www.oncotarget.com Oncotarget, 2021, Vol. 12, (No. 8), pp: 767-782

Research Paper

Controlling for cellular heterogeneity using single-cell deconvolution of gene expression reveals novel markers of colorectal tumors exhibiting microsatellite instability

Matthew A.M. Devall[1] **and Graham Casey**[1]

[1]Center for Public Health Genomics, Department of Public Health Sciences, University of Virginia, Charlottesville, VA, USA

Correspondence to: *Graham Casey*, ***email***: *gc8r@virginia.edu*

Keywords: colorectal cancer; single-cell deconvolution; microsatellite instability; RNA-sequencing; enteroendocrine

Received: *January 23, 2021* ***Accepted***: *March 22, 2021* ***Published***: *April 13, 2021*

ABSTRACT

Approximately 15% of colorectal cancer (CRC) cases present with high levels of microsatellite instability (MSI-H). Bulk RNA-sequencing approaches have been employed to elucidate transcriptional differences between MSI-H and microsatellite stable (MSS) CRC tumors. These approaches are frequently confounded by the complex cellular heterogeneity of tumors. We performed single-cell deconvolution of bulk RNA-sequencing on The Cancer Genome Atlas colon adenocarcinoma (TCGA-COAD) dataset. Cell composition within each dataset was estimated using CIBERSORTx. Cell composition differences were analyzed using linear regression. Significant differences in abundance were observed for 13 of 19 cell types between MSI-H and MSS/MSI-L tumors in TCGA-COAD. This included a novel finding of increased enteroendocrine (q = 3.71E^{-06}) and reduced colonocyte populations (q = 2.21E^{-03}) in MSI-H versus MSS/MSI-L tumors. We were able to validate some of these differences in an independent biopsy dataset. By incorporating cell composition into our regression model, we identified 3,193 differentially expressed genes (q = 0.05), of which 556 were deemed novel. We subsequently validated many of these genes in an independent dataset of colon cancer cell lines. In summary, we show that some of the challenges associated with cellular heterogeneity can be overcome using single-cell deconvolution, and through our analysis we highlight several novel gene targets for further investigation.

INTRODUCTION

Colorectal cancer (CRC) is a complex, heterogenous disease. At least two broad molecular pathways contribute to the development of CRC tumors: microsatellite instability (MSI) and chromosomal instability (CIN). Microsatellite instability-high (MSI-H) tumors account for ~15% of CRC tumors, and are driven by a dysregulation of mismatch repair (MMR) [1]. The majority of MSI-H tumors (80%) occur via acquired epigenetic silencing of the MMR gene *MLH1*. In contrast, microsatellite stable (MSS) tumors account for the majority (~85%) of CRC tumors, and are defined by increased loss or gain and rearrangement of chromosomes (CIN phenotype) [2]. MSI-H and MSS tumors have been shown to differ with regards to survival [3] and response to treatment [4], but the molecular mechanisms driving the differences between these tumor types remain poorly understood. One commonly employed approach to interrogate differences between tumor types is through a comparative analysis of RNA-sequencing (RNA-seq) data [5]. However, the cellular heterogeneity of tumors can mask important gene expression differences identified by RNA-seq. To improve understanding of the molecular mechanisms across tumor subtypes, cellular composition must be adequately controlled.

Various methodological advances have been made to address the problems arising from tumor cellular heterogeneity including the sorting of cell populations prior to RNA-seq and the development of single-cell RNA-seq (scRNA-seq) approaches. However, these studies are often limited by factors such as availability of sufficient material or high cost associated with scRNA-seq. While these methods increase resolution,

they are often hindered by small study designs, which reduces generalizability. The application of single-cell deconvolution approaches using bulk RNA-seq data therefore provides an opportunity to infer cell composition differences of large tumor datasets at reduced cost [6]. Indeed, our group has previously employed single-cell deconvolution to quantify and account for variation in cell composition in a large colon organoid study [7].

In this study, we aim to identify differences in cell composition between MSI-H and MSS/microsatellite instability-low (MSI-L) tumors in The Cancer Genome Atlas Colon Adenocarcinoma (TCGA-COAD) dataset [8]. We achieve this by using a machine learning approach [6] and by incorporating scRNA-seq data derived from normal colon biopsies [9] to deconvolute bulk RNA-seq data [8]. We estimate cell type abundance and identify novel cellular composition differences between MSI-H and MSS/MSI-L tumors, which we then validate in an independent cohort of CRC tumors [10]. Following adjustment for cell composition, we identified novel differentially expressed genes (DEG)s between MSI-H and MSS/MSI-L tumors in TCGA-COAD and replicate many of these DEGs in an independent analysis of colon cancer cell lines derived from Cancer Cell Line Encyclopedia (CCLE) [11]. Together, we provide data showing that single-cell deconvolution analysis of tumors can be used to address cellular heterogeneity, and has the potential to reveal novel insight into tumor biology.

RESULTS

Differential expression of genes specific to immune cell types are commonly overexpressed in MSI-H tumors

MSS and MSI-L tumors were merged for comparisons to MSI-H tumors in all downstream analysis based on the similarities of expression profiles observed between MSI-L and MSS tumors (Supplementary Figure 1A). Further, a total of 89 significant DEGs (q = 0.05) were identified in our preliminary regression analysis of MSI-L vs MSS tumors. This was in stark contrast to the extent of transcriptomic variation observed between MSI-H versus MSI-L (5,472 DEGs) or MSI-H versus MSS tumors (8,641 DEGs), where 61.48- and 97.01-fold more DEGs were reported than in MSI-L vs MSS tumors, respectively (Supplementary Figure 1B). This is in line with clinical practice, where MSI-L tumors are often considered to be similar to MSS tumors [12]. This grouping has also been used in other studies [13].

Differential expression analysis of RNA-seq data from MSI-H versus MSS/MSI-L tumors identified 8,693 FDR corrected DEGs (q = 0.05). Our preliminary analysis aimed to determine the potential impact of cell composition on the DEGs reported in an RNA-seq analysis of MSI-H versus MSS/MSI-L tumors. We found that 17.92% (1,558) of these DEGs were potential markers of specific cell types [9]. In total, 72.62% (515/656) of significant DEGs corresponding to immune cell markers were expressed at higher levels in MSI-H versus MSS/MSI-L tumors (Figure 1). This finding is in line with reports that MSI-H tumors are most frequently associated with increased immune cell populations [14]. We extended this analysis by increasing the resolution of the cell types considered (Supplementary Table 1). We found that expression markers of both transit amplifying (TA) and $CD8^+$T cell populations were consistently higher in MSI-H versus MSS/MSI-L tumors. Further, significant reductions in expression were identified for 41 of 50 FDR corrected stem cell-related genes in MSI-H tumors, including *LGR5* ($q = 2.67E^{-04}$). Significant reductions in expression were also identified for 101 of 146 FDR corrected colonocyte expression markers in MSI-H tumors. Thus, many of the significant differences identified in standard regression analysis of tumor biopsies are reflective of variation in cell composition across tumors.

To address the challenge of cellular heterogeneity and to accurately capture cell composition of TCGA-COAD tumors, we employed single-cell deconvolution using publicly available scRNA-seq data generated from normal colon biopsies [9]. A signature matrix was generated from scRNA-seq data, which stratified cell populations based upon the average gene expression of select genes (defined by CIBERSORTx) across the 19 cell types considered (Figure 2A). Cell scores were then generated for these cell types using this signature matrix [6]. Regression analysis was performed on each cell score to determine whether cell scores capture known expression markers of relevant cell types. One-way Fisher's exact tests determined significant enrichments for known expression markers in the DEGs identified in these cell score regressions (Figure 2B) [9]. Glial cell gene expression markers were not provided within the scRNA-seq dataset. As such, we identified canonical gene expression markers for glial cells using an online database [15]. Of the eight canonical markers of enteric glial cells identified, *ALDH1A3* ($q = 1.49E^{-25}$), *SLC18A2* ($q = 3.39E^{-11}$), *S100B* ($q = 9.69E^{-11}$), *FOXD3* ($q = 5.98E^{-10}$), *SLC18A2* ($q = 3.39E^{-11}$), *GFAP* ($q = 3.58E^{-09}$) and *GFRA3* ($q = 1.24E^{-04}$) were significantly upregulated (q = 0.05) in glial cell regressions of TCGA-COAD tumors. These findings highlight that the deconvolution approach used here accurately captures the expected expression of relevant cell types.

We used immunedeconv [16] to generate stromal and immune cell scores in matched tumor samples using four additional deconvolution methods [17–20]. We were able to estimate the relative performance of each method by correlating cell type expression markers to cell scores generated from each approach. We considered performance as the difference (shift) in median correlation between cell score and expression markers of that cell type, and the median correlation of the same cell score to

expression markers of other cell types. The performance of our cell scores was comparable to other approaches (Supplementary Table 2).

Finally, we compared the correlation of CRC stem cell markers *LGR5*, *CD24* and *EPCAM* to both a previously generated stemness score [21] and to the stem cell scores generated in our analysis. We found that *LGR5*, *CD24* and *EPCAM* were more positively correlated with stem cell scores generated in our method (r = 0.26; r = 0.45; r = 0.38 respectively) than with the stemness score (r = 0.08; r = 0.14; r = 0.30 respectively). We extended this approach to determine the relative ability of stemness scores, stem cell scores and cycling TA cell scores to capture the expression of markers of normal colon stem cells and TA cells [9] (Figure 2C). We find that stem cell scores generated here are better able to distinguish TA and stem cell gene maker expression than markers of stemness. These results are perhaps unsurprising given that the stemness score was originally designed as a pan cancer score of global dedifferentiation, rather than a colon-specific marker of stem cell content.

