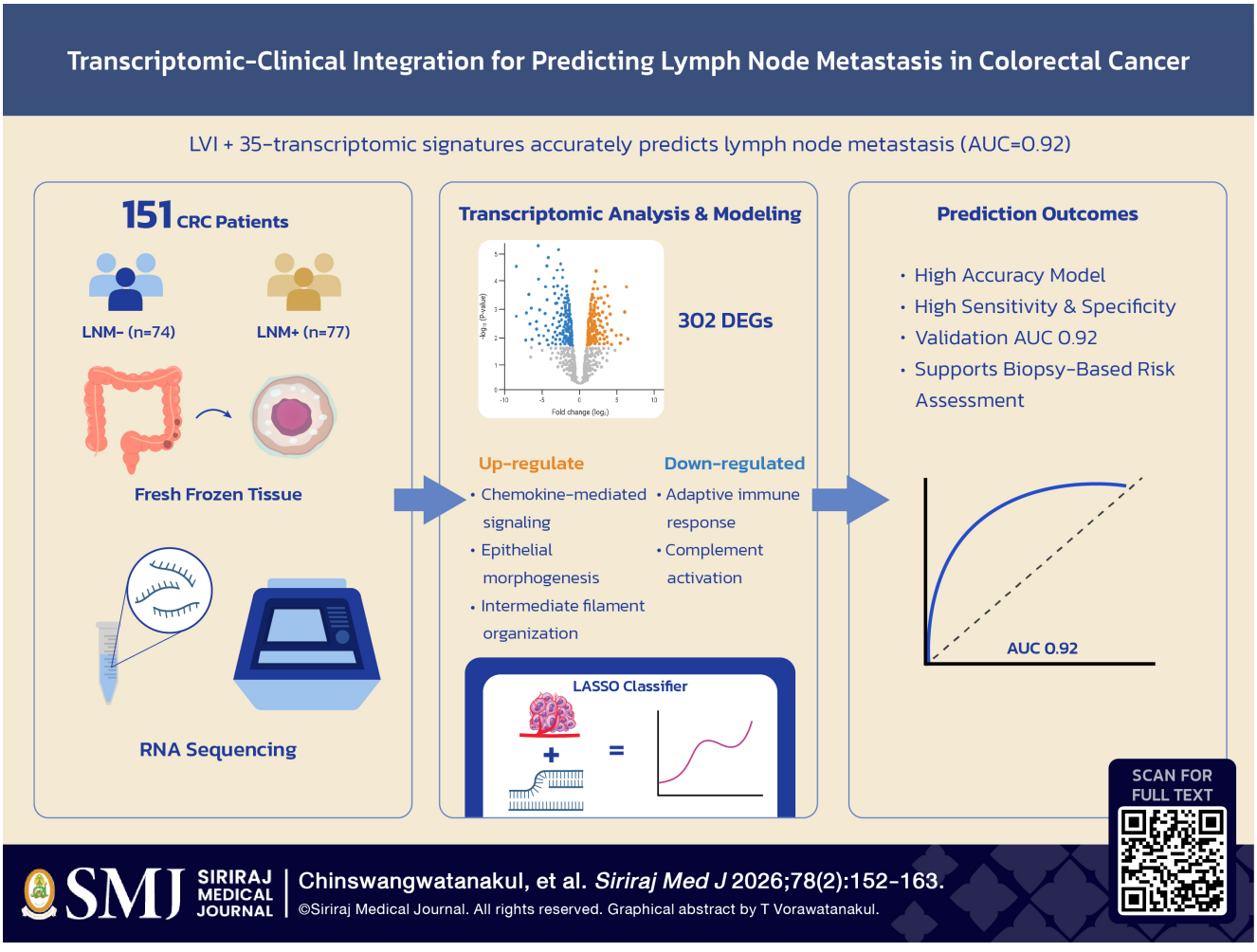
Combining Histopathologic and Gene-Expression Profiling for Risk Stratification of Nodal Metastasis in Colorectal Cancer

Watsaphon Tangkullayanone, M.D.[1,2], Nutchavadee Vorasan, M.Sc.[3], Amphun Chaiboonchoe, Ph.D.[4], Atthaphorn Trakarnsanga, M.D.[2], Pariyada Tanjak, Ph.D.[2,5], Thanawat Suwatthanarak, Ph.D.[2,5], Woramin Riansuwan, M.D.[2], Kullanist Thanormjit, M.Sc.[2,5], Onchira Acharayothin, B.Sc.[2], Asada Methasate, M.D., Ph.D.[2], Yusuke Kinugasa, M.D., Ph.D.[8], Bhoom Suktitipat, M.D., Ph.D.[6,7], Vitoon Chinswangwatanakul, M.D., Ph.D.[2,5,*]

[1]Graduate School of Medical and Dental Sciences, Joint Degree Doctoral Program in Medical Sciences between Institute of Science Tokyo, Japan and Mahidol University, Bangkok, Thailand, [2]Division of General Surgery, Department of Surgery, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand, [3]Genetic Epidemiology Research Cluster, Research Division, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand, [4]Siriraj Center of Research Excellent for Systems Pharmacology, Department of Pharmacology, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand, [5]Siriraj Cancer Center, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand, [6]Department of Biochemistry, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand, [7]Integrative Computational BioScience (ICBS) Center, Mahidol University, Nakhon Pathom, Thailand, [8]Gastrointestinal Surgery, Institute of Science Tokyo, Bunkyo-ku, Japan.



*Corresponding author: Vitoon Chinswangwatanakul
E-mail: vitoon.chi@mahidol.ac.th
Received 31 December 2025 Revised 13 January 2026 Accepted 13 January 2026
ORCID ID:http://orcid.org/0000-0001-9662-1669
https://doi.org/10.33192/smj.v78i2.279649

ABSTRACT

Objective: To identify gene-expression features associated with lymph node metastasis (LNM) in colorectal cancer (CRC) and to develop a transcriptomic-clinical predictive model for preoperative nodal assessment.

Materials and Methods: A total of 151 CRC tissue samples (74 LNM– and 77 LNM+) were analyzed using RNA sequencing. Differentially expressed genes (DEGs) were identified with DESeq2, and functional enrichment analyses were performed using the Database for Annotation, Visualization, and Integrated Discovery (DAVID). A Least Absolute Shrinkage and Selection Operator (LASSO) logistic regression model integrating gene-expression features with clinical variables was developed to predict LNM status. Model performance was evaluated using the area under the receiver operating characteristic curve (AUC), sensitivity, and specificity.

Results: A total of 302 DEGs were identified in LNM+ CRC, including 178 upregulated and 124 downregulated genes. Upregulated genes were enriched in chemokine-mediated signaling, epithelial morphogenesis, and intermediate filament organization, whereas downregulated genes were associated with adaptive immune response and complement activation. In multivariate analysis, lymphovascular invasion (LVI) was the only clinical variable independently associated with LNM. The optimized LASSO model, combining LVI with selected transcriptomic features, demonstrated excellent discriminatory performance (AUC ≈ 0.92). Key upregulated genes included *CCL21, CCL26, DEFB1, LST1, KANK4, TNNC1, PFDN6, TENM1, CST6,* and *PADI3,* while *IGHV2-26* was downregulated.

