Integrative machine learning reveals potential signature genes using transcriptomics in colon cancer

Mostafa Amir Hamza[1] & Saiful Islam[2]

[1] Department of Biotechnology and Genetic Engineering, University of Development Alternative (UODA), Dhanmondi R/A, Dhaka 1209, Bangladesh

[2] Department of Anatomy, Sher-e-Bangla Medical College, Barishal 8200, Bangladesh

Correspondence: Saiful Islam, Curator, Department of Anatomy, Sher-e-Bangla Medical College, Band Rd, Barishal 8200, Bangladesh. Email: saiful_sb31st@yahoo.com

Received: March 11, 2025 DOI: 10.14295/bjs.v4i9.745

Accepted: June 27, 2025 URL: https://doi.org/10.14295/bjs.v4i9.745

Abstract

Colon cancer is a significant health burden in the world and the second leading cause of cancer-related deaths. Despite advancements in diagnosis and treatment, identifying potential biomarkers for early detection and therapeutic targets remains challenging. This study used an integrative approach combining transcriptomics and machine learning to identify signature genes and pathways associated with colon cancer. RNA-Seq data from The Cancer Genome Atlas- Colon Adenocarcinoma (TCGA-COAD) project, comprising 485 samples, were analyzed in this study. Differential gene expression analysis revealed 657 upregulated and 8,566 downregulated genes. Notably, EPB41L3, TSPAN7, and ABI3BP were identified as highly upregulated, while LYVE1, PLPP1, and NFE2L3 were significantly downregulated in tumor samples. Gene Set Enrichment Analysis (GSEA) identified dysregulated pathways, including E2F targets, MYC targets, and G2M checkpoints, underscoring cell cycle regulation and metabolic reprogramming alterations in colon cancer. Machine learning models-Random Forest, Neural Networks, and Logistic Regression-achieved high classification accuracy (97–99%). Key genes consistently identified across these models highlight their potential translational relevance as biomarkers. This study integrates differential expression analysis, pathway enrichment, and machine learning to uncover critical insights into colon cancer biology. The study lays the groundwork for developing diagnostic and therapeutic strategies, with the identified genes and pathways serving as potential candidates for further validation and clinical applications. This approach exemplifies the potential of precision medicine to advance colon cancer research and improve patient outcomes.

Keywords: cancer genome atlas, colon cancer, machine learning, transcriptomics

Aprendizado de máquina integrativo revela genes assinatura potenciais utilizando transcriptômica no câncer de cólon

Resumo

O câncer de cólon representa um importante problema de saúde pública no mundo, sendo a segunda principal causa de mortes relacionadas ao câncer. Apesar dos avanços no diagnóstico e tratamento, a identificação de biomarcadores potenciais para detecção precoce e alvos terapêuticos ainda é um desafio. Este estudo utilizou uma abordagem integrativa combinando transcriptômica e aprendizado de máquina para identificar genes assinatura e vias associadas ao câncer de cólon. Foram analisados dados de RNA-Seq provenientes do projeto The Cancer Genome Atlas – Colon Adenocarcinoma (TCGA-COAD), compreendendo 485 amostras. A análise de expressão gênica diferencial revelou 657 genes superexpressos e 8.566 genes com expressão reduzida. Notavelmente, EPB41L3, TSPAN7 e ABI3BP foram altamente superexpressos, enquanto LYVE1, PLPP1 e NFE2L3 apresentaram redução significativa da expressão nos tumores. A Análise de Enriquecimento de Conjuntos Gênicos (GSEA) identificou vias desreguladas, incluindo alvos de E2F, MYC e pontos de checagem G2/M, evidenciando alterações na regulação do ciclo celular e no reprograma metabólico do câncer de cólon. Modelos de aprendizado de máquina – Random Forest, Redes Neurais e Regressão Logística – alcançaram alta acurácia de classificação (97–99%). Genes-chave identificados de forma consistente entre os modelos

demonstram potencial relevância translacional como biomarcadores. Este estudo integra análise de expressão diferencial, enriquecimento de vias e aprendizado de máquina para revelar insights críticos sobre a biologia do câncer de cólon. Os resultados fornecem uma base para o desenvolvimento de estratégias diagnósticas e terapêuticas, com os genes e vias identificados servindo como candidatos potenciais para validação futura e aplicações clínicas. Essa abordagem exemplifica o potencial da medicina de precisão para avançar na pesquisa do câncer de cólon e melhorar os desfechos clínicos dos pacientes.utilizando transcriptômica no câncer de cólon

Palavras-chave: atlas do genoma do câncer, câncer de cólon, aprendizado de máquina, transcriptômica

1. Introduction

Colon cancer is a major global health concern, ranking as the third most common cancer worldwide and accounting for approximately 10% of all cancer cases. It is also the second leading cause of cancer-related deaths globally. Despite advancements in diagnosis and treatment, the lack of robust biomarkers for early detection and therapeutic targeting remains a critical challenge. In 2020, an estimated 1.9 million new cases of colon cancer and over 930,000 related deaths were reported worldwide (Sawicki et al., 2021; Siegel et al., 2023; Xi; Xu, 2021).

Colon cancer is one of the most frequently diagnosed cancers in both men and women. It ranks third in cancer-related deaths among men and fourth among women, but collectively, it is the second leading cause of cancer mortality. In 2024, the American Cancer Society projects approximately 106,590 new cases of colon cancer (54,210 in men and 52,380 in women) (Walter Reed National Military Medical Center, 2024). Importantly, the incidence of colon cancer is increasing among younger adults, where it has become the leading cause of cancer-related deaths in men under 50 and the second leading cause in women under 50, following breast cancer. Colon cancer is projected to contribute substantially to the estimated 53,010 colon cancer-related deaths in the United States in 2024. (Augustus; Ellis, 2018; Siegel et al., 2024).

To address this growing burden, molecular tools have become essential in identifying underlying drivers of disease. Advancements in RNA sequencing (RNA-Seq) have revolutionized transcriptomics, enabling comprehensive profiling of gene expression across various biological conditions, including cancer (Wang et al., 2009, Tomczak et al., 2015).

