8-2018

GENOMIC AND TRANSCRIPTOMIC LANDSCAPE OF COLORECTAL PREMALIGNANCY

Kyle Chang

GENOMIC AND TRANSCRIPTOMIC LANDSCAPE OF COLORECTAL PREMALIGNANCY

by

Kyle Chang, B.S.

APPROVED:

____

Eduardo Vilar-Sanchez, M.D., Ph.D.

Advisory Professor

____

Ken Chen, Ph.D.

____

James E. Hixson, Ph.D.

____

Donald W. Parsons, M.D., Ph.D.

____

Paul A. Scheet, Ph.D.

APPROVED:

____

Dean, The University of Texas

MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences

GENOMIC AND TRANSCRIPTOMIC LANDSCAPE OF COLORECTAL PREMALIGNANCY

A

DISSERTATION

Presented to the Faculty of

The University of Texas

MD Anderson Cancer Center UTHealth

Graduate School of Biomedical Sciences

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

by

Kyle Chang, B.S.
Houston, Texas

August, 2018

To my family, David Tai Wai Chang, Man Ching Tang and John Lik Chun Chang

Acknowledgement

I would like to thank my advisor, Eduardo Vilar, for sharing his knowledge and supporting my scientific career development for the past 5 years. Under his guidance, I have learnt a great deal about colorectal premalignancy and prevention strategies. I am also thankful for the opportunities to participate in multiple external collaborations, for guiding my priorities when I was juggling with multiple projects, and for his constant encouragement when I was facing challenges.

I would like to thank my lab peers, Gita Bhatia, Hong Wu, Ester Borras, Laura Reyes, Lewins Walters, and Prashant Bommi for their advice and discussions. I also thank Anthony San Lucas for countless suggestions on bioinformatics analyses and Jerry Fowler for building a wonderful analysis pipelining tool that has saved me a lot of time in rerunning analyses.

I thank Shine Chang and Carrie Cameron for accepting me into the R25T cancer prevention research fellowship, which has supported large portion of the research presented in this dissertation. They have given me helpful advice in developing my scientific career and becoming a better communicator.

I thank my advisory committee members - Ken Chen, James Hixson, Will Parsons, and Paul Scheet for helping shape my dissertation project. I thank them for their valuable insights that have helped me better explain my research.

Lastly, I would like to thank my parents David Tai Wai Chang and Man Ching Tang, and my younger brother, John Lik Chun Chang for their love and support, and my girlfriend Claudia Wee, who has been supporting me throughout ups and downs of my PhD journey.

GENOMIC AND TRANSCRIPTOMIC LANDSCAPE OF COLORECTAL PREMALIGNANCY

Kyle Chang, B.S.

Advisory Professor: Eduardo Vilar, M.D., Ph.D.

Abstract

Colorectal cancer (CRC) is the third most commonly diagnosed cancer among men and women in the United States, with 3 to 5 percent of the cases diagnosed in the background of a hereditary form of the disease. Biologically, CRC is divided into two groups: microsatellite instable (MSI) and chromosomally unstable (CIN). Genomic and transcriptomic characterization of CRC has emerged from large-scale studies in recent years due to the advancement of next-generation sequencing technologies. These studies have identified key genes and pathways altered in CRC and provided insights to the discovery of therapeutic targets. Despite the wealth of knowledge acquired in the carcinoma stage, there have been insufficient efforts to systematically characterize premalignant lesions at the molecular level, which could lead to a better understanding of neoplastic initiation, risk prediction, and the development of targeted chemoprevention strategies. The challenge in characterizing premalignancy has always been the limited availability of sample material. This challenge is tackled by getting more samples, integrating public datasets, deploying better technology that use less amount of nucleic acids and in-silico tools to extract multi-layer information from the same experiment.

My genomic study consisted of whole exome sequencing (WES) and high-depth targeted sequencing on 80 premalignant lesions bulk tissue and crypts to assess clonality and mutational heterogeneity. WES results showed the presence of multiple clone in premalignancy based on clustering somatic mutation allele frequency. In addition, I determined that multiple clones originate from independent crypts harboring distinct APC and KRAS alterations. In my second study, I performed immune expression profiling and assessment of mutation and neoantigen rate of 28 premalignant lesions with DNA mismatch repair (MMR) deficient and proficient background using RNAseq. My results showed an activated immune profile despite low mutational and neoantigen rate, which challenges the canonical view in MMR-deficient carcinoma stage that immune activation is largely due to high mutation and neoantigen rate. In the last study, I performed transcriptomic sub-classifications of 398 premalignant lesions that associate them with different carcinomas subtypes, and clinical and histopathological features. My results revealed two major findings: prominent immune activation and WNT and MYC activation in premalignancy.

In summary, my large-scale genomic and transcriptomic analyses of colorectal adenomas have identified key molecular characteristics in early colorectal tumorigenesis and provide a foundation for discovering novel preventive strategies.

Table of Contents

References

VITA

List of Illustrations

List of Tables

1 INTRODUCTION

1.1 Colorectal Cancer

Colorectal cancer (CRC) is the third most commonly diagnosed cancer among men and women in the United States[1]. The lifetime incidence of CRC among individuals with average risk is 5%. However, individuals diagnosed with hereditary CRC syndromes face significantly higher risk at developing CRC at a young age. In fact, hereditary CRC syndromes account for 35 percent of young adult cancers[2]. There are two main focus of hereditary CRC syndromes: Familial Adenomatous Polyposis Syndrome (FAP) and Lynch Syndrome (LS), which account for 5% of the total of CRC cases. Both syndromes have an autosomal dominant inheritance and predispose individuals to develop colorectal polyps, premalignant lesions to CRC. However, FAP and LS exhibit different molecular pathway alterations and phenotype during carcinogenesis[3].

1.2 Familial Adenomatous Polyposis (FAP)

FAP is characterized by germline mutations in the adenomatous polyposis coli *(APC)* gene*,* a tumor suppressor[3]. FAP predisposes patients to develop hundreds of adenomatous polyps (adenomas) along the gastrointestinal tract and results in 100% lifetime risk of developing CRC if left untreated, with majority of diagnoses before the age of 35[3, 4]. The majority of the germline mutations are frame-shift or nonsense mutations that lead to premature truncation of protein synthesis and the rest manifest as loss of 5q[3]. The germline mutations are mostly distributed in the 5' half of the gene, with two hot spots at codon 1061 and 1309 of the β-catenin-binding sites. Germline inactivation of one *APC* allele greatly increases the rate of adenoma formation. It has been observed that FAP patients also acquire somatic *APC* alterations (loss of 5q or mutations), with the majority of the mutations located in the mutation cluster region between codon 1309 and 1450. As a result, biallelic loss of *APC* is a prominent feature in premalignant lesions and carcinomas of FAP patients, as well as 80% of sporadic CRC patients[3]. *APC* acts as a binding partner and regulator of β-catenin in the Wnt signaling pathway. Then, it leads to phosphorylation and subsequent degradation of β-catenin[3].

However, the loss of *APC* allows free β-catenin to translocate to the nucleus and activate downstream Wnt target genes, which include proto-oncogenes (such as *c-MYC* and *CCND1*), growth factors (such as *FGF20* and *FGF9*[3]), and continuous renewal of colonic crypts. This shows that *APC* plays a key role in adenoma initiation and growth promotion. Additional somatic mutations in key oncogenes and tumor suppressor genes such as *KRAS*, *BRAF*, *SMAD4*, *TP53* are somatically acquired and chromosomal instability develops during adenoma to carcinoma transformation. This process is called the “adenoma to carcinoma” transition model[5]. Therefore, FAP has been study widely as a model to understand CRC.

1.3 Lynch Syndrome (LS)

LS is caused by germline mutations in one of the DNA mismatch repair (MMR) genes: mutL homolog 1 (*MLH1*), mutS homolog 2 (*MSH2*), mutS homolog 6 (*MSH6*), PMS1 homolog2 (*PMS2*), or deletions in the epithelial cell adhesion molecule gene (*EPCAM*)[3, 6]. Approximately 90% of germline mutations are found in *MLH1* and *MSH2*. It is the most common hereditary CRC syndrome, which accounts for 3% of total CRC cases. It is estimated 1 in 280 of the population is affected by LS, affecting a total of 1.1 million people in the US[7]. LS patients are also at risk of developing endometrial, ovarian, and small intestine cancers[8]. *MLH1*, *MSH2* and *EPCAM* germline mutation carriers have similar lifetime risks for developing CRC, at approximately 70%[8],[9]. Although *MSH6* carriers have markedly lower risk in developing CRC (~20%), they still have 40% risk of developing LS-associated cancers[8]. Cancer develops in LS patients when they acquire a somatic mutation in the alternate allele of the same gene that carries the germline mutation. The resulting biallelic inactivation of MMR gene causes a rapid accumulation of mutations, particularly insertions and deletions, in microsatellites across the genome due to the fact that they are particularly susceptible to DNA polymerase slippage[10]. A subset of these mutations inevitably inactivates tumor suppressor genes or activates oncogenes, which accelerates the progression from adenoma to carcinoma. This molecular phenotype is called microsatellite instability (MSI) and occurs in about 15% of CRC[11, 12]. It is estimated that the transformation only take a few years at most in LS patients, compared to decades in average-risk population[3]. Although screening with annual colonoscopy has been demonstrated to decrease cancer incidence in LS[13], many patients continue developing CRC at a young age due to poor compliance with screening recommendations or the rapid development of interval cancers due to not yet well defined biologic reasons[14]. Interestingly, MSI carcinomas have relatively good prognosis and lower frequency of distant metastases compared to non-MSI carcinomas[12].

1.3.1 Microsatellite Instability and Immune Microenvironment

Carcinomas from LS patients or sporadic cases exhibit microsatellite instability due to MMR deficiency. Thus, they accumulate an excessive number of frame-shift mutations, which result in high amounts of mutated frameshift peptides (or neoantigens). The neoantigens are presented on cancer cells through binding the histocompatibility complex I or II (MHC I, MHC II). They are subsequently recognized by the immune system CD8+ cytotoxic T cells for MHC I antigens or by CD4+ helper T cells for MHC II antigens. As a consequence, neoantigens will activate the host immune response and cause T cell infiltration[15]. Although the activated T cells induce destruction tumor cells either by secreting cytotoxic molecules or pro-inflammatory cytokines, these neoantigens also upregulate immune inhibitory molecules such as PD-1, PD-L1, LAG-3 and other checkpoint inhibitors to counterbalance the infiltrating immune T cells[16]. As a result, the antitumor immune response is significantly impaired. Additional impairment of antitumor immune response includes mutations in genes involved in antigen presentation, such as beta2-mciroglobulin *(B2M)*, and antigen regulators such as Class II Major Histocompatibility Complex Transactivator *(CIITA*) or Regulatory Factor X5 *(RFX5)* have been identified in MSI CRC which impair antigen expression to T cells[17]. Therefore, restoring antitumor immune response has the potential to eliminate MSI carcinomas. In fact, treatment of MSI carcinomas with the checkpoint inhibitor Pembrolizumab or Nivolumab has been shown to recover antitumor function of T cells and demonstrated clinical benefit in terms of prolongation of progression free and overall survival[18-20].

1.4 Serrated Pathway

An alternative route of colorectal carcinogenesis is the serrated pathway and potentially accounts for up to one third of all CRC[3, 21]. This pathway is associated with high CpG island methylation phenotype (CIMP^{hi}) and *BRAF*^{V600E} mutations as the major driving mechanisms in both sporadic and familial cases[22] [23]. The serrated pathway has been linked to serrated polyps, which are different from traditional adenomas due to the “saw-toothed” appearance. This group of polyps are divided into hyperplastic polyps (HP), traditional serrated adenomas (TSA), and sessile serrated adenomas (SSA). HP is the most common type of polyp and is typically found in rectosigmoid colon. Although HP is considered as non-neoplastic, there is some evidence that it can progress into SSA, which has higher risk of developing into CRC[21, 22]. While SSA is more commonly found in proximal colon, TSA is more commonly found in rectosigmoid colon. Due to the malignant potential of TSA/SSA, they are both managed like traditional adenomas in the clinic[21, 22]. Serrated/hyperplastic polyposis syndrome (SPS) is a condition characterized by multiple and/or large HP or TSA/SSA in the colon. Approximately 33-59% of the individuals diagnosed with SPS have a family history of CRC but no germline mutation has been identified[21].

1.5 Clonality in Premalignancy

Intestinal epithelium undergoes continuous renewal of epithelial cells in crypts[24]. This process is maintained by stem cells at the base of crypts. The colorectal carcinogenesis model assumes that a single stem cell in a crypt (founder clone) acquires an *APC* alteration and accelerates mutated crypt expansion in a process called crypt fission[24, 25]. Therefore, neoplastic initiation is described as “monocryptal” in origin[24]. The founder clone, driven by the *APC* alteration, grows into a dominant population or clone. Consequently, *APC* alteration becomes a truncal or clonal event present in the majority of CRC. Subsequent acquisition of alterations, such as *KRAS* mutations, provide additional growth advantage and drive further tumor progression. *KRAS* mutations are present in a restricted, although significant, proportion of dysplastic cells, and generate subclones with different mutations that contribute to another level of intra-tumoral heterogeneity (ITH) early in carcinogenesis.

Advancement in next-generation sequencing and bioinformatics tools allow investigating such heterogeneity at high resolution. Computation tools such as ABSOLUTE and EXPANDS provides estimation on clonal composition based on somatic mutations generated from bulk tissue sequencing[26, 27]. These methods estimate the fraction of cancer cells containing a mutation by integrating mutation allele frequency, copy number variation, and tumor purity. For example, if all cancer cells contain a heterozygous mutation (clonal mutation) in a copy neutral region, then half of the reads are expected to support the mutation (0.5 allele frequency). Alternatively, if only half the cancer cells contain a heterozygous mutation (subclonal mutation) in a copy neutral region, then only 25 percent of the reads are expected to support the mutation (0.25 allele frequency). Therefore, clustering of the mutations' allele frequency, in a background of ploidy and purity, allows inference on the number of tumor clones.

1.6 The Consensus Molecular Subtypes of Colorectal Cancers

The canonical genetic pathways underlying the malignant transformation of colonic mucosa have been well characterized by Vogelstein et al[3, 5]. Specifically, they described a step-wise cascade of somatic mutations in tumor suppressor genes (e.g. *APC, TP53, SMAD4*) and oncogenes (e.g. *KRAS, PIK3CA*) and additional epigenetic aberrations that are now strongly implicated in colorectal cancer (CRC) initiation and growth[11]. For example, at the structural level, it has been shown that chromosomal instability (CIN) in the context of these driver mutations helps to promote tumor invasion. Indeed, imbalances in chromosome number (aneuploidy) and loss-of-heterozygosity are seen in 85% of invasive CRC tumors[28]. However, CRC is a heterogeneous disease with diverse biological characteristics that influence a patient's treatment options and prognosis that cannot simply be characterized by genetic and epigenetic alterations. Therefore, additional layers of molecular data by gene expression profiling, linked with genetic, epigenetic and clinical features, have been performed extensively to identify biological subtypes of CRC to provide a robust molecular characterization of CRC and a basis for clinical and translational stratification[29-33]. Despite various efforts in subtype identification, there have been a lack of consensus results due to different data platforms, processing methods, and algorithm designs implemented in these studies and impeded adoption in clinical and translational settings[29-32]. To reduce the discrepancies seen in different subtype identification strategies, Guinney et al provided a general framework by integrating various subtyping strategies[34]. They performed network analyses to study the associations of subtype labels used in different subtype classification system and identified four groups of subtype labels with significant associations which were termed "Consensus Molecular Subtypes" (CMSs). Using CMS labels as the gold standard, they developed a novel CMS prediction model using a random forest classifier on different gene expression platforms (Illumina RNAseq, Affymetrix and Ailgent arrays) and sample types (FFPE, fresh-frozen). Furthermore, CMS groups were associated with key genomic and epigenomic markers, clinical and pathological features, and prognosis[34]. In summary, each CMS group displays the following distinguished biological features: (1) CMS1, characterized by hypermutation, MSI, enrichment for *BRAF*^{V600E} mutation, and strong immune activation; (2) CMS2, a "canonical" subtype with marked WNT and MYC signaling activation and EGFR dependence; (3) CMS3, enriched for *KRAS* mutations and evident metabolic dysregulation; and (4) CMS4, a mesenchymal subtype, with prominent TGFβ activation, an immunosuppressive microenvironment, stromal invasion and angiogenesis activation. Notably, CMS tumor classification was shown to stratify CRC patients into distinct prognostic subgroups[34] and provide valuable insights for managing both early-stage and advanced disease.

1.7 Dissertation Objective

Genomic and transcriptomic characterization of CRC has emerged from large-scale studies in recent years due to the advancement of next-generation sequencing technologies[11, 34]. These studies have identified key genes and pathways altered in CRC and provided insights to the discovery and development of therapeutic targets[35, 36]. Despite our knowledge of the carcinoma stage, we have not fully explored the molecular characteristics of colorectal premalignancy, which can improve our understanding of neoplastic initiation and development of chemoprevention strategies. Therefore, there is an urgent need to establish a comprehensive molecular annotation of the biological pathways and a multi-layer model of initiation in colorectal premalignancy.

My long-term goal is to improve and develop novel CRC chemoprevention strategies by inhibiting initiation of premalignant lesions and their transformation into carcinomas, as well as risk prediction of individuals with premalignancy. The objective of this dissertation is to establish a model of initiation, characterize the immune expression profile and its relation to mutation and neoantigen load in MMR-deficient carcinogenesis, and analyze the sub-classification of premalignancy in association with different CRC subtypes to provide a comprehensive molecular characterization of colorectal premalignancy. My central hypothesis is that such large-scale genomic and transcriptomic characterization effort of premalignant lesions using next-generation sequencing technologies can provide insights into the molecular mechanisms of early colorectal tumorigenesis. The basis of this dissertation is derived from (i) the monocryptal polyclonal model of initiation in CRC carcinogenesis, (ii) the dependency of immune activation on high mutation rate and neoantigen load in MMR-deficient CRC, (iii) the established molecular classification of CRC using transcriptomic-based subgroups with prognostic and therapeutic features associated. Therefore, I will test the central hypothesis with the following specific aims:

1. To assess the presence and the origin of clonality in premalignancy.

1.1. To determine the presence of clonality in premalignancy. My working hypothesis is that estimating the fraction of cells containing somatic mutations and clustering their frequencies will allow inferring the number of tumor clones. I will use the bioinformatics tool ABSOLUTE to estimate tumor clones of premalignant lesions using data derived from whole exome sequencing.

1.2. To assess the origin of clonality in premalignancy. My working hypothesis is that there are independent clones derived from independent crypts (polyclonal polycryptal model) versus independent clones initiated from one ancestral crypt (polyclonal monocryptal model). I will detect distinct initiating *APC and* driver *KRAS* alterations in tissue biopsies and crypts using high-depth targeted sequencing and digital array technologies.

2. To determine the temporal changes in the immune expression profile and its relation to mutation and neoantigen rates from MMR-deficient premalignancy to carcinoma stages. My working hypothesis is that mutation and neoantigen rate arise late in the MMR-deficient carcinogenesis and are independent of the immune expression profile displayed in LS premalignancy. I will measure the expression of genes linked to the immune microenvironment (Th1/Tc1, CTL, Th17, Treg, proinflammation, and metabolism) and calculate mutation and neoantigen rates using whole transcriptome sequencing.

3. To determine the transcriptomic sub-classifications of premalignant lesions that associate them with different carcinomas subtypes, clinical and histopathological features. My working hypothesis is that the analysis of the transcriptomic profiles of premalignant lesions in the context of carcinomas subtypes can define specific events driving initial steps of carcinogenesis. I will apply CRC subtype classifier on the transcriptome data of premalignant lesions generated from microarray and whole transcriptome sequencing.

The research design and above specific aims are presented in the following chapters. In chapter 2, I present the research design and methods used in all the chapters of this dissertation. In chapter 3, I present the results of the project on the origin of clonality in premalignancy. In chapter 4, I present the results of the project studying the immune expression, mutation and neoantigen rates. Lastly, in chapter 5, I present the results of the project studying the transcriptomic sub-classification and pathway enrichments in premalignancy.

2 MATERIALS AND METHODS

Chapter 2 MATERIALS AND METHODS

The content of this chapter is based on the following publications:

Gausachs M, Borras E, **Chang K**, Gonzalez S, Azuara D, Delgado Amador A, Lopez-Doriga A, San Lucas FA, Sanjuan X, Paules MJ, Taggart MW, Davies GE, Ehli EA, Fowler J, Moreno V, Pineda M, You YN, Lynch PM, Lazaro C, Navin NE, Scheet PA, Hawk ET, Capella G, Vilar E. Mutational Heterogeneity in APC and KRAS Arises at the Crypt Level and Leads to Polyclonality in Early Colorectal Tumorigenesis. Clin Cancer Res. 2017 Oct 1;23(19):5936-5947. doi: 10.1158/1078-0432.CCR-17-0821. Epub 2017 Jun 23. PubMed PMID: 28645942; PubMed Central PMCID: PMC5626604.

Borras E, San Lucas FA, **Chang K**, Zhou R, Masand G, Fowler J, Mork ME, You YN, Taggart MW, McAllister F, Jones DA, Davies GE, Edelmann W, Ehli EA, Lynch PM, Hawk ET, Capella G, Scheet P, Vilar E. Genomic Landscape of Colorectal Mucosa and Adenomas. Cancer Prev Res (Phila). 2016 Jun;9(6):417-27. doi: 10.1158/1940-6207.CAPR-16-0081. Epub 2016 May 24. PubMed PMID: 27221540; PubMed Central PMCID: PMC4941624.

Copyright permissions are not required, since AACR states "Authors of articles published in AACR journals are permitted to use their article or parts of their article in the following ways without requesting permission from the AACR - Submit a copy of the article to a doctoral candidate's university in support of a doctoral thesis or dissertation."

Chang K, Taggart MW, Reyes-Uribe L, Borras E, Riquelme E, Barnett RM, Leoni G, San Lucas FA, Catanese MT, Mori F, Diodoro MG, You YN, Hawk ET, Roszik J, Scheet P, Kopetz S, Nicosia A, Scarselli E, Lynch PM, McAllister F, Vilar E. Immune Profiling of Premalignant Lesions in Patients with Lynch Syndrome. JAMA Oncol. 2018 Apr 16. doi: 10.1001/jamaoncol.2018.1482.

2.1 Genomic Profiling

2.1.1 Patients and Samples

A total of 80 colorectal adenomas from individuals with hereditary cancer syndromes and 6 tumors from sporadic cases were at collected at either the Catalan institute of Oncology or The University of Texas MD Anderson Cancer Center (UTMDACC). 25 of the total adenomas were analyzed with whole exome sequencing (Table 1), 37 of the total adenomas were analyzed with high-depth Ampliseq targeted sequencing and SNP arrays, and 18 of the total adenomas and 6 tumors were sectioned into multiple bulk biopsies and crypts and analyzed with digital genotyping technology (Table 2).

Table 1. Clinical characteristics of patients analyzed with WES

Patient ID	Gender	Age	Race	Pr Sx	No. of polyps detected	Cancer Dx	APC germline	No. Adenomas analyzed
CATA01	F	28	W	P	NA		c.3927_3931delAAAGA	4
CATA02	F	22	W	IRA	NA	D	c.4393_4394delAG	3
CATA03	M	38	W	P	>100		c. [1958+3G>A(;)c.1959G>A]	1
CATA04	M	47	W	P	NA	CC	c.1412delG	4
MDAC01	M	35	W	IRA	>100		c.1880dupA	2
MDAC02	M	33	W	IRA	<5		c.3810T>A	1
MDAC03	F	63	AA	IRA	<50		del exons 11-12	-
MDAC08	M	42	W	P	>100	HCC, D	c.622C>T	2
MDAC10	F	22	W	P	<5	HepBl	c.3440dupA	1
MDAC14	F	40	W*	IRA	<5		del 8-9	3
MDAC17	M	37	W	IRA	<5		c.1658G>A	2
MDAC18	M	27	W	P	>100	D	c.4393_4394delAG	1
MDAC20	M	65	W	IRA	<50		c.477C>G	1
MDAC21	F	70	W	IRA	<5		c.487C>T	-
MDAC24	F	25	AA	NP	>100	D	c.4733_4734del	-
MDAC26	F	28	W	IRA	>100	EC, OC	c.847C>T	-
MDAC29	F	25	W	P	>100		c.3810T>A	-
MDAC32	F	58	AA	NP	<50		c.1620insA	-
MDAC33	M	43	W	NP	>100		c.2894delA	-
MDAC34	M	29	W	NP	<50		c.5936delA	-

Abbreviations: Pr Sx, Prophylactic Surgery; M, male; F, Female; W, white; AA, african-american; IRA, Ileorectal anastomosis; P, Pouch; NP, not performed; HCC, Hepatocellular carcinoma; HepBI, Hepatoblastoma; D, Desmoid; EC, Endometrial cancer; OC, Ovarian cancer.

Table 2. Clinical characteristics of patients analyzed with Ampliseq, SNP arrays, and digital genotyping.

Patient			Germline information	Adenomas	Tumor characteristics			
Patient	**Gender**	**Age**	**Gene**		**Location**	**Grade**	**Size (cm)**	**Stage**
FAP1	M	19	*APC*	>150	-	-	-	-
MAP1	F	46	*MUTYH*	80-90	-	-	-	-
UFP1	M	54	*APC* and *MUTYH*	30-40	-	-	-	-
UFP2	M	48	*APC* and *MUTYH*	120	-	-	-	-
LYN1	F	51	*MLH1*	-	left	high	5.5	pT2 N0 mx/R0
CMMRD1	F	44	*PMS2*	-	-	-	-	-
SP1	F	67	-	-	left	low	3.5	pT3ab N1
SP2	M	73	-	-	right	low	5	pT3ab N2 Mx/R0
SP3	M	72	-	-	right	high	6	pT3cd (2) N0 Mx/R0
SP4	F	83	-	-	right	low	6	pT2 N0 Mx/R0
SP5	M	75	-	-	right	high	9.5	pT2 N0 Mx/R0
SP6	M	75	-	-	right	low	6.5	pT3 N0 Mx/R0
MDAC02	M	33	*APC*	<5	-	-	-	-
MDAC03	F	63	*APC*	<50	-	-	-	-
MDAC10	F	22	*APC*	<5	-	-	-	-
MDAC14	F	40	*APC*	<5	-	-	-	-
MDAC17	M	37	*APC*	<5	-	-	-	-
MDAC18	M	27	*APC*	>100	-	-	-	-
MDAC20	M	65	*APC*	<50	-	-	-	-
MDAC21	F	70	*APC*	<5	-	-	-	-
MDAC24	F	25	*APC*	>100	-	-	-	-
MDAC26	F	28	*APC*	>100	-	-	-	-
MDAC29	F	25	*APC*	>100	-	-	-	-
MDAC32	F	58	*APC*	<50	-	-	-	-
MDAC33	M	43	*APC*	>100	-	-	-	-
MDAC34	M	29	*APC*	<50	-	-	-	-

Whole Exome Sequencing

A total of 25 colorectal adenomas and matched blood samples from 12 patients diagnosed with FAP at the Catalan Institute of Oncology and UTMDACC at the time of endoscopic surveillance and were either fully resected or biopsied with snare forceps (Table 1). Tissues were retrieved from the endoscopy suite or the operating room and immediately flash-frozen or preserved in RNAlater storage solution (Life Technologies) and then stored at –80°C according to internal protocols. Blood was collected in EDTA tubes and stored appropriately for subsequent extraction of germline DNA. Genomic DNA was extracted from whole blood using the Blood & Cell Culture DNA Mini Kit (Qiagen) and from tissues using the QIAmp DNA Mini Kit (Qiagen). Exome DNA was captured using the SeqCap EZ Human Exome library v3.0 capture chip (Roche NimbleGen), which has a target capture region of 64 Mb. Samples were sequenced on an Illumina HiSeq 2000 sequencer with 76-base paired-end reads at the MD Anderson Sequencing Core Facility. Reads were aligned with the Burrows-Wheeler Alignment (BWA)[37] software to reference human genome version hg19. The initial alignment results were further processed with local realignment, duplicate read marking, and base quality recalibration by using Picard and the Genome Analysis Toolkit (GATK)[38] and by applying recommended best practices for sequence analysis from the Broad Institute.