Conclusion: Integration of LVI with transcriptomic signatures enables accurate prediction of lymph node metastasis in CRC and supports biopsy-based risk assessment to guide clinical decision-making.

Keywords: Colorectal neoplasm; Lymphatic metastasis; RNA sequencing; Gene expression regulation; Logistic Models (Siriraj Med J 2026;78(2):152-163)

INTRODUCTION

Colorectal cancer (CRC) is one of the most prevalent malignancies of the gastrointestinal tract and ranks among the top three most commonly diagnosed cancers globally as reported in the GLOBOCAN 2020 database.[1] The global incidence of CRC is approximately 19.5 per 100,000 population and continues to increase despite advances in screening programs and treatment modalities. Mortality remains significant, with an estimated rate of 9 deaths per 100,000 population, underscoring the growing clinical burden of the disease.

The five-year overall survival rate for screen-detected CRC is approximately 83.4%.[2] Survival outcomes in CRC are strongly influenced by clinical staging, which is determined by tumor invasion depth (T stage) and lymph node involvement (N stage). According to the AJCC 8th edition[3], the presence of lymph node metastasis (LNM) marks the transition to Stage III, also known as locally advanced disease. While patients with Stage I and II CRC typically have excellent five-year survival rates—approaching 90%—this rate drops to around 80% in Stage III and falls below 50% in Stage IV disease. At Siriraj Hospital in Thailand, reported five-year survival rates for CRC are 89.1% for Stage I, 78.6% for Stage II, and 57.9% for Stage III disease.[4]

In clinical practice, LNM in CRC is evaluated using a combination of tumor characteristics and imaging findings. First, the depth of primary tumor invasion (T stage), typically assessed by endoscopic examination and preoperative imaging, is a key determinant of nodal involvement, with reported LNM rates increasing from 14.3% in T1 tumors to 25.6% in T2, 61.2% in T3, and 65.6% in T4 disease.[5] Second, direct assessment of lymph node status is performed using preoperative imaging, most commonly computed tomography (CT), which infers nodal metastasis based on lymph node size (>9 mm), morphology, margin characteristics, and nodal clustering. However, the diagnostic accuracy of CT remains limited (approximately 60–70%) and is subject to both false positive and false negative results.[6] Third, pathological features obtained from biopsy specimens, particularly lymphovascular invasion (LVI), are used as indicators of nodal spread. Although LVI is associated with an increased risk of LNM, its predictive value is constrained by the fact that histological LVI is identified in only about one-third of patients with confirmed nodal metastasis.[7,8]

Given the limitations of current diagnostic approaches, there is growing interest in integration of molecular biomarkers into the staging process. Transcriptomic profiling, particularly through RNA sequencing, has emerged as a powerful tool for identifying gene expression patterns associated with metastatic behavior. The differentially expressed genes (DEGs) analysis may reveal molecular

signatures linked to lymph node involvement, offering a path to more precise risk stratification.

In this study, we aim to characterize the biological functions and signaling pathways of DEGs and to establish a predictive model for LNM in CRC. By integrating RNA-sequencing–based gene expression profiles with relevant clinical variables, we applied the Least Absolute Shrinkage and Selection Operator (LASSO) regression to identify genes most strongly correlated with nodal involvement. The resulting gene signature is intended to improve the preoperative risk assessment and facilitate personalized management of patients with CRC.

MATERIALS AND METHODS

Patients aged 18 years or older with a histopathologically confirmed diagnosis of CRC who underwent upfront surgical resection at Siriraj Hospital between 2011 and 2024 were included. Diagnosis and preoperative staging were established by colonoscopic biopsy and contrast-enhanced CT imaging. During surgery, representative primary tumor tissue was collected and preserved as fresh frozen samples, while the entire specimen underwent standard pathological assessment. Patients were excluded if they had suspected hereditary CRC syndromes, multiple synchronous primary CRCs, received neoadjuvant chemotherapy or radiotherapy prior to surgery, or presented with distant metastatic disease at diagnosis, as defined by the American Joint Committee on Cancer (AJCC) Cancer Staging Manual, 8th Edition.[9]

There are 207 CRC cases in the tissue bank. After applying exclusion criteria, 2 samples from patients who had received neoadjuvant chemoradiation, 3 samples with synchronous tumors, and 9 samples with confirmed hereditary CRC were excluded. As this study focused on LNM, 40 stage IV samples were excluded. RNA-sequencing quality control identified two additional outlier samples, which were removed from the analysis. The final cohort comprised 151 samples, which 74 were the LNM-negative (LNM-) group and 77 were the LNM-positive (LNM+) group.

Total RNA was extracted from fresh CRC tissue specimens preserved in RNAlater using a previously described protocol[10] and purified with the RNeasy Mini Kit (Qiagen). RNA concentration, purity, and integrity were assessed using NanoDrop spectrophotometry and the Agilent 2100 Bioanalyzer; only samples with RNA integrity numbers (RIN) >7 and total RNA yield >1 μg were included. Messenger RNA was enriched using poly-T oligo–attached magnetic beads, followed by strand-specific library construction with dUTP incorporation. Library quality was evaluated by Qubit, qPCR, and Bioanalyzer profiling. Paired-end sequencing (2 × 150 bp) was performed on an Illumina NovaSeq platform, generating approximately 8 Gb of data per sample. Library preparation and sequencing were outsourced to Novogene Co. Ltd. (Singapore).

Statistical analysis

RNA-sequencing data pre-processing and transcript abundance quantification

Raw RNA-sequencing data (FASTQ format) were assessed for quality using FastQC. Adapter sequences and low-quality reads were trimmed using FastP, followed by post-filtering quality verification with FastQC and summary reporting using MultiQC. Quality-controlled reads were quantified using Salmon with pseudo-alignment against the GRCh38 human reference transcriptome. Transcript-level abundances were summarized at the gene level using the R package TXimport.[11-15]

Quality control and normalization

RNA-sequencing data from primary CRC tumors comprising 37,788 genes were analyzed. Quality control at the expression level was assessed using Relative Log Expression (RLE) plots. Samples whose median RLE deviated substantially from zero were considered outliers and excluded before analysis, yielding the final dataset used for downstream procedures. Low-count genes were removed before normalization and testing. We retained genes with ≥10 total reads across all samples and detectable expression in >30 samples. Normalization was performed using the median-of-ratios method in DESeq2, and variance stabilizing transformation (VST) was subsequently applied to normalized counts for downstream analyses.

DEGs and predictive modeling

The DEGs between patients with lymph node–positive and lymph node–negative CRC were tested using the Wald test in DESeq2 under a negative binomial framework. Multiple testing was controlled with the Benjamini–Hochberg false discovery rate (FDR); genes with FDR-adjusted $p < 0.05$ were considered statistically significant. For interpretability, we prespecified a threshold of $|\log_2 \text{fold change}| > 1$ to denote biologically meaningful effects.

