8-2014

Genomic Characterization Of Polyps In Familial Adenomatous Polyposis Patients And Identification Of Candidate Chemopreventive Drugs

Francis A. San Lucas

GENOMIC CHARACTERIZATION OF POLYPS IN FAMILIAL ADENOMATOUS POLYPOSIS PATIENTS AND IDENTIFICATION OF CANDIDATE CHEMOPREVENTIVE DRUGS

by

Francis Anthony San Lucas, M.S.

APPROVED:

____
Advisory Professor: Paul Scheet, Ph.D.

____
Secondary Advisor: Eduardo Vilar, M.D., Ph.D.

____
James Hixson, Ph.D.

____
Yin Liu, Ph.D.

____
Ignacio Wistuba, M.D.

APPROVED:

____
Dean, The University of Texas
Graduate School of Biomedical Sciences at Houston

This dissertation is dedicated to my family. To my parents, Roy and Maria San Lucas, for their love and guidance through my life. To my daughters Sarai, Abbey and Natale San Lucas who continually help me to maintain perspective, who teach me to appreciate all of life's little discoveries, who constantly put a smile on my face, and who motivate me to help build a better future for this world. Most importantly, to my wife and best friend, Tien San Lucas, for her unwavering support in every way imaginable and for her encouragement and love over the years.

Acknowledgements

I thank my advisor and friend Paul Scheet for his extreme generosity in supporting my scientific development, for his wisdom and mentoring in preparing me for a career in academic research, for his desire to instill statistical intuition in me, for allowing me to lead many aspects of our research, and for devoting time to our countless research-related and personal discussions. I also thank my secondary advisor Eduardo Vilar for all of the time he dedicated to teaching me about *Familial Adenomatous Polyposis* (FAP), cancer prevention and drug discovery, for his guidance in planning, executing and publishing translational research, for his steady positive attitude when things go wrong, for treating me in many ways like a peer, and for his constant encouragement to strive for a career as a principal investigator.

I thank all of my lab peers for their countless research meetings and discussions. I especially thank Selina Vattathil for sharing so much of what she has learned in her development of *hapLOH* with me and everyone in our research group, which has provided a substantial foundation for my development of *hapLOHseq*. I thank Jerry Fowler for his expertise and creativity in many software development projects, such as developing a framework to automate sequencing analysis pipelines, and for his helpful discussions and advice on algorithms development. I thank Bo Peng for his leadership and his intense work ethic during the development of variant tools, and I thank Gao Wang for his very positive attitude and for his support and development of variant tools.

I thank my supervisory committee members for helping to shape my dissertation project and my future. Specifically, I thank Ignacio Wistuba for his insightful research questions and suggestions, for his career advice and for his upcoming support of my postdoctoral research. I thank James Hixson for helping me to better explain many aspects of my research and for being my professor in *Molecular and Cellular Approaches to Human Genetics*, which has been a very important foundational course in my graduate studies. I also thank Yin Liu for being a research advisor to me early in my graduate studies, for being my professor in *Statistical Methods in Bioinformatics*, which has been a foundational course for many aspects of my dissertation research, and for her bioinformatics advice, encouragement and friendship.

Finally, I thank the Schissler Foundation and the UT GSBS for providing fellowships that have funded me for a significant portion of the research presented in this dissertation.

GENOMIC CHARACTERIZATION OF POLYPS IN FAMILIAL ADENOMATOUS POLYPOSIS PATIENTS AND IDENTIFICATION OF CANDIDATE CHEMOPREVENTIVE DRUGS

A

DISSERTATION

Presented to the Faculty of

The University of Texas

Health Science Center at Houston

and

The University of Texas

MD Anderson Cancer Center

Graduate School of Biomedical Sciences

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

by

Francis Anthony San Lucas, M.S.

Houston, Texas

August, 2014

GENOMIC CHARACTERIZATION OF POLYPS IN FAMILIAL ADENOMATOUS POLYPOSIS PATIENTS AND IDENTIFICATION OF CANDIDATE CHEMOPREVENTIVE DRUGS

Francis Anthony San Lucas, M.S.

Advisory Professor: Paul Scheet, Ph.D.

Secondary Advisor: Eduardo Vilar, M.D., Ph.D.

Familial adenomatous polyposis (FAP) is an autosomal dominant disease characterized by *APC* germline mutations and the development of hundreds to thousands of premalignant adenomas in the gastrointestinal tract at a young age. If left untreated, these patients inevitably develop *colon cancer* (CRC) and small bowel tumors. We performed exome sequencing of samples from 12 FAP patients to characterize adenomas and to identify candidate genes of adenoma development that may serve as potential targets for chemoprevention drug development. From each patient, a blood and at least one polyp were sequenced with a total of 25 polyps analyzed. In some cases, normal mucosa samples were also sequenced. We characterized point mutations, insertions, deletions and chromosomal allelic imbalance. In addition, we performed RNA sequencing of 8 polyps and 4 normal mucosa samples from the colon and small bowel of 2 additional FAP patients.

Somatic *APC* truncating mutations and loss of chromosome 5q were recurrent across polyps, although we found no recurrent intra-patient somatic *APC* point mutations, indicating intra-patient polyp heterogeneity. Oncogenic driver events such as activating *KRAS* mutations were identified in multiple polyps. Further, analysis of mutation allele fractions suggests that several of the polyps studied are multi-clonal in nature. Excluding the known genes *APC* and *KRAS*, 50 candidate genes were identified that are putatively involved in the early development of CRC. These genes could play a role in future chemoprevention strategies. Most of these genes have been previously associated with CRC. In addition, a gene fusion in *PTEN* was detected and a novel, recurrent *REG3A* fusion was identified in duodenum polyps. The WNT signaling pathway, aberrant in 92% of CRCs, was recurrently altered in 80% of polyps.

We identified colon and duodenum gene expression signatures of FAP patients and screened them against drug-induced signatures using our *Cancer in-silico Drug Discovery* (CiDD) software. CiDD identified Celecoxib, a COX-2 inhibitor that has already been clinically tested as a chemopreventive drug, providing validity to our drug development approach. CiDD also identified a novel candidate compound, TTNPB, which targets the Retinoid pathway as a potential drug for chemopreventive treatment of FAP patients.

Table of Contents

List of Figures

List of Tables

Abbreviations

AI	allelic imbalance
CCLE	Cancer Cell Line Encyclopedia
CiDD	Cancer in silico Drug Discovery
CMap	Connectivity Map
CNA	copy number alteration
COSMIC	Catalogue of Somatic Mutations In Cancer
CRC	colorectal cancer
FAP	Familial Adenomatous Polyposis
GATK	Genome Analysis Toolkit
indel	insertion or deletion
LOH	loss of heterozygosity
MSigDB	Molecular Signatures Database
NGS	next-generation sequencing
NSAID	non-steroidal anti-inflammatory drug
PETACC3	Pan-European Trials in Alimentary Tract Cancers Clinical Trial
SNP	single nucleotide polymorphism
TCGA	The Cancer Genome Atlas
TSP	top-scoring pair
vtools	Variant Tools

1 Introduction

1.1 Background

1.1.1 Genetic basis and clinical description of *Familial Adenomatous Polyposis* (FAP)

Colorectal cancer (CRC) is the second leading cause of mortality in the United States[1] and fourth worldwide[2]. Nearly half of the general population will develop at least one benign colonic polyp in their lifetime with approximately 3% going on to develop CRC[3]. Symptoms are rare until late stages, thus most sporadic CRC cases go undetected. There are two primary forms of hereditary CRC: *Familial Adenomatous Polyposis* (FAP) and *Lynch Syndrome* (or hereditary non-polyposis colorectal cancer, HNPCC). These are both autosomal dominant diseases where patients are predisposed to cancers due to germline mutations in key genes. FAP is characterized by mutations in the *Adenomatous Polyposis Coli* (*APC*) gene, whereas Lynch Syndrome patients have germline mutations in genes involved in the DNA mismatch repair pathway including *MLH1*, *MSH2*, *TACSTD1*, *MSH6* and *PMS2*[4]. As such, FAP patients are born with only one normal allele of the *APC* tumor suppressor gene, predisposing them to the development of adenomas at younger ages compared to the general population. Similarly, Lynch Syndrome patients are born with a defect that predisposes them to higher mutation rates compared to normal individuals, giving them increased probabilities of obtaining key genomic aberrations in genes crucial for the development of carcinomas. Both of these patient populations are more likely to develop tumors compared to the general population. The focus of the research described in this dissertation will be on FAP.

CRCs are thought to progress from a normal epithelium to an adenoma, or pre-malignant lesion, and then to a carcinoma (the so-called *adenoma-to-carcinoma sequence* model[5]). This sequence describes steps of gene and pathway alterations that give abnormal cells a selective advantage to proliferate (see Figure 1). The initial step in tumorigenesis in the majority of adenomas, which are premalignant lesions, is the loss of the *APC* gene. Intermediate adenomas have activating mutations in *KRAS*, late adenomas are characterized by loss of *SMAD4*, and carcinomas have acquired *TP53* mutations among other alterations.

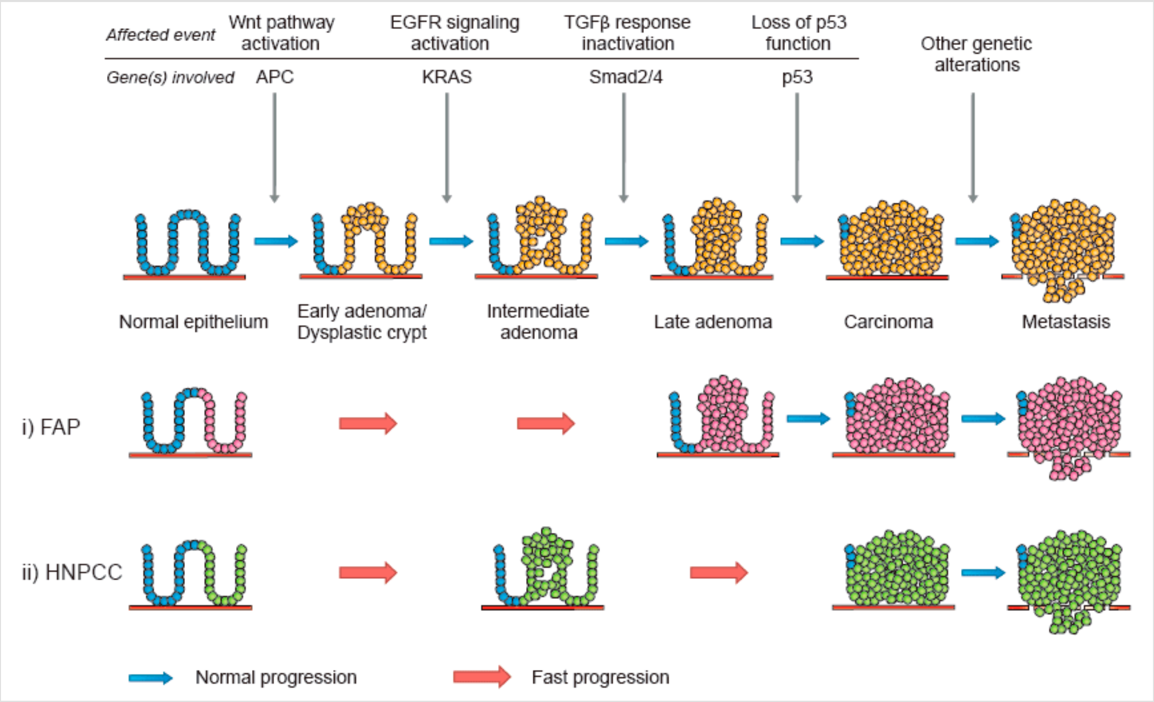


Figure 1: Step-wise progression of sporadic and inherited (FAP and HNPCC) forms of CRC from normal epithelia to adenomas and carcinomas (http://syscol-project.eu/about-syscol/).

The *APC* germline mutations of FAP patients accelerate the initiation of the adenoma-to-carcinoma sequence, resulting in the development of hundreds to thousands of polyps, generally in the colon and rectum and with lower densities in the small intestine such as the duodenum, ileum and jejunum[4] (see Figure 2). The development of these polyps in large numbers greater than 100 is termed *polyposis*. If left untreated, some polyps will inevitably progress into cancer in the lower gastrointestinal tract and less frequently in the upper gastrointestinal tract[6]. FAP accounts for less than 1 percent of all CRC cases and affects approximately 1 in 10,000 people[7]. Patients with FAP develop CRC at an average age of 35 years if left untreated, although there is variability within and between families, some of which can be explained by specific germline mutations in *APC*[4].

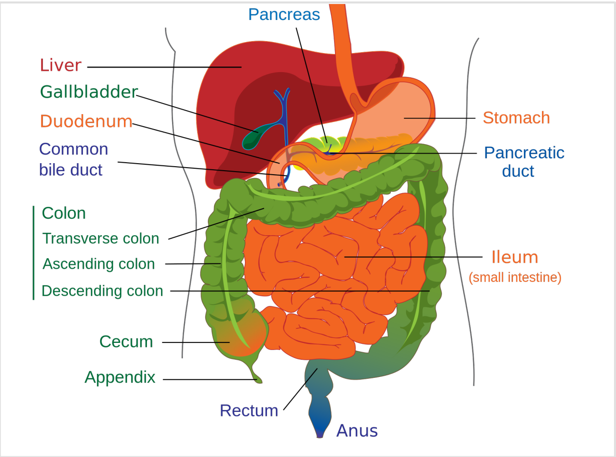


Figure 2: Overview of the gastrointestinal tract (http://en.wikipedia.org/wiki/Human_gastrointestinal_tract). In FAP patients, polyps develop and proliferate both in the lower (e.g., the colon and rectum) and the upper gastrointestinal tract (e.g., the duodenum and ileum). Malignancy rates of lower gastrointestinal tract polyps are higher than those in the upper tract. Therefore, resection of the colon and rectum are common prophylactic surgical procedures in FAP patients.

The primary cause of death for FAP patients has historically been CRC, which generally develops by the third or fourth decade of life[8]. However, the current standard of treatment, which includes regular surveillance through colonoscopy until the polyp burden becomes unmanageable, often by 20 years of age, at which point prophylactic surgical resection of the colon with colectomy or proctocolectomy, which includes removal of the rectum, is performed for cancer prevention purposes and has reduced the mortality by CRC greatly in the last 2 to 3 decades[6]. While surgery significantly improves the overall survival of FAP patients, the quality-of-life of these patients is reduced, apart from negative psychological aspects that are also associated with these surgeries. For example, reports have associated infertility with these surgeries[9]. In addition, patients with colectomies have a 25% risk of developing cancer in the preserved rectum [6]. These patients require surveillance of the rectum throughout the remainder of their lives.

The second leading cause of death after CRC in FAP patients stems from desmoid tumors and duodenal adenocarcinoma[10]. Studies have shown that duodenal adenomas have a prevalence of around 65% in FAP patients at a median age of 38 years and the lifetime risk for these patients of developing these lesions approaches 100%[11]. Although chemopreventive strategies have shown some effectiveness for inducing regression of colorectal polyps, their value in delaying or preventing duodenal polyps has been disappointing[8].

The inevitable risk of cancer in the colorectum and potentially the duodenum, and the young age at which FAP patients must undergo life-altering preventive surgery make chemoprevention in both the large and small intestine an urgent need. **Thus, it has been our**

goal to molecularly profile the differences between the at-risk normal mucosa and polyps of FAP patients to gain biological insights into the development of pre-malignant lesions, which may guide the development of chemopreventive strategies to delay or prevent the development of polyps and cancer in the colorectum and duodenum.

1.1.2 Role of the *Adenomatous Polyposis Coli* (*APC*) gene in adenoma formation

APC is a gene with 15 exons located in the long arm of chromosome 5 in band q22.2 that encodes a 2843 amino acid protein. The vast majority of *APC* mutations result in a truncated protein, where approximately one third of germline mutations in *APC* lie between codons 1061 and 1309[12]. Whereas the germline mutations are scattered throughout the 5' half of the *APC* gene, the majority of somatic mutations in both FAP and sporadic forms of CRCs are clustered between codons 1286 and 1513, in the so-called *mutation cluster region* (MCR)[13].

The *APC* gene is a tumor suppressor and promotes the degradation of beta-catenin. The regulation of beta-catenin by *APC* is accomplished by portions of the gene sequence including three 15-amino-acid repeats and seven 20-amino-acid repeats that respectively bind and downregulate beta-catenin via ubiquitination[14]. Loss of *APC* results in a constitutive activation of beta-catenin, such that beta-catenin will be translocated to the nucleus and will activate the transcription of many WNT target genes, which leads to cellular proliferation among activation of other cellular processes[15]. More than 90% of sporadic CRC patients have functional *APC* alterations[14]. Normally, for *APC* driven tumorigenesis, bi-allelic alteration initiates disease development. FAP patients have inherited only one normal functioning *APC* allele, which increases their probability and rate of disease development.

There are several FAP phenotypes that correlate with specific *APC* germline mutations. Mutations between codon 1250 and codon 1464 are associated with profuse polyposis where patients have greater than 5,000 colorectal polyps[16]. Mutations in codon 1309 are associated with early onset adenoma development (10 years earlier) and earlier CRC at ages less than 35 years of age[16]. Mutations at the 5′ and 3′ ends of the *APC* gene are associated with attenuated FAP, where patients are characterized by oligopolyposis, presenting with less than 100 colorectal polyps, with later onset of CRC at greater than 50 years of age[12].

1.1.3 Current chemopreventive strategies

Chemoprevention delays or prevents the development of cancer through the use of natural or pharmaceutical agents[17]. FAP patients are ideal for assessing the efficacy of chemopreventive agents for adenomatous polyps because FAP patients predictably develop polyps that are visible and countable prior to their transformation to cancer. Polyp counts in these patients provide a convenient measure of the effectiveness of chemopreventive agents. The potential benefits of chemoprevention in FAP consist mainly of the prevention of adenomas and the delay of tumor growth, thus giving FAP patients a longer and higher quality-of-life.

In the early 1980′s, epidemiological studies found that treatment with aspirin, a *non-steroidal anti-inflammatory drug* (NSAID), was associated with a reduced risk for CRCs[18]. Since then, NSAIDs have been extensively tested for chemoprevention of CRC in patients with hereditary predispositions, as well as the general population. NSAIDs inhibit COX, a key enzyme in the conversion of arachidonic acid to prostaglandins and other eicosanoids. There are 2 isoforms of the COX enzyme, COX-1 and COX-2. COX-1 is constitutively expressed in virtually all tissues, whereas COX-2 is absent under physiologic conditions[19]. COX-2 is induced in several clinical contexts such as inflammation and cancer. Overexpression of COX-2 has been observed in colorectal polyps and carcinomas[19,20]. The effectiveness of NSAIDs in repressing the growth of adenomas appears to be via inhibition of COX-2, although it has been suggested that NSAIDs may be effective independent of COX-2 suppression[21].

Thus, NSAIDs such as Sulindac, Celecoxib, Rofecoxib and others have been developed as chemopreventive strategies. In initial studies of the NSAID Sulindac, a substantial activity delaying the growth of polyps was observed[18]. Later, Celecoxib, gained U.S. Food and Drug Administration (FDA) approval for chemoprevention of polyps in patients with FAP[22]. However, significant cardiovascular toxic effects were observed during clinical trials in sporadic CRC populations and safety concerns led to withdrawal of the drug[23,24]. Aspirin has recently been proposed for standard chemopreventive treatment of CRC in individuals predisposed to CRC in addition to the general population. Two Phase III clinical trials evaluated aspirin as a chemopreventive agent in patients with FAP and Lynch Syndrome – the CAPP1 and CAPP2 studies, respectively. These studies showed that aspirin reduced the number of colon and rectum polyps in Lynch Syndrome patients and to a lesser extent in FAP patients. Currently, the chemopreventive agent of choice is aspirin in Lynch Syndrome and Sulindac in FAP[25,26]. However, there are currently no FDA approved chemopreventive agents available, illustrating the need for further development of chemopreventive strategies[6]. Further, the value of these agents for the prevention of polyps in the small intestine is unclear with yet no studies showing statistically significant regression of duodenal polyps in FAP patients[8].

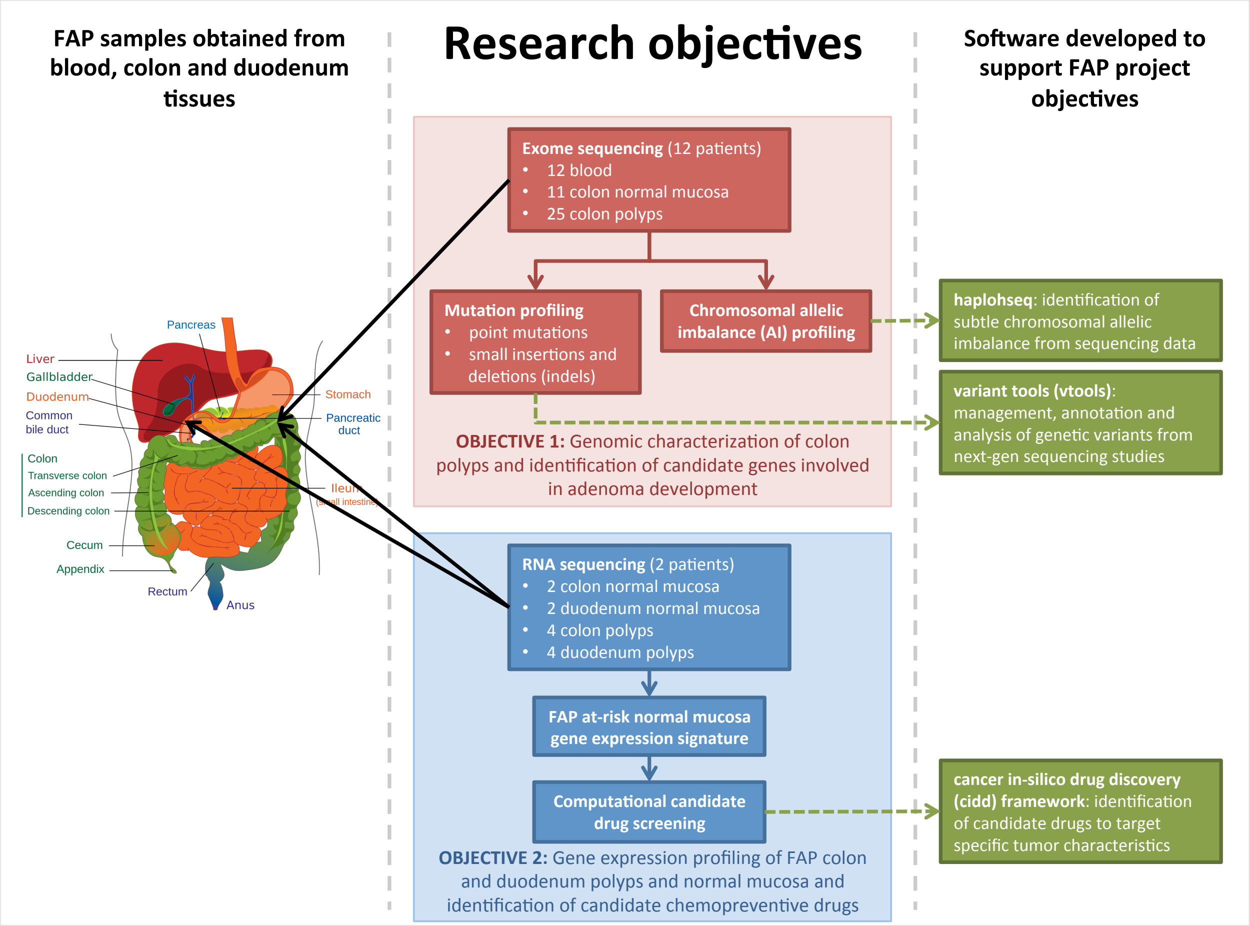
1.2 Objectives

Our long-term goal is to develop more effective chemopreventive therapies for the colon and duodenum of FAP patients. A comprehensive annotation of the genomic landscape of adenomatous polyps in FAP patients has not been previously attempted through *next-generation sequencing* (NGS) studies and is critical for enabling future targeted chemopreventive drug identification and development. In addition, aside from a single study where the normal mucosa, an adenoma and an adenocarcinoma were exome sequenced from one sporadic CRC case[27], adenomas in both the sporadic and hereditary contexts have been largely ignored with regards to NGS studies. Thus, the objective is first to molecularly characterize FAP polyps at a high-resolution by characterizing the exome and transcriptome of these lesions. As initial steps towards this goal, we are sequencing and characterizing the exomes of colon polyps and the transcriptomes of colon and duodenum polyps along with paired normal mucosa samples.

With advances in NGS technologies, the whole genomes and exomes of colorectal adenocarcinomas have been sequenced and comprehensive landscapes of genetic alterations and gene expression alterations have been characterized by *The Cancer Genome Atlas* (TCGA) consortium[28]. We will leverage these data and compare our FAP polyp genomic landscape with those of TCGA CRC tumors in an effort to identify a more comprehensive set of potential therapeutic targets that may be involved in the early development of CRC. By comparing FAP polyp somatic events to those of CRC and identifying similarities between the two, we will propose events that occur early in the development of CRCs. Conversely, by identifying CRC events that are absent in our FAP polyp data set, we can propose events that are involved in the transformation of colorectal adenomas to carcinomas.

High-throughput, high-resolution profiling through exome and RNA-sequencing of FAP adenomas may provide insights into the molecular mechanisms underlying the early development of CRC and may reveal genes or pathways that may be targeted by chemoprevention to delay or halt the tumorigenesis process in early stages. As such, our objectives are the following:

1. **To characterize the genomic landscape of FAP colon polyps** through exome sequencing analyses to compare identified recurrent somatic events to those previously associated with CRC to identify genes involved in early CRC development, which may help guide future chemoprevention strategies (see red boxes in Figure 3).
2. **To identify candidate chemopreventive drugs** to target the gene expression signatures of the at-risk normal mucosa of both the colon and duodenum in FAP patients, where the gene expression signatures are inferred from RNA sequence of colon and duodenum polyps and normal mucosa (see blue boxes in Figure 3). Ideally, chemoprevention strategies would include effective drugs or compounds that have minimal toxicity and which are inexpensive.


FAP samples obtained from blood, colon and duodenum tissues

Research objectives

Software developed to support FAP project objectives

Exome sequencing (12 patients)
- 12 blood
- 11 colon normal mucosa
- 25 colon polyps

Mutation profiling
- point mutations
- small insertions and deletions (indels)

Chromosomal allelic imbalance (AI) profiling

OBJECTIVE 1: Genomic characterization of colon polyps and identification of candidate genes involved in adenoma development

RNA sequencing (2 patients)
- 2 colon normal mucosa
- 2 duodenum normal mucosa
- 4 colon polyps
- 4 duodenum polyps

FAP at-risk normal mucosa gene expression signature

Computational candidate drug screening

OBJECTIVE 2: Gene expression profiling of FAP colon and duodenum polyps and normal mucosa and identification of candidate chemopreventive drugs

haplohseq: identification of subtle chromosomal allelic imbalance from sequencing data

variant tools (vtools): management, annotation and analysis of genetic variants from next-gen sequencing studies

cancer in-silico drug discovery (cidd) framework: identification of candidate drugs to target specific tumor characteristics

Figure 3: Overview of samples, objectives and software developed for the genomic characterization of colon polyps and the identification of candidate chemopreventive drugs for FAP patients.

In summary, effective chemopreventive strategies may reduce disease incidence, delay progression or lessen the severity of disease and disease-related secondary effects in FAP patients. Alternative benefits of chemopreventive strategies would be to postpone the need for prophylactic surgery providing patients a longer and better quality-of-life.

The research design, the analyses and biological interpretation of the FAP colon polyp genomes and the colon and duodenum polyp transcriptomes, and identification of candidate chemopreventive drugs is presented in this dissertation. In chapter 2, I describe our experimental design and patient samples, then present colon polyp somatic profiles and compare them to those of CRC from TCGA. In chapter 3, I present gene expression signatures representative of the difference between the at-risk normal mucosa and polyps in both the colon and duodenum of FAP patients, introduce a computational screening framework for candidate drug identification by illustrating its application to FAP, and describe the setup for follow-up drug testing. In chapter 4, I discuss the significance of our FAP research findings, describe future directions and discuss the potential impact of the bioinformatics tools that we developed over the course of this project. "Appendix A: Sequencing analysis pipelines" describes the pipelines created for the analysis of our exome and RNA sequencing data. In addition, the following new bioinformatics methods were developed: (1) *variant tools* for more simple annotation and analysis of identified NGS genetic variants, (2) *hapLOHseq* for the sensitive detection of chromosomal allelic imbalance events from exome sequencing data, and (3) the *Cancer in silico Drug Discovery* framework (CiDD) for the identification of candidate chemopreventive drugs. These three pieces of software (see green boxes in Figure 3) are described in detail in three subsections of "Appendix B: Bioinformatics software".

2 Genomic characterization of FAP polyps

In the *adenoma-to-carcinoma sequence*, it has been described that WNT pathway activation, which normally arises through bi-allelic loss of the *APC* gene, is the initiating event in the development of adenomas, which are pre-malignant lesions[5]. Further, *KRAS* mutations are often present when an adenoma transitions to later stages (e.g., the development of high-grade dysplasia), and subsequently, alterations in *PIK3CA* and *TP53* or other genes occur during the progression into an invasive adenocarcinoma, or cancer. We sought to characterize the genome of adenomas through exome sequencing to further refine the somatic alterations and genes that might be involved in adenoma development to identify candidate targets for chemoprevention. In this chapter, I describe the mutation and chromosomal allelic imbalance profiles of FAP polyps and compare them to profiles of CRC from TCGA. A special emphasis is placed on identifying somatic *APC* events and alterations in WNT signaling since these events are expected to be key initiating events of adenoma formation. I conclude by proposing a list of candidate genes that may contribute to the initiation or development of adenomas.

2.1 Methods

2.1.1 Available patients and samples

To characterize somatic alterations in polyps of FAP patients, we conducted a genome-wide analysis of polyps from 12 patients. Samples for 4 patients were collected at the Catalan Institute of Oncology. Colon polyp and normal mucosa samples from these patients were extracted after prophylactic surgical resection of the colon. Colon polyp and normal mucosa samples from 8 additional patients were collected at MD Anderson Cancer Center through endoscopic excision. Germline DNA was extracted from peripheral blood lymphocytes using the *Blood & Cell Culture DNA Mini Kit* (Qiagen). Polyp and normal mucosa DNA was extracted using the *QIAmp DNA Mini Kit* (Qiagen). A blood sample and one or more polyp samples were collected from each patient with a total of 25 polyps analyzed. Additionally, for 11 of the 12 patients, a normal mucosa sample was obtained.

Person	Tissue Type	Sample Name	APC Germline Mutation (cDNA)	APC Germline Mutation (protein)
CATA01	Blood	CATA01_B01_Vilar01	c.3927_3931delAAAGA	p.Glu1309Aspfs*4
	Normal Mucosa	CATA01_N01_Vilar02		
	Polyp	CATA01_P01_Vilar13		
	Polyp	CATA01_P02_Vilar16		
	Polyp	CATA01_P03_Vilar17		
	Polyp	CATA01_P04_Vilar18		
CATA02	Blood	CATA02_B01_Vilar04	c.4393_4394delAG	p.Ser1465Trpfs*3
	Normal Mucosa	CATA02_N01_Vilar05		
	Polyp	CATA02_P01_Vilar14		
	Polyp	CATA02_P02_Vilar19		
	Polyp	CATA02_P03_Vilar21		
CATA03	Blood	CATA03_B01_Vilar07	c. [1958+3G>A(;)c.1959G>A]	---
	Normal Mucosa	CATA03_N01_Vilar08		
	Polyp	CATA03_P01_Vilar15		
CATA04	Blood	CATA04_B01_Vilar10	c.1412delG	p.Gly471Aspfs*27
	Normal Mucosa	CATA04_N01_Vilar11		
	Polyp	CATA04_P01_Vilar12		
	Polyp	CATA04_P02_Vilar22		
	Polyp	CATA04_P03_Vilar23		
	Polyp	CATA04_P04_Vilar24		
MDAC01	Blood	MDAC01_B01_Vilar44	c.1880dupA	p.Ala630*
	Normal	MDAC01_N01_Vilar43		
	Polyp	MDAC01_P01_Vilar41		
	Polyp	MDAC01_P02_Vilar42		
MDAC02	Blood	MDAC02_B01_Vilar46	c.3810T>A	p.Cys1270*
	Polyp	MDAC02_P01_Vilar45		
MDAC08	Blood	MDAC08_B01_Vilar50	c.622C>T	p.Gln208*
	Normal	MDAC08_N01_Vilar49		
	Polyp	MDAC08_P01_Vilar47		
	Polyp	MDAC08_P02_Vilar48		
MDAC10	Blood	MDAC10_B01_Vilar53	c.3440dupA	p.Ser1148Thrfs*18
	Normal	MDAC10_N01_Vilar52		
	Polyp	MDAC10_P01_Vilar51		
MDAC14	Blood	MDAC14_B01_Vilar58	del 8-9	---
	Normal	MDAC14_N01_Vilar57		
	Polyp	MDAC14_P01_Vilar54		
	Polyp	MDAC14_P02_Vilar55		
	Polyp	MDAC14_P03_Vilar56		
MDAC17	Blood	MDAC17_B01_Vilar62	c.1658G>A	p.Trp553*
	Normal	MDAC17_N01_Vilar61		
	Polyp	MDAC17_P01_Vilar59		
	Polyp	MDAC17_P02_Vilar60		
MDAC18	Blood	MDAC18_B01_Vilar65	c.4393_4394delAG	p.Ser1465Trpfs*3
	Normal	MDAC18_N01_Vilar64		
	Polyp	MDAC18_P01_Vilar69		
MDAC20	Blood	MDAC20_B01_Vilar68	c.477C>G	p.Tyr159*
	Normal	MDAC20_N01_Vilar67		
	Polyp	MDAC20_P01_Vilar66		

Table 1: FAP patients, samples collected for exome sequencing and their *APC* germline mutations.

2.1.2 Data collection

Exome DNA was captured using the *SeqCap EZ Human Exome library v3.0* capture chip from 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 a mean depth of 80 reads (or "80x") at the MD Anderson Cancer Center sequencing core facility. Reads were aligned with the *Burrows-Wheeler Alignment* software (BWA)[29] to the 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)[30] and by applying recommended best practices for sequence analysis from the Broad Institute. For a complete description of the sequence alignment pipeline, see section 5.1 in Appendix A. After filtering for only those reads that map to the target exome region, the mean on-target depth was 64x (see Table 2).

Sample Name	Target Mean Depth	%@10X	%@20X	%@50X	%@100X
CATA01_B01_Vilar01	85.23	0.959	0.923	0.892	0.721
CATA01_N01_Vilar02	85.14	0.959	0.921	0.889	0.714
CATA01_P01_Vilar13	90.7	0.970	0.940	0.916	0.764
CATA01_P02_Vilar16	70.29	0.969	0.930	0.878	0.577
CATA01_P03_Vilar17	71.83	0.970	0.932	0.879	0.582
CATA01_P04_Vilar18	72.39	0.971	0.933	0.884	0.595
CATA02_B01_Vilar04	90.95	0.961	0.927	0.899	0.745
CATA02_N01_Vilar05	84.52	0.962	0.928	0.900	0.723
CATA02_P01_Vilar14	71.19	0.963	0.920	0.880	0.621
CATA02_P02_Vilar19	78.75	0.964	0.929	0.896	0.680
CATA02_P03_Vilar21	74.87	0.964	0.928	0.890	0.642
CATA03_B01_Vilar07	77.54	0.967	0.923	0.870	0.604
CATA03_N01_Vilar08	69.46	0.955	0.901	0.834	0.533
CATA03_P01_Vilar15	63.29	0.960	0.908	0.851	0.535
CATA04_B01_Vilar10	74.76	0.941	0.856	0.788	0.550
CATA04_N01_Vilar11	71.14	0.950	0.888	0.819	0.539
CATA04_P01_Vilar12	81.57	0.975	0.942	0.905	0.666
CATA04_P02_Vilar22	79.26	0.968	0.935	0.896	0.653
CATA04_P03_Vilar23	74.1	0.968	0.932	0.886	0.606
CATA04_P04_Vilar24	66.22	0.967	0.925	0.859	0.528
MDAC01_B01_Vilar44	45.65	0.978	0.921	0.812	0.294
MDAC01_N01_Vilar43	43.32	0.979	0.910	0.771	0.270
MDAC01_P01_Vilar41	45.08	0.985	0.923	0.801	0.289
MDAC01_P02_Vilar42	49.44	0.984	0.923	0.815	0.346
MDAC02_B01_Vilar46	62.46	0.983	0.941	0.881	0.507
MDAC02_P01_Vilar45	58.75	0.981	0.935	0.863	0.464
MDAC08_B01_Vilar50	71.72	0.987	0.949	0.905	0.600
MDAC08_N01_Vilar49	28.69	0.982	0.841	0.536	0.116
MDAC08_P01_Vilar47	29.4	0.979	0.862	0.574	0.109
MDAC08_P02_Vilar48	62.07	0.982	0.938	0.874	0.498
MDAC10_B01_Vilar53	63.4	0.982	0.942	0.889	0.527
MDAC10_N01_Vilar52	69.91	0.983	0.944	0.899	0.588
MDAC10_P01_Vilar51	65.94	0.983	0.942	0.890	0.548
MDAC14_B01_Vilar58	44.69	0.982	0.919	0.800	0.287
MDAC14_N01_Vilar57	42.59	0.984	0.919	0.790	0.256
MDAC14_P01_Vilar54	58.93	0.980	0.938	0.877	0.478
MDAC14_P02_Vilar55	62.26	0.984	0.942	0.885	0.514
MDAC14_P03_Vilar56	65.1	0.984	0.944	0.891	0.544
MDAC17_B01_Vilar62	61.37	0.985	0.940	0.876	0.496
MDAC17_N01_Vilar61	60.33	0.985	0.941	0.875	0.486
MDAC17_P01_Vilar59	45.32	0.985	0.921	0.801	0.294
MDAC17_P02_Vilar60	44.45	0.983	0.923	0.803	0.280
MDAC18_B01_Vilar65	66.48	0.987	0.946	0.891	0.549
MDAC18_N01_Vilar64	58.7	0.985	0.936	0.859	0.463
MDAC18_P01_Vilar69	57.13	0.980	0.910	0.793	0.426
MDAC20_B01_Vilar68	60.81	0.986	0.941	0.876	0.488
MDAC20_N01_Vilar67	60.89	0.987	0.945	0.885	0.493
MDAC20_P01_Vilar66	55.53	0.986	0.939	0.862	0.427

Table 2: Exome coverage summary for FAP samples. The on-target mean depth is 64x with an average of 92.5% of the target regions being covered by at least 20x.

Each individual and sample is characterized by a germline mutation in the *APC* gene as described in Table 1. The vast majority of FAP patients are characterized by nonsense mutations on the 5′ half of *APC* as were our samples. Germline mutations were verified in each sample through visualization of aligned sequence reads in the *Integrative Genomics Viewer* (IGV) software[31].

2.1.3 Strategies for calling somatic events

We performed exome sequencing to identify somatic events in our polyp samples including point mutations, insertions, deletions and *chromosomal allelic imbalance* (AI) events, which we define as large amplifications, deletions and *loss-of-heterozygosity* (LOH) event regions of greater than 10 MB. Data not analyzed include small *copy number variants* (CNV) less than 10 MB and structural variations such as translocations and inversions. These data are typically analyzed using SNP arrays and whole genome sequencing, respectively. With the exception of the verification of *APC* germline mutations in the blood and normal mucosa samples, the current focus of our project is on somatic variation. Thus, blood and normal samples were simply used as reference samples to allow for the characterization of polyp events as somatic (versus germline). Inspecting somatic variation in normal mucosa samples is the subject of future work and is out of scope of this dissertation. Figure 4 illustrates three general pipelines and the associated software that have been implemented for the genomic characterization of FAP polyps, which are the following: (1) alignment and quality control of sequencing reads (in black) as described previously, (2) calling and controlling of false positives for somatic point mutations and indels (in red), and (3) calling of chromosomal allelic imbalance events (in blue).

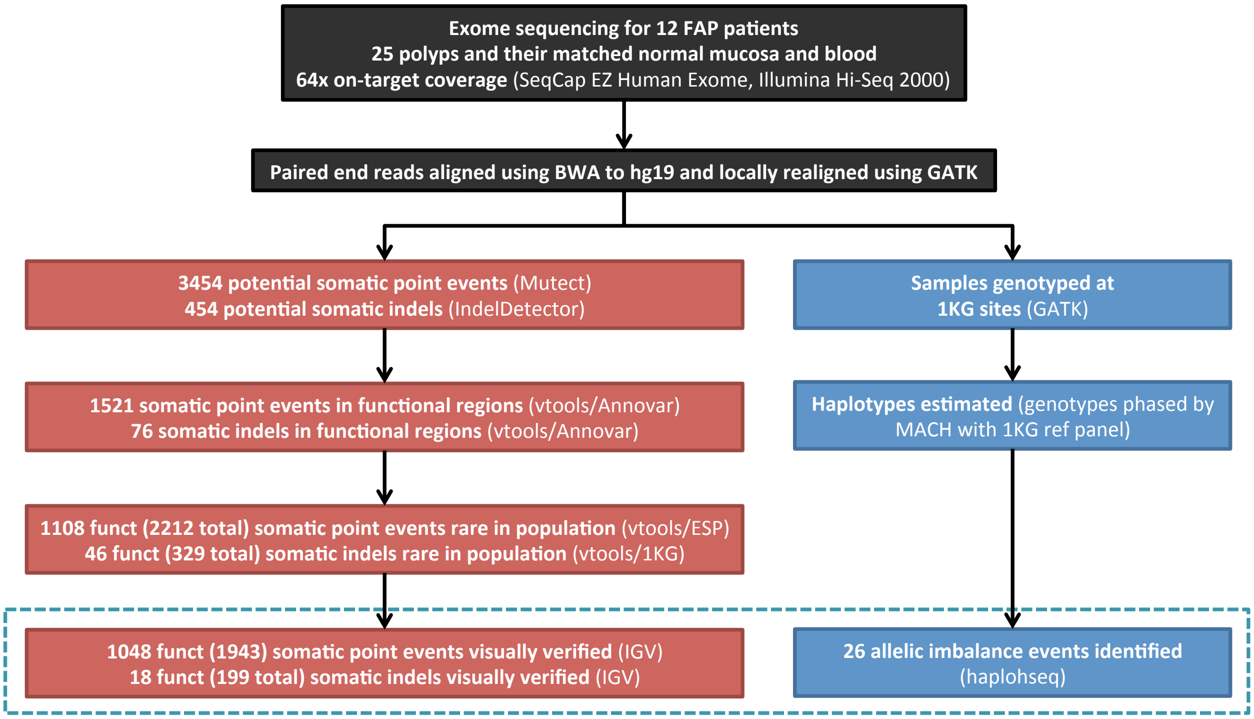

Exome sequencing for 12 FAP patients
25 polyps and their matched normal mucosa and blood
64x on-target coverage (SeqCap EZ Human Exome, Illumina Hi-Seq 2000)
Paired end reads aligned using BWA to hg19 and locally realigned using GATK
3454 potential somatic point events (Mutect)
454 potential somatic indels (IndelDetector)
Samples genotyped at
1KG sites (GATK)
1521 somatic point events in functional regions (vtools/Annovar)
76 somatic indels in functional regions (vtools/Annovar)
Haplotypes estimated (genotypes phased by
MACH with 1KG ref panel)
1108 funct (2212 total) somatic point events rare in population (vtools/ESP)
46 funct (329 total) somatic indels rare in population (vtools/1KG)
1048 funct (1943) somatic point events visually verified (IGV)
18 funct (199 total) somatic indels visually verified (IGV)
26 allelic imbalance events identified
(haplohseq)

Figure 4: Simplified pipelines for the alignment of sequencing reads (black), calling of somatic point mutations, and indels (red) and identification of chromosomal allelic imbalance events (blue) from exome sequencing data.

Mutect[32] was run for calling point mutations and Indelocator[30] was executed for calling small insertions and deletions (see Figure 4). Mutect and Indelocator were designed for calling mutations and indels in the context of low tumor purity and for identifying subclonal events making them suitable in the context of pre-cancerous lesions. To control for false positives, we annotated mutations with public databases and applied a filtering strategy to remove putative mutations that were likely to be common polymorphisms. Specifically, any somatic mutation that has been identified as a germline variant in a public sequencing project is likely to be a false positive mutation call. This "mutation" is likely a germline variant that was correctly identified in a polyp sample but failed detection in the paired blood sample.

