QUANTITATIVE APPROACHES TO UNDERSTANDING CANCER GENOMES

by

Michele Caroline Gornick

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Human Genetics)
in The University of Michigan
2011

Doctoral Committee:

Professor Stephen B. Gruber, Chair
Professor Eric Fearon
Associate Professor Julie Douglas
Assistant Professor Jun Li
Associate Professor Gad Rennert, Technion-Israel Institute of Technology

© Michele C. Gornick 2011

DEDICATION

To my ‘Godmothers’

Cathy Greco

Judith Link

Anjene Addington

Wendy Sharp

Laura Rozek

Julie Douglas

…

ACKNOWLEDGEMENTS

I'm not sure how to begin to prioritize the people who deserve to be acknowledged in my life, but this is my best attempt. First and foremost, I want to thank Levi, my best friend and husband, whom I love and respect more than all of the tennis balls in the world. Next, I would like to thank my parents, who taught me the value of hard work and perseverance. Equally important are my sisters, Mira Funari and Mary Gornick, whom I admire as strong women and strive to be a better person every day because of their example. My little brother, who understands my desire for knowledge, and has always been a scholarly inspiration. Last but not least, my beautiful nieces and nephew, who always bring a smile to my face and for whom I try to be a role model for each day.

I must also mention all of my friends, coworkers and colleagues who have played an instrumental role in this chapter of my life. It would be a lie to say that this dissertation was the product of my efforts alone. I would not have had the courage to pursue this course of study without the friendship and guidance from those who came before me, so many thanks to Nicole Scott and Cris Van Hout. With that being said, it wouldn't have been possible to get through this process without the long hours of team work, intellectual conversation and unconditional support from my friends and future colleagues Kristen Stevens, Kaanan Shah, Valerie Schaibley and Casey Herron.

Many thanks to all of the past and present members of the Gruber lab who have contributed to these projects including genetic counselors, support staff, collaborators and study participants. Specifically, I would like to acknowledge Jishu Xu, Sana Shakour, Jessica Everett, Victoria Raymond and Sherry Taylor.

In closing, I would like to offer many thanks to my dissertation committee members: Stephan B. Gruber, Eric Fearon, Julie Douglas, Jun Li and Gad Rennert.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

LIST OF ABBREVIATIONS

Abbreviation	Definition
BAF	B-allele frequency
CBS	Circular binary segmentation
CIMP	CpG island methylator phenotype
CIN	Chromosomally instable (same as MSS)
CNA	Copy number alteration
CN-LOH	Copy neutral loss of heterozygosity
CNV	Copy number variation
CRC	Colorectal cancer
EM	Expectation maximization
FAP	Familial adenomatous polyposis
GWAS	Genome-wide association study
HMM	Hidden Markov model
HMPS	Hereditary mixed polyposis syndrome
HNPCC	Hereditary nonpolyposis colorectal cancer
IBD	Identity by descent
IBS	Identity by state
LD	Linkage disequilibrium
LOH	Loss of heterozygosity
LRR	LogR ratio
MAF	Minor allele frequency
MECC	Molecular Epidemiology of Colorectal Cancer
MSI	Microsatellite instable
MSS	Microsatellite stable
MAP	MYH-associated polyposis
NGS	Next-generation sequencing
OMIM	On-line mendelian inheritance in man
PCR	Polymerase chain reaction
ROH	Run/region of homozygosity
SNP	Single nucleotide polymorphism

ABSTRACT

Recent advances in technology have enabled the systematic, genome-wide analysis of cancer genomes, providing greater insight into the genetic basis of cancer development and a deeper understanding of the human genome. The focus of the current work is to identify genomic alterations potentially conferring risk for developing colorectal and breast cancers by performing genome-wide analysis with single nucleotide polymorphism (SNP) genotyping and next-generation sequencing (NGS) platforms.

My first dissertation project involves deeply sequencing the genomes of individuals from a single family to identify novel mutations in hereditary mixed polyposis syndrome, a rare form of colorectal cancer with no known genetic basis. A novel candidate gene, *ZNF426*, was identified and decreased expression was confirmed in tumors from affected individuals.

The second part of my dissertation evaluates methods for detection of somatic copy number alterations in colorectal cancer on chromosome 18 and the application of statistical methods for utilizing poor quality tumor data. Using genotyping and expression data from tumors, a variety of structural alterations were identified on chromosome 18. Additionally, I demonstrated the utility of applying new statistical methods to identity copy number alterations in array data with high background noise.

The goal of my third project was to evaluate the contribution of consanguinity to breast cancer risk in Arab women without mutations in the *BRCA1* and *BRCA2* genes. The hypothesis in this study is that an increase in autosomal recessive genes responsible for genetic susceptibility to breast cancer is expected among families with consanguinity due to the increase

in probability of sharing alleles identical-by-descent. Six unrelated individuals with breast cancer shared a 200kb overlapping region of homozygous SNPs on chromosome 9q332-33.3, which harbors an important candidate gene for cancer risk, *LHX2*.

Whole-genome analysis allows for greater depth and higher throughput sequencing at lower costs, adding a new dimension to our understanding of cancer genetics. Future progress in these technologies and bioinformatics methods will improve the costs, sensitivity and accuracy of detecting mutations.

CHAPTER 1

Introduction

1.1 The human genome era

The human genome sequence first published in 2001 contains approximately 2.85 billion nucleotides, covering 99 percent of the genome with an error rate of 1 per 100,000 bases (Lander et al 2001). Approximately 20,000 protein-coding genes, along with thousands of non-coding RNA, regulatory sequences, enhancers, and non-coding DNA were identified. The decade following this milestone has revealed the potential of genomic maps and catalogues for biomedical research. For instance, we now can use genome-wide information to describe the structure and organization of chromosomes, methylation patterns and the extent of linkage disequilibrium between genetic markers.

The last decade has brought an influx of discoveries of disease-associated mutations in the human genome, however, these findings have explained only a small part of disease risk (Antonarakis and McKusick 2000). Although the Human Genome Project helped to propel progress forward, limitations in technology and cost restricted its application. The Human Genome Project used, for the most part, the same sequencing method involving electrophoretic separation of mixtures of randomly terminated extension products employed by Sanger in 1977, although it was dramatically improved with fluorescently labeled terminators and automated laser detectors. Optical imaging of the human reference genome using capillary-based

sequencing allowed for high quality runs. However, a major drawback of this technology is that sequence reads are on the order of 1,000 base pairs, providing a challenge to *de novo* genome sequence assembly from such short reads. For creation of the human reference, efforts were focused on placing sequenced reads into genome scaffolds of modest size (500 bp-11.5Mb), followed by integration of these scaffolds into the full, contiguous sequence (Ozawa et al 2004).

The time following the assembly of the human genome marked a transition in the field of medical genetics from a focus on single-gene disease to identifying common genes associated with complex diseases such as cancer, diabetes, and heart disease. The predominant idea surrounding complex disease was that genetic risk for common diseases is likely due to disease-predisposing alleles with relatively high frequencies and therefore only a few predominating disease alleles exist at each of the major underlying disease loci (Reich and Lander 2001). In an effort to catalogue this common variation in European, East Asian and African populations, the International Haplotype (HapMap) Project focused on defining linkage disequilibrium patterns across these genomes and found that ~500,000–1,000,000 SNPs could capture nearly 90% of common genetic variation (2005, Daly et al 2001, Reich et al 2001, Weiss and Clark 2002). These finding spurred the advent of genome-wide, array-based technology. The development of genotyping arrays (SNP arrays) allowed for the ascertainment of hundreds of thousands to a million genetic markers simultaneously, out of which emerged the Genome-Wide Association Study (GWAS). GWA studies are based on the idea that disease related variants are more common in the patients with a given disease than in healthy people (Wang et al 2005). In a GWAS, the expected result is to detect a sequence change in a SNP on a genotyping array that is statistically associated with a disease. However, the SNP itself is not necessarily the actual cause of the disease, rather a signal that there is some nearby functional gene variation. Therefore, the

study results are often just the beginning of the search for a disease-associated gene or genes (Wang et al 2010a). Recent studies have suggested that the signals in GWA studies may not always be pointing to a few common gene variants as previously thought, but instead to many rare variants each of which causes relatively few cases and may be located a distance away from the site identified in a GWAS (Dickson et al 2010).

An alternative to the GWAS is whole-genome sequencing, a method capable of identifying both common and rare variants. Whole-genome sequencing has become increasingly more practical through innovations in the amplification, sequencing, and detection of genetic variation. These second-generation technologies marketed in recent years have vastly expanded the capabilities and applications of genome analyses. Polymerase chain reaction (PCR) based amplification has now decreased the amount of time needed to perform whole genome scans from years with the Sanger technology to a few weeks with the current high-throughput platforms. With massively parallel sequencing, attention has now focused on comprehensive exome-wide and genome-wide sequencing in large numbers of samples. Along with the technology's rapid turnaround time comes an exponential increase in genetic data, which necessitates parallel progress in computing tools and bioinformatics for data handling and interpretation.

The motivation behind my dissertation is to examine the utility of three different quantitative approaches for using genome-wide platforms to understand the role of genetic variation in cancer. The second chapter applies next-generation sequencing methods to identify extremely rare, highly penetrant, novel variants in a family with an atypical form of colorectal cancer. In chapter three, I explore the role of structural variation in colorectal tumors using high-density genome-wide SNP arrays. Finally, in my fourth chapter, I identified large runs of

homozygosity enriched with autosomal recessive loci in individuals with a family history of consanguineous and breast cancer.

1.2 Insights into colorectal cancer biology

Cancer is essentially a condition of aberrant genetic programming, where changes in the genomic sequence can potentially alter the structure, function, and potentially the expression of proteins that control cellular growth and differentiation processes. This dysregulation can lead to cellular transformation, and ultimately, tumor formation. Genetic variation in a DNA sequence may be inherited in the germline or somatically acquired. Although inherited mutations may lead to familial forms of cancer, the vast majority of neoplastic disease is sporadic and arises from the progressive gain of somatic alterations. These somatic alterations often occur on a background of heritable susceptibility alleles, leading to an increased risk of developing cancer (Knudson 1993).

Colorectal cancer (CRC) is the third most prevalent cancer with nearly 1 million cases worldwide (Ferlay et al 2010). Genetic models for colorectal carcinogenesis have previously been well described based on molecular, mutational, and epigenetic patterns, including the chromosomally instable (CIN), microsatellite instability (MSI) and CpG island methylator phenotype (CIMP) pathways (Peinado et al 1992, Thibodeau et al 1993, Vilar and Gruber 2010, Vogelstein et al 1988). Mutations in central genes in these pathways have been identified in a small number of rare, highly penetrant familial syndromes, accounting for less than 5% of all CRC (Goss and Groden 2000, Kemp et al 2004, Marra and Boland 1995). Major CRC genetic syndromes include familial adenomatous polyposis (FAP), MUTYH-associated polyposis (MAP), and Lynch syndrome (hereditary nonpolyposis colorectal cancer or HNPCC). Rare syndromes include hamartomatous polyposis conditions (Peutz-Jeghers syndrome (PJS) and

juvenile polyposis syndrome (JPS)). FAP results from germline mutations in the tumor suppressor *APC* gene on chromosome 5q21, which encodes a protein that is a negative regulator in the Wnt signal transduction pathway (Benchabane and Ahmed 2009). MUTYH-associated polyposis (MAP) is an autosomal recessive disorder characterized by adenomatous polyps of the colorectum and a very high risk of CRC. MAP is associated with biallelic mutations in the *MUTYH* gene located on chromosome 1p, which encodes a protein in the DNA base excision repair pathway whose impaired function leads to increased G:C to T:A transversions (Al-Tassan et al 2002). Lynch syndrome, is an autosomal dominant condition associated with mutations in DNA mismatch repair (MMR) genes (Lynch et al 2009). Peutz-Jeghers syndrome (PJS) is an autosomal dominant disorder which leads to a predisposition to various malignancies (gastrointestinal, pancreatic, lung, breast, uterine, ovarian and testicular tumors) (Kopacova et al 2009). The majority of patients whom meet the clinical diagnostic criteria for PJS have a mutation in the tumor suppressor *STK11* gene located at 19p13.3 (Beggs et al 2010, Jenne et al 1998). Juvenile polyposis syndrome is a rare, early-onset disease, characterized by the presence of hamartomatous polyps throughout the gastrointestinal tract. It is estimated that 15%–20% of JPS patients carry autosomal dominant mutations in the *SMAD4/DPC4* on chromosome 18q21.1, and 25%–40% of the patients carry autosomal dominant mutations in the gene encoding bone morphogenetic protein receptor 1A (*BMPR1A*) on chromosome 10q22-23 (Brosens et al 2007, Howe et al 1998).

Aside from inherited genetic susceptibility, other CRC associated risk factors include the presence of large serrated polyps (serrated adenomas and hyperplastic polyps), a diet rich in total fat and meat content, cigarette smoking, male gender, the use of nonsteroidal anti-inflammatory drugs, alcohol intake, a sedentary lifestyle, high body mass index (BMI), and abdominal obesity

(Gonzalez and Riboli 2010, Hiraoka et al 2010, Hoffmeister et al 2010, Larsson and Wolk 2006). High intake of folate, vitamins and dietary fiber, colonoscopy, and postmenopausal hormone use have been associated with decreased CRC risk (Dahm et al 2010, Hildebrand et al 2009, Kim et al 2010, Rennert et al 2011).

Family history of non-syndromic CRC is associated with a two-fold increase in risk (Carstensen et al 1996) and is estimated to account for ~35% of CRC (Tenesa and Dunlop 2009). Familial non-syndromic CRC is thought to be attributed to more common, low- to moderate-penetrance mutations, mainly based on the identification of variants through candidate gene studies. Examples of these genetic variants include the I1307K mutation in the *APC* gene (Laken et al 1997) and variation in *TGFBR1 (de Jong et al 2002).* The genome-wide association study (GWAS) design has now been widely implemented in colorectal cancer, resulting in the identification of several new loci associated with CRC such as 8q23.3, 8q24, 10p14, 11q23, 15q13, and 18q21 (Broderick et al 2007, Gruber et al 2007, Houlston et al 2008, Tenesa and Dunlop 2009, Zanke et al 2007). The majority of these loci are not located within or near known genes, and the biological relevance of some of these signals is unclear (Tenesa and Dunlop 2009) (Sotelo et al 2010). Unfortunately, GWAS require large numbers of patients and are very costly. Although many common genetic variants have now been statistically associated with CRC, it appears that the majority common variants have turned out to explain only a small fraction of the genetic risk (Peters et al 2011).

Due to the limited success of the GWAS design to identify predictive variants for CRC development and the recent advancement in massively parallel sequencing technology, a newly emerging hypothesis in the field of complex diseases is that both common and rare risk variants may contribute to disease (Ley et al 2008). In the second chapter of this dissertation, I describe a

study of novel genetic variants identified in a rare familial type of colorectal cancer, hereditary mixed polyposis syndrome (HMPS). HMPS is characterized by polyps of mixed adenomatous/hyperplastic/atypical juvenile histology, and this polyp phenotype is autosomal dominantly inherited. Polyp formation will eventually lead to colorectal cancer development without appropriate monitoring and timely, preventative intervention. Little is still known about the etiology and the genetic basis of this condition.

1.3 Inheritance of breast cancer susceptibility alleles

Breast cancer is the most common cancer affecting women, accounting for approximately 30% of all incident cancer cases among women in 2009 and causing about 15% of female cancer-related deaths in the United States alone (Hayat et al 2007, Howlader et al 2010). The majority of breast cancer diagnosis occurs late in life (postmenopausal), and many environmental risk factors have been associated with the development of breast cancer including height, benign breast disease, older age at first birth, younger age at menarche, older age at menopause, high estrogen levels, ionizing radiation, and high BMI. Despite these strong environmental risk factors, the most significant predictor of breast cancer risk is family history (Couch and Weber 1996).

Highly penetrant mutations inherited in at least two susceptibility genes, *BRCA1* and *BRCA2*, are associated with an increased risk of developing breast cancer. *BRCA1* and *BRCA2* are normally involved in regulating cell growth and DNA repair and are crucial for normal cell development and differentiation (Miki et al 1994, Wooster et al 1995). Mutations in these two genes together account for approximately two thirds of familial breast cancer, or roughly 5% of all breast cancer cases. Approximately sixty percent of women who inherit mutated forms of

these genes will develop breast cancer in their lifetime, usually at relatively early ages, and women with *BRCA1* mutations also have a high risk of developing ovarian cancer (Futreal et al 1994, Metcalfe et al 2010). Studies have shown that one third of families with *BRCA2* mutations present with multiple cases of breast cancer alone, but more than 80% of families with *BRCA1* mutations have both breast cancer and ovarian cancer (Schwartz et al 2011).

In addition to *BRCA1* and *BRCA2*, other mutations have been associated with breast cancer susceptibility. A single nucleotide deletion in *CHEK2* (1100delC), a cell cycle checkpoint kinase also implicated in DNA repair, results in protein truncation and is estimated to confer a two-fold increase in breast cancer risk in women without *BRCA1* or *BRCA2* mutations (Meijers-Heijboer et al 2002). *ATM*, a DNA double-strand break repair protein, has recently been shown to confer a relative risk of 2.37 in heterozygous carriers of mutations (Renwick et al 2006). Other rare syndromes such as Li-Fraumeni Syndrome, Peutz-Jeghers Syndrome, and Cowden/PTEN Hamartoma Syndrome are all associated with a high lifetime risk of breast cancer, but they account for a very small fraction of inherited susceptibility to breast cancer (Garber and Offit 2005). The remaining ~75% of familial breast cancer cases not accounted for by these genes await explanation, emphasizing the need to identify mutations that are likely to exist in the population.

Hereditary breast cancer may account for at least 15-20% of all breast cancer in Ashkenazi Jewish women (Tonin et al 1996). The *BRCA1* mutation 185delAG is commonly seen in breast and ovarian cancer families. Preliminary studies have shown that about 1% of Ashkenazi Jews carry the 185delAG mutation. This genetic alteration has been estimated to account for 20% of cases of breast cancer and 39% of ovarian cancer diagnosed in Jewish women before age 50. In addition, two other *BRCA1* mutations, 188del11 and 5382insC, seem to

be overrepresented in the Ashkenazi Jewish population. Also, a specific mutation in *BRCA2* (6174delT) has been identified in Ashkenazi Jewish women with early onset breast cancer. This mutation is found in the population at similar rates to the 185delAG *BRCA1* mutation, and the 5382insC mutation is about half as frequent as the 185delAG. Approximately 1 in 40 Ashkenazi women may carry one of these three BRCA mutations (Abeliovich et al 1997, Kotsopoulos et al 2007, Rennert et al 2007).

Consistent with the previous observation of a difference in the distribution of demographic variables and risk factors between Ashkenazi Jewish and non-Ashkenazi women, including the presence of distinct risk haplotypes within each ethnic subgroups of *HMMR* (Pujana et al 2007); younger ages at diagnosis, larger primary tumor size, and lower 5-year survival rate have also been reported among Arab women (El Saghir et al 2006). This suggests that the mechanism of inheritance associated with breast cancer risk may vary between Jewish and non-Jewish women. Similar to inherited genes, cultural behaviors are passed on from generation to generation within families. The particular focus of this research is consanguinity in Arab populations. Breast cancer is a result of the interaction of behavioral, cultural, and environmental factors, or it may occur in combination with certain genetic mutations.

In the fourth chapter of my dissertation, I investigate the possibility of a common, recessive locus that may account for additional susceptibility to familial breast cancer apart from *BRCA1* and *BRCA2* in Arab and Jewish women with a family history of consanguinity and a sibling with breast cancer. My hypothesis is that the increase in probability of sharing alleles identical-by-descent expected among families with consanguinity will lead to an increase in autosomal recessive genes responsible for the genetic susceptibility to breast cancer. To examine this hypothesis I performed homozygosity mapping in consanguineous breast cancer patients

with an affected sibling and unaffected mothers, to identify shared genomic regions of disease susceptibility.

1.4 The future of the cancer genome

Numerous efforts to characterize cancer genomes have emerged in recent years (Berger et al 2010, Beroukhim et al 2010, Boehm et al 2007, Chapman et al 2011, Leary et al 2008, Parsons et al 2008). These initiatives have relied heavily on large-scale genotyping and now sequencing approaches to identify potentially pathogenic point mutations and somatic alterations. With this large amount of high-quality data, the discovery of interrelated pathways and analysis approaches has become increasingly complicated. Quantitative approaches to understanding the genome not only allow for the discovery of novel variants associated with disease, but also to better understand the organization and mutational profile of cancer genomes. The following chapters focus on applying quantitative genetics methods, theory, and informatics in the analysis of genomic data to detect genes that confer disease risk in colon and breast cancers.

CHAPTER 2

Whole genome sequencing to identify candidate genes for hereditary mixed polyposis syndrome

2.1 Introduction

Hereditary forms of polyposis associated with colorectal cancer risk, including Familial Adenomatous Polyposis, Juvenile Polyposis and Peutz-Jeghers syndrome, are typically straightforward to characterize and distinguish based on histological examination and recognizable clinical phenotypes (Whitelaw et al 1997). However, difficulty can arise when an individual has polyps of more than one histological type or individual polyps with overlapping histological features. Distinguishing between patients with an atypical form of a known polyposis syndrome and a distinct clinical disorder is difficult (Jass 2000).

An example of an uncertain hereditary form of polyposis is hereditary mixed polyposis syndrome (HMPS, OMIM ID %601228), which is characterized by a mixture of atypical juvenile polyps, hyperplasic polyps, sessile serrated adenomas and an increased risk of colorectal cancer. Polyps appear to be inherited in an autosomal dominant fashion. The putative susceptibility locus initially mapped to 15q13-14 by linkage analysis in three Ashkenazi Jewish families (Jaeger et al 2003, Thomas et al 1996). However, the genetic basis of this syndrome is not well-understood, and no mutation or associated variants have been identified to

date.Syndromes that appear to be Mendelian in nature are a valuable resource for the study of genes and gene function. However, for various reasons, they are often not amenable to traditional approaches to identifying risk variants such as linkage analysis. For example, some traits are so rare that only a small number of cases are available. Locus heterogeneity can also complicate standard genetic approaches such as linkage where unrelated genes cause a single disorder. Sporadic cases due to *de novo* variants and phenocopies can also complicate the search for Mendelian genes. For extremely rare disorders, like hereditary mixed polyposis syndrome, it is often the case that only affected siblings in one family or a few unrelated cases from different families are available for investigation.

An alternative to this analytic problem is to deeply sequence the genome of affected individuals since it is inherently more robust for detecting variants in heterogeneous disorders. Whole-genome next-generation sequencing techniques now feasibly offer opportunities to study extremely rare disorders by deeply sequencing the entire genome of affected individuals. Several recent examples of using this approach to determine the genetic basis of a Mendelian disorder have recently been published, demonstrating the utility of next generation sequencing (Kuhlenbaumer et al 2010, Ng et al 2010a, Ng et al 2010b) to discover mutations in genes that were refractory to other genetic study designs.

I hypothesize that rare variants with strong deleterious effects are responsible for hereditary mixed polyposis. The current study approach operated under the observation that rare monogenic disorders are often due to mutations that are highly penetrant, extremely rare, and strongly disrupt normal biology. In this chapter, I aim to characterize the role that these potentially deleterious mutations play in hereditary mixed polyposis syndrome by deeply sequencing affected individuals from a single family. Additionally, I aim to understand how

these novel variants may influence tumor phenotype using quantitative real time PCR to assess expression levels of candidate genes. The working hypothesis is that novel, deleterious changes in coding regions will alter expression in genes that play a role in tumorigenesis.

2.2 Subjects and Methods

2.2.1 Subjects

Family 202 was recruited as part of the original investigation of the genetic basis for hereditary mixed polyposis through the University of Michigan Cancer Genetics Clinic in 2003 (Figure 2.1). Three additional unrelated individuals with a putative diagnosis of hereditary mixed polyposis were identified at the Ohio State University and used as replication samples.

2.2.1.1 The University of Michigan Cancer Genetics Clinic—Family 202

The proband in the family seen at the University of Michigan was a 67-year-old man (Figure 2.1: II-1) who was diagnosed with colon cancer in 2002. He underwent a right hemi-colectomy and was found to have liver metastases. He was then treated with palliative chemotherapy. A significant family history was reported, including a daughter who was diagnosed with colon cancer at age 31 and a father who was reported to have had pancreatic cancer at age 52 (Figure 2.1: III-3 and I-1). The proband's personal and family history was suggestive of hereditary nonpolyposis colorectal cancer (HNPCC) but did not technically meet clinical diagnostic criteria for this condition. The proband's tumor was sent for genetic testing to evaluate mismatch repair defects. Microsatellite instability testing performed at the Mayo Clinic revealed 0 out of 5 makers were instable (BAT25, BAT26, D2S123, D5S346, and D17S250) and immunohistochemistry for 3 mismatch repair genes (*MLH1, MSH2 and MSH6*) was intact,

revealing no features of HNPCC. Also of note, the proband reported no family history of consanguinity or Ashkenazi Jewish ancestry. The proband died in 2007 due to complications from colorectal cancer.

The proband's 46-year-old son (Figure 2.1: III-5) was evaluated at the Cancer Genetics Clinic in 2009 regarding his personal history of colon polyps and family history of colon cancer. Due to his family history of colon cancer, the son has had colonoscopies every 4-5 years. In the past, he had several hyperplastic polyps of no significance, however in 2007, two sessile polyps in the sigmoid colon were removed and read by a pathologist (JG) to be sessile serrated adenoma. Genetic testing for microsatelitte instability and immunohistochemistry for mismatch repair defects was performed on the sessile serrated adenoma tissue in order to provide additional diagnostic information which may help to determine HNPCC as a possible diagnosis for this family. In addition, the patient provided a blood sample for familial adenomatous polyposis (FAP) and *MYH*-associated polyposis (MAP) testing. All results were negative; therefore HNPCC, FAP and MAP were excluded from his diagnosis. The patient's personal history of hyperplastic polyps and sessile serrated adenoma along with a family history of colon cancer arising in the setting of a mixed polyp led to a clinical and pathologic diagnosis of hereditary mixed polyposis syndrome.

Most recently, the healthy 50-year-old daughter (Figure 2.1: III-4) of the proband was seen for screening at the Cancer Genetic Clinic at the University of Michigan in 2011. Colonoscopy found a 3 mm sessile serrated adenoma in the descending colon as well as a 12 mm flat polyp in the descending colon. Tumor samples as well as a blood sample are currently out being evaluated for genetic testing of mutations in HNPCC, FAP and MAP.

It is important to note that the daughter (Figure 2.1: III-3) of the proband who was

diagnosed with colon cancer at age 31 has not yet been seen in the Cancer Genetics Clinic. This individual did provide a blood sample and consented to participate in the Cancer Genetic Family Registry, which allowed for genetic testing of FAP and MAP. Mutations in *APC* and *MYH* were subsequently ruled out. Additionally, immunohistochemistry for 3 mismatch repair genes (*MLH1, MSH2* and MSH6) was found to be intact. Tumor blocks from her prior colon cancer are is no longer available.

2.2.1.2 The Ohio State University Clinical Cancer Genetics Program

We were able to obtain genomic DNA and formalin fixed tumor block from three unrelated individuals collected at the Ohio State University Medical Center through the Clinical Cancer Genetics Program. All three individuals were consented under the University of Michigan Cancer Genetics Clinic protocol. One individual was of Ashkenazi decent and had a history of mixed hyperplastic/adenomatous polyps (Patient ID 8328-00) and the other two individuals were of European descent and had a history of hyperplastic polyps and sessile serrated adenomas (Patient ID 8333-00, ID8293-00). For additional information see Table 2.1.

2.2.2 Massively Parallel Sequencing

Massively parallel sequencing (also referred to as 'next-generation' or more recently 'second-generation' sequencing) is a high-throughput method of sequencing PCR-amplified DNA fragments. There are several commercially available platforms. The current study used the Illumina Genome Analyzer (Ansorge 2009, Suzuki et al 2011) to sequence two affected related individuals (Figure 2.1: II-2, III-5) with clinically and pathologically diagnosed HMPS at the University of Michigan Sequencing Core. The first step in whole-genome paired-end sequencing

is to create a library. The Encore™ NGS Library System I (©2010 NuGen Technologies Inc.) was used to build DNA libraries with insert sizes from 200-500 bp for paired-end sequencing. The kit provides reagents for repairing the ends of DNA that have been fragmented by nebulization. The ends are then repaired with a combination of fill-in reactions and exonuclease activity to produce blunt ends. Next, an 'A' base is added to the blunt ends followed by ligation of Illumina Paired-End Sequencing adapters. These adapters contain two unique sequencing primer hybridization sites that are used to attach the fragments to the flow cell. Additional sequences complementary to the oligonucleotides in the flow cell are added to the adapter sequences with tailed PCR primers. This is followed by gel-based size selection and purification to create libraries for cluster generation. Specifically, the Illumina Genome Analyzer platform uses what is referred to as a 'bridge-amplification,' a PCR-based technique in which the fragmented and ligated DNA samples are hybridized to the complementary primer bound to one of eight channels or 'lanes' on a microscope slide. Next, the fragments are denatured and the original bound fragment, as well as the newly copied strand, anneal to an immobilized primer complementary to the free adapter. This process of 'bridge amplification' is cyclically repeated to generate clusters of single-stranded DNA copied while anchored to a solid surface. After the amplification is performed, sequencing by synthesis is done using fluorescently labeled reversible terminator nucleotides. A laser is passed over the surface, and the emitted light and its physical position on the slide are detected. Finally the label and 3'-OH terminator are removed allowing for synthesis to continue (Figure 2.2).

2.2.2.1 Statistical and Bioinformatic Methods

The development of faster sequencing platforms has resulted in the generation of

overwhelming amounts of sequence data that necessitate computational tools that are not only efficient and robust but also difficult to use for even computationally sophisticated teams. These powerful computational methods are needed in nearly every step of processing and analyzing the data, including alignment of multiple sequenced DNA fragments and their reconstruction into an accurate contiguous sequence to single nucleotide and structural variant detection.

Several tools have been created to work with next-generation sequence data since its recent inception, from read based aligners like MAQ (Li et al 2008a), BWA (Li and Durbin 2009), and SOAP (Li et al 2008b), to single nucleotide polymorphism and structural variation detection tools like BreakDancer (Chen et al 2009) and VarScan (Koboldt et al 2009). Because the time between the advent of new sequencing technology and its implementation has been so rapid, there exists a large development gap between sequencing output and analysis results. Not only are there many challenges associated with implementing analysis tools that can answer researcher-specific questions and deal with mass amounts of data, the data in itself is laborious and difficult to interpret.

Because of this difficulty, I used the Genome Analysis Toolkit (GATK), a structured programming framework designed to aid in the efficient and robust analysis tools for next-generation DNA sequence data (McKenna et al 2010) to align data, recalibrate quality control scores, locally realign and call SNP/indels. The GATK environment is a platform-independent Java 1.6 framework in which analysis tools are constructed so that the underlying framework can easily parallelize and distribute processing to manage massive computing infrastructures in a scalable way. The core system uses the binary alignment version of sequence alignment/map (SAM) format, called binary alignment/map (BAM) (http://picard.sourceforge.net). GATK also provides a suite of tools including depth of coverage analyzers, a quality score recalibrator, a

local realigner, and a SNP/indel caller for working with human medical sequencing projects (Figure 2.3).

2.2.2.2 Alignment

Alignment and assembly are the first steps in processing the sequence data once the raw reads are obtained. For this analysis, the UCSC assembly hg19 was used as the reference sequence. Many short-read aligners are publicly available. In the current study I used BWA, which is based on Burrows-Wheeler Transform plus auxiliary data structures enabling a balance between fast performance and accuracy. However, alignment is often complicated by repetitive regions or sequencing errors. Because of this inherent complexity in the genome, three main steps are used to better facilitate aligning the raw data, which include base quality score recalibration, local realignment around indels and marking duplications (Figure 2.4). A total of 404,953,503 (III-5) and 239,525,568 (II-2) paired-end reads (35-120bp each) were able to be mapped to the reference sequence (mappable reads), resulting in an average of 10x depth of coverage of the genome.

2.2.2.3 Variant Calling and Annotation

After the raw data processing was complete, the next step of the GATK pipeline, which includes the unified genotyper, was to identify sites in the sequenced individuals that are statistically non-reference. The UnifiedGenotyper was used to calculate genotype likelihoods, only using reads with a minimum mapping quality of at least 10 and fewer than 4 mismatches within 40 bp. Candidate sites were called with a per site prior probability of a polymorphism of 0.001. The E-M algorithm was used to estimate the allele frequency at each site by maximum

likelihood, where each potential variant site must have a posterior probability greater than 0.9 (PHRED scaled quality score of 10). The UnifiedGenotyper produces a Variant Call Format (VCF) file that contains the variant sites discovered across samples and the genotypes assigned to each sample/site. The data was then further processed and carefully filtered using investigator-set criteria. Using several measures of quality control provided by BWA, in addition to standard genetic measures of quality control (Hardy-Weinberg Equilibrium of variants in dbSNP, Mendelian inconstancies between the father and son, transition to transversion ratio of 2), the following quality control criteria of ≥ 4x coverage, depth of coverage (DP) ≤ 360, heterozygote allele balance (ref/(ref+alt)) (AB) ≤ 75% reference, number of covering ready with mapping quality score zero (MQ0 <= 0.1 * depth of coverage)or MAPQ zero reads at locus < 4 were used. Only SNPs passing all of the criteria were included in the final analysis. Assuming a dominant model of inheritance, each case was required to have at least one novel (not in dbSNP or 1000 Genomes Project) variant in the same gene. The high-quality variants common to both the father and son where carried through for further analysis (Figure 2.5).

