METHODS FOR GENETIC LINKAGE ANALYSIS IN THE PRESENCE OF HETEROGENEITY

by

Gerald Bryce Christensen

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Biomedical Informatics

The University of Utah

August 2009

THE UNIVERSITY OF UTAH GRADUATE SCHOOL

SUPERVISORY COMMITTEE APPROVAL

of a dissertation submitted by

Gerald Bryce Christensen

This dissertation has been read by each member of the following supervisory committee and by majority vote has been found to be satisfactory.

[redacted] [redacted]

8 [redacted] [redacted]
[redacted] S. [redacted]

5/8/09 [redacted]
[redacted]

[redacted] Scott P. [redacted]
Scott P. Narus

[redacted] [redacted] Thomas
Alun W. Thomas

THE UNIVERSITY OF UTAH GRADUATE SCHOOL

FINAL READING APPROVAL

To the Graduate Council of the University of Utah:

I have read the dissertation of Gerald Bryce Christensen in its final form and have found that (1) its format, citations, and bibliographic style are consistent and acceptable; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the supervisory committee and is ready for submission to The Graduate School.

[redacted] [redacted]

Lisa A. [redacted]annon-Albright
Chair: Supervisory Committee

Approved for the Major Department

[redacted]

[redacted] A. Mitchell
Chair/Dean

Approved for the Graduate Council

[redacted]

David S. Chapman
Dean of The Graduate School

ABSTRACT

Genetic heterogeneity is one of the most significant obstacles to identifying the genetic basis for many common human diseases. Heterogeneity is the term used for genetic systems in which numerous genes each make a small contribution to the overall heritability of a disease. Linkage analysis has been used successfully for several decades to map disease susceptibility genes, but it lacks power to identify susceptibility genes in heterogeneous systems. The purpose of this research is to improve current methods and develop new methods for linkage analysis in the presence of genetic heterogeneity. Competency is established in conventional and emerging methodology, new methods are developed, and the newly developed methods are tested in real study data. Prostate cancer (PCa), a prime example of the problems that heterogeneity creates for genetic epidemiologists, is used as a model system throughout the research.

Studying alternate PCa phenotype definitions or PCa subtypes may improve our knowledge of the disease. Chapter 2 describes a conventional linkage analysis for aggressive PCa subtypes, the results of which confirm two previously reported PCa aggressiveness loci. Chapter 3 presents proof of concept that phenotypes based on gene expression profiles from microarray data may be useful for identifying genes associated with risk of PCa development via linkage analysis. Chapters 4 and 5 describe the development and application of the innovative sumLINK statistic, which identifies genetic regions of extreme consistency across pedigrees without regard to negative evidence from unlinked or uninformative pedigrees. Significance of the sumLINK statistic and the complimentary sumLOD statistic is determined empirically by an innovative permutation procedure that randomizes linkage information across pedigrees. Simulation testing shows that this method is reliable and powerful for finding genes in heterogeneous systems. The utility of the sumLINK method is demonstrated with exciting results using data from the International Consortium for Prostate Cancer Genetics for aggressive and general prostate cancer. The sumLINK procedure fills an important informatics role by facilitating secure interinstitutional data sharing and collaborative research. The sumLINK method is a powerful tool for combating the obstacles presented by heterogeneity, and will improve our knowledge of the genetic epidemiology of many common, complex diseases.

TABLE OF CONTENTS

Chapter

Appendices

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

The assistance, guidance and support of supervisory committee chairman Dr. Lisa A. Cannon-Albright and all other supervisory committee members is gratefully acknowledged. Recognition is also given to the faculty, staff and students from the Department of Biomedical Informatics for helping along the way. Funding for all parts of this work was provided by National Library of Medicine informatics training grant T15 LM07124. Additional grants supporting various parts of this work are listed within each chapter. Special recognition is given to the International Consortium for Prostate Cancer Genetics, headed by Dr. William B. Isaacs, for providing invaluable data to enhance this research.

CHAPTER I

INTRODUCTION

Background

Genetic Epidemiology

Genetic epidemiology is the study of the genetic contribution to biological phenomena, with specific emphasis on determining the hereditary factors of human disease. The basic purpose of genetic epidemiology is to define the relationship between genotype and phenotype. Genetic epidemiology is an interdisciplinary science that synthesizes knowledge from the fields of genetics, statistics, and bioinformatics. The traditional tools of genetic epidemiology include case-control genetic association techniques and a variety of pedigree-based analytical techniques, such as linkage analysis and transmission tests.

Genetic Epidemiology Research in Utah

Genetic epidemiology research has a rich history in Utah, due largely to the unique resources available in the Utah Population Database (UPDB). The UPDB was established in the 1970s and now contains records for over 12 million individuals. The UPDB began with computerized genealogical records from the descendants of Utah's founding pioneers. Those genealogies have since been extended and linked with extensive phenotype data from such sources as the Utah Cancer Registry, Utah death certificates and vital records, and hospital discharge data [1]. This combination of genealogical and phenotypic data makes it possible to calculate population-based risk estimates for diseases among the relatives of probands and to easily identify families with significant excesses of those diseases. Familial relative risks can be determined for phenotypes as diverse as cancer [2], intracranial aneurysm [3], and kidney disease [4]. UPDB resources have been instrumental in identifying many important disease genes [5-9].

Linkage Analysis

Genetic linkage analysis has been successfully used for many years as a tool for mapping disease susceptibility loci [10]. Linkage analysis is the process of identifying chromosomal segments that are co-inherited with disease status in pedigrees with an abundance of the disease. This technique is well suited for finding rare, highly penetrant genetic variants. The LOD score has been the principle metric used in linkage analysis for over 50 years [11,12]. LOD scores work well when all or most of the pedigrees studied are linked to a single genetic locus, but it lacks power to detect genes in heterogeneous systems where the trait is controlled by numerous genes and only a small proportion of the collected pedigrees are linked to any given risk locus [10]. Variants that account for less than 20% of the total heritability of a trait can rarely be detected with LOD analysis [13]. The low power of the traditional LOD score method for detecting linkage in heterogeneous systems is a major weakness of the approach. Many human health-related phenotypes are believed to be controlled by multiple genes, each accounting for such a small proportion of the heritability of the trait that it is unlikely to be detected by LOD analysis.

Heterogeneity Methods

Several analytical methods have been developed for dealing with the problem of heterogeneity. The most widely used metric is the Heterogeneity-LOD, or HLOD statistic [14]. HLOD analysis allows for a portion of the pedigrees to be unlinked at any given locus [15,16], but it tends to identify only large-effect loci. Other methods have been proposed and tested for incorporating interaction effects into linkage or for simultaneous linkage analysis of multiple loci [17,18], but these methods are in developmental stages and have not yet been broadly adopted. Another approach to the problem is the sumLOD statistic, which strives to identify loci in the presence of heterogeneity by focusing on the pedigrees with positive linkage information at a locus and ignoring negative information from other pedigrees. It has been used in the past as a summary measure [19-21], but has not been used as a test statistic because the distribution is unknown, making significance determination difficult. The sumLOD statistic is not widely used currently, but a procedure for testing its significance, presented in Chapter 4, may increase its utilization.

Prostate Cancer

Prostate cancer (PC) is the most commonly diagnosed cancer among American men. 186,320 new PC cases were expected in the United States in 2008, accounting for 25% of all new male cancer cases [22]. PC is also the second leading cause of cancer-related mortality in American men. Studies have repeatedly shown that PC has a strong hereditary component [23,24]. This observation holds true in Utah, where analysis of the UPDB indicates significant evidence of familiality for PC extending well beyond the limits of nuclear familes [Appendix A]. The relative risk of PC to first degree relatives of PC cases in Utah is 1.91 (95% CI: 1.85—1.97), and the relative risk to second degree relatives is 1.28 (1.24—1.32) [25]. The health burden of

PC makes it a priority for epidemiology research, and the familiality evidence makes it an excellent candidate for genetic epidemiology research as well.

Heterogeneity in Prostate Cancer

Prostate cancer (PC) is an excellent example of the negative impact of heterogeneity on genetic linkage. Results of the first genome-wide linkage analysis for PC were published in 1996 [26]. That study found highly significant linkage evidence (LOD = 5.43) at chromosome 1q23-24, a region that has come to be known as the HPC1 locus. Identification of HPC1 was an encouraging start to the pursuit of PC susceptibility genes; however, early attempts to replicate the linkage finding met little success [24]. Genes such as RNASEL [27] have since been proposed as the HPC1 gene, but there is still no consensus about the underlying source of the HPC1 linkage.

Several more linkage analyses were published soon after the HPC1 result that implicated additional PC susceptibility loci. Notable PC loci reported between 1998 and 2001 include PCAP [28], HPCX [29], CAPB [30], HPC20 [31], and HPC2 [32]. Linkage to the HPC2 locus on chromosome 17p was announced by Utah researchers in 2000 [33]. Positional cloning and mutation screening identified a gene, HPC2/ELAC2, mutations of which segregated with PC. This finding was greeted with "enormous excitement" by the research community [24]. Early confirmation studies reported that two common missense variants in the gene were strongly associated with PC risk (OR=2.37) and accounted for 5% of PC in the population of inference [34]. A meta-analysis of six studies showed a similar level of association [35]. Significant association was also reported in a study of African American PC cases [36]. However, other researchers failed to replicate either the linkage [37] or the association result [38,39]. Evidence for the HPC2 locus remains inconsistent. Confirmation attempts for other linkage regions have encountered difficulties similar to what has been described for the HPC1 and HPC2 loci. To date, over 30 genome-wide linkage analyses for PC or selected PC subtypes have been published [18,26,28-32,40-66], with little consensus in the findings [23]. Putative PC susceptibility loci have been reported on almost every human chromosome, and multiple susceptibility loci reported on several chromosomes [23,67,68]. Major findings are summarized in Table 1.1.

Addressing Prostate Cancer Heterogeneity

A 1998 review of PC genetics determined that "a large proportion of familial (prostate cancer) may not be due to segregation of a few major gene mutations, but rather to familial sharing of alleles at many loci, each contributing to a small increase in cancer risk" [69]. This observation was prescient of the future of PC genetics. Since 1999, at least 18 more reviews have been written about the ongoing pursuit of PC susceptibility genes [23,24,67,68,70-83]. These reviews have consistently identified genetic heterogeneity as the primary factor that complicates the search for PC genes. It has been proposed that the apparent heterogeneity of PC results partially from variability in the phenotype, and that different subtypes of the disease may each have a more uniform genetic etiology. The phenotypic variability is often attributed to the increased screening for prostate-specific antigen (PSA) which began in the 1980s and resulted in a dramatic increase in PC incidence rates [22]. Another frequent

Table 1.1. Selected prostate cancer (PC) linkage peaks

Locus	Maximum LOD, HLOD, NPL or p-value	Proposed Gene	Study
General PC			
1p36	3.22	CAPB	Gibbs, 1999 [30]
1q24-25	5.43	HPC1	Smith, 1996 [26]
1q42.2-q43	3.10	PCAP	Berthon, 1998 [28]
3p14.2	3.83	FHIT	Larson, 2005 [84]
3q26.31	2.48	N/A	Camp, 2005 [66]
5q21.1	2.06	N/A	Camp, 2005 [66]
7q11-21	3.01	N/A	Friedrichsen, 2004 [60]
8p22-23	1.84	MSR1	Xu, 2001 [85]
16q23.2	3.15	N/A	Suarez, 2000 [61]
17p11	4.53	HPC2/ELAC2	Tavtigian, 2001 [32]
17q22	3.16	N/A	Gillanders, 2004 [62]
19p13.3	2.87	N/A	Hsieh, 2001 [64]
20q13	3.02	HPC20	Berry, 2000 [31]
22q12	3.57	N/A	Xu, 2005 [63]
Xq27-28	3.85	HPCX	Xu, 1998 [29]
Aggressive PC			
1q24-25	3.25	HPC1	Goddard, 2001 [65]
1q42.2-q43	2.84	PCAP	Goddard, 2001 [65]
4q	2.80	N/A	Goddard, 2001 [65]
5q31-33	P=0.0053	N/A	Witte, 2000 [59]
6p22.3	3.00	N/A	Schaid, 2006 [51]
6q23.3	P=0.0009	N/A	Slager, 2006 [55]
7q21.11	4.09	N/A	Schaid, 2006 [51]
7q31-33	3.02	N/A	Paiss, 2003 [86]
7q32.2	P=0.0076	N/A	Witte, 2000 [59]
11q14.1-3	3.31	N/A	Schaid, 2006 [51]
19q12	P=0.0088	N/A	Witte, 2000 [59]
19q13	P<0.0001	N/A	Slager, 2003 [54]
20p11-q11	2.65	HPC20	Schaid, 2006 [51]
22q11-13	2.18	N/A	Stanford, 2006 [56]
22q13	2.06	N/A	Chang, 2005 [41]
Xq12-13	3.06	AR	Goddard, 2001 [65]
Xq27-28	2.54	HPCX	Chang, 2005 [41]

explanation for the lack of reproducible results is the high expected number of phenocopies, or pedigree members who have the disease but do not share the same inherited factors as other pedigree members, the presence of which suppresses linkage evidence. Eeles wrote the following in a 1999 review of PC genetics:

> The study of familial prostate cancer is complicated by the fact that there may be many sporadic cases in families, as prostate cancer is so common. In addition, (PSA) screen detected family history may behave differently from symptomatic disease. [76]

These observations regarding phenotypic variability were echoed by Ostrander and Stanford in 2000:

> It seems, therefore, that mapping and cloning of prostate cancer genes will be complicated. . . . First, there are a large number of men with sporadic disease in the population. . . . Finally, there is enormous variation in the phenotype of disease at diagnosis as well as disease progression within single families. The introduction of prostate-specific antigen (PSA) testing in the mid to late 1980s has probably contributed to that variability. [24]

Stratification of PC families into more homogeneous subgroups for linkage analysis has been suggested as a mechanism to address the problem of heterogeneity due to phenotypic variability [87]. Linkage pedigrees are commonly classified according to variables such as mean age of cases at PC diagnosis, evidence of male-to-male transmission, and the number of cases in the pedigree. This approach has met with some success, as many of the most significant and reproducible linkages reported for PC have come from subset analyses. The weakness of this approach comes in the inherent loss of statistical power resulting from multiple testing [88].

Several research groups have published linkage analyses focusing on the "aggressive" or "clinically significant" subset of prostate cancer cases

[41,42,48,51,52,55,59]. This approach is substantially different from simple subsetting of pedigrees, as it requires a thorough redefinition of the phenotype. The definitions of aggressiveness vary across studies, but generally include some combination of advanced tumor stage or grade, PSA level, early age at diagnosis, and mortality resulting from PC. An example of linkage analsysis for aggressive PC is presented in Chapter 2. A more recent innovation is the idea of linkage analysis using PC-related biomarkers to define the phenotype, with the hypothesis that a gene may be identified that is associated with the biomarker, and by extension, with the disease. An example of this is the recently reported linkage analysis for the TMPRSS2-ERG fusion, a genetic anomaly commonly observed in PC tumors [45]. RNA expression levels are another type of biomarker that has been suggested as a suitable phenotype for linkage analysis. This concept is discussed more in Chapter 3.

Another method to address genetic heterogeneity is to increase statistical power for finding genes by collecting larger numbers of pedigrees for analysis. The International Consortium for Prostate Cancer Genetics (ICPCG) was formed with the goal of improving PC research through the use of both prospective and retrospective collaboration [89], following Morton's directive to combine linkage evidence across studies [13]. The pooled pedigree resource of the ICPCG consists of over 1200 high-risk PC pedigrees from diverse areas of the world. Analysis of this resource has resulted in very promising linkage regions such as that on chromosome 22q12 [63,90], and there is great potential for future discoveries within the extensive ICPCG data. Chapter 5 contains an analysis of this ICPCG resource using newly developed statistical methods.

Genome-wide association studies (GWAS) have been proposed as an alternative to traditional pedigree-based linkage analysis. This approach, using case-control testing of several hundred thousand single nucleotide polymorphisms (SNPs), has shown some promise in clarifying the genetic basis of PC [91]. Recently published GWAS results for PC have resulted in several significantly associated SNP loci [72]. SNP associations on chromosomes 8, 10, and 17 have been reproduced in multiple data resources, including at the University of Utah [92], but these variants account for only a very small portion of all PC, and their functional significance has yet to be defined. GWAS methodology is still in developmental stages and has yet to fulfill the optimistic expectations for the procedure [93], but it holds promise for identifying genes involved in PC and other complex phenotypes.

Description of Research

The purpose of the research presented here is to improve current methodology and to develop new methods for linkage analysis in the presence of heterogeneity. Newton Morton, one of the fathers of modern linkage analysis, identifies the heterogeneity issue as one of the most significant unsolved problems in genetic epidemiology [13]. He describes collaborative research and methodological development as two key factors to make mapping of oligogenes (genes with the greatest effect in heterogeneous systems) possible. Morton writes:

> "The central problem of oligogenic mapping is to combine evidence from linkage and allelic association over many studies, each with inadequate power and differing to some extent from the others in phenotype definition, ascertainment, markers, and population. Otherwise stated, the central problem is to develop methods that bring to

> oligogenes the reliability that lods have given to linkage mapping for major loci [13]."

Methodological development, particularly the development of methods that encourage and facilitate collaborative research, is a necessary step to overcome the negative impact of heterogeneity. The development of a reliable, robust analysis method will be an important step toward understanding the genetic etiology of a plethora of complex human health phenotypes. The creation of such a method is the centerpiece of this research. Prostate cancer, a prime example of the complications presented by heterogeneity, is used as a model system throughout this dissertation.

The research presented here has three primary objectives:

1. Apply conventional genetic epidemiology methods to alternative phenotype definitions, such as clinically aggressive disease and predicted disease risk based on biomarkers, which may clarify the broader genetic basis of the disease.
2. Develop new statistical methods for linkage analysis in the presence of heterogeneity, designed to facilitate multicenter collaborative research.
3. Demonstrate the power and utility of the new methods using data from the International Consortium for Prostate Cancer Genetics.

Chapters 2 through 5 present the results of four research projects designed to fulfill the objectives above. The contents of these chapters, as well as four appendices containing supporting information, are described below.

Chapter 2

Chapter 2 is a description of a genome-wide linkage analysis for aggressive PC in Utah high-risk pedigrees based on the definition of aggressiveness set forth by the ICPCG [51]. The analysis is designed as a replication study, undertaken with the intention of confirming risk loci that were previously identified by the ICPCG and others for aggressive PC. Several regions of interest are identified, two of which support loci previously linked to PC aggressiveness. Chapter 2 is an example of the type of linkage analysis that is commonly published today. The results include a confirmation of previously published linkage evidence and contributes novel findings with regard to PC aggressiveness. It includes subset analysis intended to control for the effects of heterogeneity by analyzing subgroups with homogeneous phenotypic characteristics.

Chapter 3

Chapter 3 contains the results of an analysis prepared for the fifteenth Genetic Analysis Workshop (GAW, GAW15) [94]. GAW is a series of biannual conferences where genetic epidemiologists convene to discuss current and emerging topics in the field. One or more data sets representing contemporary research trends are made available prior to each GAW meeting, and participants are encouraged to use these data to test innovative analysis methods. One of the data sets provided for GAW15 participants included RNA expression levels for 3554 genes together with genome-wide SNP genotype data for 194 individuals from 14 CEPH (Centre d'Etude du Polymorphisme Humain) pedigrees [95]. The study in Chapter 3 uses these data to test the hypothesis that PC-related biomarkers may be used as a phenotype for linkage analysis with the intention of identifying the location of genes that cause PC. Phenotypes were assigned to all individuals based on RNA expression profiles consistent with PC, and conventional linkage analysis was then carried out. The results provide proof of concept that biomarkers such as RNA expression levels are valid phenotypes for linkage analysis.

Chapter 4

Morton wrote that overcoming the problem of heterogeneity in linkage will require collaborative efforts as well as the development of new analytical methods. Morton also criticized linkage methods that do not account for multiple testing [13]. All of these issues are addressed directly in Chapter 4. As discussed previously, the sumLOD statistic may have the ability to identify linkage in the presence of heterogeneity, but has not been used as a test statistic due to a lack of understanding of its distribution. Chapter 4 describes a novel genomic randomization method to test the empirical significance of the sumLOD statistic, as well as a similar metric, sumLINK. Both of these statistics use LOD scores from individual pedigrees to identify chromosomal regions of extreme consistency across multiple pedigrees with evidence of linkage, without regard to negative evidence from other pedigrees. Simulation results given in the chapter demonstrate that the sumLINK and sumLOD statistics are more powerful than conventional HLOD statistics to identify trait genes in polygenic systems.

This method facilitates collaborative research because it is a postprocessing procedure that uses only meta data (pedigree LOD scores), and therefore allows for pooled analysis without sharing protected, identifiable information. An important advantage of the sumLINK procedure is that loci identified with the method are excellent candidates for statistical recombinant mapping, as multiple pedigrees are linked to these loci. Recombinant mapping can delimit the precise chromosomal regions where trait genes are most likely to be found. Multiple testing effects for the sumLINK and sumLOD procedure are quantified by the use of false discovery rate (FDR) techniques. An illustrative example of sumLINK and sumLOD analysis is presented using data from 190 aggressive PC pedigrees provided by the ICPCG. Appendix D contains R [96] program code for running sumLINK and sumLOD analysis.

Chapter 5

Chapter 5 is an in-depth application of the sumLINK and sumLOD methods that are described in Chapter 4. The chapter describes an analysis of 1230 high-risk PC pedigrees from Europe and North America provided by the ICPCG. The previous report of significant linkage at chromosome 22q12 is confirmed, as well as several other previously reported linkage results. Linkage signals are localized to narrow chromosomal regions with statistical recombinant mapping. The regions identified are more precise than the regions identified by traditional 1-LOD support intervals. The application of these powerful statistics to such an extensive data resource provides a clear understanding of the genomic regions with the greatest evidence of consistent linkage information across multiple pedigrees. The results of Chapter 5 provide encouraging evidence that the sumLINK and sumLOD statistics will be beneficial for identifying the genes underlying PC and other complex phenotypes.

References

1. Cannon Albright LA: Utah family-based analysis: past, present and future. *Hum Hered* 2008, 65:209-220.

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

3. Cannon Albright LA, Camp NJ, Farnham JM, MacDonald J, Abtin K, Rowe KG: A genealogical assessment of heritable predisposition to aneurysms. *J Neurosurg* 2003, 99:637-643.

4. Goldfarb-Rumyantzev AS, Cheung AK, Habib AN, Wang BJ, Lin SJ, Baird BC, Naiman N, Cannon-Albright L: A population-based assessment of the familial component of chronic kidney disease mortality. *Am J Nephrol* 2006, 26:142-148.

5. Atkin CL, Hasstedt SJ, Menlove L, Cannon L, Kirschner N, Schwartz C, Nguyen K, Skolnick M: Mapping of Alport syndrome to the long arm of the X chromosome. *Am J Hum Genet* 1988, 42:249-255.

6. Barker D, Wright E, Nguyen K, Cannon L, Fain P, Goldgar D, Bishop DT, Carey J, Baty B, Kivlin J, et al.: Gene for von Recklinghausen neurofibromatosis is in the pericentromeric region of chromosome 17. *Science* 1987, 236:1100-1102.

7. Cannon-Albright LA, Goldgar DE, Meyer LJ, Lewis CM, Anderson DE, Fountain JW, Hegi ME, Wiseman RW, Petty EM, Bale AE, et al.: Assignment of a locus for familial melanoma, MLM, to chromosome 9p13-p22. *Science* 1992, 258:1148-1152.

8. Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, Liu Q, Cochran C, Bennett LM, Ding W, et al.: A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* 1994, 266:66-71.

9. Tavtigian SV, Simard J, Rommens J, Couch F, Shattuck-Eidens D, Neuhausen S, Merajver S, Thorlacius S, Offit K, Stoppa-Lyonnet D, et al: The complete BRCA2 gene and mutations in chromosome 13q-linked kindreds. *Nat Genet* 1996, 12:333-337.

10. Terwilliger JD, Goring HH: Gene mapping in the 20th and 21st centuries: statistical methods, data analysis, and experimental design. *Hum Biol* 2000, 72:63-132.

11. Morton NE: Logarithm of odds (lods) for linkage in complex inheritance. *Proc Natl Acad Sci USA* 1995, 93:3471-3476.

12. Morton NE: Sequential Tests for the Detection of Linkage. *Am J Hum Genet* 1955, 7:277-318.

13. Morton NE: Unsolved problems in genetic epidemiology. *Hum Hered* 2000, 50:5-13.

14. Gharani N, Waterworth DM, Batty S, White D, Gilling-Smith C, Conway GS, McCarthy M, Franks S, Williamson R: Association of the steroid synthesis gene CYP11a with polycystic ovary syndrome and hyperandrogenism. *Hum Mol Genet* 1997, 6:397-402.

15. Ott J: Linkage analysis and family classification under heterogeneity. *Ann Hum Genet* 1983, 47:311-320.

16. Hodge SE, Anderson CE, Neiswanger K, Sparkes RS, Rimoin DL: The search for heterogeneity in insulin-dependent diabetes mellitus (IDDM): linkage studies, two-locus models, and genetic heterogeneity. *Am J Hum Genet* 1983, 35:1139-1155.

17. Schaid DJ, McDonnell SK, Carlson EE, Thibodeau SN, Ostrander EA, Stanford JL: Affected relative pairs and simultaneous search for two-locus linkage in the presence of epistasis. *Genet Epidemiol* 2007, 31:431-449.

18. Chang BL, Lange EM, Dimitrov L, Valis CJ, Gillanders EM, Lange LA, Wiley KE, Isaacs SD, Wiklund F, Baffoe-Bonnie A, et al: Two-locus genome-wide linkage scan for prostate cancer susceptibility genes with an interaction effect. *Hum Genet* 2006, 118:716-724.

19. Camp NJ, Hopkins PN, Hasstedt SJ, Coon H, Malhotra A, Cawthon RM, Hunt SC: Genome-Wide Multipoint Parametric Linkage Analysis of Pulse Pressure in Large, Extended Utah Pedigrees. *Hypertension* 2003, 43:322-328.

20. Horne BD, Malhotra A, Camp NJ: Comparison of linkage analasys methods for genome-wide scanning of extended pedigrees, with application to the TG/HDL-C ratio in the Framingham Heart Study. *BMC Genetics* 2003, 4:S93.

21. Orr A, Dubé M, Marcadier J, Jiang H, Federico A, George S, Seamone C, Andrews D, Dubord P, Holland S, et al: Mutations in the UBIAD1 Gene, encoding a potential prenyltransferase, are causal for Schnyder Crystalline Corneal Dystrophy. *PLoS ONE* 2007, 2:e685.

22. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Murray T, Thun MJ: Cancer statistics, 2008. *CA Cancer J Clin* 2008, 58:71-96.

23. Schaid D: The Complex Genetic Epidemiology of Prostate Cancer. *Human Molecular Genetics* 2004, 13:R103-121.

24. Ostrander EA, Stanford JL: Genetics of prostate cancer: too many loci, too few genes. *Am J Hum Genet* 2000, 67:1367-1375.

25. Cannon-Albright LA SA, Camp NJ, Farnham JS, Thomas A: Population-based risk assessment for other cancers in relatives of hereditary prostate cancer cases. *Prostate* 2005, 64:347-355.

26. Smith JR, Freije D, Carpten JD, Gronberg H, Xu J, Isaacs SD, Brownstein MJ, Bova GS, Guo H, Bujnovszky P, et al: Major susceptibility locus for prostate cancer on chromosome 1 suggested by a genome-wide search. *Science* 1996, 274:1371-1374.

27. Carpten J, Nupponen N, Isaacs S, Sood R, Robbins C, Xu J, Faruque M, Moses T, Ewing C, Gillanders E, et al: Germline mutations in the ribonuclease L gene in families showing linkage with HPC1. *Nat Genet* 2002, 30:181-184.

28. Berthon P, Valeri A, Cohen-Akenine A, Drelon E, Paiss T, Wohr G, Latil A, Millasseau P, Mellah I, Cohen N, et al: Predisposing gene for early-onset prostate cancer, localized on chromosome 1q42.2-43. *Am J Hum Genet* 1998, 62:1416-1424.

29. Xu J, Meyers D, Freije D, Isaacs S, Wiley K, Nusskern D, Ewing C, Wilkens E, Bujnovszky P, Bova GS, et al: Evidence for a prostate cancer susceptibility locus on the X chromosome. *Nat Genet* 1998, 20:175-179.

30. Gibbs M, Stanford JL, McIndoe RA, Jarvik GP, Kolb S, Goode EL, Chakrabarti L, Schuster EF, Buckley VA, Miller EL, et al: Evidence for a rare prostate cancer-susceptibility locus at chromosome 1p36. *Am J Hum Genet* 1999, 64:776-787.

31. Berry R, Schroeder JJ, French AJ, McDonnell SK, Peterson BJ, Cunningham JM, Thibodeau SN, Schaid DJ: Evidence for a prostate cancer-susceptibility locus on chromosome 20. *Am J Hum Genet* 2000, 67:82-91.

32. Tavtigian SV, Simard J, Teng DH, Abtin V, Baumgard M, Beck A, Camp NJ, Carillo AR, Chen Y, Dayananth P, et al: A candidate prostate cancer susceptibility gene at chromosome 17p. *Nat Genet* 2001, 27:172-180.

33. Tavtigian SV, Simard J, Labrie F, Skolnick MH, Neuhausen SL, Rommens J, Cannon-Albright LA: A strong candidate prostate cancer predisposition gene at chromosome 17p. *Am J Hum Genet* 2000, Suppl 67:A7.

34. Rebbeck TR, Walker AH, Zeigler-Johnson C, Weisburg S, Martin AM, Nathanson KL, Wein AJ, Malkowicz SB: Association of HPC2/ELAC2 genotypes and prostate cancer. *Am J Hum Genet* 2000, 67:1014-1019.

35. Camp NJ, Tavtigian SV: Meta-analysis of associations of the Ser217Leu and Ala541Thr variants in ELAC2 (HPC2) and prostate cancer. *Am J Hum Genet* 2002, 71:1475-1478.

36. Robbins CM, Hernandez W, Ahaghotu C, Bennett J, Hoke G, Mason T, Pettaway CA, Vijayakumar S, Weinrich S, Furbert-Harris P, et al: Association of HPC2/ELAC2 and RNASEL non-synonymous variants with prostate cancer risk in African American familial and sporadic cases. *Prostate* 2008, 68:1790-1797.

37. Xu J, Zheng SL, Carpten JD, Nupponen NN, Robbins CM, Mestre J, Moses TY, Faith DA, Kelly BD, Isaacs SD, et al: Evaluation of linkage and association of HPC2/ELAC2 in patients with familial or sporadic prostate cancer. *Am J Hum Genet* 2001, 68:901-911.

38. Rokman A, Ikonen T, Mononen N, Autio V, Matikainen MP, Koivisto PA, Tammela TL, Kallioniemi OP, Schleutker J: ELAC2/HPC2 involvement in hereditary and sporadic prostate cancer. *Cancer Res* 2001, 61:6038-6041.

39. Severi G, Giles GG, Southey MC, Tesoriero A, Tilley W, Neufing P, Morris H, English DR, McCredie MR, Boyle P, Hopper JL: ELAC2/HPC2 polymorphisms, prostate-specific antigen levels, and prostate cancer. *J Natl Cancer Inst* 2003, 95:818-824.

40. Baffoe-Bonnie AB, Kittles RA, Gillanders E, Ou L, George A, Robbins C, Ahaghotu C, Bennett J, Boykin W, Hoke G, et al: Genome-wide linkage of 77 families from the African American Hereditary Prostate Cancer study (AAHPC). *Prostate* 2007, 67:22-31.

41. Chang BL IS, Wiley KE, Gillanders EM, Zheng SL, Meyers DA, Walsh PC, Trent JM, Xu J, Isaacs WB: Genome-wide screen for prostate cancer susceptibility genes in men with clinically significant disease. *Prostate* 2005, 64:356-361.

42. Christensen GB, Camp NJ, Farnham JM, Cannon-Albright LA: Genome-wide linkage analysis for aggressive prostate cancer in Utah high-risk pedigrees. *Prostate* 2007, 67:605-613.

43. Cunningham JM, McDonnell SK, Marks A, Hebbring S, Anderson SA, Peterson BJ, Slager S, French A, Blute ML, Schaid DJ, Thibodeau SN: Genome linkage screen for prostate cancer susceptibility loci: results from the Mayo Clinic Familial Prostate Cancer Study. *Prostate* 2003, 57:335-346.

44. Edwards S, Meitz J, Eles R, Evans C, Easton D, Hopper J, Giles G, Foulkes WD, Narod S, Simard J, et al: Results of a genome-wide linkage analysis in prostate cancer families ascertained through the ACTANE consortium. *Prostate* 2003, 57:270-279.

45. Hofer MD, Kuefer R, Maier C, Herkommer K, Perner S, Demichelis F, Paiss T, Vogel W, Rubin MA, Hoegel J: Genome-wide linkage analysis of TMPRSS2-ERG fusion in familial prostate cancer. *Cancer Res* 2009, 69:640-646.

46. Lange EM, Beebe-Dimmer JL, Ray AM, Zuhlke KA, Ellis J, Wang Y, Walters S, Cooney KA: Genome-wide linkage scan for prostate cancer susceptibility from the University of Michigan Prostate Cancer Genetics Project: suggestive evidence for linkage at 16q23. *Prostate* 2009, 69:385-391.

47. Lange EM, Gillanders EM, Davis CC, Brown WM, Campbell JK, Jones M, Gildea D, Riedesel E, Albertus J, Freas-Lutz D, et al: Genome-wide scan for prostate cancer susceptibility genes using families from the University of Michigan prostate cancer genetics project finds evidence for linkage on chromosome 17 near BRCA1. *Prostate* 2003, 57:326-334.

48. Lange EM, Ho LA, Beebe-Dimmer JL, Wang Y, Gillanders EM, Trent JM, Lange LA, Wood DP, Cooney KA: Genome-wide linkage scan for prostate cancer susceptibility genes in men with aggressive disease: significant evidence for linkage at chromosome 15q12. *Hum Genet* 2006, 119:400-407.

49. Neville PJ, Conti DV, Krumroy LM, Catalona WJ, Suarez BK, Witte JS, Casey G: Prostate cancer aggressiveness locus on chromosome segment 19q12-q13.1 identified by linkage and allelic imbalance studies. *Genes Chromosomes Cancer* 2003, 36:332-339.