Defining a high-resolution cellular roadmap of CRC tumors

We first aimed to determine differences in overall cell composition between MSI-H and MSS/MSI-L tumors. A linear regression was performed on MSI status for each cell score. Significant differences in 13 of 19 cell populations were identified (q = 0.05) (Figure 3). Increased cell abundance was observed for six of eight immune cell populations in MSI-H tumors, in line with increased immune cell content associated MSI-H tumors [14]. A cytolytic signature was generated for each sample by averaging the expression of six genes (*GZMA*, *GZMB*, *GZMH*, *GZMK*, *GZMM* and *PRF1*), as demonstrated in Rooney et al. [22]. This signature was strongly correlated to $CD8^+$T cell content (r = 0.76). Further, MSI-H $CD8^+$T cells were found to have a significantly increased cytolytic score ($P = 1.53E^{-53}$), indicating an increased potential for tumor immune cell killing in MSI-H samples. We also observed an increase in enteroendocrine cell (EEC) content ($q = 3.71E^{-06}$), in MSI-H versus MSS/MSI-L tumors. To the best of our knowledge, this analysis represents the first to report this finding. We also observed a decrease in colonocyte ($q = 2.21E^{-03}$) and stem cell content ($q = 4.23E^{-21}$) as well as an increase in the cycling TA cell population ($q = 2.32E^{-10}$) in MSI-H tumors, highlighting the importance of considering cellular heterogeneity of epithelial cells in these analyses.

We were able to validate some of these changes in cellular composition in a second, smaller cohort of MSI-H tumors (GSE146889) [10]. We were unable to capture EEC, dendritic cell or innate lymphoid cell gene signatures (Supplementary Figure 2). As a result, no analysis was performed on these cell types in this dataset. We replicated reduced stem cell (P = 0.02) as well as increased cycling TA ($P = 7.05E^{-03}$) and macrophage cell content (P = 0.029) in MSI-H versus MSS tumors. Further, we also observed trends for a significant increase in $CD8^+$T cells (P = 0.076) and a reduction in colonocytes (P = 0.08) in MSI-H versus MSS tumors (Supplementary Figure 3).

For sensitivity, we repeated our analysis of TCGA-COAD by stratifying MSS and MSI-L. Here, we found that all 13 significant cell populations remained significant (q = 0.05) in a regression of MSI-H versus MSS, while no significant differences were observed between MSS and MSI-L tumors. Of the 13 significantly different cell populations identified between MSI-H and MSS tumors, seven were also found to be significantly different between MSI-H and MSI-L tumors (Table 1). Replication of three additional cell types was confirmed at a nominal significance threshold (P = 0.05). These data further support a strong similarity between MSS and MSI-L tumors.

Differential expression following adjustment for cell composition

To correct for the effects of cell composition in our analysis of MSI-H versus MSS/MSI-L tumors, we repeated our initial regression while incorporating cell composition scores. We identified 3,193 DEGs (q = 0.05), of which 556 were not reported in our original analysis and were thus deemed novel (Figure 4A). Pathway analysis performed on the novel DEGs that displayed reduced expression in MSI-H compared to MSS/MSI-L tumors revealed an enrichment for DNA repair ($q = 2.81E^{-06}$). Indeed, 124 Gene Ontology biological processes were enriched in this analysis, of which 15 were associated with DNA damage, mismatch or repair processes (Figure 4B). Interestingly, *PMS1* ($q = 1.90E^{-03}$), *MSH2* ($q = 2.80E^{-03}$) and *MSH6* (q = 0.02) only reached significance following adjustment for cell composition. Inherited mutations of these genes are associated with Lynch syndrome, a genetic condition that greatly increases the risk of MSI-H tumor development [23]. In contrast, pathway analysis of downregulated DEGs that were no longer significant following adjustment for cell composition revealed enrichments for cell-specific processes such as T cell activation ($q = 2.40E^{-03}$), leukocyte differentiation (q = 0.03) and T cell differentiation (q = 0.03). However, a notable absence for enrichment of pathways associated with DNA repair or mismatch repair was observed (Figure 4B) [24, 25]. Together, these findings highlight that adjusting for cell composition, leads to the identification of important pathways, and reduced the reporting of findings that can be attributed to cell-specific variation. A full list of the pathways and DEGs identified within each analysis can be found in (Supplementary File 1).

Given the cellular heterogeneity of tumor biopsies and the relatively small sample size of the GSE146889,

we did not attempt to replicate these differences in this cohort. Instead, we performed a similar regression analysis to identify DEGs associated with MSI status in a dataset of colon cancer cell lines [11]. Here we generated cell scores for four epithelial cell populations (Supplementary Figure 4). Regression analysis was then performed on MSI status while adjusting for cell composition. We were able to replicate 607 of these DEGs at nominal significance (P = 0.05) and 221 following FDR correction (q = 0.05) (Supplementary Table 3). A one-way Fisher's exact test was performed, which revealed that there was a significant enrichment of overlap between the FDR corrected DEGs identified in these two datasets (Odds ratio = 1.68, P = 3.09E^{-10}). Thus, our secondary analysis of colon cancer cell lines provides an independent replication of the results identified in TCGA-COAD tumors. With regards to the 556 DEGs identified in TCGA-COAD only after adjustment for cell composition, 56 were also identified in colon cancer cell lines dataset (P = 0.05), of which 18 remained significant following FDR correction (Supplementary Table 5). The three most significant novel genes identified in MSI-H vs MSS/MSI-L tumors that were subsequently replicated in colon cancer cell lines were *AGMO, LINC02577* and *KIF1A*, all of which displayed reduced expression in MSI-H cell lines and tumors compared to MSS/MSI-L cell lines and tumors. To the best of our knowledge, roles for these three genes in MSI-H tumors have yet to be defined.

Network analysis for the identification of candidate modules associated with microsatellite instability.

We regressed out the effects of cell composition and additional covariates prior to our network analysis (see Methods) to determine patterns of differential co-expression between MSI-H and MSS/MSI-L tumors. We performed weighted gene co-expression network analysis (WGCNA) [26], which identified a network consisting of 96 modules of coordinated expression (Supplementary Figure 5). Of these modules, 35 were found to be significantly different between MSI-H and MSS/MSI-L tumors following strict Bonferroni correction (q = 0.05) (Supplementary File 2, Figure 5). We used an independent method to validate the co-expression observed within the 96 modules identified in our analysis. We uploaded gene lists for each module into STRING [27]. For each module, we calculated enrichment scores for protein-protein interaction (PPI). Of the 97 modules, 87 were enriched for PPI, including 34 of the 35 significant modules identified

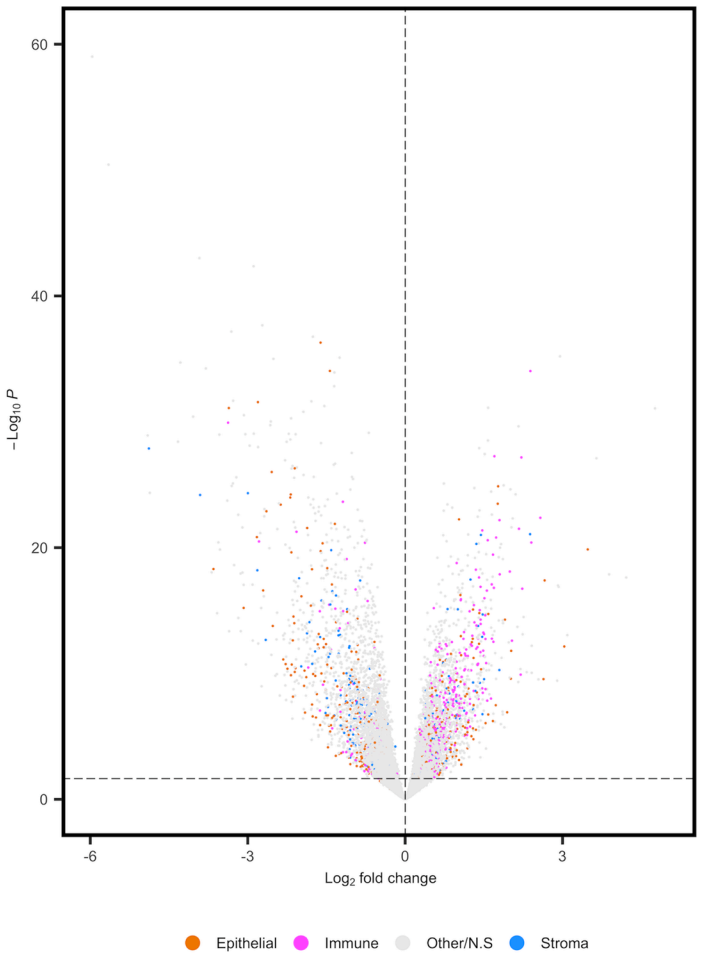

Figure 1: Differential expression analysis of TCGA-COAD prior to adjustment for cell composition. Volcano plot displaying direction of effect for bulk cell types. Positive log2fold changes correspond to increased expression in MSI-H tumors.

Table 1: Cell composition analysis of MSI-H versus MSS and MSI-L versus MSS tumors in TCGA-COAD

