

Genome-wide approaches to identifying the etiologies of complex diseases: applications in colorectal cancer and congenital heart disease

by

Stephanie Loie Stenzel

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Epidemiological Science)
in the University of Michigan
2013

Doctoral Committee:

Professor Stephen B. Gruber, Co-Chair, University of Southern California
Professor Sharon R. Kardia, Co-Chair
Associate Professor Peter J. Gruber, University of Utah
Assistant Professor Jun Li
Associate Professor Bhramar Mukherjee
Professor Patricia A. Peyser

© Stephanie Stenzel

2013

Dedication

To my parents for their love, support, and guidance.

Acknowledgements

The dissertation process has been a long and winding road, and I would not be mentally or physically where I am today without the overwhelming support from my family, friends, colleagues, and academic advisors. First and foremost, I would to thank my parents for helping me through all of the critical decisions during my academic career and for showing me by example the value of hard work. Many thanks to my sister and to the California Stenzel family who have provided relief and a level head during many stressful times. Also, I want to express my gratitude to my fiancé and best friend, Scott Schmit, who has seen me through the happy, stressful, frustrating, exciting, and difficult times during this process.

This chapter of my life would not have been the same without my great friends, including some who love science and some who just love me enough to hear about it. Thanks to each of you for lending a listening ear and for making my life simpler and happier in every way possible. A special thank you to Michele Gornick, Kaanan Shah, and Val Schaibley for being fantastic role models, constructive critics, and friends during this academic adventure. Also, I sincerely appreciate the time and energy invested by Erin Payne, Erin Ware, and Alicia Lazarus in helping to get my dissertation to the point it is today and in encouraging me every step of the way.

This dissertation research surely was not conducted in a silo and had many contributors aside from my own efforts. Many thanks go to the Kardia lab members as well as former and current members of the Gruber lab who have impacted the way that I approach research questions and have helped me with everything from running PCR reactions to discussing life

after the PhD. Among the many special lab contributors, this work would not have been possible or of the same quality without Michele Gornick, Shu-chen Huang, Kevin McDonnell, Maria Dolores Iniesta, Eduardo Vilar Sanchez, Leon Raskin, Jessica Long, Jishu Xu (Li lab), Sherry Taylor, Mai Castillo, Jennifer Smith, Wei Zhao, and Josh Buckner. Also, special thanks to Kristen Stevens for setting the foundation for the whole genome sequencing project described in Chapter 4. I would like to recognize the MECC investigators, especially Gad Rennert, for the opportunity to work on the projects in Chapters 2 and 3, as well as the Colon CFR and CORECT investigators, especially Chris Edlund and Fred Schumacher, who were instrumental for the analysis in Chapter 2. Further, I would like to acknowledge Matthew Flickinger for contributing his incredible programming skills on multiple occasions when I was at dead end and for his patience in teaching me throughout the process. Also, I would like to recognize the research participants who made all of this work possible.

In closing, I would like to express my sincere appreciation for the time, effort, and expertise dedicated by members of my dissertation committee to my work over the past three years. Stephen Gruber, Sharon Kardia, Bhramar Mukherjee, Jun Li, Peter Gruber, and Patricia Peyser have served as wonderful academic role models and advisors in the fields of epidemiology, human genetics, biostatistics, and molecular biology. While not on my dissertation committee, I owe my early genetic epidemiology training and my decision to pursue a PhD to Julia Richards, one of the most brilliant and caring mentors that a student could ask for. Finally, I would like to extend my deepest thanks to Stephen Gruber, my primary mentor and advisor, who has provided me with this amazing research experience. I cannot express enough my gratitude for your personal contributions to my training and career and for your lab that has provided me with a collaborative learning environment that feels like home.

Table of Contents

Dedication	ii
Acknowledgements	iii
List of Tables	vii
List of Figures.....	viii
List of Abbreviations	x
Abstract.....	xii
Chapter 1. Introduction	1
1.1 The burden of complex diseases	1
1.2 Heritability of CRC and CHD.....	2
1.3 Paradigm shift in genetic epidemiology study design	3
1.4 Summary of chapters	4
1.5 Overview.....	6
Chapter 2. A novel low-penetrance risk locus for colorectal cancer at 4q32.2: findings from a genome-wide association study meta-analysis.....	8
2.1 Background.....	8
2.2 Materials and methods	13
2.2.1 Study populations.....	13
2.2.1.1 Molecular Epidemiology of Colorectal Cancer Study (MECC).....	15
2.2.1.2 Colon Cancer Family Registry (CFR)	15
2.2.2 Genotyping and quality control	16
2.2.2.1 Discovery meta-analysis	16
2.2.2.2 Replication analysis	21
2.2.3 Imputation.....	22
2.2.4 Statistical analysis.....	23
2.2.4.1 Discovery GWAS meta-analysis	23
2.2.5 Replication in MECC and joint meta-analysis.....	25
2.3 Results.....	25
2.3.1 Discovery meta-analysis	25
2.3.2 Replication	31
2.4 Discussion	36
Chapter 3. MicroRNA target site polymorphisms and colorectal cancer risk in the Ashkenazi Jewish population.....	41
3.1 Introduction.....	41
3.2 Materials and methods	43
3.2.1 Study population: Molecular Epidemiology of Colorectal Cancer (MECC) Study ..	43
3.2.2 Genotyping and quality control	44
3.2.3 Gene expression quantification.....	46
3.2.4 Statistical and bioinformatic analysis	47
3.3 Results.....	48

3.3.1	Targeted genome-wide association analysis	48
3.3.2	Gene expression analysis for top association findings	52
3.3.3	Replication of previously published risk loci	54
3.3.4	Genotype concordance: miRNA targeted array vs. traditional GWAS array	57
3.4	Discussion	58
Chapter 4. Identification of potentially causal genetic variants for complex congenital heart disease through whole-genome sequencing		62
4.1	Background	62
4.2	Methods	65
4.2.1	Family recruitment	65
4.2.2	Whole genome sequence data generation	65
4.2.3	Alignment of WGS paired-end reads	66
4.2.4	SNP and insertion/deletion (indel) calling and annotation	67
4.2.5	Structural variant identification	69
4.2.6	Co-segregation analysis	69
4.2.7	Sanger sequencing	73
4.2.8	Quality control	73
4.2.9	Sequencing quality and contamination	73
4.3	Results	75
4.3.1	WGS data summary	75
4.3.2	Variant summary: SNVs, indels, and structural variants	76
4.3.3	Co-segregation analysis	80
4.4	Discussion	87
Chapter 5. Summary and Conclusions		93
References		97

List of Tables

Table 2.1. Published genome-wide significant ($p < 5 \times 10^{-8}$) CRC risk loci.....	12
Table 2.2. Demographic and clinical characteristics of genotyped participants after quality control from the Molecular Epidemiology of Colorectal Cancer (MECC; $n_{\text{discovery}} = 983$; $n_{\text{replication}} = 1,962$) and Colon Cancer Family Registry (CFR; $n_{\text{discovery}} = 2,977$) case-control studies.	14
Table 2.3. Summary of SNPs associated with CRC ($p < 3 \times 10^{-6}$) in the discovery MECC + CFR meta-analysis.....	29
Table 2.4. Discovery MECC + CFR meta-analysis association results for previously identified risk alleles..	30
Table 2.5. Top hits from MECC discovery (Omni) + CFR discovery + MECC replication (Axiom) meta-analysis.....	32
Table 2.6. Assessment of height as a potential confounder of the most significant SNP-CRC associations from the joint meta-analysis.	34
Table 3.1. Demographic and clinical characteristics of MECC participants ($n=1,025$) genotyped on the Axiom [®] miRNA Target Site Genotyping Array platform.	44
Table 3.2. Variant sources and design of the Axiom [®] miRNA Target Site Genotyping Array ...	45
Table 3.3. Summary of variants with the most statistically significant associations from the age, sex, and PC 1-2 adjusted logistic regression models.	51
Table 3.4. ANOVA results for gene expression [$\log_2(\text{normalized intensity})$] by genotype at top SNPs from the logistic regression association analyses between each SNP and CRC status.	54
Table 3.5. MECC results from logistic regression adjusted for age, sex, and PC 1-2 for previously studied miRNA variants.....	56
Table 4.1. Candidate gene list for complex CHD.....	72
Table 4.2. VerifyBamID FreeMix estimates for each individual with WGS reads combined across sequencing lanes.	75
Table 4.3. VerifyBamID FreeMix estimates for each lane of WGS reads generated for the half-sib with TOF.	75
Table 4.4. Whole-genome sequence data summary for a family affected by complex CHD.....	76
Table 4.5. SNP and small indel summary for the overall multi-sample VCF and for the per individual call sets.....	78
Table 4.6. Summary of potentially causative SNVs and small indels co-segregating according to an autosomal dominant (AD) or autosomal recessive (AR) model.	85
Table 4.7. Summary of potentially causative SNVs and small indels co-segregating according to an autosomal dominant (AD) or autosomal recessive (AR) model <i>without consideration of half-sib with truncus arteriosus' genotype due to low depth of coverage</i>	86

List of Figures

Figure 2.1. Quality control and filtering pipeline for discovery MECC samples prior to imputation.	17
Figure 2.2. Pairwise plots of principal components from the discovery phase MECC Omni samples.	18
Figure 2.3. Pairwise plots of principal components from the discovery phase CFR samples.	20
Figure 2.4. Quality control and filtering pipeline for MECC Axiom [®] replication samples prior to imputation.	21
Figure 2.5. PC1 versus PC2 plot from the replication phase MECC Axiom samples.	22
Figure 2.6. Q-Q plot of p-values derived from the inverse-variance-weighted, fixed-effects meta-analysis of adjusted logistic regression model results run on MECC Omni + CFR Illumina discovery samples.	26
Figure 2.7. Manhattan plot of $-\log_{10}(\text{p-values})$ from the inverse-variance-weighted, fixed-effects meta-analysis of discovery stage MECC and CFR adjusted association results.	27
Figure 2.8. Height by case/control status in all MECC Axiom participants.	33
Figure 2.9. LocusZoom ⁸⁵ plot of regional association results for the novel 4q32.2 genome-wide significant locus (rs17042479 +/- 1Mb).	35
Figure 2.10. LocusZoom ⁸⁵ plot of regional association results for the suggestive and previously identified 18q21 locus.	35
Figure 3.1. Quality control and filtering pipeline for MECC samples genotyped on the Affymetrix Axiom [®] miRNA Target Site Genotyping Array.	46
Figure 3.2. Plots of principal components 1 versus 2 and 1 versus 3 based on 4,736 AIMs from the MECC participants genotyped on the Axiom [®] miRNA Target Site Genotyping Array.	49
Figure 3.3. Quantile-quantile and Manhattan plots of p-values for association between each SNP and CRC from logistic regression adjusted for sex, age, and 2 PCs.	50
Figure 3.4. <i>RAPGEF2</i> gene expression in colorectal cancers measured by two separate probes in 72 MECC CRC cases by rs6827968 genotype.	54
Figure 3.5. <i>INSR</i> gene expression in 135 MECC CRC cases by rs1051690 genotype.	57
Figure 3.6. Concordance between directly measured and 1000 Genomes imputed CORECT Axiom genotypes and miRNA Target Site Array genotypes in MECC.	58
Figure 4.1. Pedigree for the index family affected by complex CHD.	65
Figure 4.2. Tranche (or truth sensitivity)-specific results from variant quality score recalibration.	77
Figure 4.3. Total number of SNVs and indels identified by chromosome for the proband with HLHS.	79
Figure 4.4. SNVs and indels shared by three affected individuals with severe CHD.	80
Figure 4.5. Visualization of SNVs and indels that follow AD patterns of inheritance.	81
Figure 4.6. SNVs and small indels co-segregating with disease according to an autosomal dominant genetic model with incomplete penetrance.	81

Figure 4.7. SNVs and small indels co-segregating with disease according to an autosomal dominant with gonadal mosaicism model.....	82
Figure 4.8. SNVs and small indels co-segregating with disease according to an autosomal recessive model.....	83

List of Abbreviations

1000G	1000 Genomes Project
A1	Effect allele
APC	Adenomatous polyposis coli
AIM	Ancestry informative marker
AJ	Ashkenazi Jewish
AD	Autosomal dominant
AR	Autosomal recessive
AVG	Average
BP	Base (nucleotide) position
BQSR	Base quality score recalibration
BAM	Binary alignment/map
CHR	Chromosome
CFR	Colon Cancer Family Registry
CORECT	Colorectal Transdisciplinary Study
CRC	Colorectal cancer
CI	Confidence interval
CHD	Congenital heart disease
CWRU	Case Western Reserve University
DP	Depth
eQTL	Expression quantitative trait locus
FS	Fisher's exact test for strand bias
FHCRC	Fred Hutchinson Cancer Research Center
FSTL5	Follistatin-like 5
GA-II	Illumina Genome Analyzer II
GxE	Gene-by-environment
GATK	Genome Analysis Toolkit
GWAS	Genome-wide association study
GC λ	Genomic control lambda
HapMap	The International HapMap Project
HWE	Hardy Weinberg Equilibrium
HLHS	Hypoplastic left heart syndrome
IBD	Identity by descent
Indel	Insertion or deletion
LD	Linkage disequilibrium
MQ	Mapping quality
MQRankSum	Mapping Quality Rank Sum Test
MAF	Minor allele frequency
MAP	MYH-associated polyposis

MECC	Molecular Epidemiology of Colorectal Cancer
mRNA	Messenger RNA
miRNA	MicroRNA
MSI-H	Microsatellite instable
NCI	National Cancer Institute
NGS	Next-generation sequencing
OR	Odds ratio
PCR	Polymerase chain reaction
PP2	Polyphen2
PC	Principal component
PCA	Principal component analysis
QD	Quality by depth ratio
QC	Quality control
Q-Q	Quantile-quantile
RNA	Ribonucleic acid
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SD	Standard deviation
TOF	Tetralogy of Fallot
TGF- β	Transforming growth factor-beta
Ti/Tv	Transition-to-transversion ratio
US	United States
UM	University of Michigan
USC	University of Southern California
VCF	Variant call format
VQSR	Variant quality score recalibration
WGS	Whole-genome sequencing

Abstract

Complex diseases, such as cancer and heart disease, have complicated genetic architectures. The genetic risk factors underlying these diseases that represent public health challenges have not been comprehensively characterized, and a more thorough understanding of the genetic contributors to their multi-factorial etiologies may offer new insights into the optimal design of prevention and early intervention strategies. Combined with the advent of lower-cost, high-throughput genomic technologies, epidemiologic studies have enhanced potential to identify subgroups of the population with unique genetic etiologies and to fill in missing knowledge regarding genetic susceptibility to these complex phenotypes. A spectrum of genetic variants exists across the continuums of effect size and allele frequency, and with this spectrum, comes a variety of etiological mechanisms that require specialized detection methods. This dissertation applies genome-wide association study (GWAS) approaches, as well as a whole-genome sequencing (WGS)-based co-segregation design, to identify susceptibility loci for two complex diseases: colorectal cancer (CRC) and severe congenital heart disease (CHD).

The first substantive chapter of this dissertation, Chapter 2, examines common genetic variation contributing to the etiology of CRC through a GWAS meta-analysis. This study, conducted among non-Hispanic whites and a founder population of Ashkenazi Jews from northern Israel, identified a novel, genome-wide significant risk locus for CRC on chromosome 4q32.2.

The second dissertation project (Chapter 3) continues the search for CRC susceptibility loci with a targeted GWAS approach in Ashkenazi Jews from northern Israel. This study was designed to expand the search for microRNA (miRNA)-related polymorphisms important in the etiology of CRC across the genome and to investigate the association between genetic variants in miRNA target sites and CRC risk using a novel genotyping platform. This chapter highlighted several suggestive risk variants with predicted miRNA binding implications. It also replicated a recent association finding between a variant in *INSR* and CRC risk and demonstrated variability in *INSR* gene expression by genotype. Most importantly, this chapter demonstrated the potential for a targeted GWAS to identify candidate risk loci and to prioritize them for functional characterization.

Finally, Chapter 4 focuses on identifying genetic risk factors contributing to the etiology of complex CHD. The primary goal of this study was to identify rare genetic variation underlying the development of complex CHD through WGS of a family affected by hypoplastic left heart syndrome (HLHS) and conotruncal defects. The hypothesis was that the clustering of these anatomically distinct lesions was attributable to a more proximal defect in cardiogenesis. Ultimately, thirty-two variants across 29 genes were identified as causal candidates for future validation, replication, and follow-up epidemiologic studies.

Taken together, these three dissertation chapters enhance our understanding of genetic contributors to two complex phenotypes through detailed investigation of both large populations and a highly informative family. Future directions include, but are not limited to, fine mapping and functional studies of identified susceptibility loci for colorectal cancer and epidemiologic studies of the population-level impact of potentially causal variants identified for complex CHD.

Chapter 1. Introduction

1.1 The burden of complex diseases

Complex diseases, such as colorectal cancer and congenital heart disease, are multi-faceted collections of pathology with complicated etiologies. Several component causes combine to increase the risk of developing these disease phenotypes at both the individual and population levels¹. Both colorectal cancer and congenital heart disease are caused by a combination of genetic factors, environmental influences, gene-by-gene interactions, and/or gene-by-environment (GxE) interactions^{2,3}. Despite their anatomical and etiologic diversity, all of these diseases have underlying pathogenic processes that can be described as complex. Their complicated natures make their etiologies challenging to study but also enrich the potential opportunities for early detection, tailored treatment, and prevention through modification of disease risk factors.

The complex disease burden is currently on the rise in the United States (US) and across the developed world, as the global population structure ages, lifestyles become more sedentary, and diets change. The number of deaths due to chronic diseases now far exceeds those attributable to communicable diseases in the Western world⁴. Congenital heart disease and colorectal cancer, the two complex conditions under study in this dissertation, are representative of the broader classes of diseases of heart disease and cancer that rank as the top one and two leading causes of death in the US, respectively⁴. Particularly pertinent to this dissertation, it is important to note that colorectal cancer (CRC) and congenital heart disease (CHD) are two conditions with long-term implications for quality of life⁵⁻⁷ and economic burden^{8,9} on the US

healthcare system. Etiologic research constitutes the first step towards informing prevention, screening, and therapeutic development.

Some recent efforts to combat complex diseases have focused on personalized or “precision” medicine that focuses on the development of tailored strategies to treat the right patient with the right therapeutic or intervention at the right time based on genomic and clinical profiles^{10,11}. In addition, a major objective from the epidemiologic perspective in the era of genomic medicine is to understand the determinants of disease in a population with the goal of informing primary or secondary prevention efforts. In order to discover methods for early intervention and ultimately prevention, we must first better understand the underlying causes of these complex diseases and their development processes. Among the multitude of potential disease drivers, the focus of this dissertation is on the genetic risk factors conferring susceptibility to CRC and CHD.

1.2 Heritability of CRC and CHD

Narrow-sense heritability is a population-level concept referring to the proportion of phenotypic variance attributable to the additive effects of genetic factors. For a binary trait, each individual is assumed to have a liability of developing disease derived from an underlying distribution in the population. In this case, heritability captures the proportion of variance in disease risk attributable to additive genetic factors. Heritability estimates for CRC and severe CHD range from 32 to 35% (95% confidence interval (CI): 10%-48%)^{12,13} and 74% (95% CI: 50%-96%) to 99% (95% CI: 59%-100%) (depending on the lesion under study)^{14,15}, respectively. These heritabilities indicate that inherited genetic factors play a substantial role in the etiologic landscape of these two complex diseases, with an even greater contribution to the phenotypic variance of risk for CHD. However, the specific genetic risk factors underlying these complex pathophysiologies remain to be comprehensively characterized, and this is typically

referred to as the missing heritability problem^{2,16}. Further, for those implicated genetic loci, information about the biological functionality of the variation remains limited.

A more thorough understanding of the genetic contributors to these complex phenotypes may offer new insights into their underlying biology. Studies of both common complex manifestations and extreme phenotypes of the same disease state may complement each other to inform potentially shared genes or pathways important for pathogenesis..

1.3 Paradigm shift in genetic epidemiology study design

Over the past two decades, major shifts in the paradigm of genetic epidemiological investigations of complex disease etiology have occurred alongside improvements in genetic technologies and advances in our understanding of genetic variants on the population scale¹⁷. Prior to the first publication of the human genome sequence in 2001¹⁸ and continuing through the early 2000s, the literature base was saturated with candidate gene single nucleotide polymorphism (SNP) association studies, often lacking statistical power. At that time, tools were not broadly available for testing agnostic hypotheses, so researchers relied solely upon existing knowledge about disease biology to drive their study designs¹⁹. Subsequently, the International Haplotype Map (HapMap) Project catalogued common genetic variation across European, African, and East Asian populations and demonstrated that between half of a million and a million SNPs could tag nearly 90% of all common variation (minor allele frequency (MAF) $\geq 5\%$) across the genome due to linkage disequilibrium patterns²⁰⁻²³. Linkage disequilibrium refers to the non-random association of specific alleles at neighboring genetic loci that results from several population genetic forces. Leveraging this concept allowed for the development of genotyping arrays that could measure a majority of common genetic variation using a single assay.

Between 2005 and 2007, genome-wide association studies (GWAS) leveraging haplotype-tagging SNP genotyping chips gained traction, allowing researchers to conduct hypothesis-free searches for common variants influencing risk of common diseases²⁴. From the mid- to late-2000's, the advent of "next-generation" sequencing (NGS) technologies took hold, and an era of even higher-throughput, whole-genome characterization and genetic variant identification began alongside GWAS. As this most recent set of technologies has driven the rapid decline in whole-genome and targeted deep sequencing costs, researchers have found themselves circling back to the importance of biologically-driven hypotheses and candidate gene or pathway studies of both common and rare genetic variation^{19,25}. It has long been recognized that a spectrum of genetic variants conferring disease risk exists across the continuums of effect size and allele frequency. However, throughout the aforementioned evolution in human genetics, statistical genetics, and genetic epidemiology, an active debate has arisen regarding whether rare or common variants contribute more substantially to complex disease etiologies^{2,26,27}. The combination genetic technologies developed over the last decade now allow for the interrogation of variation across the frequency spectrum.

1.4 Summary of chapters

Paralleling the genetic epidemiology paradigm shift outlined above, this dissertation is designed to show by example how three different quantitative approaches can be leveraged to investigate the genetic etiologies underlying complex diseases. Specifically, the following chapters demonstrate how unique study designs and quantitative approaches have the potential to characterize novel genetic contributors to CRC and CHD development, particularly in subgroups of the population.

The first two substantive chapters of this dissertation focus on elucidating novel genetic risk factors for CRC through agnostic and targeted GWAS approaches. In Chapter 2, this

dissertation explores low penetrance susceptibility loci for CRC through a GWAS meta-analysis. The common disease common variant hypothesis proposes that a multitude of common, low- to moderate-penetrance susceptibility alleles account for a substantial portion of the unidentified genetic risk factors^{24,26,28-31}. Tools that measure genetic variation across genome in an unbiased fashion offer an informative approach to identify novel risk loci when used in conjunction with association study methods. Chapter 2 expands upon the traditional agnostic GWAS approach by focusing on susceptibility locus discovery in a founder population of Ashkenazi Jews from a population-based study in northern Israel, the Molecular Epidemiology of Colorectal Cancer (MECC) study, with power for effect detection supplemented by the inclusion of non-Hispanic whites from the Colon Cancer Family Registry (CFR). I hypothesize that unidentified low penetrance susceptibility alleles confer risk to CRC.

In Chapter 3, I introduce a targeted GWAS approach to study genetic risk factors for CRC, again in the Ashkenazi Jewish population of northern Israel from the MECC study. Many targeted studies to date have focused on genetic variation in candidate genes from known biological pathways involved in CRC disease etiology. Instead, this chapter takes a new approach and studies the global impact of variation in genes affecting a pathway for regulatory control of gene expression. MicroRNAs (miRNAs) are short ribonucleic acid (RNA) molecules that act as post-transcriptional regulators of gene expression via binding to the 3' untranslated regions (3'-UTRs) of one or more messenger RNAs (mRNAs)³². This binding acts to repress translation of the messenger into protein or to signal for degradation of the targeted mRNA. SNPs found in the miRNA sequence and/or corresponding binding sites can affect the fidelity of this miRNA-mRNA interaction, and ultimately, alter the risk of tumor development³³. The study in Chapter 3 is designed to expand the search for miRNA-related genetic variants important in

cancer etiology across the genome and to investigate the association between thousands of genetic variants in miRNA target sites in 3'-UTR regions and miRNA-encoding genes and CRC risk using a novel genotyping platform. Further, this chapter bioinformatically characterizes the potential functional consequences with respect to miRNA binding of the most statistically significantly associated SNPs. Given that our research group was given early, prioritized access to a novel commercial platform to assess miRNA-related polymorphisms before it was released to the general public, I am confident that this represents the first study to examine associations between genetic variation in miRNA target sites and CRC risk using a genome-wide approach.

In the final substantive chapter (Chapter 4), this dissertation shifts to the frontier of NGS in genetic epidemiology and explores whole-genome sequencing (WGS) along with family-based co-segregation analysis as a method for identifying genes and pathways involved in CHD etiology. This approach agnostically identifies potentially causal novel and rare variants in genes or pathways important for disease development for follow-up in larger populations. I hypothesize that potentially causative genetic variants for complex CHD in a family affected by hypoplastic left heart syndrome (HLHS) and conotruncal lesions will be identifiable through WGS and informative regarding candidate genes or pathways for future investigations. This chapter leverages family data as a method to screen for and prioritize genes and/or pathways for study at the population level. Finally, Chapter 5 summarizes the key findings from this dissertation and the dissertation's contributions to the field of genetic epidemiology.

1.5 Overview

In summary, genetic epidemiological study designs have evolved over the past decade beginning with candidate gene association studies. GWAS have spurred renewed interest in biological pathway and expression quantitative trait loci (eQTL) analyses for both common and rare genetic variation. Recently, NGS-based segregation and association studies have emerged at

the forefront of rare variant study for Mendelian and complex diseases. This dissertation aims to demonstrate by example that both common and rare variants can play important roles in the genetic etiology of complex diseases. Further, I hope to convey that both population-based and family-based study designs add value to understanding the complicated milieu of genetic risk factors for complex phenotypes. Each of the study designs outlined above offer unique insights about genetic contributors to CRC and CHD, and although the focus here is on two specific complex conditions, these methods are broadly applicable to the study of genetic risk factors for other complex phenotypes.

Chapter 2. A novel low-penetrance risk locus for colorectal cancer at 4q32.2: findings from a genome-wide association study meta-analysis

2.1 Background

In the United States, colorectal cancer (CRC) has the third highest age-adjusted incidence of all cancer sites in men and women and ranks second in terms of age-adjusted mortality³⁴. Worldwide, it is the second most commonly diagnosed cancer in women and third in men; it ranks as the fourth most common cause of cancer death^{35,36}. Contributing to these summary statistics are several special populations bearing a disproportionate burden of CRC incidence. Nearly 60% of CRC cases occur in developed regions of the world^{35,36}. Among these high burden groups is the genetically homogeneous founder population of Ashkenazi Jewish individuals, with incidence rates comparable to or higher than those experienced by non-Hispanic whites in the United States, despite differences in dietary and lifestyle risk factors³⁴. Comparing population subgroups from Israel, age-standardized incidence among Ashkenazi (European or American born) Jews is 41.9 per 100,000, strikingly higher than Sephardi (Asian or African born) Jewish, Israel-born Jewish, and Israeli non-Jewish rates (25.5, 32.8, and 10.1 per 100,000, respectively)³⁷. This stratification of Israeli incidence rates, with Ashkenazi Jews experiencing the highest burden, has been maintained consistently over time³⁸.

Well-characterized epidemiologic risk factors for CRC include age, personal or family history of colorectal polyps or CRC, ethnicity, smoking, heavy alcohol use, high-fat diet, physical inactivity, and obesity³⁹. In addition to these established factors, inherited susceptibility is a major contributor to CRC risk and may help to explain the high incidence experienced by

founder populations such as Ashkenazi Jews and Newfoundlanders⁴⁰. In total, inherited genetic factors account for 32-35% of the variation in risk for developing CRC^{12,13}. However, only 2-6% of all CRC cases occur as part of well-characterized familial cancer syndromes driven by rare, highly penetrant, germline mutations^{13,41-43}. For example, the autosomal dominant condition Lynch Syndrome, formerly described as hereditary non-polyposis colorectal cancer, is caused by germline mutations in DNA mismatch repair genes (*MLH1*, *MSH2*, *MSH6*, *PMS2*, or *EPCAM*)^{44,45}. Further, the cause of familial adenomatous polyposis, an autosomal dominant syndrome characterized by excessive polyp formation in the colorectum and an exceptionally high risk of CRC, is a mutation in the adenomatous polyposis coli (*APC*) gene^{44,46}. Also, autosomal recessive conditions such as *MYH*-associated polyposis meaningfully contribute to the CRC burden.

Genetic risk factors contributing to the remaining familial and sporadic CRC cases have yet to be fully elucidated. Despite the relatively thorough characterization of genetic changes driving progression from normal epithelium to metastatic disease, it is unclear outside of the 2-6% syndromic cases which variants predispose individuals to initiate this disease course. Among the potential sources of missing heritability, the common disease common variant hypothesis proposes that a combination of many common, low- to moderate-penetrance susceptibility alleles account for a substantial portion of these unidentified genetic risk factors^{24,26,28-31}. Large populations and genetic tools that take an agnostic view of the genome offer an informative approach to address this hypothesis and to identify novel risk loci.

