Prediction of non-coding RNAs role in cancer through multi-omics data integration

Shuojingrui He
Department of Engineering
University of Bristol
Bristol, United Kingdom
eo24594@bristol.ac.uk

Weilin He
Department of Engineering
University of Bristol
Bristol, United Kingdom
te23071@bristol.ac.uk

Yibai Tang
Department of Engineering
University of Bristol
Bristol, United Kingdom
xv24324@bristol.ac.uk

Kiran Udayakumar
Department of Engineering
University of Bristol
Bristol, United Kingdom
zt24069@bristol.ac.uk

Mansi Srivastava
Department of Engineering
University of Bristol
Bristol, United Kingdom
ah24844@bristol.ac.uk

***Abstract*—To explore the regulatory mechanism of lncRNA *MIR100HG* in cancer, this study integrated multi-omics data such as gene expression, DNA methylation, and transcription factor-target genes, and systematically analyzed their functions and epigenetic regulatory relationships in five types of cancer. The samples were grouped according to the expression level of *MIR100HG*, the molecular characteristics related to or regulated by it were identified, and compared with normal tissues to reveal the cancer-specific mechanism. Meanwhile, different types of cancer are compared horizontally to explore commonalities and differences. The research adopts the random forest model for feature selection and expression state prediction, and strengthens the identification of key variables and mechanism inference. This study is conducive to clarifying the action pathway of *MIR100HG* and providing theoretical support for its use as a biomarker or therapeutic target.**

I. Introduction

Non-coding RNA, especially long non-coding RNA (lncRNA), although they do not encode proteins, play a key role in cancer regulation. *MIR100HG* is an lncRNA regulated by the TGF-β pathway and upregulated in expression in various cancers. However, its regulatory mechanism and its relationship with methylation and gene expression remain unclear.

Based on this, this study integrated gene expression, DNA methylation and TF–target gene data, focused on five types of cancer (PRAD, PAAD, LUAD, SKCM, STAD), and used logistic regression, random forest and XGBoost models to explore the predictive characteristics of *MIR100HG* expression status.

The objectives of this study are as follows: first, to investigate the molecular features associated with the expression of *MIR100HG* across five cancer types, focusing on identifying common regulators and cancer-specific transcriptional or epigenetic signatures. Second, to evaluate the expression of *MIR100HG* in healthy tissues using GTEx data, and to compare tumor tissues with matched normal tissues from TCGA. This aims to identify regulatory patterns that are specific to cancer, rather than reflecting normal physiological variation. Third, to assess the prognostic relevance of *MIR100HG* expression by performing survival analysis across cancer types. Together, these objectives form an integrative analytical framework to uncover the regulatory role of *MIR100HG* in cancer and its potential as a biomarker for diagnosis and prognosis.

II. Literature Review

Long non-coding RNAs (lncRNAs) have been increasingly recognized as key regulators in cancer. The ENCODE project showed that although protein-coding regions account for less than 2% of the genome, about 80% is transcribed, challenging the notion of noncoding genomic "inactivity" [2]. Uszczynska-Ratajczak et al. annotated tens of thousands of lncRNAs but reported that fewer than 1% were experimentally validated, underscoring a major bottleneck in functional studies [9].

Among them, *MIR100HG*, a TGF-β responsive lncRNA, has shown oncogenic potential. Ottaviani et al. found that TGF-β induces *MIR100HG* expression, upregulating intronic miR-100 and miR-125b to promote EMT and tumor progression [6]. Papoutsoglou et al. further revealed that *MIR100HG*, independent of its miRNA products, can interact with RNA-binding protein HuR (ELAVL1) to stabilize TGFB1 mRNA, thus forming a positive feedback loop that enhances autocrine TGF-β signaling [7]. However, these studies have mainly focused on a single signaling axis or individual cancer types, failing to elucidate the broader multi-omics regulatory functions of *MIR100HG*.

DNA methylation, a key epigenetic mechanism in cancer, also contributes to transcriptional regulation. Jin et al. found that DNMT1 and DNMT3B localize to gene bodies and co-occur with histone marks like H3K36me3, linking methylation to transcription elongation [4]. Brenner et al. showed that transcription factors such as MYC can recruit DNMT3A to silence genes like p21, supporting TF-directed methylation

[1]. Huang et al. summarized lncRNA-mediated methylation via DNMT recruitment, enzyme regulation, and metabolic modulation [3].

TGF-β signaling itself is context-dependent—suppressing early tumors but promoting progression later [8]. Nemeth et al. highlighted lncRNAs' roles in tumor stemness, immunity, and therapy resistance, underscoring their therapeutic potential [5].

III. Methodology

A. Data Collection and Preprocessing

1) Data Collection: This study integrated multi-omics data from TCGA, GTEx, and ENCODE to analyze the regulatory role of *MIR100HG* in five solid tumors (PRAD, PAAD, LUAD, SKCM, STAD). The data include gene expression data ($\log_2$[TPM + 0.001]), 450k methylation data (β value), and the TF–target network from ENCODE (collated by Harmonizome). The normal tissue expression data were obtained from GTEx.