To support this part of the event calling pipeline, we developed *variant tools* (vtools), a flexible annotation and analysis toolset that greatly simplifies the storage, annotation and filtering of variants and the analysis of the underlying samples[33]. Storing, annotating and analyzing variants from NGS projects can be difficult due to the availability of a wide array of data formats, tools and annotation sources, as well as the sheer size of the data. Useful tools, including the *GATK*[30], *ANNOVAR*[34] and *BEDTools*[35], can be integrated into custom pipelines for annotating and analyzing sequence variants. However, building flexible pipelines that support the tracking of variants alongside their samples, while enabling updated annotation and re-analyses is not a simple task. Using a command-line driven reporting structure, *variant tools* can be used to manage and analyze genetic variants obtained from sequence alignments, and the toolset could be used as a foundation for building more sophisticated analytical pipelines. The *variant tools* concept is illustrated in Figure 5 and its functions are described in more detail in Appendix B section 6.1.

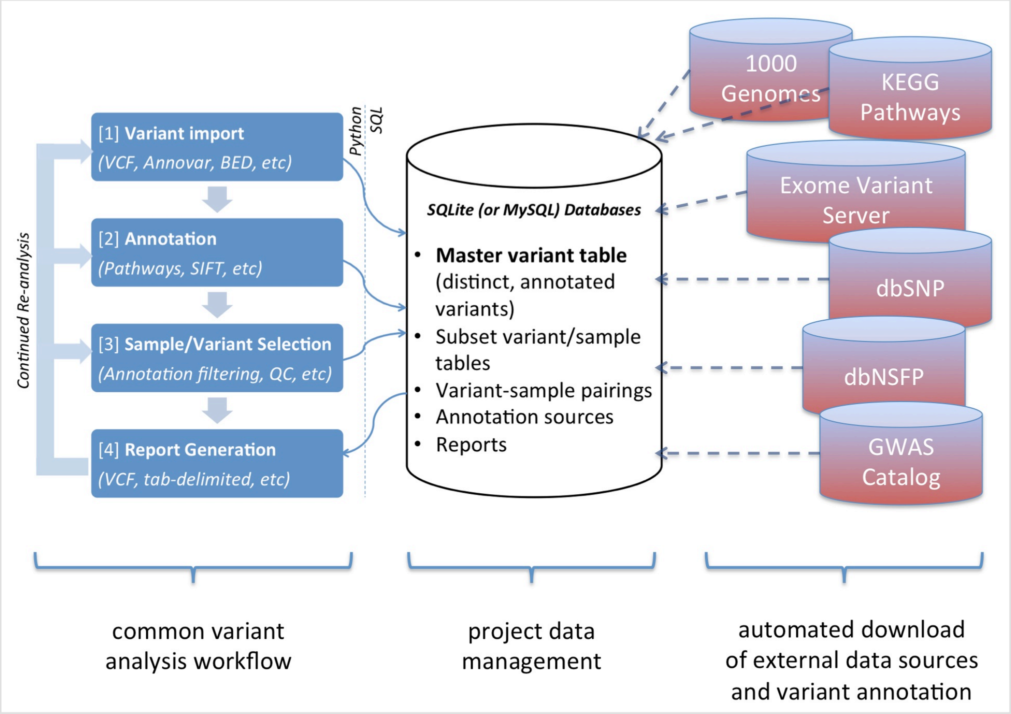

[1] Variant import
(VCF, Annovar, BED, etc)
[2] Annotation
(Pathways, SIFT, etc)
[3] Sample/Variant Selection
(Annotation filtering, QC, etc)
[4] Report Generation
(VCF, tab-delimited, etc)
Continued Re-analysis
Python
SQL
SQLite (or MySQL) Databases
• Master variant table (distinct, annotated variants)
• Subset variant/sample tables
• Variant-sample pairings
• Annotation sources
• Reports
1000 Genomes
KEGG Pathways
Exome Variant Server
dbSNP
dbNSFP
GWAS Catalog
common variant analysis workflow
project data management
automated download of external data sources and variant annotation

Figure 5: Overview of *variant tools*[33]. This software facilitates the management, annotation and analysis of genetic variants from NGS studies.

Somatic events called by Mutect and Indelocator were annotated through *variant tools* with population allele frequencies of the *1000 Genomes Project*[36] and the *Exome Sequencing Project* (version with 6,500 exomes) a for subsequent filtering of likely common polymorphisms and false positives. We excluded any candidate somatic mutations seen at 1% or greater population allele frequency in either of these projects.

In order to further limit false positive calls after filtering based on annotations, we developed a simple *sequencing read verification pipeline* to help control for systematic sequencing errors. The pipeline identifies potential errors by looking for evidence of variant reads in normal samples at all sites where somatic events were called. The verification pipeline implements the following workflow:

1. Create a master list of all the putative somatic mutations identified in all polyps.
2. Obtain a genotype call (from the *UnifiedGenotyper* of the GATK) and the *variant allele fraction* at all of the sites in the master list for all FAP samples including blood, normal mucosa and polyp samples. A variant allele fraction for a site is the number of reads harboring the variant allele divided by the total number of reads covering that site.
3. For each putative somatic mutation, mark it as "failed verification" if any of the following are true:
   a. a germline genotype call in any blood or normal mucosa sample in our project contains the variant allele (with the exception of *APC* germline mutations which may also be seen as a somatic mutation in a polyp),
   b. the variant allele fraction is 2% or greater in the paired blood sample, or
   c. the variant allele fraction is 5% or greater in the paired normal mucosa sample.

Finally, we visually verified point mutations and insertions and deletions using IGV. In total, through our filtering process we reduced the 3,454 original point mutation calls to 1,943 visually verified point mutations. For insertions and deletions, we reduced the 454 original calls to 199 visually verified calls.

A description of the pipeline for calling chromosomal *allelic imbalance* (AI) events is described in Appendix A, section 5.4. This pipeline prepares data and runs *hapLOHseq*, software that we developed, for identifying regions of chromosomal AI, which I describe in Appendix B, section 6.2.

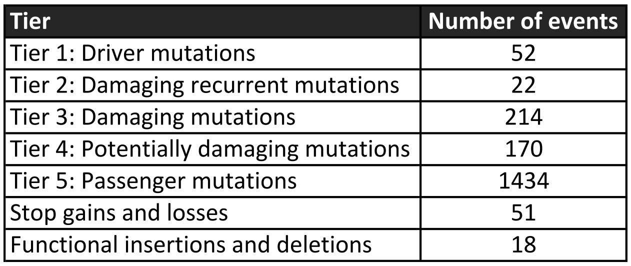
2.1.4 Prioritizing and validating mutations

For prioritization of events, mutations were annotated using *variant tools* with functional prediction statuses as determined by PolyPhen2[37], LRT[38], SIFT[39] and MutationTaster[40]. These predictions were pre-calculated by dbNSFP[41]. In addition, driver prediction statuses for each point mutation were obtained by running CHASM[42]. Finally, recurrence of events in existing cancer data sets was assessed by annotating mutations with the COSMIC database[43] through variant tools.

A strategy for categorizing mutations into functional tiers was then applied to all somatic events. First, indels and stop gain (or nonsense) and loss mutations were separated into their own tiers. Then the remaining mutations were prioritized using the following tier definitions where lower tier numbers correspond to higher priorities.

- **Tier 1:** ***Driver mutations*** – classified as a driver mutation (based on an empirical p-value ≤ 0.05 from CHASM) and seen in multiple (2 or more) other tumors in the COSMIC database
- **Tier 2:** ***Damaging recurrent mutations*** – predicted to be damaging by 2 or more algorithms and seen in multiple (2 or more) other tumors in the COSMIC database
- **Tier 3:** ***Damaging mutations*** **–** predicted to be damaging by 2 or more algorithms
- **Tier 4:** ***Potentially damaging mutations*** – predicted to be damaging by 1 algorithm
- **Tier 5:** ***Passenger mutations*** – remaining point mutations that are not stop gains or stop losses

- ***Stop gains and losses*** – any mutation that results in a nonsense mutation, such that the gene is prematurely truncated, or any mutation that alters an existing stop codon, such that the gene is elongated
- ***Functional insertions and deletions*** – any insertion or deletion in the coding region of a gene.



Tier	Number of events
Tier 1: Driver mutations	52
Tier 2: Damaging recurrent mutations	22
Tier 3: Damaging mutations	214
Tier 4: Potentially damaging mutations	170
Tier 5: Passenger mutations	1434
Stop gains and losses	51
Functional insertions and deletions	18

Table 3: Number of somatic events per tier definition

Mutations identified in *APC* or in known colorectal cancer genes or those predicted to be driver mutations in Tier 1 have been validated with Sanger sequencing in cases where we have enough DNA to perform the sequencing. Primers for amplification and sanger sequencing validation were designed by using a custom pipeline that incorporates *Primer3*[44] that targeted mutations with 50 bases of flanking DNA sequence on the 5′ and 3′ ends of the mutation. Sanger sequencing was performed on an ABI 3730 Capillary DNA Analyzer. Sequence trace files were manually inspected for the verification of point mutations, insertions and deletions. (Table for validated/failed mutations?)

2.1.5 Strategy for identifying candidate genes involved in early CRC development

Our strategy for identifying important candidate genes involved in the early development of CRC was to find recurrently mutated genes in our data set that have been previously identified as CRC genes. We identified CRC genes from 3 large-scale projects: TCGA colorectal project[28], Vogelstein *et al*[45] and Seshagiri *et al*[46]. We also interrogated pathways known to be deregulated in CRC: most importantly, the WNT, MAPK and ERBB signaling pathway. We identified the genes for these pathways using the *Molecular Signatures Database* (MSigDB)[47]. If a sample had a nonsynonymous mutation on any gene in a given pathway, that sample was labeled as having an alteration in that pathway.

2.2 FAP polyp genomic profiles

Here, I present a characterization of the genomic landscape of FAP adenomas based on the exome sequence data. I compare mutation rates and base substitution profiles between FAP polyps and CRC tumors. Somatic alterations in *APC* and other previously identified CRC genes are characterized. Additionally, recurrent chromosomal allelic imbalance events, which we define as amplifications, deletions and *copy-neutral loss-of-heterozygosity* (cn-LOH) events are identified and interpreted in the context of the adenoma-to-carcinoma sequence. In total, 52 genes that have previously been associated with CRC and that are recurrently altered in our adenomas are proposed as candidate genes involved in the early development of CRC. These genes will be followed up with functional studies in future projects.

2.2.1 Mutation profiling

To minimize batch effects in our comparisons between our FAP polyps and TCGA CRC samples, we downloaded 107 tumor/normal pairs of exomes from the TCGA project where the samples were sequenced on the Illumina Hiseq machine (the same sequencing technology used for our FAP polyps) and we ran these samples through the same bioinformatics pipelines on which we ran our FAP samples. Of the 107 CRC tumor exomes downloaded, 22.4% (24 of 107) of the tumors were stage I, 42.1% (45 of 107) were stage II, 21.5% (23 of 107) were stage III and 11.2% (12 of 107) were stage IV tumors. Three samples lacked tumor stage classifications. The same pipelines that were used with our FAP polyps to identify somatic point mutations and generate mutation reports were applied to the CRC exomes.

2.2.1.1 Mutation rates

We assessed the similarity of mutation rates and base substitution profiles from TCGA CRC samples and FAP polyps. In the TCGA CRC publication[28], the authors identified 2 classes of samples based on mutation rates: *hypermutators* (samples with greater than 10 mutations/Mb) and *nonhypermutators* (samples with less than 10 mutations/Mb). The samples that we downloaded from the TCGA were not included as part of the results in the TCGA CRC manuscript because they are newer samples, sequenced after the manuscript was published. We identified the mutation rates for the TCGA CRC samples and recapitulated the hypermutator and nonhypermutator findings of the TCGA in these newer samples and subsequently treated these two classes of samples separately (see Figure 7). Mutation rates are represented as the number of mutations per megabase for each sample.

$$mutation\ rate = \frac{mutation\ count}{callable\ bases} * (1{,}000{,}000) \tag{1}$$

For a *base*, or a specific nucleotide position in a polyp genome, to be *callable,* a minimum coverage of 12x in the polyp sample and a minimum coverage of 8x in the corresponding blood sample from the same patient is required. These minimum coverage requirements result in 80% power to identify mutations by the Mutect software[32]. The total number of callable bases is then the denominator in the mutation rate calculation.

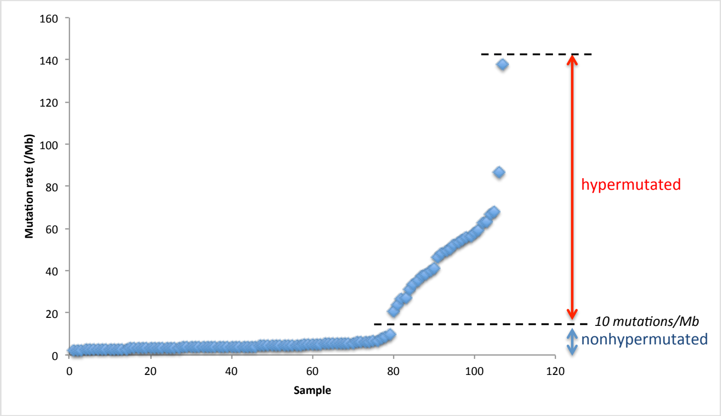


Figure 6: TCGA CRC tumors can be classified into one of two groups based on mutation rate: hypermutated for mutation rates greater than 10 mutations per megabase or nonhypermutated otherwise

Mutation rates of FAP polyps (mean mutation rate = 1.74 mutations/Mb) are lower than that of nonhypermutated CRCs (mean mutation rate = 4.26 mutations/Mb; T-Test p-value = 8.92e-13) as expected since adenomas are in an earlier stage of tumorigenesis compared to carcinomas. The mutation rates identified in the FAP polyps overlapped those of TCGA CRC nonhypermutated samples (see the mutation rate boxplot of Figure 7). The polyp mutation rate is an order of magnitude smaller than that of hypermutated CRCs (mean mutation rate = 50.88 mutations/Mb). We expect this because FAP polyps are not characterized by alterations in mismatch repair genes, which generally typify microsatellite instable and hypermutated carcinomas. A contributing factor of the mutation rate difference between polyps and nonhypermutated CRCs could be due to a lower power to detect mutations in polyps because of potentially higher normal mucosa contamination as compared to TCGA CRC samples.

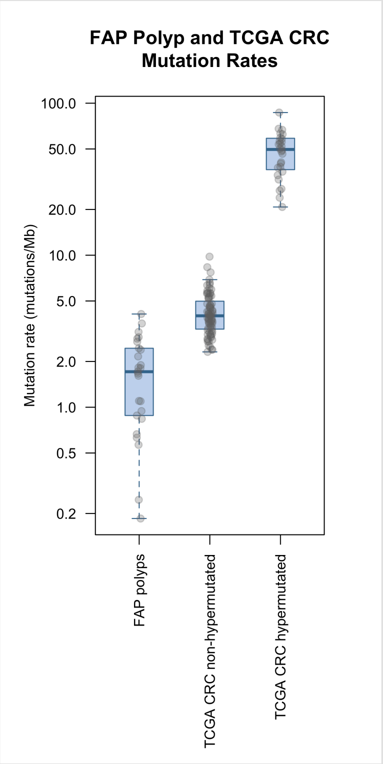


Figure 7: FAP polyp mutation rates compared to TCGA hypermutated and nonhypermutated CRCs

2.2.1.2 *Mutation base substitution signatures*

One strategy for identifying candidate mechanisms that drive tumor mutational processes is to evaluate mutation base substitution signatures. Figure 8 illustrates mutation base substitution signatures across several cancer types. In each signature plot at the bottom of Figure 8, vertical bars represent the base substitution frequencies for individual tumor samples of the corresponding cancer type. Based on these profiles, we can attempt to infer mutational processes that are the source of the base substitution signatures. For example, the high C->T substitutions in melanoma samples can be attributed to UV exposure. The high C->A substitutions in lung cancer samples are thought to be attributed to tobacco smoke exposure[48].

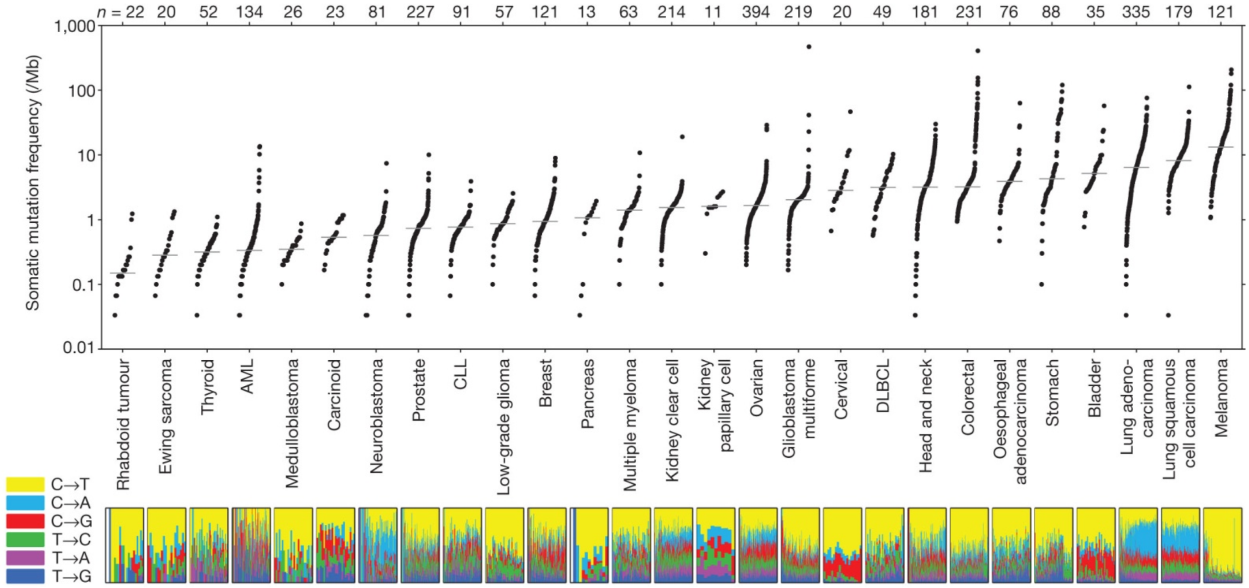


Figure 8: Mutational signatures across cancer types [49].

To refine the mutation signatures of Figure 8, Alexandrov *et al* identified base substitution profiles at a higher resolution by incorporating flanking bases for each base substitution[48]. For example, instead of identifying the numbers of C->T substitutions, they incorporated all possible combinations of flanking bases around the C and T (e.g., ACG -> ATG, ACC -> ATC, ….). After doing this, they associated their refined base substitution signatures with mutational mechanisms for several cancers. They associated CRC mutation signatures with 3 distinct mutational processes: (1) a strand-specific mutational process due to POLE mutations, (2) DNA mismatch repair deficiency and (3) aging mechanisms[48]. Using this flanking-base strategy, there are 96 possible base substitutions comprising their mutation signatures. Many of our samples have fewer than 96 mutations, thus we are not reliably able to perform a similar analysis on our polyp data. So we make a more qualitative comparison between sample types by comparing the base substitution profiles of Figure 9.

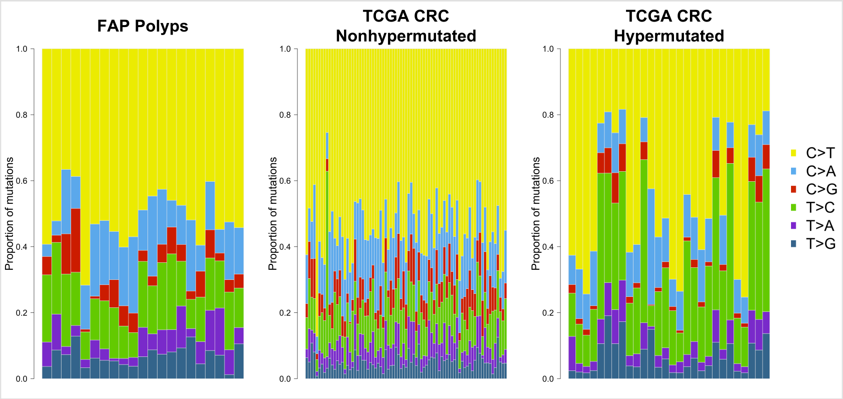


Figure 9: Base substitution profiles of FAP polyps versus TCGA hypermutated and nonhypermutated CRCs suggest that the FAP polyps have similar mutation processes underlying them compared to TCGA CRC nonhypermutated tumors

Based on base substitution profiles, the mutation signature of FAP polyps resemble that of nonhypermutated CRCs, suggesting that the mutational processes behind nonhypermutated CRCs are the same as those for FAP polyps. Hypermutators are largely microsatellite-high tumors with one or more mutations in DNA mismatch repair genes or mutations in POLE[28]. Neither of these genes was identified as mutated in analyses of our FAP polyps, leaving aging or an unidentified mutational process as the likely source of mutations in FAP adenomas.

2.2.1.3 *Variant allele fraction profiling*

Another strategy for mutation profiling is in the analysis of *variant allele fractions* (VAFs). Analysis of VAFs can provide insights into the purity and clonality of samples. The variant (or mutation) allele fraction, *f* for a particular somatic mutation *m* is:

$$f_m = \frac{number\ of\ variant\ reads_m}{total\ number\ of\ reads_m} \tag{2}$$

VAFs for mutations in chromosomal allelic imbalance regions (see section 2.2.2) were ignored in addition to mutations on chromosomes X and Y. In this way, only somatic mutations in copy neutral autosomal regions were included in the VAF profiling analyses. The resulting VAF distributions of the somatic mutations identified in 2 polyps are illustrated in Figure 10. In the *CATA01_P04_Vilar18* sample, the VAF distribution is shifted close to 0 with a mean VAF of around 0.08. All the VAF values *CATA01_P04_Vilar18* are smaller than 0.2 reflecting that these mutations all occur in a small proportion of the cells sequenced. Thus, we infer that this sample has low polyp purity. In contrast, the VAF distribution for *CATA04_P01_Vilar12* has two peaks, which represents VAFs for multiple clones. The founding clone has a mean VAF at around 0.35. To get a simple estimate of purity, we can multiply this VAF by 2 because these variants are all heterozygous genotypes, and we would expect to see the variant alleles in the founding clone in approximately one-half of these reads (assuming no copy number alterations). This would indicate that this sample has good polyp purity, with a purity estimate of 0.7 (purity = 0.35 * 2 = 0.7).

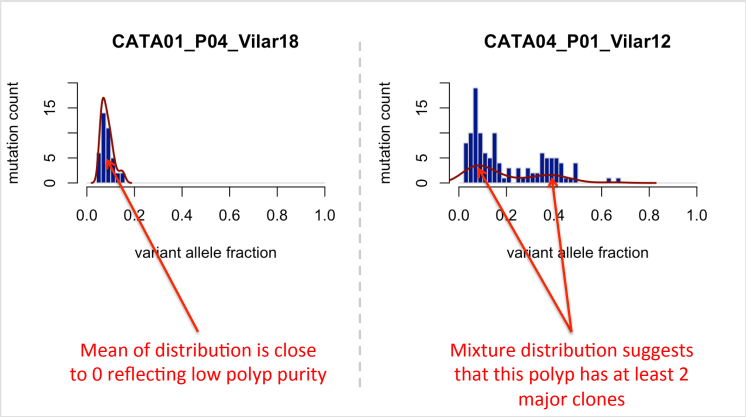


Figure 10: The distributions of variant allele fractions can be interpreted to identify samples of low polyp purity or to characterize samples exhibiting patterns of multi-clonality.

If we generate VAF distribution plots for all polyps and assess multi-clonality, we can easily identify 5 polyps that show evidence of being comprised of at least 2 major clones (see Figure 11). In addition, several polyps have VAF distributions tightly shifted near 0, indicating as we expected, that these polyps have low purity and are challenging to genomically characterize.

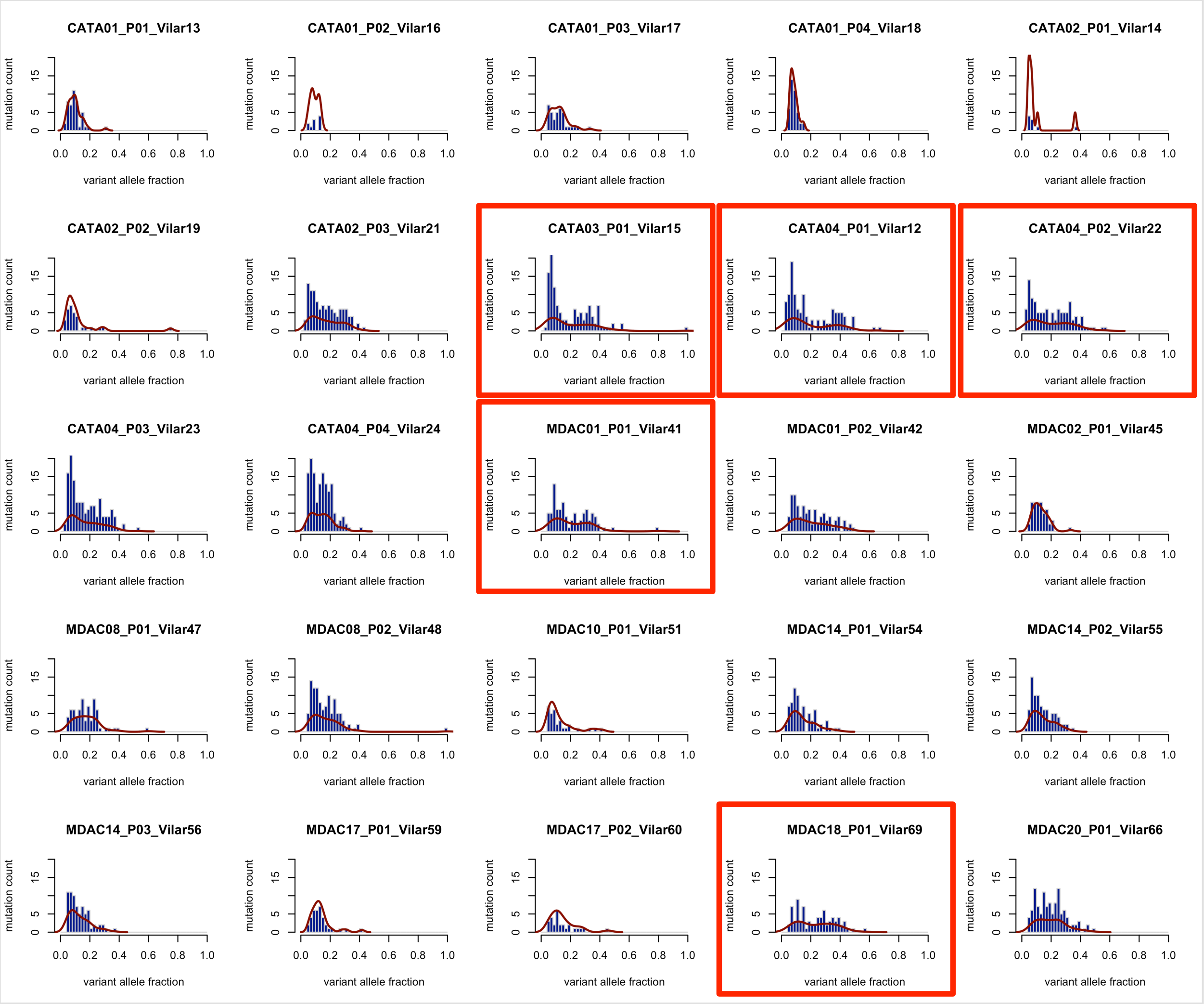


Figure 11: Distributions of variant allele fractions for all 25 polyps. Several polyps show evidence of multi-clonality (with their variant allele fraction plots boxed in red) suggesting that they are acquiring driver mutations and evolving and potentially progressing to carcinomas.

Figure 12 illustrates VAFs for all of the mutations identified in each FAP polyp using boxplots. In addition, the boxplots indicate whether or not each polyp had a somatic mutation in *APC*, and if so, a red dot indicates the VAF of that *APC* mutation. We identified somatic *APC* mutations in 52% (13 of 25) of polyps. For those samples with *APC* mutations, the *APC* mutations generally have a higher mutation allele fraction compared to other polyp mutations suggesting that somatic *APC* mutations are initiating events and reside in the founding clone of polyps.

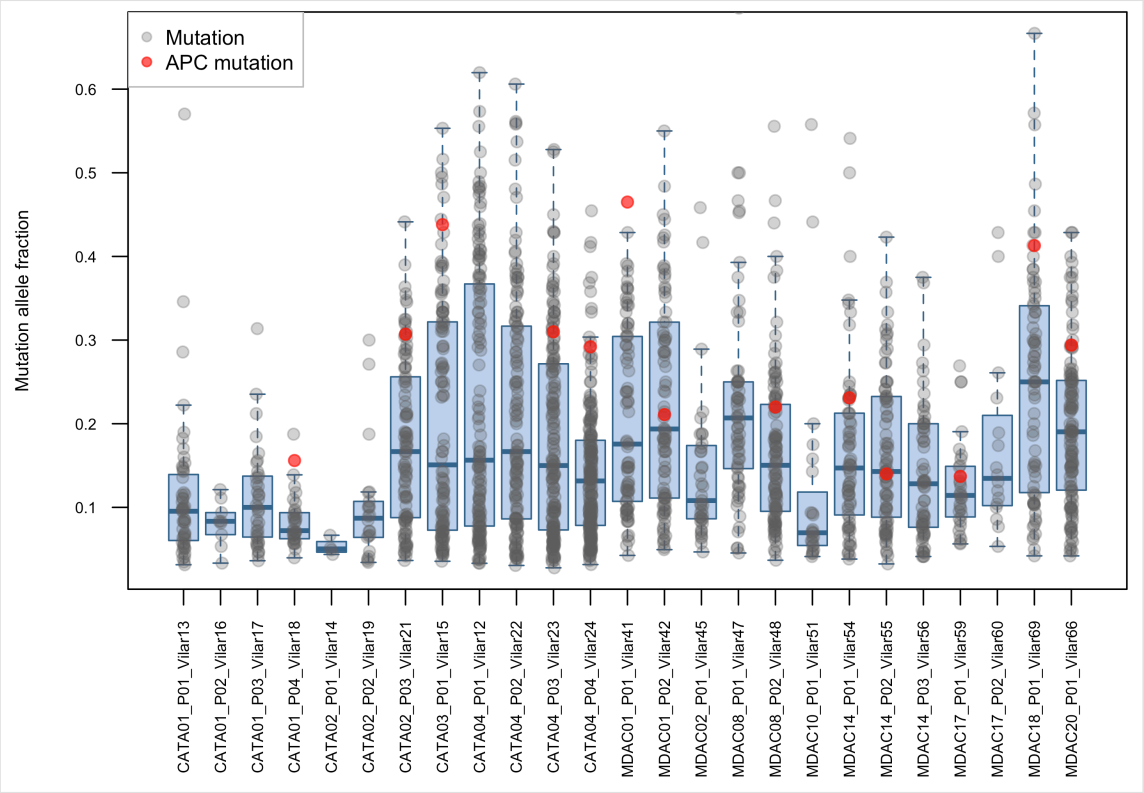


Figure 12: Mutation allele fractions for *APC* somatic events relative to those of other somatic events in the 25 FAP polyps. Samples are ordered and grouped by patient IDs, where this illuminates a potential batch effect with regards to polyp purity. For example, *CATA01* and *CATA02* samples appear to have lower allele fractions and purity relative to *CATA03*, *CATA04*, *MDAC01* and others, indicating that fluctuations in purity levels may be influenced by the persons and processes used to obtain the polyp samples.

Several of the samples without an *APC* somatic event exhibit lower mutation allele fractions, reflecting lower purity in these samples and suggesting a lower power to detect *APC* mutations. Another possibility is that a chromosomal allelic imbalance event, such as a deletion or copy-neutral LOH event may be the source of inactivation of the second *APC* allele.

2.2.2 Chromosomal allelic imbalances

A well-studied mechanism by which cancer cells alter the activity of tumor suppressor genes and oncogenes is through fluctuations in gene dosage. For FAP, deletion or LOH of chromosome 5q, where *APC* resides, is a known mechanism of *APC* loss[50]. In this project, we searched for such *chromosomal allelic imbalance* (AI) events in exome sequencing data. We define these AI events as genomic aberrations of greater than 10MB due to amplifications, deletions and cn-LOH events.

2.2.2.1 Identifying chromosomal AI events from exome sequencing data with hapLOHseq

Typically, data from *array comparative genomic hybridization* (aCGH) or *single nucleotide polymorphism* (SNP) arrays are analyzed for the identification of copy number and AI events. Often these experiments are performed in addition to exome or whole genome sequencing on tumor samples[28]. However, due to limitations in sample DNA, we were not able to run such experiments on the same polyps from which exome sequencing was performed. For this reason, we interpreted chromosomal AI events from exome sequencing data through the development of new software called *hapLOHseq*.

The traditional strategy behind identification of AI events is finding *bands of separation* as illustrated in the simulated *VAF band plots* of Figure 13. These plots show the VAFs for heterozygous sites across a region of the genome. In "normal" regions of the genome, we would expect that at heterozygous sites, 50% of reads would harbor the variant allele and the other 50% would harbor the reference allele. Thus, in "normal" regions we would expect to see a VAF band at 0.5. This is illustrated with the *normal sample* of Figure 13. In the *100% tumor sample*, there is an AI event, which results in the frequencies of the alleles of one haplotype (e.g., a configuration of alleles on one chromosome) to elevate in excess of the second haplotype where the allele frequencies decreased, resulting in 2 separate bands showing the deviation of the VAFs from 0.5. Thus these haplotypes, or alleles, are not in a 50/50 balance, rather, this is an allelic *imbalance* event.

The band separation in the *100% tumor sample* is obvious, so the event can be identified visually in these plots or via a simple algorithm. However, as the tumor purity decreases, such as in the *30% tumor* and *15% tumor* samples, it becomes harder to visually identify this event. We developed a method for exome sequencing data that addresses issues in detection of *subtle* allelic imbalances in the context of lower tumor purities. Our method, *hapLOHseq*, is a next-generation sequencing based extension of *hapLOH*[51], which is an allelic-imbalance detection method that is designed for SNP microarray data.

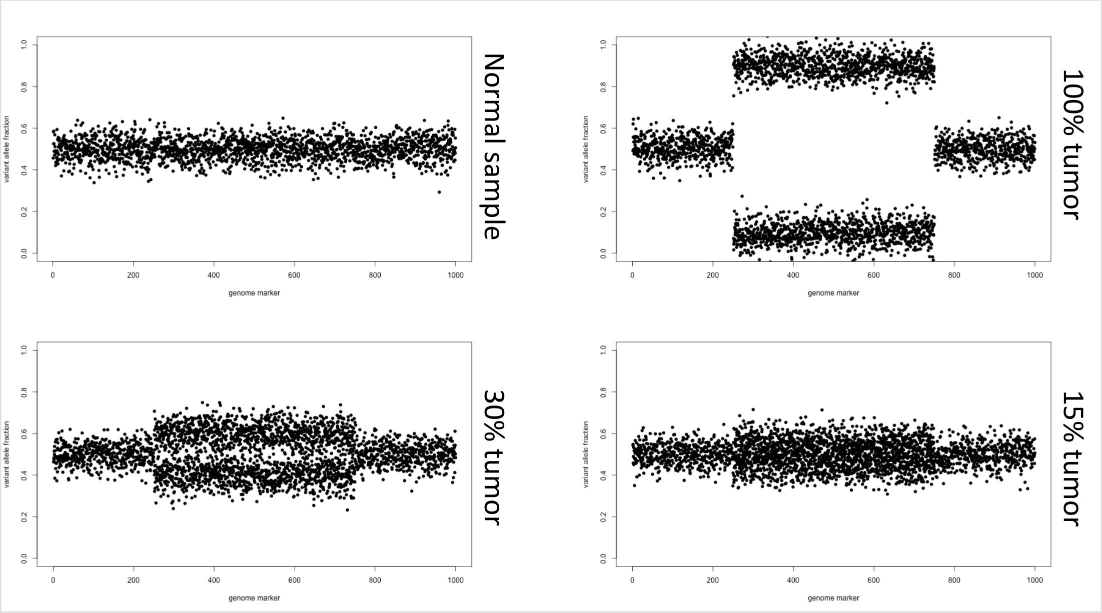


Figure 13: Methods that detect chromosomal allelic imbalance are dependent on either coverage fluctuations (not shown here) or identifying separation of allele frequency bands away from 0.5.

hapLOHseq extends the idea of searching for heterozygous sites with VAFs that deviate from 0.5 to instead look for deviation of *haplotype allele frequencies* deviating from 0.5, improving the sensitivity of identifying such events. In the simulated VAF data of Figure 14, at *10% tumor* purity, it is difficult to see any deviation of VAFs from 0.5 because the two bands overlap each other and there is no visual separation between them. But if we knew the haplotypes, and then looked for allelic imbalance between the haplotypes, the event becomes easier to discern. This is the intuition behind the method of *hapLOHseq,* i.e., to first estimate haplotypes and then check if there is allelic imbalance between the 2 haplotypes. The method is described in Appendix B section 6.2.

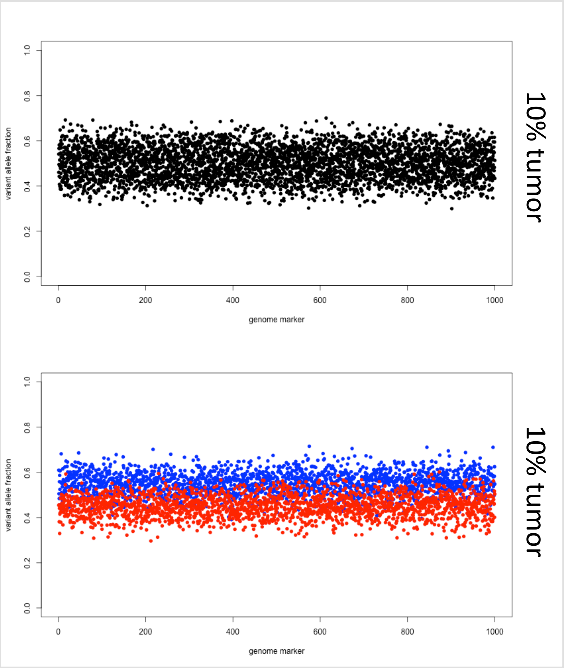
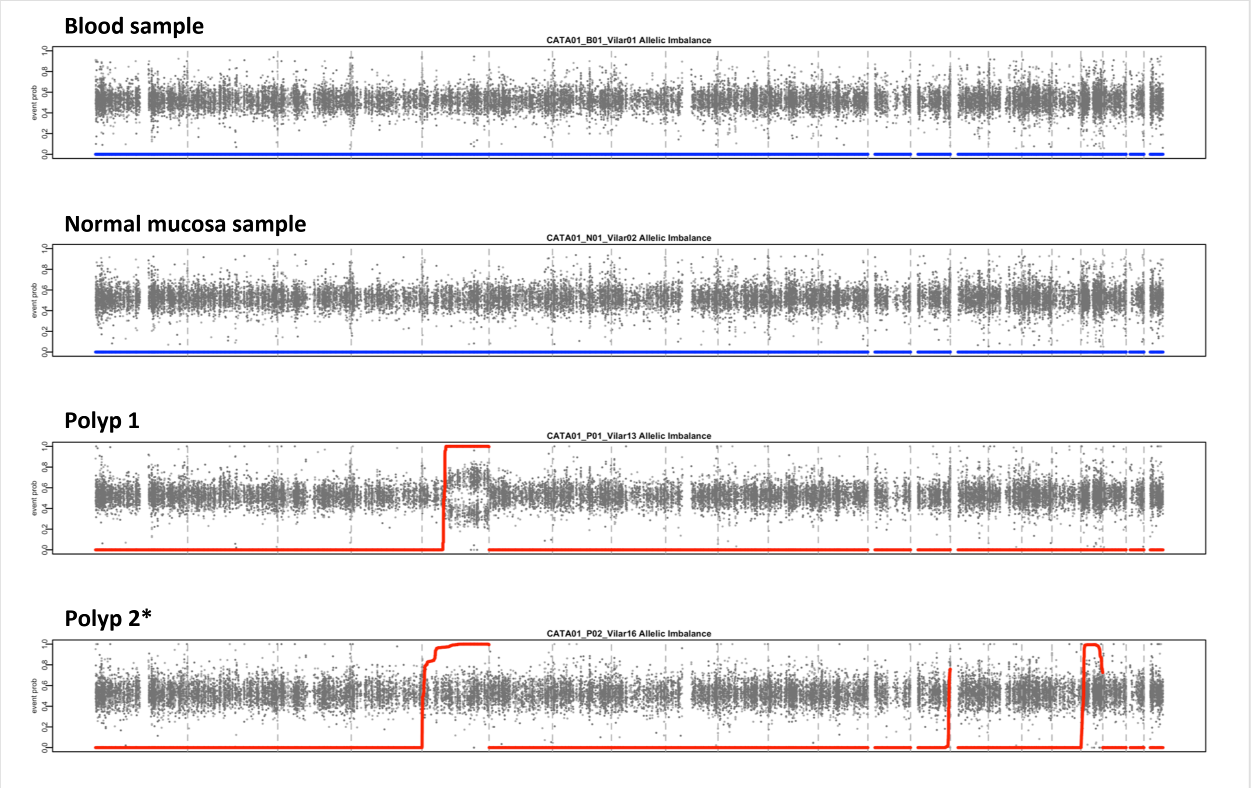


Figure 14: The intuition behind *hapLOH* and *hapLOHseq* is that these methods look for allelic imbalance of haplotypes rather than band separation of individual VAFs. Haplotype imbalances are easier to identify when the events are subtle compared to visually identifying VAF band separation.

2.2.2.2 *Chromosomal AI profiling*

We executed *hapLOHseq* with default parameters on all samples in the FAP project. *hapLOHseq plots* for 4 samples from the patient CATA01 are shown in Figure 15. For the blood, normal mucosa and 2 polyp samples, each *hapLOHseq* plot illustrates the VAFs for the corresponding sample at all genomic markers that are heterozygous in the *germline* (blood) sample for CATA01. Thus, *hapLOHseq* characterizes and identifies allelic imbalances of germline haplotypes in the polyp samples. The blue and red lines of *hapLOHseq* plots show the probabilities of chromosomal AI events across the normal and polyp genomes, respectively.



*Very low mutation rate (only 12 mutations across exome)

Figure 15: *hapLOHseq plots* for 4 samples from the CATA01 patient. The flat blue lines for the blood and normal mucosa samples reflect probabilities near 0 for AI events across these samples. The red lines represent the probabilities of events (ranging from 0 at the bottom of the plot to 1 at the top of the plot) across the polyps.

For the CATA01 blood and normal mucosa samples in Figure 15, the probabilities of AI events are virtually 0 across these genomes. For *polyp 1*, an AI event was identified with a probability near 1 at chromosome 5q. The band separation at the *hapLOHseq* identified AI region can be visually verified. *Polyp 2* is suspected of having very low polyp purity due to its very low mutation rate of 0.246 mutations per megabase. Illustrating the capabilities of identifying subtle AI at low purities, AI events were also identified in *polyp 2* at chromosomes 5 and 19.

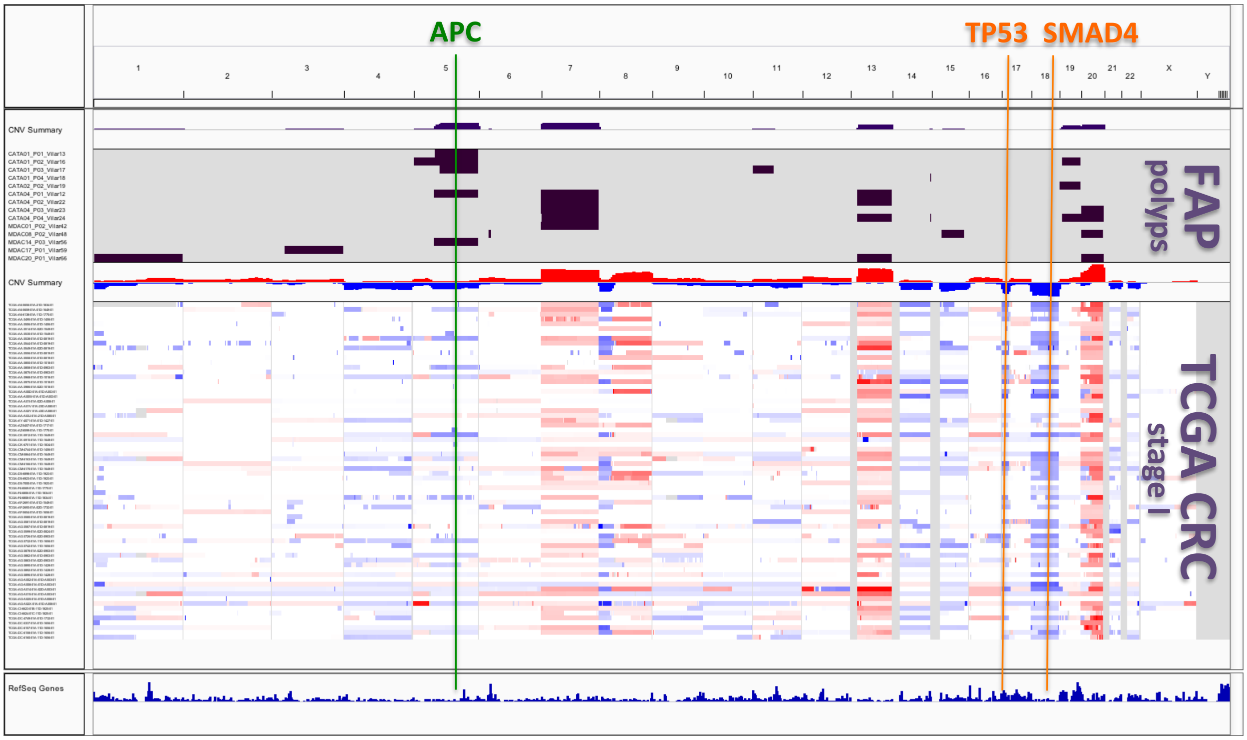


Figure 16: Summary of FAP polyp chromosomal allelic imbalance events identified by *hapLOHseq* compared to recurrent amplifications and deletions inferred from SNP arrays from the TCGA CRC project.

