

RESEARCH ARTICLE

STmiR: A Novel XGBoost-based framework for spatially resolved miRNA activity prediction in cancer transcriptomics

Jiaqi Yuan¹, Peng Xu^{1,2*}, Zheng Ye^{1*}, Wenbin Liu^{1*}

1 Institute of Computational Science and Technology, Guangzhou University, Guangzhou, China,

2 School of Computer Science of Information Technology, Qiannan Normal University for Nationalities, Duyun, Guizhou, China

* gdxupeng@gzhu.edu.cn (PX); wbliu6910@gzhu.edu.cn (WL); zheng_ye@gzhu.edu.cn (ZY)



OPEN ACCESS

Citation: Yuan J, Xu P, Ye Z, Liu W (2025) STmiR: A Novel XGBoost-based framework for spatially resolved miRNA activity prediction in cancer transcriptomics. PLoS One 20(9): e0322082. <https://doi.org/10.1371/journal.pone.0322082>

Received: March 15, 2025

Accepted: August 27, 2025

Published: September 9, 2025

Copyright: © 2025 Yuan et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data availability statement: All relevant data are within the manuscript and its [Supporting Information](#) files.

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

MicroRNAs (miRNAs) are critical regulators of gene expression in cancer biology, yet their spatial dynamics within tumor microenvironments (TMEs) remain underexplored due to technical limitations in current spatial transcriptomics (ST) technologies. To address this gap, we present STmiR, a novel XGBoost-based framework for spatially resolved miRNA activity prediction. STmiR integrates bulk RNA-seq data (TCGA and CCLE) with spatial transcriptomics profiles to model nonlinear miRNA-mRNA interactions, achieving high predictive accuracy (Spearman's $\rho > 0.8$) across four major cancer types (breast, lung, ovarian, prostate), with performance further confirmed through direct comparison with experimentally measured miRNA expression in an independent spatial transcriptomics dataset. Applied to 10X Visium ST datasets from nine cancers, STmiR identifies six pan-cancer conserved miRNAs (e.g., hsa-miR-21, hsa-let-7a) consistently ranked in the top 40 across malignancies, and uncovers cell-type-specific regulatory networks in fibroblasts, B cells, and malignant cells. A breast cancer case study demonstrates STmiR's utility in uncovering biologically relevant miRNA-target relationships and their association with key cancer pathways. By enabling spatial mapping of miRNA activity, STmiR provides a transformative tool to dissect miRNA-mediated regulatory mechanisms in cancer progression and TME remodeling, with implications for biomarker discovery and precision oncology.

Introduction

MicroRNAs (miRNAs) are small, non-coding RNA molecules that play critical roles in post-transcriptional gene regulation by binding to target messenger RNAs (mRNAs), leading to their degradation or translational repression [1,2]. These regulatory mechanisms are essential for controlling a wide range of biological processes, including cell differentiation, development, and disease progression [3,4]. Dysregulation of miRNA expression has been implicated in various diseases such as cancer, cardiovascular

disorders, and neurodegenerative conditions, highlighting the importance of miRNAs as both biomarkers and potential therapeutic targets [5,6]. As such, accurate profiling of miRNA expression is crucial for uncovering the underlying regulatory networks and understanding gene regulation in both health and disease [7].

However, measuring miRNA expression, especially in the spatial context of tissues, remains a significant technical challenge [8]. Spatial transcriptomics (ST) technologies, while offering the ability to spatially resolve mRNA expression patterns within tissues, are not optimized for capturing miRNAs due to their small size and distinct structural properties [9]. This limitation has hindered the study of miRNA expression profiles in spatial transcriptomics, leaving gaps in our understanding of the spatial organization of gene regulation by miRNAs.

Several computational approaches have been developed to address this challenge by inferring miRNA activity indirectly. Current computational approaches (e.g., GenMIR++ and MAGIA) rely heavily on linear correlation models to infer miRNA-mRNA interactions. These methods assume a simplistic inverse relationship between miRNA and target mRNA expression levels, neglecting the nonlinear dynamics inherent in biological systems. For instance, GenMIR++ fails to account for cooperative miRNA binding or context-dependent regulatory effects, leading to reduced accuracy in complex tissues such as tumors [10]. To improve on these limitations, more recent machine learning-based approaches have emerged. For example, miRLAB integrates various biological features to predict miRNA-mRNA interactions [11], and DeepMirTar applies deep learning to enhance miRNA target prediction and infer miRNA activity [12]. Additionally, single-cell RNA sequencing (scRNA-seq) has spurred the development of methods such as miRSCAPE, which leverages the reduction in target mRNA expression to infer miRNA activity in single cells [13]. Although miRSCAPE is capable of predicting miRNA activity from single-cell sequencing data, the unique nature of spatial transcriptomics, where each spot resembles bulk data composed of multiple cells, makes it more appropriate to utilize models trained on bulk RNA-seq data for miRNA activity inference. The advancement of computational methods in spatial transcriptomics also provides new avenues for inferring molecular features that are not directly measured, highlighting the potential for model-based predictions [14,15].

Here, we present STmiR, the first machine learning framework that integrates XGBoost with spatial transcriptomics to predict miRNA activity. The complete workflow of our approach is outlined in S1 Fig. The process begins by training a predictive model on large-scale, paired miRNA-mRNA bulk RNA-seq datasets from sources like TCGA and CCLE. This trained model is then applied to the mRNA expression profiles from spatial transcriptomics data to infer miRNA activity for each spatial spot. Unlike single-cell tools such as miRSCAPE, which are optimized for single-cell resolution but struggle with spatially aggregated 'spot-level' data, STmiR leverages these bulk RNA-seq-trained models to capture nonlinear miRNA-mRNA relationships across heterogeneous tissue regions. This approach enables accurate spatial mapping of miRNA activity while preserving the multicellular complexity of tumor microenvironments, facilitating downstream functional analyses such as building cell-type-specific regulatory networks.

By training a model on matched miRNA-mRNA tissue samples, STmiR can learn the underlying relationships between RNA and miRNA. The trained model is then applied to spatial transcriptomics data, allowing the prediction of miRNA activity for each spot, thereby generating spatial maps of miRNA activity. This approach provides a powerful new tool for studying miRNA dynamics in spatial contexts, offering insights into the spatial regulation of gene expression that were previously inaccessible. In parallel, we conducted a pan-cancer analysis to identify highly expressed common and specific genes across cell types in different cancer types, providing further insights into the molecular underpinnings of cancer biology.

Throughout this manuscript, we make a clear distinction between “miRNA expression/abundance,” which refers to experimentally measured values (e.g., from miRNA-seq), and “predicted miRNA activity.” We define “**predicted miRNA activity**” as the quantitative estimation of a miRNA’s regulatory function and strength, as inferred by the STmiR model from the local mRNA expression profile. It serves as a composite metric reflecting the miRNA’s regulatory impact on its target gene network.

In summary, STmiR addresses the limitations of current miRNA profiling methods in spatial transcriptomics by leveraging machine learning to accurately infer miRNA expression activity from mRNA profiles. By enabling spatially resolved miRNA activity mapping, STmiR offers a powerful tool for investigating the spatiotemporal dynamics of miRNA-mediated regulation in tumor microenvironments, with potential applications in identifying therapeutic targets and deciphering mechanisms of cancer progression.

Methods

Data Source

Bulk-RNAseq. We procured the matched bulk RNA-Seq and miRNA-Seq gene expression data for cell lines from the Cancer Cell Line Encyclopedia [16] (CCLE, as described by Barretina et al. in 2012) and Pan-cancer data from The Cancer Genome Atlas (TCGA) [17] via the URLs provided at <https://pancanatlas.xenahubs.net>. The batch effects normalized mRNA-seq and miRNA-seq data were obtained from GDC (<https://portal.gdc.cancer.gov/>). Notably, the miRNA sequencing data encompassed 744 miRNAs across 10,818 samples, while the mRNA sequencing data comprised 20,532 genes across 11,060 samples. We utilized the TPM (Transcripts Per Kilobase Million) normalized RNA-seq data from CCLE [18] (as reported by Ghandi et al. in 2019).

Initially, following outlier processing and logarithmic transformation of the original TCGA and CCLE data, we selected the matched mRNA and miRNA samples from nine distinct sites. Subsequently, considering the miRNA names, we removed the long suffixes (such as -3p), eliminated duplicates of the same miRNA name, and replaced them with the median value. Z-score normalization was then employed for data standardization. Finally, we utilized rank-based methods to integrate the sequencing data from both TCGA and CCLE sources, with the aim of eliminating batch effects.

Integration and normalization of TCGA and CCLE data

To eliminate batch effects between the TCGA and CCLE datasets, we implemented a rigorous rank-based normalization and integration pipeline. The detailed steps are as follows:

Step 1: Independent Pre-processing of Datasets

1. Data Filtering: We first processed the mRNA and miRNA expression matrices from TCGA and CCLE independently. We retained only the genes and miRNAs that were common to both datasets. Genes/miRNAs with zero expression in more than 90% of samples were removed.
2. Log Transformation: To stabilize variance and reduce the influence of outliers, we applied a $\log_2(x + 1)$ transformation to the expression values (e.g., TPM or FPKM).

Step 2: Rank-Based Normalization

1. **Intra-Sample Ranking:** Within each sample (i.e., each column of the TCGA and CCLE matrices), we ranked the log2-transformed expression values of all genes from lowest to highest and calculated their percentile rank. This step converts the expression distribution of each sample into a uniform distribution from 0 to 1, removing sample-specific distributional differences.
2. **Cross-Sample Rank-based Inverse Normal Transformation:** For each gene (i.e., each row of the matrices), we collected its percentile ranks across all samples. These ranks were then mapped to a standard normal distribution. Specifically, each rank value was replaced by the Z-score from a standard normal distribution corresponding to its quantile. This transformation ensures that each gene's expression values follow a standard normal distribution across all samples, making the TCGA and CCLE data distributionally comparable and effectively removing technical variations arising from different platforms, labs, or batches.

