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 (MSS) category accounts for more than four-fifths of colon and rectal cancer (CRC). However, studies during the last two decades in the Indian population have shown that the microsatellite instable (MSI) is more than 30% of CRC cases. We have conducted a study to explore the pathogenesis of microsatellite instability in Indian CRC. In the preliminary studies, we conducted a Nanostring Pan-Cancer pathway analysis of early-stage CRC (n = 10, MSS = 5, MSI = 5) and normal tissues (n=7). We identified the differentially expressed genes associated with the tumor and correlated them against microsatellite instability status. Among them, *AXIN2*, *ETV4*, and *RNF43* were tumor cell-specific signals that had a differential expression between MSI and MSS groups. When overlapped with the TCGA immune cell infiltration data, TIMER, these genes segregated to the tumor cells. Moreover, they were less associated with other significant genes in protein-protein interaction analysis by STRING. The expression of these genes was further validated in another set of early-stage microsatellite instable CRC (n = 15) by qPCR. The expression fold-changes of these signals suggest distinct subsets in the MSI subgroup of CRC in the Indian population.

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.*

Introduction

The average incidence rates of colon cancer across different parts of India was found to be 4.3 and 3.9 in 100,000 males and females, respectively, as per the hospital-based cancer registry by the Indian Council for Medical Research (ICMR) in 2014 [1][2]. This incidence rate is found to be rising, according to a later epidemiological study comparing all gastro-intestinal cancers [3]. The International Agency for Research on Cancer (IARC) 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 global study conducted on the Indian population showed that age-standardized incidence is 8.3 with a mortality rate of 7.5 [4]. 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 [3].

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 [5][6]. 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 another common cause, constituting roughly one-fifth of the first cause, which includes the hypermutated type in The Cancer Genome Atlas (TCGA) Classification [7]. Extensive molecular and genomic pathogenesis studies on different cancer types, such as TCGA, have not incorporated representative molecular epidemiological contributions from certain ethnic groups [8]. In India, multiple long-term series studies conducted in southern states of Tamil Nadu, Karnataka, Telangana, and Kerala showed a varying frequency of MSI in the CRC from 30% to 67% of cases [9][10][11][12][13]. With these data, however, molecular pathogenesis studies have not been conducted to explore the reasons behind the higher prevalence of MMR deficiency in the pathogenesis of CRC in the Indian population. We conducted a pilot expression analysis using Nanostring Pan-cancer pathway analysis in early-stage CRC comprising both MSI and MSS groups [14]. Based on the different gene expression patterns observed in the pilot study, we explored further with higher number of tissues and validated with orthogonal testing of those signals by RT-PCR in the current study.

Results

Samples

Before the initiation, the study was reviewed and approved by the Institutional Scientific Review Committee and Ethics Committee. The samples were obtained from tissue archives in the form of FFPE blocks stored in the Pathology department of the host institution. For Nanostring pathway analysis, eleven archived samples of Stage II colon adenocarcinoma were selected (Supplementary Dataset File S1, sheet 2). Since the MSI status was tested exclusively in early-stage CRC, tissues with available data on MSI characteristics were limited to early-stage CRC. Tumor samples were obtained from subjects with an age of onset from 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) [15]. Accordingly, six tumor tissues were categorized as deficient MMR (MSI), and five were proficient MMR (MSS). Validation trial was performed with 15 cases of exclusive early-stage MSI CRC. All 15 cases were identified as MSI-high by MMR IHC and MSI-PCR in the validation cohort. Details are given in Supplementary Dataset File S1, sheet 3. The age of onset was between 21 and 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, except two, cases had TIL.

Nanostring Expression Analysis

Sample annotation and Nanostring gene expression raw data are provided in Supplementary Dataset File S2. One of the tumor samples (SH) showed very low 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, 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 Log_2 fold-change of > 2. ($|FC\ (log_2)| > 2$); 22 genes were upregulated ($FC\ (log_2) \geq 2$), and 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.

On 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.

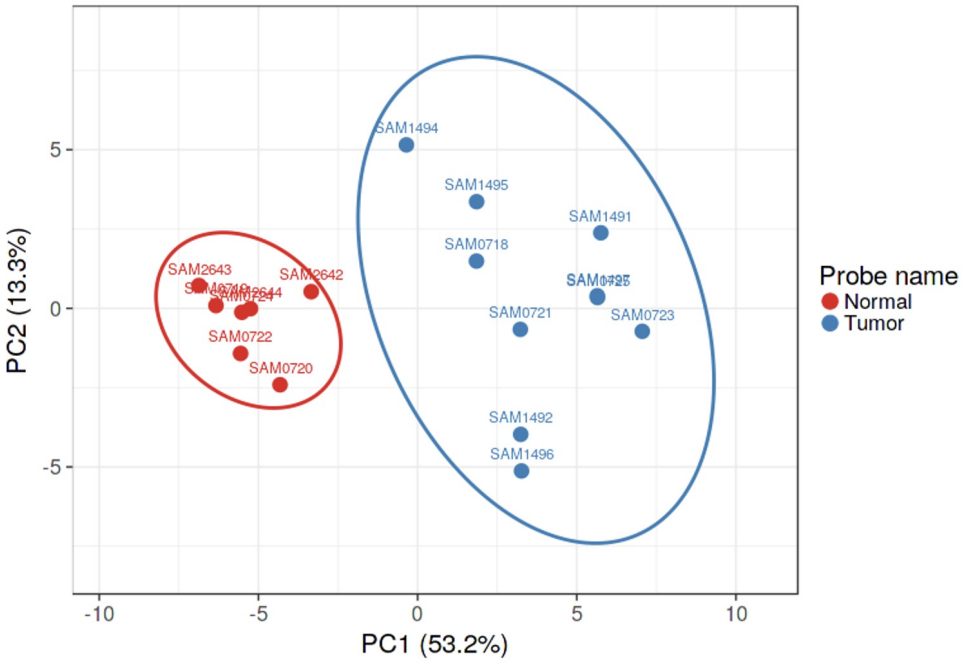

Figure 1. Nanostring Pan-cancer analysis. **1a**: Principal component analysis (PCA) showing segregation of tumor and normal from the differential expression analysis.

