Published in final edited form as:

Genes Chromosomes Cancer. 2015 September ; 54(9): 527–541. doi:10.1002/gcc.22265.

Gene Expression in Colon Cancer: A Focus on Tumor Site and Molecular Phenotype

Martha L. Slattery*, Daniel F. Pellatt, Lila E. Mullany, Roger K. Wolff, and Jennifer S. Herrick

Department of Internal Medicine, University of Utah, 383 Colorow, Salt Lake City, Utah 84117

Abstract

Hundreds to thousands of genes are differentially expressed in tumors when compared to nontumor colonic tissue samples. We evaluated gene expression patterns to better understand differences in colon cancer by tumor site and tumor molecular phenotype. We analyzed RNA-seq data from tumor/normal paired samples from 175 colon cancer patients. We implemented a cross validation strategy with nonparametric tests to identify genes which displayed varying expression characteristics related to paired tumor/nontumor tissue across proximal and distal colon sites and by tumor molecular phenotypes, that is, *TP53*, *KRAS*, CpG Island Methylator Phenotype (CIMP), and microsatellite instability (MSI). We used Ingenuity Pathway Analysis (IPA) to determine networks associated with deregulated genes in our data. Genes showed significant differences in expression characteristics at the 0.01 level in both validation groups between tumor subsite (116 genes), CIMP high versus CIMP low (79 genes), MSI versus microsatellite stable (MSS) (49 genes), *TP53*-mutated versus not mutated (17 genes), and *KRAS*-mutated versus not mutated (1 gene). Deregulated genes for CIMP high and MSI tumors were often down-regulated. In contrast to CIMP high and MSI tumors, genes that were deregulated in *TP53* were likely to be up-regulated. *ERK1*, *WNT*, growth factors and inflammation-related factors were focal points of both CIMP and MSI IPA networks. The MUC family of genes was up-regulated MSI networks. Numerous genes showed differences in expression between proximal and distal tumors, nontumor proximal and distal tissue, and tumor molecular phenotype. Deregulated mucin genes appear to play an important role in MSI tumors.

INTRODUCTION

Studies of colorectal cancer (CRC) have shown that hundreds to thousands of genes are differentially expressed in tumors when compared to normal tissue samples (Birkenkamp-Demtroder et al., 2002). Nannini et al. cited the benefit of gene expression studies as a way to better understand the carcinogenic process (Nannini et al., 2009). Studies have used gene expression data to classify tumor phenotypes (Sanz-Pamplona et al., 2012; Budinska et al., 2013; Burgess, 2013) as well as to identify genes that are uniquely expressed with microsatellite instability (MSI) (Sanz-Pamplona et al., 2011, 2012). Some studies have utilized normal or nonmutated tissue for comparison to tumor gene expression, although

*Correspondence to: Dr. Slattery; Department of Internal Medicine, University of Utah, 383 Colorow, Salt Lake City, Utah 84117. marty.slattery@hsc.utah.edu.

Additional Supporting Information may be found in the online version of this article.

many have compared tumor sites to each other without consideration for normal tissue (Alon et al., 1999; Notterman et al., 2001; Birkenkamp-Demtroder et al., 2005; Komuro et al., 2005; Kleivi et al., 2007; Lin et al., 2007). A study by Giacomini used 18 colon cancer cell lines and 61 colorectal tumors to profile gene expression and suggested that a distinct molecular signature was noticeable between MSI and micro-satellite stable (MSS) samples (Giacomini et al., 2005). Studying the molecular signatures of tumors by phenotype may provide clues into novel and specific treatment modalities.

In this study, we perform gene expression analysis using RNAseq data from tumor and adjacent nontumor tissue from individuals diagnosed with colon cancer. The goal of our study was to identify genes that were differentially expressed uniquely by tumor molecular phenotype or tumor subsite (i.e., proximal vs. distal) within the colon. We utilize various bioinformatics tool to gain insight into the carcinogenic process and unique pathways and networks in which various tumor molecular phenotypes are involved.

MATERIALS AND METHODS

RNA was available from 175 tumor and nontumor tissue pairs who were part of the Diet, Activity, and Lifestyle study, which is an incident, population-based, case-control study of colon cancer conducted in Utah, the Kaiser Permanente Medical Research Program (KPMRP), and the Twin Cities Metropolitan area. Tumor tissue for RNA extraction was available from the Utah and KPMRP sites. Cases had to have tumor registry verification of a first primary adenocarcinoma of the colon and diagnosed between October 1991 and September 1994 to be eligible for the study. Tumor tissue was obtained for 97% of all Utah cases and for 85% of all KPMRP cases (Slattery et al., 2000) and included those who signed informed consent and those retrieved by local tumor registries and sent to study investigators without personal identifiers. Individuals with known adenomatous polyposis coli (APC), Crohn's disease, or inflammatory bowel disease were not eligible for the study. Individuals with MSI high tumors were sequenced for inherited mutations in mismatch repair genes and excluded from the study if such mutations existed (Samowitz et al., 2001). The study was approved by the Institutional Review Board of the University of Utah and at KPMRP.

We have previously assessed these tumor samples for *TP53* and *KRAS* mutations, the CpG island methylator phenotype (CIMP) using the classic panel (Samowitz et al., 2005), and MSI and MSS based on the mononucleotides *BAT26* and *TGF β RII* and a panel of 10 tetranucleotide repeats that were correlated highly with the Bethesda Panel (Slattery et al., 2000); our study was done prior to the Bethesda Panel development. The classic CIMP panel consisted of five markers, *hMLH1*, *p16*, and MINT1, MINT2, and MINT31. Tumors were scored as CIMP high if two or more of the CpG islands were methylated otherwise they were classified as CIMP low.

RNA Processing

Total RNA was extracted from formalin-fixed paraffin embedded tissues. We assessed slides and tumor blocks that were prepared over the duration of the study prior to the time of RNA isolation to determine their suitability. Older slides produced RNA quality comparable with that of more recent slides; quality was not correlated with time lapse between slide

preparation and RNA preparation. The study pathologist reviewed slides to delineate tumor and nontumor tissue. Cells were dissected from 1 to 4 sequential sections on aniline blue stained slides using an H&E slide for reference. Total RNA was extracted, isolated, and purified using the RecoverAll Total Nucleic Acid isolation kit (Ambion); RNA yields were determined using a NanoDrop spectrophotometer.

Sequencing Library Preparation

Library construction was performed using the Illumina TruSeq Stranded Total RNA Sample Preparation Kit with Ribo-Zero. Briefly, Ribosomal RNA was removed from 100 ng total RNA using biotinylated Ribo-Zero oligos attached to magnetic beads that are complimentary to cytoplasmic rRNA. Following purification, the rRNA-depleted sample is fragmented with divalent cations under elevated temperatures and primed with random hexamers in preparation for cDNA synthesis. First strand reverse transcription is accomplished using Superscript II Reverse Transcriptase (Invitrogen). Second strand cDNA synthesis is accomplished using DNA polymerase I and Rnase H under conditions in which dUTP is substituted for dTTP, yielding blunt-ended cDNA fragments in which the second strand contains dUTP. An A-base is added to the blunt ends as a means to prepare the cDNA fragments for adapter ligation and block concatamer formation during the ligation step. Adapters containing a T-base overhang were ligated to the A-tailed DNA fragments. Ligated fragments were PCR-amplified (13 cycles) under conditions in which the PCR reaction enables amplification of the first strand cDNA product, whereas attempted amplification of the second strand product stalls at dUTP bases and therefore is not represented in the amplified library. The PCR-amplified library was purified using Agencourt AMPure XP beads (Beckman Coulter Genomics).

