GENETIC PREDISPOSITION TO PROSTATE CANCER: THE CONTRIBUTION OF THE HPCX LOCUS AND TGFB1 GENE

By

Brian L. Yaspan

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Cancer Biology

May, 2008

Nashville, Tennessee

Approved:

Professor Jeffrey R. Smith

Professor Robert J. Matusik

Professor William D. Dupont

Professor Scott M. Williams

To Lisa A. Weiss, my fiancée, who brings a smile to my face.

ACKNOWLEDGEMENTS

This work would not have been possible without the following financial support: a predoctoral fellowship from the Department of Defense (W81XH-06-1-0057 to Brian L. Yaspan), a Veterans Administration Merit Award (to Jeffrey R. Smith) and NIH/NCI Training Grant CA09592 (to Lynn Matrisian).

Thank you to my dissertation committee at the Vanderbilt University Medical Center for their support of me and guidance of my project: Jeffrey R. Smith (mentor), Robert Matusik (chair), William Dupont and Scott Williams. I would like to thank Ambra Pozzi, Roy Zent and Michelle Southard-Smith for their advice as I progressed in my graduate school career.

I would like to thank the men comprising our study population and their attending physicians for their generosity. Without you there would be no study.

I would like to say thank you to the many people who have aided me in the pursuit of my degree with their expertise, intellectual conversations, and kindness:

The members of the Smith laboratory: Kevin Bradley, Joan Breyer, Bradford Elmore and Kate McReynolds, and past lab members Isaac Amundson and Lelia Davis.

My family in Agoura Hills, San Diego and Davis, CA: Judy Fradkin, Amie Fradkin, Arthur Fradkin, Gary Fradkin, Reggie Fradkin, Matthew Fradkin and Kyle Fradkin. I would also especially like to thank my grandmother Shirley Fradkin who passed away during my time as a student.

My family in Nashville, TN: Natalie Levy, Rick Levy, Susan Levy, Bob Levy, Jeanmarie Levy, Marjorie Levy, Greg Levy, Madeline Levy, and the Wolpert family. I would like to thank my uncle Donald Levy who passed away during my graduate career.

My graduate student colleagues at Vanderbilt University and friends in Nashville, TN: Kelly Chandler, Ron Chandler, Shalyn Claggett, Robert Geil, Vince Gerbasi, Scott Gruver, Kristen Guglielmi, Melissa Hull, Justin Layer, Seth Ogden, Antonio Perez, Ines Macias-Perez, and Xiufeng Song.

My friends who have come from afar to visit: Stan Chia, Fan-li Chou, T.J. Cox, Aaron Fichtelberg, Dara Ghahremani, Larry Goldfinger, Anouk Jevtic, Vladimir Jevtic, Justin Maisonet, David Quinn, Eric Shilland, Melinda Soderstrom, Juliana Choy Sommer, Maritza Stanchich, Darren Su, Rachel Teichman, Nancy Vesper, Jill Weiner, Ken Weiner, Eric Weiss, and Stephanie Whitman.

Finally, a graduate student's life is an incredibly rewarding but sometimes stressful experience. During this time, I have found few things more relaxing than

watching a sleeping cat. For this, I would like to say thank you to my cats, Zina and Monty.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AD.....Autosomal dominant

AR..... Autosomal recessive

bp.....Base pair

CAPB ..... Cancer of the prostate and brain

CDCV .....Common disease common variant hypothesis

CEPH .....Centre d'Etude du Polymorphisme Humain

CEU.....Centre d'Etude du Polymorphisme Humain from Utah

CI.....Confidence interval

cM .....Centimorgan

dbSNP ..... NCBI public SNP repository

dNTP .....Deoxyribonucleotide triphosphate

EM.....Expectation-maximization

FP ..... Florescence polarization

GDB ..... Human genome database

GWAS..... Genome-wide association study

HLOD .....Heterogeneity lod

HPC..... Hereditary prostate cancer

HPC1.....Hereditary prostate cancer – Chromosome 1

HPC2.....Hereditary prostate cancer – Chromosome 2

HPC20.....Hereditary prostate cancer – Chromosome 20

HPCX ..... Hereditary prostate cancer – Chromosome X

htSNP ..... Haplotype-tagging SNP

HWE ..... Hardy-Weinberg equilibrium

ICPCG ..... International consortium for prostate cancer genetics

kb ..... Kilobase

kD ..... Kilodalton

LD ..... Linkage disequilibrium

LOD ..... Logarithm of the odds

LOH ..... Loss of heterozygosity

MAF ..... Minor allele frequency

Mb ..... Megabase

NPL-LOD ..... Non-parametric lod

OR ..... Odds ratio

PCAP ..... Predisposing for cancer of the prostate

PCR ..... Polymerase chain reaction

PrCa ..... Prostate cancer

PROGRESS ..... Prostate cancer genetic research study

PSA ..... Prostate specific antigen

RR ..... Relative risk

SBCS ..... Shanghai breast cancer study

SNP ..... Single nucleotide polymorphism

SSCP ..... Single strand conformation polymorphism

STR ..... Simple (or short) tandem repeat

CHAPTER I

INTRODUCTION

Overview

Prostate cancer (PrCa) is the most commonly diagnosed non-skin malignancy in males, with an estimated 1 in 6 men diagnosed during their lifetime[1]. It is estimated that 218,890 men will be diagnosed and 27,050 men will die from PrCa in 2007[1]. PrCa is the second leading cause of death by cancer in United States males after lung cancer. The only well-established risk factors for PrCa include age, family history and race. African Americans have a 60% higher incidence rate and a 200% higher mortality rate than that of Caucasians. All other racial/ethnic groups in North America have lower rates than African Americans, with the lowest incidence rates found in Native Americans[2]. PrCa is usually found in men over 50 years of age with about two-thirds of cases occurring after age 65. Environmental exposure risks remain unclear, but diet, occupational chemical exposures, sexually transmitted diseases and chronic prostatitis have all been implicated[3].

There is a significant genetic component to PrCa predisposition. Increased familial relative risk (RR) is observed across multiple populations (European Caucasian, Asian-American, African-American and Caucasian American)[4]. Males have a two- to three-fold increased risk of developing PrCa if they have a first or second degree relative

with PrCa[5]. Furthermore, twin concordance studies reveal a higher heritable risk for PrCa than for any other common cancer, with one study estimating heritable risk accounting for approximately 57% of variability in liability to PrCa in twins[6;7].

Genetics of Common Cancer

The first observation of inherited predisposition to common cancer dates to 100 A.D. when an unnamed Roman physician documented increased occurrence of breast cancer within a family[8]. Unfortunately, little scientific progress was made to further this observation for 1,700 years. In the late 19th century French neurologist Pierre Paul Broca documented the breast cancer occurrence in his wife's family; the cause of death for 10 of 24 women in the family over four generations (Figure 1)[8;9]. Broca also noted a high frequency of other cancers in his wife's family and surmised this observation was due to an inherited risk factor. Other reports describing families with apparent inherited predisposition to common cancers followed around this time[10;11]. Organized scientific investigation of familial aggregation of cancers did not begin until the 1960's with seminal studies by Lynch *et al*, Li and Fraumeni and Knudson[12-14].

Prostate Cancer Genetic Epidemiology

Familial Aggregation Studies

Systematic epidemiologic studies of the familial aggregation of common cancers began in the 1960's and focused heavily on cancers of the breast and colon. Unfortunately, PrCa was not the subject of intense investigation in these early studies. In

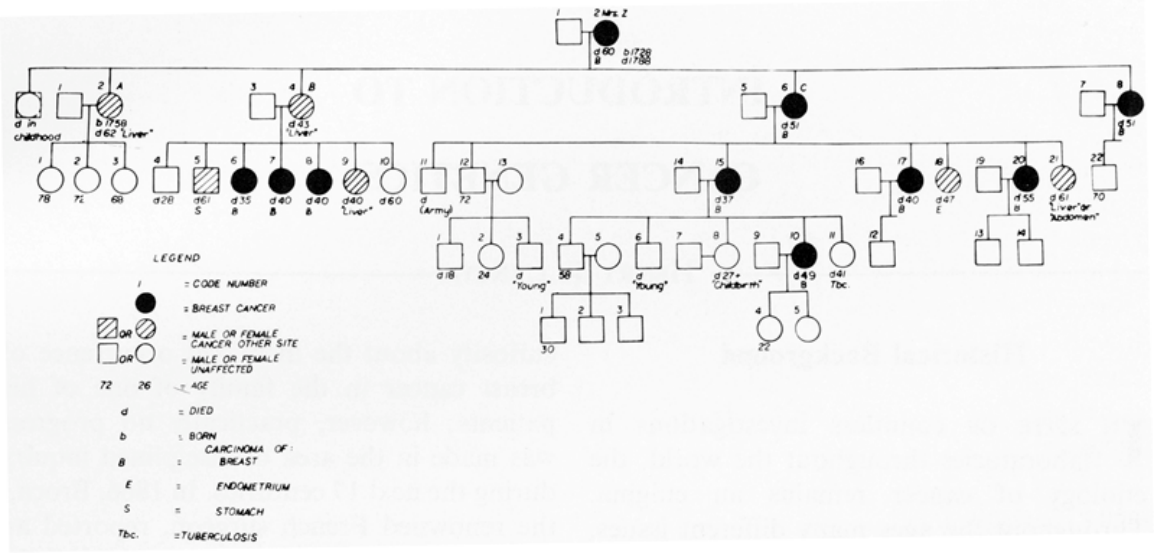


Figure 1. Pedigree of breast cancer family described by Broca in 1866 as reconstructed by Henry T. Lynch, et al. Figure taken from (8)

fact, an early review of PrCa epidemiology published in 1963 specifically noted the dearth of systematic epidemiologic investigation about PrCa despite its status as one of the most common male cancers[15]. Early studies identified the risk factors of age, race and heritability, all of which are now widely accepted. Researchers saw a substantial increase in prevalence of PrCa from age 45 onwards. They noted higher incidence in African Americans compared to Caucasians and a low incidence in Asians. Furthermore, although only two proper "case-control" epidemiology studies of familial aggregation had been reported, both concluded a genetic component to PrCa risk. The first, a report of familial aggregation of PrCa by Morganti *et al* in 1956, observed that patients with PrCa reported a higher frequency of relatives with PrCa than hospitalized controls[16]. Another, an analysis of Mormon pedigrees in Utah by Charles Woolf, found 228 men who died of PrCa were three times more likely than men who died of other causes to have a family history of PrCa[17]. Furthermore, Woolf observed a higher death rate from PrCa among brothers of PrCa cases (RR = 2.81; P=0.002) than fathers of PrCa cases (RR = 1.25; $P >$ 0.05), evidence of possible X-linked or autosomal recessive (AR) transmission of a susceptibility allele.

Subsequent studies focusing on family history of PrCa in first-degree relatives have generally found overall levels of risk due to hereditary risk factors similar to those reported in Woolf's Utah study[18]. Steinberg *et al* looked at pedigrees of 691 men with PrCa and 640 spousal controls in a study population from Johns Hopkins Hospital, finding 15% of the cases but only 8% of the controls had at least one first-degree relative affected with PrCa[19]. Furthermore, they found the odds ratio (OR) increased when more

first degree relatives were affected within the family. In a Los Angeles and Hawaiian multi-ethnic population based study, Monroe *et al* showed evidence of familial aggregation with results similar to Woolf regarding possible X-linked or AR inheritance[20]. In their study, the RR for PrCa in subjects with affected siblings was 2.07 times that of subjects with an affected parent. This RR was relatively consistent throughout all ethnicities in the study (RR range 1.85-2.47). Narod *et al* reported similar results in a Québécois population[21].

Table 1, from a review by Ola Bratt, looks at the relationship between family history and age of onset with PrCa risk[22]. Although the figures presented in Table 1 are derived from Swedish studies, risk values are approximately the same for other high risk populations such as North America, Northern Europe and Australia[22]. Bratt's review counts 27 epidemiological studies conducted of family history as a risk factor for PrCa between 1956 and 1999. All but two studies showed significantly increased risk of PrCa for first degree relatives of family members with PrCa[23].

Table 1. Effect of family history of prostate cancer on lifetime risk of clinical prostate cancer. Adapted from (23)

Family History	Relative Risk	% Absolute Risk
Negative	1	8
Father affected at 60 years or older	1.5	12
1 Brother affected affected at age 60 years. or older	2	15
Father affected before age 60 years	2.5	20
1 Brother affected before age 60 years	3	25
2 Affected male relatives*	4	30
3 or more affected male relatives	5	35-45

* Father and brother, or 2 brothers, or a brother and a maternal grandfather or uncle, or a father and a paternal grandfather or uncle

Twin Studies

Twin studies have revealed that roughly half of the risk for PrCa is heritable and much greater than that of other common cancers[24]. These studies are of interest as twins are either genetically identical (monozygotic) or share one half of their genes (dizygotic). If concordance of the outcome of interest is greater in monozygotic twins than in dizygotic twins, heritable risk factors should be of importance in the etiology of the disease. The first twin study for PrCa heritable risk factors looked at 4,840 pairs of male twins from the Swedish twin registry identifying 458 PrCa cases[25]. Researchers reported 19.2% of monozygotic twins concordant for PrCa, but only 4.3% of dizygotic twins. Another study also mined the Swedish twin registry to identify same-sex mono- and dizygotic twin pairs diagnosed with cancer between 1959 and 1992[7]. Here, investigators compared rates of stomach, colon and rectum, lung, breast, cervical, and prostate cancers. It was clear that PrCa had a strong genetic component accounting for all of the variance explained between the occurrence of PrCa in mono- and dizygotic twins, with an increased risk observed for monozygotic twins compared to dizygotic twins (RR = 6.3, 95% CI 2.5-16.0). Further evidence of heritable risk was seen in a twin study of United States World War II veterans. Concordance was significantly higher in monozygotic twins (27.1%) than in dizygotic twins (7.1%) and investigators estimated that the genetic component of PrCa susceptibility was higher than the environmental component (57% to 43% respectively)[6]. In the largest twin study to date, researchers expanded upon the Swedish study of Ahlbom *et al* to include twins identified in Danish and Finnish registries, obtaining information on cancer occurrence in 44,788 pairs of twins. They found 21% of monozygotic twins and 6% of dizygotic twins concordant for PrCa. Risk

due to heritability was calculated to be 42% of total PrCa risk, the highest of the 11 common cancers investigated[26].

Segregation Analyses

Segregation analyses support a heritable component to PrCa risk. These types of analyses aim to identify the frequency and penetrance of risk alleles as well as their mode of inheritance. While the first report of familial aggregation of PrCa was in 1956, it was not until 1992 that the first segregation analysis was completed and the concept of hereditary PrCa (HPC) was established by Carter *et al*[27]. A generally accepted definition of HPC came in a later paper from Carter *et al*; the occurrence of PrCa in each of three generations of paternal or maternal lineage, or two relatives diagnosed with PrCa before the age of 55 years, or three affected first or second degree relatives[5]. Carter *et al* suggested 9% of PrCa occurrence was caused by a rare, highly penetrant risk allele. An autosomal dominant (AD) inheritance model suggested this allele accounted for 43% of all cases occurring by age 55. The authors proposed results from the study be used as a framework for investigation by genetic linkage studies.

Subsequently, other studies of complex segregation analysis of PrCa susceptibility have been published also describing a rare, highly penetrant allele with AD inheritance[28-33]. However, a closer look at these studies indicates the possibility of genetic heterogeneity because inheritance could not be fully explained simply by an AD model. First, it should be mentioned that three of these studies, Valeri *et al*, Schaid *et al*, and the aforementioned Carter *et al* study are considered somewhat similar due to comparable

choice of probands eligible for prostatectomy having localized disease. Second, the studies from Valeri *et al* and Schaid *et al* found the AD inheritance model alone could not completely explain PrCa inheritance, suggesting other unidentified genetic or environmental components to PrCa risk. Furthermore, Schaid *et al* showed the age-adjusted RR for brothers to be greater than that of fathers. This is more consistent with an X-linked or AR mode of inheritance than AD and similar to the results found in familial aggregation studies of Woolf, Narod *et al*, and Monroe *et al*[17;20;21].

Others have found that the genetic component of PrCa risk could not be fully explained by an AD inheritance model, again suggestive of genetic heterogeneity. In a 2001 segregation analysis of 1,476 Australian men with PrCa and their first and second degree male relatives, researchers found two models of best fit: AD inherited risk for cases of younger ages, and X-linked or AR inheritance for cases of older ages. Lifetime penetrance of the AR or X-linked effect was 100% and the disease allele frequency was estimated at 0.084 (95% CI 0.067 – 0.105)[31]. Two other studies suggested a multifactorial model best identified the PrCa mode of inheritance. In the first, an analysis of 3,796 PrCa patients from 263 families from the Prostate Cancer Genetic Research Study (PROGRESS) for quantitative trait loci explaining variance in age of onset of HPC cases, the authors found evidence of 2 or 3 separate contributory loci[34]. Second, segregation analysis of a population-based sample of Canadians, and United States Caucasians, Asian Americans and African Americans showed that a multifactorial model allowing for multiple susceptibility loci each of low penetrance fit as well as the AD model. In this

study, authors argued for the multifactorial model because it contained fewer parameters than the AD model[32].

In all, familial aggregation studies, twin studies and segregation analyses each describe a significant genetic component to PrCa risk. Furthermore, many of these studies suggest that heritable risk factors of PrCa are genetically heterogeneous.

Prostate Cancer Linkage Studies

Genetic linkage analyses* have identified many genetic loci candidates across multiple chromosomes, demonstrating the genetically heterogeneous nature of heritable PrCa. Results from several linkage studies are summarized in Figure 2[35-48]. As in a similar summary in a review by Daniel Schaid, a baseline lod of 1 was used to evaluate consistency of results[49]. In all, studies have identified no fewer than 78 locations in the genome, with at least one candidate locus on 21 of 22 autosomes, as well as on chromosome X. Some loci that have been the subjects of multiple replication attempts include: 1q24-25 (HPC1)[50], 1q42.2-43 (PCAP)[51], 1p36 (CAPB)[52], Xq27-28 (HPCX)[53], 8p22-23[54], 17p (HPC2)[55] and 20q13 (HPC20)[56]. Candidate genes have been proposed for several of these loci, although each is of uncertain significance. Linkage studies stratified for factors such as disease aggressiveness have yielded other loci of interest, such as 19q12-13[57].

* Genetic linkage studies employ a framework of markers spaced across the genome to determine a region of the genome associated with a phenotype within a pedigree more often than expected by chance alone. A more detailed discussion of the advantages, disadvantages and intricacies of these studies is presented in the section entitled "The Paradigm Shift from Linkage to Association".

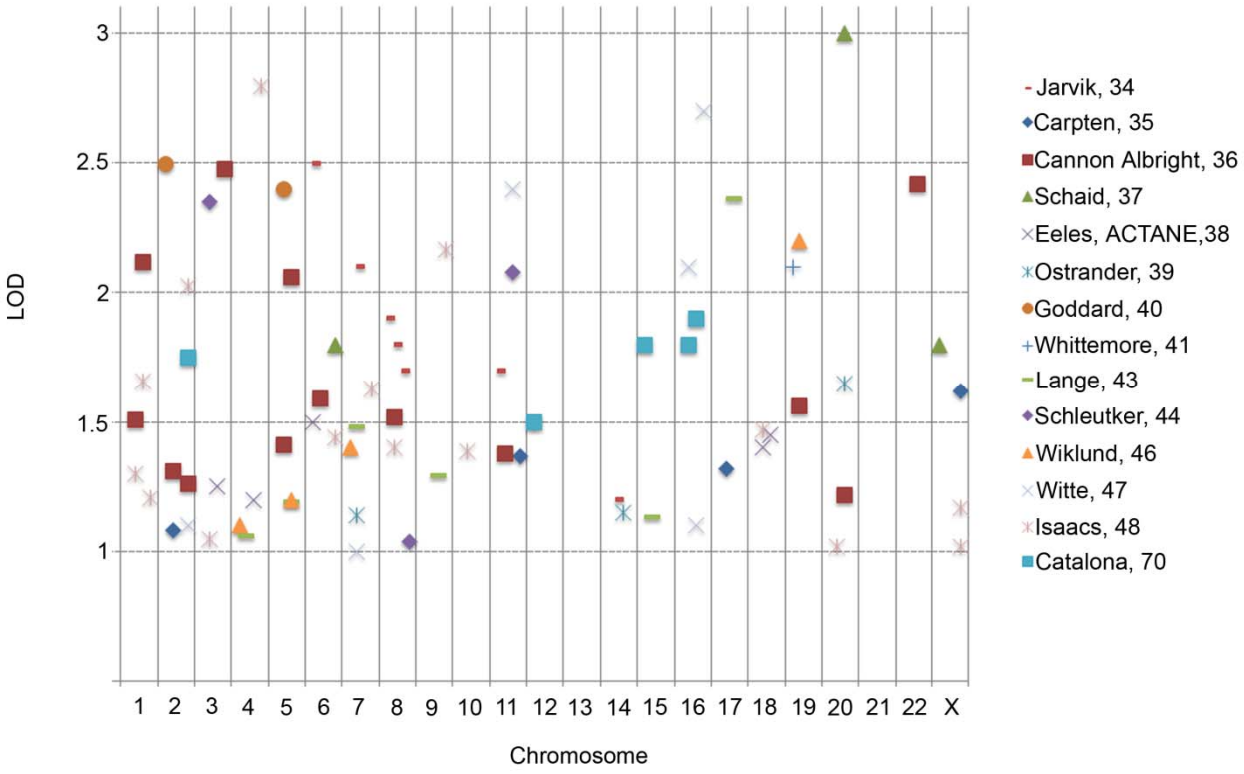


Figure 2. Summary of lod scores ≥1 from 14 PrCa genetic linkage studies conducted from 2001-2007, based on a table from Schaid, 2004. Smith *et al* (1996) and Gibbs *et al* (2000) are not included because they have been superseded by later studies included in the figure. Studies are listed in the order presented in the references section by corresponding author. Reference number is listed to the right of the corresponding author's name. Lod scores are either multipoint heterogeneity lod scores (HLODs) or multipoint model-free non-parametric linkage lod scores (NPL-LODs). If exact lod scores were unavailable, information was estimated from figures provided in their respective manuscripts.

My predoctoral training involves a candidate gene study at 19q12-13 and a thorough investigation of a candidate locus at HPCX, results of which are presented in Chapters IV and V. It is relevant to discuss these and other heavily investigated loci. These are reviewed below in the order of their appearance by year in the literature.

1q25-25/HPC1/RNASEL

Hereditary Prostate Cancer 1 (HPC1) at chromosome 1q24-25 was originally identified in 1996 in a genome-wide linkage scan[50]. The study population consisted of 79 North American and 12 Swedish families with at least three first or second degree relatives affected with PrCa. Families had an average of 4.9 affected males per pedigree and no pedigree showed evidence of bilineal inheritance. Average age of diagnosis was 65. Thirty-four males were diagnosed before age 55. Researchers typed 341 dinucleotide repeat markers in a subset of 66 North American families. The maximum lod score observed was 2.75 under a dominant inheritance model at marker D1S218, mapped to 1q24-25. Researchers then genotyped additional markers in this region, and added the remaining 13 North American families and 12 Swedish families. This provided additional evidence for linkage at marker D1S2883 (5.5 cM centromeric from D1S218) with a maximum two-point lod of 3.65 at recombination fraction $\theta = 0.18$. Non-parametric analysis was significant for multiple markers, providing additional evidence of linkage to the region. An estimated 34% of families studied linked to the region. Furthermore, authors reported two African-American families showing evidence of linkage to the region (lod = 1.4), suggesting increased risk for PrCa due to this locus over multiple populations. Subsequently, it was shown that the lod score increased to 5.10 for families

with an average age of diagnosis < 65 years and with ≥5 affected individuals in the pedigree[58]. Risk of PrCa susceptibility due to HPC1 has been confirmed or supported by several independent studies and populations[40;59-65]. However, replication has not been unanimous, and several studies have reported no evidence of linkage at HPC1[66-70].

RNASEL has been proposed as a candidate HPC1 tumor suppressor gene based on its location at the linkage peak at HPC1 and the presence of sequence variation in several of the PrCa families studied[71]. *RNASEL* encodes 2' – 5'oligoadenylate-dependent ribonuclease L (Rnase L) which regulates antiviral activities and apoptosis. Ubiquitously expressed and natively latent, Rnase L is activated through binding of double stranded RNA molecules that are present during viral replication. Activation results in large-scale RNA degradation with subsequent cellular apoptosis[72].

Carpten *et al* reported two variants in *RNASEL* tracking with two HPC1 linked families[71]. One variant, E265X, found in 0.5% of the general population and non-HPC PrCa cases, truncates Rnase L, producing a protein lacking the 2'-5' oligoadenylate binding domain. The other variant, M1I, abolishes the initiator methionine, but appears to be present only in the African American family in which it was identified, as it was not seen in a sample of 698 controls. Through functional assays, the authors showed both mutations lead to loss of function of Rnase L. A subsequent genetic association study by Rökman *et al* in a Finnish population confirmed the E265X variant (OR = 4.56, 95% CI 1.1-19.4; P = 0.04) as a risk allele as well as identified minor allele homozygotes of variant R462Q as possibly associated with risk of PrCa (OR=1.96, 95% CI 0.9–4.0;

P=0.07)[73]. Additional independent replication attempts have produced variable results. R462Q has been shown to significantly impact PrCa in multiple genetic association studies, but two of these studies identify minor allele homozygotes as *protective* against risk of PrCa, conflicting with the Rökman *et al* Finnish study[74-77]. Other studies, as well as unpublished results in 469 North American Caucasian PrCa cases with a family history of disease and 469 age- and race-matched controls from our own laboratory, have failed to find any association with *RNASEL* variants and PrCa risk[78-82].

1q42.2-43/PCAP

PCAP (Predisposing for Cancer of the Prostate) at chromosome 1q42.2-43 was originally detected in 1998 French study with a lod score of 3.30 in early-onset families (< 65 years) [67]. An estimated 40-50% of French and German families studied linked to the locus. Despite its auspicious introduction, PCAP replication attempts have seen variable results. Suggestive, but not necessarily statistically significant, evidence for linkage to the locus has been seen in several other studies in both Caucasians and African Americans, some of which describe enhanced signal from early-onset families[40;64;69;83;84]. Other studies have reported no evidence of linkage[41;45;85]. Overall, evidence at this locus from replication attempts seem merely suggestive as reviews of PrCa susceptibility loci cannot agree on whether or not most of these studies find evidence of linkage[49;86].

Xq27-28/HPCX

As previously discussed, there exists a wealth of epidemiological evidence supporting X-linked inheritance of a PrCa susceptibility allele. Therefore, reported

evidence of a PrCa locus on the X chromosome is of interest. In the genome-wide search which resulted in HPC1, a 40 cM interval from markers DXS1001 to DXS1108 was also implicated, with a maximum two-point lod = 1.08 at marker DXS1193 at chromosome Xq27-28[50]. Investigators performed a more detailed search at Xq27-28, increasing the number of pedigrees to 360. Pedigrees were collected from North America (from Johns Hopkins University and the Mayo Clinic), Finland and Sweden, and included the 79 North American families described in the HPC1 publication[53]. Researchers used a total of 33 markers at intervals of 1.2 cM in the Johns Hopkins families. A subset of 26 markers were genotyped over a 19 cM interval in the Mayo Clinic and Finland families, and a less dense 4 cM map of eight markers for the Swedish families. Twelve of these markers had lod scores > 1.0 in the combined dataset with a maximum two-point lod of 4.6 at Xq28 (marker DXS1113, $\theta = 0.26$). Interestingly, the Finnish families in this study have a peak two-point parametric lod of 2.05 at Xq27.1-2 (marker DXS1205, $\theta = 0.14$) and minimal evidence for linkage at Xq28; an indication of possible genetic heterogeneity within HPCX itself. It was estimated that 16% of North American families were linked to HPCX. This estimate increased to 40% in the Finnish families. It was also noted that the observation in the Finnish families was from a distinct subgroup of families with no evidence of male to male transmission and a late age of diagnosis[87].

There have been several replication attempts at HPCX. The first successful replication attempt was in a 1999 study from Lange *et al* of 153 PrCa pedigrees from the University of Michigan[88]. As in the original study, authors identified marker DXS1113 as suggestive of linkage (NPL Z-score = 1.20, P=0.12). Signal at this marker was strongest

in the subset of families with both no evidence of male to male transmission and early onset disease. Lange *et al* also reported a second non-overlapping peak at marker DXS294, 4.8 cM centromeric to marker DXS1205 (the peak marker for the Finnish families in the original HPCX manuscript). However, this peak signal was strongest in families with evidence of male to male transmission, counterintuitive to the idea of an X-linked locus. Further confirmation of HPCX came from analysis of a 104 family German population. Pedigrees had at least two living relatives with histologically confirmed PrCa. A peak NPL Z-score = 2.32, *P*=0.009 was seen at marker DXS984, 2.3 cM centromeric to marker DXS1205[89]. Other studies have also shown evidence of suggestive linkage to HPCX[37;84]. The first statistically significant replication of linkage to HPCX was seen in a study of 143 pedigrees from Utah[90]. Pedigrees used in this study were very large; three to eight generations with three to 62 PrCa cases in each pedigree. The authors employed a robust multipoint linkage statistic analogous to a two-point lod score but utilizing full multipoint haplotype[†] information (TLOD). A maximum TLOD of 2.72 was seen at marker DXS8069 at chromosome Xq28 which remained statistically significant after correction for multiple testing bias (*P*=0.0002). There are also independent studies showing no substantial evidence of linkage, however in each of these studies a small number of families were linked to the region[45;91].

To date, no gene has been described as an X-linked candidate, either at HPCX, or on the entirety of the X chromosome. An obvious choice for an X-linked candidate gene is the androgen receptor (*AR*), located at Xq11.2-12. *AR* has been investigated, but never

[†] A haplotype is defined as a sequential set of genetic markers that are present on the same chromosome.

identified through linkage studies. As it is separated from HPCX by over 50 cM, it is unlikely to contain the causal variant responsible for this predisposition locus.

