*Proceedings of the 12th World Congress on New Technologies (NewTech 2026)*
*August 20 – 22, 2026 – London, United Kingdom*
*Paper No. ICBB 106*
*DOI:10.11159/icbb26.106*

# LASSO-Based Gene Signature Identification for Colorectal Cancer Classification Using TCGA Expression Data

**Xia Wu[1*]**
[1]Department of Internal Medicine, Carver College of Medicine, University of Iowa, Iowa City, IA, USA
xwu76@uiowa.edu

***Abstract -*** Colorectal cancer (CRC) is the third most frequently diagnosed cancer and the second leading cause of cancer-related death worldwide. Robust molecular signatures that distinguish malignant from normal tissue are valuable for early diagnosis and biomarker development. This study developed an integrative pipeline combining differential expression analysis, unsupervised learning, and L1-regularized logistic regression using TCGA colorectal adenocarcinoma RNA-seq data from UCSC Xena. Tumor and normal samples were decoded from TCGA barcode sample-type codes, yielding 286 primary tumors and 41 adjacent normal tissues. Limma identified 1,618 dysregulated genes (358 up-regulated and 1,260 down-regulated; $|\log2FC| > 2$, BH-adjusted $p < 0.05$). Principal component analysis showed stronger tumor-normal separation after feature filtering, with PC1 explaining 40.6% of variance compared with 14.5% for the genome-wide profile. Ten-fold cross-validated LASSO logistic regression reduced the feature space to a 15-16-gene signature and achieved essentially perfect tumor-normal discrimination (AUC = 1.00; Wilcoxon $p < 1e-24$), consistent with the large transcriptomic divergence between tumor and normal colon. The signature included CRC-associated genes such as CDH3, KRT80, ETV4, ESM1, FOXQ1, OTOP2 and CDH10, and enrichment analysis linked dysregulated genes to Wnt, cadherin, Hippo, extracellular-matrix, transport and metabolic pathways. As a harder clinical benchmark, microsatellite-instability status was predicted under nested cross-validation, achieving AUC = 0.928 and identifying an MSI-H signature led by MLH1 and immune-related genes. External validation is required before diagnostic use.

***Keywords*:** colorectal cancer, LASSO, gene expression, TCGA, biomarker, machine learning

## 1. Introduction

Colorectal cancer (CRC) represents the third most diagnosed malignancy and the second leading cause of cancer-related mortality globally [1]. Early detection and accurate classification of CRC remain critical challenges in clinical oncology. High-throughput sequencing has enabled comprehensive molecular profiling of tumors and the discovery of diagnostic and prognostic biomarkers [2]. The Cancer Genome Atlas (TCGA) has generated extensive multi-omics data across cancer types, providing unprecedented opportunities for computational biomarker discovery [3].

However, the high dimensionality of transcriptomic data poses analytical challenges, requiring robust feature-selection methods to identify biologically meaningful gene signatures while avoiding overfitting [4]. Standard differential expression analysis frequently returns hundreds to thousands of significant genes, many of which are secondary changes rather than core discriminative features. Moreover, the strong multi-collinearity of gene expression, in which co-regulated genes act within shared pathways, can destabilize conventional regression models. This motivates a transition from broad genomic profiling toward a parsimonious, highly specific gene set that retains maximal discriminatory power with minimal complexity.

Least Absolute Shrinkage and Selection Operator (LASSO) regression is a powerful approach for simultaneous feature selection and model construction in high-dimensional biological data [5]. By imposing an L1 penalty on the coefficients, LASSO shrinks uninformative features to zero, yielding sparse and interpretable models, a feature particularly valuable for clinical panels, where a small set of verified biomarkers is more practical than whole-transcriptome assays.

This study presents an integrative pipeline combining differential expression analysis, unsupervised analysis, and LASSO-based feature selection to identify a robust gene signature for CRC classification using TCGA RNA-seq data. Starting from 20,530 genes, a compact signature of roughly fifteen genes was derived and evaluated by cross-validation for its ability to distinguish malignant from normal colonic tissue; the same pipeline was also applied to the harder task of predicting microsatellite-instability status.

