# Candidate-Pathway Gene Environment Interactions on Colon and Rectal Cancer Risk and Survival: Methodological Frameworks for Interaction in Genetic Association Studies

by

Noha Sharafeldin

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Public Health

Department of Public Health Sciences
University of Alberta

**ABSTRACT**

Genetic association studies have adopted for a long time a traditional analytic approach that focuses on individual genetic markers, usually single nucleotide polymorphisms (SNPs), in association with disease or phenotype. A standard single-SNP analysis that ignores combined effects of multiple SNPs and furthermore their interactions with environmental exposures, explains a small portion of disease heritability: an often cited issue of 'missing heritability'. A comprehensive approach that accounts for these interactions carries the potential for identifying novel susceptibility loci and is more suited to decipher causal relationships and underlying molecular mechanisms of disease. The overall goal of this dissertation is to develop a methodologically sound framework that examines interactions in genetic association studies that is able to represent the biologic underpinnings of disease and yield interpretations that are statistically valid and of clinical and/or public health relevance.

We first examined interactions between genetic variants at the gene level in genome-wide association study (GWAS) data of six common chronic diseases of the Wellcome-Trust-Case-Control-Consortium (WTCCC): bipolar disorder (BD); coronary artery disease (CAD); hypertension (HT); rheumatoid arthritis (RA); type 2 diabetes (T2D); and type 1 diabetes (T1D). We used logic regression to search for biologically plausible forms of SNP-set interactions within genes. Next, we extended our approach to test for gene-environment interaction (GEI) effects at the pathway level and applied it to the population-based case-control data of the Diet, Activity and Lifestyle as a Risk Factor for Colorectal Cancer Study. We focused on the candidate pathway of angiogenesis and three hypothesized environmental exposures: dietary protein intake; smoking; and alcohol consumption. Our approach consisted of 3-steps: the first

two summarized the within gene effects and the full pathway effects; and the third step modelled the GEI effects on colon and rectal cancer risk and survival.

Our interaction analysis was able to detect an appreciable number of susceptibility loci showing strong evidence of association with the six diseases in WTCCC, including novel signals supported by biologically plausible links to the diseases. The number of genes with strong evidence of association was: 13 for BD; 16 for CAD; 15 for HT; 72 for RA; 105 for T1D; and 19 for T2D. The top significant genes were: *NFIA* with BD, *CDKN2B* with CAD, *COL4A4* with HT, *BTNL2* with RA, and *TCF7L2* with T2D. The majority of strong single-SNP signals of WTCCC and on average 46% of recent GWAS meta-analyses signals were confirmed in our analysis. The results of the GEI pathway analysis also yielded an appreciable number of significant and novel interactions. Overall the magnitudes of gene interaction odds and hazard ratios increased with increasing levels of the interacting environmental exposure. This observed positive gradient supported the plausibility of the interactions. We found five statistically significant GEIs associated with colon cancer risk and three GEIs with colon cancer survival involving all three environmental exposures. For rectal cancer, we found eight significant GEIs in association with risk involving six genes and five GEIs with survival.

This dissertation showed how exploring interactions of all measured SNPs within each gene can identify appreciable numbers of novel susceptibility loci in GWAS. We also showed that GEI effects on colorectal cancer risk and survival can be identified by adopting a comprehensive candidate pathway approach that emphasizes the biologic hypothesis in the selection of the pathway genes and environmental exposures and carries that logic through the analysis.

## PREFACE

This thesis is an original work by Dr. Noha Sharafeldin with supervision of Prof. Yutaka Yasui. The research project, of which this thesis is a part, received research ethics approval from the University of Alberta Health Research Ethics Board (HREB), Project Name "Gene Pathway – Environment Interaction in Colorectal Cancer", No. Pro00026736, November 18, 2011.

The data analysis in Chapters 2, 3, and 4 is my original work with Prof. Yasui and the technical assistance of Qi Liu and Conrado Franco-Villalobos. The literature review in Chapter 2 and the concluding Chapter 5 is my original work. Data used in Chapter 2 are publically available through the "Wellcome-Trust-Case-Control-Consortium". Data used in Chapters 3 and 4 are available from the "Diet, Activity and Lifestyle as a Risk Factor for Colorectal Cancer Study" led by Dr. Martha L. Slattery, Professor at Division of Epidemiology, University of Utah. The study was funded by the US National Institutes of Health (NIH). (http://projectreporter.nih.gov/project_info_description.cfm?aid=7489405&icde=3371432)

Chapter 2 of this thesis has been submitted as Sharaf Eldin N., Liu Q., Jabbari S., Wang L., Franco-Villalobos C., Mahasirimongkol S., Yanai H., Tokunaga K., Yasui Y. Within-Gene Interactions in GWAS Identifies Novel Susceptibility Loci – The WTCCC Data Revisited. I was responsible for providing technical assistance to the analysis, interpretation and presentation of the results as well as the manuscript drafting and final composition. Qi Liu, Shahab Jabbari and Conrado Franco-Villalobos programmed and conducted the analysis. Katsushi Tokunaga, Surakameth Mahasirimongkol, and Hideki Yanai provided critical input to interpretations and reporting of the analysis results. Prof. Yasui was the senior author responsible for the research

project including the concept formation, analysis design, and supervision of the research conduct of Dr. Sharafeldin and his team members.

Chapter 3 of this thesis will be submitted as Sharaf Eldin N., Slattery M.L., Liu Q., Franco-Villalobos C., Caan B.J., Potter J.D., Yasui Y. A Candidate Pathway Approach Identifies Multiple Gene-Environment Interactions in Association with Colon Cancer Risk and Survival.

Chapter 4 of this thesis will be submitted as Sharaf Eldin N., Slattery M.L., Liu Q., Franco-Villalobos C., Caan B.J., Yasui Y. Multiple Gene-Environment Interactions on the Angiogenesis Gene-Pathway Impact Rectal Cancer Risk and Survival.

For both papers of Chapters 3 and 4, I was responsible for the development of the study hypothesis, conducting the analysis, interpretation and presentation of the results as well as the manuscript drafting and final composition. Qi Liu and Conrado Franco-Villalobos assisted in programming and conducting the analysis. Bette J. Caan and John D. Potter provided critical input to interpretations and reporting of the results. Prof. Martha L. Slattery provided the study data and together with Prof. Yasui were supervisory authors and involved in the concept formation and manuscript composition: Prof. Yasui was also responsible for analysis design and supervision of research conduct by Dr. Sharafeldin and other members of his team.

**DEDICATION**

*To my dear parents.*

*To Mohamed, Youssef, and Adam.*

**ACKNOWLEDGEMENTS**

I wish to thank the many people who, directly or indirectly, made this thesis possible.

First of all, I would like to express my endless gratitude and respect to my supervisor, Prof. Yutaka Yasui. What I have experienced and learned from him over the years is invaluable beyond what words can describe. He has guided and supported me throughout my PhD program and beyond. His inspiring research ideas, outstanding command of research tools, passion and dedication were a continuous source of motivation and I appreciate all his constructive feedback and direction during the conduct of this thesis. Prof. Yasui is an exceptional researcher and person who simply encompasses the essence of what a true scientist and mentor ought to be. It is a privilege to work closely with a researcher of his unique caliber and who has definitely helped shape my career path.

I would like to sincerely thank Dr. Martha Slattery, whose insights, advice, and encouragement from the initial steps to the final product were seminal to this work. She has made her support available in many ways and provided insightful suggestions and recommendations during the development and writing of the manuscripts. I also wish to thank Dr. Irina Dinu for her support and assistance in the completion of this work and for the opportunity of working closely with her.

I would also like to thank Dr. Marcy Winget who provided me with valuable research experience and opportunity to contribute to her research projects. My thanks to all past and current members of Prof. Yasui's research team with whom I have shared knowledge, experience and unforgettable memories over the years; with special thanks to Isac Lima, Leah Martin, Qi Liu, Xuan Wu and Conrado Franco-Villalobos. Many thanks go to my professors and colleagues at

the School of Public Health. They have been an endless source of professional and moral support and with whom I have built relationships that I will never cease to value. Special thanks to Dr. Duncan Saunders, Dr. Doug Wilson, and Dr. Ian Colman for their continuous support.

My utmost gratitude is to my loving parents, Nahed and Mohamed, for their unconditional love, support, and encouragement. They are my pillars in life and now that I became a parent myself, I came to a full understanding of their sacrifices and lifelong commitment. I would also like to thank my brother and first friend in life, Eehab, for always being there for me, sharing my ups and downs, and putting up with all the craziness only he can understand. I cannot thank enough my husband, Mohamed, for believing in me and standing by me through countless challenges. Without his understanding, encouragement and loving support, I could not have managed to complete this work. He is my companion, my confidant and soul mate. I also thank my precious sons, Youssef and Adam, whose mere presence has added meaning and reason to every step in life. *This work is for and because of you in hopes of inspiring you on your own amazing journeys in life.*

I would like to finally acknowledge the research funding and stipend support I received from a number of organizations during the course of this work including the Egyptian Government Scholarship Award; the Alberta Innovates - Health Solutions; the Alberta Cancer Foundation; and the University of Alberta Dissertation Award.

**TABLE OF CONTENTS**

**LIST OF TABLES**

**LIST OF FIGURES**

**LIST OF ABBREVIATIONS**

| | |
|---|---|
| BD | Bipolar Disorder |
| CAD | Coronary Artery Disease |
| CRC | Colorectal Cancer |
| eQTL | expression Quantitative Trait Locus |
| GEI | Gene Environment Interaction |
| GWAS | Genome Wide Association Studies |
| HR | Hazard Ratio |
| HT | Hypertension |
| IQR | Inter-Quartile Range |
| KPMCP | Kaiser Permanente Medical Care Program of Northern California |
| LD | Linkage Disequilibrium |
| MAF | Minor Allele Frequency |
| MSI | Microsatelitte Instability |
| OR | Odds Ratio |
| RA | Rheumatoid Arthritis |
| SEER | Surveillance Epidemiology and End Results |
| SNP | Single Nucleotide Polym orphism |
| T1D | Type 1 Diabetes |
| T2D | Type 2 Diabetes |
| WTCCC | Wellcome Trust Case Control Consortium |

# CHAPTER 1

# INTRODUCTION

## 1.1 Genetic association studies

Genetic association studies aim to detect associations between genotypes and a disease or trait, as well as their joint effects with social, behavioral, and environmental exposures (Cordell et al. 2005). The two approaches towards gene association studies are candidate gene or pathway studies and genome-wide association studies (GWAS). GWAS became popular following the remarkable completion of the Human Genome Project in 2003, and identification of an estimated 10 million single nucleotide polymorphisms (SNPs) transmitted across generations in blocks allowing the majority of variation within each block to be captured by 'tag' SNPs based on the linkage disequilibrium (LD) phenomenon. In GWAS, the entire human genome is explored through examination of millions of common SNPs aiming at identification of those associated with disease or phenotype (Pearson2009). A GWAS employs an agnostic data-driven approach where prior knowledge of SNP function is not required. In fact, SNPs identified through GWAS are unlikely to be the functional variants themselves and rather serve as markers for an underlying haplotype containing the functional variant (Manolio2010). Feasibility of GWAS grew with the rapid advances in genotyping technologies and is expected to increase with the emergence of next generation sequencing allowing for whole-exome and whole-genome sequencing, all coupled with a steady decline in cost (Stranger et al. 2011).

In contrast, candidate gene studies are hypothesis-driven (Rebbeck et al. 2004), which carries the potential of elucidating underlying biological mechanisms of disease. A comprehensive approach to candidate gene studies would be a focus on groups of genes in a biological pathway critical to

the development and outcome of disease. This carries the advantage of reducing the dimensionality of the search while integrating the biological hypothesis.

## 1.2 Interactions in genetic studies

Genes are inherently coordinated and when searching for genetic associations with disease, it is imperative to account for this inherent coordination among genes (Franklin et al. 1970). Traditionally in genetic association studies (GWAS or candidate gene) the individual SNPs are examined one at a time in association with a phenotype or disease. This approach has potential drawbacks: it ignores potentially larger effects of multiple functional SNPs in several genes in determining disease susceptibility; and the independent marginal effects of single-locus SNPs may overlook the effects of interacting loci whose contribution to disease susceptibility is captured only in combination with other loci. Furthermore, reported measures of association (most commonly odds ratios for case-control studies or risk estimates for cohort studies) are often of minimal clinical/ public health significance at the population level (the ORs are small despite their high statistical significance) and their reproducibility is limited. The median OR reported from GWAS is 1.28 (interquartile range (IQR) =1.17 to 1.55 for binary traits) (Witte 2010). Common alleles, however, based on their prevalence are expected to explain a larger proportion of the population attributable risk compared to the rare, high risk alleles. This is supported by the "common disease–common variant" hypothesis, which states that common diseases are a result of common genetic variants with appreciable frequency in the population at large (Reich et al. 2001). GWAS will thus have good statistical power to detect genetic variants with small to modest effects as long as they are common. This has formed the basis for GWAS and eventually led genotyping arrays to primarily measure common SNPs (e.g., MAF >5%) reducing the ability of GWAS to evaluate rare SNPs regardless of their effect size (**Figure 1.1**).

Common diseases, however, are undoubtedly also due to rare variants. Limited detection of rare SNPs and the detected small effect sizes, therefore, only explain a limited amount of the total inherited risk of disease. This has been referred to as the *missing* heritability (Goldstein2009, Manolio et al. 2009). This unexplained heritability could result from GWAS typically testing for only the marginal effects of individual SNPs, while gene-gene and gene-environment interactions still remain largely unexplored (Manolio et al. 2009, Cordell2009a).

**Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (odds ratio).**

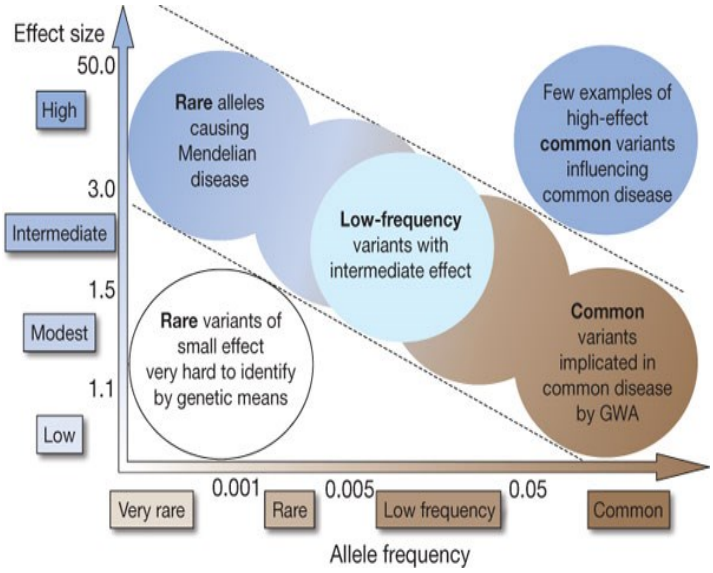



**Figure 1.1:** Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (odds ratio). Most emphasis and interest lies in identifying associations with characteristics shown within diagonal dotted lines. (Extracted from reference (Manolio et al. 2009))

### 1.2.1 Gene-Gene and Gene-Environment Interactions

An analysis approach that aims to detect only significant marginal effects of an individual SNP on a disease would be successful if that individual SNP's function is in some way biologically critical to acquire the disease. 'Epistasis' is the term used in population genetics to describe

modification of the effects of one gene by one or several other genes (gene –gene interactions) (Cordell2002). In an epistatic model, the joint effect of, for example, two SNPs will be inadequately captured by the sum of the modest effects anticipated for each SNP independently (Culverhouse et al. 2002). In fact, interactions of variants with opposite effects in the two different exposure groups (i.e. a crossing interaction) will not show a main effect and therefore will not be identified using standard approaches (Murcray et al. 2009).

In complex diseases the relationship between the phenotype and genotype is argued to fundamentally depend on the interaction between disease susceptibility loci and gene-environment interactions. That is to say the SNPs and a specific environmental exposure are working jointly to influence disease risk. Assuming genes and environment are interacting together in influencing disease is sensible based on a wide range of environmental and cultural diversity within and among human groups. Failure to incorporate genetic and environmental factors in a joint analysis potentially weakens the observed association because pools of susceptible and non-susceptible individuals are mixed together and the observed association between a true risk factor and disease occurrence tends to be shifted to the null (Khoury et al. 2009).

Studies of gene-gene and/or gene- environment interactions are relevant for several reasons. In addition to potential contribution to missing heritability, discoveries can generate new hypotheses for future replication and functional studies. Gene-environment interactions can identify environmental exposures that affect only a subpopulation of genetically susceptible individuals, which may explain failure of replication of earlier studies and the heterogeneity of main effects results across studies through showing the differences in environmental exposure distributions. Furthermore, statistical models of joint effects can be useful for individual

prediction of disease risk, prognosis, or modification of lifestyle or environmental factors that could change an individual's risk. Therefore, the discovery of relationships among genes and between genes and environmental factors has important implications for both public health and personalized medicine in targeting therapeutic and prevention strategies (Thomas2010a).

### 1.2.2 Pathway analysis

Coordination between genes can be described with a pathway structure where a pathway is composed of multiple genes with coordinated biological functions. A pathway-based analysis acknowledges the complexity of disease by accounting for multiple loci simultaneously, including gene-gene and gene-environment interactions, and relating them to biologic functions (Kraft et al. 2009). Recently, more attention is focused on several candidate genes in a single pathway perceived as more critical, or conversely depending on purely data-driven approaches from genome-wide scans to elicit important pathways. Some methods for pathway analyses have been proposed for both approaches (Thomas2010b). Several methods rely on searches of *pairwise* (two-way) interactions, including exhaustive searches (Marchini et al. 2005); Bayesian model selection (Zhang et al. 2007); and two-step analysis approaches (Wu et al. 2010,Tao et al. 2012). Another example includes modifications to pathway approaches originally developed for gene expression data such as gene set enrichment analysis (Wang et al. 2007). Nevertheless, the degree to which statistical modeling can elucidate the underlying biological mechanisms is likely to be limited. Searching for higher order interactions and moreover involving hypothesis-driven pathway-based approaches is more likely to elucidate the underlying biological mechanism of disease.

In the first study of this thesis we present an analysis that examined biologically plausible forms of SNP-set interaction effects within-genes using GWAS data for six chronic diseases. In the subsequent two studies we went a step further by considering a systematic biologically-based pathway approach to examine gene-environment interaction effects on colon and rectal cancer risk and survival. A critical biologic pathway in CRC carcinogenesis is the essential process of formation of new blood vessels from preexisting ones, known as angiogenesis. We focused on three exposures that are potentially enhancing the process of angiogenesis: dietary protein, alcohol intake and smoking.

## 1.3 Colorectal cancer

### 1.3.1 Epidemiology of Colorectal Cancer

Colorectal cancer (CRC) is one of the multi-factorial diseases where molecular approaches can help in the understanding of its complex etiology and the underlying biologic mechanisms. It is a pressing public health problem estimated worldwide as the second most prevalent cancer, with over 1 million new cases diagnosed annually (International Agency for Research on Cancer2008). The highest incidence rates are found in the western world mainly New Zealand, Australia, North America, and Europe and most recently Japan (**Figure 1.2**). The worldwide economic burden is large; and research providing a better understanding of the multi-factorial nature of the disease is crucial. Worldwide mortality attributable to CRC is approximately half that of the incidence and the five-year survival following detection and treatment is around 50% (International Agency for Research on Cancer2008).

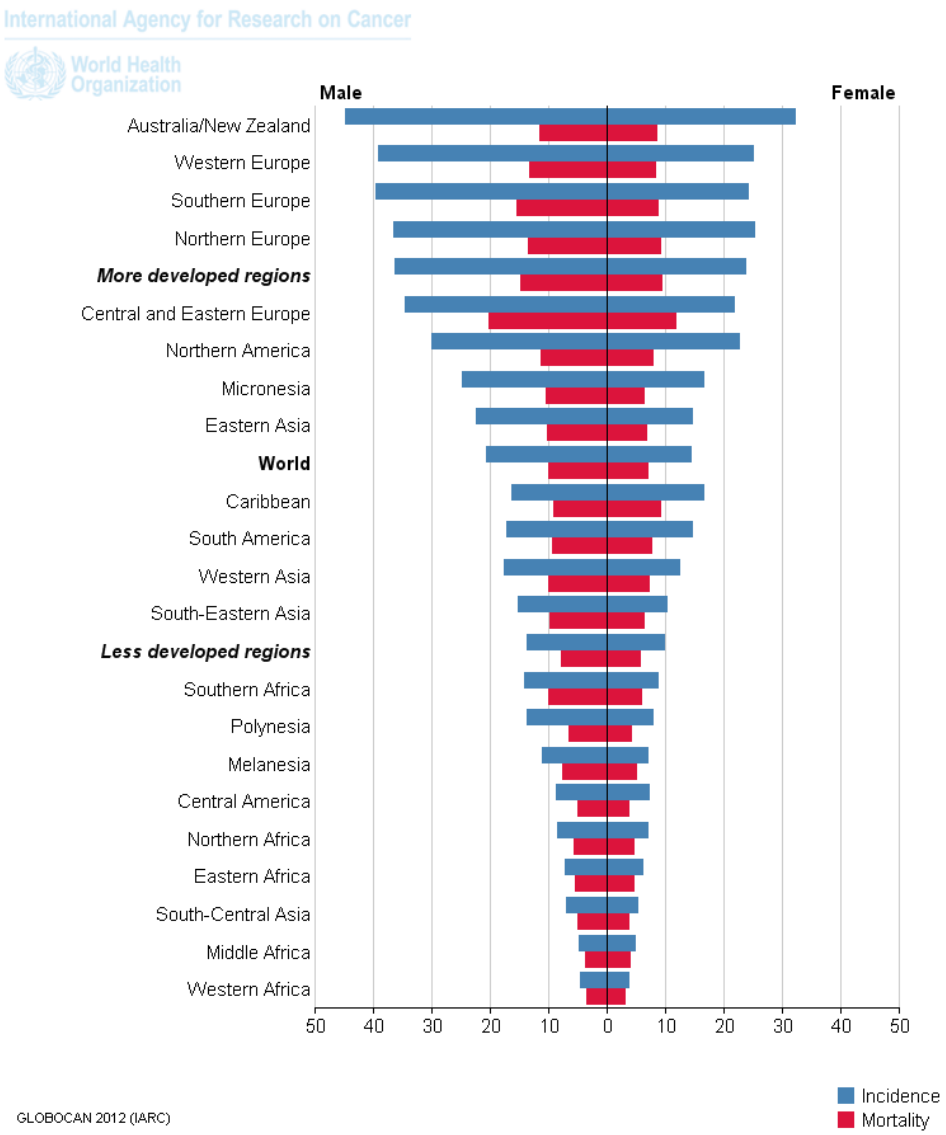



**Figure 1.2:** Estimated age-standardized colorectal cancer worldwide rates per 100,000.

Source: GLOBOCAN 2012 Colorectal Cancer Estimated Incidence, Mortality and Prevalence Worldwide in 2012, International Agency for Research on Cancer (IARC), Lyon (France), 2012. Available from http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx

### 1.3.2 Combined effects of genes and environmental exposures on colorectal cancer risk and survival

Epidemiological and molecular evidence indicate CRC is related to both genes and environment.

Majority of CRC tumors occur sporadically. An adenoma-carcinoma sequence was proposed as a

multi-hit theory involving several genetic mutations or multiple gene activating or inactivating events (e.g. mutation or loss of APC gene, mutation of KRAS gene, loss of tumor suppressor gene p53). Inherited susceptibility is present in individuals with family or personal history of CRC or adenomatous polyps; hereditary syndromes (familial adenomatous polyposis and hereditary nonpolyposis colorectal cancer); and other high risk conditions (e.g. inflammatory bowel disease and Crohn's disease).

Specific environmental exposures have been identified in the etiology pf CRC. The prominent role of environmental exposures in CRC etiology is suggested through the marked geographical variations across countries (International Agency for Research on Cancer2008) and data showing that migrant populations moving from low-risk to high risk countries adopt the disease rates of the host country. The most important factor is diet. Evidence suggests diet low in fiber, fruit and vegetables, and high in calories, refined grains, fat content and red and processed meat is associated with an increased risk of CRC. Lifestyle factors have also been suggested to increase risk including smoking and alcohol consumption, while physical activity, use of non-steroidal anti-inflammatory drugs, increased intake of vitamin D and calcium have a reduced risk of CRC (Johnson et al. 2013).

Few studies have examined the association between lifestyle environmental exposures and survival in CRC. Evidence suggests pre-diagnostic and post-diagnostic body weight and physical activity may impact CRC survival (Haydon et al. 2006). Factors with less studied effects on CRC survival include dietary patters, smoking, and alcohol consumption. Smoking was found to be associated with increased mortality risk after CRC diagnosis (McCleary et al. 2010,Phipps et al. 2011) Patients conforming to dietary guidelines and following a healthy diet was also associated with lower CRC mortality (Pelser et al. 2014). Among other key prognostic factors of CRC is

classic disease staging and microsatellite instability (MSI). The MSI phenotype results from inactivation of the DNA mismatch repair (MMR) system which leads to the shortening or lengthening of DNA by 1-6 repeating base pain units (Thibodeau et al. 1993). It has been shown that the incidence of MSI differs between stage II and stage III disease, and that its prognostic impact seems to be significantly stronger in stage II than in stage III (Saridaki et al. 2014). The limited available evidence on factors associated with CRC survival warrants more research where specific exposure effects could be better characterized through examination of their interactions with genetic CRC risk factors.

### 1.3.3 Angiogenesis pathway in colorectal cancer

Angiogenesis is one of the hallmarks of cancer as described by Hanahan and Weinberg (Hanahan et al. 2000). Although angiogenesis may not be unique to CRC, it is a key biological process in CRC carcinogenesis necessary for tumor proliferation and progression from colorectal adenoma to carcinoma (Sillars-Hardebol et al. 2010). Angiogenesis is the fundamental process of sprouting and expansion of blood vessels from preexisting vessels that provides the tumor with the blood supply it needs to grow and expand (Folkman et al. 1992). The process of transformation of normal human cells into hyperplastic then into neoplastic cells is critical for tumorigenesis. In CRC carcinogenesis, the process involves the formation of adenomatous polyps and the subsequent progression into malignancy. This multi-step process reflects sequential events of genetic alteration and activation of molecular cancer-related pathways that drive the progressive transformation of cells. Interestingly, the process of angiogenesis can be visualized in hyperplastic tissues prior to their transformation into neoplastic solid tumors, highlighting the role of angiogenesis in premalignant disease (Zhang et al. 2001). Induction of angiogenesis, therefore, seems to be an early event important for conversion of normal

epithelium into a cancer and sustained angiogenesis is essential for tumor expansion, ultimately influencing patient mortality (Ross 1989). The ability of the tumor to induce and sustain angiogenesis is acquired during its transition from a pre-vascular to a vascular phase through tumor progression, referred to as the 'angiogenic switch' (Hanahan et al. 1996). The tumor activates the angiogenic switch by balancing the effects of pro- and anti-angiogenic factors. Shifting this balance involves alteration in gene transcription, underscoring the importance and relevance of the study of polymorphisms in key genes of the angiogenesis pathway to explain variation in cancer susceptibility and survival.

A state of tissue ischemia is by nature toxic to both normal and tumor cells; cancer cells undergo genetic and adaptive changes that allow them not only to survive but also proliferate (Harris 2002). Main drivers of angiogenesis include vascular endothelial growth factor (*VEGF*) (Ferrara 1999,Lohela et al. 2009) and hypoxia-inducible factor 1 (*HIF-1*) (Semenza 2010). *VEGF* expression, potentiated in response to hypoxia, contributes to the development of solid tumors by promoting tumor angiogenesis. The VEGF isoforms bind to two tyrosine-kinase receptors, VEGFR-1 (FLT1) and VEGFR-2 (KDR), expressed almost exclusively in endothelial cells (Neufeld et al. 1999). An association between tumor angiogenesis and overall survival of CRC patients was demonstrated by identifying SNPs on the *VEGF* gene as prognostic markers for CRC (Kim et al. 2008). *HIF-1* is composed of two subunits, *HIF-1α* and *HIF-1β*, and neither is expressed in normal tissue. *HIF-1α* is the oxygen regulated subunit of *HIF-1* and its increased expression has been detected in the majority of solid tumors including colon cancer (Talks et al. 2000). *HIF1-α* was over expressed in premalignant lesions of colon cancer (Zhong et al. 1999), and increased levels of HIF-α have been associated with an aggressive phenotype and decreased patient survival (Rajaganeshan et al. 2008,Schmitz et al. 2009). Additional experimental

evidence demonstrated a loss of function of *HIF-1α* in colorectal cell lines resulting in decreased tumor growth (Imamura et al. 2009), while gain of function led to opposite results (Ravi et al. 2000). Other key angiogenic genes and angiogenesis-related genes are involved in tumor development and progression based on biological plausibility. A pathway approach is well suited to capture their interactions with one another and with environmental factors to induce tumor angiogenesis, and influence CRC risk and prognosis (Mizukami et al. 2007).

### 1.3.4 Hypothesis on environmental factors in association with colorectal cancer

We hypothesized that the effects of three specific lifestyle exposures (dietary protein intake, smoking, and alcohol consumption) on CRC risk and survival (Potter1999b, Gonzalez et al. 2010, Haggar et al. 2009) are modified by angiogenesis genes. We based our hypothesis on the biological information that a state of local tissue ischemia, mainly oxygen deprivation (hypoxia) and glucose deprivation (hypoglycemia) are a main driving force for angiogenesis (Dor et al. 2001). We hypothesized that high animal-based protein intake and heavy smoking patterns were associated with hypoxia and high alcohol intake was associated with hypoglycemia.

***Dietary protein, angiogenesis and CRC risk***

Macro level epidemiological evidence in support of associations between an increased protein diet and CRC risk comes from Japan, a country with a historically low incidence of CRC. A rapidly increasing trend of CRC incidence has been observed in recent years associated with a major shift of the traditional Japanese diet partly in the form of an increased protein intake (Potter1999b, Oba et al. 2006,Takachi et al. 2011). A higher incidence of CRC among migrant Japanese Americans compared to their white counterparts also suggests gene-environment interaction is playing an important role in their increased susceptibility (Marchand1999, Flood et

al. 2000). More recent studies have focused on specific nutrient effects on CRC including animal protein. Positive associations were observed between animal protein intake and colorectal adenoma (Yang et al. 2012) and a meat-based pattern of diet (rich in animal protein among other nutrients of red meat) and CRC (De Stefani et al. 2012).Although there are no studies available on effects of non-animal protein effects on CRC, evidence suggests a protective effect of fruit and vegetable intake on disease outcomes which is more pronounced for distal colon compared to proximal colon and rectal tumors (Voorrips et al. 2000,Annema et al. 2011).

Experimental evidence from studies performed on the *Drosophila* flies can be extended to humans based on a tight similarity of the hypoxic signaling pathways, especially the *HIF-1* pathway (Vigne et al. 2006). Cell survival was diminished in the presence of a chronic hypoxic condition, and when dietary protein was restricted, maintaining the hypoxic condition, the cell survival improved indicating an increased hypoxic tolerance (Vigne et al. 2006, Min et al. 2006). From this evidence, it can be deduced that an increased protein diet in the presence of the hypoxic condition will decrease the tolerance to the hypoxic state. Additional evidence show that dietary proteins and amino acids in the chronic hypoxic conditions can directly shorten the life of cells (Vigne et al. 2008, Grandison et al. 2009). A decrease in hypoxia tolerance, mediated by the high protein diet, is potentially enhancing angiogenesis.

***Smoking, angiogenesis and CRC risk***

Epidemiological evidence of a probable association between cigarette smoking and CRC has been suggested based on prolonged and intense smoking patterns and following a significant lag period (Slattery et al. 1997b, Luchtenborg et al. 2007, Cleary et al. 2010). Nicotine is the main bioactive component of tobacco smoke and was found to stimulate angiogenesis and tumor

growth in lung (Heeschen et al. 2001), gastric cancer (Shin et al. 2005), and colon cancer cells (Wong et al. 2007,Mousa et al. 2006). Although human data on how hypoxia, as a major driver of angiogenesis, is influenced by smoking status is not yet available (Nieder et al. 2008), experimental evidence using a hypoxic model in mice demonstrated that nicotine stimulates angiogenesis under ischemic conditions (Heeschen et al. 2006). Furthermore, both local and systemic administration of nicotine was associated with an increase in *VEGF* expression in colon cancer cells while the oral, systemic route generated increased capillary density. Interaction effects of polymorphisms on *HIF1-α* gene with tobacco smoke and alcohol were also observed to increase the risk of hepatocellular carcinoma (Hsiao et al. 2010). A similar mechanism is potentially operating in CRC. Based on experimental and epidemiological evidence, it is, therefore, plausible that cigarette smoking is interacting with angiogenesis genes and influencing risk of CRC.

***Alcohol intake, angiogenesis and CRC risk***

Epidemiological studies identified increased alcohol consumption as a major risk factor for upper alimentary tract and liver cancers (Poschl et al. 2004), and to a lesser extent in association with CRC (Potter1999b, Ferrari et al. 2007). Studies of alcohol consumption and CRC risk either report an increased or no association; an increased risk was generally reported to be more in association with distal or rectal tumors. Pure ethanol in alcohol is not carcinogenic, but may act as a solvent that enhances penetration of other carcinogens through the mucosal cells of the large intestine (Poschl et al. 2004). Experimental studies, however, showed that ethanol stimulates angiogenesis and increases *VEGF* expression in cell cultures and chick embryos (Gu et al. 2001, Gu et al. 2005), and increased the tumor growth and progression in a mammalian mouse model of melanoma, a skin cancer (Tan et al. 2007). The influence of alcohol intake on glucose

metabolism has long been studied and was shown to induce hypoglycemia especially if consumed without food (van de Wiel 2004). It is plausible that alcohol is enhancing angiogenesis under ischemic conditions.

## OBJECTIVES

### 1.4 Overall objective

The overall goal of this dissertation research is to contribute to the understanding of the biologic and causal mechanisms of complex diseases through studying gene effects and pathway gene-environment effects on disease outcomes. The general purpose is to investigate a sound methodology that is able to represent the biologic underpinnings of disease and yield interpretations that are not only statistically valid but of clinical and/or public health relevance as well.

### 1.5 Specific aims

The specific aims are:

1. To examine the SNP-set interaction effects at the gene level for six chronic diseases using genome-wide association data
2. To assess effect modifications of dietary and lifestyle factors (dietary protein intake, alcohol intake, and smoking) on colon and rectal cancer risk and survival by angiogenesis pathway genetic variants

## GENERAL METHODS

### 1.6 Study populations

The analyses in this dissertation were based on data from two studies: The Wellcome Trust Case Control Consortium (WTCCC) GWAS data (Wellcome Trust Case Control Consortium2007) and a large US-NIH-funded study entitled "Diet, Activity and Lifestyle as a Risk Factor for Colorectal Cancer" (PI: Dr. Martha L Slattery, PhD, MPH, University of Utah).

#### 1.6.1 The WTCCC

The WTCCC GWAS examined seven diseases, six of which were re-analyzed in this study: bipolar disorder (BD), coronary artery disease (CAD), hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D), type 2 diabetes (T2D). (Our work on the seventh disease, Crohn's disease, has been previously published (Dinu et al. 2012)). The WTCCC sample included individuals living within England, Scotland and Wales who were self-identified as white Europeans. The controls came from the 1958 British Birth Cohort and the UK Blood Services project. For each of the diseases studied, approximately 2000 cases and 3000 matched controls were included. We followed the WTCCC recommendations when excluding cases and controls for the analysis based on the sample call rates and evidence of recent non-European ancestry.

***Genotyping of samples and Quality Control***

All 17,000 samples were genotyped using the Affymetrix GeneChip 500K Mapping Array Set. We followed the genotype calling of WTCCC produced by its CHIAMO calling algorithm. Accordingly, we only considered genotype calls with a confidence score of 0.9, and treated the rest of the calls as missing genotypes. SNPs with SNP call rates less than 95% were removed.

We also removed SNPs based on their minor allele frequencies: the default minor allele frequency cut-off in the GenABEL R package was used (2.5/ N where N is the number of subjects), resulting in cut-offs of 0.05% for the WTCCC database. We used a cut-off of 0.2 for the Hardy-Weinberg Equilibrium test's false discovery rates, based on controls. SNP-gene mapping files were retrieved from the OpenBioinformatics website: (http://www.openbioinformatics.org/gengen/tutorial_calculate_gsea.html#_Toc210887414).

Before running the analysis, we removed SNPs within each gene sequentially, such that no pair of remaining SNPs within a gene had linkage disequilibrium ($r2 \geq 0.8$). Genes that included a single SNP following the quality control process were excluded from our SNP-set interaction analysis. Following the analysis and to ensure the quality of the genotype calls, we performed visual inspection of the SNP genotype cluster plots for SNPs in the statistically significant genes. For both cases and controls, we generated SNP genotype cluster intensity plots. SNPs whose plots indicated potential genotyping errors were excluded. This process aimed to exclude false-positive associations.

### 1.6.2 The Diet, Activity and Lifestyle as a Risk Factor for Colorectal Cancer Study

The study is a multicenter, population-based, case-control study of colon and rectal cancer. The colon cancer study was conducted at three centers: the University of Utah, Salt Lake City, Utah; the Kaiser Permanente Medical Care Program of Northern California (KPMCP), Oakland, California; and the University of Minnesota, Minneapolis, Minnesota. The rectal cancer study was conducted at the Utah and KPMCP centers only. The University of Utah served as the coordinating center of the studies.

All eligible participants were identified from residents of defined geographical areas: an eight-county area in Utah (Davis, Salt Lake, Utah, Weber, Wasatch, Tooele, Morgan, and Summit counties); KPMCP members in Northern California; and the metropolitan Twin Cities area (Anoka, Carver, Dakota, Hennepin, Ramsey, Scott, and Washington counties) in Minnesota (except for rectal cancer cases). Final case eligibility was determined by the Surveillance Epidemiology and End Results (SEER) Cancer Registries in Northern California and Utah for California and Utah study participants respectively, and through the Centers for Disease Control and Prevention funded Minnesota Tumor Registry for study participants identified in the Twin Cities Area of Minnesota. These are population- based cancer registries and all newly diagnosed cancer cases (100%) were captured by their respective registries. All registries are operating under a state law that requires cancer reporting from hospitals which have cancer registrars that report cancer cases. The cases from Kaiser in California are representative in terms of demographics as the broader state registry.

Eligiblity criteria:

- Age 30 to 79 years old at time of diagnosis;
- A tumor registry verified, first diagnosis of: primary colon cancer between October 1991 and September 1994 (International Classification Diagnosis – Oncology 2nd edition codes 18.0 and 18.2 – 18.9); or primary cancer in the rectosigmoid junction or rectum between May 1997 and May 2001;
- English speaking;
- Mentally and physically competent to complete the interview;
- Non-Hispanic white, Hispanic, or black and for the rectal cancer study Asian and American Indian people were included

Cases were identified using a rapid-reporting system with the majority of cases interviewed within four months of diagnosis. Cases with a history of previous CRC or known familial adenomatous polyposis, ulcerative colitis, or Crohn's disease as indicated on pathology reports were not eligible. Patients with Crohn's disease and ulcerative colitis have a unique pathology and their etiology differs from the broader group of colorectal cancer cases and hence they were excluded.

Criteria for eligibility for controls were the same as for cases. Controls were frequency matched to cases by sex and 5-year age groups in each geographical area. Controls were randomly selected at KPMCP of Northern California from health maintenance organization membership lists; in Utah controls aged 65 years or more from Health Care Financing Administration lists, and controls aged less than 65 years from driver's license lists; and in Minnesota from driver's license lists (Slattery et al. 1995). Of colon cancer study subjects contacted, 64.5% of cases and 63.7% of controls were interviewed. For rectal cancer study subjects contacted, 65.2% cases and 65.3% controls were interviewed. The response rates from the study were not greatly different than those reported in other epidemiologic studies (Slattery et al. 1995).

***Interview data***

A detailed in-person interview was conducted by trained and certified interviewers using laptop computers. All interviewers were trained centrally prior to the beginning of field operations. Interviewers were blinded to the case/control status of the participant and every interview was audio-taped; of the taped interviews, 1 in 10 of the first 50 and 1 in 20 after that was reviewed for quality control purposes including whether the interviewer asked the questions exactly as written and used the probes appropriately. Interviewers were provided immediate feedback and any

problems were addressed (Edwards et al. 1994). The interview lasted approximately two hours and consisted of two parts: a) the health and lifestyle questionnaire (including data on demographic characteristics, medical history, family history of cancer and polyps, meal patterns, smoking information, and alcohol consumption); and b) the diet history questionnaire (DHQ) (including data on dietary intakes). The DHQ was adapted from the validated CARDIA diet history (Slattery et al. 1994,Liu et al. 1994). Participants were asked to recall foods eaten, the frequency with which they were eaten, foods eaten as additions to other foods, and use of fats during food preparation. Nutrient values for specific foods, rather than broad grouping of foods, were calculated using the Nutrition Coordinating Center Nutrient Database version 19 (Dennis et al. 1980). The referent period for the study questionnaires was the calendar year two to three years prior to diagnosis for cases or selection for controls.

***Tumor registry data***

Local tumor registries provided data on disease stage at diagnosis, months of survival after diagnosis, cause of death, and contributing cause of death. Disease stage was categorized according to SEER cancer staging criteria (in-situ, local, regional, distant, and unknown) (Young et al. 2001). Disease staging was coded centrally by one pathologist in Utah and was missing for 3.3% of colon cancer cases and 1.5% of rectal cancer cases. Survival status was obtained for the Colon Cancer Study up to the year 2000 and for the Rectal Cancer Study up to 2007. At that time all study participants had over five years of follow-up.

***Genotyping of samples***

TagSNPs were selected using the following parameters: LD blocks using a Caucasian LD map (International HapMap Consortium2003) and an r2=0.8; MAF >0.1; LD block range= -1500 bps

from the initiation codon to +1500 bps from the termination codon; and 1 tagSNP for each LD bin. All markers were genotyped using a multiplexed bead-array assay format based on Golden Gate chemistry (Illumina Human Hap550k, San Diego, California). A genotyping call rate of 99.85% was attained. Blinded internal duplicates represented 4.4% of the total sample set; the duplicate concordance rate was 100%. *TGFβ1* gene was not included in the Illumina BeadChip platform; alternatively representative markers were genotyped using a TaqMan assay from Applied Biosystems (Foster City, California). Each 5µl PCR reaction contained 20ng of genomic DNA, primers, probes, and TaqMan Universal PCR Master Mix (containing AmpErase UNG, AmpliTaq Gold enzyme, dNTPs, and reaction buffer). PCR was carried out under the following conditions: 50$^{o}$C for two minutes to activate UNG, 95$^{o}$C for 10 min, followed by 40 cycles of 92$^{o}$C for 15 sec, and 60$^{o}$C for one minute using a 384 well duel block ABI 9700. Fluorescent endpoints of the TaqMan reactions were measured using a 7900HT sequence detection instrument. Individuals with missing genotype data were not included in the analysis for that specific marker.

Although colon and rectal cancers share many risk factors, there is evidence of differences in population characteristics and in the cancers etiology which justifies this approach (Potter1999a, Wei et al. 2004). Accordingly, colon cancer and rectal cancer data were analyzed separately based on expected site-specific associations with genetic and environmental factors. The two main outcomes of interest were cancer risk and survival and we combined the reporting of their results for each cancer.

## 1.7 Biologic interactions between genetic variants

For the WTCCC analysis we examined SNP-set interactions within genes and identified genes with statistically significant associations with the six diseases. For the colon and rectal cancer studies the objective was to model the overall interaction effects of the angiogenesis gene-pathway and the three environmental exposures. Accordingly, the SNP-set interactions within genes were considered as gene-level summaries and used as a first preliminary step to the full 3-step candidate pathway gene-environment interaction approach. More details are described in the individual studies.

Specifically, we explored two forms of SNP-set interactions: SNP *intersection* and SNP *union*. Both forms are derived from set-theory terminology. A SNP intersection is a form of interaction where disease risk is elevated only if *all* of the SNPs in a specified set (e.g., a gene) carry their respective high-risk genotype. A single SNP, or subsets, of the set carrying the high-risk genotype are insufficient to elevate disease risk. For example, for a set of three SNPs, all three SNPs (SNP 1 *and* SNP 2 *and* SNP 3) may have to carry their high-risk genotype for disease risk to be elevated. A SNP union describes a form of interaction where disease risk may be elevated through several independent ways (i.e., genetic heterogeneity) which may include a SNP intersection (e.g., SNP 1 and SNP 2) or an individual SNP carrying the high-risk genotype. We applied the logic regression to search for these biologically plausible forms of SNP-set interactions within genes (Ruczinski et al. 2003).

## 1.8 Logic regression

The logic regression methodology was developed over a decade ago (Ruczinski et al. 2003) to address the problem of detecting high-order interactions and "patterns" of these interactions

among binary predictors within a regression framework. The primary interest is revealing the interaction. Although the methodology can be applied in any setting with binary predictors, it was initially intended for genetic association study applications, e.g., to identify interacting SNPs in association with an outcome. The method employs Boolean logic and searches for Boolean combinations of binary predictors (e.g. SNPs) that improve outcome prediction.

### 1.8.1 The logic expressions

- Boolean operators are: $\Lambda$ (AND), V (OR), $^{c}$ (NOT)
- A binary predictor is a "leaf"
- The Boolean expression/statement is a combination of "leaves" joined by Boolean operators referred to as a "logic tree"; a logic tree can consist of one leaf. Since the predictors are binary, these combinations are binary as well, i.e., the logic tree takes the value of "0 and 1" or "True and False" or "Yes and No".
- The Boolean expression of a logic tree can be represented by the following equation:

$$L = (X_1 \Lambda X_2^{c}) \, V \, X_3$$

where L denotes a logic tree; $X_1$, $X_2$, $X_3$ are binary predictor variables (leaves) with 0 or 1 values. This Boolean statement is read as "$X_1$ and not $X_2$ are true or $X_3$ is true".

The Logic Tree is the graphical representation of the Boolean expression as follows:

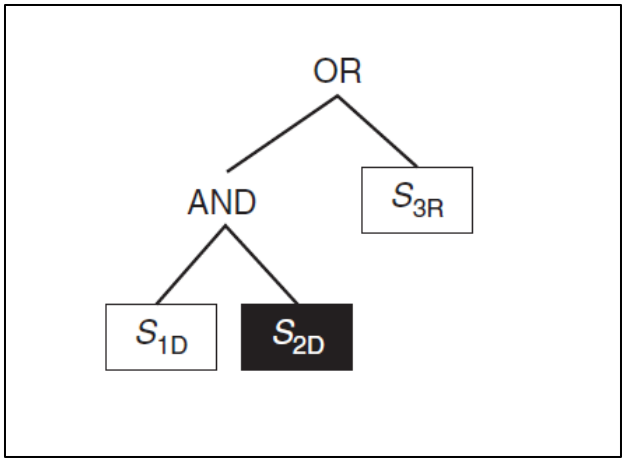



**Figure 1.3**: Graphical representation of a logic tree.

The logic tree is evaluated in a bottom-up fashion, this graph shows one logic tree with three binary variables/leaves ($S_{1D}$, $S_{2D}$ and $S_{3R}$): the top node of the tree is referred to as the "root"; leaves are in boxes. The variables connected by an operator (e.g. $S_{1D}$, $S_{2D}$ combined by AND) are each other's "siblings". The shaded box represents the complement of the variable (represented in the Boolean expression by the NOT operator). Adapted from reference (Schwender et al. 2010).

### 1.8.2 The regression model

The logic regression method involves characteristics specific to the search methodology in addition to those unique to the selected regression class. The logic model can take the form of any other 'regression' model, as long as a scoring function or performance measure reflecting the 'quality' of the model under consideration can be defined. For binary outcomes we fit logic models for logistic regression and the score is the binomial deviance, and for survival outcomes we fit exponential survival models and the score is negative log likelihood. Note that logic models for exponential survival models yielded the same results as Cox proportional hazards model yet with much less computational load.

The logic model with logit link is in the form

$$\log (Pr[Y=1]/ Pr[Y=0]) = \beta 0 + \beta 1\ L1 + \beta 2\ L2 + \ldots + \beta p\ Lp$$

where Y is a binary response variable, $\beta_0$, $\beta_1$,… $\beta_p$ are the parameters, and $L_1$, $L_2$, …, $L_p$ are the Logic Trees.

