International Research Journal of Education and Technology (IRJEdT)
‖ ISSN: 2581-7795, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 09 | Issue: 09 | Sep- 2026 | www.irjweb.com ‖

CROSS-DATASET IDENTIFICATION AND INDEPENDENT VALIDATION OF CONSERVED GENE SIGNATURES ASSOCIATED WITH COLERECTAL CANCER PROGRESSION

Ms.Pooja S

M.Sc. Bioinformatics, Department of Bioscience , Sri Krishna Arts and Science College, Coimbatore, Tamil Nadu, India [1]

Dr B. Aarthi Rashmi,School for Life Sciences,Sri Krishna University, Coimbatore, Tamil Nadu, India [1]

Dr C. Nilavamuthan ,School for Life Sciences,Sri Krishna University, Coimbatore, Tamil Nadu, India [1]

----------***----------

Abstract:*Colorectal cancer (CRC) is among the most frequently diagnosed malignancies worldwide, and its progression is driven by a complex, only partially characterized transcriptional program. Two publicly available Affymetrix microarray datasets from the Gene Expression Omnibus (GEO) were used as discovery (GSE22598; 17 paired tumour/normal colorectal tissues) and validation (GSE23878; 35 tumour and 24 non-cancerous colorectal tissues) cohorts. Differentially expressed genes (DEGs) were identified independently in each dataset using the limma linear-modelling framework (|log2 fold-change| > 1, adjusted P < 0.05, Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses were performed on the conserved DEGs, and a protein–protein interaction (PPI) network was constructed in STRING and visualised and topologically analysed in Cytoscape. The top-20 conserved candidate genes,, showed highly concordant fold-change direction and magnitude between cohorts (combined score range 20.06–29.86) and cleanly separated tumour from normal samples by unsupervised hierarchical clustering. Functional enrichment highlighted epithelial and skin-development programmes, xenobiotic transport, cell-cell junction assembly and cellular response to hypoxia (GO), together with cell-adhesion-molecule interaction and cadherin-signalling pathways (KEGG). PPI network analysis of the candidate genes and their first-shell interactors (30 nodes, 10 edges; PPI enrichment P = 0.0443) revealed only sparse direct physical connectivity among panel members, consistent with a transcriptionally rather than physically coupled programme. This two-dataset discovery-validation design identifies a compact, cross-platform-reproducible 20-gene signature associated with CRC progression that may serve as a starting point for biomarker development and mechanistic follow-up.*

Keywords:Colorectal cancer, Differential gene expression, GEO, Cross-dataset validation, limma; Gene Ontology, KEGG, Protein–protein interaction network, Cytoscape; Biomarker signature

1.INTRODUCTION

Colorectal cancer (CRC) remains one of the most commonly diagnosed cancers and a leading cause of

International Research Journal of Education and Technology (IRJEdT)
	ISSN: 2581-7795, Impact Factor 5.007	
	Peer-Reviewed	Open Access	International Journal	
	Volume: 09	Issue: 09	Sep- 2026	www.irjweb.com	

cancer-related mortality worldwide. Its clinical course is shaped by the sequential accumulation of genetic and transcriptional alterations that drive the adenoma-carcinoma progression sequence, yet the specific transcriptional programmes that are reproducibly dysregulated across patient populations remain incompletely defined. High-throughput microarray and RNA-sequencing studies deposited in the Gene Expression Omnibus (GEO) have generated a large body of colorectal transcriptomic data; however, individual studies are frequently limited by small sample sizes, platform-specific technical variation and population-specific effects, all of which can lead to false-positive or non-reproducible differential expression findings when a single dataset is analysed in isolation. A widely adopted strategy to mitigate these limitations is a discovery-validation design, in which differentially expressed genes (DEGs) identified in one cohort are cross-checked against an independent cohort, and only genes that are concordantly altered in both are carried forward as a conserved, higher-confidence signature. This approach increases the likelihood that the resulting gene panel reflects genuine tumour biology rather than dataset-specific artefacts, and it has been successfully applied to identify hub genes and candidate biomarkers in several cancers.

In this study, we applied this cross-dataset discovery-validation strategy to two independent, publicly available colorectal cancer microarray datasets, GSE22598 and GSE23878, both generated on the Affymetrix Human Genome U133 Plus 2.0 platform. We identified DEGs independently in each dataset, defined a conserved DEG set as the intersection of the two lists, ranked the conserved genes by a combined discovery-validation score to derive a top-20 candidate gene signature, and characterised the functional and network-level properties of this signature using Gene Ontology (GO) enrichment, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis, and STRING/Cytoscape-based protein–protein interaction (PPI) network topology (degree, betweenness and closeness centrality). The overall aim was to derive a compact, reproducible gene signature associated with CRC progression that is suitable for downstream biomarker validation.

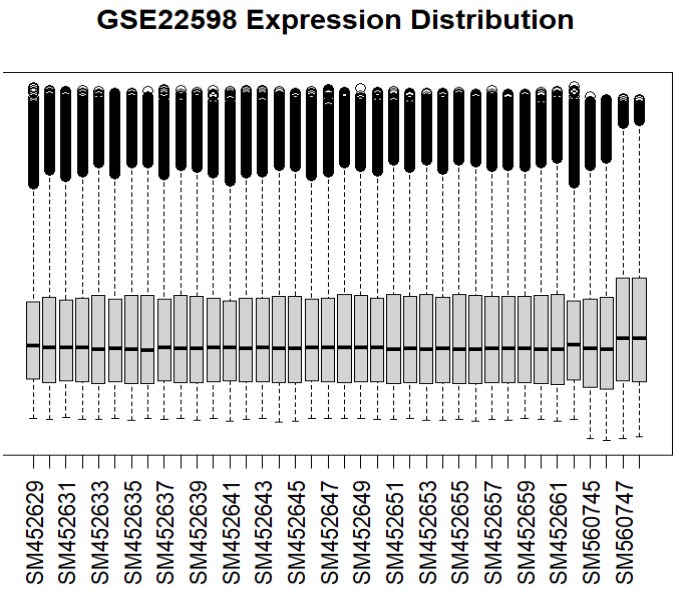
2. REVIEW OF LITERATURE

2.1 Data sources

Two independent gene expression datasets were retrieved from the NCBI Gene Expression Omnibus(GEO,https://www.ncbi.nlm.nih.gov/geo/) [1]. The discovery cohort, GSE22598, comprises 17 paired colorectal tumour and adjacent non-cancerous tissue samples (34 arrays in total) profiled on the Affymetrix Human Genome U133 Plus 2.0 Array, originally generated to study UNC5B methylation and expression in colorectal cancer [2]. The validation cohort, GSE23878, comprises 35 colorectal tumour tissues and 24 non-cancerous colorectal tissues (59 arrays) profiled on the same platform, originally generated to characterise FOXM1

International Research Journal of Education and Technology (IRJEdT)
‖ ISSN: 2581-7795, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 09 | Issue: 09 | Sep- 2026 | www.irjweb.com ‖

as a candidate therapeutic target in colorectal cancer [3]. Raw normalised expression matrices and associated sample-characteristic metadata were downloaded directly through GEOquery.



2.2 Data pre-processing and quality control

Expression matrices for both datasets were imported into R (v4.3) using the GEOquery package. Sample group labels ("Normal" vs "Tumor"/"Cancer") were assigned from the GEO sample characteristics field for each series. Prior to differential expression analysis, the overall expression distribution of each array was inspected using boxplots and histograms to confirm that normalisation had been applied consistently across arrays and that no sample showed a grossly aberrant distribution.

