Applied Medical Informatics
Original Research Vol. 47, No.4/2025, pp 121-138

Exploratory Transcriptomic Analysis of Colorectal Cancer: Identification of Highly Variable Genes and Co-expression Patterns

Amir Mohammad MAZHARI*

Faculty of Computer Science and Engineering, Shahid Beheshti University, Tehran, Iran.
E-mail: am.mazhari@mail.sbu.ac.ir

* Author to whom correspondence should be addressed;

Received: 5 July 2025/Accepted: 15 November 2025 / Published online: 12 December 12, 2025

Abstract
Background: Gene expression variability represents an essential dimension of transcriptomic complexity, reflecting biological heterogeneity and regulatory diversity across tumors. Characterizing such variability may reveal candidate biomarkers and co-regulated gene modules relevant to colorectal cancer biology. *Purpose:* The study aimed to develop a reproducible computational framework for identifying and visualizing highly variable genes within colorectal cancer transcriptomic data, providing a foundation for exploratory analysis and hypothesis generation. *Methods:* A modular Python-based pipeline was constructed to process microarray data derived from colon cancer patients included in the GSE39582 cohort. Data interrogation was performed in October 2025. Following preprocessing, probe-to-gene annotation, and log-transformation, gene-wise variance was calculated. The top 0.1% of genes ranked by variance were selected as highly variable genes. Visualization included z-score–normalized heatmaps, boxplots, and correlation matrices to illustrate heterogeneity and co-expression patterns. *Results:* Analysis revealed a small subset of genes exhibiting markedly heterogeneous expression profiles across the colorectal cancer cohort. Variability patterns suggested the existence of co-regulated gene modules and potential subtype-associated transcriptional programs. Genes previously linked to colorectal tumorigenesis, such as OLFM4, MS4A12, and CEACAM7, were among the most variable, supporting the biological relevance of variance-based selection. *Conclusions:* The developed pipeline provides a transparent and reproducible framework for rapid exploration of transcriptomic variability in colorectal cancer. Its simplicity and adaptability make it suitable for integration into diverse analytical workflows and for educational or exploratory research applications.

Keywords*:* Colorectal Neoplasms; Gene Expression Profiling; Biological Variability; Transcriptome; Data Analysis Pipeline

Introduction

Gene expression profiling through microarray or ribonucleic acid (RNA) sequencing technologies has become a fundamental approach for exploring the molecular mechanisms underlying complex biological systems, including cancer. One critical aspect of transcriptomic data analysis is identifying genes with high expression variability across samples. Genes with such variability may indicate underlying biological heterogeneity, serve as potential biomarkers, or reflect subgroup-specific regulatory activity. By leveraging advanced visualization techniques such as heatmaps, boxplots, and correlation matrices, we can uncover latent structures and co-regulated modules within the dataset, providing more granular insights into gene behavior. Gene expression variability is increasingly recognized as a core feature of biological complexity. Single-cell studies have demonstrated that genes exhibiting high variability often underlie functionally significant differences between biological states [1]. In the context of colorectal cancer (CRC), expression variability can reflect tumor heterogeneity and differential pathway activation

across tumor regions. For example, Árnadóttir et al. showed that transcriptional and proteomic profiles in CRC vary significantly depending on tumor location within the colon, emphasizing the extent of intra-tumor heterogeneity [2].

While previous large-scale transcriptomic studies in CRC have identified clinically relevant molecular subtypes using clustering techniques, they have generally focused on discovering subtype-specific markers rather than systematically analyzing gene-wise variance. Marisa et al., as part of the CIT consortium, analyzed data from 443 colon tumors (including samples from the GSE39582 dataset) and identified six robust molecular subtypes with distinct prognostic implications [3]. Such investigations highlight the value of expression-based stratification, but differ in their methodological focus from the current study. Mo et al. applied weighted gene co-expression network analysis (WGCNA) to the GSE39582 dataset, focusing on early-onset CRC, identifying gene modules associated with tumor stage and several hub genes. They reported 140 module hub genes enriched in ribosomal and extracellular matrix functions, of which seven (e.g. SPARC, DCN, FBN1) had prognostic significance for early-onset CRC [4]. Langerud et al. performed multiregional transcriptomic profiling of 692 CRC patients (including 98 primary multiregional samples and 35 primary–metastasis pairs) to assess intra-tumor heterogeneity and its impact on molecular subtypes. They found frequent discordance of consensus molecular subtypes (CMS) within tumors and showed that "intrinsic" CMS signals—derived from cancer-cell–intrinsic expression—explained more variation in patient survival than traditional bulk subtypes. The authors' findings highlighted the importance of analyzing intra-tumor variability in understanding CRC prognosis [5]. Kamal et al. investigated the role of the tumor immune microenvironment in CRC prognosis by deconvoluting bulk expression data. In four independent cohorts (including GSE39582), they inferred infiltration of 22 innate and adaptive immune cell types. High infiltration of activated CD4+ memory T cells was associated with longer relapse-free survival, independent of stage and MSI status. Their study underscored the utility of immune cell signatures as potent prognostic biomarkers in CRC, providing a potential avenue for therapeutic targeting [6].

Angius et al. provided a comprehensive review of intra-tumor heterogeneity in CRC using bulk and single-cell omics. They emphasized how genomic and transcriptomic sequencing of individual tumor cells can reveal subclonal diversity and lineage hierarchies, uncovering rare cell populations (like cancer stem cells) that drive tumor evolution and therapy resistance. The review highlighted the promise of single-cell approaches for precision oncology in CRC, illustrating how differentially expressed genes across cell clusters serve as unique markers of specific tumor cell subpopulations [7].

The Cancer Genome Atlas (TCGA) consortium conducted a seminal multi-omic characterization of colon and rectal cancers. Analyzing exomes, copy number, methylation, and transcriptomes in 276 samples, they identified ~16% hypermutated tumors (most with MSI and MLH1 silencing). They reported 24 significantly mutated genes, including expected drivers (APC, TP53, KRAS, SMAD4, PIK3CA) and novel ones (ARID1A, SOX9, FAM123B). The TCGA study also discovered recurrent copy-number alterations (e.g., ERBB2 and IGF2 amplifications) and chromosomal translocations (e.g., NAV2–TCF7L1). The analysis set the stage for subsequent biomarker identification efforts in CRC [8].

Van Dijk et al. addressed technical challenges in single-cell RNA-seq by developing the MAGIC algorithm, which uses graph-based data diffusion to impute missing transcripts ("dropouts"). Applying MAGIC across diverse systems, they demonstrated recovery of gene–gene relationships and uncovering of continuous cellular states that were obscured by noise. Their work underscores the importance of computational approaches in revealing underlying gene expression patterns in single-cell data [9]. Tirosh et al. performed single-cell RNA sequencing of 4,347 cells from six IDH-mutant human oligodendrogliomas. They reconstructed a developmental hierarchy in these tumors, finding that most cells differentiated along two glial programs. At the same time, a rare subpopulation retained a neural stem cell–like transcriptome and enriched for proliferation. Cells with stem-like signatures were inferred to be the cancer stem cell population fueling tumor growth. Notably, distinct genetic subclones (with different copy-number and point mutation profiles) showed similar expression hierarchies, suggesting that tumor architecture is primarily dictated by developmental programs rather than genetics. The findings highlight the relevance of expression variability in defining cancer stem cell populations and tumor progression [10].

Dunne et al. and Isella et al. independently tackled the problem of stromal contamination in CRC expression data by isolating cancer cell–intrinsic signals. Dunne et al. showed that CRC samples cluster more by patient-of-origin than tumor region when using an intrinsic gene signature (CRIS), indicating that cancer-cell expression is

robust to stromal heterogeneity. Similarly, Isella et al. profiled cancer-cell expression from patient-derived xenografts (where human stroma is absent) and defined five intrinsic subtypes (CRIS A–E) with distinct molecular features. Both investigations highlight the value of focusing on intrinsic expression signatures in CRC for more accurate subtype classification and prognosis [11, 12].

Joanito et al. further refined CRC taxonomy by integrating >373,000 single-cell transcriptomes (49,155 epithelial cells) with bulk data from 3,614 patients. They discovered a pervasive dichotomy of malignant epithelial cells, defining two intrinsic subtypes (termed iCMS2 and iCMS3) that underlie the bulk CMS classification. The iCMS3 comprised all MSI-high tumors and one-third of microsatellite-stable (MSS) tumors; notably, the iCMS3 MSS tumors were transcriptomically more similar to MSI-high cancers than to other MSS cases. In contrast, CMS4 tumors contained both iCMS2 and iCMS3 epithelium, with the iCMS3 cases showing the worst prognosis. Their findings proposed a novel multi-factor classification system for CRC that integrates intrinsic subtype, MSI status, and fibrosis level [13].

