PROLIFERATIVE AND INVASIVE COLORECTAL TUMORS IN PET DOGS PROVIDE UNIQUE INSIGHTS INTO HUMAN COLORECTAL CANCER

by

JIN WANG

(Under the Direction of Shaying Zhao)

ABSTRACT

Spontaneous tumors in pet dogs represent a valuable but under characterized cancer model. To better use this resource, we established the pipelines to study collaborating genomic, transcriptomic and microbiomic alterations for a canine extreme intestinal polyposis. We found, not APC mutation, but three other alteration pathways as likely reasons of this canine extreme polyposis, in comparison with 22 other dogs. First, somatic truncation mutation W411X of FBXW7, a component of an E3 ubiquitin ligase, over-activates MYC and cell cycle-promoting network, accelerating crypt cell proliferation. Second, genes of protein trafficking and localization are downregulated, likely associated with germline mutation G406D of STAMBPL1, a K63-deubiquitinase, and MYC network activation. This inhibits epithelial apical-basolateral polarity establishment, preventing crypt cell differentiation. Third, *B.uniformis*, a commensal gut anaerobe, thrives and expresses abundantly thioredoxin and nitroreductase. These bacterial products could reduce oxidative stress linked to host germline mutation R51X of CYB5RL, a cytochrome b5

reductase homologue, decreasing cell death. Our work emphasizes the close collaboration of alterations across the genome, transcriptome and microbiome in promoting tumorigenesis.

Additionally, we performed an initial global comparison between proliferative and invasive colorectal tumors from 20 canine cases and evaluated their molecular homology to human colorectal cancer (CRC). First, proliferative canine tumors harbor overactivated WNT/β-catenin pathways and recurrent CTNNB1 (β-catenin) mutations S45F/P, D32Y and G34E. Invasive canine tumors harbor prominent fibroblast proliferation and overactivated stroma. Both groups have recurrent TP53 mutations. We observed three invasion patterns in canine tumors: collective, crypt-like and epithelial–mesenchymal transition (EMT). We detected enriched *H.bilis* and *A.finegoldii* in proliferative and crypt-like tumors, but depleted mucosa-microbes in the EMT tumor. Second, guided by our canine findings, we classified 79% of 478 human colon cancers from The Cancer Genome Atlas into four subtypes: primarily proliferative, or with collective, crypt-like or EMT invasion features. Their molecular characteristics match those of canine tumors. We showed that consensus molecular subtype 4 (mesenchymal) of human CRC should be further divided into EMT and crypt-like subtypes, which differ in TGF-β activation and mucosa-microbe content.

INDEX WORDS: microbiome; germline and somatic mutation; spontaneous canine colorectal tumors; human-dog comparison; cancer cell proliferation and gene mutations; cancer cell invasion and stromal activation; CMS4 and crypt-like or EMT invasion.

PROLIFERATIVE AND INVASIVE COLORECTAL TUMORS IN PET DOGS PROVIDE UNIQUE INSIGHTS INTO HUMAN COLORECTAL CANCER

By

JIN WANG

BS, Capital University of Economics and Business, China, 2012

MS, George Washington University, 2014

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2019

PROLIFERATIVE AND INVASIVE COLORECTAL TUMORS IN PET DOGS PROVIDE UNIQUE INSIGHTS INTO HUMAN COLORECTAL CANCER

By

JIN WANG

Major Professors: Shaying Zhao

Committee: Jonathan Arnold
Kevin Dobbin
Tianming Liu

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2019

DEDICATION

This dissertation is dedicated to the people who I love and love me.

ACKNOWLEDGEMENTS

I would like to express the deepest appreciation to my advisor Shaying Zhao for her support and guidance in the five years of my doctoral study. She has been instrumental in my scientific training and taught me how to conduct scientific investigations. I would also like to thank the members of my advisory committee: Drs. Jonathan Arnold, Kevin Dobbin and Tianming Liu, for their mentorship and guidance during my graduate training. I would like to thank my fellow lab mates (past and current), Dr. Shutan Xu, Dr. Deli Liu, Tianfang Wang, Yuan Feng, Maxwell Colonna, and many more, for the academic help, fun time together and emotional support. I would like to thank my fellow graduate students, Zengyan Wang, Chuan Zhang, Qinglin Dong, Xiangran Zhao and Zhongyao Ma for their friendship and help. I would also like to thank my friends, Leng Wang, Harry Zheng, Yajun Yang, Weijia Zhan and Huaiyu Wang, for their love and encouragement to help me walk through this adventure. Special thanks to my grandparents, Dechang Wang and Qingfang Li, and my parents Huan Wang and Shuhong Zhang, for standing beside me throughout my Ph.D. studies. Lastly, I would like to thank myself that I wasn't given up during this long and hard journey.

TABLE OF CONTENTS

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

INTRODUCTION TO CANCER

Cancer results from and accumulates a series of genetic changes, from normal cells to tumor cells. Person's genetic factors interact with the external factors including physical, chemical and biological causing these genetics changes. Cancer is the abnormal cells growth more rapidly than their normal cells and metastasize into other organs through the bloodstream, which can affect any part of the body [1].

Cancer is the second leading cause of death worldwide as reported by the World Health Organization (WHO). Globally, about 1 in 6 deaths is due to cancer. In the United States, 22% of deaths in 2016 were from cancer, making it the second most common disease cause of death after heart disease among men and women [2].

COLORECTAL CANCER

Colorectal cancer (CRC) is the third most common cancer and second leading causes of death globally, as reported by the WHO. In the United States, colorectal cancer is the fourth most common cancer and the second leading cause of cancer mortality. Due to the increased screening and improvements in treatment, CRC death rate has dropped 53% from 1970 to 2016 [2]. However, new cases of CRC in adults younger than age 55 have increased almost 2% per year [2]. The

American Cancer Society estimates 145,600 people will be diagnosed with CRC in the US in 2019, and 51,020 of them will die [2].

Surgical removal is the most common treatment for patients whose cancer has not spread to lymph nodes. For all other patients, chemotherapy and radiation are the most common treatment depending on their cancer stage [3]. Targeted therapies as a new treatment, which were approved in 2004 for metastatic CRC patients, improves the overall survival [4]. The five-year survival rate was 64.9% from 2007-2013, as reported by the National Cancer Institute (NCI) [5].

CRC subtypes

For CRC subtypes, there were different independent classification systems published before the consensus molecular subtypes (CMSs) was identified in 2014. The genetic and epigenetic characteristics, and gene-expression based are two major methods for the classification [6].

There are two type of the classifications followed the genetic and epigenetic chatacteristics. Microsatellite instability (MSI) [7] chromosomal instability (CIN) [8] and the CpG island methylator phenotype (CIMP) [9] are the three major phenotypes from adenoma to carcinoma. However, it is hard to indicate the relationships among those three phenotypes. Another one is the mutation-centered classification, which is based on the step of the mutations. Begin from the APC inactivated mutations, followed by the KRAS mutated, then occur in elements of TP53, PI3K, TGF-β and other pathways [6]. However, as the sample size increasing, the mutation-centered classification is too simple to fully explain the patient diversity.

Although different gene-expression based classifications studied in different datasets, applied different clustering methods and utilized different statistics, they all contained CIN and MSI

subtypes. However, it is hard to find the relationship among those classifications. Therefore, in 2014, the CRC Subtyping Consortium (CRCSC) applied the network-based meta-analysis on all datasets and observed that the four CMSs, which are CMS1(MSI immune subtype), CMS2(canonical subtype), CMS3(metabolic subtype) and CMS4(mesenchymal subtype) [10].

DRIVER-PASSENGER DISCRIMINATION

Some of the genetic changes are cancer-causative, which are called drivers, while others are not cancer-causative, which are called passengers. Identifying drivers is a major goal of cancer research, as it will lead to a better targeted cancer therapy.

A number of strategies have been developed for cancer driver-passenger dissemination. Our lab has pioneered a novel dog-human comparison strategy to address this critical question. Specified, a former lab member Dr. Jie Tang has successful discriminate drivers from passengers for copy number abnormalities (CNAs) with array comparative genomic hybridization (aCGH) data for CRC [11]. However, besides CNAs, cancer genomes also harbor numerous sequence mutations and expression alterations, which can be drivers or passengers. Furthermore, with the advancement of next generation sequencing, more cancers are sequenced with strategies such as whole genome sequencing (WGS) and RNA sequencing (RNA-seq). As a result, significant more mutations and gene expression alterations are uncovered. Efficient driver-passenger discrimination becomes even more urgent.

Currently, driver-passenger discrimination models fall into three main categories: frequency-based, function-based and integrated frequency- and function-based models. The frequency-based

approach applies to both frequently and infrequently mutated drivers. For frequently mutated drivers, one approach assumes that driver genes are highly mutated in cancer samples compare with the background mutation rate [12]. Since the mutation rates between samples and across the genome itself vary, the background mutation frequency is difficult to estimate accurately [13]. In general, there are two ways to estimate the background mutation frequency, which include are estimating from the synonymous mutation rates and estimating from the introns and untranslated regions (UTRs). The gene-specific mutation frequencies usually have low rates of synonymous mutations that may affect the estimating accuracy [14]. Using introns and UTRs to estimate the background mutation frequency assumes that they are selectively neutral [15], which is not always correct [16-18]. Another approach uses genomic features, such as replication timing, GC content, gene density, distance to telomeres and centromeres, and nucleosome occupancy, to estimate background mutation rates [12], but a limiting factor is that the mutation frequency on those factors is less than 40% in cancer genomes [19].

While, many driver genes are infrequently mutated due to other driver genes are mutated with functional redundancy in the same pathway. For those mutated drivers, identifying those driver genes through the gene pathway is the current trend. Defining pathways by known pathway databases [20-22] or protein- protein interactions [23,24] faces the challenge that the known database is not completed [20]. This is another way to build the pathways containing mutually exclusive mutations in the same pathway and presumably functionally redundant mutations in different pathways [25]. However, mutually exclusive mutations may indicate negative

relationship [26].

Driver genes also can be detected by predicting the functional impact of mutations, which could use data from a single sample that is good for a small size study. However, the accuracy of this method across eight existing methods, which are SIFT, PolyPhen2, Condel, CHASM, mCluster, logRE, SNAP, and MutationAssessor is too low [27]. The reason is that error may be introduce into the model, such as well-conserved domains mutated that may not be the driver and poorly conserved domains mutated that may not be the passenger. Therefore, overlapping driver gene lists from frequency-based and function-based approaches is another way to detect driver genes, which can increase the accuracy [28].

MICROBIOTA ASSOCIATED WITH CRC

In microscopic studies, microbial cells, which outnumber human cells by about ten to one in a healthy human body, contribute to the human's health and behavior. However, those human-associated microbes are only partially understood in their function and dynamics in healthy or disease states [29]. Some colorectal bacteria and their metabolites could cause pro-inflammatory and DNA-damage thought to lead to colorectal cancer [30]. Additionally, comparing the microbes between the patients with CRC and healthy subjects, CRC associated bacterial co-abundance groups were differentially correlated with the expression of host immunoinflammatory response genes [31].

However, only a few studies discussed the pathogenic with CRC. Streptococcus bovis increase mostly happened in the early stage of CRC, which could be used for CRC early detection

[32,33]. Helicobacter pylori contains toxin genes, CagA and VacA, increasing the risk of CRC [34-36]. Bacteroides fragilis contains toxin gene, BFT, which cause cell proliferation and DNA damage [37-39]. Fusobacterium nucleatum, a pathogenic bacterial, increase in adenomas in the early stage that activates the Wnt/β-catenin pathway [40-42]. Escherichia coli contains toxin genes, such as afa, lpfA, eae and cyclomodulin toxins may cause DNA damage and genomic instability [43-45].

EXPERIMENTAL DESIGN AND INNOVATION

Spontaneous canine CRCs are a valuable cancer models for studying human CRCs.

Compared with other traditional animal models, such as mouse models, dog cancer serves as a valuable model analogous to humans for three reasons.

Environment. Unlike other rodent models, dog cancers are naturally occurring and heterogeneous, capturing the essence of human cancer. The dog shares the human living environment as well as the same carcinogens, such as second-hand smoke, microbiome, radiation exposure and chemical additives/preservatives in food. Besides, obesity, advancing age, and genes/improper breeding are also the risk factors causing dog cancers [46].

Genome. Comparing with other shared human environment pets, such as cats, the dog genome sequence coverage is deeper (~7.5x) and more accurate in the mammalian genome sequence [47,48]. Additionally, unlike other prevalent used model, the dog genome and genetic hallmarks are more similar to humans [47,49]. Furthermore, dog has the same types/subtypes of cancer with human and has similar clinical signs and treatment schemes [49]. Thus, the dog become the best available animal model to study cancer or other common diseases between human

and dog.

Population. In 2016, pet dogs reached 89.7 million in the US (https://www.americanpetproducts.org/press_industrytrends.asp). As the most common natural cause of death in the dog, cancer affected approximately half of dogs each year [50]. Thus, sufficient cancer and healthy pet dog samples provide a valuable resource, which is beneficial to the scientific research. In SRA database, there are above 1300 WGS samples, 760 WES samples and 700 RNA-seq samples, including more than 100 breeds.

Our study expanded the novel dog-human comparison strategy.

On the genomic level, we expanded the driver-passenger discrimination from the copy number change to the somatic missense mutation. Additionally, in order to ensure that dog and human mutations comparable, we classified canine CRCs into two major subtypes: proliferative and invasive, based on the molecular signature and microbiome analysis. Meanwhile, we studied The Cancer Genome Atlas (TCGA) database and evaluated their molecular homology to human CRC samples. Besides, since the microbiome is important for the CRC development, we developed the pipelines for microbiome discoveries with WGS and RNA-seq data.

To expand the novel dog-human comparison model in CRCs, we performed the WGS and RNA-seq on 20 canine CRC samples, found two CRC subtypes. As the known driver mutation, TP53 recurrently mutated in both proliferative and invasive group. Then, guided by our canine findings, we classified 79% of 478 human colon cancers from The Cancer Genome Atlas into four subtypes: primarily proliferative, or with collective, crypt-like or EMT invasion features. Their

molecular characteristics match those of canine tumors. Lastly, we detected enriched *Helicobacter bilis* and *Alistipes finegoldii* in proliferative and crypt-like tumors, but depleted mucosa-microbes in the EMT tumor.

CHAPTER 2

COLLABORATING GENOMIC, TRANSCRIPTOMIC AND MICROBIOMIC ALTERATIONS LEAD TO CANINE EXTREME INTESTINAL POLYPOSIS[1]

---

[1]Wang, Jin, et al. "Collaborating genomic, transcriptomic and microbiomic alterations lead to canine extreme intestinal polyposis." Oncotarget 9.49 (2018): 29162.
Reprinted here with the permission of the publisher.

ABSTRACT

Extreme intestinal polyposis in pet dogs has not yet been reported in literature. We identified a dog patient who developed numerous intestinal polyps, with the severity resembling human classic familial adenomatous polyposis (FAP), except the jejunum-ileum junction being the most polyp-dense. We investigated this dog, in comparison with 22 other dogs with spontaneous intestinal tumors but no severe polyposis, and with numerous published human cancers. We found, not APC mutation, but three other alteration pathways as likely reasons of this canine extreme polyposis. First, somatic truncation mutation W411X of FBXW7, a component of an E3 ubiquitin ligase, over-activates MYC and cell cycle-promoting network, accelerating crypt cell proliferation. Second, genes of protein trafficking and localization are downregulated, likely associated with germline mutation G406D of STAMBPL1, a K63-deubiquitinase, and MYC network activation. This inhibits epithelial apical-basolateral polarity establishment, preventing crypt cell differentiation. Third, Bacteroides uniformis, a commensal gut anaerobe, thrives and expresses abundantly thioredoxin and nitroreductase. These bacterial products could reduce oxidative stress linked to host germline mutation R51X of CYB5RL, a cytochrome b5 reductase homologue, decreasing cell death. Our work emphasizes the close collaboration of alterations across the genome, transcriptome and microbiome in promoting tumorigenesis.

INTRODUCTION

Canine cancers represent one of the best animal models of human cancers [47,51-53], because of the shared biology (e.g., intact immune system), physiology, living environment and clinical symptoms between the two species. Indeed, genomic studies from our group and others have revealed a high degree of molecular homology for histopathologically matched cancer types/subtypes between the dog and the human [54-56]. For example, the stepwise model of human colorectal tumorigenesis [57] also applies to spontaneous colorectal tumors in pet dogs [54,58]. Furthermore, our group has successfully developed a novel dog-human comparative genomics and oncology strategy for driver-passenger discrimination, a central aim of cancer research [59], for colorectal cancer (CRC) copy number alteration [11,60].

In humans, individuals with classic familial adenomatous polyposis (FAP) differ from the general population [61,62]. These individuals develop significantly more (>100) adenomatous polyps in their colon, beginning at much younger age (age 16 on average). They also develop polyps in the small intestine and other places [63,64]. The underlying pathogenic mechanism is well studied [57,61,62]. Most classic FAP patients inherit a dysfunctional copy of the APC gene, and then, a second mutation inactivates the other functional copy of APC. This results in translocation of β-catenin into the nucleus, activating WNT/β-catenin targets (e.g., MYC) [65] and accelerating cell proliferation. Furthermore, defective APC also interferes with cell adhesion, cytoskeleton and the establishment of epithelial apical-basolateral polarity All these lead to extreme colon polyposis. Other variants of FAP include attenuated FAP, which is also APC

mutation-associated, but the patients typically develop polyps at older age, and autosomal recessive FAP, which is MUTYH mutation-associated and the patients develop fewer polyps. Hereditary nonpolyposis colorectal cancer (HNPCC), another inherited condition, is caused by mutations of DNA mismatch repair genes [61,62,66-68] and others [69].

Extreme intestinal polyposis in pet dogs has not yet been reported in literature, and the underlying pathogenic mechanism is unknown. We are fortunate to identify such a case. We set out to molecularly characterize this rare canine condition and compare our findings with those of human studies, as described below.

RESULTS

N14-77 represents a rare canine case of extreme intestinal polyposis

A rare canine case of extreme intestinal polyposis (Figure 2.1) was diagnosed at the Texas A&M University Veterinary Medical Teaching Hospital and assigned "N14-77" as the case identifier. The detailed case information is provided in Supplementary Information and summarized below.

At presentation, the N14-77 patient, a 9-year-old neutered male dog of Golden Retriever-mix, had a two-month history of blood-tinged, watery diarrhea and was in poor body condition. Complete blood count revealed a microcytic, hypochromic, regenerative anemia with a severe neutrophilia and hypoalbuminemia. Abdominal radiographs and ultrasounds indicated extensive intestinal changes. A rectal scraping found numerous, degenerate neutrophils containing phagocytosed bacteria and small yeast.

Euthanasia was selected. A full necropsy indicated that, while no significant abnormalities in other organ systems, about 70% of the small intestinal mucosa was affected. Specifically, intestine, extending primarily from the mid-jejunum to the ileocecal junction, was severely thickened by innumerable, 3 mm to 1.1 cm, firm nodules that progressively coalesced into large, plaque-like, 10-30 cm-long areas with a red, granular surface. The most severe region located at the distal jejunum-ileum junction (Figure 2.1A).

Histologic examination indicated numerous single to coalescing polyps within the mucosa of sections from the jejunum to the proximal colon, and the epithelium from crypts to mucosal surface was uniformly hyperplastic (Figure 2.1B). The mucosa comprising the inter-polyp regions and within the distal colon also displayed mild to moderate hyperplasia, with variable neutrophilic infiltration and mild enterocolitis. Notably, neither malignant neoplastic transformation of nor invasion of the lamina propria by enterocytes lining the intestinal villi, crypts, or colonic glands was observed (Figure 2.1B).

Except for the location (extending primary from the mid-jejunum to the ileocecal junction and with the distal jejunum-ileum junction being the most affected), the severity of polyposis in N14-77 resembles classic FAP patients in humans.

We performed whole genome sequencing (WGS) and RNA-seq

To characterize N14-77, we performed WGS and RNA-seq analyses with frozen polyp and normal (or rather unaffected) samples. To maximally identify molecular changes associated with extreme intestinal polyposis, we chose polyps dissected from the most affected and polypdense

area, located at the distal jejunum-ileum junction (Figure 2.1A), for polyp WGS and RNA-seq. Hence, the findings represent multiple polyps but not individual ones. As controls, we performed WGS with unaffected tissue dissected from one of the inter-polyp regions of the midjejunum (Figure 2.1A), as well as RNA-seq with unaffected submucosa and muscularis propria tissue dissected away from the polyp-dense mucosa used in polyp-sequencing (Figure 2.1B). Thus, WGS and RNA-seq normal samples differ in their locations.

For WGS, we generated a 15X sequence coverage for the polyp sample and a 13X sequence coverage for the normal sample, with a fragment coverage at approximately 21X (Supplementary Table 2.1A). For RNAseq, we acquired about 80 million paired-reads for the polyp sample and 74 million paired-reads for the normal sample (Supplementary Table 2.1B). For comparison, we also performed WGS and/or RNA-seq analyses with 26 intestinal normal or tumor samples from 22 dogs with spontaneous intestinal tumors, along with a healthy dog (Supplementary Table 2.1C). Differing from N14-77, none of these dogs have this extreme polyposis phenotype. We developed a pipeline (Supplementary Figure 2.1) to interrogate the data.

We corrected genomic sequence errors in the canine APC gene

Because APC mutations characterize the human FAP condition, our initial hypothesis was that APC is mutated in N14-77. We hence first investigated canine APC but noted that it is annotated inconsistently. In the Ensembl database, APC consists of only three exons, unlike its human counterpart which has 18 exons (Figure 2.2A). In the Broad annotation [70], APC has more exons but only three are coding (Figure 2.2A). This again differs from human APC, which

has 14-16 coding exons among its transcripts (Figure 2.2A). The human "xenoRefGene" annotation shows better resemblance (Figure 2.2A). This however has been achieved by mapping human APC sequences onto the dog genome, not using dog-specific data.

To resolve this inconsistency, we examined our WGS and RNA-seq reads that are mapped onto the canine APC locus. These reads are from normal and tumor intestinal tissues of 23 dogs (Supplementary Table 2.1C), in addition to N14-77, and from other canine tissues [55,56]. We detected 5 sequence errors in exon 3 of APC in the canine reference genome [47], including two base substitutions, two base deletions and one base insertion (Figure 2.2B; Supplementary Figure 2.2). These errors result in premature stop codons and mis-annotation of APC in Broad and Ensembl databases (Figure 2.2A).

After removing the errors, we remapped our RNAseq reads (Supplementary Figure 2.2) and reassembled the canine APC gene and transcripts. Five transcripts were identified, with 16-17 exons in total of which 15-16 are coding (Figure 2.2C), better matching their human counterparts. The transcripts yield slightly different protein isoforms. One isoform is nearly identical to the canonical human APC protein (Supplementary Table 2.2B). The other isoforms have insertions or deletions, all occurring before the armadillo repeat domain (Figure 2.2C).