Based on the data provided by the course, this study further incorporated several external resources: a genome annotation table was constructed based on GRCh38 to analyze methylation levels; the pathway-level analysis referenced 3917 functional pathways from MSigDB; additionally, to ensure consistency among multi-omics datasets, UCSC Transcript IDs were mapped to HGNC symbols using the MyGene.info interface.

2) Data Preprocessing: Before the formal analysis, we standardized the structure and matched the samples across all data layers. Column names and sample ID formats were unified to ensure consistency among gene expression, methylation matrices, and *MIR100HG* grouping information. Samples that could not be matched across datasets were excluded. Since the datasets were derived from publicly standardized sources (expression as $\log_2$[TPM + 0.001], methylation as β values), no additional normalization was required.

For missing value handling, no imputation was applied due to the biological sensitivity of gene expression and methylation data, as well as the high and non-random rates of missingness (often exceeding 80%). Instead, features containing missing values were directly removed to preserve the reliability of downstream analyses.

The expression level of *MIR100HG* was grouped using the median-split method. This choice was statistically sound, as the distribution was approximately normal and the median value was close to the mean, as shown in Fig. 1. Furthermore, to reduce the influence of borderline samples in classification, samples within 10% above and below the median were excluded. This refinement enhanced the separation between high- and low-expression groups and improved the robustness of subsequent modeling.

B. Statistical Analysis

1) Differential Expression Gene Analysis (DEG) and Differential Methylation Analysis (DMA): To systematically identify the molecular characteristics associated with the expression status of *MIR100HG*, this study divided the samples of five solid tumors (PRAD, PAAD, LUAD, SKCM, STAD) into high and low expression groups based on the expression level of *MIR100HG*. Differentially expressed gene (DEG) analysis and differential methylation analysis (DMA) were performed to evaluate potential regulatory effects at the transcriptomic and epigenetic levels.

DEG analysis was conducted using the Mann–Whitney U test, combined with false discovery rate (FDR) correction and $\log_2$ fold change calculation, to identify candidate regulatory genes with significantly different expression levels. We chose the U test because it does not require the data to follow a normal distribution and is suitable for skewed data such as RNA expression. DMA was performed on the β-value matrix using independent sample t-tests to evaluate CpG site methylation differences. The independent sample t-test is adopted because the β value has the characteristics of being approximately continuous and close to a normal distribution. Especially when the sample size is sufficient, parametric tests are more suitable, which can detect mean differences more effectively and provide higher test efficacy at the same time. The screening criteria were $|\text{Mean\_Diff}| > 0.2$ and FDR < 0.05. CpG sites were mapped to promoter region genes using the UCSC annotation file and uniformly converted to HGNC gene symbols to ensure consistency across omics layers.

The results of the above analyses represent key molecular characteristics associated with *MIR100HG*, providing the basis for subsequent feature selection, model construction, and regulatory network analysis.

2) Analysis of MIR100HG-Related Methylation Regulatory Mechanism: To explore whether *MIR100HG* indirectly affects gene expression by regulating DNA methylation, this study integrated the results of differentially expressed gene (DEG) and differential methylation (DMA) analyses, focused on promoter and genomic regions, and conducted methylation localization analysis.

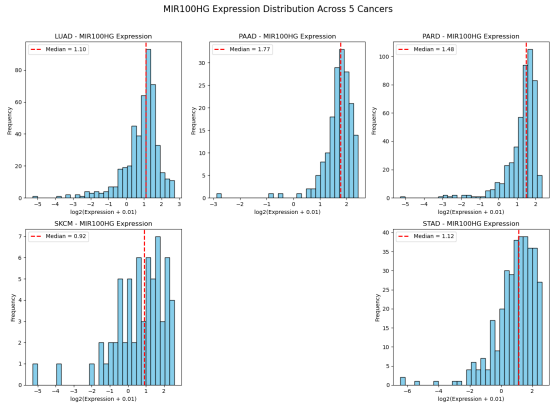

Fig. 1. Distribution of *MIR100HG* expression across samples. The median value (vertical line) was used to group samples. Those within 10% above or below the median were excluded from analysis to reduce boundary interference.

In the promoter analysis, significantly differentially methylated positions (DMPs) were mapped to genomic coordinates to determine whether they were located within the defined promoter regions. Subsequently, by integrating gene expression changes, genes showing a negative correlation between methylation and expression were identified, revealing potential downstream targets that might be indirectly regulated by *MIR100HG*.

Furthermore, DMPs outside the promoter regions were also localized. A genomic annotation table based on GRCh38 was constructed to screen for methylation sites falling into specific genomic regions. These sites were then combined with gene expression changes to explore potential target genes subject to methylation regulation at the genomic level.

3) TF Regulatory Network Identification: This study identified the key transcription factors (TFs) related to *MIR100HG* through the following process: Firstly, they were grouped based on the expression level of *MIR100HG*, and differentially expressed genes (DEGs) were screened in each type of cancer as potential regulatory target genes. Subsequently, this gene set was matched with the known TF–target network, the number of DEGs regulated by each TF was counted, and thereby its potential importance in the *MIR100HG* regulatory network was measured. The more target genes regulated, the higher the TF weight.