50. Schaid DJ, Guenther JC, Christensen GB, Hebbring S, Rosenow C, Hilker CA, McDonnell SK, Cunningham JM, Slager SL, Blute ML, Thibodeau SN: Comparison of microsatellites versus single-nucleotide polymorphisms in a genome linkage screen for prostate cancer-susceptibility Loci. *Am J Hum Genet* 2004, 75:948-965.

51. Schaid DJ, McDonnell SK, Zarfas KE, Cunningham JM, Hebbring S, Thibodeau SN, Eeles RA, Easton DF, Foulkes WD, Simard J, et al: Pooled genome linkage scan of aggressive prostate cancer: results from the International Consortium for Prostate Cancer Genetics. *Hum Genet* 2006, 120:471-485.

52. Schaid DJ, Stanford JL, McDonnell SK, Suuriniemi M, McIntosh L, Karyadi DM, Carlson EE, Deutsch K, Janer M, Hood L, Ostrander EA: Genome-wide linkage scan of prostate cancer Gleason score and confirmation of chromosome 19q. *Hum Genet* 2007, 121:729-735.

53. Schleutker J, Baffoe-Bonnie AB, Gillanders E, Kainu T, Jones MP, Freas-Lutz D, Markey C, Gildea D, Riedesel E, Albertus J, et al: Genome-wide scan for linkage in finnish hereditary prostate cancer (HPC) families identifies novel susceptibility loci at 11q14 and 3p25-26. *Prostate* 2003, 57:280-289.

54. Slager SL, Schaid DJ, Cunningham JM, McDonnell SK, Marks AF, Peterson BJ, Hebbring SJ, Anderson S, French AJ, Thibodeau SN: Confirmation of linkage of prostate cancer aggressiveness with chromosome 19q. *Am J Hum Genet* 2003, 72:759-762.

55. Slager SL, Zarfas KE, Brown WM, Lange EM, McDonnell SK, Wojno KJ, Cooney KA: Genome-wide linkage scan for prostate cancer aggressiveness loci using families from the University of Michigan Prostate Cancer Genetics Project. *Prostate* 2006, 66:173-179.

56. Stanford J, McDonnell S, Friedrichsen D, Carlson E, Kolb S, Deutsch K, Janer M, Hood L, Ostrander E, Schaid D: Prostate cancer and genetic susceptibility: a genome scan incorporating disease aggressiveness. *Prostate* 2006, 66:317-325.

57. Stanford JL, Fitzgerald LM, McDonnell SK, Carlson EE, McIntosh LM, Deutsch K, Hood L, Ostrander EA, Schaid DJ: Dense Genome-Wide SNP Linkage Scan in 301 Hereditary Prostate Cancer Families Identifies Multiple Regions with Suggestive Evidence for Linkage. *Hum Mol Genet* 2009.

58. Wiklund F, Gillanders EM, Albertus JA, Bergh A, Damber JE, Emanuelsson M, Freas-Lutz DL, Gildea DE, Goransson I, Jones MS, et al: Genome-wide

scan of Swedish families with hereditary prostate cancer: suggestive evidence of linkage at 5q11.2 and 19p13.3. *Prostate* 2003, 57:290-297.

59. Witte JS, Goddard KA, Conti DV, Elston RC, Lin J, Suarez BK, Broman KW, Burmester JK, Weber JL, Catalona WJ: Genomewide scan for prostate cancer-aggressiveness loci. *Am J Hum Genet* 2000, 67:92-99.

60. Friedrichsen DM, Stanford JL, Isaacs SD, Janer M, Chang BL, Deutsch K, Gillanders E, Kolb S, Wiley KE, Badzioch MD, et al: Identification of a prostate cancer susceptibility locus on chromosome 7q11-21 in Jewish families. *Proc Natl Acad Sci USA* 2004, 101:1939-1944.

61. Suarez BK, Lin J, Burmester JK, Broman KW, Weber JL, Banerjee TK, Goddard KA, Witte JS, Elston RC, Catalona WJ: A genome screen of multiplex sibships with prostate cancer. *Am J Hum Genet* 2000, 66:933-944.

62. Gillanders EM, Xu J, Chang BL, Lange EM, Wiklund F, Bailey-Wilson JE, Baffoe-Bonnie A, Jones M, Gildea D, Riedesel E, et al: Combined genome-wide scan for prostate cancer susceptibility genes. *J Natl Cancer Inst* 2004, 96:1240-1247.

63. Xu J, Dimitrov L, Chang BL, Adams TS, Turner AR, Meyers DA, Eeles RA, Easton DF, Foulkes WD, Simard J, et al: A combined genomewide linkage scan of 1,233 families for prostate cancer-susceptibility genes conducted by the international consortium for prostate cancer genetics. *Am J Hum Genet* 2005, 77:219-229.

64. Hsieh CL, Oakley-Girvan I, Balise RR, Halpern J, Gallagher RP, Wu AH, Kolonel LN, O'Brien LE, Lin IG, Van Den Berg DJ, et al: A genome screen of families with multiple cases of prostate cancer: evidence of genetic heterogeneity. *Am J Hum Genet* 2001, 69:148-158.

65. Goddard KA, Witte JS, Suarez BK, Catalona WJ, Olson JM: Model-free linkage analysis with covariates confirms linkage of prostate cancer to chromosomes 1 and 4. *Am J Hum Genet* 2001, 68:1197-1206.

66. Camp NJ, Farnham JM, Cannon Albright LA: Genomic search for prostate cancer predisposition loci in Utah pedigrees. *Prostate* 2005, 65:365-374.

67. Luo JH, Yu YP: Genetic factors underlying prostate cancer. *Expert Rev Mol Med* 2003, 5:1-26.

68. Langeberg WJ, Isaacs WB, Stanford JL: Genetic etiology of hereditary prostate cancer. *Front Biosci* 2007, 12:4101-4110.

69. Cussenot O, Valeri A, Berthon P, Fournier G, Mangin P: Hereditary prostate cancer and other genetic predispositions to prostate cancer. *Urol Int* 1998, 60 Suppl 2:30-34; discussion 35.

70. Simard J, Dumont M, Labuda D, Sinnett D, Meloche C, El-Alfy M, Berger L, Lees E, Labrie F, Tavtigian SV: Prostate cancer susceptibility genes: lessons learned and challenges posed. *Endocr Relat Cancer* 2003, 10:225-259.

71. Easton DF SD, Whittemore AS, Isaacs WJ: International Consortium for Prostate Cancer Genetics: Where are the prostate cancer genes?--A summary of eight genome wide searches. *Prostate* 2003, 57:261-269.

72. Witte JS: Prostate cancer genomics: towards a new understanding. *Nat Rev Genet* 2009, 10:77-82.

73. Ostrander EA, Johannesson B: Prostate cancer susceptibility loci: finding the genes. *Adv Exp Med Biol* 2008, 617:179-190.

74. Ostrander EA, Markianos K, Stanford JL: Finding prostate cancer susceptibility genes. *Annu Rev Genomics Hum Genet* 2004, 5:151-175.

75. Nupponen NN, Carpten JD: Prostate cancer susceptibility genes: many studies, many results, no answers. *Cancer Metastasis Rev* 2001, 20:155-164.

76. Eeles RA: Genetic predisposition to prostate cancer. *Prostate Cancer Prostatic Dis* 1999, 2:9-15.

77. Edwards SM, Eeles RA: Unravelling the genetics of prostate cancer. *Am J Med Genet C Semin Med Genet* 2004, 129C:65-73.

78. Rubin MA, De Marzo AM: Molecular genetics of human prostate cancer. *Mod Pathol* 2004, 17:380-388.

79. Verhage BA, Kiemeney LA: Inherited predisposition to prostate cancer. *Eur J Epidemiol* 2003, 18:1027-1036.

80. Verhage BA, Kiemeney LA: Genetic susceptibility to prostate cancer: a review. *Fam Cancer* 2003, 2:57-67.

81. Simard J, Dumont M, Soucy P, Labrie F: Perspective: prostate cancer susceptibility genes. *Endocrinology* 2002, 143:2029-2040.

82. Coughlin SS, Hall IJ: A review of genetic polymorphisms and prostate cancer risk. *Ann Epidemiol* 2002, 12:182-196.

83. Elo JP, Visakorpi T: Molecular genetics of prostate cancer. *Ann Med* 2001, 33:130-141.

84. Larson GP, Ding Y, Cheng LS, Lundberg C, Gagalang V, Rivas G, Geller L, Weitzel J, MacDonald D, Archambeau J, et al: Genetic linkage of prostate cancer risk to the chromosome 3 region bearing FHIT. *Cancer Res* 2005, 65:805-814.

85. Xu J, Zheng SL, Hawkins GA, Faith DA, Kelly B, Isaacs SD, Wiley KE, Chang B, Ewing CM, Bujnovszky P, et al: Linkage and association studies of prostate cancer susceptibility: evidence for linkage at 8p22-23. *Am J Hum Genet* 2001, 69:341-350.

86. Paiss T, Worner S, Kurtz F, Haeussler J, Hautmann RE, Gschwend JE, Herkommer K, Vogel W: Linkage of aggressive prostate cancer to chromosome 7q31-33 in German prostate cancer families. *Eur J Hum Genet* 2003, 11:17-22.

87. Verhage BA, Aben KK, Witjes JA, Straatman H, Schalken JA, Kiemeney LA: Site-specific familial aggregation of prostate cancer. *Int J Cancer* 2004, 109:611-617.

88. Risch N: A note on multiple testing procedures in linkage analysis. *Am J Hum Genet* 1991, 48:1058-1064.

89. Schaid DJ, Chang BL: Description of the International Consortium For Prostate Cancer Genetics, and failure to replicate linkage of hereditary prostate cancer to 20q13. *Prostate* 2005, 63:276-290.

90. Camp NJ, Cannon-Albright LA, Farnham JM, Baffoe-Bonnie AB, George A, Powell I, Bailey-Wilson JE, Carpten JD, Giles GG, Hopper JL, et al: Compelling evidence for a prostate cancer gene at 22q12.3 by the International Consortium for Prostate Cancer Genetics. *Hum Mol Genet* 2007, 16:1271-1278.

91. Ropers H: New Perspectives for the Elucidation of Genetic Disorders. *Am J Hum Genet* 2007, 81:199-207.

92. Camp NJ, Farnham JM, Wong J, Christensen GB, Thomas A, Cannon-Albright LA: Replication of the 10q11 and Xp11 prostate cancer risk variants: results from a Utah pedigree-based study. *Cancer Epidemiol Biomarkers Prev* 2009, 18:1290-1294.

93. Martin ER, Schmidt MA: The future is now - will the real disease gene please stand up? *Hum Hered* 2008, 66:127-135.

94. Witte JS, Schnell AH, Cordell HJ, Spielman RS, Amos CI, Miller MB, Almasy L, MacCluer JW: Introduction to Genetic Analysis Workshop 15 summaries. *Genet Epidemiol* 2007, 31 Suppl 1:S1-6.

95. Cheung VG, Spielman RS: Data for Genetic Analysis Workshop (GAW) 15, Problem 1: genetics of gene expression variation in humans. *BMC Proc* 2007, 1 Suppl 1:S2.

96. R Development Core Team: *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2006.

CHAPTER 2

GENOME-WIDE LINKAGE ANALYSIS FOR AGGRESSIVE PROSTATE CANCER IN UTAH HIGH RISK PEDIGREES

Gerald Bryce Christensen, Nicola J. Camp, James M. Farnham,

Lisa A Cannon-Albright

Published: (2007) *The Prostate*

Abstract

BACKGROUND: It has been proposed that studying alternative phenotypes, such as tumor aggressiveness, may be a solution for overcoming the apparent heterogeneity that has hindered the identification of prostate cancer genes. We present the results of a genome-scan for predisposition to aggressive prostate cancer using the Utah high-risk pedigree resource. METHODS: We identified 259 subjects with aggressive prostate cancer in 57 extended and nuclear families. Parametric and non-parametric multipoint linkage statistics were calculated for a genome-wide set of 401 microsatellite markers using the MCLINK software package. Stratification analyses by the number of affected subjects per pedigree (<5, ≥5) and the average age at diagnosis of affected subjects (<70 years, ≥70 years) were also performed. RESULTS: No significant results were observed at the genome-wide level, but suggestive evidence for linkage was observed on chromosomes 9q (HLOD=2.04) and 14q (HLOD=2.08); several pedigrees showed individual evidence for linkage at each locus (LOD > 0.58). The subset of pedigrees with earlier age at onset demonstrated nominal linkage evidence on chromosomes 3q (HLOD=1.79), 8q (HLOD=1.67), and 20q (HLOD=1.82). The late-onset subset showed suggestive linkage on chromosome 6p (HLOD=2.37) and the subset of pedigrees with fewer than five affected subjects showed suggestive linkage on chromosome 10p (HLOD=1.99). CONCLUSIONS: Linkage evidence observed on chromosomes 6p, 8q, and 20q support previously reported prostate cancer aggressiveness loci. While these results are encouraging, further research is necessary to identify the gene or genes responsible for prostate cancer aggressiveness and surmount the overarching problem of PC heterogeneity.

Introduction

Prostate cancer (PC) is the most commonly diagnosed cancer among men, and has long been recognized to occur in familial clusters. Brothers and sons of affected men have a twofold to threefold increased risk of developing prostate cancer [1-5]. Evidence that genetics plays a critical role in PC is based on results from a variety of study designs, including case-control, cohort, twin, and family-based studies [6]. However, identification of genes predisposing to prostate cancer has been difficult. In the past 10 years, investigators in the field have struggled to localize genes responsible for this common yet complex phenotype [7]. Although many candidate loci have been suggested in conventional genome-wide scans of high-risk families, successful confirmation reports have been rare. Hereditary prostate cancer is a complex disease potentially involving multiple genes and variable phenotypic expression. This genetic heterogeneity is one of the chief obstacles in understanding hereditary prostate cancer [8].

Putative prostate cancer predisposition loci identified by genetic linkage have been reported on almost all chromosomes [6]. In 2003, several investigators belonging to the International Consortium for Prostate Cancer Genetics (ICPCG) published the results of their individual linkage analyses in parallel [1,9-15]. Across these eight studies, 11 linkage peaks with LOD scores in excess of 2 were identified. However, no chromosomal region was reported as being significant at this level by more than one study and only one corresponded to a peak previously suggested by another group [7]. It has been suggested that traditional linkage analysis methods are not sufficiently powerful to localize the genes that cause complex diseases such as PC [16].

One proposed solution to the problem of heterogeneity is to use more homogeneous phenotypes in linkage analyses. An example has been the analysis of the subset of prostate cancer characterized as aggressive. Quantitative trait linkage analysis has been applied in prostate cancer using Gleason's grade, a measure of tumor aggressiveness, as the primary outcome variable, yielding evidence for regions on chromosomes 7 and 19 [17]. The region on chromosome 19q was later confirmed in another study considering Gleason's grade [18]. No genes for aggressive prostate cancer have been positively identified.

The ICPCG recently completed a genome-wide scan for PC aggressiveness defined as a qualitative trait [19]. The results of similar analyses have also been reported recently by researchers at the University of Michigan [20], Wake Forest and Johns Hopkins Universities [21], and the Fred Hutchinson Cancer Center [22]. The definition of aggressive prostate cancer considered by the ICPCG is based on a combination of clinical and pathological values including tumor stage and grade, PSA levels at diagnosis, and premature death due to PC. The ICPCG pooled analysis also required all families be only small to moderate size, to facilitate standard linkage analysis software. Hence, although the ICPCG analysis included data from the Utah prostate cancer pedigree resource, the Utah pedigrees were not analyzed in their complete form. Specifically, pedigrees were divided and trimmed before analysis, which reduced the power of the analysis to detect predisposition loci. Here we present the results of a genome-wide scan for aggressive prostate cancer predisposition loci utilizing the full Utah pedigrees.

Materials and Methods

The pedigree and genotype resources used for this analysis were described previously in a genome-wide linkage analysis of prostate cancer of 464 affected individuals in 59 Utah pedigrees [23]. For the current analysis, only cases with aggressive prostate cancer (APC) were considered affected. The phenotype data used for aggressiveness classification were obtained from Utah death certificate records and from the Utah Cancer Registry. The Utah Cancer Registry, an NCI SEER registry since 1973, contains data about all cancer events reported in the state of Utah since 1966. Prostate cancer cases were required to meet at least one of the following criteria in order to be classified as aggressive: 1) regional or distant stage; 2) poorly differentiated or undifferentiated grade; or 3) death due to metastatic prostate cancer, confirmed by death certificate. Any prostate cancer cases not meeting this criteria were classified as having unknown prostate aggressiveness status. Of the 59 Utah pedigrees analyzed previously, 57 contained at least 2 APC cases and were included in this analysis. A total of 259 APC cases were identified, 136 of whom were genotyped. Spouses and up to four children were genotyped in order to infer the genotypes of the deceased cases. All pedigrees consisted of between two and six generations, with a median of 3 generations. The mean age of prostate cancer diagnosis was 70.8 years. This is higher than the national average (about 68 years), but is similar to the mean age of diagnosis for all prostate cancer cases in the Utah Cancer Registry (70.7 years). Table 2.1 summarizes the characteristics of the pedigrees analyzed.

Genotyping was performed by the Center for Inherited Disease Research (CIDR) on a set of 401 STR markers with an average spacing of 9 cM across the 22 autosomes and the X chromosome. Details concerning laboratory methods used by CIDR are described at www.cidr.jhmi.edu. All map positions were derived from the Marshfield Genetic maps [24].

All linkage analyses were performed with MCLINK, which uses Markov Chain Monte Carlo simulation methods to sample haplotype configurations and to calculate an estimate of the LOD statistic [25]. MCLINK utilizes the robust multipoint linkage statistic proposed by Goring and Terwilliger [26], referenced hereafter as the TLOD (theta-LOD) [27]. The TLOD is analogous to a two-point LOD score, but utilizes complete multipoint inheritance information. This statistic has been successfully used to map several disease genes [23,28-31].

Three analyses were performed. Dominant and recessive parametric linkage analyses were performed with a previously published model [32]. The dominant model assumed a susceptibility allele frequency of 0.003, with penetrance of 1.00 in affected carriers, and 0.001 in noncarriers. The recessive model assumed a predisposition allele frequency of 0.15, with a penetrance of 1.0 in affected carriers and 0.001 in noncarriers. All individuals of unknown prostate aggressiveness status (all remaining individuals) were assigned a penetrance of 0.5 regardless of carrier status, making them uninformative in the analysis. The dominant and recessive models for the X chromosome differ only in the frequency of the disease allele.

Table 2.1: Summary of 57 Utah pedigrees with 2 or more aggressive prostate cancer cases

		Per Pedigree		
	Total	Mean	Min	Max
Aggressive PC cases (APC)	259	4.5	2	20
APC mean age at diagnosis	70.8	70.0	56.5	79.7
APC subjects genotyped	136	2.39	0	12
Other genotyped*	733	12.86	1	56

* connecting ancestors of cases, and spouse with up to four children were genotyped when necessary to infer genotypes

Statistics reported for these two models include the TLOD and Heterogeneity TLOD (referenced hereafter as HLOD). The third analysis calculated a nonparametric linkage (NPL) statistic for APC as a dichotomous qualitative trait. The NPL statistic was only computed for the 22 autosomes.

In addition to analyzing all APC high risk pedigrees together, we also stratified the pedigrees into selected subsets. Pedigrees were first stratified according to the average age at diagnosis of all aggressive cases, using a cutoff of 70 years. The early onset group consisted of 25 pedigrees, with 32 pedigrees in the late onset group. The pedigrees were also stratified according to the number of APC cases; 32 pedigrees had less than five APC cases and 25 pedigrees had five or more cases.

Significance of results was determined according to the standards established by Lander and Kruglyak [33]. The threshold for significant linkage is LOD = 3.30, at which level a false positive result is expected to occur with a probability of 0.05 in a full genome screen. The threshold for suggestive linkage is LOD = 1.86, which predicts 1 false positive result per genome. A threshold of LOD = 1.00 was arbitrarily selected to represent nominal linkage evidence.

Results—Parametric Analysis

The HLOD results for the dominant and recessive parametric analyses are shown in Figure 2.1. Table 2.2 summarizes the regions where at least nominal linkage was observed. No TLOD or HLOD results were statistically significant at a genome-wide level. Two regions indicated suggestive evidence for linkage: chromosome 14q (dominant HLOD=2.09 at D14S1426) and chromosome 9q (recessive HLOD=2.04 at D9S1786). Nominal evidence for linkage was also observed on chromosome 6p (recessive HLOD=1.75 at F13A1) and chromosome 3q (recessive HLOD=1.27 at D3S2460). TLOD values were generally similar to HLOD values in these regions, except on chromosome 9, where the HLOD was notably greater, with $\alpha = 0.33$.

The best evidence observed in the overall analysis was a dominant HLOD = 2.09 at D14S1426, at position 114 cM on chromosome 14q. This result was supported by six pedigrees with LOD scores greater than 0.58 ($p<0.05$). Most of these pedigrees include nonaggressive prostate cancer cases, some of which appear to share haplotypes with the linked aggressive cases. The one-LOD support interval covers a range of approximately 30 cM from about D14S1434 to the q terminus. No previous linkage results have been reported in this region.

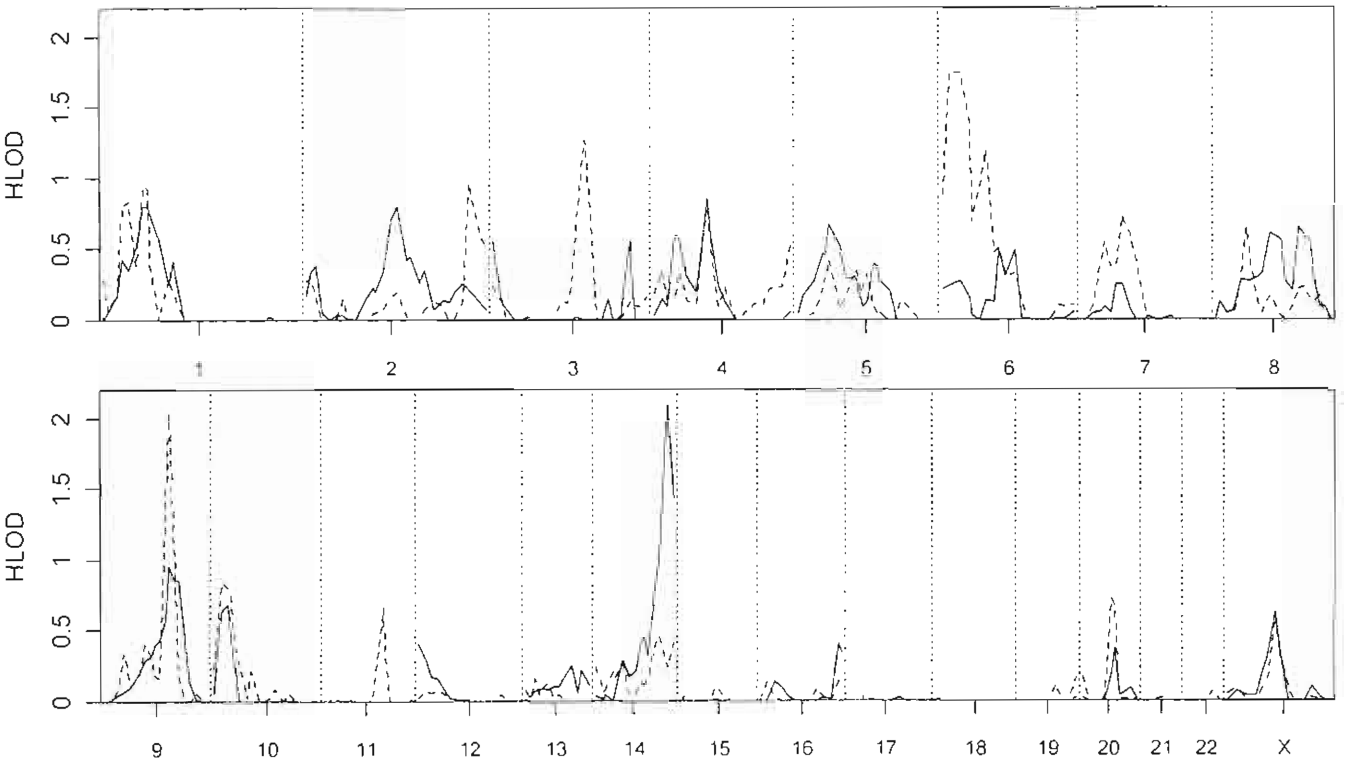


Figure 2.1. HLOD statistic for chromosomes 1-22 and X. The solid line represents the dominant model, and the broken line represents the recessive model.

Table 2.2. Summary of maximum linkage scores for each analysis model on chromosomes with at least nominal linkage evidence in the full analysis. Centimorgan positions are based on the Marshfield Genetic map.

	Dominant Model					Recessive Model					NPL	
Chr	TLOD	Pos	HLOD	α	Pos	TLOD	POS	HLOD	α	Pos	LOD	Pos
1	0	--	0.80	0.17	64	0.89	56	0.945	0.62	56	1.45	56
3	0.51	0	0.56	0.08	195	1.27	129	1.27	1.00	129	0.84	132
6	0.26	25	0.50	0.13	103	1.75	9	1.75	1.00	9	0.74	25
9	0.96	104	0.96	1.00	104	1.26	104	2.04	0.33	104	1.05	104
14	2.09	114	2.09	1.00	114	0.45	101	0.45	1.00	101	0.73	126

The second highest linkage score observed was a recessive HLOD of 2.04 at D9S1786, at map position 104 cM on chromosome 9q (TLOD = 1.26). The finding was supported by five pedigrees with LOD scores of 0.58 or greater, including a single pedigree with a LOD score of 1.63. This single pedigree consists of 3 affected siblings who share maternal and paternal haplotypes at the locus. The maternal haplotype is also shared with an affected nephew, and the paternal haplotype is shared with an affected second cousin. The pedigree also includes seven nonaggressive PC cases. Four of those cases appear to share at least one haplotype with the aggressive cases, while the remaining three appear not to share. No evidence of linkage was observed in this region in the previous genome-wide analysis of the Utah pedigrees [23].

Although it did not meet the criterion for suggestive linkage, the linkage signal on chromosome 6 is interesting because it replicates a region identified in two previous aggressive prostate cancer studies [19,20]. The signal maximized at a value of HLOD = 1.75 at a position of 9 cM from the p-terminus near F13A1. The peak is quite broad, with the one-LOD support region extending from the p-ter to about 42 cM. Six pedigrees have LOD scores exceeding 0.58 in the region. In a study of 71 families with elevated risk of prostate cancer, University of Michigan researchers reported a nonparametric LOD of 2.09 at 30cM, and a parametric HLOD = 1.52 at that position in the recessive model [20]. The ICPCG analysis [19] reported a non-parametric LOD = 3.00 at a position of 42 cM, and a recessive HLOD = 2.20 at 43 cM. The International ACTANE Consortium also reported an HLOD of 1.41 under a rare dominant model near D6S2439 (42 cM) in a study of 64 families from five countries [13], although this study was not restricted to aggressive disease.

Results—Parametric Subgroup Analysis

Genome wide HLOD results for the subset analyses are shown in Figure 2.2. Table 2.3 presents a summary of regions where at least nominal linkage evidence was observed in the subset analyses.

Early Onset

Nominal linkage evidence was observed on chromosome 20q in the early onset pedigrees. We observed a dominant HLOD = 1.82 at 52 cM, which supports the previously published HPC20 localization [34]. This region was also seen in the ICPCG pooled analysis of the aggressive prostate cancer phenotype, dominant

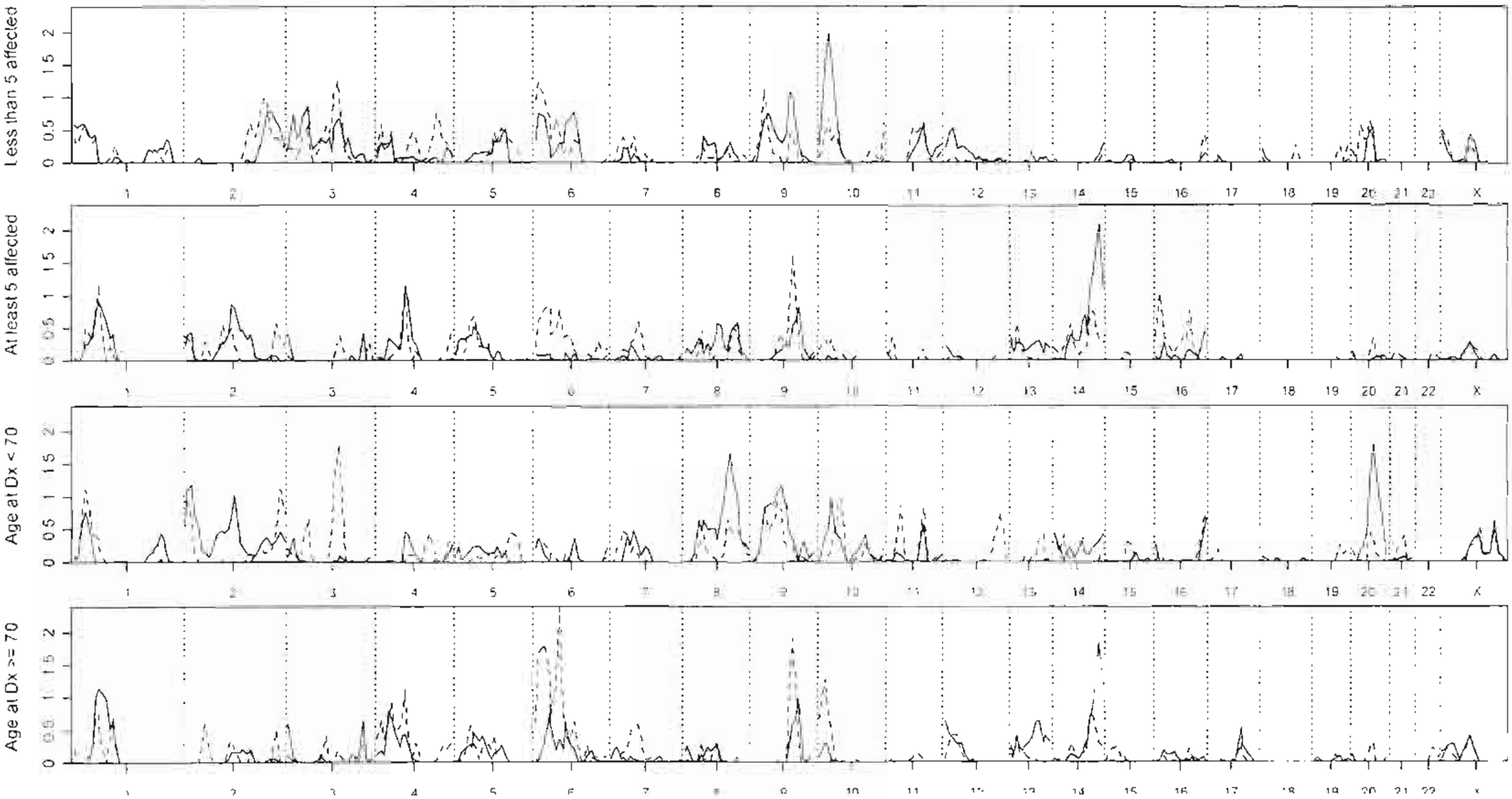


Figure 2.2. Genome-wide HLOD results for the subsets of pedigrees with less than 5 affected members, 5 or more affected members, average age at diagnosis less than 70 years, and average age at diagnosis equal to or greater than 70 years. The solid line represents results of the dominant model and the broken line represents the recessive model in each frame.

Table 2.3. Summary of chromosomes with HLOD values greater than 1.0 for subsets. Centimorgan positions are based on the Marshfield Genetic map.

Subset	Chromosome	Position (cM)	HLOD	Model
<5 Affected	3	129	1.25	Recessive
	6	9	1.23	Recessive
	9	32	1.12	Recessive
	9	98	1.08	Dominant
	10	24	1.99	Dominant
≥5 Affected	1	64	1.19	Recessive
	4	75	1.16	Dominant
	9	104	1.63	Recessive
	14	114	2.10	Dominant
	16	11	1.03	Recessive
Early onset	1	28	1.12	Recessive
	2	14	1.19	Dominant
	2	129	1.04	Dominant
	2	248	1.13	Recessive
	3	132	1.79	Recessive
	8	118	1.67	Dominant
	9	76	1.16	Dominant
	20	52	1.82	Dominant
Late onset	1	64	1.14	Dominant
	4	75	1.13	Recessive
	6	63	2.37	Recessive
	9	104	1.90	Recessive
	10	15	1.29	Recessive
	14	114	1.83	Dominant

HLOD = 2.49 at 54 cM [19]. However, contrary to our results, the ICPCG study showed slightly stronger results in the late-onset group, achieving a maximum HLOD of 2.65. Five Utah pedigrees showed individual linkage evidence (LOD > 0.58) in this region.

Two other regions on chromosomes 3q and 8q were observed in the early-onset pedigrees with LOD>1.5. On chromosome 3q, we observed a recessive HLOD = 1.79 at D3S4523 (132 cM). This finding is primarily supported by four pedigrees with individual LOD values greater than 0.58 (p<0.05). Linkage was previously reported in the Utah pedigrees on chromosome 3 with a dominant inheritance model [23], but the linkage evidence was centered around D3S2427 (182 cM), which is identified in Figure 2.3. On chromosome 8q, we observed a dominant HLOD = 1.67 at 118 cM, near D8S1132. Two recent studies have identified possible prostate cancer loci in this area [35,36]. Seven pedigrees in our resource showed nominal individual linkage (LOD > 0.58) in this region. There was no previous linkage evidence for chromosome 8 reported for the Utah pedigrees.

