Biologically Coherent MicroRNA Signature for Machine-Learning-Based Colorectal Cancer Detection

Rebaz M. Mustafa[1†], Kamaran Abdoulrahman[2] and Mohammad Hamid[3]

[1]Department of Chemistry, Faculty of Science and Health, Koya University, Koya 44023, Kurdistan Region – F.R. Iraq

[2]Department of Chemistry, College of Science, Salahaddin University-Erbil, Erbil, 44001, Kurdistan Region, Iraq

[3]Department of Molecular Medicine, Biotechnology Research Center, Pasteur Institute of Iran, Tehran, Iran

***Abstract*—Colorectal cancer (CRC) biomarker discovery is frequently limited by signature instability, methodological heterogeneity, and insufficient external validation. This study aimed to develop a reproducible bioinformatics and machine-learning framework for identifying a compact mature-microRNA (miRNA) signature capable of discriminating CRC tissues from matched normal controls. Small RNA-seq data from a paired discovery cohort (PRJNA387172; 104 tumor–normal pairs) and three independent external high-throughput sequencing cohorts (GSE125961, GSE183437, and GSE236991) were processed through uniform pre-processing, mature-miRNA quantification, counts-per-million filtering, and TMM/logCPM normalization. Within the discovery cohort, pair-preserving 80:20 partitioning was combined with repeated stratified five-fold cross-validation with 10 repeats, train-only differential-expression analysis, and consensus feature selection using Elastic Net, Boruta, and random forest recursive feature elimination. This framework identified a stable locked three-miRNA signature comprising hsa-miR-135b-5p, hsa-miR-378a-5p, and hsa-miR-549a-5p. A random forest classifier trained on this panel achieved perfect discrimination in the discovery partitions, with an area under the curve (AUC) of 1.000 in both the training and independent test subsets, and showed preliminary external performance, with panel AUCs of 0.917, 1.000, and 0.840 across three small validation cohorts and 0.961 in pooled external analysis. Expression direction was preserved across datasets, with miR-135b-5p and miR-549a-5p upregulated and miR-378a-5p downregulated in tumors. Downstream target integration analysis recognized 401 overlap-supported miRNA-target interactions spanning 393 unique genes, with enrichment in Wnt, Hippo, transforming growth factor-beta, focal adhesion, and CRC-related pathways. These findings support a biologically coherent three-miRNA random forest model for CRC biomarker discovery and prioritize candidates for future independent and experimental validation.**

***Index Terms*—Biomarker discovery, Colorectal cancer, Feature selection, microRNA, Random forest.**

I. Introduction

Colorectal cancer (CRC) remains one of the most prevalent causes of cancer incidence and mortality worldwide, and its severe clinical burden is largely due to late diagnosis and biological heterogeneity (Siegel, Giaquinto and Jemal, 2024, Li, et al., 2024). This complexity has intensified the investigation for biomarkers that can discriminate cancerous from non-cancerous tissues with high accuracy while still reflecting biologically significant disease mechanisms.

MicroRNAs (miRNAs) are small non-coding RNAs that regulate gene expression in the post-transcriptional level and are involved in multiple aspects of colonic tumorigenesis, including proliferation, invasion, metastasis, stemness, and therapy response (Włodarczyk, et al., 2025, Pająk, et al., 2025, Shirzad, et al., 2025). Although miRNAs represent promising biomarker candidates, reproducible miRNA biomarker discovery in CRC remains challenging. Reported signatures often vary substantially across studies because of differences in biospecimen type, cohort composition, pre-processing, normalization, and candidate-selection strategies (Schwab and Nonaka, 2025, Naranbat, et al., 2025). Accordingly, recent evidence suggests that multi-miRNA panels outperform single-miRNA markers and those smaller combinations, including miRNA panels, may provide favorable diagnostic performance. However, substantial between-study heterogeneity and limited validation remain major barriers to translation (Khan, Ali, and Wahab, 2025, Schwab and Nonaka, 2025). They are particularly relevant to high-dimensional sequencing data, in which correlated variables and small-to-moderate sample sizes may produce unstable biomarker lists if feature selection is performed only

ARO-The Scientific Journal of Koya University
Vol. XIV, No.2 (2026), Article ID: ARO.13036. 13 pages
DOI: 10.14500/aro.13036
Received: 24 May 2026; Accepted: 12 July 2026
Regular research paper; Published: 08 November 2026
†Corresponding author's e-mail: rebaz.muhtasm@koyauniversity.org
Copyright © 2026 Rebaz M. Mustafa, Kamaran Abdoulrahman and Mohammad Hamid. This is an open access article distributed under the Creative Commons Attribution License (CC BY-NC-SA 4.0).

once or without rigorous resampling control (Pudjihartono, et al., 2022). An initial relevance hierarchy is provided by filter methods, informative subsets are iteratively identified by wrapper methods, such as Boruta, and the removal of repetitive characteristics and selection are directly integrated into model development by incorporated methods, such as LASSO and Elastic Net (Wang, et al., 2025). Recent CRC studies have shown that Boruta-based and other machine-learning frameworks can recover diagnostically relevant miRNA signatures, but systematic evaluations still indicate that external validation is not sufficiently implemented (Wang, et al., 2025, DeGroat, et al., 2024). Consequently, robust CRC biomarker discovery increasingly depends on repeated resampling, consensus-driven feature selection, and unchanged testing in independent cohorts rather than reliance on a single optimized training result.

Moreover, pathway-centered reviews of CRC consistently highlight Wnt, transforming growth factor-beta (TGF-β), Hippo, and adhesion-related signaling networks as major components of colorectal tumor biology (Li, et al., 2024, Tejeda-Muñoz and Mei, 2024, Fasano, et al., 2024, Mohammadpour, et al., 2024), making biologically anchored miRNA signatures especially relevant for mechanistic interpretation.

A leakage-aware small RNA-sequencing framework was implemented in the present study to identify and externally validate a concise mature-miRNA signature capable of distinguishing CRC tissues from their adjacent normal tissues. To improve signature stability and reduce method-specific bias, differential-expression (DE) analysis was integrated with repeated cross-validated consensus feature selection across Elastic Net, Boruta, and random forest recursive feature elimination. The final locked panel, including three miRNAs, was then assessed without retraining or modification in independent external high-throughput sequencing cohorts, and complementary downstream analyses were performed to evaluate its association with canonical CRC pathways. This design was aimed to identify a small yet stable diagnostic miRNA signature with strong discriminatory performance and biological coherence, thereby providing a more reproducible framework for CRC biomarker discovery and validation.

II. Materials and Methods

A. Data Acquisition, Processing, and Normalization

sRNA-seq data from PRJNA387172, GSE125961, GSE183437, and GSE236991 were retrieved from the NCBI Sequence Read Archive using prefetch 3.0.10 and converted to FASTQ format with fasterq-dump 3.0.10. Read quality was assessed using FastQC 0.12.1 and summarized with MultiQC 1.30. Adapter trimming and quality filtering were performed with fastp 0.23.2, using automatic adapter detection, quality-based trimming of read termini, and a minimum retained read length of 15 nt. Processed reads were aligned to the human reference genome (hg38) with miRDeep2 (mapper.pl) and Bowtie 1.3.1, and known human miRNAs were quantified using miRDeep2 (quantifier.pl) against human mature and hairpin reference sequences. Quantification tables were imported into R 4.5.2, where counts were collapsed to the mature-miRNA level by summing precursor-specific entries. Low-abundance features were removed using a counts-per-million filter, and count-based datasets were normalized by the trimmed mean of M-values method implemented in edgeR 4.8.2, followed by log-counts-per-million transformation for downstream analyses.

To address the risk of information leakage, pre-processing was separated from outcome-aware modeling. The initial CPM filtering and TMM/logCPM normalization were applied as unsupervised count-data pre-processing steps to remove extremely low-abundance miRNAs and adjust for library-size and RNA-composition differences before any biomarker selection. These steps did not use class labels to rank miRNAs, estimate differential expression, select features, tune the classifier, or determine the final diagnostic panel. All outcome-dependent procedures, including DE testing, individual-miRNA area under the curve (AUC) estimation, consensus feature selection, random forest training, and threshold-based performance assessment, were performed only within the Train80 data or within the training folds of repeated cross-validation (Chen, Lun and Smyth, 2016, Robinson and Oshlack, 2010, Law, et al., 2014).

B. Study Design and Analytical Framework

The study was designed to derive and validate a robust CRC-associated miRNA signature from high-throughput sequencing datasets through a staged analytical workflow as below presented, comprising discovery, feature selection, locked-panel derivation, model development, external validation, and downstream systems-level interpretation (Fig. 1). The paired PRJNA387172 cohort served as the primary discovery dataset, whereas GSE125961, GSE183437, and GSE236991 were used for external validation.