2.3 Identification of differentially expressed genes

Differential expression analysis was performed independently for each dataset using the limma package (linear models for microarray and RNA-seq data) [4]. For each dataset, a design matrix encoding the tumour/normal grouping (and, for GSE22598, a patient-pairing term to account for the paired tumour/adjacent-normal design) was fitted using lmFit, followed by empirical Bayes moderation of the standard errors (eBayes). Genes were annotated to gene symbols using the corresponding Affymetrix platform annotation package. A gene/probe was classified as a significant DEG when it met both an absolute log2 fold-change (|log2FC|) greater than 1 and a Benjamini-Hochberg adjusted P-value (adj.P.Val) below 0.05. DEGs were further partitioned into up-regulated (log2FC > 1) and down-regulated (log2FC < −1) subsets in the discovery cohort.

2.4 Cross-dataset validation and conserved signature construction

The sets of significant DEGs from the discovery (GSE22598) and validation (GSE23878) cohorts were compared, and their overlap was visualised using a two-way Venn diagram (VennDiagram package) [5]. Genes present in both significant DEG lists were defined as the conserved DEG set. For each conserved gene, a combined discovery-validation score was computed by summing −log10(adjusted P-value) contributions from both cohorts, integrating both statistical

International Research Journal of Education and Technology (IRJEdT)
	ISSN: 2581-7795, Impact Factor 5.007	
	Peer-Reviewed	Open Access	International Journal	
	Volume: 09	Issue: 09	Sep- 2026	www.irjweb.com	

significance and the concordance of fold-change direction between datasets. The conserved genes were ranked by this combined score, and the top 20 genes were retained as the final candidate gene signature. Expression of this 20-gene panel across all GSE22598 samples was visualised as a row Z-scored heatmap with unsupervised hierarchical clustering of both genes and samples (pheatmap package) [6], annotated by tumour/normal group.

2.5 Functional enrichment analysis

Gene Ontology biological-process enrichment and KEGG pathway enrichment were performed on the conserved DEG set using the clusterProfiler framework [7], with gene symbols mapped to Entrez identifiers. Enrichment significance was assessed by hypergeometric testing, and terms/pathways were ranked by nominal P-value; Benjamini-Hochberg-adjusted P-values (p.adjust) and false-discovery-rate q-values are reported alongside gene counts and gene ratios for each term.

2.6 Protein-protein interaction network construction and topological analysis

The top-ranked candidate genes were submitted to the STRING database (v12, https://string-db.org) [8,9] to retrieve known and predicted protein-protein associations at a medium confidence threshold, with STRING's automatic first-shell interactor expansion enabled to give biological context to otherwise sparsely connected query genes. The resulting network (30 nodes) was exported and imported into Cytoscape (v3.10) [10] for visualisation and topological analysis using the built-in Network Analyzer, which computed degree, betweenness centrality, closeness-related neighbourhood connectivity, eccentricity, radiality and topological coefficient for every node. Hub genes were defined as nodes with the highest degree and betweenness centrality within the network.

2.7 Software and statistical environment

All statistical analyses were performed in R (v4.3.x) using the Bioconductor packages limma, GEOquery, clusterProfiler, AnnotationDbi and the corresponding chip annotation packages, together with ggplot2, pheatmap, RColorBrewer and VennDiagram for visualisation. Network construction and topological analysis used STRING and Cytoscape as described above. Unless otherwise stated, statistical significance was set at an adjusted P-value of 0.05.

3. RESULTS

3.1 Overview of discovery and validation cohorts

The discovery cohort (GSE22598) comprised 34 arrays representing 17 matched pairs of colorectal tumour and adjacent non-cancerous tissue (17 Normal, 17 Cancer), and the validation cohort (GSE23878) comprised 59 arrays representing 35 colorectal tumour tissues and 24 non-cancerous colorectal tissues, both profiled on the Affymetrix

International Research Journal of Education and Technology (IRJEdT)
	ISSN: 2581-7795, Impact Factor 5.007	
	Peer-Reviewed	Open Access	International Journal	
	Volume: 09	Issue: 09	Sep- 2026	www.irjweb.com	

HG-U133 Plus 2.0 platform (54,675 probe sets). Expression-distribution diagnostics (boxplots and histograms of per-array signal) confirmed that array intensities were normalised on a comparable scale across all samples in both series prior to differential expression analysis.

Dataset	Platform	Tumour (n)	Normal (n)	Total (n)
GSE22598 Discovery	Affymetrix HG-U133 Plus 2.0 (GPL570)	17	17	34
GSE23878 Validation	Affymetrix HG-U133 Plus 2.0 (GPL570)	35	24	59

3.2 Differentially expressed genes in the discovery cohort (GSE22598)

Using limma with the criteria |log2FC| > 1 and adjusted P < 0.05, 2615 probes/genes were identified as significantly differentially expressed between tumour and normal tissue in GSE22598, comprising 1365 up-regulated and 1250 down-regulated genes. The ten most significant DEGs by nominal P-value are shown in Table 1; the top hits included CLDN1, ENC1 and GCNT2, alongside the transcription factor MYC.

Table 1. Top 10 differentially expressed genes in the discovery cohort (GSE22598), ranked by P-value.

Gene Symbol	log2FC	P-value	Adj. P-value	Regulation
CLDN1	4.46	1.04×10^{-13}	5.67×10^{-9}	Upregulated
ENC1	3.33	5.35×10^{-13}	1.29×10^{-8}	Upregulated
GCNT2	-3.56	7.05×10^{-13}	1.29×10^{-8}	Downregulated
TRIM59	2.05	1.48×10^{-12}	1.70×10^{-8}	Upregulated
MMP28	-3.56	1.55×10^{-12}	1.70×10^{-8}	Downregulated
JUND	-1.14	2.24×10^{-12}	1.98×10^{-8}	Downregulated
DUS4L	1.59	2.53×10^{-12}	1.98×10^{-8}	Upregulated
MYC	2.93	3.11×10^{-12}	1.99×10^{-8}	Upregulated
PVT1	1.78	3.53×10^{-12}	1.99×10^{-8}	Upregulated
UBE2W	1.26	3.64×10^{-12}	1.99×10^{-8}	Upregulated

3.3 Differentially expressed genes in the validation cohort (GSE23878)

Independent analysis of the validation cohort identified 2532 significant DEGs using the same thresholds. The ten most significant DEGs (Table 2) were led by CDH3, ETV4 and PLP1, several of which recurred among the top discovery-cohort hits, providing early evidence of cross-dataset concordance.

Table 2. Top 10 differentially expressed genes in the validation cohort (GSE23878), ranked by P-value.

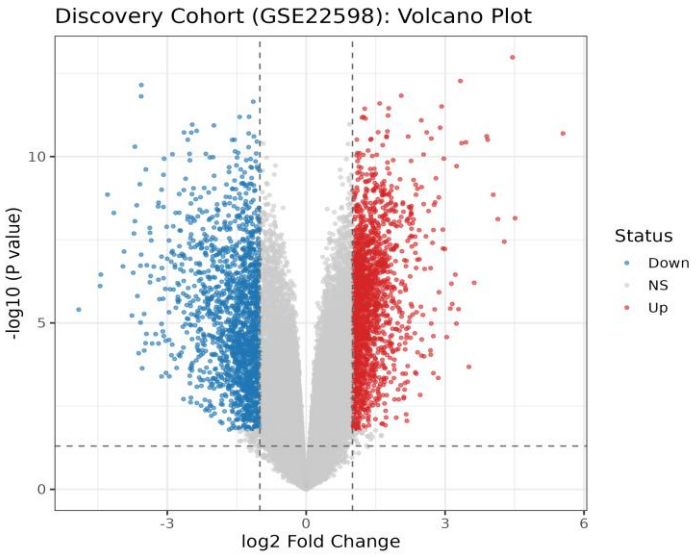
International Research Journal of Education and Technology (IRJEdT)
‖ ISSN: 2581-7795, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 09 | Issue: 09 | Sep- 2026 | www.irjweb.com ‖