## 2. Materials and Methods

### 2.1. Data and Preprocessing

Log2-normalized RNA-seq gene expression for the TCGA colorectal adenocarcinoma (COAD) cohort (dataset HiSeqV2; 20,530 genes across 329 samples) was downloaded from the UCSC Xena platform [8]. Sample type was decoded directly from the TCGA barcode: samples whose barcode ends in -01 were labeled primary tumor and those ending in -11 solid tissue normal. The two non-primary malignant samples (one recurrent, -02; one metastatic, -06) were excluded, giving a final set of 286 tumor and 41 adjacent normal samples. Genes with zero variance across samples were removed prior to analysis.

### 2.2. Differential Expression Analysis

Differential expression was assessed with the limma package [6]. A linear model was fit with a group design matrix, and the tumor-versus-normal contrast was evaluated using empirical Bayes moderation of the variance. Genes were considered differentially expressed if they satisfied $|\log2FC| > 2$ and a Benjamini-Hochberg adjusted p-value $< 0.05$.

### 2.3. Unsupervised Analysis

Principal component analysis (PCA) was performed on scaled, transposed expression matrices using singular value decomposition, both for the full gene set and for the differentially expressed genes; the proportion of variance explained by each component was recorded. PCA was used to assess global expression structure and the degree of tumor-normal separation.

### 2.4. LASSO Logistic Regression

LASSO logistic regression (binomial family, alpha = 1) was implemented with the glmnet package [7] on the differentially expressed genes. Ten-fold cross-validation was used to select the regularization parameter, reporting both lambda.min (minimum cross-validation deviance) and lambda.1se (most regularized model within one standard error of the minimum). Classification performance was assessed from the out-of-fold (pre-validated) predictions to avoid resubstitution bias, summarized by receiver operating characteristic (ROC) analysis and the area under the curve (AUC). A fixed random seed was set for reproducibility.

### 2.5. Functional Enrichment

Gene Ontology biological-process (GO BP) and KEGG pathway over-representation analyses were performed separately for the up- and down-regulated differentially expressed genes using clusterProfiler [10], with gene symbols mapped to Entrez identifiers via org.Hs.eg.db. Redundant GO terms were collapsed, and terms with a Benjamini-Hochberg adjusted p-value $< 0.05$ were retained.

### 2.6. Prediction of MSI Status

As a harder, clinically meaningful benchmark, MSI status was predicted from expression. MSI labels were taken from the cBioPortal PanCancer Atlas colorectal cohort [11]: tumors with a MANTIS score $>= 0.4$ were labeled MSI-H and the remainder MSS [12]; these calls agreed with MSIsensor ($>= 10$) for 95% of samples. After intersection with the expression matrix, 273 tumors remained (59 MSI-H, 214 MSS). To obtain an unbiased estimate, performance was assessed by nested ten-fold cross-validation in which differential-expression feature selection and LASSO fitting were performed only on the inner training samples of each fold and the held-out fold was used solely for prediction. A final signature was obtained by refitting on the full set at lambda.1se.

## 3. Results

### 3.1. Differential Expression

Differential expression analysis identified 1,618 genes meeting the significance thresholds, comprising 358 up-regulated and 1,260 down-regulated genes in tumor relative to normal tissue (Fig. 1). The predominance of down-regulated genes is consistent with the broad loss of normal colonic differentiation programs in carcinoma. Among the most strongly altered genes were markedly up-regulated CDH3, KRT80 and FOXQ1 and strongly down-regulated OTOP2.