C. Differential Expression and Repeated Cross-validated Feature Selection

Feature discovery was performed exclusively within the large discovery cohort through repeated stratified five-fold cross-validation with ten repeats to stabilize selection estimates and minimize optimistic bias. Within each training fold, low-expression miRNAs were filtered using a CPM-based rule, after which DE analysis was performed without access to held-out samples. A DGEList object was constructed in edgeR 4.8.2, normalization factors were estimated by TMM, and the retained count matrix was analyzed using voom, lmFit, and eBayes in limma 3.66.0. Differentially expressed miRNAs were prioritized through adjusted P-value with absolute log2 fold change and were then forwarded to machine-learning-based feature selection.

Three complementary feature-selection strategies were applied to the fold-specific candidate set: Elastic Net using glmnet 4.1–10, Boruta using Boruta 9.0.0, and random forest recursive feature elimination using randomForest 4.7–1.2 with utilities from caret 7.0–1. A strict intersection of features retained by all three methods was preferred

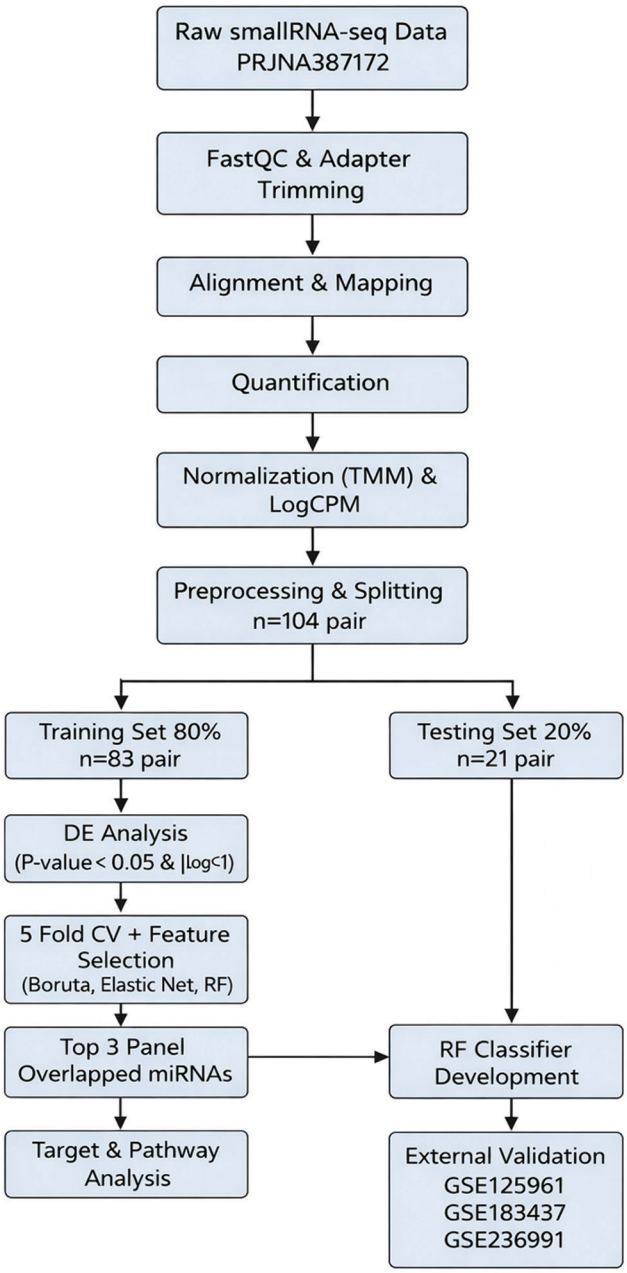


Fig. 1. Overview of the analytical workflow for derivation and validation of the colorectal cancer-associated microRNA signature.

when at least three candidates were recovered; otherwise, a fallback consensus comprising miRNAs selected by at least two methods was used. Within each training fold, the final three-miRNA panel was defined by integrating absolute log2 fold change, adjusted P-value, and individual training-fold AUC. After completion of all folds and repeats, a stable locked three-miRNA signature was derived from overall selection frequency, mean absolute log2 fold change, mean individual AUC, and consistency of expression direction across resampling iterations. This workflow developed miR-135b-5p, miR-378a-5p, and miR-549a-5p as the final locked mature-miRNA signature. Although a fallback consensus rule was pre-defined for rare folds with fewer than three strict-intersection candidates, the final locked panel was not derived from fallback-only selection because all three retained miRNAs were selected by Boruta, Elastic Net, and RF-RFE.

D. Classifier Development, Performance Assessment, and External Validation

The locked miRNA panel was used to train the definitive classifier on the discovery dataset. The finalized workflow used a random forest classifier implemented with randomForest 4.7–1.2, with supporting utilities from caret 7.0–1. Tumor-class probabilities were generated for each sample and converted to class assignments using a fixed decision threshold of 0.50. Model discrimination was evaluated by receiver operating characteristic analysis using pROC 1.19.0.1, with AUC as the principal metric. Additional performance measures included confusion-matrix counts, accuracy, specificity, sensitivity, precision, and F1-score. Data handling and graphical output were generated using data.table 1.18.2.1, dplyr 1.2.0, and ggplot2 4.0.2. The final random forest used 1000 trees; with three predictors, the default mtry was 1, and other randomForest classification settings were left at package defaults. In addition, uncertainty diagnostics were performed on the available locked panel prediction outputs and are provided in Supplementary Table S1 and Fig. 6, including confidence intervals, calibration summaries, Brier scores, and permutation-based null distributions.

For external validation, the locked feature set and discovery-derived model were applied unchanged to GSE125961, GSE183437, and GSE236991. Both individual-miRNA discrimination and panel-level performance were evaluated in each dataset. A pooled high-throughput sequencing validation analysis was additionally performed after harmonization of the locked miRNAs across external cohorts, enabling assessment of signature behavior in a larger external sample set under the same performance framework used for discovery. Because these cohorts were small, external testing was interpreted as preliminary validation of directionality and model portability rather than definitive clinical generalizability.

E. Downstream Target Retrieval and Systems-level Analysis

To define the putative regulatory landscape of the locked miRNA panel, a structured downstream bioinformatics workflow was implemented in R 4.5.2 using multiMiR 1.32.0, dplyr 1.2.0, tibble 3.3.1, tidyr 1.3.2, clusterProfiler 4.18.4, org.Hs.eg.db 3.22.0, and ggplot2 4.0.2.

Predicted and experimentally validated miRNA–target interactions were retrieved separately for each locked miRNA using get_multimir() in multiMiR 1.32.0 for human targets (org = “hsa”), without limiting returned interaction number or imposing an additional retrieval-stage score cutoff. Duplicate interactions were removed using dataset-specific composite keys. In this context, “experimentally validated” denotes database-curated interaction records retrieved through multiMiR rather than new wet-lab target validation performed in the present study.

Consensus predicted targets were defined by collapsing distinct miRNA–gene–database combinations and retaining only interactions supported by at least two independent prediction resources. Support from two, three, or four or more databases was classified as moderate, strong, or very strong, respectively. High-confidence targets were then defined by intersecting the consensus-predicted set with experimentally validated miRNA–gene interactions. Interactions supported by at least two predictive resources and at least one validated source were considered high confidence, whereas those with three or more predictive resources in addition to validation were considered very high confidence.

F. Shared-target Structure, Network Construction, and Enrichment Analysis

Shared-target architecture was analyzed from the overlap-supported target layer by reshaping distinct miRNA–gene pairs into a gene-by-miRNA membership matrix using dplyr 1.2.0 and tidyr 1.3.2. Targets were classified as unique, partially shared, or shared by all retained miRNAs. A bipartite miRNA–target network was then constructed from the overlap-supported interactions, with edge attributes capturing interaction type, prediction-support count, evidence class, predicted and validated resources, validation methods, PMIDs, and target-sharing class.

Functional enrichment was performed with clusterProfiler 4.18.4 using (org.Hs.eg.db 3.22.0) for annotation. Gene symbols were converted to Entrez identifiers using bitr, after which Gene Ontology (GO) enrichment for biological process, molecular function, and cellular component was performed with enrichGO, and pathway enrichment was performed with enrichKEGG. Significance was defined using Benjamini–Hochberg-adjusted $p < 0.05$ and $q < 0.20$.