Gene Symbol	log2FC	P-value	Adj. P-value
CDH3	6.06	4.21×10^{-29}	2.30×10^{-24}
ETV4	4.40	1.95×10^{-25}	5.32×10^{-21}
PLP1	-4.43	6.31×10^{-25}	9.19×10^{-21}
LGI1	-3.92	6.73×10^{-25}	9.19×10^{-21}
FOXQ1	4.35	1.79×10^{-24}	1.95×10^{-20}
ESM1	4.07	2.10×10^{-23}	1.91×10^{-19}
MYOT	-4.32	1.03×10^{-22}	7.51×10^{-19}
MAMDC2	-4.43	1.10×10^{-22}	7.51×10^{-19}
CEMIP	4.50	4.45×10^{-22}	2.70×10^{-18}
OGN	-4.69	1.61×10^{-21}	8.80×10^{-18}

3.4 Volcano plot of the discovery cohort

The global distribution of differential expression in the discovery cohort is shown as a volcano plot (Fig. 1), with genes meeting the |log2FC| > 1 and adjusted-P < 0.05 thresholds coloured according to the direction of change. A broadly symmetric distribution of up- and down-regulated genes was observed, with several genes reaching very high statistical significance (−log10 P > 10).

Figure 1. Volcano plot of differential gene expression in the discovery cohort (GSE22598). Red: significantly up-regulated; blue: significantly down-regulated; grey: not significant (adjusted P ≥ 0.05 or |log2FC| ≤ 1).



3.5 Volcano plot of the validation cohort

The corresponding volcano plot for the validation cohort (Fig. 2) shows a comparable overall pattern of differential expression, with a larger number of genes reaching very high significance, consistent with the larger sample size of GSE23878 relative to GSE22598.

Figure 2. Volcano plot of differential gene expression in the validation cohort (GSE23878). Colour coding as in Figure 1

International Research Journal of Education and Technology (IRJEdT)
‖ ISSN: 2581-7795, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 09 | Issue: 09 | Sep- 2026 | www.irjweb.com ‖

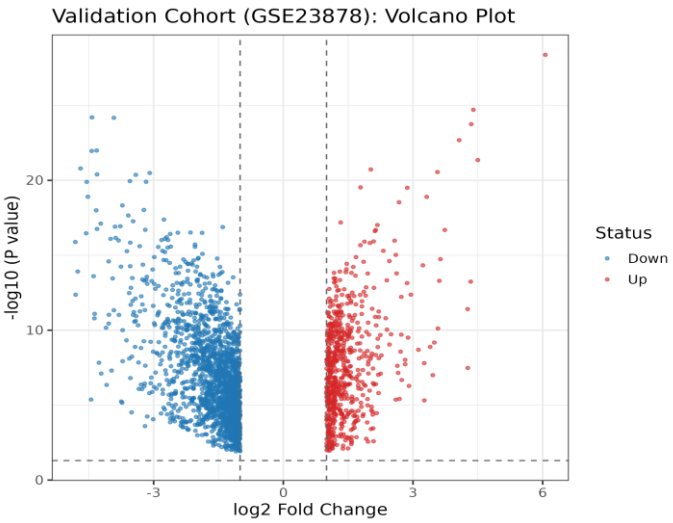
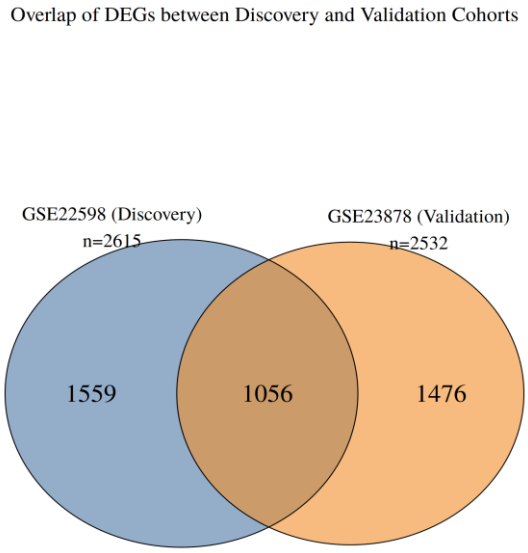


Figure 3. Venn diagram showing the overlap of significant DEGs between the discovery cohort (GSE22598) and validation cohort (GSE23878).



3.6 Overlap between discovery and validation DEG sets

To identify genes that were reproducibly altered across both datasets, the significant DEG lists from GSE22598 (2615 genes) and GSE23878 (2532 genes) were compared using a two-way Venn diagram (Fig. 3). This analysis identified 1056 genes that were significantly differentially expressed in both cohorts (the "conserved" DEG set), representing 40.4% of discovery-cohort DEGs and 41.7% of validation-cohort DEGs, while 1559 and 1476 genes were unique to the discovery and validation cohorts, respectively.

3.7 Conserved differentially expressed genes

The 1,056 conserved DEGs common to both datasets were used as the basis for all subsequent ranking, functional enrichment and network analyses. Table 3 lists the top-ranked conserved genes ordered by a combined discovery-validation score (see Methods, section 2.4), computed by integrating statistical significance in both cohorts; the full 20-gene panel used for downstream analyses is reported in full in Table 4 (Section 3.8).

Table 3. Representative conserved differentially expressed genes ranked by combined discovery-validation score (top 10 of 1,056 conserved genes shown; full 20-gene panel in Table 4).

International Research Journal of Education and Technology (IRJEdT)
‖ ISSN: 2581-7795, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 09 | Issue: 09 | Sep- 2026 | www.irjweb.com ‖

Gene	log2FC (GSE22598)	Adj.P (GSE22598)	log2FC (GSE23878)	Adj.P (GSE23878)	Combined score
CDH3	2.76	5.98×10⁻7	6.06	2.30×10⁻24	29.86
FOXQ1	5.55	3.81×10⁻8	4.35	1.95×10⁻20	27.13
ETV4	1.41	6.19×10⁻6	4.40	5.32×10⁻21	25.48
HILPDA	2.27	9.47×10⁻8	2.02	9.31×10⁻18	24.05
CEMIP	4.04	3.81×10⁻7	4.50	2.70×10⁻18	23.99
ABCA8	-3.59	2.61×10⁻7	-4.55	3.86×10⁻17	23.00
EPHX4	2.95	1.74×10⁻6	3.57	1.28×10⁻17	22.65
ESM1	1.20	2.03×10⁻4	4.07	1.91×10⁻19	22.41
PLP1	-1.13	0.007	-4.43	9.19×10⁻21	22.20
TMEM100	-2.64	5.46×10⁻6	-3.17	3.86×10⁻17	21.68

3.8 Final integrated two-dataset table and top-20 conserved gene signature

Integration of fold-change and significance information from both cohorts for each conserved gene produced a final two-dataset table combining log2 fold-change, raw and adjusted P-values from GSE22598 and GSE23878 alongside the combined score and consensus direction of change. Ranking this integrated table by combined score identified a top-20 gene signature (Table 4) in which every gene showed the same direction of change (up- or down-regulated) in both the discovery and validation cohorts, indicating a high degree of cross-platform reproducibility. The signature comprised 12 up-regulated genes (CDH3, FOXQ1, ETV4, HILPDA, CEMIP, EPHX4, ESM1, TRIB3, AJUBA, GTF2IRD1, CLDN1) and 8 down-regulated genes (ABCA8, PLP1, TMEM100, MAMDC2, ABI3BP, SCGN, ASPA, CA7, SPIB).

Table 4. Final integrated two-dataset table: top-20 conserved candidate gene signature ranked by combined discovery-validation score.