Functional enrichment of the resulting 302 DEGs was conducted using the Database for Annotation, Visualization, and Integrated Discovery (DAVID).[16] Enrichment was assessed on multiple databases, including Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), WikiPathways, and Reactome.

Statistical significance thresholds were set at p value < 0.05 and q value < 0.15 after correction for multiple testing by the Benjamini–Hochberg false discovery rate (FDR) method. Given the exploratory nature of pathway enrichment analysis, a less stringent FDR threshold (q < 0.15) was applied to identify biologically relevant pathways while minimizing false negative findings.

For downstream visualization and predictive modeling, the data were subsequently standardized to z-scores (per gene); when specified, the means and standard deviations (SDs) were estimated from the lymph node–negative reference group.

To develop a classifier for lymph node status, we used the LASSO logistic regression (glmnet package). The dataset was randomly partitioned into training (70%), testing (20%), and hold-out validation (10%) sets with stratification by outcome. The regularization parameter (λ) was selected via 10-fold cross-validation within the training set using the one-standard-error (1-SE) rule when applicable. Model performance was quantified in test and validation sets using the area under the receiver operating characteristic curve (AUC), sensitivity, and specificity. All modeling was performed in R 4.5.0.

We compared three regularized logistic regression approaches including Ridge (L2), Elastic Net ($\alpha \in [0,1]$), and LASSO (L1) using identical variable and the same outcome-stratified train/test/validation splits. For Elastic Net, α was tuned on a grid while λ was selected by 10-fold cross-validation. Model selection prioritized discrimination (AUC) and calibration on the test set, with parsimony considered when performance was comparable.

RESULTS

Baseline clinicopathological characteristics are summarized in Table 1. A total of 151 CRC samples were analyzed, including 74 LNM− and 77 LNM+ cases. Age, sex, and tumor location did not differ significantly between groups. In contrast, tumor invasion depth (T stage) was significantly higher in the LNM+ group (p < 0.001), with T3–T4 tumors predominating. LVI and perineural invasion (PNI) were strongly associated with LNM (both p < 0.001) and were largely absent in LNM− cases. Tumor differentiation was comparable between groups, with most tumors being moderately differentiated. Overall, T stage, LVI, and PNI were significantly associated with the presence of LNM, whereas other clinical features showed no significant differences.

Quality control of RNA-sequencing data from all CRC samples (37,788 genes) was performed using Relative Log Expression (RLE) plots (Fig S1).[17-19] Two samples with median RLE values deviating substantially from zero were excluded, leaving 151 samples for downstream analyses.

After sample QC and low-count filtering (retaining genes with ≥10 total reads and expression in >30 samples; 17,906 genes remained), differential expression was assessed in DESeq2. Using median-of-ratios normalization and Wald testing, we identified 302 DEGs in the primary tumors from the patients with lymph node–positive vs lymph node–negative CRC at FDR-adjusted p < 0.05 and $|\log_2$ fold change$| > 1$. Of these, 178 genes were upregulated and 124 were downregulated in lymph node–positive tumors. The overall distribution of effect sizes and significance is shown in the volcano plot (Fig 1) and the complete DEG list provided in Table S1.

Functional enrichment analysis of these DEGs using DAVID revealed distinct biological signatures between the two groups, as summarized in Table 2. The upregulated genes were mainly associated with epithelial morphogenesis, intermediate filament organization, and chemokine-mediated signaling, whereas the downregulated genes were enriched for pathways related to adaptive immune response and complement cascade.

For prediction, the 151 samples were partitioned with outcome stratification into training (n = 106; LNM−group = 51, LNM+ group = 55), test (n = 30; LNM−group = 18, LNM+ group = 12), and hold-out validation (n = 15; LNM− group = 5, LNM+ group = 10) sets (Fig 2). The candidate predictor set comprised 302 genes identified from the DEGs analysis (DESeq2; VST-transformed expression values standardized to per-gene z-scores; see Fig S2), along with the clinical covariates sex, age, primary tumor location, T stage, LVI, PNI, and tumor differentiation. Ten-fold cross-validation selected the penalty that minimized binomial deviance at λ_min = 0.01055471 ($\log \lambda \approx -4.55$) (Fig 3), yielding a 36-variable LASSO model (35 genes + LVI) with non-zero coefficients associated with lymph node positivity. Using this λ_min model, the classifier trained on per-gene z-standardized VST expression demonstrated excellent discrimination: in the training set, AUC was 1.00 with 100.0% sensitivity, 100.0% specificity, and 100.0% accuracy; in the independent test set, AUC was 0.93 with 91.7% sensitivity, 88.9% specificity, and 90.0% accuracy; and in the hold-out validation set, AUC was 0.92 with 90.0% sensitivity, 100.0% specificity, and 93.3% accuracy (Fig 4). For completeness, the one-standard-error solution is reported in Table 3 as a parsimonious comparator, but all primary performance estimates refer to the λ_min (35 genes + LVI) model. Non-zero coefficients for the λ_min model are listed in Table 3.

We compared LASSO, Ridge, and Elastic Net logistic

TABLE 1. Baseline characteristics of the patients according to lymph node status.

Characteristic	LNM− (n=74)	LNM+ (n=77)	p-values
Age, yr (mean ±SD)	66.89 ± 12.70	63.92 ± 11.60	0.135
Male, n (%)	44 (59.5)	38 (49.4)	0.213
Primary tumor location, n (%)			0.332
Right colon	11 (14.9)	10 (13.0)	
Transverse colon	4 (5.4)	2 (2.6)	
Left colon	31 (41.9)	25 (32.5)	
Rectum	28 (37.8)	40 (51.9)	
T stage, n (%)			<0.001
T1	6 (8.1)	1 (1.3)	
T2	31 (41.9)	9 (11.7)	
T3	32 (43.2)	53 (68.8)	
T4a	3 (4.1)	9 (11.7)	
T4b	2 (2.7)	5 (6.5)	
LVI, n (%)			<0.001
No	69 (93.2)	48 (62.3)	
Yes	5 (6.8)	29 (37.7)	
PNI, n (%)			<0.001
No	68 (91.9)	53 (68.8)	
Yes	6 (8.1)	24 (31.2)	
Tumor differentiation, n (%)			0.126
Well differentiated	14 (18.9)	5 (6.5)	
Moderately differentiated	58 (78.4)	67 (87.0)	
Poorly differentiated	1 (1.4)	1 (1.3)	
Signet-ring cell	0 (0)	2 (2.6)	
Mucinous	1 (1.4)	2 (2.6)	

*p-values are based on the t-test for continuous variables and the χ^2 test for categorical variables.

regression using identical features and outcome-stratified train/test/validation splits. LASSO provided the most balanced and consistent discrimination across held-out sets, achieving an AUC of 0.93 in the test set (sensitivity 91.7%, specificity 88.9%) and 0.92 in the validation set (sensitivity 90.0%, specificity 100.0%), while maintaining perfect training performance (AUC 1.00) (Fig 4). Ridge regression yielded a slightly higher test AUC (0.95) but produced a non-sparse solution that retained all features; test specificity (88.9%) and sensitivity (100.0%) were comparable, and validation AUC was 1.00 (sensitivity 90.0%, specificity 100.0%) (Fig S4). Elastic Net produced a test AUC of 0.92 with sensitivity (91.7%) at specificity (83.3%), and a validation AUC of 0.92 (sensitivity 90.0%, specificity 100.0%) (Fig S6). Considering sensitivity, specificity, and AUC jointly on the independent test and validation sets together with model sparsity and interpretability we selected LASSO as the final classifier.