Traditional genome-wide association studies (GWAS) leverage linkage disequilibrium (LD) between haplotype-tagging and/or imputed single nucleotide polymorphisms (SNPs) and small insertions or deletions (indels) on a genotyping array with the goal of identifying

associations between common genetic variants (minor allele frequency (MAF) $\geq 5\%$) and binary disease status or quantitative traits²⁴. Several GWAS of CRC risk have identified over twenty low-penetrance susceptibility loci⁴⁷⁻⁶⁰. Some risk loci have been identified in genes contributing to pathways with existing links to CRC pathogenesis, such as the Wnt and transforming growth factor-beta (TGF- β) signaling pathways⁶¹. However, these studies have also identified a number of loci in novel regions. Although the functional significance and clinical relevance of most loci remain largely undetermined, several critical genes and regulatory regions have been well defined (Table 2.1). Similar to GWAS results from other complex diseases, the loci identified thus far explain only a small proportion of CRC's heritability and have limited utility for clinical risk prediction^{2,62-65}. Although some have argued that a large number of higher-effect rare variants are the primary genetic drivers of complex diseases⁶⁶, it is likely that variants of multiple frequencies and effect sizes contribute^{2,24,26,27}. Given the large number of risk loci successfully identified for breast and prostate cancer, we hypothesize that additional common, low penetrance loci for CRC remain that will require larger sample sizes or studies in specific populations to detect. Comprehensively characterizing these risk-conferring regions could offer new insights into the complex biology of CRC.

With the goal of identifying novel susceptibility loci, we conducted a GWAS meta-analysis of case-control studies of CRC in two populations: Ashkenazi Jews from the Molecular Epidemiology of Colorectal Cancer (MECC) study and non-Hispanic whites from the Colon Cancer Family Registry (CFR). Further, we replicated the strongest findings of genetic association from the meta-analysis in an independent set of Israeli MECC cases and controls to highlight variants most likely to influence the etiology of CRC in Ashkenazi Jews (AJ). Finally, we conducted a combined meta-analysis of the three contributing sample sets to maximize our

power for detecting associated genetic variants and to fine map a genome-wide significant finding. This study design provides the potential to identify risk alleles for future functional studies that are important across European sub-populations as well as genetic variants unique to the founder population of Ashkenazi Jewish individuals.

Table 2.1. Published genome-wide significant ($p < 5 \times 10^{-8}$) CRC risk loci. Aus = Australia.

SNP	CHR	BP	Locus (nearby gene)	Study population	Reference
rs6691170	1	222045446	1q41 (<i>DUSP10</i>)	Europe	Houlston et al 2010 ⁵⁶
rs6687758	1	222164948		Europe	Houlston et al 2010 ⁵⁶
rs11903757	2	192587204	2q32.3 (<i>NABP1</i>)	Europe, US, Canada, Aus, East Asia	Peters et al 2012a ⁵⁸
rs10936599	3	169492101	3q26.2 (<i>MYNN</i>)	Europe	Houlston et al 2010 ⁵⁶
rs647161	5	134499092	5q31.1 (<i>AK026965</i>)	East Asia, Europe	Jia et al 2013 ⁶⁷
rs1321311	6	36622900	6p21 (<i>CDKN1A</i>)	Europe, US, Canada, Aus, East Asia	Dunlop et al 2012 ⁵⁹
rs7758229	6	160840252	6q25.3 (<i>SLC22A3</i>)	East Asia	Cui et al 2011 ⁶⁸
rs16892766	8	117630683	8q23.3 (<i>EIF3H</i>)	Europe	Tomlinson et al 2008 ⁵⁴
rs10505477	8	128407443	8q24 (<i>MYC</i>)	Europe, Canada, Newfoundland, US, Israel	Zanke et al 2007 ⁴⁹ Gruber et al 2007 ⁴⁷
rs6983267	8	128413305		Europe, Canada, Newfoundland, US	Zanke et al 2007 ⁴⁹ Haiman et al 2007 ⁶⁹ Tomlinson et al 2007 ⁵¹ Hutter et al 2010 ⁷⁰
rs7014346	8	128424792		Europe, Canada, East Asia, Israel	Tenesa et al 2008 ⁵⁵
rs719725	9	6365683	9p24 (<i>TPD52L3</i>)	Europe, US	Tomlinson et al 2008 ⁵⁴ Kocarnik et al 2010 ⁷¹
rs10795668	10	8701219	10p14 (<i>LOC338591</i>)	Europe	Tomlinson et al 2008 ⁵⁴
rs3824999	11	74345550	11q13.4 (<i>POLD3</i>)	Europe, US, Canada, Aus East Asia	Dunlop et al 2012 ⁵⁹
rs3802842	11	111171709	11q23 (<i>C11orf93</i>)	Europe, Canada, East Asia, Ashkenazi Jews	Tenesa et al 2008 ⁵⁵
rs10774214	12	4368352	12p13.32 (<i>CCND2</i>)	East Asia, Europe	Jia et al 2013 ⁶⁷
rs7136702*	12	50880216	12q13.12 (<i>LARP4</i>)	Europe	Houlston et al 2010 ⁵⁶
rs11169552	12	51155663		Europe	Houlston et al 2010 ⁵⁶
rs4444235	14	54410919	14q22.2 (<i>BMP4</i>)	Europe, US, Canada, Aus	Houlston et al 2008 ⁵² Tomlinson et al 2011 ⁶⁰
rs1957636	14	54560018		Europe, US, Canada, Aus	Tomlinson et al 2011 ⁶⁰
rs16969681	15	32993111	15q13.3 (<i>CRAC1, HMPS, GREM1</i>)	Europe, US, Canada,	Tomlinson et al 2011 ⁶⁰
rs4779584	15	32994756		Europe, US, Canada, Aus	Jaeger et al 2008 ⁴⁸ Tomlinson et al 2011 ⁶⁰
rs11632715	15	33004247		Europe, US, Canada, Aus	Tomlinson et al 2011 ⁶⁰
rs9929218	16	68820946	16q22.1 (<i>CDHI</i>)	Europe	Houlston et al 2008 ⁵²
rs4939827	18	46453463	18q21 (<i>SMAD7</i>)	Europe, Canada, East Asia, Israel	Broderick et al 2007 ⁵⁰ Tenesa et al 2008 ⁵⁵
rs10411210	19	33532300	19q13.1 (<i>RHPN2</i>)	Europe	Houlston et al 2008 ⁵²
rs961253	20	6404281	20p12.3 (<i>BMP2</i>)	Europe, US, Canada, Aus	Houlston et al 2008 ⁵² Tomlinson et al 2011 ⁶⁰
rs4813802	20	6699595		Europe, US, Canada, Aus, Newfoundland, Israel	Tomlinson et al 2011 ⁶⁰ Peters et al 2012b ⁵⁷
rs4925386	20	60921044	20q13.33 (<i>LAMA5</i>)	Europe	Houlston et al 2010 ⁵⁶
rs5934683	X	9751474	Xp22.2 (<i>SHROOM2</i>)	Europe, US, Canada, Aus, East Asia	Dunlop et al 2012 ⁵⁹

* rs7136702 in italics is the only marker not measured or imputed with high quality in this meta-analysis.

2.2 Materials and methods

2.2.1 Study populations

The discovery meta-analysis was conducted using germline DNA from two separate case-control studies, the Molecular Epidemiology of Colorectal Cancer (MECC) study, and the Colon Cancer Family Registry (CFR). The replication dataset consisted of an independent set of MECC participants. Demographic and clinical characteristics of participants with high quality genotype data for both the discovery and replication study phases are summarized in Table 2.2.

Table 2.2. Demographic and clinical characteristics of genotyped participants after quality control from the Molecular Epidemiology of Colorectal Cancer (MECC; $n_{\text{discovery}} = 983$; $n_{\text{replication}} = 1,962$) and Colon Cancer Family Registry (CFR; $n_{\text{discovery}} = 2,977$) case-control studies.

	Discovery Meta-Analysis				Replication	
	MECC Illumina Omni2.5M		CFR - Illumina Human 1M, 1M-Duo, and Omni1-Quad		MECC Affymetrix Axiom 1.3M	
	Cases (n=485)	Controls (n=498)	Cases (n=1977)	Controls (n=999)	Cases (n=1131)	Controls (n=831)
Age [mean(sd)]	71.6 (10.1)	72.8 (10.2)	52.7 (11.1)	59.9 (11.0)	71.0 (11.0)	73.7 (11.1)
Sex (%)						
Male	265 (54.6)	264 (53.0)	983 (49.7)	478 (47.8)	563 (49.8)	420 (50.5)
Female	220 (45.4)	234 (47.0)	994 (50.3)	521 (52.2)	568 (50.2)	411 (49.5)
Ethnicity (%)						
Ashkenazi	470 (96.9)	493 (99.0)	0	0	1129 (99.8)	804 (96.8)
Sephardi	0	0	0	0	0	9 (1.1)
Ashkenazi/ non-Jewish	8 (1.6)	2 (0.4)	0	0	1 (0.1)	3 (0.4)
Ashkenazi/ Sephardi	1 (0.2)	2 (0.4)	0	0	1 (0.1)	12 (1.4)
Non-Jewish, non-Arab	6 (1.2)	1 (0.2)	0	0	0	0
Non-Hispanic White	0	0	1977 (100)	999 (100)	0	0
Missing	0	0	0	0	0	3 (0.4)
1st degree relative with CRC (%)						
Yes	65 (13.4)	41 (8.2)	-	-	135 (11.9)	71 (8.5)
No	419 (86.4)	457 (91.8)	-	999 (100)	985 (87.1)	757 (91.1)
Missing	1 (0.2)	0	-	-	11 (1.0)	3 (0.4)
Cancer site (%)						
Left colon	269 (55.5)	-	628 (31.8)	-	402 (35.5)	-
Right colon	181 (37.3)	-	577 (29.2)	-	343 (30.3)	-
Colon (Not Other- wise Specified)	0	-	43 (2.2)	-	0	-
Rectum	31 (6.4)	-	729 (36.9)	-	300 (26.5)	-
Other	4 (0.8)	-	0	-	51 (4.5)	-
Missing	0	-	0	-	35 (3.1)	-

2.2.1.1 Molecular Epidemiology of Colorectal Cancer Study (MECC)

The Molecular Epidemiology of Colorectal Cancer Study (MECC) is a population-based, case-control study of pathologically-confirmed, incident cases of CRC recruited from a geographically-defined region of northern Israel⁷². Participant recruitment began in 1998 and remains on-going. Individually-matched controls with no prior history of CRC are selected from the same source population that gave rise to cases using the Clalit Health Services database. Matching factors include age, sex, Jewish ethnicity (Jew versus non-Jew), and primary clinic site. Subjects are interviewed to obtain demographic and clinical information, family history, and dietary habits. Biospecimens including blood, paraffin blocks, and snap frozen tumors are also collected. The discovery phase of this GWAS meta-analysis leverages data for 485 AJ cases and 498 AJ controls from MECC. Case selection for genotyping in this phase enriched for colon cancer (not rectal cancer), enriched for a specific stage distribution for a separate GWAS study of stage and prognosis, and excluded cases with microsatellite instable (MSI-H) tumors. The replication stage utilizes MECC genotypes from 1,131 AJ cases and 831 AJ controls. Replication stage cases were unselected for cancer site, stage, or MSI (Table 2.2).

2.2.1.2 Colon Cancer Family Registry (CFR)

The Colon CFR is a consortium of six centers across North America and Australia, organized to create a comprehensive resource for clinical and epidemiologic studies of CRC^{57,73}. The six centers include the University of Southern California (USC), Fred Hutchinson Cancer Research Center (FHCRC), Mayo Clinic, Cancer Care Ontario, Hawaii Cancer Registry, and the University of Melbourne. The registry contains family history, clinical history, and risk factor data on approximately 37,000 subjects (including nearly 10,000 probands and 27,000 affected or unaffected relatives and unrelated controls). Phase I (1998-2002) focused on recruitment of

incident cases of CRC via population-based cancer registries or clinical centers as well as general population or proband-identified controls. Phase II recruitment (2002-2007) included incident probands with CRC diagnosed under age 50 and additional clinic-identified families. As described previously, those selected from Phase I for the study's initial genome-wide scan (Set 1) were population-based cases and age- and sex-matched controls from the following three centers: FHCRC, Cancer Care Ontario, and Melbourne⁷⁴. Phase II participants selected for genotyping (Set 2) included population-based cases from all six study sites and same-generation family controls. However, only Phase II cases are included in this analysis, and family-based controls were excluded. Case selection for genotyping from both Sets 1 and 2 was enriched for age at onset prior to 50 or family history of CRC. All subjects self-reported as being of non-Hispanic white race, and this was verified using genotype data. Further, controls reported no family or personal CRC history. In total, the discovery GWAS meta-analysis uses 1,977 population-based cases from Phases I and II and 999 age- and sex-matched controls from Phase I.

2.2.2 Genotyping and quality control

2.2.2.1 Discovery meta-analysis

Germline DNA was extracted from peripheral blood samples for both MECC and CFR participants. MECC DNA was genotyped in two batches using the Illumina HumanOmni 2.5S-v1 BeadChip, which measures nearly 2.4 million SNPs. Batch 1 (414 cases and 155 controls) was run at Case Western Reserve University (CWRU) for the purpose of a GWAS study of stage and prognosis, and batch 2 (104 cases and 376 controls) was run at the University of Michigan (UM) to create a balanced design of cases and matched controls. Colon CFR samples were genotyped across three platforms due to chip availability: the Illumina Human 1M or 1M-Duo

(CFR Set 1), and the Omni1-Quad (CFR Set 2 controls), each containing approximately 1.2 million common loci.

MECC genotype data were cleaned based on quality control (QC) metrics at the individual subject and SNP levels (Figure 2.1). Samples with >5% missing genotypes, sex mismatches (between self-reported and genotypic predicted sex), duplicate samples, and those with excess homozygosity were identified and subsequently removed. SNPs with <95% call rate were excluded, and those inconsistent with Hardy Weinberg Equilibrium (HWE) in controls were flagged for individual review but not removed. Principal components analysis (PCA) was conducted on approximately 100,000 randomly selected markers (after LD-pruning) using the `pcaMethods` Bioconductor package⁷⁵ in R to identify ethnic outliers. These same principal components were retained to adjust for confounding due to population stratification. Pairwise plots of principal components (PCs) 1-3 on the final analysis dataset are in Figure 2.2.

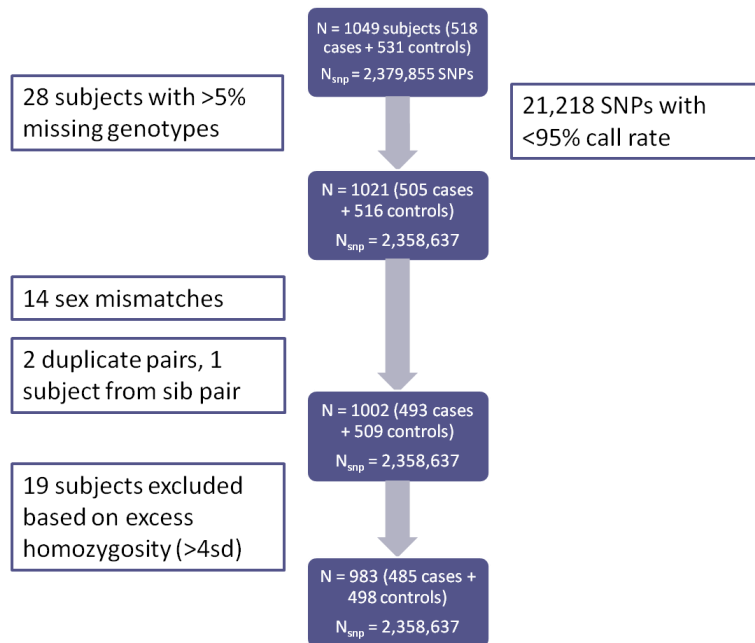


Figure 2.1. Quality control and filtering pipeline for discovery MECC samples prior to imputation.

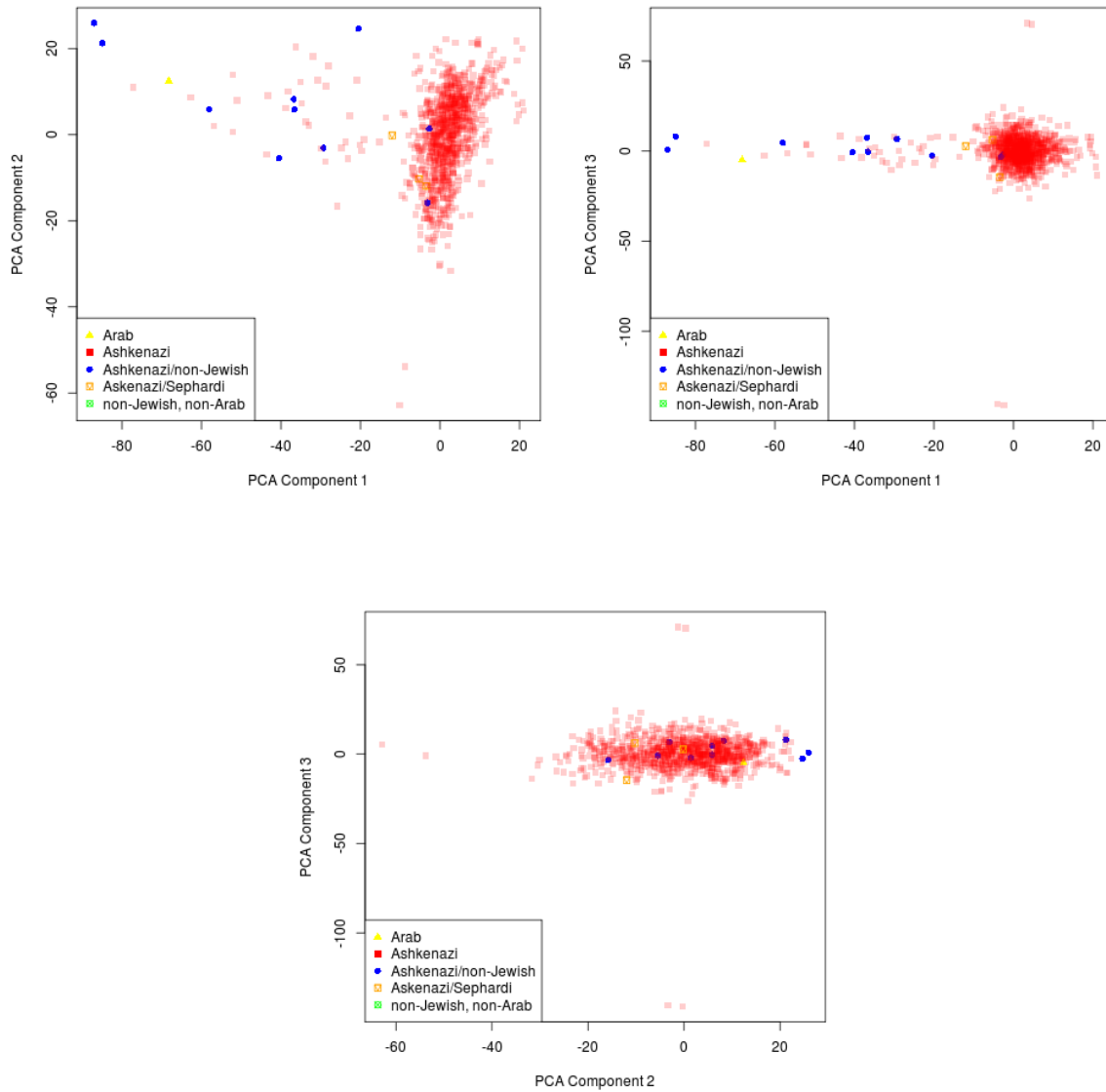


Figure 2.2. Pairwise plots of principal components from the discovery phase MECC Omni samples. Legend shows self-reported ethnicity.

CFR genotype data was cleaned using comparable parameters, and the methods applied for CFR Set 1 genotypes have been detailed elsewhere⁷⁴. Briefly, the QC criteria excluded samples based on call rate, sex mismatches, unintended duplicates, lack of concordance with

previous genotype data, and unanticipated genotype concordance or identity-by-descent (IBD) with another sample. PCA based on a panel of ancestry information markers (AIMs) was conducted using Eigenstrat⁷⁶, followed by comparison to HapMap II CEU participants from Utah, to identify and exclude ethnic outliers. SNP level exclusions were made for markers with more than two alleles, no “rs” identification number, poor genotype concordance in the same individuals across platforms, and low call rate (<90%). Pairwise plots of PCs 1-4 on the final analysis dataset are in Figure 2.3.

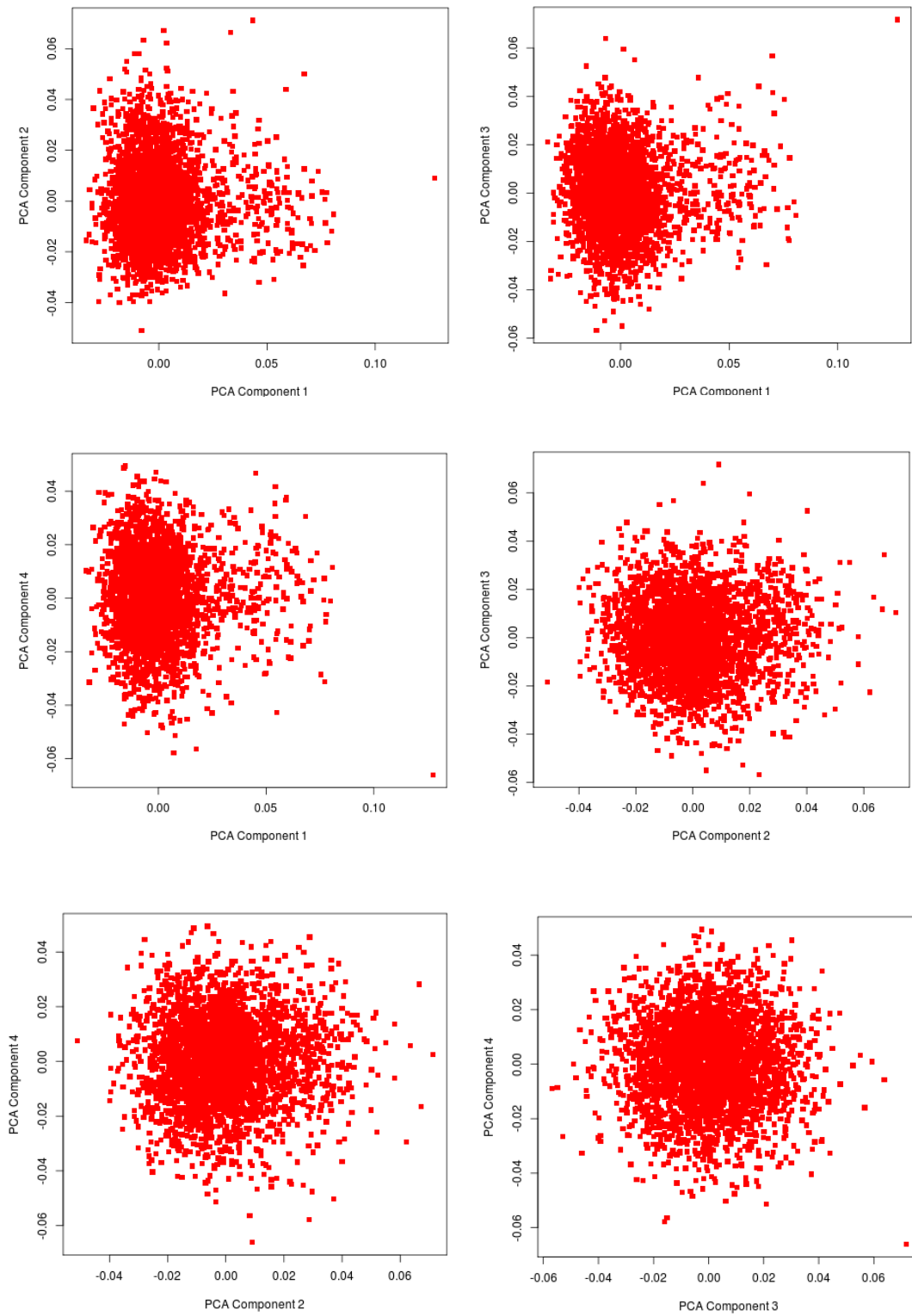


Figure 2.3. Pairwise plots of principal components from the discovery phase CFR samples. All participants self-reported as non-Hispanic White.

2.2.2.2 Replication analysis

An independent set of MECC germline DNA samples were genotyped as part of the National Cancer Institute (NCI)-sponsored Colorectal Transdisciplinary (CORECT) Study Consortium and served as the replication dataset for this study. These MECC participants had similar distributions of key demographic and clinical covariates to the discovery stage MECC subjects with the exception of cancer site and MSI (Table 2.2). Genotyping was performed via hybridization to a custom Affymetrix genome-wide platform (the Axiom[®] CORECT Set) containing approximately 1.3 million SNPs and indels spread across two physical genotyping chips (pegs). The quality control and filtering pipeline resulted in a final analysis dataset containing 1131 cases, 831 controls, and 1,230,678 genetic markers (Figure 2.4). PCA was conducted using Eigenstrat⁷⁶ based on a set of 2,884 AIMs derived from the literature and the Illumina Infinium HumanExome BeadChip and Affymetrix Axiom[®] Exome Array. Pairwise plots of PCs 1 and 2 on the final analysis dataset are in Figure 2.5.

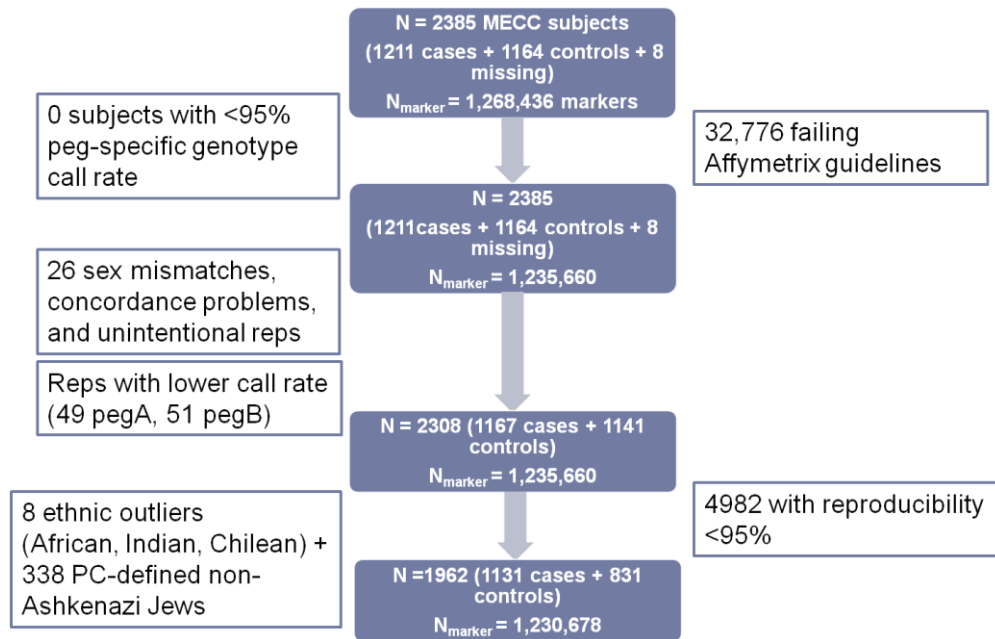


Figure 2.4. Quality control and filtering pipeline for MECC Axiom[®] replication samples prior to imputation. PC = principal component.

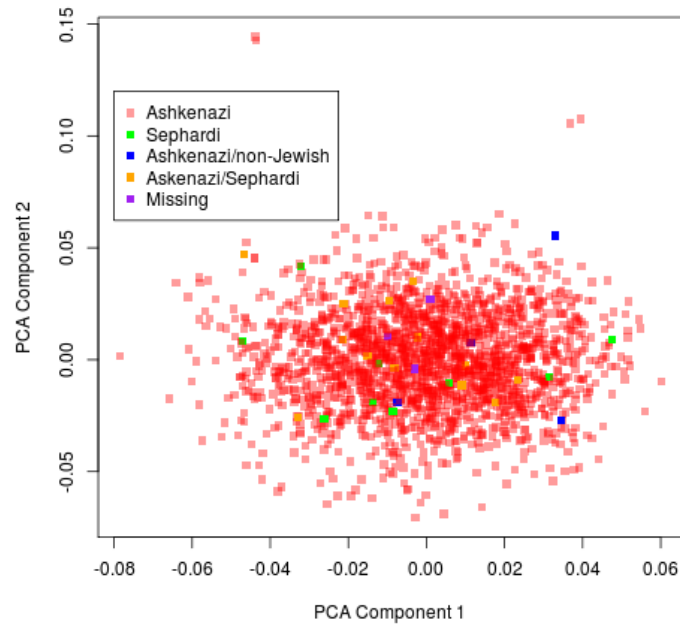


Figure 2.5. PC1 versus PC2 plot from the replication phase MECC Axiom samples. Legend shows self-reported ethnicity.