Recent reports of HPCX have narrowed the locus in the Finnish families to a 150 kb region flanked by markers D3S2390 and bG82i1.9 by LD and shared-haplotype analysis[92]. In Chapter V of my thesis, representing the majority of my predoctoral work and entitled "A Haplotype at Chromosome Xq27.2 Confers Susceptibility to Prostate Cancer", I report a thorough investigation of this candidate locus in a United States Caucasian study population.

1p36/CAPB

CAPB (Cancer Prostate Brain) was first reported in 1999 by Gibbs *et al* using the PROGRESS pedigree resource with a peak, although not statistically significant, lod score at marker D1S1597 located at chromosome 1p36[93]. The authors note that locus 1p36 is associated with loss of heterozygosity (LOH) in several types of central nervous system tumors and that prior epidemiological studies have shown a link between PrCa and brain cancer. The authors tested the hypothesis that a shared allele predisposed to both PrCa and brain cancer. When looking at a subset of 12 families in their study with evidence of a primary brain tumor, the lod score associated with marker D1S507 (~4 cM telomeric of D1S1597) increased to a statistically significant 3.22 at $\theta = 0.06$. Gibbs *et al* proposed the name CAPB to designate the link between the locus and pedigrees containing both PrCa and brain cancer cases. Authors proposed a tumor suppressor termed p73 as a candidate gene which maps to the locus and has high homology to tumor

suppressor p53. However, after extensive *de novo* single nucleotide polymorphism (SNP) discovery efforts, authors reported no variants within coding regions or intron-exon boundaries of the p73 gene associated with risk of PrCa[94]. Several independent replication attempts have shown no evidence for linkage over multiple populations. A first replication attempt in a subset of 13 HPC families with at least one instance of brain cancer from a study population from the Mayo Clinic found no evidence for linkage[69]. Nor was replication successful in a population derived from 64 southern and western European families. Six of these pedigrees had an instance of brain cancer, and authors reassessed linkage in this subset finding no evidence thereof[83]. Furthermore, evidence for linkage was not seen in a study of 33 African American families; however, none of these families had a reported case of brain cancer in first or second degree relatives[84]. In another study, four of six HPC families with at least one case of brain cancer had positive linkage results at CAPB[64]. The study population consisted of 159 HPC families including 79 from the HPC1 study by Smith *et al.*

20q13/HPC20

In 2000, evidence at 20q13 was first reported in a genome-wide scan of 162 HPC families with a maximum multipoint non-parametric linkage score of 3.02 at marker D20S887[95]. Linkage was strongest in families with a late average age of diagnosis (≥66 years) and the authors suggested the designation of this locus as HPC20. Soon afterward, a replication study using the Johns Hopkins pedigree resource confirmed linkage to 20q13, also in families with a late age of diagnosis[96]. As with all other loci, subsequent evidence has been variable. A second independent replication in 172 North American

families was only suggestive for evidence of linkage, strongest in a subset of 16 African American families[97]. Further supportive evidence has been seen in subsequent studies[37;84]. Other studies have shown no evidence of linkage, including a study from the International Consortium for Prostate Cancer Genetics (ICPCG), a collection of 1,234 pedigrees with multiple cases of PrCa[98;99].

19q12-13

Chromosome 19q12-13 was first identified in a genome-wide linkage analysis using Gleason score‡ as a quantitative trait of PrCa aggressiveness with a peak P = 0.0004 at marker D19S433[57]. Gleason score is indicative of tumor histology and the most frequently used grading system for PrCa[100;101]. The use of Gleason score as a quantitative trait may reduce phenotypic heterogeneity and thus simplify underlying genetic heterogeneity in this complex disease. Subsequently, the finding was confirmed in an independent study of HPC families from the Mayo Clinic[102]. A third independent study of affected sibling pairs from HPC families from the Fred Hutchinson Cancer Research Center has also recently confirmed linkage to the region[103]. Intriguingly, the locus is not highlighted by other linkage studies of HPC upon limiting affected status to only those men with clinically significant disease[104;105]. These seemingly conflicting results are addressed in Chapter IV of my thesis, entitled "Familial Prostate Cancer Risk, Aggressiveness, and the Transforming Growth Factor β1 T29C Polymorphism". This

‡ In 1974 Gleason and Mellinger proposed a grading system to represent the differentiation patterns of tumors within the prostate. The predominant and second most prevalent patterns are identified and graded on a scale from 1 (most differentiated) to 5 (least differentiated). These two scores are added and a resultant score from 2 (uniformly differentiated) to 10 (uniformly undifferentiated) obtained. A major shift in terms of prognosis occurs between Gleason scores 6 and 7, with scores ≥ 7 almost always requiring active treatment, as opposed to the 'wait-and-see' prognosis often employed for individuals with a score < 7.

chapter describes an association between indolent PrCa and the functional T29C polymorphism of *TGFB1* within this locus.

17p11/HPC2/ELAC2

A study of 33 large, high-risk pedigrees from Utah showed linkage to chromosome 17p at marker D17S1289 with a maximum two-point lod score of 4.53 at $\theta = 0.07$[106]. Using positional cloning and mutation screening, the authors subsequently narrowed the locus down to candidate HPC susceptibility gene *ELAC2*, located at chromosome 17p11. *ELAC2* contains homology to two protein families; PSO2/SNM1 DNA interstrand cross-link repair proteins and the 73-kD subunit of mRNA 3-prime end cleavage and polyadenylation specificity factor. It is thought to encode a metal-dependent hydrolase domain conserved among eukaryotes, archaebacteria and eubacteria. Researchers found a frameshift mutation in one large pedigree (1641insG). A second pedigree contained three variants, two of which were common in the population (S217L and A541T). This finding was corroborated by a non-family matched case-control study in which men with both 217L and 541T had an increased risk of PrCa with an odds ratio of 2.37 (95% CI 1.06-5.29). The 541T variant was only observed in men with the 217L variant, and the combination of the two was estimated to account for 5% of HPC in the study population[107]. However, this finding has been difficult to replicate, with studies showing no linkage to chromosome 17p or no association of the 217L and 541T variants with HPC[108;109].

8p22-23/MSR1

Linkage at 8p22-23 was reported in a study of 159 United States HPC pedigrees with a peak HLOD of 1.84 (P= 0.004) and an estimated 14% families linked[54]. As deletions on chromosome 8p22-23 have been seen in both PrCa cell lines and in studies of high-grade PrCa tumors, linkage to this region is very intriguing. While the study described linkage to the locus, the candidate gene proposed by the authors, *PG1,* displayed no statistical difference in allele, genotype or haplotype frequencies between case and control subjects for any SNPs or other sequence variants.

A later study of the region looked at the candidate gene macrophage scavenger receptor 1 (*MSR1*), known to be involved in prostate carcinogenesis. MSR1 is a multi-domain scavenger receptor expressed almost exclusively in macrophages, and is capable of binding a wide array of ligands including oxidized high density and low density lipoproteins, apoptotic cells, and both gram negative and gram positive bacteria[110]. Several missense mutations and one nonsense mutation were shown to be associated with PrCa risk in a United States study population consisting of Caucasians and African Americans[111]. An immediate independent confirmation of a marker 1 cM centromeric to *MSR1* followed[112]. Subsequent case-control studies in the region have focused on the nonsense mutation, R293X, which deletes most of the extracellular ligand binding domain and is of obvious functional significance. As with studies of *RNASEL*, published reports conflict. Some of these studies showed higher frequency of the nonsense mutation in cases versus controls over multiple populations; however, no association was considered statistically significant[113-115]. Other populations showed an excess of the

nonsense mutation in controls relative to cases[80;114;116]. Meta-analysis of studies through September 2005 collectively showed that R293X as well as other variants at *MSR1* do not play a major role in heritable PrCa susceptibility, but may confer moderate risk to PrCa[117].

Genetic Association Study Design

The Paradigm Shift from Linkage to Association

The diverse array of PrCa susceptibility loci and their lack of reproducibility led geneticists to question the utility of the linkage design framework when looking for causal variants of PrCa[118]. Titles of review articles summarizing PrCa linkage studies from major contributors to the field entitled "Genetics of Prostate Cancer: Too Many Loci, Too Few Genes" and "The Complex Genetic Epidemiology of Prostate Cancer" highlight the confusion surrounding linkage study results[49;119]. These troubles were not specific to PrCa research, but pervasive among study of complex, common disease in general[120].

Quite simply, the success seen from linkage analysis at uncovering the genetic basis of monogenic disorders had not been seen in the study of PrCa. It is useful to discuss the strengths and weaknesses inherent in the linkage study design. Linkage studies are strongest when the causal variant is of high penetrance. The phenotype of a highly penetrant variant will always or almost always be seen in the individual with the variant. It is likely that a successful linkage study results from the high correlation between phenotype and a highly penetrant variant. A variant of low penetrance is difficult

to find using linkage analysis. Indeed, it has been shown that linkage analysis has difficulty in identifying such variants with a risk ratio of < 3[118;121]. Furthermore, disorders identified through linkage analysis are typically of Mendelian inheritance. Mendelian disorders have well characterized models of inheritance, such as AD, AR, and X-linked. In contrast, many common diseases do not behave in this manner, often skipping generations and not adhering to one particular inheritance model. Finally, while linkage studies are largely unaffected by conditions of allelic heterogeneity, they are hindered under conditions of locus heterogeneity (also called genetic heterogeneity). Allelic heterogeneity occurs when multiple variants which result in the same phenotype are seen at the same gene. Locus heterogeneity arises from multiple variants at different genes resulting in the same phenotype. Given the nature of linkage studies, in which a connection is made between a locus in the genome and a phenotype, locus heterogeneity is problematic. In summary, locus heterogeneity and low effect size are problematic when identifying risk loci and these problems are inherent within the linkage study design.

It has been hypothesized that multiple variants of modest effect size collectively conspire to predispose to common disease. This is the central theme of the "common disease-common variant" hypothesis (CDCV) which suggests the genetic risk of common disease is due to disease loci where there are common variants[122]. The array of loci across multiple chromosomes with relatively small lod scores seen in PrCa linkage studies is supportive of this hypothesis.

In 1996, the case-control association study design was suggested as the method of choice for investigating genetic determinants of complex disease[123]. In contrast to linkage studies, which are focused on finding a *locus* in the genome associated with risk of disease, association studies identify an *allele* associated with risk of disease. Association studies test for significant differences in allele frequencies between a case population and a control population. However, in 1996, it was not feasible to replace the linkage study framework with high density, large-scale association studies. Linkage studies benefit from a pre-defined set of informative markers enabling comprehensive scan of the entire genome. A similar set was not defined for association studies. In fact, markers at the density required by a large-scale association study had not yet even been identified or catalogued. A physical map of the genome would not be available until the publication of the draft of the Human Genome Sequence in 2003, hindering assay design. Furthermore, the cost of such an undertaking was prohibitive[124]. Consequently, association studies of the time were typically of small scale and performed on a candidate gene within a locus identified through linkage analysis.

Genomic Structure and its Impact on Genetic Association Studies

While the utility of the association study to identify genetic variants predisposing to common disease was clearly evident, its practicality was hindered by factors linked to the high density of markers needed to efficiently capture genetic diversity. If the number of markers required could be reduced, large-scale association studies would potentially be a feasible alternative to genetic linkage studies in large-scale investigations. Fortunately, high-density SNP studies have demonstrated that the genome consists of discrete areas

absent of recombination which are separated by recombination hotspots and which exhibit a striking lack of diversity[125]. As such, neighboring markers in a chromosome occur together more often than expected. This is called linkage disequilibrium (LD) and substantially decreases the number of markers required to achieve comprehensive coverage. As certain alleles are inherited together, an associated variant might either be the actual causal variant or highly correlated with the causal variant. Therefore, a large-scale association study would not need to directly detect the causal variant but identify any of the variants inherited with it. Subsequent directed analysis would then identify the causal variant.

Binary and amenable to high-throughput assay, the SNP is the variant of choice for large-scale association studies. As a result of LD, most of the approximately 11 million SNPs in the genome have neighboring groups of SNPs which are correlated with each other[126]. One SNP can therefore be used as a proxy for the other correlated SNPs. Therefore, a researcher may select fewer tagging SNPs so that they capture most of the common variation within the region. This may be done on both a haplotype or single SNP basis and established methods exist for each[127;128]. An early estimate puts the number of SNPs required for comprehensive coverage on a genome-wide scale at 500,000[129]. As a practical example, our investigation of a 352 kb susceptibility locus at HPCX uses 128 tagging SNPs selected from a larger set of 246 SNPs to capture common variation in the region (Chapter V of this thesis).

Together, LD and linkage equilibrium parse the genome into a block-like structure. Over the course of many generations, ancestral chromosomes are broken up by meiotic recombination creating smaller segments of DNA. These segments are areas of high LD and are appropriately referred to as haplotype-blocks. Haplotype-blocks are passed largely intact from generation to generation, but eventually broken by recombination events over time such that block length is determined by both the age of the population and the location of recombination hotspots throughout the genome. Older populations (*e.g.* Africans) have more frequent recombination hotspots and therefore greater genetic diversity than newer populations (*e.g.* European Caucasians). Haplotype-block length is greatest and diversity lowest in founder populations derived from a limited pool of individuals (*e.g.* Finns, Icelanders). Longer blocks and lack of diversity facilitates detection of a disease associated haplotype in an association study, but hinders subsequent identification of a variant of interest. However, researchers could then turn to populations with shorter blocks to identify the causal variant on the associated haplotype.

Potential Problems for Association Studies

Population Stratification

One of the most widely discussed potential problems of association studies is population stratification and subsequent reporting of spurious associations due to this factor. Population stratification results from multiple subgroups within a population that differ in disease prevalence. This can result in a biased selection of cases from one subgroup of a population over another. If allele frequencies differ in these subgroups, spurious associations can occur. Methods have been proposed to detect and correct for

population stratification[130-133]. The effect that population stratification has upon association studies is controversial. Researchers have shown that population stratification in 'well-matched' (using self-reported ethnicity) case-control studies of admixed societies, such as those found in United States cosmopolitan areas, is unlikely to result in spurious associations[132;134;135]. Others report that minimal amounts of population stratification could result in spurious associations in studies designed to detect risk factors of modest effect size[136].

Controlling for Multiple Testing Bias

The sheer number of markers that require genotyping in association studies also may lead to the reporting of spurious associations. At a $P = 0.05$ threshold of statistical significance, 5% of markers may be falsely associated by chance alone. Very few of these markers, if any, may be the actual causal variant or its tagging SNP surrogate. It is then necessary to systematically discount the false positive associations. To circumvent, the significance threshold can be increased and it has been proposed that the Bonferroni correction be employed in this situation[123]. In Chapter III of my thesis, entitled "Haplotype Analysis of *CYP11A1* Identifies Promoter Variants Associated with Breast Cancer Risk" we use the Bonferroni correction to control for multiple testing bias. In general, however, Bonferroni correction is considered punitively conservative in association studies. High marker density, LD between markers, non-random SNP selection and redundancy between single-allele and haplotype tests strongly violate Bonferroni assumptions of independence, and other methods have been suggested[126;137]. One of these methods consists of a multi-stage approach, and it is this method we use in

our HPCX study (Chapter V). By dividing the study population into independent training and test sets, it is possible to employ the training set for the myriad of tests involved in the initial screen. Statistically significant associations would then undergo a replication attempt in the independent test set. Presumably, this would result in far fewer tests requiring correction for multiple testing bias.

Sample Size

The power of association studies to detect variants predisposing to common disease is related to both the effect size and the frequency of the variant in the population, such that a rare variant of small effect size will be most difficult to detect. Power calculations are visualized in Figure 3 using the PS program from Dupont and Plummer[138]. As the sample population increases, we are able to detect variants of smaller effect size for a variant of constant frequency (Figure 3A). For example, using a 1:1 matched case-control population with 500 cases we are able to detect a variant with frequency 5% and an odds ratio of 2.0 at 80% statistical power ($\alpha = 0.05$). Increasing the population to 2000 cases, we are able to detect a variant at an odds ratio of 1.45 ($\alpha = 0.05$). Similarly, as the allele frequency increases in a study population of constant size, we are able to detect variants of smaller effect size (Figure 3B).

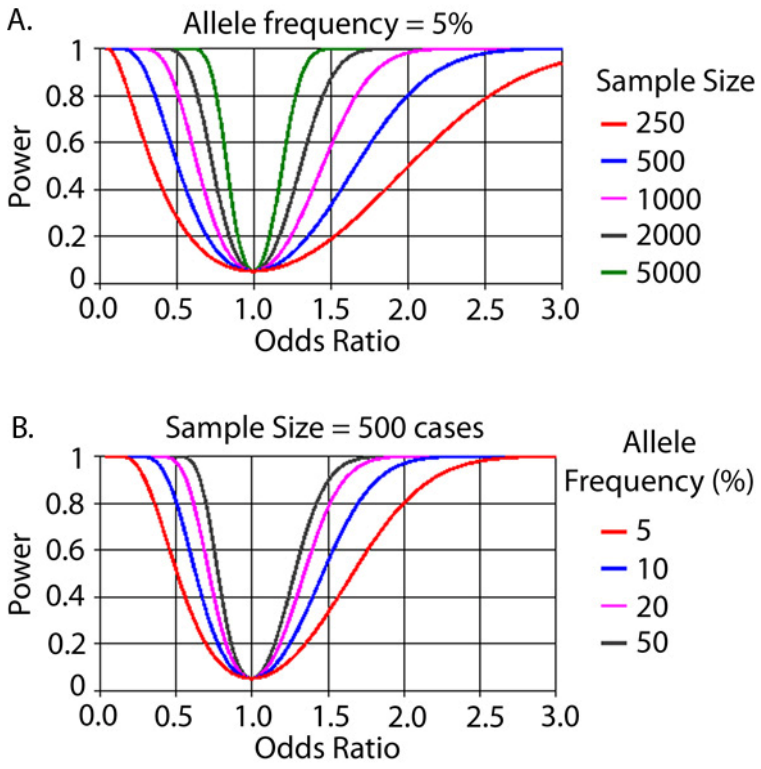


Figure 3. *Panel A.* Sample sizes needed to detect an allele of frequency = 5%, using a 1:1 matched case-control study design, with ϕ=0.0, α = 0.05, for effect sizes 0.0 – 3.0. *Panel B.* Detectable odds ratios for a population of 500 cases given allele frequencies from 5% - 50%, using a 1:1 matched case-control study design with ϕ=0.0, α = 0.05, for effect sizes 0.0 – 3.0. As an example, a population size of 500 cases is able to detect a variant of 5% frequency in the population with an odds ratio of 2.0 at 80% statistical power. Allele frequency of 5% and population sample size of 500 were chosen to approximate values of these variables in the HPCX study presented in Chapter V of this thesis.

Susceptibility Locus 8q24

In 2006, the first results from HapMap[§] based genome wide association studies (GWAS) for PrCa susceptibility loci appeared. The most intriguing results to come of these studies involve the identification of PrCa susceptibility loci on chromosome 8q24. First identified in two independent studies in 2006 from Amundadottir *et al* and Freedman *et al*, 8q24 has been replicated in several independent studies involving multiple populations[139-147]. This type of overwhelming replication had not yet been seen in the field of PrCa genetics. Studies describe three separate but contiguous loci at 8q24 responsible for PrCa risk. SNPs rs6983267 and rs1447295 result in the strongest association in most studies. Exactly how gene-poor 8q24 contributes to PrCa risk remains undefined, however it is a common location for somatic gains in PrCa[148].

Concluding Remarks

In summary, at the outset of my graduate studies researchers were beginning a shift from the linkage study design to case-control association when searching for common disease susceptibility variants in large-scale investigations. While heritability is widely accepted as the largest single factor predisposing to PrCa risk, linkage studies had failed to uncover susceptibility loci independently reproducible across study populations. We hypothesized that the reason for this failure was that heritability of PrCa was due to common variants in the population each conferring relatively lower risk than encountered

[§] The **International HapMap Project** seeks to identify common variation in humans. Currently in its second iteration, the HapMap includes information on over 3.1 million SNPs genotyped in 270 individuals from four populations (Yoruban, Japanese, Han Chinese and Caucasian Americans of western and northern European ancestry). HapMap data is publicly available (http://www.hapmap.org), allowing investigators to obtain information on relevant tagging SNPs prior to embarking on an association study.

in typical Mendelian disorders. Haplotype-based case-control association studies are better equipped to investigate this hypothesis than traditional linkage studies.

CHAPTER II

HYPOTHESIS AND SPECIFIC AIMS

Prior to embarking on a large-scale, comprehensive investigation at HPCX, we sought to test the study design and fine tune techniques and methods to be used. At the outset of my graduate studies, our PrCa study population was not yet of sufficient power and required further accrual. Therefore, we employed a heavily published study population to investigate a candidate gene of known significance with an undefined causal variant; the study population was the Shanghai Breast Cancer Study (SBCS) and the breast cancer associated gene was the rate-limiting enzyme of steroid biosynthesis, *CYP11A1*. Previously, we associated a specific allele of a simple tandem repeat (STR) upstream of *CYP11A1* with risk of breast cancer within the SBCS population[149]. We sought to use this known association as a positive control to test the ability of a haplotype-based study design to detect common risk variants and hypothesized this allele marks an uncharacterized haplotype harboring candidate functional variants and conferring breast cancer risk. Although this work was done in a breast cancer study population, the methods and techniques developed to identify an associated haplotype at *CYP11A1* were subsequently directly applicable to my PrCa thesis work. This study is discussed in detail in Chapter III.

Next, we perform a candidate gene association study as the first use of our PrCa study population. This study is described in Chapter IV. Several linkage studies have been published identifying chromosome 19q12-13 as a PrCa aggressiveness locus. Residing at chromosome 19q13.2, transforming growth factor β1 (*TGFB1*) plays a remarkable dual role in the genesis and progression of multiple cancers and is an ideal candidate gene at this locus**.** The *TGFB1* gene contains a well-studied functional T to C transition at nucleotide position 29. We focused on this common functional polymorphism and assessed the significance of *TGFB1* as a PrCa susceptibility gene. In this study, we also assessed the ability of the new study population to detect the established association at 8q24.

Our PrCa study population has grown over the course of this work and is uniquely designed to dissect the genetic component of PrCa using LD mapping. We focused on one candidate interval at HPCX derived by shared haplotype association evidence in the founder populations of Finland and Ashkenazim. We hypothesized that a gene or genes in this candidate interval at HPCX harbor common variants of modest effect size predisposing to risk of PrCa. We performed exploratory haplotype analyses in a training study population, and sought to confirm or refute statistically significant haplotypes in an independent test population. In this way we identify a haplotype within HPCX significantly associated with PrCa risk. This work is detailed in Chapter V.

Towards identifying the contribution of gene *TGFB1* and locus HPCX by identification of variants within them predisposing to PrCa susceptibility, I list the following projects and aims for my dissertation:

Project 1: Validation of Experimental Procedure Using *CYP11A1*

Specific Aims and Experimental Summary:

1) Characterization of genetic architecture at *CYP11A1*
   a. Identify all common polymorphism in the Han Chinese study population via both *de novo* SNP/STR discovery and public SNPs from dbSNP[150]
   b. Determine a set of tagging SNPs/STRs in a subset of the study population
   c. Genotype the set of tagging SNPs/STRs in the remainder of the study population

2) Determine the haplotype(s) at *CYP11A1* associated with risk of breast cancer
   a. Use haplotype and single-allele sliding window χ^2 tests of association
   b. Use age-adjusted logistic regression analysis

3) Comprehensive search for polymorphism on the associated haplotype to identify functional candidate variants
   a. Completely re-sequence associated haplotype 1.9 kb 5' to 98 bp 3' of the gene
   b. Re-sequence exons and exon-intron junctions in five most common haplotypes

4) *CYP11A1* expression analysis
   a. Test *CYP11A1* expression levels in lymphoblastoid cell lines harboring associated haplotype, relative to those harboring unassociated haplotypes

Project 2: The Transforming Growth Factor β1 T29C Polymorphism and its Association with Prostate Cancer Aggressiveness

Specific Aims and Experimental Summary:

1) Genotype SNP rs1447295 at 8q24 to assess the ability of the PrCa study population to detect an established association
   a. SNP rs1447295 has been identified and widely confirmed as associated with risk of PrCa

2) Genotype the T29C polymorphism of *TGFB1* in the study population
   a. Use dual methods to obtain accurate genotypes, with discrepancies resolved via direct-sequencing

i. Fluorogenic 5'-nuclease assay (Taqman)
ii. Single nucleotide primer extension assay with detection by florescence polarization

3) Test for association of the T29C polymorphism with PrCa
   a. Inheritance model determination
   b. Stratification by indices of aggressiveness

Project 3: Identification and characterization of an X-Linked familial PrCa gene

Specific Aims and Experimental Summary:

1) To identify and genotype all common polymorphism in an initial training PrCa population at the candidate interval
   a. Perform *de novo* SNP discovery at predicted or known genes and derive a set of survey SNPs from dbSNP spanning the broader candidate interval
   b. Genotype subset of the training population for polymorphism to determine set of tagging SNPs and type these SNPs in the remainder of the training population

2) To test haplotypes for association with PrCa risk and to determine the variant(s) within the associated haplotype(s) responsible for the significant association in the training population
   a. Perform haplotype and single-allele sliding window χ^2 tests of association
   b. Determine all haplotype windows nominally associated with risk of PrCa ($P \leq 0.05$)

3) Confirm or refute nominal statistically significant associations in an independent test study population
   a. Determine haplotype tagging SNPs (htSNPs) for associated windows
   b. Genotype htSNPs in the independent test population
   c. Confirm or refute significance using χ^2 tests of association
   d. Determine effect size in confirmed window(s) using age-adjusted conditional logistic regression

CHAPTER III

HAPLOTYPE ANALYSIS OF *CYP11A1* IDENTIFIES PROMOTER VARIANTS ASSOCIATED WITH BREAST CANCER RISK[***]

Introduction

Evidence for *CYP11A1* Involvement in Breast Cancer

Endogenous estrogen exposure in women is strongly associated with risk of breast cancer [151;152]. Genetic variation within genes encoding enzymes of the biosynthetic pathway could greatly influence estrogen exposure, and therefore associated breast cancer risk [153]. The conversion of cholesterol to pregnenolone is the common initial step in the biosynthesis of sex hormones, including estrogen, progesterone, and androgens. This rate-limiting conversion is catalyzed in steroidogenic tissues on the inner mitochondrial membrane by the cholesterol side-chain cleavage enzyme, the Cyp11A cytochrome P450 [154]. We previously demonstrated significant allelic association of a simple tandem repeat (STR) polymorphism upstream of the *CYP11A1* gene with breast cancer risk within a Chinese study population [149]. Linkage and allelic association at the marker has also been observed in the androgen-related polycystic ovary syndrome [155]. This STR is a pentanucleotide repeat (D15S520 at 15q24.1, $[TAAAA]_n$) located 487 bp upstream of the first exon of *CYP11A1*, a region not conserved between human and mouse. Three major alleles of 4-, 6- or 8-repeats account for nearly all variation at the

[***] Adapted from *Cancer Res*. 2007 Jun 16;67(12):5673-82

marker among Chinese. The 8-repeat allele is associated with a dose-dependent elevated risk of breast cancer (heterozygote OR = 1.5, 95% CI = 1.2 – 1.9, homozygote OR = 2.9, 95% CI = 1.3-6.7, trend test P <0.0001) [149].

Study Design

In this study we sought to comprehensively characterize common genetic variation at *CYP11A1*, to assess patterns of linkage disequilibrium (LD), and to refine our understanding of the contribution of *CYP11A1* genetic variation to breast cancer risk. The initial discovery of the single allele association at one STR of the *CYP11A1* gene led us to hypothesize that it marked an uncharacterized haplotype harboring candidate functional variants and conferring breast cancer risk.

Among alleles of variant sites identifying a cancer-associated haplotype, a subset that directly marks it are candidates that may be functional in the disease. Those altering transcript expression or processing, or the encoded enzyme itself remain of great interest in further delineating the role of this gene in common breast cancer. We tested this hypothesis within the Shanghai Breast Cancer Study using haplotype-based analyses to comprehensively examine the genetic architecture of *CYP11A1*. Here we demonstrate that the disease-associated haplotype is designated by multiple variants upstream of the coding region. We further observe that *CYP11A1* expression in a lymphoblastoid cell line homozygous for the disease-associated haplotype is two-fold greater than expression in lymphoblastoid cell lines harboring alternative haplotypes. We conclude that common

cis-acting variants upstream of the coding region may impact transcriptional regulation to influence breast cancer risk.

Materials and Methods

Study Population

The Shanghai Breast Cancer study has been previously described [149;156]. Briefly, study subjects were recruited between August 1996 and March 1998. All subjects were permanent residents of urban Shanghai without a prior history of any cancer and were alive at the time of interview. The study included 1,459 incident breast cancer cases diagnosed at an age between 25 and 64 years during the study period (91.1% of eligible cases). Cancer diagnoses for all patients were reviewed and confirmed by a panel of clinicians including two senior pathologists. Unaffected controls were randomly selected from the general population using the Shanghai Resident Registry, a population registry containing demographic information for all residents of urban Shanghai. Inclusion criteria for controls were identical to those for cases, with the exception of a breast cancer diagnosis. Controls were frequency matched on age (5 year intervals) to the expected age distribution of the case subjects in a 1:1 ratio. The study included 1,556 control subjects (90.3% of matched eligible controls). Blood samples for DNA extraction were collected from 1193 (82%) cases and 1310 (84%) controls. All study participants provided written informed consent under an approved institutional review board protocol.

To preserve the limited DNA from study subjects recruited in the Shanghai Breast Cancer Study, allele discovery employed DNAs obtained from Chinese cell lines of the

Coriell Institute for Medical Research (Camden, NJ). These included: NA18524, NA18526, NA18529, NA18532, NA18537, NA18540, NA18542, NA18545, NA18547, NA18550, NA18552, NA18555, NA18558, NA18561, NA18562, NA18563, NA18564, NA18566, NA18570, NA18571, NA18572, NA18573, NA18576, NA18577, NA18579, NA18582, NA18592, NA18593, NA18594, NA18603, NA18605, NA18608, NA18609, NA18611, NA18612, NA18620, NA18621, NA18622, NA18623, NA18624, NA18632, NA18633, NA18636, NA18637, NA00576, NA03433, NA13411, NA14821, NA16654, NA16688, NA16689, NA17013, NA17014, NA17015, NA17016, NA17017, NA17018, NA17019, and NA17020.