Similarly, *hapLOHseq* was applied to all of the FAP patient samples and then the results were summarized and compared with CRC AI profiles. For CRC AI events, copy number profiles inferred from SNP arrays were downloaded for 70 stage I microsatellite stable tumors. Figure 16 summarizes AI events for FAP and TCGA CRC samples across the genome. AI events in the polyps are illustrated with black bars in the top section. TCGA CRC stage I tumor events are illustrated with red and blue bars (amplifications and deletions respectively) below the FAP polyp AI events. A summary histogram for both the polyp and CRC events are shown at the tops of both sections.

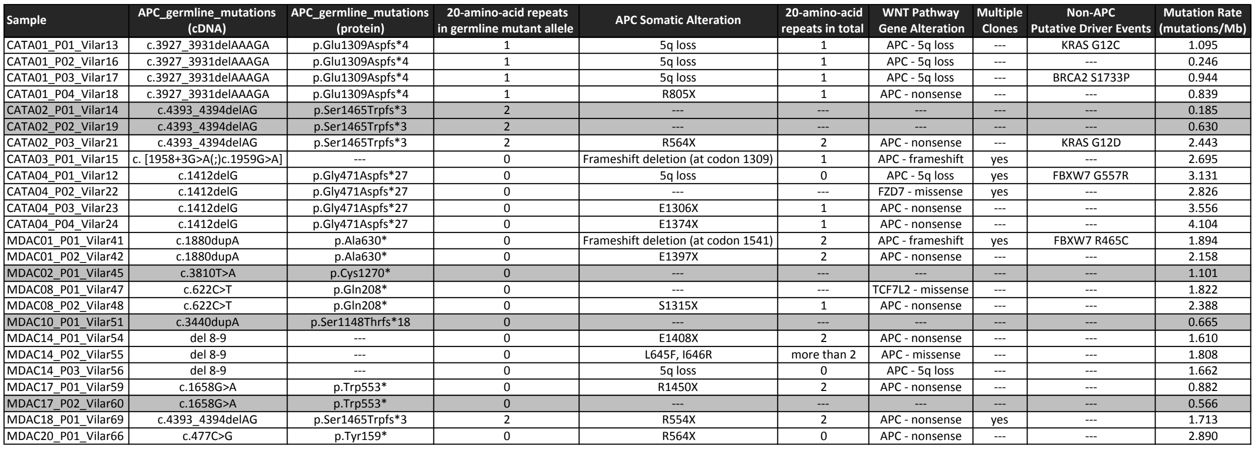
For polyp AI events, we did not distinguish among amplifications, deletions and copy-neutral LOH events in Figure 16. Some of these events were characterized visually by comparing the coverage profiles of tumor samples (at these AI events) to their paired normal samples. Amplifications, deletions and cn-LOH classifications for recurrently aberrant genes within these regions are depicted in Figure 19. However, many of the polyp AI events are too subtle to confidently determine the event type. In these cases, we classified the events as *subtle AI*. For consistency in the current comparison of these polyp AI events to CRC copy number alterations, we treat the *hapLOHseq* event calls as generic allelic imbalance events.

By comparing FAP polyp AI events to chromosomal copy number aberrations in stage 1 CRC, we can identify regions of the genome that have events in both polyps and stage 1 CRCs. Regions that are aberrant in both data sets are suggestive of events occurring early in the development of CRCs. Common aberrant regions in both data sets are loss of 5q (through deletion or cn-LOH) and amplification of chromosome 7, 13 and 20. Loss of 5q is a common mechanism of losing one copy of an *APC* gene and was observed in 25% (5 of 25) of polyps. Gains of chromosomes 7, 13 and 20 have been associated with the early development of carcinomas in previous studies[52]. Chromosome 7 increases are thought to be mechanisms of *EGFR* and *MET* oncogene amplification[53]. Chromosome 13 amplification has been associated with gains in the oncogene *CDX2*[53]. Genes associated with the colon adenoma to carcinoma progression as a result of gains in chromosome 20 include: *C20orf24*, *AURKA*, *RNPC1*, *TH1L*, *ADRM1*, *C20orf20* and *TCFL5*[54]. Gain of 20q has also been associated with a progression towards invasiveness[54]. The mutation rate of polyps with allelic imbalance of 20q is 3.235 mutations per megabase, which is higher than the overall mean mutation rate of our polyps, which is 1.75 mutations per megabase (t-test p-value = 8.575e-4). Further, the mutation rate of polyps with allelic imbalances on chromosome 20 approach the mutation rate of 3.95 mutations per megabase of the TCGC non-hypermutated CRCs, supporting the idea that these polyps have progressed further along the adenoma-to-carcinoma sequence.

Recurrent amplification and deletion regions in the CRC data set that are absent in the polyp AI data are representative of events associated with later adenoma development, potentially representing events necessary for transforming a polyp into a carcinoma. Deletions of 17p and 18q are seen in a high proportion of stage 1 CRCs in TCGA, 60% and 70% respectively. Deletion of 17p has been associated with loss of *TP53* and deletion of 18q is associated with the loss of *SMAD2* and *SMAD4* (see Figure 16). These genes have been associated with carcinoma initiation in the adenoma-to-carcinoma sequence, reflecting the utility of our FAP polyp and CRC comparative approach.

2.2.3 Patterns of *APC* somatic events

In general, events characterizing FAP polyps in our data set were also prevalent in CRC tumors. Consistent with current knowledge of FAP polyps, recurrent LOH of 5q was observed across polyps and most other identified somatic *APC* events were truncating mutations. Bi-allelic loss of *APC* was widespread and was detected in 72% (17 of 25) of polyps. Five of the 7 polyps lacking a somatic alteration of *APC* also lacked any WNT signaling pathway alterations (see section 2.2.4). These samples appeared to have lower polyp purity compared to the 80% where somatic WNT pathway alterations were identified, which results in a lower power to detect somatic events in these samples. This lower purity is reflected in the lower detectable mutation rate in these samples (0.629 mutations/Mb) versus that in WNT altered samples (2.035 mutations/Mb; t-test p-value = 0.005). As such, we believe that 72% to 100% of the FAP polyps have bi-allelic loss of *APC*. One truncating mutation at codon 564 was seen in 2 polyps from different patients: *CATA02_P03_Vilar21* and *MDAC20_P01_Vilar66*. This recurrent mutation has been documented in 21 large intestine samples in the COSMIC database[43]. In all, putative results show that 5 of the somatic *APC* events were due to deletion or LOH of 5q, 2 were single base frameshift deletions, 2 were missense mutations (identified in the same patient) and 10 were nonsense mutations. Of the 10 nonsense mutations, 5 (50%) were C->T transitions, 4 (40%) were G->T transversions and 1 (10%) was a C->A transversion.



Sample	APC_germline_mutations (cDNA)	APC_germline_mutations (protein)	20-amino-acid repeats in germline mutant allele	APC Somatic Alteration	20-amino-acid repeats in total	WNT Pathway Gene Alteration	Multiple Clones	Non-APC Putative Driver Events	Mutation Rate (mutations/Mb)
CATA01_P01_Vilar13	c.3927_3931delAAAGA	p.Glu1309Aspfs*4	1	5q loss	1	APC - 5q loss	---	KRAS G12C	1.095
CATA01_P02_Vilar16	c.3927_3931delAAAGA	p.Glu1309Aspfs*4	1	5q loss	1	APC - 5q loss	---	---	0.246
CATA01_P03_Vilar17	c.3927_3931delAAAGA	p.Glu1309Aspfs*4	1	5q loss	1	APC - 5q loss	---	BRCA2 S1733P	0.944
CATA01_P04_Vilar18	c.3927_3931delAAAGA	p.Glu1309Aspfs*4	1	R805X	1	APC - nonsense	---	---	0.839
CATA02_P01_Vilar14	c.4393_4394delAG	p.Ser1465Trpfs*3	2	---	---	---	---	---	0.185
CATA02_P02_Vilar19	c.4393_4394delAG	p.Ser1465Trpfs*3	2	---	---	---	---	---	0.630
CATA02_P03_Vilar21	c.4393_4394delAG	p.Ser1465Trpfs*3	2	R564X	2	APC - nonsense	---	KRAS G12D	2.443
CATA03_P01_Vilar15	c. [1958+3G>A(;)c.1959G>A]	---	0	Frameshift deletion (at codon 1309)	1	APC - frameshift	yes	---	2.695
CATA04_P01_Vilar12	c.1412delG	p.Gly471Aspfs*27	0	5q loss	0	APC - 5q loss	yes	FBXW7 G557R	3.131
CATA04_P02_Vilar22	c.1412delG	p.Gly471Aspfs*27	0	---	---	FZD7 - missense	yes	---	2.826
CATA04_P03_Vilar23	c.1412delG	p.Gly471Aspfs*27	0	E1306X	1	APC - nonsense	---	---	3.556
CATA04_P04_Vilar24	c.1412delG	p.Gly471Aspfs*27	0	E1374X	1	APC - nonsense	---	---	4.104
MDAC01_P01_Vilar41	c.1880dupA	p.Ala630*	0	Frameshift deletion (at codon 1541)	2	APC - frameshift	yes	FBXW7 R465C	1.894
MDAC01_P02_Vilar42	c.1880dupA	p.Ala630*	0	E1397X	2	APC - nonsense	---	---	2.158
MDAC02_P01_Vilar45	c.3810T>A	p.Cys1270*	0	---	---	---	---	---	1.101
MDAC08_P01_Vilar47	c.622C>T	p.Gln208*	0	---	---	TCF7L2 - missense	---	---	1.822
MDAC08_P02_Vilar48	c.622C>T	p.Gln208*	0	S1315X	1	APC - nonsense	---	---	2.388
MDAC10_P01_Vilar51	c.3440dupA	p.Ser1148Thrfs*18	0	---	---	---	---	---	0.665
MDAC14_P01_Vilar54	del 8-9	---	0	E1408X	2	APC - nonsense	---	---	1.610
MDAC14_P02_Vilar55	del 8-9	---	0	L645F, I646R	more than 2	APC - missense	---	---	1.808
MDAC14_P03_Vilar56	del 8-9	---	0	5q loss	0	APC - 5q loss	---	---	1.662
MDAC17_P01_Vilar59	c.1658G>A	p.Trp553*	0	R1450X	2	APC - nonsense	---	---	0.882
MDAC17_P02_Vilar60	c.1658G>A	p.Trp553*	0	---	---	---	---	---	0.566
MDAC18_P01_Vilar69	c.4393_4394delAG	p.Ser1465Trpfs*3	2	R554X	2	APC - nonsense	yes	---	1.713
MDAC20_P01_Vilar66	c.477C>G	p.Tyr159*	0	R564X	0	APC - nonsense	---	---	2.890

Table 4: Summary of polyp somatic events in *APC* and the WNT signaling pathway.

Literature has suggested that given the location within the first altered *APC* allele, one can loosely predict the location and type of the somatic alteration in the second *APC* allele. In a proposed *first-hit-second-hit model*, *APC* germline mutations near codon 1300, specifically between codons 1285 and 1378, are associated with somatic chromosome 5q loss[55]. *CATA01* is the only patient in our data set with such a germline mutation. Consistent with this *first-hit-second-hit* model, 75% (3 of 4) of *CATA01* polyps had loss of chromosome 5q, but this event was seen in only 9.5% (2 of 21) of the remaining samples where a somatic *APC* event was identified. Thus an association exists (p-value = 0.016, Fisher's exact test) between germline mutations (between codons 1285 and 1378) and chromosome 5q loss in our data set.

Further, the *first-hit-second-hit* model proposes that for patients with *APC* truncating germline mutations before codon 1264, the first repeat region of the beta-catenin binding and degradation portion of *APC*, LOH or deletion of chromosome 5q is very rare[55]. Although not at statistical significance our data set also suggests that this pattern is valid. Four of our FAP patients have germline mutations located prior to the first 20 amino-acid repeat region. Only 1 in 9 of their polyps had a somatic loss of 5q. Alternatively, 25% (4 of 16) of the remaining samples had 5q AI events.

The 20-amino-acid repeat regions of *APC* are involved in beta-catenin binding and degradation. It has been proposed that between the 2 *APC* alleles of cells there is an optimal number of beta-catenin repeats for cellular proliferation, where this theory motivates the *first-hit-second-hit-model* of *APC* somatic mutation in patients with FAP such that a polyp cell ends up with an optimal number of 1 beta-catenin repeat in the so-called *just-right* signaling model[50]. Thus, where a person has a truncating mutation that leaves *APC* with 1 beta-catenin repeat, they are more likely to have loss of *APC* (or loss of 5q) as a somatic event, leaving them with 1 beta-catenin repeat in total. This pattern is observed in our dataset. Eighteen polyps exhibited bi-allelic loss of *APC* and most commonly, adenoma cells were left with 1 beta-catenin repeat. As depicted in Table 4, 1 polyp had more than 2 of the 20 amino-acid repeats within its 2 *APC* alleles, 6 polyps had 2 of the 20 amino-acid repeat regions, 8 had 1 repeat region and 3 polyps had 0 repeat regions. In summary, our data suggests that *APC* germline mutations may be used as a predictor of future somatic events based on the *first-hit-second-hit* model [50] proposed for adenoma development.

2.2.4 Alterations in WNT signaling genes are pervasive

Activation of the WNT signaling pathway causes an accumulation of beta-catenin in the cytoplasm, leading to its eventual translocation into the nucleus where it acts as a transcriptional coactivator of *MYC* and other genes resulting in cellular proliferation. Without WNT signaling, a destruction complex would normally degrade beta-catenin. WNT pathway genes were identified as altered in 92% of TCGA CRCs[28]. Specific genes that have been identified as significantly mutated in TCGA CRCs are included in a simplified representation of the WNT signaling pathway in Figure 17.

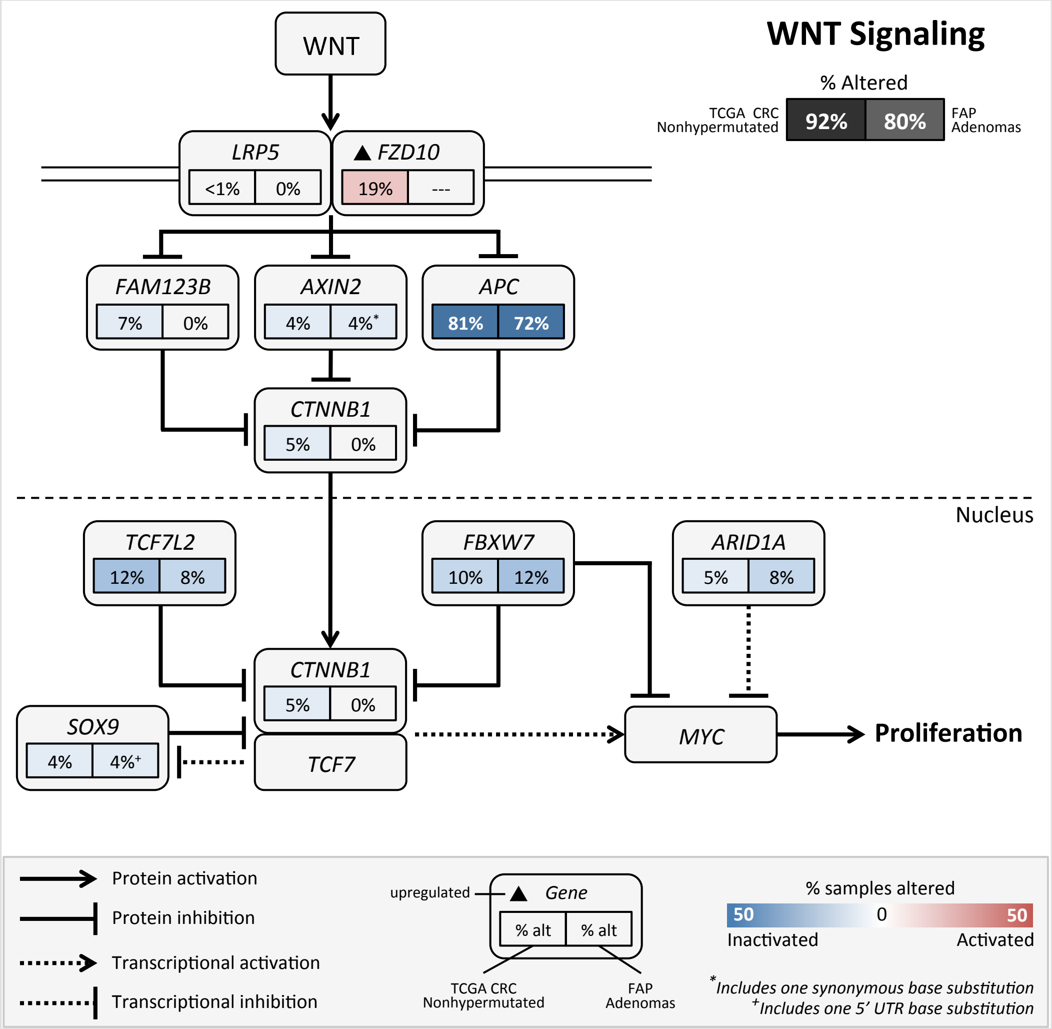


Figure 17: Significantly mutated TCGA CRC genes involved in the canonical WNT signaling pathway. Several beta-catenin inhibitors are mutated in both TCGA CRC and FAP polyps including: *APC*, *TCF7L2*, and *FBXW7*, suggesting these are key events in the early development of CRC.

Similarly, somatic alterations of WNT signaling pathway genes were seen in 80% (20 of 25) of FAP adenomas. *APC* is the predominantly mutated gene. Mutations resulting in inactivation of *APC*, *TCF7L2* and *FBXW7* appear to contribute to the proliferation of beta-catenin and the growth of polyps. Further, nonsense mutations of *ARID1A* may contribute to

MYC over-expression. Overexpression of both beta-catenin and *MYC* contribute to cellular proliferation. Additionally, putative key genes with non-coding alterations in FAP adenomas include *AXIN2* and *SOX9*, which harbored a synonymous and a 5′ UTR base substitution, respectively.

We did not detect an alteration of WNT signaling pathway genes in 5 polyps represented with rows shaded in gray in Table 4. However, these 5 polyps had low mutation rates (0.629 mutations/Mb) compared to the rest of the polyps (2.035 mutations/Mb) suggesting that we may have missed important mutations in these polyps due to a lack of power to detect them. In addition, Obrador-Hevia *et al* profiled 60 adenomas and identified WNT pathway gene aberrations at either the DNA and/or RNA level in all adenomas[56], suggesting that another source of WNT aberrations that we are missing may be hidden in the mRNA transcripts of these polyps. Additionally, TCGA identified over-expression of *FZD10* in 19% of CRCs (Figure 17). Without RNA sequencing of these 25 polyps, we lack the ability to identify transcript level WNT pathway aberrations.

Looking specifically at the WNT, MAPK and ERBB signaling pathways, which are known to be aberrant in CRC, Table 5 indicates the gene and type of mutation observed in each sample for each of these pathways. The genes for each of these pathways were obtained from the *Molecular Signatures Database* (MSigDB)[47] using the KEGG pathway names indicated on the table. The table excludes AI events and only lists somatic point mutations and small insertions and deletions. Samples with alterations in the WNT signaling pathway in addition to the MAPK and/or ERBB pathways are shaded in gray. We expect that these samples have progressed further along the adenoma-to-carcinoma sequence as compared to the other samples. This is supported by the difference in mutation rates between these samples (2.565 mutations/Mb) as compared to the remaining samples (1.552 mutations/Mb) through a T-test (p-value = 0.030).

sample name	KEGG_WNT_SIGNALING_PATHWAY	KEGG_MAPK_SIGNALING_PATHWAY	KEGG_ERBB_SIGNALING_PATHWAY	mutation rate (mutations/Mb)
CATA01_P01_Vilar13	---	KRAS:MISSENSE	KRAS:MISSENSE	1.095
CATA01_P02_Vilar16	---	---	---	0.246
CATA01_P03_Vilar17	---	---	---	0.944
CATA01_P04_Vilar18	APC:NONSENSE;TCF7L2:MISSENSE	---	---	0.839
CATA02_P01_Vilar14	---	---	---	0.185
CATA02_P02_Vilar19	---	---	---	0.630
CATA02_P03_Vilar21	APC:NONSENSE	KRAS:MISSENSE	KRAS:MISSENSE	2.443
CATA03_P01_Vilar15	CREBBP:NONSENSE;CSNK1A1:MISSENSE;APC:FRAMESHIFT	---	---	2.695
CATA04_P01_Vilar12	---	---	STAT5A:NONSENSE	3.131
CATA04_P02_Vilar22	FZD7:MISSENSE;PPP3CB:NONSENSE	PPP3CB:NONSENSE	---	2.826
CATA04_P03_Vilar23	APC:NONSENSE;TCF7L2:NONSENSE	MAP2K7:NONSENSE	MAP2K7:NONSENSE	3.556
CATA04_P04_Vilar24	APC:NONSENSE	---	---	4.104
MDAC01_P01_Vilar41	APC:FRAMESHIFT	---	---	1.894
MDAC01_P02_Vilar42	APC:NONSENSE	---	---	2.158
MDAC02_P01_Vilar45	---	---	---	1.101
MDAC08_P01_Vilar47	TCF7L2:MISSENSE	---	---	1.822
MDAC08_P02_Vilar48	APC:NONSENSE	FGFR4:MISSENSE	---	2.388
MDAC10_P01_Vilar51	---	---	---	0.665
MDAC14_P01_Vilar54	APC:NONSENSE	DUSP10:MISSENSE	---	1.610
MDAC14_P02_Vilar55	APC:MISSENSE	---	---	1.808
MDAC14_P03_Vilar56	---	---	---	1.662
MDAC17_P01_Vilar59	APC:NONSENSE	---	---	0.882
MDAC17_P02_Vilar60	---	---	---	0.566
MDAC18_P01_Vilar69	APC:NONSENSE	---	---	1.713
MDAC20_P01_Vilar66	APC:NONSENSE	---	---	2.890

Table 5: Genes aberrant in the WNT, MAPK and ERBB signaling pathways. Note that some genes are members of both the MAPK and ERBB signaling pathways. Samples are ordered by patient IDs. Those samples with alterations in WNT signaling pathway in addition to the MAPK and/or ERBB signaling pathway are shaded in gray.

2.2.5 Candidate genes of early CRC development

Consistency in our findings with current knowledge of FAP polyps provides some confidence in our results and suggests that our candidate gene list (shown on the left side of Figure 19) identified for further functional studies may indeed contain true genes involved in the development of adenomas. In addition to *APC* somatic events being identified in 72% (18 of 25) of polyps, activating mutations in *KRAS* were detected in 8% (2 of 25) of the polyps. Events in the tumor suppressors *SMAD2/4* and *TP53* which are prevalent in CRC and thought to be events that transform adenomas to carcinomas were not seen in the FAP polyps as expected since our samples are pre-malignant lesions.

Through application of *MutSig*[49], the only identified significantly mutated gene in our data set is *APC* (p-value = 1.15e-07 and q-value = 2.18e-03). The lack of statistical significance in our candidate gene findings is likely due to our small sample size. As an alternative approach to MutSig for identifying mutated genes associated with adenoma development, we looked for recurrently altered genes in our data set that have previously been associated with CRC. The motivation behind this is that we would expect that mutated genes associated with polyp development are significantly mutated in CRCs. Using this strategy, the candidate genes listed on the left of Figure 18 were identified. Excluding *APC* and *KRAS,* which are known to play important roles in adenoma development, 17 additional candidate genes have been identified.

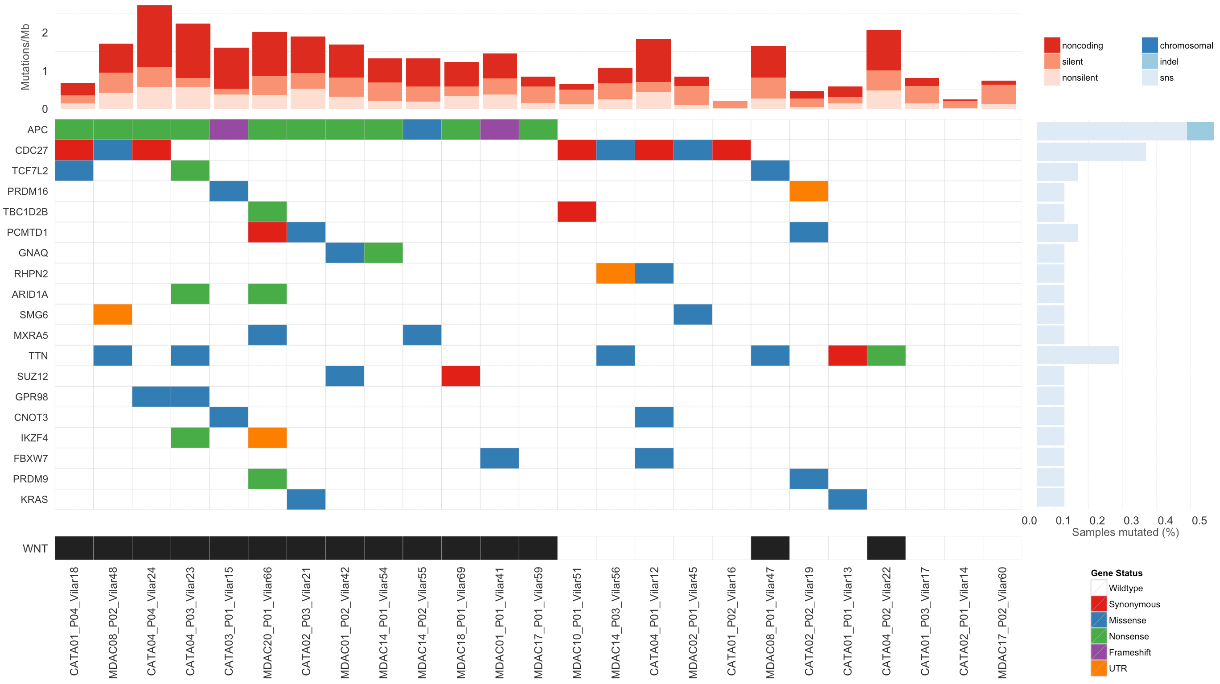


Figure 18: Map of candidate genes of early CRC development based on somatic mutation characterization.

The specific criteria for including these candidate genes are the following.

1. A candidate gene must be recurrently altered in at least 2 FAP adenomas.
2. That gene must harbor at least one mutation categorized as tier 1 and 2 or the mutation must be a nonsense mutation (see section 2.1.4 for more details on tier definitions and for details on incorporating previously identified CRC genes into our mutation prioritization).

Figure 18 is a summary characterization of our polyps and a map of our candidate genes of early CRC development. The figure shows candidate genes as rows in a categorical map, where each color corresponds to a specific type of mutation observed in each adenoma sample. The blue bar plot on the right shows the frequency at which each gene is altered in the FAP polyp samples broken down by *insertions and deletions* (indels) or *single nucleotide substitutions* (SNS). The columns of the figure correspond to the FAP polyps where the red bar plot at the top illustrates the mutation rate for each sample broken down by noncoding, silent and nonsilent alterations. Genes and samples are ordered in such a way to capture potential mutual exclusivity patterns in the data. For example, *APC* is the most frequently altered gene and *CDC27* is the next most altered gene *exclusive* of the *APC* mutated samples. Interestingly, Yu *et al* recently performed population and single-cell sequencing analyses of a bi-clonal colon cancer case and discovered that mutated *CDC27* occurs exclusively from mutated *APC* in each clone, supporting the hypotheses that each of these mutations may provide unique mechanisms of colon cancer initiation[57]. Most importantly, Figure 18 lists 17 genes (excluding *APC* and *KRAS* which are known genes involved in the early development of CRC) that meet the criteria for being candidate genes of early CRC development.

Figure 19 extends the analysis portrayed in Figure 18 to include chromosomal AI events, which we believe characterizes FAP polyps with somatic *APC* alterations more accurately. For Figure 19, the candidate gene criteria were altered such that:

1. A candidate gene must be recurrently altered in at least 2 FAP adenomas, where these alterations can include chromosomal AI.
2. That gene must harbor at least one mutation categorized as tier 1 and 2 or the mutation must be a nonsense mutation (see section 2.1.4 for more details on tier definitions and for details on incorporating previously identified CRC genes into our mutation prioritization).

In this way, even if a gene has been altered through multiple AI events in the polyps, it is not a candidate gene unless it harbors at least one highly prioritized mutation, where this prioritization is highly influenced by the mutation being predicted to be a driver event, by being recurrently altered in the COSMIC database[43], or by residing in a previously associated CRC gene (see section 2.1.4 for more details).

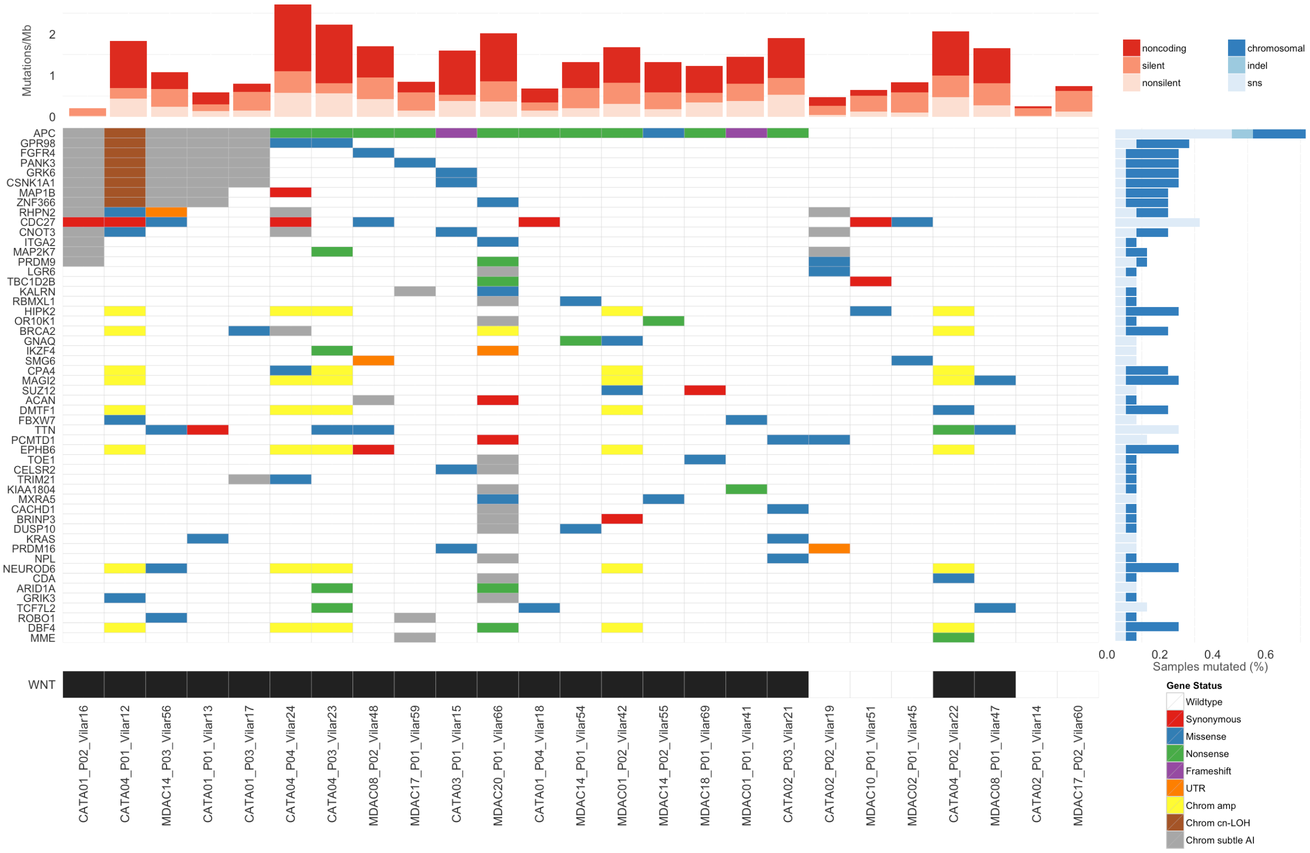


Figure 19: Map of candidate genes of early CRC development based on somatic mutation characterization considering AI events.

The aberration summary of Figure 19 now includes 5 additional *APC* alterations due to chromosomal AI events, resulting in 72% (18 of 25) polyps having a somatic *APC* alteration. Fifty-one candidate genes (32 additional candidate genes compared to those listed in Figure 18) were also identified as potentially contributing to early CRC development. The blue bar plot on the right side of Figure 19 now accounts for chromosomal AI events. In cases, where AI events could visually be categorized as amplifications, deletions or copy-neutral LOH by inspecting read coverage profiles across the genomes of these polyp samples and comparing them to coverage profiles of their paired blood samples, the events are colored accordingly. In cases of *subtle AI* (or AI events at very low cellular proportions) it is difficult to visually distinguish between amplifications, deletions or LOH events and the events are left as *subtle AI,* although in each of these cases in our FAP data set, it appears that these events are either deletions or copy-neutral LOH events.

2.3 Discussion

2.3.1 Challenges of molecularly profiling polyps

In this chapter we have characterized the genomes of polyp samples, which can be more difficult to characterize compared to tumor samples due to problems of limited amounts sample and low purity. With the goals of performing mutation and chromosomal AI profiling of these pre-malignant lesions, our project presented 2 primary challenges as compared to tumor sequencing projects:

1. The polyps obtained for experimentation were small and therefore we had limited genetic material to perform additional experiments beyond exome sequencing. We would have liked to run *comparative genomic hybridization* (CGH) or *single nucleotide polymorphism* (SNP) arrays for higher-resolution identification of copy number events for the same polyps that we performed exome sequencing on, however this was not possible due to a limited amount of quality polyp DNA. The DNA remaining after exome sequencing was conserved for Sanger sequencing validation of important identified mutations.
2. Adenomas inherently have high stromal contamination due to the fact that they develop earlier in the adenoma-to-carcinoma sequence and more closely resemble "normal epithelium" cells in biological state as compared to carcinomas.

To overcome these challenges for the genomic characterization from limited, low-purity samples, we performed analyses on the exome sequencing data using *Mutect* for the sensitive detection of point mutations and we developed our own software, *hapLOHseq*, for the detection of subtle chromosomal AI events. *hapLOHseq* was applied to the FAP exome sequencing data in section 2.2.2 and the method is described in detail in Appendix B in section 6.2.

2.3.2 Significance of findings

In this chapter, we presented the first genomic characterization of FAP adenomas performed through NGS. Through mutation profiling, we showed that FAP polyp mutation rates are lower compared to CRC mutation rates indicating that FAP adenomas are in an earlier stage of tumorigenesis compared to CRC as expected. In addition, mutational base substitution signatures of polyps appear to be identical to those of nonhypermutated CRCs, suggesting that these two types of samples have the same mutational processes driving them. These processes appear to be related to aging or other currently unidentified processes[49]. Several of the polyps also appear to be *multi-clonal*, supporting the idea that they are evolving towards carcinomas and acquiring driver mutations such as those on *KRAS*. More fundamentally, most of the polyps (72%) exhibited *second hits* of the tumor suppressor gene *APC*, of which, bi-allelic loss is thought to be the initiating event in the development of adenomas according to the adenoma-to-carcinoma sequence model. For those adenomas lacking *APC* somatic events, this could be due to lower power to detect *APC* events because of low polyp purities that these samples appear to have, or they could have alterations not detectable through exome sequencing, such as transcript level alterations or epigenetic alterations. Alternatively, these samples could be harboring other important mutations exclusive of *APC*. Eighty percent of the adenomas harbored alterations in WNT signaling. Additional WNT signaling genes besides *APC* that are altered include *AXIN2*, *TCF7L2*, *FBXW7*, *SOX9* and *ARID1A*. In total, 50 candidate genes were identified (excluding *APC* and *KRAS*) that are putatively involved in the early development of CRC. These genes are currently being functionally tested in vitro through cell-line and animal model experiments by the lab of Eduardo Vilar.

3 Identification of candidate chemopreventive drugs for FAP

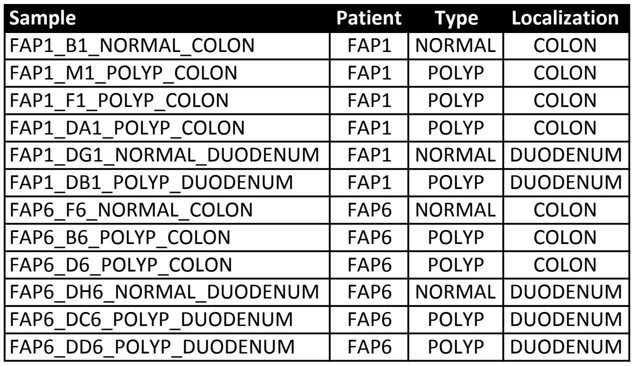
In this chapter, I present a separate but complementary phase of our project for the identification of candidate chemopreventive drugs for FAP patients (see Figure 3 for a high-level overview of the 2 phases of our FAP project). In the longer-term vision of this project, knowledge gained and candidate gene targets identified from the genomic characterization of FAP polyps (see chapter 2) will be used to inform the identification of candidate chemopreventive drugs. Indeed, that is our hope as functional studies are currently being performed to validate the candidate genes identified in the genomic characterization of colon adenomas. However, at the present time, we are taking what could be viewed as a more direct approach for identifying candidate drugs. We have performed RNA sequencing of colon and duodenum samples in FAP patients, defined gene expression signatures representative of the differences between FAP at-risk normal mucosa and polyps, and then identified candidate drugs to directly target these gene expression signatures.

Ideally, chemoprevention strategies would incorporate drugs or compounds that have minimal toxicity that are inexpensive and effective. So we take a drug repurposing approach, by screening for candidate drug compounds that include U.S. *Food and Drug Administration* (FDA)-approved drugs and nondrug bioactive compounds, which are generally considered to be safe but which may not have been shown to be effective in their originally intended purposes in addition to drugs that have been shown to be effective for various uses[58]. In this chapter, we describe this computational screening approach, apply it to FAP polyps, and propose candidate drugs for FAP patient chemoprevention.

3.1 Methods

3.1.1 Available patients and samples

Samples from 2 FAP patients followed at MD Anderson Cancer Center were collected through endoscopic biopsy (see Table 6). From each patient, normal mucosa samples from both the colon and duodenum were obtained. From patient *FAP1*, 3 colon polyps and 1 duodenum polyp were collected, and from *FAP6*, 2 colon polyps and 2 duodenum polyps were obtained. RNA from polyp and normal mucosa samples were isolated using a combined protocol with *TRIzol* reagent (Life Technologies) and the *RNeasy Mini Kit* (Qiagen).

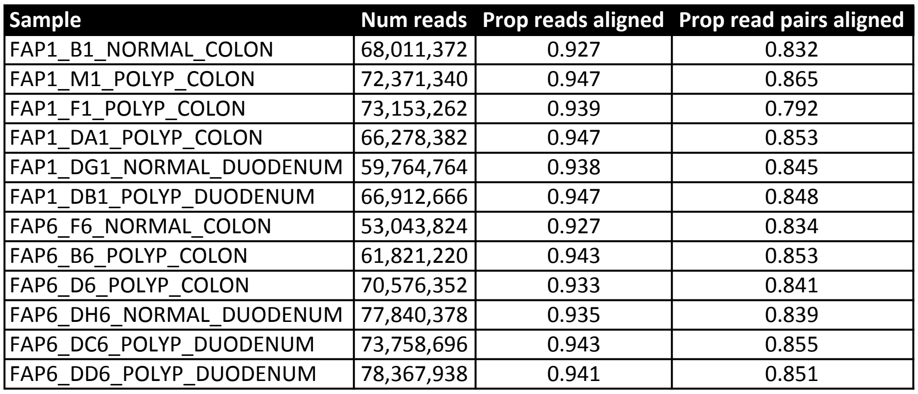


Sample	Patient	Type	Localization
FAP1_B1_NORMAL_COLON	FAP1	NORMAL	COLON
FAP1_M1_POLYP_COLON	FAP1	POLYP	COLON
FAP1_F1_POLYP_COLON	FAP1	POLYP	COLON
FAP1_DA1_POLYP_COLON	FAP1	POLYP	COLON
FAP1_DG1_NORMAL_DUODENUM	FAP1	NORMAL	DUODENUM
FAP1_DB1_POLYP_DUODENUM	FAP1	POLYP	DUODENUM
FAP6_F6_NORMAL_COLON	FAP6	NORMAL	COLON
FAP6_B6_POLYP_COLON	FAP6	POLYP	COLON
FAP6_D6_POLYP_COLON	FAP6	POLYP	COLON
FAP6_DH6_NORMAL_DUODENUM	FAP6	NORMAL	DUODENUM
FAP6_DC6_POLYP_DUODENUM	FAP6	POLYP	DUODENUM
FAP6_DD6_POLYP_DUODENUM	FAP6	POLYP	DUODENUM

Table 6: Four normal mucosa and 8 polyps were RNA sequenced from the colon and duodenum of 2 FAP patients.

3.1.2 Data collection

RNA from the FAP samples was sequenced on an Illumina Hiseq 2000 sequencer with 76 base paired-end reads at the MD Anderson Cancer Center sequencing core facility. Reads were aligned and analyzed using the *Tuxedo* protocol for differential gene expression analyses of RNA sequence data[59]. Briefly, *TopHat*[60] is used for alignment of initial reads to the human reference genome hg19. Sequence run summaries are provided in Table 7 and based on the proportion of read pairs aligned (the Tuxedo protocol specifies 0.7 as being representative of quality samples), our samples all appear to have provided high-quality sequencing reads. *Cufflinks*[61] then assembles and quantifies transcripts. Subsequently, the software package *Cuffdiff*[62] is used to identify differentially expressed genes. In a separate analysis, *Tophat-fusion*[63] is used to identify gene fusions. See Appendix A sections 5.5 and 5.6, for more details.



Sample	Num reads	Prop reads aligned	Prop read pairs aligned
FAP1_B1_NORMAL_COLON	68,011,372	0.927	0.832
FAP1_M1_POLYP_COLON	72,371,340	0.947	0.865
FAP1_F1_POLYP_COLON	73,153,262	0.939	0.792
FAP1_DA1_POLYP_COLON	66,278,382	0.947	0.853
FAP1_DG1_NORMAL_DUODENUM	59,764,764	0.938	0.845
FAP1_DB1_POLYP_DUODENUM	66,912,666	0.947	0.848
FAP6_F6_NORMAL_COLON	53,043,824	0.927	0.834
FAP6_B6_POLYP_COLON	61,821,220	0.943	0.853
FAP6_D6_POLYP_COLON	70,576,352	0.933	0.841
FAP6_DH6_NORMAL_DUODENUM	77,840,378	0.935	0.839
FAP6_DC6_POLYP_DUODENUM	73,758,696	0.943	0.855
FAP6_DD6_POLYP_DUODENUM	78,367,938	0.941	0.851

Table 7: RNA sequencing mapping statistics indicate that the sequencing quality of each of these samples is optimal. A "% read pairs aligned" > 0.70 indicates a good quality sequencing run[59].