This study focuses on identifying differentially expressed genes (DEGs) and pathways in colon cancer, a malignancy characterized by significant clinical challenges and heterogeneity. Differential expression analysis serves as a foundational step in uncovering genes with altered expression in tumors, while Gene Set Enrichment Analysis (GSEA) offers a pathway-level understanding of systemic changes in tumor biology (Subramanian et al., 2005). Additionally, gene interaction networks provide a systems biology perspective, revealing key nodes and hubs that may serve as regulatory elements or therapeutic targets (Barabasi et al., 2011).

With the growing emphasis on precision medicine, the integration of machine learning with transcriptomics has gained momentum. Machine learning approaches, such as logistic regression, artificial neural networks, and random forests, offer powerful tools for feature selection, pattern recognition, and predictive modeling (Libbrecht; Noble, 2015). In this study, we leveraged these techniques using Python's scikit-learn library to identify potential marker genes capable of distinguishing tumors from normal samples, intending to enhance diagnostic and prognostic capabilities in colon cancer. By integrating these diverse methodologies, this study provides a holistic approach to biomarker discovery in colon cancer. It identifies potential marker genes and establishes a framework for leveraging RNA-Seq data in translational cancer research. The findings hold promise for advancing our understanding of colon cancer biology and contributing to developing personalized diagnostic and therapeutic strategies.

2. Materials and Methods

2.1 Data acquisition and preprocessing

Colon adenocarcinoma (COAD) gene expression data were retrieved from the TCGA database (https://portal.gdc.cancer.gov/repository) using TCGAbiolinks in R. We analyzed 485 samples after filtering for clinical data availability. Transcriptomic data from the TCGA-COAD project were selected, comprising 485 samples, including 444 tumor tissue samples and 41 matched normal tissue samples based on clinical data availability. Clinical data for colon cancer patients were also downloaded, with key survival and staging information extracted for analysis.

2.2 Gene expression pre-filtering and normalization

Using TCGAbiolinks, transcriptomics data were downloaded as fragments per kilobase million (FPKM) unstranded normalized data (Colaprico et al., 2016; Mounir et al., 2019). The dataset initially included 60,660 Ensembl gene identifiers. These Ensembl IDs were converted to gene names, resulting in 42,225 named genes. Genes with zero expression values across all samples were excluded, leaving 10,962 genes for further expression and machine learning analysis. Since the data was normalized, a log2 transformation was applied to the FPKM data to stabilize data variance.

2.3 Gene expression analysis

Log2-transformed normalized data were used to assess gene expression alterations in tumor patients compared to those in healthy controls. A t-test was performed to identify significant gene expression changes, followed by false discovery rate (FDR) correction to adjust for multiple comparisons. Differentially expressed genes were defined as those with an FDR-adjusted p-value < 0.05. These genes were visualized using a heatmap and a volcano plot. Sample clustering patterns were examined through a Principal Component Analysis (PCA) plot to explore group separations and to understand the sample variability through clustering.

2.4 Gene set enrichment analysis (GSEA)

Based on the differentially expressed genes (p-value < 0.05), hallmark pathway analysis was performed using the Molecular Signatures Database (MSigDB) with GSEA. The identified pathways were visualized using GSEA plots to understand the gene sets associated with specific pathway enrichment patterns(Liberzon et al., 2015; Reimand et al., 2019).

2.5 Building the machine learning model

2.5.1 Data splitting and preprocessing

Machine learning models, including Random Forest (RF), Neural Network (NN), and Logistic Regression (LR), were implemented using the Scikit-learn (sklearn) package in Python to analyze the filtered TCGA RNA-seq dataset (Lopez-Cortes et al., 2020; Okoro et al., 2021; Ellrott et al., 2024). To ensure robust training and evaluation, the dataset was split into training (70%) and testing (30%) sets using stratified sampling to preserve the proportion of tumor and normal samples.

2.5.2 Model architectures and training

The Random Forest model was constructed using an ensemble learning approach with 100 decision trees, utilizing Gini impurity as the splitting criterion. Hyperparameters such as maximum tree depth and the minimum number of samples required for splits were optimized through grid search and cross-validation. The Neural Network model was designed as a multilayer perceptron with three hidden layers, each consisting of 50 neurons, and employed the rectified linear unit (ReLU) activation function. The Adam optimizer was used for weight updates, and early stopping was applied to mitigate overfitting. Hyperparameters, including the learning rate and batch size, were fine-tuned to maximize performance. The Logistic Regression model employed L2 regularization to prevent overfitting, with the maximum number of iterations set to 1,000 to ensure model convergence. The marker genes were identified based on the AUC values of each gene across the different models, and potential signature genes with consistently high AUC values were selected.

2.5.3 Model evaluation and interpretation

Model performance was assessed using accuracy, area under the receiver operating characteristic curve (AUC), and F1-score as evaluation metrics. The predictions generated by each model were further analyzed to identify significant potential signature genes associated with tumor and normal samples, providing insights into their predictive relevance and biological significance.

3. Results

3.1 Differential gene expression analysis

After pre-filtering and normalization, we identified 10,962 genes with non-zero expression values suitable for further analysis (Supplementary Table 1). Differential expression analysis revealed 657 upregulated and 8,566 downregulated genes in tumor samples compared to matched normal controls (p-value < 0.05). Notably, EPB41L3, TSPAN7, and ABI3BP were among the most upregulated genes, while LYVE1, PLPP1, and NFE2L3. were significantly downregulated. A heatmap of the differentially expressed genes demonstrated distinct segregation between tumor and normal samples, underscoring significant transcriptomic differences (Figure 1A, Supplementary Tables 2 and 3).

