Characterization of Somatic Mutations in the Normal Colon Using Duplex Sequencing to Evaluate Colorectal Cancer Risk

Alexis Blokker

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2025

Committee:

Rosana Risques

Scott Kennedy

Vera Paulson

Program Authorized to Offer Degree:

Laboratory Medicine & Pathology

University of Washington

Abstract

Characterization of Somatic Mutations in the Normal Colon using Duplex Sequencing to Evaluate Colorectal Cancer Risk

Alexis Blokker

Chair of the Supervisory Committee:

Rosana Risques

Department of Laboratory Medicine & Pathology

The accumulation of somatic mutations in normal tissues has been established over the last decade. However, we lack an understanding of how these mutations might predispose individuals to cancer. In this study we expand on previous findings to demonstrate that duplex sequencing enables the detection of variants at extremely variant allele fractions (VAFs) in normal colon from patients without cancer or polyps, with polyps, and those with colorectal cancer (CRC). By applying advanced computational techniques, we characterize the potential pathogenicity of these mutations, developing a framework for assessing positive selection in the normal colon. Our data indicates that individuals with CRC have a higher frequency of mutations in key CRC driver genes in their normal colon, specifically in *APC*, *FBXW7,* and *PIK3CA,* than patients without cancer. Furthermore, patients with cancer exhibit a higher frequency of pathogenic large clones (pathogenic variants with >1 duplex read) in *KRAS* and *TP53* in normal colon, indicating an increased prevalence of positive selection and clonal expansions driven by mutations in those genes. These mutations are not random, but cluster in colorectal cancer gene hotspots. These results reveal that the clonal landscape of CRC driver genes differs between the normal colon of individuals with cancer and without cancer, providing insight into how the somatic genome may predict a patient's risk for developing cancer.

Introduction

Colorectal cancer (CRC) is the third most common cancer diagnosis worldwide and the second leading cause of cancer related deaths[1]. Each year, colorectal cancer accounts for over 150,000 new cases in the United States alone and is projected to result in more than 52,000 deaths in 2025 [2]. Although widespread use of colonoscopies for preventive screening has improved early detection of colorectal cancer, a significant proportion of cases are still diagnosed at advanced stages, contributing to the high mortality rate [3]. Advancing our understanding of CRC initiation and progression in normal tissues is essential for the development of more effective strategies for early detection and prevention. From a pathological perspective, colorectal cancer develops through a stepwise process. First, hyperproliferation of cells initiates in the normal tissue. These early clones can escape tissue constraints and result in small or large benign growths known as polyps[4-6]. Not all polyps will become cancer, in fact only 5-10% will become cancerous[6,7,8]. However, more than half of all colorectal cancer cases originate from adenomatous polyps, and if undetected, these lesions can progress to advanced-stage disease and metastasize to distant tissues [6,7].

While the progression of colorectal cancer from adenomatous polyps to adenocarcinoma is well studied, we lack an understanding of what happens in the normal tissues prior to polyp formation that contributes to the initiation of colorectal cancer [9-12]. To study cancer at its earliest phase we need to acknowledge cancer as an evolutionary process. Cancer develops and progresses through the accumulation and subsequent selection of mutations that confer phenotypic advantage to the tumor cells[13-16]. Therefore, the investigation of the somatic landscape of normal colon tissue is likely to provide critical insights into somatic evolution and the earliest events that contribute to colorectal cancer formation.

Recent research has revealed that normal human tissues harbor abundant mutations, which accumulate through life as a consequence of aging and environmental exposures[17,18]. These studies have highlighted two main characteristics to look for when examining the somatic genome in normal tissues. The first characteristic is the type of mutations that comprise the mutational signature(s) that can arise from endogenous and exogenous factors. For example,

human aging produces two distinct mutational signatures [19-22]. The second characteristic is the clonal expansions of mutant clones, indicating the positive selection of mutations that are phenotypically advantageous [23-25]. This is observed through the enrichment for pathogenic mutations and their comparison with driver mutations found in cancer. However, while it is now well established that somatic mutations inevitably accumulate in normal tissues, including the colon, little is known about how these mutations contribute to cancer predisposition [13].

Studies on tumor mutations have provided insight regarding the earliest mutations that may have occurred in the normal tissues by identifying the earliest driver genes leading to colorectal adenocarcinoma [12,17,26,27]. Specifically, phylogenetic reconstructions of colorectal cancer evolution based on the sequencing of hundreds of tumors has revealed that the earliest driver mutations take place in *APC*, *KRAS*, *PIK3CA*, *TP53*, and *FBXW7* [12]. Most importantly, these mutations are estimated to occur very early in life, providing a rationale for the detection of these mutant clones in the normal tissue for predicting CRC risk [12]. These key genes can be evaluated in the normal colon of individuals with and without cancer to assess differences in the somatic mutation landscape.

The study of somatic mutations in normal tissue has been hampered by the small mutant clones size and their presence beneath the limit of detection by conventional sequencing methods[13,24,29]. A successful solution involves the analysis of clonal or semi-clonal tissue structures, for example, colon crypts, which can be micro-dissected and individually sequenced [27]. A main disadvantage of this approach is that, to obtain a profile of mutations per individual, hundreds of micro-dissected biopsies need to be evaluated. This can be very costly and time-consuming. An alternative approach involves the analysis of somatic mutations using ultradeep duplex sequencing[28]. Duplex sequencing employes double stranded molecular barcodes to enable independent consensus making in both strands of DNA. Mutations are only called if they are seen in a majority of sequencing reads in both strands of DNA. This protocol reduces the sequencing error rate to less than one in 10 million per nucleotide sequenced, enabling the observation of mutations occurring in normal tissues at very low variant allele fractions.

The Risques lab pioneered the use of duplex sequencing to study somatic mutations in normal tissue. In a pilot study, Matas *et al*. found somatic mutations present in the normal colon of individuals with and without colorectal cancer, with mutations occurring more frequently in

those with cancer [30]. Matas *et al.* also observed that CRC patients had multiple large clones (variants with >1 duplex read) with driver mutations distinct from those in their tumors, illustrating that multiple driver mutations can exist, but only one may result in cancer development [30]. In this study we intend to build upon these initial findings by using duplex sequencing to evaluate a larger panel of colorectal cancer driver genes, including *APC, KRAS, PIK3CA, TP53, FBXW7, BRAF*, and *SMAD4.* We analyzed mutations in the normal colon of three groups of individuals: patients without CRC or polyps, patients with polyps, and patients with CRC. To achieve very high resolution for mutation detection we increase the depth of duplex sequencing to 8000x. In addition to the targeted panel of CRC driver genes, we also assess a mutagenesis panel (TwinStrand Biosciences) composed of neutral, intergenic regions throughout the genome. This panel is included to enable the identification of mutational signature(s) by analyzing neutral mutations. Mutations were analyzed with advanced computational tools to assess both the potential pathogenicity of coding mutations and to extract mutational signatures from non-coding mutations[31, 32,33,34]. The main aim of this study is to characterize somatic mutations and their clonal expansions within the normal colon of individuals with and without colorectal cancer or polyps. By examining the mutational spectrum and burden in the normal colon, in the form of clonal expansions of mutant driver genes and mutational signatures, we aim to determine how these factors may predict an individuals' risk for developing cancer.

Results

Normal colon collection and ultradeep sequencing in patients without CRC or polyps, with polyps and with CRC

This retrospective study included 60 normal colon epithelium samples, collected from 40 patients without CRC – 20 of which had polyps – undergoing colonoscopy screening and 20 patients newly diagnosed with primary invasive colorectal adenocarcinoma undergoing surgical resection. Patients without CRC were enrolled in the GICaRes BioSample Repository at the University of Washington and patients with CRC were enrolled in the ColoCare study (Methods). Three patient groups were defined: individuals without CRC or polyps, individuals with polyps only, and individuals with CRC (Fig. 1A). Patients were selected based on normal colon tissue availability and age with the goal of having similar age ranges across 3 groups to

