EPIDEMIOLOGIC APPROACHES TO UNDERSTANDING COMPLEX DISEASES: APPLICATIONS IN CONGENITAL HEART DISEASE AND CANCER

by

Kristen N. Stevens

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Epidemiologic Sciences)
in The University of Michigan
2010

Doctoral Committee:

Professor Stephen B. Gruber, Chair
Professor Michael L. Boehnke
Professor Patricia A. Peyser
Professor Jeremy M.G. Taylor
Associate Professor Gad Rennert, Technion-Israel Institute of Technology
Assistant Professor Peter J. Gruber, University of Pennsylvania

DEDICATION

To my family

ACKNOWLEDGEMENTS

This dissertation was supported in part by grants from the National Institutes of Health (R01-CA81488 and T32-HG00040).

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

LIST OF ABBREVIATIONS

Abbreviation	Definition
AIM	Ancestral informative marker
ASE	Allele-specific expression
CHD	Congenital heart disease
CHOP	Children's Hospital of Philadelphia
CI	Confidence Interval
CNS	Central nervous system
CRC	Colorectal cancer
CT	Computerized tomography
EM	Expectation maximization
GWAS	Genome-wide association study
Gy	Gray
HR	Hazard Ratio
LD	Linkage disequilibrium
LOH	Loss of heterozygosity
MAF	Minor allele frequency
MECC	Molecular Epidemiology of Colorectal Cancer
mSv	Millisieverts
NSAIDs	Non-steroidal anti-inflammatory drugs
OR	Odds Ratio
PCA	Principal components analysis
RR	Rate ratio
SD	Standard deviation
SIR	Standardized Incidence Ratio
SNP	Single nucleotide polymorphism

ABSTRACT

Epidemiology provides a means of investigating the underlying architecture of complex diseases by combining advances in technology with our current understanding of the epidemiology and biology of disease. In this dissertation, epidemiologic methods are applied to further understand the etiology of colorectal cancer, childhood cancers, and congenital heart disease (CHD).

The low-penetrance genes that contribute to risk of familial colorectal cancers (CRCs), estimated to account for 35% of all CRCs, are mostly unknown We conducted a genome-wide association study (GWAS) of CRC using a population based case-control study in northern Israel to survey the genome for low-penetrance susceptibility genes. Two leading candidate regions resulting from the Molecular Epidemiology of Colorectal Cancer (MECC) study GWAS included rs10210149 on chromosome 2q11.2-q12 and rs16931815 on chromosome 12p11.23. After excluding potential pathogenic mutations by Sanger sequencing, allele-specific expression analyses were performed for *GPR45*, *STK38L*, and *TGFBRAP1*, three genes within the two candidate regions. *GPR45* was associated with a 27% increase in expression of one allele for each additional copy of the C allele of rs10210149 (p-trend = 0.01). Further studies are necessary to fully elucidate the mechanisms underlying these GWAS associations.

Common genetic variation and risk of congenital heart disease has not previously been studied. We investigated variation in *ISL1*, a marker of cells that contribute to specific developmental fields of the embryonic human heart, and risk of CHD in a two-

stage case-control study. Eight genic and flanking *ISL1* SNPs were significantly associated with complex CHD. A replication study analyzed the three SNPs within *ISL1* (rs3762977, IVS1+17C>T, rs1017) in 1,044 new cases and 3,934 independent controls and confirmed that genetic variation in *ISL1* is associated with risk of CHD. Our results demonstrate that two different *ISL1* haplotypes contribute to risk of CHD in white (Summary Odds Ratio (OR) =1.27, 95% Confidence Interval (CI) 1.09 – 1.48, P = 0.0018) and black/African American populations (Summary OR=1.57, 95% CI 1.07 – 2.30, P = 0.0216).

Linking epidemiologic approaches to cancer epidemiology and cardiovascular disease is stimulated by the well known association between selected forms of congenital heart disease and cancer. Investigating the relationship between CHD and childhood cancer could provide a basis for identifying novel risk factors for both sets of diseases. We conducted a retrospective cohort study of CHD at the Children's Hospital of Philadelphia (CHOP), showing that children with CHD demonstrated a 3.7-fold increase in the rate of pediatric cancer compared to the US population (Standardized Incidence Ratio (SIR) = 3.72, 95% CI = 1.53 – 9.04, p = 0.0037). Rates were higher for children with both syndromic (SIR=12.49, 95% CI 1.28 – 121.74, p=0.03) and non-syndromic (SIR=2.41, 95% CI 0.88 – 6.60, p=0.086) heart disease. Diagnostic radiation was not significantly associated with an increased rate of cancer.

CHAPTER 1

Introduction

1.1 Gaining insight into colorectal cancer biology and epidemiology

Colorectal cancer (CRC) is a significant public health problem in the United States, ranking as the third most commonly diagnosed cancer and the fourth leading cause of cancer-related deaths (Ferlay *et al.*, 2010). Furthermore, global CRC incidence is among the highest of all cancers, ranking as the fourth most commonly diagnosed cancer worldwide (WHO, 2003). The epidemiology of CRC has been studied extensively, and several confirmed environmental and lifestyle exposures have been identified that either increase risk, such as meat intake and smoking (Larsson and Wolk, 2006; Limsui *et al.*, 2010), or decrease risk, such as physical activity (Wolin *et al.*, 2009) and nonsteroidal anti-inflammatory drugs (NSAIDs) (Huls *et al.*, 2003). Genetic models for colorectal carcinogenesis have also been well described, including the chromosomal instability and microsatellite instability pathways. Autosomal dominant mutations in key genes in these two pathways have been identified in a small number of rare, highly penetrant familial syndromes, which account for less than 5% of all CRC (Goss and Groden, 2000; Kemp *et al.*, 2004; Marra and Boland, 1995). Further, family history of non-syndromic CRC is associated with a two-fold increase in risk (Carstensen *et al.*, 1996). These familial cancers are estimated to account for an additional 35% of CRC (Tenesa and Dunlop, 2009), yet the set of low-penetrance genes involved in susceptibility to these colorectal

cancers remains an elusive target

In contrast to the rare, highly penetrant mutations associated with CRC syndromes, familial non-syndromic CRC is thought to be attributed to more common, moderate- to low-penetrance mutations. A handful of these genes have been identified through the implementation of candidate gene studies, including the I1307K mutation in *APC* (Laken *et al.*, 1997) and variants in *TGFBR1* (de Jong *et al.*, 2002), which confer risks between 1.43 and 2.00. More recently, advances in genotyping technology have allowed investigators to agnostically survey the entire genome in search of additional low-penetrance susceptibility genes. The genome-wide association study (GWAS) design has now been widely implemented in colorectal cancer, resulting in the identification of several new loci strongly associated with CRC (Broderick *et al.*, 2007; Gruber *et al.*, 2007; Houlston *et al.*, 2008; Tenesa and Dunlop, 2009; Zanke *et al.*, 2007). The majority of these loci are not located within or near known genes, and the biological relevance of some of these signals is unclear. Others appear to be related to TGFB signaling (Tenesa and Dunlop, 2009) and long-range regulation of *c-MYC* (Sotelo *et al.*, 2010).

In the second chapter of this dissertation, I describe a functional study of novel associations identified in the Molecular Epidemiology of Colorectal Cancer (MECC) GWAS study, a population-based case-control study in northern Israel. While the MECC GWAS has contributed to the replication of chromosomal regions such as 8q24, 18q21, and 11q23.1 in multi-center analyses (Tenesa *et al.*, 2008), I focus here on associations that have not previously been reported. The goal of this analysis was to gain insight into the biology of colorectal cancer, thus we chose to focus on the biologically relevant

hypotheses generated from the MECC GWAS. I chose to analyze three genes- *GPR45*, *TGFBRAP1*, and *STK38L*- within two candidate regions by examining both genetic variation and expression patterns.

While searching for causal variants in genes identified from GWA studies is an important step, it is often difficult due to several constraints: 1) appropriate selection of subjects for sequencing, 2) adequate coverage of chromosomal regions captured by candidate loci, and 3) identification and interpretation of potential causal variants. The ultimate goal of these analyses is to understand the functional consequences of GWAS signals. Thus, we chose to follow our sequencing analyses by measuring differences in expression between alleles for each of the three genes mentioned above. Allele-specific expression (ASE), discussed in more detail below, has been hypothesized to play an important role in susceptibility to complex diseases but has been demonstrated in only a few examples. Here, I show that allele-specific expression of *GPR45* is associated with the risk allele at one of the MECC GWAS loci, rs10210149. *GPR45* was associated with a 27% increase in expression of one allele for each additional copy of the C allele of rs10210149 (p-trend = 0.01). ASE of *STK38L* and *TGFBRAP1* does not seem to play an important role in colorectal carcinogenesis, and further studies are necessary to fully elucidate the mechanisms underlying these GWAS associations.

1.2 Understanding the genetics of congenital heart disease

Congenital heart disease (CHD) is the most common live birth abnormality and affects an increasingly large proportion of the population (Hoffman and Kaplan, 2002; Hoffman *et al.*, 2004), yet few epidemiologic studies have been conducted to understand

the origins of this complex disease. A few maternal exposures have been associated with CHD, such as organic solvents and some illnesses, but these account for only 30% of all cardiac defects (Jenkins *et al.*, 2007). An additional 13% of CHD is attributed to large-scale chromosomal abnormalities, such as trisomy or large deletions, although this estimate varies widely by specific defect and is expected to increase with improved resolution of cytogenetic technologies (Pierpont *et al.*, 2007). However, even considering the combined attributable fractions for all of these known risk factors there still remains a large proportion of CHD with no known cause.

We are particularly interested in examining the genetic contributions to risk of CHD. Several rare single-gene disorders are associated with CHD. For example, Alagille syndrome is caused by mutations in/deletion of *JAG1*, Noonan syndrome is caused by mutations in *PTPN11*, *SOS1*, and *KRAS*, and Holt-Oram syndrome is caused by mutations in *TBX5* (Pierpont *et al.*, 2007). Patients with these disorders present with a variety of cardiac defects such as tetralogy of Fallot, septal defects, and pulmonary valve stenosis. Rare mutations in *NKX2.5* and *GATA4* have also been identified in non-syndromic patients with atrial septal defects, atrioventricular conduction delay, and ventricular septal defects (Garg *et al.*, 2003; Posch *et al.*, 2008; Schott *et al.*, 1998). These rare mutations provide evidence that single gene mutations can have a substantial phenotypic impact on cardiac development. Furthermore, studies of offspring of affected individuals have shown a significantly higher proportion of children with CHD than expected, indicating that genetics may play a larger role in CHD etiology than currently appreciated (Rose *et al.*, 1985; Whittemore *et al.*, 1982; Whittemore *et al.*, 1994).

To our knowledge, the association between common genetic variation and risk of

congenital heart disease has not yet been investigated. In this dissertation, we employ a standard candidate gene approach to investigate this association, specifically focusing on the gene *ISL1*. Two factors influenced our decision to take this approach. First, an extensive number of developmental experiments have delineated the genetic pathways involved in the regulation of cardiac development. This allowed us to identify *ISL1* as a promising candidate gene, as discussed in Chapter 3. Second, at the time the study was conducted, no data were available for a sufficiently large number of CHD cases and controls to conduct a genome-wide analysis. Given a compelling biologic rationale, genotyping costs, and our unique access to large CHD patient populations, we decided to proceed with selective genotyping of the candidate gene *ISL1* in a two-stage case-control study.

While currently described risk factors are typically associated with only subsets of CHD, our data suggest that phenotypically heterogeneous congenital heart defects may in fact have common origins. In the third chapter of this dissertation, I describe the first reported association between common genetic variation in the candidate gene *ISL1* and risk of CHD. Our results demonstrate that two different *ISL1* haplotypes contribute to risk of CHD in white (Summary OR =1.27, 95% CI 1.09 – 1.48, $P = 0.0018$) and black/African American populations (Summary OR=1.57, 95% CI 1.07 – 2.30, $P = 0.0216$).

1.3 Identifying a link between congenital heart disease and cancer

Much remains to be learned about the causes of both congenital heart disease and childhood cancers. As described above, almost 60% of all CHD is unexplained (Jenkins

et al., 2007; Pierpont *et al.*, 2007). Similarly, the causes of the majority of childhood cancers are unknown, with both genetics and the environment suggested to be risk factors (Stiller, 2004). Very few firmly established environmental risk factors have been identified for these rare cancers, including diagnostic x-rays during pregnancy, Epstein Barr virus, hepatitis B, and human immunodeficiency virus (Stiller, 2004). Additionally, some genetic syndromes are associated with risk of childhood cancer, including familial cancer syndromes, immunodeficiency and bone marrow disorders, and others (Stiller, 2004). However, the search for causes of childhood cancers is thought to have been substantially hindered by methodological issues such as participation and recall bias.

Interestingly, there is substantial evidence for a relationship between developmental abnormalities and childhood malignancies. First, several genetic disorders, such as Down syndrome and Noonan syndrome, are associated with both cardiac defects and an increased risk of pediatric cancers (Denayer *et al.*, 2008; Freeman *et al.*, 1998). Second, multiple large cohort studies have identified associations between congenital anomalies and childhood cancers (Agha *et al.*, 2005; Bjorge *et al.*, 2008; Narod *et al.*, 1997; Rankin *et al.*, 2008). These preliminary observations support the hypothesis that there is a causal link between birth defects and childhood cancer, whether the common risk factors are environmental or genetic. However, the specific association between congenital heart disease and childhood cancer has never been reported. Investigation of this relationship could provide a basis for identifying novel risk factors for both sets of diseases.

In the fourth chapter of this dissertation, I investigate the epidemiology of childhood cancers among children with congenital heart disease. The patient population

of the Children's Hospital of Philadelphia (CHOP) provides a unique opportunity to investigate this relationship. First, the cardiac phenotypes of children undergoing operations at the Cardiac Center for CHD are extremely well characterized by highly skilled physicians. Second, many of these children continue to receive long-term care at CHOP, indicating that diseases such as cancer are likely to be captured by CHOP registries and highlighting the potential for long-term follow-up of these children. Here, I describe the first report of an excess of pediatric cancers among children with CHD, suggesting that future studies of these children are warranted to elucidate the common causes of these diseases. Children with CHD demonstrated a 3.7-fold increase in the rate of pediatric cancer compared to the US population (Standardized Incidence Ratio (SIR) = 3.72, 95% CI = 1.53 – 9.04, p = 0.0037). Rates were higher for children with both syndromic (SIR=12.49, 95% CI 1.28 – 121.74, p=0.03) and non-syndromic (SIR=2.41, 95% CI 0.88 – 6.60, p=0.086) heart disease.

1.4 Conclusions

In this dissertation, I show that epidemiologic methods provide a way to understand the underlying architecture of complex diseases. Consistent with the conclusions drawn from other GWAS studies, we suggest that the underlying causal variant at rs10210149 affects expression of *GPR45* in the MECC study. We also provide strong evidence that congenital heart disease is consistent with the common disease – common variant hypothesis, and suggest a new role for known regulatory genes of cardiomyocytes in human disease. Finally, we demonstrate a link between the biology and epidemiology of both CHD and cancer, suggesting that future studies can take advantage of this relationship to further understand this common link.

CHAPTER 2

Genetic and allele-specific expression analyses of genes identified from a genome wide association study of colorectal cancer

2.1 Introduction

Colorectal cancer (CRC) is the third most commonly diagnosed cancer in the United States, and the fourth most commonly diagnosed cancer worldwide (Ferlay *et al.*, 2010). While genetic susceptibility to colorectal cancer is well described for a small number of rare, highly penetrant familial syndromes, genetic susceptibility to non-syndromic, familial CRCs is less well understood (Tenesa and Dunlop, 2009). Relatively common moderate- to low-penetrance genes are thought to be responsible for a large number of these familial colorectal cancers, which has led to a concentrated effort to identify these genes.

Recently, genome-wide association studies (GWAS) of colorectal cancer have been widely implemented in an attempt to comprehensively survey the genome for these low-penetrance susceptibility genes. These studies have identified several new loci that are strongly associated with CRC risk, including SNPs in genic regions such as *SMAD7* and *EIF3H* (8q23.3) as well as SNPs in regions far from any known genes, such as 8q24 (Broderick *et al.*, 2007; Gruber *et al.*, 2007; Houlston *et al.*, 2008; Tomlinson *et al.*, 2007; Tomlinson *et al.*, 2008; Zanke *et al.*, 2007). However, the functional significance of the majority of these associations is not yet known, other than the reported long-range regulation of *c-MYC* and enhancement of Wnt signaling at 8q24 (Pomerantz *et al.*, 2009;

Sotelo *et al.*, 2010; Tuupanen *et al.*, 2009).

We conducted a genome-wide association study of colorectal cancer using the Molecular Epidemiology of Colorectal Cancer study, a population-based case-control study in northern Israel, to identify low-penetrance susceptibility loci. In this chapter, I investigate the genetic and functional basis of associations at two candidate SNPs identified from the MECC GWAS. These two SNPs are rs10210149 on chromosome 2q11.2-q12 and rs16931835 on chromosome 12p11.23. In a log-additive model, each copy of the C allele of rs10210149 was associated with a 12% increase in risk of colorectal cancer (OR=1.12, 95% CI 1.07 – 1.18, $p = 2.0 \times 10^{-6}$) (**Figure 2.1**). Similarly, each copy of the A allele of rs16931815 was associated with an 18% increase in risk of colorectal cancer (OR=1.18, 95% CI 1.09 – 1.26, $p = 1.2 \times 10^{-4}$) (**Figure 2.2**).

Located 18.9 kb 3' of rs10210149 is *GPR45*, which is a G-protein coupled receptor. Little is known about the function of *GPR45* itself, but we can surmise potential functions of this gene based on what we know about the family to which it belongs. *GPR45* belongs to a family of proteins that mediate signals to the interior of the cell by the activation of heterotrimeric G proteins that in turn activate various effector proteins, ultimately resulting in a physiological response (Marchese *et al.*, 1999). The human genome encodes thousands of G protein coupled receptors, 150 of which still have unknown functions (Vassilatis *et al.*, 2003). *GPR109A* has been implicated as a tumor suppressor gene in colorectal cancer (Thangaraju *et al.*, 2009) and somatic mutations in colorectal tumors have been identified in *GPR112* and *GPR158* (Wood *et al*., 2007) other G-protein coupled receptors, suggesting that this family of proteins may be important in carcinogenesis.

The gene *TGFBRAP1* is also located in this candidate region, 44.2 kb 3' of rs10210149 and 23.6 kb 3' of *GPR45*. *TGFBRAP1* encodes for a protein that associates with the TGFBR complex and its primary binding partner is TGFBR2 (Wurthner *et al.*, 2001). Disruption of the TGFB pathway has been extensively reported to result in colorectal tumorigenesis, including mutations in *TGFB*, *TGFBR1*, and *TGFBR2* (Piard *et al.*, 2002). The likelihood that rs10210149 captures mutations in *TGFBRAP1* is low based on the LD structure of the region, discussed below. However, since mutations in *TGFBRAP1* could plausibly result in CRC, analysis of this gene is included in this chapter.

The second candidate region is captured by rs16931815, which is located in intron 1 of *STK38L*. This gene encodes a key positive regulatory protein of AMPK-related protein kinase 5 (*ARK5*) in the insulin-like growth factor 1 (*IGF-1*) pathway, which controls cellular processes including cell growth, mitosis, and apoptosis (Suzuki *et al.*, 2006). *IGF-1* receptor binding initiates a signaling cascade that results in the auto-phosphorylation of *STK38L,* and ultimately causes the phosphorylation and activation of *ARK5*. ARK5 has been shown to promote tumor invasion and metastasis as well as to protect tumor cells from nutrient starvation-induced death (Suzuki *et al.*, 2003; Suzuki *et al.*, 2004). Expression of *ARK5* has specifically found to be up-regulated in colorectal tumors (Kusakai *et al.*, 2004), while *STK38L* expression has been shown to be up-regulated in highly metastatic, non-small-cell lung-cancer cell lines (Hergovich *et al.*, 2006), making *STK38L* a plausible susceptibility gene for colorectal cancer.

In this chapter, I focus on the biologically relevant hypotheses resulting from the MECC GWAS. *GPR45*, *TGFBRAP1*, and *STK38L* are studied to investigate the genetic

and functional basis of the association between rs10210149, rs16931815 and risk of colorectal cancer.

2.2 Study design and methods

2.2.1 Study design

This study was separated into two parts: 1) bi-directional Sanger sequencing and 2) allele-specific expression analysis of genes within candidate loci. The candidate loci resulting from the MECC GWAS were rs10210149 on chromosome 2q11.2-q12, rs16931835 on chromosome 12p11.23, and the surrounding regions of DNA in linkage disequilibrium with these two SNPs. We chose to focus this study on only the genes within these regions to be able to conduct genetic and functional analyses. The genes of interest in this region are *GPR45* (Chr2q11.2-q12), *TGFBRAP1* (Chr2q12.1) and *STK38L* (Chr12p11.23). Sanger sequencing was performed for the exons and exon/intron boundaries for each of these genes. This approach does not allow us to detect functional variants within introns or in regulatory regions outside of the coding region. However, the goal of sequencing was to search within the coding regions for potentially pathogenic variants and to describe the extent of variation for subsequent allele-specific expression analyses.

Allele-specific expression analysis was performed for *GPR45*, *STK38L*, and *TGFBRAP1* in the second part of this study. The goal of this analysis was to assess the downstream effects of unknown functional mutations captured by rs10210149 and rs16931815 in the MECC GWAS. We hypothesized that the functional mutations captured by these SNPs may result in subtle differences in expression between alleles of

the genes in these regions, a mechanism previously described in the etiology of a subset of CRCs (Castellsague *et al.*, 2010; Yan *et al.*, 2002). This approach partially circumvents the limitations of restricting phase one sequencing to exonic regions, since it enables us to detect the effects of *cis*-acting variants that affect the relative expression of alleles for one of these genes. However, this approach would be unable to detect the effects of *trans*-acting variants at either of these two loci.

2.2.2 Subjects

Subjects for this study were selected from the Molecular Epidemiology of Colorectal Cancer (MECC) study. The Molecular Epidemiology of Colorectal Cancer (MECC) study is a population-based case-control study of incident cases of colorectal cancer in the Northern and Haifa districts of Israel (**Figure 2.3**). The MECC study is designed to take advantage of the relative ethnic homogeneity and corresponding high incidence rates of CRC found in Israel.

The risk of CRC in Israelis varies widely by ethnicity and country of origin (Fireman *et al.*, 2001). The 5-year age-standardized rate of CRC in Askenazi Jews (born in Europe or America) is 41.9/100,000 cancers compared to 25.5/100,000 for Sephardic Jews (those born in Asia or Africa). Jews born in Israel have an intermediate rate of 32.8/100,000. Non -Jews in Israel had the lowest CRC rate of 10.1/100,000 cancers. Subsequently, Ashkenazi Jews are estimated to have a 3.1-fold higher risk of CRC (95% CI 2.2 – 4.3) compared to non-Ashkenazi Jews (Bat *et al.*, 1986).

The populations of the Haifa and Northern districts of Israel serve as the source population for the MECC study. All individuals in northern Israel were eligible for the

study. MECC subjects were identified through the five main hospitals in northern Israel (Carmel, Rambam, BenZion, Nahariya, Afula) and the Kupat Holim Clalit (KHC) National Center for Cancer Control database. Cases were identified through rapid case ascertainment in the hospitals and through the KHC tumor registry by ICD code for cancer of the colon or rectum diagnosed between May 31, 1998 and March 31, 2004. The five hospitals used to identify cases in the MECC study provide care for more than 65% of individuals diagnosed with CRC in northern Israel, and cases were compared to registry incidence data to assure that ascertainment was representative of the general population.

Controls were identified through a comprehensive database of KHC enrollees and were matched to cases by age, sex, clinic, and Jewish ethnicity. Individuals with a prior CRC diagnosis were ineligible for inclusion as controls. Controls in the KHC database are assumed to be representative of the general Israeli population, which is reasonable since the database covers 60 – 65% of the population.

Data collected for the MECC study included biosamples, structured in-person interviews, and pathology reviews. Biosamples collected include blood, frozen tumor samples, and formalin-fixed paraffin embedded tumor blocks. Structured interviews were conducted with cases and controls on demographic background, medical history, family history, reproductive history, medications, health habits and nutrition (including a food frequency questionnaire). Blood samples were processed to obtain DNA and lymphocytes, of which a subset are stored at the University of Michigan.

511 Ashkenazi Jewish case-control pairs were selected for Sanger sequencing of *GPR45*. These subjects were also the same 511 pairs that comprised phase 1 of the

MECC GWAS, described above. Ashkenazi Jewish subjects were selected to increase the genetic homogeneity of the sample. 24 Ashkenazi Jewish cases that were heterozygous for rs16931815 with a self-reported family history of colorectal cancer were selected for sequencing of *STK38L*. 20 Ashkenazi Jewish cases for whom *GPR45* had been sequenced in this study and were also heterozygous at rs10210149 were sequenced for *TGFBRAP1*. For both *STK38L* and *TGFBRAP1*, only cases were selected to screen for potential mutations. Cases and not controls were selected for initial sequencing of *STK38L* and *TGFBRAP1* because the primary goal of this analysis was to identify potentially pathogenic mutations, not to estimate the association between variants and risk of colorectal cancer.

MECC subjects were chosen for allele-specific expression analysis using the following selection criteria: 1) heterozygous for an exonic SNP within the gene to be measured, 2) lymphocytes available at the University of Michigan, 3) Ashkenazi Jewish, 4) microsatellite stable tumor (cases only), and 5) no known mutations. Again, Askenazi Jewish subjects were chosen to increase the genetic homogeneity of the sample. We also excluded cases with known causes of disease. Microsatellite instability is a characteristic phenotype in tumors with defective DNA mismatch repair; thus cases with microsatellite instable tumors were excluded. Similarly, subjects with known mutations in mismatch repair genes or *APC* were also excluded. 49 cases and 64 controls were selected for analysis of ASE at *GPR45* among all eligible subjects genotyped for the GWAS SNP rs10210149. 27 cases and 22 controls were selected for analysis of ASE at *STK38L* among all eligible subjects genotyped for the GWAS SNP rs16931815. All cases (n=49) and controls (n=46) heterozygous for rs2241801 with available cDNA at the conclusion

of the *GPR45* and *STK38L* analyses were selected for analysis of ASE at *TGFBRAP1*. We had 92% power to detect a 10% difference in ASE at GPR45, 59% power to detect a 10% difference in ASE at STK38L, and 84% power to detect a 10% difference in ASE at TGFBRAP1 (**Figure 2.4**). Power was calculated under the two-sample t-test model for a two-sided hypothesis test at α=0.05., which assumes that the asymptotic relative efficiency of the two-sample t-test relative to the Wilcoxon rank sum test is close to 1.

2.2.3 MECC genome-wide association study

We conducted a GWAS study of CRC in the Molecular Epidemiology of Colorectal Cancer (MECC) study. The MECC genome-wide association study was implemented in three phases of genotyping and analysis (**Figure 2.5**). Phase 1 consisted of pooled-genotyping of 511 Ashkenazi Jewish case-control pairs, randomly selected from all Ashkenazi Jewish matched pairs in the MECC study. These subjects were divided into 6 case pools and 6 controls pools. Five cases pools were comprised of approximately 94% colon cancers cases and 6% rectal cancer cases, and one pool was specifically enriched for rectal cancers (100%). All subjects were microsatellite stable or low. Subjects were genotyped for more than 350,000 SNPs at Perlegen (Mountain View, CA). Phase 1 results were analyzed by estimating the difference in allele frequencies between case and control pools for each SNP, utilizing an inflation factor to account for the pooled genotyping. No SNPs were statistically significant at the genome-wide level after correction for multiple testing. However, we proceeded to phase 2 by selecting the 3,500 SNPs with the largest allele frequency differences.

In phase 2, we genotyped 1,500 case-control pairs for 3,500 SNPs at Perlegen

(**Figure 2.6**). The ethnic distribution of these subjects was representative of the total MECC study, including Ashkenazi Jews, Sephardi Jews, and Christian Arabs. Again, no SNPs were statistically significant in standard tests of association using a log-addtive model after correction for 3,500 tests. We prioritized 25 SNPs for phase 3 external replication, conducted at the Translational Genomics Research Institute (Phoenix, AZ) using Illumina GoldenGate genotyping technology. SNPs were prioritized based on statistical significance in phase 2, statistical significance of closely linked SNPs in GWA studies conducted by Tenesa et al. and Zanke et al., and proximity to probable candidate genes. 22 of these 25 SNPs were successfully genotyped in 1,866 additional MECC subjects, 733 subjects from Spain (Moreno *et al.*, 2006), and 6,812 subjects from Germany (Brenner et al., 2006). While the p-values resulting from the MECC GWAS analysis were not significant at the genome-wide level, we proceeded by prioritizing SNPs with evidence of association in each phase of the MECC GWAS to focus our exploration of these signals.

2.2.4 Sanger sequencing and genotyping

Sequencing of *GPR45*, *STK38L*, and *TGFBRAP1* was performed using DNA extracted from lymphocytes with the primers given in **Table 2.2**. The PCR reaction mixtures (20μL) contained 5ng of genomic DNA, 2μl of 10X PCR buffer (Applied Biosystems), 1.6μL of 25mM MgCl2 (Applied Biosystems), 0.8μL each of 10mM dNTP (New England Biolabs) and 10μM forward and reverse primers, and 1 U of AmpliTaq Gold DNA polymerase (Applied Biosystems). Cycling conditions were as follows: Initial denaturation at 95°C for 3min, 15 cycles of 95°C for 30sec, 70°C for 45sec (-1° every

cycle), 72°C for 1min10sec, 20 cycles of 95°C for 30sec, 55°C for 45sec, 72°C for 1min10sec, and a final extension at 72°C for 10min. PCR products were sequenced at the University of Michigan DNA Sequencing Core, and Mutation Surveyor Software (SoftGenetics, LLC., State College, PA, USA) was used to detect variants. Variants were analyzed for potential pathogenicity using Polyphen (http://sift.jcvi.org/) and SIFT (http://genetics.bwh.harvard.edu/pph/) for coding variants. Genotypes at rs35946826 in *GPR45*, rs10842902 in *STK38L*, and rs2241801 in *TGFBRAP1* were determined by Sanger sequencing of exon 1, the 3' UTR, and exon 2, respectively.