Late Onset

Suggestive evidence for linkage in the late-onset pedigrees was observed for chromosomes 6 and 9. The LOD scores for both chromosomes were similar to the analysis of all pedigrees combined. However, on chromosome 6, a second, independent region of linkage evidence emerged slightly downstream (HLOD = 2.37, at marker D6S1017, 63 cM) (Figure 2.3). Six pedigrees have LOD scores in excess of 0.58 (p < 0.05) in this region, although three of those also show reduced linkage (but

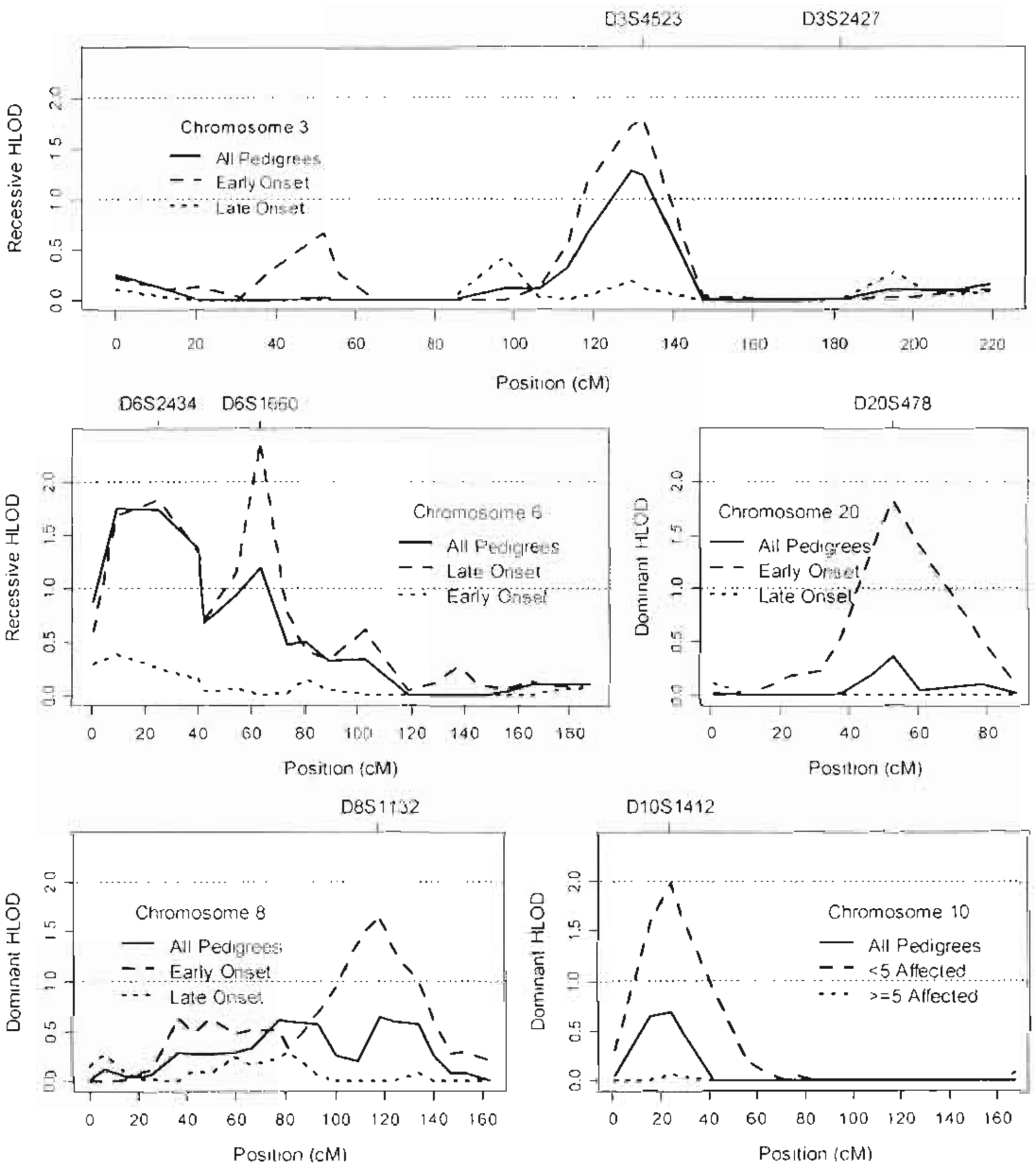


Figure 2.3. HLOD tracings for chromosomes showing at least nominal linkage evidence in subset analyses. The solid line in each figure represents the results from the analysis of all APC pedigrees combined, and the broken lines represent the indicated subsets.

still >0.58) in the upstream peak. This is interesting when compared to the ICPCG pooled analysis [19], where suggestive evidence was also identified in the recessive model (HLOD=1.98), which was strengthened in the late-onset pedigrees (HLOD=2.40). However, The ICPCG and ACTANE linkage scores both maximized near 42 cM [13,19], which represents a local minimum in our linkage graph, as shown in Figure 2.3. The chromosome 6 downstream peak that emerged in the late-onset subset analysis is closer to a locus suggested by Janer et al., who reported a dominant HLOD of 2.51 in a study of 254 families at marker D6S1281. The HLOD statistic was as high as 3.43 in one subgroup [12]. It must be noted that unlinked pedigrees can shift linkage peaks, and whether there are in fact two distinct regions is yet to be determined.

Less than Five APC

The subgroup of pedigrees with fewer than 5 affected aggressive prostate cancer cases yielded suggestive evidence on chromosome 10p (dominant HLOD = 1.99 at marker D10S1412, 24 cM). The one-LOD support interval extends from about 10 cM to 40 cM. Primary support for this peak comes from 6 pedigrees with individually significant linkage evidence (LOD > 0.58). Nominal linkage for prostate cancer was previously observed on chromosome 10p by Wake Forest/Johns Hopkins, who reported a LOD score of 1.39 at D10S249, near the p-terminus [9]. Our result does not appear to support that finding.

Five or More APC

The subgroup of pedigrees with five or more aggressive prostate cancer cases showed suggestive evidence of linkage on chromosome 14 (dominant HLOD = 2.10 at 114cM), similar to the evidence observed in the overall analysis.

Results—Nonparametric Analysis

Figure 2.4 shows the qualitative NPL statistic for the 22 autosomes. No significant linkage evidence was observed. The highest NPL statistic observed across the entire genome was 2.33 at D1S255 (56 cM) on chromosome 1p, corresponding to a LOD = 1.45. The region is about 25 cM removed from a significant chromosome 1 linkage previously reported in a single extended Utah pedigree; nominal linkage evidence was observed at the same locus in all pedigrees combined [23]. This signal is close to the CAPB locus [37], and a study using Gleason grade as a quantitative measure of prostate cancer aggressiveness also reported suggestive linkage in the region [17].

Discussion

Significant genetic heterogeneity in prostate cancer has been invoked to explain the many different published hints of linkage, as well as the failure of other studies to independently confirm most of these published linkages. A variety of approaches to select more homogeneous prostate cancer phenotypes have been attempted. One such approach has been the analysis of aggressive prostate cancer. Published analyses of the aggressive phenotype have to date not proven successful for

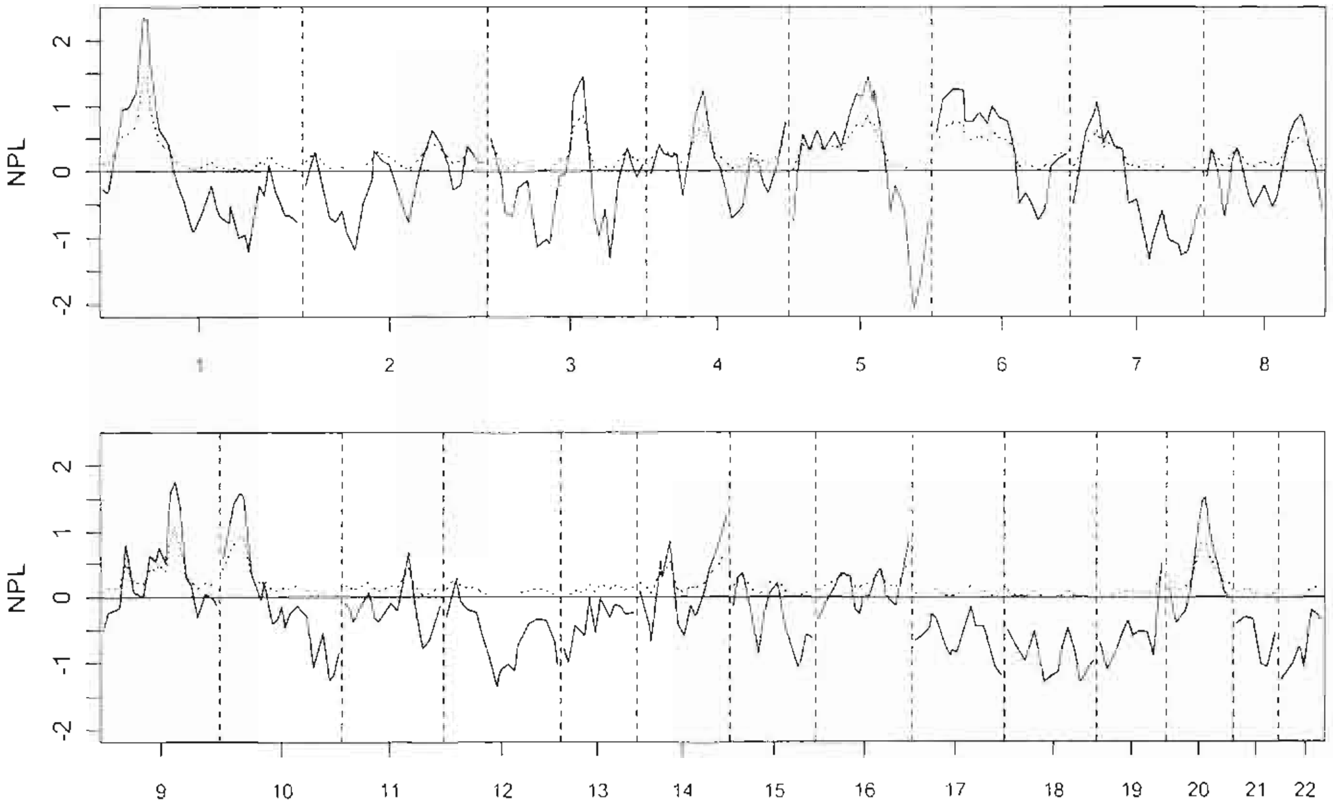


Figure 2.4. Qualitative NPL statistic for chromosomes 1-22. The solid line represents the NPL statistic, while the broken line shows the equivalent LOD value.

gene identification, but the existence of some replications across studies is encouraging.

In this analysis of aggressive prostate cancer in the extended Utah pedigrees. no significant linkages were identified, but, given the gold-standard is replication. rather than single study significance. it is encouraging that we have replicated regions also identified in other aggressive prostate cancer linkage studies. Suggestive hints of linkage were identified that showed pedigree-specific linkage support from multiple pedigrees. Among these regions. the Utah evidence for the chromosome 6p and 20q regions confirm previous suggestions of linkage to prostate cancer aggressiveness. Suggestive linkage regions on chromosomes 1p and 8q also support previously reported prostate cancer loci. As replication of linkage results is historically rare in prostate cancer genetics research, these results should not be overlooked.

The aggressiveness loci on chromosomes 6p and 20q were also identified in the ICPCG analysis of aggressive prostate cancer [19], which included data from Utah pedigrees. However, the data that were submitted to the ICPCG represent only a fraction of the data used in this study. The software used in the ICPCG pooled analysis was not capable of analyzing the complex pedigree structures in the Utah resource. Therefore, the data that were submitted consisted of smaller families and branches that had been excised from the extended pedigrees in our resource. The results we present in this report are based on significantly more data, including several large and complex pedigree structures that were not used in the ICPCG analysis. The extended pedigrees give us greater power to determine haplotype structures and inheritance patterns, especially in the case of rare alleles. This analysis also used a different phenotype definition than the ICPCG analysis, as we recognized any prostate cancer-related death to be aggressive, as opposed to only those deaths that occurred before 65 years of age.

The existence of intrafamilial heterogeneity could have affected this analysis. The pedigrees used were originally ascertained for an excess of all prostate cancers. When only the aggressive prostate cancers in each pedigree were considered, the number of affected subjects in each pedigree decreased and the genetic distance separating the cases increased. High-frequency alleles could therefore act as confounders in some cases, as it becomes difficult to determine if they were inherited from a common ancestor.

Though derived from the same data source, there is very little overlap in the results of this study and the results of the genomic scan for all prostate cancers for which this data was originally ascertained [23]. The strongest linkage signals reported in that study were on chromosomes 1, 3, 5, and 22. Some evidence of those results is visible in the present analysis, but is not noteworthy. The most striking example is chromosome 22, where the previously observed peak has completely vanished in this study using an alternate phenotype definition. Conversely, the suggestive linkage results we now report are not generally seen in the previous analysis. Most of the pedigrees that show evidence of linkage to the regions we report in this study also included individuals with nonaggressive prostate cancer. In most cases, some of those cases appeared to share haplotypes with the aggressive cases. The alternate phenotype definition clearly affected the outcome of the analysis.

Conclusion

Although this analysis did not identify any regions with significant linkage evidence at the genome-wide level, regions of interest were identified on chromosomes 9, 14, 6, and 1. The result on chromosome 6 appears to support linkage evidence reported by the ICPCG as well as the University of Michigan and the ACTANE consortium. Additionally, using pedigree subsets of the data resource identified regions of interest on chromosomes 3, 8, 10, and 20. The chromosome 20 result supports previous findings reported by researchers at Mayo Clinic and the ICPCG. We did not find sufficient evidence to support linkage regions previously reported for aggressive prostate cancer by the ICPCG for chromosome 11, by the University of Michigan for chromosome 15, or by Wake Forest/Johns Hopkins or Fred Hutchinson for Chromosome 22. Further research is necessary to identify the gene or genes responsible for prostate cancer aggressiveness and surmount the overarching problem of PC heterogeneity.

Acknowledgements

Data collection for this publication was supported by NIH National Cancer Institute grant R01 CA90752 (to L.A. Cannon Albright); a subcontract from Johns Hopkins University with funds provided by grant R01 CA89600 from the NIH National Cancer Institute (to L.A. Cannon Albright); NIH grant K07 CA98364 (to N.J. Camp); NIH National Library of Medicine training grant T15 LM07124 (to G.B. Christensen). Some data for this research was supported by the Utah Cancer Registry, which is funded by contract N01-PC-35141 from the NCI, with additional support from the Utah State Department of Heath and the University of Utah. Partial support for all datasets within the Utah Population Database (UPDB) was provided by the University of Utah Huntsman Cancer Institute. This investigation was supported by the Public Health Services research grant number M01-RR00064 from the National Center for Research Resources. Genotyping services were provided by the Center for Inherited Disease Research (CIDR). CIDR is fully funded through a federal contract from the National Institutes of Health to The Johns Hopkins University, contract number N01-HG-65403.

References

1. Cunningham JM, McDonnell SK, Marks A, Hebbring S, Anderson SA, Peterson BJ, Slager S, French A, Blute ML, Schaid DJ, Thibodeau SN: Genome linkage screen for prostate cancer susceptibility loci: results from the Mayo Clinic Familial Prostate Cancer Study. *Prostate* 2003, 57:335-346.

2. Goldgar DE ED, Cannon-Albright LA, Skolnik MH: Systematic population-based assessment of cancer risk in first-degree realtives of cancer probands. *JNCI* 1994, 86:1600-1608.

3. Steinberg GD CB, Beaty TH, Childs B, Walsh PC: Family history and the risk of prostate cancer. *Prostate* 1990, 17:337-347.

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

5. Cannon-Albright LA SA, Camp NJ, Farnham JS, Thomas A: Population-based risk assessment for other cancers in relatives of hereditary prostate cancer cases. *Prostate* 2005, 64:347-355.

6. Schaid D: The Complex Genetic Epidemiology of Prostate Cancer. *Human Molecular Genetics* 2004, 13:R103-121.

7. Easton DF SD. Whittemore AS. Isaacs WJ: International Consortium for Prostate Cancer Genetics: Where are the prostate cancer genes?--A summary of eight genome wide searches. *Prostate* 2003. 57:261-269.

8. Ostrander EA, Stanford JL: Genetics of prostate cancer: too many loci, too few genes. *Am J Hum Genet* 2000, 67:1367-1375.

9. Xu J, Gillanders EM, Isaacs SD, Chang BL, Wiley KE, Zheng SL, Jones M, Gildea D, Riedesel E, Albertus J, et al: Genome-wide scan for prostate cancer susceptibility genes in the Johns Hopkins hereditary prostate cancer families. *Prostate* 2003. 57:320-325.

10. Witte JS, Suarez BK, Thiel B, Lin J, Yu A, Banerjee TK, Burmester JK, Casey G, Catalona WJ: Genome-wide scan of brothers: replication and fine mapping of prostate cancer susceptibility and aggressiveness loci. *Prostate* 2003. 57:298-308.

11. Schleutker J, Baffoe-Bonnie AB, Gillanders E, Kainu T, Jones MP, Freas-Lutz D, Markey C, Gildea D, Riedesel E, Albertus J, et al: Genome-wide scan for linkage in Finnish hereditary prostate cancer (HPC) families identifies novel susceptibility loci at 11q14 and 3p25-26. *Prostate* 2003. 57:280-289.

12. Janer M, Friedrichsen DM, Stanford JL, Badzioch MD, Kolb S, Deutsch K, Peters MA, Goode EL, Welti R, DeFrance HB, et al: Genomic scan of 254 hereditary prostate cancer families. *Prostate* 2003, 57:309-319.

13. Edwards S, Meitz J, Eles R, Evans C, Easton D, Hopper J, Giles G, Foulkes WD, Narod S, Simard J, et al: Results of a genome-wide linkage analysis in prostate cancer families ascertained through the ACTANE consortium. *Prostate* 2003, 57:270-279.

14. Lange EM, Gillanders EM, Davis CC, Brown WM, Campbell JK, Jones M, Gildea D, Riedesel E, Albertus J, Freas-Lutz D, et al: Genome-wide scan for prostate cancer susceptibility genes using families from the University of Michigan prostate cancer genetics project finds evidence for linkage on chromosome 17 near BRCA1. *Prostate* 2003, 57:326-334.

15. Wiklund F, Gillanders EM, Albertus JA, Bergh A, Damber JE, Emanuelsson M, Freas-Lutz DL, Gildea DE, Goransson I, Jones MS, et al: Genome-wide scan of Swedish families with hereditary prostate cancer: suggestive evidence of linkage at 5q11.2 and 19p13.3. *Prostate* 2003, 57:290-297.

16. Terwilliger JD, Goring HH: Gene mapping in the 20th and 21st centuries: statistical methods, data analysis, and experimental design. *Hum Biol* 2000, 72:63-132.

17. Witte JS, Goddard KA, Conti DV, Elston RC, Lin J, Suarez BK, Broman KW, Burmester JK, Weber JL, Catalona WJ: Genomewide scan for prostate cancer-aggressiveness loci. *Am J Hum Genet* 2000, 67:92-99.

18. Slager SL, Schaid DJ, Cunningham JM, McDonnell SK, Marks AF, Peterson BJ, Hebbring SJ, Anderson S, French AJ, Thibodeau SN: Confirmation of linkage of prostate cancer aggressiveness with chromosome 19q. *Am J Hum Genet* 2003, 72:759-762.

19. Schaid D, ICPCG: Pooled genome linkage scan of aggressive prostate cancer: Results from the International Consortium for Prostate Cancer Genetics. *Hum Genet* In Press.

20. Lange E, Ho L, Beebe-Dimmer J, Wang Y, Gillanders E, Trent J, Lange L, Wood D, Cooney K: Genome-wide linkage scan for prostate cancer susceptibility genes in men with aggressive disease: significant evidence for linkage at chromosome 15q12. *Hum Genet* 2006, 119:400-407.

21. Chang BL IS, Wiley KE, Gillanders EM, Zheng SL, Meyers DA, Walsh PC, Trent JM, Xu J, Isaacs WB: Genome-wide screen for prostate cancer susceptibility genes in men with clinically significant disease. *Prostate* 2005, 64:356-361.

22. Stanford J, McDonnell S, Friedrichsen D, Carlson E, Kolb S, Deutsch K, Janer M, Hood L, Ostrander E, Schaid D: Prostate cancer and genetic susceptibility: a genome scan incorporating disease aggressiveness. *Prostate* 2006, 66:317-325.

23. Camp NJ FJ, Cannon-Albright LA: Genomic search for prostate cancer predisposition loci in Utah pedigrees. *Prostate* 2005, 65:365-374.

24. Broman K, Murray J, Sheffield V, White R, Weber J: Comprehensive human genetic maps: Individual and sex-specific variation in recombination. *Am J Hum Genet* 1998, 63:861-869.

25. Thomas A, Gutin A, Abkevich V, Bansal A: Multipoint linkage analysis by blocked Gibbs sampling. *Statistics and Computing* 2000, 10:259-269.

26. Goring HH, Terwilliger JD: Linkage analysis in the presence of errors I: complex-valued recombination fractions and complex phenotypes. *Am J Hum Genet* 2000, 66:1095-1106.

27. Abkevich V, Camp NJ, Gutin A, Farnham JM, Cannon-Albright L, Thomas A: A robust multipoint linkage statistic (tlod) for mapping complex trait loci. *Genet Epidemiol* 2001, 21 Suppl 1:S492-497.

28. Farnham J, Camp N, Neuhausen S, Tsuruda J, Parker D, MacDonald J, Cannon-Albright L: Confirmation of chromosome 7q11 locus for predispositionto intracranial aneurysm. *Hum Genet* 2004, 114:250-255.

29. Farnham JM, Camp NJ, Swensen J, Tavtigian SV, Albright LA: Confirmation of the HPCX prostate cancer predisposition locus in large Utah prostate cancer pedigrees. *Hum Genet* 2005, 116:179-185.

30. Stone S, Abkevich V, Hunt SC, Gutin A, Russell DL, Neff CD, Riley R, Frech GC, Hensel CH, Jammulapati S, et al: A major predisposition locus for severe obesity, at 4p15-p14. *Am J Hum Genet* 2002, 70:1459-1468.

31. Camp NJ LM, Richards RL, Plenk AM, Carter C, Hensel CH, Abkevich V, Skolnick MH, Shattuck D, Rowe KG, Hughes DC, Cannon-Albright LA: Genome-wide linkage analyses of extended Utah pedigrees identifies loci that influence recurrent, early-onset major depression and anxiety disorders. *Am J Med Genet B Neuropsychiatr Genet* 2005, 135B:85-93.

32. Smith JR, Freije D, Carpten JD, Gronberg H, Xu J, Isaacs SD, Brownstein MJ, Bova GS, Guo H, Bujnovszky P, et al: Major susceptibility locus for prostate cancer on chromosome 1 suggested by a genome-wide search. *Science* 1996, 274:1371-1374.

33. Lander E, Kruglyak L: Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* 1995, 11:241-247.

34. Berry R, Schroeder JJ, French AJ, McDonnell SK, Peterson BJ, Cunningham JM, Thibodeau SN, Schaid DJ: Evidence for a prostate cancer-susceptibility locus on chromosome 20. *Am J Hum Genet* 2000, 67:82-91.

35. Amundadottir LT SP, Gudmundsson J, Helgason A, Baker A, Agnarsson BA, Sigurdsson A, Benediktsdottir KR, Cazier JB, Sainz J, Jakobsdottir M, Kostic J, Magnusdottir DN, Ghosh S, Agnarsson K, Birgisdottir B, Le Roux L, Olafsdottir A, Blondal T, Andresdottir M, Gretarsdottir OS, Bergthorsson JT, Gudbjartsson D, Gylfason A, Thorleifsson G, Manolescu A, Kristjansson K, Geirsson G, Isaksson H, Douglas J, Johansson JE, Balter K, Wiklund F, Montie JE, Yu X, Suarez BK, Ober C, Cooney KA, Gronberg H, Catalona WJ, Einarsson GV, Barkardottir RB, Gulcher JR, Kong A, Thorsteinsdottir U, Stefansson K.: A common variant associated with prostate cancer in European and African populations. *Nature Genetics* 2006, 38:652-658.

36. Freedman M, Haiman C, Patterson N, McDonald G, Tandon A, Waliszewska A, Penney K, Steen R, Ardlie K, John E, et al: Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proc Natl Acad Sci* 2006, 103:14068-14073.

37. Gibbs M, Stanford JL, McIndoe RA, Jarvik GP, Kolb S, Goode EL, Chakrabarti L, Schuster EF, Buckley VA, Miller EL, et al: Evidence for a rare prostate cancer-susceptibility locus at chromosome 1p36. *Am J Hum Genet* 1999, 64:776-787.

CHAPTER 3

EXTRACTING DISEASE RISK PROFILES FROM EXPRESSION DATA FOR LINKAGE ANALYSIS: APPLICATION TO PROSTATE CANCER

Gerald Bryce Christensen, Lisa A. Cannon-Albright,
Alun Thomas, Nicola J. Camp

Published: (2007) *BMC Proceedings*

Abstract

The genetic factors underlying many complex traits are not well understood. The GAW15 Problem 1 data presents the opportunity to explore whether gene expression data from microarrays can be utilized to define useful phenotypes for linkage analysis in complex diseases. We describe a simple approach that utilizes expression profiles for multiple genes that have been associated with a disease, to develop a composite 'risk profile' that can be used to map other loci involved in the same disease process. Using prostate cancer (PCa) as our disease of interest, we identified 26 genes whose expression levels had previously been associated with PCa, and we defined three phenotypes: high, neutral, or low risk profiles, based on individual expression levels. Linkage analyses using MCLINK, a Markov chain Monte Carlo method, and MERLIN were performed for all three phenotypes. Both methods were in very close agreement. Genome-wide suggestive linkage evidence was observed on chromosomes 6 and 4. It was interesting to note that the linkage signals did not appear to be strongly influenced by the location of the original 26 genes used in the phenotype definition indicating that composite measures may have Fpotential to locate additional genes in the same process. In this example, however, extreme caution is necessary in any extrapolation of the identified loci to PCa due to the lack of data regarding the behavior of these genes' expression level in lymphoblastoid cells. Our results do indicate there exists potential to augment our current knowledge about the relationships among genes associated with complex diseases using expression data.

Background

Recent advances in biotechnology have resulted in an explosion of genotypic and phenotypic data. Millions of single nucleotide polymorphisms (SNPs) can quickly and accurately be genotyped, and microarray technology has made it possible to simultaneously assess the expression levels for many thousands of genes. The question becomes what knowledge can we extract from these extensive data sources with respect to disease susceptibility, and how? The GAW15 Problem 1 data presents a unique opportunity to explore whether gene expression data from microarrays can be used to define useful phenotypes for linkage analysis to better understand disease susceptibility. The expression data provided for Problem 1 includes 3554 genes that were previously established to have greater variation between individuals than within individuals. These expression levels are reasonable candidates for use as phenotypes in linkage analysis [1].

For the majority of complex traits, the underlying genetic factors are not fully understood, but for many, certain genes and/or genetic pathways have been implicated or related to the trait through expression experiments. The expression levels of a gene may be controlled by regulatory genes elsewhere in the genome, and the expression of multiple genes can be regulated by a common transcription factor[2]. Hence, linkage analysis of gene expression levels could conceivably identify regulatory loci associated with that gene. Further, and more related to a disease end-point, if several genes are known to be related to a given trait, it is also conceivable that their expression levels could be combined to create a phenotype to be used in linkage analysis to identify loci that are involved in disease susceptibility, perhaps through membership in the pathway or interaction (epistasis) with the known genes.

In this study, we explore whether gene expression profiles for genes that have been associated with a disease can be used to map other genes that are involved in the disease process or highlight genes within the pathways that are key factors. Here we specifically examine the approach for Prostate Cancer (PCa).

Research has consistently shown that genetics plays a critical role in PCa development, but the identification of PCa genes has proven to be very difficult. Hereditary prostate cancer is a complex disease involving numerous genes and variable phenotypic expression[3]. Recent research has demonstrated great potential for the use of proteomic profiling and other biomarkers for PCa diagnostics[4]. One such study was able to discriminate PCa from benign prostates with perfect sensitivity in men with elevated prostate specific antigen (PSA) levels using serum proteomic profiling[5]. The GAW15 Problem 1 data provide an opportunity to explore whether gene expression levels from lymphoblastoid cells can be used to develop a prostate cancer profile phenotype for use in linkage analysis. Using expression data from 26 genes whose expression levels had previously been reported to be associated with PCa [6], we defined individuals as having high, neutral, or low risk profiles based on their individual expression levels. Here we present the results of linkage analyses based on those phenotypes.

Methods

Ashida identified 21 genes that are commonly up-regulated and 63 genes that are commonly down-regulated in the transition from normal epithelium to PCa and/or prostatic intraepithelial neoplasia (PIN)[6]. Of these 84 genes, 26 were included in the data for Problem 1. These 26 genes are listed in Table 3.1. Based on the expression data for the 194 individuals in the Problem 1 data, we scaled the expression levels for each of these 26 genes to fit a standard normal distribution with mean 0 and variance 1. Two statistics, ***A*** and ***B***, were then computed for each individual. ***A*** represented the number of genes for which the expression level was greater than 1 standard deviation in the direction associated with PCa. ***B*** represented the number of genes for which the expression level was greater than 1 standard deviation in the opposite direction. One standard deviation was selected arbitrarily as a threshold to ensure that the expression values were distant from the center of the distribution while allowing for a sufficient number of informative subjects in the subsequent linkage analysis An individual was considered to be in the "high-risk profile" group if $\boldsymbol{A} \geq 4$ and $\boldsymbol{A}$-$\boldsymbol{B} \geq 2$. Individuals were classified to be in the "low-risk profile" group if $\boldsymbol{B} \geq 4$ and $\boldsymbol{B}$-$\boldsymbol{A} \geq 2$. All other subjects were classified as "neutral" and were considered as "unknown" in all linkage analyses. This classification system was devised to distribute the influence of the 26 genes on the assigned risk profiles and to prevent outlying expression levels of individual genes from having undue influence. As shown in Figure 3.1, 53 subjects (25 male and 28 female) were classified with high-risk profiles, 57 (32 male and 25 female) with low-risk profiles, and 84 (42 male and 42 female) as neutral (unknown). While women are not susceptible to PCa, they may still carry the susceptibility genes

Table 3.1: Genes used to create phenotype definition

	Gene	Location
Up-regulated	ABCC4	chr13q32
	AMACR	chr5p13.2-q11.1
	MIPEP	chr13q12
	PRC1	chr15q26.1
	SMS	chrXp22.1
Down-regulated	ANXA2	chr15q21-q22
	ARHGDIB	chr12p12.3
	ASS	chr9q34.1
	BHLHB2	chr3p26
	CD74	chr5q32
	CSPG2	chr5q14.3
	CUTL1	chr7q22.1
	CX3CL1	chr16q13
	FHL2	chr2q12-q14
	FLNA	chrXq28
	GATA3	chr10p15
	GBP2	chr1p22.2
	IER3	chr6p21.3
	IRF1	chr5q31.1
	KRT7	chr12q12-q13
	LY6E	chr8q24.3
	MMP7	chr11q21-q22
	MYL9	chr20q11.23
	SERPINB1	chr6p25
	TOP2B	chr3p24
	WFDC2	chr20q12-q13.2

for PCa; hence, in our analyses, both males and females are included. Figure 3.1 shows a scattergram of the values of *A* and ***B*** for each individual and the categorization to the high-risk, low-risk, and neutral groups.

Three phenotype models were considered. The first model ("*FULL*") included the high-risk profile individuals as "affected" and the low-risk profile individuals as "unaffected"; neutrals were "unknown." The second model ("*HIGH*") included the high-risk profile individuals as "affected" and all others as "unknown." The third model ("*LOW*") included the low-risk profile individuals as "affected" and all others as "unknown." This final phenotype model is akin to an analysis searching for protective genes. For the *FULL* and *HIGH* phenotype models, 10 of the 14 CEPH pedigrees were informative for linkage, with between 2 and 8 affected subjects per pedigree. Thirteen pedigrees were informative in the *LOW* analysis, with up to 9 affected subjects.

Dominant and recessive parametric linkage analyses were performed using MCLINK, which uses Markov chain Monte Carlo simulation methods to sample haplotype configurations to estimate the LOD statistic[7]. The inheritance model for the analysis was based on the "Smith" model used to map the HPC1 locus, but without the specificity to males[8], and assumes a population prevalence of 0.003 for the mutant allele. Genotypes for a genome-wide panel of 2,882 SNP markers were provided by GAW. The genetic map used in the analysis was based on the Rutgers genetic map, with the positions of SNPs for which genetic map position was not available interpolated from flanking markers based on physical location[9]. Any SNP located less than 0.001 cM from the preceding SNP was eliminated from the initial

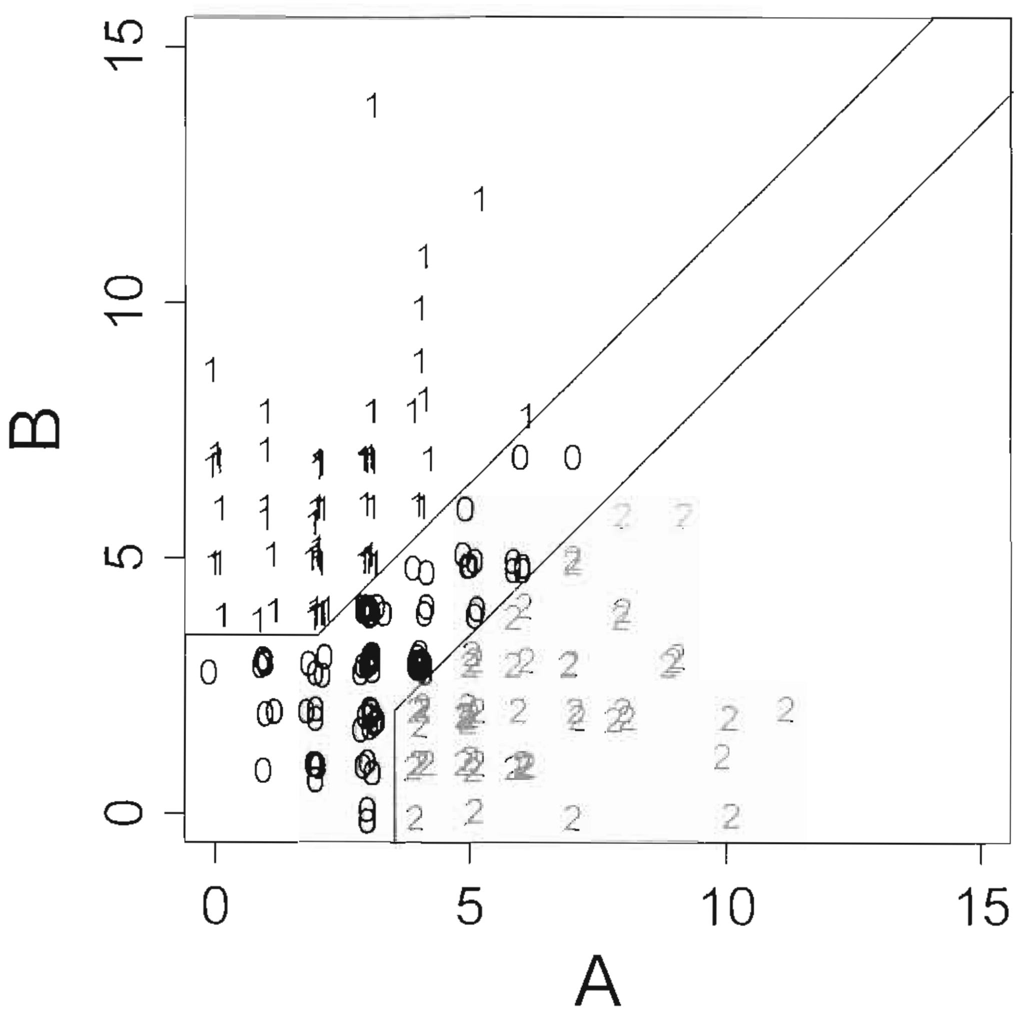


Figure 3.1. Phenotype distribution. A is the number of genes expressed more than 1 SD in the direction associated with PCa for an individual, and B is the number of genes expressed greater than 1 SD in the opposite direction. 0=neutral risk status; 1=low-risk profile; 2=high-risk profile.

analysis. After completing the initial analyses, the best linkage peaks were identified and those regions were reanalyzed using a reduced marker map, with a minimum spacing of 0.3 cM between SNPs[10]. This was done to control for the possible effects of linkage disequilibrium (LD), which may inflate LOD scores. The linkage statistics for these chromosomes were then confirmed by performing both parametric and model-free analyses with MERLIN[11]. Linked pedigrees (LOD>0.588, which represents a nominal, uncorrected p<0.05 for an individual pedigree) were identified in the regions with HLOD>1.9 (genome-wide suggestive evidence for linkage[12]) and gene expression profiles within those pedigrees were inspected to ensure that the linkage evidence was not correlated with the expression levels of any specific genes.