G. Reproducibility and Software Environment

HTS data download and primary pre-processing were conducted in Ubuntu 24.04.2 LTS using prefetch 3.0.10, fasterq-dump 3.0.10, FastQC 0.12.1, MultiQC 1.30, fastp 0.23.2, miRDeep2, and Bowtie 1.3.1, as detailed in Section 2.1. All downstream analyses were performed in R 4.5.2 on Windows 11 x64, and session information was archived to preserve the exact computational environment. The core analytical workflow used data.table 1.18.2.1, dplyr 1.2.0, stringr 1.6.0, ggplot2 4.0.2, pROC 1.19.0.1, randomForest 4.7-1.2, edgeR 4.8.2, limma 3.66.0, glmnet 4.1-10, Boruta 9.0.0, and caret 7.0-1. Downstream target and enrichment analyses additionally used multiMiR 1.32.0, tibble 3.3.1, tidyr 1.3.2, clusterProfiler 4.18.4, and org.Hs.eg.db 3.22.0. All public accession identifiers are reported in the manuscript, and the principal R scripts and Ubuntu commands are provided as supplementary material to support reproducibility.

III. Results

A. Discovery Framework

The large discovery cohort comprised 208 samples, including 104 matched tumor-normal pairs, after pre-processing and TMM/logCPM normalization. The three external high-throughput sequencing (HTS) validation datasets contained 6, 5, and 5 matched pairs (12, 10, and 10 individual samples), respectively (Table I). Counts-per-million (CPM) filtering reduced the discovery feature space from 1,476 mature miRNAs to 368 retained features without loss of samples. Pair-aware partitioning then generated a balanced 80:20 split of the filtered discovery cohort, yielding 83 pair training samples and 21 pair independent test samples while preserving the matched structure (Fig. 2).

B. Repeated Cross-validation and Feature selection

A second train-only filtering step reduced the 368 retained discovery features to 364 miRNAs in the Train80 subset. The train-only volcano plot showed a bidirectional distribution of significant features, with distinct groups of upregulated and downregulated miRNAs exceeding the combined statistical and fold-change thresholds (Fig. 3).

Repeated 5-fold cross-validation with 10 repeats yielded 50 feature-selection runs. Across these runs, 109 unique DE candidates entered the machine-learning stage, of which 75, 93, and 92 unique miRNAs were selected at least once by Elastic Net, Boruta, and RF-RFE, respectively. The strict three-method intersection contained 47 unique miRNAs overall (Fig. 4), but only three consistently exceeded the pre-defined stability threshold of 0.70: hsa-miR-549a-5p (strict intersection rate = 1.00), hsa-miR-135b-5p (0.90), and hsa-miR-378a-5p (0.78) (Table II). These three features constituted the final locked signature. The relatively large logFC values in Table II are interpreted as discovery-cohort effect-size indicators after logCPM normalization and strong tumor-normal tissue separation, rather than assay-ready quantitative fold-change estimates.

TABLE I
Cohort Composition and Retained Feature Space

Dataset	Scale	Sample pairs	Quantified microRNAs	References
Large	TMM_logCPM	104	1476	Wu, et al., 2020
Large_filtered	TMM logCPM	104	368	
Train80	TMM logCPM	83	368	
Test20	TMM logCPM	21	368	
GSE125961	TMM_logCPM	6	1476	Zhou, et al., 2020
GSE183437	TMM_logCPM	5	1476	Hozaka, et al., 2021
GSE236991	TMM_logCPM	5	1476	Moiz, et al., 2025

Large_filtered denotes the CPM-filtered discovery matrix used for pair-aware train-test splitting; Train80 and Test20 are derived from this filtered cohort. Sample pairs refer to matched tumor-normal pairs; the corresponding individual sample numbers are twice the pair counts.

TABLE II
Repeated-Cross-Validation Stability Metrics of the Final Locked Three-miRNA Signature

microRNA	Feature Selections	Average_Rate_ Threshold	logFC	adj.P.Val
hsa-miR-549a-5p	Boruta, ENet, RF-RFE	1.000	4.976	$1.158e^{-47}$
hsa-miR-135b-5p	Boruta, ENet, RF-RFE	0.900	3.445	$9.437e^{-58}$
hsa-miR-378a-5p	Boruta, ENet, RF-RFE	0.780	−2.226	$6.117e^{-49}$

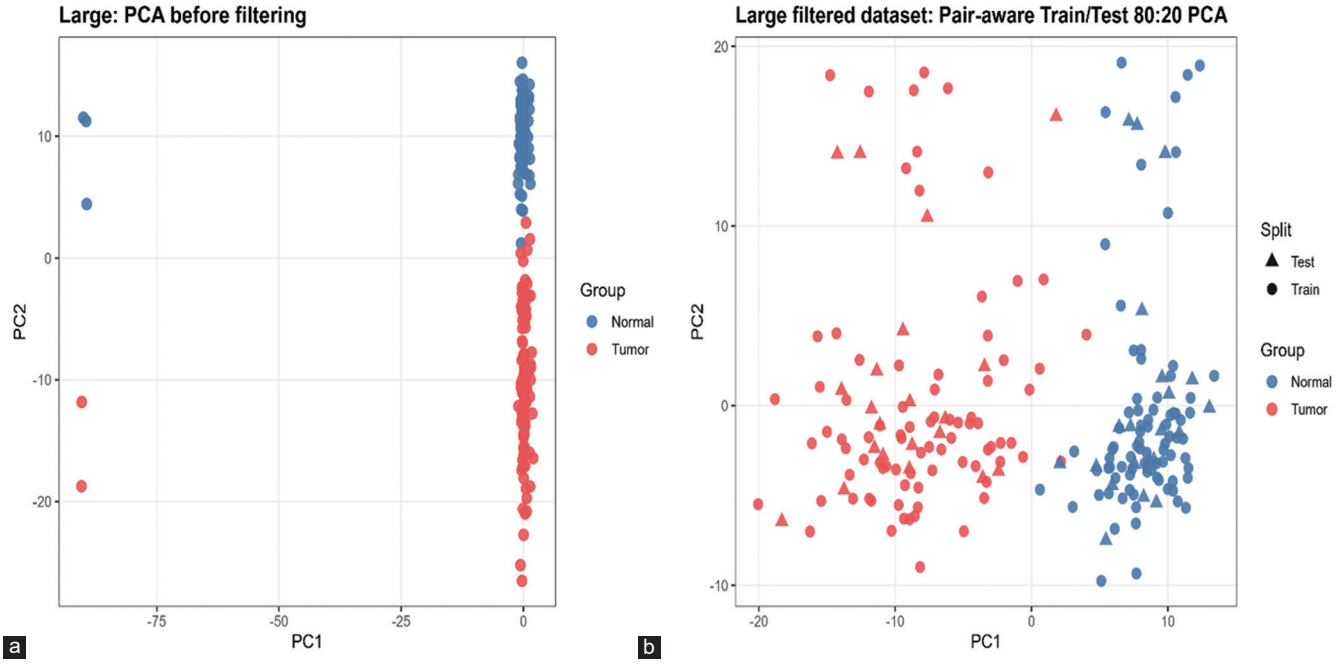


Fig. 2. Principal component analysis of the large discovery cohort. (a) Principal component analysis (PCA) plot before filtering. (b) PCA plot of the paired 80:20 train-test split.

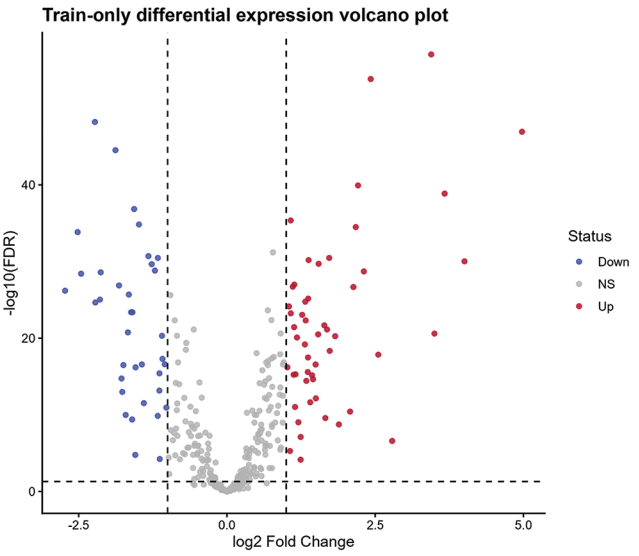


Fig. 3. Train-only differential-expression volcano plot in the filtered Train80 subset. Red and blue points denote significantly up and down regulated microRNAs, respectively.

C. Diagnostic Model Development

The three-miRNA random forest classifier achieved perfect discrimination in both discovery partitions. The locked panel produced an AUC of 1.000 in Train80 and 1.000 in Test20, with no false-positive or false-negative classifications in either subset. Because such performance is unusual in biomedical datasets, these discovery-cohort estimates were treated as descriptive and interpreted together with the pair-aware split, train-only feature selection, and external-cohort behavior. When the same panel was applied supportively to the full large cohort, performance remained perfect. At the individual-miRNA level, all three locked features retained high discrimination, with test-set AUC values of 1.000 for hsa-miR-135b-5p, 0.998 for hsa-miR-549a-5p, and 0.973 for hsa-miR-378a-5p (Table III and Fig. 5).

