

Machine learning and single-cell transcriptomics identify key lncRNAs associated with malignant transformation and immune infiltration in colorectal cancer

Han Gong

The Second Affiliated Hospital of Chengde Medical University (Chengde Central Hospital)

LiBo Liu

The 981st Hospital of the Joint Logistics Support Force of the Chinese People's Liberation Army

Chao Mi

The Second Affiliated Hospital of Chengde Medical University (Chengde Central Hospital)

BoLun Yan

The Second Affiliated Hospital of Chengde Medical University (Chengde Central Hospital)

FuYang Gao

chaichanglaotou@163.com

The Second Affiliated Hospital of Chengde Medical University (Chengde Central Hospital)

HuaYu Zhang

The Second Affiliated Hospital of Chengde Medical University (Chengde Central Hospital)

Research Article

Keywords: Colorectal cancer, Long non-coding RNA, Machine learning, Single-cell RNA sequencing, Immune infiltration

Posted Date: October 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11203434/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Machine learning and single-cell transcriptomics identify key lncRNAs associated with malignant transformation and immune infiltration in colorectal cancer

Han Gong¹, Libo Liu², Chao Mi¹, Bolun Yan¹, Fuyang Gao^{1*}, Huayu Zhang^{1*}

¹ Department of General Surgery, The Second Affiliated Hospital of Chengde Medical University (Chengde Central Hospital), Chengde 067000, Hebei, People's Republic of China

² Department of General Surgery, The 981st Hospital of the Joint Logistics Support Force of the Chinese People's Liberation Army, Chengde 067000, Hebei, People's Republic of China

* Correspondence: Fuyang Gao (E-mail: chaichanglaotou@163.com) and Huayu Zhang (E-mail: zhanghuayudoctor@163.com).

Background: Colorectal cancer (CRC) is generally thought to arise through the normal–adenoma–tumor sequence. Which long non-coding RNAs (lncRNAs) change at each step, and whether these coincide with immune-cell infiltration, remain incompletely characterized.

Methods: Three microarray datasets with normal, adenoma and tumor samples (GSE37364, GSE100179, GSE117606) were analyzed. Seven algorithms were combined into 120 models to form a two-layer classifier separating normal from diseased tissue and adenoma from tumor. Immune-cell fractions were estimated with CIBERSORT; candidates were examined in single-cell data (GSE201349; 46,562 cells) and for progression-free survival (PFS) in 357 patients from The Cancer Genome Atlas (TCGA).

Results: Seven lncRNAs were identified: UCA1, RTEL1-TNFRSF6B, LOC102724156, GATA2-AS1, SATB2-AS1, H19 and BLOC1S1-RDH5. First-layer accuracy, area under the curve (AUC) and F-score were 0.972, 0.979 and 0.978; second-layer values were 0.958, 0.958 and 0.955. M0 macrophages and activated dendritic cells were enriched in diseased tissue and correlated with five lncRNAs (all $P < 0.05$). Single-cell analysis localized GATA2-AS1 and SATB2-AS1 to stem-like epithelial cells with the highest activity in adenoma. In the TCGA cohort, expression tertile was not associated with PFS for any of the six lncRNAs measurable there; only H19 was nominally significant (log-rank $P = 0.013$), which did not remain significant after correction ($q = 0.078$).

Conclusions: These seven lncRNAs form a two-layer classifier separating normal, adenoma and tumor tissue, but none of the six measurable there was associated with PFS after correction. They are candidate tissue biomarkers for staging along the normal–adenoma–tumor sequence rather than markers of prognosis.

Keywords: Colorectal cancer; Long non-coding RNA; Machine learning; Single-cell RNA sequencing; Immune infiltration

1. Introduction

Colorectal cancer (CRC) is the third most commonly diagnosed cancer and the second leading cause of cancer-related death worldwide, with approximately 1.9 million new cases and 0.9 million cancer deaths in 2022 [1]. Most CRC cases arise through the normal–adenoma–tumor sequence, a multistep process in which the progressive accumulation of genetic and epigenetic alterations drives the malignant transformation of colorectal epithelial cells [2]. Although colonoscopy-based screening combined with endoscopic resection of adenomatous lesions has substantially reduced CRC mortality, a considerable proportion of patients with CRC are still diagnosed at an advanced stage. In addition, the sensitivity of current non-invasive screening modalities for early-stage tumors and premalignant adenomas remains suboptimal [3]. Therefore, characterizing the molecular events underlying the normal–adenoma–tumor transition and identifying reliable biomarkers for the early detection of malignant transformation are of substantial clinical significance. Such efforts may enable risk stratification of adenomas and guide individualized endoscopic surveillance, potentially facilitating earlier diagnosis and timely intervention in CRC.

Accumulating mechanistic and translational studies have gradually uncovered the critical regulatory roles of long non-coding RNAs (lncRNAs) in colorectal tumorigenesis [4,5]. Dysregulated lncRNAs act as core signaling modulators governing diverse oncogenic and immune-related pathways during CRC malignant progression [6]. By modulating cancer stem-cell properties, epithelial – mesenchymal transition (EMT) activation, and other oncogenic processes, lncRNAs exert profound effects on the initiation, invasion, and metastasis of CRC, functioning most prominently through the lncRNA-miRNA-mRNA competing endogenous RNA (ceRNA) axis, the most extensively characterized post-transcriptional regulatory paradigm in CRC [7]. Mechanistically, lncRNAs harboring multiple miRNA response elements serve as molecular sponges that sequester abundant

miRNAs, thereby relieving miRNA-mediated repression of their shared target mRNAs. Beyond miRNA sponging, lncRNAs also regulate gene expression at multiple levels through chromatin remodeling, transcriptional regulation, and scaffolding of protein complexes [8]. For example, the oncogenic lncRNA H19 sequesters miR-200a, relieving its repression of β -catenin and thereby activating the Wnt signaling cascade that drives CRC cell proliferation; in addition to its ceRNA-sponging role, H19 gives rise to miR-675 from its first exon, which promotes cell-cycle progression by repressing the RB1-E2F axis [9]. Moreover, UCA1 facilitates CRC proliferation and metastatic phenotypes through the miR-143/MYO6 signaling axis, and circulating exosomal UCA1 further augments the malignant capacity of recipient CRC cells via this ceRNA cascade, collectively constituting the complex lncRNA regulatory network of CRC [10]. Such mechanistic investigations of individual lncRNAs, however, have largely been restricted to pairwise comparisons between normal colonic mucosa and established tumor tissues. Most existing studies on CRC-associated lncRNAs center on defining their oncogenic or tumor-suppressive functions in frank malignant lesions, whereas lncRNA-driven molecular events that underpin the normal – adenoma – tumor transition and mediate concomitant remodeling of the tumor immune microenvironment (TIME) remain incompletely characterized. Therefore, it is necessary to systematically screen stage-specific lncRNA candidates associated with both the normal – adenoma – tumor transition and immune microenvironment remodeling across the full spectrum of CRC progression. Such signatures would provide novel and reliable candidate biomarkers for early detection of adenoma progression toward invasive tumor, as well as potential molecular clues for precision immunotherapy in CRC.

Traditional bulk transcriptomic profiling provides an averaged signal across all cells in a specimen and does not resolve cell-type-specific expression heterogeneity, which limits the accurate screening and functional annotation of core tumor-regulatory lncRNAs across the normal – adenoma – tumor sequence. Single-cell RNA sequencing (scRNA-seq), as a single-cell resolution transcriptomic tool, enables classification of diverse cell subpopulations in tissues and characterization of lncRNA expression profiles in discrete cell types [11]. This approach provides a cellular-level analytical framework for our study to explore lncRNA signatures related to normal – adenoma – tumor progression and immune microenvironment remodeling. In parallel, machine learning (ML) algorithms have become a mainstream approach for deciphering high-dimensional lncRNA transcriptomic data; representative

frameworks integrating multi-omics data with neural network-based information aggregation have enabled accurate prediction of cancer-related lncRNAs [12]. Multiple well-validated ML models also exhibit superior performance in filtering noisy data, screening hub lncRNA signatures, and constructing predictive models; for instance, an integrative ML procedure fitting 101 candidate models across 17 independent CRC cohorts developed a consensus immune-related lncRNA signature that outperformed traditional clinical variables and 109 previously published transcriptome signatures in predicting patient outcomes [13]. Lasso regression and ensemble XGBoost models have been widely applied to screen prognostic and tumorigenic lncRNA biomarkers from cancer transcriptome datasets [14]; for example, Lasso-Cox regression identified a six-lncRNA autophagy-related prognostic signature in CRC that was further validated by qRT-PCR in clinical specimens [15], and an XGBoost-based classifier integrating SMOTE over-sampling and Laplacian score-based feature selection identified 439 candidate cancer-related lncRNAs with high accuracy [16]. Beyond biomarker screening, support vector machines and other machine-learning algorithms have also achieved high accuracy in distinguishing cellular states such as cellular senescence from bulk transcriptome data [17], indicating the broad applicability of ML-based classification beyond conventional differential expression analysis. Collectively, the integrative analysis of scRNA-seq transcriptomic data and multi-algorithm machine learning modeling provides a reliable and efficient analytical strategy for the systematic screening and validation of stage-specific immune-associated lncRNA biomarkers across the normal – adenoma – tumor transition.

We hypothesized that a subset of lncRNAs changes progressively across the normal–adenoma–tumor sequence and that these changes are coupled to remodeling of the innate immune compartment. We therefore aimed to (i) identify lncRNAs that are consistently differentially expressed across three independent datasets spanning the normal–adenoma–tumor sequence; (ii) assemble and evaluate a two-layer machine-learning classifier that discriminates normal from diseased tissue and adenoma from tumor tissue; (iii) localize the candidates at single-cell resolution and relate them to immune-cell infiltration; and (iv) explore, in an independent cohort from The Cancer Genome Atlas (TCGA), whether their expression is associated with progression-free survival.

2. Materials and methods

2.1. Data collection

Microarray expression data were retrieved from the Gene Expression Omnibus (GEO) repository (<https://www.ncbi.nlm.nih.gov/geo/>), a MIAME-compliant database hosting array-based and sequencing-based transcriptomic profiles [18]. Sufficient sample sizes across the normal control, colonic adenoma, and colonic tumor groups were essential to guarantee stable discriminatory performance in subsequent classification modeling. Accordingly, three independent datasets (GSE37364, GSE100179, and GSE117606) with adequate adenoma and tumor specimens were selected for integrated analysis. The sample composition of each dataset is detailed as follows: GSE37364 contained 38 normal, 29 adenoma, and 14 tumor samples; an additional 13 stage-annotated colorectal cancer samples were present but not assigned to the three comparison groups. GSE100179 included 20 normal, 20 adenoma, and 20 tumor samples, whereas GSE117606 contained 65 normal, 59 adenoma, and 74 tumor samples. All bioinformatics pipelines—including raw data preprocessing, differential expression screening, and functional and pathway enrichment analyses—were implemented in R (version 4.3.0) within the RStudio environment. GSE37364 was profiled on the Affymetrix Human Genome U133 Plus 2.0 Array (GPL570), GSE100179 on the Affymetrix Human Transcriptome Array 2.0 (GPL17586) and GSE117606 on the Affymetrix HT HG-U133+ PM Array Plate (GPL25373). The single-cell dataset GSE201349 (Illumina NovaSeq 6000; GPL20301 and GPL24676) and the TCGA colorectal adenocarcinoma (COAD and READ) RNA-sequencing and clinical data used for the survival analysis are described in sections 2.8 and 2.9. The overall workflow is summarized in Fig. 1, and all plots were generated with the ggplot2 package in R.