Results

The genome-wide scan results showing the HLOD statistic for all models are shown in Figure 3.2. Significant linkage evidence was observed on chromosome 6q (HLOD = 3.51). Other peaks over HLOD=1.9 were observed on chromosomes 3, 4, and 7. Only the peaks on chromosomes 4 and 6 retained at least suggestive linkage evidence with the reduced marker set without LD.

The strongest linkage signal observed in the *FULL* analysis, and the best result overall, was HLOD=3.51 at marker rs1491074 under the dominant model on chromosome 6q. As is shown in Figure 3.3, two of the 26 genes used in creating the phenotype (*SERPINB1* and *IER3*) are located on chromosome 6, however, they are not situated near the linkage peak. Chromosome 6 was reanalyzed using a map with increased marker spacing (which reduced the number of SNPs used from 101 to 70

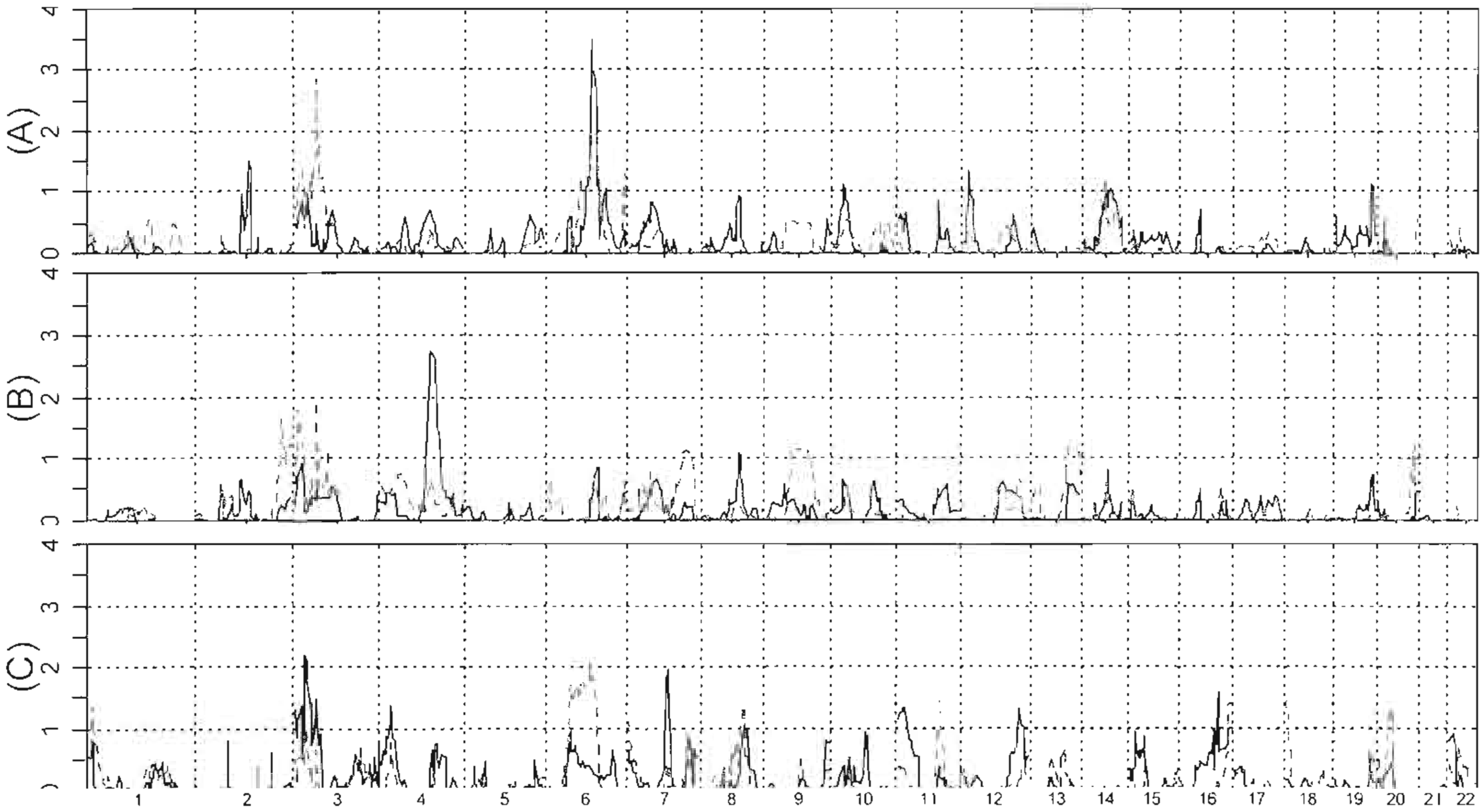


Figure 3.2. Genome-wide HLOD statistics. A) FULL analysis model B) HIGH model C) LOW (protective) model. The solid line represents the dominant inheritance model and the broken line represents the recessive model in each figure. HLOD values are shown on the vertical axis, and chromosome number is shown on the horizontal axis.

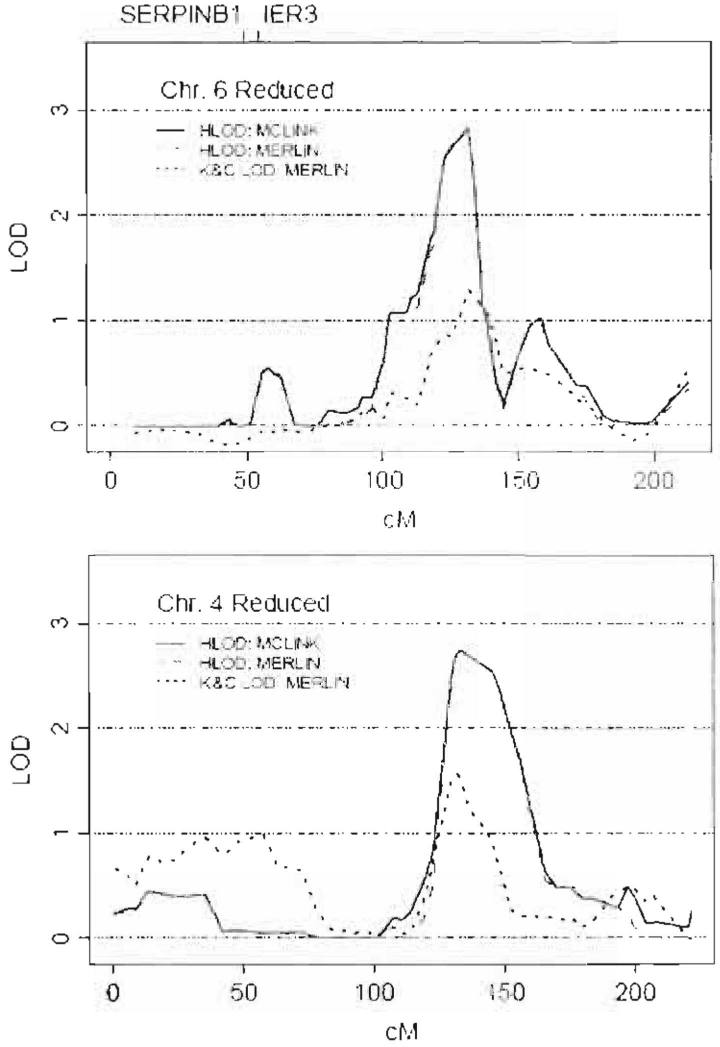


Figure 3.3. Analyses with increased marker spacing. Detail of Chr. 6 from the *FULL* phenotype model and Chr. 4 from the *HIGH* phenotype model using a minimum marker spacing of 0.3 cM. The solid line in each panel represents the dominant HLOD statistic as calculated by MCLINK, the broken line shows the dominant HLOD from Merlin, and the dotted line shows the model-free Kong and Cox lod score from MERLIN. The locations of genes included in the phenotype definition are indicated at the top of each frame.

and excluded SNP rs1491074) and the maximum HLOD fell to 2.82, suggesting the possible influence of LD in the initial result. This result was confirmed using MERLIN. The model-based HLOD statistic from MERLIN was very similar to results from MCLINK for both the full and reduced marker sets, although the model-free Kong and Cox LOD score did not perform well.

The best result in the *HIGH* analysis was HLOD=2.75 at marker rs885103 under the dominant model on chromosome 4q. Three pedigrees were linked to the locus with individual LOD scores >0.588. None of the genes used to determine the phenotype were located on chromosome 4. Linkage results were unchanged when the peak was reanalyzed with the reduced marker map, as shown in Figure 3.3. MERLIN analysis confirmed the parametric linkage result from MCLINK.

Discussion

One concern of a study based on expression levels of known genes is that a linkage analysis may simply map back to the genes used to construct the phenotype. This did not appear to be the case for this study. None of the genes were located near our best results on chromosomes 6 and 4. Our phenotype definition was simplistic, but was designed to limit the influence of individual genes on the phenotype, and thereby enhance the likelihood of identifying a locus related to the entire set. It is interesting to note that the regions we identified on chromosomes 6 [13,14] and 4q [15,16] have each been implicated in previous linkage analyses for PCa. However, it is premature to consider these as replications, as without data indicating that the expression levels seen in tumor[6] are also representative in lymphoblastoid cells, there is no evidence that the risk profiles we created are actually related to PCa. This is a major weakness of our particular PCa example, and perhaps illustrates the weakness of such approaches in general--that is, much of the experimental data is still missing and will be expensive to generate.

Because the true locations of any genes that interact with or modify the 26 we studied are not known, the statistical power of this approach can not be properly evaluated. However, with the 14 CEPH pedigrees, we were able to generate linkage peaks that appeared distinct from background noise. Further, we know that the linkage evidence observed was not influenced by the linkage analysis method chosen, as both MCLINK and MERLIN produced almost identical results. Recognizing the limitations of the data available, we present these results as proof of concept that the expression levels of several related genes can be combined to create a phenotype that can reasonably be used in linkage analysis. Such an approach could identify loci that regulate or contribute to disease pathways. More work is needed to refine and test the methodology, and more experimental data is needed to correlate tissue and lymphoblastoid expression levels, but the approach appears to have the potential to augment our current knowledge about the genetic basis of complex diseases.

Acknowledgements

The computational resources for this project have been provided by the National Institutes of Health (Grant # NCRR 1 S10 RR17214-01) on the Arches Metacluster, administered by the University of Utah Center for High Performance Computing. G. Bryce Christensen was funded via NIH NLM training grant T15

LM07124. In addition, the work was funded in part by NIH grant CA098364-01 (to NJC).

References

1. Morley M, Molony CM, Weber T, Devlin JL, Ewens KG, Spielman RS, Cheung VG: Genetic analysis of genome-wide variation in human gene expression. *Nature* 2004, 430:743-747.

2. Pennacchio L, Rubin E: Genomic strategies to identify mammalian regulatory sequences. *Nat Rev Genet* 2001, 2:100-109.

3. Schaid D: The Complex Genetic Epidemiology of Prostate Cancer. *Human Molecular Genetics* 2004, 13:R103-121.

4. Banez LL, Prasanna P, Sun L, Ali A, Zou Z, Adam BL, McLeod DG, Moul JW, Srivastava S: Diagnostic potential of serum proteomic patterns in prostate cancer. *J Urol* 2003, 170:442-446.

5. Ornstein DK, Rayford W, Fusaro VA, Conrads TP, Ross SJ, Hitt BA, Wiggins WW, Veenstra TD, Liotta LA, Petricoin EF: Serum proteomic profiling can discriminate prostate cancer from benign prostates in men with total prostate specific antigen levels between 2.5 and 15.0 ng/ml. *J Urol* 2004, 172:1302-1305.

6. Ashida S, Nakagawa H, Katagiri T, Furihata M, Iiizumi M, Anazawa Y, Tsunoda T, Takata R, Kasahara K, Miki T, et al: Molecular features of the transition from prostatic intraepithelial neoplasia (PIN) to prostate cancer: genome-wide gene-expression profiles of prostate cancers and PINs. *Cancer Res* 2004, 64:5963-5972.

7. Thomas A, Gutin A, Abkevich V, Bansal A: Multipoint linkage analysis by blocked Gibbs sampling. *Statistics and Computing* 2000, 10:259-269.

8. Smith JR, Freije D, Carpten JD, Gronberg H, Xu J, Isaacs SD, Brownstein MJ, Bova GS, Guo H, Bujnovszky P, et al: Major susceptibility locus for prostate cancer on chromosome 1 suggested by a genome-wide search. *Science* 1996, 274:1371-1374.

9. Sung Y, Di Y, Fu A, Rothstein J, Sieh W, Tong L, Thompson E, Wijsman E: Comparison of multipoint linkage analyses for quantitative traits: parametric lod scores, variance components lod scores and Bayes factors in the CEPH data. *GAW15 Proceedings* 2007, Submitted.

10. Sellick GS, Webb EL, Allinson R, Matutes E, Dyer M, Jonsson V, Langerak AW, Mauro FR, Fuller S, Wiley J, et al: A High-Density SNP Genomewide Linkage Scan for Chronic Lymphocytic LeukemiaSusceptibility Loci. *Am J Hum Genet* 2005, 77:420-429.

11. Abecasis G, Cherny S, Cookson W, Cardon L: Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* 2002, 30:97-101.

12. Lander E, Kruglyak L: Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* 1995, 11:241-247.

13. Slager S, Zarfas KE, Brown WM, Lange E, McDonnell S, Wojno KJ, Cooney KA: Genome-wide linkage scan for prostate cancer aggressiveness loci using families from the University of Michigan Prostate Cancer Genetics Project. *Prostate* 2006, 66:173-179.

14. Stanford J, McDonnell S, Friedrichsen D, Carlson E, Kolb S, Deutsch K, Janer M, Hood L, Ostrander E, Schaid D: Prostate cancer and genetic susceptibility: a genome scan incorporating disease aggressiveness. *Prostate* 2006, 66:317-325.

15. Suarez BK, Lin J, Burmester JK, Broman KW, Weber JL, Banerjee TK, Goddard KA, Witte JS, Elston RC, Catalona WJ: A genome screen of multiplex sibships with prostate cancer. *Am J Hum Genet* 2000, 66:933-944.

16. Xu J, Gillanders EM, Isaacs SD, Chang BL, Wiley KE, Zheng SL, Jones M, Gildea D, Riedesel E, Albertus J, et al: Genome-wide scan for prostate cancer susceptibility genes in the Johns Hopkins hereditary prostate cancer families. *Prostate* 2003, 57:320-325.

CHAPTER 4

THE sumLINK STATISTIC FOR GENETIC LINKAGE ANALYSIS IN THE PRESENCE OF HETEROGENEITY

Gerald Bryce Christensen, Stacey Knight, Nicola J. Camp

Published: (2009) *Genetic Epidemiology*

Abstract

We present the "sumLINK" statistic—the sum of multipoint LOD scores for the subset of pedigrees with nominally significant linkage evidence at a given locus—as an alternative to common methods to identify susceptibility loci in the presence of heterogeneity. We also suggest the "sumLOD" statistic (the sum of positive multipoint LOD scores) as a companion to the sumLINK. SumLINK analysis identifies genetic regions of extreme consistency across pedigrees without regard to negative evidence from unlinked or uninformative pedigrees. Significance is determined by an innovative permutation procedure based on genome shuffling that randomizes linkage information across pedigrees. This procedure for generating the empirical null distribution may be useful for other linkage-based statistics as well. Using 500 genome-wide analyses of simulated null data, we show that the genome shuffling procedure results in the correct type 1 error rates for both the sumLINK and sumLOD. The power of the statistics was tested using 100 sets of simulated genome-wide data from the alternative hypothesis from GAW13. Finally, we illustrate the statistics in an analysis of 190 aggressive prostate cancer pedigrees from the International Consortium for Prostate Cancer Genetics, where we identified a new susceptibility locus. We propose that the sumLINK and sumLOD are ideal for collaborative projects and meta-analyses, as they do not require any sharing of identifiable data between contributing institutions. Further, loci identified with the sumLINK have good potential for gene localization via statistical recombinant mapping, as, by definition, several linked pedigrees contribute to each peak.

Introduction

Genetic linkage analysis can be an effective tool for identifying disease susceptibility loci. However, locus heterogeneity can counter this effectiveness and is often acknowledged as the single largest issue that hinders the linkage analysis approach. Complex traits may be controlled by numerous genes and, therefore, statistics that attempt to model or recognize locus heterogeneity are required. The two common methods to address heterogeneity are the heterogeneity LOD statistic (HLOD), which statistically models the heterogeneity with an additional parameter, and phenotypic subset analysis. However, the former may fail to distinguish linked from unlinked pedigrees sufficiently to allow for substantial power increase and suffers from the lack of a precise distribution for assessing significance, and the latter requires a-priori determination of subsets. Beyond heterogeneity, localization also presents a challenge in linkage analysis. Often regions identified by linkage are large (perhaps 30-50 cM) and the boundaries ill-defined, both of which hinder follow-up studies. A method that can address locus heterogeneity and produce regions that are useful for localization would be an important addition to the tools already available.

The genetic research community has ascertained a great number of family-based resources for linkage analysis across numerous and varied complex traits. These data repositories represent a tremendous investment of time and resources, and likely contain a wealth of information–much of which has yet to be extracted. In the era of consortia efforts and with greater numbers of pedigrees available for specific diseases through collaborative efforts, new approaches and opportunities arise, especially to identify genes that may explain only a very small portion of disease that could not be identified in single studies. Here, we introduce a new approach to locus heterogeneity that focuses on individually powerful pedigrees–something that becomes possible in multicenter collaborative settings and other studies with large numbers of pedigrees. A standard HLOD analysis attempts to statistically separate linked and unlinked pedigrees through an additional parameter in the LOD calculation, α, the proportion of linked pedigrees; however, many pedigrees may be uninformative (pedigree-specific contributions surrounding 0) at a locus, and these pedigrees add statistical noise that reduces power. Our new approach uses a predefined LOD threshold to simply remove those families that are below the threshold and could be considered "noise." As such, it could be thought of as a "brute-force" approach to heterogeneity that attempts to gain power by removing noise from the statistical analysis. This method directly addresses the fact that only a small portion of the pedigrees in a data resource may be linked to any true causative locus, and in the process, identifies the informative set of families that are most useful for defining and fine mapping the locus. Several recent studies have used statistical recombinant mapping to delimit the boundaries of linkage regions [1-3]. Recombinant mapping requires that several pedigrees be linked to a region of interest, but it is unclear how many pedigrees should be linked to a locus for it to be considered a reasonable candidate for successful mapping. The sumLINK statistic can address this issue by assigning valid significance probabilities.

Our method focuses on individually powerful pedigrees that are nominally "linked" to a position in the genome and assesses whether the amount of concordance observed across the linked pedigrees at any point in the genome is more than would be expected by chance. Statistical excess of concordance is evidence for an underlying susceptibility locus. An advantage is that by the nature of the procedure, the regions of interest identified by the sumLINK statistic should have multiple pedigrees that can be used to delimit the region using statistical recombinant mapping. Further, in contrast to many other situations, the existence of different genetic marker sets (which often will occur in consortia) is not problematic and may, in fact, lead to some serendipitous pseudo-fine mapping. This method offers additional opportunities to identify disease susceptibility loci and the underlying genes using linkage-based data.

Methods

sumLINK and sumLOD

Our approach is to identify regions of the genome that display a significant excess of concordance across 'linked' pedigrees. The level of concordance is quantified by the sum of the pedigree-specific multipoint LOD scores in the identified linked pedigrees. We consider any pedigree that meets or exceeds a pedigree-specific LOD threshold of 0.588 ($p \leq 0.05$, "nominal" significant evidence) at a specific genomic position to be "linked" at that position. We have called this linkage-based statistic the "sumLINK," because it is simply the sum of multipoint LOD scores for linked pedigrees at a given point in the genome. Clearly, the distribution of the sumLINK statistic varies according to the number and structure of pedigrees in the initial resource and the parameters of the linkage model used to calculate the LOD scores. It is therefore difficult to determine the null distribution of the statistic theoretically; however, empirical methods can be employed to generate the null distribution for any data resource. The creation of the null distribution from which to test significance is outlined below.

To perform a sumLINK analysis, it is necessary that linkage results are available for each pedigree at regular intervals across the genome (Figure 4.1. A). This is possible in many standard linkage software packages that calculate multipoint LOD scores, including Merlin[4] and Genehunter-Plus [5]. The sumLINK statistic is then calculated at each position by summing the LOD scores for only those pedigrees that meet or exceed the threshold of 0.588 at each position in the genome. A simplistic example is shown in Figure 4.2. The null distribution of the sumLINK statistic must represent the chance consistency expected across linked pedigrees, matched for pedigree structure, information content, and linkage potential. We achieve this null scenario by using a genome shuffling technique. The shuffling procedure consists of a chromosome randomization step and a genome rotation step. The randomization step begins by randomizing the sequential order of chromosomes for each pedigree. Chromosomes are concatenated end-to-end in this random order to create a 'new' genome (Figure 4.1. B). The beginning and end of this new genome is connected to form a 'loop.' In the rotation step, a random position in the loop is chosen and the loop rotated such that this position becomes the new starting position and the loop is broken there. This is done for each pedigree separately, and because multipoint LODs were calculated at evenly spaced positions, these new shuffled genomes can again be aligned across pedigrees (Figure 4.1. C). A null sumLINK statistic can then be calculated at each position across the shuffled genomes. The procedure maintains the continuity and autocorrelation of marker data within chromosomes, but randomizes

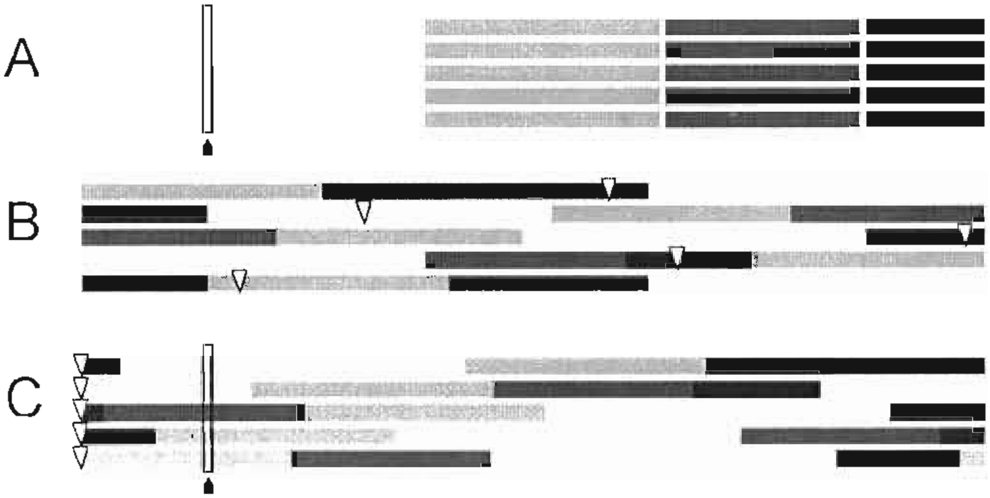


Figure 4.1. Shuffling procedure for creating null distribution. A) Raw test statistic is calculated across unshuffled data at regular intervals throughout the genome. Figure shows five pedigrees (rows) and four chromosomes (columns). B) Chromosomes are connected end-to-end in random order within each pedigree, and the resulting loop is broken at a random location to create a new starting point. C) The new starting points are aligned, and null statistics can be calculated along the shuffled genome.

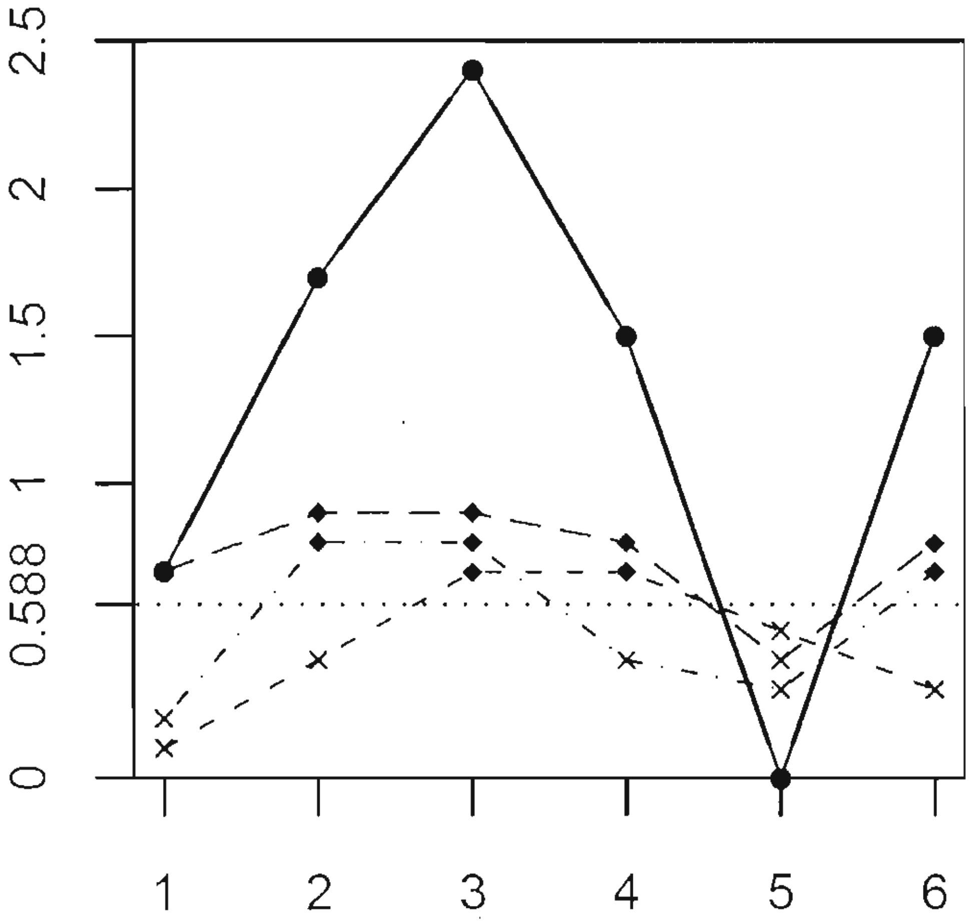


Figure 4.2. Simplistic illustration of the calculation of the sumLINK statistic. The linkage evidence for three pedigrees (broken lines) are shown across 6 loci. The sumLINK calculated from these three pedigrees in shown by the heavy black line. Pedigree LOD scores that are marked with a diamond meet the threshold for inclusion and these are summed to produce the sumLINK. Pedigree LOD scores marked with a cross do not meet the threshold for inclusion and do not contribute to the sumLINK.

consistency across pedigrees. The shuffling procedure is repeated a large number of times to determine the null distribution of the statistic for the given data.

Genome-wide significance is determined by the expected frequency of peaks of at least a certain magnitude occurring in the null sumLINK genome scans [6]. All peaks in each null genome are considered. In accordance with guidelines set by Lander and Kruglyak [7] for significance, we consider a peak height that is expected to occur with a frequency no greater than 0.05 per genome as genome-wide significant evidence for linkage, and a peak that occurs with an expected frequency of less than 1.00 per genome as genome-wide suggestive. It is important to note that a false positive rate (FPR) is not a p-value, it is a rate per genome and represents the expected frequency of peaks of at least the specified magnitude under null conditions. For example, FPR=0.6 indicates that a similar peak would be expected to occur 0.6 times per genome, which is sufficient evidence for suggestive linkage.

The advantage of the sumLINK is that regions are identified using individually powerful pedigrees, which is more intuitively appealing and perhaps convincing. Further, these independently powerful linked pedigrees can be used for localization using statistical recombinant mapping. In brief, statistical recombinant mapping uses pedigree-specific linkage evidence to estimate the positions of recombinant events on the linked segregating haplotypes, which can then be used to delimit the shared genomic region. Aligning these regions across all linked pedigrees localizes the region for further study. Another advantage of sumLINK is that it requires a minimal data set. It is not necessary to know pedigree structures or genotypes, which are required to obtain null statistics by permuting disease status [6], nor is it necessary that all pedigrees be genotyped with the same marker set. so long as the various marker sets are fitted to a common genetic map before calculating LOD scores. This property makes sumLINK ideal for multi-institutional collaborative research projects. A disadvantage is that the sumLINK will ignore some small pedigrees (e.g., sib-pairs) due to their relative lack of information content. It may therefore be attractive to additionally consider the sumLOD (sum of all positive pedigree LOD scores) as a companion to the sumLINK. The relaxed inclusion threshold of the sumLOD allows the potential for any minimally informative pedigree to contribute to the result. The sumLOD statistic is similar to the previously proposed C statistic [8], but utilizes multipoint, rather than two-point. inheritance information. The sumLOD has been used previously as a summary measure [9-11]. but has not been adopted as a test statistic due to the lack of a theoretical distribution. However. our genome shuffling procedure can be used to assess empirical significance of any statistic that is derived as a postprocessing step from pedigree-specific LOD values. including the sumLOD.

False Discovery Rate

Often for complex traits. no single significant findings are identified when using conservative family-wise multiple testing corrections and thresholds. It may therefore be useful to identify whether there exists a group of most significant findings that together indicates deviation from the null. The false discovery rate (FDR) evaluates this using the q-value. For example. if a q-value of 0.05 is assigned to the top 40 most significant findings, this indicates that 2 (0.05×40) are likely false positives, and that 38 are true-positives. In this way, a group can be identified within which true positive findings likely reside. Using our genome-shuffling method, it is possible to estimate empirical FDRs for the observed findings. In particular, this allows us to assess significance accounting for the multiple testing inherent in the multiple models and statistics.

Simulation Tests

Simulations Under the Null Hypothesis of No Linkage

We tested the sumLINK and sumLOD procedures in data simulated under null conditions in order to assess the validity of our genome shuffling procedure to generate the correct false positive rates. We created 400 two, three and four generation pedigrees typical of the families commonly used in linkage analysis. Each pedigree had a minimum of three affected subjects. Genome-wide genotypes were simulated using random gene-drops based on the genetic map and characteristics of 2,936 autosomal single nucleotide polymorphisms selected from the Illumina 6K SNP array to be free from linkage disequilibrium. This was repeated 250 times. Additional pedigree characteristics are summarized in Table 4.1. For each of the 250 replicates, multipoint parametric linkage statistics were calculated at 1cM intervals for both dominant and recessive inheritance models using Merlin [4], and results for each pedigree were extracted. Genome-wide sumLINK and sumLOD statistics were then computed for each replicate, and the empirical significance was assessed with 200 iterations of the genome shuffling procedure. Across the 250 replicates, the median number of pedigrees that contributed to the dominant sumLINK analysis was 153, and the median number that contributed to the recessive sumLINK analysis was 276.