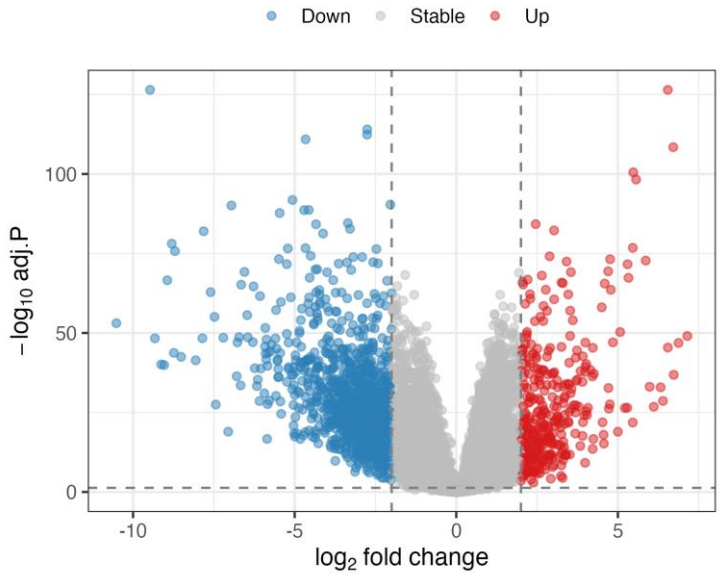



Fig. 1: Volcano plot of differential expression (tumor vs. normal). Red: up-regulated; blue: down-regulated.

### 3.2. Unsupervised Analysis

PCA of all 20,530 genes revealed partial separation between tumor and normal samples along PC1, which captured 14.5% of the total variance (Fig. 2, left). Restricting the analysis to the 1,618 differentially expressed genes substantially sharpened the separation, with PC1 explaining 40.6% of variance and the two groups forming clearly distinct clusters (Fig. 2, right). Hierarchical clustering of the top differentially expressed genes likewise partitioned the samples into tumor and normal blocks (Fig. 3).

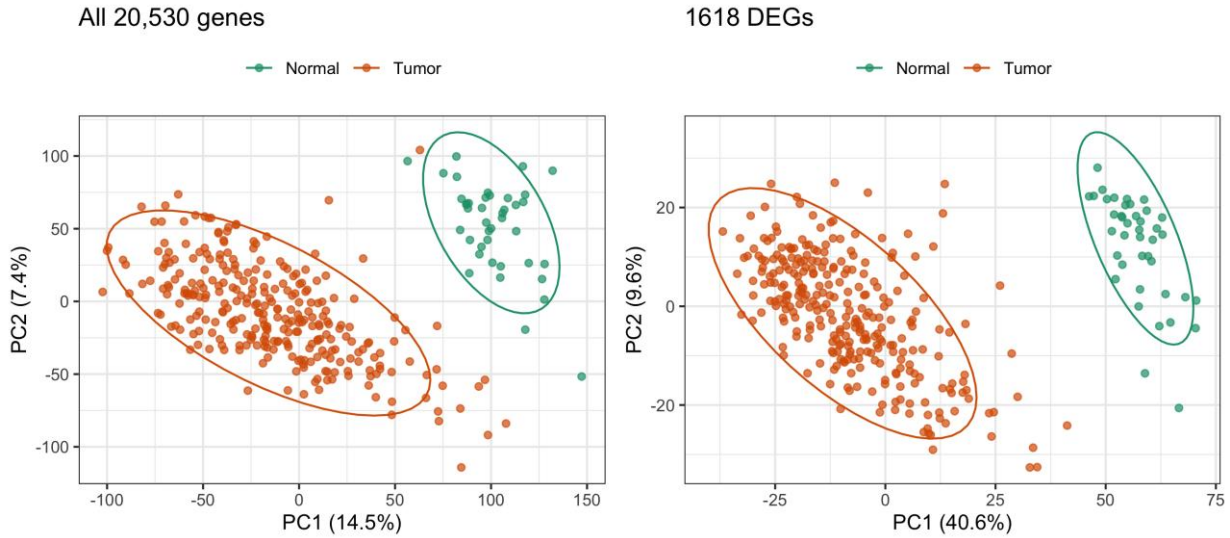



Fig. 2: PCA using all genes (left) and the differentially expressed genes (right).

