Context-Specific Co-expression Networks and Prognostic Roles of LncRNA MIR100HG

Zhanbo Wang
MSc in Data Science
University of Bristol
Bristol, UK
bb24863@bristol.ac.uk

Vidhi Vidhi
MSc in Data Science
University of Bristol
Bristol, UK
yp24973@bristol.ac.uk

Pangbo Shi
MSc in Data Science
University of Bristol
Bristol, UK
ry24085@bristol.ac.uk

Torsa Talukdar
MSc in Data Science
University of Bristol
Bristol, UK
km24473@bristol.ac.uk

***Abstract*—This study investigated the long non-coding RNA MIR100HG across five cancer types (LUAD, PAAD, PRAD, SKCM, STAD) and corresponding normal tissues using primarily gene expression and survival data. Stratification by MIR100HG expression quartiles revealed distinct co-expression networks varying significantly by expression level, cancer type, and normal versus tumor status, indicating highly context-dependent regulation. Kaplan-Meier analysis identified a significant association between high MIR100HG expression and poorer survival (OS, DSS, PFI; p¡0.05) specifically in Stomach Adenocarcinoma (STAD). A Multilayer Perceptron model successfully classified high/low MIR100HG groups (Test AUC=0.935) based on other gene expression, with feature importance highlighting transcription factors like FOXP2. These findings underscore MIR100HG's cancer-specific network alterations and its potential prognostic significance in STAD. Code is available at: https://github.com/EMATM0050-2024/dsmp-2024-groupm26.**

I. Introduction

Long non-coding RNAs (lncRNAs) are important regulatory factors in tumorigenesis and development. This study focuses on a specific lncRNA, *MIR100HG*, which has been identified as a promising candidate lncRNA because it is associated with the Transforming Growth Factor-β (TGF-β) signaling pathway and is significantly upregulated in cancer tissues compared with healthy tissues. Although preliminary studies have shown that *MIR100HG* plays a role in tumor development, its exact mechanism of action remains unclear.

Recent studies have shown that lncRNAs can affect gene expression through multiple mechanisms, including regulating DNA methylation, an important epigenetic process that often changes in cancer. Transcription factors (TFs) also play a key role by guiding gene expression and possibly interacting with methylation mechanisms.

Therefore, the main purpose of this study is to explore the mode of action of *MIR100HG* in tumors. We will utilize multi-omics methods to integrate gene expression data, DNA methylation patterns, and known transcription factor-target gene relationships from patient samples of five different cancer types (pancreatic cancer, lung cancer, melanoma, prostate cancer, and gastric cancer). Patient samples will be stratified into *MIR100HG* high-expression and low-expression groups to determine the key molecular characteristics and pathways that this lncRNA may regulate.

Furthermore, this analysis will be extended to the corresponding normal tissue data to identify the differences in the role of *MIR100HG* between healthy and cancer states. By comparing the results from cancer tissues (Task 1) and normal tissues (Task 2), our aim is to determine the cancer-specific regulatory function of *MIR100HG* (Task 3), thereby clarifying its potential role in tumorigenesis.

II. Literature review

Long non-coding RNAs (lncRNAs) have made an impact in the field of cancer biology. In particular, MIR100HG has received a lot of attention for its modulatory impact on tumor progression through gene expression, transcription factors (TFs), and epigenetic interactions in tumor biology. This review summarizes the findings on MIR100HG across five cancer types: PDAC, LUAD, SKCM, PRAD, STAD, particularly emphasizing its roles in diagnosis and treatment.

In fact, MIR100HG is overexpressed in a number of malignancies, usually exacerbating tumoral progression through TGF-β pathway activation which is instrumental in EMT (epithelial-to-mesenchymal transition) and metastasis [4]. In PDAC, MIR100HG favors tumor growth through sustaining autocrine TGF-β signaling. Preliminary RNA-seq data [3] underscored the belief that expression was indeed greater in cancerous tissues compared to adjacent non-cancerous tissues in a coherent manner across all five cancer types, promoting its value as a pan-cancer biomarker. Additionally, expression variation between histological subtypes suggests some degree of prognostic capability.

TFs and DNA methylation are the main dominators that control MIR100HG. Some of these TFs include MYC, SMAD2/3, and E2F1 (identified by ENCODE and Harmonizome), which bind to its promoter and help in transcription. In particular, MYC causes the MIR100HG to be repressed with the aid of DNA methylation transferring enzymes [1]; reinforcing feedback mechanisms between transcription and chromatin modification.

In cancer, it appears that the main cause of driving MIR100HG overexpression is hypomethylation at the promoter region, while non-cancerous tissues would exhibit suppressive hypermethylation and silencing [2].