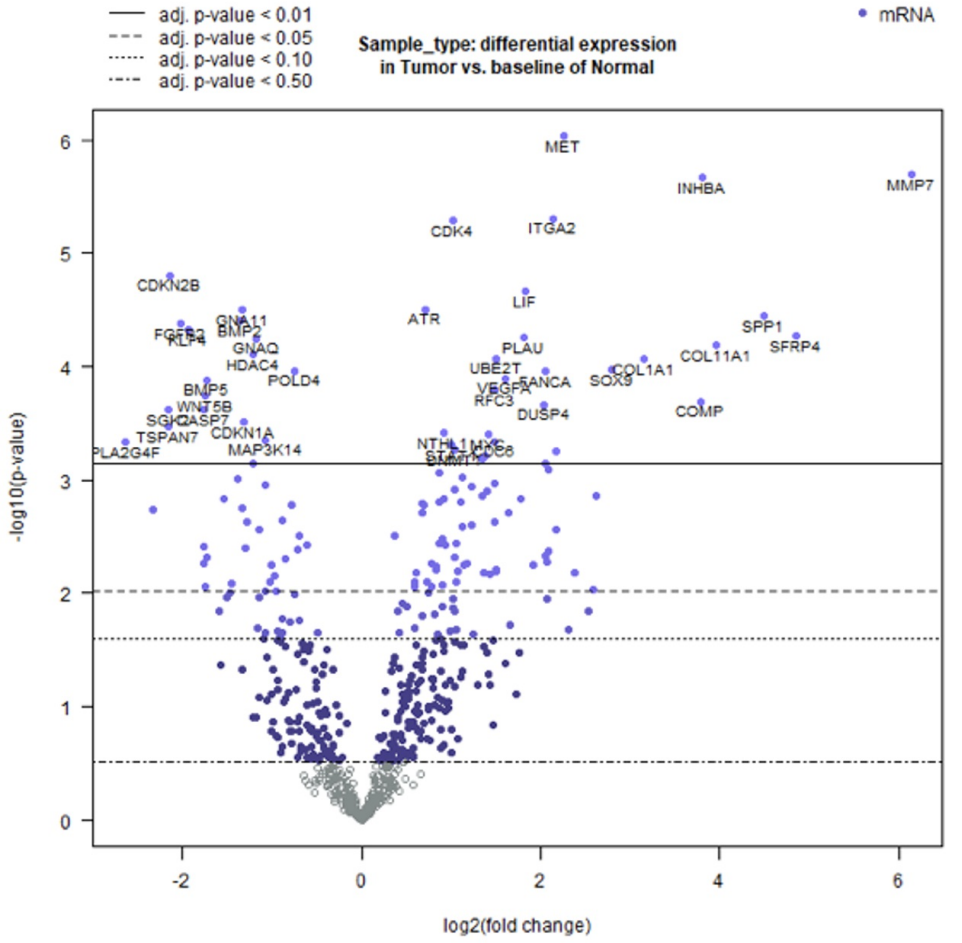

Figure 1. Nanostring Pan-cancer analysis. **1b**: Significant genes obtained from the differential expression analysis, FDR adjusted *p*-values (-log10 scale) on the Y-axis and fold-change (log2 scale)

on the X-axis.

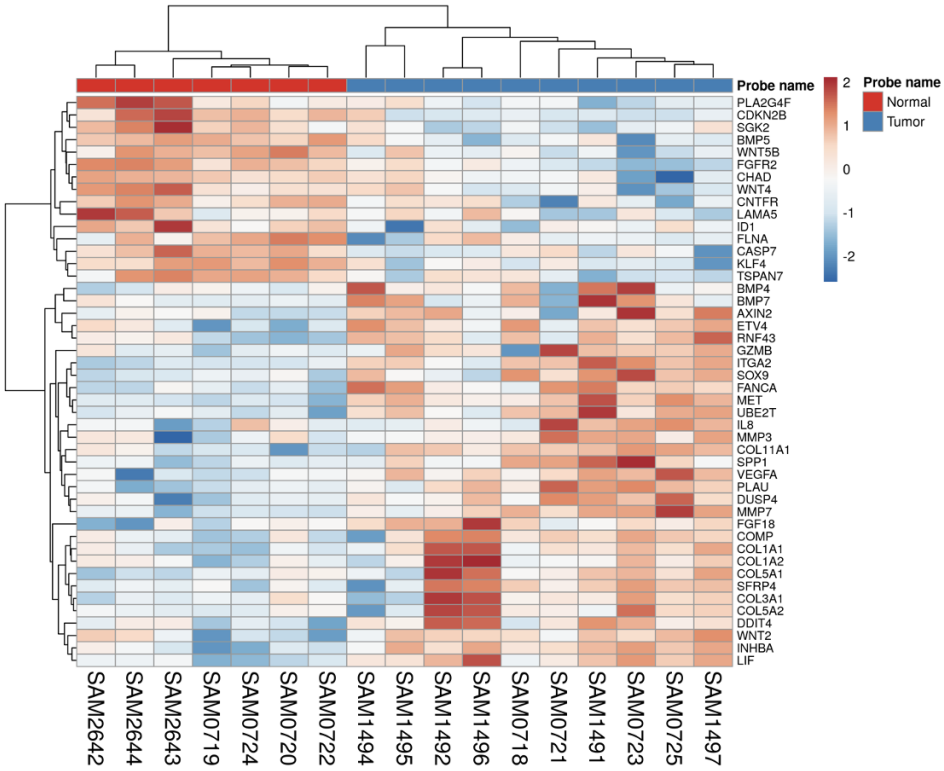

Figure 1. Nanostring Pan-cancer analysis. **1c**: Heatmap showing the clustering of significant-top 45 genes from the DE analysis (Y-axis) among tumor and normal samples (X-axis).

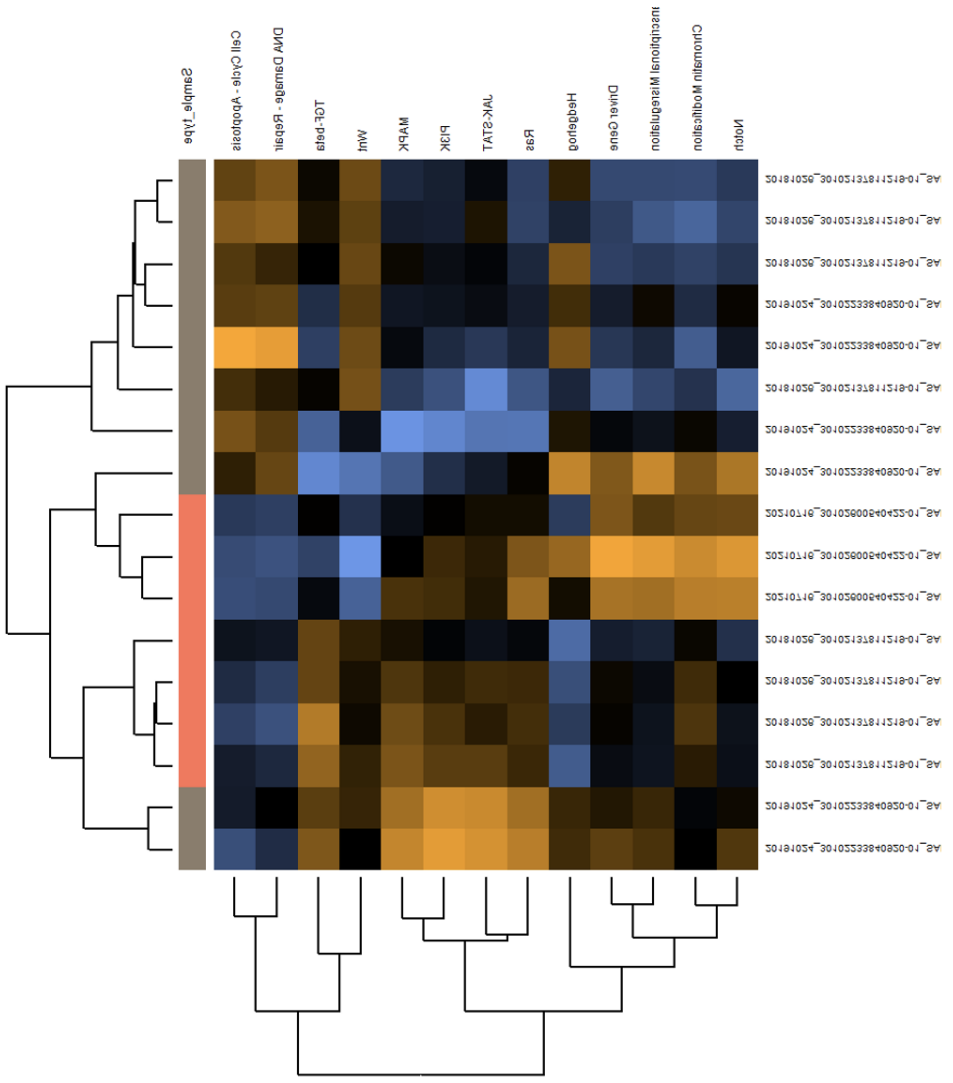

Figure 1. Nanostring Pan-cancer analysis. **1d**: Heatmap showing pathway enrichment (X-axis) against different tumor samples (Y-axis).