We found neither germline nor somatic APC mutations

To discover APC mutations in N14-77, we investigated WGS and RNA-seq reads, individually or combined, of polyp and normal samples. The combined sequence coverage reaches to 136X on average for APC coding regions (Supplementary Table 2.2C). We used

popular software tools, GATK for germline- and MuTect for somatic mutation discovery, with the corrected APC genomic sequence (Figure 2.2B). Surprisingly, neither germline nor somatic mutations were detected. To confirm this, we manually examined sequence read alignment of each of the 18 APC exons (Figure 2.2C) with IGV, a widely-used genomics viewer. No convincing mutations were noticed. None of the changes are significantly recurrent, with most found in one or two reads (Figure 2.2D; Supplementary Figure 2.3).

Given the prominence of APC mutation in human FAP, our result is somewhat unexpected. To determine if APC alters via other mechanisms, we examined its expression and found no significant alteration (Figure 2.2E). This is because the N14-77 normal sample expresses a comparable level of APC_AS1, which corresponds to the canonical human APC (Supplementary Table 2.2A), as normal colon samples of other dogs. N14-77 polyps express a lower amount of APC_AS1, but the level is still higher than colorectal tumors with APC deletion (Figure 2.2E). Notably, we did not detect significant nuclear enrichment of β-catenin in N14-77 polyp cells (Figure 2.2F), which does not support APC inactivation.

Besides APC, we also investigated other genes with germline mutations involved in human colorectal tumor development. These include AXIN2, BMPR1A, GREM1, MUTYH, PTEN, SMAD4 and STK11, as well as others (e.g., DNA mismatch repair genes) [61,62,66-69]. We examined both coding regions, with average sequence coverages ranging 136-615X after combining WGS and RNA-seq reads of both polyp and normal samples (Supplementary Table 2.2C), as well as their promoters (22-45X sequence coverage; see Supplementary Table 2.2C).

We only detected two germline mutations: L588P of AXIN2 and A161V of MUTYH. Both mutations have however occurred during evolution (Supplementary Table 2.2C) and thus are most likely natural variants - not pathogenic. We also detected three germline mutations in the AXIN2 promoter and two germline mutations in the MUTYH promoter. We however did not identify any transcription factor binding sites that are affected by these mutations. Moreover, both AXIN2 and MUTYH are expressed in N14-77 samples at a level resembling other canine intestinal normal and tumor samples (Supplementary Table 2.2D). These observations indicate that the identified promoter mutations are unlikely pathogenic. Other genes are also expressed in N14-77 polyps and/or normal sample at a comparable level as in other canine samples (Supplementary Table 2.2D), indicating that germline epigenetic silencing or activating is unlikely.

Germline mutations of other genes were identified in N14-77

STAMBPL1 G406D is the most notable germline missense mutation

After individual gene study described above, we attempted genome-wide search. GATK identified a large number of missense mutations, and we developed a pipeline (Figure 2.3A) to reduce false positives and to prioritize mutations. First, we selected mutations that were found in both normal and polyp samples and by both WGS and RNA-seq analyses. This step, ensuring that missense mutations of interest are indeed expressed, yields 4,329 mutations in total (Figure 2.3A). Second, we chose mutations that are unique to N14-77, when compared to >50 cases of sporadic canine intestinal (Supplementary Table 2.1C) and other cancers [55,56], and by

excluding canine SNPs from published studies [47,71]and databases. This is because N14-77 is the only case known to have extreme intestinal polyposis. A total of 135 mutations remain after this step (Figure 2.3A). Third, we excluded mutations located in genes with annotation issues (e.g., retrogenes or pseudogenes). To further increase the accuracy, we selected mutations with: 1) a ≥10 WGS read coverage in both normal and polyp samples; 2) a ≥30X RNA-seq read coverage in either the normal or polyp sample; and 3) a ≥0.5 variant allele frequency in the polyp sample for either WGS or RNA-seq. These selections reduce the total mutations to 21 (Supplementary Table 2.3A). For heterozygous mutations, we further prioritized those being selected in polyps, i.e. with a higher mutation rate in the polyp than in the normal sample (Figure 2.3; Supplementary Table 2.3A). Lastly, we considered evolutionary conservation and excluded mutations occurring during evolution (Supplementary Table 2.3A). In addition, we prioritized mutations predicted to alter the protein 3D structure with modeling [72] (Supplementary Table 2.3A).

After processing through our pipeline (Figure 2.3A), G406D of STAMBPL1 is the only germline missense mutation that remains. This mutation is unique to N14-77 and being selected in polyps, with the mutation rate increasing from 67% in the normal sample to 82% in the polyp sample for WGS and from 50% to 70% for RNAseq (Supplementary Table 2.3A). Furthermore, G406, the glycine residue at position 406, is conserved from fish to mammals for 100 species examined (Figure 2.3B). Based on modeling [72], the G406D change will likely destabilize the protein (Figure 2.3B).

STAMBPL1, also known as AMSH-LP, is a deubiquitinase (DUB) that cleaves K63-linked polyubiquitin chains. The crystal structure of human STAMBPL1 is determined [73]. As canine STAMBPL1 is highly homologous to its human counterpart, with the same length and sharing 92% similarity and 88% identity in amino acid sequence (Supplementary Figure 2.4A), we used the human structure to study the G406D mutation. STAMBPL1 is a zinc protease and contains two zinc centers. G406 locates in the 2nd zinc-center, neighboring the zinc-coordinating residues C402, H408 and H410 within a highly conserved peptide (C402KKK405G406F407H408PH410) (Supplementary Figure 2.4B). This peptide forms a long loop, assisting the recognition and correct binding of the proximal ubiquitin of K63-linked ubiquitin chains [73]. Importantly, K405 and F407, which flank G406, are the most frequently mutated residues of STAMBPL1 in human cancers, with frameshift mutations recurrently found (Figure 2.3B; Supplementary Table 2.3B). Based on these findings, the G406D germline mutation of STAMBPL1 may be pathogenic.

CYB5RL harbors a germline truncation mutation

We followed the same procedure of Figure 2.3A, except skipping the steps of evolutionary conservation and protein 3D structure, for germline truncation mutation discovery. We manually confirmed the results with IGV and the UCSC and Ensembl genome browsers. With these, we detected a truncation mutation, R51X, in CYB5RL (cytochrome b5 reductase like) (Figure 2.3C). This mutation is selected in the polyps, although the allele frequency is rather low (25%) based on RNA-seq reads (Supplementary Table 2.3C). There appears to be a second, but minor, alternative splicing form that is not affected by this mutation (Figure 2.3C).

No germline frameshift mutations found

We applied the same strategy described above and found no convincing germline frameshift mutations in N14-77. A small number of indels were detected, which however locate in intron regions, retrogenes, or misannotated genes based on manual examination.

Somatic mutations were identified in N14-77 polyps

Somatic truncation mutations of FBXW7, LRBA and MACF1 found

We detected 7 total somatic truncation mutations in N14-77, three of which are supported by both WGS and RNA-seq analyses (Supplementary Table 2.3D). The 1st mutation is W411X of FBXW7, occurring at a rate of 39% for WGS and 59% for RNA-seq (Figure 2.3D; Supplementary Table 2.3D). FBXW7 is one of the most frequently (~20%) mutated genes in human CRC [74]. The 2nd mutation is Q940X of LRBA (Figure 2.3D), at a rate of 90% for WGS and 38% for RNA-seq (Supplementary Table 2.3D). LRBA (lipopolysaccharide-responsive vesicle trafficking, beach- and anchor-containing) is linked to trafficking of immune molecules such as CTLA4. LRBA deficiency, a rare genetic disorder, is associated with autoimmunity, chronic diarrhea, and B-cell deficiency [75]. The 3rd mutation is W840X of MACF1 (microtubuleactin crosslinking factor 1) (Supplementary Table 2.3D), the loss of which disrupts epithelial cell polarity [76].

Somatic frameshift mutations of CBLB and other genes found

CBLB encodes an E3 ubiquitin ligase CBL-B, an immune response regulator [77]. We detected a somatic base T deletion at a rate of 50% for WGS and 29% for RNA-seq, resulting in

a frameshift mutation at residue D404 (D404fs) of CBLB (Figure 2.3E; Supplementary Table 2.3D). Frameshift indels were also uncovered in DIDO1, involved in apoptosis [78], as well as within homopolymer sites (e.g., GGGGGG) of KDM7A, IRF2BP2 and CCNB3 (Figure 2.3E; Supplementary Table 2.3D).

G→A/C→T changes dominate among somatic base substitutions

We identified 72 missense mutations in total (Supplementary Table 2.3E). Consistent with human studies [74], G→A/C→T changes dominate over other base substitutions (Figure 2.3F), indicating that C/G deamination is the major somatic mutation mechanism in N14-77 polyps. Among 72 mutations, only 7 were detected by both WGS and RNA-seq analyses (Supplementary Table 2.3E). Furthermore, quite a few mutations could be passengers, based on evolutionary conservation and molecular modeling (Supplementary Table 2.3E), as well as comparison to human mutation findings (Supplementary Figure 2.5A). However, more studies are required to determine their driver-passenger role.

Somatic whole chromosome gains detected

Our analysis (Supplementary Figure 2.1B) revealed no translocations or inversions in the N14-77 polyp or normal genome. Neither did we find focal amplifications/deletions. We did, however, detect whole chromosome gain of chromosomes 4, 7-10, 13, 15, 23, and 26. These changes are clearly somatic, because they were only found in the polyp genome but not in the normal genome (Supplementary Figure 2.5B).

Highly- and lowly expressed genes in N14-77 polyps are enriched in specific functions

With our RNA-seq data from 28 samples of canine intestinal tumor and normal tissues (Supplementary Tables 1B and 1C), we identified genes that are highly or lowly expressed in N14-77 polyps. These are defined as genes with an expression level outside the expression mean ± one standard deviation range and being the highest or lowest among the 28 samples. A total of 528 highly expressed genes were identified, of which about 474 (90%) encode proteins (Figure 2.4A; Supplementary Table 2.4A). These genes are significantly enriched in functions of cell cycle, DNA repair, as well as transcription, mRNA processing and splicing (Figure 2.4A; Supplementary Table 2.4A). Meanwhile, 621 lowly expressed genes were discovered, of which only one appears to be noncoding and 614 encode proteins with functional annotation (Figure 2.4A; Supplementary Table 2.4A). Among them, the prominently enriched functions include protein localization, trafficking and degradation, as well as cell cycle.

>50 ubiquitin-related genes are lowly expressed

Ubiquitin-related genes are enriched only among lowly expressed genes of N14-77 polyps (Figure 2.4A). Specifically, a total of 53 such genes are lowly expressed (Supplementary Figure 2.6A), of which ≥29 are associated with ubiquitin ligases and ≥4 are linked with DUBs (Supplementary Table 2.4A). About 35 genes are associated with protein degradation, including 6 encoding F-box proteins (Supplementary Table 2.4A). Interestingly, 39 genes (73%) are enriched in microRNA (miRNA) target sites.

Highly and lowly expressed cell cycle genes differ in cell cycle phase and function

Although cell cycle genes are enriched in both highly and lowly expressed gene sets (Figure 2.4A), they differ. Highly expressed ones consist of E2F targets and cycling genes with their expression peaking during the G1/S or G2 phase (Figure 2.4B; Supplementary Table 2.4B). DNA replication and/or repair genes, which primarily function in the S phase, and mitotic nuclear division genes are also among highly expressed (Figure 2.4B; Supplementary Table 2.4B). In contrast, 63 of the 92 lowly expressed cell cycle genes are associated with membrane organization, budding and trafficking; vesicle-mediated transport; protein localization, as well as ubiquitination and proteolysis (Figure 2.4B; Supplementary Table 2.4B). Furthermore, at least 12 lowly expressed genes encode cell cycle inhibitors, including RB1 and TSG101 (Supplementary Table 2.4B).

Highly and low expressed genes are enriched in different cellular locations

Approximately 41% of the highly expressed genes are annotated to be located in the nucleus (Figure 2.4C; Supplementary Table 2.4C). These include those (7%) located in the nucleolus, as well as genes functioning in DNA repair and/or replication, transcription and chromatin. This is significantly higher, when compared to both the lowly expressed genes (about 12%) and the entire gene set encoded in the genome (<25%). To the contrary, about 58% of the downregulated genes are located in the cytoplasm, associated with endosomes and other organelles (Figure 2.4C; Supplementary Table 2.4C). This is also significantly higher when compared to the highly expressed genes (~18%) and the entire gene set encoded in the genome (<40%).

Consistent with mRNA expression (Figure 2.4C), our immunohistochemistry (IHC) analysis reveals depletion of EGFR, a membrane protein, and of phosphorylated ERK and AKT, both cytoplasmic proteins, in N14-77 polyps, when compared to normal intestinal tissue samples (Figure 2.4D). This differs from MYC, a nuclear protein (Figure 2.4D), as described later.

Lowly expressed genes are enriched in miRNA target sites

About 56% (342 genes) of lowly expressed genes are enriched in putative miRNA target sites, compared to only 16% (77 genes) for highly expressed genes (Figure 2.4E). Interestingly, more noncoding RNA genes are found among highly expressed genes than among lowly expressed genes, as previously described. This is consistent with that more RNA-seq reads were mapped into intronic regions in the N14-77 polyp sample, compared to the other samples (Figure 2.4F; Supplementary Table 2.4D).

MYC network is activated in N14-77 polyps

Several analyses indicate that the MYC network is activated in N14-77 polyps. First, the MYC protein is expressed highly and more or less uniformly throughout the polyps, ranging from the bottom to the top of the intestinal mucosa (Figure 2.4D). This differs from normal intestinal tissues of other dogs where MYC is only expressed at the bottom layer of the mucosa (Figure 2.4D). Second, MYC has the highest mRNA expression level in N14-77 polyps, among 28 canine intestinal tumor and normal samples investigated (Supplementary Figure 2.6B).

Notably, MYC targets are enriched in both highly and lowly expressed gene sets of N14-77 polyps (Figure 2.4A). Specifically, 21 MYC targets, 38% of which function in DNA repair, and

33 MYCN targets, ~70% of which are E2F targets and/or associated with RNAbinding and processing, are highly expressed (Figure 2.5A; Supplementary Table 2.5A). Meanwhile, a total of 70 MYC targets are lowly expressed and, except for protein degradation, are enriched in the same functions as the entire lowly expressed gene set (Figure 2.5A; Supplementary Table 2.5A). Also like the entire gene sets, highly expressed MYC targets are enriched in the nucleus (76%) and depleted in miRNA target sites (4%), while lowly expressed ones are enriched in the cytoplasm (64%) and miRNA target sites (53%) (Figure 2.5B and 5C).

N14-77 polyps exhibit crypt proliferative progenitor signature

We investigated published gene signatures that mark different intestinal epithelial differentiation stages [79] via single sample gene set enrichment analysis (ssGSEA). Signature genes of crypt proliferative progenitors, but not of either intestinal stem cells or differentiated epithelial cells, are significantly upregulated in N14-77 polyps (Figures 5D and 5E; Supplementary Tables 5D and 5E). This agrees with that N14-77 polyps display upregulated signature of the crypt bottom and downregulated signature of the crypt top [80] (Figure 2.5F; Supplementary Table 2.5F). Furthermore, also consistent with the ssGSEA results, our IHC analysis reveals that N14-77 polyp cells lack well-established apical-basolateral polarity, unlike fully differentiated epithelial cells (Figure 2.5G; Supplementary Figure 2.6D). These observations indicate that N14-77 polyp cells are in the proliferative progenitor state (Figure 2.5H).

N14-77 intestinal microbiota is enriched in bacteroidetes

As described previously, medical examination indicates extensive bacterial infection in the N14-77 intestine. To better understand this, we utilized WGS and RNA-seq data to examine the intestinal microbiota. Briefly, we first identified WGS and RNA-seq read pairs of which neither read could be mapped onto the canine reference genome (Supplementary Tables 1A and 1B), which were then searched against three microbial databases. The 1st database is the reference genomes curated by the Human Microbiome Project [81], referred to as HMP hereafter. The 2nd database contains all bacterial genomic sequences (ABG) downloaded from the NCBI. The 3rd database is simplified from ABG, consisting of genomic sequence of the longest strain of each bacterial species with genome sequencing completed. It is hence named longest bacterial genomes (LBG). We noted that the results with LBG are somewhat skewed. We thus only focus on HMP and ABG studies, as described below.

Our analysis with WGS data reveals that N14-77 samples contain more bacteria than other intestinal tumor and normal samples which we investigated (Supplementary Table 2.1C). More importantly, bacteroidetes is the most enriched bacterial phylum, accounting for 67-72% for polyps and 45-48% for the normal sample, followed by proteobacteria and firmicutes (Figure 2.6A; Supplementary Table 2.6A). Other phyla each makes up < 1% (Supplementary Table 2.6A). Our results differ from typical microbiota of canine jejunum published [82], where bacteroidetes are less enriched than proteobacteria, firmicutes, actinobacteria and spirochaetes. Instead, with bacteroidetes predominating, N14-77 samples, which are from jejunum (Figure

2.1), better resemble colon in microbiota [82]. This is confirmed at the family level, where bacteroidaceae, enterobacteriaceae, clostridiaceae and tannerellaceae dominate (Figure 2.6B; Supplementary Table 2.6B). Again, families of bacteroidetes, i.e. bacteroidaceae and tannerellaceae, are significantly enriched, particularly in the polyp sample. At the species level, the top enriched include *Bacteroides uniformis* and *Clostridium perfringens* (Figure 2.6C; Supplementary Table 2.6C). While both bacteria can be found in the intestine of healthy individuals, they are thousands times more enriched in N14-77 tissues, compared to other canine intestinal samples investigated (Supplementary Table 2.1C).

***B. uniformis* is highly enriched and expresses thioredoxin and nitroreductase abundantly**

B. uniformis is the top enriched microbial species in both normal and polyp samples of N14-77 (Figure 2.6C; Supplementary Table 2.6C). Among its 6 strains examined, ATCC 8492 is about 7-800 times more enriched than others (Figure 2.6D; Supplementary Table 2.6D). Importantly, our RNA-seq data reveal that redox genes of *B. uniformis* are highly expressed in N14-77 polyps. Specifically, trxA which encodes thioredoxin, a redox protein, is the 2nd most abundantly expressed gene, while a nitroreductase gene ranks the third highest expressed (Figure 2.6D; Supplementary Table 2.6D).

Although being proliferative but not invasive (Figure 2.1B), N14-77 polyps appear to have a tissue redox state that better resembles invasive tumors than proliferative tumors. With host redox-related gene sets (Supplementary Figure 2.7A and Supplementary Table 2.6E), the N14-77 polyp sample clusters with invasive tumors, instead of proliferative tumors (Figure 2.6E). With

cell proliferation-related gene sets (Figure 2.6E; Supplementary Figure 2.7B) or in genome-wide expression (Supplementary Figure 2.7C), the opposite was observed.

C. perfringens is enriched and expresses α-toxin

C. perfringens is among the top few enriched species in N14-77 samples (Figure 2.6C) and is linked to conditions such as diarrhea and enteritis in dogs[83]. We hence examined *C. perfringens* in more depth. *C. perfringens* strains are classified into A, B, C, D and E types, based on major toxins produced [84]. There are 12 strains in our database: 3 type A, 2 type C, and one each for types B, D and E, plus 4 unclassified (Supplementary Table 2.6F). By counting WGS reads that are uniquely mapped to each strain, we note that strains 13 and SM101, both type A, and type E strain JGS1987 are slightly more enriched (Figure 2.6F; Supplementary Table 2.6F). Meanwhile, type B strain ATCC 3626 is the least enriched. Finally, we examined the expression of *C. perfringens* toxin genes with our RNAseq data. As strain 13 represents the reference strain for *C. perfringens* and its genome is well annotated [85], we used it to identify the toxin genes and found 25 of them (except for nanH). In the polyp sample, we detected substantial expression of α-toxin, a phospholipase C, and an enterotoxin (entC), as well as trace expression of μ-toxin (nagI), α-clostripain and hemolysin (hlyD) (Figure 2.6F; Supplementary Table 2.6F).

Lastly, although medical examination suggests yeast infection, we did not find any of our WGS reads mapped to the yeast genomes in the HMP database.

DISCUSSION

N14-77 represents the first reported case of extreme intestinal polyposis in the dog. Our analysis does not support the involvement of APC mutations. Instead, we propose that this extraordinary phenotype is possibly caused by an alteration network collaborating across the genome, transcriptome and microbiome (Figure 2.7), as discussed below.

Host ubiquitin gene alterations and MYC and cell cycle-promoting network activation keep cells proliferating

FAP and many CRCs in humans follow the pathogenic pathway of APC mutation → β-catenin accumulation in the nucleus → MYC upregulation and cell-cycle activation → cell proliferation [74]. With the lack of APC mutation, our data indicates that N14-77 has likely taken a different route: FBXW7 truncation mutation → MYC protein accumulation and cell cycle activation → cell proliferation. FBXW7, a F-box protein, constitutes the substrate-recognition subunit of the SKP1-cullin-F-box (SCF) E3 ubiquitin ligase that targets MYC and cyclin E for degradation [86]. The W411X truncation mutation of FBXW7 could render this SCF complex defective and unable to ubiquitinate MYC and cyclin E for degradation. Deletion of FBXW7 in the gut has induced intestinal adenomas in mice [87].

Interestingly, 53 ubiquitin genes are downregulated in N14-77 polyps, the significance and mechanism of which clearly need further studies. Among them are genes encoding TRPC4AP and CUL4A, which constitute the MYC-targeting DDB1-CUL4 E3 ligase complex [88]. This may further lead to MYC protein accumulation.

MYC is a master transcription factor. MYC protein accumulation accelerates the transcription of numerous cell cycle promoting genes. Indeed, E2F targets, DNA repair genes, and RNA processing and slicing genes are all upregulated in N14-77 polyps. These would keep N14-77 polyp cells proliferating (Figure 2.7).

Ubiquitin gene alteration and MYC network activation likely inhibit epithelial polarity establishment and cell differentiation

G406D of STAMBPL1 is the most significant germline missense mutation discovered in N14-77. STAMBPL1 (AMSH-LP) is a K63-specific DUB of the JAMM/MPN+ family [89]. G406 appears critical to its DUB activity, based on strong evolutionary conservation, crystal structure [73] and human cancer mutation findings. The G406D mutation may disrupt the DUB activity by destabilizing the 2nd zinc-center, affecting substrate binding.

The function of STAMBPL1 is not well understood at present. A study indicates that it potentiates TGFβ signaling by inhibiting SMAD7 [90]. However, our ssGSEA reveals no significant difference in TGFβ signaling between N14-77 polyps and other canine intestinal tumor and normal samples (Supplementary Figure 2.6C). Thus, it is possible that STAMBPL1 has other functions. Its homologue STAMBP (or AMSH) is known to participate in endosomal sorting of receptors and membrane proteins [91], e.g., STAMBP knockdown enhancing EGFR degradation. Consistent with this, we observed depletion of EGFR, pAKT and pERK proteins in N14-77 polyp cells. Interestingly, like N14-77 polyps, stomach cancers [92] that harbor STAMBPL1 F407fs or K405fs mutation also display upregulation of MYC target genes and

downregulation of trafficking genes (Supplementary Figure 2.4C). We propose that the DUB activity of STAMBPL1 is required for efficient sorting, trafficking and localization of proteins inside the cell. And this is disrupted by the G406D mutation, based on our model (Figure 2.7).