The function of lncRNAs is being uncovered in greater detail with multi-omics techniques. Tools such as iCluster and MOFA have stratified patients into subtypes according to their expression of lncRNAs. While some researchers have applied machine learning approaches (random forests, SVMs) to classify tumors based on lncRNA expression profiles, not much work has focused on building MIR100HG specific models. On the contrary, more systemic approaches such as GRN have started characterizing important TF regulators associated with MIR100HG that can be above or below on the regulatory pathway.

Lacking in vivo models or delivery mechanisms for clinical application, along with scant characterization of structural motifs for MIR100HG, remains an obstacle.

MIR100HG appears to be a predominant oncogenic regulator through TGF-β signaling and transcriptional/epigenetic mechanisms. Its multifarious, consistent upregulation across different types of cancer amplifies the possibility of using it as a diagnostic and therapeutic biomarker. There is still a long way to go before clinical translation is feasible, but multi-omics and network modeling certainly enable more in-depth exploration.

III. METHODOLOGY

A. LncRNA Mechanism Modelling

In previous biogenetic studies, long non-coding RNAs have been inferred to have certain modulating role in cancer expression, tumor promotion or inhibition. For instance, lncRNA C22orf32-1 was demonstrated to promote tumorigenesis of NPC [5];lncRNA MALAT1 has been found interaction with EZH2 enhancer, suppressing suppressing E-cadherin expression and promoting osteosarcoma metastasis [6].

Among lncRNAs class, MIR100HG, known as the host gene for the miR-100-let-7a-2-miR-125b-1 cluster, has emerged as a critical player in various cancers. Dysregulated expression of MIR100HG has been associated with oncogenic or tumor-suppressive roles, depending on the cancer type, and is implicated in processes such as cell proliferation, migration, invasion, and chemoresistance [7].

The regulatory mechanisms of lncRNAs, including MIR100HG, can be broadly categorized into transcriptional, post-transcriptional, and epigenetic levels. At the transcriptional level, lncRNAs interact with transcription factors to modulate gene expression.

To model the regulatory mechanisms of MIR100HG, a multi-omics approach integrating RNA-seq, DNA methylation, and transcription factor-target gene networks is employed. By constructing co-expression networks to identify potential regulatory relationships and performing functional enrichment analyses to reveal biological pathways influenced by MIR100HG.

The expression pattern, functions, and clinicopathological characteristics of MIR100HG in diverse cancers have been summarized in Fig1.

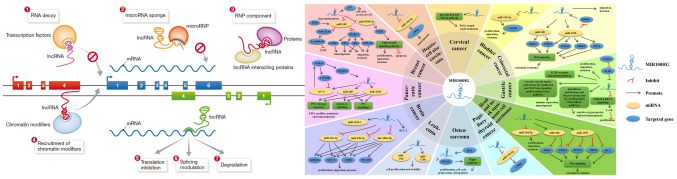

Fig. 1: LncRNA MIR100hg Mechanism (Adapted from Wu et al., 2022)

B. Sample Stratification

Related gene expression data and methylation level in cancer tissues (PDAC, LUAD, SKCM, PRAD, STAD) were stratified into high-expression (HE) and low-expression (LE) groups through dual-threshold approach:

$$\tau_c = \text{Median}(E_c) + k \cdot \text{median}(|E_c - \text{median}(E_c)|) \quad (1)$$

E_c represents $\log_2(\text{TPM} + 0.001)$ values, and k is scaling factor.

In addition, before double-thresholding approach, 25% data near mean value (the intermediate part of the quartiles) in the gene expression distribution was removed, aiming to reduce the effect of data imbalance.

C. Multi-Omics Data Integration

Genomic Region Annotation

The genomic coordinates were annotated hierarchically using *annotatr* in following steps:

- **Promoter regions**: Defined as ± 2 kb from transcription start sites (TSS), verified using ENSEMBL release 75 annotations
- **Gene bodies**: Exons and introns excluding promoter-overlapping regions, with splice variants reconciled via GENCODE v19
- **CpG islands**: Annotated according to UCSC criteria (GC content $\geq$50%, length $>$200 bp, observed/expected ratio $>$0.6), including: Shores (± 2 kb from island boundaries), Shelves (± 2 kb from shores) and Open sea (remaining genomic regions)

Transcription Factor Network Construction

A weighted graph $\mathcal{G} = (V, E, W)$ integrated three relationship classes: Coding Gene $\mathbf{G}$; methylation probes $\mathbf{M}$; TFs $\mathbf{TF}$.

