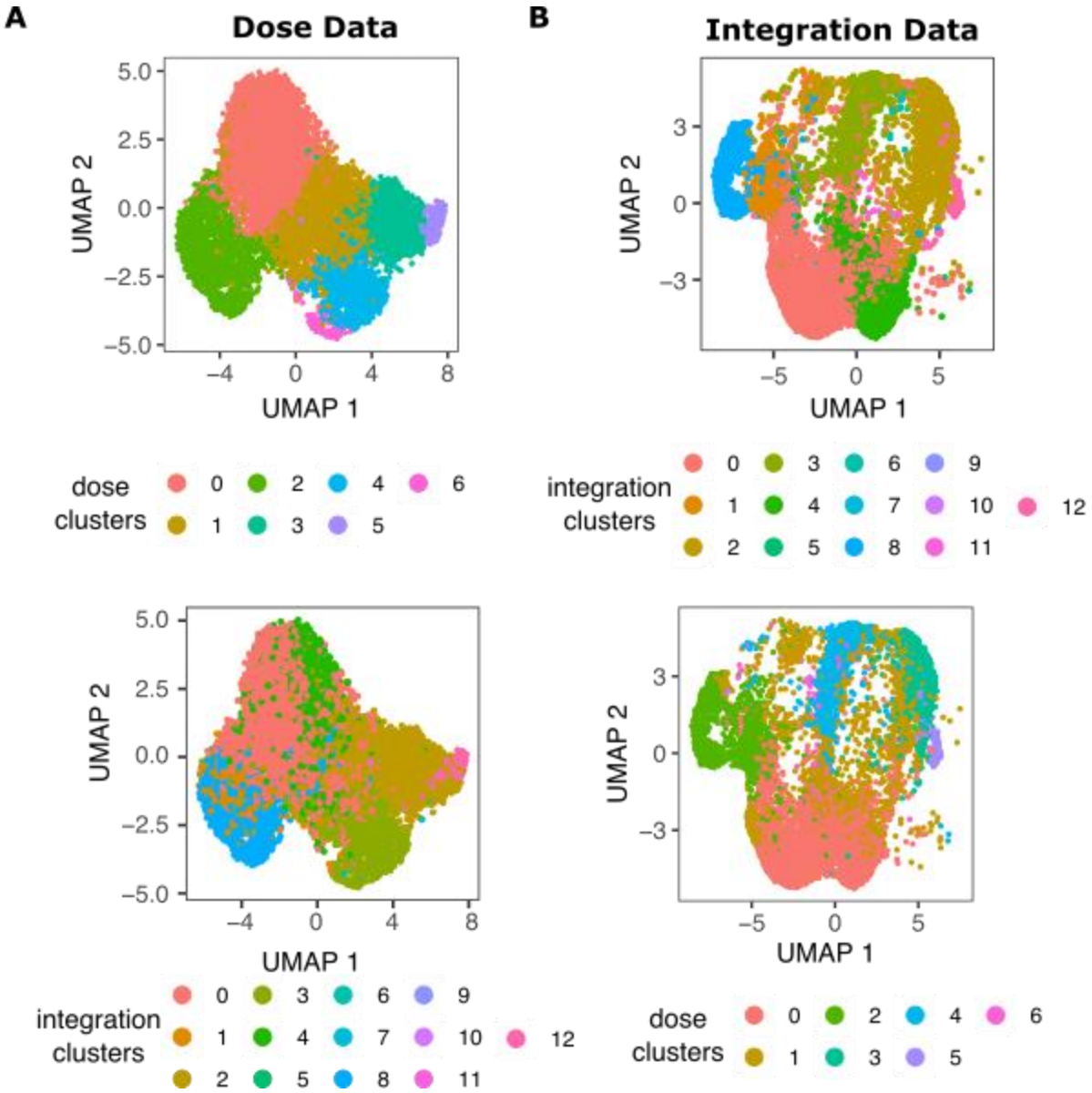
Supplemental Figure 12. Positioning of samples from different experiments in unintegrated UMAP Space. Projection of unintegrated dose and time data using UMAP. Color indicates the origin of the sample from the Days 0, 4 and 8 of the time experiment (red), the Days 0, 1, 2 and 3 of the time experiment (green), or the dose experiment (blue).



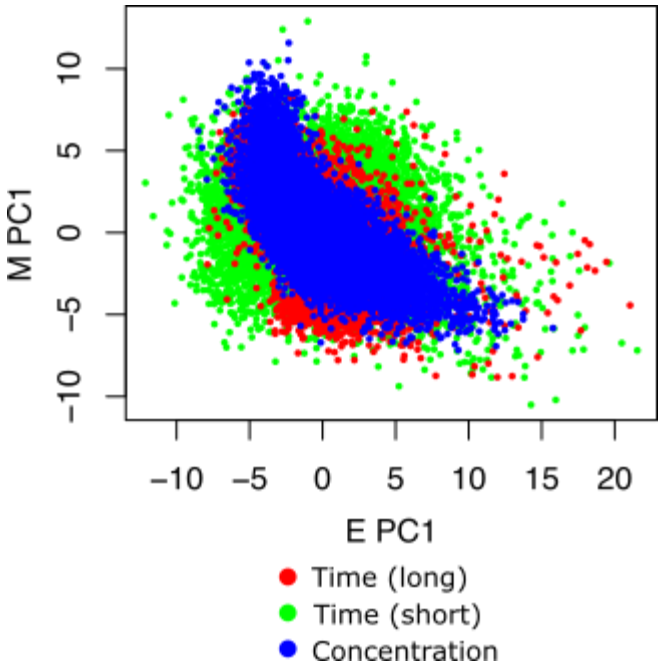
Supplemental Figure 13. Overlays of principal component scores on integrated data. (A) Projection of dose treated samples from integrated single-cell samples using UMAP which have been overlaid with scores derived from the principal components (PCs) of integrated data: E PC1 (top), M PC1 (middle), M PC2 (bottom). The color of each point indicates the value of the score from high (red) to low (blue). **(B)** Projection of integrated single-cell samples using UMAP which have been overlaid with scores derived from the components (PCs) of the unintegrated dose data: E PC1 (top), M PC2 (middle), M PC1 (bottom). Colors are as in (A).