Sequencing and Data Processing

Sequencing libraries (18 pM) were chemically denatured and applied to an Illumina TruSeq v3 single read flow cell using an Illumina cBot. Hybridized molecules were clonally amplified and annealed to sequencing primers with reagents from an Illumina TruSeq SR Cluster Kit v3-cBot-HS. Following transfer of the flowcell to an Illumina HiSeq instrument, a 50 cycle single-read sequence run was performed using TruSeq SBS v3 sequencing reagents. The single-end 50-base reads from the Illumina HiSeq2500 were aligned to a sequence database containing the human genome (build GRCh37/hg19, February 2009, from genome.ucsc.edu) plus all splice junctions generated using the USeq MakeTranscriptome application (version 8.8.1, available here: <http://useq.sourceforge.net/>). Alignment was performed using novoalign version 2.08.01 available from novocraft.com, which also trimmed any adapter sequence. Following alignment, genome alignments to splice junctions were translated back to genomic coordinates using the USeq SamTranscriptomeParser application. The resulting alignments were sorted and indexed using the Picard SortSam application (version 1.100, available here: <http://broadinstitute.github.io/picard/>). Aligned read counts for each gene were calculated using pysam (<https://code.google.com/p/pysam/>) and samtools (<http://samtools.sourceforge.net/>). A python script using the pysam library was given a list of the genome coordinates for each gene, and counts to the exons and UTRs of those genes were calculated. Gene coordinates were downloaded from <http://genome.ucsc.edu>.

We compared our data to a gene table that included 51,041 molecular features. We dropped features that were not expressed in our data or for which the expression was missing for the majority of samples. Using the BioMart tool on the Ensembl website (<http://www.ensembl.org>), we created a list of known regions linked to protein-coding genes from the human GRCh38 gene annotation dataset. Our final analysis included only the 17,384 features involved in protein coding.

Statistical Methods

Of the 197 initial tumor/nontumor tissue pairs, five subjects failed quality control (QC) based on low number of sequence counts for both tumor and nontumor tissue, and 17 were excluded because either the nontumor colonic tissue or its paired tumor tissue failed QC, leaving 175 subjects with high quality data for inclusion in the analysis. From this pool of subjects we randomly assigned people to group “A” or “B.” For each such group, the following analysis was performed. For each protein-coding gene, we measured an individual’s associated difference in expression between tumor and paired nontumor colonic tissue, subsequently referred to as differential expression, using the RPKM (Reads per Kilobase per Million Reads) data set that accounts for sequencing depth and gene size. We evaluated disparities in differential expression between sub-populations characterized by tumor site (proximal vs. distal) and tumor molecular phenotype (CIMP high vs. CIMP low; MSI vs. MSS; *TP53*-mutated vs. *TP53*-not mutated; and *KRAS*-mutated vs. *KRAS*-not-mutated) using the Wilcoxon–Mann–Whitney nonparametric test *P* value as implemented in SAS 9.4 (SAS Institute, Cary, NC). This test is sensitive to differences in location which would indicate not only a disparity in differential expression between subpopulations but also, by definition of differential expression, a difference in gene expression level between tumor and non-tumor colonic tissue in at least one of the considered subpopulations. *P* values were calculated for each protein-coding gene for each comparison between subpopulations. Similarly, the Wilcoxon–Mann–Whitney test was utilized to identify genes with significant variation in nontumor tissue expression between proximal and distal sites. To reduce Type I errors cross validation was used. A disparity in differential expression between subpopulations or in nontumor tissue expression level by site was considered significant only if the corresponding *P* values were significant both in groups “A” and “B.” We utilized a *P* value of 0.01 in both groups for general significance, although we incorporate all genes that were significant at the 0.05 level in both groups to have more information to enrich pathway analysis using bio-informatics tools.

To help describe the data, we calculated fold change values for genes that displayed significant disparities in differential expression between sub-populations. Within a given subpopulation, the fold change of a gene was defined as the ratio of its mean tumor expression to its mean nontumor expression within that subpopulation for the entire data set (groups “A” and “B” combined). As RPKM data are non-negative, a fold change greater than one indicates a positive differential expression (i.e., up-regulated) while a fold change between zero and one indicates a negative differential expression (i.e., down-regulated). To visualize the disparities in differential expression between subpopulations we created heat maps. Each heat map features the log₂ transformation of the fold changes associated with genes identified as having significant variation in differential expression between two

subpopulations. We restricted our heat maps to those genes that had a significant P value of 0.01 between the two groups. Our heat maps were created using the heatmap.2 program in the “gplots” package of R (<http://cran.r-project.org>). Distance between two vectors of log2 transformed fold changes was measured via the Euclidean metric and complete linkage was selected for this programs’ agglomerative hierarchical clustering algorithm.

Further, analysis was performed on the list of Ensemble IDs associated with genes whose differential expressions were found to vary significantly ($P < 0.05$) by subpopulation as characterized by tumor sites or molecular phenotype using QIAGEN’s Ingenuity Pathway Analysis (IPA) (2014). In identifying networks and pathways in which our genes were enriched, we used included genes from Ingenuity Knowledge Base using the option for both indirect and direct relationships. We used the IPA defaults of 35 molecules per network and 25 networks per analysis, to construct networks. All data sources were used and data sources included experimentally observed relationships. We applied the Benjamini–Hochberg (B-H) multiple testing correction to assess pathways in IPA. Of the genes deregulated in our data, 16 genes (*SPON1*, *QPRT*, *AC112721.1*, *HIST1H3A*, *PAH*, *ACSL6*, *ZBTB8B*, *AC129492.6*, *FAM188B2*, *SLC6A14*, *C7orf13*, *FOXD1*, *APLF*, *GLYATL1*, *RP11-451M19.3*, and *AC129492.6*.) did not map to annotations in IPA and were not included in the bioinformatics analysis.

RESULTS

Of these tumors, approximately 48% were proximal and 52% were distal in the colon and were similar for both Group A and B (Table 1). Evaluation of tumor molecular phenotype showed that 25.7% were CIMP high, 18.3% were MSI, 27.4% were *KRAS*-mutated, and 44.0% were *TP53*-mutated. The average age of the study participants was 65.2 years. Groups A and B were similar for most variables, although Group B had more *TP53*-mutated tumors and Group A had slightly more of the other tumor molecular phenotypes.

A total of 320 genes showed significant changes in differential expression between proximal and distal tumor subsites in both Group A and B at the 0.05 level and 116 showed significant differences at the 0.01 level (Supporting Information Table 1 shows those genes differentially expressed at the 0.01 level (Supporting Information Table 1). Many of these genes showed a striking difference in expression level between tumor and nontumor tissue, the nature of which varied by tumor location. This is illustrated by the differences in fold change values (and by differences in the associated log2 transformations) between proximal and distal tumors. (Fig. 1 Heat Map of log2 transformed fold changes of genes with disparities in differential expression between tumor sites identified at the 0.01 level). Of the 116 genes that were significantly different at the <0.01 level between tumor and normal by tumor subsite, 33 of these genes (28.6%) showed significant differences in nontumor tissue by tumor subsite. A total of 77 genes had P values of <0.01 in groups A and B for nontumor tissue (Table 2).

Assessment by tumor molecular subtype identified 250 genes where differential expression varied significantly between CIMP-high and CIMP-low tumor types at the 0.05 level and 79 were significantly differentially expressed in both groups “A” and “B” at the 0.01 level.