Moreover, uncertainty diagnostics supported the locked discovery performance while clarifying its precision (Supplementary Table S1 and Fig. 6). The locked panel showed complete separation in Train80, Test20, and the supportive full-large analysis; however, the confidence intervals explicitly quantified uncertainty around the classification metrics, particularly in the smaller Test20 subset. The permutation analyses showed that the observed AUC of 1.000 was far outside the label-permuted null distributions, which were centered near 0.50, with permutation $P = 0.001$ across the three discovery-based assessments. Calibration plots showed close agreement between mean predicted tumor probability and observed tumor fraction, and the Brier scores were low in Train80 (0.001), Test20 (0.005), and the supportive full-large analysis (0.001). These diagnostics therefore support the non-random and internally consistent behavior of the locked panel, while maintaining a cautious interpretation of complete discovery-cohort separation.

Consistent with the classifier performance, the boxplots showed marked separation of the three locked miRNAs between normal and tumor tissues in both discovery partitions. hsa-miR-135b-5p and hsa-miR-549a-5p were shifted upward in tumor samples, whereas hsa-miR-378a-5p showed a reciprocal downward shift (Fig. 7). This expression pattern was preserved in both Train80 and Test20, indicating that the locked signature retained stable directional behavior after pair-aware partitioning.

Across all samples in the full large cohort, hsa-miR-549a-5p correlated positively with hsa-miR-135b-5p (Spearman

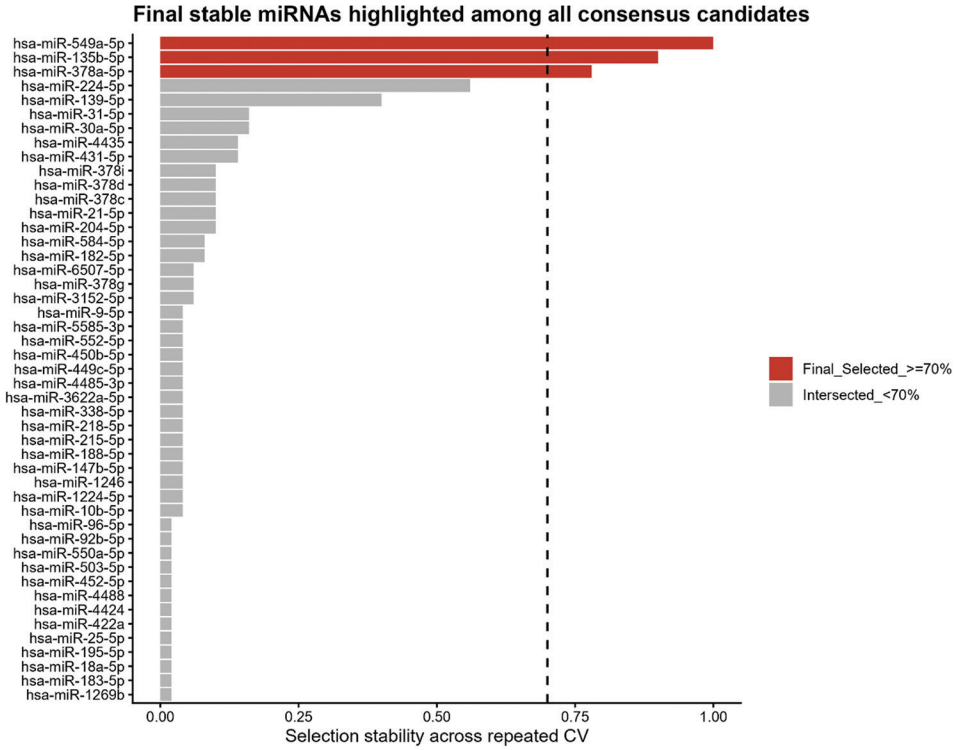


Fig. 4. Selection frequencies of strict-intersection microRNAs across 50 repeated cross-validation runs. The dominant stable features were hsa-miR-549a-5p, hsa-miR-135b-5p, and hsa-miR-378a-5p.

TABLE III
Locked-Panel Performance Across Discovery, Supportive and Validation Cohorts

Dataset	miRNA	AUC	Accuracy	Sensitivity	Specificity	Precision	F1
Train80	miRNAs panel	1.000	1.000	1.000	1.000	1.000	1.000
Test20	miRNAs panel	1.000	1.000	1.000	1.000	1.000	1.000
Train80	hsa-miR-549a-5p	1.000	0.988	0.988	0.988	0.988	0.988
Test20	hsa-miR-549a-5p	0.998	0.952	0.952	0.952	0.952	0.952
Train80	hsa-miR-135b-5p	0.998	0.982	0.976	0.988	0.988	0.982
Test20	hsa-miR-135b-5p	1.000	1.000	1.000	1.000	1.000	1.000
Train80	hsa-miR-378a-5p	0.998	0.976	0.976	0.976	0.976	0.976
Test20	hsa-miR-378a-5p	0.973	0.881	0.905	0.857	0.864	0.884

miRNAs: microRNAs, AUC: Area under the curve

rho = 0.771), whereas both miRNAs correlated inversely with hsa-miR-378a-5p (rho = −0.759 and −0.750, respectively). Thus, the locked panel not only preserved strong class separation but also showed coherent co-variation when tumor and normal samples were analyzed together.

D. External HTS Validation across Cohorts

The direction of dysregulation was preserved across all external HTS datasets: hsa-miR-135b-5p and hsa-miR-549a-5p remained increased in tumor samples, whereas hsa-miR-378a-5p remained decreased. At the panel level, discrimination was strongest in GSE183437 (AUC = 1.000), remained high in GSE125961 (AUC = 0.917), and declined in GSE236991 (AUC = 0.840), with sensitivity remaining 1.000. When the three external cohorts were pooled, the locked panel achieved an AUC of 0.961, an accuracy of 0.906, a sensitivity of 0.938, and a specificity of 0.875. At the individual-miRNA level in pooled external validation, hsa-miR-135b-5p showed the largest AUC (0.980), followed by hsa-miR-378a-5p (0.949) and hsa-miR-549a-5p (0.914) (Table IV and Fig. 8). Given the limited number of external pairs, these performance estimates should be regarded as exploratory point estimates.

For the pooled external locked-panel classification metrics, exact binomial confidence intervals were summarized in Supplementary Table S2 and were wide, as expected from the small validation size. Therefore, external AUC values were retained as exploratory point estimates and interpreted cautiously. Figure 9 presents functional enrichment. For clarity, Figure 9 is an original figure generated from the present study's target-enrichment analysis using clusterProfiler in R, with plots produced using ggplot2.

E. GO and Kyoto Encyclopedia of Genes and Genomes (KEGG) Enrichment

Downstream target retrieval yielded 15,036 predicted rows and 3,933 validated rows, and the overlap-supported