Cell-Type	MSI-H vs MSS (n = 242)			MSI-H vs MSI-L (n = 116)			MSI-L vs MSS (n = 231)			Significant in MSI-H vs MSS/MSI-L
	T-Statistic	P	FDR	T-Statistic	P	FDR	T-Statistic	P	FDR	
B	**-3.29**	**1.14E^{-03}**	**2.16E^{-03}**	-2.19	0.03	0.08	-0.73	0.47	0.68	True
CD4^{+}T	-2.08	0.04	0.053	-1.29	0.20	0.27	0.88	0.38	0.68	False
CD8^{+}T	**7.87**	**1.37E^{-13}**	**1.30E^{-12}**	**3.59**	**5.09E^{-04}**	**2.41E^{-03}**	1.05	0.30	0.68	True
Colonocytes	**-4.26**	**3.03E^{-05}**	**7.20E^{-05}**	**-2.68**	**8.58E^{-03}**	**0.02**	1.39	0.17	0.53	True
CyclingTA	**6.57**	**3.26E^{-10}**	**1.55E^{-09}**	**4.29**	**3.94E^{-05}**	**3.74E^{-04}**	0.82	0.41	0.68	True
DC	**7.11**	**1.40E^{-11}**	**8.87E^{-11}**	**3.65**	**4.14E^{-04}**	**2.42E^{-03}**	1.52	0.13	0.53	True
Enteroendocrine	**4.89**	**1.92E^{-06}**	**6.08E^{-06}**	**2.72**	**7.63E^{-03}**	**0.02**	-0.18	0.86	0.92	True
Fibroblast	-0.19	0.85	0.90	0.17	0.91	0.96	-0.89	0.37	0.68	False
Glia	**3.07**	**2.34E^{-03}**	**3.71E^{-03}**	1.28	0.20	0.27	-0.50	0.62	0.84	True
Goblet	1.92	0.06	0.08	0.46	0.65	0.82	1.85	0.07	0.53	False
ILC	-0.31	0.76	0.85	0.31	0.76	0.88	-0.73	0.47	0.68	False
Macrophage	**4.63**	**6.05E^{-06}**	**1.64E^{-05}**	2.00	0.048	0.09	-0.11	0.92	0.92	True
Mast	1.82	0.07	0.08	1.34	0.18	0.27	0.19	0.85	0.92	False
Microvascular	**-3.66**	**3.13E^{-04}**	**6.60E^{-04}**	-2.09	0.04	0.08	1.42	0.16	0.53	True
Myofibroblast	0.01	0.99	0.99	0.05	0.96	0.96	-0.86	0.39	0.68	False
NK	**2.44**	**0.02**	**0.02**	1.30	0.20	0.27	0.15	0.88	0.92	True
Pericyte	**-3.23**	**1.42E^{-03}**	**2.45E^{-03}**	-0.27	0.79	0.88	-1.72	0.09	0.53	True
Postcapillary Venule	**6.21**	**2.49E^{-09}**	**9.46E^{-09}**	**3.07**	**2.73E^{-03}**	**0.01**	-0.20	0.85	0.92	True
Stem	**-10.84**	**2.03E^{-22}**	**3.86E^{-21}**	**-6.52**	**2.41E^{-09}**	**4.58E^{-08}**	-1.60	0.11	0.53	True

Positive test statistic indicates increased cell content in tumors with greater MSI. FDR corrections were calculated for each regression analysis individually and FDR significance was set at 5%. Bond fold indicates cell-types that pass FDR correction for that regression analysis.

(Supplementary Table 4). Intramodular analysis was then performed to determine the relationship between a gene's significance and its module membership. Modules relevant to differences between MSI-H and MSS/MSI-L tumors should contain genes with a high module membership that are also significantly associated with the trait of interest. We found that these two characteristics were significantly positively correlated (P = 0.05) in 18 of the 35 modules identified (Supplementary Figure 6). To provide functional characterization, pathway analysis was performed for each of these 18 modules (Supplementary File 2).

The blueviolet module was the most significant module identified in our analysis (q = 9.07E^{-57}), and consisted of 28 genes, including central hubs *MLH1* and *EPM2AIP1* (Figure 6). This module also contained nodes for *RAB32*, *EGF* and *PTPRD*. *RAB32* is a ras proto-oncogene family member that has been previously associated with MSI-H tumors [28, 29], while both *EGF* and *PTPRD* have important roles in the regulation of cell growth and differentiation [30, 31]. Better understanding the relationship between *MLH1* and other genes in this module may provide improved insight into MSI-H tumor biology. Brown4's module eigengene was significantly positively correlated with MSI-H status (q = 5.10E^{-06}), indicating that the average expression of each gene within brown4 is increased in MSI-H versus MSS/MSI-L tumors (Figure 7). Brown4 was of particular interest given that this large module contained 81 of the 556 novel DEGs identified in our single-gene approach. We used pathway analysis to determine the molecular functionality of brown4. Here, we found that many of the Gene Ontology [24, 25] biological processes enriched in this module corresponded to apoptosis, such as positive regulation of apoptotic signaling pathways (q = 2.00E^{-03}) and intrinsic apoptotic response to DNA damage (q = 0.013). Of the top 20 genes with the greatest module membership to this module, 10 were deemed to be novel in our single-gene analysis (*ZNF628, DAPK3, TMEM259, INAFM1, RPUSD1, CAPN15, UBALD1, MAP1S, ZBTB45* and *ADAT3*). We uploaded genes within the brown4 module to CHEA3 in an attempt to identify transcription factors that may be driving this module [32]. Here, we identified *ZBTB45* as the transcription factor most likely to regulate brown4, a novel gene with high module membership (r = 0.78, P = 1.48E^{-60}). Taken together, WGCNA reveals novel insight into aberrant pathway activation between MSI-H and MSS/MSI-L tumors, which may lead to better understanding of tumor subtypes.

DISCUSSION

We demonstrate the utility of single-cell deconvolution of bulk RNA-seq to aid in the interrogation of cellular heterogeneity of tumors using single-cell RNA-

seq data from normal tissue. We first aimed to determine cell type composition differences between MSI-H and MSS/MSI-L CRC tumors. We replicated a number of known findings, including increased CD8$^+$T cells and macrophages in MSI-H tumors, which are consistent with results of a meta-analysis of MSI-H tumors across several microarray datasets [33], as well as in MMR deficient CRC tumors [34]. CD8$^+$T tumor infiltrating lymphocytes (TIL) s were also seen in greater number in MSI-H tumors [35]. Further, we were able to identify novel changes in cell composition including an increase in EEC and a reduction in stem and colonocyte cell populations in MSI-H versus MSS/MSI-L tumors. It is unclear whether these differences contribute to the etiology of MSI-H tumors. Importantly, we were able to replicate differences in cell content for three of the 13 cell types identified in TCGA-COAD (stem cell, cycling TA cells and macrophages) in independent datasets and provide some evidence for replication for an additional two cell types (CD8$^+$T cell and colonocyte) that trended in the same direction. While this provides some evidence of replication, additional validation in a larger, independent cohort should still be considered an important step to improve the generalizability of our findings.

To the best of our knowledge, changes in EEC content have not yet been described as a distinguishing feature between MSI-H and MSS/MSI-L tumors. EECs are sensory cells that play a fundamental role in the orchestration of mucosal immunity by modulating activity of several immune cell types [36]. Despite comprising only approximately 1% of the gut, these cells form the largest endocrine system in humans, while also aiding in the maintenance of the stem cell niche [36]. Previous studies have identified a subpopulation of EECs that can either migrate to the small intestinal crypt base, or remain localized there [37]. Additional research in the small intestine has shown that pre-terminal EECs are able to reconstitute LGR5$^+$ stem cells upon stem cell loss [38]. It is therefore possible that the increase in EECs reflects a population aiming to reconstitute a diminished stem cell pool, but additional work will be required to confirm this.

A reduction in colonocyte cell content was observed in MSI-H versus MSS/MSI-L tumors in both TCGA-

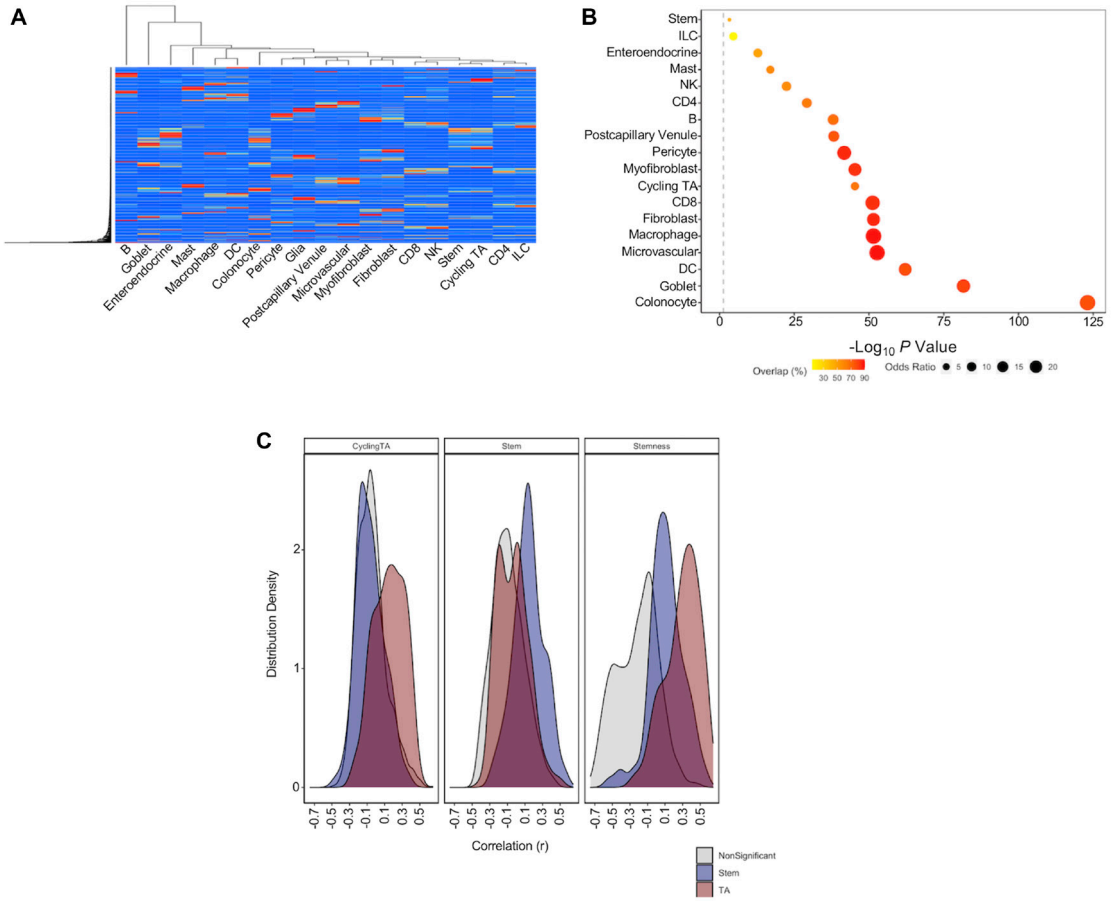

Figure 2: Single cell deconvolution of bulk RNA-seq datasets. (**A**) Heatmap to show separation of scRNA-seq cell populations based upon expression of signature genes. (**B**) Summary of enrichment analysis (one-way Fisher's exact test) for cell type markers in differential expression analysis of cell scores. Grey line represents $\log_{10}(0.05)$. Percentage overlap reflects percentage of cell type markers for a given cell type that were significant within regression of cell score. (**C**) Cell scores for cycling TA and stem cells generated using this approach, and a marker of stemness generated in a previous study were correlated to expression markers of TA cells (red), stem cells (blue) and significant markers of other colon cell types (grey).

