Open Peer Review on Qeios

Microsatellite Instability of Colon adenocarcinomas in India comprises multiple molecular subtypes

Prasanth Ariyannur[1], Veena P Menon[1], Keechilat Pavithran[1], Roopa R. Paulose[1], Damodaran M. Vasudevan[1]

1 Amrita Institute of Medical Sciences and Research Centre

Funding: Indian Council of Medical Research Amrita Vishwa Vidyapeetham

Potential competing interests: No potential competing interests to declare.

Abstract

The microsatellite stable category is more than four-fifths of global colon and rectal cancer cases (CRC). But many studies on the Indian population during the last two decades have shown that the microsatellite instable CRC was more than 30%. We have conducted a study to explore the possible pathogenesis of microsatellite instability in Indian CRC. In the preliminary investigation, we conducted a Nanostring Pan-Cancer pathway expression analysis of early-stage CRC tumor-normal pairs (n = 11, MSS = 5, MSI = 6). The differentially expressed genes associated with the tumor and correlated against microsatellite instability status were identified. Among them, *AXIN2*, *ETV4*, and *RNF43* were tumor cell-specific signals showing important differential expression patterns between MSI and MSS groups. The expression of these genes was further validated in another set of microsatellite instable early-stage CRC tissues (n = 15) by qPCR. The expression fold-changes of these signals suggest a distinct subset of MSI subgroups of CRC in the Indian population. Further validation in a larger cohort could enable a better understanding of the pathogenic mechanism as well as clinical utility of this expression pattern.

Prasanth Ariyannur[1], **Veena P. Menon**[2], **Keechilat Pavithran**[3], **Roopa Rachael Paulose**[4], **Damodaran M. Vasudevan**[1,5].

[1] *Department of Biochemistry and Molecular Biology, Amrita Institute of Medical Sciences, Amrita Vishwa Vidyapeetham, Kochi Kerala*

[2] *Department of Virology, Amrita Institute of Medical Sciences, Amrita Vishwa Vidyapeetham, Kochi, Kerala*

[3] *Department of Medical Oncology, Amrita Institute of Medical Sciences, Amrita Vishwa Vidyapeetham, Kochi Kerala.*

[4] *Department of Pathology, Amrita Institute of Medical Sciences, Amrita Vishwa Vidyapeetham, Kochi Kerala*

[5] *Department of Health Sciences Research, Amrita Institute of Medical Sciences, Amrita Vishwa Vidyapeetham, Kochi Kerala*

Keywords: colon adenocarcinoma, microsatellite instability, mismatch repair deficiency, AXIN2, ETV4, RNF43

Highlights

- Preliminary Nanostring Pan-cancer analysis of early-stage CRC identified a differential expression of a distinct set of genes.
- Superimposition of top significant differentially expressed genes on protein-protein interaction network analysis and Tumor Immune Estimation Resource filtered out distinct tumor cell-specific gene signals, *AXIN2*, *ETV4*, and *RNF43*.
- Validation of these signals on another set of early-stage MSI CRC cases identified two subcategories of MSI CRC based on the expression pattern of three gene signals.

Abbreviations

AIMS	Amrita Institute of Medical Sciences
CIN	Chromosomal Instability
COAD	Colon Adenocarcinoma
CRC	Colon and Rectal Carcinoma
DE	Differentially Expressed
FC	Fold Change
FDR	False Discovery Rate
FFPE	Formalin Fixed Paraffin Embedded
GEPIA	Gene Expression Profiling Interactive Analysis
GLOBOCAN	Global Cancer Observatory
IACR	International Association of Cancer Registries
IHC	Immunohistochemistry
LOH	Loss of Heterozygosity
MMR	Mismatch Repair
MSI	Microsatellite Instability
MSS	Microsatellite Stable
PPI	Protein-Protein Interaction
QPCR	Quantitative PCR
RT-PCR	Reverse Transcriptase Polymerase Chain Reaction
TCGA	The Cancer Genome Atlas
TIL	Tumor Infiltrating Lymphocytes
TIMER	Tumor Immune Estimation Resource
UICC	Union for International Cancer Control

Introduction

The prevalence rate of colon cancer in India is rising, according to a recent population-based cancer registry (PBCR) by the Indian Council for Medical Research (ICMR) (NCRP, 2016; Yeole, 2008). Average prevalence rates of 4.3 and 3.4 in 100,000 males and females, respectively, across different parts of India, as per the recent epidemiological reports on the colon, rectal, gall bladder, and pancreatic cancers (Shakuntala et al., 2022). The International Association of Cancer Registries (IACR) Global Cancer Observatory (GLOBOCAN 2020) showed an incidence of 4.9 per 100,000, with a higher rate of mortality in men than women (6.3 vs. 3.7). Another recent global study conducted on the Indian population showed that age-standardized incidence is 8.3. The mortality rate is 7.5 (Sharma et al., 2022). There is a wide variation in the incidence rate of CRC across India (east and south more than west and north) due to the differences in the distribution of essential but potentially modifiable lifestyle-related risk factors (Shakuntala et al., 2022).

More than thirty years of research on the molecular pathophysiology of colorectal cancer (CRC) have led to the identification of many novel mechanisms of tumorigenesis. In the western population, studies have shown that the overall pathogenicity of CRC is highest due to chromosomal instability (CIN), leading to widespread loss of heterozygosity (LOH) and gross chromosomal rearrangements, and high somatic copy number variations (Leary et al., 2008; Lin et al., 2003). The genomic instability of colonic cellular DNA can be due to genetic or non-genetic (including epigenetic, transcriptional, and post-translational) modifications of one or several signals. Defective DNA mismatch repair (MMR) and consequent microsatellite instability is the next common cause which is roughly one-fifth of the first cause, which includes the hypermutated type in the Cancer Genome Atlas Classification (TCGA, 2012). More extensive molecular and genomic pathogenesis studies on different cancer types, such as The Cancer Genome Atlas (TCGA), did not incorporate representative molecular epidemiological contributions from other ethnic groups (Hause et al., 2016). In India, multiple pioneer long-term multiple-series studies conducted in southern states, Tamil Nadu, Karnataka, Telangana, and Kerala showed a varying frequency of MSI in the CRC from 30% to 67% of cases (Dunne et al., 2016; Kanth et al., 2014; Pandey et al., 2007; Rajkumar et al., 2004; Raman et al., 2014). With these data, however, there has not been any molecular pathogenesis study to explore the reasons behind the higher prevalence of MMR deficiency in the pathogenesis of CRC in the Indian population. We have conducted a pilot expression analysis using Nanostring Pan-cancer pathway analysis in early-stage CRC comprising both MSI and MSS groups (Ariyannur et al., 2021). Upon seeing different gene expression patterns, we extended the Nanostring study more into the early-stage MSI CRC group with more samples. In the current study, we also explored orthogonal testing to validate those signals by RT-PCR.