Similarly, 187 genes showed significant variation between MSS versus MSI tumors at the 0.05 level and 49 showed differences at the 0.01 level while 114 and 23 genes, respectively, showed significant variation in differential expression between *TP53*-mutated versus nonmutated and *KRAS*-mutated versus nonmutated tumor types (17 and 1 at the 0.01 level). Supporting Information Tables 2–4 show the mean normal RPKM expression levels, mean tumor RPKM expression levels, and associated fold changes of by tumor molecular phenotype associated with these genes. Figure 2 includes heat maps of the same type as that in Figure 1 for genes that displayed significant variation in differential expression across these molecular phenotype groups at the 0.01 level in both groups “A” and “B.” (Note *KRAS*-mutated tumors are not included as only one gene was significant at the 0.01 level). As shown in the heat maps, segmenting the population into phenotype categories CIMP-high vs. CIMP-low and MSI vs. MSS identified the greatest number of genes with variable differential expression across phenotype categories. It is interesting to note the extent of overlap in differentially expressed genes for MSS versus MSI and CIMP-high vs. CIMP-low tumors. Twenty-nine of the 49 genes that were differentially expressed at the 0.01 level for MSI were also differentially expressed in CIMP (out of 79 genes). Most genes that display variable differential expression across phenotype subcategories were identified via CIMP high or MSI partitions; such genes displayed similar over- and under-expression patterns. Deregulated genes for CIMP high and MSI tumors were often down-regulated as indicated by the green coloring. In contrast, to CIMP high and MSI tumors, genes that were deregulated in *TP53* were likely to be up-regulated as shown in the red coloring.

We further evaluated genes that were deregulated in CIMP high and MSI tumors using IPA. Several networks were identified that were significantly enriched by genes that displayed variation in differential expression between tumor sites or molecular phenotype categories in our data. As shown in Table 3, 11 networks had an enrichment score of 20 or more for CIMP high tumors while there were eight networks with an enrichment score of 20 or more for MSI tumors (Table 4); enrichment values at that level are considered meaningful by IPA. These networks were enriched by genes which displayed significant disparities in differential expression by subpopulations in our data and are denoted as focus molecules in IPA. We illustrate the top three networks to better understand the relationships between genes in the networks and the large number of genes that are down-regulated in CIMP high (Figs. 3A, 3B, and 3C) and MSI tumors (Figs. 4A, 4B, and 4C). While Network 1 of both CIMP and MSI had genes that were both up (shown in red) and down-regulated (shown in green), Networks 2 and 3 were comprised primarily of genes that were down-regulated in CIMP high and MSI tumors. *ERK1* and *WNT* were focal points of both CIMP Network 1 (Fig. 3A) and MSI Network 1 (Fig. 4A). The *MUC* family of genes appeared to be more up-regulated in that pathway for MSI than for CIMP. For both Network 2 of CIMP and MSI, *P13K*, *JNK*, *Gpcr*, were central points, while *AMPK* and *VEGFA* were in the CIMP pathways, and *NFkB* complex and *MAPK* were central junctions within the MSI Network 2. Both CIMP and MSI Network 3, included growth factors and inflammation-related factors including insulin, *TGFB*, LDL, and *IL1* at key intercepts, while CIMP Network 3 had histone as a central intercept.

DISCUSSION

These data illustrate differences in gene expression between tumor and nontumor colonic tissue by tumor site as well as by tumor molecular phenotype. Although we observed differences in gene expression between colonic sites in nontumor tissue, we also identified significant differential expression by tumor subsite after taking into account expression levels in nontumor tissue. The majority of genes where tumor and nontumor expression levels varied significantly by tumor molecular phenotype were those identified by subgroups characterized by CIMP tumor type and those identified by subgroups characterized by MSI versus MSS tumor type. These genes tended to be down-regulated among CIMP high and MSI tumors.

Others have evaluated differences in gene expression by tumor subsite using different approaches from those described here. The major differences stem from sample size and comparison groups used to assess site-specific associations. The majority of studies have included few cases and few genes, making our sample of 175 one of the largest, if not the largest set of cases analyzed. Our larger sample size has given us several analytic strengths, in that we have been able to split our sample into two groups, so that cross validation of findings was possible. This also has given us greater precision in testing associations. While some studies have normal tissue for comparison, this is usually limited in numbers and not paired samples as we have here. Nonpaired samples often result in both right and left-sided tumors being compared to the same normal tissue, irrespective of tumor site (Birkenkamp-Demtroder et al., 2002; Komuro et al., 2005). Given the number of genes that were shown to have significant differences between proximal and distal tumor site, lack of consideration of tumor site would be problematic when determining differential expression with tumors. It is unclear if site differences without consideration of the underlying nontumor tissue expression are informative about the tumor differential expression or just normal tissue expression. Our results show the necessity of conducting paired samples. It is also possible that genetic, lifestyle, or dietary factors could influence expression that is site-specific; using paired samples controls for these potential lifestyle differences as the normal and tumor samples are from the same individual.

CIMP high and MSI tumors showed the greatest differential gene expression when compared to CIMP low or MSS tumors. However, CIMP high and MSI tumors were often less down-regulated than CIMP low or MSS tumors. As expected there were more similarities in differential gene expression patterns between these two phenotypes than for *TP53*-mutated and *KRAS*-mutated tumors. *KRAS*-mutated tumors had the fewest genes with differential expression levels that varied significantly compared to *KRAS*-nonmutated tumors. Examining the networks in which these genes were located could provide further insight into possible etiologic differences and disease pathways associated with these phenotypes. Networks associated with CIMP and MSI were most informative as we observed a greater number of genes with expression characteristics that varied significantly between high and low tumor phenotype groups.

Decreased gene expression would be expected in methylated tumors such as those designated CIMP high. What is less clear is if the decreased gene expression is the result of

general phenomena of methylation or is linked into specific genes and pathways. We used IPA to identify networks in which genes that were deregulated in our data were significantly enriched to better understand potentially important networks. Evaluation of the top CIMP high networks of genes whose expression characteristics vary significantly from those of CIMP low tumors shows disruption of numerous pathways. Major contributors to CIMP Network 1 evolved around *WNT*. It included *WIF1* (Wnt Inhibitory Factor 1), and other genes whose functions are related to cell growth and differentiation such as *CDX2* (caudal type homeobox2), *MSI1* (musashi RNA-binding protein), *PITX1* (paired-like homeodomain), *SOX1* (Sex Determining Region Y-Box1), and *AXIN2* (axin-like protein 2). *AXIN2* have previously been associated with CIMP in colorectal cancer (Belshaw et al., 2008) where it was down-regulated in tumor compared to normal tissue. In our data, *AXIN2* was upregulated in CIMP high tumors. In Network 2, we observed down-regulation of several G protein coupled receptors (Gpcr) that sense molecules outside of the cell and activate signal transduction pathways for cellular response. The solute carrier family (SLC) of genes that code membrane transport proteins also was down-regulated. CIMP Network 3 had numerous down-regulated genes that influenced inflammation-related factors, *TGFB*, and cholesterol; this may represent somewhat more global inactivation of genes.