Step 3: Data Integration and Final Scaling

1. **Data Concatenation:** After the normalization process, the TCGA and CCLE data matrices were directly concatenated by samples (columns) to form a single, unified expression matrix.
2. **Final Scaling of Features and Targets:** Before feeding the data into the XGBoost model, we applied Min-Max Scaling to both the mRNA expression data (features) and the miRNA expression data (targets), scaling their values to a [0, 1] range. This step helps to optimize the training efficiency and performance of the machine learning model.

scRNA-seq

We conducted an extensive analysis of nine single-cell RNA sequencing (scRNA-seq) expression profiles obtained from the TISCH2 [19] database (<http://tisch.comp-genomics.org/gallery/>), which provides comprehensive resources including differential gene expression, cell-type annotations, and associated metadata. These datasets represent diverse human cancer types, offering a broad scope for our investigation into tumor heterogeneity and microenvironment interactions.

The datasets analyzed include:

- **BRCA_EMTAB8107** (33,043 cells from 14 breast cancer patients, 10x Genomics), which comprises a substantial cohort of primary breast cancer samples.
- **CESC_GSE168652** (22,998 cells from 1 cervical cancer patient, 10x Genomics), providing a focused profile of a single primary cervical cancer sample.
- **CRC_GSE166555** (66,050 cells from 12 colorectal cancer patients, 10x Genomics), offering a large-scale dataset from primary colorectal tumors.
- **Glioma_GSE103224** (17,185 cells from 8 glioma patients, Microwell), which includes primary glioma samples sequenced using the Microwell platform.
- **NSCLC_EMTAB6149** (40,218 cells from 5 non-small cell lung cancer patients, 10x Genomics), representing primary non-small cell lung cancer samples.
- **SKCM_GSE72056** (4,645 cells from 19 metastatic melanoma patients, Smart-seq2), which provides insights into metastatic melanoma using Smart-seq2 sequencing technology.
- **OV_EMTAB8107** (24,781 cells from 5 ovarian cancer patients, 10x Genomics), capturing the cellular diversity of primary ovarian cancer samples.

- **PRAD_GSE137829** (8,640 cells from 6 prostate cancer patients, 10x Genomics), focused on primary prostate cancer tumors.
- **CRC_EM TAB8107** (23,176 cells from 7 colorectal cancer patients, 10x Genomics), offering another dataset of primary colorectal cancer samples.

To facilitate downstream analysis, scRNA-seq data were first converted from H5 to Loom format using the Seurat4.3 [20] and loomR packages. This conversion enabled efficient handling of large-scale single-cell datasets, which were subsequently subjected to comprehensive bioinformatic analysis and processing. Standardized protocols appropriate to each sequencing platform were applied to ensure the accuracy and reproducibility of our results across these diverse cancer types. This dataset collection and processing pipeline provided a robust foundation for our exploration of cancer biology at single-cell resolution.

Spatial transcriptomics

We obtained 10X Visium Spatial Transcriptomics (ST) data from 10x Genomics (<https://www.10xgenomics.com/datasets>), selecting datasets for breast cancer, non-small cell lung cancer, ovarian cancer, melanoma, cervical cancer, intestine cancer, glioma, colon cancer, and prostate cancer. For each cancer type, we downloaded both the spatial imaging data and the corresponding filtered feature matrix files. The gene-spot matrices generated from these ST and Visium samples were processed and analyzed using the scanpy package [21]. In the initial steps, we applied basic filtering to the spots based on total counts and the number of expressed genes.

To ensure high data quality, we systematically identified and removed outliers from the spatial transcriptomic datasets. Specifically, in the breast cancer dataset, we excluded five cells with expression counts exceeding 38,000, as they represented extreme outliers. Additionally, genes expressed in fewer than ten cells were filtered out, ensuring that the analysis focused only on the most robustly and consistently expressed genes.

Following these preprocessing steps, we normalized the Visium count data using the standardization method implemented in scanpy and performed a log10 transformation. This transformation compressed the dynamic range of the data, improving both the stability and sensitivity of downstream analyses. To prioritize the most informative features, we identified highly variable genes and selected the top 2,000 feature genes. These genes exhibited significant expression differences across cells, making them critical for subsequent cell type identification and analysis of biological processes.

Finally, the expression data were scaled, and we employed Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction based on gene expression profiles. Clustering was then performed, allowing us to investigate cellular heterogeneity and spatial patterns within the cancer samples, providing key insights into tumor microenvironment organization and cell-type diversity.

Model construction and performance evaluation

A schematic of the model construction and evaluation pipeline is provided in S2 Fig. In this study, we selected Extreme Gradient Boosting (XGBoost) as the primary model due to its robust performance in handling sparse data, reducing overfitting, and offering faster training times compared to other algorithms. XGBoost is an ensemble learning method based on decision trees, where a series of trees are iteratively constructed to improve predictions. Each successive tree is trained to minimize the residual errors from the previous model by optimizing a loss function. [22].

We developed a predictive model leveraging the XGBoost implementation in Python, trained on the integrated bulk RNA-seq datasets containing paired mRNA and miRNA expression profiles from TCGA and CCLE. In this framework, the rank-normalized mRNA expression levels were utilized as input features, while the corresponding miRNA expression levels served as the target output variables for training. To maximize the model's predictive performance, we performed hyperparameter tuning through an exhaustive Grid Search, systematically evaluating combinations of parameters (such as learning rate, tree depth, and number of estimators) to identify the optimal configuration [23].

The dataset used for model development was compiled from the TCGA and CCLE repositories, with analysis focused on genes common to both datasets. To mitigate batch effects, rank-based normalization was applied prior to integration, ensuring consistent gene expression profiles across datasets. For miRNA profiling, a curated database was constructed by selecting only miRNAs shared between TCGA and CCLE, following guidelines to reduce false positive annotations [24] and ensure biological relevance across the cancer types analyzed. Additionally, mRNA features included in the model were intersected with spatial transcriptomics data to enhance biological interpretability.

Preprocessing involved min-max normalization of bulk mRNA and miRNA expression data, followed by partitioning the standardized dataset into training and testing subsets. The XGBoost algorithm was implemented with default parameters and subsequently optimized using a grid search to identify the best-performing hyperparameter configuration. Model performance was evaluated by calculating the Spearman correlation coefficient between predicted miRNA activity and observed miRNA expression, quantifying the model's predictive accuracy.

To benchmark the performance of XGBoost, we also constructed and evaluated several other regression models, including ridge regression, Lasso regression, random forest regression, and a feed-forward neural network. The performance of all models was rigorously assessed using multiple metrics. We calculated the Spearman correlation coefficient between predicted and observed miRNA expression to quantify the model's accuracy in ranking miRNA levels. Additionally, we used Mean Squared Error (MSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2) to compare the predictive accuracy across all tested models. These metrics were averaged across five independent cross-validation folds to ensure the generalizability of our results and to mitigate overfitting. This comprehensive evaluation framework allowed us to demonstrate the superior performance of XGBoost for this prediction task.

External validation with measured spatial miRNA data

To directly validate the predictive accuracy of STmiR on spatial data, we utilized an independent breast cancer spatial transcriptomics dataset from the China National GeneBank (CNGB) database (accession: STDS0000038), which contains concurrent measurements for both mRNA and miRNA. The raw dataset comprised 4,325 spots and 36,601 genes.

To enhance the signal-to-noise ratio of the input data, we first applied the DeepGFT [25] tool, a deep learning method based on graph Fourier transform for denoising spatial transcriptomics data. This procedure yielded a refined expression matrix of 4,325 spots by 12,000 genes.

For this validation exercise, the STmiR model was specifically trained on breast cancer data by integrating TCGA and CCLE datasets. First, highly variable genes from the STDS0000038 spatial data were identified. These genes were then intersected with the common genes between TCGA and CCLE, resulting in a final set of 897 gene features for model input. Common miRNAs between TCGA and CCLE were selected as prediction targets (346 miRNAs). To mitigate batch effects, the expression data from TCGA and CCLE (1,090 total samples) were rank-normalized prior to training the XGBoost model.

The trained model was then applied to the denoised spatial gene expression data to predict miRNA activity. The predicted miRNA expression levels were subsequently compared against the experimentally measured miRNA expression from the STDS0000038 dataset. The concordance for the common miRNAs was evaluated using the Pearson correlation coefficient and the corresponding p-value.

Application of the model to spatial transcriptomics

To extend our bulk RNA-seq-trained model to spatial transcriptomics (ST) data, we first identify the overlapping genes between the bulk RNA-seq and ST datasets. After completing the training of the XGBoost model on bulk RNA-seq data, we input the spatial transcriptomics gene expression data into the pre-trained model to predict miRNA activity across spatially defined spots.

Our workflow integrates miRNA activity predictions with cell-type deconvolution results using a strategy based on dominant cell-type assignment. The steps are as follows:

- **Prediction of Bulk miRNA Activity per Spot:** The STmiR model was applied directly to the bulk mRNA expression profile of each spatial spot. This generated a single “predicted miRNA activity” value for each miRNA at each spot.
 - **Estimation of Cell-Type Composition:** In parallel, we used the cell2location algorithm, with an annotated single-cell RNA-seq dataset as a reference, to deconvolve the cellular composition of each spatial spot, yielding an estimated relative abundance for each cell type within each spot.
 - **Attribution of miRNA Activity to Dominant Cell Types:** To infer the cell-type-specific activity of a miRNA, we linked its predicted activity to the dominant cell type in a given spot.
- 1) First, we identified spots that were predominantly occupied by a single cell type. A spot was assigned a dominant cell type if that type’s relative abundance was the highest among all cell types within that spot (or exceeded a predefined threshold).
 - 2) Next, for this subset of spots, we attributed the high predicted miRNA activities to their corresponding dominant cell type. By identifying miRNAs that consistently exhibited high activity across all spots dominated by a particular cell type (e.g., fibroblasts), we designated them as candidate miRNAs with high activity in that cell type.

We chose this intuitive approach as it directly links the molecular features of a spatial region to its primary cellular identity. By focusing on spots with relatively homogeneous cell populations, we can more confidently assign the miRNA activities—predicted from a mixed-cell signal—to a specific cell type.

Cell2location is capable of accounting for biological variability across different cell types, while also incorporating technical factors inherent in ST data generation, as described by Kleshchevnikov et al [26].

Biofunctional analysis

We perform a comprehensive differential expression analysis across various malignant and non-malignant cell types, including cancer cells, fibroblasts, myofibroblasts, and B cells, for each cancer type. This analysis allows us to identify miRNAs with significantly differential predicted activity in these cell populations.

Following this, we construct a regulatory network of miRNAs and associated diseases using the **miRNet** platform [27] (<https://www.mirnet.ca/>), enabling us to explore the broader biological implications of these miRNAs. To further elucidate the functions of the identified miRNAs, we extract their target genes from the **HMDD** database [28], which provides curated information on miRNA-disease associations.