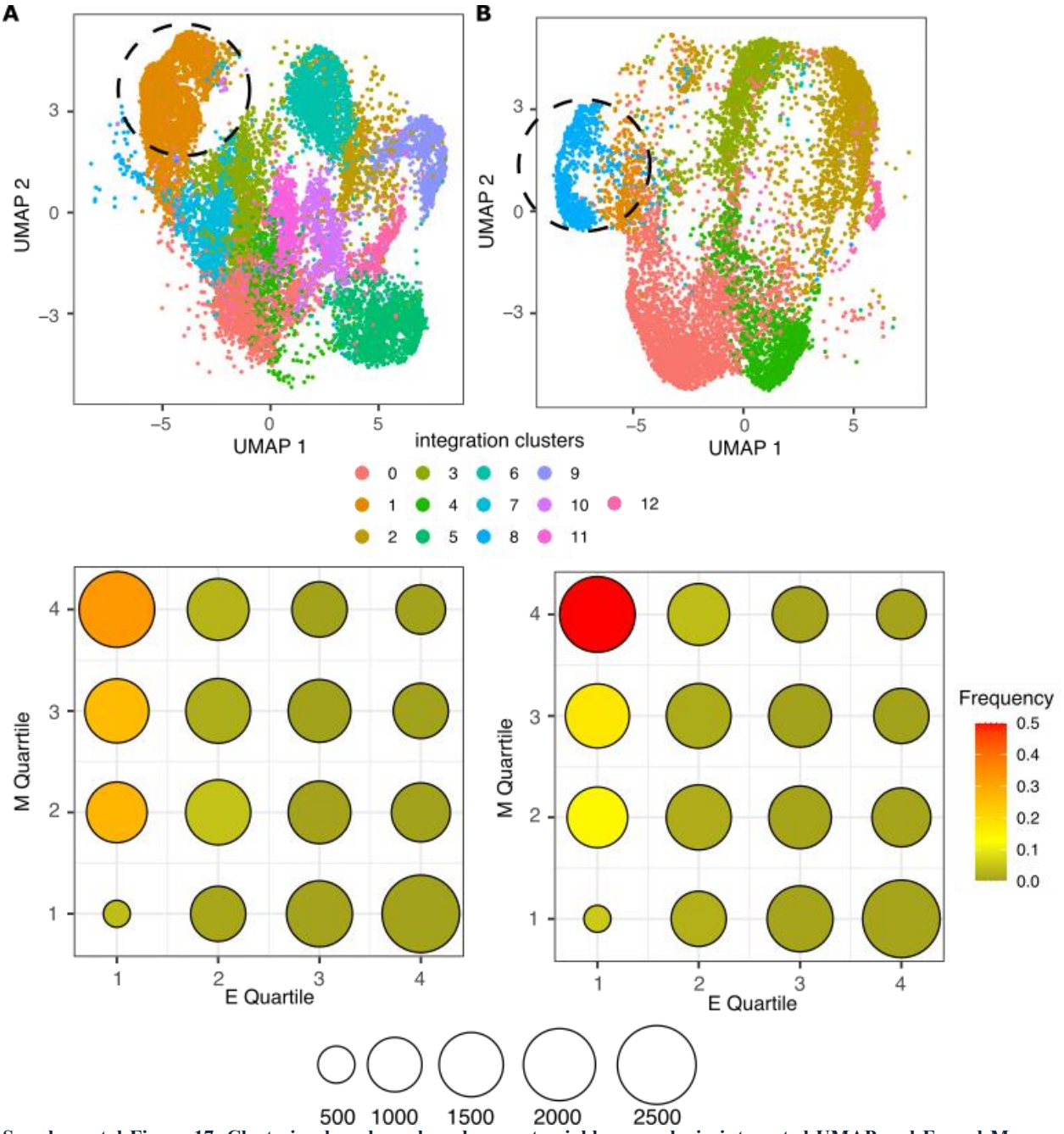
Supplemental Figure 14. Overlays of principal component scores on dose-dependent data.(A) Projection of dose-dependent single-cell samples using UMAP which have been overlaid with scores derived from the principal components (PCs) of dose-dependent data: E PC1 (top), M PC1 (middle), M PC2 (bottom). The color of each point indicates the value of the score from high (red) to low (blue). **(B)** Projection of dose-dependent single-cell samples using UMAP which have been overlaid with scores derived from the components (PCs) of the integrated data: E PC1 (top), M PC2 (middle), M PC1 (bottom). Colors are as in (A).



Supplemental Figure 15. Similarity in nearest neighbor clustering between dose-dependent and integrated data.(A) Projection of dose-dependent single-cell samples using UMAP which have been overlaid with nearest neighbor clusters from dose-dependent (top) and integrated (bottom) data. The color of each point represents the cluster of the sample respective to the legend beneath each figure. **(B)** Projection of dose treated samples from integrated single-cell samples using UMAP which have been overlaid with nearest neighbor clusters from integrated data (top) and dose-dependent data (bottom). The color of each point represents the cluster of the sample respective to the legend beneath each figure.



Supplemental Figure 16. Overlap of unintegrated E- and M-scores across time and dose data. Scatter plot of E-scores (x-axis) and M-scores (y-axis) of samples from the Days 0,4,8 of time experiment (red), Days 0,1,2,3 of time experiment (green) and dose experiment (blue) prior to full integration. Note that time samples have been integrated relative to one another as in Deshmukh et al.[13], but have not been integrated with the samples from the dose experiment.



Supplemental Figure 17. Clustering based on shared-nearest-neighbor graphs in integrated UMAP and E- and M-score space. (A) (above) Projection of time treatment single cell expression data using UMAP on integrated data. The color of individual points indicates the cluster based on shared-nearest-neighbor graphs to which the sample was assigned. The dotted circles highlight Cluster 1. (below) Bubble charts showing the fraction of samples from Cluster 1 in different segments of E- and M-score space. Each point represents a segment of E- and M-score space defined by a particular quartile of E-score (x-axis) and M-score (y-axis). The size of the point corresponds to the total number samples in the segment and the color of the point represents the fraction of samples from the corresponding cluster from 0 (black) to 0.5 (red). **(B)** (above) Projection of dose treatment single cell expression data using UMAP on integrated data. The color of individual points indicates the cluster based on shared-nearest-neighbor graphs to which the sample was assigned. The dotted circle highlights Cluster 8. (below). Bubble charts showing the fraction of samples from Cluster 8 in different segments of E- and M-score space. Point positions, size and color as are defined as in (B).

CHAPTER IV
AN EXPRESSION-BASED PROGNOSTIC MODEL FOR SMALL CELL LUNG CANCER

A version of this chapter will be published by Andrew Willems and Tian Hong.
My contributions to this work include: (1) conceptualization, (2) methodology, (3) software, (4) validation, (5) investigation, and (6) writing and editing the manuscript.

Abstract

Small cell lung cancer (SCLC) is an aggressive and recalcitrate type of lung cancer, posing a significant global health challenge with substantial morbidity. SCLC is characterized by its rapid metastasis, low survival rates, and unique histopathological features. There is therefore a critical need for novel biomarkers that can enhance patient survival prediction. In this work, we introduce an expression-based model that leverages single-cell RNA sequencing data and miRNA targeting information with a regularized Cox model. We show satisfactory accuracy of SCLC patient survival prediction. Because of the model's interpretability and lightweight nature, it is feasible to transfer the trained model across datasets. Additionally, we show that our model outperforms commonly used biomarker discovery methods, underlining its potential to enhance SCLC prognosis while also providing tangible biological insights. Finally, the 4 genes identified by our model properly stratify patient risk and have been associated with other cancers and pathways related to cancer behavior including iron uptake and transport, GTPase cycling, and NOTCH1 signaling [1-5].

Introduction

Small cell lung cancer (SCLC) is one of two major types of lung cancer which has been found to collectively kill 150,000 deaths per year and over 200,000 new cases per year [6]. SCLC comprises about 15% of all new lung cancer cases and has been characterized by a distinct morphology and aggressive metastasis [7]. SCLC has been characterized as responding positively to initial treatment with a high relapse rate characterized by tumors that become immune to standard chemotherapy and become untreatable. Sadly, an estimated 70% of all patients have metastatic

SCLC as the time of diagnosis [8]. This high rate of metastases leads to difficulties with treatment and much more grim outcomes for patients. According to The Surveillance, Epidemiology, and End Results program those who are diagnosed with metastatic SCLC have only a 3% survival rate to 5 years. While immunotherapies have shown some promise in improving these treatment outcomes there is still an urgent need to find new biomarkers that can help predict and improve patient outcomes with SCLC [9].

In the quest to find novel biomarkers that improve survival prediction and treatment outcomes much work has focused on subtyping SCLC. Rudin et al. showed that they could successfully subtype SCLC into multiple types with distinct vulnerabilities based on the expression of a few key transcription factors [10]. ASCL1, NEUROD1, POU2F3, and YAP1 lead to the foundation of their proposed A, N, P, and Y subtyping nomenclature. They found that each of these subtypes had differing levels of key cancer genes and opened the door to potentially personalizing the treatment of SCLC in a more granular way based on the gene expression profile of patient SCLC primary tumors. We used our model which has previously show good predictive performance in colorectal cancer to attempt to identify biomarkers in SCLC and explore how our model's performance might vary across various important biological factors such as tumor stage, age, smoking status, and SCLC subtype and compared it to multiple existing biomarker discovery methods [11]. We show that our model shows good predictive accuracy for overall survival of SCLC patients while outperforming other commonly used differentially expressed gene methods. Our model identifies a small subset of genes that have been implicated in other cancers and relevant cancer pathways (Figure 1).