3.1.3 Defining FAP colon and duodenum gene expression signatures

In our project, a gene expression signature is a set of up-regulated and down-regulated genes identified by comparing at-risk normal mucosa to polyp samples of FAP patients. These signatures are representative of the molecular alterations that differentiate at-risk normal-mucosa and adenomas. The normal mucosa samples of FAP patients are *at-risk* because they harbor aberrant copies of the APC gene. Given the quantified transcripts reported by Cufflinks, the software package Cuffdiff is used to identify differentially expressed genes between 2 sets of samples, such as colon polyps versus colon normal mucosa samples. To label a gene as differentially expressed (e.g., up-regulated or down-regulated) for inclusion into the gene signature, we require that Cuffdiff adjusted Benjamini-Hochberg p-values be less than 0.05. For up-regulated signature genes, we require a $\log_2$ fold-change >= 1. For down-regulated genes, we require a $\log_2$ fold-change <= -1. When applied to our computational drug-screening experiments, this gene expression signature is called a *query signature*.

3.1.4 Identifying candidate drugs to target FAP gene expression signatures

The query signature is then fed into a software application that we have developed called the *Cancer in silico Drug Discovery* framework (CiDD), where CiDD produces a report for *connections* (or *negative correlations*) between the query signature and gene expression signatures induced by candidate drug compounds. CiDD screens the FAP at-risk normal mucosa gene expression signatures against those induced by drug compounds in the *Connectivity Map* (CMap)[58] to identify candidate drugs that may target the FAP signatures.

The CMap is a collection of gene expression data for cell lines treated with bioactive small molecules paired with pattern matching algorithms that attempt to identify biologically functional connections between drugs and gene expression profiles. Thus, the CMap can be used as a database of drug-induced gene expression signatures. The CMap was designed for identifying candidate drugs for query signatures represented as Affymetrix HG-U133A gene expression microarrays because its underlying drug expression signatures are represented using Affymetrix HG-U133A data. CiDD transforms this underlying probe-based gene expression data to more generic gene-based data so that researchers can use signatures generated from RNA sequencing or other microarrays to identify candidate compounds using the CMap drug experiments. Statistical procedures provided by the CMap for computational screening of Affymetrix HG-U133A query signatures against drug-induced gene expression signatures are implemented in CiDD for generic gene-based query signatures. These procedures are rank-based and built upon Kolmogorov-Smirnov statistical tests. This makes these tests more robust to technology biases and batch effects, which is important in our case because our data were generated from RNA sequencing and not HG-U133A microarrays. Details of CMap methods can be found in Lamb *et al*[58] and a full description of CiDD is provided in Appendix B section 6.3.

3.2 FAP colon and duodenum transcription profiles

3.2.1 Gene expression signatures of at-risk normal mucosa compared to polyps

Before identifying gene expression signatures representing the difference between normal mucosa and polyp samples, we clustered the FAP samples based on their quantified gene expression data reported by Cufflinks (see Figure 20). Generally, the colon and duodenum samples cluster separately illustrating that the tissue specific differences in gene expression between the colon and duodenum are greater than the gene expression differences between polyp and normal mucosa samples. This indicates that the colon and duodenum of FAP patients may need to be targeted with different drugs. However, our goal is to identify a gene expression signature that is shared between the colon and duodenum so that we can identify candidate drug compounds that may be used for chemoprevention in both tissues.

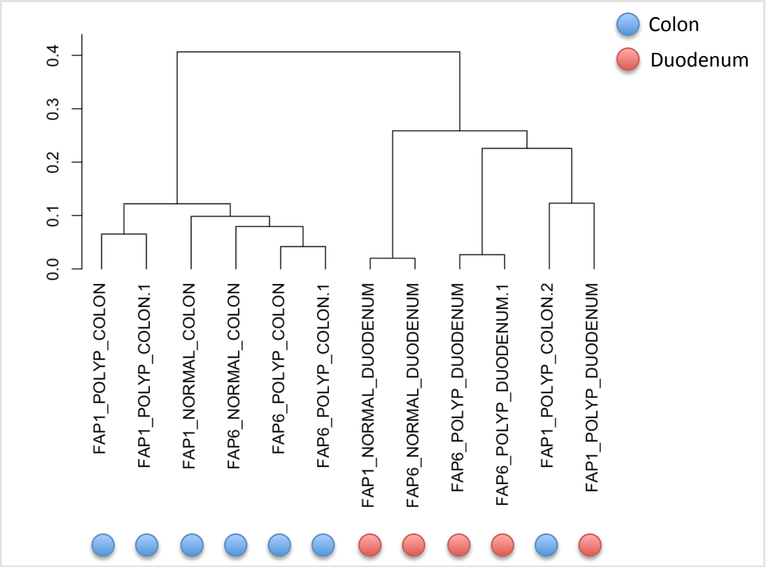


Figure 20: Unsupervised clustering of colon and duodenum samples suggest that the gene expression difference between the colon and duodenum is a stronger signature than the one that differentiates polyp and normal samples.

Our strategy for finding a common gene expression signature for chemoprevention in both the colon and duodenum of FAP patients is to (1) identify a colon or duodenum gene expression signature and then (2) check if that signature is representative of the differences between the at-risk normal mucosa and polyps in both the colon and duodenum in an unsupervised clustering analysis. An alternative strategy could have been to identify a gene expression signature by comparing all polyps versus all normal samples. We believe however that our proposed 2-step approach is more robust and less prone to overfitting to our data set.

Given that we have more *colon* polyps (5) than we have *duodenum* polyps (3), we likely would have more power to detect the true underlying biological gene expression signature in the colon compared to the duodenum in FAP patients. So Cuffdiff was run to identify differentially expressed genes between the at-risk normal mucosa and polyps using colon samples only. Using the criteria described in section 3.1.3, 131 differentially expressed genes were identified (as illustrated by the rows on the heat map of Figure 21). Contrary to Figure 20, the sample clustering in Figure 21 suggests that when limited to the genes of a colon gene expression signature, polyps of the duodenum cluster with those in the colon. This suggests that if we can identify a chemopreventive drug or compound that targets this colon gene expression signature, that compound may also be effective for chemoprevention in the duodenum.

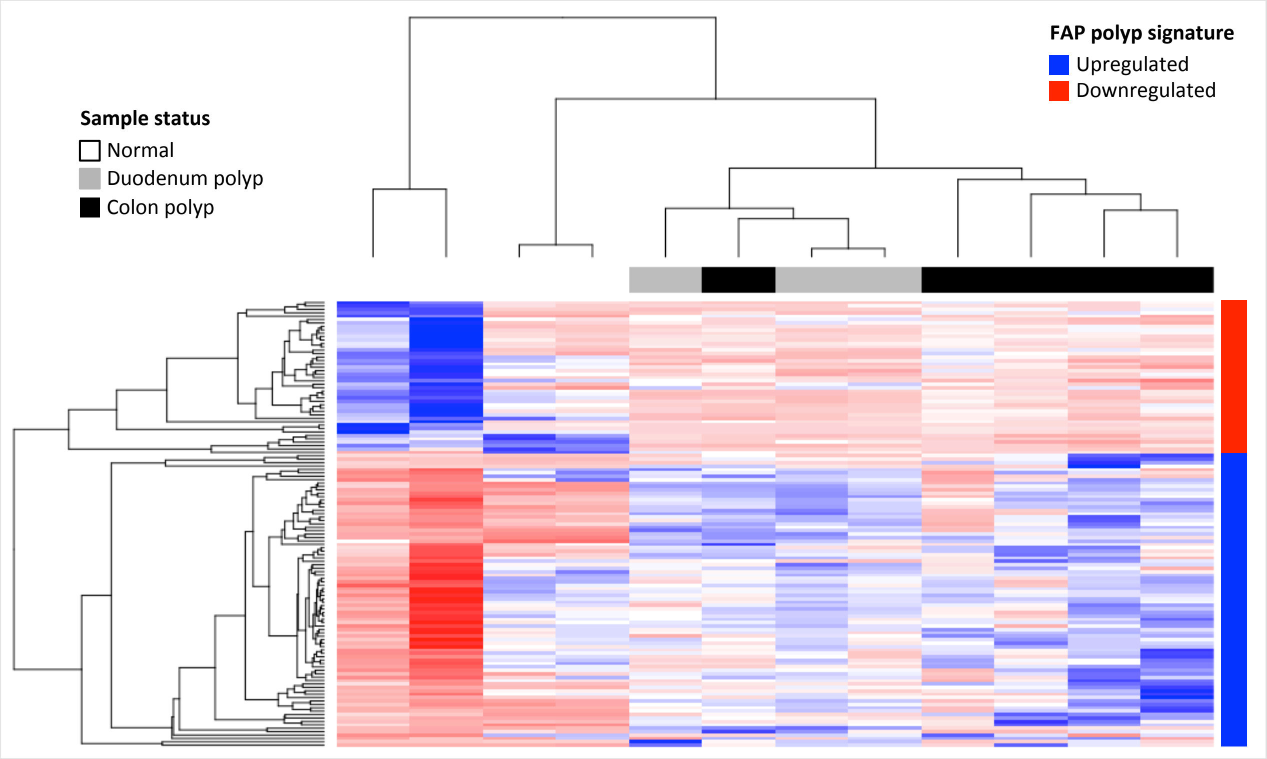


Figure 21: Unsupervised clustering of samples using an FAP colon gene expression signature that characterizes the difference between that at-risk normal mucosa and polyps in the colon of FAP patients. BH adjusted p-value < 0.05 and log2 fold-change > 1. Using this signature, the duodenum polyps cluster with the colon polyps, which suggests that if we can identify a candidate drug to target colon polyps, that drug may also target duodenum polyps.

We then characterized pathways enriched with these signature genes that are deregulated in the colon of FAP patients using *Ingenuity Pathway Analysis* (IPA). IPA identified 37 pathways that are associated with the FAP colon gene expression signature. These pathways, an association p-value and a ratio of the proportion of their member genes that are part of the expression signature are listed in Table 8. As expected, the WNT/beta-catenin pathway is associated with the colon gene expression signature (p-value = 0.021). Furthermore, several inflammatory pathways (labeled with *) were associated with the gene expression signature, which might be expected since COX-2 inhibitors are known to repress adenoma development and these inhibitors target inflammatory pathways[20]. In addition, several RXR activation pathways (labeled with **) are associated with the FAP colon gene expression signature including: (1) *FXR/RXR activation*, (2) *LPS/IL-1 mediated inhibition of RXR function*, (3) *LXR/RXR activation*, (4) *PXR/RXR activation*, and (5) *VDR/RXR activation*. The development of new chemopreventive strategies may benefit by targeting these pathways in addition to or as an alternative to inflammatory pathways.

Pathways	P-value	Ratio
FXR/RXR Activation**	2.75E-07	0.190
LPS/IL-1 Mediated Inhibition of RXR Function**	1.95E-06	0.113
Granulocyte Adhesion and Diapedesis*	1.41E-05	0.120
Agranulocyte Adhesion and Diapedesis*	1.07E-04	0.108
B Cell Development	1.10E-04	0.241
Altered T Cell and B Cell Signaling in Rheumatoid Arthritis	1.15E-04	0.140
Allograft Rejection Signaling	1.32E-04	0.153
Hepatic Cholestasis	1.41E-04	0.113
Cytotoxic T Lymphocyte-mediated Apoptosis of Target Cells*	1.55E-04	0.173
Autoimmune Thyroid Disease Signaling	2.75E-04	0.151
MIF-mediated Glucocorticoid Regulation*	3.31E-04	0.189
Graft-versus-Host Disease Signaling	3.80E-04	0.174
T Helper Cell Differentiation	3.98E-04	0.145
LXR/RXR Activation**	4.68E-04	0.111
Hepatic Fibrosis / Hepatic Stellate Cell Activation	5.37E-04	0.107
Antigen Presentation Pathway*	6.92E-04	0.175
PXR/RXR Activation**	1.07E-03	0.141
Serotonin Degradation	1.23E-03	0.154
Atherosclerosis Signaling	1.35E-03	0.099
Superpathway of Melatonin Degradation	1.38E-03	0.151
OX40 Signaling Pathway	1.58E-03	0.131
Noradrenaline and Adrenaline Degradation	1.74E-03	0.188
VDR/RXR Activation**	4.37E-03	0.115
Antioxidant Action of Vitamin C	5.37E-03	0.102
MIF Regulation of Innate Immunity	6.31E-03	0.133
Sperm Motility	7.08E-03	0.092
Phospholipases	7.24E-03	0.127
Melatonin Degradation I	1.35E-02	0.125
Nur77 Signaling in T Lymphocytes	1.82E-02	0.105
Communication between Innate and Adaptive Immune Cells	1.95E-02	0.086
Wnt/beta-catenin Signaling	2.14E-02	0.077
Dendritic Cell Maturation	2.40E-02	0.068
Ephrin B Signaling	3.09E-02	0.089
IL-4 Signaling	3.09E-02	0.093
Type I Diabetes Mellitus Signaling	3.09E-02	0.080
Calcium-induced T Lymphocyte Apoptosis	3.16E-02	0.098
Eicosanoid Signaling	3.98E-02	0.098

Table 8: Pathways associated with the FAP gene expression signature using *Ingenuity Pathway Analysis* (IPA). Notable pathways include the WNT/beta-catenin signaling pathway and several inflammatory pathways*. In addition, several RXR activation pathways** were identified.

3.2.2 Gene fusions

After identifying and characterizing differentially expressed genes between the at-risk normal mucosa and polyps in FAP patients, Tophat-fusion[63] was run on RNA sequencing reads that failed initial alignment to RefSeq gene transcripts through Cufflinks, and 269 gene fusion candidates were identified. To reduce the false positive rate, we required at least 1 spanning read (a read spanning a fusion breakpoint), 1 spanning pair (a read pair where one read resides on one gene and another resides on another gene, where the pair of reads are flanking a fusion breakpoint) and 5 total pieces of evidence (e.g., the sum of the number of spanning reads and spanning pairs) to call putative gene fusions. The remaining 22 putative gene fusions are listed in Table 9.

Sample	Gene 1	Gene 1 Position	Gene 1 pos	Gene 2	Gene 2 Position	Num Spanning Reads	Num Spanning Pairs	Strands
FAP1_DA1_POLYP_COLON	PRSS3	chr9:33798076	chr9	PRSS1	chr7:142460281	1	7	ff
FAP1_DA1_POLYP_COLON	PARL	chr3:183580484	chr3	ENSG00000217648	chr6:143663891	1	39	rf
FAP1_DA1_POLYP_COLON	GNPNAT1	chr14:53250202	chr14	PMS1	chr2:190687172	1	35	rf
FAP1_DA1_POLYP_COLON	ENSG00000159314	chr17:43511559	chr17	LOC146880	chr17:62777797	1	29	rr
FAP1_DB1_POLYP_DUODENUM	REG3G	chr2:79255058	chr2	REG3A	chr2:79384427	1	6	fr
FAP1_DB1_POLYP_DUODENUM	ENSG00000266613	chr18:8413731	chr18	RFWD2	chr1:176012385	1	8	fr
FAP1_F1_POLYP_COLON	RRN3P2	chr16:29127646	chr16	ENSG00000259807	chr16:29228801	7	6	ff
FAP1_F1_POLYP_COLON	ENSG00000248827	chr5:107061587	chr5	USP7	chr16:9009202	1	26	fr
FAP1_M1_POLYP_COLON	SLC25A11	chr17:4843394	chr17	RNF167	chr17:4843823	1	14	rr
FAP6_B6_POLYP_COLON	PTEN	chr10:89705658	chr10	RPL11	chr1:24021154	1	26	ff
FAP6_D6_POLYP_COLON	CEACAM6	chr19:42266130	chr19	CEACAM5	chr19:42221373	6	196	ff
FAP6_D6_POLYP_COLON	RNF6	chr13:26796139	chr13	FOXO1	chr13:41192773	1	12	rf
FAP6_D6_POLYP_COLON	C11orf80	chr11:66529497	chr11	C1QBP	chr17:5341442	1	48	ff
FAP6_D6_POLYP_COLON	ZNRD1-AS1	chr6:29975965	chr6	HLA-B	chr6:31323943	3	8	rf
FAP6_D6_POLYP_COLON	RPLP0P2	chr11:61404487	chr11	RPLP0	chr12:120637006	2	56	fr
FAP6_D6_POLYP_COLON	ENSG00000225630	chr1:565454	chr1	CLCA1	chr1:86950604	1	5	rf
FAP6_D6_POLYP_COLON	ENSG00000259000	chr14:45334536	chr14	DOCK11	chrX:117707777	2	5	rf
FAP6_D6_POLYP_COLON	LARP4	chr12:50856408	chr12	C15orf41	chr15:36910662	9	5	ff
FAP6_DC6_POLYP_DUODENUM	GRIN2B	chr12:13768031	chr12	C12orf36	chr12:13529226	4	4	rr
FAP6_DC6_POLYP_DUODENUM	ENSG00000232573	chr14:99439637	chr14	RPL3	chr22:39714409	1	14	rf
FAP6_DD6_POLYP_DUODENUM	ENSG00000224879	chr2:79386904	chr2	REG3A	chr2:79386554	1	50	fr
FAP6_DD6_POLYP_DUODENUM	ENSG00000232380	chr13:69560049	chr13	ZDHHC20	chr13:21961731	2	87	fr

Table 9: FAP polyp gene fusions identified with Tophat-fusion.

Interesting fusions include those involving the genes *PTEN* and *REG3A*. In CRC, *PTEN* is altered through mutations, LOH and hypermethylation, where bi-allelic inactivation of the protein is seen in 20-30% of all sporadic cases[64]. These types of *PTEN* events were not observed in our data set; however there was a *PTEN* fusion event identified in the *FAP6_B6_POLYP_COLON* sample suggesting that gene fusions may be another mechanism of

PTEN inactivation. The only gene that was recurrently altered in gene fusion events was *REG3A*. In a previous study, *REG3A* was shown to be down-regulated in 67% (20 of 30 samples) of primary human gastric cancers suggesting that *REG3A* is down-regulated in most primary human gastric cancer cells[65] and may be a relevant gene in the development of duodenum adenomas.

3.3 Candidate chemopreventive drugs for FAP patients

After performing transcriptional profiling and identifying gene expression signatures of the differences between at-risk normal mucosa and polyps in FAP patients, we used the colon gene expression signature to identify candidate chemopreventive drugs using CiDD. Briefly, as described in Appendix B section 6.3.2.2, in the normal workflow of CiDD, a user specifies a tumor characteristic of interest, such as a *BRAF* V600E mutation. CiDD then identifies samples in the TCGA harboring that tumor characteristic (e.g., CRC *BRAF* V600E samples) and a reference set of samples (e.g., CRC *BRAF* wildtype samples). Next, CiDD performs differential expression analyses on automatically downloaded RNA sequence data from these samples and assesses whether a gene expression signature is associated with the tumor characteristic for use in subsequent drug discovery screening experiments. In the case of our FAP project, we have already identified a colon gene expression signature so we can directly use it for drug screening. Thus, only steps 3 and 4 of the CiDD workflow were applied to the FAP colon gene expression signature (see Figure 22).

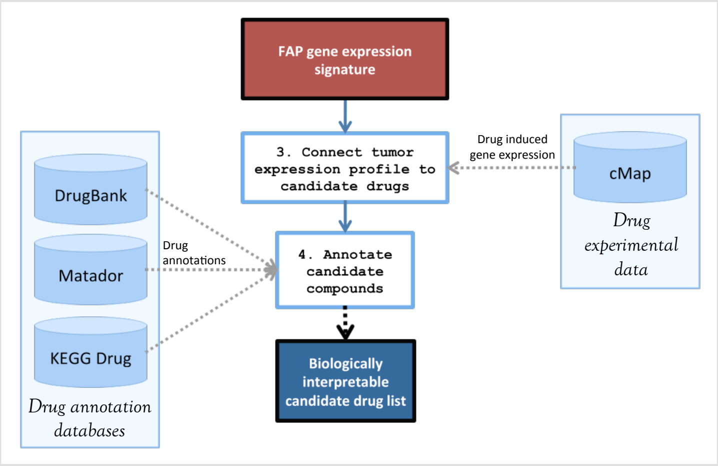
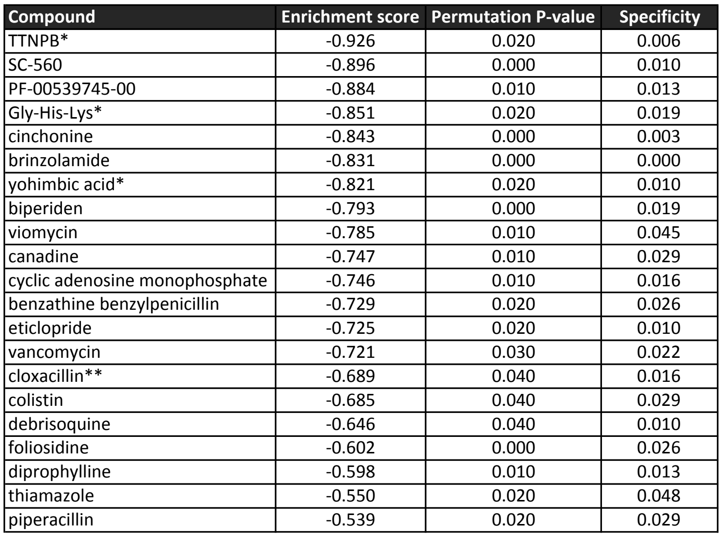

FAP gene expression signature
3. Connect tumor expression profile to candidate drugs
Drug induced gene expression
cMap
Drug experimental data
DrugBank
Matador
KEGG Drug
Drug annotations
4. Annotate candidate compounds
Biologically interpretable candidate drug list
Drug annotation databases

Figure 22: The FAP gene expression signatures were directly input into CiDD, where steps 3 and 4 of the generic workflow were run for the identification of candidate drugs.

Table 10 lists CiDD identified candidate drugs for chemoprevention in the at-risk normal mucosa of FAP patients. Three metrics of the drug screening analyses are depicted on the table:

1. *Enrichment score*: a score in the range of -1 to 1 where -1 is reflective of a drug compound being negatively correlated with the query gene expression signature and 1 representing positive correlation. The score is calculated using an algorithm that accounts for correlation of the query signature with potentially multiple instances of a drug-induced gene expression signature[58].
2. *Permutation P-value*: a measure of significance for the *enrichment score* based on calculating thousands of enrichment scores by randomly sampling enrichment scores for candidate compounds and assessing the significance of the candidate compound *enrichment score*.

3. *Specificity*: a measure of the selectivity of a drug compound for the phenotype of interest. Random query signatures are extracted from MSigDB[47] and run against the CMap to generate a background list of enrichment scores and specificity indicates how often a score equal to or more significant than the enrichment is seen.



Compound	Enrichment score	Permutation P-value	Specificity
TTNPB*	-0.926	0.020	0.006
SC-560	-0.896	0.000	0.010
PF-00539745-00	-0.884	0.010	0.013
Gly-His-Lys*	-0.851	0.020	0.019
cinchonine	-0.843	0.000	0.003
brinzolamide	-0.831	0.000	0.000
yohimbic acid*	-0.821	0.020	0.010
biperiden	-0.793	0.000	0.019
viomycin	-0.785	0.010	0.045
canadine	-0.747	0.010	0.029
cyclic adenosine monophosphate	-0.746	0.010	0.016
benzathine benzylpenicillin	-0.729	0.020	0.026
eticlopride	-0.725	0.020	0.010
vancomycin	-0.721	0.030	0.022
cloxacillin**	-0.689	0.040	0.016
colistin	-0.685	0.040	0.029
debrisoquine	-0.646	0.040	0.010
foliosidine	-0.602	0.000	0.026
diprophylline	-0.598	0.010	0.013
thiamazole	-0.550	0.020	0.048
piperacillin	-0.539	0.020	0.029

Table 10: Candidate drugs identified from the FAP colon gene expression signature that describes the differences between the at-risk normal mucosa and polyps in the colon of FAP patients. The number of asterisks following a compound indicates if the compound was identified as a candidate drug using the combined colon plus duodenum gene expression signature and/or the duodenum-only gene expression signature.

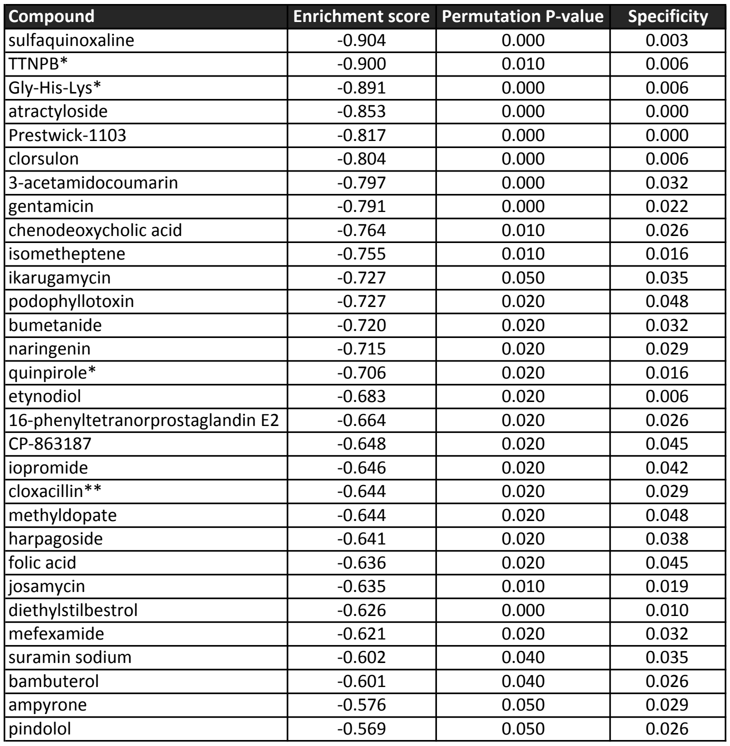
To be identified as a candidate drug in Table 10, we required a permutation p-value ≤ 0.05, an enrichment score < 0 and a specificity ≤ 0.05. These criteria define drug compounds that, at a level of statistical significance, induce gene expression signatures that are negatively correlated with the colon polyp gene expression signature in addition to inducing responses that are highly specific to the colon polyp signature. In the results tables, the drugs are ranked by their enrichment scores, where the most negatively connected drugs are ranked towards the top of the list. Of initial interest, SC-560, a COX inhibitor, is the second ranked drug in the list. This drug has been shown to be effective by inhibiting colon cancer cell proliferation with concomitant G0/G1-phase cell cycle arrest[66]. Drugs of the same class, Celecoxib and Rofecoxib have also shown activity for the prevention of adenomas in clinical trials[23,24], providing some validity to this candidate drug list.

To reinforce confidence in our findings, we generated additional gene expression signatures using different comparison classes of normal mucosa and polyp samples with the thought that drugs that appear on multiple candidate drug lists may have a better chance of being truly effective drugs for chemoprevention in both the colon and duodenum of FAP patients. In Table 11, we identified candidate drugs using a query signature for colon and duodenum samples combined. In Table 12, we identify candidate drugs to repress a gene expression signature for FAP duodenum samples exclusively.

Compound	Enrichment score	Permutation P-value	Specificity
spaglumic acid	-0.983	0.000	0.000
lycorine	-0.833	0.000	0.010
cloxacillin**	-0.779	0.000	0.010
quinpirole*	-0.817	0.000	0.000
yohimbic acid*	-0.905	0.010	0.000
arachidonyltrifluoromethane	-0.939	0.020	0.006
ketoconazole	-0.700	0.020	0.000
celecoxib	-0.652	0.020	0.000
cefotiam	-0.653	0.040	0.026
quipazine	-0.653	0.040	0.016

Table 11: Candidate drugs identified from the combined colon and duodenum gene expression signature. The number of asterisks following a compound indicates if the compound was identified as a candidate drug using the colon-only gene expression signature and/or the duodenum-only gene expression signature.

From these lists, we identified an initial pair of interesting candidate drugs for follow-up. To treat both the colon and duodenum of FAP patients in Table 11, Celecoxib was identified, which is a COX-2 inhibitor that has shown substantial activity in previous studies for repressing the development of colon polyps[23,24]. However, as explained in section 1.1.3, this drug has been associated with cardiovascular side effects and thus is not FDA approved. Nevertheless, identifying this as a candidate drug again provides some validity in the candidate drug results. Another interesting drug is TTNPB, which is the top-ranked compound for chemoprevention in the colon in Table 10 and the second-ranked compound for chemoprevention in the duodenum in Table 12. This compound was also near the top of the ranked candidate compound list (based on an *enrichment score* of -0.688) for the combined colon and duodenum signature in Table 11 although the permutation p-value did not reach statistical significance so TTNPB is not listed in this table. Additionally, TTNPB is an RXR agonist, and the pathways identified as deregulated in FAP polyps were RXR activation pathways (see Table 8), providing additional biological justification for testing this drug compound in follow-up experiments.



Compound	Enrichment score	Permutation P-value	Specificity
sulfaquinoxaline	-0.904	0.000	0.003
TTNPB*	-0.900	0.010	0.006
Gly-His-Lys*	-0.891	0.000	0.006
atractyloside	-0.853	0.000	0.000
Prestwick-1103	-0.817	0.000	0.000
clorsulon	-0.804	0.000	0.006
3-acetamidocoumarin	-0.797	0.000	0.032
gentamicin	-0.791	0.000	0.022
chenodeoxycholic acid	-0.764	0.010	0.026
isometheptene	-0.755	0.010	0.016
ikarugamycin	-0.727	0.050	0.035
podophyllotoxin	-0.727	0.020	0.048
bumetanide	-0.720	0.020	0.032
naringenin	-0.715	0.020	0.029
quinpirole*	-0.706	0.020	0.016
etynodiol	-0.683	0.020	0.006
16-phenyltetranorprostaglandin E2	-0.664	0.020	0.026
CP-863187	-0.648	0.020	0.045
iopromide	-0.646	0.020	0.042
cloxacillin**	-0.644	0.020	0.029
methyldopate	-0.644	0.020	0.048
harpagoside	-0.641	0.020	0.038
folic acid	-0.636	0.020	0.045
josamycin	-0.635	0.010	0.019
diethylstilbestrol	-0.626	0.000	0.010
mefexamide	-0.621	0.020	0.032
suramin sodium	-0.602	0.040	0.035
bambuterol	-0.601	0.040	0.026
ampyrone	-0.576	0.050	0.029
pindolol	-0.569	0.050	0.026

Table 12: Candidate drugs identified from the duodenum gene expression signature. The number of asterisks following a compound indicates if the compound was identified as a candidate drug using the combined colon plus duodenum gene expression signature and/or the colon-only gene expression signature.

3.4 Discussion

We have identified an FAP colon gene expression signature representative of the molecular differences between the at-risk normal mucosa and polyps of FAP patients and screened it against a database of drug-induced signatures using a software framework that we developed called CiDD. We have validated, in silico, the candidate celecoxib, a COX-2 inhibitor that has already been clinically tested as a chemopreventive drug in FAP, which helps support the utility of our approach. CiDD also identified the novel candidate TTNPB, which is an RXR agonist for chemoprevention in both the colon and duodenum of FAP patients.

Sulindac and bexarotene, drugs similar to celecoxib and TTNPB, have been successfully tested on cell lines and are currently being tested on *APC*$^{Min/+}$ mice by Dr. Eduardo Vilar and his lab members at MD Anderson Cancer Center. The *APC*$^{Min/+}$ model is one of the most widely used mouse models of FAP. These mice harbor a heterozygous L850X nonsense mutation in *APC*. The protocols, breeding of mice and laboratory work for the testing of these drugs have been created and managed by Dr. Eduardo Vilar and his lab members.

4 Conclusions and future directions

The long-term goal of the project described in this dissertation is to define the genomic landscape of FAP polyps, to determine their biological significance and to use this information to develop novel chemopreventive strategies for FAP patients. Of note, although hereditary forms of CRC constitute less than 5% of all cases, their study has tremendously informed the understanding of the molecular biology of CRC in general. This is highlighted by the current recommendation to use aspirin for the prevention of sporadic CRC and the approval of COX-2 inhibitors as treatment for polyps in FAP[22]. Thus, our long-term goal and the current findings presented in this dissertation, and the conclusions that follow, have the potential to impact the care of not only FAP patients but also the general population. In this chapter, I summarize our conclusions and describe possible future directions of our FAP project, and I conclude by speculating on the additional impact on cancer research that may be made through the new bioinformatics tools that we have developed.

4.1 Promising candidate genes of early CRC development

The basic strategy followed in this dissertation involved the genomic and transcriptomic profiling of FAP polyps, which are benign lesions, and the comparison of these profiles to those of CRC tumors. This strategy allowed us to identify and differentiate the events that may be crucial for the initial development of these pre-cancerous lesions versus those that might be responsible for developing these lesions into carcinomas. This strategy, which leverages TCGA data, can generally be applied to other NGS-based chemoprevention projects for any tissue type that is represented within the TCGA initiative.

In summary, somatic *APC* truncating mutations and loss of chromosome 5q were recurrent across polyps. Driver events such as activating *KRAS* mutations were identified in multiple polyps. Further, analysis of mutation allele fractions suggests that several of the polyps studied are multi-clonal and accumulating additional driver events. Excluding the known genes *APC* and *KRAS,* 50 candidate genes have been identified that could potentially play a role in future chemopreventive drug development projects. Of these genes, notable inhibitors of beta-catenin in the WNT signaling pathway were identified, namely *AXIN2*, *TCF7L2*, *FBXW7* and *SOX9* (in addition to *APC*). *ARID1A,* which is a *MYC* inhibitor that helps to control cellular proliferation was also recurrently mutated in our data set (see Figure 17). The majority of the candidate genes have been previously associated with CRC, providing additional evidence that they are important in the early development of CRC. In addition, a *PTEN* gene fusion was detected and a novel, recurrent *REG3A* fusion was identified in duodenum polyps from 2 patients. These genes are currently being biologically validated with functional studies in the lab of Eduardo Vilar at MD Anderson Cancer Center.

4.2 Next steps in the characterization of FAP adenomas and the development of FAP chemopreventive strategies

We identified a gene expression signature representative of the molecular differences between at-risk normal mucosa and polyps in the colon of FAP patients that was associated with deregulation of inflammatory pathways and RXR activation pathways (Table 8). We screened this signature against drug-induced signatures using our CiDD software. Using a combined gene expression signature representative of the differences between the at-risk normal mucosa and polyps in both the colon and duodenum of FAP patients, CiDD identified Celecoxib, a COX-2 inhibitor that targets inflammatory pathways, which has previously been clinically developed as a chemopreventive drug, thus illustrating the validity of our approach. CiDD also identified the novel chemopreventive candidate drug TTNPB, an RXR agonist, in separate analyses using FAP colon and then duodenum samples. Sulindac and bexarotene, drugs of similar function to celecoxib and TTNPB, are currently being tested on $APC^{min/+}$ mice in the lab of Eduardo Vilar at MD Anderson Cancer Center.

An additional 40 FAP samples have been RNA sequenced recently at the MD Anderson Cancer Center sequencing core facility and these data were not included in the analyses in this dissertation. These samples will be analyzed to perform a more in-depth characterization of the colon and duodenum transcriptomes in the near future. These data will help us to refine the FAP gene expression signatures and allow us to more confidently define the transcriptome differences between the colon and duodenum of FAP patients, which may provide insights into how best to treat the at-risk normal mucosa in both the colon and duodenum of FAP patients.

Additional data types may also prove useful in continuing to refine our genomic analyses of colon polyps and potentially duodenum polyps such as SNP arrays for higher-resolution copy number variant calling or whole genome sequencing for mining DNA in non-coding regions of the genome. Additionally, power to detect mutations in the polyps could be improved by deepening sequence coverage, which would be an especially useful strategy for mutation calling in low purity settings. Deeper coverage would also help in characterizing polyp clonality. Tools for characterizing clonality are dependent on accurate somatic mutation calling and the precise characterization of allele fractions for those mutations, both of which are more accurate with deeper sequencing.

To overcome problems of limited DNA in polyps, we could perform mutation calling and transcriptome characterization from the same RNA sequence reads of polyps. It is possible to detect somatic mutations in the RNA sequence data, which would allow us to characterize mutations in genes that are transcribed. Methods such as SNPiR exist for calling variants in RNA sequence data. SNPiR has shown 98% specificity and 70% sensitivity of calling coding variants in RNA sequence data that were verified using exome and whole genome sequencing[67].

Other chemoprevention clues may be hidden in the genomes of normal mucosa samples. So another focus of future work is to characterize aberrations found in the colon normal mucosa of FAP patients. This characterization may provide insights into pre- or early-adenoma development that may be very useful in the development of chemopreventive strategies.

4.3 Bioinformatics software developed for NGS-based chemopreventive research

This project required the development of several pipelines and tools. In order to annotate our data set easily and flexibly in combination with other large-scale data sets such as the 1000 Genomes Project, the Exome Sequencing Project, the COSMIC database and several others, we developed *variant tools*, which simplified the management and characterization of samples and their mutations tremendously. We also characterized chromosomal *allelic imbalance* (AI) in the pre-cancerous setting, where low "tumor" purities can make this data more difficult to analyze compared to that of tumors, using software that we developed for NGS data called *hapLOHseq*. CiDD was developed to computationally identify candidate drugs to target tumor gene expression signatures inferred from RNA sequence data.

These tools could be applied to many settings. *Variant tools* is a generic toolset for the analysis of genetic variants and can be applied to all NGS disease-research studies including cancer and non-cancer related diseases. *hapLOHseq* can be applied to a variety of settings where the detection of subtle chromosomal AI events is helpful. This includes, the early detection of cancer or metastatic disease, the sensitive detection of recurrence, the characterization of cancer evolution temporally and spatially, etc. As a complementary tool, CiDD can be used to identify an initial set of candidate drugs to target specific subtypes of cancer that might be detected. Further, CiDD may be helpful for candidate drug identification for any tumor exome or whole genome study being performed today, even in the absence of RNA sequence data in these studies. CiDD makes this possible because CiDD can obtain RNA sequence data from TCGA as a surrogate for RNA sequence of samples being genomically characterized (by identifying TCGA samples that are genomically similar to those being studied).

5 Appendix A: Sequencing analysis pipelines

In the following sections, I document the tools, versions of tools and commands executed for sequencing analysis pipelines implemented for this project. The pipelines are described using tables where the rows in the table specify an ordered list of minimal commands needed to replicate analyses described in this dissertation. In practice, these commands were implemented to run on a cluster and high-performance servers. Various steps were parallelized on a chromosome level such that for a single sample, 22 jobs (one for each chromosome excluding chromosomes X and Y) would run in parallel in a cluster environment. To simplify the description of the major steps of these pipelines, split/merge commands and intermediate reporting steps that are common to such pipelines, have been omitted.

5.1 Exome sequence alignment

Here we include the minimal commands that could be used to repeat the alignment procedure of our FAP exome sequence data. The *Burrows-Wheeler Aligner* (BWA)[29] is used for initial alignment. *Picard* is used for manipulating and cleaning up *Sequence Alignment/Map* (SAM) format files[68]. The *Genome Analysis Toolkit* (GATK)[30] was used to perform local realignment of sequencing reads. *SamTools*[68] was used for indexing bam files and generating mapping statistics. For high-level quality assessment, SamTools was used to assess the proportion of aligned reads that were aligned on-target (e.g., the number of reads aligned to the exome target region), which is a reflection of the quality of the exome capture process, and *BEDtools*[35] was run to estimate aligned sequencing depth.

Software	Version	Command	Parameters of interest	Comment
bwa	0.5.9-r16	aln		Align reads to the human reference build hg19.
bwa	0.5.9-r16	sampe		Generate sam format alignment for read pairs.
picard	1.95	CleanSam		Soft-clip alignments that hang off the end of reference sequenceand set MAPQ to 0 if a read is unmapped.
samtools	0.1.16	view		Convert from sam to bam format.
picard	1.95	SortSam	SORT_ORDER=coordinate	Sort reads by genomic position.
samtools	0.1.16	index		Create bam file index for fast read access.
gatk	2.6.4	RealignerTargetCreator	--known:dbsnp dbsnp_137.hg19.vcf --known:indels 1000G_biallelic.indels.hg19.vcf	Identify potentially problematic aligned regions around known common polymorphisms and indels.
gatk	2.6.4	IndelRealigner		Realign reads around identified regions.
picard	1.95	MarkDuplicates	REMOVE_DUPLICATES=true	Mark and remove redundant sequencing read pairs.
samtools	0.1.16	index		Re-index the realigned and cleand bam file.
gatk	2.6.4	BaseRecalibrator	-cov ReadGroupCovariate -cov QualityScoreCovariate -cov CycleCovariate -cov ContextCovariate	This is the first pass of the base quality score recalibration, which collects metrics in recalibration tables for the specified covariates used for recalibration.
gatk	2.6.4	PrintReads		Generate bam with recalibrated quality scores using the output BQSR file from the previous command.
bedtools	2.16.1	coverage		Generate coverage statistics.
samtools	0.1.16	idxstats		Get summary of mapped and unmapped reads.

Table 13: A minimal list of ordered commands required for the alignment of exome sequencing reads to the human reference hg19 build.

5.2 Calling point mutations

Given the aligned sequence files produced by our exome sequence alignment pipeline, MuTect[32] is run on polyp-blood sample pairs for the sensitive detection of point mutations. Potential false positive mutations that might be common polymorphisms are identified by cross-checking candidate mutations against population variant databases (including the 1000 Genomes[36] project and the Exome Sequencing Project) and removing those seen in 1% or more of the general population. A custom verification pipeline (described in section 2.1.4) is also run that looks for evidence of variant reads in the paired blood and normal mucosa for each candidate polyp mutation, where mutations are filtered out if variant reads are found in 2% or 5% of reads in the blood or normal mucosa, respectively. Subsequently, a subset of nonsynonymous mutations were visually verified using the *Integrative Genomics Viewer* (IGV)[31]. For each patient, we performed visual verification by inspecting the sequencing reads at each candidate mutation site across all of that patient's samples. We looked for signs of false positives, which include:

1. Mutations appearing to be located only on the ends of reads, which are lower quality base calls.
2. Observing several variant alleles around the candidate mutation, which suggests that there may be an indel in the surrounding area, resulting in poorly mapped reads and the generation of false positives.

We found that, after running our verification pipeline (described in section 2.1.4), there were very few mutations that failed visual verification.

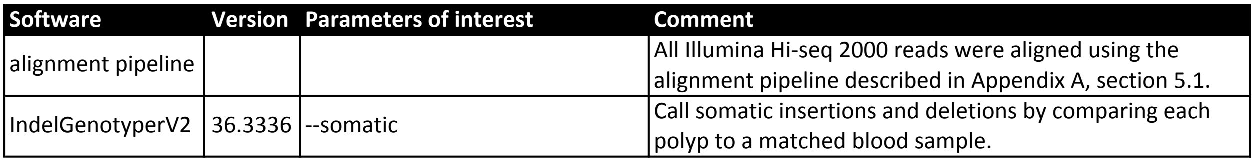
Software	Version	Parameters of interest	Comment
alignment pipeline			All Illumina Hi-seq 2000 reads were aligned using the alignment pipeline described in Appendix A, section 5.1.
mutect	1.4	--reference_sequence ucsc.hg19.fasta --cosmic hg19_cosmic_v54_120711.vcf --dbsnp dbsnp_137.hg19.vcf	Call somatic point mutations by comparing each polyp to a matched blood sample. Information in COSMIC and dbSNP is used to distinguish true somatic events from false positives.

Table 14: MuTect calls point mutations for polyp (or tumor) samples using paired sequence alignment files.

Subsequently, mutation reports were generated using variant tools[33], where we annotated our mutations with information from COSMIC[43], dbNSFP[41], the 1000 Genomes Project[36] and the Exome Sequencing Project.

5.3 Calling insertions and deletions

Similarly to point mutation calling, somatic *insertions and deletions* (indels) were detected using paired polyp and blood samples. *IndelLocator* (i.e., IndelGenotyperV2) was run to call the indels. The same verification pipeline (described in section 2.1.4) was run to identify false positives by searching for evidence of variant reads in each polyp's paired blood and normal mucosa sample. Unlike with point mutations, where we visually verified only a subset of the data (because there were very few false positives identified through their visual verification), all indels were visually verified using IGV because of their higher false-positive rate.



Software	Version	Parameters of interest	Comment
alignment pipeline			All Illumina Hi-seq 2000 reads were aligned using the alignment pipeline described in Appendix A, section 5.1.
IndelGenotyperV2	36.3336	--somatic	Call somatic insertions and deletions by comparing each polyp to a matched blood sample.

Table 15: IndelLocator calls insertions and deletions for polyp (or tumor) samples using paired sequence alignment files.

5.4 Chromosomal allelic imbalances

Chromosomal *allelic imbalances* (AI) are called using *hapLOHseq* (see Appendix B, section 6.2). *hapLOHseq* identifies regions of the genome where there is an excess of one haplotype (i.e., allelic imbalance). To do this, first germline haplotypes need to be statistically estimated (i.e., genotypes need to be *phased*). Then *hapLOHseq* is run to look for segments of the genome where allele frequencies that are higher than expected (e.g., greater than 0.5) are enriched for the alleles in an estimated germline haplotype. These are the candidate AI regions.