After identifying miRNA target genes, we conduct pathway enrichment analysis using **Metascape** [29], which helps to uncover the biological pathways in which these target genes are involved. This step is crucial for understanding the functional roles of miRNAs within different cellular contexts.

Finally, we assess the correlation between the predicted miRNA activity and the expression levels of their target genes. A positive correlation suggests that the miRNA may act as a promoter of its target gene’s expression, whereas a negative correlation indicates that the miRNA likely suppresses or dysregulates the expression of its target gene. These insights into miRNA-target interactions provide important clues about the regulatory mechanisms underlying cancer biology.

Statistical analysis

To ensure the accuracy and reliability of our findings, we employed a multifaceted statistical approach tailored to address the unique challenges of spatial transcriptomics data. The performance of the predictive model was rigorously evaluated using Spearman’s rank correlation coefficient, chosen for its robustness to non-normal distributions and outliers inherent

in sparse spatial datasets. This non-parametric measure quantified the consistency between predicted miRNA activity ranks and experimentally observed values, with higher coefficients (e.g., $\rho > 0.5$) indicating strong alignment between model predictions and biological reality. Complementing this, traditional regression metrics—including mean squared error (MSE), mean absolute error (MAE), and the coefficient of determination (R^2)—were calculated to assess prediction accuracy. These metrics were averaged across five independent cross-validation folds to mitigate overfitting and ensure generalizability, with lower MSE/MAE and higher R^2 values collectively guiding the selection of the optimal XGBoost configuration.

To dissect miRNA-target interactions, we performed Pearson correlation analysis between predicted miRNA activity levels and the expression of their experimentally validated target genes (curated from HMDD and miRNet databases). Interactions with absolute correlation coefficients exceeding 0.3 and adjusted P-values below 0.05 (Bonferroni-corrected) were deemed biologically significant. This dual threshold strategy balanced sensitivity and specificity, filtering out spurious associations while retaining mechanistically plausible links. For differential activity analysis, Wilcoxon rank-sum tests compared miRNA activity distributions between malignant and non-malignant cell populations, with Benjamini-Hochberg false discovery rate (FDR < 0.1) correction applied to account for multiple hypothesis testing.

Pathway enrichment analysis further contextualized the functional roles of differentially active miRNAs. Target genes were analyzed using Metascape, where hypergeometric testing against Gene Ontology Biological Processes and KEGG pathways identified overrepresented terms (adjusted $P < 0.05$). Visualization of regulatory networks was conducted in Cytoscape (v3.9.1), with edges weighted by correlation strength and nodes annotated by functional roles, while Matplotlib (v3.7.1) and Seaborn (v0.12.2) generated diagnostic plots such as violin plots for distribution comparisons and scatter-plots for correlation trends. All statistical workflows, including code and preprocessing pipelines, are publicly accessible on GitHub to ensure full reproducibility, adhering to FAIR principles for open science.

This integrative strategy—combining non-parametric rank comparisons, regression diagnostics, hypothesis testing with multiplicity control, and pathway-centric functional annotation—provided a robust foundation for interpreting spatially resolved miRNA activity and its implications in cancer biology.

Results

STmiR accurately predicts miRNA activity and is validated on spatial data

To construct a robust predictive model, we first evaluated a range of machine learning regression techniques on bulk RNA-seq data from TCGA and CCLE across four major cancer types: breast, lung, ovarian, and prostate. Among the tested models, XGBoost exhibited superior performance, consistently achieving the highest R^2 values (averaging ~ 0.8) and the lowest error rates (Table 1). The high concordance between the predicted miRNA activity and the observed miRNA expression is visualized in Fig 1A. Furthermore, we assessed the model's performance for each individual miRNA by calculating the Spearman correlation coefficient. The resulting distributions, shown in Fig 1B, reveal that the vast majority of miRNAs displayed a strong positive correlation (mean $\rho > 0.5$), confirming the fundamental ability of the XGBoost model to capture the complex regulatory relationships between mRNA and miRNA expression from bulk tissue data.

A critical step in our study was to confirm that this predictive accuracy translates to the unique context of spatial transcriptomics. To this end, we performed an external validation using an independent breast cancer spatial transcriptomics dataset (STDS0000038) that, crucially, contains experimentally measured ground-truth miRNA expression. This allowed for a direct, quantitative comparison between STmiR's predicted activity and actual measured miRNA levels for each spatial spot. Our analysis focused on four common miRNAs: hsa-mir-200c, hsa-mir-503, hsa-let-7b, and hsa-mir-210. As shown in S3 Fig, this direct validation yielded compelling results. We observed a strong and statistically significant positive Pearson correlation between the predicted miRNA activity and the true measured expression for all four miRNAs (R -values ranging from 0.559 to 0.708; $p < 0.001$ for all). Beyond quantitative correlation, we also assessed whether STmiR could recapitulate the spatial distribution of these miRNAs. A qualitative comparison of the UMAP plots revealed a

Table 1. Model comparison on different cancers.

Cancer Type	Model	MSE	MAE	R2
Breast Cancer	Ridge	0.028	0.115	0.709
	Lasso	0.027	0.114	0.718
	Random Forest	0.019	0.096	0.798
	Neural Network	0.022	0.105	0.771
	XGBoost	0.012	0.062	0.821
Lung Cancer	Ridge	0.026	0.116	0.712
	Lasso	0.043	0.158	0.523
	Random Forest	0.021	0.103	0.772
	Neural Network	0.023	0.112	0.742
	XGBoost	0.011	0.080	0.821
Ovarian Cancer	Ridge	0.024	0.106	0.775
	Lasso	0.029	0.130	0.730
	Random Forest	0.021	0.101	0.804
	Neural Network	0.023	0.109	0.785
	XGBoost	0.020	0.096	0.814
Prostate Cancer	Ridge	0.026	0.108	0.742
	Lasso	0.022	0.102	0.784
	Random Forest	0.020	0.096	0.801
	Neural Network	0.022	0.103	0.782
	XGBoost	0.012	0.071	0.833

<https://doi.org/10.1371/journal.pone.0322082.t001>

high degree of similarity between the spatial patterns of predicted activity and true expression. This dual quantitative and qualitative evidence provides strong support for STmiR's ability to accurately and reliably infer spatially resolved miRNA activity, establishing a solid foundation for its application in subsequent biological discovery.

Pan-cancer analysis reveals conserved regulatory hubs and tissue-specific heterogeneity

Having established the model's reliability, we applied STmiR to explore the landscape of miRNA regulation across nine distinct cancer types. This pan-cancer analysis aimed to identify miRNAs with conserved high activity, suggesting fundamental roles in tumorigenesis. By ranking the average predicted miRNA activity in each cancer type, we identified six miRNAs—hsa-miR-30a, hsa-miR-30e, hsa-miR-181a, hsa-let-7a, hsa-miR-92a, and the well-known onco-miRNA **hsa-miR-21**—that consistently ranked among the top 40 across all nine malignancies (Fig 2A, 2B). This remarkable conservation highlights their potential as core regulators in cancer biology. Indeed, these miRNAs represent key players in tumorigenesis; for instance, hsa-let-7a is a well-documented tumor suppressor that limits cell proliferation [30], whereas hsa-miR-21 is a potent oncogene that promotes tumor cell survival and growth [31]. Other identified miRNAs, including the miR-30 family, miR-181a, and miR-92a, are also widely implicated in modulating cell proliferation, differentiation, and migration within the tumor microenvironment [32–34]. The prominence of onco-miRNAs like hsa-miR-21 underscores the importance of dissecting these complex networks at the tissue level. To illustrate STmiR's capability in this regard, we performed an in-depth case study of breast cancer, focusing on its most active miRNA regulatory networks.

Our pan-cancer analysis also highlighted the diversity of miRNA regulation. While some cancers, like breast and colon cancer, showed consistent miRNA activity rankings across their constituent cell types, others, such as lung and ovarian cancer, exhibited significant heterogeneity (Fig 2C). This suggests that in certain tumors, miRNA regulation is highly context-dependent, varying significantly among different cellular communities within the tumor microenvironment.