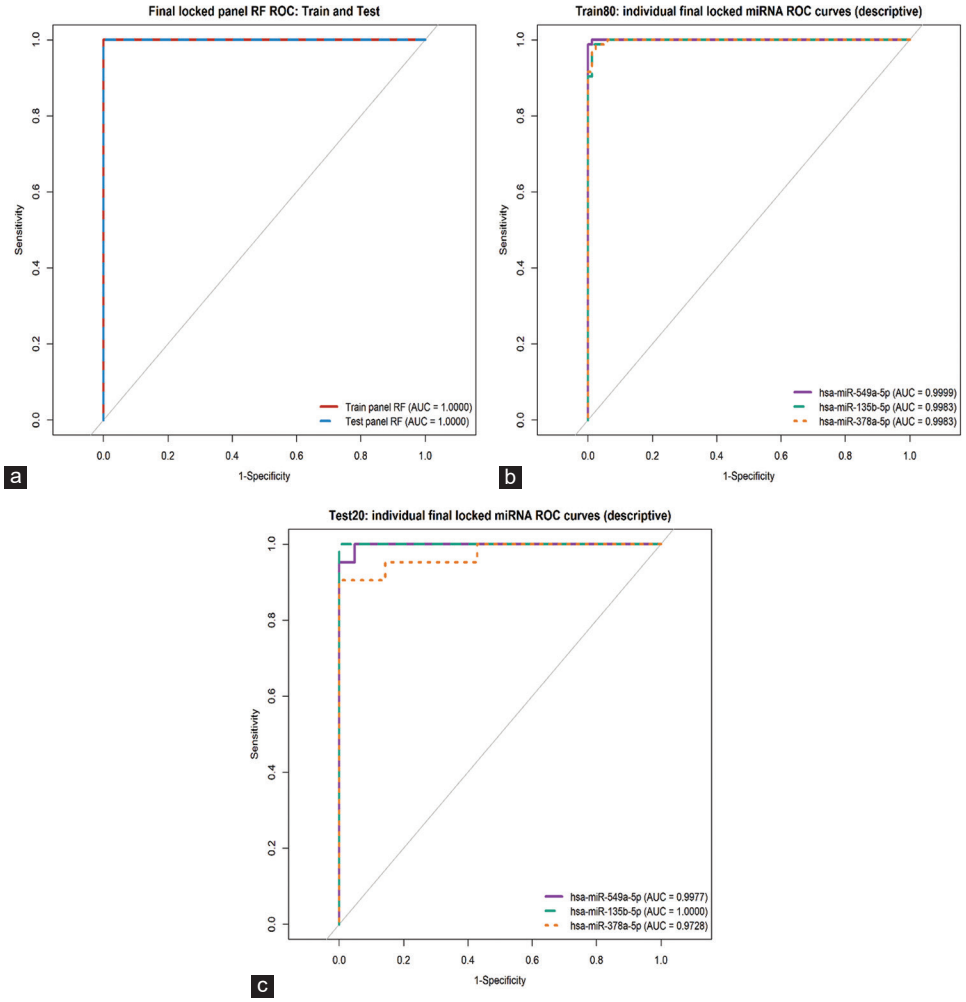


Fig. 5. Receiver operating characteristic (ROC) curves of the final locked three-microRNA (miRNA) signature in the large discovery cohort. (a) Locked random forest panel in Train80 and Test20 (area under the curve [AUC] = 1.000 for both). (b) Individual locked miRNA ROC curves in Train80. (c) Individual locked miRNA ROC curves in Test20.

TABLE IV
LOCKED-PANEL PERFORMANCE ACROSS POOLED EXTERNAL HTS COHORTS

miRNA	AUC	Accuracy	Sensitivity	Specificity	Precision	F1
miRNAs panel	0.960938	0.90625	0.9375	0.875	0.882353	0.909091
hsa-miR-549a-5p	0.914062	0.78125	0.75	0.8125	0.8	0.774194
hsa-miR-135b-5p	0.980469	0.84375	0.9375	0.75	0.789474	0.857143
hsa-miR-378a-5p	0.949219	0.9375	1	0.875	0.888889	0.941176

miRNAs: microRNAs, AUC: Area under the curve, HTS: High-throughput sequencing

target set contained 393 gene symbols, of which 390 mapped to Entrez identifiers. This set yielded 202 significant GO biological process terms, 37 molecular function terms, 10 cellular component terms, and 24 KEGG pathways. The strongest overlap-target biological process signals included regulation of supramolecular fiber organization, cellular response to peptide hormone stimulus, and the Wnt signaling pathway. Cellular component terms were led by focal adhesion and cell-substrate junction, whereas molecular function terms highlighted chromatin DNA binding, DNA-binding transcription factor binding, and beta-catenin binding. KEGG analysis recovered Hippo signaling, thyroid hormone signaling, cellular senescence, regulation of the actin cytoskeleton, TGF-β signaling, and the CRC pathway (Fig. 8). Thus, 401 refers to overlap-supported interactions, 393 refers to unique gene symbols, and 390 refers to genes successfully mapped to Entrez identifiers.

IV. DISCUSSION

The present study identified a locked three-miRNA signature composed of hsa-miR-135b-5p, hsa-miR-378a-5p, and hsa-miR-549a-5p that discriminated CRC tumor tissues from matched normal tissues. Methodologically, the work is strengthened by pair-aware partitioning, repeated cross-validation, and multi-algorithm feature selection because it is

http://dx.doi.org/10.14500/aro.13036

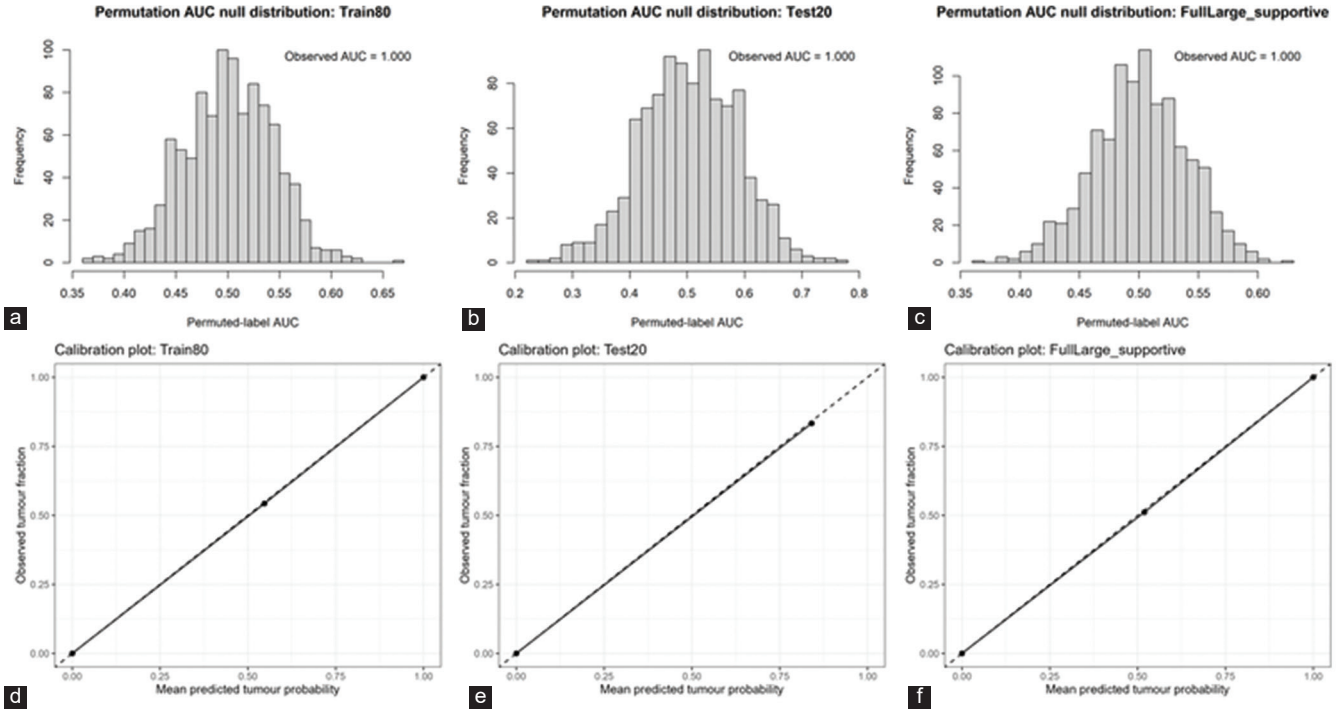


Fig. 6. Permutation and calibration diagnostics for the locked three-miRNA random forest panel. (a-c) Permutation-based null area under the curve (AUC) distributions for Train80, Test20, and the supportive full-large analysis, respectively. The observed locked-model AUC was 1.000, whereas permuted-label AUC distributions were centered near 0.50. (d-f) Calibration plots for Train80, Test20, and the supportive full-large analysis, respectively, showing agreement between mean predicted tumor probability and observed tumor fraction.