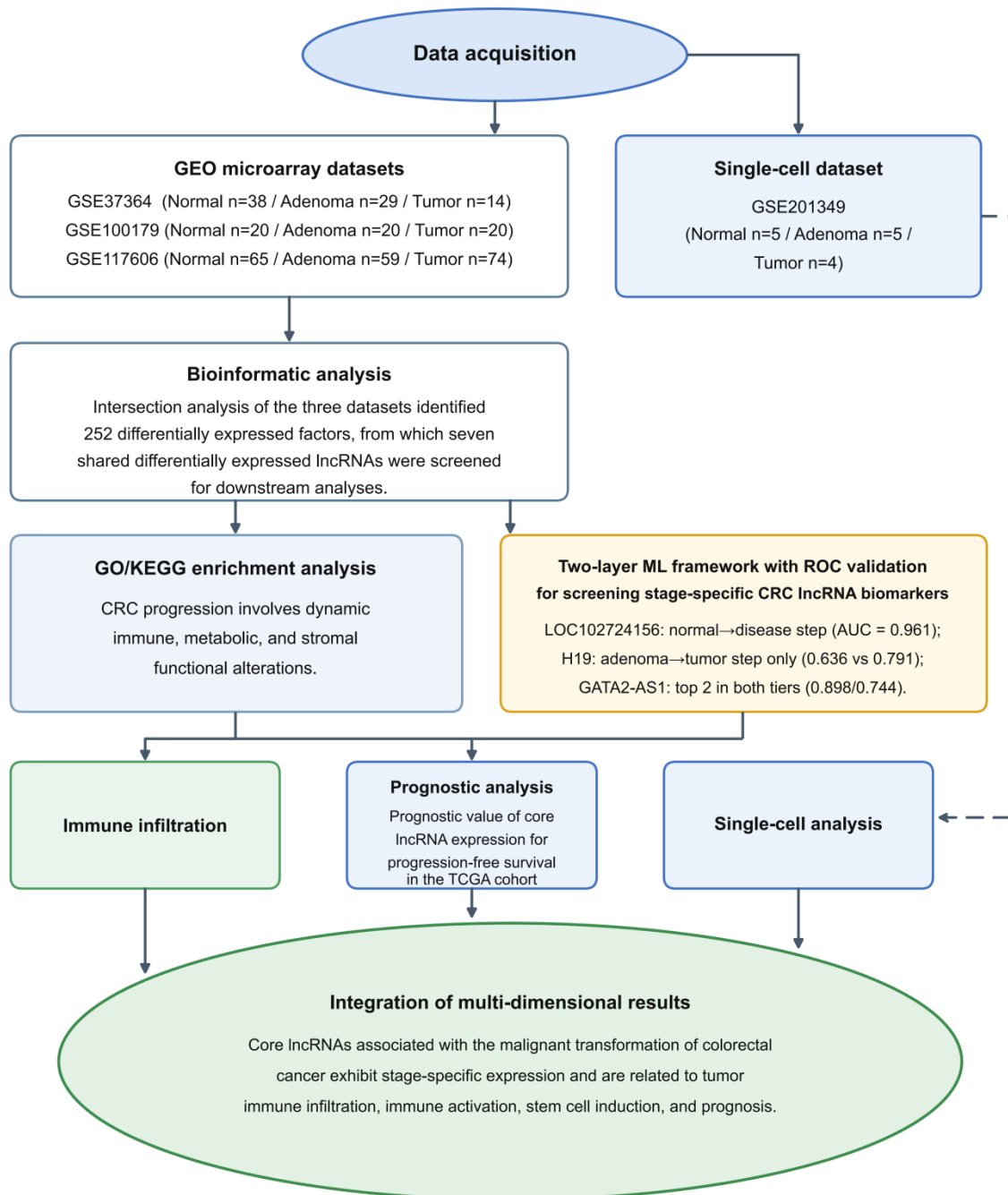


Fig. 1 Comprehensive analysis flowchart.

2.2. Differential expression analysis

The limma package (Linear Models for Microarray Analysis; version 3.58.1; <https://bioconductor.org/packages/release/bioc/html/limma.html>) [19] was applied to screen differentially expressed genes (DEGs) across the three selected microarray datasets. In the limma workflow, a linear model was fitted to the probe-level expression data to generate coefficients corresponding to the design matrix. Instead of conventional t-statistics,

moderated t-statistics, moderated F-statistics, and B-statistics (representing the log-odds of differential expression) were calculated via an empirical Bayes approach that shrinks residual standard errors toward a shared pooled value.

In the present study, DEGs were separately identified in three comparison groups: normal vs. adenoma, normal vs. tumor, and adenoma vs. tumor. The differential expression thresholds were set at false discovery rate (FDR) < 0.05 and $|\log_2 \text{fold change} (\log_2 \text{FC})| > 1$.

2.3. Screening of differentially expressed lncRNAs

Differentially expressed long non-coding RNAs (DE-lncRNAs) were identified as the lncRNA subset of the DEGs meeting FDR < 0.05 and $|\log_2 \text{FC}| > 1$.

2.4. Functional and pathway enrichment analyses

Gene Ontology (GO) functional annotation and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis of the identified DEGs were performed using the R package clusterProfiler (version 4.10.1; <https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html>) [20]. GO terms are classified into three structured ontologies: biological process (BP), cellular component (CC), and molecular function (MF). Both analyses were conducted separately for the DEGs from each of the three pairwise comparisons (normal–adenoma, normal–tumor, and adenoma–tumor). Terms with an adjusted P-value < 0.05 were considered significantly enriched, and all other parameters were kept at the package defaults.

2.5. Two-layer machine learning classification framework

A two-layer machine learning (ML) classification architecture was constructed to discriminate the sequential stages of colorectal malignant transformation. The first layer classified normal versus diseased samples (adenoma and tumor), and the second layer further differentiated adenoma from tumor among the diseased samples. All differentially expressed lncRNAs were used as features in both layers. To quantify the discriminatory capacity of each lncRNA independently, univariate logistic regression was applied to generate receiver operating characteristic (ROC) curves for normal–diseased and adenoma–tumor classification. Seven mainstream algorithms—neural network–multilayer perceptron (NN-MLP), k-nearest neighbors (KNN), linear discriminant analysis (LDA), least absolute shrinkage and selection operator (Lasso), elastic net regression (ENR), support vector machine (SVM) and extreme

gradient boosting (XGBoost)—were combined across their hyperparameter grids to establish the combinatorial ML model space [21–26]. Models yielding valid variable importance scores were retained to quantify the predictive significance of each candidate lncRNA. The modeling strategy of the first layer was fully replicated for the second layer.

2.6. Screening of core lncRNAs based on AUC

A two-layer AUC-based screening strategy was applied to identify core discriminatory lncRNAs. The first layer evaluated discrimination between normal and diseased samples, and the second layer evaluated discrimination between adenoma and carcinoma. lncRNAs with an AUC greater than 0.8 for adenoma–tumor classification were defined as high-performance biomarkers and selected for subsequent validation. ROC curves of all candidate lncRNAs were plotted for both layers to enable direct performance comparison.

2.7. Immune infiltration analysis

Immune cell infiltration patterns were evaluated using the GSE37364 dataset. CIBERSORT, implemented in the R package IOBR (version 2.2.3; <https://cran.r-project.org/package=IOBR>) [27], was applied to estimate the relative fractions of 22 immune cell subsets on the basis of the LM22 signature matrix [28], which is available at <https://cibersort.stanford.edu/>. Differences in these fractions among normal, adenoma and tumor samples were compared with the Kruskal–Wallis test followed by pairwise two-sided Mann–Whitney U tests. Associations between lncRNA expression and immune-cell fractions were assessed with Spearman rank correlation.

2.8. Single-cell transcriptome analysis

To further explore the dynamic expression patterns of key lncRNAs across distinct pathological states (normal colon, premalignant adenoma, and tumor), the single-cell transcriptome dataset GSE201349 was included for subsequent analysis. Low-quality cells were excluded after rigorous quality control, and a total of 46,562 qualified cells were subjected to normalization and log transformation. Variable genes were screened using the FindVariableFeatures function in the Seurat package (version 4.3.0; <https://satijalab.org/seurat/>) [29]. The RunHarmony function of the harmony package (version 0.1.1; <https://github.com/immunogenomics/harmony>) [30] was applied for data integration to

eliminate batch effects and unwanted technical variation. Principal component analysis (PCA) was performed to extract key dimensional features, and the top 30 principal components (PCs) were selected for t-distributed stochastic neighbor embedding (t-SNE) dimensionality reduction and visualization. Cell clustering was implemented via the FindClusters function, and cell type annotation was achieved using the SingleR package (version 2.4.1; <https://bioconductor.org/packages/release/bioc/html/SingleR.html>) [31]. Furthermore, the AUCCell package (version 1.24.0; <https://bioconductor.org/packages/release/bioc/html/AUCCell.html>) [32] was used to quantify the single-cell activity scores of key lncRNAs and evaluate their differential activity levels across diverse cell types.

2.9. Association between core lncRNA expression and progression-free survival

Progression-free survival (PFS) was analyzed in the TCGA colorectal adenocarcinoma cohort (colon adenocarcinoma [COAD] and rectal adenocarcinoma [READ] primary tumors, n = 380; 357 patients evaluable for PFS, 95 events; follow-up capped at 60 months). Only six of the seven candidate lncRNAs were assessable; LOC102724156 (Ensembl ENSG00000287750, 9q22.1, HGNC:56164) post-dates every standardized TCGA pipeline (GENCODE v22/GDC, v23/UCSC Xena Toil, v29/Recount3) and therefore has no independent lncRNA model and no measurable transcript in any public TCGA expression matrix; it can be interrogated only on the Affymetrix HG-U133 Plus 2 array (probe 230991_at), which retains the legacy Entrez nomenclature. For each remaining lncRNA, patients were ranked by tumor expression (log₂-transformed transcripts per million [TPM], UCSC Xena Toil pipeline) and the upper versus lower expression tertiles were compared, with the middle tertile excluded (n = 238 per lncRNA, except BLOC1S1-RDH5, n = 241, owing to ties at the tertile boundaries). Survival probabilities were estimated by the Kaplan–Meier method and compared with the two-sided log-rank test; hazard ratios (HR) with 95% confidence intervals were derived from univariable Cox proportional hazards regression, and the proportional hazards assumption was verified with Schoenfeld residuals. Dose–response across tertiles was tested by Cox regression on an ordinal tertile score in the full cohort (P for trend), and Benjamini–Hochberg false-discovery-rate (FDR) correction was applied across the six lncRNAs within this endpoint.

2.10. Statistical analysis

All statistical analyses were performed using R (version 4.3.0). Unless otherwise stated, a two-sided P-value < 0.05 was considered statistically significant, and P-values from analyses involving multiple comparisons were adjusted using the Benjamini–Hochberg false discovery rate (FDR) method.

3. Results

3.1. Differential expression analysis

Differential expression analysis of the GSE37364 dataset was performed at a false discovery rate (FDR) < 0.05 and $|\log_2FC| > 1$. In the normal–adenoma comparison, 1,716 DEGs (comprising both coding and non-coding transcripts) were identified, of which 602 were upregulated and 1,114 downregulated in adenoma relative to normal mucosa (Fig. 2A). In the adenoma–tumor comparison, 774 DEGs were identified, of which 469 were upregulated and 305 downregulated in tumor relative to adenoma (Fig. 2B). In the normal–tumor comparison, 2,473 DEGs were identified, of which 1,029 were upregulated and 1,444 downregulated in tumor relative to normal mucosa (Fig. 2C). DEG statistics for the GSE100179 and GSE117606 datasets are summarized in Fig. S1. Furthermore, 251 shared DEGs were identified across normal–adenoma, normal–tumor, and adenoma–tumor comparisons from the three datasets, from which seven shared differentially expressed lncRNAs were subsequently selected for downstream analysis (Fig. 2D).