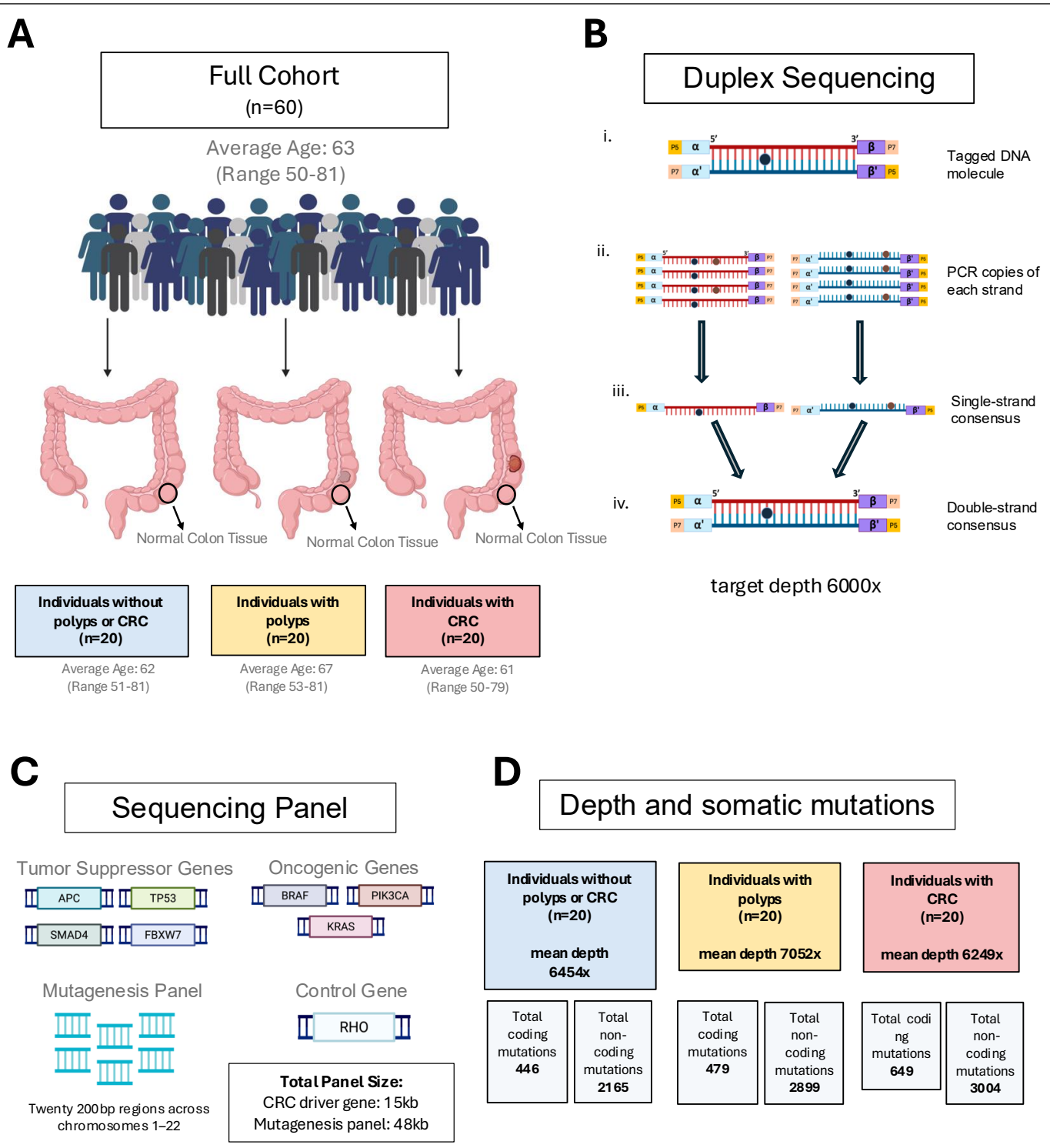

Figure 1. Study Design. **A.** Cohort description. Normal colon biopsies were collected from individuals without CRC, with polyps, and with CRC. **B.** Duplex sequencing. (i). DNA molecules are ligated to adapters with double-stranded molecular barcodes. (ii). PCR copies of each strand are generated during library prep. (iii). Single strand consensus is generated from reads sharing the same barcode. (iv). Double strand consensus is produced from comparing complementary strands containing the same barcode. **C.** Sequencing gene panel. **D.** Depth results and mutation counts by patient group.

We performed duplex sequencing of *APC, KRAS, PIK3CA, TP53, FBXW7, BRAF, SMAD4, RHO* (control gene) as well as a mutagenesis panel (non-coding regions) at a target depth of 6000x (Fig. 1B and 1C, Tables S2 and S3) in normal epithelium from the left colon of patients with and without polyps and CRC. In patients with CRC, the normal colon biopsy was collected distant from tumors (>10cm). We sequenced an average of 130 million coding nucleotides per sample resulting in an average duplex depth of 6585x (range 4909x-9962x). Duplex depth between groups (Fig. 1D) and between samples and genes (Fig. 2) were comparable. We identified more than 400 unique coding mutations and over 2,000 unique non-coding mutations in each patient group (Fig 1D). To account for variability in sequencing depth, mutation frequencies (MF) were calculated for coding mutations by dividing the total number of coding mutations for each sample by the total number of duplex nucleotides sequenced in the coding region.

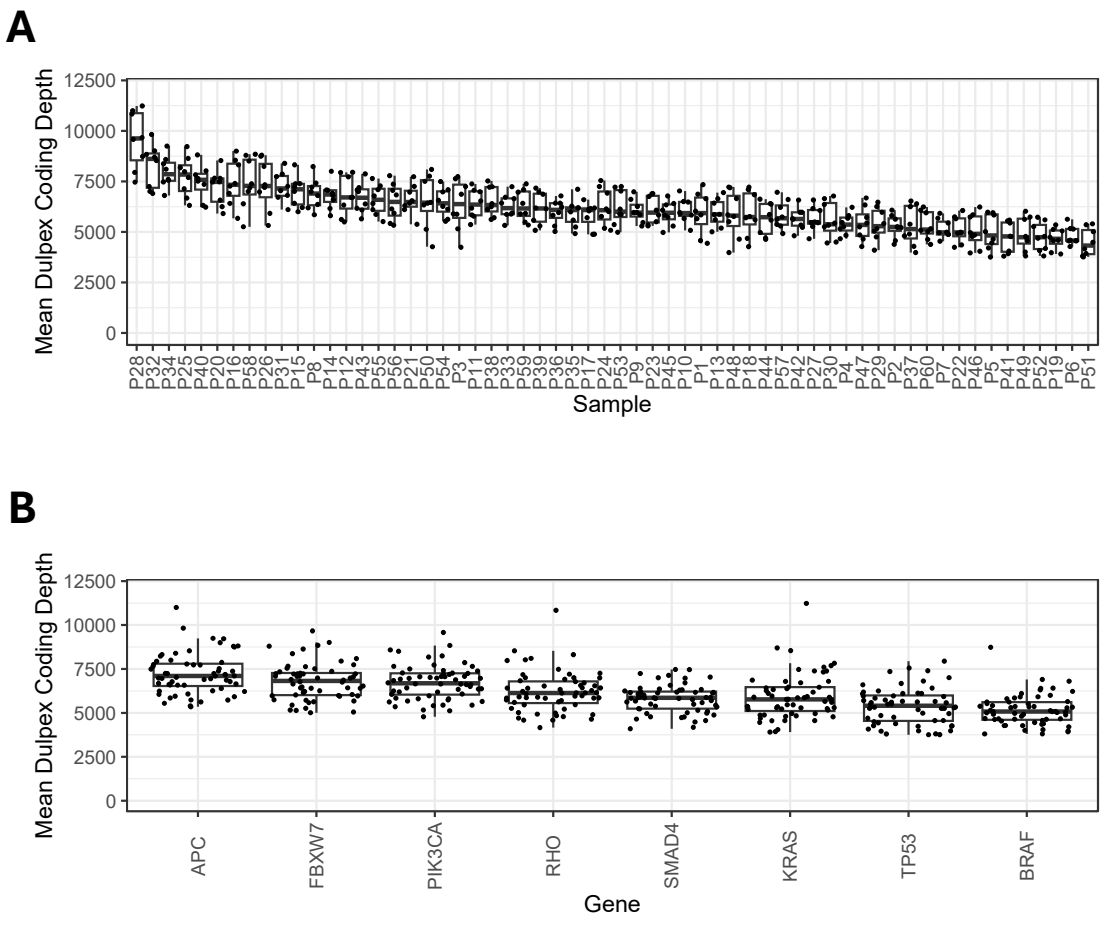

Figure 2. Homogeneous ultra-deep sequencing across normal colon samples and genes A. mean depth per normal colon sample, ordered by median depth across genes. **B.** Mean depth per gene, ordered by median depth across patient samples.

To understand the impact of cancer status on coding mutation frequency within the normal colon we plotted coding mutation frequencies by group (Fig. 3A). We observed that the cancer patient group had a significant increase in coding mutation frequency compared to both non-cancer groups ($p=0.007$, $p=0.006$ by Wilcoxon tests for no cancer or polyp subjects and polyp subjects, respectively). There was no significant difference in the coding mutation frequency between the no cancer/polyp group compared to the polyp group. Given that somatic mutations inevitably accumulate throughout a person's lifetime, we evaluated the association of MF with age using linear regression. The cancer group had a significant increase in coding mutations with age ($p=0.003$ by Spearman's rank correlation test) whereas the non-cancer groups did not display an association between coding mutation frequency and age, regardless of polyp status (Fig. 3B). To understand how other clinico-pathological factors could be contributing to mutation frequency, we performed multivariable linear regressions for coding MF and clinical variables including age, sex, BMI, and smoking status (Fig. 3C). As expected, age had a significant ($p<0.05$) effect on coding mutation frequency in the normal colon. However, the effect of CRC was larger and more significant ($p<0.0001$), independent of age and other clinical variables. None of the other variables tested (polyp status, sex, BMI, and smoking status) had a significant independent effect on MF (Fig 3C).

To test if all CRC driver genes had more mutations in the cancer group compared to the no cancer groups, we evaluated the frequency of coding mutations by gene for all patient groups (Fig. 4). Notably, only some CRC driver genes showed increased MF in the normal colon of patients with cancer compared to those without cancer. Specifically*, APC* showed significant differences between patients with polyps and cancer whereas *FBXW7* and *PIK3CA* showed significant differences between the group without cancer or polyps and the cancer group ($p=0.0016$, $p=0.029$ by Wilcoxon tests). To confirm these results were not being confounded by clinico-pathological characteristics, we performed multivariable linear regressions for coding MF and clinical variables including age, sex, BMI, and smoking status. Having cancer was associated with increased MF in the normal colon for *APC, FBXW7*, and *PIK3CA* independent