Intracellular sorting, trafficking and localization in N14-77 polyp cells are likely further disrupted by the downregulation of numerous genes associated with the system. MYC over-activation could be a contributing factor. First, many of these genes are known or putative MYC targets [93] and MYC can directly repress their transcription. Note that MYC co-repressors ZBTB17 (MIZ-1) and MXD3 are upregulated in N14-77 polyps. Alternatively, these genes could be downregulated via miRNAs, supported by that: 1) noncoding RNA genes are upregulated in N14-77 polyps; and 2) lowly expressed genes are enriched in miRNA target sites.

Intestinal epithelium develops through intestinal stem cells → proliferative progenitors → differentiated cells [79]. During the 2nd stage of differentiation, the cells exit the cell cycle and establish epithelial apical-basolateral polarity. The underlying molecular mechanisms are complex. However, the intracellular sorting, trafficking and localization system clearly plays a critical role. For example, it is required to target various proteins to appropriate places to build cell adherent junctions and signaling complex (e.g., PAR, crumbs, and scribble complex) for polarity establishment [60].

Our model (Figure 2.7) proposes the following. Because of the STAMBPL1 G406D mutation and downregulation of genes described above, the intracellular protein sorting, trafficking and location system in N14-77 polyps is defective. This deficiency inhibits epithelial

polarity establishment and cell differentiation and prevent cells from entering the G0 phase. This, in combination with cell cycle activation, keeps N14-77 polyp cells forever in the proliferative state. Consistent with our model, N14-77 polyp cells lack well-established apical basolateral polarity, and closely resemble intestinal proliferative progenitors.

Intestinally, our downregulated genes significantly overlap with transcripts enriched in the protruding pseudopodia formed by cells in response to migrating stimulus by fibronectin [94]. Whether this is a reason behind non-invasiveness of N14-77 polyp cells requires further investigation.

Bacterial redox gene expression possibly reduces oxidative stress and cell death

R51X of CYB5RL is another noteworthy germline mutation uncovered in N14-77. CYB5RL is not well studied, but its homologue cytochrome b5 reductase (CYB5R) is. The shorter soluble isoform of CYB5R is expressed in erythrocytes, catalyzing the reduction of methemoglobin (with Fe3+-heme) to hemoglobin (with Fe2+-heme). The longer isoform is expressed in other cell types. With a membrane-anchor domain, it constitutes the plasma membrane redox system which regulates the tissue redox state and reduces oxidative stress [95].

Our study reveals two isoforms of CYB5RL as well. The longer isoform is inactivated by the R51X mutation, which may cause oxidative stress, contributing to the altered tissue redox state of N14-77 polyps. Increased oxidative stress, along with bacterial toxins produced by *C. perfringens*, could result in tissue damage and lead to a faulty ileocecal valve (supported by medical examination), allowing colonic bacteria to spread to the small intestine. This may

possibly explain why the N14-77 jejunum microbiota, where bacteroidetes dominate, better resembles typical microbiota of the colon rather than the jejunum. Moreover, somatic mutations of CBLB and LRBA, two key immune regulators, could alter the host immune response. We propose that, as a result of all of these, *B. uniformis* thrives (Figure 2.7). Importantly, *B. uniformis* bacteria express thioredoxin, especially trxA, and nitroreductase genes abundantly. The trxA gene is essential for the survival of *B. fragilis* under aerobic condition by reducing oxidative stress [96]. Like *B. fragilis*, *B. uniformis* is an anaerobe and normally resides in the colon, where the O_2 level is lower than in the jejunum. Thus, we postulate that *B. uniformis* expresses thioredoxin (and nitroreductase) amply to remediate oxidative stress that is induced by the more aerobic environment of the jejunum and is exacerbated by the host CYB5RL R51X mutation. Meanwhile, host cells should also benefit. Our model proposes that by decreasing oxidative stress, these bacterial redox systems reduce host cell death and contribute to extreme polyposis (Figure 2.7).

In summary, we propose that three pathways lead to N14-77 extreme intestinal polyposis (Figure 2.7). First, MYC and cell cycle-promoting network activation, caused by a FBXW7 somatic mutation-initiated SCF E3 ubiquitin ligase defect, keeps crypt cells dividing. Second, defective intracellular trafficking and localization, originating from D406G germline mutation of STAMBPL1 and enhanced by MYC network activation, inhibit cell polarity establishment and cell differentiation, preventing cell cycle exit. Lastly, bacterial redox systems reduce the

oxidative stress caused by germline mutation R51X of CYB5RL, decreasing cell death. Lastly, we emphasize that future functional studies are required to validate our model.

MATERIALS AND METHODS

Canine tissue samples

Fresh-frozen (FF) canine intestinal normal tissues and spontaneous tumors were obtained from various Veterinary Colleges (Supplementary Table 2.1C). Samples were collected from client-owned dogs that develop the disease spontaneously, under the guidelines of the Institutional Animal Care and Use Committee for use of residual diagnostic specimens and with owner informed consent. The breed, age, histopathologic descriptions, and other information are provided in Supplementary Table 2.1C.

Tissue dissection, DNA and RNA extraction, and quality control

Cryosectioning of FF tissues, H&E staining and cryomicrodissection were performed as described [54,56] to enrich polyp/tumor cells for the polyp/tumor sample, as well as unaffected/normal cells for control/normal samples. Genomic DNA and RNA were then extracted from the dissected tissues using the AllPrep DNA/RNA Mini Kit (cat. no. 80204) from QIAGEN. Only samples with a 260/280 ratio of ~1.8 (DNA) or ~2.0 (RNA) and showing no degradation and other contaminations were subjected to further quality control with qPCR and qRTPCR analysis with a panel of genes as previously described [56,58].

Paired-end WGS and RNA-seq

Both types of sequencing were conducted using the Illumina platform, following the protocols from the manufacturer. Paired-end 125 x 125bp WGS was performed in collaboration with the BGI-America and the High Throughput Genomics Core Facility at Huntsman Cancer Center at the University of Utah. RNA-seq was performed in collaboration with the Georgia Genomics Facility at the University of Georgia.

Sequence data analyses

The overall sequence analysis pipeline was summarized in Supplementary Figure 2.1 and described in detail in Supplementary methods. Briefly, WGS reads were aligned to the dog reference genome canFam3.1 [47] with BWA [97] v0.7.10. RNA-seq reads were mapped to the same reference genome using either TopHat [98] 2.1.1 (for gene expression) or STAR [99] v2.4.1c (for mutation finding). Three canine gene annotation databases were used, including Ensembl and the Broad annotation [70], both RNA-seq based, and human xenoRefGene [56]. Known canine SNPs used include those reported in other canine samples by us [55,56] and the Broad Institute [47], as well as data from the NCBI, Ensembl, and DoGSD [71] databases. Both WGS and RNA-seq reads were used for germline mutation discovery with GATK [100] v3.6 and for somatic mutation finding with MuTect [101], following pipelines recommended by the Broad Institute. WGS data were used to identify germline and somatic inversions/translocations and chimeric fusion genes as described before [54-56]. For copy number changes, correctly and uniquely mapped WGS read pairs were used to calculate mapped pair density per 1kb tiling

window along a chromosome. Each density was normalized against the corresponding value of a control genome and then used for germline and somatic copy number change discovery as previously described [54-56]. Gene expression quantification with RNA-seq reads and other analyses were performed as previously described [55,56].

Microbiome analysis

WGS and RNA-seq read pairs that could not be placed onto the canine genome were mapped with BWA v0.7.10 to three microbial genome databases – HMP, ABG and LBG. HMP is the reference genomes curated by the Human Microbiome Project [81]. HMP consists of genomic sequences of bacteria (1751 strains from 1253 species), viruses (3683 strains from 1420 species), archaea (131 strains from 97 species) and 326 lower eukaryotic species. ABG contains all bacterial genomic sequences (ABG) downloaded from the NCBI, with 2,845,483 sequences in total from 2679 species. LBG is simplified from ABG by: 1) selecting species with complete genomic sequences; and 2) for species with multiple strains having complete genomic sequences, selecting the longest strain. LBG consists of 1,576 bacterial species.

Mapped WGS read pairs were used to estimate microbial enrichment in each sample. First, the taxonomy data downloaded from the GOLD database (gold.jgi.doe. gov) were used to classify each bacterial species. Second, mapped WGS read pairs were selected as follows. For pairs with at least one read uniquely mapped, those with mapping quality Q >0 were selected. For pairs with both reads duplicatedly mapped, those that are correctly mapped (i.e., both reads mapped to the same DNA fragment, in correct orientation and spanning a reasonable genomic

distance) were selected. Third, each selected read pair was assigned as follows. A read pair was assigned to a phylum and counted as one, if it was mapped to this phylum only and no matter how many times it was mapped within this phylum. If a read pair was mapped to ≥ 2 different phyla, it was discarded. Lastly, read pairs assigned to each phylum was tallied and used to estimate the phylum enrichment. The same procedure was followed to estimate the family, species and strain enrichment.

Bacterial genome annotation data were downloaded from NCBI and Ensembl. HTSeq [102] v0.6.1 was used to tally correctly and uniquely mapped RNA-seq reads pairs within each gene, which were then used to estimate the expression levels of bacterial genes.

Immunohistochemical analysis

Immunohistochemical (IHC) experiments were performed with 5-μm tissue sections as described [56]. Primary antibodies used include including those against E-cadherin (R&D Systems, AF648), β-catenin (Santa Cruz, sc-7199), MYC (Abcam, ab32072), EGFR (BioGenex, PU335-UP), phospho-Erk1/2 (Cell Signaling Technology, #4370) and phospho-Akt (Cell Signaling Technology, #4060). Alexa Fluor®488–, 647– or 594– conjugated secondary antibodies are from Jackson ImmunoResearch. Images were taken with a Zeiss LSM 710 confocal microscope.

Data access

Sequence data have been submitted to the NCBI SRA database with accession number PRJNA418842.

Acknowledgments

We thank Ms. Yanfang Sun, Ms. Xiong Huan, Ms. Jin Qian and Ms. Ye Wang for their contribution to the study; Mr. Roger Nilsen and the Georgia Genomics Facility, Dr. Brian Dalley at the High throughout Genomics Core of the University of Utah and the BGI for sequencing; Drs. Jan Mrazek, Michael Adams and Stephen W. Ragsdale for their help and useful discussion on the microbiome study; and Drs. William Kisseberth, Carolyn J. Henry, Susan E. Lana and Nicole C. Northrup for helping collecting canine samples. Confocal imaging was performed at the UGA Biomedical Microscopy Core.

FIGURES

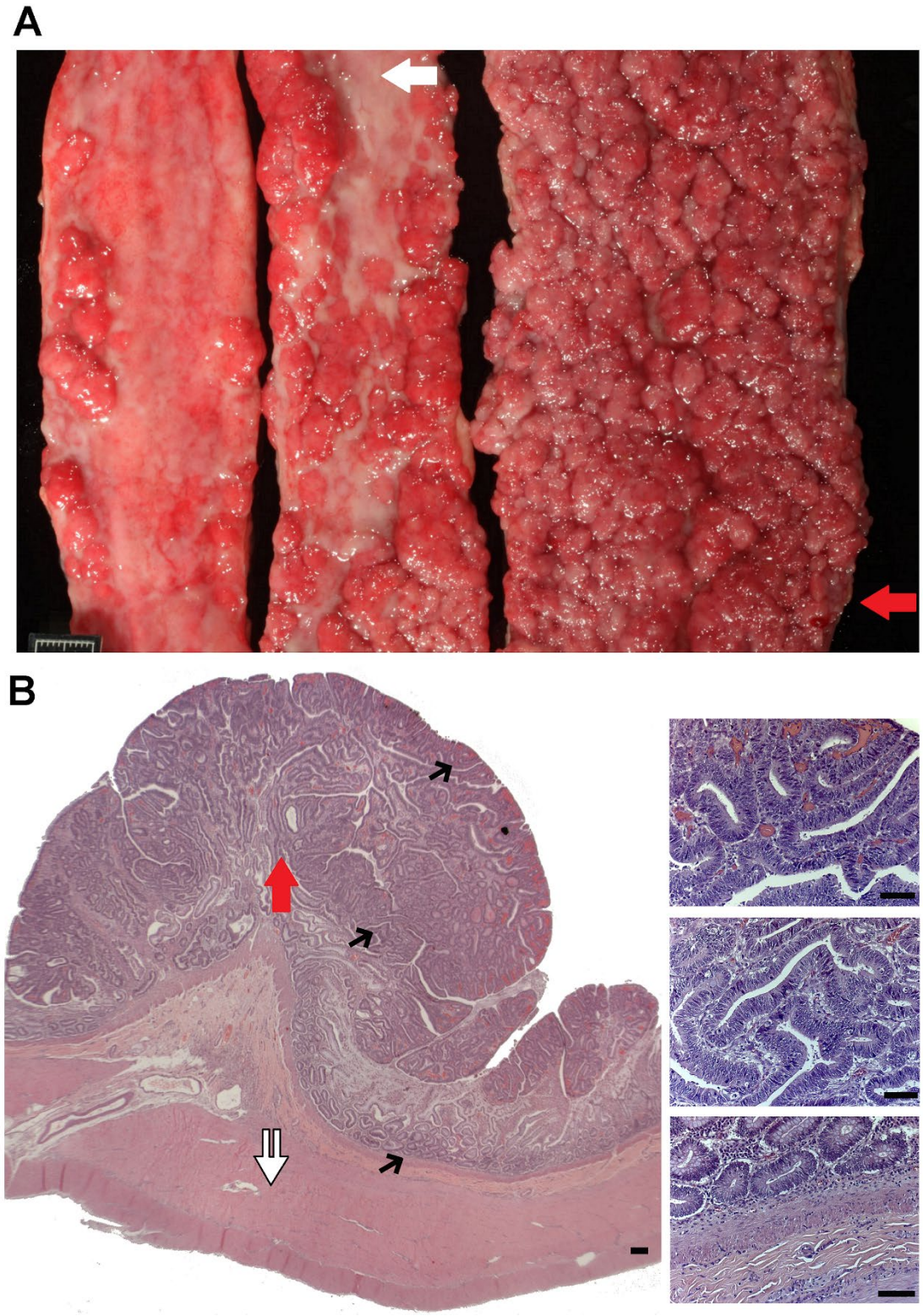


Figure 2.1: N14-77 represents a rare case of extreme intestinal polyposis in the dog.

(A) Opened small intestinal segments from left to right are from the proximal jejunum, middle jejunum and distal jejunum-ileum junction, respectively. The red arrow indicates the area used for polyp dissection and sequencing (WGS and RNA-seq). The white arrow illustrates an unaffected inter-polyp region used for normal sample WGS. The scale bar is 1cm-long.

(B) Representative H&E images of the distal jejunum-ileum junction indicate extensive cell proliferation and no invasion of proliferating enterocytes into the lamina propria or submucosa. The white double arrow exemplifies unaffected submucosa and muscularis propria tissue being dissected for normal sample RNA-seq. Images on the right are blowups of the corresponding sites pointed by black arrows on the left. Scale bar, 50μm.

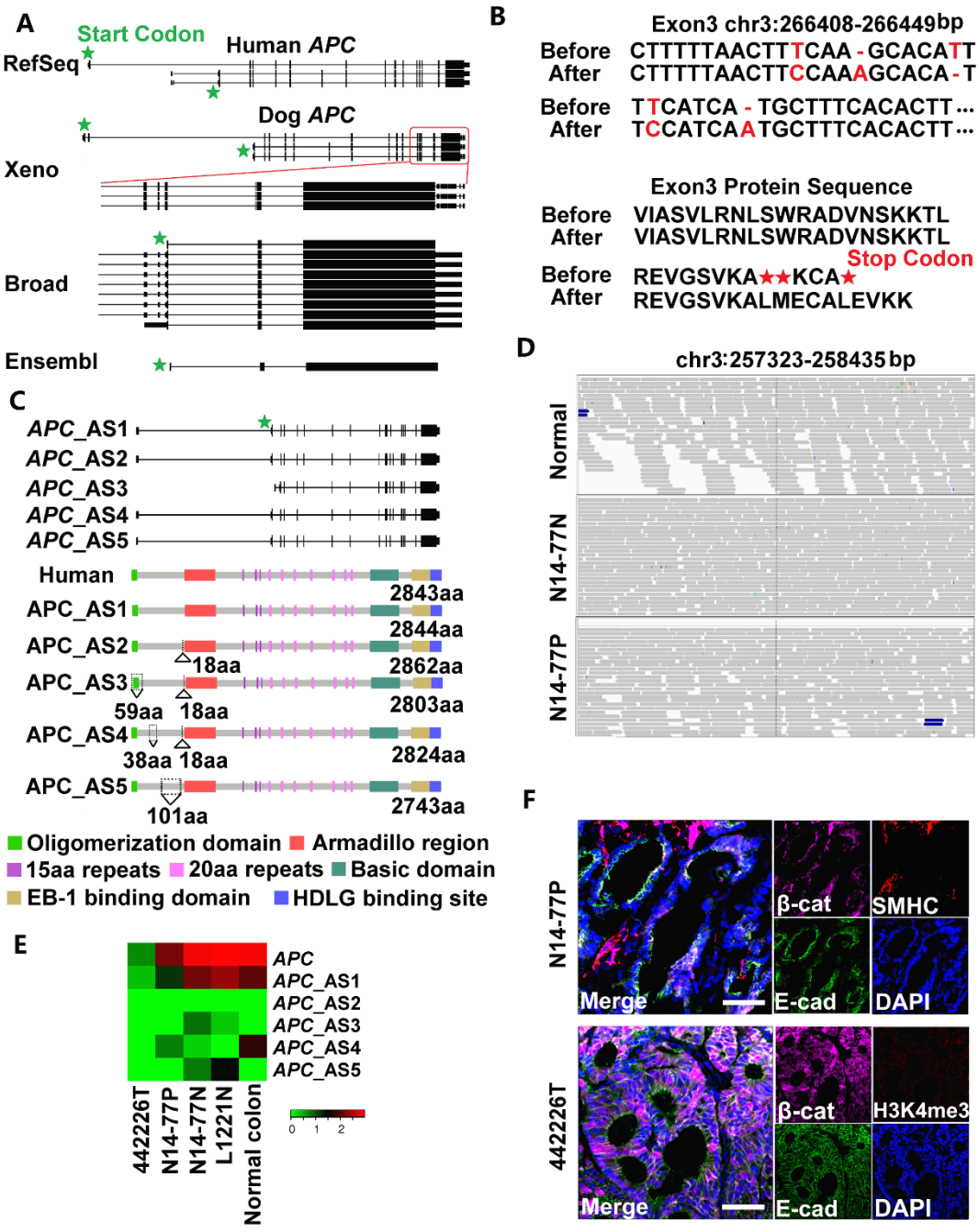


Figure 2.2: Neither germline nor somatic mutations of APC were found in N14-77.

(A) Canine APC is annotated inconsistently. "Xeno" represents human xenoRefGene (mapping human transcript or protein sequences to the canine genome). For Broad annotation [70], only the 3'–end portion enclosed by the red square is shown. Each line designates a transcript, with

coding exons, UTR exons and introns respectively represented by tall bars, short bars and the lines between the bars.

(B) Five sequence errors (red) were uncovered in exon 3 of APC in the canine genome assembly canFam3.1 (top), resulting in premature stop codons (bottom). Before and After: before and after error correction.

(C) Five alternatively spliced (AS) transcripts and protein isoforms were identified after the error correction in B. Also shown is the canonical isoform of human APC, with domains indicated. Amino acid (aa) insertions and deletions are indicated by Δ and ∇, respectively.

(D) Representative IGV images show no convincing mutations in APC. The canine genomic region shown corresponds to the human site (codons 1061-1431) that harbors some of APC mutation hotspots.

(E) The heatmap indicates the APC expression level in log2 (FPKM) (fragments per kilobase of exon per million fragments mapped). “L1221N” and “Normal colon” are normal colon epithelial tissues from two dogs, while “442226T” is a colorectal tumor with APC deletion from another dog.

(F) Representative confocal images indicate no nuclear enrichment of β-catenin (β-cat) in N14-77 polyps, unlike 442226T (with APC deletion). Scale bar, 50μm.

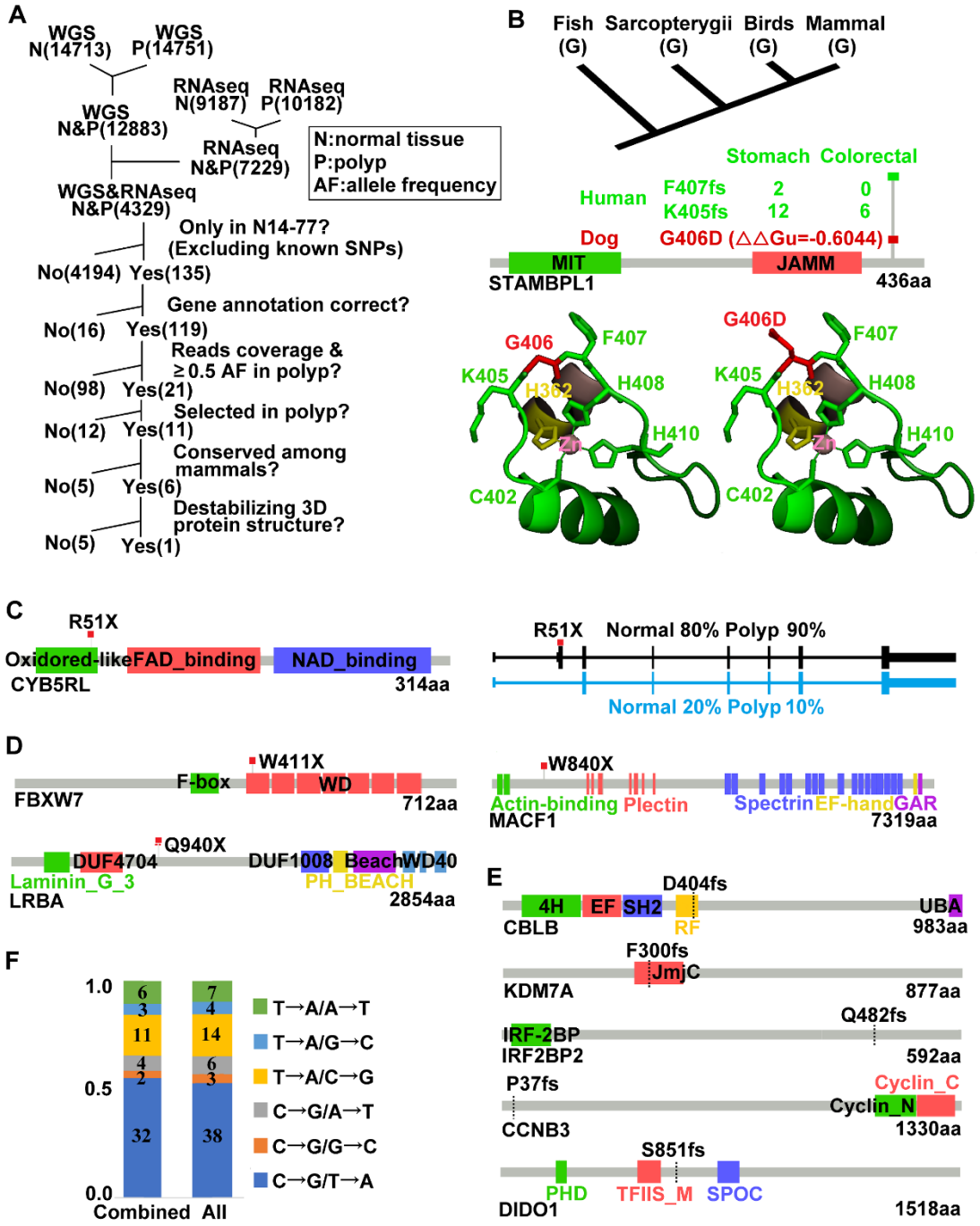


Figure 2.3: Notable germline and somatic mutations were identified in N14-77.