Variant Discovery and Confirmation

To capture genetic diversity of *CYP11A1*, database single nucleotide polymorphisms (SNPs) annotated in dbSNP were screened for common polymorphism in the study population. Fifty-three annotated SNPs spanning *CYP11A1* from 7.8 kb 5' of the 29.9 kb gene to 10 kb 3' were genotyped to assess polymorphism. This was done in quadruplicate among 15 Chinese cell line DNAs. The screening set was estimated to provide 95% power to detect a polymorphism with a minor variant frequency of 0.10, and 78% power with a frequency of 0.05.

The 15 Chinese cell lines were also employed for *de novo* SNP discovery by dual single-stranded conformation polymorphism methods (SSCP) and re-sequencing. Where either SSCP method identified a variant, conformers were re-sequenced for allele discovery. Overlapping amplimers across *CYP11A1* were employed for polymorphism

screening. This included 142 amplimers spanning from 1.9 kb 5' to 98 bp 3' of the gene. Intron 1 of the *CYP11A1* gene contains a 13.8 kb nearly contiguous interval of RepBase repeats. Select non-unique regions embedded within that interval were omitted from the survey, as outlined in Figure 4.

Characterization of haplotypes and linkage disequilibrium was conducted among a pilot subset of subjects of the Shanghai Breast Cancer Study that included 178 cases and 178 controls. Subsequent genotyping of tagging SNPs and STRs for tests of association with breast cancer was conducted among 1159 cases and 1236 controls. A total of 200 ng of DNA from the each of the pilot subjects, and 100 ng of DNA from each of the remaining study subjects was used for study.

To identify additional genetic variation on the disease-associated haplotype, Chinese cell line GM16654 that is homozygous for the disease haplotype, and comparative cell line GM10859 (CEPH 1347-02) were re-sequenced from 1.9 kb 5' to 98 bp 3' of the gene (again omitting non-unique regions within the 13.8 kb interval of intron 1), as outlined in Figure 4. All exons and exon-intron junctions were additionally re-sequenced in 5 subjects of the Shanghai Breast Cancer Study harboring five common haplotypes, including a subject homozygous for the disease-associated haplotype. Sequencing employed BigDye® terminator chemistry on a 3100 Genetic Analyzer (Applied Biosystems, Foster City, CA). These re-sequencing efforts identified five SNPs that had not previously been detected by SSCP or described in databases, two of which were not polymorphic in additional Chinese cell lines tested.

Fifteen novel SNPs discovered in the study have been submitted to dbSNP (ss68316999 - ss68317010, and ss68362647 - ss68362649). Two novel polymorphic STRs have been submitted to dbSNP and to GDB (D15S1547 and D15S1546).

SNP Genotyping

We genotyped SNPs by single nucleotide primer extension and fluorescence polarization in 384-well format [157]. Reaction processing entailed three steps: a 4.4 µl PCR reaction, addition of 4 µl of an exonuclease I (New England Biolabs, Beverly, MA) and calf intestinal alkaline phosphatase (Promega, Madison, WI) reagent mix to degrade unincorporated primer and dephosphorylate dNTPs, and a final addition of 4 µl of an Acyclopol and Acycloterminator reagent mix for the primer extension reaction (AcycloPrime^{TM} FP SNP Detection System, Perkin-Elmer, Boston, MA). Each PCR mixture included 0.1 unit AmpliTaq Gold DNA polymerase, 1x Buffer II (Applied Biosystems, Foster City, CA), 2.5 mM $MgCl_2$, 0.25 mM dNTPs, 335 nM of each primer, and 2 ng DNA template. We detected incorporation of R110- and TAMRA-labeled terminators by fluorescence polarization on a Molecular Devices / LJL Analyst HT. Both forward and reverse strand extension primers were tested to select the most robust assay. Amplimer and extension primer sequences for genotyped SNPs of Figure 5 are provided in Table 2.

Table 2. *CYP11A1* Project Assay design

Marker	Forward Primer	Reverse Primer	Allele/Dye	Strand	Extension Primer
rs3825944	CTCCACGGATGGTGAGAAC	GGAGAAACTAGTTGTTTGAC	G/A	R	TCCACTAGAGGGCAGCA
rs12438594	GACCTTGGAGAGAGGTTTC	CAGACCAAGCTCCAGGGT	G/A	R	AGACCAGGTAGGTATCCAG
rs4077585	CAATCCGAGAACCCCAGCAAC	CGCACGAGAAGAAGGTGC	G/C	R	CTTGGCTCCGGAGCCTA
rs4077582	GCTCACTGGTCTTGAATC	GCACCCTTTTACACAGGC	G/A	F	TGACATGGCATGGCATG
rs4077581	CTGGAACTGGACTCTGTC	GCTGTGAAATATTCACATGG	G/A	F	CCTTATGTGCCTGTGTAAA
rs8039957	CCTGGGCAATGTATAGAG	CATTGCAGCCCTCCTATGAG	G/A	R	TCTCCCTGCTCCTCTAA
ss68317009	GTGAGATTCTGTCTCAAAC	GTTTCACCATGTTGGCCG	G/A	R	TTCACCATGTTGGCCGG
ss68317008	GCCCTACTAAATGCCTCC	GACAGACTCAGAGCCTCAG	G/A	F	CCGCACACCTTGCAAGC
rs7174179	CAACAGAGAGAGACTTGAC	GGCTACAGACTCTAAATTC	G/A	R	TTCAGTGGTAGGAGACTGAC
rs12916765	GTTTTACGTGGGTTAGTAG	CATGGGATTGACAAAATG	G/C	F	GGTAACATATACTTAGACATTAGAATTT
rs1843090	GTAAGGTTTAAGCCCCCAG	GTGTTAGGAAAAAAACCCAC	C/T	F	TCCATTGGTTAATTCCATA
ss68317006	GTGAACATACAACCTACTCTTG	CAGTGTTTTTCCAGCTAC	G/A	F	GTTAGGGGTATGGAGCT
rs17515476	GGCCTACCTTGCAAGCTATAG	GATCACTGTTGCTGCTGTC	G/C	F	CTGTGGACAGGTGAGAAG
rs6495096	CTTGCTGGTCCATGGAAG	CTGAGTCGAGGCCCTTAAC	G/C	F	GCAACAGTGATCATAAAGCT
rs1484215	GAACGATTCCTCATCCCG	CTGGCAGAGCAATTCATC	G/A	F	CCTCTAGGTGAATCCCC
rs11632698	CTGGTCAATTTTGTTTGTGC	GAGTGAAGGGGAACAAAAC	G/A	R	CCAGGAACTGATATTCTTAGA
rs11638442	GAGGCTTGCTCTATCAG	GTACTGAGGTCTGGAAAG	G/C	R	GCCCACAGCAAATGCCT
ss68317002	CTGTATTTCATCTGGAGG	GGCAACAATGACAAGCTG	C/T	F	GCTGTGTGTTGTTTCAGTT
rs7173655	GTCATTCTGGAGTGCAATC	CATTCCATTGTCTAAAAGGC	G/A	F	CTCTACTCACTGTGGACATG
rs2279357	CTGAGGTTTGTAGACAAG	CAGCATCTGAGAAAGGCAG	G/A	F	GTCTAGGCCTAAATCAAGG
rs6495095	GGATGGAAAAGGGCTCTC	CCTGGAATCAGCTCTCAG	C/T	F	GTCCAGGTGGAGGCCAG
rs6495094	GGATGGAAAAGGGCTCTC	CCTGGAATCAGCTCTCAG	G/C	R	TGACCCCTTTTTCACCT
rs2277606	CACCTGCCTTCTCTTGGTG	GTGGAGGATTCAGCAGAGG	G/A	F	GTGAGATGGGGGAGGAG
rs1564782	GACTGTGTGAGTGTCTGTG	GGAGAGAACCGCATACTG	G/A	R	GGGGCAGGGCAAAGCCA
rs2277602	GTTCATCTCCTGTGGATC	GAACATTAGTGTGGCTGCC	G/T	F	CCACATCCACATCTACACT
rs2930306	GTGTCCTCGGACAGCATTG	CCAAATTATACCTGCCTGGG	C/T	R	GCAACACCAGGCATCTC
rs2930305	CTCAGTCTCTGCACCACAG	CAGGACTCACTCCATGAG	G/A	F	ATAACCGGGTTGTGAGC
D15S1547	Ned-AGGTAGTGGTCACTCCAG	gtgtcCAATAGAGCTGTTACCAAAC	Ned		
D15S520	gtgtCTCTGAGTCAGCTGTACTG	Hex-GAGCTATCTTGCCAGCTTG	Hex		
D15S1546	Fam-GAGACTGGTGAGGCTAAG	gtgtCCGAGTAGCTGGGATTATAG	Fam		

Simple Tandem Repeat Genotyping

5'-dye-labeled fluorescent amplimers were detected on an ABI PRISM® 3700 (Applied Biosystems, Foster City, CA). Primers were designed using a tailing strategy to promote full non-templated nucleotide addition by AmpliTaq Gold DNA polymerase (Applied Biosystems, Foster City, CA), providing unambiguous detection of alleles separated by one base pair [158]. PCR conditions were as described above. Primer sequences for polymorphic STRs are provided in Table 2. Allele fragment size estimation was accomplished using the internal size standard Genescan 400HD ROX and the local Southern algorithm of GENESCAN software. Editing of alleles was performed in GENOTYPER (Applied Biosystems, Foster City, CA).

Single Stranded Conformation Polymorphism Detection

Amplimers were electrophoresed on 0.5X MDE gels (Cambrex Biosciences, East Rutherford, NJ) at room temperature at 2W for 14 hours, and at 4°C at 4W for 14 hours. PCR conditions were as described above. Amplimers were visualized by silver staining [159]. Representative conformers were sequenced using BigDye® terminator chemistry on a 3100 Genetic Analyzer (Applied Biosystems, Foster City, CA) to identify the polymorphic sites.

Statistical Analyses

Hardy-Weinberg equilibrium (HWE) for markers was calculated using the Stata package *genassoc* of David Clayton [160]. Pairwise LD between SNPs was calculated and visualized using Haploview version 3.2 [127]. Pairwise LD for SNPs and multiallelic STRs

was calculated and visualized using MIDAS version 1 [161]. Tagging SNPs were selected using LDSelect with a minor allele frequency (MAF) threshold of 5% and an r^2 threshold of 0.7 [128]. When multiple SNPs were assigned as tagging SNPs for a particular bin, the SNP with most robust assay performance was selected for that bin.

Population haplotype frequencies were estimated by the Bayesian method implemented in PHASE version 2.1 [162-165], and by an expectation-maximization [166] (EM) algorithm implemented in custom software that we based upon a parent program written by Daniele Fallin and Nicholas Schork [167;168]. The custom EM program enabled use of multi-allelic markers, placed no hard-coded limit on the number of subjects or markers, and allowed parallel processing. Diplotypes were predicted using PHASE, and those predicted with a probability greater than 95% were used for tests of association.

The χ^2 test statistic was used to evaluate differences in allele or haplotype frequency of case and control groups. Alleles or haplotypes with an overall frequency <0.05 were grouped for analysis. A sliding window approach tested a haplotype window of *N* markers, sliding the window along the map in single marker increments [167;169]. For a given window of N adjacent markers, the profile of multiple common haplotypes and rare haplotypes as a group were evaluated in cases and controls by the χ^2 test statistic. Each N-marker haplotype and remaining haplotypes of the window as a group was also evaluated by the χ^2 test statistic. Permutation testing was used to assess significance. Subsequent estimation of effect size employed logistic regression models adjusted for age (Intercooled Stata 9, Stata Corporation, College Station, TX).

Cladistic modeling of haplotypes resolved by PHASE with ≥ 99% probability was accomplished using DNAPARS and DRAWTREE of the software package Phylip 3.6. The observed haplotype with the least number of state changes to all other observed haplotypes was designated as the outgroup for unrooted parsimony. Each multiallelic marker of *N* alleles was encoded as a series of *N*-1 binary allelic sites to allow inclusion in the model.

Expression Analyses in Lymphoblastoid Cell Lines

Expression analyses employed RNA prepared from the lymphoblastoid cell lines GM16654, GM17020, and GM17014, carrying select *CYP11A1* diplotypes (Coriell Institute for Medical Research, Camden, NJ). Cells were cultured at 37 °C under 5% CO_2 in medium containing RPMI 1640 with 2 mM L-glutamine and 15% fetal bovine serum. Total RNA from each cell line was prepared from cells in the log phase of growth using the RNeasy midi kit with on-column DNAse treatment (Qiagen, Valencia, CA). RNA quality was assessed by reverse transcriptase PCR using two different sets of intron-spanning primers, one for PGK and one for p53, with a no reverse transcriptase control to rule out DNA contamination. Nine 1 μg aliquots of total RNA of each cell line were reverse transcribed into single-stranded cDNA using High-Capacity cDNA Archive Kit (Applied Biosystems, Foster City, CA). After cDNA synthesis, RNA was degraded by alkaline hydrolysis, pH was neutralized, cDNA was purified by adsorption to silica gel (QIAquick PCR Purification Kit, Qiagen, Valencia, CA) and eluted in 60ul of 10 mM Tris Cl, pH 8.5. cDNA quantities were measured spectrophotometrically (NanoDrop ND-1000, NanoDrop Technologies, Wilmington, DE).

A fluorescently labeled TaqMan MGB probe was used to quantify *CYP11A1* expression in each of the nine reverse transcribed aliquots by real time quantitative PCR. Each assay was performed in quadruplicate. The probe spanned the exon 1 – exon 2 boundary within the coding region (Chr 15: 72424438 – 72427449, assay # Hs00167984_m1, Applied Biosystems, Foster City, CA). Five nanograms of cDNA was amplified in a 5 μL reaction using the TaqMan system (Assays-On-Demand Gene Expression Products, TaqMan Universal PCR Master Mix, 7900HT Real-Time PCR System; Applied Biosystems, Foster City, CA). For each *CYP11A1* expression assay, results were normalized to the expression of the 18S rRNA housekeeping gene in the same sample (assay # Hs99999901_s1, Applied Biosystems, Foster City, CA). Statistical comparisons were made using a one-way analysis of variance, and two-tailed Student's *t*-test.

Results

We sought common polymorphisms at the *CYP11A1* gene by screening previously annotated variation and by *de novo* variant discovery within 30 chromosomes of Chinese cell lines. We tested SNPs annotated in dbSNP across an interval from 7.8 kb upstream to 10 kb downstream of the 30 kb *CYP11A1* gene for polymorphism. We also sought previously un-described common polymorphism through survey of the gene and ~2 kb 5'-flanking sequence by SSCP and re-sequencing. Repetitive sequence was an obstacle for unique assay. A 5.6 kb window of non-unique sequence of intron 1, and several additional small repetitive intronic regions totaling under 2 kb were omitted from SNP discovery efforts (Figure 4). Collectively we identified 59 variant sites in the

CYP11A1 genomic region, positioned on the map of Figure 4. Of these, 80% were annotated in dbSNP. We developed assays for 3 STRs (including D15S1547, D15S520, and D15S1546, but omitting poly A tract indels rs3831490 and rs12899703) and 46 SNPs using Chinese cell lines. Among these markers, 3 STRs and 42 SNP assays were further genotyped in a subset of the Shanghai Breast Cancer Study population for assessment of minor allele frequency, HWE, haplotype diversity, and for selection of tagging markers. This study population subset included 178 cases and 178 controls. This yielded 3 polymorphic STRs and 27 SNPs (Figure 5) with minor allele frequencies ≥ 0.05 and in HWE ($P \geq 0.05$) for inclusion in analyses. These SNPs had MAFs that ranged from 0.49 to 0.06 among controls. STR heterozygosities were 0.79 (D15S1547), 0.52 (D15S520), and 0.70 (D15S1546).

Pairwise LD across the *CYP11A1* gene was relatively strong in the study population and without clear LD block subdivision. A Haploview plot of SNP allele pairwise D' values is presented in Figure 4. If an STR was highly mutable, one would anticipate low LD with neighboring SNPs. Instead, specific alleles of the STRs were in strong LD with select SNP alleles and efficiently tagged SNP haplotypes with few assays (Figure 5). For example, the T allele of SNP rs8039957 (associated with breast cancer risk as shown further below) had pairwise D' values of 0.93 with the 12-repeat allele of D15S1547, 0.86 with the 8-repeat allele of D15S520, and 0.58 with the 7-repeat allele of

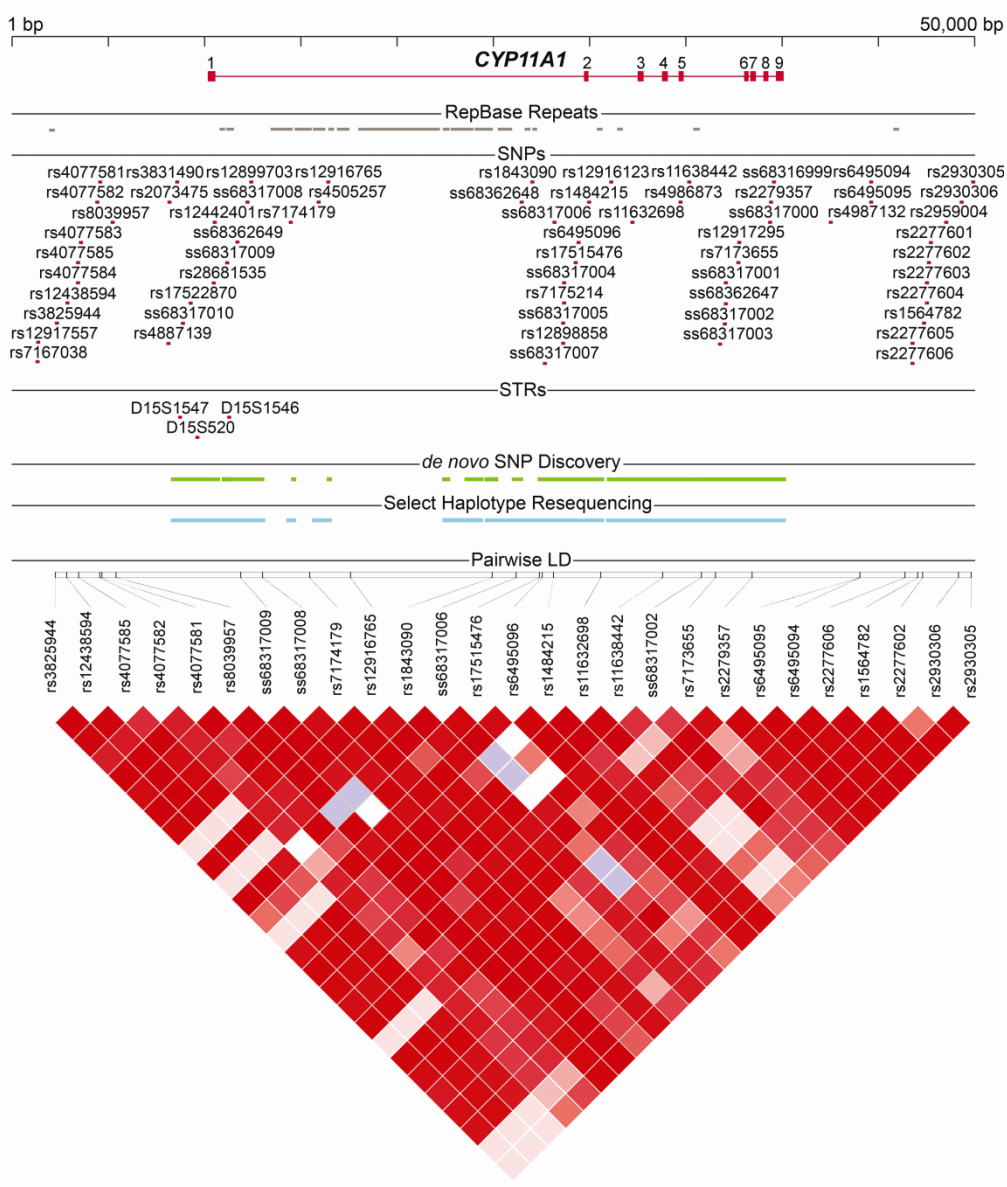


Figure 4. *CYP11A1* genetic architecture. A 50 kb interval from human chromosome 15q24.1 is depicted that encompasses the *CYP11A1* gene and 10 kb to each flank (NCBI build 36.1 from 72457199 to 72407199 bp). The gene's exons are numbered at the top with intervening introns. A 13.8 kb span of the first intron is dominated by repetitive elements. Variant sites observed among Chinese study subjects are positioned on the map. The 64 variants were identified by validation of sites annotated within dbSNP, by *de novo* discovery through SSCP / sequencing, and by re-sequencing of select population haplotypes. At bottom is a pairwise D' matrix for 356 Chinese study subjects across a subset of 27 SNPs with minor allele frequency ≥ 0.05. The matrix graph indicates relatively strong linkage disequilibrium across the locus. On the matrix, red indicates D' = 1 (LOD ≥ 2), blue indicates D' = 1 (LOD < 2), shades of pink indicate D' < 1 (LOD ≥ 2), and white indicates D' < 1 (LOD < 2).

D15S1546. Throughout the manuscript we refer to each SNP allele as that on the coding strand of the chromosome.

We sought an efficient set of tagging markers among the three STRs and 27 SNPs to capture *CYP11A1* gene diversity for tests of association with breast cancer. Eight SNPs and 3 STRs were selected as robust tagging markers, each with an allele in pairwise LD with an allele of remaining markers with an $r^2 \geq 0.70$ for the control group. This set of markers included: rs8039957, D15S1547, D15S520, D15S1546, ss68317008, ss68317006, rs1484215, rs11638442, rs7173655, rs2279357, rs2277606. Four SNPs at map ends (rs3825944, rs12438594, rs4077585, and rs2930306) were less efficiently tagged by the set, with maximal r^2 values ranging from 0.57 to 0.66.

Diplotypes of the 356 Shanghai Breast Cancer Study subjects were inferred for frequency estimation. Figure 5 illustrates haplotypes inferred by PHASE with a probability of ≥ 0.99; these are presented in an order predicted by cladistic modeling. Each haplotype has an identifying number from 1 to 57 (assigned by order of decreasing haplotype frequency). These haplotypes account for 88% of all *CYP11A1* haplotypes in this population. Only 5 haplotypes were present with greater than a frequency of 0.05.

The STR alleles marked predominant SNP haplotypes well, in concordance with the high measured pairwise LD values. Among more closely related SNP haplotypes (proximal in Figure 5), STR alleles do deviate from the principal one, and tend to do so by one or two repeat increments. This may reflect a stepwise rather than stochastic

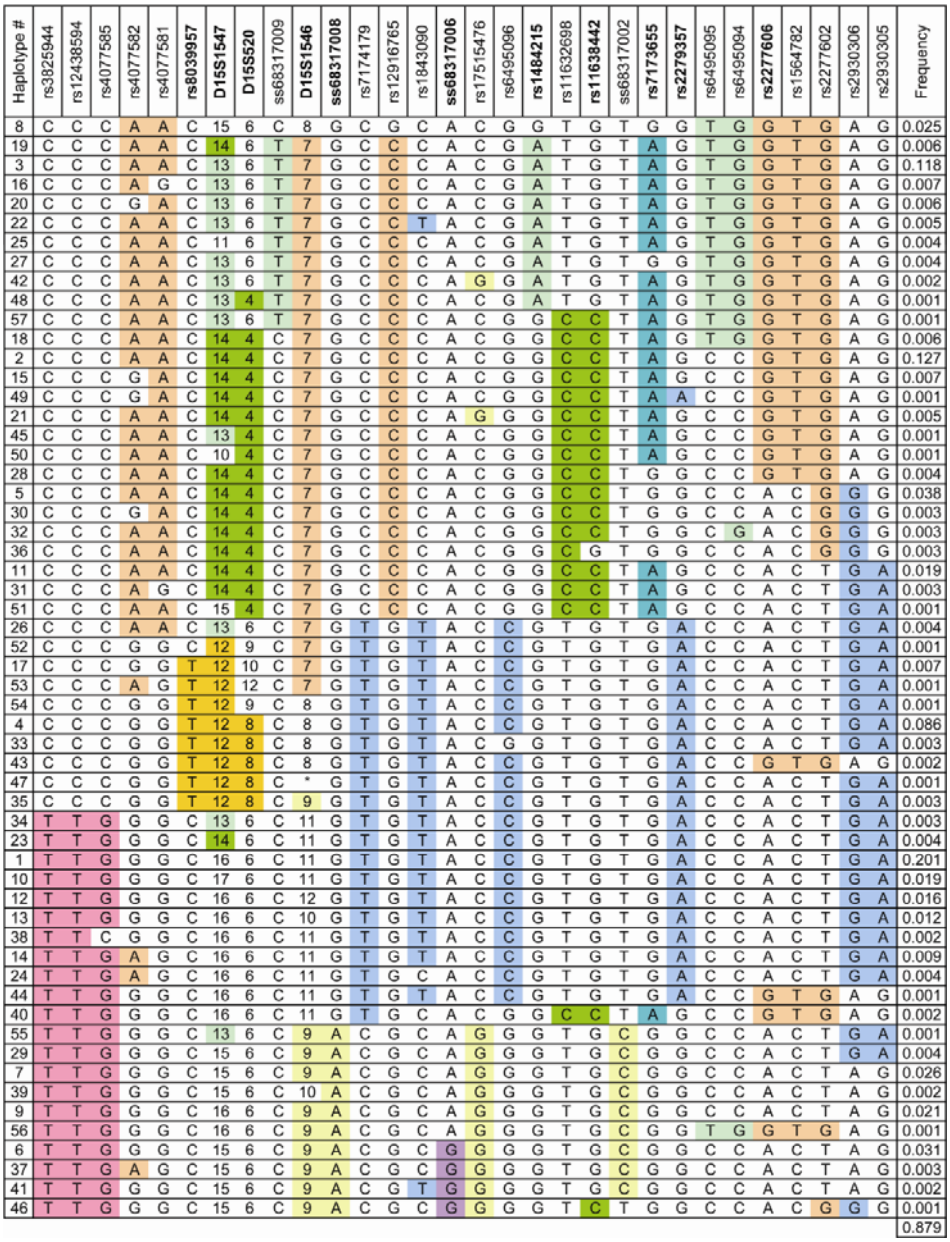


Figure 5. *CYP11A1* haplotypes among 356 Chinese study subjects, organized by cladistic similarity. 57 haplotypes are predicted among subjects with a probability ≥ 99%. Haplotypes are numbered in order of decreasing frequency. Each haplotype is designated by SNP allele and STR repeat count. Where an STR allele length was other than a multiple of the repeat unit, an asterisk is given. Among the 27 SNPs (designated by rs# or ss#) those selected as tagging SNPs are indicated in bold font. STRs (designated by D15S#) are also in bold font. Alleles are color-coded to indicate membership in LD bins where pairwise r^2 values are ≥ 0.7.

mutational mechanism [170;171]. Typical STR polymorphisms have alleles varying in increments of the repeat unit. D15S1547 is a dimer, D15S520 is a pentamer, and D15S1546 is a tetramer. However, D15S520 is distinct because it is comprised of a pentamer repeat unit while predominant population alleles are in increments of 10 bp. We subcloned and re-sequenced each of the major alleles to confirm this.

We next genotyped the set of 11 tagging markers in 1159 breast cancer cases and 1236 controls of Shanghai Breast Cancer Study population in order to explore *CYP11A1* contribution to breast cancer risk. Data was obtained on 94% of genotypes sought (per marker range 86% - 98%). Each of the tagging SNPs and STRs were in HWE ($P \geq 0.05$). Table 3 presents single allele association results comparing the case and control groups for these markers. The most significant evidence of association is observed at the three most 5' markers, each just upstream of the *CYP11A1* coding region. Significance estimates by permutation testing range from $P = 2.0 \times 10^{-5}$ to 4.1×10^{-4} for one allele at each of these markers. Each of the risk alleles observed in single allele association tests (the T allele of rs8039957, 8-repeat allele of D15S520, 12-repeat allele of D15S1547, and 7-repeat allele of D15S1546) mark closely related haplotypes in the 5' end of the *CYP11A1* gene, predominated in prevalence by haplotype #4 of Figure 5 (frequency 0.086).