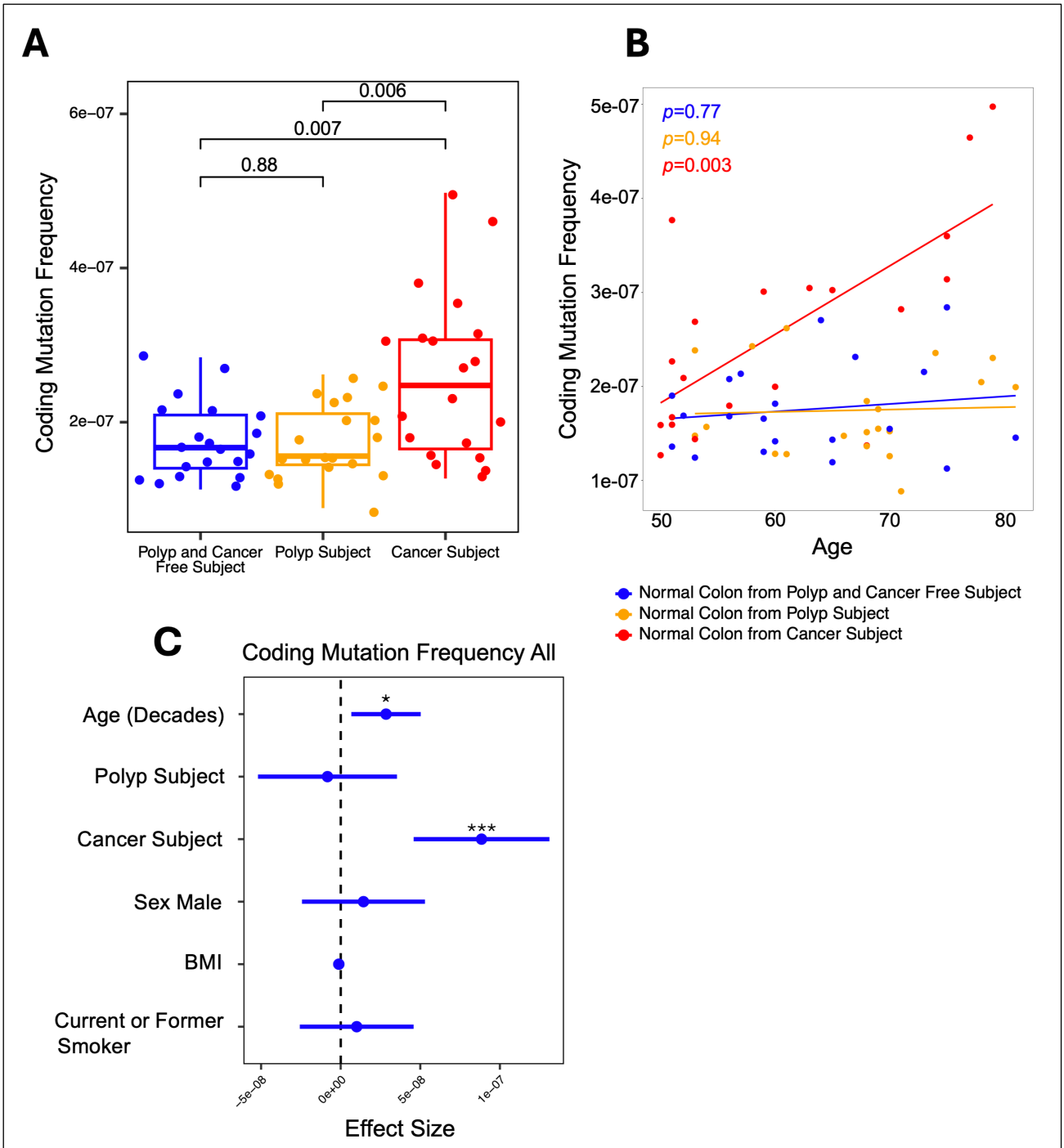

Figure 3. Normal colon of patients with CRC has higher levels of coding mutations. A. Comparison of coding MF across study groups. Box plots display the quartiles with whiskers extending to the highest and lowest data points within 1.5 times the interquartile range. p-values for Wilcoxon tests are indicated. **B.** Association of coding MF with age for the 3 study groups. p-values correspond to Spearman correlations. **C.** Multivariable linear regression model for coding MF including clinical variables. Models are adjusted by mean duplex depth and DNA integrity number. Effect sizes are indicated with solid circles with horizontal lines representing the 95% confidence intervals. *p<0.05 ***p<0.0001.

of age and other variables (Fig. 4B). *TP53, KRAS BRAF* and *SMAD4* did not have significant differences in coding MF between patient groups.

To visualize the number of coding mutations seen in each gene by group we mapped out all detected mutations per patient by gene (Fig. 5). In this display we see numerous coding mutations across the gene panel for patients without cancer or polyps, patients with polyps, and patients with cancer. The tumor suppressor genes *APC, FBXW7, and TP53* and the oncogene *PIK3CA* had more mutations in the normal colon on average. Interestingly, we observed the highest abundance of mutations in *APC* compared to any other gene, for all three patient groups. Normal colon biopsies had on average 9 mutations in *APC* (Table S4). The oncogenic genes *BRAF, KRAS*, and tumor suppressor *SMAD4* had fewer mutations on average (1-3 per patient), with several patients, particularly those without cancer or polyps, with no detectable mutations in these genes (Table S4). This may contribute to the lack of a significant difference in coding mutation frequency of these genes in the normal tissue of individuals with and without cancer.

The positive selection of clones with mutations in CRC driver genes is distinct in individuals with CRC compared to individuals without cancer

To characterize these mutations, we assigned classifications of 'likely pathogenic' or 'likely benign' based on the AlphaMissense model for predicting pathogenicity (see Methods). All coding mutations observed in each patient were color coded based on pathogenicity and clone size (Fig. 5). Clones were classified as 'large clones' when a mutation occurs in >1 duplex read at the same position. We observed that large pathogenic clones appeared more abundant in the normal colon of individuals with CRC compared to those without cancer/polyps and those with polyps (Fig. 5). Large clones are indicative of positive selection, indicating the phenotypic characteristic resulting from the mutations in these clones prompted clonal expansion. 125 large clones were detected in the patient group without cancer or polyps, compared to the 218 large clones that were detected in the cancer group. This finding demonstrates these mutations are contributing to the phenotypic advantage of cells, resulting in their subsequent selection and expansion.

To further assess positive selection, we analyzed clone size in conjunction with pathogenicity across the coding gene panel. Notably only certain genes showed enrichment of large pathogenic

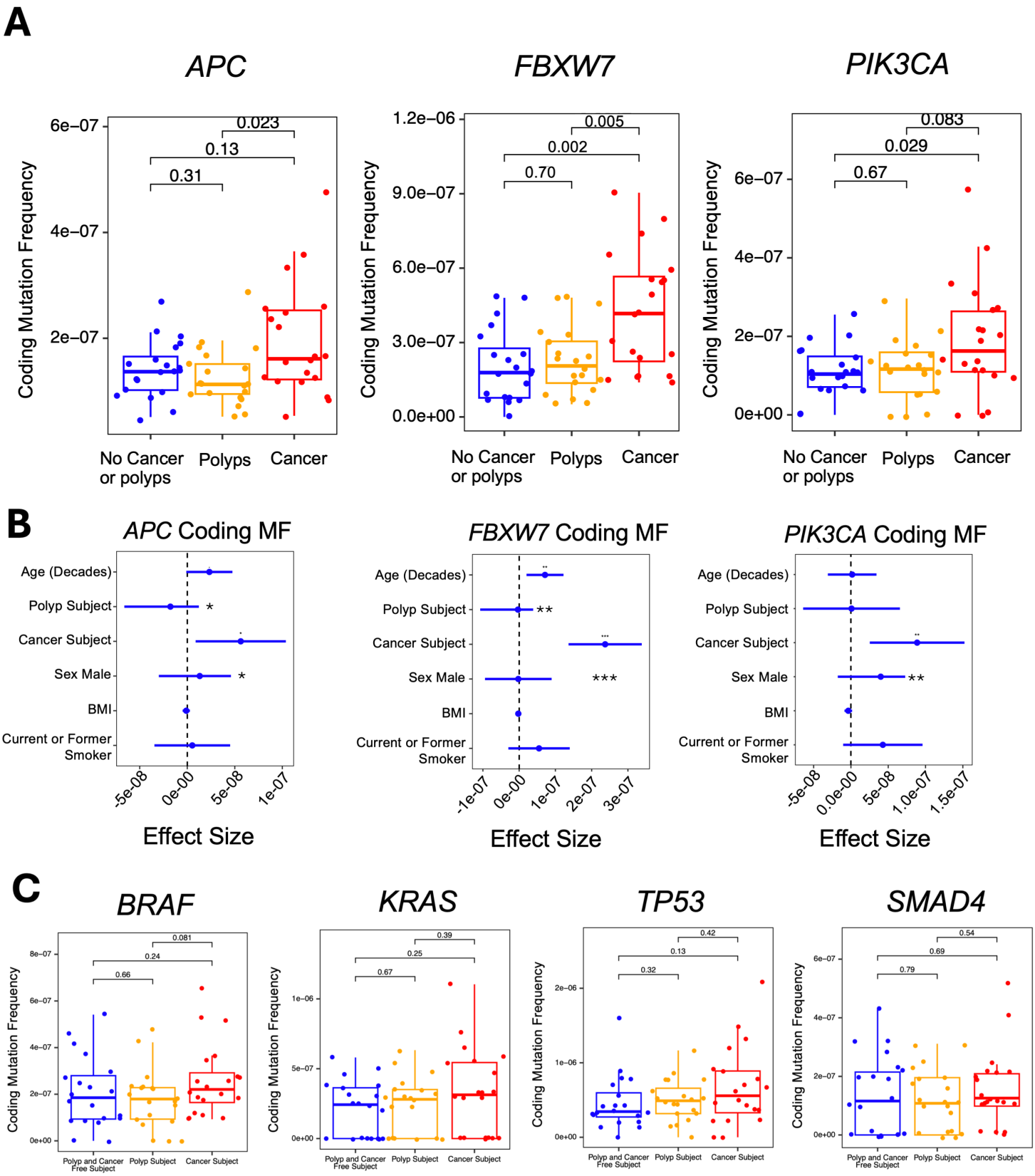

Figure 4. Normal colon of patients with CRC has higher levels of mutations in *APC, FBXW7*, and *PIK3CA* A. Coding mutation frequency of *APC, FBXW7*, and *PIK3CA* by patient group. Box plots display the quartiles with whiskers extending to the highest and lowest data points within 1.5 times the interquartile range. p-values for Wilcoxon tests are indicated. **B.** Multivariable linear regression model for coding MF including clinical variables. Models are adjusted by mean duplex depth and DNA integrity number. Effect sizes are indicated with solid circles with horizontal lines representing the 95% confidence intervals. **p<0.01, ***p<0.0001. **C.** Coding mutation frequency of *BRAF*, *KRAS, TP53,* and *SMAD4* by patient group. *B*ox plots display the quartiles with whiskers extending to the highest and lowest data points within 1.5 times the interquartile range. p-values for Wilcoxon tests are indicated.