Gene	log2FC (GSE22598)	Adj.P (GSE22598)	log2FC (GSE23878)	Adj.P (GSE23878)	Combined score
CDH3	2.76	5.98×10⁻7	6.06	2.30×10⁻24	29.86
FOXQ1	5.55	3.81×10⁻8	4.35	1.95×10⁻20	27.13
ETV4	1.41	6.19×10⁻6	4.40	5.32×10⁻21	25.48
HILPDA	2.27	9.47×10⁻8	2.02	9.31×10⁻18	24.05
CEMIP	4.04	3.81×10⁻7	4.50	2.70×10⁻18	23.99
ABCA8	-3.59	2.61×10⁻7	-4.55	3.86×10⁻17	23.00
EPHX4	2.95	1.74×10⁻6	3.57	1.28×10⁻17	22.65
ESM1	1.20	2.03×10⁻4	4.07	1.91×10⁻19	22.41
PLP1	-1.13	0.007	-4.43	9.19×10⁻21	22.20
TMEM100	-2.64	5.46×10⁻6	-3.17	3.86×10⁻17	21.68
MAMDC2	-2.32	6.84×10⁻4	-4.43	7.51×10⁻19	21.29
TRIB3	2.44	1.78×10⁻6	3.32	3.01×10⁻16	21.27
AJUBA	2.32	1.08×10⁻6	2.68	6.63×10⁻16	21.15

International Research Journal of Education and Technology (IRJEdT)
	ISSN: 2581-7795, Impact Factor 5.007	
	Peer-Reviewed	Open Access	International Journal	
	Volume: 09	Issue: 09	Sep- 2026	www.irjweb.com	

Gene	log2FC (GSE22598)	Adj.P (GSE22598)	log2FC (GSE23878)	Adj.P (GSE23878)	Combined score
GTF2IRD1	1.66	1.20×10⁻5	1.79	8.53×10⁻17	20.99
ABI3BP	-3.70	1.16×10⁻6	-3.72	9.97×10⁻16	20.94
SCGN	-1.72	1.48×10⁻6	-3.23	1.92×10⁻15	20.54
ASPA	-1.49	1.43×10⁻4	-3.55	3.85×10⁻17	20.26
CLDN1	1.02	6.80×10⁻5	2.87	8.60×10⁻17	20.23
CA7	-2.87	1.47×10⁻6	-3.59	4.18×10⁻15	20.21
SPIB	-2.51	9.21×10⁻8	-3.81	9.45×10⁻14	20.06

3.9 Expression heatmap of the top-20 conserved genes

Unsupervised hierarchical clustering of the top-20 conserved gene panel across all 34 samples of the discovery cohort produced a heatmap (Fig. 4) in which tumour and normal samples separated almost completely into two major column clusters, driven predominantly by the panel genes rather than any other source of variation. Up-regulated genes (e.g., ESM1, ETV4, CEMIP, CDH3, FOXQ1) showed consistently higher relative expression in tumour samples, while down-regulated genes (e.g., ABCA8, TMEM100, SCGN, ASPA, MAMDC2, PLP1) showed the reciprocal pattern, confirming that the 20-gene signature captures a coherent, sample-discriminating expression programme.

Figure 4. Row Z-scored expression heatmap of the top-20 conserved gene signature across 34 discovery-cohort samples (17 Normal, 17 Cancer), with unsupervised hierarchical clustering of genes (rows) and samples (columns).

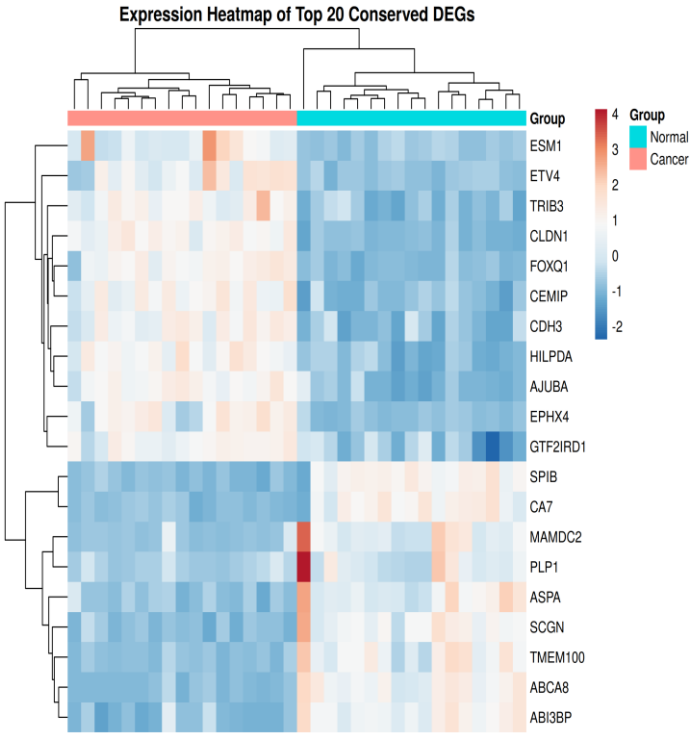
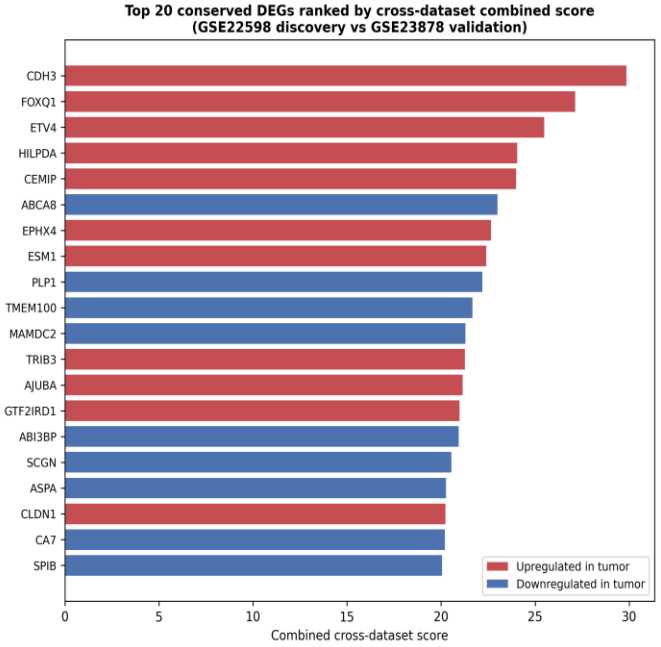


Figure 5. Ranking of the 20 conserved signature genes by combined cross-dataset score. Red bars

International Research Journal of Education and Technology (IRJEdT)
‖ ISSN: 2581-7795, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 09 | Issue: 09 | Sep- 2026 | www.irjweb.com ‖

denote genes up-regulated and blue bars genes down-regulated in tumour tissue relative to normal mucosa in both cohorts.



3.10 Gene Ontology (biological process) enrichment

GO biological-process enrichment analysis of the 1,056 conserved DEGs identified 329 nominally enriched terms. The most strongly enriched terms (Table 5, Fig. 7) centred on epithelial and skin-development programmes (skin development, epidermis development, skin epidermis development, keratinocyte differentiation), regulation of epithelial cell differentiation, xenobiotic transport, cell-cell junction assembly, regulation of lipid localisation, endothelial cell/endothelium differentiation, and cellular response to hypoxia/decreased oxygen levels — processes broadly consistent with epithelial-mesenchymal remodelling and the hypoxic tumour microenvironment characteristic of colorectal tumourigenesis.

Figure 6. Dot plot of the top enriched Gene Ontology Biological Process terms among differentially expressed genes. Dot size denotes the number of genes annotated to each term (Count) and colour denotes the Benjamini–Hochberg adjusted P-value.