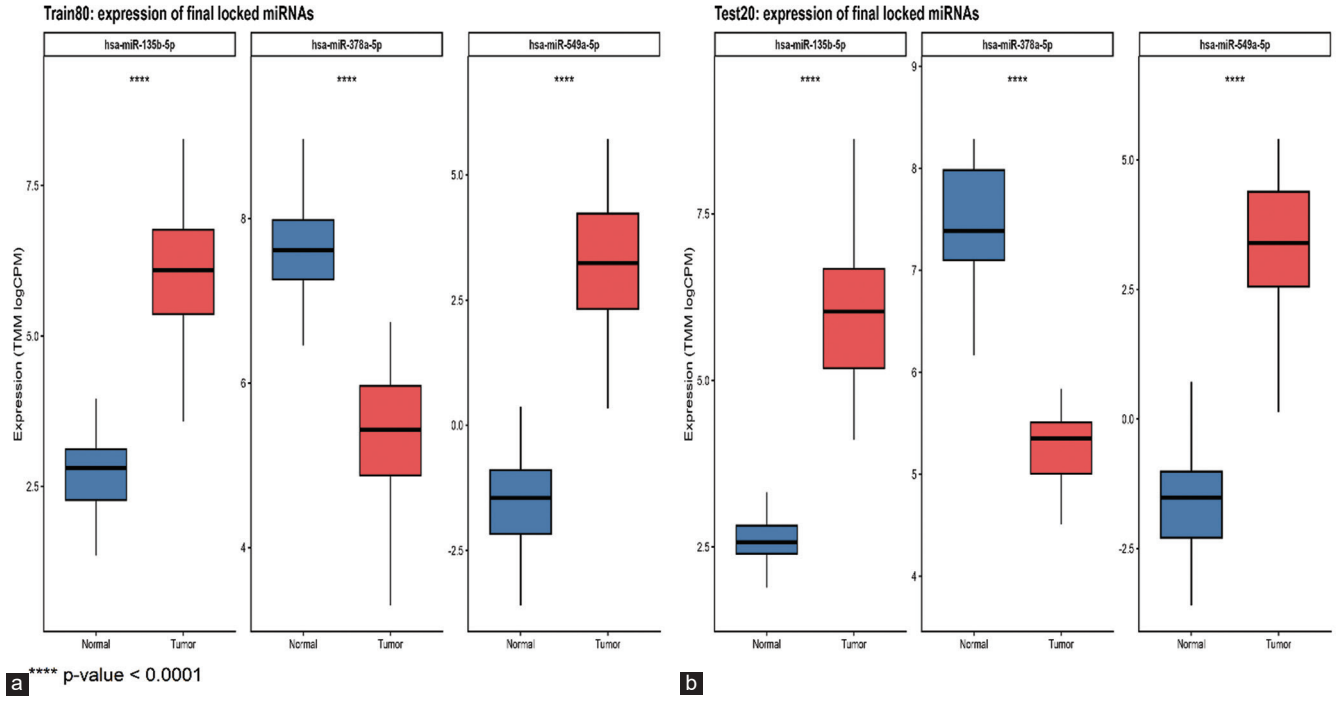


Fig. 7. Expression boxplots of the final locked microRNAs (miRNAs) in the discovery partitions. (a) Train80 boxplots of the locked miRNAs. (b) Test20 boxplots of the locked miRNAs.

a critical determinant of generalizability in high-dimensional biomedical models (Pudjihartono, et al., 2022, Wang, et al., 2025). The workflow also mentioned an important limitation of many diagnostic machine-learning studies by applying the locked panel without retraining on independent datasets, consistent with present recommendations that external testing is essential for evaluating model robustness (Beam and Kohane, 2018, Roberts, et al., 2021, DeGroat, et al., 2024).

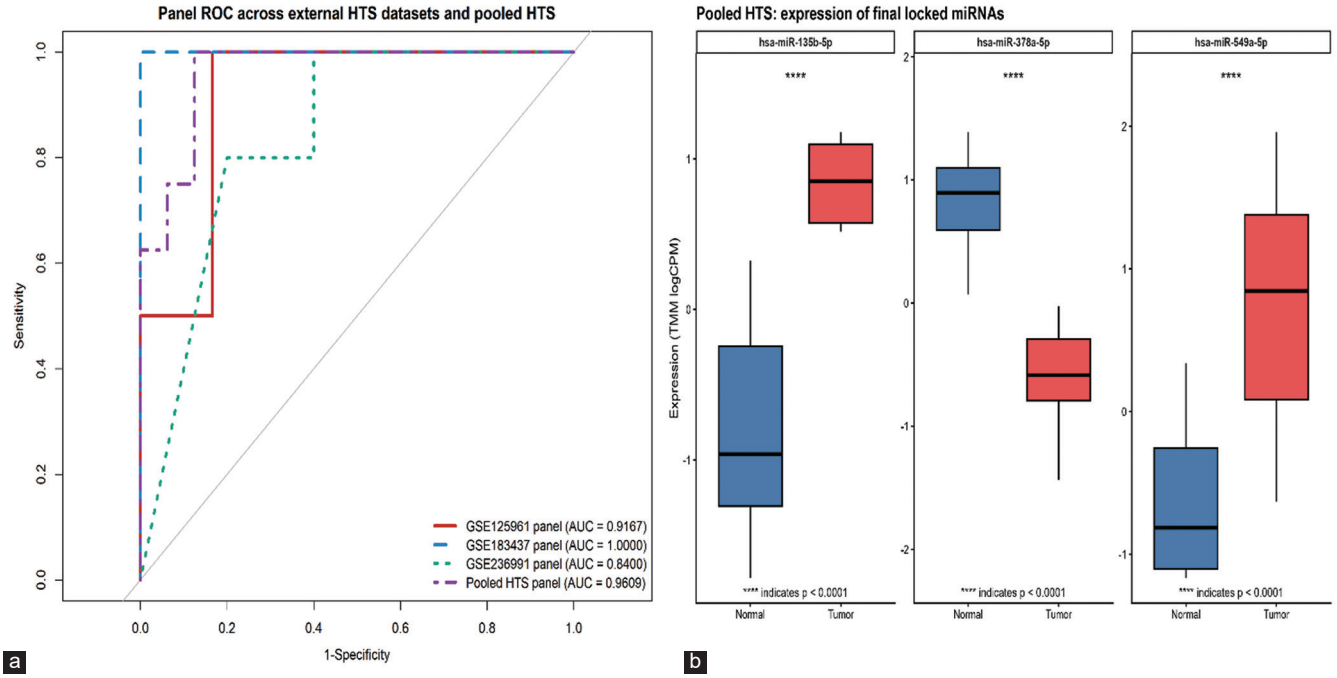


Fig. 8. External validation of the locked signature. (a) Panel receiver operating characteristic curves across the three external high-throughput sequencing (HTS) datasets and the pooled HTS cohort. (b) pooled HTS boxplots for the three locked microRNAs.

Importantly, the initial CPM filtering and TMM/logCPM normalization were used as unsupervised expression-space pre-processing, whereas DE analysis, feature selection, classifier training, and threshold-based assessment were restricted to training folds.

The perfect Train80/Test20 performance should nevertheless be interpreted cautiously. The observed AUC values of 1.000 indicate complete separation in the available discovery partitions, but they are not presented as proof of clinical readiness. To address concerns regarding potential overfitting and residual leakage, confidence intervals, calibration, Brier-score, and permutation diagnostics were incorporated into the Results (Supplementary Table S1 and Fig. 6). The permutation analyses supported the non-random nature of the locked predictions by placing the observed AUC outside null distributions centered near chance, while the calibration plots and very low Brier scores indicated internally consistent probability estimates. These findings strengthen the discovery-level robustness of the locked panel, particularly when interpreted together with the pair-aware split and the restriction of outcome-dependent procedures to Train80 or training folds. Nevertheless, these diagnostics do not replace larger independent validation and do not justify immediate clinical claims. The external cohorts were therefore framed as preliminary because public datasets that simultaneously provide paired CRC tissue design, small RNA-seq/miRNA-seq data, and compatible mature-miRNA quantification remain scarce. The pooled external analysis was used only to increase exploratory sample size, and future work should aggregate additional compatible paired tissue cohorts with within-cohort z-score standardization before pooling.

Among the retained markers, hsa-miR-135b-5p is the most biologically established. Its upregulation is consistent with evidence describing miR-135b as an oncogenic miRNA associated with cancer growth, invasion, angiogenesis, and therapy resistance (Shao, et al., 2025). The downregulation of hsa-miR-378a-5p is compatible with a tumor-suppressive interpretation and is supported by direct CRC evidence. Previous studies reported reduced miR-378a-5p/miR-378-5p expression in cell lines and CRC tissues, and functional experiments revealed that it suppresses CRC cell proliferation and promotes apoptosis through cancer-relevant targets, including BRAF and CDK1 (Wang, et al., 2015, Li, et al., 2021). Therefore, the reduced expression of hsa-miR-378a-5p in the present locked panel is biologically plausible, although further validation in independent paired CRC tissue cohorts remains necessary.

The interpretation of hsa-miR-549a-5p should remain conservative. In the present study, its high selection frequency and consistent tumor upregulation support its diagnostic contribution within the locked panel. Although recent CRC biomarker work has reported miR-549a as a candidate marker, tissue-specific mechanistic evidence for hsa-miR-549a-5p in CRC remains limited. Therefore, hsa-miR-549a-5p is best interpreted as an underexplored diagnostic candidate requiring targeted CRC-tissue validation, rather than as a mechanistically established CRC driver (Akbar, et al., 2024). The positive association between hsa-miR-135b-5p and hsa-miR-549a-5p, together with their inverse relationship with hsa-miR-378a-5p, is also consistent with the concept that miRNAs operate within coordinated regulatory networks rather than isolated one-gene switches (Chang and Xia, 2022).