Next, coding variants were annotated using the SeattleSeq annotation server (http://snp.gs.washington.edu/SeattleSeqAnnotation131/). SeattleSeq provides annotation of known and novel single nucleotide polymorphisms (SNPs) based on chromosome, base position and allelic change. Annotation includes dbSNP ‘rs’ numbers, gene names, accession numbers, SNP functions, protein positions and amino-acid changes. Because the primary hypothesis is that a rare, previously undetected variant is associated with hereditary mixed polyposis in this family, common variants were filtered out by excluding those found in dbSNP (build131), the 1000 Genomes Project, or 43 unaffected controls sequenced by hybrid capture and whole exome

sequencing (JL). Variants that lead to premature stop codons and altered splicing are more likely to have deleterious consequences on protein structure or function. In order to prioritize which novel, coding variants shared by the father and son should be followed up for further functional analysis, we used two bioinformatic prediction algorithms, Sorting Intolerant from Tolerant (SIFT) (http://sift.jcvi.org/, (Ng and Henikoff 2003)) and Polymorphism Phenotyping 2 (PolyPhen2) (http://genetics.bwh.harvard.edu/pph/, (Adzhubei et al 2010)) to determine the effects of mutations on protein function.

2.2.3 Validation and Replication

The validation of potentially interesting variants is needed to eliminate false positive variants due to genotyping errors that can occur from next-generation sequencing. Validation by Sanger sequencing was done in germline DNA from all four affected family members (Figure 2.1: II-2, III-3, III-4, III-5) and cDNA from available tumors.

2.2.3.4 Germline DNA extraction

Germline DNA was extracted from whole blood using the Puregene kit (Gentra Systems, Inc., Minneapolis, MN). DNA samples were quantified using the spectrophotometer (ND-1000, NanoDrop Technologies, Inc., Wilmington, DE) and PicoGreen assay (Molecular Probes Invitrogen Detection Technologies, Eugene, OR). The concentration for all qualified samples was normalized to 50 ng/ul.

2.2.3.5 Formalin fixed paraffin-embedded tumor DNA extraction

Paraffin-embedded tumors with adequate residual tissue for microdissection were

available for analysis for the father (II-2) and son (III-5). Tumor blocks were recut for uniform histopathologic review and microdissection, with the first slide of a series of 12 reviewed by a qualified pathologist (JKG) to confirm the original diagnosis and to circle areas for microdissection. Corresponding areas of normal tissue (with 0% tumor) from the same slide, or from another section of the same surgical resection, were circled for microdissection. DNA was extracted by scraping tissue from designated areas of slides with a clean razor blade and transferring the samples to separate non-siliconized tubes. Xylene (350 µl) was added to each sample to dissolve the paraffin, and ethanol precipitation was performed by adding 150 µl of cold 100% ethanol to each sample. Samples were next spun at 14,000 rpm at room temperature for 10 minutes. The supernatant was discarded and pellets were lyophilized in a Speed Vac for 8 min on high heat. Pellets were then resuspended in 100 µl of proteinase K buffer (200 ng/µl proteinase K in 50 mM Tris, pH 8.3) and incubated over-night at 37°C. Samples were heated at 95°C for 8 min and quickly transferred to ice for 5 min to keep the DNA from re-naturing. DNA samples were then stored at -80°C.

2.2.3.6 RNA extraction

Total RNA was extracted from single tissue isolates from available polyps, tumors and corresponding normal adjacent tissue using the Qiagen RNeasy FFPE kit (Qiagen, Germantown, MA). Adequate quantities of high-quality total RNA were isolated as determined by Agilent 2100 BioAnalyzer (Agilent Technologies, Palo Alto, CA). cDNA was synthesized using the High Capacity cDNA Reverse Transcription Kit from 200 ng of RNA (Applied Biosystems).

2.2.3.7 Polymerase chain reaction

The PCR reaction mixtures (20μL) contained 5ng of genomic DNA (of cDNA from tumors), 2μl of 10X PCR buffer (Applied Biosystems), 1.6μL of 25mM $MgCl_2$ (Applied Biosystems), 0.8μL each of 10mM dNTP (New England Biolabs) and 10μM forward and reverse primers, and 1 U of AmpliTaq Gold DNA polymerase (Applied Biosystems). Cycling conditions were as follows: Initial denaturation at 95°C for 3 minutes, 15 cycles of 95°C for 30 seconds, 70°C for 45 seconds (-1° every cycle), 72°C for 1 minute 10 seconds, 20 cycles of 95°C for 30 seconds, 55°C for 45 seconds, 72°C for 1 minute 10 seconds, and a final extension at 72°C for 10 minutes. PCR products were sequenced at the University of Michigan DNA Sequencing Core, and Mutation Surveyor Software (SoftGenetics, LLC., State College, PA, USA) was used to detect variants identified by whole genome sequencing.

Real-time quantitative PCR (qRT-PCR) was also performed using SYBR Green PCR Master Mix (Applied Biosystems) on an Applied Biosystems Prism 7900 HT Sequence Detection System. Two sets of primers were designed to assess the independent expression levels of candidate genes of interest and an endogenous control gene, *GAPDH*. All samples were tested in triplicate. The relative expression of the gene of interest was calculated by DCt normalization to the expression of *GAPDH*.

2.3 Results

2.3.1 Massively parallel sequencing

The goal of this study was to identify novel variants associated with risk of hereditary mixed polyposis syndrome. A total of 595,292,661 paired-end reads (76-120bp each) were 'mappable', with an average of 10x coverage of the haploid genome (Table 2.2). The first Solexa sequencer, the Genome Analyzer, was launched in 2007 and gave scientists the power to

sequence 1G of bases in a single run. When this study began in 2009, the Genome Analyser IIx allowed for 2 x 35bp read lengths of 600 million paired-end reads per flow cell. However, newer generations of the Illumina Genome Analyzer refined and optimized the sequencing process and allowed for more data acquisition. The samples presented here were sequenced over several years and over several sequencing platforms, therefore the output and coverage is inconsistent between runs.

2.3.1.1 Variant calling and annotation

In this study, we were interested in novel variants that are shared by the father and son. Massively parallel sequencing identified 1,162,925 such variants of which 125,460 were not present in 1000 genomes or dbSNP 131 (Figure 2.5). After quality control filtration and annotation, 64 previously unidentified nonsense, missense or splice site variants were identified in the father and the son whose whole genomes were sequenced, of those only 11 (9 missense, 2 nonsense) were predicted to be damaging by Polyphen2 and SIFT (Table 2.3).

2.3.2 Validation and replication

Nine of the 11 novel candidate variants were validated by Sanger sequencing. Six out of the 9 (67%) variants were shared by affected family members, leading to a small subset of candidate genes putatively responsible for HMPS within this family (Table 2.4). We prioritized nonsense variants first, because premature stop codons are more likely to have deleterious consequences on protein structure or function. Next-generation sequencing identified two novel, shared, nonsense variants in two candidate genes, *DMXL2* and *ZNF426*. The next step was to determine whether these variants were present in the affected family members and the unrelated

individuals with polyposis from Ohio State. Coding regions of both *DMXL2* and *ZNF426* were Sanger sequenced in the two affected sisters (Figure 2.1 III-3 and III-4) and the three unrelated individuals (Table 2.1 B: ID 8329-00, ID 8333-00, and ID 8293-00). The same nonsense variant that was present in the father (II-2) and son (III-5) in *ZNF426* was also identified in the two affected daughters (III-3 and III-4). The variant idenfied in *DMXL2* was not present in one of the affected sisters (Figure 2.1 III-4). None of the novel variants were present in the 3 unrelated polyposis individuals. Tumor DNA as well as DNA extracted from normal adjacent tissue for the father and son were sequenced for variation in *ZNF426*. The novel variant was present in the tumor, however no loss of heterozygosity was detected. Analysis from quantitative real time PCR revealed decreased expression in the carcinoma tissue from the father (II-2) but the adenomas (including son's sessile serrated adenoma) and normal adjacent tissue were expressed at comparable level in both the father (II-2) and son (III-5).

Although we attempted replicate potential variants by sequencing the exons of all candidate genes in the unrelated individuals from Ohio State, we did not find any variation in these individuals at the candidate loci. Hereditary mixed polyposis is extremely difficult to diagnosis because of the complex tumor phenotype. Therefore, we had a pathologist at the University of Michigan (JG) confirm that these individual do not have hereditary mixed polyposis, but likely have another polyposis-related syndrome.

2.4 Discussion

Hereditary mixed polyposis syndrome is characterized as an inherited form of polyposis associated with an increased colorectal cancer risk. This syndrome is extremely difficult to accurately diagnosis and the genetic basis is unknown. The hypothesis behind this project is that

a rare variant may be associated with this syndrome and I attempted to identify this variant using massively parallel sequencing which allows for the detection of previously unidentified low frequency allelic variants. Advances in sequencing technology have made this a cost-effective approach to generate large amounts of sequence data. It is now possible to obtain information at a single-base resolution in a high-throughput manner on the level of the entire human genome. We identified a novel candidate locus, *ZNF426,* using NGS.

The zinc finger protein 426 (Gene ID: 79088) is located on 19p13.2. Little is known about this gene other than variants in *ZNF426* have been found to play role in regulation of Kaposi's sarcoma-associated herpesvirus RT-mediated expression (Watanabe et al 2007). Several lines of evidence suggest that many of the zinc-finger-containing genes in the human genome are arranged in clusters (Yang et al 2009). Unfortunately, the structure or function of these zinc finger clusters is unclear except that there is some evidence that there exists a human cluster consisting >10 related Kruppel-associated box (KRAB)-containing *ZNF* genes organized in tandem over a distance of 350-450kb on chromosome 19. The information about their conservation throughout species is also limited other than the *ZNF* gene cluster in human chromosome 19q13.2 is conserved on mouse chromosome 7 (Shannon et al 1996). A study of gastrointestinal stromal tumor (GIST) patients as part of a phase II trial of neoadjuvant/adjuvant imatinib mesylate treatment for advanced primary and recurrent operable GISTs identified a gene signature that includes KRAB-ZNFs. Using gene expression profiling of tumor samples before and after imatinib mesylate therapy, they found 38 genes that were expressed at significantly lower levels in the pretreatment biopsy samples from tumors that significantly responded (>25% tumor reduction) to 8 to 12 weeks of imatinib mesylate. Eighteen of these genes encoded Krüppel-associated box (KRAB) domain containing zinc finger (ZNF)

transcriptional repressors, of which, 10 KRAB-ZNF genes mapped to chromosome 19p13.2 (Rink et al 2009). Also of interest, it was recently reported that epigenetic silencing of Krüppel-type zinc finger protein genes exists on chromosome 19q13 in oral cancer tissue in a genome-wide DNA methylation study (Lleras et al 2011).

Another point of interest is that allelic heterogeneity is likely to be ubiquitous among genes that harbor causal rare variants (Bodmer and Bonilla 2008, Pritchard 2001). It is often the case that a group of variants affecting the same gene or a set of genes with related functions is required to have a substantial probability of affecting the function of the relevant gene product. However, there are many challenges to studying heterogeneity in genes that have rare causal variants, such as the choice of candidate genes, the choice of appropriate case groups, the need for extensive DNA resequencing in large numbers of individuals, and the assessment of the functional consequences of variants. For example, for cancer, the most obvious candidates are genes that are mutated somatically or epigenetically changed in their expression in a significant proportion of cancers. Cases therefore should be enriched for the presence of rare variants. Generally these will include cases with one or more close relatives affected, but which are not clearly familial and with an early age of onset. Because hereditary mixed polyposis is a rare syndrome which is thought to be Mendelian in inheritance, we are assuming that variable expression of a single gene is responsible for multiple phenotypes. We attempted to address this by sequencing unrelated individuals from Ohio State, however, we did not find any novel variants in these individuals at any of the candidate loci. The individuals from Ohio State likely have some related polyposis syndrome that is not hereditary mixed polyposis.

Although *ZNF426*, is an attractive candidate for hereditary mixed polyposis in this family, there are several limitations to the current study. The limitation with potentially the

greatest impact is the low depth of coverage. Since the inception of this study 2009 when massively parallel sequence first because available, it has been established that 40X-50X coverage of the genome is necessary to rule out false negatives (Garner 2011). Therefore, our study is vulnerable to Type II error. For example, in the context of next-generation sequencing, if the null hypothesis is that an individual does not have some damaging variant that leads to disease, and the patient does in fact does have an undetected damaging variant, then the test fails to reject the hypothesis and in turn, incorrectly suggests that the patient does not have a damaging variant or is 'unaffected'. Because the current study focused on the identification of novel, coding variants, specifically missense, non-sense and splice variants with a high damaging potential, it is possible that we missed other equally interesting variation that certainly exists in the genome.

Yet another major challenge in next-generation sequencing analysis is distinguishing between background polymorphisms and potentially disease-associated mutations. Detailed analysis of a single individual typically requires deep sequencing, therefore we used stringent quality control filters based on preliminary analyses. Filtering on variant function and frequency and carefully selecting cases have been useful in selecting pathogenic variants in previous studies (Ng et al 2010b). However, in cases where genetic heterogeneity exists, these filters may not be as effective for identifying pathogenic variation. Relaxing these criteria will in turn increase the number of candidate genes, requiring new approaches to prioritize variants of interest. Because this disease is so rare, many individuals are required in order to combine shallow sequence data across individuals to generate accurate calls is limited. However, recent studies have demonstrated that for disease-associated variants with frequency >0.2%, sequencing 3000 individuals at 4X depth provides similar power to deep sequencing of >2000 individuals at

30X depth but requires only ~20% of the sequencing effort (Li et al 2010, Li et al 2011). Additionally, low-coverage sequencing data can be used to build a reference panel that can drive imputation into additional samples to increase power (Li et al 2010, Li et al 2011). This strategy of using low-coverage sequencing data as a reference for imputation into additional samples may be a useful and cost effective analysis of additionally identifying mutations in polyposis related colon cancer.

2.5 Future Directions

The genomes of the two affected sisters (III-3, III-4) are currently being sequenced to 40X at the University of Michigan Sequencing Core using the Genome Analyzer IIx, which is capable of paired end sequencing with 150 bp read depth and up to 640 million paired-end reads per flow cell. Whole-exome sequencing is also being performed on the father and son to increase the coverage of the exonic regions to 40X and identify additional false negatives undetected by the current shallow genome coverage.

Table 2.1: Patient descriptions from the University of Michigan and the Ohio State University

A. University of Michigan Samples

IID	Dx Age	Gender	Site	Histology
II-2	67	Male	Ascending colon	Adenocarcinoma with mixed features
III-3	31	Female	unknown	Hyperplastic polyps, juvenile polyp
III-5	46	Male	Sigmoid colon	Hyperplastic polyps, sessile serrated adenoma
III-4	50	Female	Descending colon	Hyperplastic polyps, sessile serrated adenoma

B. Ohio State University Samples*

IID	Dx Age	Gender	Site	Histology
8328-00	46, 50	Male	Transverse colon, Right colon	Hyperplastic polyps, sessile serrated adenoma
8333-00	57	Male	Sigmoid colon	Hyperplastic polyps, sessile serrated adenoma
8293-00	71, 72, 73	Male	cecum	Adenocarcinoma

** Diagnosis not confirmed by University of Michigan pathologist (JG)*

Table 2.2: Alignment of sequenced reads for each individual/lane

Lane	Total reads(bp)	Pair-Length	Aligned pairs	dup rate
Run50_s_5	15799030	36*36	15349266	1.29%
Run50_s_6	16782874	36*36	16295243	1.31%
Run50_s_7	16567174	36*36	16098798	1.26%
Run54_s_1	21335913	52*69	19985654	1.17%
Run54_s_2	21480977	52*70	20357997	1.18%
Run54_s_3	20730885	41*70	19697480	1.18%
Run54_s_4	21369840	52*70	20262894	1.15%
Run54_s_5	21062978	52*70	19974512	1.11%
Run54_s_6	20941330	52*70	19860777	1.10%
Run54_s_7	21297254	52*70	20212580	1.14%
Run54_s_8	20937268	52*70	19851061	1.14%
Run78_s_2	32767801	80*80	30464719	2.05%
Run78_s_3	30060322	80*80	30060467	2.25%
Run78_s_4	29985539	80*80	29985710	2.26%
Run78_s_5	31717446	80*80	29909069	2.25%
Run78_s_6	31967872	80*80	30231466	2.24%
Run78_s_7	30149000	80*80	30149179	2.28%
Run95_s_1	33646324	120*120	15090078	3.20%
Run95_s_2	34370879	120*120	15039776	3.20%
Run95_s_3	34841643	120*120	15331651	3.33%
Run95_s_4	34699034	120*120	15371303	3.30%
Run95_s_6	34862302	120*120	15529162	3.40%
Run95_s_7	34231103	120*120	15377865	3.30%
Run95_s_8	32874283	120*120	14805954	3.27%

Runs 50, 54, and 78 refer to patient III-5. Run 95 refers to the proband, II-2.

Table 2.3: Novel damaging variants predicted by both Polyphen2 and SIFT. Nine missense and 2 nonsense novel variants were identified using whole-genome sequencing and subsequently Sanger sequencing all affected individuals (II-2, III-3, III-4 and III-5).

Chr	Position	Allele	Function	Amino acid	Position	Gene
2	234,229,468	C/T	Missense	THR/MET	125	*SAG*
2	55,407,700	C/T	Missense	GLY/SER	444	*C2orf63*
3	100,352,109	C/T	Missense	ALA/VAL	112	*GPR128*
3	63,824,048	C/G	Missense	GLU/GLN	89	*THOC7*
5	155,771,587	G/A	Missense	ARG/GLN	30	*SGCD*
5	38,921,864	G/A	Missense	GLY/ASP	578	*OSMR*
12	54,741,568	G/A	Missense	ARG/GLN	112	*COPZ1*
18	34,349,252	C/G	Missense	GLY/ALA	1367	*FHOD3*
19	39,924,011	A/G	Missense	TYR/HIS	115	*RPS16*
15	51,809,316	A/G	Nonsense	ARG, stop	829/3037	*DMXL2*
19	**9,639,194**	**T/A**	**Nonsense**	**CYS, stop**	**509/555**	***ZNF426****

** ZNF426 was also validated using Sanger sequencing in the two affected sisters (III-3, III-4)*

Table 2.4: Primer sequences for exonic regions of the candidate gene *ZNF426*

ZNF426	Chr	Start	End	Forward Primer	Reverse Primer	bp	Tm
Exon 1	19	9646774	9647097	CCGGTGTATATGTTTGATGGC	AGCACCTCCCAAATGAATCAG	324	60
Exon 2	19	9645667	9646093	CTGCAACTGGCCTTATGTGTG	CTTTAAGAAGCCTGGAGACCG	427	60
Exon 3	19	9644317	9644739	GAATCCTGCATAGGCATTGG	CTAACAGGCAGATGGTGTTGC	423	60
Exon 4	19	9643393	9643789	AGCCTGAATTTCGGTTTAGGG	TTCCTAACTTTCTCTAAACCTTGG	397	60
Exon 5	19	9641510	9641935	GTGCTCACCACCACACC	GTTTCACAATTGGAGCATCCC	426	60
Exon 6	19	9640084	9640458	AAATAAATCTACAGGGAAGGACCT	CAATATTTGGTGTCAGGCTGAAG	375	60
Exon 6	19	9639760	9640286	CTGTGACTGTGAGCAATGTGG	TGGAATAATTGAAGGCTTTCCC	527	60
Exon 6	19	9639373	9639966	ACGAATGGAGGAATTATGGGC	AGGTGTATGGTTTCTGGGCAC	594	60
Exon 6	19	9639009	9639607	TCCTTCCTTACATCCTCACGC	GGAACAAATGAGAGCTTTCCC	599	60
Exon 6	19	9638954	9639542	ATGTGTTGAATGTGGGAAAGC	TTTCATGCAGCTTCTTCTCTCC	589	60

Figure 2.1: Family 202 pedigree. Individuals were recruited through the University of Michigan Cancer Genetics Clinic. Affected individuals are colored in black, and diagnosis is listed below each affected individual. The genomes of individuals II-2 and III-5 were sequenced. Germline DNA was collected on II-2, III-3, III-4 and III-5.

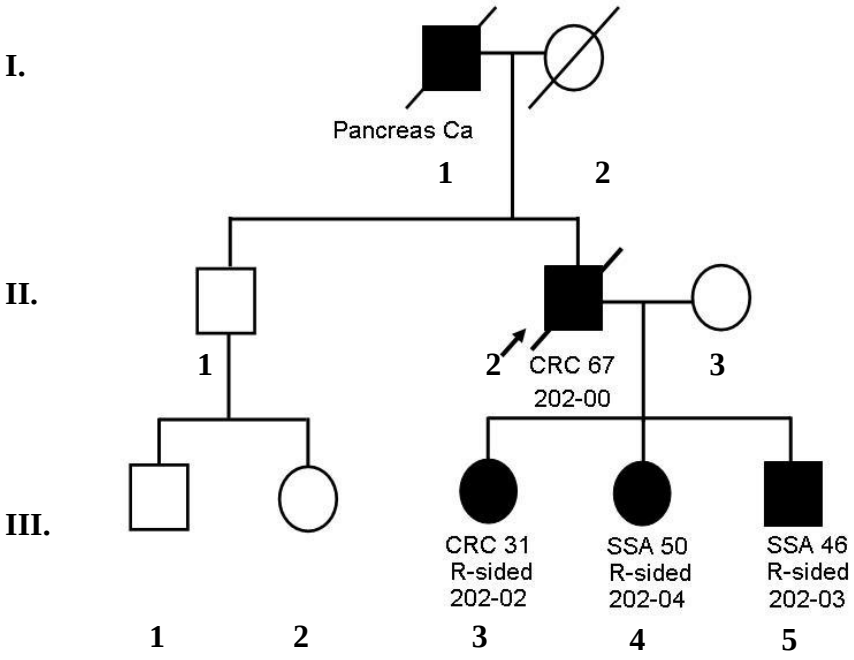
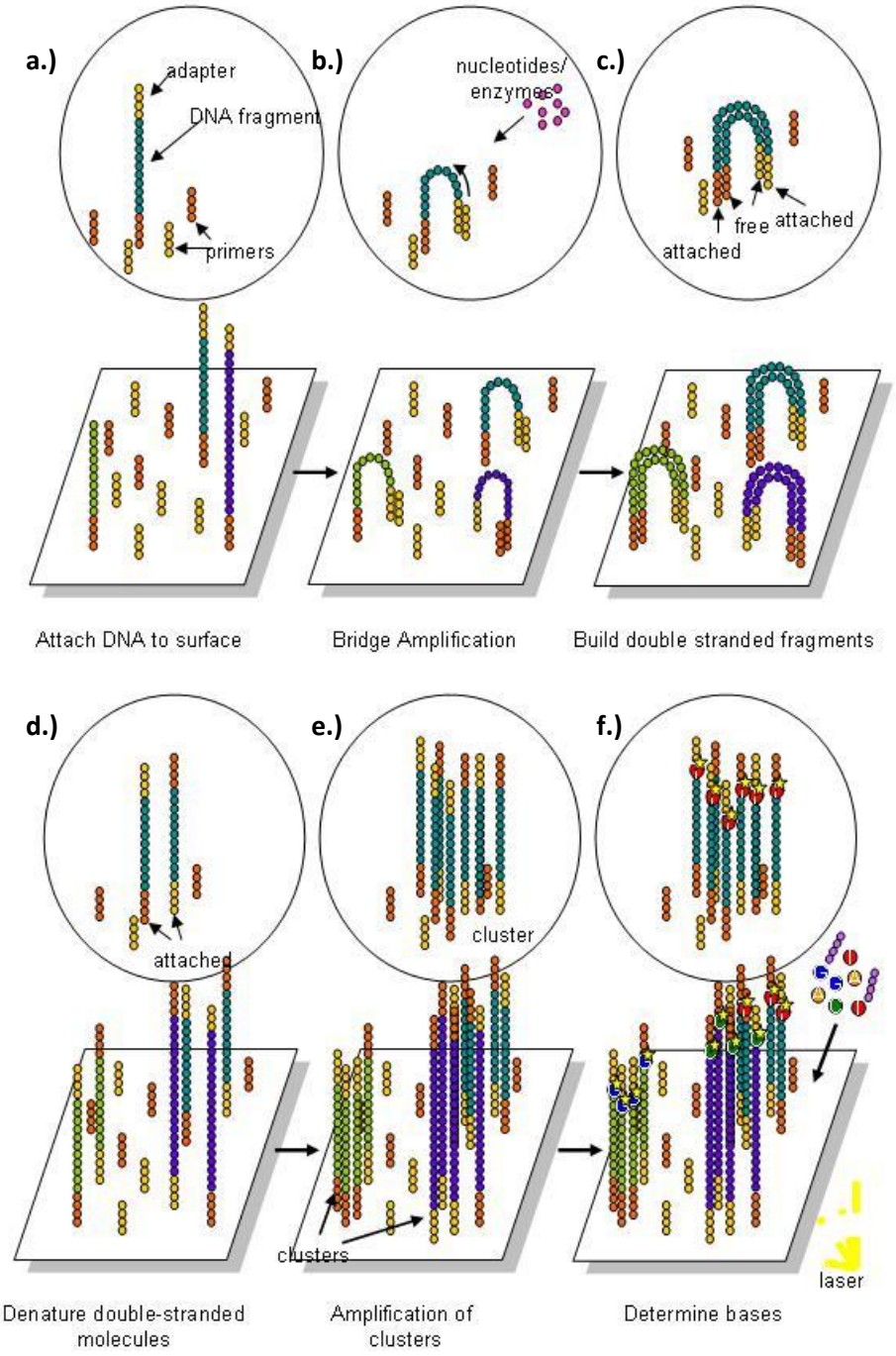
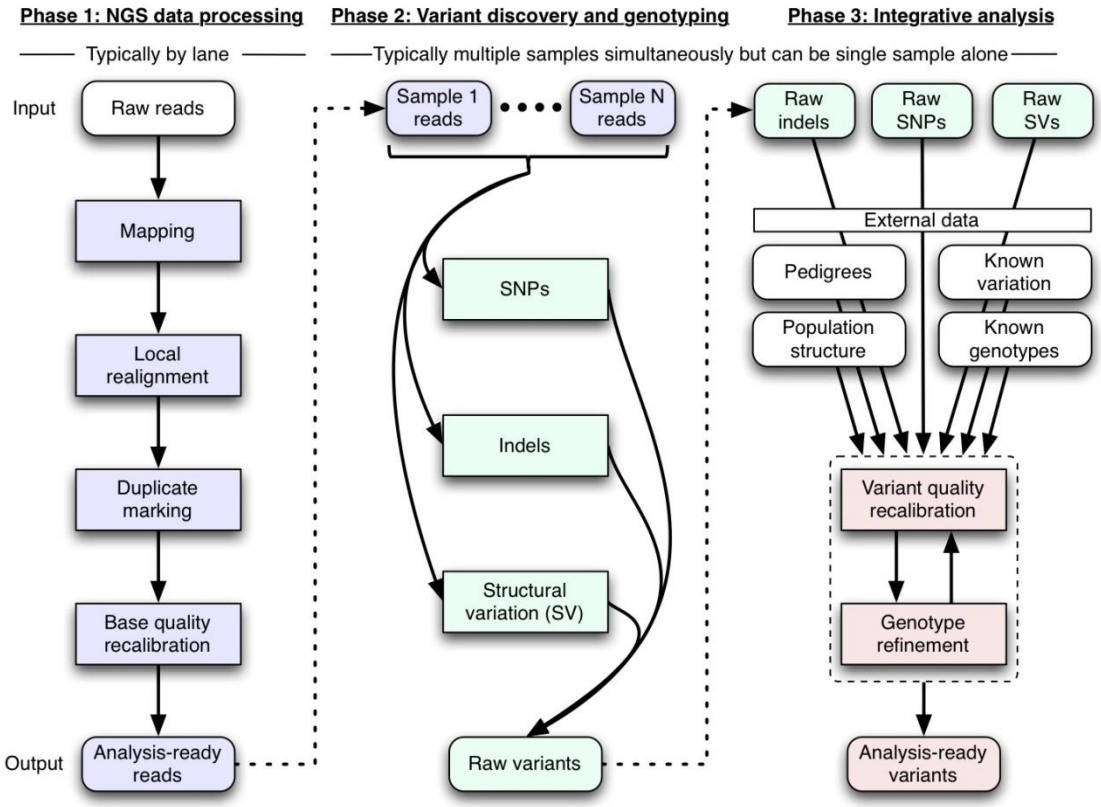


Figure 2.2: Illumina 'bridge amplification' technique. After library preparation the DNA fragments are: a.) hybridized to a lawn of primers b.) extended by polymerases c.) formation of bridge d.) bridge is denatured e.) cycle repeated f.) sequenced by synthesis



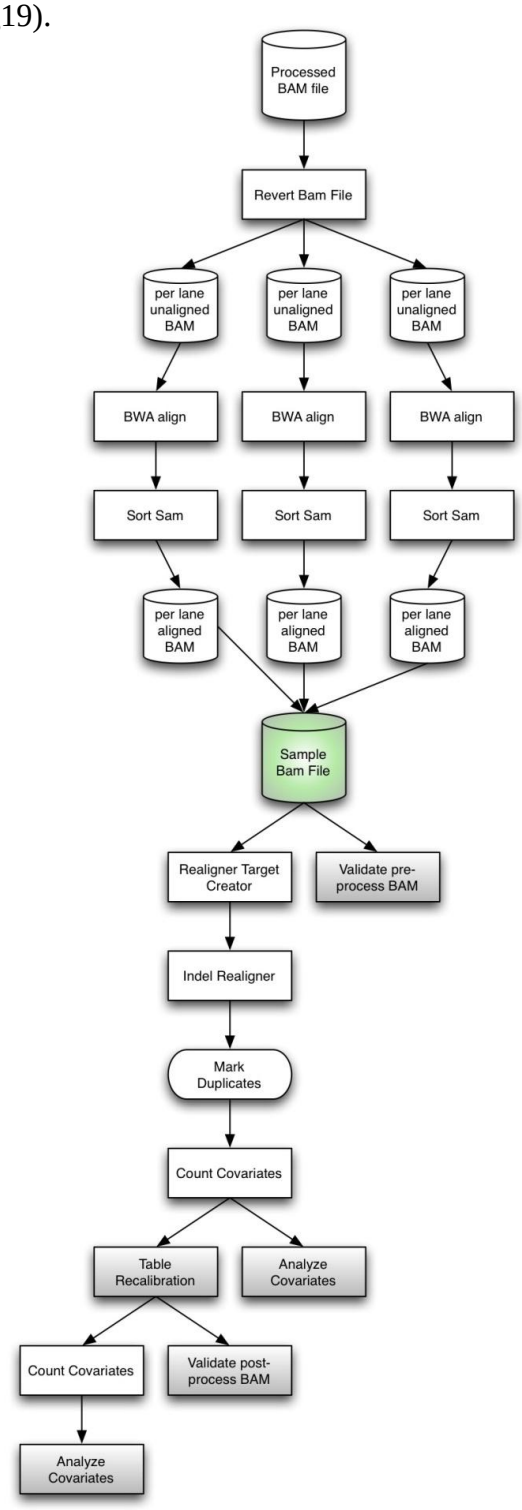
Adapted from www.illumina.com

Figure 2.3: GATK framework pipeline used for data processing and analysis



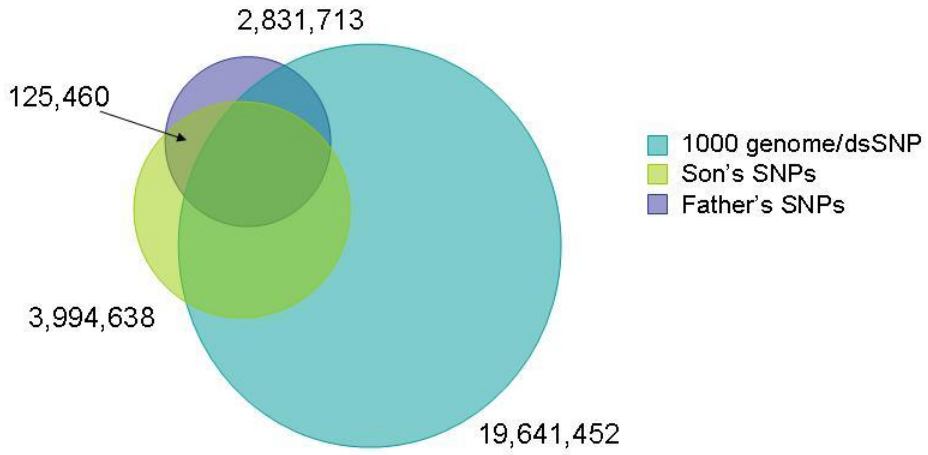
http://www.broadinstitute.org/gsa/wiki

Figure 2.4: GATK alignment pipeline. Pipeline used to align the sequence reads to the reference genome (hg19).



http://www.broadinstitute.org/gsa/wiki

Figure 2.5: Variants identified by whole genome sequencing. The figure below indicates the number of variants identified by whole-genome sequencing of a father (II-2) and son (III-5) with hereditary mixed polyposis. Approximately 125,000 *novel* variants were found in both the father and son and were used for subsequent analysis.