We explored haplotype association employing a sliding window approach across the tagging marker *CYP11A1* map. This implicates a haplotype over the 5' region of the *CYP11A1* gene in breast cancer risk. Figure 6 presents haplotype association results for a

Table 3. *CYP11A1* alleles and breast cancer risk

Marker	Allele	Cases *n* (%)		Controls *n* (%)		*P*-value
rs8039957	C	1776	(86.1)	2014	(90.0)	$5.9x10^{-5}$
	T	288	(14.0)	224	(10.0)	
D15S1547	12	292	(13.8)	226	(10.3)	$4.1x10^{-4}$
	13	356	(16.8)	415	(19.0)	0.063
	14	501	(23.7)	530	(24.2)	0.642
	15	267	(12.6)	281	(12.8)	0.783
	16	644	(30.4)	676	(30.9)	0.716
	others	58	(2.7)	60	(2.7)	
	Overall χ^2 = 13.8 (*P*-value = 0.017)					
D15S520	4	528	(23.4)	550	(23.1)	0.807
	6	1410	(62.4)	1579	(66.3)	0.005
	8	292	(12.9)	219	(9.2)	$2.0x10^{-5}$
	others	28	(1.2)	32	(1.4)	
	Overall χ^2 = 17.5 (*P*-value = $6.1x10^{-4}$)					
D15S1546	6	925	(42.5)	1012	(44.5)	0.164
	7	329	(15.1)	284	(12.5)	0.011
	8	281	(12.9)	276	(12.1)	0.414
	10	562	(25.8)	618	(27.2)	0.292
	others	81	(3.6)	84	(3.7)	
	Overall χ^2 = 7.9 (*P*-value = 0.095)					
ss68317008	A	285	(12.8)	283	(11.9)	0.370
	G	1943	(87.2)	2087	(88.1)	
ss68317006	A	2117	(94.3)	2249	(95.6)	0.042
	G	127	(5.7)	103	(4.4)	
rs1484215	A	380	(17.2)	455	(19.5)	0.042
	G	1836	(82.9)	1883	(80.5)	
rs11638442	C	569	(26.0)	573	(24.8)	0.355
	G	1623	(74.0)	1741	(75.2)	
rs7173655	A	797	(35.9)	878	(37.9)	0.166
	G	1421	(64.1)	1438	(62.1)	
rs2279357	A	947	(42.7)	963	(41.6)	0.416
	G	1269	(57.3)	1353	(58.4)	
rs2277606	A	1347	(62.3)	1342	(59.2)	0.031
	G	815	(37.7)	926	(40.8)	

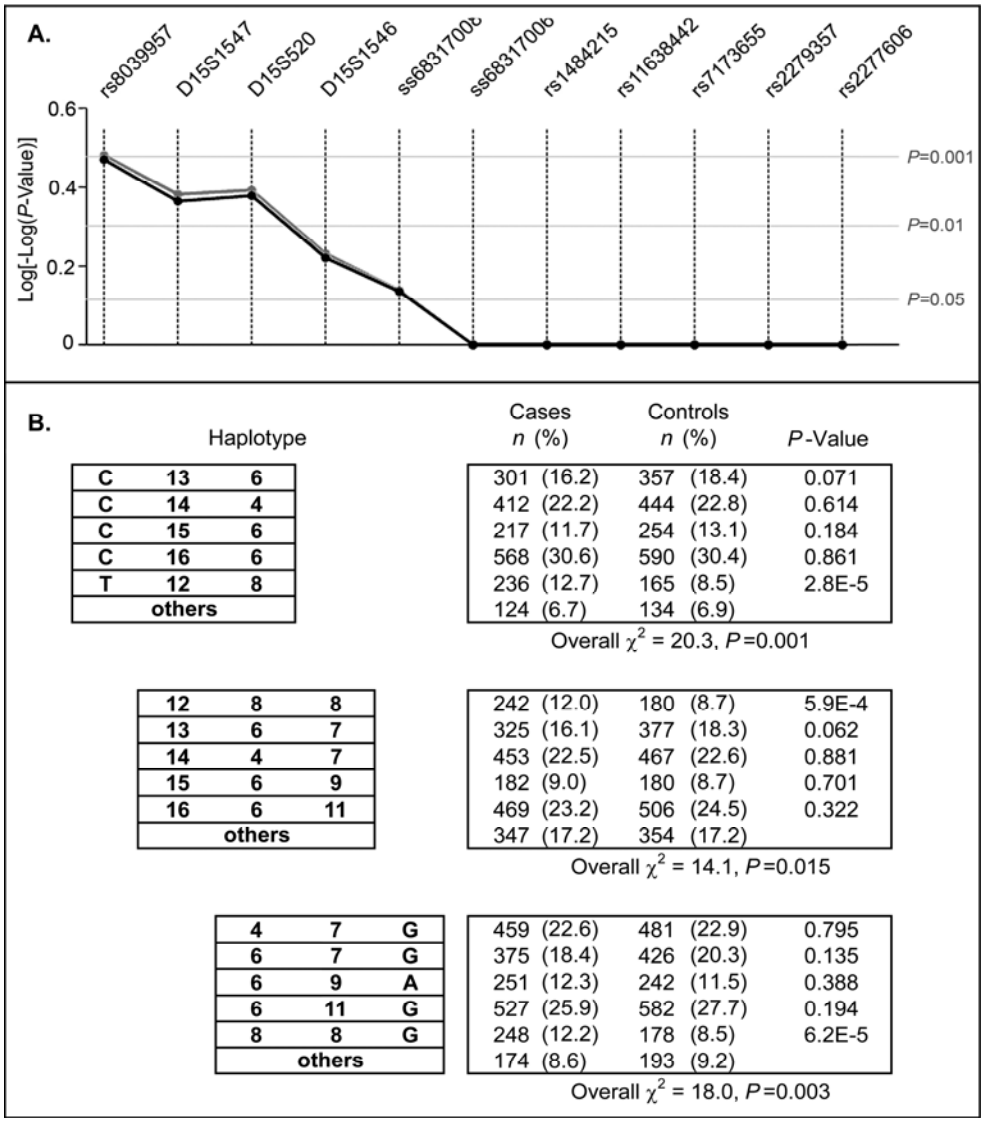


Figure 6. *CYP11A1* haplotypes and breast cancer risk. Panel A depicts the significance of the *CYP11A1* haplotype profile association with breast cancer for a window of 3 adjacent tagging markers, sliding across the map in single marker increments. The red line illustrates case and control group frequency comparisons with haplotype phase estimation by the EM algorithm. The blue line illustrates the case and control group comparison where individual study subject diplotypes were estimated by PHASE. Each graphed data point represents the average of the log[-log (*P*-value)] transformed significance levels from overall χ^2 tests that included the marker. For reference, the transformed significance levels of $P = 0.05$, 0.01, and 0.001 are provided. Panel B details evidence of association of each individual 3-marker haplotype of the *CYP11A1* promoter region with breast cancer. Each haplotype is designated by tagging SNP allele and STR repeat count.

window of 3 adjacent markers moved across the gene map in single marker increments. The overall analysis was of windows ranging in size from 2 to 11 markers. Within each window we inferred case and control haplotype frequencies by two independent methods: 1) by estimation of group frequencies using the EM algorithm, and 2) by assignment of individual study subject diplotype using PHASE. The two approaches yielded fully concordant results. Tests included those assessing overall haplotype profile differences between cases and controls (e.g. Figure 6, panel A), as well as those assessing excess of a given haplotype among cases relative to controls (e.g. Figure 6, panel B). A significant overall haplotype frequency profile difference between case and control groups was observed for all windows of width 2 - 6 markers that included any of the three most 5' markers. Significant overall haplotype profile differences were observed for 70% of windows of any width that included at least one of these markers (peak $P = 4.2 \times 10^{-4}$). A total of 55 windows were evaluated. These multiple comparisons were not independent, thus the Bonferroni corrected $P = 0.023$ is conservative. Within each significant window individual haplotype comparisons were uniformly consistent with an excess of haplotype #4 (Figure 5) in cases relative to controls (peak $P = 1.6 \times 10^{-5}$, conservatively corrected by the factor of 585 haplotypes tested at the 55 windows to $P = 0.009$). These analyses identify the promoter region of the *CYP11A1* gene as a source of breast cancer risk in the study population.

We employed logistic regression adjusted for age to assess the effect size of haplotype #4 relative to other haplotypes as a group. We evaluated the upstream promoter region of haplotype #4, delineated by markers rs8039957, D15S1547, D15S520 and

D15S1546. The resulting estimates for the risk haplotype [T_12-repeat_8-repeat_7-repeat] are presented in Table 4. Inheritance of a single copy of the haplotype confers a 1.51-fold (95% CI 1.19-1.91) significantly increased risk for breast cancer, and inheritance of two copies doubles this risk to 2.94 fold (95% CI 1.22-7.12). Evaluation of sub-haplotypes of 2 or 3 markers within this region yields similar results.

Table 4. *CYP11A1* promoter haplotype effect size upon breast cancer risk

Presence of haplotype T-12-8*	Cases *n* (%)		Controls *n* (%)		OR (95% CI)**	*P*-Value
None	710	(88.5)	811	(91.6)	1.0 (reference)	
One copy	200	(10.5)	151	(8.0)	1.51 (1.19 - 1.91)	0.001
Two copies	18	(1.0)	7	(0.4)	2.94 (1.22 - 7.12)	0.017
Trend Test						5.0×10^{-5}

* Haplotype alleles of markers rs8039957, D15S1547, D15S520
**Odds ratios are adjusted for age

We reasoned that the list of potential functional candidate variants conferring disease risk would include: 1) the alleles of the four markers above, and 2) alleles of other markers in strong LD with them, whether known or unknown. Among the 356 study subject subset, the maximum pairwise r^2 of alleles of any other known marker with the four alleles of interest was 0.23. In contrast, the allele T of rs8039957, 12-repeat of D15S1547, and 8-repeat of D15S520 had pairwise r^2 values ranging from 0.86 to 0.93, and each had weaker LD (r^2 range 0.58 to 0.66) with the 7-repeat allele of D15S1546. Based upon direct sequencing data of the promoter region, one of the database-screened SNPs (rs4887139) that had failed assay development for the 356 subjects also potentially marked the disease-associated haplotype with a C allele. Additional unknown markers that might demonstrate strong LD with alleles of the disease-associated haplotype were of concern because the SSCP methods that we employed for variant discovery at the

CYP11A1 gene lack complete sensitivity. Thus, we further searched for undiscovered variants that might also be functional candidates by re-sequencing the *CYP11A1* genomic region of a Chinese cell line and a Shanghai Breast Cancer study case that we had characterized as homozygous for haplotype #4. Two cell lines and four case subjects harboring common alternative haplotypes were also sequenced for comparison. Discovered variants were then genotyped in the 59 Chinese cell line DNAs in order to assign alleles to known haplotypes. This effort led to the discovery of an additional SNP (now designated rs12442401) in the first intron whose minor allele appeared to directly mark the disease-associated haplotype. Data to support assignment of the new SNP's minor allele to the disease haplotype was limited to that derived from sequencing, as we failed to develop a reliable genotyping assay for the marker. To summarize, the original four disease-haplotype marking variants, and the additional rs4887139 and rs12442401 each are candidates that may be functional in the phenotype. The disease-associated haplotype as defined by the full complement of observed variant sites is provided in Table 5.

The evidence that these experiments uncovered supports a role for common *CYP11A1* promoter variation in breast cancer risk. Although *CYP11A1* expression is greatest in steroidogenic tissues, it is also expressed in lymphocytes [172]. Because we had identified the *CYP11A1* diplotype for each of 59 Chinese lymphoblastoid transformed cell lines, we subsequently evaluated expression of a cell line homozygous for the disease-associated haplotype (#4 of Figure 5) and compared to expression of two cell lines homozygous for alternative common haplotypes (#'s 1 and 3). *CYP11A1* expression

Table 5. Alleles at observed variant sites of *CYP11A1* haplotype 4

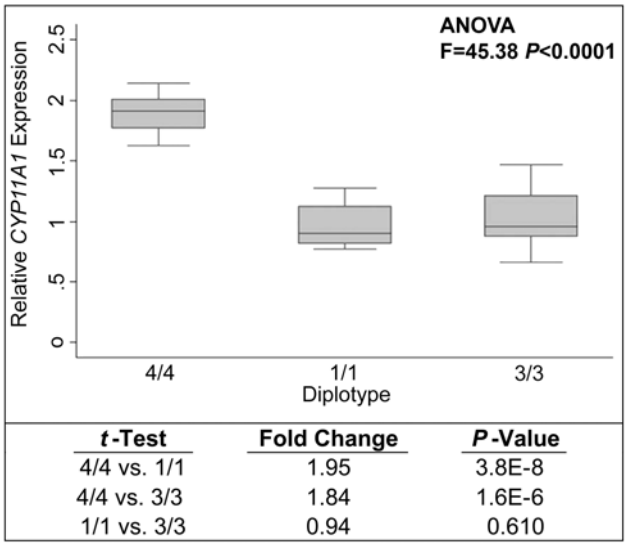
	rs7167038	rs12917557	rs3825944	rs12438594	rs4077584	rs4077585	rs4077583	rs4077582	rs4077581	rs8039957	rs4887139	rs2073475	D15S1547	ss68317010	rs17522870	D15S520	rs28681535	rs12442401	ss68317009	D15S1546	ss68362649	ss68317008	rs7174179	rs4505257	rs12916765	ss68362648	rs1843090	ss68317007	ss68317006	rs12898858	ss68317005	
5'	G	C	C	C	A	C	A	G	G	T	C	G	12	C	G	8	C	T	C	7	C	G	T	A	G	T	T	T	A	T	C	
	rs7175214	ss68317004	rs17515476	rs6495096	rs1484215	rs12916123	rs11632698	rs4986873	rs11638442	ss68317003	ss68317002	ss68362647	ss68317001	rs7173655	rs12917295	ss68317000	rs2279357	ss68316999	rs4987132	rs6495095	rs6495094	rs2277606	rs2277605	rs1564782	rs2277604	rs2277603	rs2277602	rs2277601	rs2959004	rs2930306	rs2930305		
	C	A	C	C	G	T	T	C	G	C	T	G	T	G	G	G	A	C	C	C	C	A	C	C	C	T	T	G	A	G	A	3'

was measured in total RNA prepared from the cell lines using a 5' fluorogenic nuclease quantitative real-time PCR assay, normalizing to expression of 18s rRNA. Within these cell lines the expression of the disease-associated haplotype was roughly twofold greater than that of either alternative haplotype tested (Figure 7). Increased relative expression is consistent with increased risk for breast cancer conferred by the promoter haplotype.

Discussion

We have conducted a detailed linkage disequilibrium study of the *CYP11A1* gene and demonstrated that a common promoter haplotype is associated with both increased expression and increased risk of breast cancer. Select alleles of three markers upstream of the coding region (rs8039957, D15S1547, D15S520) define the haplotype. Alleles of two additional nearby markers, rs4887139 and rs12442401, also potentially mark the haplotype of interest. An allele of D15S1546 of the first intron demonstrated less LD with alleles of the associated haplotype, and weaker association with breast cancer risk. As currently delineated, the etiologic haplotype resides in a small 4-5 kb region spanning the *CYP11A1* promoter and would have been detected in a HapMap-based study design by virtue of selection of tagging SNP rs8039957. In HapMap data of Chinese from Beijing, this SNP is in full LD with rs4887139 and with rs4278698 (a SNP that failed our assay design process).

Our observations are consistent with the important role of the cholesterol side chain cleavage enzyme in steroid sex hormone biosynthesis, and with epidemiological studies implicating estrogen biosynthesis and metabolism in breast cancer etiology [153].



t-Test	Fold Change	*P*-Value
4/4 vs. 1/1	1.95	3.8E-8
4/4 vs. 3/3	1.84	1.6E-6
1/1 vs. 3/3	0.94	0.610

Figure 7. *CYP11A1* expression in lymphoblastoid cell lines. Expression within cell line GM16654 (homozygous for the breast cancer associated haplotype 4) is shown relative to that of cell lines GM17020 and GM17014 (homozygous for common haplotypes 1 and 3). *CYP11A1* expression is normalized to 18s rRNA levels. Each box plot presents nine independent measurements of expression within a given cell line (median, box range 25^{th} - 75^{th} percentile, whiskers of data within 1.5 fold the interquartile range). Significance is presented rejecting the hypothesis that all three expression levels are the same. Pairwise comparisons of cell line expression are also made, rejecting the hypothesis that expression of the 4/4 diplotype cell line is the same as that of each other cell line.

We estimate that population-attributable risk of the 5' regulatory region haplotype of *CYP11A1* is 6.9%. This reflects an important contribution to breast cancer in the Chinese population. HapMap data for the CEU study population suggests a higher frequency of this haplotype (defined by rs8039957 allele T, rs4887139 allele C, and rs4278698 allele A) among Caucasians than among Chinese [173].

Because tissue-specific regulatory elements of *CYP11A1* are known to function in ovary and adrenal, it is conceivable that promoter haplotypes may be correlated selectively with pre- or post-menopausal breast cancer, reflecting the relative tissue origin of steroidogenesis. Cases of the Shanghai Breast Cancer Study are predominantly (71%) pre-menopausal and evidence of association is strongest in this group [149]. Intriguingly, Setiawan *et al*. also found evidence of association between a haplotype over the 5' region of the *CYP11A1* gene and breast cancer risk in the Multiethnic Cohort Study [174]. However, the risk haplotype that they identified (similar to haplotype #3 of Figure 5) is distinct from the risk haplotype (#4) identified in our study. Cases of the Multiethnic Cohort Study are predominantly (69%) post-menopausal. The risk haplotype of the Setiawan *et al*. study is tagged by rs3803463 at -7542 bp upstream of the gene, a marker also in LD with rs1484215 between exons 2 and 3 (r^2 range 0.75 to 0.87 in HapMap populations). In light of these collective findings, further epidemiological evaluation of *CYP11A1* haplotypes in pre- and post-menopausal breast cancer, and investigation of their impact on tissue-specific expression is warranted.

The structure of the *CYP11A1* gene promoter has been extensively investigated in prior studies [175-187]. The proximal promoter is comprised of a TATA box, a highly conserved SF1/LRH-1 site, and two SP1 sites. Promoter deletion mapping has also identified a negative regulatory element residing between -300 and -660 bp [175;181;185]. This region harbors non-conserved repetitive elements flanking D15S520 at -487 bp. The non-conserved CA simple sequence repeat D15S1547 at -1361 bp is also adjacent to a repetitive element. The upstream cAMP-response sequence at -1540 to -1640 bp harbors two AP1/CREB binding sites flanking an SF1 site. Further upstream are two adrenal-specific enhancers between -1840 and -1900. SNPs marking the disease-associated haplotype, rs4887139 at -2228, rs8039957 at -4884 bp, and potentially rs4278698 at -4984, do not reside within conserved regions. Both the [AAAT] simple sequence repeat of D15S1546 and SNP rs12442401 reside within non-conserved regions of the first intron. All identified variants of the disease-associated haplotype thus fall outside of conserved elements defined by vertebrate Multiz alignment in the *CYP11A1* region, but the D15S520 repeat $[TAAAA]_n$ potentially resides within a described functional promoter element.

A [TAAAA] polymorphic repeat has been demonstrated to be a negative regulatory element within the promoter of the plasma sex hormone-binding globulin gene (*SHBG*) [188]; 6- to 11-repeats reside at -726 bp of that promoter. Reporter constructs carrying 6-repeats showed significantly less transcriptional activity than constructs carrying other repeat lengths. The 6-repeat version of the *SHBG* promoter is also associated with lower SHBG levels [189]. The 6-repeat allele of D15S520 was the most

commonly observed in our study, and two cell lines homozygous for the 6-repeat allele each had significantly lower *CYP11A1* expression than a cell line homozygous for the disease-associated 8-repeat allele. The two promoter repeats may not be fully analogous however, since increased risk of breast cancer in our study was associated only with the 8-repeat allele at *CYP11A1*, not with other non-6 repeat alleles. An even and odd number of repeats could alternatively orient closely flanking transcription factors on the same or opposite DNA helical faces to influence interactions; odd repeat alleles of D15S520 were relatively rare in both the Shanghai Breast Cancer Study and in the Multiethnic Cohort Study [174].

Heritable variation of both *cis* and *trans* regulatory elements controlling expression of steroid hormone biosynthesis and metabolism genes could greatly contribute to population breast cancer risk [190;191]. Broader investigation of this large network of genes should reasonably include genetic variation of potential regulatory elements. A genome-wide or a candidate gene association study based upon tagging SNP selection from current HapMap data could have detected association of *CYP11A1* with breast cancer risk in our study population. A direct investigation of *SHBG* promoter variation in breast cancer risk has not yet been conducted, though higher plasma levels of SHBG (with corresponding lower levels of circulating estrogen) have been associated with reduced risk for breast cancer [192]. Among other genes of the steroid hormone regulatory network, a SNP within the human progesterone receptor gene promoter, located between its two alternative isoform transcript start sites, has been shown to have a direct effect on expression [193]. That promoter variant was further associated with breast

cancer risk in the Nurses' Health Study [193]. Systematic investigation of steroid hormone biosynthesis and metabolism gene variation may provide a more comprehensive picture of the role of these pathways in breast cancer risk.

Concluding Remarks

In this study, we have successfully used the association of the 8-repeat allele of D15S520 with breast cancer as a positive control to test the ability of a haplotype-based analysis to detect variants predisposing to common disease. As a result, methods and techniques developed in this study will be utilized my future thesis work identifying an unknown variant predisposing to PrCa within a candidate interval at HPCX. First we assessed methods relating to SNP selection and the utility of labor-intensive *de novo* SNP discovery. To select SNPs used in the study, we complemented a search of dbSNP with allele discovery. In this way, we were able to assess the comprehensiveness of dbSNP. One-fifth of variant sites within our genomic interval were previously undescribed, indicating that more thorough coverage of our candidate interval could be achieved through complementing SNP selection through dbSNP with allele discovery. We next examined the ability of tagging SNPs to identify a haplotype associated with risk of disease. We genotyped the full cohort of variants within a subset of our study population and determined a set of tagging SNPs and STRs across the interval. Using this set of tagging SNPs, we were able to identify a specific haplotype containing the original associated 8-repeat allele; the use of tagging SNPs allowed a reduction of workload while still enabling successful detection of a common risk allele. Finally, we successfully used sliding window haplotype analysis to implicate a haplotype over the 5' region of the gene

in breast cancer risk. This haplotype again contained the original 8-repeat allele from our previous study, indicating that a candidate risk interval could be identified by sliding window haplotype analysis.

Acknowledgements

This study was supported by a US Presidential Early Career Award for Scientists and Engineers (JRS). The Shanghai Breast Cancer Study was supported by National Cancer Institute grants R01 CA64277 and R01 CA90899 (WZ). We thank the study participants and staff of the Shanghai Breast Cancer Study.

CHAPTER IV

FAMILIAL PROSTATE CANCER RISK, AGGRESSIVENESS, AND THE TRANSFORMING GROWTH FACTOR β1 T29C POLYMORPHISM

Introduction

Family history remains the best established risk factor for PrCa. Twin studies have revealed that roughly half of the risk for PrCa is heritable – nearly twice that of other common cancers [6;24-26]. There has been intense interest in the identification of genetic risk factors that predispose to PrCa and that modify its course. Several recent pedigree-based linkage studies used Gleason score as an index of aggressiveness and concordantly identified a locus on chromosome 19q12-13 that predisposes to PrCa [57;102;103;194]. The transforming growth factor β1 gene (*TGFB1*) resides at 19q13.2 and is known to play a role in the genesis of multiple cancers [195]. Relatively little is known of its specific role in PrCa. Transforming growth factor β regulates normal prostate growth by inducing apoptosis and by inhibiting proliferation [196]. It acts as a tumor suppressor in cell culture and in mouse models of several cancers [195;197]. However, increased transforming growth factor β activity in the setting of cancer is also correlated with tumor aggressiveness [198-204]. Thus the *TGFB1* gene is a candidate tumor suppressor gene as well as a candidate oncogene that may play a role in familial PrCa.

The Contribution of the T29C Polymorphism to Common Cancers

The *TGFB1* gene harbors a well-studied T to C transition polymorphism at nucleotide position 29 that substitutes a proline residue for a leucine residue in the hydrophobic core of the signal peptide (T29C or L10P, rs1982073). This substitution alters extracellular levels of transforming growth factor β. In cell transfection studies, the proline allele functionally results in an increase in secretion relative to the leucine allele [205]. Thus, the C allele encodes the more active isoform of transforming growth factor β. Several studies have examined the *TGFB1* T29C polymorphism in breast cancer [206;207]. Meta-analyses identify T allele homozygotes as having significantly increased risk of breast cancer [208]. Dunning *et al*. estimated that 3% of all breast cancer cases may be attributable to homozygosity of the T allele [205]. Other studies have examined the role of the *TGFB1* T29C polymorphism in breast cancer aggressiveness. These studies observe increased disease aggressiveness in presence of the C allele [209-211]. Collectively, studies of the role of *TGFB1* T29C in breast cancer are supportive of *in vivo* and *in vitro* findings that transforming growth factor β modifies risk of cancer development and the aggressiveness of its course.

Evidence for *TGFB1* as a Prostate Cancer Aggressiveness Gene

The *TGFB1* gene is flanked by genetic markers associated with PrCa in multiple independent studies that employed Gleason score as a quantitative trait: D19S870, D19S875, D19S433, D19S414, D19S75, and D19S245 on the centromeric flank; and D19S178, D19S902, and D19S246 on the telomeric flank [57;102;103;194;212]. The use of Gleason score as a quantitative trait may reduce phenotypic heterogeneity and thus

simplify underlying genetic heterogeneity in this complex disease. PrCa aggressiveness was first linked to this region of 19q in a study of affected sibling pairs from Washington University, St. Louis and from Cleveland Clinic [57]. Subsequently, the finding was confirmed in an independent study of hereditary PrCa families from the Mayo Clinic [102]. A third independent study of affected sibling pairs from hereditary PrCa families from the Fred Hutchinson Cancer Research Center has also recently confirmed linkage to the region [103]. The *TGFB1* gene is a plausible candidate in this region of the genome.

Study Design

Given the weight of evidence for the role of transforming growth factor β in other cancers and the PrCa linkage evidence encompassing the *TGFB1* locus, we hypothesized that the functional T29C polymorphism of *TGFB1* modifies PrCa risk and aggressiveness. We tested the hypothesis in a study population of PrCa cases and age-matched controls consisting of Americans of Northern European descent. Each of the independent case probands came from a pedigree with at least one affected first or second degree relative. A man with a family history of PrCa has a greater potential genetic load for the disease than a man with no family history of PrCa [22]. Controls had no personal or family history of PrCa among first or second degree relatives. Our study population was divided into high and low Gleason score categories to address our hypothesis. As an assessment of study power within these two population subsets, we verified that each could detect the broadly confirmed association between PrCa and SNP rs1447295 at 8q24. This anonymous SNP dominantly confers risk of PrCa with an estimated odds ratio of ~1.7 across multiple independent Caucasian study populations [139;141;143-146;213].

We then tested for an association between the *TGFB1* variant and the PrCa Gleason score groups. Among cases with a Gleason sum ≤ 6, the more active C (proline) allele is dominantly protective (OR = 0.64) while the less active T (leucine) allele recessively confers risk (OR = 1.56). Elevated risk associated with the allele is not observed among cases with a Gleason sum ≥ 7. These observations are consistent with recent linkage analyses of hereditary PrCa employing Gleason score as a quantitative trait that have highlighted this genomic locus.

Materials and Methods

Study Population

Patients included in this study were ascertained with informed consent from Vanderbilt University Medical Center and from the VA Tennessee Valley Healthcare System with institutional review board oversight. Subjects were residents of Tennessee (75%), Kentucky (15%), Georgia (2%), Alabama (1%), Mississippi (1%), Virginia (1%), and other states (4%). Cases were ascertained at the time of treatment for the principal diagnosis of PrCa in urology clinics, and controls were ascertained at the time of routine preventative screening for PrCa in general medicine clinics. PrCa diagnoses were confirmed by review of medical records. Cases included 415 unrelated, independent Caucasian PrCa probands: 255 cases from pedigrees with two affected, 101 cases from pedigrees with three affected, and 59 from pedigrees with 4 or more affected. Each control was matched to a case on age (± 2.5 years) in a 1:1 ratio (age at screen for controls, age at diagnosis for cases). Controls included 415 unrelated, unaffected Caucasian men with no personal or family history of PrCa. Controls had a screening

prostate specific antigen (PSA) test < 4 ng/ml at the time of ascertainment, and had no record of a PSA test ≥ 4 ng/ml or abnormal digital rectal examination.

Data of personal and family history of cancer was obtained by a structured questionnaire completed by the proband, by review of their medical record, and by report from family members accompanying them at ascertainment interview. Data abstracted from the medical record included: date and results of PSA test(s), initial Gleason score grade from prostate biopsy, and subsequent prostatectomy Gleason score grade (available for 85% of cases). Analyses preferentially employed surgical specimen over biopsy data.

Genotyping

DNA was extracted from whole blood using the Puregene DNA Purification System Standard Protocol (Qiagen, Valencia, CA). DNA was quantified using the PicoGreen dsDNA Quantitation Kit (Invitrogen, Carlsbad, CA), imaged with a Molecular Devices / LJL Analyst HT (Molecular Devices, Union City, CA).

Reference SNP rs1447295 was genotyped by single nucleotide primer extension assay with detection by florescence polarization. *TGFB1* T29C (rs1982073) was genotyped by two methods: a fluorogenic 5'-nuclease assay, and a single nucleotide primer extension assay with detection by florescence polarization [157;214]. Discordant *TGFB1* T29C genotypes resulted for 17 subjects, requiring resolution by re-sequencing using an ABI 3100 automated sequencer with BigDye® terminator chemistry (Applied Biosystems, Foster City, CA).

Single nucleotide primer extension reaction processing entailed three steps: a 4.4 µl PCR reaction, addition of 4 µl of an exonuclease I (New England Biolabs, Beverly, MA) and calf intestinal alkaline phosphatase (Promega, Madison, WI) reagent mix to degrade unincorporated primer and dephosphorylate dNTPs, and a final addition of 4 µl of an Acyclopol and Acycloterminator reagent mix for the primer extension reaction (AcycloPrime^{TM} FP SNP Detection System, Perkin-Elmer, Boston, MA). Each PCR mixture included 0.1 unit AmpliTaq Gold DNA polymerase, 1x Buffer II (Applied Biosystems, Foster City, CA), 2.5 mM $MgCl_2$, 0.25 mM dNTPs, 335 nM of each primer, and 2 ng DNA template. We detected incorporation of R110- and TAMRA-labeled terminators by fluorescence polarization on a Molecular Devices / LJL Analyst HT. Primer sequences were: TGFB1 T29C (rs1982073) forward primer 5'-ACACCAGCCCTGTTCGC-3'; reverse primer 5'-CGTCAGCACCAGTAGCC-3'; extension primer 5'-GCAGCGGTAGCAGCAGC-3'. SNP rs1447295 forward primer 5'-GGTAATGAACAGTTCTGTCTC-3'; reverse primer 5'-CATGAGGAAAAGTCAACAC-3'; extension primer 5'-ATTGGGGAGGTATGTAAAA-3'.