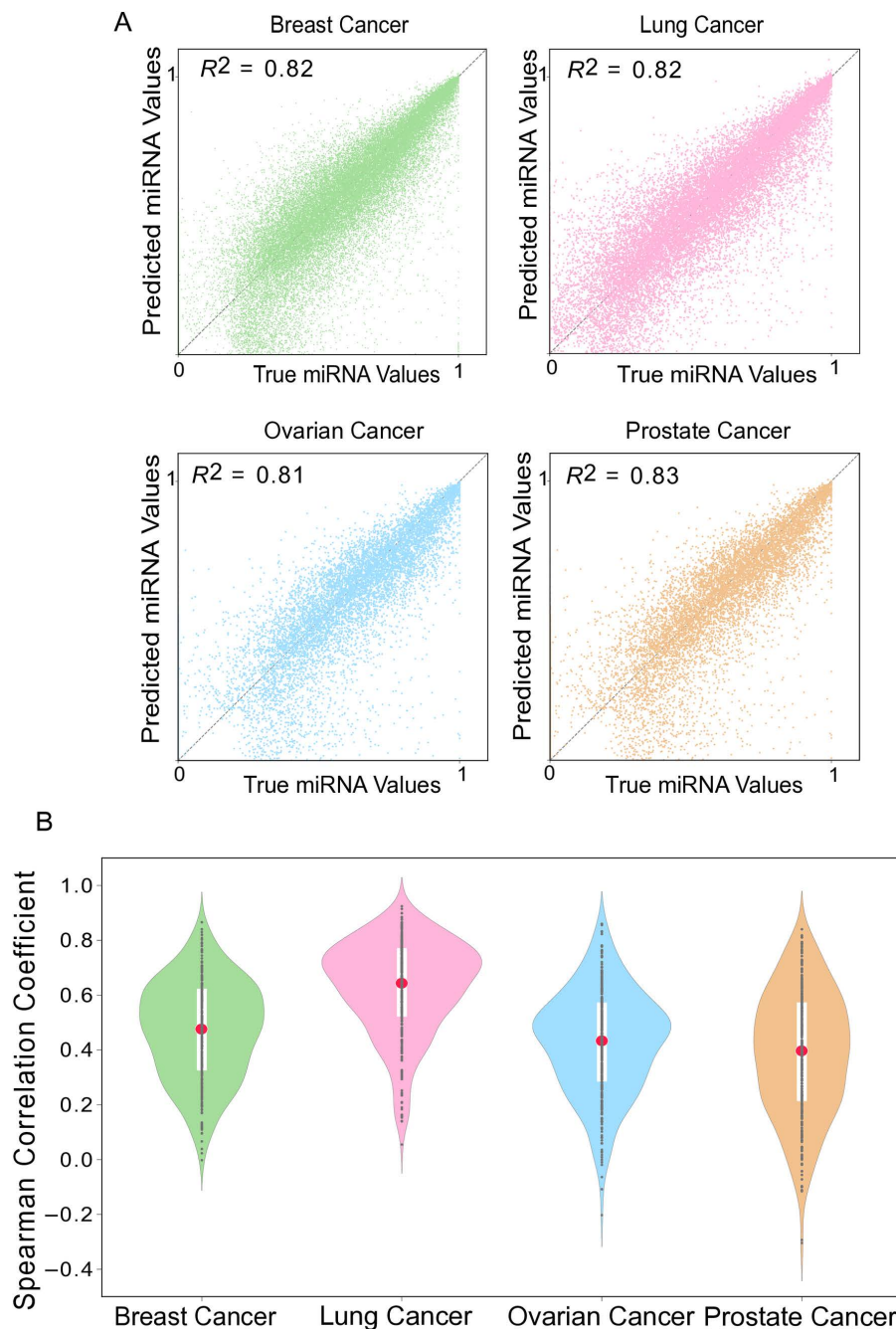


Fig 1. Performance of XGBoost in miRNA Activity Prediction Using Bulk RNA-seq Data. (A) Scatterplot of predicted versus observed miRNA expression levels. The plot illustrates the correlation between XGBoost-predicted miRNA activity (y-axis) and experimentally measured miRNA expression (x-axis) across four cancer types: breast cancer (BRCA), lung adenocarcinoma (LUAD), ovarian serous carcinoma (OV), and prostate adenocarcinoma (PRAD). Each point represents a miRNA-sample pair, with the solid line indicating the linear regression fit (slope = 0.89, 95% CI: 0.85–0.93). Shaded regions denote 95% confidence intervals. High concordance (Spearman's $\rho > 0.8$) underscores the model's accuracy in capturing miRNA expression trends. (B) Distribution of Spearman correlation coefficients across cancer types. Violin plots summarize the pairwise Spearman correlations between predicted and true miRNA expression levels for each cancer type. Individual points represent specific miRNAs, while the central red line marks the median correlation (BRCA: $\rho = 0.82$; LUAD: $\rho = 0.79$; OV: $\rho = 0.76$; PRAD: $\rho = 0.84$). The width of each violin reflects the density distribution, highlighting tighter clustering in BRCA and PRAD compared to LUAD. Coefficients closer to 1 indicate stronger rank consistency, demonstrating XGBoost's robustness in modeling miRNA activity across heterogeneous cancer datasets.

<https://doi.org/10.1371/journal.pone.0322082.g001>

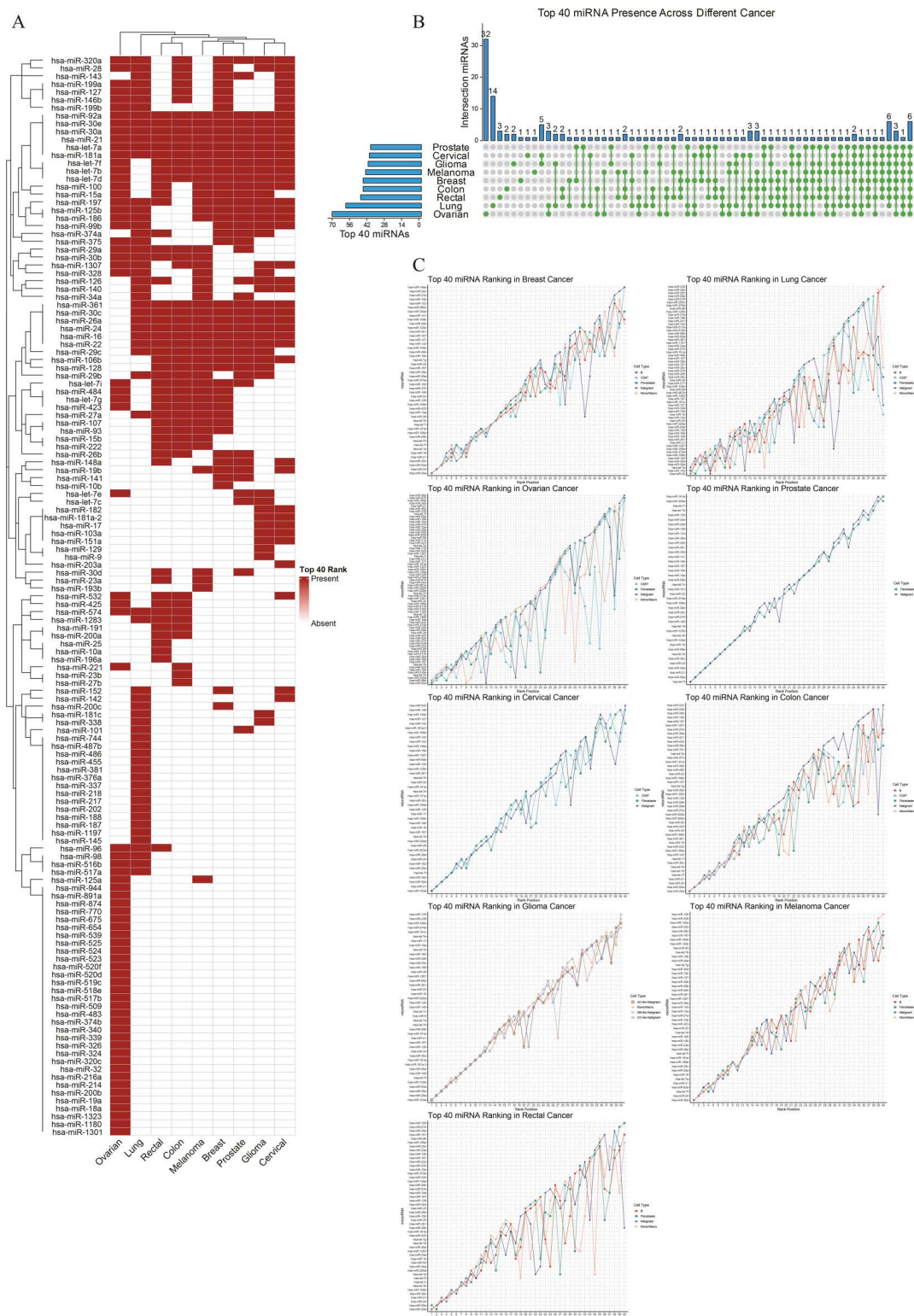


Fig 2. Pan-Cancer Spatial Comparison of miRNA Commonality and Heterogeneity. (A) Venn diagram illustrating six pan-cancer conserved miRNAs (e.g., hsa-miR-21, hsa-let-7a) consistently ranked in the top 40 across nine cancer types. (B) Heatmap displaying miRNA rankings across cancers. Red

highlights conserved miRNAs (e.g., miR-30 family), while color gradients reflect cancer-specific variations. (C) Line plot showing average miRNA expression levels across cell types (fibroblasts, B cells, malignant cells) in nine cancers. Solid lines indicate conserved rankings in breast and cervical cancers; dashed lines highlight heterogeneity in lung and ovarian cancers.

<https://doi.org/10.1371/journal.pone.0322082.g002>

Pan-cancer miRNA conservation in key cell subtypes

To further dissect these patterns, we investigated conserved miRNA activity within specific cell subtypes across the different cancers. This analysis revealed consistent signatures in key cell populations of the tumor microenvironment (Fig 3).

- **In fibroblasts**, members of the miR-30 family, particularly hsa-miR-30a and hsa-miR-30e, were consistently active. This finding is significant as these miRNAs are known modulators of fibroblast activation and fibrosis, critical processes in shaping the supportive tumor stroma [35,36].
- **In B cells**, we observed a balance of activity between the tumor-suppressive hsa-let-7a and the oncogenic hsa-miR-21. This highlights the complex immunomodulatory role of B cells, where miRNAs simultaneously regulate both tumor-suppressive and pro-tumorigenic pathways within the immune compartment [37,38].
- **Within malignant cells** themselves, hsa-let-7a and hsa-miR-30e were the most consistently active miRNAs. The sustained activity of let-7a likely reflects a persistent, albeit often overcome, cellular mechanism to restrain oncogenes like RAS, while miR-30e has been implicated in regulating cell migration and invasion [39,40].
- **In monocytes/macrophages**, hsa-miR-30e and hsa-miR-92a showed conserved activity, implicating their roles in regulating macrophage polarization and the tumor's immune landscape, potentially influencing the switch between anti-tumor M1 and pro-tumor M2 phenotypes [41–45].

Cell-type-specific miRNA signatures highlight context-dependent regulation

Beyond the conserved patterns, our analysis also identified unique miRNA activity signatures specific to certain cell types within individual cancers, underscoring the context-dependent nature of miRNA regulation (Fig 4). For instance, in lung cancer, hsa-miR-200c activity was a specific marker for malignant cells, likely related to its well-established role in suppressing the epithelial-mesenchymal transition (EMT), a key process in metastasis [46]. In contrast, in prostate cancer, hsa-miR-143 and hsa-miR-145 were highly active specifically in fibroblasts, reflecting the desmoplastic stromal reaction characteristic of that disease. These results showcase STmiR's capacity to uncover nuanced, tissue-specific regulatory programs that would be obscured in bulk analysis.

Case study: Dissecting miRNA regulatory networks in breast cancer

To demonstrate STmiR's utility in generating specific, testable hypotheses, we performed an in-depth case study in breast cancer. We analyzed the predicted activity of the most prominent miRNAs within four major cell types: malignant cells, B cells, fibroblasts, and myofibroblasts.