Hoorn et al. conducted a systematic review and meta-analysis on the clinical impact of CMS subtypes in CRC. They found that in non-metastatic CRC, CMS4 tumors had significantly worse overall survival than CMS1 or CMS2 cancers (e.g. HR≈2.6–3.3 for CMS4 vs CMS1/2). In metastatic CRC, CMS1 consistently showed poorer survival compared to CMS2–4. They also summarized subtype-specific treatment effects: adjuvant chemotherapy improved OS mainly in CMS2/3 (but not CMS4), and in metastatic CMS4 cancers, an irinotecan-based regimen yielded better outcomes than oxaliplatin. The review solidifies the clinical relevance of CMS subtypes and their potential to guide personalized treatment [14].

Kim et al. integrated heterogeneous CRC gene-expression cohorts (TCGA-COAD RNA-seq and microarray datasets GSE17536 and GSE39582) to identify prognostic biomarkers. They applied Cox regression and random forest selection to genes, ultimately nominating 28 RNA biomarkers enriched in cancer-related pathways (notably EGFR and IGF1R signaling). From this set of biomarkers, they built a prognostic model using three long noncoding RNAs (ZEB1-AS1, PI4K2A, ITGB8-AS1) along with clinical variables (stage, age), which significantly improved survival prediction compared to clinical factors alone. The approach highlights the potential of RNA biomarkers for enhancing survival prediction and therapeutic decision-making in CRC [15].

The present study introduces a computational pipeline for exploratory analysis of highly variable genes (HVGs) using publicly available microarray data from the GSE39582 [3] dataset, which comprises colon cancer samples. Our approach involves standardized preprocessing, annotation mapping, and visualization through multiple statistical plots. In contrast to classification-based pipelines, the method aims to support hypothesis generation by identifying and visualizing gene-level variability patterns that may suggest biological relevance or regulatory coordination. Although the dataset focuses on colon tumors, the discussion interprets the results in the broader context of CRC.

Materials and Methods

A modular and reproducible computational pipeline was developed in Python (v3.13) to analyze gene expression variability in colon cancer using the GSE39582 [3] microarray dataset. The pipeline was implemented with established Python libraries, where *pandas* [16], *numpy* [17], and *scikit-learn* [18] were used for statistical analysis, and *matplotlib* [19] and *seaborn* [20] for visualization. The workflow consists of five stages: data acquisition and preprocessing, probe-to-gene mapping, normalization, identification of highly variable genes (HVGs), and visualization with summary statistics. An overview of the computational workflow is illustrated in Figure 1.

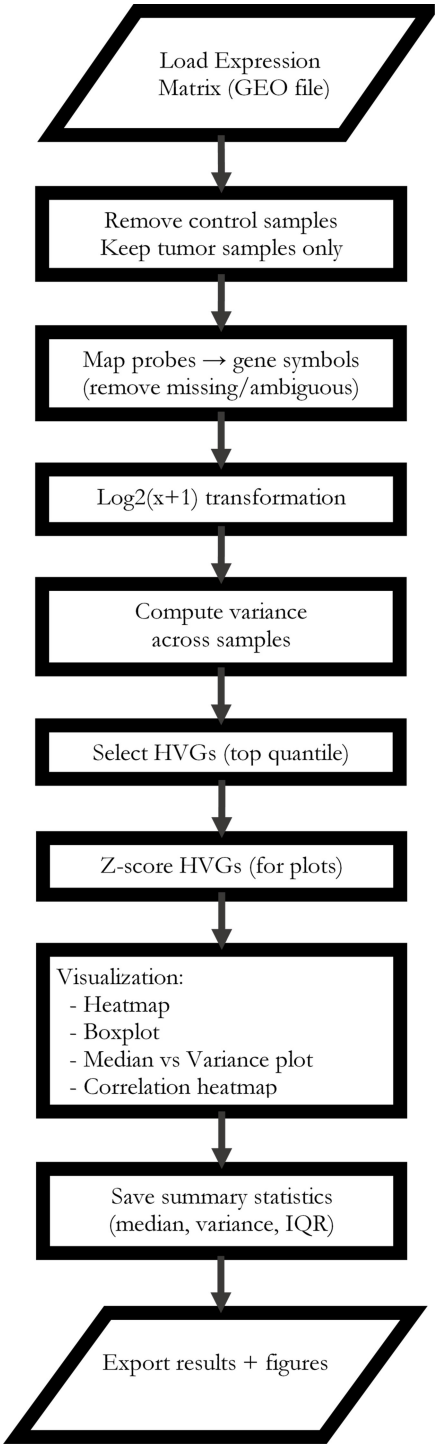

Figure 1. Workflow of the computational pipeline for preprocessing and analysis of microarray expression data. Following loading of the expression matrix and exclusion of control samples, probes were mapped to gene symbols using the corresponding platform annotation file, with entries lacking clear assignments removed. Expression values were subjected to $\log_2(x+1)$ transformation, and gene-wise variance was computed across samples. Highly variable genes (HVGs) were defined as the top 0.1% ranked by variance. HVG expression values were subsequently standardized by z-score scaling for visualization, enabling the generation of heatmaps, boxplots, variance–median scatterplots, and correlation heatmaps. Summary statistics (median, variance, and interquartile range) were calculated for HVGs as well as the full gene set, and both numerical outputs and figures were exported.

Data Acquisition and Preprocessing

The expression matrix was retrieved from the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) [21] using the GEO DataSets Advanced Search Builder. The query string applied was:

(((((("colorectal") OR "colon") OR "rectum") OR "rectal") AND "cancer") AND "homo sapiens") AND "expression profiling by array") AND molecular markers

The query returned two datasets: GSE70880 (152 samples, GPL19748 platform, including lncRNA and mRNA expression profiles across several cancer types) and GSE39582 (585 samples, GPL570 platform, focused on colon cancer with comprehensive clinical and molecular annotations). Of the two datasets, GSE39582 was selected because it provides a large, well-characterized colon cancer cohort suitable for transcriptome-based analyses [3]. The data were retrieved as a compressed series matrix file (accessed in October 2025).

Following the extraction of the relevant data section, only expression intensity values corresponding to biological tumor samples were retained, while non-tumoral controls were excluded. The resulting matrix was transposed to the conventional orientation, with rows representing samples and columns representing probe identifiers. The transposition step ensured compatibility with downstream statistical analyses.

Probe-to-Gene Mapping

Probe-to-gene relationships were defined based on the platform annotation file for GPL570 [22] (accessed in October 2025). In cases where a probe was annotated to multiple genes, only one representative gene symbol was retained to ensure a one-to-one correspondence between probes and gene symbols. This approach minimizes ambiguity in probe annotation and ensures consistency across downstream analyses.

Normalization and Gene Selection

Gene expression values were $\log_2(x+1)$ transformed to stabilize variance and approximate normality, a standard practice in microarray analysis.

The calculation of gene-wise variance was performed on log2-transformed expression data, which is methodologically preferable for gene expression analysis. Such a transformation stabilizes variance across the dynamic range of expression values and reduces the influence of extreme outliers, thereby providing a more biologically meaningful assessment of variability that is not disproportionately driven by high expressed genes.

Variance was then calculated across all tumor samples on the log-transformed scale. Highly variable genes (HVGs) were identified by applying a quantile-based threshold, corresponding to the top 0.1% of genes ranked by variance. The stringent criterion was chosen to capture the most dominant sources of biological heterogeneity while maintaining interpretability.

The focus on the top 0.1% of variance was motivated by two main considerations: (i) to capture the most pronounced changes, highlighting genes with relatively high variability across patients; and (ii) to enhance comparability and interpretability, as presenting a limited set of genes allows for clearer and more manageable tables and visualizations.

Z-score normalization was subsequently applied to HVGs exclusively for visualization purposes (heatmaps and boxplots), enabling direct comparison of expression dynamics across samples.

Visualization and Variability Analysis

The selected HVGs were analyzed using complementary visualization approaches: (i) a heatmap of z-score–normalized expression values was generated to reveal block-like expression patterns across tumor samples, (ii) a boxplot illustrated the distribution of expression values for each gene, highlighting variability and potential outliers, (iii) a scatterplot of median expression versus variance for all genes was constructed on the log-transformed scale to illustrate global patterns of variability, and (iv) a correlation heatmap was computed to evaluate

co-expression structure among HVGs. Taken together, the visualizations provide complementary insights into the heterogeneity of transcriptomic profiles in colon cancer.

Summary Statistics

Descriptive statistics, including variance, median, and interquartile range (IQR), were calculated for each gene on the log2-transformed scale. These metrics were stored in tabular format to provide a quantitative reference for exploratory analyses and potential hypothesis generation in subsequent studies.

Results

The analysis identified highly variable genes (HVGs) corresponding to the top 0.1% variance quantile across all colon cancer samples in the GSE39582 dataset. Among these, genes such as REG4, CLCA1, and PRAC1 exhibited the largest standardized variance values, suggesting potential relevance to the underlying biological heterogeneity of colorectal cancer. A comprehensive list of the HVGs, together with their variance, median expression, and interquartile range (IQR), is provided in Table 1 for reference.