2.2.3 Imputation

To analyze genotype data generated from three different platforms that measure different genetic markers and to increase the coverage of variation that is measurable across the genome, imputation of genotypes was performed for both autosomal and X chromosome markers. First, genotypes were pre-phased into best-guess haplotypes with SHAPEIT (SHAPEIT.v1.ESHG) to increase the computational efficiency of downstream steps^{77,78}. Default parameters were applied with the exception of effective-size=11,418. Then, IMPUTE2 (IMPUTE v2.2.2) was used to impute missing genotypes for study samples based on the cosmopolitan panel of reference haplotypes from Phase I of the 1000 Genomes Project (March 2012 release; n = 1092)⁷⁹. Among the panel's ethnically diverse samples, it includes 500 samples from 5 different Caucasian populations: Utah residents with Northern and Western European ancestry, Toscani in Italy, British from Scotland and England, Finnish from Finland, and Iberians from Spain. The release

contains over 39 million autosome and X chromosome SNPs, indels, and structural variants. Default parameters for IMPUTE2 were applied except for $k_{\text{hap}}=1038$ and $\text{buffer}=500$. In order to enter subsequent statistical analysis steps, genetic markers resulting from the imputation had to pass stringent imputation quality and accuracy filters ($\text{info} \geq 0.7$, $\text{certainty} \geq 0.9$, and $\text{concordance} \geq 0.9$ between directly measured and imputed genotypes after masking input genotypes (for genotyped markers only)). Further, we restricted our SNP list to those with study-specific $\text{MAF} \geq 1\%$. Following imputation and QC, 9,009,669, 8,304,060, and 9,177,523 SNPs were available for analysis in the MECC discovery, CFR discovery, and MECC replication, respectively.

2.2.4 Statistical analysis

2.2.4.1 Discovery GWAS meta-analysis

The discovery MECC and CFR datasets were first analyzed in a study-specific fashion, allowing for adjustment for appropriate covariates. Then, study-specific results were analyzed using an inverse-variance-weighted, fixed-effects meta-analysis which assumed homogeneity of effects across the two studies. In each study, to examine the association between each variant and CRC risk, we specified a log-additive genetic model, where each additional copy of the minor allele was assumed to confer the same magnitude of risk or protection. Each SNP was coded as a dosage, the expected number of effect alleles. We calculated beta coefficients and corresponding odds ratios (OR), standard errors, 95% confidence intervals, and p-values using unconditional logistic regression. For MECC, models were adjusted for age, sex, 3 PCs, and genotyping batch (in the event that differential genotyping errors occurred across the two genotyping centers). For CFR, we adjusted for age, sex, 4 PCs, and recruitment site (in 3 groups). For both studies, higher PCs were primarily driven by outliers, so adjusting for more PCs was deemed unnecessary.

These models were used to examine the odds ratio for CRC risk associated with each additional copy of the minor allele (or minor allele dosage) for a given SNP, after adjusting for all covariates in the model.

For markers with matching identification names and/or chromosome, base pair, and alleles across the two studies, we conducted a meta-analysis of point estimates and standard errors using an inverse-variance-weighted, fixed-effects approach. This method, implemented in METAL, calculates a summary estimate of effect for each marker through summation of inverse-variance weighted betas across studies, divided by the summed weights⁸⁰. A quantile-quantile (Q-Q) plot was generated to examine the distribution of meta-analysis p-values compared to the distribution under null expectations. The genomic control lambda (GC λ) associated with the observed p-value distribution was calculated to identify evidence of p-value inflation, likely due to population stratification. GC λ is the median observed squared Z statistic divided by the median expected squared Z (0.455) based on a null X^2 with 1 degree of freedom. A Manhattan plot showing p-values sorted by chromosomal position was generated to provide a visual illustration of top hit regions across the genome. SNPs with nominally significant p-values ($p < 0.05$) were selected for replication in the second phase of this study ($n_{\text{marker}} = 492,866$). Further, to ensure that our genotyping and subsequent analysis was able to detect previously published risk loci, we examined association results for 29 available out of 30 total SNPs from 22 independent loci known to be associated with CRC from previous studies (Table 2.1).

Statistical analysis and plotting were conducted using a combination of PLINK v1.07⁸¹, R v2.15.2, and METAL⁸⁰. The p-value criterion for genome-wide statistical significance of SNP coefficients was set at 5×10^{-8} .⁸² However, those SNPs with p-values less than 5×10^{-7} but larger than 5×10^{-8} were considered to have suggestive associations warranting follow-up.

2.2.5 Replication in MECC and joint meta-analysis

To replicate our discovery meta-analysis findings, we used the same logistic regression analysis methods described above to examine the marginal association between each SNP with $p < 0.05$ from the MECC+CFR discovery meta-analysis and CRC status in an independent set of MECC samples. Models were adjusted for age, sex, and 2 PCs. As for the MECC discovery set, higher PCs were outlier-driven, so additional adjustment beyond 2 PCs was not conducted. Height was also evaluated as a potential confounder given prior data for one of the SNPs that arose in our analyses, and was considered after adjusting for these four covariates using a change-in-estimate criterion approach⁸³. Because it has been demonstrated that joint analysis of two-stage GWAS designs is more efficient than replication-based analysis⁸⁴, we also conducted a fixed-effects meta-analysis of results with $p < 0.05$ from the discovery MECC/CFR meta-analysis with the results from the MECC-based replication. Subsequently, for a region identified as a novel, genome-wide significant susceptibility locus, we removed the $p < 0.05$ discovery filter to explore associations in the region in more depth through fine mapping. Regional plots of association results near the genome-wide significant finding were generated using LocusZoom with LD based on the 1000 Genomes March 2012 European population⁸⁵.

2.3 Results

2.3.1 Discovery meta-analysis

Pairwise plots of the first 3 eigenvalues for MECC and first 4 eigenvalues for CFR from PCA demonstrated that adjustment for the first 3 and 4 PCs, respectively, captured the majority of genotypic variability clearly attributable to fine-scale population substructure (Figure 2.2; Figure 2.3). Thus, the first 3 and first 4 PCs were selected as covariates for adjustment in MECC and CFR logistic regression modeling, respectively.

The Q-Q plot of p-values from the fixed-effects meta-analysis of study-specific, adjusted logistic regression model results shows the rank-ordered observed $-\log_{10}(\text{p-value})$ plotted against the rank-ordered expected $-\log_{10}(\text{p-value})$ (Figure 2.6). The p-values above the diagonal line in the upper-right quadrant of the plot demonstrate that there are a number of SNPs with associations more statistically significant than expected by chance alone, assuming a uniform distribution of p-values. Further, the associated GC λ value of 1.033 suggests that the selected covariates and PCs provide reasonable control for population stratification. The Manhattan plot displays summary results from the meta-analysis by chromosomal position and highlights a peak on chromosome 4 with six loci in tight linkage disequilibrium reaching genome-wide significance (Figure 2.7).

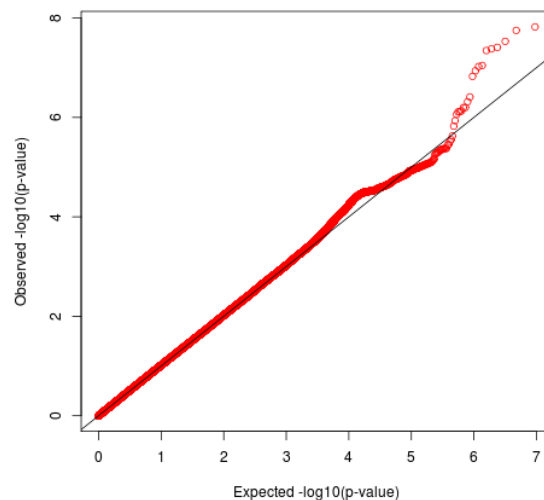


Figure 2.6. Q-Q plot of p-values derived from the inverse-variance-weighted, fixed-effects meta-analysis of adjusted logistic regression model results run on MECC Omni + CFR Illumina discovery samples. The MECC analyses were adjusted for age, sex, batch, and 3 PCs. The CFR analyses were adjusted for age, sex, recruitment site, and 4 PCs. 492,866 markers had $p < 0.05$.

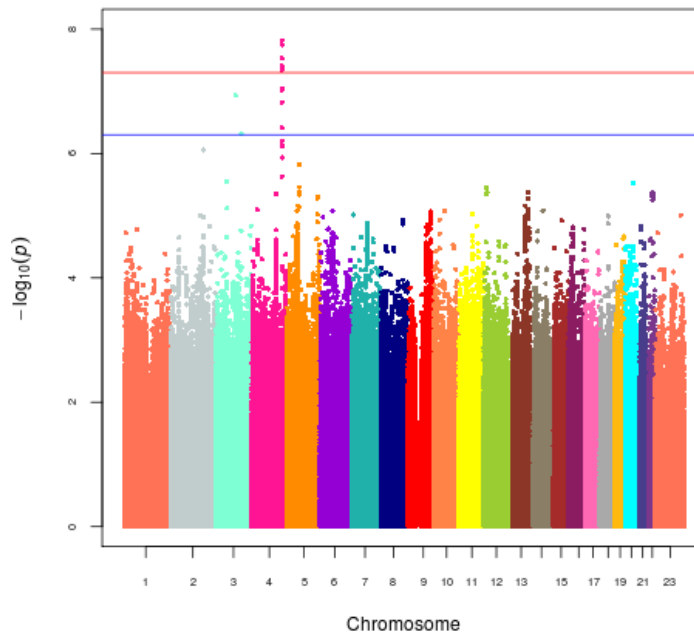


Figure 2.7. Manhattan plot of $-\log_{10}(p\text{-values})$ from the inverse-variance-weighted, fixed-effects meta-analysis of discovery stage MECC and CFR adjusted association results. Each circle represents the $-\log_{10}(p\text{-value})$ for one of 9,516,354 SNPs plotted against its chromosomal location on the x-axis. Blue line = suggestive line for genome-wide significance at 5×10^{-7} . Red line = genome-wide significance threshold at 5×10^{-8} .

rs17042479 was the SNP in the chromosome 4 genome-wide significant region with the most significant p-value (risk allele = G; OR per risk allele = 1.67; p-value = 1.5×10^{-8}). The directions of effect for MECC and CFR were consistent, with CFR exhibiting a slightly attenuated effect. Study-specific estimates demonstrate that the result was not heavily driven by either MECC or CFR findings, and the average minor allele frequency across studies was approximately 10%. This SNP was directly measured in the MECC discovery samples, CFR Set1, and MECC replication samples (imputed only in CFR Set 2). The SNP falls about 250 Kilobases (Kb) upstream of the ~800 Kb gene *FSTL5* (follistatin-like 5) and approximately 720 Kb downstream of *NAF1* (nuclear assembly factor 1 homolog (*S. cerevisiae*)). Two other loci on

chromosome 3 reached our suggestive threshold for genome-wide significance of $p < 5 \times 10^{-7}$; however, the lack of SNPs in LD with trailing p-values (i.e. attenuated association p-values for nearby SNPs as LD with the index SNP decreases) suggests that they may be false positive results. The complete list of most significant association findings ($p \leq 3 \times 10^{-6}$) between effect allele dosage and CRC status, visually indicated by the inflated tail of observed $-\log_{10}(p\text{-values})$ in the Q-Q plot (Figure 2.6) and as SNPs above the blue line in the Manhattan plot (Figure 2.7), are summarized in Table 2.3. Further, we also demonstrated that 14 out of 29 previously identified CRC risk alleles that were imputed with high quality and analyzed in this meta-analysis had nominally significant associations with $p < 0.05$ (* SE = **standard error of the beta effect estimate**

Table 2.4). Twenty-six out of 29 known susceptibility markers had a consistent risk allele and direction of effect with the previously published result. The most statistically significant risk locus was located at chromosomal region 8q24, as previously described from the same source population⁴⁷.

Table 2.3. Summary of SNPs associated with CRC ($p < 3 \times 10^{-6}$) in the discovery MECC + CFR meta-analysis. Orientation in parentheses refers to the SNP position relative to the nearest gene.

SNP	CHR	BP	Effect Allele	Avg Freq	Nearest Gene	MECC Omni OR	MECC Omni SE*	MECC Omni P	CFR OR	CFR SE*	CFR Illumina P	Meta OR	Meta SE*	Meta P
rs17042479	4	163325411	g	0.10	<i>FSTL5</i> (upstream)	2.08	0.22	6.0E-04	1.59	0.10	3.4E-06	1.67	0.09	1.5E-08
163325957:D	4	163325957	c	0.10	<i>FSTL5</i> (upstream)	2.08	0.22	6.0E-04	1.59	0.10	3.9E-06	1.67	0.09	1.8E-08
rs35509282	4	163333405	a	0.10	<i>FSTL5</i> (upstream)	1.89	0.22	3.7E-03	1.61	0.10	1.8E-06	1.66	0.09	3.0E-08
rs9998942	4	163340404	t	0.10	<i>FSTL5</i> (upstream)	1.82	0.22	5.3E-03	1.61	0.10	1.9E-06	1.65	0.09	3.9E-08
rs57336275	4	163341215	c	0.10	<i>FSTL5</i> (upstream)	1.82	0.21	5.5E-03	1.61	0.10	2.0E-06	1.64	0.09	4.2E-08
rs11736440	4	163336693	a	0.10	<i>FSTL5</i> (upstream)	1.85	0.22	4.9E-03	1.61	0.10	2.3E-06	1.65	0.09	4.5E-08
rs11100443	4	163337191	t	0.10	<i>FSTL5</i> (upstream)	1.72	0.21	1.1E-02	1.61	0.10	2.5E-06	1.62	0.09	9.0E-08
rs2122494	4	163331379	g	0.09	<i>FSTL5</i> (upstream)	1.72	0.21	1.0E-02	1.61	0.10	2.8E-06	1.64	0.09	9.5E-08
rs56856071	3	108949140	g	0.90	<i>LINC00488</i> (downstream)	1.49	0.18	2.6E-02	1.67	0.11	1.3E-06	1.61	0.09	1.2E-07
163338255:D	4	163338255	g	0.10	<i>FSTL5</i> (upstream)	1.72	0.21	1.1E-02	1.59	0.10	4.1E-06	1.61	0.09	1.5E-07
rs12508784	4	163333299	a	0.07	<i>FSTL5</i> (upstream)	2.22	0.26	2.0E-03	1.64	0.12	3.0E-05	1.72	0.11	3.9E-07
rs13063991	3	141293755	t	0.92	<i>RASA2</i> (intron)	1.90	0.25	1.1E-02	1.57	0.10	1.1E-05	1.61	0.10	4.8E-07
rs12511058	4	163326723	g	0.07	<i>FSTL5</i> (upstream)	2.22	0.26	2.1E-03	1.61	0.12	4.5E-05	1.69	0.11	6.3E-07
rs17600575	4	163329336	c	0.07	<i>FSTL5</i> (upstream)	2.22	0.26	2.1E-03	1.61	0.12	4.6E-05	1.69	0.11	6.3E-07
rs12650100	4	163330891	c	0.07	<i>FSTL5</i> (upstream)	2.27	0.26	1.9E-03	1.61	0.12	5.5E-05	1.69	0.11	7.4E-07
rs12642547	4	163337313	a	0.07	<i>FSTL5</i> (upstream)	2.08	0.25	3.6E-03	1.61	0.12	4.1E-05	1.69	0.11	7.7E-07
rs12645341	4	163337355	g	0.07	<i>FSTL5</i> (upstream)	2.08	0.25	3.6E-03	1.61	0.12	4.1E-05	1.69	0.11	7.8E-07
rs147150124	2	178654517	a	0.23	<i>PDE11A</i> (intron)	1.27	0.14	9.1E-02	1.45	0.08	2.7E-06	1.40	0.07	8.7E-07
rs59363334	4	163340796	c	0.07	<i>FSTL5</i> (upstream)	2.04	0.25	4.4E-03	1.59	0.12	5.5E-05	1.67	0.11	1.2E-06
rs115742074	5	67627365	a	0.02	<i>PIK3R1</i> (downstream)	0.27	0.32	4.9E-05	0.51	0.22	2.4E-03	0.42	0.18	1.5E-06
rs11100440	4	163324864	t	0.06	<i>FSTL5</i> (upstream)	2.44	0.29	2.1E-03	1.61	0.12	1.3E-04	1.71	0.11	2.4E-06
rs4688168	3	63439404	a	0.96	<i>SYNPR</i> (intron)	2.07	0.23	1.7E-03	1.85	0.18	4.7E-04	1.93	0.14	2.8E-06

* SE = standard error of the beta effect estimate

Table 2.4. Discovery MECC + CFR meta-analysis association results for previously identified risk alleles. SNPs in bold have nominally significant associations with CRC ($p < 0.05$).

SNP	CHR	BP	Locus	A1*	A2	Freq	OR	SE**	P	Published risk allele
rs6691170	1	222045446	1q41 (<i>DUSP10</i>)	t	g	0.37	0.99	0.05	0.84	t
rs6687758	1	222164948		a	g	0.79	1.08	0.06	0.19	g
rs11903757	2	192587204	2q32.3 (<i>NABP1</i>)	t	c	0.85	0.82	0.08	0.0097	c
rs10936599	3	169492101	3q26.2 (<i>MYNN</i>)	t	c	0.23	0.95	0.06	0.42	c
rs647161	5	134499092	5q31.1 (<i>AK026965</i>)	a	c	0.69	1.11	0.06	0.070	a
rs1321311	6	36622900	6p21 (<i>CDKN1A</i>)	a	c	0.28	1.08	0.06	0.19	a
rs7758229	6	160840252	6q25.3 (<i>SLC22A3</i>)	t	g	0.31	1.01	0.06	0.84	t
rs16892766	8	117630683	8q23.3 (<i>EIF3H</i>)	a	c	0.89	0.79	0.08	0.0052	c
rs10505477	8	128407443	8q24 (<i>MYC</i>)	a	g	0.54	1.21	0.05	0.0003	t
rs6983267	8	128413305		t	g	0.45	0.84	0.05	0.0006	g
rs7014346	8	128424792		a	g	0.40	1.14	0.05	0.012	a
rs719725	9	6365683	9p24 (<i>TPD52L3</i>)	a	c	0.62	1.17	0.05	0.0028	a
rs10795668	10	8701219	10p14 (<i>LOC338591</i>)	a	g	0.30	0.87	0.06	0.012	g
rs3824999	11	74345550	11q13.4 (<i>POLD3</i>)	t	g	0.45	0.86	0.05	0.0040	g
rs3802842	11	111171709	11q23 (<i>C11orf93</i>)	a	c	0.70	0.86	0.06	0.0068	c
rs10774214	12	4368352	12p13.32 (<i>CCND2</i>)	t	c	0.38	1.08	0.05	0.14	t
rs11169552	12	51155663	12q13.12 (<i>LARP4</i>)	t	c	0.27	0.86	0.06	0.0081	c
rs4444235	14	54410919	14q22.2 (<i>BMP4</i>)	t	c	0.50	0.97	0.05	0.50	c
rs1957636	14	54560018		t	c	0.41	1.08	0.05	0.12	t
rs16969681	15	32993111	15q13.3 (<i>CRAC1, HMPS, GREM1</i>)	t	c	0.10	1.17	0.09	0.082	t
rs4779584	15	32994756		t	c	0.20	1.12	0.07	0.086	t
rs11632715	15	33004247		a	g	0.48	1.05	0.05	0.34	a
rs9929218	16	68820946	16q22.1 (<i>CDH1</i>)	a	g	0.31	1.01	0.06	0.92	g
rs4939827	18	46453463	18q21 (<i>SMAD7</i>)	t	c	0.56	1.19	0.05	0.0010	t
rs10411210	19	33532300	19q13.1 (<i>RHPN2</i>)	t	c	0.12	0.91	0.08	0.23	c
rs961253	20	6404281	20p12.3 (<i>BMP2</i>)	a	c	0.36	1.13	0.05	0.025	a
rs4813802	20	6699595		t	g	0.66	0.88	0.05	0.014	g
rs4925386	20	60921044	20q13.33 (<i>LAMA5</i>)	t	c	0.29	0.86	0.06	0.0089	c
rs5934683	X	9751474	Xp22.2 (<i>SHROOM2</i>)	t	c	0.41	1.06	0.09	0.54	a

*A1 = effect allele for the corresponding OR; ** SE = standard error of the beta estimate

2.3.2 Replication

Genetic markers from the previously conducted MECC+CFR meta-analysis with $p < 0.05$ were carried forward into this stage ($n_{\text{marker}} = 492,866$). The combined meta-analysis of MECC discovery and CFR discovery samples together with MECC replication samples demonstrated that the region on chromosome 4q32.2 remains statistically significant at a genome-wide threshold. In the combined analysis, rs35509282 was the most strongly associated meta-analysis finding (risk allele = A; OR per risk allele = 1.54; $p\text{-value} = 8.2 \times 10^{-9}$), with the MECC replication-specific result consistent in direction with a $p\text{-value}$ of 0.033 (Table 2.5). All findings with $p < 5 \times 10^{-7}$ except for one fell within this region, and the OR estimates and average allele frequencies indicate strong LD among all chromosome 4 top hits. Height was evaluated as a potential confounder in the MECC replication samples for the top 16 hits in the intergenic *FSTL5|NAF1* region (Table 2.6). Height was not associated with CRC (Figure 2.8; $t=0.78$; $p=0.44$), and none of the genotypes were associated with height by linear regression ($p\text{-value}$ range: 0.56-1.00). Applying a change-in-estimate criterion for ORs of 10%, we showed that adjustment for height as a potential confounder in genome-wide logistic regression models above and beyond age, sex, and 2 PCs was unnecessary (Table 2.6). One marker on chromosome 18, a TA to T deletion, also reached a suggestive $p\text{-value}$ of 1.4×10^{-7} . However, this finding is not novel as it falls in the previously identified 18q21 locus (intronic region of *SMAD7*)^{50,55}.

Table 2.5. Top hits from MECC discovery (Omni) + CFR discovery + MECC replication (Axiom) meta-analysis.

SNP	CHR	BP	Avg Freq	Effect Allele	Alt Allele	MECC Omni OR	MECC Omni P	CFR OR	CFR P	MECC Axiom OR	MECC Axiom SE	MECC Axiom P	Meta OR	Meta SE*	Meta P
rs35509282	4	163333405	0.09	T	A	1.89	3.7E-03	1.61	1.8E-06	1.32	0.13	3.3E-02	1.53	0.07	8.2E-09
rs11736440	4	163336693	0.09	G	A	1.85	4.9E-03	1.61	2.3E-06	1.33	0.13	2.5E-02	1.53	0.07	8.3E-09
rs9998942	4	163340404	0.09	C	T	1.82	5.3E-03	1.61	1.9E-06	1.32	0.13	3.1E-02	1.53	0.07	9.7E-09
rs57336275	4	163341215	0.91	T	C	0.55	5.5E-03	0.62	2.0E-06	0.76	0.13	3.5E-02	0.65	0.07	1.9E-08
rs17042479	4	163325411	0.91	A	G	0.48	6.0E-04	0.63	3.4E-06	0.80	0.12	6.6E-02	0.66	0.07	1.7E-08
chr4:163325957:D	4	163325957	0.91	CAT	C	0.48	6.0E-04	0.63	3.9E-06	0.80	0.12	6.8E-02	0.66	0.07	2.0E-08
rs12508784	4	163333299	0.06	G	A	2.22	2.0E-03	1.64	3.0E-05	1.45	0.16	1.8E-02	1.64	0.09	3.3E-08
rs11100443	4	163337191	0.09	C	T	1.72	1.1E-02	1.61	2.5E-06	1.28	0.13	4.9E-02	1.49	0.07	4.0E-08
rs12511058	4	163326723	0.94	T	G	0.45	2.1E-03	0.62	4.5E-05	0.69	0.16	1.7E-02	0.62	0.09	4.7E-08
rs17600575	4	163329336	0.94	T	C	0.45	2.1E-03	0.62	4.6E-05	0.69	0.16	1.7E-02	0.62	0.09	4.8E-08
chr4:163338255:D	4	163338255	0.09	GC	G	1.72	1.1E-02	1.59	4.1E-06	1.30	0.13	4.3E-02	1.49	0.07	5.0E-08
rs12645341	4	163337355	0.94	A	G	0.48	3.6E-03	0.62	4.1E-05	0.70	0.15	1.8E-02	0.63	0.09	6.7E-08
rs12650100	4	163330891	0.06	G	C	2.27	1.9E-03	1.61	5.5E-05	1.41	0.15	2.6E-02	1.59	0.09	9.3E-08
rs12642547	4	163337313	0.06	T	A	2.08	3.6E-03	1.61	4.1E-05	1.41	0.15	2.5E-02	1.59	0.09	9.3E-08
rs2122494	4	163331379	0.91	A	G	0.58	1.0E-02	0.62	2.8E-06	0.81	0.13	8.9E-02	0.67	0.08	1.1E-07
chr18:46455468:D	18	46455468	0.43	TA	T	1.10	4.1E-01	1.28	4.5E-05	1.27	0.07	5.4E-04	1.25	0.04	1.4E-07
rs59363334	4	163340796	0.94	T	C	0.49	4.4E-03	0.63	5.5E-05	0.72	0.15	3.0E-02	0.64	0.09	1.7E-07

*SE = standard error of the beta estimate

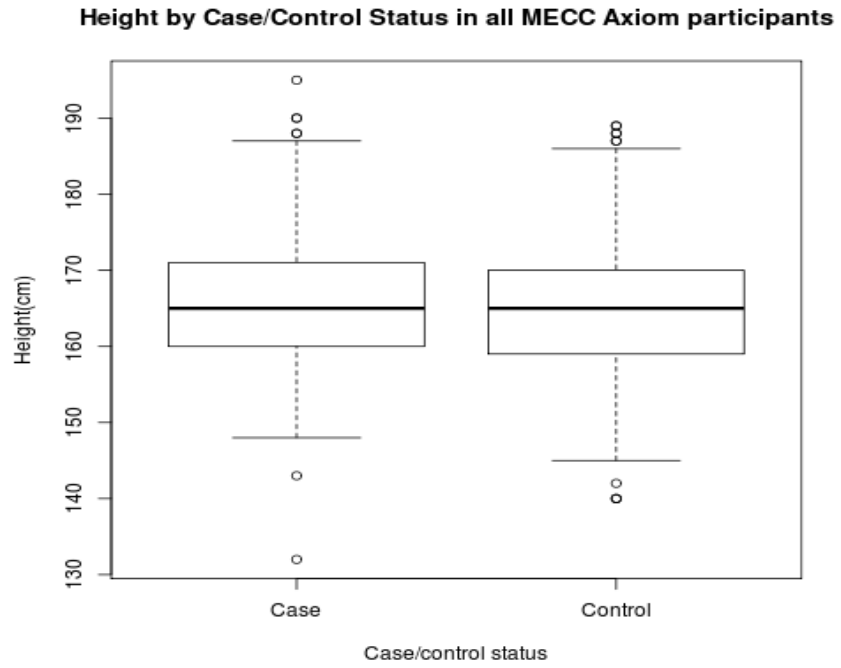


Figure 2.8. Height by case/control status in all MECC Axiom participants.

Table 2.6. Assessment of height as a potential confounder of the most significant SNP-CRC associations from the joint meta-analysis.

SNP	A1	A2	FRQ	OR*	SE*	P*	OR**	SE**	P**	% change in OR
rs17042479	A	G	0.92	0.80	0.12	0.06	0.78	0.13	0.05	-2.46%
chr4:163325957:D	CAT	C	0.92	0.80	0.12	0.07	0.78	0.13	0.06	-2.39%
rs12511058	T	G	0.95	0.69	0.16	0.02	0.68	0.17	0.02	-0.48%
rs17600575	T	C	0.95	0.69	0.16	0.02	0.68	0.17	0.02	-0.54%
rs12650100	G	C	0.95	0.71	0.16	0.03	0.71	0.17	0.04	0.11%
rs2122494	A	G	0.93	0.80	0.13	0.09	0.79	0.14	0.08	-1.77%
rs12508784	G	A	0.95	0.69	0.16	0.02	0.68	0.17	0.02	-1.16%
rs35509282	T	A	0.93	0.75	0.13	0.03	0.74	0.14	0.03	-2.12%
rs11736440	G	A	0.93	0.74	0.13	0.02	0.73	0.14	0.02	-1.92%
rs11100443	C	T	0.93	0.78	0.13	0.05	0.76	0.14	0.05	-1.75%
rs12642547	T	A	0.95	0.71	0.15	0.02	0.70	0.16	0.03	-1.16%
rs12645341	A	G	0.95	0.70	0.15	0.02	0.69	0.16	0.02	-0.85%
chr4:163338255:D	GC	G	0.93	0.77	0.13	0.04	0.75	0.14	0.04	-2.08%
rs9998942	C	T	0.93	0.75	0.13	0.03	0.74	0.14	0.03	-2.02%
rs59363334	T	C	0.95	0.72	0.15	0.03	0.71	0.16	0.03	-1.06%
rs57336275	T	C	0.93	0.76	0.13	0.03	0.74	0.14	0.03	-1.98%

* Adjusted for age, sex, 2 PCs; ** Adjusted for age, sex, 2 PCs, height
A1 = allele 1/effect allele for corresponding odds ratio; A2 = allele 2/alternate allele; FRQ = A1 frequency; SE = standard error; OR = odds ratio

Because the chromosome 4 and 18 associated SNPs reached genome-wide significance and suggestive levels, respectively, we removed the discovery p-value filter of less than 0.05 and examined the full 3-study meta-analysis results in these chromosomal locations. Regional LocusZoom plots summarize the fine mapping that is accomplishable via 1000G imputation. The association finding at 4q32.2 localizes to an approximately 250 Kb region upstream of *FSTL5* (Figure 2.9), and the 18q21 hit localizes to a 25 Kb region overlapping *SMAD7* (Figure 2.10).

Figure 2.9. LocusZoom⁸⁵ plot of regional association results for the novel 4q32.2 genome-wide significant locus (rs17042479 +/- 1Mb). The x-axis represents physical position on chromosome 4, and the y-axis shows the $-\log_{10}(\text{p-value})$ of from the meta-analysis of MECC Omni + CFR + MECC Axiom. Each circle represents one SNP's association with CRC. Purple = index SNP (rs17042479). Correlation (r^2) between the index SNP and each other SNP was calculated based on 1000 Genomes Phase I March 2012 European samples.

