cells

MDPI

Article

Using Single-Cell RNA Sequencing and MicroRNA Targeting Data to Improve Colorectal Cancer Survival Prediction

Andrew Willems [1], Nicholas Panchy [2] and Tian Hong [3,4,*]

[1] School of Genome Science and Technology, The University of Tennessee, Knoxville, TN 37916, USA
[2] Institute for Cyber-Enabled Research, Michigan State University, East Lansing, MI 48824, USA
[3] Department of Biochemistry & Cellular and Molecular Biology, The University of Tennessee, Knoxville, TN 37996, USA
[4] National Institute for Mathematical and Biological Synthesis, Knoxville, TN 37996, USA
* Correspondence: hongtian@utk.edu

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.

Keywords: colon cancer; microRNA; single-cell RNA-sequencing

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

Academic Editors: César López-Camarillo, Macrina B. Silva-Cázares and Carlos Pérez Plasencia

Received: 1 December 2022
Revised: 21 December 2022
Accepted: 22 December 2022
Published: 5 January 2023

1. 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 (22 nts), 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.

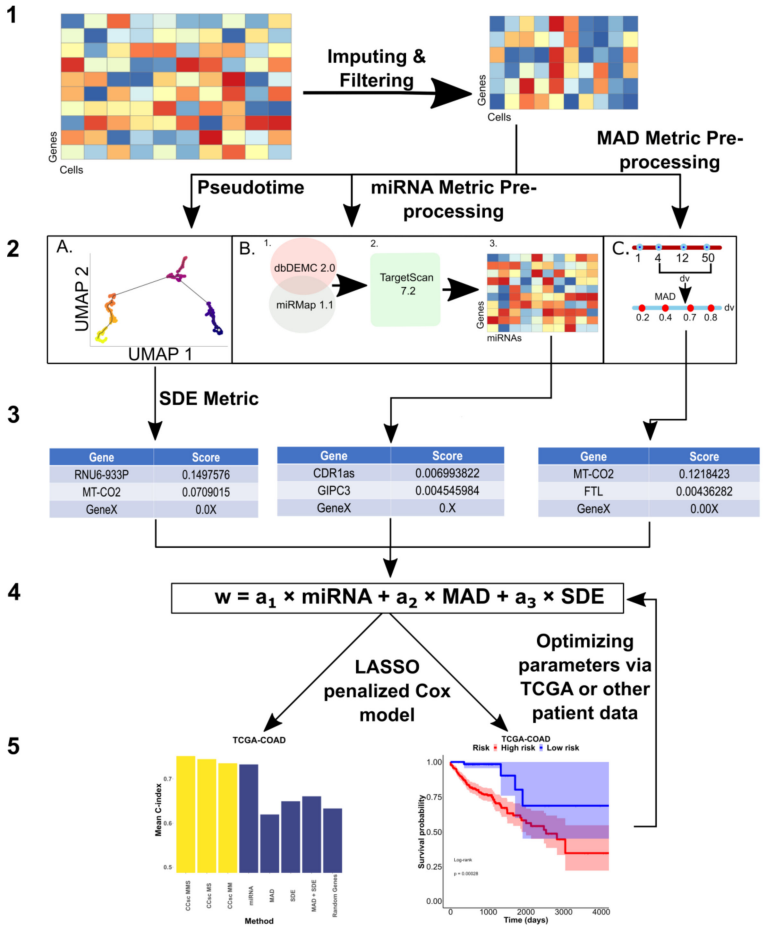

Figure 1. Overall Design of CCsc. CCsc is composed of 5 steps. (**1**) Obtaining pre-processed scRNA-seq matrix (genes by cells) and performing imputation and denoising via the MAGIC package and then filtering for low abundance samples. (**2**) Using filtered, denoised, and imputed scRNA-seq matrix to 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.

2. Materials and 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) [13,14].

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) [15,16]. 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 S1), 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 [17]. 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 [17]. 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 [18]. 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) [19]. 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 [20]. We used a scRNA-seq dataset for colorectal tumors that primarily contain epithelial cells [17]. Epithelial-mesenchymal transition is a known driver for epithelial plasticity and tumor progression for colorectal cancer and several other cancer types [17,21–24]. There, we used the expression levels of a specific mesenchymal signature gene, *VIM*, as an approximation for cells progressing through EMT [24–26]. 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 S2) [23,24]. 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

$$m_g = median(e_1, e_2, \ldots, e_c) \tag{1}$$

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

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

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

to fit the profile of a gene g concerning a pseudotime. In Equation (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 [27], but it can also include other sources of variability of gene expression, such as mutations and post-transcriptional regulations [28,29]. 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

$$w = a_1 w_{\mathrm{SDE}} + a_2 w_{\mathrm{MAD}} + a_3 w_{\mathrm{miRNA}} \tag{4}$$

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 [30–33]. We used the glmnet package in R to perform all the LASSO regularized Cox-model fittings (Hastie Lab, Stanford, CA, USA, 4.1-6) [34,35].

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

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

The C-index is equal to the concordance probability $p\,(r_i > r_j \mid z_i < z_j)$ 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 [36].

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) [37]. 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

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

O is the total of the observed events and 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 O and 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 [38].

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 [17]. 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) [39]. 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) [40]. 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.

3. 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 [17], 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 (Figures S3 and S4), 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.