The logic model for exponential survival took the form:

$$\log \lambda(c) = \beta 0 + \beta 1\ L1 + \beta 2\ L2 + \ldots + \beta p\ Lp$$

where $\lambda(c)$ is the hazard rate, a function of the marginal cumulative hazard c, $\beta 0$, $\beta 1$,…, $\beta p$ are the model parameters, and L1, L2, …, Lp are the Boolean combinations of SNPs.

### 1.8.3 The search for the optimal model

The search space, defined by the number of binary predictors and all their possible Boolean combinations, is huge in genetic association studies (whether GWAS or candidate gene studies). This requires an efficient search strategy. In logic regression this involves: the move set, the search algorithm, and the objective function that quantifies and compares logic models (Schwender et al. 2010).

***The move set***

**Figure 1.4** shows the set of six permissible moves from which, in a finite number of moves, a logic tree can be reached from another tree by a single move at a time.

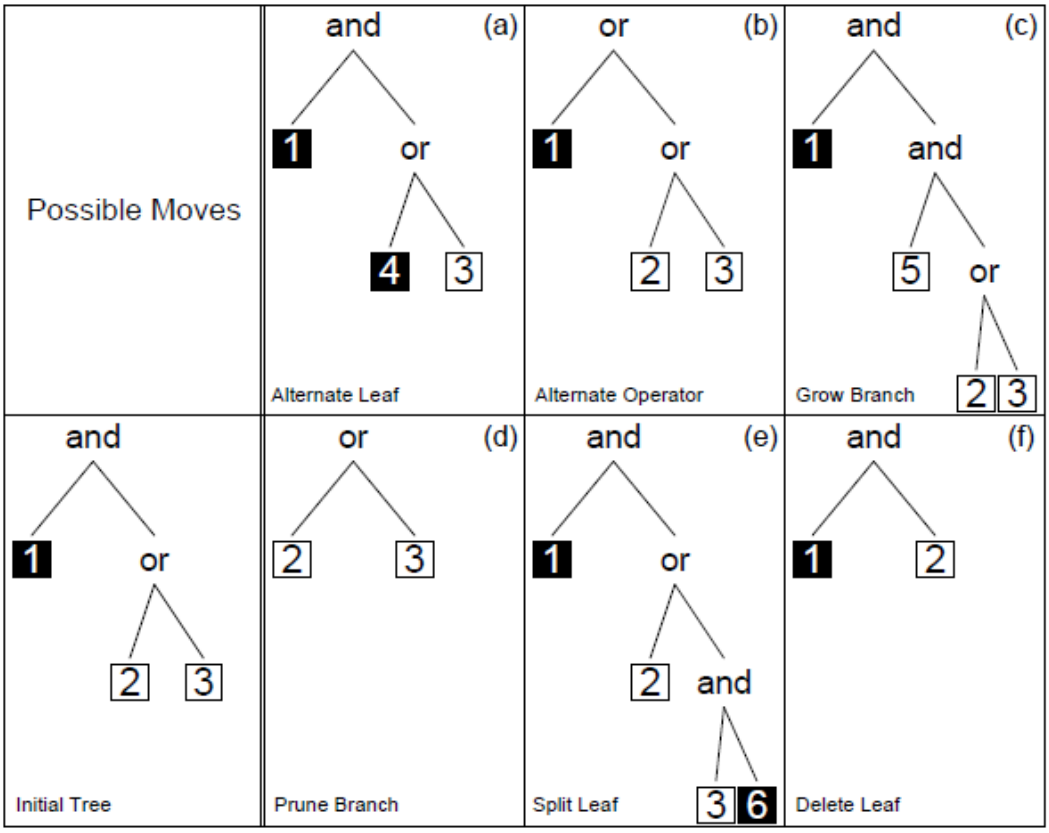



**Figure 1.4:** The move set of the logic regression algorithm. Initial tree is in the bottom left and the remaining panels represent the six permissible moves in the tree-growing process. Details can be found in reference (Ruczinski et al. 2003), figure adapted from same reference.

### *The search algorithm*

One of the search algorithms to select the logic trees implemented in logic regression is the simulated annealing algorithm. Although it has a high computational demand, the algorithm allows the search for a global optimum avoiding getting stuck in local optima (Schwender et al. 2010). It basically involves, given a certain logic tree, randomly picking a move from the set of permissible moves that leads to a new logic tree. Each move has a counter move which allows getting back from the new tree to the old tree. The acceptance probability of the new model is dependent on the scores of both the old and new models and the stage of the annealing process

(referred to as the temperature). The probability of accepting the move is one if the score is better for the new model, and is still positive if the score is worse for the new model, however, it converges to zero as the annealing process progresses and cools down. The further ahead in the annealing scheme the lower the acceptance probability if the new model has a worse score.

***Model selection***

To avoid over-fitting in logic regression models especially in the presence of noise in the data, a model selection procedure for the simulated annealing algorithm is employed. To determine the optimal model size, a definition of model size as a measure of model complexity is needed. The model size was defined as the total number of logic trees and leaves in the logic trees. For the WTCCC GWAS analysis, we used a fixed model size of maximum two logic trees and a total of five leaves (SNPs) to reduce the high computational demand of searching for the best tree/leaf combinations in a large space in GWAS. For the colon and rectal cancer analysis we applied the cross-validation method of model selection in order to find the optimal model size. Generally, logic models chosen by cross-validation or permutation tests rarely exceed sizes of four or five leaves as indicated by the developers of the logic regression. To determine the best overall model using cross-validation, the first step was to state a desired, fixed, model size. The algorithm prohibits further moves that increase the tree if the desired size is reached. This does not necessarily mean that the final model is of the stated desired size but it is up to that desired size. Usually it is of smaller size. We implemented 10-fold of cross-validations for all models with a maximum desired size of nine logic trees and 20 leaves.

***Final model iteration***

To correct for the inherent instability of the performance measure when searching a large space, we refitted the logic models using the obtained optimal model size a 20 times for the WTCCC analysis and a 100 times for the colon and rectal cancer analysis, each time with a different starting search point to reach the best solution.

### 1.9 Measure of statistical evidence

Methods of estimation of statistical evidence are described in detail under their respective studies.

### 1.10 Ethical approval

The studies included in this dissertation were submitted for review and were approved by the Human Research Ethics Board of the University of Alberta. The WTCCC analyses used existing data and no participant contact was made. There were no risks to subjects of the Colorectal Cancer study, since the studies involved analyzing data from stored DNA and did not involve any further procedures or questionnaires administered to study subjects. The analyses only included data from participants who agreed to use of their information for further studies (roughly 99%). All previously conducted study procedures were approved by ethics committees at their respective study locations.

# CHAPTER 2

## Within-Gene Interactions in GWAS Identifies Novel Susceptibility Loci – The WTCCC Data Revisited

### 2.1 Introduction

In genome-wide association studies (GWAS), up to several million *common* single nucleotide polymorphisms (SNPs) are examined in association with disease risk, comparing large numbers of disease cases and disease-free controls (Christensen et al. 2007). Feasibility of GWAS keeps growing with the continual advances in genotyping technology, increased affordability and computational power, and formation of study consortia. A typical GWAS explores SNP-disease associations following the common practice of analysis of a single SNP at a time in association with disease. In the single-SNP analysis, only marginal effects of individual SNPs are considered and effects of interacting loci to disease variability is not captured, which may contribute to the unexplained or *missing* heritability in GWAS (Manolio et al. 2009, Cordell2009b). Furthermore, the commonly reported marginal measures of association, usually odds ratios (ORs), are often very small despite their high levels of statistical significance (YasuiMay 2012, Ku et al. 2010); hence the discovered associations themselves are of limited clinical/public health significance.

In an attempt to address this missing piece of the puzzle in GWAS analysis, we focus here on examination of *SNP-set* interaction effects within-genes on disease risk. Several strategies have been developed to search for interactions on a genome-wide scale, including exhaustive searches (Marchini et al. 2005), Bayesian model selection (Zhang et al. 2007), and two-step analysis approaches (Wu et al. 2010, Tao et al. 2012). These methods rely on exhaustive searches of

*pairwise* (two-way) interactions while a search for higher order interactions, however, is more likely to elucidate the underlying biological mechanism of disease. Current searches of higher order interactions are limited to small sets of markers such as sets of tagging SNPs or markers from a candidate gene study (Schwender et al. 2010). In this report we apply the logic regression (Ruczinski et al. 2003) to search for biologically plausible forms of SNP-set interactions at the genome-wide level. We explored two forms of SNP-set interactions: SNP *intersection* and SNP *union*. Both forms are derived from set theory terminology. A SNP intersection is a form of interaction where disease risk is elevated only if *all* of the SNPs in a specified set (e.g., a gene) carry their respective high-risk genotype. A single SNP, or subsets, of the set carrying the high-risk genotype are insufficient to elevate disease risk. For example, for a set of three SNPs, all three SNPs (SNP 1 *and* SNP 2 *and* SNP 3) may have to carry their high-risk genotype for disease risk to be elevated. A SNP union describes a form of interaction where disease risk may be elevated through several independent ways (i.e., genetic heterogeneity) which may include a SNP intersection (e.g., SNP 1 and SNP 2) or an individual SNP carrying the high-risk genotype. We consider each gene as a set of SNPs and search for SNP-set interactions only within the same gene: this is justified by a recent work that found the number of significant unique expression quantitative trait locus (eQTL) SNPs are much larger than the number of significant unique eQTL-regulated genes, indicating that multiple SNPs within a gene are related to the expression level of the gene (Westra et al. 2013). SNPs within each gene in linkage disequilibrium ($r^2 \geq 0.8$) were removed sequentially before the logic regression to reduce the redundancy of the SNP sets.

We apply our method to the Wellcome Trust Case Control Consortium (WTCCC) GWAS data (Wellcome Trust Case Control Consortium2007) which examined seven diseases, six of which are re-analyzed in this study: bipolar disorder (BD), coronary artery disease (CAD), hypertension

(HT), rheumatoid arthritis (RA), type 1 diabetes (T1D), type 2 diabetes (T2D), and Crohn's disease. We have previously reported in detail on the genetic association results for Crohn's disease which has been published elsewhere (Dinu et al. 2012).

## 2.2 Methods

### 2.2.1 Study samples and genotyping

The WTCCC sample included cases self-identified as white Europeans living in Great Britain, and controls from the 1958 British Birth Cohort and the UK Blood Services project (Wellcome Trust Case Control Consortium2007). For each of the diseases studied, approximately 2000 cases and 3000 shared controls were included and samples genotyped using the Affymetrix GeneChip 500K Mapping Array Set. We followed the genotype calling of WTCCC produced by its CHIAMO calling algorithm (Wellcome Trust Case Control Consortium2007). To further ensure the quality of calls, SNP genotype clusters were visually inspected. Specifically, genotype cluster intensity plots were generated for SNPs included in the logic trees of the statistically significant genes to exclude false-positive associations. The task was performed for both cases and controls (**Supplementary Figure 2.1**). Observing clearly separated genotype clusters indicated a high quality genotype call (**Supplementary Figure 2.1A and 2.1B**), while no clear boundaries or overlapped clusters indicated a potential genotyping error that may lead to a false-positive result from the logic regression (**Supplementary Figure 2.1C and 2.1D**). **Supplementary Figure 2.2** shows the genotype cluster plots for the Logic Tree SNPs of the top significant genes associated with the six diseases. We followed the WTCCC recommendations when excluding cases and controls for the analysis; details of the quality control process is described detail elsewhere (Dinu

et al. 2012). Genes that included a single SNP following the quality control process were excluded from our SNP-set interaction analysis.

### 2.2.2 Estimation of SNP-set interaction effects

Logic regression is a methodology used primarily to detect higher-order interactions between binary predictors (Ruczinski et al. 2003, Schwender et al. 2010). The procedure forms new predictors, referred to as logic trees, from a given set of binary predictors (referred to as leaves) through combining them by Boolean operators.

A Boolean logic statement can be expressed as:

$$L = (X1 \wedge X2c) \vee X3$$

where L is a logic tree; X1, X2, X3 are binary predictor variables with 0 and 1 values; $\wedge$ (AND), V (OR), c (NOT) are operators. In the genetic association study setting the binary predictors are the SNP effects. Thus the above statement when true can be interpreted as 'IF SNP X1 carried the high-risk genotype AND SNP X2 is NOT of the high-risk genotype OR SNP X3 carried the high-risk genotype, THEN a person has a higher risk to develop a particular disease'.

Logic regression uses a simulated annealing algorithm to select the logic trees which basically involves, given a certain state, picking a move from a set of permissible moves that leads to a new state. It compares the scores of the old and new states and accepts the move if the score is better for the new state. The logic model can take the form of any other 'regression' model, as long as a scoring function assessing model goodness-of-fit can be defined. We fitted logic models for logistic regression and model fit was measured by the binomial deviance statistic. Specifically, the logic model took the form:

$$\log (E[Y]/ (1- E[Y])) = \beta 0 + \beta 1\ L1 + \beta 2\ L2 + \ldots + \beta p\ Lp$$

where Y is a binary response variable, β0, β1,…, βp are the model parameters, and L1, L2, …, Lp are the Boolean combinations of SNPs.

We studied one gene at a time, including all measured SNPs on that gene in the search. We limited the logic combinations so that each logic regression fit is set to allow for a maximum of two logic trees (Ls) and a total of five leaves (SNPs). This is done to reduce the high computational demand of searching for the best tree/leaf combinations in a large space in GWAS which can be very large depending on the number of SNPs in the gene. By limiting the number of trees and leaves, we were able to search for major SNP-interaction structures in each gene, although it may not be the full description of the interaction structure. R codes, data examples, and a ReadMe file are available for download from our website: http://www.ualberta.ca/~yyasui/yutaka.html.

### 2.2.3 Statistical significance of associations

We followed the WTCCC's framework of using two methods for measuring evidence of associations of each gene with disease risk: p-value and the Bayes Factor (BF). Each gene was considered as a set of SNPs and explored for SNP intersections and unions. We repeated the search for the best logic combinations through refitting the logic regression model 20 times each with a different random seed to ensure that the best combination is likely to be attained. The minimum deviance fit of the 20 models was selected as the best fit and represented the final model from the original dataset. We then created 20 permuted datasets that shuffle the phenotype labels of cases and controls. The above mentioned process was applied to the original dataset and to each of the 20 permuted datasets yielding a p-value and an approximate BF for each gene. The

BF is approximated by the corresponding likelihood ratio, the denominator is the median of 19 (log10) maximum likelihoods from the 19 permuted datasets (20 minus one because BF of a permuted dataset does not use its own BF in calculating the median of BF from the permuted datasets). The denominator standardizes for the higher potential for genes with larger numbers of SNPs to over fit. The comparison of the best fit of the original data to that of the permuted data for each gene takes into account the size and LD structure of that gene.

The p-value calculation properly took into account the performance of multiple testing. It was calculated for each gene as the proportion of all permuted BF values of all genes larger than the gene's observed BF. **Supplementary Figure 2.3** shows histograms of the empirical p-value distributions for the six diseases to assess their consistency with the theoretical distribution of our test. The theoretical distribution of p-values in our test is a mixture of the uniform distribution in [0,1], corresponding to the null genes, and a right-skewed distribution with a peak at zero, corresponding to the non-null susceptibility genes (Storey et al. 2003): the shape of the latter depends on the number of non-null susceptibility genes and the study power to detect them.

### 2.2.4 Statistical significance threshold

Significance thresholds above which there is strong evidence of association of a gene with a disease were set. The p-value threshold was $3.82 \times 10^{-6}$ corresponding to a p-value of 0.05 with a Bonferroni correction for multiple testing. A BF threshold was calculated for each disease based on the number of genes examined. The calculation of this threshold was based on a prior odds calculation, as follows. If we suppose N genes are investigated, of which 10 genes are assumed to be truly associated with disease risk, then the prior odds for disease risk association for any gene is 10/(N-10). To make the posterior odds of disease risk association for a gene 10 (i.e.,

probability that the gene is associated with disease risk is 10/11, or approximately 0.91), a likelihood ratio for the association over no association (i.e., the BF under the same-size logic-regression model) has to be (N-10). The number of genes we examined in the WTCCC datasets ranged from 13,083-13,106 genes, which yielded a BF threshold (in logarithm with base 10) of approximately 4.12 for all diseases under study.

## 2.3 Results

We report from our analysis many novel signals in addition to previously established associations, both in light of previous reports and in contrast with the single-SNP WTCCC analyses and most recent GWAS meta-analyses (**Table 2.1**). Genes were ranked based on strength of evidence of association using the Bayes Factor (BF) also used by WTCCC as the measure of strength of evidence in their single-SNP analysis. Genes with potential false-positive associations due to potential genotyping errors were separated. Many genes with strong evidence for association had to be removed from the list of significant genes if one SNP or more in the logic trees had a genotype call that appeared erroneous. In **Table 2.2** we present the top five genes showing strong associations with disease risk (all genes with evidence of strong disease associations are shown in **Supplementary Tables 2.1 – 2.6**). We also present the logic structures and odds ratios of association for top statistically-significant genes and show how specific risk groups could be identified from the genotypes of the logic tree SNPs: reference risk, high risk and/or low risk groups **(Table 2.3, panels A-E)**.

### 2.3.1 Bipolar disorder (BD)

Overall, a total of 13,085 genes were examined in association with BD, out of which 13 genes showed strong statistical evidence of association. The one strong signal reported from the

WTCCC single-SNP analysis was at chromosome 16p12. This signal was, however, mapped to the *PALB2* gene which included only one SNP in the WTCCC data following our quality control process: since we excluded all genes with only one SNP post quality control, this gene was excluded and was not included in our analysis. One of the 13 genes we detected as BD susceptibility loci was near the *TENM4* (Teneurin Transmembrane Protein 4) gene. This gene was reported as the single novel locus with a statistically significant association from a recent BD GWAS meta-analysis of 60,000 samples (Psychiatric GWAS Consortium Bipolar Disorder Working,Group2011) **(Table 2.1)**.

BD is a chronic recurring illness characterized by cyclic episodes of mania -extremely elevated mood and energy, disturbed thought patterns and psychotic features such as delusions and hallucinations- and depression (Anderson et al. 2012). Heritability of BD is high, with a 10-fold increased risk among first degree relatives of affected individuals (Smoller et al. 2003), yet identifying genetic variants associated with BD has been limited. Limitations have been attributed mainly to genetic heterogeneity of the disease, and possible small effects of many variants. An interaction based analysis between variants of small effect would be consistent with the nature of a multifactorial trait such as BD. Considering clinical subphenotypes of the disease may also help define genetic risk in more homogenous subsets of patients. The very recent findings of a GWAS comparing seasonal pattern subtypes of mania have identified a susceptibility locus for BD(Lee et al. 2013) that is indeed our top statistically significant gene, the *NFIA* (Nuclear Factor I/A) gene [BF=6.2]. Nuclear factor I/A is identified as a key transcription and regulatory factor involved in glial cell differentiation in the developing central nervous system (Deneen et al. 2006,Mason et al. 2009). Against initial perceptions, glial cells have been proven to play an important role in synaptic neurotransmission and neuron

communication (Volterra et al. 2005), which coincides with growing evidence involving glial cell alterations in psychiatric diseases such as BD (Rajkowska2003).

The *NFIA* gene lies on chromosome 1p31.3 –p31.2 which has been previously implicated as a region of bipolar disorder susceptibility by genome-wide linkage studies (Kremeyer et al. 2010). Other regions on chromosome 1 have also been implicated by recent GWAS as susceptibility loci for BD (Greenwood et al. 2012). The logic structure of the *NFIA* gene consisted of two logic trees: four SNPs for Logic 1 and one SNP for Logic 2. High risk groups could be identified from the SNP genotypes: a reference risk group (552 cases /1080 controls), the highest risk group (Logic 1 = No and Logic 2=Yes: 384 cases/ 413 controls; OR=1.82) (**Table 2.3, Panel A**).

### 2.3.2 Coronary artery disease (CAD)

Our interaction analysis yielded strong evidence of association with CAD for 16 genes out of a total of 13,099 genes examined. The single strong signal of association reported from the WTCCC single-SNP analysis showed the strongest evidence of association in our analysis. Recently, a meta-analysis of 14 GWAS of CAD comprising 86,995 European descent individuals reported on 23 susceptibility loci, 7 (30%) were identified in our analysis (Schunkert et al. 2011).

CAD is a leading cause of death and disability worldwide with both environmental and genetic risk factors contributing to its etiology. The pathogenesis involves buildup of an astherosclerotic plaque in the coronary arteries; narrowing of the arteries and erosion of the plaque and thrombus formation could cause angina or acute myocardial infarction (Crea et al. 2013). The strongest signal of association in our analysis was for the *CDKN2B* (Cyclin-Dependent Kinase Inhibitor 2B) gene [BF=10.8]. The 9p21.3 region, with *CDKN2B* being the nearest gene, has been repeatedly replicated in association with cardiovascular disease. Our results thus confirm the

association initially detected by the WTCCC single SNP analysis and which was confirmed in several subsequent GWAS studies not only in populations of European or Eastern Asian descent (Guo et al. 2013) but of African ancestry as well (Saade et al. 2011,Lettre et al. 2011).

Despite the consistent finding of an association between *CDKN2B* variants and CAD, the mechanism by which the expression of the *CDKN2B* gene confers risk remains unclear. The association appears to be independent of established risk factors including smoking, elevated lipid levels, and diabetes suggesting a non-inflammatory role in disease pathology (Leeper et al. 2013). The *CDKN2B* gene and its adjacent *CDKN2A* gene are tumor suppressor genes involved in the regulation of cell growth. Recent experimental studies suggest a potential role in the development of cardiovascular pathology through altered expression of these genes in the myocardial and vascular tissue. Reduced *CDKN2B* expression was detected in the astherosclerotic plaque and found to accelerate the vascular smooth muscle cell proliferation contributing to the increased risk of CAD (Pilbrow et al. 2012). The logic structure of *CDKN2B* shows the SNP genotypes identifying the different CAD risk groups (**Table 2.3, Panel B**).

### 2.3.3 Hypertension (HT)

The WTCCC single-SNP analysis was not able to identify any variants strongly associated with HT. Upon examination of 13,099 genes with HT for SNP-set interaction, our analysis, however, was able to identify 15 genes with evidence of strong associations with HT. A recent meta-analysis of HT single-SNP GWAS involving 200,000 individuals of European descent reported on a total of 27 loci out of which 8 (30%) were detected in our analysis (International Consortium for Blood Pressure Genome-Wide Association,Studies et al. 2011) **(Table 2.1)**.

HT, also referred to as arterial hypertension, is a chronic medical condition in which the arterial blood pressure is elevated. It is a widely prevalent disease and a major risk factor for cardiovascular and other related diseases (Mancia et al. 2007). The *COL4A4* (Collagen, Type IV, Alpha 4) gene showed the strongest statistical evidence of association in our analysis [BF=5.4]. It encodes one of six identified subunits of type IV collagen: the major structural component of the glomerular basement membrane which together with its capillary wall constitutes the functional glomerular filtration barrier. Mutations in *COL4A4* cause glomerular disease characterized by proteinuria and progression to renal failure (Chen et al. 2012). Specific diseases have been commonly described in association with *COL4A4* mutations: thin basement membrane neohropathy, characterized by persistent glomerular bleeding, and Alport syndrome, a genetic disorder characterized by additional cochlear and ocular involvement (Kashtan 1993). HT is an important factor in the progression of renal disease, potentially explaining the observed association of *COL4A4* variants to HT risk. Indeed the main line of treatment of Alport syndrome involves routine treatment of HT.

Collagen is among proteins accumulated in extracellular spaces in cases of fibrosis, representing the main pathologic feature of progressive fibrotic disease including renal fibrosis, heart failure and HT. Antifibrotic and antiproteinuric effects recently demonstrated for traditional antihypertensive drugs indicates a potential novel therapeutic target related to the observed Collagen Type coding genes (Gross et al. 2011).

*COL4A4* lies on chromosome 2q35-q37 which has been detected by genome-wide linkage analyses in association with the quantitative traits of HT (systolic, diastolic and pulse pressure) (Aberg et al. 2009). Despite the implication of the chromosome 2q region in linkage with the blood pressure related traits, no replicated candidate genes have been identified: this underscores

the potential discoveries of the SNP-set interaction analysis. **Table 2.3, Panel C** shows the logic structure of the *COL4A4* gene in association with increased risk of HT.

### 2.3.4 Rheumatoid arthritis (RA)

We examined 13,083 genes in association with RA, of which 72 genes showed strong evidence of association, covering the two regions of strong association detected in the WTCCC single-SNP analysis. Six (60%) out of ten RA loci reported from a RA GWAS meta-analysis of 41,282 individuals of European descent were included in our significant genes (Stahl et al. 2010) **(Table 2.1)**.

RA is a chronic inflammatory auto-immune disease causing damage to the synovial joints especially those of the hands and feet leading to pain and stiffness and could progress to joint deformity and disability (Escalante 2013). The human leukocyte antigen (HLA) region has long been established as a genetic contributor to RA susceptibility (Viatte et al. 2013). Variants have been consistently linked to the major histocompatibility complex (MHC) region including MHC class II genes *HLA-DQB1* and *HLA-DRB1*. Indeed that is confirmed by the top associations from our analysis, many of which are shared with T1D **(Table 2.2)**. Less specifically investigated is the MHC class II-associated region that yielded the top significant association in our analysis with the *BTNL2* (Butyrophilin-Like 2) gene. *BTNL2* belongs to the immunoglobulin superfamily suggested to play a role in the T-cell activation pathway that is key to the pathogenesis of autoimmune diseases such as RA and T1D (Orozco et al. 2005). Previous GWAS identifying *BTNL2* variants in association with RA have attributed it to being in strong linkage disequilibrium with the confirmed MHC genes *HLA-DQB1* and *HLA-DRB1* (Cui et al. 2009). Recently, the independent association of *BTNL2* with RA has been corroborated through exome

sequencing which identified *BTNL2* variants associated with RA independent from other RA candidate genes (*HLA-DRB* and *NOTCH4*) (Mitsunaga et al. 2013). The documented associations of *BTNL2* with other auto-immune diseases (Morais et al. 2012) renders an independent *BTNL2* – RA association worthy of further investigation. Based on the genotypes specified in the logic trees for the association of *BTNL2* and RA, carrying the high-risk genotypes infers an appreciably increased risk of RA (**Table 2.3, Panel D**).

### 2.3.5 Type 1 diabetes (T1D)

Among a total of 13,101 genes examined, 105 showed evidence of strong association with T1D. All regions reported from the WTCCC single-SNP analysis in strong association with T1D were detected in our interaction analysis. Twelve out of a total of forty one (29%) novel and known chromosomal regions reported from a combined T1D GWAS and meta-analysis of over 16,000 samples were also included in our significant genes (Barrett et al. 2009) **(Table 2.1)**.

T1D is an autoimmune disease mediated by both genetic and environmental triggers. Disease pathogenesis involves lymphocytic infiltration of pancreatic islets, destruction of beta cells, and lifelong dependency on exogenous insulin (Gan et al. 2012). The top associations were unsurprisingly for the established HLA regions on chromosome 6 including MHC genes (Erlich1991). Specifically, several MHC class II [HLA-DQ] genes showed the strongest evidence of association with T1D in our analysis. T1D and RA shared associations with HLA-DQ genes in our analysis, which have also been suggested to represent a continuous spectrum of genetic association from typical T1D, through latent autoimmune diabetes in adults, to T2D (Lin et al. 2008).

### 2.3.6 Type 2 diabetes (T2D)

A total of 19 genes among 13,083 genes examined in our interaction analysis showed strong evidence of association with T2D. Our interaction analysis detected all strong signals reported from the WTCCC single-SNP analysis (Savic et al. 2012, Herder et al. 2011). A recent meta-analysis of T2D GWAS of over 25,000 East Asian individuals identified seven new T2D loci; two (29%) of which were included in our results (Cho et al. 2012) **(Table 2.1)**.

Type 2 diabetes mellitus is a chronic metabolic disease characterized by high blood glucose levels traditionally attributed to insulin resistance (Roglic et al. 2005). Genes linked to insulin resistance, obesity, and other aspects of glucose metabolism are seldom identified. More commonly, GWAS has typically implicated genes with recognized roles in the beta cell development and the function of the adult pancreas (Florez2008). The strongest association signal reported from both the WTCCC single-SNP analysis and our analysis was for the *TCF7L2* (Transcription Factor 7-Like 2) gene [BF=8.7]. Genetic variants on *TCF7L2* were first identified from an Icelandic population and have since been replicated repeatedly across different populations (Grant et al. 2006). The gene encodes a transcription factor implicated in blood glucose homeostasis predisposing to T2D through impairment of beta cell function and insulin secretion rather than mechanisms of insulin resistance (Pearson2009).

Genetic predisposition to beta cell dysfunction could be one of the determinants of individual susceptibility to T2D; exogenous factors inducing insulin resistance are then needed for manifestation of the disease. This may include further genetic or environmental factors including lifestyle factors. In line with this hypothesis, variants on the *FTO* (Fat Mass And Obesity Associated) gene seem to affect T2D risk mediated through its clear effect on obesity risk

(Herder et al. 2011). The *FTO* gene was in fact reported among the strong signals of WTCCC single-SNP analysis and from our analysis, yet not among the top genes. T2D being one of the diseases with a complex pathogenesis, interaction between genotypes and lifestyle and treatment factors might lead to a more complete understanding of the genetic contribution to T2D risk. Some evidence indicates a suggestive reduction of a *TCF7L2* related T2D risk in response to lifestyle changes (Florez et al. 2006), however it is not consistent and calls for further investigation. The OR estimates from the logic regression are less dramatic compared to the other diseases studied which emphasizes a potential larger effect if environmental interaction effects are examined (**Table 2.3, Panel E)**.

## 2.4 Discussion

Multiple disease susceptibility loci including novel signals with biologically plausible links to six of the diseases under study by WTCCC were detected in our interaction analysis. GWAS results are based on an agnostic type of search for SNP-disease associations with no SNP having a priori higher probability of being associated with a disease (Hunter et al. 2010). Discoveries like the ones presented here from the SNP-set interactions illustrate the additional power of GWAS which has not been revealed previously by the standard single-SNP analysis. An interaction analysis does not, however, preclude the importance of association signals reported from single-SNP analyses. The majority of previously reported strong associations were detected in our analysis - almost all single-SNP WTCCC strong signals and on average 46% of the most recent GWAS meta-analysis reported loci (**Table 2.1**) - yet they were not necessarily our most significant results. The added value of an interaction analysis is the emergence of strong evidence implicating new genes that were not detected before but are supported by apparent biological links to disease. For example, strong evidence of genetic associations of the top

significant genes such as the *COL4A4* gene with HT risk, have never been reported from GWAS before. Our results provide confirmation of previous linkage analyses implicating specific chromosomal regions and diseases (e.g., 1p31 near *NFIA* gene in association with BD and 2q35 near *COL4A4* gene in association with HT). Other recent discoveries were confirmed in our analysis such as *BTNL2* with RA and *TCF7L2* with T2D and are worthy of further in depth investigation. The fact that such discoveries were detected from an interaction-based analysis in GWAS adds strength to our approach of analysis and emphasizes the importance of searching for SNP-set interaction effects, in addition to the standard single-SNP analysis in GWAS.

The ORs, a measure of SNP-disease associations, typically reported from single-SNP based GWAS analysis are in the range of 1.1-1.5 which, despite the high statistical significance, are in themselves of minimal clinical and public health importance. On the other hand, the SNP-set interaction analysis of GWAS yielded ORs with substantially larger magnitudes indicating strong associations that are more readily interpreted. Our results also allowed us to identify a larger number of genes in association with the six diseases that help determine their genetic risk. These results help to further explain the genetic roles in the pathogenesis of these diseases and opens avenues for refined risk identification and risk prediction. Of specific value are diseases with a less clearly identified genetic risk such as BD, CAD and HT. For example, the single-SNP based analysis reported from WTCCC failed to identify any strong association signals for HT, while our interaction analysis was able to detect both novel and previously reported association signals providing new insights into the genetic profile of HT.

The numbers of novel and significant genes detected from our analysis are not merely the results of lowering the significance threshold by analyzing genes rather than SNPs which reduces the number of multiple testing, or by using Bayes Factors rather than p-values as a measure of

statistical evidence for associations. Rather, they are the consequences of our analysis approach that attempts to systematically identify interactions across SNPs within each gene. The results of the SNP-set approach would not be achieved by simply lowering the threshold for single-SNP analysis. In support of the SNP-set approach compared to a single-SNP analysis, we plotted Bonferroni corrected p-values from the single-SNP analysis against our SNP-set interaction analysis (**Supplementary Figure 2.4**). All genes whose Bonferroni corrected p-values were < 0.1 by either of the two tests were plotted. These plots show that under the same criterion of Bonferroni corrected p-values, the majority of the genes were at the bottom of the plots but not necessarily near the origin, indicating greater numbers of significant signals by our SNP-set interaction approach. Thus, lowering the significance threshold for single-SNP analysis will not yield similar significant and novel signals as those detected from our SNP-set interaction approach. We also repeated all steps of the analysis on 10 permuted datasets using RA as an example. Out of the 10 permuted datasets each involving 13,083 genes, only one gene of one of the 10 datasets was statistically significant using the Bonferroni corrected p-value threshold. Thus, the huge search space for the logic regression does not explain our findings.

Our logic regression analysis, despite its utility, is not without limitations. To manage the large computational demand of the logic regression search, we had to limit the search of SNP-set interactions to a single gene at a time and fix the size of SNP interactions searched for within each gene. Our assessment of SNP-set interactions, however, has a much higher power of signal discovery compared to a SNP-set interaction test for a pair of SNPs since a significantly fewer number of tests are performed. We did not consider gene-gene interactions in the analysis; it is possible that more complex SNP interactions exist but were not discovered in our analysis. Our use of Boolean logic through logic regression may capture a subset of cis-regulatory modules if

they exist; a fuller more comprehensive modeling of the modules would require consideration of gene-gene interactions. Another potential problem is identifying false-positive associations from GWAS. The problem is mainly attributed to genotype-calling errors. Visual inspection of genotype clusters is a common approach to identify markers with errors. We further excluded markers based on the visual inspection process demonstrating the importance of adopting quality control criteria beyond algorithms designed for specific arrays such as the CHIAMO algorithm used in WTCCC.

Despite the limited form of logic regression that we applied in our analysis, searching for specific forms of SNP-set interactions is a step towards addressing the complexity of genetic associations in a GWAS compared to a marginal assessment of individual SNP effects on disease. It is important to note that although it is impossible to validate discoveries made by logic regression analysis with single-SNP analyses; our SNP-set interaction-based analysis was able to detect the majority of previous single-SNP associations including those of large meta-analysis datasets. Thus, confirmation of our novel signals need to be further investigated in larger datasets and using a gene-level interaction-based analysis.

**Table 2.1: Summary numbers of SNP-set interaction signals overlapping with WTCCC single-SNP strong signals and single-SNP Meta-analysis signals[¥]**

| Disease | Logic-Based SNP-set Interaction | WTCCC single-SNP | | Meta-Analysis | |
|---|---|---|---|---|---|
| | Number of Significant Signals | Number of Strong Signals | Overlap | Number of Loci | Overlap[*] |
| BD | 13 | 1 | 0/1 (0%) | 1 | 1/1(100%) |
| CAD | 16 | 1 | 1/1 (100%) | 23 | 7/23 (30%) |
| HT | 15 | NA | NA | 27 | 8/27 (30%) |
| RA | 72 | 2 | 2/2(100%) | 10 | 6/10 (60%) |
| T1DM | 105 | 5 | 5/5(100%) | 41 | 12/41 (29%) |
| T2DM | 19 | 3 | 3/3 (100%) | 7 | 2/7 (29%) |

[¥]Bayes Factor used as measure of evidence of association of each gene and disease risk; [*]Overlap within 100 kb or less of the reported meta-analysis locus; (BD) Bipolar disorder(Psychiatric GWAS Consortium Bipolar Disorder Working,Group2011) ; (CAD) Coronary artery disease (Schunkert et al. 2011); (HT) Hypertension (International Consortium for Blood Pressure Genome-Wide Association,Studies et al. 2011); (RA) Rheumatoid arthritis(Stahl et al. 2010); (T1D) Type 1 diabetes(Barrett et al. 2009); (T2D) Type 2 diabetes (Cho et al. 2012); (SNP) Single nucleotide polymorphism; (WTCCC) Wellcome Trust Case Control Consortium; (NA) Not Applicable.

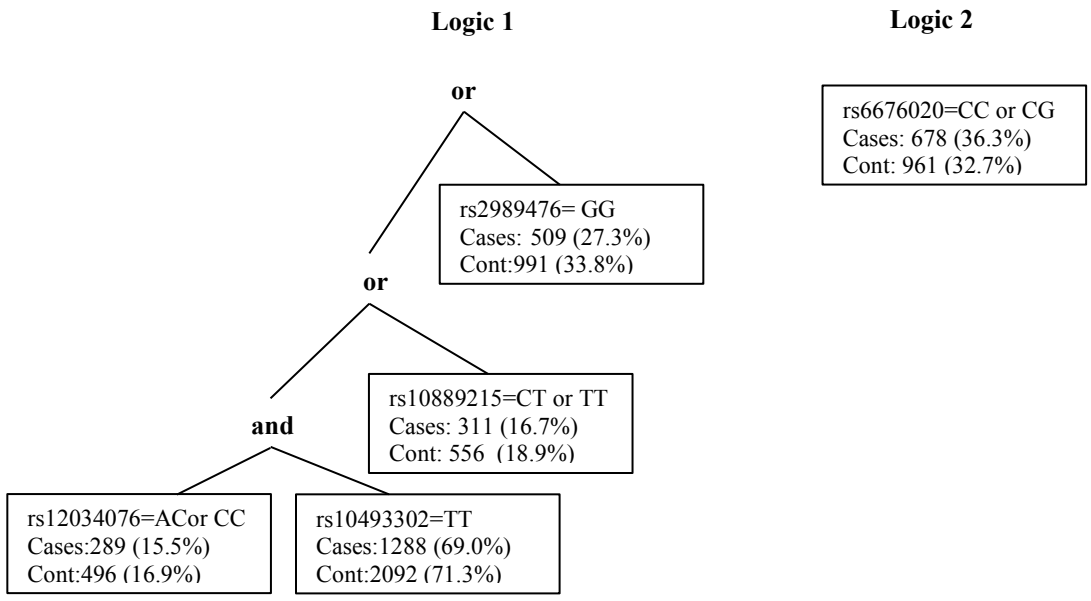
**Table 2.2: Top 5 genes showing the strongest evidence of association with each disease**

| Disease | Chromosomal Location | Gene Name | #SNPs | BF | P-value |
|---|---|---|---|---|---|
| **BD** | 1p31.3-p31.2 | *NFIA* | 113 | 6.20 | $1.15 \times 10^{-5}$ |
| | 1q44 | *NLRP3* | 9 | 5.87 | $1.53\times 10^{-5}$ |
| | 12q15 | *PTPRR* | 42 | 5.52 | $4.97\times 10^{-5}$ |
| | 2q12-q21 | *DBI* | 4 | 5.23 | $6.88\times 10^{-5}$ |
| | 5q15 | *RIOK2* | 36 | 5.06 | $8.79\times 10^{-5}$ |
| | | | | | |
| **CAD** | 9p21 | *CDKN2B* | 24 | 10.85 | $<3.81\times 10^{-6}$ |
| | 11p15.3-p14 | *TPH1* | 4 | 7.22 | $<3.81\times 10^{-6}$ |
| | 11p14.3 | *USH1C* | 22 | 5.68 | $1.91\times 10^{-5}$ |
| | 9p13.3 | *RECK* | 13 | 5.57 | $2.67\times 10^{-5}$ |
| | 3p21.31 | *CDCP1* | 26 | 5.45 | $3.44\times 10^{-5}$ |
| | | | | | |
| **HT** | 2q35-q37 | *COL4A4* | 10 | 5.40 | $4.58\times 10^{-5}$ |
| | 15q14 | *GJD2* | 28 | 5.34 | $4.58\times 10^{-5}$ |
| | 2q11.2-q12.1 | *ST6GAL2* | 77 | 5.19 | $6.49\times 10^{-5}$ |
| | 15q22.31 | *MTFMT* | 3 | 5.18 | $6.49\times 10^{-5}$ |
| | 4p16.1 | *BOD1L1* | 107 | 5.04 | $7.63\times 10^{-5}$ |
| | | | | | |
| **RA** | 6p21.3 | *BTNL2* | 10 | 95.34 | $<3.82 \times 10^{-6}$ |
| | 6p21.3 | *HLA-DRA* | 12 | 93.87 | $<3.82 \times 10^{-6}$ |
| | 6p21.3 | *C6orf10* | 13 | 91.71 | $<3.82 \times 10^{-6}$ |
| | 6p21.3 | *HLA-DQB1* | 7 | 82.23 | $<3.82 \times 10^{-6}$ |
| | 6p21.3 | *NOTCH4* | 15 | 64.9 | $<3.82 \times 10^{-6}$ |
| | | | | | |
| **T1D** | 6p21.3 | *HLA-DQB1* | 7 | Inf | $<3.82 \times 10^{-6}$ |
| | 6p21.3 | *HLA-DRA* | 10 | 230.25 | $<3.82 \times 10^{-6}$ |
| | 6p21.3 | *BTNL2* | 10 | 214.04 | $<3.82 \times 10^{-6}$ |
| | 6p21.3 | *C6orf10* | 12 | 206.30 | $<3.82 \times 10^{-6}$ |
| | 6p21.3 | *HLA-DQA1* | 4 | 203.42 | $<3.82 \times 10^{-6}$ |
| | | | | | |
| **T2D** | 10q25.3 | *TCF7L2* | 38 | 8.70 | $<3.82 \times 10^{-6}$ |
| | 4q27 | *TMEM155* | 8 | 7.58 | $<3.82 \times 10^{-6}$ |
| | 5q15 | *FAM172A* | 12 | 6.96 | $<3.82 \times 10^{-6}$ |
| | 16q13 | *GPR56* | 12 | 6.20 | $<3.82 \times 10^{-6}$ |
| | 13q12.12 | *C1QTNF9* | 4 | 5.07 | $1.22\times 10^{-4}$ |

(BD) Bipolar disorder; (CAD) Coronary artery disease, (HT) Hypertension, (RA) Rheumatoid arthritis, (T1D) Type 1 diabetes, (T2D) Type 2 diabetes, (SNP) Single nucleotide polymorphism, (BF) Bayes Factor.

**Table 2.3: Logic structures, frequencies, and associated disease odds ratios of the top significant gene**

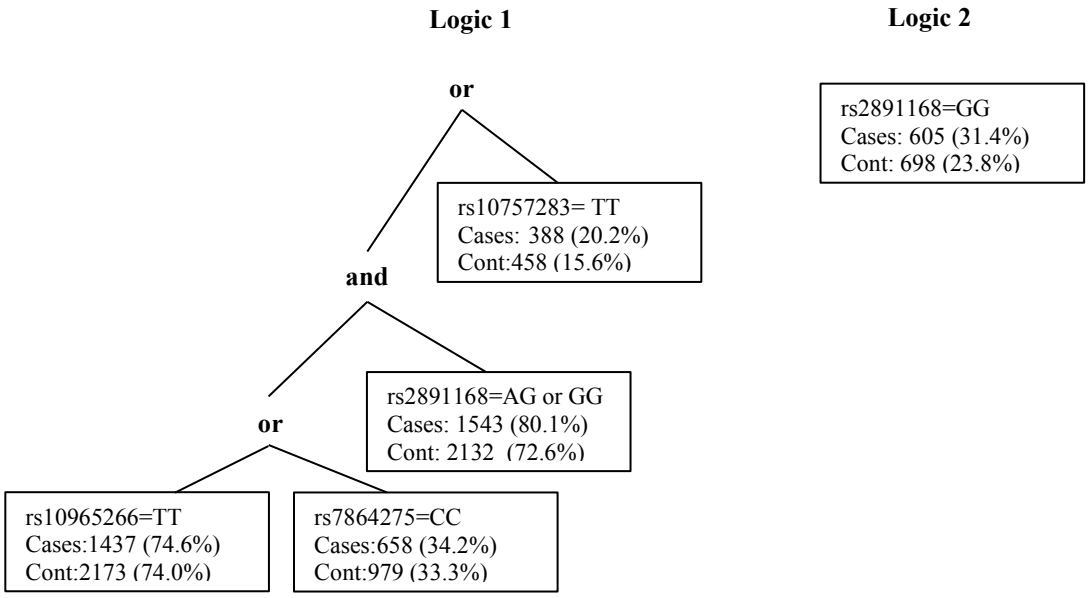
**A. Bipolar disease and *NFIA* gene**




| Logic-based Risk Groups | | Cases | Controls | Odds Ratio |
|---|---|---|---|---|
| Logic 1 | Logic 2 | | | |
| Yes | No | 552 | 1080 | 1.0 (Ref) |
| No | Yes | 384 | 413 | 1.82* |
| No | No | 638 | 895 | 1.39* |
| Yes | Yes | 294 | 548 | 1.05* |

**A**. Logic structure consisted of two logic trees with a total of five SNPs identifying two risk groups: reference risk group (552 cases / 1080 controls), *high risk groups (Logic 1= No and Logic 2=Yes: 384 cases/ 413 controls; OR=1.82; Logic 1 and 2 =No: 638 cases / 895 controls; OR = 1.39; Logic 1 and 2=Yes: 294 cases/ 548 controls; OR = 1.05).

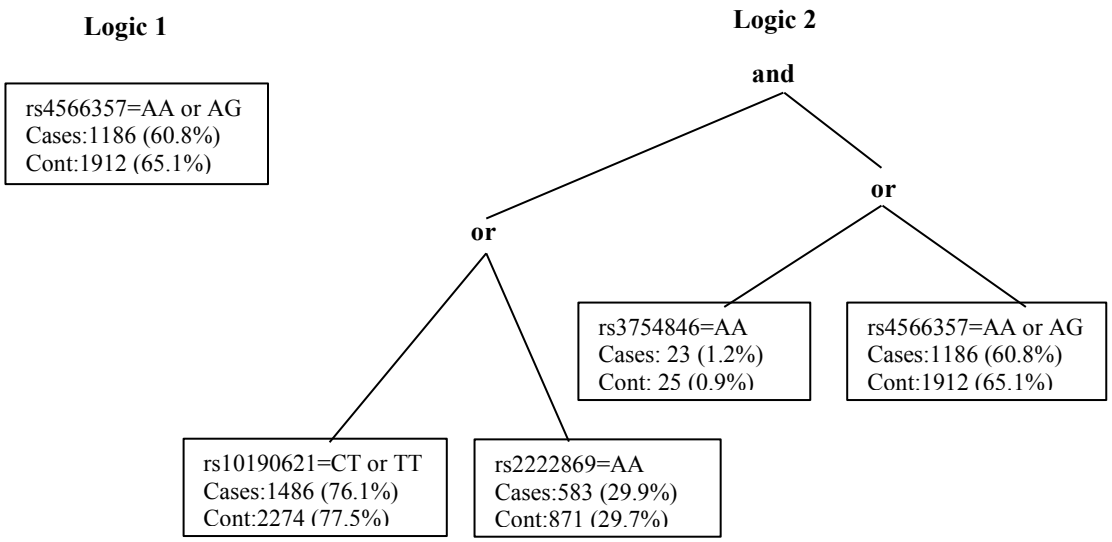
**B. Coronary artery disease and *CDKN2B* gene**




| Logic-based Risk Groups | | Cases | Controls | Odds Ratio |
|---|---|---|---|---|
| Logic 1 | Logic 2 | | | |
| Yes | No | 946 | 1381 | 1.0 (Ref) |
| Yes | Yes | 564 | 628 | 1.31* |
| No | No | 375 | 857 | 0.64¥ |
| No | Yes | 41 | 70 | 0.86¥ |

**B**. Logic structure consisted of two logic trees with a total of five SNPs identifying three risk groups: reference risk group (946 cases / 1381 controls), *high risk group (Logic 1 and 2=Yes: 564 cases/ 628 controls; OR = 1.31and ¥low risk groups (Logic 1 and 2 =No: 375 cases / 857 controls; OR = 0.64; Logic 1 = No and Logic 2=Yes: 41 cases/ 70 controls; OR=0.86).

