Title: Expression analysis of colon adenocarcinoma in the South Indian population reveals novel signals of pathogenesis

Authors:

Prasanth S. Ariyannur^{1,2*}, Reenu Anne Joy¹, Veena Menon², Roopa Rachael Paulose³, Keechilat Pavithran⁴, Damodaran M. Vasudevan^{1,2,5}.

Affiliation:

¹. Department of Biochemistry, Amrita Institute of Medical Sciences and Research Center, Amrita Vishwa Vidyapeetham, Kochi 682041

². Department of Molecular Biology, Amrita Institute of Medical Sciences and Research Center, Amrita Vishwa Vidyapeetham, Kochi 682041

³. Department of Pathology, Amrita Institute of Medical Sciences and Research Center, Amrita Vishwa Vidyapeetham, Kochi 682041

⁴. Department of Medical Oncology, Amrita Institute of Medical Sciences and Research Center, Amrita Vishwa Vidyapeetham, Kochi 682041

⁵. Department of Health Sciences Research, Amrita Institute of Medical Sciences and Research Center, Amrita Vishwa Vidyapeetham, Kochi 682041.

***Corresponding author:**

Dr. Prasanth Ariyannur
Department of Biochemistry
Amrita Institute of Medical Sciences
Amrita Vishwa Vidyapeetham
Kochi India 682041
Ph: (91) 484 6688016
Fax: (91) 484 2852020
Email: prasanthas@aims.amrita.edu

No. of Words: 3353

No. of Tables: 1

No. of Figures: 7

Keywords: colorectal cancer, Nanostring, TCGA, GEPIA, TIMER, signaling pathway

Abstract

High throughput somatic expression analyses of colon adenocarcinoma conducted so far were mostly in the western population, and no major studies are currently available in the Indian population. We have performed Nanostring PanCancer pathway panel assay in Stage II colon cancer (n = 11) and compared against normal colon tissues (n = 4). Differentially expressed (DE) genes were identified and superimposed on the Cancer Genome Atlas (TCGA) data, from the Genome Expression Profiling Interactive Analysis (GEPIA) and Tumor Immune Estimation Resource (TIMER). 83 out of 730 genes were significantly different (p -value < 0.01), 19 of which had a fold-change $|FC(\log_2)| \geq 2$. A comparison of these signals on TCGA COAD data revealed four common (MET, MCM2, ETV4, and MMP7) and 15 uncommon genes. On group comparison, ETV4 expression was significantly higher in microsatellite stable (MSS). Significant DE genes, unique in the study, were INHBA, COL1A1, COL11A1, COMP, SFRP4, and SPP1, which were clustered in STRING.db network analysis and correlated with tumor-infiltrating immune cells in TIMER. MMR functions of colon cancer pathogenesis in India may be differentiated by validating the somatic expression of most common genes identified in the study.

Introduction

More than thirty years of research on molecular pathophysiology of colorectal cancer (CRC) has led to the identification of many novel mechanisms of tumorigenesis. Based on whole-genome sequencing studies, colon adenocarcinoma (COAD) has the highest number of overall general genomic alterations per tumor compared to any other tumor tissue types, irrespective of the stage of the disease¹. This was explained by the rapid self-renewing nature of normal colonic epithelial cells, the overall additive effect of a series of small driver mutations among the background of passenger mutations creating a survival advantage. In the western population, studies have shown that overall pathogenicity of CRC is primarily due to chromosomal instability (CIN) leading to widespread loss of heterozygosity (LOH) and gross chromosomal rearrangements and high somatic copy number variations^{2,3}. Large cohort studies done in the western population attributed the occurrence of CRC primarily to sporadic multi-molecular genetic changes (70-80%), causing chromosome-wide changes⁴, which led to the classification of CRC into distinct molecular subtypes^{5,6}. Defective DNA mismatch repair (MMR) and consequent microsatellite instability (MSI), classified as hypermutated type in The Cancer Genome Atlas (TCGA), are the second common cause, accounting for roughly one-fifth of sporadic molecular genetic changes⁵. Less than 5% of CRC was found to be due to delineable hereditary molecular factors, mainly autosomal dominant inherited multi-organ cancer disorder, Lynch Syndrome (LS), due to constitutional and inherited mutations in four MMR genes-*MLH1*, *MSH2*, *MSH6*, *PMS2*, and *EPCAM*⁷. Deficient MMR comprises approximately 15% of the CRC pathogenesis, including both sporadic and hereditary causes in the western population⁸.

The complexity of MMR function due to the involvement of more than 30 genes⁹, epigenetic inactivation¹⁰, microRNA mediated regulation of MMR gene transcripts¹¹, and hetero-oligomeric co-dependence of MMR proteins¹², leading to variable phenotypic penetrance are some of the very few known factors for diagnostic and scientific challenges of this syndrome. The exorbitant diagnostic cost and subsequent lack of compliance for repeated investigations are explained partially by these challenges. However, multiple levels of diagnostic maneuvers to circumvent these issues led to a complex, expensive, indirect, and often inadequate clinical conclusion. All of these concerns suggest the lack of a better diagnostic algorithm and perhaps better biomarker(s) suited for different population create a great deal of diagnostic lacuna in the local population.

The prevalence rate of colon cancer in India, according to a recent population-based cancer registry by the Indian Council for Medical Research (ICMR), is on the rise. Average prevalence rates of 4.3 and 3.4 in 100,000 males and females respectively, across different parts of India, are further increasing, as per the recent epidemiological reports¹³. The International Association of Cancer Registries (IACR) Global Cancer Observatory (GLOBOCAN) showed an incidence of 5.8 - 9 per 100,000, with a slightly higher rate of mortality in men than women (4.1 - 5.8 vs 3.5 - 5.3). Large studies on DNA microsatellites in different cancer types, by TCGA, did not include representative people from major Indian ethnic groups¹⁴. Pioneer studies, conducted by a center in Tamil Nadu, on the Indian CRC showed that 67% of cases had deficient MMR¹⁵. Another study done in Karnataka showed a prevalence of about 10%¹⁶. Irrespective of the pathways involved in the pathogenesis of early or late-onset CRC, MSI was found to be around 40% in another study conducted in Andhra Pradesh¹⁷. More recent studies from the same region showed 48% of cases were MSI¹⁸. There have been not many studies conducted in India to address the reason behind the high prevalence of MMR deficiency occurring in India compared to other populations. Furthermore, by immunohistochemical study conducted in the our population, during the last eight years, identified deficient MMR in about 27% cases of Stage II CRC⁹. However, there has not been any molecular pathogenesis study to explore the reasons behind higher prevalence of MMR in CRC pathogenesis in Indian population.