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 differentially regulated, with p-value < 0.05 (Figure 2a). On FDR analysis, none of these genes 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 with p-value < 0.05, but FDR analysis failed to qualify them as significant (Figure 2b). None of the genes had an $|FC\ (\log_2)| > 1$. This shows that the fold changes in these genes are not significant enough to be compared 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. A correlation between the mean FC values in MSI vs. MSS is plotted in Figure 2c showing an $R^2 = 0.99$.

However, there were a few genes that spread across on either side of the trendline. The segregation of genes based on the trendline is further statistically evaluated by two-tailed *Student* t-test, showing a *p*-value <0.05 denoting a significant difference between the mean FC values between these two groups. The genes, *RNF43*, *ETV4*, *AXIN2* and *MMP7* were below the trendline are highlighted in Figure 2c. The significant genes in these two comparative analyses are given in separate sheets in Supplementary Data file S4.

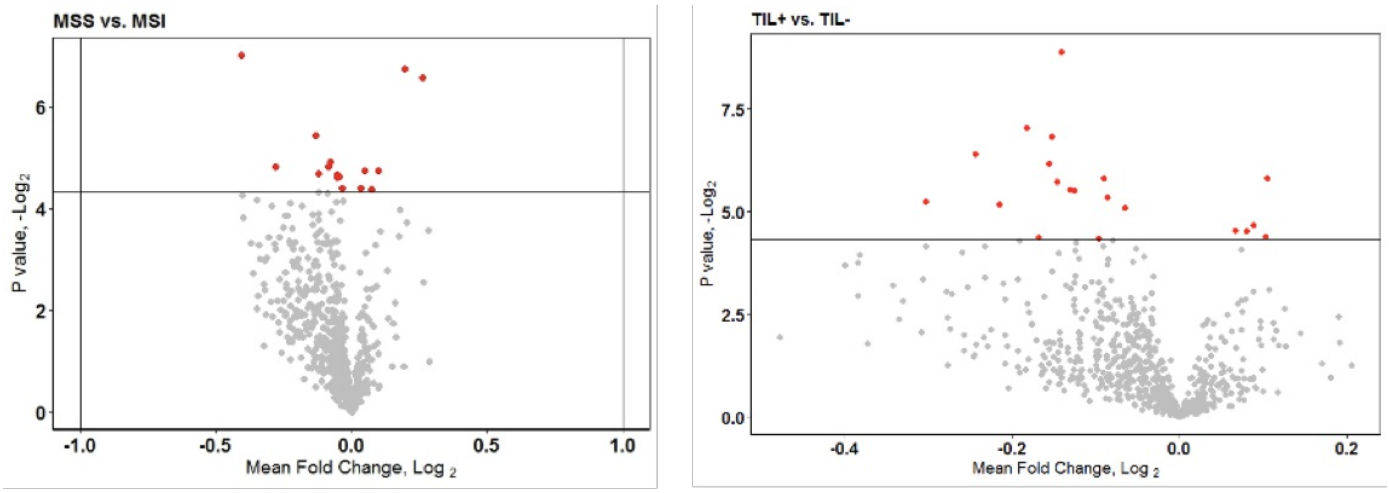

Figure 2. Scatter plots showing significant genes from the Nanostring DGE in MSI vs. MSS groups (2a) and the presence or absence of TIL (2b). In both the plots, X-axis is the fold change in the log2 scale, and Y-axis is the *p*-value in the -log2 scale. The horizontal line in the middle of the field 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).

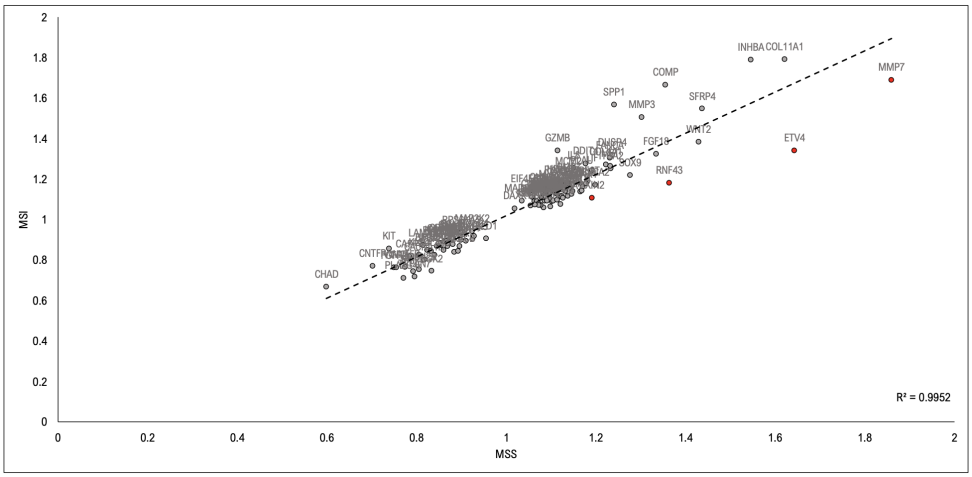

Figure 2c. Correlation scatter plot of mean FC expression of 119 gene signals between MSS and MSI groups. A trendline is drawn to show the spread of the data (R^2= 0.99), with a few genes spread on both sides of the trendline. The gene signals that fell widely below the trendline are highlighted in red dots.

GEPIA Correlation analysis

The expression profile in the current study (AIMS) was compared against TCGA colon adenocarcinoma (COAD) dataset from GEPIA. The Nanostring Pathway Analysis gene set comprises 180 of the 5078 genes included in the TCGA COAD dataset in the GEPIA database. 73 of the 180 genes in the TCGA set were included among the 119 significant genes of the AIMS study (See Figure 3a and Supplementary Dataset File S5). The most significant gene found in the TCGA from the AIMS dataset was *ETV4*; in the AIMS study, *MET* was highly significant. In the analysis, 19 of the 73 significant genes exhibited |FC (log_2)| >2, of which 16 were upregulated, and three downregulated. Comparing the adj.*p*-values between the current study and TCGA, the correlation was -0.083 (Pearson method, Figure 3b). On the contrary, when the Fold-change was compared between these two studies, the correlation coefficient was 0.77 (See Figure 3c). Genes with |FC (log_2)| > 2 that were significant (adj.*p*-value < 0.05) in both the AIMS group, as well as TCGA group, were *MMP7*, *SFRP4*, *SPP1*, *COL11A1*, *INHBA*, *COMP*, *SOX9*, *ETV4*, *MET*, *ITGA2*, *RNF43*, *FANCA*, *DUSP4*, *FGFR2*, *TSPAN7*, and *PLA2G4F*, had |FC (log_2)| >5 in the TCGA cohort. The common genes in both studies showed a similar expression pattern (upregulation/downregulation), while two genes (*PLA2G4F* and *TMPRSS2*) had reverse FC values, an upregulation in GEPIA set (1.49 and 2.27) and downregulation in AIMS set (-2.65 and -1.28 respectively).