COAD and GSE146889. Colonocytes (enterocytes of the colon) are the most abundant epithelial cell type of the colon, primarily functioning to facilitate the absorption of nutrients and water [39]. Both *CDX2* and *HNF1A* have previously been shown to play a role in enterocyte differentiation [40, 41]. Our initial analysis found that both of these genes, as well as a number of other genes required for enterocyte differentiation such as *GADD45GIP1* and

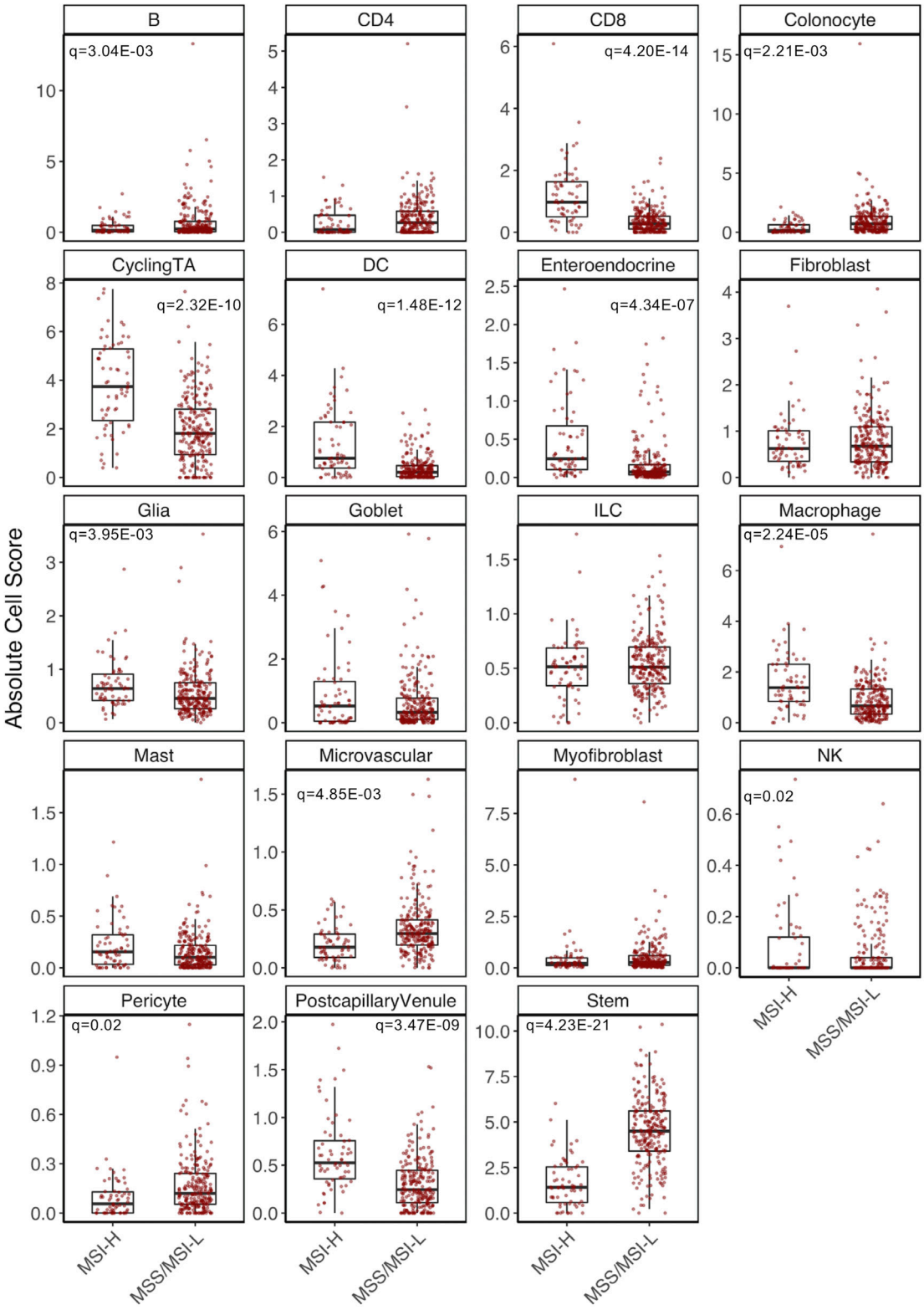

Figure 3: Cell composition analysis of TCGA-COAD dataset. Significant differences determined through linear regression analysis while adjusting for colon location, sex, consensus purity estimate and tumor stage.

ELF3 [42], were significantly downregulated in MSI-H versus MSS/MSI-L tumors prior to adjustment for cell composition. Unsurprisingly however, these cell-specific genes did not remain significant following adjustment for cell composition. It remains unclear whether the reduction in colonocyte cell population is a function of reduced expression of absorptive transcriptional activators or due to other physiological constraints such as colon location, despite efforts to correct for this in our regression models. Two intriguing additional possibilities should be considered. The first, as hypothesized with enteroendocrine cells, colonocyte precursors may undergo dedifferentiation to replenish the stem cell pool, as has been shown to occur in enterocytes of the small intestine [43]. Over time, this may significantly reduce the availability of colonocyte populations. The second is that the reduced content observed here may also contribute to, not replenish the observed stem cell reduction. Under physiological conditions, differentiated colonocytes act to metabolize butyrate, leading to the establishment of an oxygen-butryate gradient along the crypt axis. Adequately maintaining a stable oxygen-butyrate gradient is vital for the protection of stem cells, as an increase in butryare has been shown to reduce their poliferative ability [39]. Butyrate is frequently associated with reduced tumor growth and is generated through the gut microbiota. Distinct patterns of gut microbiota have been associated with MSI-H status [44]. However, it remains unknown if differences in the butyrate concentration gradient occur and if so, how they may be able to better define MSI-H tumors.

Previously, we have shown that correction for cell composition can reduce the impact of cell variation in DEG reporting of a colon organoid model exposed to ethanol [7]. Here, we use a similar approach to identify 3,193 DEGs in our regression of MSI-H versus MSS/MSI-L tumors, of which 556 were not significant prior to adjustment, and as such were deemed to be novel. Pathway analysis of DEGs displaying reduced expression in MSI-H tumors revealed an enrichment for the DNA repair pathway. Importantly, pathway analysis of genes no longer considered to be significanct following adjustment for cell compostion did not identify enrichments for repair pathways. Instead these pathways were enriched for cell-specific processes such as T-cell activation. Together, these findings indicate that adjusting for cell composition enriches for biological signals that affect the system as a whole. Hypermethylation of *MLH1* is primarily considered to be the hallmark of non-familial MSI-H tumors, while mutations in *MLH1, MSH2*, *MSH6* and to a

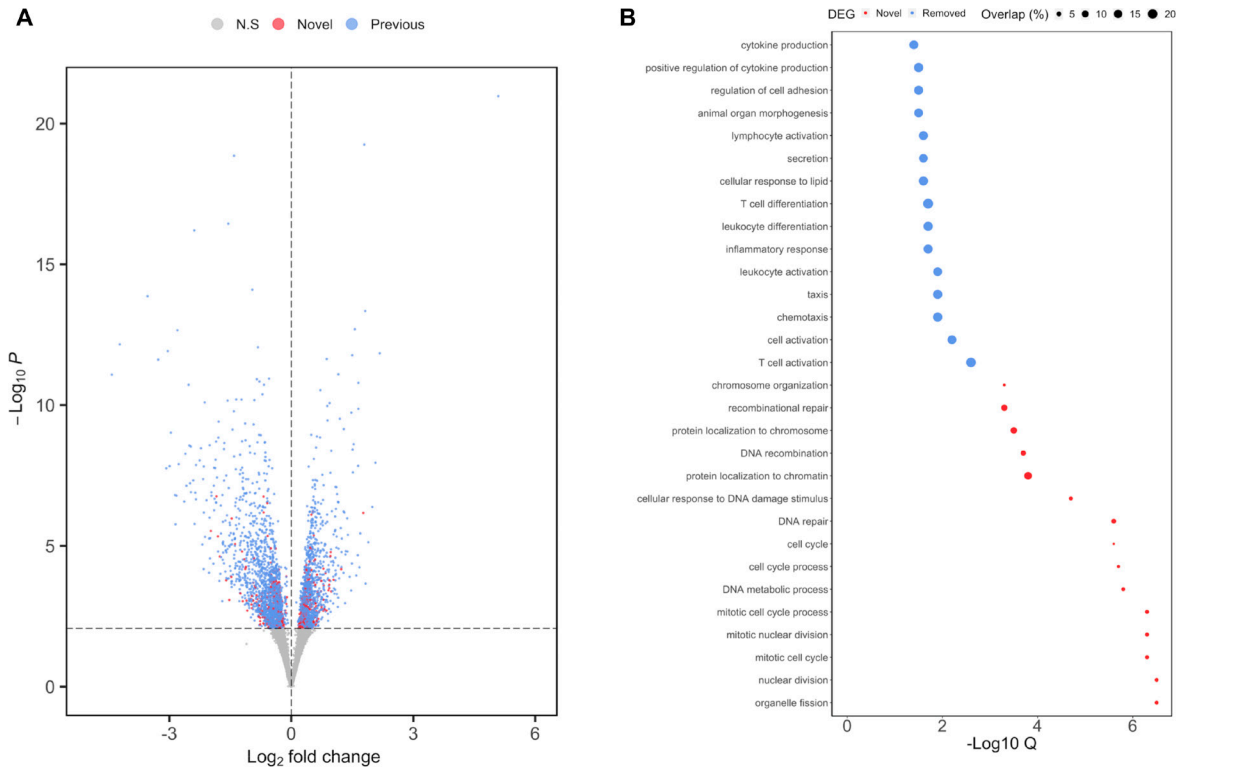

Figure 4: Differential expression analysis of MSI status in TCGA-COAD following adjustment for cell composition. (**A**) Volcano plot of results from regression on MSI status in TCGA-COAD cohort. Novel significant DEGs ($q = 0.05$) are highlighted in red, whereas DEGs identified also identified prior to adjustment for cell composition are highlighted in blue (**B**) Pathway analysis of DEGs found to be significantly reduced in MSI-H versus MSS/MSI-L tumors. Color reflects pathway analysis performed on either novel (red) or on DEGs no longer considered to be significant ($q = 0.05$) following adjustment for cell composition. Size of each circle represent the percentage of overlap between the number of DEGs and the total number of genes within a given pathway.