To address this, we selected a set of Stage II CRC from a tertiary cancer care center, including tumors from both microsatellite stable (MSS) and unstable categories, and compared the histopathological data against a high throughput expression analysis using Nanostring nCounter Pan-Cancer pathway analysis. The expression pattern and the signals obtained from this cohort were compared against the TCGA COAD data using multiple data tools such as Genome Expression Profiling Interactive Analysis (GEPIA) and Tumor Immune Estimation Resource (TIMER) to explore the unique expression signals in CRC pathogenesis in our population.

Results

Samples and screening.

Eleven samples of Stage II colon adenocarcinoma were selected. Tumor samples were obtained from subjects with an age of onset of 38 through 76 years (See Table 1 for more details). Eight samples were from the proximal colon (ileocecal and ascending colon), two from transverse colon and one from the sigmoid colon. The tumor was previously examined by histopathology for their tumor type, differentiation, the extent of tumor invasion, lymphocyte, and associated immune cell infiltration, lympho-vascular, and perineural invasion. DNA MMR was assessed by immunohistochemical reactivity of proteins MLH1, MSH2, MSH6, and PMS2. Additionally, these samples were confirmed by MSI-PCR using two mononucleotide repeat markers (BAT25, BAT26), and three quasi-monomorphic repeat markers (NR-21, NR-24, NR-27), all of which have a high correlation with low immunoreactivity to any/all four MMR protein¹⁹. Six tumor tissues had deficient MMR (both IHC and MSI-PCR), and five had proficient MMR. Five proximal colon tumors were of the MSI category. All the tumors were T3N0Mx according to the Union for International Cancer Control (UICC) staging. Six samples had moderate lymphocyte infiltration, all originating from the proximal colon region. One proximal colon tumor had severe infiltration, while three proximal and one distal had mild or no infiltration.

Table 1: Patient and sample details

ID	Age/ Sex	Location	MMR IHC	MSI - PCR	Lymphocyte Infiltration	UICC TNM group	Stage
KSN	38/F	Sig. colon	Normal	MSS	ND*	T3N0Mx, LVI+, PNI-	IIB/IIIA
SIV	47/M	Ileo-cecum	Normal	MSS	Severe	T3N0Mx, LVI-, PNI-	IIB
KSM	52/M	Asc. colon	Deficient MLH1 & PMS2	MSI-H	Moderate	T3N0Mx LVI+, PNI-,	IIB/IIIA
JCB	50/M	Asc. colon	Deficient MSH2 & MSH6	MSI-H	Moderate	T3N0Mx, LVI+, PNI-,	IIB/IIIA
AN	46/M	Asc. Colon	Deficient MSH2	MSI-H	Moderate	T3N0Mx, LVI+, PNI-,	IIB/IIIA
ES	58/F	Tr. Colon	Normal	MSS	Moderate	T3N0Mx, LVI+, PNI-,	IIB/IIIA
SH	50/M	Asc. Colon	Deficient MLH1 & PMS2	MSI-H	Moderate	T3N0Mx, LVI+, PNI-,	IIB/IIIA
SK	54/M	Asc. Colon	Normal	MSS	ND	T3N0Mx, LVI+, PNI-,	IIB/IIIA
KNR	76/F	Tr. Colon	Deficient MLH1 & PMS2	MSI-H	Mild	T3N0Mx, LVI+, PNI-,	IIB/IIIA
VVC	67/M	Asc. Colon	Deficient MSH2 & MSH6	MSI-H	ND	T3N0Mx, LVI-, PNI-,	IIB/IIIA
RV	67/F	Asc. Colon	Normal	MSS	Moderate	T3N0M0	IIB

* Not Detected.

Nanostring Expression Analysis

The sample annotation and gene expression data are provided in Supplementary Dataset File S1 and S2. Out of the 730 target signals, 83 were found to be significant (p -value < 0.01 ; FDR adj. p -value < 0.5) across the samples (Figure 1a). Of the 83 genes, 17 genes were upregulated, ($\log_2$ fold-change – $(FC(\log_2)) > 2$) and two 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. On the principal component analysis (PCA), 20 genes showed segregation between tumor and control samples (Figure 1b). FC of the top 20 genes in PCA is depicted in a heatmap in Figure 1c. The 20 top genes were clustered evenly among tumor samples, and the expression pattern was reversed in the normal samples. With the Nanostring pathway analysis, signals associated with cell cycle and apoptosis, chromatin modification, and DNA damage repair were shown to be upregulated in tumor samples and downregulated in normal tissues, with few exceptions in certain samples (Figure 1d). The overall pathway score for the chromatin modification set and DNA damage repair in different samples did not correspond to their MMR status. This may be because the overall pathway effect of the whole tissue might be influenced by the tumor cells as well as Tumor-Infiltrating Lymphocytes (TILs) and macrophages, which is further understood by the GEPIA and TIMER correlation analyses.

Figure 1:

1a:

1b:

1c:

1d:

Figure 1: Differential expression analysis: 1a:Volcano plot displaying the expression pattern of the entire gene set. Y-axis is $-\log_{10}(p\text{-value})$ and X-axis is $\log_2(\text{fold change})$ of each of the covariates. Significant genes fall at the top of the plot above the horizontal lines, and highly differentially expressed genes fall to either side. Horizontal lines indicate various False Discovery Rate (FDR) thresholds or p -value thresholds if there is no adjustment to the p -values. Genes are colored if the resulting p -value is below the given FDR or p -value threshold. Genes that had $p\text{-value} < 0.01$ (FDR adj. $p\text{-value} < 0.5$) are 83 and among them, the 40 most statistically significant genes are labeled in the plot. **1b:** Venn diagram showing the PCA of the top 20 genes of the samples in the given cohort from the ratio data estimation. **1c:** Heatmap showing the top 20 upregulated or down-regulated genes from the differential expression analysis of the Nanostring data. All these signals have a p -value of < 0.01 and a $\log_2$ fold change of at least 1. The normal and tumor samples are colored coded in the top margin. **1d:** Heatmap showing pathway score depicting an overview of how the pattern of pathway scores change across samples to understand how pathway scores cluster together and which samples exhibit similar pathway score profiles. Orange indicates high scores; blue indicates low scores. Scores are displayed on the same scale via a Z-transformation.