(A) Outlined is our pipeline for putative pathogenic germline missense mutation discovery (see text). Numbers in parentheses indicate mutation counts.

(B) G406D of STAMBPL1 is identified by the pipeline in A. The top image shows G406 conservation. Middle images (from bottom to top) indicate the protein domains, ΔΔGu (red) from modeling [72] predicting that G406D likely destabilizes the protein structure, and frameshift (fs) mutations found in human cancers (with total case numbers indicated). Bottom images are the crystal structure [73] with G406 (left) and G406D (right) in red and its flanking K405 and F407 in green, and zinc-coordinating residues C402, H362, H408 and H410 shown.

(C) R51X of CYB5RL is a heterozygous germline truncation mutation. The right image indicates the two alternative splicing forms and their proportions in each sample.

(D and **E)** Somatic truncation mutations and frameshift mutations identified.

(F) C→T/G→A changes dominate. Combined: somatic missense mutations identified by combining WGS and RNA-seq reads. All: those found by WGS alone, RNA-seq alone and combined. The numbers inside the bars specify the mutation counts.

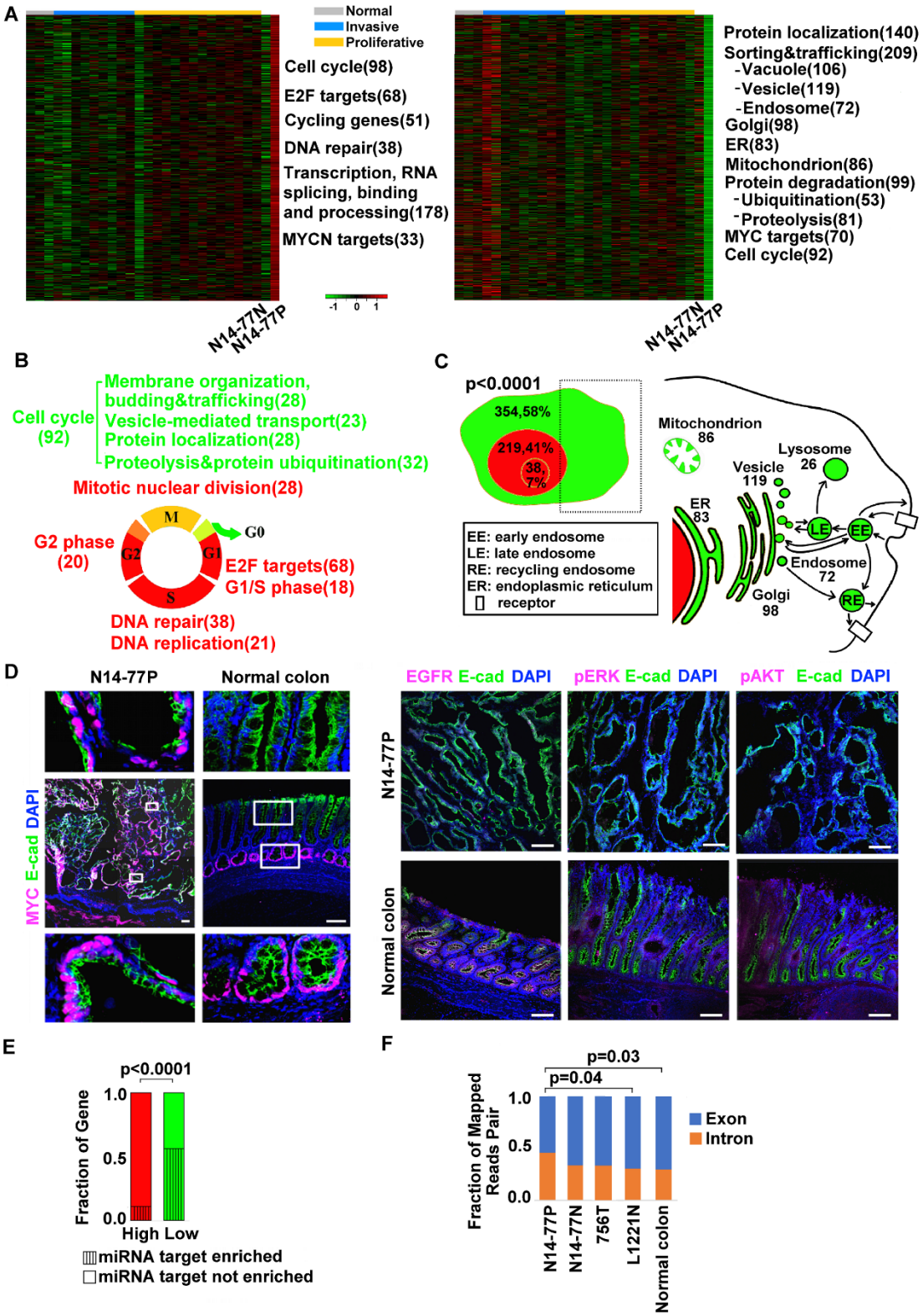

A
Normal
Invasive
Proliferative
Cell cycle(98)
E2F targets(68)
Cycling genes(51)
DNA repair(38)
Transcription, RNA splicing, binding and processing(178)
MYCN targets(33)
N14-77N
N14-77P
Protein localization(140)
Sorting&trafficking(209)
-Vacuole(106)
-Vesicle(119)
-Endosome(72)
Golgi(98)
ER(83)
Mitochondrion(86)
Protein degradation(99)
-Ubiquitination(53)
-Proteolysis(81)
MYC targets(70)
Cell cycle(92)
B
Cell cycle (92)
Membrane organization, budding&trafficking(28)
Vesicle-mediated transport(23)
Protein localization(28)
Proteolysis&protein ubiquitination(32)
Mitotic nuclear division(28)
G2 phase (20)
E2F targets(68)
G1/S phase(18)
DNA repair(38)
DNA replication(21)
C
p<0.0001
354,58%
219,41%
38, 7%
EE: early endosome
LE: late endosome
RE: recycling endosome
ER: endoplasmic reticulum
receptor
Mitochondrion 86
Lysosome 26
Vesicle 119
ER 83
Endosome 72
Golgi 98
D
N14-77P
Normal colon
MYC E-cad DAPI
EGFR E-cad DAPI
pERK E-cad DAPI
pAKT E-cad DAPI
E
p<0.0001
Fraction of Gene
High Low
miRNA target enriched
miRNA target not enriched
F
p=0.03
p=0.04
Fraction of Mapped Reads Pair
Exon
Intron
N14-77P
N14-77N
756T
L1221N
Normal colon

Figure 2.4: Highly- and lowly expressed genes in N14-77 polyps differ in function, cell cycle phase, cellular location and miRNA target site enrichment.

(A) Heatmaps from left to right indicate the log2 (FPKM) values of 474 highly- and 614 lowly expressed genes in 28 canine intestinal samples grouped as shown. Significantly enriched functions are listed next to the heatmap, with the total number of genes involved shown in the parenthesis.

(B) Red and green respectively designate highly and lowly expressed genes, along with their enriched functions and cell cycle phases. Yellow indicates cell cycle phases enriched in both highly and lowly expressed genes.

(C) Highly expressed genes are enriched in the nucleus (red) and lowly expressed genes are enriched in the cytoplasm (green), e.g., "354, 58%" indicating 354 genes, which make up 58% of all lowly expressed genes, located in the cytoplasm. The small red circle inside the nucleus designates the nucleolus. The right image illustrates that the sub cellular locations of lowly expressed genes, e.g., 83 genes are associated with ER.

(D) Representative IHC images illustrate the enrichment of nuclear protein MYC, and the depletion of membrane and cytoplasmic proteins EGFR, pERK and pAKT, in N14-77 polyps. Scale bar, 100 μm.

(E) Highly and lowly expressed genes differ in enriched miRNA target sites.

(F) More RNA-seq reads were mapped to intronic regions in N14-77 polyps. 756T is a canine jejunum tumor and others are described in Figure 2.2E.

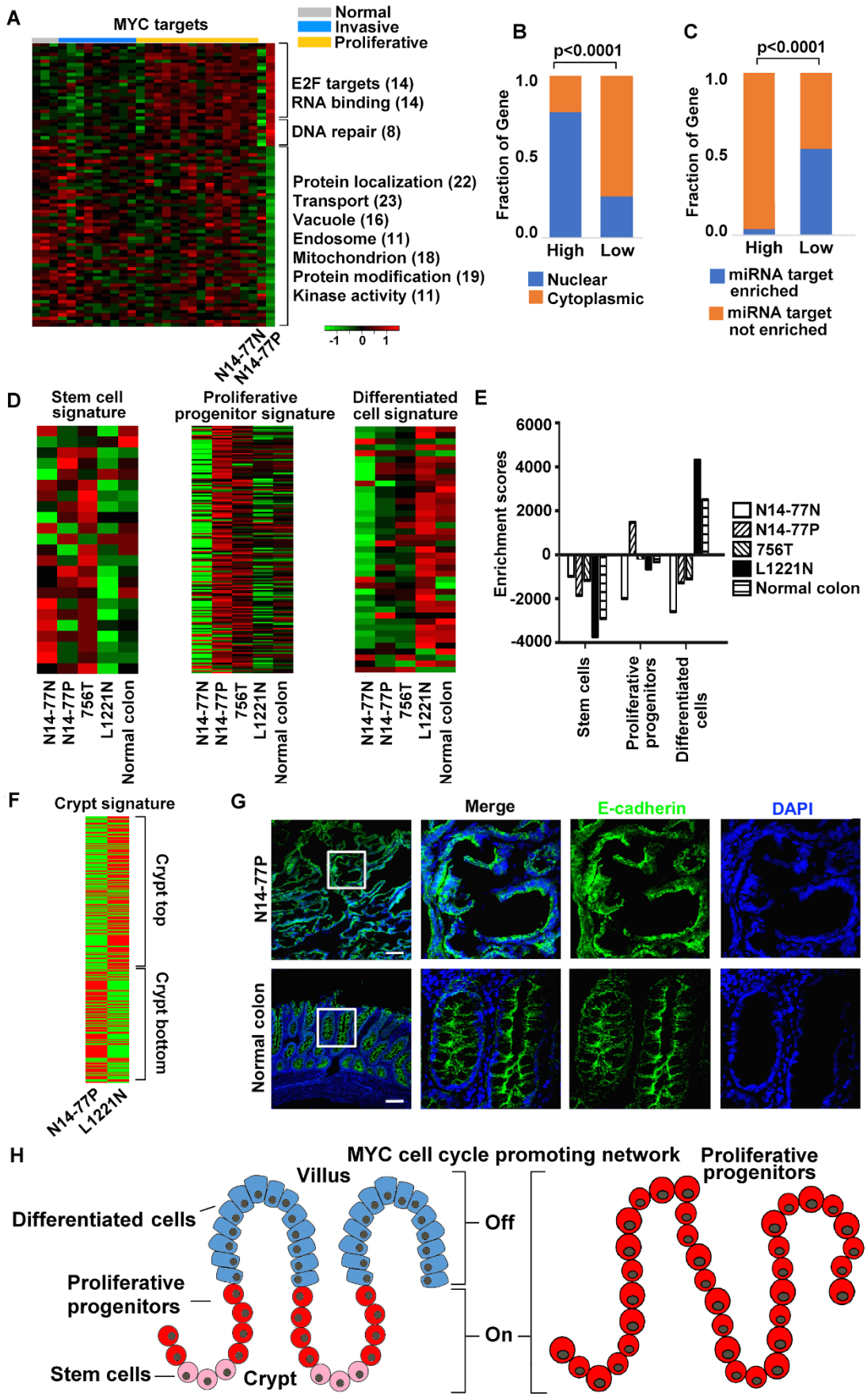

A
MYC targets
Normal
Invasive
Proliferative
E2F targets (14)
RNA binding (14)
DNA repair (8)
Protein localization (22)
Transport (23)
Vacuole (16)
Endosome (11)
Mitochondrion (18)
Protein modification (19)
Kinase activity (11)
N14-77N
N14-77P
-1 0 1
B
p<0.0001
Fraction of Gene
1.0
0.5
0.0
High Low
Nuclear
Cytoplasmic
C
p<0.0001
Fraction of Gene
1.0
0.5
0.0
High Low
miRNA target enriched
miRNA target not enriched
D
Stem cell signature
Proliferative progenitor signature
Differentiated cell signature
N14-77N
N14-77P
756T
L1221N
Normal colon
E
Enrichment scores
6000
4000
2000
0
-2000
-4000
N14-77N
N14-77P
756T
L1221N
Normal colon
Stem cells
Proliferative progenitors
Differentiated cells
F
Crypt signature
Crypt top
Crypt bottom
N14-77P
L1221N
G
Merge
E-cadherin
DAPI
N14-77P
Normal colon
H
MYC cell cycle promoting network
Proliferative progenitors
Villus
Differentiated cells
Proliferative progenitors
Stem cells
Crypt
Off
On

Figure 2.5: MYC network and crypt proliferative progenitor signature are activated in N14-77 polyps.

(A) MYC targets are enriched in both highly and lowly expressed genes. The image is presented as described for Figure 2.4A.

(B and **C)** Highly and lowly expressed MYC target genes differ in enriched cellular locations and miRNA target sites.

(D and **E)** Signature of intestinal proliferative progenitors, but not of either intestinal stem cells or differentiated cells, is activated in N14-77 polyps. The heatmaps indicate the log2 (FPKM values of signature genes [79] (D), and the bar plot indicate the corresponding ssGSEA results (E).

(F) Signature [80] of the crypt bottom, but not of the crypt top, is activated in N14-77 polyps.

(G) Representative IHC images indicate the lack of well-established epithelial apical-basolateral cell polarity in N14-77-polyp cells. Scale bar, 100 μm.

(H) Cartoons illustrate the differentiation of normal intestinal epithelium (left) and indicate that N14-77 polyp cells are in the proliferative progenitor state (right).

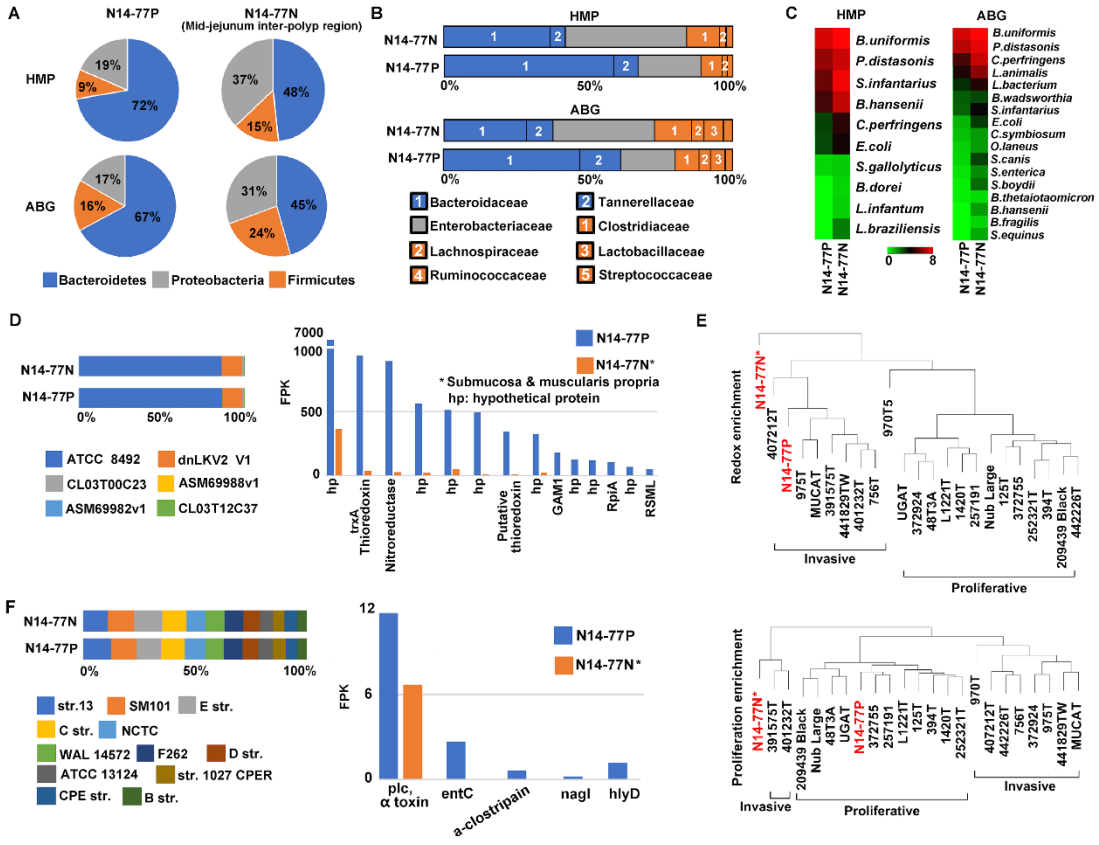


Figure 2.6: Bacteroidetes, B. uniformis and C. perfringens are significantly enriched in N14-77 intestinal microbiota.

(A) Bacteroidetes is the dominating phylum. The pie charts indicate the composition of bacterial phyla, determined by searching WGS reads against microbial genome databases HMP and ABG.

(B) Bacteriotecea is the dominant family in polyps. The colors represent bacterial phyla as shown in A.

(C) *B. uniformis* and *C. perfringens* are among the top enriched species.

(D) ATCC 8492 is the most enriched strain of *B. uniformis* (left), determined with WGS reads, and expresses abundantly thioredoxin and nitroreductase genes (right), determined with RNA-seq reads.

(E) N14-77 polyps resemble invasive tumors, but not proliferative tumors, in redox gene expression. The images indicate sample clustering based on the ssGSEA enrichment scores with indicated gene sets.

(F) *C. perfringens* strain enrichment and toxin gene expression.

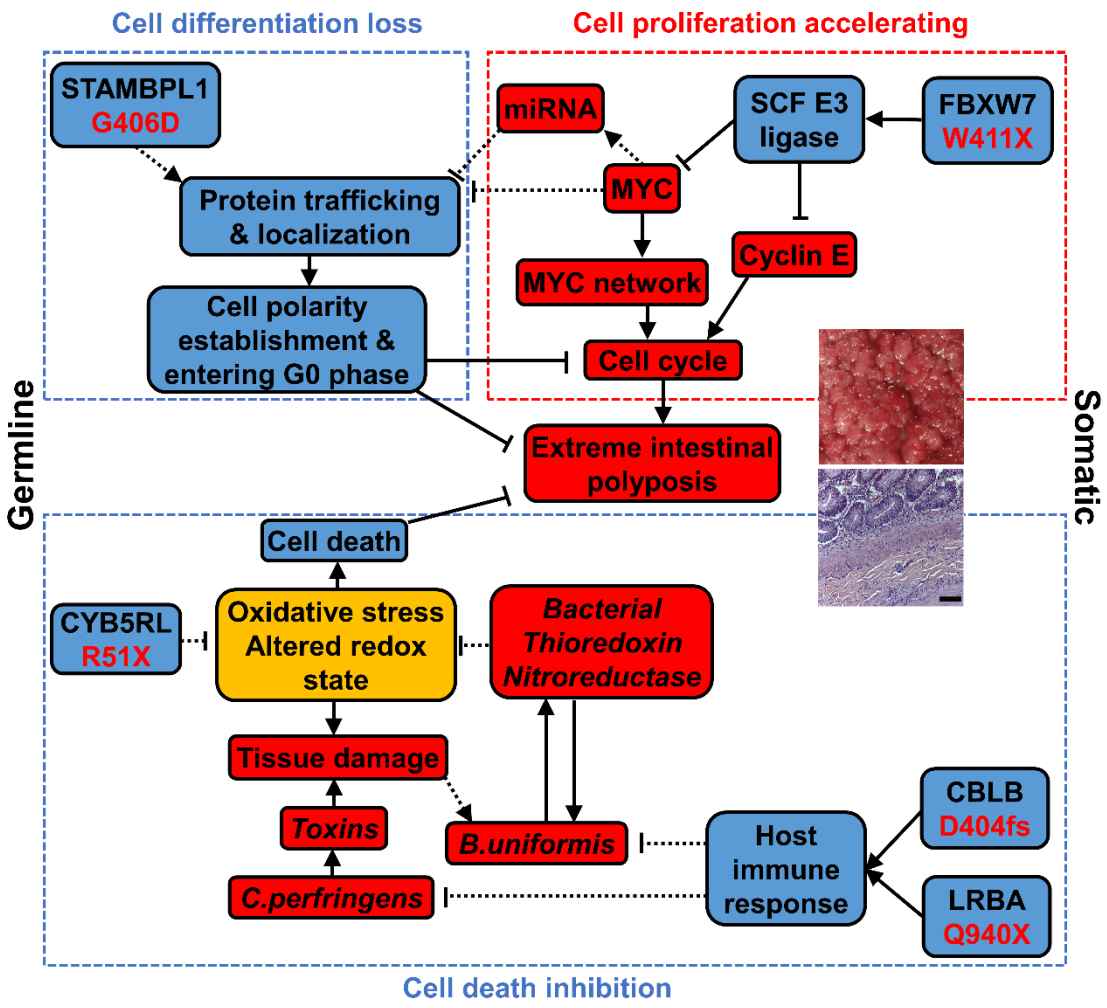


Figure 2.7: An alteration network collaborating across the genome, transcriptome and microbiome contributes to extreme intestinal polyposis of N14-77.

Bacterial elements are shown in italics. Red, blue, and yellow designate promoting, inhibiting, and both promoting and inhibiting factors, respectively. "→": promoting; "⊣": inhibiting. Solid lines indicate that the relationship is supported by published studies and our observations, while dashed lines indicate that the relationship requires future functional validation.

CHAPTER 3

PROLIFERATIVE AND INVASIVE COLORECTAL TUMORS IN PET DOGS PROVIDE UNIQUE INSIGHTS INTO HUMAN COLORECTAL CANCER[1]

[1]Wang, Jin, et al. "Proliferative and Invasive Colorectal Tumors in Pet Dogs Provide Unique Insights into Human Colorectal Cancer." Cancers 10.9 (2018): 330.
Reprinted here with the permission of the publisher.