4) Analysis of Differences Between Cancer and Normal Pathways: To evaluate the relationship between the expression status of *MIR100HG* and changes in pathway function, this study conducted pathway difference analysis on five types of cancers based on GTEx normal tissues and TCGA tumor tissues. A total of 3917 functional pathways from MSigDB were used. The $\log_2$(TPM + 0.001) expression matrix was mapped to the member genes of each pathway, and differences in pathway expression between tumor and normal groups were assessed using the Mann–Whitney U test.

The screening criteria were $p < 1 \times 10^{-5}$ and $|\log_2 \mathrm{FC}| \geq 3.0$. Pathways meeting these thresholds were considered significant and were further analyzed to identify biological pathways specifically enriched in either cancer or normal tissues, thereby exploring potential tissue-specific markers.

5) Prognostic Analysis of MIR100HG: To assess the prognostic value of *MIR100HG* expression across different cancer types, we performed comprehensive survival analyses in five cancer cohorts (PRAD, PAAD, LUAD, SKCM, STAD). Three analytical steps were included:

First, Kaplan–Meier survival curves were generated to visualize overall survival differences between high and low expression groups, with statistical significance evaluated via log-rank tests. To enhance the discriminatory power, samples within the $\pm 10\%$ range around the median expression were excluded from group assignments.

Second, multivariate Cox proportional hazards regression models were applied to determine whether *MIR100HG* expression was an independent prognostic factor after adjusting for clinical covariates such as age and tumor stage. Hazard ratios (HRs), 95% confidence intervals, and p-values were calculated for each cancer type.

Finally, we summarized sample sizes, death counts, HRs, and significance levels across all cancer types in a comparative table to assess cross-tissue prognostic patterns. This analysis provides a foundation for further exploration of the molecular mechanisms and context-specific roles of *MIR100HG* in cancer prognosis.

IV. Model Construction

A. Feature Construction and Selection

To construct a classification model for the expression status of *MIR100HG*, this study integrated multi-omics features, including genome-wide expression levels and differential methylation information, to form a structurally unified input matrix.

Specifically, the model features consist of two sources: (1) the $\log_2$(TPM + 0.001) expression values of genes, used to characterize the overall transcriptional status; and (2) the methylation levels of CpG sites that are significantly associated with *MIR100HG* expression, reflecting potential epigenetic regulatory effects. The sample label variable Y represents the expression status of *MIR100HG*, categorized into two groups: high expression and low expression.

To enhance model stability and avoid overfitting caused by high-dimensional features, a two-stage feature selection strategy combining biological priors and statistical dimensionality reduction was adopted. First, candidate features significantly associated with *MIR100HG* were initially screened based on results from differential expression analysis (DEG) and differential methylation analysis (DMA). Next, from the initially selected features, low-variance features were eliminated using variance filtering. The remaining variables were scored using the ANOVA F-test, and the top-ranked features with the highest discriminative power were retained to construct the final model input matrix.

It is worth noting that we did not adopt linear transformation techniques such as Principal Component Analysis (PCA) for dimensionality reduction. Although PCA can compress feature dimensions, the resulting components are linear combinations of multiple mixed original variables, making it difficult to trace back to specific genes or CpG sites, which hinders biological interpretation.

Taking the PRAD dataset as an example, a modeling dataset containing 495 samples and 100 selected features was ultimately obtained. Based on this dataset, classification models including logistic regression, random forest, and XGBoost were trained and evaluated.

B. Model Selection

To construct a prediction model for the expression status of *MIR100HG*, three commonly used supervised classifiers were trained for each cancer type: logistic regression (LR), random forest (RF), and extreme gradient boosting (XGBoost). The label variable Y represents the high- and low-expression groups of *MIR100HG*.

TABLE I
PERFORMANCE COMPARISON OF CLASSIFICATION MODELS USING DIFFERENT FEATURE SETS

Model	Accuracy	Recall	AUC	Feature Set
XGBoost	0.8929	0.8984	0.9610	After selection
Logistic Regression	0.9131	0.9315	0.9759	After selection
Random Forest	0.8889	0.8911	0.9707	After selection
XGBoost	0.8828	0.8782	0.9612	Original
Logistic Regression	0.8949	0.9033	0.9744	Original
Random Forest	0.8970	0.8910	0.9707	Original

Five-fold cross-validation was applied to assess model performance. Evaluation metrics included accuracy, precision, recall, F1 score, and the area under the ROC curve (AUC). The results, summarized in Table I, show that all three models achieved strong predictive performance, with AUC values above 0.96 across tumor types. Logistic regression slightly outperformed others in some metrics, though its linear nature may limit its ability to capture complex interactions. XGBoost, while slightly lower in AUC, showed better generalization in high-dimensional feature spaces. Random forests offered a strong balance between accuracy and interpretability, making them especially useful for biological interpretation and feature tracking.