Our analysis of the malignant cell population revealed a complex regulatory network (Fig 5A, left panel) where members of the oncogenic miR-200 family (miR-200a/b/c) and miR-141 showed high predicted activity. Notably, our network recapitulated several experimentally validated interactions from existing literature. For instance, the predicted negative regulation of the EMT-associated gene CDH1 by the miR-200 family is a cornerstone of cancer metastasis research [47]. Similarly, our identified link between miR-205 and the Androgen Receptor (AR) aligns with studies suggesting their expression levels could serve as predictive markers in breast cancer [48]. Furthermore, the predicted targeting of YAP1 by miR-200a is consistent with reports showing their inverse correlation in clinical specimens [49]. To understand the

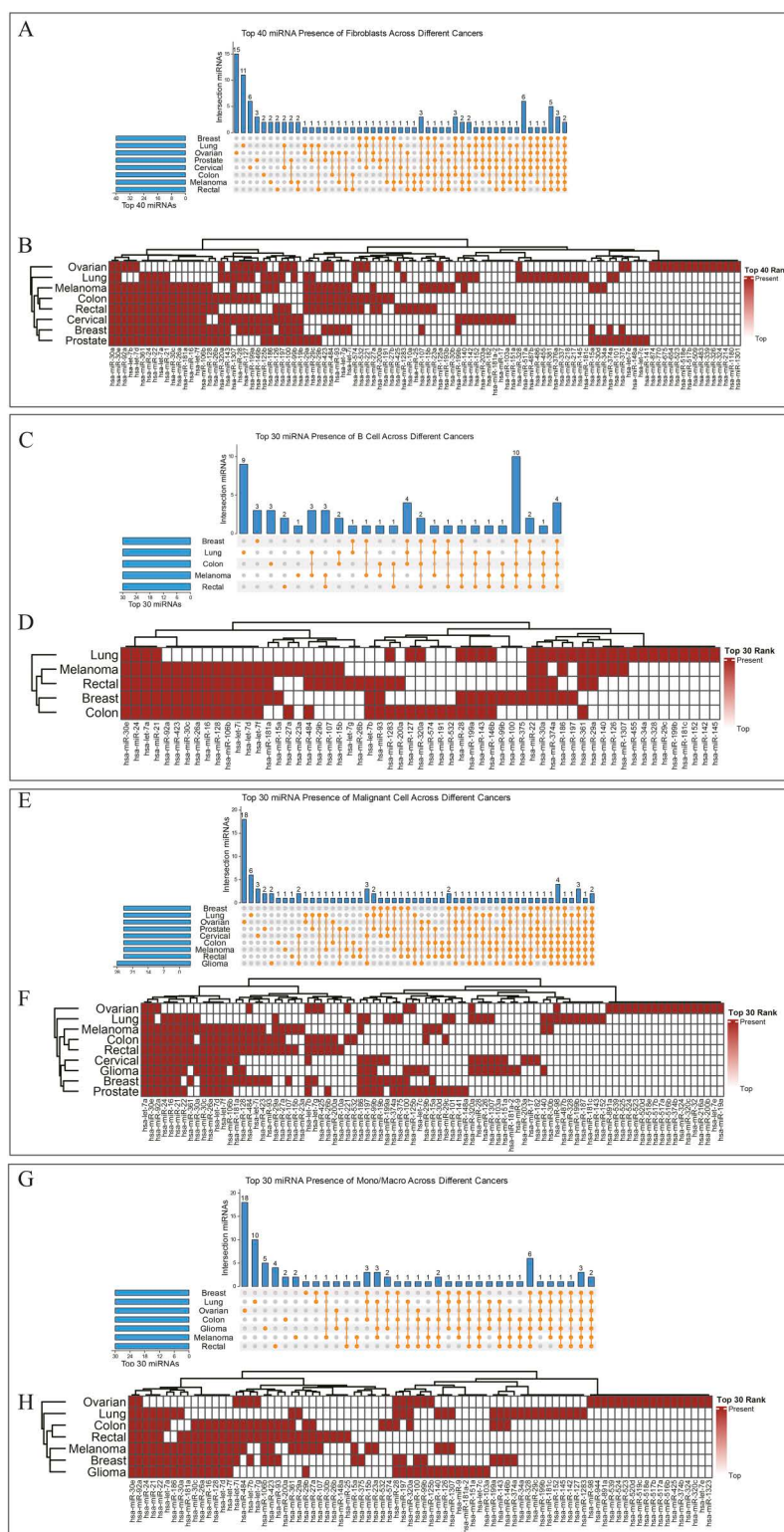


Fig 3. Pan-Cancer miRNA Conservation in Cell Subtypes. (A, B) Fibroblasts: Violin plots of hsa-miR-30a and hsa-miR-30e expression, conserved across cancers. These miRNAs regulate fibroblast activation and fibrosis. (C, D) B cells: Boxplots of hsa-let-7a (tumor suppressor) and hsa-miR-21

(oncogenic), showing balanced roles in immune modulation. (E, F) Malignant cells: Scatterplots of hsa-let-7a and hsa-miR-30e expression, linked to oncogene repression (e.g., RAS) and metastasis inhibition. (G, H) Monocytes/Macrophages: Heatmaps of miR-30e and miR-92a expression, indicating roles in macrophage polarization and immune response regulation.

<https://doi.org/10.1371/journal.pone.0322082.g003>

functional implications of these networks, we performed pathway enrichment analysis on the predicted target genes (Fig 5B). For malignant cells, the targets were significantly enriched in critical cancer pathways, including “MicroRNAs in cancer,” “PI3K-Akt signaling pathway,” and “FoxO signaling pathway,” confirming that our predictions capture biologically coherent regulatory modules.

Similarly, we identified distinct miRNA networks and associated pathways for B cells, fibroblasts, and myofibroblasts (Fig 5), highlighting how different cell types within the TME are governed by unique regulatory logic. For example, the predicted targeting of VEGFA by miR-126 in B-cells is consistent with its known role in angiogenesis [50]. This case study demonstrates how STmiR can move beyond single-molecule analysis to deconstruct the complex, cell-type-specific miRNA-mediated regulatory networks that collectively drive tumor biology, thereby generating rich, spatially-resolved hypotheses for future experimental validation.

Discussion

In this study, we developed and validated STmiR, a novel machine learning framework based on XGBoost, to infer miRNA activity from spatial transcriptomics data. By successfully bridging the gap left by technologies not optimized for direct miRNA capture, STmiR provides a powerful approach to explore the spatial dimension of post-transcriptional regulation within the native tissue context. Our work not only demonstrates the technical feasibility and accuracy of this prediction—validated against both bulk data and, crucially, ground-truth spatially resolved miRNA measurements—but also applies it to uncover biologically meaningful patterns across a pan-cancer atlas.

Our pan-cancer analysis revealed a core set of miRNAs, including the well-documented onco-miRNA hsa-miR-21 and the tumor-suppressive let-7 family, that exhibit conserved high activity across nine distinct malignancies. The consistent prominence of these miRNAs suggests their involvement in fundamental oncogenic processes that are shared across different tumor types. The dynamic balance between oncogenic miRNAs like miR-21, which drives proliferation and inhibits apoptosis [31], and tumor-suppressive miRNAs like the let-7 family, which restrain key oncogenes like RAS [39], likely represents a central regulatory battlefield in tumorigenesis. Our ability to spatially map the activity of these opposing forces provides a new lens through which to view tumor progression. Furthermore, the observed heterogeneity in miRNA activity landscapes, particularly in lung and ovarian cancers, underscores the profound impact of the tumor microenvironment (TME) [51,52]. This suggests that while core oncogenic pathways may be conserved, their regulation by miRNAs can be highly plastic and dependent on the unique cellular composition and crosstalk within a specific TME, a notion supported by recent studies on spatial intratumor heterogeneity.

A key strength of STmiR is its ability to deconvolve miRNA activity within specific cell types, moving beyond the bulk-level view. Our analyses consistently highlighted distinct miRNA regulatory programs operating in malignant cells versus various stromal and immune components of the TME. For example, the high activity of the miR-30 family in fibroblasts across multiple cancers points to a conserved role in regulating stromal responses and extracellular matrix remodeling, potentially driving the desmoplastic reaction seen in many solid tumors [35]. Similarly, the specific miRNA signatures identified in B cells and macrophages highlight the intricate role of miRNA-mediated regulation in shaping tumor immunology [45]. The breast cancer case study further solidified this utility. By dissecting the complex regulatory networks within four distinct cell populations, we moved from a single-molecule view to a systems-level understanding. We demonstrated how STmiR can generate highly specific, data-driven hypotheses—such as the interplay between the miR-200 family in malignant cells and pathways like PI3K-Akt signaling—that are primed for subsequent experimental validation.

