Multilocus Approaches to the Detection of Disease Susceptibility Regions: Methods and Applications

Julia Grant Ciampa
Lincoln College, Oxford

A thesis submitted to the University of Oxford
for the degree of Doctor of Philosophy in the
Division of Mathematical and Physical Sciences

Hilary Term, 2011

Multilocus Approaches to the Detection of Disease Susceptibility Regions: Methods and Applications

Julia Grant Ciampa

Lincoln College
University of Oxford

Thesis submitted for the degree of Doctor of Philosophy at the University of Oxford
Hilary Term, 2011

Abstract

This thesis focuses on multilocus methods designed to detect single nucleotide polymorphisms (SNPs) that are associated with disease using case-control data. I study multilocus methods that allow for interaction in the regression model because epistasis is thought to be pervasive in the etiology of common human diseases. In contrast, the single-SNP models widely used in genome wide association studies (GWAS) are thought to oversimplify the underlying biology. I consider both pairwise interactions between individual SNPs and modular interactions between sets of biologically similar SNPs. Modular epistasis may be more representative of disease processes and its incorporation into regression analyses yields more parsimonious models. My methodological work focuses on strategies to increase power to detect susceptibility SNPs in the presence of genetic interaction. I emphasize the effect of gene-gene independence constraints and explore methods to relax them. I review several existing methods for interaction analyses and present their first empirical evaluation in a GWAS setting. I introduce the innovative retrospective Tukey score test (RTS) that investigates modular epistasis. Simulation studies suggest it offers a more powerful alternative to existing methods. I present diverse applications of these methods, using data from a multi-stage GWAS on prostate cancer (PRCA). My applied work is designed to generate hypotheses about the functionality of established susceptibility regions for PRCA by identifying SNPs that affect disease risk through interactions with them. Comparison of results across methods illustrates the impact of incorporating different forms of epistasis on inference about disease association. The top findings from these analyses are well supported by molecular studies. The results unite several susceptibility regions through overlapping biological pathways known to be disrupted in PRCA, motivating replication study.

Acknowledgments

I am thankful for the invaluable instruction given to me by Drs. Nilanjan Chatterjee and Chris Holmes who are my respective mentors at the National Cancer Institute and University of Oxford. I am also grateful to the NIH-Oxford-Cambridge Scholars Program for enabling me to conduct research at two renowned institutions.

Declaration

I declare that this thesis contains my own original work except where otherwise stated with reference to the literature. The research was undertaken in the Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute and in the Department of Statistics, University of Oxford.

Contents

Glossary

- ACS: American Cancer Society cancer prevention study
- ATBC: Alpha-Tocopheral, Beta-Carotene prevention study
- CGEMS: Cancer GEnetic Markers of Susceptibility
- CI: Confidence Interval
- CML(-ME/-SI): Constrained Maximum Likelihood (analysis of logistic model for Marginal Effect or Saturated Interaction)
- CTS: Composite Tukey Score test
- EB: Empirical Bayes
- EPIC: European Prospective Investigation into Cancer and nutrition
- FPCC: French Prostate cancer Case-Control study
- GWAS: Genome Wide Association Study
- HPFS: Health Professionals Follow-up Study
- LD: Linkage Disequilibrium
- MAF: Minor Allele Frequency
- OR: Odds Ratio
- PRCA: PRostate CAncer
- PTS: Prospective Tukey Score test
- Region E: sub-region of chromosomal region 8q24 associated with Epithelial cancers
- Region P: collective sub-regions of chromosomal region 8q24 associated with Prostate cancer
- RTS: Retrospective Tukey Score test

- SNP: Single Nucleotide Polymorphism
- UML(-ME/-SI): Unconstrained Maximum Likelihood (analysis of logistic model for Marginal Effect or Saturated Interaction)

Chapter 1

Introduction

Section 1.1: Road Map for Thesis

This thesis explores methods to detect disease association in the presence of genetic interaction (epistasis) using case-control data, with an emphasis on genome wide association studies (GWAS, Section 1.2). The methodological work focuses on strategies to increase power in multilocus analyses, with an emphasis on the effect of a gene-gene independence constraint. The applied work aims to generate hypotheses about the functionality of established susceptibility regions for prostate cancer (PRCA, Section 1.4) through interaction analysis of data from Cancer Genetic Markers of Susceptibility (CGEMS, Section 1.5). This introductory chapter provides succinct background information to motivate and contextualize my dissertation research. I outline the additional chapters herein.

Chapter 2 explores existing methods for modeling epistasis. First, I consider interactions between two genetic markers. I review three methods to detect pairwise epistasis: standard logistic analysis and two methods that can improve power by exploiting an assumption of gene-gene independence in the underlying population for the candidate interacting markers. I present the first empirical GWAS evaluation of these methods, comparing their strengths and weaknesses. I highlight a recently proposed empirical Bayes method for logistic analysis and discuss one strategy to improve case-only type methods. Second, I consider modular epistasis: interactions between sets of biologically similar markers. I know of only one published method for these analyses: the prospective Tukey score test (PTS). It offers a statistical advantage

in that PTS statistics have reduced degrees of freedom and a biological advantage in that modular epistasis is arguably more representative of human disease processes.

Chapters 3 and 4 introduce an innovative and powerful multilocus test: the retrospective Tukey score test (RTS). It is designed to detect disease association in the presence of modular epistasis under an assumption of gene-gene independence in the underlying population for two sets of genetic markers. Chapter 3 details the derivation of RTS and Chapter 4 presents simulations that characterize the test's size and power. I address the merits, limitations and practical implementation of RTS. I also discuss several areas for future work with RTS and the Tukey model. In Chapter 5, I focus on the composite Tukey score test, which is an extension of RTS and PTS motivated by the empirical Bayes method for pairwise interaction.

Chapter 6 presents one of the first multi-stage GWAS explorations of pairwise epistasis. The project demonstrates some methodological challenges of large-scale interaction analyses. I present an extensive list of genetic markers worthy of replication for interaction with established PRCA susceptibility regions, prioritizing top results for biological plausibility through literature reviews. The most notable finding suggests an oncogenic mechanism for the chromosomal region 8q24 that may explain its association with a variety of epithelial cancers.

Chapter 7 presents the first genome wide exploration of modular epistasis. It includes four applications that naturally follow from the pairwise interaction analyses. RTS is the focus, but I also use PTS and follow-up with more traditional methods. These analyses permit comparison of how modeling different forms of epistasis can affect inference on disease association. The projects of Chapter 6 and 7 identify promising interactions that suggest mechanisms with substantial support from

molecular studies. The results unite several susceptibility regions in PRCA and warrant replication study that my supervisors and I have begun with collaborators.

The concluding Chapter 8 provides a succinct summary of my thesis work, which has been incorporated into several manuscripts. The methodological (Section 2.1.4) and applied (Chapter 6) work on pairwise interactions that focus on CGEMS Stage II were consolidated into a paper published in *Cancer Research* [1]. One methodological finding motivated the study of an innovative multilocus method that uses principal components to relax the gene-gene independence assumption of case-only methods (Section 2.1.5). This work was published in *American Journal of Human Genetics* [2]. The top result for pairwise interaction within the 8q24 susceptibility region (Section 6.3) was included in a *Nature Genetics* article [3]. The derivation, evaluation (Chapters 3-4) and applications (Chapter 7) of RTS have been consolidated into a single paper that will be submitted for review shortly. One RTS application (Section 7.1) is the motivating example in a methodological paper on an efficient algorithm to compute p-values for non-standard test statistics (Section 4.5), published in *Biostatistics* [4]. A second RTS application involves a supplemental main effects analysis of fine-mapping data for the susceptibility region 11q13 (Section 7.2.1a) that is part of a manuscript published by *Human Molecular Genetics*. For the paper, I also investigated pairwise interactions with the top marker in 11q13, using the fine-mapping data in an analysis similar to that of Chapter 6.[5]

Section 1.2: Background on Genome Wide Association Studies

Genetic epidemiology has undergone a revolution in the last decade. Its foundation was the completion of the Human Genome Project in 2001 [6,7]. The scientific feat challenged researchers to identify the genetic causes of human disease

from vast DNA sequences. A guiding principle in the search has been the "common variant, common disease" hypothesis that states common diseases are due, at least in part, to common genetic variants (present in at least 5% of the population) [8]. Although the "common variant, common disease" hypothesis is difficult to prove empirically, theoretical calculations based on human mutation and population genetics offer support [9]. This thesis focuses on the detection of disease association in the most common form of genetic variation: the single nucleotide polymorphism (SNP), a DNA locus at which an allele can take one of two forms (major or minor) [10].

The "common variant, common disease" hypothesis shaped the International HapMap Project that was designed to provide a public database of common genetic variants in diverse populations [11]. HapMap specifically examines linkage, or patterns of correlation, among SNPs with minor allele frequencies (MAFs) of at least 5%. Linkage disequilibrium (LD) reflects the tendency of two SNPs to be inherited jointly, whereas linkage equilibrium reflects the tendency of two SNPs to be inherited independently. An accepted measure of LD is r^2 that represents the population frequency with which two alleles are observed on the same chromosome. It ranges (0, 1) with 1 representing complete LD.[12] The LD patterns published by HapMap are based on a total of 270 subjects of European, African, Chinese and Japanese ancestry. The current database includes well over 3 million SNPs from which "tag" SNPs for genetic association studies are selected. Tag SNPs exhibit high LD with multiple markers in a region, reducing the total number of SNPs for testing in exploratory studies of disease association.[11,13] A tag SNP demonstrating disease association is not expected to be the causal SNP but rather to be in LD with the casual SNP, necessitating follow-up with fine-mapping.[14] Researchers were able to implement

HapMap's extensive tagging map within a few years due to technological advances that reduced genotying costs [15,16].

GWAS were built on the pillars of the Human Genome Project, the International HapMap Project and advances in genotyping technology [14,17]. GWAS are studies that search the entire human genome in order to identify common genetic variants associated with an observable phenotype [18]. Within three years of the first GWAS (2005) [19], markers of genetic risk had replicated in more than 40 complex diseases and traits [20]. The GWAS era is one of the greatest "bursts of discovery … in the history of medical research" [21].

Scientists championed the extraordinary potential of GWAS to expand our understanding of disease etiology and physiologic genome function even before the first GWAS report [17,22]. To facilitate discovery, scientists adapted an agnostic approach to GWAS analyses, testing SNPs for disease association irrespective of prior knowledge on their biological function. This practice enables GWAS to detect genetic risk factors that may not be studied due to low prior probability of disease association. [20] The approach has proven itself well motivated, with success stories including the complement pathway and macular degeneration [19]. GWAS have also detected robust associations for SNPs in regions without genes, including the chromosomal region 5p13 with Crohn's disease [23]. A focus of this thesis is the gene-poor region 8q24 that demonstrates associations with many cancers, including the most common gender-specific cancers of colon, prostate and breast [24].

With upwards of 500K SNPs to analyze in each GWAS, false positive results are a serious concern. An accepted, though conservative, approach to minimize false positives is to enforce a strict significance threshold set by a Bonferroni correction for multiple testing $\left(p \leq 10^{-7}\right)$ [21]. For conclusive reports on disease susceptibility, the

field now requires replication of top findings in independent samples [25]. Accordingly, many GWAS follow a multi-stage design in which a proportion of top SNPs in an early stage are followed-up in a larger, independent sample in the next stage (see for example Figure 1.1). Early findings that prove robust to replication in subsequent stages are likely to represent true susceptibility SNPs [10]. A common strategy to increase power to detect susceptibility SNPs in GWAS is to form multi-site collaborations or consortia that share data [25]. The Cancer Genetic Markers of Susceptibility (CGEMS) project, central to this thesis, is a well known, large-scale, collaborative, multi-stage GWAS [26].

One expectation for GWAS was personalized medicine, which involves efficient and accurate pairing of high risk individuals with prevention strategies and of affected individuals with treatment options [27,28]. This goal has remained elusive largely because GWAS has identified susceptibility SNPs that, despite strong associations [29], are poor classifiers of high and low risk populations [30] due to modest odds ratios (OR < 2.0) [31]. More traditional risk factors such as age and family history remain the primary tools with which physicians evaluate disease risk in patients. A prime example involves the well known and widely used Breast Cancer Risk Assessment Tool [32,33] that was updated to incorporate seven susceptibility SNPs with minimal benefit [34]. Despite small odds ratios, susceptibility SNPs can elucidate etiologic mechanisms and propose novel drug targets [35]. Classic examples are *PPARG* (OR=1.25) and *KCNJ11* (OR=1.2) that code for receptors targeted by well established drugs for diabetes [36–39]. Accordingly, the focus in GWAS has shifted from advances in personalized medicine to advances in knowledge of disease etiology [40].

A great challenge in genetic research is deriving functional significance from GWAS results, irrespective of whether the susceptibility SNP is in a gene. The first step is often to search a dense map of highly correlated SNPs in a susceptibility region. These fine-mapping efforts may detect stronger evidence of disease association in a SNP with more direct functional implication, such as a risk allele that alters protein structure. More controlled experiments to determine functionality of a putative causal variant can involve characterizing genetic manipulations or quantifying gene expression in cell lines or animal models [10,35]. Although GWAS results are many steps removed from clinical application, thorough follow-up may expedite therapeutic advances. Furthermore, medical interventions that develop from common genetic risk factors could benefit a substantial portion of the population [41].

Section 1.3: Background on Interaction Analysis in Genome Wide Association Studies

Although GWAS have identified numerous susceptibility SNPs, the potential of the vast amount of data they generate is not fully realized. From both a biological and statistical perspective, one can argue single-SNP[1] analyses that dominate GWAS literature are inadequate. In this thesis, I develop, evaluate and apply multilocus methods for modeling epistasis using case-control data, with an emphasis on GWAS.

Epistasis is a term that itself carries several meanings within and across disciplines [42,43]. Throughout this thesis, interactions between sets of one or more SNPs will be discussed in two general contexts. First, SNPs will be said to interact in the statistical sense when logistic analysis detects a significant deviation from

[1] Single-SNP analyses assess the marginal effect of a SNP on a phenotype. These results are often described as main effects in the literature. In this thesis, main effects is used exclusively to describe the effect of a single SNP in a multilocus model.

multiplicative effects. Second, SNPs will be described as interacting in the biological sense when at least one of the (gene) regions they represent is known or presumed to affect the other's function. Biological interactions can take many forms, for example synergistic or antagonistic.

Complex human diseases are likely to involve the interplay of genetic and environmental risk factors [44]. Epistasis has long been known to mediate genotype-phenotype relationships [45], with the classic example being eye color in Drosophila [46]. Studies in model organisms (organisms likely to share elements of the human genetic network) suggest epistasis is pervasive in human biology [47–50]. Studies of evolutionary genetics also suggest an important role for epistasis in human biology [45]. For example, epistasis increases robustness to mutations [51]. From a biological perspective, therefore, single-SNP models may oversimplify disease models [42,52,53].

From a statistical perspective, single-SNP analyses lose power to detect susceptibility SNPs that affect disease risk primarily through epistasis [54–59]. I focus on pairwise interactions in this overview because two-SNP models are most often studied in theoretical and applied work. Some researchers advise that GWAS analyses include exhaustive pairwise interaction searches [58]. One motivation is the phenomenon in which SNPs that affect disease risk through epistasis exhibit no main effects and only minimal marginal effects [55]. While it is difficult to quantify the frequency of such interactions, animal studies provide numerous examples of genetic risk markers detected only through interaction analysis [53,60–64]. An alternative approach is to investigate epistasis only in SNPs that reach a specified significance threshold in single-SNP analyses. Although this design reduces the burdens of multiple testing and computational intensity, the filtered selection can exclude true susceptibility SNPs.[56] A more flexible alternative is INTERSNP, which selects SNPs for pairwise

interaction testing based on prior information about marginal effects, genomic location and biology [65].

The option for GWAS interaction analyses central to this thesis is the conditional scan. The general protocol is to test individual "scan" SNPs across the genome for disease association through logistic analysis of a regression model that includes main effects for the scan SNP, main effects for a "conditioning" SNP and their pairwise interaction (2.2). The conditioning SNPs are assumed to be associated with the disease, selected for strong marginal effects. When the conditioning SNP is an established susceptibility marker, the regression model controls for the variance it explains, increasing power to detect susceptibility scan SNPs with relatively weak marginal effects [17]. The conditional scan allows for direct testing of interaction effects, as well as simultaneous "omnibus" testing of main- and epistatic effects for the scan SNP.

Omnibus testing is well suited to exploratory GWAS analyses because it can detect disease association when a SNP affects risk through main- or epistatic effects. Published simulation studies characterize omnibus testing as a powerful multilocus tool. When epistasis drives disease association, omnibus testing gains power over single-SNP analyses that overlook weak marginal effects [57,66]. These power gains tend to increase as significance thresholds become more stringent, a particularly relevant finding for GWAS analyses [54]. When epistasis is absent or relatively weak, omnibus testing loses only minimal power relative to single-SNP analysis. It pays only a small penalty in the additional degree of freedom on its test statistic.[66] For weak and strong epistatic effects, omnibus testing tends to outperform interaction-only testing. This result holds whether the interaction test involves case-control or case-only data.[57] The power to detect modest epistasis is somewhat limited for both interaction

and omnibus tests because sample sizes needed to detect an interaction are at least four times greater than to detect a marginal effect of equal magnitude [67].

By detecting interactions that affect disease risk, multilocus methods have the potential to generate hypotheses about the disease process. An interaction between SNPs in genes of the same pathways, for example, may be the first evidence that the disease process involves the pathway. The potential to generate hypotheses about the function of susceptibility SNPs in regions without genes or in genes without known functions is particularly valuable. For example, an interaction between an intergenic SNP and a SNP in a well characterized gene may be the first step in determining the elusive function.

While the hypothesis-generating potential of GWAS interaction analyses is valuable [40], caution must be used when inferring biological meaning from statistical interaction [68,69]. Options for follow-up on observed statistical interactions include computer simulations to investigate biochemical systems consistent with the proposed biological interaction, as well as molecular study to characterize the biological interaction [70]. Bioinformatics tools can also be beneficial. One is STRING (Search Tool for the Retrieval of INteracting Genes/proteins), which integrates information on genomic context, expression data, high throughput experiments and existing literature to return known and predicted interactions for the candidate genes [71]. A second resource is GRAIL (Gene Relationships Across Implicated Loci), which calculates a p-value for the functional connectivity of multiple SNPs or genomic regions based on similarities between published reports for the markers [72]. A third tool is the Pathway Interaction Database that allows researchers to search for pathways that involve one or all candidate genes [73]. These tools are most useful when the SNPs under study are in genes, which is often not true for GWAS. When candidate interacting SNPs are

intergenic, traditional literature reviews can yield greater rewards. Literature reviews are also important when the SNPs are in genes because researchers can investigate whether the genes participate in similar processes or in processes that relate to the phenotype of interest. I found literature reviews to be most useful in following-up on the results of my interaction analyses.

Section 1.4: Background on Prostate Cancer and Susceptibility Regions

Prostate cancer (PRCA) carries tremendous public health concern. It is the most common cause of cancer and the second most common cause of cancer deaths in American men. In 2009 nearly 200K new cases and 30K deaths were estimated for American men due to PRCA.[74] The need for good screening options and early intervention programs is clear: 5-year survival plummets from 100% to 32% in men diagnosed with metastastic (advanced) rather than local (early) PRCA [75]. Screening practices with the biomarker PSA have improved early detection [76] and helped reduce PRCA mortality rates [77] in recent years. Other clinical end points used to assess PRCA risk are age (older), race (African ancestry) and family history (affected first degree relatives) [78].

Although PRCA pathophysiology is a vast topic, I summarize key elements pertinent to the findings of my applied work. Nearly all cases of PRCA are adenocarcinomas that originate from organ epithelium [79]. PRCA is a hormone-sensitive disease, centered on the male androgens such as testosterone. The androgen receptor signaling axis is thought to contribute to all stages of PRCA progression [80]. WNT signaling is also disrupted in PRCA [81], as well as in other cancers [82]. Its physiological functions include embryonic development and adult homeostasis [83].

An uncharacterized "reactivation" of embryonic pathways is thought to contribute to PRCA progression [84].

Evidence to suggest a genetic contribution to PRCA risk and prognosis has been accumulating for decades [85–90]. A consistent result across ethnic groups is that individuals with an affected first degree relative tend to have slightly more than twice the odds of disease than individuals without positive family histories for PRCA [91–95]. Good and poor survival rates also aggregate in families [96]. CGEMS and other PRCA GWAS have identified several susceptibility loci for PRCA [3,97–106], but they remain relatively isolated findings with little functional significance. My applied work aims to identify interactions that suggest functionality for PRCA susceptibility regions. I focus on susceptibility SNPs identified in CGEMS, selecting the top marker in each region. I emphasize SNPs near genes to ease interpretation of observed interactions, but also study gene deserts with robust associations. In total there are five gene and two intergenic regions that I briefly summarize herein (Table 1.1).

Table 1.1: **Summary of nine conditioning SNPs for interaction analyses**.

Conditioning SNP	Conditioning Region	SNP-Region Proximity	Chr	Minor Allele	MAF	OR	95% CI	P-value
rs4962416	*CTBP2*	Intron	10	C	0.25	1.18	(1.11, 1.26)	1.05e-07
rs4857841	*EEFSEC*	Intron	3	A	0.28	1.15	(1.08, 1.22)	6.79e-06
rs4430796	*HNF1B*	Intron	17	G	0.48	0.82	(0.77, 0.87)	4.56e-12
rs10486567	*JAZF1*	Intron	7	A	0.24	0.83	(0.78, 0.89)	4.89e-08
rs10993994	*MSMB*	Near	10	T	0.38	1.23	(1.16, 1.30)	7.47e-13
rs4242382	8q24, Region P	Within	8	A	0.10	1.48	(1.36, 1.61)	<1.0e-15
rs6983267	8q24, Region E	Within	8	T	0.50	0.80	(0.76, 0.84)	1.55e-15
rs620861	8q24, Region H	Within	8	T	0.37	0.90	(0.84, 0.95)	5.62e-04
rs10896449	11q13	Within	11	A	0.50	0.84	(0.80, 0.89)	8.30e-10

Reported values are based on single-SNP analyses of data from CGEMS Stages I and II.
Chr=chromosome; MAF=minor allele frequency among controls in sample; OR=odds ratio; CI=confidence interval.

The chromosomal region 8q24 contains susceptibility SNPs that demonstrate highly significant and robust associations with diverse cancers, including prostate, colon, breast and kidney. The 1.18Mb span can be divided into sub-regions that differ in cancer susceptibility profiles. I focus on three in the thesis. Region "P" (prostate) contains three sub-regions associated specifically with PRCA [24]. Region "E" (epithelium) includes the sub-region associated with several epithelial cancers, including prostate and colon [24,107]. "Region H" (hormone) includes the sub-region associated with the hormone sensitive cancers of prostate and breast [24].

Extensive effort is devoted to identifying the elusive function(s) of 8q24. The flanking genes of 8q24 are *MYC* on the telomeric end and *FAM84b* on the centromeric end. The function of *FAM84b* is not well known, but *MYC* is well characterized. It is a WNT-targeted transcription factor that participates in the regulation of 15% of all genes, affecting cell division, growth and death [108]. *MYC* has been suggested to mediate the cancer associations of 8q24, demonstrating oncogenic properties in prostate, breast and colon cancers [82]. It has been studied most extensively in terms of Region E. The risk allele for the conditioning SNP rs6983267 has only a tenuous association with *MYC* [109–111] but more recent reports describe cis-upregulation of *MYC* [112–114].

The only gene within the extended 8q24 chromosomal region is *POU5F1B* (also called *POU5F1P1*). It specifically resides in Region E, ~15Kb from the top marker rs6983267 [115,116]. Until very recently, *POU5F1B* was classified as a pseudogene of *POU5F1* (also called *OCT4* or *OCT3*) [117], a WNT-targeted gene [118] central to the regulation of stem cell pluripotency [119]. *POU5F1B* now draws more attention as a plausible candidate gene to explain the cancer associations of 8q24 [105,117,120,121]. One study detected disease association in rs871135, a intronic SNP

to *POU5F1B*, 2 base pairs from a binding site for the transcription factor NKx-2.5 [122]. Its risk allele is associated with altered binding of the transcription factor CREB [122] that participates in tumor initiation, progression and metastasis [123]. These observations are particularly noteworthy because dysregulation of CREB binding is associated with PRCA progression and prostate tumors demonstrate excessive methylation of NKx-2.5 [122].

The second chromosomal region, 11q13, is a gene desert. It is flanked by *MYEOV* on the telomeric end and *TPCN2* on the centromeric end. *TPCN2* is associated with hair color [124]. *MYEOV* is an oncogene first identified for its over-expression in myeloma [125]. Its function and regulation are not well understood.

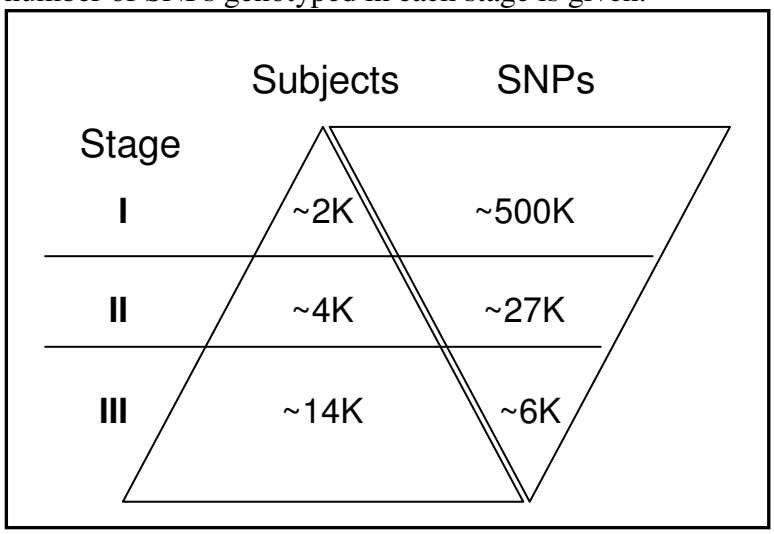
A brief description of the five gene regions studied in this thesis follows. *CTBP2* is highly expressed in prostate tissue and its expression inversely correlates with that of *PTEN*, a tumor suppressor [103]. *EEFSEC* is a translation factor that functions in protein synthesis with amino acid specificity [126]. *HNF1B* is a transcription factor that may participate in epithelial differentiation [127]. *JAZF1* is a transcriptional repressor moderately expressed in healthy prostate tissue [128]. Both *HNF1B* and *JAZF1* demonstrate association with type 2 diabetes, which itself is associated with PRCA [129]. *MSMB* encodes a protein whose loss is associated with PRCA recurrence after radical prostatectomy [103]. Its most significant risk allele, rs10993994, has been shown to increase binding affinity of CREB to *MSMB* [130,131], reminiscent of the *POU5F1B* marker.

Section 1.5: Background on Cancer Genetics Markers of Susceptibility

The National Cancer Institute initiated CGEMS in order to detect common genetic variants that affect PRCA risk, as a first step towards improved prevention and

intervention strategies [132]. CGEMS is one of the largest multi-stage GWAS. It includes several international studies, although subjects are almost exclusively of European ancestry. The general framework for CGEMS is that each successive stage includes a smaller number of SNPs genotyped in a larger sample (Figure 1.1). Detailed descriptions of the three stages have been published [3,103,104]. I present brief summaries herein.

Figure 1.1: Simplified schematic of study design for multi-stage CGEMS genome wide association study. An approximate number for total sample size in each stage is given. It represents the additional subjects who are independent of previous stages. The numbers of cases and controls were roughly equal. An approximate total for the number of SNPs genotyped in each stage is given.



The Stage I sample was derived from the Prostate, Colon, Lung and Ovarian Screening Trial. The CGEMS cohort consists of nearly 30K men age 55- to 74-years from the screening arm, which is a random assignment. An advantage of a nested case-control design is that the use of incident cases rather than prevalent cases promotes detection of risk factors that reflect risk of developing the disease rather than elements of disease progression or survival [14]. No subjects had a diagnosis of PRCA prior to

enrollment which spanned 1993-2001. Follow-up was extended to 2003 in order to enrich cases for aggressive PRCA, classified on the basis of severity at diagnosis. Controls were selected through incidence-density sampling with a target case-control ratio of one. The final dataset, after quality control checks, includes 1175 cases and 1100 controls, genotyped on 523,841 autosomal SNPs (chromosomes 1-22). Quality checks for all studies involved call rates (% SNPs assigned genotypes), Hardy-Weinberg Equilibrium in controls and genotype concordance rates with a reference dataset for SNP assays [104].

Stage II is central to this thesis. It includes four studies. Two were conducted in America. Health Professionals Follow-up Study (HPFS) is an on-going prospective cohort study of over 50K men age 40- to 75-year that began in 1986. Controls were matched on age and ethnicity, selected from men who were cancer-free at the time of case diagnosis and had undergone PSA screening. In total 596 cases and 611 controls were included in CGEMS. The American Cancer Society Cancer Prevention Study II Nutrition Cohort (ACS) follows a cohort of nearly 90K men age 40- to 92-year that was formed in 1992. Controls were matched on age, ethnicity and DNA sample in terms of collection data and specimen type. They were selected from men who were cancer free (except for melanoma) at the start of the observational interval that preceded case diagnosis. Over-sampling of aggressive cases occurred among those whose DNA testing involved buccal samples. In total, 1760 cases and 1775 controls were included in CGEMS. The third study was based in Finland: Alpha-Tocopheral, Beta-Carotene Cancer Prevention Study (ATBC). It began in 1985 as a randomized controlled trial of almost 30K male smokers age 50- to 69-year. Its objective was to investigate the preventative effects of Vitamin E and beta-carotene on cancer incidence. It continued as a longitudinal cohort study. After a 16-year follow-up, the nested case-control

sample for CGEMS was formed with matching based on age, intervention assignment and blood draw date. In total there were 929 cases and 921 controls. The fourth study is the French PRCA Case-Control Study (FPCC) that began in 1994. It recruited cases from three hospitals and matched controls on geography and age at enrollment. In total there are 656 cases and 657 controls. The entire Stage II includes 3941 cases and 3964 controls genotyped on 27,383 autosomal SNPs with evidence of disease association in the independent Stage I sample (marginal $p<0.05$). The single-SNP analyses for selection differ from those I present in this thesis. My analysis involves a dichotomous disease outcome rather than a trichotomous outcome that differentiates aggressive and non-aggressive cases. Also, my analysis codes SNP data based on allele counts, assuming the alleles at each locus affect disease risk in an additive fashion on the logistic scale. The alternative used in past CGEMS publications is to make no assumption on the structure of genetic risk and code SNP data as indicators for observed genotypes [103]. Advantages to the allele count approach is that it retains flexibility, while increasing power to detect association due to the reduced degrees of freedom on resultant test statistics [54].

Stage III includes an additional five studies. The Cancer Prostate in Sweden Study is a large population-based case-control study. Cases were recruited through regional cancer registries and controls were randomly recruited concurrently, matched on geography and expected age distribution of cases. In total there were 2314 cases and 1362 controls. The Multiethnic Cohort Study is a population-based prospective cohort study of men younger than 80-year with enrollment between 1993-1996 in Hawaii and California. The nested case-control sample of CGEMS includes incident PRCA cases and randomly sampled controls. In total there were 676 case and 682 controls. The European Prospective Investigation into Cancer and Nutrition (EPIC) is

an on-going prospective study of over 150K men recruited in 1992-2000, with regional centers in ten European countries. Cases and controls were matched on study center and blood draw details (age, fasting-status, time of day). In total there were 682 cases and 990 and controls. The Johns Hopkins University Study recruited men older than 55-year from hospitals, with cases based on radical prostectomy surgical cases and controls based on screening. In total there were 990 cases and 451 controls. The Cohort of Norway Study is a collaboration between six population-based cohorts. The incident cases were ascertained through cancer registries. Controls were matched on age and study cohort. In total there were 606 cases and 662 controls. On the total 10,272 cases and 9,123 controls in the combined Stages I-III, 5,796 autosomal SNPs were genotyped. Just over half of the SNPs were selected based on single-SNP analyses in Stage II ($p<1.0e-3$). Fine-mapping was performed for top regions with varying degrees of density in those maps based on the strength of observed associations. 8q24 was most densely mapped.[3]

Chapter 2

Methods for Modeling Epistasis

This chapter examines a variety of methods available to model epistasis using case-control data. First, I consider pairwise interactions. I review standard logistic regression analysis and two alternatives that can gain power by exploiting a valid assumption of gene-gene independence in the underlying population for the candidate interacting markers. I present the first empirical assessment of these methods in the setting of a genome wide association study (GWAS), discussing their strengths and weaknesses. The final sections of this chapter focus on modular epistasis between sets of single nucleotide polymorphisms (SNPs). In them, I review the Tukey model and prospective Tukey score test (PTS).

Section 2.1: Logistic Model and Pairwise Interactions

Case-control studies are used widely in epidemiology. They can be much more efficient than cohort studies with respect to time and money, especially for rare diseases such as prostate cancer (PRCA). Case-control studies collect data retrospectively, having already observed the outcome, whereas cohort studies collect data prospectively. Due to this difference in sampling design, case-control studies estimate odds ratios (ORs) and cohort studies estimate relative risks. An odds ratio is the odds of disease among subjects with an exposure relative to the odds of disease among unexposed subjects.[133] In this setting, the exposure is a variant allele. More generally, the odds ratio for binary variables X and Y in a generic 2x2 contingency table is:

$$OR_{XY} = \frac{\dfrac{P(Y=1 \mid X=1)}{P(Y=0 \mid X=1)}}{\dfrac{P(Y=1 \mid X=0)}{P(Y=0 \mid X=0)}} \tag{2.1}$$

This thesis focuses on the use of case-control data to estimate odds ratios through logistic regression analysis, with an emphasis on multiplicative interactions. A representative two-SNP model follows:

$$\text{logit}[P(D=1 \mid S, C, A)] = \beta_0 + \beta_1 S + \beta_2 C + \beta_3 SC + \sum_{w=1}^{W} \beta_{4w} A_w \tag{2.2}$$

These variable designations hold throughout the thesis:

D is the disease status for cases (1) and controls (0);

S is the minor allele count $(0,1,2)$ for the "scan" SNP being tested for disease association;

C is the minor allele count for the "conditioning" SNP assumed to be associated with the disease;

$\mathbf{A} = \{A_1, ..., A_W\}$ is the set of adjusting covariates; this vector notation holds throughout the thesis.

Before I explore this model, I review some notational conventions. First, I present summations (or products) over an index from its minimum to maximum value by notating only the index such that $\sum_{c=0}^{2} c$ is equivalent to $\sum_{c} c$ and I present summations taken over multiple indices using a single sigma with multiple subscripts such that $\sum_{s}\sum_{c} sc$ is equivalent to $\sum_{sc} sc$. Second, I present the linear combination that defines a logistic model as a function of covariates and parameters, excluding the

intercept. In this setting, the function $m(s,c,\mathbf{a};\boldsymbol{\beta})$ is the right hand side of (2.2) minus the intercept. Third, I use subscripts on regression parameters and odds ratios to indicate their covariate(s) (for example, $\beta_3 = \beta_{SC}$ and $\exp\{\beta_3\} = OR_{SC}$).

The regression model (2.2) can be modified in many ways to allow for more complex regressions, but the projects of Chapters 2 and 4 utilize this form to investigate pairwise interactions. This chapter focuses on methods to estimate β_{SC} which captures any non-multiplicative effects of S and C on the logistic scale. If one assumes binary genetic variables and no adjusting covariates, the data can be summarized through Table 2.1, which will be referenced throughout this chapter. In this setting,

$$\exp(\beta_{SC}) = OR_{SC} = \frac{OR_{D\text{-}SC}}{OR_{D\text{-}S,C=0}OR_{D\text{-}C,S=0}} \tag{2.3}$$

where $OR_{D\text{-}SC}$ is the odds ratio of disease associated with S and C among subjects with level one responses for both covariates relative to subjects with level zero responses for both covariates; $OR_{D\text{-}S,C=0}$ is the odds ratio of disease associated with S among subjects for whom $C = 0$ and analogously for $OR_{D\text{-}C,S=0}$ [134].

Table 2.1: Contingency table for Model (2.2) with binary disease outcome (D) and genetic variables (S, C) and no adjusting covariates.

	C=0		C=1		
	S=0	S=1	S=0	S=1	Total
D=0	*n_{000}	n_{001}	n_{010}	n_{011}	n_{0++}
D=1	n_{100}	n_{101}	n_{110}	n_{111}	n_{1++}

* n_{dsc} = number subjects with $D = d, S = s, C = c$ for $d = 0,1;\ s = 0,1;\ c = 0,1$.

Section 2.1.1: Unconstrained Maximum Likelihood Logistic Analysis

Traditionally, case-control studies are analyzed as though the data were collected prospectively. I refer to this "standard" method as the unconstrained maximum likelihood (UML) logistic analysis. Its prospective likelihood is:

$$P(D \mid S, C, \mathbf{A}) = \prod_i \pi\left(s_i, c_i, \mathbf{a}_i\right)^{d_i} \left(1 - \pi\left(s_i, c_i, \mathbf{a}_i\right)\right)^{1-d_i} \tag{2.4}$$

$$\text{with} \quad \pi(s_i, c_i, \mathbf{a}_i) = \frac{\exp\left(\hat{\beta}_0 + m\left(s_i, c_i, \mathbf{a}_i; \hat{\boldsymbol{\beta}}\right)\right)}{1 + \exp\left(\hat{\beta}_0 + m\left(s_i, c_i, \mathbf{a}_i; \hat{\boldsymbol{\beta}}\right)\right)} \tag{2.5}$$

the probability of being a case.

In the above equations and throughout the thesis:

i subscript indicates individual subjects, $i = 1, ..., n$;

n is the total number of subjects;

"^" indicates an estimated value.

The likelihood (2.4) imposes no constraints on the distribution of covariates. Prentice and Pyke demonstrated that, when this distribution is unspecified, a prospective analysis of case-control data yields valid point and variance estimates for the regression parameters, despite the retrospective study design [135]. This important result stems from the identity $P(X \mid Y) = \frac{P(Y \mid X) P(X)}{P(Y)}$ that allows one to estimate a generic odds ratio (2.1) using either $P(X \mid Y)$ or $P(Y \mid X)$ [136]. UML analyses are robust in that results are valid irrespective of the covariates' distribution. In the context of Table 2.1:

$$\hat{\beta}_{SC}^{UML} = \ln\left(\frac{n_{111} * n_{010} * n_{001} * n_{100}}{n_{001} * n_{110} * n_{000} * n_{101}}\right) \tag{2.6}$$

Section 2.1.2: Constrained Maximum Likelihood Logistic Analysis

Although UML is widely used in epidemiology, methods that impose a valid constraint of independence in the underlying population for the candidate interacting variables are well known to improve power to detect interactions. The classic example is the case-only design that estimates interaction odds ratios using data from only cases [137]. The literature discusses it in terms of gene-environment interactions most often but the method naturally extends to gene-gene interactions for unlinked loci [138]. A case-only analysis gives valid estimates of interaction odds ratio when the assumption of gene-gene independence holds. The result stems from (2.3) which gives the following identity with some algebra:

$$OR_{SC} = \frac{OR_{S-C,ca}}{OR_{S-C,co}} \tag{2.7}$$

where OR_{S-C} is the odds ratio of a level one S response associated with C among either cases (ca) or controls (co).

The assumptions of gene-gene independence and rare disease set the denominator in (2.7) to unity. In the context of Table 2.1:

$$\hat{\beta}_{SC}^{CML} = \ln\left(\frac{n_{111} * n_{100}}{n_{110} * n_{101}}\right) \tag{2.8}$$

This estimate is more precise than $\hat{\beta}_{SC}^{UML}$ (2.6) because its variance is the sum of inverse cell counts for only cases $\left(\sum_{sc} \frac{1}{n_{1sc}}\right)$ rather than for all subjects $\left(\sum_{dsc} \frac{1}{n_{dsc}}\right)$ [137].

Subsequently, the case-only method gained favor among epidemiologists [139] and the cancer research literature contains numerous applications [140–150].

The efficiency gains have trade-offs in bias. One can see from (2.7) that $\hat{OR}_{sc}^{CML}$ is biased by a magnitude of $\left(OR_{S\text{-}C,co}\right)^{-1}$ when model assumptions are invalid. The sensitivity of case-only methods to violations of gene-gene independence can substantially inflate type I error. A two-stage alternative can diminish but not eliminate this bias. The method calls for researchers to test for gene-gene independence among controls in order to determine whether a case-only or UML analysis is more appropriate.[151] The residual bias is thought to be due to the model selection uncertainty that hinders variance estimation for interaction estimates rather than to low power to detect dependence in small samples of controls [59,151].

Chatterjee and Carroll introduced the constrained maximum likelihood (CML) logistic analysis that is central to this thesis. It is a case-only type method, meaning it assumes independence in the underlying population for candidate interacting variables. One advantage of CML over the case-only method is that it estimates main- and epistatic effects, using data from cases and controls. Log-linear analysis can also estimate both values under an assumption of gene-gene independence, but the logistic model of CML is more flexible than a log-linear model. For example, it can incorporate continuous covariates.[152]

CML uses a retrospective likelihood to analyze a logistic model. Retrospective analyses are a general technique to improve efficiency by imposing valid constraints on the covariates' distribution. In this setting, the constraint is independence between the scan and conditioning SNPs in the underlying population, which the rare disease assumption allows the study controls to represent. The specific likelihood for model (2.2) without adjusting covariates is:

$$P(S,C \mid D) = \frac{P(D \mid S,C)P(S)P(C)}{P(D)} \tag{2.9}$$

In the numerator, the gene-gene independence constraint sets $P(S,C)$ to $P(S)P(C)$. This constraint can be relaxed through stratification, a procedure that assumes gene-gene independence holds only within subsets of subjects. Factors that can influence linkage disequilibrium, such as ethnicity and geography, define these strata. When they are the adjusting covariates $(\mathbf{A})$, the retrospective CML likelihood is

$$P(S,C,\mathbf{A} \mid D) = \frac{P(D \mid S,C,\mathbf{A})P(S \mid \mathbf{A})P(C,\mathbf{A})}{P(D)}$$

for a parametric $P(S \mid \mathbf{A})$ and any $P(C,\mathbf{A})$.[152]

Chatterjee and Carroll compared CML and UML through simulations that honored the gene-gene independence assumption.[2] They evaluated performance through mean squared error, which is the sum of the variance and squared bias of a parameter estimate. They found that UML and CML produce unbiased parameter estimates and that the efficiency of CML is markedly increased for $\hat{\beta}_{SC}$, slightly increased for $\hat{\beta}_S$ and comparable for $\hat{\beta}_C$. In a second set of simulations, Chatterjee and Carroll simulated the genetic data under gene-gene independence conditional on an additional risk factor (A), using the regression function $m(S,C,A;\boldsymbol{\beta})$ that included additional terms for the main effects of A and an interaction between A and S. When the analysis erroneously assumed independence on a population level through (2.9), bias was substantial for $\hat{\beta}_S$ and slight for $\hat{\beta}_C$ and $\hat{\beta}_{SC}$. In contrast, efficiency gains were substantial for $\hat{\beta}_{SC}$ and modest for $\hat{\beta}_S$ and $\hat{\beta}_C$ in analyses that assumed conditional independence.[152]

[2] Chatterjee and Carroll investigated gene-environment interactions. For convenience, I treat their genetic variable as a scan SNP and their environmental variable as a conditioning SNP.

These results correspond to those for log-linear analyses that impose independence in the underlying population for candidate interacting variables. Specifically, Umbach and Weinberg compared the performance of constrained and unconstrained analyses on the basis of asymptotic relative efficiency (ARE), which reflects the necessary sample size to achieve the same power level for a given significance level. Although efficiency gains were substantial for interaction estimates (ARE > 2), they were modest for main effects (ARE < 1.2).[153]

Section 2.1.3: Empirical Bayes Logistic Analysis

The empirical Bayes (EB) logistic analysis is a compromise between UML and CML. It relaxes the gene-gene independence assumption of CML in a data-adaptive fashion through a nuisance parameter that captures uncertainty about the assumption: $\theta_{SC,co} = \ln\left(OR_{S-C,co}\right)$. For comparison, UML allows $\theta_{SC,co}$ to be unspecified; CML sets it to zero; and the two-stage method explicitly tests $\theta_{SC,co} = 0$ in controls before selecting UML or CML for analysis. In the EB analysis, $\theta_{SC,co} \sim N\left(0, \tau^2\right)$, where τ^2 reflects uncertainty about gene-gene independence. A conservative choice for τ^2 is $\theta^2_{SC,co}$.[134] It can be interpreted as the squared bias of $\hat{\beta}^{CML}_{SC}$, since

$$\beta^{CML}_{SC} = \beta^{UML}_{SC} + \theta_{SC,co} \tag{2.10}$$

By setting τ^2 to $\theta^2_{SC,co}$, this EB analysis resembles a semi-Bayes method studied by Greenland. The semi-Bayes method was more efficient than UML when τ^2 was conservative in simulation studies of both small- and large-scale in terms of sample size and covariate dimension [154]. The EB method constructs an estimate for β_{SC} using estimates for β^{UML}_{SC} and its variance σ^2_{UML} and estimates for β^{CML}_{SC} and its bias $\theta_{SC,co}$ via:

$$\hat{\beta}_{SC}^{EB} = \hat{\beta}_{SC}^{CML}\left(\frac{\hat{\sigma}_{UML}^{2}}{\hat{\theta}_{SC,Co}^{2} + \hat{\sigma}_{UML}^{2}}\right) + \hat{\beta}_{SC}^{UML}\left(\frac{\hat{\theta}_{SC,co}^{2}}{\hat{\theta}_{SC,Co}^{2} + \hat{\sigma}_{UML}^{2}}\right) \quad (2.11)$$

The weights in are intuitive: $\hat{\beta}_{SC}^{EB} \to \hat{\beta}_{3}^{CML}$ as evidence for gene-gene independence increases and $\hat{\theta}_{SC,Co}^{2} \to 0$, whereas $\hat{\beta}_{SC}^{EB} \to \hat{\beta}_{SC}^{UML}$ as evidence against independence increases and $\hat{\theta}_{SC,Co}^{2} \to \infty$.[134]

Mukherjee and Chatterjee compared UML, case-only, two-stage and EB analyses on the basis of mean squared error. The first set of simulations evaluated performance under varying degrees of dependence between the scan and conditioning SNPs. Under gene-gene independence, the case-only method was preferable; EB and two-stage analyses were slightly less efficient, and UML was substantially less efficient. As gene-gene dependence increased, CML became least efficient; EB and UML performed comparably, and the two-stage method was less efficient than both. In a second set of simulations, Mukherjee and Chatterjee enforced gene-gene independence at a population level but allowed for variation in the dependence of S and C across strata. EB maintained its efficiency gains over UML consistently and over CML particularly in larger samples. These findings are relevant to GWAS that often involve subjects recruited in different countries.[134]

Mukherjee and Chatterjee also ran simulations for three-SNP models that demonstrate a unique feature of the EB method. The model included a bivariate $\mathbf{S} = \{S_1, S_2\}$ and its pairwise interactions with C. With respect to C, S_1 was simulated under gene-gene independence but S_2 was simulated under dependence. The results highlight the flexibility of EB in that $\hat{\beta}_{S_1C}$ was weighted towards the CML estimate and

$\hat{\beta}_{S_2C}$ towards the UML estimate. This flexibility is advantageous because interaction analyses may involve SNP pairs in varying degree of dependence, making a uniform analysis by CML or UML less preferable. This feature is particularly relevant for large-scale interaction studies.[134]

Mukherjee et al. performed additionally simulations to evaluate EB that reinforce and extend previous findings. I first review the results for type I error. The case-only method tightly controlled type I error when gene-gene independence held, but it had the greatest rates of false positives when the constraint was invalid. The two-stage method consistently demonstrated substantial inflation, but its performance improved as $\theta_{SC,co}$ and total sample size increased. This behavior may reflect an increased power to detect dependence among controls in these setting, suggesting a role for large external databases to assess gene-gene independence in the first stage of the method. The results suggest the two-stage and case-only methods are not optimal for large-scale interaction analyses that include SNP pairs in modest linkage disequilibrium that may go undetected in controls. In contrast, UML consistently controlled type I error and EB demonstrated comparable levels. Specifically, EB had empirical alpha levels in the range (0.04, 0.10) for nominal alpha 0.05. EB maintained tightest control of type I error in settings of no $\left(OR_{SC,co}=1.0\right)$, weak $\left(OR_{SC,co}=1.1\right)$ or strong $\left(OR_{SC,co}=2\right)$ dependence between the scan and conditioning SNPs. This observation suggests ambiguous levels of dependence $\left(OR_{SC,co}=1.2,1.5\right)$ hinder the EB weight scheme (2.11). With respect to power, EB consistently outperformed UML and lost only minimal power relative to CML when independence held.[59] These results characterize EB as a robust and powerful alternative to CML and UML for interaction analyses.