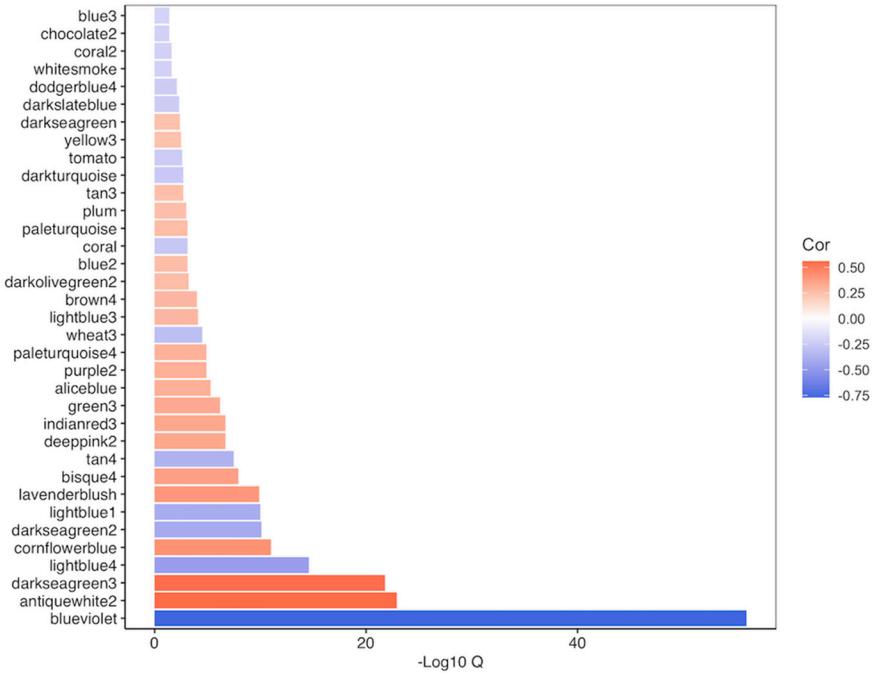

Figure 5: Overview of relationship of significant modules (q = 0.05) identified through WGCNA of TCGA-COAD following adjustment for cell composition to MSI status. Negatively correlated modules are indicative of modules primarily consisting of genes that were reduced in MSI-H versus MSS/MSI-L tumors. Of these modules, 35 were found to be significantly different between MSI-H and MSS/MSI-L tumors following strict Bonferroni correction (q = 0.05) (Supplementary File 2, Figure 5).

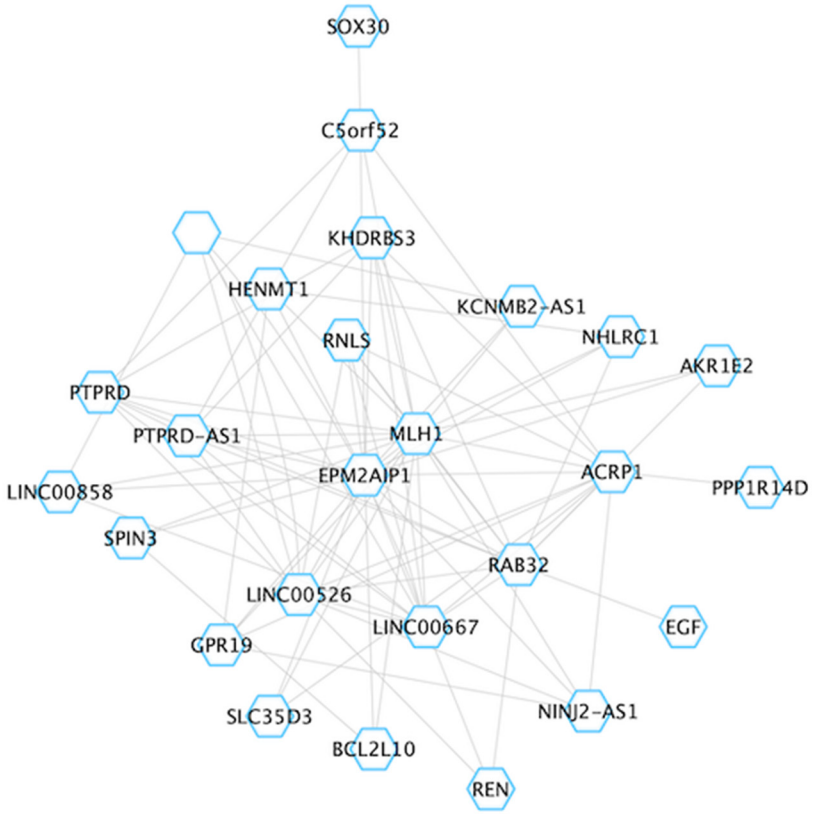

Figure 6: Overview of the blueviolet module. For visualization, the network was imported into Cytoscape. Grey lines reflect edges (connections) between hubs (genes).

lesser extent, *PMS1*, are known to drive Lynch syndrome, an inherited condition that increases the risk primarily for MSI-H-related CRC [23]. However, two recent studies have shown that reduced expression of *MSH2* and *MSH6* protein have also been identified in sporadic CRC [45, 46]. While we were unable to replicate differences in these genes in colon cancer cell lines, their identification in TCGA-COAD following the adjustment for cell composition does reflect the importance of performing deconvolution methods.

Beyond differential expression of DNA repair genes, highly significant reductions in the expression of *AGMO*, *LINC02577* and *KIF1A* in MSI-H tumors may be worth further consideration. These genes represent the three most significant novel genes that were found to be replicated in our analysis of colon cancer cell lines. Differential expression of *LINC02577* has recently been associated with CRC [47, 48], though to the best of our knowledge this gene has not been found to be differentially expressed in MSI-H tumors. Long non-coding RNAs have a variety of molecular functions, but are frequently regarded as a "sponge" for microRNAs, thus reducing their bioavailabilty to regulate the expression of downstream mRNA targets [49]. To further understand the role that this gene may play in MSI-H tumor biology, further studies should look to incorporate additional omic layers. Little is also known about the role *AGMO* may play in cancer, which primarily functions to aid in the synthesis of membrane lipids. However, recent studies have indicated a potential role for this gene in the regulation of Wnt secretion [50, 51]. Correct regulation of Wnt signaling is vital to the maintenance of healthy rates of stem cell proliferation and differentiation. Aberrant activation of the Wnt signaling pathway has been shown to be a major driver of colon cancer [52]. Further, MSS tumors are more likely to be driven through aberrant activation of Wnt signaling genes [53]. Reduced expression of *AGMO* may therefore have an important role in distinguishing MSI-H from MSS/MSI-L tumors and may contribute to the reduction in stem cell content observed by reducing Wnt activation. *KIF1A* has been associated with head and neck squamous cell carcinoma [54] and has an important role in cell division. Pathway analysis of the novel downregulated DEGs identified in our analysis revealed a number of

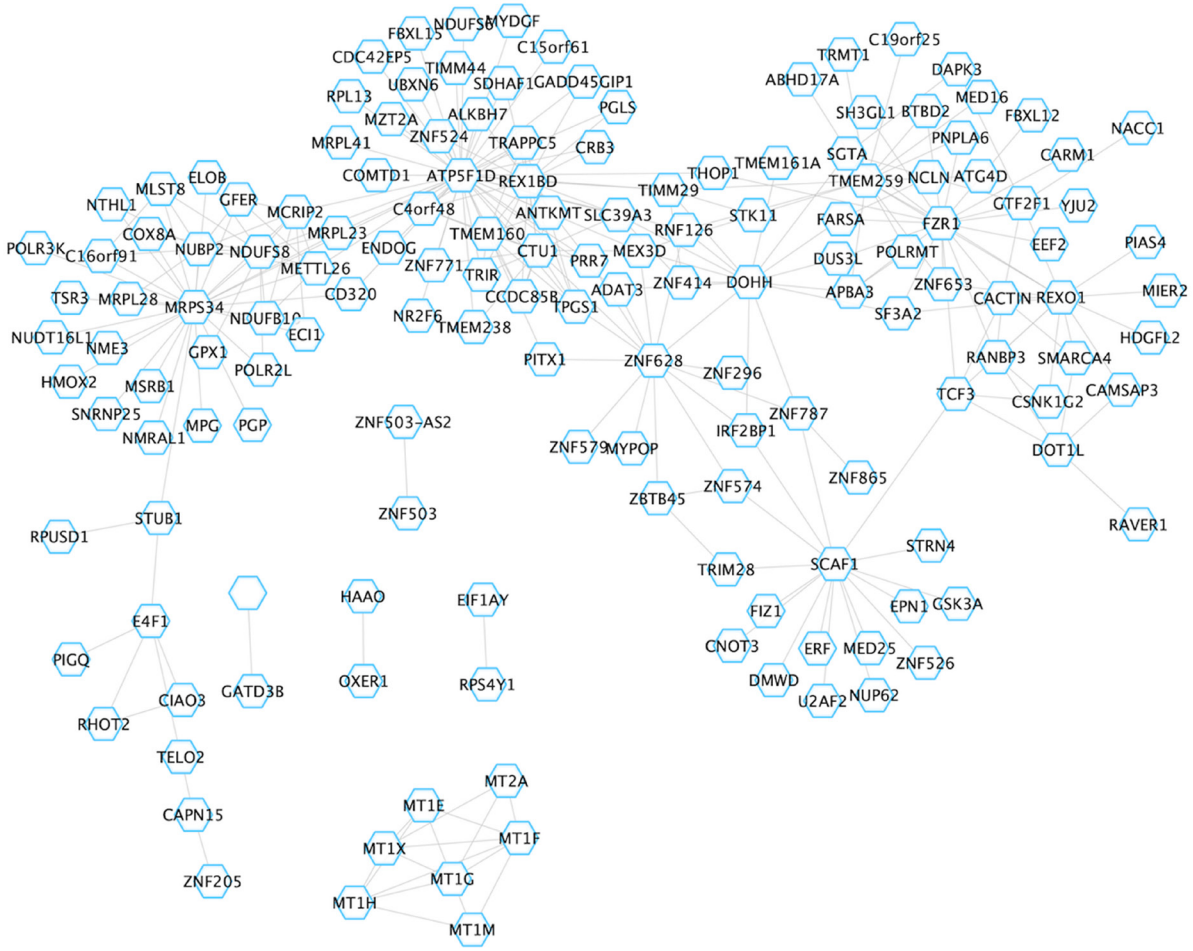

Figure 7: Overview of the brown4 module. For visualization, the network was imported into Cytoscape. Grey lines reflect edges (connections) between hubs (genes).

enrichments for cell division processes, indicating that variation in this pathway may be somewhat important in distinguishing MSI-H and MSS/MSI-L tumors. Indeed, a failure of the mismatch repair system to verify microsatellite repeat counts during cell division leads to variation in the length of their sequences. However, it remains unclear as to how reduced expression in *KIF1A* may affect MSI-H tumors.