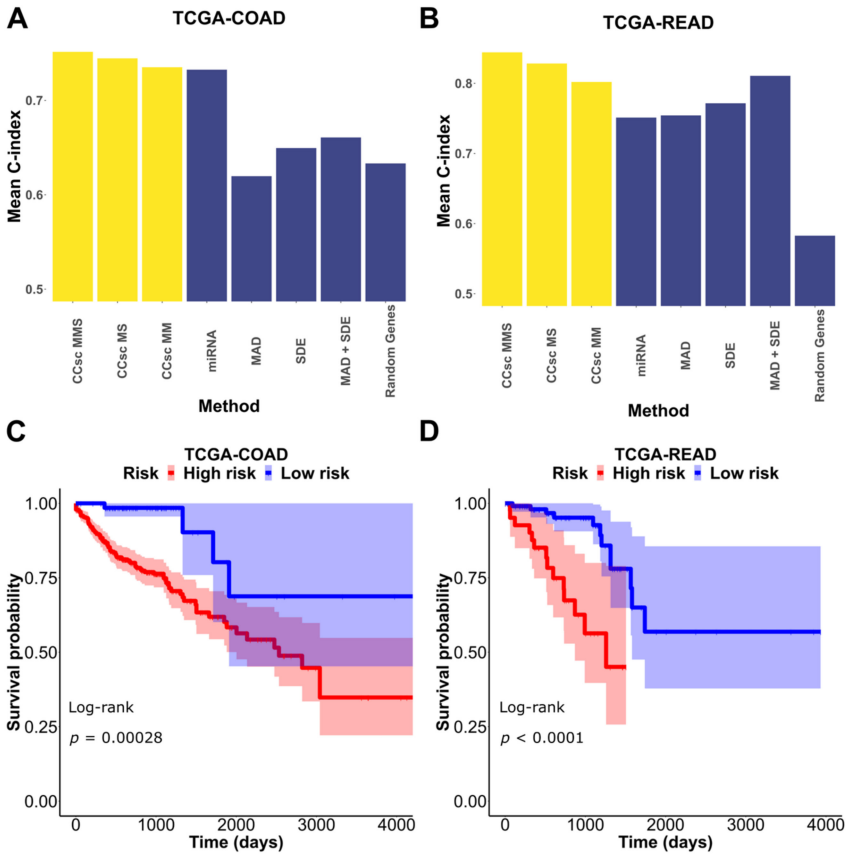

Figure 2. Combination of metrics is better than any individual metric. (**A**) We tested each of our models' metrics and compared them to CCsc MS, CCsc MM, and CCsc MMS and a set of randomly selected genes, equal to the number of genes used by our method on the TCGA-COAD dataset. (**B**) We made the 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).

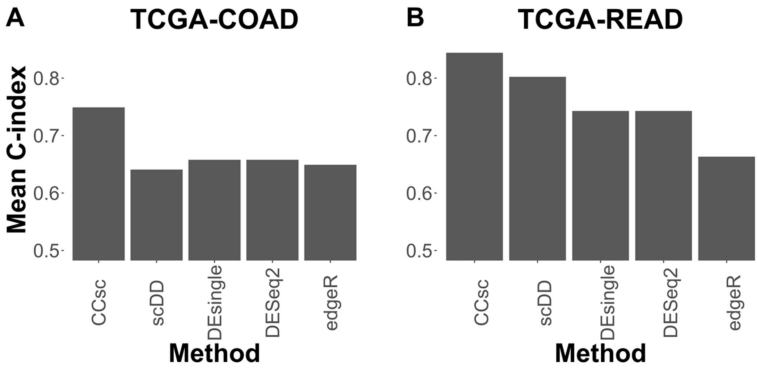

Figure 3. CCsc MMS Outperforms Other Methods. (**A**) We tested CCsc against several well-established tools on the TCGA-COAD dataset. We compared the mean concordance index performance of CCsc with its ideal weighting to DESeq2, edgeR, scDD, and DEsingle. We gave each method its optimal number of genes 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).