Section 2.1.4: Empirical Evaluation of Three Methods to Detect Pairwise Interactions

These promising simulation results motivate my empirical evaluation of UML, CML and EB in a GWAS setting. I analyzed data from Stage II of Cancer Genetic Markers of Susceptibility (CGEMS, Section 1.5) through a series of conditional genome scans, using the CaseControl.Genetics R package [155]. The data selection is a compromise between a large number of SNPs (Stage I) or subjects (Stage III). The regression models include the main effects of one scan SNP and one conditioning SNP, their pairwise interaction and covariates for study that were also used for stratification in CML (2.2). The conditioning SNPs represent nine susceptibility regions in PRCA (Table 1.1, Section 1.4). The scan SNPs were all CGEMS SNPs at least 500Kb from the conditioning SNP because CML is sensitive to violations of gene-gene independence and an accepted practice is to assume SNPs that distant in the genome are free of linkage disequilibrium. I assess the interaction parameter of these models (β_{SC}) through a Wald test, which is standard statistical test that measures the deviation of a maximum likelihood estimate from its null value [156]. This evaluation excludes omnibus tests because the three methods estimate main effects similarly.

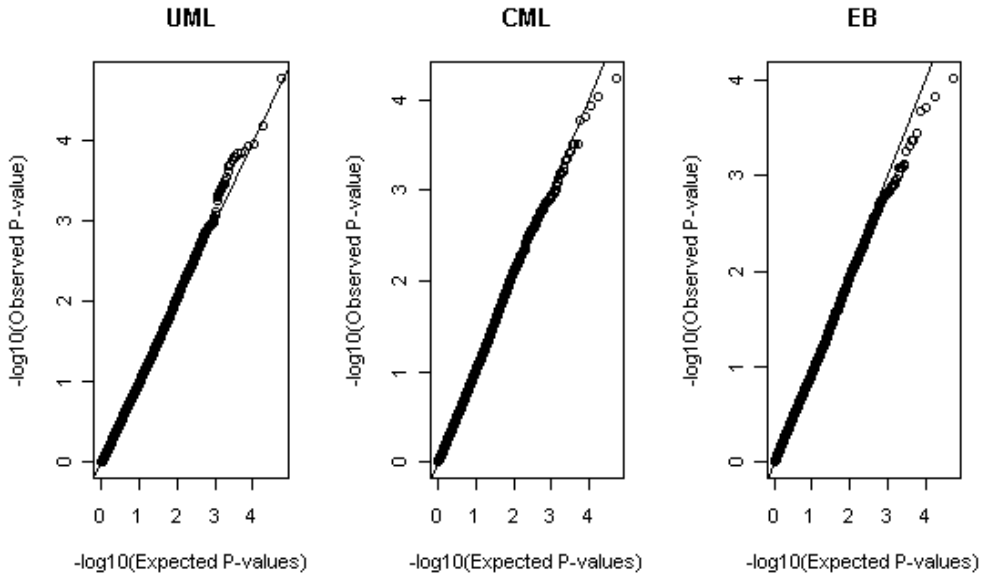
Quantile-quantile plots of interaction p-values suggest that, in general, no method is affected by large-scale systematic bias (for example, Figure 2.1a). Neither the UML nor the EB method shows evidence of bias. The robustness of the UML method has long been established, while the robustness of the EB method is credited to the data-adaptive way it incorporates gene-gene independence. In many scans, the CML method does not demonstrate bias. This finding suggests gene-gene independence can be a valid assumption in large-scale studies of pairwise interaction, provided researchers minimize bias due to linkage disequilibrium by excluding SNP

pairs in close proximity (≤500Kb). However, CML does demonstrate an excess of statistically significant associations than expected by chance in some scans (for example, Figure 2.1b). This excess is unlikely to reflect increased power to detect interactions because most GWAS markers are expected to be null. Instead, it may reflect an inflated type I error due to violations of gene-gene independence. The excess of statistically significant associations only diminishes when analyses exclude scan SNPs on the same chromosome as the conditioning SNP (for example, Figure 2.2). This observation suggests substructure in the data induces long-range dependence, possibly due to population stratification. The abundance of SNP pairs that demonstrate long-range dependence is unknown but their existence motivates methodological work to improve case-only type methods.

Figure 2.1: **Quantile-quantile plots for interaction p-values from two genome scans.** P-values were computed through Wald tests on the estimate of a multiplicative interaction between the susceptibility SNP and each of ~27K scan SNPs in CGEMS Stage II. Estimates were computed via three methods: unconstrained maximum likelihood (left), constrained maximum likelihood (middle) and an empirical Bayes method (right). Plots exclude SNPs within 500Kb of the conditioning. The conditioning region was 8q24 Region E (top) or *MSMB* (bottom). P-values less than 1.0e-6 were treated as 1.0e-6.

A)

8q24 Region E Conditional Scan



B)

***MSMB* Conditional Scan**

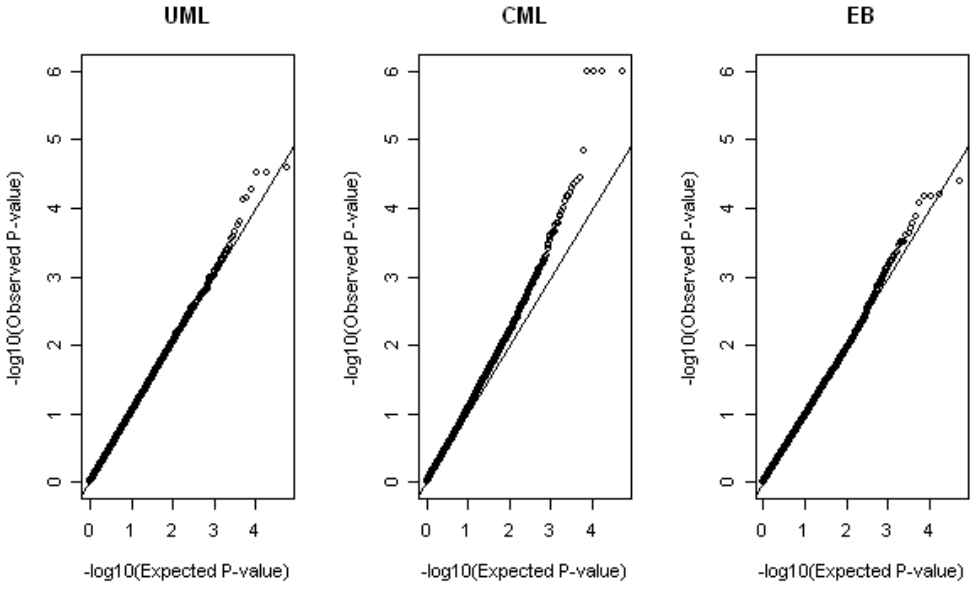
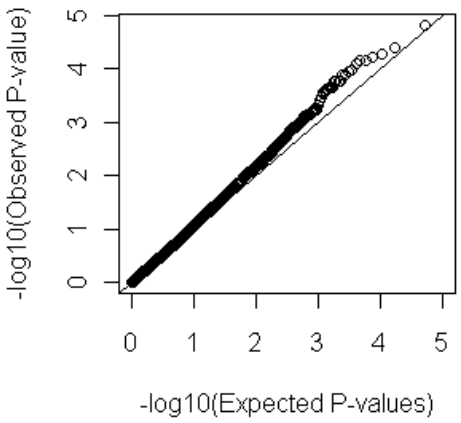


Figure 2.2: **Quantile-quantile plot for interaction p-values from the constrained maximum likelihood logistic analysis of genome scan conditional on one susceptibility SNP near *MSMB*.** P-values were computed through Wald tests on the estimate of a multiplicative interaction between a susceptibility locus near *MSMB* and each of 27,053 scan SNPs in CGEMS Stage II. Plots exclude SNPs on the same chromosome as *MSMB* to minimize violations of gene-gene independence assumption.

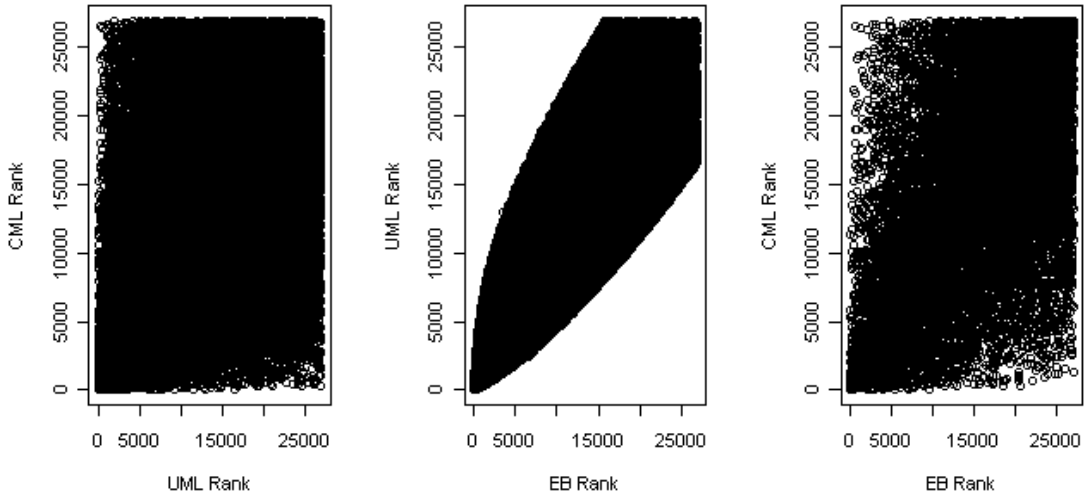


I examine consistency of results in each conditional scan by comparing the rank of scan SNPs on interaction p-value across methods. I present representative examples for scans in which CML does and does not demonstrate bias (Figure 2.3). The plots suggest CML and UML analyses yield discordant results, with essentially no visual indication of consistent rankings. The discordance is slightly less extreme for CML and EB analyses, and the UML and EB rankings are fairly concordant. These results suggest the data-adaptive gene-gene independence constraint of the EB method modifies UML rankings, whereas the rigid gene-gene independence constraint of the CML method disregards UML rankings. I cannot determine if the power advantages of CML and EB enabled either to detect true susceptibility SNPs that UML missed because this evaluation is empirical. However, the biological plausibility of top EB SNPs is promising (Section 6.4).

Figure 2.3: **Plots for ranks based on interaction p-values from two conditional genome scans.** P-values were computed through Wald tests on the estimate of a multiplicative interaction between the susceptibility SNP and each of ~27K scan SNPs in CGEMS Stage II. Estimates were computed via three methods: unconstrained maximum likelihood (UML), constrained maximum likelihood (CML) and empirical Bayes (EB). Plots exclude SNPs within 500Kb of the conditioning SNP. The conditioning region was 8q24 Region E (top) or *MSMB* (bottom).

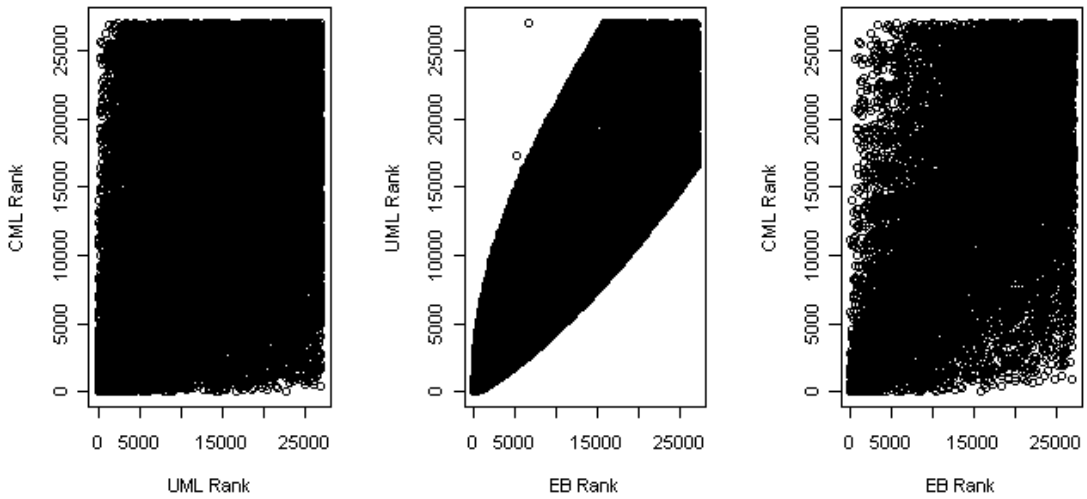
A)

8q24 Region E Conditional Scan



B)

MSMB Conditional Scan



Section 2.1.5: Follow-up on Constrained Maximum Likelihood Logistic Analysis

I collaborated with a colleague, Dr. Samsiddhi Bhattacharjee, to study the hypothesis that population stratification can induce long-range dependence in the genome, inflating type I error in large-scale studies of epistasis that use case-only type methods. Population stratification is a type of confounding unique to genetic epidemiology. It is discussed most commonly in terms of single-SNP analyses, which I review first. Population stratification occurs when the frequency of a genetic variant varies between cases and controls due to systematic differences in their ethnicity. It biases results because diseases with higher prevalence in one ethnic group will show association with SNPs whose minor allele frequencies are also more common in the ethnic group. Consequently, SNPs that reflect ancestral differences between cases and controls can be erroneously classified as susceptibility SNPs. A didactic example is a study of chopstick ability among residents of San Francisco, CA, USA. One may find an association with HLA-A1 because its SNP has a high minor allele frequency in subjects of Asian ancestry who are expected to be more adept with chopsticks.[157]

Several methods have been proposed to control for population stratification in genetic epidemiology. An early method is Genomic Control that adjusts test statistics through a uniform inflation factor that reflects deviation from an expected null distribution [158]. The main disadvantage of this method is that uniform adjustment may be too extreme in either direction for numerous SNPs [159]. Several methods exploiting principal components analysis (PCA) have been introduced more recently and are used more commonly. PCA is a statistical technique to project multi-dimensional data onto an orthogonal state space, using independent axes known as principal components that account for decreasing amounts of variation in the data [160].

In genetic association studies, PCA is performed on a large set (~15K) of reference SNPs not associated with the disease in order to cluster subjects by similar ancestry [159,161]. An excess of variation in the first principal component suggests the sample has substructure that should be accounted for in tests of disease association [2]. To adjust for population stratification the eigenvectors corresponding to the top principal components are included in the regression model [159,161]. An alternative for international studies is to incorporate indicator variables for study location in the regression model because PCA clusters often have geographical interpretations [159,162]. This adjustment also addresses non-genetic differences that may exist across centers. I use this approach in the CGEMS analyses.

Despite cautionary tails of bias [163–166], some researchers argue population stratification does not substantially affect well designed studies [162,167]. Theoretical calculations demonstrate that matching on population stratification variables (i.e. race, nationality and ancestry) in case-control studies reduces bias in parameter estimates at least partially and perfect matching completely eliminates the bias [168]. Of note, an extensive study of CGEMS determined it was free of substantial bias due to the well designed sampling strategy [162].

In the context of epistasis, population stratification takes a different form. I use the term "epistatic" population stratification to refer to the phenomenon in which the allele frequencies of candidate interacting SNPs covary within ethnic groups. Unlike population stratification that biases marginal effects, this confounding does not involve disease rates. If it is ignored in case-only analyses, interaction estimates can suffer substantial bias because the correlation can be mistaken as evidence of an epistatic effect on disease risk [169]. This sensitivity highlights the value of relaxing the gene-gene independence assumption to hold only within strata of subjects, as CML can.

The hypothesis of Bhattacharjee et al. is that imposing conditional gene-gene independence within increasingly small and homogenous subsets of subjects will reduce the observed bias of CML by decreasing residual epistatic population stratification. Bhattacharjee et al. consider several matching strategies for case-control and case-only designs that use PCA on a large panel of reference population stratification SNPs (12K and 7K) [2]. Bhattacharjee et al. investigate case-control and nearest-neighbor matching strategies, using an algorithm that examines genetic distance between subjects similar to that of Luca et al.[170]. For case-control matching, Bhattacharjee et al. consider both unconstrained (UMLC-CC) and constrained (CMLC-CC) conditional logistic analyses that differ in the gene-gene independence constraint. For nearest-neighbor matching that ignores disease status, Bhattacharjee et al. consider a constrained likelihood that incorporates the joint distribution of the disease and scan SNP prior to matching (HML-NN). A unique feature of this likelihood is the parameter for population-specific baseline disease risk. When the parameter is properly specified, the likelihood can increase efficiency by conditioning on the number of cases in each matched set, but there is a risk of confounding when the parameter is misspecified. Bhattacharjee et al. compare these methods to a variety of unconditional logistic analyses: UML, case-only and case-only with principal component eigenvectors as adjusting covariates (CO-PC).[2] CO-PC is motivated by previous work on multivariate models that reduce bias in case-only analyses by expanding the regression models to include variables that promote dependence between scan and conditioning SNPs [171].

The simulations include three scenarios: a) no population stratification or gene-gene dependence, b) gene-gene dependence and c) both gene-gene dependence and epistatic population stratification. Under a valid assumption of gene-gene

independence, all methods control type I error but constrained methods are more powerful. The case-only method demonstrates greatest asymptotic relative efficiency, UMLC-CC the least. When the gene-gene independence assumption is violated, type I error is unacceptably high for only CO and elevated for only CO-ADJ. The rank order of alternative methods on empirical power is consistent whether population stratification is present or absent. CMLC-CC and HML-NN have comparably high power; CO-ADJ is intermediate and CO-PC and UMLC-CC have comparably low power. Bhattacharjee et al. also present an empirical GWAS evaluation, using a conditioning SNP affected by population stratification in a conditional scan for pairwise interaction. Bhattacharjee et al. report genomic control inflation factors (IF) based on interaction p-values for ~500K SNPs on different chromosomes than the conditioning SNP. The results suggest only the case-only method is vulnerable to pervasive inflation of type I error in a GWAS setting. Its inflation factor is 1.32, whereas all others range (0.99, 1.02).[2] This project motivates an area for future work on the EB method. Specifically, one could investigate the performance of EB when its estimate for interaction incorporates an estimate from either HCL-NN or CCML-CC rather than CML.

Section 2.2: Modular Epistasis and the Tukey Model

The concept of modular biology was described more than a decade ago. Its building blocks are function modules, discrete units of diverse molecules. Each module accomplishes a relatively autonomous function and higher-level functions result from connections between them.[172] Both healthy and disease states involve interconnected biological processes [45]. Modular biology is consistent with evolutionary genetics because functional modules can be robust to internal change and

yet promote diversity through changes in intermodular connections [51,172]. Modular biology has clear implications for genetic epidemiology, suggesting epistasis should be investigated on a systems-level between functional modules rather than between single genetic markers.

Two studies in yeast (a model organism) motivate study of modular epistasis in humans. Both found that biological interactions occurred more frequently than expected by chance between genes that either had similar functions or encoded proteins with similar properties [173,174]. The focused study of yeast metabolism also found that genes with similar functions could be grouped into modules with consistent patterns of biological interactions with genes in other modules [174]. These results suggest that evidence of statistical interaction may be useful in assigning putative function to uncharacterized genes, consistent with my objective of generating hypotheses about the biological relevance of susceptibility regions through interaction analysis. These results further suggest that biological interactions between sets of functionally similar genes (modular epistasis) may affect risk of human disease.

A disease model involving modular epistasis is consistent with the Tukey model that Chatterjee et al. introduced [175]. The regression allows for interaction between sets of biologically similar SNPs. Subsequently, one could argue it is capable of modeling the underlying biology of human disease more closely than standard multilocus logistic models. The general form of the Tukey model is:

$$\text{logit}[P(D=1|\mathbf{S},\mathbf{C},\mathbf{A})]=\beta_0+\sum_{m=1}^{M}\beta_{1m}S_m+\sum_{z=1}^{Z}\beta_{2z}C_z+\theta\left(\sum_{m=1}^{M}\beta_{1m}S_m\sum_{z=1}^{Z}\beta_{2z}C_z\right)+\sum_{w=1}^{W}\beta_{3w}A_w \qquad (2.12)$$

where variables D, $\mathbf{S}$, $\mathbf{C}$ and $\mathbf{A}$ retain their meanings from previous sections (see Section 2.1.1).

The Tukey model differs from traditional logistic models (Section 2.1) through the interaction term. A single parameter (θ) captures the interaction of two SNP sets rather than individual parameters for each pairwise interaction between scan and conditioning SNPs.

The conceptual framework for this parsimonious regression model is detailed elsewhere [175]. The model lends itself to the Tukey 1-degree of freedom test for interaction [176] because it involves an assumption that each set of SNPs represents a single biological mechanism that affects disease risk. When this mechanism is unknown or unobserved, it can be modeled as a latent variable (V) using data from the corresponding SNP set. Chatterjee et al. define:

$$V_{\mathbf{C}} = \kappa_0 + \sum_z \kappa_z C_Z + \varepsilon_C \tag{2.13}$$

for the conditioning SNP set $(\mathbf{C})$ and $V_{\mathbf{S}}$ analogously for the scan SNP set $(\mathbf{S})$. They demonstrate that inference on epistasis is valid when the SNP sets substitute for the latent variables in the regression, using a Taylor series expansion around the error terms in the model for $V_{\mathbf{S}}$ and $V_{\mathbf{C}}$ (2.13). The latent variable framework allows one to interpret the interaction between the SNP sets as an interaction between two latent variables, making a Tukey 1-degree of freedom test for interaction [176] intuitive.[175] The latent variable framework may be most natural for SNPs that are similar in terms of genomic location or biological function. They include SNPs in the same chromosomal region or in genes of the same pathway. A common mechanism for SNPs in the same gene could be reduced activity of the gene's protein. Researchers must construct the SNP sets thoughtfully, but the Tukey model lends itself to diverse applications.

Section 2.2.1: Prospective Tukey Score Test

Using the Tukey model, Chatterjee et al. introduced the prospective Tukey score test (PTS) [175] . Score tests were first introduced by Rao in 1948 and are widely used in applied research [177]. They evaluate the derivative of a log-likelihood function evaluated under the null, as a measure of the deviation in the observed data from the null [178]. They are asymptotically equivalent to both Wald and likelihood ratio tests but they have the advantage of needing to fit only the null model [178,179]. This feature is particularly beneficial in analyses of the non-standard Tukey model.

The null hypothesis in PTS is that the scan SNP set is not associated with the disease: $\boldsymbol{\beta}_1 = \{\beta_{1m}, ..., \beta_{1M}\} = \mathbf{0}$. The parameter of interest $\boldsymbol{\beta}_1$ represents "general" disease association, incorporating the main effects and elements of the interaction effect for the scan SNPs. This "omnibus" character of PTS offers flexibility over single-SNP analyses, an advantage particularly in exploratory studies. However, it precludes the separate assessment of main- and epistatic effects, for scan SNPs.[175] I advise that researches assess the relative contribution of an epistatic effect to the disease association of a scan SNP by comparing its ranks in standard single-SNP and Tukey analyses. A substantial increase in rank would suggest epistasis drives the RTS signal.

The null Tukey model introduces complexity in PTS. It includes only main effects because the interaction term drops out of the model. Consequently, the null value of the nuisance parameter θ cannot be estimated from the data and a PTS statistic can be calculated only for a given θ. Chatterjee et al. advise calculating PTS statistics over a plausible range of θ (for example -5 to 5 on a 0.2 grid) and basing inference on the maximum test statistic [175], as has been advocated elsewhere [54].

Chatterjee et al. evaluated PTS through simulation study. They generated data for six markers in each SNP set, assuming a single causal variant was in linkage

disequilibrium. Chatterjee et al. varied the pattern of interaction between the SNP sets, allowing for a wide range of epistatic effects (multiplicative, additive, pure epistasis and cross-over). They compared the performance of PTS to UML for two Wald tests: main effects and omnibus. The first Wald test assessed the main effects of all scan SNPs, using a logistic model that included only scan SNP main effects. The second Wald test assessed main- and epistatic effects of the scan SNPs, using a saturated interaction model that included main effects of all SNPs and all pairwise interactions between scan and conditioning SNPs (4.1). PTS consistently gained power over UML analyses of saturated interaction models, presumably through the reduction in degrees of freedom. PTS equaled or improved the power of main effects Wald tests when the scan and conditioning SNP sets demonstrated epistasis and lost only minimal power in the absence of an interaction (multiplicative). The greatest power gains for PTS corresponded to models with the strongest relative epistatic effects (pure epistasis and cross-over).[175] These results highlight the omnibus nature of PTS, resembling the power profile of omnibus and marginal tests of disease association for studies of pairwise interaction (Section 1.3). They suggest the Tukey model is a promising framework through which to investigate interactions in genetic epidemiology. I extend the work of Chatterjee et al. through the retrospective Tukey score test that I derive, evaluate and apply in subsequent chapters.

Chapter 3

Retrospective Tukey Score Test:

Derivation

Section 3.1: Motivation for Retrospective Tukey Score Test

This chapter introduces the innovative retrospective Tukey score test (RTS). RTS is designed to exploit the power advantages of the prospective Tukey score test (PTS, Section. 2.2.1) [175] and constrained maximum likelihood logistic analysis (CML, Section 2.1.3) [152], thereby offering a more powerful alternative to existing methods. RTS assesses the disease association of a set of biologically similar single nucleotide polymorphisms (SNPs), allowing for interaction with a set of biologically similar susceptibility makers. It imposes an assumption of gene-gene independence in the underlying population between the SNP sets in the likelihood function under a Tukey model (3.1). RTS is expected to gain power over interaction analyses of standard logistic models because the parsimonious Tukey model reduces degrees of freedom on the test statistic, as for PTS. RTS is expected to improve power over PTS through the gene-gene independence constraint, as CML gains power over unconstrained maximum likelihood logistic analysis (UML, Section 2.1.2).

Section 3.2: Derivation of Retrospective Tukey Score Test

I reproduce the Tukey model (Section 2.2) because it is critical to RTS:

$$\text{logit}[P(D=1|\mathbf{S},\mathbf{C},\mathbf{A})]=\beta_0+\sum_{m=1}^{M}\beta_{1m}S_m+\sum_{z=1}^{Z}\beta_{2z}C_z+\theta\left(\sum_{m=1}^{M}\beta_{1m}S_m\sum_{z=1}^{Z}\beta_{2z}C_z\right)+\sum_{w=1}^{W}\beta_{3w}A_w \quad (3.1)$$

The model variables retain the following meanings throughout the text:

D is a n x 1 vector of disease status (1, case; 0, control)

$\mathbf{S}=(S_1,...,S_M)$ is an n x M matrix of minor allele counts $(0,1,2)$ for a set of scan SNPs being tested for disease association

$\mathbf{C}=(C_1,...,C_Z)$ is an n x Z matrix of risk allele counts $(0,1,2)$ for the conditioning SNPs that are presumed to be associated with the disease and may interact with scan SNPs, for $M \geq 1$ and $Z \geq 1$ with the constraint $M+Z>2$

$\mathbf{A}=(A_1,...,A_W)$ is an n x W matrix of adjusting covariates, such as gender, for $W \geq 0$

I introduce $\mathbf{K}$ to represent population stratification variables, such as ethnicity or geography, that may be a subset of $\mathbf{A}$. The purpose of the stratification variable $\mathbf{K}$ in RTS analyses is to guard against epistatic population stratification by imposing conditional gene-gene independence between the SNP sets within strata set by $\mathbf{K}$. This stratification is similar to that allowed by CML.

I derive RTS in stages. I consider a "simple" Tukey analysis first, building to a "full" analysis. They differ in variable selection: a simple analysis includes a single scan SNP and multiple conditioning SNPs, whereas a full analysis includes additional scan SNPs, adjusting covariates and stratification variables. The derivations follow each other closely and I provide fewer details in the second section. I highlight key

equations by shading the corresponding equation numbers in grey. It is important to note that this work constructs an RTS statistic for a given θ. Inference in RTS, as in PTS, is based on the maximum score test statistic for a range of θ (Section 2.2.1).

Section 3.2.1: Simple Tukey Analysis

Consider a Tukey model (3.1) that includes a single scan SNP and multiple conditioning SNPs ($M=1$, $Z>1$, $W=0$ and K is the empty set) and let the following notation hold throughout the text:

$n_{ds\mathbf{c}}$ = number subjects with $D=d, S=s, \mathbf{C}=\mathbf{c}$

$n_{d..}$ = number subjects with $D=d$, analogously for alternate indices

$$P_{ds} = P(D=d \mid S=s, \mathbf{C}=\mathbf{c}) = \frac{\exp\{d(\beta_0 + m(s,\mathbf{c};\beta_1,\boldsymbol{\beta}_2))\}}{1+\exp\{\beta_0 + m(s,\mathbf{c};\beta_1,\boldsymbol{\beta}_2)\}}$$

with $m(\mathbf{x};\boldsymbol{\gamma})$ an arbitrary function of covariates $(\mathbf{x})$ and their parameters $(\boldsymbol{\gamma})$ in the regression model under study

$$q(s,\mathbf{f}) = P(S=s) = \sum_s I_{s_i=s} f_s$$

with $I_{s_i=s} = \begin{Bmatrix} 0 \text{ if } s_i \neq s \\ 1 \text{ if } s_i = s \end{Bmatrix}$; this convention applies to all indicators.

$\mathbf{f} = \{f_u = P(S=u); u=1,2\}$, setting $f_0 = 1 - \sum_u f_u$ without an assumption of Hardy-Weinberg Equilibrium

$\delta = P(\mathbf{C}=\mathbf{c})$

$\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_2)$, or generally all parameters of null model that data estimate

$\boldsymbol{\psi} = (\boldsymbol{\beta}, \mathbf{f})$, nuisance parameters of RTS for given θ

$\boldsymbol{\eta} = (\boldsymbol{\beta}, \beta_1, \mathbf{f})$, all parameters of RTS for given θ

$\boldsymbol{\eta}_{\mathbf{0}} = (\boldsymbol{\beta}, \beta_1 = 0, \mathbf{f})$, null values of $\boldsymbol{\eta}$

The retrospective likelihood and log-likelihood function for a Tukey model in the simple analysis are:

$$l = \prod_{i=1}^{n} P(s_i, \mathbf{c}_i \mid d_i) = \prod_{i=1}^{n} \frac{P(D = d_i \mid S = s_i, \mathbf{C} = \mathbf{c}_i) P(S = s_i) P(\mathbf{C} = \mathbf{c}_i)}{\sum_{\mathbf{c}=\mathbf{0}}^{\mathbf{2}} \sum_{s=0}^{2} P(D = d_i \mid S = s_i, \mathbf{C} = \mathbf{c}_i) P(S = s_i) P(\mathbf{C} = \mathbf{c}_i)} \tag{3.2}$$

$$L = \sum_i \ln[P(d_i \mid s_i, \mathbf{c}_i)] + \ln[P(s_i)] + \ln[P(\mathbf{c}_i)] - \ln[P(d_i)] \tag{3.3}$$

$$= \sum_{ds\mathbf{c}} n_{ds\mathbf{c}} \ln(P_{ds}) + \sum_s n_{.s.} \ln[q(s, \mathbf{f})] + \sum_{\mathbf{c}} n_{..\mathbf{c}} \ln(\delta) - \sum_d n_{d..} \ln\left(\sum_{s\mathbf{c}} P_{ds} q(s, \mathbf{f}) \delta \right) \tag{3.4}$$

It may be computationally infeasible to maximize (3.4) when the conditioning factor $(\mathbf{C})$ corresponds to a high dimensional nuisance parameter (δ). In the setting of a retrospective analysis of a standard logistic model, Chatterjee and Carroll overcame this challenge through the use of a profile likelihood that does not involve high-dimensional nuisance parameters and retains the maximum likelihood estimates of a prospective analysis [152]. To motivate the profile-likelihood in this setting, note that the first two terms in (3.3) satisfy the following equality when S and $\mathbf{C}$ are independent:

$$\ln[P(d_i \mid s_i, \mathbf{c}_i)] + \ln[P(s_i)] = \ln[P(d_i, s_i \mid \mathbf{c}_i)] \tag{3.5}$$

Let R be an indicator of whether an individual in the underlying population is selected for the study. Define $\boldsymbol{\mu} = (\mu_0, \mu_1)$ for $\mu_d = P(R = 1 \mid D = d)$, which is the probability for selection of either a case or control from the underlying population. At the center of the profile likelihood is:

$$P^*_{ds} = P(D,S \mid \mathbf{C}, R=1) = \frac{P(D,S,\mathbf{C},R=1)}{\sum_{ds} P(D,S,\mathbf{C},R=1)} = \frac{P(R=1 \mid D,S,\mathbf{C})P(D \mid S,\mathbf{C})P(S \mid \mathbf{C})P(\mathbf{C})}{P(\mathbf{C})\sum_{ds} P(R=1 \mid D,S,\mathbf{C})P(D \mid S,\mathbf{C})P(S \mid \mathbf{C})}$$

$$= \frac{P(R=1 \mid D)P(D \mid S,\mathbf{C})P(S)}{\sum_{ds} P(R=1 \mid D)P(D \mid S,\mathbf{C})P(S)} = \frac{\mu_d P_{ds} q(s,\mathbf{f})}{\sum_{ds} \mu_d P_{ds} q(s,\mathbf{f})}$$

$$= \frac{\exp\{h_{ds}(\mathbf{c};\boldsymbol{\beta},\mathbf{f},\boldsymbol{\mu})\}}{\sum_{ds} \exp\{h_{ds}(\mathbf{c};\boldsymbol{\beta},\mathbf{f},\boldsymbol{\mu})\}} \tag{3.6}$$

for $h_{ds}(\mathbf{c};\boldsymbol{\beta},\beta_1,\mathbf{f},\boldsymbol{\mu}) = \ln\left(\frac{P^*_{ds}}{P^*_{00}}\right)$

$$= \ln\left(\frac{\mu_d}{\mu_0}\right) + \ln\left(\frac{P_{ds}}{P_{00}}\right) + \ln\left(\frac{q(s,\mathbf{f})}{q(0,\mathbf{f})}\right)$$

$$= d\left[\ln\left(\frac{\mu_1}{\mu_0}\right) + \beta_0 + m(s,\mathbf{c};\beta_1,\boldsymbol{\beta}_2)\right] + \ln\left(\frac{1+\exp\{\beta_0 + m(0,\mathbf{c};\beta_1,\boldsymbol{\beta}_2)\}}{1+\exp\{\beta_0 + m(s,\mathbf{c};\beta_1,\boldsymbol{\beta}_2)\}}\right) + \ln\left(\frac{q(s,\mathbf{f})}{q(0,\mathbf{f})}\right) \tag{3.7}$$

The second term in (3.7) contributes little information to the profile likelihood because it approximates zero for small effect sizes, as expected with rare diseases in genetic epidemiology [152]. Subsequently, I define

$$h^*_{ds} = d\left[\ln\left(\frac{\mu_1}{\mu_0}\right) + \beta_0 + m(s,\mathbf{c};\beta_1,\boldsymbol{\beta}_2)\right] + \ln\left(\frac{q(s,\mathbf{f})}{q(0,\mathbf{f})}\right) \tag{3.8}$$

$$\text{with } m(s,\mathbf{c};\beta_1,\boldsymbol{\beta}_2) = \beta_1 s\left(1+\theta\sum_z \beta_{2z}c_z\right) + \sum_z \beta_{2z}c_z$$

The corresponding log-likelihood for maximization is:

$$L^* = \sum_{d\mathbf{sc}} n_{d\mathbf{sc}} \ln\left(P^*_{ds}\right) = \sum_{d\mathbf{sc}} n_{d\mathbf{sc}}\left[h^*_{ds} - \ln\left(\sum_{d"s"} \exp\left(h^*_{d"s"}\right)\right)\right] \tag{3.9}$$

This log-likelihood incorporates the first two terms of (3.4) as suggested by (3.5) whereas the last two terms of (3.4) have been dropped due to the conditioning on $\mathbf{C}$.

The score function is the first derivate of this log-likelihood with respect to $\boldsymbol{\eta}$:

$$S(\boldsymbol{\eta}) = \sum_{ds\mathbf{c}} n_{ds\mathbf{c}} \left(\frac{\partial h^*_{ds}}{\partial \boldsymbol{\eta}} - \sum_{d"s"} \frac{\exp\{h^*_{d"s"}\} * \frac{\partial h^*_{d"s"}}{\partial \boldsymbol{\eta}}}{\sum_{d"s"} \exp\{h^*_{d"s"}\}} \right)$$

$$= \sum_i \frac{\partial h^*_{ds}}{\partial \boldsymbol{\eta}} - E\left[\frac{\partial h^*_{ds}}{\partial \boldsymbol{\eta}} \mid \mathbf{C} = \mathbf{c}_i, R = 1 \right] \tag{3.10}$$

The final equality in (3.10) holds through the identity $E\left[g(X) \right] = \sum_x g(x) P(x)$ for $g(X)$, a generic function of a random variable [116,180]. To compute the score function, one needs both the matrix of first derivatives for h^*_{ds} with respect to $\boldsymbol{\eta}$ (3.8) and its conditional expectation with respect to $\mathbf{C}$ and R:

$$\frac{\partial h^*_{ds}(\boldsymbol{\eta})_i}{\partial \boldsymbol{\eta}} = \begin{pmatrix} d_i \\ d_i \mathbf{c}_i^T (1 + \theta \beta_1 s_i) \\ d_i s_i \left(1 + \theta \sum_z \beta_{2z} c_{zi} \right) \\ \frac{I_{s_i \neq 0}}{1 - f_1 - f_2} + \frac{I_{s_i = 1}}{f_1} \\ \frac{I_{s_i \neq 0}}{1 - f_1 - f_2} + \frac{I_{s_i = 2}}{f_2} \end{pmatrix}^T \tag{3.11}$$

$$E\left[\frac{\partial h^*_{ds}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} \mid \mathbf{C} = \mathbf{c}_i, R = 1 \right] = \begin{pmatrix} p^*_i \\ p^*_i \mathbf{c}_i^T \left[1 + \theta \beta_1 \left(f^*_1 + 2 f^*_2 \right) \right] \\ p^*_i \left(f^*_1 + 2 f^*_2 \right) \left(1 + \theta \sum_z \beta_{2z} c_{zi} \right) \\ \frac{f^*_1 + f^*_2}{1 - f_1 - f_2} + \frac{f^*_1}{f_1} \\ \frac{f^*_1 + f^*_2}{1 - f_1 - f_2} + \frac{f^*_2}{f_2} \end{pmatrix}^T \tag{3.12}$$

The entries for $\mathbf{f}$ in (3.11) involve indicators because the derivative of $\ln\left(\frac{q(s,\mathbf{f})}{q(0,\mathbf{f})}\right)$ depends on the observed value of S. I used the following expectations to compute the entries of (3.12):

$$E\left[D \mid \mathbf{C}=\mathbf{c}_i, R=1\right] = p_i^*, \text{ which is set by the null model}$$

$$E\left[I_{s\neq u} \mid \mathbf{C}=\mathbf{c}, R=1\right] = 1-f_u^*$$

$$E\left[S \mid \mathbf{C}=\mathbf{c}, R=1\right] = \sum_s s f_s^*$$

$$E\left[I_{s=u} \mid \mathbf{C}=\mathbf{c}, R=1\right] = \sum_d P\left(S=u, D=d \mid \mathbf{C}=\mathbf{c}, R=1\right) = P\left(S=u \mid \mathbf{C}=\mathbf{c}, R=1\right)$$

$$= \sum_d P_{du}^* = f_u^* \qquad (3.13)$$

Given these results, the score function evaluated under the null is:

$$S\left(\boldsymbol{\eta}_0\right) = \sum_i \begin{pmatrix} d_i - p_i^* \\ \mathbf{c}_i^T\left(d_i - p_i^*\right) \\ \left(1+\theta\sum_z \beta_{2z} c_{zi}\right)\left(s_i d_i - \left(f_1+2f_2\right)p_i^*\right) \\ \frac{I_{s_i\neq 0} - f_1 - f_2}{f_0} + \frac{I_{s_i=1}}{f_1} - 1 \\ \frac{I_{s_i\neq 0} - f_1 - f_2}{f_0} + \frac{I_{s_i=2}}{f_2} - 1 \end{pmatrix}^T \qquad (3.14)$$

The entries for $\mathbf{f}$ in (3.14) hold because under the null $f_u^* = f_u$. Specifically, the null constraint of $\beta_1 = 0$ gives $f_u^* = \frac{q(u,\mathbf{f})\sum_d \mu_d P_{d+}}{\sum_d \mu_d P_{d+}} = f_u$ with algebra from (3.13). The entries for all nuisance parameters in (3.14) are standard results that easily can be computed in practice [181]. First,

$$S_{\boldsymbol{\beta}}(\boldsymbol{\eta}_0) = \sum_i (1, \mathbf{c}_i)^T \left(d_i - p_i^*\right)$$

is the standard score function for $\text{logit}(P(D=1|\mathbf{C})) = \beta_0 + \boldsymbol{\beta}_2\mathbf{C}$. It matches that of PTS. Second,

$$S_{\mathbf{f}}(\boldsymbol{\eta}_0) = \begin{pmatrix} \frac{n_{.1.}}{f_1} - \frac{n_{.0.}}{1-f_1-f_2} \\ \frac{n_{.2.}}{f_2} - \frac{n_{.0.}}{1-f_1-f_2} \end{pmatrix}$$

is the standard score function for probabilities associated with a random multi-categorical sample. There is no PTS equivalent, as the gene-gene independence assumption necessitates the nuisance parameter $\mathbf{f}$.

The score function converges over n to a multivariate normal distribution with mean zero and covariance $\mathbf{I}$, Fisher's information matrix, under suitable regularity conditions [182]. In this setting, the information matrix is:

$$\mathbf{I}^*(\boldsymbol{\eta}) = E\left[-\frac{\partial^2 L^*(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T} | \mathbf{C} = \mathbf{c}_i, R = 1 \right] \tag{3.15}$$

It can be classified as an observed information matrix because it is conditioned on the random variables $\mathbf{C}$ and R. The observed information matrix is routinely used in applied work. It is identical to the expected information when analyses involve canonical link functions and it is less computationally intensive. In general, neither information matrix is uniformly preferred because both have the same asymptotic mean squared error for variance estimation.[182] Some reports demonstrate the observed information matrix can outperform the expected information [183–185]. Others caution that score test statistics can be negative and invalid for inference when the observed information matrix is used, particularly when the null model fits the data poorly [186–

188]. Given the novelty of RTS, I present simulations that investigate this potential short-coming, after the derivation.