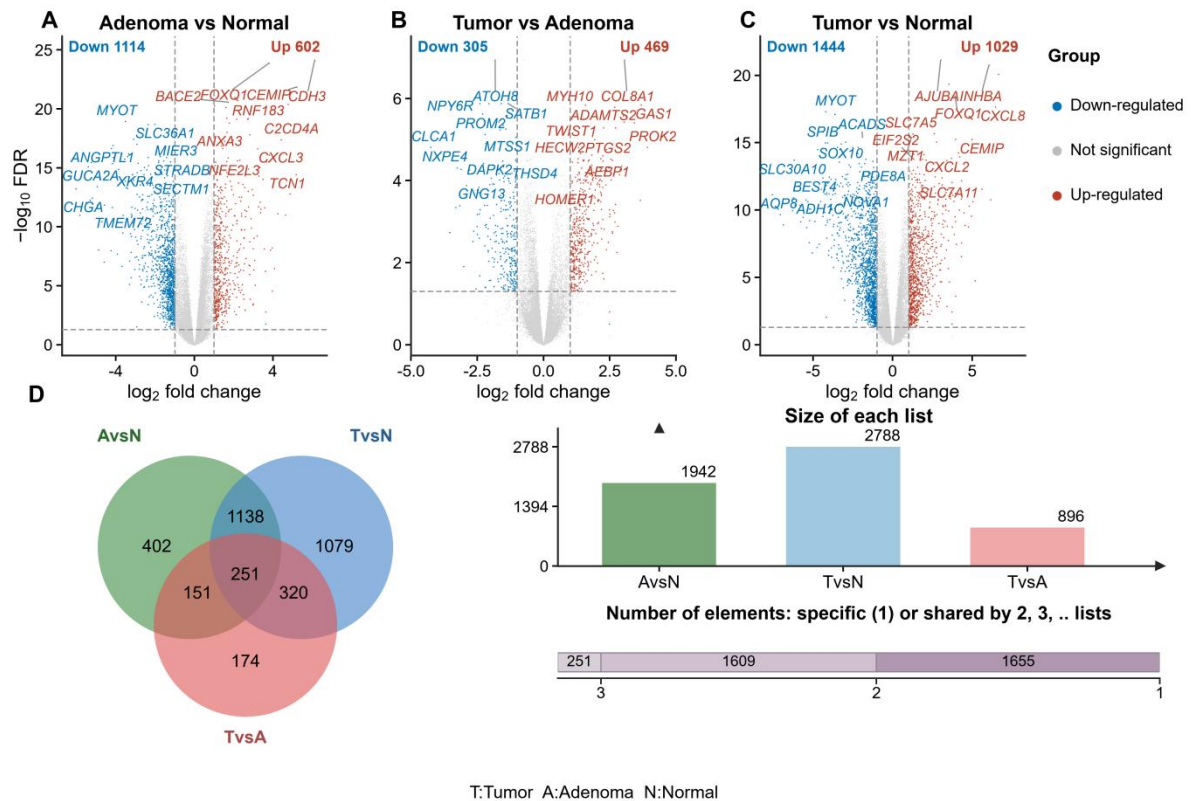


Fig. 2 Differential expression analysis of the GSE37364 dataset. (A) DEGs between adenoma and normal samples; (B) DEGs between tumor and adenoma samples; (C) DEGs between tumor and normal samples; (D) Venn diagram showing 251 shared DEGs across the three comparisons. N, normal; A, adenoma; T, tumor.

3.2. GO and KEGG enrichment analyses of the DEGs

To elucidate the potential biological functions of the differentially expressed genes (DEGs) associated with CRC malignant progression, Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses were performed for the DEGs identified in each of the three pairwise comparisons (normal–adenoma, normal–tumor, and adenoma–tumor). GO terms were classified into three ontologies: biological process (BP), cellular component (CC), and molecular function (MF).

In the normal–adenoma comparison, the enriched BP terms were predominantly related to immune regulation, including humoral immune response, immunoglobulin-mediated immune response, and B cell-mediated immunity. The corresponding CC terms were mainly enriched in collagen-containing extracellular matrix, immunoglobulin complex, and apical membrane structures; for MF, the DEGs were predominantly annotated with extracellular matrix structural constituent, glycosaminoglycan binding, and antigen binding (Fig. 3A). KEGG pathway analysis further revealed enrichment of these DEGs in protein digestion and

absorption and in bile secretion, suggesting that immune dysregulation and metabolic disturbance are already detectable in early adenoma lesions (Fig. 3B).

For the normal–tumor comparison, the DEGs remained significantly enriched in immune-related BP terms such as immune response and complement activation, and the enriched CC and MF terms were closely linked to immunoglobulin complex composition and receptor ligand activity. KEGG analysis showed prominent enrichment in cytokine–cytokine receptor interaction, together with protein digestion and absorption and bile secretion, suggesting persistent immune microenvironment dysfunction and aberrant signal transduction in CRC tissues (Fig. S2A–B).

By contrast, in the adenoma–tumor comparison a distinct enrichment pattern was observed. The most significantly enriched BP terms were related to extracellular matrix organization and leukocyte chemotaxis, reflecting active stromal remodeling and immune cell recruitment during malignant transformation; the CC and MF terms further corroborated the predominant alterations in extracellular matrix structure and receptor ligand activity. Notably, KEGG analysis identified enrichment of tumor progression-related pathways, including IL-17 and chemokine signaling, which have been implicated in CRC malignant progression (Fig. S2C–D).

Collectively, these results are consistent with stage-specific molecular dysregulation throughout CRC tumorigenesis and progression: early adenoma lesions are characterized mainly by immune and metabolic abnormalities, whereas the transition from benign adenoma to tumor is dominated by extracellular matrix remodeling and pro-tumor inflammatory pathway activation.

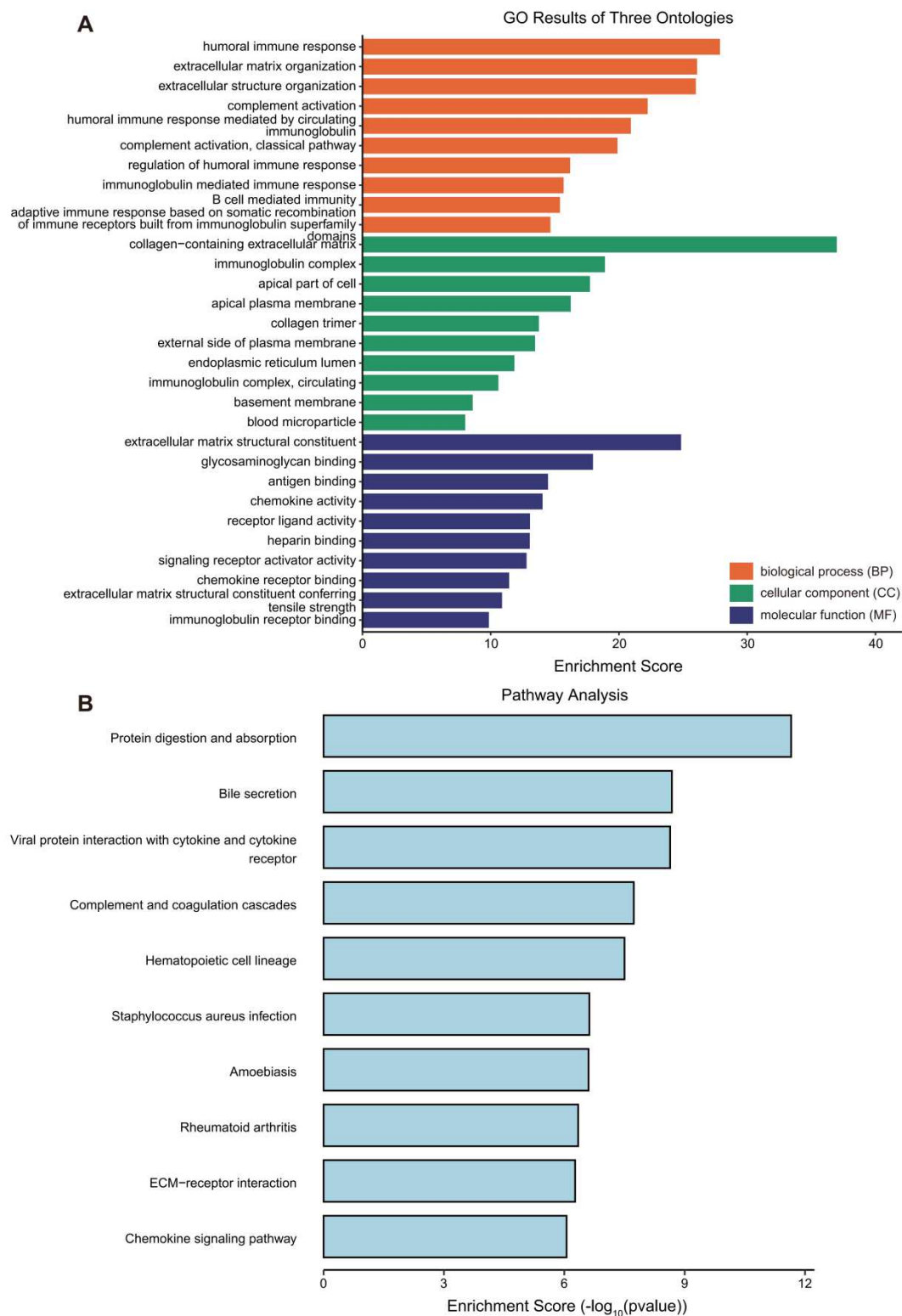


Fig. 3 GO and KEGG enrichment analyses of DEGs in colorectal cancer malignant progression. (A) Bar plot showing enriched GO terms (BP, CC and MF) of DEGs from the normal–adenoma group. (B) Bar plot showing enriched KEGG pathways of DEGs from the

normal–adenoma group. BP, biological process; CC, cellular component; MF, molecular function.

3.3. ML procedure

A two-layer machine learning (ML) framework integrating seven algorithms was constructed to screen stage-specific lncRNA biomarkers and to discriminate the sequential pathological stages of colorectal cancer (CRC) malignant transformation (Fig. S3A–H).

The first layer discriminated normal from diseased tissue (adenoma and tumor); the optimal combinatorial model achieved an accuracy of 0.972, an AUC of 0.979 and an F-score of 0.978 (Fig. S3A – D; Fig. S4A). Consensus variable-importance ranking across 21 valid combinatorial models prioritized LOC102724156, GATA2-AS1, RTEL1-TNFRSF6B, BLOC1S1-RDH5 and UCA1 (Fig. 4A).

The second layer further separated adenoma from tumor; the optimal model achieved an accuracy of 0.958, an AUC of 0.958 and an F-score of 0.955 (Fig. S3E–H; Fig. S4B), and 22 valid combinatorial models prioritized GATA2-AS1, LOC102724156, SATB2-AS1, H19 and BLOC1S1-RDH5 (Fig. 4B).