Table 1. Highly variable genes (top 0.1% variance quantile) identified from the GSE39582 dataset following $\log_2$ transformation and variance-based ranking. The listed genes exhibit the highest levels of expression heterogeneity across colon cancer samples and are considered potential candidates for further functional characterization.

Gene	Variance	Median	IQR
XIST	0.38	2.12	1.17
EIF1AY	0.34	2.23	1.20
SLC26A3	0.33	3.25	0.83
UGT2B17	0.33	2.64	1.03
PRAC1	0.31	2.65	1.10
SI	0.30	2.28	0.98
REG4	0.30	3.24	0.94
REG3A	0.29	2.86	0.99
OLFM4	0.27	3.64	0.50
REG1B	0.26	2.73	0.87
CLCA4	0.26	2.48	0.89
CLCA1	0.26	3.11	0.97
REG1A	0.25	3.11	0.90
DEFA5	0.25	2.61	0.80
DUSP27	0.25	2.27	0.89
MS4A12	0.25	2.45	0.87
CEACAM7	0.25	2.98	0.80
HLA-DQB1	0.24	2.72	0.92
ZIC2	0.22	2.19	0.85
MAGEA6	0.22	2.03	0.15
CXCL11	0.21	2.34	0.75
HEPACAM2	0.21	2.73	0.82

Figure 2 shows a heatmap of the highly variable genes. In this heatmap, MAGEA6 (Gene ID: 4105) shows mostly blue or white values, with a few distinct red spots indicating increased expression in a limited number of samples. According to the NCBI Gene, MAGEA6 belongs to the MAGEA gene family located at Xq28 and can be expressed under variable transcriptional control [23].

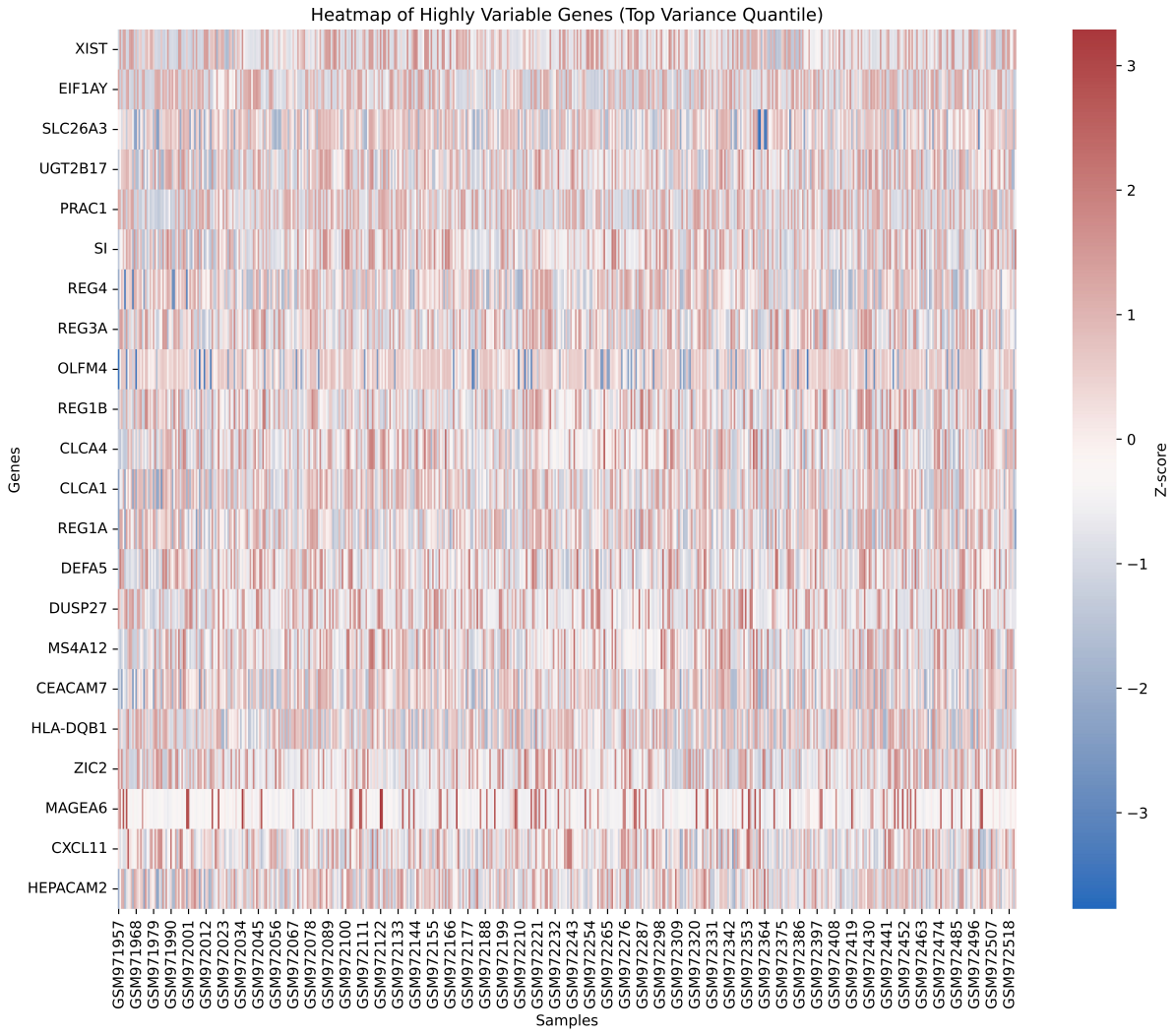

Figure 2. Heatmap showing expression patterns of highly variable genes (top 0.1% variance quantile) across all samples. The visualization reveals modular expression blocks, suggesting latent structure in the data.

The distribution of expression values for the highly variable genes was further examined using a horizontal boxplot. As shown in Figure 3, the horizontal boxplot displays the Z-scored expression distributions of the selected highly variable genes across all CRC samples. Because values were standardized by gene, most gene medians are centered near zero; however, several genes show notably wider interquartile ranges and extended whiskers, and some genes present numerous outliers. These observations indicate that a subset of genes exhibits substantially greater between-sample dispersion than the remainder. This increased dispersion could reflect true biological heterogeneity among the tumor samples, residual technical variation, or a combination of both; the plot alone does not distinguish these possibilities.