2.2.5 Allele-specific expression quantification

RNA was extracted from frozen lymphocytes of MECC subjects using a trizol-chloroform extraction protocol. RNA quality was assessed by gel electrophoresis and UV spectrometry. All RNA samples were treated with 133 units of DNase-I at 65°C for 10 minutes to eliminate DNA contamination. cDNA was generated in a 20μL reverse transcriptase reaction in the following proportions: 500ng of RNA, 4μL 5X buffer (Invitrogen), 2μL each of 2mM dNTP (New England Biolabs), p[dN]6 random primers (Roche), and DTT (Invitrogen), 0.5μL RNase out (Invitrogen), and 1uL of M-MLV reverse transcriptase (Invitrogen). Quality of cDNA was assessed by PCR amplification of a 238bp product from the *GAPDH* transcript, with the forward primer (GAGTCAACGGATTTGGTCCT) specific to the junction between exons 2 and 3 and the reverse primer (TTGATTTTGGAGGGATCTCG) specific to exon 5 of *GAPDH*.

cDNA and gDNA were PCR amplified in triplicate using the primers given in **Table 2.3**. ASE was measured at rs35946826 for *GPR45*, rs10842902 for *STK38L*, and

rs2241801 for *TGFBRAP1*. The PCR reaction mixtures (25µL) contained 5ng of genomic DNA or 1µL of cDNA, 2.5µl of 10X PCR buffer (Applied Biosystems), 2µL of 25mM MgCl2 (Applied Biosystems), 1.25µL of 2.5mM dNTP (New England Biolabs), 0.5µL of each 10µM primer, and 130.75 U of AmpliTaq Gold DNA polymerase (Applied Biosystems). Cycling conditions are as follows: Initial denaturation at 95°C for 3min, 50 cycles of 95°C for 345sec, 60°C for 45sec, 72°C for 45sec, and a final extension at 72°C for 10min. 5µL of PCR product was used for Pyrosequencing according to the standard Streptavadin- Sepharose bead-capture protocol provided by Qiagen. Results were included in analysis only if assigned a quality score of "check" or "pass" by the PyroMarkMD software using default settings.

2.2.6 Statistical methods

Descriptive characteristics were assessed using frequency tables and the means procedure in SAS (version 9.2). Statistically significant differences in clinical and epidemiologic characteristics between cases and controls were determined using chi-square tests for categorical variables and analysis of variance (ANOVA) for continuous variables.

Allele-specific expression was calculated as

$$\mathrm{ASE} = \frac{\left[\sum_{i=1}^{n}(\%\ \text{Major Allele}/\%\ \text{Minor Allele})_{\mathrm{cDNA}i}\right]/n}{\left[\sum_{j=1}^{m}(\%\ \text{Major Allele}/\%\ \text{Minor Allele})_{\mathrm{gDNA}_j}\right]/m},$$

where n indicates the number of pyrosequencing replicates performed for cDNA and m indicates the number of pyrosequencing replicates performed for genomic DNA (gDNA).

Allele-specific expression was analyzed as a continuous variable using the Wilcoxon rank-sum statistic by case-control status and the Kruskal-Wallis Test by GWAS SNP genotype. Trend in allele-specific expression by GWAS SNP genotype was measured using linear regression.

2.3 Results

2.3.1 Identification of candidate regions

The goal of this study was to gain insight into the functional consequences of genetic variants associated with risk of colorectal cancer in the MECC GWAS. Of the 22 SNPs genotyped in phase 3, none reached statistical significance in an analysis including samples from phases 2 and 3 after Bonferroni correction for 3,500 tests (**Table 2.1**). We proceeded by prioritizing SNPs based on those with evidence for association in each phase of the MECC GWAS. The top two most significant SNPs were rs10210149 on chromosome 2 and rs16931815 on chromosome 12. In a log-additive model, each copy of the C allele of rs10210149 was associated with a 12% increase in risk of colorectal cancer (OR=1.12, 95% CI 1.07 – 1.18, $p = 2.0 \times 10^{-6}$). Similarly, each copy of the A allele of rs16931815 was associated with an 18% increase in risk of colorectal cancer (OR=1.18, 95% CI 1.09 – 1.26, $p = 1.2 \times 10^{-4}$). Rs10210149 on chromosome 2q11.2-q12, rs16931835 on chromosome 12p11.23, and the surrounding regions of DNA in linkage disequilibrium with these two SNPs are considered the candidate loci from the MECC GWAS in this chapter.

2.3.2 Sequencing of genes within candidate regions

We first examined the coding regions of three genes (*GPR45*, *TGFBRAP1*, *STK38L)* within these candidate loci in an attempt to identify the functional variants captured by rs10210149 and rs16931815. *GPR45*, a 1.725 kb gene comprised of a single exon, is located on chromosome 2q11.2-q12. In HapMap samples of European ancestry (http://hapmap.ncbi.nlm.nih.gov/), variants flanking *GPR45* are in strong linkage disequilibrium with the GWAS SNP rs10210149. No polymorphisms within *GPR45* were genotyped in HapMap samples of European ancestry. Rs10210149 lies 18.9 kb 5' of *GPR45* and is in high linkage disequilibrium with SNPs spanning a 25.8 kb region that includes *GPR45* (**Figure 2.7**). Among 29 SNPs within this region, the average D' with rs10210149 is 0.931 and the average R^2 is 0.232. This indicates that mutations within this region are likely to be captured by variation in rs10210149.

We first fully sequenced *GPR45* as well as 447 bp 5' and 209 bp 3' of the gene to search for mutations in the 511 Ashkenazi Jewish case-control pairs from Phase 1 of the MECC GWAS. We identified 10 variants within and around *GPR45* using Sanger sequencing (**Table 2.4**). Eight of the ten variants identified were relatively rare with minor allele frequencies (MAF) no greater than 2%, and five of the ten variants had been previously reported in dbSNP (www.ncbi.nlm.nih.gov/projects/SNP/). Only three variants were within the coding region, and of these three only rs35946826 coded for a non-synonymous amino acid change (L312F). This change from leucine to phenylalanine was not predicted to be damaging to the GPR45 protein by either SIFT or PolyPhen.

To better understand the role of L132F as a candidate mutation, we estimated the association between this SNP and risk of colorectal cancer in the sequenced subjects.

Rs35946826 was associated with a 1.34-fold increase in risk (95% CI 1.05 – 1.71, p=0.02), which is consistent in magnitude and direction with the GWAS SNP rs10210149. We next conducted a haplotype analysis using the expectation-maximization algorithm to describe the relationship between these two SNPs. A single haplotype was associated with risk of CRC, which was perfectly captured by the C allele of rs102101049. The C-C (rs10210149-rs35946826) haplotype was associated with a 1.45-fold increase in risk of CRC (95% CI 1.12 – 1.87, p=0.004) compared to the T-T haplotype. This analysis indicates that the C allele of the GWAS SNP rs10210149 better captures risk of colorectal cancer than does the C allele of rs39546826.

We next looked at *TGFBRAP1*, which is located 23.6 kb 3' of *GPR45* and 44.2 kb 3' of the GWAS SNP rs10210149 (**Figure 2.7**). *TGFBRAP1* is comprised of 12 exons on the reverse strand of chromosome 2q12.1. In HapMap samples of European ancestry, variation within *TGFBRAP1* is not strongly associated with rs10210149. Among 92 SNPs in this 62.6 kb gene, the average D' with rs10210149 is 0.340 and the average R^2 is 0.017. Thus, variants within *TGFBRAP1* are less likely to be captured by rs10210149 compared to variants within *GPR45*, and *GPR45* remains the most likely susceptibility gene in this region of chromosome 2.

We next sequenced the 11 exons and exon/intron boundaries of *TGFBRAP1* in 20 Ashkenazi Jewish MECC cases for whom *GPR45* had been sequenced to both search for functional variants and to characterize the variation in this gene. A total of 15 variants were identified among these 20 cases (**Table 2.4**), of which 12 had been previously reported in dbSNP or the Ensembl database (http://ensembl.org/). Six variants were in the coding region of the gene, and only one of these coded for a non-synonymous amino

acid change. This previously unreported variant in exon 11, EX11-43C>T, coded for a change from histidine to argine and was present in a single subject. Both the SIFT and PolyPhen programs predicted this amino acid change to be benign. Due to the limited number of cases sequenced for *TGFBRAP1*, it is likely that we would not have been able to identify any functional variants in this gene. However, we decided to proceed with functional analysis of *TGFBRAP1* rather than continue an exhaustive search for a functional mutation in this low-priority gene.

The third gene of interest identified from the MECC GWAS, *STK38L*, is located on chromosome 12p11.23. The GWAS SNP rs16931815 is located within intron 1 of this gene and is subsequently in high linkage disequilibrium with *STK38L* variation (**Figure 2.8**). *STK38L* consists of 14 exons and spans 81.4 kb. Fifty-one SNPs were genotyped in HapMap samples of European ancestry, and among these the average D' with rs16931815 is 0.83 and the average R^2 is 0.088. Rs16931815 is also in linkage disequilibrium with the gene *ARNTL2*, located 7.5 kb 3' of *STK38L*. However, we chose to focus our analyses on *STK38L* since rs16931815 lies within the gene.

We next sequenced all 14 exons and exon/intron boundaries of *STK38L* in 24 Ashkenazi Jewish MECC cases with a family history of colorectal cancer. Only 5 variants were identified, and four of these SNPs had been previously reported in dbSNP (**Table 2.4**). Although all 5 variants were relatively common with MAFs between 0.125 and 0.210, none were located in the coding region of the gene. Again, it is likely that we were unable to identify a mutation in this gene due to the limited number of cases sequenced. However, we decided to also proceed with functional analysis of *STK38L* rather than continue a comprehensive search for a functional variant in this gene.

2.3.3 Allele-specific expression analyses of GPR45, TGFBRAP1, & STK38L

Since we were unable to detect any mutations in *GPR45, TGFBRAP1,* or *STK38L*, we decided to pursue functional analysis of these genes. This approach allows us to ascertain whether *GPR45, TGFBRAP1,* or *STK38L* are targets of the functional variants captured in the MECC GWAS without identifying the functional variants themselves. Thus, functional analysis in this study offers a method of identifying genes and their corresponding regulatory regions for future high-coverage sequencing to identify causal mutations. While it is possible that the genes affected by functional variants in these two loci lie outside of the candidate regions on chromosomes 2 and 12, we chose to first proceed with analyses of *GPR45*, *TGFBRAP1*, and *STK38L*.

Given the modest magnitudes of association observed at both rs10210149 and rs16931815, we hypothesized that the effects of the mutations captured by these SNPs may also be modest. More specifically, we proposed that the underlying functional variants could cause subtle changes in gene expression that result in the attenuation of expression of one allele. This process, known as allele-specific expression (ASE), is widespread in the normal human genome (Lo *et al.*, 2003) and has been proposed as a mechanism involved in susceptibility to complex diseases (Knight, 2005). Several examples have been reported, including *DAPK1* and chronic lymphocytic leukemia (Lynch *et al.*, 2002) and *APC* and familial adenomatous polyposis (Yan *et al.*, 2002). In this chapter, we conducted allele-specific expression analyses for *GPR45*, *TGFBRAP1*, and *STK38L* in MECC subjects.

The primary goal of these allele-specific expression analyses was to identify

changes in expression associated with variation at the GWAS SNPs. To implement this method, we took advantage of the characterization of genetic variation by Sanger sequencing to identify exonic variants. To be informative for ASE, subjects had to be heterozygous at the exonic SNP measured. A total of 113 Ashkenazi Jewish MECC participants with no known mutations and microsatellite stable tumors (cases only) were informative for ASE analysis of rs35946826 in exon 1 of *GPR45* (**Table 2.5**). As described above, rs35946826 codes for a leucine to phenylalanine change at amino acid 312 of the GPR45 protein. Rs35946826 is in linkage disequilibrium with the GWAS SNP rs10210149 with a D' of 0.958 and an R^2 of 0.47; specifically, the C allele of rs35946826 is in LD with the C (risk) allele of rs10210149.

Among these 113 informative subjects, 49 were colorectal cancer cases (43.4%) and 64 were controls (56.6%). The mean age of these subjects was 74.2 years (standard deviation (SD) = 10.1 years) and did not significantly differ by GWAS SNP genotype (p=0.12). Allele-specific expression may be influenced by both genetic environmental factors. While the subjects analyzed for ASE were not randomly selected from the MECC study and the sample sizes are small, we were interested in identifying any obvious environmental confounders of the relationship between GWAS SNP genotype and ASE values. The variable examined here were smoking, vegetable intake, and aspirin/NSAID use. Sex, smoking (ever and pack-years), vegetable intake, and daily aspirin/NSAID use were not significantly different by GWAS SNP genotype.

Allele-specific expression was measured in cDNA generated from lymphocyte mRNA for *GPR45* at rs35946826 using pyrosequencing technology. The relative expression of the C allele compared to the T allele in cDNA was normalized to the

relative expression of these alleles in genomic DNA to account for any assay-specific differences in allelic amplification. Among all 113 subjects, the mean increase in expression of the C allele compared to the T allele was 32% (ASE = 1.32, SD = 0.68). The minimum value of ASE corresponded to a 0.36-fold decrease in expression of the C allele and the maximum value of ASE corresponded to a 4.62-fold increase in the expression of the C allele. When analyzed by rs10210149 genotype, a significant difference in median ASE values was observed (Kruskal-Wallis p = 0.03). For each additional copy of the C allele of rs10210149, there was a 27% increase in the expression of the C allele of rs35946826 (p trend = 0.01) (**Figure 2.9a)**. No difference was observed in ASE by case-control status (Wilcoxon p=0.28) (**Figure 2.9b**). The relationship between GPR45 ASE, case-control status, and rs10210149 genotype were consistent when analyzed on a log scale. No difference in median ASE values was observed between cases and controls (Wilcoxon p = 0.28). Rs10210149 was highly significantly associated with log-ASE (Kruskal-Wallis p = 0.016; p trend = 0.0049).

We next measured allele-specific expression for *TGFBRAP1* at rs2241801 to investigate whether rs10210149 affected expression at both *GPR45* and *TGFBRAP1*. A total of 95 Ashkenazi Jewish MECC participants that met selection criteria were informative for ASE analysis of rs2241801 in exon 2 of *TGFBRAP1* (**Table 2.6**). rs2241801 displays low LD with the GWAS SNP rs10210149 with a D' of 0.098 and an R^2 of 0.005. Among these 95 subjects, 49 were colorectal cancer cases (51.6%) and 46 were controls (48.4%). The mean age of these subjects was 72.1 years (SD = 9.8 years) and did not significantly differ by GWAS SNP genotype (p=0.31). Subjects heterozygous for the risk allele were 5.16 times more likely to be ever smokers (95% CI 1.28 – 20.77,

p=0.02) and subjects homozygous for the risk allele were 3.11 more likely to be ever smokers (95% CI 0.73 – 13.20, p=0.12) compared to subjects homozygous for the non-risk allele of rs10210149. Sex, vegetable intake, and daily aspirin/NSAID use were not significantly different by GWAS SNP genotype.

Allele-specific expression was measured for *TGFBRAP1* at rs2241801 using the method described above, quantified as the relative expression of the A allele compared to the G allele. Among all 95 subjects, the mean increase in expression of the A allele compared to the G allele was 7% (ASE = 1.07, SD = 0.29). The minimum value of ASE corresponded to a 0.03-fold decrease in expression of the A allele and the maximum value of ASE corresponded to a 2.2-fold increase in the expression of the A allele. No significant difference in ASE values was observed by either rs10210149 genotype (Kruskal-Wallis p = 0.98) or case-control status (Wilcoxon p=0.29) (**Figure 2.10**). ASE values were also not significantly different by smoking history (Wilcoxon p=0.32). Thus, we conclude that rs10210149 variation is associated with differences in expression of *GPR45* but not *TGFBRAP1*, which is consistent with the LD structure in this region of chromosome 2. The relationship between *TGFBRAP1* ASE, case-control status, and rs10210149 genotype were consistent when analyzed on a log scale. No difference in median ASE values was observed between cases and controls (Wilcoxon p = 0.29) or by rs10210149 genotype (Kruskal-Wallis p = 0.98).

We then evaluated allele-specific expression of genes within candidate loci identified from the MECC GWAS by examining *STK38L*. A total of 49 Ashkenazi Jewish MECC participants that met selection criteria were informative for ASE analysis of rs10842902 in the 3' UTR of *STK38L* (**Table 2.7**). rs10842902 displays very high LD

with the GWAS SNP rs16931815 with a D' of 1 and an R^2 of 0.14. Among these 49 informative subjects, 27 were colorectal cancer cases (55.1%) and 22 were controls (44.9%). The mean age of these subjects was 73.0 years (SD = 8.0 years) and did not significantly differ by GWAS SNP genotype (p=0.31). Sex, vegetable intake, smoking (history and pack-years) and daily aspirin/NSAID use were also not significantly different by GWAS SNP genotype.

Allele-specific expression was quantified for *STK38L* at rs10842902 as the relative expression of the A allele compared to the G allele. Among all 49 subjects, the mean increase in expression of the A allele compared to the G allele was 10% (ASE = 1.10, SD = 0.60). The minimum value of ASE corresponded to a 0.01-fold decrease in expression of the A allele and the maximum value of ASE corresponded to a 4.0-fold increase in the expression of the A allele. No significant difference in ASE values was observed by either rs16931815 genotype (Kruskal-Wallis p = 0.41) or case-control status (Wilcoxon p=0.44) (**Figure 2.11**). The relationship between *STK38L* ASE, case-control status, and rs10210149 genotype were consistent when analyzed on a log scale. No difference in median ASE values was observed between cases and controls (Wilcoxon p = 0.44) or by rs10210149 genotype (Kruskal-Wallis p = 0.41). Thus, we conclude that rs16931815 variation is not strongly associated with differences in expression of *STK38L* alleles and does not explain the GWAS signal in this sample of the MECC population. Further analyses are required to understand the functional consequences associated with variation on chromosome 12p11.23 and risk of colorectal cancer.

2.4 Discussion

Genome-wide association studies of colorectal cancer in England, Scotland, Canada, and international replication populations have identified ten new low-penetrance susceptibility loci (Tenesa and Dunlop, 2009), but the causal variants at any these loci have yet to be uncovered despite extensive fine mapping and resequencing. This suggests that the underlying causal variants affect gene expression. Five of these ten GWAS variants capture genes in the TGFB signaling pathway, such as *SMAD7* (Broderick *et al.*, 2007; Tenesa *et al.*, 2008), *BMP2* (Jaeger *et al.*, 2008), and *BMP4* (Houlston *et al.*, 2008), suggesting a key role for this pathway in colorectal cancer susceptibility. Additionally, one of these GWAS variants on 8q24, rs6983267, has been shown to affect the regulation of c-MYC via a long-range enhancer (Sotelo *et al.*, 2010). Elucidating the functional consequences of these GWAS variants has been challenging, but provides an opportunity for understanding the mechanisms of colorectal cancer.

The results presented in this chapter are consistent with the conclusions drawn from the GWAS studies and subsequent functional analyses described above. We have identified a significant trend in allele-specific expression of *GPR45* that is associated with rs1021019. To better understand the significance of ASE at *GPR45*, there are several methods we could employ. First, we could examine the colorectal tumors of the MECC patients with extreme ASE values for loss of heterozygosity (LOH) to identify whether there is loss of the normally expressed allele. This would argue for a direct functional relationship between over-expression of *GPR45* and risk of colorectal cancer. We could additionally perform high-coverage sequencing at this locus to better define the ASE-associated haplotype and to potentially fine-map the disease-causing variant. To better understand risk of CRC associated with rs16931815, we could perform sequencing

in a larger number of subjects for both *STK38L* and *ARNTL2*, including the intronic and 5' regions for these genes.

While ASE analysis was successful in identifying subtle changes in *GPR45* associated with a GWAS SNP, it is unlikely that this method will be widely employed for large-scale GWAS candidate gene screening. This is due not only to the time and resource-intensive nature of this analysis, but also to the fact that many of the variants identified in GWA studies are in "gene-deserts". These variants are likely to be located in regulatory regions or non-coding RNA gene, but identifying genes to analyze as potential targets of these regulatory SNPs would be extremely challenging. Nevertheless, we will continue to investigate the genetic and functional basis of these associations.

One limitation of this study is that Sanger sequencing was performed for *TGFBRAP1* and *STK38L* in a small number of subjects and was restricted to the coding regions of these genes. The sample size for this analysis was too small to confidently conclude that no functional variants exist in the coding regions of these two genes. Additionally, we would have missed any variants that were located in the promoter region or other proximal regulatory regions, as well as functional variants located within introns.

A second limitation of this study is that we only used one SNP per gene to measure ASE. It is possible that SNP location influences gene expression patterns. However, our choice to measure only one SNP per gene was in large part determined by the patterns of variation within these genes. Specifically, there were very few exonic SNPs with minor allele frequencies in *GPR45* and *STK38L* greater than 5%. Selecting SNPs for analysis with small minor allele frequencies would have led to small sample

sizes, limited by the number of informative subjects that also met the other selection criteria.

Finally, we had little power to detect very small changes in ASE in this study. For example, we had only 40% power to detect a 5% difference in allele-specific expression of *GPR45* and 13% power to detect a 2.5% difference in allele-specific expression for this gene. Similarly, we had only 30% power to detect a 5% difference in allele-specific expression of *TGFBRAP1* and 10% power to detect a 2.5% difference in allele-specific expression for this gene. We had the lowest power for the *STK38L* analyses, where we had only 20% power to detect a 5% difference in allele-specific expression of *GPR45* and 6.7% power to detect a 2.5% difference in allele-specific expression for this gene. Clearly, very subtle changes in the relative expression of alleles would not be detected in this study, which could still potentially have a causal effect on colorectal carcinogenesis.

Table 2.1 22 SNPs genotyped in Phase III of the MECC GWAS

			Phase 2		Phase 2 + 3 combined	
SNP	Gene	Description	OR (95%CI)	p	OR (95%CI)	p
rs544670		1MB 5' of *PI3K*	1.04 (0.95 - 1.14)	0.47	1.07 (1.00 – 1.16)	$6.0x10^{-2}$
rs10210149		5' of *GPR45* and *TGFBRAP1*	1.07 (1.01 - 1.14)	$7.1x10^{-3}$	1.12 (1.07 - 1.18)	$2.0x10^{-6}$
rs2193075	PARD3B	PARD3B par-3 partitioning defective 3 homolog B	1.04 (0.98 - 1.11)	0.15	1.03 (0.98 – 1.09)	$7.5x10^{-2}$
rs2016993		near zinc finger proteins	1.02 (0.95 - 1.08)	0.61	1.05 (1.00 – 1.11)	$1.0x10^{-3}$
rs313587		gene dessert	1.04 (0.99 - 1.10)	0.14	1.09 (1.04 – 1.14)	$8.4x10^{-4}$
rs7733404	MCC	mutated in colorectal cancers	1.07 (1.00 - 1.13)	$4.0x10^{-2}$	1.06 (1.00 – 1.13)	$8.2x10^{-4}$
rs17012429	CNTN3	contactin 3 (plasmacytoma associated)	1.10 (1.00 - 1.21)	$5.7x10^{-2}$	1.14 (1.04 – 1.25)	$5.2x10^{-3}$
rs2576794		5' of *GPR45* and *TGFBRAP1*	1.06 (1.01 - 1.13)	$2.8x10^{-2}$	1.08 (1.03 – 1.14)	$8.8x10^{-4}$
rs4631835	C10orf81	chromosome 10 open reading frame 81	1.05 (0.98 - 1.11)	0.15	1.06 (1.01 – 1.13)	$2.7x10^{-2}$
rs2068452	CDH12	cadherin 12, type 2 (N-cadherin 2)	1.06 (0.89 - 1.26)	0.54	1.06 (0.92 – 1.23)	$3.4x10^{-1}$
rs6034187	C20orf133	chromosome 20 open reading frame 133	1.01 (0.95 - 1.07)	0.76	1.04 (0.99 – 1.10)	$7.7x10^{-2}$
rs3773966	IL1RAP	interleukin 1 receptor accessory protein	1.06 (0.96 - 1.17)	0.27	1.01 (0.92 – 1.10)	8.0x10-1
rs255153		5' of *INMT*	1.03 (0.97 - 1.10)	0.30	1.06 (1.01 – 1.12)	$1.6x10^{-2}$
rs17383284	PSD3	pleckstrin and Sec7 domain containing 3	1.00 (0.94 - 1.07)	0.92	1.03 (0.97 – 1.09)	$4.2x10^{-1}$
rs11647078		gene dessert	1.09 (1.02 - 1.16)	$7.6x10^{-3}$	1.02 (0.97 – 1.07)	$4.8x10^{-1}$
rs9385571	AKAP7	A kinase (PRKA) anchor protein 7	1.04 (0.98 - 1.10)	0.21	1.06 (1.01 – 1.11)	$2.9x10^{-2}$
rs6980478		3' of *CYP7B1*	1.06 (0.98 - 1.15)	0.13	1.09 (1.02 – 1.16)	$7.2x10^{-3}$
rs17159640	IFRD1	interferon-related developmental regulator 1	1.04 (0.91 - 1.18)	0.57	1.04 (0.95 – 1.15)	$3.4x10^{-1}$
rs16931815	STK38L	serine/threonine kinase 38 like	1.19 (1.10 - 1.29)	$1.0x10^{-4}$	1.18 (1.09 - 1.26)	$1.2x10^{-4}$
rs8049247	CDH3	cadherin 3, type 1, P-cadherin (placental)	1.08 (1.00 - 1.16)	$5.1x10^{-2}$	1.02 (0.96 – 1.09)	$5.8x10^{-1}$
rs10816788	EPB41L4B	erythrocyte membrane protein band 4.1 like 4B	1.06 (0.98 - 1.15)	0.16	1.00 (0.94 – 1.07)	$9.7x10^{-1}$
rs2689264	FTO	fatso	1.04 (0.95 - 1.13)	0.39	1.05 (0.99 – 1.12)	$9.8x10^{-2}$

Table 2.2 Sequencing primers for *GPR45*, *STK38L*, and *TGFBRAP1*

Gene	Exon	Forward primer	Reverse primer	Product size
GPR45	1	CCTTTCTCTTGTGGAGCAGG	ACAGCATGATGTCGGAGAAG	814
	1	GCAACACTGTGGTCTGCATC	ACTTTGGGGAGGATTTGGAA	897
	1	CCCATCGTCTACTGCTGGA	CGGATGTGCTTCTCACTTCA	859
STK38L	1	ACAGGTTTGGCGTAAAAACG	CTGGACACCCAAAGACACCT	488
	2	TAACTCCTGGTTTTGCCACC	AATTGGCATGTCATACGGGT	592
	3	TTTAGGCAGGAGCGTGAAGT	ATGAAAAGTCACTGGGGGTG	352
	4	TGTGAAAGAGCAACCTTGGA	ATATATTTGCAGCAGTAGTGACTTTT	556
	5	GAGCTTTTGGAGAGGTGTGC	ACTGCGTCAGTGGATGCTC	800
	6	AGAGGCCTCAGCTTCACGTA	CAGTGAACCCAGAACGGTAA	599
	7,8,9	TCCACCTTTGAGGCATTTTC	GAACCAAGGCATAAAATTCTCTT	940
	10,11	GGTTTGACGAGTTGCTCCTT	CTTCCCCCACAAAAGTGAAA	795
	12	TGATAATTTCTCTGTTTCCATGTTG	ACCTTTCCCATTCAAAGCCT	154
	13,14	AGGATGAGAAAGCCTTGGGT	TGGTGATGCAGCTACCTGAG	885
	3'UTR	ACATGACCATGAAGGCTGCT	TCACATTGAGAAATCCCCAG	957
TGFBRAP1	1	CCCTCCTCCTGTGTAGGTGA	CCTGAGTGTGACCCGAATTT	996
	2	GCAGCCTCTGTTTCTGCTTC	TCATCAAGCACTGGTCAAGC	926
	3	CCATGCTTATTTGGAAGCCT	TGCATCCTTAAAGGTTTGGC	555
	4	CAGTTTGGGGAAAGCAGTGT	GCAGTGCCTTCTCAGTCACA	442
	5	GACGTGCATTTGGGAAAGAT	ACCGTATCCACCTGAAGCAC	417
	6	ACATGATTACCCTGTCCCCA	GAGCCTCAGTAAGGGTGCAG	537
	7	GCTGATGGGGAGAGGTTGTA	AATGTACCCAGCTCATTGGC	488
	8	CCTGGCTGATGGTTGTAGGT	CAGACTTCTGAGGGGTCGAG	864
	9,10	GGTTTGGGAAGCACAGTCAC	CCACCCGCTTGATATGAGTT	934
	11	AGGGAGCCAGGTTGACTTTT	CTCTGCCTCTGCTCACACAG	722
	12	CTCAGCCAGACAAGCAACAA	CCAGGCAAGAGAGGACTTTG	872

Table 2.3 Pyrosequencing primers for *GPR45, STK38L,* and *TGFBRAP1*

Gene	SNP	Primer	Sequence
GPR45	rs35946826	Forward	Biotin-TACAGCCTCCTGTCTGTGTTTAGC
		Reverse	GATGGGGTTGAAGACGGACT
		Sequencing	CGGACTTGAGGTAACTGA
TGFBRAP1	rs2241801	Forward	Biotin-ACAGCTGCAGAGACACTTGG
		Reverse	ATGGTTCTGCGTTTGACAGAG
		Sequencing	TGAGTGCTGAGGCCG
STK38L	rs10842902	Forward	Biotin-TTTCCTGTGGGCATGCTGT
		Reverse	TTGCCCTTTAATAAGCTGACCTC
		Sequencing	CCTGTGGGCATGCTG

Table 2.4 Variants identified in GPR45, STK38L, and TGFBRAP1 by Sanger sequencing

GENE	VARIANT	LOCATION	EFFECT	MAF
GPR45	rs17636399	5’		0.223
	5’-160C>G	5’		0.014
	rs17030715	5’		0.014
	5’-31G>A	5’		0.002
	rs2576727	Exon 1	T27T	0.005
	rs56355385	Exon 1	T168T	0.014
	rs35946826	Exon 1	L312F	0.167
	EX1-227G>A	3’ UTR		0.006
	EX1-185G>A	3’ UTR		0.019
	3’+65T>G	3’		0.006
STK38L	rs1615928	Intron 1		0.132
	rs10771336	Intron 10		0.125
	rs2242185	Intron 10		0.167
	IVS10-41C>T	Intron 10		0.125
	rs4369500	Intron 12		0.210
TGFBRAP1	rs2241801	Exon2	R82R	0.158
	ENSSNP5509498	Intron 2		0.155
	rs6709616	Intron 2		0.100
	rs12476720	Exon 3	R240R	0.211
	IVS4-32InsT	Intron 4		0.050
	rs2679833	Intron 5		0.158
	rs2241799	Exon 6	N432N	0.370
	IVS7+191C>T	Intron 7		0.025
	rs2304543	Intron 9		0.211
	rs2250659	Intron 9		0.083
	rs2250658	Intron 9		0.083
	rs11676273	Exon 10	L643L	0.550
	rs2241798	Intron 10		0.211
	rs2241797	Exon 11	H724R	0.100
	EX11-43C>T	Exon 11	L787L	0.025