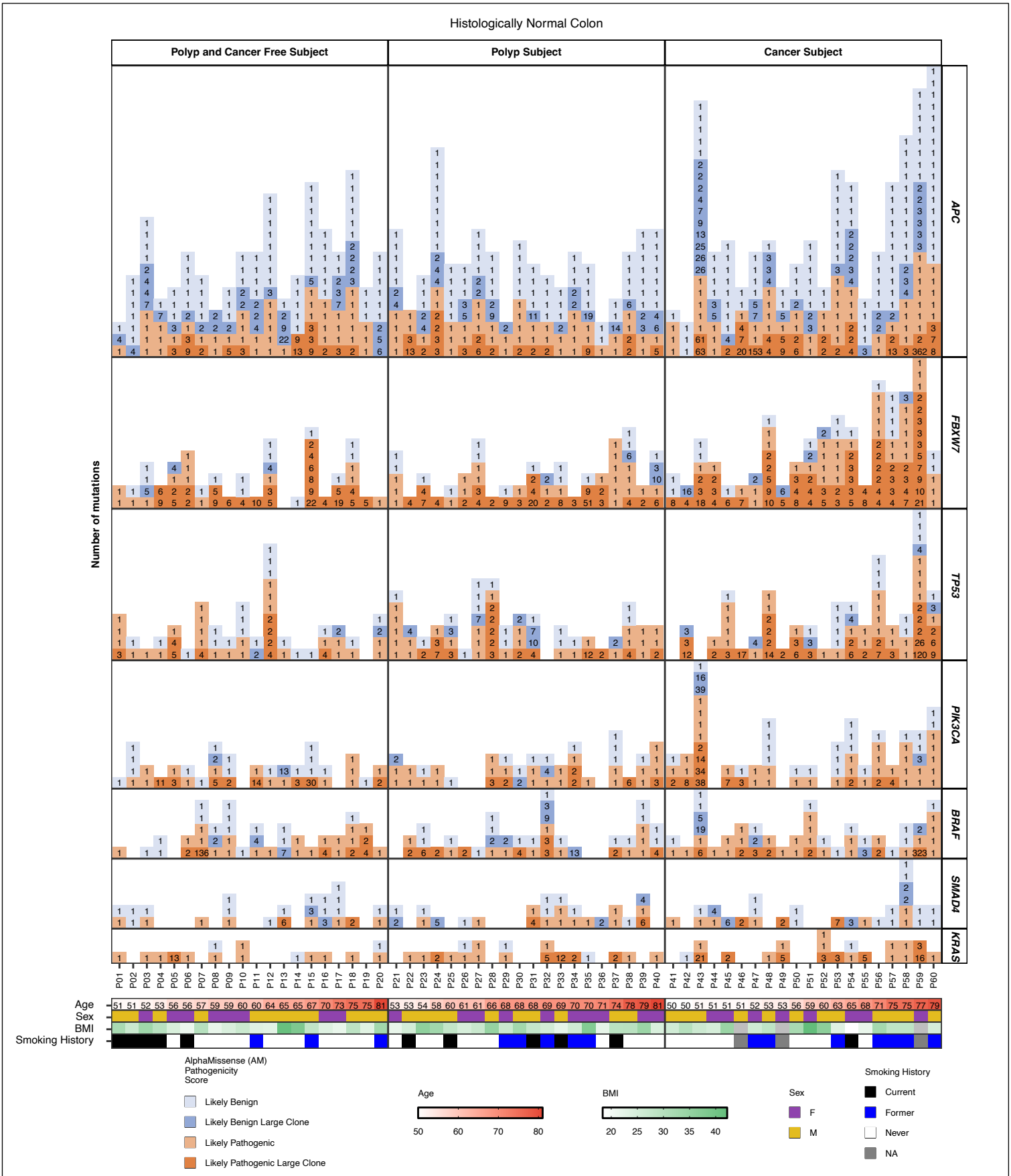

Figure 5. Coding mutations in normal colon tissue are found in all patients, including large pathogenic clones, with a higher incidence in CRC patients. Catalog of mutations detected in normal colon samples. Each square corresponds to a single mutation. The number in the square indicates the number of duplex reads carrying the mutation. Mutations are color-coded by pathogenicity and the number of mutant duplex reads >1, which is indicative of larger clones.

clones in patients with CRC, while others harbored more mutations but not enriched in pathogenicity. For example, *APC* had numerous large clones that were both pathogenic and non-pathogenic, with the frequency of large clones increasing only moderately between the no cancer/polyps group and the cancer group (Fig. 5 and 6). Most of the mutations observed in *APC* were small benign clones, meaning that, while mutations in *APC* in normal colon tissue are abundant, many mutations are not positively selected. Conversely, genes such as *KRAS* and *TP53* display a clear increase in the positive selection of pathogenic clones in individuals with cancer compared to those without cancer. Specifically, in *KRAS* we observed only 1 large pathogenic clone out of 15 total mutations detected across the patient group without cancer or polyps, compared to the 8 large pathogenic clones out of 21 total mutations detected in the cancer group ($p=0.05$, Fig. 6). In *TP53* we observed 9 large pathogenic clones in the group without cancer or polyps, 15 in those with polyps, and 27 large clones in individuals with cancer. While the increase between individuals without cancer or polyps and individuals with polyps is not significant, the increase in patients with cancer was significant ($p=0.01$ by Chi-squared test) (Fig.6). These findings indicate more positive selection of clones carrying pathogenic mutations in *KRAS* and *TP53* in the normal colon of individuals with CRC than those without cancer, closely replicating the results from our prior study [30].

Somatic mutations in CRC driver genes of the normal colon mimic known gene hotspots observed in cancer.

After characterizing the somatic landscape of the normal colon in individuals without cancer or polyps, with polyps, and with cancer, we sought to compare how the somatic landscape of the normal colon compares to the mutagenesis landscape reported in CRC. We first evaluated the proportion of mutation types seen in the three patient groups compared to mutations in a "No Selection Model". The No Selection Model was built based on the distribution of all possible mutations across all the genes sequenced, corrected by the mutational signature profile in the normal colon[49]. We also compared to the mutations registered in the COSMIC database for CRC (Fig. 7A). A larger proportion of the mutations observed in the normal colon of cancer subjects were nonsense mutations, similar to what is observed CRC (Fig. 7A). The spectrum of mutations was enriched in C>T in all groups, which is concordant with no selection but also with cancer mutations (Fig. 7B). Regarding pathogenicity, we found a significant increase of pathogenic mutations in the 3 groups over the no selection model and also a significant increase in the

cancer group compared to the no cancer group (Fig. 7C). Since AlphaMissense provides a quantitative value of pathogenicity (pathogenicity scores), we used these values to provide a more detailed comparison of the pathogenicity of the mutations in each group. To do so, we ranked the mutations for each gene in each group based on pathogenicity

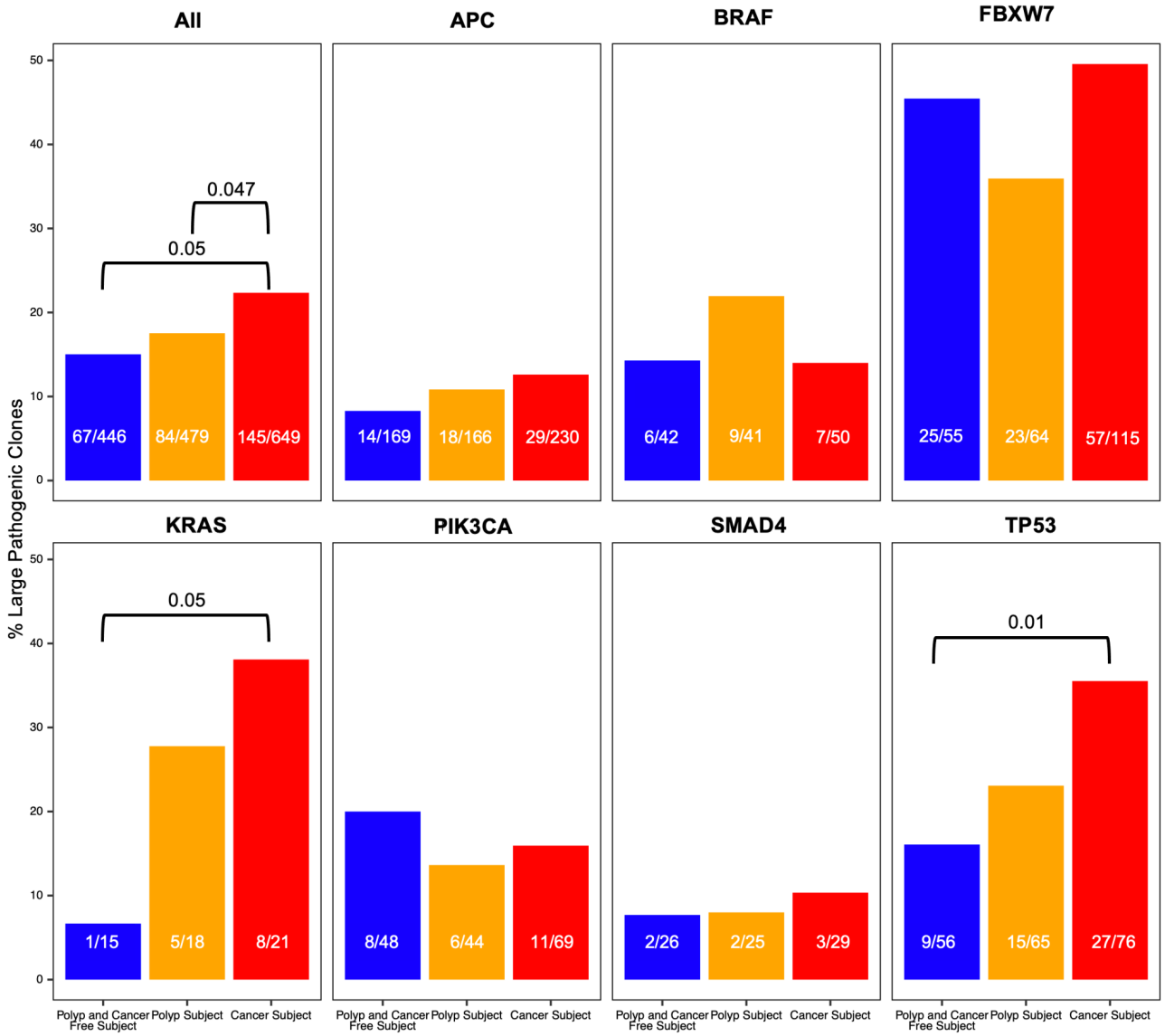

Figure 6. Pathogenic large clones were more frequent in the normal colon of individuals with cancer than those without caner or polyps. Proportion of mutations identified in normal colon samples that were large pathogenic clones. Clones are classified as large clones when a mutation occurs in more than 1 duplex read at the same position. Pathogenicity of mutations based on mutation type and AlphaMissense scores. p-values correspond to Chi-squared test for all groups with the exception of *KRAS* where p-values correspond to Fisher's exact tests.

and normalized to a 0-1 scale (Fig. 7D). The same analysis was performed for the no selection model and mutations identified in CRC from COSMIC. These plots confirmed that higher levels

of selection of pathogenic mutations occur only in certain genes (Fig. 7D). Specifically, *FBXW7, TP53*, and *KRAS* showed a distinct shift to the left (more pathogenic) of mutations detected in patients with CRC compared to those with polyps or without cancer or polyps. In contrast, mutations detected in *APC* show no selection on rank-order plots by pathogenicity, with the curves of the 3 patient groups overlapping with each other (Fig. 7D).