To compute the information matrix (3.15) one needs the second derivative of L^* (3.9) with respect to $\boldsymbol{\eta}$:

$$\frac{\partial^2 L^*(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T} =$$

$$\sum_{ds\mathbf{c}} n_{ds\mathbf{c}} \left[\frac{\partial^2 h_{ds}^*}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T} - \sum_{d''s''} \frac{\exp\left\{h_{d''s''}^*\right\}}{\sum_{d''s''}\exp\left\{h_{d''s''}^*\right\}} \left(\left[\left(\frac{\partial h_{d''s''}^*}{\partial\boldsymbol{\eta}} \right)^T \frac{\partial h_{d''s''}^*}{\partial\boldsymbol{\eta}} + \frac{\partial^2 h_{d''s''}^*}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T} \right] - \frac{\frac{\partial h_{d''s''}^*}{\partial\boldsymbol{\eta}} \sum_{d''s''}\exp\left\{h_{d''s''}^*\right\} \frac{\partial h_{d''s''}^*}{\partial\boldsymbol{\eta}}}{\sum_{d''s''}\exp\left\{h_{d''s''}^*\right\}} \right) \right]$$

$$= \sum_{ds\mathbf{c}} n_{ds\mathbf{c}} \left[\frac{\partial^2 h_{ds}^*}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T} - E\left[\frac{\partial^2 h_{ds}^*}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T} \mid \mathbf{C}=\mathbf{c}_i, R=1 \right] \text{-Var}\left(\frac{\partial h_{ds}^*}{\partial\boldsymbol{\eta}} \mid \mathbf{C}=\mathbf{c}_i, R=1 \right) \right] \qquad (3.16)$$

Chatterjee and Carroll demonstrated the difference, given by the first two terms in (3.16), goes to zero in probability [152]. Consequently, I use the following information matrix to derive the RTS statistic:

$$\mathrm{I}(\boldsymbol{\eta}) = \sum_i \mathrm{Var}\left(\frac{\partial h_{ds}^*}{\partial\boldsymbol{\eta}} \mid \mathbf{C}=\mathbf{c}_i, R=1 \right)$$

$$= \sum_i E\left[\left(\frac{\partial h_{ds}^*}{\partial\boldsymbol{\eta}} \right)^T \left(\frac{\partial h_{ds}^*}{\partial\boldsymbol{\eta}} \right) \mid \mathbf{C}=\mathbf{c}_i, R=1 \right] - E^2\left[\left(\frac{\partial h_{ds}^*}{\partial\boldsymbol{\eta}} \right) \mid \mathbf{C}=\mathbf{c}_i, R=1 \right] \qquad (3.17)$$

To compute this information matrix, one needs $\left(\frac{\partial h^*_{ds}}{\partial \boldsymbol{\eta}}\right)^T \left(\frac{\partial h^*_{ds}}{\partial \boldsymbol{\eta}}\right)$. It is a triangular matrix with diagonal elements:

$$\begin{pmatrix} d_i^2 \\ d_i^2 \mathbf{c}_i^T \mathbf{c}_i \left(1+\theta\beta_1 s_i\right)^2 \\ d_i^2 s_i^2 \left(1+\theta\sum_z \beta_{2z} c_{zi}\right)^2 \\ \frac{I^2_{s_i\neq 0}}{\left(1-f_1-f_2\right)^2} + \frac{2I_{s_i\neq 0} I_{s_i=1}}{f_1\left(1-f_1-f_2\right)} + \frac{I^2_{s_i=1}}{f_1^2} \\ \frac{I^2_{s_i\neq 0}}{\left(1-f_1-f_2\right)^2} + \frac{2I_{s_i\neq 0} I_{s_i=2}}{f_2\left(1-f_1-f_2\right)} + \frac{I^2_{s_i=2}}{f_2^2} \end{pmatrix}$$

The off-diagonal elements of $\left(\frac{\partial h^*_{ds}}{\partial \boldsymbol{\eta}}\right)^T \left(\frac{\partial h^*_{ds}}{\partial \boldsymbol{\eta}}\right)$ are:

$$\begin{pmatrix} d_i^2 \mathbf{c}_i (1+\theta\beta_1 s_i) & d_i^2 s_i \left(1+\theta\sum_z \beta_{2z} c_{zi}\right) & \left(\frac{d_i I_{s_i\neq 0}}{f_0}+\frac{d_i I_{s_i=1}}{f_1}\right) & \left(\frac{d_i I_{s_i\neq 0}}{f_0}+\frac{d_i I_{s_i=2}}{f_2}\right) \\ & d_i^2 s_i \mathbf{c}_i^T \left(1+\theta\sum_z \beta_{2z} c_{zi}\right)(1+\theta\beta_1 s_i) & d_i \mathbf{c}_i^T (1+\theta\beta_1 s_i)\left(\frac{I_{s_i\neq 0}}{f_0}+\frac{I_{s_i=1}}{f_1}\right) & d_i c_i^T (1+\theta\beta_1 s_i)\left(\frac{I_{s_i\neq 0}}{f_2}+\frac{I_{s_i=2}}{f_2}\right) \\ & & d_i s_i \left(1+\theta\sum_z \beta_{2z} c_{zi}\right)\left(\frac{I_{s_i\neq 0}}{f_0}+\frac{I_{s_i=1}}{f_1}\right) & d_i s_i \left(1+\theta\sum_z \beta_{2z} c_{zi}\right)\left(\frac{I_{s_i\neq 0}}{f_0}+\frac{I_{s_i=2}}{f_2}\right) \\ & & & \frac{I^2_{s_i\neq 0}}{f_0^2}+\sum_u \frac{I_{s_i\neq 0} I_{s_i=u}}{f_u f_0}+\frac{I_{s_i=1} I_{s_i=2}}{f_1 f_2} \end{pmatrix}$$

I computed its conditional expectation, using the following equalities:

$$E\left[I_{s_i=1} I_{s_i=2} \mid \mathbf{C}=\mathbf{c}_i, R=1\right]=0;$$

$$E\left[I_{s_i\neq 0} I_{s_i=u} \mid \mathbf{C}=\mathbf{c}_i, R=1\right]=f_u^*$$

$$E\left[s_i * I_{s_i\neq 0} \mid \mathbf{C}=\mathbf{c}_i, R=1\right]=f_1^*+2f_2^*$$

$$E\left[s_i * I_{s_i=u} \mid \mathbf{C}=\mathbf{c}_i, R=1\right]=\begin{cases} u f_u^* & \text{if } s_i=u,\ u=1,2 \\ 0 & \text{else} \end{cases}$$

The diagonal elements of $E\left[\left(\frac{\partial h_{ds}^{*}}{\partial \boldsymbol{\eta}}\right)^{T}\left(\frac{\partial h_{ds}^{*}}{\partial \boldsymbol{\eta}}\right) \mid \mathbf{C}=\mathbf{c}_i, R=1\right]$ are:

$$\begin{pmatrix} p_i^{*} \\ p_i^{*}\mathbf{c}_i^{T}\mathbf{c}_i\left(1+2\theta\beta_1\left(f_1^{*}+2f_2^{*}\right)+\theta^2\beta_1^2\left(f_1^{*}+4f_2^{*}\right)\right) \\ p_i^{*}\left(f_1^{*}+4f_2^{*}\right)\left(1+\theta\sum_z \beta_{2z}c_{zi}\right)^2 \\ \frac{f_1^{*}+f_2^{*}}{f_0^2}+\frac{2f_1^{*}}{f_1 f_0}+\frac{f_1^{*}}{f_1^2} \\ \frac{f_1^{*}+f_2^{*}}{f_0^2}+\frac{2f_2^{*}}{f_2 f_0}+\frac{f_2^{*}}{f_2^2} \end{pmatrix}$$

Its off-diagonal elements are:

$$\begin{pmatrix} p_i^{*}\mathbf{c}_i\left(1+\theta\beta_1\left(f_1^{*}+2f_2^{*}\right)\right) & p_i^{*}\left(f_1^{*}+2f_2^{*}\right)\left(1+\theta\sum_z \beta_{2z}c_{zi}\right) & \frac{p_i^{*}\left(f_1^{*}+f_2^{*}\right)}{f_0}+\frac{p_i^{*}f_1^{*}}{f_1} & \frac{p_i^{*}\left(f_1^{*}+f_2^{*}\right)}{f_0}+\frac{p_i^{*}f_2^{*}}{f_2} \\ & p_i^{*}\left(f_1^{*}+2f_2^{*}\right)c_i\left(1+\theta\sum_z \beta_{2z}c_{zi}\right)\left(1+\theta\beta_1\left(f_1^{*}+2f_2^{*}\right)\right) & p_i^{*}\mathbf{c}_i^{T}\left[\left(\frac{f_1^{*}+f_2^{*}}{f_0}+\frac{f_1^{*}}{f_1}\right)+\theta\beta_1\left(\frac{f_1^{*}+2f_2^{*}}{f_0}+\frac{f_1^{*}}{f_1}\right)\right] & p_i^{*}\mathbf{c}_i^{T}\left[\left(\frac{f_1^{*}+f_2^{*}}{f_0}+\frac{f_2^{*}}{f_2}\right)+\theta\beta_1\left(\frac{f_1^{*}+2f_2^{*}}{f_0}+\frac{f_2^{*}}{f_2}\right)\right] \\ & & p_i^{*}\left(1+\theta\sum_z \beta_{2z}c_{zi}\right)\left(\frac{f_1^{*}+2f_2^{*}}{f_0}+\frac{f_2^{*}}{f_2}\right) & p_i^{*}\left(1+\theta\sum_z \beta_{2z}c_{zi}\right)\left(\frac{f_1^{*}+2f_2^{*}}{f_0}+\frac{f_1^{*}}{f_1}\right) \\ & & & \frac{f_1^{*}+f_2^{*}}{f_0^2}+\sum_u \frac{f_u^{*}}{f_u f_0} \end{pmatrix}$$

The final term needed to compute the information matrix (3.17) is $E^T\left[\frac{\partial h_{ds}^*(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} \mid \mathbf{C}=\mathbf{c}_i, R=1\right] E\left[\frac{\partial h_{ds}^*(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} \mid \mathbf{C}=\mathbf{c}_i, R=1\right]$ that follows from (3.12). The information matrix itself is also triangular. Under the null, it can be written without repetition of off-diagonal elements as:

$$\mathbf{I}(\boldsymbol{\eta}_0)=\begin{pmatrix} p_i^*(1-p_i^*) & p_i^*(1-p_i^*)\mathbf{c}_i & p_i^*(1-p_i^*)(f_1+2f_2)\left(1+\theta\sum_z \beta_{2z}c_{zi}\right) & \mathbf{0}_{(Z+1)x2} & \\ & p_i^*(1-p_i)\mathbf{c}_i^T\mathbf{c}_i & p_i^*(1-p_i^*)\mathbf{c}_i^T(f_1+2f_2)\left(1+\theta\sum_z \beta_{2z}c_{zi}\right) & & \\ & & \left(p_i^*(f_1+4f_2)-p_i^{2*}(f_1+2f_2)^2\right)\left(1+\theta\sum_z \beta_{2z}c_{zi}\right)^2 & p_i^*\left(1+\theta\sum_z \beta_{2z}c_{zi}\right) & 2\,p_i^*\left(1+\theta\sum_z \beta_{2z}c_{zi}\right) \\ & & & \frac{1}{f_0}+\frac{1}{f_1} & \frac{1}{f_0} \\ & & & & \frac{1}{f_0}+\frac{1}{f_2} \end{pmatrix} \quad (3.18)$$

As with the score function, the information matrix for the nuisance parameters can be deconstructed into two standard results:

$$\mathbf{I}_{\beta\beta}(\boldsymbol{\eta}_0)=\sum_i \left(1,\mathbf{c}_i^T\right)\left(1,\mathbf{c}_i^T\right)^T p_i^*\left(1-p_i^*\right)$$

$$\mathbf{I}_{\mathrm{ff}}(\boldsymbol{\eta}_0)=n\begin{pmatrix}\frac{1}{f_0}+\frac{1}{f_1} & \frac{1}{f_0}\\ \frac{1}{f_0} & \frac{1}{f_0}+\frac{1}{f_2}\end{pmatrix}$$

These results define the RTS statistic for inference on the disease association of a scan SNP as follows:

$$T=\max_\theta\left(T(\theta)\right)$$

$$\text{for } T(\theta)=S_{\beta_1}^T\left(\hat{\boldsymbol{\eta}}_0\right)\mathbf{I}^{\beta_1\beta_1}\left(\hat{\boldsymbol{\eta}}_0\right)S_{\beta_1}\left(\hat{\boldsymbol{\eta}}_0\right) \tag{3.19}$$

$$\text{and } \mathbf{I}^{\beta_1\beta_1}\left(\hat{\boldsymbol{\eta}}_0\right)=\left[\mathbf{I}_{\beta_1\beta_1}\left(\hat{\boldsymbol{\eta}}_0\right)-\mathbf{I}_{\beta_1\psi}\left(\hat{\boldsymbol{\eta}}_0\right)\mathbf{I}_{\psi\psi}^{-1}\left(\hat{\boldsymbol{\eta}}_0\right)\mathbf{I}_{\beta_1\psi}^T\left(\hat{\boldsymbol{\eta}}_0\right)\right]^{-1} \tag{3.20}$$

As this score statistic does not have a standard null distribution, permutation is appropriate for its p-value computation (Sections 4.3-4.5). The score function and information matrix used to calculate the RTS statistics are evaluated using the maximum likelihood estimates of the nuisance parameters under the null. Consequently, the score function is not a summation of independent contributions from each subject and the asymptotic representation of $\mathbf{I}_{\beta_1\beta_1}(\boldsymbol{\eta}_0)$ (3.20) is used to compute the RTS statistic [178]. The maximum likelihood estimates for $\boldsymbol{\beta}$ can be obtained through standard logistic analysis because the null Tukey model has standard form, including only main effects for the conditioning SNPs (and adjusting covariate in the

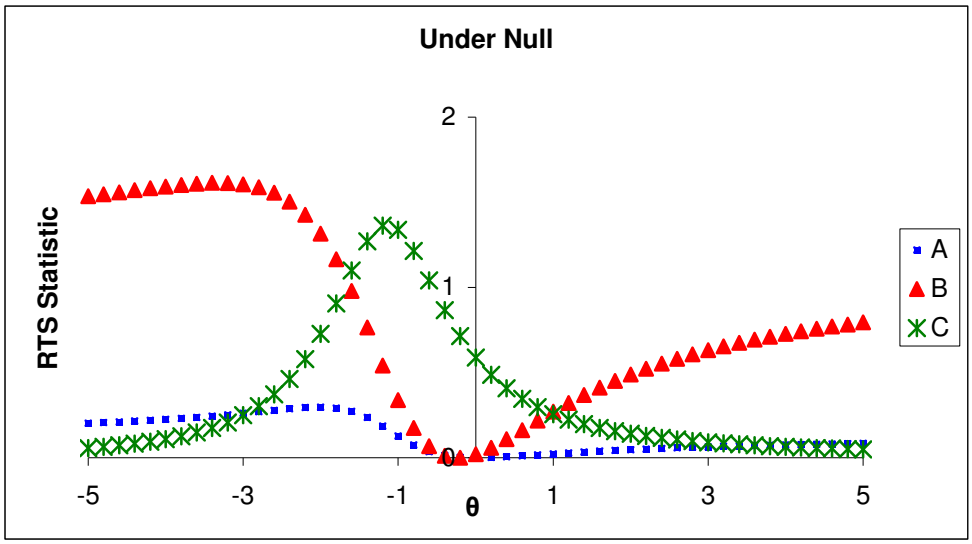
full analysis). The estimates for **f** are $\hat{f}_u = \frac{n_{.u.}}{n}$ because the genotype frequencies do not differ between cases and controls under the null.

Section 3.2.1a: Theta Maximization in Simple Tukey Analysis

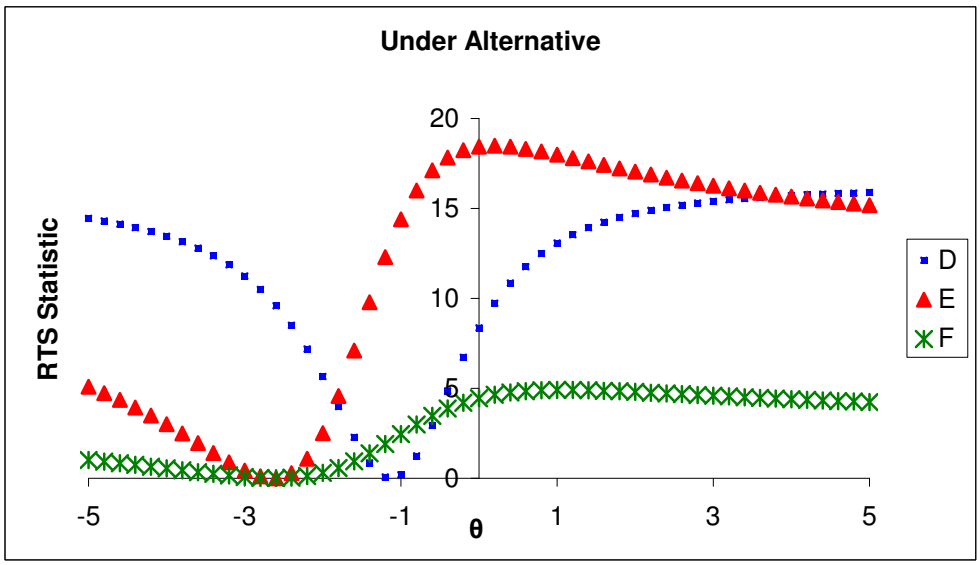
Before proceeding to the derivation of RTS in the full Tukey analysis, I examined the θ grid over which RTS statistics are maximized: -5 to 5 by 0.2 intervals. Inference based on $\max_\theta \left(T\left(\theta\right)\right)$ would be jeopardized by negative $T\left(\theta\right)$'s. Also, if a distribution of positive RTS statistics were not smooth over θ, inference would be highly sensitive to the selected θ grid. I generated ten sets of data under the null and alternative hypotheses, recording all $T\left(\theta\right)$'s. (See Section 4.1 for details on data simulation under the Tukey model.) The results indicate that the RTS statistics are consistently positive and that the θ grid is sufficiently dense. The plots also illustrate that $T\left(\theta\right)$ is not monotone over θ and $\max_\theta \left(T\left(\theta\right)\right)$ does not preferentially occur at the boundaries of θ. (Figure 3.1)

Figure 3.1: Retrospective Tukey score test statistics over range of θ . Three representative examples (legend) are given for simulations under the null (top) and alternative (bottom) hypotheses. Simulations included a total sample size of 1000 with equal cases and controls, eight conditioning SNPs and a single scan SNP with MAF=0.12. Under the alternative, $\beta_1 = \ln(1.12)$ and $\theta = 1.2$.

A)



B)



Section 3.2.2: Full Tukey Analysis

Consider the Tukey model (3.1) with multiple scan and conditioning SNPs, as well as adjusting covariates ($M > 1$, $Z > 1$ and $W > 1$). I impose two constraints of gene-gene independence in the likelihood function. First, I assume the scan and conditioning SNP sets are independent in the underlying population conditional on $\mathbf{K}$. Second, I assume the markers within the scan SNP set are independent in the underlying population. The second assumption should not impede investigations of scan SNPs from unlinked regions, such as those from different genes within a pathway. In this derivation I set K to be a binary variable, although more complex stratification is possible. For example, a trichotomous regression allows estimation of $P(\mathbf{S} \mid \mathbf{K})$ for continuous stratification variables. I assume K is not used for adjusting in the regression model and I update notation for parameters with $\boldsymbol{\beta} = \{\boldsymbol{\beta}_0, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3\}$.

The retrospective likelihood in this setting is:

$$l = \prod_i \frac{P(D \mid \mathbf{S} = \mathbf{s}_i, \mathbf{C} = \mathbf{c}_i, \mathbf{A} = \mathbf{a}_i) P(\mathbf{C} = \mathbf{c}_i, \mathbf{A} = \mathbf{a}_i) \prod_m P(S_m = s_{mi} \mid K = k_i)}{P(D)} \qquad (3.21)$$

Because I take the conditional probability of $\mathbf{S}$ with respect to only K, this likelihood enforces independence of $\mathbf{S}$ and $\mathbf{A}$. This assumption is relaxed when the conditional expectation is taken with respect to adjusting and stratification variables, for example when $\mathbf{A}$ sets the strata defined by K, as in the simulations of Chapter 4.

The corresponding profile log-likelihood is:

$$L = \sum_{d\mathbf{sca}k} n_{d\mathbf{sca}k}\left[h_{d\mathbf{s}} - \ln\left(\sum_{d'\mathbf{s}'} \exp\left(h_{d'\mathbf{s}'}\right)\right)\right]$$

where $n_{d\mathbf{sca}k}$ = number subjects with $D = d, \mathbf{S} = \mathbf{s}, \mathbf{C} = \mathbf{c}, \mathbf{A} = \mathbf{a}, K = k$

$$h_{d\mathbf{s}} = d\left[\ln\left(\frac{\mu_1}{\mu_0}\right) + \beta_0 + m\left(\mathbf{s},\mathbf{c},\mathbf{a};\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{\beta}_3,\theta\right)\right] + \sum_m \ln\left(\frac{q\left(s,m,k,\mathbf{f}\right)}{q\left(0,m,k,\mathbf{f}\right)}\right) \qquad (3.22)$$

$$q\left(j,m,k,\mathbf{f}\right) = \sum_j I_{s_m = j} * f_{jmk}$$

$$\mathbf{f} = \left\{ f_{umk} = I_{K=ki} P\left(S_m = u \mid K = k\right) \text{ for } u = 1,2;\ k = 0,1;\ f_{0mk} = 1 - \left(f_{1mk} + f_{2mk}\right)\right\}.$$

Note that independence of scan SNPs establishes independence of f_{jmk} and $f_{jm'k}$ for $m \neq m'$, simplifying calculations for the information matrix.

In this derivation, I highlight the elements needed to compute the RTS statistic. First, the score function for a given θ is:

$$S\left(\boldsymbol{\eta}\right) = \sum_i \left(\frac{\partial h_{d\mathbf{s}}}{\partial \boldsymbol{\eta}} - E\left[\frac{\partial h_{d\mathbf{s}}}{\partial \boldsymbol{\eta}} \mid \mathbf{C}_i = \mathbf{c}_i, \mathbf{A}_i = \mathbf{a}_i, K_i = k_i, R = 1\right]\right)$$

Under the null, its form is:

$$S_{\boldsymbol{\beta}_1}\left(\boldsymbol{\eta}_0\right) = \sum_i \left(1 + \theta \sum_z \beta_{2z} c_{zi}\right)\left(\mathbf{s}_i d_i - E\left(\mathbf{S}\right) p_i^*\right)$$

where $p_i^* = E\left[D \mid \mathbf{C}_i = \mathbf{c}_i, \mathbf{A}_i = \mathbf{a}_i, K_i = k_i, R = 1\right]$;

for ease of presentation $E\left(\mathbf{S}\right)$ denotes $E\left[\mathbf{S} \mid \mathbf{C}_i = \mathbf{c}_i, \mathbf{A}_i = \mathbf{a}_i, K_i = k_i, R = 1\right]$

with $E\left(S_m \mid \mathbf{C}_i = \mathbf{c}_i, \mathbf{A}_i = \mathbf{a}_i, K_i = k_i, R = 1\right) = f_{1mk} + 2 f_{2mk}$

The increased dimensionality of $\mathbf{S}$ differentiates this score function from that of the simple Tukey analysis (3.14). Next, I consider the observed information matrix:

$$\mathbf{I}\left(\boldsymbol{\eta}\right) = \sum_i \mathrm{Var}\left(\frac{\partial h_{d\mathbf{s}}}{\partial \boldsymbol{\eta}} \mid \mathbf{C} = \mathbf{c}_i, \mathbf{A}_i = \mathbf{a}_i, K_i = k_i, R = 1\right)$$

As before, I partition it into three components for the parameter of interest $(\boldsymbol{\beta}_1)$ and nuisance parameters $(\boldsymbol{\psi} = \{\boldsymbol{\beta}, \mathbf{f}\})$. For ease of presentation, I do not repeat entries involving $\mathbf{f}$ that differ only in m or k subscripts, but rather denote the entries with appropriate matrix dimensions for a general formula. Subjects contribute non-zero entries to $\mathbf{I}_{\mathbf{ff}}$ only for their respective strata because $\mathbf{f}$ involves indicators for stratification levels. The information matrix, evaluated under the null, is:

$$\mathbf{I}_{\boldsymbol{\beta}_1\boldsymbol{\beta}_1}(\boldsymbol{\eta}_0) = \sum_i \left(1 + \theta \sum_z \beta_{2z} c_{zi}\right)^2 \left(p_i^* E(\mathbf{S}^T\mathbf{S}) - p_i^{2*} E(\mathbf{S}^T) E(\mathbf{S})\right)$$

where $E(S_m S_m \mid \mathbf{C} = \mathbf{c}, \mathbf{A} = \mathbf{a}, K = k, R = 1) = f_{1mk} + 4 f_{2mk}$ and

$E(S_{m'} S_{m''} \mid \mathbf{C} = \mathbf{c}, \mathbf{A} = \mathbf{a}, K = k, R = 1) = E(S_{m'}) E(S_{m''})$ for $m' \neq m''$

given the assumed independence of the scan SNPs.

$$\mathbf{I}_{\boldsymbol{\psi}\boldsymbol{\beta}_1}(\boldsymbol{\eta}_0) = \sum_i \begin{pmatrix} p_i^*(1-p_i^*) E(\mathbf{S}) \left(1 + \theta \sum_z \beta_{2z} c_{zi}\right) \\ p_i^*(1-p_i^*) (\mathbf{c}_i^T, \mathbf{a}_i^T) E(\mathbf{S}) \left(1 + \theta \sum_z \beta_{2z} c_{zi}\right) \\ I_{K=k_i} p_i^* \left(1 + \theta \sum_z \beta_{2z} c_{zi}\right)_{2MxM} \\ 2 * I_{K=k_i} p_i^* \left(1 + \theta \sum_z \beta_{2z} c_{zi}\right)_{2MxM} \end{pmatrix}$$

$$\mathbf{I}_{\psi\psi}(\boldsymbol{\eta}_0) = \sum_i \begin{pmatrix} \left(1, \mathbf{c}_i^T, \mathbf{a}_i^T\right)\left(1, \mathbf{c}_i^T, \mathbf{a}_i^T\right)^T p_i p_i^* \left(1 - p_i^*\right) & \mathbf{0}_{(Z+W+1)x4M} \\ \mathbf{0}_{4Mx(Z+W+1)} & \begin{bmatrix} \left(\frac{1}{f_{0mk}} + \frac{1}{f_{1mk}}\right)_{2Mx2M} & \left(\frac{1}{f_{0mk}}\right)_{2Mx2M} \\ \left(\frac{1}{f_{0mk}}\right)_{2Mx2M} & \left(\frac{1}{f_{0mk}} + \frac{1}{f_{2mk}}\right)_{2Mx2M} \end{bmatrix} \end{pmatrix}$$

The complete information matrix differs from that of the simple Tukey analysis (3.17) through the increased dimensionality of both $\mathbf{S}$ and the design matrix $(\mathbf{C}, \mathbf{A})^T$. Given these results, the RTS statistic for inference is:

$T = \max_\theta T(\theta)$

for $T(\theta) = S_{\beta_1}^T\left(\hat{\boldsymbol{\eta}}_0\right)\mathbf{I}^{\beta_1\beta_1}\left(\hat{\boldsymbol{\eta}}_0\right)S_{\beta_1}\left(\hat{\boldsymbol{\eta}}_0\right)$

where $\mathbf{I}^{\beta_1\beta_1}\left(\hat{\boldsymbol{\eta}}_0\right) = \left[\mathbf{I}_{\beta_1\beta_1}\left(\hat{\boldsymbol{\eta}}_0\right) - \mathbf{I}_{\beta_1\psi}\left(\hat{\boldsymbol{\eta}}_0\right)\mathbf{I}_{\psi\psi}^{-1}\left(\hat{\boldsymbol{\eta}}_0\right)\mathbf{I}_{\beta_1\psi}^T\left(\hat{\boldsymbol{\eta}}_0\right)\right]^{-1}$ (3.23)

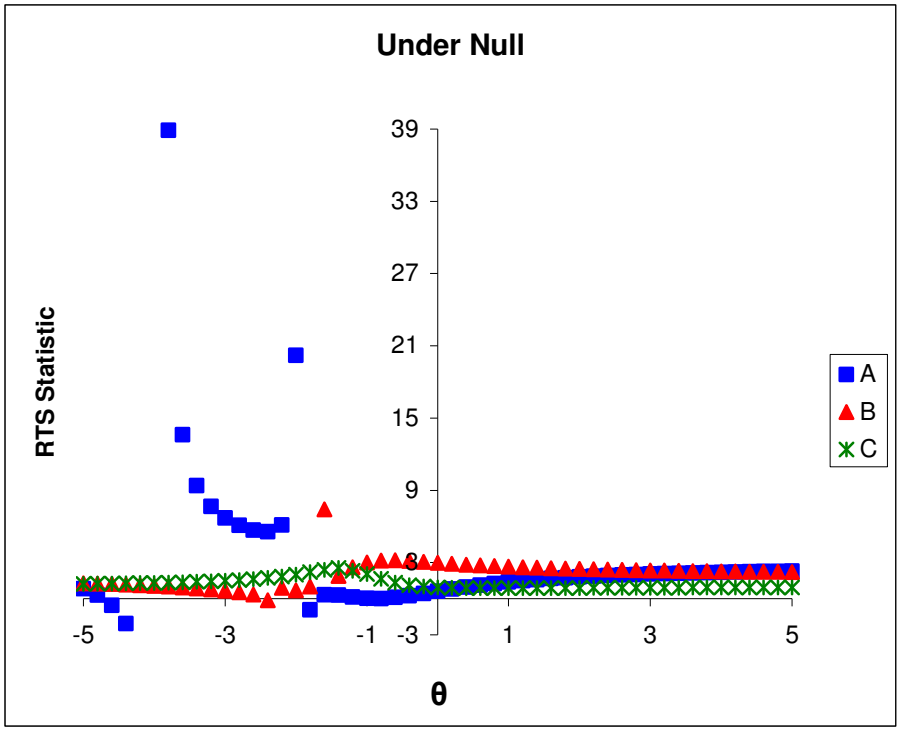
Section 3.2.2a: Theta Maximization in Full Tukey Analysis

Following a procedure similar to that in the simple Tukey analysis, I examined the θ grid over which the RTS statistic is maximized. The caption of Figure 3.2 gives a basic description of the relevant simulations. The graphs demonstrate that the full Tukey analysis is inconsistent because it can produce negative $T(\theta)$'s. I observed three scenarios: a) RTS statistics are positive for all values θ; b) RTS statistics are negative for only one value of θ; c) RTS statistics are negative for several values of θ. When the analysis involved data generated under the alternative rather than the null,

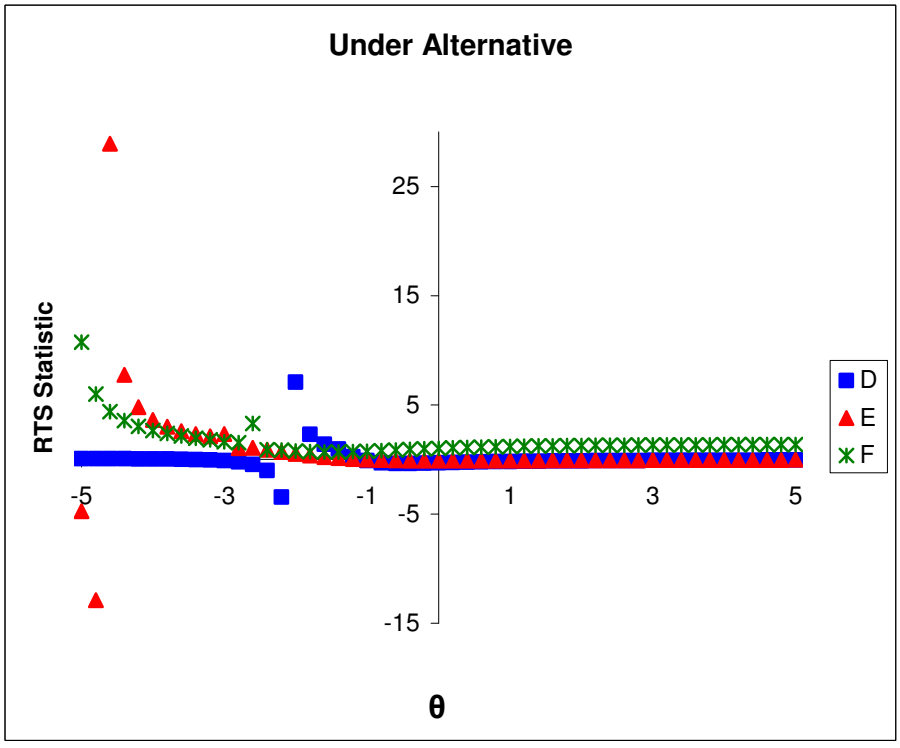
these patterns shifted, with more θ values and more simulations producing negative $T(\theta)$'s. The increased frequency of negative $T(\theta)$'s for data generated under the alternative suggests this irregularity may be due to the use of an observed information matrix. Still, an area for future work is assessing the asymptotic properties of the information matrix (3.23) through simulation.

Figure 3.2: Retrospective Tukey score test statistics over range of θ . Three representative examples (legend) are given for simulations under the null (top) and alternative (bottom) hypotheses. Simulations included a total sample size of 1000 with equal cases and controls, eight conditioning SNPs and four independent scan SNPs. The scan SNP MAFs=(0.12, 0.22, 0.28, 0.18) with ± 0.03 variation allowed between strata. Under the alternative, $\theta = 1.2$ and $\exp\{\boldsymbol{\beta}_1\} = (1.12, 1.15, 1.10, 1.13)$.

A)



B)



Section 3.2.2b: Alternate Approach for Full Tukey Analysis

An alternative to using the observed information matrix in score tests that ensures positive score test statistics is using a variance-covariance matrix based on individual score functions evaluated under the null [186]. I explore this approach. In the full Tukey analysis, the contribution of each subject to the score function evaluated under the null $\left(S(\boldsymbol{\eta}_0) = \sum_i S_i(\boldsymbol{\eta}_0) \right)$ is:

$$S_i(\boldsymbol{\eta}_0) = \begin{pmatrix} \left(\mathbf{1}, \mathbf{c}_i^T, \mathbf{a}_i^T\right)\left(d_i - p_i^*\right) \\ \left(1 + \theta \sum_z \beta_{2z} c_{zi}\right)\left(\mathbf{s}_i d_i - E(\mathbf{S})\, p_i^*\right) \\ \dfrac{I_{s_{mi}\neq 0} - f_{1mk} - f_{2mk}}{f_{0mk}} + \dfrac{I_{s_{mi}=1}}{f_{1mk}} - 1 \\ \dfrac{I_{s_{mi}\neq 0} - f_{1mk} - f_{2mk}}{f_{0mk}} + \dfrac{I_{s_{mi}=2}}{f_{2mk}} - 1 \end{pmatrix}^T$$

The alternate information matrix $\left(\mathbf{I}^*_{\boldsymbol{\eta}\boldsymbol{\eta}}\right)$ has entries satisfying the general formula:

$$\sum_{i=1}^{n_{ca}} \left(S_{i,\eta_x} S_{i\eta_y} - \overline{S_{ca,\eta_{0x}}}\,\overline{S_{ca,\eta_{0y}}} \right) + \sum_{i=n_{ca}+1}^{n} \left(S_{i,\eta_x} S_{i\eta_y} - \overline{S_{co,\eta_x}}\,\overline{S_{co,\eta_y}} \right) \tag{3.24}$$

where η_x represents a parameter in $\boldsymbol{\eta}_0$; bars indicate averages;

and data are ordered with cases ("ca") before controls ("co").

As in the simple Tukey analysis, the asymptotic representation of the information matrix is used to compute the RTS statistic:

$T = \max_\theta T(\theta)$

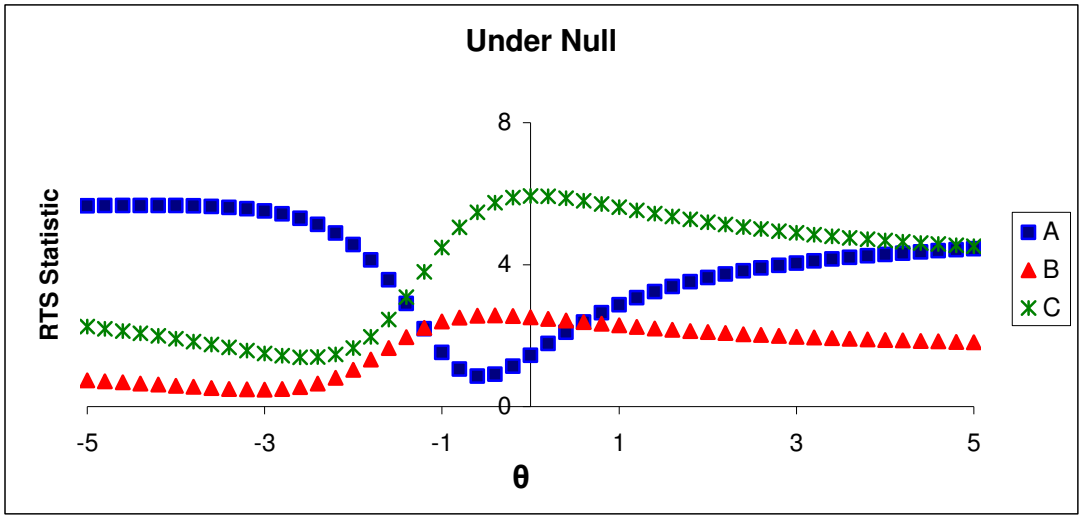
for $T(\theta) = S^T_{\boldsymbol{\beta}_1}\left(\hat{\boldsymbol{\eta}}_0\right) \mathbf{I}^{\boldsymbol{\beta}_1\boldsymbol{\beta}_1}\left(\hat{\boldsymbol{\eta}}_0\right) S_{\boldsymbol{\beta}_1}\left(\hat{\boldsymbol{\eta}}_0\right)$

where $\mathbf{I}^{\boldsymbol{\beta}_1\boldsymbol{\beta}_1}\left(\hat{\boldsymbol{\eta}}_0\right) = \left[\mathbf{I}^*_{\boldsymbol{\beta}_1\boldsymbol{\beta}_1}\left(\hat{\boldsymbol{\eta}}_0\right) - \mathbf{I}^*_{\boldsymbol{\beta}_1\boldsymbol{\psi}}\left(\hat{\boldsymbol{\eta}}_0\right) \left[\mathbf{I}^*_{\boldsymbol{\psi}\boldsymbol{\psi}}\left(\hat{\boldsymbol{\eta}}_0\right) \right]^{-1} \mathbf{I}^*_{\boldsymbol{\psi}\boldsymbol{\beta}_1}\left(\hat{\boldsymbol{\eta}}_0\right) \right]^{-1}$ (3.25)

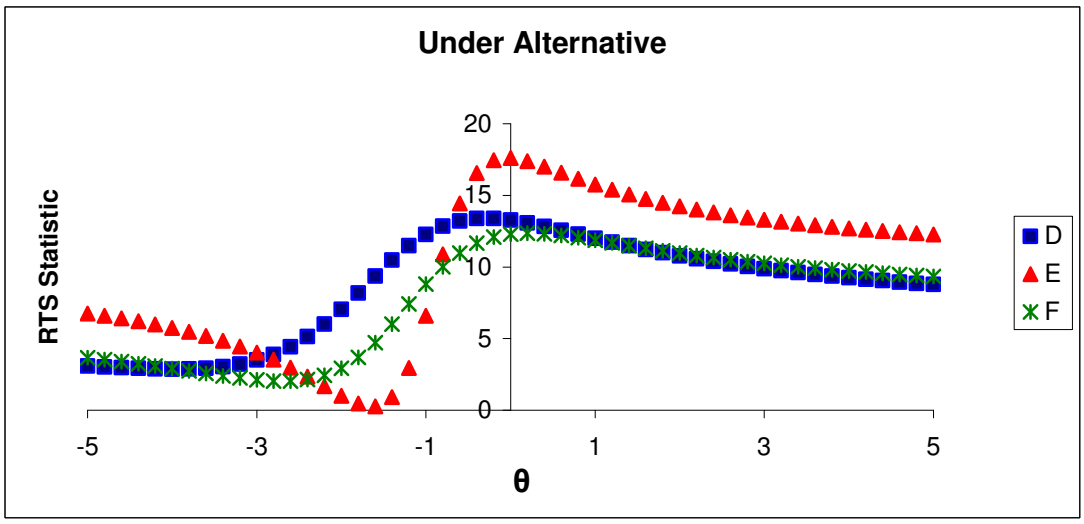
Using this formulation, I once again examined the θ grid over which RTS statistics are maximized. The results indicate that the full RTS analysis is consistent, producing only positive score test statistics, and that the θ grid is sufficiently dense for inference (Figure 3.3). Consequently, I use this alternate information matrix for both simulation studies and applied work with RTS that involve multiple scan SNPs.

Figure 3.3: Retrospective Tukey score test statistics over range of θ . Three representative examples (legend) are given for simulations under the null (top) and alternative (bottom) hypotheses. Simulations included a total sample size of 1000 with equal cases and controls, eight conditioning SNPs and four independent scan SNPs. The scan SNP MAFs=(0.12, 0.22, 0.28, 0.18) with ± 0.03 variation allowed between strata. Under the alternative, $\theta = 1.2$ and $\exp\{\boldsymbol{\beta}_1\} = (1.12, 1.15, 1.10, 1.13)$.

A)



B)



Section 3.3: Discussion

A practical consideration distinguishes the two stages of the RTS derivation. Specifically, the assumption of shared biology within a SNP set requires apriori information only when the SNP set includes multiple markers. A singular SNP set does not require its own hypothesized latent variable to validate the parsimonious interaction term in the Tukey model. Singular scan SNP sets, therefore, are well suited for exploratory analyses, such as genome scans. In contrast, more focused analyses, such as studies of candidate pathways, lend themselves to larger scan SNP sets. This distinction applies to conditioning SNP sets as well.

I compare my RTS derivation to the PTS derivation of Chatterjee et al. [175] to understand how the score test statistics are affected by the constraint of gene-gene independence in the underlying population between the scan and conditioning SNP sets. The constraint improves efficiency by reducing variance of the score function through substitution of $E[\mathbf{S}]$ for $\mathbf{S}$. A direct comparison of the score function under the null in a simple Tukey analysis makes this difference explicit:

$$S_{\beta_1}^{RTS}(\boldsymbol{\eta}_0) = \sum_i \left(1 + \theta \sum_z \beta_{2z} c_{zi}\right)\left(d_i s_i - p_i^* E[S]\right) \tag{3.26}$$

$$S_{\beta_1}^{PTS}(\boldsymbol{\eta}_0) = \sum_i \left(1 + \theta \sum_z \beta_{2z} c_{zi}\right) s_i \left(d_i - p_i^*\right) \tag{3.27}$$

The increased efficiency of RTS is reflected in the information matrix as well, with an additional substitution of $E\left[\mathbf{S}^T\mathbf{S}\right]$ for $\mathbf{S}^T\mathbf{S}$. A direct comparison for the simple Tukey analysis follows:

$$\mathbf{I}^{RTS}_{\beta_1\beta_1}(\boldsymbol{\eta}_0)=\sum_{\mathrm{i}}\left(1+\theta\sum_{z}\beta_{2z}c_{zi}\right)^2\left(p_i^{*}E\left[S^TS\right]-p_i^{2*}E^T[S]E[S]\right)$$

$$\mathbf{I}^{PTS}_{\beta_1\beta_1}(\boldsymbol{\eta}_0)=\sum_{\mathrm{i}}\left(1+\theta\sum_{z}\beta_{2z}c_{zi}\right)^2\left(p_i^{*}s_i^2-p_i^{2*}s_i^2\right)$$

$$\mathbf{I}^{RTS}_{\beta_1\boldsymbol{\beta}}(\boldsymbol{\eta}_0)=\sum_{\mathrm{i}}\left(1+\theta\sum_{z}\beta_{2z}c_{zi}\right)p_i^{*}\left(1-p_i^{*}\right)\left(\mathbf{1},\mathbf{c}_i^T\right)^T E[S]$$

$$\mathbf{I}^{PTS}_{\beta_1\boldsymbol{\beta}}(\boldsymbol{\eta}_0)=\sum_{\mathrm{i}}\left(1+\theta\sum_{z}\beta_{2z}c_{zi}\right)p_i^{*}\left(1-p_i^{*}\right)\left(\mathbf{1},\mathbf{c}_i^T\right)^T s_i$$

The substitution of $E[\mathbf{S}]$ for $\mathbf{S}$ in the score function can reduce variance substantially. In the simple case of a binary S with probability of exposure f_s, efficiency would improve by a factor of n, as $\mathrm{Var}(S)=f_s(1-f_s)$ and $\mathrm{Var}\left(E[S]\right)=\frac{1}{n}f_s(1-f_s)$.

I derived RTS assuming a multinomial distribution for the scan SNP data in terms of minor allele counts $(0,1,2)$. RTS requires that all three genotypes are observed in the sample because the inverse of $\hat{\mathbf{f}}$ is taken to compute the score test statistic. This requirement should not limit RTS analyses except in the case of very rare scan SNPs, very small studies or studies with many strata. Still, the introduction of a genotype frequency parameter that modeled scan SNP data with a binomial count $(0,1)$ is an area for future work with RTS. The modified genotype frequency parameter would apply to more than the special case of only two observed minor allele counts. It would permit RTS analysis in studies of dominant (recessive) disease models in which scan SNP data represent presence (absence) of a minor allele. Alternatively, one could introduce an assumption of Hardy-Weinberg Equilibrium into RTS analyses and use a scalar f to represent the minor allele frequency of each scan SNP.

A second area for future work on the Tukey model is specific to the analysis of a single scan SNP. It centers on a re-parameterized Tukey model:

$$\text{logit}[P(D=1 \mid S,\mathbf{C},\mathbf{A})]=\beta_0+\beta_1^* S+\sum_{\text{z}}\beta_{2z}C_z+\beta_4^*\left(S\sum_{\text{z}}\beta_{2z}C_z\right)+\sum_{\text{w}}\beta_{3w}A_w \qquad (3.28)$$

where β_1^* represents the main effect of the scan SNP and β_4^* represents the interaction between the scan and conditioning SNP sets.

This approach offers two advantages. First, it eliminates maximization of the score test statistic over θ. Second, it permits separate assessment of epistatic effects. For example, two null hypotheses that could be tested are no general disease association $\left(\{\beta_1^*,\beta_4^*\}=\mathbf{0}\right)$ and no interaction $\left(\beta_4^*=0\right)$.

A third area for future work on RTS is specific to the analysis of multiple scan SNPs. The objective is to allow for analysis of dependent scan SNPs, as may be desirable in the fine-mapping of a candidate gene. The first step is to reformulate the retrospective likelihood (3.21), replacing $\prod_m P(S_m=s_{mi} \mid K=k_i)$ with $P\left(\mathbf{S}=\mathbf{s}_i \mid K=k_i\right)$. This modification necessitates a second substitution in the log-likelihood (3.22): $\ln\left(\frac{q(\mathbf{s},k,\mathbf{f})}{q(\mathbf{0},k,\mathbf{f})}\right)$ for $\sum_m \ln\left(\frac{q(s,m,k,\mathbf{f})}{q(0,m,k,\mathbf{f})}\right)$. The score function and information matrix would need to be recalculated in this framework. One reason to prefer the current version of RTS is that $\ln\left(\frac{q(\mathbf{s},k,\mathbf{f})}{q(\mathbf{0},k,\mathbf{f})}\right)$ requires very large sample sizes to ensure that all combinations of minor allele counts for the scan SNPs are observed.

Chapter 4

Retrospective Tukey Score Test:

Evaluation

I present a variety of simulations to characterize the retrospective Tukey score test (RTS) for both simple and full Tukey analyses (Table 4.1). In general, the simple Tukey analysis includes one scan SNP (single nucleotide polymorphism) and multiple conditioning SNPs, whereas the full Tukey analysis includes four scan SNPs, multiple conditioning SNPs and a four-level factor variable for adjusting and stratification. I ran simulations under the null hypothesis of no disease association for scan SNPs $(\boldsymbol{\beta}_1 = \mathbf{0})$ to assess type I error and under the alternative to assess power. Empirical alpha levels are assessed for both the retrospective and prospective (PTS) Tukey score tests. I compare empirical power across additional methods that include three Wald tests on scan SNP parameters in standard logistic models. The first set of Wald tests evaluates the marginal effect (ME) from an unconstrained maximum likelihood analysis of a single-SNP model (UML-ME). For simulations with multiple scan SNPs, UML-ME involves only S_1. The additional Wald tests involve saturated interaction (SI) models with the general form:

$$\text{logit}[P(D=1|\mathbf{S},\mathbf{C},\mathbf{A})] = \beta_0 + \sum_m \beta_{1m} S_m + \sum_z \beta_{2z} C_z + \sum_{mz} \beta_{3mz} S_m C_z + \sum_w \beta_{4w} A_w \quad (4.1)$$

These regression models are saturated with main effects of the scan and conditioning SNPs and their pairwise interactions. They also include main effects for adjusting covariates in simulations for the full Tukey analysis. The omnibus tests on the main effects and interaction parameters of the scan SNP parameters in saturated interaction

models are based on either an unconstrained (UML-SI, Section 2.1.1) or constrained (CML-SI, Section 2.1.2) maximum likelihood analysis. CML-SI incorporates the stratification variable of the full Tukey analysis. The CML-SI omnibus test most closely resembles the RTS test of general disease association because it involves all scan SNP parameters and imposes gene-gene independence on the scan and conditioning SNPs.