Table 2.5 Epidemiologic characteristics of subjects in *GPR45* ASE analysis (n=113)

		rs10210429			
	All subjects n (%)	TT n (%)	TC n (%)	CC n (%)	P value
Case	49 (43.4)	16 (44.4)	31 (48.4)	2 (15.4)	
Control	64 (56.6)	20 (55.6)	33 (51.6)	11 (84.6)	0.09*
Age (years)	Mean=74.2 SD=10.1	Mean=71.8 SD=9.9	Mean=74.8 SD=10.6	Mean=78.1 SD=6.6	0.12†
Male	56 (50.0)	15 (41.7)	36 (57.1)	5 (38.5)	
Female	56 (50.0)	21 (58.3)	27 (42.9)	8 (61.5)	0.21*
Ever smoked	51 (46.0)	14 (40.0)	30 (47.6)	7 (53.8)	
Never smoked	60 (54.0)	21 (60.0)	33 (52.4)	6 (46.2)	0.62*
Pack years	Mean=40.8 SD=29.6	Mean=27.4 SD=16.5	Mean=46.3 SD=34.4	Mean=39.0 SD=17.3	0.27†
Veg tert 3	33 (29.2)	12 (33.3)	15 (23.4)	6 (46.1)	
Veg tert 2	37 (32.7)	7 (19.4)	28 (43.8)	2 (15.4)	
Veg tert 1	43 (38.1)	17 (47.2)	21 (32.8)	5 (38.5)	0.06*
Aspirin/NSAID daily yes					
Aspirin/NSAID daily no	51 (46.0) 60 (54.0)	15 (42.9) 20 (57.1)	27 (42.9) 36 (57.1)	9 (69.2) 4 (30.8)	0.23*

* Calculated by Fisher's exact test
† Calculated by ANOVA

Table 2.6 Epidemiologic characteristics of subjects in *TGFBRAP1* ASE analysis (n=49)

		rs10210149			
	All subjects n (%)	TT n (%)	TC n (%)	CC n (%)	P value
Case	49 (51.6)	10 (55.6)	20 (50.0)	15 (50.0)	
Control	46 (48.4)	8 (44.4)	20 (50.0)	15 (50.0)	0.92*
Age (years)	Mean=72.1 SD=9.8	Mean=72.3 SD=10.2	Mean=70.1 SD=10.4	Mean=73.7 SD=7.9	0.31†
Male	49 (53.3)	6 (33.3)	24 (60.0)	17 (60.7)	
Female	43 (46.7)	12 (66.7)	16 (40.0)	11 (39.3)	0.14*
Ever smoked	38 (40.4)	3 (17.7)	21 (52.5)	12 (40.0)	
Never smoked	56 (59.6)	14 (82.3)	19 (47.5)	18 (60.0)	0.05*
Pack years	Mean=38.5 SD=35.2	NA	Mean=35.1 SD=30.7	Mean=56.9 SD=43.4	0.19†
Veg tert 3	33 (34.7)	6 (33.3)	12 (30.0)	13 (43.3)	
Veg tert 2	30 (31.6)	6 (33.3)	18 (45.0)	6 (20.0)	
Veg tert 1	32 (33.7)	6 (33.3)	10 (25.0)	11 (36.7)	0.30*
Aspirin/NSAID daily yes					
Aspirin/NSAID daily no	37 (40.7) 54 (59.3)	7 (41.2) 10 (58.8)	16 (40.0) 24 (60.0)	11 (39.3) 17 (60.7)	0.99*

* Calculated by Fisher's exact test
† Calculated by ANOVA

Table 2.7 Epidemiologic characteristics of subjects in *STK38L* ASE analysis (n=95)

		rs16931815			
	All subjects n (%)	GG n (%)	GA n (%)	AA n (%)	P value
Case	27 (55.1)	2 (100)	15 (51.7)	10 (55.6)	
Control	22 (44.9)	0 (0)	14 (48.3)	8 (44.4)	0.67*
Age (years)	Mean=73.0 SD=8.0	Mean=67.7 SD=13.4	Mean=73.0 SD=7.9	Mean=73.6 SD=7.8	0.60†
Male	28 (58.3)	0 (0)	17 (60.7)	11 (61.1)	
Female	20 (41.7)	2 (100)	11 (39.3)	7 (38.9)	0.31*
Ever smoked	18 (36.7)	0 (0)	10 (34.5)	8 (44.4)	
Never smoked	31 (63.3)	2 (100)	19 (65.5)	10 (55.6)	0.52*
Pack years	Mean=31.5 SD=22.5	NA	Mean=27.2 SD=18.9	Mean=35.8 SD=26.2	0.46†
Veg tert 3	13 (26.6)	0 (0)	8 (27.6)	5 (27.8)	
Veg tert 2	18 (36.7)	1 (50.0)	13 (44.8)	4 (22.2)	
Veg tert 1	18 (36.7)	1 (50.0)	8 (27.6)	9 (50.0)	0.39*
Aspirin/NSAID daily yes					
Aspirin/NSAID daily no	16 (33.3)	0 (0)	11 (39.3)	5 (27.8)	
	32 (66.7)	2 (100)	17 (60.7)	13 (72.2)	0.49*

* Calculated by Fisher's exact test
† Calculated by ANOVA

Figure 2.1 GWAS associations at chromosome 2 The –log10 p-values around the signal at rs10210149 are shown by study. The orange dot indicates the p-value for rs10210149 in the combined Phase 2 and 3 analyses (MECC and replication studies). The white dots indicate p-values from Phase 1 MECC GWAS genotyping, the green dots indicate p-values from Phase 2 MECC GWAS genotyping, the blue dots indicate p-values from a Canadian GWAS (Zanke, et al. 2007), and the red dots indicate p-values from the Greman and Spanish replication samples in Phase 3 of the MECC GWAS.

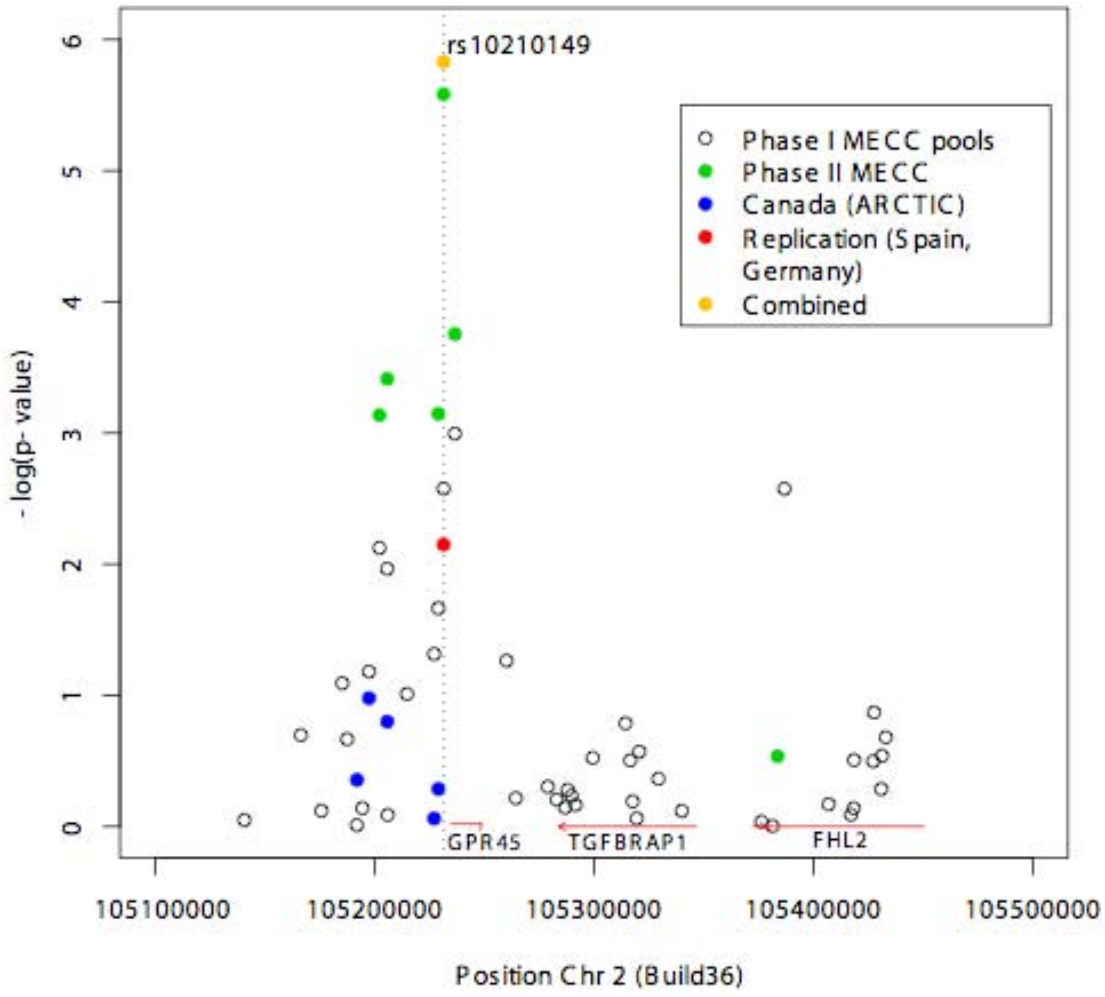


Figure 2.2 GWAS associations at chromosome 12 The –log10 p-values around the signal at rs16931815 are shown by study. The orange dot indicates the p-value for rs16931815 in the combined Phase 2 and 3 analyses (MECC and replication studies). The white dots indicate p-values from Phase 1 MECC GWAS genotyping, the green dots indicate p-values from Phase 2 MECC GWAS genotyping, the blue dots indicate p-values from a Canadian GWAS (Zanke, et al. 2007), and the red dots indicate p-values from the Greman and Spanish replication samples in Phase 3 of the MECC GWAS.

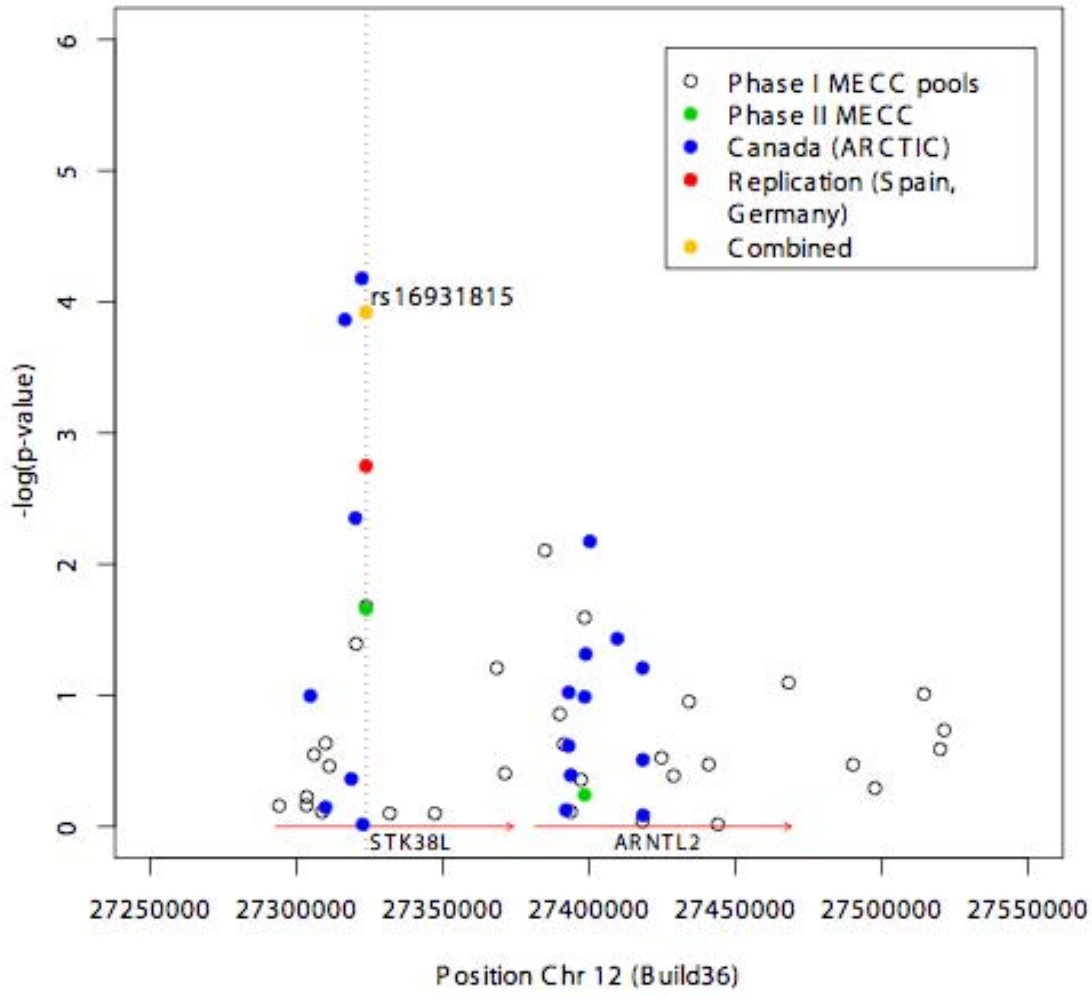


Figure 2.3 Molecular Epidemiology of Colorectal Cancer study design The design of the MECC study, a population-based case-control study of colorectal cancer in the Northern & Haifa ditricts of Israel is shown.

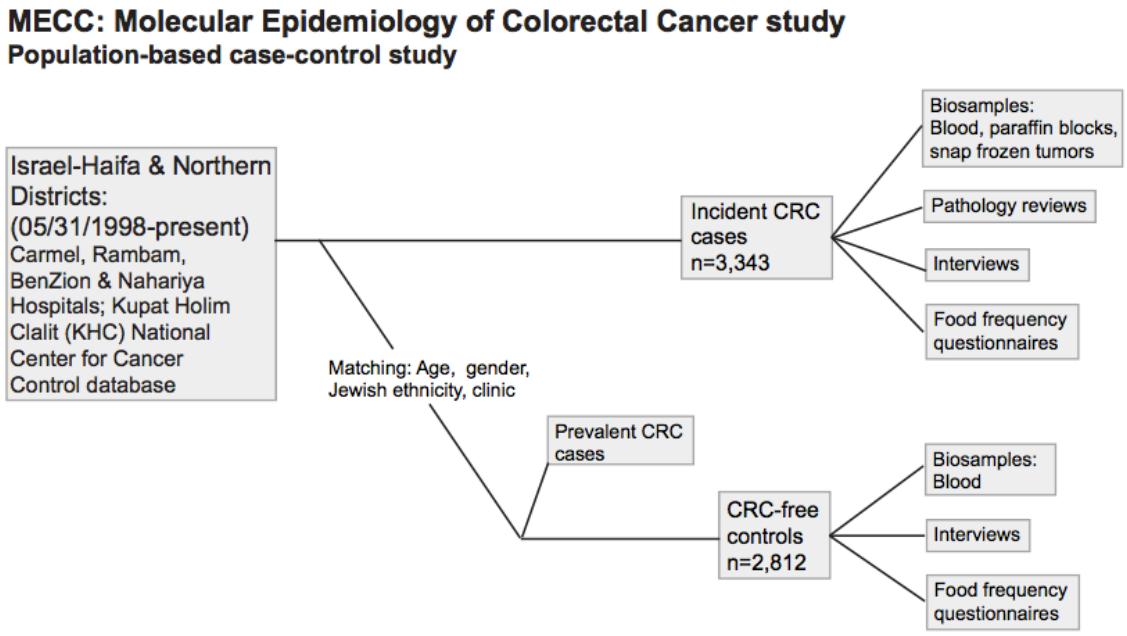


Figure 2.4 Power to detect differences in allele-specific expression means The power for detecting various differences in the mean ASE values for *GPR45* (n=113), *TGFBRAP1* (n=49), and *STK38L* (n=95) are shown.

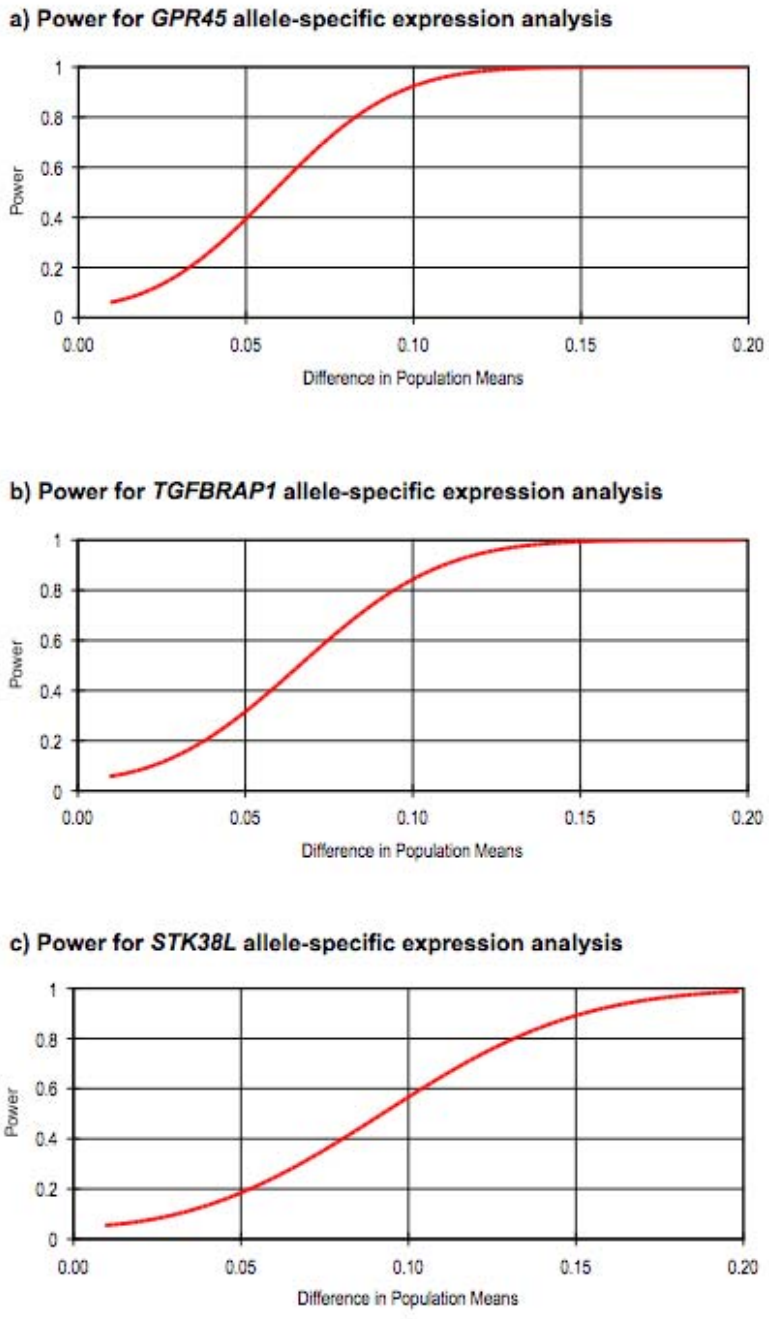


Figure 2.5 MECC genome-wide association study design The design of the MECC GWAS is shown below, conducted in three phases. Phase I was comprised of a whole genome scan and pooled analysis of 511 Ashkenazi Jewish MECC case-control pairs followed by individual-level genotyping to confirm results of the top 5,000 SNPs. Phase II was comprise of individual-level genotyping of 3,500 SNPs for 1,500 MECC case-control pairs. Phase III was comprised of replication of 22 SNPs in an independent set of cases and controls from MECC, Spain, and Germany.

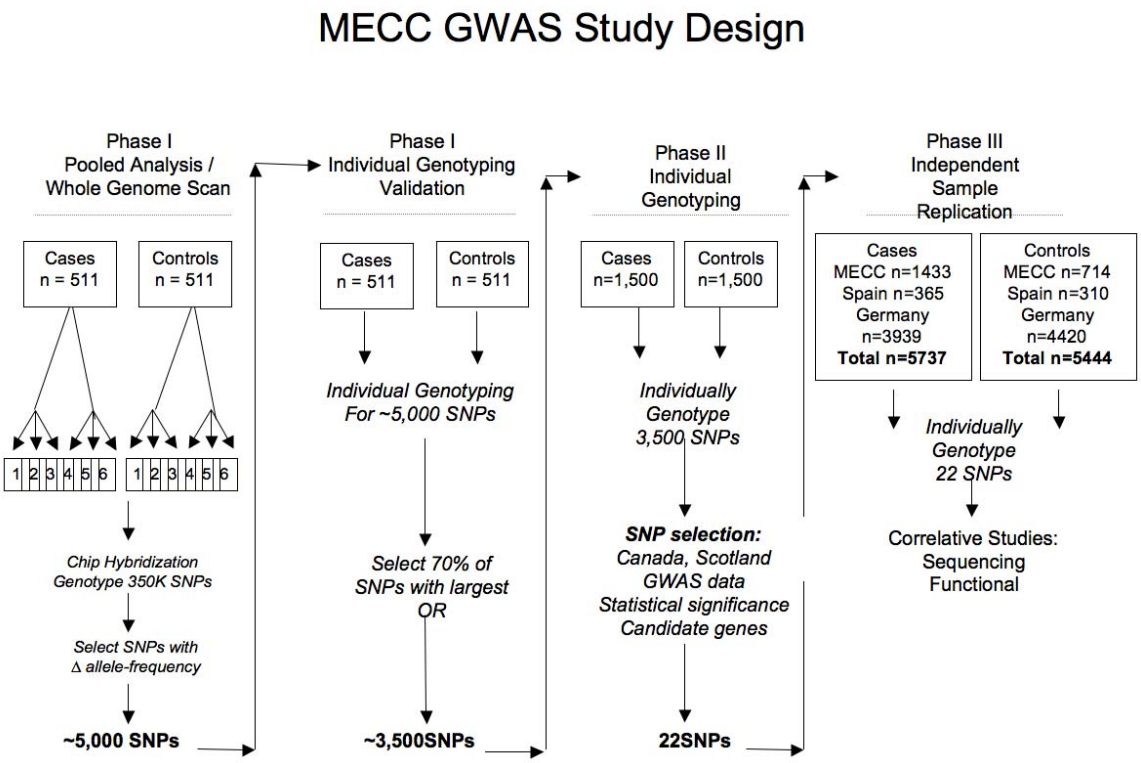


Figure 2.6 Manhattan plots of Phase II and III MECC GWAS Top panel displays the $-\log_{10}$ p-values for 3,500 SNPs genotyped in Phase II of the MECC GWAS by chromosome and position. The bottom panel shows results of the combined phase 2 and Phase 3 analysis of the 22 SNPs selected for phase 3 genotyping are shown. $-\text{Log}_{10}$ p-values are displayed in navy according to their location on chromosomes 1 to 22.

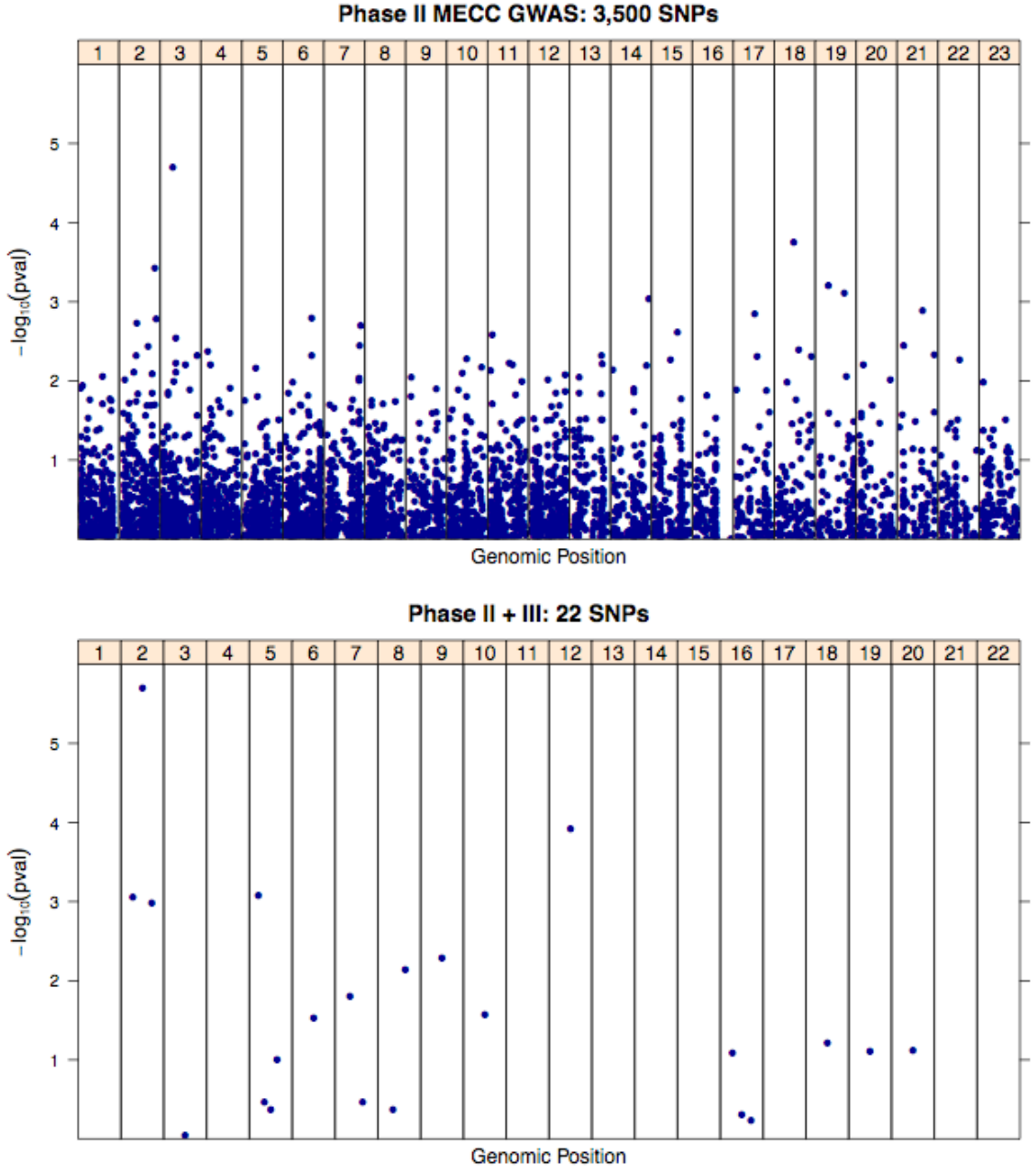


Figure 2.7 Linkage disequilibrium patterns on chromosome 2 at rs10210149 A linkage disequilibrium (LD) heatmap measured by D' is shown for SNPs along 122.3 kb of chromosome 2, measured using data from HapMap for CEPH subjects. Red indicates high LD while yellow indicates low LD. The GWAS SNP rs10210149 is indicated at its position on chromosome 2 as well as the genes *GPR45* and *TGFBRAP1*.

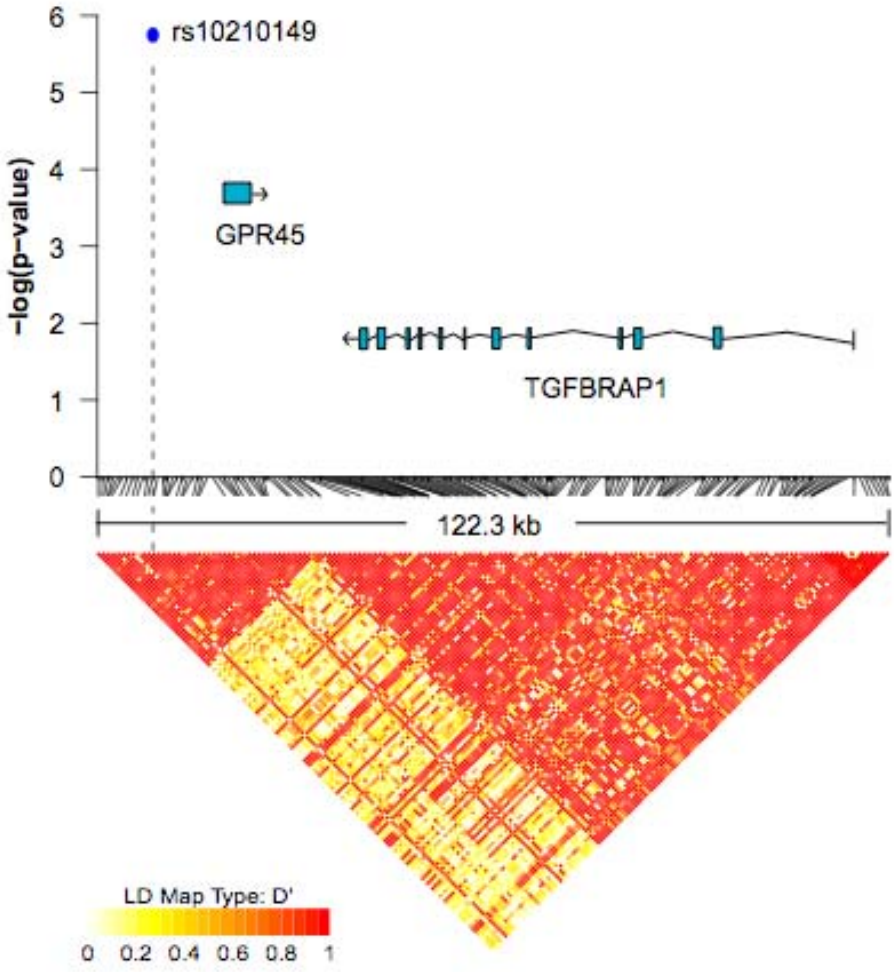


Figure 2.8 Linkage disequilibrium patterns on chromosome 12 at rs16931815 A linkage disequilibrium (LD) heatmap measured by D' is shown for SNPs along 216.9 kb of chromosome 12, measured using data from HapMap for CEPH subjects. Red indicates high LD while yellow indicates low LD. The GWAS SNP rs16931815 is indicated at its position on chromosome 12 as well as the genes *STK38L* and *ARNTL2*.