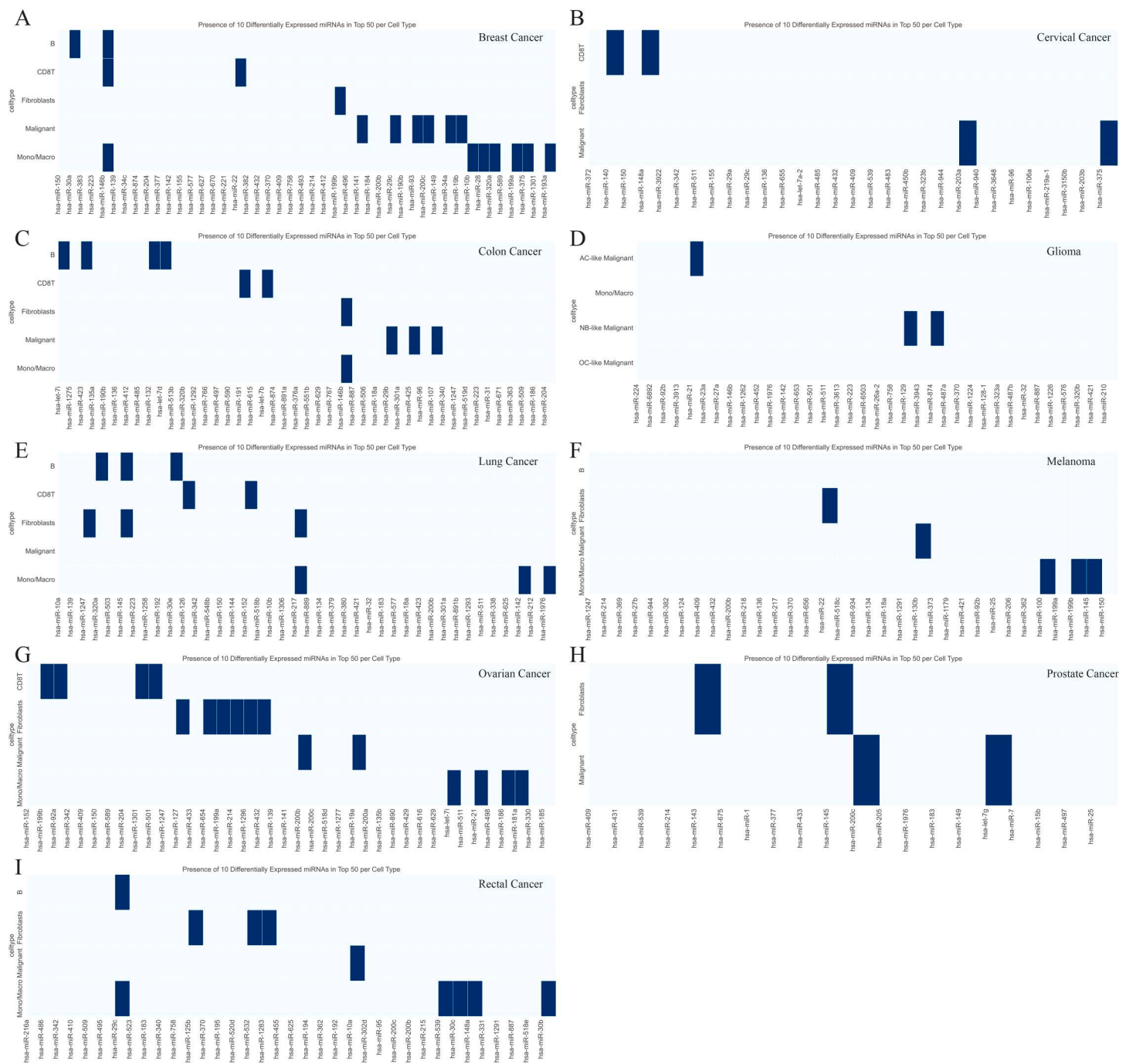


Fig 4. Cell-Type-Specific miRNA Signatures Across Cancers. (A) Breast cancer: hsa-miR-30a (B cells), hsa-miR-146b (B cells, CD8+ T cells, monocytes/macrophages), hsa-miR-22 (CD8+ T cells), and hsa-miR-199b (fibroblasts) delineate immune-stromal crosstalk. (B) Cervical cancer: CD8+ T cells express hsa-miR-140 and hsa-miR-148a (immune surveillance), while hsa-miR-203a (differentiation) and hsa-miR-375 (proliferation) dominate malignant cells. (C) Colon cancer: B cells exhibit hsa-let-7i, miR-423, and miR-132 (tumor suppression), with hsa-miR-146b shared across fibroblasts and monocytes/macrophages (inflammation/remodeling). (D) Glioma: Subtype-specific signatures include hsa-miR-21 (AC-like) and hsa-miR-129 (NB-like), indicating divergent oncogenic pathways. (E) Lung cancer: hsa-miR-320a/hsa-miR-145 (B cells), hsa-miR-217 (fibroblast activation), and hsa-miR-200c (malignant cells, EMT) highlight microenvironmental interplay. (F) Melanoma: hsa-miR-22 (fibroblasts) and hsa-miR-130b (malignant cells, metastasis) underscore stromal-tumor dynamics. (G) Ovarian cancer: Fibroblast-enriched miRNAs (hsa-miR-127, hsa-miR-654, hsa-miR-199a) and hsa-miR-21 (dual stromal/immune roles) reflect TME complexity. (H) Prostate cancer: hsa-miR-143/hsa-miR-145 mark fibroblasts, linked to desmoplastic stroma. (I) Rectal cancer: hsa-miR-125b/hsa-miR-532 (fibroblasts), hsa-miR-10a (malignant cells), and hsa-miR-539 (monocytes/macrophages) implicate inflammation-driven TME remodeling.

<https://doi.org/10.1371/journal.pone.0322082.g004>

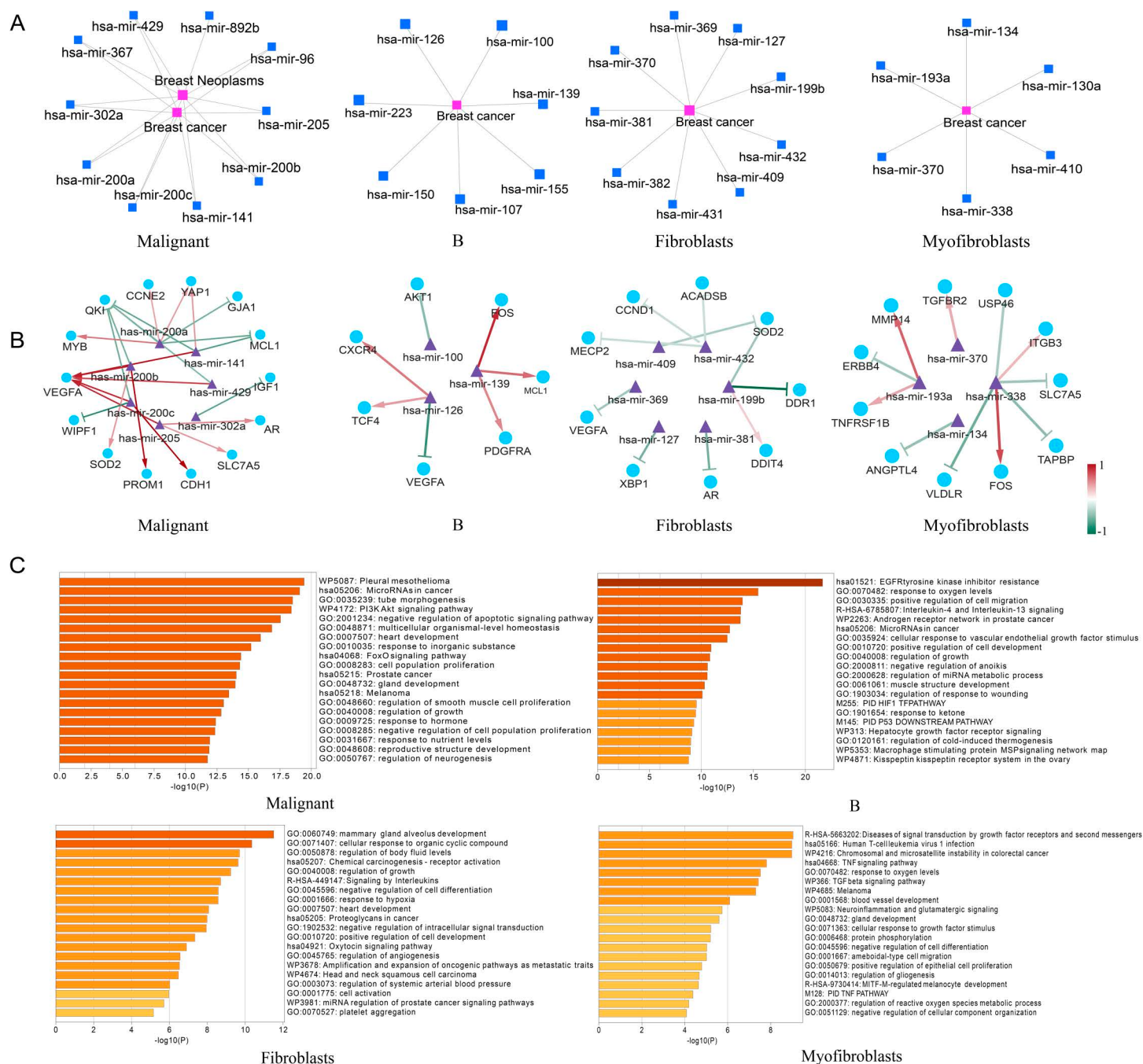


Fig 5. Breast cancer top miRNAs-target gene and miRNAs-disease regulatory network. (A) Network diagram linking top differentially expressed miRNAs (malignant cells, B cells, fibroblasts, myofibroblasts) to breast cancer pathways. (B) miRNA-target interaction map: Red edges (positive Pearson correlation, e.g., miR-205-AR), green edges (negative correlation, e.g., miR-200b-LOX). Line darkness reflects correlation strength. (C) Pathway enrichment bar charts: Malignant cells enrich "PI3K-Akt signaling," B cells associate with "EGFR resistance," fibroblasts link to "mammary gland development," and myofibroblasts to "TNF signaling."

<https://doi.org/10.1371/journal.pone.0322082.g005>

Methodologically, STmiR offers a robust and computationally efficient alternative to other advanced techniques. While deep learning models may offer advantages in capturing highly complex, non-linear patterns or integrating multi-modal data like histology images [53, 54], our XGBoost-based approach provides a balance of high accuracy, interpretability, and robustness to the inherent sparsity and noise of ST data. The successful validation against measured spatial miRNA data confirms that a model trained on large-scale bulk datasets can indeed learn generalizable regulatory principles that are transferable to the spot-level microenvironment. This aligns with a growing trend of using computational methods to deconvolve complex tissue architecture and molecular interactions from spatial data [55,56] and positions STmiR as a valuable and accessible tool for researchers to conduct large-scale exploratory analyses of miRNA function. The potential clinical relevance of identifying such spatially-defined regulatory hubs is significant [57].

Limitations of the study

Despite the promising results, we acknowledge several limitations inherent to our study. A primary limitation is the domain shift between the bulk RNA-seq data used for training and the spot-level data used for prediction. We must critically consider the potential pitfalls of this approach. Bulk tissue samples represent an averaged expression profile from millions of cells, whereas an ST spot, while not single-cell, captures a unique microenvironment with specific cellular compositions and interactions. This difference in cellular heterogeneity and technical noise profiles could potentially introduce biases into the predictions.

We have sought to mitigate this issue through several strategies. The XGBoost algorithm is inherently robust to noise and data sparsity. More importantly, our key biological conclusions are not derived from single-spot predictions but from statistical trends and correlations across thousands of spots. Our “predict-then-associate” method for inferring cell-type-specific activity is a conservative approach that avoids applying the bulk-trained model directly to deconvoluted profiles. However, the gold standard for validation remains direct experimental measurement. While our computational validation against the STDS0000038 dataset provides strong support, future work should aim to validate specific findings—such as the cell-type-specific activity of hsa-miR-21 in breast cancer—using techniques like multiplexed *in situ* hybridization (e.g., RNAScope) or immunofluorescence co-localization for target proteins, as has been demonstrated as feasible in recent spatial multi-omics studies [14].