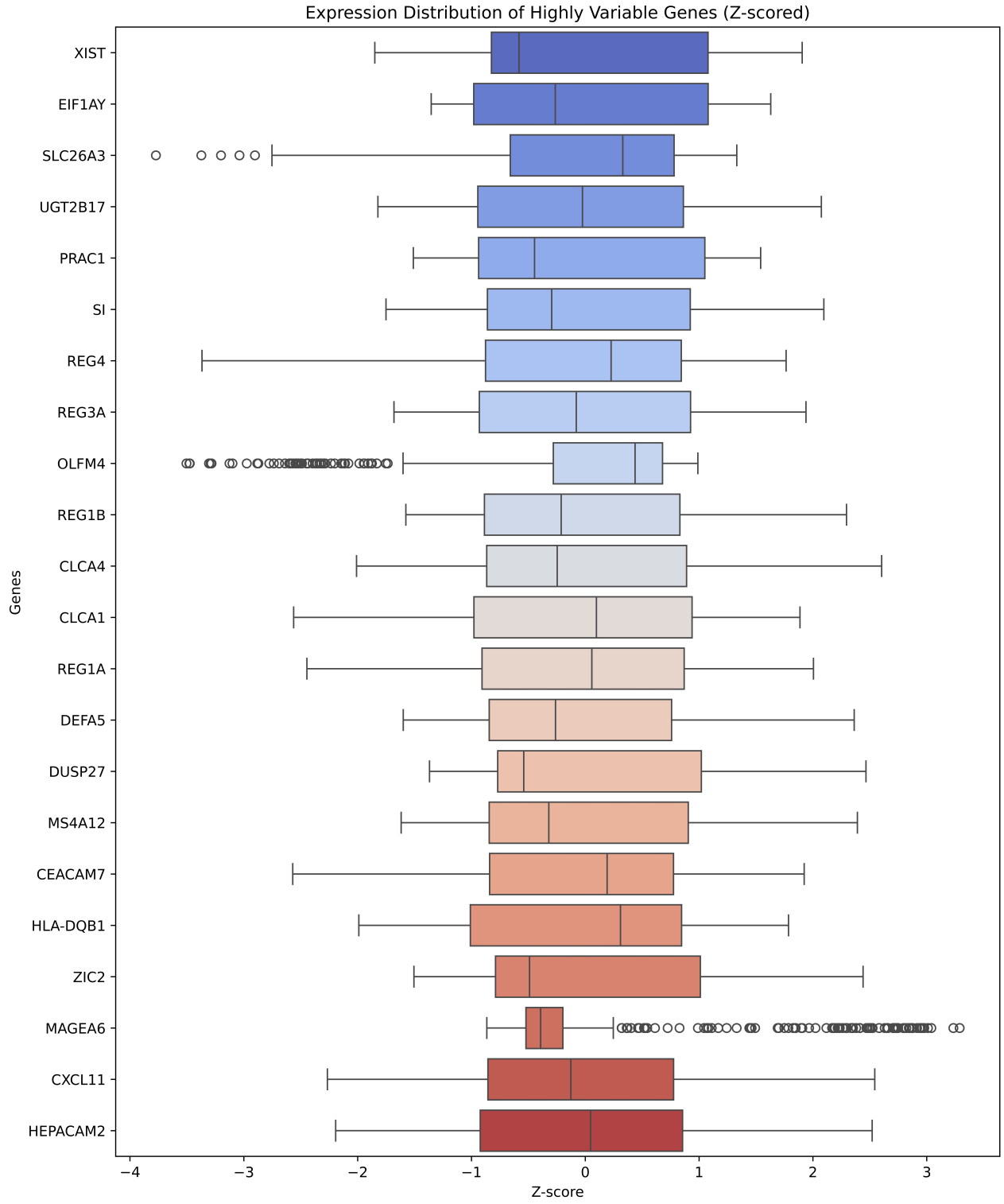

Figure 3. Boxplot of z-scored expression distribution for highly variable genes. Many features showed substantial skewness and a high number of outliers, which may correspond to conditionally expressed biomarkers or stochastic variability in transcription.

To investigate the relationship between gene expression magnitude and variability, a scatterplot of median expression versus variance was generated for all mapped genes. As depicted in Figure 4, the median versus variance scatterplot (computed on $\log_2$-transformed expression) shows that variance is not uniform across expression levels: many genes with low median expression have low variance, whereas a number of genes with intermediate

to higher median expression display elevated variance. A small fraction of genes lie well above the bulk of points. This pattern indicates which genes contribute most to sample-to-sample variability, but does not by itself identify the source or biological meaning of that variability.

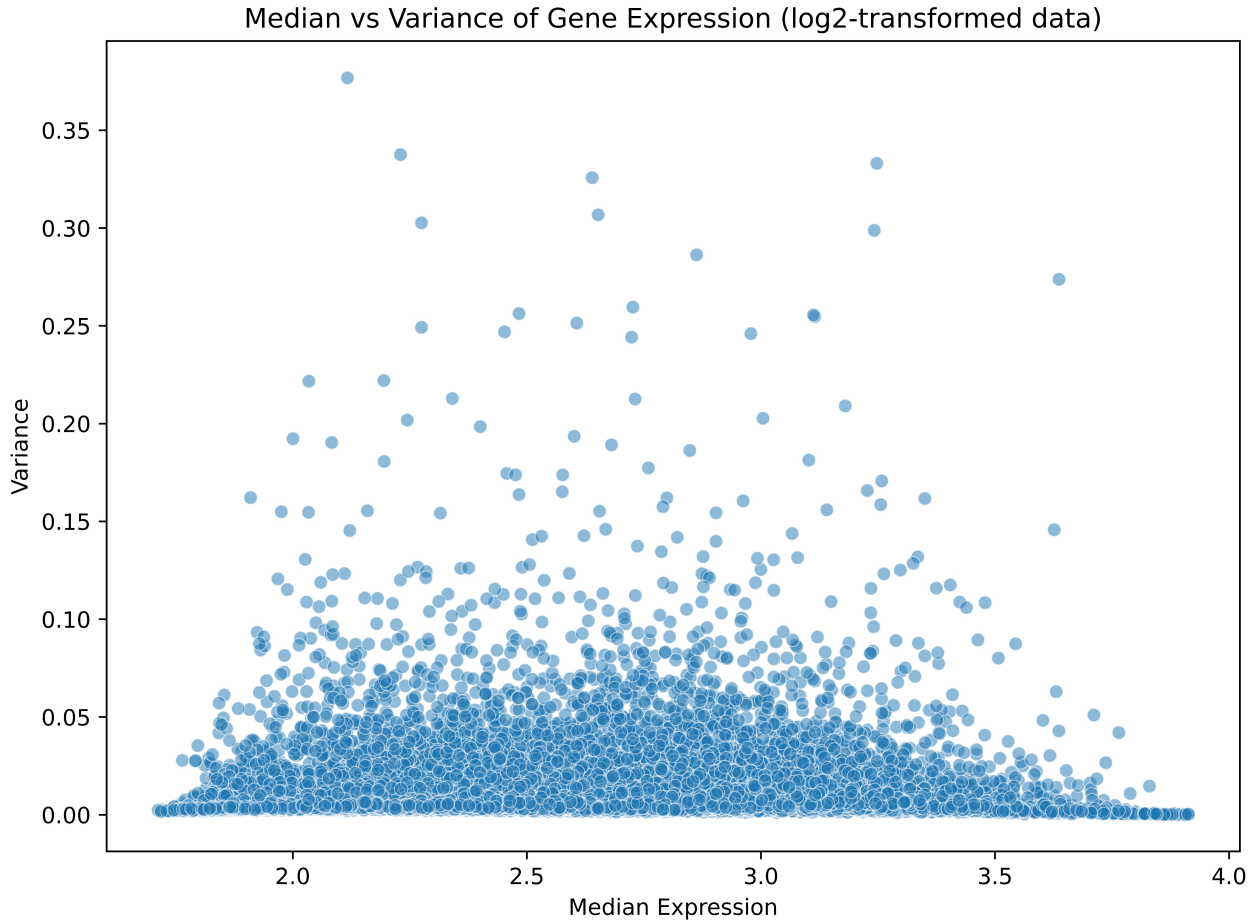

Figure 4. Scatterplot of median expression versus variance for all genes on log2-transformed data. The trend supports the notion that higher expression levels are associated with greater variance in expression across samples.

Finally, a correlation heatmap was constructed to assess co-expression among highly variable genes. As shown in Figure 5, the correlation heatmap (pairwise correlations among highly variable genes) reveals multiple blocks of strong positive correlations, interspersed with regions of weaker or negative correlations. These blocks indicate groups of genes whose expression levels co-vary across the sample set, while other genes appear to be largely independent or anticorrelated with those groups. Such co-expression structure suggests the presence of distinct expression modules within the HVG set.

Overall, the combined analyses provide a multifaceted view of gene expression variability in colorectal cancer, highlighting a subset of genes with potential biological and clinical relevance for further exploration. The findings underscore the utility of variance-guided approaches in uncovering structure and functional relationships within large-scale gene expression data, even in the absence of class labels or supervised endpoints.