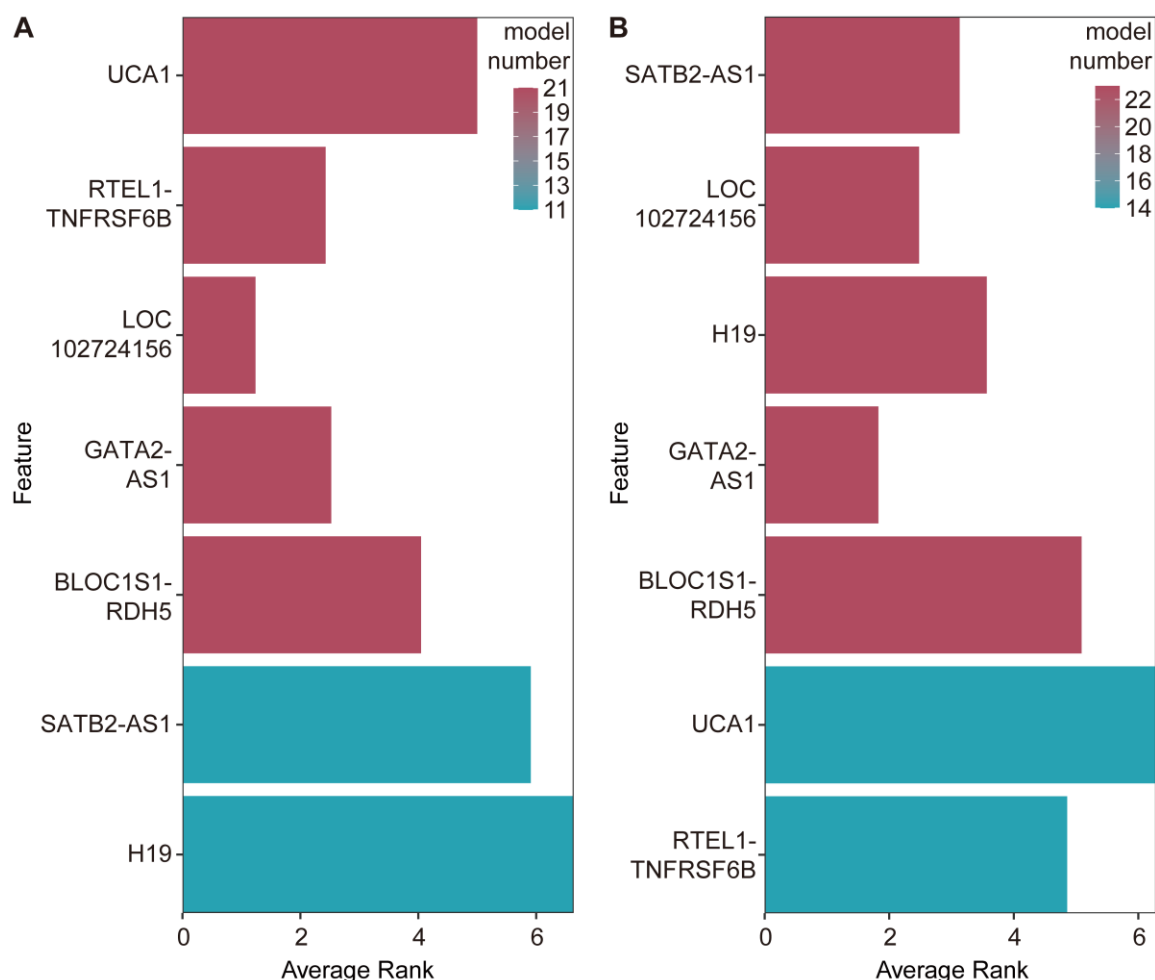


Fig. 4 Consensus variable importance of candidate lncRNAs in the two-layer machine learning framework. (A) First layer, normal versus diseased. (B) Second layer, adenoma versus carcinoma. Within each combinatorial model yielding valid feature weights, lncRNAs were ranked by absolute importance (rank 1 = most important). Bars show the mean within-model rank (average rank; lower = higher consensus importance); bar color denotes the number of models assigning a non-zero weight (model number).

3.4. AUC-ROC screening of core lncRNAs

For the discrimination of normal versus diseased (adenoma and tumor) samples in the GSE37364 dataset, the AUC values (with 95% confidence intervals) of the five core lncRNAs were as follows: RTTEL1-TNFRSF6B (0.912; 95% CI 0.852–0.973), UCA1 (0.824; 95% CI 0.733–0.914), LOC102724156 (0.961; 95% CI 0.894–0.986), GATA2-AS1 (0.898; 95% CI 0.823–0.973) and BLOC1S1-RDH5 (0.920; 95% CI 0.862–0.977) (Fig. 5A). For the classification of adenoma versus tumor samples, the AUC values were SATB2-AS1 (0.825;

95% CI 0.703–0.948), LOC102724156 (0.791; 95% CI 0.647–0.935), H19 (0.791; 95% CI 0.657–0.924), GATA2-AS1 (0.744; 95% CI 0.580–0.908) and BLOC1S1-RDH5 (0.670; 95% CI 0.477–0.863) (Fig. 5B). In addition, the expression patterns of these candidate lncRNAs across normal, adenoma and tumor samples are shown in Fig. 5C and Fig. S5. Notably, the two-layer machine learning resolved distinct biomarker roles: H19 was not retained among the top five in the first layer (AUC 0.636) but was informative specifically for the adenoma–tumor transition (AUC 0.791), whereas GATA2-AS1 ranked among the top two in both layers (AUC 0.898 and 0.744), indicating a contribution spanning both transitions.

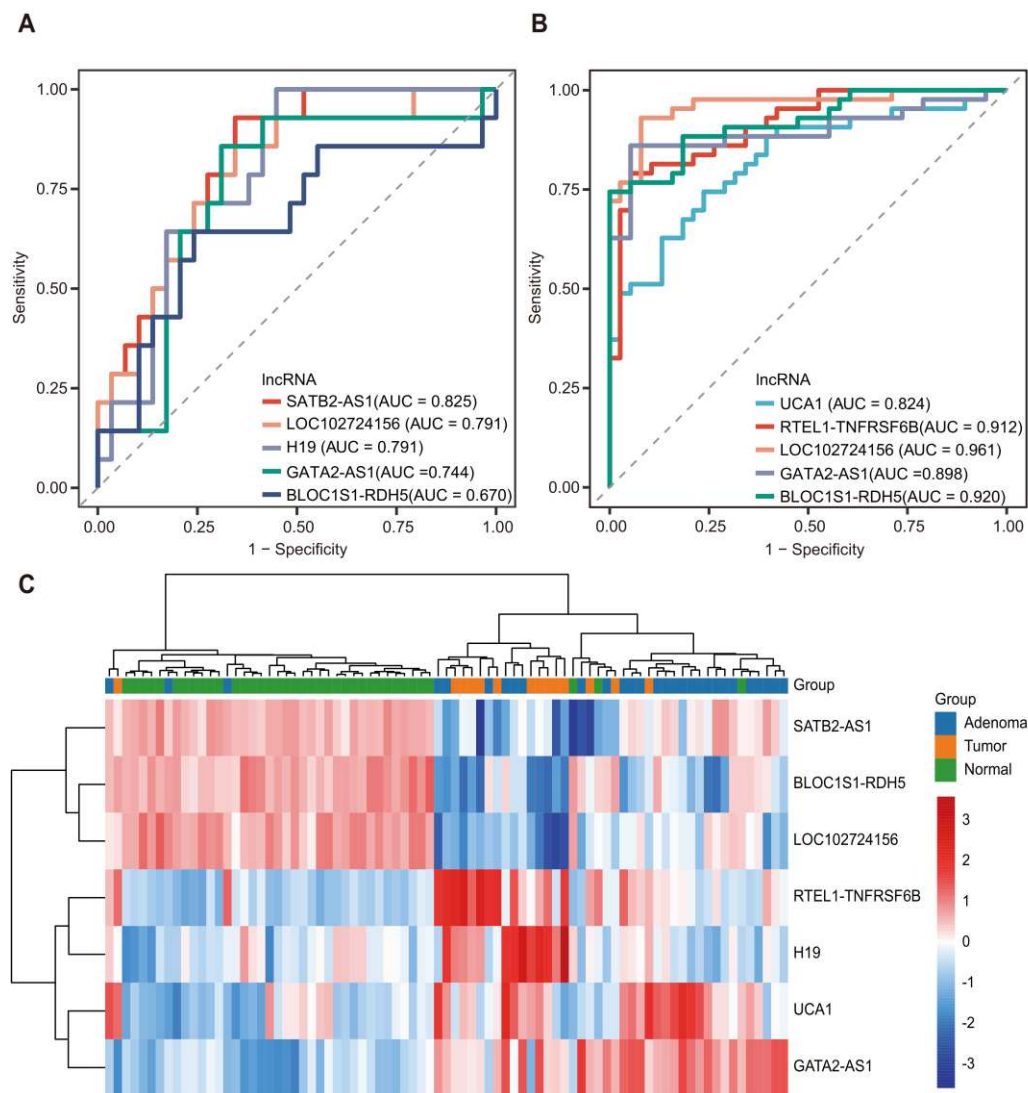


Fig. 5 ROC curves and expression heatmap of core lncRNA biomarkers based on the GSE37364 cohort. (A) ROC curves of five core lncRNAs for distinguishing normal versus diseased (adenoma and tumor) samples, with AUC values annotated for each lncRNA. (B) ROC curves of five core lncRNAs for adenoma–tumor classification. (C) Heatmap with

hierarchical clustering illustrating the expression patterns of candidate lncRNAs in normal, adenoma and tumor samples. Red denotes upregulated expression and blue denotes downregulated expression.

3.5. Immune infiltration

To characterize the immune microenvironment during colorectal carcinogenesis, CIBERSORT was used to compare the estimated fractions of 22 immune cell subsets among normal, adenoma and tumor samples in GSE37364. Compared with normal samples, adenoma and tumor samples showed higher infiltration of activated CD4 memory T cells, M0 macrophages, activated dendritic cells, activated mast cells and neutrophils. In contrast, normal samples exhibited higher infiltration of plasma cells, CD8 T cells, resting dendritic cells, resting mast cells and M2 macrophages. Plasma cells were the most abundant immune cell subset across all samples (Fig. 6). Correlation analysis further revealed that RTEL1-TNFRSF6B and H19 were positively correlated with M0 macrophages, whereas LOC102724156, SATB2-AS1 and BLOC1S1-RDH5 were negatively correlated with M0 macrophages (all $P < 0.05$). RTEL1-TNFRSF6B was also positively associated with activated dendritic cells, while LOC102724156 and BLOC1S1-RDH5 were negatively associated with activated dendritic cells (all $P < 0.05$). These results indicate that the expression of the candidate lncRNAs is associated with the estimated fractions of specific immune-cell subsets in this cohort (Fig. 7).