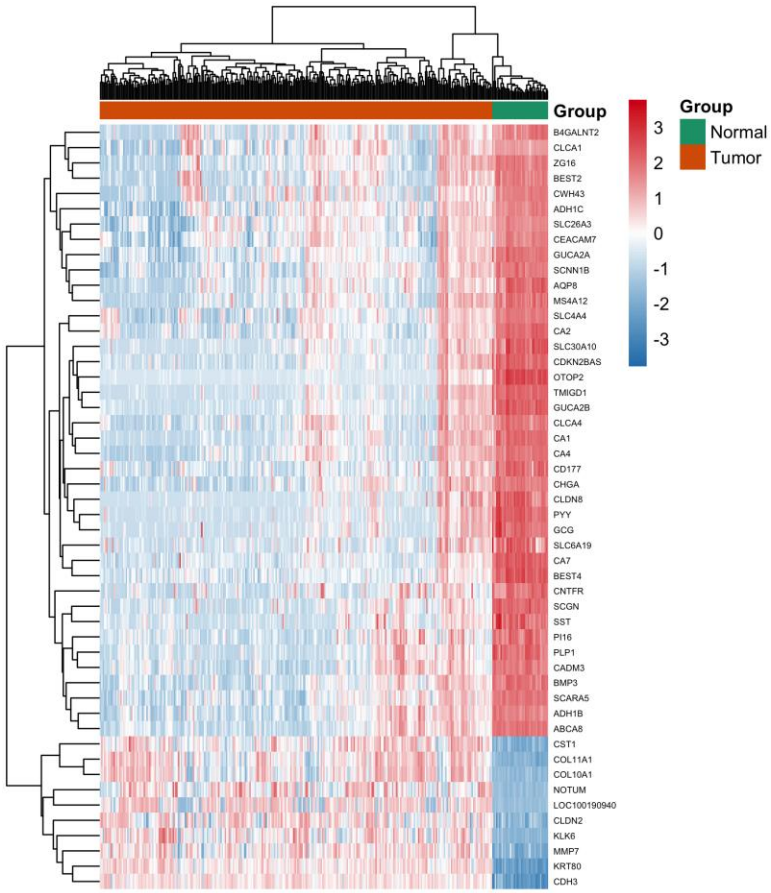



Fig. 3: Heatmap of the top 50 differentially expressed genes (row-scaled).

### 3.3. LASSO Model Performance

Ten-fold cross-validated LASSO logistic regression identified lambda.min = 0.0030, retaining 16 genes with non-zero coefficients, and lambda.1se = 0.0046, retaining 15 genes (Table 1; Fig. 4A). Using out-of-fold predictions, both models achieved near-perfect discrimination between tumor and normal tissue (AUC = 1.00; Fig. 4B), and the predicted tumor probabilities were almost completely separated between the two groups (Wilcoxon p = 3.96e-25). Because tumor and normal colon differ so extensively at the transcriptome level, this near-perfect separation is the expected outcome rather than evidence of a difficult classification problem; it serves here as an internal sanity check confirming that the compact signature retains the discriminative information of the full DEG set.

Table 1: LASSO-selected gene signatures and cross-validated performance.

| Model | Selected genes | CV AUC |
|---|---|---|
| lambda.min (16) | CDH3, OTOP2, DHRS7C, AADACL2, KRT80, ETV4, ESM1, FAM180B, RXRG, GLP2R, KHDRBS2, SGCG, FOXQ1, CMTM5, CDH10, GRIN2D | 1.00 |
| lambda.1se (15) | CDH3, OTOP2, DHRS7C, AADACL2, KRT80, ETV4, ESM1, FAM180B, RXRG, KHDRBS2, SGCG, FOXQ1, CMTM5, CDH10, GRIN2D | 1.00 |