First, the GATK is used to call genotypes for blood samples at sites that are polymorphic in the 1000 Genomes project. Then, *MaCH*[69], a Markov Chain based haplotyper, is used to statistically estimate haplotypes (using a reference panel of 200 European haplotypes). Of note, *hapLOHseq* also includes a phasing algorithm that we developed called *pairwise-phasing* (see Appendix B, section 6.2.2.1.2), where the main benefit of pairwise-phasing is that one can phase a *variant call format* (VCF) file directly without the need of processing sequencing read files. Finally, *hapLOHseq* is run using each polyp's VCF file and the corresponding estimated germline haplotypes. To classify each putative *hapLOHseq* event as somatic, we verified that the events did not exist in the blood samples by running *hapLOHseq* on the blood and normal mucosa samples as well.

Software	Version	Command	Parameters of interest	Comment
alignment pipeline				All Illumina Hi-seq 2000 reads were aligned using the alignment pipeline described in Appendix A, section 5.1.
gatk	2.6.4	UnifiedGenotyper	-gt_mode GENOTYPE_GIVEN_ALLELES -out_mode EMIT_ALL_SITES -stand_call_conf 0.0 --annotation AlleleBalance --annotation DepthPerAlleleBySample --annotation Coverage --annotation AlleleBalanceBySample	Call genotypes (within the exome target region) at the sites that are polymorphic in the 1000 genomes project. The positions for the 1000 genomes SNP sites were obtained by downloading the EUR reference panel data from http://www.sph.umich.edu/csg/abecasis/MACH/download/ and then intersecting these coordinates with the target region of the Nimblegen SeqCap EZ3 capture chip.
mach	1.0.18		--rounds 30 --states 50 --phase -h EUR.200.haplotypes	Custom scripts were run to transform the VCF generated by the GATK into PED format for MaCH. A reference panel of 200 European haplotypes (-h) were downloaded from and http://www.sph.umich.edu/csg/abecasis/MACH/download/ were used as input into the phasing.
haplohseq	0.1		--est_aberrant_emissions --num_states 2 --initial_param_normal 0.5 --initial_param_event 0.51 --event_prevalence 0.001 --event_length 50 --vcf_min_depth 10	Identify allelic imbalance events using the estimated haplotypes and the polyp VCF files. *hapLOHseq* is described in Appendix B, section 6.2.

Table 16: Minimal commands executed for the calling genotypes, estimating haplotypes and then detecting allelic imbalance events from exome sequence data.

5.5 Quantifying transcripts and identifying differentially expressed genes

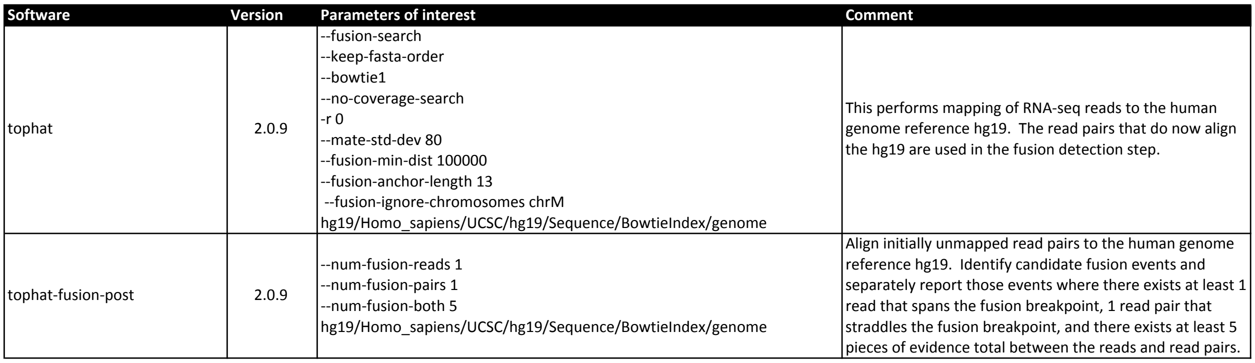
The *Tuxedo* protocol[59] was implemented to perform RNA sequence transcript quantification and for the analysis of differential gene expression. *TopHat*[60] is used for alignment of initial reads to the human genome reference hg19. *Cufflinks*[61] then assembles and quantifies isoform and gene-level transcripts. Subsequently, the software package *Cuffdiff*[62] is used to identify differentially expressed genes, where genes with an adjusted Benjamini-Hochberg p-values less than 0.05 and a log2 fold-change >= 1 or log2 fold-change <= -1 are labeled as differentially expressed. *CummerBund* (not shown below) is an R package that is part of the Tuxedo protocol, designed for the interrogation of *CuffDiff* results, which we used to explore the expression data.

Software	Version	Parameters of interest	Comment
tophat	2.0.9	--fusion-search --keep-fasta-order --bowtie1 --no-coverage-search -r 0 --mate-std-dev 80 --fusion-min-dist 100000 --fusion-anchor-length 13 --fusion-ignore-chromosomes chrM hg19/Homo_sapiens/UCSC/hg19/Sequence/BowtieIndex/genome	This performs mapping of RNA-seq reads to the human genome reference hg19.
cufflinks	2.1.1		This assembles transcripts and quatifies isoform-level and gene-level expression using FPKM.
cuffmerge	1.0.0	-g hg19/Homo_sapiens/UCSC/hg19/Annotation/Genes/genes.gtf -s hg19/Homo_sapiens/UCSC/hg19/Sequence/Chromosomes	The hg19 gene definition file (-g) and the reference sequence (-s) are resources that were downloaded from the cufflinks website (http://cufflinks.cbcb.umd.edu/igenomes.html).
cuffdiff	2.1.1	-b hg19/Homo_sapiens/UCSC/hg19/Sequence/Chromosomes	The reference fasta for bias correction (-b) is the same parameter value used in the previous cuffmerge step (-s).

Table 17: Minimal ordering of commands to quantify transcripts and identify differentially expressed genes for a single 2-class comparison.

5.6 Detecting gene fusions

RNA sequence reads were aligned using *TopHat* to the human genome reference hg19. The remaining unmapped read pairs were used as input to *Tophat Fusion*[63]. Tophat Fusion identifies putative gene fusions. We filtered these gene fusions and required putative fusion events to have at least 1 read mapped to a fusion breakpoint, 1 read pair with reads flanking the fusion breakpoint and a total of 5 (reads plus read pairs) providing such pieces of evidence for candidate fusion events.



Software	Version	Parameters of interest	Comment
tophat	2.0.9	--fusion-search --keep-fasta-order --bowtie1 --no-coverage-search -r 0 --mate-std-dev 80 --fusion-min-dist 100000 --fusion-anchor-length 13 --fusion-ignore-chromosomes chrM hg19/Homo_sapiens/UCSC/hg19/Sequence/BowtieIndex/genome	This performs mapping of RNA-seq reads to the human genome reference hg19. The read pairs that do now align the hg19 are used in the fusion detection step.
tophat-fusion-post	2.0.9	--num-fusion-reads 1 --num-fusion-pairs 1 --num-fusion-both 5 hg19/Homo_sapiens/UCSC/hg19/Sequence/BowtieIndex/genome	Align initially unmapped read pairs to the human genome reference hg19. Identify candidate fusion events and separately report those events where there exists at least 1 read that spans the fusion breakpoint, 1 read pair that straddles the fusion breakpoint, and there exists at least 5 pieces of evidence total between the reads and read pairs.

Table 18: Minimal commands to generate a list of candidate gene fusions from RNA sequence data.

6 Appendix B: Bioinformatics software developed and applied in this project

6.1 NGS variant management, annotation and analysis: `vtools`

The contents of this chapter are based on the following article, reprinted with permission, from the journal Bioinformatics:

San Lucas, F. A., Wang, G., Scheet, P. & Peng, B. *Integrated annotation and analysis of genetic variants from next-generation sequencing studies with variant tools*. Bioinformatics 28, 421–422 (2011).

6.1.1 Introduction

Tracking samples and predicted variants from *next-generation sequencing* (NGS) projects often requires building custom analysis pipelines. Data standards such as the *BED*[70] and *VCF*[71] file specifications can be used to represent these variants in a common format, simplifying integration of tools and the construction of these analysis pipelines. Difficulties include the integration of diverse annotation sources and the management of many large intermediate files containing millions of predicted variants and millions more associated annotations for each sample. These annotation sources and intermediate files often have fundamental inconsistencies using either 0- or 1-based coordinates and potentially different genomic builds, which can complicate their management and integration.

For biologists or analysts who have familiarity with programming and running tools from the command line, there are many useful tools that can be integrated into custom pipelines to annotate and filter variants. These tools include *ANNOVAR*[34] and *BEDTools*[35]. However, building effective pipelines that relate variants to their samples and sample attributes (such as cases and controls), while applying multiple annotation sources requires a large customization effort. A framework for building pipelines that facilitate simple, reproducible and recurrent analyses is currently lacking. Therefore, we have developed *variant tools*, a flexible, open-source toolset upon which custom pipelines can be easily constructed. This toolset facilitates the storage of variants (alongside their sample details) as well as the annotation, filtering and reporting of these variants at multiple levels – starting with variant reports based on individual samples to project-wide variant reports.

6.1.2 Methods

Python scripting language and it can incorporate either SQLite or MySQL as the backend database engine. The toolset is designed around a master variant table that often consists of millions of variants for all of the samples in a sequencing project along with variant attributes (called *fields* in *variant tools*). Variant fields can include sample statistics, which *variant tools* can generate, or information provided by annotation data sources. Regardless of the source of these fields, they can be used to select, output and analyze genetic variation from the project. As illustrated in Figure 5, analyzing genetic variants from NGS projects typically involves four steps, namely importing, annotating, filtering, and reporting:

Sample and variant import: *variant tools* accommodates a variety of variant file formats. It supports import of VCF files or other tab-delimited formats such as intermediate output from ANNOVAR or BEDTools. It is capable of annotating and reporting on all types of variants, including indels, as long as annotation sources are available. The toolset also supports annotation and reporting of project variants using multiple genomic builds, by automatically downloading and integrating the UCSC *liftOver* tool[70]. As an example, if variants are imported to a project using build hg18, they can be annotated using annotation sources designed for build hg19, and exported based on either hg18 or hg19 coordinates.

Annotation: *variant tools* can incorporate databases that annotate individual variants, genomic locations and regions, such as genes, and other annotation fields. A growing number of annotation sources such as *dbNSFP*[41] or *KEGG pathways*[72] can be downloaded automatically by *variant tools* whereas customized annotation databases could be created following a well-documented procedure. Any genomic data source can be imported into the database as long as project variants can be linked to the annotation source through an annotation attribute (such as genomic coordinates or a gene name).

Select variants of interest: Variants can be selected by read depth (if provided by the imported VCFs), by any of the available annotation fields (such as variant type, gene, pathway or predicted damaging effects) or by sample frequency across subsets of samples, which is useful for comparing variants across populations such as cases and controls. Complex criteria involving multiple fields from different annotation sources can be used to select or filter variants. Selected variants can be counted, exported, or saved to separate underlying tables, where they can be annotated and filtered separately from other variants.

Export reports: Variants from a table can be exported with an arbitrary number of fields, regardless of their sources. This allows users to output sample statistics such as numbers of homozygous and heterozygous genotypes in samples for selected variants alongside annotation information. More interestingly, arithmetic operations and aggregate functions can be used to output summary statistics of variants such as the average depth of coverage for a particular set of variants.

variant tools installs easily and sets up a working environment with human genome annotation sources that can be downloaded automatically. Because *variant tools* manages project variants and annotation sources for the user, it is easier to reanalyze variants as genomic builds change and as annotation sources are updated or become available. The burden of tracking VCF files, annotation files and numerous scripts is reduced. *variant tools* is freely available at http://varianttools.sourceforge.net. This website includes source code, documentation, tutorials and a description of available public annotation data sets.

6.1.3 Discussion

Despite an intuitive command-line interface, some high-level reports, such as calculating sample transition/transversion ratios or reporting the number of variants per gene, involve several vtools commands. To simplify the use of *variant tools*, we provide a reporting command vtools_report that generates example summary reports. These reports make the use of *variant tools* more practical, and the vtools_report source code provides examples of how to combine and further customize vtools commands.

Within *variant tools*, variants are linked to but stored separately from their annotations within a relational database removing the need to store large, repetitive, intermediate annotation files, which helps to conserve disk space. To improve query performance in annotation and filtering, database indexes are automatically created for the users. These indexes do add to the storage needs of variant tools.

For an example, we created a vtools project with 44 whole genome VCF files with 161 million predicted sample variants. This required 3.3G of disk space to store the variants within an SQLite database compared to 2G of disk space for the VCF files compressed or 9G uncompressed. As an added benefit of the vtools approach, these variants were stored using both hg18 and hg19 genomic coordinates within SQLite. When using a MacPro workstation with 2 Quad-Core Intel Xeon Processors at 2.26GHz and 8G of RAM, the project creation required 3.5 hours. This time can be reduced to an hour if variants are processed in parallel by vtools on a cluster system before they are merged to a larger project. The time required for subsequent annotation and filtering of these variants ranged from 1 to 10 minutes. Additional details and other examples can be found in the tutorials section of the software website.

We have provided a pre-configured but customizable framework for the analysis of predicted variants from NGS data. Although our efforts were motivated by a desire to produce initial, non-statistical analyses, we are currently expanding our software to include a suite of powerful tests for association studies. Our general framework will allow the direct comparison and implementation of a wide array of analytical methods.

6.2 Detection of allelic imbalance events: hapLOHseq

6.2.1 Introduction

A well-studied mechanism by which cancer cells alter the activity of tumor suppressor genes and oncogenes is through *copy number alteration* (CNA) events. One such aberration class that we focus on in this project is chromosomal *allelic imbalance* (AI). We define chromosomal AI as genomic aberrations of greater than 10 megabases due to amplification, deletion or *copy neutral loss-of-heterozygosity* (cn-LOH) events. The detection and characterization of these chromosomal events has many potential applications in cancer studies. For example, characterizing tumor samples with specific chromosomal amplifications and deletions can be used to inform therapeutic decision-making as these events provide insights into the progression of aberrant cells to cancer and even to metastasis[73]. Further, the *sensitive* detection of specific AI biomarkers can be used for the early detection of cancer and for the management of cancer resistance[74].

The traditional strategies for identifying CNA or AI events employ cytogenetic technologies, such as karyotyping and *fluorescence in situ hybridization* (FISH). In the last decade, *array-based comparative hybridization* (aCGH) and *single-nucleotide polymorphism* (SNP) array based approaches have become popular technologies for CNA detection[75]. One drawback of these methods, however, is that their probes are sparsely located along the genome and they are pre-defined, making it challenging to pinpoint event boundaries and detect novel and rare AI events. More recently, methods based on whole genome sequencing data have gained in popularity, due to their higher-resolution, more precise detection of CNA boundaries, and ability to identify novel CNAs[76]. In addition, some tools exist for the detection of CNA from exome sequencing and are designed to address some of the issues inherent in this data. Examples of these tools include *ExomeDepth*[77] and *ExomeCNV*[78]. ExomeDepth models the relative coverage between a sample of interest and its expected coverage based on a statistical model generated from a reference panel of exome sequenced samples. The ExomeDepth statistical model is designed to account for the capture bias in coverage that is common to exome sequence data. ExomeCNV takes a similar approach, but also incorporates allele frequencies at heterozygous sites for detection of cn-LOH events. These methods are designed to identify exon and gene-level copy number events from either paired samples or from a sample and a reference panel of exomes.

No method exists however for detecting *subtle* chromosomal AI events in exome sequence data which could be extremely valuable in cancer studies, especially where limited tissue availability exists and renders surveys with other technologies (e.g., SNP arrays) impossible. Subtle AI events are those amplification, deletion or cn-LOH events that exist in a small proportion of the cells sequenced, potentially in 20% or less of the cells. *hapLOHseq* is a software tool that we developed especially for the sensitive detection of chromosomal AI events from *next-generation sequencing* (NGS) data, and more specifically from exome sequencing data for the project described in this dissertation. *hapLOHseq* is a NGS-based adaptation of a method called *hapLOH*[51], which was designed for the subtle detection of AI events inferred from SNP array data. Inputs to *hapLOHseq* include *variant call format* (VCF) files generated from either whole-genome or exome sequencing and optionally, statistically phased germline haplotypes for these samples. The output for each *hapLOHseq* run is a report of putative AI regions of the genome along with a detailed report that includes the probability of each polymorphic heterozygous site being in a region of AI.

hapLOHseq relies on AI, or the deviation from the expected one-to-one allele ratio at heterozygous sites in germline DNA. For example, consider a heterozygous site with arbitrarily labeled alleles of A and B. A duplication event over the site results in either an AAB or ABB genotype with a corresponding imbalanced allele ratio of 2:1 or 1:2. A deletion event results in an A- or B- genotype with severely imbalanced allele ratios of 1:0 or 0:1. Similarly, cn-LOH events result in AA or BB genotypes with allele ratios of 2:0 or 0:2. With high-purity tumor samples and characterization of germline genotypes from paired normal samples, one can directly compare genotypes of the tumor and normal samples and clearly characterize the tumor genome and infer copy number changes using existing methods such as *ExomeCNV*[78]. However, when the sample has low tumor purity and contains a high-proportion of normal cells and a small proportion of tumor cells, the called genotypes will reflect those of the germline, not the tumor. Thus, to characterize the tumor we must make inferences of aberrations using subtle signals of AI inferred from the lower-level allele read counts.

6.2.2 Methods

The *hapLOHseq* method works by capturing subtle AI signals whenever there exists imbalances in haplotypes rather than simply relying on imbalances observed at independent, heterozygous sites. The method consists of 3 general steps, which include the following:

1. Estimate germline haplotypes using statistical, population genetics, from the called genotypes. These haplotypes can be estimated from a germline sample or alternatively, directly from a tumor sample in situations where the tumor purity is very low (e.g., the proportion of normal cells is very high).
2. Assess similarity between the observed reference allele frequencies from the sequencing reads and the estimated haplotypes.
3. Identify AI regions where this similarity is higher than expected indicating haplotype imbalance.

These steps are described in further detail in the following sections.

6.2.2.1 Estimation of germline haplotypes

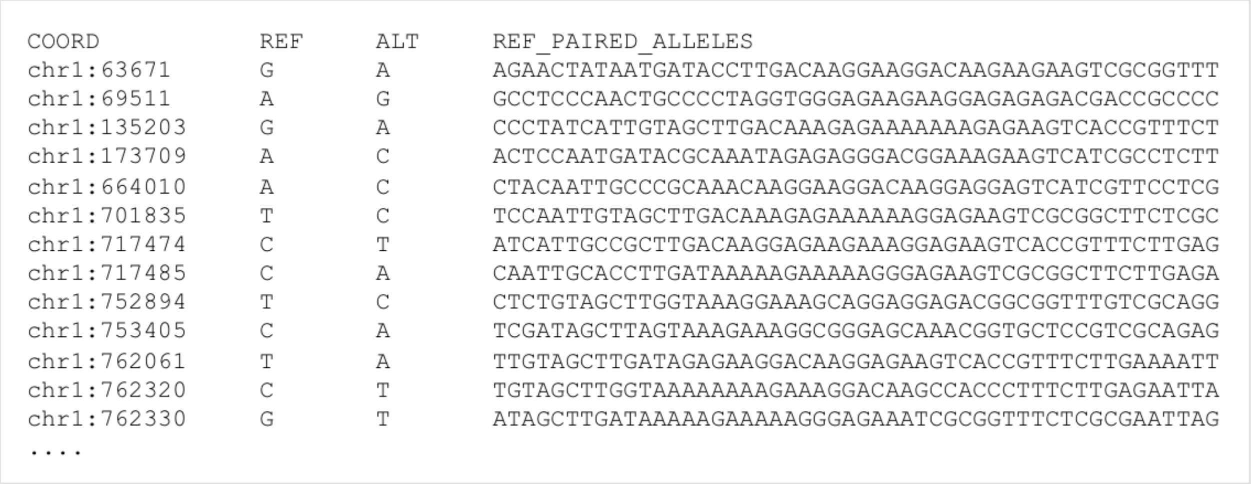
6.2.2.1.1 Existing statistical software

Several statistical software packages can be used to estimate haplotypes for use in *hapLOHseq* such as *MaCH*[69], *Beagle*[79] or *fastPHASE*[80]. In this project, we used MaCH as described in Appendix A, section 5.4. In order to properly estimate haplotypes using this strategy, it is necessary to have and process *binary alignment map* (BAM) files. The reason for this is that from variant call format (VCF) files, which is usually the end product for a sequenced sample, one cannot distinguish between missing genotypes and homozygous reference sites (because VCFs typically do not report homozygous reference genotypes). Going back to reads in a BAM file to make homozygous reference calls can be resource intensive, requiring large amounts of disk space for storage and a high-number of processors often in a distributed computing environment for analysis. In addition, users typically have VCF files of sequence variant calls but they may not have access to the low-level read data (i.e., BAM files).

To address this, we have developed a computationally efficient method of phasing, called *pairwise phasing*, that is embedded in *hapLOHseq* and runs on VCF files (without the need for BAM files). We describe the pairwise phasing method and compare its phasing accuracy to MaCH's phasing accuracy in the subsequent sections.

6.2.2.1.2 Pairwise phasing

The basic idea behind *pairwise phasing* is to estimate haplotypes using genotypes provided in a VCF file in an iterative, pairwise fashion. Given all of the heterozygous sites in a VCF file that are polymorphic in the *1000 Genomes* (1KG) *Project* (we call these sites *informative heterozygous sites*), the pairwise phasing algorithm walks across these sites and iteratively appends an allele from each heterozygous site that is more likely to be phased, or *paired*, with an allele of the current informative heterozygous site based on pre-computed pairwise *linkage disequilibrium* (LD) calculations. These paired alleles are provided with *hapLOHseq* in a *pairwise phase map*.



```
COORD           REF     ALT     REF_PAIRED_ALLELES
chr1:63671      G       A       AGAACTATAATGATACCTTGACAAGGAAGGACAAGAAGAAGTCGCGGTTT
chr1:69511      A       G       GCCTCCCAACTGCCCCTAGGTGGGAGAAGAAGGAGAGAGACGACCGCCCC
chr1:135203     G       A       CCCTATCATTGTAGCTTGACAAAGAGAAAAAAAGAGAAGTCACCGTTTCT
chr1:173709     A       C       ACTCCAATGATACGCAAATAGAGAGGGACGGAAAGAAGTCATCGCCTCTT
chr1:664010     A       C       CTACAATTGCCCGCAAACAAGGAAGGACAAGGAGGAGTCATCGTTCCTCG
chr1:701835     T       C       TCCAATTGTAGCTTGACAAAGAGAAAAAAGGAGAAGTCGCGGCTTCTCGC
chr1:717474     C       T       ATCATTGCCGCTTGACAAGGAGAAGAAAGGAGAAGTCACCGTTTCTTGAG
chr1:717485     C       A       CAATTGCACCTTGATAAAAAGAAAAAGGGAGAAGTCGCGGCTTCTTGAGA
chr1:752894     T       C       CTCTGTAGCTTGGTAAAGGAAAGCAGGAGGAGACGGCGGTTTGTCGCAGG
chr1:753405     C       A       TCGATAGCTTAGTAAAGAAAGGCGGGAGCAAACGGTGCTCCGTCGCAGAG
chr1:762061     T       A       TTGTAGCTTGATAGAGAAGGACAAGGAGAAGTCACCGTTTCTTGAAAATT
chr1:762320     C       T       TGTAGCTTGGTAAAAAAAAGAAAGGACAAGCCACCCTTTCTTGAGAATTA
chr1:762330     G       T       ATAGCTTGATAAAAAGAAAAAGGGAGAAATCGCGGTTTCTCGCGAATTAG
....
```

Figure 23: A snippet of a *pairwise phase map*. Each row corresponds to a 1KG polymorphic site and is defined with a genomic coordinate and reference and alternate alleles. The ref-paired alleles specify the alleles at subsequent polymorphic sites that are more likely to be paired with the reference allele in the current row based on pre-computed LD values.

The pairwise phase map contains a row for each polymorphic site in the 1KG data set (see Figure 23). Each polymorphic site is defined by a coordinate and a reference and alternate allele. The *ref-paired alleles* are strings of alleles at subsequent polymorphic sites that are more likely to be paired with the reference allele for the genotype in the current row in the map based on pre-computed LD values. The length of the ref-paired alleles character string is referred to as the *depth* of the pairwise phase map. An *exhaustive* pairwise phase map would contain ref-paired alleles for all subsequent polymorphic sites on the current chromosome. This would result in a very large pairwise phase map and an inefficient phasing algorithm because at some genetic distance LD becomes negligible and accounting for these data becomes detrimental to the performance of the phasing algorithm. Optimally, the pairwise phase map would be deep enough to pair up the vast majority of adjacent informative heterozygous sites but the map would not be unnecessarily deep resulting in an inefficient algorithm. Thus, *depth* can be thought of as a tuning parameter of our pairwise phasing algorithm.

To estimate an optimal depth for the pairwise phase map, we generated a histogram of the numbers of polymorphic markers in between each pair of adjacent *informative het sites* in a 1KG sample, inferred from a whole genome VCF file and an exome-only-version of the VCF file (see Figure 24). The distributions in the 2 plots are similar, where the vast majority of adjacent informative het pairs have less than 50 polymorphic sites between them. This suggests that a pairwise phase map depth of 50 may be optimal for capturing most of the pairwise information within these data types without adding inefficiency to the pairwise phasing process.

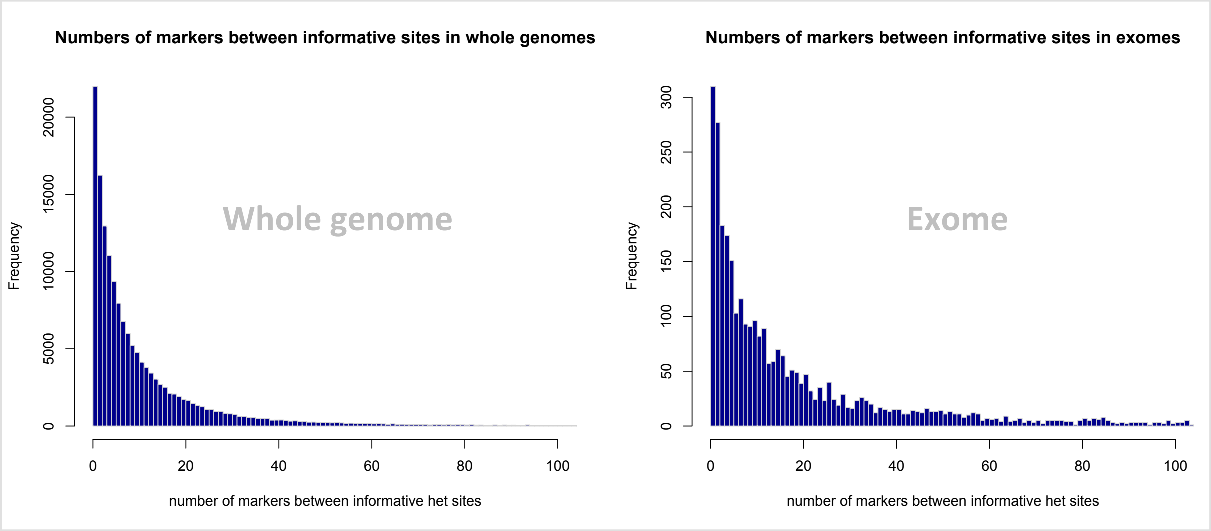


Figure 24: Distribution of the number of 1KG polymorphic sites that are in between adjacent informative heterozygous sites in a TCGA germline sample (TCGA-19-2620).

Here I describe the algorithm for determining the ref-paired alleles. Given *m* polymorphic sites in a *reference panel* (e.g., a set of 1KG haplotypes), and a depth of *d*, for each polymorphic site, we assess its LD with *d* subsequent sites. Let *i* and *j* (*i* = 1,…,*m*; j = i+1,…,i+*d*) refer to 2 polymorphic sites in the reference panel, where *i* is the *current* heterozygous site and *j* is a *paired* heterozygous site. We determine which allele at each site *j* is phased with the reference allele at the current site *i*, by computing D_{RR} and D_{RA}. D_{RR} is a measure of LD between the reference allele at site *i* and the reference allele at site *j* and represents how often reference alleles at sites *i* and *j* co-occur on the same haplotypes in the reference panel relative to the expectation based on allele frequencies. Similarly, D_{RA} is a measure of the LD between the reference allele at site *i* and the alternate allele at site *j*. Here are the simple 6 lines of pseudo code.

```
$D_{RR} = P_{RiRj} - P_{RiPRj}$
$D_{RA} = P_{RiAj} - P_{RiPAj}$
if $D_{RR} > D_{RA}$ then
    $R_i$ paired with $R_j$
else
    $R_i$ paired with $A_j$
```

This algorithm calculates LD values and determines allele pairings for each polymorphic site with a predetermined number of subsequent polymorphic sites. This number is the *depth* of the pairwise phase map discussed previously. By default, our *hapLOHseq* pairwise phase maps are constructed to a depth of 50. Of note, in cases where this depth is not large enough to accommodate pairs of informative het sites that are genomically spaced with more than 50 polymorphic sites between them, phasing is assigned randomly. A whole genome and an exome version of the pairwise phase maps will be made available with *hapLOHseq*.

Given a pairwise phase map, phasing simply consists of incrementing along each informative het site in a sample of interest, identifying the next informative het site, and determining which allele at the next het site is phased with the reference allele at the current site through a simple lookup in the pairwise phase map. As the algorithm iterates over all informative het pairs, it builds up haplotype estimates one pair at a time. The performance of the phasing algorithm scales linearly with the number of informative het sites that are available in the input VCF file. Of importance, although this was not our original motivation, it takes a few minutes to run pairwise phasing and *hapLOHseq* on hundreds of exome sequenced samples as opposed to days or weeks to run the MaCH pipeline and *hapLOHseq*.

To assess the performance of the pairwise phasing algorithm at varying depths, we generated pairwise phase maps at varying depths and then performed pairwise phasing on a 1000 Genome sample's whole genome VCF file and an exome-only-version of the VCF file. Figure 25 illustrates the phasing performance of our algorithm by comparing our estimated haplotypes to the haplotype published by the 1000 Genomes project. The figure indicates that, as expected, phasing whole genome VCFs provides more accurate estimates of haplotypes as compared to exome VCFs due to the much higher number of data points available in whole genome sequencing data. In addition, the accuracies of the haplotype estimates peak at around a depth of 40 or 50 in the pairwise phase map which further supports our decision to use a pairwise depth of 50 by default in the construction of pairwise phase maps.

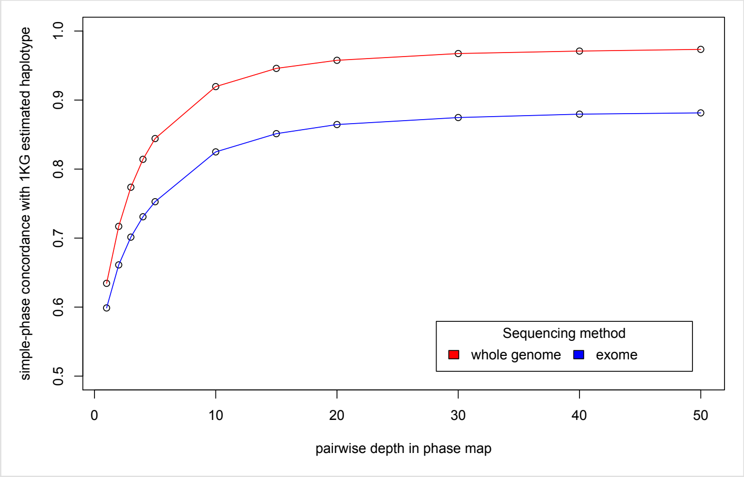


Figure 25: Phase concordance (an estimate haplotype similarity) for whole genome and exome-estimated haplotypes compared to the haplotype published by the 1000 Genomes project for a sample using pairwise phase maps at varying depths. This suggests that there is negligible gain in phasing accuracy at a pairwise depth of 30 or more.

6.2.2.1.3 Performance of MaCH and pairwise phasing

To assess the phasing accuracy of MaCH and pairwise phasing, we took a tumor/normal pair of samples from a TCGA patient (TCGA-19-2620) and phased its germline haplotypes using both strategies on three chromosomes (8, 9 and 10) where there are obvious chromosomal AI events in its paired tumor sample. In these regions, we can easily identify the true overrepresented and underrepresented haplotypes because the allele frequency bands are separate and distinct (see Figure 26). To identify these true haplotypes, at each heterozygous site in these tumor sample chromosomes, we identify the overrepresented allele (i.e., the allele at greater than 0.5 allele frequency) and assign it to the overrepresented haplotype and assign the underrepresented allele (i.e., the allele at a less than 0.5 allele frequency) to the underrepresented haplotype.

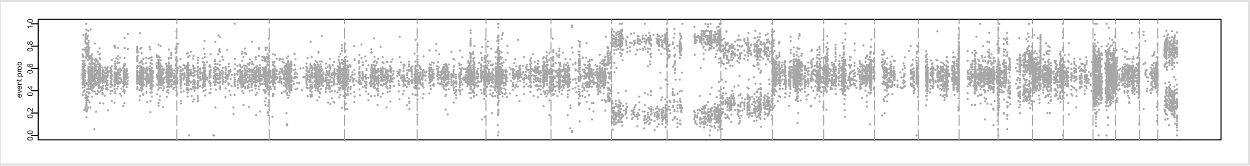


Figure 26: Variant allele frequencies across the exome at heterozygous sites in a tumor sample for a TCGA patient (TCGA-19-2620). A clear separation of allele frequencies, producing 2 bands, can be visually identified across chromosomes 8, 9 and 10.

The two resulting haplotypes define the true germline haplotypes on these chromosomes for this patient. This process was repeated for both a whole genome and an exome VCF for tumor samples of the same patient. We then ran our MaCH phasing pipeline and our pairwise phasing algorithms on the germline samples for this patient on chromosomes 8, 9 and 10. To assess the accuracy of the phasing strategies, we calculated the *switch accuracy*[81], a measure of the similarity between the true haplotypes and those estimated, for each phasing analysis and report them in Table 19. If two haplotypes have no resemblance, we would expect a switch accuracy of around 0.5. If two haplotypes were identical, the switch accuracy would be 1.

Phasing strategy	Sequencing tech	Num markers	Switch accuracy
Pairwise phasing	Exome	669	0.777
MaCH phasing	Exome	770	0.784
Pairwise phasing	Whole genome	103388	0.943
MaCH phasing	Whole genome	104606	0.942

Table 19: Performance of pairwise and MaCH phasing in the exome and whole genome contexts. The switch accuracies are nearly equal comparing pairwise and MaCH phasing.

Table 19 also indicates the number of informative het sits (i.e., *Num markers*) available to each phasing strategy. The MaCH phasing pipeline requires access to the raw sequence files and calling genotypes at sites in the 1000 Genomes reference panel. By processing the BAM files directly as opposed to using only variant sites (which are available in VCF files), this strategy allows for homozygous genotypes being available for phasing and additional heterozygous sites being available for *hapLOHseq* event calling. The MaCH phasing strategy is much more computationally demanding with the potential benefit of better phasing and subsequently more accurate AI event calling. There were 101 additional heterozygous markers in the exome analysis and 1,218 additional heterozygous markers in the whole genome analysis using the MaCH strategy versus pairwise phasing. Switch accuracies however between MaCH and pairwise phasing are almost identical, suggesting that the substantial computational gains and the ease of use of pairwise phasing come at a minimal cost when switch errors are the relevant loss function. For hapLOHseq and hapLOH this is indeed the case. We assess the improvement of *hapLOHseq* event calling using MaCH versus pairwise phasing (due to the increased number of informative markers associated with MaCH phasing) in section 6.2.3.

6.2.2.2 Phase concordance with frequency-based phasing

After statistical estimation of germline haplotypes, we assessed whether or not these haplotypes were in allelic imbalance in regions of the genome where there appear to exist an *excess haplotype* and an underrepresented haplotype based on allele frequencies. First, we determine what VCF allele frequencies indicate to be the excess haplotype by applying a threshold at each marker independently in a frequency-based phasing algorithm. The threshold is defined as the median variant allele frequency in the data set. The alleles above the threshold constitute one haplotype and the alleles below the threshold constitute the other. If no imbalance exists, these allele-frequency based haplotypes reflect only stochastic deviation. Otherwise, if there exists some true level of AI, the frequency-based haplotypes should bear some resemblance to the estimated germline haplotypes. This resemblance is quantified with *phase concordance*, a measure of similarity between 2 haplotypes. *Switch accuracy*[81], or more appropriately, *switch consistency* accommodates errors in statistical haplotype reconstructions and is used as the metric for phase concordance in *hapLOHseq*. This is the same phase concordance metric that is used by *hapLOH* and is described in detail in the corresponding manuscript[51].

Briefly, I describe how phase concordance is calculated. At each pair of adjacent informative het sites, if the overrepresented alleles (e.g., the alleles with allele frequencies larger than the frequency threshold) reside on the same statistically estimated haplotype, a "1" is recorded. If the alleles reside on different statistically estimated haplotypes, a "0" is recorded. In normal regions of the genome, we do not expect that an excess haplotype exists. In this case, we expect to record "1"s and "0"s at random, and the running average of these numbers, the *phase concordance*, would be 0.5. In regions where there is allelic imbalance and where an excess haplotype exists, we expect that the frequency-based haplotypes have some resemblance to the statistically estimated haplotypes, and we expect to observe a higher number of "1"s (e.g., we expect a phase concordance > 0.5). This string of "1"s and "0"s is referred to as *switch enumerations*.

6.2.2.3 *Identification of allelic imbalance regions with a hidden Markov model*

To identify regions of the genome with higher than expected phase concordance, we implemented a simple HMM that was proposed and implemented in *hapLOH*[51]. The observed data in the HMM are the aforementioned *switch enumerations*. Let $i = 1,\ldots,m$ represent informative heterozygous sites in the genome. Let L_i be an indicator for whether or not the interval between i and $i+1$ is contained within a region of AI in the tumor genome. $L_i,\ldots,L_{M-1}$ form a Markov chain on two or more states. By default, *hapLOHseq* implements a 2-state HMM where state 0 represents no AI and state 1 represents AI. If there are multiple AI event states (e.g., HMMs with 3 or more states), each event state corresponds to a different degree of AI. The transition probabilities are constructed as shown in Figure 27.

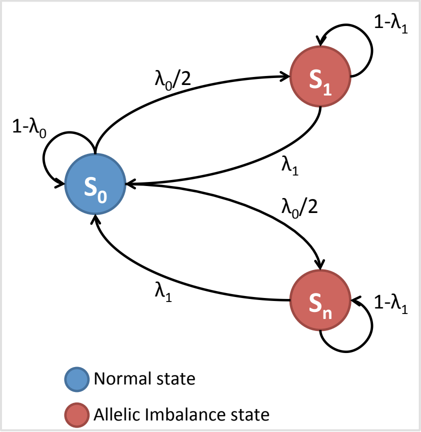


Figure 27: *hapLOHseq* HMM state transition diagram. There is one normal state and one or more AI event states where if there exists more than one AI state, each AI state represents events with different degrees of AI. By default, hapLOHseq uses one event state (i.e., n = 1).

We let α_l ($l = 0,1,\ldots,n$) denote the emission probability $Pr(x_i{=}1 | L_i{=}l)$. The emission probability α_0 for the normal state is set to 0.5 by default, which is the expected phase concordance in normal regions of the genome. In these regions, the frequency-based haplotypes have no resemblance to the statistically estimated haplotypes because there is no haplotype imbalance. For other states, the parameter will be larger than 0.5 and will be estimated using the *Baum-Welch* algorithm[82] which is an algorithm that attempts to find the HMM parameters that maximize the likelihood of the observed switch enumerations. These parameters are estimated separately for each chromosome. The probability that the process is in state *l* at marker interval *i* is calculated using the standard forward and backward algorithm[82] and these are reported in *hapLOHseq* output files.

Initial values for λ_0 and λ_1 are set from 2 user parameters that represent the expected event prevalence (p) and the expected event length (y) in megabases. By default, *hapLOHseq* takes the default size of a genome analyzed (3 billion for humans) and divides that by the number of informative het sites in the sample being analyzed. This value represents the average distance between informative het sites in megabases. The expected event length y is then divided by this distance and represented in terms of numbers of het sites with an x in the parameterizations below. Prevalence is represented with a p.

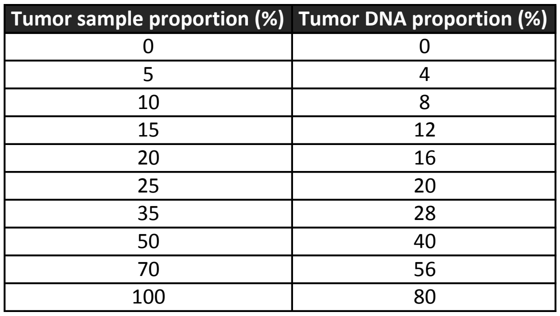
$$\lambda_0 = \frac{1}{x\left(\frac{1-p}{p}\right)} \tag{3}$$

$$\lambda_1 = \frac{1}{x} \tag{4}$$

For further details of this HMM including its performance in simulated settings please see the *hapLOH* manuscript[51]. In the next section, we apply and assess the performance of *hapLOHseq* using the MaCH and pairwise phasing strategies on a TCGA sample at various tumor purities.

6.2.3 Results

We obtained the whole genome and exome sequencing reads for the tumor (brain tissue) and normal (blood) sample of a patient with glioblastoma (TCGA-19-2620) from TCGA. Additionally, the published LOH and CNA calls for this tumor sample inferred from SNP arrays were obtained for comparison to *hapLOHseq* AI calls. To assess the performance of *hapLOHseq* at different levels of tumor purity, we created computational mixtures of the reads, which represent tumor sequencing at various purity levels. Sequencing read pairs from the tumor and normal BAM files were randomly sampled and merged into mixed BAM files at the tumor proportions listed in the first column of Table 20. TCGA estimates that the proportion of tumor DNA that is in the tumor sample is 80%. Thus, an estimate of the actual proportion of tumor DNA in each of these mixtures is listed in the second column of Table 20.



Tumor sample proportion (%)	Tumor DNA proportion (%)
0	0
5	4
10	8
15	12
20	16
25	20
35	28
50	40
70	56
100	80

Table 20: Tumor mixtures generated by mixing random sequence read pairs from a tumor and normal sample from a TCGA patient (TCGA-19-2620). Reads were mixed using the sample proportions specified. TCGA estimates that the tumor sample has 80% tumor DNA and 20% normal DNA. The estimated tumor DNA proportion of each mixture is specified.

hapLOHseq was then run on the tumor mixture BAM files for both whole genome and exome sequencing using two phasing strategies – MaCH and *pairwise phasing* as described in section 6.2.2.1. *hapLOHseq* was run on all of these samples using an estimated event length of 20 megabases. The event prevalence parameter was set to 0.001 for the exome sequencing runs and set to 0.00001 for the whole genome sequencing runs. The prevalence for the whole genome analyses was set lower to reduce noise in the whole genome sequencing results.

hapLOHseq plots were then generated for 4 sets of runs. Figure 28 and Figure 29 show the results of running *hapLOHseq* on the exome sequencing tumor mixture samples using MaCH and pairwise phasing, respectively. Figure 30 and Figure 31 show the results of running *hapLOHseq* on the whole genome sequencing tumor mixture samples using MaCH and pairwise phasing, respectively. These figures include at the top, 2 panels that show the LOH and CNA calls across the genome published by the TCGA. The intensity of the red color in the LOH plot reflects the degree of LOH observed at these sites. In the CNA plot, red represents amplification events and blue represents deletion events. The intensity of those colors reflects the copy number change observed in these regions. The subsequent *hapLOHseq* plots show (with gray dots) the *variant allele fractions* (VAFs) observed in the various tumor mixture samples at polymorphic sites across the genome that are heterozygous in the germline sample. In addition, there is a red line that shows the probability of regions of the genome being in allelic imbalance based on probabilities (from 0 to 1) reported by *hapLOHseq*.

As can be seen in these plots, at higher tumor purities, it is easier to identify AI events. In the exome sequencing analyses, the more prominent events, which were identified by TCGA on chromosomes 8, 9 and 10 can be found in the *hapLOHseq* results at 20% tumor purity with some signal of the events showing in purities as low as 8%. The MaCH phasing strategy appears to be performing better than the pairwise phasing at these lower tumor purities. Given that the phasing accuracy of MaCH and pairwise phasing are almost identical, the improvement in *hapLOHseq* sensitivity is likely due to the increased number of informative het sites (770 using the MaCH strategy versus 669 using the pairwise phasing strategy) that are available to *hapLOHseq* using the MaCH phasing strategy.