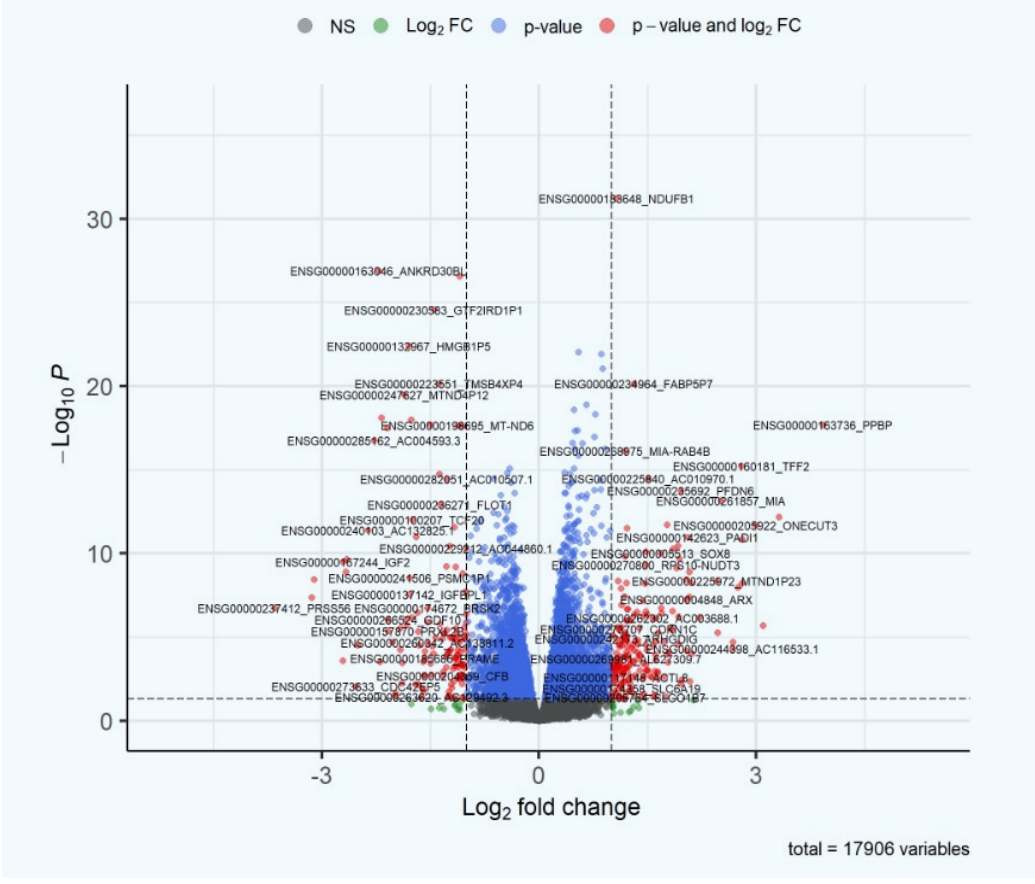


Fig 1. Volcano plot displaying differential expression between lymph node–positive and lymph node–negative CRC (DESeq2, Wald test). The x-axis shows $\log_2$ fold change (LNM+ vs LNM–) and the y-axis shows $-\log_{10}$ (FDR-adjusted P). The points in red indicate significantly DEGs ($|\log_2 FC| > 1$, FDR < 0.05; n = 302), and the points in blue represent genes with FDR < 0.05 but fold changes within ±1 (less than two-fold change). Non-significant genes are shown in grey.

TABLE 2. Functional Pathways and Representative Genes in LNM+ CRC.

Upregulation in LNM+ CRC	Representative Genes	Downregulation in LNM+ CRC	Representative Genes
Pancreatic cancer subtypes	SCEL, TFF2, KRT7, CTSE, CST6, KRT6A	Adaptive immune response	FGB, FGA, IGHV3-72, IGKV3-7, IGHV2-26, IGKV1D-13, IGKV1D-12, IGHG2, IGKV6D-21, TRBC2, IL17F, IL17A, IGKV2D-24
Morphogenesis of the epithelium	KRT16, SOX8, SOX10, KRT6A	Complement system	FGB, FGA, CFHR2, CFB, CPN1, C2
Chemokine-mediated signaling	CCL13, CCL21, TFF2, PPBP, CCL26	Regulation of complement cascade	IGHG2, CFHR2, CFB, IGKV1D-12, CPN1, C2
Intermediate filament organization	KRT16, KRT7, KRT6B, KRT6A, PRPH	Complement cascade	IGHG2, CFHR2, CFB, IGKV1D-12, CPN1, C2
Cell maturation	CCL21, REN, SOX8, SOX10	Complement and coagulation cascades	FGB, FGA, CFHR2, CFB, C2