Materials and Methods

Formalin-fixed paraffin-embedded (FFPE) tissue blocks from 30 archived early-stage colon cancer tissues which were previously identified as microsatellite unstable-high (MSI-H) using Immunohistochemistry (IHC) and MSI-PCR methods during the past five years were included in the study. We conducted Nanostring nCounter analysis for initial comparative screening in eleven archived tumor tissues and seven adjacent normal tissues. Subsequently, we selected another fifteen tumor-normal paired samples for the validation analysis using RT-PCR. Individual case details for the two arms of the study are included in Supplementary Dataset file S1. The Scientific Research and Institutional Ethics Committee approved the study before collecting pathology specimens and processing them. There were separate FFPE blocks of tumor tissues and proximal or distal resected colon tissues, which were histologically verified to be expected. MSI-PCR further characterized these tissues to confirm their MSI status after obtaining the sections. Cases that did not have normal archived tissues available were removed from the study.

Nanostring nCounter assay

The Nanostring nCounter assay was conducted by a contract research organization (Theracues Pvt Ltd, Bangalore, India), using validated commercial methods on a pay-by-service basis. Briefly, RNA was extracted from three-to-four 5μm FFPE tissue sections using a commercial FFPE nucleic acid isolation kit (Roche Molecular Diagnostics), quantified, and analyzed for fragment distribution using a bioanalyzer (Agilent). About 140 ng of RNA was used for each probing assay. The fluorescence count obtained from the nCounter machine (cutoff value ≥ 40) was normalized according to the positive controls and internal housekeeping reference genes. After standardization with standard expression sets, the median values of the probe set expression and clustering analysis was employed using appropriate package modules in R statistical program (Dunne et al., 2016). The differentially expressed (DE) genes in the tumor were calculated by taking the geometric mean of the six normal tissues as the denominator and the geometric mean of all the eleven tumor samples as the numerator. The Fold Changes (FC) of each gene was calculated from this ratio data, ratio > 1 | FC = Ratio; ratio < 1 | FC = -1*(1/Ratio). The top 20 up-and down-regulated genes were used in R-program to generate the principal component analysis (PCA) using ClustVis Field (Metsalu and Vilo, 2015). The Pathway analysis was done for each sample separately using DE genes (DE Call - Yes) using the WEB-based GEneSeTAnaLysis Toolkit (Web Gestalt) (Wang et al., 2017). The evaluation of differential expression values by group difference between microsatellite instability status, infiltration status, and the anatomical location was done using Student's *t-test* analysis, assuming unequal variance, and the Benjamini-Hochberg procedure was used to obtain the false discovery rate (FDR) adjusted *p*-value. DE genes with $|FC(\log_2)| > 1$ and *p*-value < 0.05 were considered significant.

Gene expression comparative analysis

Network Analysis

Protein-protein interaction (PPI) of the top DE genes was explored using the STRING database (string-db.org), an integrated, publicly available resource, based on laboratory experiments in protein-protein interactions, conserved co-expression data, genomic context predictions, and text mining from previous research (Szklarczyk et al., 2019). The top 20 highly significant genes from the Nanostring expression analysis and their FC(log2) values were entered into the STRING database online to explore the possible PPI among the signals. The ranking of each of these proteins was determined from the FC values.

Co-expression analysis

The significant DE genes identified were evaluated for co-expression analysis by web-based tools, Gene Expression Profiling Interactive Analysis (GEPIA) (http://gepia.cancer-pku.cn/) (Tang et al., 2019) and Tumor Immune Estimation Resource (TIMER 2.0: http://timer.cistrome.org/) (Li et al., 2017; Li et al., 2020). GEPIA and TIMER are online tools used to explore and compare the differential expression analysis results obtained from the Nanostring study against the transcriptomic data generated from more extensive studies such as The Cancer Genome Atlas (TCGA), and Genotype-Tissue Expression (GTEx) projects (Buhard et al., 2006; Li et al., 2017). Using both databases, the correlation of each of the selected genes, found to be expressed together in the test samples, were analyzed for their co-expression correlation coefficient in the TCGA COAD subset of these databases. The genes with correlation coefficient > 0.5 and $p < 0.05$ were correlated with the FC($\log_2$) and p-values of those genes in the current study using Pearson correlation using the *corr* function in the R-statistical program.

Immune cell infiltrates Analysis

TIMER is a web-based database tool that can also analyze and compare tumor-infiltrating immune cells such as B-cells, $CD4^+$ T cells, $CD8^+$ T cells, neutrophils, macrophages, and dendritic cells from gene expression profiles of TCGA data (Buhard et al., 2006) against the experimental data. The selected genes from the study samples were compared against the immune cell infiltrate to identify the type of immune cells that were maximally correlated in the TCGA COAD cohort from the given set of unique genes identified in the local population. The Spearman correlation coefficient (ρ) values of each cell type against enriched genes were obtained from the TIMER database (https://cistrome.shinyapps.io/timer/) in the Gene module tool. Hierarchical clustering analysis was performed on the ordered variance of the ρ to create a heatmap analysis using the *hclust* function in the R statistical program.

Quantitative RT-PCR analysis

Tissue processing

Another cohort of early-stage MSI CRC cases with matched tumor-normal paired archived FFPE blocks (n = 15) was assigned as the validation set for RT-PCR orthogonal testing. From the FFPE tissue blocks, 10μ thick 3-5 sections were made in a microtome and taken in a microcentrifuge tube for RNA extraction. RNA extraction and isolation were done using RNAeasy® FFPE kit (Qiagen, Germany), according to the manufacturer's instruction. Briefly, FFPE tissue sections were deparaffinized using xylene, permeabilized using 90% ethanol, lysed using proteinase K (at 56°C for 15 min), and nuclear separation (20,000 x g for 15min). The supernatant containing RNA was pipetted out and mixed with DNAase booster buffer to a tenth of the total volume. After incubation for 15min at room temperature, with 320μl of buffer and 720μl of 100% ethanol, the solution was passed through a min-Elute spin column and placed inside a 2ml collection tube. Subsequently, the centrifuge was done at 8000 x g for 15 s. The column was washed thrice with 500μl buffer RPE. Finally, 14μl of RNase-free water was added into the column attached to a fresh microcentrifuge tube and then centrifuged at 15000 x g for 1 min to elute RNA into the collection tube. RNA was quantified using Qubit RNA HS assay according to the manufacturer’s instruction in a fluorometer (Qubit 4.0, Thermo Scientific, USA).

cDNA synthesis

Using the RevertAid First Strand cDNA Synthesis Kit (Thermo Scientific, USA), first strand cDNA synthesis was done by methods described by the manufacturer. Before cDNA synthesis, RNA can be treated with RNase-free DNase I to remove trace amounts of DNA. A control (RTminus) reaction was included in all components of RTPCR except for the reverse transcriptase enzyme. We used 50ng of total RNA to generate the first strand of cDNA as an initial step for the two-step RT-PCR protocol. In a 20μl assay, 20ng of the template RNA was mixed with 1μl of oligo(dT)18 primer, 1μl of 10mM dNTP mix, 1μl of 20U/μl RiboLock RNase inhibitor, 4μl of 5x reaction buffer and made-up with 12μl of nuclease-free water. After mixing and a short centrifuge, the mixture was incubated for 60min at 42°C and terminated the reaction by heating up to 70°C for 5 min. The cDNA product was cooled in ice and stored at -20°C or immediately taken for further application.