MSI tumors have been studied more extensively with gene expression than other tumor phenotypes (Giacomini et al., 2005; Pastrello et al., 2005; Duldulao et al., 2012). The study by Giacomini et al. evaluated approximately 21,000 genes using 18 colon cancer cell lines and 61 colorectal tumors and compared expression difference in tumors that were MSI and MSS; no normal tissue was used (Giacomini et al., 2005). They identified 217 genes with a false discovery rate of <0.1 that were different between eight MSI and 10 MSS samples. As normal tissue was not used some of the differences could be from site differences as most MSI tumors are located in the proximal colon. However there was overlap in several genes identified and what has previously been reported in the literature, most notably the metallothionein genes *MT1X* and *MT2A*. These genes were upregulated in our data and contributed to MSI Network 1. Most investigators have focused on specific genes that are differentially expressed between MSI and MSS tumors (Pastrello et al., 2005; Duldulao et al., 2012). Another group of genes that were previously identified as being up-regulated in MSI tumors were mucins (Pastrello et al., 2005), which are a family of genes that are involved in cell adhesion and metastasis. Mucinous colorectal cancers have mucin overexpression and *MUC2* and *MUC5AC* have been shown to be up-regulated in MSI tumors (Biemer-Huttmann et al., 2000; Pastrello et al., 2005). In our data *MUC2*, *MUC5AC*, *MUC4*, and *MUC5B* were all highly up-regulated in MSI tumors and comprised a major component of MSI Network 1. As in CIMP Network 2, MSI Network 2 involved several G-coupled protein receptors, while MSI Network 3 had several down-regulated genes that were associated with inflammation, insulin, and energy-related pathways.

As shown in Figure 5, our data clearly illustrate the overlap in gene expression patterns between CIMP high and CIMP low, MSI and MSS, and proximal and distal tumor location. Almost all of the genes differentially expressed between MSI and MSS tumors overlapped with differentially expressed genes between CIMP high and CIMP low and between tumor subtypes. There were more differences in gene expression between CIMP high and CIMP low tumors, that were unique than we observed for MSI tumors. As we previously stated,

tumor location differences in gene expression patterns contributed to differences by tumor phenotype. These patterns of overlapping deregulated gene expression illustrate similarities and differences in molecular phenotype profiles.

There are many strengths and potential limitations in our study. As we previously mentioned, our sample was large enough to cross validate findings within our study. However, we still made comparisons between multiple subpopulations on a large number of genes. While cross validation is a powerful tool in protecting against type I error, some of our findings could be from chance. Strength of our study was that we had gene expression data on paired tumor and nontumor tissue for each individual in our data set. This allowed for direct accounting of each individual's differential expression rather than comparing tumor tissue with nontumor tissue that may not have the same site or molecular phenotype as can occur with nonpaired data. We also had tumor molecular phenotype data that allowed us to evaluate expression differences by tumor type and we were able to explore networks as well as individual genes to improve our understanding of the overall carcinogenic process.

In conclusion, we identified numerous genes with expression characteristics that varied in proximal and distal tumors as well as nontumor colonic tissue. We also identified genes whose tumor and nontumor expression characteristics varied by molecular phenotype; more genes were identified among CIMP high and MSI tumors than either *TP53* or *KRAS*. The majority of genes identified among CIMP high and MSI tumors were down-regulated. Deregulated mucin genes appear to play an important role in MSI tumors. Overlap in deregulated genes between CIMP high, MSI, and colonic subsite re-enforces the similarity in genes shared by these pathways.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The contents of this manuscript are solely the responsibility of the authors and do not necessarily represent the official view of the National Cancer Institute. We would like to acknowledge the contributions of Dr. Bette Caan, Judy Morse and Donna Schaffer and the Kaiser Permanente Medical Research Program, and Sandra Edwards for data collection and organization, Erica Wolff and Michael Hoffman for RNA extraction, Wade Samowitz for slide review, and Brett Milash at the Bioinformatics Core Facility at the University of Utah.

Supported by: NCI; Grant number: CA48998.

References

- Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci USA*. 1999; 96:6745–6750. [PubMed: 10359783]
- Belshaw NJ, Elliott GO, Foxall RJ, Dainty JR, Pal N, Coupe A, Garg D, Bradburn DM, Mathers JC, Johnson IT. Profiling CpG island field methylation in both morphologically normal and neoplastic human colonic mucosa. *Br J Cancer*. 2008; 99:136–142. [PubMed: 18542073]
- Biemer-Huttmann AE, Walsh MD, McGuckin MA, Simms LA, Young J, Leggett BA, Jass JR. Mucin core protein expression in colorectal cancers with high levels of microsatellite instability indicates a novel pathway of morphogenesis. *Clin Cancer Res*. 2000; 6:1909–1916. [PubMed: 10815915]

- Birkenkamp-Demtroder K, Christensen LL, Olesen SH, Frederiksen CM, Laiho P, Aaltonen LA, Laurberg S, Sorensen FB, Hagemann R, TF OR. Gene expression in colorectal cancer. *Cancer Res.* 2002; 62:4352–4363. [PubMed: 12154040]
- Birkenkamp-Demtroder K, Olesen SH, Sorensen FB, Laurberg S, Laiho P, Aaltonen LA, Orntoft TF. Differential gene expression in colon cancer of the caecum versus the sigmoid and rectosigmoid. *Gut.* 2005; 54:374–384. [PubMed: 15710986]
- Budinska E, Popovici V, Tejpar S, D'Ario G, Lapique N, Sikora KO, Di Narzo AF, Yan P, Hodgson JG, Weinrich S, Bosman F, Roth A, Delorenzi M. Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. *J Pathol.* 2013; 231:63–76. [PubMed: 23836465]
- Burgess DJ. Gene expression: Colorectal cancer classifications. *Nat Rev Cancer.* 2013; 13:380–381.
- Duldulao MP, Lee W, Le M, Chen Z, Li W, Wang J, Gao H, Li H, Kim J, Garcia-Aguilar J. Gene expression variations in microsatellite stable and unstable colon cancer cells. *J Surg Res.* 2012; 174:1–6. [PubMed: 21816436]
- Giacomini CP, Leung SY, Chen X, Yuen ST, Kim YH, Bair E, Pollack JR. A gene expression signature of genetic instability in colon cancer. *Cancer Res.* 2005; 65:9200–9205. [PubMed: 16230380]
- Kleivi K, Lind GE, Diep CB, Meling GI, Brandal LT, Nesland JM, Myklebost O, Rognum TO, Giercksky KE, Skotheim RI, Lothe RA. Gene expression profiles of primary colorectal carcinomas, liver metastases, and carcinomatoses. *Mol Cancer.* 2007; 6:2. [PubMed: 17201907]
- Komuro K, Tada M, Tamoto E, Kawakami A, Matsunaga A, Teramoto K, Shindoh G, Takada M, Murakawa K, Kanai M, Kobayashi N, Fujiwara Y, Nishimura N, Hamada J, Ishizu A, Ikeda H, Kondo S, Katoh H, Moriuchi T, Yoshiki T. Right- and left-sided colorectal cancers display distinct expression profiles and the anatomical stratification allows a high accuracy prediction of lymph node metastasis. *J Surg Res.* 2005; 124:216–224. [PubMed: 15820251]
- Lin HM, Chatterjee A, Lin YH, Anjomshoa A, Fukuzawa R, McCall JL, Reeve AE. Genome wide expression profiling identifies genes associated with colorectal liver metastasis. *Oncol Rep.* 2007; 17:1541–1549. [PubMed: 17487416]
- Nannini M, Pantaleo MA, Maleddu A, Astolfi A, Formica S, Biasco G. Gene expression profiling in colorectal cancer using microarray technologies: Results and perspectives. *Cancer Treat Rev.* 2009; 35:201–209. [PubMed: 19081199]
- Notterman DA, Alon U, Sierk AJ, Levine AJ. Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. *Cancer Res.* 2001; 61:3124–3130. [PubMed: 11306497]
- Pastrello C, Santarosa M, Fornasarig M, Sigon R, Perin T, Giannini G, Boiocchi M, Viel A. MUC gene abnormalities in sporadic and hereditary mucinous colon cancers with microsatellite instability. *Dis Markers.* 2005; 21:121–126. [PubMed: 16276005]
- QIAGEN's Ingenuity Pathway Analysis. 2014.
- Samowitz WS, Albertsen H, Herrick J, Levin TR, Sweeney C, Murtaugh MA, Wolff RK, Slattery ML. Evaluation of a large, population-based sample supports a CpG island methylator phenotype in colon cancer. *Gastroenterology.* 2005; 129:837–845. [PubMed: 16143123]
- Samowitz WS, Curtin K, Lin HH, Robertson MA, Schaffer D, Nichols M, Gruenthal K, Leppert MF, Slattery ML. The colon cancer burden of genetically defined hereditary nonpolyposis colon cancer. *Gastroenterology.* 2001; 121:830–838. [PubMed: 11606497]
- Sanz-Pamplona R, Berenguer A, Cordero D, Riccadonna S, Sole X, Crous-Bou M, Guino E, Sanjuan X, Biondo S, Soriano A, Jurman G, Capella G, Furlanello C, Moreno V. Clinical value of prognosis gene expression signatures in colorectal cancer: A systematic review. *PLoS One.* 2012; 7:e48877. [PubMed: 23145004]
- Sanz-Pamplona R, Cordero D, Berenguer A, Lejbkiewicz F, Rennert H, Salazar R, Biondo S, Sanjuan X, Pujana MA, Rozek L, Giordano TJ, Ben-Izhak O, Cohen HI, Trougouboff P, Bejhar J, Sova Y, Rennert G, Gruber SB, Moreno V. Gene expression differences between colon and rectum tumors. *Clin Cancer Res.* 2011; 17:7303–7312. [PubMed: 21976543]
- Slattery ML, Curtin K, Anderson K, Ma KN, Ballard L, Edwards S, Schaffer D, Potter J, Leppert M, Samowitz WS. Associations between cigarette smoking, lifestyle factors, and microsatellite instability in colon tumors. *J Natl Cancer Inst.* 2000; 92:1831–1836. [PubMed: 11078760]