Table 4.1: Basic overview of simulation studies for evaluation of retrospective Tukey score test. All simulations include 1000 subjects with equal cases and controls, unless otherwise indicated. All scan SNP data are generated under the RTS assumptions of gene-gene independence, unless otherwise indicated. Simple Tukey analyses involve a single scan SNP. Full Tukey analyses include four scan SNPs. Simulations under the Tukey model involve eight conditioning SNPs whereas simulations under models of pure epistasis involve four conditioning SNPs. Full Tukey analyses include variables for adjusting and stratification only when data are generated under the Tukey model.

Feature	**RTS Analysis**	**Model for Scan SNP Data**	**Variable Parameter**	**Methods* for Comparison**
Variance	Simple	Null TM	θ	-
Type I Error	Simple, Full	Null TM	-	PTS
	Simple	Null TM	LD of Scan and Conditioning SNPs	-
Power	Simple, Full	Alternative TM	θ, Sample Size	PTS, CML-SI, UML-SI, UML-ME
	Simple, Full	Pure Epistasis	Sample Size	PTS, CML-SI, UML-SI, UML-ME
	Simple	Misspecified TM	No. Conditioning SNPs in Interaction	-
	Full	Misspecified TM	No. Null Scan SNPs	-

*Details on methods are in introductory paragraph of this chapter.
TM=Tukey Model; LD=linkage disequilibrium; No.=number

Section 4.1.1: Design of Simulations under Tukey Model

All simulations involving the Tukey model (3.1) (Section 2.2) rather than the model of pure epistasis follow a general protocol. First, I sampled an equal number of cases and controls from Stage II of Cancer Genetic Markers of Susceptibility (CGEMS, Section 1.5). The incorporation of existing data preserves the disease association $(\boldsymbol{\beta}_2)$ and linkage disequilibrium (LD) pattern of the conditioning SNPs, as well as the disease association of the adjusting covariates $(\boldsymbol{\beta}_3)$. The conditioning SNP set for both the simple and full Tukey analysis includes eight SNPs from the extended 8q24 susceptibility region (Table 4.2). Specifically, it includes the top two markers, rs4242382 and rs6983267, as well as six additional loci in low LD (all pairwise $r^2 < 0.42$ in Stage II controls) to maximize coverage of the sub-regions and to minimize multi-collinearity in the null model. The adjusting covariates for the full Tukey analysis are indicators for study that define four strata through K (stratification variable, Section 3.2.2).

Table 4.2: Summary of 8q24 conditioning SNPs for simulation study. Results are based on CGEMS Stage II single-SNP analyses.

SNP	Risk Allele Frequency	OR	P-value
rs10505476*	0.27	1.13	3.96e-4
rs6983267*	0.50	1.24	2.22e-11
rs6999921	0.08	1.14	1.50e-2
rs1447293*#	0.38	1.16	7.24e-6
rs921146#	0.22	1.20	1.86e-6
rs13253127#	0.46	1.23	2.71e-4
rs4242382*	0.10	1.47	3.46e-15
rs6991990	0.66	1.14	1.70e-4

* conditioning SNP was used in simulations under model of pure epistasis.
# conditioning SNP was used to simulate dependence between scan and conditioning SNPs.
OR=odds ratio.

Section 4.1.1: Generation of Scan SNP Data

I generated scan SNP data under the null or alternative hypothesis, as appropriate. The method for data generation under the null hypothesis applies to all subjects. Specifically, I assigned minor allele counts through random draws from a multinomial distribution with probabilities set by Hardy-Weinberg Equilibrium [189] for a minor allele frequency f_s:

$$\left\{P(S=0)=(1-f_s)^2, P(S=1)=2f_s(1-f_s), P(S=2)=f_s^2\right\} \tag{4.2}$$

This method of generating scan SNP data applies to controls under the alternative hypothesis, as they represent the underlying population through the rare disease assumption. I repeated the procedure to simulate multiple scan SNPs, allowing the minor allele frequency of each to vary over a ± 0.03 range across strata (studies).

The method for generating scan SNP data for cases under the alternative is more involved. I give details of the method for the simple Tukey analysis, as few modifications are necessary for the full Tukey analysis. The fundamental data-generating equation incorporates the disease association of the scan and conditioning SNP sets, given by the Tukey model, as follows:

$$\begin{aligned} P(S \mid \mathbf{C}, D=1) &= \frac{OR_D(S \mid \mathbf{C}) P(S \mid \mathbf{C}, D=0)}{\sum_s OR_D(S=s \mid \mathbf{C}) P(S \mid \mathbf{C}, D=0)} \\ &= \frac{OR_D(S \mid \mathbf{C}) P(S \mid \mathbf{C})}{\sum_s OR_D(S=s \mid \mathbf{C}) P(S \mid \mathbf{C})} \qquad (4.3) \\ &= \frac{OR_D(S \mid \mathbf{C}) P(S)}{\sum_s OR_D(S=s \mid \mathbf{C}) P(S)} \qquad (4.4) \end{aligned}$$

$$\text{where } OR_D(S \mid \mathbf{C}) = \frac{P(D=1 \mid S=s, \mathbf{C})}{P(D=0 \mid S=s, \mathbf{C})} \Big/ \frac{P(D=1 \mid S=0, \mathbf{C})}{P(D=0 \mid S=0, \mathbf{C})}$$

$$= \exp\left\{ s\beta_1 \left(1 + \theta \sum_z \beta_{2z} c_z \right) \right\} \tag{4.5}$$

The equality (4.3) holds through the rare disease assumption and the equality (4.4) holds through the gene-gene independence assumption. The above formulas specify the probabilities of the multinomial distribution that I used to assign minor allele counts to the cases: $\{P(S = j \mid \mathbf{C} = \mathbf{c}_i, D = 1); j = 0,1,2\}$. In simulations to evaluate power, I varied these probabilities by altering θ in (4.5). Its value is proportional to the epistatic effect in the Tukey model.

I generated scan SNP data in simulations designed to assess the robustness of RTS to misspecification in the conditioning SNP set. I assumed the conditioning SNP set included true susceptibility markers, but I varied the number of conditioning SNPs that interacted with the scan SNP. Specifically, I varied the summation in (4.5) to include two, four, six or all eight conditioning SNPs. I randomly selected conditioning SNPs for each subset but used the same subset in the corresponding simulations. In all simulations, the null model and RTS analysis included all eight conditioning SNPs.

I generated multiple scan SNPs under the alternative by repeating the procedure for a single scan SNP. This approach is valid for the following reasons. First, the assumed independence of scan SNPs allows one to consider the conditional probability of each S_m separately. Second, the conditional odds ratio for data generation simplifies to $\exp\left\{ s_m \beta_{1m} \left(1 + \theta \sum_z \beta_{2z} c_z \right) \right\}$, as in (4.5). I varied this procedure to assess the robustness of RTS to a second form of model specification, simulating scan SNP sets with null markers.

For the simple Tukey analysis, I assessed the sensitivity of RTS to violations of the gene-gene independence assumption for the scan and conditioning SNP sets. I simulated scan SNP data under the null, using an ordered trichotomous logistic regression:

$$\text{logit}\left(P\left(S=s\,|\,\mathbf{C}\right)\right)=\exp\left(\nu_s+\boldsymbol{\lambda}_s{}^{\mathrm{T}}\mathbf{C'}\right)$$

with constraints $\nu_0=0$, $\boldsymbol{\lambda}_0=\mathbf{0}$ and $\boldsymbol{\lambda}_2=2\boldsymbol{\lambda}_1$ and $\mathbf{C'}$ a subset of $\mathbf{C}$.

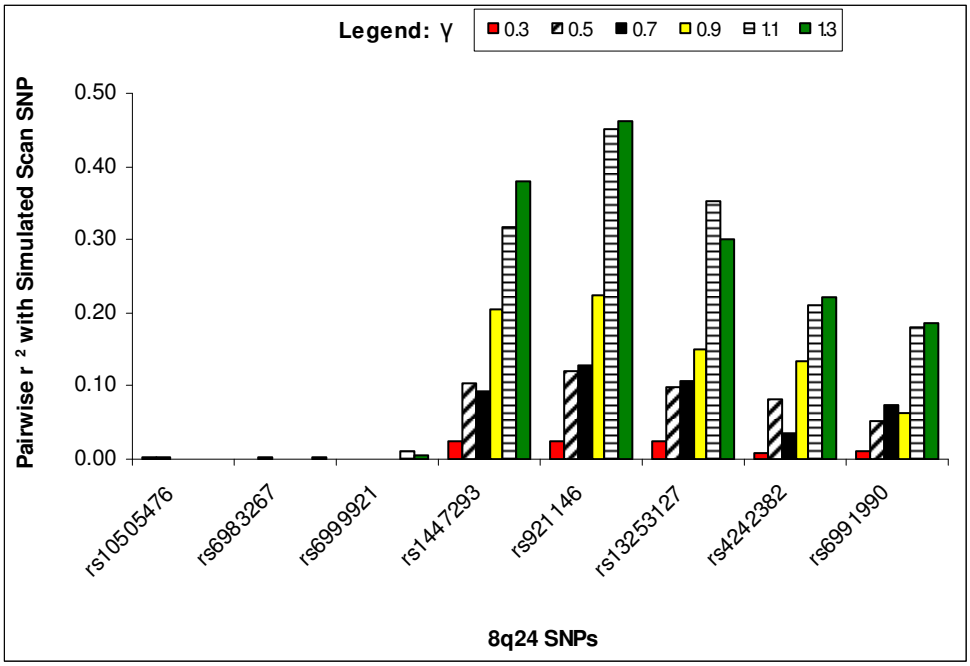
The regression model defines the probabilities of the multinomial distribution from which I randomly drew minor allele counts for all subjects:

$$P\left(S=s\,|\,\mathbf{C}\right)=\frac{\exp\left(\nu_s+s*\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{C'}\right)}{\sum_s\exp\left(\nu_s+s*\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{C'}\right)} \tag{4.6}$$

$$\text{with}\quad \nu_1=\ln\left(\frac{f_1+f_2\exp\left(\nu_2\right)}{1-f_1}\right)\quad\text{and}\quad \nu_2=\ln\left(\frac{f_2+f_1f_2}{1-f_1-f_1f_2}\right)$$

This data-generating equation depends on a subset of the conditioning SNPs because the low LD of the conditioning SNPs prevents scan SNPs from being in strong LD with even a majority of them. I selected rs1447293, rs921146 and rs13253127 for $\mathbf{C'}$ (Table 4.2) because they allow for a broad range of maximum r^2 between scan and conditioning SNPs. I varied the degree of LD between the sets through $\boldsymbol{\gamma}$ in (4.6) (Figure 4.1). I randomly sampled CGEMS subjects for these simulations from all studies other than ATBC in order to minimize potential epistatic population stratification because the minor allele frequencies of conditioning SNPs differed among controls in that study relative to all others (see for example, Figure 7.2).

Figure 4.1: Pairwise r^2 for each 8q24 conditioning SNP with simulated scan SNP. Scan SNP data were sampled from a multinomial distribution with probabilities defined by (4.3) for six γ in (4.6) (legend).



Section 4.2: Design of Simulations under Model of Pure Epistasis

Pure epistasis describes the phenomenon in which two genetic exposures affect disease risk only when both are present [189]. The relevant standard logistic regression model would include an interaction for the genetic variables but no main effects. Given the framework of the Tukey model, I set these genetic variables to be the latent variables of the scan and conditioning SNP sets.

I incorporated data from CGEMS Stage III to simulate an underlying population from which to draw cases and controls. I recorded data on four 8q24 conditioning SNPs in 9,012 controls: rs10505476 and rs6983267 of Region H (r^2 =0.35 in Stage II controls) and rs1447293 and rs4242382 of Region P (r^2 =0.18 in Stage II controls) (Table 4.2). I examined the minor allele counts of these loci to identify a genetic "risk exposure" with a frequency ~15%. The genotype 0110 in terms of presence (1) or absence (0) of a risk allele at each 8q24 locus has an exposure of 13% in the CGEMS

controls. This exposure rate corresponds to the presence of a minor allele for a causal SNP with minor allele frequency 0.07. The "0110" exposure defines the latent variable for the conditioning SNP set in both the simple and full Tukey analyses: $V_C = I_{\mathbf{C}i=0110}$.

For scan SNPs, I generated data assuming Hardy-Weinberg Equilibrium through (4.2). In the simple Tukey analysis, I set $V_S = I_{s>0}$, assuming a dominant disease model. I used a minor allele frequency of 0.30 for an exposure rate ~51%. In the full Tukey analysis, I generated scan SNPs with minor allele frequencies of 0.12, 0.26, 0.21 and 0.08. I observed a total minor allele count for scan SNPs above three in ~15% of simulated subjects and set the latent indicator variable to $V_{\mathbf{S}} = I_{\sum_m s_{mi}>3}$.

The regression model capturing the pure epistasis disease model is:

$$\text{logit}\left(P\left(D=1\right)\right) = \alpha + \beta_3^* \left(V_{\mathbf{S}} V_{\mathbf{C}}\right)$$

where α and β_3^* correspond to a disease prevalence of 1.5% (Section 6.2). This logistic model defines the probability of being a case for each simulated subject, as in (2.5). I assigned case-control status to each subject through a random draw of the corresponding Bernoulli distribution. I repeated this process until enough subjects were generated to permit sampling of 4000 cases. In each simulation to assess power, I randomly sampled an equal number of cases and controls from this underlying population and computed test statistics using data for the scan and conditioning SNPs rather than the latent variables.

Section 4.3: P-value Computation

I assessed type I error and power for two nominal alpha levels: 0.05 and 0.01. For the Wald tests, I computed p-values using asymptotic theory that a Wald test statistic follows a χ^2 distribution with degrees of freedom equal to the number of

parameters being tested [178]. For the Tukey score tests, I used three methods to compute p-values because standard theory does not apply (Section 2.2.1). The methods I used were a χ^2_{M+1} approximation, a permutation method [175] and an asymptotic approximation [190]. I constructed reference sets of 5000 null test statistics for the permutation- and asymptotic-based methods. The sets more than adequately represent the tails of the distribution, since the expected number of trials needed to observe one p-value $\leq \alpha$ is $(\alpha)^{-1}$.

The first method for p-value computation is a χ^2_{M+1} approximation, motivated by classical theory that score test statistics on parameters of a standard logistic model have χ^2 distribution with degrees of freedom equal to the number of parameters being tested under the null [182]. This method highlights the reduced degrees of freedom on a Tukey score test statistic $(M+1)$ relative to an omnibus Wald statistic $(M+MZ)$ from the analysis of a saturated interaction model (4.1). The χ^2_2 approximate p-value is: $p_{approx} = P\left(T \geq \chi^2_2\right)$. I included PTS in simulations to assess type I error because Chatterjee et al. did not consider this approximation in their report [175].

Chatterjee et al. used permutation-based p-values in PTS simulations. The method reassigns scan SNP minor allele counts rather than case-control status, preserving the disease association of the conditioning SNPs under the null.[175] A permutation p-value is: $p_{perm} = \dfrac{\#\text{ permuted } T \;\geq\; \text{simulated } T}{5000}$.

Chatterjee et al. also used an asymptotic approximation to calculate p-values in their PTS simulations [175]. The method was developed in the context of linkage analysis and has been shown to reduce the computational burden of standard

permutation methods [190]. It uses the efficient score function, which is the asymptotic, independent-identically distributed form of the score function:

$$U_{\boldsymbol{\beta}_1}(\boldsymbol{\eta}_0) = \sum_i S_{\boldsymbol{\beta}_1}(\boldsymbol{\eta}_0)_i - \mathbf{I}_{\boldsymbol{\beta}_1\boldsymbol{\psi}}(\boldsymbol{\eta}_0)_i \mathbf{I}^{-1}_{\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\eta}_0) S_{\boldsymbol{\psi}}(\boldsymbol{\eta}_0)_i \tag{4.7}$$

The efficient score function is based on a Taylor series expansion centered at the true value of $\boldsymbol{\psi}$: $S_{\beta_1}\left(\hat{\boldsymbol{\eta}}_0\right) \approx \sum_i S_{\beta_1}(\boldsymbol{\eta}_0)_i + \left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}\right)\frac{\partial S_{\beta_1}(\boldsymbol{\eta}_0)_i}{\partial \boldsymbol{\psi}}$. This approximation simplifies to (4.7) through $-S_{\boldsymbol{\psi}}(\boldsymbol{\eta}_0)\mathbf{I}^{-1}_{\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\eta}_0) \approx \hat{\boldsymbol{\psi}} - \boldsymbol{\psi}$, which holds because maximum likelihood estimates correspond to zero-value score functions. More specifically,

$$0 = S_{\boldsymbol{\psi}}\left(\hat{\boldsymbol{\eta}}_0\right)_i \approx S_{\boldsymbol{\psi}}(\boldsymbol{\eta}_0) + \left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}\right) I_{\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\eta}_0) = S_{\boldsymbol{\psi}}(\boldsymbol{\eta}_0)\mathbf{I}^{-1}_{\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\eta}_0) + \left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}\right)$$

In this asymptotic approximation, the efficient score function is modified through $U_0(\boldsymbol{\eta}_0) = \sum_i U_{\boldsymbol{\beta}_1}(\boldsymbol{\eta}_0)_i w_i$ for $W \sim N(0,1)$. Since W is independent of the data, $U_0(\boldsymbol{\eta}_0)$ is a sum of independent terms and its variance-covariance matrix can be estimated by $\mathbf{I}^{\beta_1\beta_1}(\boldsymbol{\eta}_0)$ (3.25). Using this framework, I generated an asymptotic null distribution for the Tukey score test statistics by calculating $T_{asymp} = \max_\theta \left[T_{asymp}(\theta)\right]$ for independent sets of W that altered $U_0(\boldsymbol{\eta}_0)$ in

$$T_{asymp}(\theta) = U_0^T(\boldsymbol{\eta}_0) I^{\beta_1\beta_1}(\boldsymbol{\eta}_0) U_0(\boldsymbol{\eta}_0)$$

The corresponding p-value is: $p_{asymp} = \frac{\#\, T_{asymp} \;\geq\; \text{simulated } T}{5000}$.

Section 4.4: Results

Figure legends and table captions provide details on the parameters for data generation. All results are based on 1000 simulations. The one exception is

comparison of $\mathbf{I}^{\beta_1\beta_1}$ (3.20) to empirical variance estimates for $S_{\beta_1}\left(\hat{\boldsymbol{\eta}}_0\right)$ (3.14). In each of 10,000 simulations, I recorded the score function and its asymptotic variance for a given θ to calculate the ratio of the empirical variance for $S_{\beta_1}\left(\hat{\boldsymbol{\eta}}_0\right)$ to the mean $\mathbf{I}^{\beta_1\beta_1}$. The results suggest $\mathbf{I}^{\beta_1\beta_1}$ is appropriate for the simple Tukey analysis given the range of θ studied (-5, 5) (Table 4.3).

Table 4.3: Ratio of observed to mean analytic variance of score function in retrospective Tukey score test. Values are based on 10,000 simulations under the null for six θ values and two total sample sizes with equal cases and controls. Scan SNP MAF=0.25.

	θ	**-3.4**	**-1.8**	**-0.5**	**0.7**	**1.2**	**5**
Total Sample Size	**1000**	1.00	1.02	1.06	1.03	1.02	1.01
	4000	1.01	0.99	1.02	1.03	1.02	1.01

Simulations under the null for both the simple and full Tukey analyses suggest RTS controls type I error when the independence assumption holds (Table 4.4). The χ^2_{M+1} approximation is a valid approach when the scan SNP set is singular (Figure 4.2) but deteriorates as the number of scan SNPs increases. The asymptotic approach to p-value calculations is permissible, but it can be conservative when scan SNP sets contain more than one marker. Permutation is consistently a valid method for p-value computation. Given the computational advantages, I used the χ^2_2 approximation in all simulations and applications involving a singular scan SNP set but I used permutation based p-values in work involving multiple scan SNPs. RTS does not control type I error under violations of the independence between the scan and conditioning SNP sets in the underlying population (Table 4.5).

Table 4.4: Empirical alpha levels for retrospective (RTS) and prospective (PTS) Tukey score tests. Three methods compute p-values at two nominal alpha levels in 1000 null simulations with eight conditioning SNPs. Simple Tukey analysis includes a single scan SNP with MAF=0.12 and no adjusting covariates. The full Tukey analysis includes four scan SNPs with MAFs=(0.12, 0.22, 0.28, 0.18) and a four-level factor variable for adjusting and stratification.

	Nominal Alpha	**0.05**		**0.01**	
Tukey Analysis		**RTS**	**PTS**	**RTS**	**PTS**
Simple	χ^2_2 **Approximation**	0.049	0.048	0.008	0.010
	Permutation	0.048	0.048	0.009	0.011
	Asymptotics	0.050	0.051	0.011	0.010
Full	χ^2_5 **Approximation**	0.130	0.093	0.035	0.021
	Permutation	0.047	0.051	0.009	0.008
	Asymptotics	0.043	0.047	0.008	0.006

Figure 4.2: χ^2_2 quantile-quantile plots of 1000 null retrospective (RTS, left) and prospective (PTS, right) Tukey score test statistics. Results are based on simple Tukey analysis. Total sample size is 1000 with equal number of cases and controls. Scan SNP MAF=0.12.

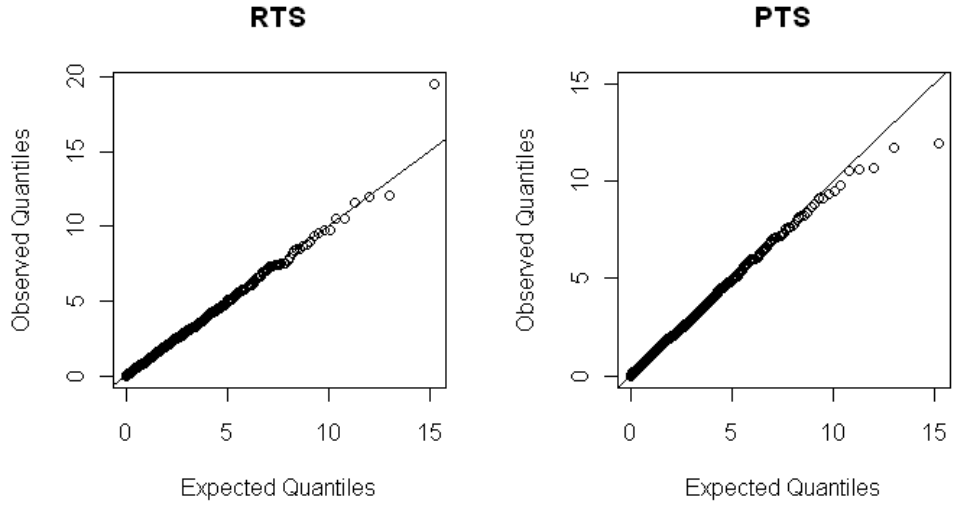


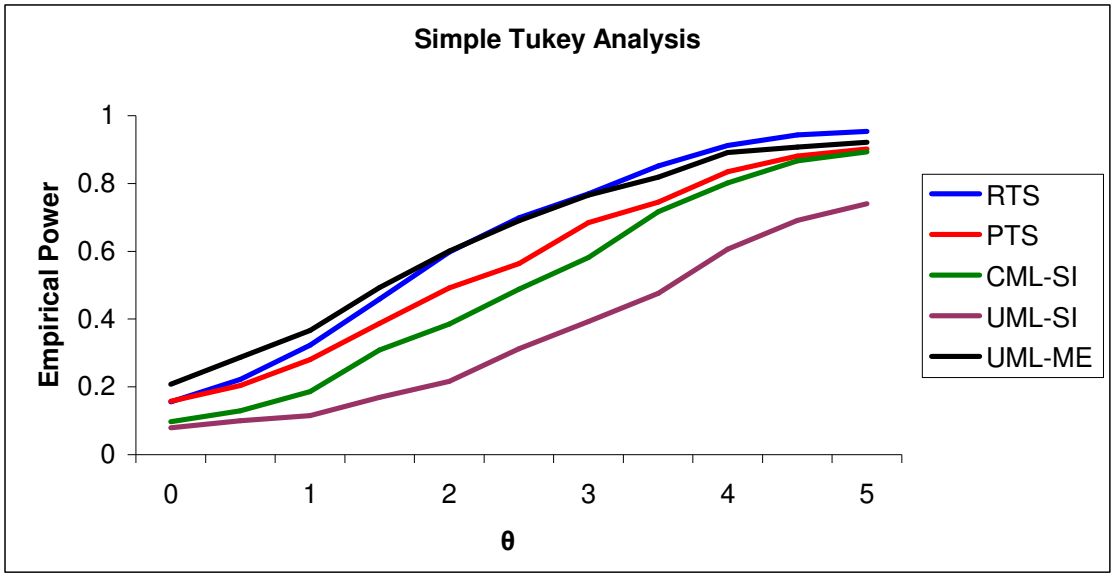
Table 4.5: Empirical alpha levels for retrospective Tukey score test when gene-gene independence assumption is violated. Results are based on 1000 null χ^2_2-approximate p-values for nominal α=0.01.

Maximum r^2 for scan and conditioning SNPs	0.05	0.12	0.25	0.31	0.44
Empirical Alpha	0.47	0.84	0.93	0.95	0.96

RTS consistently gains power over alternative methods in diverse settings (Figures 4.3-4.5). The rank order of alternative methods depends on the model for generating scan SNP data. RTS is robust to moderate misspecifications of either the scan or conditioning SNP set (Figure 4.6).

Figure 4.3: Power curves for retrospective Tukey score test and alternative methods over a range of θ values for data generation under Tukey model. θ is proportional to the epistatic effect of the SNPs. Five tests for disease association are considered: retrospective (RTS, blue) and prospective (PTS, red) Tukey score tests, a standard marginal effects Wald test (UML-ME, green) and an omnibus Wald test (main effects and interactions) based on a retrospective logistic analysis with a gene-gene independence assumption (CML-SI, purple) and a standard prospective logistic analysis (UML-SI, black) of a saturated interaction model. Each power calculation involves 1000 simulations for sample size 1000 with equal cases and controls, eight conditioning SNPs and nominal $\alpha = 0.05$. Panel A: Simple Tukey analysis includes a single scan SNP with $\beta_1 = \ln(1.15)$ and MAF=0.15. Panel B: Full Tukey analysis includes four scan SNPs with $\exp(\boldsymbol{\beta}_1) = (1.12, 1.15, 1.18, 1.20)$ and MAFs=$(0.12, 0.22, 0.28, 0.18)$ that can vary ± 0.03 across strata, as well as a four-level factor variable for stratification and adjusting.

A)



B)

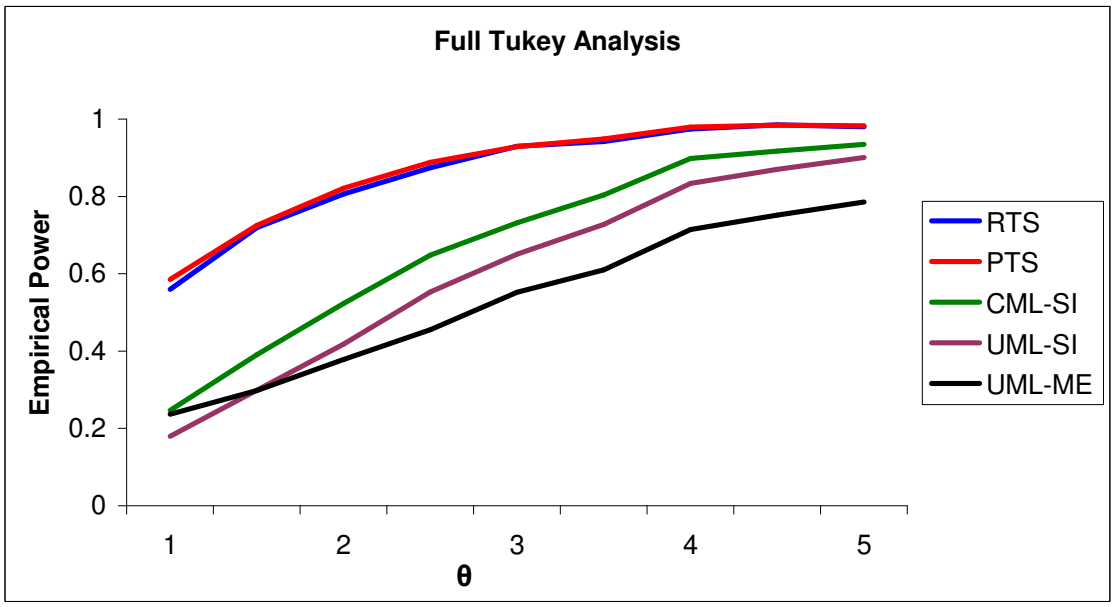
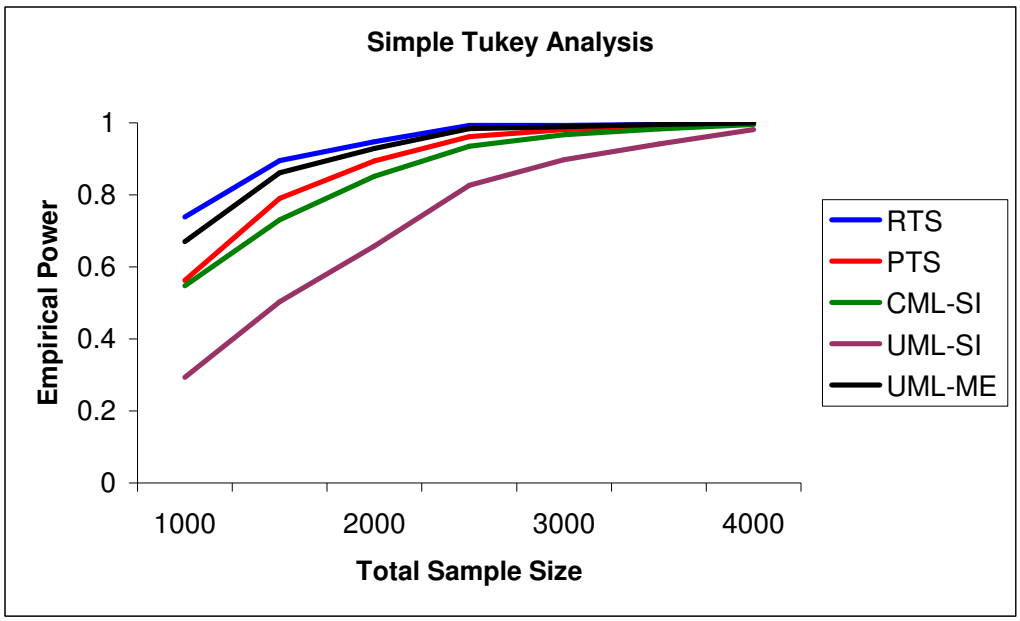


Figure 4.4: Power curves for retrospective Tukey score test and alternative methods over a range of total sample size for data generation under Tukey model. Five tests for disease association are considered: retrospective (RTS, blue) and prospective (PTS, red) Tukey score tests, a standard marginal effects Wald test (UML-ME, green) and an omnibus Wald test (main effects and interactions) based on a retrospective logistic analysis with a gene-gene independence assumption (CML-SI, purple) and a standard prospective logistic analysis (UML-SI, black) of a saturated interaction model. Power calculations involve 1000 simulations for sample size 1000 with equal cases and controls and eight conditioning SNPs. Panel A: Simple Tukey analysis involves one scan SNP with $\beta_1 = \ln(1.15)$ and MAF=0.15; $\theta = 3.5$ and nominal $\alpha = 0.01$. Panel B: Full Tukey analysis involves scan SNPs with $\exp\{\boldsymbol{\beta}_1\} = \{1.11, 1.08, 1.09, 1.06\}$ and MAFs=$(0.12, 0.22, 0.28, 0.18)$ that can vary ± 0.03 across strata; $\theta = 1.6$ and nominal $\alpha = 0.05$. Analysis also includes four-level factor variable for stratification and adjusting.

A)



B)

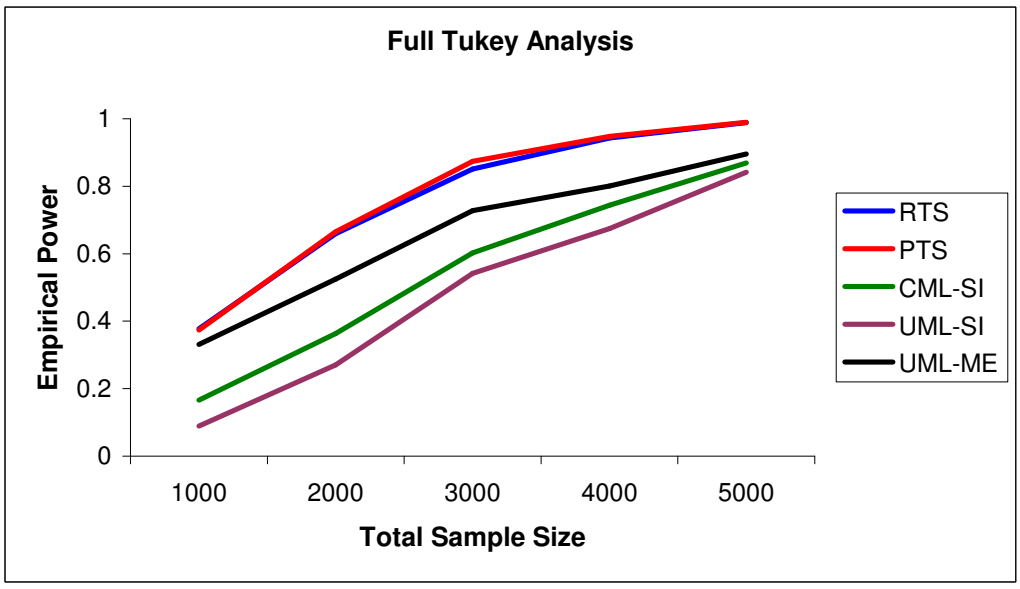
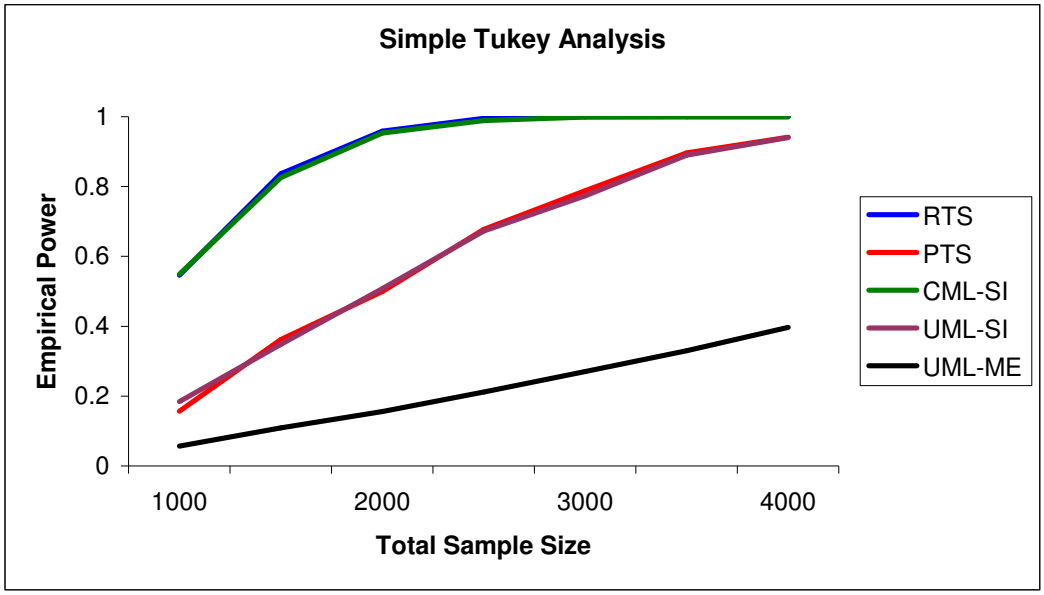


Figure 4.5: Power curves for retrospective Tukey score test and alternative methods over a range of total sample size for data generation under pure epistasis model. Five tests for disease association are considered: retrospective (RTS, blue) and prospective (PTS, red) Tukey score tests, a standard marginal effects Wald test (UML-ME, green) and an omnibus Wald test (main effects and interactions) based on a retrospective logistic analysis with a gene-gene independence assumption (CML-SI, purple) and a standard prospective logistic analysis (UML-SI, black) of a saturated interaction model. Power calculations involve 1000 simulations for sample size 1000 with equal cases and controls and an exposure based on eight conditioning SNPs with frequency ~15%. Panel A: A simple Tukey analysis assumes a dominant disease model for scan SNP with MAF=0.30, sets the marginal odds ratio for the conditioning exposure to 2.0 and uses nominal $\alpha = 0.01$. Panel B: A full Tukey analysis involves an exposure based on four scan SNPs with ~15% frequency, sets the marginal odds ratio for the conditioning exposure to 1.6 and uses nominal $\alpha = 0.05$.

A)



B)

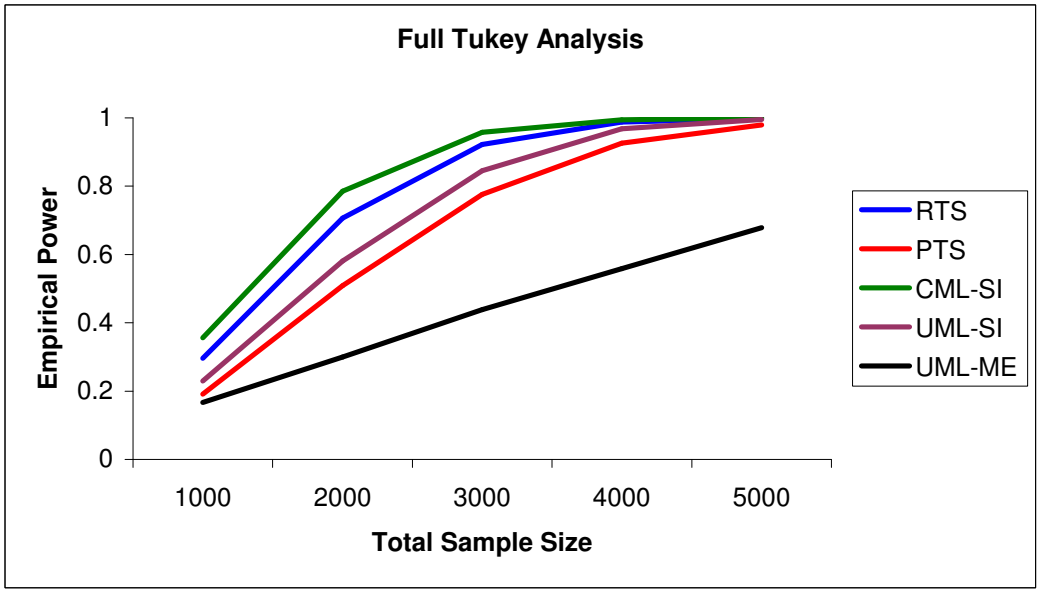
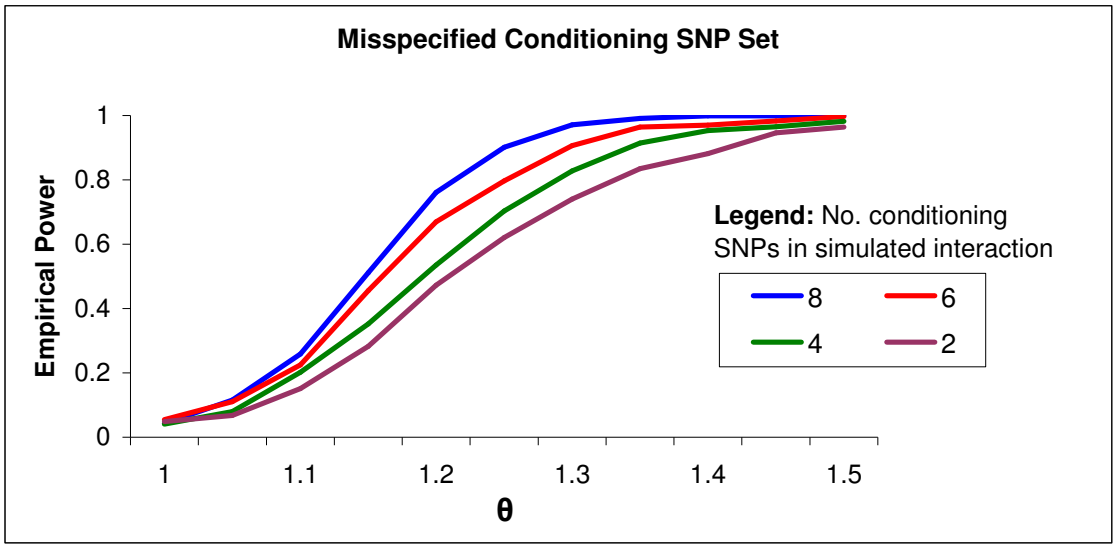
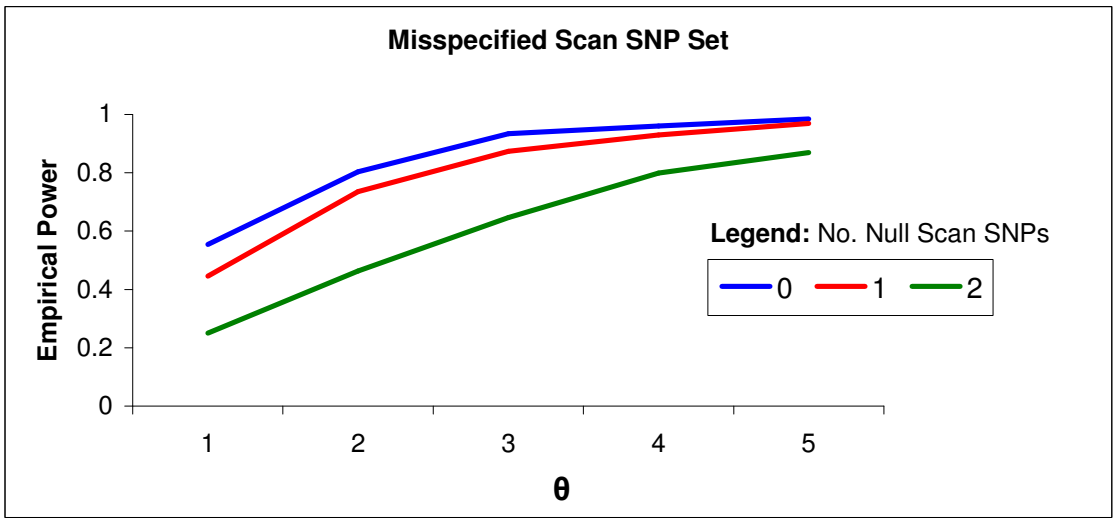


Figure 4.6: Comparison of empirical power for retrospective Tukey score test under varying levels of misspecification in the conditioning (Panel A) or scan (Panel B) SNP set. Power calculations involve 1000 simulations for sample size 1000 with equal cases and controls, eight conditioning SNPs and nominal $\alpha = 0.05$. Panel A: Scan SNP data were generated using an interaction effect with different numbers of conditioning SNPs (legend), $\beta_1 = \ln(1.15)$, MAF=0.15 and $\theta = 1.2$. Panel B: Data were generated with ≤ 2 null markers among four scan SNPs (legend), using $\exp(\boldsymbol{\beta}_1) = (1.12, 1.15, 1.18, 1.20)$ and MAFs=$(0.12, 0.22, 0.28, 0.18)$ that can vary ± 0.03 across strata. Analysis also includes a four-level factor variable for stratification and adjusting.

A)



B)



Section 4.5: Discussion

RTS is an important multilocus tool for the analysis of case-control data in genetic epidemiology. The underlying Tukey model (3.1) has several advantages. It incorporates patterns of epistasis likely to contribute to human disease, offers a parsimonious alternative to saturated interaction models (4.1) and promotes a flexible test of general disease association. RTS is the most powerful method either singularly or with one comparable alternative (Figures 4.3-4.5) in diverse settings and it controls type I error when model assumptions hold (Table 4.4). Its power profile generalizes to sample sizes commonly seen in GWAS (Figures 4.4, 4.5).

RTS gains power over PTS in a variety of settings, particularly for simple Tukey analyses (Figures 4.3a, 4.4a, 4.5a). These gains appear proportional to the epistatic effect of the scan SNP(s). For example, in the simple Tukey analysis, RTS gains power over PTS for all but the smallest values of θ that correspond to the weakest epistatic effects (Figures 4.3a). RTS and PTS have comparable power in the full Tukey analysis of multiple scan SNPs generated under the Tukey model, a setting that may correspond to a relatively weak epistatic contribution to the general disease association parameter $\boldsymbol{\beta_1}$ (Figures 4.3b, 4.4b). In contrast, RTS gains power over PTS in both simple and full Tukey analyses when scan SNP data demonstrate a large epistatic effect in a model of pure epistasis (Figure 4.5b). These findings suggest the improved efficiency of RTS over PTS has limits as epistatic effects diminish relative to main effects, particularly for larger scan SNP sets. This observation is similar the finding that constraints of gene-gene independence increase efficiency for interaction estimates more than for main effects estimates in analyses of standard logistic [152] and log-linear models [153].

RTS (PTS) gains substantial power over CML-SI (UML-SI) for analysis of data generated under the Tukey model (Figures 4.3-4.4). The increased efficiency is attributed to the reduced degrees of freedom on the score test statistics. The power gains, however, dissipate when scan SNP data are generated under a model of pure epistasis (Figure 4.5). This finding may be due to the simulation design in which the Tukey model included twice as many conditioning SNPs as the pure epistasis model (eight vs. four), allowing for a greater reduction in degrees of freedom. For example, in the simple Tukey analysis, the degrees of freedom decreased by seven (nine to two) rather than three (five to two). The comparable power of RTS (PTS) and CML-SI (UML-SI) under a model of pure epistasis could also indicate that the efficiency gains of a gene-gene independence constraint dominate those of a test statistic with reduced degrees of freedom in settings of large epistatic effects and small main effects. This interpretation follows from the generalization that the independence constraint is a specific strategy to reduce variance on interaction estimates, whereas a reduction in degrees of freedom is a general strategy to increase power. Of note, RTS consistently gains power over UML-SI which has neither a parsimonious regression model nor a gene-gene independence constraint. In practice, researchers seldom know of strong epistatic effects apriori, making the flexible Tukey score tests preferable to Wald tests on saturated interaction models, particularly in exploratory analyses.