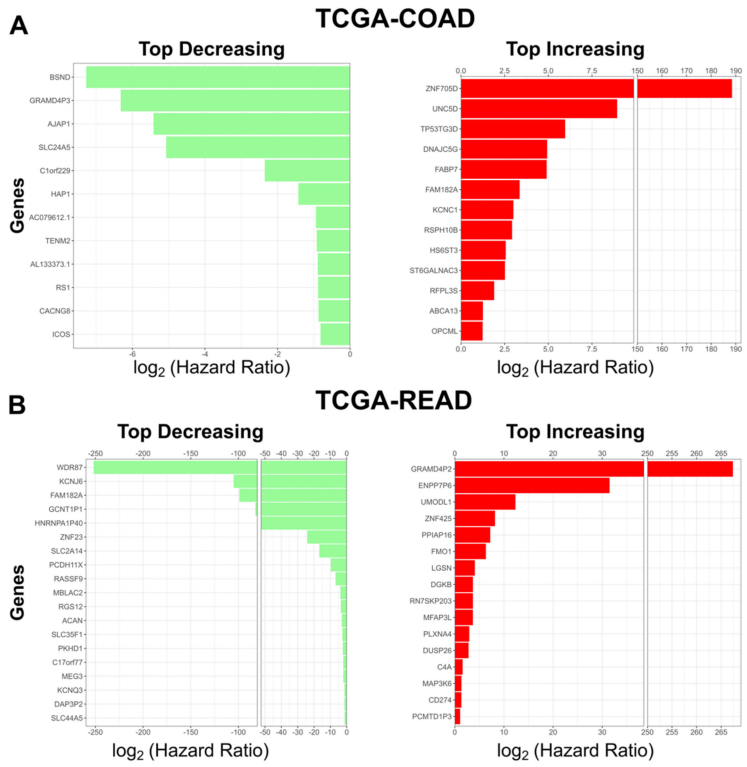

Figure 4. Top coefficients identified by our model. (**A**) Top risk decreasing (green, left) and top risk increasing (red, right) genes identified by our Cox model in TCGA-COAD. (**B**) Top risk decreasing genes (green, left) and top risk increasing (red, right) genes identified by our Cox model in 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	[41]
UNC5D	[42]
TP53TG3D	[43,44]
ST6GALNAC3	[45,46]
RSPH10B	[47]
KCNC1	[48,49]
HS6ST3	[50,51]
FAM182A	None
FABP7	[52,53]
DNAJC5G	[54]
ABCA13	[55]
OPCML	[56]
RFPL3S	[57]

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	[58]
UMODL1	[59]
RN7SKP203	Pseudogene
PPIAP16	Pseudogene
PLXNA4	[60,61]
MFAP3L	[62,63]
LGSN	[64]
GRAMD4P2	Pseudogene
FMO1	[65]
ENPP7P6	Pseudogene
DUSP26	[66,67]
DGKB	[68,69]
MAP3K6	[70]
CD274	[71]
C4A	[72]
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	[73]
SLC24A5	None (potassium-dependent sodium/calcium exchanger)
CACNG8	[74]
C1orf229	None
GRAMD4P3	Pseudogene
ICOS	[75]
HAP1	[76]
TENM2	[77]
AC079612.1	[78]
AL133373.1	None
BSND	[79]
RS1	[80]

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	[81]
RGS12	[82]
PKHD1	[83]
GCNT1P1	Pseudogene
KCNJ6	None (potassium channel)
RASSF9	[84]
DAP3P2	Pseudogene
WDR87	None
KCNQ3	[85]
PCDH11X	[86]
MEG3	[87]
MBLAC2	[88]
SLC44A5	None (Predicted to enable transmembrane transporter activity)
HNRNPA1P40	Pseudogene
ZNF23	[89]
ACAN	[90]
SLC2A14	[91]

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 (Tables 5 and 6). The pathways impacted by the top genes have been implicated in CRC progression and metastasis [92–97].

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	[92]	1.98×10^{-2}
Voltage gated Potassium channels	[93]	1.98×10^{-2}
Signaling by NOTCH3	[94]	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	[98]	1.48×10^{-3}
STAT3 nuclear events downstream of ALK signaling	[99]	4.74×10^{-3}
Signaling by ALK	[100]	1.72×10^{-2}
FOXO-mediated transcription of oxidative stress, metabolic and neuronal genes	[96]	1.72×10^{-2}

4. 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).

5. 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.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/cells12020228/s1, Figure S1: miRNA metric; Figure S2: Pseudotemporal ordering analysis; Figure S3: LASSO regularization gives the best performance; Figure S4: Determination of Ideal miRNA and miRNA target values; Figure S5: CCsc MMS Uses Comparable Gene Set Size to Existing Methods; Figure S6: CCsc MMS Outperforms other Methods in Non-TCGA CRC Data; Figure S7: Highly Ranked Gene Targets Show Reduced Expression Level. Table S1: Genes Identified by Model Regulated by miRNA. Genes are either directly or indirectly regulated by miRNAs.

Author Contributions: Conceptualization, A.W. and T.H; methodology, A.W.; software, A.W.; validation, A.W.; investigation, A.W.; resources, N.P.; data curation, A.W.; writing—original draft preparation, A.W. and T.H.; writing—review and editing, A.W. and T.H.; visualization, A.W.; supervision, T.H.; project administration, T.H.; funding acquisition, T.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Institutes of Health, grant number R01GM140162 (to T.H.).

Data Availability Statement: All computer code of this study can be found at https://github.com/compbiolover/CC-Singlecell.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **2014**, *15*, 550. [CrossRef] [PubMed]
2. Robinson, M.D.; McCarthy, D.J.; Smyth, G.K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **2010**, *26*, 139–140. [CrossRef] [PubMed]
3. Lander, E.S.; Linton, L.M.; Birren, B.; Nusbaum, C.; Zody, M.C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W. Initial sequencing and analysis of the human genome. *Nature* **2001**, *409*, 860–921. [PubMed]
4. Venter, J.C.; Adams, M.D.; Myers, E.W.; Li, P.W.; Mural, R.J.; Sutton, G.G.; Smith, H.O.; Yandell, M.; Evans, C.A.; Holt, R.A. The sequence of the human genome. *Science* **2001**, *291*, 1304–1351. [CrossRef] [PubMed]
5. Nagalakshmi, U.; Wang, Z.; Waern, K.; Shou, C.; Raha, D.; Gerstein, M.; Snyder, M. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **2008**, *320*, 1344–1349. [CrossRef]
6. Qian, M.; Wang, D.C.; Chen, H.; Cheng, Y. Detection of single cell heterogeneity in cancer. In *Seminars in Cell & Developmental Biology*; Academic Press: Cambridge, MA, USA, 2017; pp. 143–149.
7. Tellez-Gabriel, M.; Ory, B.; Lamoureux, F.; Heymann, M.-F.; Heymann, D. Tumour heterogeneity: The key advantages of single-cell analysis. *Int. J. Mol. Sci.* **2016**, *17*, 2142. [CrossRef]
8. Zhao, N.; Rosen, J.M. Breast cancer heterogeneity through the lens of single-cell analysis and spatial pathologies. In *Seminars in Cancer Biology*; Academic Press: Cambridge, MA, USA, 2021.
9. Li, X.; Liu, L.; Goodall, G.J.; Schreiber, A.; Xu, T.; Li, J.; Le, T.D. A novel single-cell based method for breast cancer prognosis. *PLoS Comput. Biol.* **2020**, *16*, e1008133. [CrossRef]
10. Bartel, D.P. MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell* **2004**, *116*, 281–297. [CrossRef]
11. Di Leva, G.; Garofalo, M.; Croce, C.M. MicroRNAs in cancer. *Annu. Rev. Pathol. Mech. Dis.* **2014**, *9*, 287–314. [CrossRef]
12. Lu, T.X.; Rothenberg, M.E. MicroRNA. *J. Allergy Clin. Immunol.* **2018**, *141*, 1202–1207. [CrossRef]