CHAPTER 3

Elucidating the complexity of chromosome 18 loss in chromosomally instable colorectal cancer

3.1 Introduction

The accumulation of somatic alterations transform normal colonic epithelial cells and instigate the development of colorectal cancer (CRC) (Fearon and Vogelstein 1990, Jass 2007). In particular, structural alterations that lead to genomic instability, such as amplifications and deletions, play an important role in pathogenesis. Genomic instability in colorectal cancer develops through two major pathways based on molecular and mutational patterns, including the chromosomally instable (CIN) and microsatellite instability (MSI) pathways (Peinado et al 1992, Thibodeau et al 1993, Vilar and Gruber 2010, Vogelstein et al 1988). The majority of colorectal cancer is characterized by chromosomal instability (CIN, also referred to as microsatellite stable or MSS) arising through the loss or gain of chromosomes and other structural rearrangements (Cahill et al 1998, Vogelstein et al 1989). By contrast, microsatellite instable (MSI) CRCs arise as a consequence of the loss of DNA mismatch repair (MMR) function and result in the accumulation of insertion and deletion mutations, particularly in microsatellite sequences. In general, tumors with similar molecular characteristics arise and behave similarly, however a third group of tumors classified as microsatellite instabilty low (MSI-L) exists in about 10% of sporadic colorectal cancers (Jass et al 1999).

Somatic loss of heterozygosity (LOH) on the long arm of chromosome 18 (18q) has been

shown to be inversely associated with microsatellite instability (MSI), and is an important molecular classifier in colorectal cancer (Jass 2007, Watanabe et al 2001). Loss of heterozygosity is thought to occur later in the development from adenoma to carcinoma, and to contribute to the inactivation of several genes with potential importance in pathogenesis (Vogelstein et al 1988). Among the proposed candidate genes located on 18q are the *DCC, SMAD4,* and *SMAD22* genes. The *DCC* gene encodes a neutrin-1 receptor, which is important in apoptosis, cell adhesion, and tumor suppression (Fearon et al 1990). *SMAD4* codes for a nuclear transcription factor in transforming growth factor-β1 signaling, and is involved in tumor suppression (Zhou et al 1998). Although candidate tumor suppressor genes located on 18q have been proposed to be targets of this LOH, the specific breakpoints and how they are involved are often imprecisely defined (Fearon and Vogelstein 1990, Park do et al 2007, Thiagalingam et al 1996).

Additionally, several studies of CRC have suggested allelic loss of chromosome 18q is associated with high metastatic potential and reduced patient survival (Diep et al 2003, Jen et al 1994, Ogunbiyi et al 1998, Watanabe et al 2001, Zhou et al 2002). Several studies have focused on the impact of LOH at 18q on the prognosis of early-stage CRC (Alazzouzi et al 2005, Barratt et al 2002, Bisgaard et al 2001, Carethers et al 1998, Chang et al 2005, Choi et al 2002, Diep et al 2003, Font et al 2001, Halling et al 1999, Lanza et al 1998, Laurent-Puig et al 1992, Martinez-Lopez et al 1998, Ogunbiyi et al 1998, Pietra et al 1998, Watanabe et al 2001, Zhang et al 2003). Of these sixteen studies, 8 found that CRC patients with LOH at 18q had asignificantly lower survival than those who were heterozygous in this region. Additionally, four of the eight found that chromosomally instable tumors had a significantly worse prognosis in a multivariate analysis with hazard ratios for death from 2.0 (95% CI: 0.27-5.10) - 7.30 (95% CI: 1.14 to 7.35)

(Carethers et al 1998, Ogunbiyi et al 1998, Watanabe et al 2001) and for recurrence a hazard ratio of 9.60 (Pietra et al 1998). Three reports did not find loss of 18q to be independently associated with prognosis (Bisgaard et al 2001, Choi et al 2002, Diep et al 2003) and one did not perform multivariate analysis (Chang et al 2005). LOH was not found to be prognostic in a univariate or multivariate analysis, yet LOH at 18q did yield a poor prognosis in stage II disease (Martinez-Lopez et al 1998). While many groups have studied LOH at 18q with respect to CRC, conflicting results warrant a more comprehensive analysis.

With recent advances in genomic technology, systematic analysis of inter- and intra-tumor heterogeneity in cancer at a high resolution is now possible using high-density SNP arrays. In this study we evaluated genomic alterations in tumors of chromosomally instable colorectal cancers and compared those to microsatellite instable cancers by performing genome-wide analysis of copy number alterations (CNAs) and focused on the somatic heterogeneity on chromosome 18. We used two different SNP genotyping platforms, the Affymetrix Genome-Wide Human SNP Array 6.0 and the Illumina Human 1M-Duo BeadChip. Germline and tumor DNA from 21 cases with colorectal cancer and germline DNA from 21 matched controls were run on both genome-wide SNP array platforms. Our original study design included thirty tumor DNA and thirty germline DNA (extracted from whole blood) samples from the same case, as well as 30 germline DNA samples from matched controls to be genotyped on the Affymetrix platform. However, after experiencing technical difficulties associated with the arrays at the University of Michigan Comprehensive Cancer Center Genomics Core, we decided to re-run 9 tumor, 9 germline and 9 matched control DNA samples on the Illumina arrays. After genotying was completed, I incorporated existing gene expression data on colorectal tumors from the

U133APlus Affymetrix expression array (n=21) to determine whether copy number alterations alter expression of genes associated with colorectal cancer (Vilar et al 2009).

Examining copy number heterogeneity in tumors in conjunction with genomic instability is instrumental in elucidating the mechanisms underlying tumorigenesis, and potentially useful for targeted clinical intervention. The identification of structural alterations on chromosome 18 and their implication for prognosis of CRC is currently an important topic of interest in clinical treatment of CRC (Jen et al 1994, Jernvall et al 1999, Ogino et al 2009, Ogunbiyi and Ogunbiyi 1998). A potential for clinical intervention is adjuvant chemotherapy which has recently been suggested to improve survival in patients with colon cancer (Andre et al 2009). Furthermore, a previous study suggested that 18q loss may be a significant prognostic factor for patients with colorectal cancer who received fluorouracil-based adjuvant chemotherapy (Watanabe et al 2001).

3.2 Subjects and Methods

3.2.1 Subjects

Patients were recruited as part of the Molecular Epidemiology of Colorectal Cancer (MECC) study, a population-based case-control study in northern Israel of all incident cases of colorectal cancer between March 31, 1998 and March 31, 2004. Incident colorectal cancer cases were ascertained from five hospitals in northern Israel, and all cases for this study have histologically confirmed cancer of the colon or rectum. The controls were individually matched for exact year of birth, sex, clinic, and Jewish versus non-Jewish heritage. The study was approved by all relevant IRBs at the University of Michigan and Carmel Medical Center in Haifa, and study participants gave written informed consent. Detailed descriptions of this study have previously been published (Poynter et al 2005). Tumor samples were obtained at the time of

surgical resection through the MECC study. Tumors were snap-frozen in liquid nitrogen, shipped on dry ice, and stored at –140C°. Frozen tumors were embedded in freezing media, cryotome sectioned (5µm) and evaluated by routine hematoxylin and eosin (H&E) stains by a surgical pathologist (JG). Areas of at least 70% tumor cells were selected for DNA and RNA isolation.

3.2.2 DNA and RNA isolation

3.2.2.1 Tumor DNA

DNA was precipitated from the organic phase by adding 0.3 ml of 100% ethanol per 1 ml of TRIZOL Reagent. Samples were stored at 15-30°C for 2-3 minutes and then centrifuged at 2,000 × g for 5 minutes at 28°C. The phenol-ethanol aqueous phase was removed and the DNA pellet was washed twice in a solution containing 0.1 M sodium citrate in 10% ethanol. At each wash, the DNA pellet was stored in the washing solution for 30 minutes at 15=30°C and centrifuged at 2,000 × g for 5 minutes at 2-8°C. Following these two washes, the DNA pellet was suspended in 75% ethanol (1.5-2 ml of 75% ethanol per 1 ml TRIzol Reagent), and then stored for 10-20 minutes at 15-30°C and centrifuge at 2,000 × g for 5 minutes at 2-8°C. The DNA was air dried for 15 minutes in an open tube and dissolved by adding 300 – 600 µl of 8 mM NaOH to DNA isolated from 107 cells or 50 – 70 mg of tissue such that the concentration of DNA was between 0.2 – 0.3 µg/µl. The insoluble material was removed by centrifugation at >12,000 × g for 10 minutes. Next the supernatant containing the DNA was transferred to a new tube. For prolonged storage, the DNA was solubilized by adding 8 mM NaOH and samples were adjusted with HEPES to pH 7-8 and supplemented with 1 mM EDTA. Once the pH was adjusted, DNA was stored at– 20°C.

3.2.2.2 Tumor RNA

Total RNA was extracted from single tissue isolates using the TRIzol protocol (Invitrogen, Carlsbad, CA) and Qiagen RNeasy purification kit (Qiagen, Germantown, MA). Adequate quantities of high-quality total RNA was assessed by 1% agarose gel electrophoresis, and samples were included only if the 18S and 28S bands were discrete and an rRNA ratio >2.0 as measured by the Agilent 2100 BioAnalyzer (Agilent Technologies, Palo Alto, CA).

3.2.2.3 Germline DNA

Germline DNA was extracted from whole blood using the Puregene kit (Gentra Systems, Inc., Minneapolis, MN). DNA samples were quantified using the spectrophotometer (ND-1000, NanoDrop Technologies, Inc., Wilmington, DE) and PicoGreen assay (Molecular Probes Invitrogen Detection Technologies, Eugene, OR). The concentration for all high quality samples was normalized to 50 ng/µl.

3.2.3 Genome-wide arrays

3.2.3.1 Affymetrix 6.0 single nucleotide polymorphism based array

The Affymetrix 6.0 Genome-Wide Human SNP Array (Affymetrix, Santa Clara, CA) contains 1.8 million probes designed to interrogate 906,600 SNPs and more than 946,000 invariant sites for the detection of copy number changes. We ran 21 sets of samples (tumor DNA, case germline DNA, and matched control DNA) on arrays at the University of Michigan Comprehensive Cancer Center Genomics Core. 250ng of DNA were digested with the StyI and NspI restriction enzymes. After the digestion, samples were PCR amplified and subsequently

labeled with biotin before hybridization. All steps were performed according to the manufacturer's protocol. Birdsuite v1.5.5, a four-stage analytical framework software program, was used to derive copy number and SNP genotypes (Korn et al 2008). The four-stage analytical framework includes Canary, a routine used to determine the copy number of each individual at regions of known copy number polymorphisms. Next, SNP calling is done using Birdseed. At each SNP locus, samples expected to have two copies are assigned canonical SNP genotypes of AA, AB or BB. In the third step information across neighboring probes is integrated using Birdseye, a hidden Markov model (HMM) based algorithm to discover rare or de novo CNAs informed by probe-specific mean and variance intensities estimated in the second step. Finally, copy number and SNP allele information are combined using Fawkes ('fast analysis with kopy-number et SNPs') to provide an integrated view of the genetic variation in each sample (Korn et al 2008).

3.2.3.2 Illumina Human 1M-Duo DNA array

The Illumina Human1M-Duo DNA Analysis BeadChip assesses nearly 1.2 million loci per sample and was run on a subset of 8 samples pairs (germline DNA from cases and germline DNA from a matched control) as well as 8 tumor samples from the same case. Genotyping was performed at the University of Michigan DNA Sequencing Core. This subset of samples was also run on the Affymetrix 6.0 platform. The Human1M-Duo BeadChip was designed to focus on tag SNPs, variants in genes, and polymorphic markers in known and copy number variation (CNV) regions. Genotyping was performed according to the Infinium HD protocol (Illumina). In summary, 200ng of genomic or tumor DNA was whole genome amplified, fragmented, precipitated, and re-suspended. Samples were then hybridized to the Illumina Human1M-Duo

BeadChips. After hybridization, the oligonucleotides were extended by a single labeled base, which was detected by fluorescence imaging with an Illumina Bead Array Reader. For each sample, Illumina's GenomeStudio software was used to convert the fluorescence intensities into SNP genotypes. We then normalized the data using a proprietary algorithm and created genotype clusters using GenomeStudio by clustering the matched normal individuals. We converted the intensity data to polar coordinates (R, Theta). The $\log_2(R_{subject}/R_{reference})$ or logR ratio (LRR) compares the direct intensity, R, between a subject sample and a paired reference sample (germline DNA from the sample person). We also transformed allelic intensities into the B-allele frequency (allelic composition) using linear interpolation of the canonical clusters. The B-allele frequency is derived from the three canonical genotyping clusters created from the GenomeStudio software (Peiffer et al 2006).

3.2.3.3 Quality control

Genotype calls from both genotyping platforms were subject to rigorous quality control before analysis. We excluded SNPs if they had a call rate lower than 99% in cases or controls, a minor allele frequency <1% in the population, or significant deviation from Hardy-Weinberg equilibrium in the controls ($P \leq 10^{-7}$). We removed all the samples with overall genotyping rate less than 98%, We ran 2 samples in duplicate to further insure data quality. Study duplicate reproducibility was 99.98% for the Illumina platform and 99.24% for the Affymetrix platform. A HapMap CEU trio was also genotyped to check for Mendelian inconsistencies. The median number of Mendelian errors was was > 0.02% and 0.1% for the Illumina and Affymetrix platforms, respectively. SNPs that showed Mendelian inconsistencies were also removed from

analysis. In the end, a total of 902,760 and 1,199,187 SNPs from the Affymetrix and Illumina platform, respectively, were used in the analyses.

3.2.3.4 Affymetrix U133A expression array

From the tumor RNA samples, preparation of cDNA, expression analysis with the U133A array was performed according to manufacturer's protocols at the University of Michigan Comprehensive Cancer Center Genomics Core. The cDNA was prepared from 50ng of total tumor RNA for each sample using Nugen WT-Ovation™ amplification method and was subsequently hybridized to the Affymetrix GeneChip Human Genome U133A PLUS 2.0 Array (Affymetrix, Santa Clara, CA) at the University of Michigan Comprehensive Cancer Center Microarray Core. Arrays were scanned using Affymetrix protocols and GeneChip scanners. Expression values were calculated using Affymetrix GeneChip analysis software MAS 5.0 (http://www.affymetrix.com/partners_programs/genechip_compatible/genechip_compatible.affx) .

3.2.4 Statistical analysis

3.2.4.1 Identification of copy number changes

3.2.4.1.1 LogR ratio and B allele frequency

The B Allele Frequency (BAF), which is a measure of normalized allelic intensity ratio, along with the logR ratio (LRR) can be useful for identifying regions of copy number change. A normal chromosome has three BAF genotype clusters, (AA, AB, and BB genotypes). Heterozygous loci are distributed either as one track around BAF=0.5, or two separate tracks above and below 0.5. Regions of 'normal' LRR values are centered around zero, but do not have

the characteristic the AB genotype cluster. The LRR and BAF can be used in combination to determine several different copy numbers and to differentiate LOH regions from normal state regions. Normal copy number regions and copy-neutral LOH regions can be distinguished from each other based on the patterns of the LRR ratio and BAF, demonstrating the utility of combining of LRR and BAF to generate copy number alteration calls (Figure 3.1). Additionally, for each segment in the final copy number alteration set we calculated the median LRR and mean "folded BAF", which is the absolute value of (BAF-0.5), for segmentation. We performed segmentation on folded BAF data using the Circular Binary Segmentation (CBS) algorithm as described below (section 3.2.4.2.2).

3.2.4.1.2 Circular binary segmentation

Segmentation algorithms have been successful in identifying regions of copy number change for analysis of data from array-based comparative genomic hybridization (aCGH) platforms (Lai et al 2005, Willenbrock and Fridlyand 2005). We used the Circular Binary Segmentation (CBS) algorithm, a robust non-parametric method for dividing the genome into regions of equal DNA copy number in order to identify chromosomal regions of gain or loss (Olshen et al 2004, Venkatraman and Olshen 2007). The CBS algorithm tests for change-points, where the change-points are defined as the genomic locations of copy number transition using a maximal t-statistic with a permutation-derived null distribution to obtain the corresponding P-value. The algorithm starts with the whole chromosome and segments it recursively by testing for change-points and stops when no more segments can be found. CBS was run on both the LRR intensity data and the BAF data for each sample using the implementation in the Bioconductor package (R version 2.9).

3.2.4.1.3 Mixed Gaussian models

As previously described in section 3.2.4.2.1, BAF values at heterozygous loci can be distributed either as one track around 0.5, or two separate tracks above and below 0.5. To aid in distinguishing true copy number alterations in tumors from a normal copy number state in data with high background noise, we fit the distribution of heterozygous BAF values in each segment of copy alteration as either one Gaussian distribution or the summation of two Gaussian distributions. The hypothesis is that when a normal copy state exists, the BAF will resemble a single Gaussian distribution, and when heterogeneous data (representing copy number change) will be a mixture of two normal distributions. When the separation between the two tracks is small, the summed distribution may resemble a single Gaussian distribution. We used empirical data to calculate the variance of BAF in order to set the criterion that distinguishes between segments. BAF standard variation ≥ 0.1 is considered to be a two-track segment and two separate Gaussian distributions. For segments with one track, we obtain the best-fit distribution as $N(\mu,\sigma^2)$, and define the folded BAF value as $|\mu-0.5|$. For segments with two tracks, the best fit distribution is $(\mu_1,\sigma^2_1;\mu_2,\sigma^2_2)$, and the folded BAF value is $|\mu_1-\mu_2|/2$. We used the maximum likelihood estimates for the mean and variance parameters as implemented in the "nlm" (non-linear minimization) function in the R Stats (R version 2.9). Non-linear minimization is a Newton-type algorithm, where at each iteration one approximates around the estimate by a quadratic function, and then takes a step towards the maximum or minimum of that quadratic function.

3.2.4.2 Analysis of expression data

We used the Bioconductor package in R to analyze the expression of genes on chromosome 18 in CIN tumors to determine whether loss of heterozygosity could be detected by a decrease in gene expression. Tumors with loss of heterozygosity cannot be distinguished from tumors with copy-neutral loss of heterozygosity by LRR. Therefore, I wanted to investigate where expression data could be used to distinguish these two mechanisms of copy number loss. Expression data from MSI tumors was used as a control set, since the vast majority of MSI-H tumors have of normal copy states. We assessed the quality of the data by evaluating the density plots of the log-intensity and RNA degradation plots corresponding to each sample. For all samples, MAS 5.0-calculated signal intensities were quantile normalized using the microarray analysis software and the normalized data were $\log_2$ transformed. Additionally, we median-centering and scaled by the standard deviation for each sample separately. Filtering excluded probe sets that were not expressed and those that exhibited low variability across samples. Expression values were required to be above the lower quartile of all expression measurements in at least 25% of samples, and the interquartile range across the samples on $\log_2$ scale was required to be at least 0.5. After preprocessing and quality assessment, 21 samples (corresponding to the 21 tumor samples successfully genotyped on the Illumina and Affymetrix platforms) and 54,677 probe sets (654 probe sets on chromosome 18) were included further analysis.

3.2.4.2.1 Hierarchical clustering

Agglomerative hierarchical clustering analysis using Wards' method (Ward and Morris 1963) and Partitioning Around Mediods (PAM) (Kaufman and Rousseeuw 1990), a more robust version of K-means clustering, were performed on the expression data as implemented by the

Cluster package in R (R version 2.9). Briefly, Ward's method defines the distance between two clusters, "A" and "B", as the sum of squares. When clusters are merged, this value increases. Hierarchical clustering starts with the sum of squares equal to zero, and increases as the clusters are merged. Ward's method minimizes this growth to keep clusters as small as possible. The PAM algorithm computes *k* representative objects, called *medoids*, defined as that object of a cluster whose average dissimilarity to all the objects in the cluster is minimal. After finding the set of medoids, each object of the data set is then assigned to the nearest medoid. As part of the analysis, visualization plots were created to display the hierarchical clustering solution based on the expression data that were determined by Ward's method as well as heat maps of the expression intensity levels in all 21 tumors in order to visualize regions of loss and gain (created in "Heatmap" and "Heatmap2" packages in R version 2.9).

3.3 Results

3.3.1 Copy number alterations on chromosome 18

Copy number alterations were identified using circular binary segmentation, which incorporated both the LRR and BAF information from SNP genotyping arrays. We also visually inspected the LRR and BAF plots to examine the extent of the region of loss. Based on this information, we determined that 12 tumors demonstrated instability on chromosome 18, resulting in loss of whole chromosomes or deletion of only several kilobases. Plots of the LRR and BAF in regions with copy number change seen in tumor samples on chromosome 18 are depicted in Figure 3.2 (A-H). Here, loss of at least one arm of chromosome 18 occurred in the tumor samples but not the genomic DNA of cases or matched controls. Eleven tumors from either the Illumina or the Affymetrix genotyping platform demonstrated copy neutral loss of

heterozygosity (CN-LOH), and 3 tumors had a complete deletion of at least one arm of chromosome 18 (Table 3.1). The deletion of a chromosomal region was defined and detected by the lack of the AB genotype cluster, as well as LRR values below zero. The combined measurement of both allelic ratios and normalized intensities provides enhanced delectability of genetic aberrations and allows for the identification of copy neutral genetic anomalies such as UPD and mitotic recombination.

We encountered several challenges when analyzing the Affymetrix 6.0 Genome-Wide Human SNP Array data. Most notably, the background noise was extremely high (Figure 3.3, A-H). To aid in distinguishing copy number alterations from tumors with a normal copy state in data with high background noise, we fit the distribution of BAF values in each segment of copy alteration as either one Gaussian distribution or the summation of two Gaussian distributions. This approach is based on the idea that the BAF data measurements are normally distributed when the chromosomes shows normal copy number and is a mixture of normal distributions when there are copy number changes on the chromosome. We took this approach to gain some insight about the high background samples, and how they could be "unmixed" into separate distributions and used to identify copy number change. When the separation between the two tracks is small, the summed distribution should resemble a single Gaussian distribution and $N(\mu,\sigma^2)$, and the folded BAF value is $|\mu-0.5|$. However, if two (or more) Gaussian distributions are present then the best fit distribution is $(\mu_1,\sigma_1^2;\mu_2,\sigma_2^2)$, and the folded BAF value is $|\mu_1 - \mu_2|/2$. Eight tumor samples were run on both genotyping platforms, so we were able to use the high quality Illumina data to verify the alterations detected even in the high background samples (Figure 2.4: A,B,C,D). The ‘useable’ samples size was increased from 8 tumors to 21 tumors using this approach.

3.3.2 Gene Expression in chromosomally instable CRC on chromosome 18

Using previously existing expression data on colorectal tumors from the U133APlus expression array, the same 21 tumor samples with existing SNP array data from both the Illumina and Affymetrix genotyping array platforms were used to examine whether loss of heterozygosity could be detected by a decrease in gene expression. Mean expression across all probes on chromosome 18 was significantly lower (p-value = 0.000137) for samples with complete loss of one arm of chromosome 18 compared to the tumor samples with 'normal' copy numbers (including the tumors samples with copy neutral LOH and the microsatellite stable tumors). Hierarchical clustering analysis and PAM clustering was performed on the expression data as implemented by the 'Cluster' package available in R. Using Wards' method which minimizes the loss associated with each grouping during clustering, and which quantifies loss in terms of an error sum-of-squares. MSI status for the tumor samples was independently confirmed by 8 fluorescent markers and immunohistochemsity results for mismatch repair genes. The hierarchical clustering analysis produced a 2-cluster solution, clearly separating MSI from MSS, with the exception of two misclassified MSI tumors. To address the question of whether expression data could be used to distinguish between tumors with loss of heterozygosity and tumors with copy-neutral loss of heterozygosity, the clustering analysis revealed distinct groups with complete deletion of at least one arm of chromosome 18 and the samples with normal copy numbers (including CN-LOH) (Figure 3.5). In addition to looking at overall loss of chromosome 18, we looked at expression levels in previously reported critical regulatory candidate genes that appear to underlie the biology and clinical features of a subset of CRCs. In the 18q21-18q21.1 region several tumor suppressor genes have been mapped (Fearon et al 1990), including the gene

SMAD4 (Thiagalingam et al 1996, Zhou et al 1998). Chromosomal instability is an mechanism, which leads to the physical loss of a wild-type copy of a tumor suppressor gene, whose normal function is to suppress the malignant phenotype. Probes to measure the expression of 3 candidate tumor suppressor genes located on 18q (*SMAD2*, *SMAD4* and *DCC)* were present on the Affy expression arrays. Mean expression levels were not significantly different in the tumors with loss of 18q compared to the CRCs that showed MSI (Table 3.2).

3.4 Discussion

The current study examines colon tumor heterogeneity on chromosome 18 by determining copy number alterations using high-density SNP genotyping arrays. Previous investigations of loss of heterozygosity in colorectal cancer used low-resolution microsatellite markers to determine loss on chromosome 18 (Mao et al 2006, Ogino et al 2009, Pilozzi et al 2011, Wang et al 2010b), however, it is now feasible to characterize tumors with greater precision at a much higher density.

The findings from this study are similar to previous reports in which the most prevalent copy number alteration identified in colorectal cancers are chromosomal losses. New insights include the statistical methods that can improve yield of poor quality and high background data and the further refinement of the regions on chromosome 18 that distinguish chromosomally instable colorectal cancer. Across chromosome 18, the most prevalent somatic copy number changes were very short (focal) or the entire length of a chromosome arm or whole chromosome (Figure 3.2). This finding is consistent with other studies that have observed a favored loss of whole arm (Beroukhim et al). Additionally, I observed that the frequency of arm-level somatic copy number alterations decreases with the length of chromosome arms (all chromosomes), and

that the majority of chromosome arms exhibit strong evidence of preferential gain or loss, but rarely both. This suggests that loss of chromosome 18 in CRC may not be due purely to the size of chromosome 18. Additionally, the finding that the predominant mutational event in tumors tend to be arm-level, compared to focal events, reflects the lack of difficulty in which these events occur during tumorogenesis (Baudis 2007). However, the high frequency of arm-level somatic copy number changes makes it difficult to determine specific genes or targets involved in loss. Furthermore, the mechanism for this gross chromosomal instability, which results in aneuploidy in the majority of cancers, is still unknown.

It has been shown that approximately 30% of genes in the human genome code for proteins that regulate DNA replication fidelity. This implies mechanisms such as mitotic checkpoint regulation and telomere shortening could lead to instability. Another potential mechanism of chromosome instability is through defects in the mechanisms controlling the numeral integrity of centrosomes, leading to centrosome amplification resulting in defective mitosis and inevitably promoting chromosome instability in tumors (Fukasawa 2005, Fukasawa 2011, Wang et al 2004).

In colorectal cancer specifically, there has been great interest in identifying the mechanisms responsible for the chromosomal instability (CIN) phenotype. Microsatellite stable tumors often show a defect in chromosome segregation, resulting in excess gains or losses per chromosome per generation. Cell fusion experiments between CIN and transfected HT-29 cells demonstrated that the CIN phenotype acts dominantly at the cellular level, suggesting that it may arise via gain-of-function mutations (Lengauer et al 1997). Whether CIN is a consequence of chromosome mis-segregation or structural rearrangement has been studied by karyotyping near-diploid colorectal cancers. Mis-segregation of normal chromosomes and structural

rearrangements were not randomly associated within tumors, defining two major pathways of CIN, including chromosome gains by mis-segregation of normal chromosomes or chromosomal losses by both mis-segregation of normal chromosomes and structural rearrangements (Muleris et al 2008). Recently, a comparison of the karyotypes of 345 cases of adenocarcinoma of the large intestine in the Mitelman Database of Chromosome Aberrations in Cancer and the karyotypes of abnormalities observed in 15 established colorectal cancer cell lines found that there were no recurrent translocations in either tumors or cell lines; isochromosomes were the most common abnormalities; and breakpoints occurred most frequently at the centromeric/pericentromeric and telomere regions. They concluded that copy number alterations appear to be the major mechanism for transcriptional deregulation of cancer genes in CRC (Knutsen et al 2010).

Previous studies have identified genes involved in chromosome instability using bioinformatic approaches, by compared 102 human genes highly related to 96 yeast CIN genes and showed that down regulation or disruption of genes involved in sister chromatid cohesion (*MRE11A* and *CDC4*) play a major role in the CIN phenotype in colorectal tumors (Barber et al 2008, Jorissen et al 2008). Another study in CRC cell lines used gene expression and array-CGH to show major difference between MSI-associated genes in MSI tumors (near-diploid) and MSS tumors (aneuploid). These data suggests that copy number change have profound effects on expression of genes in cancer cells (Barber et al 2008, Jorissen et al 2008).

A limitation to the assessment of arm-level copy number alterations seen in the current study and many previous reports is the difficulty in identifying specific breakpoints, genes, or gene targets due to the large size of the event. In this case, nucleotide sequencing may be required to help identify point mutations, where heterozygous deletions are present.

A second limitation is that with algorithms used for identifying regions of copy number changes for high-throughput analysis are prone to false positives especially with low minor allele frequency SNPs (Hariharan 2003, Jung 2005, Lin et al 2010, Tabangin et al 2009). Therefore, validation and replication of findings is important when using array-based techniques for clinical applications, where reliability and performance are critical. To address this issue, we used independent genotyping platforms, confirmed suspected loss by quantifying decreases in expression levels from the same tumors, and Sanger sequenced candidate gene regions on chromosome 18 to verify the lack of heterozygous variation in regions of deletion.

This study demonstrates that SNP genotyping arrays can detect chromosomal alterations from frozen tumors at a high resolution and that these copy number alterations point towards candidate mechanisms and pathways for further study. Additionally, the integration of expression data with copy number data can be useful in elucidating whether copy number alterations effect expression of genes associated with colorectal cancer.

Table 3.1: Summary of alterations on chromosome 18 for all tumor samples.

		Illumina1M-Duo BeadChip		Affymetrix 6.0 SNP Array		Affymetrix Expression U133Aplus array
		Copy Number		Copy Number		Cluster Analysis
Tumors	Microsatellite instability status	LRR[¥]	BAF*	LRR	BAF[$]	Expression Data
10772	Stable	CN-LOH[$]	4	na‡	2	2
10779	Stable	CN-LOH	4	na‡	2	2
11463	Stable	Deletion	2	na‡	2	1
13197	Stable	CN-LOH	4	na‡	2	2
12572	Instable	Normal	3	na‡	1	na
13030	Stable	Del 18q	3	na‡	2	2
11110	Stable	CN-LOH	4	na‡	4	2
10570	Stable	Deletion	2	na‡	2	1
11349	Stable	na†	2	na‡	na‡	2
11370	Stable	na†	2	na‡	na‡	2
10790	Stable	na†	2	na‡	na‡	2
10808	Stable	na†	4	na‡	na‡	MLE[£] did not converge
10811	Stable	na†	4	na‡	na‡	MLE did not converge
10858	Stable	na†	3	na‡	na‡	MLE did not converge
11806	Stable	na†	2	na‡	na‡	2
11813	Stable	na†	3	na‡	na‡	MLE did not converge
11877	Stable	na†	2	na‡	na‡	MLE did not converge
11300	Instable	na†	3	na‡	na‡	na
11907	Instable	na†	3	na‡	na‡	na
12572	Instable	na†	3	na‡	na‡	na
13045	Instable	na†	3	na‡	na‡	na
13120	Instable	na†	3	na‡	na‡	na

Run on both genotyping platforms

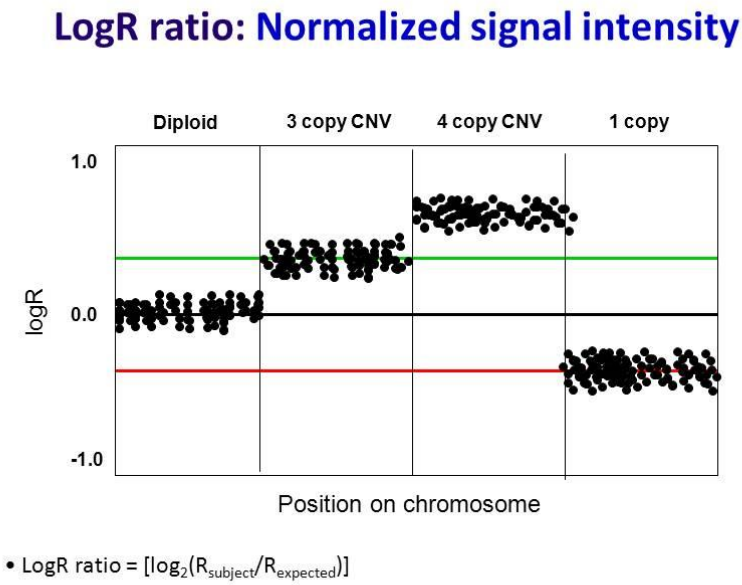
Run on Affymetrix 6.0 platform only

**BAF determined by Illumina genotype data (see Figure 1.1)*
†Tumor samples not run on Illumina platform
$BAF determined by mixture modeling (see Figure 1.3)
‡Data not usable due to high background noise
$Copy-neutral loss of heterozygosity
[¥]Log-R Ratio
[£]Maximum likelihood estimate

Table 3.2: Gene expression in candidate genes in tumors with LOH on 18q. There was no significant difference between expression of MSI tumors versus MSS tumors that had loss of chromosome 18.