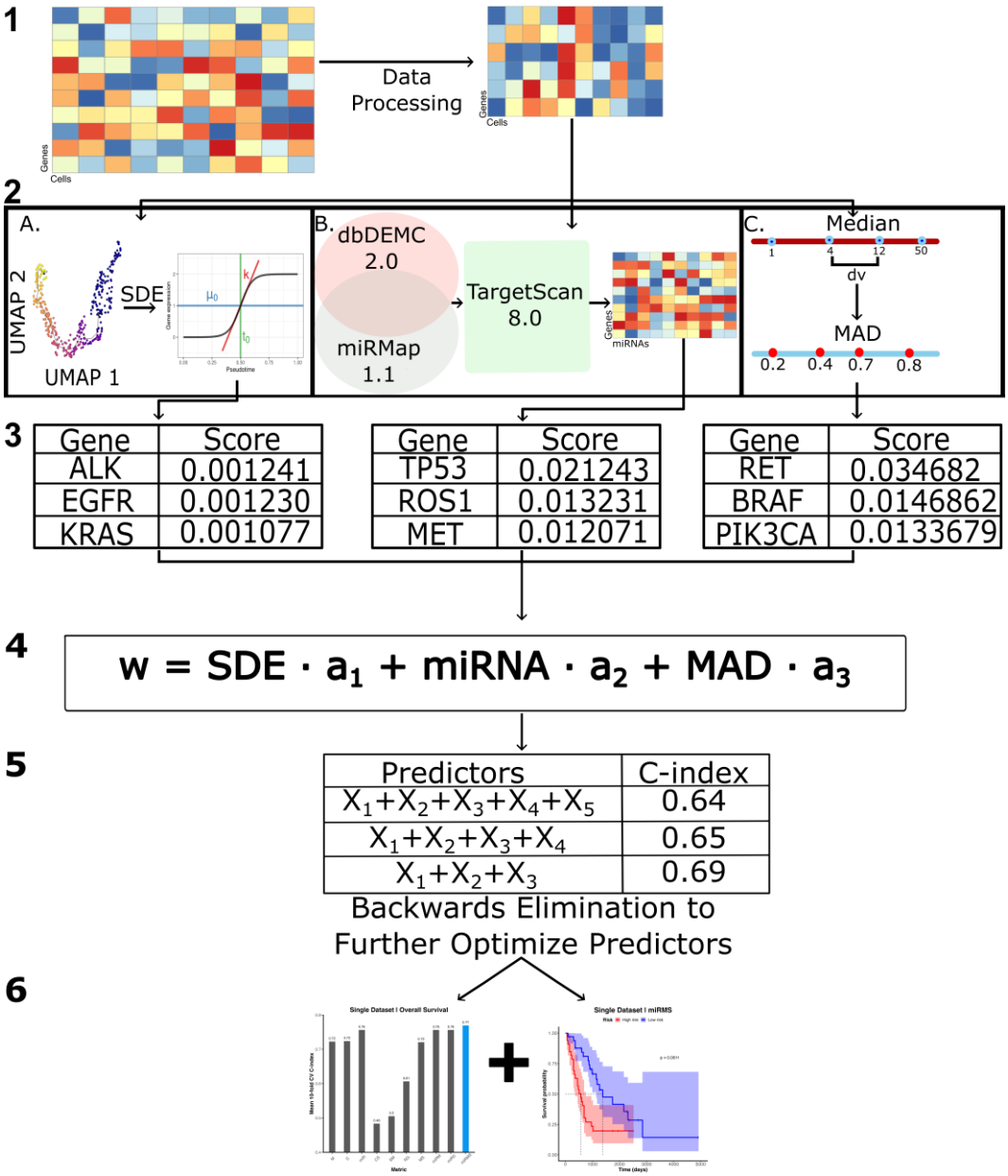

1
Genes
Cells
Data Processing
Genes
Cells
2
A.
UMAP 2
UMAP 1
SDE
Gene expression
Pseudotime
B.
dbDEMC 2.0
miRMap 1.1
TargetScan 8.0
Genes
miRNAs
C.
Median
1
4
12
50
dv
MAD
0.2
0.4
0.7
0.8
3
Gene
Score
ALK
0.001241
EGFR
0.001230
KRAS
0.001077
Gene
Score
TP53
0.021243
ROS1
0.013231
MET
0.012071
Gene
Score
RET
0.034682
BRAF
0.0146862
PIK3CA
0.0133679
4
w = SDE · a1 + miRNA · a2 + MAD · a3
5
Predictors
C-index
X1+X2+X3+X4+X5
0.64
X1+X2+X3+X4
0.65
X1+X2+X3
0.69
Backwards Elimination to Further Optimize Predictors
6
Single Dataset | Overall Survival
Mean 10-fold CV C-index
Metric
Single Dataset | miRMS
Risk
Survival probability
Time (days)

Figure 11: Flowchart of the Prognostic Model Development Process. The six-step methodology for constructing our prognostic model is depicted. Step 1: Data Preparation entails the preprocessing of bulk and single-cell RNA sequencing (scRNA-seq) datasets, consistent with methods outlined in our previous publication. Step 2: Gene Prioritization incorporates a tripartite metric approach— Single-cell Differential Expression (SDE), miRNA-target gene interactions, and Median Absolute Deviation (MAD)—to refine an integrated gene set, which is subsequently processed through a LASSO penalized Cox regression model in steps 3 and 4. Step 5: Predictor Optimization involves a backward elimination process tailored to adhere to the event-per-variable ratio of 1:10, thereby minimizing overfitting. This step iteratively excludes genes to maximize the concordance index (c-index), retaining a gene set with optimal predictive power. Step 6: Validation assesses the performance of the identified gene set using 10-fold cross-validation and the Kaplan-Meier risk estimator, ensuring robustness and reliability of the prognostic model. Note: the SDE plot seen next to the UMAP in step 2A is from Campbell et al.

Materials & Methods

Model

Detailed explanation of all metrics used by our model and our model itself can be found in our previous publication that focused on identifying an interpretable gene signature in colorectal cancer [11].

Backwards elimination procedure

We took the initial 6 gene signature identified by our model and then performed a standard backwards elimination procedure. This procedure involved taking the six identified genes and then subjecting them to an additional round of 10-fold cross validation fitting while dropping one of the 6 genes from the signature. Whichever set of 5 genes gave us the highest c-index were kept for the second round of eliminations. We then repeated this procedure a second time starting with the 5 genes and dropping them out individually and took 4 genes that generated the highest c-index [12].

Synthetic Dataset Bootstrap procedure

We used R code to generate 1,000 synthetic datasets that were based on the distribution of our survival information in our training dataset. Specifically, we generated the synthetic datasets on the distribution of the vital statues of the patients (e.g. alive or dead). We did this by first generating a random seed that was then set before creating the synthetic datasets. This ensured that the datasets

we generated are reproducible. We then calculated the distribution of vital statues in our initial training dataset with the table command and used our custom `bootstrap_func` to generate a single synthetic sample with replacement by randomly sampling from samples that matched our vital status of interest. The `sample_n` function from the dplyr package was used to perform the random sampling with replacement. We then do this for every unique value of vital status up to the number counted by the table function. Once all samples have been generated for this dataset, we then bind all the samples together and return the finished data frame. We repeat this process 1,000 times to get the 1,000 synthetic datasets that we used for our analysis [13, 14]. See our code for more details.

Power calculations

We performed power calculations for all biological variables in Figure 5. These analyses were performed with the `pwr` package in R [15]. Specifically, we used the `pwr.t.test` function for 2 group comparisons and `pwr.anova.test` for 3 group comparisons. The parameters for both functions were as follows: effect size of 0.2, power of 0.8, and alpha threshold of 0.05 [16].

Synthetic Sample Bootstrap procedure

For the generation of synthetic samples for Figure 5 we carried out a similar procedure that of how we generated our synthetic datasets. We subset our data to just the variable of interest (e.g. a particular sex, tumor stage, chemotherapy status, etc.) and then used the `slice_sample` function to randomly sample with replacement up to the number of samples determined by our power calculation for that variable.