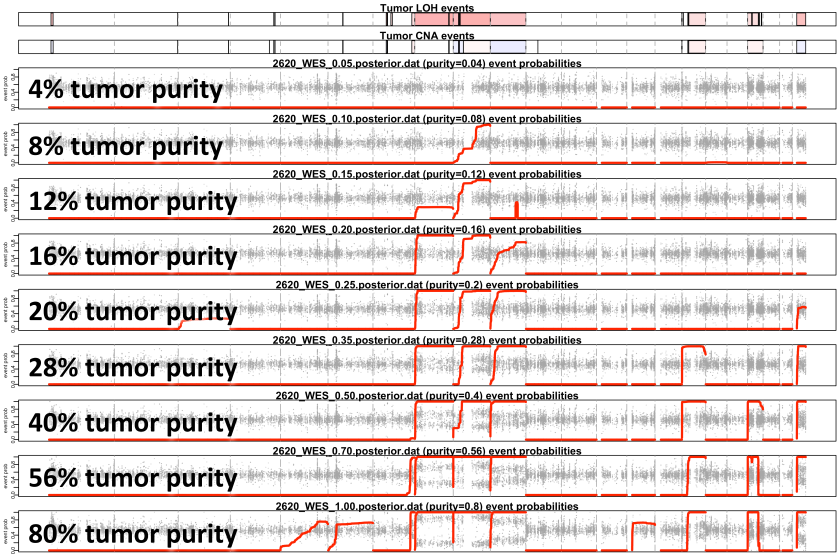


Figure 28: *hapLOHseq* calls at different computational dilutions for a single TCGA sample derived from exome sequencing data using MaCH to statistically estimate germline haplotypes. Calls made and published from SNP microarray analysis by the TCGA are represented in the top two bars (LOH and CNA events, respectively).

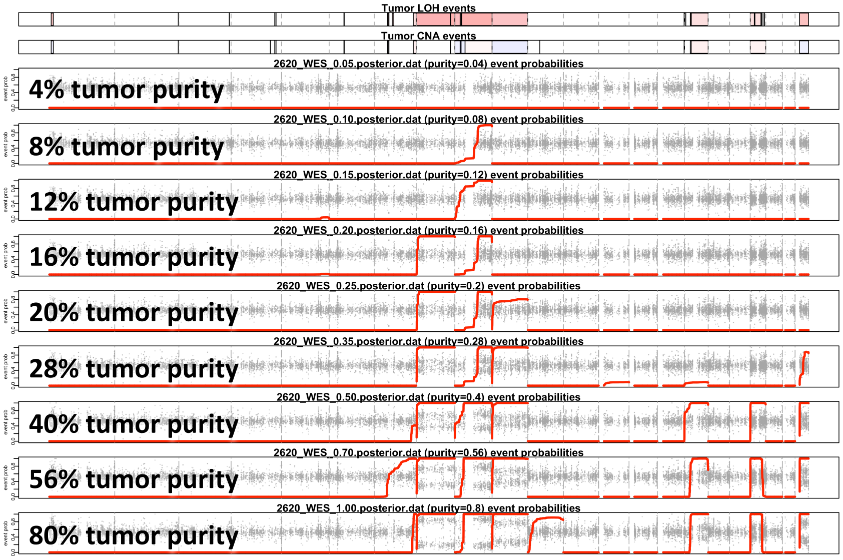


Figure 29: *hapLOHseq* calls at different computational dilutions for a single TCGA sample derived from exome sequencing data using pairwise phasing to estimate germline haplotypes. Calls made and published from SNP microarray analysis by the TCGA are represented in the top two bars (LOH and CNA events, respectively).

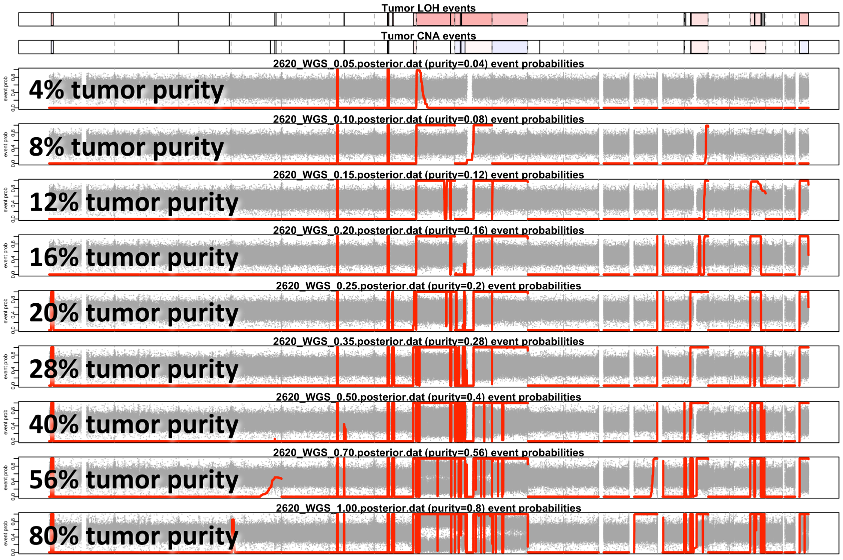


Figure 30: *hapLOHseq* calls at different computational dilutions for a single TCGA sample derived from whole genome sequencing data using MaCH to statistically estimate germline haplotypes. Calls made and published from SNP microarray analysis by the TCGA are represented in the top two bars (LOH and CNA events, respectively).

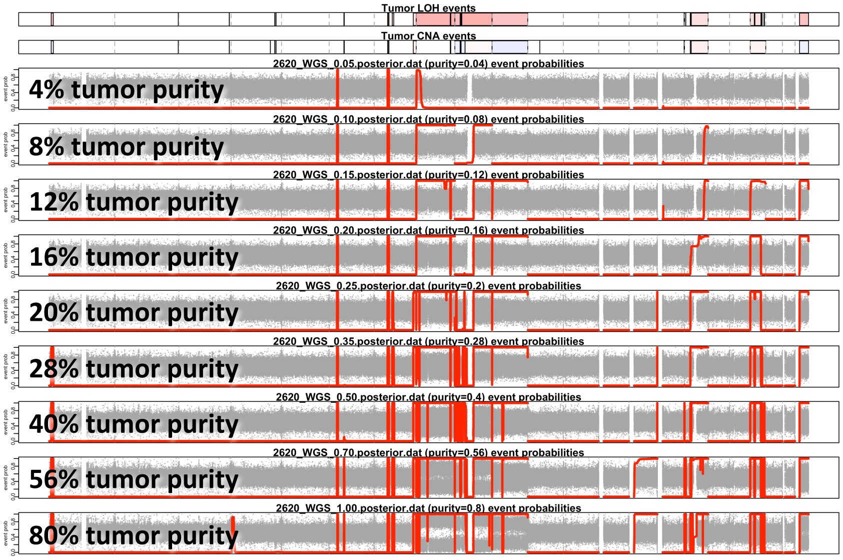


Figure 31: *hapLOHseq* calls at different computational dilutions for a single TCGA sample derived from whole genome sequencing data using pairwise phasing to estimate germline haplotypes. Calls made and published from SNP microarray analysis by the TCGA are represented in the top two bars (LOH and CNA events, respectively).

The sensitivity of *hapLOHseq* on whole genome sequencing data is much better at tumor purities in the range of 8% to 29%. The boundaries of events are more well-defined and the less prominent events identified by TCGA on chromosomes 16, 19 and 22 are identified at 12% tumor purity with some signal also seen at 8% tumor purity. The choice of phasing strategy with whole genome sequencing appears to have little to no effect on the sensitivity of *hapLOHseq*. This is likely because the phasing accuracy of MaCH and pairwise phasing are virtually identical and the number of informative het sites is so large (greater than 100,000) that the difference in these numbers (see Table 19) between the 2 strategies is negligible.

To summarize *hapLOHseq* performance, *receiver operating characteristic* (ROC) curves in Figure 32 and Figure 33 show the sensitivity and specificity of *hapLOHseq* on calling events larger than 10 megabases on the computational dilutions of the TCGA sample. Table 21 lists the *area under the curve* (AUC) corresponding to these ROC curves. At purities between 12% and 16% *hapLOHseq* is able to detect some events in the exome data. At these purities, as observed in the *hapLOHseq* plots, MaCH phasing is identifying events more precisely, likely due to the increased number of informative het sites available. Applied to whole genome sequencing, *hapLOHseq* is able to pick up events at tumor purities in between 4% and 8% where the choice of phasing strategy has little effect on the accuracy of identifying AI regions.

Of note, the AUC at 80% tumor purity is lower than that compared to many of the AUC values at lower tumor purities. We believe that at 80% tumor purities, *hapLOHseq* is detecting low-proportion chromosomal events on chromosome 14. We believe these are true events that were not detected in TCGA analyses. We believe this to be the case because we see the same event across different technologies and phasing strategies but only at high tumor purities.

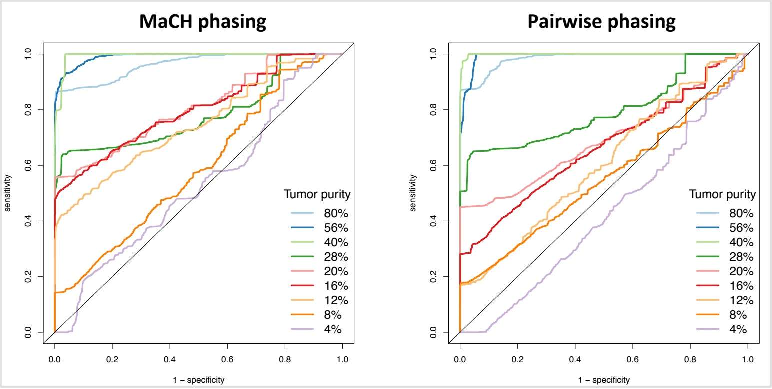


Figure 32: ROC comparison for MaCH and pairwise phasing for exome sequencing data.

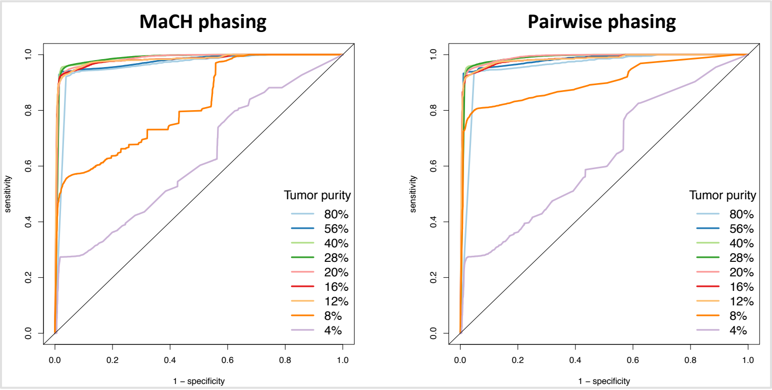


Figure 33: ROC curves for MaCH versus pairwise phasing for whole genome sequence data.

Tumor purity	Pairwise phasing Exome (AUC)	MaCH phasing Exome (AUC)	Pairwise phasing Whole genome (AUC)	MaCH phasing Whole genome (AUC)
80	0.984	0.963	0.956	0.958
56	0.992	0.991	0.980	0.974
40	0.999	0.994	0.986	0.985
28	0.794	0.784	0.985	0.985
20	0.696	0.809	0.988	0.984
16	0.665	0.799	0.982	0.981
12	0.606	0.749	0.982	0.981
8	0.551	0.590	0.902	0.819
4	0.420	0.531	0.635	0.625

Table 21: AUC for *hapLOHseq* calling strategies using pairwise and MaCH phasing for exome and whole genome samples at varying levels of tumor purity.

6.2.4 Discussion

We have presented a new method for the detection of subtle AI events in NGS data called *hapLOHseq*. We have also implemented a very efficient pairwise-phasing algorithm that allows for the estimation of haplotypes directly from VCF files, allowing users to run *hapLOHseq* without the need for low-level sequencing read files, which may either not be available, or which may be too resource intensive to run efficiently on a large number of samples. In summary, *hapLOHseq* is able to detect AI events from exome sequencing data, where these events exist in 12% to 16% of the cells sequenced. Applied to whole genome sequencing data, *hapLOHseq* has more sensitivity, being able to detect events occurring in 4% to 8% of the cells sequenced. *hapLOHseq* may be useful for the detection and profiling of AI in tumor samples that are either heavily diluted with normal tissue cells or in heterogeneous tumor samples.

6.3 Identification of candidate drugs: `cidd`

Cancer in silico Drug Discovery (CiDD) is a software framework for the identification of candidate drugs to target tumors with specific molecular characteristics. The description of CiDD in this section is part of a manuscript entitled *Cancer in silico Drug Discovery: a systems biology tool for identifying candidate drugs to target specific molecular tumor subtypes* (authored by F. Anthony San Lucas, Jerry Fowler, Kyle Chang, Scott Kopetz, Eduardo Vilar and Paul Scheet) that is currently under review at the journal *Molecular Cancer Therapeutics*.

6.3.1 Introduction

Selection of targeted therapies for cancer drug development has traditionally been based on the presence or absence of specific somatic mutations and this has been shown to be an effective strategy to improve patient outcomes[83–86]. However, a large number of targeted drugs and other compounds that have anti-tumor properties have not been linked to specific mutations, or biomarkers, that could be used to predict their selective efficacy[87]. Although next-generation sequencing (NGS) allows researchers to rapidly and comprehensively profile tumor mutations, the vast majority of these data have not been useful in the clinical setting since only a small number of mutations have been used to inform prognosis or guide therapeutic decisions[88–90].

Several computational approaches exist and have been implemented to predict the functional impact of mutations, and even to predict whether a specific mutation is a driver of the carcinogenesis process, based on several factors such as evolutionary conservation, predicted effects on protein structure and observed recurrence in existing cancer data sets[39,42,91].

However, these computational predictions provide little insight into how cellular processes are altered as a consequence of the mutations. One strategy to assess whether or not specific mutations are influential on cellular processes is to determine whether or not a mutation induces a signature of gene expression changes[92]. Gene expression signatures associated with an individual mutation could then be examined to characterize its cellular impact[47] and the signature could be used as a target for candidate drug therapies[58]. We have developed the *Cancer in silico Drug Discovery* (CiDD) platform for the purposes of characterizing tumors with specific mutations, or more generally tumors with specific clinicopathological or molecular characteristics, based on their putative effects on gene expression, and to identify candidate drugs to treat these tumors.

Here, we describe the general framework and integrated data sets of this novel platform. CiDD has been designed to generate hypotheses for the following three general problems: 1) to determine if particular clinical or molecular characteristics are functional and therefore induce unique gene expression signatures; 2) to find candidate drugs to treat specific tumor subgroups based on these expression changes; and 3) to identify cell lines that resemble the tumors being studied for subsequent *in vitro* experimentation.

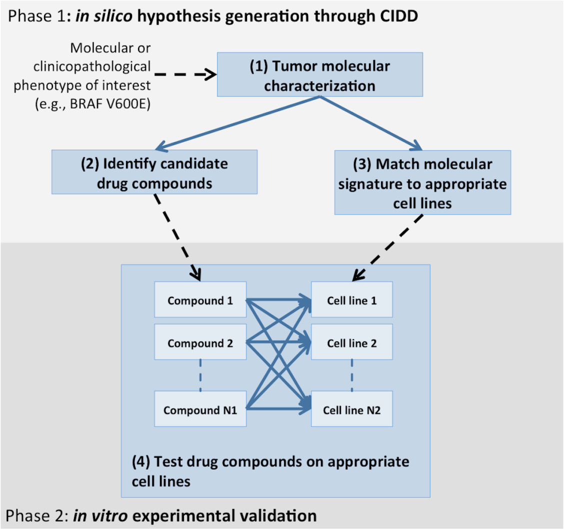

Phase 1: *in silico* hypothesis generation through CIDD
Molecular or clinicopathological phenotype of interest (e.g., BRAF V600E)
(1) Tumor molecular characterization
(2) Identify candidate drug compounds
(3) Match molecular signature to appropriate cell lines
Compound 1
Compound 2
Compound N1
Cell line 1
Cell line 2
Cell line N2
(4) Test drug compounds on appropriate cell lines
Phase 2: *in vitro* experimental validation

Figure 34: Overview of CiDD. The primary objective of CiDD is to specify initial candidate drug compounds and cell lines for laboratory drug experiments for a tumor characteristic being researched.

In addition, to illustrate the use of CiDD, we have applied it to a clinically relevant context in cancer drug development. We report the *in silico* identification of candidate drug therapies for *colorectal cancers* (CRCs) harboring the *BRAF* V600E mutation. Approximately 10% of CRCs harbor the *BRAF* V600E mutation, which confers a poor prognosis and presents a therapeutic challenge[86,93]. We describe the analyses performed with CiDD that have identified novel targets for *BRAF* mutant CRCs and have validated drugs that have already been identified as agents that target this tumor subtype such as *EGFR* inhibitors.

6.3.2 Methods

CiDD is a systematic drug discovery platform that integrates and analyzes large-scale cancer data sets with the primary goal of identifying candidate drugs and cell lines to be validated experimentally *in vitro* (see Figure 34). The core data sets used by CiDD include *The Cancer Genome Atlas* (TCGA), the *Connectivity Map* (CMap) and the *Cancer Cell Line Encyclopedia* (CCLE). CiDD is purely computational and depends on publicly available clinical and experimental datasets, as well as annotation databases. CiDD is written in Python, has R package dependencies and is command-line driven allowing it to be integrated into bioinformatics pipelines. The software and code are freely available at http://scheet.org/software.

6.3.2.1 Data assembly

Required experimental data sets for performing CiDD analyses are TCGA[28] and the CMap[58]. The CCLE[94] is required to identify cell-lines most appropriate for subsequent experimentation. TCGA includes clinical, mutation and gene expression data for thousands of samples across multiple cancer types. CiDD provides commands to download, query and analyze these data. The CMap is a collection of gene expression data for cell lines treated with small molecules paired with pattern-matching algorithms that attempt to identify biologically functional connections between drugs and gene expression profiles[58]. CiDD utilizes CMap build 02, which contains more than 7,000 expression profiles representing the effects of 1,309 compounds. The CCLE provides molecular profiles for 947 cancer cell lines which include DNA copy number, gene expression and DNA mutation data[94].

The experimental data from CMap consists of rank-based gene expression values from the Affymetrix HG-U133A microarray. Thus CMap is designed for the analysis of Affymetrix gene expression data only, which hinders using CMap with gene expression data collected from non-Affymetrix platforms. To overcome this limitation, CiDD transforms bulk-downloaded CMap data from Affymetrix probe-based rank values to Entrez gene-based ranks. Gene-based ranks are determined by taking the mean probe rank for each gene, sorting the mean rank values and then assigning a rank for each gene based on the sorted values. This allows results from RNA sequencing and Agilent microarray technologies, such as those provided by TCGA, to be analyzed with the drug-perturbed data of the CMap in a standardized way at the gene level. A similar strategy has been applied in the R package *gCMAP*[95] that allows users to query the CMap using Affymetrix probe identifiers or gene symbols. Gene-expression signatures derived from both Agilent gene expression microarrays and RNA sequencing have identified validated candidate drugs when analyzed with the Affymetrix-based drug signatures of CMap[96–98] demonstrating the feasibility of a cross-platform approach.

CiDD also uses optional annotation data sets, which include the *Molecular Signatures Database* (MSigDB)[47] for characterizing gene sets and drug databases including *DrugBank*[99], *Matador*[100] and *KEGG Drug*[72] for annotating candidate drugs. These drug databases provide information such as drug pharmacology, gene and pathway targets to make the drug reports produced by CiDD more informative for researchers. Public data from TCGA are automatically downloaded by CiDD, while data from CMap, CCLE and MSigDB require registration at their respective websites prior to downloading. Upon download, CiDD automatically prepares and manages all of the data sets for drug discovery analyses. Further descriptions of the contents of these data sets along with installation and pre-processing details are provided in section 6.3.3.1.

6.3.2.2 CiDD workflow

A common workflow using the CiDD framework is illustrated in Figure 35. Initially, a CiDD project based on a TCGA cancer type is created and clinical, mutation and gene expression data for TCGA samples are automatically downloaded. For an analysis, CiDD first identifies samples for use in computational experiments from TCGA based on user-defined clinicopathological or molecular phenotypes, such as specific gene mutations, microsatellite instability status, tumor stage, or a variety of other patient or tumor characteristics reported through TCGA projects. Based on the defined phenotype, CiDD identifies 2 classes of samples to compare. For a mutation-based phenotype, CiDD establishes one class containing samples with a defined mutation or set of mutations and a second class containing samples that are wild-type for the genes of interest. For a clinical phenotype, the user specifies both classes explicitly, such as the two classes corresponding to microsatellite instable and microsatellite stable tumors. CiDD then attempts to identify a gene-expression signature that is associated with the defined patient or tumor characteristic. If a gene expression signature exists for the phenotype of interest, that signature is characterized with gene sets defined in MSigDB and the signature is used to identify candidate drug therapies through pattern-matching algorithms proposed by the CMap. Subsequently, CiDD characterizes candidate drugs using databases such as DrugBank, Matador and KEGG Drug. Finally, CiDD identifies candidate cell lines on which to test the drugs *in vitro* by analyzing experimental data from the CCLE. The primary results of a CiDD execution are a biologically annotated candidate drug list and candidate cell lines for subsequent drug experimentation.

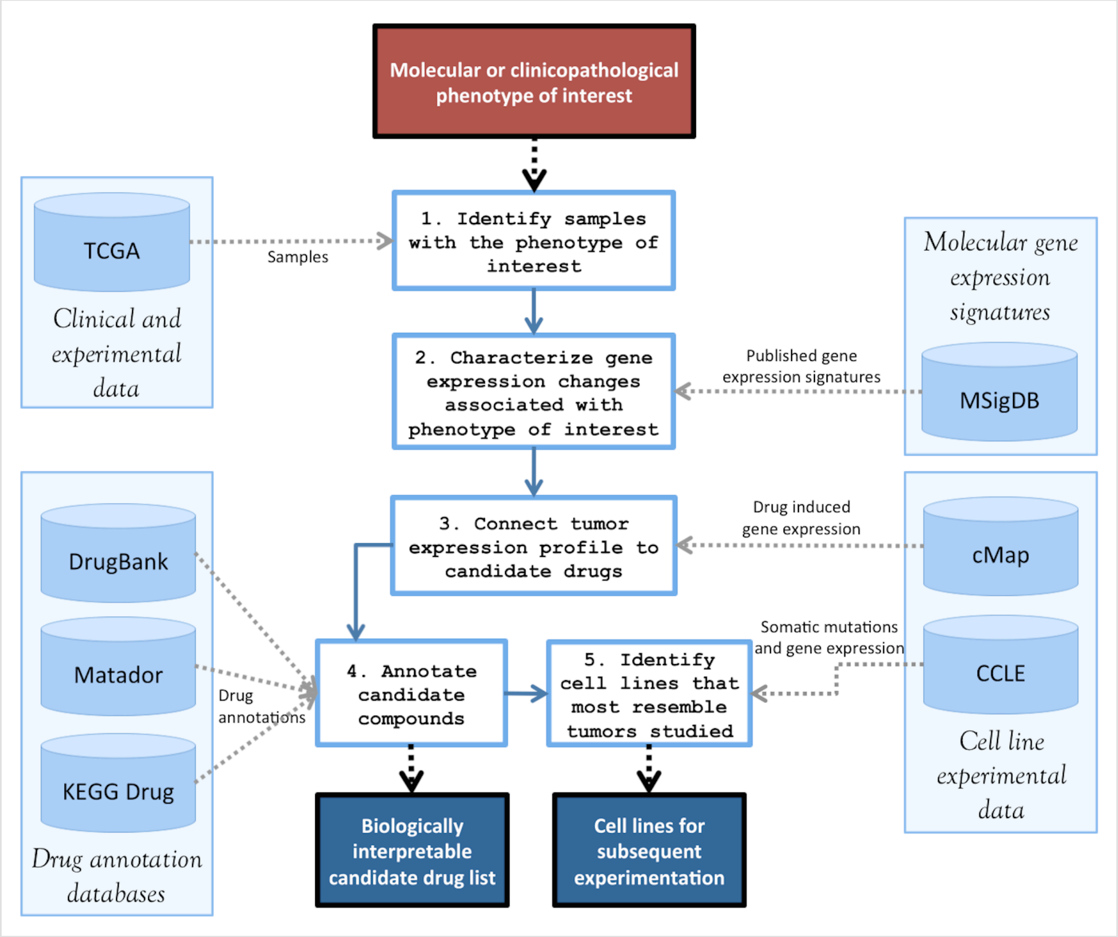

Molecular or clinicopathological phenotype of interest
1. Identify samples with the phenotype of interest
2. Characterize gene expression changes associated with phenotype of interest
3. Connect tumor expression profile to candidate drugs
4. Annotate candidate compounds
5. Identify cell lines that most resemble tumors studied
Biologically interpretable candidate drug list
Cell lines for subsequent experimentation
TCGA
Clinical and experimental data
Samples
DrugBank
Matador
KEGG Drug
Drug annotations
Drug annotation databases
Molecular gene expression signatures
MSigDB
Published gene expression signatures
Drug induced gene expression
cMap
Somatic mutations and gene expression
CCLE
Cell line experimental data

Figure 35: Steps and data sets of a basic CiDD workflow that identifies candidate drugs for a given molecular or clinicopathological phenotype of interest.

6.3.2.3 *Gene signature identification*

TCGA provides gene expression data from Agilent microarrays, Illumina GA RNA sequencing and Illumina HiSeq RNA sequencing. The gene expression data type to analyze can be specified as a parameter to CiDD. By default, CiDD will choose the technology that provides data for the largest number of samples with the phenotype of interest. Using the R package *Limma*[101] which is designed for both microarray and RNA sequencing differential expression analyses, CiDD identifies up- and down-regulated genes. CiDD characterizes differential expression results with known biological pathways by performing gene set tests from the *piano* Bioconductor package[102], while using gene sets defined by MSigDB.

6.3.2.4 *Generation of a k-top scoring pairs (k-TSP) classifier*

For generating a classifier that is robust across gene expression technologies, CiDD takes a non-parametric approach to classification and adopts an extension of the *top scoring pairs* (TSP) method[103]. Using the R package *ktspair*[104], CiDD generates a k-TSP classifier for predicting the status of the phenotype of interest on independent samples. The algorithm works by first ranking gene expressions for each sample and then identifying pairs of genes whose relative orderings within each sample class are opposite of one another. By default, to improve computational performance, CiDD limits the number of genes considered for inclusion in the classifier to only those genes in the gene expression signature. For each gene in a pair, g_1 and g_2, and for each sample s, the algorithm keeps track of whether the expression of g_1 in sample s is less than the expression of g_2 in sample s. The pairs that most consistently maintain their relative expression ordering in class 1 while having a reverse ordering in class 2 become gene pairs in the classifier. A score is assigned to each pair by the *ktspair* algorithm that represents the percentage of samples in the two classes that exhibit the expected ordering of g_1 and g_2. CiDD chooses the k pairs that meet a default threshold score of at least 0.8. The prediction is class 1 if the average expression value for the g_1 genes is lower than the average expression value for the g_2 genes; it is class 2, otherwise.

6.3.2.5 Candidate drug identification

CiDD connects the gene expression changes associated with the phenotype of interest with candidate drug compounds that induce a negatively correlated gene expression profile. CiDD compares the phenotype gene expression changes, termed a query signature, to rank-based gene expression profiles induced by CMap compounds. To compare rank-based gene expression profiles, CiDD implements a nonparametric pattern-matching algorithm based on the Kolmogorov-Smirnov statistic as described by Lamb *et al*[58]. Briefly, where up-regulated query genes tend to appear near the bottom of a compound's ranked gene expression profile and down-regulated query genes appear near the top of the ranked gene expressions, this suggests *negative connectivity*. *Positive connectivity* refers to the reverse scenario, where up-regulated query genes appear near the top of a compound's ranked gene expression profile and down-regulated query genes appear near the bottom of the ranked profile. There are multiple connectivity scores for each drug, one for each experiment where that drug was tested against an individual cell line. Connectivity scores range from -1 to +1 corresponding to negative and positive connectivity. *Enrichment* is a measure that aggregates the connectivity scores for all instances of a drug experiment to determine if they collectively have a negative or positive connectivity (ranging from -1 to 1) with the phenotype of interest. To assess the significance of the enrichment score, we have implemented the permutation procedure used by CMap, where the Kolmogorov-Smirnov statistic is computed for a set of CMap expression profiles generated from a single compound of interest within an ordered list of all the CMap expression profiles.

This provides an empirical p-value (CMap refers to it as a *permutation p*-value), a measure of the proportion of times the observed enrichment of a set of instances, or one more striking, would happen by chance. The metric *specificity* is a measure of the selectivity of a drug compound for the phenotype of interest. To determine specificity, random query signatures are extracted from MSigDB and run against the CMap to generate a background list of enrichment scores. Candidate drug compounds are deemed to have high specificity if results from these random query signatures do not identify the same candidate drug compounds as those from the query signature. CiDD then queries data downloaded from drug databases to annotate the candidate drug compounds with meaningful clinical and biological information to facilitate the biological interpretation of the list of candidate drugs.

6.3.2.6 Cell line identification

CiDD first selects CCLE cell lines based on user-specified tissue types. Then, CiDD optionally identifies cell lines that contain a user-specified mutation by interrogating CCLE mutation annotation files derived from either targeted sequencing of common cancer genes or from Oncomap 3.0, which is a SNP array that genotypes samples at the most common cancer mutation sites. Finally, CiDD runs its k-TSP classifier on CCLE gene expression data, as described in the previous section, to predict if a cell line's gene expression profile is representative of the phenotype being studied. Cell lines that meet these criteria are reported as candidates for use in subsequent drug experiments.

6.3.3 CiDD software description

6.3.3.1 Installation

Cancer in silico Drug Discovery (CiDD) is designed to run on Linux or Mac OS X environments with a recommended minimum of 100 GB of free disk space and 4 GB of memory, making it runnable on most bioinformatics desktop computers. CiDD is written in Python and has software and data dependencies. To use CiDD, users should follow the software and data set installation procedures described here.

6.3.3.1.1 Software installation

Software pre-requisites

The following should be installed before installing CiDD.

- Python 2.7 or greater
- Python libraries: numpy and lxml
- R 3.0 or greater
- R packages: edgeR, Limma, piano
- firehose_get (https://confluence.broadinstitute.org/display/GDAC/Download)
- tcga_util: a companion Python module for use with CiDD (http://scheet.org/software)

CiDD software installation

To install CiDD, download the source code from http://scheet.org/software and install the cidd command-line tool by running the following:

```
sudo python setup.py install
```

6.3.3.1.2 Data set installation

CiDD stores and manages data within a local data store. A data store is a directory on the user file system where data sets used by the CiDD framework are stored. A user can create a single data store that is shared between multiple CiDD projects and analyses. To do this, a user can set an environment variable called $DATA_STORE to the full path of the data store directory. This will tell CiDD where to find the default data store. As an alternative, a user can specify the location of their data store with each CiDD command through a --data_store parameter. This alternative approach works best if a user wants to keep TCGA cancer type data separate and manage multiple data stores. The following steps are required to initialize a CiDD data store:

1. Create the directory structure:

   ```
   mkdir ccle cmap drug_annotations msigdb tcga custom
   ```

   Alternatively, if no data store exists when the cidd setup command is run to initialize a CiDD project (see section 6.3.3.2.1) CiDD will create an empty data store directory structure and automatically populate the data store with an initial TCGA data set.

2. Download and store CiDD required data sets within the data store subdirectories. Project-specific TCGA data are automatically downloaded and managed by CiDD in the local data store. The user must download other required data sets into the data store (details follow).

3. Download and install optional drug annotation data sets within the data store subdirectories. These files are only needed for annotating candidate drug reports (details follow).

Here we describe the external data sets and files that the CiDD framework uses. There are 3 datasets listed here that are required to run CiDD and need to be downloaded manually and stored in the local data store. Other data sets are optional.

Data set descriptions

TCGA data[105] are automatically downloaded and managed by tcga_util, which is a companion Python module for use with CiDD. For a specified cancer project in TCGA, CiDD automatically downloads the clinical data, somatic mutations and gene expression data from RNA-sequencing and Agilent microarrays. The amount of data downloaded from TCGA is dependent on the cancer and data type being studied. In addition to the default data downloaded, data such as protein expression or miRNA expression could also be downloaded by tcga_util explicitly. The estimated download size of TCGA data ranges from 100 – 200 MB for most cancer projects.

The *Connectivity Map* (CMap)[58], *MSigDB*[47] and the *Cancer Cell Line Encyclopedia* (CCLE)[94] data downloads require user registration at their respective websites as detailed below. Data from *DrugBank*[99], *MATADOR*[100] and *KEGG*[72] *Drug* are not required by CiDD, but if available, they provide annotation sources for candidate drugs. Details of non-TCGA data dependencies are listed here.

CONNECTIVITY MAP (required)

requires registration: yes

website: http://www.broadinstitute.org/cmap

install location: $DATA_STORE/cmap

data files:

- instance inventory: cmap_instances_02.xls (1.6 MB)
- data matrix: rankMatrix.txt.zip (309 MB)
- gene sets: msigdb_gene_sets.zip (270 KB)

MSIGDB (required)

requires registration: yes

website: http://www.broadinstitute.org/gsea/msigdb/collections.jsp

install directory: $DATA_STORE/msigdb

data files: C2 curated gene sets: c2.all.v4.0.symbols.gmt (3.1 MB)

CANCER CELL LINE ENCYCLOPEDIA (required)

requires registration: yes

website: http://www.broadinstitute.org/ccle/data/browseData

install directory: $DATA_STORE/ccle

data files:

- mRNA expression: CCLE_Expression_Entrez_2012-09-29.gct (167.2 MB)
- Cell Line Annotations: CCLE_sample_info_file_2012-10-18.txt (196 KB)
- Oncomap mutations: CCLE_Oncomap3_2012-04-09.maf (318 KB)

- Hybrid capture sequencing mutations: CCLE_hybrid_capture1650_hg19_NoCommonSNPs_NoNeutralVariants_CDS_2012.05.07.maf (56.5 MB)

DRUGBANK (optional)

requires registration:	no
website:	http://www.drugbank.ca/downloads
install directory:	$DATA_STORE/drug_annotations/drugbank
data files:	full database in XML format: drugbank.xml.zip (16 MB)

MATADOR (optional)

requires registration:	no
website:	http://matador.embl.de
install directory:	$DATA_STORE/drug_annotations/matador
data files:	drug-protein interactions: matador.tsv.gz (419 KB)

KEGG DRUG (optional)

requires registration:	no
FTP site:	ftp://ftp.genome.jp/pub/kegg/medicus
install directory:	$DATA_STORE/drug_annotations/keggdrug
data files:	drug-molecule interactions: drug.kegg (21.7 MB)

6.3.3.2 CiDD commands

The CiDD framework is flexible for incorporation into custom workflows. The workflow can be modified and steps can be replaced with user-defined scripts. All workflow steps are executed with simple CiDD commands that are described here, where the intermediate input and output files of each command are stored and used in subsequent steps. Separating steps in this way allows users to more easily replace steps in the workflow with their own preferred methods or scripts. As an example, in a common workflow, a user specifies a molecular or clinicopathological phenotype of interest for a cancer type. CiDD would then identify sample IDs for 2 classes of samples for subsequent gene expression analyses. Alternatively, instead of specifying a clinical characteristic or mutation as the phenotype of interest, a user can perform their own analyses to identify samples with phenotypes that might not be directly supported by CiDD that she is interested in. This might include a class of samples with a particular methylation profile for a phenotype. The user could run an externally generated classifier based on methylation data and then identify their own subsets of samples based on the classifier's predictions. She can then supply her sample identifiers with class labels to CiDD at step 2, bypassing step 1. Similarly, other steps in the workflow can be replaced with user-created scripts. Here we describe the CiDD commands. For a description of parameters for each CiDD command, users can specify the -h flag with each command (e.g., cidd setup -h).

6.3.3.2.1 cidd setup

A CiDD project is initialized with the command cidd setup. Upon execution, clinical data, somatic mutations and gene expression data are automatically downloaded into the user data store for a specified TCGA cancer type specified with a --cohort parameter. Other TCGA data such as methylation or protein expression data can be downloaded explicitly into the user data store with the tcga_util command. Multiple cidd setup commands can be run for the same CiDD project to install data for multiple TCGA cancer types or multiple data release versions at anytime. By default, subsequent CiDD analyses will use the latest data sets downloaded unless otherwise specified through command parameters.

6.3.3.2.2 cidd clinical_signature, cidd_mutation_signature and cidd custom_signature

For a specified clinical or mutation-based phenotype, these commands can be used to identify gene expression signatures and then to characterize gene expression signatures using gene set tests. CiDD uses the R package *Limma*[101] to identify differentially expressed genes between the two classes as defined by command parameters. Limma supports both continuous expression measurements of microarray data and count measurements from RNA sequencing[106] which is appropriate for analysis of the TCGA data, which consists of gene expression data from Agilent microarrays, Illumina GA RNA sequencing and Illumina HiSeq RNA sequencing. The choice of expression data type can be specified through an --expression_type parameter. By default, CiDD will choose the expression technology that has data available for the largest number of samples with the phenotype of interest. CiDD requires a Benjamini-Hochberg adjusted p-value to be less than or equal to 0.05 and a $\log_2$ fold change greater than 1 to label a gene as being differentially expressed for inclusion in the gene signature, although these default parameters can be modified. The resulting sets of up- and down-regulated genes comprise the gene expression signature, which putatively represents the functional consequence of the mutation or phenotype being studied. A third, more generic, signature-based command cidd signature can be executed if a user wants to generate a signature based on two classes of samples where they have determined sample class membership external to CiDD. This command can also be used if users want to generate a signature using their own gene expression data set external to the TCGA.

This signature is then characterized with *MSigDB*[47] using the Bioconductor package *piano*[102]. By default, the reports generated by CiDD identify KEGG pathways that are associated with the phenotype of interest. Other MSigDB options for gene set groupings are also supported by CiDD, such as those defined by BIOCARTA, REACTOME, and GO.

6.3.3.2.3 cidd classifier

This command supports the generation of a mutation or phenotype classifier through cidd classifier generate and the application of the classifier for prediction with the cidd classifier predict command. CiDD constructs the gene expression classifier with an extension of the widely used non-parametric, rank-based algorithm, *top scoring pairs (TSP)*[103,107] called *k-TSP*[104] that is later applied downstream on CCLE samples to help identify candidate cell lines on which to test drug compounds. If an independent data set is available, one can also apply the classifier on this data set to assess the performance of the classifier.

6.3.3.2.4 cidd drugs

The cidd drugs command takes the signature generated from the cidd signature commands and finds candidate drugs from the CMap that induce a gene expression signature in the opposite direction to the one associated with the phenotype of interest. The CMap is a collection of gene expression data for cell lines treated with bioactive small molecules paired with pattern-matching algorithms that attempt to identify biologically functional connections between drugs and gene expression profiles[58]. CiDD utilizes CMap build 02, which contains more than 7,000 expression profiles representing the effects of 1,309 compounds. After identifying candidate drugs, the cidd drugs command annotates the candidate drugs using drug databases.

Before running cidd drugs, CiDD performs pre-processing of the CMap data. CMap (http://www.broadinstitute.org/cmap) is designed to allow users to upload a list of up- and down-regulated Affymetrix probe IDs that comprise a gene expression signature under study. The underlying gene expression data for CMap were collected from Affymetrix HG-U133A gene expression microarrays and are designed for use on Affymetrix gene expression data, which hinders using CMap with gene expression data collected from non-Affymetrix platforms. Thus, CiDD transforms bulk-downloaded CMap data from Affymetrix probe-based rank values to Entrez gene-based ranks. Gene-based ranks are determined by taking the mean probe rank for each gene, sorting the mean rank values and then assigning a rank for each gene based on the sorted values. This allows results from RNA sequencing and Agilent microarray technologies, such as those provided by TCGA, to be analyzed with the drug-perturbed data of the CMap in a standardized way at the gene level. A similar strategy has been applied in the R package *gCMAP*[95] that allows users to query the CMap using microarray probe identifiers or gene symbols. Gene-expression signatures derived from both Agilent gene expression microarrays and RNA sequencing have identified validated candidate drugs when analyzed with the Affymetrix-based drug signatures of CMap[96–98] demonstrating the feasibility of a cross-platform approach.

Further supporting a cross-platform approach, the pattern matching algorithms of the CMap are rank-based and robust to distributional assumptions of the data and to differences in normalization procedures across multiple data sets[58]. The command cidd drugs compares the phenotype gene expression changes, termed a query signature, to rank-based gene expression profiles induced by CMap compounds. The algorithm connects the gene expression changes induced by the phenotype of interest with candidate drug compounds that induce a negatively correlated gene expression profile. To compare rank-based gene expression profiles, CiDD implements the nonparametric pattern-matching algorithms based on the Kolmogorov-Smirnov statistic as described by Lamb *et al*[58].

Briefly, the "enrichment score" describes the connectivity between a drug and a query signature. This score ranges from -1 to +1 where a score near -1 reflects negative connectivity and a score near +1 reflects positive connectivity. A "permutation p-value" ranging from 0 to 1 provides a measure of significance for this score. A "specificity" value ranging from 0 to 1 describes how specific the drug is for the query signature, where a value close to 0 reflects high specificity. After calculating these metrics, CiDD then queries data downloaded from drug databases to annotate the candidate drug compounds with meaningful clinical and biological information to facilitate the biological interpretation of the list of candidate drugs. CiDD annotates candidate CMap drug compounds with *DrugBank*[99], *KEGG Drug*[72] and *Matador*[100]. CiDD attempts to link CMap provided compound names to drug database names, identifiers and drug aliases. Where links exist, the CMap identifiers are annotated with drug pharmacology, drug gene targets and drug pathway targets to help put the drugs into a biological context for clinical researchers. Additionally, Matador provides known drug-mRNA and drug-protein interactions.

6.3.3.2.5 cidd cell_lines

The CCLE provides molecular profiles for 947 cancer cell lines which include DNA copy number, gene expression and DNA mutation data[94]. From this data, cidd cell_lines searches the CCLE to identify cell lines that most closely resemble the cancer subtype being studied based on a specified tissue, a possible mutation of interest and the gene expression classifier generated by CiDD. Cell lines that fulfill criteria based on these characteristics are recommended for use in subsequent drug experiments. If a specific mutation is being studied (e.g., *BRAF* V600E), that mutation is searched for in the CCLE data set by querying mutations detected from Oncomap arrays and capture sequencing. These mutations are limited to 381 specific mutations across 33 genes using Oncomap 3.0 and to the coding regions of the 1651 genes defined in the CCLE target capture region. Search criteria can be relaxed through parameters of the command.

6.3.3.2.6 tcga_util

TCGA datasets are sufficiently useful and complex to warrant their own tool for downloading, querying, pre-processing and managing them. For this purpose, we developed tcga_util, a Python package, for use within the CiDD framework; however, tcga_util can also be useful as a stand-alone tool for generalized TCGA analyses. tcga_util manages TCGA data locally and has been designed for simple use at the command-line, which allows bioinformaticians to integrate TCGA data into their own repeatable analyses or custom applications and pipelines. CiDD uses this package directly for the automated download and management of TCGA data. These data sets are the source of the clinical and molecular data for CiDD to perform molecular characterization of the phenotypes of interest. Examples of clinical information includes age at diagnosis, gender and tumor stage as well as molecular diagnostics such as the presence of specific *KRAS* or *BRAF* mutations and microsatellite instability status. Available clinical data varies across TCGA tumor types. Molecular data available include whole genome and exome sequencing, methylation profiling, and gene and protein expression profiling, among other data types. Alternatively, direct TCGA data download through URLs and web forms is available through the NCI's TCGA Data Portal[105]. Another more user-friendly alternative for downloading and exploring subsets of TCGA data is the cBioPortal[108], which includes visual tools for browsing and analyzing TCGA data. An option for bulk TCGA data download is the utility firehose_get. Firehose is a large-scale data analysis pipeline that automatically performs standard pre-processing of TCGA data, easing the integration of data across cancer types and making the data more amenable to downstream analyses (https://confluence.broadinstitute.org/display/GDAC). The main goal of tcga_util is to help users query, download and filter through analysis-ready TCGA data for use in downstream analyses. To avoid duplication of effort, tcga_util leverages firehose_get for the majority of its TCGA data download, while adding functionality for the filtering and querying of downloaded data. tcga_util provides the following functionality:

1. download of TCGA clinical and experimental data into a local data store organized by cancer and data type simplifying repeat or new analyses,

2. sample query tools to easily find samples of interest based on clinical and mutational criteria,
3. creation of filtered data matrices that are composed of data for samples of interest that are easier to work with in downstream analysis tools such as R,
4. ability to update the local data store with the latest TCGA data releases, and

support of version tracking downloaded TCGA data for analysis reproducibility.

6.3.3.3 CiDD file descriptions

The files described here are generated by the previously described CiDD commands. Several of these files are intermediate results, being output by one CiDD command and then used as input in subsequent CiDD commands.