Gene	Location	Probe Position Start	Probe Position End	MSI Mean (n=6)	MSS Mean (n=7)
SMAD2	18q21.1	43620623	43677180	11.2423	11.3400
SMAD2	18q21.1	43621619	43677180	9.5854	9.4672
SMAD2	18q21.1	43622069	43677160	10.6510	10.4526
SMAD4	18q21.1	46810609	46861111	7.8732	7.8658
SMAD4	18q21.1	46827286	46859729	9.9604	9.8722
DCC	18q23.1	48121155	49311286	7.2770	7.3041

Figure 3.1: Visual representation of detection of copy number changes using logR ratio and B allele frequency. The top panel demonstrated what the LRR signal intensity looks like for various copies of a chromosome. The formula for the LRR is below the figure. The lower panel shows how BAF or allelic copy ratio is used to infer copy number.



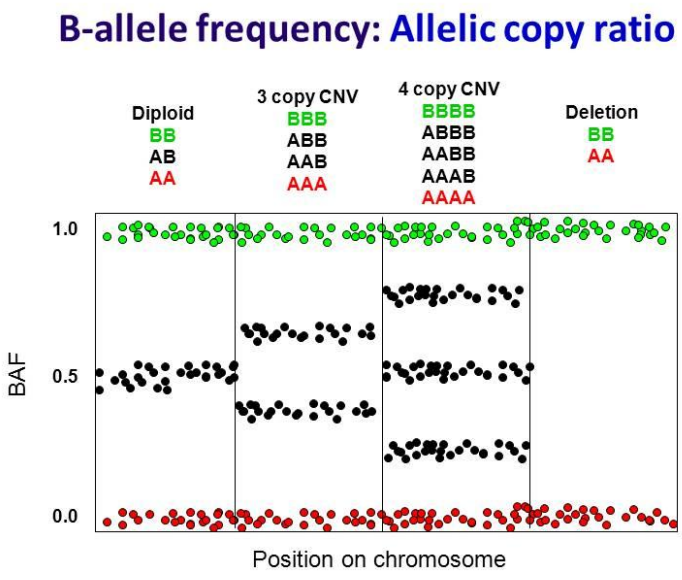
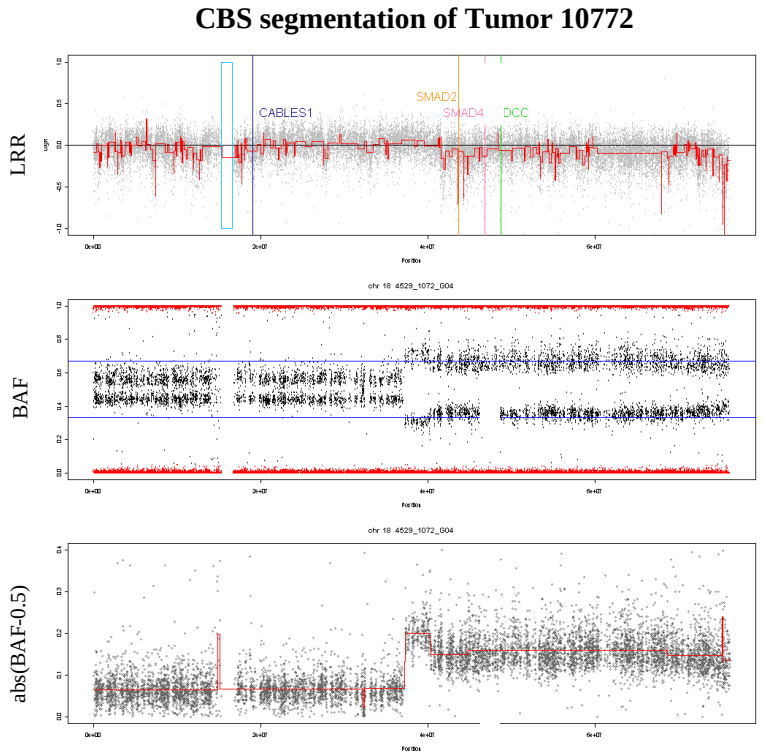
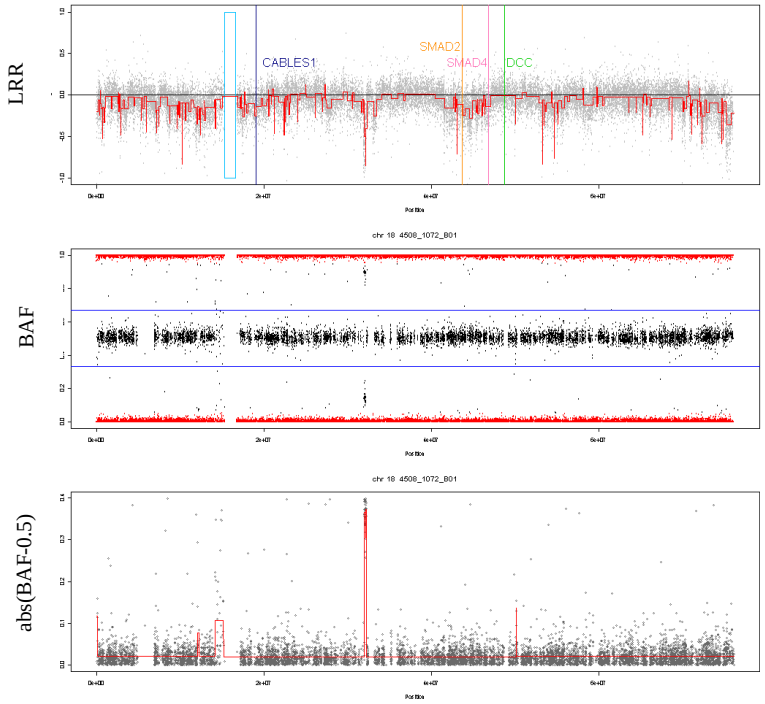


Figure 3.2 (A-H): Circular binary segmentation (CBS) on LRR, BAF and folded-BAF from 8 tumors samples run on Illumina genotyping platform. The following figures show results on chromosome 18 from segmentation for all tumors samples. The teal-colored box represents the separation of the p and q arms. The positions of candidate tumor suppressor genes are also shown on the figure.

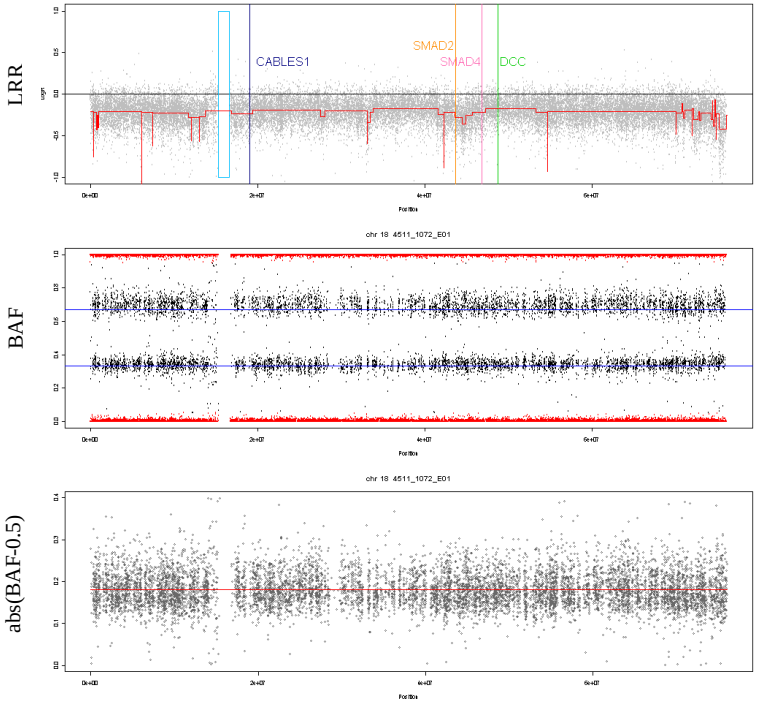
A.)



B.) **CBS segmentation of Tumor 12752**

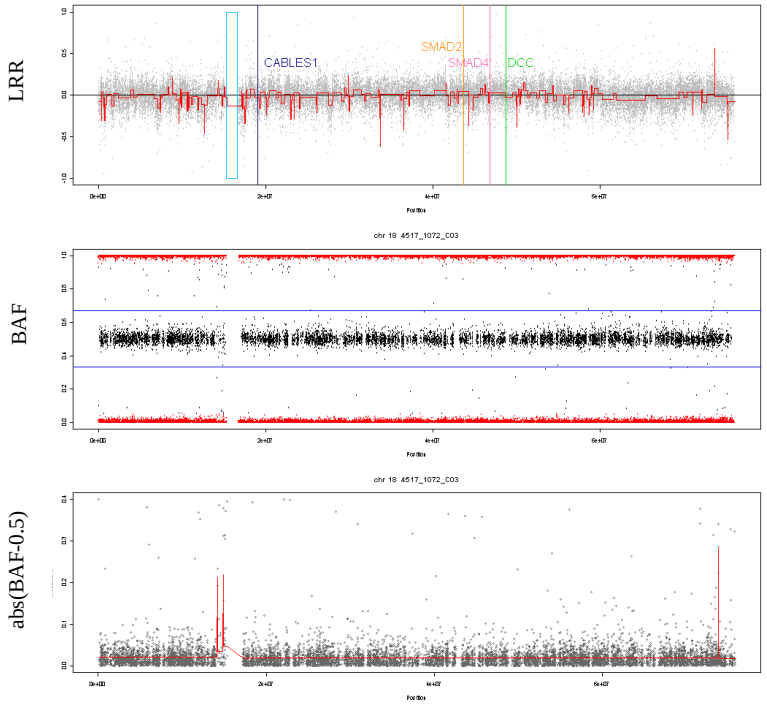


C.) **CBS segmentation of Tumor 11110**



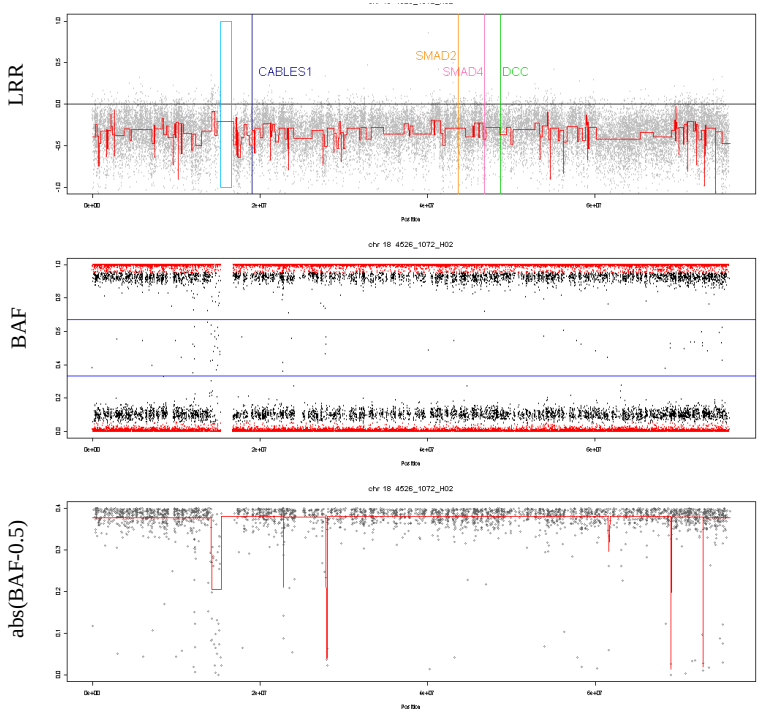
D.)

CBS segmentation of Tumor 12752



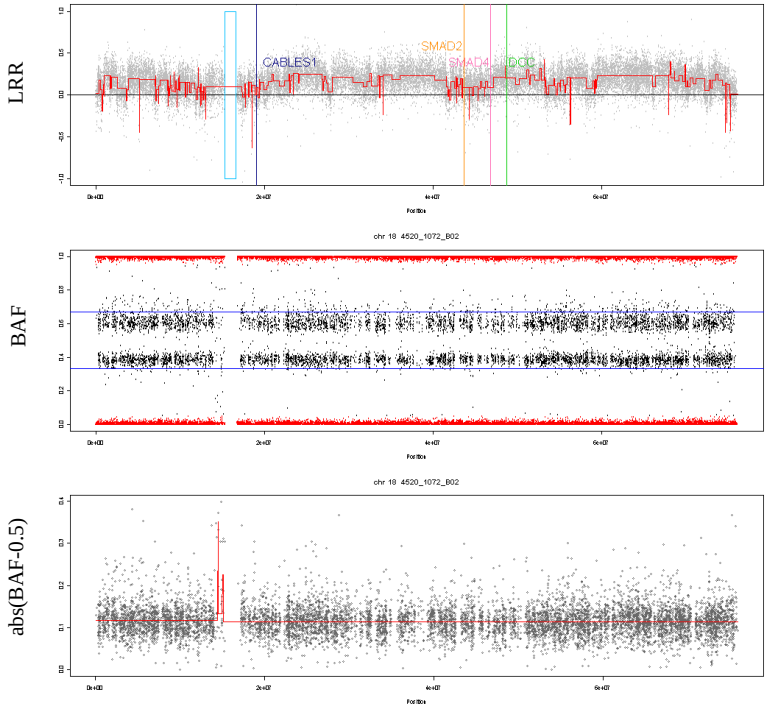
E.)

CBS segmentation of Tumor 10570



F.)

CBS segmentation of Tumor 10779



G.

CBS segmentation of Tumor 13030

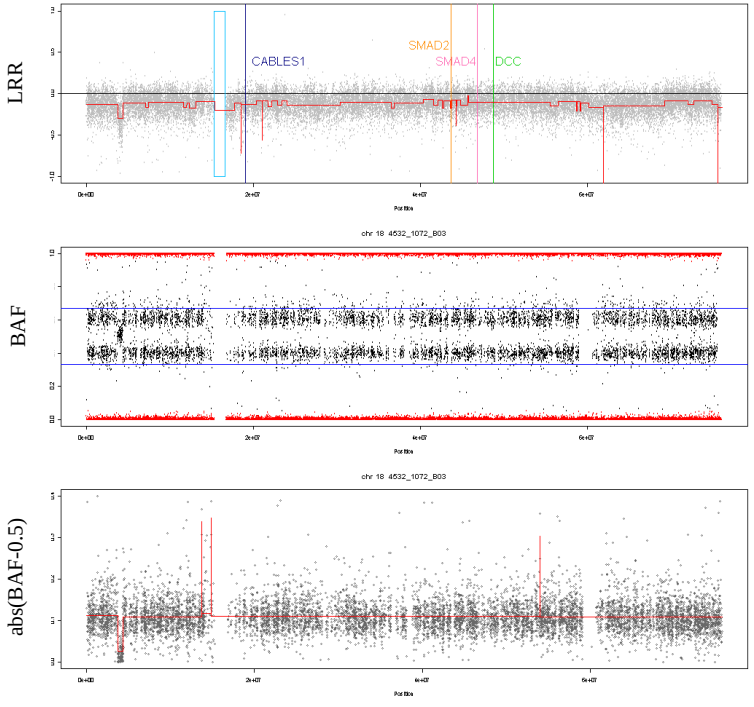
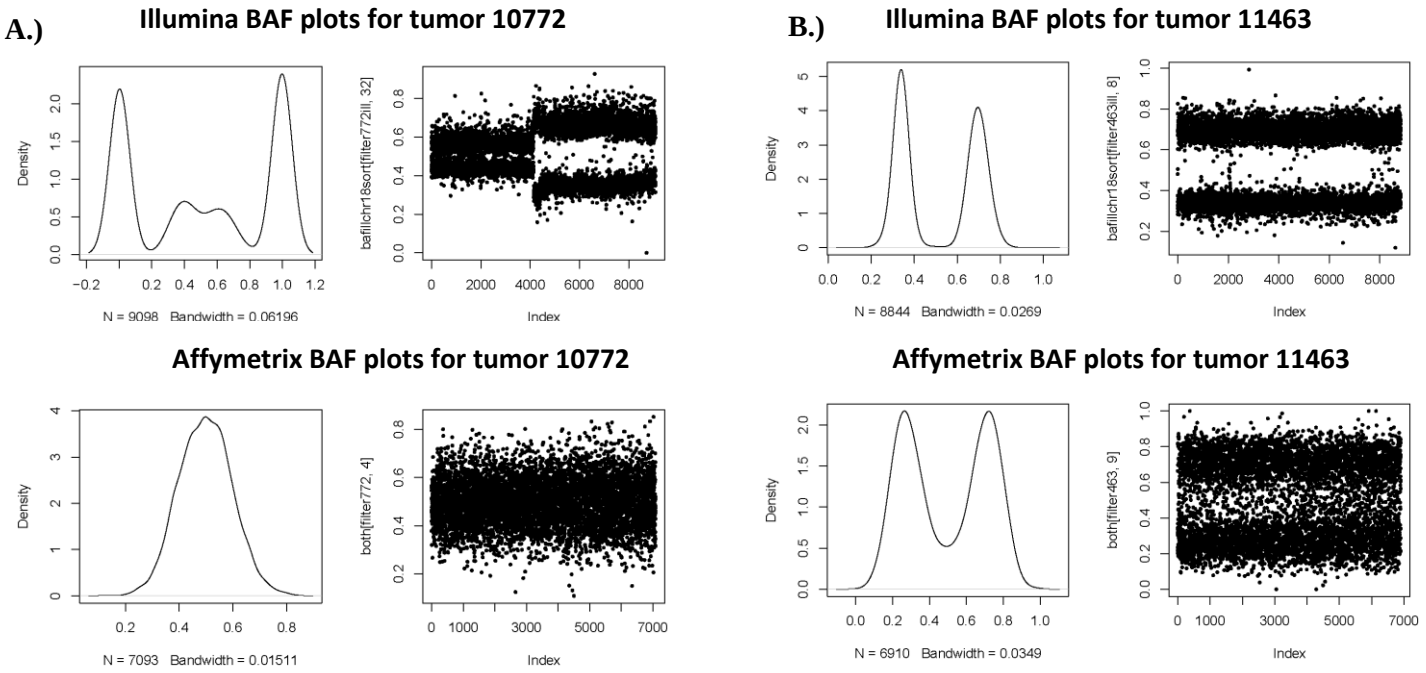
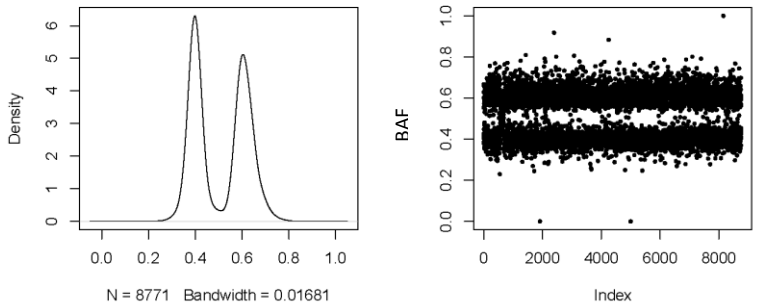


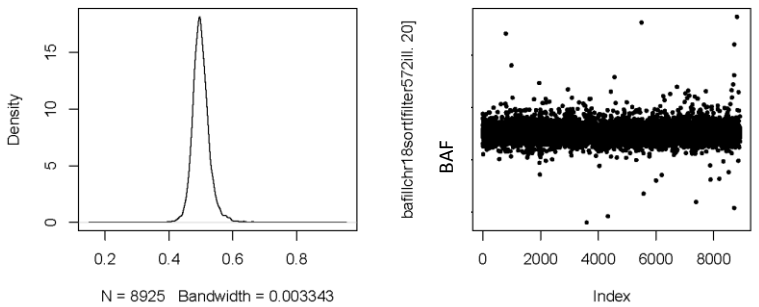
Figure 3.3 (A-C): Comparison of B allele frequencies in tumor samples from Affymetrix and Illumina genotyping platforms. The density plots of the BAF-heterozygous SNPs and the chromosome 18 plots for each tumor sample are shown below. The panels below show first the tumor run on the Illumina platform and then the same tumor sample underneath that was run on the Affymetrix platform. The Affymetrix data clearly has much higher background fluorescence intensities compared to the Illumina samples, making it very difficult to assess copy number change.



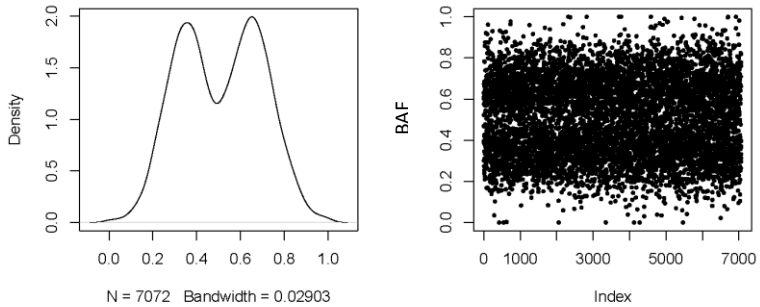
C.) **Illumina BAF plots for tumor 10779**



D.) **Illumina BAF plots for tumor 12572**



Affymetrix BAF plots for tumor 12572



Affymetrix BAF plots for tumor 12572

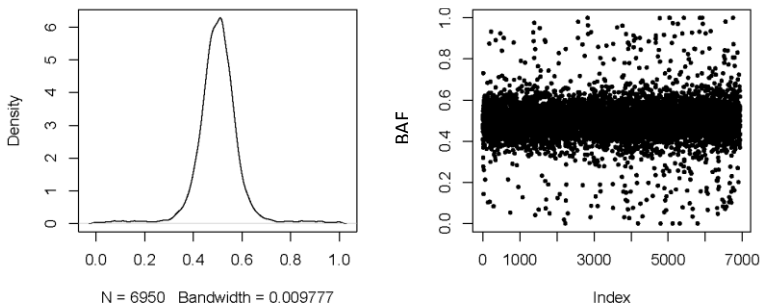
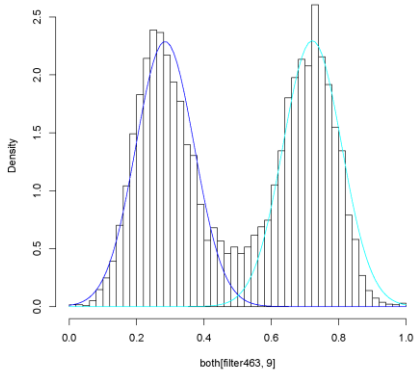
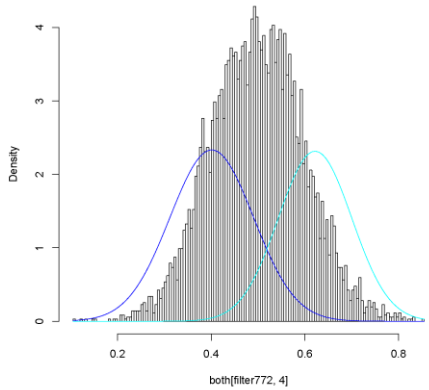


Figure 3.4: Detection of mixed Gaussian distributions based on BAF in tumor samples run on the Affymetrix platform with high background noise and the same tumor sample run on the Illumina platform. For the samples with loss of heterozygosity (10779, 11463 and 10772) there are two distinct normal distributions. The MSI tumor 12572 (with no LOH) has only one distribution.

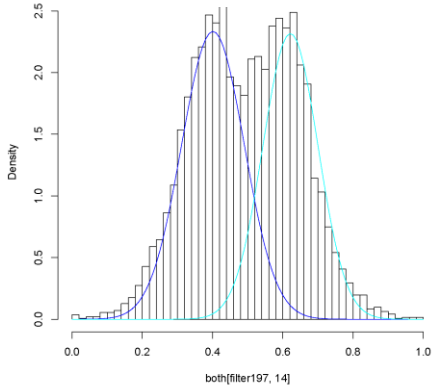
BAF data from tumor 11463



BAF data from tumor 12572



BAF data from tumor 10772



BAF data from tumor 10779

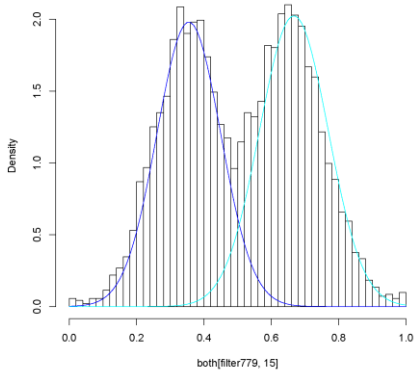
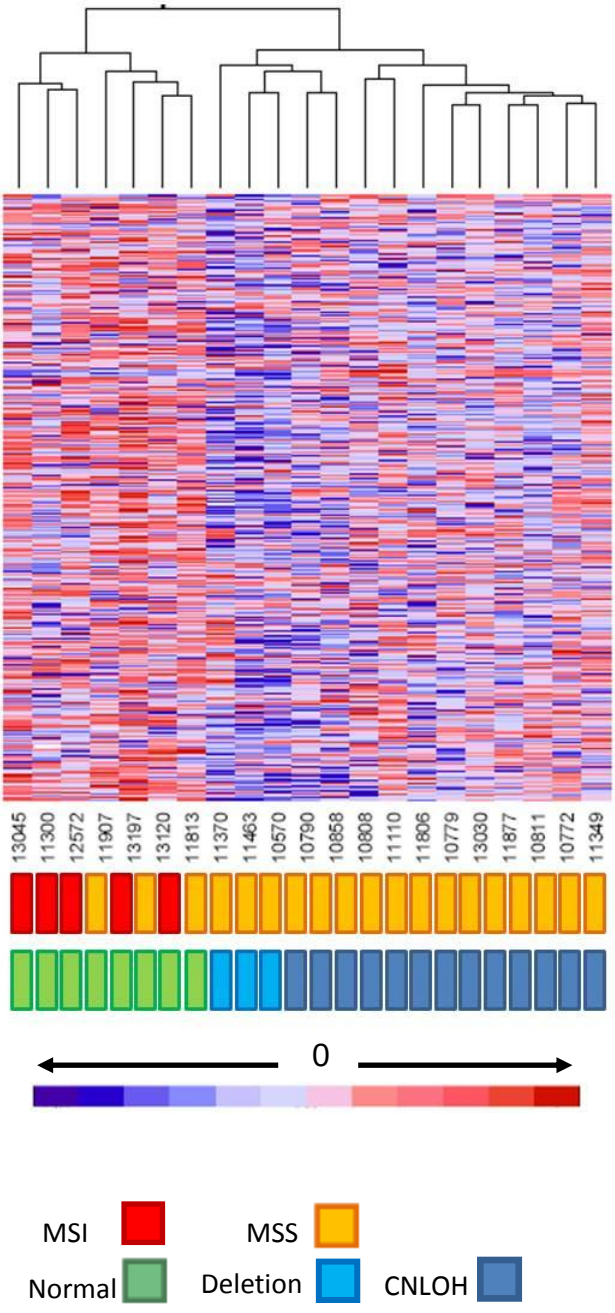


Figure 3.5: Hierarchical clustering of tumors based on chromosome 18 expression data. The figure below shows the two clusters solution (MSI vs. MSS). Below the heatmap of the expression data, the tumor instability is indicated in red (MSI) and yellow (MSS), and copy number in green (normal), blue (deletion) and purple (CN-LOH). Decreased expression can be seen in the heatmap in the three deleted samples.



CHAPTER 4

Genetic susceptibility to breast cancer among consanguineous individuals of Arab and Jewish ancestry

4.1 Introduction

Breast cancer is associated with several classes of risk factors, including hormonal and reproductive patterns however, family history is arguably one of the most important risk factors for the disease. Highly penetrant mutations in the known breast cancer genes *BRCA1* and *BRCA2* only account for 20–40% of familial breast cancer and less than 10% of breast cancer overall. These genes certainly do not explain all of the genetic variation observed in familial breast cancer, which suggests that familial breast cancer may be due to other genes (Ford et al 1998, Wacholder et al 2010).

Several studies have attempted to predict modes of familial cancer inheritance using segregation analysis. Most analyses of breast cancer inheritance support a model in which susceptibility to breast cancer is explained by a rare dominant disease allele with a high lifetime risk of the disease (Claus et al 1991). This model was confirmed by the mapping of *BRCA1* and *BRCA2* using genetic linkage analysis (Easton et al 1993, Hall et al 1990, Wooster et al 1994). Further evidence based on population studies has suggested that mutations in these genes only account for the minority of the overall familial risk of breast cancer (Peto et al 1999). Instead these studies propose that low penetrance, common genetic variants contribute to the risk of familial breast cancer (Pharoah et al 2002). For example, a population based series of breast

cancer cases investigated genetic models underlying familial breast cancer in individuals diagnosed in patients under the age of 55 (Antoniou et al 2001). The basic model indicated that either a common, recessive locus or a number of common, low-penetrance genes with additive effects likely account for residual non-*BRCA1/2* familial breast.

Breast cancer is the most commonly diagnosed cancer in Israel, and a leading cause of death among women regardless of their ethnic origin. Younger ages at diagnosis, larger primary tumor size, and lower 5-year survival rate have been reported among Arab women compared with the Jewish women (El Saghir et al 2006). According to the most recent reports of the Israel Cancer Registry data from 2007, age standardized rates of breast cancer are higher among Jewish women (87.72 per 100,000) than among Arab women (73.19 per 100,000) (http://www.health.gov.il). Variability in rates and characteristics of disease could partially be explained by differences in lifestyle factors between Arab and Jewish women. Another explanation for discrepancies such as the younger mean age at diagnosis among Arab women is that there may be previously undescribed mutations in the *BRCA1/2* genes or in other genes that lead to early-onset breast cancer.

Breast cancer among Arab women in Israel is characterized by high rates of parental consanguinity and large families, which are advantageous for family-based studies. Parental consanguinity ranges from 25% to 60% in different countries in the Middle East (Bener and Alali 2006). Endogamy in Arab populations is a result of cultural and historical practices, rather than religion. Consanguineous marriages are considered more stable in terms of maintaining family finances as well as family structure (Teebi and Teebi 2005) .

The contribution of parental consanguinity to cancer risk is an understudied area. However, there is some evidence that consanguinity augments autosomal recessive alleles due to

the increase in probability of sharing alleles identical-by-descent. The suggestion that cancer might have a non-*BRCA1/2* basis is based on several reports and other examples in cancer genetics such as, *CHEK2*, *MYH*-associated polyposis, Bloom syndrome and Fanconi anemia (Assie et al 2008, Ellis et al 1995, Lebel and Gallagher 1989, Meijers-Heijboer et al 2002, Sampson et al 2003). A study evaluating the effect of inbreeding on cancer incidence on isolated islands in Croatia found an increase in breast cancer rates primarily in the younger age groups (Rudan 1999, Rudan et al 1999). Several studies specific to Arab populations have suggested that consanguinity alters the genotype frequencies in offspring of consanguineous parents and as a result change the risk of early-onset breast cancer and possibly other malignancies (Bener et al 2001, Liede et al 2002, Rudan 1999). There is a report that inbreeding may reduce breast cancer risk, however this study was only done on women ages 40-65, and when examining families with a history of breast cancer and consanguinity, the risk was not significantly different in families without consanguinity ($P = 0.29$) (Denic and Bener 2001). The role of inbreeding on the risk of complex diseases may therefore be population-dependent and/or disease-dependent.

One way of locating human genes that are associated with recessive traits in related families is by homozygosity mapping. This method allows for the detection of disease locus because adjacent regions influencing the disease will preferentially be homozygous by descent in offspring of a consanguineous mating if the disease is recessive. From a statistical standpoint, the length of a run of homozygosity depends on the degree of parental consanguinity, because it is reduced by recombination which breaks up chromosomal segments over several generations (Wang et al 2009). For example, a single affected child of a first-cousin marriage is shown to contain the same total information about linkage as a nuclear family with three affected children (Lander and Botstein 1987). However, even in outbred populations homozygous regions

exceeding 1Mb in length have been detected (Gibson et al 2006, Li et al 2006).

In this study, we use genome-wide single nucleotide polymorphism arrays for homozygosity mapping in Arab women without mutations in *BRCA1* and *BRCA2*, to identify candidate loci within high-risk individuals with familial breast cancer. The question I am attempting to address in this chapter is whether recessive inheritance to breast cancer exists in consanguineous individuals as a risk factor for disease. My hypothesis is that an increase in autozygous alleles in consanguineous individuals with breast cancer will lead to enrichment of autosomal recessive loci. The prior success of strategies using linkage analysis in highly informative families that do not carry *BRCA1*/*BRCA2* mutations highlights the value of this approach to identify susceptibility variants in familial breast cancer (Meijers-Heijboer et al 2002). The identification of recessive loci suggests that penetrant common genetic variants may contribute to the risk of familial breast cancer.

4.2 Subjects and Methods

4.2.1 Subjects

The Breast Cancer in Northern Israel Study is an ongoing population-based case control study in northern Israel of all incident cases of female breast cancer since January 1, 2000. Incident breast cancer cases were ascertained from five hospitals in northern Israel, where diagnoses of breast cancer were made independently by the diagnosing hospital. The controls were collected from the Clalit Health Services (CHS), located in the same geographical area as cases and individually matched for exact year of birth, sex, clinic, and ethnic group (Jewish versus Arab). Patients were excluded if they had a former diagnosis of a breast cancer in the same breast and controls were excluded if found to have had a prior diagnosis of breast cancer

(including ductal carcinoma in situ). Participants were given an in-person interview, which provided information about their personal and family history of cancer, reproductive history, medical history, exposure to radiation, medication use, and a dietary questionnaire. The study was approved by IRBs at the University of Michigan and Carmel Medical Center in Haifa, and study participants gave written informed consent.