Slattery ML, Edwards SL, Palmer L, Curtin K, Morse J, Anderson K, Samowitz W. Use of archival tissue in epidemiologic studies: Collection procedures and assessment of potential sources of bias. *Mutat Res.* 2000; 432:7–14. [PubMed: 10729707]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1.

Heat Map of differential gene expression between distal and proximal tumors. Genes in green are down-regulated while genes in red are up-regulated.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 2. Heat Map of differential gene expression by tumor molecular phenotype. (A) CIMP high vs. CIMP low; (B) MSI vs. MSS. Figure 2C. *TP53*-mutated versus *TP53*-nonmutated. Genes in green are down-regulated while genes in red are up-regulated.

Figure 3.

Top Networks in CIMP high identified through IPA networks. Genes in green are down-regulated while those in red are up-regulated. The intensity of the color of the nodes (genes) indicates the degree of up-regulation. Genes uncolored or gray were not differentially expressed in the dataset but were integrated into the computationally generated networks based on stored information in IPA knowledge memory suggesting its relevance to the network.

Figure 4.

Top Networks in MSI tumors identified through IPA networks. Genes in green are down-regulated while those in red are up-regulated. The intensity of the color of the nodes (genes) indicates the degree of up-regulation. Genes uncolored or gray were not differentially expressed in the dataset but were integrated into the computationally generated networks based on stored information in IPA knowledge memory suggesting its relevance to the network.

Figure 5.
Venn diagram of gene expression patterns by tumor sub-site, CIMP high tumors, and MSI tumors.

TABLE 1

Description of Study Population

	Overall		Group A		Group B	
	N	%	N	%	N	%
Sex						
Male	94	53.7	52	59.8	42	47.7
Female	81	46.3	35	40.2	46	52.3
Center						
Kaiser	106	60.6	54	62.1	52	59.1
Utah	69	39.4	33	37.9	36	40.9
Site ^a						
Proximal	78	47.9	39	48.8	39	47.0
Distal	85	52.1	41	51.3	44	53.0
TP53						
Not mutated	98	56.0	52	59.8	46	52.3
Mutated	77	44.0	35	40.2	42	47.7
KRAS						
Not mutated	127	72.6	61	70.1	66	75.0
Mutated	48	27.4	26	29.9	22	25.0
MSI						
Stable (MSS)	143	81.7	68	78.2	75	85.2
Unstable (MSI)	32	18.3	19	21.8	13	14.8
CIMP						
Low	130	74.3	63	72.4	67	76.1
High	45	25.7	24	27.6	21	23.9
	Mean	STD	Mean	STD	Mean	STD
Age	65.2	10.2	65.3	10.1	65.1	10.4

^a 12 people are missing site information.

TABLE 2

Genes Differentially Expressed Between Proximal and Distal Nontumor Tissue at <0.01 Level in Both Groups A and B

Gene name	Average RPKM expression	Proximal	Fold change (Proximal/distal)	Group A	Group B	P-Value
<i>SLC13A2</i>	0.19	0.47	2.44	5.15E-03	1.49E-04	
<i>NR1H4</i>	0.36	0.08	0.23	8.33E-04	4.30E-04	
<i>SLC9A3</i>	8.11	1.17	0.14	3.62E-05	1.18E-05	
<i>NXPE1</i>	5.14	9.22	1.79	2.70E-03	6.78E-03	
<i>WFDC2</i>	0.24	1.30	5.32	9.38E-07	1.24E-05	
<i>HOXA13</i>	0.46	0.73	1.57	2.70E-03	1.08E-03	
<i>AKI</i>	0.37	1.99	5.33	2.36E-07	4.12E-06	
<i>CYP2C18</i>	0.99	0.29	0.30	6.77E-06	4.68E-03	
<i>ST3GAL4</i>	0.36	1.18	3.27	1.58E-07	8.61E-05	
<i>APOA4</i>	0.21	0.00	0.00	7.30E-03	3.53E-03	
<i>GCG</i>	0.08	0.30	3.91	6.67E-04	2.62E-03	
<i>MUC5B</i>	2.37	5.54	2.34	1.45E-04	7.10E-04	
<i>TRPM6</i>	0.71	1.39	1.95	8.66E-04	2.57E-03	
<i>PI3</i>	0.93	2.37	2.56	1.17E-04	1.18E-03	
<i>SLPI</i>	0.92	2.20	2.39	3.90E-03	4.56E-03	
<i>FOXA2</i>	0.38	0.89	2.34	1.50E-04	2.71E-03	
<i>HOXD13</i>	0.01	0.22	16.00	5.42E-04	9.49E-05	
<i>PYY</i>	0.36	1.27	3.54	7.00E-04	8.74E-05	
<i>SLC14A2</i>	0.24	0.02	0.07	1.24E-04	9.17E-04	
<i>SPINK5</i>	0.15	0.45	3.01	2.70E-03	3.94E-03	
<i>CAI</i>	1.43	2.77	1.94	7.81E-03	6.77E-03	
<i>SLC37A2</i>	2.42	0.30	0.12	1.98E-03	1.35E-03	
<i>CHST5</i>	0.66	2.15	3.26	1.81E-06	2.04E-04	
<i>GALNT5</i>	0.79	1.40	1.78	2.78E-03	6.78E-03	
<i>SLC28A2</i>	0.30	0.69	2.31	3.73E-03	2.29E-03	
<i>SLC20A1</i>	4.51	1.68	0.37	7.20E-03	7.00E-03	
<i>AHSG</i>	0.03	0.00	0.05	5.51E-03	1.83E-03	