The edges of $\mathcal{G}$ is defined as :

- **Co-expression**: Significant Pearson correlations (ρ) with FDR-adjusted $p < 0.05$:

$$w_{ij}^{\text{exp}} = \frac{1 + \rho(g_i, \text{MIR100HG})}{2} \quad \forall g_i \in \mathbb{G}$$

- **Methylation-gene links**: Probe-gene associations weighted by distance decay:

$$w_{ij}^{\text{meth}} = \beta_{ij} \cdot e^{-\lambda d_{ij}} \quad \lambda = 0.01 \text{ bp}^{-1}$$

- **TF-target**: ENCODE binding evidence transformed to weights:

$$w_{ij}^{\text{TF}} = -\log_{10}(\text{ChIP-seq } p\text{-value}) \cdot \mathbb{I}(p < 10^{-5})$$

And the edge weight is normalizated as follow:

$$w_{ij}^{\text{final}} = \frac{w_{ij}^{\text{raw}} - \min(W_k)}{\max(W_k) - \min(W_k)} \quad k \in \{\text{exp}, \text{meth}, \text{TF}\}$$

D. Feature Extraction

Transcriptional regulators of MIR100HG were identified through correlation analysis followed by functional characterization. All ENCODE-annotated transcription factors (TFs) were initially screened using Pearson correlation with MIR100HG expression across the five cancer types. TFs demonstrating strong association ($|\rho| \geq 0.6$, FDR < 0.01) were retained, with robustness confirmed through 1000 bootstrap resamples requiring stability ($|\rho^*| \geq 0.55$) in $\geq 95\%$ iterations. Cancer-specific regulators were further selected by requiring $\rho \geq 0.7$ in at least one cancer type.

E. Comparative Analysis between Cancerous and Normal Tissues

Predictive Modeling of Cancer-Specific Transcriptional Programs

A MLP neural network model was constructed to assess ability of key TFs and MIR100HG expression feature to distinguish cancers and normal sample. The input features consisted of the expression levels of each sample. The dataset was spited into training (80%) and testing (20%) sets. The architecture of the MLP includes an input layer matching features, two hidden layers with 128 and 64 neurons respectively, each activated by ReLU function, and an output layer with five neurons corresponding to the five cancer types, using a softmax activation function. The model uses Adam optimizer with initial learning rate of 0.01 and categorical cross-entropy loss function.

In addition, the trained model was applied to the normal tissue datasets to investigate separability between cancerous and non-cancerous samples based on the learned transcriptional signatures.

Survival Analysis of Key Transcription Factors and MIR100HG

To explore prognostic relevance of MIR100HG and identified transcription factors, survival analyses were performed using clinical follow-up data extracted from the Survival SupplementalTable dataset. Kaplan-Meier survival curves were generated to visualize differences in overall survival (OS) between groups.

IV. DATA DESCRIPTION/ PREPARATION

The following datasets were provided for this project, primarily sourced from The Cancer Genome Atlas (TCGA) and the Genotype-Tissue Expression (GTEx) project via the UCSC Xena Browser, as well as ENCODE:

Cancer Gene Expression (TCGA): Normalized gene expression levels (log2(TPM + 0.001)) for five cancer types: Pancreatic Ductal Adenocarcinoma (PAAD), Lung Adenocarcinoma (LUAD), Skin Cutaneous Melanoma (SKCM), Prostate Adenocarcinoma (PRAD), and Stomach Adenocarcinoma (STAD). Data was pre-processed to include HGNC gene symbols as row identifiers and patient sample IDs (standardized to TCGA-XX-XXXX format) as columns. This data formed the basis for *MIR100HG* expression stratification and correlation analyses within cancer cohorts.

Normal Tissue Gene Expression (GTEx): Normalized gene expression levels (log2(TPM + 0.001)) for five corresponding normal tissues: pancreas, lung, skin, prostate, and stomach. HGNC gene symbols were added. This data allowed for comparative correlation analysis between normal and cancerous tissues. Note the GTEx sample ID format: GTEX-[donor ID]-[tissue site ID]-SM-[aliquot ID].

Transcription Factor-Target Associations (ENCODE): A dataset detailing experimentally validated or predicted associations between transcription factors (TFs) and their target genes, sourced from the ENCODE project via the Ma-ayan Lab Harmonizome portal. This information was used to interpret correlation results in the context of potential regulatory relationships.

Clinical and Survival Data (TCGA Pan-Cancer): Patient clinical information, including histological subtype, age, gender, stage, and survival endpoints – Overall Survival (OS), Disease-Specific Survival (DSS), and Progression-Free Interval (PFI) – with corresponding time-to-event data (OS.time, DSS.time, PFI.time). This dataset encompassed multiple TCGA cohorts; data for the five specific cancer types of interest was extracted and matched to expression samples (specifically primary tumor samples ending in '-01') for survival analyses. Filtering by common histological subtypes within each cancer was recommended to ensure cohort consistency.

Gene Methylation Data (TCGA, 450k): Genome-wide methylation values (Beta values from Illumina HumanMethylation450 BeadChip) for the five cancer types.