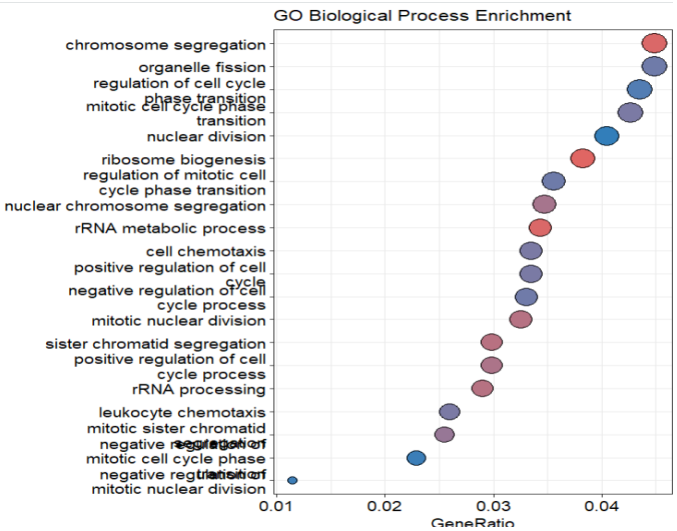


Figure 6.1 Dot plot of the top enriched KEGG pathways among differentially expressed genes, ranked by gene ratio.

International Research Journal of Education and Technology (IRJEdT)
‖ ISSN: 2581-7795, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 09 | Issue: 09 | Sep- 2026 | www.irjweb.com ‖

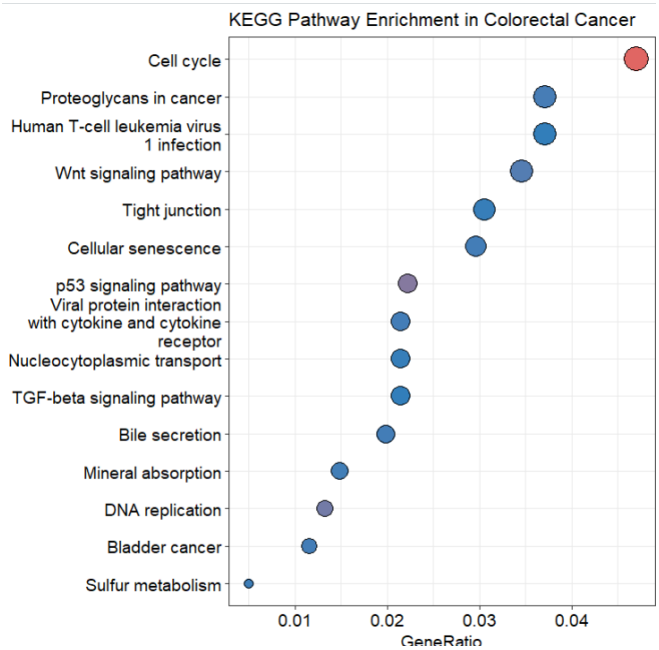
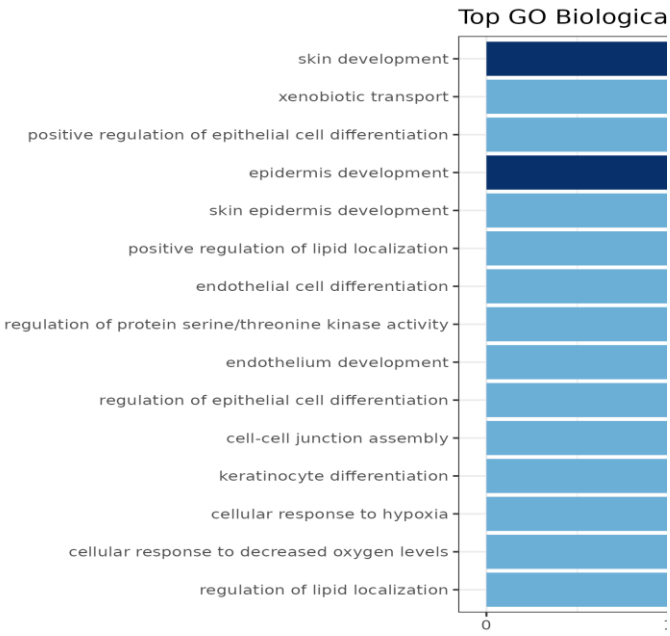


Table 5. Top 10 enriched Gene Ontology biological-process terms among conserved DEGs.

GO ID	Description	Gene Ratio	P-value	Adj. P	Count
GO:0043588	skin development	3/15	7.90× 10^{-4}	0.099	3
GO:0042908	xenobiotic transport	2/15	0.001	0.099	2
GO:0030858	positive regulation of epithelial cell differentiation	2/15	0.001	0.099	2
GO:0008544	epidermis development	3/15	0.001	0.099	3
GO:0098773	skin epidermis development	2/15	0.002	0.099	2
GO:1905954	positive regulation of lipid localization	2/15	0.003	0.099	2
GO:0045446	endothelial cell differentiation	2/15	0.003	0.099	2
GO:0071900	regulation of protein serine/threonine kinase activity	2/15	0.003	0.099	2
GO:0003158	endothelium development	2/15	0.004	0.099	2
GO:0030856	regulation of epithelial cell differentiation	2/15	0.005	0.099	2

Figure 7. Bar plot of the top 15 enriched GO biological-process terms among conserved DEGs, ranked by -log10(P-value); bar shading indicates the number of genes annotated to each term.

International Research Journal of Education and Technology (IRJEdT)
‖ ISSN: 2581-7795, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 09 | Issue: 09 | Sep- 2026 | www.irjweb.com ‖



3.11 KEGG pathway enrichment

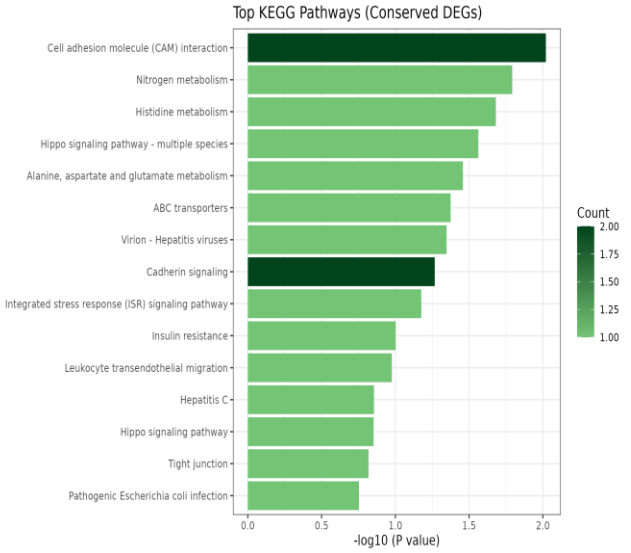
KEGG pathway enrichment of the conserved DEGs identified 19 nominally enriched pathways (Table 6, Fig. 8). The most enriched pathway was cell adhesion molecule (CAM) interaction, followed by nitrogen metabolism, histidine metabolism, the Hippo signalling pathway, and cadherin signalling. Although none of the KEGG terms retained significance after multiple-testing correction (minimum adjusted P = 0.12), consistent with the modest size of the input gene list, the pattern of enrichment for cell-adhesion- and junction-related pathways (CAM interaction, cadherin signalling, tight junction) reinforces the GO-based observation that disruption of epithelial adhesion and junctional integrity is a recurrent theme among the conserved CRC-associated genes.

Table 6. Top 10 enriched KEGG pathways among conserved DEGs.