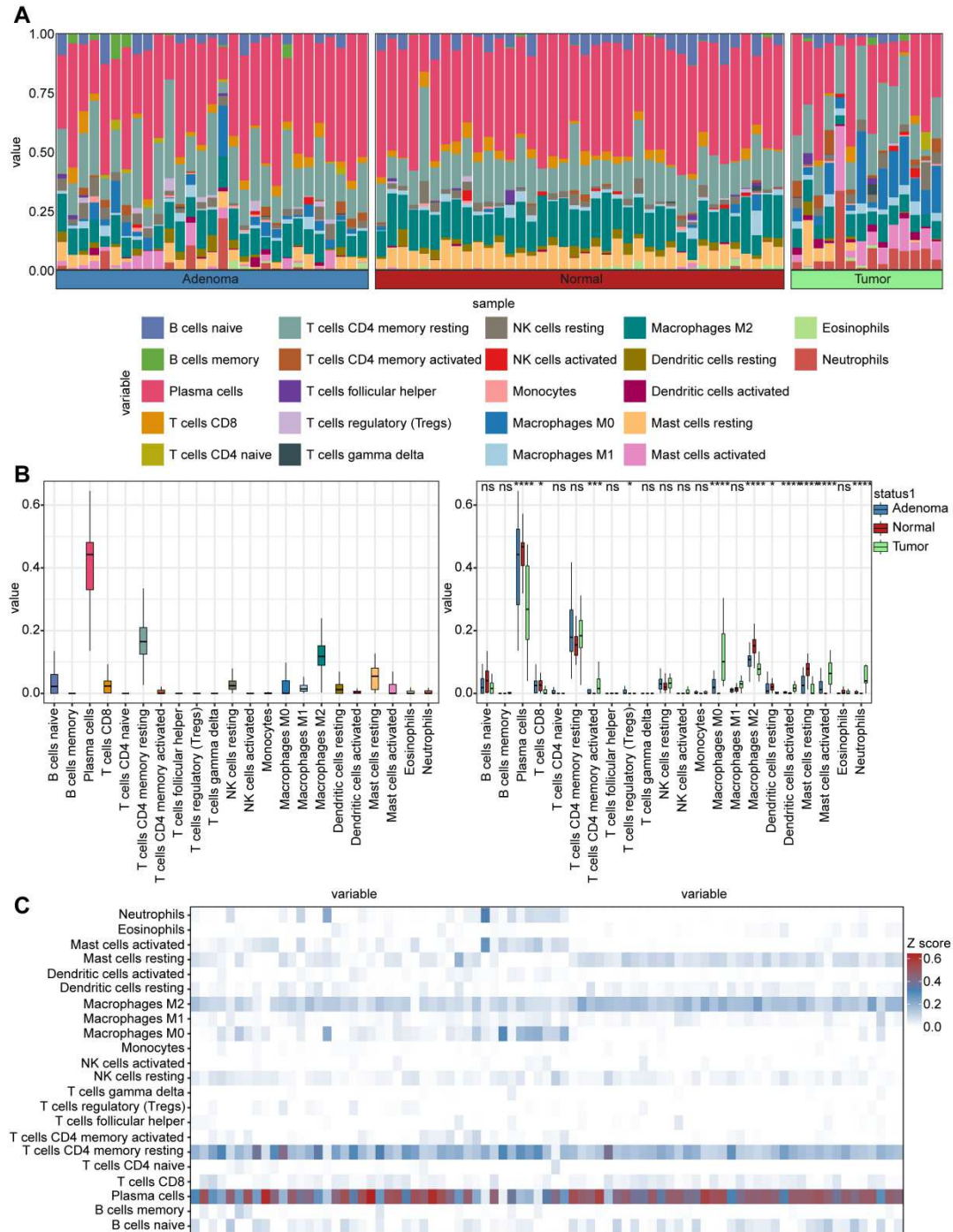


Fig. 6 Immune cell infiltration in GSE37364. (A) Stacked bar plot showing the relative fractions of 22 immune cell subsets in individual normal, adenoma and tumor samples. (B, C) Box plots comparing the estimated fractions of each immune cell subset among normal, adenoma and tumor samples. (D) Heatmap of the relative fractions of immune cell subsets across samples. N, normal; A, adenoma; T, tumor. * $P < 0.05$, ** $P < 0.01$; ns, not significant.

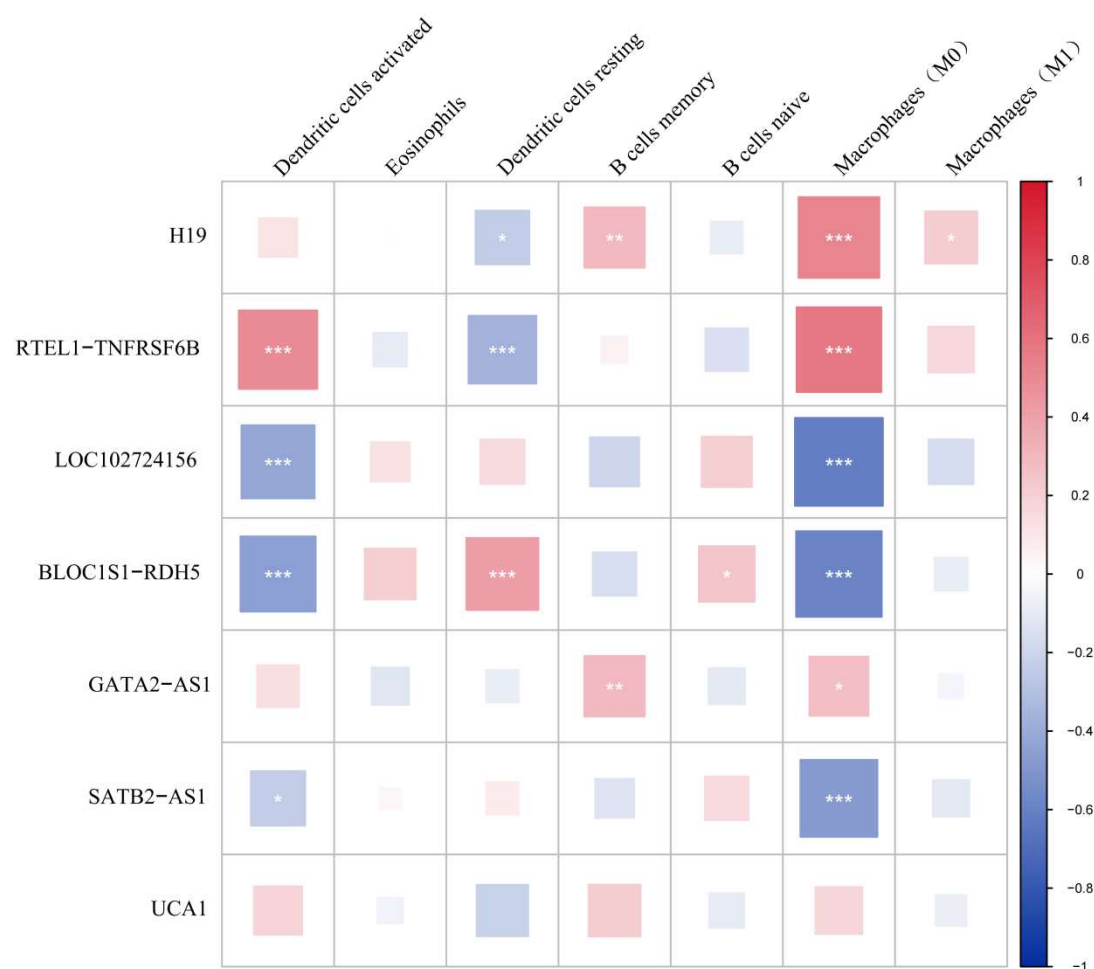


Fig. 7 Correlation between lncRNA expression and the estimated fractions of M0 macrophages and activated dendritic cells in GSE37364. Scatter plots with fitted regression lines; Spearman correlation coefficients (ρ) and two-sided P-values are shown for each lncRNA. Asterisks denote statistical significance (* $P < 0.05$; ** $P < 0.01$).

3.6. Single-cell analysis

Single-cell RNA sequencing (scRNA-seq) data of 46,562 cells derived from 5 normal, 5 adenoma and 4 tumor samples were obtained from the GSE201349 dataset and subjected to rigorous quality control. Gene counts, sequencing depth, and mitochondrial and hemoglobin gene percentages were assessed for each sample. After normalization, 3,000 highly variable genes were selected for downstream dimensionality reduction and clustering. Principal component analysis (PCA) was performed, and the top 30 principal components (PCs) were

used for t-distributed stochastic neighbor embedding (t-SNE) visualization. t-SNE partitioned the qualified cells into 21 clusters, which were annotated into 14 major cell types with SingleR, including B cells, chondrocytes, embryonic stem cells, endothelial cells, epithelial cells, fibroblasts, gametocytes, hepatocytes, induced pluripotent stem cells (iPSCs), neurons, smooth muscle cells, T cells, tissue stem cells and unclassified cells (Fig. 8A–C); the proportions of these cell types in individual samples are shown in Fig. 8D and Fig. S6. Because SingleR was queried against a broad multi-tissue reference, a subset of these labels reflects transcriptional states rather than discrete lineages resident in colonic tissue. The clusters annotated as embryonic stem cells and induced pluripotent stem cells (iPSCs) were contiguous with the epithelial compartment and are hereafter referred to as stem-like epithelial cells, whereas the chondrocyte, gametocyte and hepatocyte labels were regarded as reference mismatch and were not interpreted biologically.

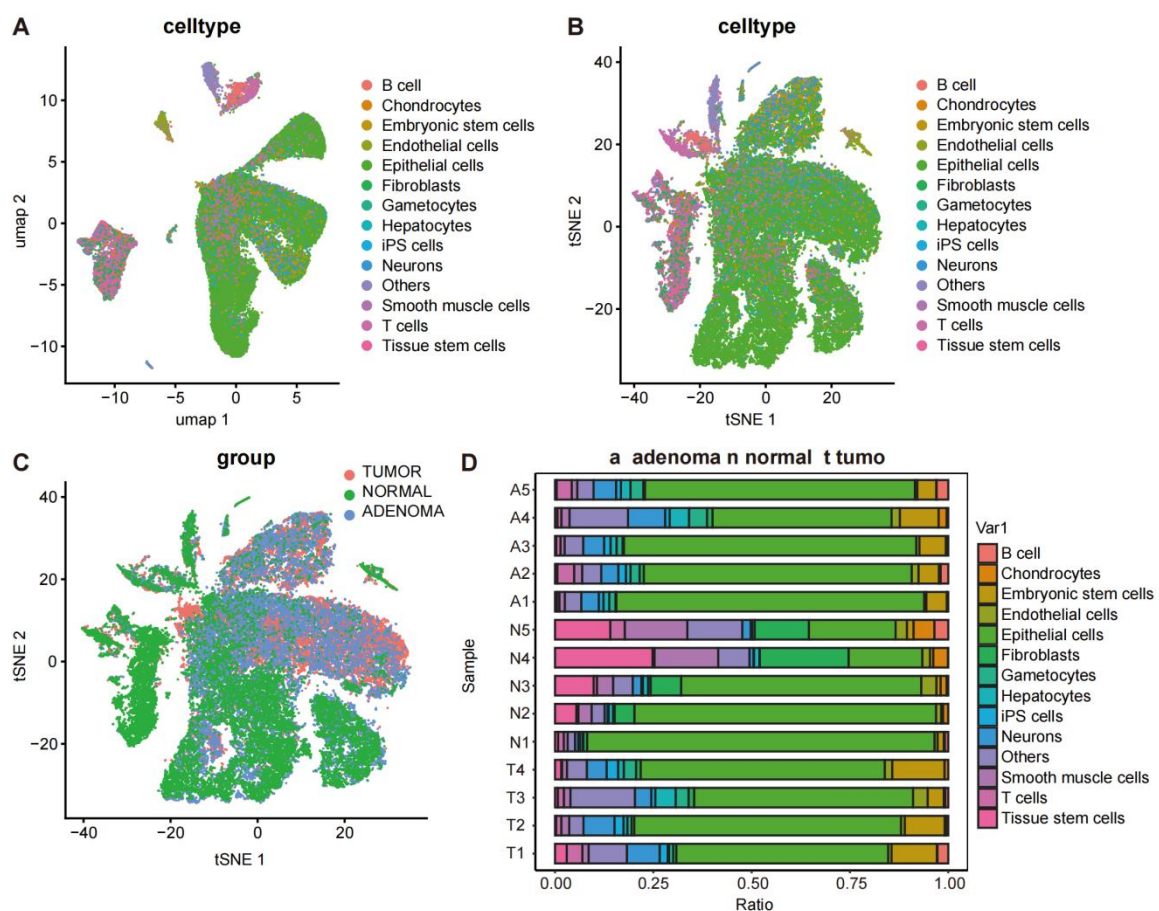


Fig. 8 Single-cell transcriptomic atlas of colorectal cancer. (A, B) Uniform manifold approximation and projection (UMAP) and t-SNE plots of 46,562 cells from 5 normal, 5 adenoma and 4 tumor samples in the GSE201349 dataset, colored by 14 cell types annotated

by SingleR. (C) t-SNE plot colored by pathological group (normal, adenoma, tumor). (D) Proportions of the 14 cell types in each individual sample. N, normal; A, adenoma; T, tumor. Further expression profiling demonstrated that three core long non-coding RNAs (lncRNAs), UCA1, GATA2-AS1 and SATB2-AS1, showed distinct expression patterns across cell types, with GATA2-AS1 and SATB2-AS1 predominantly enriched in stem-like epithelial cells within the epithelial compartment (Fig. 9).