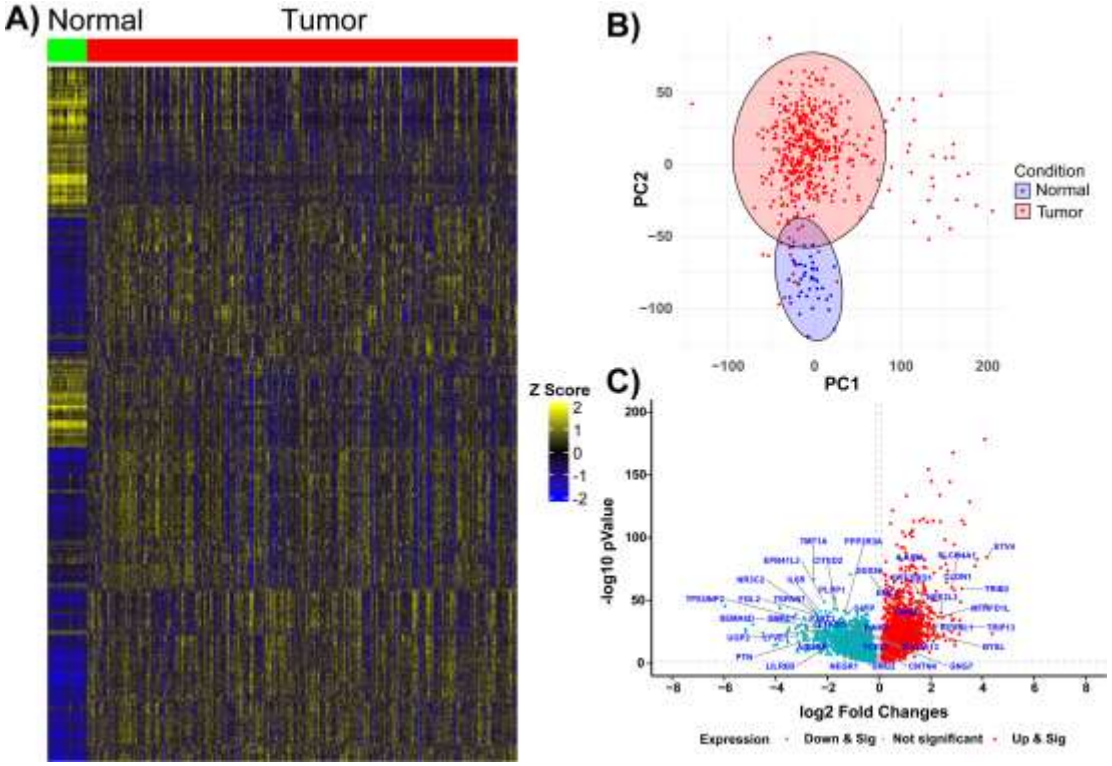

Figure 1. Differentially Expressed Genes in Tumor Tissue Samples of Colon Cancer. A). The heatmap illustrates differentially expressed genes in tumor tissues (n = 444) compared to normal tissues (n = 41) from colon cancer patients. The gradient color scale represents increased expression (yellow) and decreased expression (blue). B). The PCA plot depicts the clustering of tumor and normal tissue samples from colon cancer patients. C). The volcano plot visualizes the distribution of differentially expressed genes between tumor and normal samples. The highlighted genes represent significantly altered genes identified using machine learning models- Random Forest (RF), Neural Network (NN), and Logistic Regression (LR)-all of which achieved AUC values exceeding 97%. Source: Authors, 2025.