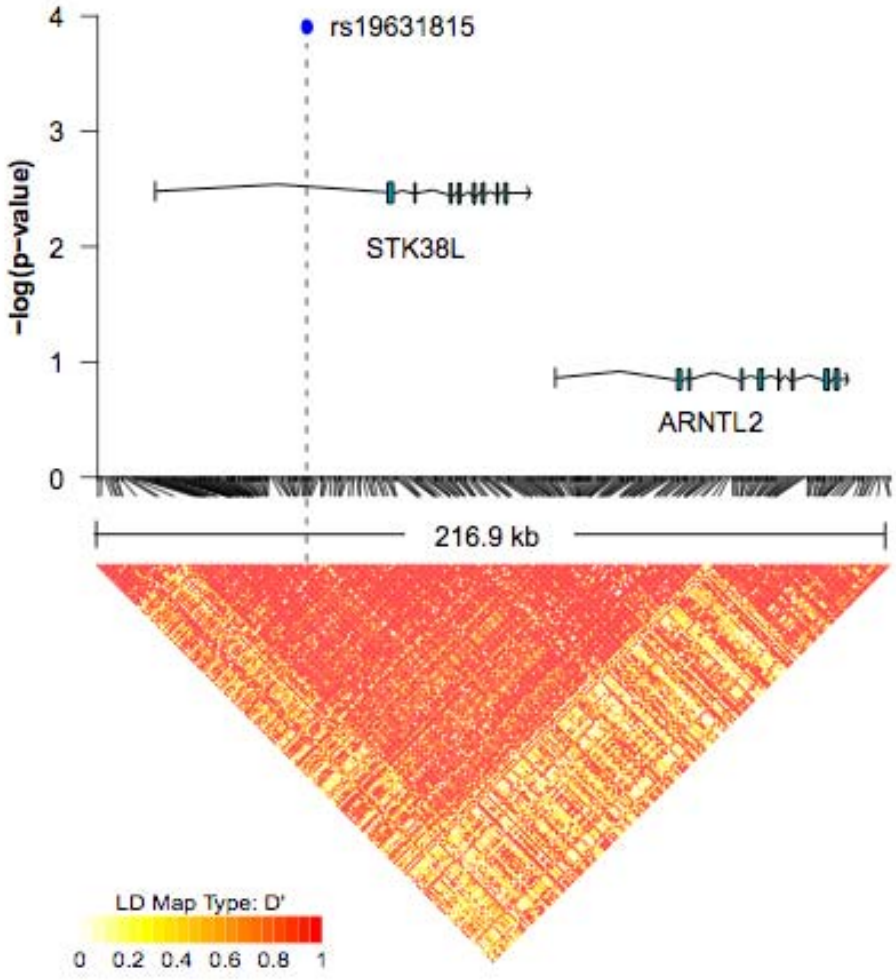


Figure 2.9 Allele-specific expression of *GPR45* Distribution of ASE measured at rs35946826 in exon 1 of *GPR45* by a) rs10210149 genotype and b) case-control status. Mean ASE values are shown within boxes.

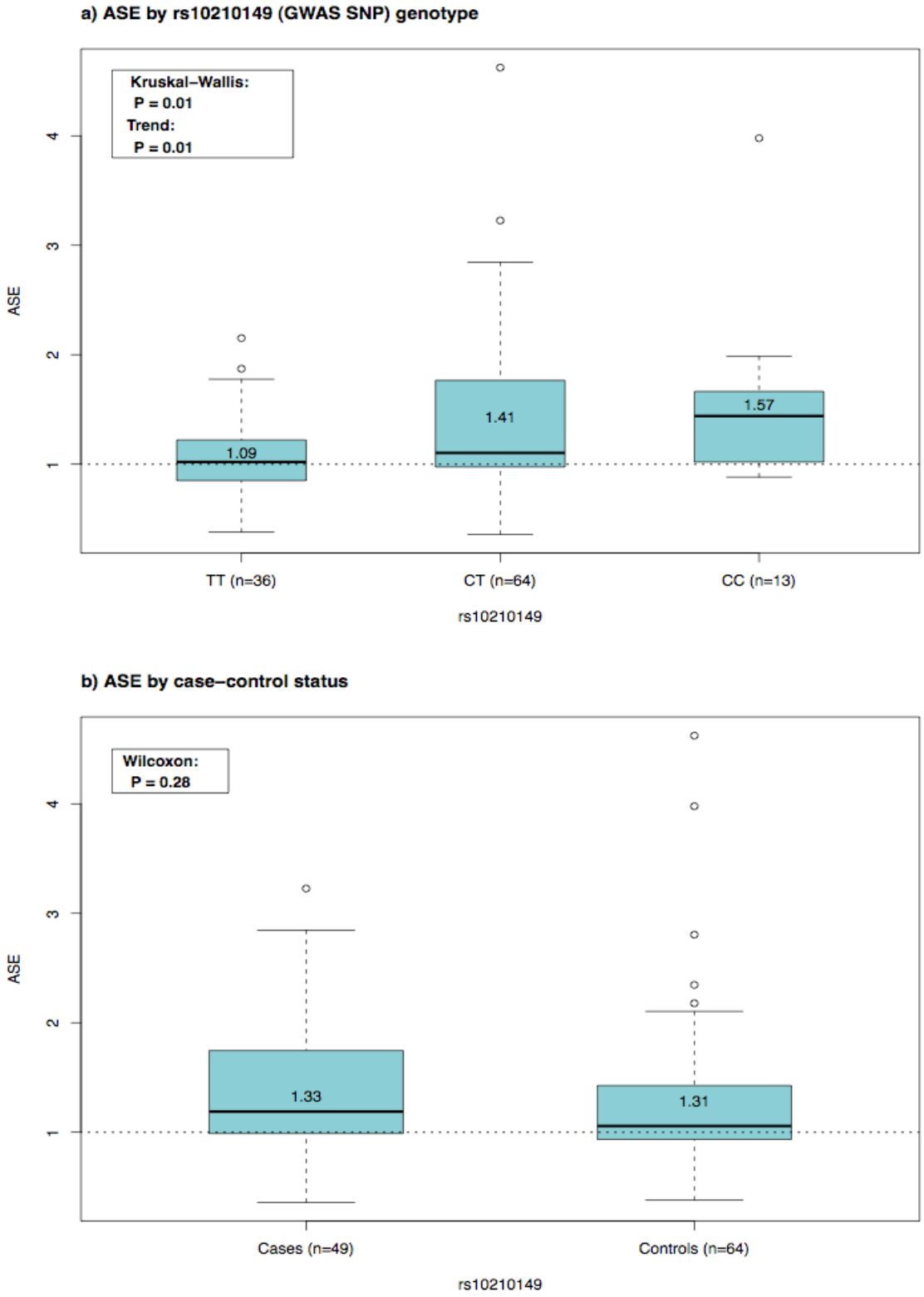


Figure 2.10 Allele-specific expression of *TGFBRAP1* Distribution of ASE measured at rs2241801 in exon 2 of *TGFBRAP1* by a) rs10210149 genotype and b) case-control status. Mean ASE values are shown within boxes.

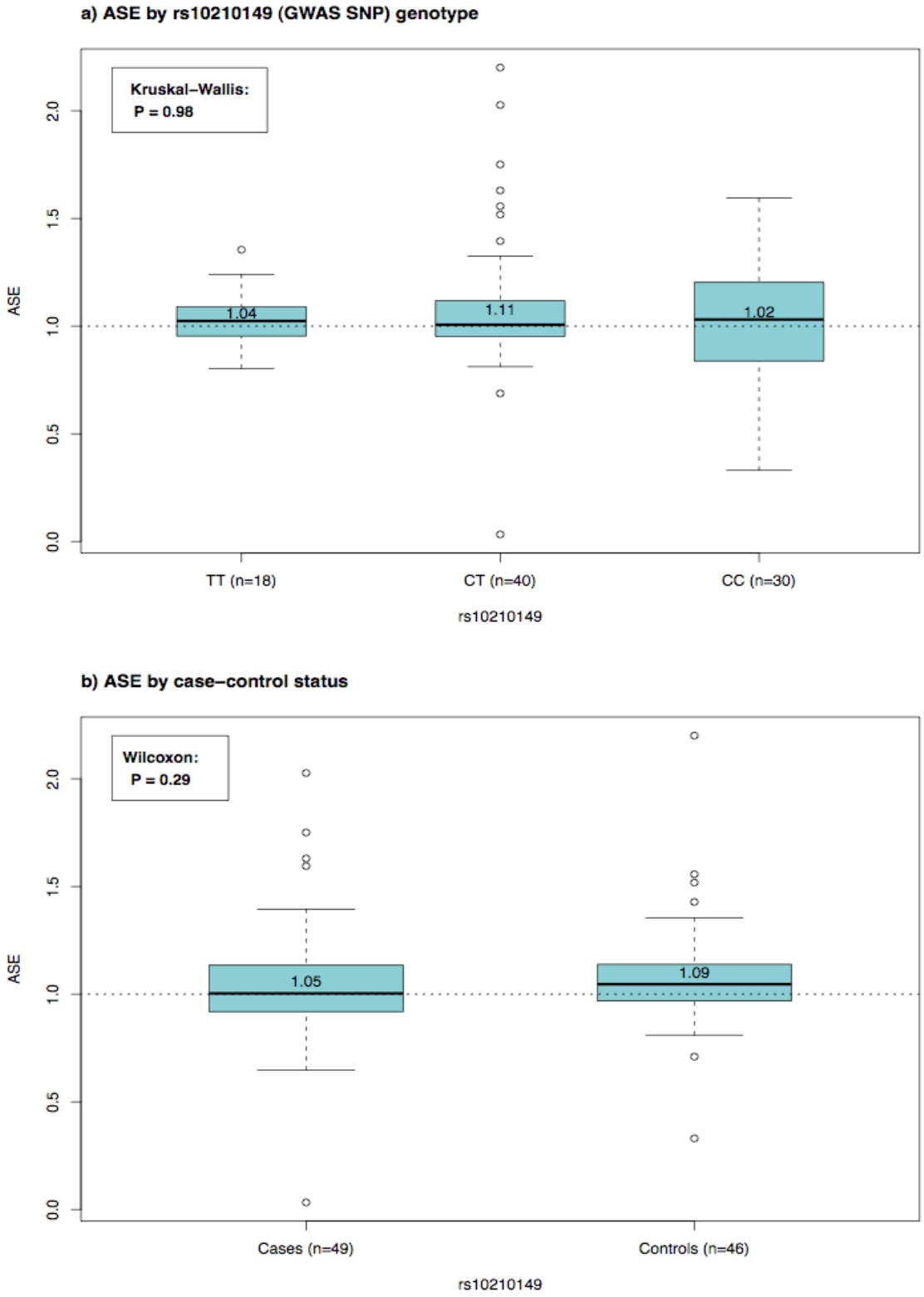
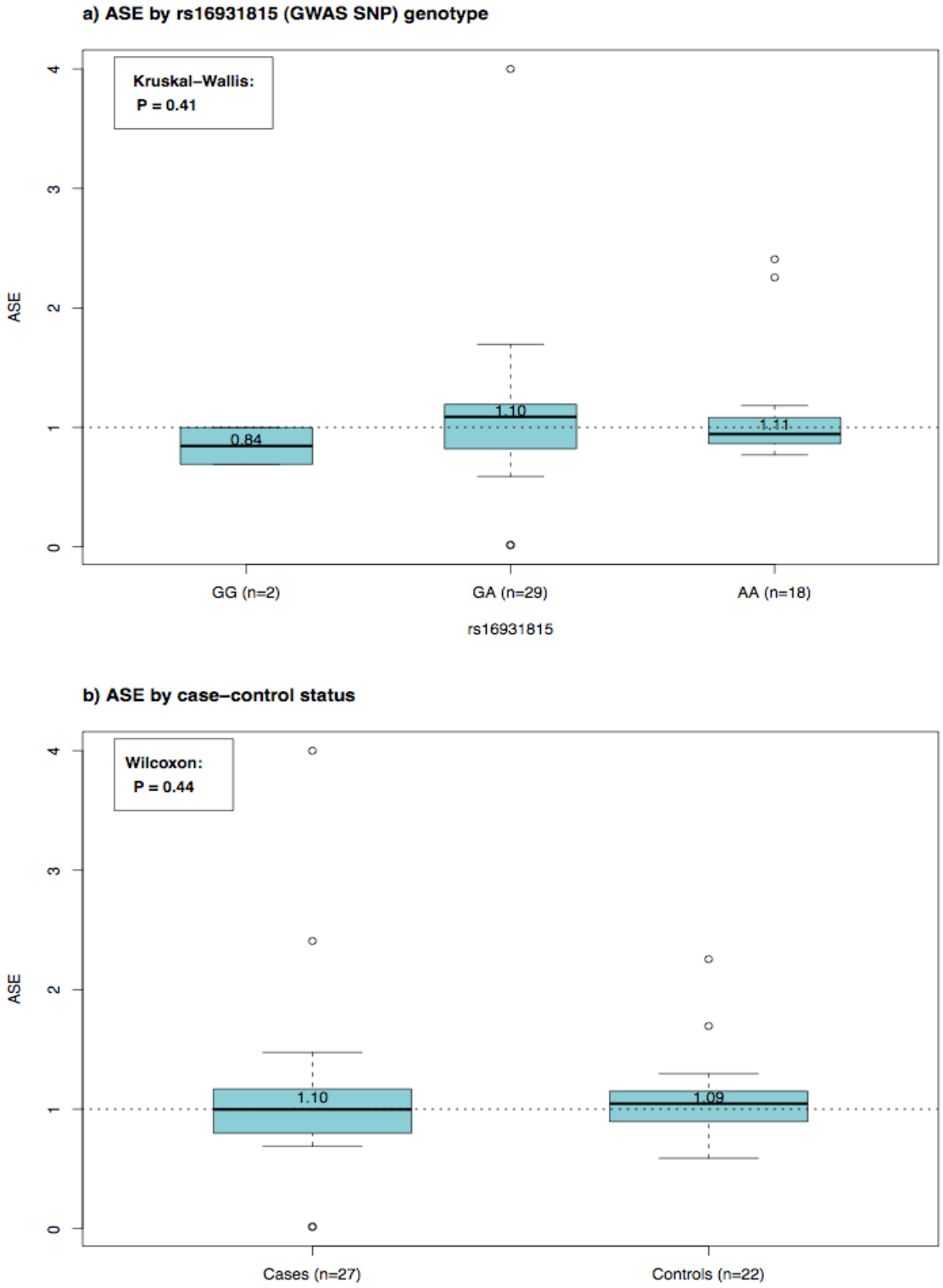


Figure 2.11 Allele-specific expression of *STK38L* Distribution of ASE measured at rs10842902 in the 3' UTR of *STK8L* by a) rs16931815 genotype and b) case-control status. Mean ASE values are shown within boxes.



CHAPTER 3

Common variation in *ISL1* confers genetic susceptibility for human congenital heart disease

3.1 Introduction

Congenital heart disease (CHD) is the most common live birth defect in the United States, affecting 1 in 20 live births (Hoffman and Kaplan, 2002; Hoffman *et al.*, 2004). 1 out of 100 CHD cases requires an intervention, reflecting the wide range in severity and morphology of these types of defects. The causes of many cases of congenital heart disease remain unknown. Those risk factors that have been identified, such as maternal exposures or chromosomal abnormalities, are typically associated with specific subsets of CHD (Jenkins *et al.*, 2007; Pierpont *et al.*, 2007). Perhaps contributing to this lack of identifiable CHD risk factors, congenital heart defects are often investigated as etiologically and morphologically separate diseases. However, examination of the gene pathways that control early human cardiac development may reveal significant insight into the common origins of a broad subset of congenital heart disease.

Currently, 13% of CHD is attributed to chromosomal abnormalities detectable by chromosome analysis (Pierpont *et al.*, 2007), although this estimate varies widely by type of defect. Improvements in the resolution of these technologies are revealing that the proportion of CHDs with chromosomal abnormalities may be even higher than this original estimate. Furthermore, studies of the offspring of affected individuals have

shown a significantly higher proportion of children with CHD than expected, suggesting an important role for genetic susceptibility to CHD (Rose *et al.*, 1985; Whittemore *et al.*, 1982; Whittemore *et al.*, 1994). However, no common genetic variants have been robustly associated with the risk of complex CHD, and only a few rare mutations have been identified in non-syndromic CHD patients. These include mutations in *NKX2.5* and *GATA4*, which have been identified in individuals with atrial septal defects, atrioventricular conduction delay, and ventricular septal defects (Garg *et al.*, 2003; Posch *et al.*, 2008; Schott *et al.*, 1998).

The functions of *NKX2.5* and *GATA4* are well described and are centrally involved in the regulation of a subset of cardiomyocytes known as the primary heart field. During vertebrate cardiac development, the 3-dimensional structure of the heart is formed from the differentiation and interaction of multiple tissue derivatives, or fields (Gruber and Epstein, 2004). The primary and secondary heart fields of the embryonic disc give rise to the intracardiac structures of the heart under the influence of adjacent tissues (Buckingham *et al.*, 2005). The first lineage contributes to the formation of both ventricles, the atrioventricular canal, and both atria. *NKX2.5* and *GATA4* are key myocardial regulatory genes in this process, and the primary heart field is marked by expression of either *TBX5* or the first wave of *NKX2.5* (Wu *et al.*, 2008). The secondary heart field provides an especially important source of cells, contributing to the outflow tract and essentially all heart regions other than the left ventricle (Buckingham *et al.*, 2005). This second population of cells is marked by expression of the *ISL1* gene (Laugwitz *et al.*, 2005; Moretti *et al.*, 2006; Wu *et al.*, 2008).

Considering that rare *NKX2.5*, *GATA4*, and *TBX5* mutations have been reported in

CHD patients, it is plausible that other genes that are critical to cardiomyocyte regulation and differentiation are also involved in CHD etiology. Specifically, *ISL1* is a likely candidate susceptibility gene for human CHD. Consistent with the integral role of *ISL1* in the regulation of the secondary heart field, *Isl1-/-* mouse embryos display distinct cardiac abnormalities: dysmorphic hearts with abnormal looping, ventricular misidentity, hypoplastic outflow tracts, and hypoplastic atrial structures (Ahlgren *et al.*, 1997; Cai *et al.*, 2003). Although mice deficient in *Isl1* harbor defects in cardiac morphogenesis, the role of *ISL1* in human congenital heart disease is unknown. We hypothesized that genetic defects in *ISL1* disrupt early human cardiac development, resulting in congenital defects in secondary heart field-derived structures. We conducted a two-stage case-control study of CHD to test this hypothesis, and in this dissertation we describe the first association between common genetic variation and risk of non-syndromic, complex CHD.

3.2 Subjects and methods

3.2.1 Study design

This study of the candidate gene *ISL1* was performed in three parts. First, I performed Sanger sequencing of *ISL1* in 99 children with CHD from the Children's Hospital of Philadelphia (CHOP). All 6 exons and exon/intron boundaries were sequenced to search for pathogenic mutations and to characterize the variation in this gene.

The second part of this study consisted of a case-control study at the Children's Hospital of Philadelphia, denoted as the stage 1 study. In this study, we estimated the

association between common genetic variation in *ISL1* and the risk of congenital heart disease. Cases and controls were derived exclusively from CHOP and are referred to in this dissertation as United States (US) subjects. Analyses were conducted separately for whites and blacks/African Americans to understand the patterns of risk in these separate groups and as one strategy to adjust for genetic differences between these two groups. Common genetic variation was examined in all subjects at 3 SNPs within *ISL1*, identified from Sanger sequencing. Data were available for 27 SNPs surrounding *ISL1* for only a subset of subjects.

In the third part of this study, we sought to replicate the results from the initial US case-control study. To accomplish this, we conducted a second case-control study denoted as the stage 2 study. This analysis was again conducted separately for whites and blacks/African Americans. Stage 2 white subjects were comprised of additional cases and controls from CHOP, cases and controls from CONCOR in the Netherlands, and CHD cases ascertained at SickKids Hospital in Toronto, Canada. All stage 2 white subjects were newly identified and distinct from stage 1 whites. Common genetic variation was examined at the 3 SNPs within *ISL1* for all subjects. Data were available for the 27 *ISL1*-flanking SNPs for US white subjects only. Stage 2 black/African American subjects were ascertained exclusively at CHOP, the only children's hospital in our consortium with a large percentage of black/African American patients. These black/African American patients were also completely distinct from stage 1 subjects. Only the 3 SNPs within *ISL1* were examined for these subjects. Analyses were conducted separately for stage 2 subjects, followed by a combined analysis of the stage 1 and stage 2 studies.

3.2.2 Subjects

United States cases and controls were recruited from the Children's Hospital of Philadelphia (CHOP) between 12/12/2003 and 08/25/2008 on a protocol approved by the Institutional Review Boards of CHOP and the University of Michigan, and parents provided written informed consent. The proportion of all eligible cases seen at the CHOP Cardiac Center in this time period that participated in this study was 31.6% (613/1939). US cases were children with complex congenital heart disease requiring surgical repair. Guided by lineage-tracing analyses in rodents, cases were defined by diseases representative of secondary heart field defect phenotypes (Black, 2007; Cai *et al.*, 2003; Sun *et al.*, 2007). These include defects of atrial septation, ventricular septation, conus positioning, and great vessel alignment (**Figure 3.1**). US controls were patients without congenital heart disease recruited through the CHOP Health Care Network by CHOP clinicians and nursing staff. The controls were screened by nurse practitioners who evaluated medical records for surgical repair of a cardiac defect. All cases and controls were evaluated by a physician.

Ethnicity for US cases was determined by self-report in stage 1. Self-reported ethnicity was not available for stage 2 cases, so ethnicity was determined by principal components analysis (PCA) in stage 2. Ethnicity for US controls was determined by PCA in both stages, also because self-reported ethnicity was not available. PCA was performed at the Center for Applied Genomics at CHOP for stage 1 controls. A different method was used to determine ethnicity for stage 2 controls. First, the first two principal components were plotted for stage 1 cases of known ethnicity, which demonstrated that the first principal component distinguished between white and black/African American

cases (**Figure 3.2**). This distinction was made using a cutoff of PC1≤0.025 to define ethnicity as white and PC1>0.025 to define ethnicity as black/African American. Similarly, the first principal component sufficiently distinguished between stage 1 white and blacks/African American controls, using a cutoff of PC1≤0.0059 to define ethnicity as white and PC1>0.0059 to define ethnicity as black/African American. The implications of the probable misclassification bias resulting from this classification method are discussed below.

Stage 2 cases and controls from the US, Toronto, and the Netherlands were recruited on institution-specific protocols, and were also approved by the IRBs of CHOP and the University of Michigan. Stage 2 US cases and controls were ascertained as described for stage 1. Dutch and Canadian cases were also children with complex congenital heart disease requiring surgical repair. The distribution of cardiac defects among these cases differed slightly from US cases, though were selected using the same second heart field defect criteria (**Figure 3.1**). Dutch controls were patients without congenital heart disease recruited through UMC Utrecht. All stage 2 subjects were evaluated by a medical doctor, and ethnicity was determined by self-report.

3.2.3 Genotyping

The 6 exons and exon/intron boundaries of *ISL1* were sequenced in the first 99 cases using bidirectional Sanger sequencing, accomplished at the University of Michigan sequencing core. The primers and cycling conditions are given in **Table 3.1**. Variants were identified using Mutation Surveyor software (State College, PA). GeneSplicer (Pertea *et al.*, 2001) and NetGene2 (Brunak *et al.*, 1991; Hebsgaard *et al.*, 1996) were

used to bioinformatically predict splice site variants.

Stage 1 and stage 2 genotypes were requested for 27 *ISL1*-flanking SNPs from the Center for Applied Genomics at CHOP that had been performed using the Illumina HumanHap 550 SNP array. Genotypes for only these 27 SNPs were obtained from the Center for Applied Genomics. At the time of the study, no additional genotypes on this platform were obtained or analyzed. Data was available for these 27 SNPs for white US cases and controls in both stage 1 and stage 2.

Genotypes for the 3 SNPs within *ISL1* (rs3762977, IVS1+17C>T, rs1017) were determined using one of two methods: 1) bidirectional sequencing of *ISL1* exons 1 and 6 or 2) genotype imputation. Stage 1 genotyping of these 3 SNPs was performed using bidirectional Sanger sequencing as described above. Stage 2 Canadian and Dutch cases were also genotyped using bidirectional sequencing. Imputation was performed for the 3 SNPs within *ISL1* for stage 2 US cases and all controls using 97 SNPs surrounding these 3 SNPs. Haplotypes were reconstructed for all 100 SNPs in 484 controls using FastPHASE14. These phased, reconstructed haplotypes were then used as the reference haplotypes for genotype imputation using the MACH program (Li Y, 2006; Scheet and Stephens, 2006). Each genotype at each SNP was associated with a QC score, interpreted as the posterior probability that the imputed genotype represents the true genotype.

Genotyping accuracy by sequencing was assessed with repeat sequencing of a subset of genotypes. Two measures were used to assess imputation error for the three *ISL1* SNPs. The first measure, εj, captures genotyping error, discrepancies with the reference panel, and recurrent mutation (Li Y, 2006; Scheet and Stephens, 2006). Slightly lower data quality is observed for larger estimates of εj. Values of εj were small

for each of the three SNPs: rs3762977 εj = 0.0229, IVS1+17C>T εj = 0.0007, rs1017 εj = 0.0434. The second measure of imputation error was the agreement between genotypes determined by sequencing and genotypes determined by imputation for a subset of stage 1 cases and controls for whom both genotypes were available. Agreement was measured using the Kappa statistic in SAS (version 9.1) at various QC cutoffs. No QC values for any of the 3 imputed SNPs were less than 0.5. Inclusion of all imputed genotypes (regardless of QC value) resulted in Kappa statistics of at least 0.889 for each of the 3 SNPs, confirming that the genotype imputation method is robust. Agreement was also measured among subjects carrying at least one minor allele for each of the three SNPs. For rs3762977 the Kappa statistic was 0.614 and for rs1017 the Kappa statistic was 0.918. For IVS1+17C>T, 75% of all cases were called heterozygotes by both Sanger sequencing and imputation. The implications of genotype misclassification are discussed below. All genotype frequencies were assessed for departure from Hardy-Weinberg equilibrium in controls.

3.2.4 Statistical methods

Single SNP analyses were conducted using unconditional logistic regression to calculate odds ratios as implemented in SAS (version 9.1). Haplotypes were estimated and tested for association with CHD using the haplo.stats package in R (http://cran.r-project.org). Significance testing was adjusted for multiple comparisons with Bonferroni correction, and 95% confidence intervals were calculated from the parameters estimated in logistic regression.

3.3 Results

3.3.1 Characterization of ISL1 variation

We hypothesized that genetic mutations or variants in the gene *ISL1* disrupt early cardiac development, resulting in a wide variety of congenital defects in the heart structures derived from the secondary heart field. To address this hypothesis, we first sequenced *ISL1* in 99 cases of complex congenital heart disease. This was done in an attempt to identify either mutations or functional variants within the gene and to characterize the overall variability at this locus. 15 polymorphisms were identified (**Table 3.2)**. Most variants were rare (8/15) and 6 had been previously reported in dbSNP. Variation was observed only in exons 1, 4, and 6 and in introns 1 and 5 in the initial 99 cases; thus, further genotyping was restricted to these regions. All exonic polymorphisms were either synonymous mutations or occurred in noncoding regions that were not predicted to be splice-site abnormalities. We concluded that none of these variants were of clear functional significance; therefore, we proceeded by examining the association between common variation in *ISL1* and risk of CHD in a case-control study.

3.3.2 Stage 1: US case-control study in white subjects

We conducted a two-stage candidate gene study to test the hypothesis that germline common genetic variants in *ISL1* confer susceptibility to non-syndromic human CHD. The stage 1 case-control study was comprised of 300 CHD cases (white n=160, black/African American n=70, other/unknown=70) and 2,201 CHD-free controls (white n=2091, black/African American n=110). Cases were children diagnosed with complex, non-syndromic CHD, all of which required operative repair. Guided by lineage-tracing

analyses in rodents, cases were defined by diseases representative of secondary heart field defect phenotypes (Black, 2007; Cai *et al.*, 2003; Sun *et al.*, 2007). Defects of the second heart field are potentially pathogenic in anatomic defects of both the right and left sides of the normal heart due the contribution of secondary heart field derivatives in both inflow and outflow tracts. These include defects of atrial septation, ventricular septation, conus positioning, and great vessel alignment.

We analyzed 30 SNPs spanning a 237 kb region around *ISL1* on chromosome 5q11.1, selected to capture variation in this region based upon linkage disequilibrium patterns in subjects of European ancestry (http://www.hapmap.org). No genome-wide data were available for this hypothesis-driven, candidate gene study. Eight individual SNPs (rs6867206, rs4865656, rs6869844, rs2115322, rs6449600, IVS1+17C>T, rs1017, rs6449612) were significantly associated with risk of CHD at the α=0.05 level (**Figure 3.3**) located within a single LD block. Indeed, HapMap data demonstrate D' =1 between three of these SNPs (rs6869844, rs6449600, rs6449612) and each of the four Hapmap published SNPs within *ISL1* (rs3792733, rs2288468, rs3811911, rs991216). The moderate magnitudes of association seen at these SNPs (OR = 1.32 – 2.30) were consistent with those expected under the common disease – common variant hypothesis (Reich and Lander, 2001; Wang *et al.*, 2005). Furthermore, the closest gene to *ISL1* is located in a different LD block more than 540 kb upstream (*PARP8*), reducing the likelihood that these SNPs are capturing an association between a gene other than *ISL1* and risk of CHD. Of the six *ISL1*-flanking SNPs, rs6869844 remained statistically significant after adjustment for multiple testing (P = 0.039 with Bonferroni correction for 30 SNPs). Located 15.7 kb 5' of *ISL1*, rs6869844 was associated with a 50% increase in

risk for each additional T allele in a log-additive model (Odds ratio (OR) = 1.51, 95% confidence interval (CI) = 1.18-1.95).