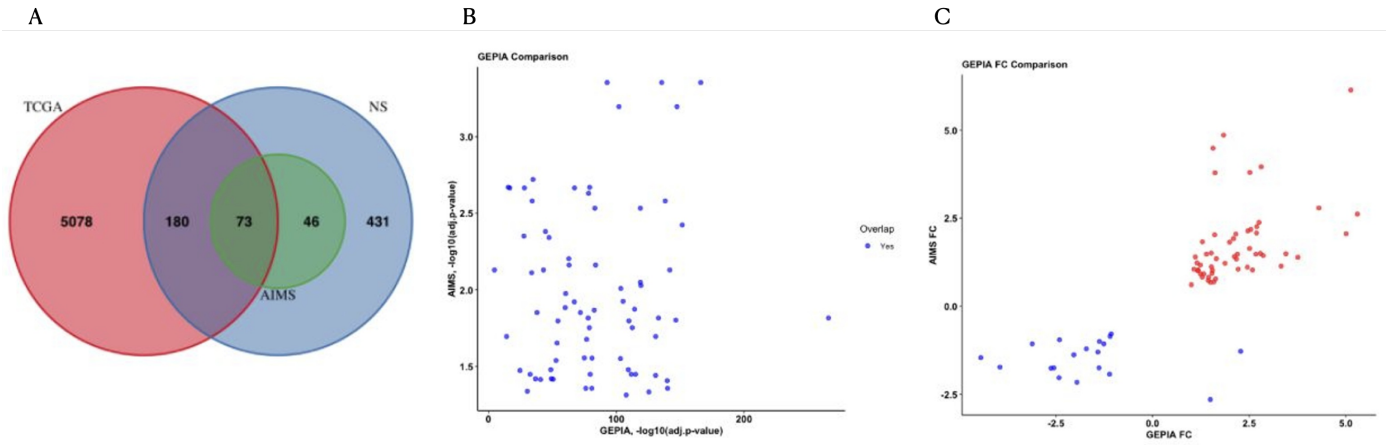

Figure 3: The Venn diagram (3a) showing the number of genes from GEPIA (as "TCGA" in red), Nanostring gene set ("NS" in blue), and significant genes in the current study ("AIMS" in green). Scatter plot of the significant genes (adj p-value < 0.05) in AIMS data overlapped with the p-value of the identical genes in TCGA data in **3b**. The Pearson correlation coefficient was -0.083. Scatter plot of fold-change of expression of significant genes in AIMS data to that of the TCGA data obtained from GEPIA in **3c**. 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 respective protein-protein interaction, gene co-occurrence, and gene-neighborhood in STRING.db between the top 20 genes is shown. 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, RNF43 and DDIT4, were found to have no interaction with the above proteins. Protein homology-based interaction was found between COL1A1, COL1A2,

COL3A1, COL11A1, and ITGA2. ITGA2 has homology-based interaction with COMP, SPP1, and MET. SPP1 and MET have homology-based interaction with MMP7 and FGF18 respectively. Though all of them are associated by gene-neighborhood criteria, above proteins also have additional experimental evidence of interaction. The strong multi-level interaction network found between these sets of genes cause them to be clustered together. Some of the proteins in this network, though may not be homology-based, but are directly networking with this group of proteins are INHBA, BMP7, and SOX9, and indirectly with ETV4 via SOX9. On the other hand, certain other genes, such as SFRP4, CXCL8, GZMB, have only evidence of gene neighborhood from curated databases. It is interesting to note that, many of the genes found to be differentially interacting in PPI network, such as ETV4, and RNF43, were also found to be segregated on the group comparison of MSI vs. MSS in the pilot Nanostring study (See Figure 2c). Moreover, the correlation cluster that was found in the TIMER correlation analysis and the genes clustered in the STRING.db network analysis corresponded to many genes to each other.

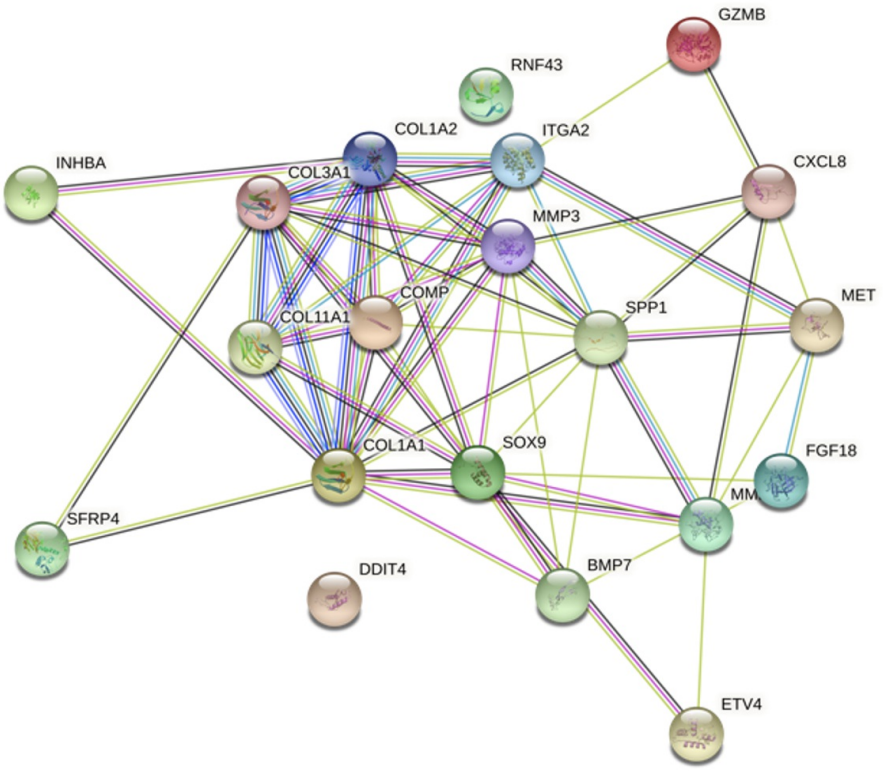

Figure 4. Protein-Protein Interaction Network analysis in STRING.db. Color codes of interaction lines represent gene neighborhood (green), gene fusions (red), gene co-occurrence (dark blue), from curated databases (teal), experimentally determined (pink), text mining (yellow), co-expression (black),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

Association of immune cell infiltration with gene expression

To explore and segregate genes associated with tumor cells vs. immune cells, the significant genes in the current study were compared against the TCGA Tumor Immune Estimation Resource (TIMER) data. A hierarchical clustering analysis (HCA) of the correlation among the top 30 significantly upregulated genes in the Nanostring analysis (adj. p-value < 0.05)

was explored according to their immune cell infiltration status data obtained from TIMER (see Figure 5 and Supplementary Dataset File S7). All the genes, except four, were associated with the immune cell infiltrates, dendritic cells, macrophages, neutrophils, CD4 T, 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 1.64, 1.05, 2.62, and 2.06. These gene signals correlate more with the tumor cells than any other cell type in the TIMER database.