TGFB1 T29C fluorogenic 5'-nuclease assay primers included a forward primer (5'-CGCGCTCTCGGCAGT-3'), a reverse primer (5'-AGGCGTCAGCACCAGTAG-3'), a VIC probe (5'-CAGCAGCGGCAGCA-3'), and a FAM probe (5'-CAGCAGCAGCAGCA-3'). Each 5 µl reaction included 5 ng genomic DNA, 2.5 µL TaqMan 2x Universal PCR Master Mix No AmpErase UNG (Applied Biosystems, Foster City, CA), 900 nM each primer, 200 nM each probe, and 1M betaine.

Statistical Analysis

Conditional logistic regression analyses were used to estimate odds ratios and 95% confidence intervals (Intercooled Stata 9, Stata Corporation, College Station, TX). The matching variable, age at diagnosis or screening, was included as a raw covariate in the model. Gleason score was compared for the left and right sum, and the largest sum was dichotomized as ≤ 6 (moderately or well differentiated) or ≥ 7 (poorly differentiated). χ^2 contingency test was used to compare the frequency of Gleason subsets of cases dichotomized by family history (2 affected only, or ≥ 3 affected). A *P* value ≤ 0.05 was considered statistically significant.

Results

We ascertained independent cases with a family history of PrCa, and we ascertained age-matched controls that were free of a personal or family history of PrCa. The distribution of case family history by number of affected first or second degree relatives was: 14% with ≥ 4 affected, 24% with 3 affected, and 61% with 2 affected. Characteristics of the study population are presented in Table 6.

Table 6. Study Demographics

		Controls	Cases
No.		415	415
Mean Age*, y		61.1	61
Median PSA*		0.95	5.6
Median Gleason Sum		-	6
Gleason Sum ≤ 6, No.		-	220
Gleason Sum ≥ 7, No.		-	188
Pedigree Structure (# Affected)†	0	415	-
	2	-	255
	3	-	101
	≥4	-	59

*At diagnosis for cases, at entry screen for controls.
†Inclusive of proband and first and second degree relatives.

We obtained *TGFB1* T29C genotypes for all study participants, and rs1447295 genotypes for 96% of study participants. The distribution of *TGFB1* T29C and rs1447295 genotypes is presented in Table 7. The variants were in Hardy-Weinberg equilibrium among control subjects (*TGFB1* T29C $P = 0.916$; rs1447295 $P = 0.471$). The observed SNP minor allele frequencies were similar to those previously reported in samples of Caucasian populations.

Table 7. Genotype Distribution – *TGFB1* T29C and 8q24

	TGFB1 T29C (rs1982073)					
	T/T		**T/C**		**C/C**	
Controls*	148	(35.7)	204	(49.2)	63	(15.2)
Cases	174	(41.9)	188	(45.3)	53	(12.8)
	8q24 (rs1447295)					
	C/C		**C/A**		**A/A**	
Controls	334	(83.3)	65	(16.2)	2	(0.5)
Cases	296	(74.4)	95	(23.9)	7	(1.8)

* *n* (%)

The established effect of the 8q24 rs1447295 variant as a determinant of PrCa risk suggested that we would have a power of 0.91 to detect the association within our study population [139;141;143-146;213]. We assessed its role in the study population using conditional logistic regression incorporating the matched study design (1:1 matching of each case to a control by age) (Table 8). In a dominant inheritance model there was a significant association of the minor allele with risk of PrCa (OR = 1.86, 95% CI 1.30-2.67, *P* = 0.001). In a recessive inheritance model the major allele significantly conferred protection (OR = 0.54, 95% CI 0.37-0.77, *P* = 0.001). With this reassurance of the ability to detect the known PrCa risk variant in the study population, we further addressed the study hypothesis at the variant of *TGFB1*.

Table 8. *TGFB1* T29C and 8q24 in Prostate Cancer

Case Strata	*n*	Minor Allele Dominant Model*	Major Allele Recessive Model†	*P*
	TGFB1 T29C (rs1982073)			
All	830	0.79 (0.60 - 1.04)	1.27 (0.97 - 1.68)	0.088
Gleason Sum ≤ 6	440	0.64 (0.43 - 0.93)	1.56 (1.08 - 2.33)	0.020
Gleason Sum ≥ 7	376	1.05 (0.69 - 1.58)	0.95 (0.63 - 1.45)	0.831
	8q24 (rs1447295)			
All	770	1.86 (1.30 - 2.67)	0.54 (0.37 - 0.77)	0.001
Gleason Sum ≤ 6	406	1.93 (1.12 - 3.33)	0.52 (0.30 - 0.89)	0.017
Gleason Sum ≥ 7	352	2.01 (1.22 - 3.34)	0.50 (0.30 - 0.82)	0.007

*AA = Reference Group (Odds Ratio = 1.00)
†AB & BB = Reference Group (Odds Ratio = 1.00)

We investigated inheritance models of the *TGFB1* T29C variant in the PrCa study population using conditional logistic regression incorporating the matched study design. Results did not meet statistical significance for the study population prior to stratification for disease aggressiveness. In the overall study population the C allele did trend toward protection against PrCa when analyzed in a dominant inheritance model (OR = 0.79, 95% CI 0.60–1.04, *P* = 0.088). Conversely, the T allele trended toward risk for PrCa when analyzed in a recessive model (OR = 1.27, 95% CI 0.97-1.68. *P* = 0.088). These observations are consistent with relatively greater biological activity of the tumor suppressor previously established for the version encoded by the C allele, relative to that encoded by the T allele.

We then evaluated the role of *TGFB1* T29C in modifying risk for PrCa by comparing more indolent or aggressive cases to their age-matched controls. We

employed Gleason score as the index of relative aggressiveness. Gleason score was dichotomized as either poorly differentiated (≥ 7), or well to moderately differentiated (≤ 6) histopathology. This stratification of the study population retained sufficient power to detect an association between the reference SNP rs1447295 and PrCa within each respective Gleason score subgroup (Table 8). PrCa risk was significantly associated with *TGFB1* genotype among cases where Gleason score was ≤ 6 (C allele dominant OR = 0.64 (95% CI 0.43-0.93); T allele recessive OR = 1.56 (95% CI 1.08-2.33); $P = 0.020$). No evidence of association was observed among cases with a Gleason score ≥ 7. Our results suggest a significant role for *TGFB1* in modifying risk specifically for a more indolent PrCa.

Low Gleason score cases comprise 58% of the study population among those from pedigrees with ≥ 3 affected, relative to 49% among those from pedigrees with only 2 affected, a significantly different distribution ($P = 0.033$). Thus, low Gleason score and greater family history also appeared to be associated within our study population. However, among cases with a family history of only 2 affected (proband and an additional first or second degree relative), PrCa risk remained significantly associated with *TGFB1* genotype in the low but not the high Gleason score subset (C allele dominant OR = 0.56 (95% CI 0.33-0.94); T allele recessive OR = 1.79 (95% CI 1.06-3.03); $P = 0.028$).

Discussion

Three independent studies have employed Gleason score as a quantitative trait to concordantly identify a locus at 19q12-13 that is likely to carry a genetic determinant of PrCa [57;102;103]. However, the locus is not highlighted by other linkage studies of hereditary PrCa upon limiting affected status to only those men with clinically significant disease [104;105]. Our study investigated the candidate functional polymorphism T29C (L10P) of the *TGFB1* gene at 19q13.2 for its role in predisposing to PrCa and for its potential influence on disease aggressiveness. The more active proline variant of the transforming growth factor β signal peptide is encoded by the C allele in the 10th codon and is known to suppress tumor initiation, where the less active leucine variant (T allele) is associated with risk for several common cancers. We observed that the *TGFB1* variant is associated specifically with more indolent PrCa. The more prevalent T allele recessively confers risk, and conversely the less prevalent C allele dominantly confers protection.

The PrCa cases evaluated in this study were limited to those with a family history of PrCa. In a multifactorial model, the genetic load for PrCa of a given individual may increase with an increasing family history of the disease. Comparison of cases with a strong family history of PrCa to controls with none may prove more powerful in detecting these effects than an alternative case-control study design without regard to family history. Nonetheless, a recent study of the separate variant C-509T (rs1800469) upstream of the *TGFB1* gene was also recently shown to be associated with a decreased risk of aggressive (Gleason ≥ 7) PrCa in a study unselected for family history[215]. The authors did not evaluate its effect on indolent PrCa. Variants T29C and C-509T are in

partial linkage disequilibrium in genotype data submitted to dbSNP (ss15356536 and ss15356531, $r^2 = 0.71$), raising the possibility that observed clinical effects too are correlated. Additional studies are warranted to shed further light on the role of this gene in determining aggressiveness of PrCa.

In this study we evaluated SNP rs1447295 at 8q24, validated for its contribution to PrCa risk in multiple independent study populations, as a means of confirming the ability of this study population and its two Gleason subsets to detect a known risk variant. Actual power to detect an effect at *TGFB1* is a function of the known allele frequency, and the *a priori* unknown effect size. The risk allele of *TGFB1* is more frequent than the risk allele of rs1447295. *TGFB1* recessive risk allele (T/T) homozygotes comprised 36% of control and 42% of case subjects. By comparison, carriers of the dominant risk allele of rs1447295 comprised 17% of control and 25% of case subjects. Our results indicate a lesser effect size for the *TGFB1* variant than for the 8q24 variant among low Gleason score cases (OR of 1.56, versus 1.93). The risk effect of *TGFB1* is observed only within the low Gleason subset, while that of the reference SNP is observed in both low and high Gleason subsets.

Our findings at *TGFB1* T29C are consistent with prior linkage studies that have focused on PrCa aggressiveness, although other additional compelling candidate genes in the region remain to be investigated. The identification of a linkage peak at this genomic locus for Gleason score as a quantitative trait does not in itself describe the direction of the association (a high versus a low score). Given our observed association with less

aggressive PrCa, restriction of the definition of a case within a hereditary PrCa pedigree to only those men with the most clinically significant disease could reduce rather than augment a linkage signal at 19q. Such an effect is apparent in the data of Chang *et al* [104]. This raises an intriguing issue regarding the severity of PrCa in the context of hereditary PrCa. Although intuitively one might anticipate more aggressive disease in hereditary PrCa, in our study population the proportion of cases with more indolent disease increases with increasing family history of PrCa. This could be due to a bias of earlier screening among those with a more extensive family history of PrCa. Nonetheless, results suggest that genetic variants may exist that specifically predispose to *less* aggressive PrCa, and warrant future studies within independent study populations. Distinction of patients likely to suffer an aggressive course from those who will not is particularly salient in this disease.

Acknowledgements

We extend particular thanks to the study participants and to Drs Joseph Smith, Michael Cookson, Sam Chang, Richard Hock, William Maynard, Jason Pereira, and William Dupont. This work was supported by an award from the V Foundation, by a MERIT grant from the US Department of Veterans Affairs, by grant W81XWH-06-1-0057 from the Department of the Army, and by General Clinical Research Center grant M01 RR-00095 from the National Center for Research Resources, National Institutes of Health.

CHAPTER V

A HAPLOTYPE AT CHROMOSOME Xq27.2 CONFERS SUSCEPTIBILITY TO PROSTATE CANCER

Introduction

Linkage and genetic epidemiological data support the existence of genetic variants on the X chromosome that predispose to PrCa[17]. PrCa loci on both arms of the X chromosome have been identified, including the HPCX locus at Xq27-28[37;53;84;87-90;104;216]. The ~14 Mb linkage interval of HPCX was originally delineated within US, Swedish, and Finnish hereditary PrCa pedigrees[53]. Further shared haplotype analysis among Finnish probands refined the locus to a candidate interval flanked on either side by a notable 113 kb inverted repeat[92;217]. The 352 kb area between these inverted repeats was the candidate interval for the present study, which sought evidence of association with PrCa among Americans of Northern European descent. Our study population was uniquely comprised of independent familial PrCa probands, matched to controls with no personal or family history of PrCa. These two groups represent extremes of potential genetic load for PrCa. Our study included a training set of 292 case-control pairs to identify nominal associations, and a test set of 215 case-control pairs to confirm or to refute observations within the training set. We conducted extensive allele discovery and validation within the study population, characterized study population linkage disequilibrium (LD) patterns, and selected tagging SNPs for tests of association by

haplotype-based methods. Our investigation comprehensively tested association of the candidate interval with PrCa, and included non-unique genomic regions that are not amenable to current high-throughput techniques.

Materials and Methods

Study Population

Study subjects were Americans of Northern European descent, ascertained with informed consent between 2002 and 2007 from Vanderbilt University Medical Center and from the VA Tennessee Valley Healthcare System with institutional review board oversight. Subjects were residents of Tennessee (75%), Kentucky (15%), Georgia (2%), Alabama (1%), Mississippi (1%), Virginia (1%), and other states (4%). Familial PrCa cases were ascertained at the time of treatment for the principal diagnosis of PrCa, and controls were ascertained at the time of routine preventative screening for PrCa. All PrCa probands included in the study are from pedigrees with a family history of PrCa, and all control probands are from pedigrees without a family history of PrCa. Family history included first and second degree relatives. Controls had a screening prostate specific antigen (PSA) test < 4 ng/ml at the time of ascertainment, and had no record of a PSA test ≥ 4 ng/ml or abnormal digital rectal examination. Each control was matched to a case on age ($\pm$ 2.5 years; age at screen for controls, age at diagnosis for cases). Case and control pedigrees were of comparable size. The mean number of at-risk male siblings was 1.8 for controls, and 1.7 for cases. Initial accruals included 292 unrelated, independent familial PrCa probands and 292 age-matched controls, comprising a training study group. Subsequent accruals included 215 additional unrelated, independent PrCa probands and

215 additional age-matched controls, comprising a separate test study group. Analyses preferentially employed prostatectomy specimen over biopsy Gleason score (available for 87% of cases). Table 9 provides characteristics of the study population.

Table 9. Study Population

		Training		Test		Combined	
		Controls	**Cases**	**Controls**	**Cases**	**Controls**	**Cases**
No.		292	292	215	215	507	507
Mean Age*, y		63.4	61.3	60.7	60.6	62.3	61.0
Median PSA*		0.95	5.7	0.92	5.6	0.92	5.7
Median Gleason Sum		-	6	-	6	-	6
Gleason Sum ≤ 6, No.		-	145	-	114	-	259
Gleason Sum ≥ 7, No.		-	130	-	96	-	226
Pedigree # Affected†	0	292	-	215	-	507	-
	2	-	184	-	142	-	326
	≥3	-	108	-	73	-	181

*At diagnosis for cases, at entry screen for controls.

†Proband plus 1st and 2nd degree affected relatives

SNP Genotyping

DNA was extracted from whole blood using the Puregene DNA Purification System Standard Protocol (Qiagen, Valencia, CA). DNA was quantified using the PicoGreen dsDNA Quantitation Kit (Invitrogen, Carlsbad, CA), imaged with a Molecular Devices / LJL Analyst HT (Molecular Devices, Union City, CA). We genotyped SNPs by single nucleotide primer extension and fluorescence polarization, as previously described[218]. Both forward and reverse strand extension primers were tested to select the

most robust assay. Amplimer and extension primer sequences for tagging SNPs are provided in Table 10.

SNP Selection

To capture genetic diversity across the candidate interval, database SNPs annotated in dbSNP were screened for common polymorphism in the study population. This included 415 annotated SNPs on chromosome X between positions 140,036,557 and 140,388,361 (NCBI Build 36.1). These were genotyped to assess polymorphism in a screening set of 40 familial PrCa probands. The screening set was estimated to provide 98% power to detect a polymorphism with a minor variant frequency of 0.10, and 87% power with a frequency of 0.05. These 40 PrCa cases were also used for *de novo* SNP discovery at known and predicted genes within the candidate interval: 4.6 kb 5' to 0.2 kb 3' of *LDOC1*; 1.6 kb 5' to 4.4 kb 3' of *SPANXC*; 3.0 kb 5' to 1.4 kb 3' of a predicted coding region containing homology to ribosomal protein L44 ("*hRPL44"*); and 2.3 kb 5' to 0.8 kb 3' of a predicted pseudogene containing homology to *RBMX2 ("RBMX2P1").* The latter two annotations were identified with custom software. We employed two single-stranded conformation polymorphism methods (redundant) and re-sequencing for SNP discovery, as previously described[218]. Exons of the four genes were also re-sequenced for all 40 PrCa cases in the screening set.

Nested Amplification of Non-unique Regions

Non-unique regions of *SPANXC*, *hRPL44* and *RBMX2P1* were assayed using a nested reaction strategy. Using custom software, unique priming sites were identified

Table 10. Tagging SNP Assays

Marker	Forward Primer	Reverse Primer	Allele	Strand	Extension Primer
rs11095852	AATGGTCACTTGGCCAC	AGGGTGCTCTATGGTGTG	C/A	F	TGGTTTTCTCTCCTAAACAC
rs5907823	CAAAACAATTCACTGCC	TCTCTTAATGTTGCTGTTGC	C/A	F	ACTCCGTCTCAAAAAAAA
rs7880499	GGTTCCAACTCATACTCTG	CTTGCTTAGAGAGCATACAA	G/A	F	CTGCCAAAAGAAGATTTT
rs1016824	CCTAGGGTTATTATGTAGC	CATAGGGAAAGAGGTATATAGAAAG	C/T	F	TAGGGTTATTATGTAGCAGGTAC
rs12156848	GGACCATATGAGAAGAAGCTC	TAGCCAGTAGCTGTGTAGTGG	G/A	F	TTGGGAAATGCAAATTATA
rs7885649	CATGTTACTGAGTAAAAG	CACCTGGCCCATAATTTC	G/A	F	ATAGCTAGAATTGTACTGGTTCTAC
rs5953563	CTGTAAATGCAGGAGTGTC	TAGCACATATCACAGTGTG	A/T	F	GCATTAACCATATGCTGTATTT
rs5954218	CAACTGAGCTCAGTGTGAC	GGCAATTTGTGAAACACC	A/T	F	AACAGACTATAAGAGTCAGCATT
rs5954222	TTATGAGAAAGACCCACTAG	CTTTTCACCAACCAGTTTC	C/T	F	AAGAAGAAAATTATAGGAGGG
rs5907828	CCCGGGAGTTTGAGGCTA	GTGGTTTTAATAGAACTAC	C/T	F	CCCTGCCTGTAAATAAATAAA
rs7051363	CCCTCAAGTCTTCTGAAA	CTGCTGGTCTTGTTCAAA	G/A	F	GAAAGGGAAAGATAGAGTCTC
rs769077	CTTAATACAAGCTCCAACG	GACTTCCACATTCTCTTTC	C/T	F	AAAAGAGGTGGGGAAAA
rs12862529	GAACTACTACCTCATACTG	GATTGTGCTTTTCATGTC	C/T	F	GTGTTAATTGTTGTGACCC
rs7883897	GCTATGTTAAAACAAATGG	GCCTTTAAATAGTCCAGACA	G/A	F	TAATGATGTCACTTATAAAGTTGATA
rs710106	GTGCTTGGAAGTGTCATC	CTCGAAAGGTGGTAGTCTG	G/A	F	AACTGCGCAGACCACCC
rs4824993	GTGAAAGGATAAGCAGTG	GGCCAGAGACCTACTTT	C/T	F	TTACTAGCTGCTCCGTAAA
rs6636233	GGCCTAGAGCACACTTTC	GCATGAAGGTATAGCACC	C/T	F	GGCTGTCACAATGACTCA
rs3761561	CGCTTCAGTTTCTTAGATGA	ATCTAGTGCAGCACACTG	G/A	F	ACTGGAAAATCAGCTTTCT
rs5954233	GTTCTCTCCTCCAGCAGA	AGGGCTGTAGAAGCTCAG	C/T	F	GTCGCTGACGGTTTCTA
rs12392927	CATGTGGTGTTAAGGCTG	TGTGAGTGCTCCAATCC	G/A	F	AAGTCTGCTGCAGGGGT
rs11095854	CATGTGGTGTTAAGGCTG	TAGGAGAATTGCTTGATG	C/T	F	AGAATGATATAGTTTGAATTTGTG
rs1012777	GCCATCTACAAGCTAAGG	GCTGCTATAACAAAGAAGC	C/T	F	TTAAGCCACCCAGTCTG
rs845173	CCTAGGTTTGCAGAGAAAATAG	CATAAAGGTCCTTTTGC	C/T	F	TGGATACTTAAAGGTAAAATTAAAG
rs845171	TTTTCTGCTGCTTTACTCCTC	CCTTGAATTGCAGGTGATAT	C/T	F	GCTCCAAAGTGAAATGG
rs1099501	CTAATTGCTACGTGTGAG	GACCTTGAGACGTTTAAG	G/A	F	TCTTTATATATATTACTGATTGATTCATT
rs845169	CTAATTGCTACGTGTGAG	GACCTTGAGACGTTTAAG	C/T	F	TTCTGTTTTGCTTAATATTGATAT

Table 10 (continued). Tagging SNP Assays

Marker	Forward Primer	Reverse Primer	Allele	Strand	Extension Primer
rs1884417	CATCTCTAGGTTTCTGGAGAC	GCTATGGAATAGTAGCTGG	G/A	F	ACATTTGTACACACATACCCT
rs5954252	ATAGGTAGCTTTTCAACCCTC	CGGAATACTATGCAGCC	G/A	R	TTTATAAGCAAGAGCTAAACATC
rs5953618	CTTTGCAGGTATTTCAACC	CTGAGTCCTCGACCATAC	G/C	F	TCGGGCGTGGTCATTCA
rs2933670	CTTTGCAGGTATTTCAACC	CTGAGTCCTCGACCATAC	A/C	F	TCTGGAGATGTTCTTTTCA
ss78456785	GCGTGGTCATTCAGCAGTTCCTC	CTTCTCTGGATCAAACC	G/T	R	GACCCGCAACCTGCTCC
ss78456788	GCGTGGTCATTCAGCAGTTCCTC	CTTCTCTGGATCAAACC	G/C	F	GCGGGTCTGAGTCCCCA
rs5953547	CAGGATAGAGACTGGATAGC	ACTTTGACCAAGGTCTG	C/T	R	AAACCCCTTCCTCAACC
rs2057217	ATCCTGCCTAACCACCTG	TGGGGTGCTTGTAGGTAG	G/A	F	ATATTCCACCAGAAAAAGG
ss78456791	CACAATGGTCTGCAATATTC	GCAGACATTGAAGAACC	G/A	F	CTTCACTTCAGAACCTAACA
ss78456793	CACAATGGTCTGCAATATTC	GCAGACATTGAAGAACC	G/A	R	AATCCAACGAGGTGAAT
ss78456795	TTGGATTCACAGGGGAC	TGGGACACTGCCTGTATG	A/T	F	TGTATATATTGGTTCTTCAATGTC
rs2144605	GCAAATTTCAACCCATG	GGGCAACAACAGTGAAAC	G/A	R	GCCATTATTTCAGATTGGA
rs3976442	TTTGGTTACTTCTGTAGC	CCAAAAGAGACATATTGGTC	G/A	F	CATACAACAATTTAACATAAAGTTTAT
ss78456800	TTTGGTTACTTCTGTAGC	CCAAAAGAGACATATTGGTC	G/C	F	TTAACATAAAGTTTATAATTAATAACATACA
rs5953578	TTTGGTTACTTCTGTAGC	CCAAAAGAGACATATTGGTC	G/T	F	AATAACATACAGAAGTTATTTCAGAA
rs2208264	TTTGGTTACTTCTGTAGC	CCAAAAGAGACATATTGGTC	G/C	F	CCCAAAGAAAGTCAAACA
rs845144	CCTATGAGGCCAAGTTTG	GAATTAAATGGGCAGTGTC	C/T	R	TCCCGGGTTCAAGCAAT
rs714075	CCTATGAGGCCAAGTTTG	CCCATTGCTACAAACTCG	G/A	F	GAAGAGATTTATGGGACCA
rs714076	CCAGTGATGACATTTCAG	CCAAGTTTGTAAACAGGG	G/A	F	CCATCCAGAAATGCTCT
rs845150	CCAGTGATGACATTTCAG	CGTGATAGAACTGCCAGC	G/A	F	TTGTTACCATCTTCAAATGAC
rs5907844	GTTGGAGGATAACTCATAC	CTCTTCACACACACCATG	G/A	F	AAGTCGTCGTGGACATAC
rs881223	TTAGGGATGACATCACTG	GGTTATCCAACATAACC	G/T	R	GGATGACATCACTGTGTGTA
rs881221	TTAGGGATGACATCACTG	CTAATGGCCATCTGCCCA	G/A	R	GCTTACGCAATTGTCTTTT
rs881222	GTCAAAGTCCCATTAGGTG	GCGGCATTTCTGTTTGG	G/A	R	TTTAACATAAAATCAAATGGC
rs881219	GGGCTGTGCTTTTAATCC	CATCTTCAACTGGGGTC	G/A	F	CTGCAGTTTCAACAGCTAG
rs2864937	GCTGAACAGTCTTCAGTG	CCAGGATATCTAGCTGTTG	C/T	R	TGCATTAAAAAAAATAATTATTTC

Table 10 (continued). Tagging SNP Assays

Marker	Forward Primer	Reverse Primer	Allele	Strand	Extension Primer
rs5907848	TCAAGCCAACATTGACTTAG	GCACGTTCTGCACATGTA	G/A	F	GCTCCTACGTCTTTTAAAAAA
rs2201245	GACTGCATAGTGTTCAGG	CTCTGTTCTGTTCCATTG	G/A	F	TGGTACCAAAACAGATATGTAG
rs5907851	CAGGTTCAAGCGATTCTC	GAATCGCTTGAACCTGGG	G/A	F	CTCGCTGTGTTGCCCAG
rs5907853	ATTAGCTATCAGGGTGAG	CTCCCAAAGTGCTGGGAT	G/C	R	CCTGTAATCCCAGCTACT
ss78456818	ATTAGCTATCAGGGTGAG	GTTGCAGTGAGCCAAGAT	C/T	F	TTTGTTTGTTTGTTTGAGTCT
rs5907858	CTGAGCCAAGATAATTGAC	TGTCTACCCTTAATGCTC	A/T	F	CATCCAAATAATGCTATGAG
rs5907859	CTGAGCCAAGATAATTGAC	TGTCTACCCTTAATGCTC	C/T	F	CATCCAAATAATGCTATGAGA
rs1389194	CAATGTCACTTGTACAAA	TGAATGGTGCCTGACAC	C/T	F	CAGTCACCCAGGTATCTGT
rs861508	GAACATAGGACACACAACAG	TGGCTCCTAATAGTAGGC	C/T	F	ATTCCCTAGCCTAGACCTT
rs845163	ACATGGGTGCATTCACC	CCAGTTCATTTATCTCAGCA	G/A	F	ACCTCTGGTGGCTCCAC
rs845164	GAGGAGGCATGTCTTCAT	ACGTGACTCTCCTAATTC	C/T	F	CATGCTACTTCTCTTTTTAGGA
rs845165	CCTGATGGTAGTAAGGAGG	GTACTACCTTAGAAGCACTG	G/A	F	GCTGCATGACCTTGGAT
rs845190	CACAACAGCTCCAAATAAGT	GGCACGTTGTAGTAGTTA	G/T	F	AGTCTCAGATCCTTTAAGTAACTC
rs845188	AGAAGTTCCCACAGCTG	AGGTTAATCCATAACAGC	C/T	F	TCAAAGGCTTCCGTATTA
rs845187	CAGTGCTCTCTCATTTGG	AAATTATTAGCGAGCCG	G/C	F	ATGAAGAAACTGAAAATTAAATG
rs845186	AATCCTCTCTGGTAAGGG	GTGTTTGAGAGAGCTTTC	C/T	F	GTAAGGGAAACCAAATAACT
rs5907874	TTTCTTGGCTTTGGTAC	GCAAGAAAGCTGTTCAGT	C/A	F	CAATTTATTACAACAACAAGC
rs845182	GTACAGTTCCTTGTATTGTG	CAGTCTTCAGTAGTTCTGAG	G/C	F	ACAGCCCTAGAACCTACTTT
rs1493189	TCTACATGGGTCCTGATG	TGTTCAGATTTGGCAAG	G/A	F	GGATTCTGATGACATTTCTCT
rs710104	TTGTCATAGCCCATTTG	CAGAAATGTGCTTAACC	G/A	F	AAAATGCCATGTGCACA
rs5954267	GCCATAGACATGATGTTC	TCAACCATGTGATCAAG	C/T	R	TTGCAGCCTTTGCCGAA
rs5953588	CTGAGTATCCATCACCTGA	CTATTGTTGAAGGCACAG	G/C	F	CACTCAGTTCCTAGGTTAATAAG
rs911483	TGCAGAGCATGACTGTAC	GTGATGATGGCATAGTAA	A/T	R	TGATGGAAGAAAGAAAGAAG
rs5907876	GATGGGGTATTCAACTCTCA	CTGCCTCAAGATGTTAAAAC	G/A	F	ATTCATTCAGTGTCTTGTGAT
rs5954270	GAAGACCACCTTCCCAT	AATCAGCTCAATTGGGTG	G/C	F	ATACAGAAAGCCCTCTGTC
rs5907131	GCCTCTAGTTTCTCATTAGAGA	CCCAGTACACCATGCTT	C/T	F	GCTGACTAAGCCTGAGAAA