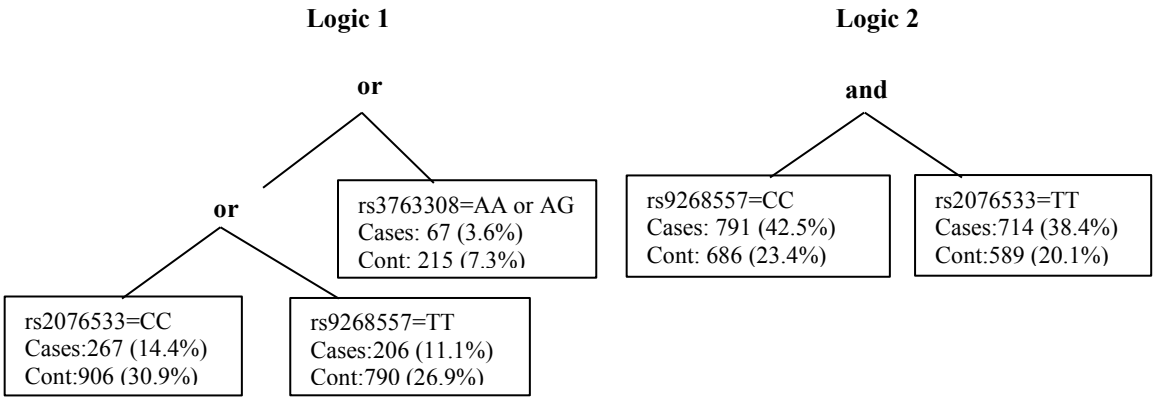
## C. Hypertension and *COL4A4* gene




| Logic-based Risk Groups | | Cases | Controls | Odds Ratio |
|---|---|---|---|---|
| Logic 1 | Logic 2 | | | |
| Yes | Yes | 1185 | 1877 | 1.0 (Ref) |
| No | Yes | 8 | 1 | 12.7* |
| No | No | 758 | 1023 | 1.17 |
| Yes | No | 1 | 35 | 0.05¥ |

**C.** Logic structure consisted of two logic trees with a total of five SNPs identifying three risk groups: reference risk group (1185 cases / 1877 controls), *high risk groups (Logic 1 =No and Logic 2 = Yes: 8 cases/ 1 control; OR = 12.7; Logic 1 and Logic 2 =No: 758 cases/1023 controls; OR=1.17), and a ¥low risk group (Logic 1 =Yes and Logic 2 =No: 1 case/ 35 controls; OR=0.05).

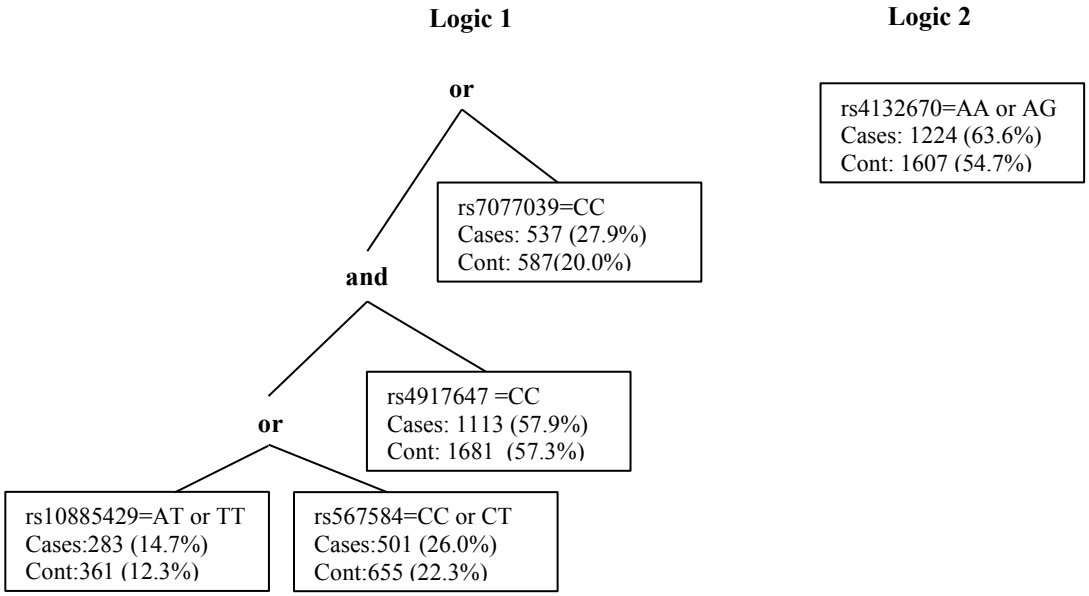
**D. Rheumatoid arthritis and *BTNL2* gene**




| Logic-based Risk Groups | | Cases | Controls | Odds Ratio |
|---|---|---|---|---|
| Logic 1 | Logic 2 | | | |
| Yes | No | 381 | 1321 | 1.0 (Ref) |
| No | Yes | 574 | 313 | 6.36* |
| No | No | 905 | 1300 | 2.41* |
| Yes | Yes | 0 | 2 | 0 |

**D.** Logic structure consisted of two logic trees with a total of five SNPs identifying two risk groups: reference risk group (381 cases and 1321 controls) and high risk groups (Logic 1 = No and Logic 2 = Yes: 574 cases/ 313 controls; OR=6.36 and Logic 1 = No and Logic 2 = No: 905 cases/ 1300 controls; OR =2.41).

**E. Type 2 diabetes and *TCF7L2* gene**



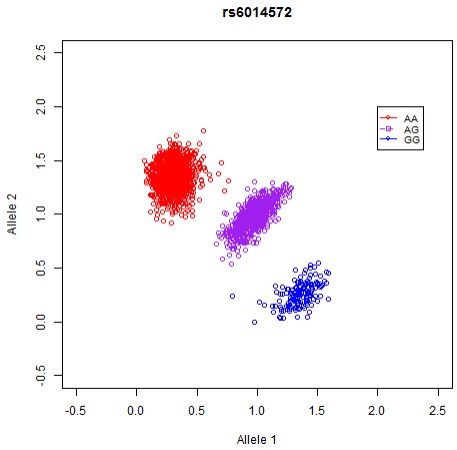

| Logic-based Risk Groups | | Cases | Controls | Odds Ratio |
|---|---|---|---|---|
| Logic 1 | Logic 2 | | | |
| No | No | 586 | 1196 | 1.0 (Ref) |
| Yes | Yes | 591 | 636 | 1.90* |
| Yes | No | 114 | 133 | 1.75* |
| No | Yes | 633 | 971 | 1.33* |

**E.** Logic structure consisted of two logic trees with a total of five SNPs identifying two risk groups: reference risk group (586 cases / 1196 controls), high risk groups (Logic 1 and 2=Yes: 591 cases/ 636 controls; OR=1.90; Logic 1 = Yes and Logic 2 =No: 114 cases / 133 controls; OR = 1.75; Logic 1= No and Logic 2=Yes: 633 cases/ 971 controls; OR = 1.33).

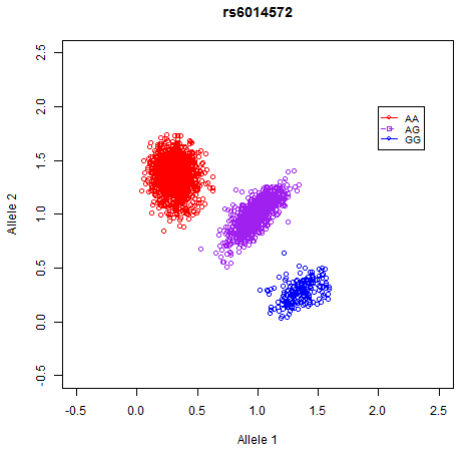
**Within-Gene Interactions in GWAS Identifies Novel Susceptibility Loci – The WTCCC Data Revisited**

**Supplementary Information**

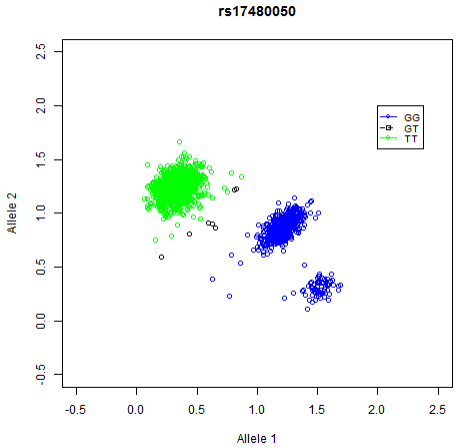
**A**



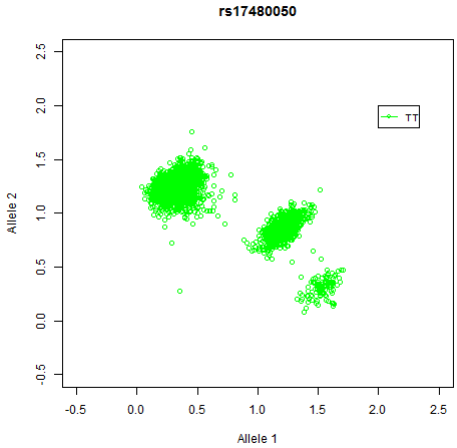

**B**




**C**

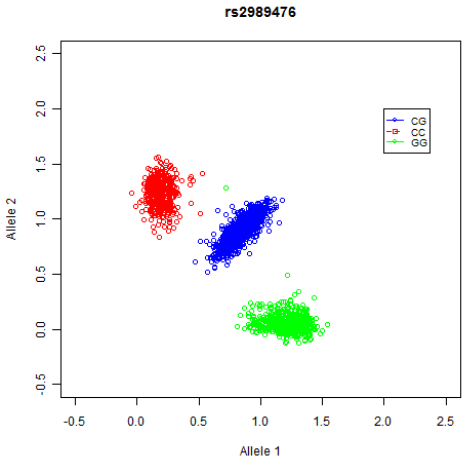


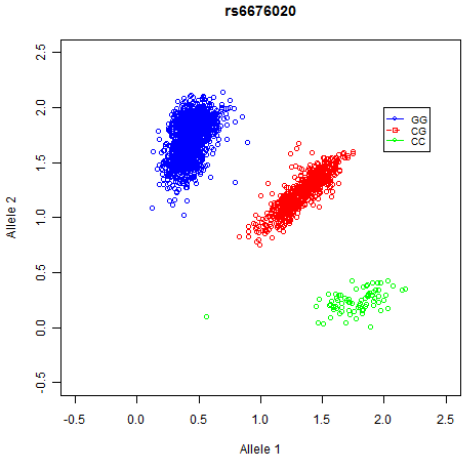

**D**

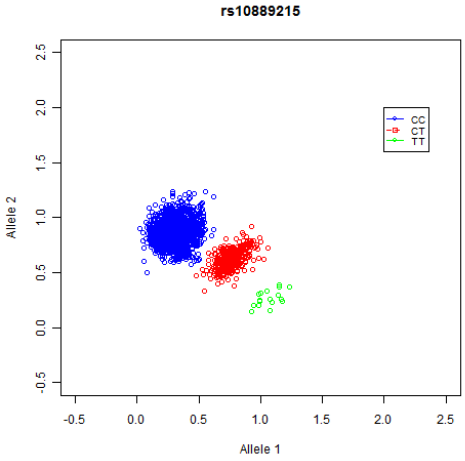



**Supplementary Figure 2.1:** Visual inspection of genotype cluster intensity plots. Genotype cluster intensity plots were generated for SNPs included in the logic trees of the statistically significant genes. The task was performed for both cases and controls. Plots on the left are for Bipolar Disease cases (A& C) and plots on the right are for controls (B & D). The x and y axes on the plots denote the intensity measurements for the two alleles at the SNP. Each point represents the measurement for a single individual. Plots A & B show an example of correct genotype calling for SNP rs6014572 on *CBLN4* gene indicating a high quality marker. In Plots C & D incorrect genotype calling was observed for SNP rs17480050 on *CSGALNACT1*gene with clear overlap of the homozygous alleles (GG and TT) indicating a genotyping error that may lead to a false-positive association.

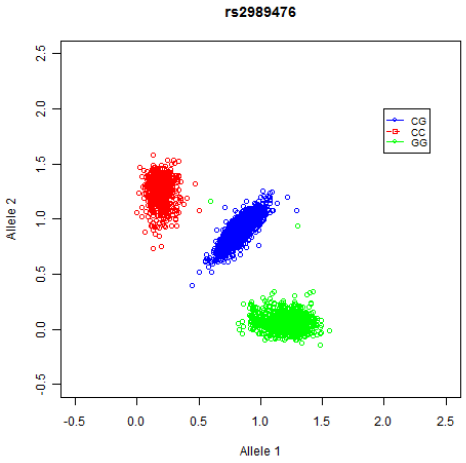
**BD Cases**

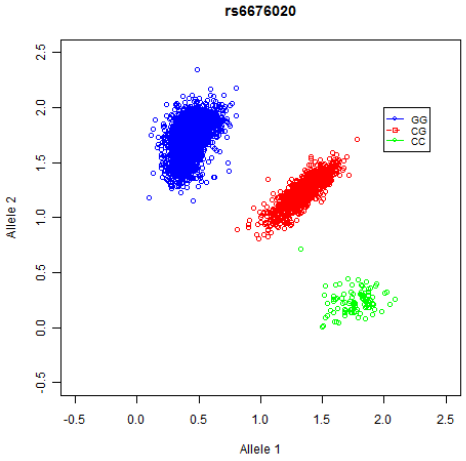


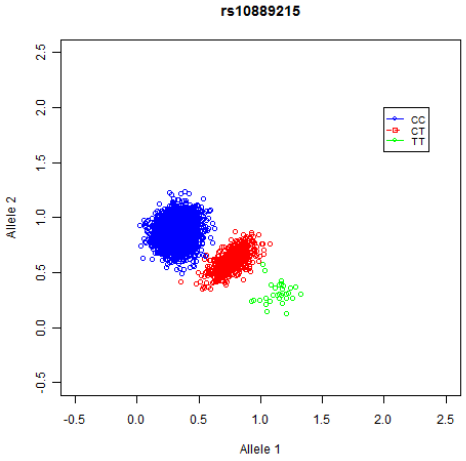



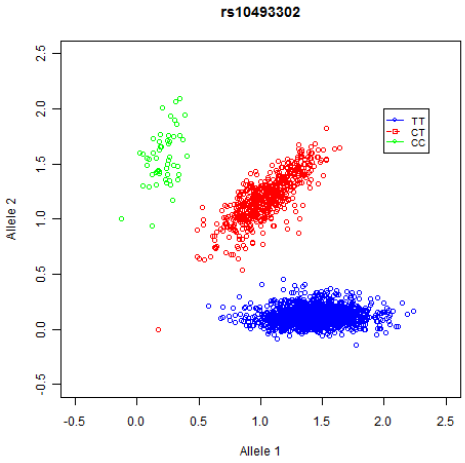



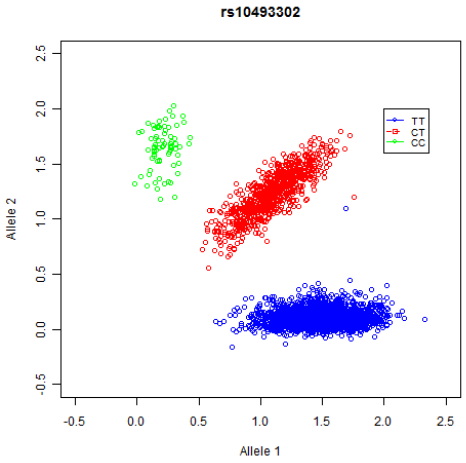

**BD Controls**

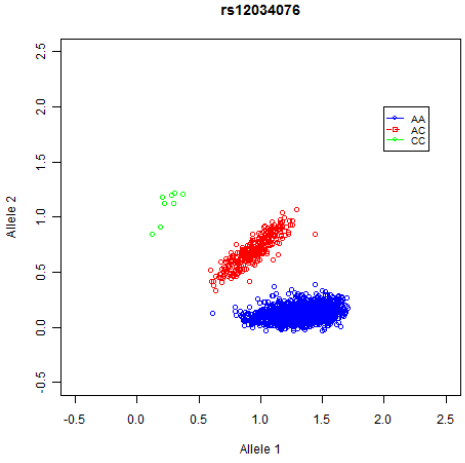


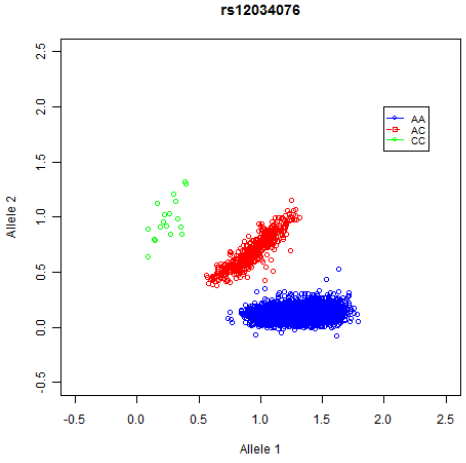





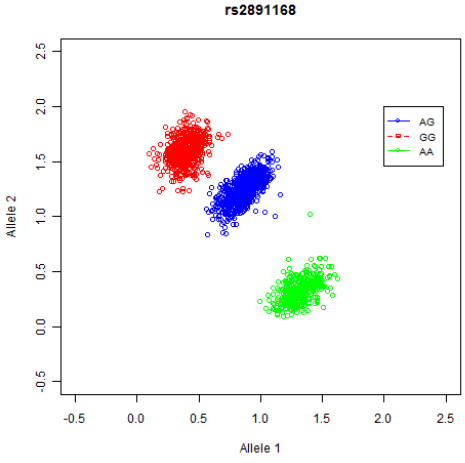



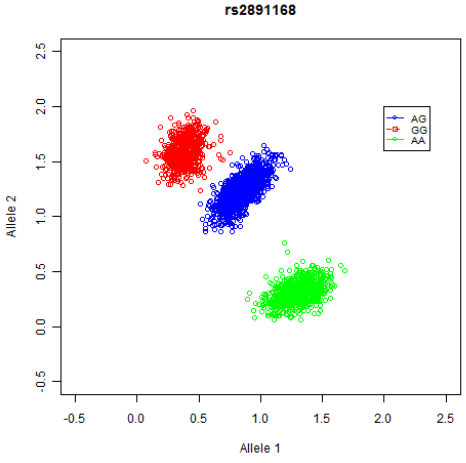



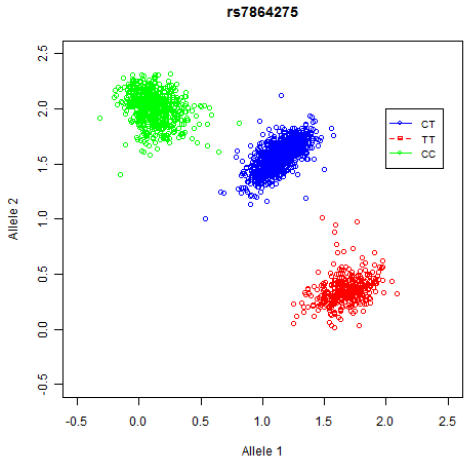



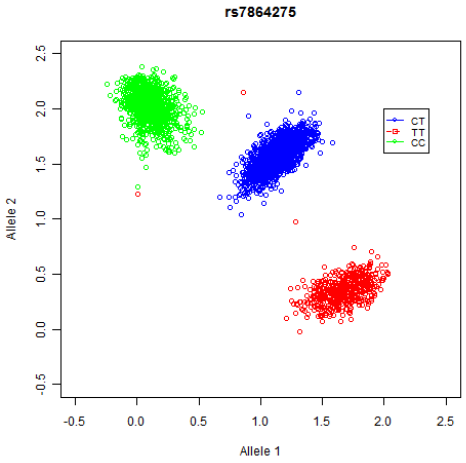

**CAD Cases**

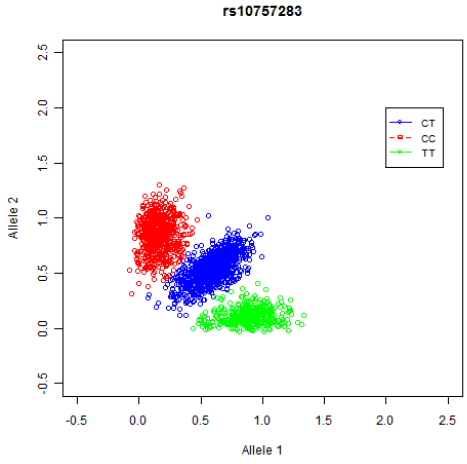
**CAD Controls**

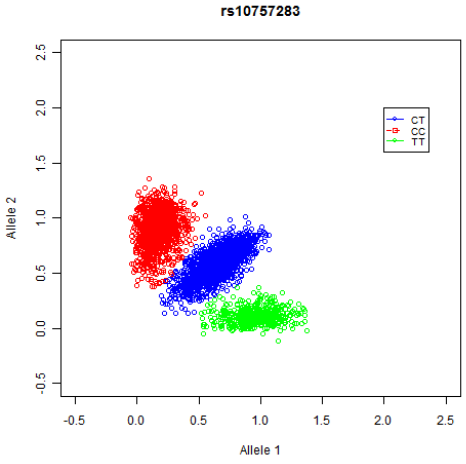


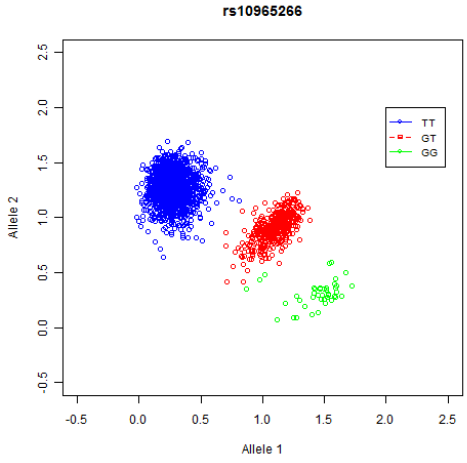



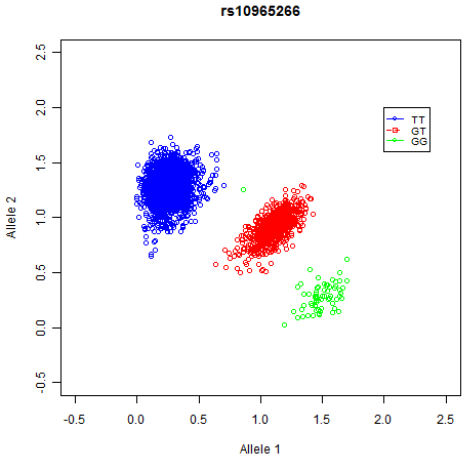


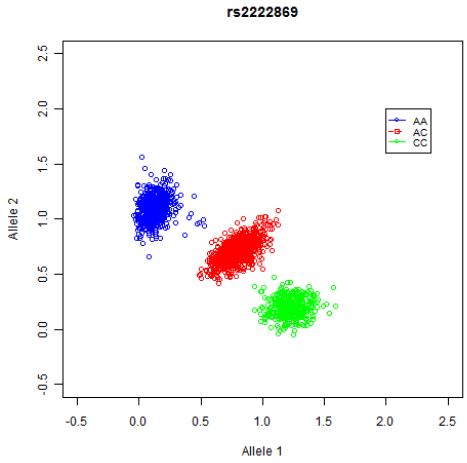



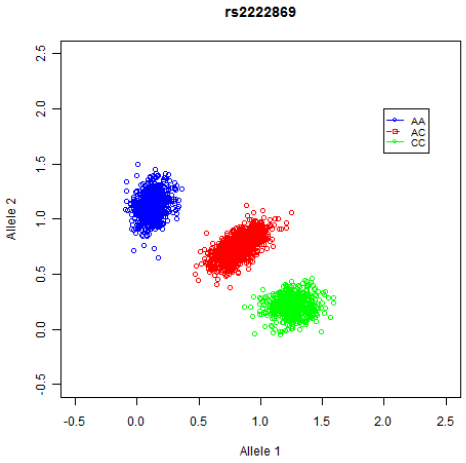



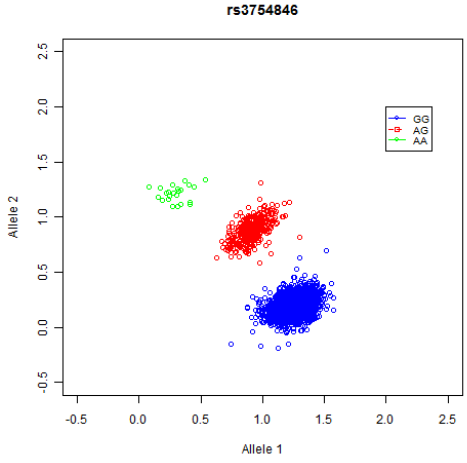



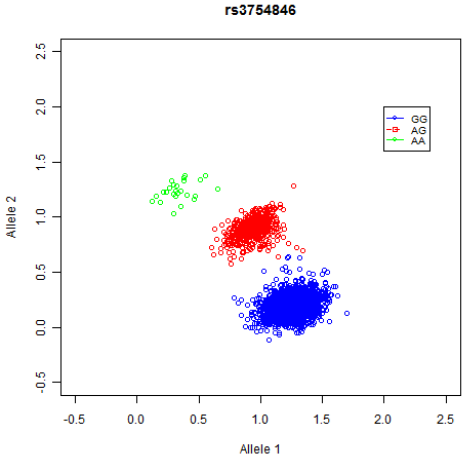



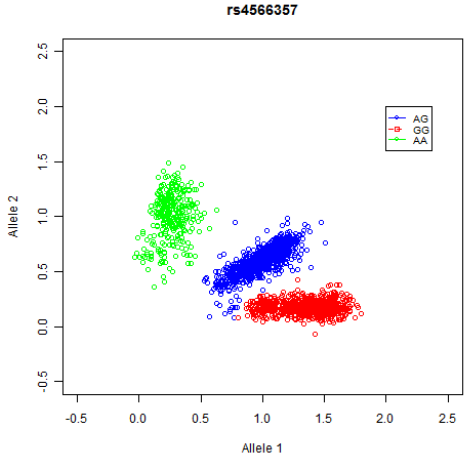

**HT Cases** **HT Controls**

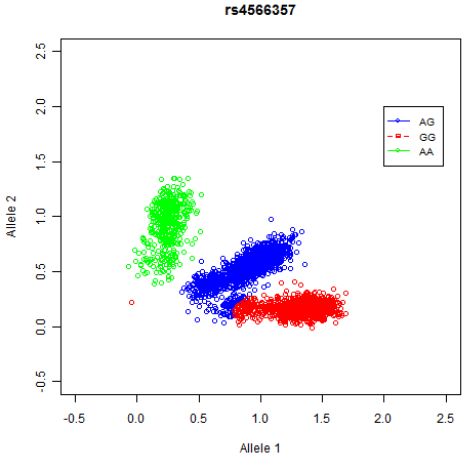

rs2222869
AA
AC
CC
Allele 2
Allele 1

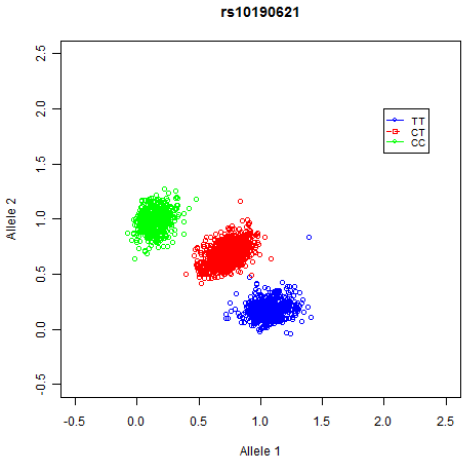

rs2222869
AA
AC
CC
Allele 2
Allele 1

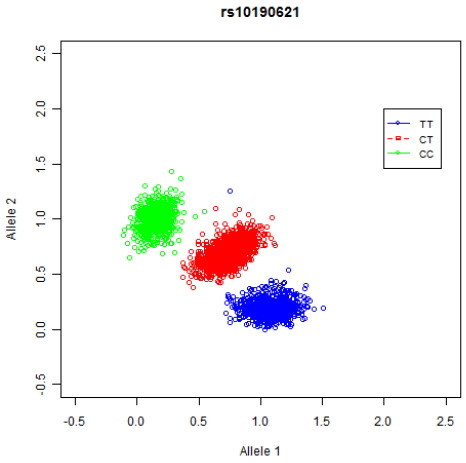

rs3754846
GG
AG
AA
Allele 2
Allele 1


rs3754846
GG
AG
AA
Allele 2
Allele 1


rs4566357
AG
GG
AA
Allele 2
Allele 1

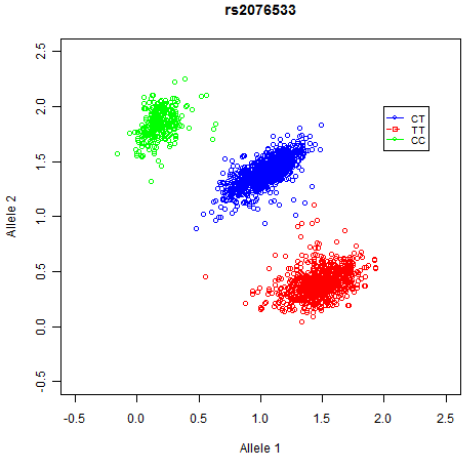

rs4566357
AG
GG
AA
Allele 2
Allele 1

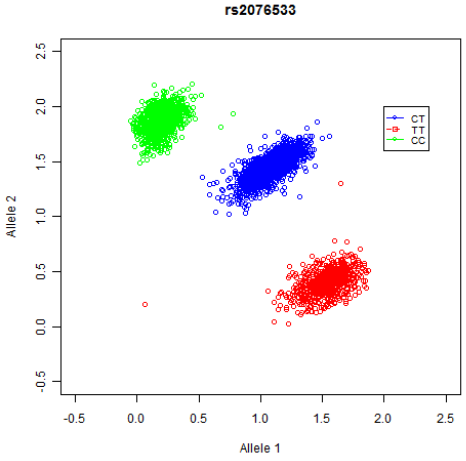


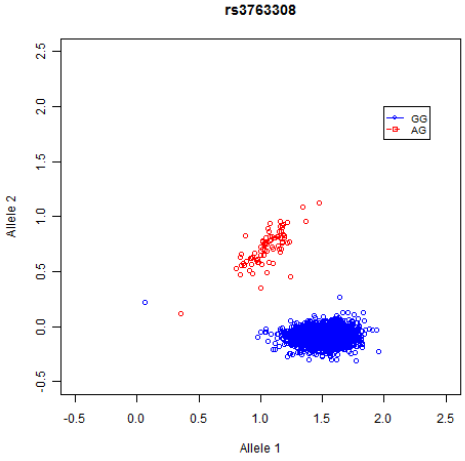



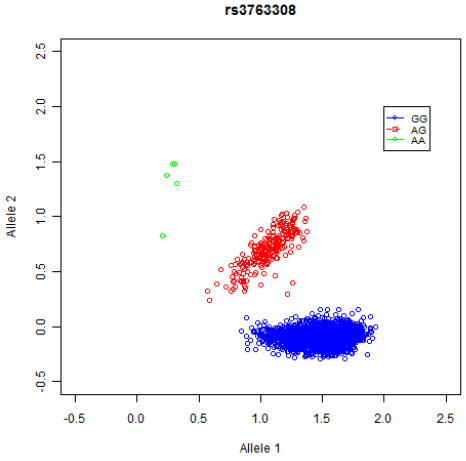

**RA Cases**

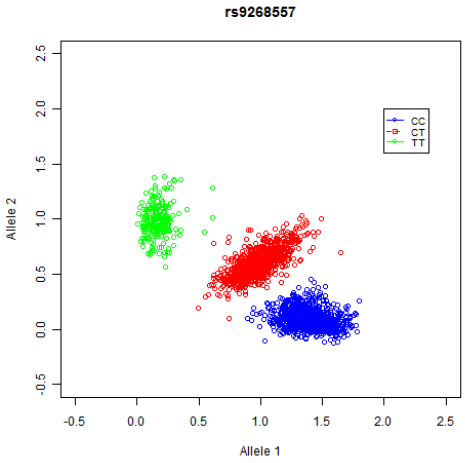
**RA Controls**

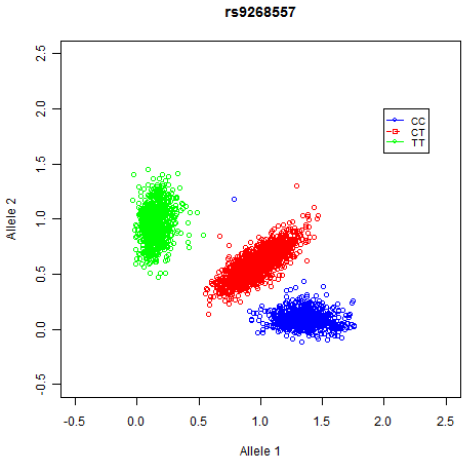


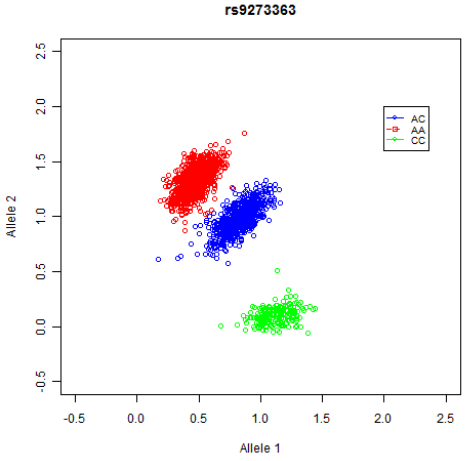



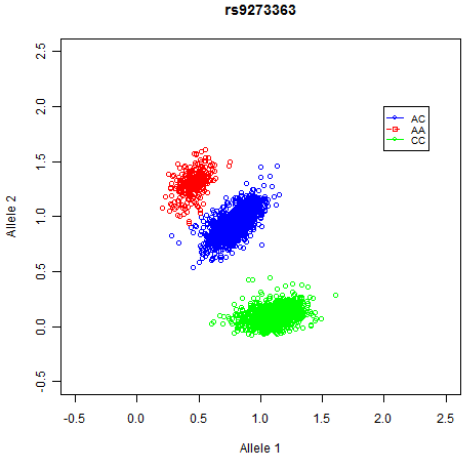


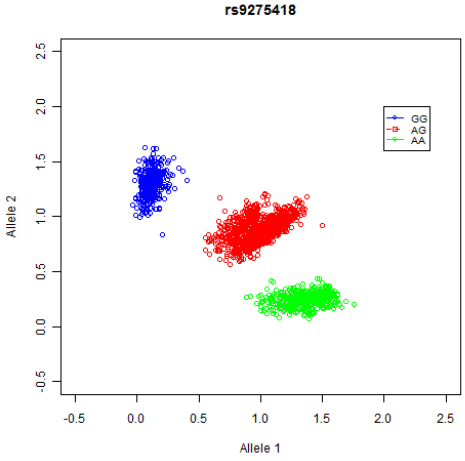



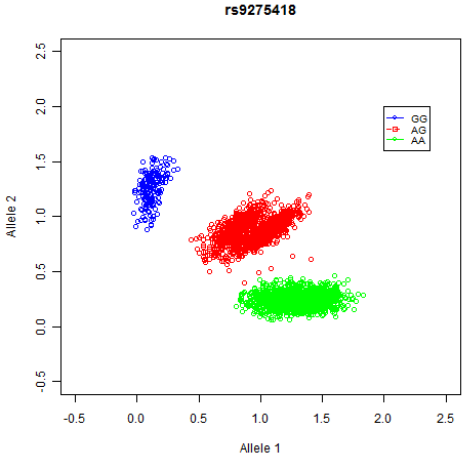

**T1D Cases**

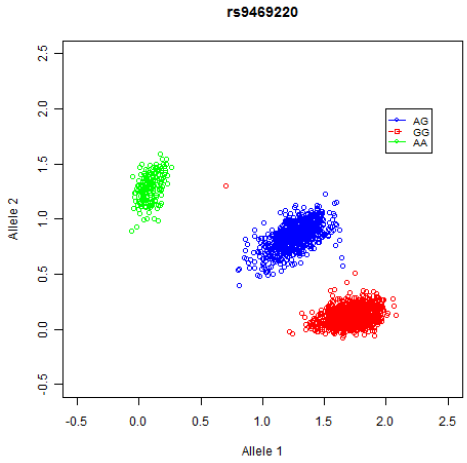
**T1D Controls**

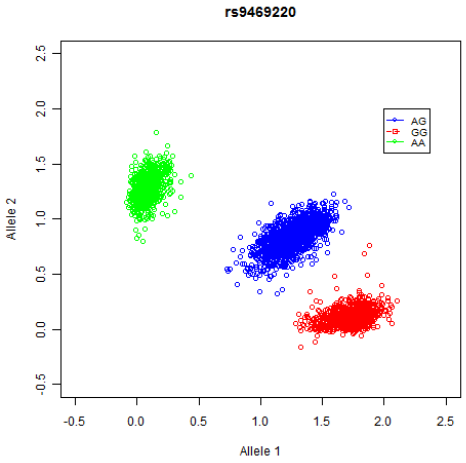


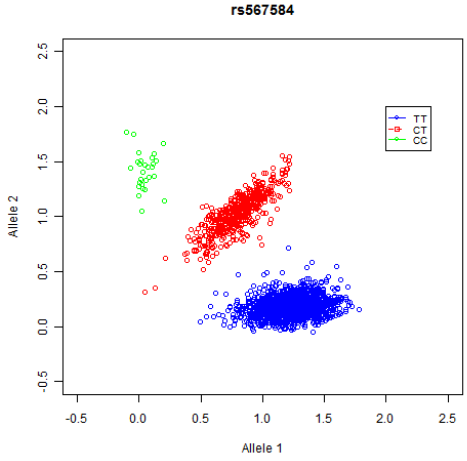



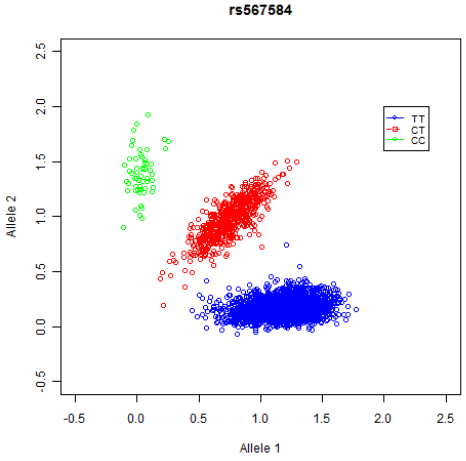


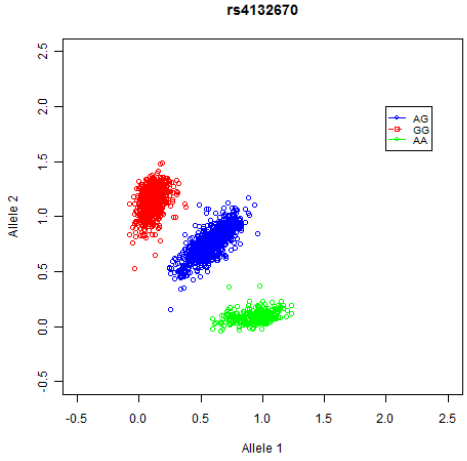



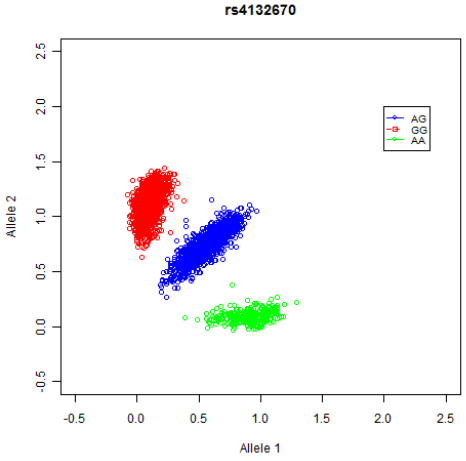

**T2D Cases**

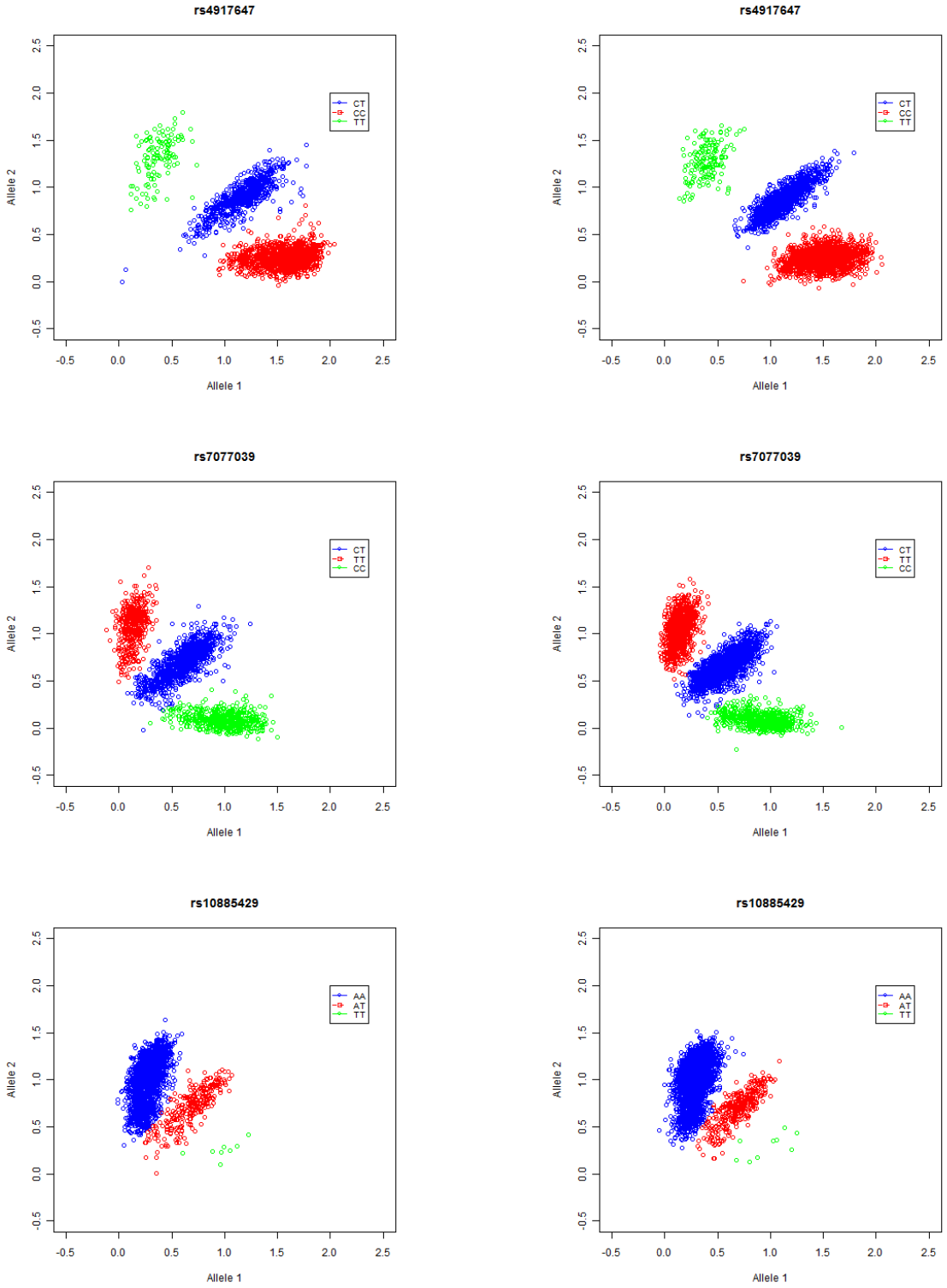
**T2D Controls**



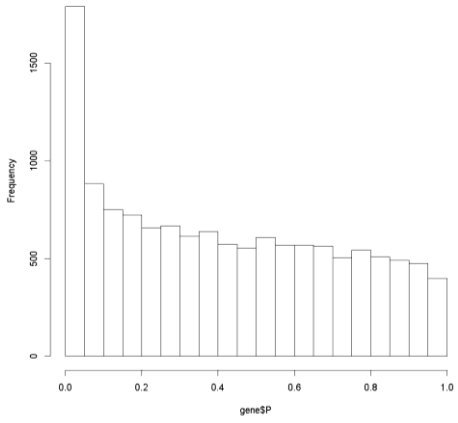



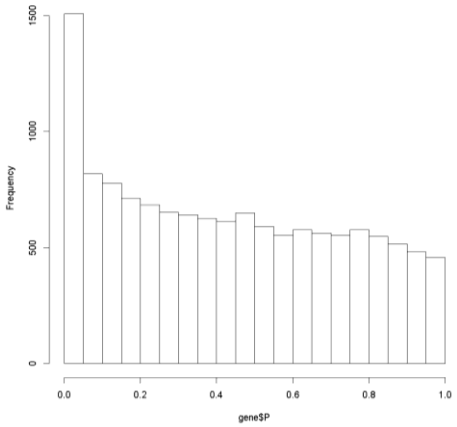



**Supplementary Figure 2.2:** Genotype cluster plots for SNPs in the logic trees of the top significant genes by disease.