The omnibus character of RTS is most evident in power comparisons with UML-ME [57]. For example, the rank order depends on the strength of the interaction effect. RTS loses minimal power to UML-ME only for weak epistatic effects in the simple Tukey analysis (Figure 4.3a) but UML-ME loses substantial power relative to all methods when pure epistasis generates strong interaction effects (Figure 4.5). UML-ME also loses substantial power in the full Tukey analysis with multiple scan

SNPs relative to the simple Tukey analysis with a single scan SNP. This observation reflects the potential for single-SNP analyses to overlook susceptibility SNPs when multilocus models capture the underlying disease process more closely.

RTS appears robust to misspecification of either the scan and or conditioning SNP set (Figure 4.6). I specifically consider Tukey models misspecified in terms of the number of null markers in the scan SNP set and number of non-interacting markers in the conditioning SNP set. In both settings, RTS loses substantial power only when at least half the markers in a SNP set are misspecified. This feature is particularly beneficial because an assumption of shared biology for the SNP sets may depend on scarce apriori information, particularly for SNPs in non-coding regions.

A practical consideration for implementing Tukey score tests is computing p-values. Permutation is the preferred method (Table 4.4), although a χ^2_2 approximation is valid for singular scan SNP sets (Figure 4.2). The challenge is the computational burden of permutation in large-scale studies with low significance thresholds. A general guideline for estimating p-values on the order of 10^{-x} is to use 10^{x+2} permutations. Yu et al. recently proposed an efficient algorithm to compute p-values that substantially reduces this burden, being suitable for GWAS analysis.[4] It essentially is a Markov Chain Monte Carlo (MCMC) algorithm, a standard sampling technique to approximate a complex distribution by simulating random samples [191]. The algorithm of Yu et al. builds on an innovative MCMC of Liang et al.: stochastic approximation-based Monte Carlo algorithm (SAMC). SAMC partitions the sample space for random draws into sets of monotonically increasing real numbers (sub-regions) and specifies the sampling frequency for each. Consequently, it can handle rare-event sampling more efficiently than standard MCMC or permutation.[192] Yu et al. adapt SAMC to over-sample the tails of a test statistic's null distribution. They

define SAMC sub-regions with the final set containing all permuted test statistics at least as large as the observed. In the each iteration (b), 5-10% of the data is permuted and a test statistic $t^{(b)}$ is computed using all the data. Yu et al. use a Metropolis-Hastings update [193] to determine if $t^{(b)}$ is incorporated into the series of random draws rather than $t^{(b-1)}$. They demonstrate that the algorithm reduces the number of iterations required for reliable p-value estimates so dramatically that efficiency in terms of computing time efficiency can improve 500K fold.[4]

The efficiency gains of RTS carry a risk of bias, as RTS is sensitive to violations in the assumption of gene-gene independence in the underlying population between scan and conditioning SNP sets (Table 4.5). This behavior is similar to that of case-only type methods for pairwise interaction analysis (Section 2.1.2). It motivates future work to relax the independence constraint. One approach would be to adopt methods that minimize bias in case-only type analyses of standard logistic models. In the spirit of Bhattacharjee et al. [2], one could derive RTS in the setting of conditional likelihood under the Tukey model that imposes gene-gene independence within small subsets of homogenous subjects clustered through the principal components analysis of population stratification SNPs. The first retrospective analysis of the innovative Tukey model (3.1), RTS may prove to be the starting point for research on robust methods that exploit independence in studies of modular interactions.

Chapter 5

Further Work with Tukey Model

Further work with the Tukey model (3.1) is motivated by the performance of the retrospective Tukey score test (RTS) in simulation and the literature on modular biology. In previous chapters, I indicated areas for future work involving the Tukey model. In the following sections, I address two additional areas for research.

Section 5.1: Composite Tukey Score Test

Due to the potential for bias in RTS, I constructed the composite Tukey score test (CTS) that relaxes the gene-gene independence constraint between the scan and conditioning SNP sets. CTS could have broad applications. It would be appropriate for large-scale studies that test for interaction between scan and conditioning SNP sets demonstrating varied degrees of linkage disequilibrium. It would also be appropriate for studies of gene-environment interactions for which it is often difficult to establish independence between the corresponding variable.

CTS is motivated by empirical Bayes (EB) logistic analysis (Section 2.1.3). An important methodological difference is that EB involves the weighting of point estimates from unconstrained and constrained logistic analyses, whereas CTS involves the weighting of score functions from the retrospective (3.26) and prospective (3.27) Tukey analyses. This project contributes to the field of statistics, as I am unaware of any published reports on using an empirical Bayes approach to the computation of score test statistics.

Section 5.1.1: Derivation

In this derivation, I consider a simple Tukey analysis with a single scan SNP, multiple conditioning SNPs and no variables for adjusting or stratification. The formula for the CTS score function weights the RTS and PTS score functions $\left(S_{\beta_1}^{RTS}, S_{\beta_1}^{PTS}\right)$, reminiscent of the formula for the EB pairwise interaction parameter estimate: $\hat{\beta}_{SC}^{EB}$ (2.11). The motivation for this approach is that the difference between the RTS and PTS score functions is the use of an expected or observed value for the scan SNP minor allele counts (Section 3.3). Specifically, I define the CTS score function as:

$$S_{\beta_1}^{CTS} = g\left(S_{\beta_1}^{RTS}, S_{\beta_1}^{PTS}\right) = \frac{S_{\beta_1}^{PTS}\left(S_{\beta_1}^{RTS} - S_{\beta_1}^{PTS}\right)^2 + S_{\beta_1}^{RTS} I_{PTS}^{\beta_1\beta_1}}{\left(S_{\beta_1}^{RTS} - S_{\beta_1}^{PTS}\right)^2 + I_{PTS}^{\beta_1\beta_1}} \tag{5.1}$$

The difference given by $\left(S_{\beta_1}^{RTS} - S_{\beta_1}^{PTS}\right)$ is a measure of the bias for $S_{\beta_1}^{RTS}$, and $\mathbf{I}_{PTS}^{\beta_1\beta_1}$ (3.23) is an estimate of the variance of $S_{\beta_1}^{PTS}$. For comparison, in EB, $\left(\hat{\beta}_{SC}^{CML} - \hat{\beta}_{SC}^{UML}\right)$ is the estimated bias of $\hat{\beta}_{SC}^{CML}$ and $\hat{\sigma}_{UML}^2$ is the estimated variance of $\hat{\beta}_{SC}^{UML}$. In practice, $S_{\beta_1}^{CTS}$ is computed using the observed score functions and information matrices from the retrospective and prospective Tukey analyses.

The delta method can be used to characterize $S_{\beta_1}^{CTS}$, a function of two summations for constant $\mathbf{I}_{PTS}^{\beta_1\beta_1}$. The delta method uses a Taylor expansion of a function of averages, $g(\mathbf{X})$, around $E[\mathbf{X}]$ to approximate the mean and variance of the function [180]. In this setting, the delta method estimates the variance of $S_{\beta_1}^{CTS}$ through:

$$\mathrm{Var}\left(S_{\beta_1}^{CTS}\right)=\begin{pmatrix}\dfrac{dg\left(S_{\beta_1}^{RTS},S_{\beta_1}^{PTS}\right)}{dS_{\beta_1}^{RTS}}\\ \dfrac{dg\left(S_{\beta_1}^{RTS},S_{\beta_1}^{PTS}\right)}{dS_{\beta_1}^{PTS}}\end{pmatrix}^{T}\mathrm{Cov}\left(S_{\beta_1}^{RTS},S_{\beta_1}^{PTS}\right)\begin{pmatrix}\dfrac{dg\left(S_{\beta_1}^{RTS},S_{\beta_1}^{PTS}\right)}{dS_{\beta_1}^{RTS}}\\ \dfrac{dg\left(S_{\beta_1}^{RTS},S_{\beta_1}^{PTS}\right)}{dS_{\beta_1}^{PTS}}\end{pmatrix} \qquad (5.2)$$

The first derivates are:

$$\frac{dg\left(S_{\beta_1}^{RTS},S_{\beta_1}^{PTS}\right)}{dS_{\beta_1}^{RTS}}=$$

$$\left(2S_{\beta_1}^{RTS}S_{\beta_1}^{PTS}-2S_{\beta_1}^{2PTS}+I_{PTS}^{\beta_1\beta_1}\right)*\left(S_{\beta_1}^{2RTS}-2S_{\beta_1}^{RTS}S_{\beta_1}^{PTS}+S_{\beta_1}^{2PTS}+I_{PTS}^{\beta_1\beta_1}\right)^{-1}-$$

$$2\left(S_{\beta_1}^{2RTS}-2S_{\beta_1}^{RTS}S_{\beta_1}^{PTS}+S_{\beta_1}^{2PTS}+I_{PTS}^{\beta_1\beta_1}\right)^{-2}\left(S_{\beta_1}^{RTS}-S_{\beta_1}^{PTS}\right)\left(S_{\beta_1}^{2RTS}S_{\beta_1}^{PTS}-2S_{\beta_1}^{RTS}S_{\beta_1}^{2PTS}+S_{\beta_1}^{3PTS}+S_{\beta_1}^{RTS}I_{PTS}^{\beta_1\beta_1}\right)$$

$$\frac{dg\left(S_{\beta_1}^{RTS},S_{\beta_1}^{PTS}\right)}{dS_{\beta_1}^{RTS}}=$$

$$\left(S_{\beta_1}^{2RTS}-4S_{\beta_1}^{RTS}S_{\beta_1}^{PTS}+3S_{\beta_1}^{2PTS}\right)*\left(S_{\beta_1}^{2RTS}-2S_{\beta_1}^{RTS}S_{\beta_1}^{PTS}+S_{\beta_1}^{2PTS}+I_{PTS}^{\beta_1\beta_1}\right)^{-1}-$$

$$2\left(S_{\beta_1}^{2RTS}-2S_{\beta_1}^{RTS}S_{\beta_1}^{PTS}+S_{\beta_1}^{2PTS}+I_{PTS}^{\beta_1\beta_1}\right)^{-2}\left(S_{\beta_1}^{PTS}-S_{\beta_1}^{RTS}\right)\left(S_{\beta_1}^{2RTS}S_{\beta_1}^{PTS}-2S_{\beta_1}^{RTS}S_{\beta_1}^{2PTS}+S_{\beta_1}^{3PTS}+S_{\beta_1}^{RTS}I_{PTS}^{\beta_1\beta_1}\right)$$

The variance-covariance matrix is:

$$\mathrm{Cov}\left(S_{\beta_1}^{RTS},S_{\beta_1}^{PTS}\right)=\begin{pmatrix}\mathbf{I}_{RTS}^{\beta_1\beta_1} & \mathrm{Cov}\left(S_{\beta_1}^{RTS},S_{\beta_1}^{PTS}\right)\\ \mathrm{Cov}\left(S_{\beta_1}^{RTS},S_{\beta_1}^{PTS}\right) & \mathbf{I}_{PTS}^{\beta_1\beta_1}\end{pmatrix}$$

The asymptotic representations of the information matrices (3.20) are the diagonal entries for variance. The off-diagonal entries for covariance depend on the efficient score functions (U) from the Tukey analyses (4.7) so that the estimates involve summations of independent terms. The general formula for covariance is:

$$\mathrm{Cov}\left(S_{\beta_1}^{RTS}, S_{\beta_1}^{PTS}\right) = \sum_{i=1}^{n_{ca}} \left(U_{i\beta_1}^{RTS} U_{i\beta_1}^{PTS} - \overline{U_{\beta_1,ca}^{RTS}}\,\overline{U_{\beta_1,ca}^{PTS}}\right) + \sum_{i=n_{ca}+1}^{n} \left(U_{i\beta_1}^{RTS} U_{i\beta_1}^{PTS} - \overline{U_{\beta_1,ca}^{RTS}}\,\overline{U_{\beta_1,ca}^{PTS}}\right)$$

where the data are organized with cases (ca) before controls (co).

These values define the CTS test statistic for a given θ as:

$$T(\theta) = \left(S_{\beta_1}^{CTS}\left(\hat{\boldsymbol{\eta}}_0\right)\right)^T \mathrm{Var}^{-1}\left(S_{\beta_1}^{CTS}\left(\hat{\boldsymbol{\eta}}_0\right)\right) S_{\beta_1}^{CTS}\left(\hat{\boldsymbol{\eta}}_0\right)$$

Inference is based on $T = \max_\theta \left(T(\theta)\right)$, as for RTS and PTS.

Section 5.1.2: Simulations and Discussion

I ran simulations to evaluate the performance of CTS under varying degrees of dependence for the scan and conditioning SNP sets. To advocate for CTS in large-scale studies of epistasis it should be robust relative to RTS and powerful relative to PTS. I generated scan SNP data under the Tukey model as before (Section 4.1). The one exception involves an additional set of simulations in which I generated data for a scan SNP in linkage disequilibrium with the conditioning SNPs under the alternative hypothesis, using the generating equations (4.3) and (4.6).

Before I assessed type I error and power, I examined two fundamental features of CTS. First, I investigated whether the delta method produced a valid estimate of the variance for $S_{\beta_1}^{CTS}$. I computed the ratio of empirical to analytic variance estimates for several θ values and degrees of gene-gene dependence over 10,000 simulations. The results support the use of the analytic $\mathrm{Var}\left(S_{\beta_1}^{CTS}\right)$ (5.2) in CTS (Figure 5.1). Second, I assessed the θ grid over which the CTS statistic is maximized for inference. The plots illustrate that $T^{CTS}(\theta)$ can fluctuate more than $T^{RTS}(\theta)$ between θ values on a 0.2 grid (Figures 5.2, 3.1). A unique feature of $T^{CTS}(\theta)$ is that its distribution can have

global and local minimums and maximums when gene-gene independence does not hold (Figure 5.2b). These observations suggest one direction for future work on CTS is to use the re-parameterized Tukey model that can eliminate maximization over θ in the analysis of a single scan SNP (3.28).

Figure 5.1: Ratio of observed to mean analytic variance of score function in composite Tukey score test. Values are based on 10,000 simulations under the null for six θ values and six levels of dependence between scan and conditioning SNPs (legend). Sample size was 1000 with equal cases and controls. Scan SNP MAF=0.25.

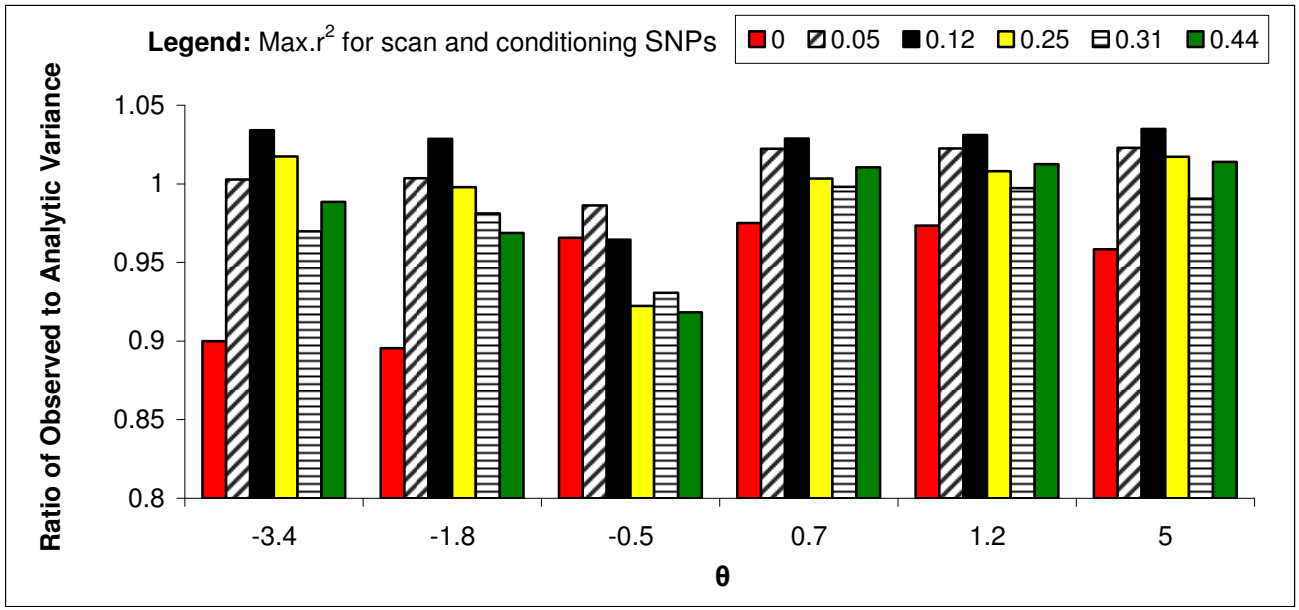
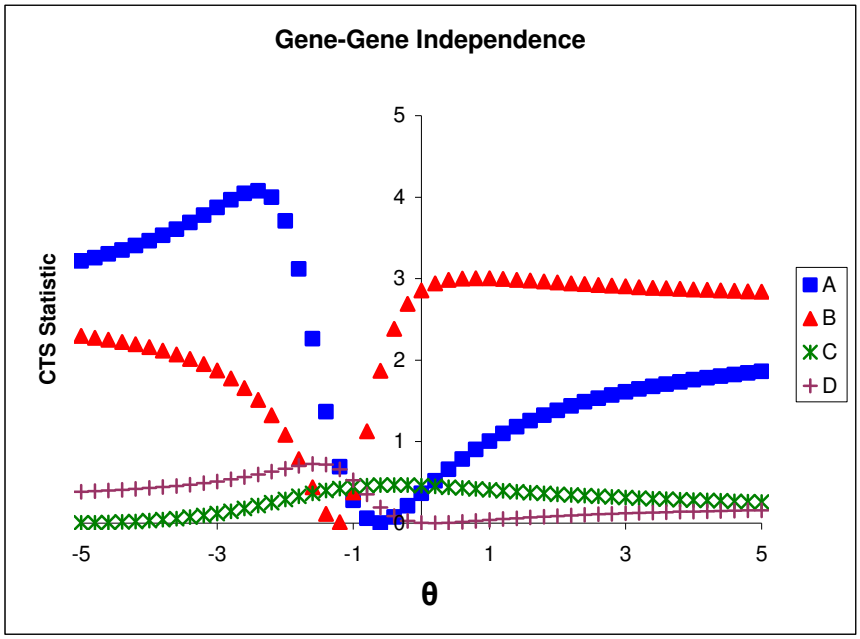
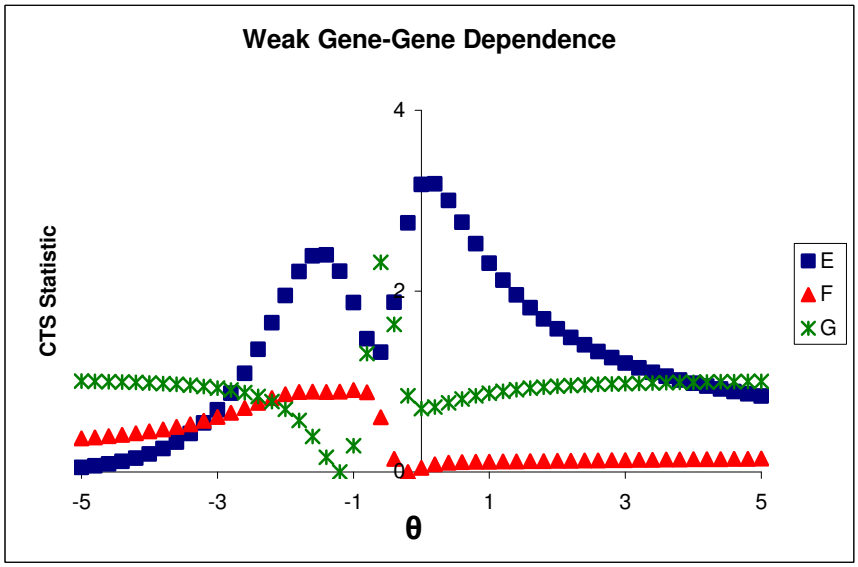


Figure 5.2: Composite Tukey score test statistics over range of θ. Representative examples (legend) are given for simulations under the null, assuming gene-gene independence (top) or weak dependence (maximum r^2=0.05, bottom) between the scan and conditioning SNP sets. Simulations include a total sample size of 1000 with equal cases and controls, eight conditioning SNPs and a single scan SNP with MAF=0.12.

A)



B)



I computed p-values though permutation (Section 4.3), given the non-standard form of the CTS statistic. I constructed a reference set of 5000 permuted CTS statistics for each linkage disequilibrium level for the scan and conditioning SNP sets to guard against differences in the null distribution of the CTS statistic. I computed summary statistics for each reference set of permuted CTS statistics (Table 5.1) and tested whether one was stochastically different from the others through the non-parametric Kruskal-Wallis test [180]. The results suggest the distribution of CTS statistics depends on whether the gene-gene independence assumption holds but not on the degree of dependence between the scan and conditioning SNPs (p-value=2.72e-08 for all reference sets; p=0.12 for only reference sets generated under gene-gene dependence). This behavior may reflect the high sensitivity of RTS to small violations of the gene-gene independence assumption. It reveals a minor obstacle to the practical implementation of CTS in large-scale studies. Specifically, one may wish to rank scan SNPs based on CTS statistics and calculate permuted p-values for only a certain percent of top results to reduce computing time. While this method is appropriate for SNP pairs that demonstrate gene-gene dependence, it will exaggerate the rank of scan SNPs that are independent of the conditioning SNP set because their critical value based on permutation is smaller (Table 5.1). This observation suggests researchers should rank scan SNPs based on permuted p-values, a requirement made less computationally intensive through the efficient algorithm for p-value computation of Yu et. al (Section 4.5) [4].

Table 5.1: Summary of composite Tukey score test statistics under varying degrees of gene-gene dependence for scan and conditioning SNP sets. Each value is based on 5000 permutations under the null.

Maximum r^2 for scan and conditioning SNPs	**Critical Value, $\alpha = 0.05$**	**Mean**	**Variance**
0.00	5.66	1.86	3.64
0.05	6.32	2.06	4.24
0.12	6.11	2.03	4.04
0.25	6.29	2.06	4.41
0.31	6.16	1.99	3.98
0.44	6.12	2.08	4.02

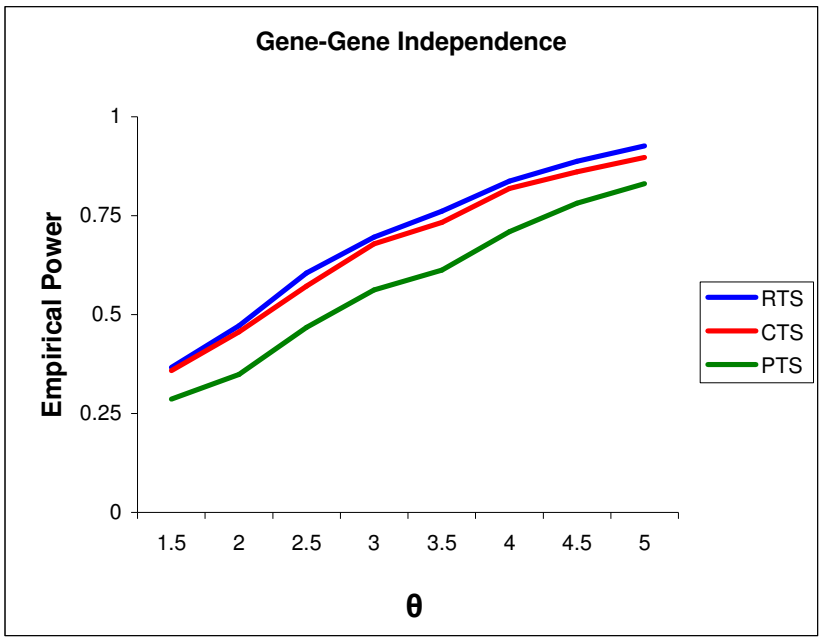
Although CTS is much more robust than RTS against violations of the gene-gene independence assumption, CTS does not control type I error as tightly as PTS (Table 5.2). The bias is similar in magnitude to that of the EB method for pairwise interaction (Section 2.1.3). It may be acceptable in an efficiency-bias trade-off, as CTS consistently gains power over PTS and loses only minimal power to RTS when gene-gene independence holds (Figure 5.3).

Table 5.2: Empirical alpha levels for composite (CTS), prospective (PTS) and retrospective (RTS) Tukey score tests under valid and invalid assumptions of gene-gene independence. Values are based on 1000 simulations for two nominal alpha levels.

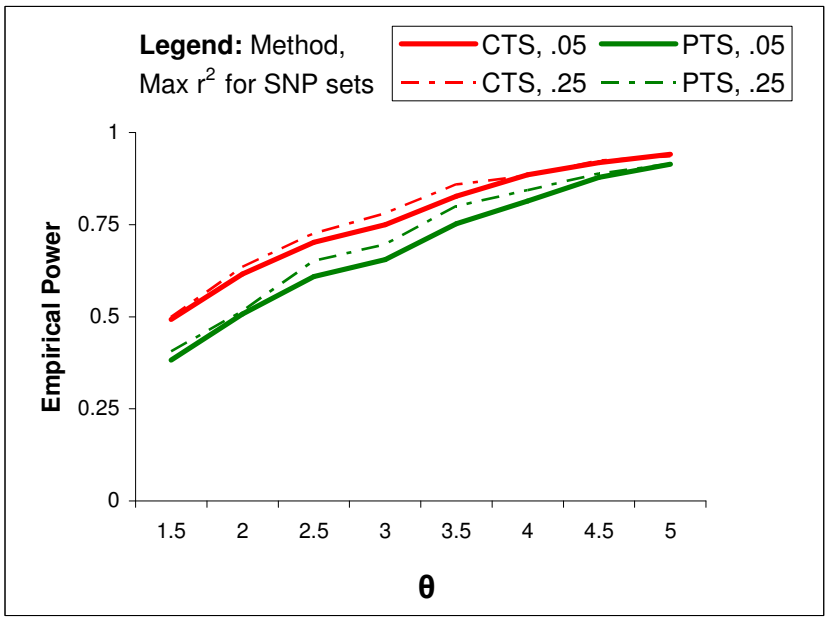
Maximum r^2 for scan and conditioning SNPs		0.00	0.05	0.12	0.25	0.31	0.44
	Nominal Alpha						
CTS	**0.05**	0.055	0.064	0.073	0.065	0.071	0.068
	0.01	0.010	0.016	0.017	0.012	0.015	0.017
PTS	**0.05**	0.049	0.051	0.049	0.049	0.048	0.047
	0.01	0.010	0.010	0.009	0.010	0.011	0.009
RTS	**0.05**	0.049	0.67	0.90	0.95	0.97	0.98
	0.01	0.008	0.47	0.84	0.93	0.95	0.96

Figure 5.3: Power curves for the Tukey score tests over a range of θ. θ is proportional to the epistatic effect of the SNPs. Each power calculation involves 1000 simulations with 500 cases and 500 controls; $\beta_1 = \ln(1.15)$; scan SNP MAF=0.15 and nominal $\alpha = 0.05$. Panel A: Retrospective (RTS), composite (CTS) and prospective (PTS) Tukey score tests are compared under gene-gene independence. Panel B: CTS and PTS are compared under two levels of gene-gene dependence between the scan and conditioning SNP sets.

A)



B)



These results suggest CTS may become a powerful and robust framework for analysis of modular epistasis, motivating future work. One area to investigate is the performance of CTS when it incorporates the score function from an RTS analysis that involves a more relaxed gene-gene independence assumption (Section 4.5). A similar topic is the bias estimate of $S_{\beta_1}^{RTS}$ in (5.1). Its current form

$$\sum_i \left(1+\theta\sum_z \hat{\beta}_{2z}\, c_{zi}\right)\hat{p}_i\left(s_i - E[S]\right) \tag{5.3}$$

does not have an intuitive interpretation in terms of uncertainty about the gene-gene independence assumption. The expectation is zero under gene-gene independence, as the final term in (5.3) is zero, and the expectation is non-zero when independence is violated, as $E[S] \neq E[S \mid \mathbf{C}]$. But, the absolute value of the estimated bias need not be proportional to the degree of dependence between the scan and conditioning SNP sets. In contrast, the estimated bias for $\hat{\beta}_{SC}^{CML}$ in EB has a clear interpretation through (2.10). One may also wish to extend CTS to the full Tukey analysis.

Decision theory is an alternative to the CTS framework that offers a similar compromise between RTS and PTS. This standard technique allows researches to assess the "risk" of basing inference on either RTS or PTS when the methods disagree. The risk is a function of the loss suffered when inference is based on the statistic that contradicts the true state of nature [180]. Examples include the use of resources in an unnecessary follow-up study or a missed opportunity to implement effective prevention policies. A measure of uncertainty on the gene-gene independence assumption is also beneficial in this approach for assigning false positive probabilities to RTS results.

Section 5.2: Bayesian Work with Tukey Model

Future work on the Tukey model could also explore strategies to relax the 1 degree-of-freedom assumption in the Tukey interaction. I am beginning work with my supervisor Dr. Chris Holmes and his fellow Dr. Joanna Davies on a Bayesian analysis for this purpose. It centers on the "Bayes-Tukey" model:

$$\text{logit}[P(D=1 \mid S, \mathbf{C})] = \beta_0 + \beta_1 S + \sum_z \beta_{2z} C_z + \sum_z \theta_z (\beta_1 S) \beta_{2z} C_z \qquad (5.4)$$

with D, S, and $\mathbf{C}$ defined as in the Tukey model (3.1) for disease outcome, scan and conditioning SNPs.

It differs from the Tukey model through the multi-dimensional $\boldsymbol{\theta}$ that replaces the scalar θ. This modification allows for more diversity in the scan SNP's interaction with the conditioning SNP set and may increase robustness to model misspecification. In practice, the Bayes-Tukey model is approximated by a linear model that improves computational efficiency. The parameter of interest in the Bayes-Tukey analysis is "interaction" parameter $\boldsymbol{\theta}$. Its maximum likelihood estimate is modeled through the prior distribution $\hat{\boldsymbol{\theta}} \sim MVN(\boldsymbol{\theta}, \boldsymbol{\lambda}\mathbf{I}_Z)$ with $\boldsymbol{\theta} \sim MVN(\mathbf{0}, \mathbf{v})$ and it is assessed through the computation of an approximate Bayes Factor [194]. In contrast, in the Tukey analyses, the 1-degree of freedom assumption sets $\boldsymbol{\lambda}$ to zero and θ is a nuisance parameter in the Tukey score tests for the general disease association parameter β_1.

Chapter 6

Genome Wide Exploration of Pairwise Interactions in Prostate Cancer

The applied work of this chapter follows naturally from the empirical evaluation of methods to detect pairwise interactions in case-control studies (Section 2.1.4). It includes a series of conditional scans that use data from each stage of Cancer Genetic Markers of Susceptibility (CGEMS, Section 1.5). My objective is to identify single nucleotide polymorphisms (SNPs) that affect prostate cancer (PRCA) risk through interaction with established susceptibility regions in order to generate hypotheses about disease etiology. This project also addresses general methodological concerns for studies of epistasis in PRCA and other phenotypes because it is one of the first interaction analyses in a multi-stage genome wide association study (GWAS).

Section 6.1: Analysis Plan

The regression models for the conditional scans are the same as in the empirical evaluation (2.2). They include one scan SNP, one conditioning SNP and their pairwise interaction. The Stage II and III analyses include indicator variables of international studies (and centers) for adjustment in the regression model and stratification in the CML analysis because geographical variables can be proxies for population stratification [159,162]. The Stage I analysis differs because the data come from a single American study. The nine conditioning regions are also the same as in the empirical evaluation (Table 1.1, Section 1.4). The same conditioning SNP applies to

all relevant scans with one exception: rs7077275 is the *CTBP2* conditioning SNP in the Stage III scan because it is the most significant regional marker in Stage III single-SNP analysis (p=4.39e-7).

Scan SNP selection differed between Stage III and earlier stages. In Stages I and II, I analyzed all possible scan SNPs. The respective sets of ~500K and ~27K SNPs set Bonferroni significance thresholds to $p \leq 1.0e-7$ and $p \leq 1.85e-6$. In Stage III, I analyzed a subset of potential scan SNPs in order to minimize the burden of multiple testing and retain the power advantages of a large sample. I limited the Stage III scan SNPs first on evidence of marginal effects in analysis of the joint data ($p<0.05$). I further limited the subset by including only the top SNP within each fine-mapping region, reducing the number of SNPs in linkage disequilibrium. These filters selected 263 SNPs for interaction analysis.

I conducted the Stage III analysis in three phases. First, I tested all pairwise interactions among the nine conditioning SNPs, using a Bonferroni threshold of $p \leq 1.11e-3$. Second, I performed a conditional scan for each of the nine susceptibility regions and the remaining 254 scan SNPs, using a Bonferroni threshold of $p \leq 1.97e-4$. Third, I tested all pairwise interactions among the 254 scan SNPs, using a Bonferroni threshold of $p \leq 1.57e-6$. For highly significant interactions involving SNPs in fine-mapping regions, I searched for nearby SNP pairs with stronger interaction signals.

I performed two sets of Wald tests on the scan SNP parameters in the regression models: interaction and omnibus (main effects and interaction). I base inference on disease association for the scan SNPs on empirical Bayes (EB) logistic analysis (Section 2.1.3) because my empirical evaluation suggests it is robust to violations of a gene-gene independence constraint and previous simulations demonstrate it gains power over standard analyses when the constraint holds at least approximately [59]. I

consider results of the unconstrained (UML, Section 2.1.1) and constrained (CML, Section 2.1.2) maximum likelihood logistic analyses to investigate how inference differs across the methods. I analyzed data from Stages I and II separately to prevent selection bias because omnibus tests assess main effect signals. In contrast, I jointly analyzed data from Stages I, II and III in the final phase to maximize on the power advantages of a large sample.

Section 6.2: Power Evaluation

I evaluated power to detect pairwise interactions at genome wide significance (Bonferroni correction) in each CGEMS stage (Stage I, 1.0e-7; Stage II, 1.85e-6; Stage III, 1.97e-4) because this project is one of the first explorations of epistasis in a multi-stage GWAS. Power calculations involve the following assumptions for simplification: a) scan and conditioning SNPs are independent; b) scan and conditioning SNPs exhibit pure epistasis (an interaction effect but no main effects); c) scan and conditioning SNPs affect disease risk through a dominant disease model; d) each sample includes an equal number of cases and controls (total n: Stage I, 2200; Stage II, 8K; Stage III, 20K). I set the probability of disease to 1.5% to reflect the prevalence among white American men [116]. I consider a variety of realistic settings over a range of marginal odds ratios (ORs) for the conditioning SNP and of minor allele frequencies (MAFs) for the scan SNP.

The power to detect an interaction in Stage II or III also depends on the power of earlier stage(s) to select the scan SNP based on its marginal effect. It is relevant, then, that interaction effects in models of pure epistasis are proportional to the SNPs' marginal effects [189]. In this framework, earlier stages would have relatively high power to detect scan SNPs that demonstrate large epistatic effects. In contrast, scan

SNPs that demonstrate large epistatic effects in models of cross-over interactions can have small or even non-existent marginal effects. A cross-over interaction describes the phenomenon in which the effect of an allele at one locus reverses in the presence of the risk allele at a second locus.[189] Early CGEMS stages that base follow-up selection on single-SNP analyses would have relatively low power to detect scan SNPs involved in cross-over interactions with conditioning SNPs.

The following terms will ease presentation of the formulas necessary for this power evaluation, which assumes a binary disease outcome (D), binary genetic variables for the scan (S) and conditioning (C) SNPs, and no adjusting covariates:

M_C = marginal OR for conditioning SNP

f_s = MAF for scan SNP	f_c = MAF for conditioning SNP
$A = P(S=1 \mid D=0) = 2f_s - f_s^2$	$B = P(S=0 \mid D=0) = 1-A$
$E = P(C=1 \mid D=0) = 2f_c - f_c^2$	$F = 1-E$
$G = P(S=1, C=1 \mid D=0) = A*E$	$H = 1-G$
$pDSC = P(D=d \mid S=s, C=c)$	

The relevant interaction parameter can be written as $\beta_{SC} = \ln\left(\frac{p111}{p011}\right) - \ln\left(\frac{p100}{p000}\right)$.

Power to detect an interaction of this magnitude at genome wide significance (α) is:

$$1 - P\left(\chi^2_{1,\mathrm{NCP}} \geq F^{-1}_{\chi^2_1}(1-\alpha)\right) \tag{6.1}$$

with non-centrality parameter $\mathrm{NCP} = \frac{\beta^2_{SC}}{\mathrm{Var}(\beta_{SC})} = \frac{\beta^2_{SC} * n/2}{\sum_{sc} \frac{1}{p0sc} + \frac{1}{p1sc}}$

and the critical value set by $F^{-1}_{\chi^2_1}$, the cumulative distribution for a central χ^2_1.

In order to calculate power, I first solved for $p111$ through:

$$M_c = \frac{(p101B + p111A)*(1 - p100B - p110A)}{(p100B + p110A)*(1 - p101B - p111A)} \tag{6.2}$$

$$p1.. = P(D=1) = \sum_{sc} p1sc * P(S=s \mid D=0) * P(C=c \mid D=0) \qquad (6.3)$$

Equation (6.2) simplifies because the pure epistasis model has no main effects, setting $pd00 = pd01 = pd00$ for all values of d. With algebra (6.3) gives the identity $p100 = \frac{p1.. - G * p111}{H}$. These equalities allow for substitutions in (6.2) that yield a quadratic formula to solve for $p111$ and, by extension, $p100$:

$$0 = p^2 111\left(\frac{GB}{H}\right)$$

$$+ p111\left(A + \frac{(M_C - 1)Ap1..}{H} - \frac{(B - M_C)G}{H} + \frac{(1 - M_C)2BG * p1..}{H^2}\right)$$

$$+ \frac{p1..}{H}\left(B - \frac{B * p1..}{H} - M_C + \frac{BM_C * p1..}{H}\right)$$

These calculations are sufficient for only Stage I, as they do not consider power of previous stages to select SNPs for follow-up. The overall power for Stages II and III is the product of the power to select the SNP for its marginal effect in all preceding stages and the power to detect the interaction in the stage at hand. This approach is permissible even for the joint analysis in Stage III because the marginal and interaction parameters of the scan SNP are asymptotically independent when the candidate interacting SNPs are independent [195]. The previous calculations for p111 and p100 set the marginal effect to:

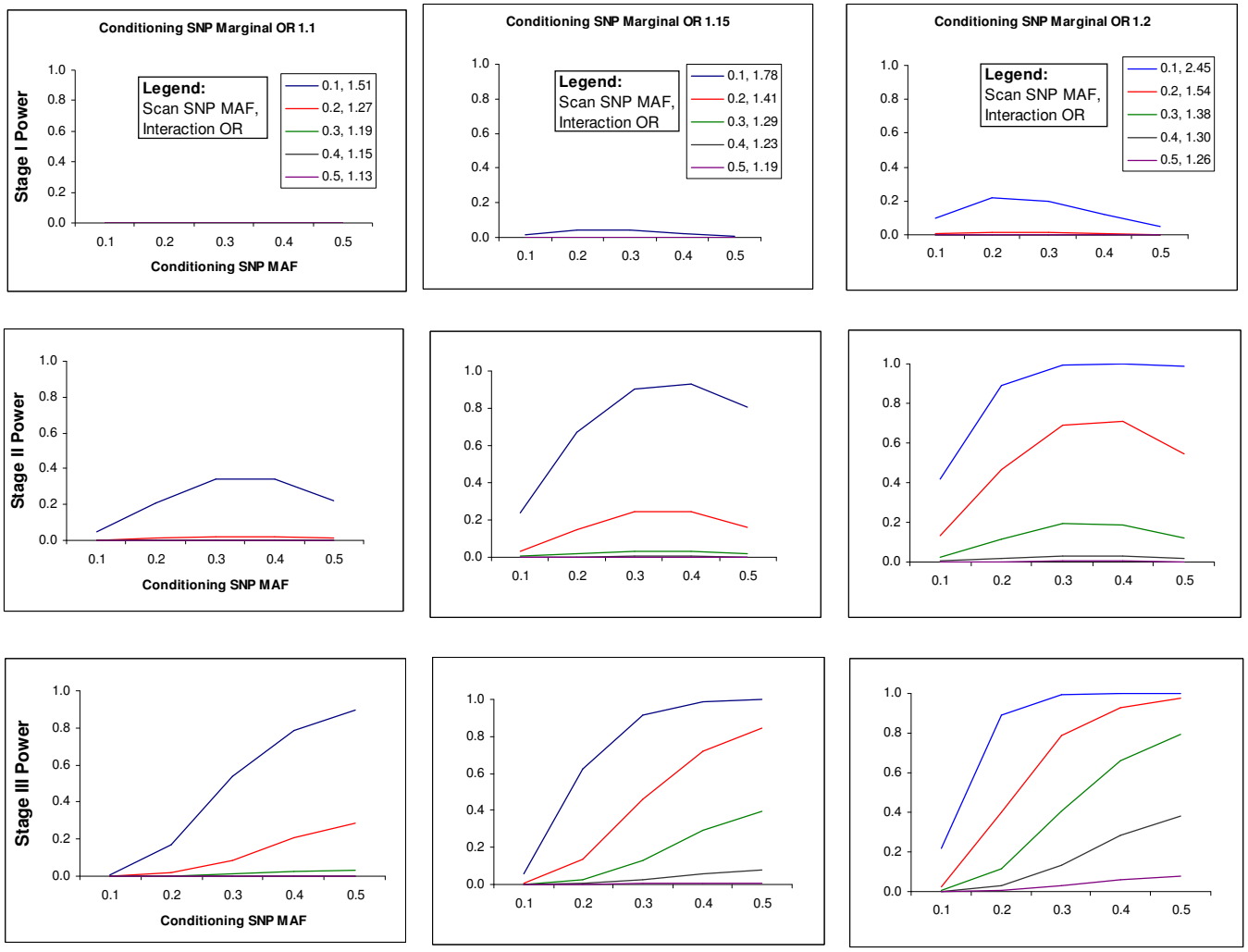
$$\beta_S = \ln\left[\frac{(p110F + p111E) * (1 - p100F - p101E)}{(p100F + p101E) * (1 - p110F - p111E)}\right]$$

This value determines power through (6.1) for the appropriate significance thresholds (Stage I, 0.05; Stage II, 0.001) and non-centrality parameter: $\frac{\beta_S^2 * n/2}{\sum_s \frac{1}{p0s.} + \frac{1}{p1s.}}$.

Section 6.3: Results

Figure 6.1 summarizes the findings on statistical power. Stage I of CGEMS has virtually no power to detect interactions in settings of multiplicative odds ratios in the range $(1.13, 2.05)$. In contrast, CGEMS Stages II and III have high power to detect modest to large interactions (OR > 1.7) for common conditioning SNPs (MAF > 20%), even after accounting for the power lost through selection of SNPs on marginal effect in earlier stages. Power is not monotone over scan SNP minor allele frequency because, in a model of pure epistasis with a fixed marginal odds ratio for the conditioning SNP, the minor allele frequency of the scan SNP is inversely proportional to the interaction odds ratio.

Figure 6.1: Power to detect interactions in CGEMS at genome wide significance. Results are for Stages I (top, p≤1.0e-7), II (middle, p≤1.85e-6), III (bottom, p≤1.97e-4). Conditioning SNP marginal ORs are 1.1 (left), 1.15 (middle), 1.2 (right). Legends apply to columns.



In the analysis of Stage I data, one SNP demonstrates genome wide significance for interaction. The conditioning SNP is 8q24 Region H and the scan SNP is rs2002865, intergenic to *DPP6* and *PAXIP1* on chromosome 7 (p=9.14e-10). The signal is driven by CML (p=9.53e-10) but also highly significant in the UML analysis (p=1.05e-5), conferring robustness to the finding. A second SNP, rs4960563, in strong linkage disequilibrium with rs2002865 (r^2 =0.68 in controls), is highly significant for interaction with the same 8q24 conditioning region (p=3.79e-6). Both interactions failed to replicate in a subset of the CGEMS Stage II sample (2439 cases and 2241 controls) that were additionally genotyped.

In the Stage II conditional scans, no SNP demonstrates genome wide significance for interaction. A list of top ranking SNPs (p<1.0e-4) from each conditional interaction scan is shown in Table 6.1. The most notable finding for biological plausibility and statistical evidence is an interaction between rs6983267 in 8q24 Region E and rs4953347 in the first intron of *EPAS1* (also known as *HIF-2A*) on chromosome 2 (p=9.69e-5, Table 6.2). Its signal is driven by CML (p=9.28e-5) but also significant in the UML analysis (p=3.05e-3). The empirical joint odds ratios suggest each rs4953347 minor allele attenuates the effect of each rs6983267 risk allele (Figure 6.2). The marginal p-value of rs4953347 is not statistically significant (p=0.38).

Table 6.1: Summary of SNPs with interaction p-values < 1.0e-4 in empirical Bayes logistic analysis of CGEMS Stage II data.