13. Campbell, K.R.; Yau, C. switchde: Inference of switch-like differential expression along single-cell trajectories. *Bioinformatics* **2017**, *33*, 1241–1242. [CrossRef]
14. Howell, D.C. Median absolute deviation. *Encycl. Stat. Behav. Sci.* **2005**.
15. Yang, Z.; Wu, L.; Wang, A.; Tang, W.; Zhao, Y.; Zhao, H.; Teschendorff, A.E. dbDEMC 2.0: Updated database of differentially expressed miRNAs in human cancers. *Nucleic Acids Res.* **2017**, *45*, D812–D818. [CrossRef]
16. Vejnar, C.E.; Zdobnov, E.M. MiRmap: Comprehensive prediction of microRNA target repression strength. *Nucleic Acids Res.* **2012**, *40*, 11673–11683. [CrossRef]
17. Li, H.; Courtois, E.T.; Sengupta, D.; Tan, Y.; Chen, K.H.; Goh, J.J.L.; Kong, S.L.; Chua, C.; Hon, L.K.; Tan, W.S. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet.* **2017**, *49*, 708–718. [CrossRef]
18. Lab, K. Rmagic EMT Tutorial. Available online: http://htmlpreview.github.io/?https://github.com/KrishnaswamyLab/MAGIC/blob/master/Rmagic/inst/examples/emt_tutorial.html (accessed on 1 December 2020).
19. Van Dijk, D.; Sharma, R.; Nainys, J.; Yim, K.; Kathail, P.; Carr, A.J.; Burdziak, C.; Moon, K.R.; Chaffer, C.L.; Pattabiraman, D. Recovering gene interactions from single-cell data using data diffusion. *Cell* **2018**, *174*, 716–729.e27. [CrossRef]
20. Cao, J.; Spielmann, M.; Qiu, X.; Huang, X.; Ibrahim, D.M.; Hill, A.J.; Zhang, F.; Mundlos, S.; Christiansen, L.; Steemers, F.J. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **2019**, *566*, 496–502. [CrossRef]
21. Rokavec, M.; Öner, M.G.; Li, H.; Jackstadt, R.; Jiang, L.; Lodygin, D.; Kaller, M.; Horst, D.; Ziegler, P.K.; Schwitalla, S. IL-6R/STAT3/miR-34a feedback loop promotes EMT-mediated colorectal cancer invasion and metastasis. *J. Clin. Investig.* **2014**, *124*, 1853–1867. [CrossRef]
22. Vu, T.; Datta, P.K. Regulation of EMT in colorectal cancer: A culprit in metastasis. *Cancers* **2017**, *9*, 171. [CrossRef]
23. Panchy, N.; Azeredo-Tseng, C.; Luo, M.; Randall, N.; Hong, T. Integrative transcriptomic analysis reveals a multiphasic epithelial–mesenchymal spectrum in cancer and non-tumorigenic cells. *Front. Oncol.* **2020**, *9*, 1479. [CrossRef]
24. 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*, lqac072. [CrossRef] [PubMed]
25. Wang, Z.; Divanyan, A.; Jourd'heuil, F.L.; Goldman, R.D.; Ridge, K.M.; Jourd'heuil, D.; Lopez-Soler, R.I. Vimentin expression is required for the development of EMT-related renal fibrosis following unilateral ureteral obstruction in mice. *Am. J. Physiol.-Ren. Physiol.* **2018**, *315*, F769–F780. [CrossRef] [PubMed]
26. Loboda, A.; Nebozhyn, M.V.; Watters, J.W.; Buser, C.A.; Shaw, P.M.; Huang, P.S.; Van't Veer, L.; Tollenaar, R.A.; Jackson, D.B.; Agrawal, D. EMT is the dominant program in human colon cancer. *BMC Med. Genom.* **2011**, *4*, 9. [CrossRef] [PubMed]
27. Zhang, J.; Nie, Q.; Zhou, T. Revealing dynamic mechanisms of cell fate decisions from single-cell transcriptomic data. *Front. Genet.* **2019**, *10*, 1280. [CrossRef] [PubMed]
28. Hu, X.; Stern, H.M.; Ge, L.; O'Brien, C.; Haydu, L.; Honchell, C.D.; Haverty, P.M.; Peters, B.A.; Wu, T.D.; Amler, L.C. Genetic alterations and oncogenic pathways associated with breast cancer subtypes. *Mol. Cancer Res.* **2009**, *7*, 511–522. [CrossRef]
29. Nordick, B.; Yu, P.Y.; Liao, G.; Hong, T. Nonmodular oscillator and switch based on RNA decay drive regeneration of multimodal gene expression. *Nucleic Acids Res.* **2022**, *50*, 3693–3708. [CrossRef]
30. Zhao, P.; Yu, B. On model selection consistency of Lasso. *J. Mach. Learn. Res.* **2006**, *7*, 2541–2563.
31. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.* **1996**, *58*, 267–288. [CrossRef]
32. Yamada, M.; Jitkrittum, W.; Sigal, L.; Xing, E.P.; Sugiyama, M. High-dimensional feature selection by feature-wise kernelized lasso. *Neural Comput.* **2014**, *26*, 185–207. [CrossRef]
33. Muthukrishnan, R.; Rohini, R. LASSO: A feature selection technique in predictive modeling for machine learning. In Proceedings of the 2016 IEEE International Conference on Advances in Computer Applications (ICACA), Coimbatore, India, 24 October 2016; pp. 18–20.
34. Simon, N.; Friedman, J.; Hastie, T.; Tibshirani, R. Regularization paths for Cox's proportional hazards model via coordinate descent. *J. Stat. Softw.* **2011**, *39*, 1–13. [CrossRef]
35. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2005**, *67*, 301–320. [CrossRef]
36. Uno, H.; Cai, T.; Pencina, M.J.; D'Agostino, R.B.; Wei, L.-J. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat. Med.* **2011**, *30*, 1105–1117. [CrossRef]
37. Goel, M.K.; Khanna, P.; Kishore, J. Understanding survival analysis: Kaplan-Meier estimate. *Int. J. Ayurveda Res.* **2010**, *1*, 274.
38. Bland, J.M.; Altman, D.G. The logrank test. *Bmj* **2004**, *328*, 1073. [CrossRef]
39. Colaprico, A.; Silva, T.C.; Olsen, C.; Garofano, L.; Cava, C.; Garolini, D.; Sabedot, T.S.; Malta, T.M.; Pagnotta, S.M.; Castiglioni, I. TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* **2016**, *44*, e71. [CrossRef]
40. Chatila, W.K.; Kim, J.K.; Walch, H.; Marco, M.R.; Chen, C.-T.; Wu, F.; Omer, D.M.; Khalil, D.N.; Ganesh, K.; Qu, X. Genomic and transcriptomic determinants of response to neoadjuvant therapy in rectal cancer. *Nat. Med.* **2022**, *28*, 1646–1655. [CrossRef]
41. Pratama, R.; Hwang, J.J.; Lee, J.H.; Song, G.; Park, H.R. Authentication of differential gene expression in oral squamous cell carcinoma using machine learning applications. *BMC Oral Health* **2021**, *21*, 281. [CrossRef]
42. Zhu, Y.; Li, Y.; Nakagawara, A. UNC5 dependence receptor family in human cancer: A controllable double-edged sword. *Cancer Lett.* **2021**, *516*, 28–35. [CrossRef]