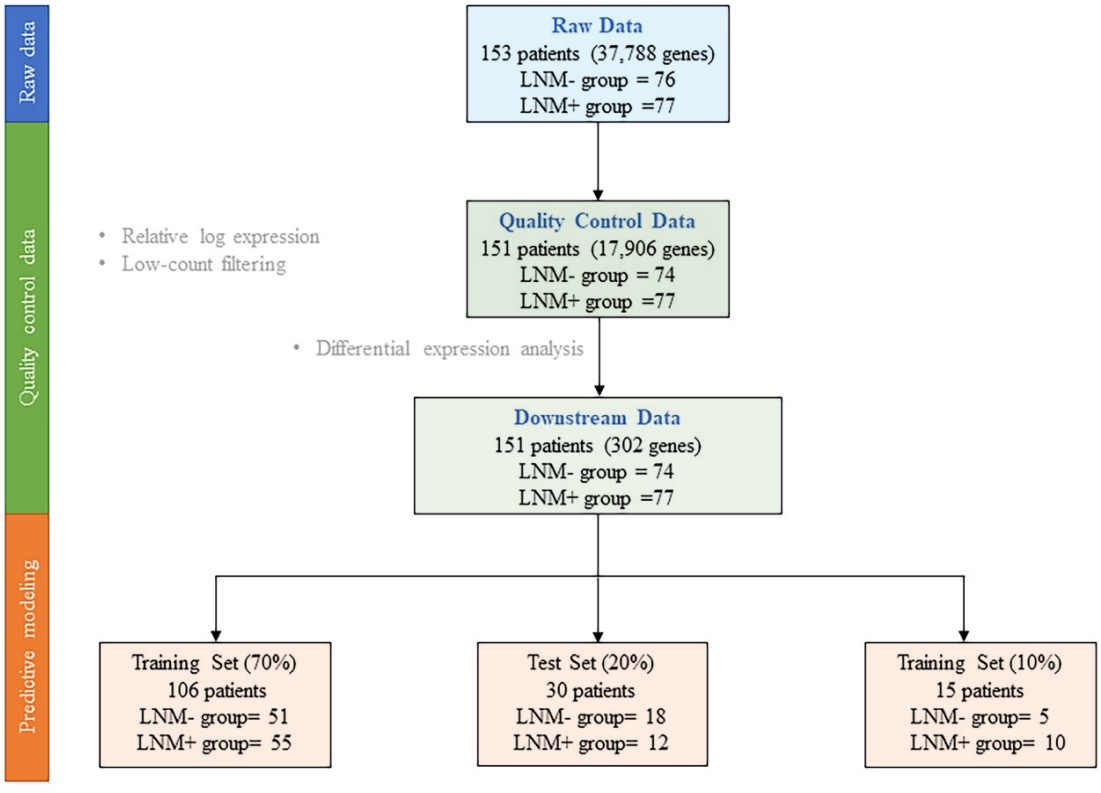


Fig 2. Diagram dataset split for predictive modeling

The full cohort (n = 151 tumors; 17,906 genes after low-count filtering) was partitioned by outcome into training (70%, n = 106; LNM−group = 51, LNM+ group = 55), test (20%, n = 30; LNM−group = 18, LNM+-group = 12), and hold-out validation (10%, n = 15; LNM− group = 5, LNM+ group = 10) sets using stratified random sampling.

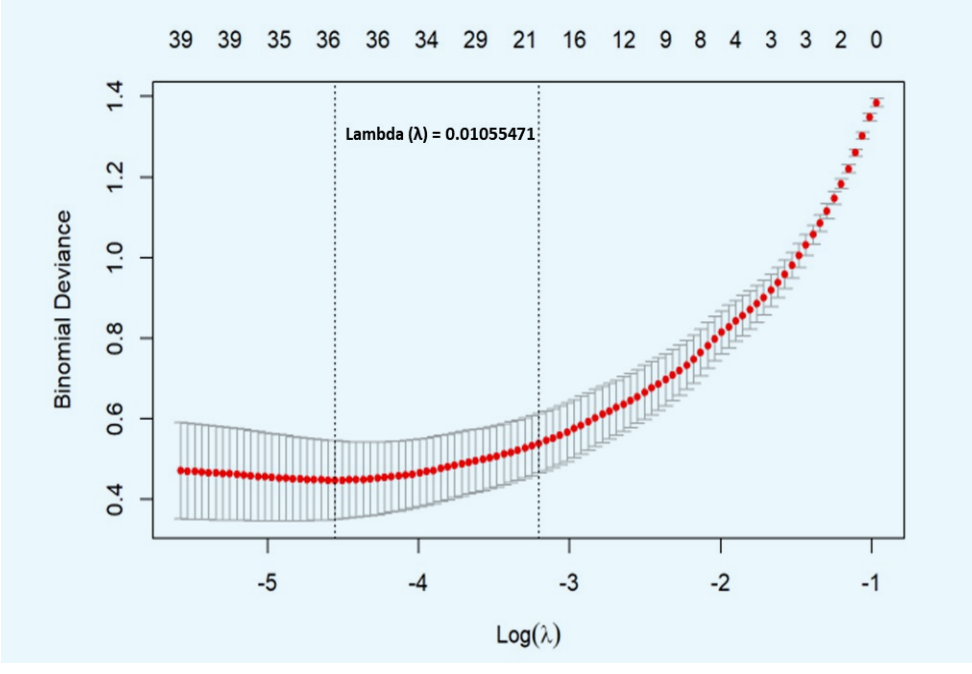


Fig 3. Ten-fold cross-validation for the LASSO logistic model.

Ten-fold cross-validation curve showing mean binomial deviance (red points) ± 1 SE (gray bars) across log(λ) values. The vertical dotted lines mark λ_min (left) and λ_1SE (right). The selected penalty at λ_min = 0.01055471 (log λ ≈ −4.55) minimizes deviance. The numbers along the top indicate the number of non-zero coefficients in the model at each λ.

TABLE 3. Final LASSO genes and clinical variable to predict lymph-node positive status.

Type	Regulation	Ensembl ID	Gene symbol/variable	Coefficient (β)
Clinical			LVI	1.367
Gene	Downregulation	ENSG00000285162	AC004593.3	-0.659
Gene		ENSG00000244694	PTCHD4	-0.396
Gene		ENSG00000205578	POM121B	-0.278
Gene		ENSG00000066032	CTNNA2	-0.266
Gene		ENSG00000143627	PKLR	-0.225
Gene		ENSG00000170892	TSEN34	-0.153
Gene		ENSG00000177238	TRIM72	-0.151
Gene		ENSG00000262771	SSBP1	-0.124
Gene		ENSG00000196656	AC004057.1	-0.096
Gene		ENSG00000167244	IGF2	-0.057
Gene		ENSG00000147255	IGSF1	-0.049
Gene		ENSG00000135744	AGT	-0.044
Gene		ENSG00000282344	IGHV2.26	-0.042
Gene		ENSG00000259848	AC097374.1	-0.025
Gene	Upregulation	ENSG00000270136	MINOS1.NBL1	0.551
Gene		ENSG00000174473	GALNTL6	0.526
Gene		ENSG00000276725	CEP170	0.519
Gene		ENSG00000226182	LST1	0.438
Gene		ENSG00000267022	AC067968.1	0.429
Gene		ENSG00000278622	TSEN34	0.392
Gene		ENSG00000006606	CCL26	0.290
Gene		ENSG00000114854	TNNC1	0.284
Gene		ENSG00000204542	C6orf15	0.280
Gene		ENSG00000280571	AC006059.2	0.270
Gene		ENSG00000283707	AC275455.1	0.239
Gene		ENSG00000175315	CST6	0.151
Gene		ENSG00000132854	KANK4	0.132
Gene		ENSG00000268975	MIA.RAB4B	0.096
Gene		ENSG00000137077	CCL21	0.073
Gene		ENSG00000164825	DEFB1	0.065
Gene		ENSG00000142619	PADI3	0.057
Gene		ENSG00000009694	TENM1	0.054
Gene		ENSG00000206283	PFDN6	0.042
Gene		ENSG00000234964	FABP5P7	0.027
Gene		ENSG00000178343	SHISA3	0.022