To further interrogate the transcriptomic variation between MSI-H and MSS/MSI-L tumors, we performed WGCNA [26]. This method employes a level of guilt by association. For example, *MLH1* was central to the structure of blueviolet, the most significantly different module identified between MSI-H and MSS/MSI-L tumors. Thus, genes within this module may be critical to distinguishing MSI-H from MSS/MSI-L tumor biology. *PTPRD* was also found within this module. This gene plays an important role in the regulation of cell growth and differentiation [30], which make it an interesting target for further consideration in poorly differentiated tumors. Further, mutations in *PTPRD* have recently been shown to be frequent in T-cell rich B-cell lymphomas that display *MLH1* haploinsufficiency [55], adding weight to the validity of our co-expression analysis. Together, these data highlight the potential importance of this gene and its interplay with *MLH1* in MSI-H tumors.

The identification of the brown4 module was of particular interest given that it contained 15.64% of the novel DEGs identified within our single-gene analysis, including *ZBTB45*. *ZBTB45* was identified as one of the top 20 most significant hub genes for brown4, indicating an important role for this gene in a module enriched for apoptosis-related pathways. This result was confirmed using CHEA3 [32], an online tool that aims to identify, rank and prioritize regulatory transcription factors that may be affecting a given set of genes. Increased rates of apoptosis have previously been reported in MSI-H versus MSS or MSI-L tumors [56], where the authors were unable to fully attribute the increased apoptotic index to an increase in TILs. Improving our understanding of transcriptional networks that drive changes in apoptosis between tumor subtypes may help to explain the survival advantage generally observed in MSI-H tumors [56].

There are a number of limitations to this study. In our analysis, we used a consensus measure of tumor purity to control for tumor heterogeneity across samples. However, while we are able to infer the cell types present within TCGA-COAD and simulatenously account for tumor heterogeneity, we were not able to specify the origins of these cell types, i.e., intra-tumoral versus intra-epithelial. In many instances, our definition of cell type populations is also limited to either the resolution defined by the scRNA-seq study used for deconvolution [9], or the resolution of cell populations that could be delineated accurately through deconvolution. Subpopulations of EECs could not be defined using our approach. EECs consists of multiple sub-lineages, often classified by their principal hormone product. These secretory cell populations vary in density across the gastrointestinal tract [57]. We also do not consider the potentially confounding effects of other cancer related molecular pathways in our analysis, such as CpG Island Methylator Phenotype (CIMP). There is considerable overlap between MSI-H status and high levels of CIMP (CIMP-H). However, CIMP-H has also been observed in a subset of MSS tumors [58]. Finally, we do not consider the role of somatic mutations in driving expression. However, adopting such approaches in future, larger studies may provide additional insight into MSI-H tumors.

In summary, we employ a machine learning approach to deconvolute MSI-H and MSS/MSI-L tumor gene expression across TCGA-COAD and two additional cohorts. We identify novel changes in cell composition for EECs and colonocytes that suggests previously uncharacterized roles for these cell populations in contributing to MSI-H tumor development. Finally we use both single-gene and network analysis to identify several novel genes that may play an important role in MSI-H tumor biology.

MATERIALS AND METHODS

RNA-seq data pre-processing

HT-Seq count and phenotype data were downloaded from the R package TCGAbiolinks [59]. For data collection, pre-processing and alignment details, please refer to the original publications [8, 59, 60]. Single-cell deconvolution of bulk RNA-seq has been shown to perform best on larger datasets [6]. Thus, we first used a total 409 samples to estimate cell populations in CRC tumors. For the analysis of gene expression differences in MSI-H versus MSS/MSI-L tumors, a total of 294 samples were considered (MSI-H = 63, MSS = 178, MSI-L = 53). Samples were removed if they had missing phenotype information for MSI status, consensus purity estimates, tumor stage, or lacked specific colon location information, i.e., colon location data was labelled either “NA” or “colon, NOS”. Cancer stages were broadly categorized into main hierarchical groupings (stage 1–4). Samples were also broadly categorized into one of three location groupings: left (descending, sigmoid, splenic flexure), right (ascending, cecum, hepatic flexure) and transverse, which were considered based upon the developmental origins of colon tissue. Given that the transverse colon is derived from either midgut or hindgut (depending on which region of the transverse colon), we considered this a distinct colon segment. Consensus purity estimates were downloaded from a previously published analysis of TCGA-COAD [61].

We identified a second, smaller CRC cohort with available MSI data on June 1st, 2020, by searching

gene expression omnibus (GEO) [62] using keywords "MSI" and "colorectal cancer" and only considered data generated using RNA-seq. This dataset also contained RNA-seq count data for endometrial cancer, which was not considered here. Of note, the majority of MSI-H individuals considered in this dataset were putative Lynch syndrome (32/36), while the majority of TCGA-COAD is considered to be sporadic MSI-H tumors. Raw counts were downloaded from GEO, accession: GSE146889 for downstream analysis.

For further validation, we also downloaded RNA-seq count and TPMs from the Broad CCLE website (https://portals.broadinstitute.org/ccle). Details for RNA-seq library generation and pre-processing can be found in the original article [11].

Single-cell deconvolution of bulk RNA-seq data

We downloaded publicly available scRNA-seq data derived from normal colon biopsies [9]. To reduce heterogeneity in single-cell expression, only cells derived from healthy colon were considered for this analysis. Transcripts per million (TPM)s were generated using scater [63]. Given the size of the dataset, cells were randomly downsized to permit upload to CIBERSORTx [6]. For model evaluation, cell composition scores were correlated to known gene expression markers in an attempt to determine relative performance.

The final dataset consisted of 19,567 genes across 5,412 cells. Multiple similar cell types were merged to aid in this analysis, for example: B cells (plasma, germinal center, follicular). M cells were removed due to low abundance in the original analysis ($n = 10$). Tuft cells were removed after multiple attempts to define population led to inadequate identification of cell population markers. Secretory TA cells were removed given their similar transcriptional profile to mature secretory cell populations and cycling TA cells. TA1 and TA2 cell populations were also not considered, given their similarity to cycling TA cells. Epithelial progenitor cells of goblets and colonocytes were also removed in favor of their mature cell populations to aid in their distinction from cycling TA cells. A total of 19 distinct cell types were considered in the final analysis. We note that this represents a reduction in granularity from the 51 unique cell types identified in scRNA-seq analysis of normal colon [9].

Following upload to CIBERSORTx [6], single cells were clustered based upon overall similarity of expression using default parameters, with the following notable exceptions: minimum expression = 0; number of significant genes to define cell type = 150; sampling = 1, $q = 0.001$. Following this, TPMs from TCGA-COAD samples were imported and cell composition scores were estimated. For deconvolution the following parameters were set: 500 permutations; quantile normalization was disabled; S-mode was set for batch correction; scores were generated in absolute mode. Cell scores were then centered and scaled about the mean prior to incorporation as covariates in a regression model. The same parameters were also used for deconvolution of GSE146889 [10].

For colon cancer cell lines [11], a total of 3,988 cells across four epithelial cell types were considered for deconvolution (cycling TA, stem cell, colonocyte progenitors and immature goblets). Deconvolution was performed as above, with one exception: the number of genes used to define cell types was set to a range of 100–600.

Regression analysis

Differentially expressed genes DEGs were identified through regression analysis performed in DESeq2 [64]. Several regression models were used in this study.

Differential expression

For the analysis of cell type agnostic differential expression we used the following model:

Expression ~ *Stage* + *Sex* + Consensus Purity Estimate + *Colon Location* + *Scores* + *MSI*

Where: expression = gene expression of each gene for each individual; score = cell score (continuous variable); stage = cancer stage (factored 1–4); sex = biological sex; consensus purity estimate = tumor purity (continuous variable); colon location = location of biopsy taken (factor); MSI = microsatellite instability status.

Cell composition

To analyze differences in absolute cell scores between MSI-H and MSS/MSI-L a linear regression model was used.

Score ~ *Stage* + *Sex* + *Consensus Purity Estimate* + *Colon Location* + *MSI*

Where: score = cell score for each cell type and individual (continuous); stage = cancer stage (factored 1-4); sex = biological sex; consensus purity estimate = tumor purity (continuous variable); colon location = location of biopsy taken (factor); MSI = microsatellite instability status.

WGCNA

Genes with a count of less than 10 in 150 samples were filtered, leaving a total of 17,361 genes for downstream network analysis. Raw counts were converted into counts per million and the effects of tumor stage (factor), sex, colon location (factor), tumor purity and cell compositon were regressed out prior to WGCNA [26] using limma [65]. Heirarchical clustering analysis was used to determine outliers based on their average dissimilarity, which led to the removal of four samples. A total of 289 samples and 17,361 genes were therefore used to construct the network. WGCNA was performed under default settings with the exception of the following parameters: a soft power of four was chosen, where the degree of independence was

determined to be 0.876; blocksize was set to the number of genes used; signed hybrid and pearson correlation were preferred; minimum module size was set to 10; deep split was set to 3 and strongly correlated modules ($r = 0.7$) were merged prior to association testing.