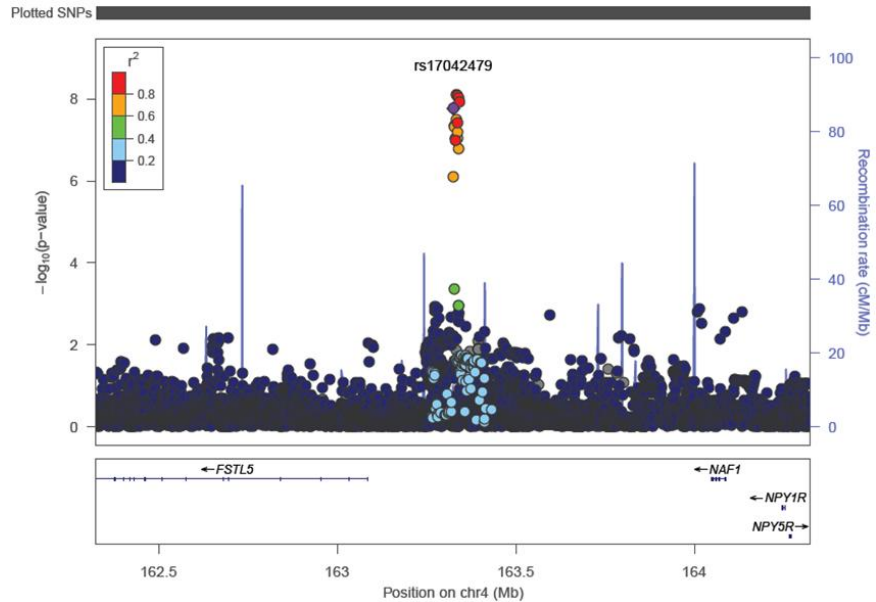
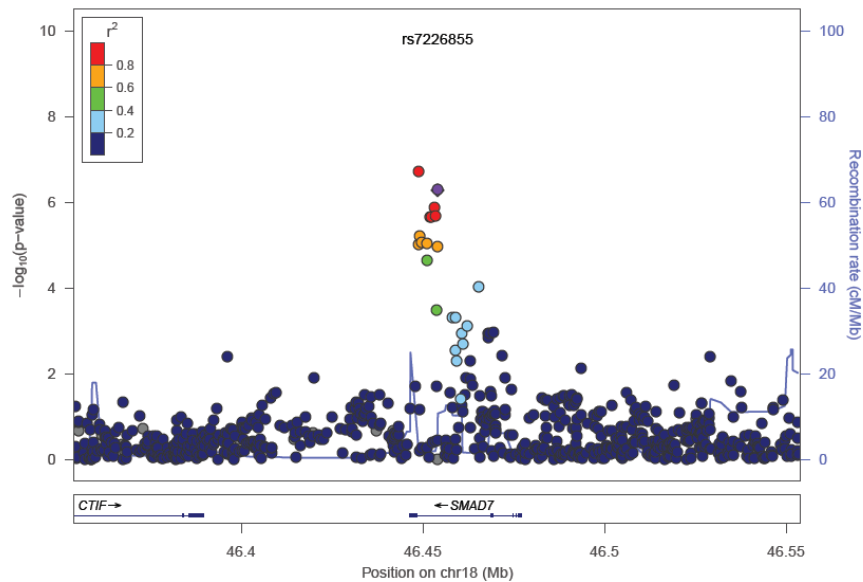


Figure 2.10. LocusZoom⁸⁵ plot of regional association results for the suggestive and previously identified 18q21 locus. The x-axis represents physical position on chromosome 4, and the y-axis shows the $-\log_{10}(\text{p-value})$ of from the meta-analysis of MECC Omni + CFR + MECC Axiom. Each circle represents one SNP's association with CRC. Purple = index SNP (rs7226855). Correlation (r^2) between the index SNP and each other SNP was calculated based on 1000 Genomes Phase I March 2012 European samples.



2.4 Discussion

This GWAS meta-analysis with independent replication was designed to identify novel, low-penetrance susceptibility loci among sub-populations with European ancestry. In the discovery meta-analysis, we identified a novel, genome-wide significant, CRC susceptibility locus on 4q32.2 with a MAF of approximately 9%. To date, no other CRC susceptibility loci have been identified on chromosome 4. This finding was replicated in an independent set of MECC cases and controls with a consistent direction of effect and nominal p-value less than 0.05. Evidence that known CRC risk loci were also identifiable with our study design (combining Ashkenazi Jewish and non-Hispanic whites) increased our confidence in the reproducibility and generalizability of this novel finding. Twenty-six of 29 previously published loci replicated with a consistent direction of effect, and for those with inconsistent direction of effect or for those not statistically significantly associated at $p < 0.05$, the results might be attributed to modest sample size or to differences in LD structure between our sample and other samples used in other GWAS.

The 4q32.2 region localized based on LocusZoom plots falls approximately 250 Kb upstream of the *FSTL5* gene. The frequencies of our lead SNPs matched closely with their reported MAFs for Europeans from HapMap (CEU) according to the dbSNP database. Interestingly, this SNP has a MAF close to 50% for those of Asian and African descent in HapMap. Preliminary bioinformatic analysis based on ENCODE⁸⁶ data in the UCSC Genome Browser (<http://genome.ucsc.edu/cgi-bin/hgGateway>) reveals that the broader region around the lead discovery SNP (rs17042479) may overlap with an H3K27Ac histone mark (a feature often located near active regulatory elements), a DNaseI hypersensitivity region (a chromatin accessibility feature common to cis-regulatory sequences), and/or a transcription factor binding site. While we do not yet have experimental evidence to confirm any functional regulatory role

of genetic variation in this region, it is possible that an intergenic SNP or the genetic element that it tags exerts a regulatory effect on the nearest gene, *FSTL5*. *FSTL5* encodes an extracellular matrix protein that interacts with metalloproteases and may be structurally similar to some collagen-degrading matrix metalloproteinases (MMPs) and MMP inhibitor TIMP1 that are critical for normal physiology⁸⁷. Little is known about the gene's function, but some evidence suggests potential links to known etiologic pathways involved in CRC development. In general, follistatins bind activins, regulate cellular differentiation, and neutralize TGF- β superfamily members (http://www.nlm.nih.gov/cgi/mesh/2011/MB_cgi?mode=&term =Follistatin). The TGF- β signaling pathway's role in CRC development has been clearly characterized.

Although there is no literature suggesting a direct role in CRC pathogenesis, genetic variation in and expression of *FSTL5* have been implicated in association with other complex phenotypes. Genetic variation in *FSTL5* has been associated with bone marrow suppression following thiopurine treatment for inflammatory bowel disease⁸⁷. *FSTL5* expression measured by immunohistochemistry has been described as a biomarker of poor prognosis in medulloblastoma⁸⁸. Among the minimal amount of published literature on the biological relevance of this gene and gene region, the NHGRI Catalog of Published Genome-Wide Association Studies⁸⁹ has shown effects of markers at 4q32.2 on complex traits, however, on completely unrelated phenotypes. An intergenic SNP between *FSTL5* and *NAF1* has been associated with hair morphology⁹⁰. SNPs downstream of *FSTL5* have been associated with response to amphetamines⁹¹ and diabetic retinopathy⁹², and an intronic *FSTL5* SNP has been associated with height⁹³ via GWAS. Importantly, we provided evidence that height is not likely to confound the genetic marker-CRC associations for our most statistically significant findings upstream of *FSTL5*.

Although power calculations would suggest a relatively high probability of identifying novel susceptibility loci, the data reveal only a single locus that reaches nominal levels of genome-wide significance. Given the selection for cases with a family history of CRC within the CFR study, it is possible that this genome wide-signal represents a combination of a low penetrance susceptibility allele in the Ashkenazi Jews (MECC) and a higher penetrance, rare effect among family history-positive, young age at onset non-Hispanic white (CFR) individuals may be driving our result. Further, potential selection bias resulting from CFR case enrichment for younger age at onset and family history may have led to overestimates of effect, enabling the novel chromosome 4 locus's detection here but not in previous GWAS. A similar phenomenon occurred when the penetrance of *BRCA1* and *BRCA2* mutations for breast cancer were initially overestimated by studies of high-risk families⁹⁴. False positive findings also remain a threat to the interpretation of GWAS studies, even when setting a reasonably high threshold for GWAS significance. Additional replication studies will help to clarify this possibility. We are currently investigating these associations in an additional ~4,500 CRC cases and ~4,500 controls from the CORECT study and a new study of 30,000 cases and 15, 000 controls that are planned to be studied within the NCI-funded GAME-ON consortium.

To determine if this locus has any biological relevance, there are a number of next steps to consider. First, fine mapping and screening of coding regions for mutations using standard methods offers an advantage to localize the most strongly associated SNPs within the region⁹⁵. Next, if justified by fine mapping exercises and more extensive searches of ENCODE data, functional work is warranted, leveraging expression quantitative trait loci (eQTL) analyses and experimental studies such as chromatin immunoprecipitation and sequencing (ChIP-seq) and electro-mobility shift assays.

While this study has multiple strengths, it is also limited by some of the same considerations common to most GWAS studies. First, the assumption that all SNPs conform to a log-additive genetic model creates an over-simplified view of the genome. However, it is not feasible to identify the appropriate genetic model on a SNP-by-SNP basis for millions of genetic markers. Though it is possible to run multiple model forms (recessive, co-dominant, dominant), log-additive models were chosen here because they provide a parsimonious approach to the data that yield the greatest statistical power while attempting to limit multiple testing. Second, the sample size limited our power to detect an effect, and particularly, the power to examine the effects of rarer variants either directly measured on the arrays or imputed ($MAF < 1\%$). However, it is clear that the combination of Ashkenazi Jewish individuals and non-Hispanic whites enriched for family history comprised a unique study sample for detection of a novel result. Further, the choice of the non-Ashkenazi Jewish Colon CFR as our second population for discovery may decrease the ability to detect significant variants that may be specific to the Ashkenazi Jewish founder population..

Finally, multiple levels of control influence the development of a disease phenotype including but not limited to known epidemiologic risk factors, sequence variation, transcriptional regulation, gene expression, DNA methylation, and chromatin modifications. Here we only focus on germline genetic variation, while environmental factors, gene-gene interactions, and gene-by-environment interactions are known to play roles in the development of CRC. Interactions were not the focus here because of the prohibitively large samples sizes needed for their study. In addition, GWAS have been criticized for multiple reasons, among which is the inability to explicitly determine functionality of SNPs identified to be associated with disease.

As discussed above, next steps including fine mapping of regions with significant hits and association studies with *cis* gene expression will be critical to glean initial insights into the function of the genetic region in relation to CRC development. Further, the data from this chapter contribute to a larger CORECT consortium effort, which will ultimately have much stronger power for revealing common variants with extremely low effect size and low/intermediate-frequency variants with larger effect sizes. Each of these variants will have the potential to explain some proportion of the missing heritability for CRC. The combined contribution of newly-identified susceptibility variants and known risk loci to explaining the missing heritability of CRC will be a focus of future investigations using risk modeling. Here, we did not calculate the population attributable risk for each suggestive variant, as this parameter can be difficult to interpret due to consideration of each factor in isolation without accounting for gene-gene or gene-environment interactions.

Chapter 3. MicroRNA target site polymorphisms and colorectal cancer risk in the Ashkenazi Jewish population

3.1 Introduction

In addition to protein-encoding messenger RNAs (mRNAs), other classes of small RNA molecules exist with specialized regulatory and processing functions. Among these types of regulatory RNAs are microRNAs (miRNAs), short (18-24 nucleotide) non-protein-coding molecules that act as post-transcriptional regulators of gene expression³². The biogenesis of a miRNA begins with transcription from a small, stand-alone gene or an intron or exon of a known protein-coding gene and transitions through a series of conversion steps from hairpin precursors to duplexed pre-miRNA intermediates, and finally, to mature, single-stranded miRNAs^{96,97}. MiRNAs exert their regulatory effects via binding to complementary ~6-8 nucleotide target seed sites in the 3' untranslated regions (3'-UTRs) of one or more mRNAs. Depending on the fidelity and context of the interaction, this binding acts to repress translation of the messenger into protein or to signal for degradation of the targeted mRNA^{32,98,99}. Each miRNA typically binds multiple, even thousands, of messenger targets, offering the potential for widespread downstream effects^{42,100}.

Deregulated miRNA profiles have been described across a range of hematologic malignancies and solid tumors including colorectal cancer (CRC)^{101,102}. Further, it has been suggested by some that miRNA biology can be integrated into the molecular sub-typing of colorectal tumors and into the traditional model of genetic alterations accompanying progression from normal mucosa to carcinoma, particularly among tumors that develop through the

chromosomal instability pathway^{100,103-106}. As an extension of this work, several miRNAs have been proposed as biomarkers for CRC early detection, prognosis, and progression¹⁰⁷⁻¹⁰⁹.

Despite the extensive miRNA profiling in colorectal tumors, the factors driving aberrations in miRNA expression and their impact on CRC development remain less clear. One hypothesis proposes that single nucleotide polymorphisms (SNPs) found in genes encoding the miRNA sequence or 3'-UTR regions of the corresponding binding sites affect miRNA transcription, miRNA processing, and/or the fidelity of the miRNA-mRNA interaction. In turn, any of these alterations could plausibly impact target mRNA translation into proteins critical for cellular differentiation and proliferation. Evidence from studies of candidate miRNA-related genetic alterations supports this hypothesis and suggests that such SNPs may alter expression of some miRNAs in CRC¹¹⁰ and increase or decrease the risk of tumor development³³. Target site polymorphisms that confer risk of CRC in specific populations have been identified in *INSR*³³, *CD86*³³, *IL16*¹¹¹, *RPA2*¹¹², and *GTF2H1*¹¹²; however, replication of these findings has been limited with the exception of rs1051690 in *INSR* and rs17281995 in *CD86*¹¹³. To date, published studies have been limited to candidate miRNA-related SNP analyses and have not comprehensively investigated polymorphisms implicated in the post-transcriptional miRNA regulatory pathway across the genome.

In this study, we expanded the search for miRNA-related genetic variants important in the etiology of CRC across the genome and investigated the association between thousands of genetic variants in miRNA target sites in 3'-UTR regions and miRNA-encoding genes and CRC risk using a novel genotyping platform. As opposed to a classical genome-wide association study (GWAS) approach which relies on linkage disequilibrium (LD) with haplotype-tagging SNPs, we leveraged genotyping of SNPs bioinformatically predicted to have functional implications

specific to the miRNA regulatory pathway. Further, we bioinformatically characterized the potential miRNA binding consequences of our most significantly associated SNPs and further explored these associations with expression quantitative trait loci (eQTL) analyses. This study was designed to evaluate the feasibility of a targeted GWAS approach for identifying lead candidates and prioritizing them for functional characterization based on biologically relevant hypotheses. The genetically homogeneous founder population of Ashkenazi Jewish individuals experiences a high burden of CRC and served as the focus of this study³⁷.

3.2 Materials and methods

3.2.1 Study population: Molecular Epidemiology of Colorectal Cancer (MECC) Study

MECC is a population-based, case-control study of pathologically-confirmed, incident cases of CRC recruited from a geographically-defined region of northern Israel. Subject recruitment began in 1998 and remains on-going. Individually-matched controls with no prior history of CRC are selected from the same source population that gave rise to cases based on the Clalit Health Services database. Matching factors include age, sex, Jewish ethnicity (Jewish versus non-Jewish), and primary clinic site. Subjects are interviewed to obtain demographic data, clinical information, family history, and dietary habits. Also, biospecimens including blood, paraffin blocks, and snap frozen tumors are collected. This genome-wide analysis included genotype data on 596 cases and 429 controls from MECC (Table 3.1). Case selection for genotyping in this phase was enriched for those of self-reported Ashkenazi Jewish ancestry.

Table 3.1. Demographic and clinical characteristics of MECC participants (n=1,025) genotyped on the Axiom[®] miRNA Target Site Genotyping Array platform.

	Cases (n=596)	Controls (n=429)
Age [mean(sd)]	70.9 (10.7)	74.3 (10.6)
Sex (%)		
Male	291 (48.8)	221 (51.5)
Female	305 (51.2)	208 (48.5)
Self-reported race/ethnicity (%)		
Ashkenazi	595 (99.8)	413 (96.3)
Ashkenazi/Sephardi	1 (0.2)	4 (0.9)
Sephardi	0	7 (1.6)
Ashkenazi/non-Jewish	0	2 (0.5)
Missing	0	3 (0.7)
Cancer site (%)		
Left colon	198 (33.2)	-
Right colon	189 (31.7)	-
Rectum	168 (28.2)	-
Other	33 (5.5)	-
Missing	8 (1.3)	-
Stage at diagnosis (%)		
I	95 (15.9)	-
II	125 (21.0)	-
III	118 (19.8)	-
IV	48 (8.1)	-
Missing	210 (35.2)	-

3.2.2 Genotyping and quality control

Germline DNA was extracted from peripheral blood samples and sent to Affymetrix for genotyping. Genotyping was conducted using a novel Affymetrix Axiom[®] miRNA Target Site Genotyping array with 237,858 probes (Table 3.2). The chip was originally designed by Affymetrix using four different bioinformatic miRNA prediction algorithms, including PolymiRTS, dPORE, Patrocles, and microRNA.org. These algorithms were leveraged to select polymorphic sites for the array that overlap genes encoding miRNAs, genes encoding proteins important for miRNA processing, and/or target seed sites¹¹⁴⁻¹¹⁷. In addition, the array included a

panel of ancestry informative markers (AIMs) and loci with known complex trait associations from the August 16, 2011 National Human Genome Research Institute (NHGRI) GWAS Catalog¹¹⁸. The same MECC samples used in this study were also genotyped on a custom Affymetrix Axiom[®] genotyping platform with ~1.3 million SNPs and indels for a GWAS as part of the Colorectal Transdisciplinary (CORECT) Study to compare concordance across the genotyping platforms.

Table 3.2. Variant sources and design of the Axiom[®] miRNA Target Site Genotyping Array. Data provided by Affymetrix.

Source	Type	Array
PolymiRTS	Mature miRNA, processing proteins, miRNA	440
	Target seed sites	85,900
dPORE	miRNA gene regulatory regions	10,400
Patrocles	miRNA genes & target seed sites	1,200
microRNA.org	Conserved & un-conserved target seed sites	158,400
Other	Ancestry Informative Markers	4,470
	Genome-wide association study compatibility	5,010
Total		237,858

MECC genotype data was cleaned based on quality control metrics at the individual subject and SNP levels (Figure 3.1). Samples with >5% missing genotypes, sex mismatches (between self-reported and genotypic predicted sex), and duplicate samples were identified and subsequently removed. Monomorphic SNPs and those with <95% call rate were excluded as well as those inconsistent with Hardy Weinberg Equilibrium (HWE) in controls. Principal components analysis (PCA) was conducted on a panel of 4,736 AIMs using the `pcaMethods` Bioconductor package^{75,119} in R to identify ethnic outliers for removal and to later adjust for

confounding due to population stratification. Plots of principal components (PCs) 1-3 on the final analysis dataset were generated to examine population structure. Because almost all of the self-reported Sephardi Jews were controls, making PC adjustment for population structure nearly impossible, we eliminated PC-defined non-Ashkenazi Jews in the same way as in Chapter 2.

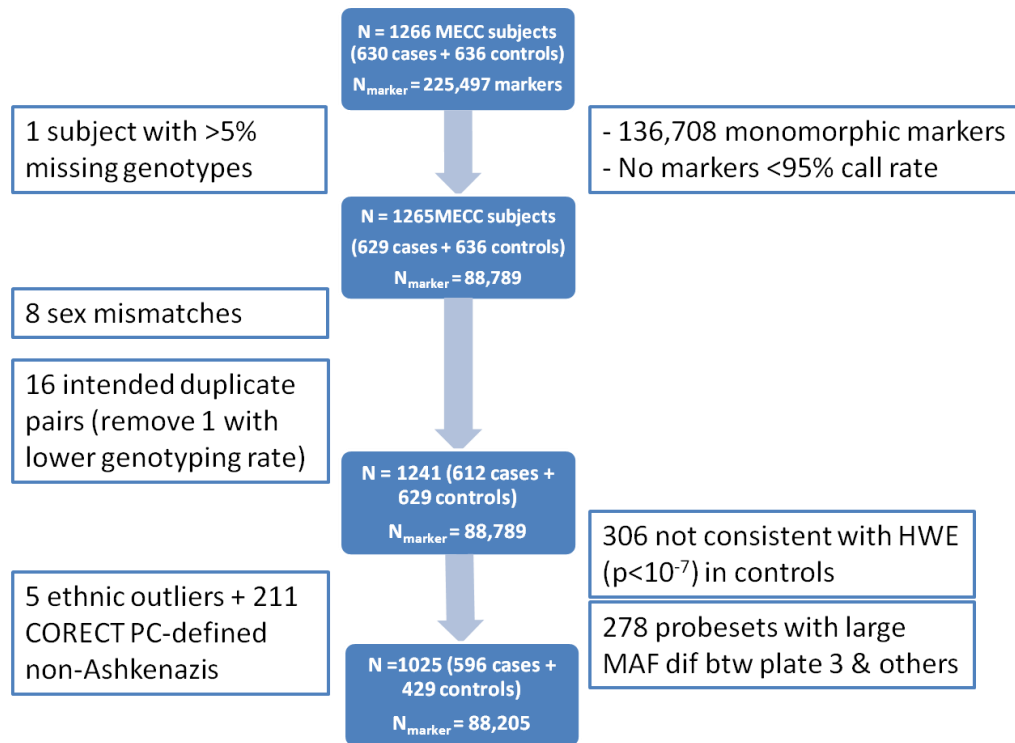


Figure 3.1. Quality control and filtering pipeline for MECC samples genotyped on the Affymetrix Axiom[®] miRNA Target Site Genotyping Array. PC = principal component. MAF = minor allele frequency.

3.2.3 Gene expression quantification

Gene expression levels from 419,473 probe sets derived from two Affymetrix expression arrays were quantified on RNA isolated from snap frozen tumors of 331 MECC CRC cases. Of these 331 cases, 135 also had high-throughput genotype data available (63 on the Affymetrix Axiom[®] CORECT custom array and 72 on the Illumina HumanOmni 2.5S-v1 BeadChip). Methods for gene expression quantification via hybridization to GeneChip[®] Human Genome U133A 2.0 and Human Genome U133 Plus 2.0 Arrays have been described elsewhere¹²⁰.

Briefly, expression was measured in two batches (one for each array) followed by quantile normalization and \log_2 transformation of MAS 5.0-calculated signal intensities. Data from the two batches were aligned after individual batch preprocessing and quality control.

3.2.4 Statistical and bioinformatic analysis

Logistic regression was employed to examine the marginal association between each marker on the miRNA target site array with $MAF \geq 1\%$ ($n_{\text{marker}} = 55,208$) and CRC risk assuming a log-additive genetic model. Here, each additional copy of the minor allele was assumed to confer the same magnitude of risk or protection. Each model was run both unadjusted and adjusted for sex, age, and the first two PCs (see parameterization below). We calculated beta coefficients, standard errors, odds ratios (OR) with associated 95% confidence intervals, and p-values from unconditional logistic regression. The Bonferroni-corrected alpha level was set at 9.0×10^{-7} ($0.05/55,406$ SNPs).

Unadjusted:

$$\text{logit}[P(\text{CRC})] = \beta_0 + \beta_1 * \text{SNP} + \varepsilon$$

Adjusted:

$$\text{logit}[P(\text{CRC})] = \beta_0 + \beta_1 * \text{SNP} + \beta_2 * \text{PC1} + \beta_3 * \text{PC2} + \beta_4 * \text{AGE} + \beta_5 * \text{SEX} + \varepsilon$$

where $\text{SNP} = 0$ for AA, 1 for AB, and 2 for BB when B is the minor allele

PC1 = value of principal component 1

PC2 = value of principal component 2

AGE = age at diagnosis (cases) or age at study recruitment (controls)

SEX = 1 for male, 2 for female

After taking this genome-wide approach, we then examined previously published SNPs from three studies in the candidate miRNA-related polymorphism literature to assess our ability to replicate purported risk loci^{33,111,112}.

To begin the bioinformatic characterization of functional consequences of our most significantly associated SNPs, we investigated predicted changes in miRNA binding using a combination of algorithms: microrna.org, miRBase, PolymiRTS, and dPORE^{114,115,119,121-124}. In addition, we conducted analysis of variance (ANOVA) to compare differences in gene expression by genotype for all SNPs with association p-values less than 5×10^{-4} , where expression and genotype data permitted. Expression of the gene nearest to each SNP was considered.

3.3 Results

3.3.1 Targeted genome-wide association analysis

The distributions of demographic and clinical characteristics of the final analysis dataset were comparable across case and control groups (Table 3.1). Plots of the first 3 eigenvectors from MECC PCA indicated that the original samples selected for analysis included some non-Ashkenazi Jewish individuals (almost exclusively among controls) that inhibited our ability to control for confounding due to population stratification through PC adjustment (Figure 3.2). Following removal of 5 ethnic outliers and 211 PC-defined non-Ashkenazis, the first 2 PCs were sufficient to control for population stratification, as indicated by genomic control lambda (GC λ) values shown below.

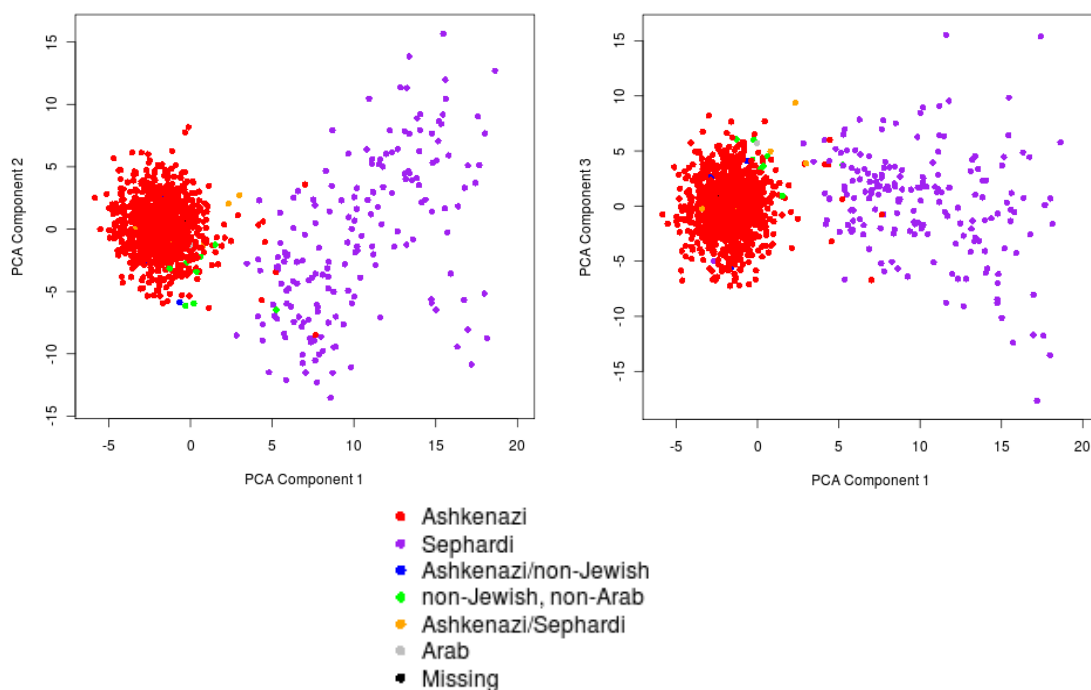


Figure 3.2. Plots of principal components 1 versus 2 and 1 versus 3 based on 4,736 AIMs from the MECC participants genotyped on the Axiom[®] miRNA Target Site Genotyping Array.

Quantile-quantile (Q-Q) and Manhattan plots visually display $-\log_{10}(\text{p-values})$ resulting from the logistic regression models adjusted for age, sex, and 2 PCs (Figure 3.3). The Q-Q plot in the left panel plots the rank-ordered observed $-\log_{10}(\text{p-value})$ against the rank-ordered expected $-\log_{10}(\text{p-value})$. It demonstrates that, on average, we did not observe SNPs with associations more statistically significant than expected under a uniform distribution of p-values. The GC λ value of 1 suggests that PCs 1 and 2 were sufficient to control for population stratification in our ethnically homogenous study sample. The Manhattan plot displays the summary results by ordered chromosomal position and shows that our lowest p-values are in the 10^{-5} range with none reaching genome-wide statistical significance after correction for multiple testing.

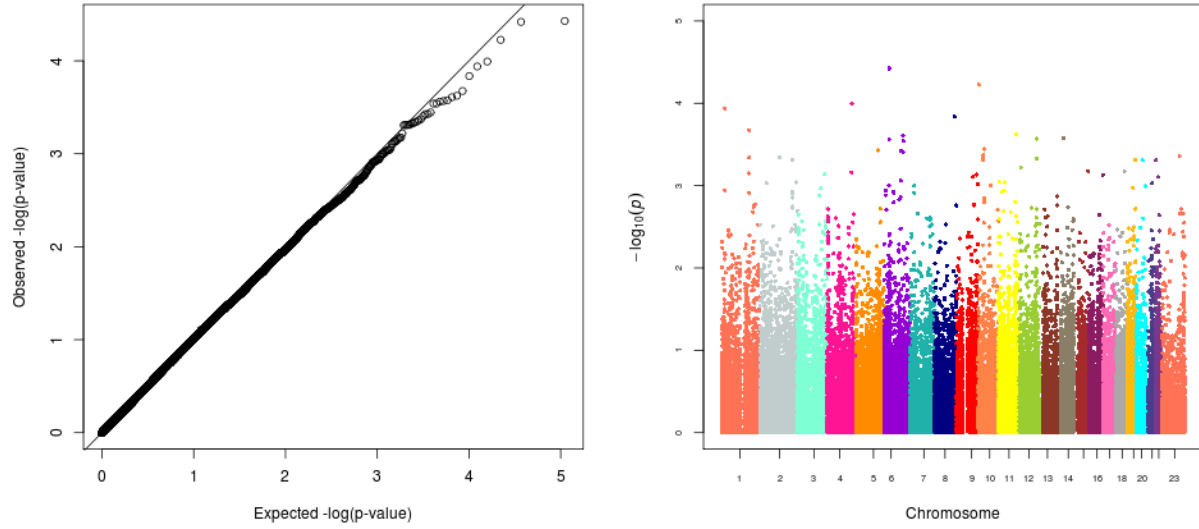


Figure 3.3. Quantile-quantile and Manhattan plots of p-values for association between each SNP and CRC from logistic regression adjusted for sex, age, and 2 PCs. MAF \geq 1%; $n_{\text{marker}}=55,406$. Genomic control $\lambda = 1$.