In summary, random forest was selected as the primary modeling method for subsequent analyses.

V. CONCLUSION OF COMPARATIVE ANALYSIS OF CANCER TISSUES

A. The Main Differential Characteristic Model Results of MIR100HG High and Low Expression Groups in Different Cancer Types

Importance scores were based on the Gini impurity metric in the random forest, the top 20 most discriminative key genes in each cancer type were identified using a random forest model. We then performed functional enrichment analysis based on the KEGG pathway database to uncover core regulatory genes

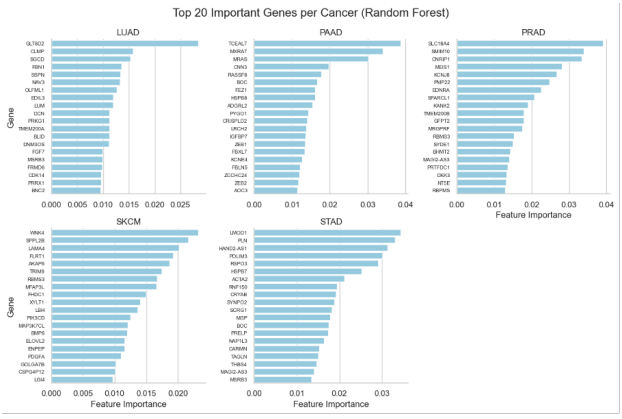

Fig. 2. Top 20 most important genes identified by the random forest model and their importance distribution across five cancer types.

TABLE II
TOP ENRICHED KEGG PATHWAYS AND CORRESPONDING HIT GENES ACROSS CANCERS

KEGG Pathway	Cancer Count	Hit Genes
Calcium signaling pathway	4	EDNRA, FGF7, PDGFA, PLN
Transcriptional misregulation in cancer	4	CDK14, MEIS1, PDGFA, ZEB1
Regulation of actin cytoskeleton	3	FGF7, MRAS, PIK3CD, PDGFA
Vascular smooth muscle contraction	3	ACTA2, EDNRA, PRKG1
Pathways in cancer	3	EDNRA, FGF7, LAMA4, PIK3CD, PDGFA
Diabetic cardiomyopathy	3	GFPT2, PIK3CD, PLN

and signaling pathways involved in each cancer. We also investigated the recurrent enrichment patterns across different cancer types.

As shown in Fig. 2, the genes GLT8D2 in LUAD, TCEAL7 in PAAD, and SLC16A4 in PRAD demonstrate significantly higher importance compared to other genes, suggesting their dominant regulatory roles. In contrast, the importance scores of the top 10 genes in SKCM and STAD are relatively even, implying a synergistic contribution from multiple genes.

Fig. 3 illustrates the intersection of the top 20 characteristic genes among the five cancers. Most genes are specific to a single cancer type, indicating strong cancer-type specificity. Only a limited number of overlapping genes are found between certain cancer combinations. In the next step, we conduct enrichment analysis of the top 20 genes in each cancer to gain deeper biological insights.

Table II summarizes the KEGG pathways significantly enriched in at least three cancer types, along with the correspond-

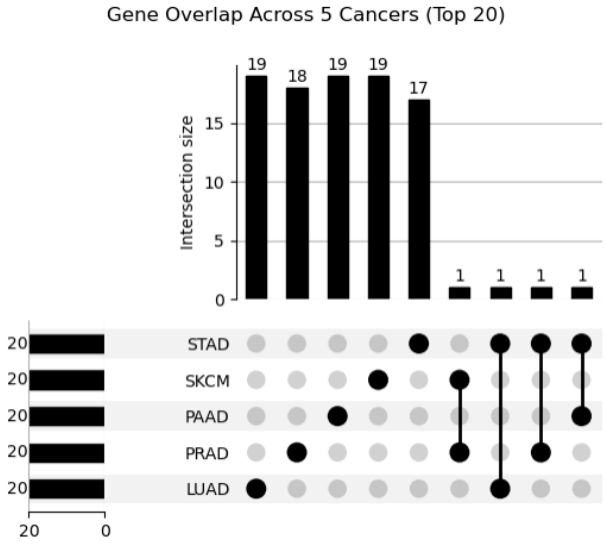

Fig. 3. Overlap of top 20 important genes across five cancer types.

ing hit genes. Among them, the calcium signaling pathway, PI3K–Akt signaling, and Rap1 signaling appeared repeatedly across multiple cancers, suggesting that they may constitute core components of a pan-cancer regulatory network.

Furthermore, the hit genes identified—such as PIK3CD, FGF7, and PDGFA—were involved in multiple enriched pathways. This recurrence indicates that these genes could serve as potential cross-cancer pathway regulators.

By integrating machine learning-based gene importance ranking, pathway enrichment analysis, and cross-cancer comparison, this study not only identified critical regulators specific to each cancer type but also uncovered shared signaling mechanisms and candidate key drivers across cancers.