The Familial Cancer Consultation Service, run by the Cancer Control Center and directed by Dr. Gad Rennert, provides counseling to patients, families, and health care providers regarding inherited susceptibility to cancer. Genetic testing for the three Ashkenazi founder mutations, *BRCA1* 185delAG, *BRCA1* 6174delT, and *BRCA2* 5382insC is offered at no cost to the patient or family member through a certified molecular diagnostics clinical laboratory. Mutation-negative families served as the resource for further gene discovery in this study.

The population size of Israel is approximately 7.4 million people. Seventy six% are Jews, 19.5% are Arabs, and 4.3% belong to other ethnic groups, including Bedouin and Druze. The proportion of Arabs in northern Israel is 39% vs. 19% in the total Israeli population. Among Arab women, nearly 75% are Moslems, 16.5% are Christians and 8.5% are Druze (Central Bureau of Statistics, http://www1.cbs.gov.il). The Arab population has distinct age distributions compared to the Jewish population. In the Jewish population, 12.5% of women are over age 65 as compared with 3.5% of Arab women. The percentage of participants in the Breast Cancer in Northern Israel Study who have self-reported related parents is 4.5%, 28.5%, and 30.1% among Jewish, Christian Arabs, and Muslim Arabs respectively (Table 4.1).

4.2.1.2 Pilot study

As part of a pilot study, 9 cases with breast cancer from the Breast Cancer in Northern Israel Study that were referred to the Familial Cancer Consultation Service and identified as having a sibling with breast cancer and a family history of consanguinity (Table 4.2 and Figure 4.1, A-J) Genetic testing for Ashkenazi founder mutations did not reveal mutations in *BRCA1/2* genes. These 9 individuals were genotyped using the Affymetrix 6.0 Genome-Wide Human SNP Array (see section 4.2.3.1).

4.2.1.3 Arab family expansion

An additional 50 DNA samples from 10 Arab families with a history of breast cancer and consanguinity were recently collected in Israel, in an attempt to expand the sample set (Table 4.3). Families were contact after being identified by having a family history of consanguinity and a sibling with breast cancer. A priority of collecting of sibling DNA (affected, and unaffected) and parents was taken, followed by any additional affect relatives.

4.2.2 Genomic DNA Isolation

Genomic DNA was extracted from whole blood using the Puregene kit (Gentra Systems, Inc., Minneapolis, MN). DNA samples were quantified using the ND-8000 spectrophotometer and PicoGreen assay (Molecular Probes Invitrogen Detection Technologies, Eugene, OR). The concentration for all qualified samples was normalized to 50 ng/ul.

4.2.3 Genome-wide single nucleotide polymorphism array

4.2.3.1 Affymetrix 6.0 array

The Affymetrix 6.0 Genome-Wide Human SNP Array (Affymetrix, Santa Clara, CA) arrays contain 1.8 million probes, including 906,600 SNP probes. Nine samples were run on arrays at the University of Michigan Comprehensive Cancer Center Affymetrix and Microarray Core as part of a pilot analysis. Two hundred and fifty nanograms of DNA were digested with StyI and NspI. After the restriction digestion, samples were PCR amplified and subsequently labeled with biotin before hybridization. All steps were performed according to the manufacturer's protocol.

4.2.3.2 Illumina HumanCytoSNP-12 BeadChip

The HumanCytoSNP-12 BeadChip is a whole-genome scanning panel designed for efficient, high-throughput analysis of genetic and structural variation. The BeadChip includes a complete panel of genome-wide tag SNPs including 200,000 SNPs with the highest tagging power. DNA samples from 22 cases and 24 unaffected family members were obtained and genotyped using the Illumina HumanCytoSNP-12 BeadChip. Probands we screened for mutations in *BRCA1/2* using a CLIA certified lab, Myriad Genetics (2011 Myriad Genetic Laboratories, Inc., 320 Wakara Way, Salt Lake City, UT 84108-1214). Comparative analysis of the coding sequence, intro-exon boundaries and deletion/duplication analysis of *BRCA1/2* was performed on affected individuals. Forty-six individuals from 5 families were genotyped using this platform at the University of Michigan DNA Sequencing Core. Two of these probands were genotyped as part of the pilot study and used for further quality control.

4.2.4 Homozygosity Mapping

Using identity-by-descent (IBD) methods to map genes in Mendelian disorders has proven to be a useful strategy in clinical genetics and may have potential in complex diseases. Homozygosity mapping identifies autosomal recessive genes in consanguineous families by detecting chromosomal regions that show homozygous IBD segments. I used a sliding window approach as implemented in PLINK for the analysis of IBD. The whole genome association analysis toolset, PLINK v1.05, was used to screen for runs of homozygous genotypes in all cases and unaffected family members (Purcell et al 2007). The data were filtered based on several quality control measures. Individuals were required to have a genotype for at least 95% of the loci, and an individual SNP was considered a failure if <95% of the samples generated a genotype. The overall genotyping call rate was 99.91%. One individual was genotyped twice on the Illumina array for quality control. To identify samples showing relatedness, identity-by-state (IBS) values were calculated for all pairs of individuals (Figure 3.2). Analysis was only performed on the autosomal SNPs. We excluded SNPs on the basis of deviation from Hardy-Weinberg equilibrium using a threshold of $P < 1 \times 10^{-5}$ in either the cases or controls. We also removed SNPs with a minor allele frequency of <0.05. Next, the remaining high quality SNPs were pruned for strong local linkage disequilibrium (LD) ($r^2 > 0.8$) and removed in order to find long segments that are more likely to represent homozygosity by descent (i.e. autozygosity) reducing the chance of an exaggeration of small random differences and in turn the production of false-positive results (Abecasis et al 2005). Approximately 122,000 SNPs were carried through for identifying runs of homozygosity. Fifty homozygous SNPs spanning a1000 kilobase distance were required for a homozygous region to be called. The algorithm for detecting runs of homozygosity (ROH) in PLINK uses a sliding window approach. Briefly, a window of 50 SNPs is taken and moved incrementally across the genome. At each window position, the region is

determined to be 'homozygous' based on the required criteria, allowing for 2 heterozygous or missing calls (due to genotyping error). Then, for each SNP, the proportion of 'homozygous' windows that overlap that position is calculated. The "homozyg-group" option was used to produce a file of the overlapping ROH regions separated into pools containing the number of cases and controls carrying the ROH. Pools with more than five samples were considered as recurrent ROHs. A consensus SNP set representing the minimal overlapping region across all samples in the pool was used to define the recurrent ROH regions.

4.2.4.2 Sanger sequencing

Candidate genes in regions identified as having a run of homozygosity were sequenced using the genomic DNA (Table 4.5). The PCR reaction mixtures (20μL) contained 5ng of genomic DNA, 2μl of 10X PCR buffer (Applied Biosystems), 1.6μL of 25mM MgCl2 (Applied Biosystems), 0.8μL each of 10mM dNTP (New England Biolabs) and 10μM forward and reverse primers, and 1 U of AmpliTaq Gold DNA polymerase (Applied Biosystems). Cycling conditions were as follows: Initial denaturation at 95°C for 3 minutes, 15 cycles of 95°C for 30 seconds, 70°C for 45 seconds (-1° every cycle), 72°C for 1 minute 10 seconds, 20 cycles of 95°C for 30 seconds, 55°C for 45 seconds, 72°C for 1 minute 10 seconds, and a final extension at 72°C for 10 minute. PCR products were sequenced at the University of Michigan DNA Sequencing Core, and Mutation Surveyor Software (SoftGenetics, LLC., State College, PA, USA).

4.3 Results

4.3.1 Pilot study homozygosity mapping

Nine DNA samples with Affymetrix 6.0 genome-wide SNP array data were analyzed as part of a pilot study using the program PLINK to screen for runs of homozygosity of at least 1000 kilobases in size. Five individuals had large runs of homozygous SNPs, including two Arab subjects. The largest region of homozygosity detected was on chromosome 10q23.1-25.3, which harbors the known tumor suppressor gene *PTEN*. All analyses were repeated after additional samples were genotyped on the Illumina platform (see section 4.3.2 and 4.3.3). This region was detect on the Illumina platform, but did not meet the criteria for preliminary analysis that priority regions of overlapping homozygosity must occur in at least 6 different individuals from different families.

4.3.2 Arab family expansion homozygosity mapping

In an attempt to expand the sample size of Arab families with a history of consanguinity and a sibling with breast cancer, additional DNA samples were collected from both affected and unaffected family members (see section 4.2.1.3). Forty-six DNA samples with Illumina Human CytoSNP-12 genotypes, including the 9 samples from the pilot study were analyzed using PLINK to detect large stretches of homozygosity. Runs of homozygosity (ROH) thresholds were set based on genomic regions in which a minimum number of consecutive, non-missing SNPs were homozygous. ROH was measured per individual in terms of their total length or the sum of the lengths of the ROHs found in each person (Spain et al 2009). The frequencies of detected ROH of ≥5 Mb were calculated in cases and control. Among affected individuals, 20 of 22 (90%) had at least one ROH of >5Mb and 15 of 25 (60%) unaffected family members had ROHs of >5Mb (P = 0.05242, Fisher's exact test), similar to previous findings in consanguineous individuals with cancer. (Bacolod et al 2009, Spain et al 2009).

The primary question of interest in this study is whether homozygous regions harbor recessive loci associated with breast cancer risk in consanguineous individuals. Using runs of homozygosity with ≥1 Mb of consecutive homozygous SNPs determined minimal overlapping ROH regions that were found in at least six affected individuals from different families (Table 4.4). A 242 kilobase region on chromosome 9q33.2-33.3 was present in affected cases only and not in any of the unaffected family members (Figure 4.3). The only known gene or coding sequence present in this overlapping, homozygous region is *LHX2* (Figure 4.4).

4.3.3 Investigation of candidate locus using Sanger sequencing

All coding region (+/- 200 bp) of the gene *LHX2* were sequenced in the six individuals with the overlapping run of homozygosity. No variation differing from the reference CEPH sample sequence was detected in any of the individuals. Sequence was analyzed using both Mutation Surveyor and Sequencer software (*section 4.2.4.2*).

4.4 Discussion

Recent studies have reported an increased frequency in runs of homozygosity in cancer cases (Assie et al 2008, Bacolod et al 2009) Additionally, genetic modeling suggests that either a polygenic model of common, low-penetrant genes or Mendelian inheritance of an autosomal recessive allele account for non-*BRCA1/2* familial aggregation of breast cancer. The idea that dominant disorders currently outnumber recessive disorders in humans, may represent an artifact of the clinical appearance of genetic disorders in the outbred population of Western society, is one potential explanation for there have not been many reports of autosomal recessive genes for breast cancer (Teebi and Farag 1997). However, breast cancer among Arab women in Israel is

characterized by high rates of parental consanguinity and large families, which are advantageous for family-based studies and allow for the enrichment of recessive loci, making this population ideal for identifying recessive genes associated with breast cancer. Families were also selected not only on the basis of having a family history of consanguinity but also a sibling with breast cancer and an unaffected mother, increasing the likelihood that genes associated with disease in these families are inherited in an autosomal recessive fashion.

Homozygosity mapping allows for the detection of recessive genes associated with the disease locus due to the fact that adjacent regions on chromosomes will preferentially be homozygous by descent in offspring of a consanguineous mating. In the current study, we identified a run of homozygosity in Arab and Jewish women with a family history of consanguinity and a sibling with breast cancer on 9q33.2-33.3. This region only contains one gene, *LHX2,* a putative transcription factor containing two cystein-rich (LIM) motifs and a homeobox (HOX) DNA-binding domain. Recent genome-wide methylation studies have suggested a role for *LHX2* methylation in breast and lung cancer (Kamalakaran et al 2011, Rauch et al 2006). Analysis of these tumors using follow-up survival data identified differential methylation of islands proximal to genes involved in cell fate commitment, including *LHX2,* as having prognostic value independent of subtypes and other clinical factors associated with breast cancer. Changes in methylation are commonly seen in human tumors, and several studies have implicated a role for DNA methylation in cancer pathogenesis (Laird and Jaenisch 1994, Laird et al 1995). It has also been found that methylation varies in different tumors. For example, some loci tend to show increased levels of DNA methylation while others have found a decrease in levels of methylation (Issa et al 1994, Ohtani-Fujita et al 1993, Wahlfors et al 1992). It is suggested that changes in DNA methylation that contribute to

oncogenesis affect the expression levels by increasing the expression of oncogenes and that hypermethylation silences tumor-suppressor genes (Baylin et al 1991).

Although we did not detect any mutations in the exonic regions of *LHX2*, there are several possibilities for why dysregulation of this gene may be associated with in breast cancer. Enhancers can activate transcription independent of their location, distance or orientation with respect to the promoters of genes (Ong and Corces 2011). Homozygous variation occurs in the enhancers or promoter region may affect their interaction and in turn the expression of *LHX2*. Another possibility is that the homozygous mutations in enhancers or other sequence-specific DNA binding proteins may lead to the inability of histone chaperones or modifying enzyme to be recruited and therefore they will be unable to modifiy chromatin or other epigenetic marks (Cui et al 2009, Heintzman et al 2009). Methods such as CHIP-seq may be useful for determining alteration in expression of non-coding regions of *LHX2* (Johnson et al 2007).

One criticism of the current study may be the relatively small sample size due to the limited number of individuals in an isolated community or a single large family with a high level of inbreeding. Therefore, the relevance of inbreeding to the population risk of cancer is unclear and inbreeding and founder effects may be confounded. A limitation to the homozygosity mapping analysis is that it is difficult to determine the exact window size and thresholds, relative to the SNP density and expected size of homozygous segments, even though this is obviously important. This study is limited to detection of large segments. Another criticism is that the overlapping ROH on chromosome 9 identified in the current study may be a false positive due to large amounts of IBS across the genome. One strategy to quantify this possibility would be to conduct a simulation study to assess the statistical significance of the type of ROH that we are

observing arising by chance (Clark 1999, Wang et al 2005). For more accurate detection of smaller segments it might make more sense to use an approach that also takes population parameters such as allele frequency and recombination rate into account (Scharpf et al 2008).

Although there may be some inconsistency in evidence for a role of consanguinity in risk of cancer, the study that demonstrated a tendency of consanguinity to decrease the risk of breast cancer should be interpreted with caution (Denic and Bener 2001, Denic et al 2005). The first study was restricted to women aged 40 to 65, and found no significant difference between women with a family history of breast cancer and consanguinity, and the second study lacked the power to detect an association. Additionally, one cannot exclude the possibility that the inconsistencies in findings are due to different genetic backgrounds.

In summary, the goal of the current study was to evaluate the contribution of consanguinity to breast cancer risk in Arab women without mutations in *BRCA1* and *BRCA2*. An increase in autosomal recessive genes responsible for genetic susceptibility to breast cancer is expected among families with consanguinity due to the increase in probability of sharing alleles identical-by-descent. Homozygosity mapping was performed in consanguineous breast cancer patients to identify shared genomic regions of disease susceptibility. Six individuals with breast cancer had 242kb overlapping run of homozygous SNPs on chromosome 9q33.2-33.3 which harbors a potentially important candidate gene in cancer, *LHX2*.

4.5 Future Directions

A tumor block from one individual with breast cancer and the candidate, homozygous region on chromosome 9q33.2-33.3 is available for future analysis. We plan to perform IHC and RT-PCR to determine if *LHX2* is expressed in the tumor, and if expression levels are reduced in

tumor as compared to normal breast tissue. Expression of *LHX2*, is has previously been detected in breast tissue, however how it is expressed in tumors from individuals with breast cancer has not been determined.

Table 4.1: Participants in the Breast Cancer in Northern Israel Study

	Jewish		Christian Arab		Muslim Arab	
	Cases (n=1418)	**Controls (n=1118)**	**Cases (n=97)**	**Controls (n=59)**	**Cases (n=177)**	**Controls (n=161)**
Age-mean (yrs)±SD	60.5±12.8	60.9±12.7	55.25±12.6	56.95±12.5	49.0±10.6	50.4±11.3
Post menopausal (%)	81.8	77.1	77.3	66.1	58.8	39.1
Age at menarche- mean (yrs)±SD	14.4±2.7	13.7±2.9	14.1±2.8	13±2.4	14.0±2.5	13.5±1.6
Age at first pregnancy- mean (yrs) ±SD	23.5±4.5	23.0±4.2	22.6±3.9	23.2±2.8	22.5±5.5	21.8±4.1
Number of children- mean±SD	2.5±1.3	2.8±1.4	4.2±1.9	4.5±2.1	5.2±2.9	6.0±2.7

Table 4.2 Pilot study breast cancer cases and consanguineous relationships

	Muslim Arab	**Christian Arab**
Cases (n=9)	5 (55.4%)	4 (44.5%)
1st Cousins (n=5)	4 (80%)	1(25%)
2nd Cousins (n=4)	1 (20%)	3(75%)

Table 4.3 Breast cancer cases and unaffected relatives recruited as part of the expansion sample

	Muslim Arab	Bedouin	Christian Arab	Sephardi Jewish	Ashkenazi Jewish	Males	Females
Cases (n=22)	8 (36%)	3 (14%)	8 (36%)	2 (9%)	1 (5%)	0	22
Unaffected (n=25)	18 (72%)	na	7 (28%)	na	na	6	19
Genotyping failure (n=1)	na	na	1	na	na	na	1
Full sib pairs (n=100)	81 (81%)	3 (3%)	16 (16%)	na	na	na	na
Parent/Offspring pairs (n=18)	12 (67%)	6 (33%)	na	na	na	na	na
Avuncular pairs (n=24)	24 (100%)	na	na	na	na	na	na

Table 4.4: Homozygous region of overlap in breast cancer cases

FID	IID	CHR	SNP1	SNP2	KB	NSNP
1001	8874	9	rs10759229	rs10858284	27686.5	1301
1005	8880	9	rs7851754	rs1614329	20898.9	875
1007	8882	9	rs7037974	rs2789769	15390.1	610
1004	8879	9	rs4836798	rs7034837	4154.24	114
4958	8855	9	rs10984840	rs11793247	15319.1	721
4955	8872	9	rs10760302	rs11792632	12425.8	715
overlap	**6**	**9**	**rs10760302**	**rs7034837**	**242.472**	**20**

FID = family ID, IID = individual ID, CHR= chromosome

Table 4.5: Primers for sequencing exonic regions of *LHX2* on chromosome 9.

LHX2	Start	End	Forward Primer	Reverse Primer	bp
Exon 1	125,814,193	125,814,708	CTTGTGACCCTGGCTTTGG	GCCTTGCATTCTGACCGAG	516
Exon 2	125,815,969	125,816,359	CACAGAGGGAGTTGTGGGTG	CTCCTGGGACTAAACGGAAAG	391
Exon 3	125,817,174	125,817,709	GTACCCAACCGTGTGTTCCC	CAGAGATTCAATCCAGCTCCC	536
Exon 4	125,823,104	125,823,515	GGATTGAAATGTTTGGCAGTG	AGAGAAGCAGACACAGGGTGG	412
Exon 5	125,834,444	125,834,942	GAGCTCTGAGTGAAGCAGTCG	TTACCTCTGTTTCCAGGCGAG	499

Figure 4.1: Pedigree 41983. This Muslim Arab family has a history of consanguinity (1st cousins). The proband has a sibling with a diagnosis with breast cancer, and an unaffected mother.

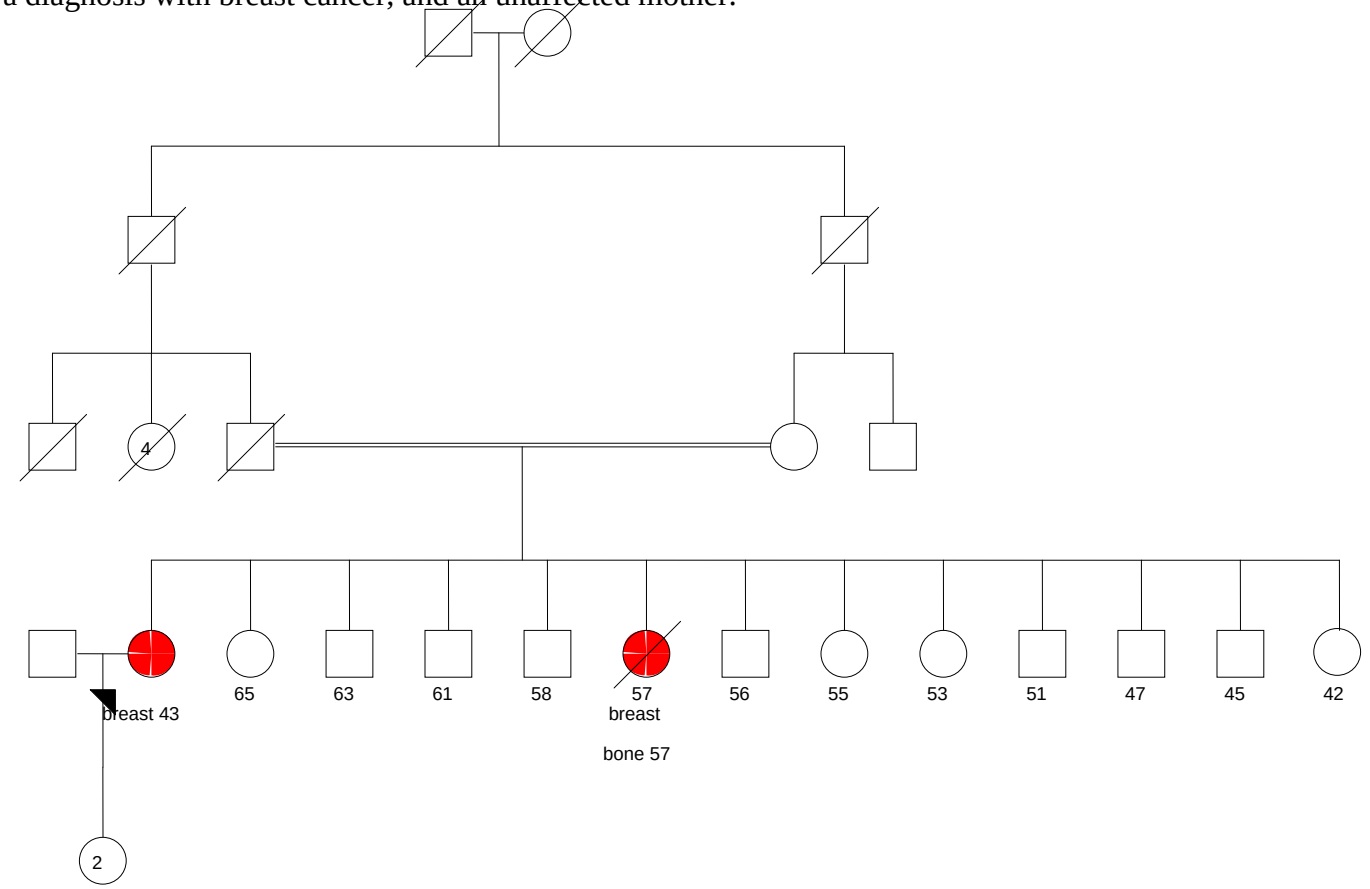


Figure 4.2: Pairwise (1-IBS) distance matrix of family expansion sample. This is the proportion of alleles IBS across all SNPs for a given pair of individuals. Values of 0 would indicate sample duplication. The distance matrix identifies clusters of individuals of greater similarity, corresponding to members of the same family.

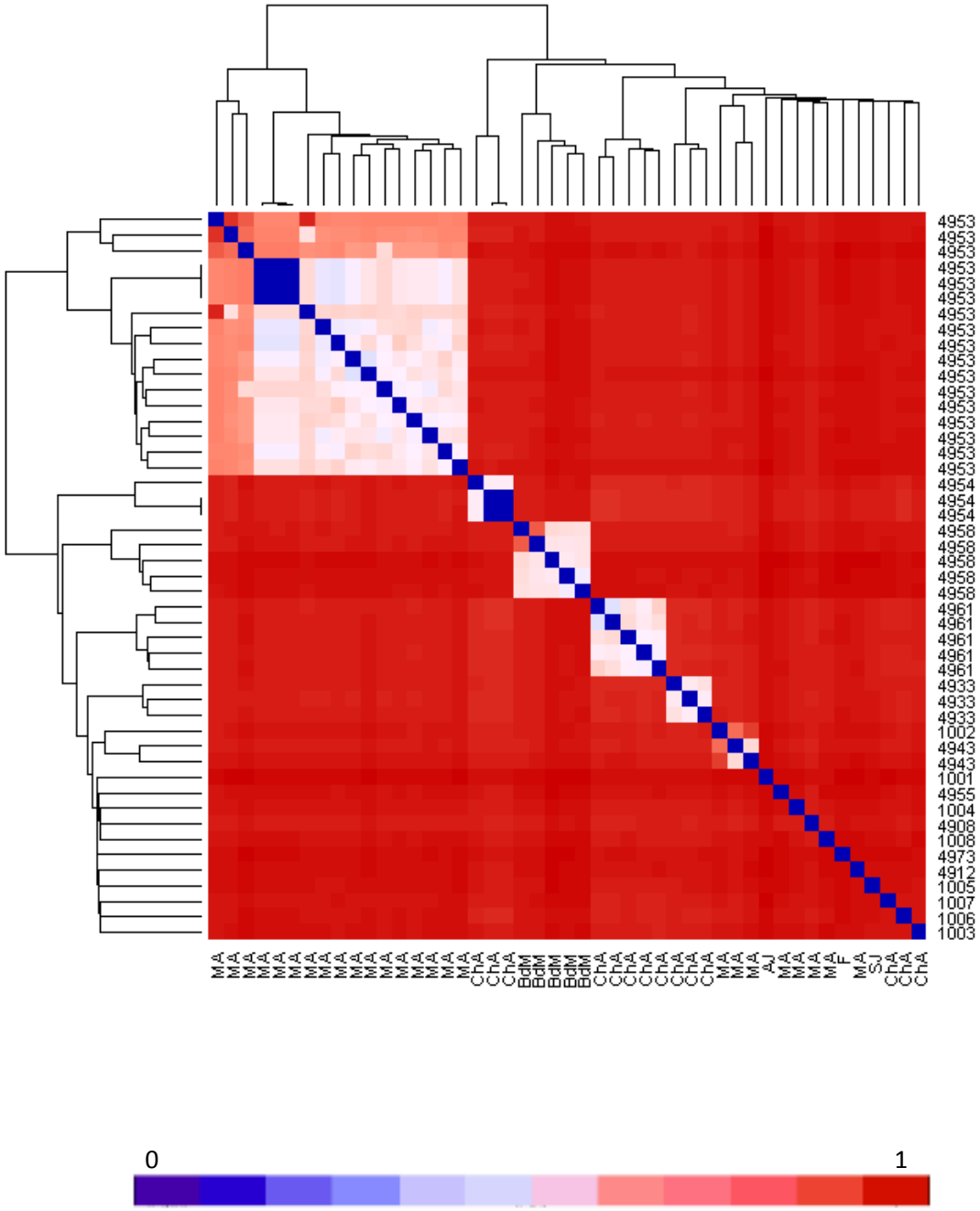


Figure 4.3: Overlapping runs of homozygosity by affection status and family. This figure shows the overlapping ROH in more than 5 individuals by affection status where affected is blue and unaffected is green. Runs of the same color are in the same family. I focused on regions of overlap in individuals from different families. It can be seen that the region on chromosome 9 has 6 different affected individuals from 6 different families.

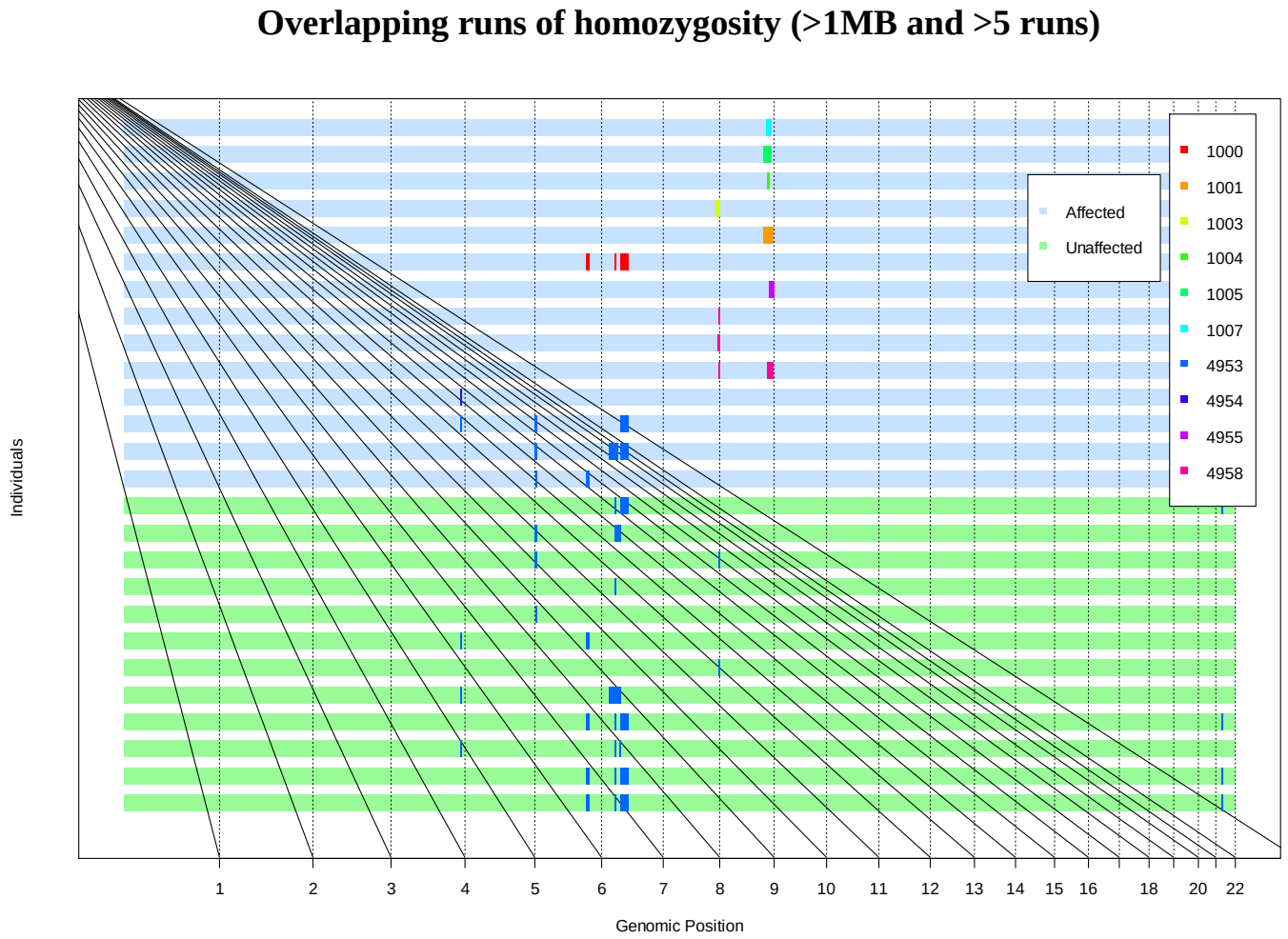
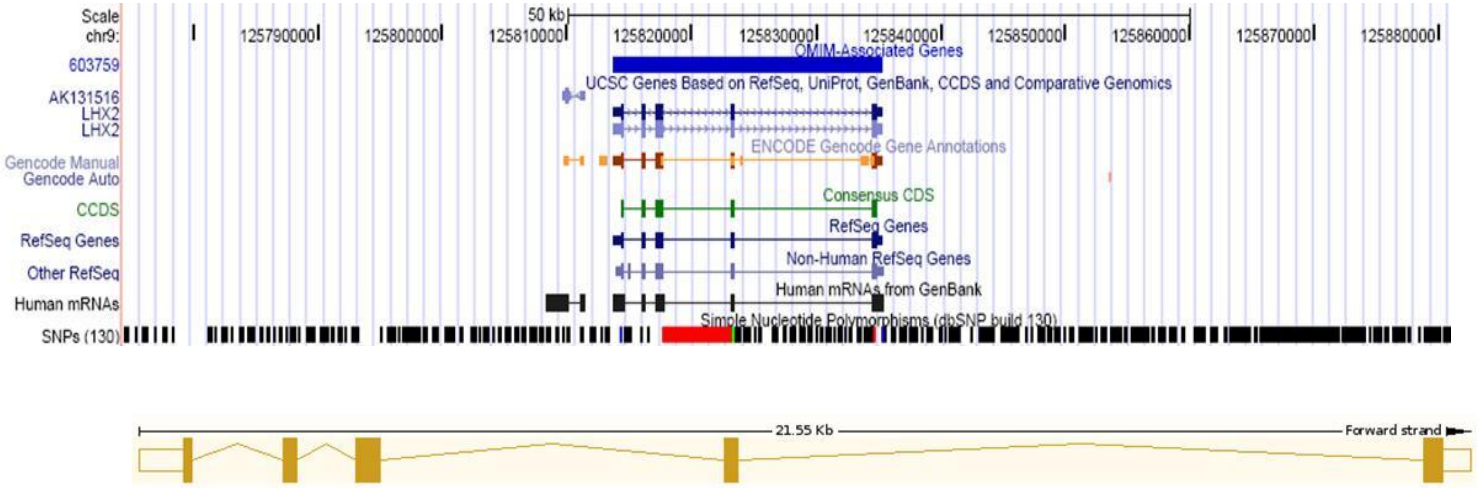


Figure 4.4: Overlapping region of homozygosity on Chromosome 9q33.2-33.3 which contains the candidate gene *LHX2*. This figure is taken from UCSC genome browser in the region where 6 individuals with breast cancer had an overlapping run of homozygosity. The only gene present in this region is *LHX2*. Below the UCSC figure is a schematic of *LHX2* from Ensemble, showing that this gene has 5 exons.