Abbreviations

5-FU: 5-fluorouracil; CCLE: Cancer Cell Line Encyclopedia; CIMP: CpG Island Methylator Phenotype; CIMP-H: (CpG Island Methylator Phenotype high); CIN: chromosomal instability; CRC: colorectal cancer; DEG: differentially expressed gene; EEC: enteroendocrine cell; MMR: mismatch repair; MSI: microsatellite instability; MSI-H: microsatellite instability high; MSI-L: microsatellite instability low; MSS: microsatellite stable; RNA-seq: RNA-sequencing; scRNA-seq: single-cell RNA-sequencing; TA: transit amplifying; TCGA-COAD: The Cancer Genome Atlas Colon Adenocarcinoma; TIL: tumor infiltrating lymphocyte; TPM: transcript per million.

Author contributions

MD was responsible for conception, design, data visualization and analysis of data within the project. MD and GC were responsible for interpretation of results and drafting and revising the manuscript.

CONFLICTS OF INTEREST

Authors have no conflicts of interest to declare.

FUNDING

This work was supported by funding through NIH grants: NIH/NCI CA143237 and NIH/NCI CA204279.

REFERENCES

1. Markowitz SD, Bertagnolli MM. Molecular origins of cancer. Molecular basis of colorectal cancer. N Engl J Med. 2009; 361:2449–60. https://doi.org/10.1056/NEJMra0804588. [PubMed]
2. Vargas-Rondon N, Villegas VE, Rondon-Lagos M. The Role of Chromosomal Instability in Cancer and Therapeutic Responses. Cancers (Basel). 2017; 10:4. https://doi.org/10.3390/cancers10010004. [PubMed]
3. Zaanan A, Shi Q, Taieb J, Alberts SR, Meyers JP, Smyrk TC, Julie C, Zawadi A, Tabernero J, Mini E, Goldberg RM, Folprecht G, Van Laethem JL, et al. Role of Deficient DNA Mismatch Repair Status in Patients With Stage III Colon Cancer Treated With FOLFOX Adjuvant Chemotherapy: A Pooled Analysis From 2 Randomized Clinical Trials. JAMA Oncol. 2018; 4:379–83. https://doi.org/10.1001/jamaoncol.2017.2899. [PubMed]
4. Le DT, Uram JN, Wang H, Bartlett BR, Kemberling H, Eyring AD, Skora AD, Luber BS, Azad NS, Laheru D, Biedrzycki B, Donehower RC, Zaheer A, et al. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. N Engl J Med. 2015; 372:2509–20. https://doi.org/10.1056/NEJMoa1500596. [PubMed]
5. Narayanan S, Kawaguchi T, Peng X, Qi Q, Liu S, Yan L, Takabe K. Tumor Infiltrating Lymphocytes and Macrophages Improve Survival in Microsatellite Unstable Colorectal Cancer. Sci Rep. 2019; 9:13455. https://doi.org/10.1038/s41598-019-49878-4. [PubMed]
6. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, Khodadoust MS, Esfahani MS, Luca BA, Steiner D, Diehn M, Alizadeh AA. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol. 2019; 37:773–82. https://doi.org/10.1038/s41587-019-0114-2. [PubMed]
7. Devall M, Plummer SJ, Bryant J, Jennelle LT, Eaton S, Dampier CH, Huyghe JR, Peters U, Powell SM, Casey G. Ethanol exposure drives colon location specific cell composition changes in a normal colon crypt 3D organoid model. Sci Rep. 2021; 11:432. https://doi.org/10.1038/s41598-020-80240-1. [PubMed]
8. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012; 487:330–37. https://doi.org/10.1038/nature11252. [PubMed]
9. Smillie CS, Biton M, Ordovas-Montanes J, Sullivan KM, Burgin G, Graham DB, Herbst RH, Rogel N, Slyper M, Waldman J, Sud M, Andrews E, Velonias G, et al. Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis. Cell. 2019; 178:714–30.e722. https://doi.org/10.1016/j.cell.2019.06.029. [PubMed]
10. Diguardo M, Zarei S, Jackson R, Nair A, Davila J, Halling K. 25. RNAseq analysis of PD-L1 expression in colorectal and endometrial tumors and correlation with microsatellite instability. Cancer Genetics. 2018; 45:226–27.
11. Ghandi M, Huang FW, Jane-Valbuena J, Kryukov GV, Lo CC, McDonald ER 3rd, Barretina J, Gelfand ET, Bielski CM, Li H, Hu K, Andreev-Drakhlin AY, Kim J, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature. 2019; 569:503–08. https://doi.org/10.1038/s41586-019-1186-3. [PubMed]
12. Li K, Luo H, Huang L, Luo H, Zhu X. Microsatellite instability: a review of what the oncologist should know. Cancer Cell Int. 2020; 20:16. https://doi.org/10.1186/s12935-019-1091-8. [PubMed]
13. Liu Q, Zhang B. Integrative Omics Analysis Reveals Post-Transcriptionally Enhanced Protective Host Response in Colorectal Cancers with Microsatellite Instability. J Proteome Res. 2016; 15:766–76. https://doi.org/10.1021/acs.jproteome.5b00847. [PubMed]
14. Guinney J, Dienstmann R, Wang X, de Reynies A, Schlicker A, Soneson C, Marisa L, Roepman P, Nyamundanda G, Angelino P, Bot BM, Morris JS, Simon IM, et al. The

consensus molecular subtypes of colorectal cancer. Nat Med. 2015; 21:1350–56. https://doi.org/10.1038/nm.3967. [PubMed]

15. Franzen O, Gan LM, Bjorkegren JLM. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database (Oxford). 2019; 2019:baz046. https://doi.org/10.1093/database/baz046. [PubMed]

16. Sturm G, Finotello F, Petitprez F, Zhang JD, Baumbach J, Fridman WH, List M, Aneichyk T. Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. Bioinformatics. 2019; 35:i436–45. https://doi.org/10.1093/bioinformatics/btz363. [PubMed]

17. Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol. 2017; 18:220. https://doi.org/10.1186/s13059-017-1349-1. [PubMed]

18. Finotello F, Mayer C, Plattner C, Laschober G, Rieder D, Hackl H, Krogsdam A, Loncova Z, Posch W, Wilflingseder D, Sopper S, Ijsselsteijn M, Brouwer TP, et al. Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. Genome Med. 2019; 11:34. https://doi.org/10.1186/s13073-019-0638-6. [PubMed]

19. Racle J, de Jonge K, Baumgaertner P, Speiser DE, Gfeller D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. Elife. 2017; 6:e26476. https://doi.org/10.7554/eLife.26476. [PubMed]

20. Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, Selves J, Laurent-Puig P, Sautes-Fridman C, Fridman WH, de Reyniès A. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biol. 2016; 17:218. https://doi.org/10.1186/s13059-016-1070-5. [PubMed]

21. Malta TM, Sokolov A, Gentles AJ, Burzykowski T, Poisson L, Weinstein JN, Kaminska B, Huelsken J, Omberg L, Gevaert O, Colaprico A, Czerwińska P, Mazurek S, et al. Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. Cell. 2018; 173:338–54. https://doi.org/10.1016/j.cell.2018.03.034. [PubMed]

22. Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. Cell. 2015; 160:48–61. https://doi.org/10.1016/j.cell.2014.12.033. [PubMed]

23. Carneiro da Silva F, Ferreira JR, Torrezan GT, Figueiredo MC, Santos EM, Nakagawa WT, Brianese RC, Petrolini de Oliveira L, Begnani MD, Aguiar-Junior S, Rossi BM, Ferreira FDO, Carraro DM. Clinical and Molecular Characterization of Brazilian Patients Suspected to Have Lynch Syndrome. PLoS One. 2015; 10:e0139753. https://doi.org/10.1371/journal.pone.0139753. [PubMed]

24. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000; 25:25–29. https://doi.org/10.1038/75556. [PubMed]

25. Carbon S, Douglass E, Good BM, Unni DR, Harris NL, Mungall CJ, Basu S, Chisholm RL, Dodson RJ, Hartline E, Fey P, Thomas PD, Albou LP, et al, and Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res. 2021; 49:D325–34. https://doi.org/10.1093/nar/gkaa1113. [PubMed]

26. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol. 2005; 4:Article17. https://doi.org/10.2202/1544-6115.1128. [PubMed]

27. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ, von Mering C. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019; 47:D607–13. https://doi.org/10.1093/nar/gky1131. [PubMed]

28. Shibata D, Mori Y, Cai K, Zhang L, Yin J, Elahi A, Hamelin R, Wong YF, Lo WK, Chung TK, Sato F, Karpeh MS Jr, Meltzer SJ. RAB32 hypermethylation and microsatellite instability in gastric and endometrial adenocarcinomas. Int J Cancer. 2006; 119:801–06. https://doi.org/10.1002/ijc.21912. [PubMed]

29. Mori Y, Yin J, Sato F, Sterian A, Simms LA, Selaru FM, Schulmann K, Xu Y, Olaru A, Wang S, Deacu E, Abraham JM, Young J, et al. Identification of genes uniquely involved in frequent microsatellite instability colon carcinogenesis by expression profiling combined with epigenetic scanning. Cancer Res. 2004; 64:2434–38. https://doi.org/10.1158/0008-5472.can-03-3508. [PubMed]

30. Funato K, Yamazumi Y, Oda T, Akiyama T. Tyrosine phosphatase PTPRD suppresses colon cancer cell migration in coordination with CD44. Exp Ther Med. 2011; 2:457–63. https://doi.org/10.3892/etm.2011.231. [PubMed]

31. Schuldiner M, Yanuka O, Itskovitz-Eldor J, Melton DA, Benvenisty N. Effects of eight growth factors on the differentiation of cells derived from human embryonic stem cells. Proc Natl Acad Sci U S A. 2000; 97:11307–12. https://doi.org/10.1073/pnas.97.21.11307. [PubMed]

32. Keenan AB, Torre D, Lachmann A, Leong AK, Wojciechowicz ML, Utti V, Jagodnik KM, Kropiwnicki E, Wang Z, Ma'ayan A. ChEA3: transcription factor enrichment analysis by orthogonal omics integration. Nucleic Acids Res. 2019; 47:W212–24. https://doi.org/10.1093/nar/gkz446. [PubMed]