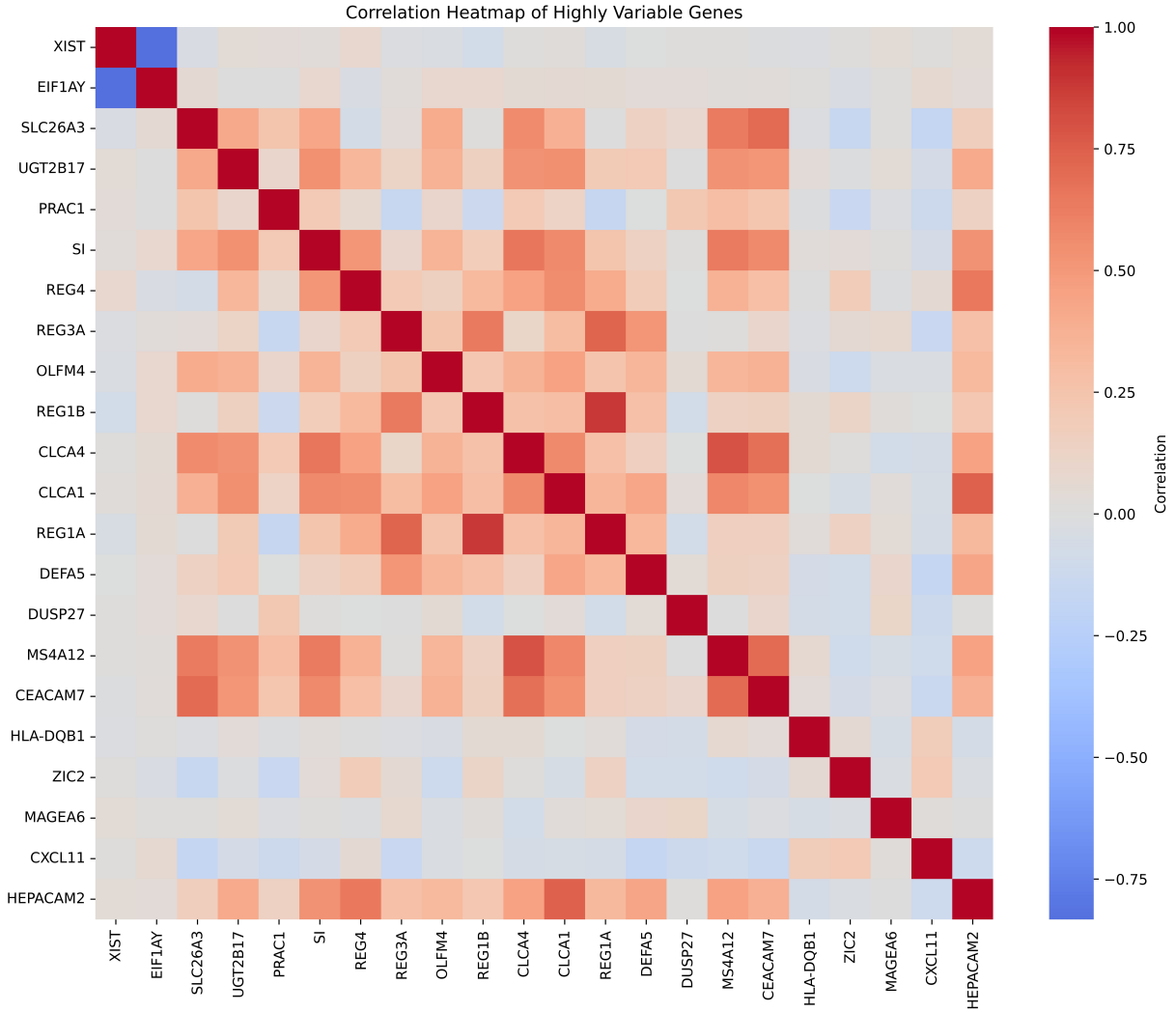

Figure 5. Correlation heatmap among highly variable genes. Warm and cool colors respectively represent positive and negative correlations, while white indicates weak or no correlation. Correlated blocks suggest transcriptional coordination, potentially driven by shared regulatory elements or epigenetic states.

Among the highly variable genes identified in the present analysis, as listed in Table 2, OLFM4, MS4A12, and CEACAM7 are notable due to their established associations with colorectal cancer (CRC) biology. The gene types are protein-coding, with the exception of XIST, which is an ncRNA.

Discussion

In the present study, a variance-based strategy was applied to identify a subset of highly variable genes (HVGs) was identified, capturing pronounced heterogeneity across patients and revealing co-expression structures indicative of transcriptional diversity. Rather than aiming to establish novel molecular subtypes or prognostic biomarkers, the primary objective of this analysis was to illustrate how a variance-guided pipeline can facilitate hypothesis generation and provide an accessible approach for early-stage exploration of gene expression variability, with the broader implications interpreted in the context of colorectal cancer biology.

In the following sections, we review the identified genes based on prior studies that have investigated their functional roles and clinical relevance in CRC.

Table 2. Functional annotation of representative highly variable genes (HVGs) identified in the present study. The information compiled in this table was obtained from the NCBI Gene database (https://www.ncbi.nlm.nih.gov/gene, accessed October 2025) [23].

Gene name	Symbol - Type	RefSeq summary
X inactive specific transcript	XIST	Encodes a non-coding RNA that initiates X-chromosome inactivation and is expressed only from the inactive X.
eukaryotic translation initiation factor 1A Y-linked	EIF1AY	Located on Y chromosome, encodes a protein related to EIF1A, stabilizes initiator Met-tRNA binding.
solute carrier family 26 member 3	SLC26A3	Encodes a chloride/bicarbonate exchanger; localized to the lower intestinal tract mucosa; mutations cause congenital chloride diarrhea.
UDP glucuronosyltransferase family 2 member B17	UGT2B17	Enzyme catalyzes glucuronidation of steroids and drugs; CNV associated with osteoporosis susceptibility.
PRAC1 small nuclear protein	PRAC1	Specifically expressed in prostate, rectum and distal colon; may play a nuclear regulatory role.
sucrase-isomaltase	SI	Encodes sucrase-isomaltase enzyme, essential for carbohydrate digestion; mutations cause congenital deficiency.
regenerating family member 4	REG4	Enables heparin and mannan binding; acts upstream in bacterial response; located in cytoplasm.
regenerating family member 3 alpha	REG3A	Encodes a pancreatic secretory protein; involved in proliferation and antimicrobial function; gene expression upregulated in pancreatic inflammation and liver carcinogenesis.
olfactomedin 4	OLFM4	This gene was originally cloned from human myeloblasts and found to be selectively expressed in inflammed colonic epithelium. This gene encodes an antiapoptotic factor that promotes tumor growth.
regenerating family member 1 beta	REG1B	Encodes secreted protein; related to islet regeneration and lithogenesis; clustered with other REG genes.
chloride channel accessory 4	CLCA4	Calcium-sensitive chloride conductance protein; spliced variants, one protein-coding.
chloride channel accessory 1	CLCA1	Encodes calcium-activated chloride channel protein; expressed in intestine; cleaved into 2 subunits.
regenerating family member 1 alpha	REG1A	Secreted protein involved in islet regeneration and pancreatic lithogenesis; clustered with REG family genes.
defensin alpha 5	DEFA5	Encodes antimicrobial peptide; highly expressed in Paneth cells; part of defensin family clustered on chr8.
serine/threonine/tyrosine interacting like 2	STYXL2 (DUSP27)	Predicted MAPK phosphatase; involved in negative regulation of MAPK cascade; cytoplasmic localization.
membrane spanning 4-domains A12	MS4A12	Silencing of this gene in colon cancer cells has been shown to suppress proliferation, cell motility, and chemotactic invasion. The gene belongs to a cluster of related genes located on chromosome 11. In addition, two transcript variants that encode distinct isoforms have been identified.
CEA cell adhesion molecule 7	CEACAM7	Expression of this gene may be downregulated in colon and rectal cancer, and lower expression can be predictive of rectal cancer recurrence.
major histocompatibility complex, class II, DQ beta 1	HLA-DQB1	Encodes HLA class II beta chain; central in immune response; polymorphisms important in transplantation.
Zic family zinc finger 2	ZIC2	Encodes transcriptional repressor; mutations cause holoprosencephaly type 5; polymorphisms linked to NTD risk.
MAGE family member A6	MAGEA6	Member of MAGEA family clustered on Xq28; alternative splicing; implicated in hereditary disorders.
C-X-C motif chemokine ligand 11	CXCL11	Encodes CXC chemokine; induces T-cell chemotaxis; IFN-γ strongly induces expression.
HEPACAM family member 2	HEPACAM2	Encodes immunoglobulin superfamily protein involved in mitosis; chromosomal deletion including this gene may be associated with myeloid leukemia and myelodysplastic syndrome.

Olfactomedin 4

Our analysis highlighted olfactomedin 4 (OLFM4) as one of the most variably expressed genes in colorectal cancer (CRC), consistent with its previously reported role in tumor progression and metastasis. Notably, Huang et al. (2012) demonstrated that OLFM4 overexpression was significantly correlated with liver metastasis in Taiwanese CRC patients, with an odds ratio of 3.438 [24]. Their study employed a weighted enzymatic chip array platform and validated findings through reverse transcriptase–PCR, underscoring the robustness of OLFM4 as a biomarker for metastatic potential. Importantly, the work by Huang and colleagues not only confirmed OLFM4 upregulation in primary CRC tissues compared with adjacent normal mucosa but also reported its association with circulating tumor cells and suggested its potential prognostic relevance [24]. The evidence lends further support to our identification of OLFM4 among the top highly variable genes, suggesting that aberrant regulation of this gene may play a critical role in promoting tumor cell dissemination and metastatic behavior in CRC [24]. The convergence of our findings with the work of Huang et al. strengthens the hypothesis that OLFM4 can serve as a clinically relevant marker for disease aggressiveness and progression.

In addition to the above clinical associations, earlier work has further explored OLFM4's biological roles in colorectal cancer, particularly its link to stem cell identity and tumor initiation.

Supporting the interpretation of OLFM4 as a functionally relevant biomarker in colorectal cancer, the seminal work by van der Flier et al. (2009) established OLFM4 as a robust marker of intestinal stem cells and a subset of colorectal cancer cells [25]. Their study demonstrated that OLFM4 is highly expressed in crypt base columnar cells in the human small intestine and colon, where it co-localizes with Lgr5-positive stem cells, and is markedly upregulated in subsets of colorectal adenocarcinoma cells [25]. Importantly, OLFM4 expression in malignant tissues was substantially higher than that observed in adjacent normal crypt base cells, highlighting its potential role in marking tumor-initiating or cancer stem cell populations [25]. By integrating molecular signatures of Wnt signaling and stemness, the authors proposed that OLFM4 serves not only as a faithful stem cell marker but also as a putative identifier of cells with tumor-initiating capacity in human colorectal cancer [25]. In light of the findings, the presence of OLFM4 among the top highly variable genes in our dataset strongly suggests that its deregulation reflects alterations in stem cell biology and tumor initiation pathways [25]. Together, our results and the work of van der Flier and colleagues reinforce the concept that OLFM4 expression bridges the biology of normal intestinal stem cells with the pathogenesis of colorectal carcinoma.