Although none of the individual SNPs achieved genome-wide significance, our top findings are detailed in Table 3.3. Interestingly, seven out of our nine most statistically significant SNPs yield a predicted change in miRNA binding in an allele-specific manner. Each of these seven significant variants predict either a change from no miRNA binding to one or more miRNAs binding or from one set of miRNAs to a different set. None of the most significant miRNA SNPs has been previously reported as significantly associated with risk of CRC. *RAPGEF2* and 4 other nearest genes for SNPs in Table 3.3 were selected for detailed expression analysis based on availability of companion probe set data.

Table 3.3. Summary of variants with the most statistically significant associations from the age, sex, and PC 1-2 adjusted logistic regression models. A1 miRNA and A2 miRNA refer to miRNAs predicted to bind in the presence of each alternative allele. AIM = ancestry informative marker. GWAS = GWAS compatibility marker. NA = no prediction available. Grey indicates the availability of high-throughput genotype data and gene expression data for the nearest gene.

rsID	CHR	BP	A1*	MAF	OR	SE**	P	Gene	A1 miRNA	A2 miRNA
rs2985	6	34845648	C	0.30	0.66	0.10	3.7E-05	<i>UHRF1BP1</i> (intron)	None	miR-885-5p
rs1139139	10	5020625	T	0.27	1.54	0.11	6.0E-05	<i>AKR1C1</i> (downstream)	None	miR-451b miR-556-5p
rs6827968	4	161399652	A	0.34	1.48	0.10	1.0E-04	<i>RAPGEF2</i> (downstream) <i>FSTL5</i> (downstream)	AIM	AIM
rs12130051	1	19545053	T	0.06	2.27	0.21	1.2E-04	<i>KIAA0090</i> (3'-UTR)	None	miR-222 miR-1244 miR-3129
rs80350662	1	178819064	A	0.04	2.63	0.26	2.1E-04	<i>RALGPS2</i> (intron)	miR-32-3p miR-4775	miR-1277- 5p miR-889
rs1834481	11	112023827	G	0.12	0.60	0.14	2.4E-04	<i>IL18</i> (intron)	None	miR-637 miR-5009- 5p miR-541-3p
rs1044724	6	125412231	C	0.08	0.54	0.17	2.5E-04	<i>RNF217</i> (downstream) <i>TPD52L1</i> (upstream)	miR-3978	miR-3978
rs4766991	12	113137384	T	0.16	1.58	0.13	2.7E-04	<i>PTPN11</i> (downstream) <i>RPH3A</i> (upstream)	In promoter for miR- 1302-1	In promoter for miR- 1302-1
rs7746892	6	125408263	G	0.08	0.54	0.17	2.9E-04	<i>RNF217</i> (downstream) <i>TPD52L1</i> (upstream)	miR-545	miR-1252 miR-4476 miR-4533 miR-873
rs7746860	6	125408221	G	0.08	0.54	0.17	2.9E-04	<i>RNF217</i> (downstream) <i>TPD52L1</i> (upstream)	miR-2681	miR-1295 miR-4747- 3p
rs2489495	10	38502333	T	0.20	0.66	0.12	3.6E-04	<i>LOC1001290</i> 55 (exon)	miR-635	NA
rs853158	5	142605172	C	0.36	0.71	0.10	3.7E-04	<i>ARHGAP26</i> (3'-UTR)	miR-3926	miR-4480
rs9374072	6	109591586	G	0.31	0.71	0.10	3.7E-04	<i>CEP57L1</i> (downstream) <i>CCDC162</i> (upstream)	miR-605	NA
rs471429	6	125409031	G	0.09	0.57	0.16	3.9E-04	<i>RNF217</i> (downstream) <i>TPD52L1</i> (upstream)	miR-3126- 5p miR-3174 miR-3591- 5p	miR-4270 miR-4441

									miR-3606 miR-4419a miR-4510 miR-921	
rs12268559	10	32856746	C	0.07	0.54	0.18	4.3E-04	<i>CCDC7</i> (coding)	miR-578	NA
rs142004998	23	119760042	C	0.08	0.51	0.19	4.4E-04	<i>CIGALT1C1</i> (3'-UTR)	miR-1284 miR-337-3p miR-520d-5p	NA
rs3180466	2	129023866	G	0.03	0.38	0.28	4.5E-04	<i>HS6ST1</i> (3'-UTR)	miR-4758-5p miR-574-5p miR-615-3p	NA
rs79029362	1	178516408	G	0.03	2.91	0.30	4.6E-04	<i>Clorf220</i> (exon)	miR-455-3p	NA
rs4766992	12	113137531	A	0.26	1.45	0.11	4.7E-04	<i>PTPN11</i> (downstream) <i>RPH3A</i> (upstream)	Upstream of miR-1302-encoding gene	Upstream of miR-1302-encoding gene
rs56391924	10	32745248	C	0.07	0.54	0.18	4.7E-04	<i>CCDC7</i> (missense)	miR-4273	NA
rs117299563	19	52095600	C	0.02	0.28	0.37	4.9E-04	<i>FLJ30403</i> (exon,3'-UTR)	NA	miR-3684
rs6072275	20	39743905	A	0.15	1.59	0.13	4.9E-04	<i>TOP1</i> (intron)	GWAS	NA
rs107321	22	18512282	T	0.41	0.72	0.09	4.9E-04	<i>FLJ41941</i> (exon)	miR-1284 miR-337-3p miR-374a-5p	NA
rs1972820	2	212243422	G	0.36	1.41	0.10	4.9E-04	<i>ERBB4</i> (3'-UTR)	miR-4633-5p miR-532-5p	miR-3144-3p miR-875-5p
rs12247495	10	32802829	C	0.07	0.54	0.18	4.9E-04	<i>CCDC7</i> (intron)	miR-539-5p	NA

*A1 = effect allele for the corresponding OR; ** SE = standard error of the beta estimate

3.3.2 Gene expression analysis for top association findings

Of the top 25 SNPs that met our p-value threshold of 5×10^{-4} from the association analysis in 3.3.1, 11 corresponding nearest genes had at least 1 matching probe in our gene expression dataset. Among the 21 total probes quantifying gene expression from these 11 genes (some genes had multiple probes), 13 probes for 6 genes had a corresponding genotype measured in MECC cases from the custom Affymetrix Axiom and/or Illumina Omni platforms.

Table 3.4 enumerates ANOVA results for gene expression [$\log_2(\text{normalized intensity})$] by genotype for the 6 represented nearest genes with appropriate data availability. The most statistically significant intergenic SNP (rs6827968) falls downstream of the RAP guanine nucleotide exchange factor 2 (*RAPGEF2*) gene, with an F statistic of 5.71 and p-value of 0.02. *RAPGEF2* expression levels plotted against the number of copies of the minor allele at this SNP locus in our study sample can be visualized in Figure 3.4. Although rs6827968 is highly unlikely to exert a direct regulatory influence via the miRNA pathway on the nearest gene since it is an AIM, *RAPGEF2* encodes a protein that could plausibly be linked to CRC etiology. *RAPGEF2* activates RAS through promotion of the active GTP-bound state in a GTP/GDP-regulated signal transduction switch¹²⁵. This SNP also falls downstream of the follistatin-like 5 (*FSTL5*) gene. Given the genome-wide significance of a SNP in *FSTL5* (Chapter 2), we were interested in whether or not this miRNA-associated SNP might also lead to deregulation of *FSTL5* gene expression. However, no corresponding gene expression probe was available in our assay. Experiments are ongoing to evaluate this further.

Table 3.4. ANOVA results for gene expression [$\log_2(\text{normalized intensity})$] by genotype at top SNPs from the logistic regression association analyses between each SNP and CRC status.

SNP	Gene	Probe	F statistic	P
rs6827968	<i>RAPGEF2</i>	203096_s_at	5.71	0.02
rs6827968	<i>RAPGEF2</i>	203097_s_at	2.51	0.12
rs6827968	<i>RAPGEF2</i>	215992_s_at	3.62	0.06
rs12130051	<i>KIAA0090</i>	212394_at	0.16	0.69
rs12130051	<i>KIAA0090</i>	212395_s_at	1.60	0.21
rs12130051	<i>KIAA0090</i>	212396_s_at	0.18	0.67
rs12130051	<i>KIAA0090</i>	215991_s_at	0.33	0.57
rs853158	<i>ARHGAP26</i>	205068_s_at	1.51	0.22
rs853158	<i>ARHGAP26</i>	205069_s_at	0.05	0.82
rs853158	<i>ARHGAP26</i>	215955_x_at	0.97	0.33
rs6072275	<i>TOP1</i>	208900_s_at	0.13	0.72
rs6072275	<i>TOP1</i>	208901_s_at	1.44	0.23
rs1972820	<i>ERBB4</i>	206794_at	0.07	0.80

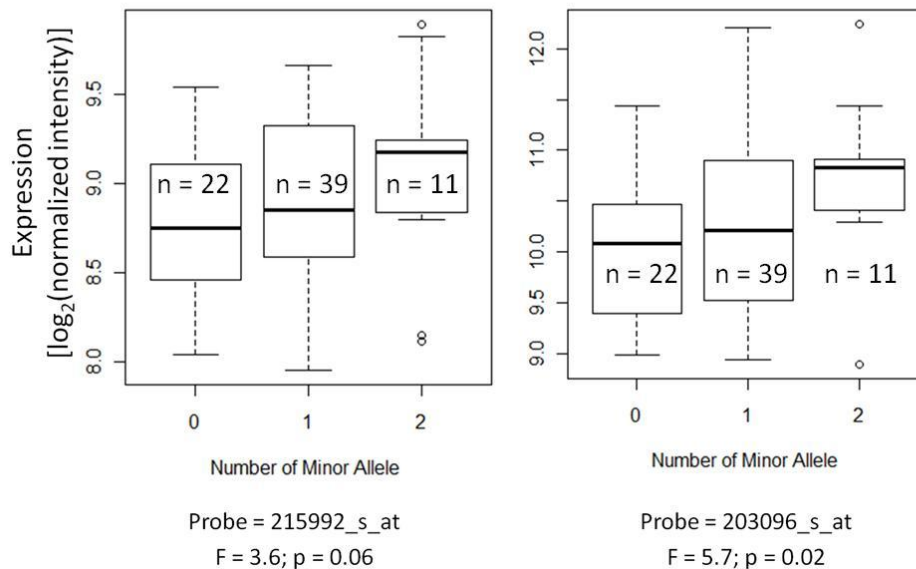


Figure 3.4. *RAPGEF2* gene expression in colorectal cancers measured by two separate probes in 72 MECC CRC cases by rs6827968 genotype.

3.3.3 Replication of previously published risk loci

We also examined the CRC association with 19 candidate SNPs previously presented in the literature (8 from Landi et al³³, 5 from Azimzadeh et al¹¹¹, and 6 from Naccarati et al¹¹²), of

which 6 were statistically significant in the original report. In our dataset, we replicated only one of the previously reported findings (Table 3.5). The single replicated variant (rs1051690), originally reported in Landi *et al*³³, falls in the 3'-UTR region of the insulin receptor gene *INSR* (Table 3.5; OR = 1.38; p = 0.03) and has predicted miRNA binding consequences. Although this SNP did not meet our association p-value threshold for eQTL analysis, we included it in order to provide additional evidence to support prioritization of a replicated locus for functional characterization. This SNP provides a clear example of how such a target site polymorphism could influence that same gene's expression in a dose-response manner (Figure 3.5; F = 21.3; p = 9.0×10^{-6}).

Table 3.5. MECC results from logistic regression adjusted for age, sex, and PC 1-2 for previously studied miRNA variants. Data row 1 shows replication of a variant in *INSR* that was suggestively associated with CRC risk in Landi et al^{33,113} (OR (95% CI) = 1.86 (0.99,3.50); P=0.052).

SNP	CHR	Position	Gene	A1 ⁺	Published OR	Published P	MECC OR	MECC SE ⁺⁺	MECC P
rs1051690*	19	7116963	<i>INSR</i>	T	1.86	0.05	1.38	0.15	0.03
rs1368439*	5	158742014	<i>ILI2B</i>	G	1.17	0.65	0.80	0.12	0.06
rs11515**	9	21968199	<i>CDKN2A</i>	G	1.16	0.71	1.14	0.12	0.30
rs1126547***	3	14186757	<i>XPC</i>	C	1.13	0.73	1.14	0.13	0.32
rs3135500*	16	50766886	<i>NOD2</i>	A	1.22	0.07	1.09	0.09	0.37
rs1131445*	15	81601782	<i>IL16</i>	C	0.99	1.00	0.93	0.09	0.41
rs1131445**	15	81601782	<i>IL16</i>	C	2.21	0.00	0.93	0.09	0.41
rs1051208**	3	12625747	<i>RAF1</i>	T	1.11	0.85	0.90	0.14	0.44
rs4596***	11	18388128	<i>GTF2H1</i>	C	0.79	0.03	0.93	0.10	0.44
rs2229090***	3	14187345	<i>XPC</i>	C	0.91	0.38	1.08	0.12	0.54
rs17281995*	3	121839641	<i>CD86</i>	C	2.93	0.01	1.05	0.12	0.66
rs11677*	1	20301964	<i>PLA2G2A</i>	T	1.02	0.97	1.06	0.14	0.70
rs7356***	1	28218100	<i>RPA2</i>	C	1.33	0.04	1.02	0.09	0.81
rs1803541***	2	128014913	<i>ERCC3</i>	T	0.96	0.70	0.97	0.14	0.86
rs16870224*	5	40692940	<i>PTGER4</i>	A	2.31	0.14	0.99	0.12	0.94
rs16870224**	5	40692940	<i>PTGER4</i>	A	0.29	0.11	0.99	0.12	0.94
rs4781563***	16	14045399	<i>ERCC4</i>	A	0.68	0.09	1.00	0.10	0.97
rs916055*	17	4534834	<i>ALOX15</i>	C	0.98	0.91	1.00	0.10	0.98
rs743554**	17	73754248	<i>ITGB4</i>	A	0.76	0.36	NA	NA	NA

* Landi et al 2008³³; ** Azimzadeh et al 2012¹¹¹; ***Naccarati et al 2012¹¹²

⁺ A1 = effect allele for the corresponding odds ratios; ⁺⁺ SE = standard error of the beta estimate

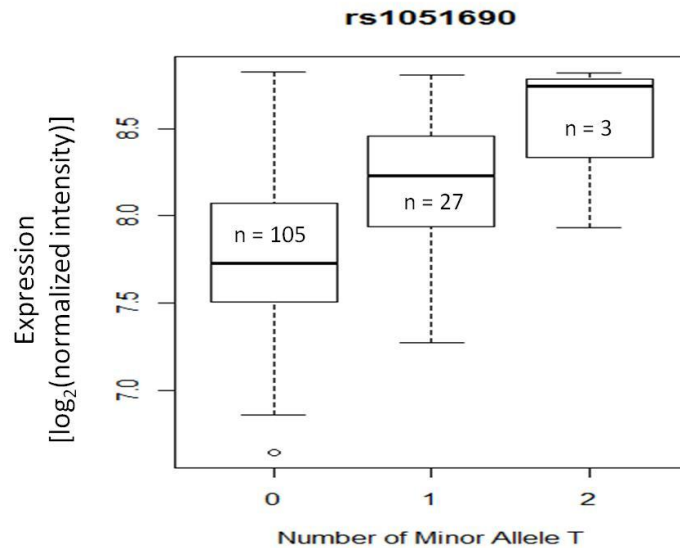


Figure 3.5. *INSR* gene expression in 135 MECC CRC cases by rs1051690 genotype. ANOVA F-statistic = 21.3 and p-value = 9.0×10^{-6} .

3.3.4 Genotype concordance: miRNA targeted array vs. traditional GWAS array

Only 14,436 markers were directly measured on the Affymetrix miRNA Target Site Genotyping Array and the CORECT Axiom 1.3M custom array. Among those variants, there was a 99.89% overall genotype concordance across arrays. After leveraging imputation to 1000 Genomes for the CORECT Axiom 1.3M custom array, 63,407 out of 88,205 markers passing quality control filters from the targeted array were imputed with high quality. A comparison of concordance between the Affymetrix miRNA Target Site Genotyping Array genotypes and directly measured and imputed best call genotypes from the CORECT Axiom 1.3M custom array showed that the targeted miRNA array has added value over the GWAS array for many markers from this regulatory pathway, as heterozygote genotype concordance is severely depressed for SNPs with $MAF \leq 5\%$ (Figure 3.6).

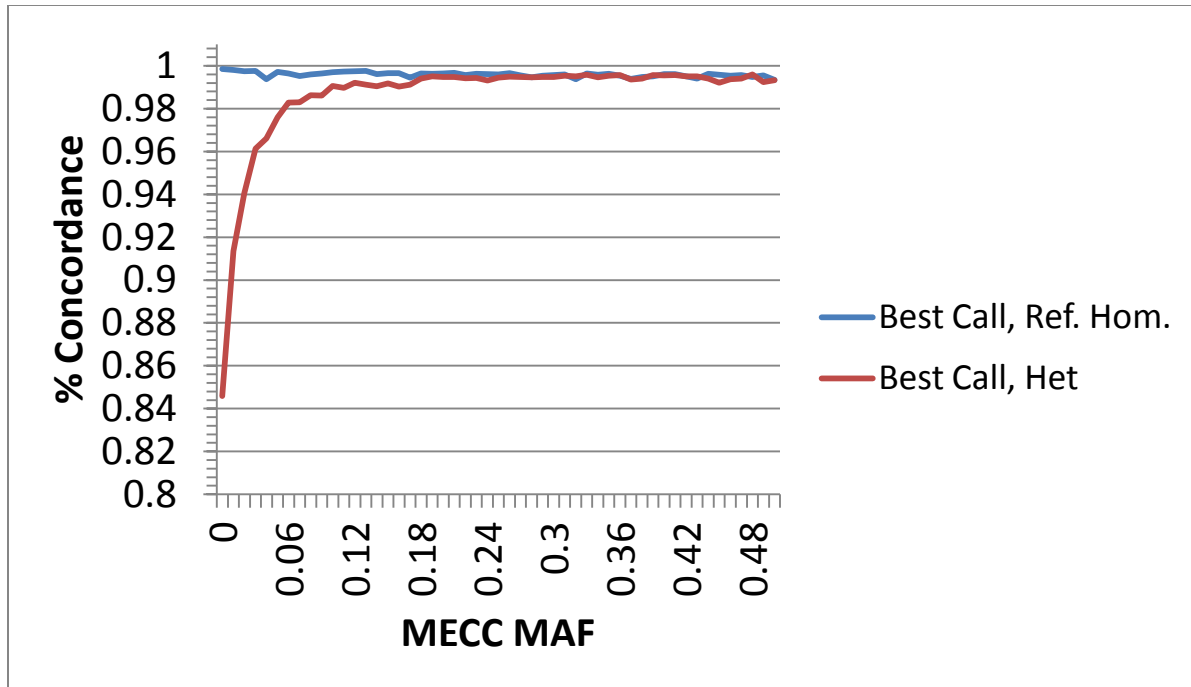


Figure 3.6. Concordance between directly measured and 1000 Genomes imputed CORECT Axiom genotypes and miRNA Target Site Array genotypes in MECC. Best call refers to the conversion of genotype dosages from imputation into best genotype calls based on a genotype probability threshold of 0.9. Ref. Hom. = reference homozygote; Het = heterozygote. Data provided by Affymetrix.

3.4 Discussion

To our knowledge, this is the first study to examine associations between genetic variations in miRNA genes or target seed sites and CRC risk using a genome-wide approach informed by bioinformatic miRNA prediction algorithms. While we did not identify any genome-wide significant associations meeting the traditional threshold of 5×10^{-8} , this study did highlight suggestive variants with predicted miRNA binding implications. Also, we replicated a recent association finding between rs1051690 in *INSR* and CRC risk and demonstrated variability in *INSR* gene expression by genotype at this locus. While limited with respect to power, this pilot study demonstrated the potential for a targeted GWAS approach to identify candidate susceptibility loci and to prioritize them based on biological insights for further functional characterization.

Alterations of expression from miRNA targets may be mediated by seed site polymorphisms that strengthen or weaken the miRNA-mRNA interaction. We illustrate a relevant example in this study from an association finding through eQTL analysis. The *INSR* 3'-UTR variant (rs1051690) association with CRC had previously been detected in both Czech Republic and Spanish case-control studies assuming a co-dominant model^{33,113}. We were able to replicate this risk locus based on a log-additive genetic model assumption. Further, in the Chapter 2 analysis which included more MECC samples as well as CFR samples, the T allele at rs1051690 had a consistent direction of effect (OR=1.11) with a p-value of 0.07. To date, few studies have examined the functional consequences miRNA-related SNPs. However, *INSR* is a notable exception. The same group that originally identified the *INSR* association later conducted in vitro luciferase reporter assays to show that the minor allele leads to differential regulation of reporter gene expression¹¹³. Evidence from our eQTL analysis corroborates this finding, and a link between insulin resistance and CRC has long been recognized¹²⁶. It is possible that each additional copy of the minor/risk allele reduces miRNA-mRNA binding to the point of inhibiting mRNA degradation, which is what may lead to the increased *INSR* gene expression observation. Or, it is possible that the SNP exerts an effect analogous to haploinsufficiency, such that one copy of the major allele is not sufficient to appropriately repress *INSR* protein expression. Further functional work is necessary to elucidate this SNP's mechanism of action.

Sethupathy and Collins suggested in 2008 that future studies involved in elucidating the role of miRNA-related polymorphisms in complex diseases such as CRC should focus on all three of these domains: genetic, functional (testing altered miRNA targeting mediated by genetic variation), and mechanistic (testing the mechanism by which altered miRNA leads to tumor development)¹²⁷. We provide evidence with respect to genetic and functional studies. In

accordance with their recommendations, we also minimize confounding by population stratification that limits many candidate gene studies. The next step is to expand our genotyped dataset to increase power for detecting novel risk loci. With respect to genotyping platform selection for future studies, this chapter highlighted an advantage of this novel genotyping array over a traditional GWAS array with imputation based on the 1000 Genomes Project haplotypes, particularly when rare variants are of interest. Further, replication and fine-mapping in the CORECT Study will strengthen our confidence in both novel and previously published findings. Also, functional studies are underway to identify SNP effects on miRNA binding fidelity (for rs1051690 as well as other top association findings) and to find the best in vitro model for allele-specific effects.

This study has limitations with respect to power and modeling assumptions. Our sample size is limited to only 1,025 samples. However, this analysis, which was able to replicate a previously identified miRNA risk locus and characterize preliminary functionality, provides justification for study in a larger sample. Our lack of genome-wide significant findings is likely to be attributable to a lack of power, and our sample size did not permit the investigation of effects for rare variants with $MAF < 1\%$. Also, not all SNPs exert their effects according to the assumed log-additive genetic model, and this choice made to restrict multiple testing could inhibit our ability to identify risk loci that are consistent with a recessive, dominant, or co-dominant model. Further, we did not consider interactions between these potentially risk-conferring variants or variant effects in the context of other genetic or environmental risk factors. Finally, our ability to examine gene expression was limited by both data availability and restriction to studying the SNP's nearest gene. SNPs in genic or intergenic regions involved with regulatory functions may or may not act in a *cis* manner.

Despite these limitations, we provide evidence that a targeted genome-wide approach for studying germline susceptibility can be extended beyond known or purported cancer biology pathways to the exploration of a regulatory pathway with widespread post-transcriptional effects. A better understanding of the mechanisms by which aberrations in miRNA expression and binding impact CRC development and progression may offer critical insights for prevention and targeted therapeutic approaches. Specifically, the *INSR* variant warrants further investigation in a functional setting to elucidate its role in the alteration of CRC risk.

Chapter 4. Identification of potentially causal genetic variants for complex congenital heart disease through whole-genome sequencing

4.1 Background

In the United States, approximately 1 out of every 100 live births is affected by congenital heart disease (CHD)¹²⁸. The prevalence rate of complex CHD cases that require intensive cardiac treatment and care is estimated to be 2.5 to 3 per 1000 live births¹²⁸. It is worthwhile to note that the literature is inconsistent with respect to reporting incidence rate versus prevalence rate, and our choice to use the term prevalence rate is consistent with the recognition that early fetal loss leads to a potentially major difference in incidence and prevalence rates. Further, prevalence estimates of complex CHD lesions among those advancing beyond neonates range from 30,000 to 180,000, depending on an assumption regarding the proportion of cases treated¹²⁹. Among the multiple forms of severe CHD, three are the lesions relevant to this dissertation chapter (prevalence per million live births): hypoplastic left heart syndrome (HLHS, 266 per million), truncus arteriosus (107), and tetralogy of Fallot (407). Hypoplastic left heart syndrome (HLHS) is a complex disorder characterized by underdevelopment of the left ventricle and narrowed or closed mitral and aortic valves¹³⁰. Truncus arteriosus is a congenital heart defect typically characterized by an anatomically abnormal truncal valve and a ventricular septal defect¹³¹. Tetralogy of Fallot (TOF) typically consists of four anatomical features resulting in low blood oxygenation: a ventricular septal defect, right ventricular hypertrophy, pulmonary stenosis, and an overriding aorta. Because we found evidence of these three complex CHD lesions clustering in a single pedigree, it is likely that all

three have the same underlying etiology in the family under study in this chapter. We hypothesize that the clustering of these anatomically distinct left-sided and conotruncal lesions (truncus arteriosus and TOF) is attributable to a more proximal defect in cardiogenesis.

The narrow sense heritability of severe CHD has been characterized for specific lesions and estimated to be 99% for HLHS alone or 74% for HLHS with associated cardiovascular malformations¹⁴. Despite the known heritability of CHD, specific genetic risk factors contributing to its etiology have remained poorly understood. Chromosomal anomalies, such as Turner syndrome, trisomy 13, trisomy 18, DiGeorge syndrome, and Jacobsen's syndrome (terminal 11q deletion) occur in 5-12% of cases^{130,132-135}. Also, several critical regions have been identified through a recent linkage study of 208 individuals from 33 families with HLHS and associated cardiovascular malformations: 6q23.3, 10q22.1, 7q31.2, 11q22.1, 12q13.1, 14q23.2, and 20q12¹³⁶. Additional candidate genes associated with HLHS in smaller studies include *GJAI*¹³⁷, *NOTCH1*¹³⁸, *HAND1*¹³⁹, and *NKX2-5*¹⁴⁰. Dozens of other candidate genes for complex CHD have been implicated in the recent literature as well¹⁴¹. Clearly, multiple genetic loci, possibly in combination with established environmental risk factors are implicated in HLHS and other forms of complex CHD. These environmental risk factors include, but are not limited to, maternal exposures such as maternal rubella, pre-gestational diabetes, phenylketonuria, therapeutic and non-therapeutic drugs, and organic solvents¹⁴². Further identification of risk-conferring variants and description of the heterogeneous genetic landscape of severe CHD have potential to fill in some of the phenotype's missing heritability, to increase our understanding of heart development mechanisms, and to guide the discovery of intervention strategies. It has been hypothesized that impairment of blood flow into or out of the developing left ventricle may lead

to left heart hypoplasia; therefore, variants identified in association with HLHS may provide insight into genes important for fetal blood flow control and primary heart development¹⁴³.

Next-generation sequencing technologies, including whole-genome sequencing (WGS), hold great promise for enhancing our understanding of the genetic basis underlying complex CHD by taking an agnostic approach to identifying genomic variation. NGS refers to a set of technologies that use massively parallel sequencing by synthesis to generate thousands to millions of short reads (~25-120 nucleotides) with error rates <1%. Paired-end reads are reconstructed into a linear sequence of the whole genome by leveraging the human reference genome as a scaffold. Benefits of this sequencing approach include profoundly lower per base cost of sequencing data as compared to direct Sanger sequencing and the ability to capture novel and extremely rare variation that is not typically assessed with SNP genotyping arrays.

This chapter brings together the rapidly evolving technology of WGS with family-based co-segregation analysis to identify rare, potentially causative variants for severe CHD that can inform follow-up prevalence studies in cohorts of unrelated individuals. We hypothesize that potentially causative genetic variants for complex CHD co-segregating with disease in an affected family will be identifiable through WGS and informative regarding candidate genes or pathways for future investigations. The index family is particularly unique in that multiple CHD cases cluster in a single generation with the same father but different mothers. The implication of this methodology, shown by example in this chapter, is that by treasuring your exceptions and leveraging informative families, one can narrow the list of candidate disease susceptibility genes and pathways to study epidemiologically in larger populations, which can be cost-prohibitive to explore agnostically.

4.2 Methods

4.2.1 Family recruitment

The family under study was recruited in 2009 after the proband (IS-1065/SHFG-12005) underwent surgical treatment for HLHS. Her two affected half-sisters, one with TOF (SHFG-12366/CT7-3) and one with truncus arteriosus (SHFG-12463/CT7-4) were also recruited into the study. Peripheral blood was drawn from these 3 affected individuals. Later, peripheral blood samples were collected from the proband's unaffected father (CGN-8845-00) in August 2011 and unaffected full brother (SHFG-23129) in December 2011. The pedigree depicts the presence of affected offspring of the unaffected father from two different maternal lineages (Figure 4.1). Informed consent was obtained according to Institutional Review Board-approved protocols at the Children's Hospital of Philadelphia and the University of Michigan.