B. Significantly Different Transcription Factors and Their Target Genes

To investigate the core structure and cancer-specific features of the *MIR100HG*-associated regulatory network, differentially expressed genes (DEGs) from five cancer types were matched to the TF–target network obtained from the ENCODE database. The number of DEGs regulated by each transcription factor (TF) within individual cancers and across cancer types was calculated as a proxy for regulatory importance. This analysis enabled the identification of key TFs that consistently appeared across cancers (Fig. 4).

Further, a heatmap was generated to visualize the regulatory intensity of the top 10 high-frequency TFs across the five cancers, as shown in Fig. 5. The results revealed that POLR2A, CTCF, EZH2, H2AFZ, and EP300 regulate a substantial number of DEGs in multiple cancer types, indicating their roles in a potential cross-cancer core regulatory network. Notably, POLR2A and CTCF demonstrated consistently high regulatory strength, suggesting their positions as central hub nodes in the *MIR100HG*-related pathway. Additionally, the involvement of EZH2 and HDAC2 suggests that *MIR100HG*-related networks may be coupled with epigenetic silencing mechanisms.

Moreover, MYC, MAX, CHD1, and HDAC2 exhibited tissue-specific regulatory activity, particularly in PRAD and STAD, while LUAD presented fewer core TFs. This may reflect a tissue-specific regulatory architecture in LUAD, or the presence of alternative non-transcriptional mechanisms. These findings highlight the heterogeneous regulatory patterns of *MIR100HG* among different cancer types.

C. Genes Co-regulated by DNA Methylation

1) Genome-wide Patterns of Methylation Regulation in Different Cancers: To assess the impact of DNA methylation on gene expression, both gene expression and methylation levels of all DEGs were analyzed.

In promoter regions, a general trend of "decreased methylation + increased expression" was observed across most cancer types. In PRAD, 94% of differential probes were hypomethylated, and 343 genes were significantly upregulated in the *MIR100HG* high-expression group. Similar patterns were observed in PAAD and STAD, with the most pronounced change in STAD ($\log_2$ FC up to 4.73). In contrast, SKCM exhibited a bidirectional regulatory mode, with enrichment in both expression activation and "increased methylation + decreased expression" pairs. However, LUAD did not show a dominant pairing pattern, implying that *MIR100HG* may act through alternative nonmethylation pathways in this context.

Methylation regulation within non-promoter (genomic) regions appeared limited. Only a few genes in some cancer types followed the "decreased methylation + increased expression" trend, suggesting sparse and context-specific regulation.

In summary, promoter methylation remains the primary mechanism associated with gene expression changes across cancers in relation to *MIR100HG*.

2) Association Analysis of Methylation and Expression in MIR100HG Promoter Region: Based on the clear trend of overall methylation regulation, we further analyzed the methylation status of the *MIR100HG* promoter itself to investigate whether it participates in cancer regulation through the classical mechanism of "hypermethylation inhibiting expression."

Among the five cancer types, PAAD displayed a typical regulatory pattern. Approximately 75% of the probes showed a significant negative correlation (Spearman correlation coefficient, FDR-adjusted $p < 0.05$) with *MIR100HG* expression, supporting the hypothesis that its expression is suppressed by promoter methylation. In contrast, in STAD, PRAD, and LUAD, most probes exhibited a positive correlation with gene expression, suggesting that the upregulation of *MIR100HG* in these cancers might be independent of promoter methylation. This finding warrants further validation in combination with other multi-omics data, as summarized in Table III.

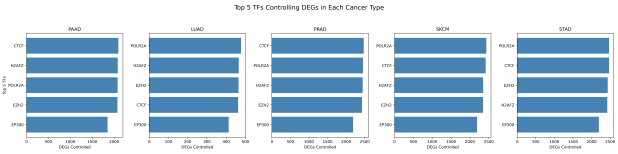

Fig. 4. Number of differentially expressed genes regulated by top transcription factors in each cancer type.

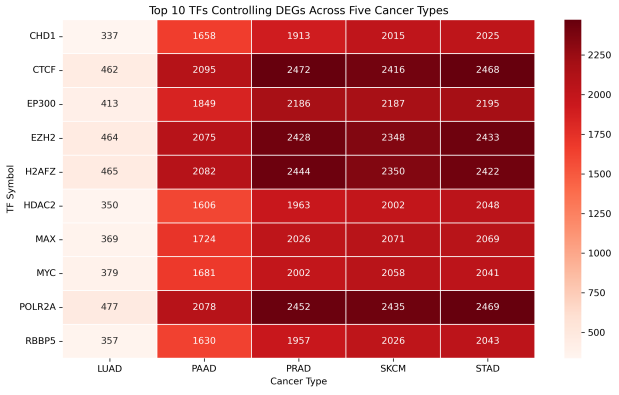

Fig. 5. Regulatory intensity of the top 10 transcription factors across five cancer types. Heatmap color represents the number of DEGs regulated.

