

AWARD NUMBER: W81XWH-06-1-0057

TITLE: Identification and Characterization of an X-Linked Familial Prostate Cancer Gene

PRINCIPAL INVESTIGATOR: Brian Yaspan

CONTRACTING ORGANIZATION: Vanderbilt University Medical Center
Nashville, TN 37203

REPORT DATE: November 2008

TYPE OF REPORT: Annual Summary

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE 1 November 2008		2. REPORT TYPE Annual Summary		3. DATES COVERED 24 Oct 2005 – 23 Oct 2008	
4. TITLE AND SUBTITLE Identification and Characterization of an X-Linked Familial Prostate Cancer Gene				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER W81XWH-06-1-0057	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) Brian Yaspan E-Mail: brian.yaspan@vanderbilt.edu				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Vanderbilt University Medical Center Nashville, TN 37203				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Medical Research and Materiel Command Fort Detrick, Maryland 21702-5012				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION / AVAILABILITY STATEMENT Approved for Public Release; Distribution Unlimited					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT Prostate cancer is the most commonly diagnosed non-skin malignancy in males, with as many as one in 5 males living in developed nations being diagnosed with prostate cancer in their lifetime. Despite the medical significance of prostate cancer our understanding of predisposition and progression in the disease remains rudimentary. Prostate cancer is estimated to have the largest heritable component of all common cancers. We will explicitly characterize ancestral versions of a gene region originally implicated in prostate cancer through study of families with multiple cases of prostate cancer (HPC) to enable us to test the hypothesis that a common disease-predisposing genetic mutation conferring modest risk is shared among present-day prostate cancer cases in the broader population by virtue of inheritance from an ancient founder. We hope the findings of this proposal will offer a promising inroad for predicting disease predisposition, for tailoring the most effective current therapy to each individual patient, and for developing rational new therapies.					
15. SUBJECT TERMS No Subject Terms provided.					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER (include area code)
U	U	U	UU	15	

Table of Contents

Introduction.....	4
Body.....	4- 8
Key Research Accomplishments.....	9
Reportable Outcomes.....	9
Conclusions.....	9
References.....	n/a
Appendices.....	n/a

Introduction

There is a significant heritable component of prostate cancer. Increased familial relative risk is observed across multiple populations. Male first degree relatives of prostate cancer patients have a two- to three-fold increased risk. Segregation analyses support genetic rather than shared environmental risk. Twin cancer concordance studies reveal a higher heritable risk for prostate cancer than for any other common cancer. Additional epidemiological studies have been consistent with X-linked transmission, identifying higher risk for a man with an affected brother relative to one with an affected father. Despite the overwhelming genetic predisposition evidence, the identification of prostate cancer susceptibility genes has been difficult. Linkage studies have resulted in the identification of several loci difficult to confirm across study populations. However, summary studies of genome-wide scans for prostate cancer susceptibility loci in general confirm two loci, HPC-1 and HPC-X.

Our study seeks to identify a candidate gene or genes conferring prostate cancer susceptibility at locus HPC-X in a US Caucasian study population. We hypothesize that a gene or genes at HPC-X harbor common moderate-penetrance variants predisposing to prostate cancer. We looked at shared haplotypes in founder populations and found two intervals likely to harbor prostate cancer susceptibility genes. We have chosen to first focus on one interval at locus HPC-X (termed HPC-X region A) due to shared haplotype association evidence in the founder populations of Finland, Iceland and Ashkenazim.

Body

Accomplishments

Results from Tasks 1-3 have been published in the journal Human Genetics (Hum Genet. 2008 May;123(4):379-86). A summary by task appears below and the manuscript is attached. Please note that Tables referenced are from the attached manuscript.

Task 1. To identify and genotype all common polymorphism in our study population at potential genes of the candidate interval (HPC-X Region A). (Months 1-12):

- a. Perform *de novo* SNP discovery at predicted or known genes and derive a set of survey SNPs spanning the HPC-X locus and a density of 3-5 kb from dbSNP
- b. Genotype a subset of the study population for all SNPs in 1a
- c. Analyze genotypes to determine genetic architecture of HPC-X

Task 1a-1c has been completed.