Auxiliary Annotation Files: Supporting files included ID/Gene mapping for methylation probes and hg19 gene annotations, used primarily for data processing and interpretation stages (not directly analyzed in the results presented here).

A. Data Integration and Preprocessing

RNA sequencing data, DNA methylation profiles, and transcription factor (TF)-target gene association datasets were collected from publicly available repositories. Corresponding normal tissue data were acquired from the Genotype-Tissue Expression (GTEx) project [8]. TF-target relationships were retrieved from the ENCODE project [9] through the Harmonizome platform [10].

The integrated datasets included normalized gene expression levels (log-transformed TPM), genome-wide DNA methylation, and curated lists of transcription factor-target gene pairs. Gene expression values were normalized using log transformation:$\log_2(\text{TPM} + 0.001)$.

Then TF-target association files were preprocessed with target genes filtered based on their presence in both RNA-seq and methylation datasets

For DNA methylation data processing, β values representing the methylation fraction at each CpG site were utilized. Probes with detection p-values greater than 0.01, those overlapping single nucleotide polymorphisms (SNPs), or those located on sex chromosomes were excluded. Remaining β values were normalized using quantile normalization to account for technical biases across samples.

To address batch effects and ensure cross-platform comparability, an enhanced ComBat framework is applied for normalizing TCGA (cancer) and GTEx (normal) datasets:

$$E_{ij}^{*} = \frac{E_{ij} - \gamma_i - X_{ij}\beta_i}{\delta_i} + \alpha_i \tag{2}$$

- γ_i, δ_i are platform-specific location and scale parameters to adjust the center and range of the data.
- α_i is a tissue-specific covariate adjustment term to account for systematic differences across tissue types.
- $X_{ij}\beta_i$ is a linear correction matrix for covariates such as age and sex.

Through the normalization, cross-platform data is standardized for integration, with biological features preserved while removing technical artifacts

V. RESULTS AND DISCUSSIONS

A. Sample Stratification based on MIR100HG Expression

To investigate expression-dependent roles of *MIR100HG*, the Lung Adenocarcinoma (LUAD) cohort (N=513) from TCGA was stratified based on *MIR100HG* expression levels. Quartiles were employed to define distinct patient groups, enabling comparative analyses focused on expression extremes. Patients were categorized as Low expression ($\leq$Q1) or High expression ($\geq$Q3), often excluding the middle group (Q1 $<$ Expr. $<$ Q3) to enhance contrast.

The calculated first (Q1) and third (Q3) quartile expression values for *MIR100HG* were 1.2333 and 2.6895, respectively. This stratification resulted in equally sized Low and High expression groups, each containing 129 patients (25.15% of the cohort), as detailed in Table I. These defined Low and High groups formed the basis for subsequent correlation analyses presented in this report.

TABLE I: Patient Stratification based on *MIR100HG* Expression in LUAD Cohort (N=513).

Expression Group	Definition	Sample Count (%)
Low	$\leq$ Q1 (1.2333)	129 (25.15%)
Middle	Q1 $<$ Expr. $<$ Q3	255 (49.71%)
High	$\geq$ Q3 (2.6895)	129 (25.15%)

We conducted the same stratification operation for the other four types of cancer.

B. Correlation Analysis of MIR100HG Expression Groups

Following patient stratification into Low ($\leq$Q1) and High ($\geq$Q3) *MIR100HG* expression groups for each cancer type (LUAD, PAAD, PRAD, SKCM, STAD), Pearson correlation analysis was performed. This analysis aimed to identify genes co-expressed with *MIR100HG* specifically within these extreme expression strata, potentially revealing context-dependent regulatory networks.

Pearson correlation analysis was performed within the stratified High and Low *MIR100HG* expression groups for each cancer type. This aimed to identify context-dependent co-expression patterns. Figures 2 and 3 visualize the top $\pm$10 correlated genes in the LUAD cohort as a representative example, illustrating significant differences between the High and Low expression strata.

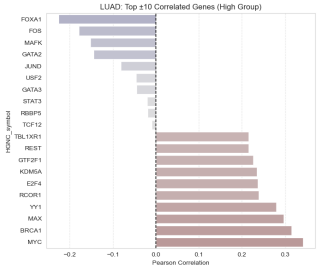

Fig. 2: Top correlated genes with *MIR100HG* in LUAD high-expression group.

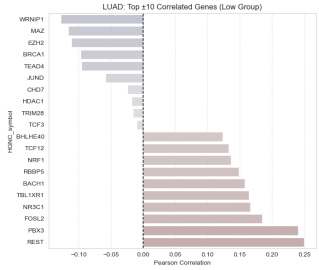

Fig. 3: Top correlated genes with *MIR100HG* in LUAD low-expression group.