Datasets

Our whole-genome sequencing samples came from George et al [17]. The 120 samples contained matched SCLC tumor samples with matched normal samples. We subset the 120 samples to include only samples that contained both genomic expression information and meta data information. This left us with 72 samples (Table 1). We accessed the whole-genome sequencing samples through cBioPortal web interface [18-20]. The scRNA-seq dataset used is from Tian et al [21].

Implementation

We implemented our metric and model in R. All code and data for this manuscript are available on Github (https://github.com/compbiolover/sclc). Detailed code on the implementation of our model and metrics are also available on Github (https://github.com/compbiolover/CC-Singlecell).

Results

Table 9. Whole Genome Sequencing Sample Characteristics. This table presents a modified version of the whole genome sequencing data from George et al., summarizing key clinical and demographic variables for 72 cancer patients. The variables include vital status, treatment modalities (radiation and chemotherapy), tumor staging (TNM), smoking status, sex, and age distribution. For N and M staging, 'X' indicates that the staging could not be determined. The age classification of 'Old' or 'Young' is based on the median age of patients in our dataset; patients above the median age are classified as 'Old', while those below are classified as 'Young'.

Vital Status		Treatment	
Alive	26	Radiation (Yes/No)	37/25
Deceased	46	Chemotherapy (Yes/No)	48/17
Tumor Stage		**N Staging**	
I/II/III/IV	32/12/20/8	0/1/2/3/X	32/12/20/2/1
Smoker		**M Staging**	
Current/Former	40/32	0/1/X	55/8/4
Sex		**Age**	
Female/Male	21/51	Old/Young	34/38

Our model heavily focuses on gene prioritization to help reduce the complex, high-dimensional space of large sequencing datasets down to just a handful of easy to interpret genes. It does this through three distinct metrics. One, the median absolute deviation (MAD) metric. This metric evaluates the variability of a particular gene's expression profile. Two, the Switch-like differential expression across single-cell trajectories (SDE), and three the association between a disease-causing miRNA and its gene target (miR). We first sought to see how well each of the individual metrics performed and how well they performed in combination. We found that the metrics had variables levels of performance with it consistently performed in the low to mid-0.7s of the concordance index. We found that taking all three metrics together (miRMS) gave us the best overall performance slightly edging out the miR metric alone and several two metric combinations. (Figure 2A). Given that there is a paucity of SCLC genomic data that also includes survival information we sought to get a robust sense of how well our combined signature performed. We did this by generating 1,000 synthetic bootstrapped datasets that mirrored the distribution of our existing dataset's survival times. (see Materials & Methods for details) We saw that once again the miRMS metric was our best performer with a mean c-index of 0.74 across the 1,000 datasets (Figure 2B). Here, miRMS had more separation in its performance compared to individual metrics or two metric combinations. Next, we sought to see if our miRMS performed when predicting progress-free survival as this has been found to be a better survival metric for some cancers. We took the same 1,000 bootstrapped datasets and predicted progress free instead of overall survival. We observed that once again the miRMS metric had the best performance when assessing progress-free survival. Interestingly this time it tied with two, two metric combinations and the miR and SDE metrics alone which all performed at the same level (Figure 2C). This shows the complexity in predicting cancer survival. Despite this nuance give that it tied for top performer we continued

to investigate the performance of miRMS. We found that it was able to stratify patients into high and low risk groups when using the Kaplan-Meier risk estimator. In particular we found that the median survival of a patient in the low-risk group lived an estimated ~1,380 days vs. ~510 days for a patient in the high risk group. This large difference in survival times implies our identified signature has meaningful potential patient benefit as the difference in median survival is 870 days (Figure 2D).

Next, we sought to identify and better validate the genes associated with miRMS. We found that it was a six-gene signature that had a very large hazard ratio associated with a single gene (GRIN2B). We used backwards elimination to determine the 2 genes that should be removed to ensure that our number of predictors followed the 10:1 rule commonly used in the field of statistics to ensure that our model's number of predictors is appropriate for the number of events in our dataset. After backwards elimination we were left with 4 genes[12]. These four genes were TFRC, FAM83f, DLK1, and GNG13 (Figure 3A). These 4 genes have been found to be associated with various cancers including lung cancer [1, 22-25]. Next, we were interested in what pathways these genes might play roles in. We used Enrichr to perform this gene enrichment analysis [26-28]. We found these genes were associated with a variety of cancer associated pathways including iron metabolism, beta cell proliferation, notch signaling (Figure 3B) [3, 29-34]. They were also found to be associated with various aspects of Gamma signaling, glucagon-like peptide-1, and G-protein activation (Figure 3C)[35-38]. We then sought to compare our model's performance to that of other commonly used methods for determining differentially expressed genes. We observed that miRMS outperformed all other methods with a mean 10-fold cross-validated c-index of 0.724 compared to 0.652 for DEsingle, 0.646 for scDD, 0.683 for DESeq2, and 0.646 for edgeR (Figure 3D). We then followed this with a robust assessment of the other methods c-index performance by performing

the bootstrapping procedure described earlier with miRMS once again proving to be the best performing method (Figure 3E). Finally, we wanted to see if our 4 gene signature varied across sex as much literature has found women, on average, have better survival rates compared to me for SCLC. We subset our training data to just patients annotated as female and found that our signature did indeed do a better job of predicting the survival compared to male patients. We still had good performance for both groups with a mean of 0.77 for the males and 0.96 for the females (Figure 3F).

Next, we sought to see how miRMS performed across a variety of important biological and sociological variables including the TNM classification of malignant tumors (TNM) and other important variables associated with cancer severity such as age and smoking status. We performed power analyses to ensure we could adequately detect actual differences. Based on these calculations we generated synthetic samples from the relevant subgroups with a random sampling procedure up to the number determined by the power calculation (see Materials & Methods for details). We found that in all cases our signature had differential performance across primary tumor size (T), degree of spread to regional lymph nodes (N), and presence of distance metastasis (M) (Figures 4A-4C). This was also true when examining the age of patients (Figure 4D), the smoking status of patients (Figure 4E), and if they had been given chemotherapy previously (Figure 4F).

Finally, we compared the performance of our method to two deep-learning methods. Due to some debate about how deep-learning and classical statistical models compare in performance we sought to see if using either a Random Forest survival model (randomForestSRC) or a Cox Proportional Hazards deep neural network (deepSurv) lead to better c-index performance than our model[39, 40]. We tested this by taking the top 20 genes identified by the gene prioritization step of our

pipeline and then giving them to each of the methods. For each method we performed either 10-fold cross validation for the c-index (miRMS, deepSurv, random genes) or we calculated the c-index directly (random forest). We found that even when these deep learning methods select their optimal set of predictors miRMs still performs better speaking to the power of our approach. In addition, we show that gene prioritization makes a difference to overall performance as all methods have better performance than just selecting random genes (Figure 5).