2.1.2 Mutation Detection in WES

We used MuTect 1.14[39] for calling point mutations and Indelocator[40] for calling small

insertions and deletions. Somatic events called by MuTect and Indelocator were annotated with Annovar[41] and Variant Tools[42] with population allele frequencies of the 1000 Genomes Project and the Exome Variant Server (data release version ESP6500) for subsequent filtering of likely common polymorphisms and false positives. We excluded candidate somatic mutations seen at 1% or greater population allele frequency in either of these projects. All point mutations, small insertions, and deletions were visually verified using the Integrated Genome Viewer[43].

2.1.3 Mutational Signature Analysis

Mutation signatures were detected and plotted with deconstructSigs_1.8.0[44]

2.1.4 Allelic Imbalance Analysis

HapLOH[45] was used to detect allelic imbalance in SNP array. It combines the haplotype estimate, the SNP array's B allele frequencies (BAFs), and log R ratios (LRRs) to detect the deviations of phase concordance expected in blood samples. Then, a hidden Markov model (HMM) was applied to identify regions of subtle AIs. The over-represented alleles were called to classify AI regions into amplifications, deletions, cn-LOH, or indeterminable AI. HapLOHSeq[46], an extension to hapLOH, was used to detect allelic imbalance in whole exome sequencing and RNA sequencing data. To classify allelic imbalance (AI) events as amplification, deletion, or copy neutral LOH (cn-LOH), we used a hypothesis testing framework, with a null hypothesis of no presence of copy number changes between the blood and the adenoma samples in the region of AI (thus, cn-LOH) and an alternative that there are changes (deletion or amplification). We first applied a coverage normalization factor, for each adenoma/blood pair, across the exome to harmonize total coverage between the paired samples. Next, we calculated the mean coverage difference between the paired samples within each AI region. To establish significance of the putative deletion or amplification, we then applied a permutation-based test as follows: for each event, 10,000 permutations were performed, in which we randomly assigned sample labels at each genomic site and calculated a mean coverage difference between the adenoma and blood. The mean coverage difference obtained using the true sample labels was then compared with those generated from the permuted labels, and a permutation P-value was calculated as the frequency of permuted coverage differences that had a more extreme value than the true value. The null hypothesis of cn-LOH was rejected for permutation P-values ≤0.05. In such cases, the copy number alteration was determined to be the putative call (depending on the direction of coverage differential).

2.1.5 Clonality Estimation with WES

Somatic point mutations detected by MuTect 1.14[39] using default settings were filtered for “PASS” mutations. Copy number variations were detected using VarScan 2.3.7[47] with *P* = .01 and data-ratio=adenoma and matched blood reads ratio. VarScan log2ratio output was segmented with DNAcopy[48].We estimated the number of clones in adenomas by constructing kernel density estimation (KDE) using mutation cancer cell fractions (CCFs) and solving for the number of modes in the KDE. In the analysis of stage I CRC from TCGA, ABSOLUTE inputs are somatic point mutations obtained from the TCGA data portal and copy number variations obtained from the SNP Array 6.0 calls of the Broad Institute’s Firehose portal.

2.1.6 Ion Torrent Ampliseq Panel Analysis

A total of 37 adenoma samples from FAP patients with an IT Personal Genome Machine (PGM; Life Technologies) by the Sequencing and Non-Coding RNA Program at MD Anderson using the Ion PGM 200 Sequencing Kit on an Ion 318 Chip Kit (Life Technologies) (Table 2). IT Variant Caller v4.2 was run in the somatic low stringency mode to detect variants in *APC* and *KRAS* against reference human genome version hg19 on each adenoma and matched normal samples. Then, normal variants were subtracted from matched adenoma variants to create a list of somatic candidates for each adenoma and normal pair. Events located within the first and last 15% of the bases of the read were excluded. Then, a list of somatic candidates went through the following quality control steps: 1) Mutation allele frequencies were re-evaluated after removing variant reads where the mutation lies within the first 15% or last 15% of the bases of the reads; 2) Mutations with more than 2 variant alleles were excluded; 3) Mutations must be covered by a minimum of 100 reads. If a mutation allele frequency is 2-5% at least 10 reads must show the variant allele. If a mutation allele frequency

> 5% at least 25 reads must show the variant allele. Validation of somatic mutations and indels were performed using Sanger sequencing.

2.1.7 SNP Arrays

SNP arrays from 37 FAP adenoma samples (Table 2) were performed using the HumanOmniExpressExome beadChip array (Illumina, San Diego, CA). This array contains 964,193 markers including 273,246 markers specific to the exome. The DNA input for each sample was 200 ng. Intensity files (.idat) were used to make genotyping calls in GenomeStudio software utilizing the Illumina-supplied cluster file.

2.1.8 Colony genotyping assay and power calculation

We performed a colony genotyping assay to determine whether the two *APC* somatic mutations [Somatic #1 (S1) and Somatic #2 (S2)] detected in one of the adenomas belonged to different alleles and therefore to independent clones or if they could be present in the same allele. We performed a long-range PCR (LR-PCR) in the adenoma sample using LaTaq polymerase (TaKaRa Bio Inc, Shiga, Japan). Then, the PCR product was purified from the gel using the QIAquick gel extraction kit (Qiagen), ligated to the plasmid pGEM-Teasy Vector System II with T4 DNA ligase (Promega, Madison, WI) and genotyped using one Taqman assay per mutation (AHN1X23, AHPAV65 and AHZAHK3, Applied Biosystems, Maryland, USA). All samples were genotyped by triplicate and positive and negative controls for all mutations were included in every single plate. Controls were colonies that were validated previously using Sanger Sequencing. We performed a power calculation to determine the minimum number of assays (n) required to observe Somatic #1 and Somatic #2 separately at least 3 times with 95% power and determined that at least 82 clones were needed. The estimation of the tumor purity in our adenoma was consistent with 30% of dysplastic cells and 70% of normal cells[49]. Our Ampliseq sequencing results observed that S1 and S2 have similar allelic frequencies. Based on these premises, the probability of observing a heterozygous S1 (Pr(s1)) and a heterozygous S2 (Pr(S2)) are 0.075. Thus, we can define our power calculation as the following: Probability of observing S1 >2 times and S2 >2 times in n assays.

$$= \mathrm{Pr}(S1 > 2\ and\ s2 > 2)$$

$$= \sum_{i=2}^{n} \mathrm{Pr}(\ S1 > 2\ and\ S2 > i)$$

$$= \sum_{i=2}^{n} \mathrm{Pr}(S1 > 2|S2 = i)\,\mathrm{Pr}(S2 = i)$$

$$= \sum_{i=2}^{n} [1 - \mathrm{Pr}(S1 = 0|S2 = i) - \mathrm{Pr}(S1 = 1|S2 = i) - \mathrm{Pr}(S2 = 2|S2 = i)]\,\mathrm{Pr}(\mathrm{S2 = i})$$

$$\begin{aligned} = \sum_{i=2}^{n} \Bigg[1 - (1 - \mathrm{Pr(S1)})^{n-i} - \binom{n-i}{1} \mathrm{Pr}(S1)^{1}\,(1 - \mathrm{Pr(S1)})^{n-1-i} \\ - \binom{n-i}{2} \mathrm{Pr}(S1)^{2}\,(1 - \mathrm{Pr(S1)})^{n-2-i} \Bigg] \mathrm{Pr(S2 = i)} \end{aligned}$$

2.1.9 Crypt isolation

A total of 18 adenomas and 6 tumors were minced into 3 mm pieces, incubated in 30mM EDTA in HBSS for 20 min, and vigorously shaken to obtain a supernatant enriched for crypts (Table 3). The same day of the sample collection individual crypts were picked up under microscope and were placed into 0.5 mL microfuge tubes. Normal crypts were easily distinguishable from tumor crypts based on their characteristics (normal crypts were small and nonbranched, while tumor crypts were large and displayed sheets of epithelium with thick clusters of different sizes)[50]. This method limits the contamination of non-neoplastic cells by an efficient separation of lamina propria mucosa or stroma[51]. Isolation of individual crypts was verified by microscopy in a selection of cases to ensure that the molecular analysis was representative of individual crypts and that the methodology for isolation was consistent.

Table 3 Number of adenomas, bulk biopsies and crypts isolated for *APC* and *KRAS* mutation analysis.

Patient ID	# Sample	# Bulk Biopsies	# Crypts	# Sample	# Bulk Biopsies	# Crypts
CMMRD1	1	1	10	0	0	0
FAP1	3	8	30	0	0	0
LYN1	0	0	0	1	1	10
MAP1	3	2	30	0	0	0
SP1	0	0	0	1	4	10
SP2	0	0	0	1	6	10
SP3	0	0	0	1	6	10
SP4	0	0	0	1	6	10
SP5	0	0	0	1	5	10
SP6	0	0	0	1	5	10
UFP1	4	5	43	0	0	0
UFP2	7	10	62	0	0	0
total	18			6		

2.1.10 *APC* mutations in bulk and crypt samples

A fragment of 1650 base pairs (bp) of the mutational cluster region (MCR) was amplified and then sequenced in biopsy sections. For crypt analysis, a nested PCR approach was performed to increase the amount of DNA template. All detected alterations were validated in at least 3 independent reactions and verified by two independent reviewers. Normal mucosa biopsies and normal crypts from the same cases were also analyzed in parallel to rule out random acquisition of *APC* mutations or PCR artifacts.

2.1.11 *KRAS* hotspot mutations in bulk and crypt samples

In bulk biopsies, genotyping of *KRAS* hotspot mutations was performed using allele-specific probes for seven *KRAS* mutational hotspots (G12A, G12C, G12D, G12R, G12S, G12V and G13D) in the Digital Array platform (Fluidigm Corporation, South San Francisco, CA; with an analytical sensitivity from 0.05% to 0.1%, depending on the mutation analyzed [52]. In crypts, the same assays were performed using the 48.48 BioMark Dynamic Array (Fluidigm Corporation) after a pre-amplification step. The analytical sensitivity of the Dynamic Array for the detection of the mutant allele oscillated between 5-12.5%. In the absence of detectable DNA contamination, consistent results were obtained in sixtuplicates allowing for robust observations. Median values of replicates were used to quantify the amount of mutant alleles. The quantitation of the signal was extrapolated from standard curves generated from reconstitution experiments that also allowed the definition of conservative thresholds.
Of note, we carefully ruled out the introduction of systematic DNA contamination as a consequence of performing the pre-amplification of DNA using a nested PCR approach by using multiple independent steps. First, mastermix preparation, DNA pipetting, first PCR amplification, and second PCR amplification were all performed in four separate rooms, where different instrumentation was used. Second, we always kept a unidirectional flow for all the samples. Third, controls for contamination were included in every single PCR reaction. Contamination controls for the first external *KRAS* PCR reaction were re-amplified together

including always specific positive and wild-type controls. Fourth, random sets of amplicons were tested specific *KRAS* mutation probes using conventional real-time PCR analysis by LightCycler 480 to add an additional contamination control, always ruling out the presence of contamination. Finally, we performed a total of six replicates for each assay using the 48.48 Dynamic Array on the BioMark platform (Fluidigm Corporation) in order to re-assure on the robustness of our results.

2.1.12 Metapopulations analysis in crypts

The percentage of *KRAS* mutant cells for each crypt was extrapolated from the allele frequencies in each crypt. Hierarchical clustering of crypt data was performed to evaluate the relationship between metapopulations of crypts in adenomas and carcinomas using Euclidean distance for measuring similarity across crypts and complete linkage methods for clustering. The grouping of the metapopulations was performed using the “pvclust” package[53], which generates probability values using the bootstrap resampling technique.

2.2 Immune Profiling

2.2.1 Patients and Samples

A total of 28 colorectal polyps (26 tubular adenomas and 2 hyperplastic polyps) from 21 patients diagnosed with FAP (n=21) and LS (n=11) at UTMDACC and 3 early stage colorectal tumors (one stage I and two stage II) from 3 patients diagnosed with LS obtained by Nouscom SRL from “Regina Elena” National Cancer Institute, Rome, Italy. All of the patients had their diagnosis confirmed by genetic testing performed at a Clinical Laboratory Improvement Amendments (CLIA) laboratory and carried mutations in *APC* (n=10), *MLH1* (n=3), *MSH2* (n=5), *MSH6* (n=5) and *PMS2* (n=1) (Table S1). Polyps were collected at the time of endoscopic surveillance and were either fully resected or biopsied with snare forceps. Tumors were collected at the time of surgical resection. All tissues were immediately flash-frozen or preserved in RNAlater storage solution (Life Technologies, Carlsbad, CA) and then stored at –80°C according to internal protocols. Total RNA was extracted using TRIzol (Invitrogen, Waltham, MA) and the RNeasy kit (Qiagen). RNA was quantified using Qubit (LifeTechnologies) and RNA quality with the Agilent 2100 Bioanalyzer in order to select samples for analysis based on RNA integrity numbers for analysis. The diagnosis of adenomatous versus hyperplastic polyps and tumors was confirmed by an expert gastrointestinal pathologist in all samples that rendered enough tissue for both nucleic acid extraction and pathologic confirmation (Table 4, and Supplementary Table S2**Error! Reference source not found.** in Chang, Kyle, et al. "Immune Profiling of Premalignant Lesions in Patients With Lynch Syndrome." JAMA oncology (2018)). Informed consent was obtained from all individuals, and the IRB approved this study (UTMDACC Protocol ID number: PA12-0327). In addition, a total of 47 CRC (6 hyper-mutants and 41 non-hypermutants) were downloaded from The Genomics Data Commons13 for comparative analysis with carcinomas.

Table 4. Clinical characteristics of FAP and LS patients

Patient ID	Gender	Age	Race	Colorectal Sx	No. of polyps detected	Cancer Dx	Germline mutation	No. of normal mucosa sequenced	No. of adenomas sequenced	No. of tumors sequenced
							Gene			
FAP_1	M	35	W	IRA	>100		APC	1	2	0
FAP_2	M	33	W	IRA	<5		APC	1	2	0
FAP_3	F	42	W	IRA	5		APC	1	2	0
FAP_4	M	42	W	P	>100	HCC, D	APC	1	3	0
FAP_5	F	40	W	IRA	5		APC	1	1	0
FAP_6	M	37	W	IRA	<5		APC	1	1	0
FAP_7	M	65	W	IRA	<50		APC	1	1	0
FAP_8	F	25	AA	NP	>100	D	APC	1	2	0
FAP_9	F	28	W	IRA	>100	EC, OC	APC	1	1	0
FAP_10	F	25	W	P	>100		APC	1	1	0
LS_1	M	53	W	RH	1	CC	MSH6	1	1	0
LS_2	F	46	W	NP	3	EC	MSH2	1	1	0
LS_3	M	52	W	LH	1	CC	PMS2	1	1	0
LS_4	F	37	W	NP	1		MSH2	1	1	0
LS_5	F	53	W	NP	1		MSH2	1	1	0
LS_6	M	58	W	NP	2		MSH2	1	1	0
LS_7	M	76	W	LH	2	CC	MSH2	1	1	0
LS_8	F	68	W	NP	4	EC	MSH6	2	1	0
LS_9	F	62	W	NP	1	EC	MSH6	1	1	0
LS_10	F	43	W	RH	3	CC	MSH6	1	1	0
LS_11	F	63	W	TCR	3	CC, EC	MSH6	1	1	0
LS_12	F	35	W	SC	0	CC	MLH1	1	0	1
LS_13	F	76	W	SC	0	CC, BC, GIST	MLH1	1	0	1
LS_14	F	61	W	RH	0	CC	PMS2	1	0	1

Abbreviations: M, male; F, Female; W, white; AA, African-American; IRA, ileorectal anastomosis; P, pouch; RH, right hemicolectomy; LH, left hemicolectomy; TCR, transverse colon resection; SC, subtotal colectomy; NP, not performed; NA, not available; CC, colon cancer; HCC, hepatocellular carcinoma; HepBI, hepatoblastoma; D, desmoid; EC, endometrial cancer; OC, ovarian cancer. *Denotes Hispanic ancestry.

2.2.2 RNA Sequencing

Sample preparation, library construction, and sequencing were performed at UTMDACC Sequencing Core Facility and the Center for Genomics and Transcriptomics (Tuebingen, Germany). For polyps, transcriptome analysis was performed with the Illumina HiSeq 2000 instrument generating paired-end 75-base pair (bp) reads when 8 samples are multiplexed using barcodes. Each sample had an average of 40 million read pairs and an average alignment rate of 84.83% (SD=2.98%). For tumors, transcriptome analysis was performed with the Illumina HiSeq 4000 instrument generating paired-end 100-bp reads. Each sample had an average of 27.5 million read pairs and an average alignment rate of 76% (SD=0.04%). Reads were mapped to human genome assembly b37 using STAR v2.5.1b_modified in 2-PASS mode. During STAR first-pass, reads are aligned against transcript definition from gencode.v19. Splice junctions output during STAR first-pass were used in STAR second-pass mapping to generate aligned BAM[54].

2.2.3 Gene Expression Analysis

We used rsem-calculate-expression with default options in RSEM v1.2.21[55] to quantify mRNA expression in LS and FAP samples. RSEM transcript counts of all polyp and tumor samples were combined into a matrix and quantile-normalized using batch as a covariate. Normalized counts were transformed into log2 counts-per-million (CPM) with limma_3.30.9[56]. Genes linked to the immune microenvironment of CRC and MMR deficiency were grouped by lineage and/or function (Th1/Tc1, CTL, Th17, Treg, proinflammation, and metabolism) as previously reported[16] to distinguish which genes were differentially expressed on the basis of LS and FAP status. Differential expression of immune-related genes between FAP and LS polyps, and between LS hyper-mutant and LS non-hyper-mutant were assessed with Welch’s t-test and Benjamini & Hochberg multiple corrections. Significant genes were denoted as False Discovery Rate (FDR) <0.05. Differential expression of immune-related genes among FAP, LS polyps and LS carcinomas were performed with one-way ANOVA and

Tukey's test. All statistical analyses were performed with stats package in R version 3.3.1. T-cell signature enrichment score was calculated with GSVA_1.24.1[57] for FAP and LS polyps, LS tumors. Significant differences between groups were performed with one-way ANOVA and Tukey's test.

2.2.4 Mutation Analysis

We followed GATK best practices on mutation discovery with RNA-Seq. We performed mark-duplicates, SplitNCigarReads and base-recalibration with GATK[58] v3.6 on STAR-aligned BAM files. We called somatic mutations with MuTect2[39]. We decided to keep for further analysis as somatic mutation candidates all mutations denoted as "PASS" by MuTect2 and that were not present in the thousand genome phase 3 data[59], the NHLBI GO Exome Sequencing Project (EVS; http://evs.gs.washington.edu/EVS/) or ExAC database [60], for which the reference allele and variant allele count were >2 and the allele frequency >5%. Mutations were annotated by Annovar[41] and Variant Tools[42]. Mutation rates were calculated by dividing the number of somatic mutations by the number of callable bases (defined as >10x in polyp and matched normal mucosa sample). Mutation signatures were detected and plotted with deconstructSigs_1.8.0[44] and hierarchical clustering of premalignant and malignant samples from our cohort and TCGA were performed with the same package.

2.2.5 Neoantigen Discovery

We used seq2HLA v2.213[61] with default settings to generate 4-digits HLA typing for MHC Class I and II on FAP and LS normal mucosa. Then, we ran pvac-seq v4.0.814[62] to generate neoantigen predictions on each polyps or tumor sample using the following parameters: (i) somatic mutation candidates from MuTect2, (ii) class I and II HLA 4-digit typing, (iii) NetMHCpan[63] and NetMHCpanII[64] for class I and II prediction, and (iv) epitope length of 8, 9, 10, 11 amino acids for class I and 15 for class II. We selected those neoantigens that required a binding affinity <500 nM, had allele frequency >5%, were expressed at a level >10

Transcripts per Million (TPM) and were not present in the 1,000 genomes database. We classified neoantigens based on their affinity in the following categories: strong if the binding affinity was <50 nM, medium if affinity was <100 nM, and weak if affinity was <500 nM. We performed canonical pathway comparison of neoantigens genes discovered in FAP and LS polyps with IPA. Selected pathways with P-values <0.05 were plotted. We assessed the difference in the number of neoantigens among FAP, LS non-hypermutant and hyper-mutant polyps, and LS carcinomas using Kruskal-Wallis test and Dunn’s multiple comparison test.

2.3 Consensus Molecular Subtyping

2.3.1 Patients and Samples

We collected adenomatous (N=301) and serrated polyps (N=28 HPs and 60 SSAs) from sporadic and hereditary populations from a variety of sources: (1) an original institutional cohort from UTMDACC, (2) eight publicly available data sets from prior publications (GSE10714[65], GSE19963, GSE4183[66], GSE45270[29], GSE8671[67], GSE79462[68], GSE46513[69], GSE76987[70], GSE88945[71], GSE106500[72], GSE108317), and (3) proprietary data from Janssen Oncology (Table 5). Clinical information and raw gene expression data from publicly available data sets were retrieved from their original publications, data repositories, or by contacting the primary investigators. All patients and samples obtained from hereditary patients of UTMDACC were collected under approval of the institutional review board and written informed consent was obtained from all individuals. Archival samples across the CRC progression axis were obtained by Janssen Oncology from Avaden Biosciences and Asterand Biosciences under approved protocols from several institutions (Table 6).

Table 5 Data sets used in transcriptomic subtype analysis

	Source	**Platform**	**AP**	**HP**	**SSA**	**Total**
GSE10714	public	Affy 2.0	5	11	0	16
GSE19963	public	Affy 2.0	0	5	0	5
GSE4183	public	Affy 2.0	14	0	0	14
GSE45270	public	Affy 2.0	7	0	6	13
GSE8671	public	Affy 2.0	31	0	0	31
Avaden 1	proprietary	Affy PM	60	0	11	71
Avaden 2	proprietary	Affy PM	136	0	9	145
GSE79462	public	Affy PM	9	0	7	16
GSE46513	public	Illumina HiSeq	0	0	7	7
GSE76987	public	Illumina HiSeq	10	10	20	40
MDACC-FAP	public	Illumina HiSeq	16	0	0	16
MDACC-LS	public	Illumina HiSeq	13	2	0	15
		Total	**301**	**28**	**60**	

Abbreviations: AP, adenomatous polyp; HP, hyperplastic polyp; SSA, sessile serrated polyp

Table 6. Clinical and pathological characteristics of patients belonging to cohorts Avaden 1 and Avaden 2

Characteristics		Avaden 1	Avaden 2	Total
Samples	n	71	145	216
Gender	female	36	74	110
	male	35	71	106
Age (yr)	mean	69.63	62.13	64.60
	sd	11.72	13.33	13.27
Pathology	TA	20	26	46
	TVA	28	74	102
	TA with HGD	6	1	7
	TA with CA	2	0	2
	TVA with HGD	4	31	35
	TVA with CA	0	4	4
	SSA	11	9	20
Location	left	19	91	110
	right	45	50	95
	NA	7	4	11
Size (mm)	mean	20.49	12.53	15.12
	sd	13.76	6.80	10.30

Abbreviations: Tubular adenoma, TA; Tubulovillous adenoma, TVA; Sessile serrated adenoma, SSA; High grade dysplasia, HGD; Carcinoma, CA;

2.3.2 Targeted Panel Sequencing

Targeted panel sequencing on *KRAS* and *BRAF* was performed by Janssen Oncology in the Avaden 1 and 2 data set. Sequencing reads were aligned with BWA-mem[37] and mutations were detected with HaplotypeCaller from GATK version 3.4-46-gbc02625[73]. Mutations were annotated with Oncotator version 1.9.6.1[74].

2.3.3 Expression Data Processing

We constructed cohort aggregated gene expression matrices for each data platform prior to downstream analysis. First, for RNAseq data, gene-level counts were generated using RSEM[55] and subsequently log transformed and quantile-normalized. We kept genes that were expressed in at least two-thirds of the samples. Second, for Affymetrix HGU133 Plus 2.0 array data, each studies' CEL files were aggregated and normalized with the single-sample frozen robust multi-array average (RMA) method from fRMA[75]. Probes were annotated with hgu133plus2frmavecs from R Bioconductor[76]. Multiple probes mapping to the same Entrez gene were aggregated using the 1st eigenvector from singular value decomposition (SVD). Third, for Affymetrix HGU133 Plus PM Array, each studies' CEL files were normalized using the RMA method from the Bioconductor affy package[77]. Probes were annotated with HT_HG-U133_Plus_PM.na36.annot.csv from the Affymetrix website. Multiple probes mapped to the same gene are combined using SVD. To correct for systematic batch differences across studies, we applied the ComBat[78] method implemented in the SVA[79] R package including polyp type as a covariate. For Affymetrix data, we used arrayQualityMetrics[80] R package to exclude outliers based on gene expression distribution and either distances between array or MA plot. For RNAseq data, we used a method based on Euclidean distances between samples. Outlier samples were excluded if their summed distances were > 2.5 standard deviations. We removed genes that were not common across the three aggregated expression matrices.

2.3.4 Consensus Molecular Subtype Classification

We performed CMS classification on the polyp samples using the CMSclassifier R package[34]. CMS classification is assigned to the subtype with highest posterior probability. A list of curated gene sets previously identified by the CRC CMS consortium which consisted of ESTIMATE algorithm[81], curated signatures, canonical gene sets, immune activation, and metabolic action were used in the analysis[34]. For each aggregated data set, we first calculated differentially expressed genes between each CMS subtype as compared to other subtypes using limma_3.30.13[56]. Genes were ranked by –log(p-value) * fold change direction (1 if log fold change > 1 and -1 if log fold change <1). Then, we performed gene set enrichment analysis (GSEA) for each subtype using pre-ranked GSEA method from fgsea R package[82] with 10,000 permutations (Table S5-7). Gene sets with Benjamin-Hochberg adjusted p-value < 0.05 were plotted (Fig S2-4). Benjamin-Hochberg adjusted p-value per gene set across aggregated data sets were combined using Fisher's method to produce a combined enrichment analysis. Clinical, pathological and molecular associations with CMS groups were performed using Fisher's exact test by comparing each CMS subtype vs others.

3 CLONALITY IN PREMALIGNANT LESIONS

Chapter 3 CLONALITY IN PREMALIGNANT LESIONS

Content of this chapter is based on the following publications:

Gausachs M, Borras E, **Chang K**, Gonzalez S, Azuara D, Delgado Amador A, Lopez-Doriga A, San Lucas FA, Sanjuan X, Paules MJ, Taggart MW, Davies GE, Ehli EA, Fowler J, Moreno V, Pineda M, You YN, Lynch PM, Lazaro C, Navin NE, Scheet PA, Hawk ET, Capella G, Vilar E. Mutational Heterogeneity in APC and KRAS Arises at the Crypt Level and Leads to Polyclonality in Early Colorectal Tumorigenesis. Clin Cancer Res. 2017 Oct 1;23(19):5936-5947. doi: 10.1158/1078-0432.CCR-17-0821. Epub 2017 Jun 23. PubMed PMID: 28645942; PubMed Central PMCID: PMC5626604.

Borras E, San Lucas FA, **Chang K**, Zhou R, Masand G, Fowler J, Mork ME, You YN, Taggart MW, McAllister F, Jones DA, Davies GE, Edelmann W, Ehli EA, Lynch PM, Hawk ET, Capella G, Scheet P, Vilar E. Genomic Landscape of Colorectal Mucosa and Adenomas. Cancer Prev Res (Phila). 2016 Jun;9(6):417-27. doi: 10.1158/1940-6207.CAPR-16-0081. Epub 2016 May 24. PubMed PMID: 27221540; PubMed Central PMCID: PMC4941624.

Copyright permissions are not required, since AACR states “Authors of articles published in AACR journals are permitted to use their article or parts of their article in the following ways without requesting permission from the AACR - Submit a copy of the article to a doctoral candidate's university in support of a doctoral thesis or dissertation.