43. 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**, arXiv:2021.07.30.454550.
44. Lu, Q.; Guo, Q.; Xin, M.; Lim, C.; Gamero, A.M.; Gerhard, G.S.; Yang, L. LncRNA TP53TG1 Promotes the Growth and Migration of Hepatocellular Carcinoma Cells via Activation of ERK Signaling. *Non-Coding RNA* **2021**, *7*, 52. [CrossRef]
45. Dong, Y.; Zhang, T.; Li, X.; Yu, F.; Guo, Y. Comprehensive analysis of coexpressed long noncoding RNAs and genes in breast cancer. *J. Obstet. Gynaecol. Res.* **2019**, *45*, 428–437. [CrossRef] [PubMed]
46. Kasper, B.T.; Koppolu, S.; Mahal, L.K. Insights into miRNA regulation of the human glycome. *Biochem. Biophys. Res. Commun.* **2014**, *445*, 774–779. [CrossRef] [PubMed]
47. Brim, H.; Boulares, H.; Daremipouran, M.; Lee, E.; Ashktorab, H. Streptococcus sp. VT_162 infection of colon cancer cell lines induces mRNAs that associate with poor prognosis. *Cancer Res.* **2020**, *80*, 6102. [CrossRef]
48. Chen, S.; Xiao, L.; Peng, H.; Wang, Z.; Xie, J. Methylation gene KCNC1 is associated with overall survival in patients with seminoma. *Oncol. Rep.* **2021**, *45*, 73. [CrossRef] [PubMed]
49. Gu, X.-Y.; Jin, B.; Qi, Z.-D.; Yin, X.-F. MicroRNA is a potential target for therapies to improve the physiological function of skeletal muscle after trauma. *Neural Regen. Res.* **2022**, *17*, 1617.
50. Iravani, O.; Bay, B.-H.; Yip, G.W.-C. Silencing HS6ST3 inhibits growth and progression of breast cancer cells through suppressing IGF1R and inducing XAF1. *Exp. Cell Res.* **2017**, *350*, 380–389. [CrossRef]
51. Guo, Y.; Min, Z.; Jiang, C.; Wang, W.; Yan, J.; Xu, P.; Xu, K.; Xu, J.; Sun, M.; Zhao, Y. Downregulation of HS6ST2 by miR-23b-3p enhances matrix degradation through p38 MAPK pathway in osteoarthritis. *Cell Death Dis.* **2018**, *9*, 1–15. [CrossRef]
52. Ma, R.; Wang, L.; Yuan, F.; Wang, S.; Liu, Y.; Fan, T.; Wang, F. FABP7 promotes cell proliferation and survival in colon cancer through MEK/ERK signaling pathway. *Biomed. Pharmacother.* **2018**, *108*, 119–129. [CrossRef]
53. Tian, X.; Yang, H.; Fang, Q.; Quan, H.; Lu, H.; Wang, X. Circ_ZFR affects FABP7 expression to regulate breast cancer progression by acting as a sponge for miR-223-3p. *Thorac. Cancer* **2022**, *13*, 1369–1380. [CrossRef]
54. Misawa, K.; Imai, A.; Matsui, H.; Kanai, A.; Misawa, Y.; Mochizuki, D.; Mima, M.; Yamada, S.; Kurokawa, T.; Nakagawa, T. Identification of novel methylation markers in HPV-associated oropharyngeal cancer: Genome-wide discovery, tissue verification and validation testing in ctDNA. *Oncogene* **2020**, *39*, 4741–4755. [CrossRef]
55. Araújo, T.; Seabra, A.; Lima, E.; Assumpção, P.; Montenegro, R.; Demachki, S.; Burbano, R.; Khayat, A. Recurrent amplification of RTEL1 and ABCA13 and its synergistic effect associated with clinicopathological data of gastric adenocarcinoma. *Mol. Cytogenet.* **2016**, *9*, 52. [CrossRef]
56. Li, C.; Tang, L.; Zhao, L.; Li, L.; Xiao, Q.; Luo, X.; Peng, W.; Ren, G.; Tao, Q.; Xiang, T. OPCML is frequently methylated in human colorectal cancer and its restored expression reverses EMT via downregulation of smad signaling. *Am. J. Cancer Res.* **2015**, *5*, 1635.