ABSTRACT

Spontaneous tumors in pet dogs represent a valuable but undercharacterized cancer model. To better use this resource, we performed an initial global comparison between proliferative and invasive colorectal tumors from 20 canine cases and evaluated their molecular homology to human colorectal cancer (CRC). First, proliferative canine tumors harbor overactivated WNT/β-catenin pathways and recurrent CTNNB1 (β-catenin) mutations S45F/P, D32Y and G34E. Invasive canine tumors harbor prominent fibroblast proliferation and overactivated stroma. Both groups have recurrent TP53 mutations. We observed three invasion patterns in canine tumors: collective, crypt-like and epithelial–mesenchymal transition (EMT). We detected enriched *Helicobacter bilis* and *Alistipes finegoldii* in proliferative and crypt-like tumors, but depleted mucosa-microbes in the EMT tumor. Second, guided by our canine findings, we classified 79% of 478 human colon cancers from The Cancer Genome Atlas into four subtypes: primarily proliferative, or with collective, crypt-like or EMT invasion features. Their molecular characteristics match those of canine tumors. We showed that consensus molecular subtype 4 (mesenchymal) of human CRC should be further divided into EMT and crypt-like subtypes, which differ in TGF-β activation and mucosa-microbe content. Our canine tumors share the same pathogenic pathway as human CRCs. Dog-human integration identifies three CRC invasion patterns and improves CRC subtyping.

INTRODUCTION

Spontaneous canine cancers represent one of the best animal models of human cancers [11,47,51,53-56,103-106]. Being naturally occurring, heterogeneous and with an intact immune system, they capture the essence of human cancer, unlike most genetically modified or xenograft rodent models. Furthermore, dogs better resemble humans in biology, e.g., similar telomere and telomerase activities [107] and more frequent spontaneous cancers of epithelial origin [54-56,103,106], unlike mice [108]. Notably, dogs share the same environment as humans and hence are exposed to the same carcinogens. Indeed, similar risk factors for cancer development (advancing age, obesity, diet, etc.) and numerous dog-human anatomic/clinical homologies for the same type of cancer have been noted [53,103,104,106].

Unlike human cancers where hundreds of thousands of cancer cases have been characterized with genome-wide approaches [65,74,109-114], far fewer canine cancers have been studied. As a result, we have a limited molecular understanding of canine cancers, which makes this immensely valuable resource significantly understudied and underused.

With 140,250 new cases and 50,630 deaths estimated in 2018 [115], colorectal cancer (CRC) is the third most common cancer in the US. Thus, to better understand and treat CRC is important. We have previously characterized copy number abnormalities (CNAs) in canine CRC genomes [54], which supports the dog-human molecular homology. Furthermore, we have successfully developed a novel dog-human comparison strategy for cancer driver-passenger discrimination for amplified/deleted genes [11,60].

To further understand colorectal carcinogenesis mechanisms in pet dogs and their homology/difference with their human counterparts, we set out to investigate gene expression alteration, mutations, and microbiota changes of intestinal tumors from 22 pet dogs, as described below.

RESULTS

RNA-Seq Analysis Clusters the Tumors into Two Major Groups

We performed RNA-seq on 26 intestinal samples collected from dogs with spontaneous tumors in the large intestine (20 dogs) and the small intestine (two dogs), and without any intestinal tumors detected (one dog) (Table S3.1). Among the samples, 23 are tumors consisting of colorectal adenomas from four dogs, adenocarcinomas (12 colorectal and one each for duodenum and jejunum) from 17 dogs, and two colonic stromal tumors from one dog (Table S3.1). Three samples are normal colonic epithelial tissues from two canine patients described above and one normal dog. Histologically, the 4 adenomas and 17 adenocarcinomas can be largely classified into two groups: highly proliferative (4 adenomas and 9 adenocarcinomas) or highly invasive (8 tumors). Highly proliferative tumors are characterized by prominent proliferation of epithelial cells that are clearly marked by E-cadherin staining (Figure 3.1A). Highly invasive tumors are characterized by: (1) the spread of tumor cells into submucosa and muscle layers of the intestine; and (2) the lack of prominent proliferation of clearly marked epithelial cells (Figure 3.1A).

Some of the tumors clearly have more stromal cell content (Figure 3.1A). To reduce variations, we maximally dissected away stromal regions without any tumor or epithelial cells, and only used sections enriched in tumor or epithelial cells for RNA-seq and other genomic analyses. We performed RNA-seq analysis with digested tissues of the 26 samples (Table S3.1A).

We then conducted non-negative matrix factorization (NMF) [116] clustering analysis with 10,618 total genes that are expressed in at least one sample (Table S3.1B). The analysis identified four metagene sets and four sample clusters (Figure 3.1B). First, the two stromal tumors form one NMF cluster, and the three normal samples and one tumor constitute another (Figure 3.1B). The remaining two clusters nicely separate highly proliferative tumors from highly invasive tumors: one cluster consisting of 12 (out of 13 total) proliferative tumors, while the other containing 7 (out of 8 total) invasive tumors and one proliferative tumor (Figure 3.1B). Thus, the results are consistent with the histopathological classification as illustrated in Figure 3.1A. Metagenes identified are also informative. Metagenes of the normal group are significantly enriched in functions that characterize differentiated colon epithelial cells (Figure 3.1B, Table S3.1C). These include β-catenin-downregulated targets, APC-upregulated targets, epithelial cell polarity and others. The opposite was noted for metagenes of the proliferative group. Metagenes of the invasive group are significantly enriched in features of cell invasion, e.g., extracellular matrix, etc. (Figure 3.1B).

NMF clusters are supported by unsupervised hierarchical clustering with various numbers of top most variable genes in expression across the 26 samples. The analysis consistently separates invasive tumors from proliferative tumors (Figure S3.1). The same is achieved with the principle component analysis (PCA) with the entire transcriptome.

In summary, histopathological and three gene expression clustering strategies have consistently classified the tumors into two major groups, highly proliferative or highly invasive. Below are our molecular characterizations of each group.

Canonical CRC Pathways Are Activated in Proliferative Tumors

To better understand the difference between proliferative and invasive tumors identified in Figure 3.1, we performed gene set enrichment analysis (GSEA) with signature gene groups used in subtyping [10,117-120] [121]and characterization [79,80,122-124] of human CRCs. These include 17 canonical CRC signatures, 11 cancer pathways, 9 stromal signatures, 33 specific immune processes and 14 specific metabolic processes, totaling to 3881 genes (Table S3.2A, B). The analysis reveals that activation of WNT/β-catenin signaling → cell cycle and proliferation is the most significant feature of proliferative tumors (Figure S3.2A, Table S3.2A). Indeed, intestinal WNT/β-catenin/TCF signature [123], crypt proliferation signature [79] and cell cycle activation signature [10] are all significantly ($p < 0.05$) enriched in proliferative tumors (Figure 3.2A). Although not as significant, MYC targets are also upregulated in proliferative tumors (Figure 3.2A). These results are consistent with single sample GSEA (ssGSEA) (Figure 3.2B). In

summary, these canine proliferative tumors share similar molecular features as canonical CRCs in humans [10,65,74,109,111,112].

CTNNB1 and TGF-β Signaling Genes Were Recurrently Mutated in Proliferative Tumors

To understand the mechanisms underlying the observed WNT/β-catenin → cell proliferation activation (Figure 3.2A, B), we performed Whole Genome Sequencing (WGS) for 15 tumor and matching normal samples of 10 canine colorectal tumor cases (Table S3.1). We then combined WGS and RNA-seq data (Table S3.1A), which significantly increases the sequence coverage for mutation finding.

CTNNB1, which encodes β-catenin, is the most noteworthy. It is mutated in 7 (50%) proliferative tumors. Importantly, the mutations are S45P, S45F, G34E and D32Y (Figure 3.2C, Table S3.2C). All locate in the N-terminal peptide of D(32)S(33)G(34)IHSGATTTAPS(45)LS of β-catenin, where phosphorylation of the Ser/Thr residues initiates β-catenin degradation [125]. These mutations are likely gain of function and colorectal tumorigenesis drivers. First, S45P/F mutations would prevent S45 phosphorylation and hinder β-catenin degradation. Indeed, our IHC experiment reveals increased accumulation of β-catenin inside tumor cells that harbor S45F mutation (Figure 3.2D). Furthermore, these tumor cells also express substantially more MYC protein in their nucleus (Figure 3.2D).

Following N-terminal Ser/Thr phosphorylation, β-catenin is targeted for ubiquitination and degradation. D32Y and G34E mutations are likely to affect this process, as "D(32)pS(33)G(34)IHpS" marks the "DpSGφXpS" destruction motif [126]. To better understand

this, we studied the crystal structure of human β-TrCP1/Skp1/β-catenin [126], an E3 ligase complex that ubiquitinates β-catenin (note that except for a T60S change, canine β-catenin is identical to human β-catenin; see Figure S3.2B). The N-terminal phosphorylated peptide of β-catenin binds β-TrCP1 via hydrogen bonds and electrostatic interactions [126] (Figure 3.2E and Figure S3.2C), some of which would be disrupted by the D32Y mutation. Likewise, G34 locates in a positively charged environment [126] (Figure 3.2E), and the G34E mutation would change the electrostatic interaction. Indeed, our substrate docking modeling indicates that both mutations alter the binding of β-catenin to β-TrCP1 (Figure 3.2F). Our IHC experiment reveals substantial accumulation of β-catenin and MYC in tumor cells with D32Y and G34E mutations (Figure 3.2D).

Even though we did not find notable APC mutations, APC is recurrently downregulated, especially in proliferative tumors (Table S3.2D). Its lowest expression level was observed in a proliferative tumor (Figure 3.2C) that harbors neither CTNNB1 mutations nor mutations described below.

Besides CTNNB1, we also uncovered mutations in ACVR2A and ACVR1B, which encode receptors of activin, a member of the TGF-β superfamily, currently in proliferative tumors (Figure 3.2C).

Cancer-Associated Fibroblast (CAF) and Stromal Signatures Are Activated in Invasive Tumors

Besides canonical CRC pathways (Figure 3.2A), we also investigated tumor microenvironment. We found that stromal signatures derived from human CRC [122] are activated in canine invasive tumors, compared to proliferative tumors. Specifically, CAF and endothelial cell signatures are significantly enriched, while the leukocyte signature is not (Figure 3.3A, B and Table S3.2A, B). Interestingly, similar conclusions were reached with stromal signatures derived from single cell RNA-seq analysis of human melanoma [127]. CAF, macrophage and endothelial cell signatures are enriched in invasive tumors, whereas signatures of B-cells and T-cells are not (Figure S3.3A, B). Lastly, consistent with CAF signature enrichment, fibroblast activation markers are upregulated in invasive tumors (Figure 3.3A, Table S3.3).

Epithelial–mesenchymal transition (EMT) has been extensively studied in human CRC. To investigate EMT in these canine tumors, we examined its signatures from human CRC [117,124]. As expected, the epithelial signature is significantly enriched in proliferative tumors, whereas the mesenchymal signature is significantly enriched in invasive tumors (Figure 3.3A, B). The EMT activation signature is upregulated in invasive tumors, albeit not as significantly (Figure 3.3A, B).

Three Modes of Cancer Cell Invasion Were Observed

Both EMT and CAF signatures are upregulated in invasive tumors (Figure 3.3A, B). To better understand this, we performed IHC experiments with vimentin, a mesenchymal cell marker frequently used for fibroblast and CAF identification, as well as E-cadherin, an epithelial marker. The result supports the gene signature analysis shown in Figure 3.3A, B. Fibroblast proliferation is clearly more prominent in invasive tumors than in proliferative tumors, while no difference was found for pSTAT3 (Figure 3.3C), a marker often associated with immune response.

Importantly, the IHC study reveals three modes of tumor cell invasion: collective, crypt-like and EMT (Figure 3.3C). Collective and EMT invasions are both well studied in human cancers [128,129]. We observed collective invasion in canine proliferative tumors, with masses that consist of predominantly epithelial cells, with far fewer fibroblasts, found in submucosa and muscularis layers of the colon (Figure 3.3C). We also observed EMT in canine invasive tumors, with numerous tumor cells expressing E-cadherin and vimentin simultaneously (Figure 3.3C).

"Crypt-like" is another invasion mode frequently observed in our canine invasive tumors. In this mode, crypt-like structures, consisting of a monolayer of epithelial cells that are surrounded by densely populated and multilayered fibroblasts, were found in submucosa and muscularis layers of the colon (Figure 3.3C and Figure S3.3C). Crypt-like invasion differs from collective invasion in: (1) no significant epithelial cell proliferation (monolayer versus multilayer); and (2)

very prominent fibroblast proliferation. Crypt-like invasion also differs from EMT invasion, as epithelial cells and mesenchymal cells are easily distinguishable.

Crypt-like invasion is not as extensively reported as collective or EMT invasion; we hence use an invasive cancer, 407212T, as an example for further illustration. Tumor cells of 407212T have penetrated through the colon and have likely metastasized to the lung (Table S3.1). Our IHC staining reveals clear tracks of crypt-like structures (Figure 3.3C and Figure S3.3C), as if tumor cells have been walking through the colon. Crypt-like structures vary considerably in size. Each has a monolayer of epithelial cells with distinct cell-cell junction, as indicated by E-cadherin and β-catenin staining (Figure 3.3C), resembling colonic crypts. However, they also differ from normal crypts. First, not just surrounded by extensive fibroblasts, many crypt-like structures harbor fibroblasts inside their lumen (Figure 3.3C). Second, their epithelial cells are nearly all MYC-positive, matching crypt stem cells or progenitors but not fully differentiated cells (Figure 3.3C). This is supported by their activated signatures of hypoxia and cellular response to oxidative stress (Figure S3.3D). Thus, these crypt-like structures consist of cells with colon stem cell or progenitor features.

Crypt-Like Invasion Tumor Harbors Mucosa-Like Microbiome

Gut microbiome has gained increasing attention in human CRC research. For an initial understanding of the microbiomes of our canine tumors, we searched for microbial sequences in their WGS data (Table S3.1A), as previously described [106]. As expected, canine colorectal samples contain >100-fold more bacterial sequences than skin samples (Figure 3.4A–C, Table

S3.4A–C). Importantly, these colorectal samples are enriched in three bacterial phyla: bacteroidetes, proteobacteria and firmicutes (Figure 3.4A, Table S3.4A). This is supported at the family level, where the top enriched families include bacteroidaceae, enterobacteriaceae, rikenellaceae and helicobacteraceae (Figure 3.4B, Table S3.4B). At the species level, top abundant bacteria also belong to these three phyla, although the actual species vary in each sample (Figure 3.4C, Table S3.4C). One difference between our findings and published human and canine colon microbiota data [82,130] is that proteobacteria, but not fusobacteria, is among the top 3 most enriched phyla.

Tumor 407212T, which exemplifies crypt-like invasion (Figure 3.3C), is especially noteworthy. Although located in the muscularis layers of the colon and distant from the mucosa, this tumor harbors a microbiome with enrichment and diversity values as high as those of mucosa samples, including normal tissues and proliferative tumors (Figure 3.4D, Table S3.4D). One species, *Alistipes finegoldii*, a commensal gut microbe and belonging to the phylum of bacteroidetes, is abnormally enriched (Figure 3.4C). Please note that *A. finegoldii* has been detected in blood samples of human CRC patients [131]. On the contrary, the EMT tumor (391575T; see Figure 3.3C) is significantly depleted in bacteria (Figure 3.4D).

Helicobacter bilis has been linked to inflammatory bowel disease (IBD) and CRC in mouse models [132]. We noted that *H. bilis* is significantly enriched in a proliferative tumor (372755T) (Figure 3.4C). Among its strains examined, ATCC43879 is >16-fold more enriched than others, with its top expressed genes encoding flagellin A and others (Figure 3.4E, Table S3.4E).

TP53 Is Recurrently Altered in Both Proliferative and Invasive Tumors

Unlike CTNNB1 and ACVR2A/1B (Figure 3.2C), we detected TP53 mutations (whole gene deletion, indels and missense mutations) in both proliferative and invasive tumors (Figure 3.5A, D and Table S3.5A). Missense mutations identified are all located in the DNA binding domain and are also common in human cancer. For example, through protein alignment (Figure 3.5B, Table S3.5B), canine R162H and R261C/H are equivalent to human R175H and R273C/H, respectively. Both are among the top three most frequent TP53 mutations in human CRC (Figure S3.4) and are known cancer drivers [133].

Finally, we observed intron 6 retention in a fraction of TP53 transcripts in both proliferative and invasive tumors (Figure 3.5C, D). Intron 6-retention will create two stop codons within the TP53 DNA binding domain (Figure 3.5A, C).

We Identified Three Types of Invasion in Human Colon Cancers

To further evaluate the dog-human molecular homology, we tried to identify the four molecular subtypes illustrated in Figure 3.3C (i.e., proliferative and three types of invasion: collective, crypt-like and EMT) among the 478 human colon cancers from The Cancer Genome Atlas (TCGA) [74]. Guided by our canine findings (Figure S3.5A), we studied the distribution and clustering of ssGSEA enrichment scores of CRC signatures of: (1) proliferation [79]; (2) EMT (epithelial, mesenchymal, and EMT activation) [117,124]; (3) CAF and stroma [122]; and (4) central tumor and invasive front [121] (Figure S3.5B). We also included developmental signatures on: (1) colonic stem cells, progenitors and differentiated cells [79]; and (2) colon crypt

and top [80]. We identified 74 proliferative tumors, 159 tumors of collective invasion, 79 tumors of crypt-like invasion, and 67 tumors of EMT invasion (Table S3.6A). These total 379 tumors, accounting for 79% of all TCGA colon cancers examined. Proliferative tumors show the largest differences from other tumors in all CRC signatures examined except for epithelial signature (Figure 3.6A). Among the three invasive subtypes, collective invasion displays more suppressed signatures related to stroma (mesenchymal, EMT activation, CAF, stromal and invasive front) (Figure 3.6A). Finally, crypt-like and EMT invasions are similar, except that the former has more activated signatures of central tumor and cell proliferation (Figure 3.6A).

The four subtypes differ significantly in several aspects in canonical CRC pathway alterations and gene mutations. First, proliferative and collective invasion subtypes both harbors more activated WNT pathway and MYC targets, but more suppressed PI3K/AKT signaling (Figure 3.6B, Table S3.6B). Second, crypt-like invasion and EMT invasion both have a higher mutation rate of TP53 (Figure 3.6C and Figure S3.5C, Table S3.6C). Yet, TP53 signaling is enhanced in crypt-like invasion (Figure 3.6B). Third, EMT invasion harbors the most activated TGF-β signaling and overall the fewest mutations in relevant genes, a clear difference from other subtypes (Figure 3.6B, C and Figure S3.5C, Table S3.6B, C).

We also investigated the difference in microbiome among the four subtypes. First, we identified WGS data from TCGA that are available to 51 proliferative tumors, 98 tumors of collective invasion, 29 tumors of crypt-like invasion and 31 tumors of EMT invasion (Table S3.6D). Then, we performed the same analysis as described for canine tumors (Figure 3.5). We

noted that crypt-like invasion tumors have similar or even higher bacterial enrichment and diversity, when compared to proliferative and collective invasion tumors (Figure 3.6D, Table S3.6D). EMT invasion tumors, however, consistently harbor fewer bacteria (Figure 3.6D, Table S3.6D). The observations agree with our canine findings (Figure 3.4D).

We Classified Consensus Molecular Subtype 4 (CMS4) into Crypt-Like and EMT Invasions

We examined the relationship between our subtypes and the four CRC consensus molecular subtypes (CMSs) from a well-cited study [10]. A total of 419 TCGA colon cancers were investigated by both methods (Table S3.6E). We noted a significant overlap (>50%) between our collective invasion and CMS1 (Figure 3.6E, Table S3.6E). CMS1 also harbors smaller fractions of proliferative, crypt-like and EMT subtypes of ours. CMS1 is characterized by hypermutation, microsatellite instability, and strong immune activation [10]. We also observed a significant overlap between our proliferative subtype and CMS3 (Figure 3.6E, Table S3.6E). CMS3 also contains collective and unclassified colon cancers by us (Figure S3.5, Table S3.6E). CMS3 is epithelial and has evident metabolic dysregulation [10]. CMS2 is also epithelial and is characterized with WNT and MYC signaling activation (thus the canonical subtype) [10]. It consists of our proliferative and collective invasion subtypes and cancers that are not classified by us (Figure 3.6E, Table S3.6E), none of which is enriched. The most interesting finding, however, is that our crypt-like and EMT invasion subtypes are both highly enriched in CMS4, accounting for 87% of all CMS4 tumors (Figure 3.6E, Table S3.6E). CMS4, being mesenchymal and with stromal invasion, is featured with prominent TGF-β activation [10]. Yet, our study

further classified CMS4 into EMT invasion and crypt-like invasion, with TGF-β activation found only in the EMT invasion subtype (Figure 3.6B).

DISCUSSION

Canine Colorectal Tumors Follow Canonical Pathogenic Pathways of Human CRC

Alteration of WNT signaling pathway [134], observed in >90% human CRCs [74], leads to MYC activation, cell proliferation and ultimately tumorigenesis [74]. We have reached the same conclusion for proliferative colorectal tumors in dogs. One interesting difference lies in CTNNB1, which is mutated in <10% of human CRCs [74] but in >60% of our canine proliferative tumors. Please note that CTNNB1 mutations detected in our canine tumors are S45P/F, D32Y and G34D, which interfere with β-catenin ubiquitination and degradation, yielding the same outcome as APC mutation. Intriguingly, we did not find frequent APC mutation in these canine samples, unlike human CRC [74], although we noted recurrent downregulation of APC. We do not know if this is related to the local genomic environment of APC. While canine APC locates at the chromosome end (near heterochromatin), human APC lies in the middle of chromosome 5 (euchromatin). Future study with a larger sample size is clearly required to answer the question. We nonetheless emphasize that whether it is APC mutation or CTNNB1 mutation, the outcome remains the same-activation of WNT signaling.

Alteration of TGF-β signaling pathway also leads to MYC activation and cell proliferation in human CRC [74]. Analogous to human CRC [74], we found recurrent mutation in TGF-β signaling genes ACVR2A and ACVR1B in our canine proliferative tumors.

Alteration of TP53 pathway occurs in more than half of human CRCs [74]. Comparable to this, TP53 is recurrently mutated in our canine tumors. Moreover, most mutations detected have been reported in human CRC, with some already classified as drivers [133]. Please note that TP53 mutations are found in both proliferative and invasive canine tumors, unlike CTNNB1. This is consistent with the Vogelstein model that places TP53 mutation at a later carcinogenesis stage of human CRC [57]. Lastly, we have detected a stop-codon-creating intron-retention in canine tumors. More studies are needed to determine if a truncated TP53 protein is indeed generated and, more importantly, how this has happened. For example, is it due to mis-splicing, and/or because nonsense-mediated mRNA decay is off or dysfunctional?