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 change of all signals were compared separately against MSI status (MSS vs. MSI groups) and TIL status (detected or not in FFPE sections)

using the Student t-test. The results are shown in Figures 2a & 3. The expression FC of 730 genes from the Nanostring DE analysis was compared against MSI and MSS groups to reveal 13 genes that were significantly different, p -value < 0.05 . Among these, four genes had a mean $|\text{FC}(\log_2)| > 1$ between the two groups (Figure 2a). Of the four, highest differential expression was shown by ETV4 (mean $\text{FC}(\log_2) = 12.38$; p -value = 0.035) in MSS group compared to MSI (Figure 2b). Other significant genes with $\text{FC} > 1$ were PLCB4 (mean $\text{FC}(\log_2) = 2.43$; p -value = 0.014), PROM1 (mean $\text{FC}(\log_2) = 1.39$; p -value = 0.025) and BIRC3 (mean $\text{FC}(\log_2) = -1.13$; p -value = 0.049). All these genes were significantly different from similar FC in expression in the TCGA as well (as shown in GEPIA analysis below). When signals were compared against groups with the two groups of TIL, the only signal that was most significant with the highest mean FC was ETV4 (mean $\text{FC}(\log_2) = 12.82$; p -value = 0.035) (Figure 3a & 3b). Other signals that were significant between the groups, but $\text{FC}(\log_2) \approx 1$ were BCL2A1 (mean $\text{FC}(\log_2) = -1.061$, p -value = 0.018), CDC7 (mean $\text{FC}(\log_2) = -1.42$, p -value = 0.016), TNFRSF10A (mean $\text{FC}(\log_2) = -1.14$, p -value = 0.013). Notably, ETV4 expression fold is significantly higher in TIL ND group. In both these group comparisons, ETV4 was found to be the only signal that is conspicuously higher in the MSS group with no TIL. FDR adjustment resulted in no significant difference in expression in any gene signals across either of these groups. There was no significant difference in the gene expression pattern when compared against age, sex, or anatomical location of the tumor.

Figure 2

Figure 2: Comparison of microsatellite instability status in gene expression patterns. 2a: Group comparison of expression fold change in MSI and MSS group using t-test. The X-axis shows the mean difference between the two groups for each of the signals and the y-axis shows the $-\log_2(p\text{-value})$. Horizontal intercept above 4 in the Y-axis is the line of significance, signals above which have a p -value < 0.05 . The vertical line with X-intercept is at a $\log_2(\text{FC})$ of one. There were only three signals with $|\log_2(\text{FC})| > 1$ and p -value < 0.05 . Four genes satisfied p -value < 0.05 and $\text{FC}(\log_2) > 1$ - ETV4, PLCB4, and PROM1 (marked by red dot). **2b:** This figure shows the expression fold change in highly significant genes identified in figure 2a. Among them, ETV4 has the highest expression in the MSS group compared to MSI.

Figure 3

Figure 3: Comparison of tumor immune cell infiltration status in gene expression pattern. 3a: Group comparison of expression fold change in TIL vs. none using t-test. Axis annotations and intercepts are the same as that of figure 2a. Four signals with $|\log_2(\text{FC})| > 1$ and $p\text{-value} < 0.05$ were ETV4, TNFRSF10A, CDC7, and BCL2A1 (marked by red dots). ETV4 is the common signal in both these categories with maximum fold change. **3b:** Boxplot depicting the expression fold change in the four genes identified in 4a. Among them, ETV4 is associated with high expression in the TIL ND group.

TCGA COAD Correlation Analysis

GEPIA

The expression profile in the current study (AIMS) was compared against the TCGA COAD dataset on GEPIA. 45 out of 83 significant genes in the AIMS study were correlated with the TCGA COAD dataset (common set) in the GEPIA database (See Figure 4 and Supplementary Dataset File S3). However, 38 genes were found to be not significantly different in the TCGA dataset (unique set). The most significant gene in the TCGA from the current dataset is ETV4, while overall in the AIMS study, HDAC4 was highly significant (see volcano plot 1b). Analysis of the top 19 of the 83 signals with $|\text{FC}(\log_2)| > 2$ revealed 14 genes were significant and five genes were from the unique set. Furthermore, eight genes with $|\text{FC}(\log_2)| > 2$, were significant ($p\text{-value} < 0.01$) in both AIMS study groups as well as TCGA group (common set). These genes were MMP7, COL11A1, INHBA, ETV4, RNF43, MET, GZMB, and FLNA. Among them, MMP7, ETV4, and RNF43 were found to have $|\text{FC}(\log_2)| > 5$ in the TCGA cohort (See Supplementary Dataset File S3). Several genes that were significant in the TCGA were found to be either not significant or had $\text{FC}(\log_2) < 1$, in the current study. These were CCNB1, CDC6, PKMYT1, UBE2T, AINX2, ITGA2, RFC3, FANCA, MYC, IL6, and FGFR2. Signals that showed similar FC expression in both these studies were PLAUI, EZH2, TNFRSF10A, BRCA2, DUSP4, MCM7, PRKDC, and RIN1. One signal, HRAS, showed an upregulation, $\text{FC}(\log_2) = 1.01$ in our study and down-regulation in TCGA GEPIA database, $\text{FC}(\log_2) = -1.407$. Significantly down-regulated signals in

both the studies were FLNA, KLF4, FGFR2, IL6R, TSPAN7, WNT5B, HDAC4, CDKN2B, and COL4A5.

Figure 4

Figure 4: Comparison of 83 significant genes in the current study correlated with TCGA COAD expression data analyzed by GEPIA. The X-axis shows the p -values ($-\log_{10}$) of the current study and the Y-axis depicts the p -values ($-\log_{10}$) from the TCGA COAD data. 45 out of 83 genes had a significant difference in TCGA, while 38 genes were not significant in the TCGA data (as zero values in the graph).

TIMER

The list of highly significant 83 genes obtained from the Nanostring was fed into the TIMER TCGA COAD correlation module. Correlation module showed expression scatter plots between a pair of genes from the list in TCGA COAD cancer type with Spearman's rho value with statistical significance (Figure 5a). Partial correlation with age and tumor purity was also analyzed in the given set of gene samples. Correlation data is given in Supplementary Dataset file S4. As shown in the correlation matrix, Figure 5a, from TIMER and the chord diagram, Figure 5b, expressions of COL1A1, COL11A1, COMP, INHBA, and SFRP4 had a score of 4.5, and SPP1 had a score of 3.5, all of them were found to be significantly correlated (Figure 5c). While DDIT4, RNF43, and ETV4 had a very low score (0.0 – 0.5) suggesting a low correlation. From the correlation heatmap (figure 5c), ETV4 and RNF43 expressions were significantly correlated to each other according to cluster analysis of the TIMER correlation database.