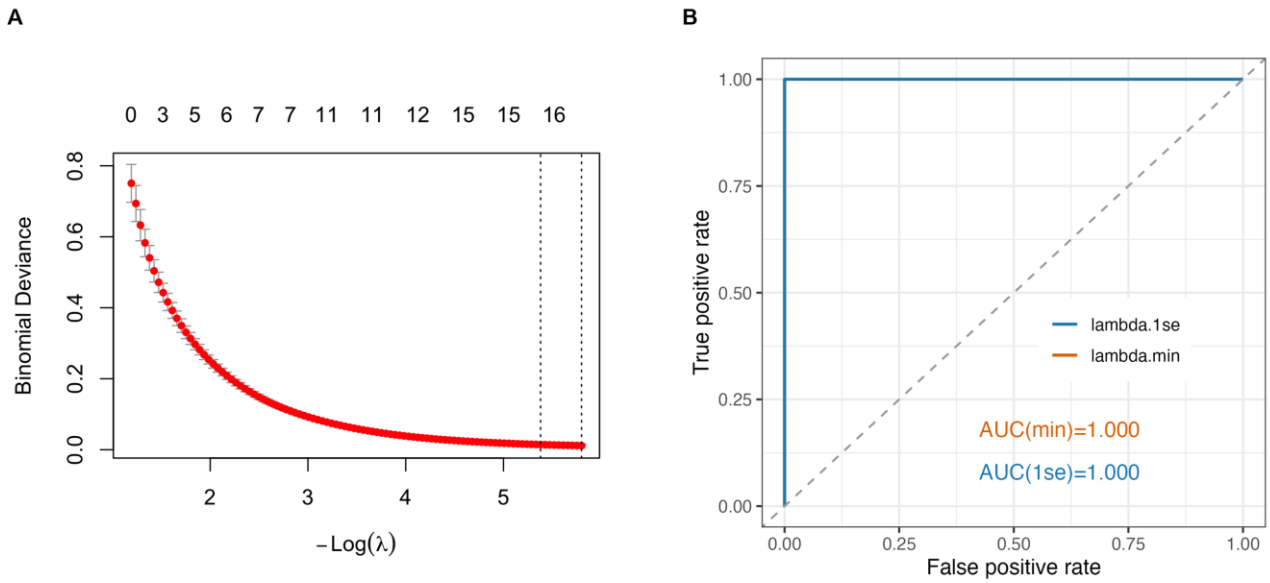



Fig. 4: LASSO model tuning and classification performance. A: ten-fold cross-validation curve; dashed lines mark lambda.min and lambda.1se. B: ROC curves for the lambda.min and lambda.1se models (cross-validated AUC = 1.00).

### 3.4. Functional Enrichment

To characterize the biology underlying the differentially expressed genes, GO and KEGG over-representation analyses were performed separately for the up- and down-regulated sets (Fig. 5). Up-regulated genes were enriched for epidermis and keratinocyte differentiation, intermediate-filament organization and extracellular-matrix remodeling, and, at the pathway level, for the Wnt, cadherin and Hippo signaling pathways, canonical drivers of colorectal tumorigenesis. Down-regulated genes were dominated by terms reflecting loss of normal colonic epithelial function, including regulation of membrane potential, sodium-ion transmembrane transport, hormone metabolism, neuroactive ligand-receptor interaction and cytochrome-P450 drug metabolism. These enrichments are concordant with the direction of the LASSO signature (e.g. up-regulated CDH3 and KRT80, down-regulated OTOP2) and support its biological relevance.