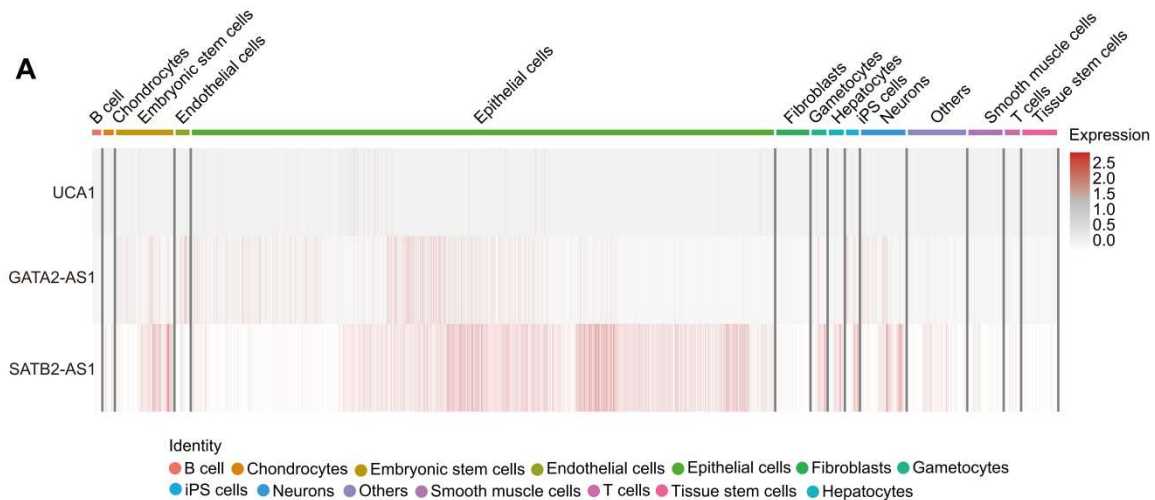


Fig. 9 Cell-type-specific expression patterns of key lncRNAs in the CRC malignant transformation continuum. Heatmap showing the expression levels of UCA1, GATA2-AS1 and SATB2-AS1 across individual cells, grouped by cell identity. Cells are color-coded by cell type. The color scale represents normalized expression levels, with red indicating higher expression and gray indicating lower expression.

The AUCell algorithm was subsequently applied to quantify cell-type-specific lncRNA activity scores and characterize their differential expression patterns across distinct pathological groups. In addition, UCA1 was detected only sporadically across cell subsets, and its expression was relatively lower in adenoma samples than in tumor counterparts. Notably, adenoma tissues exhibited markedly higher AUCell scores for GATA2-AS1 and SATB2-AS1 in stem-like epithelial cells, indicating enhanced transcriptional activity of these lncRNAs within this progenitor subpopulation. Collectively, these single-cell profiling results demonstrated cell-type- and disease-stage-specific expression patterns of the key lncRNAs during CRC malignant transformation (Fig. 10).

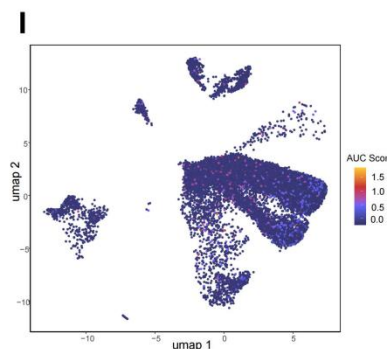
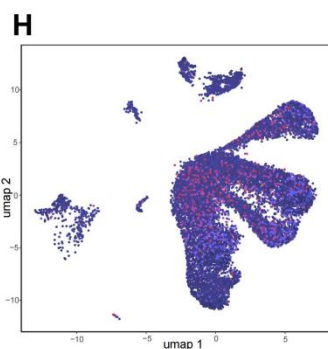
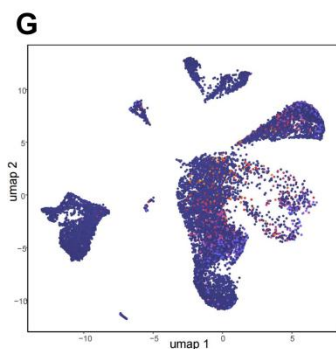
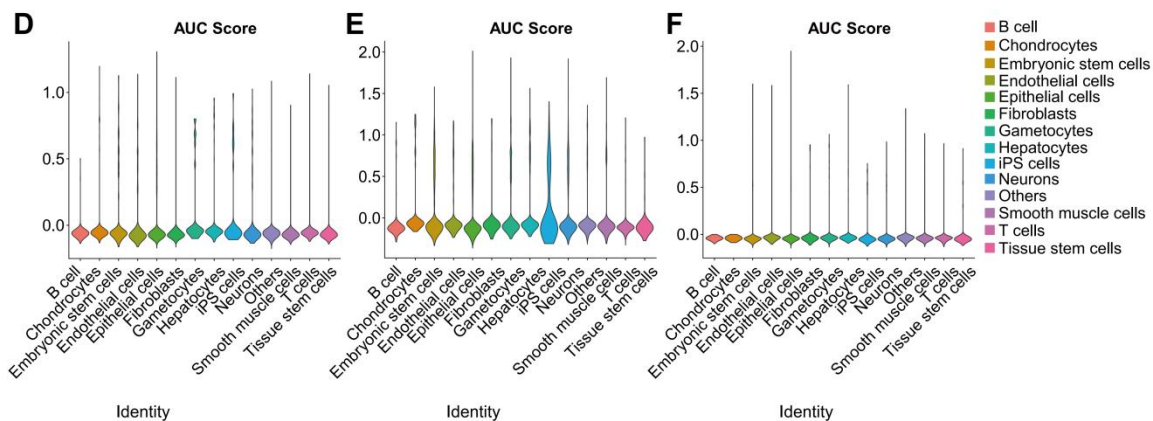
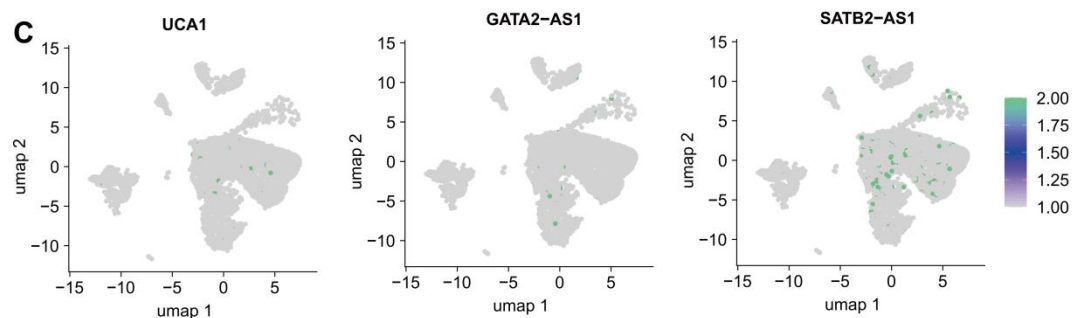
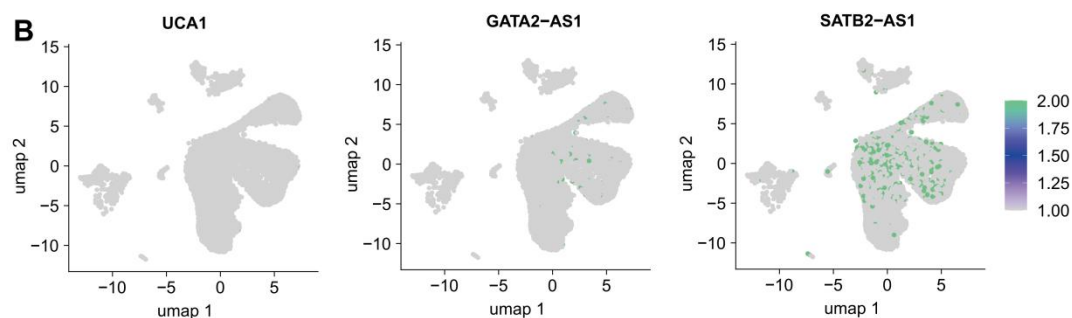
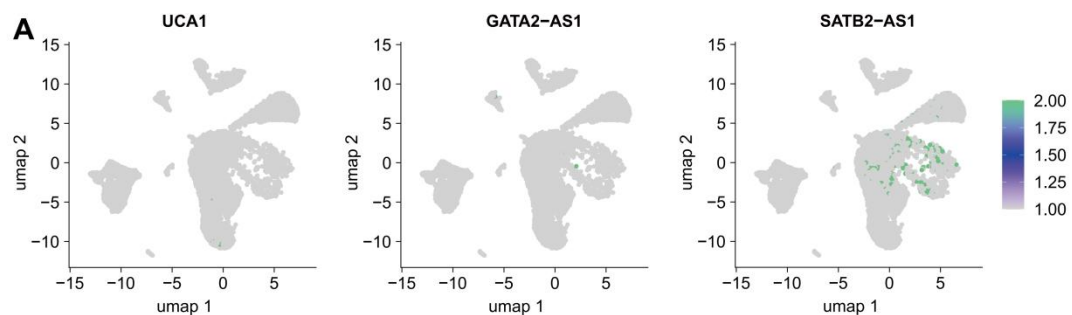


Fig. 10 Single-cell expression and AUCell activity of the core lncRNAs in the GSE201349 dataset. (A–C) UMAP projections of cells from normal (A), adenoma (B) and tumor (C) samples, colored by the expression of UCA1, GATA2-AS1 and SATB2-AS1. (D–F) Violin plots showing AUCell activity scores across the 14 annotated cell types in normal, adenoma and tumor samples, respectively. (G–I) UMAP projections colored by AUCell activity scores. Color intensity reflects the expression level in (A–C) and the transcriptional activity in (G–I).

3.7. Association between core lncRNA expression and progression-free survival

Six of the seven lncRNAs in the two-layer classifiers were measurable in the TCGA colorectal adenocarcinoma cohort; LOC102724156 was absent from public TCGA expression matrices. Among the 357 patients evaluable for progression-free survival (PFS) (95 events; follow-up capped at 60 months), patients were ranked within each lncRNA according to tumor expression, and the upper (T3) and lower (T1) tertiles were compared with the middle tertile excluded; 238 patients were compared for each lncRNA and 241 for BLOC1S1-RDH5. Kaplan–Meier curves for the six lncRNAs are shown in Fig. 11, and Benjamini–Hochberg adjusted P values (q) were calculated across the six lncRNAs. PFS did not differ between the upper and lower expression tertiles for UCA1 (log-rank $P = 0.420$, $q = 0.840$; hazard ratio [HR] for the upper versus lower tertile 1.22, 95% CI 0.75–1.96), RTEL1-TNFRSF6B ($P = 0.863$, $q = 0.863$; HR 0.95, 95% CI 0.56–1.63), GATA2-AS1 ($P = 0.837$, $q = 0.863$; HR 0.95, 95% CI 0.58–1.55) or SATB2-AS1 ($P = 0.767$, $q = 0.863$; HR 0.93, 95% CI 0.56–1.54). Patients in the upper H19 tertile had shorter PFS than those in the lower tertile ($P = 0.013$, $q = 0.078$; HR 1.83, 95% CI 1.13–2.99). The upper BLOC1S1-RDH5 tertile showed a similar direction of effect that did not reach statistical significance ($P = 0.064$, $q = 0.192$; HR 1.61, 95% CI 0.95–2.68). No lncRNA retained a

significant association with PFS after correction for multiple testing.