Three SNPs analyzed in stage 1 (rs3762977, IVS1+17C>T, rs1017) were located within the *ISL1* gene in the 5'UTR, intron 1, and the 3' UTR, respectively (**Figure 3.4**). To diminish the potential for population stratification in this sample, we first restricted our stage 1 analyses to white cases with non-syndromic CHD and white controls (n=100 cases, 576 controls) with genotype data available for these SNPs. IVS1+17C>T was associated with a more than two-fold increase in risk among whites with the C/T genotype (OR = 2.30, 95% CI 1.12 – 4.70, $P = 0.023$) (**Table 3.3a**). Rs1017 was highly significant in a log additive model, with an 81% increase in risk associated with each additional copy of the T allele (OR = 1.81, 95% CI 1.29 – 2.54, $P = 0.0007$). Dominant and recessive models for rs1017 were also highly significant. Children with the A/T or T/T genotype had a 2.28-fold increase in risk compared to children with the A/A genotype (OR = 2.28, 95% CI 1.35 – 3.87, $P = 0.002$). Similarly, children with the T/T genotype had a 2.21-fold increase in risk compared to children with the A/A or A/T genotype (OR = 2.11, 95% CI 1.17 – 3.80, $P = 0.013$).

We then delineated the patterns of risk in these subjects by using the expectation maximization (EM) method to estimate haplotypes and risk of CHD from the 6 *ISL1*-flanking SNPs (rs6867206, rs4865656, rs6869844, rs2115322, rs6449600, rs6449612) and the 3 SNPs within *ISL1* (rs3762977, IVS1+17C>T, rs1017). The three SNPs within *ISL1* most effectively captured risk of CHD. In stage 1 whites, an additive model fit the data well (global haplotype association $P = 0.0008$). Two haplotypes, A-C-T and A-T-T (rs3762977- IVS1+17C>T -rs1017), were strongly associated with CHD risk (**Table**

3.3b). A child's risk of CHD was 2.01 times greater with each copy of the A-C-T haplotype compared to the A-C-A haplotype (95%CI 1.35 – 2.99, $P = 0.0006$) and 3.30 times greater with each copy of the A-T-T haplotype (95% CI 1.52 – 7.18, $P = 0.0026$).

3.3.3 Stage 2: US, Canadian, and Dutch case-control study in white subjects

To understand the role of *ISL1* variation and risk of CHD in other populations, we studied *ISL1* variation in a second, independent analysis of samples from the US, Canada, and the Netherlands. Stage 2 cases and controls were completely distinct from those in initial stage 1. The stage 2 white subjects consisted of 995 cases (US n=265, Canada n=94, Netherlands n=636) and 2089 controls (North America n=1446, Netherlands n=643). For the purpose of this analysis, cases from the US and Canada were combined and compared to US controls, indicated as North American cases and controls. The allele frequencies at each of the three *ISL1* SNPs were comparable between US and Canadian cases (**Table 3.4**). This indicates that estimates resulting from the North American analysis are not confounded by differences in allele frequencies between the US and Canadian cases.

Data were available for the 27 *ISL1*-flanking SNPs for only US whites in stage 2. Single SNP analyses in the stage 2 US white population confirmed the association at 10 of these SNPs within and around *ISL1* (**Figure 3.5**). Further investigating this relationship in stage 2, we next examined the 3 SNPs within *ISL1* in all North American whites. Rs1017 was significantly associated with risk of CHD in a log-additive model, where each copy of the T allele at rs1017 increased a child's risk of CHD by 22% (OR = 1.22, 95% CI 1.05 – 1.44, $P = 0.022$) (**Table 3.5a**). Also consistent with the stage 1

analysis, the A-C-T haplotype was significantly associated with risk among North American whites (**Table 3.5b)**. Each copy of the A-C-T haplotype conferred a 33% increase in risk (OR = 1.33, 95% CI 1.08 – 1.62, *P* = 0.0065) compared to the A-C-A haplotype.

We next estimated the association between the three *ISL1* SNPs and risk of CHD using Dutch cases and controls from stage 2. In single SNP analyses, none of the three SNPs were associated with risk of congenital heart disease (**Table 3.6a**). Similarly, there was no significant association between *ISL1* variation and CHD risk in a haplotype analysis. However, a haplotype analysis that combined all stage 2 whites demonstrated that the A-C-T haplotype is significantly associated with risk among whites (Global *P* = 0.00003) (**Table 3.7**). Each copy of the A-C-T haplotype conferred an 18% increase in risk (OR = 1.18, 95% CI 1.00 – 1.39, *P* = 0.0485) compared to the A-C-A haplotype. We next performed a combined analysis of both stage 1 and stage 2 for the A-C-T haplotype in whites, which was highly significant (**Table 3.8, Figure 3.6**). Each copy of the A-C-T haplotype was associated with a 27% increase in risk of CHD (95% CI 1.09 – 1.48, *P* = 0.0018).

The precise distribution of CHD diagnoses was different between stage 1 and stage 2 populations (**Figure 3.1**). However, hypoplastic left heart syndrome (HLHS) and D-transposition of the great arteries (D-TGA) were the most common diagnoses in both stages, with HLHS accounting for 19.4.0% in Stage 1 and 7.0% in Stage 2 and D-TGA accounting for 14.0% in Stage 1 and 28.4% in Stage 2. To ensure that the stage 2 replication of our original findings was not influenced by the differences among case populations, we performed a subset analysis to include only the most frequent diagnoses

in both stages, HLHS and D-TGA. Associations with rs1017 and the A-C-T haplotype in stage 1, stage 2, and combined analyses were consistent with analyses utilizing cases of all secondary heart field defects in both magnitude and significance of association (**Table 3.9**). This indicates that risk of CHD is consistently associated with common genetic variation in *ISL1* in whites whether considering all secondary heart field defects combined or subsets of the two most common diagnoses.

3.3.4 Stage 1: US case-control study in black/African American subjects

To understand the role of *ISL1* variation in an ethnically distinct sample, we investigated these 3 SNPs in the stage 1 black/African American cases and controls using the exact same phenotypic definitions for cases with non-syndromic CHD (n=54 cases, 110 controls). Compared to whites, analysis at these three loci demonstrated a different pattern of association between *ISL1* and risk of CHD (**Table 3.10a**). While no association was observed at rs3762977 in whites, black/African American children were at a more than 2-fold increase in risk for each additional copy of the G allele at this locus (OR = 2.21, 95%CI 1.15-4.24; *P*=0.017). Variation at IVS1+17C>T was extremely rare among blacks/African Americans with only 1 heterozygous control and 0 heterozygous cases, and no association between rs1017 and risk of CHD was observed (OR = 1.08, 95%CI 0.66-1.76; *P*=0.756). However, as with the single SNP analyses, haplotype analysis showed that the black/African American sample demonstrated a distinct pattern of risk at the *ISL1* locus (**Figure 3.6, Table 3.10b**). The A-C-T haplotype was not associated with increased risk of CHD among blacks, and the A-T-T haplotype was not identified in any cases or controls of black/African American ancestry. In contrast, the G-

C-T haplotype was associated with a 2-fold increase in risk of CHD (OR = 1.99, 95%CI 1.02-3.87; *P*=0.043).

3.3.5 Stage 2: US case-control study in black/African American subjects

To determine whether our findings in stage 1 blacks/African Americans would be consistent in a new set of samples, we analyzed a distinct set of 49 US black/African American cases and 1,845 US black/African American controls (**Table 3.11a**). In this stage 2 sample, the relative risk for rs3762977 was consistent with that seen in stage 1 blacks/African Americans (OR = 1.20, 95% CI 0.74 – 1.95, *P* = 0.457), although not statistically significant. Similarly, the G-C-T haplotype did not reach statistical significance among blacks/African Americans in stage 2 (**Table 3.11b**), but the relative risk for this haplotype was consistent with that seen in stage 1 (OR = 1.28, 95% CI 0.75 – 2.19, *P* = 0.359). The G-C-T haplotype was significantly associated with risk of CHD in a summary analysis of stage 1 and stage 2 blacks/African Americans, where each copy of this haplotype conferred a 57% increase in risk (95% CI 1.07 – 2.30, *P* = 0.0216) (**Table 3.12, Figure 3.6**). These data provide evidence that genetic variation in *ISL1* is associated with risk of CHD in blacks/African Americans, and this risk is characterized by a pattern of variation distinct from that in whites.

3.4 Discussion

Our results demonstrate that two different *ISL1* haplotypes contribute to risk of CHD in white and black/African American samples. These data provide strong evidence that congenital heart disease is consistent with the common disease – common variant

hypothesis in two ethnically distinct samples. Further work is necessary to determine whether these two haplotypes capture ancestrally distinct causative mutations or are in linkage disequilibrium with a single disease-causing mutation. Our observations of different risk haplotypes in whites and black/African Americans is intriguing and suggests that different risk alleles are present in the *ISL1* locus within these groups. This provides an opportunity for identifying causal variants through subsequent studies with admixture mapping or deep sequencing within these two patient samples.

One limitation of this study is that there is likely to be misclassification of ethnicity among the stage 2 US subjects for whom ethnicity was determined by principal components analysis. In stage 2, we expect misclassification to be non-differential since ethnicity was determined using SNPs that should be independent of case-control status, although we are unable to directly measure this. This type of measurement error due to population stratification could be important if the degree of misclassification was very large and differed among cases and controls. However, we do not anticipate this to meaningfully affect our results since misclassification of ethnicity is expected to be independent of case-control status and *ISL1* genotype status. Further, if misclassification of ethnicity is in fact non-differential, then we would expect this type of measurement error to bias our results towards the null.

To determine whether using principal components analysis to classify ethnicity produces different results compared to another method of classifying ethnicity, we employed the ANCESTRYMAP (Patterson *et al.*, 2004) program as an alternative method. This program uses genotype information from two ancestral populations to estimate admixture in a test population. We used 26 of 136 ancestral informative markers

(AIMs) on chromosome 5 for which genotype information for stage 2 subjects was available (Smith *et al.*, 2004). None of these 26 SNPs were in linkage disequilibrium with the *ISL1* locus based upon LD patterns in subjects of European ancestry (http://www.hapmap.org). The 26 AIM genotypes for the two ancestral populations, the Centre d'Etude du Polymorphisme Humain (European) and Yoruban (African) HapMap samples, were downloaded from http://www.hapmap.org. We ran the ANCESTRYMAP program using default parameters to obtain estimates of the percent European ancestry for all US stage 2 subjects of unknown ancestry as well as a subset of US stage 1 subjects of known ancestry (**Figure 3.7**). A bimodal distribution of the percent European ancestry was observed among all subjects, which was highly correlated with self-reported ethnicity among stage 1 subjects. Subjects with greater than 65% European ancestry were defined as white, and subjects with less then 65% European ancestry were defined as black/African American. Single SNP and haplotype analyses were performed using ANCESTRYMAP-defined ethnicity, and results were qualitatively similar to those described above. This suggests that using PCA to define ethnicity does not produce results that are substantially different from another method of classifying ethnicity.

Another limitation of this study is that there is likely to be genotype misclassification in stage 2 US subject and Dutch controls due to imputation error. First considering stage 2 US subjects, we would expect genotype misclassification to be non-differential with respect to case-control status since SNPs used for imputation were not linked to *ISL1* variation. We would not anticipate this to substantially bias the effect estimates for the three *ISL1* SNPs, although this would depend on the degree of misclassification for each SNP. Estimates of genotyping error, εj, were small for each of

the three SNPs: rs3762977 εj = 0.0229, IVS1+17C>T εj = 0.0007, rs1017 εj = 0.0434. It is possible that stage 2 US analyses of rs1017 and rs3762977 were influenced by genotyping error; however, the degree of misclassification is estimated to be minor and should not significantly affect the estimates of association in analyses of stage 2 US subjects.

Genotype misclassification among stage 2 Dutch samples would be differential with respect to case-control status, since cases were genotyped using Sanger sequencing while genotypes for controls were imputed. As mentioned above, estimates of genotyping error were small. Whether this would bias estimates of association for the three *ISL1* SNPs towards or away from the null depends on whether controls were more or less likely to be classified as having the risk allele for rs1017. This could partially explain the differences in the direction and magnitude of effect estimates for rs1017 between the stage 2 North American and Dutch analyses.

Finally, we had limited power to detect an association between rs3762977 and risk of CHD in stage 2 blacks/African Americans for a two-sided hypothesis test at α=0.05. In this population of 49 cases and 1,845 controls, if we assume that the minor allele frequency for rs3762977 is equal to 18.4%, then the minimum detectable odds ratio with 80% power is 2.42. Furthermore, we only have 25% power to detect an odds ratio of 1.5 and 60% power to detect an odds ratio of 2.0. In other words, if the true association between *ISL1* variation and risk of CHD in blacks/African Americans is modest (i.e. OR < 2), then we had insufficient power to detect this association in stage 2. This is consistent with the analysis of stage 2 blacks/African Americans, and further argues that a combined analysis of stage 1 and stage 2 subjects is appropriate.

The biologic rationale is compelling: ISL1 is a transcription factor that marks cardiac progenitor cells and controls secondary heart field differentiation, and new evidence suggests that purified populations of ISL1+ progenitor cells are capable of self-renewal and expansion into cardiomyocytes, smooth muscle, and endothelial lineages (Bu *et al.*, 2009). In addition to providing new insight into the variety of congenital heart disease phenotypes that can be produced from second heart field defects in humans, our observations also may provide the basis for a more integrated understanding of the molecular basis of human congenital heart disease.

Table 3.1 PCR primers & conditions for *ISL1* sequencing

Exon	Primer Sequence	T_a°*
1	F: 5' GAG CAG CGC CAC AGG AGG C 3' R: 5' CTT GGC ACC TCA GCC TGT GC 3'	62
2	F: 5' GTA GGA AGT AAA CGG TTA GTC 3' R: 5' CTT GTA TGA CTA CAC TGA GGC 3'	56
3	F: 5' AGT GCC GGC CTG AAG TGA C 3' R: 5' ACA GGC TGG CTT AAC CTG G 3'	62
4	F: 5' AAG CGA GCC TCC AGC CCA G 3' R: 5' GTG CGA TCC TGC GTA CCA G 3'	62
5	F: 5' AAC ATG TTG GGA TTG GTT GGG 3' R: 5' TTC CAT CTG GGA GCT GAC AC 3'	56
6	F: 5' ATG AAT ACT ATT CCA GTG TCC 3' R: 5' GTT TGG CAA GGC AAT GAC C 3' F: 5' TCT AGT CCA TCC TAA TCT G 3' R: 5’ AAA GTG GCA AGT CTT CCG AC 3’	56 56

***** Cycling conditions for all primer sets: Initial denaturation at 95°C for 10min, 30 cycles of 95°C for 30sec, T_a°C for 30sec, 72°C for 1min, final extension at 72°C for 10min.

Table 3.2 *ISL1* variation identified by Sanger sequencing

Location	Polymorphism	Controls			Cases		
Exon 1	rs3762977	AA	AG	GG	AA	AG	GG
		329	102	6	136	39	4
Exon 1	EX1+67G>C	GG	GC	CC	GG	GC	CC
		432	4	0	176	1	0
Exon 1	EX1+192C>G	CC	CG	GG	CC	CG	GG
		438	0	0	179	1	0
Exon 1	rs36216897	AA	AG	GG	AA	AG	GG
		418	15	0	175	3	0
Exon 1	EX1-269G>A	GG	GA	AA	GG	GA	AA
		424	9	0	178	2	0
Exon 1	EX1-215T>G	TT	TG	GG	TT	TG	GG
		432	0	0	178	1	0
Exon 1	rs3917084	AA	AG	GG	AA	AG	GG
		404	22	0	172	6	0
Intron 1	IVS+17C>T	CC	CT	TT	CC	CT	TT
		402	30	0	163	15	0
Exon 4	rs2303751	AA	AG	GG	AA	AG	GG
		NA	NA	NA	49	21	11
Exon 4	EX4+89C>T	CC	CT	TT	CC	CT	TT
		NA	NA	NA	87	1	0
Intron 5	IVS5-105T>A	TT	TA	AA	TT	TA	AA
		299	0	0	111	2	0
Exon 6	EX6+96A>T	AA	AT	TT	AA	AT	TT
		298	0	0	156	1	0
Exon 6	EX6+483T>C	TT	TC	CC	TT	TC	CC
		427	0	0	185	0	1
Exon 6	rs41268421	GG	GT	TT	GG	GT	TT
		383	34	1	170	9	1
Exon 6	rs1017	AA	AT	TT	AA	AT	TT
		182	192	51	68	82	35

Table 3.3 *ISL1* and risk of congenital heart disease in stage 1 US whites

a) Single SNP associations

Genotypes	Controls [n (%)]	Cases [n (%)]	OR [95% CI]	P value
Stage 1				
rs3762977				
A/A	329 (75.3)	65 (79.3)	1.00	
A/G	102 (23.3)	15 (18.3)	0.74 (0.41 / 1.36)	0.338
G/G	6 (1.4)	2 (2.4)	1.68 (0.33 / 8.55)	0.527
		log-additive:	0.87 (0.52 / 1.47)	0.607
IVS1+17C>T				
C/C	402 (93.1)	70 (85.4)	1.00	
C/T	30 (6.9)	12 (14.6)	2.30 (1.12 / 4.70)	0.023
rs1017				
A/A	182 (42.8)	21 (25.3)	1.00	
A/T	192 (45.2)	43 (51.8)	1.94 (1.11 / 3.40)	0.020
T/T	51 (12.0)	19 (22.9)	3.23 (1.61 / 6.46)	0.0009
		log-additive:	1.81 (1.29 / 2.54)	0.0007

b) Haplotype associations

Haplotypes	rs3762977	IVS1+17C>T	rs1017	Frequency (%)	OR [95% CI]	*P* value
1	A	C	A	0.622	1.00	
2	A	C	T	0.208	2.01 (1.35 / 2.99)	0.0006
3	G	C	T	0.126	1.12 (0.64 / 1.95)	0.700
4	A	T	T	0.038	3.30 (1.52 / 7.18)	0.0026
Global haplotype association						0.0008

Rare estimated haplotypes (cumulative frequency = 0.0053) not shown.

Table 3.4 Minor allele frequencies of 3 ISL1 SNPs in stage 2 US and Canadian cases

SNP	US	Canada	P value*
rs3762977	0.128	0.109	0.756
IVS1+17C>T	0.036	0.054	0.270
rs1017	0.415	0.372	0.556

* calculated by Fisher's exact test

Table 3.5 *ISL1* and risk of congenital heart disease in stage 2 North American whites (US + Canada)

a) Single SNP associations

Genotypes	Controls [n (%)]	Cases [n (%)]	OR [95% CI]	P value
rs3762977				
A/A	1128 (78.1)	281 (77.4)	1.00	
A/G	289 (20.0)	75 (20.7)	1.04 (0.78 / 1.38)	0.777
G/G	28 (1.9)	7 (1.9)	1.00 (0.43 / 2.32)	0.993
		log-additive:	1.03 (0.81 / 1.31)	0.815
IVS1+17C>T				
C/C	1334 (92.3)	334 (92.0)	1.00	
C/T	111 (7.0)	29 (8.0)	1.04 (0.68 / 1.60)	0.843
rs1017				
A/A	591 (40.9)	129 (35.3)	1.00	
A/T	672 (46.5)	177 (48.5)	1.21 (0.94 / 1.55)	0.144
T/T	182 (12.6)	59 (16.2)	1.49 (1.05 / 2.11)	0.027
		log-additive:	1.22 (1.03 / 1.44)	0.022

b) Haplotype associations

Haplotypes	rs3762977	IVS1+17C>T	rs1017	Frequency (%)	OR [95% CI]	*P* value
1	A	C	A	0.630	1.00	
2	A	C	T	0.211	1.33 (1.08 / 1.62)	0.0065
3	G	C	T	0.119	1.09 (0.85 / 1.40)	0.502
4	A	T	T	0.037	1.11 (0.71 / 1.72)	0.653
Global haplotype association						0.087

Rare estimated haplotypes (cumulative frequency = 0.0021) not shown.

Table 3.6 *ISL1* and risk of congenital heart disease in stage 2 Dutch whites

a) Single SNP associations

Genotypes	Controls [n (%)]	Cases [n (%)]	OR [95% CI]	P value
Stage 1				
rs3762977				
A/A	499 (77.6)	486 (78.1)	1.00	
A/G	139 (21.6)	124 (19.9)	0.92 (0.70 / 1.20)	0.528
G/G	5 (0.8)	12 (2.0)	2.46 (0.86 / 7.04)	0.093
		log-additive:	1.03 (0.81 / 1.31)	0.808
IVS1+17C>T				
C/C	571 (88.8)	560 (91.2)	1.00	
C/T	72 (11.2)	51 (8.3)	0.72 (0.50 / 1.05)	0.091
T/T	0 (0)	3 (0.5)	NA	0.975
		log-additive:	0.82 (0.57 / 1.17)	0.277
rs1017				
A/A	204 (31.7)	229 (36.4)	1.00	
A/T	319 (49.6)	297 (47.1)	0.83 (0.65 / 1.06)	0.134
T/T	120 (18.7)	104 (16.5)	0.77 (0.56 / 10.7)	0.117

b) Haplotype associations

Haplotypes	rs3762977	IVS1+17C>T	rs1017	Frequency (%)	OR [95% CI]	*P* value
1	A	C	A	0.668	1.00	
2	A	C	T	0.165	0.96 (0.73 / 1.25)	0.762
3	G	C	A	0.064	1.01 (0.72 / 1.42)	0.958
4	G	C	T	0.052	0.95 (0.62 / 1.45)	0.798
5	A	T	A	0.025	1.04 (0.60 / 1.82)	0.887
6	A	T	T	0.025	0.58 (0.31 / 1.07)	0.082
Global haplotype association						0.473

Table 3.7 *ISL1* haplotype association with risk of CHD in stage 2 whites (US, Canada, Netherlands)

Haplotypes	rs3762977	IVS1+17C>T	rs1017	Frequency (%)	OR [95% CI]*	*P* value*
1	A	C	A	0.649	1.00	
2	A	C	T	0.190	1.18 (1.00 / 1.39)	0.0485
3	G	C	T	0.093	1.04 (.084 / 1.28)	0.722
4	A	T	T	0.033	0.86 (0.61 / 1.23)	0.423
5	G	C	A	0.025	1.08 (0.77 / 1.52)	0.655
Global haplotype association						0.00003

Rare estimated haplotypes (cumulative frequency = 0.010) not shown.

* Controlling for geographical region (North American vs. Dutch)

Table 3.8 Summary *ISL1* haplotype association with risk of CHD in all whites (stage 1 & stage 2)

Haplotypes	rs3762977	IVS1+17C>T	rs1017	Frequency (%)	OR [95% CI]	*P* value
1	A	C	A	0.645	1.00	
2	A	C	T	0.192	1.27 (1.09 / 1.48)	0.0018
3	G	C	T	0.098	1.07 (0.88 / 1.30)	0.5068
4	A	T	T	0.034	1.04 (0.75 / 1.44)	0.8216
5	G	C	A	0.022	1.10 (0.78 / 1.53)	0.5928
Global haplotype association						0.000004

Rare estimated haplotypes (cumulative frequency = 0.0099) not shown.

Table 3.9 *ISL1* associations with risk of HLHS and D-TGA in white populations

	Stage 1 OR (95% CI)	P value	Stage 2* OR (95% CI)	P value	Combined† OR (95% CI)	P value
rs1017	2.04 (1.17 – 3.56)	0.012	1.36 (0.96 – 1.94)	0.086	1.48 (1.08 – 2.04)	0.016
A-C-T	2.27 (1.22 – 4.24)	0.010	1.35 (0.90 – 2.04)	0.15	1.62 (1.12 – 2.35)	0.0099

* Analyses controlled for center
† Analyses controlled for center and stage

Table 3.10 *ISL1* and risk of congenital heart disease in stage 1 US blacks/African Americans

a) Single SNP associations

Genotypes	Controls [n (%)]	Cases [n (%)]	OR [95% CI]	P value
Stage 1				
rs3762977				
A/A	46 (67.7)	21 (45.7)	1.00	
A/G	20 (29.4)	21 (45.7)	2.30 (1.03 / 5.12)	0.042
G/G	2 (2.9)	4 (8.6)	4.38 (0.74 / 25.8)	0.103
		log-additive:	2.21 (1.15 / 4.23)	0.017
IVS1+17C>T				
C/C	104 (99.0)	46 (100)	1.00	
C/T	1 (1.0)	0 (0)	NA	0.507
rs1017				
A/A	18 (22.0)	10 (21.7)	1.00	
A/T	37 (45.1)	19 (41.3)	0.92 (0.36 / 2.39)	0.871
T/T	27 (32.9)	17 (37.0)	1.13 (0.42 / 3.03)	0.803
		log-additive:	1.08 (0.66 / 1.76)	0.756

b) Haplotype associations

Haplotypes	rs3762977	IVS1+17C>T	rs1017	Frequency (%)	OR [95% CI]	*P* value
1	A	C	A	0.443	1.00	
2	A	C	T	0.336	0.65 (0.34 / 1.24)	0.195
3	G	C	T	0.218	1.99 (1.02 / 3.87)	0.044
Global haplotype association						0.051

Rare estimated haplotypes (cumulative frequency = 0.0032) not shown.

Table 3.11 *ISL1* and risk of congenital heart disease in stage 2 US blacks/African Americans

a) Single SNP associations

Genotypes	Controls [n (%)]	Cases [n (%)]	OR [95% CI]	P value
Stage 1				
rs3762977				
A/A	1235 (66.9)	31 (63.3)	1.00	
A/G	540 (29.3)	15 (30.6)	1.11 (0.59 / 2.07)	0.751
G/G	70 (3.8)	3 (6.1)	1.71 (0.51 / 5.72)	0.386
		log-additive:	1.20 (0.74 / 1.95)	0.457
IVS1+17C>T				
C/C	1803 (97.7)	49 (100)	1.00	
C/T	42 (2.3)	0 (0)	NA	0.564
rs1017				
A/A	476 (25.8)	11 (22.5)	1.00	
A/T	901 (48.8)	22 (44.9)	1.06 (0.51 / 2.20)	0.883
T/T	468 (25.4)	16 (32.6)	1.48 (0.68 / 3.22)	0.324
		log-additive:	1.23 (0.83 / 1.83)	0.306

b) Haplotype associations

Haplotypes	rs3762977	IVS1+17C>T	rs1017	Frequency (%)	OR [95% CI]	*P* value
1	A	C	A	0.496	1.00	
2	A	C	T	0.308	1.29 (0.82 / 2.03)	0.270
3	G	C	T	0.180	1.28 (0.75 / 2.19)	0.359
Global haplotype association						0.464

Rare estimated haplotypes (cumulative frequency = 0.015) not shown.

Table 3.12 Summary *ISL1* haplotype association with risk of CHD in all blacks/African Americans (stage 1 & stage 2)

Haplotypes	rs3762977	IVS1+17C>T	rs1017	Frequency (%)	OR [95% CI]	*P* value
1	A	C	A	0.492	1.00	
2	A	C	T	0.310	1.16 (0.81 / 1.66)	0.427
3	G	C	T	0.183	1.58 (1.08 / 2.31)	0.019
Global haplotype association						0.343

Rare estimated haplotypes (cumulative frequency = 0.015) not shown.

Figure 3.1 Diagnosis distribution in stage 1 and stage 2 case-control studies. Cases were chosen *a priori* to represent a wide variety of developmental phenotypes that include developmental structures aberrantly formed as derivatives of the secondary heart field. These diagnostic choices were informed from lineage tracing analyses of *Isl1*+ progenitor cells in rodents. See appendix for definitions of diagnoses.

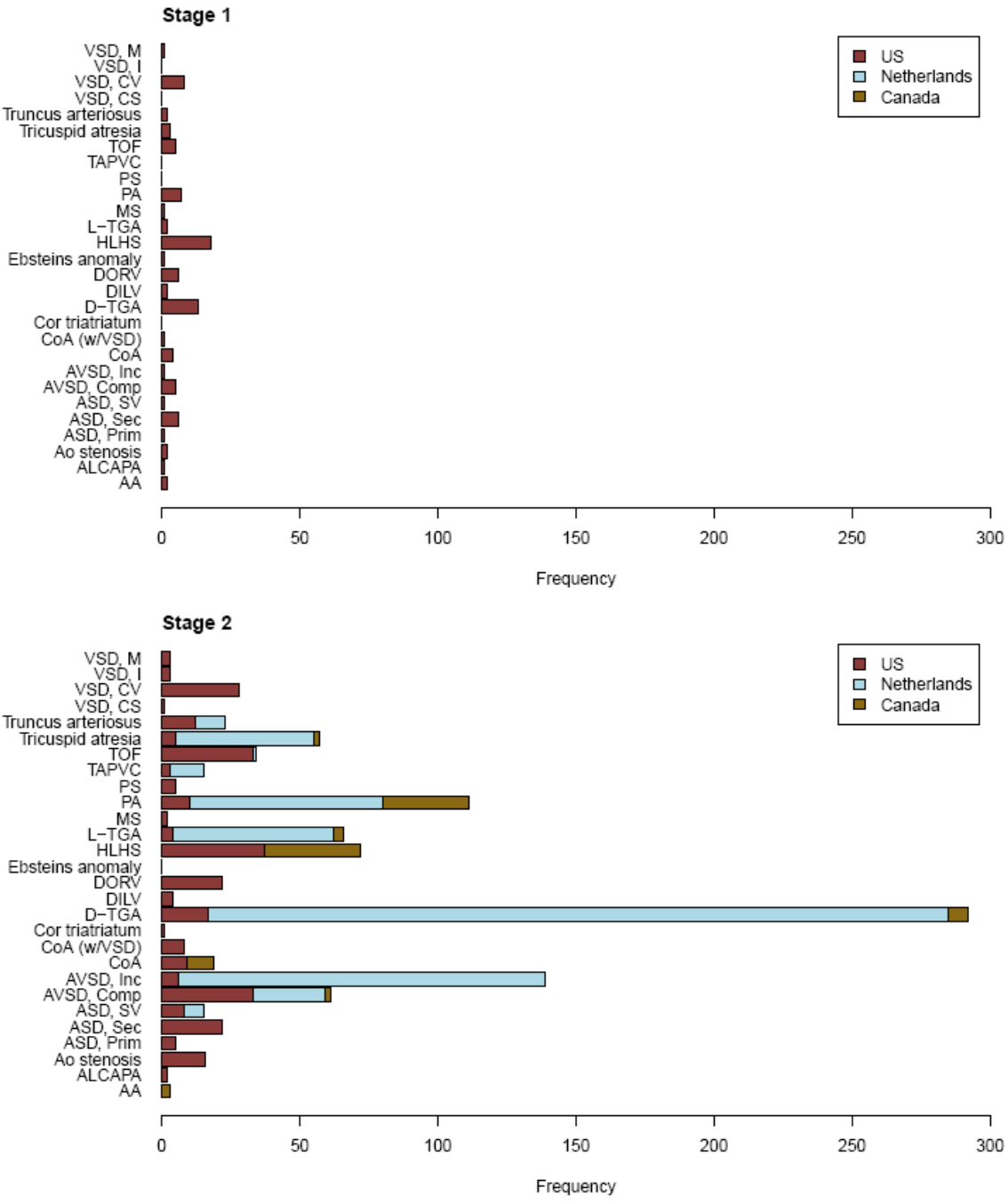


Figure 3.2 Ethnic distribution of cases and controls by cluster analysis. The first two principal components from a principal components analysis utilizing all SNPs on chromosome 5 that are contained within the Illumina HumanHap550 array are plotted for a) stage 1 cases of known ethnicity, where PC1≤0.025 captures white cases and PC1>0.025 captures black/African American cases; b) stage 2 cases of unknown ethnicity, where PC1≤0.025 defines white cases and PC1>0.025 defines black/African American cases; c) stage 1 controls of known ethnicity, where PC1≤0.0059 captures white controls and PC1>0.0059 captures black/African American controls; b) stage 2 cases of unknown ethnicity, where PC1≤0.0059 defines white controls and PC1>0.0059 defines black/African American controls.