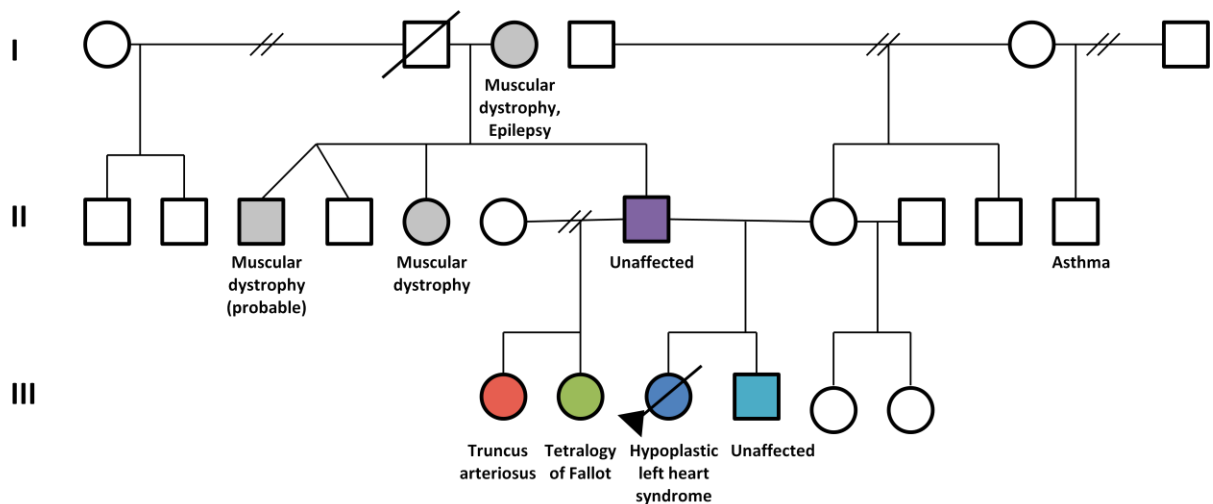


Figure 4.1. Pedigree for the index family affected by complex CHD. Proband is indicated by the black arrow. Individuals represented in color had DNA samples available for whole-genome sequencing. HLHS = hypoplastic left heart syndrome.

4.2.2 Whole genome sequence data generation

Germline DNA was extracted from the peripheral blood samples of 5 individuals from the index family using standard methods (Figure 4.1). Paired-end, whole-genome sequence data was generated on a combination of the Illumina Genome Analyzer II (GA-II) and HiSeq

platforms for germline DNA from 3 affected and 2 unaffected individuals spanning two generations of this family. These individuals included the proband with HLHS (III:3); her affected half-sisters with TOF (III:2) and truncus arteriosus (III:1), respectively; her unaffected full brother (III:4); and their shared, unaffected father (II:7).

Initial sequencing included 16 lanes of paired-end reads (120 nucleotides) from the Illumina Genome Analyzer II that were run at the University of Michigan (UM) DNA Sequencing Core. Sequencing libraries were prepared according to the manufacturer's instructions at UM using the NuGEN Encore™ NGS Library System I. The proband's DNA was run across 8 lanes, and each of the half-sib's DNA samples was run on 4 lanes. Because the ability to discover single nucleotide variants (SNVs) significantly increases with depth of coverage until approximately 15X for homozygous sites and 33X for heterozygous sites, additional sequencing depth for the proband and half-sib with TOF was generated on an Illumina HiSeq at the University of Southern California¹⁴⁴. DNA availability for the half-sibling (half-sib) with truncus arteriosus was limited and did not permit supplemental sequencing. In addition, new sequence data from the proband's unaffected father and full brother was obtained in order to offer new insights about variants segregating in this family. At USC, libraries were prepared using the Mondrian SP workstation (NuGen) with NEBNext DNA Library Prep reagents and the Mondrian SP Universal cartridge according to the Library Preparation – Method1 protocol.

4.2.3 Alignment of WGS paired-end reads

Sequencing reads were aligned to the human genome reference sequence, followed by local realignment around indels and recalibration. A combination of Burrows-Wheeler Aligner (v0.6.2) and Picard (v1.71; <http://picard.sourceforge.net>) were used to globally align paired-end reads in raw FASTQ file format to the human genome reference sequence NCBI37.2 and to mark

duplicate read pairs that commonly arise from PCR amplification of the same original fragments¹⁴⁵. Several tools from the Genome Analysis Toolkit (GATK; v1.6) were employed to convert raw aligned reads to analysis-ready reads¹⁴⁶. Reads were aggregated on a per-sample basis into a single binary alignment/map (BAM) file, duplicate reads were removed, and remaining reads underwent multi-sample realignment with known sites. Sample-level local realignment was accomplished using the RealignerTargetCreator and IndelRealigner tools from GATK and was informed by highly confident, known indels from the 1000 Genomes Project (1000G) and a 2011 *Genome Research* publication 2011^{79,147}. Following realignment, GATK's Base Quality Score Recalibration (BQSR) was used to adjust quality scores to more accurately reflect the probability of read mismatch to the reference. Average depths of coverage across the exome and across the genome for each sample were calculated using the GATK DepthOfCoverage tool.

4.2.4 SNP and insertion/deletion (indel) calling and annotation

The pipeline developed for SNP and small indel (~10 nucleotides or less) calling was modeled after the published GATK pipeline for multiple samples in a cohort, which allows variant discovery across samples in the index family¹⁴⁶. UnifiedGenotyper was run with a minimum Phred-scaled ($-10\log_{10}$) confidence score threshold of 20 to account for the range of coverage depths across samples and for the filtering steps to follow. Raw variant calls were then statistically filtered to remove predicted false positives using the Variant Quality Score Recalibrator (VQSR). This tool generated a Gaussian mixture model trained on highly-confident, known variant sites from dbSNP, 1000G, and HapMap to determine the probability that other sites in our dataset were true. For SNPs, the following annotations were used for training: quality by depth ratio for variant confidence (QD), site consistency with exactly two haplotypes

(HaplotypeScore), Mapping Quality Rank Sum Test (MQRankSum), Read Position Rank Sum Test (ReadPosRankSum), Fisher's exact test for strand bias (FS), mapping quality (MQ), and depth (DP). For indels, the following annotations were applied: QD, FS, HaplotypeScore, and ReadPosRankSum. VQSR added annotations to the "FILTER" field of the variant call format (VCF) file, and the user selected the truth sensitivity level at which to start filtering out variants based on a balance between likely true and false positives and on the novel transition-to-transversion (Ti/Tv) ratio. Variants not passing the VQSR filter at the data-driven sensitivity level selected were removed using VCFtools¹⁴⁸ (v0.1.10) prior to co-segregation analyses.

Standard variant quality indicators were calculated on the resulting multi-sample VCF file, as well as on an individual sample basis, using the GATK VariantEvalWalker. The comparison file contained variants from dbSNP build 129, the build just prior to when the deposition of 1000G variants began. Quality metrics were checked for consistency with expectations. Then, genomic annotations were assigned to each variant, where applicable and where data was available, using SnpSift¹⁴⁹ (v3.1), SnpEff¹⁵⁰ (v2.0.5), and GATK's VariantAnnotator. Among these annotations were allele frequency in 1000G European samples, gene name, effect type (e.g. frameshift, non-synonymous coding, 3'-untranslated region (UTR), synonymous coding, etc.), impact (high impact = splice site acceptor, splice site donor, start lost, exon deleted, frameshift, stop gained, or stop lost), SIFT¹⁵¹ prediction score, and Polyphen-2¹⁵² predictions for variant impact on protein structure and function. PolyPhen-2 and SIFT were employed to illuminate potentially damaging variants based on predicted protein structure and between-species conservation^{151,152}. Variants considered to be potentially damaging were those with a SIFT score less than or equal to 0.05 and/or Polyphen-2 prediction outcome of possibly damaging (P) or probably damaging (D). In addition, the VCF was annotated by SnpSift using

dbSNP build 137 to highlight rare variants without minor allele frequency (MAF) $\geq 1\%$ in at least one of over 50 major populations (with ≥ 2 two unrelated carrying the minor allele).

4.2.5 Structural variant identification

To provide evidence that larger structural variation commonly driving syndromic CHD was not segregating in this family, we ran a Cytogenomic SNP Microarray (Affymetrix CytoScan HD[®] Array) on the proband's germline DNA in a CLIA-certified laboratory (ARUP Laboratories). In addition to clinical genetic testing, we evaluated structural variants that might co-segregate in this family but would not be detectable by the clinical microarray using BreakDancer¹⁵³.

4.2.6 Co-segregation analysis

The final SNV and indel variant calls from the parsed, multi-sample VCF were analyzed in R for co-segregation according to one of three genetic models that could theoretically be consistent with the pedigree under study: 1) autosomal dominant (AD) with incomplete penetrance, 2) AD with gonadal mosaicism in the father, and 3) autosomal recessive. Mosaicism refers to the existence of two or more genetically distinct groups of cells in a single individual that arise following mutation and cell division. If this phenomenon occurs in the sperm cell population, a genetic variant passed by a father to his offspring may only be present in the contributing subpopulation of his germ cells and undetectable (or below the threshold for detection) in his peripheral blood. For the AD model with incomplete penetrance, we considered variants that were a) carried with one copy by all 3 affected individuals and the unaffected father (initially ignoring the unaffected brother's genotype) or b) carried with one copy by high-coverage affected individuals (proband and half-sib with TOF) and the unaffected father (initially ignoring the unaffected brother's genotype). Here, we assumed that incomplete

penetrance needed to be invoked for the father but not necessarily the brother. However, upon generation of our final candidate list, an additional assumption was made that the variant was not present in the unaffected brother to avoid generating a candidate list of variants shared by the entire family. For the AD model with mosaicism, we identified variants with one copy that were a) shared by all 3 affected individuals and not shared by the unaffected father or brother (genotype calls allowed to be missing) or b) shared by high-coverage affected individuals and not shared by the unaffected father or brother (genotype calls allowed to be missing). Finally, for the autosomal recessive pattern of inheritance, we aggregated variants with two copies shared by the proband and high-coverage affected with TOF that were also present with 1 copy in the unaffected father. In all three scenarios, we allowed for the possibility that a variant may not be called even if present in the half-sib with truncus arteriosus due to low sequencing depth of coverage.

Considering each genetic model separately, we next restricted the respective candidate variant lists to markers with $MAF \leq 1\%$ (for AD models) or $\leq 10\%$ (for AR model), and subsequently, to those in coding regions. A lenient 10% threshold was selected for the AR model because this approximated the expected frequency that would be consistent with the allele frequencies of a Hardy Weinberg Equilibrium $q^2 (p^2 + 2pq + q^2 = 1)$ approximating a disease affecting 1 in 100 live births. Rare variants co-segregating according to one of these models and predicted to be potentially damaging by SIFT and/or Polyphen-2 were considered causal candidates.

Further, for each genetic model, we carefully examined variants falling under previously identified linkage peaks (6q23.3, 10q22.1, 7q31.2, 11q22.1, 12q13.1, 14q23.2, 20q12) or within 320 candidate gene regions (

Table 4.1). We relaxed our allele frequency filters to $MAF \leq 10\%$ for all three models and examined each coding variant manually. Resulting SNVs and indels from the genome-wide and candidate region analyses for each model were collapsed into a single candidate list of variants that could potentially alter gene function but required further validation. For the final variant list, an updated version of SIFT and PROVEAN¹⁵⁴ (a new prediction algorithm similar to SIFT but allowing for the annotation of indels) were run to obtain more accurate functional prediction scores.

Table 4.1. Candidate gene list for complex CHD.

<i>ACTC1</i>	<i>DVL2</i>	<i>GATA6</i>	<i>KIT</i>	<i>MYST1</i>	<i>RARA</i>	<i>SMAD9</i>	<i>TNNI1</i>
<i>ACVR1</i>	<i>DYNC1H1</i>	<i>GDF1</i>	<i>KITLG</i>	<i>NEUROD1</i>	<i>RARB</i>	<i>SMARCD3</i>	<i>TNNI2</i>
<i>ACVR2B</i>	<i>ECE1</i>	<i>GJA1</i>	<i>KRIT1</i>	<i>NF1</i>	<i>RARG</i>	<i>SMO</i>	<i>TNNI3</i>
<i>ADAM17</i>	<i>EDN1</i>	<i>GLI2</i>	<i>LBX1</i>	<i>NFATC1</i>	<i>RELA</i>	<i>SMYD1</i>	<i>TTC8</i>
<i>ALDH1A2</i>	<i>EDNRA</i>	<i>GLI4</i>	<i>LEF1</i>	<i>NFATC2</i>	<i>RORC</i>	<i>SOS1</i>	<i>TTN</i>
<i>APC</i>	<i>EGFR</i>	<i>GSC</i>	<i>LEFTY1</i>	<i>NFATC3</i>	<i>RXRA</i>	<i>SOX4</i>	<i>TWIST1</i>
<i>APEX1</i>	<i>EP300</i>	<i>GSK3A</i>	<i>LEFTY2</i>	<i>NFATC4</i>	<i>RXRB</i>	<i>SOX7</i>	<i>TWIST2</i>
<i>ATP1A1</i>	<i>EPHB2</i>	<i>GSK3B</i>	<i>LRRC10</i>	<i>NKX2-5</i>	<i>RXRG</i>	<i>SOX9</i>	<i>UGDH</i>
<i>ATP1A2</i>	<i>EPO</i>	<i>HAND1</i>	<i>MAML1</i>	<i>NK3-2</i>	<i>SALL4</i>	<i>SOX10</i>	<i>VANGL1</i>
<i>AXIN1</i>	<i>ERBB2</i>	<i>HAND2</i>	<i>MAML2</i>	<i>NODAL</i>	<i>S1PR2</i>	<i>SOX18</i>	<i>VANGL2</i>
<i>BBS4</i>	<i>ERBB3</i>	<i>HAS1</i>	<i>MAML3</i>	<i>NOG</i>	<i>SEMA3A</i>	<i>SRF</i>	<i>VCAM1</i>
<i>BBS5</i>	<i>ETS1</i>	<i>HAS2</i>	<i>MAP3K7IP1</i>	<i>NOTCH1</i>	<i>SEMA3B</i>	<i>STIL</i>	<i>VCAN</i>
<i>BBS7</i>	<i>EVX1</i>	<i>HDAC1</i>	<i>MED13L</i>	<i>NOTCH2</i>	<i>SEMA3C</i>	<i>SUPT5H</i>	<i>VEGFA</i>
<i>BMP1</i>	<i>FGF10</i>	<i>HDAC2</i>	<i>MEF2A</i>	<i>NR2F2</i>	<i>SEMA3D</i>	<i>SUPT6H</i>	<i>VEGFB</i>
<i>BMP2</i>	<i>FGF4</i>	<i>HDAC3</i>	<i>MEF2B</i>	<i>NRG1</i>	<i>SEMA3E</i>	<i>TAL1</i>	<i>VEGFC</i>
<i>BMP4</i>	<i>FGF8</i>	<i>HDAC4</i>	<i>MEF2C</i>	<i>NRP1</i>	<i>SEMA3F</i>	<i>TAZ</i>	<i>WIF</i>
<i>BMP5</i>	<i>FIGF</i>	<i>HDAC5</i>	<i>MESP1</i>	<i>NRP2</i>	<i>SEMA3G</i>	<i>TBX1</i>	<i>WNT1</i>
<i>BMP6</i>	<i>FLT4</i>	<i>HDAC6</i>	<i>MESP2</i>	<i>NTF3</i>	<i>SEMA4A</i>	<i>TBX10</i>	<i>WNT10A</i>
<i>BMP7</i>	<i>FN1</i>	<i>HDAC7</i>	<i>MIB1</i>	<i>NTRK3</i>	<i>SEMA4B</i>	<i>TBX15</i>	<i>WNT10B</i>
<i>BMP10</i>	<i>FOLR1</i>	<i>HDAC8</i>	<i>MIRN1-1</i>	<i>NUMBL</i>	<i>SEMA4C</i>	<i>TBX18</i>	<i>WNT11</i>
<i>BMPR1A</i>	<i>FOXC1</i>	<i>HDAC9</i>	<i>MIRN1-2</i>	<i>PAX3</i>	<i>SEMA4D</i>	<i>TBX19</i>	<i>WNT16</i>
<i>BNIP3L</i>	<i>FOXC2</i>	<i>HEG1</i>	<i>MKKS</i>	<i>PCAF</i>	<i>SEMA4E</i>	<i>TBX2</i>	<i>WNT2</i>
<i>CACNA1C</i>	<i>FOXH1</i>	<i>HEY1</i>	<i>MKL1</i>	<i>PCSK6</i>	<i>SEMA4F</i>	<i>TBX20</i>	<i>WNT2B</i>
<i>CCM2</i>	<i>FOXJ1</i>	<i>HHEX</i>	<i>MKL2</i>	<i>PDGFRA</i>	<i>SEMA4G</i>	<i>TBX21</i>	
<i>CER1</i>	<i>FOXP1</i>	<i>HIF1A</i>	<i>MEOX2</i>	<i>PHC1</i>	<i>SEMA5A</i>	<i>TBX22</i>	
<i>CFC1</i>	<i>FOXP2</i>	<i>HOPX</i>	<i>MSX1</i>	<i>PITX2</i>	<i>SEMA5B</i>	<i>TBX3</i>	
<i>CHD7</i>	<i>FOXP3</i>	<i>HOXA3</i>	<i>MSX2</i>	<i>PLXNA1</i>	<i>SEMA6A</i>	<i>TBX4</i>	
<i>CHRD</i>	<i>FOXP4</i>	<i>HSPG2</i>	<i>MTHFR</i>	<i>PLXNA2</i>	<i>SEMA6B</i>	<i>TBX5</i>	
<i>CITED2</i>	<i>FURUIN</i>	<i>IFT52</i>	<i>MYBPC3</i>	<i>PLXNA3</i>	<i>SEMA6C</i>	<i>TBX6</i>	
<i>CREBBP</i>	<i>FZD1</i>	<i>IGF1R</i>	<i>MYC</i>	<i>PLXNA4</i>	<i>SEMA6D</i>	<i>TEAD1</i>	
<i>CRELD1</i>	<i>FZD10</i>	<i>IHH</i>	<i>MYCN</i>	<i>PLXNB1</i>	<i>SEMA7A</i>	<i>TEK</i>	
<i>CRKL</i>	<i>FZD2</i>	<i>IRX3</i>	<i>MYH6</i>	<i>PLXNB2</i>	<i>SHH</i>	<i>TFAP2B</i>	
<i>CSRP1</i>	<i>FZD3</i>	<i>IRX4</i>	<i>MYH7</i>	<i>PLXNB3</i>	<i>SMAD1</i>	<i>TGFA</i>	
<i>CTNNB1</i>	<i>FZD4</i>	<i>ISL1</i>	<i>MYL2</i>	<i>PLXNC1</i>	<i>SMAD2</i>	<i>TGFB1</i>	
<i>CXADR</i>	<i>FZD5</i>	<i>ISL2</i>	<i>MYL3</i>	<i>PLXND1</i>	<i>SMAD3</i>	<i>TGFB2</i>	
<i>DHFR</i>	<i>FZD6</i>	<i>JAG1</i>	<i>MYL4</i>	<i>PSEN1</i>	<i>SMAD4</i>	<i>TGFB3</i>	
<i>DKKL1</i>	<i>FZD7</i>	<i>JUN</i>	<i>MYL7</i>	<i>PTK2</i>	<i>SMAD5</i>	<i>TGFBR1</i>	
<i>DLC1</i>	<i>FZD9</i>	<i>KDR</i>	<i>MYLK2</i>	<i>PTPN11</i>	<i>SMAD5OS</i>	<i>TGFBR2</i>	
<i>DLL1</i>	<i>GATA4</i>	<i>KHDRBS1</i>	<i>MYLK3</i>	<i>PTPN21</i>	<i>SMAD6</i>	<i>TGFBR3</i>	
<i>DVL1</i>	<i>GATA5</i>	<i>KIF3A</i>	<i>MYOCD</i>	<i>RAF1</i>	<i>SMAD7</i>	<i>TM7SF2</i>	

4.2.7 Sanger sequencing

In our initial analysis, we attempted to validate all variants that co-segregated according to one of our 3 pre-specified genetic models and that met the genome-wide filtering criteria described in 4.2.6. The goal was to validate variants using a technique independent of NGS to eliminate false-positive results and yield a list of genes and/or pathways to carry forward into future functional and population-based studies. Primer sets were designed for polymerase chain reaction (PCR) amplification of the 28 identified regions of interest. We attempted to validate ten variants by Sanger sequencing first in the half-sib with truncus arteriosus prior to moving onto the whole sample set, as calls were absent for this individual (likely due to low coverage). For another 18 variants, direct sequencing was conducted in either the proband and half-sib with truncus arteriosus or the full complement of 5 samples.

4.2.8 Quality control

We employed quality control (QC) methods to verify sample purity represented in our sequencing reads and to rule out the presence of sample contamination. The VerifyBamID tool was employed to detect whether reads in a given indexed BAM file were actually generated based on reads from a mixture of more than one individual sample¹⁵⁵. The tool's FreeMix parameter measured the percent contamination in a BAM file, with greater than or equal to 0.03 serving as the threshold for evidence of contamination.

4.2.9 Sequencing quality and contamination

After completion of the full analysis pipeline, including attempts to validate by Sanger sequencing, we detected consistent signs of variant call discrepancies across sequencing methods. Although validation rates for NGS are typically low, the validation of some WGS-called variants by Sanger sequencing in the “incorrect” individual indicated the potential for

sample contamination or swap. Further, identity by descent (IBD) estimates calculated using VCFtools and PLINK⁸¹ indicated that the relationships between individuals according to genotype data were not consistent with the verified pedigree. For example, the proband and half-sib with TOF appeared to be more consistent with full-siblings instead of half-siblings. The VerifyBamID FreeMix estimates for all reads aggregated according to their initial sample assignment indicated the presence of read contamination in the half-sib with TOF (Table 4.2). Each individual lane generated for the half-sib with TOF was subsequently investigated, but FreeMix estimates indicated no within-lane contamination from a second sample (Table 4.3). All 4 lanes from the GA-II when considered together were uncontaminated (FreeMix = 0.0082), and all 3 lanes from the Hi-Seq when considered together were uncontaminated (FreeMix = 0.0031). However, based on clues from IBD analysis and Sanger sequencing results on select variants with apparent swaps, we combined GA-II lanes from the TOF half-sib with Hi-Seq lanes run on the “unaffected brother”, yielding no contamination (FreeMix = 0.0053). Thus, we concluded that the sample labels were swapped between these two individuals when supplementary sequencing at USC was completed. We later confirmed with the sequencing facility at USC that this bioinformatically-identified sample swap had indeed occurred in the lab but had not been reported to our group. After the sample mislabeling issue was discovered and reads from the appropriate lanes were reassigned to the correct individuals, the full pipeline from 4.2.3 through 4.2.6 (with the exception of alignment) was re-run, and the results presented below are based on the corrected read assignments.

Table 4.2. VerifyBamID FreeMix estimates for each individual with WGS reads combined across sequencing lanes.

Sample – data combined across lanes	FreeMix
Proband (HLHS)	0.0060
Unaffected father	0.0028
Unaffected brother	0.0031
Half-sib (truncus arteriosus)	0.0079
Half-sib (TOF)	0.2565

Table 4.3. VerifyBamID FreeMix estimates for each lane of WGS reads generated for the half-sib with TOF. GA-II/UM = Illumina Genome Analyzer II/University of Michigan. Hi-Seq/USC = Illumina Hi-Seq/University of Southern California.

Lane	Sequencing	FreeMix
1	GA-II/UM	0.0066
2	GA-II/UM	0.0106
3	GA-II/UM	0.0076
4	GA-II/UM	0.0069
5	Hi-Seq/USC	0.0034
6 (failed lane)	Hi-Seq/USC	NA
7	Hi-Seq/USC	0.0033
8	Hi-Seq/USC	0.0035

4.3 Results

4.3.1 WGS data summary

Paired-end WGS reads were generated for the proband with HLHS (total average depth of genome coverage ~36X); her affected half-sisters with TOF (~38X) and truncus arteriosus (~10X), respectively; her unaffected full brother (~30X); and their shared unaffected father

(~33X). When considering data from both platforms combined, average depths of coverage across the exome and across the genome for each sample were calculated using GATK's DepthOfCoverage tool (Table 4.4)¹⁴⁶.

Table 4.4. Whole-genome sequence data summary for a family affected by complex CHD. % of reads aligned refers to the percentage of all generated reads per sample that were successfully aligned to the human genome reference sequence.

	Affected status	Sequencing Platform	Total # of reads (% of reads aligned)	Exome coverage	Genome coverage
Proband	Affected (HLHS)	GA-II + HiSeq	1,167,587,159 (85.4%)	30.31	36.24
Half-sister	Affected (TOF)	GA-II + HiSeq	1,193,994,664 (89.2%)	31.34	38.56
Half-sister	Affected (Truncus arteriosus)	GA-II	265,932,629 (89.0%)	7.59	9.76
Full brother	Unaffected	HiSeq	929,212,340 (91.6%)	24.07	29.57
Father	Unaffected	HiSeq	1,041,630,573 (91.0%)	26.57	32.95

4.3.2 Variant summary: SNVs, indels, and structural variants

The number of variant calls across four novel false-discovery rate levels resulting from VQSR filtering was examined, and a truth sensitivity level of 99% was selected for generating the final variant call set. This decision was made to limit false positives and to match the novel Ti/Tv ratio as closely as possible to our expectations for WGS data (Figure 4.2). A summary of VQSR-filtered variant calls suggested high concordance with variants in dbSNP build 129, Ti/Tv ratios near the expected 2.1 level for WGS data, and heterozygous to non-reference homozygous genotype ratios in the typical 1.8-1.9 range (

Table 4.5). Between 4.1 and 4.2 million variants were present in each individual sample, with the exception of the half-sib with truncus arteriosus who had fewer calls (~3.7 million). This was anticipated due to lower sequencing depth. As expected, the number of variants identified declined with increasing chromosome number, corresponding to cytogenetic ordering of the autosomes by size (example for proband in Figure 4.3).

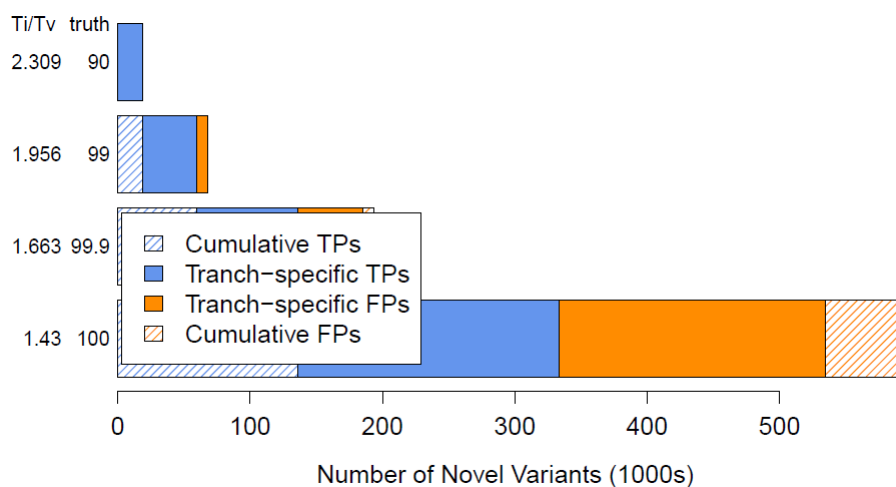


Figure 4.2. Tranche (or truth sensitivity)-specific results from variant quality score recalibration. X-axis = Total number of novel variants called. Y-axis: novel Ti/Tv ratio and overall truth sensitivity. TP = true positive. FP = false positive.

Table 4.5. SNP and small indel summary for the overall multi-sample VCF and for the per individual call sets.

Sample	Called Loci	Concordance Rate	Ti/Tv Ratio	HetHom Ratio
Multi-sample	5,776,723 (4,619,827 SNPs; 916,577 indels)	96.27%	2.13	1.77
Proband	4,212,911 (3,254,953 SNPs; 724,353 indels)	95.28%	2.13	2.01
Half-sib (TOF)	4,159,646 (3,198,101 SNPs; 726,254 indels)	95.22%	2.13	1.90
Half-sib (Truncus arteriosus)	3,676,053 (3,044,213 SNPs; 467,835 indels)	96.25%	2.12	1.33
Unaffected brother	4,180,504 (3,201,208 SNPs; 742,436 indels)	95.2%	2.14	1.85
Unaffected father	4,144,233 (3,167,282 SNPs; 739,814 indels)	95.17%	2.14	1.83

Concordance rate = the percentage of variants with a matching alternate allele when the locus is shared between the file under evaluation and the comparison file from dbSNP. Ti/Tv ratio = transition-to-transversion ratio. HetHom ratio = ratio of heterozygous genotypes to non-reference homozygous genotypes.