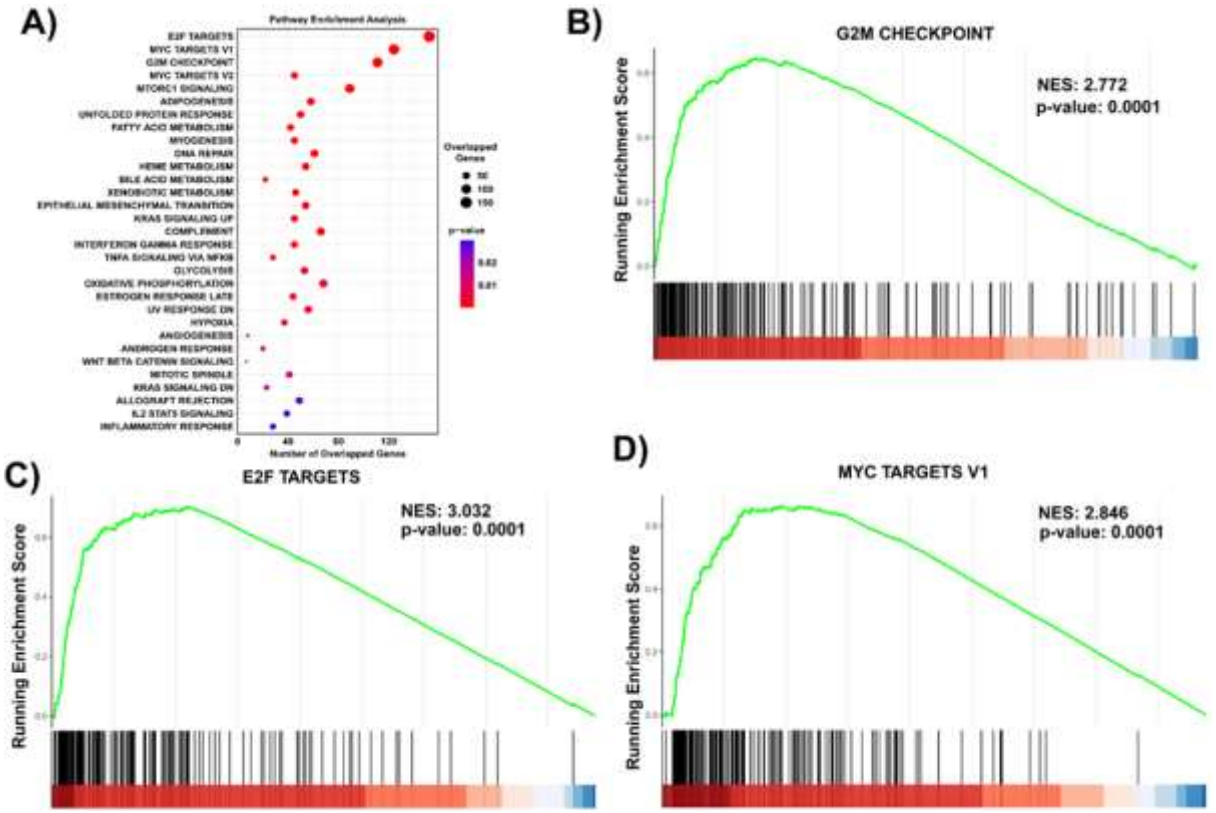

Figure 2. Gene Set Enrichment Analysis (GSEA) Pathway Analysis revealed the enriched biological pathways in colon cancer patients' tumor samples. A). The dot plot represents the enriched pathways identified using GSEA. Pathways are ranked based on enrichment scores, with dot size corresponding to the number of overlapping genes and color indicating statistical significance (adjusted p-value). The enrichment plots display the top three significantly enriched pathways: G2M Checkpoints (B), E2F Targets (C), and MYC Targets (D). The green curve represents the running enrichment score, indicating the accumulation of gene hits as ranked by their differential expression. Vertical black lines denote the positions of pathway genes within the ranked gene list. The normalized enrichment score (NES) and p-values are displayed for each pathway. Source: Authors, 2025.

Principal Component Analysis (PCA) further validated these differences, with tumor and normal samples forming distinct clusters along the principal components, reflecting the unique transcriptional landscapes of each group (Figure 1B, Supplementary Table 2). A volcano plot highlighted the most notable genes, with the top 14 upregulated genes being Bystin like protein (BYSL), Solute Carrier Family 2 Member 13 (SLC2A13), Ectodermal Neural Cortex 1 (ENC1), Ajuba LIM Protein (AJUBA), Tumor Protein p53 Inducible Nuclear Protein 2 (TP53INP2), DEAD Box Helicase 56 (DDX56), Cbp/p300 Interacting Transactivator 2 (CITED2), PTEN Induced Kinase 1 (PINK1), Guanine Nucleotide Binding Protein G(I)/G(S)/G(O) Subunit Gamma 2 (GNG2), Semaphorin 6D (SEMA6D), Claudin 1 (CLDN1).

Erythrocyte Membrane Protein Band 4.1 Like 3 (EPB41L3), Tetraspanin 7 (TSPAN7), ABI Family Member 3 Binding Protein (ABI3BP), and the top 27 downregulated genes including Lymphatic Vessel Endothelial Hyaluronan Receptor 1 (LYVE1), Phospholipid Phosphatase 1 (PLPP1), Nuclear Factor, Erythroid 2 Like 3 (NFE2L3), Electron Transfer Flavoprotein Dehydrogenase (ETFDH), Nuclear Receptor Subfamily 3 Group C Member 2 (NR3C2), Solute Carrier Organic Anion Transporter Family Member 4A1 (SLCO4A1), GTF2I Repeat Domain Containing 1 (GTF2IRD1), Twinkle mtDNA Helicase (TWNK).

ETS Variant Transcription Factor 4 (ETV4), Transcription Factor 21(TCF21), Protein Phosphatase 2 Regulatory Subunit 3 Alpha (PPP2R3A), Sphingomyelin Phosphodiesterase 1 (SMPD1), Glycolipid Transfer Protein (GLTP), RuvB Like AAA ATPase 1 (RUVBL1), Purinergic Receptor P2Y1 (P2RY1), Thyroid Hormone Receptor Interactor 13 (TRIP13), Contactin 4 (CNTN4), Methylenetetrahydrofolate Dehydrogenase (NADP+ Dependent) 1 Like (MTHFD1L), Fibrinogen Like 2 (FGL2), Neuronal Growth Regulator 1 (NEGR1), Interleukin 6 Receptor (IL6R), Thiol Methyltransferase 1A (TMT1A), UDP-Glucose Pyrophosphorylase 2

(UGP2), Tribbles Pseudokinase 3 (TRIB3), Pleiotrophin (PTN), Guanine Nucleotide-Binding Protein G(I)/G(S)/G(O) Subunit Gamma-7 (GNG7), and Leukocyte Immunoglobulin-Like Receptor Subfamily B Member 5 (LILRB5). These genes exhibited highly significant (p-value < 0.05) (Figure 1C, Supplementary Table 2). The differential expression patterns of these genes provide valuable insights into potential biomarkers and therapeutic targets in colon cancer.