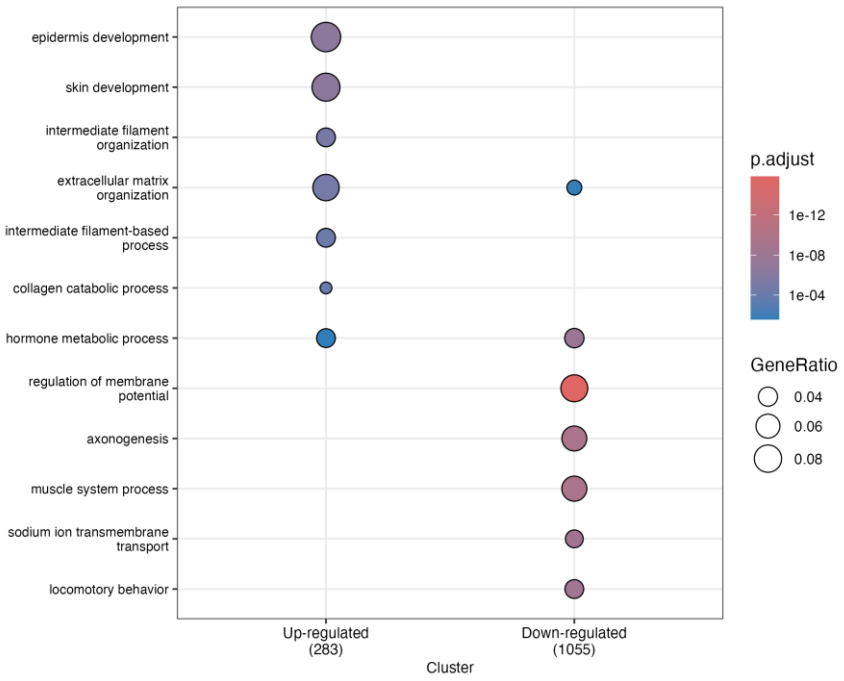


Fig. 5: GO biological-process and KEGG over-representation analysis of the up- and down-regulated DEGs (top terms; dot size = gene ratio, color = adjusted p-value).

### 3.5. Prediction of Microsatellite-Instability Status

Unlike tumor-versus-normal discrimination, distinguishing MSI-H from MSS tumors is a genuinely difficult task with direct clinical consequences, since MSI-H tumors qualify for immune-checkpoint therapy. Under nested cross-validation, in which feature selection never sees the held-out samples, the model achieved an AUC of 0.928 (Fig. 6), a realistic level of performance for this problem. The out-of-fold predicted MSI-H probabilities were strongly separated between the two groups (Wilcoxon p = 8.6e-24) while retaining the partial overlap expected of a non-trivial classifier. The final 22-gene signature was led by MLH1, whose epigenetic silencing is the predominant cause of sporadic MSI-H colorectal cancer, together with a mismatch-repair gene (MSH4) and immune genes (IFNG, CIITA, TNFRSF18, KIR3DL1), mirroring the hypermutated, immune-infiltrated biology of MSI-H tumors.