3.1 Introduction

Progression along the normal mucosa-adenoma-carcinoma sequence in CRC is fostered by progressive accumulation of genomic events in well-known driver genes with the truncal events being the acquisition of *APC* alterations in approximately 85% of cases [5]. The “Big Bang” model of colorectal carcinogenesis states that the vast majority of the genomic alterations accumulate during early stages of carcinogenesis and assumes that this massive accumulation happens after an initiating driver (*APC)* has occurred in a single crypt (founder clone)[83, 84]. Subsequent acquisition of driver mutations, such as *KRAS,* foster tumor progression and generation of subclones that will acquire additional mutations. Therefore, the “big bang” model of colorectal carcinogenesis implies a “monocryptal polyclonal” origin. However, the presence of multiple founder clones (lineages) from distinct crypts observed in both animal models and human samples [85-87] and challenges the monocryptal founder clone as origin of ITH.

Next-generation sequencing technologies and bioinformatics tools have been widely adopted to deconvolve ITH[26, 27]. I hypothesize that inference on the number of tumor populations based on somatic mutations will determine the presence of clonality in premalignancy. In addition, I hypothesize origin of clonality in premalignant lesions is derived from multiple independent crypts that acquire distinct alterations in *APC* and *KRAS*.

3.2 Results

3.2.1 Evidence of clonality in WES

I searched for evidence of clonality in *APC*-driven adenomas by visually inspecting the distribution of mutations ranked by allelic frequency. I observed 2 clusters of mutation allelic frequencies in 5 samples, suggesting the presence of 1 major clone and other minor subclones (Figure 1A). However, this type of analysis could be confounded by tumor purity and copy number variation. To mitigate the effect of variations in these 2 factors, I applied computational tools to transform the frequency distribution of mutations in each sample and to infer the number of clones. I compared *in silico* estimations of purity, ploidy, and number of clones between adenomas and stage I CRC. The purity was significantly lower in adenomas compared with stage I CRCs (0.35 vs. 0.58, *P* < .0001; Figure 1B), as was the ploidy (1.93 vs. 2.61, *P* < .0001; Figure 1C). Then, my analysis revealed the presence of multi-clonality in 18 (72%) of 25 adenomas, with at least 1 major and 1 minor subclone per lesion. Moreover, more than 50% of the polyps from the same individual has estimated to have multiclonality. Interestingly, the number of clones estimated in adenomas was not significantly different from that for stage I CRC (1.72 vs. 2.06, *P* = .165; Figure 1D).

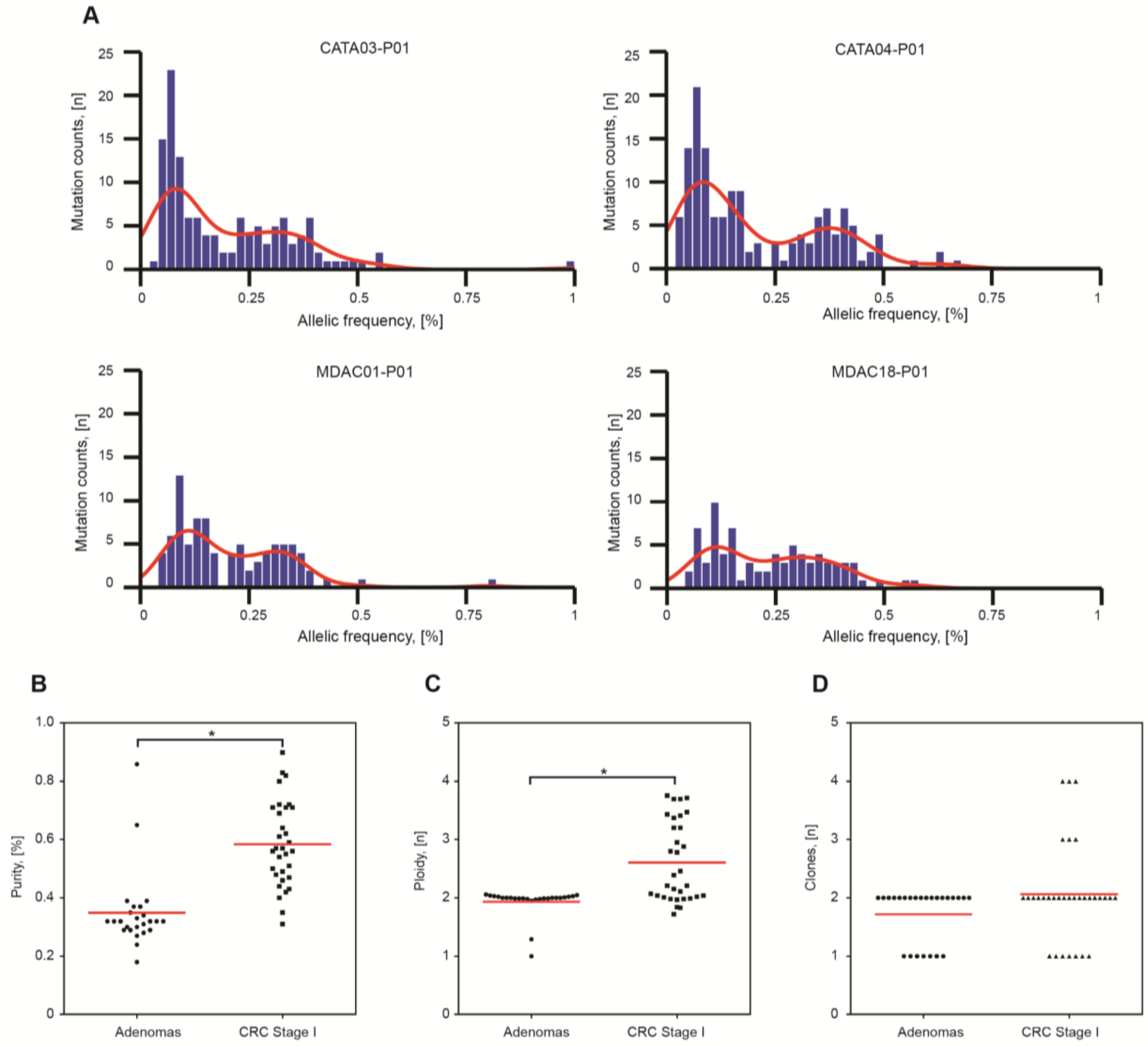


Figure 1. Clonality analysis of colorectal adenomas. A. Mutation counts detected by whole-exome sequencing in 4 different adenomas, ordered by allelic frequency and provided as examples presenting evidence of clonality. **B, C.** Purity and ploidy of adenomas and stage I CRCs were estimated by the ABSOLUTE computational method (*P < .0001). **D.** Numbers of clones in adenomas were inferred by clustering cancer cell fraction of mutations estimated by ABSOLUTE.

3.2.2 Multiple somatic events detected in FAP Adenomas with high depth sequencing

Next-generation sequencing at high depth using Ion Torrent AmpliSeq was performed in a total of 37 adenomas from 14 patients diagnosed with FAP to detect public somatic mutations in *APC* and *KRAS* (**Figure 2**). An advantage of studying adenomas of FAP patients with confirmed germline alterations of *APC* to detect hints of multiclonality is that one *APC* event is already given from the germline, so any new additional *APC* events will be acquired (somatic events under the assumption that adenomas are diploid). Therefore, accumulation of multiple somatic events either by mutation or allelic imbalance (loss of 5q) will be suggestive of polyclonality. The average sequencing depth obtained per sample for *APC* and *KRAS* was 3,640x and 6,685x, respectively. I identified a total of 32 somatic events in *APC* with an average allelic frequency of 21%. The majority were mutations generating a stop codon (29 out of 32); although we detected other events such as a splicing mutation (c.1744-1G>A) and two missense mutations (c.2438A>G; p.Asn813Ser and c.2258A>G; p.His753Arg), which were all predicted to have a functional impact (). Using SNP arrays, loss of 5q was identified in 4 adenomas with an average allelic frequency of 4% (Table 7). Overall, I was able to detect a second hit in *APC* in 81% of adenomas (30/37). Six adenomas (16%) presented double somatic events in *APC* in the form of two different mutations or the combination of one mutation associated with a 5q loss (Table 7). In addition, I detected 15 *KRAS* somatic mutations in 14 adenomas with an average allelic frequency of 29%.

One adenoma presented two different *KRAS* mutations: c.35G>A (p.Gly12Asp) and c.40G>A (p.Val14Ile) with allelic frequencies of 5% and 16%, respectively. Interestingly, this adenoma also presented 2 independent somatic events in *APC*: 5q loss and c.4464_4467delInsGTAAT (p.Leu1489*) (Table 7). *APC* alterations detected in one adenoma (MDAC33_P04) were selected to validate the multiclonal origin of these events using colony analysis. Of note, this adenoma was euploid and the region of *APC* containing the two somatic mutations (c.4348C>T and c.4267_4280del) and the germline mutation (c.2894delA) was amplified and cloned into pGEM-T plasmid. A total of 100 colonies were genotyped and each colony carried a single mutational event: 25% were wild-type, 36% carried the germline mutation (c.2894del), 15% one of the somatic hits c.4348C>T and 24% the other somatic hit c.4267_4280del (Figure 3). These cloning results confirmed that the somatic *APC* mutations arose from separate and independent clones (polyclonal adenomas).

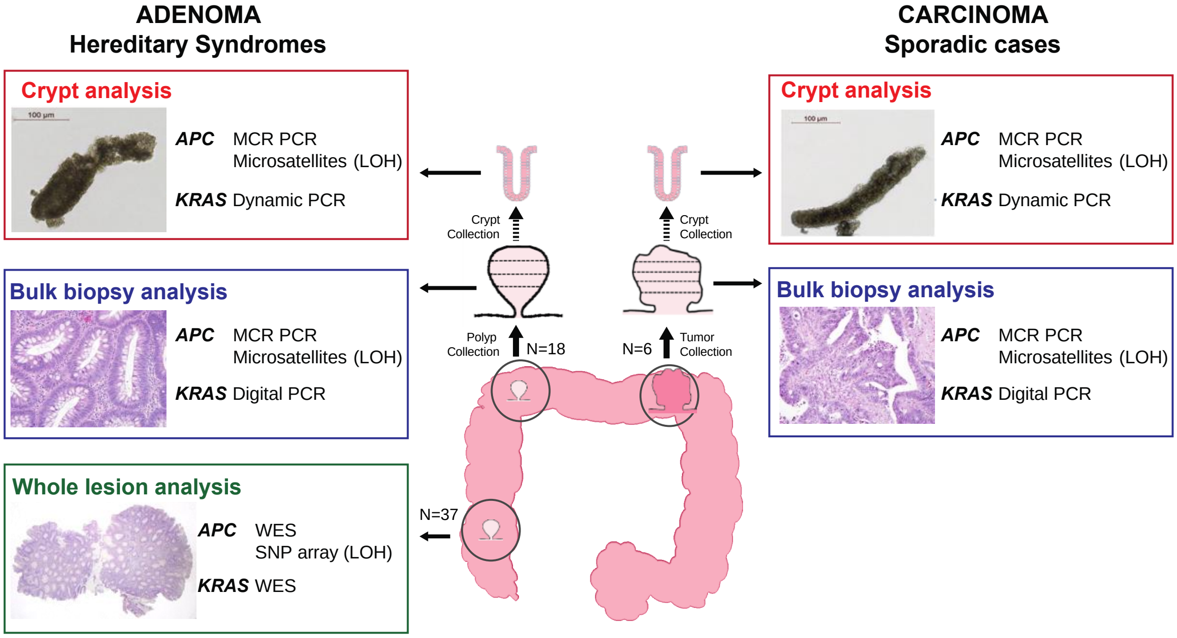


Figure 2. Experimental design and molecular analysis performed in bulk biopsies, crypts and whole lesion extracts of adenomas and colorectal carcinomas. Adenoma analyses were performed only in samples obtained from patients diagnosed with hereditary colorectal cancer syndromes. Carcinoma analysis were performed in a total of 6 tumors from sporadic cases and one tumor from a hereditary case. Note that both adenomas and carcinomas were analyzed from different locations of the colon. WES, whole exome sequencing.

Table 7. Analysis of *APC, KRAS* and 5q loss using AmpliSeq and SNP array.

Sample	*APC* somatic alterations		Frequency	*KRAS* somatic alterations		Frequency
	cDNA	protein		cDNA	protein	
MDAC02_P02	c.2438A>G	p.Asn813Ser	0.11			
	c.4348C>T	p.Arg1450*	0.12			
MDAC02_P03	c.4234G>T	p.Gly1412*	0.07			
MDAC03_P01				c.351A>T	p.Lys117Asn	0.4
MDAC03_P02	c.4348C>T	p.Arg1450*	0.24	c.38G>A	p.Gly13Asp	0.31
MDAC03_P03	c.4348C>T	p.Arg1450*	0.16	c.38G>A	p.Gly13Asp	0.18
MDAC10_P02	c.4330C>T	p.Gln1444*	0.08			
MDAC14_P04	c.4189G>T	p.Glu1397*	0.07			
MDAC14_P05	c.4219_4220del	p.Ser1407*	0.13			
MDAC17_P03	c.3925_3928del	p.Glu1309Argfs*11	0.13			
MDAC17_P04						
MDAC18_P02	c.1660C>T	p.Arg554*	0.06			
MDAC20_P02						
MDAC21_P01				c.35G>T	p.Gly12Val	0.09
MDAC24_P01						
MDAC24_P02	5q loss		0.02			
MDAC24_P03	c.2258A>G	p.His753Arg	0.04	c.35G>T	p.Gly12Val	0.37
MDAC24_P04	c.847C>T	p.Arg283*	0.23			
MDAC24_P05	c.1744-1G>A	p.?	0.13			
	c.3340C>T	p.Arg1114*	0.04			
MDAC26_P01	c.4135G>T	p.Glu1379*	0.46	c.35G>T	p.Gly12Val	0.09
	5q loss		0.03			
MDAC26_P02	c.4464_4467delinsGTAAT	p.Leu1489*	0.36	c.40G>A	p.Val14Ile	0.05
	5q loss		0.04	c.35G>A	p.Gly12Asp	0.16
MDAC26_P03	c.4508C>A	p.Ser1503*	0.37			
MDAC26_P04	c.4099C>T	p.Gln1367*	0.45			
	5q loss		0.06			
MDAC26_P05	c.4348C>T	p.Arg1450*	0.39	c.35G>A	p.Gly12Asp	0.31
MDAC29_P01	c.4037C>A	p.Ser1346*	0.30			
MDAC29_P02	c.4037C>A	p.Ser1346*	0.06			
MDAC32_P01						
MDAC32_P02	c.3916G>T	p.Glu1306*	0.27			
MDAC32_P03	c.4348C>T	p.Arg1450*	0.27	c.35G>T	p.Gly12Val	0.46
MDAC33_P01	c.4267_4280del	p.Leu1423Trpfs*10	0.40	c.57_58delinsTT	p.Leu19_Thr20delinsPheSer	0.62
MDAC33_P02	c.4267_4280del	p.Leu1423Trpfs*10	0.40	c.57_58delinsTT	p.Leu19_Thr20delinsPheSer	0.57
MDAC33_P03	c.4749_4754delinsCACGT	p.Met1583Ilefs*3	0.33	c.436G>A	p.Ala146Thr	0.09
MDAC33_P04	c.4267_4280del	p.Leu1423Trpfs*10	0.14	c.57_58delinsTT	Leu19_Thr20delinsPheSer	0.26
	c.4348C>T	p.Arg1450*	0.16			
MDAC33_P05	c.4267_4280del	p.Leu1423Trpfs*10	0.32	c.57_58delinsTT	Leu19_Thr20delinsPheSer	0.35
MDAC34_P01	c.3766C>T	p.Gln1256*	0.08			
MDAC34_P02	c.4348C>T	p.Arg1450*	0.15			
MDAC34_P03	c.3340C>T	p.Arg1114*	0.17			
MDAC34_P04						

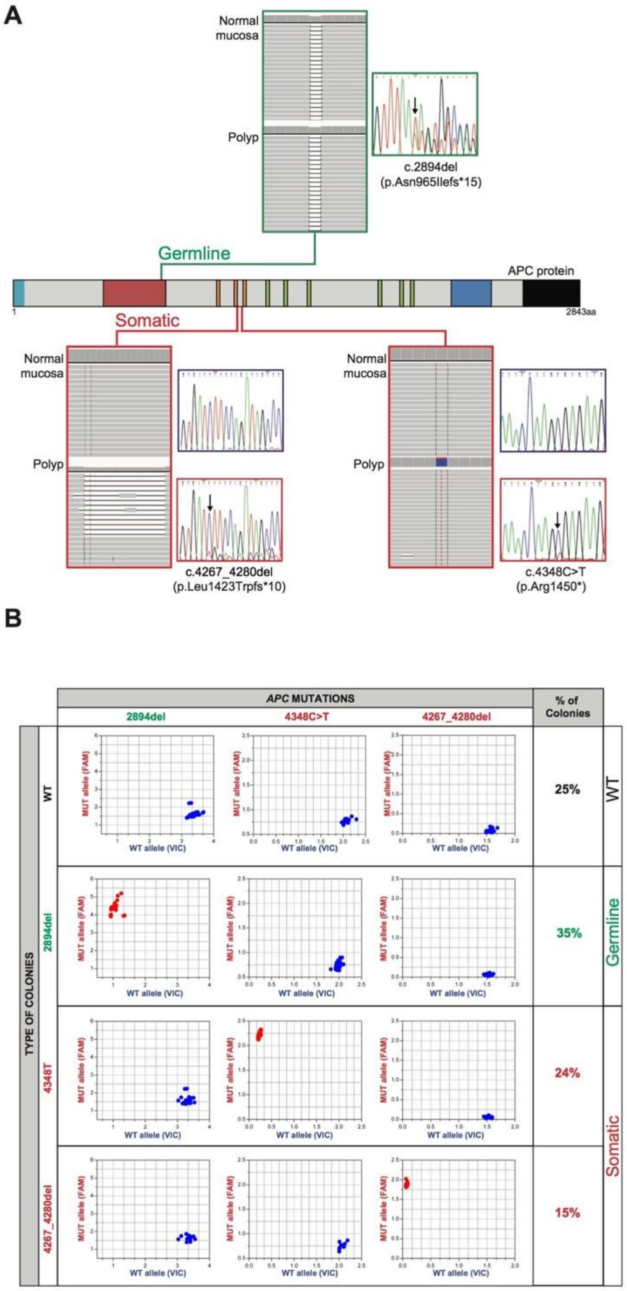


Figure 3. A. Schematic representation of the multiple APC mutations identified in the adenoma selected for assessment of polyclonality by a cloning approach. The germline event is depicted in green and the somatic events are depicted in red; **B.** Frequency of colonies harboring wild-type (wt), germline, and somatic mutated genotypes (mut). X-axis represents intensity of FAM fluorescence (wt allele) and Y-axis VIC fluorescence (mut allele). On the scatter plot diagrams, end signals from each sample have been presented as a single dot. (Cloning experiment was performed by Ester Borras.)

3.2.3 Paired analysis of multi-region biopsies and crypts of adenomas and carcinomas reveal high degree of heterogeneity

Next-generation sequencing results from whole lesion DNA extracts provided evidence that the mutational heterogeneity observed in adenomas was derived from multiple different clones containing distinct mutations in truncal driver genes. This leads to a hypothesis that ITH emerges early on colorectal carcinogenesis due to the interaction of multiple mutated crypts (independent lineages) that is not captured by bulk biopsies, thus requiring individual crypt analysis. Therefore, I decided to assess the presence of multiple somatic events in paired multi-region bulk biopsies and crypts extracted from the same lesions by applying ultrasensitive genotyping techniques within mutational hotspots of the *APC* and *KRAS* genes.

APC analysis. Multi-region biopsy analysis revealed the presence of somatic *APC* mutations in almost all of the regions of 2 adenomas from the FAP and UFP2 cases (Figure 4, and Supplementary Figure S3 and Table S7 in Gausachs M, Chang K, et al. "Mutational Heterogeneity in *APC* and *KRAS* Arises at the Crypt Level and Leads to Polyclonality in Early Colorectal Tumorigenesis". Clin Cancer Res. (2017)). Crypt analysis of these adenomas revealed mixtures of novel somatic *APC* alleles, which were not present in the bulk biopsies, and also an abundance of wild-type alleles of the MCR of *APC*. The percentage of crypts with *APC* mutations varied across adenomas but overall there was a predominance of wild-type ones (average 20%, range 13-100%). In contrast, adenomas from the MAP and UPF1 cases showed a good correlation between bulk biopsy and crypt analyses with all the biopsies and crypts displaying a wild-type status for the MCR region of *APC* (Figures S1-2 and Table S7 in Gausachs M, Chang K, et al. "Mutational Heterogeneity in *APC* and *KRAS* Arises at the Crypt Level and Leads to Polyclonality in Early Colorectal Tumorigenesis". Clin Cancer Res. (2017)).

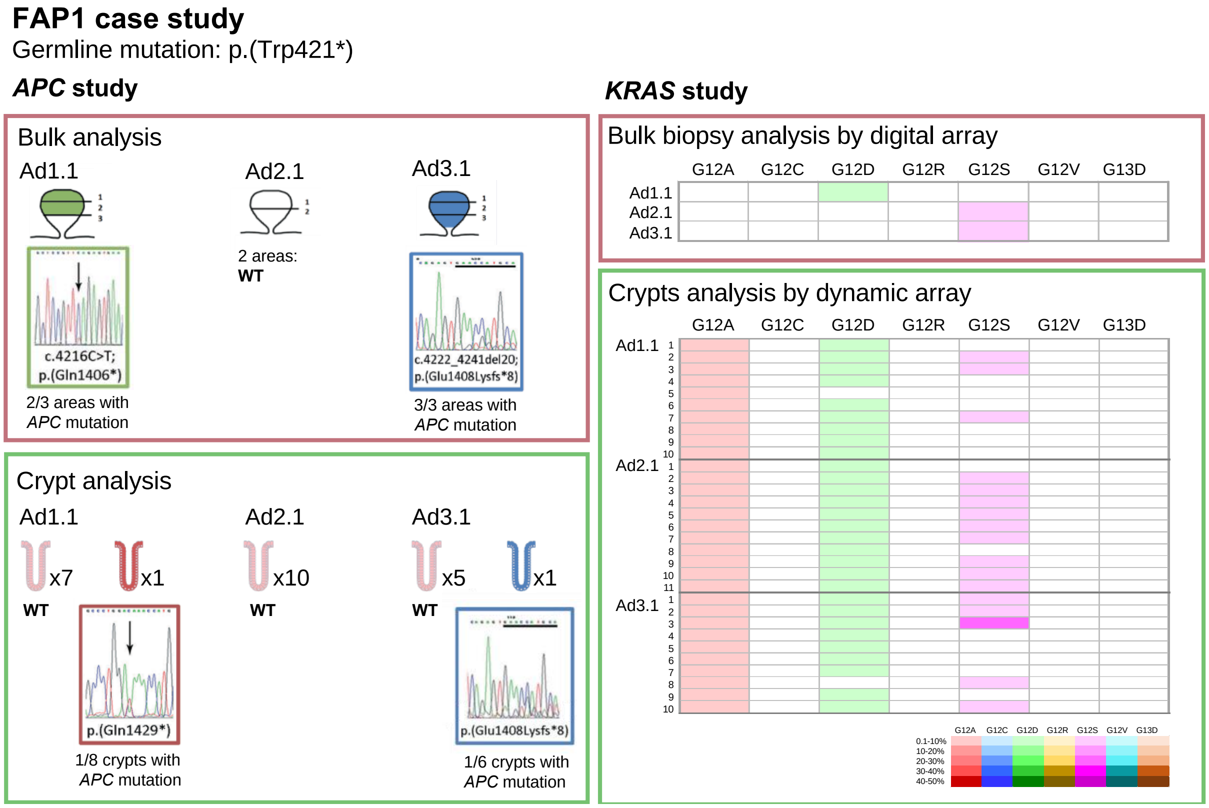


Figure 4. APC and KRAS genotyping of crypts and bulk biopsies in adenomas from the FAP case. On the left side is displayed the study of the mutator cluster region (MCR) of *APC* performed in 3 adenomas (Ad1.1, Ad2.1, and Ad3.1). The upper panel shows the results of the bulk biopsy analysis and the lower panel the results obtained from crypts that correspond to the same lesions. On the right side are presented the results of the *KRAS* genotyping. Each column represents one of the *KRAS* hotspot mutations that were tested by the digital or the dynamic array, and each row represents the results of bulk biopsies (one biopsy per adenoma, upper panel) or single crypts (10 crypts per adenoma, lower panel). The upper panel shows the digital PCR results from bulk biopsies, which identify a single *KRAS* mutation at low frequency in each adenoma.

KRAS analysis. In general, all bulk biopsies analyzed in adenomas from the MAP, UFP1 and UFP2 cases unveiled multiple co-existing *KRAS* mutations (mutational load average 11%) while FAP adenomas only harbored single mutations (Figure 4, Supplementary Figure S1-3 and Table S8 in Gausachs M, Chang K, et al. "Mutational Heterogeneity in *APC* and *KRAS* Arises at the Crypt Level and Leads to Polyclonality in Early Colorectal Tumorigenesis". Clin Cancer Res. (2017)). However, when I studied crypts obtained from the same lesions a striking intra-crypt mutational heterogeneity was evident in 76% of adenomas (13/17). An average of 44% of the crypts per adenoma displayed multiple *KRAS* mutations (range 10-100%) and the proportion of mutant alleles was much higher compared to the bulk biopsies (average 26%, versus 7.6%). Moreover, I also observed *KRAS* mutational heterogeneity in the matched surrounding normal mucosa that may reflect a ‘genetic field effect’, which may be acquired prior to the *APC* alterations or as a consequence of underlying deficiencies in DNA repair that are the basis of these syndromes.

A high great degree of consistency between the results of bulk biopsies and crypts in terms of *KRAS* mutational diversity and proportion of mutated alleles was observed among carcinomas (29% in biopsies versus 22% in crypts). In this regard, no additional mutations were detected in the crypts analyzed. However, in terms of mutational load, the proportion of *KRAS* mutant alleles among crypts was variable (Figure 5, supplementary Table S8 in Gausachs M, Chang K, et al. "Mutational Heterogeneity in *APC* and *KRAS* Arises at the Crypt Level and Leads to Polyclonality in Early Colorectal Tumorigenesis". Clin Cancer Res. (2017)). Intriguingly, several carcinomas (SP2, SP3 and SP6) have a variable fraction of wild-type *KRAS* crypts that co-existed with crypts with a high *KRAS* mutational load, depicting another level of genetic heterogeneity.