Furthermore, our cell-type attribution method, which links miRNA activity to the dominant cell type in a spot, is a conservative simplification. In spots with highly mixed populations, this approach may overlook signals from sub-dominant but biologically important cell types. Future methods could aim to deconvolve activity in a more proportional manner.

Conclusion

In conclusion, STmiR represents a significant step forward in our ability to understand the spatial regulation of miRNAs in cancer. By accurately inferring miRNA activity in a spatially resolved manner, this approach provides a powerful resource for exploring the intricate regulatory networks that govern gene expression in complex tissues. The work presented here not only offers novel biological insights into pan-cancer miRNA regulation but also lays a robust methodological foundation for future research. By generating rich, spatially-contextualized, and testable hypotheses, STmiR opens new avenues for investigating the roles of miRNAs in tumor progression, microenvironment remodeling, and therapeutic response.

Resources and methods

Supporting information

S1. Supporting information. (DOCX)

S1 Fig. The overall workflow of the STmiR framework. This Fig illustrates the three main stages of the STmiR methodology. The process begins with model construction, where paired miRNA-mRNA bulk RNA-seq data from TCGA and CCLE are integrated to train an XGBoost model that uses mRNA expression to predict miRNA activity. Next, in the spatial transcriptomics application stage, this pre-trained model is used to infer miRNA activity from the mRNA expression profiles of spatial spots, while cell-type deconvolution is concurrently performed using cell2location. In the final stage, the framework enables downstream functional analysis by identifying cell-type-specific miRNA activity to construct miRNA-target gene and miRNA-disease regulatory networks for biological interpretation.

(PDF)

S2 Fig. Schematic of the model construction and evaluation pipeline. This diagram outlines the procedure for building and validating the predictive model. The workflow starts with paired miRNA-mRNA profiles derived from bulk RNA-seq samples. The mRNA expression data serves as the input for model construction, which generates a predicted miRNA profile. This predicted profile is then compared against the actual miRNA profile from the same samples, which serves as the ground truth. The predictive performance of the model is quantitatively assessed by calculating the Spearman correlation between the predicted and actual miRNA profiles.

(PDF)

S3 Fig. Direct Validation of STmiR's Predictive Accuracy on Spatial miRNA Data. The Fig demonstrates the model's performance by comparing predicted miRNA expression to true measured expression from an independent spatial transcriptomics dataset. (A-D) Scatterplots show the correlation between predicted (y-axis) and true (x-axis) scaled expression for four miRNAs: (A) hsa-mir-200c ($R=0.708$, $p=0.000$), (B) hsa-mir-503 ($R=0.608$, $p=0.000$), (C) hsa-let-7b ($R=0.642$, $p=0.000$), and (D) hsa-mir-210 ($R=0.559$, $p=0.000$). The red line indicates the linear regression fit. (E-H) UMAP plots illustrate the spatial distribution of the true measured expression for (E) hsa-mir-200c, (F) hsa-mir-503, (G) hsa-let-7b, and (H) hsa-mir-210. (I-L) UMAP plots illustrate the spatial distribution of the STmiR-predicted expression for (I) hsa-mir-200c, (J) hsa-mir-503, (K) hsa-let-7b, and (L) hsa-mir-210. The similarity in spatial patterns between the true (E-H) and predicted (I-L) plots confirms the model's ability to capture spatial localization of miRNA activity.

(PDF)

S4. Raw data.

(ZIP)

S5. Supporting information.

(DOCX)

Author contributions

Conceptualization: Jiaqi Yuan.

Data curation: Jiaqi Yuan.

Formal analysis: Peng Xu.

Funding acquisition: Peng Xu, Wenbin Liu.

Investigation: Zheng Ye.

Methodology: Jiaqi Yuan.

Project administration: Peng Xu.

Software: Jiaqi Yuan, Wenbin Liu.

Supervision: Peng Xu, Zheng Ye.

Validation: Wenbin Liu.

Visualization: Zheng Ye.

Writing – original draft: Jiaqi Yuan.

Writing – review & editing: Wenbin Liu, Zheng Ye.

References

1. Sriyothi L, Ponne S, Prathama T, Ashok C, Baluchamy S. Roles of non-coding RNAs in transcriptional regulation. *Transcriptional and Post-transcriptional regulation*. 2018;55.
2. Santosh B, Varshney A, Yadava PK. Non-coding RNAs: biological functions and applications. *Cell Biochem Funct*. 2015;33(1):14–22. <https://doi.org/10.1002/cbf.3079> PMID: [25475931](https://pubmed.ncbi.nlm.nih.gov/25475931/)
3. Hill M, Tran N. miRNA interplay: mechanisms and consequences in cancer. *Disease models & mechanisms*. 2021;14(4):dmm047662. <https://doi.org/10.1242/dmm047662>
4. Lin Z, Tang X, Wan J, Zhang X, Liu C, Liu T. Functions and mechanisms of circular RNAs in regulating stem cell differentiation. *RNA biology*. 2021;18(12):2136–49.
5. Li S, Lei Z, Sun T. The role of microRNAs in neurodegenerative diseases: a review. *Cell Biology and Toxicology*. 2023;39(1):53–83.
6. Perdoncin M, Konrad A, Wyner JR, Lohana S, Pillai SS, Pereira DG, et al. A Review of miRNAs as Biomarkers and Effect of Dietary Modulation in Obesity Associated Cognitive Decline and Neurodegenerative Disorders. *Front Mol Neurosci*. 2021;14:756499. <https://doi.org/10.3389/fnmol.2021.756499> PMID: [34690698](https://pubmed.ncbi.nlm.nih.gov/34690698/)
7. Ye Z, Sun B, Mi X, Xiao Z. Gene co-expression network for analysis of plasma exosomal miRNAs in the elderly as markers of aging and cognitive decline. *PeerJ*. 2020;8:e8318. <https://doi.org/10.7717/peerj.8318> PMID: [31934508](https://pubmed.ncbi.nlm.nih.gov/31934508/)
8. Kiessling P, Kuppe C. Spatial multi-omics: novel tools to study the complexity of cardiovascular diseases. *Genome Med*. 2024;16(1):14. <https://doi.org/10.1186/s13073-024-01282-y> PMID: [38238823](https://pubmed.ncbi.nlm.nih.gov/38238823/)
9. Piñeiro AJ, Houser AE, Ji AL. Research Techniques Made Simple: Spatial Transcriptomics. *J Invest Dermatol*. 2022;142(4):993–1001.e1. <https://doi.org/10.1016/j.jid.2021.12.014> PMID: [35331388](https://pubmed.ncbi.nlm.nih.gov/35331388/)
10. Sass S, Pitea A, Unger K, Hess J, Mueller NS, Theis FJ. MicroRNA-Target Network Inference and Local Network Enrichment Analysis Identify Two microRNA Clusters with Distinct Functions in Head and Neck Squamous Cell Carcinoma. *Int J Mol Sci*. 2015;16(12):30204–22. <https://doi.org/10.3390/ijms161226230> PMID: [26694379](https://pubmed.ncbi.nlm.nih.gov/26694379/)
11. Le TD, Zhang J, Liu L, Liu H, Li J. miRLAB: An R Based Dry Lab for Exploring miRNA-mRNA Regulatory Relationships. *PLoS One*. 2015;10(12):e0145386. <https://doi.org/10.1371/journal.pone.0145386> PMID: [26716983](https://pubmed.ncbi.nlm.nih.gov/26716983/)
12. Wen M, Cong P, Zhang Z, Lu H, Li T. DeepMirTar: a deep-learning approach for predicting human miRNA targets. *Bioinformatics*. 2018;34(22):3781–7. <https://doi.org/10.1093/bioinformatics/bty424> PMID: [29868708](https://pubmed.ncbi.nlm.nih.gov/29868708/)
13. Olgun G, Gopalan V, Hannehalli S. miRSCAPE - inferring miRNA expression from scRNA-seq data. *iScience*. 2022;25(9):104962. <https://doi.org/10.1016/j.isci.2022.104962> PMID: [36060076](https://pubmed.ncbi.nlm.nih.gov/36060076/)
14. Luo Z, Mei J, Wang X, Wang R, He Z, Geffen Y, et al. Voluntary exercise sensitizes cancer immunotherapy via the collagen inhibition-orchestrated inflammatory tumor immune microenvironment. *Cell Rep*. 2024;43(9):114697. <https://doi.org/10.1016/j.celrep.2024.114697> PMID: [39217611](https://pubmed.ncbi.nlm.nih.gov/39217611/)
15. Ji X, Tian X, Feng S, Zhang L, Wang J, Guo R, et al. Intermittent F-actin Perturbations by Magnetic Fields Inhibit Breast Cancer Metastasis. *Research (Wash D C)*. 2023;6:0080. <https://doi.org/10.34133/research.0080> PMID: [36939445](https://pubmed.ncbi.nlm.nih.gov/36939445/)
16. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. Addendum: The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*. 2019;565(7738):E5–6. <https://doi.org/10.1038/s41586-018-0722-x> PMID: [30559381](https://pubmed.ncbi.nlm.nih.gov/30559381/)
17. Hutter C, Zenklusen JC. The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. *Cell*. 2018;173(2):283–5. <https://doi.org/10.1016/j.cell.2018.03.042> PMID: [29625045](https://pubmed.ncbi.nlm.nih.gov/29625045/)
18. Ghandi M, Huang FW, Jané-Valbuena J, Kryukov GV, Lo CC, McDonald ER 3rd, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature*. 2019;569(7757):503–8. <https://doi.org/10.1038/s41586-019-1186-3> PMID: [31068700](https://pubmed.ncbi.nlm.nih.gov/31068700/)
19. Han Y, Wang Y, Dong X, Sun D, Liu Z, Yue J, et al. TISCH2: expanded datasets and new tools for single-cell transcriptome analyses of the tumor microenvironment. *Nucleic Acids Res*. 2023;51(D1):D1425–31. <https://doi.org/10.1093/nar/gkac959> PMID: [36321662](https://pubmed.ncbi.nlm.nih.gov/36321662/)
20. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol*. 2018;36(5):411–20. <https://doi.org/10.1038/nbt.4096> PMID: [29608179](https://pubmed.ncbi.nlm.nih.gov/29608179/)
21. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol*. 2018;19(1):15. <https://doi.org/10.1186/s13059-017-1382-0> PMID: [29409532](https://pubmed.ncbi.nlm.nih.gov/29409532/)
22. Ester M, Kriegl H, Xu X. XGBoost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (vol, pg 785, 2016)*. GEOGRAPHICAL ANALYSIS. 2022.