Although van der Flier et al. underscored OLFM4's role in stemness and tumor initiation, other investigations have revealed stage-specific dynamics that further refine its clinical interpretation.

Adding further weight to the role of OLFM4 in colorectal cancer, Besson et al. (2011) provided a comprehensive proteomic analysis across distinct stages of colorectal tumor progression, identifying OLFM4 as a nonmetastatic tumor marker [26]. Using laser capture microdissection and quantitative proteomic approaches, the authors established that OLFM4 expression is significantly upregulated in adenomas and early-stage colorectal cancers (stages I–II), but markedly reduced in advanced, metastatic stages (III–IV) [26]. Immunohistochemistry performed on 126 patient tissues confirmed this stage-specific regulation, revealing pronounced cytoplasmic and nuclear OLFM4 expression during dysplasia and noninvasive stages, in contrast to diminished levels in late-stage disease [26]. Mechanistically, their data also implicated the Ras–NF-κB2 signaling axis as a regulator of OLFM4 expression, with enhanced expression observed in Ras-mutated tumors [26]. Importantly, OLFM4 was shown to be secreted in a glycosylated form, further underscoring its translational potential as a biomarker for early detection and patient stratification [26]. Together, the findings support the interpretation that OLFM4 acts as a critical regulator during the early, nonmetastatic phases of colorectal cancer, and its subsequent downregulation may be a prerequisite for tumor progression and metastasis [26]. In the context of our results, the consistency between transcriptomic and proteomic evidence underscores OLFM4's robustness as a biologically relevant marker of tumor initiation and early-stage progression in colorectal cancer.

Extending this stage-specific view, organoid-based models have provided direct evidence of how OLFM4 expression shifts during the transition from primary to metastatic CRC.

Further support for OLFM4 as a key regulator of colorectal cancer biology comes from the study of Okamoto et al. (2021), which applied a comparative analysis of patient-matched primary and metastatic colorectal cancer organoids [27]. Their integrative transcriptomic and single-cell RNA-sequencing approach revealed a striking reduction in OLFM4-associated stem-like clusters in metastatic and recurrent lesions relative to their

corresponding primary tumors [27]. OLFM4 was not only the most differentially expressed gene distinguishing primary from metastatic PDOs but was also shown to define a stem-like subpopulation (cluster C1) with organoid-initiating and differentiation capacity [27]. Functional validation confirmed that OLFM4+ cells were indispensable for efficient reconstitution of primary PDOs, whereas metastatic PDOs could sustain growth independent of OLFM4+ cells, indicating a fundamental shift in cellular requirements during disease progression [27]. Immunohistochemical analyses further validated that OLFM4+ cells were enriched in primary tumor specimens but markedly diminished in metastatic lesions, mirroring the organoid data [27].

The findings highlight two key dimensions of OLFM4's role in colorectal cancer: first, its association with stem-like cellular subpopulations that underlie tumor initiation and structural hierarchy; and second, its progressive loss in metastatic disease, consistent with a phenotypic transition toward less differentiated, proliferative states. Together with other proteomic and transcriptomic studies, the data from Okamoto et al. reinforce OLFM4 as a robust, stage-dependent biomarker in colorectal cancer, underscoring both its utility for early tumor characterization and its relevance in understanding the cellular remodeling events that enable metastasis [27].

In parallel, molecular studies of regulatory networks have demonstrated that OLFM4 is not merely a passive marker but also an active downstream effector of oncogenic drivers.

In line with previous studies, Yan et al. (2021) provided additional mechanistic evidence underscoring the central role of OLFM4 in colorectal cancer progression [28]. Their study identified the super-enhancer-associated long noncoding RNA (SE-lncRNA) AC005592.2 as a novel oncogenic driver in CRC, demonstrating that its overexpression promoted proliferation, invasion, and migration of colorectal cancer cells while inhibiting apoptosis [28]. Importantly, through integrative RNA-seq and functional validation, OLFM4 emerged as a major putative downstream effector of AC005592.2, with both transcriptional and protein-level upregulation confirmed in patient samples and CRC cell lines [28].

The findings reinforce the notion that OLFM4 is not a passive biomarker but rather a transcriptionally regulated mediator of tumor biology, operating within complex regulatory networks such as SE-lncRNA–driven oncogenic circuits [28]. The observation that OLFM4 expression correlated with clinical parameters including TNM stage and tumor differentiation further strengthens its clinical relevance [28]. Notably, the study also cited previous research highlighting the context-dependent role of OLFM4, wherein it may be highly expressed in early or well-differentiated CRC but progressively downregulated in advanced or metastatic disease [28].

Together with evidence from clinical analyses reported in the study, the results of Yan et al. underscore the multifaceted nature of OLFM4 regulation in CRC and provide an additional layer of support for its consideration as both a diagnostic biomarker and a potential therapeutic target [28]. Specifically, its integration into super-enhancer–lncRNA–associated regulatory networks highlights new opportunities for therapeutic intervention targeting upstream regulatory elements, thereby broadening the translational significance of OLFM4 in colorectal cancer [28].

Beyond tumor-intrinsic signaling, immune-related research has further uncovered a role for OLFM4 in shaping the tumor microenvironment and immune evasion.

The present findings on the role of OLFM4 in disease progression are further substantiated by evidence from Chen et al. (2022), who demonstrated that OLFM4 expression is markedly elevated in polymorphonuclear myeloid-derived suppressor cells (PMN-MDSCs) during the transition from colitis to colorectal cancer [29]. Their study revealed that OLFM4 positively correlates with disease severity and promotes PMN-MDSC recruitment through the NF-κB/PTGS2 signaling axis, thereby contributing to the establishment of an immunosuppressive tumor microenvironment [29]. Importantly, genetic ablation of OLFM4 in myeloid cells reduced PMN-MDSC infiltration, delayed colitis-associated tumorigenesis, and enhanced responsiveness to anti-PD-1 immunotherapy, underscoring its pivotal role in modulating immune escape mechanisms in colorectal cancer [29]. The observations align with our data, reinforcing OLFM4 as a key mediator linking chronic inflammation to malignant transformation, and highlighting its potential as both a biomarker of disease progression and a therapeutic target in inflammation-driven colorectal carcinogenesis (Chen et al., 2022).

Taken together, the evidence consistently positions OLFM4 as a multifaceted regulator of CRC biology, with roles spanning stemness, inflammation, and metastasis. To broaden the discussion of highly variable genes identified in our dataset, we next turn to MS4A12, another gene with established functional links to colorectal cancer.

Membrane Spanning 4-Domains A12

Our identification of membrane spanning 4-domains A12 (MS4A12) among the most variable genes in the GSE39582 cohort is consistent with earlier mechanistic studies that have highlighted its role as a colon cancer–associated gene [30]. In particular, Koslowski and colleagues (2009) demonstrated that MS4A12 expression is strictly confined to colonic epithelial cells and is frequently maintained during malignant transformation of colon tissue [30]. They further elucidated that the transcription factor CDX2 is an essential activator of MS4A12, thereby providing a molecular rationale for its tissue-specific expression pattern and oncogenic activity [30]. Functional assays in colon cancer cell lines revealed that MS4A12 acts as a component of store-operated calcium entry, sensitizing tumor cells to epidermal growth factor (EGF) signaling and thereby promoting proliferation, motility, and chemotactic invasion [30]. Such observations not only support the biological plausibility of MS4A12 variability in colorectal cancer but also underscore its potential contribution to tumor progression through growth factor–dependent pathways [30]. By aligning with the prior findings, our results reinforce the view that MS4A12 variability is not a stochastic artifact of transcriptomic profiling but rather a reflection of its central role in the biology of colon carcinogenesis.

Complementing the above observations, earlier work from the same group had already characterized MS4A12 as a colon-selective differentiation gene with oncogenic potential.

The functional significance of MS4A12 in colorectal cancer is further reinforced by the seminal work of Koslowski et al. (2008), who first described MS4A12 as a colon-selective store-operated calcium channel with strong potential as a therapeutic target [31]. Their study established that MS4A12 is a surface-expressed differentiation gene restricted to colonocytes, and importantly, its expression persists in a substantial proportion of colon cancers (approximately 63%) [31]. Functional assays demonstrated that MS4A12 directly contributes to EGFR-mediated calcium signaling, thereby enhancing proliferative capacity, motility, and invasive potential of colon cancer cells [31]. Silencing of MS4A12 markedly impaired the processes, supporting its role as an essential mediator of tumor-promoting pathways [31]. Moreover, the structural similarity of MS4A12 to CD20—a clinically validated target in hematological malignancies—underscores its potential druggability for antibody-based therapies in colon cancer [31]. Taken together, the findings indicate that variability in MS4A12 expression observed in our dataset does not merely represent stochastic noise, but rather reflects the involvement of a tissue-specific oncogenic driver with both mechanistic relevance and therapeutic promise.