The target and enrichment analyses provide biological plausibility for the locked panel. Overlap-supported targets were enriched in pathways central to colorectal tumorigenesis,

http://dx.doi.org/10.14500/aro.13036

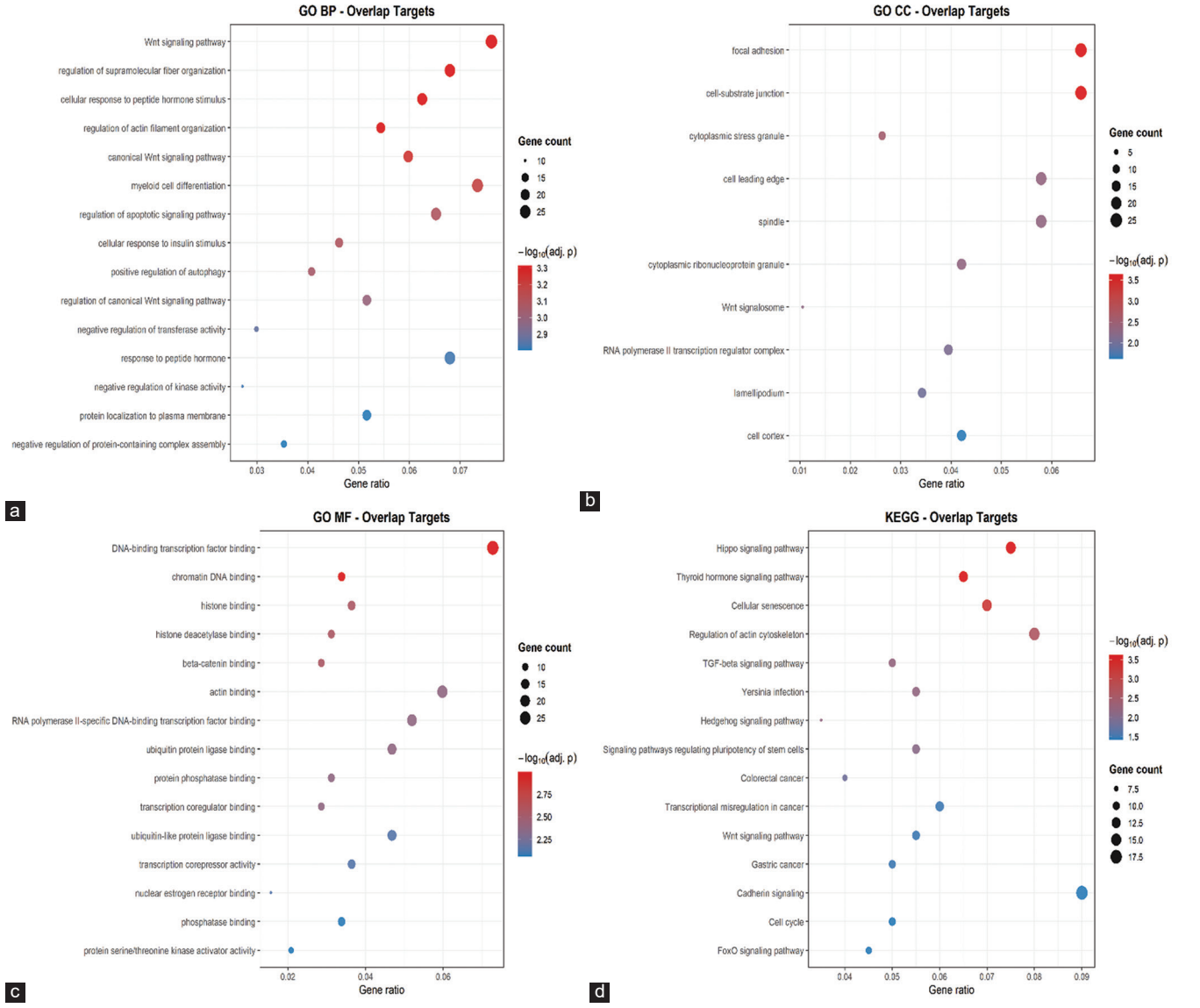


Fig. 9. Overlap-supported target genes Functional enrichment analysis. (a) Gene ontology (GO) biological process enrichment of overlap-supported targets. (b) GO cellular component enrichment. (c) GO molecular function enrichment. (d) Kyoto Encyclopedia of Genes and Genomes pathway enrichment.

including Wnt signaling, Hippo signaling, TGF-β-related programs, focal adhesion, actin-cytoskeleton regulation, and CRC pathway annotations (Li, et al., 2024, Tejeda-Muñoz and Mei, 2024, Fasano, et al., 2024, Mohammadpour, et al., 2024, Saha, et al., 2024). These pathways are linked to tumor initiation, invasion, stemness, microenvironmental interaction, and therapeutic response. The recovery of shared-core genes, including NUCKS1 and Metadherin (MTDH), further supports the biological coherence of the network, because these genes have been associated with CRC invasion, metastasis, progression, and stemness-related phenotypes (Zhu, et al., 2025, Todorović and Amedei, 2024). However, database-supported target enrichment remains hypothesis-generating and does not prove direct regulation by the selected miRNAs in CRC tissues.

From a translational perspective, a three-miRNA tissue signature is attractive because compact panels are easier to standardize and may be more feasible for qPCR, digital PCR, or targeted assays. Still, this study used tissue RNA-seq data only; any extension to liquid biopsy or circulating-miRNA testing should be considered a future hypothesis rather than a conclusion of the present work (Ma, et al., 2024, Naranbat, et al., 2025). Finally, the findings support a biologically coherent candidate diagnostic model in which hsa-miR-135b-5p and hsa-miR-549a-5p mark tumor-associated activation, whereas hsa-miR-378a-5p reflects loss of tumor-suppressive regulation. Further prospective, assay-level, and functional validation is required before clinical translation.

V. Conclusion

This study identified a stable and biologically interpretable three-miRNA signature, hsa-miR-135b-5p, hsa-miR-378a-5p, and hsa-miR-549a-5p, that discriminated CRC tissues

from matched normal controls with strong discovery performance and preliminary support from independent external high-throughput sequencing cohorts. By combining leakage-aware preprocessing, repeated cross-validated consensus feature selection, and unchanged external testing, the study established a reproducible framework for miRNA biomarker discovery that extends beyond single-pass model optimization. The preserved expression direction of the locked miRNAs across cohorts, together with enrichment of overlap-supported targets in canonical CRC-related pathways, such as Wnt, Hippo, TGF-β, focal adhesion, and CRC signaling, further supports the biological relevance of the identified panel. Although additional prospective, assay-level, and functional validation remains necessary, these findings position the three-miRNA set as a candidate diagnostic biomarker panel with providing a reproducible analytical foundation for future translational studies in CRC.

A. Data and Code Availability

All analyzed raw sequencing datasets are publicly available from the NCBI SRA/GEO repositories under the accession numbers PRJNA387172, GSE125961, GSE183437, and GSE236991. The principal R scripts and Ubuntu commands are provided in the Supplementary zip file.

V1. Acknowledgment

The authors acknowledge that this work has been carried out in part in accordance with the requirements for Rebaz M. Mustafa's doctoral degree at the University of Koya. The authors express their sincere gratitude to the Department of Chemistry of the University of Koya for their continuing academic support and the provision of the facilities necessary for this research. The authors also express their sincere appreciation to the Pasteur Institute of Iran for its valuable scientific collaboration during the course of this work.

References

Akbar, S., Mashreghi, S., Kalani, M.R., Valanik, A., Ahmadi, F., Aalikhani, M. and Bazi, Z., 2024. Blood miRNAs miR-549a, miR-552, and miR-592 serve as potential disease-specific panels to diagnose colorectal cancer. *Heliyon*, 10, p. e28492.

Beam, A.L., and Kohane, I.S., 2018. Big data and machine learning in health care. *JAMA,* 319**,** pp.1317-1318.

Chang, L.M., and Xia, J., 2022. MicroRNA regulatory network analysis using miRNet 2.0. In: *Transcription Factor Regulatory Networks.* Springer, Berlin.

Chen, Y., Lun, A.T., and Smyth, G.K., 2016. From reads to genes to pathways: Differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Research,* 5**,** p.1438.

Degroat, W., Abdelhalim, H., Peker, E., Sheth, N., Narayanan, R., Zeeshan, S., Liang, B.T., and Ahmed, Z., 2024. Multimodal AI/ML for discovering novel biomarkers and predicting disease using multi-omics profiles of patients with cardiovascular diseases. *Scientific Reports,* 14**,** p.26503.

Fasano, M., Pirozzi, M., Miceli, C.C., Cocule, M., Caraglia, M., Boccellino, M., Vitale, P., De Falco, V., Farese, S., Zotta, A., Ciardiello, F., and Addeo, R., 2024. TGF-β modulated pathways in colorectal cancer: New potential therapeutic opportunities. *International Journal of Molecular Sciences,* 25**,** p.7400.