We Have Detected Three Invasion Modes of Canine Cancer Cells

Microenvironment is important in cancer development and invasion [135,136]. Stromal signatures reported for human CRC [122] are activated in our invasive canine tumors, supporting the dog-human molecular homology. Importantly, we have detected three modes of cancer cell invasion in our canine tumors: collective, crypt-like and EMT. Collective and EMT invasions are both well studied in human cancers [128,129]. Collective invasion is largely defined as migration of a group of cells while maintaining cell-cell contacts. These cells are often epithelial in nature and thus can be readily distinguished from the microenvironment. This is unlike EMT invasion, where many cancer cells have acquired stromal cell features.

To our knowledge, crypt-like invasion, where cancer cells spread via crypt-like structures, is not as extensively reported as collective or EMT invasion. Our study indicates that these cancer

cells are MYC-positive, resembling crypt stem cells or progenitors. We propose that they are capable of crypt development in non-mucosa locations because of prominent fibroblast proliferation, which has remodeled the microenvironment to be more mucosa-like (supported by their microbiota that resembles mucosa samples). Whether this is true and how this occurs of course need more research. For example, the origin of the proliferating fibroblasts is unclear. Are they derived from some types of crypt mesenchymal stem cells that migrate with the cancer cells? Or are they local?

Human CMS4 Colon Cancers Consist of Crypt-Like and EMT Invasion Subtypes that Differ in TGF-β Signaling

Most human CRCs can be classified as one of the four consensus molecular subtypes (CMS1, CMS2, CMS3 and CMS4), each with distinct molecular features [10]. CMS4 is the "mesenchymal" subtype, characterized with TGF-β activation, stromal invasion and angiogenesis [10]. Our analysis indicates that CMS4 actually consists of two subtypes, EMT and crypt-like invasion. Although EMT and crypt-like invasions are indeed very similar molecularly, our analysis reveals a few differences. First, only EMT invasion harbors TGF-β activation, likely due to less frequent mutation of TGF-β signaling genes. Crypt-like invasion, meanwhile, displays more activated signature of central tumor and, as discussed previously, may harbor some types of stem cells [65]. We plan to validate this finding using a large sample size, including rectum cancers, in the future. In addition, we plan to include more signatures and

parameters, including the consensus Immunoscore calculated based on the density of CD3+ and CD8+ T-cells within central tumor and invasive front from a recent publication [137].

Microbiome could represent another difference. In crypt-like invasion, the tumors appear to retain the mucosa microbiota after spreading to foreign locations. In EMT invasion, however, the tumors seem to have lost the mucosa microbiota. A recent publication [138] reports that *Fusobacterium nucleatum* and other microorganisms of human colorectal tumors are retained in metastatic sites, and that antibiotic treatment inhibits tumor growth in mouse models. Thus, it would be useful to perform deeper microbiome comparison between EMT and crypt-like invasions, including metagenomics data from stool samples.

Although more studies are needed, our findings shed more light on the molecular mechanisms of human CRC invasion. Importantly, because of the molecular differences, different treatment may be considered between EMT and crypt-like invasion subtypes. For example, a recent publication has elegantly shown that the efficacy of the PD-1/PD-L1 blockade therapy of several cancers is influenced by gut microbiome [139].

Dog-Human Comparison Could Be Effective for Driver-Passenger Discrimination for Missense Mutations

Driver-passenger discrimination has always been a central aim of cancer research. We have previously shown that our human-dog comparative genomics and oncology strategy is effective for driver-passenger discrimination for amplified/deleted genes in CRCs [11,54]. Our work here also indicates the potential of this approach on missense mutations. Indeed, known and putative

drivers of CTNNB1 and TP53 are among the most frequent missense mutations detected in our canine tumors. The comparison can be expanded to numerous other genes that harbor one or multiple missense mutations, once the corresponding amino acid residues between the dog and human proteins are established.

Stromal drivers and microbial drivers are harder to identify, with fewer efficient approaches available. Our discovery of prominent fibroblast proliferation in canine invasive tumors, as well as significant enrichment of *H. bilis* and *A. finegoldii* in canine tumors may open a new avenue to address these important but difficult questions. Indeed, fibroblasts are known to play an important role in human CRC and other cancers [122,135,140], *H. bilis* is linked to human IBD and CRC [132], and *A. finegoldii* is detected in blood samples of human CRC patients [131].

Lastly, we acknowledge our current canine sample size is small. Because of the vast heterogeneity, a much larger sample size is required for efficient driver-passenger discrimination via dog-human comparison. Also note that our current study has relied on WGS of tumor samples for microbiome analysis, which may fail to detect less abundant bacterial species. Metagenomics data from stool samples should also be examined.

MATERIALS AND METHODS

Canine Samples

Fresh-frozen (FF) canine tissues and spontaneous tumors were obtained from various veterinary colleges (Table S3.1). Samples were collected from client-owned dogs that develop the disease spontaneously, under the guidelines of the Institutional Animal Care and Use

Committee for use of residual diagnostic specimens and with owner informed consent. The breed, age, histopathologic description and other information are provided in Table S3.1. The research received the ethical approval from the Institutional Animal Care and Use Committee (A2017 01-025-R1, approved on 8 February 2018 for University of Georgia; 2010A0015-R2, approved on 2 December 2017 for Ohio State University; and 16-6532A, approved on 22 March 2018 for Colorado State University).

Tissue Dissection, DNA and RNA Extraction, and Quality Control

Cryosectioning of FF tissues, H&E staining and cryomicrodissection were performed as described [54,56,106] to enrich tumor cells for tumor samples, and unaffected/normal epithelial cells for control/normal samples. Genomic DNA and RNA were extracted from the dissected tissues using the AllPrep DNA/RNA Mini Kit (cat. no. 80204) from QIAGEN (Germantown, MD, USA). Only samples with a 260/280 ratio of ~1.8 (DNA) or ~2.0 (RNA) and showing no degradation and other contaminations were subjected to further quality control with qPCR and qRT-PCR analysis with a panel of genes [54,56,106].

Immunohistochemical (IHC) Analysis

IHC experiments were performed with 5 μM tissue sections and with antibodies as described [56,106]. Images were taken with a Zeiss LSM 710 confocal microscope (ZEISS, Oberkochen, Germany).

Paired-End WGS and RNA-Seq

Illumina sequencing was conducted. Paired-end 125 × 125 bp WGS was performed in collaboration with the BGI-America and the High Throughput Genomics Core Facility at Huntsman Cancer Center at the University of Utah (Salt Lake City, UT, USA). RNA-seq was performed in collaboration with the Georgia Genomics Facility at the University of Georgia.

Sequence Data Analyses

Sequence data were analyzed following pipelines as described [55,56,106]. Briefly, WGS reads were aligned to the dog reference genome canFam3.1 with BWA v0.7.10 (bio-bwa.sourceforge.net). RNA-seq reads were mapped to the same reference genome using either TopHat 2.1.1 (ccb.jhu.edu/software/tophat/index.shtml) (for gene expression) or STAR v2.4.1c (github.com/alexdobin/STAR) (for mutation finding). Both RNA-seq-based canine gene annotation [106] and human xenoRefGene [56] annotation were used. Both WGS and RNA-seq reads were used for mutation discovery with GATK v3.6 and MuTect. Known canine single nucleotide polymorphisms (SNPs) were excluded as described [106]. WGS data were used to identify inversions/translocations and chimeric fusion genes [54-56,106]. For CNA discovery, correctly and uniquely mapped WGS read pairs were used [54-56,106]. Gene expression quantification with RNA-seq reads and other analyses were performed as described [55,56,106].

Microbiome Analysis

Microbiome analysis was performed as described [106]. Briefly, WGS and RNA-seq read pairs that could not be placed onto the canine genome were mapped with BWA v0.7.10 to two

microbial genome databases: HMP (the reference genome database curated by the Human Microbiome Project) and ABG (all bacterial genomic sequences) [106]. The bacterial diversity was calculated D by:

(1) Simpson's Diversity:

$$D = 1 - \frac{\sum n_i(n_i-1)}{N(N-1)};$$

(2) Shannon-Wiener Diversity:

$$D = -\sum p_i * lnp_i\,;\; p_i = \frac{n_i}{N}.$$

In both methods, n_i is the total number of reads mapped to the i^{th} species, and N is the total number reads mapped to all species.

TCGA Data Analysis

RNA-seq expression and WGS data of TCGA human colon cancers were obtained from the NCI GDC data portal (portal.gdc.cancer.gov). The mutation data were downloaded from the cBioportal Cancer Genomics database (www.cbioportal.org). Subtyping was performed using ssGSEA enrichment scores of CRC signatures as summarized in Materials.

Data Access

Sequence data have been submitted to the NCBI SRA database with accession number PRJNA418842.

Conclusions

Consistent with our previous CNA study [54], our current findings support that dogs share the same CRC development and progression pathways as humans. Furthermore, our study sheds

light on the molecular features unique to proliferative and invasive canine tumors. Importantly, we identified three modes of CRC cell invasion in dogs and humans. Our work reveals that CMS4 human colon cancers consist of two subtypes, EMT and crypt-like invasion, that differ in TGF-β signaling and microbe content.

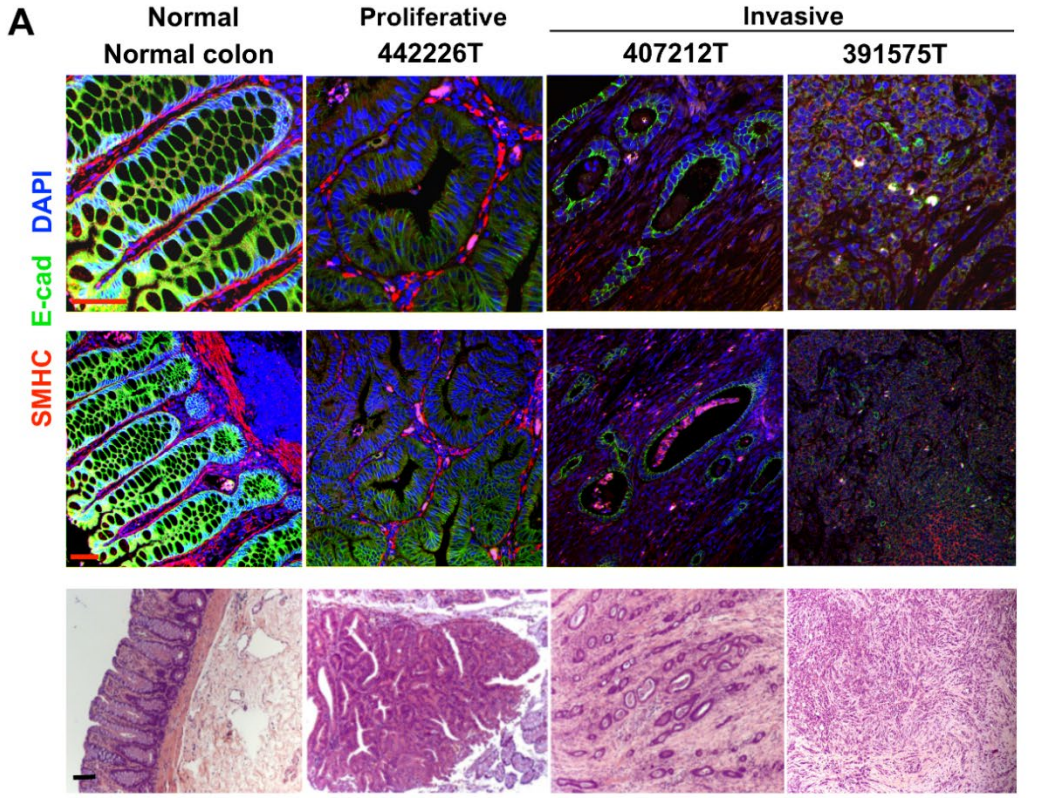
Supplementary Materials

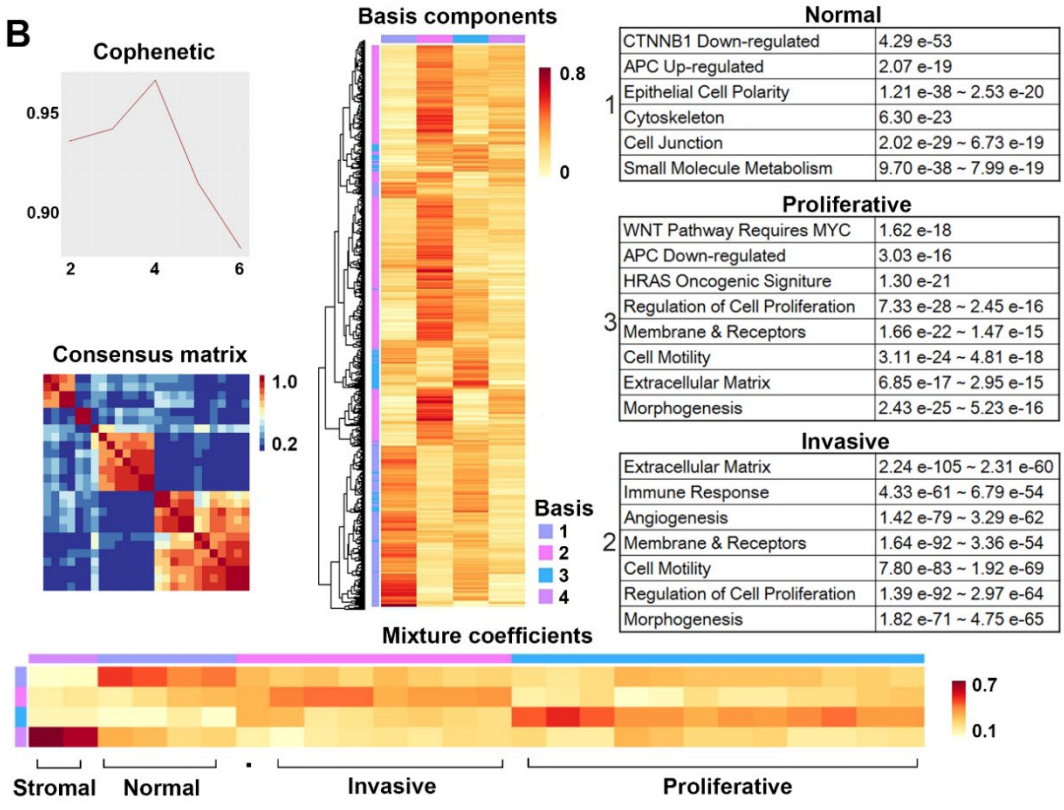
The following are available online at https://www.mdpi.com/2072-6694/10/9/330/s1, Figure S3.1: Hierarchical clustering of 26 canine samples, Figure S3.2: GSEA of canine tumors with canonical CRC signatures and canine CTNNB1 mutations, Figure S3.3: Stromal signature gene expression and representative IHC images of canine crypt-like and EMT invasion, Figure S3.4: TP53 mutations in human CRCs from TCGA, Figure S3.5: Subtyping of human colon cancer from TCGA and mutation signatures, Table S3.1: Canine case information and clustering analysis, Table S3.2: Canine GSEA and gene mutation of canonical CRC pathways, Table S3.3. Gene expression of fibroblast activation markers, Table S3.4: Canine microbiome analyses, Table S3.5: TP53 mutations, Table S3.6: Human colon cancer subtyping and molecular features.

Acknowledgements

We thank Huan Xiong; Jin Qian and Ye Wang for their contribution to the study; Holly Borghese and Alison G. Meindl for their help in sample collection and manuscript editing; the Georgia Genomics Facility, Brian Dalley of the University of Utah and the BGI for sequencing. Confocal imaging was performed at the UGA Biomedical Microscopy Core.

FIGURES





Normal

CTNNB1 Down-regulated	4.29 e-53
APC Up-regulated	2.07 e-19
Epithelial Cell Polarity	1.21 e-38 ~ 2.53 e-20
Cytoskeleton	6.30 e-23
Cell Junction	2.02 e-29 ~ 6.73 e-19
Small Molecule Metabolism	9.70 e-38 ~ 7.99 e-19

1

Proliferative

WNT Pathway Requires MYC	1.62 e-18
APC Down-regulated	3.03 e-16
HRAS Oncogenic Signiture	1.30 e-21
Regulation of Cell Proliferation	7.33 e-28 ~ 2.45 e-16
Membrane & Receptors	1.66 e-22 ~ 1.47 e-15
Cell Motility	3.11 e-24 ~ 4.81 e-18
Extracellular Matrix	6.85 e-17 ~ 2.95 e-15
Morphogenesis	2.43 e-25 ~ 5.23 e-16

3

Invasive

Extracellular Matrix	2.24 e-105 ~ 2.31 e-60
Immune Response	4.33 e-61 ~ 6.79 e-54
Angiogenesis	1.42 e-79 ~ 3.29 e-62
Membrane & Receptors	1.64 e-92 ~ 3.36 e-54
Cell Motility	7.80 e-83 ~ 1.92 e-69
Regulation of Cell Proliferation	1.39 e-92 ~ 2.97 e-64
Morphogenesis	1.82 e-71 ~ 4.75 e-65

2

Figure 3.1. RNA-seq analysis clusters the tumors into two major groups-proliferative or invasive.

(A) Representative confocal (top two panels) and H&E staining (bottom panel) images of canine colon normal tissues, proliferative and invasive tumors. E-cad: E-cadherin; SMHC: smooth muscle myosin heavy chain (SMHC). Scale bar: 100 μM.

(B) Non-negative matrix factorization (NMF) clustering identifies four metagenes (top left and middle columns), three of which have significantly enriched functions (top right column). The metagenes cluster the samples into four groups (bottom), with invasive and proliferative ones being the largest. See also Table S3.1 and Figure S3.1.

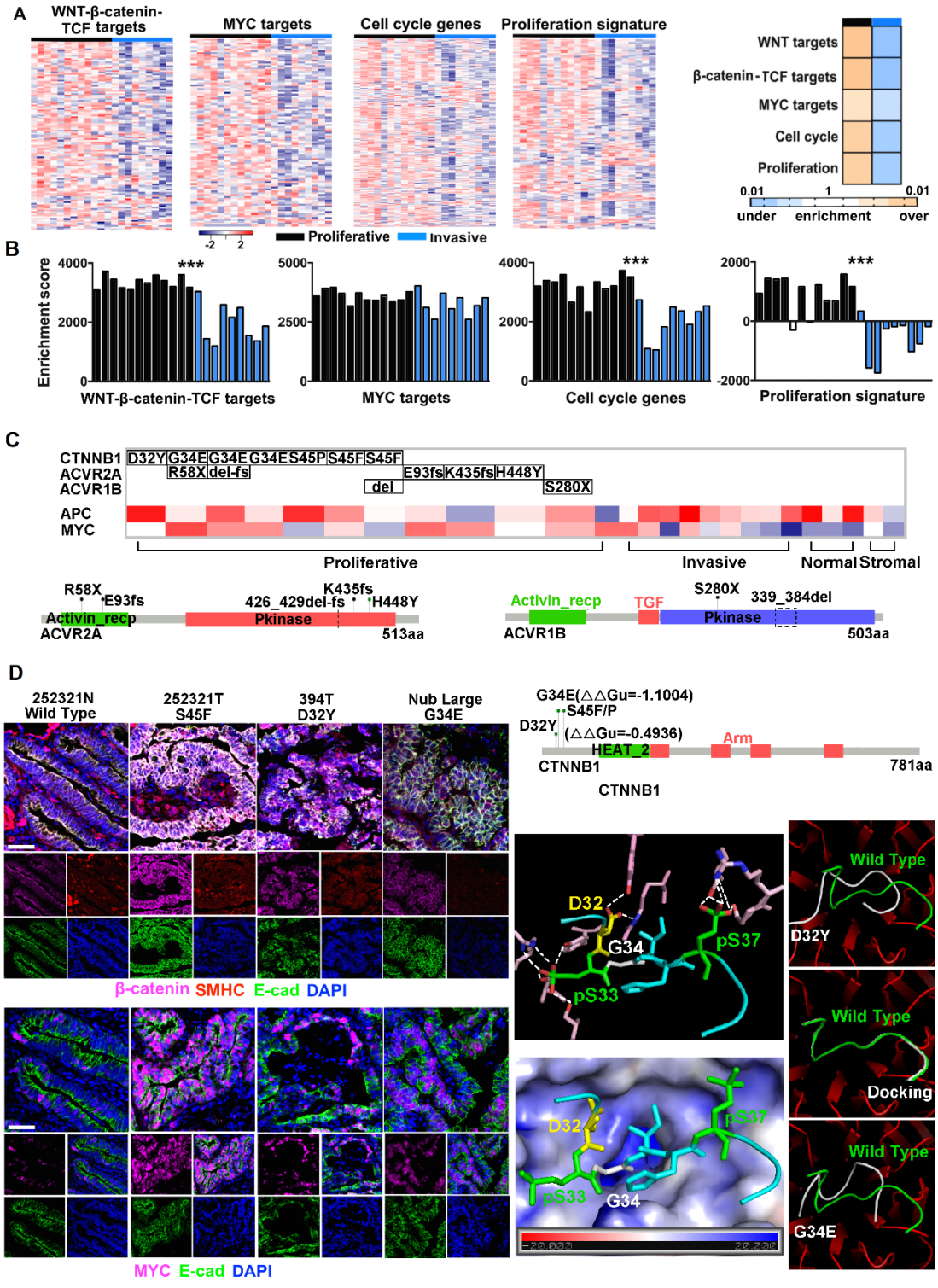


Figure 3.2. Proliferative tumors harbor activated WNT-β-catenin cell proliferation pathways and underlying gene mutations.

(A) Left four heatmaps indicate higher expression levels, represented by log2(FPKM), of gene signatures shown in proliferative tumors than invasive tumors, with their GSEA p-values specified by the right heatmap.

(B) The bar plots indicate ssGSEA enrichment scores of the same gene signatures, reaching the same conclusion as (A), ***: $p < 0.001$.

(C) Proliferative tumors harbor recurrent mutations of CTNNB1, ACVR2A and ACVR1B, as well as recurrent APC downregulation and MYC upregulation. Mutations of the three genes are shown at the bottom with their protein domains indicated. ΔΔGu, estimated as previously described [106], predicts if a missense mutation will alter the protein 3D structure.

(D) Representative IHC images indicate the enrichment of cellular and nuclear β-catenin (top) and nuclear MYC (bottom) in tumor cells harboring CTNNB1 mutations.

(E) The top 3D structure indicates that the phosphorylated N-terminal peptide of β-catenin binds β-TrCP1 through hydrogen bonds (dashed white lines) via D32, pS33 and pS37 of β-catenin. The bottom 3D structure indicates that the binding site locates in a positively charged pocket formed by β-TrCP1, and G34 of β-catenin locates at the center of the pocket.

(F) Docking of β-catenin peptides to β-TrCP1 indicates that D32Y and G34E mutations alter substrate binding. The ground truth peptide binding in the crystal structure [126] is shown green, while peptide docking is shown in white. See also Figure S3.2, Table S3.2.