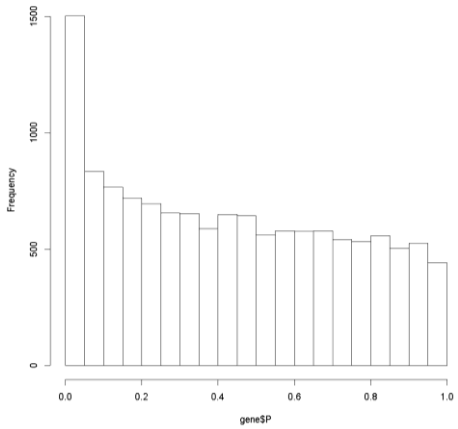
**Bipolar Disorder**



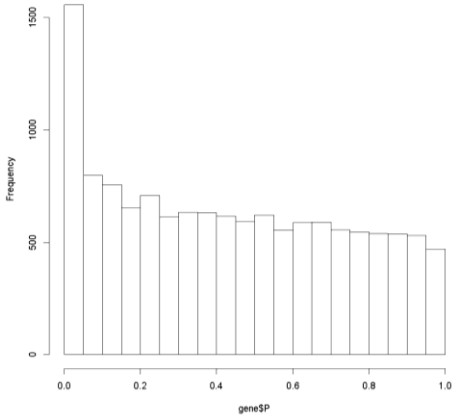

**Coronary Artery Disease**



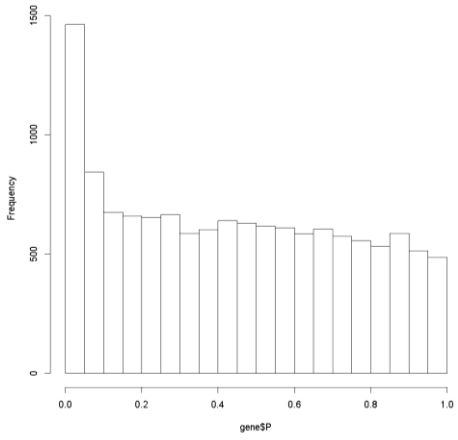

**Hypertension**



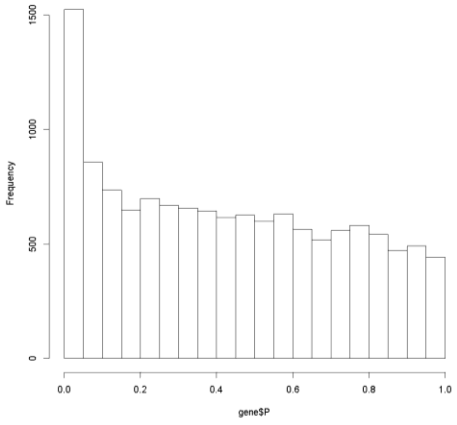

**Rheumatoid Arthritis**



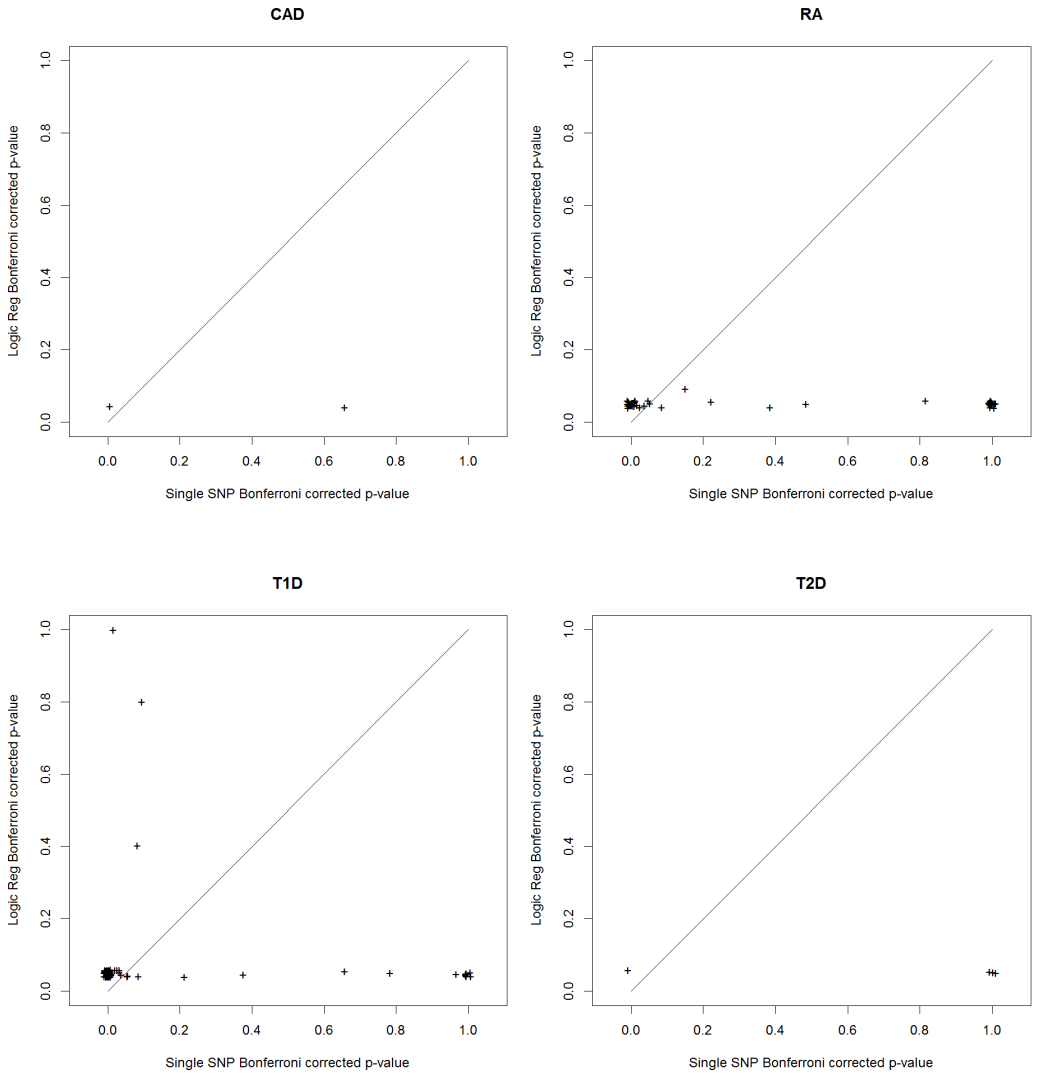

**Type I Diabetes**




**Type II Diabetes**




**Supplementary Figure 2.3** Histograms of the empirical p-value distributions for the six diseases. The theoretical distribution of p-values in our test is a mixture of the uniform distribution in [0,1], corresponding to the null genes, and a right-skewed distribution with a peak at zero, corresponding to the non-null susceptibility genes: the shape of the latter depends on the number of non-null susceptibility genes and the study power to detect them. The histograms of p-values allow assessment of the consistency of the empirical distributions of p-values with the theoretical mixture distribution.




**Supplementary Figure 2.4** Plots of Bonferroni corrected p-values from the single-SNP analysis and our SNP-set interaction analysis in the abscissa and ordinate, respectively. For a given gene, the abscissa plotted the smallest p-value of all SNPs in the gene, testing each SNP's dominant, recessive, and co-dominant associations and selecting the most significant association. All genes whose Bonferroni corrected p-values were < 0.1 by either of the two tests were plotted. These plots indicate that the majority of the significant genes detected from our SNP-set interaction analysis were at the bottom of the plots but not necessarily near the origin, showing greater numbers of significant signals by our SNP-interaction approach, under the same criterion of Bonferroni correction applied to both methods. Note that Bipolar disorder and Hypertension had no SNPs with a Bonferroni corrected p-value <0.1 by either of the two tests.

**Supplementary Table 2.1: Significant genes showing the strongest evidence of association with Bipolar Disease (N=13 genes)**

| Chromosome Location | Gene Name | #SNPs | Bayes Factor (BF) | P-value |
|---|---|---|---|---|
| 1p31.3-p31.2 | *NFIA* | 113 | 6.20 | $1.15 \times 10^{-5}$ |
| 1q44 | *NLRP3* | 9 | 5.87 | $1.53 \times 10^{-5}$ |
| 12q15 | *PTPRR* | 42 | 5.52 | $4.97 \times 10^{-5}$ |
| 2q12-q21 | *DBI* | 4 | 5.24 | $6.88 \times 10^{-5}$ |
| 5q15 | *RIOK2* | 36 | 5.06 | $8.79 \times 10^{-5}$ |
| 1p22.1 | *GCLM* | 7 | 5.00 | $9.17 \times 10^{-5}$ |
| 8p23-p22 | *BLK* | 24 | 4.70 | 0.000198701 |
| 2q37.3 | *COPS8* | 32 | 4.60 | 0.000248376 |
| 21q22.3 | *B3GALT5* | 18 | 4.58 | 0.000256018 |
| 12q24 | *P2RX7* | 13 | 4.56 | 0.000256018 |
| 11q14.1 | *TENM4* | 299 | 4.27 | 0.000420328 |
| 7q36.1 | *GIMAP2* | 7 | 4.14 | 0.000576997 |
| 20p13 | *CDC25B* | 6 | 4.13 | 0.000596102 |

**Supplementary Table 2.2: Significant genes showing the strongest evidence of association with Coronary Artery Disease (N=16 genes)**

| Chromosome Location | Gene Name | #SNPs | Bayes Factor (BF) | P-value |
|---|---|---|---|---|
| 9p21 | *CDKN2B* | 24 | 10.85 | $<3.82 \times 10^{-6}$ |
| 11p15.3-p14 | *TPH1* | 4 | 7.23 | $<3.82 \times 10^{-6}$ |
| 11p14.3 | *USH1C* | 22 | 5.69 | $1.91 \times 10^{-5}$ |
| 9p13.3 | *RECK* | 13 | 5.58 | $2.67 \times 10^{-5}$ |
| 3p21.31 | *CDCP1* | 26 | 5.46 | $3.44 \times 10^{-5}$ |
| 1p22 | *TTF2* | 5 | 5.38 | $4.96 \times 10^{-5}$ |
| 5q31.3 | *SPRY4* | 21 | 5.31 | $5.34 \times 10^{-5}$ |
| 2q37.1 | *RNF7* | 7 | 5.20 | $7.63 \times 10^{-5}$ |
| 2q35 | *ALLC* | 143 | 4.98 | $9.92 \times 10^{-5}$ |
| 2p11.2 | *DNAH6* | 5 | 4.97 | $9.92 \times 10^{-5}$ |
| 10q25.1-q25.2 | *ACSL5* | 8 | 4.65 | 0.000171768 |
| 11q24 | *SIAE* | 4 | 4.54 | 0.000240476 |
| 5p13.1 | *MROH2B* | 14 | 4.52 | 0.000248111 |
| 15q22.3-q23 | *NEO1* | 18 | 4.49 | 0.00027483 |
| 3p12.3 | *GBE1* | 123 | 4.24 | 0.000473319 |
| 6p21 | *PPARD* | 11 | 4.20 | 0.000507672 |

**Supplementary Table 2.3: Significant genes showing the strongest evidence of association Hypertension (N=15 genes)**

| Chromosome Location | Gene Name | #SNPs | Bayes Factor (BF) | P-value |
|---|---|---|---|---|
| 2q35-q37 | *COL4A4* | 10 | 5.40 | $4.58 \times 10^{-5}$ |
| 15q14 | *GJD2* | 28 | 5.34 | $4.58 \times 10^{-5}$ |
| 2q11.2-q12.1 | *ST6GAL2* | 77 | 5.19 | $6.49 \times 10^{-5}$ |
| 15q22.31 | *MTFMT* | 3 | 5.18 | $6.49 \times 10^{-5}$ |
| 4p16.1 | *BOD1L1* | 107 | 5.03 | $7.63 \times 10^{-5}$ |
| 11q13.4 | *PGM2L1* | 10 | 4.89 | 0.000114513 |
| 20p13 | *DEFB129* | 7 | 4.67 | 0.000175586 |
| 7q21.11 | *SEMA3E* | 50 | 4.56 | 0.000213757 |
| 12q23.2 | *MYBPC1* | 29 | 4.55 | 0.000225208 |
| 3p24.1 | *ZCWPW2* | 44 | 4.50 | 0.000248111 |
| 6q14-q15 | *TBX18* | 61 | 4.40 | 0.000305366 |
| 10p11.21 | *CREM* | 5 | 4.34 | 0.000362623 |
| 2q11.2 | *NCAPH* | 2 | 4.24 | 0.000450416 |
| 3p24.1 | *EOMES* | 41 | 4.21 | 0.000477136 |
| 14q32.33 | *C14orf80* | 3 | 4.16 | 0.000519124 |

**Supplementary Table 2.4: Significant genes showing the strongest evidence of association with Rheumatoid Arthritis (N=72 genes)**

| Chromosome Location | Gene Name | #SNPs | Bayes Factor (BF) | P-value |
|---|---|---|---|---|
| 6p21.3 | *BTNL2* | 10 | 95.34 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *HLA-DRA* | 12 | 93.87 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *C6orf10* | 13 | 91.71 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *HLA-DQB1* | 7 | 82.23 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *NOTCH4* | 15 | 64.91 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *HLA-DRB1* | 2 | 45.91 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *HLA-DQA1* | 4 | 45.15 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *TNXB* | 3 | 31.67 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *TAP2* | 7 | 22.99 | $<3.82 \times 10^{-6}$ |
| 1p13.2 | *RSBN1* | 4 | 22.72 | $<3.82 \times 10^{-6}$ |
| 6p21.33 | *APOM* | 2 | 21.34 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *HLA-DOA* | 16 | 20.08 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *PRRC2A* | 3 | 19.89 | $<3.82 \times 10^{-6}$ |
| 1p12-p11.2 | *MAGI3* | 21 | 17.55 | $<3.82 \times 10^{-6}$ |
| 6q12-q13 | *RIMS1* | 75 | 16.18 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *BAG6* | 6 | 15.69 | $<3.82 \times 10^{-6}$ |

| | | | | |
|---|---|---|---|---|
| 3p13 | *GXYLT2* | 9 | 14.49 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *HLA-B* | 15 | 13.60 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *HCP5* | 19 | 13.24 | $<3.82 \times 10^{-6}$ |
| 6p21.33 | *MICA* | 11 | 13.07 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *HLA-DPA1* | 7 | 10.45 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *COL11A2* | 12 | 10.45 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *TRIM26* | 8 | 10.34 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *HLA-F* | 18 | 10.30 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *HLA-DPB1* | 11 | 10.13 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *C6orf15* | 12 | 9.92 | $<3.82 \times 10^{-6}$ |
| 14q23.1 | *DACT1* | 52 | 9.90 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *TRIM31* | 7 | 9.28 | $<3.82 \times 10^{-6}$ |
| 11p15 | *ART1* | 3 | 9.21 | $<3.82 \times 10^{-6}$ |
| 6p21 | *SKIV2L* | 2 | 8.78 | $<3.82 \times 10^{-6}$ |
| 6p22.1 | *TRIM40* | 7 | 8.72 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *DDR1* | 7 | 8.43 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *AGPAT1* | 2 | 8.23 | $<3.82 \times 10^{-6}$ |
| 11q13.4 | *LRP5* | 10 | 8.16 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *HLA-A* | 17 | 7.88 | $<3.82 \times 10^{-6}$ |
| 6p21.33 | *VARS* | 3 | 7.59 | $<3.82 \times 10^{-6}$ |
| 1p13.2 | *PTPN22* | 3 | 7.55 | $<3.82 \times 10^{-6}$ |
| 6q23 | *TNFAIP3* | 25 | 7.54 | $<3.82 \times 10^{-6}$ |
| 6p21.33 | *DPCR1* | 3 | 7.48 | $<3.82 \times 10^{-6}$ |
| 6p22.1 | *RPP21* | 18 | 7.39 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *HLA-DMB* | 7 | 7.38 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *TRIM39* | 4 | 7.21 | $<3.82 \times 10^{-6}$ |
| 17q23.2 | *NACA2* | 8 | 7.10 | $<3.82 \times 10^{-6}$ |
| 6p21.32 | *MUC21* | 10 | 6.99 | $<3.82 \times 10^{-6}$ |
| 12q12 | *ADAMTS20* | 50 | 6.30 | $3.82 \times 10^{-6}$ |
| 6p21.3 | *BAK1* | 6 | 6.16 | $3.82 \times 10^{-6}$ |
| 6p22.1 | *FLJ45422* | 7 | 6.11 | $3.82 \times 10^{-6}$ |
| 6p21.3 | *UBD* | 6 | 6.02 | $3.82 \times 10^{-6}$ |
| 1p13 | *PHTF1* | 2 | 5.92 | $7.64 \times 10^{-6}$ |
| 6p21.33 | *LY6G6C* | 3 | 5.60 | $2.29 \times 10^{-5}$ |
| 6p22.1 | *ZFP57* | 3 | 5.59 | $2.29 \times 10^{-5}$ |
| 6p22 | *TRIM27* | 9 | 5.36 | $2.68 \times 10^{-5}$ |
| 6p21.31 | *UQCC2* | 7 | 5.31 | $3.06 \times 10^{-5}$ |
| 2p24 | *ASAP2* | 25 | 5.11 | $5.73 \times 10^{-5}$ |
| 19p13.2-p13.1 | *NOTCH3* | 5 | 5.03 | $6.88 \times 10^{-5}$ |
| 4p15.2 | *ANAPC4* | 14 | 4.88 | 0.000110831 |
| 1p13.2 | *HIPK1* | 2 | 4.82 | 0.00012994 |

| | | | | |
|---|---|---|---|---|
| 1p36.3 | *TP73* | 4 | 4.80 | 0.000133761 |
| 6p21.31 | *IP6K3* | 4 | 4.70 | 0.000171979 |
| 6p21.3 | *RNF39* | 2 | 4.66 | 0.000183444 |
| 2p22-p21 | *THUMPD2* | 18 | 4.63 | 0.000198731 |
| 10p15-p14 | *IL2RA* | 16 | 4.61 | 0.000210196 |
| 1p21.3 | *GPR88* | 10 | 4.54 | 0.000252236 |
| 19p12 | *ZNF254* | 14 | 4.53 | 0.000263701 |
| 1q21.1 | *PRKAB2* | 17 | 4.47 | 0.000290453 |
| 7q32-q33 | *PODXL* | 45 | 4.45 | 0.000309562 |
| 7q22-qter | *CNOT4* | 17 | 4.37 | 0.00037071 |
| 6p21.3 | *KIFC1* | 2 | 4.35 | 0.000378354 |
| 20q13.2-q13.3 | *EDN3* | 37 | 4.31 | 0.000416571 |
| 8p23.2 | *CSMD1* | 746 | 4.28 | 0.000431858 |
| 20p13 | *DEFB129* | 7 | 4.21 | 0.000485363 |
| 20q11.2-q12 | *EPB41L1* | 6 | 4.19 | 0.00050065 |

**Supplementary Table 2.5: Significant genes showing the strongest evidence of association with Type 1 Diabetes (N=105 genes)**

| Chromosome Location | Gene Name | #SNPs | Bayes Factor (BF) | P-value |
|---|---|---|---|---|
| 6p21.3 | *HLA-DQB1* | 7 | Inf | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *HLA-DRA* | 10 | 230.25 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *BTNL2* | 10 | 214.04 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *C6orf10* | 12 | 206.30 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *HLA-DQA1* | 4 | 203.42 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *NOTCH4* | 15 | 177.52 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *HLA-DRB1* | 2 | 163.01 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *TNXB* | 3 | 108.40 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *PRRC2A* | 3 | 94.87 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *HLA-DOB* | 8 | 92.12 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *BAG6* | 6 | 90.17 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *HCP5* | 19 | 78.27 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *MSH5* | 4 | 75.48 | $<3.82 \times 10^{-6}$ |
| 6p21.33 | *MICA* | 11 | 73.98 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *AIF1* | 4 | 73.81 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *C6orf15* | 12 | 57.05 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *MICB* | 5 | 47.59 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *HLA-C* | 18 | 44.60 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *BRD2* | 2 | 43.44 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *DDR1* | 7 | 40.67 | $<3.82 \times 10^{-6}$ |
| 6p21 | *SKIV2L* | 2 | 38.08 | $<3.82 \times 10^{-6}$ |

| | | | | |
|---|---|---|---|---|
| 6p21.3 | *HLA-A* | 16 | 37.74 | $<3.82 \times 10^{-6}$ |
| 6p22.1 | *RPP21* | 18 | 36.22 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *HLA-E* | 7 | 34.52 | $<3.82 \times 10^{-6}$ |
| 6p21.32 | *MUC21* | 10 | 30.85 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *AGPAT1* | 2 | 27.94 | $<3.82 \times 10^{-6}$ |
| 6p21.33 | *TUBB2A* | 2 | 27.80 | $<3.82 \times 10^{-6}$ |
| 6p21.33 | *DPCR1* | 3 | 26.31 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *HLA-DOA* | 16 | 26.31 | $<3.82 \times 10^{-6}$ |
| 1p13.2 | *RSBN1* | 4 | 22.66 | $<3.82 \times 10^{-6}$ |
| 6p21 | *MAS1L* | 7 | 20.34 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *OR2H2* | 5 | 20.04 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *TAP1* | 4 | 19.63 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *TRIM10* | 2 | 19.07 | $<3.82 \times 10^{-6}$ |
| 6p22.1 | *FLJ45422* | 7 | 18.65 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *HLA-G* | 19 | 18.58 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *HLA-F* | 17 | 17.58 | $<3.82 \times 10^{-6}$ |
| 6p22.1 | *OR10C1* | 3 | 16.77 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *TRIM26* | 8 | 16.47 | $<3.82 \times 10^{-6}$ |
| 6p22.1 | *OR14J1* | 3 | 14.48 | $<3.82 \times 10^{-6}$ |
| 6p22.1 | *ZNF311* | 4 | 14.22 | $<3.82 \times 10^{-6}$ |
| 6p22.2-p21.31 | *OR12D2* | 4 | 14.12 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *ZNF165* | 7 | 13.73 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *ZKSCAN8* | 5 | 13.50 | $<3.82 \times 10^{-6}$ |
| 6p22.1 | *OR12D3* | 3 | 13.12 | $<3.82 \times 10^{-6}$ |
| 6p22.1 | *HIST1H2BL* | 7 | 13.12 | $<3.82 \times 10^{-6}$ |
| 6p22.1 | *BTN3A2* | 8 | 12.64 | $<3.82 \times 10^{-6}$ |
| 6p21.31 | *GABBR1* | 6 | 12.57 | $<3.82 \times 10^{-6}$ |
| 6p22.1 | *OR2J3* | 4 | 12.32 | $<3.82 \times 10^{-6}$ |
| 6p22.1 | *HIST1H2BJ* | 10 | 12.23 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *ZSCAN9* | 4 | 11.97 | $<3.82 \times 10^{-6}$ |
| 6p21.33 | *ATAT1* | 2 | 11.82 | $<3.82 \times 10^{-6}$ |
| 12q24.13 | *NAA25* | 5 | 11.72 | $<3.82 \times 10^{-6}$ |
| 6p22.1 | *BTN2A1* | 6 | 11.52 | $<3.82 \times 10^{-6}$ |
| 6p22.2-p21.31 | *OR2J2* | 2 | 11.13 | $<3.82 \times 10^{-6}$ |
| 6p22.1 | *OR2W1* | 4 | 11.12 | $<3.82 \times 10^{-6}$ |
| 6p22.1 | *ZSCAN16* | 3 | 11.04 | $<3.82 \times 10^{-6}$ |
| 6p21.33 | *VARS* | 3 | 11.01 | $<3.82 \times 10^{-6}$ |
| 1p13.2 | *PTPN22* | 3 | 10.57 | $<3.82 \times 10^{-6}$ |
| 6p22.1 | *HIST1H4H* | 3 | 9.86 | $<3.82 \times 10^{-6}$ |
| 6p21.33 | *PSORS1C3* | 2 | 9.18 | $<3.82 \times 10^{-6}$ |
| 12q13 | *ERBB3* | 3 | 9.08 | $<3.82 \times 10^{-6}$ |

| | | | | |
|---|---|---|---|---|
| 6p21.3 | *TCF19* | 4 | 8.83 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *DAXX* | 2 | 8.43 | $<3.82 \times 10^{-6}$ |
| 1p13 | *PHTF1* | 2 | 7.90 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *ZNF184* | 5 | 7.51 | $<3.82 \times 10^{-6}$ |
| 12q24.13 | *TMEM116* | 4 | 7.26 | $<3.82 \times 10^{-6}$ |
| 6p22.3-p22.1 | *ZSCAN31* | 6 | 7.16 | $<3.82 \times 10^{-6}$ |
| 6p22.1 | *VN1R10P* | 9 | 7.14 | $<3.82 \times 10^{-6}$ |
| 12q13.2 | *SUOX* | 2 | 7.07 | $<3.82 \times 10^{-6}$ |
| 6p22.1 | *LOC651503* | 2 | 6.99 | $<3.82 \times 10^{-6}$ |
| 12q13 | *RAB5B* | 3 | 6.95 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *BTN3A3* | 2 | 6.94 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *SLC17A3* | 10 | 6.90 | $<3.82 \times 10^{-6}$ |
| 6p21.3 | *SYNGAP1* | 2 | 6.57 | $<3.82 \times 10^{-6}$ |
| 6p21.31 | *POU5F1* | 3 | 6.56 | $<3.82 \times 10^{-6}$ |
| 1p13.2 | *BCL2L15* | 2 | 5.88 | $1.14 \times 10^{-5}$ |
| 6p21.3 | *MLN* | 20 | 5.80 | $1.14 \times 10^{-5}$ |
| 6p21 | *ITPR3* | 10 | 5.70 | $1.91 \times 10^{-5}$ |
| 1p13.2 | *HIPK1* | 2 | 5.60 | $3.05 \times 10^{-5}$ |
| 6p21 | *PRSS16* | 5 | 5.54 | $3.05 \times 10^{-5}$ |
| 1p13.2 | *AP4B1* | 3 | 5.43 | $4.20 \times 10^{-5}$ |
| 3p21.31 | *CCR2* | 3 | 5.36 | $4.20 \times 10^{-5}$ |
| 12p13 | *CLEC2D* | 12 | 5.19 | $6.11 \times 10^{-5}$ |
| 6p22.1 | *GPX5* | 3 | 5.16 | $6.87 \times 10^{-5}$ |
| 12q24 | *SH2B3* | 4 | 5.13 | $8.01 \times 10^{-5}$ |
| 10p15-p14 | *IL2RA* | 16 | 5.08 | $8.01 \times 10^{-5}$ |
| 12q24 | *PTPN11* | 8 | 5.04 | $9.54 \times 10^{-5}$ |
| 6p21.3 | *DHX16* | 2 | 4.92 | 0.000129761 |
| 12q24.2 | *ALDH2* | 3 | 4.86 | 0.00015266 |
| 6p21.3 | *UBD* | 6 | 4.85 | 0.00015266 |
| 12q24.11 | *MYL2* | 3 | 4.81 | 0.000167926 |
| 6p21.3 | *HIST1H1E* | 2 | 4.78 | 0.000179376 |
| 12q14.1 | *KCNC2* | 91 | 4.76 | 0.000187009 |
| 16p13.13 | *CLEC16A* | 20 | 4.69 | 0.000206091 |
| 3p21.31 | *FLJ78302* | 6 | 4.63 | 0.000217541 |
| 4q27 | *ADAD1* | 3 | 4.53 | 0.000270972 |
| 1p31 | *SERBP1* | 23 | 4.52 | 0.000270972 |
| 12q24.12 | *ACAD10* | 3 | 4.50 | 0.000270972 |
| 4q12 | *POLR2B* | 6 | 4.49 | 0.000270972 |
| 3p21.3 | *CCR3* | 8 | 4.47 | 0.000282421 |
| 4q27 | *KIAA1109* | 5 | 4.29 | 0.000435081 |
| 3p25.2 | *TSEN2* | 10 | 4.20 | 0.000488512 |

| | | | | |
|---|---|---|---|---|
| 12q24 | *BRAP* | 2 | 4.20 | 0.000488512 |
| 6p22.1 | *HIST1H3I* | 3 | 4.15 | 0.000572475 |

**Supplementary Table 2.6: Significant genes showing the strongest evidence of association with Type 2 Diabetes (N=19 genes)**

| Chromosome Location | Gene Name | #SNPs | P-value | Bayes Factor (BF) |
|---|---|---|---|---|
| 10q25.3 | *TCF7L2* | 38 | $<3.82 \times 10^{-6}$ | 8.69 |
| 4q27 | *TMEM155* | 8 | $<3.82 \times 10^{-6}$ | 7.58 |
| 5q15 | *FAM172A* | 12 | $<3.82 \times 10^{-6}$ | 6.96 |
| 16q13 | *GPR56* | 12 | $<3.82 \times 10^{-6}$ | 6.20 |
| 13q12.12 | *C1QTNF9* | 4 | 0.000122 | 5.07 |
| 6p12.1 | *GFRAL* | 9 | 0.000122 | 5.06 |
| 10q11.22-q11.23 | *ZNF239* | 10 | 0.000141 | 4.99 |
| 9q34.3 | *PPP1R26* | 22 | 0.000203 | 4.81 |
| 6p21.1 | *TNFRSF21* | 28 | 0.000233 | 4.72 |
| 1q24 | *NME7* | 10 | 0.000252 | 4.67 |
| 12q21.2 | *BBS10* | 18 | 0.000252 | 4.67 |
| 2q24.2 | *RBMS1* | 31 | 0.000264 | 4.65 |
| 9p21.3 | *PTPLAD2* | 8 | 0.000355 | 4.49 |
| 12p11.22 | *OVCH1* | 12 | 0.000394 | 4.43 |
| 1q31.3 | *KCNT2* | 123 | 0.000489 | 4.28 |
| 11q23 | *ABCC8* | 14 | 0.000615 | 4.17 |
| 8p11.22 | *PLEKHA2* | 12 | 0.000623 | 4.15 |
| 11p15 | *TRIM22* | 6 | 0.000631 | 4.14 |
| 16q12.2 | *FTO* | 69 | 0.000646 | 4.12 |

# CHAPTER 3

## A Candidate Pathway Approach Identifies Multiple Gene-Environment Interactions in Association with Colon Cancer Risk and Survival

### 3.1 Introduction

Colon cancer is a good example of a multi-factorial disease with well-documented genetic and non-genetic risk factors (Potter1999a). Several lines of evidence indicate a prominent role of dietary and lifestyle factors in colon cancer etiology including wide geographical variations in incidence across countries (International Agency for Research on Cancer2008), and migrant populations, especially of Asian descent, moving from low-risk to high risk countries adopting the host country's high levels of risk (Marchand1999,Flood et al. 2000). Additional evidence comes from Japan, a country with historically one of the lowest incidence of colon cancer becoming one of the highest incidence in the world over several decades (Oba et al. 2006,Takachi et al. 2011,Potter1999a). Although evidence on lifestyle environmental exposures effects on colon cancer survival is limited, some evidence suggests pre- and/or post-diagnostic dietary patterns, physical activity, smoking, and alcohol consumption may have an impact on CRC mortality (Pelser et al. 2014).

Considerable research efforts have been made towards identifying highly- and moderately-penetrant rare variants in association with colon cancer, and more recently common low-penetrance risk alleles through genome-wide association studies (GWAS), yet with limited success (Whiffin et al. 2014). This has enforced the hypothesis that the large unexplained hereditary component of colon cancer risk referred to as "*missing heritability*" may be partially explained by epistatic and/or gene-environment interactions (GEIs) (Manolio et al. 2009). A

standard “marginal” approach of analysis in genetic association studies that does not take into account the possibility of interaction between individual genetic variants by analyzing single nucleotide polymorphisms (SNPs) one at a time will, therefore, either fail to observe or detect weak associations. This practice ignores the inherent coordination between genes better described by a pathway structure composed of multiple genes with related biologic functions contributing to risk in different environmental contexts (Kraft et al. 2009).

It is essential to focus on a biological pathway relevant to the disease, and environmental exposures relevant to the pathway. One of the genetic pathways of special interest to colon cancer outcomes is the angiogenesis pathway which mediates the process of sprouting of blood vessels from existing ones, allowing for tumor growth and progression. An ischemic tumor microenvironment with poor oxygen and nutrient supply is an important trigger of the angiogenesis process (Folkman et al. 1992). Several proteins are involved in tumor angiogenesis including the vascular endothelial growth factor (VEGF) which acts as one of the most potent angiogenic factors (Ferrara1999, Lohela et al. 2009). Another important factor that regulates gene expression in angiogenesis is the hypoxia-inducible factor 1 (HIF-1) (Semenza 2010). Activation of the *HIF-1α* signaling pathway under glucose deprivation has also been shown recently to lead to colon cancer cells acquiring anti-apoptosis properties (Nishimoto et al. 2014). We selected three environmental exposures with evidence of associations with colon cancer and relevant to the angiogenesis pathway: dietary protein intake; cigarette smoking; and alcohol consumption (Gonzalez et al. 2010, Poynter et al. 2009, Cheng et al. 2014). We hypothesized the three environmental exposures are biologically stimulating tumor angiogenesis under ischemic conditions (hypoxia and hypoglycemia) (Vigne et al. 2006, Wong et al. 2007,Gu et al. 2001).

In this study we examined GEIs of the angiogenesis gene-pathway and the three environmental factors in association with both colon cancer susceptibility and survival. We applied a candidate-pathway approach that considered gene rather than individual SNP effects, and selected the candidate genes and the environmental variables based on biologic hypothesis. We also emphasized biologic plausibility in the form the SNP-set interactions within each gene might take. Our candidate-pathway approach involved three steps that summarized the individual gene and gene-gene interaction effects in the first two steps and modeled pathway GEIs in the third step.

## 3.2 Methods

### 3.2.1 Data Sources

***Study population***

The "Diet, Activity and Lifestyle as a Risk Factor for Colon Cancer" is a multicenter, population-based, case-control study of colon cancer conducted at three geographical areas in the United States: Utah, Northern California, and Minnesota (Slattery et al. 1997a). Colon cancer cases were identified using a rapid-reporting system during the period between October 1991 and September 1994 with the majority of cases interviewed within four months of diagnosis. Final case eligibility was determined by the Surveillance Epidemiology and End Results (SEER) Cancer Registries in Northern California and Utah for California and Utah study participants, respectively, and through the Minnesota Tumor Registry for study participants identified in the Twin Cities Area of Minnesota. Eligible cases were 30 to 79 years old at time of diagnosis, English speaking, and mentally and physically competent to complete the interview. Cases with a previous history of colorectal cancer or known familial adenomatous polyposis, ulcerative

colitis, or Crohn's disease (as indicated on pathology reports) were not eligible. Controls were frequency matched to cases by sex and 5-year age groups in each geographical area.

***Interview data***

A detailed in-person interview was conducted by trained and certified interviewers using laptop computers (Edwards et al. 1994). The interview took approximately two to three hours and consisted of two parts: a) the health and lifestyle questionnaire (including data on demographic characteristics, medical history, meal patterns, smoking and alcohol consumption among other information); and b) a diet history questionnaire adapted from the validated CARDIA diet history to be used as a computer-assisted questionnaire in case-control studies (Slattery et al. 1994,Liu et al. 1994). The referent period for the study questionnaires was the calendar year two to three years prior to diagnosis for cases or to selection for controls.

***Tumor registry data***

Data obtained from local tumor registries were used to determine disease stage at diagnosis, months of survival after diagnosis, and vital status. Disease stage was categorized using the SEER staging criteria (in-situ, local, regional, distant, and unknown) (Young et al. 2001). The follow-up time was considered from the date of diagnosis up to date of last follow-up or death. Follow-up was terminated at the end of the year 2000 and all study participants had over five years of follow-up.

***TagSNP selection and genotyping***

TagSNPs were selected using the following parameters: LD blocks using a Caucasian LD map (International HapMap Consortium2003) and $r^2 \geq 0.8$; minor allele frequency (MAF)>0.1; LD

block range= -1500 bps from the initiation codon to +1500 bps from the termination codon; and 1 tagSNP for each LD bin. All markers were genotyped using a multiplexed bead-array assay format based on Golden Gate chemistry (Illumina Human Hap550k, San Diego, California). A genotyping call rate of 99.85% was achieved. Blinded internal duplicates represented 4.4% of the total sample set; the duplicate concordance rate was 100%. *TGFβ1* gene was not included in the Illumina BeadChip platform; alternatively representative markers were genotyped using a TaqMan assay from Applied Biosystems (Foster City, California). Each 5µl PCR reaction contained 20 ng of genomic DNA, primers, probes, and TaqMan Universal PCR Master Mix (containing AmpErase UNG, AmpliTaq Gold enzyme, dNTPs, and reaction buffer). PCR was carried out under the following conditions: 50$^{o}$C for 2 minutes to activate UNG, 95$^{o}$C for 10 min, followed by 40 cycles of 92$^{o}$C for 15 sec, and 60$^{o}$C for 1 minute using 384 well duel block ABI 9700. Fluorescent endpoints of the TaqMan reactions were measured using a 7900HT sequence detection instrument. Individuals with missing genotype data were not included in the analysis for that specific marker.

***Candidate Gene-Pathway***

**Figure 3.1** shows the components of the angiogenesis pathway that we hypothesized to be relevant to colon cancer. We constructed it using several sources of information in an attempt to include important and relevant genes to best represent the angiogenesis pathway in relation to colon cancer. We extracted information from three recognized web-based resources using the search term "angiogenesis":

- The BioCarta Pathways: "VEGF, Hypoxia, and Angiogenesis Pathway" (http://www.biocarta.com/pathfiles/h_vegfPathway.asp)

- KEGG (Kyoto Encyclopedia of Genes and Genomes) Pathway database: "VEGF Signaling Pathway" available from the KEGG (http://www.kegg.jp/kegg-bin/highlight_pathway?scale=1.0&map=map04370&keyword=angiogenesis)
- Cell Signaling Technologies Pathways: the "Angiogenesis Signaling Pathway" from the CST pathways (http://www.cellsignal.com/common/content/content.jsp?id=pathways-angiogenesis) (See Supplementary Figures 1 - 3).

We supplemented the information by reviewing available evidence on the biologic activity and function of the candidate genes and on experimental observations of biologic activities of the genes in relation to tumor angiogenesis through online gene databases and PubMed. The working pathway figure was used as a guide to the analysis, and genes were described as either major drivers of the angiogenesis process or interacting inflammatory genes (**Table 3.1**).

***Environmental Variables***

*Smoking:*

Smoking status was based on regular cigarette smoking defined as smoking at least 100 cigarettes during a lifetime. For smokers, the total years of smoking were determined by taking into account start and stop dates of smoking. Pack-years of cigarettes smoked were used to represent patterns of smoking and derived by multiplying the usual number of cigarettes smoked per day by total years of smoking cigarettes, and dividing by 20 (a pack of cigarettes). For this analysis, subjects were categorized as having 20 or more pack-years, less than 20 pack-years, or having never smoked.

*Alcohol:*

Alcoholic beverages were defined in the diet history questionnaire as beer, wine, and hard liquor including alcoholic cocktails, whiskey, gin, vodka, scotch, bourbon, or rum. Participants were asked to report usual amounts consumed during the weekdays and during weekend days separately to better capture total consumption. Additionally, participants were asked about alcohol consumption 10 and 20 years ago as part of the health and lifestyle questionnaire. Participants who responded with "no" to the question, "Did you ever drink an average of one or more alcoholic beverages a month for a year or longer?" were considered never to have drunk alcohol. Participants who responded "yes" to this question were then asked the usual number of 12-ounce bottles of beer, 4-ounce glasses of wine, and 1.5-ounce shots of hard liquor consumed during the referent year, 10 and 20 years ago. Long-term exposure to alcohol, based on consumption of any type of alcoholic beverage 10 and 20 years prior to the referent year, was categorized in two levels (none to moderate and high alcohol consumption, cut-off was 20gms/week for men and 10gms/week for women).

*Dietary Protein:*

Participants were asked to recall foods eaten, the frequency with which they were eaten, serving size, and whether fats were added in the preparation. Nutrient information was obtained by converting food-intake data into nutrient data using the Minnesota Nutrition Coordinating Center nutrient database.(Dennis et al. 1980) Total protein intake is the sum intake of animal proteins (meats, poultry, fish, dairy, and eggs) and vegetable proteins (legumes, tofu). We calculated an animal/vegetable protein intake ratio and used a cut-off corresponding to the median of animal protein proportion of total protein intake (i.e. 60% of total protein intake is animal protein equivalent to a 1.5 animal/vegetable protein intake ratio, which means 50% more animal protein

intake than vegetable protein intake). This resulted into two categories (Low and high animal/vegetable protein intake ratio).

### 3.2.2 Statistical analysis

***Three step approach of candidate pathway-based gene-environment interaction analysis***

Our approach to examining gene-environment interactions (GEIs) at the pathway level, adjusting for gene and gene-gene interaction effects, attempted to integrate biologic and logical reasoning using a three-step analysis approach. The analysis was conducted for colon cancer risk and colon cancer survival separately, but the three-step procedure used for the two analyses was identical: thus, we will describe the procedure focusing on colon cancer risk below, supplementing specific differences for colon cancer survival as needed.

Each step provided a "product" to be used in the following steps. Step 1: for each gene on the pathway we summarized SNP-set interactions within the gene that are relevant for colon cancer risk for the susceptibility analysis and colon cancer survival for survival analysis. Specifically, we developed *gene-specific trees* (GSTs) that captured SNP-set interactions in the gene using logic regression (see Supplementary Materials for details). Step 2: epistatic interactions of genes in the pathway (gene-gene interactions) were modeled using the GSTs from Step 1 to develop *pathway tree(s)*. Pathway trees represented interactions of the genes without considering the environmental exposures, and were used as adjustment variable(s) in the GEI models of the next third step. Step 3: we modelled pathway GEIs between the GSTs and the three environmental exposures. We divided the full pathway into 9 sub-pathways and summarized GEIs in each sub-pathway using backward selection. The GEIs that remained in the sub-pathway summary models at the 5% significance level were jointly tested in the final GEI model for the entire pathway. A

summary of the three steps of the analysis approach are shown in **Tables 3.2A and 3.2B** for colon cancer risk and survival, respectively.

***SNP-set interactions within a gene on the pathway: Step 1***

We started the analysis exploring two forms of SNP-set interactions, SNP *intersection* and SNP *union*, within each gene in order to summarize the gene's SNP profile relevant to colon cancer risk. Both forms are derived from set-theory terminology reflecting biologically plausible interaction forms. A SNP intersection is a form of interaction where disease risk is elevated only if *all* of the SNPs in a specified set (e.g., a gene) carry their respective high-risk genotype. A single SNP, or subsets, of the set carrying the high-risk genotype are insufficient to elevate disease risk. For example, for a set of three SNPs, all three SNPs (SNP 1 *and* SNP 2 *and* SNP 3) may have to carry their high-risk genotype for disease risk to be elevated. A SNP union describes a form of interaction where disease risk may be elevated through several independent ways (i.e., genetic heterogeneity) which may include a SNP intersection (e.g., SNP 1 and SNP 2) or an individual SNP carrying the high-risk genotype. We applied logic regression (Ruczinski et al. 2003) to search for these biologically plausible forms of SNP-set interactions within genes (Dinu et al. 2012). For each gene on the pathway, this step produced one or more GSTs that represented profiles of combinations of SNPs in the gene relevant to colon cancer risk and survival. The GSTs, instead of individual SNPs, were used as building blocks of gene-gene and gene-environment interactions in the subsequent two steps.

***Pathway gene-set interactions: Step 2***

We used logic regression to search for gene-set interactions among all GSTs of the pathway developed in Step 1, except, instead of using individual SNPs as the binary predictors in the logic

regression models we used the GSTs. Note that each GST is a binary variable defined by multiple SNPs of the same gene. The final logic combinations of GSTs were considered as *pathway tree(s)* which provided a summary of the gene-gene interactions in the full pathway and were used as adjustment variable(s) in GEI modeling in Step 3.

***Modeling pathway gene-environment interaction (GEI) effects: Step 3***

The third and final step was guided by the working pathway figure (**Figure 3.1**) and used standard epidemiological modeling methods, i.e., logistic regression for colon cancer risk and Cox proportional hazards regression for colon cancer survival. The pathway genes were grouped into nine mutually exclusive sub-pathways of closely related genes (e.g., a gene and its gene receptors) as illustrated in the figure. To summarize GEIs within each sub-pathway, we used a backward variable selection procedure for model building (Harrell et al. 1996). For each sub-pathway, the procedure started with a model including the interaction terms of the corresponding GSTs and the three environmental factors of interest and eliminated the least significant term(s) in a stepwise fashion. All interaction terms significant at p-value $< 0.05$ from each sub-pathway were tested simultaneously in the final pathway GEI model. All models were adjusted for the pathway trees from Step 2 in addition to age at diagnosis or selection, sex, race (white, Hispanic, or black race), and study center (University of Utah; the Kaiser Permanente Medical Care Program of Northern California; and the University of Minnesota). Baseline hazards for Cox proportional hazards models were stratified by colon cancer stage at diagnosis.

## 3.3 Results

The study included data on 1,541 colon cancer cases and 1,934 controls. Follow-up data and vital status for use in the survival analysis were available for only 1,408 of the 1,541 cases. Cases with

missing follow-up belonged mainly to the Northern California and Minnesota study centers; patients may have moved out of state or were not able to be tracked by their respective local tumor registry. They, however, did not differ from cases with follow-up information with regards to baseline variables (age, sex, race, or cancer stage).

The angiogenesis candidate gene-pathway included a total of 257 SNPs in 34 genes **(Table 3.1)**. The results of the first step of the analysis involved only the genetic components of the pathway: SNP-set interactions within each of the 34 genes. Details of the logic models that yielded the GSTs of the 34 pathway genes in association with colon cancer risk and survival are shown in **Supplementary Tables 3.1 and 3.2**, respectively. The tables show the optimal model size for each gene as determined by the cross-validation, the model score and structure of the final logic model. Some genes had more than one GST; hence the total number of logic trees for the pathway exceeded the number of the pathway genes. For example, the logic regression yielded 4 GSTs for the *FLT1* gene in association with colon cancer risk **(Supplementary Table 3.1)**.

Results of the second step of the analysis summarized the pathway by searching for GST interactions (gene-gene interactions). This process yielded one pathway tree in association with colon cancer risk **(Supplementary Figure 3.4)** and 4 pathway trees in association with colon cancer survival **(Supplementary Figure 3.5)**. The figures show the GST of each pathway tree including the at-risk genotypes and frequencies.

The third step of the analysis involved modeling for interactions between the GSTs and the three environmental exposures of interest. Only the statistically significant GEI results are displayed in **Tables 3.3 and 3.4** in association with colon cancer risk and survival, respectively. Statistically significant interactions were observed between specific components of the angiogenesis gene-

pathway and all three environmental exposures. Overall, the magnitude of the main effects of the significant GSTs increased with increasing levels of animal/vegetable protein intake ratio, smoking, and alcohol consumption.

Genes among the major drivers of angiogenesis interacted with the three exposures in association with colon cancer risk. The *FLT1* gene showed statistically significant interactions with ≥ 20 pack-years of smoking (interaction odds ratio ($OR_{INT}$) = 1.64, 95% confidence interval (CI) (1.11, 2.41), p-value=0.013) and high animal/vegetable protein intake ($OR_{INT}$=1.69, 95% (1.03, 2.77), p-value=0.037); and the *KDR* gene showed a statistically significant interaction with long-term alcohol consumption ($OR_{INT}$=1.53, 95% CI (1.10, 2.13), p-value=0.012). Among the inflammatory genes, interactions between *BMP4* gene and ≥ 20 pack-years of smoking ($OR_{INT}$=1.60, 95% CI (1.10, 2.32), p-value=0.013) and *TLR2* gene and long-term alcohol consumption ($OR_{INT}$=1.59, 95% CI (1.05, 2.38), p-value=0.027) were statistically significant.

Three genes among the inflammatory genes had significant GEIs in association with colon cancer survival, each with one of the three exposures: *TNF* gene and high animal/vegetable protein intake ratio (interaction hazard ratio $HR_{INT}$=1.74, 95% CI (1.09, 2.76), p-value=0.019); *BMP1* gene and ≥ 20 pack-years of smoking ($HR_{INT}$=1.79, 95% CI (1.03, 3.10), p-value=0.039); and *BMPR2* gene and long-term alcohol consumption ($HR_{INT}$=7.91, 95% CI (1.57, 39.74), p-value=0.012).

### 3.4 Discussion

Prior to the advent of GWAS, candidate gene studies specified genes to be investigated *a priori* based on their biologic functional significance to the disease. An approach to investigate the entire pathway systematically, however, has been lacking and seldom has the biologic reasoning

used for the selection of candidate genes been carried through to the analysis (Thomas et al. 2009). We developed a novel candidate-pathway framework to assess GEIs and illustrated its use for colon cancer risk and survival. We focused on only one gene-pathway, the angiogenesis pathway, and three angiogenesis-related lifestyle risk factors and identified several novel GEIs. Our framework emphasized the biologic hypothesis throughout the process starting from the selection of the candidate genes and the specific lifestyle exposures, and carried the logic to the three steps of the analysis. We started by developing GSTs that captured biologically plausible forms of SNP-set interactions within each gene, hence, our building blocks of gene-gene and gene-environment analysis represented the genes rather than individual SNPs. Our next step provided a summary of the full pathway's genetic effects. Since the same environmental exposure could be interacting with different genes on the same pathway, whether through similar or different mechanisms (Cordell 2009a), guided by the working pathway figure, we dissected the pathway into mutually exclusive sub-pathways involving groups of genes sharing the same function or that are closely related. This grouping allowed for genes in the sub-pathways to interact with all three exposures and avoid potentially missing important GEIs. Indeed, we observed interactions between the same environmental exposure and different components of the angiogenesis gene-pathway. Our approach to use gene-level summaries rather than individual SNPs is a step ahead of a typical interaction analysis that considers pairwise interactions between SNPs for gene-gene interactions or interactions between an individual SNP and an environmental exposure for GEI testing.

Interest in identifying GEIs in colorectal cancer has been on the rise. In genome-wide settings, GEI has been examined through genome-wide scans and/or a candidate approach focusing on previously identified GWAS loci and known colorectal cancer risk/survival factors. One GWAS

that used 3 methods to test GEI (a traditional case-control test, a case-only test and a 2-step method proposed by Murcray and colleagues that involves a screening test followed by a traditional case-control test of GEI) did not identify any genome-wide significant GEIs, yet using a candidate approach of analyzing previously reported colorectal cancer GWAS susceptibility loci they identified 7 nominally significant GEIs one of which was between alcohol and a SNP on *CHD1* gene (chromosome 16q22.1) (Figueiredo et al. 2011). Another study that examined 10 published colorectal cancer GWAS loci and 12 environmental risk factors identified a single interaction with vegetable consumption and a SNP on chromosome 8q23.3 (Hutter et al. 2012). A third study was able to identify an interaction with being overweight and a SNP on chromosome 11q23.3 but none were identified from the candidate approach (Siegert et al. 2013). In contrast to focusing on previous empirical GWAS findings as candidate genes for GEI testing, our approach to selecting candidate genes and the pathway was based on biologic relevance and hypothesis. Despite analyzing genes in only one pathway, we were able to identify a considerably large number of significant interactions with all three exposures on both cancer risk and survival with a magnitude of the interaction OR ranging between 1.53 to 1.69 for risk and 1.80 to 7.78 for survival. We believe our findings could be increased by focusing on more colon-cancer-related pathways and their relevant environmental exposures.