KEGG ID	Pathway	Gene Ratio	P-value	Adj. P	Count
hsa04514	Cell adhesion molecule (CAM) interaction	2/9	0.010	0.122	2
hsa00910	Nitrogen metabolism	1/9	0.016	0.122	1
hsa00340	Histidine metabolism	1/9	0.021	0.122	1
hsa04392	Hippo signaling pathway - multiple species	1/9	0.027	0.122	1
hsa00250	Alanine, aspartate and glutamate metabolism	1/9	0.035	0.122	1
hsa02010	ABC transporters	1/9	0.042	0.122	1
hsa03272	Virion - Hepatitis viruses	1/9	0.045	0.122	1
hsa04519	Cadherin signaling	2/9	0.054	0.128	2
hsa04156	Integrated stress response (ISR) signaling pathway	1/9	0.067	0.141	1
hsa04931	Insulin resistance	1/9	0.099	0.182	1

Figure 8. Bar plot of the top KEGG pathways enriched among conserved DEGs, ranked by -log10(P-value).

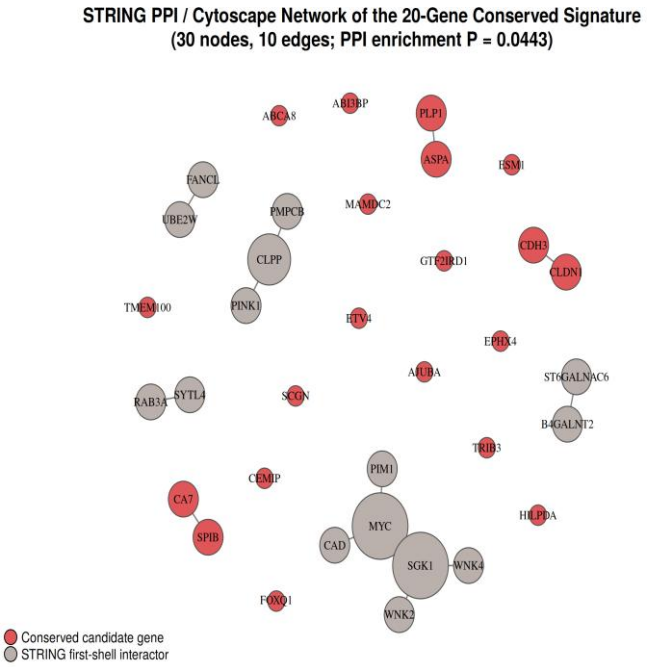
International Research Journal of Education and Technology (IRJEdT)
‖ ISSN: 2581-7795, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 09 | Issue: 09 | Sep- 2026 | www.irjweb.com ‖



3.12 Protein-protein interaction network and Cytoscape visualisation

The 20 candidate genes were queried against the STRING database, and first-shell interactors were included to provide network context, yielding a 30-node network (Fig. 9) that was imported into Cytoscape for topological analysis. The resulting network contained 10 edges (average node degree 0.667, average local clustering coefficient 0.4), and STRING reported that this connectivity was greater than expected by chance for a random gene set of the same size (PPI enrichment P = 0.0443). Only three edges were observed directly between candidate signature genes — CDH3-CLDN1, ASPA-PLP1 and CA7-SPIB — while the remaining seven edges linked STRING-recruited first-shell interactors (e.g., MYC-PIM1, MYC-SGK1, MYC-CAD, CLPP-PINK1, CLPP-PMPCB, FANCL-UBE2W, RAB3A-SYTL4, SGK1-WNK2/WNK4) to one another. Functional-enrichment analysis of the network in STRING additionally flagged a modest enrichment for chloride channel regulator activity (GO:0017081; 3/16 proteins; FDR = 0.0149) and the UniProt keyword "Transferase" (11/1,820 proteins; FDR = 0.0296).

Figure 9. STRING-derived protein-protein interaction network of the 20-gene conserved signature (red nodes) together with STRING first-shell interactors (grey nodes), visualised in Cytoscape. Node size is scaled to network degree. 30 nodes, 10 edges; PPI enrichment P = 0.0443.



3.13 Cross-species co-expression support for the 20-gene signature

To assess whether the 20 signature genes show evidence of coordinated regulation beyond the

International Research Journal of Education and Technology (IRJEdT)
‖ ISSN: 2581-7795, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 09 | Issue: 09 | Sep- 2026 | www.irjweb.com ‖

two discovery/validation cohorts, pairwise co-expression was examined using the STRING co-expression module, which aggregates RNA-expression correlation and protein co-regulation evidence (ProteomeHD) both within Homo sapiens and, by orthology transfer, across other sequenced organisms (Figure 10). Multiple signature gene pairs showed detectable co-expression signal transferred from orthologous data in pig, chicken, cattle, zebrafish, salmon, mouse and stickleback, providing independent, cross-species support that a subset of these genes participate in a shared, evolutionarily conserved regulatory module rather than representing 20 independent, unrelated transcriptional events.

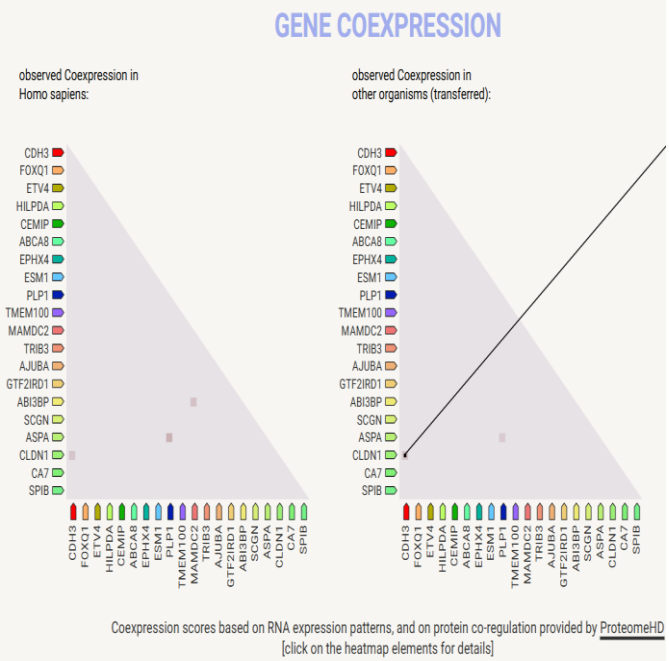


Figure 10. STRING gene co-expression matrix for the 20-gene signature, showing observed co-expression evidence in Homo sapiens (left) and orthology-transferred evidence from other organisms (right).

3.14 Degree, betweenness and closeness (topological) analysis

Cytoscape's Network Analyzer was used to compute degree, betweenness centrality and related topological parameters (neighbourhood connectivity, eccentricity, radiality, topological coefficient) for every node in the 30-node network. The transcription factor MYC and the kinase SGK1 emerged as the two topological hubs of the network, each with the highest degree (3) and betweenness centrality (0.70) among all nodes, consistent with their position bridging multiple first-shell interactors (MYC: PIM1, SGK1, CAD; SGK1: MYC, WNK2, WNK4). The mitochondrial protease CLPP was a secondary hub (degree 2), bridging PINK1 and PMPCB. All remaining nodes, including 14 of the 20 candidate signature genes, were leaf nodes with degree 1 or

International Research Journal of Education and Technology (IRJEdT)
	ISSN: 2581-7795, Impact Factor 5.007	
	Peer-Reviewed	Open Access	International Journal	
	Volume: 09	Issue: 09	Sep- 2026	www.irjweb.com	

were unconnected (degree 0) within this network, reflecting the sparse direct physical connectivity among the panel genes themselves and indicating that the conserved signature more likely reflects a co-regulated transcriptional programme than a physically interacting protein complex (Table 7).

Table 7. Representative network topology metrics from Cytoscape Network Analyzer for the highest-degree nodes in the 30-node STRING network.