1) Moderate correlation strength: Among the five types of cancer, the Pearson correlation between *MIR100HG* and its top co-expressed gene was generally moderate (the most absolute $R < 0.4$), as shown in LUAD (fig:luad_high and fig:luad_low). This indicates complex multifactorial mediation, although a stronger positive correlation ($R > 0.5$) was observed in the PAAD Low group (data not shown).

2) Context-dependent (high and low) : In each type of cancer, there were significant differences in the correlation patterns between the high and low *MIR100HG* groups. For LUAD (fig:luad_high, fig:luad_low), The highest positive correlation shifted from *MYC* and *BRCA1* ($R \approx 0.3$) in the high group to *REST* and *PBX3* ($R \approx 0.24$) in the low group. The negative correlation of *FOXA1* ($R \approx -0.2$) is unique to the high group. This context dependence is widely visible (not shown in the data), indicating that the regulatory effect of *MIR100HG* is affected by its expression level.

3) When comparing all five cancer types, considerable heterogeneity was observed (data for PAAD, PRAD, SKCM, and STAD were not shown). A few genes remained at the highest correlation throughout, highlighting the cancer-specific association of *MIR100HG*, and the PAAD Low group showed a unique pattern dominated by positive correlation.

4) Functional significance: Despite heterogeneity, the most important related genes usually include key regulatory factors such as transcription factors (*MYC*, *REST*, *FOS*) and epigenetic modification factors (*EZH2*, *HDAC1*). This repetitive

co-expression supports the potential involvement of textit-MIR100HG in regulating gene expression programs through environment-dependent mechanisms such as protein interactions or chromatin state changes, which are influenced by their abundance and specific cancer types.

C. Survival Analysis based on MIR100HG Expression

To assess the clinical relevance of *MIR100HG* expression levels, Kaplan-Meier survival analysis was performed comparing patients in the High ($\geq$Q3) versus Low ($\leq$Q1) expression groups previously defined for each cancer type (LUAD, PAAD, PRAD, SKCM, STAD). Survival differences were evaluated using the log-rank test for three endpoints: Overall Survival (OS), Disease-Specific Survival (DSS), and Progression-Free Interval (PFI).

Across most cancer types analyzed (LUAD, PAAD, PRAD, SKCM), no statistically significant differences in OS, DSS, or PFI were observed between the High and Low *MIR100HG* expression groups (log-rank $p > 0.05$ for all endpoints in these cohorts, data plots not shown). This suggests that within these specific cancer contexts, the level of *MIR100HG* expression, when dichotomized by quartiles, may not be a strong independent prognostic indicator for these survival outcomes.

However, a notable and statistically significant association was found in the Stomach Adenocarcinoma (STAD) cohort. As illustrated in Figures 4, patients with High *MIR100HG* expression exhibited significantly poorer survival outcomes compared to those with Low expression across all three endpoints analyzed. Specifically, the log-rank test yielded p-values indicating significant differences for OS ($p = 0.0003$, Fig. 4), DSS ($p = 0.0114$, Fig. 5), . The Kaplan-Meier curves for STAD clearly show the High expression group (red line) having a lower survival probability over time compared to the Low expression group (blue line).

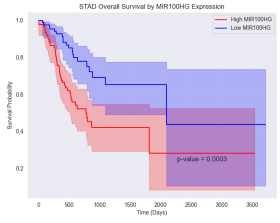

Fig. 4: Overall survival in STAD by *MIR100HG* expression.

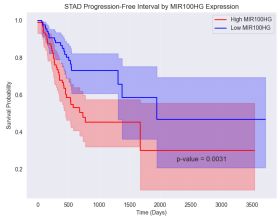

Fig. 5: Disease-specific survival in STAD by *MIR100HG* expression.

This pronounced difference specifically in STAD suggests a potentially significant prognostic role for *MIR100HG* in this cancer type, where high expression is associated with adverse outcomes. The lack of significant association in LUAD, PAAD, PRAD, and SKCM further underscores the cancer-type specific nature of *MIR100HG*'s potential functions and clinical implications, aligning with the heterogeneity observed in the correlation analyses. Further investigation, potentially involving multivariate analysis adjusting for other clinical factors, would be needed to confirm *MIR100HG* as an independent prognostic marker in STAD.

D. MLP Model for Predicting MIR100HG Expression Group

To further explore the relationship between *MIR100HG* status and gene expression more broadly, we used the Multilayer Perceptron (MLP) classifier. The neural network model was trained on combined, scaled expression data for all five cancer types (excluding *MIR100HG* itself) to predict whether a sample belongs to a high (1) or low (0)*MIR100HG* expression group, defined by quartile. The model architecture combines L1 regularization and dropout layers to mitigate overfitting to identify potential predictive features.