Gene name	Distal	Proximal	Fold change	Group A	Group B
	Average RPKM expression	(Proximal/distal)			P-Value
<i>CDHR1</i>	1.34	3.51	2.62	1.88E-04	4.82E-05
<i>ADRA2A</i>	0.67	0.21	0.32	1.12E-03	5.65E-03
<i>SPON1</i>	0.82	1.63	1.99	7.20E-03	5.06E-03
<i>CLDN8</i>	0.15	0.97	6.60	3.74E-07	5.27E-07
<i>HKDC1</i>	0.49	0.17	0.36	6.36E-03	3.49E-03
<i>B3GNT7</i>	2.22	9.88	4.45	4.69E-08	5.89E-06
<i>LRRRC43</i>	0.04	0.01	0.17	8.14E-03	3.36E-03
<i>PRAC1</i>	0.07	1.72	24.45	5.35E-12	2.34E-09
<i>HOXB13</i>	0.12	2.02	16.61	3.59E-12	7.24E-10
<i>ST6GALNAC6</i>	1.23	6.55	5.33	6.82E-10	6.84E-06
<i>NLRP4</i>	0.02	0.01	0.25	6.67E-03	1.43E-03
<i>CYP11B1</i>	0.03	0.01	0.20	2.37E-03	5.83E-03
<i>NPHS1</i>	0.05	0.01	0.19	1.06E-03	3.19E-03
<i>CAPN13</i>	0.37	0.85	2.27	2.82E-03	7.06E-04
<i>PITX2</i>	0.35	0.02	0.07	2.02E-06	5.09E-11
<i>C7orf57</i>	0.03	0.01	0.33	7.88E-03	9.02E-04
<i>CPA6</i>	0.13	0.31	2.40	6.58E-03	2.14E-03
<i>CYP2C19</i>	0.56	0.15	0.26	3.23E-06	3.77E-03
<i>CKB</i>	4.04	10.24	2.54	7.63E-05	2.16E-03
<i>ANPEP</i>	10.21	1.44	0.14	8.51E-06	9.42E-05
<i>B4GALNT2</i>	2.02	0.58	0.29	8.72E-03	1.06E-04
<i>DPCR1</i>	0.03	0.01	0.18	6.73E-03	7.22E-03
<i>DRD5</i>	0.11	0.01	0.07	1.60E-03	7.10E-05
<i>PKHD1</i>	0.07	0.02	0.27	1.00E-04	2.53E-04
<i>KRT15</i>	0.14	0.38	2.76	6.18E-04	8.74E-04
<i>REG3A</i>	0.70	0.16	0.23	1.78E-04	8.93E-04
<i>INSL5</i>	0.01	0.23	15.53	5.79E-03	3.26E-04
<i>HOXC5</i>	0.05	0.01	0.29	9.21E-05	1.72E-03
<i>MFSH4</i>	0.32	1.03	3.20	1.98E-06	3.55E-04
<i>C2orf73</i>	0.02	0.01	0.37	4.32E-03	5.08E-03

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Gene name	Distal		Proximal	Fold change (Proximal/distal)	Group A		Group B	
	Average RPKM expression				P-Value		P-Value	
ARSI	0.14	0.39	0.39	2.77	2.89E-04	2.92E-04		
B3GALT5	0.91	3.35	3.35	3.68	6.30E-06	1.45E-04		
FAM3B	0.33	0.01	0.01	0.03	3.49E-09	8.33E-05		
C21orf88	0.06	0.72	0.72	12.60	3.43E-10	7.80E-07		
FFAR4	0.25	0.66	0.66	2.63	3.30E-04	8.45E-03		
GLDN	0.04	0.27	0.27	7.21	9.51E-05	1.97E-05		
LHFPL3	0.70	0.21	0.21	0.30	8.75E-04	1.49E-05		
TMEM72	0.21	0.42	0.42	2.02	4.12E-03	3.59E-03		
NWD1	0.05	0.09	0.09	1.98	8.24E-03	3.55E-03		
WNT7B	0.03	0.00	0.00	0.08	8.01E-03	3.75E-03		
CTD-2228K2.5	8.24	1.56	1.56	0.19	5.23E-04	6.05E-05		
PLA2G2A	4.51	13.72	13.72	3.04	3.00E-04	1.58E-04		
ESRRG	0.07	0.02	0.02	0.28	3.98E-05	2.96E-03		
SULT1C2	0.27	0.31	0.31	1.14	3.96E-03	7.68E-03		
HSD3B2	0.23	0.04	0.04	0.18	7.71E-03	9.82E-05		
NPY4R	0.64	1.58	1.58	2.46	3.85E-03	3.05E-03		
MUC12	4.82	16.21	16.21	3.37	9.48E-07	6.61E-05		
GBP7	0.05	0.01	0.01	0.15	4.57E-03	5.48E-03		
PRAC2	0.10	0.93	0.93	9.34	3.52E-09	1.13E-09		
LITDI	0.95	0.34	0.34	0.35	8.35E-03	7.94E-04		

TABLE 3

Networks Associated with CIMP High Tumors the Differentially Expressed Genes in our Data [Color table can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