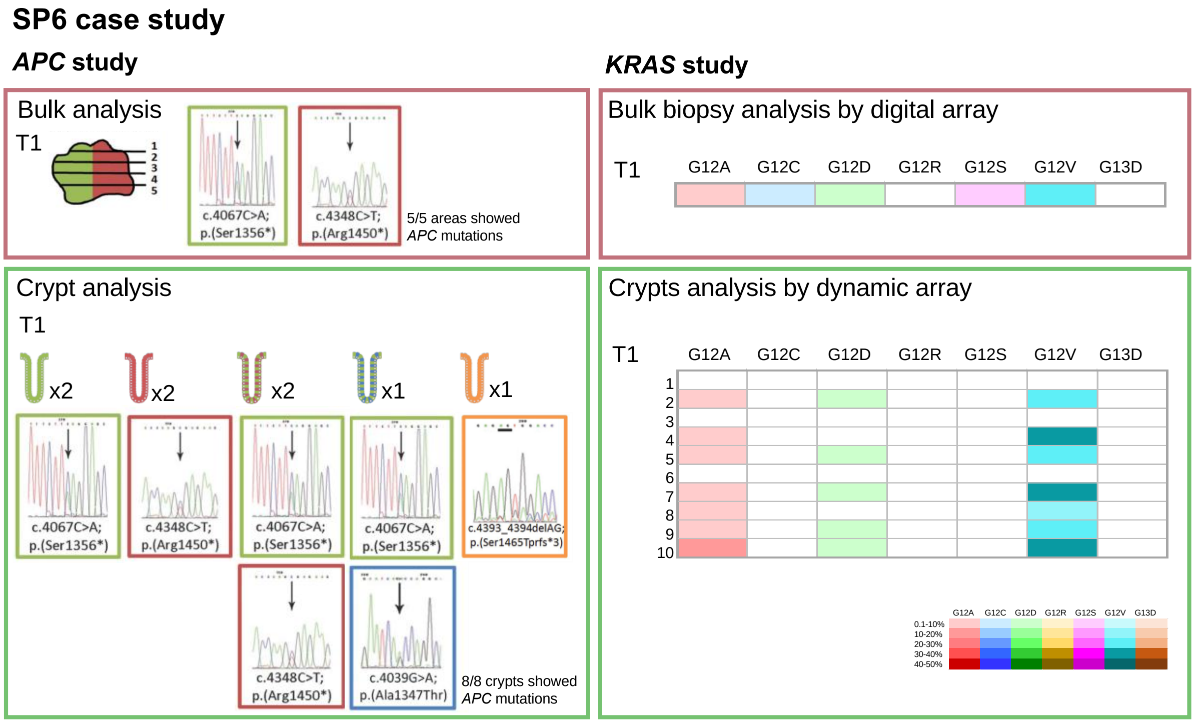


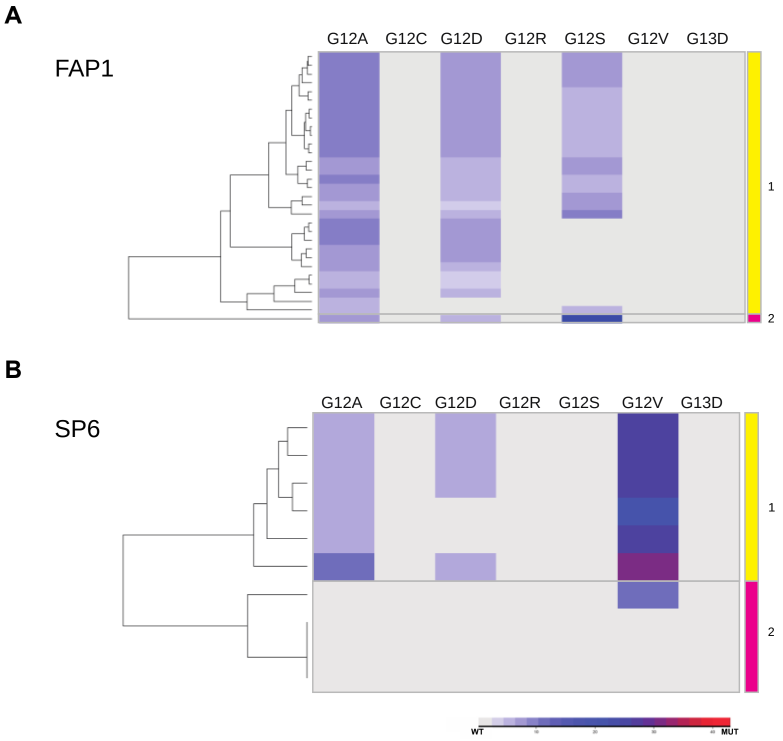
Figure 5. APC and KRAS genotyping of crypts and bulk biopsies from sporadic colorectal carcinoma case #6 (SP6).

In an effort to provide a complete picture of the evolutionary dynamics of *APC* and *KRAS* mutations in all pathways of carcinogenesis in CRC, two different hereditary cases displaying MMR deficiency were included. We analyzed one adenoma from a patient with CMMR-D and one carcinoma from an unrelated Lynch Syndrome patient (Table S2). The bulk biopsy analysis in these two samples was limited to one biopsy rather than being multiregional as in the rest of hereditary adenomas and sporadic tumors included in this report. The analysis of the MCR of *APC* in both cases did not reveal any mutation at the biopsy and crypt level (Figure S4 and Table S7 in Gausachs M, Chang K, et al. "Mutational Heterogeneity in *APC* and *KRAS* Arises at the Crypt Level and Leads to Polyclonality in Early Colorectal Tumorigenesis". Clin Cancer Res. (2017)). Absence of heterogeneity with a predominance of wild-type *KRAS* crypts was observed in the adenomas while the carcinoma showed a relatively low level of heterogeneity among the crypts (Figure S4 and Table S8 in Gausachs M, Chang K, et al. "Mutational Heterogeneity in *APC* and *KRAS* Arises at the Crypt Level and Leads to Polyclonality in Early Colorectal Tumorigenesis". Clin Cancer Res. (2017)). Therefore, *APC* and *KRAS* analysis did not capture mutational heterogeneity in MMR deficient lesions but these results do not rule out that polyclonality that may be observed by alternative studies analyses other drivers more relevant to MMR deficient carcinogenesis. Nonetheless, these results reinforce the robustness of our technical approach as it did not observe any random variation among the MMR deficient samples, thus providing a negative control.

3.2.4 Hierarchical clustering of crypts reveals a striking degree of non-random inter-crypt heterogeneity with several crypt metapopulations

Unsupervised clustering of all crypts based on *KRAS* mutations for each patient revealed a non-random pattern (Figure 5 and Figures S8-11). A discrete and similar number of metapopulations was evident in both adenomas (mean 3.80 metapopulations per case) and carcinomas (mean 2.84 metapopulations per case). This fact reflects that a stable number of founder clones emerge early in carcinogenesis and remains stable during progression of carcinogenesis.

Crypts from MAP, UFP1 and UFP2 cases clustered around 9, 2 and 4 metapopulations, respectively, compared to FAP, which showed 2 main ones (Figure 5 and Figure S8). The difference observed between the number of clones in MAP and FAP may be related to the mechanism of genetic instability (driven by deficiency in base excision repair versus chromosomal instability). Particularly striking was the fact that crypts from adenomas that were physically located at different parts of the colon clustered together, once again pointing towards the existence of a ‘genetic field’ effect.



3.3 Discussions

In this chapter, I used computational tools to identify the presence of polyclonality in FAP adenomas. The number of tumor populations between adenomas and stage I CRC is not significantly different, which suggests that polyclonality originates early in carcinogenesis and not from late clonal expansions. This observation is consistent with the “big bang” model of colorectal carcinogenesis.

In addition, I have uncovered the presence of multiple co-occurring somatic *APC* mutations using whole lesion extracts from colorectal adenomas of FAP patients, which is consistent with previous observations[88]. However, using cloning approaches we have confirmed for the first time that they belong to independent clones. Furthermore, the fact that is observed only in pre-malignant samples, which have not acquired yet chromosomal instability and are still diploid, confirms that the theory of the presence of ‘three-hits’ inactivating *APC* is still plausible in carcinomas that may have acquired multiple copies of *APC*[89].

Furthermore, I have detected the presence of mutational heterogeneity in *APC* at the crypt level among early stage adenomas with only a minority of the crypts harboring the clonal *APC* mutation identified in their paired bulk biopsies. The majority of crypts from adenomas were *APC* wild-type. Both results are in line with previous observations made in more advanced lesions such as adenoma-in-carcinoma samples, where the degree of heterogeneity was higher, and only a minority of crypts were wild-type[87]. This is not unexpected and reflects further progression into the clonal evolution presented in more advanced stages of carcinoma progression. The fact that we have consistently identified wild-type *APC* crypts in the intraluminal surface of carcinomas and adenomas distal to normal mucosa is compatible with the coexistence of normal crypts inside of the dysplastic cancerous mass that, based on the work of Thliveris et al, could represent wild-type crypts recruited by the mutant *APC* founder clone to foster neoplastic growth through transformation later on[87].

The *KRAS* genotyping analysis using digital PCR has detected a striking degree of heterogeneity at the biopsy and crypt levels within adenomas and carcinomas, which is in line previous results that used less sensitive techniques with lower resolution[52, 90]. Clustering analysis has revealed that the observed KRAS mutational pattern in crypts was not random and depicted the presence of a limited number of crypt metapopulations in adenomas and carcinomas. It was notable that crypts obtained from distinct adenomas arising from hereditary patients displayed a similar mutational pattern, irrespective of the location of the lesion consistent with a ‘genetic field effect’. There are several factors that could explain this observation: (i) the asymmetric expansion associated to crypt fission or other mechanisms of crypt interactions yet to be determined[91, 92]; (ii) the underlying mechanism of genetic instability present in hereditary cases such as chromosomal instability in FAP, base excision repair deficiency in the case of MAP, and MMR deficiency in Lynch syndrome and CMMRD (38), the latter being notorious for an absence of *KRAS* heterogeneity. However, we cannot rule out that at least part of these *KRAS* mutations emerge prior to the acquisition of *APC* mutations and the establishment of a founder clone. In fact, *KRAS* mutations have been detected in normally appearing tissues and in aberrant crypt foci that may not progress into adenomas (39). Lastly, this diversity of KRAS mutations acquired early in carcinogenesis has multiple ramifications for the design of chemoprevention strategies and to explain the emergence of resistance to therapies.

Our study has several limitations. First, the mechanical isolation of crypts is subject to sampling bias and selection of specific cell populations. However, we believe that we have minimized this problem by involving dedicated expert gastrointestinal pathologists to collect samples following established procedures that are unlikely to be contaminated by normal crypts. Second, performing genomic analysis in relatively small groups of cells is challenging due to the limited amounts of DNA rendered. We have evaluated different approaches (whole genome amplification versus nested PCR) to amplify the nucleic acid material prior to our ultrasensitive genotyping and demonstrated that our approach using nested PCR is capable to render robust analytical results. Furthermore, all of our tests have been performed in sextuplicate and included multiple internal controls, thus minimizing uncertainty. Therefore, we are confident that this rigorous and reproducible approach minimizes concerns for contamination. Third, in our crypts analysis we have only assessed the MCR of APC and therefore the possibility that we have missed mutations located outside of this region reflecting additional heterogeneity is obvious. In the future, single cell sequencing analysis will be the most appropriate tool to interrogate the dynamics and clonal heterogeneity of crypts in premalignancy, thus helping to clarify the level of mutational diversity within the crypt and establishing a cellular hierarchy based on mutational events[93].

Taken together, the body of evidence presented here demonstrates that the presence of ITH in colorectal premalignancy is abundant and secondary to the presence of multiple independent lineages derived from crypt progenitors, which is highlighted by the mutational heterogeneity detected in *APC* and *KRAS*.

4 IMMUNE PROFILING OF PREMALIGNANT LESIONS IN PATIENTS WITHLYNCH SYNDROME

Chapter 4 IMMUNE PROFILING OF PREMALIGNANT LESIONS IN PATIENTS WITHLYNCH SYNDROME

The content of this chapter is based on the following publication[72]:

Chang K, Taggart MW, Reyes-Uribe L, Borras E, Riquelme E, Barnett RM, Leoni G, San Lucas FA, Catanese MT, Mori F, Diodoro MG, You YN, Hawk ET, Roszik J, Scheet P, Kopetz S, Nicosia A, Scarselli E, Lynch PM, McAllister F, Vilar E. Immune Profiling of Premalignant Lesions in Patients With Lynch Syndrome. JAMA Oncol. 2018 Apr 16. doi: 10.1001/jamaoncol.2018.1482. PubMed PMID: 29710228.

4.1 Introduction

LS is the most common hereditary CRC syndrome and represents a model to study carcinogenesis in the background of DNA mismatch repair deficiency, which is the basis of approximately 15% of sporadic CRC. MMR deficiency causes an excessive number of frame-shift mutations that generates neoantigens and infiltration by immune T-cells. Neoantigens are thought to induce an upregulation of checkpoint molecules such as PD-1, PD-L1, and LAG3 to counter balance infiltrating T-cells and allow the continual progression of tumor[16]. In fact, checkpoint inhibitors Pembrolizumab and Nivolumab have demonstrated clinical utility in extending progression free and overall survival of patients with MMR-deficient tumors[18-20]. In order to investigate the potential opportunity for immuno-prevention in LS patients, we proposed this study to assess the expression levels of immune checkpoints and T-cell infiltration in premalignant lesions of LS patients. We hypothesized that MMR-deficient lesions display an up-regulation of immune checkpoints compared to MMR-proficient lesions. I addition, we proposed that mutational load and neoantigen formation arise late in MMR-deficient carcinogenesis (either advanced premalignant lesions or carcinomas) and they are independent of the immune-profile displayed by premalignant lesions.

4.2 Results

4.2.1 LS premalignancy display a unique immune signature

RNA-sequencing (RNA-seq) was performed in a total of 28 colorectal polyps (Table 8). All of the polyps analyzed from FAP patients (n=17) were confirmed to be early tubular adenomas, smaller than 1 cm in diameter, and without signs of high-grade dysplasia. All LS polyps (n=11) were early adenomas of 1 cm in diameter, with the exception of 2 that were hyperplastic polyps. A total of 4 LS polyps displayed MMR deficiency by loss of staining in MSH2 and/or MSH6, and the rest were MMR proficient (Table 9). Overall, LS polyps showed a significantly higher expression of *CD4*, *IFNG*, *LAG3* and *CD274/PDL1* (Checkpoints), *IL12A* and *TNF* (Pro-inflammatory) when compared with FAP polyps and displayed a consistent trend among the genes integrated in these pathways regardless of their MMR status (Figure 6). Interestingly, *LAG3* was observed to be the most significantly upregulated. Then, we analyzed the evolution of immune activation in MMR carcinogenesis by comparing LS polyps to carcinomas and observed additional consistent activation among genes in the Proinflammatory and Metabolism pathways that were absent in premalignancy (Figure 7A). Of note, LS premalignancy showed activation of both *PD-L1* and *LAG-3* and carcinomas showed deregulation of additional checkpoints such as *CTLA4*. This expression pattern displayed by LS polyps and carcinomas is consistent with a strong enrichment for additional immune related gene sets such as Immune Activation, Immune Response, PD-1 activation and T-cell reaction (Figure 7B)[94]. These results suggest the existence of a robust immune microenvironment in LS premalignancy secondary to T-cell infiltration[16].

Table 8 Clinical characteristics of FAP and LS patients

Patient	Gender	Age	Race	Colorectal Sx	Polyps detected	Cancer Dx	Germline mutation	
							Gene	gDNA
FAP_1	M	35	W	IRA	>100		APC	c.1880dupA
FAP_2	M	33	W	IRA	<5		APC	c.3810T>A
FAP_3	F	42	W	IRA	5		APC	del exon 14
FAP_4	M	42	W	P	>100	HCC, D	APC	c.622C>T
FAP_5	F	40	W	IRA	5		APC	del exons 8-9
FAP_6	M	37	W	IRA	<5		APC	c.1658G>A
FAP_7	M	65	W	IRA	<50		APC	c.477C>G
FAP_8	F	25	AA	NP	>100	D	APC	c.4733_4734del
FAP_9	F	28	W	IRA	>100	EC, OC	APC	c.847C>T
FAP_10	F	25	W	P	>100		APC	c.3810T>A
LS_1	M	53	W	RH	1	CC	MSH6	c.3744_3773del30
LS_2	F	46	W	NP	3	EC	MSH2	c.687delA
LS_3	M	52	W	LH	1	CC	PMS2	del exon 14
LS_4	F	37	W	NP	1		MSH2	c.1034G>A
LS_5	F	53	W	NP	1		MSH2	c.1661+1G>A
LS_6	M	58	W	NP	2		MSH2	del exons 1-6
LS_7	M	76	W	LH	2	CC	MSH2	c.1216C>T
LS_8	F	68	W	NP	4	EC	MSH6	c.3238_3239delCT
LS_9	F	62	W	NP	1	EC	MSH6	c.3860ins4
LS_10	F	43	W	RH	3	CC	MSH6	c.2645_2653delTTAAGTCTA
LS_11	F	63	W	TCR	3	CC, EC	MSH6	c.3699_3702delAGAA
LS_12	F	35	W	SC	0	CC	MLH1	c.1279C>T
LS_13	F	76	W	SC	0	CC, BC, GIST	MLH1	c.1918C>T
LS_14	F	61	W	RH	0	CC	PMS2	c.2059C>T

Abbreviations: M, male; F, Female; W, white; AA, African-American; IRA, ileorectal anastomosis; P, pouch; RH, right hemicolectomy; LH, left hemicolectomy; TCR, transverse colon resection; SC, subtotal colectomy; NP, not performed; NA, not available; CC, colon cancer; HCC, hepatocellular carcinoma; HepBl, hepatoblastoma; D, desmoid; EC, endometrial cancer; OC, ovarian cancer. *Denotes Hispanic ancestry.

Table 9 Pathology characteristics of FAP and LS polyps

Patient ID	Sample ID	Type	Pathology	Location	Size/TNM	IHC-MSH2	IHC-MSH6	IHC-MLH1	IHC-PMS2
FAP_1	FAP_G2	FAP	TA	R	5-10	NA	NA	NA	NA
FAP_1	FAP_G3	FAP	TA	R	5-10	NA	NA	NA	NA
FAP_2	FAP_G53	FAP	TA	R	<5	NA	NA	NA	NA
FAP_2	FAP_G55	FAP	TA	R	<5	NA	NA	NA	NA
FAP_3	FAP_G5	FAP	TA	R	<5	NA	NA	NA	NA
FAP_3	FAP_G6	FAP	TA	R	<5	NA	NA	NA	NA
FAP_4	FAP_G13	FAP	TA	R	5-10	NA	NA	NA	NA
FAP_4	FAP_G14	FAP	TA	R	5-10	NA	NA	NA	NA
FAP_4	FAP_G15	FAP	TA	R	5-10	NA	NA	NA	NA
FAP_5	FAP_G25	FAP	TA	R	<5	NA	NA	NA	NA
FAP_6	FAP_G29	FAP	TA	R	NA	NA	NA	NA	NA
FAP_7	FAP_G39	FAP	TA	R	<5	NA	NA	NA	NA
FAP_8	FAP_G47	FAP	TA	R	<5	NA	NA	NA	NA
FAP_8	FAP_G48	FAP	TA	R	<5	NA	NA	NA	NA
FAP_9	FAP_G67	FAP	TA	R	5-10	NA	NA	NA	NA
FAP_10	FAP_G61	FAP	TA	NA	NA	NA	NA	NA	NA
LS_1	LS_EBL1	LS	TA	D	5-10	Intact	Intact	NA	NA
LS_2	LS_EBL3	LS	TA	A	<5	Intact	Intact	NA	NA
LS_3	LS_EBL5	LS	HP	D	5-10	NA	NA	Intact	Intact
LS_4	LS_EBL7	LS	TA with HGD	D	>10	Lost	Lost	NA	NA
LS_5	LS_EBL9	LS	TA	R	5-10	Lost	Lost	NA	NA
LS_6	LS_EBL11	LS	TA	A	>10	Lost	Lost	NA	NA
LS_7	LS_EBL13	LS	TA	A	5-10	Intact	Intact	NA	NA
LS_8	LS_EBL18	LS	HP	D	5-10	Intact	Intact	NA	NA
LS_9	LS_EBL26	LS	TA	D	5-10	NA	Intact	NA	NA
LS_10	LS_EBL20	LS	TA	D	5-10	Intact	Lost	NA	NA
LS_11	LS_EBL23	LS	TA	A	<5	Intact	Intact	NA	NA
LS_12	LS_T3	LS	CA	D	pT3N0	Intact	Intact	Lost	Lost
LS_13	LS_T2	LS	CA	D	pT4aN0	Intact	Intact	Lost	Lost
LS_14	LS_T1	LS	CA	A	pT2N0	Intact	Intact	Intact	Lost

Abbreviations: TA, tubular adenoma; HP, hyperplastic polyp; HGD, high grade dysplasia; A, cecum/ascending/hepatic flexure; D, transverse/descending/sigmoid; R, rectum/rectosigmoid; NA, not assessed. Assessment of expression of MMR proteins in LS polyps was based on the already known germline mutations. Therefore, only the protein pair matching the germline mutation was performed.

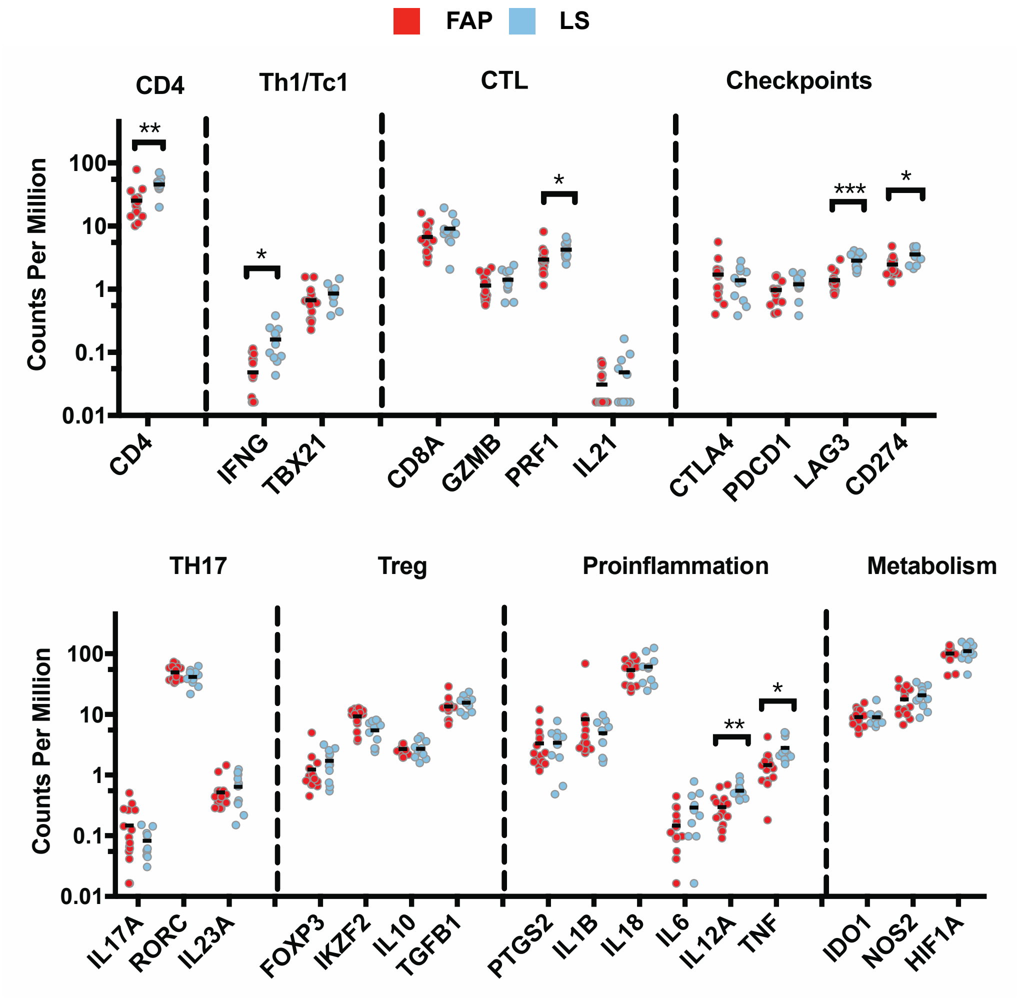


Figure 6 mRNA expression levels of immune-related genes involved in CD4, Th1/Tc1, CTL, checkpoint response, TH17, Treg, Proinflammation and Metabolism comparing LS and FAP polyps. The graphs display means for each group and statistically significant difference between FAP and LS (*P<0.05, **P<0.01, ***P<0.001), using Welch's t-test and multiple comparisons by Benjamini-Hochberg method.

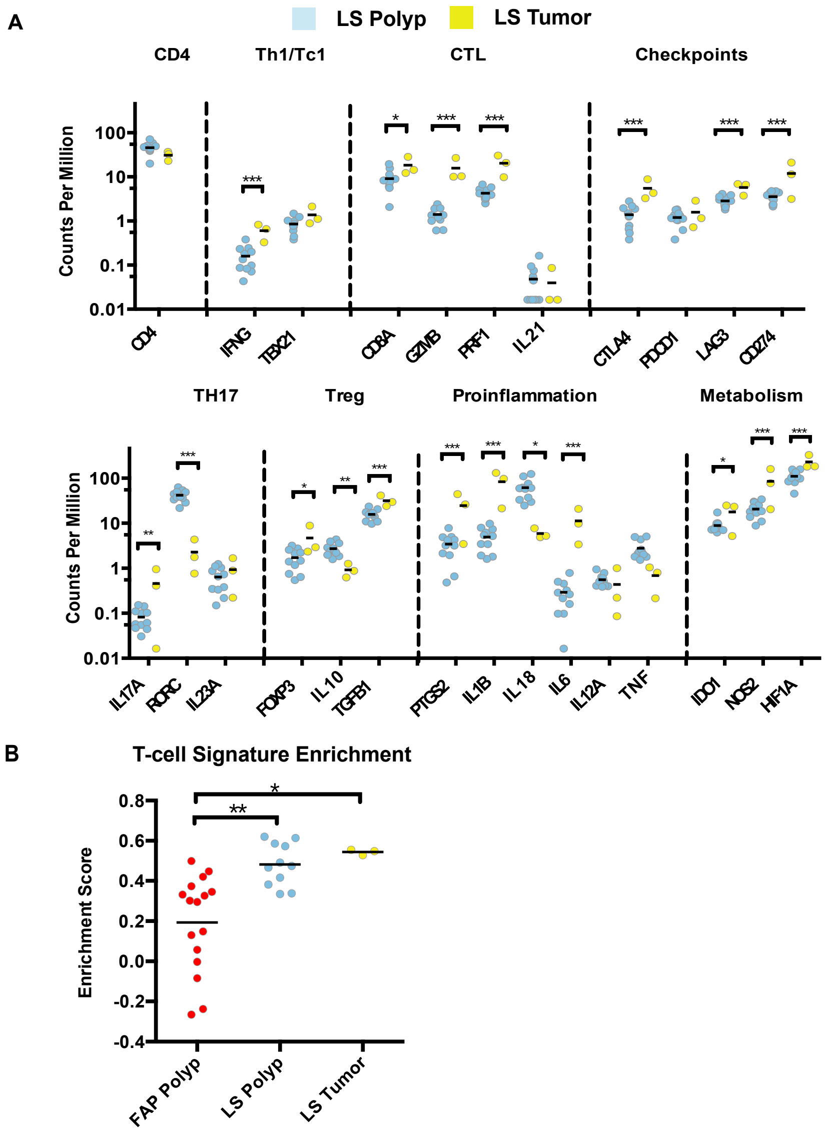


Figure 7. A. mRNA expression levels of immune-related genes involved in CD4, Th1/ Tc1, CTL, checkpoint response, TH17, Treg, Proinflammation and Metabolism comparing LS polyps and LS tumors; B. T-cell signature enrichment score comparing LS and FAP polyps, LS tumors. The graphs display means for each group and statistically significant difference between each group using ANOVA and Tukey's Test, and multiple comparisons by Benjamini & Hochberg method (*P<0.05, **P<0.01, ***P<0.001).

4.2.2 Hypermutated LS polyps are associated with mismatch repair-deficient mutation signature

Based on the previous findings made in carcinomas, we hypothesized that one possible explanation for the immune deregulation observed in LS premalignancy is the acquisition of high levels of somatic mutations (hypermutation). In order to assess the mutation rate in our samples, we called mutations from RNA-seq data and compared the results to hyper- and non-hypermutant sporadic carcinomas from The Cancer Genome Atlas (TCGA) and our LS carcinomas. To this end, we first demonstrated the feasibility of using RNA-seq to estimate somatic mutation rates by observing a statistically significant correlation between mutation rates called from whole exome sequencing (WES) data and RNA-seq in 47 TCGA samples with matched data (R^2=0.339, *P*-value<0.001). Overall, polyps displayed low mutation rates compared to carcinomas; however, among LS polyps 3 were found to be hypermutated (Figure 8A), and exhibited a mutation signature with distinct C>T changes that are associated with deficiency in the DNA MMR system[95]. These 3 hyper-mutant LS polyps clustered with sporadic hyper-mutant CRCs from TCGA and LS carcinomas based on mutation signature 6 (Figure 8B, Figure 9) and displayed loss of staining of MMR proteins (i.e. MMR deficiency, Table 9). At the same time FAP polyps and non-hypermutated LS polyps shared a similar mutation spectrum lacking the distinct MMR-deficient pattern. Furthermore, a comparative analysis of the immune profile of hyper-mutant and non-hypermutant LS polyps was only significant for the Treg-related gene *FOXP3* and the immune checkpoint *CTLA4* (Table 10). We confirmed with immunohistochemistry the prominent infiltration by FOXP3 positive T-cells of hyper-mutant LS polyps (Figure 10). This fact suggests that the immune activation program that is displayed by all LS polyps of this cohort (both hyper- and non-hypermutant) is independent of the mutation rate.