Table 4.1. Simulated data characteristics

	Null Simulation (×250)	Selected GAW13 Simulation Pedigrees		
		All	**SumLINK set**	**SumLOD-only set**
Pedigrees	400	5232	1056	4176
Individuals	2673	44,326	13,568	30,758
Persons per ped	6.68 (4 to 16)	8.47 (4 to 25)	12.85 (6 to 25)	7.37 (4 to 22)
Mean generations	2.44 (2 to 4)	2.38 (2 to 4)	2.83 (2 to 4)	2.27 (2 to 4)
Total Affected	1627 (60.9%)	20,889 (47.1%)	6363 (46.9%)	14,526 (47.2 %)
Mean Affected	4.07 (3 to 11)	3.42 (2 to 12)	4.78 (2 to 12)	3.07 (2 to 11)
Total Genotyped	1627 (60.9%)	25808 (58.2%)	7590 (55.9%)	18218 (59.2%)
Mean Genotyped	4.07 (3 to 11)	4.93 (2 to 15)	7.19 (3 to 15)	4.36 (2 to 14)
Marker Type	SNP	STRP	STRP	STRP
Marker Number	2936	399	399	399

Simulations Under the Alternative Hypothesis of Linkage

We illustrate the power of the sumLINK and sumLOD statistics, in comparison to the more standard HLOD, by applying these to data based on the well-documented, simulated genome-wide data from Genetic Analysis Workshop 13 (GAW13) [12,13]. The GAW13 data were designed to represent random pedigrees (not ascertained for specific disease) and contained several simulated 'heart disease' traits. Fifty trait-related genes were simulated, most with common underlying susceptibility alleles and low effect sizes. There were 100 replicates of 330 simulated pedigree structures available, resulting in a total of 33,000 independent pedigrees. Genotypes were simulated for 399 microsatellite markers across the 22 autosomes. We chose to analyze an obesity trait as defined by [14] and sampled from the full set of pedigrees to better represent a linkage resource ascertained for disease. From the full set of 33,000 pedigrees, we extracted 5232 independent and minimally informative unilineal pedigrees (those with at least two genotyped subjects classified as obese). For each of the 5232 pedigrees, multipoint parametric linkage statistics were calculated at 1cM intervals for a simple dominant inheritance model using Merlin [4]. The pedigrees were divided into two groups based on whether they would be included in a sumLINK analysis (that is, that a minimum LOD score of 0.588 was observed at least once across the genome). Of the 5232, 1056 pedigrees were useful for sumLINK analysis; the remaining 4176 pedigrees were not suitable for sumLINK analysis but remained useful for sumLOD analysis. Pedigree characteristics for each group are summarized in Table 4.1. By sampling from the two groups of pedigrees, we created 100 replicates each containing 200 pedigrees (100 useful for sumLINK and 100 that were not); all sampling was performed with replacement. We then calculated genome-wide sumLOD and sumLINK statistics for each of the 100 replicates, with empirical significance determined by 200 repetitions of the genome shuffling procedure. HLODs were calculated with Merlin. Thresholds of 1.9 and 3.3 were used to determine suggestive and significant HLOD results.

Aggressive Prostate Cancer Case Study

We performed a sumLINK and sumLOD analysis on 190 pedigrees provided by the International Consortium for Prostate Cancer Genetics (ICPCG) [15] with clinically aggressive prostate cancer. A conventional linkage study of this resource and description of the data was published previously [16]. Dominant and recessive multipoint LOD scores were computed for each pedigree at 1-cM increments throughout the 22 autosomes by Genehunter-Plus using the models (dominant and recessive) as described by Schaid [15]. The sumLINK and sumLOD statistics were then calculated at each of the cM positions for both models, and empirical significance of the observed peaks was determined by 1000 repetitions of the genome shuffling procedure. Of the total 190 ICPCG pedigrees, 125 pedigrees achieved linkage evidence of at least 0.588 at some point in the genome with the dominant model and 127 for the recessive model. Hence, only these numbers of pedigrees contribute to the sumLINK analysis. All 190 pedigrees reached a LOD score greater than zero at least once in each inheritance model, allowing them all to contribute to the sumLOD analysis under both models.

Results

Simulation Tests

Simulations Under the Null Hypothesis of No Linkage

Table 4.2 illustrates the false positive rates observed under the dominant and recessive models for the sumLINK and sumLOD. All of these results fall within the 95% confidence interval based on 250 replicates and for a Poisson process with rates of 0.05 or 1.0, and illustrate that the genome shuffle procedure to determine significance is valid.

Table 4.2. False positive rates estimated from 250 genome-wide replicates under the null hypothesis of no linkage

Statistic	**Analysis Model**	**Statistical FPR threshold surpassed**	
		Significant (0.05)*	**Suggestive (1.00)†**
sumLINK	**Dominant**	0.056	0.936
	Recessive	0.060	1.124
sumLOD	**Dominant**	0.052	1.020
	Recessive	0.052	0.912
OVERALL		0.055	0.998

*95% CI for 0.05 in 250 replicates is [0.024.0.080]
†95% CI for 1.00 in 250 replicates is [0.876.1.124]

Simulations Under the Alternative Hypothesis of Linkage

Results of the power testing are summarized in Table 4.3. All genes that were identified with suggestive evidence at least five times by any one of the three statistics (HLOD, sumLINK, or sumLOD) are summarized. The simulated data included only one gene, *Gb11*, which affected baseline weight with a reasonably large effect size. This gene was identified with excellent power with all three statistics (all ≥99% power). Of the remaining genes, all had very common susceptibility alleles (minor allele frequencies ≥0.15) and low effect sizes. Subsequently, none were identified particularly well. However, it is interesting to note that of these lower effect size genes that were identified at least 5 times out of the 100 replicates, the sumLOD and/or sumLINK were always superior, and exhibited significantly more power than the HLOD at eight of the 10 loci. There were no genes that were significantly better identified with the HLOD.

Table 4.3. Power to detect at least suggestive linkage evidence in 100 simulations. All loci detected at least 5 times by any of the statistics are shown.

		Power (%)		
Gene	**Affected Trait**	**sumLINK**	**sumLOD**	**HLOD**
Gb11.Gb4*	Weight-baseline. Height-baseline	99	100	100
Gb2	Height-baseline	17†	22†	3
Gb20	HDL-baseline	9	14†	4
Gb22,Gs3**	HDL. Triglycerides	6†	8†	1
Gb15.Gs10.	HDL. Triglycerides. Glucose.	7†	7†	1
Gs12.Gb37**	SBP. DBP			
Gs4	Triglycerides, Glucose.	4	7†	1
Gs2	Weight-slope	6	5	3
Gs8	Cholesterol-slope	5	6†	1
Gb24	Triglycerides-baseline	2	6†	0
Gb5	Height-baseline	1	5†	0

Gene characteristics can be found in [Daw, et al. 2003].

*Gb4 was observed independently several times, but was often obscured by the broad peak at Gb11.

**Genes were positioned too closely on the chromosome to discern which was responsible for the linkage signal;

† significantly greater power than the HLOD. None of the differences observed between the sumLINK and sumLOD are statistically significant.

Aggressive Prostate Cancer Case Study

In our real data aggressive prostate cancer case study example, the sumLINK and sumLOD analyses identified significant linkage evidence at two loci (chromosomes 20q and 11q) and suggestive evidence at a third locus (chromosome 2), as shown in Table 4.4. The peak on chromosome 20 was significant under the dominant inheritance model for both the sumLINK (sumLINK = 13.848, number of linked pedigrees = 17, expected false positive rate (FPR) =0.005) and the sumLOD (sumLOD = 30.311, number of positive pedigrees = 83, FPR=0.028). The peak on chromosome 11 was significant in the recessive sumLOD analysis (FPR=0.007), with suggestive evidence in other analyses. The sumLINK analysis also identified suggestive linkage evidence on chromosome 2 under both dominant and recessive models (FPR = 0.628 and 0.897, respectively). Figure 4.3 shows the genome-wide sumLINK results for the dominant model and sumLOD results for the recessive model.

In an attempt to consider the multiple testing inherent from performing both the sumLINK and the sumLOD, both for dominant and recessive models, we considered the false discovery rates. Each centimorgan position in the genome search data (N=3502) was considered as an individual observation and p-values were calculated for every position based on the respective empirical distribution for each of the analyses. When the results of all four analyses are pooled, the top 54 ranked cM positions collectively attained an FDR of 0.1. An FDR of 0.1 indicates that the expected ratio of false:true positives is 1:9. That is, one tenth of these 54 (or, 5-6 positions) are likely from the null (false positives), but the remaining are likely true

Table 4.4. Summary of significant and suggestive linkage peaks

sumLINK Model		cM Position	sumLINK	Expected Frequency (FPR)	Number of contributing pedigrees	FDR q-value
Dominant	20	59	13.848	0.005	17	0.017
	2	69	10.837	0.628	13	0.175
Recessive	11	89	10.941	0.598	14	0.128
	2	68	10.624	0.897	14	0.209
sumLOD Model	**Chrom.**	**cM Position**	**sumLOD**	**Expected Frequency (FPR)**	**Number of contributing pedigrees**	**FDR q-value**
Recessive	11	89	27.975	0.007	87	0.017
Dominant	20	59	30.311	0.028	83	0.017
	11	89	27.650	0.657	81	0.213

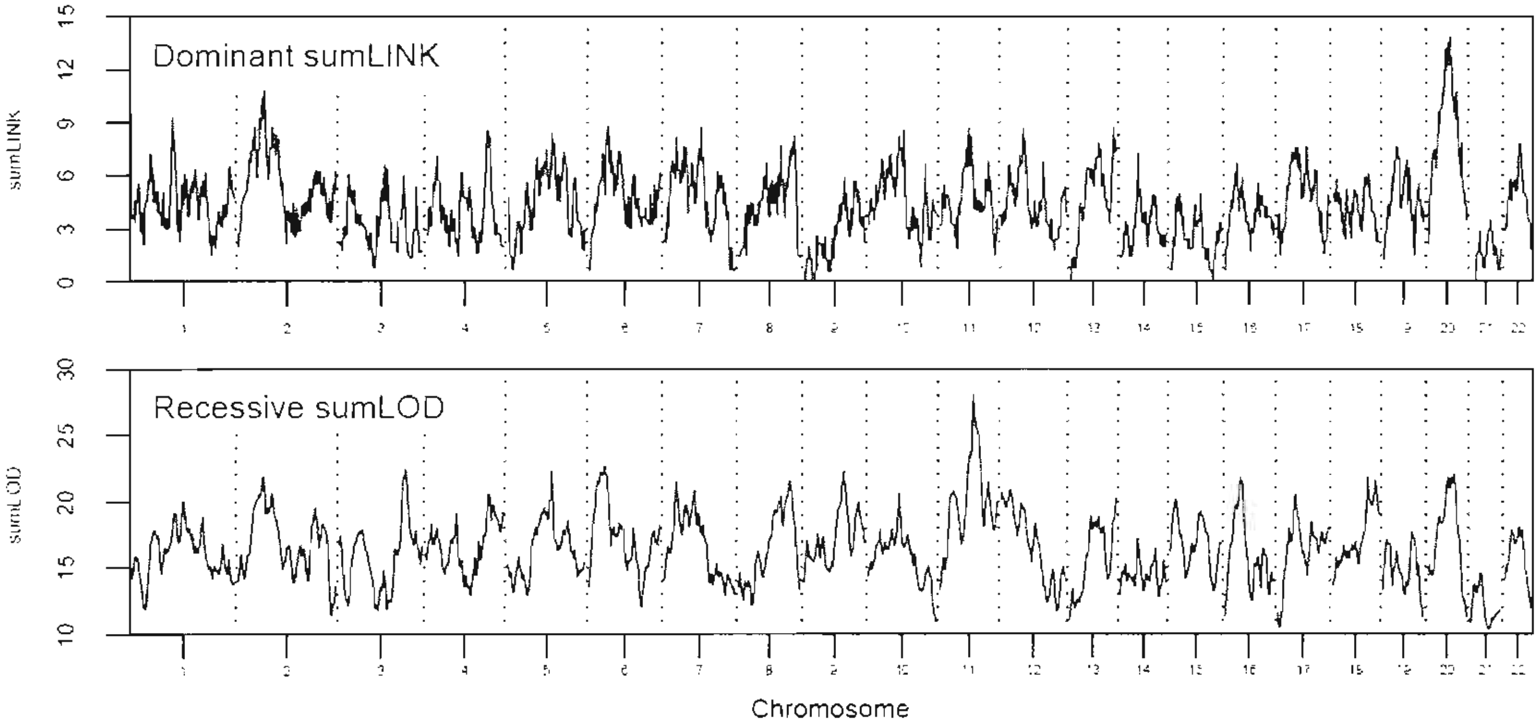


Figure 4.3. Genome-wide multipoint sumLINK results (dominant model) and sumLOD results (recessive model) for the ICPCG aggressive prostate cancer data.

positives. FDR will not differentiate which are which; however, in this case example, all 54 positions fall under one of the significant linkage peaks (19, 16, and 19 positions on chromosomes 20 (sumLINK), 20 (sumLOD), and 11 (sumLOD), respectively). Hence, even if all 5-6 false positive findings were from one region, it is still expected to have a true positive in each. In conclusion, the FDR suggests that the linkage peaks on chromosomes 11 and 20 are likely true positive findings after correction for multiple testing.

We applied statistical recombinant mapping to all three regions with at least suggestive genome-wide evidence to delimit the regions of interest. The genotypes used in this multicenter collaborative analysis were derived from several diverse sets of microsatellite markers, generally with an average spacing of 10 cM. On average, therefore, most pedigrees have a genotyped marker within 5 cM of any given cM position on the genetic map. Pedigrees were therefore included in a localization analysis if they achieved LOD ≥ 0.588 within 5 cM of the observed peak. Figure 4.4 illustrates the by pedigree LOD tracings used in the recombinant mapping for the three regions of interest. Recombinant events are estimated to be at the outermost point of a sharp decline in LOD score, as these positions indicate statistical evidence for a loss of genetic sharing. This point is a conservative estimate for the outer limit of the region where a susceptibility variant may be found. A region bounded by two recombinant events on each side represents an approximate 95% confidence interval for the

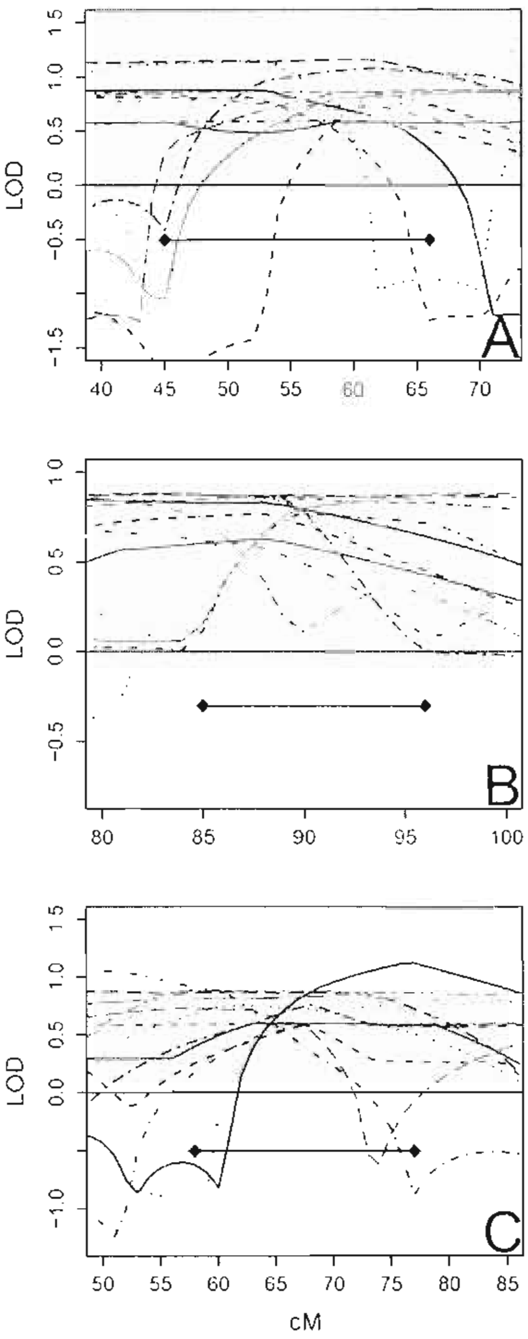


Figure 4.4. LOD traces for each pedigree contributing to the linkage results on chromosomes A) 20-dominant, B) 11-recessive, and C) 2-dominant. Black bars indicate the two-recombinant localization regions.

consensus region [1]. As seen in Figure 4.4, the linkage peaks on Chromosomes 20, 11, and 2 can each be conservatively localized to regions of 21, 21, and 19 cM, respectively.

Discussion

The sumLINK statistic is a new method aimed at addressing both heterogeneity and localization. The procedure is designed to identify the genomic regions for which an excessive number of powerful pedigrees are concordant. It is an ideal approach for multicenter collaborations or large single-site studies where a large number of pedigrees are available. A distinct advantage of this method is that it does not require collaborating centers to share raw data such as pedigree structures or genotypes, and does not require that each center use the same marker set. Provided a common genetic map is used for analysis, each center can perform their own analyses, calculating multipoint LOD scores at the same equally-spaced increment across the genome. It is only necessary to share these meta data (a multipoint genome scan for each pedigree), which enhances data privacy and security.

An important advantage of the sumLINK is the ability to identify loci that have good potential for gene localization, as several linked pedigrees exist beneath each peak identified. An unexpected benefit of compiling data across centers that used different marker maps is that the resolution of the localization can be higher than any of the individual genetic maps due to the overlaying of data. In our example, even with a low density 10 cM marker map, we were able to localize each region to approximately 20 cM, and these localized regions would be greatly refined with fine-mapping. This method of using the limits of sharing observed within extended pedigrees is intuitively appealing for localization, but may also have theoretical advantages over other common methods. Often, so-called "1-LOD" support intervals are reported for linkage peaks generated from a HLOD analysis; however, support intervals should strictly be applied to parameter estimates (the recombination fraction parameter, θ, in the case of linkage statistics) and are relevant in the context of two-point maxLOD statistics that are directly analogous to likelihood ratio tests. The standard practice of a 1-LOD support interval using the value of the statistic itself (usually HLOD) rather than a parameter is not statistically well-grounded, although since θ is a distance parameter, it has intuitive appeal. In particular, in a HLOD analysis, it is not clear whether the statistical noise generated by "unlinked" pedigrees may mask or shift positive linkage evidence. Hence, these "1-LOD" intervals can only be considered as a rough guide.

The shuffling method we have implemented to determine the null distribution is a particularly innovative element of the sumLINK procedure, and may be especially useful to the broader research community. We used the procedure to assess the significance of two genome-wide linkage statistics (sumLOD and sumLINK), but it may have broader applications for testing the significance of other statistics with unknown distributions. It is a simple, elegant, and quick way to create null data for assessing significance. It accounts for variations in pedigree structure as well as the autocorrelation of consecutive loci inherent in genetic linkage data. We developed a postprocessing script written in R [17] that calculates the sumLINK and sumLOD statistics, performs the genome-shuffling, generates the empirical null distribution, and tests the significance of observed linkage peaks. Computational time is dependent upon the number of pedigrees and the length of the genomic region being analyzed. The ICPCG data, comprised of 190 pedigrees and 3502 data points from 22 chromosomes. required 21.3 seconds per iteration with a 3.0 GHz Intel Xeon Duo Core 64-bit CPU running R v2.4.1 on Red Hat Enterprise Linux v5. One iteration consists of shuffling all pedigrees, calculating the null sumLINK, sumLOD. and number of pedigrees contributing to each statistic at all data points, and writing out a text file containing these values. Significance is computed in a later step after all shuffling iterations are complete. Our simulated null data (400 pedigrees, 3550 cM. 22 chromosomes) required 60.8 seconds per iteration. and the simulated data sets from GAW13 (200 pedigrees, 3604 cM. 22 chromosomes) required 22.9 seconds per iteration.

Analysis of simulated null data illustrated that the type-I error rate for the sumLINK and sumLOD statistics were all within acceptable boundaries. indicating that the genome shuffling procedure is valid for significance testing. It is interesting to note that the sumLINK and sumLOD statistics did not frequently agree with regard to the locations of statistically significant peaks in the null data, nor did they generally agree with the HLOD. This perhaps indicates that the three statistics are sensitive to different characteristics of the null data.

Analysis of simulated alternative hypothesis data was based on a GAW13 complex model. An obesity phenotype was selected because it is a complex trait simulated with extensive locus heterogeneity. One major weight gene, *Gb11* was easily identified, with both the sumLOD and sumLINK showing good comparability with the HLOD. Power was low for all other genes, but this was not unexpected. Others who analyzed these data reported that the simulated obesity-related genes, particularly those genes affecting change over time, were very difficult to find [18,19]. The data creators intentionally made many of the genes challenging and perhaps even impossible to find [13]. Although the power was low, the sumLINK and sumLOD statistics consistently outperformed the HLOD in identifying the minor genes. However, we do not believe that these new statistics should replace the HLOD; rather, our investigation indicates a proof-of-principle that the sumLINK and/or sumLOD are useful companion measures to help identify the best loci for further testing.

Potential limitations of our method include that the genome-shuffling procedure to create the null distribution may not be useful for studies including only a small number of pedigrees due to the limited number of shuffled genomes that can be generated. The shuffling procedure also assumes that information content is approximately constant across the genome, an assumption that may be violated at the telomeres where multipoint information and information content is reduced systematically. We tested robustness to this by removing all the telomeric regions from the ICPCG data and repeating the analysis. We found that because these regions are such a small part of the entire genome, they do not substantially bias the shuffled null genomes and the results were extremely robust. However, given the difference in information content between the sex chromosomes and the autosomes, we suggest the method for autosomal genome scans only. The term "genome-wide" as used in this manuscript refers only to the 22 autosomes. All of the sumLINK and sumLOD analyses we presented were performed using sex-averaged genetic maps. The effect of this assumption on the characteristics of these new statistics has not been investigated here..

In our example case study of the ICPCG aggressive prostate cancer data, we identified 3 regions of interest for further follow-up: two with genome-wide significant evidence supported by FDR analysis, and one with suggestive evidence. This performance is very encouraging. A prior linkage study of these data using conventional LOD/HLOD procedures indicated suggestive linkage evidence at the same loci that we identified on chromosomes 11 and 20 (HLODs of 2.40 for a recessive model and 2.49 for a dominant model, respectively) [16]. Our method finds superior levels of significance; both loci are genome-wide significant. However, it is certainly notable that Schaid et al. reported that the evidence on chromosome 11 increased to HLOD=3.31 in subset analyses for early age-at-onset pedigrees, and the region on chromosome 20 increased to HLOD= 2.65 in the subset of pedigrees with mean age-at-onset greater than 65 years [16]. Without necessitating the increased multiple testing inherent from subset analyses, the sumLINK was able to identify the more powerful pedigrees and the superior evidence. Our suggestive region on chromosome 2 was not identified using conventional linkage statistics in the previous study.

Conclusion

We have proposed a new statistic to identify linkage regions that have promise for localization and follow-up to gene identification. An R-script is available from the authors that can be used to calculate the sumLINK and sumLOD statistics and generate the null distributions to assess significance of each. We do not claim that these statistics are superior, but that there is evidence that they are useful companion statistics to the HLOD. This method is of particular use within the framework of large collaborative data as it requires neither the sharing of raw data nor the use of common marker sets. We believe this is an important additional statistical tool for identifying linkage regions likely to harbor disease predisposition genes.

Acknowledgements

The computational resources for this project have been provided by the National Institutes of Health (Grant # NCRR 1 S10 RR17214-01) on the Arches Metacluster, administered by the University of Utah Center for High Performance Computing. The International Consortium for Prostate Cancer Genetics is funded by NCI CA89600 (to William B. Isaacs, Head ICPCG). The Genetic Analysis Workshop is funded by R01 GM031575. This work was also supported by NCI CA98364 (to Nicola J. Camp). Bryce Christensen was supported by National Library of Medicine training grant NLM T15 LM07124.

References

1. Camp NJ, Farnham JM, Cannon-Albright LA: Localization of a prostate cancer predisposition gene to an 880-kb region on chromosome 22q12.3 in Utah high-risk pedigrees. *Cancer Research* 2006, 66:10205-10212.

2. Camp NJ, Farnham JM, Allen-Brady K, Cannon-Albright LA: Statistical recombinant mapping in extended high-risk Utah pedigrees narrows the 8q24 prostate cancer locus to 2.0 Mb. *Prostate* 2007, 67:1456-1464.

3. Johanneson B, McDonnell SK, Karyadi DM, Hebbring S, Wang L, Deutsch K, McIntosh L, Kwon EM, Suuriniemi M, Stanford JL, et al: Fine mapping of

familial prostate cancer families narrows the interval for a susceptibility locus on chromosome 22q12.3 to 1.36 Mb. *Hum Genet* 2008, 123:65-75.

4. Abecasis G, Cherny S, Cookson W, Cardon L: Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* 2002, 30:97-101.

5. Kong A, Cox NJ: Allele-sharing models: LOD scores and accurate linkage tests. *Am J Hum Genet* 1997, 61:1179-1188.

6. Chen L, Storey JD: Relaxed Significance Criteria for Linkage Analysis. *Genetics* 2006, 173:2371-2381.

7. Lander E, Kruglyak L: Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* 1995, 11:241-247.

8. MacLean CJ, Ploughman LM, Diehl SR, Kendler KS: A new test for linkage in the presence of locus heterogeneity. *Am J Hum Genet* 1992, 50:1259-1266.

9. Camp NJ, Hopkins PN, Hasstedt SJ, Coon H, Malhotra A, Cawthon RM, Hunt SC: Genome-Wide Multipoint Parametric Linkage Analysis of Pulse Pressure in Large, Extended Utah Pedigrees *Hypertension* 2003, 43:322-328.

10. Horne BD, Malhotra A, Camp NJ: Comparison of linkage analasys methods for genome-wide scanning of extended pedigrees, with application to the TG/HDL-C ratio in the Framingham Heart Study. *BMC Genetics* 2003, 4:S93.

11. Orr A, Dubé M, Marcadier J, Jiang H, Federico A, George S, Seamone C, Andrews D, Dubord P, Holland S, et al: Mutations in the UBIAD1 Gene, encoding a potential prenyltransferase, are causal for Schnyder Crystalline Corneal Dystrophy. *PLoS ONE* 2007, 2:e685.

12. Almasy L, Amos CI, Bailey-Wilson JE, Cantor RM, Jaquish CE, Martinez M, Neuman RJ, Olson JM, Palmer LJ, Rich SM, et al: Genetic Anlaysis Workshop 13: Analysis of Longitudinal Family Data for Complex Diseases and Related Risk Factors. *BMC Genetics* 2003, 4:S1.

13. Daw EW, Morrison J, Zhou X, Thomas DC: Genetic Analysis Workshop 13: Simulated longitudinal data on families for a system of oligogenic traits. *BMC Genetics* 2003, 4:S3.

14. Klein AP, Kovac I, Sorant AJM, A. B-B, Doan BQ, Ibay G, Lockwood E, Mandal D, Santhosh L, Weissbecker K, et al: Importance sampling method of correction for multiple testing in affected sib-pair linkage analysis. *BMC Genetics* 2003, 4:S73.

15. Schaid DJ, Chang BL: Description of the international consortium for prostate cancer genetics, and failure to replicate linkage of hereditary prostate cancer to 20q13. *Prostate* 2004.

16. Schaid D, ICPCG: Pooled genome linkage scan of aggressive prostate cancer: Results from the International Consortium for Prostate Cancer Genetics. *Hum Genet* 2006, 120(4):471-485.

17. R Development Core Team: *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2006.

CHAPTER 5

GENOME-WIDE LINKAGE ANALYSIS OF 1233 PROSTATE CANCER PEDIGREES FROM THE INTERNATIONAL CONSORTIUM FOR PROSTATE CANCER GENETICS USING NOVEL sumLINK AND sumLOD ANALYSES

G. Bryce Christensen, Lisa A. Cannon-Albright, Nicola J. Camp and the International Consortium for Prostate Cancer Genetics.

For submission to: *The Prostate*

Abstract

BACKGROUND: Prostate cancer is generally believed to have a strong inherited component, but the search for susceptibility genes has been hindered by the effects of genetic heterogeneity. The recently developed sumLINK and sumLOD statistics are powerful tools for linkage analysis in the presence of heterogeneity. METHODS: We performed a secondary analysis of 1233 prostate cancer pedigrees from the International Consortium for Prostate Cancer Genetics (ICPCG) using two novel statistics, the sumLINK and sumLOD. For both statistics, dominant and recessive genetic models were considered. False discovery rate (FDR) analysis was conducted to assess the effects of multiple testing. RESULTS: Our analysis identified significant linkage evidence at chromosome 22q12, confirming previous findings by the initial conventional analyses of the same ICPCG data. Twelve other regions were identified with genomewide suggestive evidence for linkage. Seven regions (1q23, 5q11, 5q35, 6p21, 8q12, 11q13, 20p11-q11) are near loci previously identified in the initial ICPCG pooled data analysis or the subset of aggressive prostate cancer (PC) pedigrees. Three other regions (1p12, 8p23, 19q13) confirm loci reported by others, and two (2p24, 6q27) are novel susceptibility loci. FDR testing indicates that over 70% of these results are likely true positive findings. Statistical recombinant mapping narrowed regions to an average of 9 cM. CONCLUSIONS: Our results represent genomic regions with the greatest consistency of positive linkage evidence across a very large collection of high-risk prostate cancer pedigrees using new statistical tests that deal powerfully with heterogeneity. These regions are excellent candidates for further study to identify prostate cancer predisposition genes.

Introduction

Prostate cancer (PC) is believed to have a complex environmental and genetic etiology potentially involving numerous genes [1]. The identification of PCa genes has proven to be very difficult: genetic heterogeneity is a major issue that hinders progress [2]. Confirmations of reported PC susceptibility loci are infrequent and some of the loci that have been confirmed by multiple researchers are in chromosomal regions with very few promising candidate genes [3,4]. Luo and Yu reported in 2003 that evidence for PC susceptibility variants had been reported on all but two human chromosomes [5]. These two remaining chromosomes, 21 [6,7] and 22 [8,9], have subsequently both been implicated. The International Consortium for Prostate Cancer Genetics (ICPCG) was formed by a large and diverse group of researchers who have pooled their resources with the intent of deciphering the principal genetic factors underlying this pervasive disease [10]. The ICPCG published the findings of a conventional linkage analysis using the well-known heterogeneity LOD (HLOD) statistic and multiple subset analyses based on 1233 high-risk prostate cancer pedigrees. The study identified several susceptibility loci for further study [8].

Here we present the results of a secondary analysis of the ICPCG pooled pedigree resource using new genome-wide linkage-based statistics, the sumLINK and sumLOD, to identify PC susceptibility loci. These new statistics have been shown in simulation studies to be powerful and robust tools for identifying susceptibility loci in the presence of genetic heterogeneity [11]. The sumLINK/sumLOD approach is well-suited to analysis of pooled data resources such as this, because it requires only summary data from each constituent group, which is logistically easier to attain (there are often data privacy and confidentiality concerns associated with sharing individual raw genotype data and pedigree structures). Secondary analyses of existing data that are more powerful at addressing genetic heterogeneity have the potential to refine the original analyses, and identify additional evidence for PC predispostion genes.

Methods

The sumLINK approach focuses on 'linked' pedigrees, which we define to be a pedigree-specific LOD$\geq$0.588 (p$\leq$0.05). The aim is to identify regions with extreme consistency of linkage evidence across pedigrees. The sumLINK statistic is the sum of multipoint LOD scores for all pedigrees that meet the threshold of LOD$\geq$0.588 at a given point in the genome. This value is computed at intervals of one centimorgan throughout the genome. We assess the significance of the sumLINK empirically using a unique genome randomization and shuffling method that simulates the expected consistency of linked pedigrees under null conditions [11]. Briefly, for each pedigree, the vectors of LOD scores for each chromosome are connected in random order, with the first and last values connected to form a 'loop,' and the loop is broken at a random position to create a randomized, shuffled 'genomewide' vector of LOD scores. These vectors are then aligned across pedigrees and values of the sumLINK statistic are calculated. This procedure is designed to maintain each pedigree's potential for linkage signals across the genome, but randomizes consistency of linkage evidence across pedigrees. Observed peaks are compared with peaks occurring in 1000 iterations of the randomized data in order to establish the expected frequency of peaks with a similar or greater magnitude for the data in question. This expected frequency may be called a false positive rate, or FPR.

The sumLOD statistic is a complimentary companion to the sumLINK statistic. The sumLOD statistic is similar to the sumLINK statistic, but with a reduced inclusion threshold; all positive pedigree LOD scores at each point in the genome are summed to calculate the sumLOD statistic. Significance of the sumLOD is determined empirically by the same genome randomization procedure that is used for the sumLINK. In accordance with the standards for significant linkage evidence set by Lander and Kruglyak [12], peak sumLINK and sumLOD values are considered to represent significant evidence of linkage if the expected frequency of peaks of similar magnitude under null conditions is less than 0.05 per genome. Peak values are considered to be suggestive evidence of linkage if the expected frequency is less than one per genome.

We applied the sumLINK and sumLOD procedures to the 1233 PC pedigrees in the ICPCG pooled pedigree resource. Pedigree characteristics and genotyping details have been described previously [8]. The two statistics were computed at 1-cM increments (N=3502) throughout the 22 autosomes based on LOD scores from the dominant and recessive inheritance models that were used in the original ICPCG analysis. The sex chromosomes were not included in the analysis. 572 pedigrees achieved a maximum LOD score of at least 0.588 at some point in the genome under the dominant inheritance model, and 533 pedigrees achieved a LOD score of at least 0.588 under the recessive model. Only these pedigrees contributed to the sumLINK analyses. 1230 pedigrees contributed to the sumLOD analyses under each model.

Empirical significance was computed based on 1000 iterations of the genome randomization technique.

False positive rates were calculated based on the empirical distributions for each of the four analyses (dominant and recessive, sumLINK and sumLOD). False discovery rate (FDR) q-values were estimated to account for the effects of multiple testing that are inherent in the usage of multiple models and statistics. Application of FDR methods to multipoint LOD scores have been shown to be valid provided no fine-mapping markers are used [13]. This requirement is met in the present analysis. The empirical FDR q-value represents the probability that a given result is a false positive based on the pooled distributions of all four analyses.

Localization

Loci identified with the sumLINK approach have natural potential for subsequent gene localization using statistical recombinant mapping [14], as, by definition, there exist a statistical excess of linked pedigrees contributing to each peak. Hence, for all significant and suggestive sumLINK peaks, we will pursue localization using statistical recombinant mapping. The genetic marker sets for which pedigrees were genotyped varied between institutions. Even though the resolution of each separate linkage study map was an average spacing of 10 cM, the disparity of different marker maps helps fine-mapping efforts. If pedigrees from different resources are linked to the same region, they can identify regions smaller than the resolutions of each independent marker map. These genomic segments are the most probable locations for finding a PC susceptibility gene.

Given that the linkage evidence for each pedigree is based on a 10 cM map, most pedigrees will have a genotyped marker within 5 cM of any given cM position on the genetic map. Hence, when selecting pedigrees to consider 'linked' to a significant or suggestive region, we identified all pedigrees that achieved LOD≥0.588 within 5 cM of the observed sumLINK peak. We then examined the LOD score curves for each of these pedigrees and determined the probable location of recombination events that mark the outer limits of the segregating chromosomal segment within each pedigree. Recombinant events are estimated to be at the outer point of an abrupt drop in LOD score, as these positions are statistical evidence for a loss of genetic sharing by affected pedigree members. A shared chromosomal region bounded by two recombinant events on each side is an approximate 95% confidence interval for the consensus region [14].