Figure 5:
Figure 5a

5b

5c:

Figure 5a: Co-expression correlation scatterplot of top 10 upregulated genes in the study having $\log_2(\text{FC}) > 1$ from the TCGA TIMER database. The correlation coefficient and the p -value are given in the top right corner of each of the scatterplot. As seen from the image, highly correlated signals would have a higher correlation. **5b: Chord diagram showing the correlation of TIMER co-expression data.** Top 10 genes that are correlated in the TIMER TCGA COAD database are arranged according to their correlation. The overall contribution of the correlation coefficient was found to be highest among six genes (COL11A1, COL1A1, COMP, SPP1, SFRP4, and INHBA; a score of 3.5 – 4.5). Genes that contributed less to overall correlation were DDIT4, RNF43, and ETV4 (score of 0.0-1.5). **5c: Heatmap showing the co-expression of the top 10 genes in the TIMER database.** Added to the high co-expression correlation with the six genes (COL11A1, COL1A1, SPP1, SFRP4, COMP, and INHBA), there is a separate correlation between RNF43 and ETV4 in a different cluster. Details are described in the main body of the text.

Network Analysis:

The top 20 highly significant genes from the Nanostring expression analysis were fed into the STRING.db online tool to explore the possible protein-protein interactions among these gene signals. The ranking of each of these proteins was determined from the fold-change values obtained from the differential expression ratio data from the current study. As illustrated in figure 6, eleven genes from the top 20 were found to be interacting with each other according to several databases included in STRING.db (gene co-occurrence and gene-neighborhood). Detailed list of interactions is given in Supplementary Dataset File S6). These were INHBA, COL1A1, COL11A1, COMP, SFRP4, SPP1, MMP7, and WNT2. These were indirectly connected to KLF4, CDKN2B, and ETV4. Among these, experimentally proved interactions were projected in three clusters. These were 1) INHBA, COL1A1, COL11A1, and COMP, 2) SFRP4, WNT2, SPP1, MMP7, KLF4, and CDKN2B, and 3) CACNA1H and CACNA2DH. Protein homology was found between COL1A1 and COL11A1, SFRP4 and WNT2, MMP7 and SPP1, and FLNA and FLNC. Interactions between different subunits of the interacting

proteins (FLNC and FLNA, CACNA1H, and CACNA2D3) were also identified with the network analysis. There were no previous records of any gene fusions between them. Apart from these interactions, all the other genes from the top 20 genes were not shown to have any interaction with each other. The correlation cluster is seen with the TIMER TCGA database (and the genes clustered in the STRING.db network analysis was corresponding to many genes to each other.

Figure 6:

Figure 6: Network by STRING.db of the top 20 genes that were identified from the expression analysis. All the interactions were highly significant (FDR adj. p -value < 0.01). All colored nodes are the first shell of the interaction of each of the proteins. Edges represent protein-protein interaction, not necessarily connected. 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). The structures seen inside the nodes are the 3D structures of the known proteins.

Association of immune cell infiltration with gene expression

Figure 7 shows a hierarchical clustering analysis of the correlation among the top 10 genes, identified by the Nanostring analysis from the current study cohort, with the immune cell infiltration data available in the TCGA COAD data obtained via TIMER. The immune cell correlation to the 83 significant genes are given in Supplementary Dataset File S5. Macrophages were aligned with the maximum number of genes within the given set. Five of the six highly co-expressed genes were associated with the tissue macrophage infiltration. Among them, COL11A1, COL1A1, INHBA, SPP1,

SFRP4, and COMP were clustered with macrophages. Macrophages, CD4⁺ T lymphocytes, dendritic cells as well as neutrophils had a typically similar association with a small variance across the different genes. However, the first three genes were clustered with CD4⁺ T-cells, dendritic cells, as well as neutrophils. These signals were less associated with CD8⁺ T cells as well as B-cells. This shows that the high expressions of the first six signals may have originated from the tumor infiltrates. Signals like MMP7 and DDIT4 were not associated with any specific cell type. However, ETV4 and RNF43 were clustered with tumor purity and away from all other infiltrating cell types. This suggests that the expressions of ETV4 and RNF43 might be represented by the tumor cells.

Figure 7

Figure 7: Comparison of the top 10 genes from the current study to immune cell infiltration in the TCGA TIMER database. Tissue Macrophages (Mph) are highly associated with the majority of the genes upregulated, followed by CD4-T lymphocytes, Dendritic cells (Dcell), Neutrophils (Nphil), CD8-T lymphocytes, and B-cells in the descending order. The tumor content (represented by “Purity”) is highly correlated with ETV4 and RNF43.

Discussion

One of the major considerations of the large scale TCGA/Pan-Cancer Atlas studies is to correlate molecular pathogenesis data with the gross pathological and clinical parameters of cancerogenesis in different populations. The molecular classifications prescribed by the TCGA/ Pan-Cancer Atlas are derived from tumor samples of a wide variety of tissues at various levels of tumor stages, infiltrating

cellular types, and different ethnic groups. Correlation of different data in TCGA to clinical medicine requires clinical molecular validation studies and analysis for their translational capacity in different populations. To that end, molecular signals from tumor samples of different populations that are annotated and stratified can be overlaid on to the TCGA at multiple data layers to examine the prognostic and pathogenic role of individual signals in a particular population. Using this method, the strength of the exhaustive analysis of TCGA can be applied directly to clinical implementation. In the current study, we have shown the first expression profile of colon adenocarcinoma in the South Indian population, a population that has not been widely represented in most of the previous large cohort studies. To validate these results, we showed preliminary methods of overlapping the results from the current study with TCGA data using three different methods. First is the comparison of the significant genes identified in the two datasets, second is the co-expression correlation between the genes in both datasets and thirdly is the comparison of differentially expressed genes in the study cohort with immune cell co-localization.

Among all the differentially expressed genes, in multiple group comparisons and the TCGA database by GEPIA and TIMER analysis, ETV4 was the most significant gene identified in our study. ETV4 expression was particularly high in tissues that have proficient MMR. This may be the first study to show a high correlation of ETV4 with MSS CRCs. ETV4 or ETS Variant Transcription factor 4 (E1A enhancer-binding protein – E1AF or Polyoma Enhancer Activator 3 Homolog – PEA3) is a known transcription factor, upregulated and activated in colon adenocarcinoma²⁰⁻²². Overexpression of ETV4 was shown in breast cancer via induction of HER2/Neu^{23,24}. ETV4 or other PEA3 factors (ETV1 and ETV5) were shown to initiate tumorigenesis by chromosomal translocation with other genes regions such as EWS in Ewing Sarcoma²⁵ or with TMPRSS2 in prostate cancer²⁶. ETV4 has also been shown to upregulate the expression of MMP7 in many gastrointestinal tumors²⁷. Although in the current study, MMP7 expression was not highly significant across the samples, it showed the highest fold-change in DE (Figures 1). The reduced expression of ETV4 in MSI tumors could be due to the presence of dinucleotide GT repeats in the third intron of the *ETV4* gene, causing frame-shift mutation in the coding region. In MSS tumors, ETV4, along with the β -catenin-induced Tcf-mediated *MMP7* promoter activation in intestinal and colon cancer cells²⁸. However, a specific driving factor for the upregulation of ETV4 in MSS tumors is unclear from previous studies. ETV4 binding motif is located in the upstream-144 of the *MMP7* promoter region. However, the domain of the ETV4 that binds to the promoter region is unknown yet. Matrix-metalloproteinase-7 (MMP7) is a commonly upregulated gene in colon cancers^{27,29-31}. ETV4 induced MMPs (MMP1 and MMP7) are involved in collagen catabolism, both being molecular markers for the epithelial-mesenchymal transition³². In the current study, though ETV4 was found to be elevated in MSS samples, it was also elevated in cases of MSI with TILs. As an offshoot of ETV4 upregulation mediated MMP-7 activation, proMMP-9 can be activated to attach to a cellular G-protein coupled receptor mechanism for the activation of the Neuraminidase-1 enzyme, which causes cleavage of α 2,3-sialic acid causing immune recognition has been recently identified as a mechanism in many cancers^{33,34}. Though post-translational modification and activation of ETS proteins by either phosphorylation³⁵ or transcriptional co-activation via acetylation or sumoylation²⁶ have already been demonstrated, signals that induce the expression of ETV4 in MSS tumor cells are still not completely understood.