ID	Molecules in network	Score	Focus molecules ^a	Top diseases and functions
1	<i>AGR2</i> , Alpha catenin, <i>ANO1</i> , <i>AQP5</i> , <i>ATP9A</i> , <i>AXIN2</i> , <i>B3GALT5</i> , <i>CDX2</i> , Cg, <i>CREB3L1</i> , <i>DAPK1</i> , <i>DUSP4</i> , <i>ELAVL2</i> , <i>EREG</i> , ERK1/2, FSH, GUCY2C, <i>KCTD1</i> , <i>LGALS2</i> , Lh, <i>MSI1</i> , <i>MT2A</i> , <i>MUC5B</i> , <i>NKD1</i> , <i>PDE4D</i> , <i>PITX2</i> , <i>PLK2</i> , Proinsulin, <i>PYY</i> , <i>RPRM</i> , <i>SOX1</i> , <i>TDGF1</i> , <i>WIF1</i> , <i>Wnt</i> , <i>ZIC2</i>	54	24	Organismal development, cellular development, cellular growth and proliferation
2	<i>ADCY</i> , <i>AMPK</i> , <i>Calmodulin</i> , <i>Creb</i> , <i>DRD5</i> , F Actin, <i>FFAR4</i> , GHRHR, GLP1R, Gpcr, <i>GPR132</i> , <i>GPR143</i> , GPR160, GRM8, Jnk, <i>MFSD4</i> , <i>MLPH</i> , PARP, <i>PHACTR3</i> , <i>PI3K</i> (complex), <i>Pka</i> , <i>Pkc(s)</i> , PP1 protein complex group, <i>PPP1R14C</i> , <i>RAB27B</i> , <i>RAB3B</i> , <i>RAMP1</i> , <i>SLC14A2</i> , <i>SLC38A3</i> , <i>SLC39A4</i> , <i>SLC39A12</i> , <i>SLC6A4</i> , <i>SNCAIP</i> , <i>SPDEF</i> , <i>VEGF</i>	37	18	Cell signaling, molecular transport, nucleic acid metabolism
3	<i>ACE2</i> , <i>Akt</i> , <i>Alp</i> , <i>AMACR</i> , <i>ASCL2</i> , <i>CEACAM6</i> , Collagen(s), <i>CTTNBP2</i> , <i>DDC</i> , <i>DPEP1</i> , <i>ELOVL5</i> , <i>ESRRG</i> , Growth hormone, HDL, HDL-cholesterol, hemoglobin, Histone h4, IL1, Laminin, LDL, LDL-cholesterol, <i>LPL</i> , N-cor, <i>OPTC</i> , <i>PMEPA1</i> , Ppp2c, <i>SERPINF2</i> , <i>SLC2A8</i> , <i>TFF1</i> , <i>TG</i> , TGFB, <i>TINAG</i> , <i>TRPV6</i> , <i>VNN1</i> , <i>XPNPEP2</i>	33	13	Cell cycle, lipid metabolism, molecular transport
4	<i>ADRA2A</i> , <i>Ap1</i> , <i>ARID3A</i> , <i>BEX2</i> , <i>CD3</i> , <i>CDC42EP1</i> , <i>DLEU7</i> , <i>E2f</i> , <i>ERK</i> , estrogen receptor, <i>F7</i> , <i>IFI6</i> , IFN Beta, <i>INPP5D</i> , <i>KCNK5</i> , <i>KIAA0226L</i> , <i>KRT13</i> , Mapk, NFkB (complex), <i>PDGF BB</i> , <i>PLA2</i> , <i>PLA2G12B</i> , <i>PLA2G2A</i> , <i>POFUT1</i> , Rac, <i>RARRES3</i> , <i>Ras</i> , <i>RNF125</i> , <i>ROR1</i> , Shc, Sos, TCR, <i>TNFRSF11A</i> , <i>VAV3</i> , <i>XAF1</i>	32	14	Cardiovascular disease, developmental disorder, cell-to-cell signaling and interaction
5	<i>ADIRF</i> , <i>BSN</i> , <i>CA8</i> , CECR2, CENPB, DHX34, DNAJC13, ETV7, EXT2, FAM49A, FITM2, FZD5, <i>GALNT5</i> , <i>HOXC6</i> , <i>KRT13</i> , <i>KRT23</i> , <i>L3MBTL1</i> , <i>LAPTM4B</i> , MICAL1, <i>NKD2</i> , <i>PCMTD2</i> , <i>PLCB4</i> , PLXDC1, <i>QSOX1</i> , RNF25, <i>RNF43</i> , <i>RPS6KA6</i> , SAP30BP, <i>SEMA5A</i> , SPTBN4, SRSF5, <i>ST3GAL4</i> , <i>UBC</i> , <i>UFM1</i> , <i>ZSCAN2</i>	27	15	Cellular assembly and organization, cellular development, cellular growth and proliferation
6	<i>ACSL5</i> , <i>ATP9A</i> , <i>BARX2</i> , Bmyo, <i>CLDN4</i> , <i>COL6A2</i> , <i>COL6A3</i> , <i>CREB3</i> , <i>DUSP4</i> , FGF5, FGF12, GPD5, GPCPD1, <i>GPR132</i> , IL13, IL22RA2, KLK1, LIF, LPCAT3, <i>LY6G6C</i> , <i>MUC5B</i> , MYL3, OSM, <i>PLAGL2</i> , <i>PLL</i> , <i>QSOX1</i> , SELE, <i>SLC30A2</i> , <i>SLC6A4</i> , SPINT1, SRF, <i>ST6GALNAC6</i> , TGFB1, <i>TMEM40</i> , <i>TPBG</i>	27	12	Hereditary disorder, skeletal and muscular disorders, antimicrobial response
7	<i>ADK</i> , <i>AK1</i> , APP, BCKDK, CAB39L, CCDC144NL, CKMT1A/CKMT1B, CXXC1, ENO3, <i>ESRRG</i> , FGF12, GSPT2, H2AFJ, H2AFY, HSP90AA1, LAMC1, MAST2, <i>MOGAT3</i> , <i>NOL4L</i> , NSRP1, <i>PRR15</i> , <i>R3HDM1</i> , <i>RPS6KA6</i> , <i>SLC13A3</i> , SMG1, SMG5, <i>SMTNL2</i> , SMYD3, SRPK1, STK26, STRADA, <i>TRIM7</i> , <i>TTI2</i> , <i>TTPA</i> , <i>UPF3A</i>	26	13	Molecular transport, DNA replication, recombination, and repair, nucleic acid metabolism
8	26s Proteasome, 3,20-pregnanedione, <i>ABAT</i> , <i>ABCA3</i> , <i>AQP5</i> , calpain, <i>CAPN9</i> , <i>CAPN10</i> , <i>CAPN11</i> , caspase, <i>CEBPA</i> , CEL, CFTR, <i>CYP2B6</i> , <i>FAM3B</i> , <i>GAS2</i> , <i>GATA5</i> , Histone h3, <i>Hsp90</i> , <i>IL12</i> (complex), Immunoglobulin, Insulin, Interferon alpha, <i>LST1</i> , <i>MLPH</i> , <i>NRI12</i> , <i>OXCT1</i> , P38 MAPK, <i>PRAP1</i> , RNA polymerase II, <i>SLC9A3</i> , <i>Sult2a2</i> , <i>TNNC2</i> , <i>Troponin C</i> , Ubiquitin	25	11	Lipid metabolism, small molecule biochemistry, cell morphology
9	<i>ADAM9</i> , <i>AKAP11</i> , <i>ANO9</i> , ATP2B2, CPNE9, DLG1, DLGAP4, EIF5AL1, <i>ELAVL2</i> , ENGASE, FZD4, GAS1, GLI1, GUCY1A2, <i>MAP7D2</i> , MARCH2, MPP1, MPP2, PAPSS2, PHLDA2, <i>PLD5</i> ,	25	13	Nucleic acid metabolism, small molecule biochemistry,

ID	Molecules in network	Score	Focus molecules ^a	Top diseases and functions
10	PLOD1, RAB34, <i>REG4</i> , RNASEL, <i>SDR42E1</i> , SULF2, <i>SVOPL</i> , UBC, <i>ZIC2</i> , <i>ZNF529</i> , <i>ZNF737</i> , <i>ZNF773</i> , <i>ZNF813</i> , <i>ZNF518B</i>	23	13	<i>embryonic development</i>
	<i>B3GNT7</i> , DLX1, DUSP5, EED, ELAVL1, ETF1, EVX1, FAM217B, FARP1, FOXE1, <i>HOXB13</i> , IYD, <i>LRRRC2</i> , Macf1, MBD3, MGA, NCKAP1, NFIB, <i>NKD1</i> , <i>NR6A1</i> , PEG3, POU5F1, PPHLN1, RBM15, <i>REC114</i> , <i>SLC16A14</i> , <i>SPIN3</i> , TBX3, <i>TC2N</i> , TEAD2, <i>TP53RK</i> , <i>UNC45B</i> , WNT8B, <i>ZIC2</i> , <i>ZSWIM1</i>			<i>Gene expression, embryonic development, organismal development</i>
11	<i>ALDOC</i> , <i>BAG3</i> , <i>CACNA1G</i> , <i>CACNG4</i> , <i>CEND1</i> , <i>CTNBL1</i> , <i>DDX27</i> , <i>DKK3</i> , dopamine, <i>DRD5</i> , <i>EPDR1</i> , ERP27, GABRG2, GNG3, <i>GNG4</i> , GNG7, GNG12, GNG13, HTT, KCNK2, <i>KRT38</i> , <i>MOCS3</i> , NCKAP1, <i>OC90</i> , OPRK1, <i>PGBD5</i> , SPP1, SYN2, <i>SYN3</i> , <i>TMEM252</i> , UBD, UBQLN4, UCHL3, YWHAG, <i>ZMYND8</i>	20	10	<i>Neurological disease, cell-to-cell signaling and interaction, drug metabolism</i>

^aFocus molecules are those significantly differentially expressed with CIMP high tumors in our data. Red font indicates upregulated genes, while green font indicates down-regulated genes.