TABLE III
CORRELATION ANALYSIS BETWEEN PROMOTER METHYLATION PROBES AND *MIR100HG* EXPRESSION ACROSS CANCER TYPES

Cancer Type	Valid Probes	Significant Negative Correlation Probes	Significant Positive Correlation Probes	Dominant Trend	Sample Size
STAD	16	2 (12.5%)	14 (87.5%)	Primarily positive	371
LUAD	7	0 (0%)	7 (100%)	Primarily positive	453
PRAD	55	6 (10.9%)	36 (65.5%)	Primarily positive	495
PAAD	4	3 (75%)	1 (25%)	Primarily negative	178
SKCM	56	2 (3.6%)	5 (8.9%)	Both minimal	102

VI. CONCLUSION OF NORMAL TISSUE ANALYSIS

A. Differential Expression of MIR100HG in Normal and Tumor Tissues

To examine the expression differences of *MIR100HG* between normal and tumor tissues, we mapped the normal tissue samples to corresponding cancer types and conducted independent sample t-tests between *MIR100HG* expression levels in tumor versus normal tissues.

As shown in Table IV, *MIR100HG* was significantly downregulated in PRAD, LUAD, and SKCM, while it was upregulated in PAAD. All corresponding p-values were statistically significant ($p < 0.001$), indicating that *MIR100HG* may participate in distinct regulatory roles across tumor types. In STAD, however, no significant difference was observed ($p \approx 0.097$), suggesting potential involvement of more complex, tissue-dependent mechanisms. This might be due to the high heterogeneity of the STAD subtypes or the low expression difference of *MIR100HG* in STAD.

B. Differences in Pathway Activation Patterns Across Cancer Types

Enrichment analysis was performed on DEGs between tumor and normal tissues to identify systematic changes in functional pathways. The results showed that PAAD and STAD demonstrated enhanced activity in several cancer-related pathways, including DNA unwinding, cell cycle progression, and epithelial–mesenchymal transition (EMT), indicating strong functional activation.

In contrast, LUAD, PRAD, and SKCM predominantly exhibited downregulation in pathways such as PI3K–ERBB4 signaling, lipid metabolism, and immune response, suggesting suppression of core functional modules during tumorigenesis.

Cross-cancer comparisons revealed several shared pathways—including cell cycle regulation, PI3K–AKT signaling, and metabolic remodeling—were involved in all five cancer types. The shared but variably regulated nature of these pathways indicates the context-specific reprogramming of typical carcinogenic circuits. PAAD exhibited the highest number of tumor-specific activated pathways (83), indicating substantial transcriptomic reprogramming, while SKCM and LUAD had the most normal tissue-specific pathways (84 and 30, respectively), suggesting extensive pathway silencing.

C. Differential Regulation and Key Marker Pathway Identification

By comparing gene expression patterns between normal and tumor tissues, several cancer-specific markers were identified. In LUAD, the most significantly downregulated genes—KANK3, BTNL9, and FCN3—are associated with immune barrier functions. In PAAD, STAD, and SKCM, multiple lncRNAs (e.g., *LINC02577*, *HOTTIP*) and pseudogenes (e.g., *RPS27P13*) were upregulated in tumor tissues, implying roles in immune evasion and transcriptional stress. In PRAD, ribosome-related pseudogenes such as *RPL17P16* were upregulated, whereas DNA repair-related genes (*PARP6*, *SEC31B*) were downregulated, indicating dysfunction in ribosome and repair mechanisms.

D. Comparison with TF–Target Network and Biomarker Exploration

To determine whether *MIR100HG* influences pathway reprogramming through transcription factors, we compared enrichment results from tumor and normal tissues with pathways derived from the TF–target network. The TF–based predictions showed lower fold changes and weaker statistical significance, implying that *MIR100HG* may act indirectly by modulating

TABLE IV
COMPARISON OF *MIR100HG* EXPRESSION BETWEEN NORMAL AND TUMOR SAMPLES

Cancer Type	Normal Mean	Tumor Mean	T-Statistic	p-Value
PRAD	5.06	2.58	27.78	1.94×10^{-71}
LUAD	3.78	2.01	17.65	7.02×10^{-56}
PAAD	1.64	3.30	-11.55	7.72×10^{-26}
SKCM	3.86	0.88	13.19	1.23×10^{-24}
STAD	2.28	1.90	1.67	9.67×10^{-2}

TABLE V
SURVIVAL ANALYSIS OF *MIR100HG* ACROSS CANCER TYPES

Cancer Type	Sample Size	Deaths	HR	p-Value	Significant
PAAD	183	95	0.98	0.92	No
PRAD	563	10	0.92	0.90	No
STAD	402	151	1.51	0.01	**Yes**
SKCM	100	28	1.22	0.61	No
LUAD	529	197	0.95	0.73	No

transcription factors, rather than directly regulating broad downstream gene expression.

Several pathways also emerged as potential biomarkers. REACTOME_UNWINDING_OF_DNA and KEGG_TYPE_I_DIABETES_MELLITUS were significantly upregulated in PAAD, supporting their utility in tumor screening. The RNA interference pathway was activated in STAD, potentially aiding in the early diagnosis of digestive system tumors. Meanwhile, inhibition of IRS_ACTIVATION in SKCM may serve as a marker for immune escape subtypes.