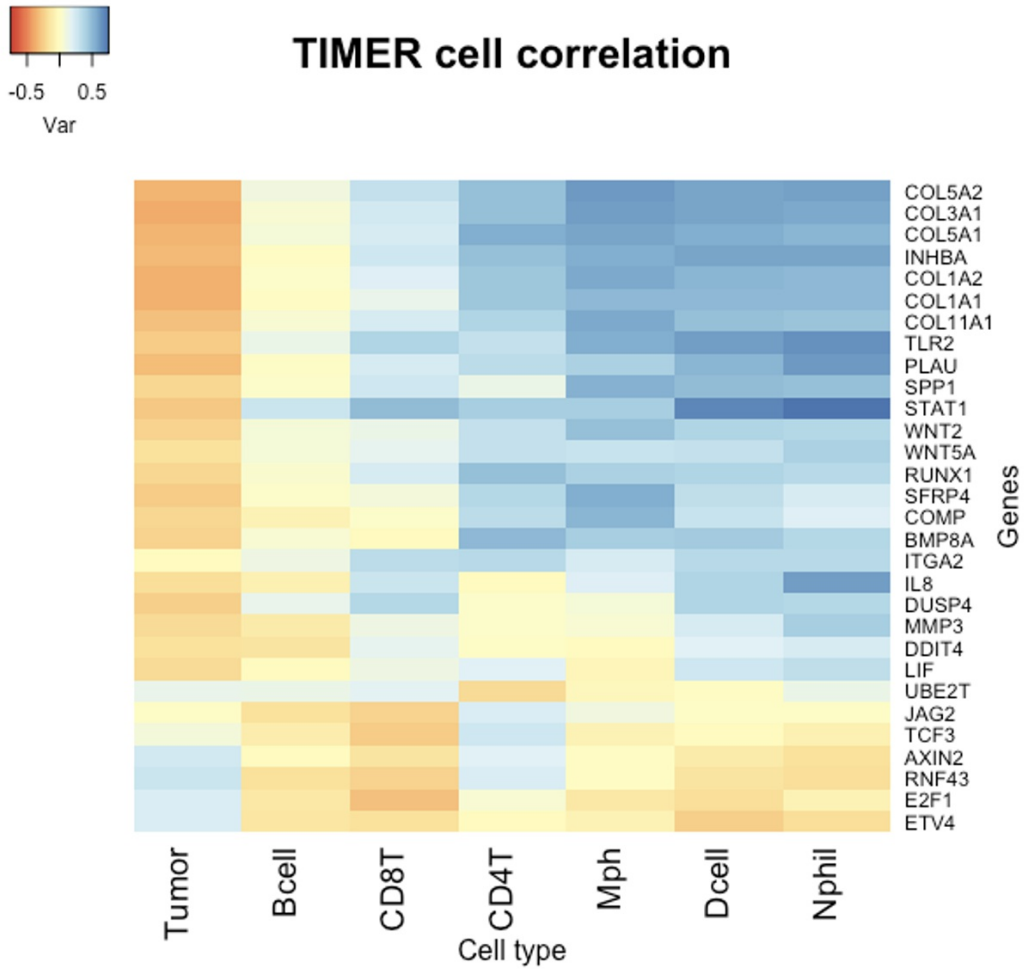

Figure 5. HCA correlation of top 30 significant genes identified in the Nanostring analysis against immune cell infiltration. Clustering based on a negative ordering of the variance of Partial correlation coefficients (ρ) of each of the top 30 significant genes from current study to the infiltrating 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 paired tumor-normal tissue samples of early-stage microsatellite instability-high (MSI-H) colon cancer (n = 15) was obtained from the archive for an RT-PCR. The MSI status of these samples were confirmed with IHC and MSI-PCR as well. An overall comparison of the differential

expression in terms of copy number fold-change in tumor shows broad variation in expression levels of these genes across different MSI groups. The variance of gene expression across the samples and across the genes was compared and assessed by two-way ANOVA. The Ct values and the relative fold-change based on copy number evaluation is detailed in Supplementary Data File S8. The general expression pattern of the three genes curated from the discovery set is given in Figure 6.

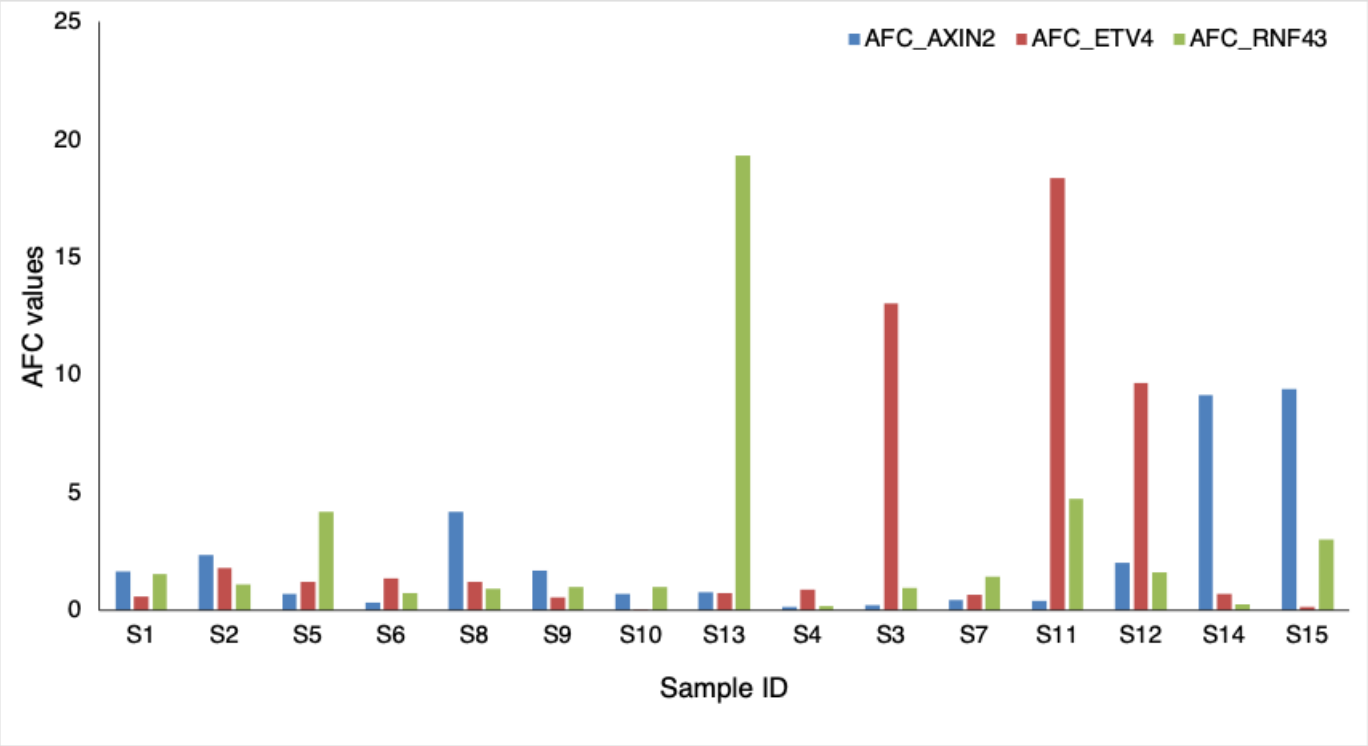

Figure 6. Comparison of expression by RT-PCR validation. Barplot showing the absolute expression fold-change (AFC) of the three genes, *AXIN2*, *ETV4* and *RNF43*. Values of the fold-change is given in Supplementary Dataset File S8.

Figure 7 shows the variance in gene signal fold-changes for the validation set of samples. There was a significant overlap in the overall variations in expression comparing ΔCt of tumor vs. control on principal component analysis (Figure 7a). However, *RNF43* and *ETV4* are positively correlated to PC1 and PC2. *AXIN2* has a positive correlation with PC1 and negative to PC2. At the same time, *CTNNB1* has an opposite trend to *AXIN2*. *TLR4* has a negative trend in both the principal components suggesting that the variance differs from all the other targets. On Hierarchical clustering analysis on variance, *TLR4* clustered differentially from that of the other signals (Figure 7b). 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.