33. Wang H, Wang X, Xu L, Zhang J, Cao H. Analysis of the transcriptomic features of microsatellite instability subtype colon cancer. BMC Cancer. 2019; 19:605. https://doi.org/10.1186/s12885-019-5802-2. [PubMed]

34. Nakajima T, Uehara T, Iwaya M, Kobayashi Y, Maruyama Y, Ota H. Characterization of LGR5 expression in poorly differentiated colorectal carcinoma with mismatch repair protein deficiency. BMC Cancer. 2020; 20:319. https://doi.org/10.1186/s12885-020-06791-8. [PubMed]

35. Duhen T, Duhen R, Montler R, Moses J, Moudgil T, de Miranda NF, Goodall CP, Blair TC, Fox BA, McDermott JE, Chang SC, Grunkemeier G, Leidner R, et al. Co-expression of CD39 and CD103 identifies tumor-reactive CD8 T cells in human solid tumors. Nat Commun. 2018; 9:2724. https://doi.org/10.1038/s41467-018-05072-0. [PubMed]

36. Worthington JJ, Reimann F, Gribble FM. Enteroendocrine cells-sensory sentinels of the intestinal environment and orchestrators of mucosal immunity. Mucosal Immunol. 2018; 11:3–20. https://doi.org/10.1038/mi.2017.73. [PubMed]

37. Sei Y, Lu X, Liou A, Zhao X, Wank SA. A stem cell marker-expressing subset of enteroendocrine cells resides at the crypt base in the small intestine. Am J Physiol Gastrointest Liver Physiol. 2011; 300:G345–56. https://doi.org/10.1152/ajpgi.00278.2010. [PubMed]

38. Jadhav U, Saxena M, O'Neill NK, Saadatpour A, Yuan GC, Herbert Z, Murata K, Shivdasani RA. Dynamic Reorganization of Chromatin Accessibility Signatures during Dedifferentiation of Secretory Precursors into Lgr5+ Intestinal Stem Cells. Cell Stem Cell. 2017; 21:65–77.e5. https://doi.org/10.1016/j.stem.2017.05.001. [PubMed]

39. Allaire JM, Crowley SM, Law HT, Chang SY, Ko HJ, Vallance BA. The Intestinal Epithelium: Central Coordinator of Mucosal Immunity. Trends Immunol. 2018; 39:677–96. https://doi.org/10.1016/j.it.2018.04.002. [PubMed]

40. Grainger S, Savory JG, Lohnes D. Cdx2 regulates patterning of the intestinal epithelium. Dev Biol. 2010; 339:155–65. https://doi.org/10.1016/j.ydbio.2009.12.025. [PubMed]

41. D'Angelo A, Bluteau O, Garcia-Gonzalez MA, Gresh L, Doyen A, Garbay S, Robine S, Pontoglio M. Hepatocyte nuclear factor 1alpha and beta control terminal differentiation and cell fate commitment in the gut epithelium. Development. 2010; 137:1573–82. https://doi.org/10.1242/dev.044420. [PubMed]

42. Kwon MC, Koo BK, Kim YY, Lee SH, Kim NS, Kim JH, Kong YY. Essential role of CR6-interacting factor 1 (Crif1) in E74-like factor 3 (ELF3)-mediated intestinal development. J Biol Chem. 2009; 284:33634–41. https://doi.org/10.1074/jbc.M109.059840. [PubMed]

43. Tetteh PW, Basak O, Farin HF, Wiebrands K, Kretzschmar K, Begthel H, van den Born M, Korving J, de Sauvage F, van Es JH, van Oudenaarden A, Clevers H. Replacement of Lost Lgr5-Positive Stem Cells through Plasticity of Their Enterocyte-Lineage Daughters. Cell Stem Cell. 2016; 18:203–13. https://doi.org/10.1016/j.stem.2016.01.001. [PubMed]

44. Purcell RV, Visnovska M, Biggs PJ, Schmeier S, Frizelle FA. Distinct gut microbiome patterns associate with consensus molecular subtypes of colorectal cancer. Sci Rep. 2017; 7:11590. https://doi.org/10.1038/s41598-017-11237-6. [PubMed]

45. Westwood A, Glover A, Hutchins G, Young C, Brockmoeller S, Robinson R, Worrilow L, Wallace D, Rankeillor K, Adlard J, Quirke P, West N. Additional loss of MSH2 and MSH6 expression in sporadic deficient mismatch repair colorectal cancer due to MLH1 promoter hypermethylation. J Clin Pathol. 2019; 72:443–47. https://doi.org/10.1136/jclinpath-2018-205687. [PubMed]

46. Wang T, Stadler ZK, Zhang L, Weiser MR, Basturk O, Hechtman JF, Vakiani E, Saltz LB, Klimstra DS, Shia J. Immunohistochemical null-phenotype for mismatch repair proteins in colonic carcinoma associated with concurrent MLH1 hypermethylation and MSH2 somatic mutations. Fam Cancer. 2018; 17:225–28. https://doi.org/10.1007/s10689-017-0031-9. [PubMed]

47. Berral-Gonzalez A, Riffo-Campos AL, Ayala G. OMICfpp: a fuzzy approach for paired RNA-Seq counts. BMC Genomics. 2019; 20:259. https://doi.org/10.1186/s12864-019-5496-5. [PubMed]

48. Wan J, Deng D, Wang X, Wang X, Jiang S, Cui R. LINC00491 as a new molecular marker can promote the proliferation, migration and invasion of colon adenocarcinoma cells. Onco Targets Ther. 2019; 12:6471–80. https://doi.org/10.2147/OTT.S201233. [PubMed]

49. Militello G, Weirick T, John D, Doring C, Dimmeler S, Uchida S. Screening and validation of lncRNAs and circRNAs as miRNA sponges. Brief Bioinform. 2017; 18:780–88. https://doi.org/10.1093/bib/bbw053. [PubMed]

50. Petko J, Thileepan M, Sargen M, Canfield V, Levenson R. Alternative splicing of the Wnt trafficking protein, Wntless and its effects on protein-protein interactions. BMC Mol Cell Biol. 2019; 20:22. https://doi.org/10.1186/s12860-019-0208-1. [PubMed]

51. Duncan AR, Gonzalez DP, Del Viso F, Robson A, Khokha MK, Griffin JN. Alkylglycerol monooxygenase, a heterotaxy candidate gene, regulates left-right patterning via Wnt signaling. Dev Biol. 2019; 456:1–7. https://doi.org/10.1016/j.ydbio.2019.07.019. [PubMed]

52. Schatoff EM, Leach BI, Dow LE. Wnt Signaling and Colorectal Cancer. Curr Colorectal Cancer Rep. 2017; 13:101–10. https://doi.org/10.1007/s11888-017-0354-9. [PubMed]

53. Pai P, Rachagani S, Dhawan P, Batra SK. Mucins and Wnt/beta-catenin signaling in gastrointestinal cancers: an unholy nexus. Carcinogenesis. 2016; 37:223–32. https://doi.org/10.1093/carcin/bgw005. [PubMed]

54. Demokan S, Chang X, Chuang A, Mydlarz WK, Kaur J, Huang P, Khan Z, Khan T, Ostrow KL, Brait M, Hoque MO, Liegeois NJ, Sidransky D, et al. KIF1A and EDNRB are differentially methylated in primary HNSCC and salivary rinses. Int J Cancer. 2010; 127:2351–59. https://doi.org/10.1002/ijc.25248. [PubMed]

55. Patel R, Zhang L, Desai A, Hoenerhoff MJ, Kennedy LH, Radivoyevitch T, Ban Y, Chen XS, Gerson SL, Welford SM. Mlh1 deficiency increases the risk of hematopoietic malignancy after simulated space radiation exposure. Leukemia. 2019; 33:1135–47. https://doi.org/10.1038/s41375-018-0269-8. [PubMed]

56. Michael-Robinson JM, Biemer-Huttmann A, Purdie DM, Walsh MD, Simms LA, Biden KG, Young JP, Leggett BA, Jass JR, Radford-Smith GL. Tumour infiltrating lymphocytes and apoptosis are independent features in colorectal cancer stratified according to microsatellite

instability status. Gut. 2001; 48:360–66. https://doi.org/10.1136/gut.48.3.360. [PubMed]

57. Beumer J, Puschhof J, Bauza-Martinez J, Martinez-Silgado A, Elmentaite R, James KR, Ross A, Hendriks D, Artegiani B, Busslinger GA, Ponsioen B, Andersson-Rolf A, Saftien A, et al. High-Resolution mRNA and Secretome Atlas of Human Enteroendocrine Cells. Cell. 2020; 181:1291–1306.e19. https://doi.org/10.1016/j.cell.2020.04.036. [PubMed]

58. Barault L, Charon-Barra C, Jooste V, de la Vega MF, Martin L, Roignot P, Rat P, Bouvier AM, Laurent-Puig P, Faivre J, Chapusot C, Piard F. Hypermethylator phenotype in sporadic colon cancer: study on a population-based series of 582 cases. Cancer Res. 2008; 68:8541–46. https://doi.org/10.1158/0008-5472.CAN-08-1171. [PubMed]

59. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot TS, Malta TM, Pagnotta SM, Castiglioni I, Ceccarelli M, Bontempi G, Noushmehr H. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 2016; 44:e71. https://doi.org/10.1093/nar/gkv1507. [PubMed]

60. Knijnenburg TA, Wang L, Zimmermann MT, Chambwe N, Gao GF, Cherniack AD, Fan H, Shen H, Way GP, Greene CS, Liu Y, Akbani R, Feng B, et al, and Cancer Genome Atlas Research Network. Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. Cell Rep. 2018; 23:239–54. https://doi.org/10.1016/j.celrep.2018.03.076. [PubMed]

61. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. Nat Commun. 2015; 6:8971. https://doi.org/10.1038/ncomms9971. [PubMed]

62. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Research. 2013; 41:D991–95. https://doi.org/10.1093/nar/gks1193. [PubMed]

63. McCarthy DJ, Campbell KR, Lun ATL, Wills QF. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinformatics. 2017; 33:1179–86. https://doi.org/10.1093/bioinformatics/btw777. [PubMed]

64. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014; 15:550. https://doi.org/10.1186/s13059-014-0550-8. [PubMed]

65. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015; 43:e47. https://doi.org/10.1093/nar/gkv007. [PubMed]