Quantitative RT-PCR

Following the Nanostring studies, genes identified specific to tumor purity and correlating with the MSI characteristics were validated by RT-PCR. Along with the particular genes, common genes associated with Colon cancer (β-Catenin gene CTNNB1) and marker for immune cells (Toll-like receptor 4, TLR4) were also explored in the same tissues. From previous studies, the PCR primer sequences for the genes *AXIN2* (Pangestu et al., 2021), *ETV4* (Dumortier et al., 2018), *RNF43* (Pangestu et al., 2021), *TLR4* (Yang et al., 2010), and *CTNNB1* (Janitschke et al., 2020; Xu et al., 2019) were obtained and are given in Table 1. *C1or43* was taken as internal housekeeping gene control as this was found to be better than other housekeeping genes in colonic cells (Janitschke et al., 2020; Xu et al., 2019). The primers were custom-synthesized and sourced from a commercial agency (SIGMA, Bengaluru). The assay was established per the previous publications with minor modifications. Briefly, the assay comprises 5μl of 2X master mix, with 2μl of forward and reverse primers, 1μl of template DNA, and 2μl of Nuclease Free water added. PCR includes an initial denaturation step at 95°C for 10min followed by 40 cycles of denaturation at 95°C for 15s and annealing /extension at 55°C for 1min. Real-time PCR was done in Qiagen Rotogene Q Realtime PCR system. Cq values were obtained with mean Cq and standard deviation. Standard deviations were below 0.2, which are permissible for the reliability of results. The Cq values of C1orf43 were found to be equal across the samples. Relative Fold-change (RFC) was calculated among fifteen paired samples using the formula CN = $2^{(-\Delta\Delta Cq)}$, where ΔCq is the difference between the Cq values of the target gene and housekeeping gene control (C1orf43) and ΔΔCq is the difference between ΔCq of tumor and normal. A limit of detection assay was performed to identify the relationship between log values of copy numbers [In(CN)] and Cq values. From the Cq values of the target genes, the CN was resolved from the standard curve. The ratio of absolute CN (Absolute fold-change) of each target in tumor to normal was calculated to estimate and compare the RFC.

Table 1. Primer sequences for the RT-PCR validation

Gene	Forward	Reverse
AXIN2	5'-CAAACTTTCGCCAACCGTGGTTG	5'-GGTGCAAAGACATAGCCAGAACC
C1orf 43	5'-AGCTCTGGATGCCATTCGTACC	5'-GTGTTTCGCAGATCCAGCAGGT
CTNNB1	5'-CACAAGCAGAGTGCTGAAGGTG	5'-GATTCCTGAGAGTCCAAAGACAG
ETV4	5'-AGGAACAGACGGACTTCGCCTA	5'-CTGGGAATGGTCGCAGAGGTTT
RNF43	5'-GGTTACATCAGCATCGGACTTGC	5'-ATGCTGGCGAATGAGGTGGAGT
TLR4	5'-CCCTGAGGCATTTAGGCAGCTA	5'-AGGTAGAGAGGTGGCTTAGGCT

Results and Discussion

Patients and samples

For Nanostring pathway analysis, eleven samples of Stage II colon adenocarcinoma were selected (Supplementary Dataset File 1, sheet 2). Tumor samples were obtained from subjects with an age of onset from the late-30s through mid-70s. Eight cases were from the right/proximal colon (cecum and ascending colon), two from the transverse colon, and one from the sigmoid colon. All the tumors were $T_3N_0M_0$, according to the American Joint Committee on Cancer (AJCC) staging. Six samples from the right colon had moderate lymphocyte infiltration (TIL). DNA MMR was assessed by IHC reactivity to four standard MMR proteins (MLH1, MSH2, MSH6 & PMS2). These samples were confirmed by MSI-PCR using two mononucleotide repeat markers (BAT25, BAT26) and three quasi-monomorphic mononucleotide repeat markers (NR-21, NR-24, NR-27) (Buhard et al., 2006). Six tumor tissues had deficient MMR (MSI), and five had proficient MMR (MSS). For the validation trials, 15 cases of exclusive MSI early-stage CRC were selected. All 15 cases were identified as MSI by MMR IHC and MSI-PCR in the discovery cohort. Details are given in Supplementary Dataset File 1, sheet 3. The age of onset was 21-75 years, with four cases from the right-side colon, four from the hepatic flexure region, two from the transverse colon, two from splenic flexure, and three from the left colon. All the cases had TIL, except two cases.

Nanostring Expression Analysis

Sample annotation and Nanostring gene expression raw data are provided in Supplementary Dataset File S2. One of the samples (SH) yielded very less hybridization signal and was removed as an outlier. The principal component analysis (PCA) of the whole gene set is shown in figure 1a. Differential expression analysis and the significant genes are listed in Supplementary Dataset File S3. In the segregated samples, out of the 730 target signals, 155 genes had an FDR adj. *p*-value < 0.1, 119 genes had an FDR adj. *p*-value < 0.05 across the samples and 45 genes across the samples had an FDR adj. p-value < 0.01 (Figure 1b). The FC values of 45 top genes were clustered among tumor samples, and the expression pattern was reversed in the normal tissues (Figure 1c). Of these 45 genes, 28 had an absolute fold-change of > 2. ($|(FC(\log_2)| > 2$) and 22 genes were upregulated ($FC(\log_2) \geq 2$), while six genes were downregulated ($FC(\log_2) \leq -2$).

These genes are listed in Supplementary Dataset File S3 with their corresponding *p*-values and FC($\log_2$) values.

Nanostring Pathway analysis: Gene signals associated with cell cycle and apoptosis, chromatin modification, and DNA damage repair were upregulated in tumor samples and downregulated in normal tissues, with a few exceptions in specific cases (Figure 1d). The FC and overall pathway score for the chromatin modification set and DNA damage repair in different models did not correspond to their MMR status. This might be influenced by the tumor cells, TILs, and macrophages, which are further explored by the GEPIA and TIMER correlation analyses.

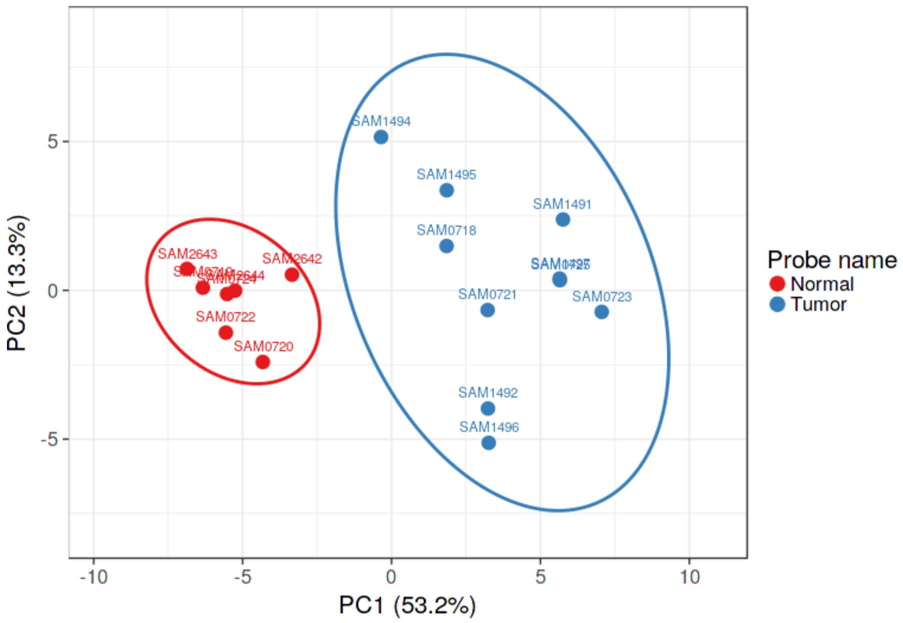


Figure 1. Nanostring Pan-cancer analysis. *1a: Principal component analysis (PCA) showing segregation of tumor and normal from the differential expression analysis.*