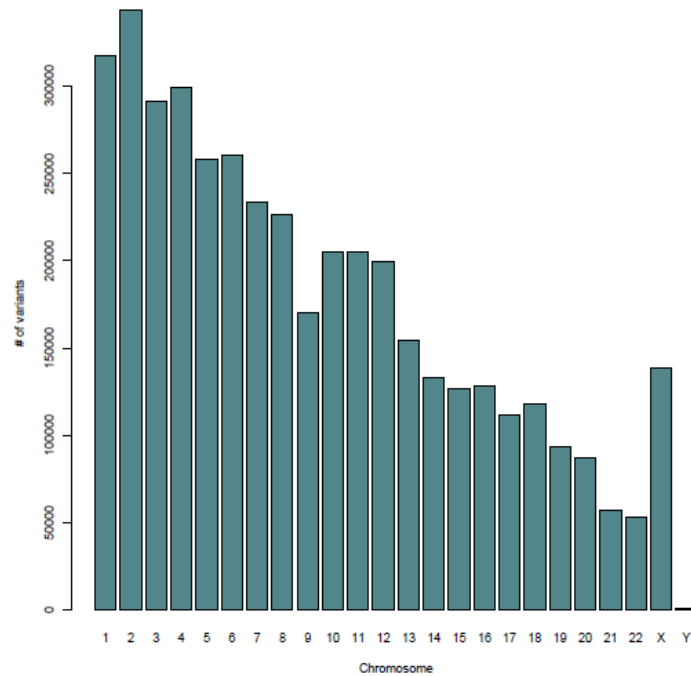


Figure 4.3. Total number of SNVs and indels identified by chromosome for the proband with HLHS.

After filtering based on truth sensitivity level, there were a maximum of 2,508,045 shared variants between all three affected offspring (Figure 4.4). The proportion of all variants that were novel or rare (MAF \leq 1%) ranged from 2.4% to 5.3% across the three individuals. Total variant counts in this figure did not match exactly with those reported in

Table 4.5 because variants from unlocalized sequences (i.e. “GL” variants with no chromosome assignment) were removed prior to co-segregation analysis.

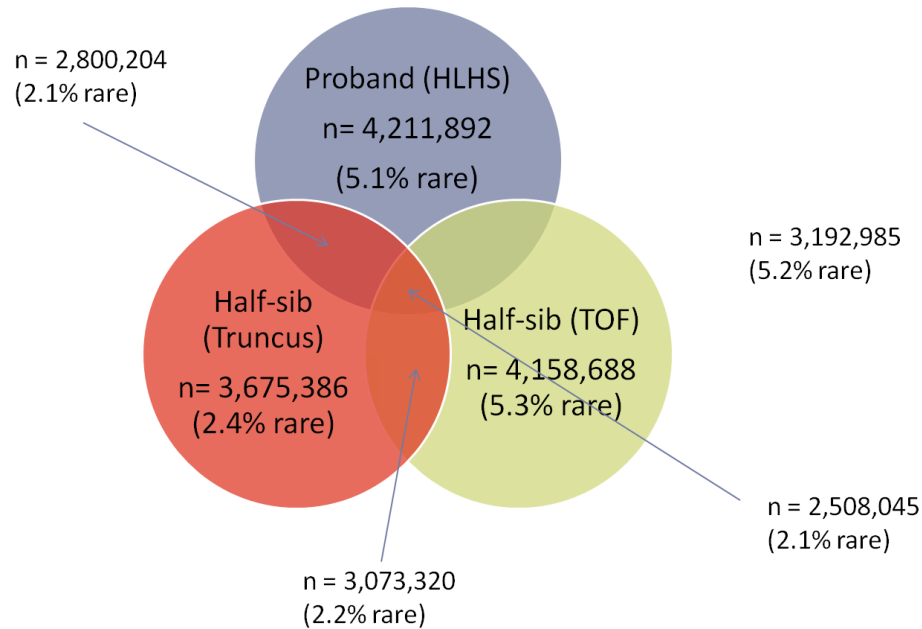


Figure 4.4. SNVs and indels shared by three affected individuals with severe CHD. Rare = novel or $MAF \leq 1\%$.

In terms of structural variation, cytogenomic SNP microarray test results on the proband showed no clinically significant chromosomal abnormalities. Due to DNA quality issues, no deletion calls $<100\text{kb}$ or duplication calls $<400\text{kb}$ were made. BreakDancer results were uninformative.

4.3.3 Co-segregation analysis

Ignoring genotypes from the low-coverage, half-sib with truncus arteriosus, we visualized sharing of more confident variant calls in the remainder of the family (Figure 4.5). Red stripes represent potentially causative variants that co-segregate according to an AD model with incomplete penetrance (i.e. shared by high-coverage affecteds and the unaffected father but not the unaffected brother). Purple stripes denote potentially causative variants that co-segregate according to an AD model with gonadal mosaicism (i.e. shared by high-coverage affecteds but not shared by the father or brother).

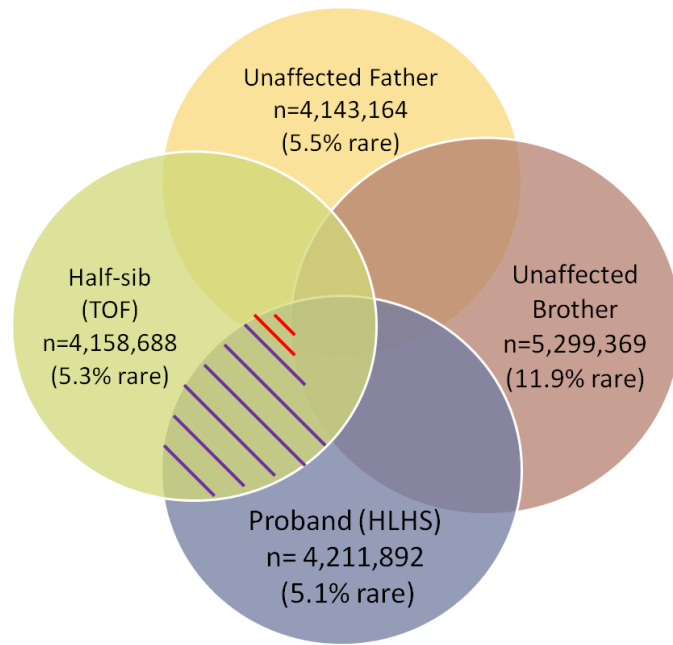


Figure 4.5. Visualization of SNVs and indels that follow AD patterns of inheritance. Red = AD with incomplete penetrance. Purple = AD with gonadal mosaicism. Rare = novel or MAF \leq 1%.

When considering variants segregating according to an AD model with incomplete penetrance in the father, we initially identified nearly 3 million candidate variants (Figure 4.6). However, this number was reduced to less than 500,000 when a stringent MAF filter of 1% was applied. Only about 0.6% of these remaining variants fell in coding regions, and among those, approximately 2.8% were predicted to be damaging.

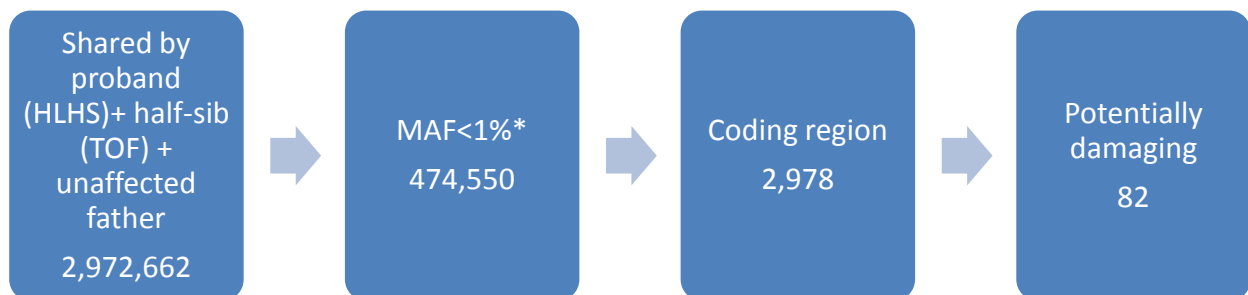


Figure 4.6. SNVs and small indels co-segregating with disease according to an autosomal dominant genetic model with incomplete penetrance. * = in at least 1 of over 50 major populations in dbSNP with at least 2 individuals carrying the minor allele or in HapMap European samples from Utah.

When restricting the analysis to previously identified linkage peaks and candidate genes but with a relaxed variant frequency filter, the analysis yielded an additional 399 variants falling in coding regions. This list of 82 potentially damaging variants plus 399 candidate region variants was further reduced by eliminating overlap between these two lists, removing variants present in the unaffected brother, and filtering out variants present with two copies. If the variant is sufficiently rare to cause this rare phenotype in an AD fashion, it is unlikely to be seen twice in the same individual.

Examining variants that segregate with CHD according to an AD model with gonadal mosaicism in the father, we observed 57,514 variants shared by the high coverage affecteds but absent in the unaffected family members (Figure 4.7). Only about 6,000 SNPs and indels were novel or rare, and of these, 30 fell in coding regions. None of these variants were predicted to be damaging by SIFT and/or Polyphen-2. Because none were predicted as damaging, we considered all coding variants co-segregating according to this model as potential candidates. When we relaxed the frequency filter and examined previously identified candidate regions, the analysis yielded an additional 10 variants falling in coding regions. The list of 30 coding variants plus 10 candidate region variants was further reduced by eliminating overlap between these two lists and variants present with two copies in a single individual.

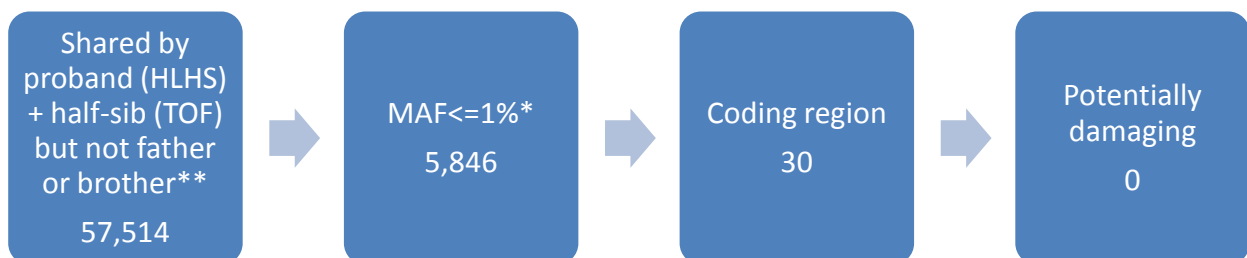


Figure 4.7. SNVs and small indels co-segregating with disease according to an autosomal dominant with gonadal mosaicism model. * = in at least 1 of 50 major populations with at least 2

individuals with the minor allele or in HapMap European samples from Utah. **Unaffected brother and father were allowed to have missing calls.

An autosomal recessive pattern of inheritance was the final genetic model considered.

Approximately 900,000 variants had two copies of the alternate allele in high-coverage affecteds and 1 copy of the alternate allele in the unaffected father (Figure 4.8). After applying MAF and coding region filters, there were 120 unique variants that were predicted by SIFT and/or Polyphen-2 to be possibly or probably damaging. When restricting the analysis to published linkage peaks and candidate genes, 157 more variants falling in coding regions were identified. The list of 120 potentially damaging plus 157 candidate region variants was further reduced by eliminating overlap between these two lists, removing variants present in the unaffected brother, and variants present with two copies in the father and/or brother.

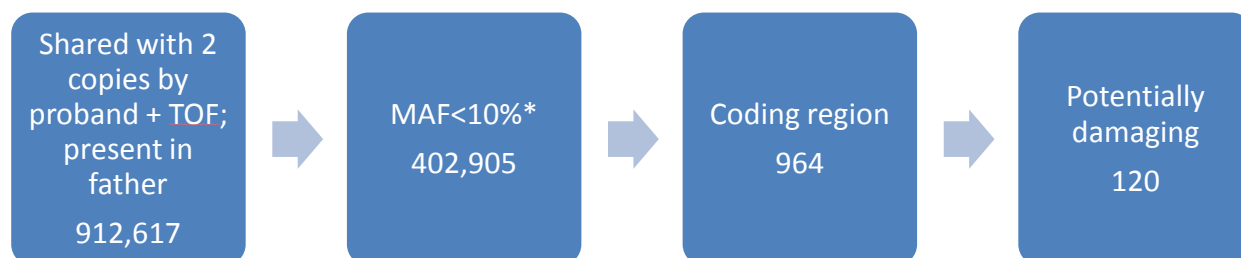


Figure 4.8. SNVs and small indels co-segregating with disease according to an autosomal recessive model. * = in at least 1 of 50 major populations with at least 2 individuals with the minor allele or in HapMap European samples from Utah. **Unaffected brother and father were allowed to have missing calls.

The final summary of all 32 co-segregating variants across the three genetic models is compiled in Table 4.6. This table assumes that the low-coverage, half-sib with truncus arteriosus either has a genotype call that's consistent with the genetic model of interest or a missing call. Of these 32 variants, 8 were predicted to be damaging by at least 1 of the 3 algorithms employed. A

secondary list of candidate variants can be found in Table 4.7, where variant calls for only the low-coverage, half-sib with truncus arteriosus do not fit with the specified genetic model. While not of primary interest, these variants could also be critical contributors to the list for validation, as our call set for this individual is relatively unreliable. In both tables, a variant on the X chromosome appeared to segregate according to an autosomal dominant model with mosaicism. In the case of rs201522041 (Table 4.6), this is because the alternate allele was called to be present with one copy for the proband (female) and half-sib with TOF (female), with two copies for the half-sib with truncus arteriosus (female), and with zero copies for the unaffected father (a call was missing for the unaffected brother). For rs2230488 (Table 4.7), the proband and half-sib with TOF were called to have one copy of the variant allele, whereas the half-sib with truncus arteriosus as well as the unaffected father and brother showed no copies.

Table 4.6. Summary of potentially causative SNVs and small indels co-segregating according to an autosomal dominant (AD) or autosomal recessive (AR) model. Variants in red are predicted to be damaging by SIFT, PROVEAN, and/or Polyphen-2 (PP2).

CHR	POSITION	Variant	REF	ALT	CEU MAF	GENE	EFFECT	SIFT	PROVEAN	PP2	Genetic Model	Cand. region
1	16333142	rs34643567 rs146554621	GAATA	G	0.00	<i>C1orf64</i>	3'-UTR	NA	NA	NA	AD mosaicism	No
1	109465165	rs35029887	ACTT	A	0.00	<i>GPSM2</i>	DELETION	NA	N,N,D	NA	AD mosaicism	No
1	152975715	rs63405761	C	T	0.00	<i>SPRR3</i>	SYN	1	N	NA	AD mosaicism	No
1	172502479	rs2285144	T	C	0.00	<i>C1orf9</i>	SYN	1	N	NA	AD mosaicism	No
2	161350080	.	C	T	NA	<i>RBMS1</i>	START GAIN	NA	NA	NA	AD mosaicism	No
3	30656065	rs56402737	C	T	0.08	<i>TGFBR2</i>	START GAIN	NA	NA	NA	AD mosaicism	Yes
3	125286291	.	A	C	NA	<i>OSBPL1</i>	NON SYN	0.006	N	P	AD inc pen	No
6	117622233	rs529038	C	T	0.00	<i>ROS1</i>	NON SYN	0.28	N	B	AD mosaicism	No
8	22570773	.	AT	A	NA	<i>PEBP4</i>	3'-UTR	NA	NA	NA	AD mosaicism	No
10	70101417	rs16925347	G	C	0.09	<i>HNRNPH3</i>	NON SYN	0.41	N	B	AD inc pen	Yes
10	70405541	rs72799515	A	G	0.04	<i>TET1</i>	NON SYN	0.001	N	B	AD inc pen	Yes
10	70856963	rs67852477	C	G	0.01	<i>SRGN</i>	NON SYN	0.12	D	NA	AD inc pen	Yes
10	72520738	rs16927931	G	A	0.05	<i>ADAMTS14</i>	3'-UTR	NA	NA	NA	AD inc pen	Yes
10	72520835	rs41278004	C	T	0.09	<i>ADAMTS14</i>	3'-UTR	NA	NA	NA	AD inc pen	Yes
10	73059104	rs74761108	G	A	0.03	<i>UNC5B</i>	3'-UTR	NA	NA	NA	AD inc pen	Yes
10	73059407	rs58848310	C	CT	NA	<i>UNC5B</i>	3'-UTR	NA	NA	NA	AR	Yes
10	73062534	rs2275578	A	G	0.03	<i>UNC5B</i>	3'-UTR	NA	NA	NA	AD inc pen	Yes
11	58949518	.	C	T	NA	<i>DTX4</i>	NON SYN	0.002	N	P	AD inc pen	No
11	60567525	rs71997819	CCTGGG	C	0.00	<i>MS4A10</i>	3'-UTR	NA	NA	NA	AD mosaicism	No
11	61908440	rs2277283	T	C	0.00	<i>INCENP</i>	NON SYN	0.03	D	P,D,D	AD inc pen	No
11	89608808	.	GC	G	NA	<i>TRIM64B</i>	FRAME SHIFT	NA	NA	NA	AD mosaicism	No
13	37420610	rs146502953	CAA	C	NA	<i>SMAD9</i>	3'-UTR	NA	NA	NA	AD inc pen	Yes
15	96881264	rs34353956	C	CAACA	NA	<i>NR2F2</i>	3'-UTR	NA	NA	NA	AD mosaicism	Yes
16	50267603	rs11319165	TA	T	0.00	<i>PAPD5</i>	3'-UTR	NA	NA	NA	AD mosaicism	No
19	36435928	rs79027052	GC	G	0.00	<i>LRFN3</i>	3'-UTR	NA	NA	NA	AD mosaicism	No
19	40882548	.	G	A	NA	<i>PLD3</i>	NON SYN	0.17	N	P	AD inc pen	No
19	44006360	.	C	G	NA	<i>PHLDB3</i>	NON SYN	0.06	N	P	AD inc pen	No
19	45316804	rs3810141	C	T	0.00	<i>BCAM</i>	SYN	1	N	NA	AD mosaicism	No
19	50045878	rs34654230	C	T	0.00	<i>RCN3</i>	NON SYN	0.02	D	B	AD inc pen	No
21	43805637	rs2839501	C	T	0.00	<i>TMPRSS3</i>	SYN	1	N	NA	AD mosaicism	No
22	19748690	rs72646953	G	A	0.02	<i>TBX1</i>	SYN	1	N	NA	AD mosaicism	Yes
X	84343215	rs201522041	TATAA	T	0.00	<i>APOOL</i>	SPLICE SITE ACCEPTOR	NA	NA	NA	AD mosaicism	No

Missing ID = variant has MAF <1% in all dbSNP populations. REF = reference allele. ALT = alternate allele. CEU MAF = minor allele frequency in CEPH Utah HapMap samples. SIFT = updated SIFT scores where <0.05 indicates a damaging prediction. PROVEAN: N = neutral, D = damaging. PP2: D = probably damaging, P = possibly damaging; B = benign. Cand. region = previously identified linkage peak or candidate gene.

Table 4.7. Summary of potentially causative SNVs and small indels co-segregating according to an autosomal dominant (AD) or autosomal recessive (AR) model *without consideration of half-sib with truncus arteriosus' genotype due to low depth of coverage*. Variants in red are predicted to be damaging by SIFT, PROVEAN, and/or Polyphen-2 (PP2).

CHR	BP	Variant	REF	ALT	CEU MAF	GENE	EFFECT	SIFT	PROVEAN	PP2	Genetic Model	Cand. Region
1	100376325	rs12043139	G	A	0.00	AGL	NON SYN	0.07	N	B,B,B	AD inc pen	No
1	119427467	rs61730011	A	C	0.09	TBX15	NON SYN	0.02	D	NA	AD inc pen	Yes
2	65299330	rs141499084	T	A	0.01	CEP68	NON SYN	0.01	N	P	AD inc pen	No
2	128389835	rs190699169	G	A	0.00	MYO7B	NON SYN	0.005	D	B	AD inc pen	No
2	131218821	.	TC	T	NA	POTE1	3'-UTR	NA	NA	NA	AD mosaicism	No
2	140989650	rs200444225 rs150474267	GTTC	G	0.00	LRP1B	3'-UTR	NA	NA	NA	AD mosaicism	No
2	179418318	.	T	C	NA	TTN	NON SYN	0.39	D	B	AD inc pen	Yes
2	179549131	rs116676813	C	T	0.04	TTN	NON SYN	1	N	B	AD inc pen	Yes
2	206641239	.	TCGCA	T	NA	NRP2	FRAME SHIFT	NA	NA	NA	AD inc pen	Yes
2	206641245	rs200483574	T	TA	0.00	NRP2	FRAME SHIFT	NA	NA	NA	AD inc pen	Yes
3	195453017	rs144288174	GC	G	0.00	MUC20	FRAME SHIFT	NA	NA	NA	AD mosaicism	No
4	36068192	.	TA	T	NA	ARAP2	3'-UTR	NA	NA	NA	AD mosaicism	No
5	115811273	rs17432496	C	T	0.06	SEMA6A	NON SYN	0.47	N	NA	AD inc pen	Yes
5	178510271	rs147516432	GAT	G	0.00	ZNF354C	3'-UTR	NA	NA	NA	AD mosaicism	No
6	135523805	.	C	T	NA	MYB	NON SYN	0	N	NA	AD inc pen	Yes
6	137519097	rs55665036	C	A	0.00	IFNGR1	3'-UTR	NA	NA	NA	AD inc pen	Yes
8	28428141	.	CA	C	NA	FZD3	3'-UTR	NA	NA	NA	AD inc pen	Yes
10	72015573	rs3812694	T	G	0.08	NPFRI	NON SYN	0.006	D	D	AD inc pen	Yes
11	27517042	.	CA	C	NA	LIN7C	3'-UTR	NA	NA	NA	AD mosaicism	No
11	56019973	.	G	T	NA	OR5T3	NON SYN	0	D	D	AD inc pen	No
13	28499583	rs139276646	CCCCT CCTCT	C	0.00	PDX1	3'-UTR	NA	NA	NA	AD mosaicism	No
13	37419247	rs57525591	GTGTGTA	G	NA	SMAD9	3'-UTR	NA	NA	NA	AD mosaicism	Yes
14	23999627	rs35092523	TA	T	0.00	ZFH2	FRAME SHIFT	NA	NA	NA	AD mosaicism	No
16	338189	rs34015754	C	T	0.02	AXINI	NON SYN	0.12	N	P	AD inc pen	Yes
19	11459665	rs78738753	G	GA	0.00	CCDC159	FRAME SHIFT	NA	NA	NA	AD mosaicism	No
19	17932190	rs1047233	T	C	0.00	INSL3	SYN	1	N	NA	AD mosaicism	No
19	35785147	rs56138004	C	T	0.00	MAG	SYN	1	N	NA	AD mosaicism	No
20	3627840	rs11481874	C	CA	0.00	ATRN	3'-UTR	NA	NA	NA	AD mosaicism	No
X	20204461	rs2230488	G	T	0.00	RPS6KA3	SYN	1	N	NA	AD mosaicism	No

Missing ID = variant has MAF <1% in all dbSNP populations. REF = reference allele. ALT = alternate allele. CEU MAF = minor allele frequency in CEPH Utah HapMap samples. SIFT = updated SIFT scores where <0.05 indicates a damaging prediction. PROVEAN: N = neutral, D = damaging. PP2: D = probably damaging, P = possibly damaging; B = benign. Cand. region = previously identified linkage peak or candidate gene

4.4 Discussion

This dissertation chapter presented an initial exploration into the genetic etiology underlying severe CHD in a multiplex family affected by both HLHS and conotruncal lesions. We leveraged WGS on affected and unaffected family members to examine the potential for a rare mutation in a single gene to co-segregate with three anatomically distinct CHD phenotypes. The heterogeneous etiologic landscape contributing to CHD development remains under-characterized, and by studying this family in depth, we had the opportunity to investigate the hypothesis that some cases could be driven by a rare variant of large effect in a gene important for early cardiogenesis. While the population-wide heritability of complex CHD is high, this fact is less relevant for a pedigree with clear evidence for familial clustering and a strong genetic component. If a single variant is indeed responsible for disease in this family, that gene, gene product, or regulatory element would deserve attention as an important component in the process of heart development prior to separation of the left and right chambers. The goals of this study were two-fold: 1) to generate a list of candidate variants that may be responsible for CHD in this family, and 2) to provide proof of principle that this study design has potential to highlight genes and/or pathways for study of these phenotypes at the population level.

This study identified a list of 32 variants in 29 genes that might contribute to the genetic landscape underlying severe CHD etiology for follow-up studies at the bench and in an epidemiologic setting. This list includes another 29 variants in 27 genes if flexibility in genotypes called for the low-coverage affected with truncus arteriosus was allowed. In generating this list of co-segregating variants, we considered the possibility of three genetic models that were consistent with the pedigree. Whether or not CHD in this family is driven by variation at a single locus remains to be determined, but if it is, the most plausible pattern of inheritance is autosomal dominant. Autosomal recessive inheritance is highly unlikely given that

a causal variant would have to be carried by two unrelated mothers and the father; nonetheless, it is a theoretically viable model that deserved attention. Among SNVs and indels in the variant lists generated, the most plausible functional candidates are those co-segregating according to one of the two AD models and those predicted to be damaging by at least one bioinformatic algorithm. Further, although 3'-UTR and synonymous coding variants still have the potential for impactful functional consequences, frame shift and non-synonymous coding variants would have undeniably simpler interpretations in relation to affecting gene expression and/or protein function. Regardless of these speculative criteria for prioritization, all 61 of them are viable causal candidates. It is important to note that only a handful of these variants have been previously implicated in heart development pathways. This finding was encouraging, indicating that we have prioritized several novel genes and/or pathways for CHD development by studying this family.

Ultimately, we generated a candidate list that is too large to individually validate variants with extremely limited DNA resources for key affected family members, and family dynamics preclude additional sample acquisition. We have not yet validated any of these variants by Sanger sequencing or another array-based technology due to limitations regarding DNA sample availability. Some resources were exhausted through Sanger sequencing in our original analysis that highlighted a sample swap problem. The proband with HLHS is now deceased, and additional DNA from the two affected half-siblings is not available. Thus, the remaining resources are precious even after whole genome amplification, and additional strategic conversations need to take place to optimize DNA usage for validation and functional studies moving forward.

Next steps in addition to variant validation on an independent platform include both quantitative and experimental approaches. First, we have existing Illumina 550K genotype array data on several hundred unrelated individuals with these three phenotypes as well as healthy controls. Leveraging imputation to 1000G, we have the potential to conduct prevalence and case-control association studies to examine if these variants also exert effects on severe CHD in a population context. However, a limitation is that many of these variants are quite rare; a vast majority were not directly measured on this platform, and high quality imputation will likely be difficult. Also, it is possible that the same variant would not be responsible for disease in other families or in the general population. Because variants may not be captured on commercially-available arrays or imputable with high quality, a second option is to directly measure these variants using a TaqMan[®] type of approach or to sequence the full genes including upstream and downstream areas to scan for variants in additional cases. Because of the potential for allelic heterogeneity in the general population, it will probably be most fruitful to perform targeted sequencing of these genes. In addition to likely underpowered single variant association tests, it may be worthwhile to consider gene-level association tests like the sequence kernel association test¹⁵⁶ to examine the joint effects of rare variation across pre-specified regions. An experimental approach under consideration is the use of morpholinos in zebrafish to knock down each individual gene's function from our candidate list and to screen for developmental cardiac phenotypes.

An apparent limitation to our study is that we focused our efforts solely on Mendelian models, concentrating on an analytic strategy including coding regions of the genome. This approach may be over-simplified, since regulatory regions of the non-coding genome may also segregate in a Mendelian fashion. The disease may also be etiologically complex by nature, with

multiple genetic and environmental factors contributing to CHD development. It is possible that none of the identified variants will replicate in other patients or populations, reflecting either biological truth of a family-specific cryptic mutation, methodological limitations, or genetic locus heterogeneity. Complex CHD in this family could be driven by a constitutional epimutation (heritable gene expression silencing), gene-gene interaction, or gene-environment interaction that is not measurable using this study design. Environmental influences are less likely to be primary contributors since the heart defects manifest at birth, and the affected offspring were born to two unrelated mothers with different prenatal environments. One might also speculate that there could be some link between the three severe CHD phenotypes and muscular dystrophy in the father's side of the family; however, there is no literature to date that supports an association between these two disease phenotypes, and we did not detect any mutations in genes responsible for muscular dystrophy in any patient with CHD.

Further limitations relate to variant calling and the technical aspects leading to generation of our final candidate list. NGS technologies remain notorious for higher error rates than traditional Sanger sequencing unless the depth of coverage is overwhelming, so one must rely on bioinformatic algorithms to predict false-positivity based on sequence quality metrics. Variant calling is a probabilistic science that depends on sequencing quality, depth of coverage, estimates of the per base mutation rate, and many other factors. The accuracy of calls increases with increasing coverage, but the number of calls is highly sensitive to parameter settings in the realignment and recalibration steps of the pipeline. Different variant calling pipelines could yield quite different final variant lists for validation. Thus, it is possible that a limited number of the identified variants will validate in this family using a different technology.

In addition, our search for potentially causative variants may have been narrowed by first running a linkage analysis in this family. This approach would have highlighted specific regions to focus our variant search, but the index family did not have enough informative meioses to make the dedication of previous DNA resources to a linkage chip justifiable. However, as an alternative, we did leverage previously published linkage peaks and candidate genes to inform our variant search. Finally, our main focus, with the exception of the clinical cytogenomic microarray, was on variants in coding, or “gene”, regions. However, we know from candidate gene and genome-wide association studies that a vast majority of disease-predisposing alleles do not fall in these regions but in other sequences with regulatory implications. Ignoring the remainder of the genome which was measured via WGS is a key limitation. The literature surrounding methods to mine the rest of the genome for potentially functional variants is evolving and serves as a future direction of this study.