	Interacting Region				Multiplicative Interaction	
Conditioning Region	**SNP**	**MAF**	**Chr, Nearby Gene**	**Marginal* OR (95% CI)**	**OR (95% CI)**	**P-value**
CTBP2	rs7765379	0.12	6	0.99 (0.90, 1.09)	0.80 (0.71, 0.89)	9.83E-05
EEFSEC	rs12489404	0.42	3, *GRM7*	0.99 (0.93, 1.05)	0.85 (0.79, 0.91)	3.10E-06
	rs10458466	0.49	1	0.97 (0.91, 1.03)	1.15 (1.07, 1.24)	6.52E-05
HNF1B	rs10506678	0.43	12	1.00 (0.94, 1.07)	1.14 (1.07, 1.22)	5.35E-05
	rs617182	0.46	17, *PHOSPHO1*	0.95 (0.89, 1.01)	1.15 (1.07, 1.23)	7.14E-05
	rs4691238	0.41	4	0.95 (0.90, 1.02)	1.14 (1.07, 1.22)	8.37E-05
JAZF1	rs745720	0.19	2	0.96 (0.88, 1.03)	1.23 (1.12, 1.35)	2.28E-05
	rs2899748	0.35	15, *GLCE*	0.96 (0.90, 1.02)	0.84 (0.77, 0.91)	3.04E-05
MSMB	rs10935317	0.42	3	0.98 (0.92, 1.05)	1.14 (1.07, 1.22)	3.93E-05
	rs12605415	0.38	18	0.99 (0.93, 1.06)	1.14 (1.07, 1.22)	6.20E-05
	rs11083271	0.32	18	1.02 (0.96, 1.09)	1.16 (1.08, 1.24)	6.67E-05
	rs9880831	0.31	3, *LOC391524*	1.04 (0.97, 1.12)	0.87 (0.82, 0.93)	6.90E-05
	rs7921651	0.23	10, *FAM107B*	1.03 (0.97, 1.12)	0.86 (0.79, 0.93)	8.62E-05
8q24 Region P	rs4660403	0.21	1, *LOC388621*	0.93 (0.86, 1.01)	1.24 (1.12, 1.38)	7.57E-05
8q24 Region E	rs4953347	0.48	2, *EPAS1*	0.97 (0.91, 1.04)	1.13 (1.06, 1.21)	9.69E-05
8q24 Region H	rs11589338	0.10	1, *TMCO4*	0.94 (0.84, 1.05)	1.30 (1.15, 1.47)	1.57E-05
11q13	rs1240224	0.30	12	1.03 (0.96, 1.10)	1.15 (1.07, 1.23)	9.00E-05

*Marginal results are based on single-SNP analyses. Chr=chromosome; OR=odds ratio; CI=confidence interval**.**

Figure 6.2: Empirical joint odds ratios for rs4953347 in *EPAS1* and rs6983267 in 8q24 Region E. Interaction p-value is 9.69e-5, ranking first among 27,226 SNPs in the conditional scan. Bold values indicate odds ratios significant at the 5% level.

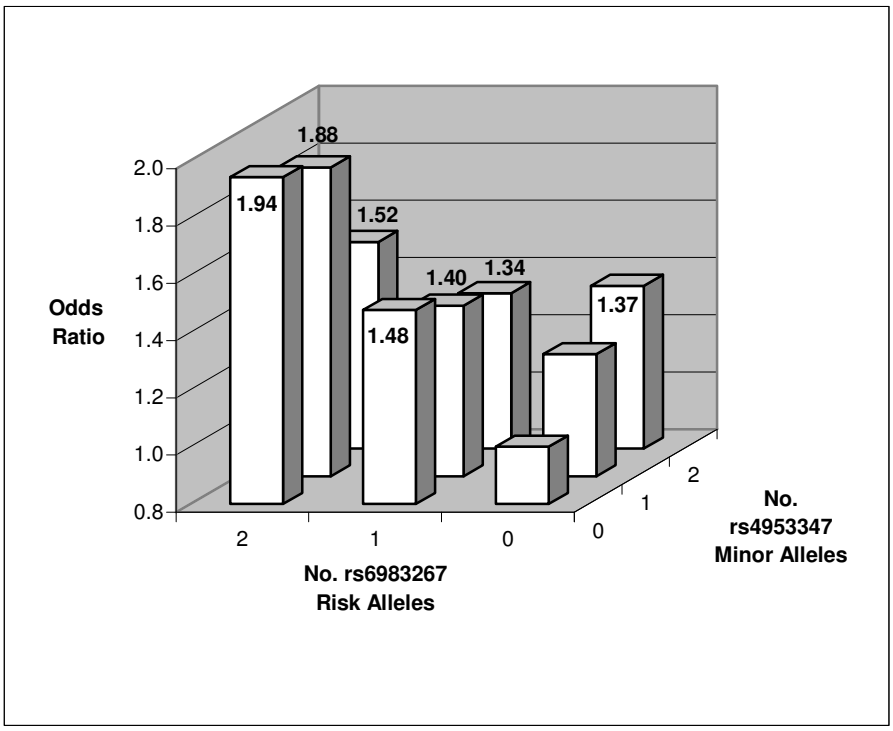


Table 6.2: Results of empirical Bayes analysis of logistic model for pairwise interaction between rs4953347 in *EPAS1* and rs6983267 in 8q24 Region E.

Covariate	OR	95% CI	P-value
rs6983267	0.72	(0.66, 0.78)	5.48e-14
rs4810671	0.87	(0.80, 0.95)	1.33e-3
rs6983267 : rs4810671	1.13	(1.06, 1.21)	9.69e-5

A colon designates an interaction.
OR=odds ratio; CI=confidence interval.

The Stage III results highlight one SNP pair for interaction in each phase of the analysis. In the first phase, the top pair of conditioning SNPs for interaction includes rs620861 of 8q24 Region H and rs4242382 of 8q24 Region P (p=7.90e-3). This result was published recently [3]. I searched for stronger evidence of epistasis between all SNP pairs in these fine-mapping regions (66 SNPs in Region P, 57 in Region H). The top SNP pair of rs6986543 and POU5F1P1-173 is significant after adjusting for multiple testing (p=4.27e-7). The signal is driven by CML (p=2.52e-7) but also significant in the UML analysis (p=7.33e-4). In the second analysis phase, the *MSMB* conditional scan highlights a biologically interesting scan SNP: rs4656538 near *POU2F1* on chromosome 1 (p=0.02, rank 2). The CML result is also significant but the UML signal is much stronger (p=3.65e-3). In the final analysis phase, the second overall SNP pair for interaction includes rs9920331 in an intron of *NR2F2* on chromosome 15 and rs1859962 in a gene desert on chromosome 17 (p=8.28e-5). rs1859962 demonstrates genome wide significance for marginal effect in Stage III (p=9.78e-18). rs9920331 demonstrates a statistically significant interaction with 8q24 Region H as well (p=6.01e-3, rank 2). Both interaction signals for rs9920331 are driven by CML but also significant in the UML analysis.

The CGEMS selection for follow-up based on marginal effects excluded 15,917 SNPs after Stage I and 271 SNPs after Stage II that demonstrate evidence of disease association in omnibus testing (Stage I $p<0.005$, Stage II $p<0.001$). However, omnibus testing revealed no new susceptibility regions.

Section 6.4: Discussion

Although the literature includes reports on large-scale interaction analyses [196], this project is the first published exploration of gene-gene interactions in the setting of a multi-stage GWAS to my knowledge [1]. It provides insight into both the possible role of gene-gene interactions in the etiology of PRCA and methodological challenges that large-scale studies of epistasis in other traits will face. With respect to the later, my power evaluation suggests that very large sample sizes are required to detect pairwise interactions at GWAS significance. The exploration of gene-gene interactions in smaller studies is unlikely to lead to definitive findings but can be useful in generating lists of SNP pairs for replication studies. This project provides a list of SNPs that warrant follow-up study with varying degrees of priority (Table 6.1). All top EB results are significant in the CML and UML analyses. This consistency demonstrates robustness in the findings. It also suggests inference on disease association can be similar between CML and UML analyses, although my empirical evaluation revealed discordant ranks for scan SNPs based on interaction p-values in the CML and UML analyses (Figure 2.3).

The omnibus tests did not detect novel genome wide significant susceptibility SNPs. However, a large number of SNPs excluded from follow-up in CGEMS due to weak marginal effects have omnibus signals that surpass the significant threshold for follow-up selection. This observation demonstrates the potential of multilocus analyses to alter the trajectory of multi-stage GWAS. Researchers could adopt a two-stage selection process. First, a genome scan selects a portion of top SNPs for follow-up based on single-SNP analysis. Second, a series of conditional scans that involve SNPs with strong marginal effects selects a set of top scan SNPs for follow-up based on omnibus testing. A similar strategy has been advocated previously [40].

The interaction results of this project should be interpreted with regard to the power calculations, particularly because they involve conditioning SNPs with comparable marginal effects and minor allele frequencies (Figure 6.1). The power evaluation suggests Stage I is underpowered to detect SNPs that affect PRCA risk through pure epistasis. In contrast, Stages II and III are well powered to detect large multiplicative effects. The scarcity of genome wide significant results may indicate that, if pure epistasis contributes to PRCA risk, it is through more modest multiplicative effects that would correspond to the range of odds ratios detected in the analysis. I acknowledge that this project may have missed scan SNPs involved in strong cross-over effects with the susceptibility regions, as they may have been excluded from CGEMS follow-up due to weak marginal effects. This project may also have missed higher order interactions through its focus on pairwise epistasis. The task of searching for such complex interactions is a computationally daunting task with a great multiple testing burden in large-scale studies. Investigation of higher order interactions through logistic analysis involves a regression model that quickly becomes saturated with interaction terms. One strategy to overcome this challenge is sequential modeling of higher order interactions and simultaneous testing of all interactions not statistically significant in simpler models [197]. An alternative to logistic analysis is multifactor-dimensionality reduction, a model-free approach that uses cross-validation and permutation testing to assess the ability of a one dimensional predictor (for example, a genotype) to predict disease status [198].

The one genome wide significant result in this application failed to replicate. It involves 8q24 Region H and rs2002865 in Stage I. The result is unlikely to be due to genotyping error as a second SNP in strong linkage disequilibrium with the original signal also showed strong significance. The result was interesting biologically because

the conditioning region is also associated with breast cancer and the scan SNP is located near *PAXIP1*, whose protein shares features with the protein of the well characterized breast cancer susceptibility gene *BRCA* [199]. Its failure to replicate underscores the challenges of employing rank p-values to prioritize interaction results and of establishing significance thresholds for conclusive findings. These results reinforce the need for methods that consider both the power to detect a SNP and its biological plausibility. I briefly review two.

The False Positive Report Probability (FPRP) is defined as the probability that a SNP is a null marker conditional on a significant test statistic. In an unbiased study, the value is determined by the test's power (ω), its size (α) and the prior probability on the alternative hypothesis (π). The relevant formula is:

$$\text{FPRP} = \frac{\alpha(1-\pi)}{\alpha(1-\pi)+(1-\omega)\pi}$$

Wacholder et al. advise that power calculations be based on a fixed odds ratio (for example 1.5) under the alternative and that the observed p-value be used for α. Their simulation studies demonstrate it can be challenging to categorize a statistically significant SNP as a true positive when π is very low. The results also suggest the benefit to FPRP from increases in sample size has a practical limit, particularly for studies with low π.[200] These findings are relevant to GWAS in which most SNPs are expected to be null.

Wacholder et al. discuss practical applications of FPRP. They acknowledge that one obstacle to its broad implementation is hesitance among scientists to specify π. They advise that FPRP be reported over a range of π, allowing external researchers to base inference on a different π, should they disagree with the publishing authors. Wacholder et al. similarly advocate flexibility in FPRP interpretation. They

note that a universal threshold for FPRP is not practical given the varied designs, purposes and consequences of genetic studies. They contrast definitive and exploratory studies as examples of a setting in which either a stringent or relaxed criterion is appropriate.[200]

Wakefield introduced the Bayesian False Discovery Probability that has the same objective as FPRP but incorporates information and establishes thresholds differently. It requires investigators to define a prior distribution for the odds ratio under study, whereas Wacholder et al. assign it a fixed value. Wakefield uses that prior distribution to compute a posterior probability for the null, which is intrinsically linked to false positive probabilities. Wakefield defines an approximate Bayes Factor and uses a cost analysis to establish a flexible threshold for the "noteworthiness" of a significant finding. Wakefield argues that a significant finding should be labeled noteworthy only if the posterior expected cost of a false negative exceeds that of a false positive.[201]

Perhaps the most noteworthy result of our study is the top SNP for interaction with 8q24 Region E in Stage II: rs4953347. This SNP is an intronic variant of *EPAS1*, which is biologically interesting both individually and in conjunction with 8q24 Region E. The gene is a member of the hypoxia-inducible factor family that has been shown to promote key carcinogenic processes, including angiogenesis and metastasis [202]. Under hypoxic conditions that are common in tumor micro-environments, EPAS1 directly binds *POU5F1* and activates its expression [203,204]. By activating *POU5F1*, EPAS1 has been shown to promote tumorigenesis [205]. Both *EPAS1* and *POU5F1B* are over-expressed in PRCA, but *POU5F1* is not expressed in healthy or malignant prostate tissue [120,202]. Given these data, I propose that *POU5F1B* mediates the observed *EPAS1*-8q24 Region E interaction. This hypothesis makes an assumption that

EPAS1 participates in the regulation of *POU5F1B*, which is currently poorly understood. Of note, a regulatory element containing the conditioning SNP rs6983267 has been described [109,112]. Its risk allele increases local binding affinity of β-catenin-TCF complexes up to fourfold. Although studies have focused on TCF4, it is possible that the results generalize to the β-catenin-TCF3 complex [110] that promotes *POU5F1* expression [206] and shares one of two *POU5F1* promoter regions with EPAS1.

Given the literature, I further propose that the PRCA association of 8q24 Region E involves a type of pluripotency network centered on *POU5F1B* rather than on *POU5F1* (Figure 7.8). Over-expression of *POU5F1B* in PRCA may mimic ectopic *POU5F1* expression because the genes produce proteins with similar function [121]. It may also promote growth of PRCA tumors, which typically originate in epithelium [79], because ectopic *POU5F1* expression promotes epithelial tumors [207]. This hypothesis aligns well with reports that PRCA progression involves the reactivation of embryonic pathways [84] because *POU5F1* is central to the regulation of stem cell pluripotency in embryogenesis [119].

The *EPAS1*-Region E interaction warrants replication not only in studies of PRCA but also in studies of additional epithelial cancers associated with the sub-region, including colon, kidney, larynx and thyroid. The motivation stems from molecular evidence that, in at least a subset of these cancers, *EPAS1* is over-expressed [202], mRNA transcripts of *POU5F1B* are present [117] and embryonic pathways are implicated [208]. All these features characterize colon cancer, the third most common cancer in American men and women [74]. The hypothesis that *POU5F1B* mediates, at least in part, the association of 8q24 Region E with several epithelial cancers carries great significance, as relatively few genes promote cancer in multiple organs [209]. If

replication study is confirmatory, it will underscore the importance of incorporating interaction analyses into studies of complex human diseases because rs4953347 was overlooked in single-SNP analyses.

The Stage III analysis is essentially an all pairwise interaction analysis among 263 SNPs with evidence of association in single-SNP analyses ($p<0.05$). Although no SNP pairs reach Bonferroni significance, several top-ranking SNP pairs are noteworthy for their biology. The top SNP pair for interaction among the nine susceptibility SNPs involves 8q24 Regions P and H, both gene-deserts. While it is difficult to suggest functional relevance for the observed interaction, the result underscores the complexity of 8q24.

The second SNP for interaction in the Stage III *MSMB* conditional scan is rs4656538, intronic to *POU2F1*. *POU2F1* has been shown to enhance activity of cyclin D1, which controls the expression of many genes involved in the cell cycle and is over-active in several cancers. In a breast cancer study, POU2F1 was shown to form a complex with CREB bound to *cyclin D1* and to enhance CREB-dependent transcription of *cyclin D1* [47]. Given the literature, this observed interaction suggests a functional relevance for the molecular finding that the *MSMB* conditioning SNP alters CREB binding affinity [130,131]. The potential biological importance of this finding is far greater than its modest p-value suggests.

The second overall SNP pair for interaction in Stage III includes rs1859962 of the 17q24 gene desert and rs9920331 of *NR2F2*. rs1859962 has no known function but is reported to be a susceptibility SNP for PRCA [100]. *NR2F2* encodes a transcription factor that regulates expression of diverse genes. It is thought to participate in organogenesis through mesenchymal-epithelial interactions [210]. This function is interesting given the embryonic model for PRCA pathogenesis. Although *NR2F2* has

not been studied in terms of PRCA, it is thought to promote the progression of both lung [211] and breast [212] cancers. Its expression influences the activity of the well known tumor suppressor p53 [213]. The scan SNP rs9920331 also demonstrates interaction with 8q24 Region H. This result supports a functional relationship between *NR2F2* and gene deserts in PRCA etiology, although it is difficult to suggest a mechanism for the observed interaction. Long-range gene regulation, which is often proposed for intergenic regions, is unlikely because these SNPs are on different chromosomes.

Chapter 7

Exploration of Modular Epistasis in Prostate Cancer

This chapter presents several applications of the Tukey score tests that follow from the preceding exploration of pairwise interactions in prostate cancer (PRCA). I investigate modular epistasis in PRCA through two genome wide scans and analyses of a candidate pathway and gene, using data from Stage II of Cancer Genetics Markers of Susceptibility (CGEMS). I focus on the retrospective Tukey score test (RTS) but also use the prospective Tukey score test (PTS) to assess consistency of results. I examine comparable results from the empirical Bayes analysis of pairwise interactions to assess how incorporation of different forms of epistasis affects inference on disease association for specific single nucleotide polymorphisms (SNPs). I also consider standard single-SNP analyses (UML-ME) to assess how the incorporation of either form of epistasis affects inference on disease association.

I used more traditional methods to follow-up on promising RTS results. First, I analyzed saturated interaction models (4.1) with main effects for the SNPs and study covariates, as well as pairwise interactions for the scan and conditioning SNPs through either an unconstrained (UML-SI) or constrained (CML-SI) maximum likelihood. The constrained analysis assumes gene-gene independence for the scan and conditioning SNPs in the underlying population and stratifies on study to minimize epistatic population stratification. Omnibus tests on the main effects and interaction parameters of the scan SNP(s) in these models correspond to Tukey score tests of general disease association. Second, I constructed empirical joint odds ratios (ORs). These

supplemental analyses allow one to assess the robustness of findings and to explore individual interactions between specific combinations of scan and conditioning SNPs.

Section 7.1: Genome Wide Scans

Section 7.1.1: Analysis Plan

The objective of this project is to generate hypotheses about the functionality of susceptibility regions in PRCA, as in the exploration of pairwise interaction. I conducted two genome scans for modular epistasis with distinct conditioning SNP sets (Table 7.1).

The first scan focuses on 8q24 Region P. I designed the conditioning SNP set under an assumption that Region P susceptibility SNPs represent a single casual mechanism specific to the prostate in contrast to 8q24 SNPs in sub-regions associated with additional or different cancers. Region P contains the most statistically significant marker in the extended 8q24 region: rs4242382, a conditioning SNP in the pairwise interaction analysis. rs4242382 resides in a sub-region of Region P that contains androgen receptor (AR) binding sites and enhancer elements responsive to androgens [214]. It is in complete linkage disequilibrium with the Region P susceptibility SNP most extensively studied for molecular function: rs11986220 (not genotyped in CGEMS) [215]. The risk allele of rs11986220 is associated with increased androgen-dependent enhancer activity in Region P and with increased binding affinity of the androgen receptor protein through the co-regulator FoxA1 [214]. These results are particularly noteworthy because the androgen receptor signaling axis is central to the etiology of PRCA [80,216] and because *AR* mutations that affect androgen receptor co-regulators are sufficient to cause aggressive PRCA in mice [217]. The conditioning SNP set for this genome scan includes rs4242382 and six additional loci from 8q24

Region P that demonstrate low pairwise linkage disequilibrium, maximizing coverage of tag SNPs and minimizing multi-collinearity in the null model (Figure 7.1).

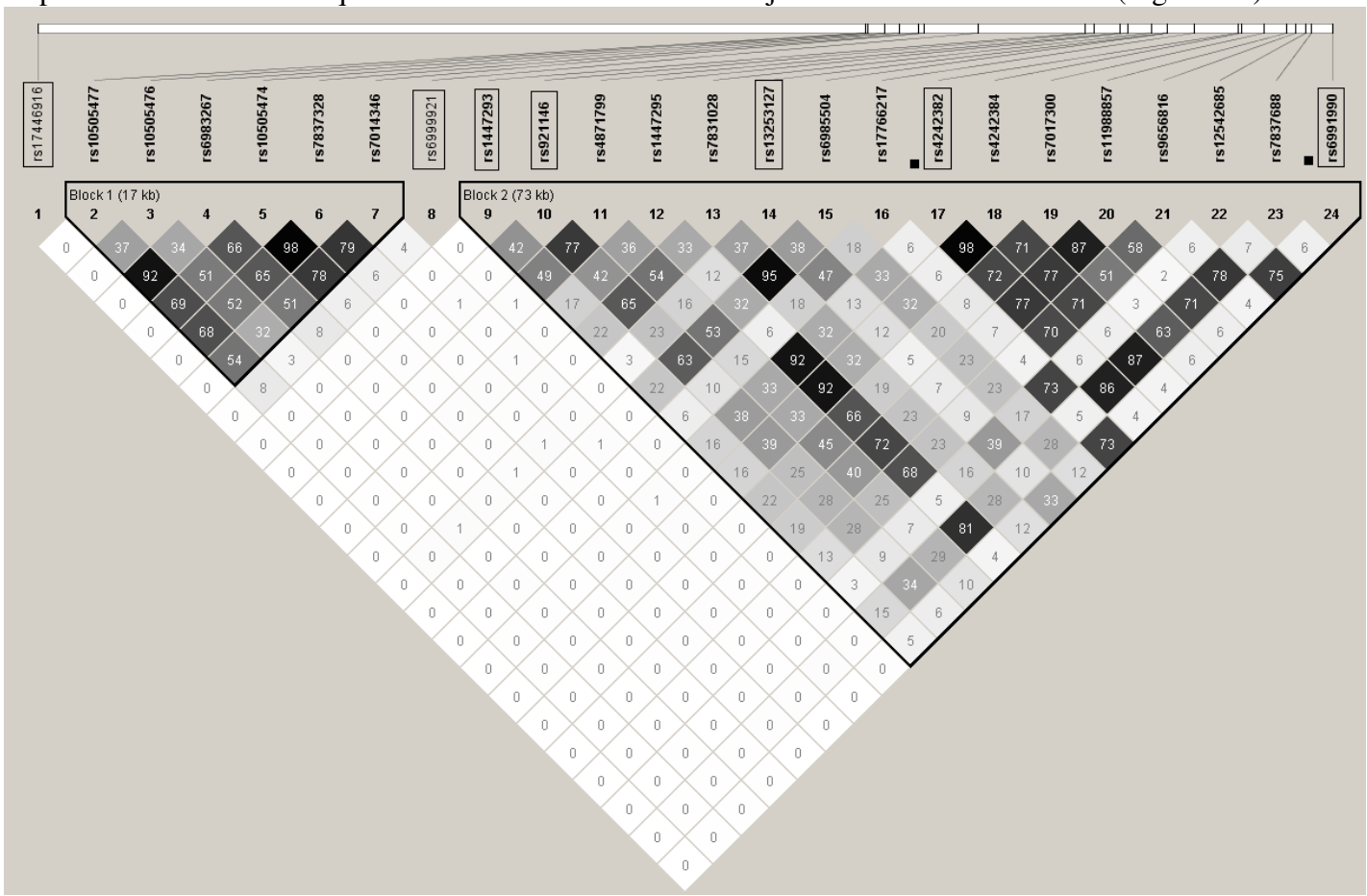
The second genome scan investigates the association of PRCA with type 2 diabetes [129,218–220]. The conditioning SNP set spans *HNF1B* and *JAZF1*, regions that demonstrate associations with both diseases [100,103,129]. The SNP set specifically includes the *HNF1B* and *JAZF1* conditioning SNPs from the pairwise interaction scans, as well as an independent PRCA susceptibility SNP intronic to *HNF1B* [100] ($r^2 = 0.01$ in controls) (Table 7.1). I constructed this conditioning SNP set assuming the markers affect PRCA risk through a common mechanism that involves a diabetic phenotype unobserved in CGEMS. Of note, increased PRCA risk is associated with a decreased genetic risk of type 2 diabetes [220].

Table 7.1: Summary of conditioning SNP sets for two genome scans.

Genome Scan	Conditioning SNP (Gene)	Frequency* (Allele)	OR	P-value
8q24 Region P	rs17446916	0.43 (T)	1.07	0.04
	rs6999921	0.08 (G)	1.15	0.02
	rs1447293	0.37 (G)	1.14	1.54e-4
	rs921146	0.22 (C)	1.21	2.13e-6
	rs13253127	0.46 (T)	1.14	1.22e-4
	rs4242382	0.10 (A)	1.47	3.74e-13
	rs6991990	0.66 (C)	1.15	1.02e-4
Diabetes	rs4430796† (*HNF1B*)	0.52 (A)	1.26	1.85e-11
	rs1164973 (*HNF1B*)	0.81 (G)	1.18	2.70e-4
	rs10486567† (*JAZF1*)	0.75 (G)	1.24	2.21e-7

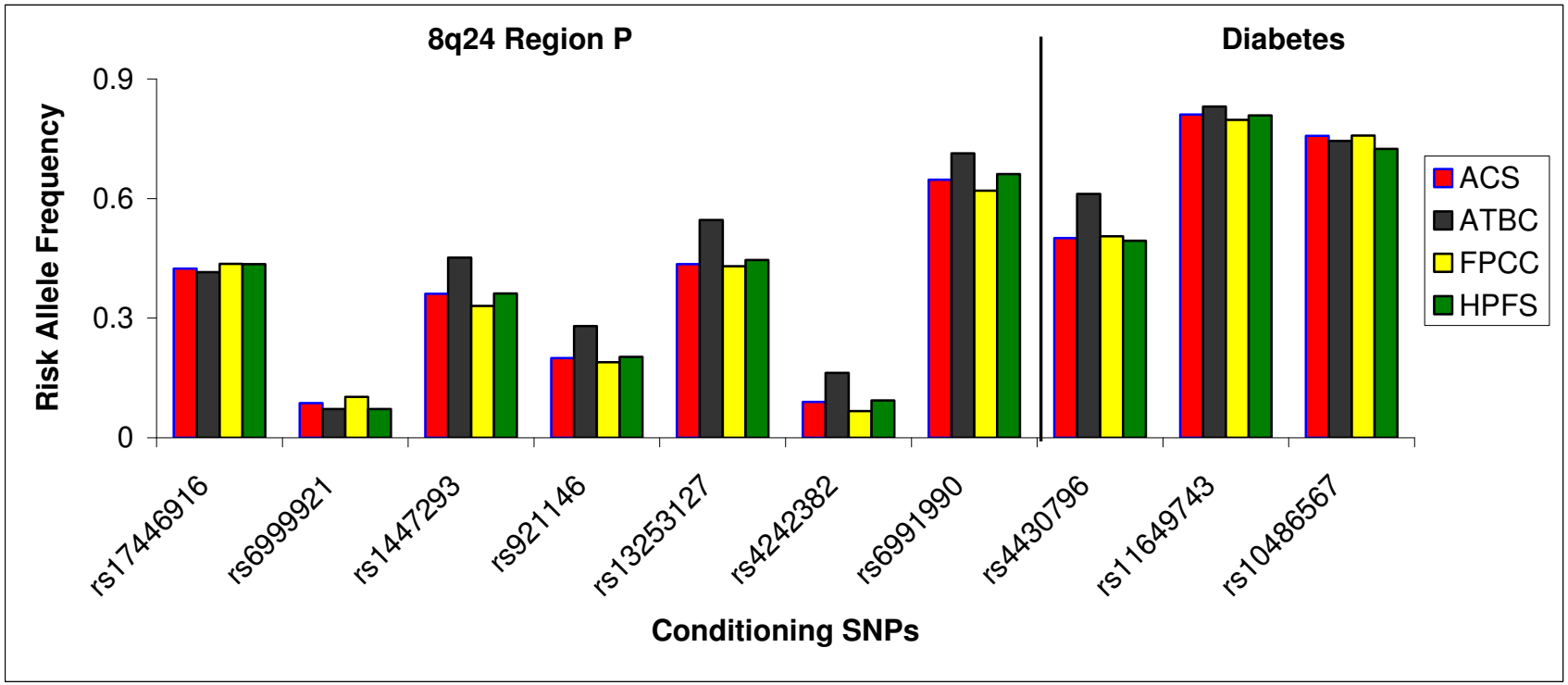
*Risk allele frequency is based on CGEMS Stage II controls.
†SNP is associated with type 2 diabetes.
OR=odds ratio.

Figure 7.1: Linkage disequilibrium (LD) plot for twenty-four 8q24 SNPs genotyped in CGEMS Stage II. The seven boxed loci represent conditioning SNPs in the 8q24 Region P genome scan. The outlined LD block with five of these SNPs contains three binding sites for the androgen receptor. The smaller outlined LD block is 8q24 Region E (Section 1.4). This LD plot spans 354,599 base pairs. The color scheme reflects LD with white corresponding to low r^2 and black to high. Pairwise r^2 based on 3887 study controls are written as percentages in respective blocks. Black squares indicate SNP contributed to joint odd ratios calculations (Figure 7.6).



In each genome scan, I tested individual scan SNPs for disease association, allowing for interaction with the conditioning SNP set. I excluded scan SNPs within 500Kb of any conditioning SNP to minimize violations of gene-gene independence due to physical proximity. This filter selected 27,042 scan SNPs for the 8q24 Region P scan and 27,203 scan SNPs for the "Diabetes" scan, setting the Bonferroni threshold to $p \leq 1.83e\text{-}6$. I included indicators for study in the regression models and stratified the RTS analyses by participation in ATBC because the minor allele frequencies (MAFs) of the conditioning SNPs differ among controls in that study relative to all others (Figure 7.2).

Figure 7.2: Risk allele frequencies for two conditioning SNP sets by study. Values are based on controls in each CGEMS Stage II study (legend). Conditioning SNP sets were used in genome scans to explore the prostate cancer associations of 8q24 Region P and type 2 diabetes.



Section 7.1.2: Results

Figure 7.3 presents quantile-quantile plots of the RTS and PTS p-values for ~27K scan SNPs in the two genome scans for modular epistasis. The top SNPs in all analyses represent established susceptibility regions. The top six SNPs in the 8q24 Region P scan and the top 13 SNPs in the Diabetes scan met genome wide significance in single-SNP analyses of CGEMS Stage II data. The deflation in their p-values from the single-SNP to the Tukey analysis may indicate that the simpler model is more appropriate for these SNPs. Their top ranks are merely shuffled between the RTS and PTS analyses. This consistency contrasts the discrepancies in RTS versus PTS ranks observed for the vast majority of scan SNPs (Figure 7.4).

Figure 7.3: Quantile-quantile plots of χ^2_2-approximate p-values from retrospective (right) and prospective (left) Tukey score tests in 8q24 Region P (top) and Diabetes (bottom) scans with ~27K SNPs. SNPs within 500Kb of the conditioning SNPs are excluded.

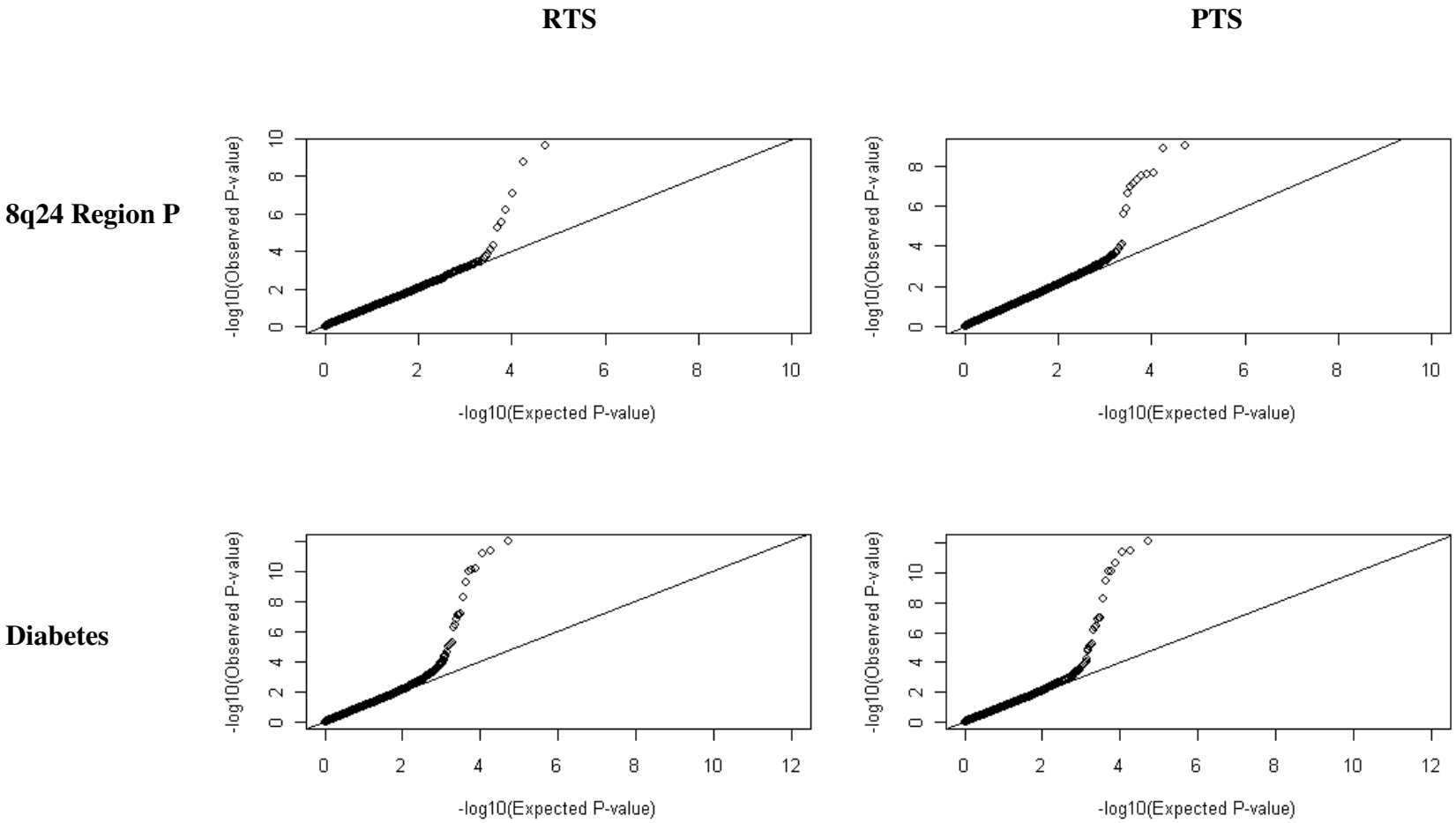
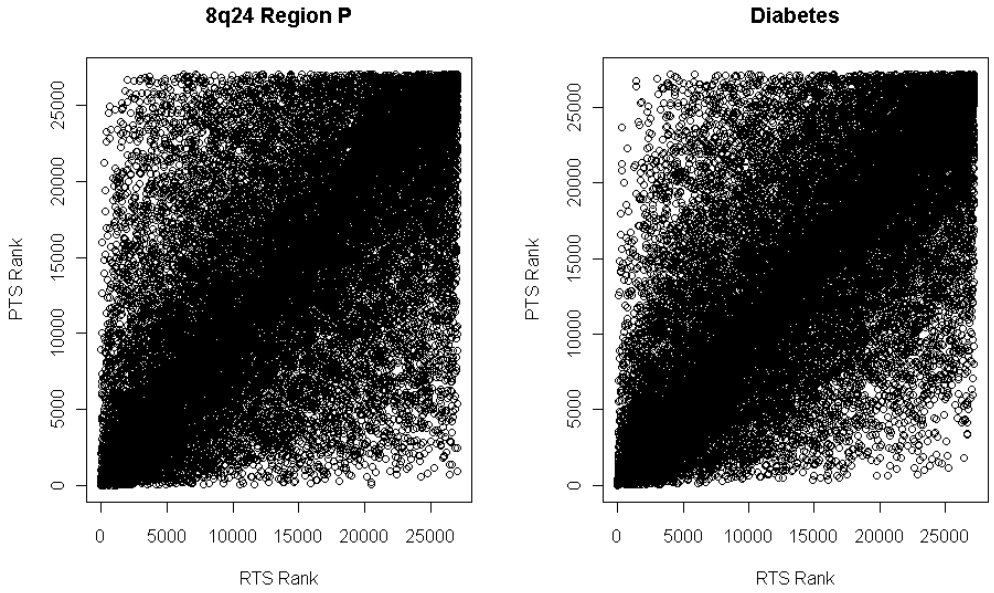


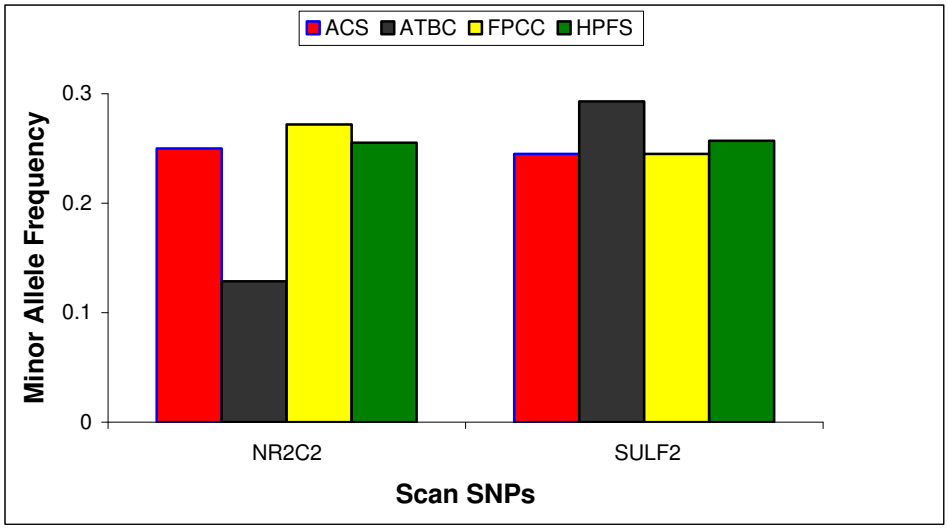
Figure 7.4: **Ranks of ~27K SNPs in two genome scans based on retrospective (RTS) and prospective (PTS) Tukey score tests.** The conditioning SNP sets for the analyses represented either 8q24 Region P (left) or Diabetes (right). SNPs within 500Kb of the conditioning SNPs were excluded.



For inference on modular epistasis in PRCA, I focus on the set of high-ranking SNPs in the RTS analyses that have not been implicated in previous GWAS. Each genome scan highlights a SNP noteworthy for biology. In the 8q24 Region P scan, rs748120 of *NR2C2* (also known as *TR4*, p=4.98e-5, MAF=0.23) ranks 7th overall and 1st among "novel" susceptibility SNPs. In the Diabetes scan, rs4810671 of *SULF2* (p=4.84e-5, MAF=0.48) ranks 23rd overall and 4th among novel susceptibility SNPs. In standard single-SNP analyses, the overall ranks of these scan SNPs are notably lower (rs748120, 191; rs4810671, 252), suggesting interaction drives the RTS signals. Since RTS assesses the general disease association of a scan SNP, it follows that RTS ranks for these scan SNPs are more similar to the empirical Bayes omnibus ranks in comparable conditional scans for pairwise interaction: rs748120 ranks 9th in the 8q24 Region P pairwise scan; rs4810671 ranks 62nd in the *JAZF1* and 109th in *HNF1B* scan. Both scan SNPs demonstrate statistically significant associations in the PTS analysis,

although p-values decrease in magnitude from an order of 1.0e-5 to 1.0e-3. This observed deflation may reflect the increased efficiency of RTS due to the gene-gene independence constraint. The stratification procedure in the RTS analysis appears appropriate for these scan SNPs because their minor allele frequencies differ among controls in ATBC relative to other studies (Figure 7.5).

Figure 7.5: Minor allele frequencies of two scan SNPs by study. Values are based on controls in each CGEMS Stage II study (legend). *NR2C2* is a top result in the 8q24 Region P genome scan and *SULF2* is a top result in the Diabetes genome scan.



More traditional analyses support the RTS results and provide further information on the nature of the observed interactions. First, the CML-SI omnibus tests on all scan SNP parameters are statistically significant for both rs748120 (p=2.67e-5) and rs4810671 (p=4.32e-4); the UML-SI omnibus tests are also significant. In these models, each scan SNP demonstrates a significant interaction ($p<0.05$) with two conditioning SNPs (Tables 7.2, 7.3). Empirical joint odds ratios involve total risk allele counts of these conditioning SNPs and minor allele counts of the scan SNP (Figure

7.6). The cell counts for odds ratios are generally large, with few below 100. In the 8q24 Region P scan, the exceptions are the subset of no risk alleles at conditioning SNP loci and the subset of two minor alleles at the scan SNP locus. In the Diabetes scan, the exceptions are the subset of four risk alleles at conditioning SNP loci and the subset of two minor alleles at the scan SNP locus.

Table 7.2: Results of CML analysis of saturated interaction model with scan SNP rs748120 and seven 8q24 Region P conditioning SNPs.

	Main Effects			Interaction with rs748120		
Covariate	**OR**	**95% CI**	**P-value**	**OR**	**95% CI**	**P-value**
rs748120 (*NR2C2*)	0.99	(0.84, 1.16)	0.90			
rs17446916	1.08	(1.01, 1.16)	0.03	0.98	(0.90, 1.06)	0.58
rs6999921	1.17	(1.04, 1.33)	0.01	1.02	(0.89, 1.16)	0.82
rs1447293	1.01	(0.91, 1.12)	0.81	1.11	(0.99, 1.24)	0.07
rs921146	0.90	(0.77, 1.05)	0.18	1.06	(0.90, 1.24)	0.51
rs13253127	0.98	(0.88, 1.09)	0.67	1.05	(0.94, 1.18)	0.39
rs4242382	1.70	(1.48, 1.96)	2.56e-13	0.71	(0.61, 0.84)	2.29e-5
rs6991990	1.13	(1.03, 1.25)	0.01	0.89	(0.80, 0.99)	0.03

Omnibus test on all rs748120 parameters gives p=2.67e-5.
CML=constrained maximum likelihood logistic analysis; OR=odds ratio; CI=confidence interval.

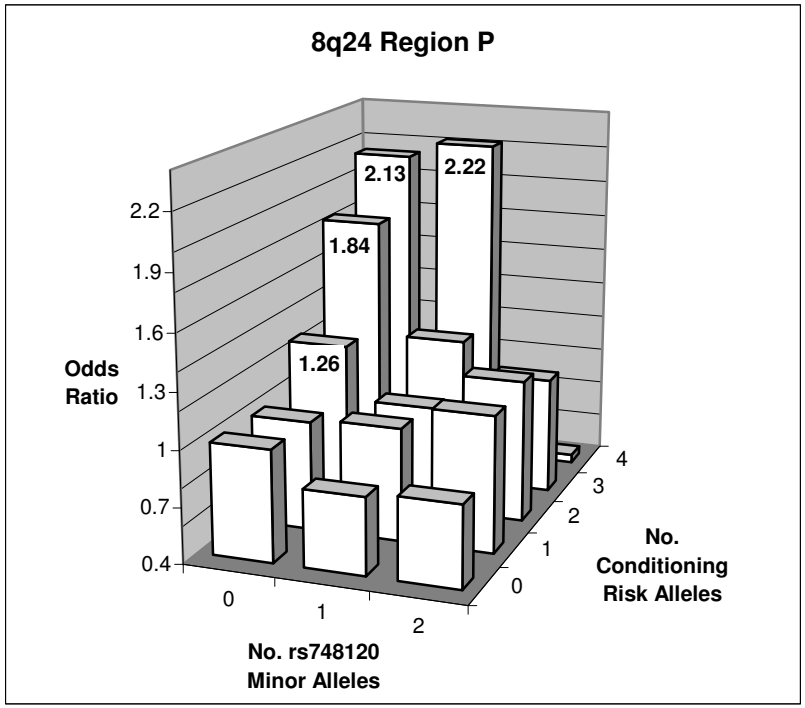
Table 7.3: Results of CML analysis of saturated interaction model with scan SNP rs4810671 and conditioning SNPs in *HNF1B* and *JAZF1*.

	Main Effects			Interaction with rs748120		
Covariate	**OR**	**95% CI**	**P-value**	**OR**	**95% CI**	**P-value**
rs4810671 (*SULF2*)	0.75	(0.56, 1.00)	0.053			
rs4430796 (*HNF1B*)	1.20	(1.11, 1.29)	2.45e-6	1.13	(1.03, 1.23)	0.010
rs11649743 (*HNF1B*)	1.14	(1.04, 1.26)	0.007	1.00	(0.88, 1.13)	0.974
rs10486567 (*JAZF1*)	1.19	(1.09, 1.30)	8.30e-5	1.14	(1.02, 1.28)	0.019

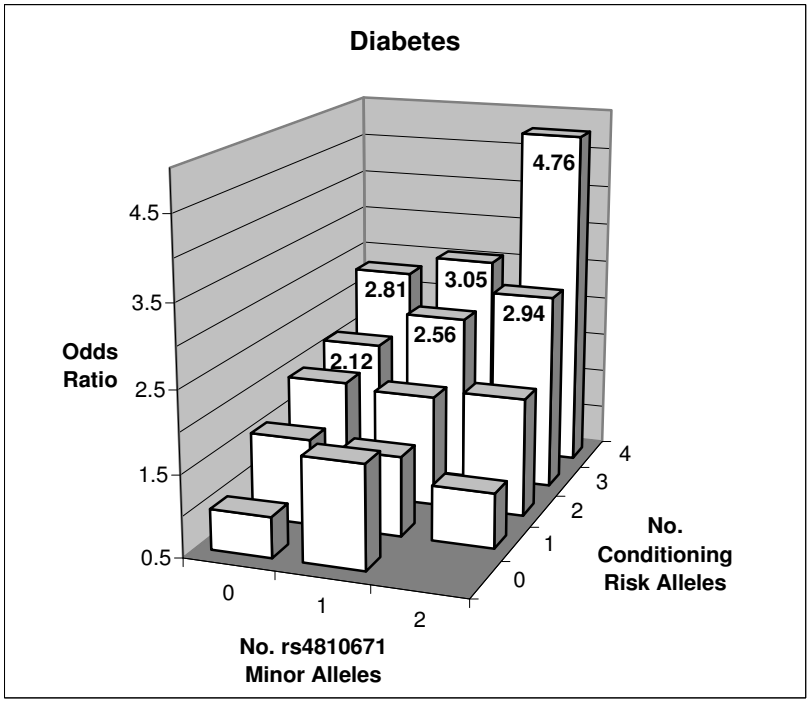
Omnibus test on all rs4430796 parameters gives p= 4.32e-4.
CML=constrained maximum likelihood logistic analysis; OR=odds ratio; CI=confidence interval.

Figure 7.6: Empirical joint odds ratios. Results are based on the minor allele count of a top scan SNP and the total risk allele count of conditioning SNPs that demonstrate significant interaction with it in a saturated interaction model. For the 8q24 Region P scan (left), the scan SNP is rs748120 with MAF=0.23 and the conditioning SNPs are rs4242382 and rs6991990. For the Diabetes scan (right), the scan SNP is rs4810671 with MAF=0.48 and the conditioning SNPs are rs4430796 (*HNF1B*) and rs10486567 (*JAZF1*). Values are given for odds ratios significant at the 5% level.