6.3.3.3.1 Sample files: {analysis_name}_{cases|controls}.samples

For a specified mutation or clinical characteristic, CiDD defines two classes of samples – a case class and a control class. The sample files list the TCGA identifiers that are a part of each class. These are generated by the cidd signature commands and used by tcga_util to construct case and control gene expression matrices.

6.3.3.3.2 RNA sequencing read count matrices: {analysis_name}_{cases|controls}.readcounts

CiDD downloads TCGA level 3 RNA sequencing data by default. The downloaded read count data has been RSEM normalized. These files are tab-delimited where genes correspond to rows and samples correspond to columns. Values in the files correspond to read counts. One file is generated for each class of samples by the cidd signature commands.

6.3.3.3.3 Agilent expression matrices: {analysis_name}_{cases|controls}.expr

CiDD downloads TCGA level 3 Agilent gene expression data by default. These data have been Lowess normalized. These files are tab-delimited where genes correspond to rows and samples correspond to columns. Values in the matrix correspond to a gene expression level. One file is generated for each class of samples by the cidd signature commands.

6.3.3.3.4 Differential expression results: {analysis_name}.diff_exp

Differential expression results from Limma are output in the diff_exp tab-delimited file. These results are produced by the cidd signature commands. Each row corresponds to a gene and differential expression metrics are represented in the columns. These columns include:

- logFC: $\log_2$ fold-change corresponding to the phenotype of interest
- AveExpr: average $\log_2$-expression value
- t: t-statistic
- P.Value: differential expression raw p-value
- adj.P.Val: Benjamini-Hochberg adjusted p-value or q-value
- B: log-odds that the gene is differentially expressed

Documentation describing Limma can be found at http://www.bioconductor.org/packages/release/bioc/html/limma.html.

6.3.3.3.5 Gene expression signature files: {analysis_name}_{up|down}.sig

The signature files contains a list of up and down-regulated gene identifiers based on fold-change and significance thresholds that define differentially expressed genes. These parameters can be specified in the cidd signature commands. The signature files produced by the cidd signature commands are input to the cidd drugs command.

6.3.3.3.6 Signature heatmap: {analysis_name}_heatmap.png

A sample and gene clustered heatmap using the case and control samples and signature genes is generated by the cidd signature commands. See Figure 36 for an example heatmap. The clustering of case samples based on signature genes is illustrated with a dendrogram at that top of the heatmap where black bars label the case samples.

6.3.3.3.7 Gene set analysis results: {analysis_name}.gsa

Gene set analysis results from the Bioconductor package *piano*[102] are output to the gsa file. Each row corresponds to a gene set. By default, CiDD uses KEGG gene sets defined in MSigDB[47] for the gene set tests. Columns in the file specify the numbers of genes in each gene set along with test statistic values and p-values that indicate whether or not each gene set is associated with the phenotype of interest. The software and documentation are available at http://bioconductor.org/packages/devel/bioc/html/piano.html.

6.3.3.3.8 Candidate drug report: {analysis_name}.drugs

The candidate drug report is a tab-delimited file produced by the cidd drugs command. Each row corresponds to a drug, and the column data are described below. See Lamb *et al*[58] for algorithm details for calculating values for mean_connectivity_score, enrichment and permutation_p.

- num_instances: the number of times that this drug was tested on a cell line

- mean_connectivity_score: the average connectivity score ranging from -1 to 1 across all instances of a drug. The closer the connectivity score is to -1 (i.e., the stronger the negative connectivity), the more we might expect that the drug will negate the gene expression signature of the phenotype of interest.
- enrichment: a measure of enrichment of all of the instances of a drug having a negative or positive connectivity (ranging from -1 to 1) with the gene expression signature of the phenotype of interest
- permutation_p: an estimate of the likelihood that the enrichment of a set of instances in the list of all instances in a given result would be observed by chance
- non_null_percentage: the number of instances with non-zero connectivity scores
- specificity: a measure of the selectivity of a drug compound for the phenotype of interest. Random query signatures are extracted from MSigDB and run against the CMap to generate a background list of enrichment scores and specificity indicates how often a score equal to or smaller than the enrichment is seen.
- pharmacology_drugbank: general description of the drug potentially including the drug origin, composition, pharmacokinetics, pharmacodynamics, therapeutic use and toxicology
- pathways_keggdrug: pathway targets of the drug
- targets_keggdrug: gene targets of the drug

interactions_matador: protein and gene interactions of which the drug is known to be a part of

6.3.3.3.9 Cell line report: {analysis_name}.cell_lines

The cell line report lists cell lines that are similar to the case samples identified by CiDD. Each row in the report corresponds to a candidate cell line.

6.3.4 Results: application of CiDD to BRAF V600E colorectal cancer

We applied CiDD to identify candidate drugs to treat CRCs harboring *BRAF* V600E mutations using mutation and RNA-sequencing data from the TCGA colon and rectum projects. We also identified cell lines from the CCLE that are representative of colorectal tumors with *BRAF* mutations, thus making them candidates for *in vitro* drug testing. We refer to these analyses as the *TCGA-derived* analyses. We then compared our systematic *TCGA-derived* analyses generated with CiDD with analyses performed using a previously published gene expression signature for *BRAF* V600E generated from CRC samples of the *PETACC3* (Pan-European Trials in Alimentary Tract Cancers) clinical trial[93]. We refer to these previously published gene expression analyses as the *PETACC3-derived* analyses.

The following commands were run to perform the *TCGA-derived* expression analyses and can be run to replicate the analysis using the same version of TCGA data as described in the main manuscript:

```
[1] cidd setup -dr 2014_01_15 -ar 2013_09_23 \
      -c coadread crc_brafv600e_proj
[2] cidd mutation_signature -dr 2014_01_15 -ar 2013_09_23 \
      -c coadread -g BRAF -aac V600E \
      -gnc 20 -lfc 2 -lperm 1000 -gperm 1000 \
      -gsm samplePermutation -n crc_brafv600e
[3] cidd classifier generate -n crc_brafv600e
```

```
[4] cidd drugs -np 1000 -nt 20 -n crc_brafv600e

[5] cidd cell_lines -g BRAF -aac V600E \

     -t LARGE_INTESTINE \

     -n crc_brafv600e
```

The following describes the commands for the analysis:

1. Setup a CiDD project called crc_brafv600e_proj and initialize it with data from the TCGA colon and rectum (i.e., coadread) project using data released on 2014_01_15 and analyses released on 2013_09_23.
2. Generate a mutation gene expression signature and characterize that signature with KEGG pathways. The mutation is specified to be in the BRAF gene with an amino acid change of V600E. By specifying the analysis name crc_brafv600e, output files of this command are prefixed with crc_brafv600e and can be automatically identified by CiDD in subsequent steps by specifying the analysis name. A minimum log fold change of 2 is specified for identifying differentially expressed genes. By default, a Benjamini Hochberg p-value of 0.05 is required for identifying differentially expressed genes. To assess significance for KEGG gene set tests, this command specifies the use of permuting sample labels (as opposed to permuting gene labels) 1000 times.
3. Generate a k-TSP classifier to predict *BRAF* V600E CRC status based on the sample class files generated in [2] by using the same analysis name crc_brafv600e.
4. Identify candidate drugs from the CMap using the mutation signature generated in [2] by specifying the same analysis name crc_brafv600e. The command specifies the use of

20 threads and 1000 permutations for assessing the significance (a permutation_p value) for the reported connectivity scores.

5. Identify candidate cell lines that are derived from a LARGE_INTESTINE tissue type, that harbor a BRAF V600E mutation and that exhibit a gene expression profile similar to *BRAF* V600E CRCs based on the classifier generated for the crc_brafv600e analysis.

6.3.4.1 Identification of a BRAF V600E gene expression signature

Among all TCGA CRC samples, we used CiDD to identify 20 samples with a *BRAF* V600E mutation and 149 *BRAF* wild-type samples with available Illumina GA RNA sequencing data. Then, CiDD identified 63 up-regulated and 170 down-regulated genes (*log fold-change* >= 2 and *Benjamini Hochberg adjusted p-value* <= 0.05) that generated a clustering of samples representative of *BRAF* mutation status as shown in Figure 36.

Figure 3

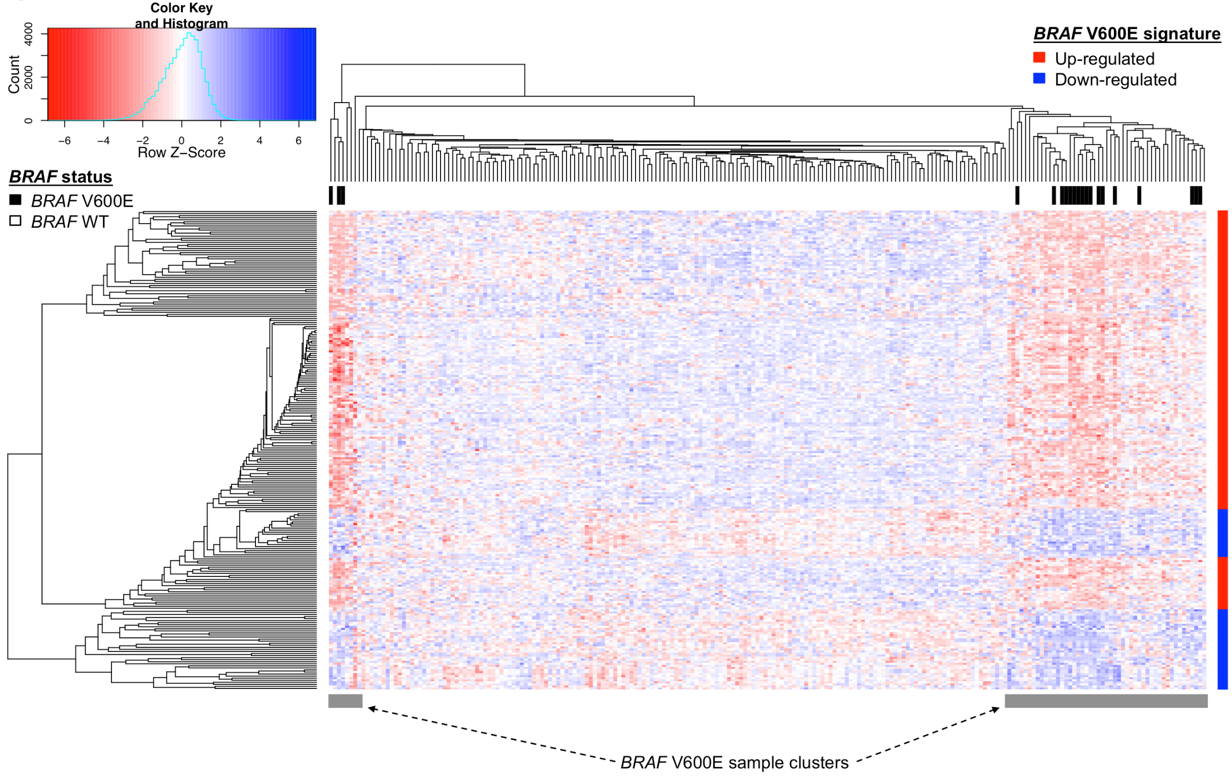
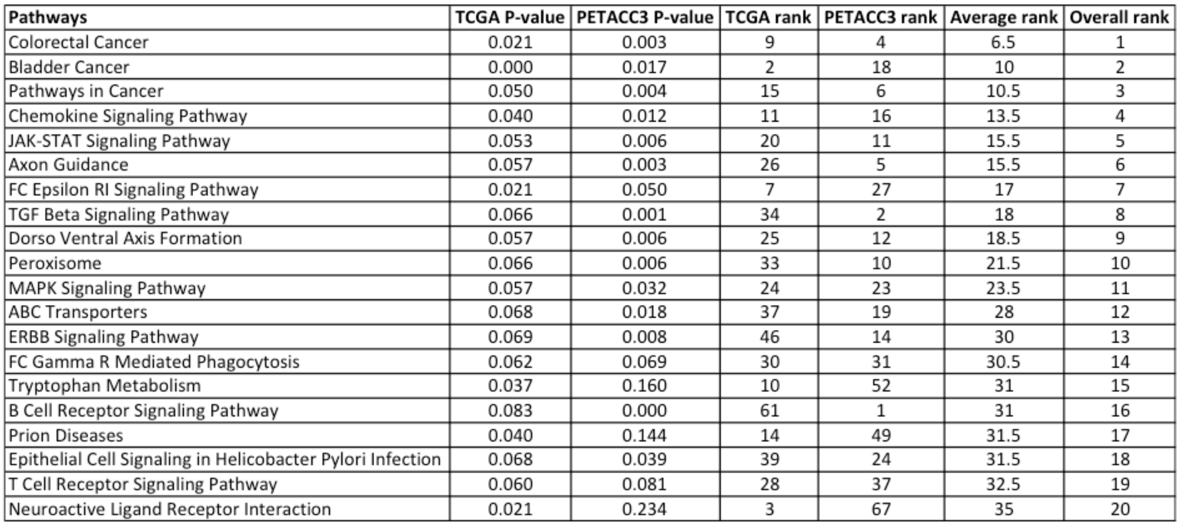


Figure 36: CiDD-generated heat map and clustering of BRAF V600E mutated CRCs based on TCGA Illumina GA RNA sequencing data. Differentially expressed genes comparing BRAF V600E and BRAF wildtype samples were identified using the Limma package in R and required to have a Benjamini Hochberg adjusted p-value <= 0.05 and a minimum log fold change >= 2. Hierarchical clustering of the samples and genes were performed using hclust with a "pearson" distance measure in R. The BRAF V600E gene expression signature is represented with the vertical colored bar on the right side of the figure, where red represents down-regulated genes and blue up-regulated genes. BRAF V600E mutant samples all reside within 2 sample clusters of the heatmap, which suggests that the BRAF V600E signature captures the gene expression response of BRAF V600E mutations.

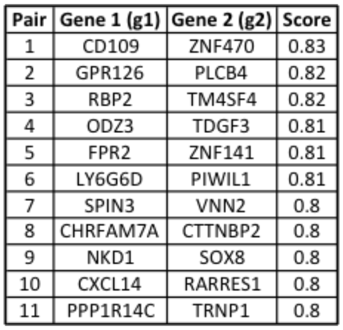
Then, we identified pathways associated with the *BRAF* signature through CiDD using Wilcoxon-based gene set tests[102]. For assessing significance of the gene set tests, CiDD performed 1000 runs of the differential expression analyses, permuting the *BRAF* mutant status of samples within each run. Fifteen KEGG gene sets were associated with the *BRAF* V600E status (*FDR adjusted p-value* <= 0.05). To incorporate *PETACC3-derived* pathways as part of the pathway analysis, a list of the top 20 pathways based on an average ranking within the *TCGA* and *PETACC3-derived* pathway lists is provided in Table 22. Because raw gene expression data was not available for the *PETACC3-derived* signature, gene set tests were not performed. Instead, for the *PETACC3-derived* analysis, hypergeometric tests were applied to identify KEGG pathways enriched with genes from this signature. Twenty-seven KEGG pathways are enriched with genes from the *PETACC3-derived* signature (*p-value* <= 0.05). The pathway ordering in Table 1 reflects the average of the *p-value* ranks within each set. Full reports are provided in Supplementary Results (see the tcga_gsa and petacc3_hyper sheets for the *TCGA-derived* and *PETACC3-derived* reports respectively). These pathways are consistently related to CRC biology such as the top ranked pathway ("Colorectal Cancer") and other pathways related to TGFβ signaling ("TGF Beta Signaling Pathway"), which are well known for their role in CRC. Additionally, it is known that the *BRAF* gene plays a role in controlling cellular proliferation and differentiation through regulation of the MAP kinase signaling pathway[109], and the "*MAPK* Signaling Pathway" is also represented in the top ranked pathways.



Pathways	TCGA P-value	PETACC3 P-value	TCGA rank	PETACC3 rank	Average rank	Overall rank
Colorectal Cancer	0.021	0.003	9	4	6.5	1
Bladder Cancer	0.000	0.017	2	18	10	2
Pathways in Cancer	0.050	0.004	15	6	10.5	3
Chemokine Signaling Pathway	0.040	0.012	11	16	13.5	4
JAK-STAT Signaling Pathway	0.053	0.006	20	11	15.5	5
Axon Guidance	0.057	0.003	26	5	15.5	6
FC Epsilon RI Signaling Pathway	0.021	0.050	7	27	17	7
TGF Beta Signaling Pathway	0.066	0.001	34	2	18	8
Dorso Ventral Axis Formation	0.057	0.006	25	12	18.5	9
Peroxisome	0.066	0.006	33	10	21.5	10
MAPK Signaling Pathway	0.057	0.032	24	23	23.5	11
ABC Transporters	0.068	0.018	37	19	28	12
ERBB Signaling Pathway	0.069	0.008	46	14	30	13
FC Gamma R Mediated Phagocytosis	0.062	0.069	30	31	30.5	14
Tryptophan Metabolism	0.037	0.160	10	52	31	15
B Cell Receptor Signaling Pathway	0.083	0.000	61	1	31	16
Prion Diseases	0.040	0.144	14	49	31.5	17
Epithelial Cell Signaling in Helicobacter Pylori Infection	0.068	0.039	39	24	31.5	18
T Cell Receptor Signaling Pathway	0.060	0.081	28	37	32.5	19
Neuroactive Ligand Receptor Interaction	0.021	0.234	3	67	35	20

Table 22: The top 20 ranked pathways associated with BRAF V600E status based on systematic TCGA gene expression analyses presented with those derived from the independent PETACC3-based analyses. The table is ordered by the overall rank of each pathway where the overall rank represents an average rank across both the TCGA- and PETACC-derived analyses. P-values and ranks for pathways associated for both the TCGA- and PETACC-derived analyses are shown. These pathways are consistently related to CRC biology such as the top-ranked pathway "Colorectal Cancer" and the "TGF Beta Signaling Pathway" in addition to the "MAPK Signaling Pathway" which is known to play a role in BRAF-mutant CRC.

Finally, we used CiDD to identify an 11-pair k-TSP classifier for predicting the *BRAF* V600E status of independent samples using the TCGA data set. The classifier gene pairs are listed in Table 23. For prediction, a default predictive score of 0.8 on the TCGA data set is required for inclusion into the classifier. If the average value or rank for the *g_1 genes* is less than that of the *g_2 genes*, the sample is predicted to harbor a *BRAF* V600E mutation and otherwise the sample is predicted to be *BRAF* wild type.



Pair	Gene 1 (g1)	Gene 2 (g2)	Score
1	CD109	ZNF470	0.83
2	GPR126	PLCB4	0.82
3	RBP2	TM4SF4	0.82
4	ODZ3	TDGF3	0.81
5	FPR2	ZNF141	0.81
6	LY6G6D	PIWIL1	0.81
7	SPIN3	VNN2	0.8
8	CHRFAM7A	CTTNBP2	0.8
9	NKD1	SOX8	0.8
10	CXCL14	RARRES1	0.8
11	PPP1R14C	TRNP1	0.8

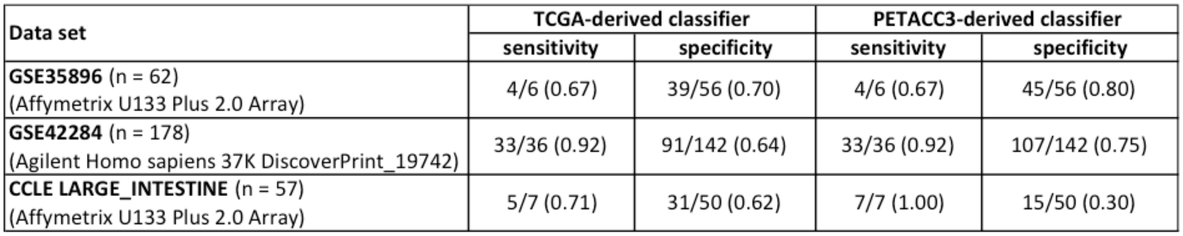
Table 23: TCGA-derived k-TSP classifier for predicting BRAF V600E status

6.3.4.2 *Validation of the TCGA-derived gene-pair classifier for predicting BRAF V600E status*

In order to validate the *TCGA-derived* gene expression analyses, we compared the performance of a previously reported *BRAF* V600E gene expression classifier derived from the *PETACC3* clinical trial [93] against the gene expression classifier that we identified from the TCGA data set.

The *PETACC3-derived* gene expression signature for our drug analyses consisted of 193 up-regulated and 92 down-regulated probes. These probes correspond to 224 unique genes. The research group also developed a 32-pair TSP classifier based on Affymetrix probe IDs for predicting the BRAF V600E status of CRCs. We translated these probe IDs to Entrez gene IDs so the classifier could be applied to RNA sequencing and Agilent test data sets. To assess the robustness of their gene expression results, we applied the gene-based *PETACC3-derived* classifier to TCGA samples that were retrieved and annotated with *BRAF* mutation statuses by CiDD. When applied to TCGA RNA sequencing data, the *PETACC3-derived* classifier resulted in 93.3% sensitivity and 83.5% specificity for detecting *BRAF* V600E samples.

To assess the quality of the systematic *TCGA-derived* classifier generated by CiDD, we compared the performance of the *TCGA-* and *PETACC3-derived* classifiers on 3 independent data sets (see Table 24) – two have been previously published and are available in the *Gene Expression Omnibus*[110,111] and the third is the CCLE data set. The sensitivity and specificity of both classifiers are comparable on the GSE35896 and GSE42284 data sets with the *PETACC3-derived* classifier exhibiting small improvements in specificity. The *PETACC3-derived* classifier achieved 100% sensitivity but only 30% specificity for *BRAF* status prediction on the CCLE large intestine data set. The *TCGA-derived* classifier had lower sensitivity (71%) but achieved better specificity (62%). These results suggest that the systematically obtained *BRAF* V600E classifier from CiDD is comparable to the published *PETACC3-derived* signature and that the *TCGA-derived* classifier may even have improved specificity for distinguishing *BRAF* wild-type cell lines from the *BRAF* mutant cell lines.



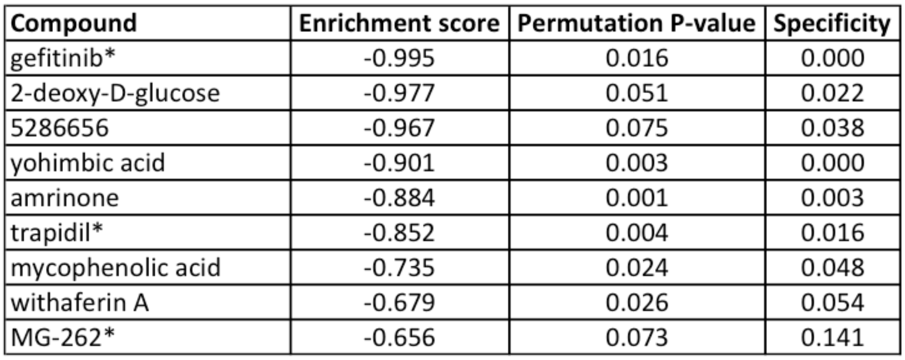
Data set	TCGA-derived classifier		PETACC3-derived classifier	
	sensitivity	specificity	sensitivity	specificity
GSE35896 (n = 62) (Affymetrix U133 Plus 2.0 Array)	4/6 (0.67)	39/56 (0.70)	4/6 (0.67)	45/56 (0.80)
GSE42284 (n = 178) (Agilent Homo sapiens 37K DiscoverPrint_19742)	33/36 (0.92)	91/142 (0.64)	33/36 (0.92)	107/142 (0.75)
CCLE LARGE_INTESTINE (n = 57) (Affymetrix U133 Plus 2.0 Array)	5/7 (0.71)	31/50 (0.62)	7/7 (1.00)	15/50 (0.30)

Table 24: Performance of the TCGA- and PETACC3-derived BRAF V600E CRC classifiers when applied to independent gene expression data sets. The sensitivity and specificity of both classifiers are comparable with the PETACC3-derived classifier exhibiting small improvements in specificity on the GSE35896 and GSE42284 data sets. The TCGA-derived classifier had lower sensitivity (71%) but achieved better specificity (62%) on the CCLE data set. These results suggest that the systematically obtained BRAF V600E classifier from CiDD is comparable to the

published PETACC3-derived signature and that the TCGA-derived classifier may even have improved specificity for distinguishing BRAF wild-type cell lines from the BRAF mutant cell lines.

6.3.4.3 *Candidate drug therapies for BRAF V600E CRC*

Using both the *TCGA* and *PETACC3-derived* gene expression signatures, CiDD identified potentially novel candidate drugs to treat *BRAF* V600E CRCs. Drugs with a negative enrichment score and a permutation *p*-value less than 0.1 using the *TCGA* gene expression signature are listed in Table 3. Three compounds, Gefitinib, MG-262 and Trapidil, were identified using both the *TCGA* and *PETACC3-derived* gene expression signatures. Independent research groups have recently shown that *EGFR* inhibitors such as Gefitinib and proteosome inhibitors such as MG-262 are effective drugs for treatment of colorectal tumors with *BRAF* mutations[86,112]. Trapidil is a novel candidate drug that inhibits *phosphodiesterase* and *TXA2*.



Compound	Enrichment score	Permutation P-value	Specificity
gefitinib*	-0.995	0.016	0.000
2-deoxy-D-glucose	-0.977	0.051	0.022
5286656	-0.967	0.075	0.038
yohimbic acid	-0.901	0.003	0.000
amrinone	-0.884	0.001	0.003
trapidil*	-0.852	0.004	0.016
mycophenolic acid	-0.735	0.024	0.048
withaferin A	-0.679	0.026	0.054
MG-262*	-0.656	0.073	0.141

Table 25: Candidate drug compounds identified systematically by CiDD for BRAF V600E CRC based on the TCGA-derived gene expression signature. Nine drugs were identified having both a negative enrichment score and a maximum permutation P-value of 0.1. Three of these drugs (*) were also identified using the PETACC3-derived gene expression signature.

6.3.4.4 Cancer cell lines that most resemble BRAF V600E CRC

Finally, in order to identify candidate cell lines for *in vitro* testing, CiDD analyzed data from the CCLE. From 947 cell lines in the CCLE, CiDD identified 48 large intestine samples that we consider to be representative of colorectal tumors. Then CiDD reduced this number to 7, representing those large intestine cell lines that have *BRAF* V600E mutations. Finally, using the 11 gene-pair k-TSP classifier generated by CiDD, 5 of these cell lines were predicted to be *BRAF* V600E on the basis of having similar gene expression profiles to the TCGA *BRAF* V600E mutated CRCs. The five identified cell lines include RKO, SNUC5, CL34, COLO205 and HT29. OUMS23 and SW1417 are the two *BRAF* V600E large intestine cell lines that are predicted to be *BRAF* wild-type by the *TCGA-derived* gene expression classifier.

6.3.5 Discussion

As genomic technologies have ushered in the potential for targeted drug development, large-scale public genomic databases have matured in size, scope and information content to complement this effort. It is thus advantageous, and indeed possibly necessary, to apply computational genomics to inform the drug discovery process. While subgroup classification for prognostic assessment and therapeutic planning has been applied clinically for decades, especially among hematologic malignancies and in some solid tumors such as breast cancers, other tumor types such as CRCs appear phenotypically homogenous and are thus clinically indistinguishable. In order to reveal subclasses for these tumors and to generalize their genome-based classification, the use of genetic and transcriptomic analyses may prove essential. Systems biology tools such as CMap, and we believe CiDD as well now, help fill this need of identifying candidate interventions that target specific pathways deregulated in these tumor subclasses. In this regard, CMap provided the original approach to guide drug development based on transcriptomic data. CiDD is taking this systems biology approach further by extending the CMap with the clinical and molecular data of TCGA along with the high-throughput experiments of the CCLE for the purposes of systematic cancer drug discovery. While current public resources such as that of TCGA are impressive, they are likely just a beginning. The basic logic of CiDD naturally extends to utilization of forthcoming, larger-scale databases from drug perturbation experiments and genetic and transcriptomic sequencing of tumors of a wider array of sizes and associated clinical outcomes.

We believe CiDD is the first framework that supports systematic drug discovery based on user-specified TCGA clinical and molecular phenotypes. CiDD allows researchers to perform the following: (1) assess whether or not a mutation or clinical phenotype is associated with a gene expression signature, (2) identify candidate drugs to target this gene expression signature, and (3) identify cell lines for subsequent *in vitro* drug experimentation. We have illustrated the power of such an approach in a meaningful application to CRCs with somatic mutations in *BRAF*. CiDD also offers utility to researchers simply wishing to interrogate and organize TCGA data, as it can be applied to create an inventory of available TCGA data with particular clinical or genomic features, such as available data sets or patients with particular mutations, independently of its drug identification capabilities.

One of the most crucial steps in the *BRAF* V600E analysis was identifying a gene expression signature associated with the *BRAF* V600E mutation and generating a classifier for predicting mutation status. In both of these cases, we showed that the signature and classifier of the CiDD framework are comparable to those identified from the published *PETACC3-derived* analyses[93]. Similarly to the *PETACC3-derived* signature and classifier, the CiDD-generated signature was composed of genes representative of known pathways associated with the *BRAF* V600E mutation, most notably the "*MAPK* Signaling Pathway", and the performance of the classifier on independent data sets generated from orthogonal gene expression technologies showed robustness. The advantage of CiDD analyses is that they are systematic studies of generally available datasets. We did not have to generate any of our own experimental data, and the gene expression analyses can be relatively easily replicated and repeated for other mutation or clinical phenotypes.

Once we validated the gene expression signature, we used CiDD to identify candidate compounds for tumors harboring the well-known *BRAF* V600E mutation. Since the initial communication of the presence of mutations in the kinase *BRAF* in cancer[113], activating mutations have been described in several malignancies with different frequencies such as hairy cell leukemia (100%), melanoma (50-60%), thyroid carcinoma (30-50%) and CRC (10%)[114]. The most frequently identified mutation is a valine-to-glutamic acid substitution at codon 600 (V600E) that activates the signaling cascade downstream of *MEK* and *ERK*[113]. Other mutations have been found at the same codon and are considered equivalents in terms of oncogenic activation[114]. Therefore, substantial efforts were invested on developing ATP-competitive *RAF* inhibitors such as Vemurafenib and Dabrafenib to specifically target the *MAPK* pathway. Yet, the clinical success of *BRAF* inhibition has been variable and highly dependent on the tumor context. In this regard, Vemurafenib has demonstrated improvement in survival in patients diagnosed with stage IV melanomas harboring the *BRAF* V600E mutation[115]. However, this degree of clinical benefit has not been observed in the same molecular context in CRCs[116]. This is probably secondary to the intrinsic mechanisms of resistance to *BRAF* inhibition that are specific to the tumor context[114]. *BRAF* mutations in CRCs have been associated with poor prognosis and an aggressive disease course, and a characteristic clinical phenotype consistent with older age at diagnosis, female gender, right-sided location and the presence of high levels of microsatellite instability[117,118].

Two strategies have been suggested to overcome the primary resistance to *BRAF* inhibition in CRC biology. One strategy that has been supported independently by two different groups is the inhibition of the *EGFR* pathway by using monoclonal antibodies against *EGFR* (such as Cetuximab) or kinase inhibitors (such as Gefinitib and Erlotinib) in combination with *BRAF* inhibitors. *EGFR* is activated by feedback mechanisms upon *BRAF* inhibition, thus reactivating *ERK* via *RAS* and *CRAF*, therefore combinations of *EGFR* and *BRAF* inhibition will synergize in terms of activity[86,119–121]. The second strategy is based on targeting the proteasome pathway. This has demonstrated specific activity against *BRAF* V600E mutant CRC cell lines and tumor xenografts. This set of experiments was performed using classical (Bortezomib) and novel (Carfilzomib) proteasome inhibitors and demonstrated similar activity. However, as opposed to *EGFR* feedback, proteasome inhibition seems to function independently of *BRAF* inhibition [112]. CiDD has been able to identify both types of compounds (*EGFR* and proteasome inhibitors) as candidate drugs through an agnostic approach, thus providing a biological validation of the value of CiDD as an screening tool to identify novel drugs to be tested and further developed in specific tumor subtypes.

CiDD also addresses the important issue of identifying the most appropriate cell lines as pre-clinical models for cancer researchers. Systematic comparisons between cancer cell lines and tumor samples from human tissues have documented substantial differences between the two, emphasizing the importance of making genomically informed choices when identifying cell lines as pre-clinical models of a tumor subtype [122]. The CCLE provides mutation and gene expression data that allow CiDD to make these molecularly informed decisions in selecting cell lines. In our *BRAF* V600E analysis, CiDD identified 7 large intestine cell lines harboring the *BRAF* V600E mutation. However, only 5 of the 7 were predicted to be *BRAF* V600E based on CiDD's gene expression classifier, suggesting heterogeneity among the *BRAF* V600E mutated cell lines. Helpfully, CiDD prioritized those cell lines into 2 groups for *in vitro* testing, proposing that 5 of the 7 *BRAF* V600E mutated large intestine cell lines more closely resemble the TCGA CRC *BRAF* V600E tumors at a gene expression level.

CiDD has some limitations that could restrict its application in specific situations. Primarily, CiDD is dependent on identifying a gene expression signature representative of a phenotype of interest. In some cases, a clinical phenotype or mutation may not actually induce a gene expression response. In other clinical contexts, such as for rare mutations and infrequent clinical phenotypes, CiDD may not have the power to identify the true underlying gene expression signature associated with the phenotype, because CiDD is limited by the number of samples available in TCGA with that specific phenotype. In these rare-phenotype analyses, CiDD may fail to identify a statistically significant gene expression signature representative of the phenotype of interest. Researchers interested in rare clinical or molecular subgroups will need to consider alternative strategies for increasing their sample sizes. These strategies may include aggregating TCGA tumor types or grouping mutations or clinical phenotypes in biologically meaningful ways, such as aggregating rare mutations at a gene or pathway level to increase the sample size. The CiDD command that generates gene expression signatures based on defined mutations provides support for aggregating mutations by listing amino acid substitutions explicitly, specifying types of mutations (such as Nonsense mutations) or by defining sets of mutations based on gene and gene set membership. Additionally, the CiDD framework does not support the identification of candidate drug combinations to target tumor subtypes. The CMap provides drug-perturbed data that were generated by applying compounds to cell lines one compound at a time. If future drug-perturbed data sets provide gene expression data of multiple compounds being applied to cell lines, incorporation of this data into CiDD should be relatively straightforward. As an alternative, the computational identification of multiple interacting candidate drugs based on current data sets is a potential area for future CiDD development.

Of course, these limitations apply more generally for these difficult scenarios and are not unique to CiDD. In fact, CiDD helps address these limitations by being easy to run and repeat to test multiple hypotheses quickly. Further, CiDD is a framework rather than a specific method *per se*. As public databases evolve and expand, and as robust statistical methodologies mature for cross-platform expression-based signature identification, CiDD could be adapted to incorporate these improved components. In this sense, what we have demonstrated here is a "lower bound" of sorts, and we expect more powerful findings to emerge from such efficient systems-based computation. Finally, the field of gene expression analysis, particularly for identifying signatures of cancer subtypes, has been criticized for failing to adhere to standards of repeatability [123]. Our software facilitates repeatability and even enables replication of findings with external data sets. In all of these aspects, we expect the community of cancer genomic researchers to benefit from, and further contribute to, this framework.

7 Bibliography

1. Siegel, R., Ma, J., Zou, Z. & Jemal, A. Cancer statistics, 2014: Cancer Statistics, 2014. *CA. Cancer J. Clin.* **64,** 9–29 (2014).
2. Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E. & Forman, D. Global cancer statistics. *CA. Cancer J. Clin.* **61,** 69–90 (2011).
3. Kwong, L. N. & Dove, W. F. APC and Its Modifiers in Colon Cancer. (2009). at <http://www.landesbioscience.com/pdf/08Nathke_Kwong.pdf>
4. Shussman, N. & Wexner, S. D. Colorectal polyps and polyposis syndromes. *Gastroenterol. Rep.* **2,** 1–15 (2014).
5. Fearon, E. R. & Vogelstein, B. A Genetic Model for Colorectal Tumorigenesis. *Cell* **61,** 759–767 (1990).
6. Kim, B. & Giardiello, F. M. Chemoprevention in familial adenomatous polyposis. *Best Pract. Res. Clin. Gastroenterol.* **25,** 607–622 (2011).
7. Haggar, F. & Boushey, R. Colorectal Cancer Epidemiology: Incidence, Mortality, Survival, and Risk Factors. *Clin. Colon Rectal Surg.* **22,** 191–197 (2009).
8. Brosens, L. A. A. Prevention and management of duodenal polyps in familial adenomatous polyposis. *Gut* **54,** 1034–1043 (2005).
9. Gorgun, E., Remzi, F. H., Goldberg, J. M., Thornton, J., Bast, J., Hull, T. L., Loparo, B. & Fazio, V. W. Fertility is reduced after restorative proctocolectomy with ileal pouch anal anastomosis: A study of 300 patients. *Surgery* **136,** 795–803 (2004).

10. Sarre, R., Frost, A., Jagelman, D., Petras, R., Sivak, M. & McGannon, E. Gastric and duodenal polyps in familial adenomatous polyposis: a prospective study of the nature and prevalence of upper gastrointestinal polyps. *Gut* **28,** 306–314 (1987).

11. Bulow, S. Duodenal adenomatosis in familial adenomatous polyposis. *Gut* **53,** 381–386 (2004).

12. Trimbath, J. & Giardiello, F. Review article: genetic testing and counselling for hereditary colorectal cancer. *Aliment Pharmacol Ther* **16,** 1843–1857 (2002).

13. Mori, Y., Nagse, H., Ando, H., Horii, A., Ichii, S., Nakatsuru, S., Aoki, T., Miki, Y., Mori, T. & Nakamura, Y. Somatic mutations of the APC gene in colorectal tumors: mutation cluster region in the APC gene. *Hum. Mol. Genet.* **1,** 229–233 (1992).

14. Walther, A., Johnstone, E., Swanton, C., Midgley, R., Tomlinson, I. & Kerr, D. Genetic prognostic and predictive markers in colorectal cancer. *Nat. Rev. Cancer* **9,** 489–499 (2009).

15. Plawski, A., Banasiewicz, T., Borun, P., Kubaszewski, L., Krokowicz, P., Skrzypczak-Zielinska, M. & Lubinski, J. Familial adenomatous polyposis of the colon. *Hered. Cancer Clin. Pract.* **11,** 15 (2013).

16. Friedl, W., Caspari, R., Sengteller, M., Uhlhaas, S., Lamberti, C., Jungck, M., Kadmon, M., Wolf, M., Fahnenstich, J., Gebert, J., Moslein, G., Mangold, E. & Propping, P. Can APC mutation analysis contribute to therapeutic decisions in familial adenomatous polyposis? Experience from 680 FAP families. *Gut* **48,** 515–521 (2001).

17. Sporn, M. & Suh, N. Chemoprevention of cancer. *Carcinogenesis* **21,** 525–530 (2000).

18. Waddell, W. R. & Loughry, R. W. Sulindac for polyposis of the colon. *J. Surg. Oncol.* **24,** 83–87 (1983).

19. Tsujii, M., Kawano, S., Tsuji, S., Sawaoka, H., Hori, M. & DuBois, R. N. Cyclooxygenase regulates angiogenesis induced by colon cancer cells. *cell* **93,** 705–716 (1998).

20. Wang, D. & DuBois, R. N. The role of COX-2 in intestinal inflammation and colorectal cancer. *Oncogene* **29,** 781–788 (2010).

21. Baek, S. J., Kim, K.-S., Nixon, J. B., Wilson, L. C. & Eling, T. E. Cyclooxygenase inhibitors regulate the expression of a TGF-β superfamily member that has proapoptotic and antitumorigenic activities. *Mol. Pharmacol.* **59,** 901–908 (2001).

22. Grosch, S., Maier, T. J., Schiffmann, S. & Geisslinger, G. Cyclooxygenase-2 (COX-2)-Independent Anticarcinogenic Effects of Selective COX-2 Inhibitors. *JNCI J. Natl. Cancer Inst.* **98,** 736–747 (2006).

23. Arber, N., Eagle, C. J., Spicak, J., Rácz, I., Dite, P., Hajer, J., Zavoral, M., Lechuga, M. J., Gerletti, P., Tang, J. & undefined, others. Celecoxib for the prevention of colorectal adenomatous polyps. *N. Engl. J. Med.* **355,** 885–895 (2006).

24. Solomon, S. D., Pfeffer, M. A., McMurray, J. J. V., Fowler, R., Finn, P., Levin, B., Eagle, C., Hawk, E., Lechuga, M., Zauber, A. G., Bertagnolli, M. M., Arber, N., Wittes, J. & for the APC and PreSAP Trial Investigators. Effect of Celecoxib on Cardiovascular Events and Blood Pressure in Two Trials for the Prevention of Colorectal Adenomas. *Circulation* **114,** 1028–1035 (2006).

25. Burn, J., Gerdes, A.-M., Macrae, F., Mecklin, J.-P., Moeslein, G., Olschwang, S., Eccles, D., Evans, D. G., Maher, E. R., Bertario, L. & undefined, others. Long-term effect of aspirin on cancer risk in carriers of hereditary colorectal cancer: an analysis from the CAPP2 randomised controlled trial. *The Lancet* **378,** 2081–2087 (2012).

26. Burn, J., Bishop, D. T., Chapman, P. D., Elliott, F., Bertario, L., Dunlop, M. G., Eccles, D., Ellis, A., Evans, D. G., Fodde, R., Maher, E. R., Moslein, G., Vasen, H. F. A., Coaker, J., Phillips, R. K. S., Bulow, S., Mathers, J. C. & for the International CAPP consortium. A Randomized Placebo-Controlled Prevention Trial of Aspirin and/or Resistant Starch in Young People with Familial Adenomatous Polyposis. *Cancer Prev. Res. (Phila. Pa.)* **4,** 655–665 (2011).

27. Zhou, D., Yang, L., Zheng, L., Ge, W., Li, D., Zhang, Y., Hu, X., Gao, Z., Xu, J., Huang, Y., Hu, H., Zhang, H., Zhang, H., Liu, M., Yang, H., Zheng, L. & Zheng, S. Exome Capture Sequencing of Adenoma Reveals Genetic Alterations in Multiple Cellular Pathways at the Early Stage of Colorectal Tumorigenesis. *PLoS ONE* **8,** e53310 (2013).