To derive a set of survey SNPs spanning the interval at HPC-X we assayed SNPs found in dbSNP for polymorphism in a subset of 40 prostate cancer cases from the training dataset. Out of 415 SNPs culled from database entries, we identified 194 as polymorphic and assayable in our study population. To augment this set of SNPs, we undertook *de novo* SNP discovery in the same subset of 40 prostate cancer cases at

known and predicted genes in the region, identified in **Figure 1**. In addition to known genes *SPANXC* and *LDOC1*, custom software identified a coding region containing homology to *RPL44* and a pseudogene containing homology to *RBMX2*. *De novo* SNP discovery at these four features resulted in 52 additional SNPs for a total of 246 SNPs. Linkage disequilibrium (LD) patterns at HPC-X Region A for these 246 SNPs typed in a subset of 141 controls of our training population are presented in **Figure 1**. Four major blocks of LD are apparent. Block “A” contains all four known and predicted genes.

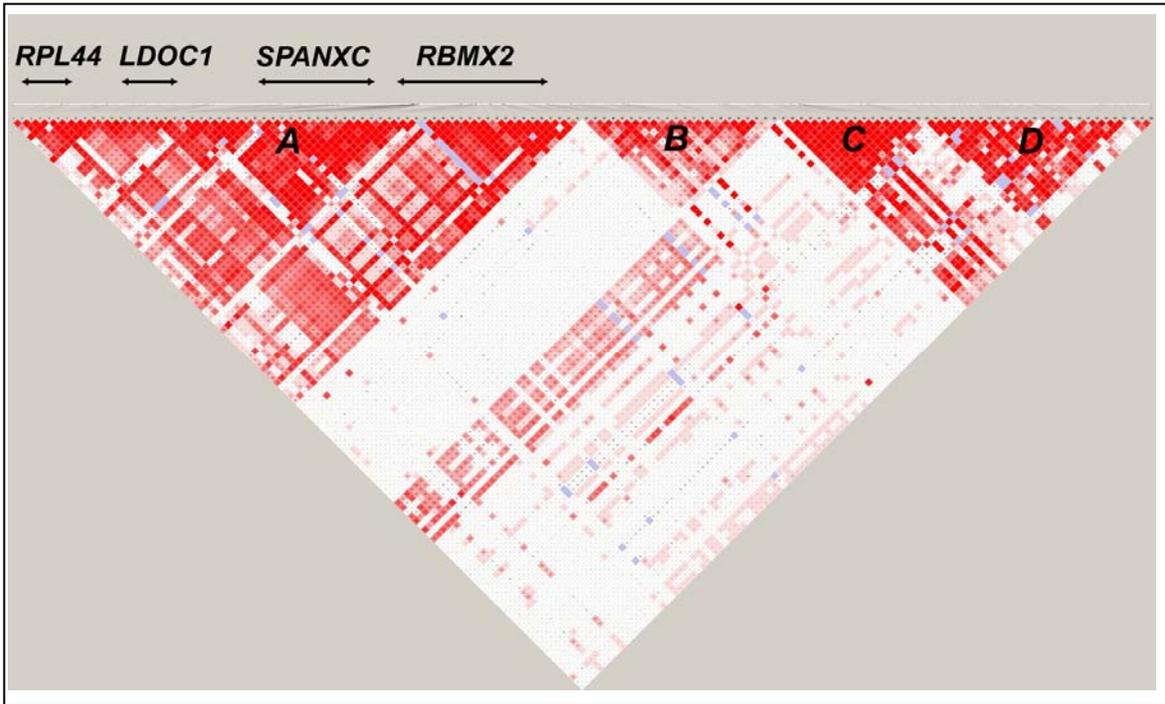


Figure 1: Linkage disequilibrium (LD) patterns at HPC-X Region A for 128 tagging SNPs for 141 controls. SNPs encompassing genes and 10 kb to each flank are positioned at top of the figure. Four major blocks are observed and marked as A, B, C and D. Red, $D' = 1$ ($LOD \geq 2$); blue, $D' = 1$ ($LOD < 2$); pink, $D' < 1$ ($LOD \geq 2$); white, $D' < 1$ ($LOD < 2$).

Task 2. To determine a set of tagging SNPs across our candidate interval at HPC-X, to genotype them in the training dataset and to test for significant association with risk of prostate cancer

- a. Determine a set of tagging SNPs across our candidate interval at HPC-X from the 246 SNPs typed in *Task 1*
- b. Genotype tagging SNPs in the remainder of the training dataset population
- c. Using single allele and sliding window haplotype analysis, determine haplotype windows statistically associated with risk of prostate cancer

Task 2 a-c has been completed.