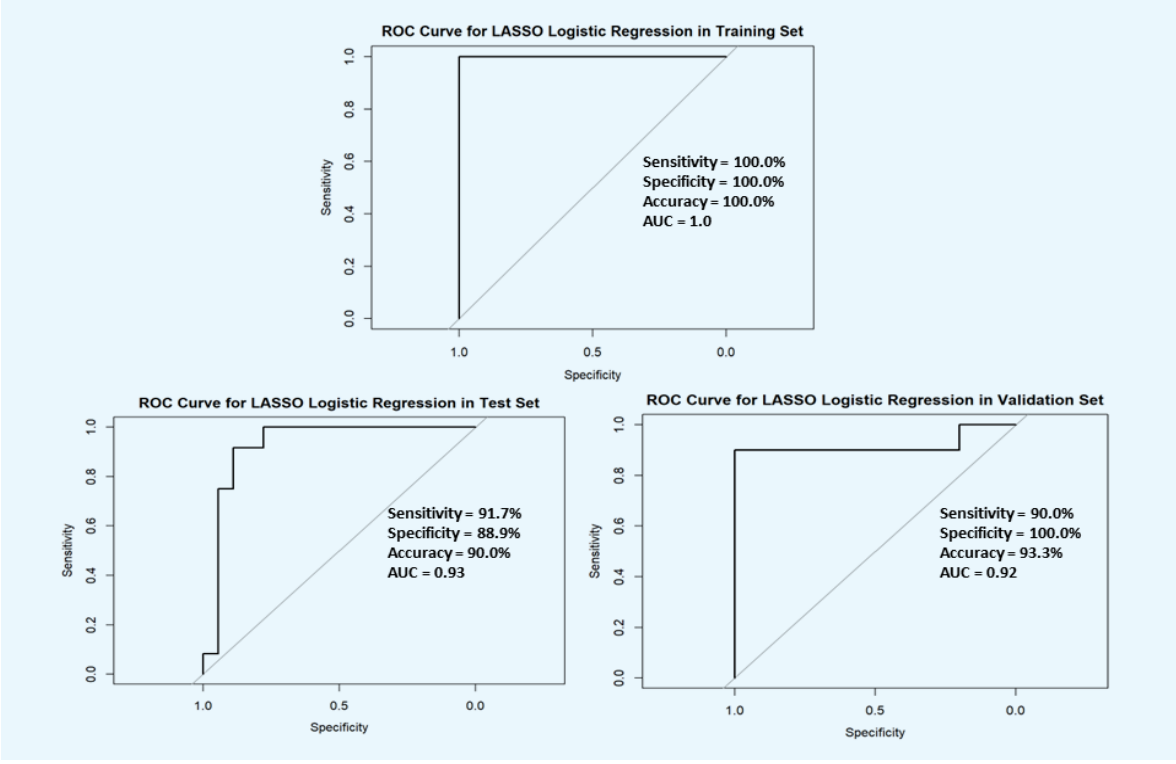


Fig 4. ROC curves for the 35 genes + LVI LASSO classifier
Receiver operating characteristic (ROC) curves for the LASSO logistic regression model selected at λ_min = 0.01055471, evaluated in the training (top), test (bottom-left), and hold-out validation (bottom-right) sets. Performance summaries are shown within each panel.

DISCUSSION

LVI is a well-established histopathological marker of aggressive colorectal cancer and was significantly associated with lymph node metastasis (LNM) in our cohort. After integration into a multivariate LASSO model with transcriptomic variables, LVI remained the only independent pathological predictor, underscoring its role as a critical early step in lymphatic dissemination. Functional analysis of 302 differentially expressed genes revealed that LNM-positive tumors were enriched for pathways related to chemokine signaling, epithelial morphogenesis, and cytoskeletal organization, while downregulated genes were associated with adaptive immune responses and complement activation. These findings suggest that nodal metastasis in CRC arises from coordinated epithelial remodeling alongside suppression of antitumor immune mechanisms.

Among these, Sciellin (SCEL) promotes cancer cell stiffness and tumor colonization, partly through activation of the Wnt/β-catenin pathway and enhancement of mesenchymal-to-epithelial transition (MET). Elevated SCEL expression in late-stage CRC suggests its involvement in tumor progression.[20] Chemokine-mediated signaling directs lymphocyte trafficking to lymph nodes, and it has been hypothesized that cancer cells may exploit this mechanism to invade lymphatic tissue.[21] In this pathway, CCL21 and CCL26 (C-C motif chemokine ligand 21 and 26) are implicated. CCL21 promotes nodal spread through matrix metalloproteinase-9 (MMP9) activation and extracellular matrix remodeling[22,23], thereby enhancing cancer cell migration.[24] Additionally, increased expression of epithelial markers such as Keratin (KRT) or cytokeratin (CK) has been associated with tumor progression, increased migratory capacity, and epithelial–mesenchymal transition (EMT).[25] The transcription factor SOX8 (SRY-box) further contributes to aggressive tumor behavior by activating Wnt/β-catenin signaling, leading to enhanced proliferation, reduced apoptosis, and increased EMT activity.[26]

Tumor-infiltrating B lymphocytes (TIBLs) are associated with favorable outcomes in CRC, and downregulation of IGHV2-26 may reflect reduced immune infiltration and enhanced immune evasion.[27] Additionally, diminished expression of the complement cascade may impair immune surveillance and complement-dependent cytotoxicity, facilitating LNM.[28]

The LASSO regression model identified a concise panel of 35 genes together with LVI as an independent clinical predictor of lymph-node metastasis. Although the selected genes did not fully overlap with the DAVID-

enriched set, many shared functional relevance with pathways involved in chemokine signaling, epithelial remodeling, cytoskeletal dynamics, and immune regulation. This complementarity reflects the distinct aims of the two analyses: DAVID reveals global biological themes, while LASSO isolates the smallest gene set that best predicts metastatic potential.

Among the upregulated genes, several were linked to chemokine-mediated signaling and immune modulation. CCL21 and CCL26 may guide tumor-cell migration toward lymphatic channels through MMP9-dependent matrix remodeling. DEFB1 (Defense beta1), an innate-immunity gene with context-specific oncogenic or suppressive roles, correlates with immune-checkpoint activation and poorer outcomes.[29,30] LST1 (Leukocyte specific transcript 1) mediates inflammatory crosstalk between tumor and stromal cells, promoting invasion and proliferation.[31,32]

KANK4 (KN motif and ankyrin repeat domains 4)[33,34], TNNC1 (Troponin C1)[35-37], PFDN6 (Prefoldin 6)[38], and TENM1 (Teneurins 1)[39], regulate cytoskeletal organization and epithelial structure, supporting increased motility and adhesion changes characteristic of metastatic cells.

PADI3 (Protein arginine deaminase 3)[40,41] and CST6 (Cystatin 6)[42,43] participate in EMT control; their dysregulation may favor metastatic colonization. GALNT6 (Polypeptide N-acetylgalactosaminyltransferase 6)[44] promotes proliferation and migration by altering mucin glycosylation and epithelial polarity, while TSEN34 and CEP170[45]—involved in RNA processing and cell-division control—likely reflect heightened proliferative activity in advanced tumors.

Among the downregulated genes, several act as tumor suppressors or immune mediators. TRIM72 (Tripartite motif containing 72)[46] and PTCHD4 (Patched domain containing 4)[47,48], a negative regulator of Hedgehog (HH) signaling. Loss of Patched-mediated inhibition may lead to HH signaling activation and promote EMT induction. Reduced CTNNA2 (Catenin alpha 2) disrupts epithelial adhesion and favors mesenchymal transformation.[49] The suppression of IGHV2-26, a key immunoglobulin heavy-chain gene, mirrors the weakened adaptive-immune and complement pathways observed in the DAVID analysis.