Table 10 (continued). Tagging SNP Assays

Marker	Forward Primer	Reverse Primer	Allele	Strand	Extension Primer
rs7060108	TGTTGGAAAGGTAACCTG	GTGGTAACCTTCAACCTG	G/A	F	AATTTTATTCGGGATAGTGTC
rs5907135	ATGTGAACAACCAAGGC	ACATTCAGTACATACCTG	G/A	F	TTTAAGAAATGCAATTCAAAT
rs6636266	CTGAAGCATCAGGTTCC	GAAGCTTCTGGGGTTTT	G/A	F	AACAAGAATGTTTAGTAGTAATGCT
rs5907890	CAATAGGCAATATTGGAGTC	AAACATACACTGATACTG	A/T	R	AAATGTAATTCCATCTTTTGTC
rs4825002	CTTAAACGCATATGCAC	CTGGTTGATAGCATTGTTCT	G/T	F	GAGGAAACAATTGCTAGAATT
rs5907891	CAGGAGATCAATGAAACCAA	GTGTTTGAATTTTGTCAA	C/T	F	GGAATGCACAACCAGATA
rs4824867	CATATGAGGTGCCATTCAC	GATCTCCTGACCTCATGAC	A/C	R	TTTGTGGTGGGCCAAAA
rs844971	GTGCTTTCACACATTATAGC	GTCTCTTGGGGAAGCAGA	G/A	R	ACCACTTATTGGATGTGTTT
rs5954277	CCCTTCTCAAGTGAAAAG	GAGAGAAGTGCAGTTGTTG	A/C	R	AAGTGCAGTTGTTGGTACTA
rs844964	GTAAGTTGATGGGCATGTT	CAGTGACCTGAGTTGCAC	C/G	F	GAAAGCAAAGTACAATTTACAAC
rs844963	GCCCAACATTAATTACTCTGTG	CACTCGTCTGGTGTCTAAG	A/G	F	TGGCAAGCCCAGTGGAT
rs844961	AAAAGGTATTCTCTACGCAC	CAAGTGGTAAACACGGC	A/G	F	AAAAAAGATTTTAATGATGTGTAG
rs844957	CTTAAAGGGACTGGGCATTA	CAGAACAAATCATGTGCAG	G/A	F	TCCCCACCAGAAAGTGG
rs844956	AATACATCCATGACCAGC	CATGTTGGTTACGTTGGTG	A/G	F	AGAGATGCTCCTTGCAA
rs844953	GGACATACAATGGAATGAAG	GGAAGTGGCTGATAGAG	A/T	F	GCCATATTGTGATTCATTTTA
rs6636273	GGACATACAATGGAATGAAG	GGAAGTGGCTGATAGAG	T/A	R	CAGAGGTAAAACCTATTTATCAT
rs844952	GAGGTATTAAAGAAAGCTGGTGTAG	GCCATTACACTGTCGTTC	A/G	F	CTCTCTGACGTTCTAAAATTAGA
rs844946	GTTGAGACTTGTGCCCAA	CCATAGTCACTGCTGAGG	G/C	F	CACCAGTAGCACCACATA
rs1493192	GGAGCCACATTTGATTTG	CACTTCAAGAGGCAAGTT	T/C	F	TATTTGCTGTCTGGTGTATATAC
rs926809	CTTTTGTACTGCAGCTG	CCTTGGGCTACATCATATTA	C/T	R	TGTCATCTACTTTTATCATTATTAAAC
rs5953592	CTGCAAGCAAGAACAAAC	ATGTTACCCCACATTGTG	G/C	R	TAAGGTTTGCAGATACCAA
rs5954285	GCAGACTCTCTCAAGCAG	TCTGCTCTGATCTTTATCTC	C/T	F	AAAGTATCCAGTCAAAGGAA
rs6636281	CAACCTACCAAGACTGAATT	GGTAATGCTGACCTTGAC	A/G	F	AAGTAAAGGTATTAAATCAGTAATAAGA
rs2864951	GGAAGAAGTGAAATTGTC	GGATAGTTTATTGTTAAGC	A/G	F	AAATACAATAATCCTGCAGAAT
rs5907905	CTGCCCAAATTTGTTCA	CTGTAAGGTTCTTGGAAAG	A/G	F	CACCCCTCTAACACCTAG
rs5954291	CCCATAAACATAGTTGCAG	TGTATCACACCCTTGATG	A/G	F	AAAATCAATTAATGCAATCC

Table 10 (continued). Tagging SNP Assays

Marker	Forward Primer	Reverse Primer	Allele	Strand	Extension Primer
rs5954292	AGACTGAGAGGTGTGAGTTC	CTAGACTATCGTGGGAATTA	G/C	R	GACTCTGGAATCCTATGTAATTA
rs4824870	AATATGTTGAAAGGGGC	CCTCACAGTTATAGAGATTG	T/C	R	TGTCCTCACATGGTGGA
rs5953596	GTCACCAAAATGCTATG	CATGATTTAAAAGGCAAC	G/A	R	GGAGTTTTATAGTCTCTAAGAGAGC
rs5907914	GTGTATGCCTATATTACGAC	CACTGAACATTGAATACCTG	A/C	F	TGTCACATAAGTGAGTGGAA
rs6528850	TTATGAAGTGGCAATGC	CTTCCCTTTCCCTGAACA	A/G	F	TGACTGTTCATGAAGGTCA
rs5907150	GCTGTAAACTACAACATGGT	CTTGGTTCTGATTTGATC	A/G	F	ATGGTATACAAATAGGAATAATTG
rs5907916	CTACCTGTATCTTTGCACACT	CAAATATGACTGTGTGGTG	A/G	F	ATAAGAAAACTGAGGATCAGAG
rs5907918	GTTTGTGAAGTTCAACAGTC	GGCTGTGAAAACATGTAAAC	T/A	R	AAACAAAGATCTCAAGTTGATTA
rs6528854	CACTTCAAAGTAGCCATT	GTACTGTTGAATCCCTCTA	A/G	F	AAAAGAAGCTATGAGAAAAAAA
rs5953599	AGTCAAACTCATGGAAGC	CTGTCTTCTTCACAAATT	A/C	F	GGCTGGGGAGAGAGGTA
rs5954305	GGAACATCATTCAGCTA	CCCATTAAGAAGTTAGTCTC	T/C	R	CTTTCTGTGTCTTTGGACTTA
rs5907921	CAAGTTCAAATCTCATCTCC	CTGTCCTTACAGACACG	T/C	R	TACAAGAAGATTGTATCATGAAA
rs5953606	AACAGAATACACAAGCACAC	GTTAAAACACTGCCTATGC	G/A	R	ATAACATCAGTGCCCTTAAG
rs6636314	ATTTAAAGGGCCAGTCC	CTTGGGAAAAGTTGGTAC	A/G	F	GAAAGGGAATCTAGAAAATACA
rs6636316	TTTACAAGACACTTCTGCAC	AGGGACTAACTCCTTACTC	A/G	F	CGATGATTGATGATTAATTG
rs5953608	ACAGGATCCAAAGAAAG	CCATAAAGTTTTGTAGTTC	A/C	F	AGAAGAAAATGCAATGAAA
rs5907944	GTTGCGCAATATTGTGA	GGTCTTGGACAGTGAGAC	G/A	R	TTTTAACCAACTTAAAATGTGA
rs5954322	GATCATTCCATGTGAGGC	TTGGATAGGGAATTACG	T/A	R	GAAGCCTTCAGATTTTTTT
rs5907945	GAAGAGATGGGAATACACAC	GCAGTCAACTCACTTTCTAGTGATA	A/G	F	GCTTCCTGCTAAAAATGTT
rs6636335	GCAACAAATGACAGTGGT	TCCACTAACTCCTACACAATAC	A/G	F	TTCTTTATGAAATAAAAGAATTTGT
rs6528868	GAGCAGAAGCCAGATTTA	GAAATACTGTAGTTCCCGC	C/A	R	TTCCTCAGTCAAGATTCTAGTT
rs6528869	GGCTGCTATAACAAAAGC	GGGATTAGCAGCCTTATA	G/A	R	GACTTCCCAGTCTCCAG
rs11095879	CTAGGTTGGTGAAGTTG	GCCAGGCATACCATTTT	A/T	F	ACTCTTTTCTTATCTTTGATTTTC
rs4825014	CTGACTTGTCAAGTATGACTG	TGCCAATCAGCTCTATC	T/C	R	GGGAAATTTCTTTAGTGTCTAA

flanking non-unique regions and amplified using the Expand Long Template PCR System (Roche Diagnostics, Indianapolis, IN). Long-range PCR product was then diluted 1:5,000,000 to dilute carry-over genomic template to non-amplifiable levels while retaining the ability to amplify from the long-range PCR product. This was verified by successful test of amplimers nested within the template long-range PCR product, but failure of amplimers elsewhere in the genome. High pairwise LD between SNPs within a given non-unique region and flanking unique SNPs supported correct non-unique copy assay (visible in Figure 8). All nested assays within long-range amplimers yielded one allele per subject, concordant with a unique X-chromosomal region for a male. Amplimer primers for long range PCR are seen in Table 11.

Table 11. Long Range Amplimers

Region	Forward Primer	Reverse Primer
hRPL44	AAGCACAACATGGATAGG	CAAGTTGAGGATCATCATG
SPANXC	CACTCTCTAGGGTCTAC	GGAGCATTAACTCACTCCTTA
RBMX2P1	GCTGAGAATTAGCATTGTC	GGGTATATCCACAGCCTAAG

Tag SNP Determination

Tagging SNP determination was conducted in a subset of 141 training set control subjects that were genotyped at 246 SNPs (including 194 validated from dbSNP and 52 identified by *de novo* discovery efforts). Pairwise LD was visualized using Haploview v4[127]. LDSelect was used for tagging SNP selection, specifying a MAF threshold of 0.05 and an r^2 threshold of 0.9[128]. A total of 128 tagging SNPs were selected for assay in the remaining subjects of the training set (totaling 292 independent familial case probands and their 292 age-matched controls). Data was obtained for 96.2% of the 74,752 tagging

genotypes sought in the training subjects, with a per marker range from 88.4% to 100%. SNPs in this tagging set and their assay primers are listed in Table 10.

Statistical Analyses

A sliding window approach tested a haplotype window of *N* markers, sliding the window along the map in single marker increments[167;169;218]. Each *N*-marker haplotype was compared to the remaining haplotypes as a group among the training group of 292 cases and 292 matched controls. The resulting 2 x 2 contingency table was evaluated by the χ^2 test statistic. Haplotype windows of 1-10 markers were evaluated in the exploratory analyses of training subjects. In a given map region that was nominally associated with PrCa within the training subjects ($P \leq 0.05$), haplotype tagging SNPs (htSNPs) were selected that most efficiently distinguished the risk haplotype from others in the region. Nominally significant tagged haplotypes (two observed) were genotyped in a subsequently ascertained independent test group of 215 cases and their 215 matched controls to address multiple comparisons. Significance for a given tagged haplotype candidate was adjusted for the two comparisons among test subjects through permutation testing. We generated 5,000 copies of the data set in which case status was permuted. A χ^2 value for each tagged haplotype was calculated for each simulated data set, as it was for the real data. Since the null hypothesis is true for each randomized subject set, the proportion of simulated χ^2 values greater than the real χ^2 value was used as a *P* value for the association, adjusted for multiple comparisons. Unless specifically noted, *P* values are unadjusted for multiple comparisons.

A risk haplotype that was significant after adjustment for multiple comparisons among test subjects was subsequently modeled by conditional logistic regression to obtain an estimate of effect size (Intercooled Stata 9, Stata Corporation, College Station, TX). The matching variable, age at diagnosis or screening, was included as a raw covariate in the model. Permutation testing was employed to assign significance.

Results

Our allele discovery and characterization was done within a screening set of 40 familial PrCa probands. We evaluated 415 SNPs annotated in dbSNP, 194 of which were polymorphic in these subjects. We also undertook *de novo* SNP discovery efforts across *LDOC1* and *SPANXC*, as well as across an *RPL44* homolog and *RBMX2P1* pseudogene. Among these, only *LDOC1* resides within a region of unique genomic sequence. We devised a nested amplification system to allow assay of non-unique genomic sequence flanked by unique sequence. Collectively, we discovered 52 common SNPs amenable to assay. The 246 polymorphic SNPs (194 from dbSNP, 52 newly discovered) of the screening subjects were genotyped in a subset of the training study population (141 cases and 141 controls) for assessment of allele frequency and for tagging SNP selection based upon LD patterns. Within this data, 220 SNPs had a minor allele frequency ≥ 0.05 for inclusion in subsequent analyses. Pairwise LD across the candidate interval for these SNPs highlights four LD blocks (denoted A, B, C, and D in Figure 8). Block A contains all four candidate gene regions.

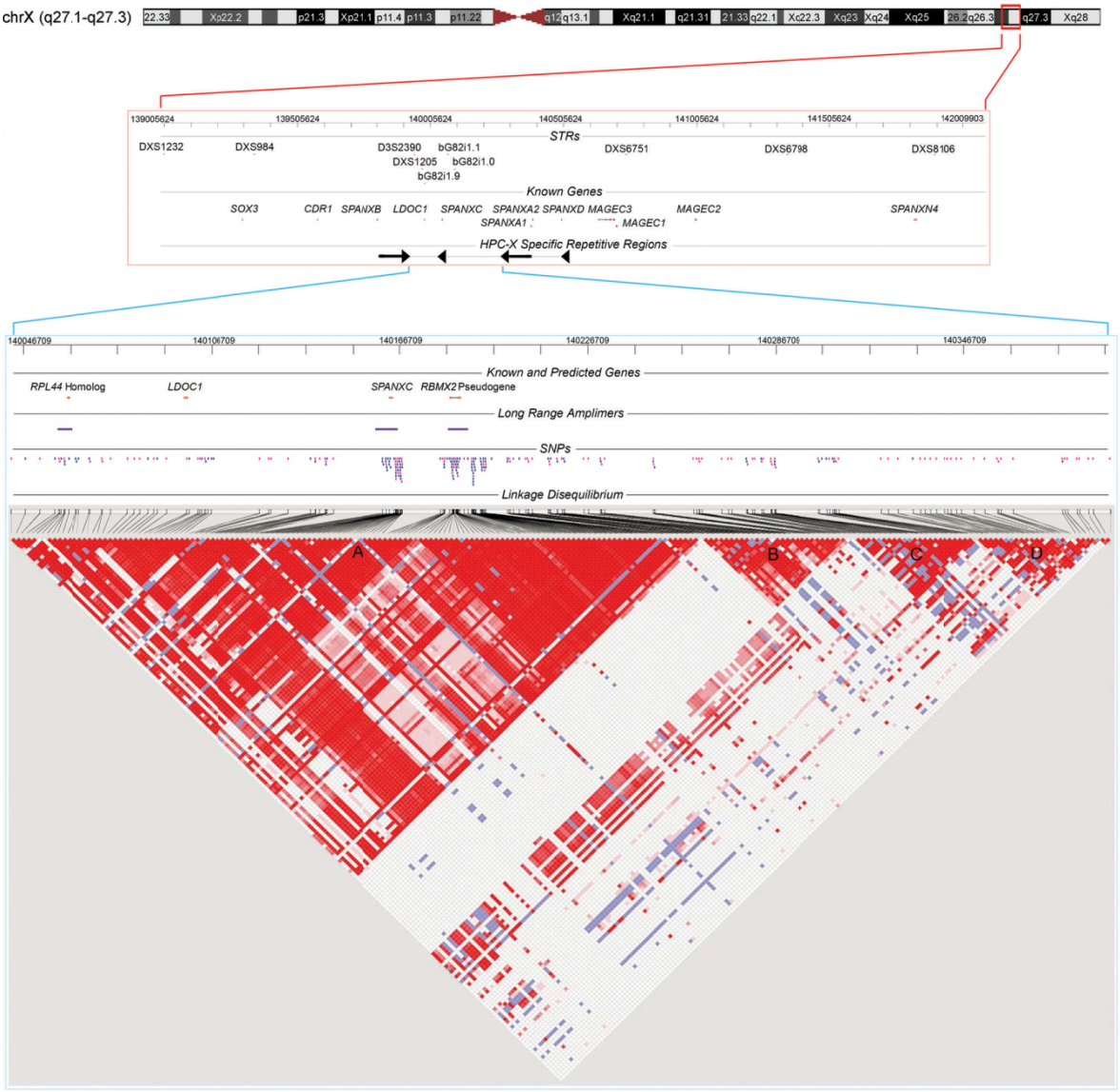
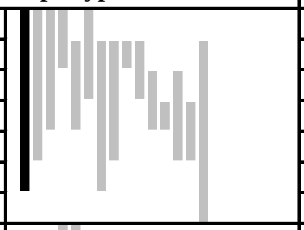
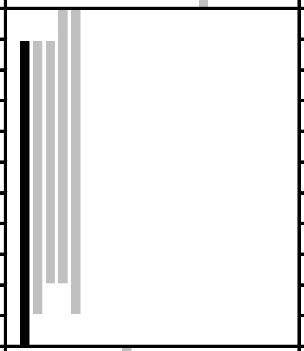
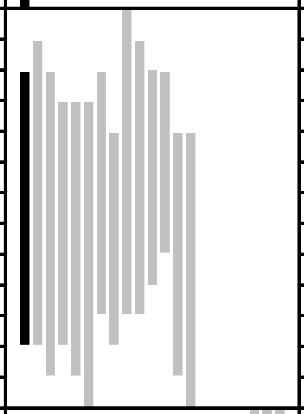
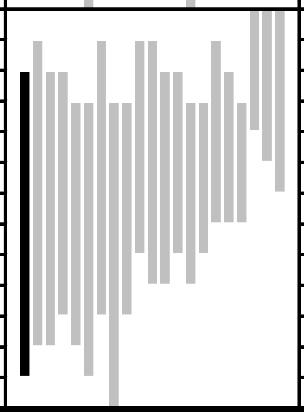


Figure 8. HPCX Candidate Interval. 3 MB of Xq27.1-27.3 depicting previously genotyped STR markers, annotated genes, and a complex HPCX specific repeat is shown at top (red bounding box, NCBI build 36.1 ChrX: 139005624-142009903 bp). The candidate interval for study is zoomed at bottom (blue bounding box, ChrX:140046709-140391709 bp). The interval contains *SPANXC* and *LDOC1*, as well as a predicted *RPL44* homolog and a pseudogene of *RBMX2*. As members of larger gene families, unique long range amplimers (denoted) were required to ensure site-specific assays. Polymorphic SNPs (N=246) are positioned on the map (tagging SNPs in pink). At bottom is a pairwise LD matrix for 141 controls across the subset of 220 SNPs with MAF ≥0.05. Red, $D' = 1$ (lod ≥ 2); blue, $D' = 1$ (lod < 2); pink, $D' < 1$ (lod ≥ 2); white, $D' < 1$ (lod < 2). Blocks of LD are denoted A, B, C and D.

Among the 220 informative SNPs, we selected 128 tagging SNPs for genotyping in the full group of training subjects (292 familial case probands and 292 age-matched controls). We explored evidence of association with PrCa using a haplotype-based sliding window approach. This entailed evaluation of 1,235 haplotype windows across the candidate interval. All haplotype windows of statistical significance were from four distinct regions. At each of the four regions, multiple overlapping windows were consistent with the redundant identification of one haplotype associated with PrCa risk. These four candidate risk haplotypes are numbered 1 to 4 in Table 12.

Only a subset of SNPs in each of the four regions was required to distinguish the candidate risk haplotype from remaining haplotypes. We identified haplotype-tagging SNPs (htSNPs) efficiently capturing the four candidate risk haplotypes (full span of windows $P \leq 0.05$) of Table 12. As our analysis required complete data for each subject across the multiple SNPs of the haplotype, the restricted set of htSNPs provided a better estimate of haplotype frequency. Only two of the four haplotypes were nominally significant when assessed by htSNPs among training subjects (Table 13, haplotype 1 ($\chi^2 = 5.24$, $P = 0.023$) and haplotype 3 ($\chi^2 = 5.08$, $P = 0.020$)). We evaluated evidence of association between these two htSNP haplotypes and PrCa in a second, independent study group of 215 familial PrCa probands and 215 age-matched controls. These subjects were accrued after the training subjects over the course of the ongoing study. Numerous exploratory tests were conducted among training subjects, but only two tests were conducted among test subjects, a greatly restricted number of comparisons. Only haplotype 3 was significant among test subjects (Table 13, $\chi^2 = 3.73$, $P = 0.040$).

Table 12. Sliding Window Risk Haplotypes at Xq27 - Training Subjects

Location	Significant Haplotype Windows*		Allele	Cases	Controls	*P* †
Haplotype 1 (*hRPL44*)	rs11095852		A	92 (39.8)	63 (27.3)	0.003
	rs5907823		C			
	rs7880499		G			
	rs1016824		T			
	rs12156848		G			
	rs7885649		A			
	rs5953563		A			
Haplotype 2 (*RBMX2P1*)	rs714076		G	15 (8.6)	5 (2.9)	0.021
	rs845150		A			
	rs5907844		G			
	rs881223		A			
	rs881221		T			
	rs881222		C			
	rs881219		A			
	rs2864937		A			
	rs5907848		A			
	rs2201245		G			
	rs5907851		A			
Haplotype 3 (ChrX: 140190766-140213636)	rs5907859		T	15 (6.9)	3 (1.4)	0.003
	rs1389194		T			
	rs861508		C			
	rs845163		A			
	rs845164		C			
	rs845165		A			
	rs845190		T			
	rs845188		C			
	rs845187		C			
	rs845186		A			
	rs5907874		C			
	rs845182		C			
	rs1493189		A			
Haplotype 4 (ChrX: 140266943-140295222)	rs844971		T	21 (8.9)	9 (3.8)	0.024
	rs5954277		T			
	rs844964		G			
	rs844963		A			
	rs844961		G			
	rs844957		A			
	rs844956		C			
	rs844953		T			
	rs6636273		A			
	rs844952		G			
	rs844946		C			
	rs1493192		A			
	rs926809		C			

* Sliding haplotype windows of $P \leq 0.05$, graphically ordered as most (black) to least significant (left to right).

† *P* for haplotype designated in black, with corresponding numbers of cases and controls, and haplotype frequencies (%).

Table 13. Tagged Risk Haplotypes at Xq27 -Training and Test Subjects

Location	htSNP	Allele	Training			Test		
			Case	Control	*P*	Case	Control	*P*
Haplotype 1 (*hRPL44*)	rs5907823	C	95 (38.6)	71 (28.9)	0.023	73 (39.3)	79 (2.5)	0.536
	rs7880499	G						
	rs1016824	T						
	rs12156848	G						
	rs7885649	A						
Haplotype 2 *(RBMX2P1)*	rs845150	A	19 (6.6)	9 (3.1)	0.062			
Haplotype 3 (ChrX: 140190766-140213636)	rs861508	C	18 (6.8)	7 (2.6)	0.020	13 (6.6)	5 (2.5)	0.040
	rs845165	A						
	rs845190	T						
	rs845187	C						
	rs845186	A						
	rs1493189	A						
Haplotype 4 (ChrX: 140266943-140295222)	rs844963	A	22 (7.7)	14 (4.9)	0.168			
	rs844956	C						

Shown are the corresponding numbers of cases and controls, and tagged haplotype frequencies (%).

Permutation testing was used to correct this value for the two comparisons conducted in test subjects, yielding $P = 0.048$. Our study identifies haplotype 3 as the most likely genetic variant of the interval to be associated with familial PrCa, with a nominal significance of $P = 0.003$ in the combined training and test subjects.

Under logistic regression modeling to assess effect size, haplotype 3 was associated with PrCa with an odds ratio of 3.41 (95% CI 1.04-11.17, $P = 0.034$) among test subjects, and an odds ratio of 2.52 (95% CI 1.25 – 5.10, $P = 0.006$) among combined training and test subjects. The effect size was more marked among the subset of 284 cases with aggressive PrCa (Gleason score ≥ 7), with an odds ratio of 4.06 (95% CI 1.15 – 14.31, $P = 0.021$). Gleason score is among the most important criteria in defining clinically significant disease. Our results are consistent with linkage data at the locus under stratification for clinically significant disease[104].

Discussion

The location of this haplotype coincides with that described through prior high-density simple tandem repeat (STR) mapping within a Finnish study population[92;217]. Among the STRs, bG82i1.1 was most significantly associated with PrCa in the prior study. The peak associated haplotype in the Finnish study was comprised of alleles at bG82i1.1 (centromeric) and bG82i1.0 (telomeric), $P = 0.0014$. Haplotype 3 of our study directly overlays the recombination hotspot between LD blocks A and B of Figure 8. The most centromeric SNP of the associated haplotype (rs5907859) is 4.0 kb from bG82i1.1. The most telomeric SNP of the associated haplotype (rs1493189) is 1.9 kb from

bG82i1.0. The same genomic region is highlighted by our present study of Americans of Northern European descent and the prior study of Finns. We further note that among SNPs evaluated in the genome wide association study of PrCa recently published by Thomas et al., rs845189 has a Whole Genome Rank of 1135 out of 527,869 SNPs accessed with a significance of $P = 0.002$[219]. This SNP resides at the LD break centered within the disease-associated haplotype of our study.

The associated haplotype region does not harbor known genes. All missense variants of potential interest in the entire candidate interval of 352 kb were within *SPANXC*, 30 kb from the associated haplotype. These missense variants clustered into two groups. The first group (all in exon 1) included D17E, A21V, and M24T. The second group (all in exon 2) included P29S, T30S, D32Y, and M42L. Within a group, a male subject had either each first or each second allele as listed. Additional *SPANXC* missense variants, E23K, V59F and L68V, did not appear to be in these two LD groups. This allele structure in *SPANXC* is also evident in data of an independent study[220]. That study also found no evidence to support an association between *SPANXC* alleles and risk of PrCa. The coding regions of *hRPL44* and *LDOC1* were without missense variants. We denote *RBMX2P1* as a pseudogene, having a mutated initiator methionine, multiple frameshift mutations, and an internal *Alu* insertion. Thus, the missense variants at *SPANXC* were among the best potential candidates for association with PrCa at Xq27.

The haplotype significantly associated with PrCa in this study straddles an LD break, potentially detecting a pair of contributing components located within each of the

two bounding LD blocks (*e.g.* a gene and a long-range regulatory element). In a sliding window haplotype analysis, a haplotype overlapping the two blocks would be particularly suited to detect such a combination. A single allele analytic approach failed to detect it. We considered the possibility that causal variants are a pair of non-contiguous SNPs within each LD block. We divided haplotype 3 so that those SNPs in LD block A comprised sub-haplotype 3A, and those in block B comprised sub-haplotype 3B. Eighteen of the SNPs within block A and only one SNP within block B (rs5907874) had an $r^2 \geq 0.8$ with the respective sub-haplotypes (Table 14). A matrix depiction of pairwise r^2 values between these is illustrated in Figure 9. The T30S (ss78456788) variant of *SPANXC* exon 2 also demonstrated modest LD with sub-haplotype 3A ($r^2 = 0.73$). Another variant altering an open reading frame within *RBMX2P1* (rs1968987) directly marked sub-haplotype 3A ($r^2 = 1$).

Only a subset of the SNPs demonstrating LD with the two sub-haplotypes had been genotyped as tagging SNPs in the training study population to enable an assessment of disease association. These included rs1012777, ss78456788 (T30S), rs12394263, ss78456800, rs5953578, rs845144, rs714076, rs881223, and ss78456818 in LD block A, and rs5907874 in LD block B. χ^2 tests of association for these block A SNP--block B SNP pair haplotypes were each associated with PrCa in our training group (*P* range 0.0008 to 0.030), with one exception (rs12394263 - rs5907874, $P = 0.065$). Results for each comparison are presented in Table 14. Therefore, these SNP pairs and the original sliding window haplotype spanning the LD break each detect the association with PrCa with varying efficiencies.

Table 14. SNPs with $r^2 > 0.8$ With Sub-Haplotype 3A

SNP*	r^2 with Sub-Haplotype 3A	*P*-value (Virtual haplotype with rs5907874[#])
rs1012777	0.95	0.003
ss78456783¶	0.81	0.030
rs34173722	0.85	‡
rs12394263	0.90	0.065
ss78456800	0.90	0.007
rs5953578	0.82	0.009
rs845144	0.90	0.006
rs714076 rs714074 rs714073	0.82	0.007
rs881223 rs982033 rs1389194 rs1968987 rs2864928†	0.90	0.0008
ss78456806	1.00	§
ss78456818 rs5907849	0.84	0.002

SNPs in **bold** used to calculate *P*-value within the training dataset

* $r^2 = 1.0$ among multiple SNPs listed in a row

rs5907874 has an $r^2 = 0.800$ with Haplotype 3B

¶ was not typed as a tagging SNP in our study, but was captured with an $r^2 = 0.974$ with **ss78456785**

† rs2864928 has an $r^2 = 1.0$ in HapMap CEU data, and was not included in our study

‡ was not typed as a tagging SNP in our study, but was captured with an $r^2 = 0.945$ with rs12394263

§ was not typed as a tagging SNP in our study, but was captured with an $r^2 = 0.946$ with rs845144

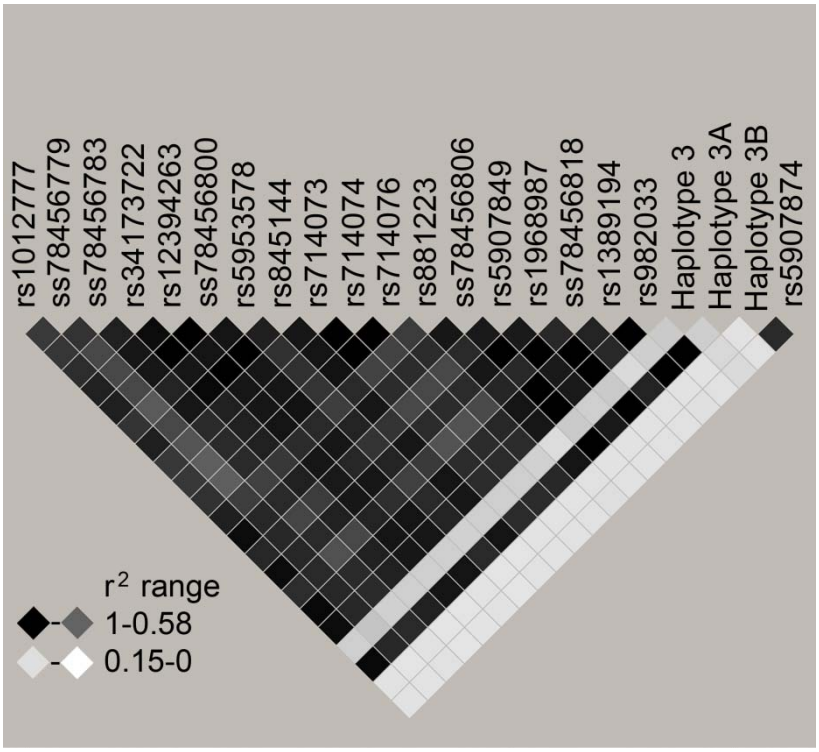


Figure 9. SNPs marking PrCa associated haplotype 3, sub-haplotypes A and B (portions in LD blocks A and B, respectively). A matrix is depicted of SNPs ordered centromeric to telomeric for those with an $r^2 > 0.8$ with the sub-haplotypes in 141 training set controls. A haplotype was dichotomized (present or absent) for pairwise r^2 calculation. Haplotype 3A contains htSNPs rs861508 and rs845165. Haplotype 3B contains htSNPs rs845190, rs845187, rs845186 and rs1493189.