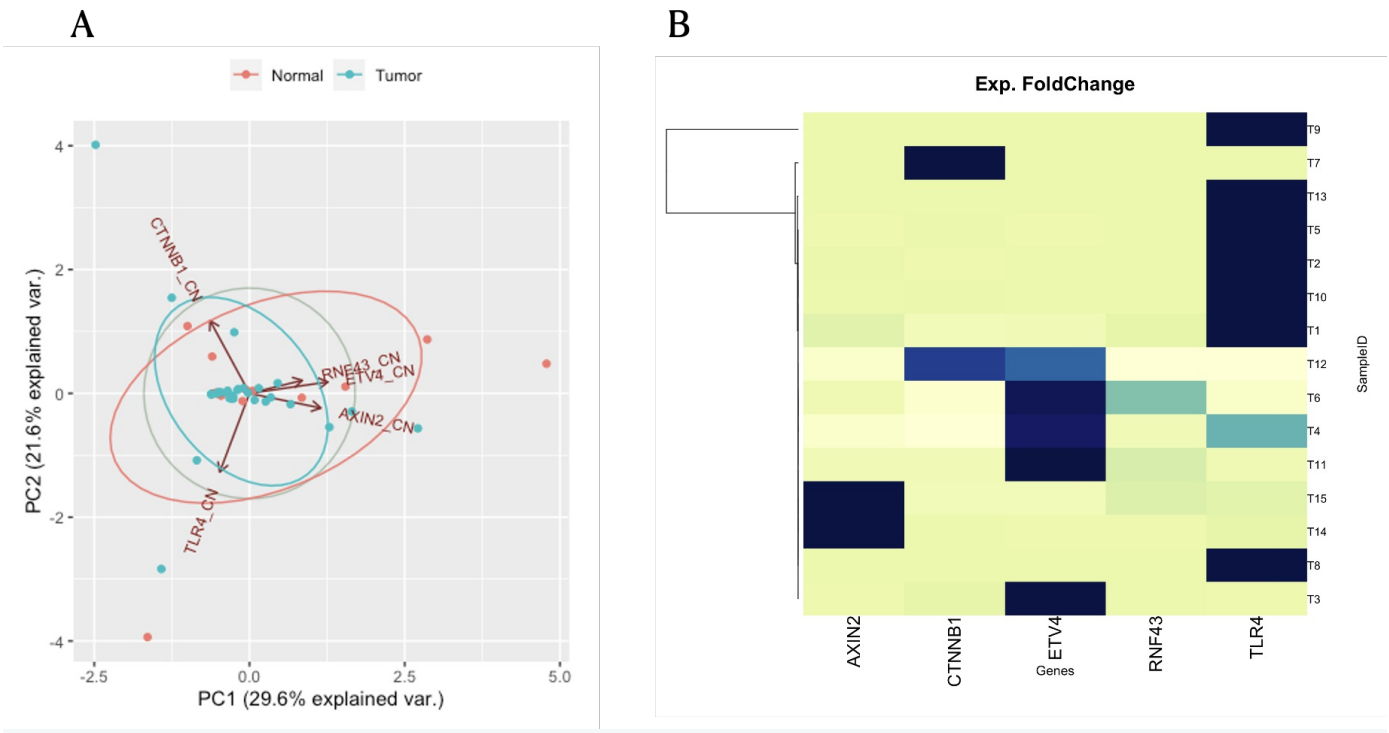

Figure 7. Expression fold change comparison. 7a: PCA of individuals compared against their expression levels. Note that the correlation trends are distinctly diverse for the three genes (*AXIN2*, *ETV4*, *RNF43*), *TLR4*,and *CTNNB1*. **7b**: The HCA of the variance of each sample across different target 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.

Discussion

In the current study, we have identified distinct categorizations of a subset of three genes,*AXIN2*, *ETV4*, and *RNF43* based on the variations in their expression in early-stage MSI-H CRC. Multiple comparisons of fold-changes in MSI vs. MSS group failed to provide a set of significant genes with high fold-change of > 2 (Figure 2a). However, the mean fold-change correlation between MSI and MSS group showed a distinct separation of three genes *AXIN2*, *ETV4*, and *RNF43* on one-side of the trendline of the correlation plot (Figure 2c). Surprisingly, on PPI network analysis of the top significant interaction set, *ETV4* and *RNF43* (genes below the trendline in Figure 2c) were less interacted with the other group of proteins (that were above the trendline in Figure 2c). Genes above the trendline in Figure 2c, namely, *COL1A1*, *SPP1*, *COMP*, *SFRP4*, *MMP3*, *GZMB*, etc., were clustered more closely with immune cells (Macrophages, Dendritic cells, NK cells), while *AXIN2*, *ETV4,* and *RNF43* were clustered with tumor cell purity. We also found differential clustering of these two set of genes in TIMER TCGA COAD data (Figure 5). Because of the peculiar grouping of these three genes among the 119 significant genes in the discovery cohort, from multiple sources of data comparison, these genes were selected for the validation study in the MSI-H set.

One of the major deficiencies of the study is the paucity of MSI-H specific samples in study cohorts. This is due to the lack of availability of early stage MSI-H CRC in the archived samples in the host institution. The other deficiency was the limited number of genes included in the Nanostring Pancancer Pathway analysis. Part of the reason for MSI-H in study

patients could be due to a congenital deficiency of DNA MMR gene expression such as Lynch syndrome. India, being a very populated country, with complex multi-ethnic cross-cultural population, identification of a distinct population with congenital MMR deficiency could be better stemmed from a more comprehensive population based large sequencing projects by exome or whole genome sequencing, which is beyond the scope of this study. Since this was a retrospective study on archived samples, the somatic mutational status of driver genes such as *BRAF*, *KRAS*, and *NRAS* was not available for the study patients. This is because the screening of these mutations requires more focused prospective clinical studies suited for therapeutic recommendations. But for comprehensive analysis of impact on MSI in the study, the role of these drivers could be critical. However, in the pilot Nanostring analysis, most of these driver genes were included in the probe set and their expression was also assessed along with the other genes included in the Pancancer pathway.

AXIN2 and *RNF43* are the negative regulators of the Wnt pathway, and their upregulation may suggest a Wnt ligand-independent (LI) pathogenesis [16]. Fusion mediated suppression of *RNF43*, by R-Spondin protein, facilitating the Frizzled protein to bind to Wnt ligand is associated with Wnt Ligand Dependent (LD) pathogenesis of CRC [17]. Similarly, epigenetic suppression of *AXIN2* was shown to be associated with Wnt LD pathogenesis[17]. Additionally, *AXIN2* is emerging to be an appropriate early biomarker for Wnt LI CRC with therapeutic potential [17]. *CTNNB1* (β-Catenin) is known to be upregulated in the canonical Wnt pathway (LI or LD) CRC [17]. In the current study, the variance in expression of *CTNNB1* is unrelated to the expression of *AXIN2*, *ETV4*, *RNF43* or *TLR4*, suggesting an overlapping role of β-Catenin in Wnt LD/LI MSI CRC.

As shown in Figure 7a, the expression of *ETV4* is positively correlated with that of *RNF43* and *AXIN2* in PC1. This may suggest its possible role in Wnt LI MSI CRC, which has not been previously explored. One of the major repressors of *ETV4* is a transcriptional regulator called Capicua (CIC), a High Mobility Group (HMG) box containing protein, and act as a tumor suppressor in colorectal cancers [18]. A previous study has shown that one of the HMG box proteins, HMGB1, causes DNA MMR deficiency [19]. HMGB1 is a cytokine released on lipopolysaccharide stimulation by colonic epithelial cells or normally by necrotic cells and acts via TLR4 receptor / MyD88 pathway inducing pro-inflammatory genes to block apoptosis in colon cancer [20][21]. Because of this relationship, TLR4 expression was taken into consideration in the validation cohort of the current study. In coherence with the previous studies, the variance in expression of both *ETV4* and *TLR4* are contrary to each other in the current study as shown in Figure 7. Furthermore, the relative expression of *TLR4* is negatively correlated to that of *RNF43* in PC1 and PC2 and to that of *AXIN2* in PC1. This may suggest that, with respect to the expression of *AXIN2*, *RNF43* and *ETV4*, the lower expression of *TLR4* could be a critical factor in deciding the Wnt ligand independence in MSI-H CRC pathogenesis. On the other hand, in the *TLR4* high-variant set as seen in Figure 7b, the contrasting expression of *TLR4* to that of *AXIN2*, *RNF43* and *ETV4* may indicate a Wnt ligand dependent MSI CRC. Based on the variation in the expression of these four genes, *AXIN2*, *ETV4*, *RNF43* and *TLR4*, the MSI-H group could be considered as a diverse group exhibiting different pathogenesis or at different stages of evolution of the cancer progression.