Among the previously identified GEIs focusing on candidate variants was the interaction between low folate intake and an MTHFR polymorphism on increased risk of colorectal adenoma (Ulrich et al. 1999). A recent meta-analysis identified GEIs of heavy smoking and heavy alcohol drinking with another MTR polymorphism on increased colorectal cancer risk (Ding et al. 2013). Recent evidence also supported an association between prolonged cigarette smoking and colorectal cancer that is modified by specific variants in carcinogen metabolism

genes (Cleary et al. 2010). These findings suggest that choice and definition of specific components or patterns of the assessed environmental exposures are important elements in characterizing a strong GEI (Prentice 2011). For diet, for example, we focused on the specific nutrient effect of animal protein previously shown to be associated with colorectal adenoma (Yang et al. 2012) and colorectal cancer (De Stefani et al. 2012). Our approach, indeed, identified GEI of a high animal/vegetable protein intake ratio and both arms of the pathway: *FLT1* among major angiogenesis genes on colon cancer risk and *TNF* among interacting inflammatory genes on colon cancer survival.

Other significant GEIs that we detected included genes that have demonstrated strong associations with colorectal cancer in previous reports. One example is for major drivers of the angiogenesis process: the *VEGF* receptor 1 and 2 genes (*FLT1* and *KDR*, respectively) (Jang et al. 2013, Slattery et al. 2014).We were able to characterize significant GEIs of the *FLT1* gene with smoking and animal protein intake, and the *KDR* gene with alcohol consumption in association with colon cancer risk. Our study is among few others that examined GEI in association with colon cancer survival after diagnosis. Previous studies that examined loci associated with colorectal cancer prognosis identified genetic variants affecting survival and recurrence in patients receiving chemotherapy (Xing et al. 2011, Dai et al. 2012). A study of postmenopausal women identified an association with a locus in *SMAD7* and pre-diagnostic non-steroidal anti-inflammatory drug use (Passarelli et al. 2011). In our results, we detected interactions between three genes *TNF*, *BMP1* and *BMPR2* genes and animal protein intake, smoking, and alcohol, respectively. These interactions have not been reported previously in association with colon cancer survival.

BMPs (bone morphogenetic proteins) are multi-functional growth factors part of the TGFβ superfamily (Chen et al. 2004). A special interest in *BMP* genes and colorectal cancer developed over the last decade with several studies demonstrating their tumor suppressor properties (Beck et al. 2006, Nishanian et al. 2004). Specifically, BMP2 is a cell differentiation and proliferation factor that has been shown in *in vitro* studies to inhibit colon epithelial cell growth inducing apoptosis and inhibiting cell proliferation (Hardwick et al. 2004). Evidence from gene expression studies indicates that expression levels of *BMP2* were significantly lower in colorectal adenocarcinomas compared to adenomatous polyps since loss of *BMP2* caused hyperplasia of intestinal epithelial cells and tumorigenesis (Xiang et al. 2012). *BMP4* loci were also previously identified in association with colorectal cancer from a GWAS meta-analysis (Houlston et al. 2008) and a fine mapping study of susceptibility loci in the BMP pathway including *BMP2* and *BMP4* genes (Tomlinson et al. 2011). We have previously published associations between SNPs on BMP-signaling pathway genes and colon cancer risk including *BMP2*, *BMP4*, and *BMPR2* (Slattery et al. 2012b). In this analysis, we identified GEIs with BMP genes in association with both colon cancer risk and survival. We observed GEIs between *BMP4* gene and smoking on colon cancer risk; and *BMP1* gene and smoking and *BMPR2* gene and alcohol on colon cancer survival. Some of these interactions display clear dose response associations as shown by an increasing magnitude of gene OR with increasing levels of smoking. Our results, thus, provide additional evidence of the potential importance of BMP-related genes as components of the angiogenesis pathway, and their interactions in colon cancer etiology and outcomes.

A GWAS employs an agnostic data-driven approach where prior knowledge of SNP function is not required and most GWASs have not investigated GEI, primarily due to lack of data on environmental exposures (Stranger et al. 2011). On the other hand, a candidate pathway in a

candidate gene study is based on biological hypotheses derived from existing knowledge of candidate pathway, genes and/or SNPs defining that pathway. Approaches based on informed candidate gene selection may be more suited to examining GEI effects compared to GWAS loci. The low-penetrance GWAS loci are harder to identify, have smaller effect sizes, and are unlikely to be the functional variants themselves alone. The associated SNPs are markers of an underlying haplotype that includes the functional variants (Stranger et al. 2011). In contrast, our candidate pathway approach may carry several advantages over an empirical data-driven approach: (1) candidate genes and exposures were selected based on biologic relevance; (2) it allowed for the interaction of multiple SNPs within each gene potentially capturing the full gene effect; and (3) a multiple testing adjustment for testing many non-hypothesized associations in GWAS was not required for our candidate gene-pathway analyses because the associations were biologically hypothesized *a priori* (Tomlinson et al. 2011). Furthermore, GWAS analyses that focus only on SNPs with significant marginal effects will miss interactions with variants with weak or no marginal effects.

A few limitations of our study are related to the design of the case-control study which suffers from inherent forms of bias such as recall bias. This was minimized by: using a rapid-reporting system to identify cases; conducting the majority of interviews within 4 months of diagnosis; and limiting the referent period of the study questionnaires to two to three years prior to diagnosis. With regards to our environmental exposures of interest, we obtained long-term alcohol consumption and cigarette smoking history. The diet history used was extensive and able to capture more detail compared to that obtained when using self-administered questionnaires. We considered all colon cancer cases and did not stratify by distal and proximal site. We selected the genes of the angiogenesis pathway using a candidate approach. Not all genes in the human

genome have been characterized and, therefore, their pathway information may not be available; if such genes with little or no characterization existed and were relevant to the angiogenesis process they would have been missed. Due to the large size of the GEI models involving a large number of GST-environment interactions to be tested across the pathway, we limited the adjustment variables to select CRC risk and survival predictors. This limitation led to not adjusting for effects of other CRC-relevant factors such as tumor microsatellite instability (MSI) status. Our analysis was based on fixed datasets of available data and we had no control over the sample size. In addition, our study was not a null study and we did detect statistically significant GEIs and thus a power calculation was not needed. The novel GEIs detected from our analysis in association with colon cancer risk and survival emphasize the need to employ an approach based on biologic hypotheses when examining GEIs. The low-penetrance markers and the environmental exposures they interact with are common in the population, and the magnitude of the interaction is often larger than their individual effects. Identification of these interactions could potentially explain a large portion of the risk variation in the population (Le Marchand et al. 2008). Knowledge of the (theoretically modifiable) lifestyle factors influencing colon cancer risk and survival as modified by the individuals' genetic susceptibility can be directly translated into practical public health applications. It also helps shift the notion of the deterministic role of susceptibility genes into one that reflects interacting effects of genes and lifestyle habits, and portrays colon cancer and death from colon cancer as a potentially avoidable disease while providing new insights into its prevention strategies.

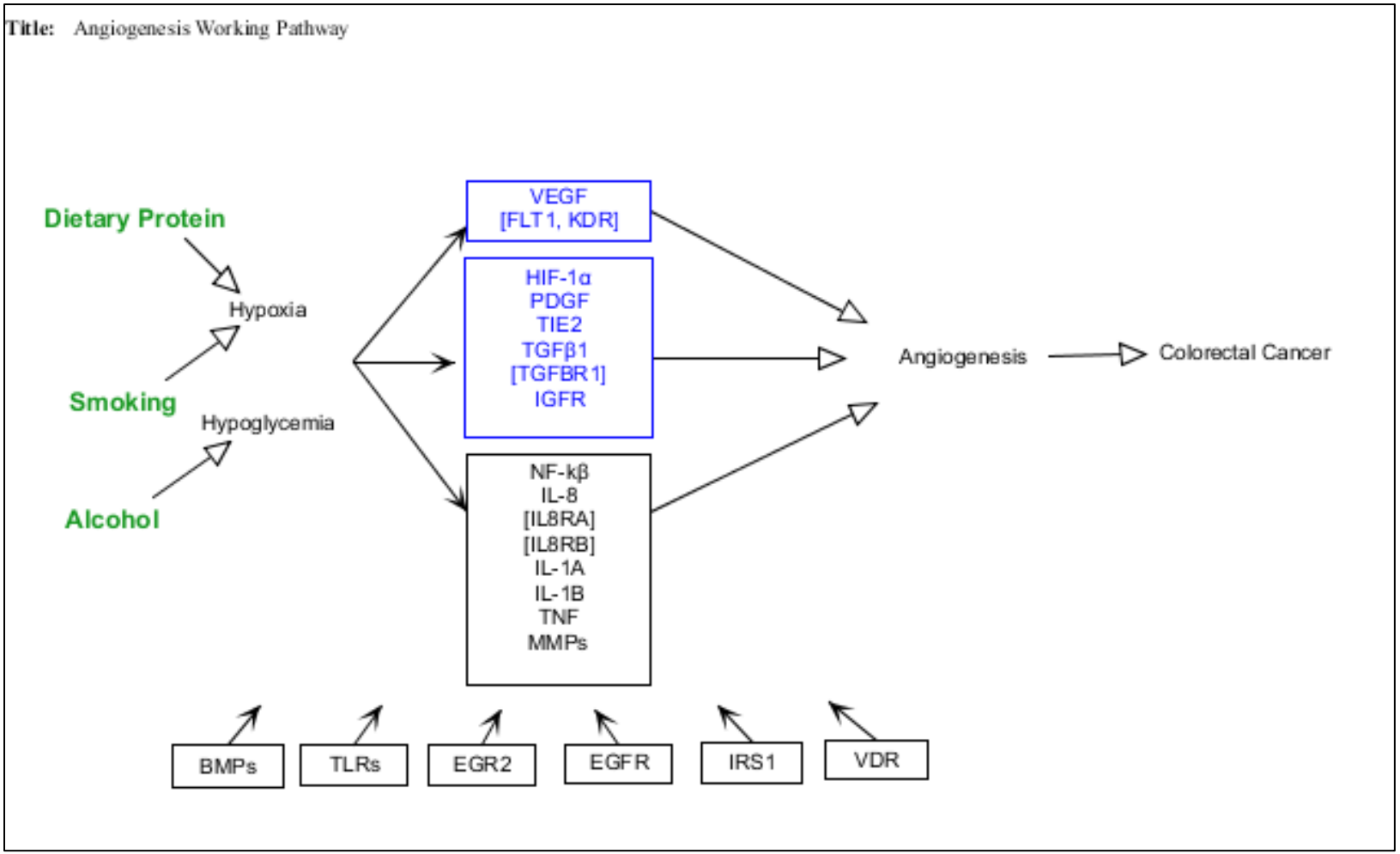



**Figure 3.1: Working figure of the angiogenesis pathway genes.**

Key gene components of the pathway are in blue frames; secondary genes are in black frames; environmental factors are in green text.

**Table 3.1: Angiogenesis pathway gene list**

| Genes | Name |
|---|---|
| **Major drivers of angiogenesis** | |
| *VEGFA* | Vascular endothelial growth factor A |
| *FLT1* | Vascular endothelial growth factor receptor 1 |
| *KDR* | Vascular endothelial growth factor receptor 2 |
| *HIF-1α* | Hypoxia-inducible factor 1, alpha |
| *PDGF* | Platelet-derived growth factor |
| *TIE2* | Tyrosine-protein kinase receptor |
| *TGFβ* | Transforming growth factor, beta |
| *TGFβR* | Transforming growth factor, beta receptor |
| *IGF-IR* | Insulin-like growth factor-I receptor |
| **Interacting inflammatory genes** | |
| *NFKB1* | Nuclear factor of kappa light polypeptide gene enhancer in B-cells 1 |
| *IL8* | Interleukin-8 |
| *IL8RA* | Interleukin-8 receptor, alpha |
| *IL8RB* | Interleukin-8 receptor, beta |
| *IL1A* | Interleukin-1, alpha |
| *IL1B* | Interleukin-1, beta |
| *TNF* | Tumor necrosis factor |
| *MMPs* | Matrix metallopeptidases (*MMP1, MMP3, MMP7, MMP9*) |
| *BMPs* | Bone morphogenetic proteins (*BMP1, BMP2, BMP4, BMPR1A, BMPR1B, BMPR2*) |
| *TLRs* | Toll-like Receptors (*TLR2, TLR3, TLR4*) |
| *EGR2* | Early Growth response 2 |
| *EGFR* | Epidermal growth factor receptor |
| *IRS1* | Insulin receptor substrate 1 |
| *VDR* | Vitamin D Receptor |

**Table 3.2: Summary of the 3-step candidate-pathway gene-environment interaction approach**

**A. Colon cancer risk analysis steps**

| Analysis Step | Interaction of interest | Variable of interest | Model | Specific Procedures | Product |
|---|---|---|---|---|---|
| **Step 1:** Summarize gene effects | SNP-set interaction within gene | SNPs on each gene separately | Logic regression with logit link | Cross-validation to determine optimal model size | Gene-Specific trees (GSTs) |
| **Step 2:** Summarize pathway effects | Gene-set interaction within pathway | All GSTs on the pathway | Logic regression with logit link | Cross-validation to determine optimal model size | Pathway Trees |
| **Step 3:** Test gene-environment interaction | Gene-environment interaction within pathway | a. Sub-pathway specific GSTxE* b. Full pathway GSTxE | Logistic regression model[¥] | Statistical significance testing | Pathway GEIs |

* GSTxE, gene-specific tree - environment interaction
[¥]Models adjusted for age, sex, race, study center, pathway tree

**B. Colon cancer survival analysis steps**

| Analysis Step | Interaction of interest | Variable of interest | Model | Specific Procedures | Product |
|---|---|---|---|---|---|
| **Step 1:** Summarize gene effects | SNP-set interaction within gene | SNPs on each gene separately | Logic regression fitting exponential survival models | Cross-validation to determine optimal model size | Gene-Specific trees (GSTs) |
| **Step 2:** Summarize pathway effects | Gene-set interaction within pathway | All GSTs on the pathway | Logic regression fitting exponential survival models | Cross-validation to determine optimal model size | Pathway Trees |
| **Step 3:** Test gene-environment interaction | Gene-environment interaction within pathway | a. Sub-pathway specific GSTxE* b. Full pathway GSTxE | Cox Proportional Hazards model[¥] | Statistical significance testing | Pathway GEIs |

* GSTxE, gene-specific tree - environment interaction
[¥]Models adjusted for age, sex, race, study center, pathway trees, stratified by cancer stage

**Table 3.3: Effects of gene-environment interactions significant at 5% level between colon cancer gene-specific trees and environmental factors on colon cancer risk**

| Gene-Specific Tree | Gene | Chr | Cases (%) | Controls (%) | Gene OR* (95%CI) | Env Factor | Category | N (%) | Gene OR by Env Factor* (95%CI) | $OR_{INT}$* (95%CI) | $P_{INT}$* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| rs678714 (TA or AA) | *FLT1* | 13q12 | 276 (18.1%) | 417 (21.5%) | **0.82 (0.69, 0.97)** | Smoking | Non | 1563 (44.6%) | **0.72 (0.55, 0.94)** | Ref | |
| | | | | | | | < 20 PY | 668 (19.1%) | **0.58 (0.40, 0.86)** | 0.80 (0.50, 1.28) | 0.350 |
| | | | | | | | ≥ 20 PY | 1272 (36.3%) | 1.16 (0.88, 1.54) | **1.64 (1.11, 2.41)** | **0.013** |
| rs2387632 (CC or CT) **OR** rs9513070 (GG) | *FLT1* | 13q12 | 1,333 (85.6%) 238 (15.3%) | 1,738 (88.8%) 335 (17.11%) | **0.64 (0.51, 0.80)** | Animal/ Vegetable Protein Ratio | Low | 1010 (28.7%) | **0.40 (0.26, 0.61)** | Ref | |
| | | | | | | | High | 2506 (71.3%) | **0.76 (0.59, 0.99)** | **1.69 (1.03, 2.77)** | **0.037** |
| rs6838752 (TT) | *KDR* | 4q11-q12 | 925 (59.9%) | 1,098 (56.7%) | 1.11 (0.97, 1.27) | Alcohol | Non/Moderate | 2744 (77.0%) | 1.01 (0.86, 1.18) | Ref | |
| | | | | | | | Heavy | 819 (23.0%) | **1.53 (1.14, 2.04)** | **1.53 (1.10, 2.13)** | **0.012** |
| rs17563 (CC or CT) | *BMP4* | 14q22-q23 | 1,193 (76.6%) | 1,568 (80.2%) | **0.84 (0.71, 0.99)** | Smoking | Non | 1562 (44.6%) | **0.70 (0.54, 0.89)** | Ref | |
| | | | | | | | < 20 PY | 668 (19.1%) | 0.74 (0.51, 1.09) | 1.05 (0.67, 1.65) | 0.821 |
| | | | | | | | ≥ 20 PY | 1271 (36.3%) | **1.12 (0.85, 1.47)** | **1.60 (1.10, 2.32)** | **0.013** |
| rs3804099 (TT or TC) | *TLR2* | 4q32 | 1,257 (80.7%) | 1,531 (78.2%) | **1.20 (1.02, 1.42)** | Alcohol | Non/ Moderate | 2714 (77.3%) | 1.08 (0.89, 1.31) | Ref | |
| | | | | | | | Heavy | 798 (22.7%) | **1.72 (1.21, 2.45)** | **1.59 (1.05, 2.38)** | **0.027** |

Abbreviations: Chr, Chromosome; Env, Environmental; PY, pack-years; OR, odds ratio; P, p-value; INT, interaction
*Adjusted for age, sex, race, study center, pathway tree

**Table 3.4: Effects of gene-environment interactions significant at 5% level between colon gene-specific trees and environmental factors on colon cancer survival**

| Gene-Specific Tree | Gene | Chr | Cases (%) | Gene HR* (95%CI) | Env Factor | Category | N (%) | Gene OR by Env Factor* (95%CI) | $HR_{INT}$* (95%CI) | $P_{INT}$* |
|---|---|---|---|---|---|---|---|---|---|---|
| rs1800630 (CA or AA) | *TNF* | 6p21.3 | 466 (31.53%) | 0.93 (0.77, 1.14) | Animal/ Vegetable Protein Ratio | Low | 399 (27.0%) | **0.62 (0.41, 0.93)** | Ref | |
| | | | | | | High | 1079 (73.0%) | 1.07 (0.86, 1.35) | **1.74 (1.09, 2.76)** | **0.019** |
| rs13257482 (GG) | *BMP1* | 8p21 | 850 (58.3%) | **1.45 (1.13, 1.87)** | Smoking | Non | 614 (41.7%) | 1.12 (0.76, 1.65) | Ref | |
| **OR** rs4075478 (TC or CC) | | | 902 (61.8%) | | | < 20 PY | 279 (19.0%) | 1.39 (0.68, 2.84) | 1.51 (0.68, 3.36) | 0.317 |
| | | | | | | ≥ 20 PY | 578 (39.3%) | **2.04 (1.37, 3.04)** | **1.79 (1.03, 3.10)** | **0.039** |
| rs12477602 (GG or GA) | *BMPR2* | 2q33-q34 | 1389 (98.7%) | 1.09 (0.55, 2.18) | Alcohol | Non/Moderate | 1060 (74.9%) | 0.60 (0.28, 1.29) | Ref | |
| | | | | | | Heavy | 356 (25.1%) | 2.64 (0.57, 11.33) | **7.91 (1.57, 39.74)** | **0.012** |

Abbreviations: Chr, Chromosome; Env, Environmental; PY, pack-years; HR, hazard ratio; P, p-value; INT, interaction

*Adjusted for age, sex, race, study center, pathway trees, baseline hazard stratified by cancer stage

**A Candidate Pathway Approach Identifies Multiple Gene-Environment Interactions in Association with Colon Cancer Risk and Survival**

**Supplementary Information**

## Logic Regression

Logic regression involves a method that detects high-order interactions and patterns of interactions among (binary) predictors in association with an outcome within a regression framework. The method employs the Boolean logic searching for Boolean combinations of binary predictors (e.g., SNPs). The SNPs in a Boolean combination are referred to as "leaves" and the combination of the SNPs joined by the Boolean operators, ⊔(AND), ⊔ (OR), and $^{c}$ (NOT), is referred to as a "logic tree". The logic trees are also binary taking the value of "0" or "1", or "Yes" or "No".

We used the logic regression implemented in R version 3.0.0 using the "LogicReg" R package (Charles Kooperberg and Ingo Ruczinski (2013). LogicReg: Logic Regression. R package version 1.5.5. http://CRAN.R-project.org/package=LogicReg). We used logic regression models fitting logistic models to assess colon cancer risk; and exponential survival models to assess colon cancer survival.

Specifically, the logic model with logit link took the form:

$$\log (\Pr[Y=1]/ \Pr[Y=0]) = \beta 0 + \beta 1\ L1 + \beta 2\ L2 + \ldots + \beta p\ Lp$$

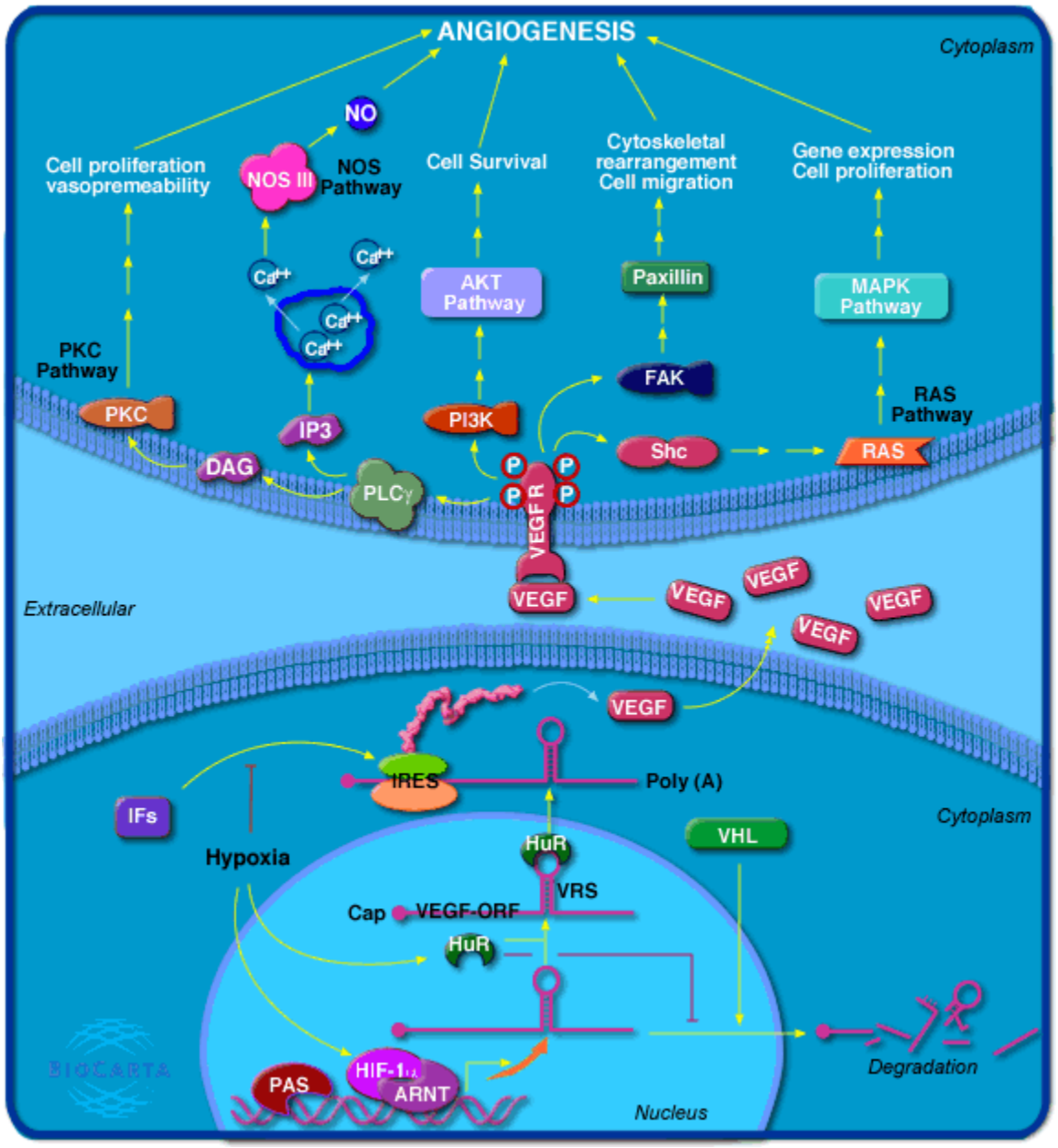
where Y is a binary response variable, β0, β1,…, βp are the model parameters, and L1, L2, …, Lp are the Boolean combinations of SNPs.

The logic model for exponential survival is equivalent to the proportional-hazards form:

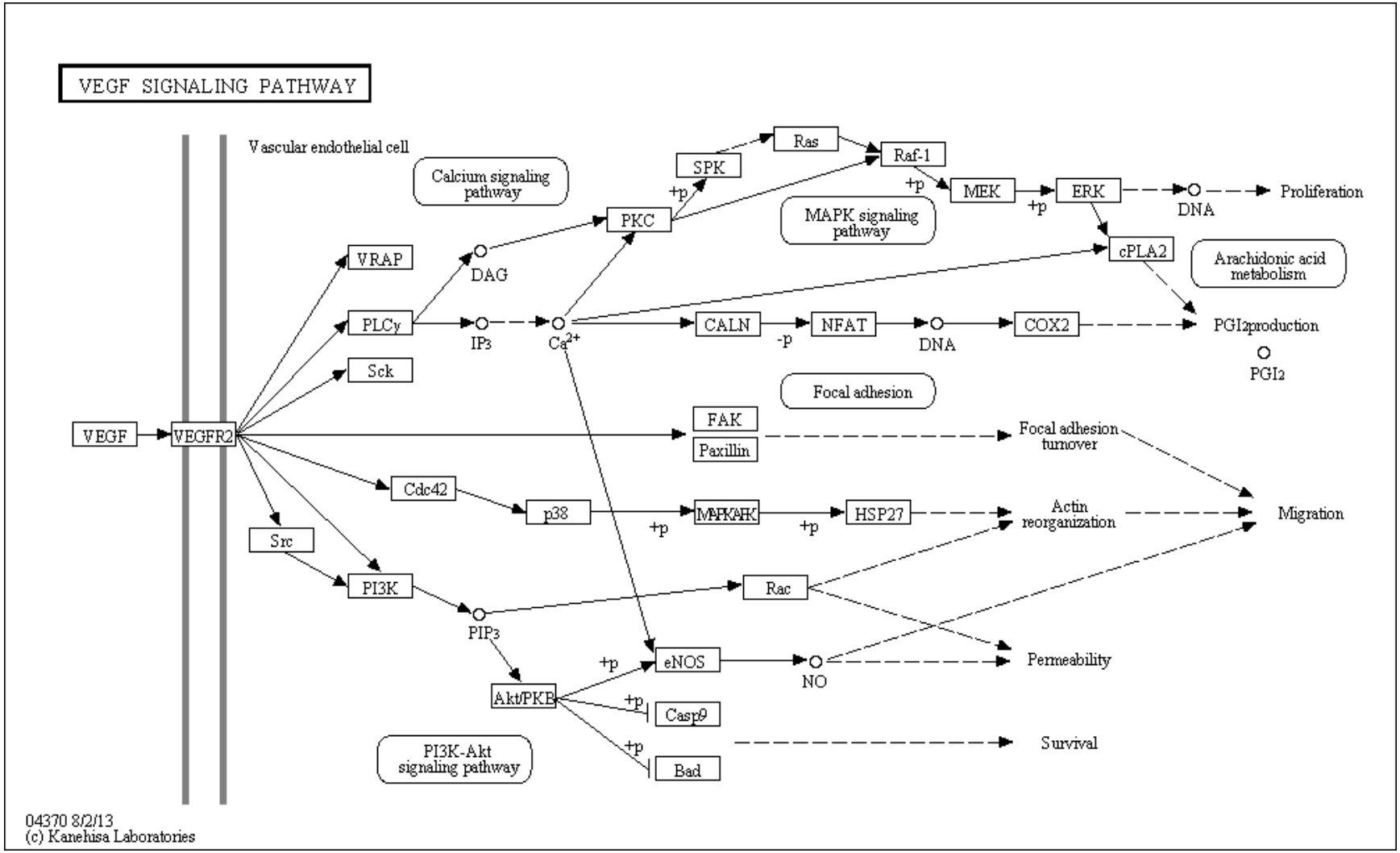
$$\log \lambda(c) = \beta 0 + \beta 1\ L1 + \beta 2\ L2 + \ldots + \beta p\ Lp$$

where $\lambda(c)$ is the hazard rate, a function of the marginal cumulative hazard c, $\beta 0$, $\beta 1$,…, $\beta p$ are the model parameters, and L1, L2, …, Lp are the Boolean combinations of SNPs.

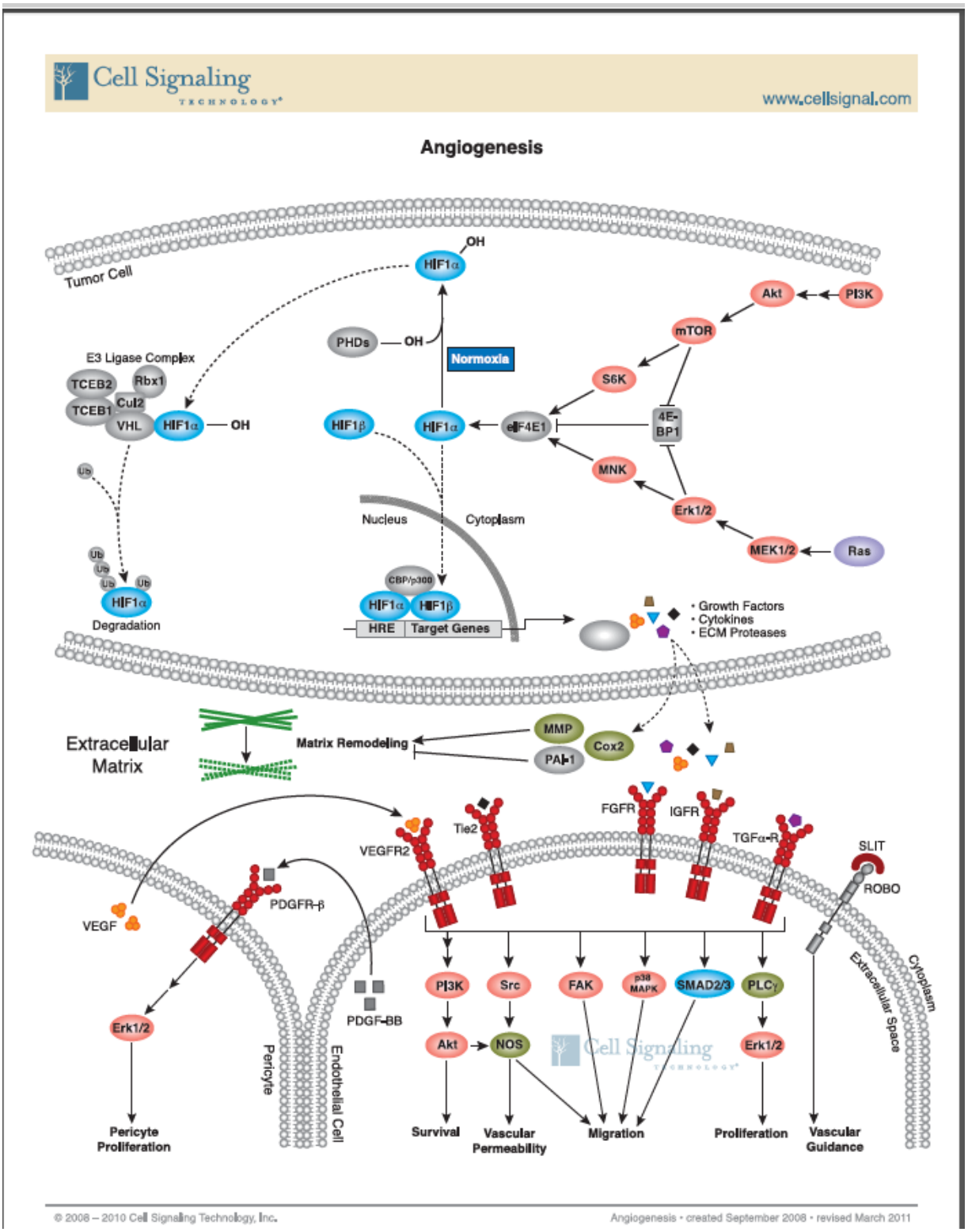
Considering the large search space, defined by the number of SNPs and all their possible combinations, the logic regression needs to employ an efficient search strategy. The LogicReg package in R uses a simulated annealing search algorithm that involves, given a certain model, randomly picking a move from a set of six permissible moves leading to a new model. A model selection procedure that determines model size for the simulated annealing algorithm (i.e., the number of combinations of SNPs or trees and the number of SNPs in a combination or leaves) is necessary to avoid over fitting. We derived the optimal size of the logic regression model using 10-fold cross-validation up to a desired maximum size of 9 trees and 20 leaves. The search algorithm is designed to prohibit further moves if the desired maximum size is reached and more often the final model size is smaller than the desired size. We fitted the optimal-size model 100 times, each with a different random seed (i.e. starting point for the search), and the best solution (based on lower deviance or negative log-likelihood function) was reached. Note that logic models for exponential survival models yield the same results as Cox proportional hazards models yet with much less computational burden.




**Supplementary Figure 3.1:** VEGF, Hypoxia, and Angiogenesis Pathway. Illustration reproduced courtesy of The BioCarta Pathways (http://www.biocarta.com/pathfiles/h_vegfPathway.asp).



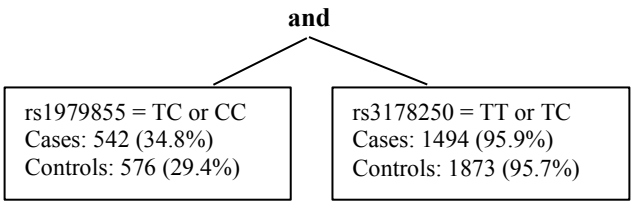

**Supplementary Figure 3.2:** VEGF Signaling Pathway. Illustration reproduced courtesy of KEGG, (Kyoto Encyclopedia of Genes and Genomes) Pathway database (http://www.kegg.jp/kegg-bin/highlight_pathway?scale=1.0&map=map04370&keyword=angiogenesis).



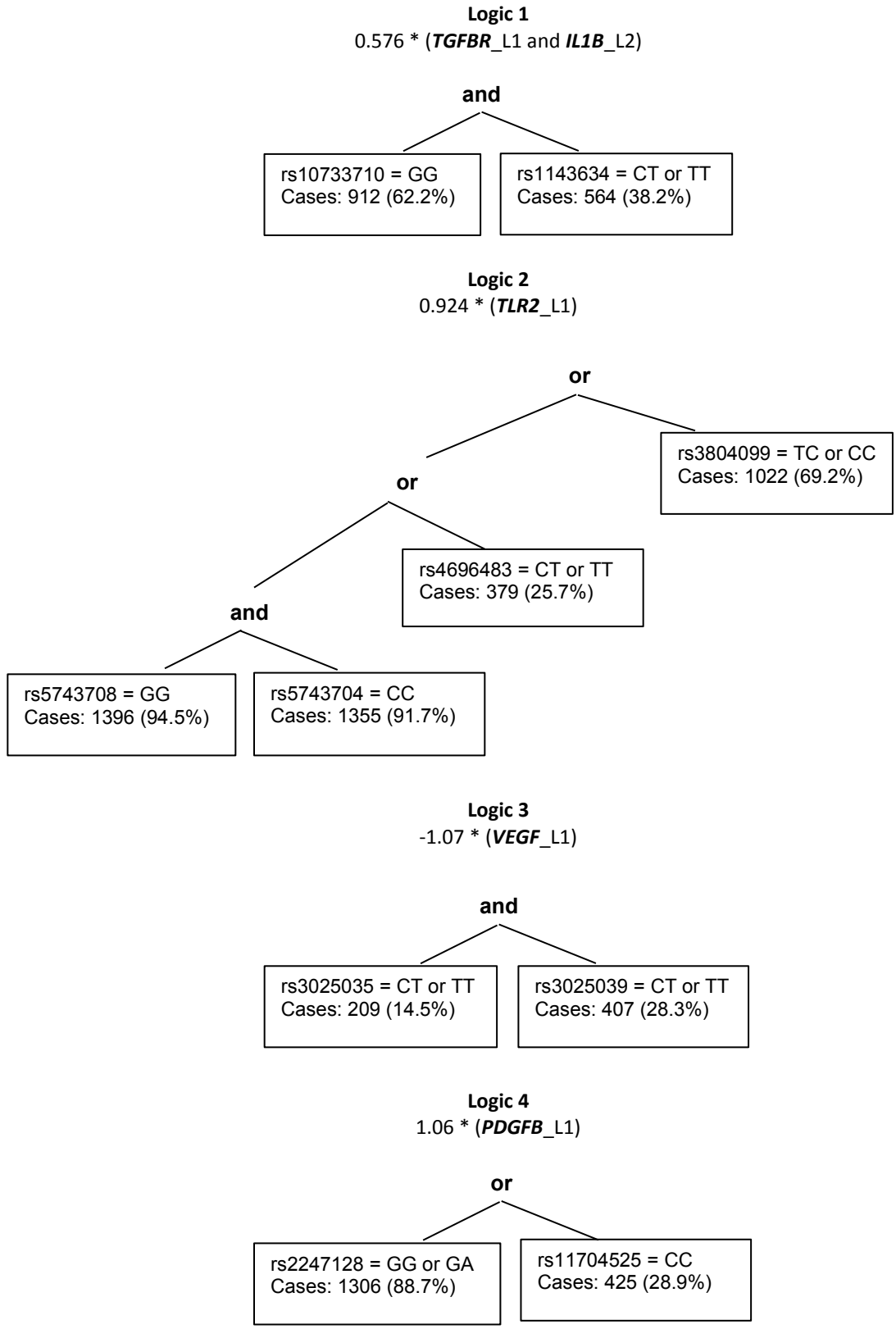

**Supplementary Figure 3.3:** Angiogenesis Signaling Pathway. Illustration reproduced courtesy of Cell Signaling Technology, Inc. (www.cellsignal.com).
(http://www.cellsignal.com/common/content/content.jsp?id=pathways-angiogenesis).

**Logic 1**

**-0.369 * (*BMP2*_L1)**




**Supplementary Figure 3.4: Gene-Pathway Tree in association with Colon Cancer Risk.**




**Supplementary Figure 3.5: Gene-Pathway Tree in association with Colon Cancer Survival.**

**Supplementary Table 3.1: Colon cancer gene-specific trees and colon cancer risk**

| Gene | N of Trees | N of Leaves | Model score | Model |
|---|---|---|---|---|
| *VEGF* | 1 | 3 | 4682.03 | -0.234 +2.18 * ((rs3025033 and (not rs3025030)) and rs833069) |
| *FLT1* | 5 | 5 | 4709.211 | -1.44 +1.16 * (rs2387632 and (not rs9513070)) +0.275 * rs12858139 +0.726 * (not rs2387632) +0.317 * (not rs678714) |
| *KDR* | 2 | 2 | 4770.181 | -0.185 +0.549 * rs1870377 -0.641 * rs6838752 |
| *HIF1* | 1 | 2 | 4813.678 | -0.252 +0.631 * (rs2301113 and (not rs11549465)) |
| *PDGFB* | 1 | 1 | 4805.092 | -0.221 -1.39 * rs6001512 |
| *TEK* | 1 | 1 | 4891.394 | -0.179 -0.148 * rs603085 |
| *TGFB* | 2 | 2 | 4858.819 | -0.537 +0.164 * rs4803455 +0.302 * (not rs1800469) |
| *TGFBR* | 3 | 3 | 4776.543 | 0.2 -0.287 * (not rs1571590) -0.16 * rs6478974-0.0976 * rs1571590 |
| *IGF1R* | 1 | 1 | 4885.171 | -0.196 -0.0785 * rs2139924 |
| *NFKB1* | 1 | 1 | 4821.9 | -0.215 -0.42 * rs4648110 |
| *IL8* | 2 | 2 | 4818.652 | -0.182 +0.472 * (not rs2227307) -0.666 * (not rs4073) |
| *IL8RA* | 1 | 1 | 4794.547 | -0.261 +0.0929 * (not rs1008563) |
| *IL8RB* | 1 | 1 | 4822.595 | -0.158 -0.0972 * rs4674258 |
| *IL1A* | 1 | 2 | 4826.072 | -0.298 +0.0838 * ((not rs3783546) or (not rs1878321)) |
| *IL1B* | 3 | 4 | 4814.134 | -0.0784 -0.683 * (rs1143623 or ((not rs1143627) and (not rs1143633))) +0.505 * (not rs1143633) |
| *TNF* | 1 | 1 | 4823.367 | -0.279 +0.17 * rs1800630 |
| *MMPS* | 1 | 2 | 4872.371 | -0.556 +0.371 * ((not rs1996352) or (not rs3918261)) |
| *BMP1* | 2 | 2 | 4760.828 | -0.241 +0.898 * (rs12114940 and rs7592) |
| *BMP2* | 2 | 4 | 4802.057 | -0.327 +0.321 * (rs1979855 and (not rs3178250)) +1.58 * ((not rs1979855) and rs3178250) |
| *BMP4* | 1 | 1 | 4818.195 | -0.273 +0.209 * rs17563 |
| *BMPR1A* | 1 | 3 | 4776.84 | -0.119 -0.264 * ((rs6586034 or rs7088641) and (not rs7895217)) |
| *BMPR1B* | 1 | 1 | 4785.95 | -0.293 +0.2 * (not rs9307147) |
| *BMPR2* | 1 | 1 | 4610.991 | -0.273 +0.16 * rs6751210 |
| *GDF10* | 1 | 1 | 4806.873 | -0.216 -0.196 * rs7093975 |
| *TLR2* | 1 | 1 | 4822.488 | -0.197 -0.153 * rs3804099 |
| *TLR3* | 1 | 1 | 4819.283 | -0.582 +0.369 * (not rs3775292) |

| | | | | |
|---|---|---|---|---|
| *TLR4* | 1 | 1 | 4812.316 | -0.916 +0.7 * (not rs11536898) |
| *EGR2* | 1 | 1 | 4820.473 | -0.353 +0.142 * (not rs2295814) |
| *EGFR* | 2 | 2 | 3797.056 | 0.0183 -0.249 * (not rs2472520) -0.271 * rs6944906 |
| *IRS1* | 1 | 1 | 5120.865 | -0.141 +0.22 * IRS1 |
| *VDR* | 1 | 1 | 4453.781 | -0.306 +0.205 * VDRFok1 |

**Supplementary Table 3.2: Colon cancer gene-specific trees and colon cancer survival**

| Gene | N of Trees | N of Leaves | Model score | Model |
|---|---|---|---|---|
| *VEGF* | 1 | 2 | 476.336 | -0.0166 +0.626 * (rs3025035 and rs3025039) |
| *FLT1* | 2 | 4 | 470.782 | 0.0524 -0.27 * rs17537653 +1.56 * (rs2256849 and ((not rs3936415) and (not rs9554320))) |
| *KDR* | 1 | 1 | 484.679 | 0.159 -0.269 * (not rs11941492) |
| *HIF1* | 2 | 4 | 368.132 | -2.25 +2.35 * (rs2301113 and rs1951795) +2.25 * ((not rs1951795) and (not rs11549465)) |
| *PDGFB* | 1 | 2 | 489.5 | -0.799 +0.82 * ((not rs2247128) or (not rs11704525)) |
| *TEK* | 1 | 1 | 500.886 | -0.0129 +0.0482 * rs603085 |
| *TGFB* | 1 | 1 | 497.861 | -0.0234 +0.0473 * rs1800469 |
| *TGFBR* | 1 | 1 | 484.8 | -0.152 +0.24 * (not rs10733710) |
| *IGF1R* | 1 | 1 | 499.475 | -0.00841 +0.249 * rs2139924 |
| *NFKB1* | 1 | 1 | 492.632 | -0.0109 +0.483 * rs4648072 |
| *IL8* | 2 | 3 | 490.763 | -0.131 +1.47 * (not rs4073) -1.29 * ((not rs4073) or (not rs2227307)) |
| *IL8RA* | 1 | 4 | 485.987 | 0.31 -0.335 * (rs1008562 or ((rs16858808 and rs1008563) or rs1008563)) |
| *IL8RB* | 1 | 1 | 492.01 | 0.119 -0.152 * (not rs4674258) |
| *IL1A* | 1 | 2 | 491.212 | -0.135 +0.219 * (rs1878321 or rs3783546) |
| *IL1B* | 2 | 2 | 489.699 | -0.205 +0.22 * rs1143627 +0.199 * rs1143634 |
| *TNF* | 1 | 1 | 493.9038 | 0.0132 -0.0428 * rs1800630 |
| *MMPS* | 1 | 4 | 488.361 | 0.155 -0.383 * (rs470215 or (((not rs470215) or rs3025066) and (not rs1996352))) |
| *BMP1* | 1 | 2 | 483.36 | -0.348 +0.418 * ((not rs13257482) or rs4075478) |
| *BMP2* | 1 | 3 | 487.648 | -0.239 +0.349 * (((not rs1979855) and (not rs1005464)) or (not rs3178250)) |
| *BMP4* | 1 | 2 | 492.0516 | 0.12 -0.185 * (rs17563 and (not rs2761887)) |
| *BMPR1A* | 1 | 2 | 483.277 | -0.372 +0.461 * (rs6586034 or (not rs12765929)) |
| *BMPR1B* | 1 | 1 | 489.477 | -0.0779 +0.208 * (not rs2719176) |
| *BMPR2* | 1 | 1 | 467.852 | 0.547 -0.556 * (not rs12477602) |
| *GDF10* | 1 | 1 | 490.678 | 0.0314 -0.346 * rs762454 |
| *TLR2* | 2 | 5 | 483.326 | -0.353 -0.558 * ((((not rs5743708) and (not rs5743704)) or rs4696483) or rs3804099) +0.884 * (not rs4696483) |
| *TLR3* | 1 | 2 | 492.134 | 0.0351 -0.258 * (rs3775291 and rs3775292) |

| | | | | |
|---|---|---|---|---|
| *TLR4* | 1 | 1 | 488.965 | 0.173 -0.243 * (not rs10759932) |
| *EGR2* | 2 | 3 | 488.955 | 3.88 -3.87 * rs9990 -3.88 * ((not rs9990) or rs2295814) |
| *EGFR* | 1 | 1 | 396.168 | -0.182 +0.234 * (not rs845552) |
| *IRS1* | 1 | 1 | 595.49 | 0.0169 -0.12 * IRS1 |
| *VDR* | 1 | 3 | 445.842 | 0.376 -0.434 * ((VDRBsm1 or (VDRCdx2)) or (not VDRFok1)) |

# CHAPTER 4

# Multiple Gene-Environment Interactions on the Angiogenesis Gene-Pathway Impact Rectal Cancer Risk and Survival

## 4.1 Introduction

Studying genetic variants in epidemiologic studies is of great value for identifying disease risk and outcomes. Compared to environmental exposures they are less sensitive to bias and represent valid, time-independent, biologically representative markers of disease (Le Marchand et al. 2008). Coupled with rapidly evolving technological advances, they are increasingly attractive tools to researchers especially with the emergence of genome-wide association studies (GWAS) a decade ago typically linking individual single nucleotide polymorphisms (SNPs) to disease or phenotype. GWAS have identified 18 susceptibility loci associated with colorectal cancer in European populations (Whiffin et al. 2014). Only a few of the identified SNPs have clear functional roles relevant to disease mechanisms. This is because SNP selection in GWAS is guided by linkage disequilibrium rather than functionality and it is difficult to determine whether the identified SNPs are causal or merely surrogates of the true causal variants (Hindorff et al. 2009). GWAS can be viewed as a discovery tool, without any specific hypothesis of genetic associations or any biological relevance. In addition, SNPs are typically of low-penetrance risk and despite being common, their effects are usually small and of limited preventive impact. It is possible however, that individual loci are contributing to risk through a multi-gene model.

A multi-gene model is best approached using a pathway of biological relevance to the disease of interest. One of the critical cancer-related biological processes necessary for tumor proliferation and progression in rectal cancer is the angiogenesis process: the fundamental process of

sprouting and expansion of blood vessels from preexisting vessels (Hanahan et al. 2011). Induction of angiogenesis seems to be an early event important for conversion of normal epithelium into cancer cells that influence risk of developing the disease, while sustained angiogenesis is essential for tumor expansion which may ultimately influence mortality (Ross1989, Folkman et al. 1989). In this study we focused on angiogenesis-related genes and aimed to construct a working pathway that captures the important genes working together to influence rectal cancer susceptibility and survival. The full risk model underlying rectal cancer, however, essentially involves interactions of genes with lifestyle environmental exposures (Lichtenstein et al. 2000). Identifying genetic risk modifying environmental exposure effects is, thus, critical from a prevention point of view (Giarelli et al. 2005).