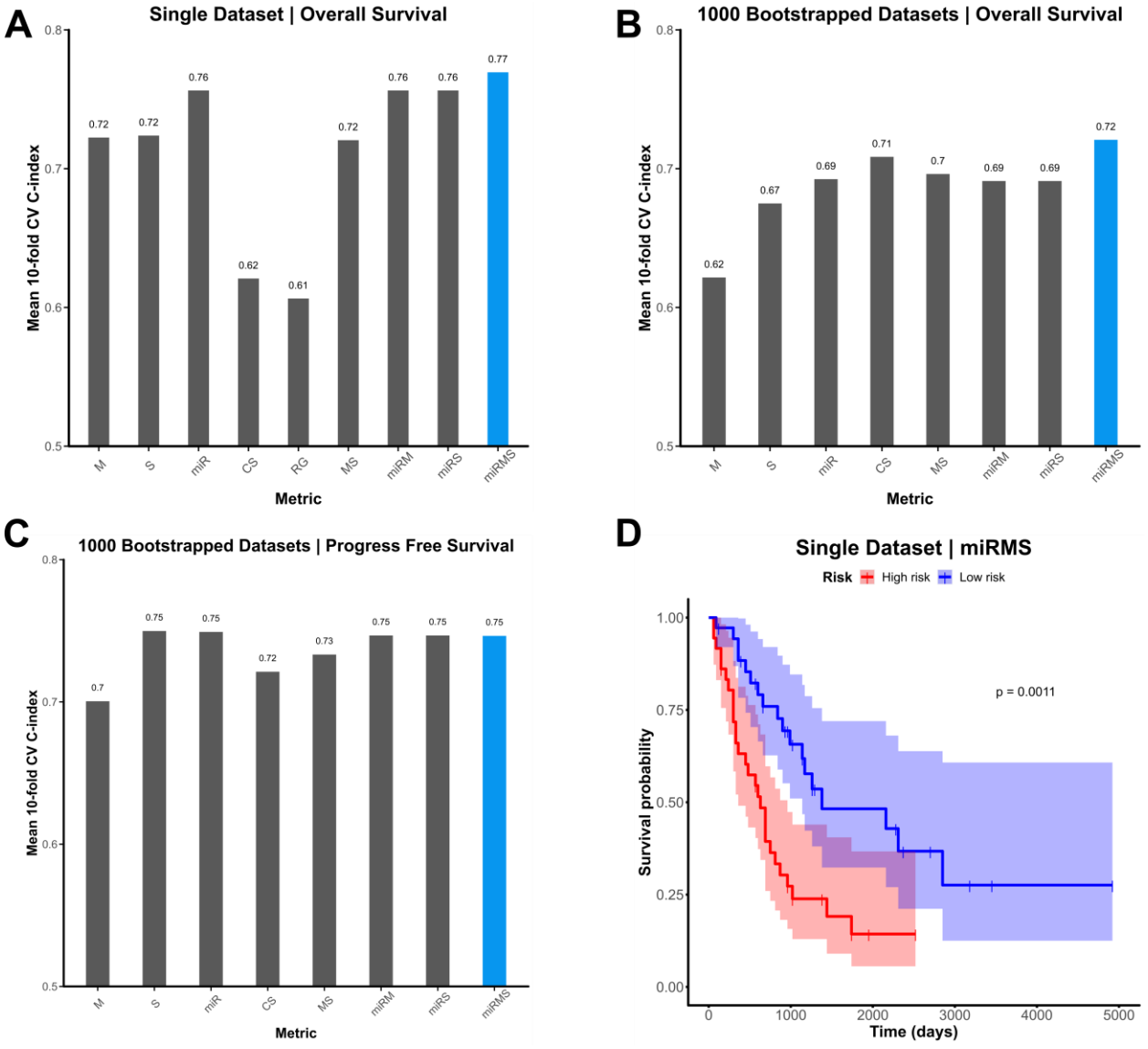


Figure 12: Composite Metrics Outperform Individual Measures. We evaluated the performance of individual metrics (M = MAD, S = SDE, miR = miRNA) and compared them to linear combinations of these metrics in (A) a single dataset and (B) 1,000

simulated datasets based on the survival distribution of the dataset used in (A). Combining all three metrics yielded the highest overall concordance index (c-index) in 10-fold cross-validation. To validate the signature identified in (A), we applied it to 1,000 datasets and calculated the mean of the mean 10-fold cross-validation across all 1,000 datasets in (B). We also assessed the performance of the signature in progression-free survival. In this analysis, both SDE and miRNA alone performed as well as the three-metric model. In (D), we demonstrated that the six-gene signature identified by the three-metric model effectively stratifies patients into high-risk and low-risk groups. Significance in (D) was determined using a p-value threshold of < 0.05 using the log-rank test.

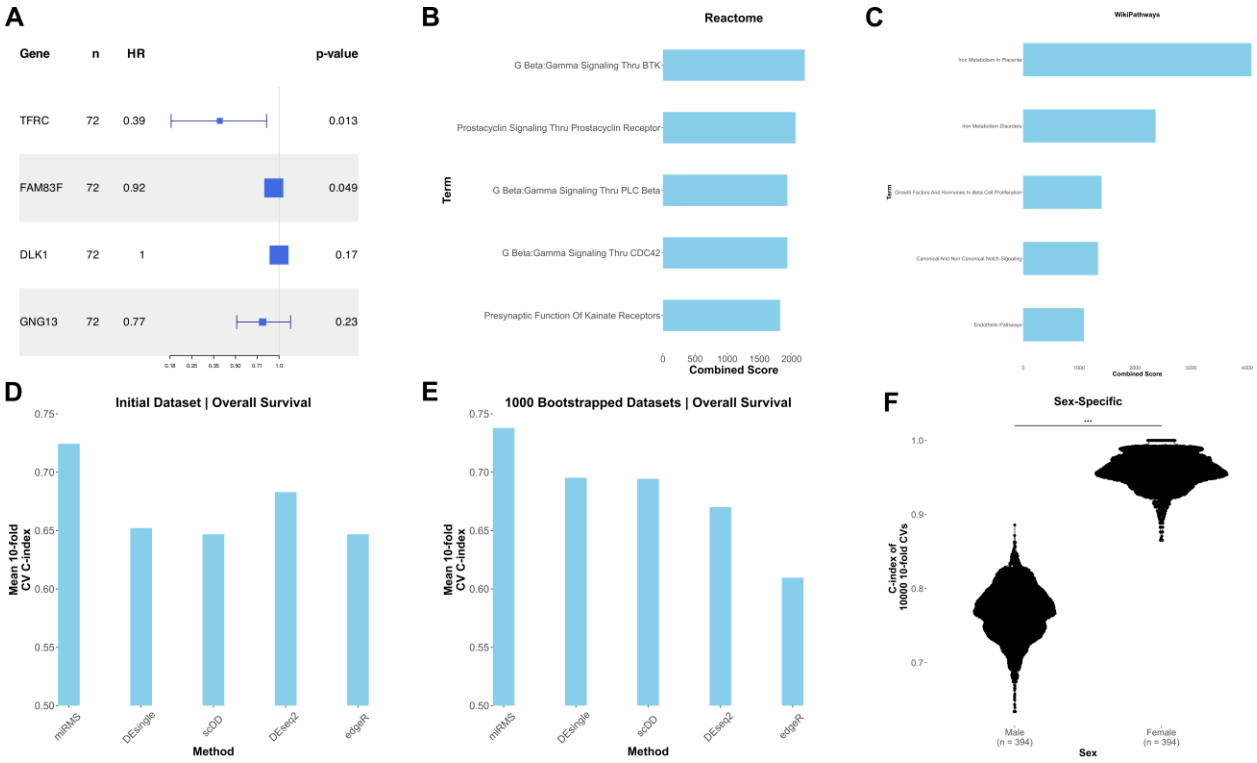


Figure 13: Validation of miRMS model. In (A) we show the the performance of our signature on the training dataset with the hazard ratios of each of the genes in the model. In (B), we have the top 5 enriched pathways from the Reactome database from the genes in our model. In (C) we have same but for the enriched pathways from the Wikipathways database. Both pathway enrichment tools reveal that our genes are associated with pathways related to iron metabolism, NOTCH signaling, and various aspects of Gamma signaling. Thes pathways have been shown to play important roles in various cancers. In (D) we performed 10-fold cross validation on the initial dataset and took the mean of this performance for miRMS along with several commenly used diffeential-expression methods. We see that miRMS outperforms all methods. In (E) we perform the same analysis as in (D) but on our 1,000 synthetic datasets to compare the methods to further validate miRMS's performance benefits compared to other methods. Finally, in panel (F) we investigated whether miRMS performance was influenced by sex and saw that it performed well for both sexes but that it has better mean performance for females. This is inline with literature that finds women have better outcomes from SCLC. *** p-value < 0.001. Combined Score is defined as the -log(p-value) * odds ratio. log is the natural logorithm. p-value threshold is < 0.05. n in (F) based on 0.2 effect size at 80% power.