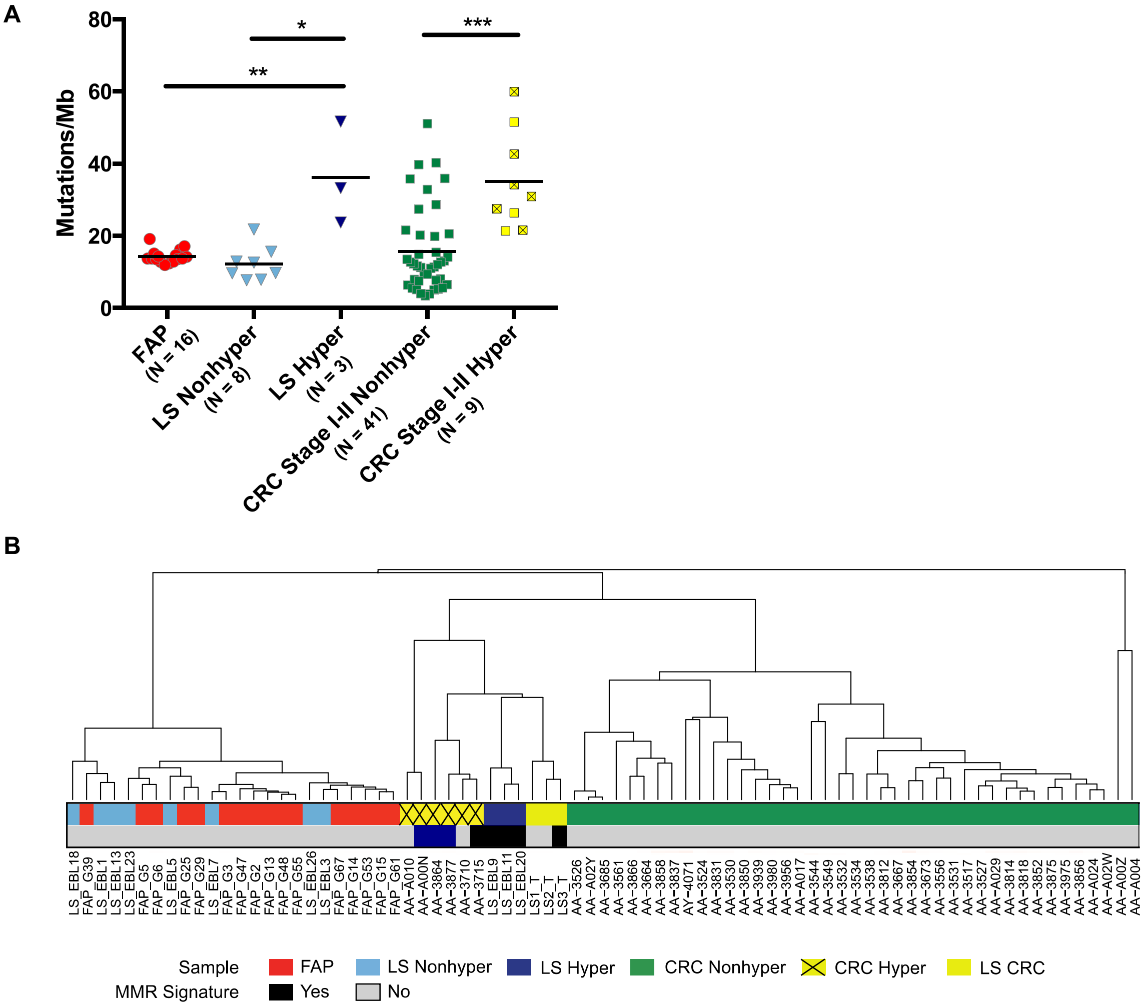


Figure 8. Mutational rate and mutation signature distribution in Lynch syndrome premalignancy. A. Comparisons of mutation rate among FAP, hypermutated and non-hypermutated LS polyps, hypermutated LS, and TCGA tumors. The graphs display means for each group and statistically significant differences between groups (*P<0.05, **P<0.01, ***P<0.001); B. Hierarchical Clustering of mutation spectrum of FAP and LS polyps, LS Tumors and sporadic TCGA CRC stage I and II colorectal tumors with known MSI and hypermutation status. A total of 3 hypermutated LS polyps, 1 LS tumor and 4 hypermutated TCGA tumors display mutational signature 6, which is caused by defective DNA MMR.

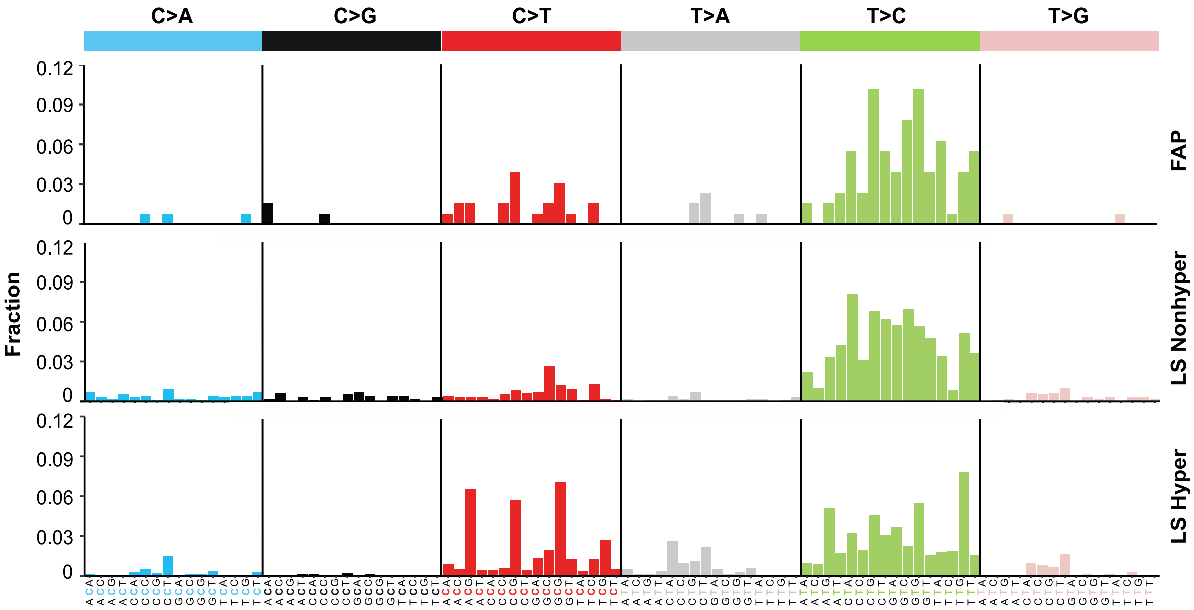


Figure 9. Mutation Spectrum of FAP and LS polyps. Note that LS adenomas with hypermutation displayed a higher proportion of C>T changes as well as a different profile of T>C changes.

Table 10. Immune profile of LS polyps classified by mutational rate (hyper- versus non-hypermutant. Genes linked to the immune microenvironment of CRC and MMR deficiency were grouped by lineage and/or function (Th1/Tc1, CTL, Th17, Treg, proinflammation, and metabolism) as previously reported by Llosa et al, Cancer Discovery (2016). A total of 3 LS polyps were found to be hypermutant and 8 non-hypermutant. Expression values are expressed in counts per million. The statistical tests were performed on log2-transformed CPM. Welch's t-Test and multiple correction by Benjamini & Hochberg were used (FDR<0.05). Significantly different genes are in bold. Abbreviations: Hyper, Hypermutant.

Gene	LS Hypermutant Mean	LS Non-hypermutant Mean	LS Hyper- vs Non-hypermutant FDR	LS Hyper- vs Non-hypermutant p-value
CD4	42.268	47.382	0.8453	0.6574
IFNG	0.141	0.167	0.9375	0.8565
TBX21	0.693	0.918	0.7547	0.3808
CD8A	6.919	9.970	0.7547	0.4472
GZMB	1.342	1.435	0.8860	0.7219
PRF1	4.051	4.313	0.9375	0.8613
IL21	0.039	0.052	0.9375	0.9375
IL17A	0.081	0.083	0.9375	0.9242
RORC	37.285	43.126	0.8295	0.5372
IL23A	0.573	0.667	0.9375	0.9101
FOXP3	2.839	1.291	**0.0466**	**0.0017**
IL10	1.880	3.036	0.0651	0.0145
TGFB1	18.171	14.693	0.7204	0.2935
PTGS2	5.660	2.590	0.0789	0.0205
IL1B	4.935	4.907	0.8295	0.5530
IL18	33.313	71.735	0.0520	0.0058
IL6	0.446	0.233	0.7547	0.4226
IL12A	0.439	0.599	0.2647	0.0882
TNF	2.024	3.085	0.3795	0.1405

CTLA4	2.281	1.045	**0.0495**	**0.0037**
PDCD1	1.283	1.163	0.7547	0.3759
LAG3	2.654	2.925	0.8453	0.6361
CD274	3.039	3.752	0.7547	0.3989
IDO1	10.619	8.295	0.8453	0.6166
NOS2	27.165	18.278	0.1003	0.0297
HIF1A	142.719	98.423	0.0651	0.0129

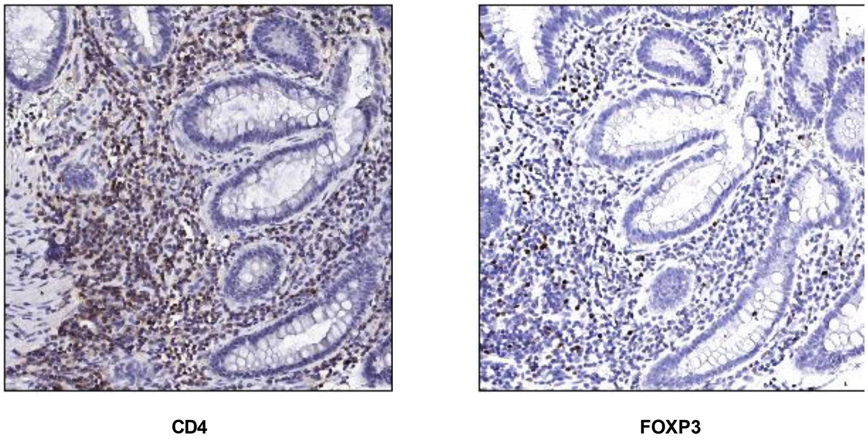


Figure 10. Immunohistochemical staining of CD4 and FOXP3 of a representative hypermutant Lynch syndrome polyp showing abundant infiltration by CD4 positive/FOXP3 positive T-cells.

4.2.3 The number of neoantigens is correlated with mutation rates but not associated with the immune expression profile

We postulated that the immune profile observed in all LS polyps could be secondary to an increase in neoantigen rate that is independent from the global mutational rate. To determine this, we performed MHC class I and II typing and detected tumor-specific neoantigens using bioinformatic methods[62, 96]. The 3 hypermutated LS polyps displayed a neoantigen burden that was similar to that for LS carcinomas and higher than for the non-hypermutant LS and FAP polyps (Figure 11). This difference was statistically significant for both high- and low-binding affinity neoantigens binding MHC class I and II (Figure 12A, B) and secondary to the accumulation of indels (Figure 12C, D). Overall, the total number of neoantigens in FAP samples did not vary as a function of the mutational rate (R^2=0.02) but it did correlate well in LS polyps (R^2=0.8, *P*-value<0.001, Figure 12E, F). These analyses confirm that a higher neoantigen load is secondary to an increase in mutational rate in hyper-mutant LS polyps but this is not responsible for the overall immune profile displayed by LS premalignancy.

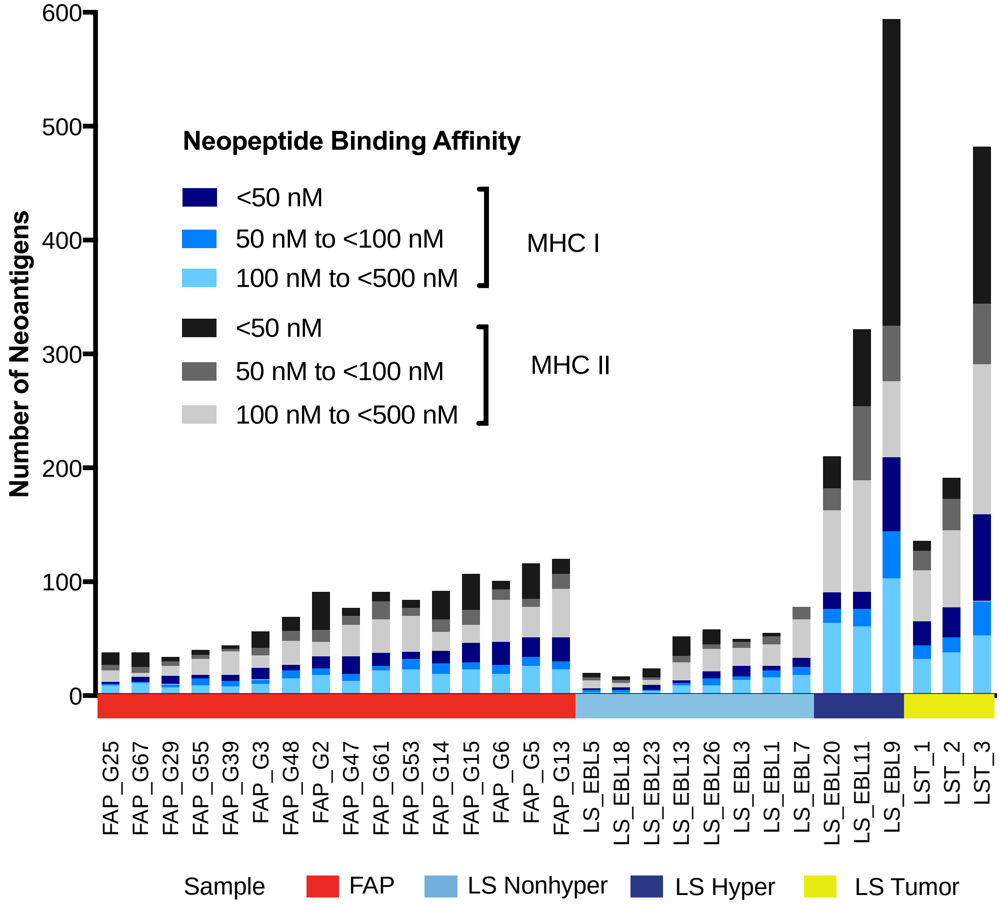


Figure 11. Total number of MHC class I and II neoantigens in FAP, LS polyps and LS tumors. Each column represents a sample with MHC class I neoantigens in shades of blue stratified by binding affinity and MHC class II in shades of grey. Neoantigens selected had <500 nM >5% allele frequency, were expressed at >10 Transcripts per Million (TPMs), and were not found in 1,000 genomes database.

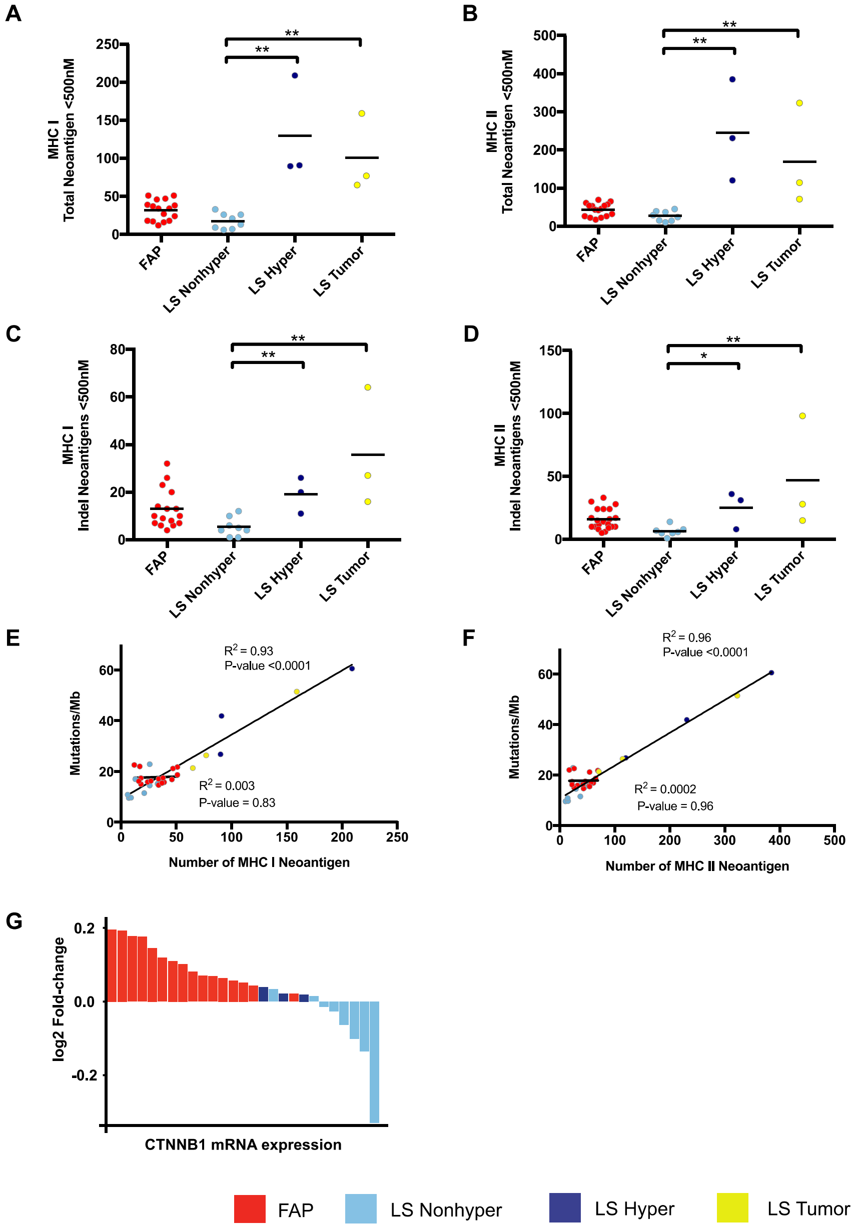


Figure 12. MHC Class I and II neoantigens in Lynch syndrome and FAP premalignancy, and LS carcinomas. A,B Predicted number of MHC class I and II neoantigens (<500nM and >5% AF and not in 1kg and >10 TPM) of FAP and LS polyps, LS tumors; **C,D** Number of InDels in MHC class I and II predicted neoantigens; **E,F** Mutation rate vs predicted number of MHC class I and II neoantigens; Statistical differences between each groups are performed with

Kruskal-Wallis and Dunn’s Test. (*P<0.05, **P<0.01, ***P<0.001); **G**. fold-change of CTNNB1 mRNA level between FAP, LS polyps and matched normal mucosa.

4.2.4 LS premalignancy display neoantigens in additional DNA repair pathways and FAP in WNT/β-catenin

We proceeded to discover gene pathways affected by emerging neoantigens unique to LS premalignancy using Ingenuity Pathway Analysis (IPA). Among the most significant pathways enriched by both class I and II neoantigens, we identified alterations in DNA repair mechanisms that could contribute to additional accumulation of somatic mutations in advanced LS polyps and carcinomas such as the role of BRCA1 in DNA damage response and *ATM* signaling (Figure 13). On the other side, FAP polyps acquired neoantigens in the WNT pathway (Figure 13) and also accumulated somatic genomic events in *APC* (Table 11). As it has been recently suggested, the activation of β-catenin (*CTNNB1*) secondary to deregulation of the WNT pathway is responsible for immune exclusion in carcinomas[94]; therefore, we decided to assess the expression levels of *CTNNB1* in our samples and found that all FAP polyps presented with WNT/β-catenin activation compared to normal adjacent mucosa. In contrast LS polyps did not display any significant activation of *CTNNB1*, thus supporting the contribution of WNT/β-catenin to immune exclusion in FAP premalignancy (Figure 12G). However, in the absence of high neoantigen rates and MMR deficiency the mechanism responsible for the immunoactivation in LS premalignancy remains elusive.

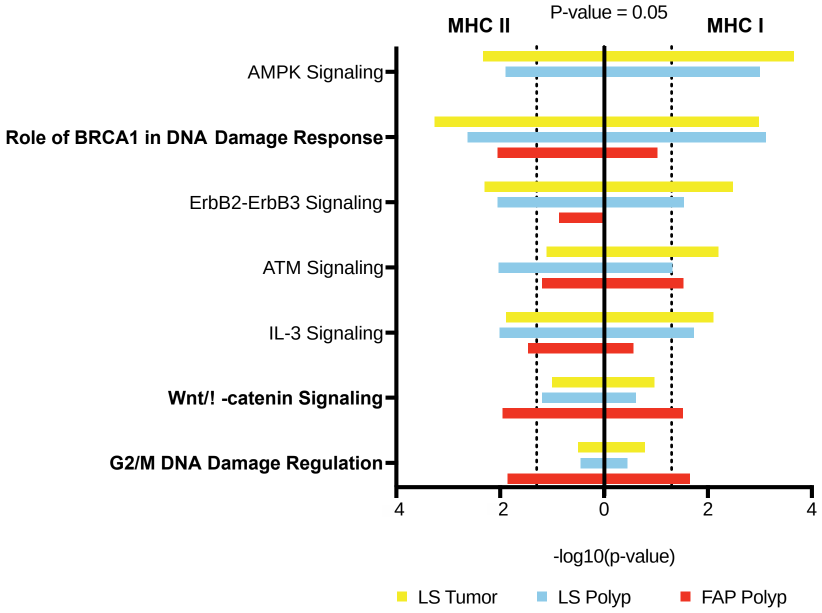


Figure 13. Pathway analysis of neoantigens present in FAP, LS polyps and LS Tumors. MHC class I and II neoantigens were selected with <500nM binding affinity and >5% allele frequency and expressed at >10 TPMs and were not found in 1,000 genomes database. Only selected pathways with enrichment for both MHC class I and II neoantigens are displayed. Complete list of pathways can be found eTable9 in Chang K, et al. “Immune Profiling of Premalignant Lesions in Patients With Lynch Syndrome”. JAMA Oncol. 2018 Apr 16. doi: 10.1001/jamaoncol.2018.1482.

Table 11. Germline and somatic mutation detected in FAP and LS polyps. Mutational data derived from RNA-seq analysis. Sample FAP_G53 and LS_EBL26 harbored APC LOH in 5q in addition to the somatic mutations.

Sample ID	Type	Patient ID	Mutation	Gene	cDNA	Protein	Germline Validation
FAP_G2	FAP	FAP_1	Germline	APC	c.1880dupA	p.Ala630*	Yes
FAP_G3	FAP	FAP_1	Germline	APC	c.1880dupA	p.Ala630*	Yes
			Somatic	APC	c.4473_4474delinsGC	p.(Phe1491_Ala1492delinsLeuPro)	
			Somatic	KRAS	c.519T>C	p.(=)	
FAP_G53	FAP	FAP_2	Germline	APC	c.3810T>A	p.Cys1270*	Yes
			Somatic	APC	c.2438A>G	p.(Asn813Ser)	
			Somatic	APC	c.4348C>T	p.(Arg1450*)	
FAP_G55	FAP	FAP_2	Germline	APC	c.3810T>A	p.Cys1270*	Yes
FAP_G5	FAP	FAP_3	Germline	APC	del exon 15	p.?	No
			Somatic	APC	LOH 5q		
FAP_G6	FAP	FAP_3	Germline	APC	del exon 16	p.?	No
FAP_G13	FAP	FAP_4	Germline	APC	c.622C>T	p.Gln208*	Yes
			Somatic	APC	c.4057G>T	p.(Glu1353*)	
			Somatic	APC	c.4135G>T	p.(Glu1379*)	
FAP_G14	FAP	FAP_4	Germline	APC	c.622C>T	p.Gln208*	No
FAP_G15	FAP	FAP_4	Germline	APC	c.622C>T	p.Gln208*	Yes
			Somatic	APC	c.4189_4190del	p.(Arg1399Phefs*9)	
			Somatic	APC	LOH 5q		
FAP_G25	FAP	FAP_5	Germline	APC	del exons 8-10	p.?	No
FAP_G29	FAP	FAP_6	Germline	APC	c.1658G>A	p.Trp553*	Yes
FAP_G39	FAP	FAP_7	Germline	APC	c.477C>G	p.Tyr159*	No
FAP_G47	FAP	FAP_8	Germline	APC	c.4733_4734del	p.Cys1578Tyrfs*13	No
			Somatic	APC	LOH 5q		
FAP_G48	FAP	FAP_8	Germline	APC	c.4733_4734del	p.Cys1578Tyrfs*14	No
FAP_G67	FAP	FAP_9	Germline	APC	c.847C>T	p.Arg283*	No
			Somatic	APC	c.4135G>T	p.(Glu1379*)	
			Somatic	KRAS	c.35C>A	p.(Gly12Val)	
FAP_G61	FAP	FAP_10	Germline	APC	c.3810T>A	p.Cys1270*	Yes
LS_EBL1	LS	LS_1	Germline	MSH6	c.3744_3773del	p.His1248_Ser1257del	Yes
LS_EBL3	LS	LS_2	Germline	MSH2	c.687del	p.Ala230Leufs*16	Yes
LS_EBL5	LS	LS_3	Germline	PMS2	del exon 14	p.?	No
			Somatic	MLH1	c.342_343del	p.(Ile115Tyrfs*6)	
LS_EBL7	LS	LS_4	Germline	MSH2	c.1034G>A	p.Trp345*	Yes
LS_EBL9	LS	LS_5	Germline	MSH2	c.1661+1G>A	p.?	No
			Somatic	MSH6	c.1340T>C	p.(Leu447Pro)	
			Somatic	MSH6	c.3410T>C	p.(Met1137Thr)	
LS_EBL11	LS	LS_6	Germline	MSH2	del exons 1-6	p.?	No

			Somatic	MSH2	c.2172G>A	p.(=)	
			Somatic	APC	c.2626C>T		
			Somatic	KRAS	c.485A>G	p.(Glu162Gly)	
			Somatic	CIITA	c.1436C>T	p.A479V	
LS_EBL13	LS	LS_7	Germline	MSH2	c.1216C>T	p.Arg406*	Yes
LS_EBL18	LS	LS_8	Germline	MSH6	c.3238_3239del	p.Leu1080Valfs*12	No
LS_EBL26	LS	LS_9	Germline	MSH6	c.3860_3861insATTA	p.Y1287*	Yes
			Somatic	MSH2	c.2646_2647del	p.(Lys882Asnfs*16)	
			Somatic	APC	c.4348C>T	p.(Arg1450*)	
LS_EBL20	LS	LS_10	Germline	MSH6	c.2645_2653del	p.Phe882*	Yes
			Somatic	APC	c.4234G>T	p.	
			Somatic	APC	LOH 5q		
LS_EBL23	LS	LS_11	Germline	MSH6	c.3699_3702del	p.Lys1233Asnfs*6	Yes
LS_T3	LS	LS_12	Germline	MLH1	c.1279C>T	p.Gln427*	Yes
			Somatic	APC	c.C4549T	p.Q1517X	
			Somatic	BRAF	c.1208delC	p.P403fs	
			Somatic	CTNNB1	c.C134T	p.S45F	
			Somatic	AXIN1	c.1922delA	p.K641fs	
			Somatic	RFX5	c.56delC	p.P19fs	
			Somatic	RFXAP	c.297delG	p.P99fs	
LS_T2	LS	LS_13	Germline	MLH1	c.1918C>T	p.Pro640Ser	Yes
			Somatic	APC	c.2540dupA	p.E847fs	
			Somatic	KRAS	c.G38A	p.G13D	
			Somatic	MLH1	c.C350T	p.T117M	
			Somatic	MSH6	c.666_667insGATGGAGG	p.D222fs	
			Somatic	MSH6	c.668_669insGGCACAACTT ACGTAAC	p.N223_E224delinsKAQ LTX	
LS_T1	LS	LS_14	Germline	MLH1	c.2059C>T	p.Arg687Trp	Yes

4.3 Discussion

Our results show a distinct immune profile in LS polyps, independent of the DNA mutation rate, the emergence of neoantigens that is secondary to frameshift mutations, and the MMR status. Among the immune checkpoints upregulated in polyps stand *LAG3*, which constitutes a promising target for immune interception in this patient population. Therefore, emergence of high mutation burdens and neoantigens cannot simply be applied as a biomarker to guide implementation and development of immuno-prevention strategies. In addition, we observed that neoantigen formation correlates with a high mutational rate present in the subgroup of LS polyps that are hyper-mutants. The acquisition of additional MHC class I and II neoantigens by hypermutated LS polyps was associated with the introduction of alterations in DNA damage repair pathways, which could further explain how MMR-deficiency increases neoantigen formation leading to hypermutation in carcinomas.