A)



B)



Section 7.1.3: Discussion

These applications highlight important features of the Tukey score tests, as well as identify SNPs that suggest functionality for established PRCA susceptibility regions. The RTS analyses specifically highlight rs748120 of *NR2C2* in the 8q24 Region P scan and rs4810671 of *SULF2* in the Diabetes scan. Although the scan SNPs do not meet genome wide significance, both rank in the top 25 of more than 27K SNPs. Dominating the top ranking SNPs in both the RTS and PTS analyses are markers from established susceptibility regions that meet genome wide significance in single-SNP CGEMS analyses. These results support the use of Tukey score tests in exploratory GWAS analyses, highlighting the flexible nature of a test for general disease association. Although top results are consistent, the RTS and PTS analyses produce discordant ranks for the vast majority of scan SNPs (Figure 7.4). Consistency of Tukey score test results may suggest robustness in a finding.

The 8q24 Region P scan highlights rs748120 of *NR2C2*. An epistatic effect appears to drive the RTS signal because rs741820 ranks much lower in single-SNP analysis. The scan SNP demonstrates evidence of epistasis with multiple conditioning SNPs in a saturated interaction model (Table 7.2). The two conditioning SNPs with statistically significant pairwise interactions ($p \leq 0.05$) in the model reside in the same sub-region of Region P that contains AR binding sites and enhancer elements responsive to androgens. They are rs4242382, the most significant SNP in marginal analyses, and rs6991990. These SNPs are independent ($r^2 = 0.04$ in controls), suggesting there is a region-wide interaction between 8q24 Region P and rs748120. The empirical joint odds ratio that incorporate the conditioning SNPs rs4242382 and rs6991990 suggest the latent causal mechanism of 8q24 Region P acts almost exclusively in the absence of minor rs748120 alleles (Figure 7.6a).

NR2C2 is a co-regulator of the androgen receptor protein (AR); it encodes a protein that can form a complex with AR and decrease expression of both their target genes [221]. A target gene of *NR2C2* is *NANOG*, one of two additional genes at the center of stem cell pluripotency with *POU5F1* [222]. This relationship is particularly noteworthy because PRCA pathogenesis is thought to involve the reactivation of embryonic pathways [84]. It suggests a mechanism by which embryonic pathways can be activated in PRCA in addition to the hypothesized regulation of *POU5F1B* by *EPAS1* (Section 6.3). NR2C2 has also been shown to reduce synthesis of Vitamin D [223], which influences cellular differentiation and proliferation in the prostate [224,225]. Observational studies identified an association of Vitamin D deficiency with increased PRCA risk [226]. Given the literature, I hypothesize that impaired AR-NR2C2 binding contributes to the increased androgen responsiveness of 8q24 Region P in the presence of a risk allele at rs11986220 (or rs4242382 by complete linkage disequilibrium). Impaired binding would also correspond to increased NR2C2 activity that could activate *NANOG* expression and repress Vitamin D synthesis at levels outside the physiologic range. Both sequelae would increase risk of PRCA, consistent with previous reports.

The Diabetes scan highlights rs4810671 of *SULF2*. The RTS signal appears to involve a relatively large epistatic effect, given the comparatively low rank of rs4810671 in single-SNP analysis. The scan SNP demonstrates evidence of epistasis with two of the three conditioning SNPs in a saturated interaction model: rs4430796 of *HNF1B* and rs10486567 of *JAZF1*, the conditioning SNPs from the pairwise interaction scans. These results suggest the observed association of rs4810671 with PRCA involves the proposed latent diabetic phenotype rather than a gene specific function of a single conditioning region. The empirical joint odds ratios that incorporate these

conditioning SNPs suggest rs4810671 affects PRCA risk only when at least three conditioning SNP risk alleles are present (Figure 7.6b).

SULF2 has been well characterized in molecular studies. It demonstrates connections to cancer, specifically PRCA, as well as type 2 diabetes. In several cancers, including pancreas, breast, lung and liver, *SULF2* demonstrates oncogenic properties [227–230]. The gene encodes an enzyme that modifies the sulfation and function of heparin sulfate proteoglycans [231]. Abnormalities in these sugars and particularly in HSPG2 (Perlecan) contribute to the disease progression of both type 2 diabetes [232] and PRCA [233,234]. Heparin sulfate proteoglycans influence the FGF and WNT-signaling axes [228] that are disrupted in PRCA [81,235,236]. A positive regulator of *SULF2* is insulin, the hormone at the center of diabetes [237]. Cellular response to insulin is thought to mediate the association of type 2 diabetes and PRCA [238–241]. Given the literature and observed modular interaction, I propose that abnormal insulin function affects PRCA risk through dysregulation of *SULF2*. This hypothesis aligns well with the claim that a consequence of type 2 diabetes mediates its association with PRCA [220]. The RTS finding is worthy of replication, particularly because the SULF2 enzyme would be good candidate for a drug target and biomarker [228] in PRCA.

If replication study confirms rs748120 and rs4810671 are susceptibility SNPs for PRCA, it will underscore the benefit of incorporating epistasis and particularly modular epistasis into GWAS analyses. For comparison, both scan SNPs were excluded from CGEMS Stage III due to weak marginal effects ($p>0.001$). A recent report refutes a link between the associations of *HNF1B* and *JAZF1* with PRCA and type 2 diabetes based on null results in main effects analyses [242]. My own pairwise

interaction analysis (Chapter 6) overlooked rs4810671 in both the *HNF1B* and *JAZF1* conditional scans, but it did highlight rs4242382 in the 8q24 Region P conditional scan.

Section 7.2: Candidate Regions

Section 7.2.1: Candidate Gene *MYEOV* Application

MYEOV has been implicated in PRCA because it flanks the 11q13 susceptibility region and functions as an oncogene in other cancers [103]. I investigated the hypothesis that these regions affect PRCA risk through epistasis using PTS because physical proximity invalidates the RTS assumption of independence for the regions. The regression model includes a single scan SNP, rs11605162 that ranks 1st out of 23 SNPs in the *MYEOV* gene region for marginal effect in Stage I (p=0.058). This filter ensures data for selection and analysis are independent. The regression model also includes covariates for study. I constructed the conditioning SNP set through the following fine-mapping analysis.

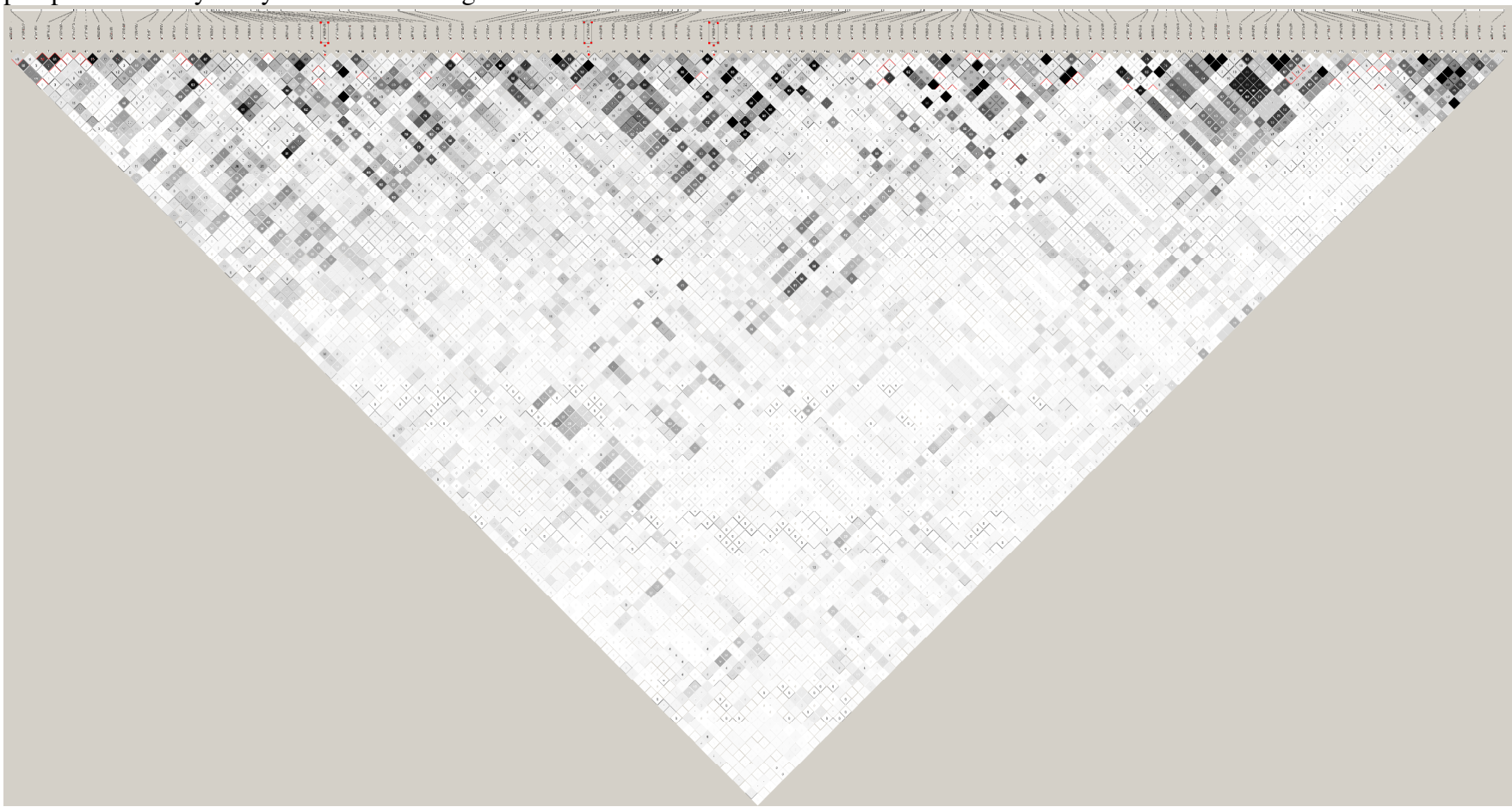
Section 7.2.1a: Fine-Mapping of Chromosomal Region 11q13

The fine-mapping data of CGEMS Stage III include 120 SNPs from the 11q13 susceptibility region (Figure 7.7). The conditioning SNP from the 11q13 scan for pairwise interaction, rs10896449, demonstrates the most significant marginal effect among them [5]. An objective of this supplemental analysis is to determine if 11q13 harbors independent risk markers in addition to the top marker, as seen with 8q24 [3,105] and *HNF1B* [100]. The logistic models for this analysis include rs10896449, study covariates and, as appropriate, one or more 11q13 SNPs. I analyzed the models using the standard UML method. I considered two-SNP main effects models first, testing the additional 11q13 SNP for PRCA association. I assessed the statistical

significance of the top SNP in the analysis through a parametric permutation method, since the high correlation of SNPs made a Bonferroni correction too conservative. The permutation procedure preserves the linkage disequilibrium of the SNPs in each model by permuting the outcome data rather than the genetic data. It also preserves the PRCA association of rs10896449 by assigning case-control status based on random draws from a Bernoulli distribution with probability of "case" set by the null model, as in (2.5). I assigned an adjusted p-value to the top SNP, using minimum p-values from each of 100K permutations to define $p_{\text{adj}} = P\left(\min\left(\text{permuted } p\right) \leq \text{ observed } p\right)$.

If $p_{\text{adj}} \leq 0.05$, I defined the top SNP as an independent susceptibility SNP in 11q13, incorporated it into the null model and repeated the process for a 3-SNP alternative model. This sequential testing of multilocus models continued until the top SNP had $p_{\text{adj}} > 0.05$. To reduce computing time, I collaborated with a colleague, Mr. William Wheeler, on the permutations.

Figure 7.7: Linkage disequilibrium (LD) plot of chromosomal region 11q13. Plot spans 241,804 base pairs and denotes SNPs included in CGEMS Stage III fine-mapping. The color scheme reflects LD with white corresponding to low r^2 and black to high. Pairwise r^2 is written as percentages in respective blocks. Estimates are based on HapMap data for subjects of European ancestry [243]. Three SNPs identified as independent susceptibility loci for prostate cancer are boxed in red. They define the conditioning SNP sets in the prospective Tukey analysis for candidate gene *MYEOV*.



The results suggest 11q13 contains three independent PRCA susceptibility SNPs: rs10896449, rs12793759 and rs10896438 (Table 7.4, Figure 7.7). The top SNP for the analysis of two-SNP models is rs12793759, a novel susceptibility marker for PRCA (p_{adj}=0.004). The top SNP for the analysis of three-SNP models is rs10896438. I included it in the final model, despite a borderline adjusted p-value (p_{adj}=0.054), because it is a proxy for the PRCA susceptibility SNP rs12418451 [106] (r^2=0.96) [115]. These SNPs demonstrate low pairwise r^2 in CGEMS controls: rs10896449 and rs12793759, 0.17; rs10896449 and rs10896438, 0.10; rs12793759 and rs10896438, 0.12.

Table 7.4: Summary of 3-SNP model from sequential logistic regression analysis of SNPs in 11q13 susceptibility region.

SNP	OR	95% CI	P value
rs10896449	1.14	(1.09, 1.20)	8.69e-9
rs12793759	1.11	(1.04, 1.18)	1.41e-3
rs10896438	1.07	(1.02, 1.12)	5.92e-3

OR=odds ratio; CI=confidence interval.

A relevant methodological issue is that the final model of a forward-selection analysis can be highly dependent on the data. Alternative approaches that demonstrate increased predictive accuracy include non-parametric bootstrapping and sub-sampling. In the first, forward selection is performed multiple times on subsets that have been sampled with replacement from the full dataset. In the latter, forward selection is performed on subsets that have been sampled without replacement. Bootstrapping can involve datasets as large as the original, but sub-sampling is restricted to datasets smaller than the original. Both approaches construct a series of multilocus models that

can be examined to determine the probability that a given SNP is included in the final model.[244]

Section 7.2.1b: Results

The PTS results do not support an association of *MYEOV* with PRCA risk (p=0.68). Similarly, no association was detected for this *MYEOV* SNP in the conditional scan for pairwise interaction with 11q13 (empirical Bayes omnibus p=0.75) or in the single-SNP analysis (p=0.13).

Section 7.2.2: Ectopic *POU5F1* Expression

This project is motivated by the top result for pairwise interaction in the conditional scan for 8q24 Region E (Table 6.2, Figure 6.2). I specifically investigate the hypothesis that *POU5F1B* mediates the PRCA association of 8q24 Region E through functional similarities with *POU5F1*. Despite substantial support in the literature (Section 6.4), this hypothesis has not been tested in published work.

The Tukey model for this analysis differs from those previous. Its conditioning SNP set has a single marker and its scan SNP set has multiple (four) markers. The conditioning SNP is the top marker in 8q24 Region E, rs6983267, which is ~15Kb from *POU5F1B* and has been implicated as a causative locus [245]. I constructed the scan SNP set to represent ectopic *POU5F1* expression, which is tumorigenic in epithelium [205] where the vast majority of PRCA tumors originate [79]. The pool of potential scan SNPs includes all CGEMS SNPs in the regions of *POU5F1* target genes. A recent bioinformatics study provides a thorough list of these genes, which are too numerous for a single regression model [246]. I limited the pool to target genes involved in pluripotency regulation because progenitor cells are central to the

development of PRCA [84] and of epithelial tumors due to ectopic *POU5F1B* expression [205]. Also, the selection of scan SNPs within a functional group of a pathway provides stricter adherence to the latent variable framework of the Tukey model. CGEMS includes four of these genes, *ESRRB*, *SALL4*, *KLF4* and *DPPA3*, for which I provide brief information (Table 7.5, Figure 7.8). Both *ESRRB* and *SALL4* are positive regulators, as well as target genes, of *POU5F1* [247]. One function of *ESRRB* is to up-regulate *NANOG*, reminiscent of *NR2C2* [248]. *DPPA3* expression has been detected in germ cell tumors [249]. *KLF4* is a transcription factor that regulates the cell cycle and cell differentiation in many tissue types, including epithelium. It can act as a tumor suppressor or oncogene depending on the tissue.[250,251] The scan SNP set includes the top SNP in each gene based on a Stage I single-SNP analysis, ensuring the data for selection and analysis are independent. As the scan and conditioning SNPs are on different chromosomes, I assume the gene-gene independence constraint is valid.

I included study indicators in the Tukey model for this analysis but did not stratify by study because the allele frequencies of the scan and conditioning SNPs do not appear to covary within study among CGEMS controls (Figure 7.9). No correction for multiple testing is necessary with a single hypothesis.

Table 7.5: Summary of scan SNP set for candidate pathway application. Results from Stage II are given for the 8q24 Region E conditional scan.

Scan SNP	Nearby Gene	SNP-Gene Proximity	Chr	MAF	Stage I Marginal P-value	Stage II Omnibus P-value	Stage II Interaction P-value
rs10772662	*DPPA3*	Nearby	12	0.32	0.02	0.07	0.82
rs7155416	*ESRRB*	Intron	14	0.12	0.17	0.08	0.08
rs1832741	*KLF4*	Nearby	9	0.47	0.11	0.10	0.18
rs6021460	*SALL4*	Intron	20	0.44	0.77	0.21	0.09

Chr=chromosome; MAF=minor allele frequency.

Figure 7.8: Simplified schematics of pluripotency network centered on *POU5F1*. The four genes listed in the blue box as major pluripotency-related genes were jointly tested for disease association through the prospective Tukey score test with the conditioning SNP rs6983267 of 8q24 Region E, yielding a significant result (p=0.01). An intronic SNP of *EPAS1* is the top pairwise interacting SNP with rs6983267 in a genome wide scan (p=9.69e-5). rs6983267 is associated with altered binding of β-catenin to 8q24. Sharp (blunt) arrows represent excitatory (inhibitory) relationships between genes; WNT, FZD and SMAD respectively represent the gene families of wingless type protein, frizzled, and similar to mothers against decapentaplegi; Ids = inhibitors of differentiation. Adapted from previous publications [118,246].

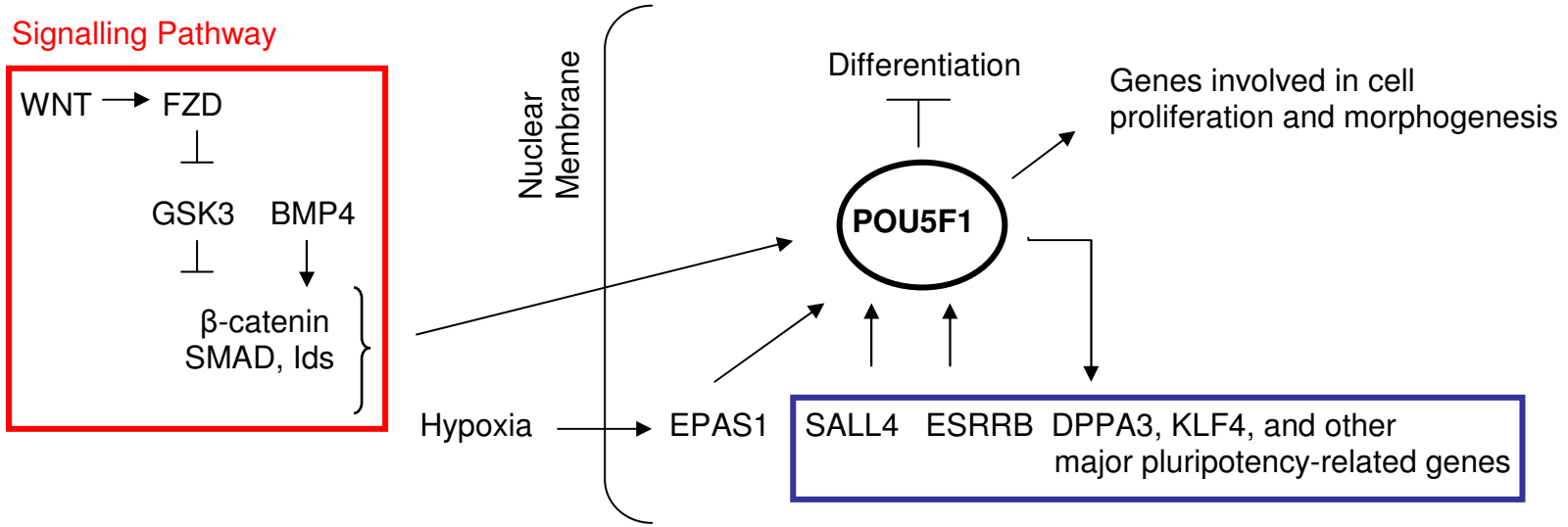
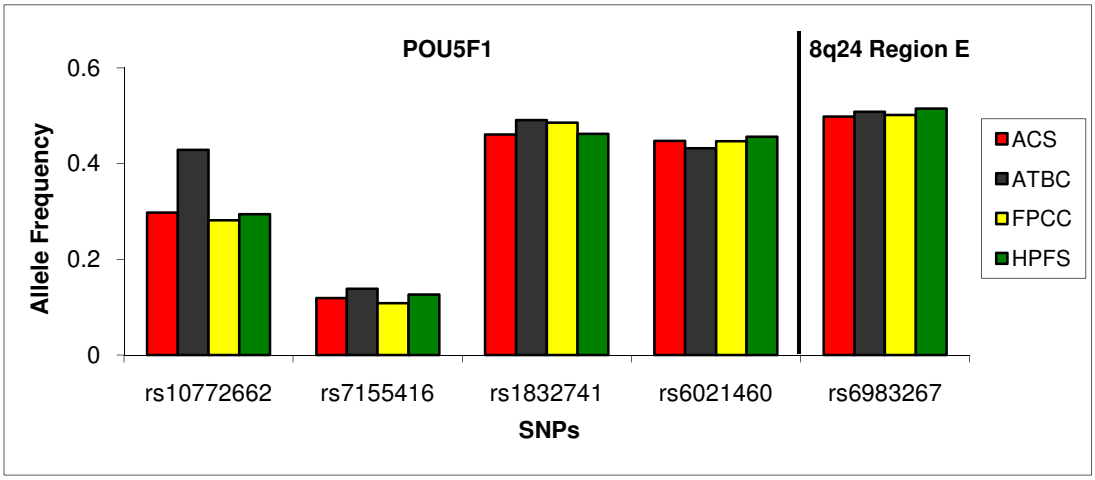


Figure 7.9: Allele frequency for scan and conditioning SNP sets of candidate pathway application by study. Minor allele frequencies are given for the scan SNP set (POU5F1). Risk allele frequencies are given for conditioning SNP (8q24 Region E). Values are based on controls in each CGEMS Stage II study (legend).



Section 7.2.2a: Results

The scan SNP set representing ectopic *POU5F1* expression demonstrates weak evidence of disease association in RTS (p=0.14). The set demonstrates a significant association in PTS (p=0.01), and more traditional analyses support an association. The omnibus test on all scan SNP parameters in a saturated interaction model is significant in the CML-SI analysis (p=0.03), more so in the UML-SI analysis (p=4.32e-3, Table 7.6). In these analyses, rs10772662 of *DPPA3* is the only scan SNP that does not show evidence of PRCA association for main effect or interaction (p<0.05). Figure 7.10 gives the empirical joint odds ratios for the remaining scan SNPs and the 8q24 Region E conditioning SNP based on a total risk allele count. The "risk" alleles for the scan SNPs correspond to higher PRCA risk in Stage II single-SNP analysis, although no marginal effect is significant. They are the major allele for rs1832741 and rs6021460 and the minor allele for rs7155416. The maximum total count of risk alleles for the scan SNPs is five because the subset of six risk alleles has several empty cells. The

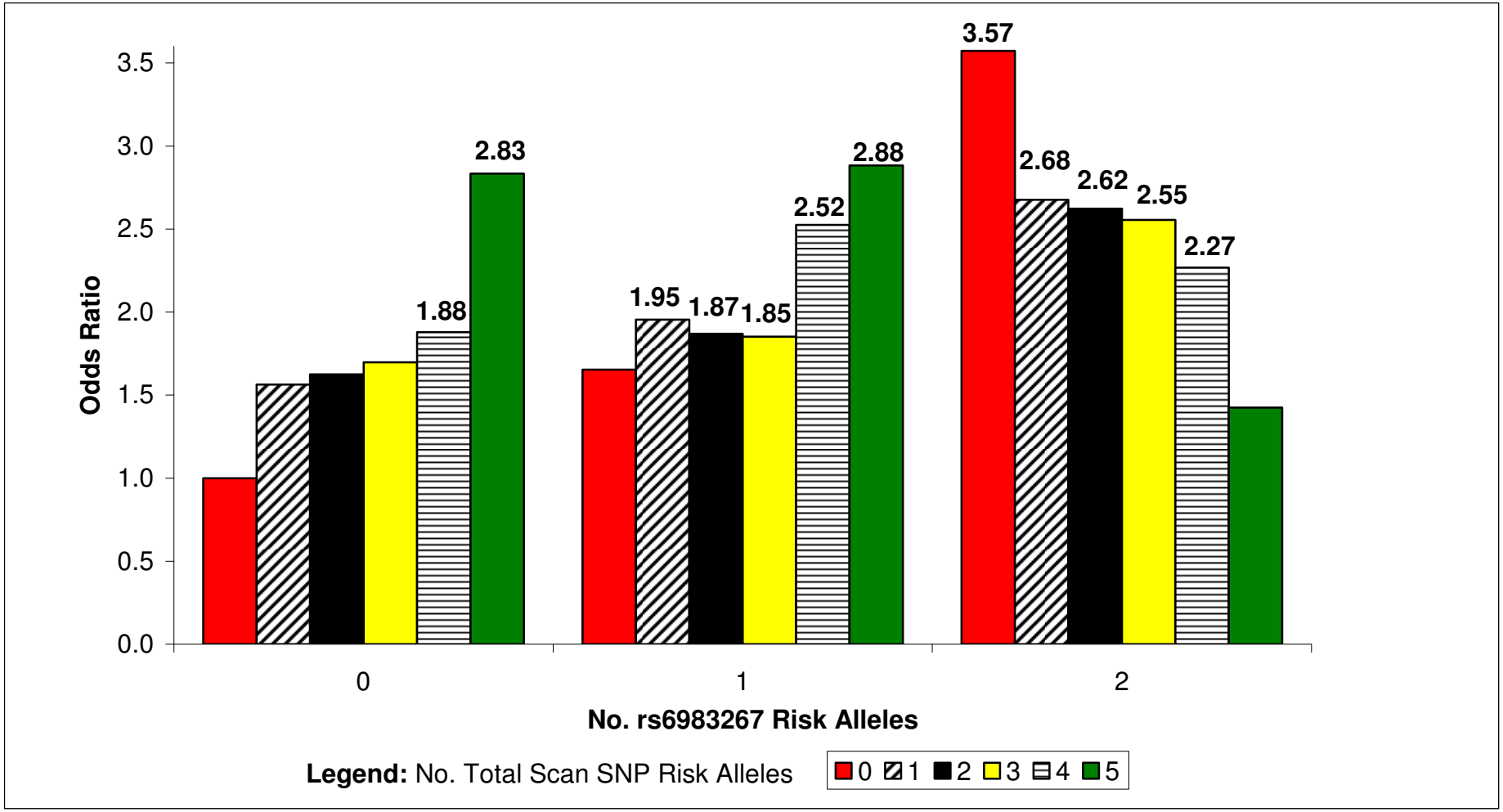
only additional cells with counts less than 100 are within the subsets of zero or five scan SNP risk alleles.

Table 7.6: Results of UML analysis of saturated interaction model with one 8q24 Region E SNP and four scan SNPs from target genes of *POU5F1*.

	Main Effects			Interaction with rs6983267		
Covariate	**OR**	**95% CI**	**P-value**	**OR**	**95% CI**	**P-value**
rs6983267	1.08	(0.94, 1.26)	0.26			
rs10772662 (*DPPA3*)	1.07	(0.95, 1.20)	0.29	1.01	(0.92, 1.11)	0.77
rs7155416 (*ESRRB*)	1.24	(1.05, 1.47)	0.01	0.87	(0.76, 0.99)	0.04
rs1832741 (*KLF4*)	0.88	(0.79, 0.99)	0.03	1.07	(0.98, 1.17)	0.14
rs6021460 (*SALL4*)	0.89	(0.79, 0.996)	0.04	1.11	(1.02, 1.21)	0.02

Omnibus test on all scan SNP parameters gives p=4.23e-3.
UML=unconstrained maximum likelihood logistic analysis; OR=odds ratio; CI=confidence interval.

Figure 7.10: Empirical joint odds ratios. Risk allele counts for scan SNPs involve rs7155416 of *ESRRB*, rs1832741 of *KLF4* and rs6021460 of *SALL4*. Bold values indicate odds ratios significant at the 5% level.



Section 7.2.2b: Discussion

The results of this candidate pathway application provide sufficient evidence to warrant follow-up on the hypothesis that *POU5F1B* mediates the PRCA association of 8q24 Region E through its similarities with *POU5F1*. Specifically, the results suggest that *POU5F1* target genes involved in pluripotency regulation affect PRCA risk through interaction with 8q24 Region E. If replication confirms a PRCA association for these scan SNPs, it would underscore the benefit of modeling modular epistasis because all scan SNPs were overlooked in single-SNP analyses and in the 8q24 Region E conditional scan for pairwise interaction.

The SNP sets demonstrate a cross-over interaction in their empirical joint odds ratios (Figure 7.10). When both conditioning SNP risk alleles are present, the greatest risk is observed in the subset of no scan SNP risk alleles but, when less than two conditioning SNP risk alleles are present, the greatest risk is observed in the subset of maximum scan SNP risk alleles. This pattern is a more extreme version of the one involving 8q24 Region E and the *EPAS1* SNP rs4953347 (Figure 6.2). The consistency supports my hypothesis that the PRCA association of 8q24 Region E involves a type of pluripotency network centered on *POU5F1B* rather than *POU5F1*. More specifically, I hypothesize *EPAS1* promotes over-expression of *POU5F1B* that mimics tumorigenic, ectopic expression of *POU5F1*. I advocate for follow-up study not only in PRCA but in all epithelial cancers associated with 8q24 Region E.

Chapter 8

Conclusion

Multilocus methods can increase power to a detect disease association in single nucleotide polymorphisms (SNPs), capturing the biology of complex human diseases more accurately than single-SNP models [42,52–58]. My dissertation research contributes to the field of statistical genetics by evaluating and developing methods to detect disease association in the presence of epistasis using case-control data. My applied work with these methods contributes to the field of cancer genetics by identifying statistical interactions that suggest functional relevance for susceptibility regions in prostate cancer (PRCA).

Chapter 2 presents the first empirical evaluation of standard logistic analysis and two methods that can gain power to detect pairwise interactions in the setting of a genome wide association study (GWAS). The alternate methods impose a constraint of gene-gene independence in the underlying population between candidate interacting SNPs. The results reveal a weakness in case-only type methods and highlight the promise of a recently proposed empirical Bayes (EB) method. The rigid constraint of case-only type methods inflates type I error when candidate interacting SNPs demonstrate linkage disequilibrium or epistatic population stratification (Figure 2.1b). Researchers can minimize the first type of bias that is likely more common by selecting SNP pairs for analysis based on genomic location. Recent modifications to the case-only method minimize the second type of bias, retaining efficiency over standard analyses [152,153,171]. For example, Bhattacharjee et al. proposed a method that enforces the independence constraint in small, homogenous subsets of subjects

clustered through principal component analysis [2]. The EB method minimizes bias in a more general way, using the data to assess uncertainty about independence and to weight interaction estimates from constrained and unconstrained analyses (2.11). This data-adaptive approach makes EB more robust than case-only type methods (Figure 2.1) and more powerful than standard methods.[59]

Chapters 3 and 4 introduce the innovative retrospective Tukey score test (RTS). The underlying Tukey model (3.1) allows for interaction between sets of biologically similar SNPs [175], representing a movement in statistical genetics towards modular biology. The regression model also offers statistical advantages, motivating a flexible test of disease association that incorporates main- and epistatic effects and reducing degrees of freedom on the test statistic. RTS further improves efficiency by imposing a constraint of gene-gene independence in the underlying population between the SNP sets. When independence holds, RTS demonstrates high power to detect disease association in the presence of epistasis (Figures 4.3-4.5) and controls type I error (Table 4.4). It is a powerful alternative to existing methods for analysis of case-control data in genetic epidemiology. Furthermore, RTS provides a foundation for work on robust and efficient methods that exploit gene-gene independence in studies of modular epistasis, as the case-only method does for pairwise interactions. I explore an extension of RTS in Chapter 5: the composite Tukey score test (CTS). CTS can be considered an EB equivalent for studies of modular epistasis because it uses the data to relax the RTS independence assumption. Simulation studies motivate further work on CTS, suggesting it may become a powerful and robust framework for analyses of modular interactions, both gene-gene and gene-environment.

Chapters 6 and 7 present numerous applications of EB and RTS to a PRCA GWAS. The results underscore the potential of multilocus methods to generate

hypotheses about disease etiology, uniting several susceptibility regions through overlapping pathways known to be disrupted in PRCA. The pairwise interaction between *EPAS1* and 8q24 Region E (Figure 6.2) implicates *POU5F1B* and suggests a mechanism for its over-expression in PRCA. The follow-up RTS analysis suggests *POU5F1B* may be at the center of a reactivated pluripotency network that, in healthy states, is centered on *POU5F1*, *NANOG* and *SOX2* (Figure 7.8). The interaction results for *NR2C2* and 8q24 Region P (Figure 7.6a) suggest a second mechanism for reactivation of embryonic pathways in PRCA: increased *NANOG* expression. *NR2C2* also influences the androgen receptor signaling axis [80] that is central to PRCA progression [221]. *NR2C2* can affect WNT signaling, which is disrupted in PRCA, through the Vitamin D pathway that is associated with PRCA risk [81,223,226,252,253]. *JAZF1* regulates *NR2C2* [128]. *JAZF1* and *HNF1B* are associated with PRCA and type 2 diabetes [129]. Increasing evidence suggests insulin mediates the association of PRCA and type 2 diabetes [238–240]. An observed modular interaction suggests the mechanism may be insulin's regulation of *SULF2*, an oncogene that also affects WNT signaling [228] (Figure 7.6b). These results have substantial support in the literature, but all require replication in large studies of independent subjects. My supervisors and I have established collaborations for this purpose. Confirmatory results would underscore the advantages of incorporating epistasis into exploratory GWAS analyses because these candidate SNPs were overlooked in single-SNP analyses.

Reference List

1. Ciampa J, Yeager M, Amundadottir L, Jacobs K, Kraft P, Chung C, Wacholder S, Yu K, Wheeler W, Thun MJ, Divers WR, Gapstur S, Albanes D, Virtamo J, Weinstein S, Giovannucci E, Willet WC, Cancel-Tassin G, Cussenot O, Valeri A, Hunter D, Hoover R, Thomas G, Chanock S, Chatterjee N (2011) Large Scale Exploration of Gene-Gene Interactions in Prostate Cancer Using a Multi-stage Genome-wide Association Study. Cancer Res 71: 3287-3295.

2. Bhattacharjee S, Wang Z, Ciampa J, Kraft P, Chanock S, Yu K, Chatterjee N (2010) Using Principal Components of Genetic Variation for Robust and Powerful Detection of Gene-Gene Interactions in Case-Control and Case-Only Studies. Am J Hum Genet 86: 331-342.

3. Yeager M, Chatterjee N, Ciampa J, Jacobs KB, Gonzalez-Bosquet J, Hayes RB, Kraft P, Wacholder S, Orr N, Berndt S, Yu K, Hutchinson A, Wang Z, Amundadottir L, Feigelson HS, Thun MJ, Diver WR, Albanes D, Virtamo J, Weinstein S, Schumacher FR, Cancel-Tassin G, Cussenot O, Valeri A, Andriole GL, Crawford ED, Haiman CA, Henderson B, Kolonel L, Le ML, Siddiq A, Riboli E, Key TJ, Kaaks R, Isaacs W, Isaacs S, Wiley KE, Gronberg H, Wiklund F, Stattin P, Xu J, Zheng SL, Sun J, Vatten LJ, Hveem K, Kumle M, Tucker M, Gerhard DS, Hoover RN, Fraumeni JF, Jr., Hunter DJ, Thomas G, Chanock SJ (2009) Identification of a new prostate cancer susceptibility locus on chromosome 8q24. Nat Genet 41: 1055-1057.

4. Yu K, Liang F, Ciampa J, Chatterjee N (2011) Efficient p-value evaluation for resampling-based tests. Biostatistics Epub: 1-11.

5. Chung CC, Ciampa J, Yeager M, Jacobs KB, Berndt SI, Hayes RB, Gonzalez-Bosquet J, Kraft P, Wacholder S, Orr N, Yu K, Hutchinson A, Boland J, Chen Q, Feigelson HS, Thun MJ, Diver WR, Albanes D, Virtamo J, Weinstein S, Schumacher FR, Cancel-Tassin G, Cussenot O, Valeri A, Andriole GL, Crawford ED, Haiman CA, Henderson BE, Kolonel L, Le ML, Siddiq A, Riboli E, Key TJ, Kaaks R, Isaacs WB, Isaacs SD, Grunberg H, Wiklund F, Xu J, Vatten LJ, Hveem K, Njolstad I, Gerhard DS, Tucker M, Hoover RN, Fraumeni JF, Jr., Hunter DJ, Thomas G, Chatterjee N, Chanock SJ (2011) Fine Mapping of a Region of Chromosome 11q13 Reveals Multiple Independent Loci Associated with Risk of Prostate Cancer. Hum Mol Genet Epub.

6. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M,

Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la BM, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de JP, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.

7. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, bu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di F, V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch

DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M (2001) The sequence of the human genome. Science 291: 1304-1351.

8. Collins FS, Guyer MS, Charkravarti A (1997) Variations on a theme: cataloging human DNA sequence variation. Science 278: 1580-1581.

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

10. Pearson TA, Manolio TA (2008) How to interpret a genome-wide association study. JAMA 299: 1335-1344.

11. The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851-861.

12. Thomas, D. C. (2004) Statistical Methods in Genetic Epidemiology. New York: Oxford University Press.

13. Manolio TA, Collins FS (2009) The HapMap and genome-wide association studies in diagnosis and therapy. Annu Rev Med 60:443-56.: 443-456.

14. Khoury MJ, Yang Q (1998) The future of genetic studies of complex human diseases: an epidemiologic perspective. Epidemiology 9: 350-354.

15. Syvanen AC (2005) Toward genome-wide SNP genotyping. Nat Genet 37 Suppl:S5-10.: S5-10.

16. Steemers FJ, Chang W, Lee G, Barker DL, Shen R, Gunderson KL (2006) Whole-genome genotyping with the single-base extension assay. Nat Methods 3: 31-33.

17. Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6: 95-108.

18. US Department of Health and Human Services (2011) GWAS (Genome Wide Assocation Studies).

19. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308: 385-389.

20. Manolio TA, Brooks LD, Collins FS (2008) A HapMap harvest of insights into the genetics of common disease. J Clin Invest 118: 1590-1605.

21. Hunter DJ, Kraft P (2007) Drinking from the fire hose--statistical issues in genomewide association studies. N Engl J Med 357: 436-439.

22. Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273: 1516-1517.

23. Libioulle C, Louis E, Hansoul S, Sandor C, Farnir F, Franchimont D, Vermeire S, Dewit O, de VM, Dixon A, Demarche B, Gut I, Heath S, Foglio M, Liang L, Laukens D, Mni M, Zelenika D, Van GA, Rutgeerts P, Belaiche J, Lathrop M, Georges M (2007) Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. PLoS Genet 20: 58.

24. Ghoussaini M, Song H, Koessler T, Al Olama AA, Kote-Jarai Z, Driver KE, Pooley KA, Ramus SJ, Kjaer SK, Hogdall E, DiCioccio RA, Whittemore AS, Gayther SA, Giles GG, Guy M, Edwards SM, Morrison J, Donovan JL, Hamdy FC, Dearnaley DP, rdern-Jones AT, Hall AL, O'Brien LT, Gehr-Swain BN, Wilkinson RA, Brown PM, Hopper JL, Neal DE, Pharoah PD, Ponder BA, Eeles RA, Easton DF, Dunning AM (2008) Multiple loci with different cancer specificities within the 8q24 gene desert. J Natl Cancer Inst 100: 962-966.

25. Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, Hirschhorn JN, Abecasis G, Altshuler D, Bailey-Wilson JE, Brooks LD, Cardon LR, Daly M, Donnelly P, Fraumeni JF, Jr., Freimer NB, Gerhard DS, Gunter C, Guttmacher AE, Guyer MS, Harris EL, Hoh J, Hoover R, Kong CA, Merikangas KR, Morton CC, Palmer LJ, Phimister EG, Rice JP, Roberts J, Rotimi C, Tucker MA, Vogan KJ, Wacholder S, Wijsman EM, Winn DM, Collins FS (2007) Replicating genotype-phenotype associations. Nature 447: 655-660.

26. National Cancer Institute (2011) Cancer Genetic Markers of Susceptibility; cgems.cancer.gov.

27. Brand A, Brand H, Schulte in den BT (2008) The impact of genetics and genomics on public health. Eur J Hum Genet 16: 5-13.

28. Guttmacher AE, Collins FS (2005) Realizing the promise of genomics in biomedical research. JAMA 294: 1399-1402.

29. Jakobsdottir J, Gorin MB, Conley YP, Ferrell RE, Weeks DE (2009) Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers. PLoS Genet 5: e1000337.

30. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P (2004) Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. Am J Epidemiol 159: 882-890.

31. The 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.

32. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ (1989) Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 20: 1879-1886.

33. National Cancer Institute (2011) Breast Cancer Risk Assessment Tool; www.cancer.gov/bcrisktool.

34. Gail MH (2008) Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk. J Natl Cancer Inst 100: 1037-1041.

35. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9: 356-369.

36. McCarthy MI (2004) Progress in defining the molecular basis of type 2 diabetes mellitus through susceptibility-gene identification. Hum Mol Genet 13 Spec No 1: R33-R41.

37. Grant SF, Hakonarson H (2007) Recent development in pharmacogenomics: from candidate genes to genome-wide association studies. Expert Rev Mol Diagn 7: 371-393.

38. Florez JC, Jablonski KA, Sun MW, Bayley N, Kahn SE, Shamoon H, Hamman RF, Knowler WC, Nathan DM, Altshuler D (2007) Effects of the type 2 diabetes-associated PPARG P12A polymorphism on progression to diabetes and response to troglitazone. J Clin Endocrinol Metab 92: 1502-1509.

39. Colca JR (2007) Future directions for insulin sensitizers in disease prevention. Curr Opin Investig Drugs 8: 707-710.

40. Hunter DJ, Thomas G, Hoover RN, Chanock SJ (2007) Scanning the horizon: what is the future of genome-wide association studies in accelerating discoveries in cancer etiology and prevention? Cancer Causes Control 18: 479-484.

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

42. Phillips PC (2008) Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. Nat Rev Genet 9: 855-867.

43. Rothman KJ, Greenland S, Walker AM (1980) Concepts of interaction. Am J Epidemiol 112: 467-470.

44. Mackay TF, Stone EA, Ayroles JF (2009) The genetics of quantitative traits: challenges and prospects. Nat Rev Genet 10: 565-577.

45. Moore JH (2003) The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Hum Hered 56: 73-82.

46. Bridges CB (1919) VERMILION-DEFICIENCY. J Gen Physiol 1: 645-656.

47. Boulon S, Dantonel JC, Binet V, Vie A, Blanchard JM, Hipskind RA, Philips A (2002) Oct-1 potentiates CREB-driven cyclin D1 promoter activation via a phospho-CREB- and CREB binding protein-independent mechanism. Mol Cell Biol 22: 7769-7779.

48. Brem RB, Storey JD, Whittle J, Kruglyak L (2005) Genetic interactions between polymorphisms that affect gene expression in yeast. Nature 436: 701-703.

49. Peripato AC, De Brito RA, Vaughn TT, Pletscher LS, Matioli SR, Cheverud JM (2002) Quantitative trait loci for maternal performance for offspring survival in mice. Genetics 162: 1341-1353.

50. Routman EJ, Cheverud JM (/10) Gene Effects on a Quantitative Trait: Two-Locus Epistatic Effects Measured at Microsatellite Markers and at Estimated QTL. Evolution 51: 1654-1662.

51. Wagner A (2000) Robustness against mutations in genetic networks of yeast. Nat Genet 24: 355-361.

52. Franklin I, Lewontin RC (1970) Is the gene the unit of selection? Genetics 65: 707-734.

53. Templeton AR, Sing CF, Brokaw B (1976) The unit of selection in Drosophila mercatorum. I. The interation of selection and meiosis in parthenogenetic strains. Genetics 82: 349-376.

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

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

56. Evans DM, Marchini J, Morris AP, Cardon LR (2006) Two-stage two-locus models in genome-wide association. PLoS Genet 2: e157.

57. Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ (2007) Exploiting gene-environment interaction to detect genetic associations. Hum Hered 63: 111-119.

58. Marchini J, Donnelly P, Cardon LR (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 37: 413-417.

59. Mukherjee B, Ahn J, Gruber SB, Rennert G, Moreno V, Chatterjee N (2008) Tests for gene-environment interaction from case-control data: a novel study of type I error, power and designs. Genet Epidemiol 32: 615-626.

60. Carlborg O, Kerje S, Schutz K, Jacobsson L, Jensen P, Andersson L (2003) A global search reveals epistatic interaction between QTL for early growth in the chicken. Genome Res 13: 413-421.

61. Hsueh WC, Cole SA, Shuldiner AR, Beamer BA, Blangero J, Hixson JE, MacCluer JW, Mitchell BD (2001) Interactions between variants in the beta3-adrenergic receptor and peroxisome proliferator-activated receptor-gamma2 genes and obesity. Diabetes Care 24: 672-677.

62. Leamy LJ, Routman EJ, Cheverud JM (2002) An epistatic genetic basis for fluctuating asymmetry of mandible size in mice. Evolution 56: 642-653.

63. Tripodis N, Hart AA, Fijneman RJ, Demant P (2001) Complexity of lung cancer modifiers: mapping of thirty genes and twenty-five interactions in half of the mouse genome. J Natl Cancer Inst 93: 1484-1491.

64. van WT, Ruivenkamp CA, Stassen AP, Moen CJ, Demant P (1999) Four new colon cancer susceptibility loci, Scc6 to Scc9 in the mouse. Cancer Res 59: 4216-4218.

65. Herold C, Steffens M, Brockschmidt FF, Baur MP, Becker T (2009) INTERSNP: genome-wide interaction analysis guided by a priori information. Bioinformatics 25: 3275-3281.

66. Chatterjee N, Mukherjee B (2008) Statistical approaches to studies of gene-gene and gene-environment interactions. In: Rebbeck T, Ambrosone C, Shields P, editors. Molecular Epidemiology: Applications in Cancer and Other Human Disease. New York: Informa Healthcare USA, Inc.

67. Smith PG, Day NE (1984) The design of case-control studies: the influence of confounding and interaction effects. Int J Epidemiol 13: 356-365.

68. Siemiatycki J, Thomas DC (1981) Biological models and statistical interactions: an example from multistage carcinogenesis. Int J Epidemiol 10: 383-387.