High-risk genotypes modifying the effects of high-risk environmental exposures on cancer outcomes, referred to as Gene-Environment Interactions (GEIs), have become a recent focus of molecular epidemiologic studies (Thomas 2010b). Identifying GEIs may well explain an important component of the "missing heritability" (Manolio et al. 2009, Thomas2010a). For specific exposures such as cigarette smoking and alcohol consumption, the evidence of association with rectal cancer has been inconclusive (Slattery et al. 1997b,Ferrari et al. 2007,Potter1999a,Poynter et al. 2009,Cheng et al. 2014,Gong et al. 2012) and could be potentially strengthened by examining their association with rectal cancer in genetically susceptible individuals. Specifically, experimental evidence has shown that nicotine in tobacco smoke and ethanol stimulates angiogenesis under ischemic conditions (Heeschen et al. 2006, Gu et al. 2001). In addition, certain dietary patterns, specifically those that contain high consumption of red and processed meat, are associated with a moderate increased risk of rectal cancer (Gonzalez et al. 2010, Larsson et al. 2006, Chan et al. 2011). Diet, however, is a complex

mixture of many nutrients and characterization of GEI could help determine the specific nutrients affecting the cancer risk and cancer-related mortality. We focused on protein as a nutrient based on its documented stimulatory effect on angiogenesis (Vigne et al. 2006, Grandison et al. 2009).

Our approach to testing pathway GEI effects on rectal cancer risk and survival involved selection of candidate genes in the angiogenesis pathway and three environmental exposures relevant to angiogenesis. We hypothesized that high animal/vegetable protein intake ratio and prolonged intense pattern of cigarette smoking influence hypoxia (oxygen deprivation) and long-term alcohol intake influence hypoglycemia (glucose deprivation), both of which are ischemic conditions that enhance angiogenesis (Harris2002,Dor et al. 2001,Nishimoto et al. 2014). We carried the logic of the biological hypothesis to the analysis by searching for biological forms of SNP-set interactions at the gene level, gene-set interactions at the pathway level, and modeled the pathway GEI effects guided by the hypothesized pathway structure.

## 4.2 Methods

### 4.2.1 Data Sources

***Study Population***

This analysis was based on a multicenter, population-based, case-control study of rectal cancer (The Diet, Activity and Lifestyle as a Risk Factor for Rectal Cancer) conducted at two geographical areas in the United States: Utah and Northern California (Slattery et al. 2003). Rectal cancer cases were identified using a rapid-reporting system during the period between May 1997 and May 2001. Case eligibility was determined according to the Surveillance

Epidemiology and End Results (SEER) Cancer Registries in Northern California and Utah. Eligibility criteria were: being 30 to 79 years of age at time of diagnosis, speaking English, and being mentally and physically competent to complete the interview. Cases with history of previous colorectal cancer or known familial adenomatous polyposis (as indicated on pathology reports), ulcerative colitis, or Crohn's disease were not eligible. Controls were frequency matched to cases by sex and 5-year age groups in each geographical area.

***Interview data***

Trained and certified interviewers conducted a detailed computerized in-person interview that took approximately 2 to 3 hours to complete (Edwards et al. 1994). Participants completed two questionnaires: a) the health and lifestyle questionnaire (among data collected were demographic characteristics, medical history, meal patterns, smoking and alcohol consumption information); and b) a diet history questionnaire on dietary intakes. Dietary intake was ascertained using an adaptation of the CARDIA diet history (Slattery et al. 1994, Liu et al. 1994,McDonald et al. 1991). The referent period for the study questionnaires was the calendar year two to three years prior to diagnosis or from selection for controls.

***Tumor registry data***

Tumor registry data was obtained from local tumor registries to determine disease stage at diagnosis, months of survival after diagnosis, and vital status. Disease stage was categorized using the SEER staging criteria (in-situ, local, regional, distant, and unknown) (Young et al. 2001). Follow-up was obtained for all study participants and was terminated at the end of the year 2007. At that time all study participants had over five years of follow-up.

***TagSNP selection and genotyping***

TagSNPs were selected using the following parameters: LD blocks using a Caucasian LD map (International HapMap Consortium2003) and $r^2 \geq 0.8$; minor allele frequency (MAF)>0.1; LD block range= -1500 bps from the initiation codon to +1500 bps from the termination codon; and 1 tagSNP for each LD bin. All markers were genotyped using a multiplexed bead-array assay format based on Golden Gate chemistry (Illumina Human Hap550k, San Diego, California). A genotyping call rate of 99.85% was achieved. Blinded internal duplicates represented 4.4% of the total sample set; the duplicate concordance rate was 100%. TGFβ1 gene was not included in the Illumina BeadChip platform; alternatively, representative markers were genotyped using a TaqMan assay from Applied Biosystems (Foster City, California). Each 5µl PCR reaction contained 20ng of genomic DNA, primers, probes, and TaqMan Universal PCR Master Mix (containing AmpErase UNG, AmpliTaq Gold enzyme, dNTPs, and reaction buffer). PCR was carried out under the following conditions: 50$^{o}$C for 2 minutes to activate UNG, 95$^{o}$C for 10 min, followed by 40 cycles of 92$^{o}$C for 15 sec, and 60$^{o}$C for 1 minute using a 384 well duel block ABI 9700. Fluorescent endpoints of the TaqMan reactions were measured using a 7900HT sequence detection instrument. Individuals with missing genotype data were not included in the analysis for that specific marker.

***Candidate Gene-Pathway***

We constructed a working figure of the angiogenesis gene-pathway relevant to rectal cancer to guide the analysis (**Figure 4.1**). The process involved extracting information from the standard pathway maps and pathway text descriptions from three recognized web-based resources: The BioCarta organization, KEGG (Kyoto Encyclopedia of Genes and Genomes), and Cell Signaling

Technologies (CST). We specifically searched these resources using the keyword "angiogenesis" and extracted information from the "VEGF, Hypoxia, and Angiogenesis Pathway" from the BioCarta Pathways (http://www.biocarta.com/pathfiles/h_vegfPathway.asp); the "VEGF Signaling Pathway" available from the KEGG Pathway database (http://www.kegg.jp/kegg-bin/highlight_pathway?scale=1.0&map=map04370&keyword=angiogenesis); and the "Angiogenesis Signaling Pathway" from the CST pathways (http://www.cellsignal.com/common/content/content.jsp?id=pathways-angiogenesis) (See **Supplementary Figures 4.1-4.3**). We also conducted supplementary searches of online gene databases and PubMed for information on biological function of the candidate genes and experimental observations of biological activities of genes in relation to tumor angiogenesis. Examination of the molecular interactions as illustrated in the pathway maps along with their descriptions, and the information on the biological activity of the genes guided the candidate gene selection and provided rationale for grouping genes in specific sub-pathways. Genes were included in the working pathway figure as either major drivers of the angiogenesis process or interacting inflammatory genes (**Table 4.1**).

***Environmental variables***

*Smoking*

An individual was considered a regular cigarette smoker if smoked at least 100 cigarettes during a lifetime, and otherwise was classified as never having smoked. For smokers, pack-years of cigarettes smoked was determined by multiplying the usual number of cigarettes smoked per day by total years of smoking cigarettes (determined by taking into account start and stop dates of smoking), and dividing by 20 (a pack of cigarettes). For this analysis, subjects were categorized

using a cut-off of 20 pack-years (20 or more pack-years, less than 20 pack-years, and never smoked).

*Alcohol*

Participants were asked to report usual amounts consumed during the weekdays and during weekend days to better capture total alcohol consumption. Additionally, participants were asked about alcohol consumption 10 and 20 years ago as part of the health and lifestyle questionnaire. Alcoholic beverages were defined as beer, wine, and hard liquor including alcoholic cocktails, whiskey, gin, vodka, scotch, bourbon, or rum. Participants who responded with "no" to the question, "Did you ever drink an average of one or more alcoholic beverages a month for a year or longer?" were considered as never to have drunk alcohol. Participants who responded "yes" to this question were then asked the usual number of 12-ounce bottles of beer, 4-ounce glasses of wine, and 1.5-ounce shots of hard liquor consumed 10 and 20 years ago. Long-term exposure to alcohol, based on consumption of any type of alcoholic beverage 10 and 20 years prior to the referent year, was categorized in two levels (none to moderate and high alcohol consumption, cut-off was 20gms/week for men and 10gms/week for women).

*Dietary Protein*

Nutrient information was obtained by converting food-intake data into nutrient data using the Minnesota Nutrition Coordinating Center nutrient database (Dennis et al. 1980). Total protein intake included animal proteins (meats, poultry, fish, dairy, and eggs) and vegetable proteins (legumes, tofu). We calculated an animal/vegetable protein intake ratio and used a cut-off corresponding to the median of animal protein proportion of total protein intake (i.e. 60% of total protein intake is animal protein equivalent to a 1.5 animal/vegetable protein intake ratio, which

means 50% more animal protein intake than vegetable protein intake). This resulted in two categories (low and high animal/vegetable protein intake ratio).

### 4.2.2 Statistical analysis

We applied a 3-step analysis framework to modeling candidate pathway GEIs consisting of the following steps:

1. Step 1, developing a summary profile for each gene on the candidate pathway referred to as *gene-specific tree* (GST). We used logic regression (Ruczinski et al. 2003) to search for SNP-set interactions within each gene. Details on the method are provided in supplementary materials. The GSTs, rather than individual SNPs, were used as building blocks for the next two steps.
2. Step 2, modelling gene-set interactions across the full pathway referred to as *pathway tree(s)* by searching for GST-set interactions using logic regression. Pathway trees are adjusted for in the GEI models of the next step.
3. Step 3, modelling pathway GEIs between the GSTs and the three environmental exposures. Guided by the pathway figure (**Figure 4.1**), we first divided the full pathway into nine sub-pathways (grouped in boxes in the figure) and summarized GEIs in each sub-pathway using backward selection that eliminated the least significant interaction term(s) in a stepwise fashion. The GEIs that remained in the sub-pathway summary models at the 5% significance level were jointly tested in the final GEI model for the entire pathway. We fitted logistic regression models for rectal cancer risk and Cox proportional hazard regression models for rectal cancer survival. All models in addition to adjusting for the pathway trees, were also adjusted for age at diagnosis or selection, sex, race (white, Hispanic, or black race), and study center (University of Utah and the

Kaiser Permanente Medical Care Program of Northern California). Baseline hazards for Cox proportional hazards models were stratified by rectal cancer stage at diagnosis. The GEI models were fitted using Stata version 12.

A summary of the three steps of the analysis approach are shown in **Tables 4.2A and 4.2B** for rectal cancer risk and survival, respectively.

## 4.3 Results

We analyzed data of 747 rectal cancer cases and 956 controls. The angiogenesis candidate gene-pathway included a total of 257 SNPs belonging to 34 angiogenesis-related genes (**Table 4.1**). The GSTs developed from the first step of the analysis are shown in Supplementary **Tables 4.1 and 4.2** for rectal cancer risk and survival, respectively. The tables include a list of the genes, the corresponding optimal model size as determined by the cross-validation, the score of the final model, and the SNPs forming the GSTs. The pathway trees resulting from the second step of the analysis are shown in **Supplementary Figures 4.3 and 4.4.** The third and final step of the analysis modeled the pathway GEIs; results are displayed in **Tables 4.3** for rectal cancer risk and **Table 4.4** for rectal cancer survival. For all significant GEIs, we observed a positive gradient in the magnitude of the main GST effects with increasing levels of animal protein intake, smoking and alcohol consumption.

Eight significant GEIs were associated with rectal cancer risk involving six genes. Two genes were among the major drivers of angiogenesis: *PDGFB* rs4821877 with high animal/vegetable protein intake (interaction odds ratio ($OR_{INT}$) = 1.75, 95% confidence interval (CI) (1.04, 2.92), p-value= 0.034) and *IGF1R* rs2139924 with long-term alcohol consumption ($OR_{INT}$ =1.69, 95% CI (1.04, 2.72), p-value= 0.033). Other significant GEIs were: *TNF* rs1800630 ($OR_{INT}$ =1.85,

95% CI (1.10, 3.11), p-value= 0.021) with ≥ 20 pack-years of smoking and *MMP1* rs470215 ($OR_{INT}$ =2.44, 95% CI (1.24, 4.81), p-value= 0.010). Both complementary GSTs for *TLR4* and *EGR2* genes were interacting with both smoking and alcohol consumption. *TLR4* rs1927911 AND rs11536889 with ≥ 20 pack-years of smoking ($OR_{INT}$ =2.34, 95% CI (1.38, 3.98), p-value= 0.002), *TLR4* rs1927911 OR rs11536889 with long-term alcohol consumption ($OR_{INT}$ =2.10, 95% CI (1.22, 3.60), p-value= 0.007); *EGR2* rs2295814 with ≥ 20 pack-years of smoking ($OR_{INT}$ =2.23, 95% CI (1.04, 4.78), p-value= 0.040) with long-term alcohol consumption ($OR_{INT}$ =2.12, 95% CI (1.01, 4.46), p-value= 0.048).

Five GEIs were associated with survival, four of which were interactions with high animal/vegetable protein intake: *KDR* rs6838752 (interaction hazard ratio $HR_{INT}$=4.12, 95% CI (1.52, 11.13), p-value= 0.005), *TLR2* rs7656411 ($HR_{INT}$=8.69, 95% CI (1.09, 69.12), p-value= 0.041), *EGR2* rs224082 ($HR_{INT}$=2.41, 95% CI (1.40, 4.15), p-value= 0.002), and *EGFR* rs17151957 ($HR_{INT}$=5.84, 95% CI (1.80, 18.94), p-value= 0.003). The fifth significant interaction was *IL8RA* rs1008562 with ≥ 20 pack-years of smoking ($HR_{INT}$=2.05, 95% CI (1.12, 3.76), p-value= 0.019).

## 4.4 Discussion

We focused on the angiogenesis pathway, a biologic pathway relevant to rectal cancer outcomes, selected three environmental exposures relevant to angiogenesis and applied a gene-environment interaction analysis approach that captured underlying biologic forms of interaction within genes and between genes and environmental exposures. Specifically, we constructed a working pathway figure of select angiogenesis-related genes and used it to guide the analysis. We used logic regression to summarize SNP-set interactions within each gene of the angiogenesis

pathway and gene-set interactions across the full pathway, and modeled gene-environment interactions crossing the gene-level summaries from logic regression with dietary protein intake, smoking, and alcohol consumption. Eight interactions for rectal cancer risk and five for rectal cancer survival were statistically significant at the 5% level.

A recent approach to evaluating gene-environment interaction in cancer is through investigating interactions between known common susceptibility loci (i.e., strong and statistically significant GWAS or candidate-gene findings) and established risk factors of the cancer. These studies are generally of large sample size and may involve combined case-control and/or nested case-control samples. Despite the advantages of their size and design they have provided limited evidence of GEI in colorectal cancer (Figueiredo et al. 2011,Hutter et al. 2012,Siegert et al. 2013) and similarly in breast (Campa et al. 2011,Travis et al. 2010) and prostate cancers (Lindstrom et al. 2011). It has been argued that these studies may be missing interactions that would have been found had they considered different environmental exposures, measured them differently, and/or used different models (Prentice 2011). Another possibility is relying on a hypothesis-driven approach that focuses on environmental factors relevant to the studied genes.

Our choice of environmental variables was based on the hypothesis that protein intake, smoking and alcohol are enhancing angiogenesis and potentially interacting with the angiogenesis genes in the state of tumor ischemia. Lifestyle factors such as smoking and alcohol consumption could be considered more strictly "environmental", having less genetic influence compared to other complex risk factors with more pronounced genetic influence (e.g., body mass index). We also focused on intense and long-term patterns of smoking and alcohol consumption (i.e., considering both amount and duration of exposure) as compared to never exposed participants which maximizes power of detecting an interaction with the gene and avoids dilution of risk by short-

term and/or distant former users. It has also been suggested that modelling SNP genotypes as indicator variables for one and two minor alleles rather than the number of SNP minor alleles strengthens the detected association (Prentice et al. 2009). It is possible that certain environmental effects are localized to a specific SNP genotype (Prentice 2011). Logic regression models interactions of binary predictors and we used indicators of SNP genotypes to develop the gene-specific trees. These considerations embedded in the framework of our approach potentially enhanced its capacity to detect GEI in rectal cancer.

Several studies supported a potential role of high intake of red and processed meat on colorectal cancer risk, yet the evidence remains insufficient (Alexander et al. 2010, Alexander et al. 2011). Plausible mechanisms were related to the content of meat (protein, iron) (Sun et al. 2012b)or compounds generated by the cooking process (N-nitroso compounds, heterocyclic aromatic amines) (Zhu et al. 2014). These compounds that are not absorbed by the small intestine are transferred to the lumen of the large intestine lumen and, when in excess, may have toxic effects on the large intestine mucosa (Kim et al. 2013). The higher intake of protein and a decrease in its digestibility leads to more undigested proteins reaching the colon and being fermented by colonic bacteria. Increased rate of protein fermentation may promote DNA damage and loss of large intestine epithelial cell homeostasis leading to an imbalance between new and dying cells and ultimately tumor growth (Kim et al. 2013). Protein fermentation mainly occurs in the distal parts of the colon and rectum (Silvester et al. 1995, Chao et al. 2005) and previously reported associations of meat intake have been generally stronger for distal colon and rectal cancer. High energy intake has also been linked with increased risk of colorectal cancer but not diet high in protein (Sun et al. 2012a); however animal protein intake specifically has been previously associated with colorectal adenoma (Yang et al. 2012).

In our results, four of the five observed GEIs on rectal cancer survival were with high animal protein intake. The mechanisms relating animal protein to rectal cancer and the specific gene functions could plausibly explain such observations. One of the observed interactions is with the *KDR* gene. *KDR* is the *VEGF* receptor that mediates *VEGF-A* induced production of Nitric Oxide (NO) by endothelial cells (Kroll et al. 1998). Apart from the processing of meat, a high protein diet leads to a high amine concentration (due to the excess intake and increased fermentation) which in the presence of NO yields the potentially carcinogenic nitrosoamines (Nitric Oxide (NO) added to the amine). *TLR2* and *EGR2* genes are both related to the same signaling pathway, and evidence has shown *TLR* expression and signaling mediates the response of intestinal epithelial cells to bacterial antigens possibly increasing the rate of protein fermentation (Singh et al. 2005). Western style diet has been linked to colorectal cancer (Slattery2000,Slattery et al. 2000,Murtaugh et al. 2004) and experimental evidence showed *EGFR* was required for tumor promotion by the western style diet and fat rich diet (Dougherty et al. 2009,Dougherty et al. 2011). Western diet is rich in meat and although these studies describe the high fat content, this may help explain the *EGFR* and animal protein interaction effects on rectal cancer.

Recently, a genome-wide analysis identified an interaction between a SNP on chromosome 10p14 near the *GATA3* gene and processed meat that modified colorectal cancer risk (Figueiredo et al. 2014). The authors suggested GATA3 transcription triggers a pro-tumorigenic inflammatory response of processed meat on colorectal cancer. In our results, we identified interactions of high animal protein intake on rectal cancer risk with *MMP1* gene previously implicated in inflammatory mediated pathological processes including tumor progression (Brinckerhoff et al. 2000). Based on evidence that GATA3 transcription factor was found to

potentially mediate different expression levels of *MMP1* (Affara et al. 2011), it is possible that our observed MMP1-animal protein GEI is providing further characterization of the GATA3-processed meat interaction. This also supports a candidate pathway approach in identifying biologically plausible GEI effects.

Toll-like receptors (TLRs) play a key role in the innate immune system and are important mediators of inflammation in the gut, potentially modulating colorectal cancer risk. Tissue expression studies suggest involvement of TLR2 (Nihon-Yanagi et al. 2012) and TLR4 (Tchorzewski et al. 2014) in colorectal carcinogenesis. Functional polymorphisms in both genes were also recently found in association with colorectal cancer risk, and their effects were modified by obesity and smoking (Pimentel-Nunes et al. 2013). We have previously reported on associations of *TLR2* and *TLR4* SNPs with colon cancer risk and survival (Slattery et al. 2012a). In this analysis, we observed interactions of *TLR4* with smoking and alcohol on rectal cancer risk and *TLR2* with animal protein on rectal cancer survival. Tumor promotion by triggering an inflammatory response provides biological plausibility to these interactions. For example, cigarette smoke has been shown to activate NF-κB (Anto et al. 2002), one pathway through which the inflammatory response of TLRs is mediated. Inhibition of NF-κB dependent intestinal inflammation was recently attained by targeting an enteroglial-specific protein/TLR4 axis demonstrating therapeutic effects in ulcerative colitis (Esposito et al. 2013). Potential similar effects could provide insights into new drug targets for colorectal cancer. Further research implicating TLR genes and their interactions with lifestyle factors could provide important insights into drug targets for colorectal cancer.

We performed secondary analysis of available case-control study data of rectal cancer and had no control over the sample size. , We detected statistically significant GEIs in our study and thus

a power calculation was not needed. Data were collected through a standardized interview process to minimize interviewer bias, long-term exposure information for smoking and alcohol were collected, and availability of confounding variables for model adjustment were available. The interviewer-administered questionnaires were extensive and captured more detailed exposure information than is available from self-administered questionnaires. The major strength of the analysis, however, was our integration of the relevant biologic information in the construction of the pathway that was carried throughout the analysis process. The GEI models were large in size involving a large number of GST-environment interactions across the pathway; accordingly we limited the adjustment variables to select CRC risk and survival predictors. Although we adjusted for important predictors (age, sex, race, study center, and cancer stage), we did not adjust for further CRC-relevant factors such as tumor microsatellite instability (MSI) status. Although it is possible that we missed important angiogenesis genes when developing the working pathway figure, our candidate pathway has involved a relatively large number of genes implicated in rectal carcinogenesis. We used cross-validation to specify model size for the logic regression models and as such summarization of the gene effects was limited by the specified model size in addition to the number of tagSNPs on each gene. Our candidate approach compared to a pure empirical approach to examining GEI, however, was able to detect an appreciable number of novel GEIs. Furthermore, since the candidate associations were biologically hypothesized a priori, a multiple testing adjustment for testing many non-hypothesized associations in GWAS was not required for our candidate gene-pathway analyses.

Our approach to pathway analysis provides a powerful tool to elucidate the overall effects of the angiogenesis pathway genes and their interaction with the three exposures on rectal cancer outcomes. The angiogenesis pathway is one of the hallmarks of cancer, and findings could be

potentially informative to other solid tumors. The diet and lifestyle factors are modifiable factors and are theoretically preventable and also considering the large magnitude of the detected GEIs the potential preventive impact is increased. In addition to essential insights for preventive strategies, GEI studies are useful for identifying drug targets and opens avenues for personalized preventive and treatment strategies.

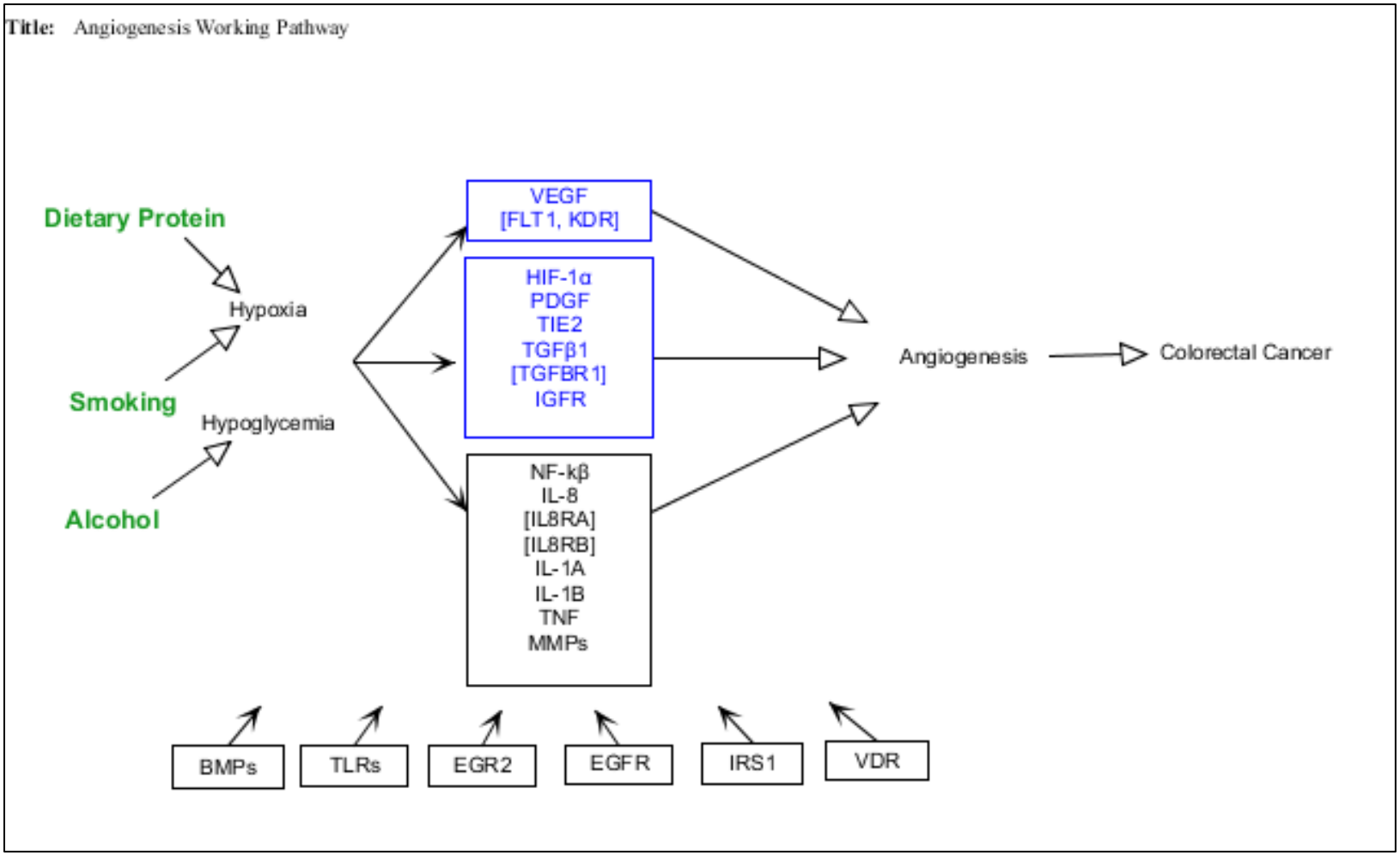



**Figure 4.1: Working figure of the angiogenesis pathway genes.**

Key gene components of the pathway are in blue frames; secondary genes are in black frames; environmental factors are in green text.

**Table 4.1:** Gene list in angiogenesis pathway

| Genes | Name |
|---|---|
| **Key components of angiogenesis pathway** | |
| *VEGFA* | Vascular endothelial growth factor A |
| *FLT1* | Vascular endothelial growth factor receptor 1 |
| *KDR* | Vascular endothelial growth factor receptor 2 |
| *HIF-1α* | Hypoxia-inducible factor 1, alpha |
| *PDGF* | Platelet-derived growth factor |
| *TIE2* | Tyrosine-protein kinase receptor |
| *TGFβ* | Transforming growth factor, beta |
| *TGFβR* | Transforming growth factor, beta receptor |
| *IGF-IR* | Insulin-like growth factor-I receptor |
| **Interacting inflammatory genes** | |
| *NFKB1* | Nuclear factor of kappa light polypeptide gene enhancer in B-cells 1 |
| *IL8* | Interleukin-8 |
| *IL8RA* | Interleukin-8 receptor, alpha |
| *IL8RB* | Interleukin-8 receptor, beta |
| *IL1A* | Interleukin-1, alpha |
| *IL1B* | Interleukin-1, beta |
| *TNF* | Tumor necrosis factor |
| *MMPs* | Matrix metallopeptidases (*MMP1, MMP3, MMP7, MMP9*) |
| *BMPs* | Bone morphogenetic protein (*BMP1,BMP2,BMP4,BMPR1A,BMPR1B,BMPR2*) |
| *TLRs* | Toll-like Receptor (*TLR2, TLR3, TLR4*) |
| *EGR2* | Early Growth response 2 |
| *EGFR* | Epidermal growth factor receptor |
| *IRS1* | Insulin receptor substrate 1 |
| *VDR* | Vitamin D Receptor |

**Table 4.2: Summary of the 3-step candidate pathway gene-environment interaction approach**

**C. Rectal cancer risk analysis steps**

| Analysis Step | Interaction of interest | Variable of interest | Model | Specific Procedures | Product |
|---|---|---|---|---|---|
| **Step 1:** Summarize gene effects | SNP-set interaction within gene | SNPs on each gene separately | Logic regression with logit link | Cross-validation to determine optimal model size | Gene-specific trees (GSTs) |
| **Step 2:** Summarize pathway effects | Gene-set interaction within pathway | All GSTs on the pathway | Logic regression with logit link | Cross-validation to determine optimal model size | Pathway Trees |
| **Step 3:** Test gene-environment interaction | Gene-environment interaction within pathway | a. Sub-pathway specific GSTxE* b. Full pathway GSTxE | Logistic regression model[¥] | Statistical significance testing | Pathway GEIs |

* GSTxE, gene-specific tree - environment interaction
[¥]Models adjusted for age, sex, race, study center, pathway trees

**D. Rectal cancer survival analysis steps**

| Analysis Step | Interaction of interest | Variable of interest | Model | Specific Procedures | Product |
|---|---|---|---|---|---|
| **Step 1:** Summarize gene effects | SNP-set interaction within gene | SNPs on each gene separately | Logic regression fitting exponential survival models | Cross-validation to determine optimal model size | Gene-specific trees (GSTs) |
| **Step 2:** Summarize pathway effects | Gene-set interaction within pathway | All GSTs on the pathway | Logic regression fitting exponential survival models | Cross-validation to determine optimal model size | Pathway Trees |
| **Step 3:** Test gene-environment interaction | Gene-environment interaction within pathway | a. Sub-pathway specific GSTxE* b. Full pathway GSTxE | Cox Proportional Hazards model[¥] | Statistical significance testing | Pathway GEIs |

* GSTxE, gene-specific tree - environment interaction
[¥]Models adjusted for age, sex, race, study center, pathway tree, stratified by cancer stage

**Table 4.3: Effects of gene-environment interactions significant at 5% level between rectal cancer gene-specific trees and environmental factors on rectal cancer risk**

| Gene-Specific Tree | Gene | Chr | Cases (%) | Control (%) | Gene OR* (95%CI) | Env Factor | Category | N (%) | Gene OR by Env Factor* (95%CI) | $OR_{INT}$* (95%CI) | $P_{INT}$* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| rs4821877 (CC or CT) | *PDGFB* | 22q13.1 | 610 (80.7%) | 746 (77.6%) | 1.21 (0.95, 1.54) | Animal/ Vegetable Protein Ratio | Low | 612 (35.6%) | 0.85 (0.57, 1.26) | Ref | |
| | | | | | | | High | 1106 (64.4%) | **1.47 (1.08, 2.00)** | **1.75 (1.04, 2.92)** | **0.034** |
| rs2139924 (AA) | *IGF1R* | 15q26.3 | 243 (30.4%) | 287 (28.5%) | 0.93 (0.77, 1.00) | Alcohol | Non/Moderate | 1396 (77.4%) | 0.82 (0.66, 1.02) | Ref | |
| | | | | | | | Heavy | 408 (22.6%) | 1.36 (0.91, 2.05) | **1.69 (1.04, 2.72)** | **0.033** |
| rs1800630 (CA or AA) | *TNF* | 6p21.3 | 240 (31.8%) | 267 (27.8%) | 1.19 (0.96, 1.47) | Smoking | Non | 834 (48.7%) | 0.95 (0.70, 1.30) | Ref | |
| | | | | | | | < 20 PY | 400 (23.4%) | 1.13 (0.71, 1.81) | 1.14 (0.65, 2.01) | 0.644 |
| | | | | | | | ≥ 20 PY | 477 (27.9%) | **1.68 (1.11, 2.54)** | **1.85 (1.10, 3.11)** | **0.021** |
| rs470215 (TT or TC) | *MMP1* | 11q22.3 | 715 (90.3%) | 880 (87.9%) | 1.23 (0.89, 1.69) | Animal/ Vegetable Protein Ratio | Low | 640 (35.7%) | 0.67 (0.40, 1.14) | Ref | |
| | | | | | | | High | 1153 (64.3%) | **1.78 (1.18, 2.70)** | **2.44 (1.24, 4.81)** | **0.010** |

*(Table continues)*

**Table 4.3 (Continued).**

| Gene-Specific Tree | Gene | Chr | Cases (%) | Control (%) | Gene OR* (95%CI) | Env Factor | Category | N (%) | Gene OR by Env Factor* (95%CI) | $OR_{INT}$* (95%CI) | $P_{INT}$* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| rs1927911 (CC) **AND** rs11536889 (GG) | *TLR4* | 9q32-q33 | 396 (52.4%)<br>546 (72.2%) | 495 (51.5%)<br>684 (71.1%) | 0.93 (0.76, 1.15) | Smoking | Non<br>< 20 PY<br>≥ 20 PY | 834 (48.7%)<br>400 (23.4%)<br>477 (27.9%) | 0.80 (0.59, 1.08)<br>0.65 (0.41, 1.04)<br>1.33 (0.90, 1.98) | Ref<br>0.99 (0.57, 1.74)<br>**2.34 (1.38, 3.98)** | <br>0.980<br>**0.002** |
| rs1927911 (CT or TT) **OR** rs11536889 (GC or CC) | *TLR4* | 9q32-q33 | 360 (47.6%)<br>210 (27.8%) | 467 (48.5%)<br>278 (28.9%) | 1.07 (0.87, 1.32) | Alcohol | Non/Moderate<br>Heavy | 1326 (77.2%)<br>408 (22.6%) | 0.95 (0.75, 1.21)<br>1.58 (1.01, 2.47) | Ref<br>**2.10 (1.22, 3.60)** | <br>**0.007** |
| rs2295814 (GA or AA) | *EGR2* | 10q21.1 | 106 (14.0%) | 115 (12.0%) | 1.11 (0.83, 1.49) | Smoking | Non<br>< 20 PY<br>≥ 20 PY | 834 (48.7%)<br>400 (23.4%)<br>477 (27.9%) | 0.90 (0.58, 1.37)<br>1.21 (0.65, 2.28)<br>1.53 (0.89, 2.65) | Ref<br>1.84 (0.83, 4.09)<br>**2.23 (1.04, 4.78)** | <br>0.130<br>**0.040** |
| rs2295814 (GG) | *EGR2* | 10q21.1 | 650 (86.0%) | 847 (88.0%) | 0.90 (0.67, 1.20) | Alcohol | Non/Moderate<br>Heavy | 1326 (77.2%)<br>391 (22.8%) | 0.81 (0.57, 1.14)<br>1.21 (0.69, 2.11) | Ref<br>**2.12 (1.01, 4.46)** | <br>**0.048** |

Abbreviations: Chr, Chromosome; Env, Environmental; PY, pack-years; OR, odds ratio; P, p-value; INT, interaction
*Adjusted for age, sex, race, study center, pathway trees

**Table 4.4: Effects of gene-environment interactions significant at 5% level between rectal gene-specific trees and environmental factors on rectal cancer survival**

| Gene-Specific Tree | Gene | Chr | Cases (%) | Gene HR* (95%CI) | Env Factor | Category | N (%) | Gene OR by Env Factor* (95%CI) | $HR_{INT}$* (95%CI) | $P_{INT}$* |
|---|---|---|---|---|---|---|---|---|---|---|
| rs6838752 (TT or TC) | *KDR* | 4q11-q12 | 705 (93.6%) | 0.89 (0.55, 1.45) | Animal/ Vegetable Protein Ratio | Low | 258 (32.4%) | **0.44 (0.21, 0.91)** | Ref | |
| | | | | | | High | 538 (67.6%) | 1.43 (0.73, 2.83) | **4.12 (1.52, 11.13)** | **0.005** |
| rs1008562 (GG) | *IL8RA* | 2q35 | 211 (27.9%) | 1.17 (0.89, 1.53) | Smoking | Non | 348 (46.2%) | 1.04 (0.68, 1.60) | Ref | |
| | | | | | | < 20 PY | 160 (21.2%) | 0.88 (0.44, 1.75) | 0.96 (0.46, 1.98) | 0.905 |
| | | | | | | ≥ 20 PY | 245 (32.5%) | **1.88 (1.20, 2.95)** | **2.05 (1.12, 3.76)** | **0.019** |
| rs7656411 (GG) | *TLR2* | 4q32 | 61 (8.1%) | 0.83 (0.48, 1.44) | Animal/ Vegetable Protein Ratio | Low | 244 (32.3%) | **0.13 (0.02, 0.98)** | Ref | |
| | | | | | | High | 512 (67.7%) | 1.33 (0.74, 2.38) | **8.69 (1.09, 69.12)** | **0.041** |
| rs224082 (GA or AA) | *EGR2* | 10q21.1 | 455 (60.2%) | **0.72 (0.56, 0.92)** | Animal/ Vegetable Protein Ratio | Low | 244 (32.3%) | **0.39 (0.25, 0.62)** | Ref | |
| | | | | | | High | 512 (67.7%) | 0.93 (0.68,1.23) | **2.41 (1.40, 4.15)** | **0.002** |
| rs17151957 (AA) | *EGFR* | 7p12 | 41 (6.5%) | **1. 82 (1.16, 2.88)** | Animal/ Vegetable Protein Ratio | Low | 244 (32.3%) | 0.54 (0.19, 1.53) | Ref | |
| | | | | | | High | 512 (67.7%) | **3.37 (1.95, 5.82)** | **5.84 (1.80, 18.94)** | **0.003** |

Abbreviations: Chr, Chromosome; Env, Environmental; PY, pack-years; HR, hazard ratio; P, p-value; INT, interaction
*Adjusted for age, sex, race, study center, pathway tree, baseline hazard stratified by cancer stage

**Supplementary Information**

***Logic Regression***

Logic regression is a methodology that searches for Boolean combinations of binary predictors (e.g., SNPs) to detect high-order interactions and their patterns within a regression framework. The SNPs in a Boolean combination are referred to as "leaves" and the combination of the SNPs joined by the Boolean operators, □ (AND), □ (OR), and $^{c}$ (NOT), is referred to as a "logic tree". The logic trees are also binary variables taking the value of "0" or "1", or "Yes" or "No".

We implemented the logic regression in R version 3.0.0 using the "LogicReg" R package (Charles Kooperberg and Ingo Ruczinski (2013). LogicReg: Logic Regression. R package version 1.5.5. http://CRAN.R-project.org/package=LogicReg). We used logic regression models fitting logistic models to assess rectal cancer risk (scoring function: deviance); and exponential survival models (scoring function: negative log-likelihood) to assess rectal cancer survival. Categorical SNP genotype variables (coded: 0 for major-allele homozygotes (reference category), 1 for heterozygotes, and 2 for minor-allele homozygotes) were transformed into two binary dummy/indicator variables for having one and two minor SNP alleles. Specifically, the logic model with logit link took the form:

$$\log (\Pr[Y=1]/ \Pr[Y=0]) = \beta 0 + \beta 1\ L1 + \beta 2\ L2 + \ldots + \beta p\ Lp$$

where Y is a binary response variable, β0, β1,…, βp are the model parameters, and L1, L2, …, Lp are the Boolean combinations of SNPs.

The logic model for exponential survival took the form:

$$\log \lambda(c) = \beta 0 + \beta 1\ L1 + \beta 2\ L2 + \ldots + \beta p\ Lp$$

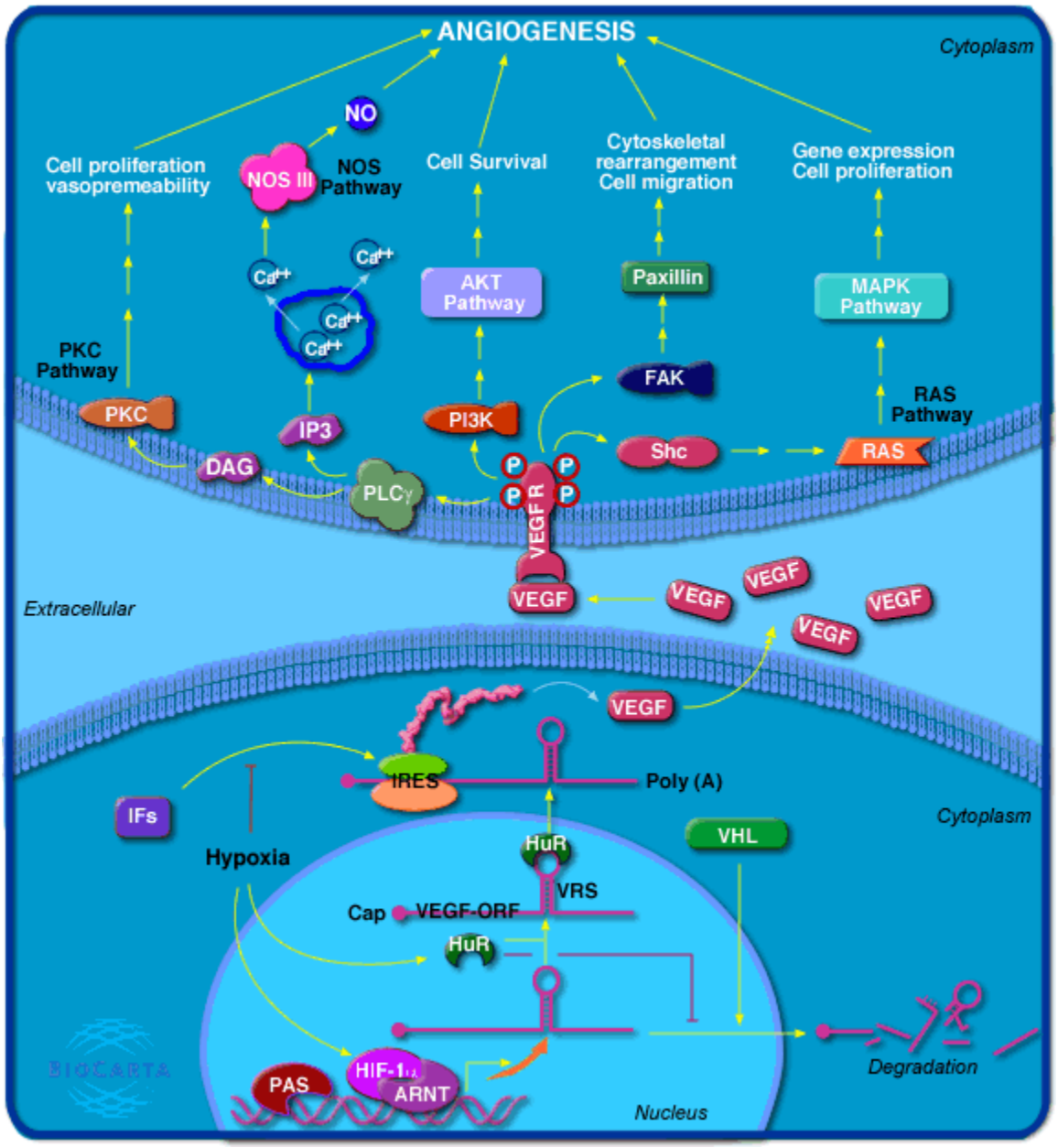
where $\lambda(c)$ is the hazard rate, a function of the marginal cumulative hazard c, $\beta 0$, $\beta 1$,…, $\beta p$ are the model parameters, and L1, L2, …, Lp are the Boolean combinations of SNPs.

Considering the large search space, defined by the number of SNPs and all their possible combinations, the logic regression needs to employ an efficient search strategy. One of the search algorithms to select the logic trees implemented in logic regression is the simulated annealing algorithm (Schwender et al. 2010). It basically involves, given a certain tree, picking a single move at a time from a set of six permissible moves (and counter moves) that leads to a new logic tree. The acceptance probability of the new model is dependent on the scores of both the old and new models and the stage of the annealing process. The further ahead in the annealing scheme the lower the acceptance probability if the new model has a worse score. To avoid over fitting in logic regression models it is necessary to employ a model selection procedure for the simulated annealing algorithm. Model selection involves determining the optimal model size defined as the number of logic trees and number of leaves in the logic trees. One of the methods implemented in the 'LogicReg' R package to derive the optimal model size is cross-validation. A desired maximum fixed size is indicated and if reached the search algorithm prohibits further moves that increase the trees/leaves over the desired size. The final model size is usually smaller. We implemented 10-fold of cross-validations for all models with a maximum desired size of 9 logic trees and 20 leaves. We fitted the optimal-size model a 100 times, each with a different random seed (i.e. starting point for the search), and the model with the smallest scoring function was considered as the best solution.