While the above mechanistic studies provided molecular insights, large-scale bioinformatics and clinical analyses have subsequently confirmed the prognostic implications of MS4A12 variability.

Further evidence supporting the clinical relevance of MS4A12 in colorectal cancer comes from the study by Han et al. (2020), who used large-scale bioinformatics and clinical validation to identify MS4A12 as a hub gene in primary colorectal cancer (PCRC) [32]. Their integrated analysis of multiple GEO datasets revealed MS4A12 as one of ten significantly dysregulated genes, with consistently downregulated expression in tumor tissues compared to normal colorectal epithelium [32]. Importantly, survival analysis demonstrated that patients with low expression of MS4A12 exhibited significantly poorer overall survival, thereby positioning MS4A12 not only as a diagnostic marker but also as a prognostic biomarker in PCRC [32]. Validation by RT-qPCR in a cohort of 192 patients confirmed these findings, showing reduced MS4A12 expression in tumors relative to adjacent normal tissues [32]. The study further emphasized the potential of MS4A12, alongside CLCA4, as a candidate for personalized medicine approaches, suggesting that its expression profile could be leveraged for prognosis, therapeutic stratification, and even targeted intervention [32]. The observations directly complement our findings by demonstrating that MS4A12 variability in colorectal tumors has tangible clinical consequences, linking molecular heterogeneity to patient outcome [32].

Further mechanistic studies have refined the picture by linking MS4A12 expression to tumor cell differentiation and patient outcome, thereby underscoring its context-dependent roles in CRC.

The role of MS4A12 in colorectal cancer is further substantiated by the study of He et al. (2017), who investigated its involvement in tumor cell differentiation and patient prognosis [33]. Their experimental data showed that MS4A12 expression increases during sodium butyrate–induced differentiation of colon cancer cells, and silencing of MS4A12 variant-1 significantly inhibited differentiation markers such as alkaline phosphatase (ALP) and E-cadherin [33]. Furthermore, suppression of MS4A12 attenuated the ability of butyrate to induce cell cycle arrest and apoptosis, indicating that MS4A12 is a functional mediator of differentiation-linked tumor

suppression pathways [33]. Clinically, analysis of large-scale datasets (GSE39582 and GSE38832) revealed that low MS4A12 expression correlated with worse survival in early-stage colon cancer, although its prognostic value diminished in advanced-stage disease [33]. These findings emphasize a dual role of MS4A12: on the one hand, as a differentiation-related regulator whose reduced expression promotes aggressive tumor phenotypes; and on the other, as a stage-specific prognostic biomarker capable of stratifying risk in early colon cancer patients [33]. Taken together, the evidence complements our findings by demonstrating that variability in MS4A12 expression reflects not only molecular heterogeneity, but also clinically relevant differences in tumor differentiation and survival outcomes [33].

Whereas MS4A12 highlights oncogenic mechanisms tied to calcium signaling and differentiation, CEACAM7 provides a contrasting example of a gene whose downregulation appears to drive loss of adhesion and early tumorigenesis in CRC.

CEA Cell Adhesion Molecule 7

Our findings regarding the downregulation of CEA cell adhesion molecule 7 (CEACAM7) in colorectal cancer are consistent with prior evidence highlighting its prognostic relevance in disease progression [34]. Messick et al. (2010) demonstrated that CEACAM7 expression is markedly reduced in rectal cancer tissues compared to normal mucosa, and more importantly, its loss was significantly associated with disease recurrence, particularly in stage II patients [34]. Their data revealed that decreased CEACAM7 expression served as an independent prognostic marker for recurrence-free survival, underscoring its potential clinical utility in stratifying patients for adjuvant therapy [34]. The observation aligns with our results, reinforcing the notion that CEACAM7 downregulation represents an early molecular event in colorectal tumorigenesis and may contribute to unfavorable outcomes by impairing cellular differentiation and adhesion [34]. Taken together, the findings strengthen the argument that CEACAM7 is not only a diagnostic marker of malignant transformation but also a prognostic biomarker with clinical implications for treatment decision-making [34].

Beyond the above prognostic associations, structural studies have offered a deeper understanding of how CEACAM7 contributes to epithelial integrity.

In addition to clinical evidence supporting CEACAM7 downregulation in colorectal cancer, structural studies have provided insights into its molecular properties. Bonsor et al. (2015) resolved the crystal structure of the N-terminal dimerization domain of CEACAM7 and showed that the protein forms a homodimer with approximately tenfold higher affinity compared with CEACAM5 [35]. This increased affinity was attributed to structural differences, including a buckled C" strand that creates an additional hydrogen bond in the dimer interface [35]. Consistent with previous reports, the authors noted that CEACAM7 is expressed on differentiated epithelial cells of the colon and rectum and is downregulated in colorectal cancer [35]. Although the study primarily focused on structural characterization rather than functional analysis, these findings support the idea that CEACAM7's structural stability and dimerization properties could be relevant to its role in epithelial cell adhesion.

Consistent with the structural and functional insights, earlier studies have also documented the progressive loss of CEACAM7 expression during the adenoma–carcinoma sequence in colorectal tumorigenesis.

The current findings on CEACAM7 downregulation in colorectal neoplasia are consistent with earlier reports that have demonstrated a similar loss of expression in early stages of tumorigenesis. Thompson and colleagues (1997) provided compelling evidence that CEACAM7, referred to as CGM2 in their study, is strongly expressed in differentiated colonocytes located in the upper third of normal colorectal crypts but is markedly reduced in adenomas and completely absent in advanced carcinomas and metastases [36]. Using in situ hybridization and monoclonal antibody approaches, they confirmed that CEACAM7 expression is restricted to mature epithelial cells and is lost during the adenoma–carcinoma sequence [36]. Importantly, the study emphasized that CEACAM7 downregulation represents an early molecular event in colorectal tumorigenesis, distinguishing it from other CEA family members such as CEACAM5 and CEACAM6, which are maintained or even upregulated in malignant tissue [36].

By corroborating our findings with the work of Thompson et al., it becomes evident that CEACAM7 plays a critical role as a putative tumor suppressor within the CEA family [36]. Its selective loss in precancerous lesions suggests that CEACAM7 contributes to the maintenance of normal epithelial differentiation and that its absence may facilitate the transition to malignant phenotypes [36]. The alignment of the independent observations

reinforces the robustness of CEACAM7 as a biomarker for early colorectal neoplasia and highlights its potential utility in distinguishing normal mucosa from preneoplastic and neoplastic states [36].

Despite providing valuable insights into gene expression variability in colorectal cancer, several limitations should be acknowledged. First, the analysis relied exclusively on a single microarray dataset (GSE39582), which may limit the generalizability of the findings across other cohorts or platforms. Second, the study focused on the top 0.1% highly variable genes, potentially overlooking biologically relevant genes with moderate variability.

Overall, our findings demonstrate that variance-based transcriptomic exploration can highlight biologically relevant genes and co-expression structures in colorectal cancer, even without prior assumptions about subtype classification or clinical outcomes. The consistency of highly variable genes such as OLFM4, MS4A12, and CEACAM7 with previous reports underscores the potential of this approach to recover meaningful biological signals from public datasets. At the same time, the analysis presented here should be regarded as a starting point for generating hypotheses rather than a definitive framework for biomarker discovery. Future investigations should aim to validate the findings and explore how they may be applied in clinical contexts.

Conclusions

The pipeline is particularly valuable in pedagogical contexts, where it can introduce students to transcriptomic data analysis with minimal technical overhead. It also provides a practical tool for rapid preliminary screening of newly available datasets, enabling researchers to identify patterns of variability before investing in more resource-intensive analyses. Finally, its simplicity makes it accessible to investigators without extensive bioinformatics expertise, offering an approachable entry point into transcriptomic data exploration.

In addition to its practical advantages, the results obtained through the pipeline highlight biological patterns that merit further examination.

The findings offer a basis for exploring gene modules, regulatory influences, or expression-based patient stratification in CRC. Genes of such type may act as putative biomarkers or represent transcriptionally co-regulated modules with biological or clinical relevance.

Variance-guided exploration of transcriptomic data represents a useful preliminary step in uncovering functional diversity within tumor populations. Identification of genes with exceptionally high variability can support hypothesis generation regarding regulatory mechanisms, cellular heterogeneity, and tumor evolution. Integration of variance-based metrics with additional layers of molecular data, such as epigenetic or proteomic profiles, may further clarify the biological processes driving gene expression dispersion in colorectal cancer.

Overall, the approach highlights the interpretive value of transcriptomic variability and underscores the potential of quantitative variance analysis as a foundation for future investigations aimed at understanding intra-tumor heterogeneity and identifying functionally coordinated gene networks in colorectal malignancy.

Author Contributions: A.M.M.: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review and editing.

Funding: This research received no funding.