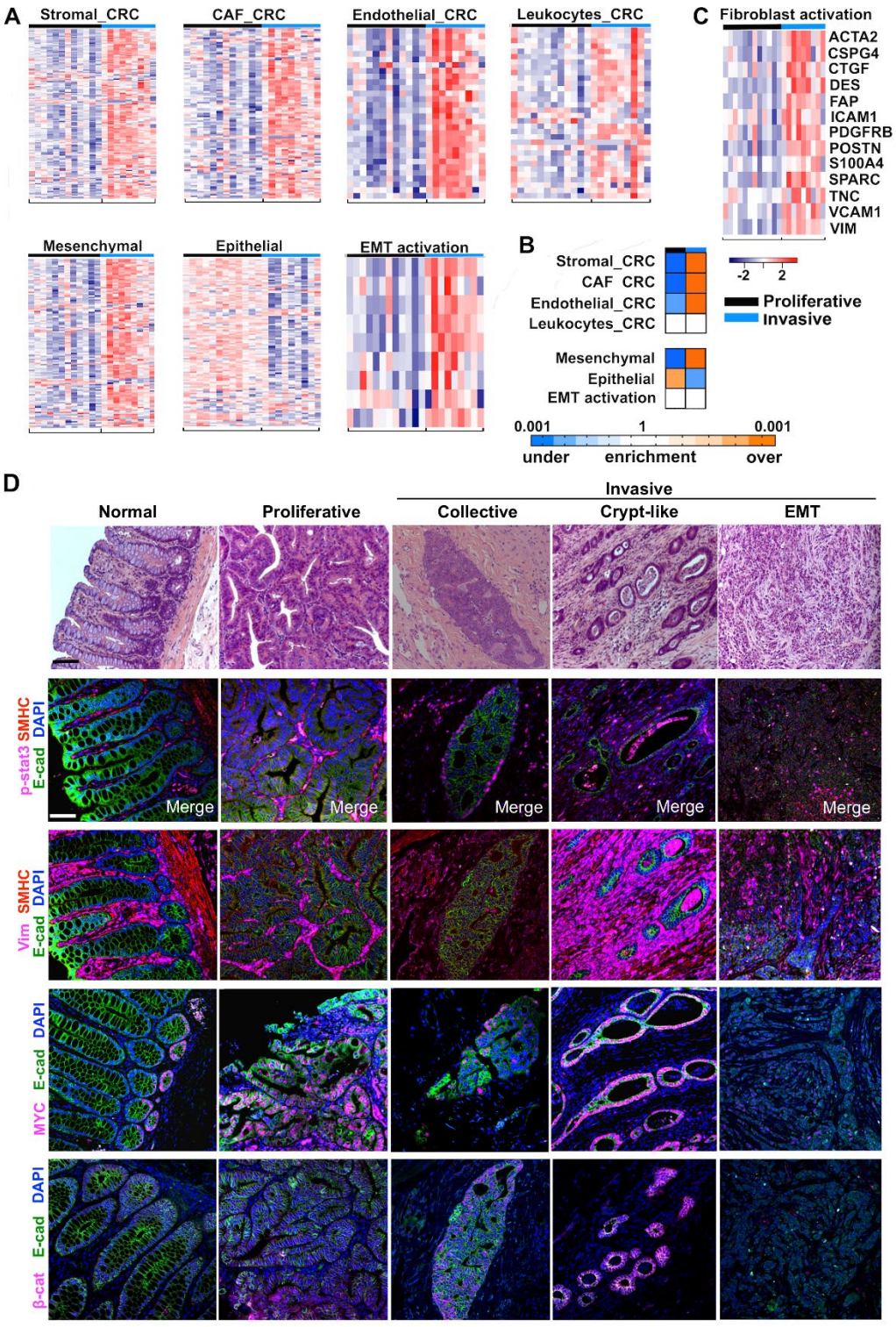


Figure 3.3. Stromal signatures are activated in invasive tumors and three invasion modes are observed.

(A) Heatmaps indicate higher expression levels of stromal and EMT signature genes in invasive tumors than in proliferative tumors. Heatmaps are presented as described for Figure 3.2A.

(B) Heatmaps of the GSEA p-values of the signatures indicated.

(C) Representative IHC images of normal colon, proliferative tumor, and tumors of three invasion modes. In crypt-like invasion (407212T), tracks of MYC-positive crypt-like structures are surrounded by dense and multilayers of fibroblasts. β-cat: β-catenin. See also Figure S3.3, Tables S3.2 and S3.3.

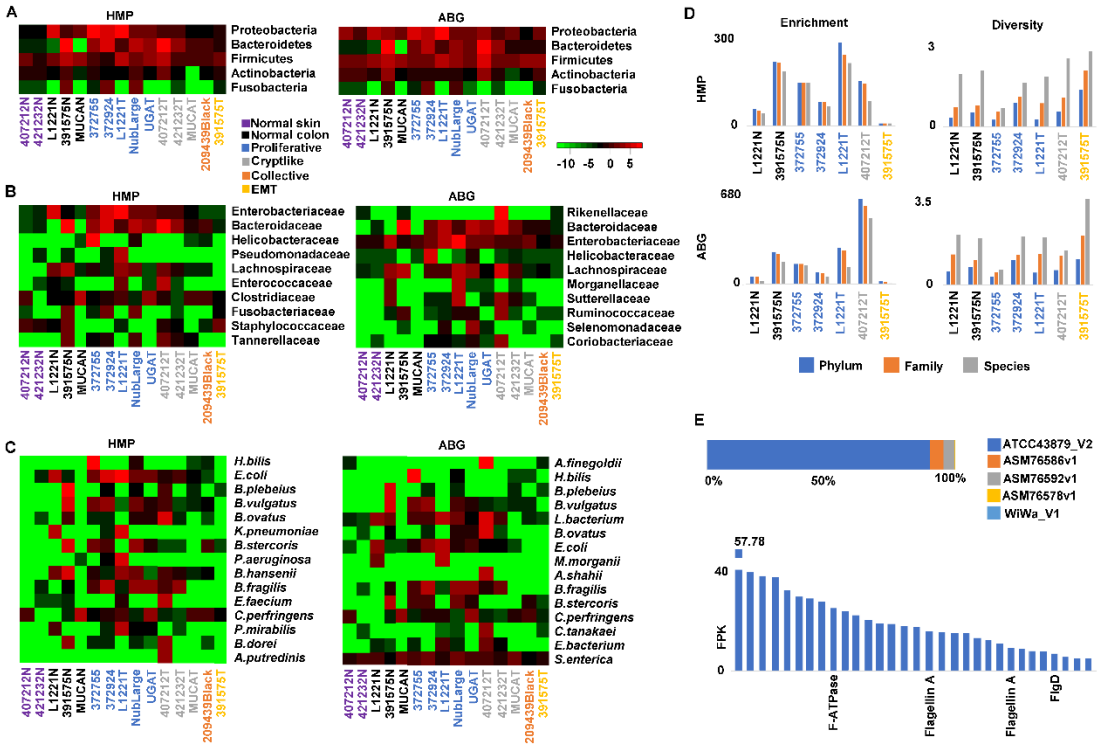
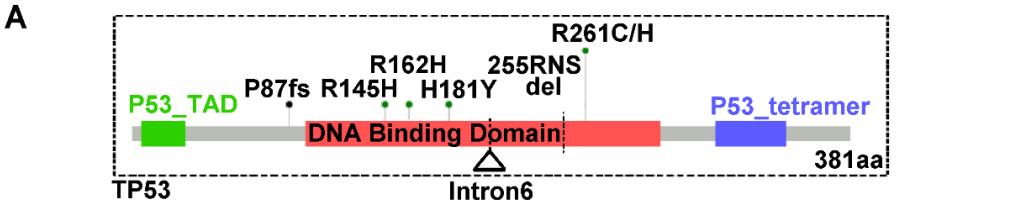


Figure 3.4. A crypt-like invasion tumor harbors a microbiome that resembles mucosa samples.

(A–C) Heatmaps indicate enrichment levels (Red: enriched; green: depleted) of bacterial phylum, family and species in each sample. HMP (Human Microbiome Project) and ABG (all bacterial genomic sequences) are the two microbial databases used. Sample types are specified by the colors as indicated.

(D) Crypt-like invasion tumor 407212T (gray), but not EMT invasion tumor 391575T (yellow), resembles normal colorectal mucosal samples (black) and proliferative tumors (blue) in bacterial enrichment and diversity.

(E) *H. bilis* strain ATCC 43879 is enriched in proliferative tumor 372755 (top) and expresses genes including those encoding flagellin A (bottom). See also Table S3.4.



B

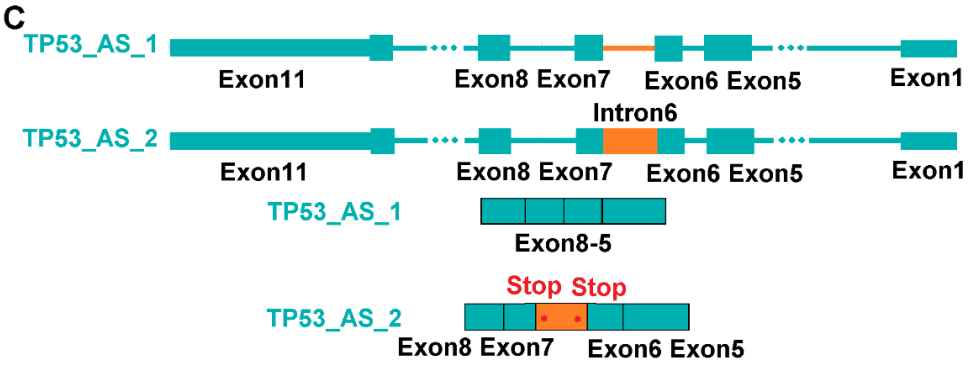
```
Human MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQP
Dog   MQEPQSELN IDPPLSQETFSELWNLLPENNVLSSELCPA VDEL L - LPESVVN

Human WFTEDPGPDEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQ
Dog   WLDEDS - -DDAPRMPAT - - - - - - - - - - SAPTAPGPAPSWPLSSSVPSPKTYP
                                                          P87fs

Human GSYGFRLGFLHSGTAKSVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPPGTRV
Dog   GTYGFRLGFLHSGTAKSVTWTYSPLLNKLFCQLAKTCPVQLWVSSPPPPNTCV

      R158H/S/C          R175H                    H193Y/R/L
Human RAMAIYKQSQHMTEVVRRCPHHERCSD - SDGLAPPQHLIRVEGNLRVEYLDDR
Dog   RAMAIYKKS E F VTEVVRRCPHHERCSDSSDGLAPPQHLIRVEGNLRAKYLDDR
      R145H              R162H                    H181Y
Human NTFRHSVVVPYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGN
Dog   NTFRHSVVVPYEPPEVGSDYTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGN
          R273C/H
Human LLGRNSFEVRVCACPGRDRRTEEENLRKKGEPHHELPPGSTKRALPNNTSSSP
Dog   VLGRNSFEVRVCACPGRDRRTEEENFHKKGEPCPEPPPGSTKRALPPSTSSSP
      255del-fs  R261C/H
Human QPKKKPLDGEYFTLQIRGRERFEMFRELNEALELKDAQAGKEPGGSRAHSSH
Dog   PQKKKPLDGEYFTLQIRGRERYEMFRNLNEALELKDAQSGKEPGGSRAHSSH

Human LKSKKGQSTSRHKKLMFKTEGPDSD 393
Dog   LKAKKGQSTSRHKKLMFKREGPDSD 381
```



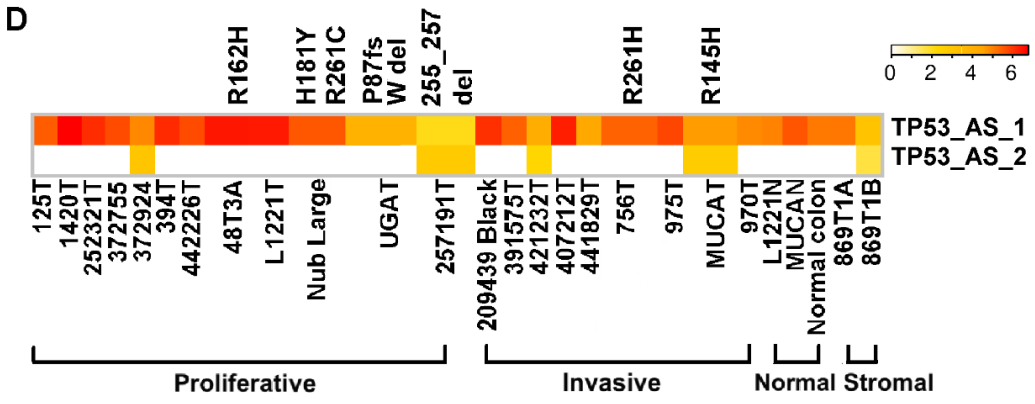


Figure 3.5. TP53 is recurrently altered in both proliferative and invasive tumors, with some being known drivers.

(A) TP53 mutations include whole gene deletion, indicated by the dashed lines, and other changes shown.

(B) Human and dog TP53 protein alignment, with canine mutations and some of their human counterparts (e.g., R175H and R273C/H) indicated below and above the alignment respectively.

(C) Intron 6 retention, yielding two premature stop codons, was detected.

(D) TP53 is altered in both proliferative and invasive tumors. The heatmap indicates the abundance of the two transcripts shown in (C). See also Table S3.5 and Figure S3.4.

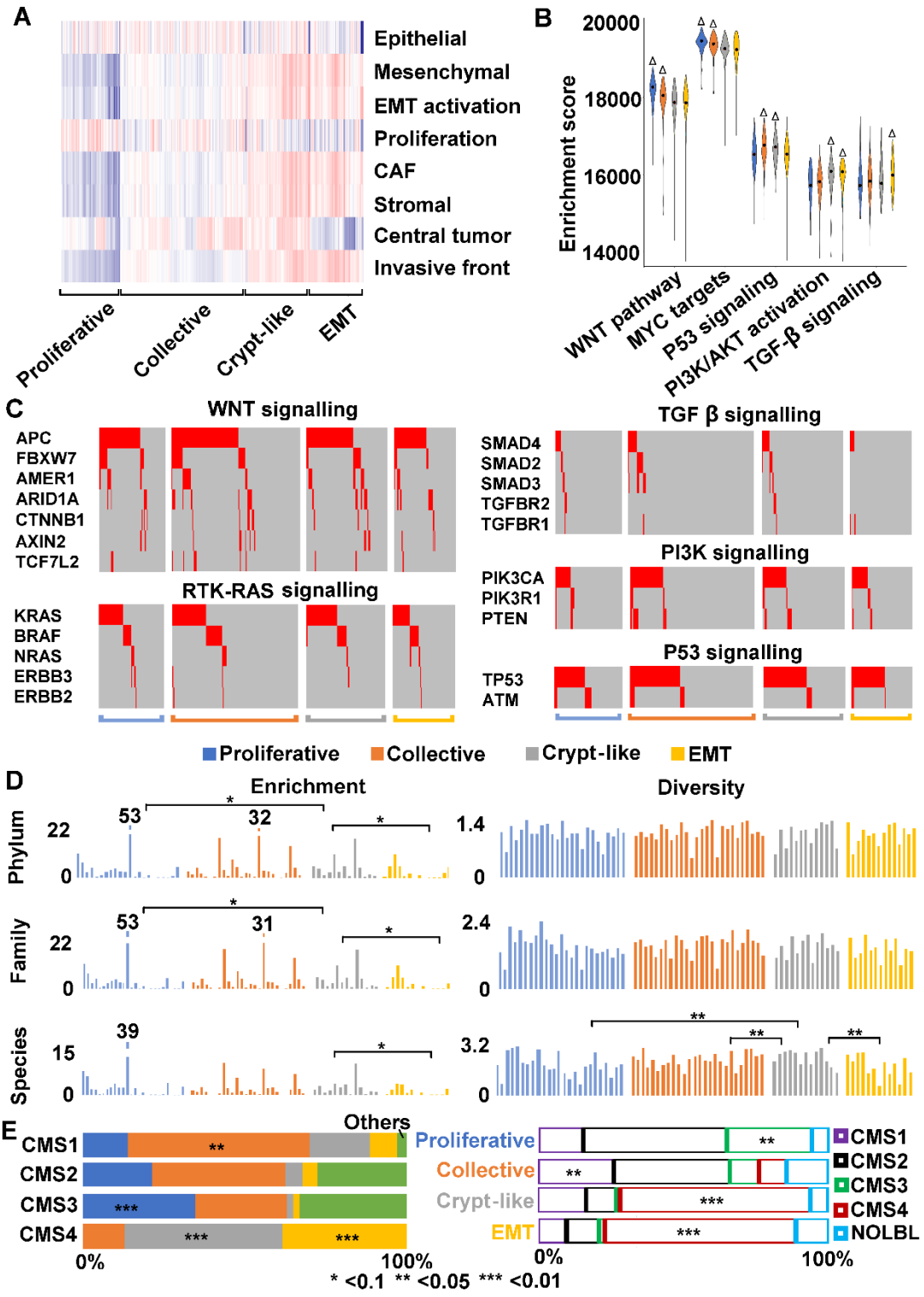


Figure 3.6. We have classified TCGA colon cancers and found that consensus molecular subtype 4 (CMS4) consists of crypt-like and EMT invasion cancers.

(A) TCGA colon cancers (379 out of 478 total) were classified into four subtypes, with their ssGSEA enrichment scores of signatures indicated by the heatmap. Red: enriched; blue: depleted. **(B)** The four subtypes differ in canonical CRC pathways, as indicated by the distribution of ssGSEA enrichment scores represented by the violin plot. Δ: significant activation. **(C)** TGF-β signaling genes are less frequently mutated in EMT invasion compared to other subtypes. Each column represents a tumor sample, and only driver mutations (red) are shown. **(D)** EMT tumors overall harbor fewer bacteria, compared to tumors of other subtypes. **(E)** CMS4 can be further divided into crypt-like invasion and EMT invasion subtypes. Left bars indicate the distribution of our subtypes among the CMS subtypes, while right bars indicate the opposite. Others: not classified by us; NOLBL: no label (not classified [10]). See also Figure S3.5 and Table S3.6.

CHAPTER 4

CONCLUSIONS

In the past five years, our study was focused on expanding dog-human comparison strategy utilizing our canine CRC samples and TCGA human CRC samples. We reported the first case of extreme intestinal polyposis in the dog. Additionally, we performed an initial global comparison between proliferative and invasive colorectal tumors from 20 canine cases and evaluated their molecular homology to human colorectal cancer (CRC). Also, successful expand dog-human comparison strategy on driver somatic missense mutations. Lastly, we established the pipeline to discover the microbiome from tumor samples.

Firstly, we utilized the extreme intestinal polyposis as an example to build pipelines for germline mutations and microbiome discovery. We propose that three pathways lead to the dog extreme intestinal polyposis. First, MYC and cell cycle-promoting network activation, caused by a FBXW7 somatic mutation-initiated SCF E3 ubiquitin ligase defect, keeps crypt cells dividing. Second, defective intracellular trafficking and localization, originating from D406G germline mutation of STAMBPL1 and enhanced by MYC network activation, inhibit cell polarity establishment and cell differentiation, preventing cell cycle exit. Lastly, bacterial redox systems reduce the oxidative stress caused by germline mutation R51X of CYB5RL, decreasing cell death.

Secondly, consistent with our previous CNA study [54], our current findings support that dogs share the same CRC development and progression pathways as humans. Furthermore, our study sheds light on the molecular features unique to proliferative and invasive canine tumors.

Importantly, we identified three modes of CRC cell invasion in dogs and humans. Our work reveals that CMS4 human colon cancers consist of two subtypes, EMT and crypt-like invasion, that differ in TGF-β signaling and microbe content.

Besides, driver-passenger discrimination has always been a central aim of cancer research. Stromal drivers and microbial drivers are harder to identify, with fewer efficient approaches available. Unlike CTNNB1 mutation, both proliferative and invasive groups have recurrent TP53 mutations, with some being known drivers detecting by our expanded dog-human comparison strategy. Also, based on our microbiome analysis, prominent fibroblast proliferation in canine invasive tumors, as well as significant enrichment of *H. bilis* and *A. finegoldii* in canine tumors may open a new avenue to address these important but difficult questions. Indeed, fibroblasts are known to play an important role in human CRC and other cancers [122,135,140], *H. bilis* is linked to human IBD and CRC [132], and *A. finegoldii* is detected in blood samples of human CRC patients [131].

Lastly, we acknowledge our current canine sample size is small. Because of the vast heterogeneity, a much larger sample size is required for efficient driver-passenger discrimination via dog-human comparison. Also note that our current study has relied on WGS of tumor samples for microbiome analysis, which may fail to detect less abundant bacterial species. Metagenomics data from stool samples should also be examined. In the future, including the deep learning analysis on the H&E image is helpful to validate the microbiome analysis.

REFERENCES

1. Wong KM, Hudson TJ, McPherson JD. Unraveling the genetics of cancer: genome sequencing and beyond. *Annu Rev Genomics Hum Genet.* 2011;12:407-430.
2. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. *CA Cancer J Clin.* 2019;69(1):7-34.
3. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. *CA Cancer J Clin.* 2015;65(2):87-108.
4. Hagan S, Orr MCM, Doyle B. Targeted therapies in colorectal cancer—an integrative view by PPPM. *EPMA Journal.* 2013;4(1).
5. Howlader N, Ries LA, Mariotto AB, Reichman ME, Ruhl J, Cronin KA. Improved estimates of cancer-specific survival rates from population-based data. *J Natl Cancer Inst.* 2010;102(20):1584-1598.
6. Wang W, Kandimalla R, Huang H, et al. Molecular subtyping of colorectal cancer: Recent progress, new challenges and emerging opportunities. *Semin Cancer Biol.* 2019;55:37-52.
7. Ionov Y, Peinado MA, Malkhosyan S, Shibata D, Perucho M. Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. *Nature.* 1993;363(6429):558-561.

8. Okugawa Y, Grady WM, Goel A. Epigenetic Alterations in Colorectal Cancer: Emerging Biomarkers. *Gastroenterology.* 2015;149(5):1204-1225 e1212.

9. Pino MS, Chung DC. The chromosomal instability pathway in colon cancer. *Gastroenterology.* 2010;138(6):2059-2072.

10. Guinney J, Dienstmann R, Wang X, et al. The consensus molecular subtypes of colorectal cancer. *Nat Med.* 2015;21(11):1350-1356.

11. Tang J, Li Y, Lyon K, et al. Cancer driver-passenger distinction via sporadic human and dog cancer comparison: a proof-of-principle study with colorectal cancer. *Oncogene.* 2014;33(7):814-822.

12. Pon JR, Marra MA. Driver and passenger mutations in cancer. *Annu Rev Pathol.* 2015;10:25-50.

13. Lawrence MS, Stojanov P, Polak P, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature.* 2013;499(7457):214-218.

14. Evans P, Avey S, Kong Y, Krauthammer M. Adjusting for background mutation frequency biases improves the identification of cancer driver genes. *IEEE Trans Nanobioscience.* 2013;12(3):150-157.