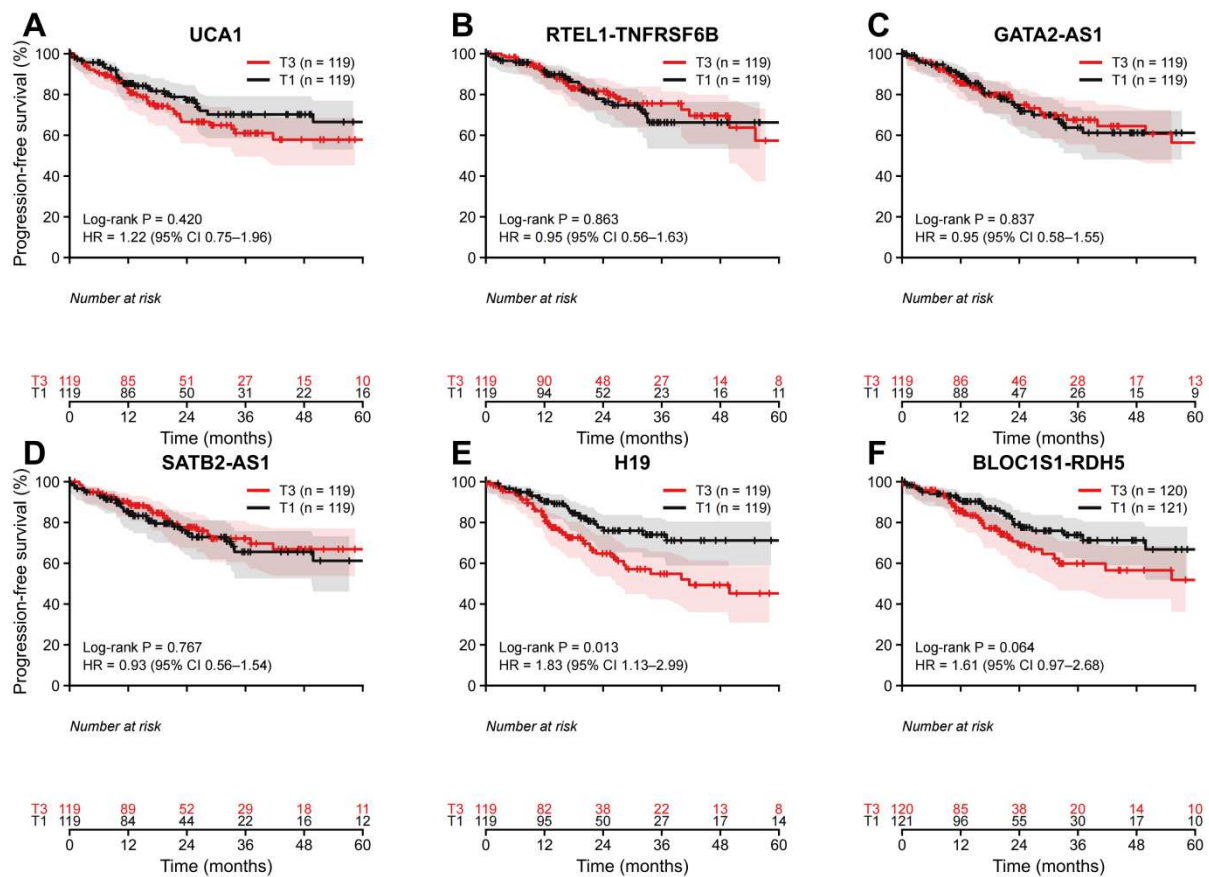


Fig. 11 Kaplan – Meier analysis of progression-free survival by core lncRNA expression tertile. (A–F) Curves for patients with upper- versus lower-tertile expression of UCA1 (A), RTKL1-TNFRSF6B (B), GATA2-AS1 (C), SATB2-AS1 (D), H19 (E) and BLOC1S1-RDH5 (F). Shaded areas indicate 95% confidence intervals; numbers at risk are shown beneath each plot.

4. Discussion

Colorectal malignant transformation is a multistep pathological cascade that evolves from normal mucosa through premalignant adenoma to invasive tumor, accompanied by cumulative genetic and epigenetic alterations and dynamic remodeling of the tumor immune microenvironment [33]. Long non-coding RNAs have been implicated in CRC initiation and progression through oncogenic signaling, epithelial–mesenchymal transition, and immune-microenvironment reprogramming [34]. Most existing studies, however, compared lncRNA expression only between normal and tumor tissues. Kalmár and colleagues previously profiled lncRNA expression across the adenoma–tumor transition in a single cohort [35]. The present study extends this work in three respects: candidate lncRNAs were

required to be differentially expressed in three independent datasets rather than one; the final panel was derived from a multi-algorithm machine-learning framework and evaluated as a two-layer classifier distinguishing normal, adenoma and tumor tissue rather than by differential expression alone; and the candidates were further examined at single-cell resolution and related to immune-cell infiltration. In total, seven core lncRNAs were identified and assembled into two-layer signatures that distinguished normal from diseased tissue and adenoma from tumor tissue.

The two-layer framework combined seven machine-learning algorithms into 120 candidate models, so that the final signatures were selected according to cross-validated classification performance rather than differential-expression statistics alone. The first-layer signature, consisting of UCA1, RTEL1-TNFRSF6B, LOC102724156, GATA2-AS1 and BLOC1S1-RDH5, distinguished normal from diseased samples with an accuracy of 0.972 and an AUC of 0.979, and the second-layer signature, including SATB2-AS1, LOC102724156, H19, GATA2-AS1 and BLOC1S1-RDH5, separated adenoma from tumor samples with an accuracy of 0.958 and an AUC of 0.958. Individually, the AUC values of the core lncRNAs ranged from 0.670 to 0.961. Several of these lncRNAs have been reported in CRC. Consistent with earlier studies, the oncogenic lncRNA UCA1 contributes to CRC proliferation and metastasis through ceRNA axes [36]; in the present cohort, its expression increased progressively from normal mucosa through adenoma to tumor (Fig. 5C; Fig. S5). H19 can activate Wnt signaling during colorectal carcinogenesis [37]; our results additionally showed a positive correlation between H19 and M0 macrophage infiltration, pointing to an immune-related function during the normal–adenoma–tumor transition. In contrast, LOC102724156 and BLOC1S1-RDH5 have been little studied in CRC. Their high variable-importance scores, first-layer AUC values of 0.961 and 0.920, and significant correlations with infiltrating immune cells suggest that they may participate in CRC malignant transformation and TIME remodeling; the mechanistic basis warrants further investigation. Notably, LOC102724156 was shared by both signatures, indicating that it may act throughout the adenoma–tumor progression.

Functional enrichment analysis of the DEGs delineated stage-specific biological alterations along the normal–adenoma–tumor transition. Immune-response and metabolic processes predominated in early adenoma lesions, whereas extracellular-matrix remodeling and pro-inflammatory pathways such as IL-17 and chemokine signaling became prominent during the transition from adenoma to tumor. These findings are consistent with previous

pathological observations that immune dysregulation is an early event in premalignant colonic lesions, and that subsequent stromal remodeling and chronic inflammatory activation facilitate the acquisition of invasive phenotypes [38]. CIBERSORT analysis further showed higher proportions of M0 macrophages and activated dendritic cells in adenoma and tumor tissues than in normal mucosa [39]. Correlation analysis indicated that the core lncRNAs were significantly associated with these innate immune-cell populations: RTEL1-TNFRSF6B and H19 were positively correlated with M0 macrophages, whereas LOC102724156, SATB2-AS1 and BLOC1S1-RDH5 were negatively correlated. Several studies have shown that lncRNAs can regulate macrophage polarization and dendritic-cell function and thereby shape the immune microenvironment in CRC [40]. Our results provide computational evidence linking these lncRNAs to innate immune-cell infiltration during the normal–adenoma–tumor transition, suggesting that they may be involved in premalignant progression through their association with tumor-infiltrating immune cells.

By integrating bulk-tissue screening with single-cell transcriptome analysis, this study partly overcomes a limitation of bulk sequencing, namely that it masks cell-type-specific expression heterogeneity. Re-analysis of the GSE201349 dataset showed that GATA2-AS1 and SATB2-AS1 were predominantly enriched in stem-like epithelial cells, and that SATB2-AS1 was upregulated in adenoma samples. AUCell analysis further indicated higher transcriptional activity of the core lncRNAs in stem-like epithelial cells of adenoma specimens. Stem-like epithelial cells in premalignant adenoma are considered a possible cellular origin of subsequent malignant transformation, and aberrant activation of lncRNAs in these cells may contribute to the acquisition of malignant potential [41]. These single-cell observations complement the bulk-tissue results and indicate that the identified lncRNAs show both stage- and cell-type-dependent expression patterns, whereas UCA1 was detected only sporadically and at lower levels in adenoma than in tumor samples.

Whether expression of the seven lncRNAs is related to clinical outcome is a question distinct from the discrimination of normal, adenoma and tumor tissue addressed by the two-layer signatures. In the TCGA colorectal adenocarcinoma cohort (357 patients evaluable for PFS, 95 events, follow-up capped at 60 months), patients in the upper H19 expression tertile had shorter PFS than those in the lower tertile (log-rank $P = 0.013$, $q = 0.078$; HR 1.83, 95% CI 1.13–2.99), whereas no difference between tertiles was observed for UCA1, RTEL1-TNFRSF6B, GATA2-AS1, SATB2-AS1 or BLOC1S1-RDH5 (all log-rank $P > 0.05$ and $q \geq 0.192$). After correction for the six lncRNAs tested, no association remained

significant. The direction of the H19 result is compatible with the oncogenic functions reported for this lncRNA in colorectal carcinogenesis [37]; whether it reflects the positive correlation between H19 and M0 macrophage infiltration cannot be judged from these data. Several constraints apply to this analysis: tertile boundaries were defined within this cohort rather than pre-specified, the comparison was restricted to the upper and lower tertiles, the number of progression events was modest, and LOC102724156 could not be assessed because it is absent from public TCGA expression matrices. The seven lncRNAs were therefore assembled and evaluated as classifiers of the normal–adenoma–tumor sequence, and the present data do not support their use as prognostic markers; any prognostic value remains to be tested in independent cohorts with larger numbers of events.

From a translational perspective, the two-layer lncRNA signatures have potential clinical implications. Colonoscopy remains the reference standard for CRC screening, but its use is limited by its invasive nature and restricted population access. Molecular biomarkers for identifying high-risk adenoma and early tumor are therefore needed for auxiliary risk stratification [4]. In this study, the lncRNA panels distinguished normal from diseased tissue with an AUC of 0.979 and adenoma from tumor tissue with an AUC of 0.958, supporting their potential value as tissue markers for identifying premalignant and malignant lesions. In addition, given their correlations with immune-cell infiltration, these lncRNAs may provide clues for immune-targeted strategies in CRC. For instance, lncRNAs associated with macrophage or dendritic-cell abundance may represent candidate targets for modulating the immune microenvironment in premalignant lesions and early-stage CRC [42], although this remains to be tested experimentally.

5. Conclusion

Seven lncRNAs associated with the normal–adenoma–tumor sequence of CRC were identified by integrating multi-dataset transcriptomics, multi-algorithm machine learning and single-cell transcriptome profiling, and were assembled into a two-layer classifier that discriminated normal from diseased tissue and adenoma from tumor tissue. These lncRNAs showed both stage- and cell-type-dependent expression, with preferential activity in stem-like epithelial cells of adenoma samples, and were consistently associated with innate immune-cell infiltration. After correction for multiple testing, none of the six lncRNAs measurable in the TCGA cohort was associated with progression-free survival. The seven lncRNAs are therefore candidate tissue biomarkers for staging along the normal–adenoma–

tumor sequence rather than prognostic markers; their diagnostic value requires prospective clinical validation, and their biological functions remain to be tested experimentally.

6. Limitations

Several limitations of this study should be acknowledged. First, all biomarker screening and immune-correlation analyses were based on public transcriptomic datasets. Although multiple GEO cohorts were included for discovery, the three microarray datasets were generated on different platforms and do not constitute a fully independent validation cohort; prospective clinical cohorts with well-annotated clinical information are required for external validation. Second, the sample size for the adenoma–tumor comparison was relatively small, which may limit the stability of the second-layer signature. Third, CIBERSORT estimates relative fractions of immune cell subsets that sum to one, and correlation analyses based on such compositional data should be interpreted with caution. Fourth, our conclusions regarding immune-regulatory effects are derived from computational correlation analysis; the mechanisms by which LOC102724156, BLOC1S1-RDH5 and other candidates modulate immune-cell phenotypes remain unclear, and *in vitro* and *in vivo* experiments are needed. Fifth, the single-cell data were obtained from poly(A)-enriched libraries, which may miss non-polyadenylated lncRNA transcripts. Finally, this work focused on transcript-level expression; epigenetic alterations such as DNA methylation at these lncRNA loci warrant further exploration.