The Nanostring PanCancer panel includes many targets that are associated with tumor microenvironment and invasion. The current study has depicted a cluster of genes that are upregulated in the tumors, compared to normal tissues, although not clustered along with the tumor cells. Signals included are COL1A1, COL11A1, SPP1, COMP, SFRP4, and INHBA. Numerous studies have shown the

proliferative role of collagen in gastric^{32,36}, breast³⁷, and colon cancers^{38,39}. Previous studies have also demonstrated that the upregulation of MMPs (MMP7 and 9) along with other collagen proteins are important for tumor vascular invasion⁴⁰. Along with MMPs and other signaling markers of Epithelial-Mesenchymal Transition (EMT), COMP⁴¹ and SFRP4⁴² were also found to be upregulated in colon cancers and associated with poor overall survival. Inhibin β A (INHBA), is a member of Transforming Growth Factor β (TGF- β) superfamily, whose increased expression has been shown earlier in colon adenocarcinoma cells⁴³, and was recently demonstrated to have a prognostic significance as well⁴³.

The clustering of genes from the network analysis (Figure 5) and TIMER correlation analysis (Figure 6), in the current study, showed two distinct clusters of gene signals, possibly residing in two different cell types/groups of the tumor. The first cluster contained COL1A1, COL11A1, COMP, SFRP4, SPP1, and INHBA which were mainly associated with samples with higher infiltration of tumor-associated macrophages (TAM), TILs (CD4⁺ cells), dendritic cells and neutrophils in TGCA TIMER data. The second cluster contained ETV4 and RNF43 whose expression was associated with the purity of tumor cells in TIMER data. MMP7 and DDI4 were mostly associated with cells that were not associated with high purity of tumor cells. This may implicate that the ETV4 and RNF43 might be directly expressed by tumor cells, while others were expressed in infiltrating non-tumor cells. Since the proliferation of both tumor cells and infiltrating immune cells are required for the tumor evolution and progression, the proliferative role of these signals may apply to these cell types. Though the cell types associated with these signals were previously reported in several studies^{44,45}, the signals correlated in the current study have not been associated with any of the tumor immune cells. Hence, these results may imply that the current set of signals, might present in the immune cells as well. The high-throughput expression analysis approach is limited in its application, in this context, to delineate the location of these signals in a complex tumor microenvironment. More elaborate studies such as immunohistochemical colocalization experiments are required to identify the cellular location of these signals in tumors.

Conclusions

We have shown, in this study, a method to bridge the gap between the large cohort molecular pathogenesis discovery studies, such as TCGA, to low-throughput and uncharacterized pathological tumor samples from routine clinical medicine, often from unique or highly focused ethnic population. We explored a high throughput expression analysis using the Nanostring PanCancer pathway panel in 11 stage-II colon and rectal adenocarcinoma tumor from South Indian (Kerala) population. 83 significant genes enriched cell cycle- apoptosis, chromatin modification, and DNA damage repair pathways. The signals revealed in the study were overlapped with TCGA at three different data layers. Among the top 20 differentially expressed genes with largest FC, two sets of gene signals were clustered, according to the comparative co-expression analysis using GEPIA and protein interaction network analysis. Similar segregation of signals was found in tumor immune infiltration status according to TIMER. The former mRNA signal cluster was found to be associated with macrophages, CD4⁺ T cells, dendritic cells, and neutrophils. Of the two signals in the second cluster, one signal, ETV4, was significantly higher in MSS tumors. The significance of the high expression of ETV4 with respect to MSS and its associated cause and/or effect in Indian CRC is an unexplored area to date. The signals revealed in this study need to be further validated in larger patients' samples in the same population in order to check pathogenic role and predictive and prognostic utility. We think that methods like this, if implemented and validated in different populations in larger numbers, would help in the identification of more precise prognostic predictive markers and effective agents that can counter the pathogenesis to provide better patient care and effective intervention.

Methods

Samples

Archived tumor tissues, which were formalin-fixed paraffin-embedded (FFPE) during the previous year, were utilized for the study. We selected 11 tumor tissues of Stage II CRC, and four normal tissues for the Nanostring nCounter analysis. Institutional scientific and ethics committee approval was obtained before the study. All the samples were previously diagnosed by histopathology as well as immunohistochemistry and/or MSI-PCR for diagnosing MMR deficiency. The normal tissues were obtained from the normal regions of the tumor tissue, which were also histopathologically identified. The study was conducted as per the guidelines of the Institutional Ethics Committee of Amrita Institute of Medical Sciences (AIMS). The project was presented to the Institutional Regulatory Board (IRB), which comprises the Scientific Review Committee and the Institutional Ethics Committee. All the tumor samples included in this study were deidentified. The tissues were obtained as part of the usual diagnostic and treatment policies of the institution. This study did not change the clinical management of those patients. Considering all these facts, the committee decided that there is no need for separate consent from each patient. The study was reviewed and approved by the IRB on Nov 24th 2017 (IRB-AIMS-2017-124).

Sectioning and Microdissection

Serial 5µm histological sections of FFPE tissue block of normal and tumor were prepared. The first and fifth sections were stained with hematoxylin and eosin (H&E), sandwiched tissue levels were mounted unstained on slides, verified by light microscopy for each block using levels 1 & 5, and areas to be micro-dissected were marked. Normal or tumor tissue was micro-dissected from the unstained slides for each case by overlaying the unstained slides in a laminar flow tissue culture hood. Alternatively, one section from the trimmed block is sampled for a tumor and a normal area that is marked in the block. The block is cut in those areas and re-embedded in separate blocks and sectioned into separate sample tubes for further processing.