E. Prognostic Value of MIR100HG Expression

We performed a survival analysis of *MIR100HG* across five cancer types using Kaplan–Meier curves, log-rank tests, and multivariate Cox regression to evaluate its independent prognostic value.

As shown in Table V, only in STAD was *MIR100HG* significantly associated with poor survival outcomes (HR = 1.51, $p = 0.01$), serving as an independent risk factor, This indicates that the functional role of *MIR100HG* may be particularly active or uniquely related in the tumor biology of gastric cancer. In other cancers, no significant correlation was observed, possibly due to tumor heterogeneity, variability in *MIR100HG* expression, or limited event counts.

VII. Further Work and Improvement

This study constructed a comprehensive multi-omics framework centered on the long non-coding RNA *MIR100HG* and identified potential regulatory pathways and key factors across five cancer types, whereas several limitations were identified during the analysis. Future research can focus on the following directions to improve the exploration of MIR100HG's functional mechanisms and enhance its potential in clinical transformation.

A. Expanding Sample Coverage and Cancer Types

During the proccess of model construction, Significant differences in data quality and expression dynamics across cancer types were observed, which directly impacted performance and interpretability. For instance, in STAD, MIR100HG expression did not differ significantly between tumor and normal tissues ($p \approx 0.097$), suggesting a potential role for non-transcriptional regulation or tissue-specific mechanisms. Meanwhile, the insufficient number of normal tissue samples in cancer types such as SKCM and LUAD constrained the statistical ability for differential analysis and model evaluation.

To address these issues, future efforts can integrate public datasets from GEO, ICGC, and ArrayExpress to enhance both tumor and normal sample coverage—particularly in underrepresented subtypes. For cancers with weak or inconsistent *MIR100HG* signals, functionally related lncRNAs (e.g., from co-expression or co-regulation modules) can be incorporated to build more generalizable and robust regulatory models.

B. Expanding Regulatory Dimensions and Network Reconstruction

While this study focused on transcriptional regulation through gene expression, DNA methylation, and transcription factor interactions, *MIR100HG* may participate in more regulatory processes. Future work could consider integrating additional molecular layers such as non-coding RNA interactions, post transcriptional modulation, and chromatin accessibility data to explore its overall function more comprehensively. Understanding of context specific regulatory roles may be improved by combining the data through network based or multi omics integration strategies.

C. Model Strategy Optimization and Interpretability Enhancement

This study adopted a "single-cancer optimization, cross-cancer transfer" strategy, where model construction and feature selection were initially performed in PRAD and then applied to other cancer types to evaluate regulatory transferability. While this approach improved model consistency and comparability, it may underperform in adapting to cancer-specific regulatory architectures. For example, in SKCM, the flat distribution of feature importance revealed a more polygenic regulation mode, for which multi-label or hierarchical learning methods may be more suitable.

Notably, the classification models using default parameters has achieved a strong predictive performance ($AUC > 0.96$), indicating that the data itself has robust discriminatory power. Nevertheless, future work may investigate hyperparameters optimization to further improve robustness and generalizability, particularly in borderline cases.

Currently, the model outputs and pathway enrichment results have not been fully integrated and still rely on manual annotation of top-ranked features. Future studies can combine GSEA, pathway-based embedding, and multi-modal learning strategies to to achieve the automation of the reasoning chain: "genes - pathways - mechanisms." We also suggest building a double layer modeling architecture to capture both cancer-specific and cross cancer regulatory patterns, and apply transfer learning, attention mechanisms, and multitask modeling to more effectively reveal regulatory structures and enhance generalization across diverse cancer contexts.

VIII. Conclusions

This study systematically characterized the potential regulatory mechanisms of *MIR100HG* across five representative solid tumors by integrating transcriptomic data, DNA methylation profiles, and TF–target regulatory networks. The results demonstrate that the expression state of *MIR100HG* in tumor tissues is significantly associated with widespread molecular alterations, suggesting its multi-layered and synergistic roles in cancer development and progression.

At the transcription factor regulation level, we found that high *MIR100HG* expression robustly activates a set of recurrent, cross-cancer TFs such as ZEB1, EBF1, and FOXP2. The downstream targets of these TFs exhibited consistently

upregulated expression and were significantly enriched in key oncogenic processes including epithelial-mesenchymal transition (EMT), immune response, and stemness maintenance. This implies that *MIR100HG* may function as a transcriptional hub, orchestrating tumor cell fate reprogramming. Moreover, cancer-specific TF modules (e.g., FOXP2 in STAD, REST and TCF12 in SKCM) highlight the tissue-dependent regulatory responses of *MIR100HG*.

At the DNA methylation level, we identified a large number of promoter-associated CpG sites that were significantly demethylated in *MIR100HG*-high groups and strongly associated with the upregulation of corresponding genes. This supports the classical "demethylation–activation" model, particularly evident in STAD and PAAD. Furthermore, the methylation level of the *MIR100HG* promoter itself showed cancer-specific correlations with its expression: most probes in PAAD exhibited negative correlations, supporting repression via methylation, while LUAD and STAD showed predominantly positive correlations, suggesting alternative epigenetic regulatory mechanisms.