57. Guo, J.; Wang, S.; Jiang, Z.; Tang, L.; Liu, Z.; Cao, J.; Hu, Z.; Chen, X.; Luo, Y.; Bo, H. Long Non-Coding RNA RFPL3S Functions as a Biomarker of Prognostic and Immunotherapeutic Prediction in Testicular Germ Cell Tumor. *Front. Immunol.* **2022**, *13*, 859730. [CrossRef]
58. Wang, Y.; Ye, X.; Zhou, J.; Wan, Y.; Xie, H.; Deng, Y.; Yan, Y.; Li, Y.; Fan, X.; Yuan, W. A novel human KRAB-related zinc finger gene ZNF425 inhibits mitogen-activated protein kinase signaling pathway. *BMB Rep.* **2011**, *44*, 58–63. [CrossRef]
59. Wang, W.; Tang, Y.; Ni, L.; Kim, E.; Jongwutiwes, T.; Hourvitz, A.; Zhang, R.; Xiong, H.; Liu, H.; Rosenwaks, Z. Overexpression of Uromodulin-like1 accelerates follicle depletion and subsequent ovarian degeneration. *Cell Death Dis.* **2012**, *3*, e433. [CrossRef]
60. Kigel, B.; Rabinowicz, N.; Varshavsky, A.; Kessler, O.; Neufeld, G. Plexin-A4 promotes tumor progression and tumor angiogenesis by enhancement of VEGF and bFGF signaling. *Blood J. Am. Soc. Hematol.* **2011**, *118*, 4285–4296. [CrossRef]
61. Mawaribuchi, S.; Aiki, Y.; Ikeda, N.; Ito, Y. mRNA and miRNA expression profiles in an ectoderm-biased substate of human pluripotent stem cells. *Sci. Rep.* **2019**, *9*, 1–13. [CrossRef]
62. Lou, X.; Kang, B.; Zhang, J.; Hao, C.; Tian, X.; Li, W.; Xu, N.; Lu, Y.; Liu, S. MFAP3L activation promotes colorectal cancer cell invasion and metastasis. *Biochim. Biophys. Acta* **2014**, *1842*, 1423–1432. [CrossRef]
63. Ye, J.; Luo, W.; Luo, L.; Zhai, L.; Huang, P. MicroRNA-671-5p inhibits cell proliferation, migration and invasion in non-small cell lung cancer by targeting MFAP3L. *Mol. Med. Rep.* **2022**, *25*, 1–8. [CrossRef]
64. Nakatsugawa, M.; Hirohashi, Y.; Torigoe, T.; Asanuma, H.; Takahashi, A.; Inoda, S.; Kiriyama, K.; Nakazawa, E.; Harada, K.; Takasu, H. Novel spliced form of a lens protein as a novel lung cancer antigen, Lengsin splicing variant 4. *Cancer Sci.* **2009**, *100*, 1485–1493. [CrossRef]
65. Zhang, T.; Yang, P.; Wei, J.; Li, W.; Zhong, J.; Chen, H.; Cao, J. Overexpression of flavin-containing monooxygenase 5 predicts poor prognosis in patients with colorectal cancer. *Oncol. Lett.* **2018**, *15*, 3923–3927. [CrossRef] [PubMed]
66. Yu, W.; Imoto, I.; Inoue, J.; Onda, M.; Emi, M.; Inazawa, J. A novel amplification target, DUSP26, promotes anaplastic thyroid cancer cell growth by inhibiting p38 MAPK activity. *Oncogene* **2007**, *26*, 1178–1187. [CrossRef] [PubMed]
67. Thompson, E.M.; Stoker, A.W. A review of DUSP26: Structure, regulation and relevance in human disease. *Int. J. Mol. Sci.* **2021**, *22*, 776. [CrossRef] [PubMed]
68. Zhang, Z.Y.; Yao, Q.Z.; Liu, H.Y.; Guo, Q.N.; Qiu, P.J.; Chen, J.P.; Lin, J.Q. Metabolic reprogramming-associated genes predict overall survival for rectal cancer. *J. Cell. Mol. Med.* **2020**, *24*, 5842–5849. [CrossRef] [PubMed]
69. Kefas, B.; Floyd, D.H.; Comeau, L.; Frisbee, A.; Dominguez, C.; Dipierro, C.G.; Guessous, F.; Abounader, R.; Purow, B. A miR-297/hypoxia/DGK-α axis regulating glioblastoma survival. *Neuro-Oncol.* **2013**, *15*, 1652–1663. [CrossRef]