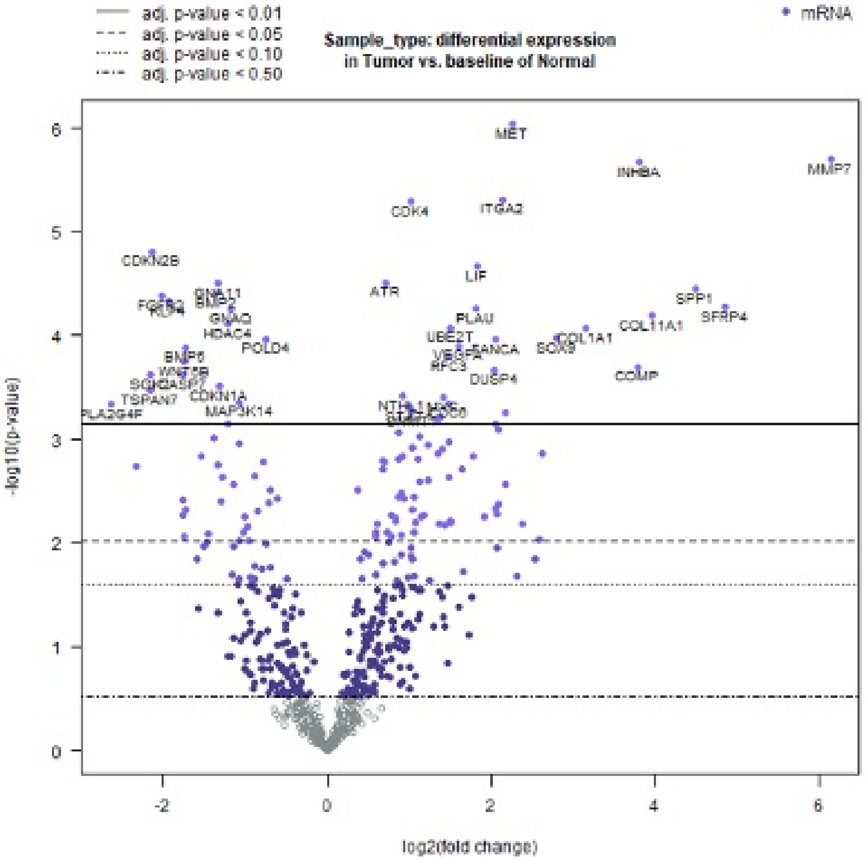


Figure 1. Nanostring Pan-cancer analysis. *1b: Significant genes obtained from the differential expression analysis, FDR adjusted p-values (-log10 scale) on the Y-axis and fold-change (log2 scale) on the X-axis.*

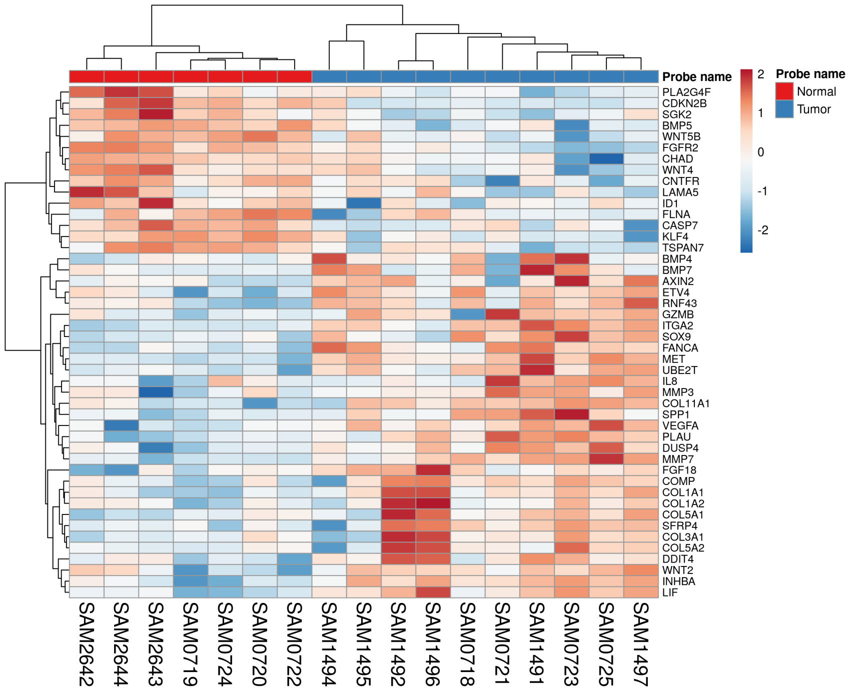


Figure 1. Nanostring Pan-cancer analysis. *1c: Heatmap showing the clustering of significant-top 45 genes from the DE analysis (Y-axis) among tumor and normal samples*

(X-axis).

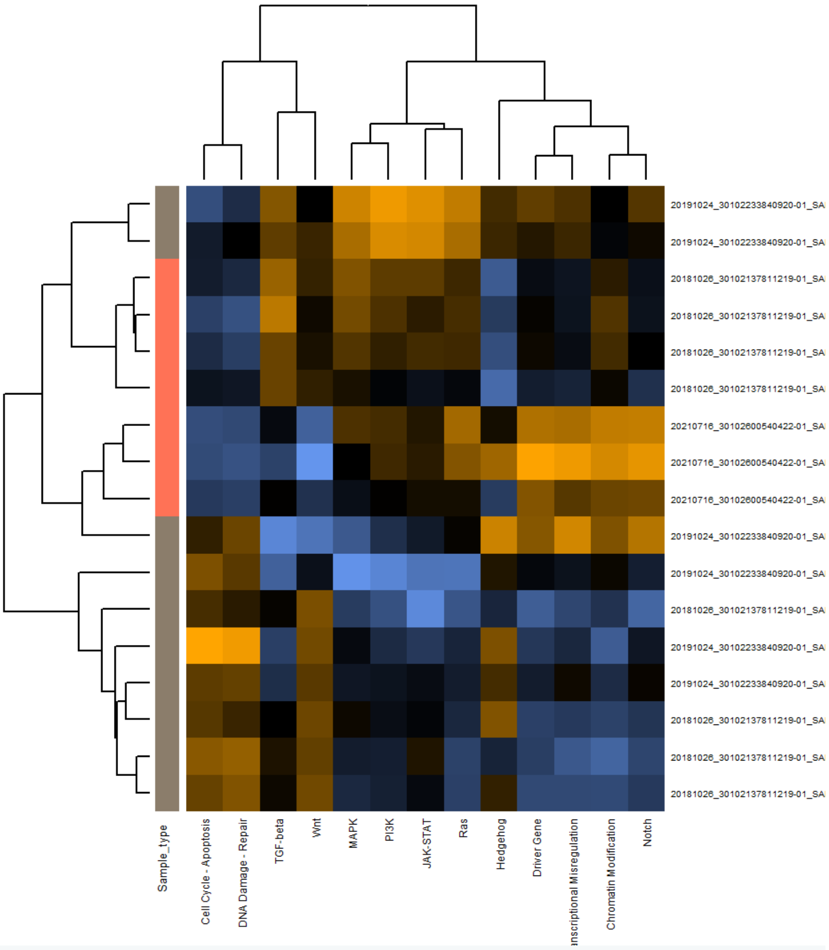


Figure 1. Nanostring Pan-cancer analysis. *1d: Heatmap showing pathway enrichment (X-axis) against different tumor samples (Y-axis).*