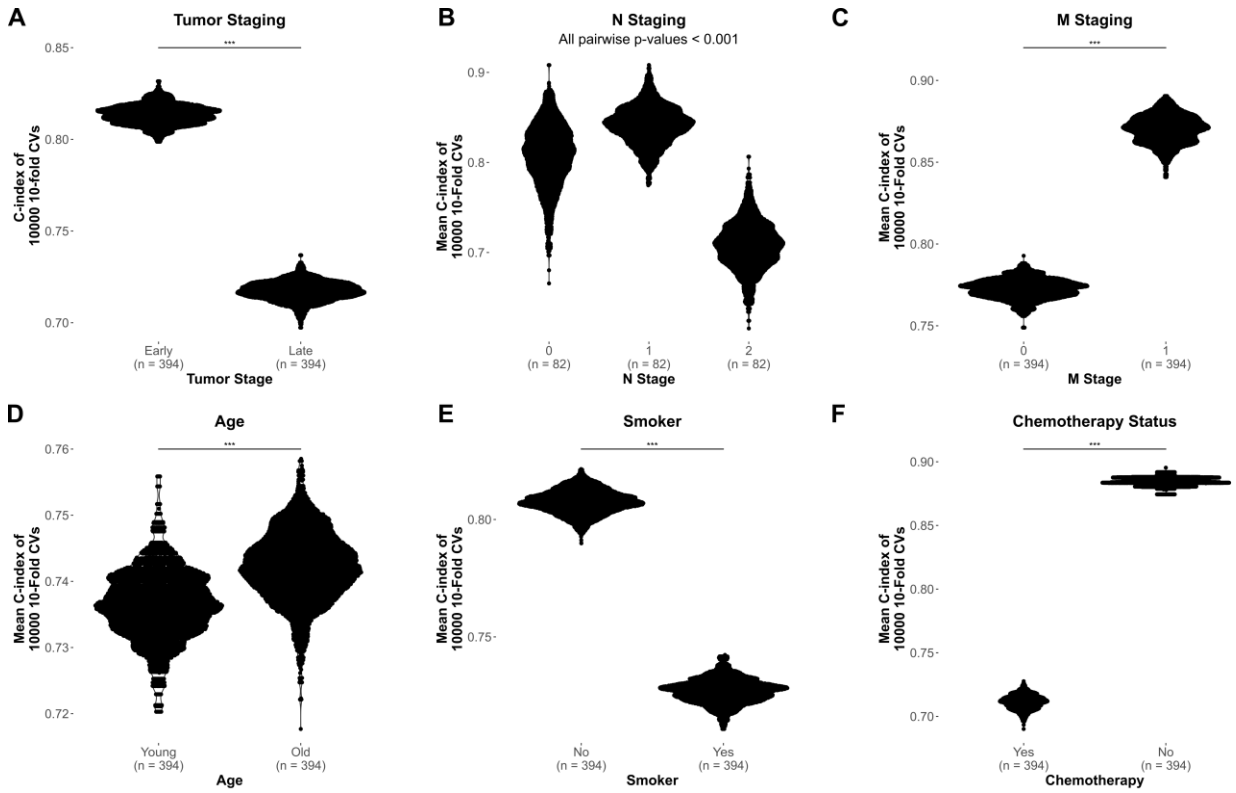


Figure 14: Differential Predictive Ability of the miRMS Signature Across Biological & Sociological Variables. Here we employed a four-gene miRMS signature and subjected it to rigorous evaluation using distinct subsets of the initial training dataset. These subsets were generated via simulated sampling based on power calculations. Our primary objective was to scrutinize the signature's predictive performance across the spectrum of the TNM Classification of Malignant Tumors (TNM) standard and other importnat biological or sociological variables. Specifically, the training dataset was meticulously stratified to encompass only those samples that conformed to specific TNM staging criteria. Subsequently, simulated samples were meticulously generated in accordance with the predetermined power calculations. Following this, we conducted a comprehensive assessment of our four-gene signature, leveraging the power of 10,000 iterations of the 10-fold cross-validation technique. (A) When categorizing all patients with stage I or II cancer as "early" and those annotated as stage III or IV as "late," our findings align with expectations, revealing a higher model accuracy for patients with less advanced tumor staging. (B) An exploration of N stages 0, 1, and 2 demonstrated comparable model performance for lymph node stages 0 and 1. However, a noticeable decline in performance was observed when predicting samples characterized by the most advanced lymph node staging. (C) Interestingly, our results exhibit superior predictive performance for samples classified as "metastatic" (1) compared to those categorized as "non-metastatic" (0). This observation may be attributed to the limitations inherent in a binary classification of samples into metastatic or non-metastatic categories. (D) Upon categorizing patients as "old" or "young" based on the median age of the entire cohort, our model displayed marginally enhanced performance for older patients. This phenomenon could be attributed to the typicality of older patients, while younger SCLC patients may present with additional complexities, potentially influencing earlier cancer diagnoses and, consequently, predictive accuracy. (E) Intriguingly, our model exhibited heightened accuracy for patients with no history of smoking. (F) Finally, patients who had not previously undergone chemotherapy were found to be better predicted by our signature. All p-values reported as < 0.05. P-values were calculated employing the Wilcoxon Ranked Sum test (A, C-F) or ANOVA (B). Pairwise p-values were further assessed using Dunn's test (B). The "n" represents the number of samples within each group, as determined by power calculations based on an effect size of 0.2 at 80% power.

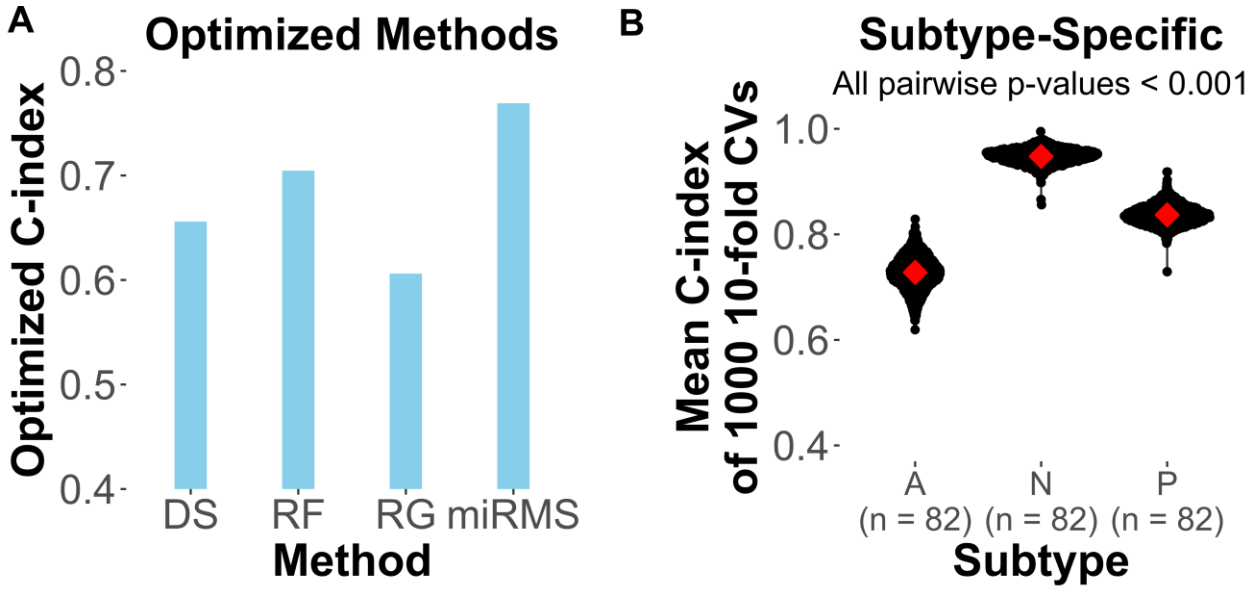


Figure 15: Comparative Performance of miRMS Against Deep Learning Methods and Across SCLC Subtypes (A) Performance comparison of miRMS with deep learning methods using the top 20 genes identified by our gene prioritization pipeline. miRMS outperforms both deepSurv (DS) and Random Forest (RF), as well as a set of randomly selected genes (RG). This demonstrates the effectiveness of our gene prioritization approach and the superiority of miRMS even when all methods are optimized. (B) Differential performance of our identified signature across SCLC subtypes, based on the ANPY subtyping system proposed by Rudin et al. The most common subtype (A) shows a predictive accuracy of approximately 0.72. The Y subtype was excluded from analysis due to insufficient sample size (n=2) in our dataset. Statistical significance was assessed using ANOVA followed by Dunn's test for multiple comparisons, with p-values adjusted using the Benjamini-Hochberg method ($\alpha = 0.05$).