70. Gaston, D.; Hansford, S.; Oliveira, C.; Nightingale, M.; Pinheiro, H.; Macgillivray, C.; Kaurah, P.; Rideout, A.L.; Steele, P.; Soares, G. Germline mutations in MAP3K6 are associated with familial gastric cancer. *PLoS Genet.* **2014**, *10*, e1004669. [CrossRef]
71. Kula, A.; Dawidowicz, M.; Kiczmer, P.; Seńkowska, A.P.; Świętochowska, E. The role of genetic polymorphism within PD-L1 gene in cancer. Review. *Exp. Mol. Pathol.* **2020**, *116*, 104494. [CrossRef]
72. Wang, Y.; Li, C.; Qi, X.; Yao, Y.; Zhang, L.; Zhang, G.; Xie, L.; Wang, Q.; Zhu, W.; Guo, X. 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. *Int. J. Mol. Sci.* **2022**, *23*, 8799. [CrossRef]
73. Han, J.; Xie, C.; Pei, T.; Wang, J.; Lan, Y.; Huang, K.; Cui, Y.; Wang, F.; Zhang, J.; Pan, S. Deregulated AJAP1/β-catenin/ZEB1 signaling promotes hepatocellular carcinoma carcinogenesis and metastasis. *Cell Death Dis.* **2017**, *8*, e2736. [CrossRef]
74. He, Z.; Li, B. Recent progress in genetic and epigenetic profile of diffuse gastric cancer. *Cancer Transl. Med.* **2015**, *1*, 80–93.
75. Shamsdin, S.A.; Karimi, M.H.; Hosseini, S.V.; Geramizadeh, B.; Fattahi, M.R.; Mehrabani, D.; Moravej, A. Associations of ICOS and PD. 1 gene variants with colon cancer risk in the Iranian population. *Asian Pac. J. Cancer Prev.* **2018**, *19*, 693.
76. Peng, L.; Liu, Y.; Chen, J.; Cheng, M.; Wu, Y.; Chen, M.; Zhong, Y.; Shen, D.; Chen, L.; Ye, X. APEX1 regulates alternative splicing of key tumorigenesis genes in non-small-cell lung cancer. *BMC Med. Genom.* **2022**, *15*, 147. [CrossRef]
77. Peppino, G.; Ruiu, R.; Arigoni, M.; Riccardo, F.; Iacoviello, A.; Barutello, G.; Quaglino, E. Teneurins: Role in Cancer and Potential Role as Diagnostic Biomarkers and Targets for Therapy. *Int. J. Mol. Sci.* **2021**, *22*, 2321. [CrossRef]
78. Sun, Y.; Peng, P.; He, L.; Gao, X. Identification of lnc RNAs related to prognosis of patients with colorectal cancer. *Technol. Cancer Res. Treat.* **2020**, *19*, 1533033820962120. [CrossRef]
79. Shinmura, K.; Igarashi, H.; Kato, H.; Koda, K.; Ogawa, H.; Takahashi, S.; Otsuki, Y.; Yoneda, T.; Kawanishi, Y.; Funai, K. BSND and ATP6V1G3: Novel immunohistochemical markers for chromophobe renal cell carcinoma. *Medicine* **2015**, *94*, e989. [CrossRef]
80. Zhang, T.; Cheng, G.; Chen, P.; Peng, Y.; Liu, L.; Li, R.; Qiu, B. RS1 gene is a novel prognostic biomarker for lung adenocarcinoma. *Thorac. Cancer* **2022**, *13*, 1850–1861. [CrossRef]
81. Winter, G.E.; Radic, B.; Mayor-Ruiz, C.; Blomen, V.A.; Trefzer, C.; Kandasamy, R.K.; Huber, K.V.; Gridling, M.; Chen, D.; Klampfl, T. The solute carrier SLC35F2 enables YM155-mediated DNA damage toxicity. *Nat. Chem. Biol.* **2014**, *10*, 768–773. [CrossRef]
82. Wang, Y.; Wang, J.; Zhang, L.; Karatas, O.F.; Shao, L.; Zhang, Y.; Castro, P.; Creighton, C.J.; Ittmann, M. RGS12 Is a Novel Tumor-Suppressor Gene in African American Prostate Cancer That Represses AKT and MNX1 ExpressionRGS12 in African American Prostate Cancer. *Cancer Res.* **2017**, *77*, 4247–4257. [CrossRef]
83. Ward, C.J.; Wu, Y.; Johnson, R.A.; Woollard, J.R.; Bergstralh, E.J.; Cicek, M.S.; Bakeberg, J.; Rossetti, S.; Heyer, C.M.; Petersen, G.M. Germline PKHD1 mutations are protective against colorectal cancer. *Hum. Genet.* **2011**, *129*, 345–349. [CrossRef]
84. Xu, Q.; Yang, H.; Fan, G.; Zhang, B.; Yu, J.; Zhang, Z.; Jia, G. 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. *Pathol.-Res. Pract.* **2022**, *238*, 154092. [CrossRef]