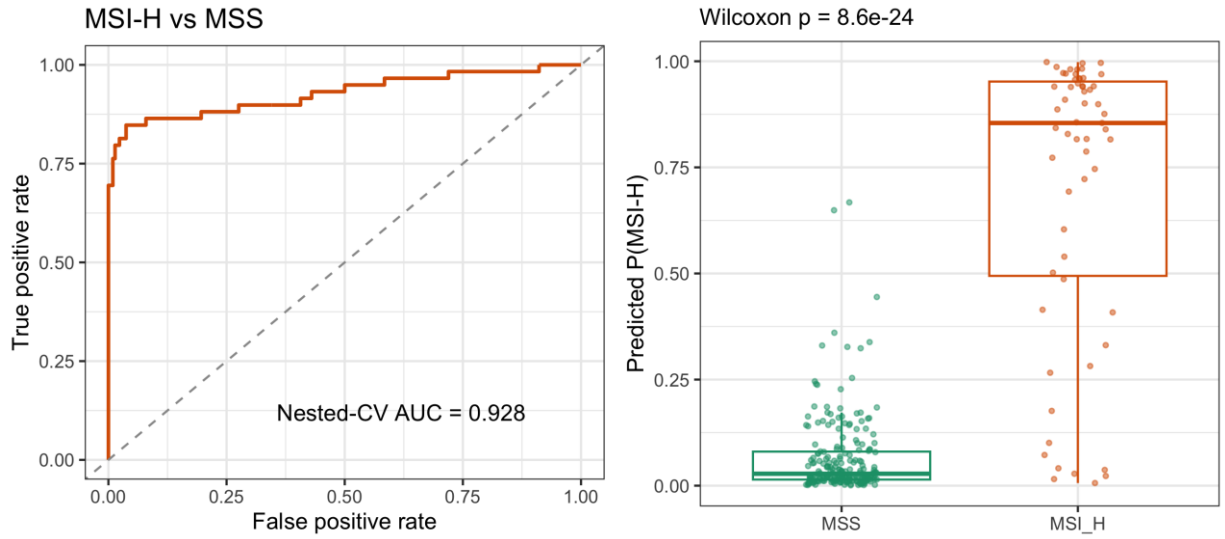



Fig. 6: MSI-status prediction (MSI-H vs MSS). Left: nested cross-validation ROC (AUC = 0.928). Right: out-of-fold predicted MSI-H probability by group.

## 4. Discussion

Integrating differential expression analysis with LASSO logistic regression yielded a compact 15-16-gene signature that distinguishes colon adenocarcinoma from normal colonic tissue with near-perfect cross-validated accuracy. Reducing more than 20,000 features to roughly fifteen genes while retaining full discriminatory power illustrates the value of L1 regularization for building interpretable, low-dimensional classifiers from transcriptomic data.

The selected genes are concordant with established CRC biology. Among the up-regulated members, CDH3 (P-cadherin) is aberrantly expressed in colorectal tumors and linked to invasion and poor prognosis; ETV4, an oncogenic ETS-family transcription factor, promotes proliferation and metastasis; FOXQ1 drives tumorigenicity and epithelial-mesenchymal transition in CRC [9]; ESM1 (endocan) is associated with tumor angiogenesis; and KRT80 supports tumor-cell proliferation. Conversely, the down-regulated genes include candidate tumor suppressors and markers of normal colonic differentiation: OTOP2, the most strongly down-regulated gene in the signature, is a p53-responsive proton channel silenced in CRC, while CDH10 is a cadherin frequently lost in colorectal tumors. The coherent up/down structure of the signature supports its biological plausibility rather than reflecting purely statistical artifacts.

Functional enrichment reinforced this interpretation. Up-regulated genes mapped onto Wnt, cadherin and Hippo signaling and extracellular-matrix remodeling, hallmark oncogenic programs in colorectal cancer, whereas down-regulated genes corresponded to the loss of differentiated colonic transport, secretory and xenobiotic-metabolism functions. The signature therefore captures coherent tumor biology rather than statistically incidental features.

The contrast between the two classification tasks is instructive. Tumor-versus-normal separation is essentially saturated (AUC 1.00) and mainly serves to validate the pipeline, whereas MSI-status prediction, evaluated without feature-selection leakage, yields a realistic AUC of 0.928 and recovers the established biology of MSI-H tumors, from MLH1 loss

to interferon-driven immune activation. The MSI signature is therefore both a stricter test of the method and a more clinically actionable output, given the role of MSI-H status in selecting patients for immunotherapy.

Several limitations should be emphasized. Discriminating bulk tumor from adjacent normal tissue is an intrinsically easy task because the two states differ profoundly in their transcriptomes; the near-perfect AUC therefore reflects the strength of this biological contrast and should not be read as evidence of clinical-grade diagnostic performance. The model was trained and evaluated within a single cohort by cross-validation; independent external validation is required to establish generalizability and to guard against batch- and platform-specific effects. Finally, clinically useful problems beyond the MSI analysis presented here, including molecular subtyping, tumor staging, and prognosis, remain to be addressed. Future work should integrate clinical covariates and validate both signatures on external cohorts before any diagnostic application.