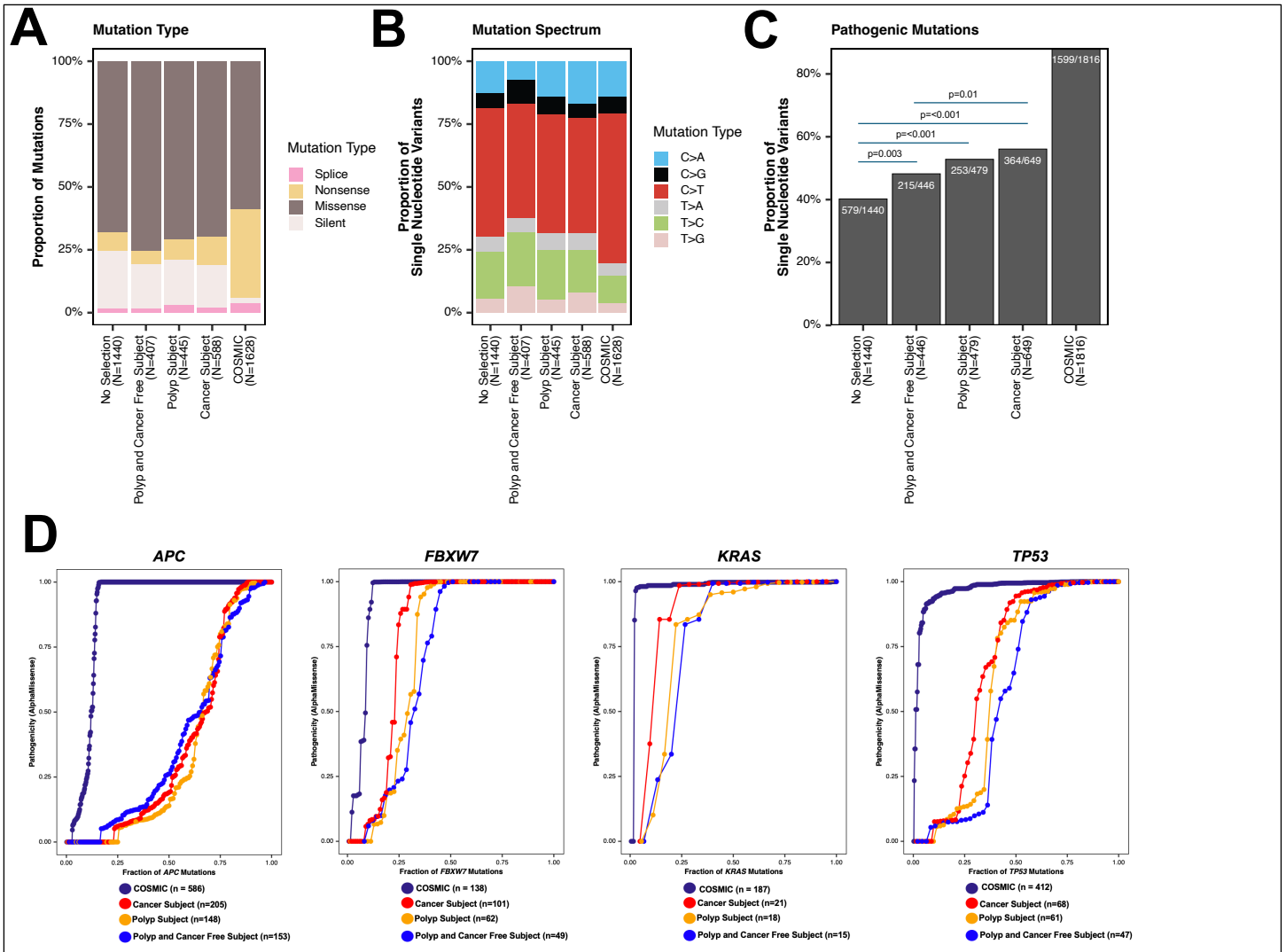

Figure 7. Positive selection of pathogenic mutations is more apparent in normal tissue of cancer patients compared to no cancer patients and only occurs in some CRC driver genes. A. Distribution of mutation type across all patient groups, COSMIC mutations, and no selection model. Number of mutations per group is indicated on the x-axis. COSMIC mutations refer to genome-wide studies across various human cancers for all the genes in the panel. The "no selection" model includes a random distribution of mutations in the coding region across the genes panel. **B**. Mutation spectrum distribution, including only single-nucleotide variants. Groups as in A, **C.** Pathogenicity of mutations based on mutation type and AlphaMissense scores. Groups as in A. p-values correspond to Chi-square tests. **D.** Pathogenicity scores compared by study group in 4 representative genes (*APC, FBXW7, TP53*, and *KRAS*). Pathogenicity scores are indicated in the Y-axis and are based on mutation type and AlphaMissense scores. For each group of individuals and gene, mutations in the normal colon are ranked from the lowest to the highest pathogenicity score and normalized by the total number of mutations in the group (indicated in brackets in the legend). COSMIC mutations include gene variants reported in genome wide screens across human cancers. No selection mutations include all possible substitutions in the coding region of each gene.

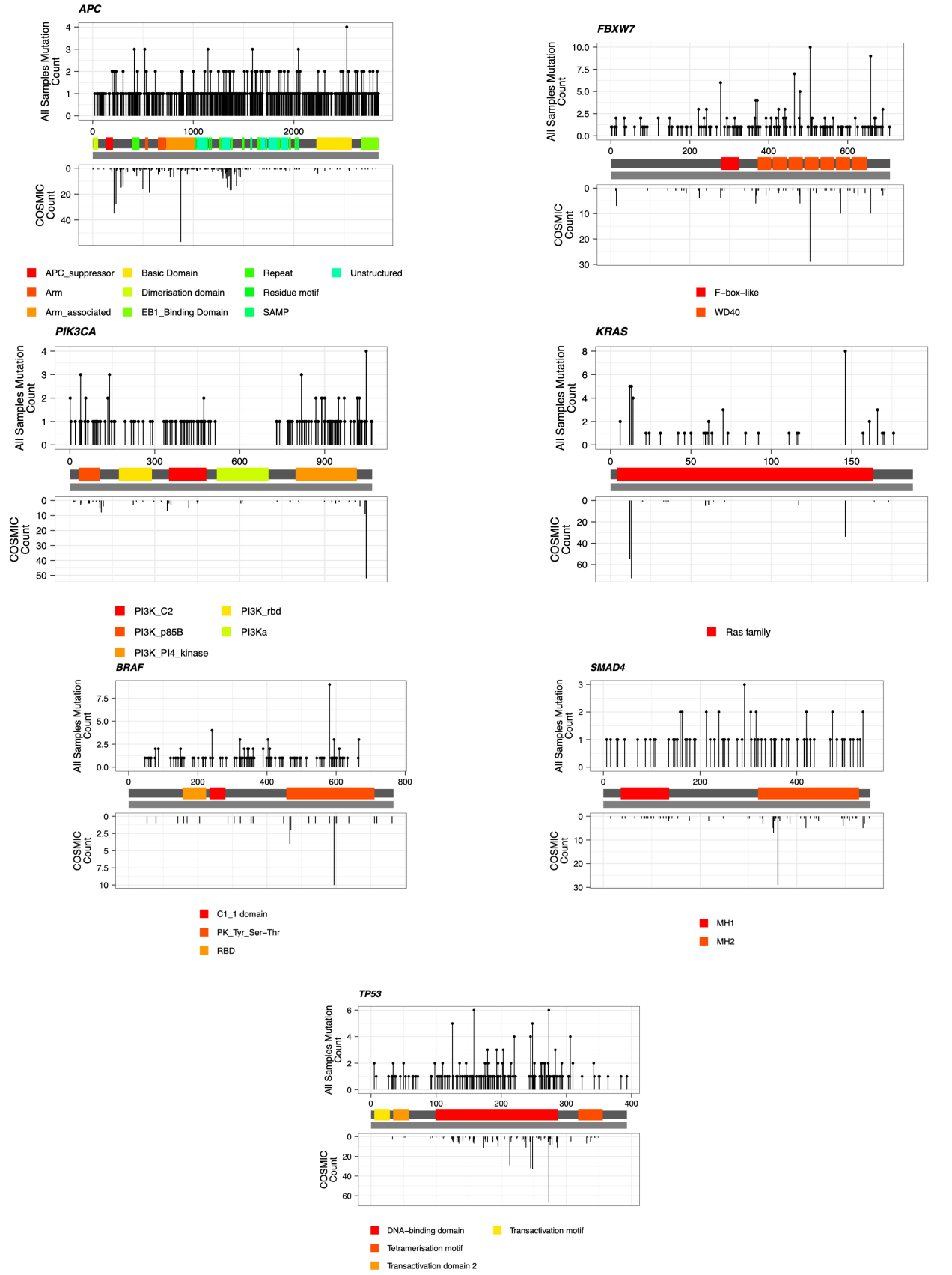

Figure 8. Somatic mutations observed in the normal tissue of individuals with and without cancer cluster in known gene hotspots for CRC. *APC, SMAD4, FBXW7, BRAF, PIK3CA, TP53,* and *KRAS* mutational distribution maps for normal colon samples compared to COSMIC CRC mutations. Upper panels show the location by codon number of gene mutations identified in normal colon samples with the mutational count on the y-axis. Bottom panels show the codon location of gene CRC mutations from the COMSIC database. Gene domains are color-coded and indicated in the legend.