Group comparison

To compare the effects of microsatellite instability (MSI) status and tumor immune cell infiltration (TIL) on the expression fold of mRNA signals, the expression fold changes of all signals were compared separately against MSI status (MSS vs. MSI groups) and TIL status. The expression FC of 730 genes from the Nanostring DE analysis was compared between the MSI and MSS groups to reveal 16 genes that were significantly (p-value < 0.05) differentially regulated (Figure 2a). On FDR analysis, none of them were significant (adj. p-value < 0.05) and none of them had a mean $|FC(log_2)| > 1$. In comparison to TIL status, 20 genes were found to be significant at p-value < 0.05, but on FDR analysis failed to be significant. None of the genes had an $|FC(log_2)| > 1$. This shows that the fold changes in these genes are not significant enough to compare with one gene or group of genes directly. There was no significant difference in the gene expression pattern when compared to the age and gender of the patient or the anatomical location of the tumor. The genes that are significant in these two comparative analyses are given in Supplementary Data file S4.

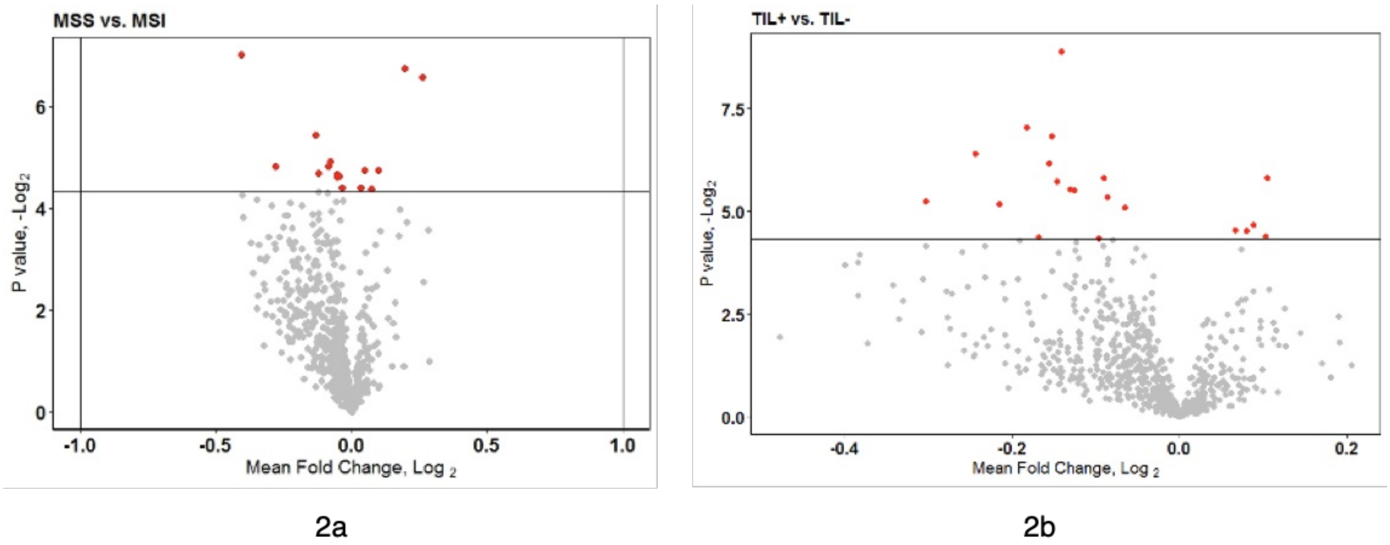


Figure 2. Scatter plots showing significant genes from the Nanostring DGE in MSI vs. MSS groups (2a) and the presence or absence of TIL (2b). In both the plots, X-axis is the foldchange in the log2 scale, and Y-axis is the p-value in the -log2 scale. The horizontal line in the middle of the field corresponds to the p-value = 0.05; dots above the horizontal line represent genes that have a p-value < 0.05 (red-colored), and dots below the horizontal line represent genes that have a p-value > 0.05 (grey-colored dots).

GEPIA Correlation analysis

The expression profile in the current study (AIMS) was compared against the TCGA COAD dataset from GEPIA. Out of 119 significant genes in the AIMS study, 73 genes are shared with the TCGA COAD dataset in the GEPIA database (See Figure 3 and Supplementary Dataset File S5). All the common genes were found to be significant in both datasets. The most significant gene in the TCGA from the AIMS dataset is ETV4; in the AIMS study, MET was highly significant. In the analysis of the top 19 of the 73 signals with |FC(log$_2$)| >2, 16 genes were upregulated, and three genes were downregulated. Comparing the adj. *p*-values between the current study vs. TCA, the correlation was -0.083 (Pearson method, Figure 4b). On the contrary, when the Fold-change was compared between these two studies, the correlation coefficient was 0.77 (See figure 4c). Genes with |FC(log$_2$)| > 2 were significant (adj. *p*-value < 0.05) in both AIMS study groups as well as TCGA group

were *MMP7*, *SFRP4*, *SPP1*, *COL11A1*, *INHBA*, *COMP*, *SOX9*, *ETV4*, *MET*, *ITGA2*, *RNF43*, *FANCA*, *DUSP4*, *FGFR2*, *TSPAN7*, *PLA2G4F*, and *MMP7* were found to have |FC(log$_2$)| >5 in the TCGA cohort. Both data's common genes showed a similar expression pattern (upregulation/downregulation), while two genes (*PLA2G4F* and *TMPRSS2*) had a reversal in FC values. The GEPIA comparison shows that the current study represents the TCGA data overall.