We have demonstrated that polyps that arise in LS are enriched for CD4-positive T-cells, which are responsible for the upregulation of the immune checkpoints *PD-L1* and *LAG-3*. This is consistent with recently reported transcriptomic profiles detected in normal mucosa samples of LS patients that harbored a CRC, which showed strong immune response associated to invasion of CD4-positive T-cells, expression of immune checkpoints, and HLA[97]. Furthermore, we found that this transcriptional program was displayed across all polyp types regardless of major clinico-pathological features such as histology (it was observed both in adenomas and hyperplastic polyps), size, location (right versus left), the presence of advanced features (high-grade dysplasia), and MMR status; but, most importantly, it was independent of the accumulation of somatic mutations. In fact, the only difference in terms of immune activation displayed by the 3 polyps that were found to be hyper-mutant and MMR-deficient compared to the rest of LS polyps was the upregulation of *CTLA4,* which was shared with LS carcinomas, and *FOXP3*, which is consistent with development of immune tolerance upon progression of carcinogenesis. In fact, the results of the analysis of immune-cell infiltrates from colorectal carcinomas diagnosed in participants of the Nurses' Health Study and the Health Professionals Follow-up Study reported a correlation between neoantigen load and density of *FOXP3* positive T-cell infiltrates[98]. In addition, neoantigens accumulated along with the acquisition of additional indels that generate new open reading frames in hyper-mutant polyps are more immunogenic than single nucleotide variants. Overall, these observations are consistent with the results from pan-TCGA analysis that indicated that indel load is more closely associated with overall immunogenicity and response to checkpoint inhibition[99]. Therefore, our results challenge the concept that immune activation in LS is a consequence of the excessive accumulation of somatic variation secondary to MMR deficiency, since all polyps analyzed presented a consistent immune profile regardless of the mutation rate or abundance of high-affinity binding neoantigens. This canonical concept could be the case at later stages of premalignancy (advanced polyps) and progression into carcinoma. However, immune deregulation could precede the accumulation of genomic aberrations and neoantigen formation in initial steps of carcinogenesis (Figure 14). Finally, this observation will advocate for the development of vaccine strategies to prevent the progression of carcinogenesis by priming T-cells to antigens displayed by early lesions that will be cleared at the premalignant stage. Furthermore, combinations of immune checkpoint inhibitors and vaccines could be exploiting both components displayed by MMR-deficient premalignancy.

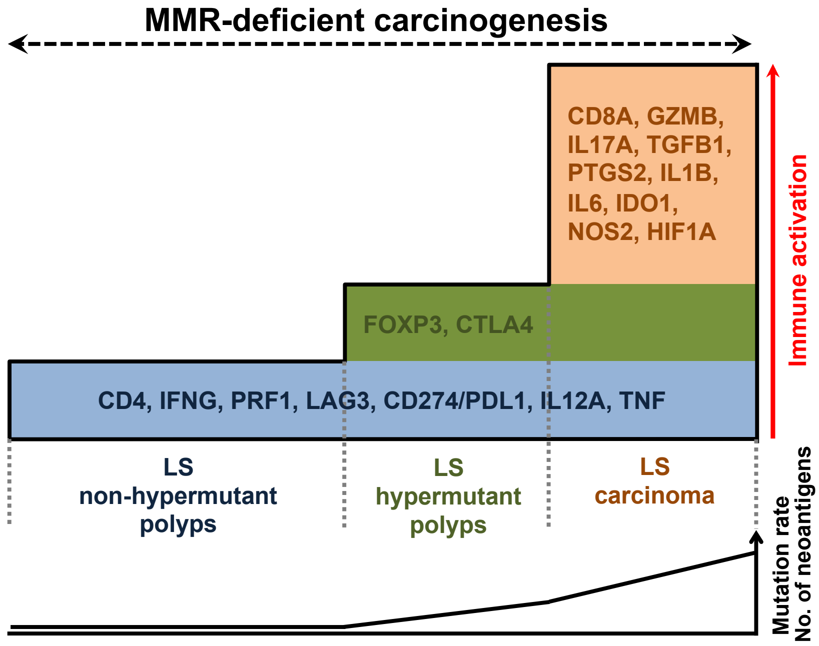

MMR-deficient carcinogenesis
CD8A, GZMB, IL17A, TGFB1, PTGS2, IL1B, IL6, IDO1, NOS2, HIF1A
FOXP3, CTLA4
CD4, IFNG, PRF1, LAG3, CD274/PDL1, IL12A, TNF
Immune activation
LS non-hypermutant polyps
LS hypermutant polyps
LS carcinoma
Mutation rate
No. of neoantigens

Figure 14. Schematic model of the immune activation in LS carcinogenesis. LS polyps display a marked immune activation profile characterized by CD4 T-cells, pro-inflammatory, and checkpoint molecules that is independent of mutational rates, neoantigen formation, and MMR status at early stages of carcinogenesis. Progression of mutational rate and acquisition of invasive features with evolution into carcinomas activate additional immune pathways with eventual development of immune tolerance (advanced lesions) and evasion (carcinomas).

5 CONSENSUS MOLECULAR SUBTYPE OF SPORADIC AND HEREDITARY PREMALIGNANT LESIONS

Chapter 5 CONSENSUS MOLECULAR SUBTYPE OF SPORADIC AND HEREDITARY PREMALIGNANT LESIONS

5.1 Introduction

The canonical genetic pathway of step-wise cascade of somatic mutations in tumor suppressor genes and oncogenes have been described extensively in colorectal carcinogenesis[5, 28]. At the transcriptomic level, used large-scale gene expression-based profiling of primary CRC tumors to identify four distinct consensus molecular subtypes (CMS) with distinguished biological features and prognostic subgroups[34].

Up to 5% of all CRC cases arise in the setting of well-defined inherited syndromes, such as familial adenomatous polyposis (FAP) and Lynch syndrome (LS), among others. FAP results from germline mutations in *APC*, which constitutively activates WNT/β-catenin mediated-transcription, driving the transformation of intestinal crypts to conventional precursor lesions (tubular, tubulovillous or villous adenomas). FAP-related adenomas have a spectrum of molecular features similar to CIN-positive CRCs[3]. LS-associated CRC results from germline mutations in the MMR genes (*MLH1*, *MSH2*, *MSH6*, *PMS2* and *EPCAM*), with conventional adenomas representing the majority of precursor lesions[100]. Even though MSI is present in only half of LS-associated polyps at diagnosis, MMR becomes universally detected in more advanced precursor lesions, when larger in size[101]. An alternative route of colorectal carcinogenesis is the serrated pathway has been shown to be an alternative colorectal carcinogenetic route potentially accounting for up to one third of all CRCs. Serrated/Hyperplastic polyposis syndrome (SPS) is characterized by numerous sessile serrated adenomas (SSA), predominantly located in the right side of the colon, in addition to hyperplastic polyps (HP)[102]. SSA carcinogenesis pathway is associated with high CpG island methylation phenotype (CIMP^{hi}) and *BRAF*^{V600E} mutations as the major driving mechanisms in both sporadic and familial cases[22] [23].

Collectively, such molecular insights of sporadic and familial CRC have been crucial in the development of clinical standards for managing both early-stage and advanced disease. On the contrary, applications towards disease risk prediction and targeted prevention still remain limited. This is due to a relative paucity of information regarding the full spectrum of genetic, epigenetic, and transcriptomic changes in benign colon polyps or pre-malignant adenomas. We hypothesize that filling this knowledge gap, particularly at the transcriptomic level, may lead to new approaches for CRC disease prevention not only in the general population, but also among high-risk groups such as those with familial CRC. Therefore, to more broadly understand the transcriptomic landscape of premalignant polyps, we have applied CMS classification on several cohorts of sporadic and familial adenomas with gene expression data. We hypothesize that: (i) most FAP-related and sporadic conventional adenomatous polyps (AP) have a CMS2-like (epithelial canonical) phenotype; (ii) LS polyps display a CMS1-like (MSI-Immune) phenotype; (iii) SSA and HP are enriched for both CMS1-like and CMS4-like (mesenchymal) phenotypes.

5.2 Results

5.2.1 Consensus molecular subtyping of a large cohort of polyps revealed CMS2 and CMS1 as major subtypes in premalignancy

We first analyzed the distribution of CMS groups across polyps from different clinical contexts (sporadic versus hereditary syndrome) and pathologic subtype (AP, HP, and SSA). Overall, the majority of sporadic polyps (n=311) were classified as either CMS2 (69.5%) or CMS1 (21.9%), while CMS3 (5.1%) and CMS4 (1.6%) classifications were less abundant. Furthermore, within sporadic polyps, the majority of AP (80%) were classified as CMS2, whereas the majority of HP (57.1%) and SSA (76.5%) were classified as CMS1 (Figure 15A).

Similarly, hereditary polyps (n=78) were mostly distributed between CMS2 (52.6%) and CMS1 (38.5%). CMS3 (2.6%) and CMS4 (6.4%) again accounted for a small percentage of total hereditary polyps (Figure 15B). AP from a hereditary background were predominantly

(86.7%) CMS2. Surprisingly, the majority (86.7%) of AP from LS patients were also classified as CMS2, which is in contrast to our *a priori* assumption. HP (71.4%) and SSA (96.2%) from a hereditary background were mostly classified as CMS1. Overall, these results suggest that the CMS2 (canonical) and CMS1 (MSI-Immune) molecular subtypes play dominant roles in early conventional adenomas and serrated polyps, respectively.

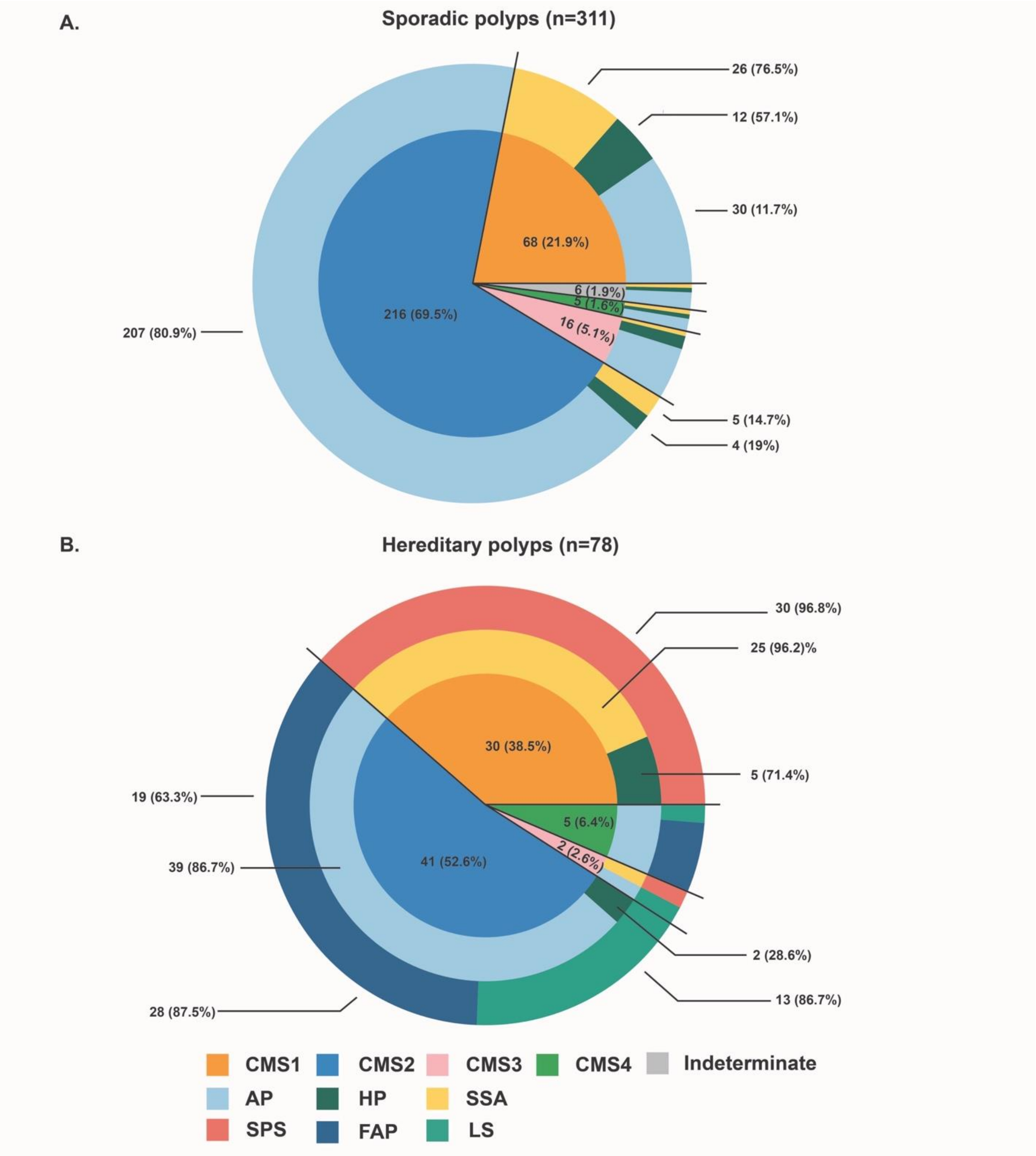


Figure 15. Circos plots presenting the distributions of consensus molecular subtype (CMS) groups in sporadic (A) and hereditary polyps (B).

5.2.2 Pathway enrichment analysis of CMS showed immune activation and classical WNT and MYC targets as dominant signatures in premalignancy

We performed GSEA using previously described biological pathways and expression signatures pertinent to CRC carcinogenesis[34]. CMS1-like polyps were characterized by significant enrichment of genes involved in immune and stromal infiltration as well as pathways implicated in immune cytotoxicity. They also showed strong activation in *JAK-STAT* and *MAPK* signaling (Figure 16A-C). CMS2-like polyps displayed strong enrichment for *WNT* and *MYC* targets, which are classical carcinogenesis pathways in CRC (Figure 16A-C). For the small number polyps classified as CMS3, we did not observe significant enrichment for glutamine and fatty acid pathways, thus making their activation a molecular feature that arise in advanced adenoma or carcinoma (Figure 16A-C). Lastly, although the number of CMS4 polyps was small, they showed significant enrichment of mesenchymal and stromal signatures along with *TGFβ* activation (Figure 16A-C). Taken together, these analyses confirm that immune activation and classical carcinogenesis pathways were the main transcriptomic events in colorectal premalignancy.

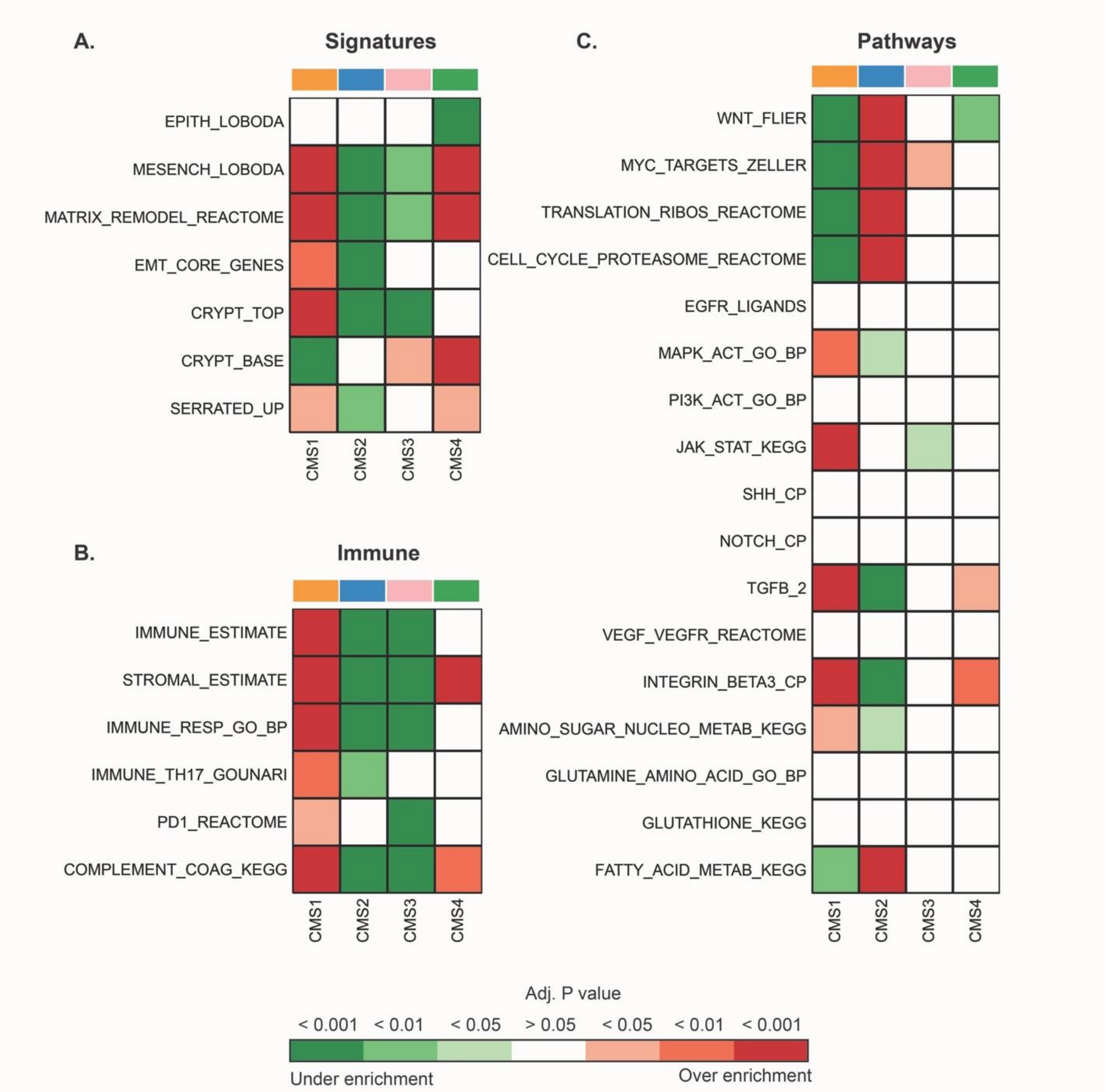


Figure 16. Aggregated gene set enrichment analysis of the different consensus molecular subtype (CMS) groups showing signatures of interest in colorectal carcinogenesis (A), immune signatures (B) and canonical pathways (C).

5.2.3 Associations of CMS with polyp location and KRAS and BRAF mutations

We next explored CMS distributions across various clinical-pathologic and molecular features. We found that CMS1 and CMS2 polyps have similar proportions in both males and females (Figure 17A). No statistically significant association was found between the presence of high-grade dysplasia/carcinoma in situ and CMS classification (Figure 17B). Interestingly, CMS1 polyps were more frequently presented in right colon in both sporadic ($P < 0.01$) and hereditary ($P < 0.0001$) cohorts. On the contrary, CMS2 polyps are more frequently presented in the left colon in both sporadic ($P < 0.005$) and hereditary ($P < 0.005$) cases (Figure 17C). These results suggest that CMS1 carcinomas may be largely derived from HP and SSA that are often found in the right colon. Next, we investigated mutations associated with CMS groups and found that *BRAF*^{V600E} was more frequently present in CMS1 polyps ($P < 0.0001$). Furthermore, *KRAS* codon 12 and 13 mutations showed a trend to occur occurred less frequently among CMS1 polyps compared to CMS2 polyps ($P < 0.05$) did not reach statistical significance. Due to the small number of CMS3 polyps in our sample, we did not observe significant over-representation of *KRAS*.

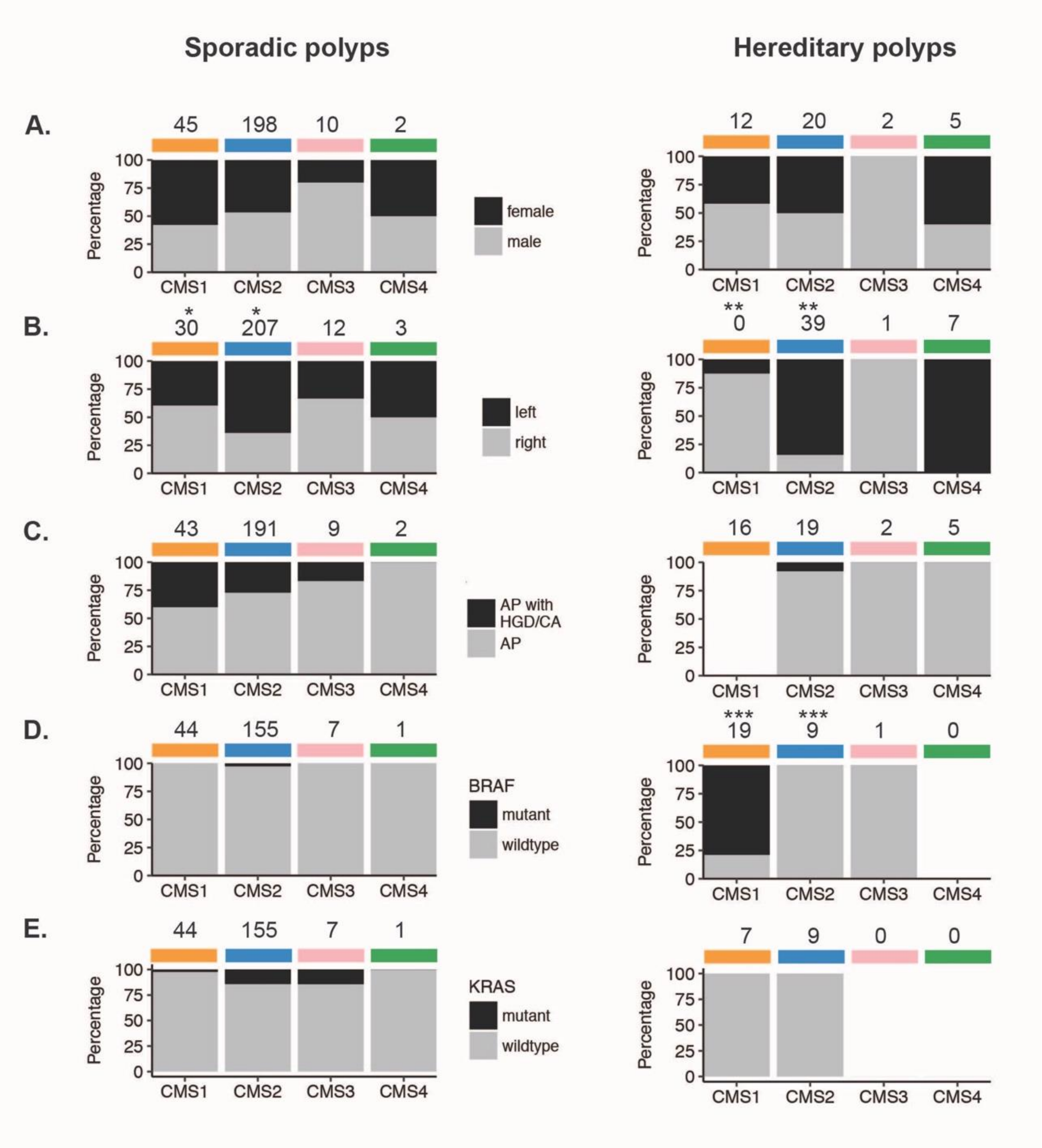


Figure 17. Clinical, pathological and molecular associations of consensus molecular subtype (CMS) groups; A. Distribution by gender; **B.** Presence of large-grade dysplasia or carcinoma in situ for the subset of APs; **C.** polyp location**; D.** BRAF mutation status; **E.** KRAS mutation status. (*P<0.01, **P<0.001)

5.3 Discussion

In the United States, the overall incidence of CRC has steadily decreased over the past decade[1] owing to increased utilization of screening colonoscopies. Yet, CRC still remains the third most common newly diagnosed malignancy in both men and women. Furthermore, despite its decreased overall incidence, an alarming trend towards younger age-of-onset (< 55 years old) has also emerged[103]. These observations highlight an important challenge to better understand the molecular diversity of colonic polyps and to develop targeted approaches for disease prevention. Towards this end, we performed a large-scale transcriptomic analysis of both sporadic and familial colon polyps by applying the CMS framework.

Taken together, our results allow the proposal of a new model for pathway activation driving premalignancy (Figure 18). First, we found that FAP, LS and sporadic conventional adenomas display an epithelial canonical CMS2 phenotype with strong WNT and MYC downstream targets enrichment. Secondly, we found that HP and SSA were both enriched for MSI-Immune CMS1 phenotype. While they displayed strong enrichment of immune and JAK-STAT activation, they also showed enrichment in TGFβ activation and stromal signature. Our results are consistent with previous studies showing that TGFβ activation play an important role in colorectal carcinogenesis. Using human organoid cultures and genome editing technology, Fessler et al investigators have shown that the genetic background of premalignant lesions dictates the dominating response to TGFβ, changing it from a largely apoptotic response in WNT pathway-activated conventional tubular adenomas to a dominant epithelial-mesenchymal transition response in *BRAF*^{V600E}-mutated SSA[68]. Depending on the level of TGFβ on the microenvironment, SSA could progress to either poor-prognosis CMS4 tumors (high in TGFβ signaling) or the good-prognosis CMS1 tumors (low in TGFβ signaling).

Finally, the ability to classify resected premalignant lesions into indolent versus aggressive molecular subtypes may have important clinical utility for colon cancer screening in the future. Specifically, CMS4 carcinomas have an aggressive clinical phenotype as defined by a higher proportion of advanced stage at diagnosis and worse outcomes after surgery and adjuvant chemotherapy[34]. Extrapolating from these data, we hypothesize that CMS4-like premalignant lesions evolve more rapidly in the adenoma-to-carcinoma transformation process and may indicate the need for closer follow-up evaluation of the at-risk normal mucosa than would otherwise be pursued based on current surveillance guidelines. Certainly, several key studies are needed to explore this hypothesis and firmly establish the prognostic utility of premalignant CMS classification. Towards this end, analysis of longitudinal patient cohorts with sufficient clinical outcomes data (e.g. diagnosis of advanced adenoma or carcinoma) would be highly valuable. Although current CMS RF classifier has robust performance on archived tissue specimens, the classifier contains more than 200 genes. A new classifier requiring fewer genes would be ideal in clinical settings. Recently, a new CMS classifier using 38 genes derived from Nanostring platform shown to be suitable for FFPE samples[104].

We acknowledge that our study has several limitations. First, although we have demonstrated that CMS classification is technically feasible for premalignant tissue, it is also important to recognize that the classifier was derived specifically from carcinoma. As such, it is possible that our classification method performs sub-optimally for premalignant tissue and may not accurately describe its transcriptomic landscape. In future studies, the alternative approach would be to derive a premalignancy-specific classifier. Second, our study did not include comprehensive analysis of somatic single-nucleotide mutations or copy-number alterations in the polyp samples. These additional analyses would have allowed us to correlate various known CRC drivers with CMS classification of polyps. Third, we did not have information on the MSI status of the polyps in our study. We attempted to assess the MSI status of polyps using gene signatures derived from carcinoma data and hierarchical clustering but we were unable to observe distinct groups of samples. These limitations are primarily driven by the diminutive size of polyps and the requirement to prioritize tissue for gene expression analysis. Lastly, to help maximize the number of adenomas included in our study, we opted to classify samples according to the CMS subtype with the highest posterior probability. By comparison, setting a minimum threshold of 0.5 posterior probability would reduce the number of analyzable samples significantly but would not change the overall conclusions of our study (Table 12).