Using LDSelect, we have determined a set of 128 tagging SNPs ($r^2 > 0.9$) and typed them in the remainder of our training dataset ($N = 292$ cases, 292 age-matched controls). All windows of statistical significance conferring risk of prostate cancer (nominal $P \leq 0.05$) within the candidate interval were from four distinct regions and are seen in **Table 2 of the attached manuscript**. Regions are listed from 5' to 3' across the candidate interval. Individual SNPs in each haplotype are identified with an internal seven-digit code. Case and control frequencies are seen for the haplotype window with the most significant P value, colored in black. All other haplotype windows showing statistical significance are colored in grey.

Task 3. To confirm associated prostate cancer gene variants in a second study population, and to extend investigation in an African American Study population. (Months 24-36)

- a. Ascertainment of independent population to be used as a test dataset
- b. Confirm or refute areas of statistical association from Task 2 in test dataset
- c. Extend findings into an African American study population, currently under ascertainment

Task 3 a, b has been completed. Task 3c is in progress.

We have recently ascertained an independent test dataset of 215 prostate cancer probands with a family history of disease and 215 age-matched controls. We used this dataset to confirm or refute statistical associations seen in the training dataset. We identified haplotype tagging SNPs (htSNP) for each of the four candidate risk haplotypes encompassing the entirety of each associated region and tested for association with risk of prostate cancer in our test dataset (**Table 3 in the attached manuscript**). Case and control frequencies differ from those reported in Table 2 due to the use of only htSNPs to define the haplotype, allowing inclusion of some samples previously dropped from analysis due to missing data. Haplotype 3 was statistically significant within the test population (6.6% of cases, 2.5% of controls; $P = 0.04$). This haplotype spans areas “A” and “B” as seen in Figure 1, identifying a recombination hotspot. Using permutation testing, this result was still significant after adjustment for multiple testing ($P = 0.048$).

We currently do not have sufficient power in our African American dataset and as a result, we are actively increasing our ascertainment efforts.

Key Research Accomplishments

1. Ascertainment of a US Caucasian study population with statistical power to detect common variants that may predispose prostate cancer risk.
2. *De novo* SNP discovery leading to discovery of 52 unpublished SNPs (at time of discovery) in a US Caucasian population. At time of writing 32 are still unpublished and have been submitted to dbSNP.
3. Identification of a 22.8 kb area spanning a recombination hotspot tagged by 6 htSNPs associated with risk of prostate cancer in both test and training datasets.

Reportable Outcomes

A Haplotype at Xq27.2 Confers Susceptibility to Prostate Cancer

Hum Genet. 2008 May;123(4):379-86; PMID 18350320

Yaspan BL, McReynolds KM, Elmore JB, Breyer JP, Bradley KM, Smith JR

Investigation of a candidate locus at HPC-X in familial prostate cancer

Poster presentation at IMPaCT meeting, Hyatt Regency Atlanta 2007

Yaspan B, McReynolds K, Elmore JB, Breyer J, Bradley K, Smith JR

No association with risk of prostate cancer for *LDOC1* and *SPANX-C* candidate genes within the HPC-X locus in a US Caucasian study population

Poster presentation at the American Society of Human Genetics Meeting, New Orleans, LA 2006

Yaspan B, Elmore JB, Breyer J, Bradley K, McReynolds K, Smith JR

Conclusions

Over the past two years, we have been systematically dissecting HPC-X to uncover the variant or variants responsible for its association with risk of prostate cancer. Starting with region A, whose boundaries we identified through shared haplotype analysis of founder populations, we have identified one haplotype tagged by 6 htSNP spanning a 22.8 kb region pinpointing a recombination hotspot associated with risk of prostate cancer. This haplotype was statistically significant for risk of prostate cancer in both our test and training datasets.

A haplotype at chromosome Xq27.2 confers susceptibility to prostate cancer

Brian L. Yaspan · Kate M. McReynolds ·
J. Bradford Elmore · Joan P. Breyer ·
Kevin M. Bradley · Jeffrey R. Smith

Received: 29 January 2008 / Accepted: 29 February 2008 / Published online: 19 March 2008
© Springer-Verlag 2008

Abstract We conducted an association study to identify risk variants for familial prostate cancer within the HPCX locus at Xq27 among Americans of Northern European descent. We investigated a total of 507 familial prostate cancer probands and 507 age-matched controls without a personal or family history of prostate cancer. The study population was subdivided into a set of training subjects to explore genetic variation of the locus potentially impacting risk of prostate cancer, and an independent set of test subjects to confirm the association and to assign significance, addressing multiple comparisons. We identified a 22.9 kb haplotype nominally associated with prostate cancer among training subjects (292 cases, 292 controls; $\chi^2 = 5.08$, $P = 0.020$), that was confirmed among test subjects (215

cases, 215 controls; $\chi^2 = 3.73$, $P = 0.040$). The haplotype predisposed to prostate cancer with an odds ratio of 3.41 (95% CI 1.04–11.17, $P = 0.034$) among test subjects. The haplotype extending from rs5907859 to rs1493189 is concordant with a prior study of the region within the Finnish founder population, and warrants further independent investigation.