Our study sought to identify the genetic variant predisposing to familial PrCa at Xq27, a locus initially identified by linkage study of American, Swedish, and Finnish hereditary PrCa pedigrees, and subsequently refined by linkage disequilibrium analysis of the Finnish familial PrCa cases. After a comprehensive effort in the present study, we identified a single candidate haplotype that was associated with familial PrCa within independent training and test study subjects. Although the replication was encouraging, the sample size of our test group was sufficiently small that an independent assessment of significance is warranted. Population structure is unlikely to represent a confounding factor within our study, as self-described ethnicity has recently been shown to accurately represent genetic ancestry among Americans of Northern European descent[132;133]. We believe that this haplotype represents the best candidate within the region for further investigation within additional study populations. If confirmed, these findings should begin to clarify the X-linked heritable component of PrCa risk.

Acknowledgements

We extend particular thanks to the study participants and to Drs Joseph Smith, Michael Cookson, Sam Chang, Richard Hock, William Maynard, and Jason Pereira. This work was supported by an award from the V Foundation, by a US Presidential Early Career Award for Scientists and Engineers, by a MERIT grant from the US Department of Veterans Affairs.

CHAPTER VI

CONCLUSIONS AND FUTURE DIRECTIONS

We began with a systematic analysis of the *CYP11A1* gene and its association with risk of breast cancer. The goals of the study were to comprehensively characterize common genetic variation at *CYP11A1*, to assess patterns of linkage disequilibrium (LD) and to refine our understanding of the contribution of *CYP11A1* genetic variation to breast cancer risk. In relation to the body of work as presented in this thesis, the goals of the study were to design and test the techniques and methods to be used in our comprehensive evaluation of an HPCX candidate interval. This preliminary project identified a haplotype containing promoter variants significantly linked to breast cancer risk. Building off previous work identifying a specific STR allele conferring risk of breast cancer, we described a specific haplotype which results in an increase in *CYP11A1* expression in lymphocytes. Future studies may be able to show if this expression increase extends into steroidogenic tissues where the rate-limiting step of steroid biosynthesis catalyzed by *CYP11A1* occurs. Using the odds ratio as an estimate for the relative risk, and the frequency of the haplotype in controls as an estimate of that of the general population, it is estimated that the population attributable risk associated with this haplotype is 6.9%. Our laboratory is currently investigating the entire steroid biosynthesis pathway and its role in breast cancer to continue this work.

We deemed the *CYP11A1* study a successful test of the methods to be used in the PrCa project. Over the time of the *CYP11A1* project, we ascertained a unique study population powered to identify common variants of PrCa susceptibility. We employed this study population in a focused search for causal variants at two well published linkage peaks; aggressiveness locus 19q12-13 and HPCX. First we used the candidate gene approach to uncover an association between the functional T29C polymorphism of *TGFB1* located within the peak at 19q12-13. We observed that this association was between the genotype and occurrence of *indolent* disease. This association specific to cases with low Gleason scores (≤ 6) could explain the confusion surrounding reporting of 19q12-13 as a susceptibility locus. Studies using Gleason score as a quantitative trait see 19q12-13 as significantly associated with risk of PrCa and have described it as a PrCa aggressiveness locus. However, linkage studies looking at only aggressive cases do not see significance at this locus. While it could be assumed that the aggressive cases were driving the association, including Gleason score as a quantitative trait does not indicate the direction of the association. Although the effect size observed in our study population was not exceptionally large, to my knowledge it is the first report of an association specific to indolent PrCa. Treatment for PrCa cases of indolent or aggressive nature are radically different, ranging from a 'wait-and-see' approach to prostatectomy. As such, knowledge of patient genotypes such as that of *TGFB1* T29C would be particularly salient when choosing a treatment path. To continue this work, our laboratory is currently investigating the role of a polymorphism of similar functional significance in TGF-β receptor *TGFBR1*. *TGFBR1* contains a repeat which encodes for a 9-alanine repeat (*TGFBR1**9A). A common variant which encodes a version of the protein containing a 6-

alanine repeat (*TGFBR1**6A) has reduced signaling functionality when compared to *TGFBR1**9A[221-223]. Furthermore, it has been shown that the interaction of the two significantly impacts risk of breast cancer[207]. We will investigate the role of this polymorphism and that of a potential interaction with *TGFB1* T29C in PrCa.

Finally, we report a comprehensive investigation of a 352 kb candidate locus within HPCX. Previous studies of HPCX have identified two separate, non-overlapping peaks of interest; at Xq27.1-2 and Xq28. We comprehensively evaluated an interval at Xq27.2 for association with risk of PrCa, and internally confirmed our results using discrete training and test study populations. Our potential risk haplotype maps to the same location as identified in our prior study of a Finnish population. This independent study employed a low-density STR marker strategy rather than the high-density SNP analysis presented in this thesis. Since different markers were used in the two studies, it is yet unknown if both identify the same haplotype. In our population, this haplotype spans two LD blocks and is seen in 6.7% of cases and 2.6% of controls. Due to the existence of a second, non-overlapping peak at Xq28 it is possible that there are multiple genomic variations at HPCX that confer PrCa risk. Therefore, we are currently extending our HPCX investigation to encompass the entire 14 Mb region from Xq27.1-Xq28.

Following up on linkage analysis data, we have identified potential predisposing variants at 19q13 and HPCX. It is somewhat likely that other linkage peaks harbor associated variants similar to the ones described in this body of work; however, the confusion surrounding most linkage results makes choosing one peak over another for

directed analysis problematic. As a result, it appears that most current PrCa genetic research is abandoning pursuance of linkage results in favor of genome-wide association. As with linkage studies, multiple loci have been identified via genome-wide association; however unlike linkage results, these associations have been replicated across multiple study populations.

An overall conclusion I derive from my thesis work is that the genetic component of PrCa risk is genetically heterogeneous, comprised of multiple common variants with a moderate effect size and supportive of the CDCV hypothesis. Furthermore, specific variants may drive phenotypic differences within PrCa cases. Genetic heterogeneity is clearly evident from the results presented in this body of work in combination with recent published results describing three unlinked PrCa risk loci at 8q24, as well as loci at 17q12 and 17q24.3[139;140;224]. Our laboratory and others have also seen statistically significant associations for overall risk at Xp11.22 and 2p15[225]. The role of these variants in the etiology of PrCa to date is unknown. The risk of any one of the variants described is moderate, with no reported odds ratio > 2. However, a recent study has shown that when more than one of the five published risk variants at 8q24 and 17q are present, risk of PrCa increases in a manner proportional to the number of high risk loci (up to an OR = 9.46 for individuals with all five risk loci and a family history of disease)[226]. There was no difference in risk for indolent or aggressive cases for the cumulative effect of these loci, and as such the authors surmise these loci may play an important role in the early etiology of PrCa. While it appears that these five variants together do not discriminate between indolent and aggressive disease, it is apparent from other studies that phenotypic

differences in PrCa cases may result from genetic variation. Evidence of this is seen not only in the *TGFB1* T29C work in this thesis, but other genome-wide association studies focusing on aggressive cases[227]. It is these associations specific to indolent or aggressive disease that will be of greatest utility to clinicians determining how to correctly treat an individual with PrCa.

In our PrCa study population, we restricted cases to those with a family history of disease, and through screening, selected controls without a family history of disease. We reasoned that we could reduce confounding due to genetic heterogeneity by the comparison of PrCa cases with a high likelihood of genetic risk factors to controls with no known genetic risk factors. This is unique to other published PrCa study populations. While unused in our matched study population, we have ascertained 156 Caucasian individuals with sporadic PrCa; a PrCa case with no family history of disease. To examine a potential difference between sporadic PrCa cases and those with a family history of disease at a locus of known significance, we again turn to 8q24 SNP rs1447295. We can compare genotype frequencies among cases with a family history of disease, sporadic cases and controls; the latter two groups both having no family history of disease (Table 15). It is notable that genotype frequencies of the sporadic cases at rs1447295 are nearly identical to those of the control population. Furthermore, the difference of allele frequencies between sporadic cases and cases with a family history of disease is statistically significant ($P = 0.01$). Previously published results also see no statistical difference with regards to rs1447295 allele frequency between sporadic cases and controls but a significant statistical difference comparing cases with a family history

of disease and controls[146]. Similar comparisons for both *TGFB1* T29C and HPCX haplotype 3 do not provide similarly striking results; at both loci the allele frequencies of the sporadic cases fall between those of the cases with a family history of disease and the controls. Loci on 1q25 (located within linkage peak HPC1) and 7p21 have been identified as risk variants in a large population consisting entirely of sporadic cases[65]. It is unknown if these two loci replicate in populations without sporadic cases, but authors point to HPC1 as evidence that 1q25 is associated with inherited forms of PrCa as well. Although including sporadic PrCa cases in a study population would increase sample size, based on our data at rs1447295, including such cases in the study population could potentially confound results by diluting the risk signal from familial and hereditary cases. If this trend is pervasive to other loci confirmed over multiple study populations, it could indicate that greater power to find associated variants is seen through analysis of familial and hereditary cases.

Table 15. 8q24 Genotype Distribution

	8q24 (rs1447295)					
	C/C		**C/A**		**A/A**	
Controls*	334	(83.3)	65	(16.2)	2	(0.5)
Sporadic Cases	130	(83.0)	25	(16.0)	1	(1.0)
Cases with Family History	296	(74.4)	95	(23.9)	7	(1.8)

* *n* (%)

Technological advances within the past several years have dramatically changed the way genetic association studies are designed and conducted. Truly, the execution and scope of my thesis work would be vastly different if it were proposed in 2008 rather than 2003. As a practical example of what is now possible due to technological advances, over

a period of almost six years using our florescence polarization based SNP assay system our lab generated close to 600,000 unique genotypes for various projects including the *CYP11A1*, *TGFB1* T29C and HPCX projects outlined in this body of work. In contrast, our high throughput Illumina GoldenGate® system, produced 976,303 unique genotypes in one month. While it is noted that the florescence polarization system was not running constantly through those six years, this is still representative of the huge amount of data that can be produced in a short time with these new genotyping technologies. This level of throughput is not without sacrifice; analysis of non-unique regions, such as those at HPCX, requires extensive assay customization not available to a proprietary system. However, as costs decrease, sample sizes and overall genomic coverage will increase allowing ever more reliable high-throughput identification of variants of all effect sizes.

The design and execution of genetic association studies have also been changed by recent scientific advances. During the time of my graduate studies the International HapMap was proposed, and Phases I and II completed. The HapMap was announced in 2003 with the goal of "determining the common patterns of DNA sequence variation in the human genome and to make this information freely available in the public domain"[228]. Phase I of the HapMap was released in 2005 with genotypes of 269 samples over four populations covering at least one common SNP (MAF $\geq$ 5%) every five kb (over 1 million total SNPs) across the genome[173]. Phase II was released in 2007 and contains a total of 3.1 million SNPs[229]. Researchers can now turn to this resource when selecting tagging SNPs for an initial screen in an association study, no longer having to identify a set of tagging SNPs in their individual population. The effect of the HapMap

has been so profound it has been suggested that it become the collection of markers used by most researchers as a reference and that its population data provide a framework for the genetic structure of a sample population[230].

The research presented in this body of work has been directed at identification of genetic variants predisposing to PrCa. We undertook a focused analysis of loci identified and independently replicated by PrCa linkage studies. Resultantly, we have identified two specific variants as candidates for risk of PrCa. As with all genetic association studies, the ultimate validity of these associations will no doubt be judged by the ability of other investigators with independent study populations to replicate our results. Nevertheless, this study represents part of a worldwide effort to uncover elusive variants predisposing to PrCa risk.

REFERENCES

1. Ries LAG et al., Available from http://seer.cancer.gov/csr/1975_2004/.

2. Stanford, J. L, Stephenson, R. A., Coyle, L. M., Cerhan, J., Correa, R., Eley, J. W., Gilliland, F., Hankey, B., Kolonel, L. N., Kosary, C., Ross, R., Severson, R., and West, D. Prostate Cancer Trends 1973-1995. 99-4543. 1999. Bethesda, MD, NIH Pub. 1999.
Ref Type: Report

3. M. A. Rubin and A. M. De Marzo, "Molecular Genetics of Human Prostate Cancer," *Mod.Pathol.* 17, no. 3 (2004): 380-388.

4. A. S. Whittemore et al., "Family History and Prostate Cancer Risk in Black, White, and Asian Men in the United States and Canada," *Am.J.Epidemiol.* 141, no. 8 (1995): 732-740.

5. B. S. Carter et al., "Hereditary Prostate Cancer: Epidemiologic and Clinical Features," *J.Urol.* 150, no. 3 (1993): 797-802.

6. W. F. Page et al., "Heredity and Prostate Cancer: a Study of World War II Veteran Twins," *Prostate* 33, no. 4 (1997): 240-245.

7. A. Ahlbom et al., "Cancer in Twins: Genetic and Nongenetic Familial Risk Factors," *JNCI Journal of the National Cancer Institute* 89, no. 4 (1997): 287-293.

8. H. T. Lynch, *Cancer Genetics*, ed. Lynch HT. (Springfield: Charles C Thomas, 1976).

9. Broca P.P., *Traité des Tumeurs*, vol. 1 (Paris: P. Asselin, 1866).

10. "Classics in Oncology. Heredity With Reference to Carcinoma As Shown by the Study of the Cases Examined in the Pathological Laboratory of the University of Michigan, 1895-1913. By Aldred Scott Warthin. 1913," *CA Cancer J.Clin.* 35, no. 6 (1985): 348-359.

11. Wolff J., *Die Lehre von der Krebskrankheit von den äldsten Zeiten bis zur Gegenwart* (Jena: Gustav Fleischer, 1907).

12. H. T. Lynch et al., "Hereditary Factors in Cancer. Study of Two Large Midwestern Kindreds," *Arch.Intern.Med.* 117, no. 2 (1966): 206-212.

13. F. P. Li and J. F. Fraumeni, Jr., "Soft-Tissue Sarcomas, Breast Cancer, and Other Neoplasms. A Familial Syndrome?," *Ann.Intern.Med.* 71, no. 4 (1969): 747-752.

14. A. G. Knudson, Jr., "Mutation and Cancer: Statistical Study of Retinoblastoma," *Proc.Natl.Acad.Sci.U.S.A* 68, no. 4 (1971): 820-823.

15. H. King, E. Diamond, and A. M. Lilienfeld, "Some Epidemiological Aspects of Cancer of the Prostate," *J.Chronic.Dis.* 16 (1963): 117-153.

16. G. Morganti et al., "[Clinico-Statistical and Genetic Research on Neoplasms of the Prostate.]," *Acta Genet.Stat.Med.* 6, no. 2 (1956): 304-305.

17. C. M. Woolf, "An Investigation of the Familial Aspects of Carcinoma of the Prostate," *Cancer* 13 (1960): 739-744.

18. Ross R.K. and Schottenfeld D., "Prostate Cancer," in *Cancer Epidemiology and Prevention*, ed. Schottenfeld D. and Fraumeni J.F. 2nd ed. (New York: Oxford University Press, 1996), 1180-1206.

19. G. D. Steinberg et al., "Family History and the Risk of Prostate Cancer," *Prostate* 17, no. 4 (1990): 337-347.

20. K. R. Monroe et al., "Evidence of an X-Linked or Recessive Genetic Component to Prostate Cancer Risk," *Nat.Med.* 1, no. 8 (1995): 827-829.

21. S. A. Narod et al., "The Impact of Family History on Early Detection of Prostate Cancer," *Nat.Med.* 1, no. 2 (1995): 99-101.

22. O. Bratt, "Hereditary Prostate Cancer: Clinical Aspects," *J.Urol.* 168, no. 3 (2002): 906-913.

23. O. Bratt, "Hereditary Prostate Cancer," *BJU.Int.* 85, no. 5 (2000): 588-598.

24. N. Risch, "The Genetic Epidemiology of Cancer: Interpreting Family and Twin Studies and Their Implications for Molecular Genetic Approaches," *Cancer Epidemiol.Biomarkers Prev.* 10, no. 7 (2001): 733-741.

25. H. Gronberg, L. Damber, and J. E. Damber, "Studies of Genetic Factors in Prostate Cancer in a Twin Population," *J.Urol.* 152, no. 5 Pt 1 (1994): 1484-1487.

26. P. Lichtenstein et al., "Environmental and Heritable Factors in the Causation of Cancer--Analyses of Cohorts of Twins From Sweden, Denmark, and Finland," *N.Engl.J.Med.* 343, no. 2 (2000): 78-85.

27. B. S. Carter et al., "Mendelian Inheritance of Familial Prostate Cancer," *Proc.Natl.Acad.Sci.U.S.A* 89, no. 8 (1992): 3367-3371.

28. H. Gronberg et al., "Segregation Analysis of Prostate Cancer in Sweden: Support for Dominant Inheritance," *Am.J.Epidemiol.* 146, no. 7 (1997): 552-557.

29. D. J. Schaid et al., "Evidence for Autosomal Dominant Inheritance of Prostate Cancer," *Am.J.Hum.Genet.* 62, no. 6 (1998): 1425-1438.

30. B. A. Verhage et al., "Autosomal Dominant Inheritance of Prostate Cancer: a Confirmatory Study," *Urology* 57, no. 1 (2001): 97-101.

31. J. Cui et al., "Segregation Analyses of 1,476 Population-Based Australian Families Affected by Prostate Cancer," *Am.J.Hum.Genet.* 68, no. 5 (2001): 1207-1218.

32. G. Gong et al., "Segregation Analysis of Prostate Cancer in 1,719 White, African-American and Asian-American Families in the United States and Canada," *Cancer Causes Control* 13, no. 5 (2002): 471-482.

33. A. Valeri et al., "Segregation Analysis of Prostate Cancer in France: Evidence for Autosomal Dominant Inheritance and Residual Brother-Brother Dependence," *Ann.Hum.Genet.* 67, no. Pt 2 (2003): 125-137.

34. E. M. Conlon et al., "Oligogenic Segregation Analysis of Hereditary Prostate Cancer Pedigrees: Evidence for Multiple Loci Affecting Age at Onset," *Int.J.Cancer* 105, no. 5 (2003): 630-635.

35. A. B. Baffoe-Bonnie et al., "Genome-Wide Linkage of 77 Families From the African American Hereditary Prostate Cancer Study (AAHPC)," *Prostate* 67, no. 1 (2007): 22-31.

36. N. J. Camp, J. M. Farnham, and L. A. Cannon Albright, "Genomic Search for Prostate Cancer Predisposition Loci in Utah Pedigrees," *Prostate* 65, no. 4 (2005): 365-374.

37. J. M. Cunningham et al., "Genome Linkage Screen for Prostate Cancer Susceptibility Loci: Results From the Mayo Clinic Familial Prostate Cancer Study," *Prostate* 57, no. 4 (2003): 335-346.

38. S. Edwards et al., "Results of a Genome-Wide Linkage Analysis in Prostate Cancer Families Ascertained Through the ACTANE Consortium," *Prostate* 57, no. 4 (2003): 270-279.

39. D. M. Friedrichsen et al., "Identification of a Prostate Cancer Susceptibility Locus on Chromosome 7q11-21 in Jewish Families," *Proc.Natl.Acad.Sci.U.S.A* 101, no. 7 (2004): 1939-1944.

40. K. A. Goddard et al., "Model-Free Linkage Analysis With Covariates Confirms Linkage of Prostate Cancer to Chromosomes 1 and 4," *Am.J.Hum.Genet.* 68, no. 5 (2001): 1197-1206.

41. C. L. Hsieh et al., "A Genome Screen of Families With Multiple Cases of Prostate Cancer: Evidence of Genetic Heterogeneity," *Am.J.Hum.Genet.* 69, no. 1 (2001): 148-158.

42. M. Janer et al., "Genomic Scan of 254 Hereditary Prostate Cancer Families," *Prostate* 57, no. 4 (2003): 309-319.

43. E. M. Lange 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* 57, no. 4 (2003): 326-334.

44. J. Schleutker et al., "Genome-Wide Scan for Linkage in Finnish Hereditary Prostate Cancer (HPC) Families Identifies Novel Susceptibility Loci at 11q14 and 3p25-26," *Prostate* 57, no. 4 (2003): 280-289.

45. B. K. Suarez et al., "A Genome Screen of Multiplex Sibships With Prostate Cancer," *Am.J.Hum.Genet.* 66, no. 3 (2000): 933-944.

46. F. Wiklund et al., "Genome-Wide Scan of Swedish Families With Hereditary Prostate Cancer: Suggestive Evidence of Linkage at 5q11.2 and 19p13.3," *Prostate* 57, no. 4 (2003): 290-297.

47. J. S. Witte et al., "Genome-Wide Scan of Brothers: Replication and Fine Mapping of Prostate Cancer Susceptibility and Aggressiveness Loci," *Prostate* 57, no. 4 (2003): 298-308.

48. J. Xu et al., "Genome-Wide Scan for Prostate Cancer Susceptibility Genes in the Johns Hopkins Hereditary Prostate Cancer Families," *Prostate* 57, no. 4 (2003): 320-325.

49. D. J. Schaid, "The Complex Genetic Epidemiology of Prostate Cancer," *Hum.Mol.Genet.* 13 Spec No 1 (2004): R103-R121.

50. J. R. Smith et al., "Major Susceptibility Locus for Prostate Cancer on Chromosome 1 Suggested by a Genome-Wide Search," *Science* 274, no. 5291 (1996): 1371-1374.

51. P. Berthon et al., "Predisposing Gene for Early-Onset Prostate Cancer, Localized on Chromosome 1q42.2-43," *Am.J.Hum.Genet.* 62, no. 6 (1998): 1416-1424.

52. M. Gibbs et al., "Evidence for a Rare Prostate Cancer-Susceptibility Locus at Chromosome 1p36," *Am.J.Hum.Genet.* 64, no. 3 (1999): 776-787.

53. J. Xu et al., "Evidence for a Prostate Cancer Susceptibility Locus on the X Chromosome," *Nat.Genet.* 20, no. 2 (1998): 175-179.

54. J. Xu et al., "Linkage and Association Studies of Prostate Cancer Susceptibility: Evidence for Linkage at 8p22-23," *Am.J.Hum.Genet.* 69, no. 2 (2001): 341-350.

55. S. V. Tavtigian et al., "A Candidate Prostate Cancer Susceptibility Gene at Chromosome 17p," *Nat.Genet.* 27, no. 2 (2001): 172-180.

56. R. Berry et al., "Evidence for a Prostate Cancer-Susceptibility Locus on Chromosome 20," *Am.J.Hum.Genet.* 67, no. 1 (2000): 82-91.

57. J. S. Witte et al., "Genomewide Scan for Prostate Cancer-Aggressiveness Loci," *Am.J.Hum.Genet.* 67, no. 1 (2000): 92-99.

58. H. Gronberg et al., "Characteristics of Prostate Cancer in Families Potentially Linked to the Hereditary Prostate Cancer 1 (HPC1) Locus," *JAMA* 278, no. 15 (1997): 1251-1255.

59. K. A. Cooney et al., "Prostate Cancer Susceptibility Locus on Chromosome 1q: a Confirmatory Study," *J.Natl.Cancer Inst.* 89, no. 13 (1997): 955-959.

60. C. L. Hsieh et al., "Re: Prostate Cancer Susceptibility Locus on Chromosome 1q: a Confirmatory Study," *J.Natl.Cancer Inst.* 89, no. 24 (1997): 1893-1894.

61. S. L. Neuhausen et al., "Prostate Cancer Susceptibility Locus HPC1 in Utah High-Risk Pedigrees," *Hum.Mol.Genet.* 8, no. 13 (1999): 2437-2442.

62. J. Xu, "Combined Analysis of Hereditary Prostate Cancer Linkage to 1q24-25: Results From 772 Hereditary Prostate Cancer Families From the International Consortium for Prostate Cancer Genetics," *Am.J.Hum.Genet.* 66, no. 3 (2000): 945-957.

63. E. L. Goode et al., "Linkage Analysis of 150 High-Risk Prostate Cancer Families at 1q24-25," *Genet.Epidemiol.* 18, no. 3 (2000): 251-275.

64. J. Xu et al., "Linkage of Prostate Cancer Susceptibility Loci to Chromosome 1," *Hum.Genet.* 108, no. 4 (2001): 335-345.

65. R. K. Nam et al., "A Genome-Wide Association Screen Identifies Regions on Chromosomes 1q25 and 7p21 As Risk Loci for Sporadic Prostate Cancer," *Prostate Cancer Prostatic.Dis.* (2007).

66. R. A. McIndoe et al., "Linkage Analysis of 49 High-Risk Families Does Not Support a Common Familial Prostate Cancer-Susceptibility Gene at 1q24-25," *Am.J.Hum.Genet.* 61, no. 2 (1997): 347-353.

67. P. Berthon et al., "Predisposing Gene for Early-Onset Prostate Cancer, Localized on Chromosome 1q42.2-43," *Am.J.Hum.Genet.* 62, no. 6 (1998): 1416-1424.

68. R. A. Eeles et al., "Linkage Analysis of Chromosome 1q Markers in 136 Prostate Cancer Families. The Cancer Research Campaign/British Prostate Group U.K. Familial Prostate Cancer Study Collaborators," *Am.J.Hum.Genet.* 62, no. 3 (1998): 653-658.

69. R. Berry et al., "Linkage Analyses at the Chromosome 1 Loci 1q24-25 (HPC1), 1q42.2-43 (PCAP), and 1p36 (CAPB) in Families With Hereditary Prostate Cancer," *Am.J.Hum.Genet.* 66, no. 2 (2000): 539-546.

70. B. K. Suarez et al., "Replication Linkage Study for Prostate Cancer Susceptibility Genes," *Prostate* 45, no. 2 (2000): 106-114.

71. J. Carpten et al., "Germline Mutations in the Ribonuclease L Gene in Families Showing Linkage With HPC1," *Nat.Genet.* 30, no. 2 (2002): 181-184.

72. J. Sun et al., "Genetic Variability in Inflammation Pathways and Prostate Cancer Risk," *Urol.Oncol.* 25, no. 3 (2007): 250-259.

73. A. Rokman et al., "Germline Alterations of the RNASEL Gene, a Candidate HPC1 Gene at 1q25, in Patients and Families With Prostate Cancer," *Am.J.Hum.Genet.* 70, no. 5 (2002): 1299-1304.

74. L. Wang et al., "Analysis of the RNASEL Gene in Familial and Sporadic Prostate Cancer," *Am.J.Hum.Genet.* 71, no. 1 (2002): 116-123.

75. H. Nakazato et al., "Role of Genetic Polymorphisms of the RNASEL Gene on Familial Prostate Cancer Risk in a Japanese Population," *Br.J.Cancer* 89, no. 4 (2003): 691-696.

76. G. Casey et al., "RNASEL Arg462Gln Variant Is Implicated in Up to 13% of Prostate Cancer Cases," *Nat.Genet.* 32, no. 4 (2002): 581-583.

77. S. J. Shook et al., "Association of RNASEL Variants With Prostate Cancer Risk in Hispanic Caucasians and African Americans," *Clin.Cancer Res.* 13, no. 19 (2007): 5959-5964.

78. F. Wiklund et al., "Genetic Analysis of the RNASEL Gene in Hereditary, Familial, and Sporadic Prostate Cancer," *Clin.Cancer Res.* 10, no. 21 (2004): 7150-7156.

79. C. Maier et al., "Mutation Screening and Association Study of RNASEL As a Prostate Cancer Susceptibility Gene," *Br.J.Cancer* 92, no. 6 (2005): 1159-1164.

80. H. Rennert et al., "Association of Susceptibility Alleles in ELAC2/HPC2, RNASEL/HPC1, and MSR1 With Prostate Cancer Severity in European American and African American Men," *Cancer Epidemiol.Biomarkers Prev.* 14, no. 4 (2005): 949-957.

81. C. Cybulski et al., "DNA Variation in MSR1, RNASEL and E-Cadherin Genes and Prostate Cancer in Poland," *Urol.Int.* 79, no. 1 (2007): 44-49.

82. S. E. Daugherty et al., "RNASEL Arg462Gln Polymorphism and Prostate Cancer in PLCO," *Prostate* 67, no. 8 (2007): 849-854.

83. G. Cancel-Tassin et al., "PCAP Is the Major Known Prostate Cancer Predisposing Locus in Families From South and West Europe," *Eur.J.Hum.Genet.* 9, no. 2 (2001): 135-142.

84. W. M. Brown et al., "Hereditary Prostate Cancer in African American Families: Linkage Analysis Using Markers That Map to Five Candidate Susceptibility Loci," *Br.J.Cancer* 90, no. 2 (2004): 510-514.

85. A. S. Whittemore et al., "No Evidence of Linkage for Chromosome 1q42.2-43 in Prostate Cancer," *Am.J.Hum.Genet.* 65, no. 1 (1999): 254-256.

86. J. Simard et al., "Prostate Cancer Susceptibility Genes: Lessons Learned and Challenges Posed," *Endocr.Relat Cancer* 10, no. 2 (2003): 225-259.

87. J. Schleutker et al., "A Genetic Epidemiological Study of Hereditary Prostate Cancer (HPC) in Finland: Frequent HPCX Linkage in Families With Late-Onset Disease," *Clin.Cancer Res.* 6, no. 12 (2000): 4810-4815.

88. E. M. Lange et al., "Linkage Analysis of 153 Prostate Cancer Families Over a 30-CM Region Containing the Putative Susceptibility Locus HPCX," *Clin.Cancer Res.* 5, no. 12 (1999): 4013-4020.

89. S. Bochum et al., "Confirmation of the Prostate Cancer Susceptibility Locus HPCX in a Set of 104 German Prostate Cancer Families," *Prostate* 52, no. 1 (2002): 12-19.