The performance of the model during training is summarized in Figure 6. The training accuracy was improved steadily, which was about 95%, and the verification accuracy was about 87%, which showed good learning ability. Correspondingly, both the training loss and the validation loss decrease significantly, and the validation loss tends to be stable, indicating that early stopping effectively prevents a large amount of overfitting. Evaluation of the test set confirmed the predictive ability of the model with an accuracy of 84.2% and an area under the curve (AUC) of 0.935 (detailed metrics and confusion matrix are shown previously). These results indicate that the expression profiles of the selected genes contain predictive information about the expression status of *MIR100HG*.

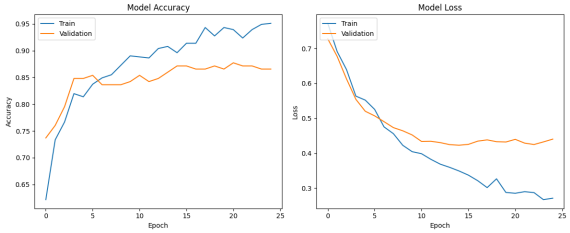

Fig. 6: MLP model training history. (Left) Training and validation accuracy per epoch. (Right) Training and validation loss per epoch.

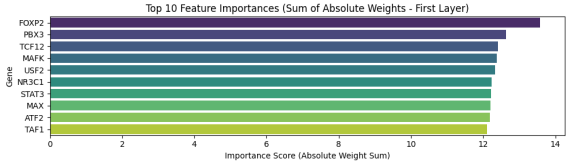

Fig. 7: Top 10 feature importances based on the sum of absolute weights from the MLP's first dense layer.

Feature importance was estimated by summing the absolute weights of the connections of each input gene to the neurons of the first dense layer of the trained MLP. This approach takes advantage of the tendency of L1 regularization to push less relevant feature weights towards zero. The top 10 most important features identified by this method are shown in Figure 7. Notably, the transcription factors *FOXP2*, *PBX3*,

TCF12, *MAFK* and *USF2* contributed the most to the prediction of this model. Several of these genes also appeared in the correlation analysis, although of moderate strength, reinforcing their potential association with *MIR100HG*. The prominence of regulators among the most predictive features further supports the hypothesis that *MIR100HG* is integrated into the gene regulatory network.

In conclusion, the MLP model successfully distinguished the *MIR100HG* high and low expression groups based on the expression of other genes, and achieved good performance on unseen data. Feature importance analysis highlighted several key transcription factors as potentially significant predictors or interactors associated with *MIR100HG* status in concurrent cancer types

E. Analysis of MIR100HG Correlations in Normal Tissues

To investigate whether the observed correlation patterns were cancer-specific, a similar analysis was conducted using gene expression datasets from five corresponding normal tissues obtained from the Genotype-Tissue Expression (GTEx) project (pancreas, lung, skin, prostate, stomach). Samples within each normal tissue type were stratified into Low ($\leq$Q1) and High ($\geq$Q3) *MIR100HG* expression groups using tissue-specific quartiles. Pearson correlation coefficients were then computed between *MIR100HG* and potential transcription factor (TF) or target genes within these strata.

The analysis revealed distinct correlation patterns that varied significantly across the different normal tissues, often differing from the patterns seen in the corresponding cancer types. As a representative example, Figure 8 displays the top $\pm$10 TFs correlated with *MIR100HG* in the High expression group of normal pancreas tissue. In this group, several TFs exhibited strong positive correlations with *MIR100HG*, notably *FOXA1* ($R \approx 0.8$), *TEAD4* ($R \approx 0.7$), and *MAX* ($R \approx 0.75$). This specific pattern highlights potentially important tissue-specific interactions.

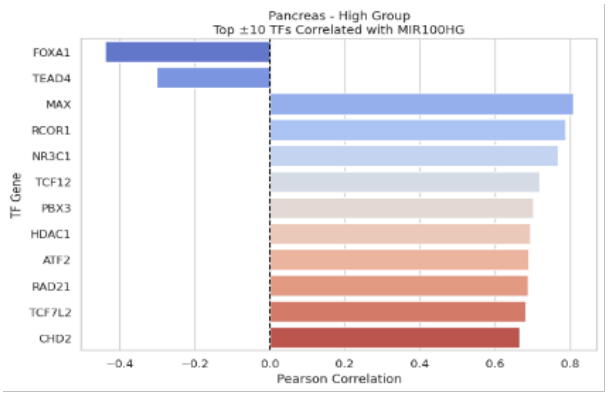

Fig. 8: Top $\pm$10 TFs correlated with *MIR100HG* in the normal Pancreas High expression group. Note the strong positive correlations for *FOXA1*, *TEAD4*, and *MAX*.