3.2 Gene set enrichment analysis (GSEA)

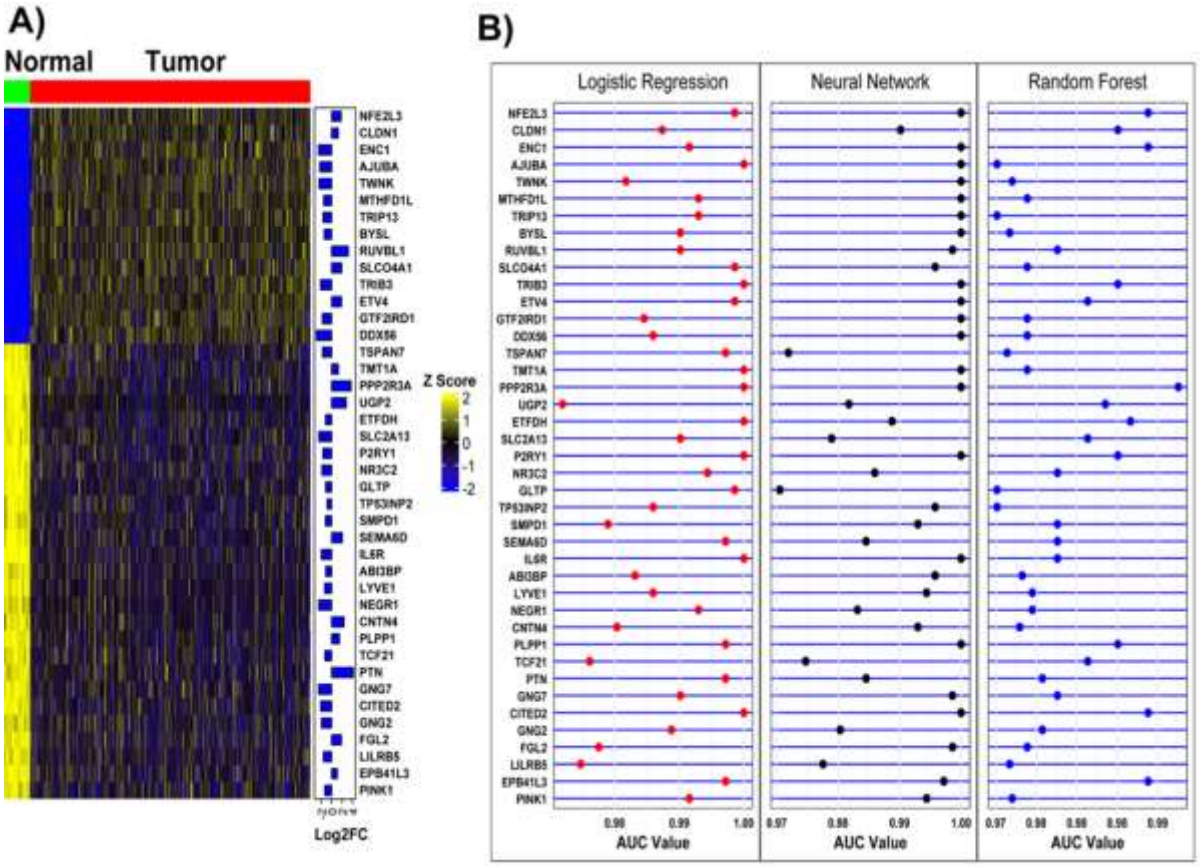

Figure 3. Machine Learning Models Classify Signature Genes in Tumor Patients. A). The heatmap displays differentially expressed genes and their fold changes in tumor tissues compared to normal tissues from colon cancer patients. These genes were identified using machine learning models-Random Forest (RF), Neural Network (NN), and Logistic Regression (LR)-all of which achieved AUC values exceeding 97%. B). The dot plot illustrates the AUC distribution across the three machine learning models in tumor tissue samples, highlighting their classification performance. Source: Authors, 2025.

Gene Set Enrichment Analysis (GSEA) identified several hallmark pathways significantly enriched in tumor samples compared to healthy controls. The top 10 pathways included E2F targets, MYC targets V1, G2M checkpoint, MYC targets V2, MTORC1 signaling, adipogenesis, unfolded protein response, fatty acid metabolism, myogenesis, and DNA repair (Figure 2A, Supplementary Table 4). The top-ranked pathway, E2F targets, involved critical cell cycle regulatory genes such as CDC25B, MYBL2, MYC, TRIP13, and UBE2S (Figure 2A).

These genes are important for cell cycle progression and are transcriptionally regulated by E2F transcription factors, highlighting their significant role in colon cancer pathogenesis. Enrichment plots of key pathways revealed a distinct concentration of altered genes at one end of the ranked gene list, supporting their biological significance in colon cancer progression (Figure 2B-2D). The top three pathways-E2F targets, MYC targets V1, and G2M checkpoint-showed strong positive enrichment in tumor samples relative to normal controls,

reinforcing their critical involvement in tumorigenesis (Figure 2B-2D). These findings underscore the dysregulation of core processes such as cell cycle control, metabolic signaling, and stress responses in colon cancer. The enrichment of pathways like MTORC1 signaling and unfolded protein response points to metabolic rewiring and cellular stress adaptation as key features of tumor biology. Additionally, the dysregulation of fatty acid metabolism and adipogenesis suggests a potential link between lipid metabolism and tumor progression.

3.3 Machine learning analysis

The machine learning models, including Random Forest (RF), Neural Network (NN), and Logistic Regression (LR), achieved robust classification performance in distinguishing tumors from normal samples. The RF model demonstrated the highest accuracy (97%) and AUC (1.00), followed by NN (99%, AUC = 1.00) and LR (99%, AUC = 1.00). Area under the curve (AUC > 0.97) analysis from the RF, NN, and LR models identified 41 top contributors' genes to classification performance. These genes were consistently highlighted across models, indicating their potential as robust biomarkers, and were significantly altered in tumor samples compared to healthy samples in colon cancer (Figure 3, Supplementary Tables 5 and 6).

The AUC-ROC curves demonstrate the models' performance in distinguishing between classes, with the area under the curve (AUC) serving as a metric for classification accuracy. Higher AUC values indicate better model performance. The k-fold cross-validation approach ensures robustness and generalizability by partitioning the dataset into multiple subsets for training and validation (Supplementary Figure 1). The top 3 genes ($\log^2$ FC > 3.00) of EPB41L3, TSPAN7, and ABI3BP genes were significantly increased, whereas the other top 3 genes ($\log^2$ FC < -2.5) of LYVE1, PLPP1, and NFE2L3 were significantly decreased in tumor samples compared to normal samples in colon cancer (Figure 3).

3.4 Identification of potential biomarkers

The integrative analysis, leveraging differential gene expression, Gene Set Enrichment Analysis (GSEA), and machine learning approaches, identified a subset of genes with high predictive potential in colon cancer. The top three upregulated genes were EPB41L3, TSPAN7, and ABI3BP, which exhibited significantly increased expression in tumor samples compared to normal samples (Figure 3). These genes are likely involved in tumor progression and could serve as potential targets for therapeutic intervention. Conversely, the top three downregulated genes were LYVE1, PLPP1, and NFE2L3, which were significantly suppressed in tumor samples (Figure 3).