Discussion

SCLC has proven to be a difficult to treat cancer with high mortality. Our prognostic model shows an efficient and light-weight way to prioritize and determine a small subset of genes that are important for predicting patient survival. We show that the signature is robust in its performance through many synthetically generated datasets and that our method identifies genes that appear to play roles in biologically important pathways that are relevant to cancer. These include pathways such as G:beta gamma signaling and iron metabolism which have both been found to play important roles in the proliferation and miniatous of various cancers.

On top of identifying a biologically easy to interpret set of predictors that can be further tested at the bench, we show that our method outperforms several popular differentially expressed genes methods such as DESeq2, edgeR, and DESingle which are popular in the literature for finding novel cancer biomarkers. We show that the genes identified by our model separate patients into

high and low risk groups and that those groups have meaningful differences in median survival time.

Our method is computationally lightweight and can be run on a standard desktop or laptop computer, which stands in stark contrast to some methods that require substantially more computational power to fit. Additionally, our prognostic model allows for a straightforward swap between classical statistical approaches like the Cox model and deep learning-based approaches like deepSurv and randomForestSRC. This "plug and play" type of interface in our model allows for multiple model types to be used in our pipeline and could in theory lead to more comprehensive benchmarking for certain types of tasks, as one could imagine swapping in many different models while holding the inputs (same prioritized gene list) and output types (concordance index and Kaplan-Meier risk estimator) constant.

However, it's important to address the limitations of our study. Given that our model is both trained and validated on a single small SCLC dataset of 72 samples, its generalizability is limited. We acknowledge the risk of overfitting to our dataset. To alleviate this concern, we performed both 10-fold cross validation and synthetic data generation. While these approaches strengthen the robustness of our model's survival accuracy and confidence in the HRs identified, it cannot be guaranteed that some overfitting did not occur in our model fitting process.

The true generalizability of our model could only be sufficiently assessed with more real-world datasets being collected and analyzed with our identified signature. However, we do know from our previous publication that our pipeline is well suited to work across at least a few other types of cancer, and that many of the conclusions we see here in SCLC are also seen in colon and rectal cancer. Panchy et al. used our approach to show that our method found genes within their E-M space work that had varying levels of survival performance across 27 different kinds of cancer

from TCGA. They found that many of their highest performing gene signatures were in their high M space, which helps strengthen the notion that our method, while only being validated on a single SCLC dataset, would offer a consistent and valid approach for future SCLC datasets [41].

We note that we did not exclude patient samples with "x" tumor staging given that there was only 1 N stage and 4 M stage samples whose staging information could not be determined. Given our already small sample size, we opted to include them as the concerns about statistical power seemed more pressing.

Our model used only genomic expression information to make patient survival predictions. In theory, incorporating other important clinical variables (such as patient treatment regime, comorbidities of patients, etc.) would lead to even better overall model accuracy. We hope to take this work in that direction in the future. It would also be exciting to take this pipeline in a direction focused on trying to optimize treatment outcomes or recommendations. One could image that instead of using Cox models to predict survival time we could swap it out for another model and predict treatment effectiveness for example.

Conclusion

In conclusion, we have shown that our previously published model is adept at predicting SCLC patient survival while identifying an interpretable and robust set of genes. The genes identified by our model appear to be biologically relevant and have a tangible impact on patient survival. Finally, our identified signature outperforms several commonly used methods for biomarker discovery while using similar numbers of genes.

Data Availability

All code generated and used in this manuscript can be found at two GitHub repositories (https://github.com/compbiolover/sclc) and (https://github.com/compbiolover/CC-Singlecell). The bulk RNA-seq dataset can be downloaded from the cBioPortal for Cancer Genomics (https://www.cbioportal.org/study/summary?id=sclc_ucologne_2015). The single-cell RNA-seq dataset can be downloaded from the OMIX database (https://ngdc.cncb.ac.cn/omix/release/OMIX002441).

References

1. Pittaway, J.F., et al., *The role of delta-like non-canonical Notch ligand 1 (DLK1) in cancer.* Endocrine-Related Cancer, 2021. **28**(12): p. R271-R287.
2. Haga, R.B. and A.J. Ridley, *Rho GTPases: Regulation and roles in cancer cell biology.* Small GTPases, 2016. **7**(4): p. 207-221.
3. Gharaibeh, L., et al., *Notch1 in cancer therapy: possible clinical implications and challenges.* Molecular pharmacology, 2020. **98**(5): p. 559-576.
4. Sun, J., et al., *Identification of key pathways and genes in PTEN mutation prostate cancer by bioinformatics analysis.* BMC medical genetics, 2019. **20**(1): p. 1-9.
5. Zhang, W., et al., *Combined treatment of polygonatum and Scutellaria baicalensis suppresses lung cancer cell proliferation through inducing ferroptosis.* Phytomedicine Plus, 2023. **3**(4): p. 100482.
6. Basumallik, N. and M. Agarwal, *Small cell lung cancer*, in *StatPearls [internet]*. 2022, StatPearls Publishing.
7. Jain, A.S., et al., *Everything old is new again: Drug repurposing approach for non-small cell lung cancer targeting MAPK signaling pathway.* Frontiers in Oncology, 2021. **11**: p. 741326.
8. Evans, W., et al., *VP-16 and cisplatin as first-line therapy for small-cell lung cancer.* Journal of Clinical Oncology, 1985. **3**(11): p. 1471-1477.
9. Esposito, G., et al., *Immunotherapy in small cell lung cancer.* Cancers, 2020. **12**(9): p. 2522.
10. Rudin, C.M., et al., *Molecular subtypes of small cell lung cancer: a synthesis of human and mouse model data.* Nature Reviews Cancer, 2019. **19**(5): p. 289-297.
11. Willems, A., N. Panchy, and T. Hong, *Using Single-Cell RNA Sequencing and MicroRNA Targeting Data to Improve Colorectal Cancer Survival Prediction.* Cells, 2023. **12**(2): p. 228.
12. Derksen, S. and H.J. Keselman, *Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables.* British Journal of Mathematical and Statistical Psychology, 1992. **45**(2): p. 265-282.
13. Wickham, H., et al., *Welcome to the tidyverse: Journal of Open Source Software, v. 4.* doi, 2019. **10**: p. 1686.