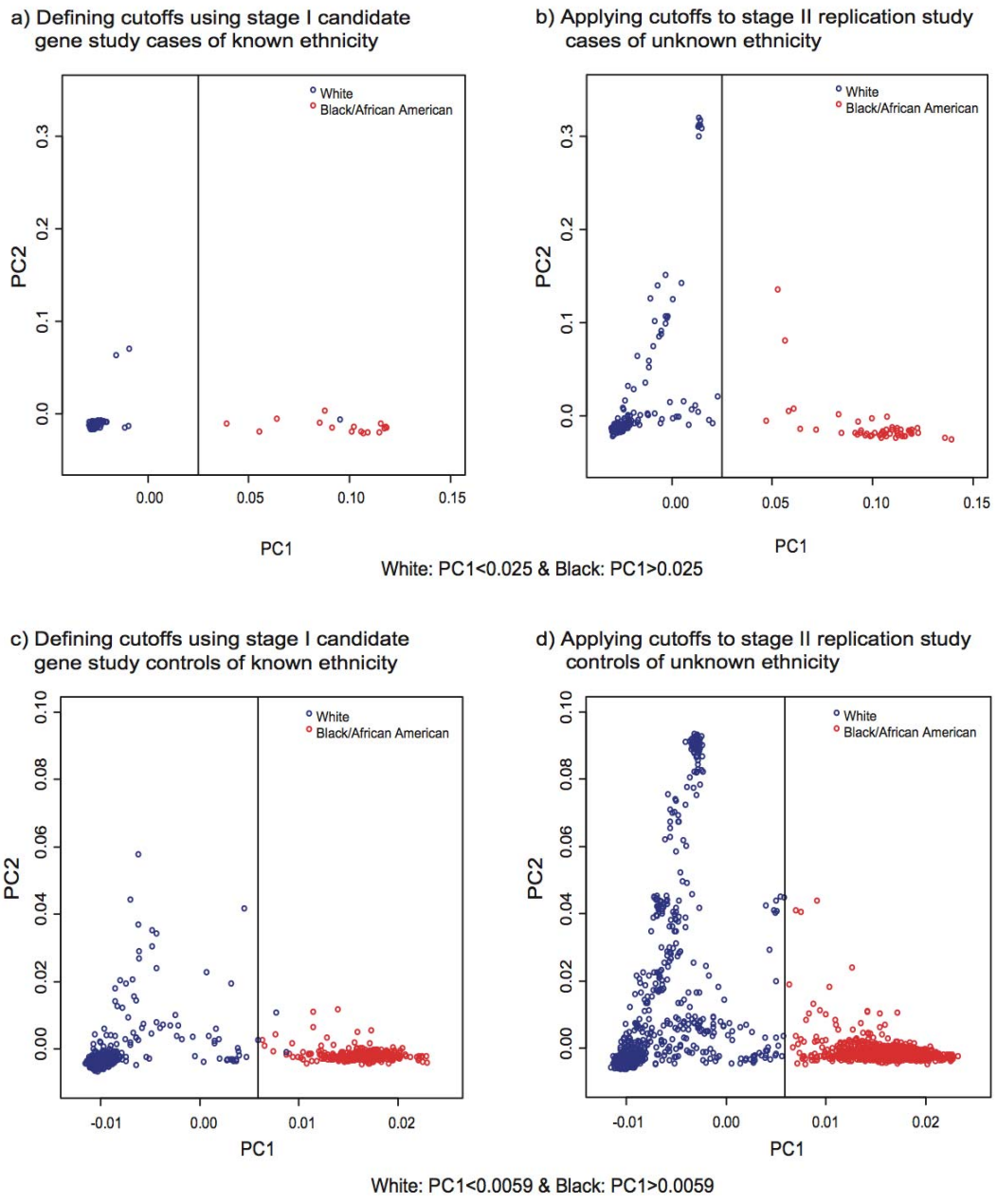


Figure 3.3 Stage 1 *ISL1* SNP associations with CHD on chromosome 5. Analysis of SNP data within and surrounding *ISL1* in stage 1 yielded 8 SNPs that were significantly associated with CHD in an ethnically heterogeneous US population. ORs, 95%CIs and P values significant at = 0.05 are depicted in black. Non-significant ORs, 95% CIs and P values are depicted in grey. The yellow highlighted region indicates the location of ISL1 on chromosome 5. Labeled SNPs: (a) rs6867206, (b) rs4865656, (c) rs6869844, (d) rs2115322, (e) rs6449600, (f) rs3762977, (g) IVS1+17C>T, (h) rs1017, (i) rs6449612.

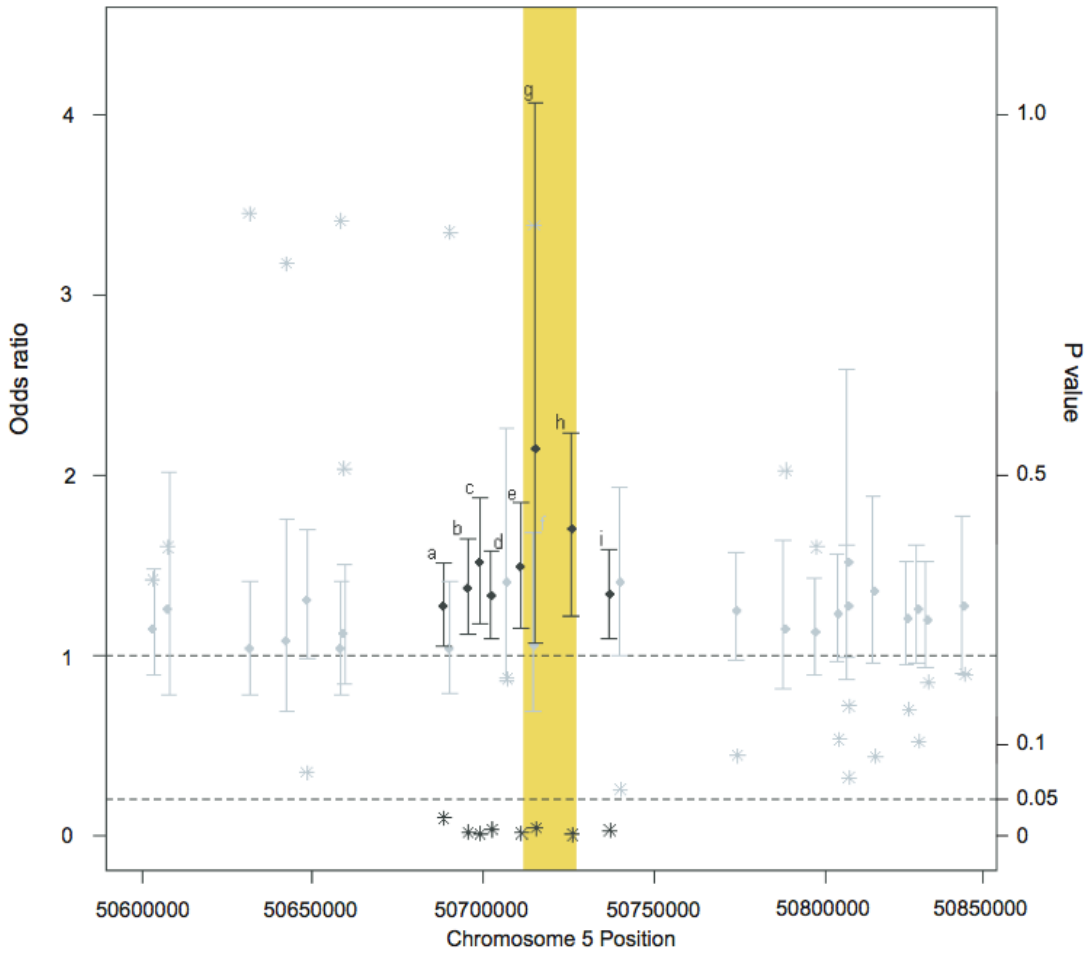


Figure 3.4 Chromosome 5 variation in the *ISL1* region. The location of ISL1 on chromosome 5 (Build 36) is depicted, where exons of the *ISL1* gene are depicted as shaded boxes, the 5' UTR and 3' UTR are depicted as white boxes, and introns are represented as black lines. The three SNPs within *ISL1* studied in stage 1 and stage 2 are depicted with respect to their location in the gene. The six SNPs flanking *ISL1* identified as significantly associated with risk of CHD in stage 1 are indicated along chromosome 5.

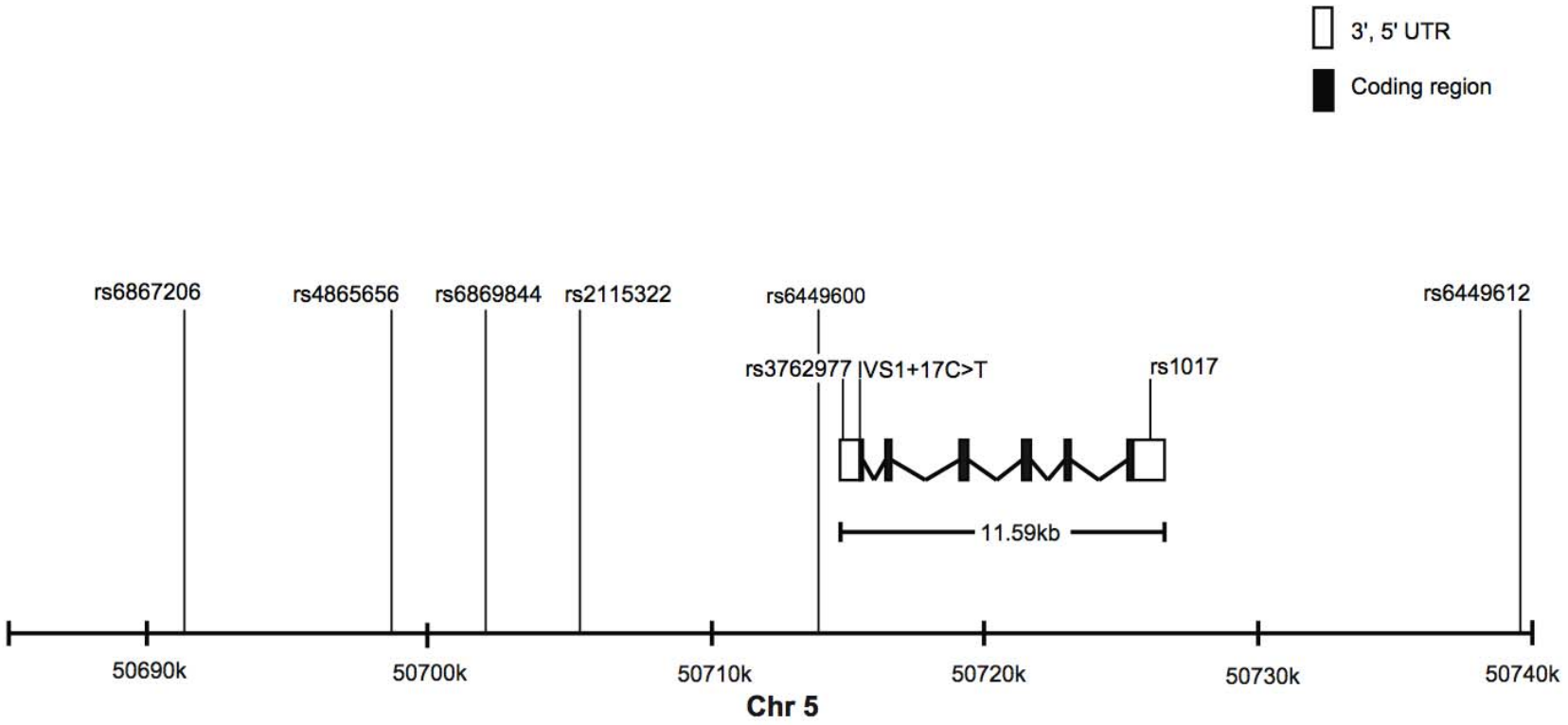


Figure 3.5 Stage 2 *ISL1* SNP associations with CHD on chromosome 5. Analysis of SNP data within and surrounding *ISL1* in stage 2 US whites yielded 10 SNPs that were significantly associated with CHD in an initial analysis of an ethnically heterogeneous US population. ORs, 95%CIs and P values significant at = 0.05 are depicted in black. Non-significant ORs, 95% CIs and P values are depicted in grey. The yellow highlighted region indicates the location of ISL1 on chromosome 5. Labeled SNPs: a) rs6867206, b) rs4865656, c) rs6869844, d) rs2115322, e) rs6449600, f) rs3762977 †, g) IVS1+17C>T †, h) rs1017 †, i) rs6449612. † SNP genotypes determined by imputation.

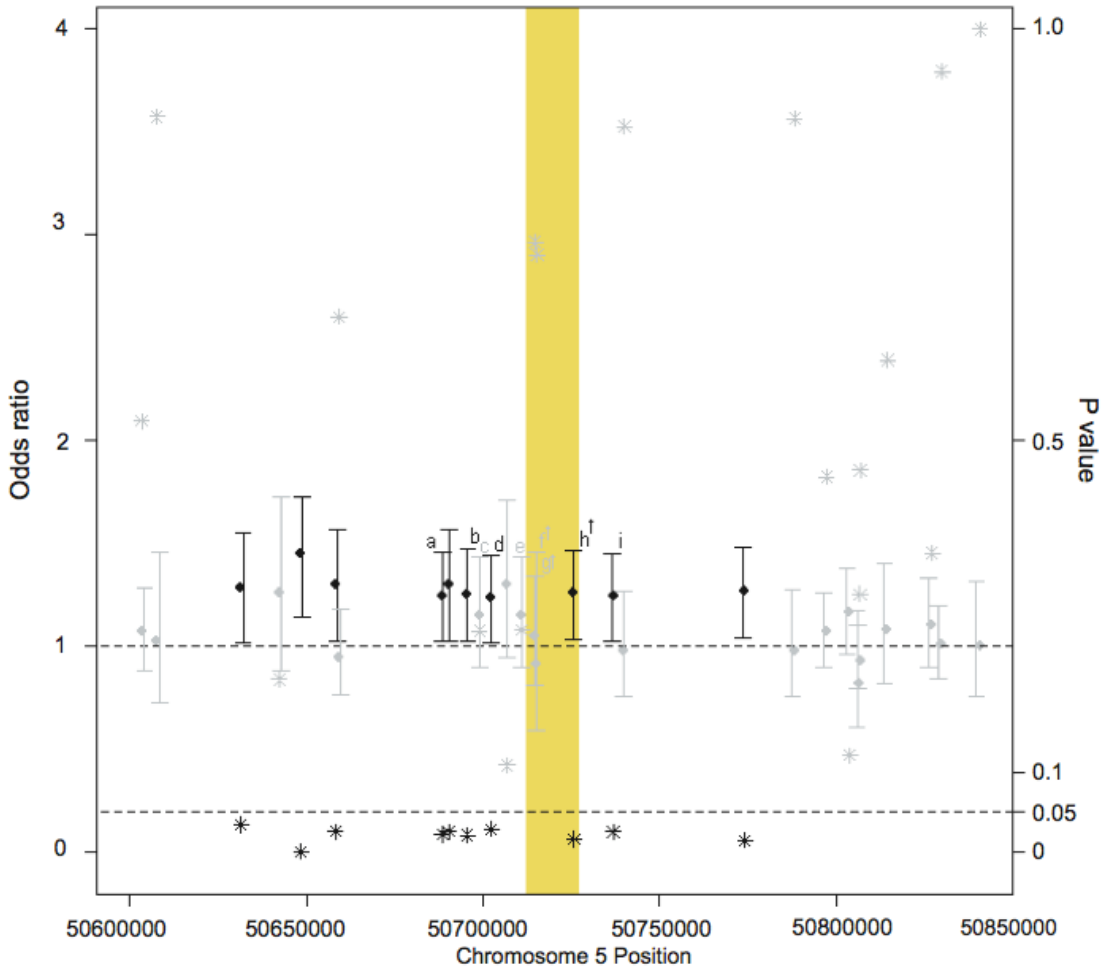


Figure 3.6 *ISL1* haplotypes and risk of congenital heart disease by race/ethnicity. a) The A-C-T risk haplotype in white stage 1 (US) and stage 2 (US, Canada, Netherlands) populations. Odds ratios (95% CIs) for each stage are denoted by black boxes (gray lines). Summary OR estimates are represented by black diamonds, where diamond width corresponds to 95% CI bounds. Box and diamond heights are inversely proportional to precision of the OR estimate. b) The G-C-T risk haplotype in black/African American stage 1 (US) and stage 2 (US) populations. Odds ratios (95% CIs) are denoted as in 2a.

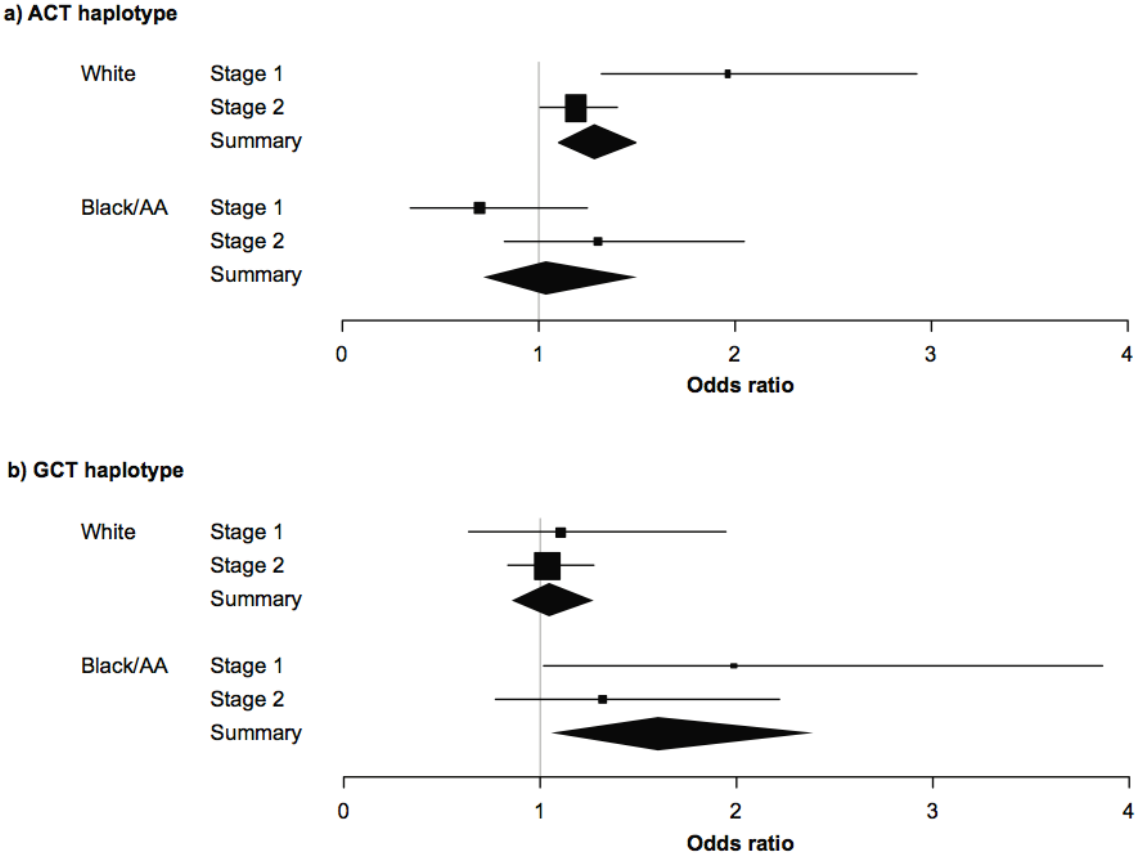
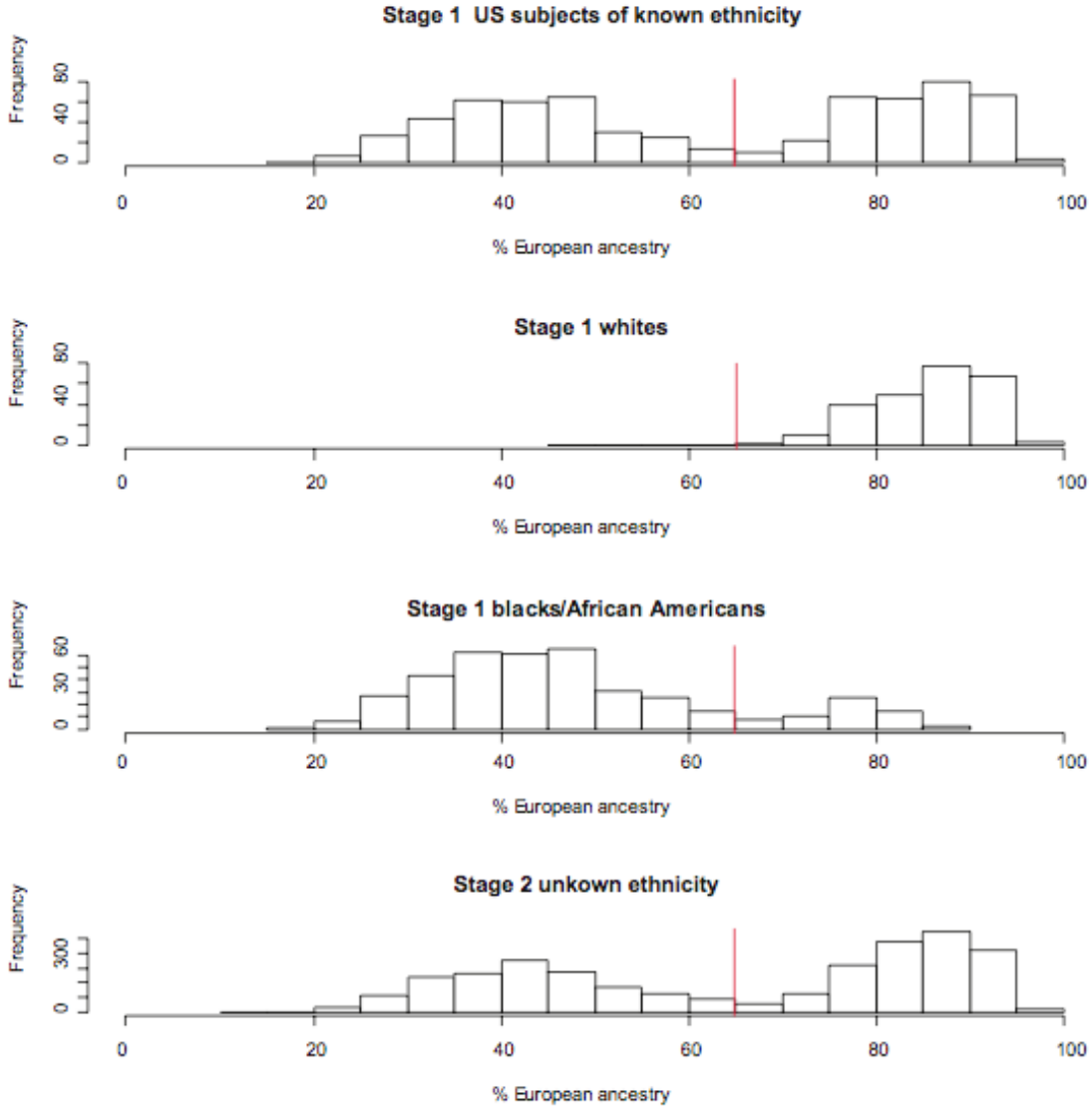


Figure 3.7 ANCESTRYMAP admixture estimation using 26 Ancestral Informative Markers. The distribution of estimated percent European ancestry for a) all stage 1 US subjects of known ethnicity (n=650), b) stage 1 US whites (n=251), c) stage 2 US blacks/African Americans (n=399), and d) stage 2 US subjects of unknown ethnicity (n=3610). 65% cutoff is represented by a red line. Individuals above 65% European ancestry were defined as white in stage 2 US subjects and below 65% were defined as black/African American.



CHAPTER 4

Pediatric cancer epidemiology among children with congenital heart disease

4.1 Introduction

Survivors living with congenital heart disease (CHD) are a large and growing population, with an estimated 1.3 million people currently living with CHD in the United States (Hoffman *et al.*, 2004). These individuals have been closely observed to understand clinical outcomes related to their specific cardiac defects, with particular emphasis on mortality, additional heart complications, overall functional status, and quality of life (Connor *et al.*, 2004; Hickey *et al.*, 2009; Schultz and Wernovsky, 2005; Verheugt *et al.*, 2008). Despite these extensive studies investigating CHD-related outcomes, the impact of other chronic diseases, such as childhood cancer, experienced by this patient population is not well understood.

Several genetic disorders are associated with both cardiac defects and an increased risk of pediatric cancers. Children with Down syndrome (trisomy 21) are at a higher risk of developing acute myeloid leukemia (AML) or acute lymphoblastic leukemia (ALL) than the general population (0-4 years: Standardized incidence ratio (SIR) (95% CI) = 56 (38 -81); 5-29 years: SIR (95% CI) = 10 (4 - 20)) (Freeman *et al.*, 1998). Approximately 44% of Down syndrome children also have various congenital heart defects, most of which are atrioventricular (45%) or ventricular septal defects (35%) (Roizen and Patterson, 2003). The majority of children with Noonan and Costello syndromes also have characteristic cardiac phenotypes, such as pulmonary valve stenosis,

hypertrophic cardiomyopathy, and atrial tachycardia (Denayer *et al.*, 2008; Tartaglia and Gelb, 2005). These disorders are caused by mutations in the *RAS* pathway (Gripp, 2005; Tartaglia and Gelb, 2005), a set of genes commonly found to be mutated in multiple cancer types (Bos, 1989; Davies *et al.*, 2002). Children with Noonan syndrome are at an increased risk of juvenile myelomonocytic leukemia (JMML) and rhabdomyosarcoma (Denayer *et al.*, 2008), while children with Costello syndrome are at an increased risk of rhabdomyosarcoma, neuroblastoma, and bladder carcinoma (Denayer *et al.*, 2008; Gripp, 2005).

These genetic syndromes provide evidence that events occurring early in development, such as chromosomal abnormalities or mutations, can result in both congenital anomalies and an increased susceptibility to cancer. We hypothesize that this phenomenon may extend beyond these currently described disorders, affecting a larger population of children with non-syndromic birth defects. Indeed, associations between various birth defects and childhood cancers have been identified. A large population-based study of childhood cancers in Great Britain found congenital malformations in 4.4% of children with solid tumors and 2.6% of children with leukemia or lymphoma (Narod *et al.*, 1997). Furthermore, several large cohort studies of children with congenital anomalies in Toronto, Northern England, Norway, and Sweden have identified associations between subsets of birth defects and pediatric cancers, including hepatic cancers among children with digestive system anomalies, lymphomas and bone tumors among children with musculoskeletal deformities, and kidney tumors among children with genitourinary malformations (Agha *et al.*, 2005; Bjorge *et al.*, 2008; Rankin *et al.*, 2008).

Although these studies provide evidence for an association between developmental abnormalities and cancer risk, causal links have yet to be identified. Studying cancer incidence among children with specific subsets of birth defects may provide a way to unravel the complex biology underlying these relationships, particularly considering that most of the excess cancers identified in these studies were related to only specific types of birth defects. Congenital heart disease is an ideal candidate for such an analysis, as most cases of CHD are not explained by known risk factors such as maternal exposures or chromosomal abnormalities. The relationship between childhood cancer and congenital heart disease has not previously been investigated, yet this would provide an opportunity to gain significant insight into the etiologies of both sets of diseases. In this chapter, I investigate this relationship and demonstrate an excess of pediatric cancers in a large, hospital-based cohort of children with congenital heart disease.

4.2 Subjects and methods

4.2.1 Study design

To investigate the relationship between childhood cancers and congenital heart disease, we conducted a retrospective cohort study of CHD at the Children's Hospital of Philadelphia. The source population for this study consists of any child with a congenital heart defect requiring operative repair at CHOP between the ages of 0 and 18 years. Although the majority of individuals in this population reside in Pennsylvania and New Jersey (77.7%), this is not a geographically restricted source population since patients from across the country are referred to the Cardiac Center at CHOP. Follow-up for the diagnosis of any malignant neoplasm was complete through Jul 22, 2009 when data were

ascertained through the CHOP cancer registry. Incidence rates of cancer were first calculated within the CHOP cohort, taking into consideration potential risk factors such as genetic syndromes and diagnostic radiation exposure. We then compared the observed number of cancers in this cohort to the number of expected cancers based on US pediatric cancer incidence rates, standardized by age and stratified by potential risk factors where appropriate.

4.2.2 Subjects

Congenital heart disease cases were ascertained for this retrospective cohort study at the Children's Hospital of Philadelphia on a protocol approved by the Institutional Review Boards of CHOP and the University of Michigan. Cases were selected for inclusion in the study if they had undergone operative repair of a cardiac defect from January 1, 2001 to July 22, 2009 at the CHOP Cardiac Center. Subjects also had to be 18 years of age or less at the time of operative repair. Cases were ineligible for the study if the primary or secondary indication for operation at the time of ascertainment was not classified as a congenital heart defect (**Table 4.1)**.

4.2.3 Identification of incident cancers

Incident cancers were identified from the CHOP cancer registry at the end of follow-up. All patients in the CHOP cancer registry on July 22, 2009 were queried by medical record number and date of birth to identify patients from the CHD cohort. Date of cancer diagnosis and cancer histology / behavior were obtained for each member of the CHD cohort also present in the CHOP cancer registry. Only malignant diagnoses were

classified as incident cancers.