The table lists the top 20 genes selected based on their classification performance in distinguishing colon cancer from benign samples using multiple machine learning models. Gene selection was based on their contribution to model accuracy, such as area under the curve (AUC) (Table 1). Their decreased expression may indicate disruption in pathways critical for maintaining normal cellular homeostasis and immune response. These findings underscore the potential of these genes as robust biomarkers for distinguishing tumors from normal samples. Furthermore, their consistent identification across multiple analytical approaches lays a strong foundation for subsequent validation studies. Ultimately, these biomarkers hold promise for advancing colon cancer diagnostics and therapeutics, paving the way for personalized medicine strategies.

Table 1. Top 20 differentially expressed genes identified in colon cancer compared to benign tissues using TCGA RNA-seq data and machine learning models.

Genes	AUC Values		
	Logistic Regression	Neural Network	Random Forest
CITED2	1.000000	1.000000	0.988806
CLDN1	0.987360	0.990169	0.985075
ENC1	0.991573	1.000000	0.988806
EPB41L3	0.997191	0.997191	0.988806
ETFDH	1.000000	0.988764	0.986629
ETV4	0.998596	1.000000	0.981343
GNG7	0.990169	0.998596	0.977612
IL6R	1.000000	1.000000	0.977612
NFE2L3	0.998596	1.000000	0.988806
NR3C2	0.994382	0.985955	0.977612
P2RY1	1.000000	1.000000	0.985075
PLPP1	0.997191	1.000000	0.985075
PPP2R3A	1.000000	1.000000	0.992537
RUVBL1	0.990169	0.998596	0.977612
SEMA6D	0.997191	0.984551	0.977612
SLC2A13	0.990169	0.978933	0.981343
SMPD1	0.978933	0.992978	0.977612
TCF21	0.976124	0.974719	0.981343
TRIB3	1.000000	1.000000	0.985075
UGP2	0.971910	0.981742	0.983520

Source: Authors, 2025.

4. Discussion

This study demonstrates the power of integrating transcriptomics and machine learning to uncover robust biomarkers and pathways in colon cancer. Our findings reveal distinct transcriptional and pathway-level alterations that differentiate tumors from normal samples, providing critical insights into colon cancer biology. By combining differential gene expression analysis, GSEA, and machine learning approaches, we uncovered distinct transcriptional and pathway-level alterations that differentiate tumors from normal samples. These findings not only deepen our understanding of colon cancer biology but also lay a foundation for the development of diagnostic and therapeutic strategies.

The differential gene expression analysis revealed a substantial number of genes with altered expression in tumor samples, including 657 upregulated and 8,566 downregulated genes. Among these, EPB41L3, TSPAN7, and ABI3BP emerged as the most upregulated genes ($\log^2$ FC > 3.00), suggesting their potential involvement in tumor progression. These genes have been implicated in cellular adhesion, signaling, and modulation of the tumor microenvironment, all critical processes in cancer (Barabasi et al., 2011). Conversely, the top three downregulated genes, LYVE1, PLPP1, and NFE2L3 ($\log^2$ FC < -2.5), suggest disrupted immune signaling and lipid metabolism in tumor tissues. For instance, LYVE1 is linked to lymphatic vessel integrity and immune regulation (Subramanian et al., 2005), and its suppression may contribute to immune evasion. Similarly, PLPP1 and NFE2L3, associated with lipid metabolism and oxidative stress (Tomczak et al., 2015), underscore the metabolic vulnerabilities of tumor cells.

Gene Set Enrichment Analysis (GSEA) provided additional insights into the pathways disrupted in colon cancer. Key pathways, including E2F targets, MYC targets, and G2M checkpoints, were significantly enriched in tumor samples. These findings emphasize dysregulated cell cycle control, proliferation, and metabolic rewiring as hallmarks of colon cancer (Wang et al., 2009). The identification of pathways such as MTORC1 signaling and unfolded protein response highlights the ability of colon cancer cells to adapt to metabolic stress and optimize survival in nutrient-limited environments (Libbrecht; Noble, 2015). Enrichment of pathways like fatty acid metabolism and adipogenesis further underscores the interplay between lipid metabolism and tumor progression.

Machine learning analyses reinforced the robustness of the identified biomarkers. The Random Forest, Neural Network, and Logistic Regression models achieved high classification accuracy (97-99%) and AUC values

(1.00), demonstrating their predictive power in distinguishing tumors from normal samples. Importantly, identifying key genes across models consistently underscores their translational relevance. Genes such as EPB41L3 (Son et al., 2020), TSPAN7 (Qi et al., 2020), and ABI3BP (Horpaopan et al., 2017; Latini et al., 2011; Chen et al., 2021; Nong et al., 2021), as well as LYVE1 (Parr; Jiang, 2003; Sundov et al., 2013; Capuano et al., 2019), PLPP1 (Tang; Brindley, 2020), and NFE2L3 (Palma et al., 2012; Aono et al., 2019; Bury et al., 2019; Saliba et al., 2022), were consistently highlighted as significant contributors to classification performance. These results confirm the utility of integrating transcriptomics with machine learning for biomarker discovery (Libbrecht; Noble, 2015).

The observed transcriptional and pathway alterations align with previous studies while also providing novel insights into the molecular underpinnings of colon cancer (Dunne; Arends, 2024). For instance, the enrichment of E2F and MYC target pathways supports the role of dysregulated transcriptional networks in driving tumor growth (Subramanian et al., 2005; Johnson et al., 2016; Oshi et al., 2020). Additionally, the suppression of immune-related genes, such as LYVE1 (Parr; Jiang, 2003; Sundov et al., 2013; Capuano et al., 2019), highlights potential mechanisms of immune evasion, which may be critical for tumor survival and progression (Viudez-Pareja et al., 2023).

While this study provides valuable insights, some limitations warrant further investigation. Validation in independent cohorts and experimental models is essential to confirm the identified biomarkers and pathways. Additionally, functional studies are needed to elucidate the precise roles of these genes and pathways in colon cancer pathogenesis. Integrating additional omics datasets, such as proteomics or metabolomics, could offer a more comprehensive understanding of tumor biology and uncover additional therapeutic opportunities (Barabasi et al., 2011).