## 5. Conclusion

This study developed a reproducible, integrative bioinformatics pipeline for CRC biomarker discovery and identified a compact gene signature that cleanly separates tumor from normal colorectal tissue. The selected genes are biologically interpretable and consistent with known CRC pathophysiology, and the framework is readily transferable to other cancer types and molecular-classification tasks. External validation and extension to finer-grained clinical endpoints are the natural next steps.

## Data and Code Availability

The gene expression data are publicly available from the UCSC Xena platform (TCGA Colon Cancer cohort, dataset TCGA.COAD.sampleMap/HiSeqV2), and MSI annotations from the cBioPortal PanCancer Atlas colorectal cohort. All analysis code (R) needed to reproduce the results and figures is openly available at https://github.com/candicewu0515/crc-lasso-tcga.

## References


[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA Cancer J. Clin., vol. 71, no. 3, pp. 209-249, 2021.

[2] The Cancer Genome Atlas Network, "Comprehensive molecular characterization of human colon and rectal cancer," Nature, vol. 487, no. 7407, pp. 330-337, 2012.

[3] The Cancer Genome Atlas Research Network, J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Mills Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart, "The Cancer Genome Atlas Pan-Cancer analysis project," Nat. Genet., vol. 45, no. 10, pp. 1113-1120, 2013.

[4] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," Bioinformatics, vol. 23, no. 19, pp. 2507-2517, 2007.

[5] R. Tibshirani, "Regression shrinkage and selection via the lasso," J. R. Stat. Soc. Series B, vol. 58, no. 1, pp. 267-288, 1996.

[6] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, "limma powers differential expression analyses for RNA-sequencing and microarray studies," Nucleic Acids Res., vol. 43, no. 7, p. e47, 2015.

[7] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," J. Stat. Softw., vol. 33, no. 1, pp. 1-22, 2010.

[8] M. J. Goldman, B. Craft, M. Hastie, K. Repečka, F. McDade, A. Kamath, A. Banerjee, Y. Luo, D. Rogers, A. N. Brooks, J. Zhu, and D. Haussler, "Visualizing and interpreting cancer genomics data via the Xena platform," Nat. Biotechnol., vol. 38, no. 6, pp. 675-678, 2020.

[9] H. Kaneda, T. Arao, K. Tanaka, D. Tamura, K. Aomatsu, K. Kudo, K. Sakai, M. A. De Velasco, K. Matsumoto, Y. Fujita, Y. Yamada, J. Tsurutani, I. Okamoto, K. Nakagawa, and K. Nishio, "FOXQ1 is overexpressed in colorectal cancer and enhances tumorigenicity and tumor growth," Cancer Res., vol. 70, no. 5, pp. 2053-2063, 2010.

[10] T. Wu, E. Hu, S. Xu, M. Chen, P. Guo, Z. Dai, T. Feng, L. Zhou, W. Tang, L. Zhan, X. Fu, S. Liu, X. Bo, and G. Yu, "clusterProfiler 4.0: A universal enrichment tool for interpreting omics data," The Innovation, vol. 2, no. 3, p. 100141, 2021.

[11] E. Cerami, J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, B. A. Aksoy, A. Jacobsen, C. J. Byrne, M. L. Heuer, E. Larsson, Y. Antipin, B. Reva, A. P. Goldberg, C. Sander, and N. Schultz, "The cBio Cancer Genomics Portal: An

open platform for exploring multidimensional cancer genomics data," Cancer Discov., vol. 2, no. 5, pp. 401-404, 2012.
[12] E. A. Kautto, R. Bonneville, J. Miya, L. Yu, M. A. Krook, J. W. Reeser, and S. Roychowdhury, "Performance evaluation for rapid detection of pan-cancer microsatellite instability with MANTIS," Oncotarget, vol. 8, no. 5, pp. 7452-7463, 2017.