28. Muzny, D. M., Bainbridge, M. N., Chang, K., Dinh, H. H., Drummond, J. A., Fowler, G., Kovar, C. L., Lewis, L. R., Morgan, M. B., Newsham, I. F., Reid, J. G., Santibanez, J., Shinbrot, E., Trevino, L. R., Wu, Y.-Q., Wang, M., Gunaratne, P., Donehower, L. A., Creighton, C. J., Wheeler, D. A., Gibbs, R. A., Lawrence, M. S., Voet, D., Jing, R., Cibulskis, K., Sivachenko, A., Stojanov, P., McKenna, A., Lander, E. S., Gabriel, S., Getz, G., Ding, L., Fulton, R. S., Koboldt, D. C., Wylie, T., Walker, J., Dooling, D. J., Fulton, L., Delehaunty, K. D., Fronick, C. C., Demeter, R., Mardis, E. R., Wilson, R. K., Chu, A., Chun, H.-J. E., Mungall, A. J., Pleasance, E., Gordon Robertson, A., Stoll, D., Balasundaram, M., Birol, I., Butterfield, Y. S. N., Chuah, E., Coope, R. J. N., Dhalla, N., Guin, R., Hirst, C., Hirst, M., Holt, R. A., Lee, D., Li, H. I., Mayo, M., Moore, R. A., Schein, J. E., Slobodan, J. R., Tam, A., Thiessen, N., Varhol, R., Zeng, T., Zhao, Y., Jones, S. J. M., Marra, M. A., Bass, A. J., Ramos, A. H., Saksena, G., Cherniack, A. D., Schumacher, S. E., Tabak, B., Carter, S. L., Pho, N. H.,

Nguyen, H., Onofrio, R. C., Crenshaw, A., Ardlie, K., Beroukhim, R., Winckler, W., Getz, G., Meyerson, M., Protopopov, A., Zhang, J., Hadjipanayis, A., Lee, E., Xi, R., Yang, L., Ren, X., Zhang, H., Sathiamoorthy, N., Shukla, S., Chen, P.-C., Haseley, P., Xiao, Y., Lee, S., Seidman, J., Chin, L., Park, P. J., Kucherlapati, R., Todd Auman, J., Hoadley, K. A., Du, Y., Wilkerson, M. D., Shi, Y., Liquori, C., Meng, S., Li, L., Turman, Y. J., Topal, M. D., Tan, D., Waring, S., Buda, E., Walsh, J., Jones, C. D., Mieczkowski, P. A., Singh, D., Wu, J., Gulabani, A., Dolina, P., Bodenheimer, T., Hoyle, A. P., Simons, J. V., Soloway, M., Mose, L. E., Jefferys, S. R., Balu, S., O'Connor, B. D., Prins, J. F., Chiang, D. Y., Neil Hayes, D., Perou, C. M., Hinoue, T., Weisenberger, D. J., Maglinte, D. T., Pan, F., Berman, B. P., Van Den Berg, D. J., Shen, H., Triche Jr, T., Baylin, S. B., Laird, P. W., Getz, G., Noble, M., Voet, D., Saksena, G., Gehlenborg, N., DiCara, D., Zhang, J., Zhang, H., Wu, C.-J., Yingchun Liu, S., Shukla, S., Lawrence, M. S., Zhou, L., Sivachenko, A., Lin, P., Stojanov, P., Jing, R., Park, R. W., Nazaire, M.-D., Robinson, J., Thorvaldsdottir, H., Mesirov, J., Park, P. J., Chin, L., Thorsson, V., Reynolds, S. M., Bernard, B., Kreisberg, R., Lin, J., Iype, L., Bressler, R., Erkkilä, T., Gundapuneni, M., Liu, Y., Norberg, A., Robinson, T., Yang, D., Zhang, W., Shmulevich, I., de Ronde, J. J., Schultz, N., Cerami, E., Ciriello, G., Goldberg, A. P., Gross, B., Jacobsen, A., Gao, J., Kaczkowski, B., Sinha, R., Arman Aksoy, B., Antipin, Y., Reva, B., Shen, R., Taylor, B. S., Chan, T. A., Ladanyi, M., Sander, C., Akbani, R., Zhang, N., Broom, B. M., Casasent, T., Unruh, A., Wakefield, C., Hamilton, S. R., Craig Cason, R., Baggerly, K. A., Weinstein, J. N., Haussler, D., Benz, C. C., Stuart, J. M., Benz, S. C., Zachary Sanborn, J., Vaske, C. J., Zhu, J., Szeto, C., Scott, G. K., Yau, C., Ng, S., Goldstein, T., Ellrott, K., Collisson, E., Cozen, A. E., Zerbino, D., Wilks, C., Craft, B., Spellman, P., Penny, R., Shelton, T., Hatfield, M., Morris, S., Yena, P., Shelton, C., Sherman, M., Paulauskis, J., Gastier-Foster, J. M., Bowen, J., Ramirez, N. C., Black, A., Pyatt, R., Wise, L., White, P., Bertagnolli, M., Brown, J., Chan, T. A., Chu, G. C., Czerwinski, C., Denstman, F., Dhir, R., Dörner, A., Fuchs, C. S., Guillem, J. G., Iacocca, M., Juhl, H., Kaufman, A., Kohl III, B., Van Le, X., Mariano, M. C., Medina, E. N., Meyers, M., Nash, G. M., Paty, P. B., Petrelli, N., Rabeno, B., Richards, W. G., Solit, D., Swanson, P., Temple, L., Tepper, J. E., Thorp, R., Vakiani, E., Weiser, M. R., Willis, J. E., Witkin, G., Zeng, Z., Zinner, M. J., Zornig, C., Jensen, M. A., Sfeir, R., Kahn, A. B., Chu, A. L., Kothiyal, P., Wang, Z., Snyder, E. E., Pontius, J., Pihl, T. D., Ayala, B., Backus, M., Walton, J., Whitmore, J., Baboud, J., Berton, D. L., Nicholls, M. C., Srinivasan, D., Raman, R., Girshik, S., Kigonya, P. A., Alonso, S., Sanbhadti, R. N., Barletta, S. P., Greene, J. M., Pot, D. A., Mills Shaw, K. R., Dillon, L. A. L., Buetow, K., Davidsen, T., Demchok, J. A., Eley, G., Ferguson, M., Fielding, P., Schaefer, C., Sheth, M., Yang, L., Guyer, M. S., Ozenberger, B. A., Palchik, J. D., Peterson, J., Sofia, H. J. & Thomson., E. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487,** 330–337 (2012).

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

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

31. Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14,** 178–192 (2013).

32. Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. S. & Getz, G. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31,** 213–219 (2013).

33. San Lucas, F. A., Wang, G., Scheet, P. & Peng, B. Integrated annotation and analysis of genetic variants from next-generation sequencing studies with variant tools. *Bioinformatics* **28,** 421–422 (2011).

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

35. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).

36. Durbin, R. M., Altshuler, D. L., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Collins, F. S., De La Vega, F. M., Donnelly, P., Egholm, M., Flicek, P., Gabriel, S. B., Gibbs, R. A., Knoppers, B. M., Lander, E. S., Lehrach, H., Mardis, E. R., McVean, G. A., Nickerson, D. A., Peltonen, L., Schafer, A. J., Sherry, S. T., Wang, J., Wilson, R. K., Gibbs, R. A., Deiros, D., Metzker, M., Muzny, D., Reid, J., Wheeler, D., Wang, J., Li, J., Jian, M., Li, G., Li, R., Liang, H., Tian, G., Wang, B., Wang, J., Wang, W., Yang, H., Zhang, X., Zheng, H., Lander, E. S., Altshuler, D. L., Ambrogio, L., Bloom, T., Cibulskis, K., Fennell, T. J., Gabriel, S. B., Jaffe, D. B., Shefler, E., Sougnez, C. L., Bentley, D. R., Gormley, N., Humphray, S., Kingsbury, Z., Koko-Gonzales, P., Stone, J., McKernan, K. J., Costa, G. L.,

Ichikawa, J. K., Lee, C. C., Sudbrak, R., Lehrach, H., Borodina, T. A., Dahl, A., Davydov, A. N., Marquardt, P., Mertes, F., Nietfeld, W., Rosenstiel, P., Schreiber, S., Soldatov, A. V., Timmermann, B., Tolzmann, M., Egholm, M., Affourtit, J., Ashworth, D., Attiya, S., Bachorski, M., Buglione, E., Burke, A., Caprio, A., Celone, C., Clark, S., Conners, D., Desany, B., Gu, L., Guccione, L., Kao, K., Kebbel, A., Knowlton, J., Labrecque, M., McDade, L., Mealmaker, C., Minderman, M., Nawrocki, A., Niazi, F., Pareja, K., Ramenani, R., Riches, D., Song, W., Turcotte, C., Wang, S., Mardis, E. R., Wilson, R. K., Dooling, D., Fulton, L., Fulton, R., Weinstock, G., Durbin, R. M., Burton, J., Carter, D. M., Churcher, C., Coffey, A., Cox, A., Palotie, A., Quail, M., Skelly, T., Stalker, J., Swerdlow, H. P., Turner, D., De Witte, A., Giles, S., Gibbs, R. A., Wheeler, D., Bainbridge, M., Challis, D., Sabo, A., Yu, F., Yu, J., Wang, J., Fang, X., Guo, X., Li, R., Li, Y., Luo, R., Tai, S., Wu, H., Zheng, H., Zheng, X., Zhou, Y., Li, G., Wang, J., Yang, H., Marth, G. T., Garrison, E. P., Huang, W., Indap, A., Kural, D., Lee, W.-P., Fung Leong, W., Quinlan, A. R., Stewart, C., Stromberg, M. P., Ward, A. N., Wu, J., Lee, C., Mills, R. E., Shi, X., Daly, M. J., DePristo, M. A., Altshuler, D. L., Ball, A. D., Banks, E., Bloom, T., Browning, B. L., Cibulskis, K., Fennell, T. J., Garimella, K. V., Grossman, S. R., Handsaker, R. E., Hanna, M., Hartl, C., Jaffe, D. B., Kernytsky, A. M., Korn, J. M., Li, H., Maguire, J. R., McCarroll, S. A., McKenna, A., Nemesh, J. C., Philippakis, A. A., Poplin, R. E., Price, A., Rivas, M. A., Sabeti, P. C., Schaffner, S. F., Shefler, E., Shlyakhter, I. A., Cooper, D. N., Ball, E. V., Mort, M., Phillips, A. D., Stenson, P. D., Sebat, J., Makarov, V., Ye, K., Yoon, S. C., Bustamante, C. D., Clark, A. G., Boyko, A., Degenhardt, J., Gravel, S., Gutenkunst, R. N., Kaganovich, M., Keinan, A., Lacroute, P., Ma, X., Reynolds, A., Clarke, L., Flicek, P., Cunningham, F., Herrero, J., Keenen, S., Kulesha, E., Leinonen, R., McLaren,

W. M., Radhakrishnan, R., Smith, R. E., Zalunin, V., Zheng-Bradley, X., Korbel, J. O., Stütz, A. M., Humphray, S., Bauer, M., Keira Cheetham, R., Cox, T., Eberle, M., James, T., Kahn, S., Murray, L., Chakravarti, A., Ye, K., De La Vega, F. M., Fu, Y., Hyland, F. C. L., Manning, J. M., McLaughlin, S. F., Peckham, H. E., Sakarya, O., Sun, Y. A., Tsung, E. F., Batzer, M. A., Konkel, M. K., Walker, J. A., Sudbrak, R., Albrecht, M. W., Amstislavskiy, V. S., Herwig, R., Parkhomchuk, D. V., Sherry, S. T., Agarwala, R., Khouri, H. M., Morgulis, A. O., Paschall, J. E., Phan, L. D., Rotmistrovsky, K. E., Sanders, R. D., Shumway, M. F., Xiao, C., McVean, G. A., Auton, A., Iqbal, Z., Lunter, G., Marchini, J. L., Moutsianas, L., Myers, S., Tumian, A., Desany, B., Knight, J., Winer, R., Craig, D. W., Beckstrom-Sternberg, S. M., Christoforides, A., Kurdoglu, A. A., Pearson, J. V., Sinari, S. A., Tembe, W. D., Haussler, D., Hinrichs, A. S., Katzman, S. J., Kern, A., Kuhn, R. M., Przeworski, M., Hernandez, R. D., Howie, B., Kelley, J. L., Cord Melton, S., Abecasis, G. R., Li, Y., Anderson, P., Blackwell, T., Chen, W., Cookson, W. O., Ding, J., Min Kang, H., Lathrop, M., Liang, L., Moffatt, M. F., Scheet, P., Sidore, C., Snyder, M., Zhan, X., Zöllner, S., Awadalla, P., Casals, F., Idaghdour, Y., Keebler, J., Stone, E. A., Zilversmit, M., Jorde, L., Xing, J., Eichler, E. E., Aksay, G., Alkan, C., Hajirasouliha, I., Hormozdiari, F., Kidd, J. M., Cenk Sahinalp, S., Sudmant, P. H., Mardis, E. R., Chen, K., Chinwalla, A., Ding, L., Koboldt, D. C., McLellan, M. D., Dooling, D., Weinstock, G., Wallis, J. W., Wendl, M. C., Zhang, Q., Durbin, R. M., Albers, C. A., Ayub, Q., Balasubramaniam, S., Barrett, J. C., Carter, D. M., Chen, Y., Conrad, D. F., Danecek, P., Dermitzakis, E. T., Hu, M., Huang, N., Hurles, M. E., Jin, H., Jostins, L., Keane, T. M., Quang Le, S., Lindsay, S., Long, Q., MacArthur, D. G., Montgomery, S. B., Parts, L., Stalker, J., Tyler-Smith, C., Walter, K., Zhang, Y., Gerstein, M. B., Snyder, M., Abyzov, A., Balasubramanian, S., Bjornson, R., Du, J., Grubert, F., Habegger, L., Haraksingh, R., Jee, J., Khurana, E., Lam, H. Y. K., Leng, J., Jasmine Mu, X., Urban, A. E., Zhang, Z., Li, Y., Luo, R., Marth, G. T., Garrison, E. P., Kural, D., Quinlan, A. R., Stewart, C., Stromberg, M. P., Ward, A. N., Wu, J., Lee, C., Mills, R. E., Shi, X., McCarroll, S. A., Banks, E., DePristo, M. A., Handsaker, R. E., Hartl, C., Korn, J. M., Li, H., Nemesh, J. C., Sebat, J., Makarov, V., Ye, K., Yoon, S. C., Degenhardt, J., Kaganovich, M., Clarke, L., Smith, R. E., Zheng-Bradley, X., Korbel, J. O., Humphray, S., Keira Cheetham, R., Eberle, M., Kahn, S., Murray, L., Ye, K., De La Vega, F. M., Fu, Y., Peckham, H. E., Sun, Y. A., Batzer, M. A., Konkel, M. K., Walker, J. A., Xiao, C., Iqbal, Z., Desany, B., Blackwell, T., Snyder, M., Xing, J., Eichler, E. E., Aksay, G., Alkan, C., Hajirasouliha, I., Hormozdiari, F., Kidd, J. M., Chen, K., Chinwalla, A., Ding, L., McLellan, M. D., Wallis, J. W., Hurles, M. E., Conrad, D. F., Walter, K., Zhang, Y., Gerstein, M. B., Snyder, M., Abyzov, A., Du, J., Grubert, F., Haraksingh, R., Jee, J., Khurana, E., Lam, H. Y. K., Leng, J., Jasmine Mu, X., Urban, A. E., Zhang, Z., Gibbs, R. A., Bainbridge, M., Challis, D., Coafra, C., Dinh, H., Kovar, C., Lee, S., Muzny, D., Nazareth, L., Reid, J., Sabo, A., Yu, F., Yu, J., Marth, G. T., Garrison, E. P., Indap, A., Fung Leong, W., Quinlan, A. R., Stewart, C., Ward, A. N., Wu, J., Cibulskis, K., Fennell, T. J., Gabriel, S. B., Garimella, K. V., Hartl, C., Shefler, E., Sougnez, C. L., Wilkinson, J., Clark, A. G., Gravel, S., Grubert, F., Clarke, L., Flicek, P., Smith, R. E., Zheng-Bradley, X., Sherry, S. T., Khouri, H. M., Paschall, J. E., Shumway, M. F., Xiao, C., McVean, G. A., Katzman, S. J., Abecasis, G. R., Blackwell, T., Mardis, E. R., Dooling, D., Fulton, L., Fulton, R., Koboldt, D. C., Durbin, R. M., Balasubramaniam, S., Coffey, A., Keane, T. M., MacArthur, D. G., Palotie, A., Scott, C., Stalker, J., Tyler-Smith, C., Gerstein, M. B., Balasubramanian, S., Chakravarti, A., Knoppers, B. M., Abecasis, G. R., Bustamante, C. D., Gharani, N., Gibbs, R. A., Jorde, L., Kaye, J. S., Kent, A., Li, T., McGuire, A. L., McVean, G. A., Ossorio, P. N., Rotimi, C. N., Su, Y., Toji, L. H., Tyler-Smith, C., Brooks, L. D., Felsenfeld, A. L., McEwen, J. E., Abdallah, A., Juenger, C. R., Clemm, N. C., Collins, F. S., Duncanson, A., Green, E. D., Guyer, M. S., Peterson, J. L., Schafer, A. J., Abecasis, G. R., Altshuler, D. L., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E. & McVean, G. A. A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).

37. Ramensky, V. E., Bork, P. & Sunyaev, S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* **30,** 3894–3900 (2002).

38. Chun, S. & Fay, J. C. Identification of deleterious mutations within three human genomes. *Genome Res.* **19,** 1553–1561 (2009).

39. Ng, P. C. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31,** 3812–3814 (2003).

40. Schwarz, J. M., Rödelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods* **7,** 575–576 (2010).

41. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP v2.0: A Database of Human Non-synonymous SNVs and Their Functional Predictions and Annotations. *Hum. Mutat.* **34,** E2393–E2402 (2013).

42. Carter, H., Chen, S., Isik, L., Tyekucheva, S., Velculescu, V. E., Kinzler, K. W., Vogelstein, B. & Karchin, R. Cancer-Specific High-Throughput Annotation of Somatic Mutations: Computational Prediction of Driver Missense Mutations. *Cancer Res.* **69,** 6660–6667 (2009).

43. Forbes, S. A., Bindal, N., Bamford, S., Cole, C., Kok, C. Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., Teague, J. W., Campbell, P. J., Stratton, M. R. & Futreal, P. A. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **39,** D945–D950 (2010).

44. Koressaar, T. & Remm, M. Enhancements and modifications of primer design program Primer3. *Bioinformatics* **23,** 1289–1291 (2007).

45. Wood, L. D., Parsons, D. W., Jones, S., Lin, J., Sjoblom, T., Leary, R. J., Shen, D., Boca, S. M., Barber, T., Ptak, J., Silliman, N., Szabo, S., Dezso, Z., Ustyanksky, V., Nikolskaya, T., Nikolsky, Y., Karchin, R., Wilson, P. A., Kaminker, J. S., Zhang, Z., Croshaw, R., Willis, J., Dawson, D., Shipitsin, M., Willson, J. K. V., Sukumar, S., Polyak, K., Park, B. H., Pethiyagoda, C. L., Pant, P. V. K., Ballinger, D. G., Sparks, A. B., Hartigan, J., Smith, D. R., Suh, E., Papadopoulos, N., Buckhaults, P., Markowitz, S. D., Parmigiani, G., Kinzler, K. W., Velculescu, V. E. & Vogelstein, B. The Genomic Landscapes of Human Breast and Colorectal Cancers. *Science* **318,** 1108–1113 (2007).

46. Seshagiri, S., Stawiski, E. W., Durinck, S., Modrusan, Z., Storm, E. E., Conboy, C. B., Chaudhuri, S., Guan, Y., Janakiraman, V., Jaiswal, B. S., Guillory, J., Ha, C., Dijkgraaf, G. J. P., Stinson, J., Gnad, F., Huntley, M. A., Degenhardt, J. D., Haverty, P. M., Bourgon, R., Wang, W., Koeppen, H., Gentleman, R., Starr, T. K., Zhang, Z., Largaespada, D. A., Wu, T. D. & de Sauvage, F. J. Recurrent R-spondin fusions in colon cancer. *Nature* **488,** 660–664 (2012).

47. Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P. & Mesirov, J. P. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27,** 1739–1740 (2011).

48. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A.-L., Boyault, S., Burkhardt, B., Butler, A. P., Caldas, C., Davies, H. R., Desmedt, C., Eils, R., Eyfjörd, J. E., Foekens, J. A., Greaves, M., Hosoda, F., Hutter, B., Ilicic, T., Imbeaud, S., Imielinsk, M., Jäger, N., Jones, D. T. W., Jones, D., Knappskog, S., Kool, M., Lakhani, S. R., López-Otín, C., Martin, S., Munshi, N. C., Nakamura, H., Northcott, P. A., Pajic, M., Papaemmanuil, E., Paradiso, A., Pearson, J. V., Puente, X. S., Raine, K., Ramakrishna, M., Richardson, A. L., Richter, J., Rosenstiel, P., Schlesner, M., Schumacher, T. N., Span, P. N., Teague, J. W., Totoki, Y., Tutt, A. N. J., Valdés-Mas, R., van Buuren, M. M., van 't Veer, L., Vincent-Salomon, A., Waddell, N., Yates, L. R., Zucman-Rossi, J., Andrew Futreal, P., McDermott, U., Lichter, P., Meyerson, M., Grimmond, S. M., Siebert, R., Campo, E., Shibata, T., Pfister, S. M., Campbell, P. J. & Stratton, M. R. Signatures of mutational processes in human cancer. *Nature* **500,** 415–421 (2013).

49. Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., Kiezun, A., Hammerman, P. S., McKenna, A., Drier, Y., Zou, L., Ramos, A. H., Pugh, T. J., Stransky, N., Helman, E., Kim, J., Sougnez, C., Ambrogio, L., Nickerson, E., Shefler, E., Cortés, M. L., Auclair, D., Saksena, G., Voet, D., Noble, M., DiCara, D., Lin, P., Lichtenstein, L., Heiman, D. I., Fennell, T., Imielinski, M., Hernandez, B., Hodis, E., Baca, S., Dulak, A. M., Lohr, J., Landau, D.-A., Wu, C. J., Melendez-Zajgla, J., Hidalgo-Miranda, A., Koren, A., McCarroll, S. A., Mora, J., Lee, R. S., Crompton, B., Onofrio, R., Parkin, M., Winckler, W., Ardlie, K., Gabriel, S. B., Roberts, C. W. M., Biegel, J. A., Stegmaier, K., Bass, A. J., Garraway, L. A., Meyerson, M., Golub, T. R.,

Gordenin, D. A., Sunyaev, S., Lander, E. S. & Getz, G. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–218 (2013).

50. Albuquerque, C., Breukel, C., van der Luijt, R., Fidalgo, P., Lage, P., Slors, F., Leitao, C. N., Fodde, R. & Smits, R. The 'just-right' signaling model: APC somatic mutations are selected based on a specific level of activation of the beta-catenin signaling cascade. *Hum. Mol. Genet.* **11,** 1549–1560 (2002).

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

52. Hirsch, D., Camps, J., Varma, S., Kemmerling, R., Stapleton, M., Ried, T. & Gaiser, T. A new whole genome amplification method for studying clonal evolution patterns in malignant colorectal polyps. *Genes. Chromosomes Cancer* **51,** 490–500 (2012).

53. Xie, T., d' Ario, G., Lamb, J. R., Martin, E., Wang, K., Tejpar, S., Delorenzi, M., Bosman, F. T., Roth, A. D., Yan, P., Bougel, S., Di Narzo, A. F., Popovici, V., Budinská, E., Mao, M., Weinrich, S. L., Rejto, P. A. & Hodgson, J. G. A Comprehensive Characterization of Genome-Wide Copy Number Aberrations in Colorectal Cancer Reveals Novel Oncogenes and Patterns of Alterations. *PLoS ONE* **7,** e42001 (2012).

54. Carvalho, B., Postma, C., Mongera, S., Hopmans, E., Diskin, S., van de Wiel, M. A., Van Criekinge, W., Thas, O., Matthäi, A., Cuesta, M. A. & others. Multiple putative oncogenes at the chromosome 20q amplicon contribute to colorectal adenoma to carcinoma progression. *Gut* **58,** 79–89 (2009).

55. Crabtree, M., Sieber, O. M., Lipton, L., Hodgson, S. V., Lamlum, H., Thomas, H. J. W., Neale, K., Phillips, R. K. S., Heinimann, K. & Tomlinson, I. P. M. Refining the relation between 'first

hits' and 'second hits' at the APC locus: the 'loose fit' model and evidence for differences in somatic mutation spectra among patients. *Oncogene* **22,** 4257–4265 (2003).

56. Obrador-Hevia, A., Chin, S.-F., González, S., Rees, J., Vilardell, F., Greenson, J. K., Cordero, D., Moreno, V., Caldas, C. & Capellá, G. Oncogenic KRAS is not necessary for Wnt signalling activation in APC-associated FAP adenomas. *J. Pathol.* **221,** 57–67 (2010).

57. Yu, C., Yu, J., Yao, X., Wu, W. K., Lu, Y., Tang, S., Li, X., Bao, L., Li, X., Hou, Y. & others. Discovery of biclonal origin and a novel oncogene SLC12A5 in colon cancer by single-cell sequencing. *Cell Res.* (2014).

58. Lamb, J. The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science* **313,** 1929–1935 (2006).

59. Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L. & Pachter, L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7,** 562–578 (2012).

60. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. & Salzberg, S. L. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14,** R36 (2013).

61. Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* **27,** 2325–2329 (2011).

62. Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L. & Pachter, L. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31,** 46–53 (2012).

63. Kim, D. & Salzberg, S. L. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol* **12,** R72 (2011).

64. Molinari, F. & Frattini, M. Functions and Regulation of the PTEN Gene in Colorectal Cancer. *Front. Oncol.* **3,** (2014).

65. Choi, B., Suh, Y., Kim, W.-H., Christa, L., Park, J. & Bae, C.-D. Downregulation of regenerating islet-derived 3 alpha (REG3A) in primary human gastric adenocarcinomas. *Exp. Mol. Med.* **39,** 796–804 (2007).

66. Wu, W. K. K., Sung, J. J. Y., Wu, Y. C., Li, H. T., Yu, L., Li, Z. J. & Cho, C. H. Inhibition of cyclooxygenase-1 lowers proliferation and induces macroautophagy in colon cancer cells. *Biochem. Biophys. Res. Commun.* **382,** 79–84 (2009).

67. Piskol, R., Ramaswami, G. & Li, J. B. Reliable Identification of Genomic Variants from RNA-Seq Data. *Am. J. Hum. Genet.* **93,** 641–651 (2013).

68. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

69. Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34,** 816–834 (2010).

70. Karolchik, D. The UCSC Genome Browser Database. *Nucleic Acids Res.* **31,** 51–54 (2003).

71. Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R. & 1000 Genomes Project Analysis Group. The variant call format and VCFtools. *Bioinformatics* **27,** 2156–2158 (2011).

72. Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* **42,** D199–D205 (2013).

73. Diep, C. B., Kleivi, K., Ribeiro, F. R., Teixeira, M. R., Lindgjaerde, O. C. & Lothe, R. A. The order of genetic events associated with colorectal cancer progression inferred from meta-analysis of copy number changes. *Genes. Chromosomes Cancer* **45,** 31–41 (2006).

74. Bettegowda, C., Sausen, M., Leary, R. J., Kinde, I., Wang, Y., Agrawal, N., Bartlett, B. R., Wang, H., Luber, B., Alani, R. M., Antonarakis, E. S., Azad, N. S., Bardelli, A., Brem, H., Cameron, J. L., Lee, C. C., Fecher, L. A., Gallia, G. L., Gibbs, P., Le, D., Giuntoli, R. L., Goggins, M., Hogarty, M. D., Holdhoff, M., Hong, S.-M., Jiao, Y., Juhl, H. H., Kim, J. J., Siravegna, G., Laheru, D. A., Lauricella, C., Lim, M., Lipson, E. J., Marie, S. K. N., Netto, G. J., Oliner, K. S., Olivi, A., Olsson, L., Riggins, G. J., Sartore-Bianchi, A., Schmidt, K., Shih, l.-M., Oba-Shinjo, S. M., Siena, S., Theodorescu, D., Tie, J., Harkins, T. T., Veronese, S., Wang, T.-L., Weingart, J. D., Wolfgang, C. L., Wood, L. D., Xing, D., Hruban, R. H., Wu, J., Allen, P. J., Schmidt, C. M., Choti, M. A., Velculescu, V. E., Kinzler, K. W., Vogelstein, B., Papadopoulos, N. & Diaz, L. A. Detection of Circulating Tumor DNA in Early- and Late-Stage Human Malignancies. *Sci. Transl. Med.* **6,** 224ra24–224ra24 (2014).

75. Carter, N. P. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat. Genet.* **39,** S16–S21 (2007).

76. Zhao, M., Wang, Q., Wang, Q., Jia, P. & Zhao, Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics* **14,** S1 (2013).

77. Plagnol, V., Curtis, J., Epstein, M., Mok, K. Y., Stebbings, E., Grigoriadou, S., Wood, N. W., Hambleton, S., Burns, S. O., Thrasher, A. J., Kumararatne, D., Doffinger, R. & Nejentsev, S. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics* **28,** 2747–2754 (2012).

78. Sathirapongsasuti, J. F., Lee, H., Horst, B. A. J., Brunner, G., Cochran, A. J., Binder, S., Quackenbush, J. & Nelson, S. F. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics* **27,** 2648–2654 (2011).

79. Browning, S. R. & Browning, B. L. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *Am. J. Hum. Genet.* **81,** 1084–1097 (2007).

80. Scheet, P. & Stephens, M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* **78,** 629–644 (2006).

81. Lin, S., Cutler, D. J., Zwick, M. E. & Chakravarti, A. Haplotype inference in random population samples. *Am. J. Hum. Genet.* **71,** 1129–1137 (2002).

82. Rabiner, L. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. IEEE* **77,** (1989).

83. Sosman, J. A., Kim, K. B., Schuchter, L., Gonzalez, R., Pavlick, A. C., Weber, J. S., McArthur, G. A., Hutson, T. E., Moschos, S. J. & Flaherty, K. T. Survival in BRAF V600–mutant advanced melanoma treated with vemurafenib. *N. Engl. J. Med.* **366,** 707–714 (2012).

84. Lievre, A., Bachet, J.-B., Boige, V., Cayre, A., Le Corre, D., Buc, E., Ychou, M., Bouche, O., Landi, B., Louvet, C., Andre, T., Bibeau, F., Diebold, M.-D., Rougier, P., Ducreux, M.,

Tomasic, G., Emile, J.-F., Penault-Llorca, F. & Laurent-Puig, P. KRAS Mutations As an Independent Prognostic Factor in Patients With Advanced Colorectal Cancer Treated With Cetuximab. *J. Clin. Oncol.* **26,** 374–379 (2008).

85. Pao, W., Wang, T. Y., Riely, G. J., Miller, V. A., Pan, Q., Ladanyi, M., Zakowski, M. F., Heelan, R. T., Kris, M. G. & Varmus, H. E. KRAS Mutations and Primary Resistance of Lung Adenocarcinomas to Gefitinib or Erlotinib. *PLoS Med.* **2,** e17 (2005).

86. Prahallad, A., Sun, C., Huang, S., Di Nicolantonio, F., Salazar, R., Zecchin, D., Beijersbergen, R. L., Bardelli, A. & Bernards, R. Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR. *Nature* **483,** 100–103 (2012).

87. McDermott, U. & Settleman, J. Personalized Cancer Therapy With Selective Kinase Inhibitors: An Emerging Paradigm in Medical Oncology. *J. Clin. Oncol.* **27,** 5650–5659 (2009).

88. Rehm, H. L. Disease-targeted sequencing: a cornerstone in the clinic. *Nat. Rev. Genet.* **14,** 295–300 (2013).

89. Hansen, A. R. & Bedard, P. L. Clinical application of high-throughput genomic technologies for treatment selection in breast cancer. *Breast Cancer Res.* **10,** 11 (2013).

90. Kim, T.-M. Clinical applications of next-generation sequencing in colorectal cancers. *World J. Gastroenterol.* **19,** 6784 (2013).

91. Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S. & Sunyaev, S. R. A method and server for predicting damaging missense mutations. *Nat. Methods* **7,** 248–249 (2010).

92. Miller, L. D., Smeds, J., George, J., Vega, V. B., Vergara, L., Ploner, A., Pawitan, Y., Hall, P., Klaar, S. & Liu, E. T. An expression signature for p53 status in human breast cancer predicts

mutation status, transcriptional effects, and patient survival. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 13550–13555 (2005).

93. Popovici, V., Budinska, E., Tejpar, S., Weinrich, S., Estrella, H., Hodgson, G., Van Cutsem, E., Xie, T., Bosman, F. T., Roth, A. D. & Delorenzi, M. Identification of a Poor-Prognosis BRAF-Mutant-Like Population of Patients With Colon Cancer. *J. Clin. Oncol.* **30,** 1288–1295 (2012).

94. Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., Sonkin, D., Reddy, A., Liu, M., Murray, L., Berger, M. F., Monahan, J. E., Morais, P., Meltzer, J., Korejwa, A., Jané-Valbuena, J., Mapa, F. A., Thibault, J., Bric-Furlong, E., Raman, P., Shipway, A., Engels, I. H., Cheng, J., Yu, G. K., Yu, J., Aspesi, P., de Silva, M., Jagtap, K., Jones, M. D., Wang, L., Hatton, C., Palescandolo, E., Gupta, S., Mahan, S., Sougnez, C., Onofrio, R. C., Liefeld, T., MacConaill, L., Winckler, W., Reich, M., Li, N., Mesirov, J. P., Gabriel, S. B., Getz, G., Ardlie, K., Chan, V., Myer, V. E., Weber, B. L., Porter, J., Warmuth, M., Finan, P., Harris, J. L., Meyerson, M., Golub, T. R., Morrissey, M. P., Sellers, W. R., Schlegel, R. & Garraway, L. A. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483,** 603–307 (2012).

95. Sandmann, T., Kummerfeld, S. K., Gentleman, R. & Bourgon, R. gCMAP: user-friendly connectivity mapping with R. *Bioinformatics* **30,** 127–128 (2013).

96. Zhang, X.-Z., Yin, A.-H., Lin, D.-J., Zhu, X.-Y., Ding, Q., Wang, C.-H. & Chen, Y.-X. Analyzing Gene Expression Profile in K562 Cells Exposed to Sodium Valproate Using Microarray Combined with the Connectivity Map Database. *J. Biomed. Biotechnol.* **2012,** 1–8 (2012).

97. Heinonen, H., Nieminen, A., Saarela, M., Kallioniemi, A., Klefström, J., Hautaniemi, S. & Monni, O. Deciphering downstream gene targets of PI3K/mTOR/p70S6K pathway in breast cancer. *BMC Genomics* **9,** 348 (2008).

98. McArt, D. G., Dunne, P. D., Blayney, J. K., Salto-Tellez, M., Van Schaeybroeck, S., Hamilton, P. W. & Zhang, S.-D. Connectivity Mapping for Candidate Therapeutics Identification Using Next Generation Sequencing RNA-Seq Data. *PLoS ONE* **8,** e66902 (2013).

99. Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., Djoumbou, Y., Eisner, R., Guo, A. C. & Wishart, D. S. DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs. *Nucleic Acids Res.* **39,** D1035–D1041 (2010).

100. Gunther, S., Kuhn, M., Dunkel, M., Campillos, M., Senger, C., Petsalaki, E., Ahmed, J., Urdiales, E. G., Gewiess, A., Jensen, L. J., Schneider, R., Skoblo, R., Russell, R. B., Bourne, P. E., Bork, P. & Preissner, R. SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Res.* **36,** D919–D922 (2007).

101. Smyth, G. K. in *Bioinforma. Comput. Biol. Solut. Using R Bioconductor* 397–420 (Springer, 2005).

102. Varemo, L., Nielsen, J. & Nookaew, I. Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic Acids Res.* **41,** 4378–4391 (2013).

103. Tan, A. C., Naiman, D. Q., Xu, L., Winslow, R. L. & Geman, D. Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics* **21,** 3896–3904 (2005).

104. Damond, J. ktspair. (2013). at <http://www.inside-r.org/packages/cran/ktspair/docs/ordertsp>

105. McLendon, R., Friedman, A., Bigner, D., Van Meir, E. G., Brat, D. J., M. Mastrogianakis, G., Olson, J. J., Mikkelsen, T., Lehman, N., Aldape, K., Alfred Yung, W. K., Bogler, O., VandenBerg, S., Berger, M., Prados, M., Muzny, D., Morgan, M., Scherer, S., Sabo, A., Nazareth, L., Lewis, L., Hall, O., Zhu, Y., Ren, Y., Alvi, O., Yao, J., Hawes, A., Jhangiani, S., Fowler, G., San Lucas, A., Kovar, C., Cree, A., Dinh, H., Santibanez, J., Joshi, V., Gonzalez-Garay, M. L., Miller, C. A., Milosavljevic, A., Donehower, L., Wheeler, D. A., Gibbs, R. A., Cibulskis, K., Sougnez, C., Fennell, T., Mahan, S., Wilkinson, J., Ziaugra, L., Onofrio, R., Bloom, T., Nicol, R., Ardlie, K., Baldwin, J., Gabriel, S., Lander, E. S., Ding, L., Fulton, R. S., McLellan, M. D., Wallis, J., Larson, D. E., Shi, X., Abbott, R., Fulton, L., Chen, K., Koboldt, D. C., Wendl, M. C., Meyer, R., Tang, Y., Lin, L., Osborne, J. R., Dunford-Shore, B. H., Miner, T. L., Delehaunty, K., Markovic, C., Swift, G., Courtney, W., Pohl, C., Abbott, S., Hawkins, A., Leong, S., Haipek, C., Schmidt, H., Wiechert, M., Vickery, T., Scott, S., Dooling, D. J., Chinwalla, A., Weinstock, G. M., Mardis, E. R., Wilson, R. K., Getz, G., Winckler, W., Verhaak, R. G. W., Lawrence, M. S., O'Kelly, M., Robinson, J., Alexe, G., Beroukhim, R., Carter, S., Chiang, D., Gould, J., Gupta, S., Korn, J., Mermel, C., Mesirov, J., Monti, S., Nguyen, H., Parkin, M., Reich, M., Stransky, N., Weir, B. A., Garraway, L., Golub, T., Meyerson, M., Chin, L., Protopopov, A., Zhang, J., Perna, I., Aronson, S., Sathiamoorthy, N., Ren, G., Yao, J., Wiedemeyer, W. R., Kim, H., Won Kong, S., Xiao, Y., Kohane, I. S., Seidman, J., Park, P. J., Kucherlapati, R., Laird, P. W., Cope, L., Herman, J. G., Weisenberger, D. J., Pan, F., Van Den Berg, D., Van Neste, L., Mi Yi, J., Schuebel, K. E., Baylin, S. B., Absher, D. M., Li, J. Z., Southwick, A., Brady, S., Aggarwal, A., Chung, T., Sherlock, G., Brooks, J. D., Myers, R. M., Spellman, P. T., Purdom, E., Jakkula, L. R., Lapuk, A. V., Marr, H., Dorton, S.,

Gi Choi, Y., Han, J., Ray, A., Wang, V., Durinck, S., Robinson, M., Wang, N. J., Vranizan, K., Peng, V., Van Name, E., Fontenay, G. V., Ngai, J., Conboy, J. G., Parvin, B., Feiler, H. S., Speed, T. P., Gray, J. W., Brennan, C., Socci, N. D., Olshen, A., Taylor, B. S., Lash, A., Schultz, N., Reva, B., Antipin, Y., Stukalov, A., Gross, B., Cerami, E., Qing Wang, W., Qin, L.-X., Seshan, V. E., Villafania, L., Cavatore, M., Borsu, L., Viale, A., Gerald, W., Sander, C., Ladanyi, M., Perou, C. M., Neil Hayes, D., Topal, M. D., Hoadley, K. A., Qi, Y., Balu, S., Shi, Y., Wu, J., Penny, R., Bittner, M., Shelton, T., Lenkiewicz, E., Morris, S., Beasley, D., Sanders, S., Kahn, A., Sfeir, R., Chen, J., Nassau, D., Feng, L., Hickey, E., Zhang, J., Weinstein, J. N., Barker, A., Gerhard, D. S., Vockley, J., Compton, C., Vaught, J., Fielding, P., Ferguson, M. L., Schaefer, C., Madhavan, S., Buetow, K. H., Collins, F., Good, P., Guyer, M., Ozenberger, B., Peterson, J. & Thomson, E. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455,** 1061–1068 (2008).

106. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. Voom! Precision weights unlock linear model analysis tools for RNA-seq read counts. (2013). at <http://www.statsci.org/smyth/pubs/VoomPreprint.pdf>

107. Leek, J. T. The tspair package for finding top scoring pair classifiers in R. *Bioinformatics* **25,** 1203–1204 (2009).

108. Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E., Antipin, Y., Reva, B., Goldberg, A. P., Sander, C. & Schultz, N. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discov.* **2,** 401–404 (2012).

109. Cantwell-Dorris, E. R., O'Leary, J. J. & Sheils, O. M. BRAFV600E: Implications for Carcinogenesis and Molecular Therapy. *Mol. Cancer Ther.* **10,** 385–394 (2011).

110. Roepman, P., Schlicker, A., Tabernero, J., Majewski, I., Tian, S., Moreno, V., Snel, M. H., Chresta, C. M., Rosenberg, R., Nitsche, U., Macarulla, T., Capella, G., Salazar, R., Orphanides, G., Wessels, L. F., Bernards, R. & Simon, I. M. Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition: Molecular subtypes in colorectal cancer. *Int. J. Cancer* **134,** 552–562 (2014).

111. Schlicker, A., Beran, G., Chresta, C. M., McWalter, G., Pritchard, A., Weston, S., Runswick, S., Davenport, S., Heathcote, K. & Castro, D. A. Subtypes of primary colorectal tumors correlate with response to targeted treatment in colorectal cell lines. *BMC Med. Genomics* **5,** 66 (2012).

112. Zecchin, D., Boscaro, V., Medico, E., Barault, L., Martini, M., Arena, S., Cancelliere, C., Bartolini, A., Crowley, E. H., Bardelli, A., Gallicchio, M. & Di Nicolantonio, F. BRAF V600E is a determinant of sensitivity to proteasome inhibitors. *Mol. Cancer Ther.* **12,** 2950–61 (2013).

113. Davies, H., Bignell, G. R., Cox, C., Stephens, P., Edkins, S., Clegg, S., Teague, J., Woffendin, H., Garnett, M. J. & Bottomley, W. Mutations of the BRAF gene in human cancer. *Nature* **417,** 949–954 (2002).

114. Lito, P., Rosen, N. & Solit, D. B. Tumor adaptation and resistance to RAF inhibitors. *Nat. Med.* **19,** 1401–1409 (2013).

115. Chapman, P. B., Hauschild, A., Robert, C., Haanen, J. B., Ascierto, P., Larkin, J., Dummer, R., Garbe, C., Testori, A. & Maio, M. Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *N. Engl. J. Med.* **364,** 2507–2516 (2011).

116. Kopetz, S., Desai, J., Chan, E., Hecht, J. R., O'Dwyer, J., Lee, R. G., Nolop, K. B. & Saltz, L. PLX4032 in metastatic colorectal cancer patients with mutant BRAF tumors. *J. Clin. Oncol.* **28,** 3534 (2010).

117. Tanaka, H., Deng, G., Matsuzaki, K., Kakar, S., Kim, G. E., Miura, S., Sleisenger, M. H. & Kim, Y. S. BRAF mutation, CpG island methylator phenotype and microsatellite instability occur more frequently and concordantly in mucinous than non-mucinous colorectal cancer. *Int. J. Cancer* **118,** 2765–2771 (2006).

118. Tran, B., Kopetz, S., Tie, J., Gibbs, P., Jiang, Z.-Q., Lieu, C. H., Agarwal, A., Maru, D. M., Sieber, O. & Desai, J. Impact of BRAF mutation and microsatellite instability on the pattern of metastatic spread and prognosis in metastatic colorectal cancer. *Cancer* **117,** 4623–4632 (2011).

119. Corcoran, R. B., Ebi, H., Turke, A. B., Coffee, E. M., Nishino, M., Cogdill, A. P., Brown, R. D., Della Pelle, P., Dias-Santagata, D., Hung, K. E., Flaherty, K. T., Piris, A., Wargo, J. A., Settleman, J., Mino-Kenudson, M. & Engelman, J. A. EGFR-Mediated Reactivation of MAPK Signaling Contributes to Insensitivity of BRAF-Mutant Colorectal Cancers to RAF Inhibition with Vemurafenib. *Cancer Discov.* **2,** 227–235 (2012).

120. Yang, H., Higgins, B., Kolinsky, K., Packman, K., Bradley, W. D., Lee, R. J., Schostack, K., Simcox, M. E., Kopetz, S., Heimbrook, D., Lestini, B., Bollag, G. & Su, F. Antitumor Activity

of BRAF Inhibitor Vemurafenib in Preclinical Models of BRAF-Mutant Colorectal Cancer. *Cancer Res.* **72,** 779–789 (2012).

121. Mao, M., Tian, F., Mariadason, J. M., Tsao, C. C., Lemos, R., Dayyani, F., Gopal, Y. N. V., Jiang, Z.-Q., Wistuba, I. I., Tang, X. M., Bornman, W. G., Bollag, G., Mills, G. B., Powis, G., Desai, J., Gallick, G. E., Davies, M. A. & Kopetz, S. Resistance to BRAF Inhibition in BRAF-Mutant Colon Cancer Can Be Overcome with PI3K Inhibition or Demethylating Agents. *Clin. Cancer Res.* **19,** 657–667 (2013).

122. Domcke, S., Sinha, R., Levine, D. A., Sander, C. & Schultz, N. Evaluating cell lines as tumour models by comparison of genomic profiles. *Nat. Commun.* **4,** (2013).

123. Baggerly, K. Disclose all data in publications. *Nature* **467,** 401–401 (2010).

Vita

Francis “Anthony” San Lucas is the son of Roy and Maria San Lucas, husband of Tien San Lucas and father of Sarai, Abigail and Natale San Lucas. He and his wife are blessed to have another child due in October 2014. Anthony grew up in Houston, TX where he graduated from Langham Creek High School and developed a strong interest in engineering and computer science. He attended The University of Texas at Austin obtaining a Bachelor’s Degree in Mechanical Engineering and City College of the City University of New York where he obtained a Master’s Degree in Computer Science. He has worked as a Project Engineer for Chevron Chemical Company, as a Researcher/Software Engineer on projects for the United States Department of Defense Air Force Research Laboratories, as an Applications Developer for JP Morgan Chase Investment Banking Technologies and as a Lead Scientific Programmer for the Human Genome Sequencing Center at Baylor College of Medicine. Anthony’s research interests are focused on helping doctors make more molecularly informed treatment decisions for individuals with chronic illnesses such as heart disease, diabetes and cancer. He expects to get his Doctor of Philosophy Degree in Bioinformatics from The University of Texas at Houston Graduate School of Biomedical Sciences in August 2014. He will become a Postdoctoral Fellow in the Department of Translational Molecular Pathology at MD Anderson Cancer Center in September 2014.