Background

Linkage and genetic epidemiological data support the existence of genetic variants on the X chromosome that predispose to prostate cancer (Woolf 1960). Prostate cancer loci on both arms of the X chromosome have been identified, including the HPCX locus at Xq27–28 (Bochum et al. 2002; Brown et al. 2004; Chang et al. 2005; Cunningham et al. 2003; Farnham et al. 2005; Gillanders et al. 2004; Gudmundsson et al. 2008; Lange et al. 1999; Schleutker et al. 2000; Xu et al. 1998). The ~14 Mb linkage interval of HPCX was originally delineated within US, Swedish, and Finnish hereditary prostate cancer pedigrees (Xu et al. 1998). 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 (Baffoe-Bonnie et al. 2005a, b). The 352 kb area between these inverted repeats was the candidate interval for the present study (Fig. 1), which sought evidence of association with familial prostate cancer among Americans of Northern European descent. Our study population was uniquely comprised of independent familial prostate cancer probands, matched to controls with no personal or family history of prostate cancer. These two groups represent extremes of potential genetic load for prostate cancer. Our study included a training set of 292 case-control pairs to identify

Electronic supplementary material The online version of this article (doi:10.1007/s00439-008-0486-8) contains supplementary material, which is available to authorized users.

B. L. Yaspan · J. R. Smith
Department of Cancer Biology,
Vanderbilt-Ingram Cancer Center,
Vanderbilt University School of Medicine,
Nashville, TN, USA

K. M. McReynolds · J. B. Elmore · J. P. Breyer · K. M. Bradley ·
J. R. Smith (✉)
Department of Medicine, Vanderbilt-Ingram Cancer Center,
Vanderbilt University School of Medicine, 529 Light Hall,
2215 Garland Avenue, Nashville, TN 37232-0275, USA
e-mail: jeffrey.smith@vanderbilt.edu