Analysis of other normal tissues (data not shown) revealed further heterogeneity. Skin tissue, for instance, displayed inconsistent correlation patterns between its high and low *MIR100HG* expression groups, with different sets of TFs dominating each stratum. Normal stomach tissue exhibited correlation profiles that shared some similarities with normal lung and prostate tissues, suggesting potential shared regulatory mechanisms in these organs, though still distinct from pancreas or skin. Prostate tissue itself showed strong positive correlations with factors like *TCF7L2* and *FOSL2* in the low expression group, contrasting with the pancreas high group.

Overall, these findings in normal tissues strongly suggest that the regulatory context associated with *MIR100HG* expression is highly tissue-specific. The differences observed not only between tissue types but also when compared to their cancerous counterparts imply potential alterations in *MIR100HG*-associated regulatory networks during carcinogenesis.

F. Comparison between Cancer and Normal Tissues

To identify potential cancer-specific functional or regulatory alterations associated with *MIR100HG*, the correlation patterns of high and low expression groups in each cancer species were compared with those in the corresponding normal tissues (from the GTEx project)."

The main finding of this comparison was that the *MIR100HG* co-expression network was significantly different between cancer tissues derived from the same organ and normal tissues. "Although normal tissues themselves exhibit distinct tissue-specific correlation features (see Section V-E for details), these patterns are often quite different from those in matched cancer tissues." This suggests that the gene network associated with *MIR100HG* may undergo significant remodeling during tumorigenesis.

For example, in the high expression group of normal pancreatic tissue (Figure 8), *FOXA1* showed a very strong positive correlation ($R \approx 0.8$) with *MIR100HG*. However, this strong correlation was completely absent in pancreatic ductal adenocarcinoma (PAAD), where *FOXA1* did not enter the list of genes highly correlated with *MIR100HG* in either the high or low expression groups of cancer tissues (data not shown). In contrast, in the cancer context, different transcription factors, such as *PBX3* in the PAAD high expression group or *BHLHE40* in the low expression group ($R \approx 0.6$), became the genes most closely associated with *MIR100HG* expression.

These persistent differences between paired normal and cancer tissues strongly suggest that the functional role of *MIR100HG* and its regulatory interactions are significantly altered in the cancer state. Genes that are tightly co-regulated with *MIR100HG* under normal physiological conditions may be lost in cancer; At the same time, new potential pro-or anti-tumor interactions may emerge in the tumor microenvironment. This remodeling process may involve novel interactions of *MIR100HG* with different protein partners, such as transcription factors or epigenetic modifiers, or be influenced by aberrant signaling pathways that are unique to cancer cells. Identifying these cancer-specific patterns of association will be important to reveal the mechanism of action of *MIR100HG*

in tumorigenesis or progression, as well as to evaluate its potential as a therapeutic target.

VI. Challenges and Limitations

*1*During stratification, by focusing majorly on the top and bottom 25% of *MIR100HG* expressors, many available samples were excluded, which significantly affected the tissues with lower initial counts. This in return reduced the statistical power and increased variability, particularly in stomach, where the sample sizes post stratification was notably small.

*2*Aligning data across datasets posed as a new challenge. Since all the three concerned datasets for this task (phenotype data, TF dataset, and expression data) were obtained from different sources, it required meticulous alignment by sample IDs and gene symbols. Issues related to naming conventions and missing data were handled with extensive pre-processing to ensure reliable integration of datasets.

*3*Pearson correlation coefficients consider linear relationships between TF and target genes, which is convenient and interpretable. However, it may neglect the non-linear interactions that fail to correspond to simple linear outlines. Alternative methods, such as mutual information metrics, could be used to uncover complex dependencies.

*4*Lastly, visualization and interpretations posed a few challenges. Despite offering valuable insights, scatter plots and correlation matrices needed careful selection of TF-target pairs to minimize noise. It was proved vital to emphasize on those TFs that had the strongest correlation changes to highlight meaningful biological patterns.

VII. Further Work and Improvement

By integrating gene expression and clinical data, this study provided a preliminary understanding of the background role and potential significance of *MIR100HG* in five cancer types. However, there are several avenues for future work and improvement:

1) Alternative stratification strategy: The current stratification based on quartiles ($\leq$Q1 vs. $\geq$Q3) effectively highlights differences between extremes of expression but excludes the pooled 50

2) Enhanced network and regulatory analysis: While Pearson correlation identifies co-expressed genes, more complex network inference algorithms (e.g., graphical models, mutual information-based approaches) can be used to infer underlying regulatory relationships and directionality rather than simple linear associations. Explicit integration of transcription factor-target gene association data and potential TF binding site information provided would significantly enrich these network models.