The *TLR4* high-variant set, having low variance in *AXIN2, RNF43* and *ETV4* expression, may reflect a better immune cell mounting, suggesting a better prognosis, at least for the MSI-H early-stage CRC. But previous studies have shown that *TLR4* expression is associated with progression of CRC irrespective of the MSI status[22]. To delineate this, more

frequency and mechanistic studies are required to understand the MMR deficiency occurring in *TLR4* upregulated colon cancer epithelial cells. Secondly, survival analysis on samples stratified based on *TLR4* expression and MSI status might be needed to understand the overall effect on the natural course. Similarly, indications from other studies suggest a better-to-poor prognosis for the Wnt LI CRC. This could be because the pathogenesis of ligand dependency (based on the expression profiles of *AXIN2* and *RNF43*) is spread across the of samples between CMS1 (good prognosis) and CMS4 (bad prognosis) subtypes of CRC, as explained earlier [16]. Because of this, a comparative expression study of these genes along with the MSI status, in a larger set of cases, might be helpful to understand how Wnt Ligand dependence contributes to the natural history and progression of the MSI-H early-stage CRC. In summary, expression studies of these four genes in an MSI stratified early-stage CRC might better help understand the underlying molecular pathogenesis of MSI in Indian early-stage CRC, which can help stratify cases for better treatment and prognosis outcomes.

Methods

Ethical Considerations

The Scientific Research and Institutional Ethics Committee reviewed and approved the study (IRB-AIMS-2017-124) before preparing samples from pathologically characterized archived specimens. The committee observed that no human subjects directly participated in the study, thus no requirement for a patient consent form. The tissue samples were anonymized for the compilation of archived data, for pilot and validation experiments, as well as for publication of results. The committee reviewed before, during, and after the completion of the study and approved the adherence to the ethical standards stipulated by the institutional ethics committee. The study was designed according to the consensus guidelines and statements for a molecular pathological biomarker explorative study [23][24].

Tissue selection and processing

Formalin-fixed paraffin-embedded (FFPE) tissue blocks from archived early-stage colon cancer tissues which were previously identified as high microsatellite instability (MSI-H) using Immunohistochemistry (IHC) and MSI-PCR methods during the previous five years were included in the study. There were separate FFPE blocks of tumor tissues and proximal or distal resected colon tissues, which were histologically verified. MSI-PCR further characterized these tissues to confirm their MSI status after obtaining the sections. FFPE blocks that did not contain normal non-tumor regions were removed from the validation study. Individual anonymized case details for the two arms of the study are included in Supplementary Dataset file S1. The explorative study included Nanostring nCounter Pan-Cancer pathway analysis in 11 archived tumor tissues and seven adjacent normal tissues. Subsequently, we selected another 15 tumor-normal paired archived specimens for the quantitative RT-PCR validation analysis.

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 [13]. The differentially expressed (DE) genes in the tumor were calculated by taking the geometric mean of the seven normal tissues as the denominator and the geometric mean of all 10 tumor samples as the numerator. The Fold Change (FC) of each gene was calculated from this ratio data, per the guidelines given by the Nanostring technologies, USA (MAN-C0011-04), ratio > 1 | FC = Ratio; ratio < 1 | FC = -1*(1/Ratio) [25]. The entire set of DE genes were imported into R-program to generate the principal component analysis (PCA) using ClustVis [26]. The Pathway analysis was done for each sample separately using DE genes (DE Call - Yes) using the WEB-based GEneSeTAnaLysis Toolkit (Web Gestalt) [27]. 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 [28]. The top 20 highly significant genes from the Nanostring expression analysis and their FC ($\log_2$) 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/) [29] and Tumor Immune Estimation Resource (TIMER 2.0: http://timer.cistrome.org/) [30][31]. 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 The Cancer Genome Atlas (TCGA), and Genotype-Tissue Expression (GTEx) [15][30]. The correlation of each of the selected genes in the test samples were analyzed for their co-expression correlation coefficient in the TCGA COAD subset of both these databases. The fold change values of the genes with $p < 0.05$ in the database 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

Tumor Immune Estimation Resource (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 [15] 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.

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, three to five 10µm sections were 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 15 mins 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, centrifugation 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 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 20ng 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 1 hr at 42°C and the reaction was then

terminated 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 that were identified specific to tumor purity and correlating with the MSI characteristics were validated by RT-PCR. Along with the selected genes, common genes associated with Colon cancer (β-Catenin, *CTNNB1*) and marker for immune cells (Toll-like receptor 4, *TLR4*) were also explored in the same tissues. The PCR primer sequences for the genes *AXIN2* [32], *ETV4* [33], *RNF43* [32], *TLR4* [34], and *CTNNB1* [35][36] were obtained from previous studies and are given in Table 1. *C1orf43* was taken as internal housekeeping gene control as this was found to be better than other housekeeping genes in colonic cells [35][36]. The primers were custom-synthesized and sourced from a commercial agency (SIGMA, Bengaluru INDIA). 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 is 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 15 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 [ln (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.

Gene	Forward	Reverse
AXIN2	5'-CAAACTTTCGCCAACCGTGGTTG	5'-GGTGCAAAGACATAGCCAGAACC
C1orf43	5'-AGCTCTGGATGCCATTCGTACC	5'-GTGTTTCGCAGATCCAGCAGGT
CTNNB1	5'-CACAAGCAGAGTGCTGAAGGTG	5'-GATTCCTGAGAGTCCAAAGACAG
ETV4	5'-AGGAACAGACGGACTTCGCCTA	5'-CTGGGAATGGTCGCAGAGGTTT
RNF43	5'-GGTTACATCAGCATCGGACTTGC	5'-ATGCTGGCGAATGAGGTGGAGT
TLR4	5'-CCCTGAGGCATTTAGGCAGCTA	5'-AGGTAGAGAGGTGGCTTAGGCT

Table 1. Primer sequences for the RT-PCR validation

Acknowledgments

This work has been supported by the ICMR (Grant number 33/7/2019), New Delhi, India, and Intramural Seed Grant by Amrita Vishwa Vidyapeetham (AVVP), Coimbatore Tamil Nadu to PA. We want to thank Ms. Reenu Anne Joy from Karkinos Healthcare Pvt Ltd, for handling of the discovery cohort samples, review of review and manuscript, Dr. Divya, Dr. Aditi, Dr. Pooja, Dr. Monica from the GI Oncopathology fellowship program, and Ms. Sheeba, Ms. Bindu, Ms. Nisha from the Department of Pathology, Amrita Institute of Medical Sciences, Kochi, for their valuable contribution on tissue handling and processing. Dr. Gopalakrishna Ramasamy from Theracues Private Ltd, Bangalore, for Nanostring analysis. Dr. Beena from OmicsGen Laboratory at Kakkanad, Kochi, for their contribution to the PCR processes. We would also like to thank Dr. Sreekumar Kannoth, Dr. Manu Raj, Mr. Sivakumar Venugopal, Mr. Karthikeyan, Dr. Apsy, and Ms. Sonu from the Department of Health Sciences Research at Amrita Institute of Medical Sciences for their valuable contribution on office proceedings and functioning of the project.