Funding

This work was supported by the Chengde Municipal Science and Technology Bureau (Chengde Basic Research Project “Dysregulation of lncRNAs during the malignant transformation of colon cells and its relationship with immune cell profiles”, Grant No. 202301A113).

Data availability

The datasets analyzed during the current study are publicly available in the Gene Expression Omnibus (GEO) repository under the following accession numbers and links: GSE37364 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE37364>), GSE100179 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE100179>), GSE117606 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE117606>) and GSE201349 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE201349>). The TCGA colorectal

adenocarcinoma (COAD and READ) RNA-sequencing and clinical data used for the survival analysis were downloaded from the UCSC Xena repository at the following cohort pages:

TCGA Colon Cancer (COAD)

([https://xenabrowser.net/datapages/?cohort=TCGA%20Colon%20Cancer%20\(COAD\)\)](https://xenabrowser.net/datapages/?cohort=TCGA%20Colon%20Cancer%20(COAD))) and

TCGA Rectum Cancer (READ)

([https://xenabrowser.net/datapages/?cohort=TCGA%20Rectum%20Cancer%20\(READ\)\)](https://xenabrowser.net/datapages/?cohort=TCGA%20Rectum%20Cancer%20(READ)))).

The LM22 signature matrix used for the CIBERSORT analysis was obtained from the CIBERSORT web portal (<https://cibersortx.stanford.edu/>), where it is available upon free registration, and is also distributed with the IOBR R package (<https://github.com/IOBR/IOBR>) used in this study. All other data generated or analyzed during this study are included in this published article and its supplementary information files.

Declarations

Ethics approval and consent to participate

Not applicable. This study analyzed only publicly available, de-identified datasets deposited in the Gene Expression Omnibus and The Cancer Genome Atlas; no human participants, human data or biological material were directly collected by the authors.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing financial or non-financial interests.

Author contributions

Han Gong and Fuyang Gao conceived and designed the study. Han Gong, Libo Liu, Chao Mi and Bolun Yan acquired, analyzed and interpreted the data. Han Gong drafted the manuscript. Fuyang Gao and Huayu Zhang critically revised the manuscript for important intellectual content and supervised the work. All authors read and approved the final version of the manuscript.

References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2024;74(3):229–263.
<https://doi.org/10.3322/caac.21834>
2. Gharib E, Robichaud GA. From crypts to cancer: a holistic perspective on colorectal carcinogenesis and therapeutic strategies. *Int J Mol Sci.* 2024;25(17):9463.
<https://doi.org/10.3390/ijms25179463>
3. Kumar SN, Kiruthiga SS, Ramasamy M. Integrated perspectives on colorectal carcinogenesis: molecular pathogenesis, genomic alterations, diagnostic paradigms, therapeutic interventions and AI-driven directions in precision oncology. *Front Oncol.* 2026;16:1770430. <https://doi.org/10.3389/fonc.2026.1770430>
4. Wu Y, Xu X. Long non-coding RNA signature in colorectal cancer: research progression and clinical application. *Cancer Cell Int.* 2023;23(1):28.
<https://doi.org/10.1186/s12935-023-02867-0>
5. Lin Y, Zhao W, Lv Z, Xie H, Li Y, Zhang Z. The functions and mechanisms of long non-coding RNA in colorectal cancer. *Front Oncol.* 2024;14:1419972.
<https://doi.org/10.3389/fonc.2024.1419972>
6. Hamidi AA, Khalili-Tanha G, Nasrpour Navaei Z, Moghbeli M. Long non-coding RNAs as the critical regulators of epithelial mesenchymal transition in colorectal tumor cells: an overview. *Cancer Cell Int.* 2022;22(1):71. <https://doi.org/10.1186/s12935-022-02501-5>
7. Li R, Xu H, Gao X. The ceRNA network regulates epithelial-mesenchymal transition in colorectal cancer. *Heliyon.* 2023;9(3):e14143.
<https://doi.org/10.1016/j.heliyon.2023.e14143>
8. Chowdhary A, Satagopam V, Schneider R. Long non-coding RNAs: mechanisms, experimental, and computational approaches in identification, characterization, and their biomarker potential in cancer. *Front Genet.* 2021;12:649619.
<https://doi.org/10.3389/fgene.2021.649619>

9. Chowdhury PR, Salvamani S, Gunasekaran B, Peng HB, Ulaganathan V. H19: an oncogenic long non-coding RNA in colorectal cancer. *Yale J Biol Med.* 2023;96(4):495–509. <https://doi.org/10.59249/tdbj7410>
10. Luan Y, Li X, Luan Y, Zhao R, Li Y, Liu L, et al. Circulating lncRNA UCA1 promotes malignancy of colorectal cancer via the miR-143/MYO6 axis. *Mol Ther Nucleic Acids.* 2020;19:790–803. <https://doi.org/10.1016/j.omtn.2019.12.009>
11. Huang D, Ma N, Li X, Gou Y, Duan Y, Liu B, et al. Advances in single-cell RNA sequencing and its applications in cancer research. *J Hematol Oncol.* 2023;16(1):98. <https://doi.org/10.1186/s13045-023-01494-6>
12. Yuan L, Zhao J, Sun T, Shen Z. A machine learning framework that integrates multi-omics data predicts cancer-related LncRNAs. *BMC Bioinformatics.* 2021;22(1):332. <https://doi.org/10.1186/s12859-021-04256-8>
13. Liu Z, Liu L, Weng S, Guo C, Dang Q, Xu H, et al. Machine learning-based integration develops an immune-derived lncRNA signature for improving outcomes in colorectal cancer. *Nat Commun.* 2022;13(1):816. <https://doi.org/10.1038/s41467-022-28421-6>
14. Fu X, Ma W, Zuo Q, Qi Y, Zhang S, Zhao Y. Application of machine learning for high-throughput tumor marker screening. *Life Sci.* 2024;348:122634. <https://doi.org/10.1016/j.lfs.2024.122634>
15. Cheng L, Han T, Zhang Z, Yi P, Zhang C, Zhang S, et al. Identification and validation of six autophagy-related long non-coding RNAs as prognostic signature in colorectal cancer. *Int J Med Sci.* 2021;18(1):88–98. <https://doi.org/10.7150/ijms.49449>
16. Zhang X, Li T, Wang J, Li J, Chen L, Liu C. Identification of cancer-related long non-coding RNAs using XGBoost with high accuracy. *Front Genet.* 2019;10:735. <https://doi.org/10.3389/fgene.2019.00735>
17. Mahmud S, Zheng C, Santiago FE, Zhang L, Robbins PD, Dong X. A machine learning approach identifies cellular senescence on transcriptome data of human cells in vitro. *GeroScience.* 2025;47(3):5287–5301. <https://doi.org/10.1007/s11357-024-01485-6>

18. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* 2013;41:D991–D995. <https://doi.org/10.1093/nar/gks1193>
19. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 2015;43(7):e47. <https://doi.org/10.1093/nar/gkv007>
20. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation (Camb).* 2021;2(3):100141. <https://doi.org/10.1016/j.xinn.2021.100141>
21. Cover T, Hart P. Nearest neighbor pattern classification. *IEEE Trans Inf Theory.* 1967;13(1):21–27. <https://doi.org/10.1109/TIT.1967.1053964>
22. Fisher RA. The use of multiple measurements in taxonomic problems. *Ann Eugen.* 1936;7(2):179–188. <https://doi.org/10.1111/j.1469-1809.1936.tb02137.x>
23. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B.* 1996;58(1):267–288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>
24. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B.* 2005;67(2):301–320. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>
25. Cortes C, Vapnik V. Support-vector networks. *Mach Learn.* 1995;20(3):273–297. <https://doi.org/10.1023/A:1022627411411>
26. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM; 2016. p. 785–794. <https://doi.org/10.1145/2939672.2939785>
27. Zeng D, Fang Y, Qiu W, Luo P, Wang S, Shen R, et al. Enhancing immuno-oncology investigations through multidimensional decoding of tumor microenvironment with IOBR 2.0. *Cell Rep Methods.* 2024;4(12):100910. <https://doi.org/10.1016/j.crmeth.2024.100910>
28. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration

- of cell subsets from tissue expression profiles. *Nat Methods*. 2015;12(5):453–457.
<https://doi.org/10.1038/nmeth.3337>
29. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al. Comprehensive integration of single-cell data. *Cell*. 2019;177(7):1888–1902.e21.
<https://doi.org/10.1016/j.cell.2019.05.031>
30. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods*. 2019;16(12):1289–1296. <https://doi.org/10.1038/s41592-019-0619-0>
31. Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol*. 2019;20(2):163–172. <https://doi.org/10.1038/s41590-018-0276-y>
32. Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, et al. SCENIC: single-cell regulatory network inference and clustering. *Nat Methods*. 2017;14(11):1083–1086. <https://doi.org/10.1038/nmeth.4463>
33. Silinskaite U, Valciukiene J, Jakubauskas M, Poskus T. The immune environment in colorectal adenoma: a systematic review. *Biomedicines*. 2025;13(3):699.
<https://doi.org/10.3390/biomedicines13030699>
34. Li Z, Wang D, Zhu X. Unveiling the functions of five recently characterized lncRNAs in cancer progression. *Clin Transl Oncol*. 2025;27(2):458–465.
<https://doi.org/10.1007/s12094-024-03619-w>
35. Kalmár A, Nagy ZB, Galamb O, Csabai I, Bodor A, Wichmann B, et al. Genome-wide expression profiling in colorectal cancer focusing on lncRNAs in the adenoma-carcinoma transition. *BMC Cancer*. 2019;19(1):1059.
<https://doi.org/10.1186/s12885-019-6180-5>
36. Yao F, Wang Q, Wu Q. The prognostic value and mechanisms of lncRNA UCA1 in human cancer. *Cancer Manag Res*. 2019;11:7685–7696.
<https://doi.org/10.2147/CMAR.S200436>

37. Wu B, Zhang Y, Yu Y, Zhong C, Lang Q, Liang Z, et al. Long noncoding RNA H19: a novel therapeutic target emerging in oncology via regulating oncogenic signaling pathways. *Front Cell Dev Biol.* 2021;9:796740. <https://doi.org/10.3389/fcell.2021.796740>
38. Schmitt M, Greten FR. The inflammatory pathogenesis of colorectal cancer. *Nat Rev Immunol.* 2021;21(10):653–667. <https://doi.org/10.1038/s41577-021-00534-x>
39. Zhong J, Qin Y, Yu P, Xia W, Gu B, Qian X, et al. The landscape of the tumor-infiltrating immune cell and prognostic nomogram in colorectal cancer. *Front Genet.* 2022;13:891270. <https://doi.org/10.3389/fgene.2022.891270>
40. Li W, Zhang H, You Z, Guo B. LncRNAs in immune and stromal cells remodel phenotype of cancer cell and tumor microenvironment. *J Inflamm Res.* 2024;17:3173–3185. <https://doi.org/10.2147/JIR.S460730>
41. Xiao X, Chen H, Yang L, Xie G, Shimizu R, Murai A. Concise review: cancer cell reprogramming and therapeutic implications. *Transl Oncol.* 2022;24:101503. <https://doi.org/10.1016/j.tranon.2022.101503>
42. Kadian LK, Verma D, Lohani N, Yadav R, Ranga S, Gulshan G, et al. Long non-coding RNAs in cancer: multifaceted roles and potential targets for immunotherapy. *Mol Cell Biochem.* 2024;479(12):3229–3254. <https://doi.org/10.1007/s11010-024-04933-1>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementarymaterial.docx](#)