Ethics Statement: Not applicable.

Data Availability Statement: The GSE39582 dataset is available at: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE39582. The GPL570 platform is available at: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL570

Conflict of Interest: The author declares no conflict of interest.

References

1. Gatlin V, Gupta S, Romero S, Chapkin RS, Cai JJ. Exploring cell-to-cell variability and functional insights through differentially variable gene analysis. npj Systems Biology and Applications. 2025;11(1):29. doi:10.1038/s41540-025-00507-z

2. Árnadóttir SS, Mattesen TB, Vang S, Madsen MR, Madsen AH, Birkbak NJ, et al. Transcriptomic and proteomic intra-tumor heterogeneity of colorectal cancer varies depending on tumor location within the colorectum. PLoS One. 2020;15(12):e0241148. doi:10.1371/journal.pone.0241148
3. Marisa L, de Reyniès A, Duval A, Selves J, Gaub MP, Vescovo L, et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. PLoS Med. 2013;10(5):e1001453. doi:10.1371/journal.pmed.1001453
4. Mo X, Su Z, Yang B, Zeng Z, Lei S, Qiao H. Identification of key genes involved in the development and progression of early-onset colorectal cancer by co-expression network analysis. Oncol Lett. 2020;19(1):177-86. doi:10.3892/ol.2019.11073
5. Langerud J, Eilertsen IA, Moosavi SH, Klokkerud SMK, Reims HM, Backe IF, et al. Multiregional transcriptomics identifies congruent consensus subtypes with prognostic value beyond tumor heterogeneity of colorectal cancer. Nat Commun. 2024;15(1):4342. doi:10.1038/s41467-024-48706-2
6. Kamal Y, Dwan D, Hoehn HJ, Sanz-Pamplona R, Alonso MH, Moreno V, et al. Tumor immune infiltration estimated from gene expression profiles predicts colorectal cancer relapse. Oncoimmunology. 2021;10(1):1862529. doi:10.1080/2162402x.2020.1862529
7. Angius A, Scanu AM, Arru C, Muroni MR, Carru C, Porcu A, et al. A Portrait of Intratumoral Genomic and Transcriptomic Heterogeneity at Single-Cell Level in Colorectal Cancer. Medicina (Kaunas). 2021;57(11). doi:10.3390/medicina57111257
8. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012;487(7407):330-7. doi:10.1038/nature11252
9. van Dijk D, Sharma R, Nainys J, Yim K, Kathail P, Carr AJ, et al. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. Cell. 2018;174(3):716-29.e27. doi:10.1016/j.cell.2018.05.061
10. Tirosh I, Venteicher AS, Hebert C, Escalante LE, Patel AP, Yizhak K, et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. Nature. 2016;539(7628):309-13. doi:10.1038/nature20123
11. Dunne PD, Alderdice M, O'Reilly PG, Roddy AC, McCorry AMB, Richman S, et al. Cancer-cell intrinsic gene expression signatures overcome intratumoural heterogeneity bias in colorectal cancer patient classification. Nat Commun. 2017;8:15657. doi:10.1038/ncomms15657
12. Isella C, Brundu F, Bellomo SE, Galimi F, Zanella E, Porporato R, et al. Selective analysis of cancer-cell intrinsic transcriptional traits defines novel clinically relevant subtypes of colorectal cancer. Nat Commun. 2017;8:15107. doi:10.1038/ncomms15107
13. Joanito I, Wirapati P, Zhao N, Nawaz Z, Yeo G, Lee F, et al. Single-cell and bulk transcriptome sequencing identifies two epithelial tumor cell states and refines the consensus molecular classification of colorectal cancer. Nat Genet. 2022;54(7):963-75. doi:10.1038/s41588-022-01100-4
14. Ten Hoorn S, de Back TR, Sommeijer DW, Vermeulen L. Clinical Value of Consensus Molecular Subtypes in Colorectal Cancer: A Systematic Review and Meta-Analysis. J Natl Cancer Inst. 2022;114(4):503-16. doi:10.1093/jnci/djab106
15. Kim Z, Lee J, Yoon YE, Yun JW. Unveiling Prognostic RNA Biomarkers through a Multi-Cohort Study in Colorectal Cancer. International Journal of Molecular Sciences. 2024;25(6):3317. doi:10.3390/ijms25063317
16. The pandas development t. pandas-dev/pandas: Pandas. 2020;latest. doi:10.5281/zenodo.3509134
17. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature. 2020;585(7825):357-62. doi:10.1038/s41586-020-2649-2
18. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825-30.
19. Hunter JD. Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering. 2007;9(3):90-5. doi:10.1109/MCSE.2007.55
20. Waskom ML. seaborn: statistical data visualization. Journal of Open Source Software. 2021;6(60):3021. doi:10.21105/joss.03021
21. NCBI Gene Expression Omnibus (GEO), https://www.ncbi.nlm.nih.gov/geo/. 2025.
22. Affymetrix I. [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array, Platform GPL570. 2003.
23. National Center for Biotechnology I. Gene, https://www.ncbi.nlm.nih.gov/gene/. 2025.

24. Huang MY, Wang HM, Chang HJ, Hsiao CP, Wang JY, Lin SR. Overexpression of S100B, TM4SF4, and OLFM4 Genes Is Correlated with Liver Metastasis in Taiwanese Colorectal Cancer Patients. DNA and Cell Biology. 2012;31(1):43-9. doi:10.1089/dna.2011.1264
25. Van der Flier LG, Haegebarth A, Stange DE, Van de Wetering M, Clevers H. OLFM4 Is a Robust Marker for Stem Cells in Human Intestine and Marks a Subset of Colorectal Cancer Cells. Gastroenterology. 2009;137(1):15-7. doi:10.1053/j.gastro.2009.05.035
26. Besson D, Pavageau AH, Valo I, Bourreau A, Bélanger A, Eymerit-Morin C, et al. A Quantitative Proteomic Approach of the Different Stages of Colorectal Cancer Establishes OLFM4 as a New Nonmetastatic Tumor Marker. Molecular & Cellular Proteomics. 2011;10(12). doi:10.1074/mcp.M111.009712
27. Okamoto T, DuVerle D, Yaginuma K, Natsume Y, Yamanaka H, Kusama D, et al. Comparative Analysis of Patient-Matched PDOs Revealed a Reduction in OLFM4-Associated Clusters in Metastatic Lesions in Colorectal Cancer. Stem Cell Reports. 2021;16(4):954-67. doi:10.1016/j.stemcr.2021.02.012
28. Yan LP, Chen HH, Tang L, Jiang P, Yan F. Super-enhancer-associated long noncoding RNA AC005592.2 promotes tumor progression by regulating OLFM4 in colorectal cancer. Bmc Cancer. 2021;21(1). doi:10.1186/s12885-021-07900-x
29. Chen ZY, Zhang XG, Xing Z, Lv SJ, Huang LX, Liu JP, et al. OLFM4 deficiency delays the progression of colitis to colorectal cancer by abrogating PMN-MDSCs recruitment. Oncogene. 2022;41(22):3131-50. doi:10.1038/s41388-022-02324-8
30. Koslowski M, Türeci Ö, Huber C, Sahin U. Selective activation of tumor growth-promoting Ca2+ channel MS4A12 in colon cancer by caudal type homeobox transcription factor CDX2. Molecular Cancer. 2009;8. doi:10.1186/1476-4598-8-77
31. Koslowski M, Sabin U, Dhaene K, Huber C, Türeci Ö. MS4A12 is a colon-selective store-operated calcium channel promoting malignant cell processes. Cancer Research. 2008;68(9):3458-66. doi:10.1158/0008-5472.Can-07-5768
32. Han J, Zhang X, Liu Y, Jing L, Liu YB, Feng L. CLCA4 and MS4A12 as the significant gene biomarkers of primary colorectal cancer. Bioscience Reports. 2020;40. doi:10.1042/bsr20200963
33. He L, Deng HY, Wang XC. Decreased expression of MS4A12 inhibits differentiation and predicts early stage survival in colon cancer. Neoplasma. 2017;64(1):65-73. doi:10.4149/neo_2017_108
34. Messick CA, Sanchez J, DeJulius KL, Hammel J, Ishwaran H, Kalady MF. CEACAM-7: A predictive marker for rectal cancer recurrence. Surgery. 2010;147(5):713-9. doi:10.1016/j.surg.2009.10.056
35. Bonsor DA, Beckett D, Sundberg EJ. Structure of the N-terminal dimerization domain of CEACAM7. Acta Crystallographica Section F-Structural Biology Communications. 2015;71:1169-75. doi:10.1107/s2053230x15013576
36. Thompson J, Seitz M, Chastre E, Ditter M, Aldrian C, Gespach C, et al. Down-Regulation of Carcinoembryonic Antigen Family Member 2 Expression Is an Early Event in Colorectal Tumorigenesis1,2. Cancer Research. 1997;57(9):1776-84.