TABLE 4

Networks Associated with MSI Differentially Expressed Tumors [Color table can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

ID	Molecules in Network	Score	Focus molecules ^a	Top diseases and functions
1	<i>AGR2</i> , <i>AKAP5</i> , Alpha catenin, <i>ATOH1</i> , <i>ATP9A</i> , <i>AXIN2</i> , <i>Cg</i> , Creb, <i>CREB3L1</i> , <i>DUSP4</i> , <i>EREG</i> , <i>ERK1/2</i> , <i>FSH</i> , <i>GNLY</i> , <i>HOXB13</i> , Lh, <i>LRRN1</i> , <i>MT1X</i> , <i>MT2A</i> , <i>MUC2</i> , MUC4, <i>MUC5AC</i> , <i>MUC5B</i> , Mucin, <i>NKD1</i> , <i>NR6A1</i> , <i>PDE4D</i> , <i>PITX2</i> , <i>PYY</i> , RAP1GAP, <i>RPRM</i> , SOX1, <i>T</i> , <i>TDGF1</i> , <i>Wnt</i>	55	26	Cell death and survival, cell morphology, organ morphology
2	<i>ADRA2A</i> , <i>ASCL2</i> , caspase, <i>CD3</i> , <i>CEL</i> , <i>DLEU7</i> , <i>ERK</i> , <i>F7</i> , F Actin, <i>FABP6</i> , <i>FAM46A</i> , <i>FFAR4</i> , <i>FRAS1</i> , FREM1, <i>FREM2</i> , Gpcr, <i>GPR143</i> , <i>GPR157</i> , <i>GRM8</i> , <i>GTF2IRD1</i> , Histone h3, Jnk, <i>KIAA0226L</i> , <i>LRRC2</i> , Mapk, NFkB (complex), <i>NUDT6</i> , <i>PI3K</i> (complex), Pka, Pkc(s), <i>POFUT1</i> , <i>SPDEF</i> , <i>TPH1</i> , VEGF, <i>XAF1</i>	39	20	Developmental disorder, hereditary disorder, ophthalmic disease
3	<i>ABAT</i> , <i>ABCA3</i> , <i>DCY</i> , Akt, <i>Ap1</i> , <i>ARID3A</i> , <i>CEBPA</i> , CYP27A1, <i>CYP2B6</i> , <i>DDC</i> , <i>DGAT2</i> , <i>DPEP1</i> , E2f, <i>FAM3B</i> , <i>HDL</i> , <i>IL1</i> , Immunoglobulin, Insulin, <i>IP6K3</i> , LDL, <i>LPL</i> , <i>MLPH</i> , Nr1h, <i>PLA2G2A</i> , <i>PMEP1</i> , <i>PRAP1</i> , <i>RAB27B</i> , Ras, <i>REN</i> , <i>SALL4</i> , <i>SERPINE2</i> , Sos, Tgfbeta, <i>TNFRSF11A</i> , trypsin	38	20	Cell morphology, cellular assembly and organization, tissue morphology
4	<i>ADK</i> , <i>AGR2</i> , APP, ARID3B, <i>B3GNT7</i> , CHMP3, COL4A6, CST3, <i>CTTNBP2</i> , <i>CXCL14</i> , <i>DUSP4</i> , EXT2, FBXO38, FXYD1, GBP4, <i>GDPD5</i> , <i>HYPK</i> , <i>KLF7</i> , Macf1, miR-224-5p (miRNAs w/seed AAGUCAC), MYH9, <i>MYOM3</i> , <i>PKNOX1</i> , <i>PLAGL2</i> , plasminogen activator, <i>POU5F1</i> , <i>PPP2CA</i> , <i>PRR15</i> , <i>PTPRO</i> , SELE, SMCHD1, <i>SMTNL2</i> , <i>ST6GALNAC6</i> , <i>STRIP2</i> , <i>TGFB1</i>	24	15	Cardiovascular disease, developmental disorder, hereditary disorder
5	<i>ADK</i> , <i>AK1</i> , ARIH1, <i>CAB39L</i> , <i>CLCN5</i> , dopamine, <i>FARP1</i> , <i>GGT7</i> , <i>HECW1</i> , <i>KCTD20</i> , <i>KIF3C</i> , <i>KRT23</i> , <i>MAPRE3</i> , <i>MARK1</i> , <i>NOL4L</i> , NUDT2, <i>PALD1</i> , <i>PHACTR3</i> , PPHLN1, PPP1CA, <i>RNF43</i> , RNF19A, <i>SLC9A3</i> , <i>SNCAIP</i> , SNCB, SRPK1, STK11IP, STRADA, STXBP2, <i>SYN3</i> , SYT3, TROAP, UBC, USP48, <i>ZSWIM3</i>	23	14	Molecular transport, cell-to-cell signaling and interaction, nervous system development and function
6	12-hydroxyeicosatetraenoic acid, <i>APOLD1</i> , <i>ARSJ</i> , CELSR3, COL12A1, <i>COL9A1</i> , COL9A2, <i>COL9A3</i> , Collagen type ix, <i>COMP</i> , <i>CYP4F2</i> , <i>FMOD</i> , <i>FNI</i> , <i>FZD8</i> , G-protein gamma, <i>GNG4</i> , GNG5, GNG7, GRB2, HNRNPA2B1, HSP90AA1, <i>HUNK</i> , <i>LY6G6F</i> , MYRIP, NCK1, <i>NKD2</i> , PHACTR2, <i>PLCB4</i> , RAF1, <i>RNF125</i> , <i>SGSM2</i> , <i>SHROOM2</i> , <i>SLC22A3</i> , STK32C, <i>TRIM7</i>	23	14	Connective tissue disorders, developmental disorder, hereditary disorder
7	<i>AMACR</i> , <i>AQP5</i> , CATSPER1, CATSPER4, <i>CATSPERB</i> , <i>CTSE</i> , CYP27A1, <i>DDX27</i> , DMD, <i>EPDR1</i> , Flg, FXYD1, GFER, HSPA2, <i>IGFN1</i> , IL31, <i>IP6K3</i> , <i>LY6G6C</i> , Mt3, nitrite, <i>NR3C1</i> , <i>OSM</i> , <i>P38 MAPK</i> , <i>PLLP</i> , PPARA, <i>PPP1R14C</i> , PPP1R16A, RBP4, SIRT1, SLC10A2, <i>SLC1A7</i> , <i>SNTB1</i> , Stat3-Stat3, <i>TLR5</i> , <i>TNNC2</i>	23	14	Lipid metabolism, small molecule biochemistry, vitamin and mineral metabolism
8	<i>ABLIM2</i> , ADH6, ADH1B, APCDD1, B3GALT1, <i>B3GALT5</i> , <i>CA8</i> , CEBPB, COL4A5, Cyp2c11l, <i>CYP2W1</i> , <i>ELAVL1</i> , F Actin, <i>FOXJ3</i> , <i>GAS2</i> , <i>GOLT1A</i> , HNF1A, HNF4A, <i>KLHL3</i> , <i>MT1X</i> , MTF2, PCNP, <i>PLA2G12B</i> , Rbp, <i>SLC22A11</i> , Sprr1b, <i>ST3GAL4</i> , <i>STRA6</i> , <i>TMEM176A</i> , <i>TNFRSF19</i> , tretinoin, <i>UGT2B7</i> , <i>UGT2B15</i> , <i>XPNPEP2</i> , <i>ZNF556</i>	22	14	Cellular development, digestive system development and function, hepatic system development and function

^aFocus molecules are those significantly differentially expressed with MSI tumors in our data. Red font indicates upregulated genes, while green font indicates down-regulated focus genes.