These data suggest that mutations in *FBXW7, TP53*, and *KRAS* observed in the normal tissues of patients with CRC have positive selection similar to mutations seen in cancer. Building on this, we mapped out the mutations we detected across each gene and compared them to mutations registered in the COSMIC database for CRC (Fig. 8). This analysis revealed a high degree of overlap between the mutations found in the normal colon and the mutations found in CRC only for certain genes, specifically, *FBXW7*, *TP53, and KRAS,* the same genes where we observed the highest rates of positive selection of pathogenic clones (Fig. 7, Fig. 8*).* For these genes, the mutations in the normal tissue cluster around known CRC hotspots (Fig. 8). This demonstrates that these mutations are not random but are being selected for, similarly to how mutations in cancer are selected. Moreover, in the genes where positive selection did not occur, such as *APC*, there are numerous mutations detected throughout the gene, not just in locations of known hotspots (Fig. 8).

Detection of non-coding mutations with mutagenesis panel enables mutational signature evaluation

Lastly, we were able to detect >100 non-coding mutations in most samples (including the mutagenesis panel and non-coding regions surrounding the target genes) (Table S4), which enabled the performance of mutational signature analysis. There are 96 possible nucleotide changes when considering all the possible permutations of nucleotides including the one before and after the mutation. Therefore, a minimum of 100 mutations detected are required for complete mutational signature analysis [42,43,44]. Mutational signatures were extracted from non-coding mutations using SigProfilerExtractor [51]. 11 unique signatures were present in the normal colon of our study cohort (Fig. 9). The majority of mutations were attributed to SBS1 and SBS5, both of which are known signatures associated with aging [45]. Notably, we also detected SBS6, SBS10a, and SBS10b in the normal colon tissue of individuals with CRC. SBS6 is associated with mismatch repair deficiency and SBS10a/SBS10b are associated with hypermutators, specifically mutations in DNA polymerase epsilon, a key component of DNA synthesis [44]. Both processes are involved in colorectal carcinogenesis [45,47]. These data demonstrate that somatic mutations in normal colon can be used to evaluate endogenous and exogenous factors that are associated with carcinogenesis.

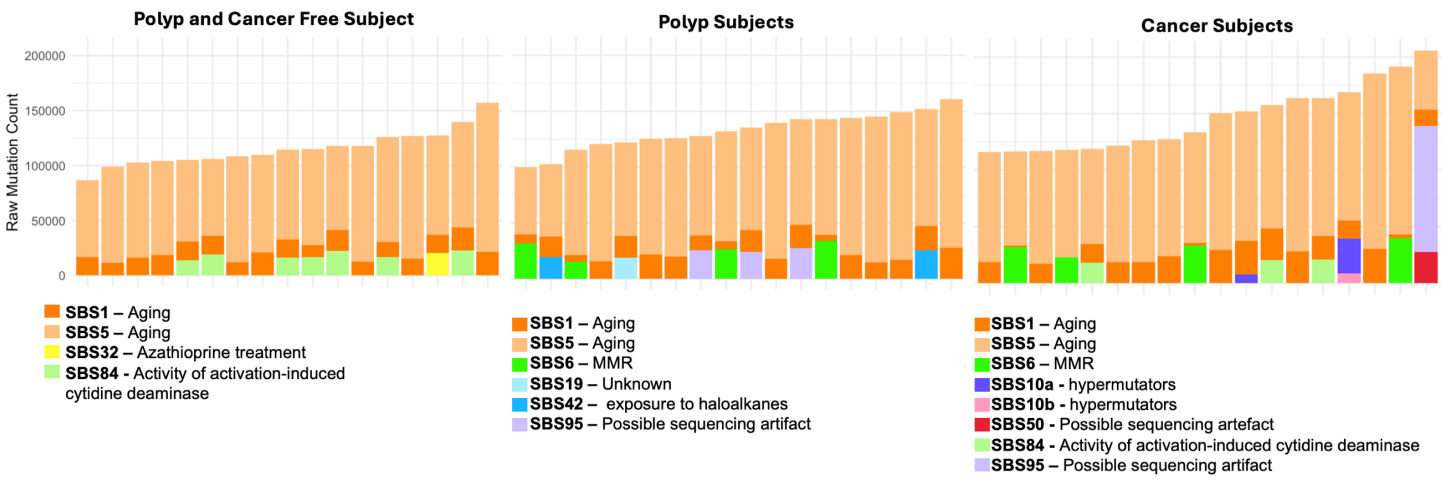

Figure 9. Mutational signatures extracted using SigProfilerExtractor. Non-coding mutations detected in the mutagenesis panel were used to perform mutational signature evaluation. SigProfilerExtractor performs allows de novo extraction of mutational signatures to identify identifying operative signatures, their activities and their probability to cause a specific mutation type.

Discussion

In this study we have demonstrated that patients with CRC have a higher frequency of somatic mutations in their normal colon tissue and significantly more clonal expansions of mutant cells carrying pathogenic mutations in key CRC driver genes. These clonal expansions also mimic the mutagenesis landscape seen in CRC and could potentially predict a patient's risk for developing cancer. We have validated previously published findings on the somatic landscape of the normal colon and have demonstrated that ultra-deep sequencing is a viable method of detection for mutations with low variant allele fractions from a single biopsy[18]. We have also illustrated how new deep-learning models can be applied to evaluate the pathogenicity of detected mutations and to study positive selection in normal tissues.

The main finding of this study is that patients with colorectal cancer have a higher frequency of cancer driver gene mutations in the normal colon than patients without cancer. This is the case for *APC*, which is the most mutated gene in the normal colon and shows a high prevalence of large clones carrying both pathogenic and non-pathogenic mutations, especially in patients with CRC. This is notable as *APC* is often the first gene mutated in CRC[37, 38]. Similarly, we observed significantly more mutations, specifically pathogenic mutations, in *FBXW7* and *PIK3CA* in individuals with cancer. Compared to studies such as Lee-Six et al., which analyzed hundreds of

individual crypts per patient and reported low frequencies of *APC* driver mutations, our approach detected a higher frequency of these mutations from a single biopsy [13]. This suggests that duplex sequencing enables higher-resolution analysis of the somatic landscape in normal colon tissue.

Not all genes studied showed significant differences in mutation frequency across patient groups. *SMAD4* did not present a clear increase in pathogenic mutation frequency in the cancer group compared to the non-cancer groups. This is consistent with the fact that *SMAD4* is a CRC driver gene shown to mutate later in colorectal cancer development[12]. Studies have shown that *SMAD4* mutations cannot initiate carcinogenesis alone but can enhance the progression of tumor development initiated by other genes, such as *APC* [33,34].

After mapping out all the mutations seen in our patient groups, we were able to visualize that the clones carrying pathogenic mutations were more frequently expanded in patients with cancer than those without cancer. Interestingly, not all CRC driver genes displayed clear positive selection of pathogenic mutations, as this was most notably observed in *KRAS* and *TP53.* While these genes did not see an increase in the number of mutations between study groups, there was a clear increase in large pathogenic clones in these genes in patients with CRC. This suggests that this process is not random but specific to these pathogenic mutations, which are being positively selected for and expanded likely due to the phenotypic advantage that they offer.

Numerous mutations detected in the normal colon tissue cluster around known hotspots for CRC in multiple driver genes, including *FBXW7, KRAS,* and *TP53*. The gene with the most distinct increase of pathogenic large clones in the normal tissue of individuals with CRC was *FBXW7*. 6-10% of standard CRC cases report *FBXW7* mutations, however this number increases to 20-30% in early-onset cases[39,38]. *FBXW7* is an essential tumor suppressor that when dysregulated, can result in the accumulation of its substrate cyclin E, leading to aberrant cell proliferation and abnormal chromosome congression during the cell cycle [40]. The increased selection of mutated *FBXW7* clones in patients with CRC observed in conjunction with these mutations corresponding to CRC hotspots indicates this molecular process might play a role in the development of CRC.

In this study we included a mutagenesis panel to detect neutral mutations to build mutational signatures without the bias of selection introduced by cancer gene mutations. While the mutagenesis panel provided at least 100 or more additional mutations per patient, it was still a relatively low number for mutational signature deconvolution. Studies have shown at least 100

mutations are needed for accurate mutational signature analysis [50]. Similar to Lee-six et al. many of the non-coding mutations detected in the normal colon were attributed SBS1 and SBS5, both of which are caused by aging [13]. However, Lee-Six et al. were able to detect 14 known mutational signatures within the normal colon and six novel signatures, whereas here we were only able to detect 11 unique signatures, two of which – SBS50 and SBS95 – could be due to sequencing artifacts [13]. In Lee-Six et al. whole genome sequencing was performed enabling the detection of more non-coding mutations and higher resolution for signature deconvolution. Newer duplex sequencing pipelines currently under development with less stringent consensus criteria might enable us to recover more mutations while preserving sufficient accuracy for more in-depth mutational signature analysis in the future.

There are a few limitations to the methods used in this study. We were able to sequence at an average depth of 6585x. While this allowed for the observation and characterization of numerous somatic mutations, we were limited in our study of mutations of the oncogenes *KRAS, PIK3CA,* and *BRAF*. This is because oncogenic mutations typically occur in a few specific codons and thus, they are less common overall than mutations in tumor suppressor genes, where multiple truncating mutations along the gene can lead to loss of function. Nevertheless, sequencing at a greater depth would allow for a greater detection of mutations in these genes, potentially providing more significance to the observed trend. Another caveat is that for the CRC patient group in this study, we only had tumor sequencing data for a small subset of patients (not reported). Therefore, we could not compare the mutations observed in the patients' normal colon tissue to those seen in their tumors. In addition, tumors could not be evaluated for microsatellite instability and the presence of germline mutations predisposing to CRC was unknown. However, familial CRC has an incidence of only ~15% making it unlikely to significantly contribute to our cohort [50]. The limited number of genes tested is also a caveat of the study. While we observed interesting patterns of mutation and clonal expansion in normal colon for the selected genes, expanding to a larger set of cancer driver genes would enable a clearer picture of the somatic mutational landscape of the normal colon.

In conclusion, this study demonstrates that patients with colorectal cancer not only carry a greater frequency of driver mutations in their normal colon, mostly contributed by *APC, FBXW7* and *PIK3CA*, but also experience increased positive selection of clones harboring pathogenic

mutations in *KRAS* and *TP53*. By employing new computational methods, such as pathogenicity scoring using deep-learning AI models and mutational signature analysis, we have begun to develop a framework for assessing CRC risk based on the somatic landscape of a patient's normal tissue. More studies involving deeper sequencing (>10000x), larger panel of genes, and the inclusion of tumor data from CRC patients will be essential to better understand the role of large pathogenic clones on CRC development and their potential as a biomarker for CRC risk.