Author Contributions

PA has designed the experiments with VM and DMV. PA conducted molecular biology experiments. RRP has conducted histopathology and immunohistochemistry examination. PA compiled the results and analyzed them with DMV and VM. PA wrote the manuscript, and DMV, VM, KP, and RRP reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Data availability

The raw data and supplementary files are available at https://osf.io/tcdk9/

References

1. ^ *Yeole, B. B. Trends in cancer incidence in esophagus, stomach, colon, rectum and liver in males in India. Asian Pac J Cancer Prev 9, 97-100 (2008).*
2. ^ *NCRP. in Three Year Report of Population Based Cancer Registeries 2012-2014 National Cancer Registry Programme, (ed NCDIR) Ch. 2, 9-26 (NCDIR-NCRP, ICMR, 2016).*
3. a, b *Shakuntala, S. T. et al. Descriptive Epidemiology of Gastrointestinal Cancers: Results from National Cancer Registry Programme, India. Asian Pac J Cancer Prev 23, 409-418 (2022). https://doi.org:10.31557/APJCP.2022.23.2.409*
4. ^ *Collaborators, G. B. D. C. C. 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. Lancet Gastroenterol Hepatol 7, 627-647 (2022).*

https://doi.org:10.1016/S2468-1253(22)00044-9

5. ^*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 (2003). https://doi.org:10.1245/Aso.2003.12.014*
6. ^*Leary, R. J. et al. Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers. Proc Natl Acad Sci U S A 105, 16224-16229 (2008). https://doi.org:10.1073/pnas.0808041105*
7. ^*TCGA. Comprehensive molecular characterization of human colon and rectal cancer. Nature 487, 330-337 (2012). https://doi.org:10.1038/nature11252*
8. ^*Hause, R. J., Pritchard, C. C., Shendure, J. & Salipante, S. J. Classification and characterization of microsatellite instability across 18 cancer types. Nat Med 22, 1342-1350 (2016). https://doi.org:10.1038/nm.4191*
9. ^*Rajkumar, T. et al. Mutation analysis of hMSH2 and hMLH1 in colorectal cancer patients in India. Genet Test 8, 157-162 (2004). https://doi.org:10.1089/gte.2004.8.157*
10. ^*Pandey, V. et al. Assessment of microsatellite instability in colorectal carcinoma at an Indian center. Int J Colorectal Dis 22, 777-782 (2007). https://doi.org:10.1007/s00384-006-0241-3*
11. ^*Raman, R. et al. Evidence for possible non-canonical pathway (s) driven early-onset colorectal cancer in India. Mol Carcinog 53 Suppl 1, E181-186 (2014). https://doi.org:10.1002/mc.21976*
12. ^*Kanth, V. V. et al. Microsatellite instability and promoter hypermethylation in colorectal cancer in India. Tumour Biol 35, 4347-4355 (2014). https://doi.org:10.1007/s13277-013-1570-9*
13. [a, b]*Dunne, P. D. et al. Challenging the Cancer Molecular Stratification Dogma: Intratumoral Heterogeneity Undermines Consensus Molecular Subtypes and Potential Diagnostic Value in Colorectal Cancer. Clin Cancer Res 22, 4095-4104 (2016). https://doi.org:10.1158/1078-0432.CCR-16-0032*
14. ^*Ariyannur, P. S. et al. Pilot Nanostring PanCancer pathway analysis of colon adenocarcinoma in a tertiary healthcare centre in Kerala, India. Ecancermedicalscience 15, 1302 (2021). https://doi.org:10.3332/ecancer.2021.1302*
15. [a, b, c]*Buhard, O. et al. 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 (2006). https://doi.org:10.1200/jco.2005.02.7227*
16. [a, b]*Kleeman, S. O. & Leedham, S. J. Not All Wnt Activation Is Equal: Ligand-Dependent versus Ligand-Independent Wnt Activation in Colorectal Cancer. Cancers (Basel) 12, 3355 (2020). https://doi.org:10.3390/cancers12113355*
17. [a, b, c, d]*Kleeman, S. O. et al. Exploiting differential Wnt target gene expression to generate a molecular biomarker for colorectal cancer stratification. Gut 69, 1092-1103 (2020). https://doi.org:10.1136/gutjnl-2019-319126*
18. ^*Lee, J. S. et al. Capicua suppresses colorectal cancer progression via repression of ETV4 expression. Cancer Cell Int 20, 42 (2020). https://doi.org:10.1186/s12935-020-1111-8*
19. ^*Yuan, F., Gu, L., Guo, S., Wang, C. & Li, G. M. Evidence for involvement of HMGB1 protein in human DNA mismatch repair. The Journal of biological chemistry 279, 20935-20940 (2004). https://doi.org:10.1074/jbc.M401931200*
20. ^*Lotze, M. T. & Tracey, K. J. High-mobility group box 1 protein (HMGB1): nuclear weapon in the immune arsenal. Nat Rev Immunol 5, 331-342 (2005). https://doi.org:10.1038/nri1594*
21. ^*Makkar, S. et al. Hyaluronic Acid Binding to TLR4 Promotes Proliferation and Blocks Apoptosis in Colon Cancer. Mol*

Cancer Ther 18, 2446-2456 (2019). https://doi.org:10.1158/1535-7163.MCT-18-1225

22. ^*Yesudhas, D., Gosu, V., Anwar, M. A. & Choi, S. Multiple roles of toll-like receptor 4 in colorectal cancer. Front Immunol 5, 334 (2014). https://doi.org:10.3389/fimmu.2014.00334*
23. ^*Masucci, G. V. et al. Validation of biomarkers to predict response to immunotherapy in cancer: Volume I - pre-analytical and analytical validation. J Immunother Cancer 4, 76 (2016). https://doi.org:10.1186/s40425-016-0178-1*
24. ^*Duffy, M. J. et al. Validation of new cancer biomarkers: a position statement from the European group on tumor markers. Clin Chem 61, 809-820 (2015). https://doi.org:10.1373/clinchem.2015.239863*
25. ^*Nanostring. Gene Expression Data Analysis Guidelines, (2017).*
26. ^*Metsalu, T. & Vilo, J. ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. Nucleic Acids Res 43, W566-570 (2015). https://doi.org:10.1093/nar/gkv468*
27. ^*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 Res 45, W130-W137 (2017). https://doi.org:10.1093/nar/gkx356*
28. ^*Szklarczyk, D. et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 47, D607-D613 (2019). https://doi.org:10.1093/nar/gky1131*
29. ^*Tang, Z., Kang, B., Li, C., Chen, T. & Zhang, Z. GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. Nucleic Acids Res 47, W556-W560 (2019). https://doi.org:10.1093/nar/gkz430*
30. [a, b]*Li, T. et al. TIMER: A Web Server for Comprehensive Analysis of Tumor-Infiltrating Immune Cells. Cancer Res 77, e108-e110 (2017). https://doi.org:10.1158/0008-5472.CAN-17-0307*
31. ^*Li, T. et al. TIMER2.0 for analysis of tumor-infiltrating immune cells. Nucleic Acids Res 48, W509-w514 (2020). https://doi.org:10.1093/nar/gkaa407*
32. [a, b]*Pangestu, N. S., Chueakwon, P., Talabnin, K., Khiaowichit, J. & Talabnin, C. RNF43 overexpression attenuates the Wnt/beta-catenin signalling pathway to suppress tumour progression in cholangiocarcinoma. Oncol Lett 22, 846 (2021). https://doi.org:10.3892/ol.2021.13107*
33. ^*Dumortier, M. et al. ETV4 transcription factor and MMP13 metalloprotease are interplaying actors of breast tumorigenesis. Breast Cancer Res 20, 73 (2018). https://doi.org:10.1186/s13058-018-0992-0*
34. ^*Yang, H. et al. Reduced expression of Toll-like receptor 4 inhibits human breast cancer cells proliferation and inflammatory cytokines secretion. J Exp Clin Cancer Res 29, 92 (2010). https://doi.org:10.1186/1756-9966-29-92*
35. [a, b]*Janitschke, D. et al. 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 (2020). https://doi.org:10.3390/ijms21239015*
36. [a, b]*Xu, L. et al. 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 (2019). https://doi.org:10.1186/s12881-019-0867-y*