CHAPTER 5

Conclusions

In this dissertation, I have applied several quantitative approaches to understanding the cancer genome, which allow for the discovery of the organization and function. I have gained insight into the role of genetic variation in the colon and breast cancer genomes through the recent advances in technology, and bioinformatic methods.

I investigated the genetic basis of extremely rare, highly penetrant, novel variants in a family with hereditary mixed polyposis syndrome using next-generation sequencing techniques and the functional role of a novel variant in the candidate gene, *ZNF426*. Specifically, using the Illumina Genome Analyzer to sequence two affected related individuals with clinically and pathologically diagnosed HMPS, a total of 595,292,661 paired-end reads (76-120bp each) were mapped to the UCSC assembly hg19, with an average of 10x coverage of the haploid genome. The Genome Analysis Toolkit (GATK), which is a structured programming framework designed to aid in the efficient and robust analysis tools, was used for DNA sequence data analysis. I identified 1,162,925 variants were in common between the father and son, of which 125,460 were not present in 1000 genomes or dbSNP 131. After quality control filtration and annotation, 64 previously unidentified, nonsense, missense or splice site variants were identified between the two family members. Nine of the 11 novel candidate variants were next validated by Sanger sequencing. Six out of the 9 (67%) variants were shared by four of the affected family members, leading to a small subset of candidate genes putatively responsible for HMPS within this family.

One nonsense variant that was identified by next-generation sequencing in the father and son in *ZNF426* was also present in the two affected daughters. Tumor DNA as well as DNA extracted from normal adjacent tissue for the father and son were sequenced for variation in *ZNF426*. The novel variant was present in the tumor samples, however no loss of heterozygosity was detected. Analysis from quantitative real time PCR revealed decreased expression in the carcinoma tissue from the father, yet the adenomas and normal adjacent tissue were expressed at comparable level in both the father and son. This study offers new insight on a novel region on 19p13.2, that may be involved with susceptibility HMPS.

Chapter three examines the role of colon tumor heterogeneity on chromosome 18 in CRC, with much more depth and insight then previous low-resolution microsatellite markers. In this chapter, I determined copy number alterations using high-density, genome-wide SNP arrays using circular binary segmentation, which incorporated both the LRR and BAF information from SNP genotyping arrays. I also visually inspected the LRR and BAF plots to see the extent of the region of loss. Based on this information, I determined that 14 out of 21 tumors demonstrated instability on chromosomes 18, resulting in loss of whole chromosomes or deletion of only several kilobases. Eleven tumors from either the Illumina or the Affy genotyping platform demonstrated copy neutral loss of heterozygosity, 3 tumors had a complete deletion of at least one arm of chromosome 18. Due to poor data quality on the Affy platform, resulting in high background noise, we fit the distribution of BAF values in each segment of copy alteration as either one Gaussian distribution or the summation of two Gaussian distributions. Eight tumor samples were run on both genotyping platforms, so I was able to use those samples as confirmation of alterations detected even in the high background samples. The ‘useable’ samples size was increased from 8 tumors to 21 tumors using this approach. Additionally expression data

on colorectal tumors from the U133APlus expression array was assessed for decrease in expression levels in regions of loss on chromosome 18 in CIN tumors. Mean expression across all probes on chromosome 18 was significantly lower (p-value = 0.000137) for samples with complete loss of one arm of chromosome 18 compared to the tumor samples with 'normal' copy numbers (including the tumors samples with copy neutral loss of heterozygosity and the microsatellite stable tumors). Hierarchical clustering analysis and K-means clustering was performed on the expression data. The hierarchical clustering analysis produced a 2-cluster solution, clearly separating microsatellite instable from chromosomally instable tumors and suggested grouping of the 3 samples with complete deletion of at least one arm of chromosome 18 and the samples with 'normal' copy numbers (including tumors with CN-LOH and neutral copy state). New insights revealed in this chapter are the statistical methods that can improve yield of poor quality, high background data and the further refinement of the regions on chromosome 18 that distinguish chromosomally instable colorectal cancer.

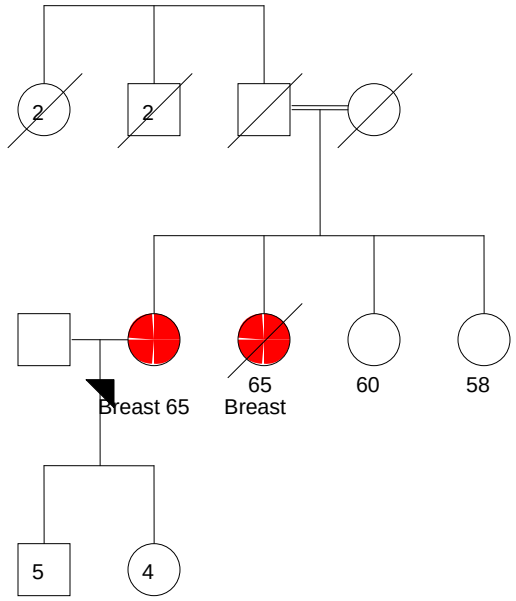
Finally, in the fourth chapter, I examined large runs of homozygosity in consanguineous individuals with a family history of breast cancer for autosomal recessive loci associated with disease. I identified large stretches of homozygous SNPs using a genome-wide genotyping approach in Arab and Jewish women with breast cancer and no mutations in *BRCA1* and *BRCA2*. The whole genome association analysis toolset, PLINK v1.05, was used to screen for runs of homozygous genotypes in all cases and unaffected family members. Runs of homozygosity with ≥1 Mb of consecutive homozygous SNPs defined minimal overlapping ROH regions. Six individuals with breast cancer had 242 kb overlapping run of homozygous SNPs on chromosome 9q33.2-33.3 which harbors a potentially important candidate gene in cancer, *LHX2*. All coding region (+/- 200 bp) of the gene *LHX2* were sequenced in the six individuals with the overlapping

run of homozygosity. *LHX2* is a putative transcription factor containing two cystein-rich (LIM) motifs and a homeobox (HOX) DNA-binding domain. Previous analysis of the sequence including the coding regions revealed a CpG-rich region, implicating a role for methylation in this gene.

APPENDIX:

40094

Muslims



40597

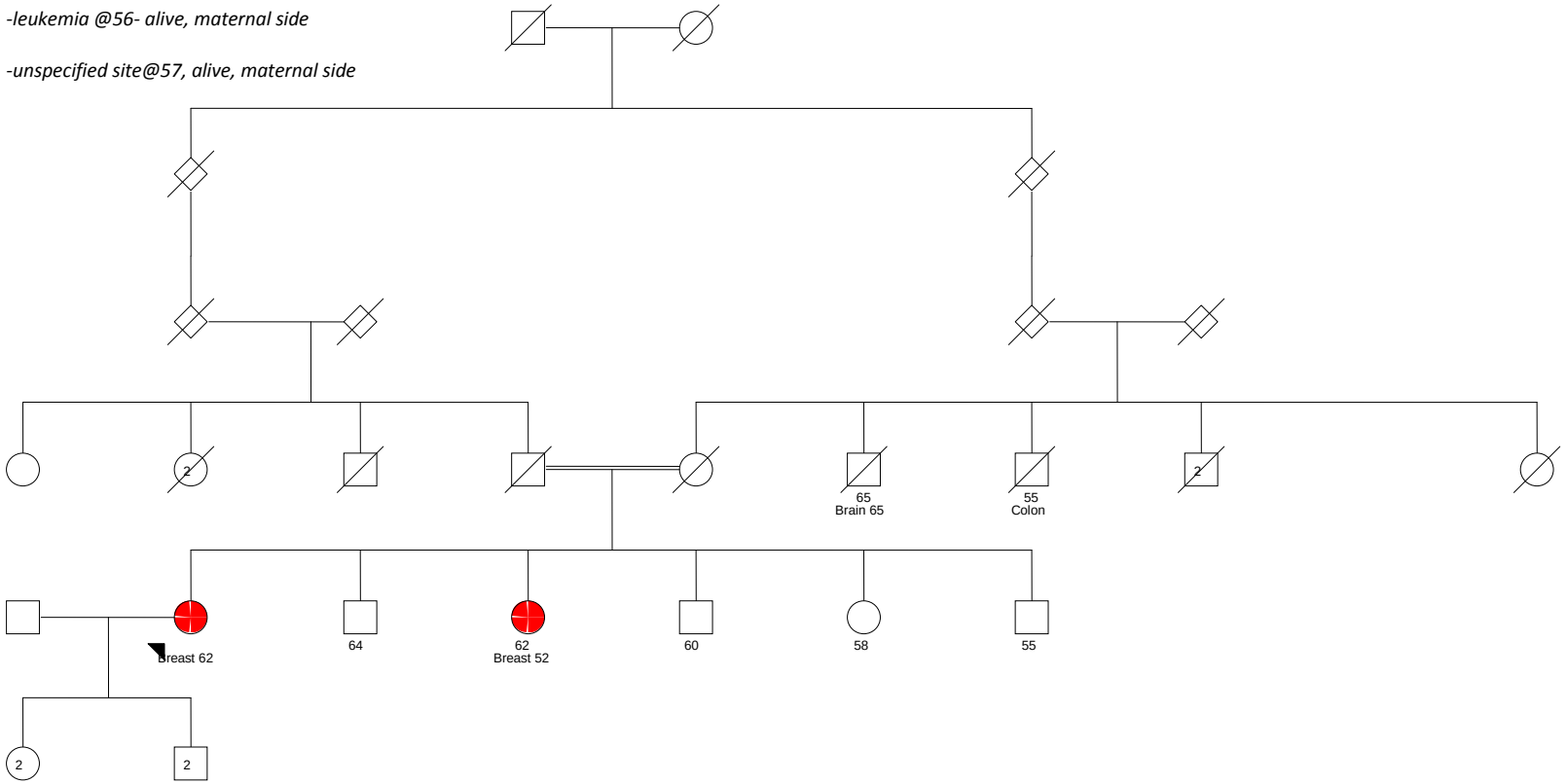
Christians

3 other 2nd degree family members had cancer

-uterus @ 45-deceased maternal side

-leukemia @56- alive, maternal side

-unspecified site@57, alive, maternal side



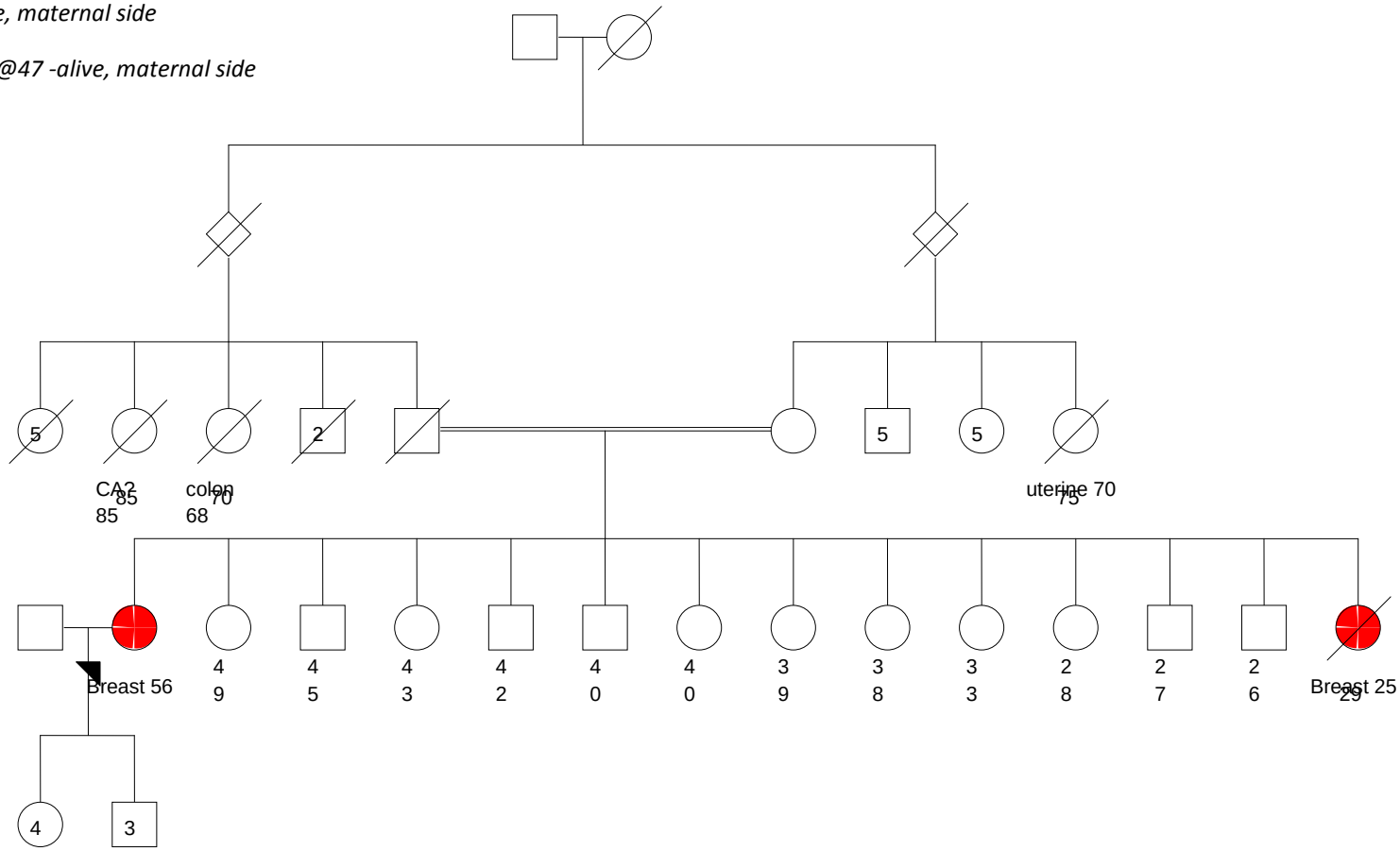
40609

Muslim

-one of the proband's sisters (33) is marked as a twin- not clear with whom- (deceased sister??)
-two 2nd degree relatives had cancer

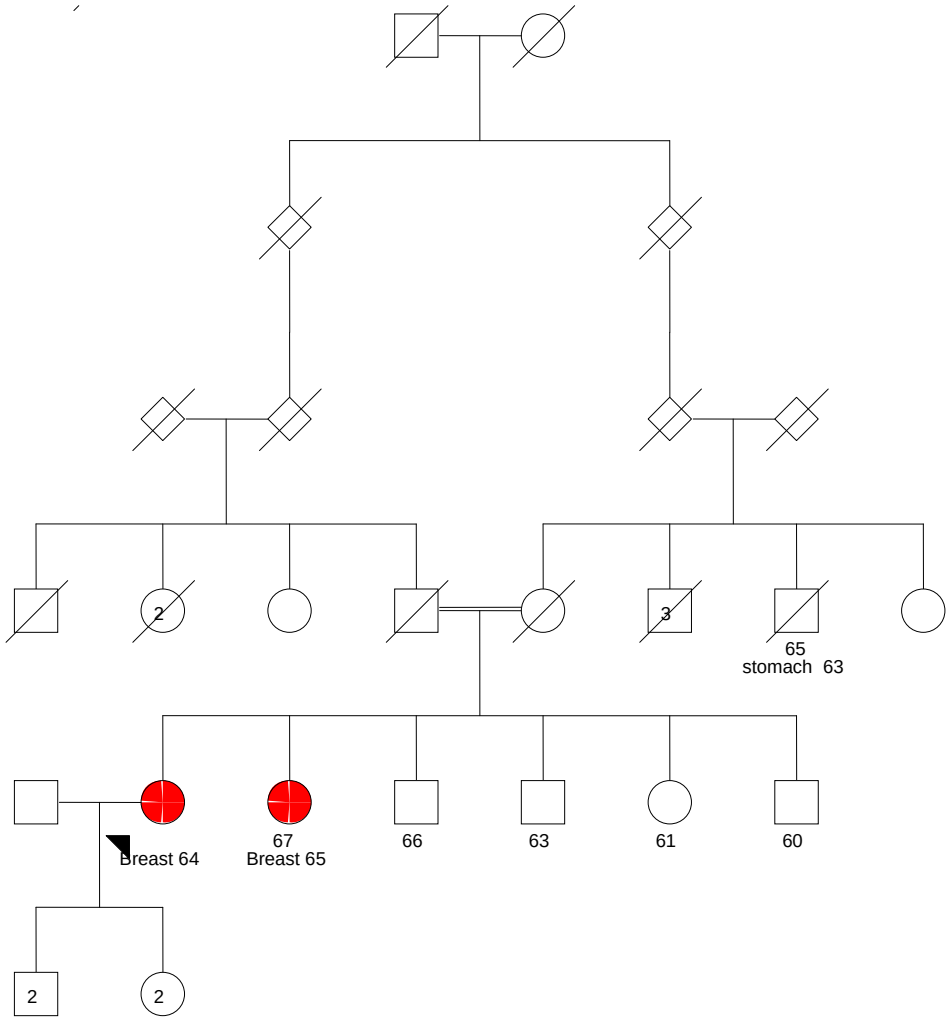
leukemia- alive, maternal side

Breast cancer @47 -alive, maternal side



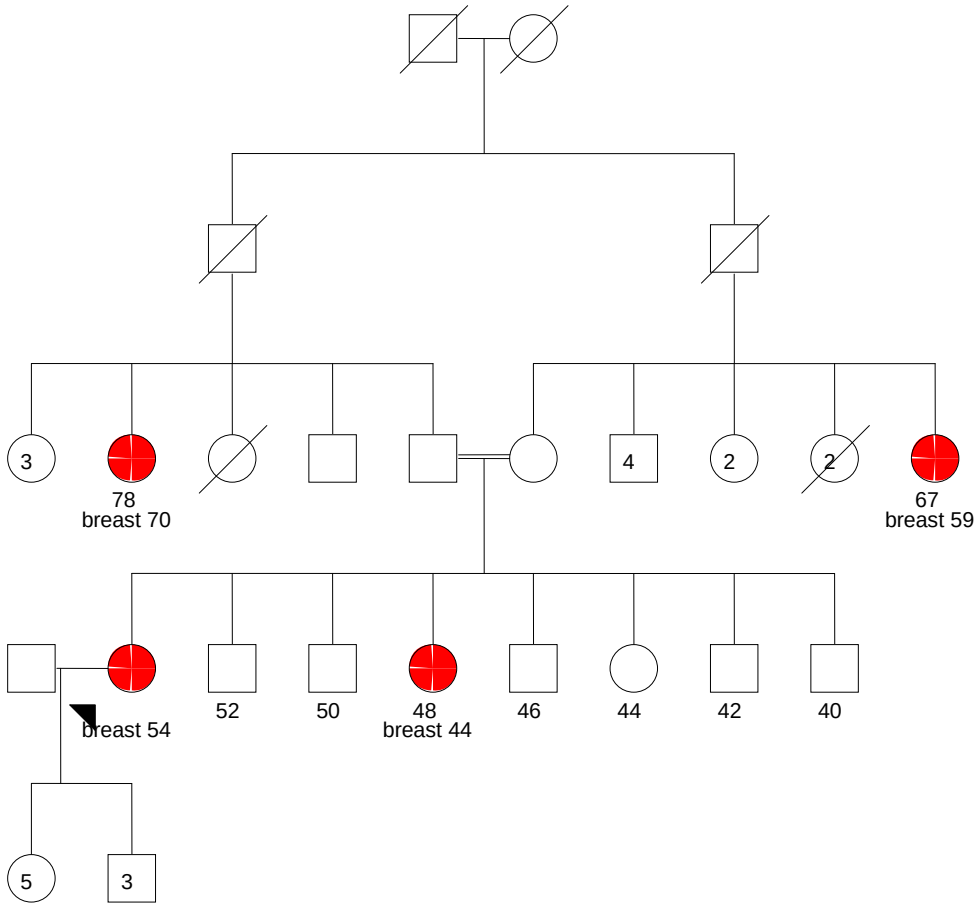
41128

2nd degree relative had uterus at 46-deceased, maternal side (Christians)



41622

(Muslims)

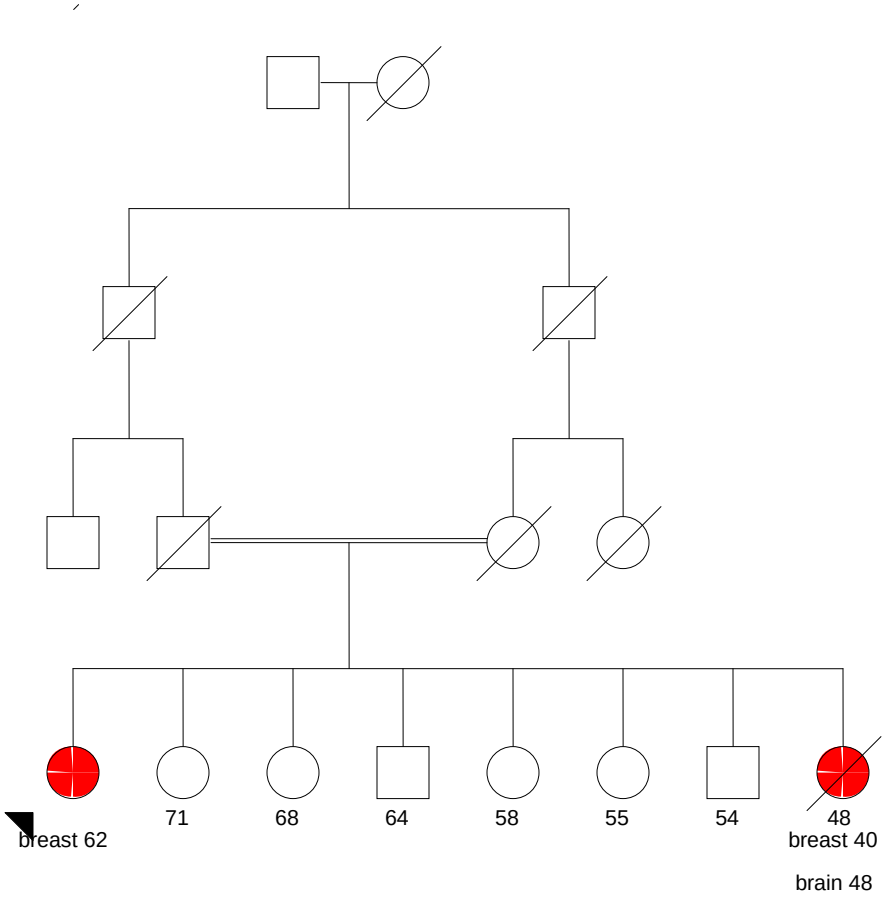


41677

Christians

-2nd degree family member had brain cancer @44-deceased, paternal side

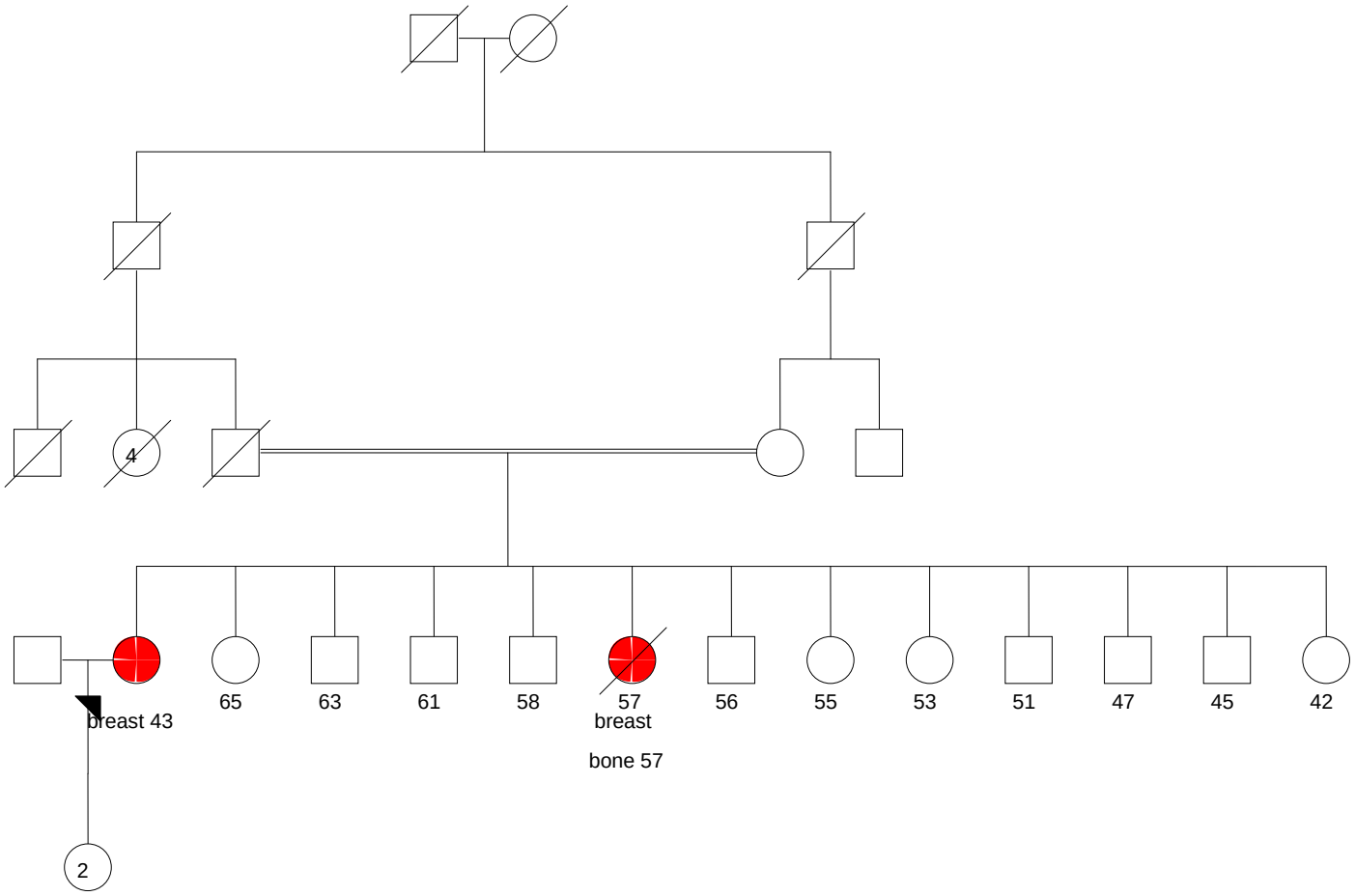
-two of the proband's sisters are marked as having unknown twin status (9)!!



41983

Muslims

two 2nd degree family members had cancer 999, one maternal and on paternal side- both are deceased



42424

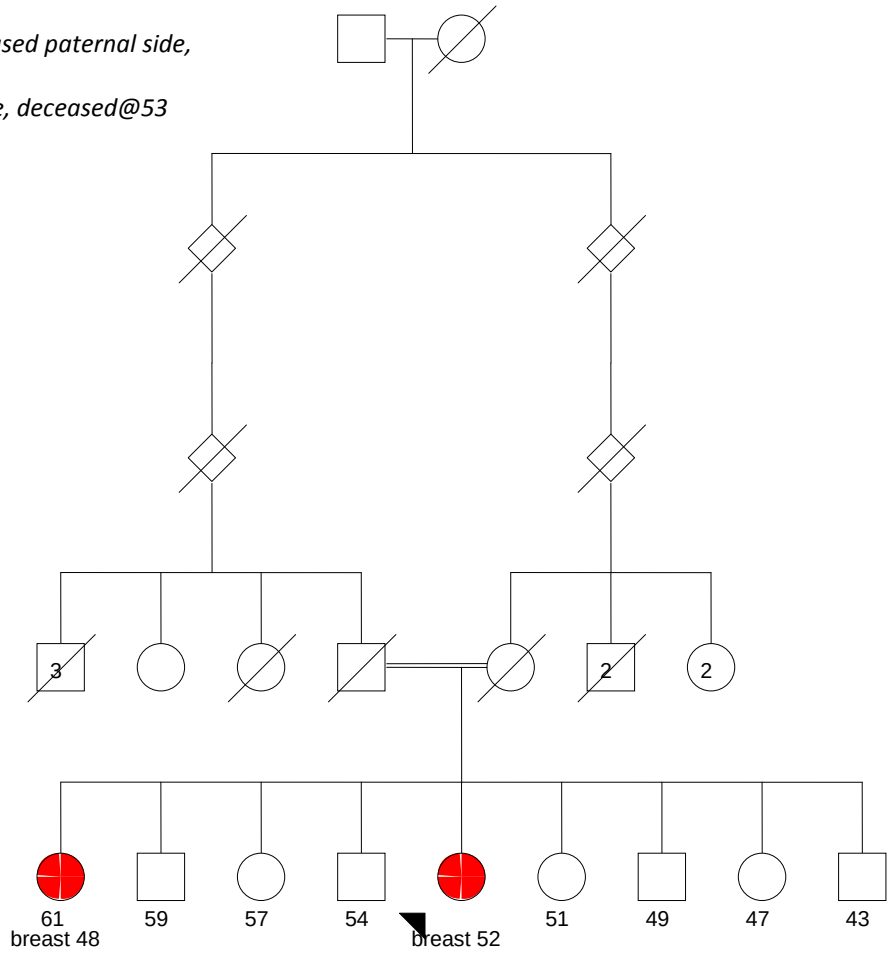
Christians

3 other 2nd degree family members had cancer:

-stomach @52- alive, maternal side,

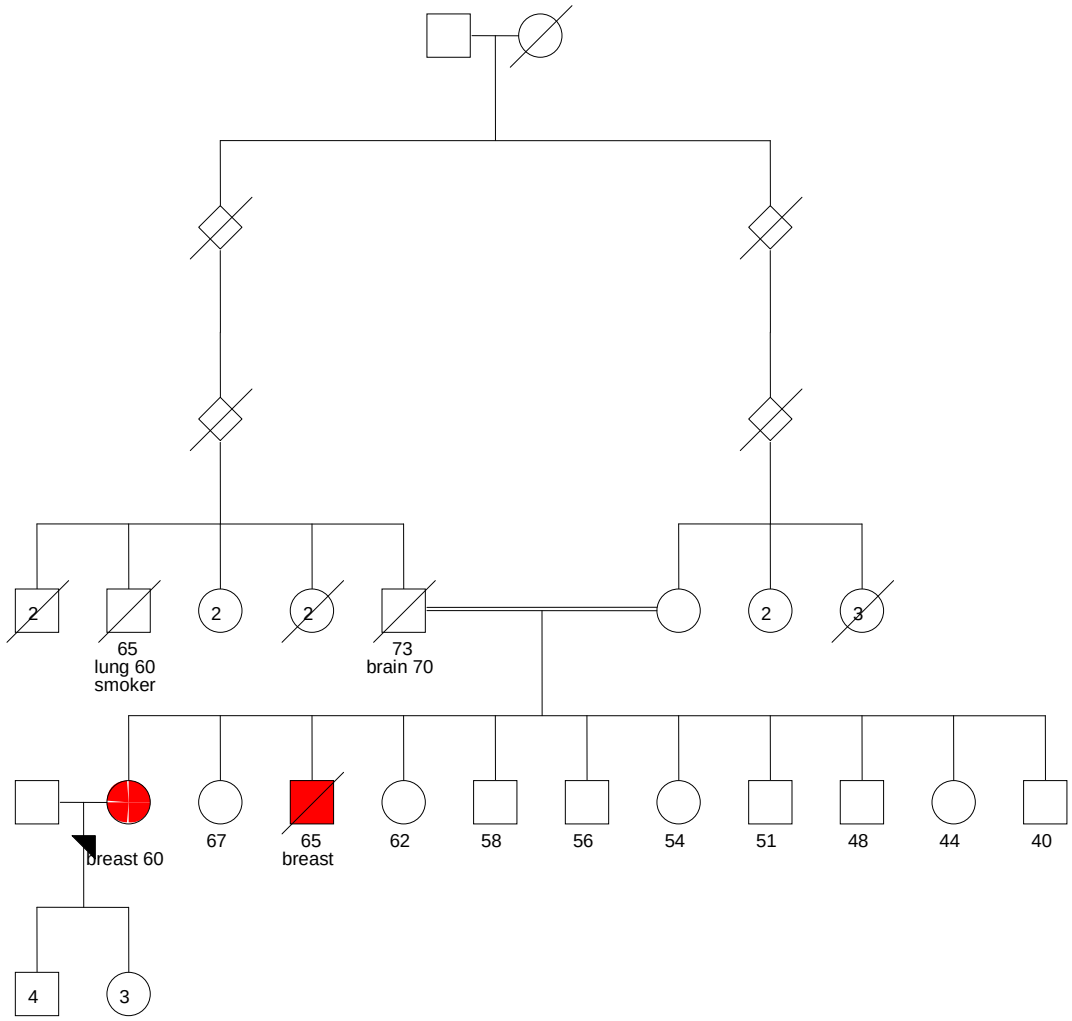
-stomach@61-deceased paternal side,

- colon, paternal side, deceased@53



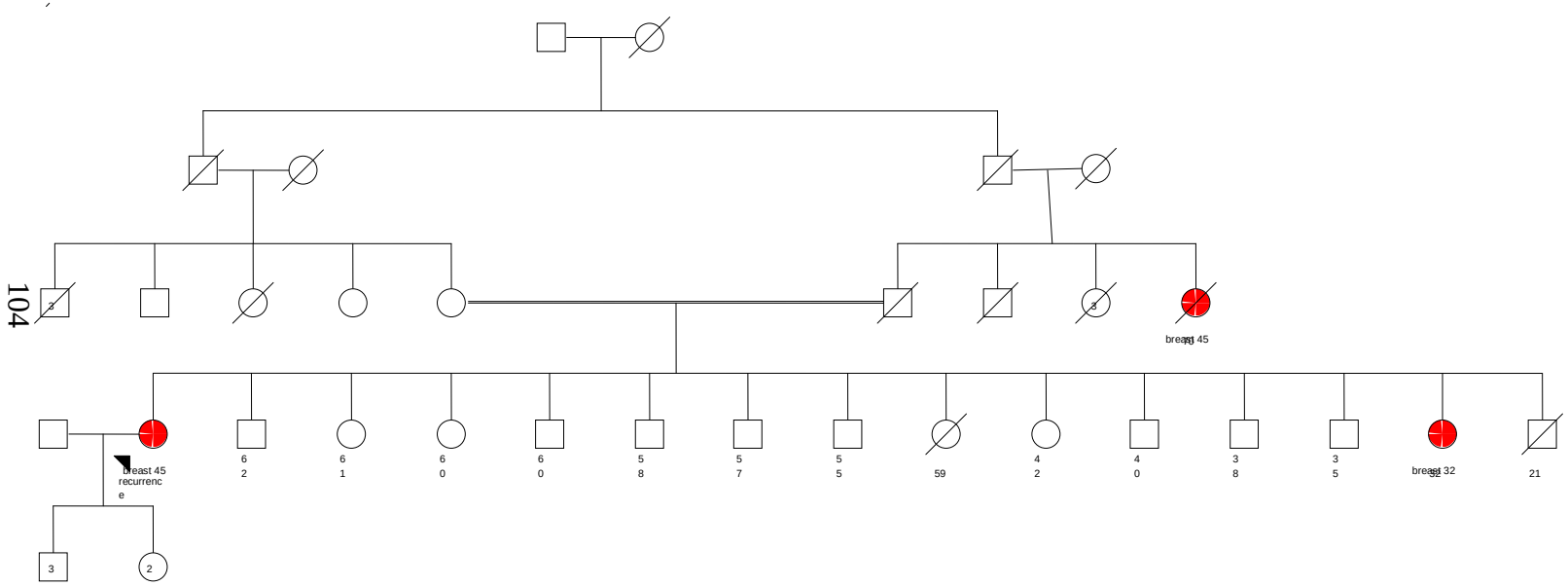
42527

Muslims



42791

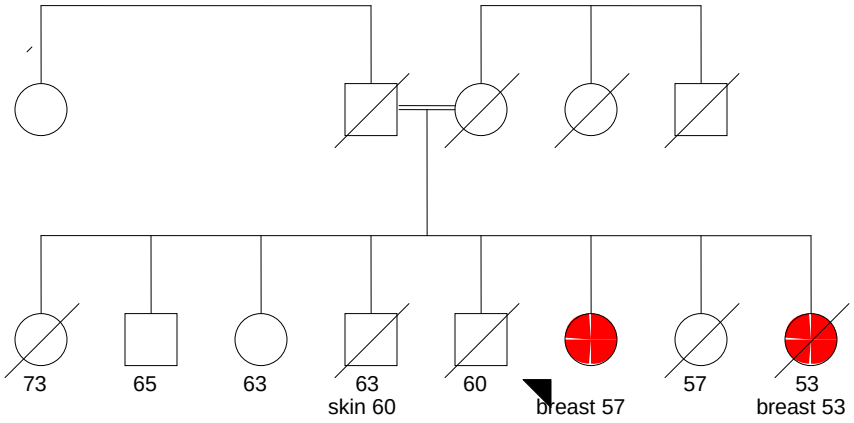
Muslims



43749

Christians

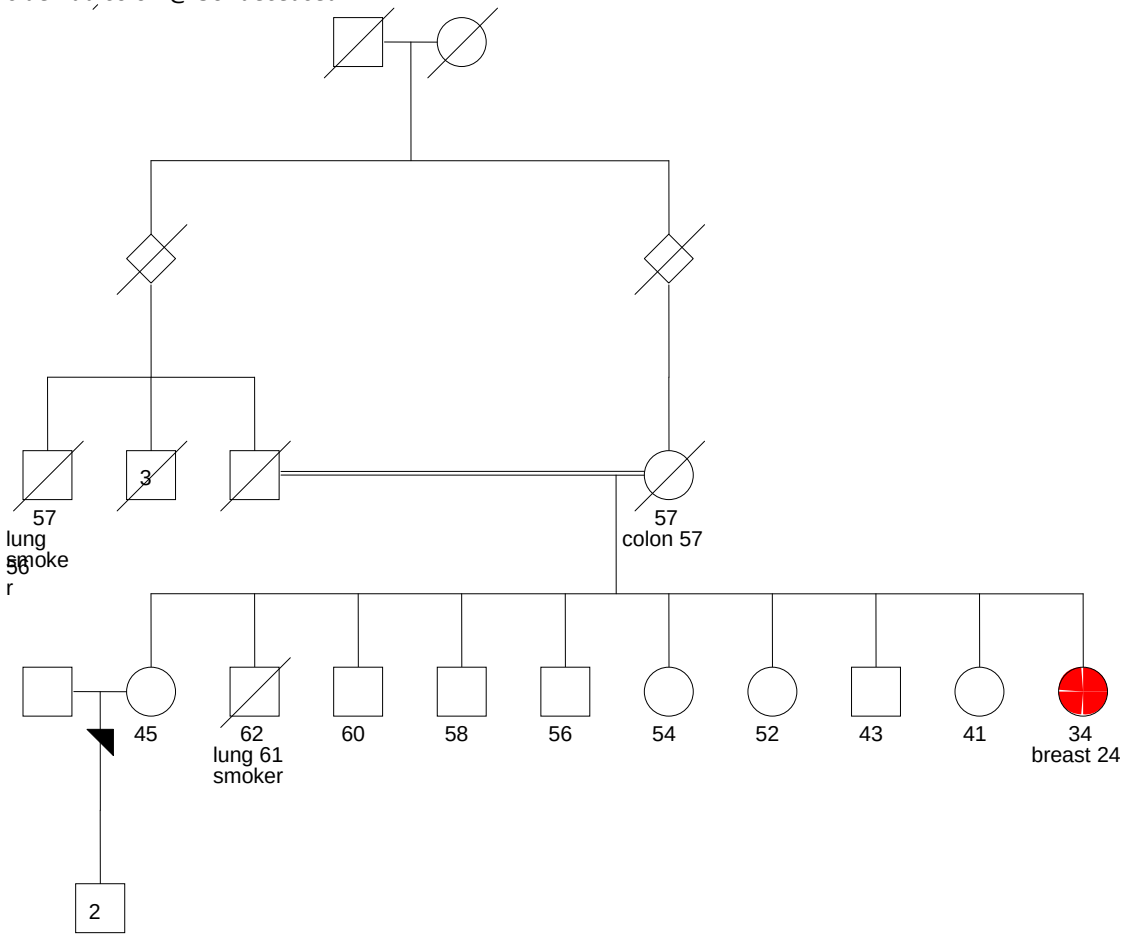
2nd degree family member had colon cancer @ 33-deceased



50436

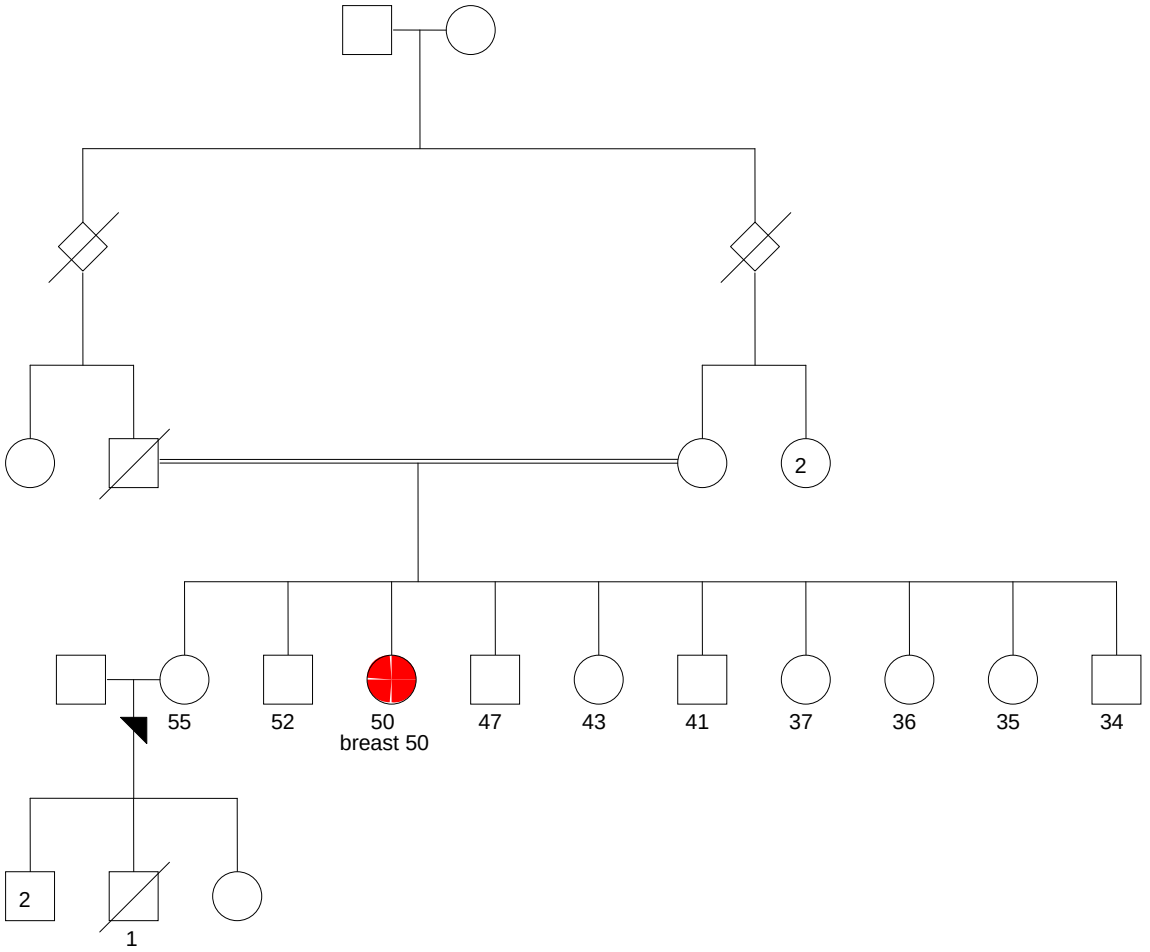
Christians

2nd degree family member- paternal side had colon @ 30- deceased



50718

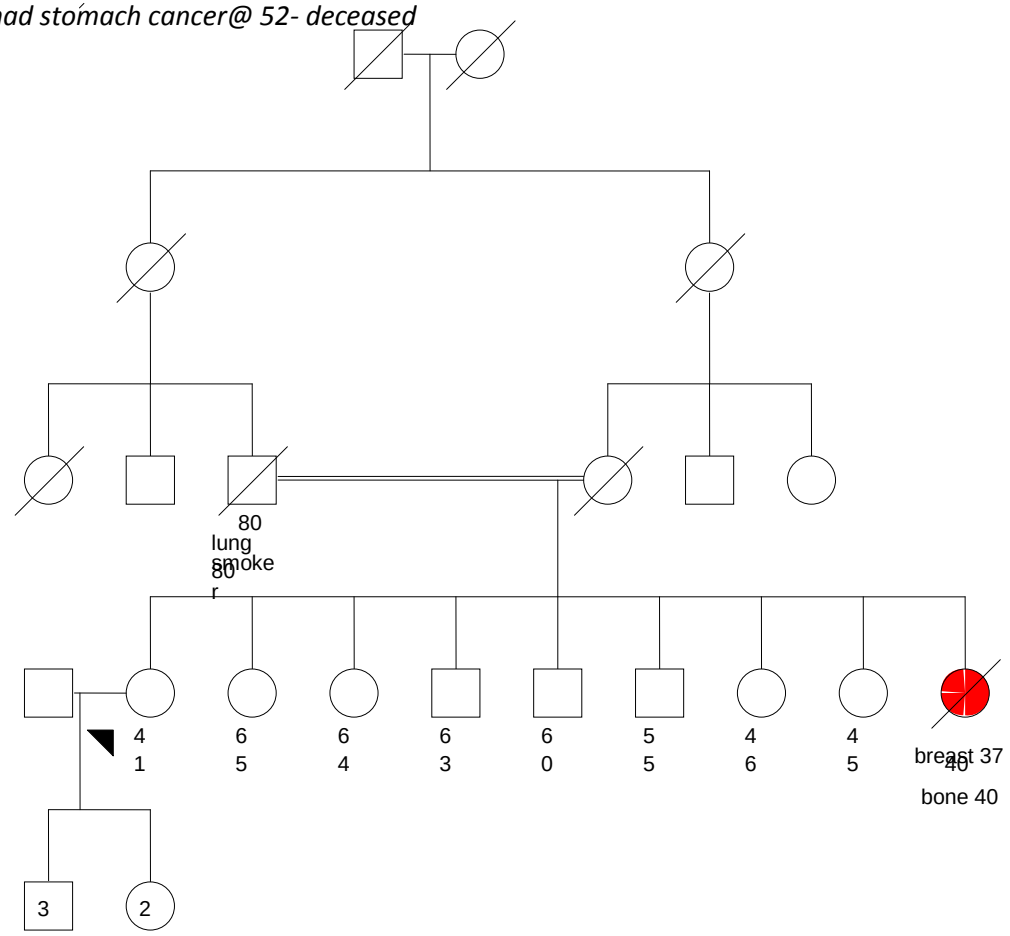
Muslims



51301

Muslims

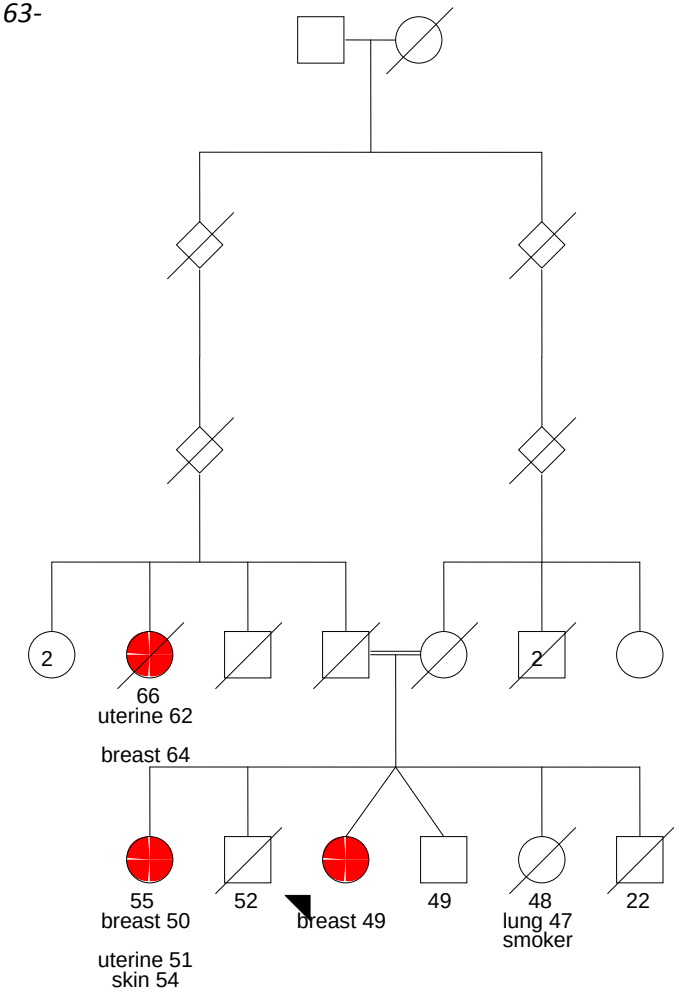
2nd degree relative, maternal side, had stomach cancer@ 52- deceased



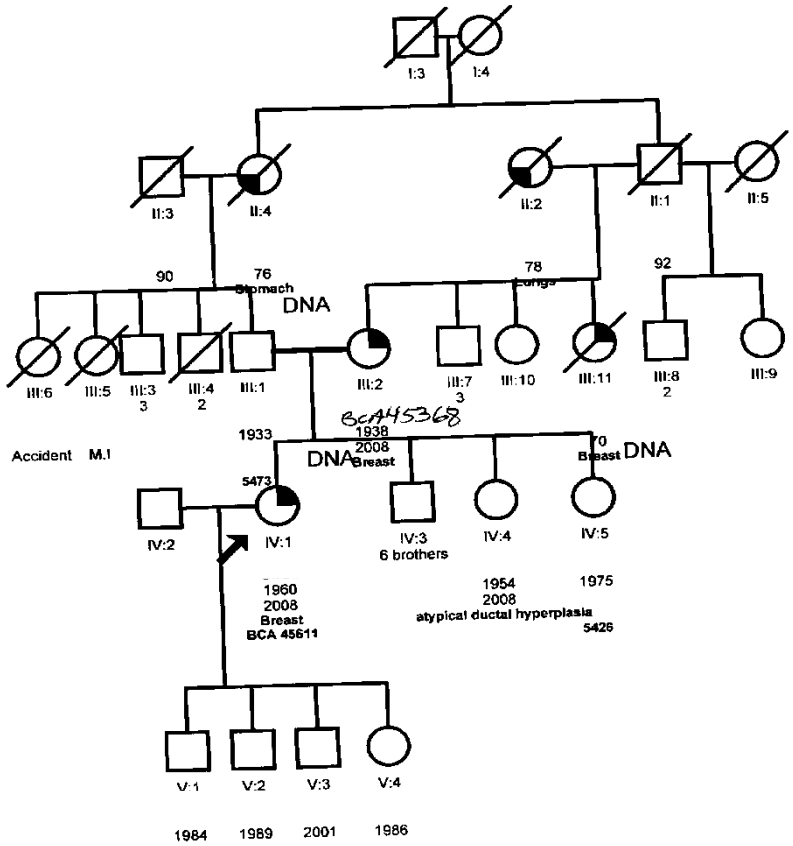
90049

Christians

2nd degree family member had brain cancer @ 63- deceased, paternal side

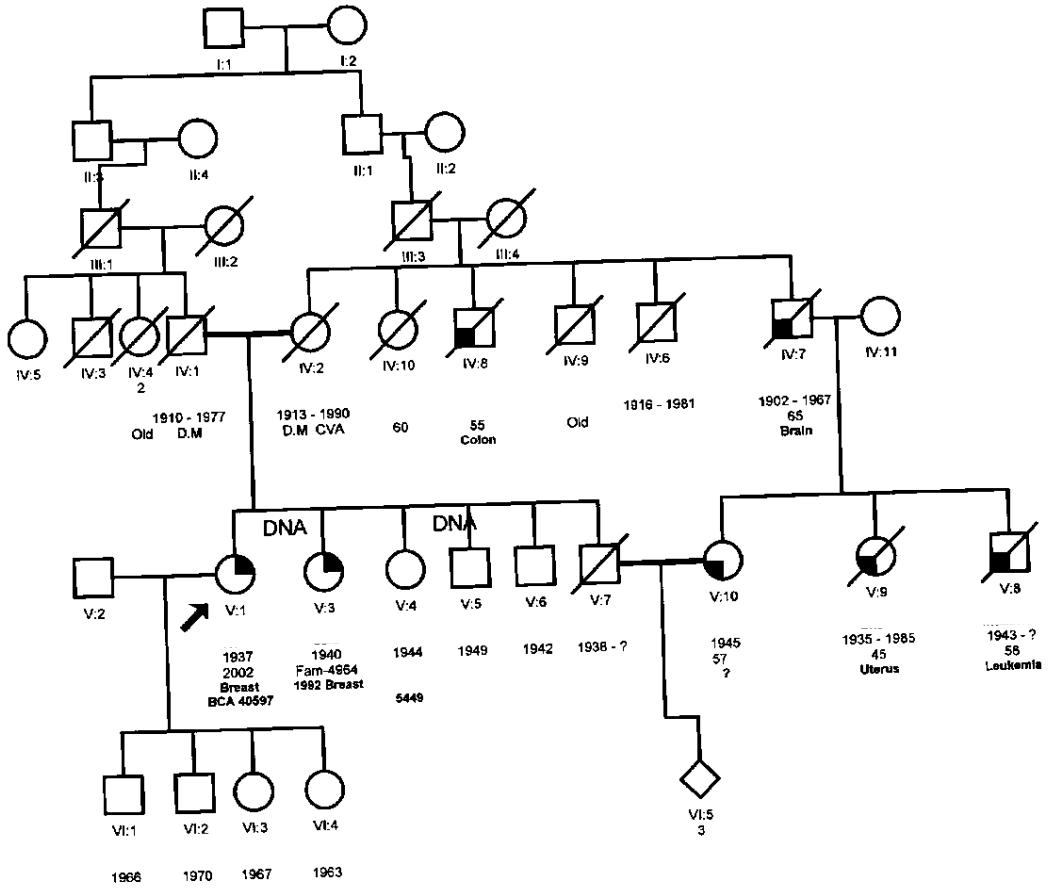


Number : 4958
Creation date : 15/6/2009
Arab-Muslem BCA 45611



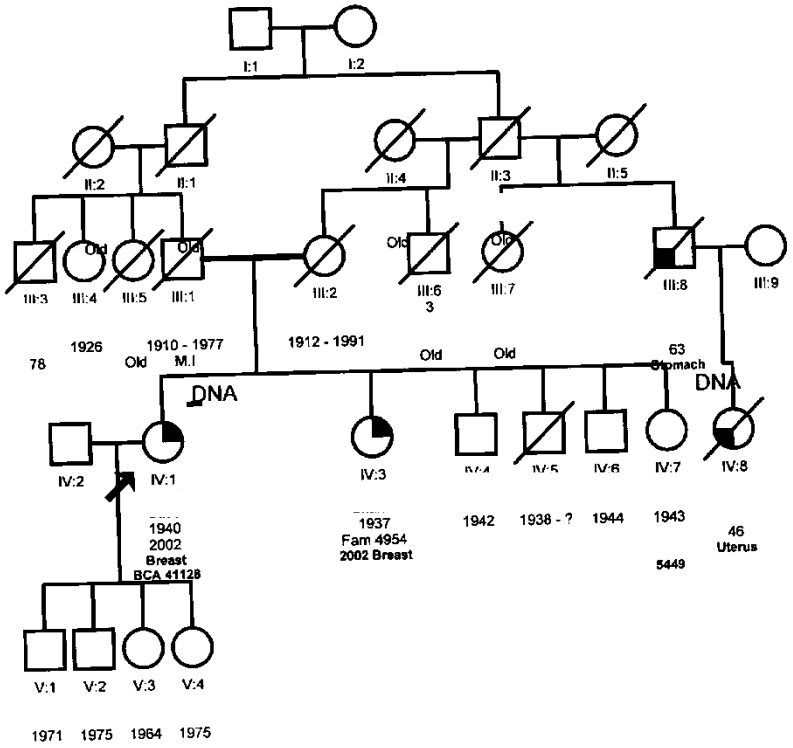
Printed : 29-10-2009

Number : 4954 +4964
Creation date : 15/6/2009
Arab-Christian BCA 40597



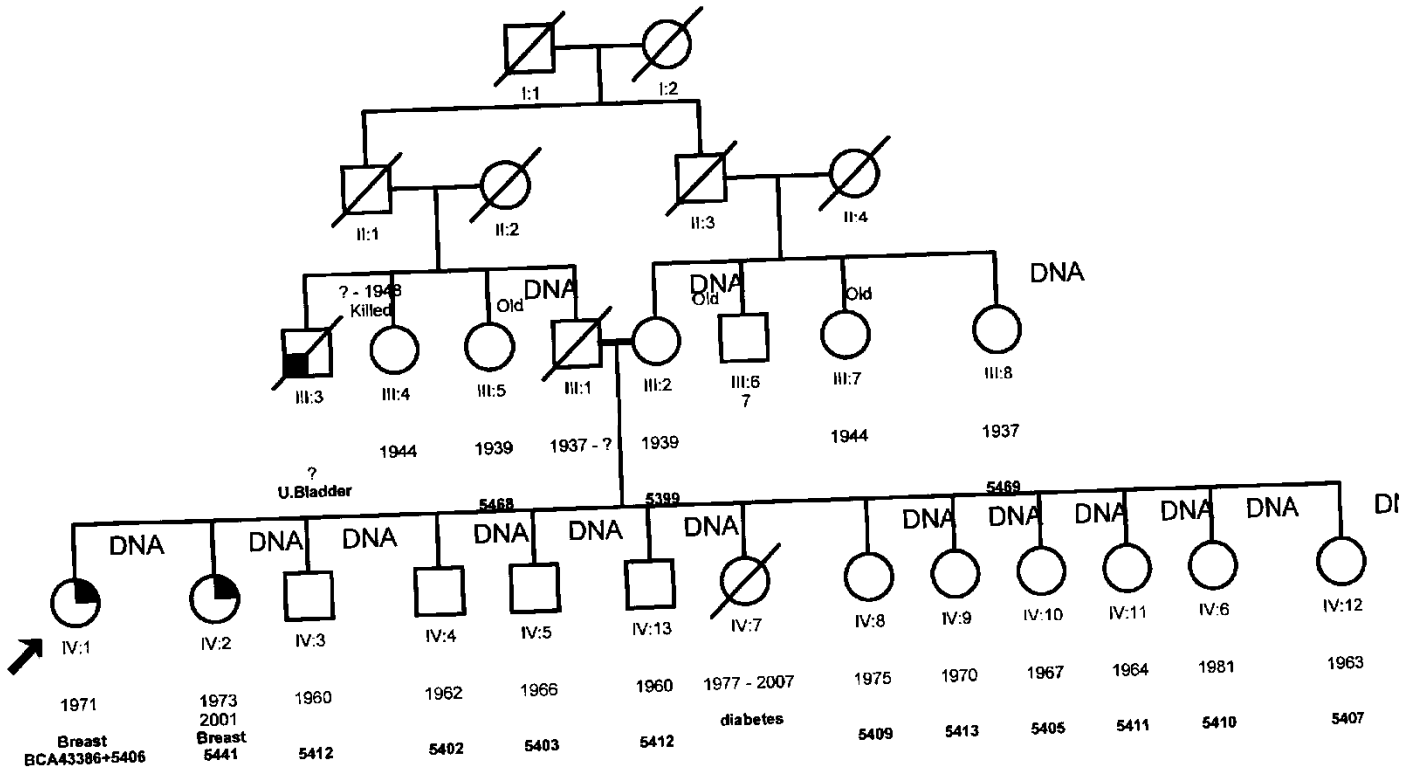
Printed : 29-10-2009

Number : 4964 +4954
Creation date : 15/6/2009
Arab-Christian BCA 41128



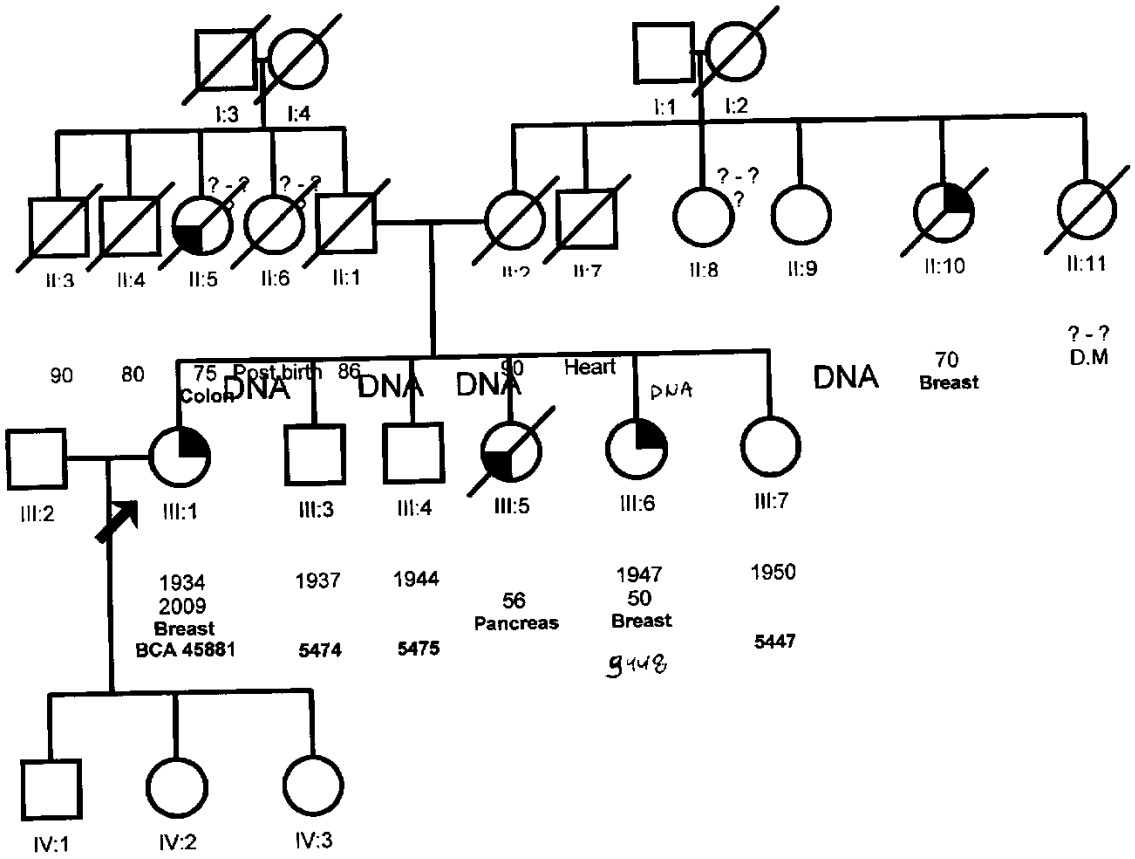
Printed : 29-10-2009

Number : 4953
Creation date : 15/6/2009
Arab-Muslem BCA 43386



Number : 4933
Creation date : 5/6/2009
Arab Christian BCA 45881

Printed : 29-10-2009



REFERENCES

(2005). A haplotype map of the human genome. *Nature* **437:** 1299-1320.

Abecasis GR, Ghosh D, Nichols TE (2005). Linkage disequilibrium: ancient history drives the new genetics. *Hum Hered* **59:** 118-124.

Abeliovich D, Kaduri L, Lerer I, Weinberg N, Amir G, Sagi M *et al* (1997). The founder mutations 185delAG and 5382insC in BRCA1 and 6174delT in BRCA2 appear in 60% of ovarian cancer and 30% of early-onset breast cancer patients among Ashkenazi women. *Am J Hum Genet* **60:** 505-514.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P *et al* (2010). A method and server for predicting damaging missense mutations. *Nat Methods* **7:** 248-249.

Al-Tassan N, Chmiel NH, Maynard J, Fleming N, Livingston AL, Williams GT *et al* (2002). Inherited variants of MYH associated with somatic G:C-->T:A mutations in colorectal tumors. *Nat Genet* **30:** 227-232.

Alazzouzi H, Alhopuro P, Salovaara R, Sammalkorpi H, Jarvinen H, Mecklin JP *et al* (2005). SMAD4 as a prognostic marker in colorectal cancer. *Clin Cancer Res* **11:** 2606-2611.

Andre T, Boni C, Navarro M, Tabernero J, Hickish T, Topham C *et al* (2009). Improved overall survival with oxaliplatin, fluorouracil, and leucovorin as adjuvant treatment in stage II or III colon cancer in the MOSAIC trial. *J Clin Oncol* **27:** 3109-3116.

Ansorge WJ (2009). Next-generation DNA sequencing techniques. *N Biotechnol* **25:** 195-203.

Antonarakis SE, McKusick VA (2000). OMIM passes the 1,000-disease-gene mark. *Nat Genet* **25:** 11.

Antoniou AC, Pharoah PD, McMullan G, Day NE, Ponder BA, Eas D (2001). Evidence for further breast cancer susceptibility genes in addition to BRCA1 and CA2 in a population-based study. *Genet Epidemiol* **21:** 1-18.

Assie G, LaFramboise T, Platzer P, Eng C (2008). Frequency of germline genomic homozygosity associated with cancer cases. *JAMA* **299:** 1437-1445.

Bacolod MD, Schemmann GS, Giardina SF, Paty P, Notterman DA, Barany F (2009). Emerging paradigms in cancer genetics: some important findings from high-density single nucleotide polymorphism array studies. *Cancer Res* **69:** 723-727.

Barber TD, McManus K, Yuen KW, Reis M, Parmigiani G, Shen D *et al* (2008). Chromatid cohesion defects may underlie chromosome instability in human colorectal cancers. *Proc Natl Acad Sci U S A* **105:** 3443-3448.

Barratt PL, Seymour MT, Stenning SP, Georgiades I, Walker C, Birbeck K *et al* (2002). DNA markers predicting benefit from adjuvant fluorouracil in patients with colon cancer: a molecular study. *Lancet* **360:** 1381-1391.

Baudis M (2007). Genomic imbalances in 5918 malignant epithelial tumors: an explorative meta-analysis of chromosomal CGH data. *BMC Cancer* **7:** 226.

Baylin SB, Makos M, Wu JJ, Yen RW, de Bustros A, Vertino P *et al* (1991). Abnormal patterns of DNA methylation in human neoplasia: potential consequences for tumor progression. *Cancer Cells* **3:** 383-390.