Insights derived from the present analysis have suggested a number of previously unrecognized candidate genes and mutations that may be causally related to the development of CHD. This work will guide future family-based, population-based, and functional studies of the genetics of congenital heart disease. As seen in other diseases such as lipid disorders, familial hypertension, and Lynch Syndrome, there can be functional overlap between family-based and population-based cases with respect to genes and pathways contributing to their etiology. Studies of one-of-a-kind families, like the one in this dissertation chapter, have the potential to highlight new genes or pathways that warrant study on an epidemiologic scale in conjunction with other genetic and environmental risk factors. Further, we have the opportunity to study the potentially diverse manifestations of the same genetic defect, which has implications for general gene hunting for rare diseases. Seemingly unrelated conditions may share the same genetic defect; this

family affected by both HLHS and conotruncal defects is an example of such a diverse manifestation.

Chapter 5. Summary and Conclusions

This dissertation applied three different quantitative approaches to identify novel genetic risk factors contributing to the etiology of colorectal cancer (CRC) and complex congenital heart disease (CHD). It showed by example how extensions of traditional epidemiological and human genetic study designs can be implemented to characterize unique genetic risk factors for CRC and CHD. Also, this dissertation demonstrated the value of both population-based and family-based studies over the full range of genetic variant frequencies to advance our understanding of germline susceptibility for these complex diseases. While the first two chapters focused on CRC and the latter chapter addressed complex CHD, the quantitative approaches used in this dissertation are applicable across a wide spectrum of complex phenotypes.

Chapter 2 explored risk loci for CRC through a genome-wide association study (GWAS) meta-analysis. This project combined germline DNA samples from two case-control studies of high CRC incidence populations, Ashkenazi Jews from the Molecular Epidemiology of Colorectal Cancer (MECC) study and non-Hispanic whites from the Colon Cancer Family Registry (CFR). The discovery-replication design leveraged directly measured and 1000 Genomes Project imputed genotypes from 485 cases and 498 controls from MECC (discovery), 1977 cases and 999 controls from CFR (discovery), and 1131 cases and 831 controls from MECC (replication). In the discovery stage, an inverse-variance-weighted, fixed-effects meta-analysis was conducted on study-specific results from adjusted logistic regression models run to assess marginal genetic associations. This analysis replicated 14 out of 29 previously reported

risk loci at a minimum statistical significance level of 0.05. Further, the discovery meta-analysis identified a novel, genome-wide significant risk locus upstream of *FSTL5* at 4q32.2, which was sustained upon replication in an independent set of MECC samples from the CORECT consortium. The most statistically significant SNP from the joint discovery plus replication meta-analysis was rs35509282 (effect allele = A; OR=1.54; $p=8.2 \times 10^{-9}$; MAF~9%). This chapter also fine-mapped the genome-wide significant signal from the combined discovery plus replication meta-analysis. Future studies will include further replication in other populations, deeper fine mapping with the GAME-ON consortium's OncoChip, bioinformatic analysis of potentially overlapping biofeatures indicative of regulatory activity, and functional studies to better characterize this new risk locus.

In Chapter 3, the search for CRC susceptibility loci continued with a targeted GWAS approach in Ashkenazi Jews from northern Israel. MicroRNAs (miRNAs) act as key post-transcriptional regulators of gene expression via binding to the 3' untranslated regions (UTRs) of mRNAs. Single nucleotide polymorphisms (SNPs) found in the miRNA sequence and/or corresponding binding sites can affect the fidelity of this interaction, and evidence from candidate miRNA studies suggests that such SNPs may increase or decrease risk of tumor development. This study was designed to expand the search for miRNA-related polymorphisms important in the etiology of CRC across the genome and to investigate the association between genetic variants in miRNA target sites and colorectal cancer (CRC) risk using a novel genotyping platform, the Axiom[®] miRNA Target Site Genotyping Array (237,858 markers). The final study sample after quality control filtering included 596 cases and 429 controls from MECC. The marginal association between each marker with minor allele frequency $\geq 1\%$ and CRC risk was examined assuming a log-additive genetic model using logistic regression adjusted

for sex, age, and the first two principal components (PC) to capture fine-scale population structure. There were 23 markers with p-values less than 5×10^{-4} , and the most statistically significant association involved rs2985 (chr6: 34845648; intron of *UHRF1BP1*) with an OR of 0.66 and p-value of 3.7×10^{-5} . Further, this study provided evidence for replication of a previously published locus, rs1051690 in the 3'UTR region of the insulin receptor gene *INSR* (OR: 1.38; p = 0.03), and demonstrated variability in *INSR* gene expression by genotype at this locus. To our knowledge, this was the first study to examine the association between genetic variation in miRNA target sites and cancer risk using a genome-wide approach. Future directions include expansion of the genotyped dataset, replication and fine-mapping in the GAME-ON and CORECT consortia, and functional studies to identify the allele-specific effects on miRNA binding and to find the best in vitro model.

Finally, Chapter 4 explored genetic risk factors contributing to the etiology of complex CHD. The goal of this study was to identify rare genetic variation underlying the development of complex CHD through whole-genome sequencing (WGS) of a family affected by hypoplastic left heart syndrome (HLHS) and conotruncal defects. We hypothesized that the clustering of these anatomically distinct lesions was attributable to a more proximal defect in cardiogenesis. Paired-end, WGS data was generated on the Illumina Genome Analyzer II and HiSeq platforms for germline DNA from a proband with HLHS (~36X average depth of genome coverage); her affected half-sisters with tetralogy of Fallot (~39X) and truncus arteriosus (~10X), respectively; her unaffected full brother (~30X); and their shared unaffected father (~33X). Sequencing reads were aligned to the human genome reference sequence, followed by realignment and recalibration. Single nucleotide variant and short indel calls were made using the Genome Analysis Toolkit's UnifiedGenotyper. Variants with <1% minor allele frequency and all variants

located in previously-published CHD-associated gene regions were identified. These variants were analyzed for co-segregation according to one of three genetic models that were consistent with the pedigree: 1) autosomal dominant (AD) with incomplete penetrance, 2) AD with gonadal mosaicism, and 3) autosomal recessive. Thirty-two variants in 29 genes co-segregating according to one of these models and defined as high impact and/or potentially damaging by SIFT, PROVEAN, and/or Polyphen-2 were highlighted as causal candidates for future validation and follow-up epidemiologic studies.

The combination of epidemiologic study design, state of the art human genetic techniques, advanced bioinformatics tools, and quantitative methods enabled these applied studies at the frontier of genetic/genomic epidemiology. In this era of genomics research, there exists a wide range of genotyping and sequencing technologies to measure genetic variation. Also, we have a multitude of techniques at our disposal to evaluate the functional consequences of genetic variation prior to undertaking in vitro studies. As in this dissertation, decisions regarding which tools and methodologies to use must be incorporated as a critical aspect of study design.

Together, these three studies advance our understanding of genetic contributions to complex phenotypes through detailed investigation of highly informative individual patients, families, and large populations. This dissertation demonstrates the value of both population-based studies of common complex diseases and family-based studies of rare abnormalities, genetic studies in subgroups of the population, a focus on the full range of variant frequencies, and the use of unbiased gene hunting to focus future functional characterization efforts.

References

1. McPherson R. From genome-wide association studies to functional genomics: new insights into cardiovascular disease. *The Canadian journal of cardiology* 2013;29:23-9.
2. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature* 2009;461:747-53.
3. Chatterjee N, Wheeler B, Sampson J, Hartge P, Chanock SJ, Park JH. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nature genetics* 2013.
4. Hoyert DL, Xu J. Deaths: Preliminary Data for 2011. *National Vital Statistics Reports* 2012;61.
5. Peery AF, Dellon ES, Lund J, et al. Burden of gastrointestinal disease in the United States: 2012 update. *Gastroenterology* 2012;143:1179-87 e1-3.
6. Soerjomataram I, Lortet-Tieulent J, Parkin DM, et al. Global burden of cancer in 2008: a systematic analysis of disability-adjusted life-years in 12 world regions. *Lancet* 2012;380:1840-50.
7. Feinstein JA, Benson DW, Dubin AM, et al. Hypoplastic left heart syndrome: current considerations and expectations. *Journal of the American College of Cardiology* 2012;59:S1-42.
8. Yabroff KR, Lund J, Kepka D, Mariotto A. Economic burden of cancer in the United States: estimates, projections, and future research. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 2011;20:2006-14.
9. Mariotto AB, Yabroff KR, Shao Y, Feuer EJ, Brown ML. Projections of the cost of cancer care in the United States: 2010-2020. *Journal of the National Cancer Institute* 2011;103:117-28.
10. Ginsburg GS, Willard HF. Genomic and personalized medicine: foundations and applications. *Translational research : the journal of laboratory and clinical medicine* 2009;154:277-87.
11. Hamburg MA, Collins FS. The path to personalized medicine. *The New England journal of medicine* 2010;363:301-4.
12. Lichtenstein P, Holm NV, Verkasalo PK, et al. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *The New England journal of medicine* 2000;343:78-85.
13. Burt R. Inheritance of Colorectal Cancer. *Drug discovery today Disease mechanisms* 2007;4:293-300.
14. Hinton RB, Jr., Martin LJ, Tabangin ME, Mazwi ML, Cripe LH, Benson DW. Hypoplastic left heart syndrome is heritable. *Journal of the American College of Cardiology* 2007;50:1590-5.
15. Cripe L, Andelfinger G, Martin LJ, Shooner K, Benson DW. Bicuspid aortic valve is heritable. *Journal of the American College of Cardiology* 2004;44:138-43.

16. Eichler EE, Flint J, Gibson G, et al. Missing heritability and strategies for finding the underlying causes of complex disease. *Nature reviews Genetics* 2010;11:446-50.
17. Mardis ER. A decade's perspective on DNA sequencing technology. *Nature* 2011;470:198-203.
18. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature* 2001;409:860-921.
19. Tomlinson I. Colorectal cancer genetics: from candidate genes to GWAS and back again. *Mutagenesis* 2012;27:141-2.
20. A haplotype map of the human genome. *Nature* 2005;437:1299-320.
21. Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. High-resolution haplotype structure in the human genome. *Nature genetics* 2001;29:229-32.
22. Reich DE, Cargill M, Bolk S, et al. Linkage disequilibrium in the human genome. *Nature* 2001;411:199-204.
23. Weiss KM, Clark AG. Linkage disequilibrium and the mapping of complex human traits. *Trends in genetics : TIG* 2002;18:19-24.
24. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. *American journal of human genetics* 2012;90:7-24.
25. KA. W. DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program. In.
26. Gibson G. Rare and common variants: twenty arguments. *Nature reviews Genetics* 2011;13:135-45.
27. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. *Nature genetics* 2008;40:695-701.
28. Lander ES. The new genomics: global views of biology. *Science* 1996;274:536-9.
29. Pritchard JK, Cox NJ. The allelic architecture of human disease genes: common disease-common variant...or not? *Human molecular genetics* 2002;11:2417-23.
30. Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996;273:1516-7.
31. Reich DE, Lander ES. On the allelic spectrum of human disease. *Trends in genetics : TIG* 2001;17:502-10.
32. Bartel DP. MicroRNAs: target recognition and regulatory functions. *Cell* 2009;136:215-33.
33. Landi D, Gemignani F, Naccarati A, et al. Polymorphisms within micro-RNA-binding sites and risk of sporadic colorectal cancer. *Carcinogenesis* 2008;29:579-84.
34. Howlander N NA, Krapcho M, Neyman N, Aminou R, Altekruse SF, Kosary CL, Ruhl J, Tatalovich Z, Cho H, Mariotto A, Eisner MP, Lewis DR, Chen HS, Feuer EJ, Cronin KA (eds). SEER Cancer Statistics Review, 1975-2009 (Vintage 2009 Populations), based on November 2011 SEER data submission. In. National Cancer Institute, Bethesda, MD; 2012.
35. Ferlay J SH, Bray F, Forman D, Mathers C and Parkin DM. GLOBOCAN 2008 v1.2, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 10 [Internet]. In. Lyon, France: International Agency for Research on Cancer; 2010.
36. Bray F RJ, Masuyer E, Ferlay J. Estimates of global cancer prevalence in 2008 for 27 sites in the adult population, submitted. . In.
37. Fireman Z, Sandler E, Kopelman Y, Segal A, Sternberg A. Ethnic differences in colorectal cancer among Arab and Jewish neighbors in Israel. *The American journal of gastroenterology* 2001;96:204-7.

38. Barchana M, Liphshitz I, Rozen P. Trends in colorectal cancer incidence and mortality in the Israeli Jewish ethnic populations. *Familial cancer* 2004;3:207-14.
39. Hagggar FA, Boushey RP. Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors. *Clinics in colon and rectal surgery* 2009;22:191-7.
40. Woods MO, Younghusband HB, Parfrey PS, et al. The genetic basis of colorectal cancer in a population-based incident cohort with a high rate of familial disease. *Gut* 2010;59:1369-77.
41. Kemp Z, Thirlwell C, Sieber O, Silver A, Tomlinson I. An update on the genetics of colorectal cancer. *Human molecular genetics* 2004;13 Spec No 2:R177-85.
42. Fearon ER. Molecular genetics of colorectal cancer. *Annual review of pathology* 2011;6:479-507.
43. Rustgi AK. The genetics of hereditary colon cancer. *Genes & development* 2007;21:2525-38.
44. Lynch HT, de la Chapelle A. Hereditary colorectal cancer. *The New England journal of medicine* 2003;348:919-32.
45. Vasen HF, Moslein G, Alonso A, et al. Guidelines for the clinical management of Lynch syndrome (hereditary non-polyposis cancer). *Journal of medical genetics* 2007;44:353-62.
46. Galiatsatos P, Foulkes WD. Familial adenomatous polyposis. *The American journal of gastroenterology* 2006;101:385-98.
47. Gruber SB, Moreno V, Rozek LS, et al. Genetic variation in 8q24 associated with risk of colorectal cancer. *Cancer biology & therapy* 2007;6:1143-7.
48. Jaeger E, Webb E, Howarth K, et al. Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk. *Nature genetics* 2008;40:26-8.
49. Zanke BW, Greenwood CM, Rangrej J, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nature genetics* 2007;39:989-94.
50. Broderick P, Carvajal-Carmona L, Pittman AM, et al. A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. *Nature genetics* 2007;39:1315-7.
51. Tomlinson I, Webb E, Carvajal-Carmona L, et al. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nature genetics* 2007;39:984-8.
52. Houlston RS, Webb E, Broderick P, et al. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nature genetics* 2008;40:1426-35.
53. Neklason DW, Kerber RA, Nilson DB, et al. Common familial colorectal cancer linked to chromosome 7q31: a genome-wide analysis. *Cancer research* 2008;68:8993-7.
54. Tomlinson IP, Webb E, Carvajal-Carmona L, et al. A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nature genetics* 2008;40:623-30.
55. Tenesa A, Farrington SM, Prendergast JG, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nature genetics* 2008;40:631-7.
56. Houlston RS, Cheadle J, Dobbins SE, et al. Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. *Nature genetics* 2010;42:973-7.
57. Peters U, Hutter CM, Hsu L, et al. Meta-analysis of new genome-wide association studies of colorectal cancer risk. *Human genetics* 2012;131:217-34.

58. Peters U, Jiao S, Schumacher FR, et al. Identification of Genetic Susceptibility Loci for Colorectal Tumors in a Genome-Wide Meta-Analysis. *Gastroenterology* 2012.
59. Dunlop MG, Dobbins SE, Farrington SM, et al. Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk. *Nature genetics* 2012;44:770-6.
60. Tomlinson IP, Carvajal-Carmona LG, Dobbins SE, et al. Multiple common susceptibility variants near BMP pathway loci GREM1, BMP4, and BMP2 explain part of the missing heritability of colorectal cancer. *PLoS genetics* 2011;7:e1002105.
61. Tenesa A, Dunlop MG. New insights into the aetiology of colorectal cancer from genome-wide association studies. *Nature reviews Genetics* 2009;10:353-8.
62. Chung CC, Chanock SJ. Current status of genome-wide association studies in cancer. *Human genetics* 2011;130:59-78.
63. Kraft P, Hunter DJ. Genetic risk prediction--are we there yet? *The New England journal of medicine* 2009;360:1701-3.
64. Park JH, Wacholder S, Gail MH, et al. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature genetics* 2010;42:570-5.
65. Jo J, Nam CM, Sull JW, et al. Prediction of Colorectal Cancer Risk Using a Genetic Risk Score: The Korean Cancer Prevention Study-II (KCPS-II). *Genomics & informatics* 2012;10:175-83.
66. Goldstein DB. Common genetic variation and human traits. *The New England journal of medicine* 2009;360:1696-8.
67. Jia WH, Zhang B, Matsuo K, et al. Genome-wide association analyses in East Asians identify new susceptibility loci for colorectal cancer. *Nature genetics* 2013;45:191-6.
68. Cui R, Okada Y, Jang SG, et al. Common variant in 6q26-q27 is associated with distal colon cancer in an Asian population. *Gut* 2011;60:799-805.
69. Haiman CA, Le Marchand L, Yamamoto J, et al. A common genetic risk factor for colorectal and prostate cancer. *Nature genetics* 2007;39:954-6.
70. Hutter CM, Slattery ML, Duggan DJ, et al. Characterization of the association between 8q24 and colon cancer: gene-environment exploration and meta-analysis. *BMC cancer* 2010;10:670.
71. Kocarnik JD, Hutter CM, Slattery ML, et al. Characterization of 9p24 risk locus and colorectal adenoma and cancer: gene-environment interaction and meta-analysis. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 2010;19:3131-9.
72. Poynter JN, Gruber SB, Higgins PD, et al. Statins and the risk of colorectal cancer. *The New England journal of medicine* 2005;352:2184-92.
73. Newcomb PA, Baron J, Cotterchio M, et al. Colon Cancer Family Registry: an international resource for studies of the genetic epidemiology of colon cancer. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 2007;16:2331-43.
74. Figueiredo JC, Lewinger JP, Song C, et al. Genotype-environment interactions in microsatellite stable/microsatellite instability-low colorectal cancer: results from a genome-wide association study. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 2011;20:758-66.

75. Stacklies W, Redestig H, Scholz M, Walther D, Selbig J. *pcaMethods*--a bioconductor package providing PCA methods for incomplete data. *Bioinformatics* 2007;23:1164-7.
76. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* 2006;38:904-9.
77. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics* 2009;5:e1000529.
78. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. *Nature methods* 2012;9:179-81.
79. A map of human genome variation from population-scale sequencing. *Nature* 2010;467:1061-73.
80. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 2010;26:2190-1.
81. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* 2007;81:559-75.
82. Dudbridge F, Gusnanto A. Estimation of significance thresholds for genomewide association scans. *Genetic epidemiology* 2008;32:227-34.
83. McNamee R. Confounding and confounders. *Occupational and environmental medicine* 2003;60:227-34; quiz 164, 234.
84. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nature genetics* 2006;38:209-13.
85. Pruim RJ, Welch RP, Sanna S, et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* 2010;26:2336-7.
86. Dunham I, Kundaje A, Aldred SF, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;489:57-74.
87. Zabala W, Cruz R, Barreiro-de Acosta M, et al. New genetic associations in thiopurine-related bone marrow toxicity among inflammatory bowel disease patients. *Pharmacogenomics* 2013;14:631-40.
88. Remke M, Hielscher T, Korshunov A, et al. FSTL5 is a marker of poor prognosis in non-WNT/non-SHH medulloblastoma. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2011;29:3852-61.
89. Hindorff LA, MJEBI, Morales J (European Bioinformatics Institute), Junkins HA, Hall PN, Klemm AK, and Manolio TA. A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies. April 16, 2013.
90. Medland SE, Nyholt DR, Painter JN, et al. Common variants in the trichohyalin gene are associated with straight hair in Europeans. *American journal of human genetics* 2009;85:750-5.
91. Hart AB, Engelhardt BE, Wardle MC, et al. Genome-wide association study of d-amphetamine response in healthy volunteers identifies putative associations, including cadherin 13 (CDH13). *PloS one* 2012;7:e42646.
92. Huang YC, Lin JM, Lin HJ, et al. Genome-wide association study of diabetic retinopathy in a Taiwanese population. *Ophthalmology* 2011;118:642-8.
93. Croteau-Chonka DC, Marvelle AF, Lange EM, et al. Genome-wide association study of anthropometric traits and evidence of interactions with age and study year in Filipino women. *Obesity (Silver Spring)* 2011;19:1019-27.

94. Mavaddat N, Peock S, Frost D, et al. Cancer Risks for BRCA1 and BRCA2 Mutation Carriers: Results From Prospective Analysis of EMBRACE. *Journal of the National Cancer Institute* 2013;105:812-22.
95. Pittman AM, Webb E, Carvajal-Carmona L, et al. Refinement of the basis and impact of common 11q23.1 variation to the risk of developing colorectal cancer. *Human molecular genetics* 2008;17:3720-7.
96. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 2004;116:281-97.
97. Ambros V. The functions of animal microRNAs. *Nature* 2004;431:350-5.
98. Chendrimada TP, Gregory RI, Kumaraswamy E, et al. TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. *Nature* 2005;436:740-4.
99. Hutvagner G, Zamore PD. A microRNA in a multiple-turnover RNAi enzyme complex. *Science* 2002;297:2056-60.
100. Vilar E, Tabernero J, Gruber SB. Micromanaging the classification of colon cancer: the role of the microRNAome. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2011;17:7207-9.
101. Calin GA, Croce CM. MicroRNA signatures in human cancers. *Nature reviews Cancer* 2006;6:857-66.
102. Cummins JM, He Y, Leary RJ, et al. The colorectal microRNAome. *Proceedings of the National Academy of Sciences of the United States of America* 2006;103:3687-92.
103. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. *Cell* 1990;61:759-67.
104. Olson P, Lu J, Zhang H, et al. MicroRNA dynamics in the stages of tumorigenesis correlate with hallmark capabilities of cancer. *Genes & development* 2009;23:2152-65.
105. Balaguer F, Moreira L, Lozano JJ, et al. Colorectal cancers with microsatellite instability display unique miRNA profiles. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2011;17:6239-49.
106. Bartley AN, Yao H, Barkoh BA, et al. Complex patterns of altered MicroRNA expression during the adenoma-adenocarcinoma sequence for microsatellite-stable colorectal cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2011;17:7283-93.
107. Schee K, Fodstad O, Flatmark K. MicroRNAs as biomarkers in colorectal cancer. *The American journal of pathology* 2010;177:1592-9.
108. Liu M, Chen H. The role of microRNAs in colorectal cancer. *Journal of genetics and genomics = Yi chuan xue bao* 2010;37:347-58.
109. Ju J. miRNAs as biomarkers in colorectal cancer diagnosis and prognosis. *Bioanalysis* 2010;2:901-6.
110. Vinci S, Gelmini S, Mancini I, et al. Genetic and epigenetic factors in regulation of microRNA in colorectal cancers. *Methods* 2013;59:138-46.
111. Azimzadeh P, Romani S, Mohebbi SR, et al. Association of polymorphisms in microRNA-binding sites and colorectal cancer in an Iranian population. *Cancer genetics* 2012;205:501-7.
112. Naccarati A, Pardini B, Stefano L, et al. Polymorphisms in miRNA-binding sites of nucleotide excision repair genes and colorectal cancer risk. *Carcinogenesis* 2012;33:1346-51.
113. Landi D, Moreno V, Guino E, et al. Polymorphisms affecting micro-RNA regulation and associated with the risk of dietary-related cancers: a review from the literature and new evidence

for a functional role of rs17281995 (CD86) and rs1051690 (INSR), previously associated with colorectal cancer. *Mutation research* 2011;717:109-15.

114. Ziebarth JD, Bhattacharya A, Chen A, Cui Y. PolymiRTS Database 2.0: linking polymorphisms in microRNA target sites with human diseases and complex traits. *Nucleic acids research* 2012;40:D216-21.

115. Schmeier S, Schaefer U, MacPherson CR, Bajic VB. dPORE-miRNA: polymorphic regulation of microRNA genes. *PloS one* 2011;6:e16657.

116. Hiard S, Charlier C, Coppieters W, Georges M, Baurain D. Patrocles: a database of polymorphic miRNA-mediated gene regulation in vertebrates. *Nucleic acids research* 2010;38:D640-51.

117. Betel D, Koppal A, Agius P, Sander C, Leslie C. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome biology* 2010;11:R90.

118. Hindorff LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America* 2009;106:9362-7.

119. Betel D, Wilson M, Gabow A, Marks DS, Sander C. The microRNA.org resource: targets and expression. *Nucleic acids research* 2008;36:D149-53.

120. Vilar E, Bartnik CM, Stenzel SL, et al. MRE11 deficiency increases sensitivity to poly(ADP-ribose) polymerase inhibition in microsatellite unstable colorectal cancers. *Cancer research* 2011;71:2632-42.

121. Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic acids research* 2011;39:D152-7.

122. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. miRBase: tools for microRNA genomics. *Nucleic acids research* 2008;36:D154-8.

123. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic acids research* 2006;34:D140-4.

124. Griffiths-Jones S. The microRNA Registry. *Nucleic acids research* 2004;32:D109-11.

125. Rebhun JF, Castro AF, Quilliam LA. Identification of guanine nucleotide exchange factors (GEFs) for the Rap1 GTPase. Regulation of MR-GEF by M-Ras-GTP interaction. *The Journal of biological chemistry* 2000;275:34901-8.

126. Giovannucci E. Insulin and colon cancer. *Cancer causes & control : CCC* 1995;6:164-79.

127. Sethupathy P, Collins FS. MicroRNA target site polymorphisms and human disease. *Trends in genetics : TIG* 2008;24:489-97.

128. Hoffman JI, Kaplan S. The incidence of congenital heart disease. *Journal of the American College of Cardiology* 2002;39:1890-900.

129. Hoffman JI, Kaplan S, Liberthson RR. Prevalence of congenital heart disease. *American heart journal* 2004;147:425-39.

130. Barron DJ, Kilby MD, Davies B, Wright JG, Jones TJ, Brawn WJ. Hypoplastic left heart syndrome. *Lancet* 2009;374:551-64.

131. Gruber PJ, Epstein JA. Development gone awry: congenital heart disease. *Circulation research* 2004;94:273-83.

132. Allan LD, Sharland GK, Milburn A, et al. Prospective diagnosis of 1,006 consecutive cases of congenital heart disease in the fetus. *Journal of the American College of Cardiology* 1994;23:1452-8.

133. Raymond FL, Simpson JM, Sharland GK, Ogilvie Mackie CM. Fetal echocardiography as a predictor of chromosomal abnormality. *Lancet* 1997;350:930.
134. Grossfeld PD, Mattina T, Lai Z, et al. The 11q terminal deletion disorder: a prospective study of 110 cases. *American journal of medical genetics Part A* 2004;129A:51-61.
135. Pierpont ME, Basson CT, Benson DW, Jr., et al. Genetic basis for congenital heart defects: current knowledge: a scientific statement from the American Heart Association Congenital Cardiac Defects Committee, Council on Cardiovascular Disease in the Young: endorsed by the American Academy of Pediatrics. *Circulation* 2007;115:3015-38.
136. Hinton RB, Martin LJ, Rame-Gowda S, Tabangin ME, Cripe LH, Benson DW. Hypoplastic left heart syndrome links to chromosomes 10q and 6q and is genetically related to bicuspid aortic valve. *Journal of the American College of Cardiology* 2009;53:1065-71.
137. Dasgupta C, Martinez AM, Zuppan CW, Shah MM, Bailey LL, Fletcher WH. Identification of connexin43 (alpha1) gap junction gene mutations in patients with hypoplastic left heart syndrome by denaturing gradient gel electrophoresis (DGGE). *Mutation research* 2001;479:173-86.
138. Garg V, Muth AN, Ransom JF, et al. Mutations in NOTCH1 cause aortic valve disease. *Nature* 2005;437:270-4.
139. Reamon-Buettner SM, Ciribilli Y, Inga A, Borlak J. A loss-of-function mutation in the binding domain of HAND1 predicts hypoplasia of the human hearts. *Human molecular genetics* 2008;17:1397-405.
140. McElhinney DB, Geiger E, Blinder J, Benson DW, Goldmuntz E. NKX2.5 mutations in patients with congenital heart disease. *Journal of the American College of Cardiology* 2003;42:1650-5.
141. Fahed AC, Gelb BD, Seidman JG, Seidman CE. Genetics of congenital heart disease: the glass half empty. *Circulation research* 2013;112:707-20.
142. Jenkins KJ, Correa A, Feinstein JA, et al. Noninherited risk factors and congenital cardiovascular defects: current knowledge: a scientific statement from the American Heart Association Council on Cardiovascular Disease in the Young: endorsed by the American Academy of Pediatrics. *Circulation* 2007;115:2995-3014.
143. Grossfeld PD. The genetics of hypoplastic left heart syndrome. *Cardiology in the young* 1999;9:627-32.
144. Bentley DR, Balasubramanian S, Swerdlow HP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 2008;456:53-9.
145. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25:1754-60.
146. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* 2011;43:491-8.
147. Mills RE, Pittard WS, Mullaney JM, et al. Natural genetic variation caused by small insertions and deletions in the human genome. *Genome research* 2011;21:830-9.
148. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. *Bioinformatics* 2011;27:2156-8.
149. Cingolani P, Patel VM, Coon M, et al. Using *Drosophila melanogaster* as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. *Frontiers in genetics* 2012;3:35.

150. Cingolani P, Platts A, Wang le L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* 2012;6:80-92.
151. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols* 2009;4:1073-81.
152. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nature methods* 2010;7:248-9.
153. Chen K, Wallis JW, McLellan MD, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods* 2009;6:677-81.
154. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. *PloS one* 2012;7:e46688.
155. Jun G, Flickinger M, Hetrick KN, et al. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *American journal of human genetics* 2012;91:839-48.
156. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *American journal of human genetics* 2011;89:82-93.