Results

Figure 5.1 shows the genome-wide sumLINK and sumLOD statistics for each model, together with lines representing the thresholds for significant and suggestive linkage as determined by the randomization procedure. Results are summarized in Table 5.1. We identified one locus with significant linkage evidence, and twelve loci with suggestive linkage evidence. There were no significant or suggestive linkage peaks identified by the recessive sumLINK analysis.

Significant linkage evidence was observed at chromosome 22q12 by both the dominant sumLINK (FPR=0.010, 46 contributing pedigrees) and the dominant sumLOD (FPR=0.032, 454 contributing pedigrees). In addition to both of these

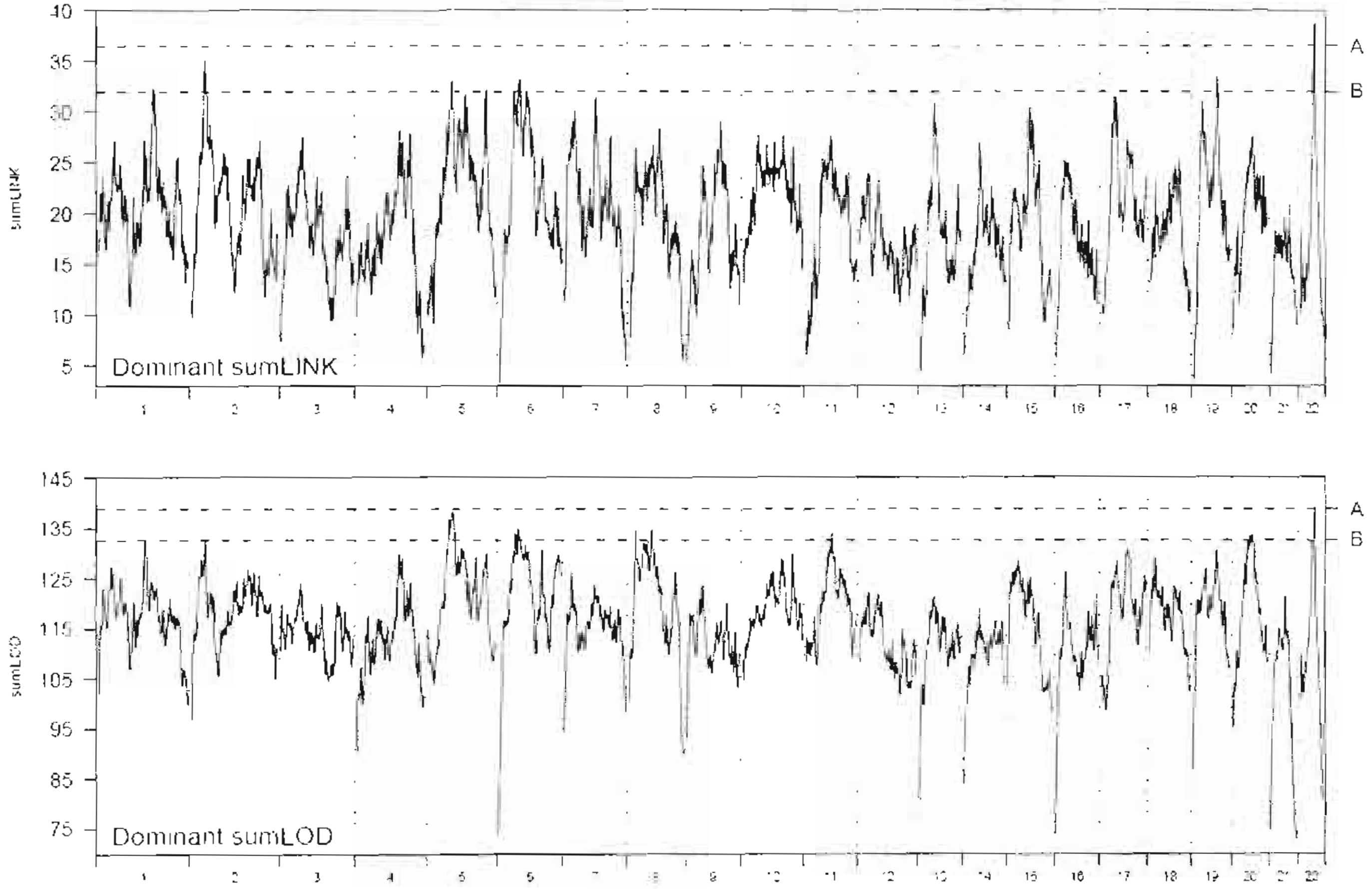


Figure 5.1. Genome-wide sumLINK and sumLOD values for dominant and recessive inheritance models. The line marked "A" in each figure represents the threshold for significant linkage evidence determined by the genome shuffling process. The line marked "B" shows the threshold for suggestive linkage evidence.

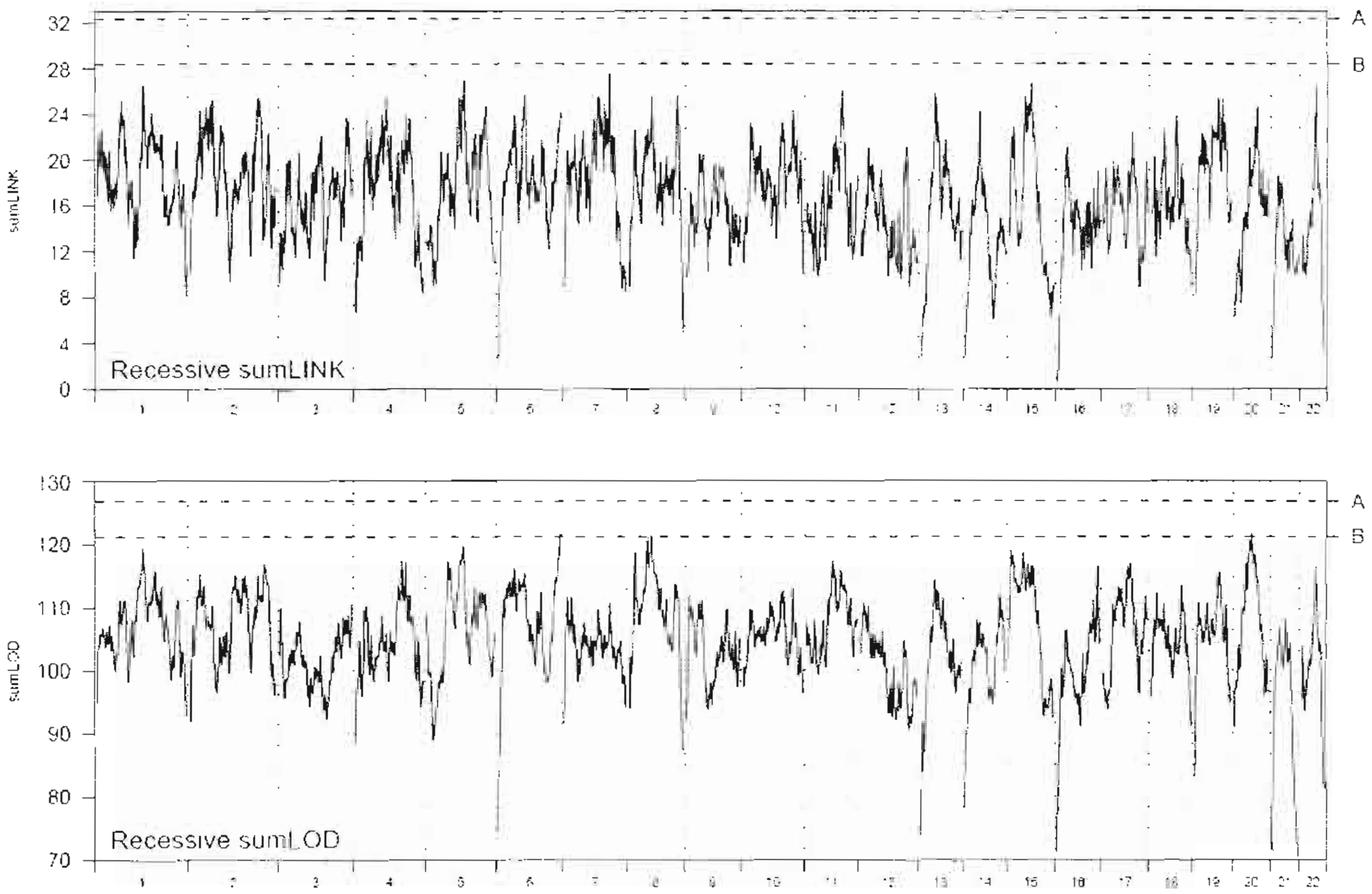


Figure 5.1. continued.

Table 5.1. Chromosomal regions with at least suggestive linkage evidence.

							FDR**		
Chr	Nearest marker	cM	Total peds contributing	Analysis	Model	FPR*	q-val	Obs peaks	Exp peaks
22q12	D22S283	42	46	sumLINK	Dom	0.010	0.115	1	0.1
	D22S283	42	454	sumLOD	Dom	0.032	0.186	2	0.4
5q11	D5S2500	75	507	sumLOD	Dom	0.059	0.200	4	0.8
	D5S407	72	43	sumLINK	Dom	0.529	0.259	17	4.4
2p24	D2S1360	39	45	sumLINK	Dom	0.089	0.200	4	0.8
6p21	D6S2427	59	495	sumLOD	Dom	0.350	0.259	17	4.4
	D6S1017	64	40	sumLINK	Dom	0.445	0.259	17	4.4
19q13	D19S900	70	43	sumLINK	Dom	0.379	0.259	17	4.4
8q12	D8S285	68	487	sumLOD	Dom	0.393	0.259	17	4.4
	D8S285	68	449	sumLOD	Rec	0.851	0.259	17	4.4
8p23	D8S1130	18	467	sumLOD	Dom	0.442	0.259	17	4.4
11q13	D11S1314	79	491	sumLOD	Dom	0.558	0.259	17	4.4
20p11-q11	D20S912	51	484	sumLOD	Rec	0.688	0.259	17	4.4
	D20S195	58	489	sumLOD	Dom	0.736	0.259	17	4.4
6q27	D6S281	189	450	sumLOD	Rec	0.740	0.259	17	4.4
1q23	D1S2628	164	44	sumLINK	Dom	0.822	0.259	17	4.4
5q35	D5S400	177	43	sumLINK	Dom	0.852	0.259	17	4.4
1p12	D1S534	140	491	sumLOD	Dom	0.935	0.262	18	4.7

* False positive rate, or FPR, refers to the expected frequency of peaks of similar or greater magnitude based on the results of 1000 repetitions of the genome randomization procedure.

** False discovery rate (FDR) results are based on the cumulative distribution of null p-values from all analyses. The q-value indicates the proportion of all results of similar or greater significance that are expected to be false positives.

findings being genome-wide significant in their respective single genomewide screens (FPRs < 0.05), after correction for all four genomewide analyses, the FDR was 0.186. This indicates that under the null hypothesis. the expected number of peaks at least as extreme as these two is only 0.4 (–0.186×2), and therefore. that 1.6 of these 2 peaks are not likely to be from the null distribution. Since both peaks are at 22q12, this indicates that even after correction for the four genomewide analyses performed here, there is excellent evidence that the 22q12 locus is a true positive.

Suggestive peaks are those that in a single genomewide screen would only be expected once per genome under the null hypothesis. Twelve loci were identified within their respective single genomewide analyses to have suggestive evidence for linkage. In decreasing order of significance, these regions were at chromosomes 5q11 (dominant sumLOD and sumLINK), 2p24 (dominant sumLINK), 6p21 (dominant sumLOD and sumLINK), 19q13 (dominant sumLINK), 8q12 (dominant and recessive sumLOD), 8p23 (dominant sumLOD), 11q13 (dominant sumLOD), 20p11-q11 (dominant and recessive sumLOD), 6q27 (recessive sumLOD), 1q23 (dominant sumLINK), 5q35 (dominant sumLINK), and 1p12 (dominant sumLOD). Loci at 5q11 and 2p24. are perhaps worthy of particular note because although strictly only suggestive, both were borderline significant (FPRs of 0.059 and 0.089, respectively). Accounting for the four genomewide analyses, the FDR value associated with these 18 suggestive and significant peaks (distributed across 13 regions) was 0.262, indicating that only 4.7 (18×0.262) peaks would have been expected under the null. That is, we observed 13.3 more peaks than expected and thus. 13.3 are likely not from the null.

Hence, there is good evidence that many, although not all, of these loci with suggestive evidence for linkage are also true positive findings.

Table 5.2 shows the results of our localization analysis for the seven significant and suggestive regions identified with the sumLINK analyses. Estimated regions are based on the observation of two recombination events at each end, indicating an approximate 95% support interval. The microsatellite markers flanking the two-recombinant region are also reported. These two-recombinant localization intervals range from 5 to 17 cM, with a mean of 9.1 cM. Since we included information from all pedigrees with a LOD≥0.588 within 5 cM of the peak, there were some instances where pedigrees showed conflicting evidence about the location of the shared chromosomal region. In these instances, we selected the region where the greatest number of pedigrees agreed, and reported the number of conflicting pedigrees in the table together with the number of supporting pedigrees.

Table 5.2. Localization intervals for sumLINK regions

Locus	Peak (cM)	2-recomb. Interval (cM)	Supporting Pedigrees*	Conflicting Pedigrees	Flanking markers
1q23	164	161—170	59	0	D1S1677—D1S452
2p24	39	29—40	53	0	D2S1400 D2S1360
5q11	72	72—79	52	2	D5S407—D5S647
5q35	177	168—185	52	0	D5S422—D5S1960
6p21	64	65—74	50	0	D6S1582—D6S1280
19q13	70	63—69	55	2	D19S570—D19S420
22q12	42	37—42	57	0	D22S280—D22S683

* Number of pedigrees with LOD ≥0.588 within 5cM of the peak.

Discussion

We have performed a secondary analysis of data from the largest collection of high-risk prostate cancer pedigrees ever assembled with new multipoint linkage-based statistics, sumLINK and sumLOD, which are specifically designed to address genetic heterogeneity. Three of the thirteen loci that we identified in the present analysis (5q11, 5q35, and 22q12) correspond directly to peaks that were reported in the original ICPCG analysis using the conventional HLOD statistic [8]. In that analysis, a dominant LOD score of 1.95 was observed at 22q12, which increased to 3.57 in the subset of pedigrees with at least five affected family members. Additionally, a non-parametric LOD of 2.28 was reported at 5q12, and a dominant LOD of 2.05 was reported at 5q35 in the subset of families with mean age at diagnosis ≤65 years. Two other loci (1q23 and 8q12) are near peaks that were reported in the first analysis [8]. The loci on chromosomes 6p21, 11q13, and 20p11-q11 correspond to susceptibility loci previously identified in the ICPCG data resource in linkage scans for aggressive prostate cancer [11,15]. The remaining loci have not previously been identified in pooled ICPCG data, though many of them correspond to findings reported elsewhere in linkage studies by individual institutions.

The dominant and recessive sumLOD peaks on chromosome 20 appear to be supportive of the HPC20 locus [16], although it should be noted that the original HPC20 linkage peak was at 20q13, about 20-30 cM downstream from the peaks we report here. Our tentative replication of HPC20 is in contrast to an earlier ICPCG study using the same data and a conventional HLOD approach that failed to replicate this locus [10], although a later ICPCG study concentrating on aggressive prostate cancer pedigrees did find linkage evidence [15]. The ICPCG aggressive PC linkage study found a dominant LOD score of 2.49 midway between the dominant and recessive sumLOD peaks that we report here. The observed LOD score increased to 2.65 in the subset of pedigrees with mean age at onset >65 years. The present study includes data from most of the pedigrees that were included in the ICPCG aggressiveness analysis, but the difference in phenotype definition prevents a direct comparison of the pedigrees that contribute to the results. HPC20 was originally identified by the Mayo Clinic site [16,17]; however, of the 45 pedigrees that exhibited LOD≥0.588 within 5 cM of the dominant sumLOD peak, only 6 were from Mayo Clinic. As seen from these comparisons, one distinct advantage of the sumLINK and sumLOD statistics is that the approach inherently identifies subgroups of pedigrees that are genetically alike, and hence, one analysis can encompass what in conventional analyses may take many subset analyses and multiple testing corrections. It is therefore perhaps not surprising that our results more closely align with linkage findings for subset-based analyses such as aggressive prostate cancer [15].

In addition to the findings discussed above, three of the other suggestive linkage regions reported here support previously identified loci. Our peak at 1p12 falls within a region of interest reported by other ICPCG member-sites [18]. The peak at chromosome 1q23 approximates the HPC1 susceptibility region [19], although the *RNASEL* candidate gene proposed as the HPC1 gene [20] is located about 20 Mb beyond the boundary of our support interval. An ICPCG member-site previously reported linkage at 8p23 [21], a finding that was recently replicated and refined by combined somatic deletion and fine linkage mapping [22]. The suggestive sumLOD peak at 8p23 is about 4 Mb from the *MSR1* PC candidate gene. Our 19q13 region also corresponds to previously reported linkage for aggressive PC [23,24].

Our suggestive regions on chromosomes 2p24 and 6q27 appear to be new. Of particular interest of these new loci is perhaps 2p24. Statistical evidence for 2p24 was borderline significant. and recently. a germline copy number variant at the 2p24 locus has been associated with aggressive prostate cancer [25]. Other notable association studies have focused on regions identified in this report. Copy number variations at 8p23 and 11q13 have been implicated in aggressive PC and PC recurrence, respectively [26]. Kallikrein genes *KLK2* and *KLK3* at chromosome 19q13 have been identified as PC candidate genes [27].

We did not identify linkage evidence to regions that have recently received much attention due to highly significant and replicable association evidence with PC in genome-wide association studies. The most compelling of these results are located on chromosomes 8q24. 17q12. and 10q11 [3]. It is perhaps not surprising that we did not find any evidence to support these regions because these SNPs have common minor allele frequencies and very small effect sizes. The sumLINK and sumLOD are linkage-based statistics. and linkage is most powerful for finding rarer. more highly-penetrant variants.

The localization procedure we used here to delimit support intervals generated much more concise intervals than the 1-LOD drop regions reported previously by ICPCG for the four sumLINK peaks that overlapped with previous findings [8] . The intervals reported previously ranged from 12 to 30 cM with a mean length of 21.2 cM. substantially longer than the mean length of 9.5 cM we report here for the same 4 regions. A particularly interesting example of the narrower intervals can be seen in the putative susceptibility locus at chromosome 5q11-12. The previous analysis of this data identified a suggestive HLOD peak at 77 cM. with a reported 1-LOD support interval extending from 66—96 cM. In the present analysis. the sumLINK statistic identified a suggestive linkage peak at 72 cM and a 2-recombinant support interval of only 7 cM, which includes the original HLOD peak. This ability to more narrowly define regions using statistical recombinant mapping was also illustrated by an earlier candidate region localization study for the chromosome 22q12 susceptibility locus [9]. That report had the advantage of LOD score data from several large pedigrees with fine-mapping markers that were not included in the present results. Nonetheless. and as expected. the 2-recombinant localization region we report here supports the region previously reported in that paper.

Conclusion

A secondary reanalysis of 1233 PC pedigrees using novel linkage statistics identified 13 regions with at least genomewide suggestive evidence for linkage. Eight regions provide confirmation of loci previously identified by conventional linkage analyses in the same ICPCG data [8] or the subset of aggressive PC pedigrees [15], three are regions that confirm loci not seen in the original analyses. but are reported in other linkage studies [18,22-24], and two are novel loci. One distinct benefit of the sumLINK and sumLOD approach is that the statistics are based on the identification of pedigrees that are genetically alike at a locus. and the constituent set of pedigrees may change from locus-to-locus. This both addresses genetic heterogeneity directly and largely circumvents the need for subset and stratification analyses that are costly in terms of multiple testing. This is illustrated by the fact that several of the regions identified here replicate results that were originally found in stratification analyses. The second advantage for the sumLINK statistic is the natural progression to statistical recombinant mapping, which appears to hold much promise for narrowing linkage regions. Furthermore, the FDR approach for correction of multiple genomewide analyses can better guide interpretation and aid prioritization of findings. Evidence here suggests that these statistics have the potential to further refine the results of original analyses, and provide new directions in the pursuit for PC susceptibility genes.

Acknowledgements

We would like to express our gratitude to the many families who participated in the many studies involved in the International Consortium for Prostate Cancer Genetics (ICPCG). The ICPCG, including the consortium's Data Coordinating Center (DCC), is made possible by a grant from the National Institutes of Health U01 CA89600 (to W.B.I.). G.B.C. was supported by National Library of Medicine training grant NLM T15 LM07124. N.J.C. was supported in part by USPHS CA98364 (to N.J.C.). Additional support to participating groups, or members within groups, is as follows:

AAHPC Group: The authors would like to express their gratitude to the African-American Hereditary Prostate Cancer Study (AAHPC) families and study participants for their continued involvement in this research. We specifically name, C.

Ahaghotu. J. Bennett, W. Boykin. G. Hoke, T. Mason, C. Pettaway, S. Vijayakumar, S. Weinrich, M. Franklin. P. Roberson, J. Frost. E. Johnson. L. Faison-Smith. C. Meegan. M. Johnson, L. Kososki. C. Jones, and R. Mejia. We would also like to thank members of the National Human Genome Center (NHGC) at Howard University namely R. Kittles, G.M. Dunston. P. Furbert-Harris, and C. Royal. We would also like to acknowledge the contribution of the National Human Genome Research Institute (NHGRI) and TGen genotyping staff including E. Gillanders and C. Robbins. The AAHPC study would not have been possible without F. Collins (Director of NHGRI) and J. Trent (Director of TGen). This research was funded primarily through the NIH Center for Minority and Health Disparities (1-HG-75418). A.B.B-B and A.G also received support from USPHS CA-06927 and an appropriation from the Commonwealth of Pennsylvania. This research was also supported in part by the Intramural Research Program of the NIH (NHGRI) and USPHS RR03048 from the National Center for Research Institute and USPHS RR03048 from the National Center for Research Resources. **ACTANE Group**: Genotyping and statistical analysis for this study. and recruitment of U.K. families. was supported by Cancer Research U.K (CR-UK). Additional support was provided by The Prostate Cancer Research Foundation, The Times Christmas Appeal, and the Institute of Cancer Research. Genotyping was conducted in the 'Jean Rook Gene Cloning Laboratory' which is supported by BREAKTHROUGH Breast Cancer - Charity No. 328323. The funds for the ABI 377 used in this study were generously provided by the legacy of the late Marion Silcock. We thank S. Seal and A. Hall for kindly storing and logging the samples that were provided. D.F.E is a Principal Research Fellow of CR-UK. Funding in Australia was obtained from The Cancer Council Victoria, The National Health and Medical Research Council (grants 940934, 251533, 209057, 126402, 396407), Tattersall's and The Whitten Foundation. We would like to acknowledge the work of the study coordinator M. Staples and the Research Team B. McCudden, J. Connal, R. Thorowgood, C. Costa, M. Kevan, and S. Palmer, and to J. Karpowicz for DNA extractions. The Texas study of familial prostate cancer was initiated by the Department of Epidemiology, M.D. Anderson Cancer Center. M.B. was supported by an NCI Post-doctoral Fellowship in Cancer Prevention (R25). We would also like to specifically thank the following members of ACTANE: S. Edwards, M. Guy, Q. Hope, S. Bullock, S. Bryant, S. Mulholland, S. Jugurnauth, N. Garcia, A. Ardern-Jones, A. Hall, L. O'Brien, B. Gehr-Swain, R. Wilkinson, D. Dearnaley, The UKGPCS Collaborators, British Association of Urological Surgeons' Section of Oncology (UK Sutton); Chris Evans (UK Cambridge); M. Southey (Australia); N. Hamel, S. Narod, J. Simard (Canada); C. Amos (USA Texas); N. Wessel, T. Andersen (Norway); D.T. Bishop (EU Biomed). **BC/CA/HI Group:** USPHS CA67044. **FHCRC Group:** USPHS CA80122 (to J.L.S.) which supports the family collection; USPHS CA78836 (to E.A.O). E.A.O was supported in part by the NHGRI. **JHU Group:** Genotyping for the JHU, University of Michigan, University of Tampere, and University of Umeå groups' pedigrees was provided by NHGRI and TGen genotyping staff including E. Gillanders, MP Jones, D. Gildea, E. Riedesel, J. Albertus, D. Freas-Lutz, C. Markey, J. Carpten, and J. Trent. **Mayo Clinic Group:** USPHS CA72818. **Michigan Group:** USPHS CA079596. **University of Tampere Group:** The Competitive Research Funding of the Pirkanmaa Hospital District, Reino Lahtikari

Foundation, Finnish Cancer Organisations, Sigrid Juselius Foundation, and Academy of Finland grant 211123. **University of Ulm Group:** Deutsche Krebshilfe, grant number 70-3111-V03. **University of Umea Group:** Work was supported by the Swedish Cancer Society and a Spear grant from the Umeå University Hospital, Umeå, Sweden. **University of Utah Group:** Data collection was supported by USPHS CA90752 (to L.A.C.-A.) and by the Utah Cancer Registry, which is funded by Contract #N01-PC-35141 from the National Cancer Institute's Surveillance, Epidemiology, and End-Results Program with additional support from the Utah State Department of Heath and the University of Utah. Partial support for all datasets within the Utah Population Database was provided by the University of Utah Huntsman Cancer Institute and also by the USPHS M01-RR00064 from the National Center for Research Resources. Genotyping services were provided by the Center for Inherited Disease Research (N01-HG-65403). **DCC:** The study is partially supported by USPHS CA106523 (to J.X.), USPHS CA95052 (to J.X.), and Department of Defense grant PC051264 (to J.X.).

References

1. Schaid D: The Complex Genetic Epidemiology of Prostate Cancer. *Human Molecular Genetics* 2004, 13:R103-121.

2. Ostrander EA, Stanford JL: Genetics of prostate cancer: too many loci, too few genes. *Am J Hum Genet* 2000, 67:1367-1375.

3. Witte JS: Prostate cancer genomics: towards a new understanding. *Nat Rev Genet* 2009, 10:77-82.

4. Ghoussaini M, Song H, Koessler T, Al Olama AA, Kote-Jarai Z, Driver KE, Pooley KA, Ramus SJ, Kjaer SK, Hogdall E, et al: Multiple loci with different cancer specificities within the 8q24 gene desert. *J Natl Cancer Inst* 2008, 100:962-966.

5. Luo JH, Yu YP: Genetic factors underlying prostate cancer. *Expert Rev Mol Med* 2003, 5:1-26.

6. Yoshimoto M, Joshua AM, Chilton-Macneill S, Bayani J, Selvarajah S, Evans AJ, Zielenska M, Squire JA: Three-color FISH analysis of TMPRSS2/ERG fusions in prostate cancer indicates that genomic microdeletion of chromosome 21 is associated with rearrangement. *Neoplasia* 2006, 8:465-469.

7. Liu W, Chang B, Sauvageot J, Dimitrov L, Gielzak M, Li T, Yan G, Sun J, Sun J, Adams TS, et al: Comprehensive assessment of DNA copy number alterations in human prostate cancers using Affymetrix 100K SNP mapping array. *Genes Chromosomes Cancer* 2006, 45:1018-1032.

8. Xu J, Dimitrov L, Chang BL, Adams TS, Turner AR, Meyers DA, Eeles RA, Easton DF, Foulkes WD, Simard J, et al: A combined genomewide linkage scan of 1,233 families for prostate cancer-susceptibility genes conducted by the international consortium for prostate cancer genetics. *Am J Hum Genet* 2005, 77:219-229.

9. Camp NJ, Cannon-Albright LA, Farnham JM, Baffoe-Bonnie AB, George A, Powell I, Bailey-Wilson JE, Carpten JD, Giles GG, Hopper JL, et al: Compelling evidence for a prostate cancer gene at 22q12.3 by the International Consortium for Prostate Cancer Genetics. *Hum Mol Genet* 2007, 16:1271-1278.

10. Schaid DJ, Chang BL: Description of the International Consortium For Prostate Cancer Genetics, and failure to replicate linkage of hereditary prostate cancer to 20q13. *Prostate* 2005, 63:276-290.

11. Christensen GB, Knight S, Camp NJ: The sumLINK statistic for genetic linkage analysis in the presence of heterogeneity. *Genet Epidemiol* 2009.

12. Lander E, Kruglyak L: Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* 1995, 11:241-247.

13. Chen L, Storey JD: Relaxed Significance Criteria for Linkage Analysis. *Genetics* 2006, 173:2371-2381.

14. Camp NJ, Farnham JM, Cannon-Albright L: Localization of a Prostate Cancer Predisposition Gene to an 880-kb Region of Chromosome 22q12.3 in Utah High-Risk Pedigrees. *Cancer Res* 2006, 66:10205-10212.

15. Schaid D. ICPCG: Pooled genome linkage scan of aggressive prostate cancer: Results from the International Consortium for Prostate Cancer Genetics. *Hum Genet* 2006, 120(4):471-485.

16. Berry R, Schroeder JJ, French AJ, McDonnell SK, Peterson BJ, Cunningham JM, Thibodeau SN, Schaid DJ: Evidence for a prostate cancer-susceptibility locus on chromosome 20. *Am J Hum Genet* 2000, 67:82-91.

17. Cunningham JM, McDonnell SK, Marks A, Hebbring S, Anderson SA, Peterson BJ, Slager S, French A, Blute ML, Schaid DJ, Thibodeau SN: Genome linkage screen for prostate cancer susceptibility loci: results from the Mayo Clinic Familial Prostate Cancer Study. *Prostate* 2003, 57:335-346.

18. Slager S, Zarfas KE, Brown WM, Lange E, McDonnell S, Wojno KJ, Cooney KA: Genome-wide linkage scan for prostate cancer aggressiveness loci using families from the University of Michigan Prostate Cancer Genetics Project. *Prostate* 2006, 66:173-179.

19. Smith JR, Freije D, Carpten JD, Gronberg H, Xu J, Isaacs SD, Brownstein MJ, Bova GS, Guo H, Bujnovszky P, et al: Major susceptibility locus for prostate cancer on chromosome 1 suggested by a genome-wide search. *Science* 1996, 274:1371-1374.

20. Carpten J, Nupponen N, Isaacs S, Sood R, Robbins C, Xu J, Faruque M, Moses T, Ewing C, Gillanders E, et al: Germline mutations in the ribonuclease L gene in families showing linkage with HPC1. *Nat Genet* 2002, 30:181-184.

21. Xu J, Zheng SL, Hawkins GA, Faith DA, Kelly B, Isaacs SD, Wiley KE, Chang B, Ewing CM, Bujnovszky P, et al: Linkage and association studies of prostate cancer susceptibility: evidence for linkage at 8p22-23. *Am J Hum Genet* 2001, 69:341-350.

22. Chang BL, Liu W, Sun J, Dimitrov L, Li T, Turner AR, Zheng SL, Isaacs WB, Xu J: Integration of somatic deletion analysis of prostate cancers and germline linkage analysis of prostate cancer families reveals two small consensus regions for prostate cancer genes at 8p. *Cancer Res* 2007, 67:4098-4103.

23. Neville PJ, Conti DV, Krumroy LM, Catalona WJ, Suarez BK, Witte JS, Casey G: Prostate cancer aggressiveness locus on chromosome segment 19q12-q13.1 identified by linkage and allelic imbalance studies. *Genes Chromosomes Cancer* 2003, 36:332-339.

24. Slager SL, Schaid DJ, Cunningham JM, McDonnell SK, Marks AF, Peterson BJ, Hebbring SJ, Anderson S, French AJ, Thibodeau SN: Confirmation of linkage of prostate cancer aggressiveness with chromosome 19q. *Am J Hum Genet* 2003, 72:759-762.

25. Liu W, Sun J, Li G, Zhu Y, Zhang S, Kim ST, Sun J, Wiklund F, Wiley K, Isaacs SD, et al: Association of a germ-line copy number variation at 2p24.3 and risk for aggressive prostate cancer. *Cancer Res* 2009, 69:2176-2179.

26. Paris PL, Andaya A, Fridlyand J, Jain AN, Weinberg V, Kowbel D, Brebner JH, Simko J, Watson JE, Volik S, et al: Whole genome scanning identifies genotypes associated with recurrence and metastasis in prostate tumors. *Hum Mol Genet* 2004, 13:1303-1313.

27. Pal P, Xi H, Sun G, Kaushal R, Meeks JJ, Thaxton CS, Guha S, Jin CH, Suarez BK, Catalona WJ, Deka R: Tagging SNPs in the kallikrein genes 3 and 2 on 19q13 and their associations with prostate cancer in men of European origin. *Hum Genet* 2007, 122:251-259.

CHAPTER 6

SUMMARY AND CONCLUSIONS

Summary

The purpose of this research was to extend methodology for identifying disease genes in heterogeneous systems. Current methodology for genetic linkage analysis was examined and new methods for linkage analysis were developed. Prostate cancer (PC), the genetic etiology of which is believed to be very complicated, was used as a model system throughout the research.

Chapter 2 Review

Genetic heterogeneity has been identified as the principle factor responsible for the numerous published hints of PC linkage and the relative dearth of positive replication studies. It is encouraging that analysis of clinically defined PC subtypes, especially "aggressive" disease cases, has resulted in several replicated linkage findings. Chapter 2 is a conventional linkage analysis for aggressive PC using data ascertained from an unconventional resource, the Utah Population Database (UPDB) [1]. Analysis of the large extended pedigrees in this study did not return any significant linkage results, but previously reported linkages for the aggressive phenotype to chromosomes 6p and 20q were confirmed. Suggestive linkage regions on chromosomes 1p and 8q also support previously reported PC linkage.

It is very interesting to note the differences between the results of this study and the results of the previous linkage analysis of the same pedigrees for all PC cases [2]. The best linkage signals reported in that study, on chromosomes 1, 3, 5, and 22, practically disappeared in the analysis of the aggressive phenotype. Conversely, the suggestive linkage results in the aggressive analysis are not generally seen in the results of the original analysis. This observation supports the notion that aggressive PC may have a different genetic etiology than PC in general, or that the etiology of PC may involve too many genes for any one of them to be detected by analysis of only the broader phenotype.