5. Conclusion

Our integrative approach identifies key molecular signatures in colon cancer, offering promising candidates for diagnostic and therapeutic development. Future studies should focus on validating these biomarkers in independent cohorts and exploring their functional roles to advance precision medicine in colon cancer. Future studies focusing on experimental validation and clinical translation will be crucial for leveraging these findings to improve patient outcomes in colon cancer.

6. Authors' Contributions

Mostafa Amir Hamza: conceptualization, methodology, data collection, manuscript preparation, writing-review and editing. *Saiful Islam*: conceptualization, manuscript preparation, writing-review and editing.

7. Conflicts of Interest

The authors declare no competing interests.

8. Ethics Approval

Not applicable.

9. References

Aono, S., Hatanaka, A., Hatanaka, A., Gao, Y., Hippo, Y., Taketo, M. M., Waku, T., & Kobayashi, A. (2019). beta-Catenin/TCF4 complex-mediated induction of the NRF3 (NFE2L3) gene in cancer cells. *International Journal of Molecular Sciences*, 20(13). https://doi.org/10.3390/ijms20133344

Augustus, G. J., & Ellis, N. A. (2018). Colorectal cancer disparity in african americans: Risk factors and carcinogenic mechanisms. *The American Journal of Pathology*, 188(2), 291-303. https://doi.org/10.1016/j.ajpath.2017.07.023

Barabasi, A. L., Gulbahce, N., & Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1), 56-68. https://doi.org/10.1038/nrg2918

Bury, M., Le Calve, B., Lessard, F., Dal Maso, T., Saliba, J., Michiels, C., Ferbeyre, G., & Blank, V. (2019). NFE2L3 Controls colon cancer cell growth through regulation of DUX4, a CDK1 inhibitor. *Cell Reports*, 29(6), 1469-1481 e1469. https://doi.org/10.1016/j.celrep.2019.09.087

Capuano, A., Pivetta, E., Sartori, G., Bosisio, G., Favero, A., Cover, E., Andreuzzi, E., Colombatti, A., Cannizzaro, R., Scanziani, E., Minoli, L., Bucciotti, F., Amor Lopez, A. I., Gaspardo, K., Doliana, R., Mongiat, M., & Spessotto, P. (2019). Abrogation of EMILIN1-beta1 integrin interaction promotes experimental colitis and colon carcinogenesis. *Matrix Biology*, 83, 97-115. https://doi.org/10.1016/j.matbio.2019.08.006

Chen, W., Huang, J., Xiong, J., Fu, P., Chen, C., Liu, Y., Li, Z., Jie, Z., & Cao, Y. (2021). Identification of a Tumor Microenvironment-Related Gene Signature Indicative of Disease Prognosis and Treatment Response in Colon Cancer. *Oxidative Medicine and Cellular Longevity*, 2021, 6290261. https://doi.org/10.1155/2021/6290261

Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T. S., Malta, T. M., Pagnotta, S. M., Castiglioni, I., Ceccarelli, M., Bontempi, G., & Noushmehr, H. (2016). TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res*, 44(8), e71. https://doi.org/10.1093/nar/gkv1507

Dunne, P. D., & Arends, M. J. (2024). Molecular pathological classification of colorectal cancer-an update. *Virchows Arch*, 484(2), 273-285. https://doi.org/10.1007/s00428-024-03746-3

Ellrott, K., Wong, C. K., Yau, C., Castro, M. A. A., Lee, J. A., Karlberg, B. J., Grewal, J. K., Lagani, V., Tercan, B., Friedl, V., Hinoue, T., Uzunangelov, V., Westlake, L., Loinaz, X., Felau, I., Wang, P. I., Kemal, A., Caesar-Johnson, S. J., Shmulevich, I. & Laird, P. W. (2024). Classification of non-TCGA cancer samples to TCGA molecular subtypes using compact feature sets. *Cancer Cell*. https://doi.org/10.1016/j.ccell.2024.12.002

Horpaopan, S., Kirfel, J., Peters, S., Kloth, M., Huneburg, R., Altmuller, J., Drichel, D., Odenthal, M., Kristiansen, G., Strassburg, C., Nattermann, J., Hoffmann, P., Nurnberg, P., Buttner, R., Thiele, H., Kahl, P., Spier, I., & Aretz, S. (2017). Exome sequencing characterizes the somatic mutation spectrum of early serrated lesions in a patient with serrated polyposis syndrome (SPS). *Hereditary Cancer in Clinical Practice*, 15, 22. https://doi.org/10.1186/s13053-017-0082-9

Johnson, J., Thijssen, B., McDermott, U., Garnett, M., Wessels, L. F., & Bernards, R. (2016). Targeting the RB-E2F pathway in breast cancer. *Oncogene*, 35(37), 4829-4835. https://doi.org/10.1038/onc.2016.32

Latini, F. R., Hemerly, J. P., Freitas, B. C., Oler, G., Riggins, G. J., & Cerutti, J. M. (2011). ABI3 ectopic expression reduces in vitro and in vivo cell growth properties while inducing senescence. *BMC Cancer*, 11, 11. https://doi.org/10.1186/1471-2407-11-11

Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6), 321-332. https://doi.org/10.1038/nrg3920

Liberzon, A., Birger, C., Thorvaldsdottir, H., Ghandi, M., Mesirov, J. P., & Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst*, 1(6), 417-425. https://doi.org/10.1016/j.cels.2015.12.004

Lopez-Cortes, A., Cabrera-Andrade, A., Vazquez-Naya, J. M., Pazos, A., Gonzales-Diaz, H., Paz, Y. M. C., Guerrero, S., Perez-Castillo, Y., Tejera, E., & Munteanu, C. R. (2020). Prediction of breast cancer proteins involved in immunotherapy, metastasis, and RNA-binding using molecular descriptors and artificial neural networks. *Scientific Report*, 10(1), 8515. https://doi.org/10.1038/s41598-020-65584-y