90. J. M. Farnham et al., "Confirmation of the HPCX Prostate Cancer Predisposition Locus in Large Utah Prostate Cancer Pedigrees," *Hum.Genet.* 116, no. 3 (2005): 179-185.

91. M. A. Peters et al., "Genetic Linkage Analysis of Prostate Cancer Families to Xq27-28," *Hum.Hered.* 51, no. 1-2 (2001): 107-113.

92. A. B. Baffoe-Bonnie et al., "A Major Locus for Hereditary Prostate Cancer in Finland: Localization by Linkage Disequilibrium of a Haplotype in the HPCX Region," *Hum.Genet.* 117, no. 4 (2005): 307-316.

93. M. Gibbs et al., "Evidence for a Rare Prostate Cancer-Susceptibility Locus at Chromosome 1p36," *Am.J.Hum.Genet.* 64, no. 3 (1999): 776-787.

94. M. A. Peters et al., "Germline Mutations in the P73 Gene Do Not Predispose to Familial Prostate-Brain Cancer," *Prostate* 48, no. 4 (2001): 292-296.

95. R. Berry et al., "Evidence for a Prostate Cancer-Susceptibility Locus on Chromosome 20," *Am.J.Hum.Genet.* 67, no. 1 (2000): 82-91.

96. S. L. Zheng et al., "Evidence for a Prostate Cancer Linkage to Chromosome 20 in 159 Hereditary Prostate Cancer Families," *Hum.Genet.* 108, no. 5 (2001): 430-435.

97. C. H. Bock et al., "Analysis of the Prostate Cancer-Susceptibility Locus HPC20 in 172 Families Affected by Prostate Cancer," *Am.J.Hum.Genet.* 68, no. 3 (2001): 795-801.

98. G. Cancel-Tassin et al., "No Evidence of Linkage to HPC20 on Chromosome 20q13 in Hereditary Prostate Cancer," *Int.J.Cancer* 93, no. 3 (2001): 455-456.

99. D. J. Schaid and B. L. Chang, "Description of the International Consortium For Prostate Cancer Genetics, and Failure to Replicate Linkage of Hereditary Prostate Cancer to 20q13," *Prostate* 63, no. 3 (2005): 276-290.

100. D. F. Gleason and G. T. Mellinger, "Prediction of Prognosis for Prostatic Adenocarcinoma by Combined Histological Grading and Clinical Staging," *J.Urol.* 111, no. 1 (1974): 58-64.

101. A. M. DeMarzo et al., "Pathological and Molecular Aspects of Prostate Cancer," *Lancet* 361, no. 9361 (2003): 955-964.

102. S. L. Slager et al., "Confirmation of Linkage of Prostate Cancer Aggressiveness With Chromosome 19q," *Am.J.Hum.Genet.* 72, no. 3 (2003): 759-762.

103. D. J. Schaid et al., "Genome-Wide Linkage Scan of Prostate Cancer Gleason Score and Confirmation of Chromosome 19q," *Hum.Genet.* (2007).

104. B. L. Chang et al., "Genome-Wide Screen for Prostate Cancer Susceptibility Genes in Men With Clinically Significant Disease," *Prostate* 64, no. 4 (2005): 356-361.

105. J. L. Stanford et al., "Prostate Cancer and Genetic Susceptibility: a Genome Scan Incorporating Disease Aggressiveness," *Prostate* 66, no. 3 (2006): 317-325.

106. S. V. Tavtigian et al., "A Candidate Prostate Cancer Susceptibility Gene at Chromosome 17p," *Nat.Genet.* 27, no. 2 (2001): 172-180.

107. T. R. Rebbeck et al., "Association of HPC2/ELAC2 Genotypes and Prostate Cancer," *Am.J.Hum.Genet.* 67, no. 4 (2000): 1014-1019.

108. J. Xu et al., "Evaluation of Linkage and Association of HPC2/ELAC2 in Patients With Familial or Sporadic Prostate Cancer," *Am.J.Hum.Genet.* 68, no. 4 (2001): 901-911.

109. A. Rokman et al., "ELAC2/HPC2 Involvement in Hereditary and Sporadic Prostate Cancer," *Cancer Res.* 61, no. 16 (2001): 6038-6041.

110. N. Platt and S. Gordon, "Is the Class A Macrophage Scavenger Receptor (SR-A) Multifunctional? - The Mouse's Tale," *J.Clin.Invest* 108, no. 5 (2001): 649-654.

111. J. Xu et al., "Germline Mutations and Sequence Variants of the Macrophage Scavenger Receptor 1 Gene Are Associated With Prostate Cancer Risk," *Nat.Genet.* 32, no. 2 (2002): 321-325.

112. F. Wiklund et al., "Linkage Analysis of Prostate Cancer Susceptibility: Confirmation of Linkage at 8p22-23," *Hum.Genet.* 112, no. 4 (2003): 414-418.

113. E. H. Seppala et al., "Germ-Line Alterations in MSR1 Gene and Prostate Cancer Risk," *Clin.Cancer Res.* 9, no. 14 (2003): 5252-5256.

114. Q. Hope et al., "Macrophage Scavenger Receptor 1 999C>T (R293X) Mutation and Risk of Prostate Cancer," *Cancer Epidemiol.Biomarkers Prev.* 14, no. 2 (2005): 397-402.

115. F. Lindmark et al., "Analysis of the Macrophage Scavenger Receptor 1 Gene in Swedish Hereditary and Sporadic Prostate Cancer," *Prostate* 59, no. 2 (2004): 132-140.

116. L. Wang et al., "No Association of Germline Alteration of MSR1 With Prostate Cancer Risk," *Nat.Genet.* 35, no. 2 (2003): 128-129.

117. J. Sun et al., "Meta-Analysis of Association of Rare Mutations and Common Sequence Variants in the MSR1 Gene and Prostate Cancer Risk," *Prostate* 66, no. 7 (2006): 728-737.

118. N. J. Risch, "Searching for Genetic Determinants in the New Millennium," *Nature* 405, no. 6788 (2000): 847-856.

119. E. A. Ostrander and J. L. Stanford, "Genetics of Prostate Cancer: Too Many Loci, Too Few Genes," *Am.J.Hum.Genet.* 67, no. 6 (2000): 1367-1375.

120. J. Altmuller et al., "Genomewide Scans of Complex Human Diseases: True Linkage Is Hard to Find," *Am.J.Hum.Genet.* 69, no. 5 (2001): 936-950.

121. D. Altshuler et al., "The Common PPARgamma Pro12Ala Polymorphism Is Associated With Decreased Risk of Type 2 Diabetes," *Nat.Genet.* 26, no. 1 (2000): 76-80.

122. J. K. Pritchard and N. J. Cox, "The Allelic Architecture of Human Disease Genes: Common Disease-Common Variant...or Not?," *Hum.Mol.Genet.* 11, no. 20 (2002): 2417-2423.

123. N. Risch and K. Merikangas, "The Future of Genetic Studies of Complex Human Diseases," *Science* 273, no. 5281 (1996): 1516-1517.

124. J. R. Gulcher, A. Kong, and K. Stefansson, "The Role of Linkage Studies for Common Diseases," *Curr.Opin.Genet.Dev.* 11, no. 3 (2001): 264-267.

125. M. J. Daly et al., "High-Resolution Haplotype Structure in the Human Genome," *Nat.Genet.* 29, no. 2 (2001): 229-232.

126. J. N. Hirschhorn and M. J. Daly, "Genome-Wide Association Studies for Common Diseases and Complex Traits," *Nat.Rev.Genet.* 6, no. 2 (2005): 95-108.

127. J. C. Barrett et al., "Haploview: Analysis and Visualization of LD and Haplotype Maps," *Bioinformatics.* 21, no. 2 (2005): 263-265.

128. C. S. Carlson et al., "Selecting a Maximally Informative Set of Single-Nucleotide Polymorphisms for Association Analyses Using Linkage Disequilibrium," *Am.J.Hum.Genet.* 74, no. 1 (2004): 106-120.

129. L. Kruglyak, "Prospects for Whole-Genome Linkage Disequilibrium Mapping of Common Disease Genes," *Nat.Genet.* 22, no. 2 (1999): 139-144.

130. L. R. Cardon and L. J. Palmer, "Population Stratification and Spurious Allelic Association," *Lancet* 361, no. 9357 (2003): 598-604.

131. J. K. Pritchard and N. A. Rosenberg, "Use of Unlinked Genetic Markers to Detect Population Stratification in Association Studies," *Am.J.Hum.Genet.* 65, no. 1 (1999): 220-228.

132. H. Tang et al., "Genetic Structure, Self-Identified Race/Ethnicity, and Confounding in Case-Control Association Studies," *Am.J Hum.Genet.* 76, no. 2 (2005): 268-275.

133. D. J. Hunter et al., "A Genome-Wide Association Study Identifies Alleles in FGFR2 Associated With Risk of Sporadic Postmenopausal Breast Cancer," *Nat.Genet.* 39, no. 7 (2007): 870-874.

134. K. G. Ardlie, K. L. Lunetta, and M. Seielstad, "Testing for Population Subdivision and Association in Four Case-Control Studies," *Am.J.Hum.Genet.* 71, no. 2 (2002): 304-311.

135. J. S. Pankow et al., "Regarding "Testing for Population Subdivision and Association in Four Case-Control Studies"," *Am.J.Hum.Genet.* 71, no. 6 (2002): 1478-1480.

136. M. L. Freedman et al., "Assessing the Impact of Population Stratification on Genetic Association Studies," *Nat.Genet.* 36, no. 4 (2004): 388-393.

137. J. D. Storey and R. Tibshirani, "Statistical Significance for Genomewide Studies," *Proc.Natl.Acad.Sci.U.S.A* 100, no. 16 (2003): 9440-9445.

138. W. D. Dupont and W. D. Plummer, Jr., "Power and Sample Size Calculations. A Review and Computer Program," *Control Clin.Trials* 11, no. 2 (1990): 116-128.

139. L. T. Amundadottir et al., "A Common Variant Associated With Prostate Cancer in European and African Populations," *Nat.Genet.* 38, no. 6 (2006): 652-658.

140. J. Gudmundsson et al., "Genome-Wide Association Study Identifies a Second Prostate Cancer Susceptibility Variant at 8q24," *Nat.Genet.* 39, no. 5 (2007): 631-637.

141. M. L. Freedman et al., "Admixture Mapping Identifies 8q24 As a Prostate Cancer Risk Locus in African-American Men," *Proc.Natl.Acad.Sci.U.S.A* 103, no. 38 (2006): 14068-14073.

142. M. Yeager et al., "Genome-Wide Association Study of Prostate Cancer Identifies a Second Risk Locus at 8q24," *Nat.Genet.* 39, no. 5 (2007): 645-649.

143. F. R. Schumacher et al., "A Common 8q24 Variant in Prostate and Breast Cancer From a Large Nested Case-Control Study," *Cancer Res.* 67, no. 7 (2007): 2951-2956.

144. G. Severi et al., "The Common Variant Rs1447295 on Chromosome 8q24 and Prostate Cancer Risk: Results From an Australian Population-Based Case-Control Study," *Cancer Epidemiol.Biomarkers Prev.* 16, no. 3 (2007): 610-612.

145. M. Suuriniemi et al., "Confirmation of a Positive Association Between Prostate Cancer Risk and a Locus at Chromosome 8q24," *Cancer Epidemiol.Biomarkers Prev.* 16, no. 4 (2007): 809-814.

146. L. Wang et al., "Two Common Chromosome 8q24 Variants Are Associated With Increased Risk for Prostate Cancer," *Cancer Res.* 67, no. 7 (2007): 2944-2950.

147. C. A. Haiman et al., "Multiple Regions Within 8q24 Independently Affect Risk for Prostate Cancer," *Nat.Genet.* 39, no. 5 (2007): 638-644.

148. S. A. Savage and M. H. Greene, "The Evidence for Prostate Cancer Risk Loci at 8q24 Grows Stronger," *J.Natl.Cancer Inst.* 99, no. 20 (2007): 1499-1501.

149. W. Zheng et al., "Population-Based Case-Control Study of CYP11A Gene Polymorphism and Breast Cancer Risk," *Cancer Epidemiol.Biomarkers Prev.* 13, no. 5 (2004): 709-714.

150. E. M. Smigielski et al., "DbSNP: a Database of Single Nucleotide Polymorphisms," *Nucleic Acids Res.* 28, no. 1 (2000): 352-355.

151. R. C. Travis and T. J. Key, "Oestrogen Exposure and Breast Cancer Risk," *Breast Cancer Res.* 5, no. 5 (2003): 239-247.

152. E. J. Folkerd et al., "The Relationship Between Factors Affecting Endogenous Oestradiol Levels in Postmenopausal Women and Breast Cancer," *J.Steroid Biochem.Mol.Biol.* 102, no. 1-5 (2006): 250-255.

153. K. Mitrunen and A. Hirvonen, "Molecular Epidemiology of Sporadic Breast Cancer. The Role of Polymorphic Genes Involved in Oestrogen Biosynthesis and Metabolism," *Mutat.Res.* 544, no. 1 (2003): 9-41.

154. E. Gasteiger et al., "ExPASy: The Proteomics Server for in-Depth Protein Knowledge and Analysis," *Nucleic Acids Res.* 31, no. 13 (2003): 3784-3788.

155. N. Gharani et al., "Association of the Steroid Synthesis Gene CYP11a With Polycystic Ovary Syndrome and Hyperandrogenism," *Hum.Mol.Genet.* 6, no. 3 (1997): 397-402.

156. Y. T. Gao et al., "Association of Menstrual and Reproductive Factors With Breast Cancer Risk: Results From the Shanghai Breast Cancer Study," *Int.J.Cancer* 87, no. 2 (2000): 295-300.

157. X. Chen, L. Levine, and P. Y. Kwok, "Fluorescence Polarization in Homogeneous Nucleic Acid Analysis," *Genome Res.* 9, no. 5 (1999): 492-498.

158. M. J. Brownstein, J. D. Carpten, and J. R. Smith, "Modulation of Non-Templated Nucleotide Addition by Taq DNA Polymerase: Primer Modifications That Facilitate Genotyping," *Biotechniques* 20, no. 6 (1996): 1004-1010.

159. A. von Deimling et al., "A Rapid and Non-Radioactive PCR Based Assay for the Detection of Allelic Loss in Human Gliomas," *Neuropathol.Appl.Neurobiol.* 19, no. 6 (1993): 524-529.

160. H. J. Cordell and D. G. Clayton, "Genetic Association Studies," *Lancet* 366, no. 9491 (2005): 1121-1131.

161. T. R. Gaunt et al., "MIDAS: Software for Analysis and Visualisation of Interallelic Disequilibrium Between Multiallelic Markers," *BMC.Bioinformatics.* 7 (2006): 227.

162. J. S. Stephens et al., "Effects of Electrospinning and Solution Casting Protocols on the Secondary Structure of a Genetically Engineered Dragline Spider Silk Analogue Investigated Via Fourier Transform Raman Spectroscopy," *Biomacromolecules.* 6, no. 3 (2005): 1405-1413.

163. J. Marchini et al., "A Comparison of Phasing Algorithms for Trios and Unrelated Individuals," *Am.J.Hum.Genet.* 78, no. 3 (2006): 437-450.

164. M. Stephens, N. J. Smith, and P. Donnelly, "A New Statistical Method for Haplotype Reconstruction From Population Data," *Am.J.Hum.Genet.* 68, no. 4 (2001): 978-989.

165. M. Stephens and P. Donnelly, "A Comparison of Bayesian Methods for Haplotype Reconstruction From Population Genotype Data," *Am.J.Hum.Genet.* 73, no. 5 (2003): 1162-1169.

166. L. Excoffier and M. Slatkin, "Maximum-Likelihood Estimation of Molecular Haplotype Frequencies in a Diploid Population," *Mol.Biol.Evol.* 12, no. 5 (1995): 921-927.

167. D. Fallin et al., "Genetic Analysis of Case/Control Data Using Estimated Haplotype Frequencies: Application to APOE Locus Variation and Alzheimer's Disease," *Genome Res.* 11, no. 1 (2001): 143-151.

168. D. Fallin and N. J. Schork, "Accuracy of Haplotype Frequency Estimation for Biallelic Loci, Via the Expectation-Maximization Algorithm for Unphased Diploid Genotype Data," *Am.J.Hum.Genet.* 67, no. 4 (2000): 947-959.

169. R. A. Mathias et al., "A Graphical Assessment of P-Values From Sliding Window Haplotype Tests of Association to Identify Asthma Susceptibility Loci on Chromosome 11q," *BMC.Genet.* 7 (2006): 38.

170. M. Kayser et al., "Characteristics and Frequency of Germline Mutations at Microsatellite Loci From the Human Y Chromosome, As Revealed by Direct Observation in Father/Son Pairs," *Am.J.Hum.Genet.* 66, no. 5 (2000): 1580-1588.

171. B. Brinkmann et al., "Mutation Rate in Human Microsatellites: Influence of the Structure and Length of the Tandem Repeat," *Am.J.Hum.Genet.* 62, no. 6 (1998): 1408-1415.

172. Z. Zhou et al., "Prominent Sex Steroid Metabolism in Human Lymphocytes," *Mol.Cell Endocrinol.* 138, no. 1-2 (1998): 61-69.

173. "A Haplotype Map of the Human Genome," *Nature* 437, no. 7063 (2005): 1299-1320.

174. V. W. Setiawan et al., "A Systematic Assessment of Common Genetic Variation in CYP11A and Risk of Breast Cancer," *Cancer Res.* 66, no. 24 (2006): 12019-12025.

175. I. C. Guo, M. C. Hu, and B. C. Chung, "Transcriptional Regulation of CYP11A1," *J.Biomed.Sci.* 10, no. 6 Pt 1 (2003): 593-598.

176. Z. Liu and E. R. Simpson, "Steroidogenic Factor 1 (SF-1) and SP1 Are Required for Regulation of Bovine CYP11A Gene Expression in Bovine Luteal Cells and Adrenal Y1 Cells," *Mol.Endocrinol.* 11, no. 2 (1997): 127-137.

177. Z. Liu and E. R. Simpson, "Molecular Mechanism for Cooperation Between Sp1 and Steroidogenic Factor-1 (SF-1) to Regulate Bovine CYP11A Gene Expression," *Mol.Cell Endocrinol.* 153, no. 1-2 (1999): 183-196.

178. P. Venepally and M. R. Waterman, "Two Sp1-Binding Sites Mediate CAMP-Induced Transcription of the Bovine CYP11A Gene Through the Protein Kinase A Signaling Pathway," *J.Biol.Chem.* 270, no. 43 (1995): 25402-25410.

179. R. Ahlgren et al., "Role of Sp1 in CAMP-Dependent Transcriptional Regulation of the Bovine CYP11A Gene," *J.Biol.Chem.* 274, no. 27 (1999): 19422-19428.

180. J. Doi et al., "Salt-Inducible Kinase Represses CAMP-Dependent Protein Kinase-Mediated Activation of Human Cholesterol Side Chain Cleavage Cytochrome P450 Promoter Through the CREB Basic Leucine Zipper Domain," *J.Biol.Chem.* 277, no. 18 (2002): 15629-15637.

181. S. J. Chou, K. N. Lai, and B. Chung, "Characterization of the Upstream Sequence of the Human CYP11A1 Gene for Cell Type-Specific Expression," *J.Biol.Chem.* 271, no. 36 (1996): 22125-22129.

182. M. C. Hu et al., "Regulation of Steroidogenesis in Transgenic Mice and Zebrafish," *Mol.Cell Endocrinol.* 171, no. 1-2 (2001): 9-14.

183. Y. Huang et al., "Action of Hormone Responsive Sequence in 2.3 Kb Promoter of CYP11A1," *Mol.Cell Endocrinol.* 175, no. 1-2 (2001): 205-210.

184. B. C. Chung, I. C. Guo, and S. J. Chou, "Transcriptional Regulation of the CYP11A1 and Ferredoxin Genes," *Steroids* 62, no. 1 (1997): 37-42.

185. I. C. Guo, H. M. Tsai, and B. C. Chung, "Actions of Two Different CAMP-Responsive Sequences and an Enhancer of the Human CYP11A1 (P450scc) Gene in Adrenal Y1 and Placental JEG-3 Cells," *J.Biol.Chem.* 269, no. 9 (1994): 6362-6369.

186. I. C. Guo and B. C. Chung, "Cell-Type Specificity of Human CYP11A1 TATA Box," *J.Steroid Biochem.Mol.Biol.* 69, no. 1-6 (1999): 329-334.

187. D. Monte, F. DeWitte, and D. W. Hum, "Regulation of the Human P450scc Gene by Steroidogenic Factor 1 Is Mediated by CBP/P300," *J.Biol.Chem.* 273, no. 8 (1998): 4585-4591.

188. K. N. Hogeveen, M. Talikka, and G. L. Hammond, "Human Sex Hormone-Binding Globulin Promoter Activity Is Influenced by a (TAAAA)n Repeat Element Within an Alu Sequence," *J.Biol.Chem.* 276, no. 39 (2001): 36383-36390.

189. C. A. Haiman et al., "Common Genetic Variation in the Sex Steroid Hormone-Binding Globulin (SHBG) Gene and Circulating Shbg Levels Among Postmenopausal Women: the Multiethnic Cohort," *J.Clin.Endocrinol.Metab* 90, no. 4 (2005): 2198-2204.

190. V. G. Cheung et al., "Mapping Determinants of Human Gene Expression by Regional and Genome-Wide Association," *Nature* 437, no. 7063 (2005): 1365-1369.

191. M. Morley et al., "Genetic Analysis of Genome-Wide Variation in Human Gene Expression," *Nature* 430, no. 7001 (2004): 743-747.

192. A. Zeleniuch-Jacquotte et al., "Circulating Enterolactone and Risk of Endometrial Cancer," *Int.J.Cancer* 119, no. 10 (2006): 2376-2381.

193. G. S. Huggins et al., "GATA5 Activation of the Progesterone Receptor Gene Promoter in Breast Cancer Cells Is Influenced by the +331G/A Polymorphism," *Cancer Res.* 66, no. 3 (2006): 1384-1390.

194. S. L. Slager et al., "Genome-Wide Linkage Scan for Prostate Cancer Aggressiveness Loci Using Families From the University of Michigan Prostate Cancer Genetics Project," *Prostate* 66, no. 2 (2006): 173-179.

195. B. Bierie and H. L. Moses, "Tumour Microenvironment: TGFbeta: the Molecular Jekyll and Hyde of Cancer," *Nat.Rev.Cancer* 6, no. 7 (2006): 506-520.

196. C. Lee et al., "Transforming Growth Factor-Beta in Benign and Malignant Prostate," *Prostate* 39, no. 4 (1999): 285-290.

197. P. M. Siegel and J. Massague, "Cytostatic and Apoptotic Actions of TGF-Beta in Homeostasis and Cancer," *Nat.Rev.Cancer* 3, no. 11 (2003): 807-821.

198. H. L. Adler et al., "Elevated Levels of Circulating Interleukin-6 and Transforming Growth Factor-Beta1 in Patients With Metastatic Prostatic Carcinoma," *J.Urol.* 161, no. 1 (1999): 182-187.

199. W. Cui et al., "TGFbeta1 Inhibits the Formation of Benign Skin Tumors, but Enhances Progression to Invasive Spindle Carcinomas in Transgenic Mice," *Cell* 86, no. 4 (1996): 531-542.

200. B. I. Dalal, P. A. Keown, and A. H. Greenberg, "Immunocytochemical Localization of Secreted Transforming Growth Factor-Beta 1 to the Advancing Edges of Primary Tumors and to Lymph Node Metastases of Human Mammary Carcinoma," *Am.J.Pathol.* 143, no. 2 (1993): 381-389.

201. S. M. Kakonen et al., "Transforming Growth Factor-Beta Stimulates Parathyroid Hormone-Related Protein and Osteolytic Metastases Via Smad and Mitogen-Activated Protein Kinase Signaling Pathways," *J.Biol.Chem.* 277, no. 27 (2002): 24571-24578.

202. R. S. Muraoka et al., "Blockade of TGF-Beta Inhibits Mammary Tumor Cell Viability, Migration, and Metastases," *J.Clin.Invest* 109, no. 12 (2002): 1551-1559.

203. D. R. Welch, A. Fabra, and M. Nakajima, "Transforming Growth Factor Beta Stimulates Mammary Adenocarcinoma Cell Invasion and Metastatic Potential," *Proc.Natl.Acad.Sci.U.S.A* 87, no. 19 (1990): 7678-7682.

204. J. J. Yin et al., "TGF-Beta Signaling Blockade Inhibits PTHrP Secretion by Breast Cancer Cells and Bone Metastases Development," *J.Clin Invest* 103, no. 2 (1999): 197-206.

205. A. M. Dunning et al., "A Transforming Growth Factorbeta1 Signal Peptide Variant Increases Secretion in Vitro and Is Associated With Increased Incidence of Invasive Breast Cancer," *Cancer Res.* 63, no. 10 (2003): 2610-2615.

206. E. Ziv et al., "Association Between the T29-->C Polymorphism in the Transforming Growth Factor Beta1 Gene and Breast Cancer Among Elderly White Women: The Study of Osteoporotic Fractures," *JAMA* 285, no. 22 (2001): 2859-2863.

207. V. G. Kaklamani et al., "Combined Genetic Assessment of Transforming Growth Factor-Beta Signaling Pathway Variants May Predict Breast Cancer Risk," *Cancer Res.* 65, no. 8 (2005): 3454-3461.

208. Breast Cancer Association Consortium, "Commonly Studied Single-Nucleotide Polymorphisms and Breast Cancer: Results From the Breast Cancer Association Consortium," *J.Natl.Cancer Inst.* 98, no. 19 (2006): 1382-1396.

209. A. M. Gonzalez-Zuloeta Ladd et al., "Transforming-Growth Factor Beta(1) Leu10Pro Polymorphism and Breast Cancer Morbidity," *Eur.J.Cancer* (2006).

210. A. Shin et al., "Genetic Polymorphisms of the Transforming Growth Factor-Beta1 Gene and Breast Cancer Risk: a Possible Dual Role at Different Cancer Stages," *Cancer Epidemiol.Biomarkers Prev.* 14, no. 6 (2005): 1567-1570.

211. X. O. Shu et al., "Genetic Polymorphisms in the TGF-Beta 1 Gene and Breast Cancer Survival: a Report From the Shanghai Breast Cancer Study," *Cancer Res.* 64, no. 3 (2004): 836-839.

212. P. J. Neville et al., "Prostate Cancer Aggressiveness Locus on Chromosome Segment 19q12-Q13.1 Identified by Linkage and Allelic Imbalance Studies," *Genes Chromosomes.Cancer* 36, no. 4 (2003): 332-339.

213. C. A. Haiman et al., "A Common Genetic Risk Factor for Colorectal and Prostate Cancer," *Nat.Genet.* 39, no. 8 (2007): 954-956.

214. K. J. Livak, "Allelic Discrimination Using Fluorogenic Probes and the 5' Nuclease Assay," *Genet.Anal.* 14, no. 5-6 (1999): 143-149.

215. T. C. Brand et al., "Association of Polymorphisms in TGFB1 and Prostate Cancer Prognosis," *J.Urol.* 179, no. 2 (2008): 754-758.

216. E. M. Gillanders et al., "Combined Genome-Wide Scan for Prostate Cancer Susceptibility Genes," *J.Natl.Cancer Inst.* 96, no. 16 (2004): 1240-1247.

217. A. B. Baffoe-Bonnie et al., "A Major Locus for Hereditary Prostate Cancer in Finland: Localization by Linkage Disequilibrium of a Haplotype in the HPCX Region (Erratum)," *Hum.Genet.* 118, no. 2 (2005): 307.

218. B. L. Yaspan et al., "Haplotype Analysis of CYP11A1 Identifies Promoter Variants Associated With Breast Cancer Risk," *Cancer Res.* 67, no. 12 (2007): 5673-5682.

219. G. Thomas et al., "Multiple Loci Identified in a Genome-Wide Association Study of Prostate Cancer," *Nat.Genet.* (2008).

220. N. Kouprina et al., "Mutational Analysis of SPANX Genes in Families With X-Linked Prostate Cancer," *Prostate* 67, no. 8 (2007): 820-828.

221. B. Pasche et al., "TbetaR-I(6A) Is a Candidate Tumor Susceptibility Allele," *Cancer Res.* 59, no. 22 (1999): 5678-5682.

222. B. Pasche et al., "Type I Transforming Growth Factor Beta Receptor Maps to 9q22 and Exhibits a Polymorphism and a Rare Variant Within a Polyalanine Tract," *Cancer Res.* 58, no. 13 (1998): 2727-2732.

223. T. Chen et al., "Structural Alterations of Transforming Growth Factor-Beta Receptor Genes in Human Cervical Carcinoma," *Int.J.Cancer* 82, no. 1 (1999): 43-51.

224. J. Gudmundsson et al., "Two Variants on Chromosome 17 Confer Prostate Cancer Risk, and the One in TCF2 Protects Against Type 2 Diabetes," *Nat.Genet.* 39, no. 8 (2007): 977-983.

225. J. Gudmundsson et al., "Common Sequence Variants on 2p15 and Xp11.22 Confer Susceptibility to Prostate Cancer," *Nat.Genet.* (2008).

226. S. Lilly Zheng et al., "Cumulative Association of Five Genetic Variants With Prostate Cancer," *The New England Journal of Medicine* (2008): NEJMoa075819.

227. D. Duggan et al., "Two Genome-Wide Association Studies of Aggressive Prostate Cancer Implicate Putative Prostate Tumor Suppressor Gene DAB2IP," *J.Natl.Cancer Inst.* 99, no. 24 (2007): 1836-1844.

228. "The International HapMap Project," *Nature* 426, no. 6968 (2003): 789-796.

229. K. A. Frazer et al., "A Second Generation Human Haplotype Map of Over 3.1 Million SNPs," *Nature* 449, no. 7164 (2007): 851-861.

230. "Framework for a Fully Powered Risk Engine," *Nat.Genet.* 37, no. 11 (2005): 1153.