Overall, to the best of our knowledge, our study is the largest investigation of transcriptional drivers in colorectal premalignancy. Our results show that pathway dependencies of different CMS groups originally described in carcinomas are indeed recapitulated in adenomas, thus opening the door to more personalized development of targeted chemopreventive strategies for polyps, particularly in hereditary syndromes.

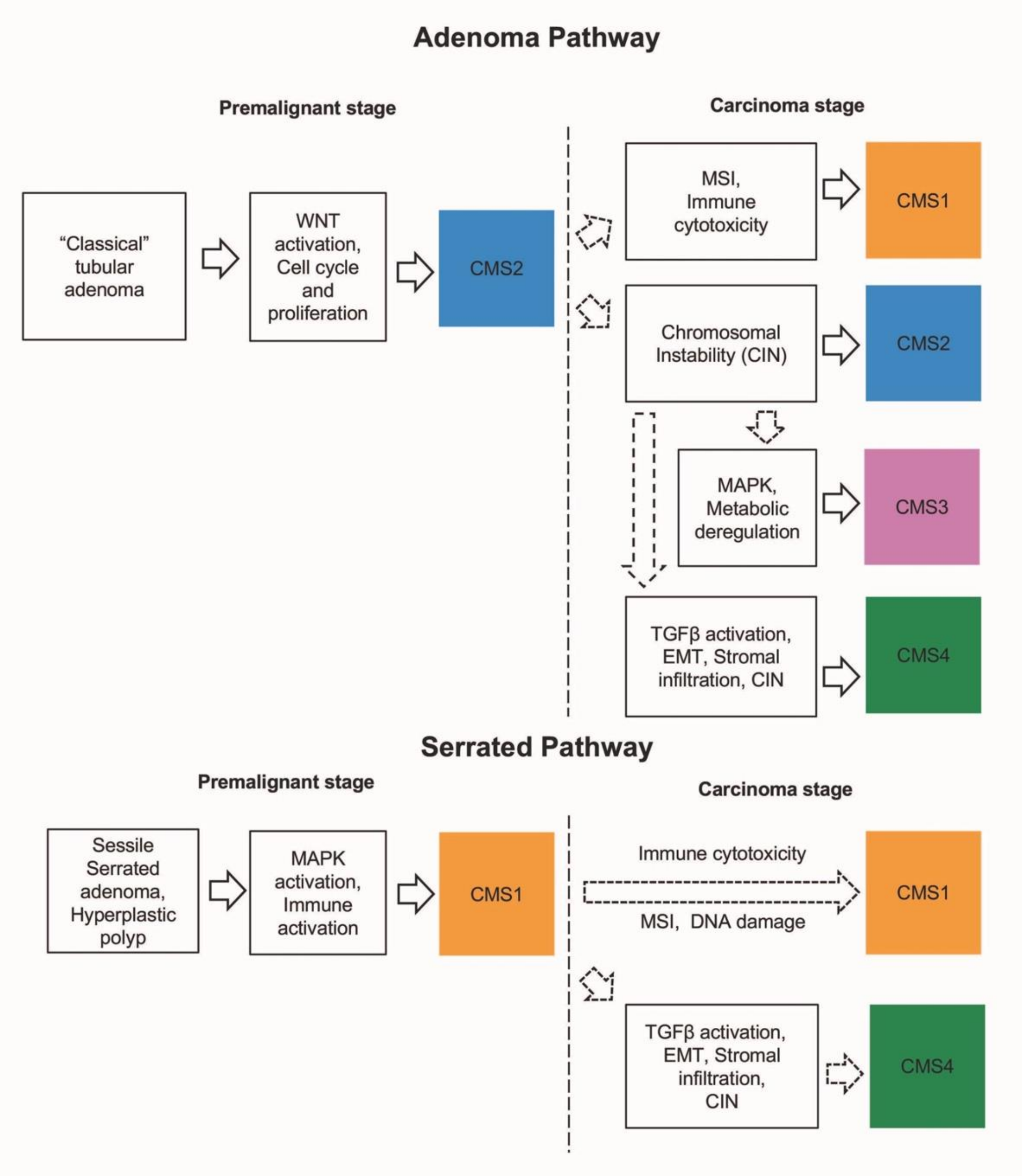

Adenoma Pathway
Premalignant stage
Carcinoma stage
"Classical" tubular adenoma
WNT activation, Cell cycle and proliferation
CMS2
MSI, Immune cytotoxicity
CMS1
Chromosomal Instability (CIN)
CMS2
MAPK, Metabolic deregulation
CMS3
TGFβ activation, EMT, Stromal infiltration, CIN
CMS4
Serrated Pathway
Premalignant stage
Carcinoma stage
Sessile Serrated adenoma, Hyperplastic polyp
MAPK activation, Immune activation
CMS1
Immune cytotoxicity
MSI, DNA damage
CMS1
TGFβ activation, EMT, Stromal infiltration, CIN
CMS4

Figure 18. Model of pathway activation driving the consensus molecular subtype (CMS) classification in adenomatous (top) and serrated polyps (bottom).

Table 12 Distribution of CMS subgroups using different probability thresholds for the CMS RF classifier

	RF P=0.5		***RF P=0.3***	
	n	***%***	***n***	***%***
CMS1	21	12.6	98	25.6
CMS2	143	85.6	257	67.1
CMS3	3	1.8	18	4.7
CMS4	0	0	10	2.6
Indeterminate	222	-	6	-

6 DISCUSSION, CONLUSIONS AND FUTURE DIRECTIONS

Chapter 6 DISCUSSION, CONLUSIONS AND FUTURE DIRECTIONS

6.1 Discussion and Conclusions

In the past decade, large scale sequencing studies have identified genomic and transcriptomic alterations of colorectal tumorigenesis[11, 34]. The results have led to development of novel therapeutic targets and improved disease management in late stage carcinomas[19, 105]. Despite the abundance of knowledge acquired in carcinoma stage, there have been not be sufficient systematic and large-scale efforts in characterizing colorectal premalignant lesions. Therefore, the purpose of this dissertation is to provide insights into the molecular mechanisms of early colorectal carcinogenesis using a large cohort of premalignant lesions and next-generation sequencing technologies.

The “Big Bang” colorectal carcinogenesis model states that the vast majority of the genomic alterations are acquired during early stages of carcinogenesis and assumes that this massive accumulation happens after an initiating driver (*APC)* has occurred in a single clone (founder clone)[83, 84]. Subsequent acquisition of driver mutations, such as *KRAS,* foster tumor progression and generation of subclones that will acquire additional mutations. Therefore, the “big bang” model of colorectal carcinogenesis implies that a polyclonal carcinoma is derived from monoclonal origin[106]. In chapter 3, we applied bioinformatics approaches to determine the presence of polyclonality in premalignant lesions using whole exome sequencing. My results showed that 72% of the lesions are polyclonal. In addition, I did not detect a significant difference in the number of clones between premalignant lesions and stage I carcinomas, thus suggesting that polyclonality originates early in carcinogenesis and not at late clonal expansions. These observations suggest a model of initiation which is based on the expansion of an *APC*-driven clone that constitutes the founder progenitor of the tumor cell population early in carcinogenesis. The mutational profile of this founder clone contains the catalog of "public" mutations that are thus subsequently present in all tumor cells and include those cooperating with *APC*. The progeny of this major clone will then acquire additional mutations that are the source of private low-frequency events that remain less abundant, as these minor subclones are marginally distributed within the geography of the tumor mass[49].

Overall, given the limited amount of sample material from premalignant lesions, bioinformatics approach such as ABSOLUTE has allowed us to uncover interesting findings regarding clonality through deconvoluting somatic mutations generated from whole exome sequencing. However, this approach also has some limitations. First, ABSOLUTE and other similar deconvolution methods assume that a locus does not mutate more than once in its evolutionary history and the mutation does not disappear or reverse, which implies a persistent phylogeny[107]. Thus, they assume the clusters of mutations are present at shared cellular frequencies and are indications of tumor populations. In simple cases, the derivation of mutation clusters will be correct; however, if there exist multiple subclones, then they may be incorrectly clustered. This error can be mitigated with high-depth targeted sequencing to more accurately detect the frequencies of mutations[107]. On the other hand, the "persistent" phylogeny assumption has been known to be violated by known phenomena such as revertant mutations and deletion of loci harboring mutations. Given that colorectal premalignant lesions are mostly diploid and harbor far fewer copy number variation events than carcinomas, this issue may not be as significant in premalignant lesions[49]. In addition, the cellular frequencies and phylogeny of mutations also rely on the tumor purity of the sample and the anatomical region of the biopsy [26]. Sequencing the sample at higher depth will improve the sensitivity of detecting low allele-frequency events in premalignant lesions because these samples have been shown to have around 30% dysplastic cell content[49], compared to 80% in carcinoma samples[108, 109]. In addition, multiple anatomical samples sequencing combined with Bayesian modeling can reduce the uncertainty of the mutation phylogeny by borrowing statistical strength across multiple datasets that is lacking in single sample data[107]. Furthermore, somatic copy number variations and somatic mutation often reside in the same region but have unknown phase or genealogical order. Therefore, tools like ABSOLUTE simplify the assumption by estimating a global ratio of aneuploid and euploid cells under a tumor-normal-two-population assuming from somatic copy number data[108].

To overcome limitation of sequencing depth in WES, low purity of premalignant lesions, and the lack of mutational resolution in single sample data. I have performed high depth sequencing and sensitive digital genotyping array on driver genes in bulk tissues and crypts to further elucidate the model of initiation and clonality in early colorectal carcinogenesis. My results reveal that polyclonality is derived from independent crypts with distinct *APC* and *KRAS* alterations demonstrates that the presence of ITH in colorectal premalignancy is abundant and secondary to the presence of multiple independent lineages derived from different crypt progenitors. This observation contrasts with the majority of tumor evolution models that are based on the expansion of a single dominant clone that harbors a truncal mutation and several subsequent driver alterations (mainly in *APC* but also *KRAS*, *TP53* and *PIK3CA*)[84]. These models ignore the degree of ITH that has been already acquired at the premalignant stage with multiple founder clones competing to get selected over the others, thus making the ITH acquired at the carcinoma stage just one snapshot of the entire tumor evolution[110]. Certainly, my observations are complementary to other tumor evolution models and precede them[84]. I speculate that polyclonality in premalignancy depends on crypt interactions by mechanisms involving crypt fission which leads to the recruitment of different independent lineages, and may not contribute further to progression but rather engulf the dominant clone[111].

In chapter 4, I have shown that LS premalignant lesions display an activated immune profile of CD4+ T cells enrichment and up-regulation of proinflammatory and checkpoint molecules. This profile is consistent with a recent report demonstrating that mucosa from LS patients with carcinoma display an activated immune profile of CD4+ T cells infiltration, HLA and checkpoint expression[97]. Interestingly, the observed immune profile is displayed across lesions regardless of histology, size, MMR status, as well as mutational and neoantigen rate. Comparison of immune activation profile between non-hypermutated and hypermutated LS polyps only demonstrated differential upregulation of *CTLA4* (checkpoint), and *FOXP3* (Treg), which is consistent with the development of immune tolerance upon CRC progression. Specifically, the density of *FOXP3*+ T cells infiltrates have been reported to be correlated with neoantigen load in CRC[98].

In addition, neoantigens accumulated along with the acquisition of additional indels that generate new open reading frames in hyper-mutant polyps are more immunogenic than single nucleotide variants. Overall, these observations are consistent with the results from pan-TCGA analysis that indicated that InDel load is more closely associated with overall immunogenicity and response to checkpoint inhibition[99]*.* Therefore, my results challenge the canonical concept that immune activation in LS is a consequence of acquisition of high levels of mutations and neoantigens secondary to MMR deficiency. This concept could be the case in later stages of premalignancy and carcinoma stage. However, immune deregulation could precede the accumulation of genomic aberrations and neoantigen formation in initial steps of carcinogenesis[72]. One possible explanation is the upregulation of proinflammatory cytokines (Interlukin 12A, *IL12A)* observed in LS polyps. *IL12A* is a T-cell stimulating factor which activates the production of Interferon gamma (*INFG)* and Tumor necrosis factor (*TNF)* from T cells. *INFG* plays a critical role in immunity as it is secreted by helper T cells and cytotoxic T cells as part of the adaptive immunity response while *TNF* is a cell signaling protein that is involved in systemic inflammation. The fact that these genes are also shown to be upregulated in LS polyps versus FAP polyps implies an immune response presence in the initial steps of carcinogenesis. However, further studies will be required to understand the mechanisms behind the upregulation of proinflammatory cytokines in the context of MMR-deficient carcinogenesis.

An additional finding of this chapter is the presence of neoantigens enriched for deregulation of additional DNA repair pathways, which may stimulate additional genomic deregulation in alternative DNA repair and other pathways. Accumulation of mutations in target genes involved in immune surveillance contribute in later stages of carcinogenesis to promote immune escape, and further progression such as mutations in the β2-microglobulin gene (*B2M*) that causes the loss of MHC class I antigen presentation[112]. Moreover, I have identified somatic mutations in genes regulating MHC class II (*CIITA*, *RFX5*, and *RFXAP*) that have been reported previously as microsatellite instability targets[17, 97, 113] in one of the LS hypermutant polyps and one of the LS carcinomas. In contrast, FAP polyps accumulated neoantigens that were enriched for the WNT pathway. This accumulation of genomic events in the WNT pathway led to activation of β-catenin in T cells, which has been identified as an important pathway related to immune evasion by Luke et al[94]. They have analyzed more than 8000 TCGA carcinomas samples and classified them into T-cell-inflamed and T-cell-non-inflamed samples. In the T-cell-non-inflamed sample group, they detected mutations in genes involved in WNT pathway, such as *CTNNB1* and inactivating mutations in negative regulators such as *Axin1, Axin2, APC1,* and *APC2*[94, 114]. In addition, immunohistochemistry demonstrated that CD8-positive T cell infiltration and β-catenin levels are inversely correlated[114]. Given that 50% of the MSI-hypermutated carcinomas arise from MMR-deficiency harbors *APC* inactivating mutations, inhibition of WNT signaling pathway can lead to higher T cell infiltration and produce a more favorable immune microenvironment for immune interception strategies.

Collectively, my findings open the field of immunoprevention in LS to checkpoint inhibitors as an immune interception strategy. This class of agents have shown high level of clinical activity in the treatment of stage IV MMR-deficient CRC[18, 19]. My data is particularly compelling for the use of *LAG3* and dual *LAG3/PD1* inhibitors in the prevention space as demonstrated by the upregulation of both molecules in LS polyps. *LAG3* is a molecule found on the cell surface that plays a role in the negative regulation of T-cells and binds MHC II molecules with high affinity[115]. Currently *LAG3* inhibitors are being developed in several clinical trials (Clinicaltrials.gov number NCT01968109, NCT02061761). In the first-in-human

Phase I trials IMP321 showed no dose-limiting toxicity and the side effects were minimal[116, 117].

An additional immunoprevention strategy proposed in this chapter is the development of cancer prevention vaccines based on the presence of frameshift peptides in LS polyps. Cancer vaccines contain cancer-specific peptides injected into patients to boost the immune system ability to recognize and eliminate cancer cells. Cancer-specific frameshift peptides have been detected in MMR-deficient carcinomas using computational tools and existing genomic data from The Cancer Genome Atlas[118-121], and they have been shown to elicit immune response in carcinomas[99, 122] . Therefore, combinations of vaccine approaches and single/dual checkpoint blockage are logical next steps in immuno-prevention development in this hereditary disease.

In chapter 5, my results show that FAP, LS and sporadic conventional adenomas display an epithelial canonical CMS2 phenotype with strong WNT and MYC downstream targets enrichment. Interestingly, MMR deficiency has been observed in only half of LS polyps at diagnosis[123, 124], but becomes nearly universal in advanced adenomas and carcinoma[123]. The majority of LS adenomas included in my study were early lesions (low grade), which likely explains the predominance of CMS2-associated signaling in these samples. Nonetheless, the data presented in chapter 5 in combination with chapter 4 immune profiling of LS polyps leads to a hypothesis that LS polyps transition from an epithelial phenotype (CMS2-like) with some degree of immune activation at early stages that does not become the complete MSI-Immune (CMS1-like) phenotype until further development of dysplasia and complete loss of MMR functioning, which occurs at advanced stages[72]. Additional correlative studies of MMR deficiency and the immune microenvironment in advanced LS lesions are warranted to explore this hypothesis using ex vivo organoid models and longitudinal samples. Secondly, I found that HP and SSA were both enriched for MSI-Immune CMS1 phenotype. While they displayed strong enrichment of immune and JAK-STAT activation, they also showed enrichment in TGFβ activation and stromal signature (Chang et al, submitted). Our results are consistent with previous studies showing that TGFβ activation play an important role in colorectal carcinogenesis. Using human organoid cultures and genome editing technology, Fessler et al investigators have shown that the genetic background of premalignant lesions dictates the dominating response to TGFβ, changing it from a largely apoptotic response in WNT pathway-activated conventional tubular adenomas to a dominant epithelial-mesenchymal transition response in *BRAF*^{V600E}-mutated SSA[68]. Depending on the level of TGFβ on the microenvironment, SSA could progress to either poor-prognosis CMS4 tumors (high in TGFβ signaling) or the good-prognosis CMS1 tumors (low in TGFβ signaling).

In addition, the ability to classify resected premalignant lesions into indolent versus aggressive molecular subtypes may have important clinical utility for colon cancer screening in the future. Specifically, CMS4 carcinomas have an aggressive clinical phenotype as defined by a higher proportion of advanced stage at diagnosis and worse outcomes after surgery and adjuvant chemotherapy[34]. Extrapolating from these data, I hypothesize that CMS4-like premalignant lesions evolve more rapidly in the adenoma-to-carcinoma transformation process and may indicate the need for closer follow-up evaluation of the at-risk normal mucosa than would otherwise be pursued based on current surveillance guidelines. Certainly, several key studies are needed to explore this hypothesis and firmly establish the prognostic utility of premalignant CMS classification. Towards this end, analysis of longitudinal patient cohorts with sufficient clinical outcomes data (e.g. diagnosis of advanced adenoma or carcinoma) would be highly valuable. Although current CMS RF classifier has robust performance on archived tissue specimens, the classifier contains more than 200 genes. A new classifier requiring fewer genes would be ideal in clinical settings.

In conclusions, my data have shed light on the model of initiation step of carcinogenesis, characterization of the immune profile and its relation to mutation and neoantigen load in MMR-deficient carcinogenesis, and sub-classification of premalignancy in association with different CRC subtypes. My findings establish a comprehensive molecular characterization of premalignant lesions and opens the field to novel development of chemoprevention strategies in colorectal premalignancy.

6.2 Future Directions

The presence of multiple independent clones derived from distinct crypts progenitors were detected by distinct *APC* and *KRAS* mutations. It is feasible that some of clones may not contribute further to progression but rather engulf the dominant clone which serves as the “founder” clone of CRC carcinogenesis. To study the clonal evolution dynamics of CRC in detail, I propose the use of single cell exome sequencing of adenoma with carcinoma in-situ to explore the lineage of major clones and inferred common ancestors and chronology of mutation from adenoma to carcinoma progression. We can perform sectioned biopsies on the sample and isolate single cell with flow cytometry or nanogrids that have almost no doublet error. Doublet errors can cause problem when reconstructing phylogenetic lineage[125]. We will use Monovar[126] for single nucleotide mutation detection and genotyping because of its ability to account for allelic dropout, sequencing false-positives errors, and non-uniform coverage. Then, we can apply SCITE[127], which uses a flexible Markov Chain Monte Carlo (MCMC) model to infer chronology of mutations and OncoNEM[128], which clusters cells into subclones and infers ancestral clones. This will provide novel insight on clonal diversity and evolution from adenoma to carcinoma progression.

My study of immune profiling of premalignant lesions from patients with LS reveal upregulation of checkpoints molecules and immunogenic neoantigens, we propose several experiments and analyses to investigate the use of checkpoint inhibitors and vaccine development as chemoprevention strategies. Firstly, neoantigens generated from frameshift peptides (FSP) in microsatellite (MS) loci are more immunogenic than single nucleotide variants[17, 119]. The detection of FSP will require more sensitive bioinformatics tools than the ones used in this dissertation due to read-length limits and sequencing errors that vary across

MS loci[129]. MSMutect[129]improves the detection of FSP by performing realignment of reads in MS loci. Then, we will identify somatic events with a statistical test to account for noise, motif and length of repeats. Secondly, despite filtering for strong neoantigens based on expression levels and binding affinity to MHC, there are additional factors that will contribute to the recognition of neoantigens by T-cells, thus triggering an immune reponse[122]. These factors include the T-cell receptors (TCR) repertoire[130], density of T cells in the microenvironment of premalignant lesions [98], the integrity of the antigen presentation system[119], expression status of checkpoints, and the clonality of the neoantigens[131, 132]. Therefore, I propose a predictive model on immunogenicity of neoantigens followed by experimental validation. My plan is to collect a large cohort of LS premalignant lesions along with clinical and histopathological annotations and perform WES and RNAseq to identify HLA types and neoantigen candidates, and TCR repertoire sequencing to identify T-cell receptor sequences. Then, design a multi-variable linear regression model on (i) binding affinity of MHC molecule to mutant peptide versus wildtype peptide, (ii) mutation status on crucial genes required for antigen presentation machinery, such as *B2M* (MHC I), *CIITA, RFX5 and RFXAP* (MCH II), (iii) sequence similarity between mutant peptide and T cell recognized epitopes given in Immune Epitope Database and Analysis Resource (IEDB)[133], (iv) expression levels of immune activation genes, and (v) the clonal or subclonal status of the neoantigens based on in-silico tools. The response variable of the model will be measurement of antigen-specific T cells response performed with enzyme-linked immunosorbent spot (ELISPOT), a widely use method for measuring antigen-specific T cells[134]. Finally, vaccines can be developed based on predicted neoantigens and evaluated using in-vitro systems such as organoids derived from patient’s adenoma tissue and co-cultured lymphocytes[135].

The large-scale transcriptomic analysis of sporadic and hereditary shows that classification of resected premalignant lesions into CMS subtypes can offer clinical utility for CRC screening. Current follow-up guidelines after initial discovery of premalignancy depends on histology, size, multiplicity, and family history[136], and CMS-like molecular subtype classification can improve risk prediction by providing important molecular features that is not captured by current morphological guidelines. Therefore, it will be crucial to perform a longitudinal study by tracking premalignant lesions or carcinomas occurrence and evaluate the prediction performance of molecular features. In addition, a new classification model may be desired as the current classifier is specifically optimized for carcinomas. I will first conduct a search for consistently expressed genes across the various types of polyps and carcinomas and retrain a random forest classifier to predict CMS-polyp subtypes in a large database of carcinoma samples with known CMS-carcinoma status (the original CMS classification was developed from 4500 carcinomas). Performance metrics will be assessed and compared with the original CMS-carcinoma classifiers to ensure that accuracy is not severely compromised with this new model. Then, I will apply the CMS-polyps categories to the polyp samples data and assess for gene enrichment across CMS-polyp groups. If strong associations between the CMS-carcinoma and -polyp categories are found, I will integrate the histopathological categorization of polyps into the novel framework and thereby associate carcinoma subtypes with their premalignant origins.

In summary, I propose using single cell sequencing technology to study clonal evolution from premalignant to carcinoma stages, designing a predictive model for immunogenicity of neoantigens, studying the efficacy of cancer prevention vaccines and checkpoint inhibitors in MMR-deficient premalignancy, designing a transcriptomic classifier optimized for premalignancy and carcinomas and evaluating the performance of subtype classification for predicting premalignant lesions or carcinoma occurrence.

References

1. Siegel RL, Miller KD, Jemal A: **Cancer statistics, 2018**. *CA Cancer J Clin* 2018, **68**(1):7-30.
2. Mork ME, You YN, Ying J, Bannon SA, Lynch PM, Rodriguez-Bigas MA, Vilar E: **High Prevalence of Hereditary Cancer Syndromes in Adolescents and Young Adults With Colorectal Cancer**. *J Clin Oncol* 2015, **33**(31):3544-3549.
3. Fearon ER: **Molecular genetics of colorectal cancer**. *Annu Rev Pathol* 2011, **6**:479-507.
4. Galiatsatos P, Foulkes WD: **Familial adenomatous polyposis**. *Am J Gastroenterol* 2006, **101**(2):385-398.
5. Fearon ER, Vogelstein B: **A genetic model for colorectal tumorigenesis**. *Cell* 1990, **61**(5):759-767.
6. Kohlmann W, Gruber SB: **Lynch Syndrome**. In: *GeneReviews(R).* Edited by Pagon RA, Adam MP, Ardinger HH, Wallace SE, Amemiya A, Bean LJH, Bird TD, Fong CT, Mefford HC, Smith RJH *et al*. Seattle (WA); 1993.
7. Spira A, Disis ML, Schiller JT, Vilar E, Rebbeck TR, Bejar R, Ideker T, Arts J, Yurgelun MB, Mesirov JP *et al*: **Leveraging premalignant biology for immune-based cancer prevention**. *Proc Natl Acad Sci U S A* 2016.
8. Bonadona V, Bonaiti B, Olschwang S, Grandjouan S, Huiart L, Longy M, Guimbaud R, Buecher B, Bignon YJ, Caron O *et al*: **Cancer risks associated with germline mutations in MLH1, MSH2, and MSH6 genes in Lynch syndrome**. *Jama* 2011, **305**(22):2304-2310.
9. Ligtenberg MJ, Kuiper RP, Geurts van Kessel A, Hoogerbrugge N: **EPCAM deletion carriers constitute a unique subgroup of Lynch syndrome patients**. *Fam Cancer* 2013, **12**(2):169-174.

10. Jiricny J: **The multifaceted mismatch-repair system**. *Nat Rev Mol Cell Biol* 2006, **7**(5):335-346.

11. Network TCGA: **Comprehensive molecular characterization of human colon and rectal cancer**. *Nature* 2012, **487**(7407):330-337.

12. Boland CR, Goel A: **Microsatellite instability in colorectal cancer**. *Gastroenterology* 2010, **138**(6):2073-2087 e2073.

13. Jarvinen HJ, Aarnio M, Mustonen H, Aktan-Collan K, Aaltonen LA, Peltomaki P, De La Chapelle A, Mecklin JP: **Controlled 15-year trial on screening for colorectal cancer in families with hereditary nonpolyposis colorectal cancer**. *Gastroenterology* 2000, **118**(5):829-834.

14. Moller P, Seppala T, Bernstein I, Holinski-Feder E, Sala P, Evans DG, Lindblom A, Macrae F, Blanco I, Sijmons R *et al*: **Cancer incidence and survival in Lynch syndrome patients receiving colonoscopic and gynaecological surveillance: first report from the prospective Lynch syndrome database**. *Gut* 2017, **66**(3):464-472.

15. Smyrk TC, Watson P, Kaul K, Lynch HT: **Tumor-infiltrating lymphocytes are a marker for microsatellite instability in colorectal carcinoma**. *Cancer* 2001, **91**(12):2417-2422.

16. Llosa NJ, Cruise M, Tam A, Wicks EC, Hechenbleikner EM, Taube JM, Blosser RL, Fan H, Wang H, Luber BS *et al*: **The vigorous immune microenvironment of microsatellite instable colon cancer is balanced by multiple counter-inhibitory checkpoints**. *Cancer discovery* 2015, **5**(1):43-51.

17. Kloor M, Michel S, von Knebel Doeberitz M: **Immune evasion of microsatellite unstable colorectal cancers**. *International journal of cancer Journal international du cancer* 2010, **127**(5):1001-1010.

18. Overman MJ, McDermott R, Leach JL, Lonardi S, Lenz HJ, Morse MA, Desai J, Hill A, Axelson M, Moss RA *et al*: **Nivolumab in patients with metastatic DNA mismatch**

repair-deficient or microsatellite instability-high colorectal cancer (CheckMate 142): an open-label, multicentre, phase 2 study. *The lancet oncology* 2017, **18**(9):1182-1191.