4.2.2 Data collection

Collection of demographic and clinical information for eligible CHD cases was performed at the Children's Hospital of Philadelphia. Medical records were reviewed for demographic variables including date of birth, sex, race, and state of residence at the time of operation. Clinical information obtained from medical records included date of operation, diagnoses made at the time of operation, operation type, and genetic syndrome or other chromosomal abnormality.

Data from diagnostic radiation exams performed at the Children's Hospital of Philadelphia during the follow-up period were obtained from the Department of Radiology at CHOP. These data included plain films, computerized tomography (CT) scans, fluoroscopy procedures, nuclear medicine exams, and the corresponding date of administration for each exam. Radiation exposure in millisieverts (mSv) resulting from each exam was calculated using a comprehensive radiology look-up-table, which is based on the age at which the exam was administered.

Additionally, data from all cardiac catheterizations performed at CHOP during the follow-up period were obtained from the Cardiology Department. These data included the date of exam, weight of the patient at the time of the exam (kg), and duration of the exam (minutes). Radiation exposure in millisieverts (mSv) resulting from each catheterization was calculated using a two-stage approach (**Table 4.2**). First, linear regression was used to estimate the relationship between fluoroscopy time and measured dose-area product (μGy•m2) using available data from CHOP patients for whom dose-

area product was measured directly. The estimates from these models were then used to calculate dose-area product for each patient based on fluoroscopy time and weight at the time of exam. Dose-area product was then converted to total effective radiation dose (mSv) using a conversion factor calculated from previous radiologic phantom studies (ATOM®, Computerized Imaging Reference Systems, Inc., Norfolk, VA) performed at CHOP in conjunction with a radiation physicist.

For patients that developed cancer during the follow-up period, we excluded any exam administered from three months prior to the date of cancer diagnosis to the end of follow-up. Cumulative radiation exposure (mSv) for each subject in the cohort was calculated by summing the individual exposures received from each exam, including any plain films, CT scan, fluoroscopy procedures, nuclear medicine exams, or cardiac catheterizations.

4.2.3 Statistical methods

All statistical analyses were performed in SAS (version 9.2) and graphics were prepared in R (version 2.10.1). Descriptive statistics were calculated for categorical variables using frequency tables and were calculated for continuous variables using the means procedure. Radiation exposure was examined as both a continuous variable and a categorical variable. Densities for radiation exposure were estimated using the density function in the stats R package.

Each participant contributed person-years to this analysis, calculated separately for cancer and non-cancer cases. For CHD patients that did not develop cancer during the follow-up period, person-years were calculated from date of birth to the end of

follow-up. For CHD patients that were identified to have developed cancer during the follow-up period, person-years were calculated from date of birth to the date of cancer diagnosis.

Rates of cancer within the cohort were calculated by Poisson regression using the genmod procedure. Pediatric cancer rates by five-year age intervals were obtained from SEER using data from 2000-2006 (http://seer.cancer.gov/statistics/), and these rates were used to calculate the expected numbers of cancer in this cohort using the means procedure. Age-standardized incidence ratios, 95% confidence intervals, and corresponding p-values were calculated using Poisson regression as implemented by the genmod procedure. Sex, genetic syndrome status, and radiation exposure were included as covariates in these models when appropriate.

4.3 Results

4.3.1 CHOP cohort

To investigate the epidemiology of pediatric cancers among children diagnosed with congenital heart disease, we conducted a retrospective cohort study of CHD at the Children's Hospital of Philadelphia (CHOP). A total of 5,162 patients underwent at least one operation at the CHOP Cardiac Center from January 1, 2001 to July 22, 2009. Of this total patient population, 4,805 (93.1%) patients underwent repair for a congenital heart defect and 4,523 (87.6%) children were also 0-18 years of age at the time of operation.

Among these 4,523 eligible CHD patients, children were nearly equally distributed among age categories, with a slightly higher proportion of neonates (38.2%)

compared to infants (29.2%) and children (32.6%) (**Table 4.3**). The median age at operation was 0.25 years (mean = 2.5, standard deviation (SD) = 4.5), reflecting that most of the cardiac defects in this cohort required operative repair early in life. Children in the cohort were slightly more likely to be male (54.6%) compared to female (45.3%). Race was missing for a substantial proportion of this cohort (34.7%); among those with known race, children were most likely to be white (66.7%) or black/African American (19.9%).

The most common congenital heart defects in this cohort were patent ductus arteriosus, ventricular septal defects, tetralogy of Fallot, hypoplastic left heart syndrome, coarctation of the aorta, and d-transposition of the great arteries (**Table 4.1**). While the CHD diagnoses observed among these children represented a range of moderate to severe defects, most were severe and required operative repair in early infancy. Although mild or moderate defects that present later in life or do not require operative repair at all are potentially associated with cancer development, these defects are not captured in this cohort.

The presence of genetic syndromes or other chromosomal abnormalities in these children was of particular interest in this study because of the association between pediatric cancer incidence and some genetic syndromes. Consistent with current estimates of the proportion of CHD attributed to chromosomal abnormalities in the literature (Pierpont *et al.*, 2007), 13.8% of all subjects had a genetic syndrome or other chromosomal abnormality (**Table 4.4**). The most common syndromes were Down syndrome (45.0%) and DiGeorge syndrome (13.0%). An additional 14.5% of these 625 children had dysmorphic features but no identified genetic syndrome. Among the identifiable genetic syndromes were several that are associated with increased rates of

cancer, including Down syndrome (trisomy 21) (Freeman *et al.*, 1998), Hirschprung disease (Sijmons *et al.*, 1998), LEOPARD syndrome (Schrader *et al.*, 2009), neurofibromatosis (Asthagiri *et al.*, 2009; Sorensen *et al.*, 1986), Noonan syndrome (Denayer *et al.*, 2008), trisomy 18 (Schnater *et al.*, 2003), and Turner syndrome (Schoemaker *et al.*, 2008).

4.3.2 Pediatric cancer incidence

Among all 5,162 children seen at the CHOP Cardiac Center from 01/01/2001 to 07/22/2009, 57 were identified as diagnosed with cancer within the CHOP cancer registry by the end of follow-up. Four of these patients were older than 18 years at the time of their cardiac operation and were excluded from our analyses. The primary indications for operation at the Cardiac Center for 31 of these 57 cancer patients were non-CHD diagnoses, including cardiac tumors (19.4%), lung diseases (19.4%), mediastinal or pleural diseases (41.9%), and heart or lung transplants (12.9%). Although excluded from our subsequent analyses, we observed that 4 of the 24 patients that underwent heart or lung transplantation developed cancer during follow-up, three with lymphoproliferative diseases and one with a rhabdomyosarcoma. This is consistent with the observation that patients who have undergone solid organ transplantation have a 5- to10-fold increase in cancer risk (Gross *et al.*, 2010), the most common being posttransplant lymphoproliferative disease (PTLD).

We next restricted our analysis to include only the 4,523 eligible patients comprising the CHD cohort, limited to those children who had undergone operative repair of a congenital heart defect at the age of 18 years or less. A total of 23 children

were diagnosed with cancer during the follow-up period, corresponding to an incidence rate of 70 per 100,000 person-years. Compared to the expected number of cancers in this cohort based on SEER-estimated rates of pediatric cancer in the United States, this represents a 3.72-fold increase in pediatric cancer incidence (Standardized Incidence Ratio (SIR) = 3.72, 95% CI = 1.53 – 9.04, p = 0.0037).

The locations and histologies of these incident cancers were variable, including brain and other nervous system tumors, hematological tumors, a neuroendocrine tumor, a soft tissue tumor, and other solid tumors (**Table 4.5**). The types of cardiac defects observed among these children were qualitatively similar to the distribution of CHD in the total cohort. Although not statistically significantly different, children who developed cancer were followed on average for 6 months longer than children who did not develop cancer during follow-up (p=0.31). Children diagnosed with cancer were also slightly older at the time of operative repair for their cardiac defect compared to cancer-free CHD patients and were more likely to be female (**Table 4.3**). The rate of cancer among females was not significantly different compared to males (Hazard ratio (HR) = 1.35, 95% CI 0.60 – 3.06, p=0.47) (**Table 4.6**).

To understand the relationship between known chromosomal abnormalities and cancer development in this cohort, we next examined the distribution of tumors by genetic syndrome diagnoses (**Figure 4.1**). Almost half (44.5%) of these 23 cancer patients had a genetic syndrome, corresponding to an estimated incidence rate of 250 per 100,000 person-years among children with any genetic syndrome (**Table 4.6**). The most common syndrome among these 10 children was Down syndrome, which was diagnosed in 7 patients with cancer and was associated exclusively with hematological cancers

including precursor B-cell leukemia, lymphoproliferative disease, and myeloproliferative disease (**Table 4.5, Figure 4.1**). The number of cancers among all 281 children with Down syndrome was 22.83 times higher (95% CI 0.61 – 849.50, p=0.09) than expected based on US pediatric cancer rates. Additionally, one patient with Turner syndrome developed a neuroblastoma, one patient with monosomy 7 developed myelodysplastic syndrome, and one patient with dysmorphic features developed a cranial teratoma. While not statistically significant, the number of cancers among the remaining 344 children with genetic syndromes other than Down syndrome was 6.07 times higher (95% CI 0.30 – 123.22, p=0.24) than expected. Considering all genetic syndromes together, there was a 12.49-fold increase (95% CI 1.28 – 121.74, p=0.03) in the number of cancers among all 625 children with a genetic syndrome compared to the 5 cancers expected based on US rates (**Table 4.7**).

An additional 13 children without an identifiable genetic syndrome were also diagnosed with cancer at CHOP, corresponding to an incidence rate of 49 per 100,000 person-years (**Table 4.6**). In general, the proportions of cancer types observed in this group were consistent with the distribution of pediatric cancers as reported by SEER, where hematological cancers were the most common (38.5%) followed by brain or central nervous system (CNS) tumors (30.8%). In addition, an embryonal rhabdomyosarcoma was diagnosed in a child with dilated cardiomyopathy and a teratoma was diagnosed in a child with patent ductus arteriosus; both teratoma and rhabdomyosarcoma are relatively common pediatric tumors. Two uncommon pediatric tumors were also observed: one paraganglioma and one hepatoblastoma. Together, these 13 cancers represent a 2.41-fold increase (95% CI 0.88 – 6.60, p=0.086) in the number of

cancers expected among children without a genetic syndrome (**Table 4.7**). Children with any genetic syndrome had a higher rate of cancer (Rate ratio (RR) = 4.13, 95% CI 1.71 – 9.96, p=0.0016) and a higher SIR (SIR ratio = 5.18, 95% CI 0.43 – 62.32, p=0.20) compared to children with no genetic syndromes. However, these data provide evidence that there is still a substantial increase in pediatric cancer rates among CHD patients that is not attributable to identifiable genetic syndromes alone.

Exposure to diagnostic radiation was also of interest in this study as a potential risk factor for cancer development. Data from 132,208 diagnostic radiation exams were available from the CHOP Radiology and Cardiology departments for 4,162 children in the cohort. The types of exams represented in this data set include plain films, CT exams, nuclear medicine exams, fluoroscopy procedures, and cardiac catheterization. We found that the total effective radiation exposure in this cohort was low but quite variable (**Figure 4.2**). The median exposure in the cohort was 1.01 millisieverts (mSv), although dosages ranged between 0.01 mSv and 518.72 mSv. The median exposure in this cohort is comparable to the estimated average annual effective dose in the U.S. population of 1.2 mSv.

Radiation exposure was marginally higher among the 14 cancer patients with available radiation data compared to the remaining patients in the cohort, though this increase was not statistically significant (Wilcoxon p = 0.62) (**Figure 4.2**). Within the cohort, the rate of pediatric cancer increased by 0.4% for every one millisievert increase in diagnostic radiation (RR=1.004, 95% CI 0.9935 – 1.1957, p=0.42). Consistent with this analysis, the number of cancers among children who received greater than 1.01 mSv of radiation was 2.88 times greater (95% CI 0.74 – 11.27, p=0.13) than the number

expected based on US pediatric cancer rates, while the number of cancers among children who received less than 1.01 mSv of radiation was 1.99 times greater (95% CI 0.50 – 7.94, p=0.33) than expected. This difference in the SIRs was not statistically significant (SIR ratio = 1.45, 95% CI 0.21 – 10.10, p=0.71). These data provide evidence for a previously unrecognized association between CHD and childhood cancer that is not fully accounted for by genetic syndromes or radiation exposure.

4.4 Discussion

We have identified an almost 4-fold increase in the rate of cancer among children with congenital heart disease. Down syndrome and other genetic syndromes were strongly associated with cancer risk, but there remains an unexplained 2.4-fold increase in risk among children without any identifiable genetic disorders (p = 0.086). Investigating the environmental exposures and genetic abnormalities that are found among the CHD patients that developed cancer may provide significant insight into the causes of these two sets of diseases.

Radiation exposure was not significantly associated with an increased rate of cancer in this study. The relationship between diagnostic radiation and childhood cancer risk is unclear, and studies are inconclusive about the dose, duration, and the induction period between radiation exposure and tumorigenesis in childhood. A classic study reported an increased risk of brain tumors among individuals treated with radiation for tinea capitis in childhood, although the overall dosage was high (1-2Gy) and it is important to note that cancers were not identified until an average of 30 years after exposure (Sadetzki *et al.*, 2005). Another study investigating postnatal x-ray exposure

and childhood cancer risk found no increase in cancer rates, but the average exposure was 7 μSv, about 1000-fold less than the median exposure in the Sadetzki study (Hammer *et al.*, 2009). Finally, a study of cancer following cardiac catheterization found a 2.3-fold increase in cancer rates, where cancers were identified from 5 to 38 years after exposure (Modan *et al.*, 2000). Considering these findings, it is not altogether surprising that we cannot detect an excess of cancers attributable to radiation exposure given that most children were exposed to low levels of radiation and were followed on average for only 4.5 years. We may also simply have insufficient power to detect a meaningful increase in cancer risk if the true effect is small. Using a Poisson model to calculate power with α=0.05, we had only 7.5% power to detect a 1.44-fold difference in pediatric cancer rates between children with less than vs. greater than median radiation. Using this same model, we had only 8.9% power to detect an SIR of 2.78 and 9.6% power to detect an SIR of 3.02 among children with less than and greater than median radiation, respectively. Long term follow-up of the CHOP cohort is warranted, especially given the exceptionally high quality radiation dosimetry data available and clinical data for these children.

One limitation of this study is that we had limited sensitivity to identify cancers diagnosed at institutions other than CHOP. Approximately 22.3% of the CHD cohort resided outside of Pennsylvania or New Jersey at the time of their CHD operation, and it is likely that cancers within this group would not have been captured by the CHOP cancer registry unless these patients returned to CHOP for long-term follow-up. This type of misclassification would be expected to result in underestimation of the actual number of incident cancers among children in this cohort. Thus, the results presented in

this dissertation are most likely underestimates of the true rates of pediatric cancer among children with CHD.

To explore this issue given the available data, we performed a subset analysis where we restricted the cohort to children who resided in Pennsylvania or New Jersey at the time of their CHD operation. The results of this analysis were qualitatively similar to those described above, but with limited power to detect associations due to the decrease in sample size. Among the 3,509 children that resided in Pennsylvania or New Jersey at the time of their CHD operation, 16 developed cancer during the follow-up period, corresponding to a 3.35-fold increase in the rate of cancer compared to the general population (SIR=3.35, 95% CI 1.21 – 9.32, p=0.020). Children with any genetic syndrome had cancer rates 7.91 times higher than children with no genetic syndrome (HR=7.91, 95% CI 2.97 – 21.08, $p < 0.0001$). Correspondingly, children with genetic syndromes had rates of cancer 14.94 times higher than the general population (SIR=14.94, 95% CI 0.94 – 237.52, $p = 0.055$). While not statistically significant, children without genetic syndromes had a 1.89-fold higher rate of cancer than the general population (SIR=1.89, 95% CI 0.58 – 6.13, p 0.029).

A related issue is that while the CHD diagnoses observed among these children represented a range of moderate to severe defects, most were severe and required operative repair in early infancy. Although mild or moderate defects that present later in life or do not require operative repair at all are potentially associated with cancer development, these defects are not captured in this cohort. The estimates in this study would be biased if the association between cancer and CHD is substantially different in magnitude between children with severe defects compared to children with milder

defects. Given this limitation, the estimates in this study are only generalizable to children with severe CHD requiring operative repair, as the association between CHD and cancer risk among children with mild CHD not requiring operative repair remains unknown.

The cohort experienced a short follow-up period, with an average follow-up time of 4.5 years. This has several implications. First, it is likely that some children in the cohort will develop cancer after the end follow-up in this study, particularly neonates and infants who could be no older than 10 years at the end of follow-up. Second, those cancers that did develop are more likely to be attributed to prenatal exposure or genetic abnormalities. The latent period for cancers attributable to exposures such as postnatal diagnostic radiation is likely to be longer than the average follow-up time, given the discussion of radiation and cancer risk above.

We identified a higher rate of cancer among children with genetic syndromes, with a particularly strong association observed between Down syndrome and risk of hematological malignancies. This is consistent with the known association between chromosomal abnormalities and cancer risk. However, it is interesting to note that cancer was only identified among children with Down syndrome, Turner syndrome, and monosomy 7 developed cancer in this study even though an additional 22 children had a disorder known to be associated with tumor development. For example, none of the 10 patients with Noonan syndrome were diagnosed with cancer over the approximately 4.5 person-years of cumulative follow-up, with the oldest patient now 25 years of age. This is certainly influenced by the limited power of our study to detect an increased risk for subtypes of genetic syndromes based on: 1) overall rarity of childhood cancer, even

among a high-risk group, 2) the short follow-up period, and 3) incomplete ascertainment of incident cancers as discussed above.

Another limitation is that the estimates of childhood cancer rates in this study may be affected by the types of defects observed in the CHOP cohort. Since children were ascertained on the basis of operative repair for their CHD, mild or moderate defects that do not require an operation were not captured. However, it is possible that these defects are also related to childhood cancer risk and should be included when investigating the relationship between CHD and pediatric cancers. Nonetheless, this study provides strong evidence for an increase in cancer risk among children with CHD, warranting further investigation to understand the basis for this relationship.

Table 4.1 Congenital heart defects of patients in CHOP cohort (n=4523)

CHD diagnosis*	n (%)
AI	64 (1.42)
AP window	9 (0.20)
ASD	280 (6.20)
AVSD	304 (6.73)
Aberrant subclavian artery	3 (0.07)
Ao aneurysm	21 (0.46)
Ao dissection	1 (0.02)
Ao stenosis	122 (2.70)
Aortic arch hypoplasia	12 (0.27)
Arrhythmia	190 (4.20)
Bilateral SVC	4 (0.09)
CCAVC, unbalanced	4 (0.09)
CDH	2 (0.04)
CoA	281 (6.22)
Conduit failure	6 (0.13)
Cor triatriatum	10 (0.22)
Coronary artery anomaly	70 (1.55)
D-TGA	256 (5.66)
DCM	34 (0.75)
DCRV	30 (0.66)
DILV	49 (1.08)
DIRV	1 (0.02)
DOLV	3 (0.07)
DORV	109 (2.41)
Ebstein's anomaly	12 (0.27)
HCM	20 (0.44)
HLHS	399 (8.83)
Heart failure	21 (0.46)
IAA	48 (1.06)
L-TGA	5 (0.11)
MR	41 (0.91)
MS	11 (0.24)
MV abnormality	3 (0.07)
MV atresia	11 (0.24)
PA	46 (1.02)
PA stenosis	14 (0.31)
PA/VSD	46 (1.02)
PAPVC	27 (0.60)
PDA	703 (15.56)
PHTN	21 (0.46)
PI	6 (0.13)
PS	4 (0.09)
PV stenosis	19 (0.42)

* See appendix for definitions

Table 4.1 continued

PVD	1 (0.02)
Pericardial disease	11 (0.24)
Pulmonary valve disease	5 (0.11)
RVOTO	2 (0.04)
Single ventricle, other	65 (1.44)
TAPVC	49 (1.08)
TOF	400 (8.85)
TOF/APV	12 (0.27)
Tracheal stenosis	5 (0.11)
Tricuspid atresia	51 (1.13)
Tricuspid stenosis	9 (0.20)
Tricuspid valve disease	15 (0.33)
Truncus arteriosus	63 (1.39)
VSD	406 (8.96)
Vascular Ring	104 (2.30)

* See appendix for definitions

Table 4.2 Cardiac catheterization conversions from fluoroscopy time to effective radiation dose (mSv)

Weight category (kg)	Dose-area product* (µGy•m2)	Effective radiation dose† (mSv)
≤ 5	(MIN†) x 12.79 + 64.11	(µGy•m2) x 0.02072
(5-15]	(MIN) x 30.95 + 171.14	(µGy•m2) x 0.00914
(15-30]	(MIN) x 67.36 + 642.18	(µGy•m2) x 0.0068
(30-60]	(MIN) x 249.46 + 332.99	(µGy•m2) x 0.00206
> 60	(MIN) x 468.41 + 607.72	(µGy•m2) x 0.00175

* Models estimated by linear regression

† Total fluoroscopy times in minutes

Table 4.3 Demographic and clinical characteristics of CHOP CHD cohort (n=4,523)

Variable		Total n=4523 n (%)	Cancer-free n=4500 n (%)	Incident cancers n=23 n (%)	P value*
Age at operation (years)		Mean=2.5 SD=4.5 Median=0.25	Mean=2.5 SD=4.5 Median=0.25	Mean=4.1 SD=6.3 Median=0.75	0.25
Age category					
	Child (1-18 years]	1473 (32.6)	1462 (32.5)	11 (50.0)	
	Infant (0-1 years]	1321 (29.2)	1314 (29.2)	7 (31.8)	
	Neonate [0 years]	1727 (38.2)	1723 (38.3)	4 (18.2)	0.10†
Sex					
	Female	2051 (45.4)	2039 (45.3)	12 (52.2)	
	Male	2470 (54.6)	2459 (54.7)	11 (47.8)	0.54
Race					
	White	1971 (43.6)	1958 (43.5)	13 (56.5)	
	Black/African American	587 (13.0)	585 (13.0)	2 (8.7)	
	Asian	40 (0.9)	40 (0.9)	0 (0)	
	Native American	1 (0.02)	1 (0.02)	0 (0)	
	Other	356 (7.9)	355 (7.9)	1 (4.3)	
	Unknown	1568 (34.7)	1561 (34.7)	7 (30.4)	0.73†
Follow-up time (years)		Mean=4.5 SD=2.3	Mean=4.5 SD=2.3	Mean=5.1 SD=2.4	0.31

* χ^2 test or t-test of difference between incidence cancers and cancer-free subjects

† Fisher's exact p-value

Table 4.4 Genetic syndromes & other chromosomal abnormalities in CHOP cohort (n=625)

Description	n (%)
22q11 deletion/DiGeorge syndrome	81 (12.96)
Alagile syndrome	5 (0.80)
Asplenia	38 (6.08)
CHARGE	4 (0.64)
Cleft lip/palate	1 (0.16)
Cystic fibrosis	1 (0.16)
Ellis van Creveld syndrome	2 (0.32)
Freeman Sheldon syndrome	1 (0.16)
Hirschprungs disease	2 (0.32)
Horner syndrome	2 (0.32)
Jacobson (11q deletion) syndrome	1 (0.16)
Joubert syndrome	1 (0.16)
Kawasaki disease	3 (0.48)
Klinefelter syndrome	1 (0.16)
Leopard syndrome	1 (0.16)
Loeys-Dietz syndrome	2 (0.32)
LQT syndrome	25 (4.00)
Marfan syndrome	6 (0.96)
Muscular dystrophy/myopathy	1 (0.16)
Neurofibromatosis	2 (0.32)
Noonan syndrome	10 (1.60)
Peters syndrome	1 (0.16)
Rubinstein-Taybi syndrome	1 (0.16)
Scimitar syndrome	9 (1.44)
Trisomy 18	4 (0.64)
Trisomy 21	281 (44.96)
Turner syndrome (45XO)	20 (3.20)
William syndrome	7 (1.12)
Wolff-Parkinson-White syndrome	5 (0.80)
X-linked chronic granulomatous disease	1 (0.16)
Dysmorphic features, no identified syndrome	91 (14.56)
Other chromosomal abnormality	15 (2.40)

Table 4.5 Cancers in CHOP cohort (n=23)

Sex	CHD diagnosis*	Cancer diagnosis	Cancer type	Genetic syndrome
Male	HLHS	Cranial teratoma	Brain/CNS	DFNS†
Female	CoA	Glioma	Brain/CNS	
Male	TOF	Neuroblastoma	Brain/CNS	
Female	CoA	Neuroblastoma	Brain/CNS	
Female	HLHS	Neuroblastoma	Brain/CNS	Turner syndrome
Female	TOF	Primitive neuroectodermal tumor	Brain/CNS	
Female	PDA	Hodgkin lymphoma	Hematological	
Female	PHTN	Lymphoma	Hematological	
Female	AVSD	Lymphoproliferative Disease/Disorder	Hematological	Trisomy 21
Female	DCM	Lymphoproliferative Disease/Disorder	Hematological	
Male	VSD	Myelodysplastic syndrome	Hematological	Monsomy 7
Male	VSD	Myeloproliferative disease	Hematological	Trisomy 21
Male	AVSD	Myeloproliferative disease	Hematological	Trisomy 21
Male	AVSD	Myeloproliferative disease	Hematological	Trisomy 21
Male	DILV	Precursor B-cell lymphoblastic leukemia	Hematological	
Female	ASD	Precursor B-cell lymphoblastic leukemia	Hematological	Trisomy 21
Female	VSD	Precursor B-cell lymphoblastic leukemia	Hematological	Trisomy 21
Male	AVSD	Precursor B-cell lymphoblastic leukemia	Hematological	Trisomy 21
Male	Heart failure	Precursor T-cell lymphoblastic lymphoma	Hematological	
Male	Ao aneurysm	Paraganglioma	Neuroendocrine	
Female	ASD	Hepatoblastoma	Other solid	
Female	PDA	Teratoma	Other solid	
Male	DCM	Embryonal rhabdomyosarcoma	Soft tissue	

* See appendix for definitions
† DFNS: dysmorphic features, no syndrome

Table 4.6 Pediatric cancer rates in the CHOP cohort (n=4,523)

	n	Cancer incidence rate (per year)	Hazard ratio (95% CI)	P value
Overall cohort cancer rate	23	0.00070		
Male	11	0.00052	1.00	
Female	12	0.00085	1.35 (0.60 – 3.06)	0.47
No genetic syndrome	13	0.00049	1.00	
Any genetic syndrome	10	0.0025	5.09 (2.20 – 11.77)	0.0001
≤ 1.0 mSv Radiation*	6	0.00039	1.00	
>1.0 mSv Radiation*	8	0.00056	1.39 (0.48 – 4.102)	0.54
Cumulative radiation (per mSv)*			0.999 (0.985 – 1.014)	0.94

* Among subjects with available radiation data n=4,162

Table 4.7 Age-standardized incidence ratios of pediatric cancer

	Observed cancers (n)	Expected cancers (n)	SIR	95% CI	P value
Overall cohort SIR	23	6.18	3.72	1.53 – 9.04	0.0037
Female	12	2.57	4.67	1.21 – 18.00	0.025
Male	11	3.62	3.04	0.93 – 9.97	0.067
Any genetic syndrome	10	0.80	12.49	1.28 – 121.74	0.030
No genetic syndrome	13	5.39	2.41	0.88 – 6.60	0.086
>1.0 mSv Radiation*	8	2.78	2.88	0.74 – 11.27	0.13
≤ 1.0 mSv Radiation*	6	3.02	1.99	0.50 – 7.94	0.33

* Among subjects with available radiation data n=4,162

Figure 4.1 Incident cancers (n=23) by genetic syndromes Cancers identified in the CHOP CHD cohort are displayed, categorized by tumor site and histology as defined in Table 4.5. Cancers are further categorized by genetic syndrome diagnoses.