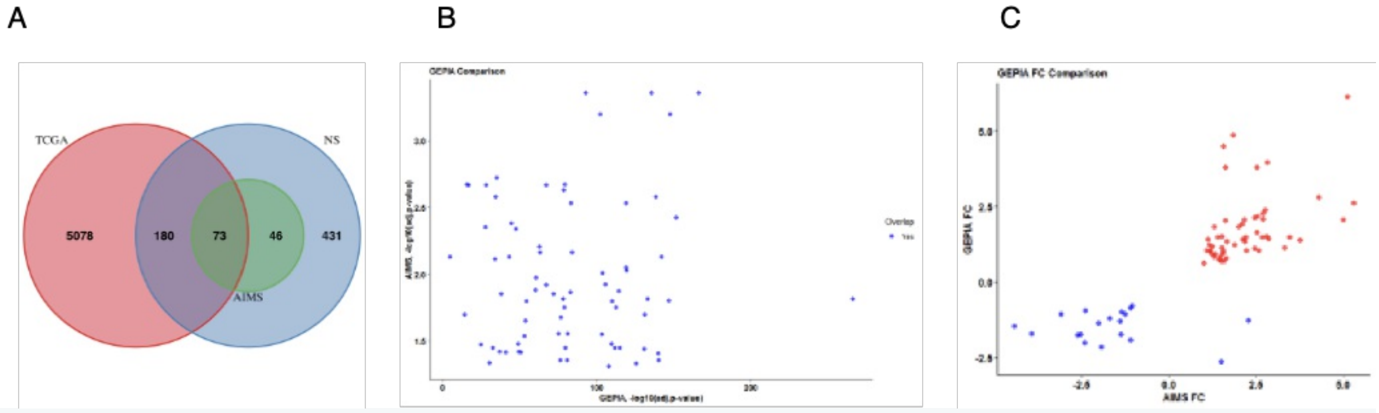


Figure 3. Comparison of study with the TCGA study. A: The Venn diagram showing the number of genes from GEPIA (as “TCGA” in red), Nanostring geneset (“NS” in blue), and significant genes in the current study (“AIMS” in green). B: Scatter plot of the significant genes (adj p-value < 0.05) in AIMS data overlapped with the p-value of the identical genes in TCGA data. Pearson correlation coefficient was -0.083. C: Scatter plot of fold-change of expression of significant genes in AIMS data to that of the TCGA data obtained from GEPIA. The Pearson correlation coefficient was 0.77. The red dots are upregulated genes, and the blue dots are downregulated genes.

Protein-Protein Interaction Network Analysis

As illustrated in Figure 4, the top 20 genes interact with each other according to their respective protein-protein interaction, gene co-occurrence, and gene-neighborhood in STRING.db. A detailed list of interactions is given in Supplementary Dataset File S6. Among these, experimentally proven interactions were projected to be a cluster containing 15 core proteins. These were COL1A1, COL3A1, COL1A2, COL11A1, COMP, SPP1, MMP7, MMP3, ITGA2, MET, SOX9, BMP7, MET, FGF18 CXCL8, and distantly interacting proteins were ETV4, SFRP4, INHBA, and GZMB. Two proteins were found to have no interaction with the above proteins, RNF43 and DDIT4. Protein homology was found between COL1A1, COL3A1, COL11A1, and SFRP4 has homology-based interaction. The correlation cluster is seen with the TIMER TCGA database, and the genes clustered in the STRING.db network analysis corresponded to many genes with each other.

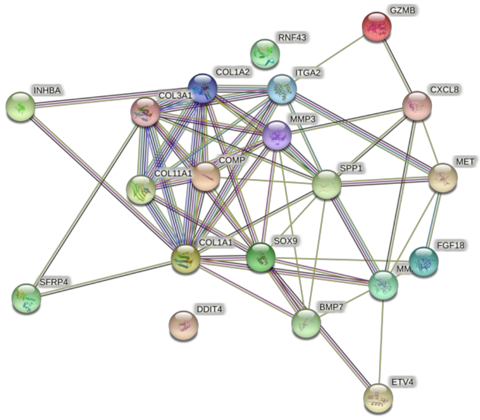


Figure 4. Protein-Protein Interaction Network analysis done in STRING.db. Color codes of interaction lines represent gene neighborhood (green), gene fusions (red), gene co-occurrence (dark blue), from curated databases (teal), experimentally determined (pink), text mining (yellow), co-expression (black), protein homology (light blue).

Association of immune cell infiltration with gene expression

As shown in Figure 5, hierarchical clustering analysis of the correlation among the top 30 from significantly upregulated genes identified by the Nanostring analysis (adj. p-value < 0.05) in the current study, with the immune cell infiltration data obtained from TIMER (Supplementary Dataset File S7). All the genes, except four, were associated with the immune cell infiltrates, Dendritic cells, Macrophages, Neutrophils, CD4, CD8 T, and B lymphocytes. *AXIN2*, *E2F1*, *ETV4*, and *RNF43* were clustered with tumor purity and away from all other infiltrating cell types. These gene signals' FC(log$_2$) were, respectively, 1.64, 1.05, 2.62, and 2.06. These gene signals correlate more to the tumor cells than any other cell type in the TIMER database.