The challenge for prostate cancer researchers is to sufficiently clarify the differences within the PC phenotype so that genes controlling each subtype can be identified. This includes improving the definitions of "aggressive" and "early onset" disease. Similar challenges exist for other heterogeneous phenotypes as well. Obtaining valid results from linkage analysis using conventional techniques requires that phenotypes be defined as concisely as possible. It is possible that new analytical methods will make it possible to find more genes with marginal effects, but any ambiguity in the analyzed phenotype will make it more difficult to establish a causative relationship.

Chapter 3 Review

To identify disease susceptibility genes in complex systems, it may be necessary find alternative methods of defining the phenotype of interest, thereby excluding statistical noise produced by phenocopies (individuals whose disease was caused by a nongenetic factor). It has been proposed that biomarkers associated with the disease, such as RNA expression levels, may lead to genes associated with disease, particularly in regulatory pathways. Chapter 3 is a conceptual study that uses gene expression profiles from randomly-ascertained individuals in pedigrees as a phenotype for linkage analysis with the hypothesis that loci may be found that are linked to the expression profile as well as the associated disease. This study concentrated on an expression profile that could be considered a marker of PC risk. The risk level for study subjects was determined based on the expression levels of 26 genes that had previously been implicated as over- or under-expressed in PC tumors. Linkage analysis of this "PC risk" phenotype identified interesting results on chromosomes 4 and 6 in areas previously linked to PC susceptibility.

A weakness of this study is that the risk profiles used as phenotypes in this chapter used expression data measured in blood cells, whereas the model for risk was based on reported RNA expression levels in prostate tumors. This presumes that blood biomarkers can, at some level, represent tumor biomarkers and can be used to predict disease risk. This is a novel concept that is receiving much attention in current cancer research. The NIH Challenge Grants in Health and Science Research initiative, part of the American Recovery and Reinvestment Act of 2009, includes requests for research proposals to identify body fluid expression biomarkers that provide early detection for the risk of cancer and age-related diseases. Understanding the genetics underlying gene expression, which is at the center of Chapter 3, has also been identified as a key to understanding human disease genetics [3].

The results in Chapter 3 stand as proof of concept that the expression levels of several related genes can be combined to create a reasonable phenotype for linkage analysis. As shown in Chapter 2, it is very important to concisely define phenotypes used in linkage analysis in order to obtain reliable results. In the event that a phenotype's clinical presentation is easily confused with a different, unrelated phenotype, the use of related gene expression levels and other biomarkers may improve the accuracy of the diagnosis. and thereby strengthen the results of the analysis. The methods presented here would be particularly useful for identifying inherited variants that may regulate genetic pathways.

Chapter 4 Review

Collaborative data sharing, appropriately accounting for multiple testing, and developing new analytical methods have all been identified as necessary factors to overcome the problems of heterogeneity in linkage analysis [4]. Chapter 4 describes an attempt to address all of these factors through the development of the novel sumLINK and sumLOD analysis method. These statistics are designed to address both heterogeneity and gene localization. Trait genes in heterogeneous systems may not be found by traditional linkage methods because only a limited number of pedigrees may be linked to the locus. and the linkage signal from those pedigrees does not rise above the statistical "background noise." The sumLINK and sumLOD procedures identify the genomic regions for which a substantial number of pedigrees give concordant linkage evidence, regardless of the negative evidence exhibited by unlinked pedigrees. The sumLINK and sumLOD methods were validated by extensive simulations. Simulation testing established that the rate of detecting false positive results was followed the expected distribution. Both statistics routinely outperformed the conventional HLOD statistic in identifying the location of trait related genes in data simulating a complex disease system.

The methods described in Chapter 4 are of particular use within the framework of large collaborative data as they require neither the sharing of raw data nor the use of common genetic marker sets. Empirical methods for false discovery rate (FDR) analysis give perspective to the significance of results across multiple analyses. These characteristics were demonstrated in a case study that analyzed 190 pedigrees with aggressive PC provided by the International Consortium for Prostate Cancer Genetics (ICPCG). The case study successfully replicated the results of a conventional linkage analysis and also identified an additional locus linked to PC aggressiveness that had not been reported previously.

Chapter 5 Review

The International Consortium for Prostate Cancer Genetics (ICPCG) has assembled a collection of 1233 high-risk PC pedigrees for linkage analysis. Chapter 5 is a practical application of the sumLINK and sumLOD methods within this large ICPCG data resource. Conventional analysis of this resource previously identified significant linkage on chromosome 22q12. and suggestive linkage evidence in several other regions [5]. Most of the previously identified regions of interest were found via subset analysis. The sumLINK and sumLOD analyses in Chapter 5 confirm the significant linkage result at 22q12 as well as suggestive loci originally reported by the ICPCG at chromosomes 1q23, 5q11, 5q35, 6p21, 8q12, 11q13, and 20p11-q11. This analysis also provided confirmation of susceptibility loci reported by other researchers on chromsomes 1p12. 8p23, and 19q13, which were not found in this data using conventional linkage statistics. All linkage regions identified by the sumLINK statistic can be delimited by statistical recombinant mapping, which can greatly reduce the length of the chromosomal segment where candidate genes may be found. The results found with sumLINK and sumLOD did not require subsetting the data, thereby limiting the effects of multiple testing on the significance of the results. The results in this chapter identify the genomic regions with the greatest consistency of positive linkage evidence across a very large collection of PC pedigrees. These regions are excellent candidates for further study to identify PC predisposition genes.

Contribution to Biomedical Informatics

Biomedical informatics is the application of computer science and information systems to healthcare and biological research. The field of genetic epidemiology relies heavily on biomedical informatics tools and resources. The Utah Population Database [1] and online resources such the UCSC Genome Browser [6] are examples of informatics resources that make genetic epidemiology research possible. Genetic epidemiologists contribute to biomedical informatics by developing phenotype and genotype databases and by creating software tools and algorithms for analyzing genetic data and testing hypotheses of genetic contributions. The present research makes a significant contribution to biomedical informatics with the development of the sumLINK analysis method. Aside from being a powerful statistic for genetic epidemiology research, the sumLINK procedure has two major benefits for informatics.

First, the novel genome shuffling algorithm used to determine sumLINK significance may have a variety of additional applications both in and out of the biological sciences. The algorithm can be applied to any type of trend data measured at regular intervals for several experimental subjects. Possible applications include such varied fields as public health surveillance (Is an apparent peak in over-the-counter sales of flu medicine at several pharmacies significantly different from the established norms, and therefore indicative of an outbreak?) and meteorology (Is it significant that several weather stations in a broad area reported recorded spikes in wind speed at the time?).

The second informatics benefit of the sumLINK method is that it facilitates collaborative research and data sharing. Data security and protecting the privacy of research subjects are very important issues in informatics. Sharing private patient data or other secure, proprietary information between institutions can be a barrier to research collaborations as institutional review boards are generally wary about allowing researchers to pass any sensitive data outside of their own institution. The sumLINK procedure uses only unidentifiable meta data as input, and can therefore be used to analyze data from multiple institutions without sharing any private data. This characteristic makes the sumLINK a very valuable tool from the perspective of biomedical informatics.

Discussion

The problems presented by genetic heterogeneity will not be easy for genetic epidemiologists to overcome. Many years of research have already been dedicated to this topic, and many more years are likely to follow. It is clear that current methods are not capable of solving the problem. New and innovative analytical methods are needed. The sumLINK and sumLOD statistics are a step in the right direction, but they must still withstand the test of time. SumLINK analysis outperforms conventional methods in simulated data, but no genes have yet been proven to exist based on sumLINK evidence, and it remains unclear whether susceptibility loci identified with this method will have a better replication rate than traditional methods.

Despite the novelty of the methods, sumLINK and sumLOD remain LOD-based linkage statistics. The LOD statistic has been used for over 50 years, and it may be that the answer to the overarching heterogeneity problem lies somewhere further outside the bounds of traditional methodology. Advances in high-throughput genotyping make it possible to simultaneously genotype over one million single nucleotide polymorphisms (SNPs), and exciting new methods are being developed to utilize this rich data. One such technique is shared genomic segment (SGS) analysis [7]. This method uses long runs of SNPs at which alleles are shared identically by state in pedigrees to localize hypothesized predisposition genes. High-density SNP genotyping has also led to the current trend of genome-wide association studies (GWAS). GWAS analyses typically use a case-control design with large numbers of unrelated subjects. This type of study is still in developmental stages, but has already demonstrated utility for finding common genetic variants that have moderate statistical association with disease status.

Although useful, GWAS cannot entirely replace pedigree-based analysis. Current GWAS methods lack the power to identify rare (allele frequency < 0.05), high-risk coding variants, which is a strength of linkage analysis [3]. The expense of collecting and genotyping the large number of subjects needed for GWAS is also a limiting factor. There are some phenotypes for which linkage will always be more powerful and efficient. A combination of linkage and association techniques is likely to give the best power in heterogeneous systems. An example of the synergy that results from combined linkage and association testing can be seen in the chromosome 8q24 PC locus. This locus was identified by linkage analysis and replicated by GWAS, but it remains unclear which gene in that region, if any, is responsible for the association with increased PC risk. In such a case, statistical recombinant mapping with linked pedigrees can refine the area most likely to harbor the true susceptibility variant [8]. Pedigree-based data is also useful for association studies, and several tools have been developed for association testing using related subjects [9-12]. Methods for association analysis using related subjects need to be refined and extended for use in genome-wide setting. This concept is addressed in Appendix C, which contains results of a combined linkage and association study in extended pedigrees. The efficiency and economy of genomic research can be greatly improved if linkage and association analysis can be performed in the same group of subjects, as the cost of subject ascertainment and genotyping can be combined.

The future of genetic epidemiology research relies on the development of analysis methods with sufficient power to identify susceptibility genes in complex, heterogeneous systems. This may require a fundamental shift in the way we think about phenotypes, as demonstrated in Chapters 2 and 3. It may also require the development of new analytical methods, such as the sumLINK and sumLOD statistics described in Chapters 4 and 5. Methods must also be developed for emerging data types. The past decade has seen an explosion in genotype data with the introduction of affordable, high throughput, high density SNP genotyping technology. Despite the advances in data collection, there has been little fundamental change in the tools used to analyze genetic data over the same period. The coming decade is likely to experience another genotyping renaissance as "next-generation" sequencing technology comes of age [13]. At the present, this technology can greatly assist with mutation screening and candidate gene identification in linkage regions, but it may have far greater implications in the future, as full genome sequencing becomes a practical reality. Full genome sequences will contain treasures of knowledge, but it will never be understood unless genetic epidemiologists are prepared with the methods needed to process and analyze such vast amounts of data. We must be ready.

References

1. Skolnick M (Ed.). *The Utah genealogical data base: a resource for genetic epidemiology*. New York: Cold Spring Harbor Laboratory; 1980.

2. Camp NJ FJ, Cannon-Albright LA: Genomic search for prostate cancer predisposition loci in Utah pedigrees. *Prostate* 2005, 65:365-374.

3. Hardy J, Singleton A: Genomewide Association Studies and Human Disease. *N Engl J Med* 2009, 360:1759-1768.

4. Morton NE: Unsolved problems in genetic epidemiology. *Hum Hered* 2000, 50:5-13.

5. Xu J, Dimitrov L, Chang BL, Adams TS, Turner AR, Meyers DA, Eeles RA, Easton DF, Foulkes WD, Simard J, et al: A combined genomewide linkage scan of 1,233 families for prostate cancer-susceptibility genes conducted by the international consortium for prostate cancer genetics. *Am J Hum Genet* 2005, 77:219-229.

6. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: The human genome browser at UCSC. *Genome Res* 2002, 12:996-1006.

7. Thomas A, Camp NJ, Farnham JM, Allen-Brady K, Cannon-Albright LA: Shared genomic segment analysis. Mapping disease predisposition genes in extended pedigrees using SNP genotype assays. *Ann Hum Genet* 2008, 72:279-287.

8. Camp NJ, Farnham JM, Allen-Brady K, Cannon-Albright LA: Statistical recombinant mapping in extended high-risk Utah pedigrees narrows the 8q24 prostate cancer locus to 2.0 Mb. *Prostate* 2007, 67:1456-1464.

9. Slager SL, Schaid DJ, Wang L, Thibodeau SN: Candidate-gene association studies with pedigree data: controlling for environmental covariates. *Genet Epidemiol* 2003, 24:273-283.

10. Allen-Brady K, Wong J, Camp NJ: PedGenie: an analysis approach for genetic association testing in extended pedigrees and genealogies of arbitrary size. *BMC Bioinformatics* 2006, 7:209.

11. Curtin K, Wong J, Allen-Brady K, Camp NJ: PedGenie: meta genetic association testing in mixed family and case-control designs. *BMC Bioinformatics* 2007, 8:448.

12. Browning SR, Briley JD, Briley LP, Chandra G, Charnecki JH, Ehm MG, Johansson KA, Jones BJ, Karter AJ, Yarnall DP, Wagner MJ: Case-control single-marker and haplotypic association analysis of pedigree data. *Genet Epidemiol* 2005, 28:110-122.

13. Mardis ER: The impact of next-generation sequencing technology on genetics. *Trends Genet* 2008, 24:133-141.

APPENDIX A

SURVEY OF EXCESS FAMILIALITY IN PROSTATE CANCER

Excerpt from a poster presented at the 2006

National Library of Medicine Informatics Training Conference

Survey of Excess Familiality in Prostate Cancer

by:

GB Christensen, JM Farnham, NJ Camp, LA Cannon-Albright

University of Utah Department of Biomedical Informatics

Background

Prostate cancer (PCa) is the most commonly diagnosed cancer among men, and has long been recognized to occur in familial clusters. However, identification of genes predisposing individuals to prostate cancer has been difficult. Putative PCa predisposition loci identified by genetic linkage have been reported on almost all chromosomes, but successful confirmation reports have been rare. PCa is a complex disease likely involving multiple genes and variable phenotypic expression. As a step toward understanding PCa heterogeneity, we used the resources of the Utah Population Database to review several PCa-related phenotypes for excess familiality. PCa subgroups that can be shown to have a strong familial component become candidates for linkage analysis and other genetic testing to determine the genetic basis for the observed phenotype.

Data Resource

- Utah Population Database (UPDB)
    - Records for approximately 2.2 million individuals
    - Up to 9 generations of genealogical data linking individuals into pedigrees

  - Linked to death certificates from Utah vital records, providing cause of death data since 1904
- Utah Cancer Registry (UCR)
  - Part of Surveillance, Epidemiology and End Results (SEER) program since 1973
  - All cancer events except basal and squamous cell carcinomas are required to be reported
  - Fully linked to UPDB
- At the time of this study, 17,379 PCa cases from UCR were linked to UPDB genealogies
- Tables A.1 --5 summarize the primary variables from UCR and UPDB used in this study

Genealogical Index of Familiality (GIF)

The GIF statistic tests the hypothesis that a set of individuals is more closely related than would be expected by chance. The statistic is computed for the cases and for 1000 sets of controls that are carefully matched to each subject in the case group based on sex, year of birth, and place of birth. An empirical p-value determines the significance of the relatedness of the cases in comparison to the repeated controls. Results of GIF analysis are summarized in Table A.6.

Table A.1: Age at PCa diagnosis

Age (years)	N
40-49	143
50-59	1283
60-69	5436
70-79	7184
80-89	3051
≥90	273

Table A.2: Age at PCa-related death

Age (years)	N
40-49	15
50-59	166
60-69	801
70-79	2028
80-89	1943
≥90	420

Table A.3: ICD cause-of-death codes for PCa

ICD Revision	Code	N
6	177	35
7	177	487
8	185	650
9	185	2719
10	C61	1487

Table A.4: PCa stage at diagnosis

Stage Code	Description	N
1	Localized	6973
2,3.4.5	Regional	6840
7	Distant metastases / Systemic disease	5206

Table A.5: PCa grade at diagnosis

Grade Code	Description	N
1	Well differentiated	7205
2	Moderately differentiated	6930
3	Poorly differentiated	3035
4	Undifferentiated, anaplastic	209

Table A.6. Summary of GIF analysis for prostate cancer subgroups

		Overall GIF			GIF without 1° and 2° relatives		
Phenotype Group	N	Cases	Controls	Empirical P	Cases	Controls	Empirical P
All PC Cases	17,379	3.75	3.02	<0.001	2.80	2.52	<0.001
Localized PCa (Stage=1)	6973	3.87	3.03	<0.001	2.80	2.48	<0.001
Metastatic PCa (Stage=7)	1506	3.54	3.01	<0.001	2.57	2.45	0.174
PCa Dx before age 60	1426	4.99	3.14	<0.001	2.98	2.76	0.034
PCa Dx after age 80	3324	3.92	2.99	<0.001	3.13	2.61	<0.001
Grade 1 PCa	7205	3.80	2.99	<0.001	2.73	2.41	<0.001
Grade 3/4 PCa	3244	3.81	3.02	<0.001	2.89	2.55	<0.001
PCa-related death	5378	3.83	2.94	<0.001	2.39	2.25	0.008
PCa death before age 65	456	5.01	2.98	<0.001	2.30	2.42	0.661
Malnutrition Death*	1481	2.97	2.98	0.528	2.28	2.45	0.918

- All PCa subgroups analyzed show greater than expected familiality
- Cases with metastatic disease show reduced significance when the contribution of close relatives is removed
- Relatedness of group who died from PCa before age 65 also loses significance beyond close relatives

Graphical GIF Results

The graphs in Figure A.1 show the contribution made to the GIF statistic by individuals with varying degrees of relatedness. Each step on the horizontal axis represents increasingly distant relatives. In familial diseases, we expect the case group to be consistently higher than the controls for several steps. In each frame the black bars represent the case group and the grey bars represent the average results of 1000 matched control sets. The GIF results of malnutrition-related death are included to show the typical behavior of a non-hereditary phenotype.

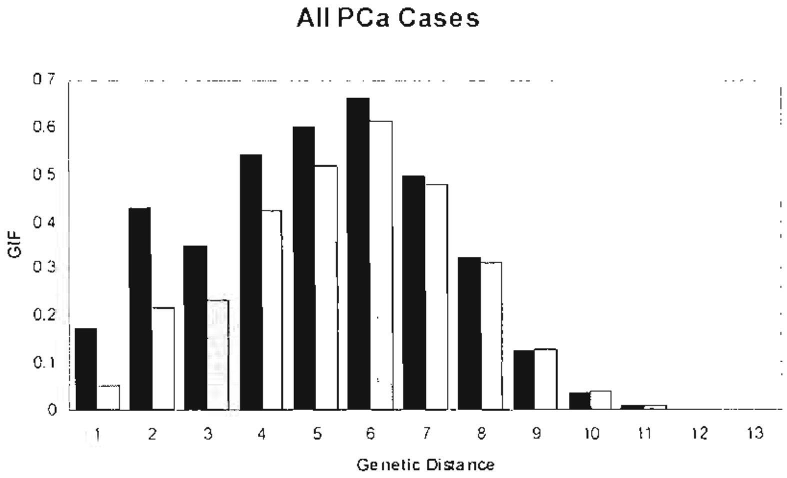
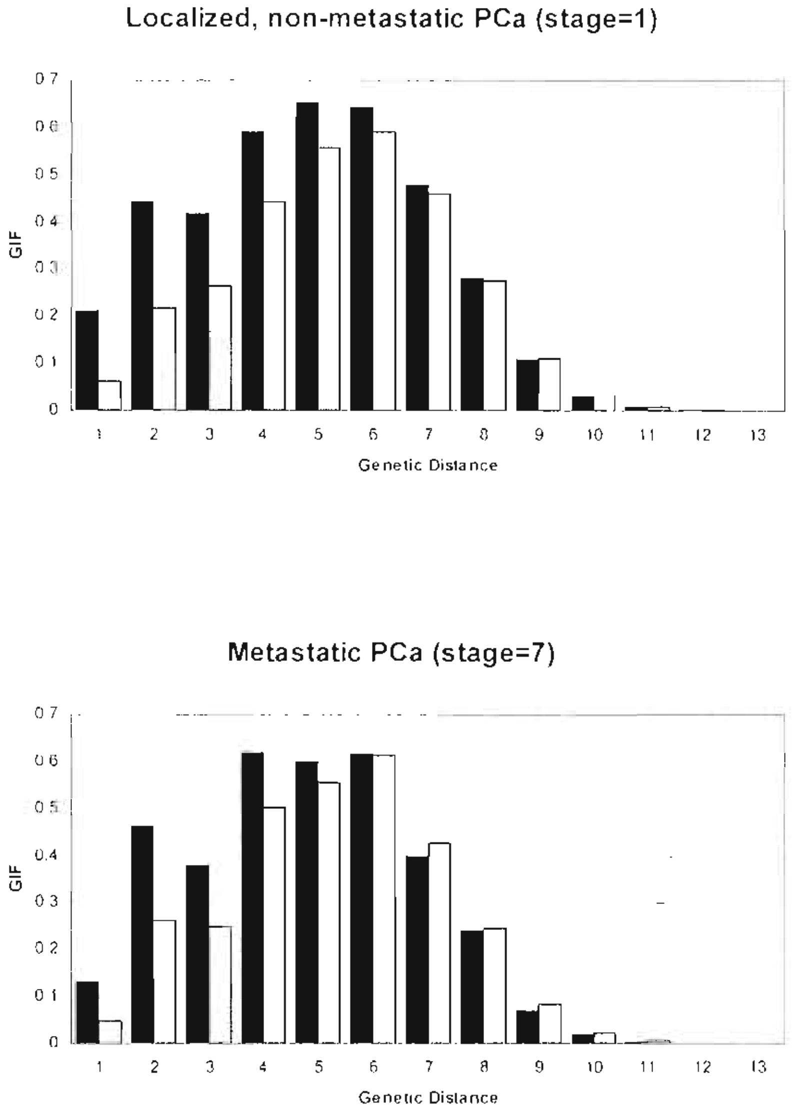


Figure A.1. Genealogical index of familiality.



Localized, non-metastatic PCa (stage=1)

Metastatic PCa (stage=7)

Figure A.1. continued.

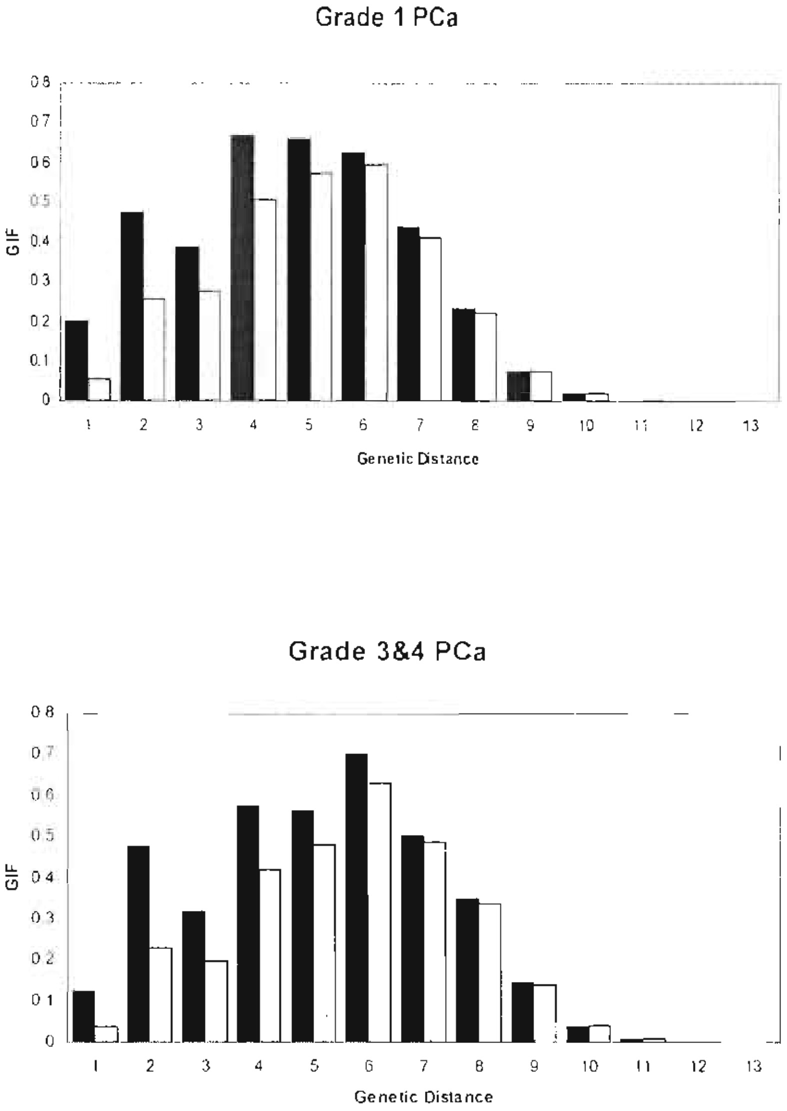
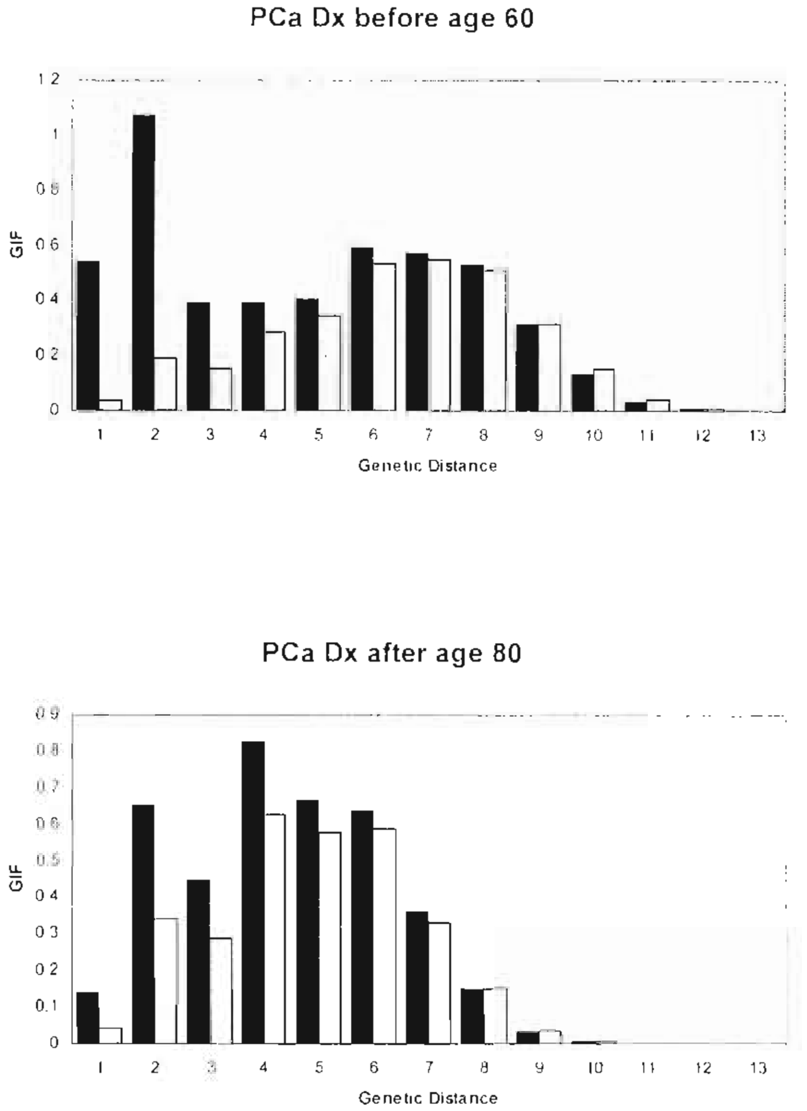


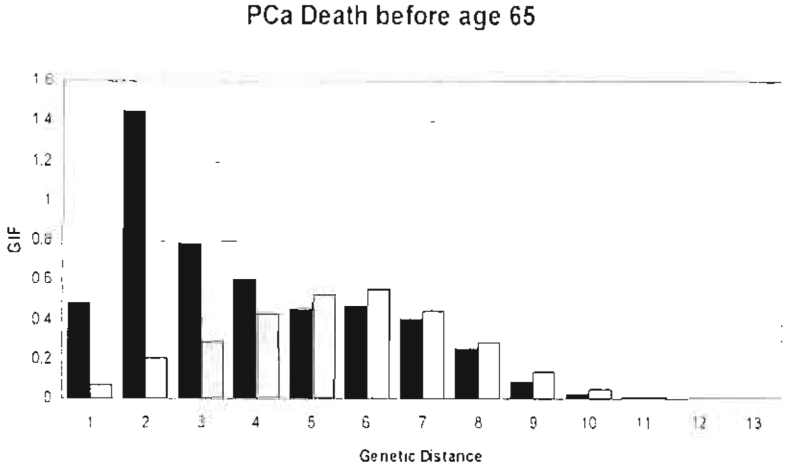
Figure A.1. continued.



PCa Dx before age 60

PCa Dx after age 80

Figure A.1. continued.



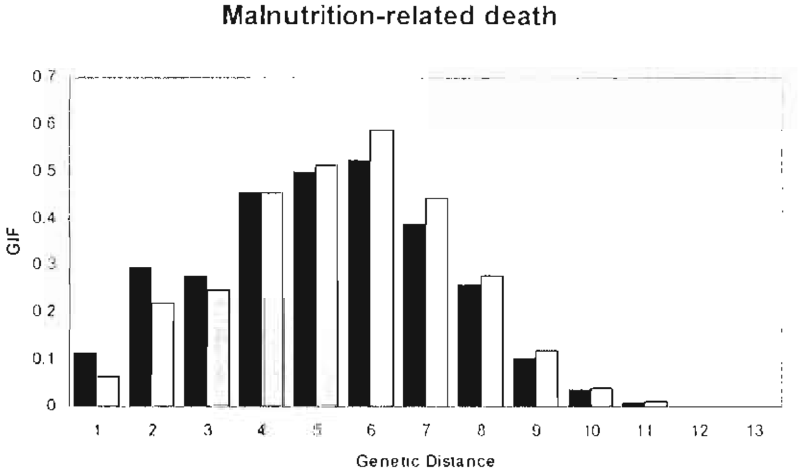


Figure A.1. continued.

Familial Relative Risk (FRR)

The resources of the UPDB allow us to make population-based estimates of relative risk for family members of individuals with specific phenotypes. Table A.7 shows the relative risk to first. second and third degree relatives of cases for developing the same PCa phenotype.

Table A.7: Summary of Familial Relative Risks for prostate cancer subgroups

Phenotype	Relationship	Subjects	Observed Cases	Expected Cases	FRR	95% CI
All PCa Cases		17,379				
	1°		5330	2762.8	1 93	1.88—1.98
	2°		5256	4071.4	1.26	1.26—1.33
	3°		9312	8402 1	1.11	1.09—1.13
Localized PCa (Stage=1)		6973				
	1°		1061	521 4	2 04	1.91—2 16
	2°		1293	898.4	1.44	1.36—1.52
	3°		2332	1979.1	1.18	1.13—1.23
Metastatic PCa (Stage=7)		1506				
	1°		51	27.3	1 87	1.39—2.46
	2°		62	40.8	1 52	1.16—1 95
	3°		161	137 3	1 17	0.99—1 37
PCa Dx before age 60		1426				
	1°		119	18.6	6.4	5.30—7 65
	2°		58	24	2 42	1.84—3.13
	3°		114	80	1 42	1 18—1 71
PCa Dx after age 80		3324				
	1°		322	157 5	2 04	1 83—2.28
	2°		288	203.8	1.41	1.25—1.59
	3°		809	700.1	1 16	1.08—1.24
Grade 1 PCa		7205				
	1°		1159	596.6	1 94	1 83—2 06
	2°		1193	922 7	1.29	1 22—1.37
	3°		2688	2271.5	1 18	1.14—1 23
Grade 3/4 PCa		3244				
	1°		238	110 4	2.16	1 89—2.45
	2°		224	154.7	1.45	1 26—1.65
	3°		655	519.7	1.26	1 17—1 36
PCa-related death		5378				
	1°		786	406 4	1.93	1.80—2.07
	2°		985	700 4	1 41	1 32—1.50
	3°		1534	1334.7	1.15	1.09—1 21
PCa death before age 65		456				
	1°		9	2.3	3 91	1 79—7 4
	2°		16	5.3	3 02	1 72—4.90
	3°		15	9.8	1.53	0.85—2.52
Malnutrition-related death		1481				
	1°		31	25.5	1 21	0 83—1 73
	2°		49	44	1 11	0.82—1.47
	3°		125	115.8	1.08	0.90—1.29

Conclusions

- All of the PCa subgroups examined show a significant familial component
- Best result was for early diagnosis group (age at Dx less than 60 years)
    - GIF = 4.99 was the second largest observed
    - FRR = 6.4 for first degree relatives was largest observed
    - FRR values for all relative groups were significantly higher than the values for general PCa
- Strong familiality for PCa-related death prior to age 65 is not observed in distant relatives.
    - May be the result of a small sample size
- Localized vs. Metastatic PCa cases
    - Metastatic PCa loses familial significance beyond first and second degree relatives
        - Result may be affected by relatively small sample size.
    - FRR values for localized cases were significantly higher than the values for general PCa in both second and third degree relatives

Our results show that early-onset PCa cases and those cases with localized disease have the strongest familial relationships. These two phenotypes may therefore be strong candidates for linkage analysis or other genetic testing to identify genes that are associated with prostate cancer.

APPENDIX B

GENETIC SUSCEPTIBILITY OF PROSTATE CANCER: GENOME-WIDE SCREEN OF "NON-AGGRESSIVE" DISEASE

Excerpt from a poster presented at the 2006 conference of the International Genetic Epidemiology Society

Genetic susceptibility of Prostate Cancer: Genome-wide screen of "non-aggressive" disease

by:

GB Christensen, NJ Camp, JM Farnham, LA Cannon-Albright

University of Utah Department of Biomedical Informatics

Background

Research has consistently shown that genetics plays a critical role in prostate cancer (CaP) development, but the identification of CaP genes has proven to be very difficult. Hereditary prostate cancer is a complex disease believed to involve numerous genes and variable penetrance. It has been proposed that studying alternative, highly homogenous phenotypes related to CaP may be a solution for overcoming the apparent heterogeneity that has hindered the identification of susceptibility genes. Several recent studies have applied this idea to "aggressive" or "clinically significant" cases of CaP. Using the resources of the Utah Population Database, we identified two phenotypes often associated with non-aggressive disease that show significant familiality. We present those results here.