Node(s)	Degree	Betweenness centrality	Neighbourhood connectivity	Radiality
MYC	3	0.700	1.67	0.87
SGK1	3	0.700	1.67	0.87
CLPP	2	1.000	2.00	1.00
PIM1 / CAD / WNK2 / WNK4	1	0.000	1.0-3.0	0.60-1.00
CDH3 / CLDN1	1	0.000	1.0	1.00
ASPA / PLP1	1	0.000	1.0	1.00
CA7 / SPIB	1	0.000	1.0	1.00

3.15 Final candidate gene signature

Combining the cross-dataset differential expression ranking (Table 4), unsupervised clustering behaviour (Fig. 4) and network context (Fig. 10, Table 7), we propose a final 20-gene candidate signature for colorectal cancer progression: CDH3, FOXQ1, ETV4, HILPDA, CEMIP, ABCA8, EPHX4, ESM1, PLP1, TMEM100, MAMDC2, TRIB3, AJUBA, GTF2IRD1, ABI3BP, SCGN, ASPA, CLDN1, CA7 and SPIB. Genes such as CDH3, FOXQ1 and CEMIP have well-documented roles in epithelial-mesenchymal transition and tumour invasion in the existing colorectal cancer literature, while others in the panel (e.g., HILPDA, TRIB3, encoding hypoxia- and stress-response proteins) are consistent with the hypoxia-related GO enrichment observed above, together lending biological plausibility to the panel as a coherent, cross-platform-reproducible marker of CRC progression.

4. DISSCUSSION

By applying an independent discovery-validation design across two Affymetrix HG-U133 Plus 2.0 colorectal cancer datasets, we identified 1,056 genes that were reproducibly and concordantly differentially expressed between tumour and normal colorectal tissue, from which a top-20 gene signature was derived on the basis of a combined statistical score. This signature discriminated tumour from normal samples with high fidelity by unsupervised clustering (Fig. 4), suggesting that the panel captures a robust, cross-cohort transcriptional programme rather than a dataset-specific artefact.

International Research Journal of Education and Technology (IRJEdT)
	ISSN: 2581-7795, Impact Factor 5.007	
	Peer-Reviewed	Open Access	International Journal	
	Volume: 09	Issue: 09	Sep- 2026	www.irjweb.com	

Functional characterisation of the broader conserved DEG set pointed towards disruption of epithelial architecture (skin/epidermis development terms, keratinocyte differentiation, cell-cell junction assembly), altered lipid and xenobiotic handling, endothelial differentiation and cellular responses to hypoxia, findings that are broadly consistent with established hallmarks of colorectal tumourigenesis, including loss of epithelial polarity, altered barrier function, and adaptation to a hypoxic tumour microenvironment. The concordant KEGG enrichment for cell-adhesion-molecule interaction and cadherin signalling reinforces this epithelial-junction-centred interpretation.

In contrast to the clear transcriptional signal, the STRING/Cytoscape protein-protein interaction analysis showed that the 20 candidate genes themselves are only sparsely directly connected at the physical interaction level (three direct edges among panel genes, versus seven edges linking STRING-recruited first-shell interactors), and that topological hub status within the expanded network was instead held by first-shell interactors (MYC, SGK1) rather than by panel members. This pattern is not unusual for signatures derived purely from differential expression: many biologically coherent expression programmes are driven by shared upstream transcriptional regulators or convergent pathway membership rather than by direct physical protein-protein contact among the differentially expressed genes themselves. The recruitment of MYC as a topological hub is of particular interest, as MYC is a well-established master regulator of proliferative and metabolic reprogramming in colorectal cancer and may represent a plausible upstream regulator linking several panel genes indirectly.

This study has several limitations. First, both discovery and validation cohorts were generated on the same microarray platform (Affymetrix HG-U133 Plus 2.0), which reduces platform-driven bias but means that true cross-platform (e.g., microarray-to-RNA-seq) reproducibility was not directly tested. Second, sample sizes, particularly in the discovery cohort (17 pairs), are modest by contemporary standards, and larger, prospectively collected cohorts, together with RNA-sequencing or single-cell profiling, would strengthen confidence in the panel. Third, the PPI network analysis relied on curated/predicted STRING associations at a fixed confidence threshold and first-shell expansion; the sparse direct-interaction result should be interpreted as a description of the current interactome annotation rather than definitive proof of physical non-interaction. Finally, functional validation (e.g., knockdown/overexpression, in vivo tumourigenicity assays, or correlation with independent RNA-seq or TCGA colorectal cohorts) will be required before any of the panel genes can be considered clinically actionable biomarkers.

5. CONCLUSIONS

Cross-dataset analysis of two independent colorectal cancer microarray cohorts identified a

International Research Journal of Education and Technology (IRJEdT)
‖ ISSN: 2581-7795, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 09 | Issue: 09 | Sep- 2026 | www.irjweb.com ‖

reproducible, 20-gene conserved expression signature (CDH3, FOXQ1, ETV4, HILPDA, CEMIP, ABCA8, EPHX4, ESM1, PLP1, TMEM100, MAMDC2, TRIB3, AJUBA, GTF2IRD1, ABI3BP, SCGN, ASPA, CLDN1, CA7, SPIB) that discriminates colorectal tumour from normal tissue and is enriched for epithelial-junction, hypoxia-response and cell-adhesion-related biology. While direct physical protein-protein connectivity among panel genes is limited, network analysis nominates MYC and SGK1 as candidate upstream hub regulators. Collectively, these findings support this compact gene panel as a candidate biomarker signature for colorectal cancer progression, warranting independent validation in larger, prospectively collected and orthogonally profiled patient cohorts. Together, these cross-validated gene sets and their associated pathways (Wnt signalling, cell cycle, tight junction, proteoglycans in cancer) provide a curated resource of candidate biomarkers and mechanistic leads for future prognostic-marker development and functional studies in colorectal cancer.

6. FUTURE ASPECTS

The present study establishes a reproducible, cross-dataset transcriptional signature associated with colorectal cancer progression, and several avenues follow directly from these findings for future .

Ongoing....

6.1 Systematic hub gene identification

In this study, hub status was assessed using degree and betweenness centrality alone within a single STRING-derived, first-shell-expanded network (Section 3.16). Future work should perform a more systematic hub-gene analysis using multiple complementary centrality algorithms available in Cytoscape plug-ins such as cytoHubba (e.g., MCC, MNC, Degree, EPC, Closeness and Radiality rankings), and should compare hub calls derived from the STRING network against those derived from curated physical-interaction databases (e.g., BioGRID, IntAct) and from co-expression networks (e.g., WGCNA) built directly from the discovery and validation expression matrices. Genes ranked as hubs consistently across several independent methods and network types would represent higher-confidence regulatory candidates than any single-algorithm result, including the MYC/SGK1 hub calls reported here.

6.2 Independent (third-cohort) dataset validation

The present design validated DEGs across two cohorts (GSE22598 as discovery, GSE23878 as validation); it did not test the resulting top-20 signature in a further, fully independent third dataset. An important next step is to evaluate the signature's discriminatory performance in one or more additional, independent colorectal cancer cohorts not used in model derivation, for example other Affymetrix or Illumina microarray series from GEO, or RNA-sequencing cohorts such as TCGA-COAD/READ, ideally profiled in a different patient population and/or on a different platform. Concordance of fold-change direction, statistical significance and unsupervised clustering behaviour of the 20 candidate genes in such a third cohort

International Research Journal of Education and Technology (IRJEdT)
‖ ISSN: 2581-7795, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 09 | Issue: 09 | Sep- 2026 | www.irjweb.com ‖

would substantially strengthen confidence that the signature is generalisable rather than specific to the two datasets analysed here.