19. Le DT, Durham JN, Smith KN, Wang H, Bartlett BR, Aulakh LK, Lu S, Kemberling H, Wilt C, Luber BS *et al*: **Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade**. *Science* 2017, **357**(6349):409-413.

20. Le DT, Uram JN, Wang H, Bartlett BR, Kemberling H, Eyring AD, Skora AD, Luber BS, Azad NS, Laheru D *et al*: **PD-1 Blockade in Tumors with Mismatch-Repair Deficiency**. *N Engl J Med* 2015, **372**(26):2509-2520.

21. Guarinos C, Sanchez-Fortun C, Rodriguez-Soler M, Alenda C, Paya A, Jover R: **Serrated polyposis syndrome: molecular, pathological and clinical aspects**. *World journal of gastroenterology* 2012, **18**(20):2452-2461.

22. Rex DK, Ahnen DJ, Baron JA, Batts KP, Burke CA, Burt RW, Goldblum JR, Guillem JG, Kahi CJ, Kalady MF *et al*: **Serrated lesions of the colorectum: review and recommendations from an expert panel**. *Am J Gastroenterol* 2012, **107**(9):1315-1329; quiz 1314, 1330.

23. Huang CS, O'Brien M J, Yang S, Farraye FA: **Hyperplastic polyps, serrated adenomas, and the serrated polyp neoplasia pathway**. *Am J Gastroenterol* 2004, **99**(11):2242-2255.

24. Humphries A, Wright NA: **Colonic crypt organization and tumorigenesis**. *Nature reviews Cancer* 2008, **8**(6):415-424.

25. Barker N, Ridgway RA, van Es JH, van de Wetering M, Begthel H, van den Born M, Danenberg E, Clarke AR, Sansom OJ, Clevers H: **Crypt stem cells as the cells-of-origin of intestinal cancer**. *Nature* 2009, **457**(7229):608-611.

26. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, Laird PW, Onofrio RC, Winckler W, Weir BA *et al*: **Absolute quantification of somatic DNA alterations in human cancer**. *Nature biotechnology* 2012, **30**(5):413-421.

27. Andor N, Graham TA, Jansen M, Xia LC, Aktipis CA, Petritsch C, Ji HP, Maley CC: **Pan-cancer analysis of the extent and consequences of intratumor heterogeneity**. *Nature medicine* 2016, **22**(1):105-113.

28. The Cancer Genome Atlas Network: **Comprehensive molecular characterization of human colon and rectal cancer**. *Nature* 2012, **487**(7407):330-337.

29. De Sousa EMF, Wang X, Jansen M, Fessler E, Trinh A, de Rooij LP, de Jong JH, de Boer OJ, van Leersum R, Bijlsma MF *et al*: **Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions**. *Nature medicine* 2013, **19**(5):614-618.

30. Sadanandam A, Lyssiotis CA, Homicsko K, Collisson EA, Gibb WJ, Wullschleger S, Ostos LC, Lannon WA, Grotzinger C, Del Rio M *et al*: **A colorectal cancer classification system that associates cellular phenotype and responses to therapy**. *Nature medicine* 2013, **19**(5):619-625.

31. Budinska E, Popovici V, Tejpar S, D'Ario G, Lapique N, Sikora KO, Di Narzo AF, Yan P, Hodgson JG, Weinrich S *et al*: **Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer**. *J Pathol* 2013, **231**(1):63-76.

32. Roepman P, Schlicker A, Tabernero J, Majewski I, Tian S, Moreno V, Snel MH, Chresta CM, Rosenberg R, Nitsche U *et al*: **Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition**. *International journal of cancer Journal international du cancer* 2014, **134**(3):552-562.

33. Schlicker A, Beran G, Chresta CM, McWalter G, Pritchard A, Weston S, Runswick S, Davenport S, Heathcote K, Castro DA *et al*: **Subtypes of primary colorectal tumors**

correlate with response to targeted treatment in colorectal cell lines. *BMC Med Genomics* 2012, **5**:66.

34. Guinney J, Dienstmann R, Wang X, de Reynies A, Schlicker A, Soneson C, Marisa L, Roepman P, Nyamundanda G, Angelino P *et al*: **The consensus molecular subtypes of colorectal cancer**. *Nature medicine* 2015, **21**(11):1350-1356.

35. Angelova M, Charoentong P, Hackl H, Fischer ML, Snajder R, Krogsdam AM, Waldner MJ, Bindea G, Mlecnik B, Galon J *et al*: **Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy**. *Genome Biol* 2015, **16**:64.

36. Dienstmann R, Vermeulen L, Guinney J, Kopetz S, Tejpar S, Tabernero J: **Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer**. *Nature reviews Cancer* 2017, **17**(2):79-92.

37. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754-1760.

38. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M *et al*: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data**. *Genome Res* 2010, **20**(9):1297-1303.

39. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G: **Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples**. *Nature biotechnology* 2013, **31**(3):213-219.

40. **Broad Institute Cancer Genome Analysis: Indelocator** [http://www.broadinstitute.org/cancer/cga/indelocator]

41. Wang K, Li M, Hakonarson H: **ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data**. *Nucleic Acids Res* 2010, **38**(16):e164.

42. San Lucas FA, Wang G, Scheet P, Peng B: **Integrated annotation and analysis of genetic variants from next-generation sequencing studies with variant tools**. *Bioinformatics* 2012, **28**(3):421-422.

43. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP: **Integrative genomics viewer**. *Nat Biotech* 2011, **29**(1):24-26.

44. Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C: **DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution**. *Genome Biol* 2016, **17**:31.

45. Vattathil S, Scheet P: **Haplotype-based profiling of subtle allelic imbalance with SNP arrays**. *Genome Res* 2013, **23**(1):152-158.

46. San Lucas FA, Sivakumar S, Vattathil S, Fowler J, Vilar E, Scheet P: **Rapid and powerful detection of subtle allelic imbalance from exome sequencing data with hapLOHseq**. *Bioinformatics* 2016, **32**(19):3015-3017.

47. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK: **VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing**. *Genome Res* 2012, **22**(3):568-576.

48. Seshan VE OA: **DNAcopy: DNA copy number data analysis**. In.; 2018.

49. Borras E, San Lucas FA, Chang K, Zhou R, Masand G, Fowler J, Mork ME, You YN, Taggart MW, McAllister F *et al*: **Genomic Landscape of Colorectal Mucosa and Adenomas**. *Cancer Prev Res (Phila)* 2016, **9**(6):417-427.

50. Habano W, Sugai T, Nakamura S, Yoshida T: **A novel method for gene analysis of colorectal carcinomas using a crypt isolation technique**. *Laboratory investigation; a journal of technical methods and pathology* 1996, **74**(5):933-940.

51. Cheng H, Bjerknes M, Amar J: **Methods for the determination of epithelial cell kinetic parameters of human colonic epithelium isolated from surgical and biopsy specimens**. *Gastroenterology* 1984, **86**(1):78-85.

52. Azuara D, Ginesta MM, Gausachs M, Rodriguez-Moranta F, Fabregat J, Busquets J, Pelaez N, Boadas J, Galter S, Moreno V *et al*: **Nanofluidic digital PCR for KRAS mutation detection and quantification in gastrointestinal cancer**. *Clinical chemistry* 2012, **58**(9):1332-1341.

53. Ryota Suzuki HS: **pvclust**. In.; 2017.

54. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: **STAR: ultrafast universal RNA-seq aligner**. *Bioinformatics* 2013, **29**(1):15-21.

55. Li B, Dewey CN: **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome**. *BMC bioinformatics* 2011, **12**:323.

56. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK: **limma powers differential expression analyses for RNA-sequencing and microarray studies**. *Nucleic Acids Res* 2015, **43**(7):e47.

57. Hanzelmann S, Castelo R, Guinney J: **GSVA: gene set variation analysis for microarray and RNA-seq data**. *BMC bioinformatics* 2013, **14**:7.

58. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J *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.

59. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA *et al*: **A global reference for human genetic variation**. *Nature* 2015, **526**(7571):68-74.

60. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB *et al*: **Analysis of protein-coding genetic variation in 60,706 humans**. *Nature* 2016, **536**(7616):285-291.

61. Boegel S, Scholtalbers J, Lower M, Sahin U, Castle JC: **In Silico HLA Typing Using Standard RNA-Seq Sequence Reads**. *Methods in molecular biology* 2015, **1310**:247-258.

62. Hundal J, Carreno BM, Petti AA, Linette GP, Griffith OL, Mardis ER, Griffith M: **pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens**. *Genome medicine* 2016, **8**(1):11.

63. Nielsen M, Andreatta M: **NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets**. *Genome medicine* 2016, **8**(1):33.

64. Jensen KK, Andreatta M, Marcatili P, Buus S, Greenbaum JA, Yan Z, Sette A, Peters B, Nielsen M: **Improved methods for predicting peptide binding affinity to MHC class II molecules**. *Immunology* 2018.

65. Galamb O, Sipos F, Solymosi N, Spisak S, Krenacs T, Toth K, Tulassay Z, Molnar B: **Diagnostic mRNA expression patterns of inflamed, benign, and malignant colorectal biopsy specimen and their correlation with peripheral blood results**. *Cancer Epidemiol Biomarkers Prev* 2008, **17**(10):2835-2845.

66. Galamb O, Gyorffy B, Sipos F, Spisak S, Nemeth AM, Miheller P, Tulassay Z, Dinya E, Molnar B: **Inflammation, adenoma and cancer: objective classification of colon biopsy specimens with gene expression signature**. *Dis Markers* 2008, **25**(1):1-16.

67. Sabates-Bellver J, Van der Flier LG, de Palo M, Cattaneo E, Maake C, Rehrauer H, Laczko E, Kurowski MA, Bujnicki JM, Menigatti M *et al*: **Transcriptome profile of human colorectal adenomas**. *Mol Cancer Res* 2007, **5**(12):1263-1275.

68. Fessler E, Drost J, van Hooff SR, Linnekamp JF, Wang X, Jansen M, De Sousa EMF, Prasetyanti PR, JE IJ, Franitza M *et al*: **TGFbeta signaling directs serrated adenomas to the mesenchymal colorectal cancer subtype**. *EMBO Mol Med* 2016.

69. Delker DA, McGettigan BM, Kanth P, Pop S, Neklason DW, Bronner MP, Burt RW, Hagedorn CH: **RNA sequencing of sessile serrated colon polyps identifies differentially expressed genes and immunohistochemical markers**. *PloS one* 2014, **9**(2):e88367.

70. Kanth P, Bronner MP, Boucher KM, Burt R, Neklason DW, Hagedorn CH, Delker DA: **Gene Signature in Sessile Serrated Polyps Identifies Colon Cancer Subtype**. *Cancer Prev Res (Phila)* 2016.

71. Crespo M, Vilar E, Tsai SY, Chang K, Amin S, Srinivasan T, Zhang T, Pipalia NH, Chen HJ, Witherspoon M *et al*: **Colonic organoids derived from human induced pluripotent stem cells for modeling colorectal cancer and drug testing**. *Nature medicine* 2017, **23**(7):878-884.

72. Chang K, Taggart MW, Reyes-Uribe L, et al.: **Immune profiling of premalignant lesions in patients with lynch syndrome**. *JAMA Oncology* 2018.

73. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, Kling DE, Gauthier LD, Levy-Moonshine A, Roazen D *et al*: **Scaling accurate genetic variant discovery to tens of thousands of samples**. *bioRxiv* 2017.

74. Ramos AH, Lichtenstein L, Gupta M, Lawrence MS, Pugh TJ, Saksena G, Meyerson M, Getz G: **Oncotator: cancer variant annotation tool**. *Human mutation* 2015, **36**(4):E2423-2429.

75. McCall MN, Jaffee HA, Irizarry RA: **fRMA ST: frozen robust multiarray analysis for Affymetrix Exon and Gene ST arrays**. *Bioinformatics* 2012, **28**(23):3153-3154.

76. McCall MN, Irizarry RA: **McCall MN and Irizarry RA. hgu133plus2frmavecs: Vectors used by frma for microarrays of type hgu133plus2. R package version 1.5.0.** In.

77. Gautier L, Cope L, Bolstad BM, Irizarry RA: **affy--analysis of Affymetrix GeneChip data at the probe level**. *Bioinformatics* 2004, **20**(3):307-315.

78. Johnson WE, Li C, Rabinovic A: **Adjusting batch effects in microarray expression data using empirical Bayes methods**. *Biostatistics* 2007, **8**(1):118-127.

79. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD: **The sva package for removing batch effects and other unwanted variation in high-throughput experiments**. *Bioinformatics* 2012, **28**(6):882-883.

80. Kauffmann A, Gentleman R, Huber W: **arrayQualityMetrics--a bioconductor package for quality assessment of microarray data**. *Bioinformatics* 2009, **25**(3):415-416.

81. Yoshihara K, Shahmoradgoli M, Martinez E, Vegesna R, Kim H, Torres-Garcia W, Trevino V, Shen H, Laird PW, Levine DA *et al*: **Inferring tumour purity and stromal and immune cell admixture from expression data**. *Nat Commun* 2013, **4**:2612.

82. Sergushichev A: **An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation**. *bioRxiv* 2016.

83. Nowell PC: **The clonal evolution of tumor cell populations**. *Science* 1976, **194**(4260):23-28.

84. Sottoriva A, Kang H, Ma Z, Graham TA, Salomon MP, Zhao J, Marjoram P, Siegmund K, Press MF, Shibata D *et al*: **A Big Bang model of human colorectal tumor growth**. *Nature genetics* 2015, **47**(3):209-216.

85. Thliveris AT, Schwefel B, Clipson L, Plesh L, Zahm CD, Leystra AA, Washington MK, Sullivan R, Deming DA, Newton MA *et al*: **Transformation of epithelial cells through recruitment leads to polyclonal intestinal tumors**. *Proc Natl Acad Sci U S A* 2013, **110**(28):11523-11528.

86. Thliveris AT, Halberg RB, Clipson L, Dove WF, Sullivan R, Washington MK, Stanhope S, Newton MA: **Polyclonality of familial murine adenomas: analyses of mouse chimeras with low tumor multiplicity suggest short-range interactions**. *Proc Natl Acad Sci U S A* 2005, **102**(19):6960-6965.

87. Thirlwell C, Will OC, Domingo E, Graham TA, McDonald SA, Oukrif D, Jeffrey R, Gorman M, Rodriguez-Justo M, Chin-Aleong J *et al*: **Clonality assessment and clonal ordering of individual neoplastic crypts shows polyclonality of colorectal adenomas**. *Gastroenterology* 2010, **138**(4):1441-1454, 1454 e1441-1447.

88. Lamlum H, Papadopoulou A, Ilyas M, Rowan A, Gillet C, Hanby A, Talbot I, Bodmer W, Tomlinson I: **APC mutations are sufficient for the growth of early colorectal adenomas**. *Proc Natl Acad Sci U S A* 2000, **97**(5):2225-2228.

89. Segditsas S, Rowan AJ, Howarth K, Jones A, Leedham S, Wright NA, Gorman P, Chambers W, Domingo E, Roylance RR *et al*: **APC and the three-hit hypothesis**. *Oncogene* 2009, **28**(1):146-155.

90. Zhu D, Keohavong P, Finkelstein SD, Swalsky P, Bakker A, Weissfeld J, Srivastava S, Whiteside TL: **K-ras gene mutations in normal colorectal tissues from K-ras mutation-positive colorectal cancer patients**. *Cancer Res* 1997, **57**(12):2485-2492.

91. Wong WM, Mandir N, Goodlad RA, Wong BC, Garcia SB, Lam SK, Wright NA: **Histogenesis of human colorectal adenomas and hyperplastic polyps: the role of cell proliferation and crypt fission**. *Gut* 2002, **50**(2):212-217.

92. Wasan HS, Park HS, Liu KC, Mandir NK, Winnett A, Sasieni P, Bodmer WF, Goodlad RA, Wright NA: **APC in the regulation of intestinal crypt fission**. *J Pathol* 1998, **185**(3):246-255.

93. Leedham SJ, Wright NA: **Expansion of a mutated clone: from stem cell to tumour**. *J Clin Pathol* 2008, **61**(2):164-171.

94. Spranger S, Luke JJ, Bao R, Zha Y, Hernandez KM, Li Y, Gajewski AP, Andrade J, Gajewski TF: **Density of immunogenic antigens does not explain the presence or absence of the T-cell-inflamed tumor microenvironment in melanoma**. *Proc Natl Acad Sci U S A* 2016, **113**(48):E7759-E7768.

95. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL *et al*: **Signatures of mutational processes in human cancer**. *Nature* 2013, **500**(7463):415-421.

96. Boegel S, Lower M, Schafer M, Bukur T, de Graaf J, Boisguerin V, Tureci O, Diken M, Castle JC, Sahin U: **HLA typing from RNA-Seq sequence reads**. *Genome medicine* 2012, **4**(12):102.

97. Binder H, Hopp L, Schweiger MR, Hoffmann S, Juhling F, Kerick M, Timmermann B, Siebert S, Grimm C, Nersisyan L *et al*: **Genomic and transcriptomic heterogeneity of colorectal tumours arising in Lynch syndrome**. *J Pathol* 2017, **243**(2):242-254.

98. Giannakis M, Mu XJ, Shukla SA, Qian ZR, Cohen O, Nishihara R, Bahl S, Cao Y, Amin-Mansour A, Yamauchi M *et al*: **Genomic Correlates of Immune-Cell Infiltrates in Colorectal Carcinoma**. *Cell Rep* 2016, **17**(4):1206.

99. Turajlic S, Litchfield K, Xu H, Rosenthal R, McGranahan N, Reading JL, Wong YNS, Rowan A, Kanu N, Al Bakir M *et al*: **Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis**. *Lancet Oncol* 2017, **18**(8):1009-1021.

100. Liljegren A, Barker G, Elliott F, Bertario L, Bisgaard ML, Eccles D, Evans G, Macrae F, Maher E, Lindblom A *et al*: **Prevalence of adenomas and hyperplastic polyps in mismatch repair mutation carriers among CAPP2 participants: report by the colorectal adenoma/carcinoma prevention programme 2**. *J Clin Oncol* 2008, **26**(20):3434-3439.

101. Yurgelun MB, Goel A, Hornick JL, Sen A, Turgeon DK, Ruffin Iv MT, Marcon NE, Baron JA, Bresalier RS, Syngal S *et al*: **Microsatellite instability and DNA mismatch repair protein deficiency in Lynch syndrome colorectal polyps**. *Cancer Prev Res (Phila)* 2012.

102. Syngal S, Brand RE, Church JM, Giardiello FM, Hampel HL, Burt RW, American College of G: **ACG clinical guideline: Genetic testing and management of hereditary gastrointestinal cancer syndromes**. *Am J Gastroenterol* 2015, **110**(2):223-262; quiz 263.

103. Bailey CE, Hu CY, You YN, Bednarski BK, Rodriguez-Bigas MA, Skibber JM, Cantor SB, Chang GJ: **Increasing disparities in the age-related incidences of colon and rectal cancers in the United States, 1975-2010**. *JAMA Surg* 2015, **150**(1):17-22.

104. Fontana E, Ragulan C, Eason K, Si-Lin K, Siew TW, Nyamundanda G, Patil Y, Poudel P, Chau I, Tan IB *et al*: **145OValidated nCounter platform to stratify colorectal cancer (CRC) into Consensus Molecular Subtypes (CMS) and CRCassigner subtypes in Asian population**. *Annals of Oncology* 2017, **28**(suppl_10):mdx659.003-mdx659.003.

105. Roh W, Chen PL, Reuben A, Spencer CN, Prieto PA, Miller JP, Gopalakrishnan V, Wang F, Cooper ZA, Reddy SM *et al*: **Integrated molecular analysis of tumor biopsies on sequential CTLA-4 and PD-1 blockade reveals markers of response and resistance**. *Sci Transl Med* 2017, **9**(379).

106. Kang H, Salomon MP, Sottoriva A, Zhao J, Toy M, Press MF, Curtis C, Marjoram P, Siegmund K, Shibata D: **Many private mutations originate from the first few divisions of a human colorectal adenoma**. *J Pathol* 2015, **237**(3):355-362.

107. Roth A, Khattra J, Yap D, Wan A, Laks E, Biele J, Ha G, Aparicio S, Bouchard-Cote A, Shah SP: **PyClone: statistical inference of clonal population structure in cancer**. *Nat Methods* 2014, **11**(4):396-398.

108. Li B, Li JZ: **A general framework for analyzing tumor subclonality using SNP array and DNA sequencing data**. *Genome Biol* 2014, **15**(9):473.

109. Yadav VK, De S: **An assessment of computational methods for estimating purity and clonality using genomic data derived from heterogeneous tumor tissue samples**. *Brief Bioinform* 2015, **16**(2):232-241.

110. Gausachs M, Borras E, Chang K, Gonzalez S, Azuara D, Delgado Amador A, Lopez-Doriga A, San Lucas FA, Sanjuan X, Paules MJ *et al*: **Mutational Heterogeneity in APC and KRAS Arises at the Crypt Level and Leads to Polyclonality in Early Colorectal Tumorigenesis**. *Clin Cancer Res* 2017, **23**(19):5936-5947.

111. Martincorena I, Roshan A, Gerstung M, Ellis P, Van Loo P, McLaren S, Wedge DC, Fullam A, Alexandrov LB, Tubio JM *et al*: **Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin**. *Science* 2015, **348**(6237):880-886.

112. Kloor M, Michel S, Buckowitz B, Ruschoff J, Buttner R, Holinski-Feder E, Dippold W, Wagner R, Tariverdian M, Benner A *et al*: **Beta2-microglobulin mutations in microsatellite unstable colorectal tumors**. *Int J Cancer* 2007, **121**(2):454-458.

113. Lee J, Li L, Gretz N, Gebert J, Dihlmann S: **Absent in Melanoma 2 (AIM2) is an important mediator of interferon-dependent and -independent HLA-DRA and HLA-DRB gene expression in colorectal cancers**. *Oncogene* 2012, **31**(10):1242-1253.

114. Pai SG, Carneiro BA, Mota JM, Costa R, Leite CA, Barroso-Sousa R, Kaplan JB, Chae YK, Giles FJ: **Wnt/beta-catenin pathway: modulating anticancer immune response**. *J Hematol Oncol* 2017, **10**(1):101.

115. Nguyen LT, Ohashi PS: **Clinical blockade of PD1 and LAG3--potential mechanisms of action**. *Nat Rev Immunol* 2015, **15**(1):45-56.

116. Brignone C, Grygar C, Marcu M, Perrin G, Triebel F: **IMP321 (sLAG-3) safety and T cell response potentiation using an influenza vaccine as a model antigen: a single-blind phase I study**. *Vaccine* 2007, **25**(24):4641-4650.

117. Brignone C, Gutierrez M, Mefti F, Brain E, Jarcau R, Cvitkovic F, Bousetta N, Medioni J, Gligorov J, Grygar C *et al*: **First-line chemoimmunotherapy in metastatic breast carcinoma: combination of paclitaxel and IMP321 (LAG-3Ig) enhances immune responses and antitumor activity**. *J Transl Med* 2010, **8**:71.

118. Leoni G, D'Alise AM, Cotugno G, Mori F, Catanese MT, Langone F, Fichera I, De Lucia M, Vitale R, Leuzzi A *et al*: **A viral vectored vaccine based on shared tumor neoantigens for prevention and treatment of microsatellite instable (MSI) cancers**. *Journal for ImmunoTherapy of Cancer* 2017, **5**((Suppl 2):86):P139.

119. Schwitalle Y, Kloor M, Eiermann S, Linnebacher M, Kienle P, Knaebel HP, Tariverdian M, Benner A, von Knebel Doeberitz M: **Immune response against frameshift-induced neopeptides in HNPCC patients and healthy HNPCC mutation carriers**. *Gastroenterology* 2008, **134**(4):988-997.

120. von Knebel Doeberitz M, Kloor M: **Towards a vaccine to prevent cancer in Lynch syndrome patients**. *Fam Cancer* 2013, **12**(2):307-312.

121. Hause RJ, Pritchard CC, Shendure J, Salipante SJ: **Classification and characterization of microsatellite instability across 18 cancer types**. *Nat Med* 2016, **22**(11):1342-1350.

122. Yarchoan M, Johnson BA, 3rd, Lutz ER, Laheru DA, Jaffee EM: **Targeting neoantigens to augment antitumour immunity**. *Nature reviews Cancer* 2017, **17**(4):209-222.

123. Yurgelun MB, Goel A, Hornick JL, Sen A, Turgeon DK, Ruffin MTt, Marcon NE, Baron JA, Bresalier RS, Syngal S *et al*: **Microsatellite instability and DNA mismatch repair protein deficiency in Lynch syndrome colorectal polyps**. *Cancer Prev Res (Phila)* 2012, **5**(4):574-582.

124. Walsh MD, Buchanan DD, Pearson SA, Clendenning M, Jenkins MA, Win AK, Walters RJ, Spring KJ, Nagler B, Pavluk E *et al*: **Immunohistochemical testing of**

conventional adenomas for loss of expression of mismatch repair proteins in Lynch syndrome mutation carriers: a case series from the Australasian site of the colon cancer family registry. *Mod Pathol* 2012, **25**(5):722-730.

125. Navin NE, Chen K: **Genotyping tumor clones from single-cell data**. *Nat Methods* 2016, **13**(7):555-556.

126. Zafar H, Wang Y, Nakhleh L, Navin N, Chen K: **Monovar: single-nucleotide variant detection in single cells**. *Nat Methods* 2016, **13**(6):505-507.

127. Jahn K, Kuipers J, Beerenwinkel N: **Tree inference for single-cell data**. *Genome Biol* 2016, **17**:86.

128. Ross EM, Markowetz F: **OncoNEM: inferring tumor evolution from single-cell sequencing data**. *Genome Biol* 2016, **17**:69.

129. Maruvka YE, Mouw KW, Karlic R, Parasuraman P, Kamburov A, Polak P, Haradhvala NJ, Hess JM, Rheinbay E, Brody Y *et al*: **Analysis of somatic microsatellite indels identifies driver events in human tumors**. *Nature biotechnology* 2017, **35**(10):951-959.

130. Rosati E, Dowds CM, Liaskou E, Henriksen EKK, Karlsen TH, Franke A: **Overview of methodologies for T-cell receptor repertoire analysis**. *BMC Biotechnol* 2017, **17**(1):61.

131. McGranahan N, Swanton C: **Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future**. *Cell* 2017, **168**(4):613-628.

132. Caswell DR, Swanton C: **The role of tumour heterogeneity and clonal cooperativity in metastasis, immune evasion and clinical outcome**. *BMC Med* 2017, **15**(1):133.

133. Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, Cantrell JR, Wheeler DK, Gabbard JL, Hix D, Sette A *et al*: **The immune epitope database (IEDB) 3.0**. *Nucleic Acids Res* 2015, **43**(Database issue):D405-412.

134. Slota M, Lim JB, Dang Y, Disis ML: **ELISpot for measuring human immune responses to vaccines**. *Expert Rev Vaccines* 2011, **10**(3):299-306.

135. Nozaki K, Mochizuki W, Matsumoto Y, Matsumoto T, Fukuda M, Mizutani T, Watanabe M, Nakamura T: **Co-culture with intestinal epithelial organoids allows efficient expansion and motility analysis of intraepithelial lymphocytes**. *J Gastroenterol* 2016, **51**(3):206-213.

136. Schreuders EH, Ruco A, Rabeneck L, Schoen RE, Sung JJ, Young GP, Kuipers EJ: **Colorectal cancer screening: a global overview of existing programmes**. *Gut* 2015, **64**(10):1637-1649.

VITA

Kyle Chang was born in Hong Kong, the son of David Tai Wai Chang and Tang Man Ching. He completed his Bachelor’s Degree in Electrical and Computer Engineering at Carnegie Mellon University. He worked as a Front-end Developer for Invesco and as a Senior Bioinformatics Programmer for the Human Genome Sequencing Center at Baylor College of Medicine before entering The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences for his PhD.