Data Resource

- Utah Population Database (UPDB)
  - Records for approximately 2.2 million individuals
  - Up to 9 generations of genealogical data linking individuals into pedigrees
  - Linked to death certificates providing cause of death data since 1904

- Utah Cancer Registry (UCR)
  - Part of Surveillance, Epidemiology and End Results (SEER) program since 1973
  - All cancer events (except basal and squamous cell carcinomas) are recorded
  - Fully linked to UPDB
- 18,894 CaP cases from UCR currently linked to UPDB genealogies

Familial Relative Risk (FRR)

- The resources of the UPDB make it possible to make population-based estimates of relative risk for family members of individuals with specific phenotypes. Considering each CaP subgroup to be a unique condition, Table B.1 shows the relative risk to first, second and third degree relatives of cases for developing the same phenotype.
- All examined subgroups have a significant familial risk component.
- Non-metastatic disease shows a greater risk to extended family than general CaP.
- Cases diagnosed before age 65 and cases surviving more than 10 years have a risk significantly greater than general CaP for all three relative groups.

Table B.1: Summary of Familial Relative Risks for selected prostate cancer subgroups					
Phenotype		Observed	Expected		
Relationship	Subjects	Cases	Cases	FRR	95% CI
All CaP Cases	18,894				
1°		5400	2815.2	1.92	1.87—1.97
2°		5336	4145.4	1.29	1.25—1.32
3°		9397	8527.3	1.10	1.08—1.12
Localized (Non-Metastatic) CaP	7563				
1°		1081	531.1	2.04	1.92—2.16
2°		1316	916.6	1.46	1.36—1.52
3°		2357	2005.0	1.18	1.13—1.22
Regional or Distant Mets.	8974				
1°		1286	670.9	1.92	1.81—2.02
2°		1128	853.5	1.32	1.25—1.40
3°		3097	2692.7	1.15	1.11—1.19
CaP Survival < 5 years	6926				
1°		802	413.0	1.94	1.81—2.08
2°		980	723.1	1.30	1.27—1.44
3°		1834	1645.0	1.11	1.06—1.17
CaP Survival > 10 years	4786				
1°		556	218.3	2.55	2.34—2.77
2°		395	258.2	1.53	1.38—1.69
3°		1302	1007.7	1.29	1.22—1.36
CaP Dx before age 65	4094				
1°		401	121.2	3.31	2.99—3.65
2°		264	157.1	1.68	1.48—1.90
3°		642	514.3	1.25	1.15—1.35
FRR significantly greater than that for general CaP					

Linkage Analysis

Dominant and recessive parametric linkage analyses were performed for the CaP subgroups with survival of greater than 10 years and with localized tumors. All analyses were performed using the MCLINK software package at the Center for High Performance Computing at the University of Utah. Genotyping was performed by the Center for Inherited Disease Research (CIDR) on a full-genome set of 401 STR markers with an average spacing of 9 cM. A summary of the pedigrees is in Table B.2. HLOD tracings are shown in Figures B.1 and B.2.

Table B.2: Summary of pedigree characteristics for linkage analyses

	Phenotype		Summary data for the original CaP linkage resource used for this study.
	Non-metastatic CaP	Survival > 10 years	
Pedigrees	47	44	59
Cases	176	150	464
Mean age at dx (yrs)	69.0	66.6	69.7
Mean case survival (months)	112.9	172.7	106.2
Case subjects genotyped	86	115	246
Other genotyped*	676	594	640

* connecting ancestors of cases, and spouse with up to four children were genotyped when necessary to infer genotypes

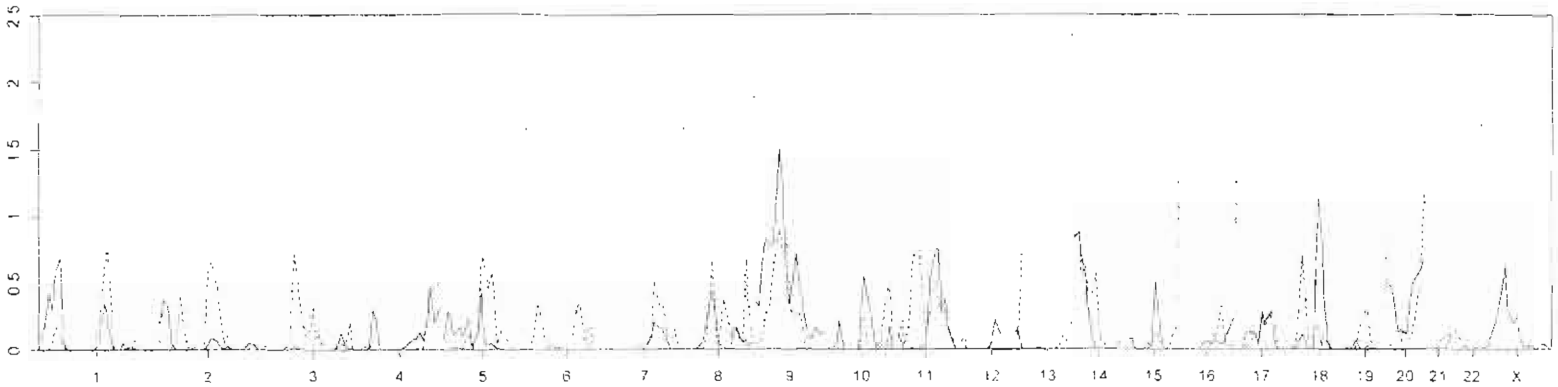


Figure B.1: HLOD statistic for linkage to non-metastatic CaP. The solid line represents the dominant model, and the broken line represents the recessive model.

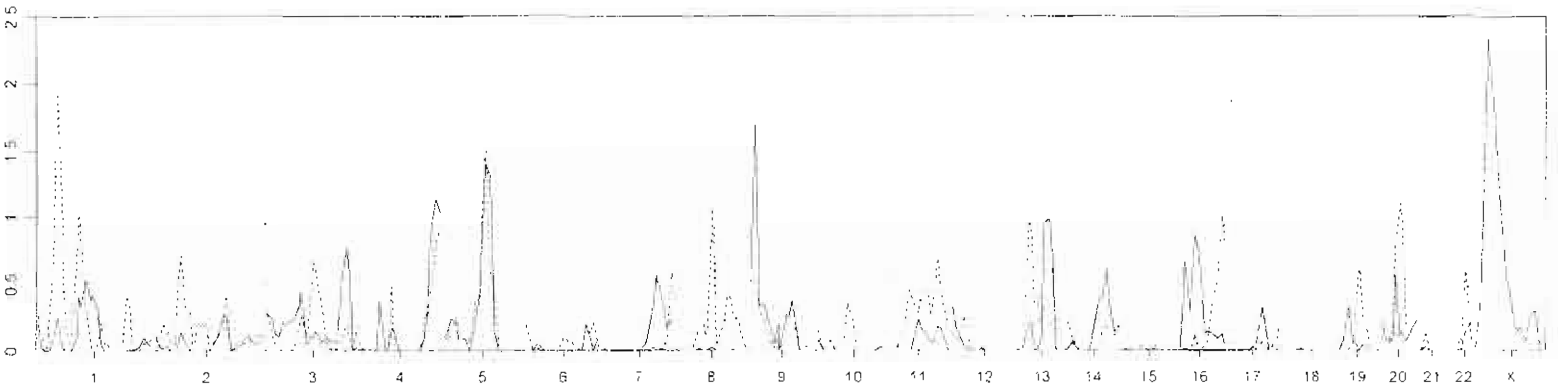


Figure B.2: HLOD statistic for linkage to CaP with survival of over 10 years. The solid line represents the dominant model, and the broken line represents the recessive model.

Discussion

- No significant linkage evidence was observed at the genome-wide level for either of the phenotypes examined.
- Best result for the non-metastatic subgroup was HLOD = 1.50 in the dominant analysis at 58 cM on chromosome 9.
- Best result for the long survival subgroup was HLOD = 2.33 in the dominant model at 40 cM on chromosome X.
  - Signal is at Xp21-22, and is not associated with the HPCX locus at Xq27-28.
- Long survival appears to be correlated with early age at diagnosis, which is generally considered to be a trait of hereditary CaP cases.
- The pedigrees and genotypes used in this study were originally ascertained for a linkage analysis of general prostate cancer. Considering only subgroups of the original cases results in fewer cases per pedigree and greater genetic distance between cases, increasing the possibility of confounding due to intra-familial heterogeneity.
- Further research is necessary to identify the genes responsible for hereditary prostate cancer and surmount the overarching problem of CaP heterogeneity.

Acknowledgements

Data collection for this research was supported by NIH National Cancer Institute grant R01 CA90752 (to L.A. Cannon Albright); a subcontract from Johns Hopkins University with funds provided by grant R01 CA89600 from the NIH National Cancer Institute (to L.A. Cannon Albright); NIH grant K07 CA98364 (to N.J. Camp); and NIH NLM training grant T15 LM07124 (to G.B. Christensen). An allocation of computer time from the Center for High Performance Computing at the University of Utah is gratefully acknowledged. The computational resources for this project have been provided by the National Institutes of Health (Grant # NCRR 1 S10 RR17214-01) on the Arches Metacluster, administered by the University of Utah Center for High Performance Computing. Research was supported by the Utah Cancer Registry, which is funded by Contract N01-PC-35141 from the National Cancer Institute with additional support from the Utah State Department of Health and the University of Utah. Partial support for all datasets within the Utah Population Database (UPDB) was provided by the University of Utah Huntsman Cancer Institute. This investigation was supported by the Public Health Services research grant number M01-RR00064 from the National Center for Research Resources. Genotyping services were provided by the Center for Inherited Disease Research (CIDR). CIDR is fully funded through a federal contract from the National Institutes of Health to The Johns Hopkins University, contract number N01-HG-65403.

APPENDIX C

COMBINED GENOME-WIDE ASSOCIATION AND LINKAGE ANALYSIS OF EXTENDED UTAH PROSTATE CANCER PEDIGREES IDENTIFIES SIGNIFICANCE AT CHROMOSOME 8q12

Excerpt from a poster presented at the 2008 conference of the International Genetic Epidemiology Society

Combined genome-wide association and linkage analysis of extended Utah prostate cancer pedigrees identifies significance at chromosome 8q12

GB Christensen, J Farnham, NJ Camp, LA Cannon-Albright

University of Utah School of Medicine Department of Biomedical Informatics, Salt Lake City, UT

Abstract

We performed genome-wide linkage and case/control association studies in 27 prostate cancer cases from 3 extended, informative, high-risk Utah pedigrees. All relationships between cases were more distant than first degree. Genotyping was performed with the Illumina 550k SNP array, after exclusion of 58,000 markers failing quality control. For controls, we selected caucasians from the Illumina iControl data set (n=1,579), also genotyped for the 550k SNPs.

A naive Fisher's Exact Test was used for the initial association screen, ignoring the familial relationships between cases, under three models: dominant, recessive, and an allele test. Fifty-four distinct markers were selected for secondary screening with a significance cut off of p<1e-5. Secondary screening was performed using Genie software, which included known relationships between cases. In the secondary screen, 1 marker reached the genome-wide significance threshold of p<3.4e-7. This marker was on chromosome band 8q12.3 (p=1e-7). Five of the top 8 associations from the secondary screening were also at 8q12.3. Other regions with markers reaching p<3e-6 included: 4p13, 2p25, 7p21, 17q22, and 21q21.

We also performed linkage analysis in the 3 pedigrees using 27,157 SNPs from the Illumina 550k set. Under the Smith (1996) inheritance model, two regions showed suggestive evidence of linkage; chromosome band 2p15 (hetLOD=2.44), and chromosome band 8q12-q21 (max hetLod = 2.28). The SNPs showing significant evidence for association are located within the linkage peak, but were not used as part of the linkage analysis.

Data

- Cases
  - 27 prostate cancer cases from two extended Utah pedigrees (Figure C.1)
  - Mostly 3-5 generations separated
  - Subjects originally ascertained for shared genomic segment analysis
- Controls
  - Caucasians from Illumina iControl database
  - N=1579
  - Male and female controls used
- Genotyping
  - All cases and controls genotyped with Illumina 550k SNP array

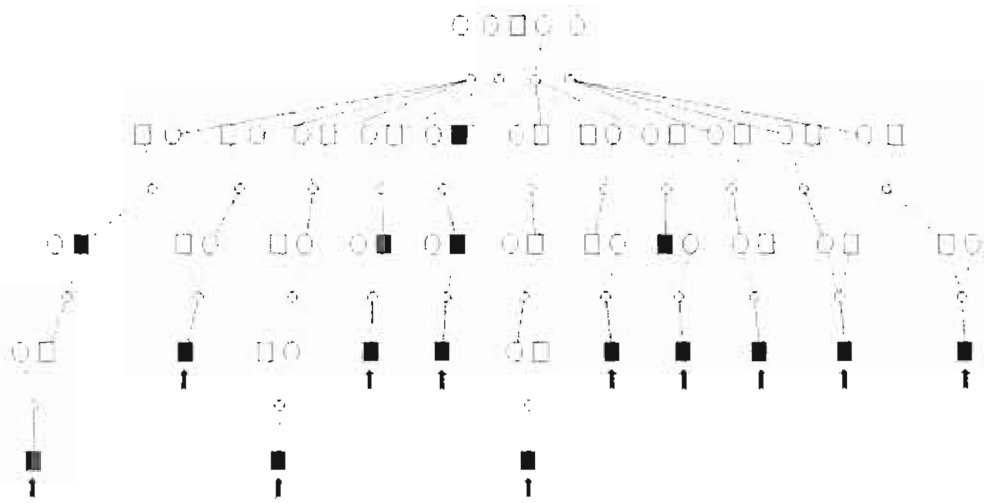


Figure C.1. Sample Pedigree. Dark boxes indicate known prostate cancer cases. Black arrows indicate genotyped subjects.

Analyses

- Association
    - Quality control
        - HWE (controls only, p<0.05)
        - All 3 genotypes must be observed in cases and controls
        - Minimum individual call rate 98%
        - Minimum marker call rate 98% in cases and controls
    - Stage1
        - Naïve Fisher exact test run on all SNPs (assuming independence)
        - Dominant, recessive, and allele tests
        - SNPs advanced to stage 2 if p<1e-5
    - Stage 2
        - Analysis with GENIE
        - Empirical significance test accounting for familial relationships
        - Significance thresholds set as in Hoggart et al. 2008
- Linkage
    - 27,157 SNPs selected from Illumina 550k set
        - Minimum spacing 0.1 cM
        - Minimum heterozygosity of 0.3
        - Maximum R^2=0.16 within 5Mb window
    - Smith (1998) inheritance model
    - MCLINK software package

Results

- Linkage (Figure C.2)
    - Two suggestive linkage peaks
        - HLOD=2.44 at Chromosome 2p15
        - HLOD=2.28 at Chromosome 8q12-21
- Association
    - Stage 1
        - 54 unique SNPs from 73 analyses passed threshold
    - Stage 2 (Table C.1)
        - 1 SNP significant at threshold of 3.4e-7
            - rs975847 at 8q12.3, p=1.0e-7
- Chromosome 8q12.3
    - Identified in both linkage and association analyses
    - Location of best overall association result and four of the top five results from the second stage
    - Nine SNPs from a 217 kb region passed first stage of association
        - None of these SNPs were included in the linkage set
        - No genes are located within the 217 kb span
        - Region includes RNA sequence AL137390, expressed in testis

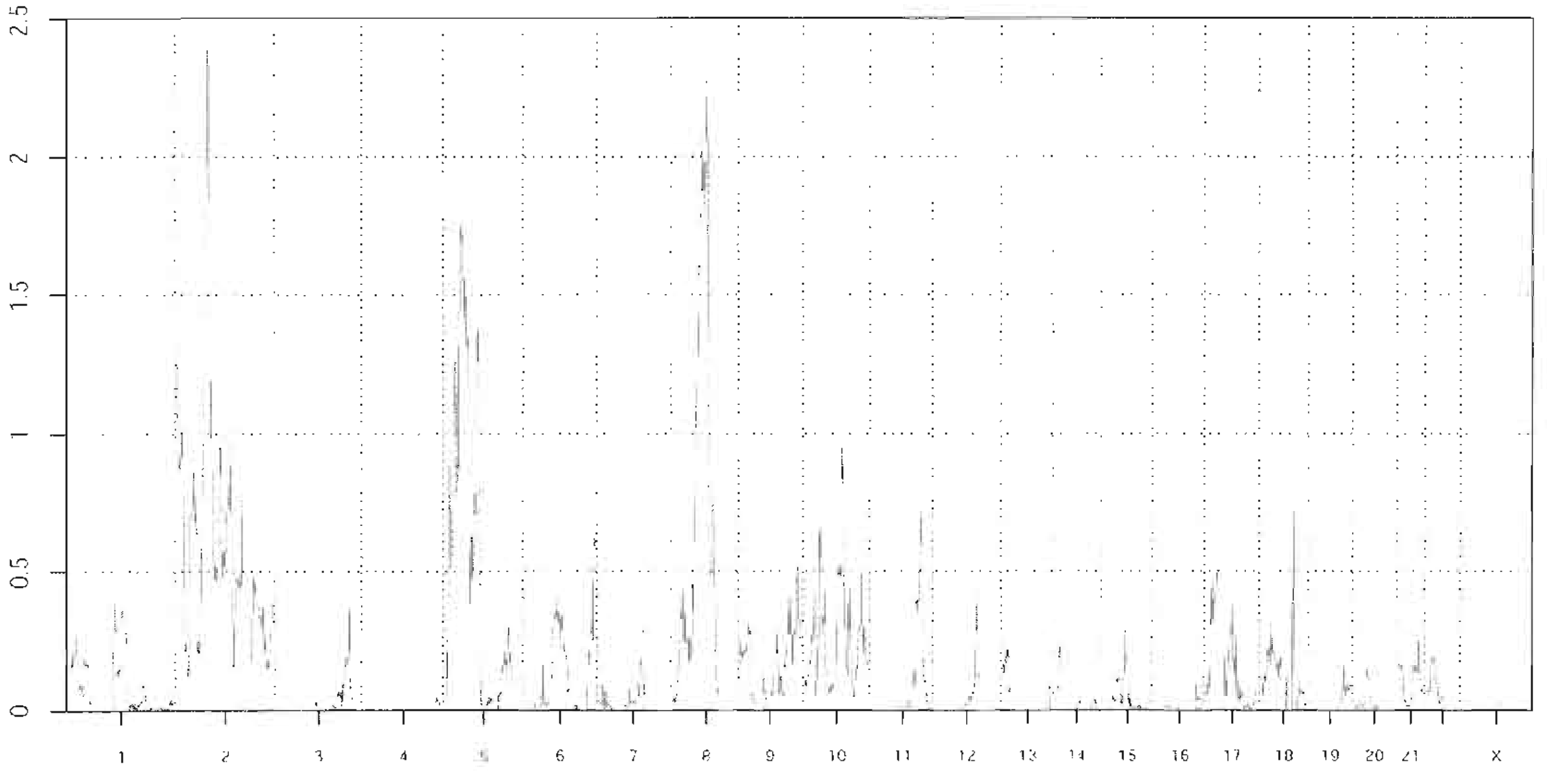


Figure C.2. Genome-wide HLOD results. The inset shows detail from chromosome 8, with an arrow indicating the relative position of the significant association finding.

Table C.1. GWA Secondary Screen

SNP	Location	Test	Emp. P-val	Notes
rs975847	8q12.3	Dominant	1.00e-7	
rs1347901	4p13	Dominant	6.00e-7	
rs6471975	8q12.3	Allele	9.56e-7	
rs823422	8q12.3	Dominant	1.30e-6	
rs344248	8q12.3	Allele	1.31e-6	
rs768447	2p25	Allele	2.00e-6	
rs2107280	7p21.3	Dominant	2.00e-6	NXPH1 intron
rs344210	8q12.3	Allele	2.46e-6	
rs8070264	17q22	Dominant	2.80e-6	
rs2826745	21q21	Dominant	2.90e-6	NCAM2 intron

Results shown for SNPs with observed P<3e-6. Only the best test result is shown for each SNP.

Discussion

- This small study shows the power and synergistic utility of using both linkage and association analysis in high risk pedigrees
- Genome-wide association and linkage can be performed with one set of genotype data, reducing costs and improving efficiency
- The 8q12 region identified as significant for prostate cancer predisposition has not been previously reported for linkage or association, but is recognized for LOH.
- No association evidence was observed at 8q24 prostate cancer locus

Acknowledgements

- Research was supported by the Utah Cancer Registry, which is funded by contract N01-PC-35141 from the National Cancer Institute's SEER program with additional support from the Utah State Department of Health and the University of Utah.
- Partial support for all datasets within the Utah Population Database (UPDB) was provided by the University of Utah Huntsman Cancer Institute.
- Some research support provided by the University of Utah General Clinical Research Center, Grant # M01 RR00064.
- International Consortium for Prostate Cancer Genetics: a subcontract from Johns Hopkins University with funds provided by grant R01 CA89600 from the NIH National Cancer Institute (to L.A. Cannon-Albright)
- GB Christensen supported by NIH NLM training grant T15 LM07124

APPENDIX D

R PROGRAM CODE FOR sumLINK PROCEDURE

slkpoint.r

```
###slkpoint.r
###B Christensen, 4/20/2009
###Calculates observed values for sumLINK and sumLOD
###Outputs results to file "points.out"
###Requires input files for each chromosome with pedigree name
###in first column, and LOD scores in subsequent columns.
###All input files must have pedigrees in same order.
###Creates basic plots of both stats


#####function zeroes out negative values
isgtz=function(x){
if(x <= 0) 0
else x
}

#####function zeroes out non-sig values (p<0.05)
issig=function(x){
if(x < 0.588) 0
else x
}

####Start chromosome loops####
for(ch in 1:22){
fil=paste("lods/c",ch,".lods.cm",sep="")  ###define data file path

d22 = read.table(fil)[,-1]      ##read file, remove first column
cm = dim(d22)[2]                ##count LOD observations (may be 1 cM
increments)

####Add basic lod scores, sumlod, and sumlink
####Count positive lods and linked lods at each point

ctpos=colSums(d22>0)      ##total peds w/ positive LOD
ctlnk=colSums(d22>=0.588)     ##total peds w/ significant LOD
sld=colSums(apply(d22,c(1,2),'isgtz'))  ##calculate sumLOD
slk=colSums(apply(d22,c(1,2),'issig'))  ## calculate sumLINK
lod=colSums(d22)   ##calculate LOD

##store chromosome data in object count*
chrom=rep(ch,cm)
pos=1:cm
assign(paste("count",ch,sep=""),as.data.frame(cbind(chrom,pos,ctpos,
ctlnk,slk,sld,lod)))

} #end chromosome loops#

###combine all chromosomes###
points=rbind(count1,count2,count3,count4,count5,count6,count7,count8,
count9,count10,count11,count12,count13,count14,count15,count16,
count17,count18,count19,count20,count21,count22)
```

```
###write to file
write.table(points,"points.out",quote=FALSE,row.names=FALSE)


##############################
##Make basic sumLINK Plots####
##############################

###Create vector "d2" with cumulative cM positions
ch1=numeric(22)
for(ch in 1:22){ch1[ch]=max(points$pos[points$chrom==ch])}
ch12=numeric(23)
ch12[1]=0
ch12[2:23]=cumsum(ch1+10)

d2=numeric(length(points$pos))
for(ch in 1:22){
d2[points$chrom==ch] = points$pos[points$chrom==ch]+ch12[ch]
}

##create vector "t1", with positions for axis tick marks###
t1=numeric(22)
for(i in 1:22){
         t1[i]=median(d2[points$chrom==i])
}

###create vector "vert" with positions of vertical chromosome
separators
vert=ch12-5


##Plot SumLINK###
pdf("SumPlot1.pdf",height=7.5,width=10)
par(mfrow=c(2,1), mar=c(3,1,1,0.5));

plot(d2,points$slk,type="n",ylim=c(0,15), xaxt="n", xlab="", ylab="")
mtext("sumLINK",2,0)
axis(side=1,labels=1:22, at=t1, cex.axis=0.75)
segments(vert[2:23],-100,vert[2:23],1000,lty=3,lwd=0.5)
for(i in 1:22){
lines(d2[points$chrom==i],points$slk[points$chrom==i],lty=1, lwd=0.5)
         }

##Plot SumLOD
plot(d2,points$sld,type="n",ylim=c(10,30),xaxt="n", xlab="", ylab="")
mtext("sumLOD",2,0)
axis(side=1,labels=1:22, at=t1,cex.axis=0.75)
segments(vert[2:23],-100,vert[2:23],1000,lty=3,lwd=0.5)
for(i in 1:22){
lines(d2[points$chrom==i],points$sld[points$chrom==i],lty=1, lwd=0.5)
         }

dev.off()
```

FindPeaks.r

```
###FindPeaks.r
###B. Christensen 4/20/2009
###Program reads "points.out" and finds peaks for
###observed sumLOD and sumLINK statistics
###Peak values saved to files


points=read.table("points.out",header=TRUE)

####SumLINK Peaks

slkbase=quantile(points$slk,.90)
max4=points[points$slk>slkbase,]

t1=NULL
peak=NULL

for(i in 1:(dim(max4)[1]-1)){
       if(abs(max4$pos[i+1]-max4$pos[i]) < 10 &&
       max4$chrom[i]==max4$chrom[i+1]){
              t1=rbind(t1,max4[i,])
       }#end if
       else{
              t1=rbind(t1,max4[i,])
              t2=t1[t1$slk==max(t1$slk),]
              peak=rbind(peak,t2[1,])
              t1=NULL
       }#end else
}#end for

##finish last loop
t1=rbind(t1,max4[i,])
t2=t1[t1$slk==max(t1$slk),]
peak=rbind(peak,t2[1,])
t1=NULL

outpeak=cbind(peak$chrom, peak$pos, peak$slk)
write.table(outpeak,"sumLinkPeaks.t",row.names=FALSE,quote=FALSE,
col.names=TRUE)


####SumLOD Peaks

sldbase=quantile(points$sld,.90)
max4=points[points$sld>sldbase,]

t1=NULL
peak=NULL

for(i in 1:(dim(max4)[1]-1)){
       if(abs(max4$pos[i+1]-max4$pos[i]) < 10 &&
       max4$chrom[i]==max4$chrom[i+1]){
              t1=rbind(t1,max4[i,])
```

```
          }#end if
       else{
             t1=rbind(t1,max4[i,])
             t2=t1[t1$sld==max(t1$sld),]
             peak=rbind(peak,t2[1,])
             t1=NULL
        }#end else
}#end for

##finish last loop
t1=rbind(t1,max4[i,])
t2=t1[t1$sld==max(t1$sld),]
peak=rbind(peak,t2[1,])
t1=NULL

outpeak=cbind(peak$chrom, peak$pos, peak$sld)

write.table(outpeak,"[illegible]",row.names=FALSE, quote=FALSE,col.na
mes=TRUE)
```

rshuff2.r

```
###rshuff2.r
###B. Christensen 4/20/2009
###Reads in raw LOD score data, performs shuffling,
###calculates sumLOD and sumLINK from shuffled data,
###writes a result file for each set "point*"


nshff=1000                 #number of shuffles#
pts="../points.out"        #path to points.out#


###Read in "points.out" to determine genome length
points=read.table(pts,header=TRUE)
genl=dim(points)[1]


###################################################
###READ IN LOD SCORES AND RANDOMIZE CHR ORDER######
###################################################
##adjust data file path as necessary


r1=read.table("../lods/c1.lods.cm")
r2=read.table("../lods/c2.lods.cm")
r3=read.table("../lods/c3.lods.cm")
r4=read.table("../lods/c4.lods.cm")
r5=read.table("../lods/c5.lods.cm")
r6=read.table("../lods/c6.lods.cm")
r7=read.table("../lods/c7.lods.cm")
r8=read.table("../lods/c8.lods.cm")
r9=read.table("../lods/c9.lods.cm")
r10=read.table("../lods/c10.lods.cm")
r11=read.table("../lods/c11.lods.cm")
r12=read.table("../lods/c12.lods.cm")
r13=read.table("../lods/c13.lods.cm")
r14=read.table("../lods/c14.lods.cm")
r15=read.table("../lods/c15.lods.cm")
r16=read.table("../lods/c16.lods.cm")
r17=read.table("../lods/c17.lods.cm")
r18=read.table("../lods/c18.lods.cm")
r19=read.table("../lods/c19.lods.cm")
r20=read.table("../lods/c20.lods.cm")
r21=read.table("../lods/c21.lods.cm")
r22=read.table("../lods/c22.lods.cm")

npeds=dim(r1)[1]  ###number of pedigrees

for(repl in 1:nshff){

rx=NULL
for(i in 1:npeds){
rx=cbind(rx,rank(runif(22),ties.method="random"))
}
```

```
###OPTIONAL to write out chrom order and/or read in saved chrom order
##write.table(rx,paste("../chord/CHord",rept,sep=""),quote=FALSE,
row.names=FALSE,col.names=FALSE)
##rx=read.table(paste("CHord",rept,sep=""))


rfull=NULL
for(i in 1:npeds){
x1=NULL
        for(j in 1:22){
        z1=eval(as.name(paste("r",rx[i,j],sep='')))
        x1=c(x1,as.numeric(z1[i,-1]))
        } ##end for j

rfull=rbind(rfull,x1)
} ##end for i


##################################################################
###RECENTER GENOME-WIDE LODSCORES#############################
##################################################################


shift=ceiling(runif(npeds,0.00001,genl))

d22=NULL
for(i in 1:npeds){
z=shift[i]
x1=rfull[i,]
x2=c(x1[z:genl],x1) ##(uneven lengths cut off below)
d22=rbind(d22,x2[1:genl])
}


#################################################################
####CALCULATE AND OUTPUT SUMMARY STATISTICS######################
#################################################################


#####function zeroes out negative values
isgtz=function(x){
if(x <= 0) 0
else x
}

#####function zeroes out non-sig values
issig=function(x){
if(x < 0.588) 0
else x
}
```

```
####Add up sumlod and sumlink
####Count positive lods and linked lods at each point


ctpos=colSums(d22>0)
ctlnk=colSums(d22>=0.588)
sld=colSums(apply(d22,c(1,2),'isgtz'))
slk=colSums(apply(d22,c(1,2),'issig'))
lod=colSums(d22)

sld=round(sld,5)
slk=round(slk,5)
lod=round(lod,5)

summ = as.data.frame(cbind( 1:geni,lod,sld,slk,ctlnk,ctpos))

write.table(summ,paste("point", rept, sep=""),row.names=FALSE)

rm(rFull)
rm(d22)
rm(rx)
}

quit("no")  ###close R, don't save workspace
```

testPeaks95p1.r

```
###testPeaks95p1.r
###B. Christensen 4/20/2009
###Identifies peaks from shuffled data results and
###determines significance of observed peaks and
###writes out result files


nshff=1000    ###number of simulations


####Read in observed points and peaks

points=read.table("points.out",header=TRUE)
gen1=dim(points)[1]

sldpeaks=read.table("sumlodPeaks2",header=TRUE)
slkpeaks=read.table("sumlinkPeaks2",header=TRUE)

slkbase=quantile(points$slk,0.95)
sldbase=quantile(points$sld,0.95)


######Read shuffled genomes and select peaks####

sldsims=NULL
slksims=NULL

for(i in 1:nshff){

path=paste("sims/point",i,sep="")
print(path)
sims=as.data.frame(cbind(rep(i,gen1),read.table(path,header=TRUE)))
names(sims)[1]="sim"
names(sims)[2]="locus"



#####Find pointwise sumLink peaks over [threshold] in full simulation
set

max4=sims[sims$slk>slkbase,]

if(dim(max4)[1]==1){slksims=rbind(slksims,max4)}

if(dim(max4)[1]>1){
t1=NULL
peak=NULL

for(i in 1:(dim(max4)[1]-1)){
if(abs(max4$locus[i+1]-max4$locus[i]) < 2 &&
max4$sim[i]==max4$sim[i+1]){
        t1=rbind(t1,max4[i,])
        }#end if
else{
t1=rbind(t1,max4[i,])
t2=t1[t1$slk==max(t1$slk),]
peak=rbind(peak,t2[1,])
t1=NULL
          }#end else
}#end for
##finish last loop
t1=rbind(t1,max4[i,])
t2=t1[t1$slk==max(t1$slk),]
peak=rbind(peak,t2[1,])
t1=NULL

slksims=rbind(slksims,peak)

       }###end if max>1


####Find pointwise summed peaks over threshold in full simulation
set

max4=sims[sims$sld>sldbase,]

if(dim(max4)[1]==1){sldsims=rbind(sldsims,max4)}

if(dim(max4)[1]>1){
t1=NULL
peak=NULL

for(i in 1:(dim(max4)[1]-1)){
if(abs(max4$locus[i+1]-max4$locus[i]) < 2 &&
max4$sim[i]==max4$sim[i+1]){
        t1=rbind(t1,max4[i,])
         }#end if
else{
t1=rbind(t1,max4[i,])
t2=t1[t1$sld==max(t1$sld),]
peak=rbind(peak,t2[1,])
t1=NULL
          }#end else
}#end for
##finish last loop
t1=rbind(t1,max4[i,])
t2=t1[t1$sld==max(t1$sld),]
peak=rbind(peak,t2[1,])
t1=NULL

sldsims=rbind(sldsims,peak)

       }###end if max>1


}###end of shuffs
```

```
########################
####assess significance
########################

###sumlink significance

ksig=function(x){dim(slksims[slksims$slk>=x,])[1]/nshff}

slksig=numeric(length(slkpeaks$slk))
for(k in 1:length(slkpeaks$slk)){slksig[k]=ksig(slkpeaks$slk[k])}

dest="sumlinkSig95p1"
slkpeaks$slk=round(slkpeaks$slk,3)
write.table(cbind(slkpeaks,slksig)[slksig<=5,],dest,quote=FALSE,row.n
ames=FALSE,sep="\t")



###sumlod significance

dsig=function(x){dim(sldsims[sldsims$sld>=x,])[1]/nshff}

sldsig=numeric(length(sldpeaks$sld))
for(k in 1:length(sldpeaks$sld)){sldsig[k]=dsig(sldpeaks$sld[k])}

dest="sumlodSig95p1"
sldpeaks$sld=round(sldpeaks$sld,3)
write.table(cbind(sldpeaks,sldsig)[sldsig<=5,],dest,quote=FALSE,row.n
ames=FALSE,sep="\t")

slkbase
sldbase
quit("no")
```