Overall, our findings highlight the central role of LVI in allowing tumor cells to access the lymphatic system and spread to lymph nodes. Lymph-node–positive CRC demonstrated increased chemokine signaling, cytoskeletal remodeling, EMT, and reduced immune activity, which are biological processes that support metastatic progression. The genes selected by the LASSO model represent these pathways and together form an expression pattern that can help estimate the likelihood of nodal involvement. This study has strengths, including the integration of gene-expression data with a routinely assessed pathological feature, LVI, producing a model that may be usable in real clinical decision-making. However, the study also has limitations. Samples were obtained from a single center, and some genes in the model have not yet been well studied in CRC. Some histopathologic features incorporated into the model, including tumor budding and LVI, are known to be subject to inter-observer variability. These parameters were derived from routine pathology reports assessed by a single pathologist, rather than independent review by multiple observers. This may introduce variability in feature classification and represents a limitation of the study. The proposed model is intended for preoperative risk stratification using colonoscopic biopsy specimens, all transcriptomic data were generated from surgically resected tumor tissues. Therefore, external validation using matched biopsy and resection samples will be essential to determine the generalizability of this model and to establish its applicability for preoperative clinical decision-making. In addition, future studies should evaluate whether a targeted expression panel derived from the model genes, such as RT-qPCR- or immunohistochemistry (IHC)-based assay can maintain predictive performance.

CONCLUSIONS

Combining LVI with transcriptomic profiling provides an effective approach to predict lymph-node metastasis in CRC. The resulting gene-based model reflects coordinated activation of chemokine signaling, cytoskeletal remodeling, and EMT, accompanied by immune suppression, and demonstrates strong potential for early nodal risk assessment. However, further validation in preoperative biopsy samples is required to confirm its real-world clinical utility. In addition, optimization of the gene set into a targeted, cost-effective assay suitable for routine pathology workflows will be necessary to support future clinical translation.

Data Availability Statement

The authors affirm that the data supporting the findings of this study are included within the article and its supplementary materials.

ACKNOWLEDGEMENTS

The authors thank all individuals and departments who contributed to this study and are grateful to Dr. Mark Simmerman for English language editing.

DECLARATIONS

Grants and Funding Information

This research study was funded by (i) Health Systems Research Institute (HSRI) of Thailand (63-117, 66-083) and (ii) Foundation for Cancer Care, Siriraj Hospital (R016241047).

Conflict of Interest

The authors declare no competing interests.

Registration Number of Clinical Trial

Not applicable.

Author Contributions

Conceptualization and methodology, W.T., P.T., T.S., B.S., and V.C. ; Investigation, W.T., P.T., T.S., A.T., W.R., A.M., K.T., and O.A. ; Formal analysis, W.T., A.C., P.T., T.S., N.V., and B.S. ; Visualization and writing – original draft, W.T. and N.V. ; Writing – review and editing, P.T., T.S., B.S., Y.K., and V.C. ; Funding acquisition, P.T. and V.C. ; Supervision, V.C. All authors have read and agreed to the final version of the manuscript.

Use of Artificial Intelligence

No artificial intelligence tools or technologies were utilized in the writing, analysis, or development of this research.

Supplementary Information

The online version contains supplementary material are available from the corresponding author on reasonable request.

Ethical Approval

This study was approved by the Siriraj Institutional Review Board (COA No. Si 105/2021 and Si 156/2011) and conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from all participants.

REFERENCES

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021;71(3):209-49.
2. Cardoso R, Guo F, Heisser T, De Schutter H, Van Damme N, Nilbert MC, et al. Overall and stage-specific survival of patients with screen-detected colorectal cancer in European countries: A population-based study in 9 countries. Lancet Reg Health Eur. 2022;21:100458.
3. Weiser MR. AJCC 8th Edition: Colorectal Cancer. Ann Surg Oncol. 2018;25(6):1454-5.
4. Mongkhonsupphawan A, Sethalao N, Riansuwan W. Long-term Oncologic Outcomes After Curative Surgery in Stage I–III Thai Colorectal Cancer Patients. Siriraj Med J. 2022;74(11):739-46.
5. Lohsiriwat V, Rungteeranont C, Saigosoom N. Incidence and Pattern of Nodal Metastasis in Colon and Rectal Cancer: a Study of 1,012 Cases from Thailand. Siriraj Med J. 2020;72(5):386-90.
6. Hong EK, Landolfi F, Castagnoli F, Park SJ, Boot J, Van den Berg J, et al. CT for lymph node staging of Colon cancer: not only size but also location and number of lymph node count. Abdom Radiol (NY). 2021;46(9):4096-105.
7. Wang X, Cao Y, Ding M, Liu J, Zuo X, Li H, Fan R. Oncological and prognostic impact of lymphovascular invasion in Colorectal Cancer patients. Int J Med Sci. 2021;18(7):1721-9.
8. Zhang L, Deng Y, Liu S, Zhang W, Hong Z, Lu Z, et al. Lymphovascular invasion represents a superior prognostic and predictive pathological factor of the duration of adjuvant chemotherapy for stage III colon cancer patients. BMC Cancer. 2023;23(1):3.
9. Weiser MR. AJCC 8th Edition: Colorectal Cancer. Ann Surg Oncol. 2018;25(6):1454-5.
10. Acharayothin O, Thiengtrong B, Juengwiwattanakitti P, Anekwiang P, Riansuwan W, Chinswangwatanakul V, et al. Impact of Washing Processes on RNA Quantity and Quality in Patient-Derived Colorectal Cancer Tissues. Biopreserv Biobank. 2023;21(1):31-7.
11. Bioinformatics B. FastQC: a quality control tool for high throughput sequence data. 2011.
12. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32(19):3047-8.
13. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14(4):417-9.
14. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884-i90.
15. Chen S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. iMeta. 2023;2(2):e107.
16. Sherman BT, Hao M, Qiu J, Jiao X, Baseler MW, Lane HC, et al. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucleic Acids Res. 2022;50(W1):W216-w21.
17. Peixoto L, Risso D, Poplawski SG, Wimmer ME, Speed TP, Wood MA, Abel T. How data analysis affects power, reproducibility and biological insight of RNA-seq studies in complex datasets. Nucleic Acids Res. 2015;43(16):7664-74.
18. Gandolfo LC, Speed TP. RLE plots: Visualizing unwanted variation in high dimensional data. PLoS One. 2018;13(2):e0191629.
19. Molania R, Gagnon-Bartsch JA, Dobrovic A, Speed TP. A new normalization for Nanostring nCounter gene expression data. Nucleic Acids Res. 2019;47(12):6073-83.
20. Chou CK, Fan CC, Lin PS, Liao PY, Tung JC, Hsieh CH, et al. Sciellin mediates mesenchymal-to-epithelial transition in colorectal cancer hepatic metastasis. Oncotarget. 2016;7(18):25742-54.
21. Mashino K, Sadanaga N, Yamaguchi H, Tanaka F, Ohta M, Shibuta K, et al. Expression of Chemokine Receptor CCR7 Is Associated with Lymph Node Metastasis of Gastric Carcinoma1. Cancer Res. 2002;62(10):2937-41.
22. Li J, Sun R, Tao K, Wang G. The CCL21/CCR7 pathway plays