Hozaka, Y., Kita, Y., Yasudome, R., Tanaka, T., Wada, M., Idichi, T., Tanabe, K., Asai, S., Moriya, S., and Toda, H., 2021. RNA-sequencing based microRNA expression signature of colorectal cancer: The impact of oncogenic targets regulated by miR-490-3p. *International Journal of Molecular Sciences,* 22**,** p.9876.

Khan, A.U., Ali, M., and Wahab, M.A., 2025. Diagnostic potential of cross-specimen microRNA panels as biomarkers for colorectal cancer: A systematic review and meta-analysis. *Molecular Diagnosis and Therapy,* 29**,** pp.707-719.

Law, C.W., Chen, Y., Shi, W., and Smyth, G.K., 2014. Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology,* 15**,** p.R29.

Li, K., Zhang, J., Zhang, M., Wu, Y., Lu, X., and Zhu, Y., 2021. miR-378a-5p inhibits the proliferation of colorectal cancer cells by downregulating CDK1. *World Journal of Surgical Oncology,* 19**,** p.54.

Li, Q., Geng, S., Luo, H., Wang, W., Mo, Y.Q., Luo, Q., Wang, L., Song, G.B., Sheng, J.P., and Xu, B. 2024. Signaling pathways involved in colorectal cancer: Pathogenesis and targeted therapy. *Signal Transduction and Targeted Therapy,* 9**,** p.266.

Ma, L., Guo, H., Zhao, Y., Liu, Z., Wang, C., Bu, J., Sun, T., and Wei, J. 2024. Liquid biopsy in cancer current: Status, challenges and future prospects. *Signal Transduction and Targeted Therapy,* 9**,** p.336.

Mohammadpour, S., Torshizi Esfahani, A., Sarpash, S., Vakili, F., Zafarjafarzadeh, N., Mashaollahi, A., Pardakhtchi, A., and Nazemalhosseini-Mojarad, E., 2024. Hippo signaling pathway in colorectal cancer: Modulation by various signals and therapeutic potential. *Analytical Cellular Pathology,* 2024**,** p.5767535.

Moiz, S., Saha, B., Mondal, V., Bishnu, D., Das, B., Bose, B., Das, S., Banerjee, N., Dutta, A., Chatterjee, K., Goswami, S., Mukhopadhyay, S., and Basu, S., 2025. Differential expression of miRNAs between young-onset and late-onset Indian colorectal carcinoma patients. *Non-Coding RNA,* 11**,** p.10.

Naranbat, D., Herdes, E., Tapinos, N., and Tripathi, A., 2025. Review of microRNA detection workflows from liquid biopsy for disease diagnostics. *Expert Reviews in Molecular Medicine***,** pp.1-41.

Pająk, W., Kleinrok, J., Pec, J., Michno, K., Wojtas, J., Badach, M., Teresińska, B., and Baj, J., 2025. Micro RNA in colorectal cancer-potential diagnostic and prognostic markers-an updated review. *International Journal of Molecular Sciences,* 26**,** p.8615.

Pudjihartono, N., Fadason, T., Kempa-Liehr, A.W., and O'Sullivan, J.M., 2022. A review of feature selection methods for machine learning-based disease risk prediction. *Frontiers in Bioinformatics,* 2**,** p.927312.

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A.I., Etmann, C., Mccague, C., Beer, L., Weir-McCall, J.R., Teng, Z., Gkrania-Klotsas, E., Rudd, J.H., Sala, E., Schönlieb, C.B., and Aix-Covnet., 2021. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence,* 3**,** pp.199-217.

Robinson, M.D., and Oshlack, A., 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology,* 11**,** p.R25.

Saha, S., Ghosh, S., Ghosh, S., Nandi, S., and Nayak, A., 2024. Unraveling the complexities of colorectal cancer and its promising therapies - an updated review. *International Immunopharmacology,* 143**,** p.113325.

Schwab, S., and Nonaka, T., 2025. Circulating miRNAs as liquid biopsy biomarkers for diagnosis in patients with colorectal cancer: A systematic review and meta-analysis. *Frontiers in Genetics,* 16**,** p.1574586.

Shao, Y., Zhang, S., Pan, Y., Peng, Z., and Dong, Y., 2025. miR-135b: A key role in cancer biology and therapeutic targets. *Non-Coding RNA Research,* 12**,** pp.67-80.

Shirzad, S., Eterafi, M., Karimi, Z., and Barazesh, M., 2025. MicroRNAs involved in colorectal cancer, a rapid mini-systematic review. *BMC Cancer,* 25**,** p.934.

Siegel, R.L., Giaquinto, A.N., and Jemal, A., 2024. Cancer statistics, 2024. *CA: A Cancer Journal for Clinicians,* 74**,** pp.12-49.

Tejeda-Muñoz, N., and Mei, K.C., 2024. Wnt signaling in cell adhesion, development, and colon cancer. *IUBMB Life,* 76**,** pp.383-396.

Todorović, N., and Amedei, A., 2024. Metadherin-driven promotion of cancer stem cell phenotypes and its effect on immunity in hepatocellular carcinoma. *World Journal of Gastroenterology,* 30**,** pp.2624-2628.

Wang, J., Zhang, Z., and Wang, Y., 2025. Utilizing feature selection techniques for AI-driven tumor subtype classification: Enhancing precision in cancer diagnostics. *Biomolecules,* 15**,** p.81.

Wang, Z., Ma, B., Ji, X., Deng, Y., Zhang, T., Zhang, X., Gao, H., Sun, H., Wu, H., and Chen, X., 2015. MicroRNA-378-5p suppresses cell proliferation and induces apoptosis in colorectal cancer cells by targeting BRAF. *Cancer Cell International,* 15**,** p.40.

Włodarczyk, M., Maryńczak, K., Burzyński, J., Włodarczyk, J., Basak, J., Fichna, J., Majsterek, I., Ciesielski, P., Spinelli, A., and Dziki, Ł., 2025. The role of miRNAs in the pathogenesis, diagnosis, and treatment of colorectal cancer and colitis-associated cancer. *Clinical and Experimental Medicine,* 25**,** p.86.

Wu, S.M., Tsai, W.S., Chiang, S.F., Lai, Y.H., Ma, C.P., Wang, J.H., Lin, J., Lu, P.S., Yang, C.Y., and Tan, B.C.M., 2020. Comprehensive transcriptome profiling of Taiwanese colorectal cancer implicates an ethnic basis for pathogenesis. *Scientific Reports,* 10**,** p.4526.

Zhou, F., Tang, D., Xu, Y., He, H., Wu, Y., Lin, L., Dong, J., Tan, W., and Dai, Y., 2020. Identification of microRNAs and their endonucleolytic cleavaged target mRNAs in colorectal cancer. *BMC Cancer,* 20**,** p.242.

Zhu, L., Zhao, T., Su, H., Li, J., Shen, X., Zhang, L., Chen, J., and Song, Y., 2025. NUCKS1 promotes invasion and metastasis of colorectal cancer by stabilizing HDAC2 and activating AKT. *Oncogenesis,* 14**,** p.19.

Supplementary Material

SUPPLEMENTARY TABLE S1
Uncertainty-Aware Performance Assessment Of The Locked Three-Mirna Random Forest Panel In The Discovery-Based Analyses

Dataset	AUC (95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Precision (95% CI)	Brier	Permutation P
Train80	1.000 (1.000-1.000)	1.000 (0.978-1.000)	1.000 (0.957-1.000)	1.000 (0.957-1.000)	1.000 (0.957-1.000)	0.0008	$9.99e^{-04}$
Test20	1.000 (1.000-1.000)	1.000 (0.916-1.000)	1.000 (0.839-1.000)	1.000 (0.839-1.000)	1.000 (0.839-1.000)	0.0082	$9.99e^{-04}$
Full large supportive	1.000 (1.000-1.000)	1.000 (0.982-1.000)	1.000 (0.965-1.000)	1.000 (0.965-1.000)	1.000 (0.965-1.000)	0.0023	$9.99e^{-04}$

AUC confidence intervals were estimated using DeLong's method. Classification metric confidence intervals are exact binomial 95% intervals. Permutation P values were derived from 1,000 label permutations while keeping locked model probabilities fixed

SUPPLEMENTARY TABLE S2
Exact Binomial Confidence Intervals For Pooled External Locked-Panel Classification Metrics

Metric	Numerator/denominator	Estimate	Exact 95% CI
Accuracy	29/32	0.906	0.750-0.980
Sensitivity	15/16	0.938	0.698-0.998
Specificity	14/16	0.875	0.617-0.984
Precision	15/17	0.882	0.636-0.985