69. Thompson WD (1991) Effect modification and the limits of biological inference from epidemiologic data. J Clin Epidemiol 44: 221-232.

70. Moore JH, Ritchie MD (2004) STUDENTJAMA. The challenges of whole-genome approaches to common diseases. JAMA 291: 1642-1643.

71. NNF Center for Protein Reserach, University of Copenhagen, European Molecular BIology Labortory, Technical University Dresden Biotechnology, Swiss Institute of Bioinformatics et al. (2011) STRING 8.3: Known and predicted protein-protein interactions; string-db.org.

72. Broad Institute of Massachusetts Institute of Technology and Harvard (2011) Gene Relationships Across Implicated Loci; www.broadinstitute.org/mpg/grail.

73. National Cancer Institute, Nature Publishing Group (2011) Pathway Interaction Database; pid.nci.nih.gov.

74. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Thun MJ (2009) Cancer statistics, 2009. CA Cancer J Clin 59: 225-249.

75. Crawford ED (2009) Understanding the epidemiology, natural history, and key pathways involved in prostate cancer. Urology 73: S4-10.

76. Farkas A, Schneider D, Perrotti M, Cummings KB, Ward WS (1998) National trends in the epidemiology of prostate cancer, 1973 to 1994: evidence for the effectiveness of prostate-specific antigen screening. Urology 52: 444-448.

77. Hankey BF, Feuer EJ, Clegg LX, Hayes RB, Legler JM, Prorok PC, Ries LA, Merrill RM, Kaplan RS (1999) Cancer surveillance series: interpreting trends in prostate cancer--part I: Evidence of the effects of screening in recent prostate cancer incidence, mortality, and survival rates. J Natl Cancer Inst 91: 1017-1024.

78. Gronberg H (2003) Prostate cancer epidemiology. Lancet 361: 859-864.

79. Miller GJ, Torkko KC (2001) Natural history of prostate cancer--epidemiologic considerations. Epidemiol Rev 23: 14-18.

80. Buchanan G, Irvine RA, Coetzee GA, Tilley WD (2001) Contribution of the androgen receptor to prostate cancer predisposition and progression. Cancer Metastasis Rev 20: 207-223.

81. Verras M, Sun Z (2006) Roles and regulation of Wnt signaling and beta-catenin in prostate cancer. Cancer Lett 237: 22-32.

82. Nesbit CE, Tersak JM, Prochownik EV (1999) MYC oncogenes and human neoplastic disease. Oncogene 18: 3004-3016.

83. MacDonald BT, Tamai K, He X (2009) Wnt/beta-catenin signaling: components, mechanisms, and diseases. Dev Cell 17: 9-26.

84. Schaeffer EM, Marchionni L, Huang Z, Simons B, Blackman A, Yu W, Parmigiani G, Berman DM (2008) Androgen-induced programs for prostate epithelial growth and invasion arise in embryogenesis and are reactivated in cancer. Oncogene 27: 7180-7191.

85. Ghadirian P, Cadotte M, Lacroix A, Perret C (1991) Family aggregation of cancer of the prostate in Quebec: the tip of the iceberg. Prostate 19: 43-52.

86. Goldgar DE, Easton DF, Cannon-Albright LA, Skolnick MH (1994) Systematic population-based assessment of cancer risk in first-degree relatives of cancer probands. J Natl Cancer Inst 86: 1600-1608.

87. Krain LS (1974) Some epidemiologic variables in prostatic carcinoma in California. Prev Med 3: 154-159.

88. WOOLF CM (1960) An investigation of the familial aspects of carcinoma of the prostate. Cancer 13:739-44.: 739-744.

89. Gronberg H, Damber L, Damber JE (1994) Studies of genetic factors in prostate cancer in a twin population. J Urol 152: 1484-1487.

90. Page WF, Braun MM, Partin AW, Caporaso N, Walsh P (1997) Heredity and prostate cancer: a study of World War II veteran twins. Prostate 33: 240-245.

91. Hayes RB, Liff JM, Pottern LM, Greenberg RS, Schoenberg JB, Schwartz AG, Swanson GM, Silverman DT, Brown LM, Hoover RN, . (1995) Prostate cancer risk in U.S. blacks and whites with a family history of cancer. Int J Cancer 60: 361-364.

92. Johns LE, Houlston RS (2003) A systematic review and meta-analysis of familial prostate cancer risk. BJU Int 91: 789-794.

93. Spitz MR, Currier RD, Fueger JJ, Babaian RJ, Newell GR (1991) Familial patterns of prostate cancer: a case-control analysis. J Urol 146: 1305-1307.

94. Steinberg GD, Carter BS, Beaty TH, Childs B, Walsh PC (1990) Family history and the risk of prostate cancer. Prostate 17: 337-347.

95. Whittemore AS, Wu AH, Kolonel LN, John EM, Gallagher RP, Howe GR, West DW, Teh CZ, Stamey T (1995) Family history and prostate cancer risk in black, white, and Asian men in the United States and Canada. Am J Epidemiol 141: 732-740.

96. Hemminki K, Ji J, Forsti A, Sundquist J, Lenner P (2008) Concordance of survival in family members with prostate cancer. J Clin Oncol 26: 1705-1709.

97. Eeles RA, Kote-Jarai Z, Giles GG, Olama AA, Guy M, Jugurnauth SK, Mulholland S, Leongamornlert DA, Edwards SM, Morrison J, Field HI, Southey MC, Severi G, Donovan JL, Hamdy FC, Dearnaley DP, Muir KR, Smith C, Bagnato M, rdern-Jones AT, Hall AL, O'Brien LT, Gehr-Swain BN, Wilkinson RA, Cox A, Lewis S, Brown PM, Jhavar SG, Tymrakiewicz M, Lophatananon A, Bryant SL, Horwich A, Huddart RA, Khoo VS, Parker CC, Woodhouse CJ, Thompson A, Christmas T, Ogden C, Fisher C, Jamieson C, Cooper CS, English DR, Hopper JL, Neal DE, Easton DF (2008) Multiple newly identified loci associated with prostate cancer susceptibility. Nat Genet 40: 316-321.

98. Gudmundsson J, Sulem P, Rafnar T, Bergthorsson JT, Manolescu A, Gudbjartsson D, Agnarsson BA, Sigurdsson A, Benediktsdottir KR, Blondal T, Jakobsdottir M, Stacey SN, Kostic J, Kristinsson KT, Birgisdottir B, Ghosh S, Magnusdottir DN, Thorlacius S, Thorleifsson G, Zheng SL, Sun J, Chang BL, Elmore JB, Breyer JP, McReynolds KM, Bradley KM, Yaspan BL, Wiklund F, Stattin P, Lindstrom S, Adami HO, McDonnell SK, Schaid DJ, Cunningham JM, Wang L, Cerhan JR, St Sauver JL, Isaacs SD, Wiley KE, Partin AW, Walsh PC, Polo S, Ruiz-Echarri M, Navarrete S, Fuertes F, Saez B, Godino J, Weijerman PC, Swinkels DW, Aben KK, Witjes JA, Suarez BK, Helfand BT, Frigge ML, Kristjansson K, Ober C, Jonsson E, Einarsson GV, Xu J, Gronberg H, Smith JR, Thibodeau SN, Isaacs WB, Catalona WJ, Mayordomo JI, Kiemeney LA, Barkardottir RB, Gulcher JR, Thorsteinsdottir U, Kong A, Stefansson K (2008) Common sequence variants on 2p15 and Xp11.22 confer susceptibility to prostate cancer. Nat Genet 40: 281-283.

99. Gudmundsson J, Sulem P, Gudbjartsson DF, Blondal T, Gylfason A, Agnarsson BA, Benediktsdottir KR, Magnusdottir DN, Orlygsdottir G, Jakobsdottir M, Stacey SN, Sigurdsson A, Wahlfors T, Tammela T, Breyer JP, McReynolds KM, Bradley KM, Saez B, Godino J, Navarrete S, Fuertes F, Murillo L, Polo E, Aben KK, van O, I, Suarez BK, Helfand BT, Kan D, Zanon C, Frigge ML, Kristjansson K, Gulcher JR, Einarsson GV, Jonsson E, Catalona WJ, Mayordomo JI, Kiemeney LA, Smith JR, Schleutker J, Barkardottir RB, Kong A, Thorsteinsdottir U, Rafnar T, Stefansson K (2009) Genome-wide association and replication studies identify four variants associated with prostate cancer susceptibility. Nat Genet 41: 1122-1126.

100. Gudmundsson J, Sulem P, Steinthorsdottir V, Bergthorsson JT, Thorleifsson G, Manolescu A, Rafnar T, Gudbjartsson D, Agnarsson BA, Baker A, Sigurdsson A, Benediktsdottir KR, Jakobsdottir M, Blondal T, Stacey SN, Helgason A, Gunnarsdottir S, Olafsdottir A, Kristinsson KT, Birgisdottir B, Ghosh S, Thorlacius S, Magnusdottir D, Stefansdottir G, Kristjansson K, Bagger Y, Wilensky RL, Reilly MP, Morris AD, Kimber CH, Adeyemo A, Chen Y, Zhou J, So WY, Tong PC, Ng MC,

Hansen T, Andersen G, Borch-Johnsen K, Jorgensen T, Tres A, Fuertes F, Ruiz-Echarri M, Asin L, Saez B, van BE, Klaver S, Swinkels DW, Aben KK, Graif T, Cashy J, Suarez BK, van Vierssen TO, Frigge ML, Ober C, Hofker MH, Wijmenga C, Christiansen C, Rader DJ, Palmer CN, Rotimi C, Chan JC, Pedersen O, Sigurdsson G, Benediktsson R, Jonsson E, Einarsson GV, Mayordomo JI, Catalona WJ, Kiemeney LA, Barkardottir RB, Gulcher JR, Thorsteinsdottir U, Kong A, Stefansson K (2007) Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. Nat Genet 39: 977-983.

101. Sun J, Purcell L, Gao Z, Isaacs SD, Wiley KE, Hsu FC, Liu W, Duggan D, Carpten JD, Gronberg H, Xu J, Chang BL, Partin AW, Walsh PC, Isaacs WB, Zheng SL (2008) Association between sequence variants at 17q12 and 17q24.3 and prostate cancer risk in European and African Americans. Prostate 68: 691-697.

102. Sun J, Zheng SL, Wiklund F, Isaacs SD, Purcell LD, Gao Z, Hsu FC, Kim ST, Liu W, Zhu Y, Stattin P, Adami HO, Wiley KE, Dimitrov L, Sun J, Li T, Turner AR, Adams TS, Adolfsson J, Johansson JE, Lowey J, Trock BJ, Partin AW, Walsh PC, Trent JM, Duggan D, Carpten J, Chang BL, Gronberg H, Isaacs WB, Xu J (2008) Evidence for two independent prostate cancer risk-associated loci in the HNF1B gene at 17q12. Nat Genet 40: 1153-1155.

103. Thomas G, Jacobs KB, Yeager M, Kraft P, Wacholder S, Orr N, Yu K, Chatterjee N, Welch R, Hutchinson A, Crenshaw A, Cancel-Tassin G, Staats BJ, Wang Z, Gonzalez-Bosquet J, Fang J, Deng X, Berndt SI, Calle EE, Feigelson HS, Thun MJ, Rodriguez C, Albanes D, Virtamo J, Weinstein S, Schumacher FR, Giovannucci E, Willett WC, Cussenot O, Valeri A, Andriole GL, Crawford ED, Tucker M, Gerhard DS, Fraumeni JF, Jr., Hoover R, Hayes RB, Hunter DJ, Chanock SJ (2008) Multiple loci identified in a genome-wide association study of prostate cancer. Nat Genet 40: 310-315.

104. Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, Wacholder S, Minichiello MJ, Fearnhead P, Yu K, Chatterjee N, Wang Z, Welch R, Staats BJ, Calle EE, Feigelson HS, Thun MJ, Rodriguez C, Albanes D, Virtamo J, Weinstein S, Schumacher FR, Giovannucci E, Willett WC, Cancel-Tassin G, Cussenot O, Valeri A, Andriole GL, Gelmann EP, Tucker M, Gerhard DS, Fraumeni JF, Jr., Hoover R, Hunter DJ, Chanock SJ, Thomas G (2007) Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat Genet 39: 645-649.

105. Zheng SL, Sun J, Cheng Y, Li G, Hsu FC, Zhu Y, Chang BL, Liu W, Kim JW, Turner AR, Gielzak M, Yan G, Isaacs SD, Wiley KE, Sauvageot J, Chen HS, Gurganus R, Mangold LA, Trock BJ, Gronberg H, Duggan D, Carpten JD, Partin AW, Walsh PC, Xu J, Isaacs WB (2007) Association between two unlinked loci at 8q24 and prostate cancer risk among European Americans. J Natl Cancer Inst 99: 1525-1533.

106. Zheng SL, Stevens VL, Wiklund F, Isaacs SD, Sun J, Smith S, Pruett K, Wiley KE, Kim ST, Zhu Y, Zhang Z, Hsu FC, Turner AR, Johansson JE, Liu W, Kim JW, Chang BL, Duggan D, Carpten J, Rodriguez C, Isaacs W, Gronberg H, Xu J (2009) Two independent prostate cancer risk-associated Loci at 11q13. Cancer Epidemiol Biomarkers Prev 18: 1815-1820.

107. Wokolorczyk D, Gliniewicz B, Sikorski A, Zlowocka E, Masojc B, Debniak T, Matyjasik J, Mierzejewski M, Medrek K, Oszutowska D, Suchy J, Gronwald J, Teodorczyk U, Huzarski T, Byrski T, Jakubowska A, Gorski B, van de WT, Walczak S, Narod SA, Lubinski J, Cybulski C (2008) A range of cancers is associated with the rs6983267 marker on chromosome 8. Cancer Res 68: 9982-9986.

108. Gearhart J, Pashos EE, Prasad MK (2007) Pluripotency redux--advances in stem-cell research. N Engl J Med 357: 1469-1472.

109. Pomerantz MM, Ahmadiyeh N, Jia L, Herman P, Verzi MP, Doddapaneni H, Beckwith CA, Chan JA, Hills A, Davis M, Yao K, Kehoe SM, Lenz HJ, Haiman CA, Yan C, Henderson BE, Frenkel B, Barretina J, Bass A, Tabernero J, Baselga J, Regan MM, Manak JR, Shivdasani R, Coetzee GA, Freedman ML (2009) The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. Nat Genet 41: 882-884.

110. Tuupanen S, Turunen M, Lehtonen R, Hallikas O, Vanharanta S, Kivioja T, Bjorklund M, Wei G, Yan J, Niittymaki I, Mecklin JP, Jarvinen H, Ristimaki A, Di-Bernardo M, East P, Carvajal-Carmona L, Houlston RS, Tomlinson I, Palin K, Ukkonen E, Karhu A, Taipale J, Aaltonen LA (2009) The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. Nat Genet 41: 885-890.

111. Zanke BW, Greenwood CM, Rangrej J, Kustra R, Tenesa A, Farrington SM, Prendergast J, Olschwang S, Chiang T, Crowdy E, Ferretti V, Laflamme P, Sundararajan S, Roumy S, Olivier JF, Robidoux F, Sladek R, Montpetit A, Campbell P, Bezieau S, O'Shea AM, Zogopoulos G, Cotterchio M, Newcomb P, McLaughlin J, Younghusband B, Green R, Green J, Porteous ME, Campbell H, Blanche H, Sahbatou M, Tubacher E, Bonaiti-Pellie C, Buecher B, Riboli E, Kury S, Chanock SJ, Potter J, Thomas G, Gallinger S, Hudson TJ, Dunlop MG (2007) Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. Nat Genet 39: 989-994.

112. Wright JB, Brown SJ, Cole MD (2010) Upregulation of c-MYC in cis through a Large Chromatin Loop Linked to a Cancer Risk-Associated Single-Nucleotide Polymorphism in Colorectal Cancer Cells. Mol Cell Biol 30: 1411-1420.

113. Ahmadiyeh N, Pomerantz MM, Grisanzio C, Herman P, Jia L, Almendro V, He HH, Brown M, Liu XS, Davis M, Caswell JL, Beckwith CA, Hills A,

Macconaill L, Coetzee GA, Regan MM, Freedman ML (2010) 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. Proc Natl Acad Sci U S A 107: 9742-9746.

114. Sotelo J, Esposito D, Duhagon MA, Banfield K, Mehalko J, Liao H, Stephens RM, Harris TJ, Munroe DJ, Wu X (2010) Long-range enhancers on 8q24 regulate c-Myc. Proc Natl Acad Sci U S A 107: 3001-3005.

115. International HapMap Project (2010) HapMap Genome Browser, Phase 1 & 2 full dataset for Utah residents with ancestry from northern and western Europe; hapmap.ncbi.nlm.nih.gov.

116. National Cancer Institute (2009) Surveillance Epidemiology and End Results, Complete Prevalence Program, Version 1.2 - Beta 1; seer.cancer.gov.

117. Suo G, Han J, Wang X, Zhang J, Zhao Y, Zhao Y, Dai J (2005) Oct4 pseudogenes are transcribed in cancers. Biochem Biophys Res Commun 337: 1047-1051.

118. Boiani M, Scholer HR (2005) Regulatory networks in embryo-derived pluripotent stem cells. Nat Rev Mol Cell Biol 6: 872-884.

119. Pan GJ, Chang ZY, Scholer HR, Pei D (2002) Stem cell pluripotency and transcription factor Oct4. Cell Res 12: 321-329.

120. Kastler S, Honold L, Luedeke M, Kuefer R, Moller P, Hoegel J, Vogel W, Maier C, Assum G (2009) POU5F1P1, a putative cancer susceptibility gene, is overexpressed in prostatic carcinoma. Prostate 70: 666-674.

121. Panagopoulos I, Moller E, Collin A, Mertens F (2008) The POU5F1P1 pseudogene encodes a putative protein similar to POU5F1 isoform 1. Oncol Rep 20: 1029-1033.

122. Pal P, Xi H, Guha S, Sun G, Helfand BT, Meeks JJ, Suarez BK, Catalona WJ, Deka R (2009) Common variants in 8q24 are associated with risk for prostate cancer and tumor aggressiveness in men of European ancestry. Prostate 69: 1548-1556.

123. Xiao X, Li BX, Mitton B, Ikeda A, Sakamoto KM (2010) Targeting CREB for Cancer Therapy: Friend or Foe. Curr Cancer Drug Targets 10: 384-391.

124. Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Jakobsdottir M, Steinberg S, Gudjonsson SA, Palsson A, Thorleifsson G, Palsson S, Sigurgeirsson B, Thorisdottir K, Ragnarsson R, Benediktsdottir KR, Aben KK, Vermeulen SH, Goldstein AM, Tucker MA, Kiemeney LA, Olafsson JH, Gulcher J, Kong A, Thorsteinsdottir U, Stefansson K (2008) Two newly identified genetic determinants of pigmentation in Europeans. Nat Genet 40: 835-837.

125. Janssen JW, Vaandrager JW, Heuser T, Jauch A, Kluin PM, Geelen E, Bergsagel PL, Kuehl WM, Drexler HG, Otsuki T, Bartram CR, Schuuring E (2000) Concurrent activation of a novel putative

transforming gene, myeov, and cyclin D1 in a subset of multiple myeloma cell lines with t(11;14)(q13;q32). Blood 95: 2691-2698.

126. Fagegaltier D, Hubert N, Yamada K, Mizutani T, Carbon P, Krol A (2000) Characterization of mSelB, a novel mammalian elongation factor for selenoprotein translation. EMBO J 19: 4796-4805.

127. Kolatsi-Joannou M, Bingham C, Ellard S, Bulman MP, Allen LI, Hattersley AT, Woolf AS (2001) Hepatocyte nuclear factor-1beta: a new kindred with renal cysts and diabetes and gene expression in normal human development. J Am Soc Nephrol 12: 2175-2180.

128. Nakajima T, Fujino S, Nakanishi G, Kim YS, Jetten AM (2004) TIP27: a novel repressor of the nuclear orphan receptor TAK1/TR4. Nucleic Acids Res 32: 4194-4204.

129. Frayling TM, Colhoun H, Florez JC (2008) A genetic link between type 2 diabetes and prostate cancer. Diabetologia 51: 1757-1760.

130. Chang BL, Cramer SD, Wiklund F, Isaacs SD, Stevens VL, Sun J, Smith S, Pruett K, Romero LM, Wiley KE, Kim ST, Zhu Y, Zhang Z, Hsu FC, Turner AR, Adolfsson J, Liu W, Kim JW, Duggan D, Carpten J, Zheng SL, Rodriguez C, Isaacs WB, Gronberg H, Xu J (2009) Fine mapping association study and functional analysis implicate a SNP in MSMB at 10q11 as a causal variant for prostate cancer risk. Hum Mol Genet 18: 1368-1375.

131. Lou H, Yeager M, Li H, Bosquet JG, Hayes RB, Orr N, Yu K, Hutchinson A, Jacobs KB, Kraft P, Wacholder S, Chatterjee N, Feigelson HS, Thun MJ, Diver WR, Albanes D, Virtamo J, Weinstein S, Ma J, Gaziano JM, Stampfer M, Schumacher FR, Giovannucci E, Cancel-Tassin G, Cussenot O, Valeri A, Andriole GL, Crawford ED, Anderson SK, Tucker M, Hoover RN, Fraumeni JF, Jr., Thomas G, Hunter DJ, Dean M, Chanock SJ (2009) Fine mapping and functional analysis of a common variant in MSMB on chromosome 10q11.2 associated with prostate cancer susceptibility. Proc Natl Acad Sci U S A 106: 7933-7938.

132. National Cancer Institute (2011) Cancer Genetic Markers of Susceptibility.

133. Agresti, A (1996) An Introduction to Categorical Data Analysis. New York: John Wiley and Sons, Inc.

134. Mukherjee B, Chatterjee N (2008) Exploiting gene-environment independence for analysis of case-control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. Biometrics 64: 685-694.

135. Prentice RL, Pyke R (1979) Logistic Disease Incidence Models and Case-Control Studies. Biometrika 66: 403-411.

136. Confield J (1956) A statistical problem arising from retrospective studies. Proc Third Berkely Symp Math, Statist, Prob 4: 135-148.

137. Piegorsch WW, Weinberg CR, Taylor JA (1994) Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. Stat Med 13: 153-162.

138. Yang Q, Khoury MJ, Sun F, Flanders WD (1999) Case-only design to measure gene-gene interaction. Epidemiology 10: 167-170.

139. Khoury MJ, Flanders WD (1996) Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls! Am J Epidemiol 144: 207-213.

140. Becher H, Schmidt S, Chang-Claude J (2003) Reproductive factors and familial predisposition for breast cancer by age 50 years. A case-control-family study for assessing main effects and possible gene-environment interaction. Int J Epidemiol 32: 38-48.

141. Chang-Claude J, Dunning A, Schnitzbauer U, Galmbacher P, Tee L, Wjst M, Chalmers J, Zemzoum I, Harbeck N, Pharoah PD, Hahn H (2003) The patched polymorphism Pro1315Leu (C3944T) may modulate the association between use of oral contraceptives and breast cancer risk. Int J Cancer 103: 779-783.

142. Egan KM, Newcomb PA, Titus-Ernstoff L, Trentham-Dietz A, Mignone LI, Farin F, Hunter DJ (2003) Association of NAT2 and smoking in relation to breast cancer incidence in a population-based case-control study (United States). Cancer Causes Control 14: 43-51.

143. Ghadirian P, Narod S, Fafard E, Costa M, Robidoux A, Nkondjock A (2009) Breast cancer risk in relation to the joint effect of BRCA mutations and diet diversity. Breast Cancer Res Treat 117: 417-422.

144. Infante-Rivard C, Labuda D, Krajinovic M, Sinnett D (1999) Risk of childhood leukemia associated with exposure to pesticides and with gene polymorphisms. Epidemiology 10: 481-487.

145. Marcus PM, Hayes RB, Vineis P, Garcia-Closas M, Caporaso NE, Autrup H, Branch RA, Brockmoller J, Ishizaki T, Karakaya AE, Ladero JM, Mommsen S, Okkels H, Romkes M, Roots I, Rothman N (2000) Cigarette smoking, N-acetyltransferase 2 acetylation status, and bladder cancer risk: a case-series meta-analysis of a gene-environment interaction. Cancer Epidemiol Biomarkers Prev 9: 461-467.

146. Modan B, Hartge P, Hirsh-Yechezkel G, Chetrit A, Lubin F, Beller U, Ben-Baruch G, Fishman A, Menczer J, Struewing JP, Tucker MA, Wacholder S (2001) Parity, oral contraceptives, and the risk of ovarian cancer among carriers and noncarriers of a BRCA1 or BRCA2 mutation. N Engl J Med 345: 235-240.

147. Pasanisi P, Hedelin G, Berrino J, Chang-Claude J, Hermann S, Steel M, Haites N, Hart J, Peled R, Gafa L, Leggio L, Traina A, Amodio R, Primic-Zakelj M, Zadnik V, Veidebaum T, Tekkel M, Berrino F (2009) Oral contraceptive use and BRCA penetrance: a case-only study. Cancer Epidemiol Biomarkers Prev 18: 2107-2113.

148. Prentice RL, Huang Y, Hinds DA, Peters U, Cox DR, Beilharz E, Chlebowski RT, Rossouw JE, Caan B, Ballinger DG (2010) Variation in the FGFR2 gene and the effect of a low-fat dietary pattern on invasive breast cancer. Cancer Epidemiol Biomarkers Prev 19: 74-79.

149. Sturmer T, Wang-Gohrke S, Arndt V, Boeing H, Kong X, Kreienberg R, Brenner H (2002) Interaction between alcohol dehydrogenase II gene, alcohol consumption, and risk for breast cancer. Br J Cancer 87: 519-523.

150. Moorman PG, Iversen ES, Marcom PK, Marks JR, Wang F, Lee E, Ursin G, Rebbeck TR, Domchek SM, Arun B, Susswein L, Isaacs C, Garber JE, Visvanathan K, Griffin CA, Sutphen R, Brzosowicz J, Gruber S, Finkelstein DM, Schildkraut JM (2010) Evaluation of established breast cancer risk factors as modifiers of BRCA1 or BRCA2: a multi-center case-only analysis. Breast Cancer Res Treat 124: 441-451.

151. Albert PS, Ratnasinghe D, Tangrea J, Wacholder S (2001) Limitations of the case-only design for identifying gene-environment interactions. Am J Epidemiol 154: 687-693.

152. Chatterjee N, Carroll RJ (2005) Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. Biometrika 92: 399-418.

153. Umbach DM, Weinberg CR (1997) Designing and analysing case-control studies to exploit independence of genotype and exposure. Stat Med 16: 1731-1743.

154. Greenland S (1993) Methods for epidemiologic analyses of multiple exposures: a review and comparative study of maximum-likelihood, preliminary-testing, and empirical-Bayes regression. Stat Med 12: 717-736.

155. 2011) The R Project for Statistical Computing; www.r-project.org.

156. Wald A (1941) Asymptotically most powerful tests of statistical hypotheses. The Annals of Mathematical Statistics 12: 1-19.

157. Lander ES, Schork NJ (1994) Genetic dissection of complex traits. Science 265: 2037-2048.

158. Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55: 997-1004.

159. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904-909.

160. Hastie A, Tibshirani R Friedman JH (2001) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

161. Li Q, Yu K (2008) Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. Genet Epidemiol 32: 215-226.

162. Yu K, Wang Z, Li Q, Wacholder S, Hunter DJ, Hoover RN, Chanock S, Thomas G (2008) Population substructure and control selection in genome-wide association studies. PLoS One 3: e2551.

163. Marchini J, Cardon LR, Phillips MS, Donnelly P (2004) The effects of human population structure on large genetic association studies. Nat Genet 36: 512-517.

164. Thomas DC, Witte JS (2002) Point: population stratification: a problem for case-control studies of candidate-gene associations? Cancer Epidemiol Biomarkers Prev 11: 505-512.

165. Helgason A, Yngvadottir B, Hrafnkelsson B, Gulcher J, Stefansson K (2005) An Icelandic example of the impact of population structure on association studies. Nat Genet 37: 90-95.

166. Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, Groop LC, Altshuler D, Ardlie KG, Hirschhorn JN (2005) Demonstrating stratification in a European American population. Nat Genet 37: 868-872.

167. Wacholder S, Rothman N, Caporaso N (2002) Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. Cancer Epidemiol Biomarkers Prev 11: 513-520.

168. Lee WC, Wang LY (2009) Reducing population stratification bias: stratum matching is better than exposure. J Clin Epidemiol 62: 62-66.

169. Wang LY, Lee WC (2008) Population stratification bias in the case-only study for gene-environment interactions. Am J Epidemiol 168: 197-201.

170. Luca D, Ringquist S, Klei L, Lee AB, Gieger C, Wichmann HE, Schreiber S, Krawczak M, Lu Y, Styche A, Devlin B, Roeder K, Trucco M (2008) On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. Am J Hum Genet 82: 453-463.

171. Gatto NM, Campbell UB, Rundle AG, Ahsan H (2004) Further development of the case-only design for assessing gene-environment interaction: evaluation of and adjustment for bias. Int J Epidemiol 33: 1014-1024.

172. Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. Nature 402: C47-C52.

173. Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, Chen Y, Cheng X, Chua G, Friesen H, Goldberg DS, Haynes J, Humphries C, He G, Hussein S, Ke L, Krogan N, Li Z, Levinson JN, Lu H, Menard P, Munyana C, Parsons AB, Ryan O, Tonikian R, Roberts T, Sdicu AM, Shapiro J, Sheikh B, Suter B, Wong SL, Zhang LV, Zhu H, Burd CG, Munro S, Sander C, Rine J, Greenblatt J, Peter M, Bretscher A, Bell G, Roth FP, Brown GW, Andrews B, Bussey H, Boone C (2004) Global mapping of the yeast genetic interaction network. Science 303: 808-813.

174. Segre D, Deluna A, Church GM, Kishony R (2005) Modular epistasis in yeast metabolism. Nat Genet 37: 77-83.

175. Chatterjee N, Kalaylioglu Z, Moslehi R, Peters U, Wacholder S (2006) Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. Am J Hum Genet 79: 1002-1016.

176. Tukey JW (1948) One Degree of Freedom for Non-Additivity. Biometrics 5: 232-242.

177. Rao CR (1948) Large sample tests of statistical hypotheses concerning sevral parameters with applications to problems of estimation. Proceedings of the Cambridge Philosophical Society 44: 50-57.

178. Cox, DR and Snell EJ (1989) Analysis of Binary Data. London: Chapman and Hall.

179. Buse A (1982) The likelihood ratio, wald and lagrange multiplier tests: an expository note. The American Statistician 36: 153-157.

180. Rice J (1995) Mathematical Statistics and Data Analysis. International Thomson Publishing.

181. Edwards, A. W. F. (1992) Likelihood: Expanded Edition. Baltimore: The Johns Hopkins University Press.

182. Lloyd CJ (1999) Statistical Analysis of Categorical Data. New York: John Wiley & Songs, Inc.

183. Catchpole EA, Morgan BJT (1996) Model selection in ring-recovery models using score tests. Biometrics 52: 672.

184. Hosking JRM (1984) Testing whether the shape parameter is zero in the generalized extreme-value distribution. Biometrika 71: 374.

185. Storer BE, Wacholder S, Breslow NE (1983) Maximum likelihood fitting of general risk models to stratified data. Applied Statistics 32: 172-181.

186. Freedman D (2007) How can the score test be inconsistent? The American Statistician 291-295.

187. Lawrance AJ (1987) The score statistic for regression transformation. Biometrika 74: 275-279.

188. Morgan B, Palmer K, Ridout M (2007) Negative score test statistic. The American Statistician 61: 285-288.

189. Khoury, M. J., Beaty, T. H., and Cohen, B. H. (1993) Fundamentals of Genetic Epidemiology. New York: Oxford University Press.

190. Lin DY, Zou F (2004) Assessing genomewide statistical significance in linkage studies. Genet Epidemiol 27: 202-214.

191. Metropolis N, Rosenblut A, Rosenbluth M, Teller A, Teller E (1953) Equations of state calculations by fast computing machines. Journal of Chemical Physics 21: 1087-1092.

192. Liang F, Liu C, Carroll R (2007) Stochastic approximation in Monte Carlo computation. Journal of the American Statistical Association 102: 305-320.

193. Hastings W (1970) Monte Carlo sampling methods using Markov chains and their application. Biometrika 57: 97-109.

194. Kaas RE, Raftery AE (1995) Bayes Factors. Journal of the American Statistical Association 90: 773-795.

195. Dai J, Kooperberg C, LeBlanc ML, Prentice R (2010) One two-stage hypothesis testing procedures via asymptotically independent statistics. UW Biostatistics Working Paper Series, Paper 366.

196. Liu Y, Xu H, Chen S, Chen X, Zhang Z, Zhu Z, Qin X, Hu L, Zhu J, Zhao GP, Kong X (2011) Genome-wide interaction-based association analysis identified multiple new susceptibility Loci for common diseases. PLoS Genet 7: e1001338.

197. Millstein J, Conti DV, Gilliland FD, Gauderman WJ (2006) A testing framework for identifying susceptibility genes in the presence of epistasis. Am J Hum Genet 78: 15-27.

198. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 69: 138-147.

199. Manke IA, Lowery DM, Nguyen A, Yaffe MB (2003) BRCT repeats as phosphopeptide-binding modules involved in protein targeting. Science 302: 636-639.

200. Wacholder S, Chanock S, Garcia-Closas M, El GL, Rothman N (2004) Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. J Natl Cancer Inst 96: 434-442.

201. Wakefield J (2007) A Bayesian measure of the probability of false discovery in genetic epidemiology studies. Am J Hum Genet 81: 208-227.

202. Rankin EB, Giaccia AJ (2008) The role of hypoxia-inducible factors in tumorigenesis. Cell Death Differ 15: 678-685.

203. Forristal CE, Wright KL, Hanley NA, Oreffo RO, Houghton FD (2010) Hypoxia inducible factors regulate pluripotency and proliferation in human embryonic stem cells cultured at reduced oxygen tensions. Reproduction 139: 85-97.

204. Simon MC, Keith B (2008) The role of oxygen availability in embryonic development and stem cell function. Nat Rev Mol Cell Biol 9: 285-296.

205. Covello KL, Kehler J, Yu H, Gordan JD, Arsham AM, Hu CJ, Labosky PA, Simon MC, Keith B (2006) HIF-2alpha regulates Oct-4: effects of hypoxia on stem cell function, embryonic development, and tumor growth. Genes Dev 20: 557-570.

206. Waltzer L, Bienz M (1999) The control of beta-catenin and TCF during embryonic development and cancer. Cancer Metastasis Rev 18: 231-246.

207. Hochedlinger K, Yamada Y, Beard C, Jaenisch R (2005) Ectopic expression of Oct-4 blocks progenitor-cell differentiation and causes dysplasia in epithelial tissues. Cell 121: 465-477.

208. Ricci-Vitiani L, Lombardi DG, Pilozzi E, Biffoni M, Todaro M, Peschle C, De MR (2007) Identification and expansion of human colon-cancer-initiating cells. Nature 445: 111-115.

209. Risch N (2001) The genetic epidemiology of cancer: interpreting family and twin studies and their implications for molecular genetic approaches. Cancer Epidemiol Biomarkers Prev 10: 733-741.

210. Pereira FA, Qiu Y, Zhou G, Tsai MJ, Tsai SY (1999) The orphan nuclear receptor COUP-TFII is required for angiogenesis and heart development. Genes Dev 13: 1037-1049.

211. Navab R, Gonzalez-Santos JM, Johnston MR, Liu J, Brodt P, Tsao MS, Hu J (2004) Expression of chicken ovalbumin upstream promoter-transcription factor II enhances invasiveness of human lung carcinoma cells. Cancer Res 64: 5097-5105.

212. Nakshatri H, Mendonca MS, Bhat-Nakshatri P, Patel NM, Goulet RJ, Jr., Cornetta K (2000) The orphan receptor COUP-TFII regulates G2/M progression of breast cancer cells by modulating the expression/activity

of p21(WAF1/CIP1), cyclin D1, and cdk2. Biochem Biophys Res Commun 270: 1144-1153.

213. Huang Q, Raya A, DeJesus P, Chao SH, Quon KC, Caldwell JS, Chanda SK, Izpisua-Belmonte JC, Schultz PG (2004) Identification of p53 regulators by genome-wide functional analysis. Proc Natl Acad Sci U S A 101: 3456-3461.

214. Jia L, Landan G, Pomerantz M, Jaschek R, Herman P, Reich D, Yan C, Khalid O, Kantoff P, Oh W, Manak JR, Berman BP, Henderson BE, Frenkel B, Haiman CA, Freedman M, Tanay A, Coetzee GA (2009) Functional enhancers at the gene-poor 8q24 cancer-linked locus. PLoS Genet 5: e1000597.

215. Jia L, Landan G, Pomerantz M, Jaschek R, Herman P, Reich D, Yan C, Khalid O, Kantoff P, Oh W, Manak JR, Berman BP, Henderson BE, Frenkel B, Haiman CA, Freedman M, Tanay A, Coetzee GA (2009) Functional enhancers at the gene-poor 8q24 cancer-linked locus. PLoS Genet 5: e1000597.

216. Culig Z, Bartsch G (2006) Androgen axis in prostate cancer. J Cell Biochem 99: 373-381.

217. Han G, Buchanan G, Ittmann M, Harris JM, Yu X, Demayo FJ, Tilley W, Greenberg NM (2005) Mutation of the androgen receptor causes oncogenic transformation of the prostate. Proc Natl Acad Sci U S A 102: 1151-1156.

218. Kaaks R, Stattin P (2010) Obesity, endogenous hormone metabolism, and prostate cancer risk: a conundrum of "highs" and "lows". Cancer Prev Res (Phila Pa) 3: 259-262.

219. Nicolucci A (2010) Epidemiological aspects of neoplasms in diabetes. Acta Diabetol .

220. Pierce BL, Ahsan H (2010) Genetic susceptibility to type 2 diabetes is associated with reduced prostate cancer risk. Hum Hered 69: 193-201.

221. Lee YF, Shyr CR, Thin TH, Lin WJ, Chang C (1999) Convergence of two repressors through heterodimer formation of androgen receptor and testicular orphan receptor-4: a unique signaling pathway in the steroid receptor superfamily. Proc Natl Acad Sci U S A 96: 14724-14729.

222. Shyr CR, Kang HY, Tsai MY, Liu NC, Ku PY, Huang KE, Chang C (2009) Roles of testicular orphan nuclear receptors 2 and 4 in early embryonic development and embryonic stem cells. Endocrinology 150: 2454-2462.

223. Lee YF, Young WJ, Lin WJ, Shyr CR, Chang C (1999) Differential regulation of direct repeat 3 vitamin D3 and direct repeat 4 thyroid hormone signaling pathways by the human TR4 orphan receptor. J Biol Chem 274: 16198-16205.

224. Hansen CM, Binderup L, Hamberg KJ, Carlberg C (2001) Vitamin D and cancer: effects of 1,25(OH)2D3 and its analogs on growth control and tumorigenesis. Front Biosci 6:D820-48.

225. Omdahl JL, Morris HA, May BK (2002) Hydroxylase enzymes of the vitamin D pathway: expression, function, and regulation. Annu Rev Nutr 22:139-66.

226. Schwartz GG, Hanchette CL (2006) UV, latitude, and spatial trends in prostate cancer mortality: all sunlight is not the same (United States). Cancer Causes Control 17: 1091-1101.

227. Lai JP, Sandhu DS, Yu C, Han T, Moser CD, Jackson KK, Guerrero RB, Aderca I, Isomoto H, Garrity-Park MM, Zou H, Shire AM, Nagorney DM, Sanderson SO, Adjei AA, Lee JS, Thorgeirsson SS, Roberts LR (2008) Sulfatase 2 up-regulates glypican 3, promotes fibroblast growth factor signaling, and decreases survival in hepatocellular carcinoma. Hepatology 47: 1211-1222.

228. Lemjabbar-Alaoui H, van ZA, Singer MS, Xue Q, Wang YQ, Tsay D, He B, Jablons DM, Rosen SD (2010) Sulf-2, a heparan sulfate endosulfatase, promotes human lung carcinogenesis. Oncogene 29: 635-646.

229. Morimoto-Tomita M, Uchimura K, Bistrup A, Lum DH, Egeblad M, Boudreau N, Werb Z, Rosen SD (2005) Sulf-2, a proangiogenic heparan sulfate endosulfatase, is upregulated in breast cancer. Neoplasia 7: 1001-1010.

230. Nawroth R, van ZA, Cervantes S, McManus M, Hebrok M, Rosen SD (2007) Extracellular sulfatases, elements of the Wnt signaling pathway, positively regulate growth and tumorigenicity of human pancreatic cancer cells. PLoS One 2: e392.

231. Dai Y, Yang Y, MacLeod V, Yue X, Rapraeger AC, Shriver Z, Venkataraman G, Sasisekharan R, Sanderson RD (2005) HSulf-1 and HSulf-2 are potent inhibitors of myeloma tumor growth in vivo. J Biol Chem 280: 40066-40073.

232. Conde-Knape K (2001) Heparan sulfate proteoglycans in experimental models of diabetes: a role for perlecan in diabetes complications. Diabetes Metab Res Rev 17: 412-421.

233. Datta MW, Hernandez AM, Schlicht MJ, Kahler AJ, DeGueme AM, Dhir R, Shah RB, Farach-Carson C, Barrett A, Datta S (2006) Perlecan, a candidate gene for the CAPB locus, regulates prostate cancer cell growth via the Sonic Hedgehog pathway. Mol Cancer 5: 9.

234. Kosir MA, Wang W, Zukowski KL, Tromp G, Barber J (1999) Degradation of basement membrane by prostate tumor heparanase. J Surg Res 81: 42-47.

235. Murphy T, Darby S, Mathers ME, Gnanapragasam VJ (2010) Evidence for distinct alterations in the FGF axis in prostate cancer progression to an aggressive clinical phenotype. J Pathol 220: 452-460.

236. Kwabi-Addo B, Ozen M, Ittmann M (2004) The role of fibroblast growth factors and their receptors in prostate cancer. Endocr Relat Cancer 11: 709-724.

237. Wang P, Keijer J, Bunschoten A, Bouwman F, Renes J, Mariman E (2006) Insulin modulates the secretion of proteins from mature 3T3-L1 adipocytes: a role for transcriptional regulation of processing. Diabetologia 49: 2453-2462.

238. Giovannucci E, Rimm EB, Stampfer MJ, Colditz GA, Willett WC (1998) Diabetes mellitus and risk of prostate cancer (United States). Cancer Causes Control 9: 3-9.

239. Kasper JS, Liu Y, Giovannucci E (2009) Diabetes mellitus and risk of prostate cancer in the health professionals follow-up study. Int J Cancer 124: 1398-1403.

240. Rodriguez C, Patel AV, Mondul AM, Jacobs EJ, Thun MJ, Calle EE (2005) Diabetes and risk of prostate cancer in a prospective cohort of US men. Am J Epidemiol 161: 147-152.

241. Zhu K, Lee IM, Sesso HD, Buring JE, Levine RS, Gaziano JM (2004) History of diabetes mellitus and risk of prostate cancer in physicians. Am J Epidemiol 159: 978-982.

242. Stevens VL, Ahn J, Sun J, Jacobs EJ, Moore SC, Patel AV, Berndt SI, Albanes D, Hayes RB (2010) HNF1B and JAZF1 genes, diabetes, and prostate cancer risk. Prostate 70: 601-607.

243. Zheng SL, Stevens VL, Wiklund F, Isaacs SD, Sun J, Smith S, Pruett K, Wiley KE, Kim ST, Zhu Y, Zhang Z, Hsu FC, Turner AR, Johansson JE, Liu W, Kim JW, Chang BL, Duggan D, Carpten J, Rodriguez C, Isaacs W, Gronberg H, Xu J (2009) Two independent prostate cancer risk-associated Loci at 11q13. Cancer Epidemiol Biomarkers Prev 18: 1815-1820.

244. Valdar W, Holmes CC, Mott R, Flint J (2009) Mapping in structured populations by resample model averaging. Genetics 182: 1263-1277.

245. Yeager M, Xiao N, Hayes RB, Bouffard P, Desany B, Burdett L, Orr N, Matthews C, Qi L, Crenshaw A, Markovic Z, Fredrikson KM, Jacobs KB, Amundadottir L, Jarvie TP, Hunter DJ, Hoover R, Thomas G, Harkins TT, Chanock SJ (2008) Comprehensive resequence analysis of a 136 kb region of human chromosome 8q24 associated with prostate and colon cancers. Hum Genet 124: 161-170.

246. Sharov AA, Masui S, Sharova LV, Piao Y, Aiba K, Matoba R, Xin L, Niwa H, Ko MS (2008) Identification of Pou5f1, Sox2, and Nanog downstream target genes with statistical confidence by applying a novel algorithm to time course microarray and genome-wide chromatin immunoprecipitation data. BMC Genomics 9:269.

247. Kang J, Shakya A, Tantin D (2009) Stem cells, stress, metabolism and cancer: a drama in two Octs. Trends Biochem Sci 34: 491-499.

248. van den Berg DL, Zhang W, Yates A, Engelen E, Takacs K, Bezstarosti K, Demmers J, Chambers I, Poot RA (2008) Estrogen-related receptor beta interacts with Oct4 to positively regulate Nanog gene expression. Mol Cell Biol 28: 5986-5995.

249. Bowles J, Teasdale RP, James K, Koopman P (2003) Dppa3 is a marker of pluripotency and has a human homologue that is expressed in germ cell tumours. Cytogenet Genome Res 101: 261-265.

250. Liu NC, Lin WJ, Kim E, Collins LL, Lin HY, Yu IC, Sparks JD, Chen LM, Lee YF, Chang C (2007) Loss of TR4 orphan nuclear receptor reduces phosphoenolpyruvate carboxykinase-mediated gluconeogenesis. Diabetes 56: 2901-2909.

251. Evans PM, Liu C (2008) Roles of Krupel-like factor 4 in normal homeostasis, cancer and stem cells. Acta Biochim Biophys Sin (Shanghai) 40: 554-564.

252. Aguilera O, Pena C, Garcia JM, Larriba MJ, Ordonez-Moran P, Navarro D, Barbachano A, Lopex de Silanes I, Ballestar E, Fraga MF, Esteller M, Gamallo C, Bonilla F, Gonzalez-Sancho JM, Munox A (2007) The Wnt antagonist DICKKOPF-1 gene is induced by 1alpha,25-dihydroxyvitamin D3 associated to the differentiation of human colon cancer cells. Carcinogenesis 28: 1877-1884.

253. Palmer HG, Anjos-Afonso F, Carmeliet G, Takeda H, Watt FM (2008) The vitamin D receptor is a Wnt effector that controls hair follicle differentiation and specifies tumor type in adult epidermis. PLoS One 23: e1483.