a key role in human colon cancer metastasis through regulation of matrix metalloproteinase-9. Dig Liver Dis. 2011;43(1):40-7.
23. Zou Y, Chen Y, Wu X, Yuan R, Cai Z, He X, et al. CCL21 as an independent favorable prognostic factor for stage III/IV colorectal cancer. Oncol Rep. 2013;30(2):659-66.
24. Pezeshkian Z, Nobili S, Peyravian N, Shojaee B, Nazari H, Soleimani H, et al. Insights into the Role of Matrix Metalloproteinases in Precancerous Conditions and in Colorectal Cancer. Cancers. 2021;13(24):6226.
25. Hosseinalizadeh H, Hussain QM, Poshtchaman Z, Ahsan M, Amin AH, Naghavi S, et al. Emerging insights into keratin 7 roles in tumor progression and metastasis of cancers. Front Oncol. 2023;13:1243871.
26. Li C, Cheng B, Yang X, Tong G, Wang F, Li M, et al. SOX8 promotes tumor growth and metastasis through FZD6-dependent Wnt/β-catenin signaling in colorectal carcinoma. Heliyon. 2023;9(12):e22586.
27. Fagarasan G, Fagarasan V, Bintintan VV, Dindelegan GC. The Role of Tumor-Infiltrating B Lymphocytes in Colorectal Cancer Patients: A Systematic Review of Immune Landscape Evolution. Cancers (Basel). 2025;17(18).
28. Afshar-Kharghan V. The role of the complement system in cancer. J Clin Invest. 2017;127(3):780-9.
29. Agoston EI, Acs B, Herold Z, Fekete K, Kulka J, Nagy A, et al. Deconstructing Immune Cell Infiltration in Human Colorectal Cancer: A Systematic Spatiotemporal Evaluation. Genes (Basel). 2022;13(4).
30. Lo JH, Battaglin F, Baca Y, Xiu J, Brodskiy P, Algaze S, et al. DEFB1 gene expression and the molecular landscape of colorectal cancer (CRC). J Clin Oncol. 2022;40(16 Suppl):3523.
31. Weidle UH, Rohwedder I, Birzele F, Weiss EH, Schiller C. LST1: A multifunctional gene encoded in the MHC class III region. Immunobiology. 2018;223(11):699-708.
32. Islam MS, Gopalan V, Lam AK, Shiddiky MJA. Current advances in detecting genetic and epigenetic biomarkers of colorectal cancer. Biosens Bioelectron. 2023;239:115611.
33. Bjorling E, Oksvold P, Forsberg M, Lund J, Ponten F, Uhlén M. Human protein atlas, version 2. Molecular & Cellular Proteomics. 2006;5(10):S328-S.
34. Tadijan A, Samarzija I, Humphries JD, Humphries MJ, Ambriovic-Ristov A. KANK family proteins in cancer. Int J Biochem Cell Biol. 2021;131:105903.
35. Yin JH, Elumalai P, Kim SY, Zhang SZ, Shin S, Lee M, et al. TNNC1 knockout reverses metastatic potential of ovarian cancer cells by inactivating epithelial-mesenchymal transition and suppressing F-actin polymerization. Biochem Biophys Res Commun. 2021;547:44-51.
36. Fang C, Zhang X, Li C, Liu F, Liu H. Troponin C-1 Activated by E2F1 Accelerates Gastric Cancer Progression via Regulating TGF-beta/Smad Signaling. Dig Dis Sci. 2022;67(9):4444-57.
37. Shang B, Qiao H, Wang L, Wang J. In-depth study of pyroptosis-related genes and immune infiltration in colon cancer. PeerJ. 2024;12:e18374.
38. Xu F, Kong L, Sun X, Hui W, Jiang L, Han W, et al. PFDN6 contributes to colorectal cancer progression via transcriptional regulation. eGastroenterology. 2024;2(2):e100001.
39. Peppino G, Ruiu R, Arigoni M, Riccardo F, Iacoviello A, Barutello G, et al. Teneurins: Role in Cancer and Potential Role as Diagnostic Biomarkers and Targets for Therapy. Int J Mol Sci. 2021;22(5):2321.
40. De Smedt L, Palmans S, Andel D, Govaere O, Boeckx B, Smeets D, et al. Expression profiling of budding cells in colorectal cancer reveals an EMT-like phenotype and molecular subtype switching. Br J Cancer. 2017;116(1):58-65.
41. Chang X, Chai Z, Zou J, Wang H, Wang Y, Zheng Y, et al. PADI3 induces cell cycle arrest via the Sirt2/AKT/p21 pathway and acts as a tumor suppressor gene in colon cancer. Cancer Biol Med. 2019;16(4):729-42.
42. Xu D, Ding S, Cao M, Yu X, Wang H, Qiu D, et al. A Pan-Cancer Analysis of Cystatin E/M Reveals Its Dual Functional Effects and Positive Regulation of Epithelial Cell in Human Tumors. Front Genet. 2021;12:733211.
43. Zhu X, Zhao L, Hu P. Predictive Values of Homeobox Gene A-Antisense Transcript 3 (HOXA-AS3), Cystatin 6 (CST6), and Chromobox Homolog 4 (CBX4) Expressions in Cancer Tissues for Recurrence of Early Colon Cancer After Surgery. Int J Gen Med. 2024;17:1-8.
44. Peng X, Chen X, Zhu X, Chen L. GALNT6 Knockdown Inhibits the Proliferation and Migration of Colorectal Cancer Cells and Increases the Sensitivity of Cancer Cells to 5-FU. J Cancer. 2021;12(24):7413-21.
45. Nunes L, Li F, Wu M, Luo T, Hammarström K, Torell E, et al. Prognostic genome and transcriptome signatures in colorectal cancers. Nature. 2024;633(8028):137-46.
46. Faleti OD, Gong Y, Long J, Luo Q, Tan H, Deng S, et al. TRIM72 inhibits cell migration and epithelial-mesenchymal transition by attenuating FAK/akt signaling in colorectal cancer. Heliyon. 2024;10(18):e37714.
47. Chung JH, Larsen AR, Chen E, Bunz F. A PTCH1 homolog transcriptionally activated by p53 suppresses Hedgehog signaling. J Biol Chem. 2014;289(47):33020-31.
48. Rossi M, Banskota N, Shin CH, Anerillas C, Tsitsipatis D, Yang JH, et al. Increased PTCHD4 expression via m6A modification of PTCHD4 mRNA promotes senescent cell survival. Nucleic Acids Res. 2024;52(12):7261-78.
49. Vite A, Li J, Radice GL. New functions for alpha-catenins in health and disease: from cancer to heart regeneration. Cell Tissue Res. 2015;360(3):773-83.