Nanostring nCounter assay:

The NanoString nCounter[®] Analysis System delivers direct, multiplexed measurement of gene expression, providing digital readouts of the relative abundance of hundreds of mRNA transcripts simultaneously^{46,47}. The nCounter Analysis System is based on gene-specific probe pairs that are hybridized to the sample in solution. The protocol eliminates any enzymatic reactions that might introduce bias in the results. The Reporter Probe carries the fluorescent signal; the Capture Probe allows the complex to be immobilized for data collection. Up to 550 pairs of probes, specific for a set of genes are combined with a series of internal controls to form a CodeSet.

The assay was conducted by a contract research organization (Theracues Pvt Ltd, Bangalore, India) on a pay-by-service basis. According to the protocol from the organization, briefly, three-to-four 5µm FFPE samples sections were taken for RNA extraction using commercial FFPE nucleic acid isolation kit (Roche Molecular Diagnostics). RNA was quantified using Nanodrop and analyzed for fragment distribution using a bioanalyzer (Agilent). Considering the profile on the bioanalyzer ~140 ng of RNA was used for each probing assay. Hybridization with PanCancer Pathway Panel codeset was performed for 18 hrs. Post- hybridization, this was separated on nCounter sprint machine on a microfluidic cartridge. The data were normalized, and fold change was generated using nSolver 4.0 software.

Analysis:

For Nanostring expression assays, data from NCBI Gene Expression Omnibus profiles (<http://www.ncbi.nlm.nih.gov/geo/>) from independent cancer gene sets (accession numbers such as GSE14333, GSE35602, etc.) were used for expression analysis. To find the fold-change, a ratio of median values of the probeset expression in each of the tumor samples to the median value of the probeset expression in the normal tissue was identified. The median values of the probe set expression, after standardization with standard expression sets, and fold-change values were drawn in data matrices, and two-dimensional hierarchical clustering analysis with different Euclidean distance clustering methods was employed using appropriate package modules in R statistical program⁹. Advanced Analysis includes the *Pathway score* and *Differential Expression* Analysis. Nearly 770 genes were probed in a single PanCancer Pathway hybridization chip, including about 606 genes were from thirteen canonical pathways, and about 125 genes were cancer driver genes. Approximately 50 genes were included as reference and housekeeping genes. The fluorescence count obtained from the nCounter machine (cutoff value of ≥ 40) was normalized according to the positive controls and internal housekeeping reference genes. Among the four tumor-normal paired samples, included in the pilot study, the Differentially expressed (DE) genes in the tumor were calculated by taking the geometric mean of the four normal tissue as the denominator and the geometric mean of all the tumor samples as the numerator. The fold changes were calculated from this ratio data. If ratio > 1 ; then Foldchange = Ratio, if ratio < 1 ; then Foldchange = $-1*(1/\text{Ratio})$. The fold changes were then reported. The Fold Change data were filtered for Fold changes ≥ 1.5 & ≤ -1.5 . The genes with counts ≥ 40 in normal or tumor samples were retained. The final list of genes was then marked in red for downregulation and green for upregulation. The top 10 up- and down-regulated genes were used to generate the principal component analysis (PCA) plot. Genes were tested for differential expression in response to each selected sample. For each gene, single linear regression was fit using all selected covariates to predict expression to eliminate confounding results due to measured covariates, and to isolate the independent association of each covariate with gene expression, measuring each variable's association with a gene after holding all other variables constant. Evaluation of differential expression values by group difference between microsatellite instability status, infiltration status as well as the anatomical location were analyzed using Student t-test analysis assuming unequal variance. The Pathway analysis was done for each sample separately using DE genes (DE Call - Yes) using WEB-based GENE SeT AnaLysis Toolkit (WebGestalt)⁴⁸. The Enrichment was carried out for functional databases Kyoto Encyclopaedia of Genes and Genomes (KEGG)⁴⁹ and Panther Pathways⁵⁰ using the Gene Ontology summary database (GOSlim). From the set of genes obtained from the WebGestalt tool, further statistical analysis using PCA was performed for dimension reduction. PCA was done using ClustVis⁵¹, an R-program based component factor analysis tool.

Network Analysis

STRING database (string-db.org) is an integrated, publicly available resource for identifying integrated protein-protein interaction networks in biology, developed by STRING consortium⁵². The interaction network was built on laboratory experiments in protein-protein interactions, conserved co-expression data, genomic context predictions, and text mining from previous research. The genes that were found to be highly significant across the samples and maximal log2 fold-change (either upregulated or downregulated) from the Nanostring nSolver analysis were selected by the appropriate R-program module. These genes, with their log2 fold-change values, were uploaded into the STRING-db.v.11 online search tool, in the "Proteins with Values/Ranks" tab.

Co-expression analysis

The network identified from the STRING database was further evaluated by co-expression analysis by GEPIA and TIMER databases. GEPIA is an online tool (<http://gepia2.cancer-pku.cn/>) to explore RNA sequencing expression analysis of 9736 tumors and 8587 normal samples from the TCGA and GTEx projects using standard processing pipeline⁵³. One-to-one gene expression analysis is possible with GEPIA. Using both databases, the correlation of each of the selected genes, found to be expressed together from the test samples, were analyzed for their co-expression correlation coefficient in the TCGA COAD using TIMER database (<https://cistrome.shinyapps.io/timer/>). The spearman correlation coefficient obtained from the TIMER is compared against the network analysis results obtained from STRING.db.

Immune cell infiltrate Analysis

TIMER (cistrome.shinyapps.io/timer) is a web-based database and computational tool to statistically analyze tumor-infiltrating immune cells from gene expression profiles¹⁹. The tool compares the immune cell infiltrates such as B-cells, CD4⁺ T cells, CD8⁺ T cells, neutrophils, macrophages, and dendritic cells from the TCGA data, which contain more than 10,000 samples from 32 different cancer types. The selected genes from the study samples were compared against the TIMER immune cell infiltrate to identify the type of immune cells that were maximally expressed in the TCGA COAD cohort from the given set of unique genes identified from the local population. The correlation coefficient of each cell type in the TIMER database against the top enriched genes from the current study is analyzed and clustered using hierarchical clustering analysis to identify the set of genes that may have an association with a particular cell type. Clustering analysis was done using base functions such as *hclust*, and *dist* functions in the R program.

Data availability

Supplementary and other data are publicly available at the Open Science Forum website. The data can be accessed at the weblink: <https://osf.io/tcdk9/>.