Mounir, M., Lucchetta, M., Silva, T. C., Olsen, C., Bontempi, G., Chen, X., Noushmehr, H., Colaprico, A., & Papaleo, E. (2019). New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx. *PLoS Computational Biology*, 15(3), e1006701. https://doi.org/10.1371/journal.pcbi.1006701

Nong, B., Guo, M., Wang, W., Songyang, Z., & Xiong, Y. (2021). Comprehensive Analysis of Large-Scale Transcriptomes from Multiple Cancer Types. *Genes (Basel)*, 12(12). https://doi.org/10.3390/genes12121865

Okoro, P. C., Schubert, R., Guo, X., Johnson, W. C., Rotter, J. I., Hoeschele, I., Liu, Y., Im, H. K., Luke, A., Dugas, L. R., & Wheeler, H. E. (2021). Transcriptome prediction performance across machine learning models and diverse ancestries. *HGG Advances*, 2(2). https://doi.org/10.1016/j.xhgg.2020.100019

Oshi, M., Takahashi, H., Tokumaru, Y., Yan, L., Rashid, O. M., Nagahashi, M., Matsuyama, R., Endo, I., & Takabe, K. (2020). The E2F Pathway Score as a Predictive Biomarker of Response to Neoadjuvant Therapy in ER+/HER2- Breast Cancer. *Cells*, 9(7). https://doi.org/10.3390/cells9071643

Palma, M., Lopez, L., Garcia, M., de Roja, N., Ruiz, T., Garcia, J., Rosell, E., Vela, C., Rueda, P., & Rodriguez, M. J. (2012). Detection of collagen triple helix repeat containing-1 and nuclear factor (erythroid-derived 2)-like 3 in colorectal cancer. *BMC Clinical Pathology*, 12, 2. https://doi.org/10.1186/1472-6890-12-2

Parr, C., & Jiang, W. G. (2003). Quantitative analysis of lymphangiogenic markers in human colorectal cancer. *International Journal of Oncology*, 23(2), 533-539. https://doi.org/10.3892/ijo.23.2.533

Qi, Y., Li, H., Lv, J., Qi, W., Shen, L., Liu, S., Ding, A., Wang, G., Sun, L., & Qiu, W. (2020). Expression and function of transmembrane 4 superfamily proteins in digestive system cancers. *Cancer Cell Internation*, 20, 314. https://doi.org/10.1186/s12935-020-01353-1

Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., Wadi, L., Meyer, M., Wong, J., Xu, C., Merico, D., & Bader, G. D. (2019). Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, cytoscape and enrichmentMap. *Nature Protocols*, 14(2), 482-517. https://doi.org/10.1038/s41596-018-0103-9

Saliba, J., Coutaud, B., Makhani, K., Epstein Roth, N., Jackson, J., Park, J. Y., Gagnon, N., Costa, P., Jeyakumar, T., Bury, M., Beauchemin, N., Mann, K. K., & Blank, V. (2022). Loss of NFE2L3 protects against inflammation-induced colorectal cancer through modulation of the tumor microenvironment. *Oncogene*, 41(11), 1563-1575. https://doi.org/10.1038/s41388-022-02192-2

Sawicki, T., Ruszkowska, M., Danielewicz, A., Niedzwiedzka, E., Arlukowicz, T., & Przybylowicz, K. E. (2021). A review of colorectal cancer in terms of epidemiology, risk factors, development, symptoms and diagnosis. *Cancers (Basel)*, 13(9). https://doi.org/10.3390/cancers13092025

Siegel, R. L., Giaquinto, A. N., & Jemal, A. (2024). Cancer statistics, 2024. *CA Cancer Journal for Clinicians*, 74(1), 12-49. https://doi.org/10.3322/caac.21820

Siegel, R. L., Wagle, N. S., Cercek, A., Smith, R. A., & Jemal, A. (2023). Colorectal cancer statistics, 2023. *CA: A Cancer Journal for Clinicians*, 73(3), 233-254. https://doi.org/https://doi.org/10.3322/caac.21772

Son, H. J., Choi, E. J., Yoo, N. J., & Lee, S. H. (2020). Mutation and expression of a candidate tumor suppressor gene EPB41L3 in gastric and colorectal cancers. *Pathology & Oncology Research*, 26(3), 2003-2005. https://doi.org/10.1007/s12253-019-00787-x

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43), 15545-15550. https://doi.org/10.1073/pnas.0506580102

Sundov, Z., Tomic, S., Alfirevic, S., Sundov, A., Capkun, V., Nincevic, Z., Nincevic, J., Kunac, N., Kontic, M., Poljak, N., & Druzijanic, N. (2013). Prognostic value of MVD, LVD and vascular invasion in lymph node-negative colon cancer. *Hepatogastroenterology*, 60(123), 432-438. https://doi.org/10.5754/hge12826

Tang, X., & Brindley, D. N. (2020). Lipid Phosphate Phosphatases and Cancer. *Biomolecules*, 10(9). https://doi.org/10.3390/biom10091263

Tomczak, K., Czerwinska, P., & Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary Oncology (Pozn)*, 19(1A), A68-77. https://doi.org/10.5114/wo.2014.47136

Viudez-Pareja, C., Kreft, E., & Garcia-Caballero, M. (2023). Immunomodulatory properties of the lymphatic endothelium in the tumor microenvironment. *Frontiers Immunology*, 14, 1235812. https://doi.org/10.3389/fimmu.2023.1235812

Walter Reed National Military Medical Center. (2024). *Colorectal Cancer Awareness Month: Early detection is the best prevention*. https://walterreed.tricare.mil/News-Gallery/Articles/Article/3719070/colorectal-cancer-awareness-month-early-detection-is-the-best-prevention#:~:text=According%20to%20the%20American%20Cancer,men%20and%2019%2C890%20in%20women).

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1), 57-63. https://doi.org/10.1038/nrg2484

Xi, Y., & Xu, P. (2021). Global colorectal cancer burden in 2020 and projections to 2040. *Translational Oncology*, 14(10), 101174. https://doi.org/10.1016/j.tranon.2021.101174

Funding

Not applicable.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.