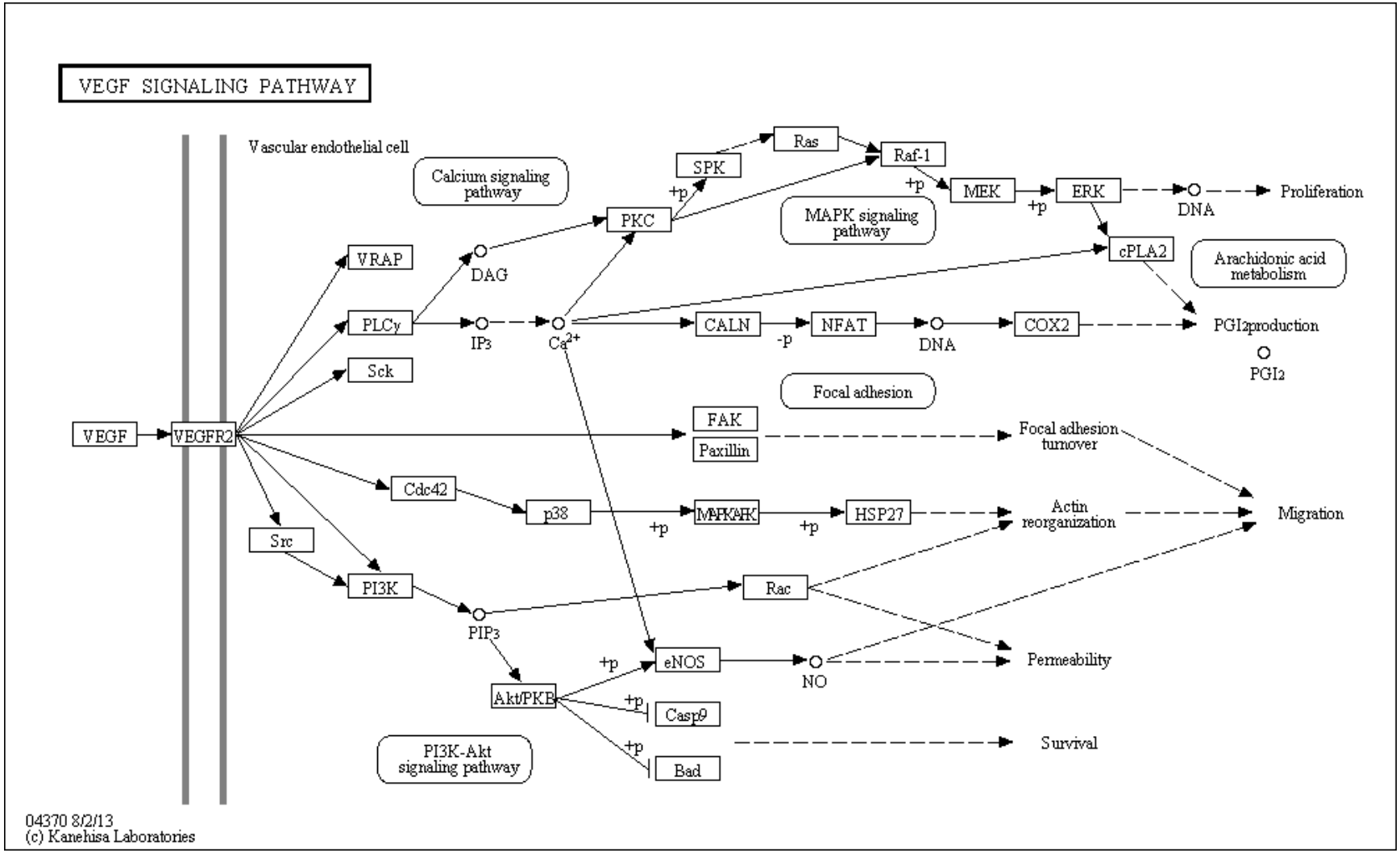
***Biological interactions between genetic variants***

We explored two forms of biologically plausible SNP-set interactions derived from set theory terminology: SNP *intersection* and SNP *union*. A SNP intersection is a form of interaction where

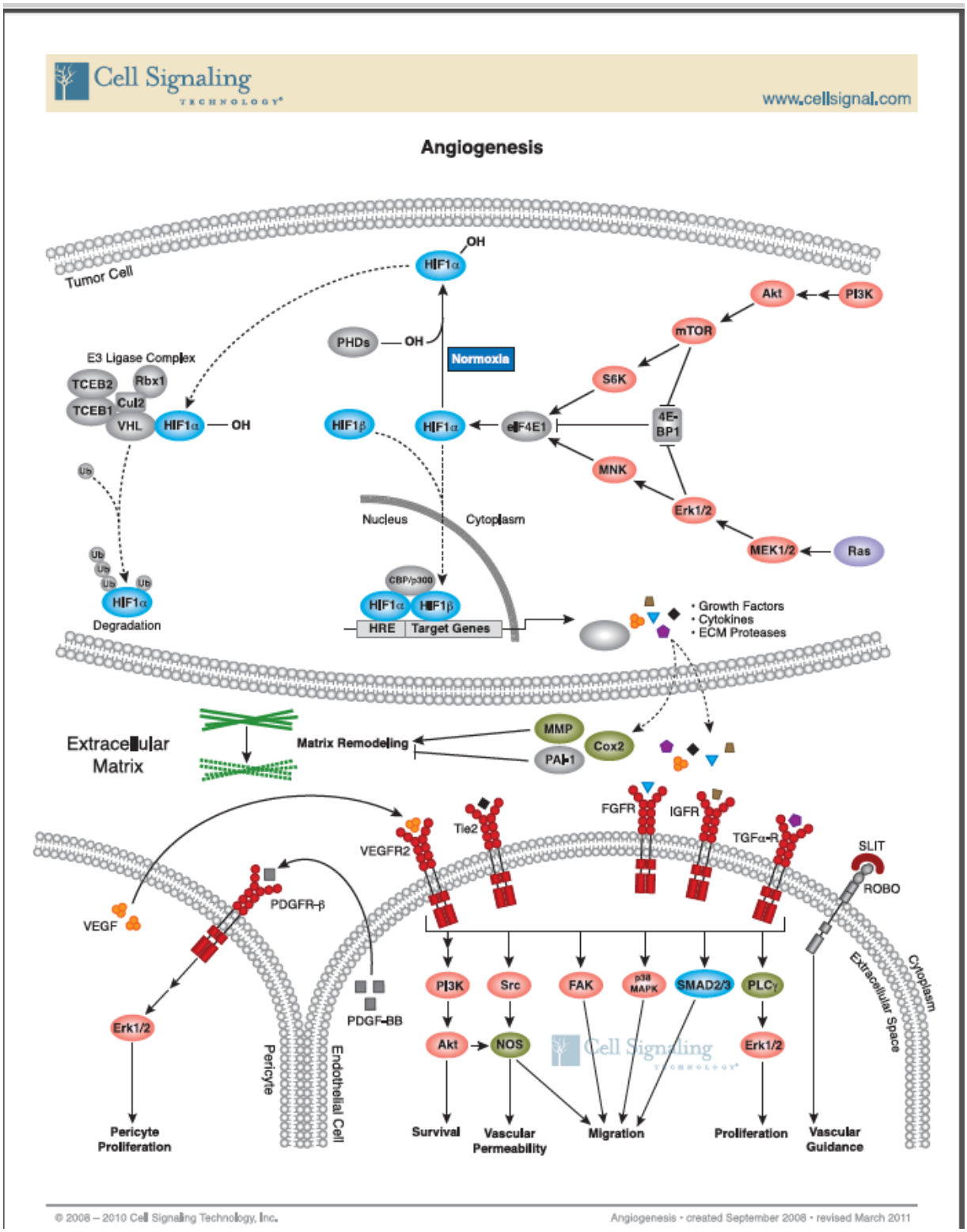
disease risk is elevated only if *all* of the SNPs in a specified set (e.g., a gene) carry their respective high-risk genotype. A single SNP, or subsets, of the set carrying the high-risk genotype are insufficient to elevate disease risk. For example, for a set of three SNPs, all three SNPs (SNP 1 *and* SNP 2 *and* SNP 3) may have to carry their high-risk genotype for disease risk to be elevated. A SNP union describes a form of interaction where disease risk may be elevated through several independent ways (i.e., genetic heterogeneity) which may include a SNP intersection (e.g., SNP 1 and SNP 2) or an individual SNP carrying the high-risk genotype. We applied logic regression (Ruczinski et al. 2003) to search for these biologically plausible forms of SNP-set interactions within genes (Dinu et al. 2012).




**Supplementary Figure 1:** VEGF, Hypoxia, and Angiogenesis Pathway. Illustration reproduced courtesy of The BioCarta Pathways (http://www.biocarta.com/pathfiles/h_vegfPathway.asp).



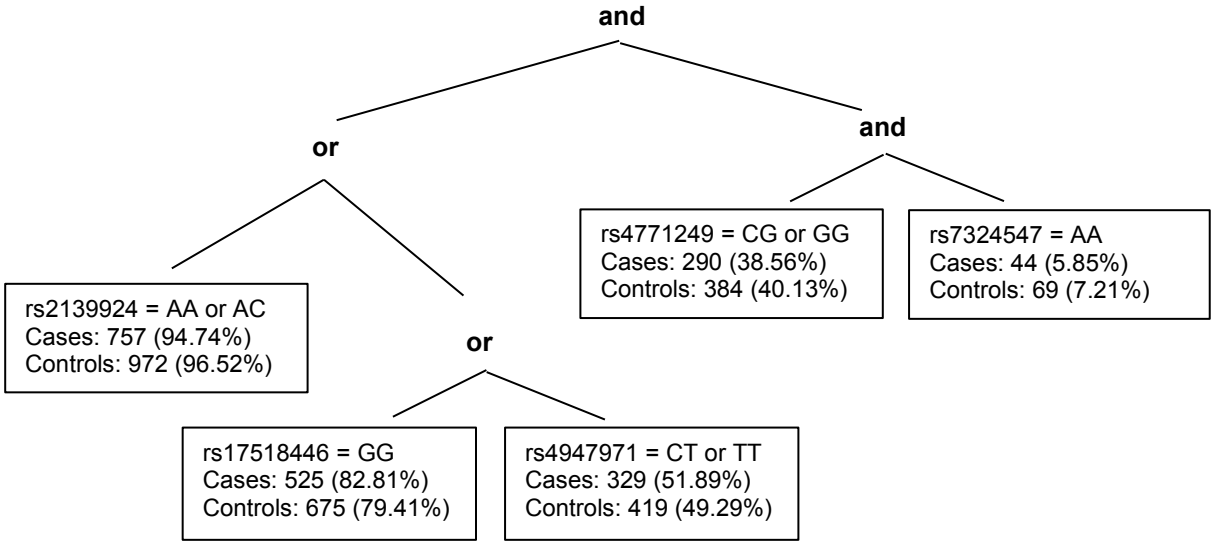

**Supplementary Figure 2:** VEGF Signaling Pathway. Illustration reproduced courtesy of KEGG, (Kyoto Encyclopedia of Genes and Genomes) Pathway database (http://www.kegg.jp/kegg-bin/highlight_pathway?scale=1.0&map=map04370&keyword=angiogenesis).




**Supplementary Figure 3:** Angiogenesis Signaling Pathway. Illustration reproduced courtesy of Cell Signaling Technology, Inc. (www.cellsignal.com). (http://www.cellsignal.com/common/content/content.jsp?id=pathways-angiogenesis).

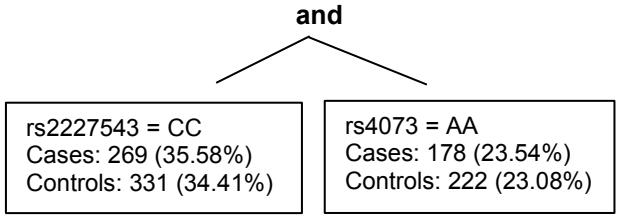
**Logic 1**

1.68 * (((***IGF1R***_1) or ***EGFR***_L1) and (***FLT1***_L1))



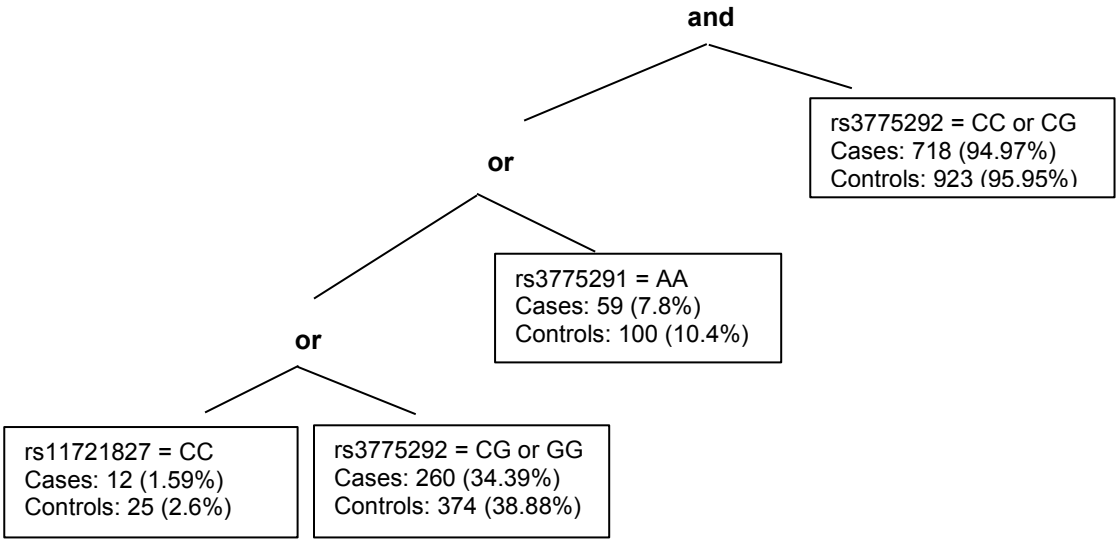

**Logic 2**

2.65 * ***IL8***_L1



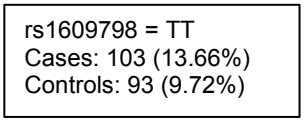

**Logic 3**

-0.475 * ***TLR3***_L1



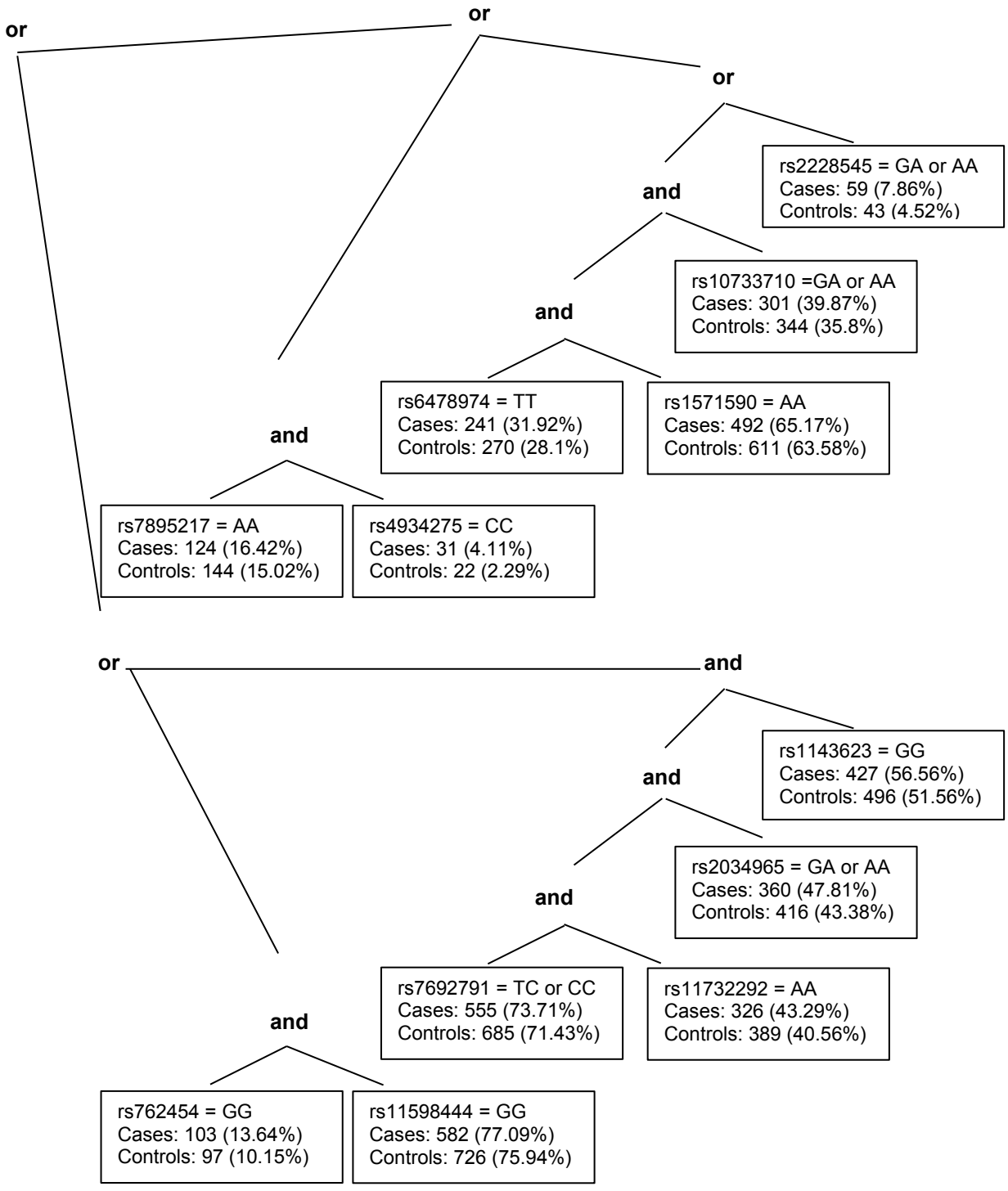

**Logic 4**

-0.555 * (not ***NFKB1***_L1)

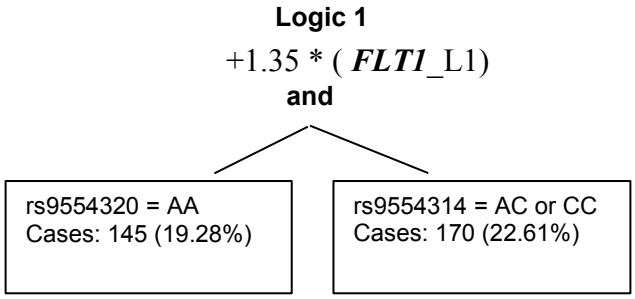


**Logic 5**

1.03 * (((***GDF10***_L1) or (***KDR***_L1 and ***IL1B***_L1)) or ((***TGFBR***_L1 or ***BMPR2***_L1) or ***BMPR1A***_L1))




**Supplementary Figure 4: Gene-Pathway Tree in association with Rectal Cancer Risk.**




**Supplementary Figure 5: Gene-Pathway Tree in association with Rectal Cancer Survival.**

**Supplementary Table 1: Rectal cancer gene-specific trees and rectal cancer risk**

| Gene | N of Trees | N of Leaves | Model deviance | Logic Model |
|---|---|---|---|---|
| *VEGF* | 1 | 1 | 2331.035 | -0.13 -0.256 * (not rs2010963) |
| *FLT1* | 1 | 2 | 2330.264 | -0.218 -1.69 * (rs4771249 and rs7324547) |
| *KDR* | 1 | 3 | 2333.108 | -0.322 +0.538 * ((rs7692791 and (not rs11732292)) and rs2034965) |
| *HIF1* | 1 | 1 | 1711.391 | -0.248 +0.335 * rs1951795 |
| *PDGFB* | 1 | 1 | 2338.464 | -0.198 -0.194 * rs4821877 |
| *TEK* | 1 | 1 | 2437.349 | -0.294 +0.177 * rs603085 |
| *TGFB* | 1 | 2 | 2404.302 | -0.241 +1.03 * (rs1800469 and rs4803455) |
| *TGFBR* | 1 | 3 | 2345.621 | -0.291 +0.452 * (((not rs6478974) and (not rs1571590)) and rs10733710) |
| *IGF1R* | 1 | 1 | 2476.206 | 0.182 -0.432 * (not rs2139924) |
| *NFKB1* | 1 | 1 | 2341.393 | -0.283 +0.385 * rs1609798 |
| *IL8* | 1 | 2 | 2352.165 | -0.255 +0.842 * ((not rs2227543) and rs4073) |
| *IL8RA* | 1 | 1 | 2327.882 | -0.315 +0.33 * rs1008562 |
| *IL8RB* | 1 | 1 | 2343.786 | 0.00499 -0.327 * (not rs1126579) |
| *IL1A* | 2 | 2 | 2352.195 | -0.114 -0.218 * rs3783546 -0.123 * (not rs2856838) |
| *IL1B* | 1 | 1 | 2350.997 | -0.351 +0.201 * (not rs1143623) |
| *TNF* | 1 | 1 | 2349.938 | -0.295 +0.184 * rs1800630 |
| *MMPS* | 1 | 1 | 2442.518 | -0.208 -0.236 * rs470215 |
| *BMP1* | 1 | 1 | 2328.169 | -0.244 +0.685 * rs3924229 |
| *BMP2* | 1 | 2 | 2347.41 | -0.325 +0.307 * ((not rs235770) and (not rs7270163)) |
| *BMP4* | 1 | 1 | 2350.241 | -0.383 +0.178 * (not rs2761887) |
| *BMPR1A* | 1 | 2 | 2346.542 | -0.259 +0.648 * (rs7895217 and rs4934275) |
| *BMPR1B* | 2 | 3 | 2331.615 | 0.14 -0.513 * rs1863652 -0.488 * (rs7694043 or rs3796442) |
| *BMPR2* | 1 | 1 | 2327.704 | -0.272 +0.588 * rs2228545 |
| *GDF10* | 1 | 2 | 2335.894 | -0.273 +0.833 * (rs762454 and (not rs11598444)) |
| *TLR2* | 1 | 2 | 2348.24 | -0.253 +1.43 * (rs1898830 and rs7656411) |
| *TLR3* | 1 | 4 | 2342.684 | -0.0836 -0.372 * ((rs3775291 or (rs11721827 or rs3775292)) and (not rs3775292)) |

| | | | | |
|---|---|---|---|---|
| *TLR4* | 4 | 6 | 2335.05 | 17 -0.553 * rs11536889 -17.9 * (rs1927911 or rs11536898) -17.3 * (not rs1927911) +0.746 * (rs1927911 or rs11536889) |
| *EGR2* | 1 | 1 | 2355.287 | -0.265 +0.183 * rs2295814 |
| *EGFR* | 1 | 2 | 2011.035 | -0.946 +0.717 * ((not rs17518446) or rs4947971) |
| *IRS1* | 1 | 1 | 2425.874 | -0.14 -0.11 * (not IRS1) |
| *VDR* | 2 | 2 | 2291.277 | -0.319 +0.289 * ((not VDRBsm1) and (not VDRFok1)) |

**Supplementary Table 2: Rectal cancer gene-specific trees and rectal cancer survival**

| Gene | N of Trees | N of Leaves | Model deviance | Logic Model |
|---|---|---|---|---|
| *VEGF* | 1 | 2 | 251.874 | -1.51 +1.54 * ((not rs3025040) or (not rs3025035)) |
| *FLT1* | 1 | 2 | 250.173 | 0.0637 -1.23 * (rs9554320 and rs9554314) |
| *KDR* | 5 | 8 | 238.811 | 0.284 -1.11 * rs2305949 -0.534 * rs2305948 +0.447 * (rs11732292 and (not rs12498529)) -0.45 * (not rs6838752) +0.995 * ((not rs2071559) and ((not rs2125489) and rs2305949)) |
| *HIF1* | 1 | 1 | 181.896 | -0.0509 +0.281 * rs1951795 |
| *PDGFB* | 1 | 1 | 256.592 | 0.0269 -0.462 * rs5750781 |
| *TEK* | 1 | 1 | 264.549 | -0.392 +0.401 * (not rs603085) |
| *TGFB* | 1 | 1 | 261.178 | -0.073 +0.259 * (not rs4803455) |
| *TGFBR* | 1 | 1 | 257.822 | 0.0687 -0.203 * rs1571590 |
| *IGF1R* | 1 | 1 | 269.653 | -0.0241 +0.414 * rs2139924 |
| *NFKB1* | 2 | 2 | 248.503 | 0.0339 -2.74 * (not rs11722146) +2.7 * (not rs1609798) |
| *IL8* | 1 | 2 | 254.513 | 0.0414 -1.04 * ((not rs2227543) and rs2227307) |
| *IL8RA* | 1 | 1 | 255.545 | 0.21 -0.299 * (not rs1008562) |
| *IL8RB* | 1 | 1 | 255.825 | 0.245 -0.343 * (not rs1126579) |
| *IL1A* | 2 | 2 | 255.057 | 0.138 -0.425 * (not rs3783546) +0.285 * (not rs2856838) |
| *IL1B* | 1 | 1 | 257.29 | -0.099 +0.232 * (not rs1143633) |
| *TNF* | 1 | 3 | 257.479 | -0.298 +0.339 * ((rs1800630 or (not rs1799964)) and (not rs1799964)) |
| *MMPS* | 2 | 3 | 261.841 | -0.353 -1.89 * (rs470215 and rs1996352) +0.424 * (not rs3025066) |
| *BMP1* | 1 | 1 | 254.868 | 0.183 -0.273 * (not rs3924231) |
| *BMP2* | 3 | 5 | 250.705 | 0.287 -0.25 * (not rs235770) +0.478 * rs235770 -0.767 * (((not s7270163) and (not rs3178250)) and rs235770) |
| *BMP4* | 1 | 1 | 257.334 | -0.0355 +0.178 * rs2761887 |
| *BMPR1A* | 2 | 3 | 253.986 | 1.2 -1.23 * (not rs6586034) -1.21 * (rs7895217 or rs7088641) |
| *BMPR1B* | 1 | 1 | 254.513 | -0.119 +0.281 * rs13134042 |
| *BMPR2* | 1 | 1 | 255.746 | 0.0839 -0.201 * rs13430786 |
| *GDF10* | 1 | 1 | 253.259 | -1.1 +1.14 * (not rs2853838) |
| *TLR2* | 2 | 2 | 256.084 | -0.366 +0.528 * (not rs7656411) -0.219 * rs1898830 |

| | | | | |
|---|---|---|---|---|
| *TLR3* | 1 | 1 | 256.736 | -0.223 +0.304 * (not rs11721827) |
| *TLR4* | 1 | 1 | 257.131 | 0.0182 -0.951 * rs11536889 |
| *EGR2* | 1 | 1 | 256.718 | -0.112 +0.268 * (not rs224082) |
| *EGFR* | 1 | 1 | 222.601 | -0.0397 +0.54 * rs17151957 |
| *IRS1* | 1 | 1 | 266.786 | 0.0151 -0.123 * IRS1 |
| *VDR* | 1 | 1 | 247.286 | 0.0292 -0.233 * (not VDRFok1) |

# CHAPTER 5

# DISCUSSION

Chronic diseases are multifactorial by nature and their complex etiology involves interplay between multiple genetic factors and with environmental factors (Dempfle et al. 2008). A simplified approach that focuses only on effects of individual markers (e.g., single nucleotide polymorphisms (SNPs)) is ignoring this inherent nature of disease, hence explaining only a small portion of its heritability while a significant part remains unaccounted for, referred to as 'missing heritability'(Manolio et al. 2009). Several possible explanations for the missing heritability include overestimation of the heritability component of complex traits, underestimation of the risk associated with currently identified alleles, or yet-to-be identified common and/or rare alleles. Another probable reason is the existence of unidentified gene-gene and gene-environment interactions (GEIs) (Culverhouse et al. 2002).

In this dissertation we presented a novel methodological strategy to examine interactions between genetic variants and between genetic and environmental factors at the gene and pathway levels. We introduced our approach by applying it to genome-wide association (GWAS) data of six common chronic diseases and searched for biologically plausible forms of SNP-Set interactions within genes. We then extended our approach to test for GEIs at the gene-pathway level and applied it to case-control data of colon and rectal cancer focusing on the candidate angiogenesis pathway and the hypothesized environmental exposures: dietary protein intake, smoking, and alcohol consumption. Our framework consisted of 3-steps: the first two summarized the gene effects within genes and across the full pathway and the third step modelled the GEI effects on colon and rectal cancer risk and survival.

## 5.1 Within-gene SNP-set interactions in GWAS

In the first study of this dissertation we showed how exploring interactions of all measured SNPs within each gene can identify appreciable numbers of novel susceptibility loci in GWAS. We re-analyzed six diseases of The Wellcome-Trust-Case-Control-Consortium (WTCCC) data (Wellcome Trust Case Control Consortium2007): bipolar disorder (BD), coronary artery disease (CAD), hypertension (HT), rheumatoid arthritis (RA), type 2 diabetes (T2D), and type 1 diabetes (T1D). We considered two biologically plausible forms of SNP-set interactions: SNP *intersection* and SNP *union*. A SNP-set included all measured SNPs on each individual gene. SNP-set interactions within each gene were searched for using logic regression. The number of genes that showed strong evidence of association was: 13 for BD, 16 for CAD, 15 for HT, 72 for RA, 105 for T1D and 19 for T2D. In addition, strong evidence emerged implicating a large number of new discoveries supported by apparent biologically plausible links to disease. Top significant genes were: *NFIA* with BD, *CDKN2B* with CAD, *COL4A4* with HT, *BTNL2* with RA, and *TCF7L2* with T2D.

### 5.1.1 Examining epistatic interactions using logic regression in GWAS

Several methodological approaches have been developed to search for interactions on a genome-wide scale including exhaustive searches of two-locus (*pairwise*) interactions (Marchini et al. 2005), two-stage methods that involve screening for marginal effects and selecting a subset of loci that pass some single-locus significance threshold which are then carried to the second stage of an exhaustive search of pairwise interactions (Wu et al. 2010,Tao et al. 2012), and Bayesian model selection (Zhang et al. 2007). Exhaustive searches for higher order interactions in a genome-wide setting means that the number of tests, the amount of time, and the computational

load increases exponentially with an increase of the order of interaction considered. These methods cannot effectively handle the high dimensionality of GWAS data, require significant marginal effects of individual markers, and/or rely on searches for *pairwise* interactions rather than *higher-order* interactions. A search for higher order interactions, however, is more likely to elucidate the underlying biological mechanism of disease.

Logic regression searches for models with binary predictors that are combined by Boolean combinations (Ruczinski et al. 2003). We used logic regression to identify epistatic interactions of biologically plausible forms of SNP interactions within each gene (*SNP intersection* which combines SNPs by "and", e.g., SNP 1 *and* SNP 2 *and* SNP 3, *SNP union* which combines SNPs by "or", e.g., SNP 1 *or* SNP 2: we can combine the two forms in one, e.g., (SNP 1 *or* SNP 2) *and* SNP 3). We demonstrated how it can be applied to GWAS data to identify novel susceptibility loci for six diseases in the WTCCC data through identification of higher order SNP-set interactions within genes. These discoveries illustrate the additional power of GWAS which has not been revealed previously by the standard single-SNP analysis.

Although replication of findings in independent samples has become the standard for assessing GWAS statistical results, this requirement may actually lead to missing real genetic effects (Greene et al. 2009). Comparing our findings to the single-SNP WTCCC analysis (Wellcome Trust Case Control Consortium2007), our interaction analysis was able to detect the majority of the previously reported strong signals. We also compared our findings to recently published GWAS meta-analyses of the six diseases detecting on average 46% of the reported loci. It is important to note that although it is impossible to validate discoveries made by logic regression analysis with single-SNP analyses, detecting the majority of previous single-SNP associations in our SNP-set interaction-based analysis is corroborating and providing support to our findings.

Proper validation of our novel signals would require further investigation in larger datasets and using a gene-level interaction-based analysis.

### 5.1.2 Assessment of strength of evidence

Exhaustive searches for interactions in GWAS raise the issue of multiple testing similar to the single-locus analysis of a GWAS. A Bonferroni correction is suggested appropriate when all tests are independent (Marchini et al. 2005) or using permutation to assess significance of testing for association while allowing for interactions and accounting for correlation of tests (Chapman et al. 2007). The latter may be computationally prohibitive for large numbers of GWAS loci. We followed the WTCCC's framework of using the Bayes Factor (BF) as the measure of evidence of association of each gene and disease risk for all six diseases. The BF was standardized taking the median fit score of 20 permuted datasets. Each gene's strength of evidence was assessed by exceeding the calculated disease BF threshold which is based on the strength of evidence in the Bayesian philosophy. We also assessed statistical significance of association using a p-value cut-off of $3.82 \times 10^{-6}$ (corresponding to a p-value of 0.05 with a Bonferroni correction for multiple testing of approximately 13,000 genes per disease). Both statistical frameworks agreed on how genes were ranked based on statistical significance, however, more signals were considered significant by BF and did not reach significance using p-value. We do not believe that lowering the significance threshold for a single-SNP analysis would yield as many signals as identified using BF. We plotted Bonferroni corrected p-values from the single-SNP analysis and our SNP-set interaction analysis that showed greater numbers of significant signals by our SNP-interaction approach, under the same criterion of Bonferroni correction applied to both methods. We also repeated our analysis on 10 permuted datasets using RA as an example. Out of the 10 permuted datasets each involving 13,083 genes, only one gene of one of the 10 datasets was statistically

significant using the Bonferroni corrected p-value threshold. Thus, the huge search space for the logic regression would not explain our findings.

### 5.1.3 Significance and interpretation of our findings

In addition to statistical support of our top findings, they were also supported by apparent biological links to disease. For example, our results provided confirmation of previous linkage analyses implicating specific chromosomal regions and diseases such as chromosome 1p31 near *NFIA* gene in association with BD and chromosome 2q35 near *COL4A4* gene in association with HT, the latter never been reported by GWAS before. We also confirmed other GWAS discoveries such as *BTNL2* with RA recently corroborated through exome sequencing (Mitsunaga et al. 2013) and found an association with sarcoidosis, another auto-immune disease (Morais et al. 2012). Another association was *TCF7L2* with T2D that has been repeatedly replicated across different populations (Grant et al. 2006) with evidence suggestive of a reduction of a *TCF7L2* related T2D risk in response to lifestyle changes (Florez et al. 2006). Such previously reported associations confirmed in our interaction analysis are worthy of further in depth investigation.

The magnitude of the odds ratios reported from our SNP-set interaction analysis of GWAS was substantially larger compared to those typically reported from single-SNP based GWAS analysis ranging between: 1.1 and 1.5. Besides adding strength to the associations, this made them more readily interpretable. We also identified a larger number of genes that may help determine the genetic risk of the six diseases and open avenues for refined risk identification and risk prediction. This is of specific value for diseases with a less clearly identified genetic risk such as BD, CAD and HT. For example, the single-SNP based analysis reported from WTCCC failed to

identify any strong association signals for HT, while our interaction analysis was able to detect both novel and previously reported association signals providing new insights into the genetic profile of a complex disease such as HT.

The novel discoveries and previously reported associations detected from our interaction-based analysis in GWAS adds strength to our approach of analysis and emphasizes the importance of searching for SNP-set interaction effects, in addition to the standard single-SNP analysis in GWAS.

## 5.2 Pathway gene-environment interaction in candidate gene studies

In the second half of this dissertation we went steps further with our approach to ultimately examine candidate pathway GEIs on colorectal cancer risk and survival. The study of GEI on disease outcomes has several motivations and advantages: (1) to better characterize gene and environmental exposure effects; (2) to increase the power to detect genes with small marginal effects especially if the effect is relevant to a subgroup defined by a certain environmental exposure; (3) similarly, it strengthens the association of disease with environmental exposures by examining its effects in genetically susceptible groups; (4) defining environmental exposures and focusing on specific lifestyle components helps to determine which element of complex exposures (e.g., diet) are important; (5) to gain insights into disease mechanisms by focusing on relevant biological pathways; and (6) the clinical relevance and public health impact is emphasized through its use in new preventive and therapeutic strategies including personalized approaches.

### 5.2.1 Methods to examine gene-environment interactions

Some approaches proposed for examining gene-gene interactions (some of which described above) could be extended to GEIs (Dempfle et al. 2008)including single-stage approaches using ordinary or penalized regression frameworks, or a two-step strategy that involves a screening test followed by a traditional case-control test of GEI (Murcray et al. 2009). The case-only design has been proposed as more powerful alternative to the standard case-control test (Clayton et al. 2001) based on the assumption that the genotype and environmental exposure are independent in the population under study. Violation of this assumption, however, can yield a severely inflated type I error (Piegorsch et al. 1994).

A recent application of the traditional case-control test, case-only test and the 2-step method proposed by Murcray and colleagues on colorectal cancer GWAS data did not identify any genome-wide significant GEIs (Figueiredo et al. 2011). These authors also used a candidate approach to analyze previously reported colorectal cancer GWAS susceptibility loci and 14 environmental exposures known to be involved in colorectal cancer etiology, they identified seven nominally significant GEIs one of which was between alcohol and a SNP on *CHD1* gene (chromosome 16q22.1).

Examining GEI at the gene-pathway level could depend on purely data-driven approaches from genome-wide scans to elicit important pathways, or through a focus on several candidate genes in a single pathway perceived as more critical. Some methods for pathway analyses have been proposed for both approaches (Thomas 2010b). Examples include data mining approaches (Kraft et al. 2009), and modifications to pathway approaches originally developed for gene expression data such as gene set enrichment analysis (Wang et al. 2007). Nevertheless, hypothesis-driven

pathway-based approaches that require prior knowledge of the underlying etiology can elucidate the underlying biological mechanisms (Thomas 2010b). Evidence on relevant pathways may involve biologic plausibility, as well as evidence available from experimental studies and prior epidemiologic reports (Jorgensen et al. 2009). Experimental studies in model organisms have, indeed, provided several evidences of interactions between genes and exposures which help suggest candidate gene-environment interactions to be examined in epidemiologic studies (Aschard et al. 2012).

### 5.2.2 A 3-step analytic framework to identify candidate pathway gene-environment interactions

We developed a novel candidate-pathway framework to assess GEIs and illustrated its use for colon and rectal cancer risk and survival. We focused on the angiogenesis pathway, and three angiogenesis-related lifestyle risk factors: dietary protein, smoking, and alcohol consumption. Our framework emphasized the biologic hypothesis throughout the process starting from the selection of the candidate genes and the specific lifestyle exposures, and carried the logic to the three steps of the analysis: a component that has been lacking in the study of candidate pathway analysis (Thomas et al. 2009). Building on our approach to examine within-gene SNP-set interactions using logic regression, the first step of our analysis framework produced gene-specific trees (GSTs). They formed the building blocks for the subsequent two steps of gene-gene and gene-environment analysis. The second step provided a summary of the full pathway's genetic effects. The third step modelled the pathway GEIs. We used standard logistic regression and Cox proportional hazards modelling that included the main effects of gene and environment variables, relevant interaction terms and adjustment variables, and built the models using stepwise backward elimination. This standard approach to testing an interaction is arguably the

most natural way given hypothesized genetic and environmental factors influencing disease outcomes (Cordell 2009a). We believe our approach to use the GSTs as a summary of each gene's SNP profile rather than individual SNPs was a step ahead of a typical interaction analysis that considers pairwise interactions between SNPs for gene-gene interactions or interactions between an individual SNP and an environmental exposure for GEI testing. We analyzed data of colon cancer and rectal cancer cases and controls from the Diet, Activity and Lifestyle as a Risk Factor for Colon Cancer Study conducted in the United States (Slattery et al. 1997a,Slattery et al. 2003). We selected a total of 257 SNPs in 34 genes of the angiogenesis candidate gene-pathway based on standard pathway maps, experimental and epidemiological evidence. We found five statistically significant GEIs associated with colon cancer risk and three GEIs with colon cancer survival involving all there environmental exposures. For rectal cancer, we found eight significant GEIs in association with risk between six genes and five GEIs with survival.

### 5.2.3 Candidate gene and GWAS approaches to examine gene-environment interactions

Most GWASs have not investigated GEI, primarily due to lack of data on environmental exposures (Stranger et al. 2011). Study consortia, although they carry the advantage of increased sample size, may face some challenges due to differences in exposure measurement protocols across studies, differences in the scale of reported gene-environment interaction effects, and differences in the distribution of exposures across studies (Aschard et al. 2012) all of which would require the investigation of between-study heterogeneity (Thompson et al. 2011). On the other hand, a candidate pathway study based on informed candidate gene selection with detailed information on environmental exposures may be more suited to examining GEI effects compared to GWAS loci which are harder to identify, have smaller effect sizes, and are unlikely to be the functional variants themselves (Stranger et al. 2011). In application of our candidate pathway

approach we selected the candidate genes and exposures based on biologic relevance; we included in the analysis all hypothesized genes in the pathway as opposed to focusing only on markers with significant marginal effects; the gene-level summaries are potentially capturing the full gene effect; and a multiple testing adjustment for testing many non-hypothesized associations in GWAS was not required for our candidate gene-pathway analyses because the associations were biologically hypothesized a priori (Tomlinson et al. 2011).

### 5.2.4 Gene-environment interaction effects on colon and rectal cancer risk and survival

Colon and rectal cancers may share genetic and environmental risk factors, yet there is evidence of differences in the characteristics of each cancer population (Potter1999a,Wei et al. 2004,Annema et al. 2011,Shin et al. 2011), suggesting different mechanisms could be influencing the development of each type of cancer (Robsahm et al. 2013). We have previously shown that colon and rectal tumors differed in somatic mutation frequencies (microsatellite instability, CpG island methylator phenotype, and Ki-ras mutations were more frequent in proximal colon tumors, and p53 mutation more in distal colon and rectal tumors) (Slattery et al. 2009). Thus, we analyzed the colon and rectal cancer data separately. Indeed our results show differences between colon and rectal cancer risk and survival GEI profiles. Although separating colon and rectal cancer analyses was justified, there was initially no specific scientific motivation to combining risk and survival reporting. Our results, however, show there were some similarities detected between the risk and survival profiles within each cancer (for example: GEIs with *BMP* genes on colon cancer risk and survival and *TLR* genes and *EGR2* gene on rectal cancer risk and survival).

One interesting finding was the predominance of interactions of four genes (*KDR*, *TLR2*, *EGR2*, and *EGFR*) with animal dietary protein out of a total of five significant interactions on rectal cancer survival. A potential role for high intake of red and processed meat on colorectal cancer risk has been supported by many, yet not all studies (Alexander et al. 2010,Alexander et al. 2011). Reported associations of meat intake have been generally stronger for distal colon and rectal cancer. A plausible explanation relates the increased rate of protein fermentation by large intestine bacteria mainly occurring in the distal parts of the colon and rectum (Silvester et al. 1995,Chao et al. 2005) and tumor promotion (Kim et al. 2013). It is possible that a characterization of GEI with a high animal/vegetable protein intake ratio that we observed provides some interpretation to the high red and processed meat associations with rectal cancer.

Overall, we observed GEIs of genes among major drivers of angiogenesis and angiogenesis-related inflammatory genes and with all three environmental exposures. Some of the interactions were with genes known to be strongly associated with colon and rectal cancer. More GEIs among those genes with all three environmental exposures, however, were associated with colon cancer risk compared to rectal cancer risk. For example, GEIs of *FLT1* gene (*VEGF* receptor 1) with smoking and animal protein intake, and *KDR* gene (*VEGF* receptor 2) with alcohol consumption in association with colon cancer risk. For rectal cancer risk, GEIs of *PDGFB* with animal protein and *IGF1R* with alcohol consumption were detected.

Other interactions were with genes that have become recently of interest in association with colon and rectal cancer such as BMP genes (Beck et al. 2006,Nishanian et al. 2004) and TLR genes (Nihon-Yanagi et al. 2012,Tchorzewski et al. 2014). We identified GEIs with BMP genes in association with both colon cancer risk and survival. We observed GEIs between BMP4 gene and smoking on colon cancer risk; and BMP1 gene and smoking and BMPR2 gene and alcohol

on colon cancer survival. On rectal cancer, we observed interactions of TLR4 with smoking and alcohol on rectal cancer risk and TLR2 with animal protein. Our results help define the full risk profile associated with these genes recently implicated in colorectal cancer carcinogenesis. Some of the GEIs we observed with smoking for both colon and rectal cancers displayed dose response associations as shown by increasing magnitudes of gene ORs with increasing levels of smoking. This observed positive gradient adds to the plausibility of the interactions.

### 5.3 Strengths and Limitations

The discoveries detected from our interaction-based analysis in GWAS add strength to our approach of analysis and emphasize the importance of searching for SNP-set interaction effects, in addition to the standard single-SNP analysis in GWAS. We used logic regression to search for interactions between SNPs within genes. The logic regression method, despite its utility, is not immune to limitations. To manage the large computational demand of the logic regression search in the WTCCC GWAS analysis, we had to limit the search of SNP-set interactions to a single gene at a time and fix the size of SNP interactions searched for within each gene (up to 2 trees and 5 leaves). It is possible that more complex SNP interactions exist but were not considered in our analysis. Our assessment of SNP-set interactions, however, has a much higher power of signal discovery compared to a pairwise SNP-SNP interaction since significantly fewer tests are performed. Despite the limited form of logic regression that we applied in our analysis, searching for specific forms of SNP-set interactions is a step towards addressing the complexity of genetic associations in a GWAS compared to a marginal assessment of individual SNP effects on disease. In this analysis, we analyzed SNPs within genes and did not consider gene-gene interactions and a more comprehensive approach would require consideration of gene-gene interactions.

The main strength of our candidate pathway GEI approach is integration of the relevant biologic information in the construction of the pathway (applied to both the pathway and environmental exposures) and throughout the analysis process. In the application of the logic regression in the first step of the colon and rectal cancer candidate pathway GEI analysis, we also searched for SNP-set interactions within each gene. Rather than fixing the size of the logic regression models, we used cross-validation to determine the optimal model sizes albeit specifying a maximum desired size (up to nine trees and 20 leaves). As such summary of the gene effects was limited by the reached model size in addition to the number of tagSNPs on each gene. We selected the genes of the angiogenesis pathway using a candidate approach and it is possible that some genes relevant to the angiogenesis process have been omitted inadvertently in developing the working pathway figure. The candidate pathway, however, involved a relatively large number of genes implicated in colorectal carcinogenesis. We also note that we focused on only one gene-pathway and three lifestyle risk factors, yet identified an appreciable number of novel strong interactions on both colon and rectal cancer risk and survival. Other limitations of our study were related to the design of the case-control study which suffers from inherent forms of bias such as recall bias. In this study this was minimized by: using a rapid-reporting system to identify cases; conducting the majority of interviews within four months of diagnosis; and focusing the referent period of the study questionnaires to two to three years prior to diagnosis. With regards to our environmental exposures of interest, we obtained long-term alcohol consumption and cigarette smoking history, and extensive diet history to capture more detail compared to self-administered questionnaires.

## 5.4 Conclusions and Public Health Implications

It is important to note that there is no consensus yet about the best statistical method to model gene-gene interactions or GEIs and more research and applications to real data are required (Figueiredo et al. 2011). I attempted in this dissertation to provide a methodologically sound and biologically plausible approach to examining interactions with an emphasis on the biological hypothesis, whether in the form that the SNP-set interactions might take or the choice of candidate genes and environmental exposures, can yield novel and biologically plausible results. In addition to statistical significance, the findings were corroborated by previous evidence and biologic links to the diseases studied.

The analysis of the WTCCC GWAS data emphasized the importance of adopting a method that can handle higher order SNP-set interactions and demonstrated its ability in discovering novel disease susceptibility loci in addition to confirming findings from standard single-SNP analysis in GWAS. This analysis formed the base for the candidate-pathway GEI framework that also yielded novel GEIs in association with colon and rectal cancer risk and survival. Knowledge generated from this research can be directly translated into practical clinical and public health applications. Susceptibility loci provide potential leads towards identifying drug targets, thereby helping to reduce the socio-economic burden of disease. Furthermore, identification of GEI loci of common genetic markers and theoretically modifiable lifestyle factors provides new insights into screening and preventive strategies and opens avenues for personalized strategies.

## 5.5 Future Directions

In the candidate approach to identifying pathway GEIs, I focused on the angiogenesis pathway which is one of the hallmarks of cancer making the findings potentially informative to other solid tumors. One future direction would be extending the framework to other solid tumors.

Personalized medicine and lifestyle recommendations based on the individual genetic profile are being promoted as the future of clinical and public health. One important application would be expanding this research to other colorectal-cancer-relevant genetic pathways and associated lifestyle risk factors to enrich and complete the findings. Primary and secondary colorectal prevention strategies could be developed and tested based on the findings. Specifically, innovative technologies for colorectal cancer primary prevention and screening would include: (1) offering genome sequence testing and profiling individual risk based on the gene-lifestyle interactions we and others identify; (2) monitoring lifestyle habits regularly. The idea would be to develop a risk profile for the individual based on their genetic-predisposition profiles and tailor specific interventions for their relevant lifestyle habits. Applications to provincial and national grant competitions for funding of future projects are planned.

**REFRENCES**


Aberg, K., Dai, F., Viali, S., Tuitele, J., Sun, G., Indugula, S. R., Deka, R., Weeks, D. E., and McGarvey, S. T. 2009. Suggestive linkage detected for blood pressure related traits on 2q and 22q in the population on the Samoan islands. BMC Med. Genet. 10: 107-2350-10-107.

Affara, M., Dunmore, B. J., Sanders, D. A., Johnson, N., Print, C. G., and Charnock-Jones, D. S. 2011. MMP1 bimodal expression and differential response to inflammatory mediators is linked to promoter polymorphisms. BMC Genomics 12: 43-2164-12-43.

Alexander, D. D. and Cushing, C. A. 2011. Red meat and colorectal cancer: a critical summary of prospective epidemiologic studies. Obes. Rev. 12: e472-93.

Alexander, D. D., Miller, A. J., Cushing, C. A., and Lowe, K. A. 2010. Processed meat and colorectal cancer: a quantitative review of prospective epidemiologic studies. Eur. J. Cancer Prev. 19: 328-341.

Anderson, I. M., Haddad, P. M., and Scott, J. 2012. Bipolar disorder BMJ 345: e8508.

Annema, N., Heyworth, J. S., McNaughton, S. A., Iacopetta, B., and Fritschi, L. 2011. Fruit and vegetable consumption and the risk of proximal colon, distal colon, and rectal cancers in a case-control study in Western Australia. J. Am. Diet. Assoc. 111: 1479-1490.

Anto, R. J., Mukhopadhyay, A., Shishodia, S., Gairola, C. G., and Aggarwal, B. B. 2002. Cigarette smoke condensate activates nuclear transcription factor-kappaB through phosphorylation and degradation of IkappaB(alpha): correlation with induction of cyclooxygenase-2 Carcinogenesis 23: 1511-1518.

Aschard, H., Lutz, S., Maus, B., Duell, E. J., Fingerlin, T. E., Chatterjee, N., Kraft, P., and Van Steen, K. 2012. Challenges and opportunities in genome-wide environmental interaction (GWEI) studies. Hum. Genet. 131: 1591-1613.

Barrett, J. C., Clayton, D. G., Concannon, P., Akolkar, B., Cooper, J. D., Erlich, H. A., Julier, C., Morahan, G., Nerup, J., Nierras, C., et al. 2009. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat. Genet. 41: 703-707.

Beck, S. E., Jung, B. H., Fiorino, A., Gomez, J., Rosario, E. D., Cabrera, B. L., Huang, S. C., Chow, J. Y., and Carethers, J. M. 2006. Bone morphogenetic protein signaling and growth suppression in colon cancer. Am. J. Physiol. Gastrointest. Liver Physiol. 291: G135-45.

Brinckerhoff, C. E., Rutter, J. L., and Benbow, U. 2000. Interstitial collagenases as markers of tumor progression. Clin. Cancer Res. 6: 4823-4830.