85. Shorthouse, D.; Zhuang, J.L.; Rahrmann, E.P.; Kosmidou, C.; Rahrmann, K.W.; Hall, M.; Greenwood, B.; Devonshire, G.; Gilbertson, R.J.; Fitzgerald, R.C. The Role of Potassium Channels in the Pathogenesis of Gastrointestinal Cancers and Therapeutic Potential. *bioRxiv* **2022**, arXiv:2020.03.10.984039.
86. Pancho, A.; Aerts, T.; Mitsogiannis, M.D.; Seuntjens, E. Protocadherins at the crossroad of signaling pathways. *Front. Mol. Neurosci.* **2020**, *13*, 117. [CrossRef]
87. Ghafouri-Fard, S.; Taheri, M. Maternally expressed gene 3 (MEG3): A tumor suppressor long non coding RNA. *Biomed. Pharmacother.* **2019**, *118*, 109129. [CrossRef] [PubMed]
88. Lechner, S.; Malgapo, M.I.P.; Grätz, C.; Steimbach, R.R.; Baron, A.; Rüther, P.; Nadal, S.; Stumpf, C.; Loos, C.; Ku, X. Target deconvolution of HDAC pharmacopoeia reveals MBLAC2 as common off-target. *Nat. Chem. Biol.* **2022**, *18*, 812–820. [CrossRef] [PubMed]
89. Zhang, X.; Ding, C.; Tian, H.; Dong, X.; Meng, X.; Zhu, W.; Liu, B.; Wang, L.; Huang, M.; Li, C. ZNF23 suppresses cutaneous melanoma cell malignancy via mitochondria-dependent pathway. *Cell. Physiol. Biochem.* **2017**, *43*, 147–157. [CrossRef]
90. Vafaeie, F.; Nomiri, S.; Ranjbaran, J.; Safarpour, H. ACAN, MDFI, and CHST1 as Candidate Genes in Gastric Cancer: A Comprehensive Insilco Analysis. *Asian Pac. J. Cancer Prev.* **2022**, *23*, 683–694. [CrossRef] [PubMed]
91. Amir Shaghaghi, M.; Zhouyao, H.; Tu, H.; El-Gabalawy, H.; Crow, G.H.; Levine, M.; Bernstein, C.N.; Eck, P. The SLC2A14 gene, encoding the novel glucose/dehydroascorbate transporter GLUT14, is associated with inflammatory bowel disease. *Am. J. Clin. Nutr.* **2017**, *106*, 1508–1513. [CrossRef]
92. Serafin, V.; Persano, L.; Moserle, L.; Esposito, G.; Ghisi, M.; Curtarello, M.; Bonanno, L.; Masiero, M.; Ribatti, D.; Stürzl, M. Notch3 signalling promotes tumour growth in colorectal cancer. *J. Pathol.* **2011**, *224*, 448–460. [CrossRef]
93. Abdul, M.; Hoosein, N. Voltage-gated potassium ion channels in colon cancer. *Oncol. Rep.* **2002**, *9*, 961–964. [CrossRef]
94. Varga, J.; Nicolas, A.; Petrocelli, V.; Pesic, M.; Mahmoud, A.; Michels, B.E.; Etlioglu, E.; Yepes, D.; Häupl, B.; Ziegler, P.K. AKT-dependent NOTCH3 activation drives tumor progression in a model of mesenchymal colorectal cancer. *J. Exp. Med.* **2020**, *217*, e20191515. [CrossRef]
95. Pardo, L.A.; Stühmer, W. The roles of K+ channels in cancer. *Nat. Rev. Cancer* **2014**, *14*, 39–48. [CrossRef]
96. Farhan, M.; Silva, M.; Xingan, X.; Huang, Y.; Zheng, W. Role of FOXO transcription factors in cancer metabolism and angiogenesis. *Cells* **2020**, *9*, 1586. [CrossRef]

97. Ma, J.; Matkar, S.; He, X.; Hua, X. FOXO family in regulating cancer and metabolism. In *Seminars in Cancer Biology*; Academic Press: Cambridge, MA, USA, 2018; pp. 32–41.
98. Rutkowski, M.J.; Sughrue, M.E.; Kane, A.J.; Mills, S.A.; Parsa, A.T. Cancer and the Complement CascadeCancer and the Complement Cascade. *Mol. Cancer Res.* **2010**, *8*, 1453–1465. [CrossRef]
99. Gu, Y.; Mohammad, I.S.; Liu, Z. Overview of the STAT-3 signaling pathway in cancer and the development of specific inhibitors. *Oncol. Lett.* **2020**, *19*, 2585–2594. [CrossRef]
100. Pilling, A.B.; Kim, J.; Estrada-Bernal, A.; Zhou, Q.; Le, A.T.; Singleton, K.R.; Heasley, L.E.; Tan, A.C.; DeGregori, J.; Doebele, R.C. ALK is a critical regulator of the MYC-signaling axis in ALK positive lung cancer. *Oncotarget* **2018**, *9*, 8823. [CrossRef]