Beggs AD, Latchford AR, Vasen HF, Moslein G, Alonso A, Aretz S *et al* (2010). Peutz-Jeghers syndrome: a systematic review and recommendations for management. *Gut* **59:** 975-986.

Benchabane H, Ahmed Y (2009). The adenomatous polyposis coli tumor suppressor and Wnt signaling in the regulation of apoptosis. *Adv Exp Med Biol* **656:** 75-84.

Bener A, Denic S, Al-Mazrouei M (2001). Consanguinity and family history of cancer in children with leukemia and lymphomas. *Cancer* **92:** 1-6.

Bener A, Alali KA (2006). Consanguineous marriage in a newly developed country: the Qatari population. *J Biosoc Sci* **38:** 239-246.

Berger MF, Levin JZ, Vijayendran K, Sivachenko A, Adiconis X, Maguire J *et al* (2010). Integrative analysis of the melanoma transcriptome. *Genome Res* **20:** 413-427.

Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J *et al* (2010). The landscape of somatic copy-number alteration across human cancers. *Nature* **463:** 899-905.

Bisgaard ML, Jager AC, Dalgaard P, Sondergaard JO, Rehfeld JF, Nielsen FC (2001). Allelic loss of chromosome 2p21-16.3 is associated with reduced survival in sporadic colorectal cancer. *Scand J Gastroenterol* **36:** 405-409.

Bodmer W, Bonilla C (2008). Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* **40:** 695-701.

Boehm JS, Zhao JJ, Yao J, Kim SY, Firestein R, Dunn IF *et al* (2007). Integrative genomic approaches identify IKBKE as a breast cancer oncogene. *Cell* **129:** 1065-1079.

Broderick P, Carvajal-Carmona L, Pittman AM, Webb E, Howarth K, Rowan A *et al* (2007). A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. *Nat Genet* **39:** 1315-1317.

Brosens LA, van Hattem A, Hylind LM, Iacobuzio-Donahue C, Romans KE, Axilbund J *et al* (2007). Risk of colorectal cancer in juvenile polyposis. *Gut* **56:** 965-967.

Cahill DP, Lengauer C, Yu J, Riggins GJ, Willson JK, Markowitz SD *et al* (1998). Mutations of mitotic checkpoint genes in human cancers. *Nature* **392:** 300-303.

Carethers JM, Hawn MT, Greenson JK, Hitchcock CL, Boland CR (1998). Prognostic significance of allelic lost at chromosome 18q21 for stage II colorectal cancer. *Gastroenterology* **114:** 1188-1195.

Carstensen B, Soll-Johanning H, Villadsen E, Sondergaard JO, Lynge E (1996). Familial aggregation of colorectal cancer in the general population. *Int J Cancer* **68:** 428-435.

Chang SC, Lin JK, Lin TC, Liang WY (2005). Loss of heterozygosity: an independent prognostic factor of colorectal cancer. *World J Gastroenterol* **11:** 778-784.

Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, Sougnez C, Schinzel AC *et al* (2011). Initial genome sequencing and analysis of multiple myeloma. *Nature* **471:** 467-472.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS *et al* (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* **6:** 677-681.

Choi SW, Lee KJ, Bae YA, Min KO, Kwon MS, Kim KM *et al* (2002). Genetic classification of colorectal cancer based on chromosomal loss and microsatellite instability predicts survival. *Clin Cancer Res* **8:** 2311-2322.

Clark AG (1999). The size distribution of homozygous segments in the human genome. *Am J Hum Genet* **65:** 1489-1492.

Claus EB, Risch N, Thompson WD (1991). Genetic analysis of breast cancer in the cancer and steroid hormone study. *Am J Hum Genet* **48:** 232-242.

Couch FJ, Weber BL (1996). Mutations and polymorphisms in the familial early-onset breast cancer (BRCA1) gene. Breast Cancer Information Core. *Hum Mutat* **8:** 8-18.

Cui K, Zang C, Roh TY, Schones DE, Childs RW, Peng W *et al* (2009). Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. *Cell Stem Cell* **4:** 80-93.

Dahm CC, Keogh RH, Spencer EA, Greenwood DC, Key TJ, Fentiman IS *et al* (2010). Dietary fiber and colorectal cancer risk: a nested case-control study using food diaries. *J Natl Cancer Inst* **102:** 614-626.

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001). High-resolution haplotype structure in the human genome. *Nat Genet* **29:** 229-232.

de Jong MM, Nolte IM, te Meerman GJ, van der Graaf WT, de Vries EG, Sijmons RH *et al* (2002). Low-penetrance genes and their involvement in colorectal cancer susceptibility. *Cancer Epidemiol Biomarkers Prev* **11:** 1332-1352.

Denic S, Bener A (2001). Consanguinity decreases risk of breast cancer--cervical cancer unaffected. *Br J Cancer* **85:** 1675-1679.

Denic S, Bener A, Sabri S, Khatib F, Milenkovic J (2005). Parental consanguinity and risk of breast cancer: a population-based case-control study. *Med Sci Monit* **11:** CR415-419.

Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB (2010). Rare variants create synthetic genome-wide associations. *PLoS Biol* **8:** e1000294.

Diep CB, Thorstensen L, Meling GI, Skovlund E, Rognum TO, Lothe RA (2003). Genetic tumor markers with prognostic impact in Dukes' stages B and C colorectal cancer patients. *J Clin Oncol* **21:** 820-829.

Easton DF, Bishop DT, Ford D, Crockford GP (1993). Genetic linkage analysis in familial breast and ovarian cancer: results from 214 families. The Breast Cancer Linkage Consortium. *Am J Hum Genet* **52:** 678-701.

El Saghir NS, Seoud M, Khalil MK, Charafeddine M, Salem ZK, Geara FB *et al* (2006). Effects of young age at presentation on survival in breast cancer. *BMC Cancer* **6:** 194.

Ellis NA, Groden J, Ye TZ, Straughen J, Lennon DJ, Ciocci S *et al* (1995). The Bloom's syndrome gene product is homologous to RecQ helicases. *Cell* **83:** 655-666.

Fearon ER, Cho KR, Nigro JM, Kern SE, Simons JW, Ruppert JM *et al* (1990). Identification of a chromosome 18q gene that is altered in colorectal cancers. *Science* **247:** 49-56.

Fearon ER, Vogelstein B (1990). A genetic model for colorectal tumorigenesis. *Cell* **61:** 759-767.

Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM (2010). Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int J Cancer* **127:** 2893-2917.

Font A, Abad A, Monzo M, Sanchez JJ, Guillot M, Manzano JL *et al* (2001). Prognostic value of K-ras mutations and allelic imbalance on chromosome 18q in patients with resected colorectal cancer. *Dis Colon Rectum* **44:** 549-557.

Ford D, Easton DF, Stratton M, Narod S, Goldgar D, Devilee P *et al* (1998). Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium. *Am J Hum Genet* **62:** 676-689.

Fukasawa K (2005). Centrosome amplification, chromosome instability and cancer development. *Cancer Lett* **230:** 6-19.

Fukasawa K (2011). Aberrant activation of cell cycle regulators, centrosome amplification, and mitotic defects. *Horm Cancer* **2:** 104-112.

Futreal PA, Liu Q, Shattuck-Eidens D, Cochran C, Harshman K, Tavtigian S *et al* (1994). BRCA1 mutations in primary breast and ovarian carcinomas. *Science* **266:** 120-122.

Garber JE, Offit K (2005). Hereditary cancer predisposition syndromes. *J Clin Oncol* **23:** 276-292.

Garner C (2011). Confounded by sequencing depth in association studies of rare alleles. *Genet Epidemiol*.

Gibson J, Morton NE, Collins A (2006). Extended tracts of homozygosity in outbred human populations. *Hum Mol Genet* **15:** 789-795.

Gonzalez CA, Riboli E (2010). Diet and cancer prevention: Contributions from the European Prospective Investigation into Cancer and Nutrition (EPIC) study. *Eur J Cancer* **46:** 2555-2562.

Goss KH, Groden J (2000). Biology of the adenomatous polyposis coli tumor suppressor. *J Clin Oncol* **18:** 1967-1979.

Gruber SB, Moreno V, Rozek LS, Rennerts HS, Lejbkowicz F, Bonner JD *et al* (2007). Genetic variation in 8q24 associated with risk of colorectal cancer. *Cancer Biol Ther* **6:** 1143-1147.

Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B *et al* (1990). Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* **250:** 1684-1689.

Halling KC, French AJ, McDonnell SK, Burgart LJ, Schaid DJ, Peterson BJ *et al* (1999). Microsatellite instability and 8p allelic imbalance in stage B2 and C colorectal cancers. *J Natl Cancer Inst* **91:** 1295-1303.

Hariharan R (2003). The analysis of microarray data. *Pharmacogenomics* **4:** 477-497.

Hayat MJ, Howlader N, Reichman ME, Edwards BK (2007). Cancer statistics, trends, and multiple primary cancer analyses from the Surveillance, Epidemiology, and End Results (SEER) Program. *Oncologist* **12:** 20-37.

Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF *et al* (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459:** 108-112.

Hildebrand JS, Jacobs EJ, Campbell PT, McCullough ML, Teras LR, Thun MJ *et al* (2009). Colorectal cancer incidence and postmenopausal hormone use by type, recency, and duration in cancer prevention study II. *Cancer Epidemiol Biomarkers Prev* **18:** 2835-2841.

Hiraoka S, Kato J, Fujiki S, Kaji E, Morikawa T, Murakami T *et al* (2010). The presence of large serrated polyps increases risk for colorectal cancer. *Gastroenterology* **139:** 1503-1510, 1510 e1501-1503.

Hoffmeister M, Schmitz S, Karmrodt E, Stegmaier C, Haug U, Arndt V *et al* (2010). Male sex and smoking have a larger impact on the prevalence of colorectal neoplasia than family history of colorectal cancer. *Clin Gastroenterol Hepatol* **8:** 870-876.

Houlston RS, Webb E, Broderick P, Pittman AM, Di Bernardo MC, Lubbe S *et al* (2008). Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat Genet* **40:** 1426-1435.

Howe JR, Roth S, Ringold JC, Summers RW, Jarvinen HJ, Sistonen P *et al* (1998). Mutations in the SMAD4/DPC4 gene in juvenile polyposis. *Science* **280:** 1086-1088.

Howlader N, Ries LA, Mariotto AB, Reichman ME, Ruhl J, Cronin KA (2010). Improved estimates of cancer-specific survival rates from population-based data. *J Natl Cancer Inst* **102:** 1584-1598.

Issa JP, Ottaviano YL, Celano P, Hamilton SR, Davidson NE, Baylin SB (1994). Methylation of the oestrogen receptor CpG island links ageing and neoplasia in human colon. *Nat Genet* **7:** 536-540.

Jaeger EE, Woodford-Richens KL, Lockett M, Rowan AJ, Sawyer EJ, Heinimann K *et al* (2003). An ancestral Ashkenazi haplotype at the HMPS/CRAC1 locus on 15q13-q14 is associated with hereditary mixed polyposis syndrome. *Am J Hum Genet* **72:** 1261-1267.

Jass JR, Biden KG, Cummings MC, Simms LA, Walsh M, Schoch E *et al* (1999). Characterisation of a subtype of colorectal cancer combining features of the suppressor and mild mutator pathways. *J Clin Pathol* **52:** 455-460.

Jass JR (2000). Familial colorectal cancer: pathology and molecular characteristics. *Lancet Oncol* **1:** 220-226.

Jass JR (2007). Classification of colorectal cancer based on correlation of clinical, morphological and molecular features. *Histopathology* **50:** 113-130.

Jen J, Kim H, Piantadosi S, Liu ZF, Levitt RC, Sistonen P *et al* (1994). Allelic loss of chromosome 18q and prognosis in colorectal cancer. *N Engl J Med* **331:** 213-221.

Jenne DE, Reimann H, Nezu J, Friedel W, Loff S, Jeschke R *et al* (1998). Peutz-Jeghers syndrome is caused by mutations in a novel serine threonine kinase. *Nat Genet* **18:** 38-43.

Jernvall P, Makinen MJ, Karttunen TJ, Makela J, Vihko P (1999). Loss of heterozygosity at 18q21 is indicative of recurrence and therefore poor prognosis in a subset of colorectal cancers. *Br J Cancer* **79:** 903-908.

Johnson DS, Mortazavi A, Myers RM, Wold B (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316:** 1497-1502.

Jorissen RN, Lipton L, Gibbs P, Chapman M, Desai J, Jones IT *et al* (2008). DNA copy-number alterations underlie gene expression differences between microsatellite stable and unstable colorectal cancers. *Clin Cancer Res* **14:** 8061-8069.

Jung SH (2005). Sample size for FDR-control in microarray data analysis. *Bioinformatics* **21:** 3097-3104.

Kamalakaran S, Varadan V, Giercksky Russnes HE, Levy D, Kendall J, Janevski A *et al* (2011). DNA methylation patterns in luminal breast cancers differ from non-luminal subtypes and can identify relapse risk independent of other clinical variables. *Mol Oncol* **5:** 77-92.

Kaufman L, Rousseeuw PJ (1990). *Finding groups in data : an introduction to cluster analysis*. Wiley: New York.

Kemp Z, Thirlwell C, Sieber O, Silver A, Tomlinson I (2004). An update on the genetics of colorectal cancer. *Hum Mol Genet* **13 Spec No 2:** R177-185.

Kim DH, Smith-Warner SA, Spiegelman D, Yaun SS, Colditz GA, Freudenheim JL *et al* (2010). Pooled analyses of 13 prospective cohort studies on folate intake and colon cancer. *Cancer Causes Control* **21:** 1919-1930.

Knudson AG, Jr. (1993). Introduction to the genetics of primary renal tumors in children. *Med Pediatr Oncol* **21:** 193-198.

Knutsen T, Padilla-Nash HM, Wangsa D, Barenboim-Stapleton L, Camps J, McNeil N *et al* (2010). Definitive molecular cytogenetic characterization of 15 colorectal cancer cell lines. *Genes Chromosomes Cancer* **49:** 204-223.

Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER *et al* (2009). VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25:** 2283-2285.

Kopacova M, Tacheci I, Rejchrt S, Bures J (2009). Peutz-Jeghers syndrome: diagnostic and therapeutic approach. *World J Gastroenterol* **15:** 5397-5408.

Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S *et al* (2008). Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* **40:** 1253-1260.

Kotsopoulos J, Lubinski J, Lynch HT, Klijn J, Ghadirian P, Neuhausen SL *et al* (2007). Age at first birth and the risk of breast cancer in BRCA1 and BRCA2 mutation carriers. *Breast Cancer Res Treat* **105:** 221-228.

Kuhlenbaumer G, Hullmann J, Appenzellerm S (2010). Novel genomic techniques open new avenues in the analysis of monogenic disorders. *Hum Mutat*.

Lai WR, Johnson MD, Kucherlapati R, Park PJ (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* **21:** 3763-3770.

Laird PW, Jaenisch R (1994). DNA methylation and cancer. *Hum Mol Genet* **3 Spec No:** 1487-1495.

Laird PW, Jackson-Grusby L, Fazeli A, Dickinson SL, Jung WE, Li E *et al* (1995). Suppression of intestinal neoplasia by DNA hypomethylation. *Cell* **81:** 197-205.

Laken SJ, Petersen GM, Gruber SB, Oddoux C, Ostrer H, Giardiello FM *et al* (1997). Familial colorectal cancer in Ashkenazim due to a hypermutable tract in APC. *Nat Genet* **17:** 79-83.

Lander ES, Botstein D (1987). Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science* **236:** 1567-1570.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J *et al* (2001). Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Lanza G, Matteuzzi M, Gafa R, Orvieto E, Maestri I, Santini A *et al* (1998). Chromosome 18q allelic loss and prognosis in stage II and III colon cancer. *Int J Cancer* **79:** 390-395.

Larsson SC, Wolk A (2006). Meat consumption and risk of colorectal cancer: a meta-analysis of prospective studies. *Int J Cancer* **119:** 2657-2664.

Laurent-Puig P, Olschwang S, Delattre O, Remvikos Y, Asselain B, Melot T *et al* (1992). Survival and acquired genetic alterations in colorectal cancer. *Gastroenterology* **102:** 1136-1141.

Leary RJ, Lin JC, Cummins J, Boca S, Wood LD, Parsons DW *et al* (2008). Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers. *Proc Natl Acad Sci U S A* **105:** 16224-16229.

Lebel RR, Gallagher WB (1989). Wisconsin consanguinity studies. II: Familial adenocarcinomatosis. *Am J Med Genet* **33:** 1-6.

Lengauer C, Kinzler KW, Vogelstein B (1997). Genetic instability in colorectal cancers. *Nature* **386:** 623-627.

Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K *et al* (2008). DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456:** 66-72.

Li H, Ruan J, Durbin R (2008a). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18:** 1851-1858.

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

Li LH, Ho SF, Chen CH, Wei CY, Wong WC, Li LY *et al* (2006). Long contiguous stretches of homozygosity in the human genome. *Hum Mutat* **27:** 1115-1121.

Li R, Li Y, Kristiansen K, Wang J (2008b). SOAP: short oligonucleotide alignment program. *Bioinformatics* **24:** 713-714.

Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* **34:** 816-834.

Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR (2011). Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res* **21:** 940-951.

Liede A, Malik IA, Aziz Z, Rios Pd Pde L, Kwan E, Narod SA (2002). Contribution of BRCA1 and BRCA2 mutations to breast and ovarian cancer in Pakistan. *Am J Hum Genet* **71:** 595-606.

Lin WJ, Hsueh HM, Chen JJ (2010). Power and sample size estimation in microarray studies. *BMC Bioinformatics* **11:** 48.

Lleras RA, Adrien LR, Smith RV, Brown B, Jivraj N, Keller C *et al* (2011). Hypermethylation of a cluster of Kruppel-type zinc finger protein genes on chromosome 19q13 in oropharyngeal squamous cell carcinoma. *Am J Pathol* **178:** 1965-1974.

Lynch HT, Lynch PM, Lanspa SJ, Snyder CL, Lynch JF, Boland CR (2009). Review of the Lynch syndrome: history, molecular genetics, screening, differential diagnosis, and medicolegal ramifications. *Clin Genet* **76:** 1-18.

Mao X, Hamoudi RA, Talbot IC, Baudis M (2006). Allele-specific loss of heterozygosity in multiple colorectal adenomas: toward an integrated molecular cytogenetic map II. *Cancer Genet Cytogenet* **167:** 1-14.

Marra G, Boland CR (1995). Hereditary nonpolyposis colorectal cancer: the syndrome, the genes, and historical perspectives. *J Natl Cancer Inst* **87:** 1114-1125.

Martinez-Lopez E, Abad A, Font A, Monzo M, Ojanguren I, Pifarre A *et al* (1998). Allelic loss on chromosome 18q as a prognostic marker in stage II colorectal cancer. *Gastroenterology* **114:** 1180-1187.

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

Meijers-Heijboer H, van den Ouweland A, Klijn J, Wasielewski M, de Snoo A, Oldenburg R *et al* (2002). Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. *Nat Genet* **31:** 55-59.

Metcalfe K, Lubinski J, Lynch HT, Ghadirian P, Foulkes WD, Kim-Sing C *et al* (2010). Family history of cancer and cancer risks in women with BRCA1 or BRCA2 mutations. *J Natl Cancer Inst* **102:** 1874-1878.

Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S *et al* (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **266:** 66-71.

Muleris M, Chalastanis A, Meyer N, Lae M, Dutrillaux B, Sastre-Garau X *et al* (2008). Chromosomal instability in near-diploid colorectal cancer: a link between numbers and structure. *PLoS One* **3:** e1632.

Ng PC, Henikoff S (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31:** 3812-3814.

Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI *et al* (2010a). Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet* **42:** 790-793.

Ng SB, Nickerson DA, Bamshad MJ, Shendure J (2010b). Massively parallel sequencing and rare disease. *Hum Mol Genet* **19:** R119-124.

Ogino S, Nosho K, Irahara N, Shima K, Baba Y, Kirkner GJ *et al* (2009). Prognostic significance and molecular associations of 18q loss of heterozygosity: a cohort study of microsatellite stable colorectal cancers. *J Clin Oncol* **27:** 4591-4598.

Ogunbiyi AO, Ogunbiyi JO (1998). Nevus depigmentosus and inflammatory linear epidermal nevus--an unusual combination with a note on histology. *Int J Dermatol* **37:** 600-602.

Ogunbiyi OA, Goodfellow PJ, Herfarth K, Gagliardi G, Swanson PE, Birnbaum EH *et al* (1998). Confirmation that chromosome 18q allelic loss in colon cancer is a prognostic indicator. *J Clin Oncol* **16:** 427-433.

Ohtani-Fujita N, Fujita T, Aoike A, Osifchin NE, Robbins PD, Sakai T (1993). CpG methylation inactivates the promoter activity of the human retinoblastoma tumor-suppressor gene. *Oncogene* **8:** 1063-1067.

Olshen AB, Venkatraman ES, Lucito R, Wigler M (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5:** 557-572.

Ong CT, Corces VG (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat Rev Genet* **12:** 283-293.

Ozawa S, Sugano K, Sonehara T, Fukuzono S, Ichikawa A, Fukayama N *et al* (2004). High resolution for single-strand conformation polymorphism analysis by capillary electrophoresis. *Anal Chem* **76:** 6122-6129.

Park do Y, Sakamoto H, Kirley SD, Ogino S, Kawasaki T, Kwon E *et al* (2007). The Cables gene on chromosome 18q is silenced by promoter hypermethylation and allelic loss in human colorectal cancer. *Am J Pathol* **171:** 1509-1519.

Parsons DW, Jones S, Zhang X, Lin JC, Leary RJ, Angenendt P *et al* (2008). An integrated genomic analysis of human glioblastoma multiforme. *Science* **321:** 1807-1812.

Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, Garcia F *et al* (2006). High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res* **16:** 1136-1148.

Peinado MA, Malkhosyan S, Velazquez A, Perucho M (1992). Isolation and characterization of allelic losses and gains in colorectal tumors by arbitrarily primed polymerase chain reaction. *Proc Natl Acad Sci U S A* **89:** 10065-10069.

Peters U, Hutter CM, Hsu L, Schumacher FR, Conti DV, Carlson CS *et al* (2011). Meta-analysis of new genome-wide association studies of colorectal cancer risk. *Hum Genet*.

Peto J, Collins N, Barfoot R, Seal S, Warren W, Rahman N *et al* (1999). Prevalence of BRCA1 and BRCA2 gene mutations in patients with early-onset breast cancer. *J Natl Cancer Inst* **91:** 943-949.

Pharoah PD, Antoniou A, Bobrow M, Zimmern RL, Easton DF, Ponder BA (2002). Polygenic susceptibility to breast cancer and implications for prevention. *Nat Genet* **31:** 33-36.

Pietra N, Sarli L, Costi R, Ouchemi C, Grattarola M, Peracchia A (1998). Role of follow-up in management of local recurrences of colorectal cancer: a prospective, randomized study. *Dis Colon Rectum* **41:** 1127-1133.

Pilozzi E, Ferri M, Onelli MR, Mercantini P, Corigliano N, Duranti E *et al* (2011). Prognostic significance of 18q LOH in sporadic colorectal carcinoma. *Am Surg* **77:** 38-43.

Poynter JN, Gruber SB, Higgins PD, Almog R, Bonner JD, Rennert HS *et al* (2005). Statins and the risk of colorectal cancer. *N Engl J Med* **352:** 2184-2192.

Pritchard JK (2001). Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* **69:** 124-137.

Pujana MA, Han JD, Starita LM, Stevens KN, Tewari M, Ahn JS *et al* (2007). Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nat Genet* **39:** 1338-1349.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D *et al* (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81:** 559-575.

Rauch T, Li H, Wu X, Pfeifer GP (2006). MIRA-assisted microarray analysis, a new technology for the determination of DNA methylation patterns, identifies frequent methylation of homeodomain-containing genes in lung cancer cells. *Cancer Res* **66:** 7939-7947.

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ *et al* (2001). Linkage disequilibrium in the human genome. *Nature* **411:** 199-204.

Reich DE, Lander ES (2001). On the allelic spectrum of human disease. *Trends Genet* **17:** 502-510.

Rennert G, Bisland-Naggan S, Barnett-Griness O, Bar-Joseph N, Zhang S, Rennert HS *et al* (2007). Clinical outcomes of breast cancer in carriers of BRCA1 and BRCA2 mutations. *N Engl J Med* **357:** 115-123.

Rennert G, Pinchev M, Rennert HS, Gruber SB (2011). Use of bisphosphonates and reduced risk of colorectal cancer. *J Clin Oncol* **29:** 1146-1150.

Renwick A, Thompson D, Seal S, Kelly P, Chagtai T, Ahmed M *et al* (2006). ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. *Nat Genet* **38:** 873-875.

Rink L, Skorobogatko Y, Kossenkov AV, Belinsky MG, Pajak T, Heinrich MC *et al* (2009). Gene expression signatures and response to imatinib mesylate in gastrointestinal stromal tumor. *Mol Cancer Ther* **8:** 2172-2182.

Rudan I (1999). Inbreeding and cancer incidence in human isolates. *Hum Biol* **71:** 173-187.

Rudan I, Ranzani GN, Strnad M, Vorko-Jovic A, John V, Unusic J *et al* (1999). Surname as 'cancer risk' in extreme isolates: example from the island of Lastovo, Croatia. *Coll Antropol* **23:** 557-569.

Sampson JR, Dolwani S, Jones S, Eccles D, Ellis A, Evans DG *et al* (2003). Autosomal recessive colorectal adenomatous polyposis due to inherited mutations of MYH. *Lancet* **362:** 39-41.

Scharpf RB, Parmigiani G, Pevsner J, Ruczinski I (2008). Hidden Markov models for the assessment of chromosomal alterations using high-throughput SNP arrays. *Ann Appl Stat* **2:** 687-713.

Schwartz MD, Isaacs C, Graves KD, Poggi E, Peshkin BN, Gell C *et al* (2011). Long-term outcomes of BRCA1/BRCA2 testing: risk reduction and surveillance. *Cancer*.

Shannon M, Ashworth LK, Mucenski ML, Lamerdin JE, Branscomb E, Stubbs L (1996). Comparative analysis of a conserved zinc finger gene cluster on human chromosome 19q and mouse chromosome 7. *Genomics* **33:** 112-120.

Sotelo J, Esposito D, Duhagon MA, Banfield K, Mehalko J, Liao H *et al* (2010). Long-range enhancers on 8q24 regulate c-Myc. *Proc Natl Acad Sci U S A* **107:** 3001-3005.

Spain SL, Cazier JB, Houlston R, Carvajal-Carmona L, Tomlinson I (2009). Colorectal cancer risk is not associated with increased levels of homozygosity in a population from the United Kingdom. *Cancer Res* **69:** 7422-7429.

Suzuki S, Ono N, Furusawa C, Ying BW, Yomo T (2011). Comparison of sequence reads obtained from three next-generation sequencing platforms. *PLoS One* **6:** e19534.

Tabangin ME, Woo JG, Martin LJ (2009). The effect of minor allele frequency on the likelihood of obtaining false positives. *BMC Proc* **3 Suppl 7:** S41.

Teebi AS, Farag TI (1997). *Genetic disorders among Arab populations*. Oxford University Press: New York.

Teebi AS, Teebi SA (2005). Genetic diversity among the Arabs. *Community Genet* **8:** 21-26.

Tenesa A, Dunlop MG (2009). New insights into the aetiology of colorectal cancer from genome-wide association studies. *Nat Rev Genet* **10:** 353-358.

Thiagalingam S, Lengauer C, Leach FS, Schutte M, Hahn SA, Overhauser J *et al* (1996). Evaluation of candidate tumour suppressor genes on chromosome 18 in colorectal cancers. *Nat Genet* **13:** 343-346.

Thibodeau SN, Bren G, Schaid D (1993). Microsatellite instability in cancer of the proximal colon. *Science* **260:** 816-819.

Thomas HJ, Whitelaw SC, Cottrell SE, Murday VA, Tomlinson IP, Markie D *et al* (1996). Genetic mapping of hereditary mixed polyposis syndrome to chromosome 6q. *Am J Hum Genet* **58:** 770-776.

Tonin P, Weber B, Offit K, Couch F, Rebbeck TR, Neuhausen S *et al* (1996). Frequency of recurrent BRCA1 and BRCA2 mutations in Ashkenazi Jewish breast cancer families. *Nat Med* **2:** 1179-1183.

Venkatraman ES, Olshen AB (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23:** 657-663.

Vilar E, Mukherjee B, Kuick R, Raskin L, Misek DE, Taylor JM *et al* (2009). Gene expression patterns in mismatch repair-deficient colorectal cancers highlight the potential therapeutic role of

inhibitors of the phosphatidylinositol 3-kinase-AKT-mammalian target of rapamycin pathway. *Clin Cancer Res* **15:** 2829-2839.

Vilar E, Gruber SB (2010). Microsatellite instability in colorectal cancer-the stable evidence. *Nat Rev Clin Oncol* **7:** 153-162.

Vogelstein B, Fearon ER, Hamilton SR, Kern SE, Preisinger AC, Leppert M *et al* (1988). Genetic alterations during colorectal-tumor development. *N Engl J Med* **319:** 525-532.

Vogelstein B, Fearon ER, Kern SE, Hamilton SR, Preisinger AC, Nakamura Y *et al* (1989). Allelotype of colorectal carcinomas. *Science* **244:** 207-211.

Wacholder S, Hartge P, Prentice R, Garcia-Closas M, Feigelson HS, Diver WR *et al* (2010). Performance of common genetic variants in breast-cancer risk models. *N Engl J Med* **362:** 986-993.

Wahlfors J, Hiltunen H, Heinonen K, Hamalainen E, Alhonen L, Janne J (1992). Genomic hypomethylation in human chronic lymphocytic leukemia. *Blood* **80:** 2074-2080.

Wang K, Dickson SP, Stolle CA, Krantz ID, Goldstein DB, Hakonarson H (2010a). Interpretation of association signals and identification of causal variants from genome-wide association studies. *Am J Hum Genet* **86:** 730-742.

Wang S, Haynes C, Barany F, Ott J (2009). Genome-wide autozygosity mapping in human populations. *Genet Epidemiol* **33:** 172-180.

Wang W, Li YF, Sun XW, Chen G, Zhan YQ, Huang CY *et al* (2010b). Correlation analysis between loss of heterozygosity at chromosome 18q and prognosis in the stage-II colon cancer patients. *Chin J Cancer* **29:** 761-767.

Wang WY, Barratt BJ, Clayton DG, Todd JA (2005). Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet* **6:** 109-118.

Wang Z, Cummins JM, Shen D, Cahill DP, Jallepalli PV, Wang TL *et al* (2004). Three classes of genes mutated in colorectal cancers with chromosomal instability. *Cancer Res* **64:** 2998-3001.

Ward AM, Morris RP (1963). *A Theological book list of works in English, French, German, Portuguese, Spanish*. Theological Education Fund ;

Distributor, Allenson's: London

Naperville, Ill.

Watanabe A, Higuchi M, Fukushi M, Ohsawa T, Takahashi M, Oie M *et al* (2007). A novel KRAB-Zinc finger protein interacts with latency-associated nuclear antigen of Kaposi's sarcoma-associated herpesvirus and activates transcription via terminal repeat sequences. *Virus Genes* **34:** 127-136.

Watanabe T, Wu TT, Catalano PJ, Ueki T, Satriano R, Haller DG *et al* (2001). Molecular predictors of survival after adjuvant chemotherapy for colon cancer. *N Engl J Med* **344:** 1196-1206.

Weiss KM, Clark AG (2002). Linkage disequilibrium and the mapping of complex human traits. *Trends Genet* **18:** 19-24.

Whitelaw SC, Murday VA, Tomlinson IP, Thomas HJ, Cottrell S, Ginsberg A *et al* (1997). Clinical and molecular features of the hereditary mixed polyposis syndrome. *Gastroenterology* **112:** 327-334.

Willenbrock H, Fridlyand J (2005). A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics* **21:** 4084-4091.

Wooster R, Neuhausen SL, Mangion J, Quirk Y, Ford D, Collins N *et al* (1994). Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. *Science* **265:** 2088-2090.

Wooster R, Bignell G, Lancaster J, Swift S, Seal S, Mangion J *et al* (1995). Identification of the breast cancer susceptibility gene BRCA2. *Nature* **378:** 789-792.

Yang Z, Wen HJ, Minhas V, Wood C (2009). The zinc finger DNA-binding domain of K-RBP plays an important role in regulating Kaposi's sarcoma-associated herpesvirus RTA-mediated gene expression. *Virology* **391:** 221-231.

Zanke BW, Greenwood CM, Rangrej J, Kustra R, Tenesa A, Farrington SM *et al* (2007). Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat Genet* **39:** 989-994.

Zhang H, Arbman G, Sun XF (2003). Codon 201 polymorphism of DCC gene is a prognostic factor in patients with colorectal cancer. *Cancer Detect Prev* **27:** 216-221.

Zhou S, Buckhaults P, Zawel L, Bunz F, Riggins G, Dai JL *et al* (1998). Targeted deletion of Smad4 shows it is required for transforming growth factor beta and activin signaling in colorectal cancer cells. *Proc Natl Acad Sci U S A* **95:** 2412-2416.

Zhou W, Goodman SN, Galizia G, Lieto E, Ferraraccio F, Pignatelli C *et al* (2002). Counting alleles to predict recurrence of early-stage colorectal cancers. *Lancet* **359:** 219-225.