References

- 1 Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546-1558, doi:10.1126/science.1235122 (2013).
- 2 Lin, J. K., Chang, S. C., Yang, Y. C. & Li, A. F. Y. Loss of heterozygosity and DNA aneuploidy in colorectal adenocarcinoma. *Ann Surg Oncol* **10**, 1086-1094, doi:10.1245/Aso.2003.12.014 (2003).
- 3 Leary, R. J. *et al.* Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 16224-16229, doi:10.1073/pnas.0808041105 (2008).
- 4 Whiffin, N. *et al.* Identification of susceptibility loci for colorectal cancer in a genome-wide meta-analysis. *Human molecular genetics* **23**, 4729-4737, doi:10.1093/hmg/ddu177 (2014).
- 5 Cancer Genome Atlas, N. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330-337, doi:10.1038/nature11252 (2012).
- 6 Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nat Med* **21**, 1350-1356, doi:10.1038/nm.3967 (2015).

- 1 7 Lynch, H. T., Lynch, J. F., Shaw, T. G. & Lubiński, J. HNPCC (Lynch Syndrome): Differential
2 Diagnosis, Molecular Genetics and Management - a Review. *Hereditary Cancer in Clinical*
3 *Practice* **1**, 1-12, doi:10.1186/1897-4287-1-1-7 (2003).
- 4 8 Aaltonen, L. A. *et al.* Incidence of hereditary nonpolyposis colorectal cancer and the feasibility of
5 molecular screening for the disease. *N Engl J Med* **338**, 1481-1487,
6 doi:10.1056/NEJM199805213382101 (1998).
- 7 9 Dunne, P. D. *et al.* Challenging the Cancer Molecular Stratification Dogma: Intratumoral
8 Heterogeneity Undermines Consensus Molecular Subtypes and Potential Diagnostic Value in
9 Colorectal Cancer. *Clin Cancer Res* **22**, 4095-4104, doi:10.1158/1078-0432.CCR-16-0032
10 (2016).
- 11 10 Kane, M. F. *et al.* Methylation of the hMLH1 Promoter Correlates with Lack of Expression of
12 hMLH1 in Sporadic Colon Tumors and Mismatch Repair-defective Human Tumor Cell Lines.
13 *Cancer Research* **57**, 808-811 (1997).
- 14 11 Valeri, N. *et al.* Modulation of mismatch repair and genomic stability by miR-155. *Proceedings*
15 *of the National Academy of Sciences* **107**, 6982-6987, doi:10.1073/pnas.1002472107 (2010).
- 16 12 Mensenkamp, A. R. *et al.* Somatic mutations in MLH1 and MSH2 are a frequent cause of
17 mismatch-repair deficiency in Lynch syndrome-like tumors. *Gastroenterology* **146**, 643-646
18 e648, doi:10.1053/j.gastro.2013.12.002 (2014).
- 19 13 Yeole, B. B. Trends in cancer incidence in esophagus, stomach, colon, rectum and liver in males
20 in India. *Asian Pac J Cancer Prev* **9**, 97-100 (2008).
- 21 14 Hause, R. J., Pritchard, C. C., Shendure, J. & Salipante, S. J. Classification and characterization
22 of microsatellite instability across 18 cancer types. *Nat Med* **22**, 1342-1350, doi:10.1038/nm.4191
23 (2016).
- 24 15 Rajkumar, T. *et al.* Mutation analysis of hMSH2 and hMLH1 in colorectal cancer patients in
25 India. *Genet Test* **8**, 157-162, doi:10.1089/gte.2004.8.157 (2004).
- 26 16 Pandey, V. *et al.* Assessment of microsatellite instability in colorectal carcinoma at an Indian
27 center. *Int J Colorectal Dis* **22**, 777-782, doi:10.1007/s00384-006-0241-3 (2007).
- 28 17 Raman, R. *et al.* Evidence for possible non-canonical pathway(s) driven early-onset colorectal
29 cancer in India. *Mol Carcinog* **53 Suppl 1**, E181-186, doi:10.1002/mc.21976 (2014).
- 30 18 Kanth, V. V. *et al.* Microsatellite instability and promoter hypermethylation in colorectal cancer
31 in India. *Tumour Biol* **35**, 4347-4355, doi:10.1007/s13277-013-1570-9 (2014).
- 32 19 Buhard, O. *et al.* Multipopulation Analysis of Polymorphisms in Five Mononucleotide Repeats
33 Used to Determine the Microsatellite Instability Status of Human Tumors. *Journal of Clinical*
34 *Oncology* **24**, 241-251, doi:10.1200/jco.2005.02.7227 (2006).
- 35 20 Horiuchi, S. *et al.* Association of ets-related transcriptional factor E1AF expression with tumour
36 progression and overexpression of MMP-1 and matrilysin in human colorectal cancer. *J Pathol*
37 **200**, 568-576, doi:10.1002/path.1387 (2003).
- 38 21 Eskandari, E., Mahjoubi, F. & Motalebzadeh, J. An integrated study on TFs and miRNAs in
39 colorectal cancer metastasis and evaluation of three co-regulated candidate genes as prognostic
40 markers. *Gene* **679**, 150-159, doi:10.1016/j.gene.2018.09.003 (2018).
- 41 22 Benz, C. C. *et al.* HER2/Neu and the Ets transcription activator PEA3 are coordinately
42 upregulated in human breast cancer. *Oncogene* **15**, 1513-1525, doi:10.1038/sj.onc.1201331
43 (1997).
- 44 23 Myers, E. *et al.* A positive role for PEA3 in HER2-mediated breast tumour progression. *British*
45 *journal of cancer* **95**, 1404-1409, doi:10.1038/sj.bjc.6603427 (2006).
- 46 24 Peter, M. *et al.* A new member of the ETS family fused to EWS in Ewing tumors. *Oncogene* **14**,
47 1159-1164, doi:10.1038/sj.onc.1200933 (1997).
- 48 25 Paulose, R. R., Ail, D. A., Biradar, S., Vasudevan, A. & Sundaram, K. R. Prognostic and
49 predictive significance of microsatellite instability in stage II colorectal carcinoma: An 8-year
50 study from a tertiary center in South India. *Indian J Cancer* **56**, 302-308,
51 doi:10.4103/ijc.IJC_365_18 (2019).