3) Integration of methylation data: A key area for expansion is the inclusion of methylation data (450k arrays) mentioned in the project description but not used in the current analysis. Studying the relationship between *MIR100HG* expression levels and methylation patterns at the promoters or regulatory regions of specific genes, especially those identified as relevant genes or known cancer driver genes, can reveal the key epigenetic mechanisms affected by *MIR100HG*.

VIII. Conclusion

In this study, gene expression and clinical data of five tumor types (LUAD, PAAD, PRAD, SKCM, STAD) and corresponding normal tissues were integrated to investigate the role of the long non-coding RNA *MIR100HG*. By stratified patients according to *MIR100HG* expression and performing correlation, survival, and predictive model analyses, we emerged with several key conclusions.

A common point we observed was that *MIR100HG* had consistent but overall moderate correlations with key regulatory genes, including transcription factors and epigenetic modifiers, across different cancer types. This suggests that *MIR100HG* is generally involved in gene expression regulation pathways.

However, significant differences and context specificity were the main findings. First, the specific genes that are most strongly associated with *MIR100HG* vary greatly in their expression levels within the same cancer type. Second, significant heterogeneity was observed across the five cancer types, each with a different set of associated genes. Third, and perhaps most critically, the correlation patterns in cancer tissues were significantly different from their corresponding normal tissues, strongly suggesting cancer-specific alterations and potential functional shifts in the regulatory network of *MIR100HG* during tumorigenesis. This specificity was further highlighted by survival analysis, which revealed that high *MIR100HG* expression was significantly associated with poor prognosis (OS, DSS, PFI) in STAD patients, indicating a potential cancer type-specific prognostic role.

Taken together, our findings suggest that *MIR100HG* operates highly dependent on the environment, that its expression levels vary according to its own expression levels, specific cancer types, and differ significantly between normal and cancer states. While its general association with regulators is evident, its specific interactions and clinical impact appear to be tissue and disease specific. This study highlights the value of integrated data science approaches in resolving lncRNA function and provides a basis for further investigation of the specific mechanisms and therapeutic potential of *MIR100HG*, especially in STAD.

References

[1] C. Brenner *et al.*, "Myc represses transcription through recruitment of DNA methyltransferase corepressor," *EMBO J.*, vol. 24, no. 2, pp. 336–346, Jan. 2005, doi: 10.1038/sj.emboj.7600509.

[2] B. Jin *et al.*, "Linking DNA methyltransferases to epigenetic marks and nucleosome structure genome-wide in human tumor cells," *Cell Rep.*, vol. 2, no. 5, pp. 1411–1424, Nov. 2012, doi: 10.1016/j.celrep.2012.10.017.

[3] S. Ottaviani *et al.*, "TGF-β induces miR-100 and miR-125b but blocks let-7a through LIN28B controlling PDAC progression," *Nat. Commun.*, vol. 9, art. no. 1845, May 2018, doi: 10.1038/s41467-018-03962-x.

[4] P. Papoutsoglou *et al.*, "The noncoding MIR100HG RNA enhances the autocrine function of transforming growth factor β signaling," *Oncogene*, vol. 40, no. 21, pp. 3748–3765, May 2021, doi: 10.1038/s41388-021-01803-8.

[5] G.-H. Nie, Z. Li, H.-F. Duan, L. Luo, H.-Y. Hu, W.-Q. Yang, L.-P. Nie, R.-F. Zhu, X.-F. Chen, and W. Zhang, "lncRNA C22orf32-1 contributes to the tumorigenesis of nasopharyngeal carcinoma," *Oncol. Lett.*, vol. 13, no. 6, pp. 4487–4492, 2017, doi: 10.3892/ol.2017.6021.

[6] M. Yang, H. Lu, J. Liu, S. Wu, P. Kim, and X. Zhou, "lncRNAfunc: a knowledgebase of lncRNA function in human cancer," *Nucleic Acids Res.*, vol. 50, no. D1, pp. D1295–D1306, 2022, doi: 10.1093/nar/gkab1035.

[7] Y. Wu, Z. Wang, S. Yu, D. Liu, and L. Sun, "LncmiRHG-MIR100HG: A new budding star in cancer," *Front. Oncol.*, vol. 12, art. no. 997532, 2022, doi: 10.3389/fonc.2022.997532.

[8] GTEx Consortium, "The Genotype-Tissue Expression (GTEx) Project," *Science*, vol. 369, no. 6509, pp. 1318–1320, 2020.

[9] ENCODE Consortium, "The ENCODE Project: Cataloging functional elements of the human genome," *Nature*, vol. 489, pp. 57–74, 2012.

[10] Harmonizome, "Harmonizome: Integrating and Mining the ENCODE Project Data," 2016. [Online]. Available: https://maayanlab.cloud/Harmonizome/.