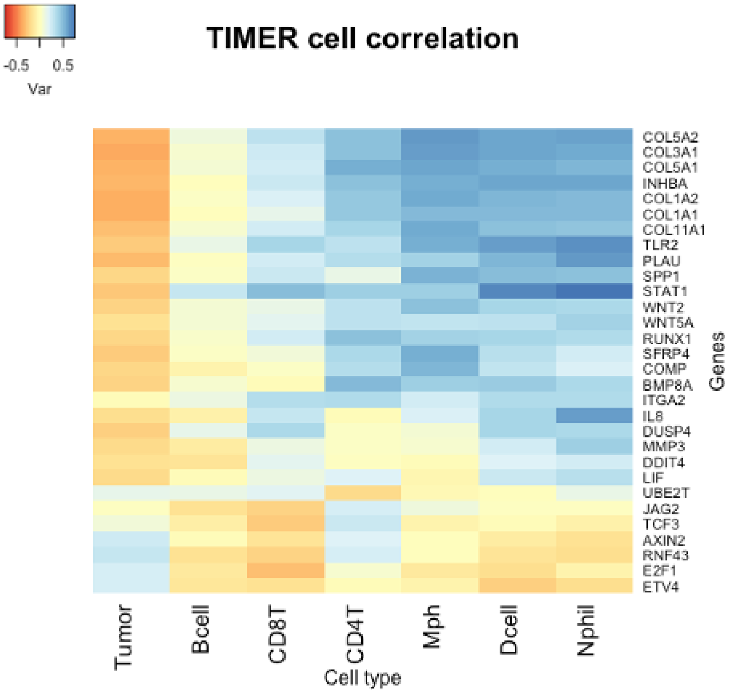


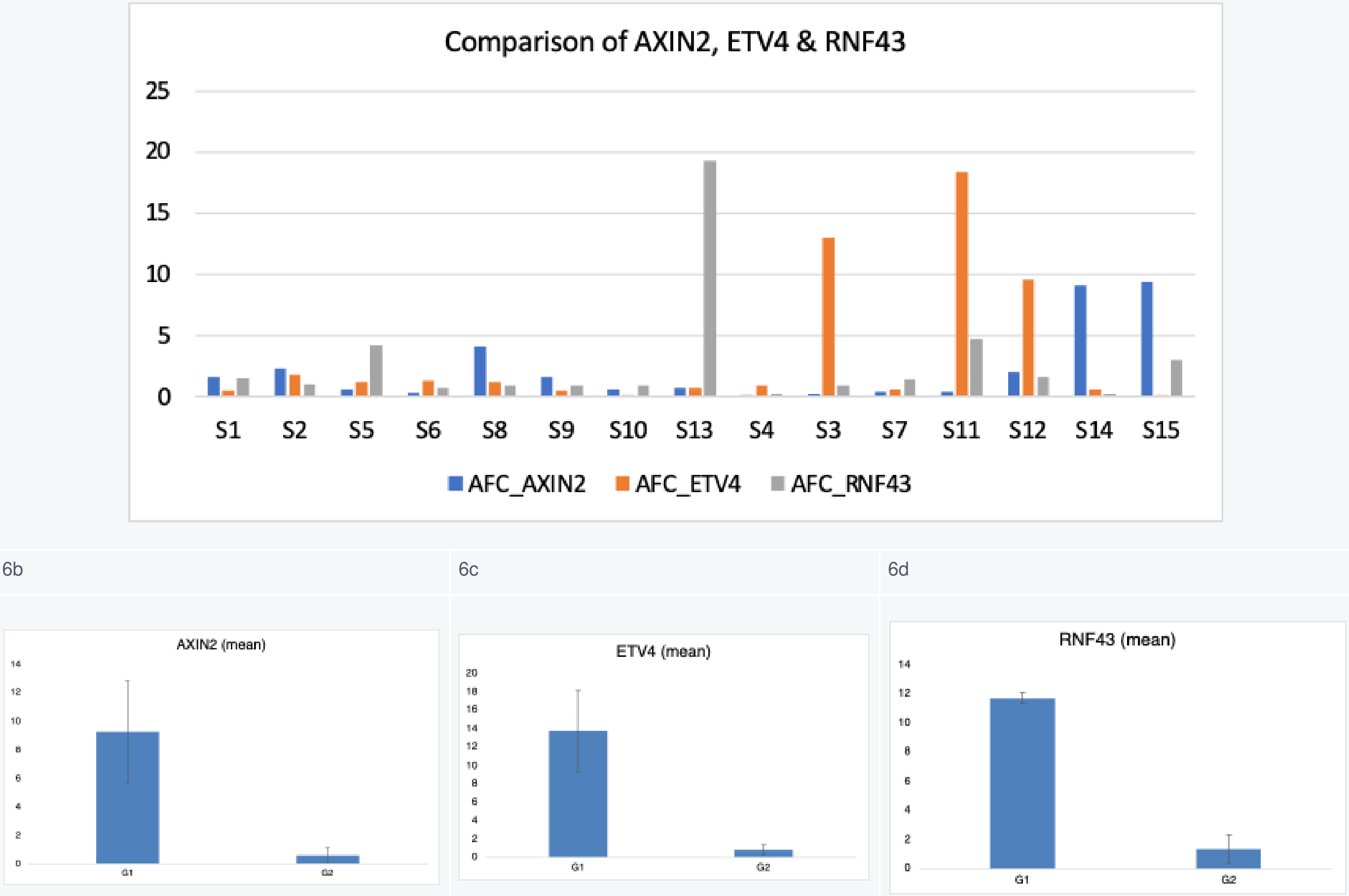
Figure 5. HCA correlation of top 30 significant genes identified in the Nanostring analysis against immune cell infiltration. Clustering based on a negative ordering of the variance of Partial correlation coefficients (ρ) of each of the top 30 significant genes from current study to the infiltrating cell type in the TCGA COAD data in TIMER. The x-axis contains different immune cell types and tumor purity. The more correlation to tumor purity means less correlated to infiltrating cell types. Bcell: B lymphocyte. CD8T: CD-8 T lymphocytes, CD4T: CD4-T lymphocytes, Mph: Macrophages, Dcell: Dendritic cells, Nphil: Neutrophil leukocytes.

Validation of gene expression

To validate the top genes correlated with the tumor cells, another set of early-stage microsatellite unstable colon cancer paired tumor-normal tissue samples (n = 15) was taken for an RT-PCR. An overall comparison shows broad variation in expression levels of these genes across different MSI groups. As given in the bar diagram in figure 6, there were distinct groupings in terms of expression foldchange using RT-PCR. After grouping the samples into G1 (high FC) & G2 (low FC) according to the AFC expression levels of *ETV4*, the group's expression difference was found to be very significant, p-value = 4.99067E-08 on the *Student t-test*. Three samples were in G1, and 12 samples were grouped into G2. After grouping the samples according to the expression fold-change of *AXIN2*, the difference in expression between the groups was found to be significant, p-value = 3.26391E-06 on the *Student t-test*. Five samples were grouped into G1, and 10 cases were grouped into G2. When grouping these samples according to the expression of *RNF43*, two cases had high expression (G1), and 13 samples had lower expression (G2). The expression fold-change was significantly different between these two groups, p-value = 2.89572E-09, on the *Student t-test*.

On PCA, comparing ΔCt of tumor vs. control, overall variations overlap (Figure 7a). However, the*RNF43* and *ETV4* had positive correlations to PC1 and PC2. *AXIN2* has a positive correlation with PC1 and a negative to PC2. At the same time, *CTNNB1* has an opposite trend to *AXIN2*. At the same time, *TLR4* has a negative trend in both the principal components suggesting that the variance differs from all the other targets. Hierarchical clustering analysis shows *TLR4* clustered differentially than all the other signals. Figure 7 shows the variance in gene signals RFC for the validation set of samples. The PCA shows that the *TLR4* expression differs from other targets' expressions. Moreover, the variance of *TLR4* expression is seen only in a subset of MSI CRC samples.

Figure 6. Comparison of expression by RT-PCR validation



6a

6b

6c

6d

Figure 6. Foldchange expression of genes that were selected for validation. 6a shows the comparative expression of the three genes, AXIN2, ETV4 and RNF43. The segregation of those samples in terms of the expression was done for the three genes in Figures 6b, 6c, and 6d. The bars graphs shown are mean ± sd.