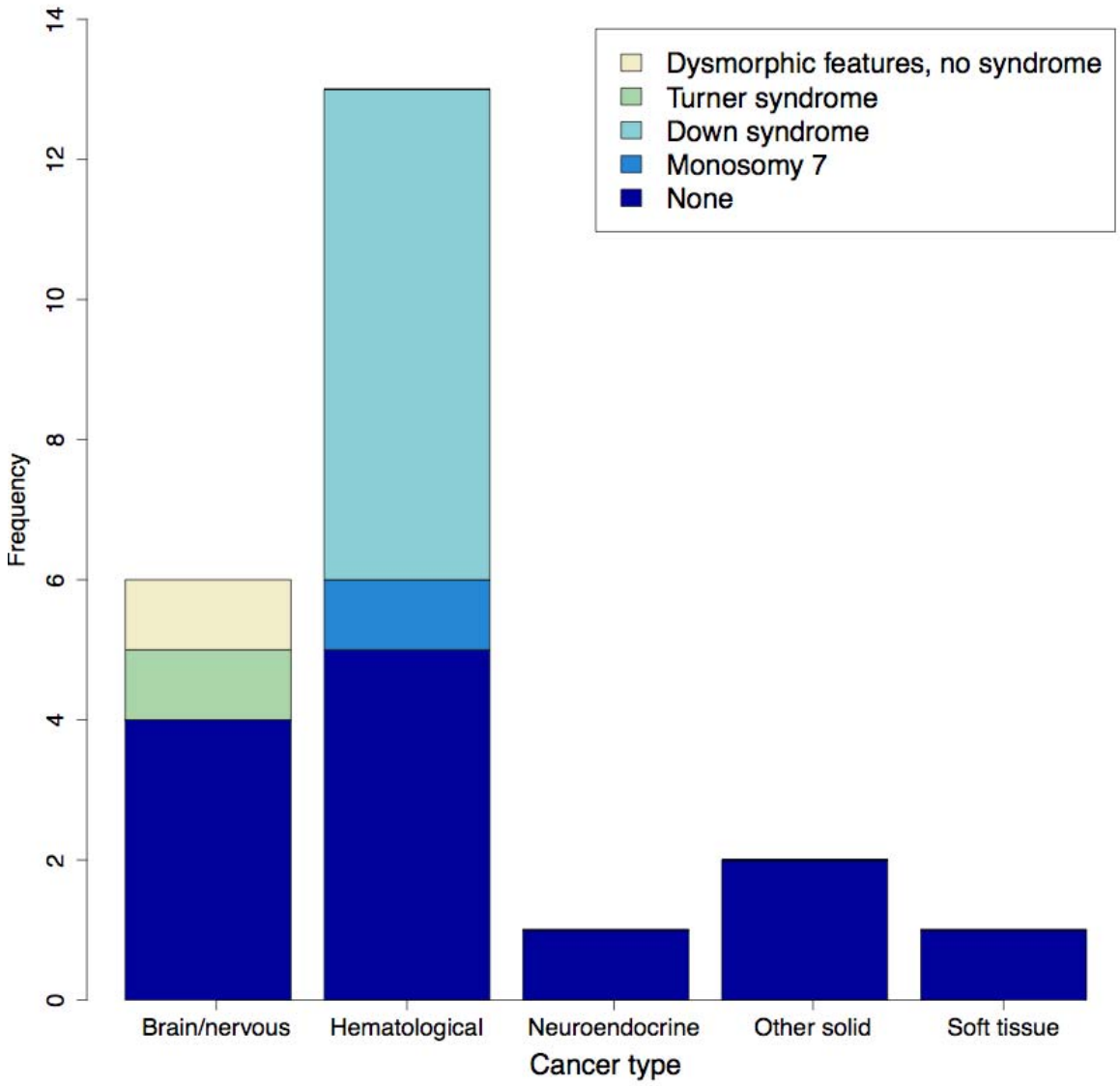
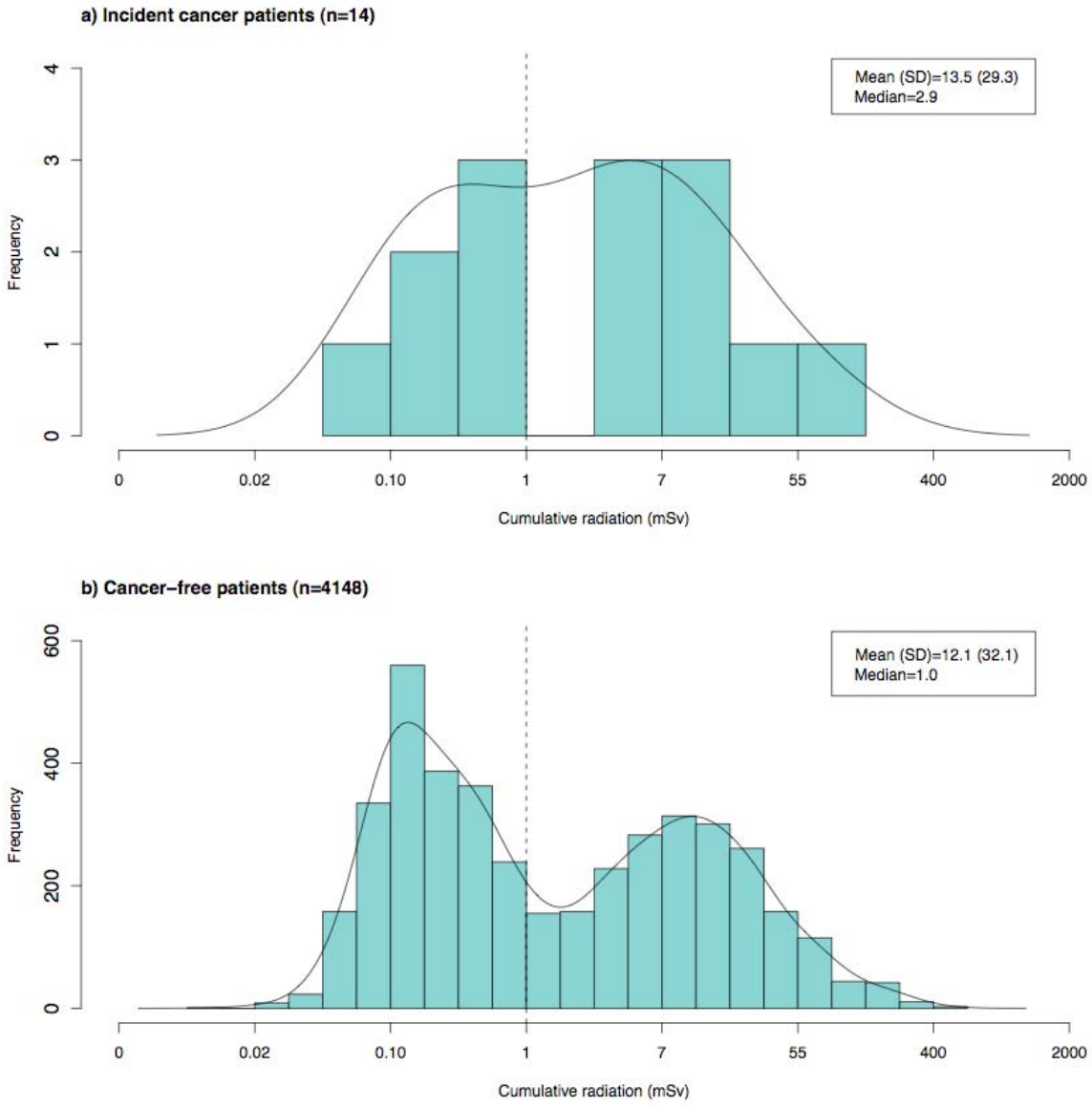


Figure 4.2 Cumulative diagnostic radiation exposure (mSv) by cancer status
Effective radiation exposure from diagnostic exams is shown on a log-scale for 4,162 CHD patients with available data. a) The distribution of radiation among children that developed cancer during follow-up with available data is shown (n=14). Mean = 13.5 mSv, Standard deviation (SD) = 29.3 mSv, Median = 2.9 mSv. b) The distribution of radiation among children that remained cancer-free during follow-up with available data is shown (n=4,148). Mean = 12.1 mSv, SD = 32.1 mSv, Median = 1.0 mSv.



CHAPTER 5

Conclusions

In this dissertation, I have shown that epidemiologic methods provide a way to understand the underlying architecture of complex diseases. We have gained significant understanding of the epidemiology of colorectal cancer, congenital heart disease, and childhood cancer by combining our current understanding of the biology of these diseases with advances in technology, genetic epidemiology, and classic epidemiologic methods. Not only have we gained insight into these diseases individually, but we have also identified a previously unrecognized link between congenital heart disease and cancer, suggesting that future studies can take advantage of this relationship to further understand the common link between the epidemiology of these two diseases.

We investigated the genetic and functional basis of associations at two candidate SNPs identified from the MECC GWAS, rs10210149 on chromosome 2q11.2-q12 and rs16931815 on chromosome 12p11.23. Specifically, MECC subjects were screened for pathogenic mutations and allele-specific expression analyses were performed for *GPR45*, *STK38L*, and *TGFBRAP1*. Of these three genes, *GPR45* was associated with a 27% increase in expression of one allele for each additional copy of the C allele of rs10210149 (p-trend = 0.01). Consistent with the conclusions drawn from other GWAS studies and subsequent functional analyses, we suggest that the underlying causal variant at rs10210149 affects gene expression. Elucidating the functional consequences of these

GWAS variants has been challenging, but provides an opportunity for understanding the mechanisms of colorectal cancer.

We next investigated variation in *ISL1* and risk of CHD in a two-stage case-control study, identifying *ISL1* as a candidate susceptibility gene for human CHD by its integral role in the regulation of the secondary heart field. Eight genic and flanking *ISL1* SNPs were significantly associated with CHD. Our results demonstrate that two different *ISL1* haplotypes contribute to risk of CHD in white (Summary Odds Ratio (OR) =1.27, 95% Confidence Interval (CI) 1.09 – 1.48, $P = 0.0018$) and black/African American populations (Summary OR=1.57, 95% CI 1.07 – 2.30, $P = 0.0216$), suggesting a new role for known regulatory genes of cardiomyocytes in human disease. These data provide strong evidence that congenital heart disease is consistent with the common disease – common variant hypothesis in two different ethnic groups. Further, we provide new insight into the variety of congenital heart disease phenotypes that can be produced from genetic abnormalities in a single source population of cardiac progenitor cells.

Our efforts to understand the epidemiology of both colorectal cancer and congenital heart disease are linked by a well known association between selected forms of congenital heart disease and cancer. In the CHOP cohort study, we showed that children with CHD demonstrated a 3.7-fold increase in the rate of pediatric cancer compared to the US population (Standardized Incidence Ratio (SIR) = 3.72, 95% CI = 1.53 – 9.04, p = 0.0037). Rates were higher for children with both syndromic (SIR=12.49, 95% CI 1.28 – 121.74, p=0.03) and non-syndromic (SIR=2.41, 95% CI 0.88 – 6.60, p=0.086) heart disease. We propose that further studies of cancer incidence among children with CHD may provide a way to unravel the complex biology underlying

this relationship, providing significant insight into the causes of both CHD and pediatric cancers.

In this dissertation, I applied epidemiologic methods to further understand the etiology of three complex diseases: colorectal cancer, childhood cancers, and congenital heart disease (CHD). The characterization of allele-specific expression of *GPR45*, *STK38L*, and *TGFBRAP1* in the MECC study demonstrated the value of studies that combine genetic epidemiology and functional data when evaluating candidate genes identified from genome-wide association studies. Our study of genetic variation in *ISL1* and risk of human congenital heart disease demonstrated a previously unidentified role for common variation in two different ethnic populations. Future studies should investigate the association between risk of CHD and common variation in other genes that are critical to cardiomyocyte regulation and differentiation, which may also be involved in susceptibility to this disease. Finally, evaluating the association between childhood cancers and CHD has demonstrated a link between the biology and epidemiology of both diseases. Understanding the environmental exposures and genetic abnormalities that are found among the CHD patients that developed cancer may provide significant insight into the causes of these two sets of complex diseases.

APPPENDIX

CHD Diagnosis	Definition
AA	Atrial abnormality, other
AI	Aortic insufficiency
AP window	Aortopulmonary window
ASD	Atrial septal defect
ASD, Prim	ASD, primum
ASD, Sec	ASD, secundum
ASD, SV	ASD, Sinus venosus
AVSD	Atrioventricular septal defect
AVSD, Comp	Atrioventricular septal defect, complete
AVSD, Inc	Atrioventricular septal defect, incomplete
Ao aneurysm	Aortic aneurysm
Ao dissection	Aortic dissection
Ao stenosis	Aortic stenosis
Bilateral SVC	Bilateral superior vena cava
CCAVC, unbalanced	Complete common atroventricular canal
CDH	Congenital diagphragmatic hernia
CoA	Coarctation of the aorta
D-TGA	D-transposition of the great arteries
DCM	Dilated cardiomyopathy
DCRV	Double chambered right ventricle
DILV	Double inlet left ventricle
DIRV	Double inlet right ventricle
DOLV	Double outlet left ventricle
DORV	Double outlet right ventricle
HCM	Hypertrophic cardiomyopathy
HLHS	Hypoplastic left heart syndrome
IAA	Interrupted aortic arch
L-TGA	Congenitally corrected transpostion of the great arteries
MR	Mitral regurgitation
MS	Mitral stenosis
MV abnormality	Mitral valve abnormality
MV atresia	Mitral valve atresia
PA	Pulmonary atresia
PA stenosis	Pulmonary artery stenosis
PA/VSD	Pulmonary atresia/ventricular septal defect
PAPVC	Partial anomalous pulmonary venous connection
PDA	Patent ductus arteriosus

CHD Diagnosis	Definition
PHTN	Pulmonary hypertension
PI	Pulmonary insufficiency
PS	Pulmonary stenosis
PV stenosis	Pulmonary vein stenosis
PVD	Pulmonary vascular disease
Pericardial disease, NOS	Pericardial disease, not otherwise specified
RVOTO	Right ventricular outflow tract obstruction
TAPVC	Total anomalous pulmonary venous connection
TOF	Tetralogy of Fallot
TOF/APV	Tetralogy of Fallot, absent pulmonary valve
VSD	Ventricular septal defect
VSD, CS	Ventricular septal defect, conoseptal
VSD, CV	Ventricular septal defect, conoventricular
VSD, I	Ventricular septal defect, inlet
VSD, M	Ventricular septal defect, muscular

References

Agha MM, Williams JI, Marrett L, To T, Zipursky A, Dodds L (2005). Congenital abnormalities and childhood cancer. *Cancer* **103:** 1939-48.

Ahlgren U, Pfaff SL, Jessell TM, Edlund T, Edlund H (1997). Independent requirement for ISL1 in formation of pancreatic mesenchyme and islet cells. *Nature* **385:** 257-60.

Asthagiri AR, Parry DM, Butman JA, Kim HJ, Tsilou ET, Zhuang Z *et al* (2009). Neurofibromatosis type 2. *Lancet* **373:** 1974-86.

Bat L, Pines A, Ron E, Rosenblum Y, Niv Y, Shemesh E (1986). Colorectal adenomatous polyps and carcinoma in Ashkenazi and non-Ashkenazi Jews in Israel. *Cancer* **58:** 1167-71.

Bjorge T, Cnattingius S, Lie RT, Tretli S, Engeland A (2008). Cancer risk in children with birth defects and in their families: a population based cohort study of 5.2 million children from Norway and Sweden. *Cancer Epidemiol Biomarkers Prev* **17:** 500-6.

Black BL (2007). Transcriptional pathways in second heart field development. *Semin Cell Dev Biol* **18:** 67-76.

Bos JL (1989). ras oncogenes in human cancer: a review. *Cancer Res* **49:** 4682-9.

Brenner H, Chang-Claude J, Seiler CM, Sturmer T, Hoffmeister M (2006). Does a negative screening colonoscopy ever need to be repeated? *Gut* **55:** 1145-50.

Broderick P, Carvajal-Carmona L, Pittman AM, Webb E, Howarth K, Rowan A *et al* (2007). A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. *Nat Genet* **39:** 1315-7.

Brunak S, Engelbrecht J, Knudsen S (1991). Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J Mol Biol* **220:** 49-65.

Bu L, Jiang X, Martin-Puig S, Caron L, Zhu S, Shao Y *et al* (2009). Human ISL1 heart progenitors generate diverse multipotent cardiovascular cell lineages. *Nature* **460:** 113-7.

Buckingham M, Meilhac S, Zaffran S (2005). Building the mammalian heart from two sources of myocardial cells. *Nat Rev Genet* **6:** 826-35.

Cai CL, Liang X, Shi Y, Chu PH, Pfaff SL, Chen J *et al* (2003). Isl1 identifies a cardiac progenitor population that proliferates prior to differentiation and contributes a majority of cells to the heart. *Dev Cell* **5:** 877-89.

Carstensen B, Soll-Johanning H, Villadsen E, Sondergaard JO, Lynge E (1996). Familial aggregation of colorectal cancer in the general population. *Int J Cancer* **68:** 428-35.

Castellsague E, Gonzalez S, Guino E, Stevens KN, Borras E, Raymond VM *et al* (2010). Allele-Specific Expression of APC in Adenomatous Polyposis Families. *Gastroenterology*.

Connor JA, Arons RR, Figueroa M, Gebbie KM (2004). Clinical outcomes and secondary diagnoses for infants born with hypoplastic left heart syndrome. *Pediatrics* **114:** e160-5.

Davies H, Bignell GR, Cox C, Stephens P, Edkins S, Clegg S *et al* (2002). Mutations of the BRAF gene in human cancer. *Nature* **417:** 949-54.

de Jong MM, Nolte IM, te Meerman GJ, van der Graaf WT, de Vries EG, Sijmons RH *et al* (2002). Low-penetrance genes and their involvement in colorectal cancer susceptibility. *Cancer Epidemiol Biomarkers Prev* **11:** 1332-52.

Denayer E, de Ravel T, Legius E (2008). Clinical and molecular aspects of RAS related disorders. *J Med Genet* **45:** 695-703.

Ferlay J, Parkin DM, Steliarova-Foucher E (2010). Estimates of cancer incidence and mortality in Europe in 2008. *Eur J Cancer* **46:** 765-81.

Fireman Z, Sandler E, Kopelman Y, Segal A, Sternberg A (2001). Ethnic differences in colorectal cancer among Arab and Jewish neighbors in Israel. *Am J Gastroenterol* **96:** 204-7.

Freeman SB, Taft LF, Dooley KJ, Allran K, Sherman SL, Hassold TJ *et al* (1998). Population-based study of congenital heart defects in Down syndrome. *Am J Med Genet* **80:** 213-7.

Garg V, Kathiriya IS, Barnes R, Schluterman MK, King IN, Butler CA *et al* (2003). GATA4 mutations cause human congenital heart defects and reveal an interaction with TBX5. *Nature* **424:** 443-7.

Goss KH, Groden J (2000). Biology of the adenomatous polyposis coli tumor suppressor. *J Clin Oncol* **18:** 1967-79.

Gripp KW (2005). Tumor predisposition in Costello syndrome. *Am J Med Genet C Semin Med Genet* **137C:** 72-7.

Gross TG, Savoldo B, Punnett A (2010). Posttransplant lymphoproliferative diseases. *Pediatr Clin North Am* **57:** 481-503, table of contents.

Gruber PJ, Epstein JA (2004). Development gone awry: congenital heart disease. *Circ Res* **94:** 273-83.

Gruber SB, Moreno V, Rozek LS, Rennerts HS, Lejbkowicz F, Bonner JD *et al* (2007). Genetic variation in 8q24 associated with risk of colorectal cancer. *Cancer Biol Ther* **6:** 1143-7.

Hammer GP, Seidenbusch MC, Schneider K, Regulla DF, Zeeb H, Spix C *et al* (2009). A cohort study of childhood cancer incidence after postnatal diagnostic X-ray exposure. *Radiat Res* **171:** 504-12.

Hebsgaard SM, Korning PG, Tolstrup N, Engelbrecht J, Rouze P, Brunak S (1996). Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information. *Nucleic Acids Res* **24:** 3439-52.

Hergovich A, Stegert MR, Schmitz D, Hemmings BA (2006). NDR kinases regulate essential cell processes from yeast to humans. *Nat Rev Mol Cell Biol* **7:** 253-64.

Hickey EJ, Veldtman G, Bradley TJ, Gengsakul A, Manlhiot C, Williams WG *et al* (2009). Late risk of outcomes for adults with repaired tetralogy of Fallot from an inception cohort spanning four decades. *Eur J Cardiothorac Surg* **35:** 156-64; discussion 164.

Hoffman JI, Kaplan S (2002). The incidence of congenital heart disease. *J Am Coll Cardiol* **39:** 1890-900.

Hoffman JI, Kaplan S, Liberthson RR (2004). Prevalence of congenital heart disease. *Am Heart J* **147:** 425-39.

Houlston RS, Webb E, Broderick P, Pittman AM, Di Bernardo MC, Lubbe S *et al* (2008). Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat Genet* **40:** 1426-35.

Huls G, Koornstra JJ, Kleibeuker JH (2003). Non-steroidal anti-inflammatory drugs and molecular carcinogenesis of colorectal carcinomas. *Lancet* **362:** 230-2.

Jaeger E, Webb E, Howarth K, Carvajal-Carmona L, Rowan A, Broderick P *et al* (2008). Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk. *Nat Genet* **40:** 26-8.

Jenkins KJ, Correa A, Feinstein JA, Botto L, Britt AE, Daniels SR *et al* (2007). 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* **115:** 2995-3014.

Kemp Z, Thirlwell C, Sieber O, Silver A, Tomlinson I (2004). An update on the genetics of colorectal cancer. *Hum Mol Genet* **13 Spec No 2:** R177-85.

Knight JC (2005). Regulatory polymorphisms underlying complex disease traits. *J Mol Med* **83:** 97-109.

Kusakai G, Suzuki A, Ogura T, Miyamoto S, Ochiai A, Kaminishi M *et al* (2004). ARK5 expression in colorectal cancer and its implications for tumor progression. *Am J Pathol* **164:** 987-95.

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

Larsson SC, Wolk A (2006). Meat consumption and risk of colorectal cancer: a meta-analysis of prospective studies. *Int J Cancer* **119:** 2657-64.

Laugwitz KL, Moretti A, Lam J, Gruber P, Chen Y, Woodard S *et al* (2005). Postnatal isl1+ cardioblasts enter fully differentiated cardiomyocyte lineages. *Nature* **433:** 647-53.

Li Y AG (2006). Mach 1.0: Rapid Haplotype Reconstruction and Missing Genotype Inference. *Am J Hum Genet*.

Limsui D, Vierkant RA, Tillmans LS, Wang AH, Weisenberger DJ, Laird PW *et al* (2010). Cigarette smoking and colorectal cancer risk by molecularly defined subtypes. *J Natl Cancer Inst* **102:** 1012-22.

Lo HS, Wang Z, Hu Y, Yang HH, Gere S, Buetow KH *et al* (2003). Allelic variation in gene expression is common in the human genome. *Genome Res* **13:** 1855-62.

Lynch HT, Weisenburger DD, Quinn-Laquer B, Watson P, Lynch JF, Sanger WG (2002). Hereditary chronic lymphocytic leukemia: an extended family study and literature review. *Am J Med Genet* **115:** 113-7.

Marchese A, Sawzdargo M, Nguyen T, Cheng R, Heng HH, Nowak T *et al* (1999). Discovery of three novel orphan G-protein-coupled receptors. *Genomics* **56:** 12-21.

Marra G, Boland CR (1995). Hereditary nonpolyposis colorectal cancer: the syndrome, the genes, and historical perspectives. *J Natl Cancer Inst* **87:** 1114-25.

Modan B, Keinan L, Blumstein T, Sadetzki S (2000). Cancer following cardiac catheterization in childhood. *Int J Epidemiol* **29:** 424-8.

Moreno V, Gemignani F, Landi S, Gioia-Patricola L, Chabrier A, Blanco I *et al* (2006). Polymorphisms in genes of nucleotide and base excision repair: risk and prognosis of colorectal cancer. *Clin Cancer Res* **12:** 2101-8.

Moretti A, Caron L, Nakano A, Lam JT, Bernshausen A, Chen Y *et al* (2006). Multipotent embryonic isl1+ progenitor cells lead to cardiac, smooth muscle, and endothelial cell diversification. *Cell* **127:** 1151-65.

Narod SA, Hawkins MM, Robertson CM, Stiller CA (1997). Congenital anomalies and childhood cancer in Great Britain. *Am J Hum Genet* **60:** 474-85.

Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, Oksenberg JR *et al* (2004). Methods for high-density admixture mapping of disease genes. *Am J Hum Genet* **74:** 979-1000.

Pertea M, Lin X, Salzberg SL (2001). GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res* **29:** 1185-90.

Piard F, Martin L, Chapusot C, Ponnelle T, Faivre J (2002). [Genetic pathways in colorectal cancer: interest for the pathologist]. *Ann Pathol* **22:** 277-88.

Pierpont ME, Basson CT, Benson DW, Jr., Gelb BD, Giglia TM, Goldmuntz E *et al* (2007). 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* **115:** 3015-38.

Pomerantz MM, Ahmadiyeh N, Jia L, Herman P, Verzi MP, Doddapaneni H *et al* (2009). The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat Genet* **41:** 882-4.

Posch MG, Perrot A, Schmitt K, Mittelhaus S, Esenwein EM, Stiller B *et al* (2008). Mutations in GATA4, NKX2.5, CRELD1, and BMP4 are infrequently found in patients with congenital cardiac septal defects. *Am J Med Genet A* **146A:** 251-3.

Rankin J, Silf KA, Pearce MS, Parker L, Ward Platt M (2008). Congenital anomaly and childhood cancer: A population-based, record linkage study. *Pediatr Blood Cancer* **51:** 608-12.

Reich DE, Lander ES (2001). On the allelic spectrum of human disease. *Trends Genet* **17:** 502-10.

Roizen NJ, Patterson D (2003). Down's syndrome. *Lancet* **361:** 1281-9.

Rose V, Gold RJ, Lindsay G, Allen M (1985). A possible increase in the incidence of congenital heart defects among the offspring of affected parents. *J Am Coll Cardiol* **6:** 376-82.

Sadetzki S, Chetrit A, Freedman L, Stovall M, Modan B, Novikov I (2005). Long-term follow-up for brain tumor development after childhood exposure to ionizing radiation for tinea capitis. *Radiat Res* **163:** 424-32.

Scheet P, Stephens M (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* **78:** 629-44.

Schnater JM, Kohler SE, Lamers WH, von Schweinitz D, Aronson DC (2003). Where do we stand with hepatoblastoma? A review. *Cancer* **98:** 668-78.

Schoemaker MJ, Swerdlow AJ, Higgins CD, Wright AF, Jacobs PA (2008). Cancer incidence in women with Turner syndrome in Great Britain: a national cohort study. *Lancet Oncol* **9:** 239-46.

Schott JJ, Benson DW, Basson CT, Pease W, Silberbach GM, Moak JP *et al* (1998). Congenital heart disease caused by mutations in the transcription factor NKX2-5. *Science* **281:** 108-11.

Schrader KA, Nelson TN, De Luca A, Huntsman DG, McGillivray BC (2009). Multiple granular cell tumors are an associated feature of LEOPARD syndrome caused by mutation in PTPN11. *Clin Genet* **75:** 185-9.

Schultz AH, Wernovsky G (2005). Late outcomes in patients with surgically treated congenital heart disease. *Semin Thorac Cardiovasc Surg Pediatr Card Surg Annu***:** 145-56.

Sijmons RH, Hofstra RM, Wijburg FA, Links TP, Zwierstra RP, Vermey A *et al* (1998). Oncological implications of RET gene mutations in Hirschsprung's disease. *Gut* **43:** 542-7.

Smith MW, Patterson N, Lautenberger JA, Truelove AL, McDonald GJ, Waliszewska A *et al* (2004). A high-density admixture map for disease gene discovery in african americans. *Am J Hum Genet* **74:** 1001-13.

Sorensen SA, Mulvihill JJ, Nielsen A (1986). Long-term follow-up of von Recklinghausen neurofibromatosis. Survival and malignant neoplasms. *N Engl J Med* **314:** 1010-5.

Sotelo J, Esposito D, Duhagon MA, Banfield K, Mehalko J, Liao H *et al* (2010). Long-range enhancers on 8q24 regulate c-Myc. *Proc Natl Acad Sci U S A* **107:** 3001-5.

Stiller CA (2004). Epidemiology and genetics of childhood cancer. *Oncogene* **23:** 6429-44.

Sun Y, Liang X, Najafi N, Cass M, Lin L, Cai CL *et al* (2007). Islet 1 is expressed in distinct cardiovascular lineages, including pacemaker and coronary vascular cells. *Dev Biol* **304:** 286-96.

Suzuki A, Kusakai G, Kishimoto A, Lu J, Ogura T, Esumi H (2003). ARK5 suppresses the cell death induced by nutrient starvation and death receptors via inhibition of caspase 8 activation, but not by chemotherapeutic agents or UV irradiation. *Oncogene* **22:** 6177-82.

Suzuki A, Ogura T, Esumi H (2006). NDR2 acts as the upstream kinase of ARK5 during insulin-like growth factor-1 signaling. *J Biol Chem* **281:** 13915-21.

Suzuki M, Igarashi R, Sekiya M, Utsugi T, Morishita S, Yukawa M *et al* (2004). Dynactin is involved in a checkpoint to monitor cell wall synthesis in Saccharomyces cerevisiae. *Nat Cell Biol* **6:** 861-71.

Tartaglia M, Gelb BD (2005). Germ-line and somatic PTPN11 mutations in human disease. *Eur J Med Genet* **48:** 81-96.

Tenesa A, Dunlop MG (2009). New insights into the aetiology of colorectal cancer from genome-wide association studies. *Nat Rev Genet* **10:** 353-8.

Tenesa A, Farrington SM, Prendergast JG, Porteous ME, Walker M, Haq N *et al* (2008). Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat Genet* **40:** 631-7.

Thangaraju M, Cresci GA, Liu K, Ananth S, Gnanaprakasam JP, Browning DD *et al* (2009). GPR109A is a G-protein-coupled receptor for the bacterial fermentation product butyrate and functions as a tumor suppressor in colon. *Cancer Res* **69:** 2826-32.

Tomlinson I, Webb E, Carvajal-Carmona L, Broderick P, Kemp Z, Spain S *et al* (2007). A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat Genet* **39:** 984-8.

Tomlinson IP, Webb E, Carvajal-Carmona L, Broderick P, Howarth K, Pittman AM *et al* (2008). A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nat Genet* **40:** 623-30.

Tuupanen S, Turunen M, Lehtonen R, Hallikas O, Vanharanta S, Kivioja T *et al* (2009). The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nat Genet* **41:** 885-90.

Vassilatis DK, Hohmann JG, Zeng H, Li F, Ranchalis JE, Mortrud MT *et al* (2003). The G protein-coupled receptor repertoires of human and mouse. *Proc Natl Acad Sci U S A* **100:** 4903-8.

Verheugt CL, Uiterwaal CS, Grobbee DE, Mulder BJ (2008). Long-term prognosis of congenital heart defects: a systematic review. *Int J Cardiol* **131:** 25-32.

Wang WY, Barratt BJ, Clayton DG, Todd JA (2005). Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet* **6:** 109-18.

Whittemore R, Hobbins JC, Engle MA (1982). Pregnancy and its outcome in women with and without surgical treatment of congenital heart disease. *Am J Cardiol* **50:** 641-51.

Whittemore R, Wells JA, Castellsague X (1994). A second-generation study of 427 probands with congenital heart defects and their 837 children. *J Am Coll Cardiol* **23:** 1459-67.

WHO (2003). World Cancer Report (eds Stewart B. W. & Kleihues P.). **13**.

Wolin KY, Yan Y, Colditz GA, Lee IM (2009). Physical activity and colon cancer prevention: a meta-analysis. *Br J Cancer* **100:** 611-6.

Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ *et al* (2007). The genomic landscapes of human breast and colorectal cancers. *Science* **318:** 1108-13.

Wu SM, Chien KR, Mummery C (2008). Origins and fates of cardiovascular progenitor cells. *Cell* **132:** 537-43.

Wurthner JU, Frank DB, Felici A, Green HM, Cao Z, Schneider MD *et al* (2001). Transforming growth factor-beta receptor-associated protein 1 is a Smad4 chaperone. *J Biol Chem* **276:** 19495-502.

Yan H, Dobbie Z, Gruber SB, Markowitz S, Romans K, Giardiello FM *et al* (2002). Small changes in expression affect predisposition to tumorigenesis. *Nat Genet* **30:** 25-6.

Zanke BW, Greenwood CM, Rangrej J, Kustra R, Tenesa A, Farrington SM *et al* (2007). Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat Genet* **39:** 989-94.