15. Hodis E, Watson IR, Kryukov GV, et al. A landscape of driver mutations in melanoma. *Cell.* 2012;150(2):251-263.

16. Jeon S, Lambert PF. Integration of human papillomavirus type 16 DNA into the human genome leads to increased stability of E6 and E7 mRNAs: implications for cervical

carcinogenesis. *Proc Natl Acad Sci U S A.* 1995;92(5):1654-1658.

17. Eick D, Piechaczyk M, Henglein B, et al. Aberrant c-myc RNAs of Burkitt's lymphoma cells have longer half-lives. *The EMBO journal.* 1985;4(13b):3717-3725.

18. Wiestner A, Tehrani M, Chiorazzi M, et al. Point mutations and genomic deletions in CCND1 create stable truncated cyclin D1 mRNAs that are associated with increased proliferation rate and shorter survival. *Blood.* 2007;109(11):4599-4606.

19. Hodgkinson A, Chen Y, Eyre-Walker A. The large-scale distribution of somatic mutations in cancer genomes. *Hum Mutat.* 2012;33(1):136-143.

20. Lin J, Gan CM, Zhang X, et al. A multidimensional analysis of genes mutated in breast and colorectal cancers. *Genome Res.* 2007;17(9):1304-1318.

21. Wendl MC, Wallis JW, Lin L, et al. PathScan: a tool for discerning mutational significance in groups of putative cancer genes. *Bioinformatics.* 2011;27(12):1595-1602.

22. Dees ND, Zhang Q, Kandoth C, et al. MuSiC: identifying mutational significance in cancer genomes. *Genome Res.* 2012;22(8):1589-1598.

23. Babaei S, Hulsman M, Reinders M, de Ridder J. Detecting recurrent gene mutation in interaction network context using multi-scale graph diffusion. *Bmc Bioinformatics.* 2013;14.

24. Cerami E, Demir E, Schultz N, Taylor BS, Sander C. Automated network analysis identifies core pathways in glioblastoma. *PLoS One.* 2010;5(2):e8918.

25. Yeang CH, McCormick F, Levine A. Combinatorial patterns of somatic gene mutations in

cancer. *FASEB J.* 2008;22(8):2605-2622.

26. Pena-Llopis S, Christie A, Xie XJ, Brugarolas J. Cooperation and antagonism among cancer genes: the renal cancer paradigm. *Cancer Res.* 2013;73(14):4173-4179.

27. Gnad F, Baucom A, Mukhyala K, Manning G, Zhang Z. Assessment of computational methods for predicting the effects of missense mutations in human cancers. *BMC Genomics.* 2013;14 Suppl 3:S7.

28. Tamborero D, Gonzalez-Perez A, Perez-Llamas C, et al. Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci Rep.* 2013;3:2650.

29. Integrative HMPRNC. The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell Host Microbe.* 2014;16(3):276-289.

30. Louis P, Hold GL, Flint HJ. The gut microbiota, bacterial metabolites and colorectal cancer. *Nat Rev Microbiol.* 2014;12(10):661-672.

31. Flemer B, Lynch DB, Brown JM, et al. Tumour-associated and non-tumour-associated microbiota in colorectal cancer. *Gut.* 2017;66(4):633-643.

32. Abdulamir AS, Hafidh RR, Bakar FA. Molecular detection, quantification, and isolation of Streptococcus gallolyticus bacteria colonizing colorectal tumors: inflammation-driven potential of carcinogenesis via IL-1, COX-2, and IL-8. *Mol Cancer.* 2010;9:249.

33. Boleij A, Tjalsma H. Gut bacteria in health and disease: a survey on the interface between intestinal microbiology and colorectal cancer. *Biological Reviews.* 2012;87(3):701-730.

34. Jones M, Helliwell P, Pritchard C, Tharakan J, Mathew J. Helicobacter pylori in colorectal neoplasms: is there an aetiological relationship? *World J Surg Oncol.* 2007;5:51.

35. Higashi H, Tsutsumi R, Fujita A, et al. Biological activity of the Helicobacter pylori virulence factor CagA is determined by variation in the tyrosine phosphorylation sites. *Proc Natl Acad Sci U S A.* 2002;99(22):14428-14433.

36. Guo Y, Li HY. Association between Helicobacter pylori infection and colorectal neoplasm risk: a meta-analysis based on East Asian population. *J Cancer Res Ther.* 2014;10 Suppl:263-266.

37. Rhee KJ, Wu S, Wu X, et al. Induction of persistent colitis by a human commensal, enterotoxigenic Bacteroides fragilis, in wild-type C57BL/6 mice. *Infect Immun.* 2009;77(4):1708-1718.

38. Sobhani I, Tap J, Roudot-Thoraval F, et al. Microbial dysbiosis in colorectal cancer (CRC) patients. *PLoS One.* 2011;6(1):e16393.

39. Boleij A, Hechenbleikner EM, Goodwin AC, et al. The Bacteroides fragilis toxin gene is prevalent in the colon mucosa of colorectal cancer patients. *Clin Infect Dis.* 2015;60(2):208-215.

40. Rubinstein MR, Wang X, Liu W, Hao Y, Cai G, Han YW. Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating E-cadherin/beta-catenin signaling via its FadA adhesin. *Cell Host Microbe.* 2013;14(2):195-206.

41. Kostic AD, Chun E, Robertson L, et al. Fusobacterium nucleatum potentiates intestinal

tumorigenesis and modulates the tumor-immune microenvironment. *Cell Host Microbe.* 2013;14(2):207-215.

42. McCoy AN, Araujo-Perez F, Azcarate-Peril A, Yeh JJ, Sandler RS, Keku TO. Fusobacterium is associated with colorectal adenomas. *PLoS One.* 2013;8(1):e53653.

43. Maddocks OD, Short AJ, Donnenberg MS, Bader S, Harrison DJ. Attaching and effacing Escherichia coli downregulate DNA mismatch repair protein in vitro and are associated with colorectal adenocarcinomas in humans. *PLoS One.* 2009;4(5):e5517.

44. Arthur JC, Jobin C. The complex interplay between inflammation, the microbiota and colorectal cancer. *Gut Microbes.* 2013;4(3):253-258.

45. Prorok-Hamon M, Friswell MK, Alswied A, et al. Colonic mucosa-associated diffusely adherent afaC+ Escherichia coli expressing lpfA and pks are increased in inflammatory bowel disease and colon cancer. *Gut.* 2014;63(5):761-770.

46. Meuten DJ. *Tumors in Domestic Animals.* Ames, Iowa: Iowa State University Press; 2002.

47. Lindblad-Toh K, Wade CM, Mikkelsen TS, et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature.* 2005;438(7069):803-819.

48. Pontius JU, Mullikin JC, Smith DR, et al. Initial sequence and comparative analysis of the cat genome. *Genome Res.* 2007;17(11):1675-1689.

49. Shearin AL, Ostrander EA. Leading the way: canine models of genomics and disease. *Dis Model Mech.* 2010;3(1-2):27-34.

50. Gardner HL, Fenger JM, London CA. Dogs as a Model for Cancer. *Annual Review of*

Animal Biosciences. 2016;4(1):199-222.

51. Boyko AR. The domestic dog: man's best friend in the genomic era. *Genome Biol.* 2011;12(2):216.

52. Hayward JJ, Castelhano MG, Oliveira KC, et al. Complex disease and phenotype mapping in the domestic dog. *Nat Commun.* 2016;7:10460.

53. Paoloni M, Khanna C. Translation of new cancer treatments from pet dogs to humans. *Nat Rev Cancer.* 2008;8(2):147-156.

54. Tang J, Le S, Sun L, et al. Copy number abnormalities in sporadic canine colorectal cancers. *Genome Res.* 2010;20(3):341-350.

55. Liu D, Xiong H, Ellis AE, et al. Canine spontaneous head and neck squamous cell carcinomas represent their human counterparts at the molecular level. *PLoS Genet.* 2015;11(6):e1005277.

56. Liu D, Xiong H, Ellis AE, et al. Molecular homology and difference between spontaneous canine mammary cancer and human breast cancer. *Cancer Res.* 2014;74(18):5045-5056.

57. Kinzler KW, Vogelstein B. Lessons from Hereditary Colorectal Cancer. *Cell.* 1996;87(2):159-170.

58. Youmans L, Taylor C, Shin E, et al. Frequent alteration of the tumor suppressor gene APC in sporadic canine colorectal tumors. *PLoS One.* 2012;7(12):e50813.

59. Eifert C, Powers RS. From cancer genomes to oncogenic drivers, tumour dependencies and therapeutic targets. *Nat Rev Cancer.* 2012;12(8):572-578.

60. Li Y, Xu J, Xiong H, et al. Cancer driver candidate genes AVL9, DENND5A and NUPL1 contribute to MDCK cystogenesis. *Oncoscience.* 2014;1.

61. Grady WM. Genetic testing for high-risk colon cancer patients1 1Abbreviations used in this paper: FAP, familial adenomatous polyposis; HMPS, hereditary mixed polyposis syndrome; HNPCC, hereditary nonpolyposis colon cancer; JPS, juvenile polyposis; MMR, mutation mismatch repair; MSI, microsatellite instability; PJS, Peutz-Jeghers syndrome; TGF, transforming growth factor. *Gastroenterology.* 2003;124(6):1574-1594.

62. Fearon ER. Molecular genetics of colorectal cancer. *Annu Rev Pathol.* 2011;6:479-507.

63. Samadder NJ, Neklason DW, Boucher KM, et al. Effect of Sulindac and Erlotinib vs Placebo on Duodenal Neoplasia in Familial Adenomatous Polyposis: A Randomized Clinical Trial. *JAMA.* 2016;315(12):1266-1275.

64. Johnson JC, DiSario JA, Grady WM. Surveillance and treatment of periampullary and duodenal adenomas in familial adenomatous polyposis. *Current Treatment Options in Gastroenterology.* 2004;7(2):79-89.

65. Di Cecilia S, Zhang F, Sancho A, et al. RBM5-AS1 Is Critical for Self-Renewal of Colon Cancer Stem-like Cells. *Cancer Res.* 2016;76(19):5615-5627.

66. Lipkin SM, Afrasiabi K. Familial colorectal cancer syndrome X. *Semin Oncol.* 2007;34(5):425-427.

67. Lipkin SM, Wang V, Jacoby R, et al. MLH3: a DNA mismatch repair gene associated with

mammalian microsatellite instability. *Nat Genet.* 2000;24(1):27-35.

68. Thompson BA, Greenblatt MS, Vallee MP, et al. Calibration of multiple in silico tools for predicting pathogenicity of mismatch repair gene missense substitutions. *Hum Mutat.* 2013;34(1):255-265.

69. Park DJ, Tao K, Le Calvez-Kelm F, et al. Rare mutations in RINT1 predispose carriers to breast and Lynch syndrome-spectrum cancers. *Cancer Discov.* 2014;4(7):804-815.

70. Hoeppner MP, Lundquist A, Pirun M, et al. An improved canine genome and a comprehensive catalogue of coding genes and non-coding transcripts. *PLoS One.* 2014;9(3):e91172.

71. Bai B, Zhao WM, Tang BX, et al. DoGSD: the dog and wolf genome SNP database. *Nucleic Acids Res.* 2015;43(Database issue):D777-783.

72. Folkman L, Stantic B, Sattar A, Zhou Y. EASE-MM: Sequence-Based Prediction of Mutation-Induced Stability Changes with Feature-Based Multiple Models. *J Mol Biol.* 2016;428(6):1394-1405.

73. Sato Y, Yoshikawa A, Yamagata A, et al. Structural basis for specific cleavage of Lys 63-linked polyubiquitin chains. *Nature.* 2008;455(7211):358-362.

74. Cancer Genome Atlas N. Comprehensive molecular characterization of human colon and rectal cancer. *Nature.* 2012;487(7407):330-337.

75. Alkhairy OK, Abolhassani H, Rezaei N, et al. Spectrum of Phenotypes Associated with Mutations in LRBA. *J Clin Immunol.* 2016;36(1):33-45.

76. May-Simera HL, Gumerson JD, Gao C, et al. Loss of MACF1 Abolishes Ciliogenesis and Disrupts Apicobasal Polarity Establishment in the Retina. *Cell Rep.* 2016;17(5):1399-1413.

77. Liyasova MS, Ma K, Lipkowitz S. Molecular pathways: cbl proteins in tumorigenesis and antitumor immunity-opportunities for cancer treatment. *Clin Cancer Res.* 2015;21(8):1789-1794.

78. Garcia-Domingo D, Ramirez D, Gonzalez de Buitrago G, Martinez-A C. Death Inducer-Obliterator 1 Triggers Apoptosis after Nuclear Translocation and Caspase Upregulation. *Molecular and Cellular Biology.* 2003;23(9):3216-3225.

79. Merlos-Suarez A, Barriga FM, Jung P, et al. The intestinal stem cell signature identifies colorectal cancer stem cells and predicts disease relapse. *Cell Stem Cell.* 2011;8(5):511-524.

80. Kosinski C, Li VS, Chan AS, et al. Gene expression patterns of human colon tops and basal crypts and BMP antagonists as intestinal stem cell niche factors. *Proc Natl Acad Sci U S A.* 2007;104(39):15418-15423.

81. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The human microbiome project. *Nature.* 2007;449(7164):804-810.

82. Schmitz S, Suchodolski J. Understanding the canine intestinal microbiota and its modification by pro-, pre- and synbiotics - what is the evidence? *Vet Med Sci.* 2016;2(2):71-94.

83. Silva RO, Lobato FC. Clostridium perfringens: A review of enteric diseases in dogs, cats

and wild animals. *Anaerobe.* 2015;33:14-17.

84. McDonel JL. Clostridium perfringens toxins (type A, B, C, D, E). *Pharmacology & Therapeutics.* 1980;10(3):617-655.

85. Shimizu T, Ohtani K, Hirakawa H, et al. Complete genome sequence of Clostridium perfringens, an anaerobic flesh-eater. *Proc Natl Acad Sci U S A.* 2002;99(2):996-1001.

86. Nakayama KI, Nakayama K. Ubiquitin ligases: cell-cycle control and cancer. *Nat Rev Cancer.* 2006;6(5):369-381.

87. Babaei-Jadidi R, Li N, Saadeddin A, et al. FBXW7 influences murine intestinal homeostasis and cancer, targeting Notch, Jun, and DEK for degradation. *J Exp Med.* 2011;208(2):295-312.

88. Choi SH, Wright JB, Gerber SA, Cole MD. Myc protein is stabilized by suppression of a novel E3 ligase complex in cancer cells. *Genes Dev.* 2010;24(12):1236-1241.

89. Komander D, Clague MJ, Urbe S. Breaking the chains: structure and function of the deubiquitinases. *Nat Rev Mol Cell Biol.* 2009;10(8):550-563.

90. Ibarrola N, Kratchmarova I, Nakajima D, et al. Cloning of a novel signaling molecule, AMSH-2, that potentiates transforming growth factor beta signaling. *BMC Cell Biol.* 2004;5:2.

91. McCullough J, Clague MJ, Urbe S. AMSH is an endosome-associated ubiquitin isopeptidase. *J Cell Biol.* 2004;166(4):487-492.

92. Cancer Genome Atlas Research N. Comprehensive molecular characterization of gastric

adenocarcinoma. *Nature.* 2014;513(7517):202-209.

93. Zeller KI, Jegga AG, Aronow BJ, O'Donnell KA, Dang CV. An integrated database of genes responsive to the Myc oncogenic transcription factor: identification of direct genomic targets. *Genome Biol.* 2003;4(10):R69.

94. Mili S, Moissoglu K, Macara IG. Genome-wide screen reveals APC-associated RNAs enriched in cell protrusions. *Nature.* 2008;453(7191):115-119.

95. Hyun DH, Lee GH. Cytochrome b5 reductase, a plasma membrane redox enzyme, protects neuronal cells against metabolic and oxidative stress through maintaining redox state and bioenergetics. *Age (Dordr).* 2015;37(6):122.

96. Reott MA, Parker AC, Rocha ER, Smith CJ. Thioredoxins in redox maintenance and survival during oxidative stress of Bacteroides fragilis. *J Bacteriol.* 2009;191(10):3384-3391.

97. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2010;26(5):589-595.

98. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics.* 2009;25(9):1105-1111.

99. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013;29(1):15-21.

100. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc*

Bioinformatics. 2013;43:11 10 11-33.

101. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol.* 2013;31(3):213-219.

102. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics.* 2015;31(2):166-169.

103. Gordon I, Paoloni M, Mazcko C, Khanna C. The Comparative Oncology Trials Consortium: using spontaneously occurring cancers in dogs to inform the cancer drug development pathway. *PLoS Med.* 2009;6(10):e1000161.

104. Rowell JL, McCarthy DO, Alvarez CE. Dog models of naturally occurring cancer. *Trends Mol Med.* 2011;17(7):380-388.

105. Parker HG, Shearin AL, Ostrander EA. Man's best friend becomes biology's best in show: genome analyses in the domestic dog. *Annu Rev Genet.* 2010;44:309-336.

106. Wang J, Wang T, Bishop MA, et al. Collaborating genomic, transcriptomic and microbiomic alterations lead to canine extreme intestinal polyposis. *Oncotarget.* 2018;9(49):29162-29179.

107. Nasir L, Devlin P, McKevitt T, Rutteman G, Argyle DJ. Telomere lengths and telomerase activity in dog tissues: a potential model system to study human telomere and telomerase biology. *Neoplasia.* 2001;3(4):351-359.

108. Rangarajan A, Weinberg RA. Opinion: Comparative biology of mouse versus human cells: modelling human cancer in mice. *Nat Rev Cancer.* 2003;3(12):952-959.

109. Chen HJ, Sun J, Huang Z, et al. Comprehensive models of human primary and metastatic colorectal tumors in immunodeficient and immunocompetent mice by chemokine targeting. *Nat Biotechnol.* 2015;33(6):656-660.

110. Liu Y, Sethi NS, Hinoue T, et al. Comparative Molecular Analysis of Gastrointestinal Adenocarcinomas. *Cancer Cell.* 2018;33(4):721-735 e728.

111. Wood LD, Parsons DW, Jones S, et al. The genomic landscapes of human breast and colorectal cancers. *Science.* 2007;318(5853):1108-1113.

112. Dihlmann S, von Knebel Doeberitz M. Wnt/beta-catenin-pathway as a molecular target for future anti-cancer therapeutics. *Int J Cancer.* 2005;113(4):515-524.

113. Inamura K. Colorectal Cancers: An Update on Their Molecular Pathology. *Cancers (Basel).* 2018;10(1).

114. Herring E, Kanaoka S, Tremblay E, Beaulieu JF. A Stool Multitarget mRNA Assay for the Detection of Colorectal Neoplasms. *Methods Mol Biol.* 2018;1765:217-227.

115. Siegel RL, Miller KD, Fedewa SA, et al. Colorectal cancer statistics, 2017. *CA Cancer J Clin.* 2017;67(3):177-193.

116. Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A.* 2004;101(12):4164-4169.

117. De Sousa EMF, Wang X, Jansen M, et al. Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. *Nat Med.* 2013;19(5):614-618.

118. Budinska E, Popovici V, Tejpar S, et al. Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. *J Pathol.* 2013;231(1):63-76.

119. Roepman P, Schlicker A, Tabernero J, et al. Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition. *Int J Cancer.* 2014;134(3):552-562.

120. Sadanandam A, Lyssiotis CA, Homicsko K, et al. A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nat Med.* 2013;19(5):619-625.

121. Dunne PD, McArt DG, Bradley CA, et al. Challenging the Cancer Molecular Stratification Dogma: Intratumoral Heterogeneity Undermines Consensus Molecular Subtypes and Potential Diagnostic Value in Colorectal Cancer. *Clin Cancer Res.* 2016;22(16):4095-4104.

122. Isella C, Terrasi A, Bellomo SE, et al. Stromal contribution to the colorectal cancer transcriptome. *Nat Genet.* 2015;47(4):312-319.

123. Van der Flier LG, Sabates-Bellver J, Oving I, et al. The Intestinal Wnt/TCF Signature. *Gastroenterology.* 2007;132(2):628-632.

124. Loboda A, Nebozhyn MV, Watters JW, et al. EMT is the dominant program in human colon cancer. *BMC Med Genomics.* 2011;4:9.

125. Liu C, Li Y, Semenov M, et al. Control of β-Catenin Phosphorylation/Degradation by a Dual-Kinase Mechanism. *Cell.* 2002;108(6):837-847.

126. Wu G, Xu G, Schulman BA, Jeffrey PD, Harper JW, Pavletich NP. Structure of a β-TrCP1-Skp1-β-Catenin Complex. *Molecular Cell.* 2003;11(6):1445-1456.

127. Tirosh I, Izar B, Prakadan SM, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science.* 2016;352(6282):189-196.

128. Thiery JP. Epithelial-mesenchymal transitions in tumour progression. *Nat Rev Cancer.* 2002;2(6):442-454.

129. Friedl P, Gilmour D. Collective cell migration in morphogenesis, regeneration and cancer. *Nat Rev Mol Cell Biol.* 2009;10(7):445-457.

130. Donaldson GP, Lee SM, Mazmanian SK. Gut biogeography of the bacterial microbiota. *Nat Rev Microbiol.* 2016;14(1):20-32.

131. Fenner L, Roux V, Ananian P, Raoult D. Alistipes finegoldii in blood cultures from colon cancer patients. *Emerg Infect Dis.* 2007;13(8):1260-1262.

132. Shomer NH, Dangler CA, Schrenzel MD, Fox JG. Helicobacter bilis-induced inflammatory bowel disease in scid mice with defined flora. *Infection and Immunity.* 1997;65(11):4858-4864.

133. Freed-Pastor WA, Prives C. Mutant p53: one name, many proteins. *Genes Dev.* 2012;26(12):1268-1286.

134. Sharma M, Castro-Piedras I, Simmons GE, Jr., Pruitt K. Dishevelled: A masterful conductor of complex Wnt signals. *Cell Signal.* 2018;47:52-64.

135. Rajaram M, Li J, Egeblad M, Powers RS. System-wide analysis reveals a complex network of tumor-fibroblast interactions involved in tumorigenicity. *PLoS Genet.* 2013;9(9):e1003789.

136. Tabassum DP, Polyak K. Tumorigenesis: it takes a village. *Nat Rev Cancer.* 2015;15(8):473-483.

137. Pagès F, Mlecnik B, Marliot F, et al. International validation of the consensus Immunoscore for the classification of colon cancer: a prognostic and accuracy study. *The Lancet.* 2018;391(10135):2128-2139.

138. Bullman S, Pedamallu CS, Sicinska E, et al. Analysis of Fusobacterium persistence and antibiotic response in colorectal cancer. *Science.* 2017;358(6369):1443-1448.

139. Routy B, Le Chatelier E, Derosa L, et al. Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors. *Science.* 2018;359(6371):91-97.

140. Marusyk A, Tabassum DP, Janiszewska M, et al. Spatial Proximity to Fibroblasts Impacts Molecular Features and Therapeutic Sensitivity of Breast Cancer Cells Influencing Clinical Outcomes. *Cancer Res.* 2016;76(22):6495-6506.