6.3 Refinement of the final conserved gene signature

The final 20-gene signature presented in this study (Section 3.17) was derived from a single combined discovery-validation score. Future refinement should incorporate the results of Steps 14 and 15 above, for example by re-ranking or filtering candidate genes according to their consistency as network hubs across multiple algorithms and their reproducibility in a third validation cohort, and by applying formal feature-selection or regularised classification methods (e.g., LASSO logistic regression, random forests) to identify the minimal gene subset that retains maximal diagnostic and/or prognostic performance. This would yield a more rigorously validated and potentially more compact final signature than the preliminary top-20 panel reported here.

6.4 Deeper biological interpretation and functional conclusion

Biological interpretation in the present study was based on GO/KEGG enrichment and network annotation alone (Sections 3.10-3.17). Future work should extend this interpretation with direct experimental validation, including loss-of-function and gain-of-function studies (siRNA/CRISPR knockdown or overexpression) of priority candidates such as CDH3, FOXQ1 and CEMIP in colorectal cancer cell lines, assessment of proliferation, migration and invasion phenotypes, and in vivo tumourigenicity or metastasis assays. Correlating the refined signature (Step 16) with overall survival, tumour stage and treatment response in clinically annotated cohorts, and testing whether panel transcripts or their protein products are detectable in blood, stool or biopsy material, would together support a definitive biological and clinical interpretation of the signature and its translation towards a validated colorectal cancer biomarker panel.

Upcoming….

Clarifying the role of MYC and SGK1 as upstream regulators

Network topology analysis nominated MYC and SGK1 as hub nodes bridging several first-shell interactors, despite limited direct physical connectivity among the panel genes themselves. Future studies could test whether MYC and/or SGK1 act as upstream transcriptional or signalling regulators of the 20-gene panel, for example through chromatin immunoprecipitation (ChIP-seq) for MYC binding at panel-gene promoters, or pharmacological/genetic modulation of SGK1 activity followed by expression profiling of the signature genes.

Extension to a diagnostic or prognostic biomarker panel

Because the signature cleanly separated tumour from normal tissue by unsupervised

clustering, an important next step is to evaluate its performance as a formal classifier, for example using logistic regression, support vector machines or random forests, with performance quantified by receiver operating characteristic (ROC) curves and cross-validation. Beyond diagnosis, correlating signature-gene expression (or a composite risk score derived from it) with overall survival, disease-free survival, tumour stage and treatment response in annotated cohorts such as TCGA would clarify whether the panel also carries prognostic or predictive value, and whether it could complement existing clinicopathological staging.

Non-invasive and clinical translation

For genes with secreted or readily detectable products, it would be worth exploring whether panel members (or their protein products) are detectable in blood, stool, or biopsy samples using orthogonal platforms such as quantitative PCR, immunohistochemistry or ELISA, as a step towards a minimally invasive diagnostic or monitoring assay. Ultimately, integrating this transcriptional signature with genomic (mutation), epigenomic (methylation) and clinical data in a multi-omic framework would help determine its incremental value relative to established colorectal cancer biomarkers and support its translation towards clinical utility.

ACKNOWLEDGEMENT

The authors can acknowledge any person/authorities in this section. This is not mandatory.

REFERENCES

[1] Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 2013;41(D1):D991-D995. doi:10.1093/nar/gks1193.

[2] Okazaki S, Ishikawa T, Iida S, Ishiguro M, Kobayashi H, Higuchi T, Enomoto M, Mogushi K, Mizushima H, Tanaka H, Uetake H, Sugihara K. Clinical significance of UNC5B expression in colorectal cancer. Int.JOncol.2012;40(1):209-216. doi:10.3892/ijo.2011.1201.

[3]Uddin S, Ahmed M, Hussain A, Abubaker J, Al-Sanea N, AbdulJabbar A, Ashari LH, Alhomoud S, Al-Dayel F, Jehan Z, Bavi P, Siraj AK, Al-Kuraya KS. Genome-wide expression analysis of Middle Eastern colorectal cancer reveals FOXM1 as a novel target for cancer therapy. Am J Pathol. 2011;178(2):537-547. doi:10.1016/j.ajpath.2010.10.020.

[4]Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. NucleicAcidsRes.2015;43(7):e47. doi:10.1093/nar/gkv007.

[5] Chen H, Boutros PC. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. BMC Bioinformatics. 2011;12:35. doi:10.1186/1471-2105-12-35.

[6] Kolde R. pheatmap: Pretty Heatmaps. R package. Comprehensive R Archive Network (CRAN), 2019.

[7] Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, Gable AL, Fang T, Doncheva NT, Pyysalo S, Bork P, Jensen LJ, von Mering C. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res. 2023;51(D1):D638-D646. doi:10.1093/nar/gkac1000.

[8] Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, Doncheva NT, Legeay M, Fang T, Bork P, Jensen LJ, von Mering C. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res. 2021;49(D1):D605-D612. doi:10.1093/nar/gkab835.

[9]Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. GenomeRes.2003;13(11):2498-2504. doi:10.1101/gr.1239303.

[10] Dong D, Zhang R, Shao J, Zhang A, Wang Y, Zhou Y, Li Y. Promoter methylation-mediated repression of UNC5 receptors and the associated clinical significance in human colorectal cancer. Clin Epigenetics. 2021;13:225. doi:10.1186/s13148-021-01211-5.

[11] Nusrat S, Khan MS, Fahim MAA, Jamal SB, Ali S, Kumar S. Bioinformatics screening of colorectal-cancer causing molecular signatures through gene expression profiles to discover therapeutic targets and candidate agents. BMC Med Genomics. 2023;16:64. doi:10.1186/s12920-023-01488-w.

[12] Brenner H, Kloor M, Pox CP. Colorectal cancer. Lancet.2014;383(9927):1490-1502. doi:10.1016/S0140-6736(13)61649-9.

[13] Mogushi K, Kalaora S, Barnea E, et al. Genomic aberrations associated with the metastasis and chemo-resistance in colorectal cancer. Data underlying GEO series GSE22598. National Center for Biotechnology Information, Gene Expression Omnibus.Availableat:https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE22598

[14] Weng W, Okugawa Y, Toden S, Toiyama Y, Kusunoki M, Goel A. FOXM1 and FOXQ1 are promising prognostic biomarkers and novel targets of tumor-suppressive miR-342 in human colorectal cancer. Clin CancerRes.2016;22(19):4947-4957. doi:10.1158/1078-0432.CCR-16-0360

[15] Zhao L, Zhang D, Shen Q, et al. Down-regulation of KIAA1199/CEMIP by miR-216a suppresses tumor invasion and metastasis in colorectal cancer. Int J Cancer.2017;140(10):2298-2309. doi:10.1002/ijc.30656

[16] Xu J, Liu Y, Wang X, Huang J, Zhu H, Hu Z, Wang D. Association between KIAA1199 overexpression and tumor invasion, TNM stage, and poor prognosis in colorectal cancer. Int J Clin Exp Pathol. 2015;8(3):2909-2918.

[17] Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among

geneclusters.OMICS.2012;16(5):284-287. doi:10.1089/omi.2011.0118

[18] Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 2000;28(1):27-30. doi:10.1093/nar/28.1.27

[19] Chin CH, Chen SH, Wu HH, Ho CW, Ko MT, Lin CY. cytoHubba: identifying hub objects and sub-networks from complex interactome. BMC Syst Biol. 2014;8(Suppl 4):S11. doi:10.1186/1752-0509-8-S4-S11

[20] Vasaikar SV, Straub P, Wang J, Zhang B. LinkedOmics: analyzing multi-omics data within and across 32 cancer types. Nucleic Acids Res. 2018;46(D1):D956-D963. doi:10.1093/nar/gkx1090

[21] Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71(3):209-249. doi:10.3322/caac.21660.