- 26 Yamamoto, H. *et al.* Expression of ets-related transcriptional factor E1AF is associated with tumor progression and over-expression of matrilysin in human gastric cancer. *Carcinogenesis* **25**, 325-332, doi:10.1093/carcin/bgh011 (2004).
- 27 Crawford, H. C. *et al.* The PEA3 subfamily of Ets transcription factors synergizes with beta-catenin-LEF-1 to activate matrilysin transcription in intestinal tumors. *Molecular and cellular biology* **21**, 1370-1383, doi:10.1128/MCB.21.4.1370-1383.2001 (2001).
- 28 Janknecht, R. Regulation of the ER81 transcription factor and its coactivators by mitogen- and stress-activated protein kinase 1 (MSK1). *Oncogene* **22**, 746-755, doi:10.1038/sj.onc.1206185 (2003).
- 29 Sole, X. *et al.* Discovery and validation of new potential biomarkers for early detection of colon cancer. *PLoS One* **9**, e106748, doi:10.1371/journal.pone.0106748 (2014).
- 30 Xu, H. *et al.* Identification and Verification of Core Genes in Colorectal Cancer. *Biomed Res Int* **2020**, 8082697, doi:10.1155/2020/8082697 (2020).
- 31 Campo, L., Zhang, C. & Breuer, E. K. EMT-Inducing Molecular Factors in Gynecological Cancers. *Biomed Res Int* **2015**, 420891, doi:10.1155/2015/420891 (2015).
- 32 Junnila, S. *et al.* Gene expression analysis identifies over-expression of CXCL1, SPARC, SPP1, and SULF1 in gastric cancer. *Genes Chromosomes Cancer* **49**, 28-39, doi:10.1002/gcc.20715 (2010).
- 33 Haxho, F., Neufeld, R. J. & Szewczuk, M. R. Neuraminidase-1: a novel therapeutic target in multistage tumorigenesis. *Oncotarget* **7**, 40860-40881, doi:10.18632/oncotarget.8396 (2016).
- 34 Abdulkhalek, S. & Szewczuk, M. R. Neu1 sialidase and matrix metalloproteinase-9 cross-talk regulates nucleic acid-induced endosomal TOLL-like receptor-7 and -9 activation, cellular signaling and pro-inflammatory responses. *Cell Signal* **25**, 2093-2105, doi:10.1016/j.cellsig.2013.06.010 (2013).
- 35 Guo, B., Panagiotaki, N., Warwood, S. & Sharrocks, A. D. Dynamic modification of the ETS transcription factor PEA3 by sumoylation and p300-mediated acetylation. *Nucleic acids research* **39**, 6403-6413, doi:10.1093/nar/gkr267 (2011).
- 36 Norton, N. *et al.* Assessment of Tumor Heterogeneity, as Evidenced by Gene Expression Profiles, Pathway Activation, and Gene Copy Number, in Patients with Multifocal Invasive Lobular Breast Tumors. *PLoS One* **11**, e0153411, doi:10.1371/journal.pone.0153411 (2016).
- 37 Fischer, H., Stenling, R., Rubio, C. & Lindblom, A. Colorectal carcinogenesis is associated with stromal expression of COL11A1 and COL5A2. *Carcinogenesis* **22**, 875-878, doi:10.1093/carcin/22.6.875 (2001).
- 38 Mojica, W. & Hawthorn, L. Normal colon epithelium: a dataset for the analysis of gene expression and alternative splicing events in colon disease. *BMC Genomics* **11**, 5, doi:10.1186/1471-2164-11-5 (2010).
- 39 St Croix, B. *et al.* Genes expressed in human tumor endothelium. *Science* **289**, 1197-1202, doi:10.1126/science.289.5482.1197 (2000).
- 40 Nfonsam, V. N. *et al.* COMP Gene Coexpresses With EMT Genes and Is Associated With Poor Survival in Colon Cancer Patients. *J Surg Res* **233**, 297-303, doi:10.1016/j.jss.2018.08.021 (2019).
- 41 Nfonsam, L. E., Jandova, J., Jecius, H. C., Omesiete, P. N. & Nfonsam, V. N. SFRP4 expression correlates with epithelial mesenchymal transition-linked genes and poor overall survival in colon cancer patients. *World journal of gastrointestinal oncology* **11**, 589-598, doi:10.4251/wjgo.v11.i8.589 (2019).
- 42 Okano, M. *et al.* Significance of INHBA expression in human colorectal cancer. *Oncol Rep* **30**, 2903-2908, doi:10.3892/or.2013.2761 (2013).
- 43 Li, X. *et al.* INHBA is a prognostic predictor for patients with colon adenocarcinoma. *BMC Cancer* **20**, 305, doi:10.1186/s12885-020-06743-2 (2020).

- 44 Zhou, R. *et al.* Immune cell infiltration as a biomarker for the diagnosis and prognosis of stage I-III colon cancer. *Cancer Immunol Immunother* **68**, 433-442, doi:10.1007/s00262-018-2289-7 (2019).
- 45 Peng, D. *et al.* An immune infiltration signature to predict the overall survival of patients with colon cancer. *IUBMB Life* **71**, 1760-1770, doi:10.1002/iub.2124 (2019).
- 46 Geiss, G. K. *et al.* Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nat Biotechnol* **26**, 317-325, doi:10.1038/nbt1385 (2008).
- 47 Balko, J. M. *et al.* Profiling of residual breast cancers after neoadjuvant chemotherapy identifies DUSP4 deficiency as a mechanism of drug resistance. *Nat Med* **18**, 1052-1059, doi:10.1038/nm.2795 (2012).
- 48 Wang, J., Vasaikar, S., Shi, Z., Greer, M. & Zhang, B. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic acids research* **45**, W130-W137, doi:10.1093/nar/gkx356 (2017).
- 49 Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research* **40**, D109-114, doi:10.1093/nar/gkr988 (2012).
- 50 Mi, H. *et al.* PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic acids research* **45**, D183-D189, doi:10.1093/nar/gkw1138 (2017).
- 51 Metsalu, T. & Vilo, J. ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. *Nucleic acids research* **43**, W566-570, doi:10.1093/nar/gkv468 (2015).
- 52 Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research* **47**, D607-D613, doi:10.1093/nar/gky1131 (2019).
- 53 Li, T. *et al.* TIMER: A Web Server for Comprehensive Analysis of Tumor-Infiltrating Immune Cells. *Cancer Res* **77**, e108-e110, doi:10.1158/0008-5472.CAN-17-0307 (2017).

Acknowledgments:

The authors would like to thank the senior and junior faculty staffs, office and laboratory staffs of Departments of Medical Oncology, Pathology, Molecular Biology, Biochemistry, Health Sciences Research and Molecular Oncology Laboratory of Amrita Institute of Medical Sciences, Kochi for the coordination of resources for the conduct of research. The authors would also like to thank the staff and CEO Mr. Gopalakrishna Ramaswamy of Theracues Innovations Pvt Ltd, Bangalore, India for the coordination and conduct of Nanostring assays. The study was supported by an internal seed grant by Amrita Vishwa Vidyapeetham to PSA.

Author contributions:

PSA has designed the experiments with RAJ, VM, RRP, and KP. VM and DMV reviewed the project and further modified. Conduct of the research was done by PSA and RAJ. Data analysis was done by PSA. The manuscript was written by PSA and RAJ, reviewed by VM, RRP, KP, and DMV.

Competing interests

The author(s) declare no competing interests.