Figure 7. Expression foldchange comparison

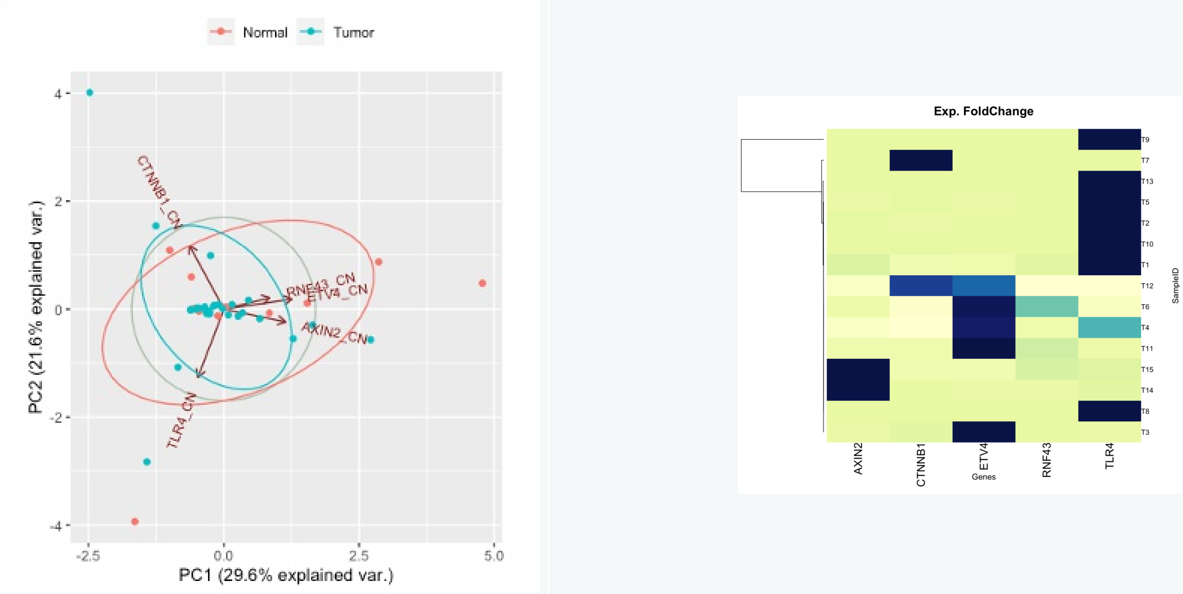


Figure 7a. PCA of individuals compared against their expression levels. Note that the correlation trends are distinctly diverse for the three genes (*AXIN2*, *ETV4*, *RNF43*), *TLR4*, and *CTNNB1*.

Figure 7b. The hierarchical clustering of the variance of each sample across different target genes is depicted in a heatmap. The cluster dendrogram of samples is given on the side. Heatmap shows a differential expression of TLR4 compared to the other signals in the validation set.

Conclusion

In the current study, we have identified distinct categorizations of early-stage MSI CRC based on the expression of three genes *AXIN2*, *ETV4*, and *RNF43*. *AXIN2* and *RNF43* are the negative regulators of the Wnt pathway, and their upregulation suggests a Wnt ligand-independent pathogenesis (Kleeman and Leedham, 2020). However, the expression of *ETV4*, which has not been previously known to be involved in the Wnt pathway, suggests its possible role in Wnt-ligand independent MSI CRC. The expression of *CTNNB1* is unrelated to the expression of the three genes, suggesting an overlapping role of β-Catenin in MSI CRC. On the other hand, the contrasting expression of *TLR4* to that of trios may indicate that the tumor immune reaction could be segregated based on the expression of these genes. The subgroup with *TLR4* upregulation may contribute to a better prognosis for the MSI pathogenesis of early-stage CRC. Further studies on these three genes in a larger number of CRC might be required to understand the delineation of these sub-groups in the molecular progression of MSI CRC and how this expression categorization would impact the treatment and prognosis outcomes.

References

- Ariyannur, P. S., et al., 2021. Pilot Nanostring PanCancer pathway analysis of colon adenocarcinoma in a tertiary healthcare centre in Kerala, India. Ecancermedicalscience. 15, 1302.
- Buhard, O., et al., 2006. Multipopulation Analysis of Polymorphisms in Five Mononucleotide Repeats Used to

Determine the Microsatellite Instability Status of Human Tumors. Journal of Clinical Oncology. 24, 241-251.

- Dumortier, M., et al., 2018. ETV4 transcription factor and MMP13 metalloprotease are interplaying actors of breast tumorigenesis. Breast Cancer Res. 20, 73.
- Dunne, P. D., et al., 2016. Challenging the Cancer Molecular Stratification Dogma: Intratumoral Heterogeneity Undermines Consensus Molecular Subtypes and Potential Diagnostic Value in Colorectal Cancer. Clin Cancer Res. 22, 4095-104.
- Hause, R. J., et al., 2016. Classification and characterization of microsatellite instability across 18 cancer types. Nat Med. 22, 1342-1350.
- Janitschke, D., et al., 2020. Unique Role of Caffeine Compared to Other Methylxanthines (Theobromine, Theophylline, Pentoxifylline, Propentofylline) in Regulation of AD Relevant Genes in Neuroblastoma SH-SY5Y Wild Type Cells. Int J Mol Sci. 21.
- Kanth, V. V., et al., 2014. Microsatellite instability and promoter hypermethylation in colorectal cancer in India. Tumour Biol. 35, 4347-55.
- Kleeman, S. O., Leedham, S. J., 2020. Not All Wnt Activation Is Equal: Ligand-Dependent versus Ligand-Independent Wnt Activation in Colorectal Cancer. Cancers (Basel). 12, 3355.
- Leary, R. J., et al., 2008. Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers. Proc Natl Acad Sci U S A. 105, 16224-9.
- Li, T., et al., 2017. TIMER: A Web Server for Comprehensive Analysis of Tumor-Infiltrating Immune Cells. Cancer Res. 77, e108-e110.
- Li, T., et al., 2020. TIMER2.0 for analysis of tumor-infiltrating immune cells. Nucleic Acids Res. 48,W509-w514.
- Lin, J. K., et al., 2003. Loss of heterozygosity and DNA aneuploidy in colorectal adenocarcinoma. Annals of Surgical Oncology. 10, 1086-1094.
- NCRP, Report on 27 PBCRs in India. In: NCDIR, (Ed.), Three Year Report of Population Based Cancer Registeries 2012-2014. NCDIR-NCRP, ICMR, Bangalore, 2016, pp. 9-26.
- Pandey, V., et al., 2007. Assessment of microsatellite instability in colorectal carcinoma at an Indian center. Int J Colorectal Dis. 22, 777-82.
- Pangestu, N. S., et al., 2021. RNF43 overexpression attenuates the Wnt/beta-catenin signalling pathway to suppress tumour progression in cholangiocarcinoma. Oncol Lett. 22, 846.
- Rajkumar, T., et al., 2004. Mutation analysis of hMSH2 and hMLH1 in colorectal cancer patients in India. Genet Test. 8, 157-62.
- Raman, R., et al., 2014. Evidence for possible non-canonical pathway(s) driven early-onset colorectal cancer in India. Mol Carcinog. 53 Suppl 1, E181-6.
- Shakuntala, S. T., et al., 2022. Descriptive Epidemiology of Gastrointestinal Cancers: Results from National Cancer Registry Programme, India. Asian Pac J Cancer Prev. 23, 409-418.
- Sharma, R., et al., 2022. Global, regional, and national burden of colorectal cancer and its risk factors, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. The Lancet Gastroenterology & Hepatology. 7, 627-647.

- Szklarczyk, D., et al., 2019. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 47, D607-D613.
- Tang, Z., et al., 2019. GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. Nucleic Acids Res. 47, W556-W560.
- TCGA, 2012. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 487, 330-7.
- Wang, J., et al., 2017. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. Nucleic Acids Res. 45, W130-W137.
- Xu, L., et al., 2019. Novel reference genes in colorectal cancer identify a distinct subset of high stage tumors and their associated histologically normal colonic tissues. BMC Med Genet. 20, 138.
- Yang, H., et al., 2010. Reduced expression of Toll-like receptor 4 inhibits human breast cancer cells proliferation and inflammatory cytokines secretion. J Exp Clin Cancer Res. 29, 92.
- Yeole, B. B., 2008. Trends in cancer incidence in esophagus, stomach, colon, rectum and liver in males in India. Asian Pac J Cancer Prev. 9, 97-100.