Methods

Patient and Sample Selection

This study included 60 individuals ages 50 years or older (mean age 63, range 50-81) who were selected to one of 3 groups: no cancer and no polyps (20 individuals), polyps but not CRC (20 individuals), and CRC (20 individuals) (Fig, 1A). Groups had similar age ranges, but the mean age was slightly higher for the polyp group. No polyp/cancer subjects and polyp subjects underwent standard screening colonoscopies, where they consented to have biopsies collected from their normal colon tissue for the GICaRes BioSample Repository at the University of Washington's Departments of Medicine and Laboratory Medicine & Pathology.

Polyp formation was defined as a diagnosis of adenoma, sessile serrated adenoma, or proximally located hyperplastic polyp made in the procedure where the sample was procured or any prior procedure. Patients with CRC were newly diagnosed with primary invasive colorectal adenocarcinoma and undergoing surgical resections, and they consented to have biopsies collected from their normal colon tissue for the ColoCare research study at the Fred Hutch Cancer Center. All normal samples from individuals with CRC were located 10 to 15 cm from their tumor. We analyzed normal left colon epithelium in individuals without CRC and normal left epithelium distant from tumor (>10cm) in individuals with CRC. Immediately after collection, samples were frozen in liquid nitrogen and stored at -80ºC until DNA extraction. A subset of normal colon samples from cancer free and cancer donors were histologically examined and revealed 70-80% epithelial content in samples from both groups. Patients provided written informed consent for study enrollment and sample collection. Clinico-pathological information available included age, sex, tumor location, polyp location, BMI, smoking history,

and diabetes status (Supplementary Table S1). The study was conducted in accordance with recognized ethical guidelines, which include but are not restricted to U.S. Common Rule, Belmont Report, Declaration of Helsinki, and Nuremberg Code, and following protocols approved by Institutional Review Board committees at the University of Washington and the Fred Hutchinson Cancer Research Center.

DNA Extraction

DNA extractions were performed by the Grady lab at the Fred Hutch Cancer Research Center. Genomic DNA was extracted from frozen normal colon tissue samples using the DNEasy Blood & Tissue Kit (Qiagen, Hilden, Germany). Forceps mucosal biopsies procured at endoscopy were approximately 6mm x 4mm x 3mm in size and the whole biopsy was used for DNA extraction. In normal colon samples from surgical resections, mucosal epithelium was selected to match the size of the endoscopic biopsies. DNA was quantified by Qubit dsDNA BR Assay Kit (ThermoFisher Scientific, Waltham, MA). DNA quality was assessed with Genomic TapeStation (Agilent Technologies, Santa Clara, CA) and demonstrated high quality in all samples (DNA integrity number (DIN) ≥7).

Duplex Sequencing Library Preparation

Duplex Sequencing libraries were prepared using commercially available kits (TwinStrand Biosciences) using 400ng of genomic DNA from normal colon samples. DNA samples were processed by enzymatic fragmentation, end-repair, A-tailing, and ligation of duplex adapters including molecular barcodes. DNA fragments were then amplified by PCR. Following PCR amplification, two hybridization captures were performed with biotinylated probes targeting *APC*, *KRAS*, *PIK3CA*, *TP53*, *FBXW7, BRAF,* SMAD4, and *RHO* (Integrated DNA Technology, Coralville, IA, USA) and a mutagenesis panel (TwinStrand Biosciences, Seattle, WA, USA) (Fig. 1C). Capture reactions were performed with probes at a concentration of 0.75uM per reaction. After each capture, samples underwent magnetic bead washes, and post-wash samples were amplified by PCR. Library concentrations were quantified using Qubit dsDNA HS Assay kits (ThermoFisher Scientific) and library fragment size was determined with Agilent 4200 TapeStation with HS D1000 tapes.

Libraries were pooled for sequencing at 150 million reads per colon sample. All samples were sequenced at the Fred Hutch Cancer Research Center Genomics Services on an Illumina NovaSeq X Plus using 2 x 150bp paired-end reads.

Duplex Sequencing Analysis

Sequencing data was processed with the duplex sequencing pipeline v2.1.4 (https://github.com/Kennedy-Lab-UW/Duplex-Seq-Pipeline). This pipeline can be divided into five key steps: consensus making, alignment, blast, post-blast clean-up, and variant calling.

Raw sequencing reads underwent barcode recognition to group PCR copies from the same source DNA molecule into groups, termed 'families'. PCR copy families underwent single-strand consensus, where unique mutations were only called when they occur in the same position in 70% or greater of the PCR reads that share the same barcode. Complementary DNA strands were then compared to produce a double-strand consensus read (called duplex read) where nucleotides from the forward and reverse strands were compared and unique mutations were only called when they occur in the same position on both strands.

Before alignment, adaptors were clipped using Cutadapt. Duplex reads were then aligned to the human genome reference hg38 using BWA. A blast filter was applied to eliminate potential reads of non-human genomes. In post-blast clean-up, the AS-XS score was set at 50 to eliminate reads that align to multiple places in the genome. End reads were trimmed 15bp from the 5' end and 5bp from the 3' end. Additionally, we filtered out any mutations seen within 10bp from either end of each read, after clipping, to eliminate artifactual mutations related to the NovaSeq X Plus chemistry.

Variant calling was done using VarDict Java and VCF file outputs were then converted to MAF files using Vcf2Maf script with Variant Effect Predictor (VEP) version 104. MAF files underwent post-processing with R version 4.4.1. Only positions with a minimum depth of 1000x and N values <1% were included in the analysis. Mean coding depth was calculated considering the depth at coding positions for each sample. Variant allele fraction (VAF) was calculated for each mutation by dividing the number of mutant reads by the total number of duplex reads at the given position. Mutations were then filtered based on VAF (>0.3) to eliminate Single Nucleotide Polymorphisms (SNPs). SNPs VAF values were used for quality check to confirm no cross-contamination.

Mutation Analysis

Using the information from MAF files, mutations were characterized by type (missense, nonsense, splice, insertion, deletion, synonymous, intronic), and mutation spectrum (C>A, C>G, C>T, T>A, T>C, and T>G). For each sample, Mutation Frequency (MF) was calculated for coding and non-coding regions as the number of unique mutations within a region divided by the total number of duplex nucleotides sequences in the corresponding region.

Pathogenicity scores for missense mutations were calculated using AlphaMissense, a deep-learning AI model that classified all 71M possible missense variants in the human proteome [31]. This model uses three main components: evaluation of mutation effect on protein structure outcome, multiplexed assays of variant effect, and training from population frequency data, to assign a pathogenicity score between 0 and 1 [31]. Missense variants are classified as likely benign (score 0–0.340), of ambiguous pathogenicity (score 0.340–0.564), or likely pathogenic (score ≥0.564) [31]. For simplification, all missense variants classified as ambiguous by AlphaMissense were grouped as benign. For other mutation types, pathogenicity classifications were assigned during post-sequencing processing. Silent mutations and splice region mutations outside of exons received a score of 0 and were classified as benign. Nonsense mutations, indels, splice site mutations, double nucleotide mutations, trinucleotide mutations, and splice region mutations within exons were assigned a score of 1 and classified as pathogenic. Pathogenic mutation frequency was calculated by dividing the number of pathogenic mutant reads by the total number of duplex nucleotides sequenced in the corresponding region.

Comparison with COSMIC data

Catalogue of Somatic Mutations in Cancer (COSMIC) data was used to determine cancer mutations for each of the genes analyzed. We retrieved all mutations corresponding to genome wide screen analyses of human carcinoma (January 18, 2023). We next re-annotated COSMIC mutations using VEP, Vcf2Maf, and AlphaMissense, as described for the study samples. Mutations (n=11174) were classified by type (missense, nonsense, silent, indels, and splice). Substitutions (n=9296) were classified by spectrum (nucleotide change).

“No Selection” Model

To determine the distribution of mutations in the absence of selection, a list of all possible single nucleotide substitutions in the coding region and splice site boundaries of each gene was

generated in silico as a VCF file, which was then converted to a MAF file and annotated using VEP, Vcf2Maf, and AlphaMissense, as described for the study samples. Only mutations present in coding regions or splice sites and not within masked regions were included in analysis for valid comparisons with study samples. We then created the No Selection model by sampling from this file a number of mutations equal to the total number of mutations present in all samples (1440) with probabilities of each mutation type (context plus nucleotide change) being set by observed signatures in normal colon crypts [49]. This process was repeated 1000 times, and mean values from those 1000 iterations were used as the No Selection model. Mutations were classified by type (missense, nonsense, silent, and splice), spectrum (nucleotide change), and AlphaMissense pathogenicity.

Mutational signatures

Mutational signatures were analyzed using SigProfilerExtractor [51]. This model uses an unsupervised machine-learning model for de novo extraction of signatures from non-coding somatic mutations [51].

Statistical Analysis

Correlations between MF and pathogenic MF and age were performed with Spearman's rank test. Comparisons of MF and pathogenic MF across groups and by genes were performed by Wilcoxon rank-sum tests. Comparisons of categorical variables across groups, such as the fraction of mutations that were pathogenic large clones, were tested with Chi-squared tests and Fisher's exact tests. Multivariable linear regression models were used to examine the associations between age, sex, BMI, smoking status, and MF outcomes. Models were adjusted for DIN and duplex depth. All tests were two-sided at α level (type 1 error rate) of 0.05.