14. R Core Team, R., *R: A language and environment for statistical computing.* 2013.
15. Champely, S., C. Ekstrom, and P. Dalgaard, *pwr: Basic Functions for Power Analysis. R package version 1.3-0*. 2020.
16. Schmidt, S.A., S. Lo, and L.M. Hollestein, *Research techniques made simple: sample size estimation and power calculation.* Journal of Investigative Dermatology, 2018. **138**(8): p. 1678-1682.
17. George, J., et al., *Comprehensive genomic profiles of small cell lung cancer.* Nature, 2015. **524**(7563): p. 47-53.
18. Cerami, E., et al., *The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data.* Cancer discovery, 2012. **2**(5): p. 401-404.
19. Gao, J., et al., *Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal.* Science signaling, 2013. **6**(269): p. pl1-pl1.
20. de Bruijn, I., et al., *Analysis and visualization of longitudinal genomic and clinical data from the AACR project GENIE biopharma collaborative in cBioPortal.* Cancer research, 2023. **83**(23): p. 3861-3867.
21. Tian, Y., et al., *Single-cell transcriptomic profiling reveals the tumor heterogeneity of small-cell lung cancer.* Signal Transduction and Targeted Therapy, 2022. **7**(1): p. 346.
22. Zhu, X., et al., *Mechanisms of gambogic acid-induced apoptosis in non-small cell lung cancer cells in relation to transferrin receptors.* Journal of Chemotherapy, 2009. **21**(6): p. 666-672.
23. Yuanlin, L., et al., *High GNG13 expression is associated with poor survival in epithelial ovarian cancer and breast cancer.* Neoplasma, 2022. **69**(1).
24. Salama, M., et al., *Fam83F induces p53 stabilisation and promotes its activity.* Cell Death & Differentiation, 2019. **26**(10): p. 2125-2138.
25. Li, L., et al., *DLK1 promotes lung cancer cell invasion through upregulation of MMP9 expression depending on Notch signaling.* PloS one, 2014. **9**(3): p. e91509.
26. Chen, E.Y., et al., *Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool.* BMC bioinformatics, 2013. **14**(1): p. 1-14.
27. Kuleshov, M.V., et al., *Enrichr: a comprehensive gene set enrichment analysis web server 2016 update.* Nucleic acids research, 2016. **44**(W1): p. W90-W97.
28. Xie, Z., et al., *Gene set knowledge discovery with Enrichr.* Current protocols, 2021. **1**(3): p. e90.
29. Wang, Y., et al., *Iron metabolism in cancer.* International journal of molecular sciences, 2018. **20**(1): p. 95.
30. Agiannitopoulos, K., et al., *Copy Number Variations (CNVs) account for 10.8% of pathogenic variants in patients referred for hereditary cancer testing.* Cancer Genomics & Proteomics, 2023. **20**(5): p. 448-455.
31. Lucito, R., et al., *Copy-number variants in patients with a strong family history of pancreatic cancer.* Cancer biology & therapy, 2007. **6**(10): p. 1592-1599.
32. Yuan, X., et al., *Notch signaling: an emerging therapeutic target for cancer treatment.* Cancer letters, 2015. **369**(1): p. 20-27.
33. Irani, S., et al., *A review of the profile of endothelin axis in cancer and its management.* Critical reviews in oncology/hematology, 2014. **89**(2): p. 314-321.
34. Heit, J.J., S.K. Karnik, and S.K. Kim, *Intrinsic regulators of pancreatic β-cell proliferation.* Annu. Rev. Cell Dev. Biol., 2006. **22**: p. 311-338.

35. Tennis, M.A., et al., *The role of prostacyclin in lung cancer.* Translational research, 2010. **155**(2): p. 57-61.
36. Xiao, X.-H., et al., *Regulating Cdc42 and its signaling pathways in cancer: small molecules and MicroRNA as new treatment candidates.* Molecules, 2018. **23**(4): p. 787.
37. Wang, J. and C.H. Kim, *Differential risk of cancer associated with glucagon-like peptide-1 receptor agonists: analysis of real-world databases.* Endocrine Research, 2022. **47**(1): p. 18-25.
38. Messex, J.K. and G.-Y. Liou, *Targeting BTK signaling in the microenvironment of solid tumors as a feasible cancer therapy option.* Cancers, 2021. **13**(9): p. 2198.
39. Katzman, J.L., et al., *DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network.* BMC medical research methodology, 2018. **18**(1): p. 1-12.
40. Ishwaran, H., et al., *Random survival forests.* 2008.
41. Panchy, N., et al., *Comparative single-cell transcriptomes of dose and time dependent epithelial–mesenchymal spectrums.* NAR Genomics and Bioinformatics, 2022. **4**(3): p. lqac072.

CONCLUSION

The overarching objectives of this thesis are twofold: firstly, to advance biomarker discovery in high-dimensional genomic spaces by identifying interpretable and readily testable gene sets; and secondly, to enhance patient survival prediction in various cancer contexts, specifically colorectal and small cell lung cancer.

To pursue these objectives, a novel miRNA gene targeting metric was developed within a comprehensive gene prioritization and integration pipeline. This pipeline integrates single-cell information, miRNA-targeting data, and bulk RNA-sequencing data in a cohesive manner. The resulting framework identified novel biomarkers that demonstrated superior patient survival prediction accuracy compared to existing methods for biomarker discovery, such as DESeq2 and edgeR, in the contexts of colorectal and small cell lung cancer (Chapters 2 and 4).

While this work does not definitively resolve the challenges associated with high-dimensional biological data, it establishes a robust framework for integrating multiple types of omic information while maintaining interpretability of results. This approach is particularly crucial in the era of big data, where many advanced models, despite their excellent predictive accuracy, often lack interpretability and intuitive connections between identified features and outcomes of interest. Furthermore, the predictors identified by such models may prove challenging to validate experimentally.

In contrast, the approach presented in this thesis, while offering good (if not superior) predictive accuracy, facilitates straightforward interpretation of each predictor's impact on patient outcomes. For instance, a predictor with a hazard ratio less than 1 is interpreted as beneficial to the patient, while a predictor with a hazard ratio greater than 1 is considered detrimental. Notably, the

predictors identified by this model are consistently enriched in pathways crucial to cancer development, progression, and immune evasion.

The flexible and modular nature of this approach allows for the incorporation of additional prioritization metrics beyond the three utilized in this thesis. For example, alternative methods such as standard deviation could be employed to rank genes by variance by simply replacing the Median Absolute Deviation (MAD) metric. Moreover, the linear model can be extended to accommodate any number of metrics through a straightforward extension of the grid search code.

This work also demonstrates potential applications beyond its primary focus. For instance, bioinformaticians interested in predicting tumor severity could replace the Cox model with alternative models such as Random Forest. Similarly, researchers aiming to assess how potential biomarkers predict treatment outcomes could substitute the Cox model with methods like k-Nearest Neighbors (Chapter 4).

Furthermore, the utility of this method extends to predicting outcomes from other types of analyses. Chapter 3 illustrates how this approach can be used to evaluate genes in various segments of the Epithelial-Mesenchymal (E-M) scoring space and determine their efficacy in predicting survival outcomes across a wide range of cancers.

It is important to acknowledge the limitations of this method. Primarily, it may not achieve the same level of predictive accuracy as some deep learning approaches, which often prioritize performance at the expense of interpretability. Additionally, while this work has shown promising results in colorectal and small cell lung cancer, its performance has not been validated across all cancer types, and its efficacy may vary in different biological contexts. While our multi-step prioritization and penalization model provides an interpretable and high-performance signature,

we do not claim it to be universally superior for identifying interpretable gene signatures in high-dimensional omic data spaces.

Despite these limitations, this work has the potential for significant immediate impact in survival analysis across various diseases with substantial available omics data. Its flexible nature and lightweight computational framework make it readily applicable to other contexts where predicting survival is of interest. Interestingly, its utility may extend beyond human studies to any genomic data with a time-to-event framework, such as time to flowering in plants or formation of microbial communities.

To build upon these findings, several immediate next steps are proposed. First, validating the method across a broader range of cancer types would further establish its generalizability. Second, expanding the model's outputs could enhance its utility. Developing a more comprehensive tool that allows researchers to input scRNA-seq matrices, bulk genome sets, and cancer-specific miRNA data to generate pathway enrichment analyses and publication-quality figures would significantly increase the method's accessibility and impact.

In conclusion, this work demonstrates that integrating and prioritizing multiple types of omics data significantly improves patient survival prediction and outperforms existing biomarker discovery tools. The flexibility of the developed framework allows for its application across multiple model types and diverse biological contexts, providing actionable and interpretable insights that are amenable to further validation. While acknowledging its limitations, this approach offers a promising direction for future research in biomarker discovery and survival prediction, with potential applications extending beyond cancer research to various fields of biological inquiry.