At the pathway and systems biology level, *MIR100HG* expression was closely linked to the remodeling of several core cancer pathways. PAAD and STAD demonstrated extensive pathway activation, including cell cycle progression, DNA damage response, and inflammatory signaling. In contrast, PRAD, SKCM, and LUAD exhibited systematic repression of functional pathways. Compared to TF–target networks, these pathway-level changes were more pronounced, indicating that *MIR100HG* may not act as a direct driver but rather fine-tunes downstream signaling through the modulation of TF networks to meet the adaptive needs of the tumor microenvironment.

In summary, *MIR100HG* does not exert its effects through a singular mechanism. Instead, it regulates tumor phenotypes through an integrative model of transcriptional regulation, epigenetic modification, and pathway activation. This combined strategy includes both conserved cross-cancer modules and cancer-type-specific regulatory patterns, highlighting its potential clinical relevance. We also identified several pathways (e.g., DNA unwinding and glycolipid-inflammatory signaling in PAAD, IRS signaling in SKCM) that were specifically activated under high *MIR100HG* expression, providing promising candidates for future validation as early diagnostic markers or therapeutic targets.

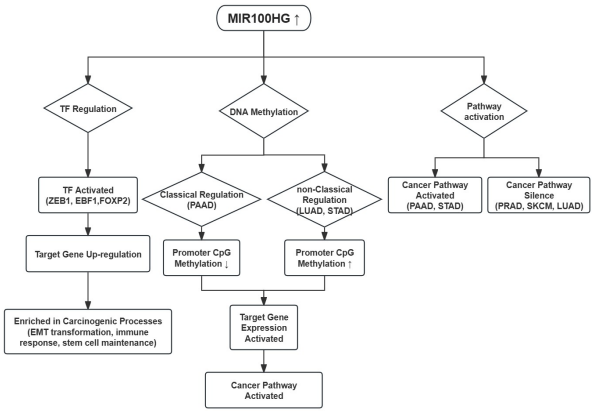

Fig. 6. Schematic diagram summarizing the multi-level regulatory mechanism of *MIR100HG* in cancer: through transcription factor regulation, DNA methylation, and pathway activation.

ACKNOWLEDGMENT

The source code for this study is available at: https://github.com/EMATM0050-2024/dsmp-2024-m28.git

REFERENCES

[1] Carmen Brenner, Rachel Deplus, Céline Didelot, Axelle Loriot, Emmanuelle Viré, Charles De Smet, Arantxa Gutierrez, Davide Danovi, David Bernard, Thierry Boon, et al. Myc represses transcription through recruitment of dna methyltransferase corepressor. *The EMBO journal*, 24(2):336–346, 2005.

[2] ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *Nature*, 447(7146):799–816, 2007.

[3] Wanxu Huang, Hua Li, Qingsong Yu, Wei Xiao, and Dan Ohtan Wang. Lncrna-mediated dna methylation: an emerging mechanism in cancer and beyond. *Journal of Experimental & Clinical Cancer Research*, 41(1):100, 2022.

[4] Bilian Jin, Jason Ernst, Rochelle L Tiedemann, Hongyan Xu, Suhas Sureshchandra, Manolis Kellis, Stephen Dalton, Chen Liu, Jeong-Hyeon Choi, and Keith D Robertson. Linking dna methyltransferases to epigenetic marks and nucleosome structure genome-wide in human tumor cells. *Cell reports*, 2(5):1411–1424, 2012.

[5] Kinga Nemeth, Recep Bayraktar, Manuela Ferracin, and George A Calin. Non-coding rnas in disease: from mechanisms to therapeutics. *Nature Reviews Genetics*, 25(3):211–232, 2024.

[6] Silvia Ottaviani, Justin Stebbing, Adam E Frampton, Sladjana Zagorac, Jonathan Krell, Alexander de Giorgio, Sara M Trabulo, Van TM Nguyen, Luca Magnani, Hugang Feng, et al. Tgf-β induces mir-100 and mir-125b but blocks let-7a through lin28b controlling pdac progression. *Nature communications*, 9(1):1845, 2018.

[7] Panagiotis Papoutsoglou, Dorival Mendes Rodrigues-Junior, Anita Morén, Andrew Bergman, Fredrik Pontén, Cédric Coulouarn, Laia Caja, Carl-Henrik Heldin, and Aristidis Moustakas. The noncoding mir100hg rna enhances the autocrine function of transforming growth factor β signaling. *Oncogene*, 40(21):3748–3765, 2021.

[8] Louise Richardson, Scott G Wilcockson, Luca Guglielmi, and Caroline S Hill. Context-dependent tgfβ family signalling in cell fate regulation. *Nature Reviews Molecular Cell Biology*, 24(12):876–894, 2023.

[9] Barbara Uszczynska-Ratajczak, Julien Lagarde, Adam Frankish, Roderic Guigó, and Rory Johnson. Towards a complete map of the human long non-coding rna transcriptome. *Nature Reviews Genetics*, 19(9):535–548, 2018.