Supplementary tables:

Supplementary Tables. xlsx

References:

1. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2024; 74(3): 229-263. doi:10.3322/caac.21834
2. American Cancer Society. (2025, January 16). *Colorectal cancer statistics: How common is colorectal cancer?* American Cancer Society. https://www.cancer.org/cancer/types/colon-rectal-cancer/about/key-statistics.html
3. Siegel RL, Wagle NS, Cercek A, Smith RA, Jemal A. Colorectal cancer statistics, 2023. *CA Cancer J Clin*. 2023; 73(3): 233-254. doi:10.3322/caac.21772
4. Colon polyps. National Institute of Diabetes and Digestive and Kidney Diseases. https://www.niddk.nih.gov/health-information/digestive-diseases/colon-polyps. July 2017.
5. Amersi F, Agustin M, Ko CY. Colorectal cancer: epidemiology, risk factors, and health services. Clin Colon Rectal Surg. 2005 Aug;18(3):133-40. doi: 10.1055/s-2005-916274. PMID: 20011296; PMCID: PMC2780097.
6. Conteduca, V., Sansonno, D., Russi, S. & Dammacco, F. Precancerous colorectal lesions. Int. J. Oncol. 43, 973–984 (2013).
7. Nguyen LH, Goel A, Chung DC. Pathways of Colorectal Carcinogenesis. Gastroenterology. 2020 Jan;158(2):291-302. doi: 10.1053/j.gastro.2019.08.059. Epub 2019 Oct 14. PMID: 31622622; PMCID: PMC6981255.
8. Turner KO, Genta RM, Sonnenberg A. Lesions of all types exist in colon polyps of all sizes. *Am J Gastroenterol*. 2018 Feb;113(2):303–306. doi: 10.1038/ajg.2017.439
9. Hossain, M. S. et al. Colorectal cancer: a review of carcinogenesis, global epidemiology, current challenges, risk factors, preventive and treatment strategies. *Cancers* 14, 1732 (2022)
10. Li, Q., Geng, S., Luo, H. et al. Signaling pathways involved in colorectal cancer: pathogenesis and targeted therapy. Sig Transduct Target Ther 9, 266 (2024). https://doi.org/10.1038/s41392-024-01953-7
11. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* 144, 646–674 (2011).
12. Gerstung, M., Jolly, C., Leshchiner, I. *et al.* The evolutionary history of 2,658 cancers. *Nature* 578, 122–128 (2020).
13. Lee-Six, H., Olafsson, S., Ellis, P. *et al.* The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* 574, 532–537 (2019). https://doi.org/10.1038/s41586-019-1672-7
14. Stratton MR. Journeys into the genome of cancer cells. EMBO Mol Med. 2013 Feb;5(2):169-72. doi: 10.1002/emmm.201202388.
15. Peter C. Nowell, The Clonal Evolution of Tumor Cell Populations.Science194,23-28(1976). DOI:10.1126/science.959840
16. Cairns, J. Mutation selection and the natural history of cancer. *Nature* 255, 197–200 (1975). https://doi.org/10.1038/255197a0
17. Martincorena, I. Somatic mutation and clonal expansions in human tissues. *Genome Med* **11**, 35 (2019). https://doi.org/10.1186/s13073-019-0648-4
18. Turner KO, Genta RM, Sonnenberg A. Lesions of all types exist in colon polyps of all sizes. *Am J Gastroenterol*. 2018 Feb;113(2):303–306. doi: 10.1038/ajg.2017.439
19. J. Vijg, Somatic mutations, genome mosaicism, cancer and aging. Curr. Opin. Genet. Dev. 26, 141–149 (2014).
20. M. O'Huallachain, K.J. Karczewski,S.M. Weissman, A.E. Urban,& M.P. Snyder, Extensive genetic variation in somatic human tissues, Proc. Natl. Acad. Sci. U.S.A. 109 (44) 18018-18023, https://doi.org/10.1073/pnas.1213736109 (2012).
21. Kennedy SR, Zhang Y, Risques RA. Cancer-Associated Mutations but No Cancer: Insights into the Early Steps of Carcinogenesis and Implications for Early Cancer Detection. Trends Cancer. 2019 Sep;5(9):531-540. doi: 10.1016/j.trecan.2019.07.007.
22. Merlo, L., Pepper, J., Reid, B. *et al.* Cancer as an evolutionary and ecological process. *Nat Rev Cancer* **6**, 924–935 (2006). https://doi.org/10.1038/nrc2013

23. Fowler, J. C., & Jones, P. H. (2022). Somatic Mutation: What Shapes the Mutational Landscape of Normal Epithelia?. Cancer discovery, 12(7), 1642–1655. https://doi.org/10.1158/2159-8290.CD-22-0145
24. Iñigo Martincorena *et al.* High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* 348,880-886(2015).DOI:10.1126/science.aaa6806
25. Muiños, F., Martínez-Jiménez, F., Pich, O. *et al.* In silico saturation mutagenesis of cancer genes. *Nature* 596, 428–432 (2021). https://doi.org/10.1038/s41586-021-03771-1
26. Wood, L. D. et al. The genomic landscapes of human breast and colorectal cancers. Science 318, 1108–1113 (2007).
27. Martincorena I, Campbell PJ. Somatic mutation in cancer and normal cells. Science. 2015 Sep 25;349(6255):1483-9.
28. Kennedy, S.R., et al., *Detecting ultralow-frequency mutations by Duplex Sequencing.* Nat Protoc, 2014. **9**(11): p. 2586-606.
29. Cagan, A., Baez-Ortega, A., Brzozowska, N. *et al.* Somatic mutation rates scale with lifespan across mammals. *Nature* 604, 517–524 (2022). https://doi.org/10.1038/s41586-022-04618-z
30. Matas J, Kohrn B, Fredrickson J, Carter K, Yu M, Wang T, Gui X, Soussi T, Moreno V, Grady WM, Peinado MA, Risques RA. Colorectal Cancer Is Associated with the Presence of Cancer Driver Mutations in Normal Colon. Cancer Res. 2022 Apr 15;82(8):1492-1502. doi: 10.1158/0008-5472.CAN-21-3607.
31. Jun Cheng *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* 381, eadg7492 (2023).DOI:10.1126/science.adg7492
32. Islam SMA, *et al*. Uncovering novel mutational signatures by *de novo* extraction with SigProfilerExtractor. Cell Genom. 2022 Nov 9;2(11):None. doi: 10.1016/j.xgen.2022.100179.
33. Fang, Tian et al. "Prognostic role and clinicopathological features of SMAD4 gene mutation in colorectal cancer: a systematic review and meta-analysis." BMC gastroenterology vol. 21,1 297. 23 Jul. 2021, doi:10.1186/s12876-021-01864-9
34. Takaku K, Oshima M, Miyoshi H, Matsui M, Seldin MF, Taketo MM. Intestinal tumorigenesis in compound mutant mice of both Dpc4 (Smad4) and Apc genes. Cell. 1998;92:645–56. doi: 10.1016/s0092-8674(00)81132-0.
35. Risques RA, Kennedy SR. Aging and the rise of somatic cancer-associated mutations in normal tissues. PLoS Genet. 2018 Jan 4;14(1):e1007108. doi: 10.1371/journal.pgen.1007108. PMID: 29300727; PMCID: PMC5754046.
36. American Cancer Society**.** "*What Causes Colorectal Cancer?*" *American Cancer Society*, 29 Jan. 2024, https://www.cancer.org/cancer/types/colon-rectal-cancer/causes-risks-prevention/what-causes.html.
37. Grant, A., Xicola, R.M., Nguyen, V. *et al.* Molecular drivers of tumor progression in microsatellite stable *APC* mutation-negative colorectal cancers. *Sci Rep* **11**, 23507 (2021). https://doi.org/10.1038/s41598-021-02806-x
38. Kemp Z, Rowan A, Chambers W, Wortham N, Halford S, Sieber O, Mortensen N, von Herbay A, Gunther T, Ilyas M, Tomlinson I. CDC4 mutations occur in a subset of colorectal cancers but are not predicted to cause loss of function and are not associated with chromosomal instability. Cancer Res. 2005 Dec 15;65(24):11361-6. doi: 10.1158/0008-5472.CAN-05-2565. PMID: 16357143.
39. Kothari, Nishi et al. "Increased incidence of FBXW7 and POLE proofreading domain mutations in young adult colorectal cancers." *Cancer* vol. 122,18 (2016): 2828-35. doi:10.1002/cncr.30082
40. Siu, Ka Tat et al. "Chromosome instability underlies hematopoietic stem cell dysfunction and lymphoid neoplasia associated with impaired Fbw7-mediated cyclin E regulation." Molecular and cellular biology vol. 34,17 (2014): 3244-58. doi:10.1128/MCB.01528-13
41. Zou, X., Owusu, M., Harris, R. et al. Validating the concept of mutational signatures with isogenic cell models. Nat Commun 9, 1744 (2018). https://doi.org/10.1038/s41467-018-04052-8
42. Alexandrov, L.B., Kim, J., Haradhvala, N.J. *et al.* The repertoire of mutational signatures in human cancer. *Nature* 578, 94–101 (2020). https://doi.org/10.1038/s41586-020-1943-3

43. COSMIC. Single Base Substitution (SBS) Mutational Signatures. Wellcome Sanger Institute, version 3.4, https://cancer.sanger.ac.uk/signatures/sbs/
44. Xing, XuanXuan et al. “Polymerase Epsilon-Associated Ultramutagenesis in Cancer.” *Cancers* vol. 14,6 1467. 12 Mar. 2022, doi:10.3390/cancers14061467
45. Koh, G., Degasperi, A., Zou, X. et al. Mutational signatures: emerging concepts, caveats and clinical applications. Nat Rev Cancer 21, 619–637 (2021). https://doi.org/10.1038/s41568-021-00377-7
46. Campbell, Brittany B et al. “Comprehensive Analysis of Hypermutation in Human Cancer.” Cell vol. 171,5 (2017): 1042-1056.e10. doi:10.1016/j.cell.2017.09.048
47. Jacob, Sandrine, et al. "The Role of the DNA Mismatch Repair System in the Cytotoxicity of the Topoisomerase Inhibitors Camptothecin and Etoposide to Human Colorectal Cancer Cells." Cancer Research, vol. 61, no. 17, 2001, pp. 6555–6562. American Association for Cancer Research.
48. Caputo, Francesco et al. “BRAF-Mutated Colorectal Cancer: Clinical and Molecular Insights.” International journal of molecular sciences vol. 20,21 5369. 28 Oct. 2019, doi:10.3390/ijms20215369
49. Moore, L., Cagan, A., Coorens, T.H.H. et al. The mutational landscape of human somatic and germline cells. Nature 597, 381–386 (2021). https://doi.org/10.1038/s41586-021-03822-7
50. Colorectal Cancer Genetics." PDQ Cancer Information Summaries, National Cancer Institute, 9 May 2024, https://www.cancer.gov/types/colorectal/hp/colorectal-genetics-pdq.
51. Islam, S.M. Ashiqul et al. Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. Cell Genomics, Volume 2, Issue 11, 100179