Campa, D., Kaaks, R., Le Marchand, L., Haiman, C. A., Travis, R. C., Berg, C. D., Buring, J. E., Chanock, S. J., Diver, W. R., Dostal, L., et al. 2011. Interactions Between Genetic Variants and

Breast Cancer Risk Factors in the Breast and Prostate Cancer Cohort Consortium. Journal of the National Cancer Institute 103: 1252-1263.

Chan, D. S., Lau, R., Aune, D., Vieira, R., Greenwood, D. C., Kampman, E., and Norat, T. 2011. Red and processed meat and colorectal cancer incidence: meta-analysis of prospective studies. PLoS One 6: e20456.

Chao, A., Thun, M. J., Connell, C. J., McCullough, M. L., Jacobs, E. J., Flanders, W. D., Rodriguez, C., Sinha, R., and Calle, E. E. 2005. Meat consumption and risk of colorectal cancer. JAMA 293: 172-182.

Chapman, J. and Clayton, D. 2007. Detecting association using epistatic information. Genet. Epidemiol. 31: 894-909.

Chen, D., Zhao, M., and Mundy, G. R. 2004. Bone morphogenetic proteins. Growth Factors 22: 233-241.

Chen, Y. M. and Miner, J. H. 2012. Glomerular basement membrane and related glomerular disease. Transl. Res. 160: 291-297.

Cheng, J., Chen, Y., Wang, X., Wang, J., Yan, Z., Gong, G., Li, G., and Li, C. 2014. Meta-analysis of prospective cohort studies of cigarette smoking and the incidence of colon and rectal cancers Eur. J. Cancer Prev. .

Cho, Y. S., Chen, C. H., Hu, C., Long, J., Ong, R. T., Sim, X., Takeuchi, F., Wu, Y., Go, M. J., Yamauchi, T., et al. 2012. Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. Nat. Genet. 44: 67-72.

Christensen, K. and Murray, J. C. 2007. What genome-wide association studies can do for medicine N. Engl. J. Med. 356: 1094-1097.

Clayton, D. and McKeigue, P. M. 2001. Epidemiological methods for studying genes and environmental factors in complex diseases. Lancet 358: 1356-1360.

Cleary, S. P., Cotterchio, M., Shi, E., Gallinger, S., and Harper, P. 2010. Cigarette smoking, genetic variants in carcinogen-metabolizing enzymes, and colorectal cancer risk Am. J. Epidemiol. 172: 1000-1014.

Cordell, H. J. 2009a. Detecting gene-gene interactions that underlie human diseases Nat. Rev. Genet. 10: 392-404.

Cordell, H. J. 2009b. Detecting gene-gene interactions that underlie human diseases. Nat. Rev. Genet. 10: 392-404.

Cordell, H. J. 2002. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans Hum. Mol. Genet. 11: 2463-2468.

Cordell, H. J. and Clayton, D. G. 2005. Genetic association studies. The Lancet 366: 1121-1131.

Crea, F. and Liuzzo, G. 2013. Pathogenesis of acute coronary syndromes. J. Am. Coll. Cardiol. 61: 1-11.

Cui, J., Taylor, K. E., Destefano, A. L., Criswell, L. A., Izmailova, E. S., Parker, A., Roubenoff, R., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., et al. 2009. Genome-wide association study of determinants of anti-cyclic citrullinated peptide antibody titer in adults with rheumatoid arthritis. Molecular Medicine 15: 136-143.

Culverhouse, R., Suarez, B. K., Lin, J., and Reich, T. 2002. A perspective on epistasis: limits of models displaying no main effect Am. J. Hum. Genet. 70: 461-471.

Dai, J., Gu, J., Huang, M., Eng, C., Kopetz, E. S., Ellis, L. M., Hawk, E., and Wu, X. 2012. GWAS-identified colorectal cancer susceptibility loci associated with clinical outcomes Carcinogenesis 33: 1327-1331.

De Stefani, E., Ronco, A. L., Boffetta, P., Deneo-Pellegrini, H., Correa, P., Acosta, G., and Mendilaharsu, M. 2012. Nutrient-derived dietary patterns and risk of colorectal cancer: a factor analysis in Uruguay. Asian Pac. J. Cancer. Prev. 13: 231-235.

Dempfle, A., Scherag, A., Hein, R., Beckmann, L., Chang-Claude, J., and Schafer, H. 2008. Gene-environment interactions for complex traits: definitions, methodological requirements and challenges. Eur. J. Hum. Genet. 16: 1164-1172.

Deneen, B., Ho, R., Lukaszewicz, A., Hochstim, C. J., Gronostajski, R. M., and Anderson, D. J. 2006. The transcription factor NFIA controls the onset of gliogenesis in the developing spinal cord Neuron 52: 953-968.

Dennis, B., Ernst, N., Hjortland, M., Tillotson, J., and Grambsch, V. 1980. The NHLBI nutrition data system J. Am. Diet. Assoc. 77: 641-647.

Ding, W., Zhou, D. L., Jiang, X., and Lu, L. S. 2013. Methionine synthase A2756G polymorphism and risk of colorectal adenoma and cancer: evidence based on 27 studies. PLoS One 8: e60508.

Dinu, I., Mahasirimongkol, S., Liu, Q., Yanai, H., Sharaf Eldin, N., Kreiter, E., Wu, X., Jabbari, S., Tokunaga, K., and Yasui, Y. 2012. SNP-SNP interactions discovered by logic regression explain Crohn's disease genetics. PLoS One 7: e43035.

Dor, Y., Porat, R., and Keshet, E. 2001. Vascular endothelial growth factor and vascular adjustments to perturbations in oxygen homeostasis Am. J. Physiol. Cell. Physiol. 280: C1367-74.

Dougherty, U., Cerasi, D., Taylor, I., Kocherginsky, M., Tekin, U., Badal, S., Aluri, L., Sehdev, A., Cerda, S., Mustafi, R., et al. 2009. Epidermal growth factor receptor is required for colonic

tumor promotion by dietary fat in the azoxymethane/dextran sulfate sodium model: roles of transforming growth factor-{alpha} and PTGS2. Clin. Cancer Res. 15: 6780-6789.

Dougherty, U., Mustafi, R., Wang, Y., Musch, M. W., Wang, C. Z., Konda, V. J., Kulkarni, A., Hart, J., Dawson, G., Kim, K. E., et al. 2011. American ginseng suppresses Western diet-promoted tumorigenesis in model of inflammation-associated colon cancer: role of EGFR. BMC Complementary & Alternative Medicine 11: 111.

Edwards, S., Slattery, M. L., Mori, M., Berry, T. D., Caan, B. J., Palmer, P., and Potter, J. D. 1994. Objective system for interviewer performance evaluation for use in epidemiologic studies Am. J. Epidemiol. 140: 1020-1028.

Erlich, H. A. 1991. HLA class II sequences and genetic susceptibility to insulin dependent diabetes mellitus Baillieres Clin. Endocrinol. Metab. 5: 395-411.

Escalante, A. 2013. Chapter 51 - Rheumatoid Arthritis. In Women and Health (Second Edition) (Anonymous ), pp. 771-784. Academic Press, .

Esposito, G., Capoccia, E., Turco, F., Palumbo, I., Lu, J., Steardo, A., Cuomo, R., Sarnelli, G., and Steardo, L. 2013. Palmitoylethanolamide improves colon inflammation through an enteric glia/toll like receptor 4-dependent PPAR-alpha activation. Gut .

Ferrara, N. 1999. Molecular and biological properties of vascular endothelial growth factor J. Mol. Med. 77: 527-543.

Ferrari, P., Jenab, M., Norat, T., Moskal, A., Slimani, N., Olsen, A., Tjonneland, A., Overvad, K., Jensen, M. K., Boutron-Ruault, M. C., et al. 2007. Lifetime and baseline alcohol intake and risk of colon and rectal cancers in the European prospective investigation into cancer and nutrition (EPIC) Int. J. Cancer 121: 2065-2072.

Figueiredo, J. C., Hsu, L., Hutter, C. M., Lin, Y., Campbell, P. T., Baron, J. A., Berndt, S. I., Jiao, S., Casey, G., Fortini, B., et al. 2014. Genome-wide diet-gene interaction analyses for risk of colorectal cancer. PLoS Genet. 10: e1004228.

Figueiredo, J. C., Lewinger, J. P., Song, C., Campbell, P. T., Conti, D. V., Edlund, C. K., Duggan, D. J., Rangrej, J., Lemire, M., Hudson, T., et al. 2011. Genotype-environment interactions in microsatellite stable/microsatellite instability-low colorectal cancer: results from a genome-wide association study Cancer Epidemiol. Biomarkers Prev. 20: 758-766.

Flood, D. M., Weiss, N. S., Cook, L. S., Emerson, J. C., Schwartz, S. M., and Potter, J. D. 2000. Colorectal cancer incidence in Asian migrants to the United States and their descendants. Cancer Causes Control 11: 403-411.

Florez, J. C. 2008. Newly identified loci highlight beta cell dysfunction as a key cause of type 2 diabetes: where are the insulin resistance genes? Diabetologia 51: 1100-1110.

Florez, J. C., Jablonski, K. A., Bayley, N., Pollin, T. I., de Bakker, P. I., Shuldiner, A. R., Knowler, W. C., Nathan, D. M., Altshuler, D., and Diabetes Prevention Program Research Group. 2006. TCF7L2 polymorphisms and progression to diabetes in the Diabetes Prevention Program. N. Engl. J. Med. 355: 241-250.

Folkman, J. and Shing, Y. 1992. Angiogenesis J. Biol. Chem. 267: 10931-10934.

Folkman, J., Watson, K., Ingber, D., and Hanahan, D. 1989. Induction of angiogenesis during the transition from hyperplasia to neoplasia Nature 339: 58-61.

Franklin, I. and Lewontin, R. C. 1970. Is the gene the unit of selection? Genetics 65: 707-734.

Gan, M. J., Albanese-O'Neill, A., and Haller, M. J. 2012. Type 1 diabetes: current concepts in epidemiology, pathophysiology, clinical care, and research. Current Problems in Pediatric & Adolescent Health Care 42: 269-291.

Giarelli, E. and Jacobs, L. A. 2005. Modifying cancer risk factors: the gene-environment interaction. Semin. Oncol. Nurs. 21: 271-277.

Goldstein, D. B. 2009. Common genetic variation and human traits N. Engl. J. Med. 360: 1696-1698.

Gong, J., Hutter, C., Baron, J. A., Berndt, S., Caan, B., Campbell, P. T., Casey, G., Chan, A. T., Cotterchio, M., Fuchs, C. S., et al. 2012. A pooled analysis of smoking and colorectal cancer: timing of exposure and interactions with environmental factors. Cancer Epidemiol. Biomarkers Prev. 21: 1974-1985.

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

Grandison, R. C., Piper, M. D., and Partridge, L. 2009. Amino-acid imbalance explains extension of lifespan by dietary restriction in Drosophila Nature 462: 1061-1064.

Grant, S. F. A., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Manolescu, A., Sainz, J., Helgason, A., Stefansson, H., Emilsson, V., Helgadottir, A., et al. 2006. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. Nat. Genet. 38: 320-323.

Greene, C. S., Penrod, N. M., Williams, S. M., and Moore, J. H. 2009. Failure to replicate a genetic association may provide important clues about genetic architecture. PLoS One 4: e5639.

Greenwood, T. A., Akiskal, H. S., Akiskal, K. K., and Kelsoe, J. R. 2012. Genome-Wide Association Study of Temperament in Bipolar Disorder Reveals Significant Associations with Three Novel Loci. Biol. Psychiatry 72: 303-310.

Gross, O., Girgert, R., Rubel, D., Temme, J., Theissen, S., and Muller, G. A. 2011. Renal protective effects of aliskiren beyond its antihypertensive property in a mouse model of progressive fibrosis Am. J. Hypertens. 24: 355-361.

Gu, J. W., Bailey, A. P., Sartin, A., Makey, I., and Brady, A. L. 2005. Ethanol stimulates tumor progression and expression of vascular endothelial growth factor in chick embryos Cancer 103: 422-431.

Gu, J. W., Elam, J., Sartin, A., Li, W., Roach, R., and Adair, T. H. 2001. Moderate levels of ethanol induce expression of vascular endothelial growth factor and stimulate angiogenesis Am. J. Physiol. Regul. Integr. Comp. Physiol. 281: R365-72.

Guo, J., Li, W., Wu, Z., Cheng, X., Wang, Y., and Chen, T. 2013. Association between 9p21.3 genomic markers and coronary artery disease in East Asians: a meta-analysis involving 9,813 cases and 10,710 controls. Mol. Biol. Rep. 40: 337-343.

Haggar, F. A. and Boushey, R. P. 2009. Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors Clin. Colon Rectal Surg. 22: 191-197.

Hanahan, D. and Folkman, J. 1996. Patterns and emerging mechanisms of the angiogenic switch during tumorigenesis Cell 86: 353-364.

Hanahan, D. and Weinberg, R. A. 2011. Hallmarks of cancer: the next generation Cell 144: 646-674.

Hanahan, D. and Weinberg, R. A. 2000. The hallmarks of cancer Cell 100: 57-70.

Hardwick, J. C., Van Den Brink, G. R., Bleuming, S. A., Ballester, I., Van Den Brande, J. M., Keller, J. J., Offerhaus, G. J., Van Deventer, S. J., and Peppelenbosch, M. P. 2004. Bone morphogenetic protein 2 is expressed by, and acts upon, mature epithelial cells in the colon. Gastroenterology 126: 111-121.

Harrell, F. E.,Jr, Lee, K. L., and Mark, D. B. 1996. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors Stat. Med. 15: 361-387.

Harris, A. L. 2002. Hypoxia--a key regulatory factor in tumour growth Nat. Rev. Cancer. 2: 38-47.

Haydon, A. M., Macinnis, R. J., English, D. R., and Giles, G. G. 2006. Effect of physical activity and body size on survival after diagnosis with colorectal cancer. Gut 55: 62-67.

Heeschen, C., Chang, E., Aicher, A., and Cooke, J. P. 2006. Endothelial progenitor cells participate in nicotine-mediated angiogenesis J. Am. Coll. Cardiol. 48: 2553-2560.

Heeschen, C., Jang, J. J., Weis, M., Pathak, A., Kaji, S., Hu, R. S., Tsao, P. S., Johnson, F. L., and Cooke, J. P. 2001. Nicotine stimulates angiogenesis and promotes tumor growth and atherosclerosis Nat. Med. 7: 833-839.

Herder, C. and Roden, M. 2011. Genetics of type 2 diabetes: pathophysiologic and clinical relevance Eur. J. Clin. Invest. 41: 679-692.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., and Manolio, T. A. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits Proc. Natl. Acad. Sci. U. S. A. 106: 9362-9367.

Houlston, R. S., Webb, E., Broderick, P., Pittman, A. M., Di Bernardo, M. C., Lubbe, S., Chandler, I., Vijayakrishnan, J., Sullivan, K., Penegar, S., et al. 2008. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. Nat. Genet. 40: 1426-1435.

Hsiao, P. C., Chen, M. K., Su, S. C., Ueng, K. C., Chen, Y. C., Hsieh, Y. H., Liu, Y. F., Tsai, H. T., and Yang, S. F. 2010. Hypoxia inducible factor-1alpha gene polymorphism G1790A and its interaction with tobacco and alcohol consumptions increase susceptibility to hepatocellular carcinoma. J. Surg. Oncol. 102: 163-169.

Hunter, D. J. and Chanock, S. J. 2010. Genome-wide association studies and "the art of the soluble". J. Natl. Cancer Inst. 102: 836-837.

Hutter, C. M., Chang-Claude, J., Slattery, M. L., Pflugeisen, B. M., Lin, Y., Duggan, D., Nan, H., Lemire, M., Rangrej, J., Figueiredo, J. C., et al. 2012. Characterization of gene-environment interactions for colorectal cancer susceptibility loci Cancer Res. 72: 2036-2044.

Imamura, T., Kikuchi, H., Herraiz, M. T., Park, D. Y., Mizukami, Y., Mino-Kenduson, M., Lynch, M. P., Rueda, B. R., Benita, Y., Xavier, R. J., et al. 2009. HIF-1alpha and HIF-2alpha have divergent roles in colon cancer Int. J. Cancer 124: 763-771.

International Agency for Research on Cancer. 2008. World Cancer Report 2008. International Agency for Research on Cancer, Albany, NY, USA.

International Consortium for Blood Pressure Genome-Wide Association,Studies, Ehret, G. B., Munroe, P. B., Rice, K. M., Bochud, M., Johnson, A. D., Chasman, D. I., Smith, A. V., Tobin, M. D., Verwoert, G. C., et al. 2011. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature 478: 103-109.

International HapMap Consortium. 2003. The International HapMap Project. Nature 426: 789-796.

Jang, M. J., Jeon, Y. J., Kim, J. W., Cho, Y. K., Lee, S. K., Hwang, S. G., Oh, D., and Kim, N. K. 2013. Association of VEGF and KDR single nucleotide polymorphisms with colorectal cancer susceptibility in Koreans Mol. Carcinog. 52 Suppl 1: E60-9.

Johnson, C. M., Wei, C., Ensor, J. E., Smolenski, D. J., Amos, C. I., Levin, B., and Berry, D. A. 2013. Meta-analyses of colorectal cancer risk factors. Cancer Causes Control 24: 1207-1222.

Jorgensen, T. J., Ruczinski, I., Kessing, B., Smith, M. W., Shugart, Y. Y., and Alberg, A. J. 2009. Hypothesis-driven candidate gene association studies: practical design and analytical considerations Am. J. Epidemiol. 170: 986-993.

Kashtan, C. E. 1993. Alport Syndrome and Thin Basement Membrane Nephropathy. In GeneReviews (eds. R. A. Pagon, M. P. Adam, T. D. Bird, C. R. Dolan, C. T. Fong, and K. Stephens), University of Washington, Seattle, Seattle (WA).

Khoury, M. J. and Wacholder, S. 2009. Invited commentary: from genome-wide association studies to gene-environment-wide interaction studies--challenges and opportunities Am. J. Epidemiol. 169: 227-30; discussion 234-5.

Kim, E., Coelho, D., and Blachier, F. 2013. Review of the association between meat consumption and risk of colorectal cancer. Nutr. Res. 33: 983-994.

Kim, J. G., Chae, Y. S., Sohn, S. K., Cho, Y. Y., Moon, J. H., Park, J. Y., Jeon, S. W., Lee, I. T., Choi, G. S., and Jun, S. H. 2008. Vascular endothelial growth factor gene polymorphisms associated with prognosis for patients with colorectal cancer Clin. Cancer Res. 14: 62-66.

Kraft, P. and Raychaudhuri, S. 2009. Complex diseases, complex genes: keeping pathways on the right track Epidemiology 20: 508-511.

Kremeyer, B., Garcia, J., Muller, H., Burley, M. W., Herzberg, I., Parra, M. V., Duque, C., Vega, J., Montoya, P., Lopez, M. C., et al. 2010. Genome-wide linkage scan of bipolar disorder in a Colombian population isolate replicates Loci on chromosomes 7p21-22, 1p31, 16p12 and 21q21-22 and identifies a novel locus on chromosome 12q Hum. Hered. 70: 255-268.

Kroll, J. and Waltenberger, J. 1998. VEGF-A induces expression of eNOS and iNOS in endothelial cells via VEGF receptor-2 (KDR). Biochem. Biophys. Res. Commun. 252: 743-746.

Ku, C. S., Loy, E. Y., Pawitan, Y., and Chia, K. S. 2010. The pursuit of genome-wide association studies: where are we now? J. Hum. Genet. .

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

Le Marchand, L. and Wilkens, L. R. 2008. Design considerations for genomic association studies: importance of gene-environment interactions Cancer Epidemiol. Biomarkers Prev. 17: 263-267.

Lee, H., Woo, H. G., Greenwood, T. A., Kripke, D. F., and Kelsoe, J. R. 2013. A genome-wide association study of seasonal pattern mania identifies NF1A as a possible susceptibility gene for bipolar disorder. J. Affect. Disord. 145: 200-207.

Leeper, N. J., Raiesdana, A., Kojima, Y., Kundu, R. K., Cheng, H., Maegdefessel, L., Toh, R., Ahn, G. O., Ali, Z. A., Anderson, D. R., et al. 2013. Loss of CDKN2B promotes p53-dependent smooth muscle cell apoptosis and aneurysm formation. Arterioscler. Thromb. Vasc. Biol. 33: e1-e10.

Lettre, G., Palmer, C. D., Young, T., Ejebe, K. G., Allayee, H., Benjamin, E. J., Bennett, F., Bowden, D. W., Chakravarti, A., Dreisbach, A., et al. 2011. Genome-wide association study of coronary heart disease and its risk factors in 8,090 African Americans: the NHLBI CARe Project PLoS Genet. 7: e1001300.

Lichtenstein, P., Holm, N. V., Verkasalo, P. K., Iliadou, A., Kaprio, J., Koskenvuo, M., Pukkala, E., Skytthe, A., and Hemminki, K. 2000. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. N. Engl. J. Med. 343: 78-85.

Lin, J., Zhou, Z. G., Wang, J. P., Zhang, C., and Huang, G. 2008. From Type 1, through LADA, to type 2 diabetes: a continuous spectrum? Ann. N. Y. Acad. Sci. 1150: 99-102.

Lindstrom, S., Schumacher, F., Siddiq, A., Travis, R. C., Campa, D., Berndt, S. I., Diver, W. R., Severi, G., Allen, N., Andriole, G., et al. 2011. Characterizing associations and SNP-environment interactions for GWAS-identified prostate cancer risk markers--results from BPC3. PLoS One 6: e17142.

Liu, K., Slattery, M., Jacobs, D.,Jr, Cutter, G., McDonald, A., Van Horn, L., Hilner, J. E., Caan, B., Bragg, C., and Dyer, A. 1994. A study of the reliability and comparative validity of the cardia dietary history Ethn. Dis. 4: 15-27.

Lohela, M., Bry, M., Tammela, T., and Alitalo, K. 2009. VEGFs and receptors involved in angiogenesis versus lymphangiogenesis. Curr. Opin. Cell Biol. 21: 154-165.

Luchtenborg, M., White, K. K., Wilkens, L., Kolonel, L. N., and Le Marchand, L. 2007. Smoking and colorectal cancer: different effects by type of cigarettes? Cancer Epidemiol. Biomarkers Prev. 16: 1341-1347.

Mancia, G., De Backer, G., Dominiczak, A., Cifkova, R., Fagard, R., Germano, G., Grassi, G., Heagerty, A. M., Kjeldsen, S. E., Laurent, S., et al. 2007. 2007 Guidelines for the Management of Arterial Hypertension: The Task Force for the Management of Arterial Hypertension of the European Society of Hypertension (ESH) and of the European Society of Cardiology (ESC). J. Hypertens. 25: 1105-1187.

Manolio, T. A. 2010. Genomewide association studies and assessment of the risk of disease N. Engl. J. Med. 363: 166-176.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., et al. 2009. Finding the missing heritability of complex diseases Nature 461: 747-753.

Marchand, L. L. 1999. Combined influence of genetic and dietary factors on colorectal cancer incidence in Japanese Americans. J. Natl. Cancer. Inst. Monogr. (26): 101-105.

Marchini, J., Donnelly, P., and Cardon, L. R. 2005. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat. Genet. 37: 413-417.

Mason, S., Piper, M., Gronostajski, R. M., and Richards, L. J. 2009. Nuclear factor one transcription factors in CNS development Mol. Neurobiol. 39: 10-23.

McCleary, N. J., Niedzwiecki, D., Hollis, D., Saltz, L. B., Schaefer, P., Whittom, R., Hantel, A., Benson, A., Goldberg, R., and Meyerhardt, J. A. 2010. Impact of smoking on patients with stage III colon cancer: results from Cancer and Leukemia Group B 89803. Cancer 116: 957-966.

McDonald, A., Van Horn, L., Slattery, M., Hilner, J., Bragg, C., Caan, B., Jacobs, D.,Jr, Liu, K., Hubert, H., and Gernhofer, N. 1991. The CARDIA dietary history: development, implementation, and evaluation J. Am. Diet. Assoc. 91: 1104-1112.

Min, K. J. and Tatar, M. 2006. Restriction of amino acids extends lifespan in Drosophila melanogaster Mech. Ageing Dev. 127: 643-646.

Mitsunaga, S., Hosomichi, K., Okudaira, Y., Nakaoka, H., Kunii, N., Suzuki, Y., Kuwana, M., Sato, S., Kaneko, Y., Homma, Y., et al. 2013. Exome sequencing identifies novel rheumatoid arthritis-susceptible variants in the BTNL2. J. Hum. Genet. 58: 210-215.

Mizukami, ,Yusuke and Chung, ,Daniel. 2007. Hypoxia, angiogenesis, and colorectal cancer. Current Colorectal Cancer Reports 71-75.

Morais, A., Lima, B., Peixoto, M. J., Alves, H., Marques, A., and Delgado, L. 2012. BTNL2 gene polymorphism associations with susceptibility and phenotype expression in sarcoidosis. Respir. Med. 106: 1771-1777.

Mousa, S. and Mousa, S. A. 2006. Cellular and molecular mechanisms of nicotine's pro-angiogenesis activity and its potential impact on cancer J. Cell. Biochem. 97: 1370-1378.

Murcray, C. E., Lewinger, J. P., and Gauderman, W. J. 2009. Gene-environment interaction in genome-wide association studies Am. J. Epidemiol. 169: 219-226.

Murtaugh, M. A., Ma, K. N., Sweeney, C., Caan, B. J., and Slattery, M. L. 2004. Meat consumption patterns and preparation, genetic variants of metabolic enzymes, and their association with rectal cancer in men and women J. Nutr. 134: 776-784.

Neufeld, G., Cohen, T., Gengrinovitch, S., and Poltorak, Z. 1999. Vascular endothelial growth factor (VEGF) and its receptors. FASEB J. 13: 9-22.

Nieder, C. and Bremnes, R. M. 2008. Effects of smoking cessation on hypoxia and its potential impact on radiation treatment effects in lung cancer patients Strahlenther. Onkol. 184: 605-609.

Nihon-Yanagi, Y., Terai, K., Murano, T., Matsumoto, T., and Okazumi, S. 2012. Tissue expression of Toll-like receptors 2 and 4 in sporadic human colorectal cancer. Cancer Immunol. Immunother. 61: 71-77.

Nishanian, T. G., Kim, J. S., Foxworth, A., and Waldman, T. 2004. Suppression of tumorigenesis and activation of Wnt signaling by bone morphogenetic protein 4 in human cancer cells Cancer. Biol. Ther. 3: 667-675.

Nishimoto, A., Kugimiya, N., Hosoyama, T., Enoki, T., Li, T. S., and Hamano, K. 2014. HIF-1alpha activation under glucose deprivation plays a central role in the acquisition of anti-apoptosis in human colon cancer cells. Int. J. Oncol. 44: 2077-2084.

Oba, S., Shimizu, N., Nagata, C., Shimizu, H., Kametani, M., Takeyama, N., Ohnuma, T., and Matsushita, S. 2006. The relationship between the consumption of meat, fat, and coffee and the risk of colon cancer: A prospective study in Japan. Cancer Lett. 244: 260-267.

Orozco, G., Eerligh, P., Sanchez, E., Zhernakova, S., Roep, B. O., Gonzalez-Gay, M. A., Lopez-Nevot, M. A., Callejas, J. L., Hidalgo, C., Pascual-Salcedo, D., et al. 2005. Analysis of a functional BTNL2 polymorphism in type 1 diabetes, rheumatoid arthritis, and systemic lupus erythematosus. Hum. Immunol. 66: 1235-1241.

Passarelli, M. N., Coghill, A. E., Hutter, C. M., Zheng, Y., Makar, K. W., Potter, J. D., and Newcomb, P. A. 2011. Common colorectal cancer risk variants in SMAD7 are associated with survival among prediagnostic nonsteroidal anti-inflammatory drug users: a population-based study of postmenopausal women Genes Chromosomes Cancer 50: 875-886.

Pearson, E. R. 2009. Translating TCF7L2: from gene to function. Diabetologia 52: 1227-1230.

Pelser, C., Arem, H., Pfeiffer, R. M., Elena, J. W., Alfano, C. M., Hollenbeck, A. R., and Park, Y. 2014. Prediagnostic lifestyle factors and survival after colon and rectal cancer diagnosis in the National Institutes of Health (NIH)-AARP Diet and Health Study. Cancer 120: 1540-1547.

Phipps, A. I., Baron, J., and Newcomb, P. A. 2011. Prediagnostic smoking history, alcohol consumption, and colorectal cancer survival: the Seattle Colon Cancer Family Registry. Cancer 117: 4948-4957.

Piegorsch, W. W., Weinberg, C. R., and Taylor, J. A. 1994. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies Stat. Med. 13: 153-162.

Pilbrow, A. P., Folkersen, L., Pearson, J. F., Brown, C. M., McNoe, L., Wang, N. M., Sweet, W. E., Tang, W. H., Black, M. A., Troughton, R. W., et al. 2012. The chromosome 9p21.3 coronary heart disease risk allele is associated with altered gene expression in normal heart and vascular tissues. PLoS One 7: e39574.

Pimentel-Nunes, P., Teixeira, A. L., Pereira, C., Gomes, M., Brandao, C., Rodrigues, C., Goncalves, N., Boal-Carvalho, I., Roncon-Albuquerque, R.,Jr, Moreira-Dias, L., et al. 2013. Functional polymorphisms of Toll-like receptors 2 and 4 alter the risk for colorectal carcinoma in Europeans. Dig. Liver Dis. 45: 63-69.

Poschl, G. and Seitz, H. K. 2004. Alcohol and cancer Alcohol Alcohol. 39: 155-165.

Potter, J. D. 1999a. Colorectal cancer: molecules and populations. J. Natl. Cancer Inst. 91: 916-932.

Potter, J. D. 1999b. Colorectal cancer: molecules and populations J. Natl. Cancer Inst. 91: 916-932.

Poynter, J. N., Haile, R. W., Siegmund, K. D., Campbell, P. T., Figueiredo, J. C., Limburg, P., Young, J., Le Marchand, L., Potter, J. D., Cotterchio, M., et al. 2009. Associations between smoking, alcohol consumption, and colorectal cancer, overall and by tumor microsatellite instability status. Cancer Epidemiol. Biomarkers Prev. 18: 2745-2750.

Prentice, R. L. 2011. Empirical Evaluation of Gene and Environment Interactions: Methods and Potential. Journal of the National Cancer Institute 103: 1209-1210.

Prentice, R. L., Huang, Y., Hinds, D. A., Peters, U., Pettinger, M., Cox, D. R., Beilharz, E., Chlebowski, R. T., Rossouw, J. E., Caan, B., et al. 2009. Variation in the FGFR2 Gene and the Effects of Postmenopausal Hormone Therapy on Invasive Breast Cancer. Cancer Epidemiology Biomarkers & Prevention 18: 3079-3085.

Psychiatric GWAS Consortium Bipolar Disorder Working,Group. 2011. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. Nat. Genet. 43: 977-983.

Rajaganeshan, R., Prasad, R., Guillou, P. J., Poston, G., Scott, N., and Jayne, D. G. 2008. The role of hypoxia in recurrence following resection of Dukes' B colorectal cancer Int. J. Colorectal Dis. 23: 1049-1055.

Rajkowska, G. 2003. Depression: what we can learn from postmortem studies Neuroscientist 9: 273-284.

Ravi, R., Mookerjee, B., Bhujwalla, Z. M., Sutter, C. H., Artemov, D., Zeng, Q., Dillehay, L. E., Madan, A., Semenza, G. L., and Bedi, A. 2000. Regulation of tumor angiogenesis by p53-induced degradation of hypoxia-inducible factor 1alpha Genes Dev. 14: 34-44.

Rebbeck, T. R., Spitz, M., and Wu, X. 2004. Assessing the function of genetic variants in candidate gene association studies Nat. Rev. Genet. 5: 589-597.

Reich, D. E. and Lander, E. S. 2001. On the allelic spectrum of human disease Trends Genet. 17: 502-510.

Robsahm, T. E., Aagnes, B., Hjartaker, A., Langseth, H., Bray, F. I., and Larsen, I. K. 2013. Body mass index, physical activity, and colorectal cancer by anatomical subsites: a systematic review and meta-analysis of cohort studies. Eur. J. Cancer Prev. 22: 492-505.

Roglic, G., Unwin, N., Bennett, P. H., Mathers, C., Tuomilehto, J., Nag, S., Connolly, V., and King, H. 2005. The burden of mortality attributable to diabetes: realistic estimates for the year 2000 Diabetes Care 28: 2130-2135.

Ross, R. 1989. Angiogenesis. Successful growth of tumours Nature 339: 16-17.

Ruczinski, I., Kooperberg, C., and LeBlanc, M. 2003. Logic Regression. Journal of Computational and Graphical Statistics 12: 475-511.

Saade, S., Cazier, J. B., Ghassibe-Sabbagh, M., Youhanna, S., Badro, D. A., Kamatani, Y., Hager, J., Yeretzian, J. S., El-Khazen, G., Haber, M., et al. 2011. Large scale association analysis identifies three susceptibility loci for coronary artery disease PLoS One 6: e29427.

Saridaki, Z., Souglakos, J., and Georgoulias, V. 2014. Prognostic and predictive significance of MSI in stages II/III colon cancer. World J. Gastroenterol. 20: 6809-6814.

Savic, D., Bell, G. I., and Nobrega, M. A. 2012. An in vivo cis-Regulatory Screen at the Type 2 Diabetes Associated TCF7L2 Locus Identifies Multiple Tissue-Specific Enhancers PLoS One 7: e36501.

Schmitz, K. J., Muller, C. I., Reis, H., Alakus, H., Winde, G., Baba, H. A., Wohlschlaeger, J., Jasani, B., Fandrey, J., and Schmid, K. W. 2009. Combined analysis of hypoxia-inducible factor 1 alpha and metallothionein indicates an aggressive subtype of colorectal carcinoma Int. J. Colorectal Dis. 24: 1287-1296.

Schunkert, H., Konig, I. R., Kathiresan, S., Reilly, M. P., Assimes, T. L., Holm, H., Preuss, M., Stewart, A. F., Barbalic, M., Gieger, C., et al. 2011. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nat. Genet. 43: 333-338.

Schwender, H. and Ruczinski, I. 2010. Logic regression and its extensions. Adv. Genet. 72: 25-45.

Semenza, G. L. 2010. Defining the role of hypoxia-inducible factor 1 in cancer biology and therapeutics Oncogene 29: 625-634.

Shin, A., Joo, J., Bak, J., Yang, H. R., Kim, J., Park, S., and Nam, B. H. 2011. Site-specific risk factors for colorectal cancer in a Korean population. PLoS One 6: e23196.

Shin, V. Y., Wu, W. K., Chu, K. M., Wong, H. P., Lam, E. K., Tai, E. K., Koo, M. W., and Cho, C. H. 2005. Nicotine induces cyclooxygenase-2 and vascular endothelial growth factor receptor-2 in association with tumor-associated invasion and angiogenesis in gastric cancer Mol. Cancer. Res. 3: 607-615.

Siegert, S., Hampe, J., Schafmayer, C., von Schonfels, W., Egberts, J. H., Forsti, A., Chen, B., Lascorz, J., Hemminki, K., Franke, A., et al. 2013. Genome-wide investigation of gene-environment interactions in colorectal cancer Hum. Genet. 132: 219-231.

Sillars-Hardebol, A. H., Carvalho, B., de Wit, M., Postma, C., Delis-van Diemen, P. M., Mongera, S., Ylstra, B., van de Wiel, M. A., Meijer, G. A., and Fijneman, R. J. 2010. Identification of key genes for carcinogenic pathways associated with colorectal adenoma-to-carcinoma progression Tumour Biol. 31: 89-96.

Silvester, K. R. and Cummings, J. H. 1995. Does digestibility of meat protein help explain large bowel cancer risk? Nutr. Cancer 24: 279-288.

Singh, J. C., Cruickshank, S. M., Newton, D. J., Wakenshaw, L., Graham, A., Lan, J., Lodge, J. P., Felsburg, P. J., and Carding, S. R. 2005. Toll-like receptor-mediated responses of primary intestinal epithelial cells during the development of colitis. Am. J. Physiol. Gastrointest. Liver Physiol. 288: G514-24.

Slattery, M. L. 2000. Diet, lifestyle, and colon cancer. Semin. Gastrointest. Dis. 11: 142-146.

Slattery, M. L., Caan, B. J., Benson, J., and Murtaugh, M. 2003. Energy balance and rectal cancer: an evaluation of energy intake, energy expenditure, and body mass index Nutr. Cancer 46: 166-171.

Slattery, M. L., Caan, B. J., Duncan, D., Berry, T. D., Coates, A., and Kerber, R. 1994. A computerized diet history questionnaire for epidemiologic studies J. Am. Diet. Assoc. 94: 761-766.

Slattery, M. L., Curtin, K., Wolff, R. K., Boucher, K. M., Sweeney, C., Edwards, S., Caan, B. J., and Samowitz, W. 2009. A comparison of colon and rectal somatic DNA alterations Dis. Colon Rectum 52: 1304-1311.

Slattery, M. L., Edwards, S. L., Caan, B. J., Kerber, R. A., and Potter, J. D. 1995. Response rates among control subjects in case-control studies Ann. Epidemiol. 5: 245-249.

Slattery, M. L., Herrick, J. S., Bondurant, K. L., and Wolff, R. K. 2012a. Toll-like receptor genes and their association with colon and rectal cancer development and prognosis. Int. J. Cancer 130: 2974-2980.

Slattery, M. L., Lundgreen, A., Herrick, J. S., Kadlubar, S., Caan, B. J., Potter, J. D., and Wolff, R. K. 2012b. Genetic variation in bone morphogenetic protein and colon and rectal cancer. Int. J. Cancer 130: 653-664.

Slattery, M. L., Lundgreen, A., and Wolff, R. K. 2014. VEGFA, FLT1, KDR and colorectal cancer: assessment of disease risk, tumor molecular phenotype, and survival Mol. Carcinog. 53 Suppl 1: E140-50.

Slattery, M. L., Potter, J., Caan, B., Edwards, S., Coates, A., Ma, K. N., and Berry, T. D. 1997a. Energy balance and colon cancer--beyond physical activity Cancer Res. 57: 75-80.

Slattery, M. L., Potter, J. D., Friedman, G. D., Ma, K. N., and Edwards, S. 1997b. Tobacco use and colon cancer Int. J. Cancer 70: 259-264.

Slattery, M. L., Potter, J. D., Ma, K. N., Caan, B. J., Leppert, M., and Samowitz, W. 2000. Western diet, family history of colorectal cancer, NAT2, GSTM-1 and risk of colon cancer. Cancer Causes Control 11: 1-8.

Smoller, J. W. and Finn, C. T. 2003. Family, twin, and adoption studies of bipolar disorder Am. J. Med. Genet. C. Semin. Med. Genet. 123C: 48-58.

Stahl, E. A., Raychaudhuri, S., Remmers, E. F., Xie, G., Eyre, S., Thomson, B. P., Li, Y., Kurreeman, F. A., Zhernakova, A., Hinks, A., et al. 2010. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat. Genet. 42: 508-514.

Storey, J. D. and Tibshirani, R. 2003. Statistical significance for genomewide studies Proc. Natl. Acad. Sci. U. S. A. 100: 9440-9445.

Stranger, B. E., Stahl, E. A., and Raj, T. 2011. Progress and promise of genome-wide association studies for human complex trait genetics. Genetics 187: 367-383.

Sun, Z., Liu, L., Wang, P. P., Roebothan, B., Zhao, J., Dicks, E., Cotterchio, M., Buehler, S., Campbell, P. T., McLaughlin, J. R., et al. 2012a. Association of total energy intake and macronutrient consumption with colorectal cancer risk: results from a large population-based case-control study in Newfoundland and Labrador and Ontario, Canada. Nutrition Journal 11: 18.

Sun, Z., Zhu, Y., Wang, P. P., Roebothan, B., Zhao, J., Zhao, J., Dicks, E., Cotterchio, M., Buehler, S., Campbell, P. T., et al. 2012b. Reported intake of selected micronutrients and risk of colorectal cancer: results from a large population-based case-control study in Newfoundland, Labrador and Ontario, Canada. Anticancer Res. 32: 687-696.

Takachi, R., Tsubono, Y., Baba, K., Inoue, M., Sasazuki, S., Iwasaki, M., and Tsugane, S. 2011. Red meat intake may increase the risk of colon cancer in Japanese, a population with relatively low red meat consumption. Asia Pac. J. Clin. Nutr. 20: 603-612.

Talks, K. L., Turley, H., Gatter, K. C., Maxwell, P. H., Pugh, C. W., Ratcliffe, P. J., and Harris, A. L. 2000. The expression and distribution of the hypoxia-inducible factors HIF-1alpha and HIF-2alpha in normal human tissues, cancers, and tumor-associated macrophages Am. J. Pathol. 157: 411-421.

Tan, W., Bailey, A. P., Shparago, M., Busby, B., Covington, J., Johnson, J. W., Young, E., and Gu, J. W. 2007. Chronic alcohol consumption stimulates VEGF expression, tumor angiogenesis and progression of melanoma in mice. Cancer. Biol. Ther. 6: 1211-1217.

Tao, S., Feng, J., Webster, T., Jin, G., Hsu, F. C., Chen, S. H., Kim, S. T., Wang, Z., Zhang, Z., Zheng, S. L., et al. 2012. Genome-wide two-locus epistasis scans in prostate cancer using two European populations. Hum. Genet. 131: 1225-1234.

Tchorzewski, M., Lewkowicz, P., Dziki, A., and Tchorzewski, H. 2014. Expression of toll-like receptors on human rectal adenocarcinoma cells. Arch. Immunol. Ther. Exp. (Warsz) 62: 247-251.

Thibodeau, S. N., Bren, G., and Schaid, D. 1993. Microsatellite instability in cancer of the proximal colon. Science 260: 816-819.

Thomas, D. 2010a. Gene-environment-wide association studies: emerging approaches Nat. Rev. Genet. 11: 259-272.

Thomas, D. 2010b. Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies Annu. Rev. Public Health 31: 21-36.

Thomas, D. C., Conti, D. V., Baurley, J., Nijhout, F., Reed, M., and Ulrich, C. M. 2009. Use of pathway information in molecular epidemiology Hum. Genomics 4: 21-42.

Thompson, J. R., Attia, J., and Minelli, C. 2011. The meta-analysis of genome-wide association studies. Briefings in Bioinformatics 12: 259-269.

Tomlinson, I. P., Carvajal-Carmona, L. G., Dobbins, S. E., Tenesa, A., Jones, A. M., Howarth, K., Palles, C., Broderick, P., Jaeger, E. E., Farrington, S., et al. 2011. Multiple common susceptibility variants near BMP pathway loci GREM1, BMP4, and BMP2 explain part of the missing heritability of colorectal cancer. PLoS Genet. 7: e1002105.

Travis, R. C., Reeves, G. K., Green, J., Bull, D., Tipper, S. J., Baker, K., Beral, V., Peto, R., Bell, J., Zelenika, D., et al. 2010. Gene-environment interactions in 7610 women with breast cancer: prospective evidence from the Million Women Study. Lancet 375: 2143-2151.

Ulrich, C. M., Kampman, E., Bigler, J., Schwartz, S. M., Chen, C., Bostick, R., Fosdick, L., Beresford, S. A., Yasui, Y., and Potter, J. D. 1999. Colorectal adenomas and the C677T MTHFR polymorphism: evidence for gene-environment interaction? Cancer Epidemiol. Biomarkers Prev. 8: 659-668.

van de Wiel, A. 2004. Diabetes mellitus and alcohol Diabetes Metab. Res. Rev. 20: 263-267.

Viatte, S., Plant, D., and Raychaudhuri, S. 2013. Genetics and epigenetics of rheumatoid arthritis. Nat. Rev. Rheumatol. 9: 141-153.

Vigne, P. and Frelin, C. 2008. The role of polyamines in protein-dependent hypoxic tolerance of Drosophila BMC Physiol. 8: 22.

Vigne, P. and Frelin, C. 2006. A low protein diet increases the hypoxic tolerance in Drosophila PLoS One 1: e56.

Volterra, A. and Meldolesi, J. 2005. Astrocytes, from brain glue to communication elements: the revolution continues Nat. Rev. Neurosci. 6: 626-640.

Voorrips, L. E., Goldbohm, R. A., van Poppel, G., Sturmans, F., Hermus, R. J., and van den Brandt, P. A. 2000. Vegetable and fruit consumption and risks of colon and rectal cancer in a prospective cohort study: The Netherlands Cohort Study on Diet and Cancer. Am. J. Epidemiol. 152: 1081-1092.

Wang, K., Li, M., and Bucan, M. 2007. Pathway-Based Approaches for Analysis of Genomewide Association Studies Am. J. Hum. Genet. 81: .

Wei, E. K., Giovannucci, E., Wu, K., Rosner, B., Fuchs, C. S., Willett, W. C., and Colditz, G. A. 2004. Comparison of risk factors for colon and rectal cancer. Int. J. Cancer 108: 433-442.

Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661-678.

Westra, H. J., Peters, M. J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M. W., Fairfax, B. P., Schramm, K., Powell, J. E., et al. 2013. Systematic identification of trans eQTLs as putative drivers of known disease associations Nat. Genet. 45: 1238-1243.

Whiffin, N. and Houlston, R. S. 2014. Architecture of inherited susceptibility to colorectal cancer: a voyage of discovery Genes (Basel) 5: 270-284.

Witte, J. S. 2010. Genome-wide association studies and beyond Annu. Rev. Public Health 31: 9-20 4 p following 20.

Wong, H. P., Yu, L., Lam, E. K., Tai, E. K., Wu, W. K., and Cho, C. H. 2007. Nicotine promotes colon tumor growth and angiogenesis through beta-adrenergic activation Toxicol. Sci. 97: 279-287.

Wu, J., Devlin, B., Ringquist, S., Trucco, M., and Roeder, K. 2010. Screen and clean: a tool for identifying interactions in genome-wide association studies. Genet. Epidemiol. 34: 275-285.

Xiang, L., Wang, S., Jin, X., Duan, W., Ding, X., and Zheng, C. 2012. Expression of BMP2, TLR3, TLR4 and COX2 in colorectal polyps, adenoma and adenocarcinoma Mol. Med. Rep. 6: 973-976.

Xing, J., Myers, R. E., He, X., Qu, F., Zhou, F., Ma, X., Hyslop, T., Bao, G., Wan, S., Yang, H., et al. 2011. GWAS-identified colorectal cancer susceptibility locus associates with disease prognosis Eur. J. Cancer 47: 1699-1707.

Yang, S. Y., Kim, Y. S., Song, J. H., Chung, S. J., Lee, I. H., Hong, K. J., Lee, E. J., Kim, D. H., Yim, J. Y., Park, M. J., et al. 2012. [Dietary risk factors in relation to colorectal adenoma]. Korean Journal of Gastroenterology/Taehan Sohwagi Hakhoe Chi 60: 102-108.

Yasui, Y. May 2012. Why odds ratio estimates of GWAS are almost always close to 1.0 COBRA Preprint Series. Working Paper 94 .

Young, J. L. J., Roffers, S. D., Ries, L. A. G., Fritz, A. G., and Hurlbut, A. A. (. 2001. SEER Summary Staging Manual - 2000: Codes and Coding Instructions. National Cancer Institute NIH Pub. No. 01-4969,: .

Zhang, X., Gaspard, J. P., and Chung, D. C. 2001. Regulation of vascular endothelial growth factor by the Wnt and K-ras pathways in colonic neoplasia Cancer Res. 61: 6050-6054.

Zhang, Y. and Liu, J. S. 2007. Bayesian inference of epistatic interactions in case-control studies. Nat. Genet. 39: 1167-1173.

Zhong, H., De Marzo, A. M., Laughner, E., Lim, M., Hilton, D. A., Zagzag, D., Buechler, P., Isaacs, W. B., Semenza, G. L., and Simons, J. W. 1999. Overexpression of hypoxia-inducible factor 1alpha in common human cancers and their metastases Cancer Res. 59: 5830-5835.

Zhu, Y., Wang, P. P., Zhao, J., Green, R., Sun, Z., Roebothan, B., Squires, J., Buehler, S., Dicks, E., Zhao, J., et al. 2014. Dietary N-nitroso compounds and risk of colorectal cancer: a case-control study in Newfoundland and Labrador and Ontario, Canada. Br. J. Nutr. 111: 1109-1117.