23. Putatunda S, Rama K, editors. A comparative analysis of hyperopt as against other approaches for hyper-parameter optimization of XGBoost. Proceedings of the 2018 international conference on signal processing and machine learning; 2018.
24. Fromm B, Høyse E, Domanska D, Zhong X, Aparicio-Puerta E, Ovchinnikov V, et al. MirGeneDB 2.1: toward a complete sampling of all major animal phyla. *Nucleic Acids Res.* 2022;50(D1):D204–10. <https://doi.org/10.1093/nar/gkab1101> PMID: 34850127
25. Sun S, Liu J, Li G, Liu B. DeepGFT: identifying spatial domains in spatial transcriptomics of complex and 3D tissue using deep learning and graph Fourier transform. *Genome Biol.* 2025;26(1):153. <https://doi.org/10.1186/s13059-025-03631-5> PMID: 40462157
26. Kleshchevnikov V, Shmatko A, Dann E, Aivazidis A, King HW, Li T, et al. Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat Biotechnol.* 2022;40(5):661–71. <https://doi.org/10.1038/s41587-021-01139-4> PMID: 35027729
27. Chang L, Zhou G, Soufan O, Xia J. miRNet 2.0: network-based visual analytics for miRNA functional analysis and systems biology. *Nucleic Acids Res.* 2020;48(W1):W244–51. <https://doi.org/10.1093/nar/gkaa467> PMID: 32484539
28. Huang Z, Shi J, Gao Y, Cui C, Zhang S, Li J, et al. HMDD v3.0: a database for experimentally supported human microRNA-disease associations. *Nucleic Acids Res.* 2019;47(D1):D1013–7. <https://doi.org/10.1093/nar/gky1010> PMID: 30364956
29. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun.* 2019;10(1):1523. <https://doi.org/10.1038/s41467-019-09234-6> PMID: 30944313
30. Kolenda T, Przybyła W, Teresiak A, Mackiewicz A, Lamperska KM. The mystery of let-7d—a small RNA with great power. *Contemporary Oncology/ Współczesna Onkologia.* 2014;18(5):293–301.
31. Singh VK, Rajak N, Singh Y, Singh AK, Giri R, Garg N. Role of MicroRNA-21 in Prostate Cancer Progression and Metastasis: Molecular Mechanisms to Therapeutic Targets. *Ann Surg Oncol.* 2024;31(7):4795–808. <https://doi.org/10.1245/s10434-024-15453-z> PMID: 38758485
32. Mao L, Liu S, Hu L, Jia L, Wang H, Guo M, et al. miR-30 Family: A Promising Regulator in Development and Disease. *Biomed Res Int.* 2018;2018:9623412. <https://doi.org/10.1155/2018/9623412> PMID: 30003109
33. Li J, Shen J, Zhao Y, Du F, Li M, Wu X, et al. Role of miR-181a-5p in cancer (Review). *Int J Oncol.* 2023;63(4):108. <https://doi.org/10.3892/ijo.2023.5556> PMID: 37539738
34. Jiang M, Li X, Quan X, Li X, Zhou B. MiR-92a Family: A Novel Diagnostic Biomarker and Potential Therapeutic Target in Human Cancers. *Front Mol Biosci.* 2019;6:98. <https://doi.org/10.3389/fmolb.2019.00098> PMID: 31632984
35. Eichelmann A, Matuszcak C, Hummel R, Haier J. Role of miRNAs in cell signaling of cancer associated fibroblasts. *The International Journal of Biochemistry & Cell Biology.* 2018;101:94–102.
36. Wu G, Xie B, Lu C, Chen C, Zhou J, Deng Z. microRNA-30a attenuates TGF- β 1-induced activation of pulmonary fibroblast cell by targeting FAP- α . *J Cell Mol Med.* 2020;24(6):3745–50. <https://doi.org/10.1111/jcmm.15020> PMID: 31991519
37. Jamshidi M, Fagerholm R, Muranen TA, Kaur S, Potdar S, Khan S, et al. High miR-30 Expression Associates with Improved Breast Cancer Patient Survival and Treatment Outcome. *Cancers (Basel).* 2021;13(12):2907. <https://doi.org/10.3390/cancers13122907> PMID: 34200751
38. Selcuklu SD, Donoghue MTA, Spillane C. miR-21 as a key regulator of oncogenic processes. *Biochem Soc Trans.* 2009;37(Pt 4):918–25. <https://doi.org/10.1042/BST0370918> PMID: 19614619
39. Lan F-F, Wang H, Chen Y-C, Chan C-Y, Ng SS, Li K, et al. Hsa-let-7g inhibits proliferation of hepatocellular carcinoma cells by downregulation of c-Myc and upregulation of p16(INK4A). *Int J Cancer.* 2011;128(2):319–31. <https://doi.org/10.1002/ijc.25336> PMID: 20309945
40. Yang T, Hu J, Zhang L, Liu L, Pan X, Zhou Y, et al. CircCUL1 inhibits trophoblast cell migration and invasion and promotes cell autophagy by sponging hsa-miR-30e-3p in fetal growth restriction via the ANXA1/PI3K/AKT axis. *J Biochem Mol Toxicol.* 2024;38(7):e23759. <https://doi.org/10.1002/jbt.23759> PMID: 39003567
41. Miranda K, Yang X, Bam M, Murphy EA, Nagarkatti PS, Nagarkatti M. MicroRNA-30 modulates metabolic inflammation by regulating Notch signaling in adipose tissue macrophages. *Int J Obes (Lond).* 2018;42(6):1140–50. <https://doi.org/10.1038/s41366-018-0114-1> PMID: 29899524
42. Ahmad I, Naqvi RA, Valverde A, Naqvi AR. LncRNA MALAT1/microRNA-30b axis regulates macrophage polarization and function. *Front Immunol.* 2023;14:1214810. <https://doi.org/10.3389/fimmu.2023.1214810> PMID: 37860007
43. Conrad O, Burgy M, Foppolo S, Jehl A, Thiéry A, Guihard S, et al. Tumor-Suppressive and Immunomodulating Activity of miR-30a-3p and miR-30e-3p in HNSCC Cells and Tumoroids. *Int J Mol Sci.* 2023;24(13):11178. <https://doi.org/10.3390/ijms241311178> PMID: 37446353
44. Li X, Guo S, Min L, Guo Q, Zhang S. miR-92a-3p promotes the proliferation, migration and invasion of esophageal squamous cell cancer by regulating PTEN. *Int J Mol Med.* 2019;44(3):973–81. <https://doi.org/10.3892/ijmm.2019.4258> PMID: 31257524
45. Nilsson S, Möller C, Jirstrom K, Lee A, Busch S, Lamb R, et al. Downregulation of miR-92a is associated with aggressive breast cancer features and increased tumour macrophage infiltration. *PloS one.* 2012;7(4):e36051.
46. Lin Z, Roche ME, Díaz-Barros V, Domingo-Vidal M, Whitaker-Menezes D, Tuluc M, et al. MiR-200c reprograms fibroblasts to recapitulate the phenotype of CAFs in breast cancer progression. *Cell Stress.* 2024;8:1–20. <https://doi.org/10.15698/cst2024.03.293> PMID: 38476765
47. Schliekelman MJ, Gibbons DL, Faca VM, Creighton CJ, Rizvi ZH, Zhang Q, et al. Targets of the tumor suppressor miR-200 in regulation of the epithelial-mesenchymal transition in cancer. *Cancer Res.* 2011;71(24):7670–82. <https://doi.org/10.1158/0008-5472.CAN-11-0964> PMID: 21987723
48. Kalinina TS, Kononchuk VV, Yakovleva AK, Alekseenok EY, Sidorov SV, Gulyaeva LF. Association between lymph node status and expression levels of androgen receptor, miR-185, miR-205, and miR-21 in breast cancer subtypes. *International Journal of Breast Cancer.* 2020;2020(1):3259393.

49. Yu S-J, Hu J-Y, Kuang X-Y, Luo J-M, Hou Y-F, Di G-H, et al. MicroRNA-200a promotes anoikis resistance and metastasis by targeting YAP1 in human breast cancer. *Clin Cancer Res*. 2013;19(6):1389–99. <https://doi.org/10.1158/1078-0432.CCR-12-1959> PMID: [23340296](https://pubmed.ncbi.nlm.nih.gov/23340296/)
50. Chen H, Li L, Wang S, Lei Y, Ge Q, Lv N. Reduced miR-126 expression facilitates angiogenesis of gastric cancer through its regulation on VEGF-A. *Oncotarget*. 2014;5(23):11873.
51. Liao K, Gong L, Yang Y, He Y, Wang F, Huang Y, et al. A comprehensive review of research progress in Chinese medicines for primary liver cancer treatment. *Tradit Med Res*. 2022;7(2):10.
52. Zhang H, Lv G, Liu S, Liu D, Wu X. The artificial intelligence watcher predicts cancer risk by facial features. *Tradit Med Res*. 2022;7(1):10.53388.
53. Guo Z, Saw PE, Jon S. Non-invasive physical stimulation to modulate the tumor microenvironment: unveiling a new frontier in cancer therapy. *BIO Integration*. 2024;5(1):986.
54. Mokashi A, Bhatia NM. Integrated network ethnopharmacology, molecular docking, and ADMET analysis strategy for exploring the anti-breast cancer activity of Ayurvedic botanicals targeting the progesterone receptor. *BIO Integration*. 2024;5(1):970.
55. Xu H, Guo Z, Li M, Chaves HV, Pinto V d PT, Du M. Copper-based nanomaterials for image-guided cancer therapy. *BIO Integration*. 2024;5(1):976.
56. Tang J, He J, Guo H, Lin H, Li M, Yang T, et al. PTBP2-Mediated Alternative Splicing of IRF9 Controls Tumor-Associated Monocyte/Macrophage Chemotaxis and Repolarization in Neuroblastoma Progression. *Research (Wash D C)*. 2023;6:0033. <https://doi.org/10.34133/research.0033> PMID: [37040518](https://pubmed.ncbi.nlm.nih.gov/37040518/)
57. Guo XF, Gu SS, Wang J, Sun H, Zhang YJ, Yu PF. Protective effect of mesenchymal stem cell-derived exosomal treatment of hippocampal neurons against oxygen-glucose deprivation/reperfusion-induced injury. *World Journal of Emergency Medicine*. 2022;13(1):46.