J. R. Smith
Medical Research Service,
VA Tennessee Valley Healthcare System,
Nashville, TN, USA

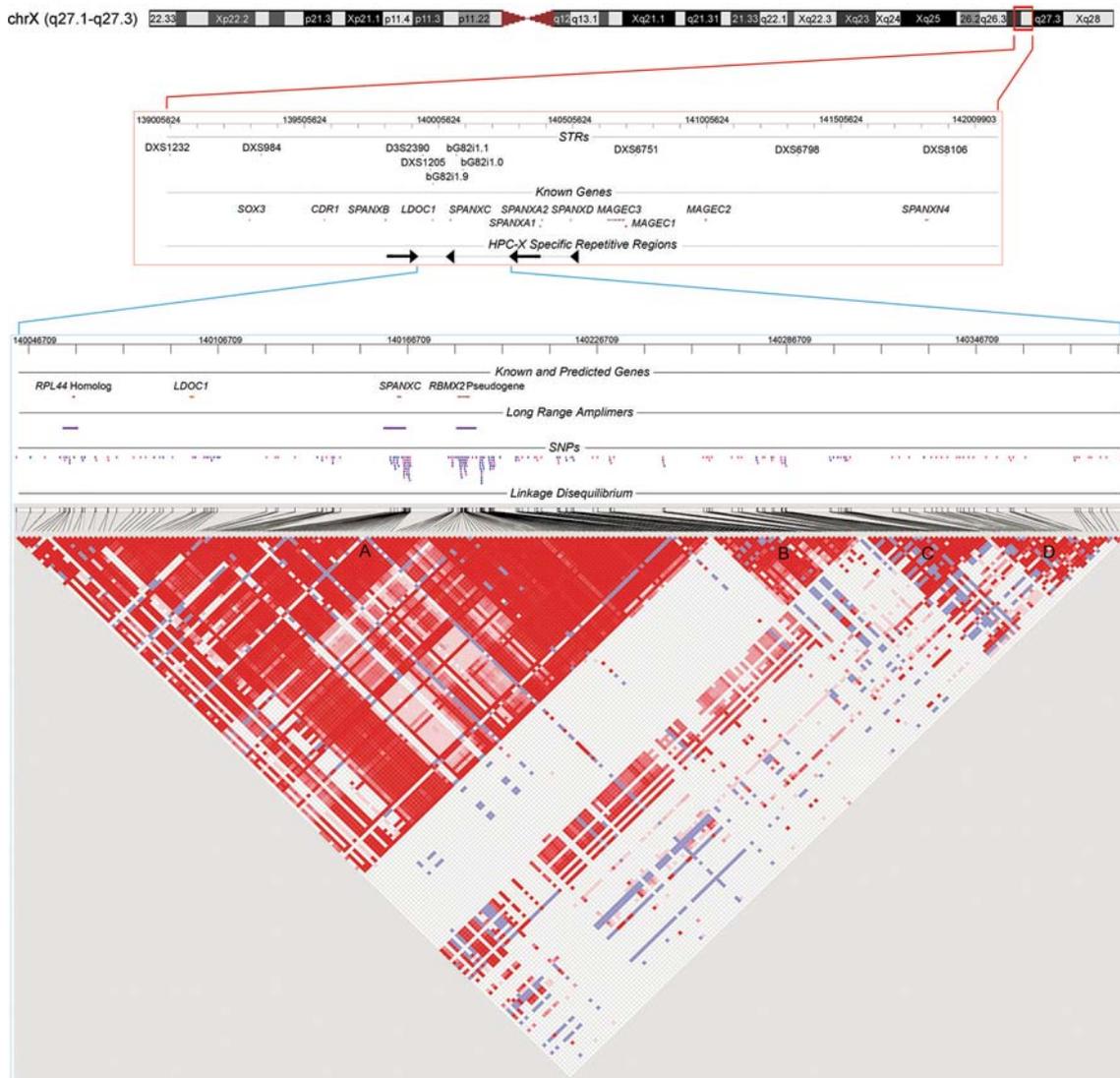


Fig. 1 *HPCX* Candidate Interval. 3 MB of Xq27.1–27.3 depicting previously genotyped STR markers, annotated genes, and a complex *HPCX* specific repeat is shown at top (red bounding box, NCBI build 36.1 ChrX: 139005624–142009903 bp). The candidate interval for study is zoomed at bottom (blue bounding 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 amplifiers (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 $MAF \geq 0.05$. Red $D' = 1$ ($\text{lod} \geq 2$); blue $D' = 1$ ($\text{lod} < 2$); pink $D' < 1$ ($\text{lod} \geq 2$); white $D' < 1$ ($\text{lod} < 2$). Blocks of LD are denoted A, B, C and D. Block A contains all four candidate gene regions

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 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 prostate cancer, and included non-unique genomic regions that are not amenable to current high-throughput techniques.

Materials and methods

Samples

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 (adjacent hospitals) 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 prostate cancer cases were ascertained at the time of treatment for the principal diagnosis of prostate cancer, and controls were ascertained at the time of routine preventative screening for prostate cancer. All prostate cancer probands included in the study are from pedigrees with a family history of prostate cancer, and all control probands are from pedigrees without a family history of prostate cancer. 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 \geq 4 ng/ml or record of abnormal digital rectal examination. Controls were matched to cases 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 1.7 for cases, and 1.8 for controls. Initial accruals included 292 unrelated, independent familial prostate cancer probands and 292 age-matched controls, comprising a training study group. Subsequent accruals included 215 additional unrelated, independent prostate cancer probands and 215 additional age-matched controls, comprising a separate test study group. The geographical distribution (by state of residence) of training and test subjects was not significantly different. Analyses preferentially employed prostatectomy specimen over biopsy Gleason score (available for 87% of cases). Table 1 provides characteristics of the study population.

Genotyping methods

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 (Yaspan et al. 2007). 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 Electronic supplementary Table 1.

Candidate polymorphisms

To capture genetic diversity across the candidate interval, database SNPs annotated in dbSNP were screened for common polymorphism in a set of 40 independent familial prostate cancer probands from the training study group. This included 415 annotated SNPs on chromosome X between positions 140,036,557 and 140,388,361 (NCBI Build 36.1). 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 prostate cancer 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 by aligning transcript evidence on the genomic map. We employed two single-stranded conformation polymorphism methods (redundant) and re-sequencing for SNP discovery, as previously described (Yaspan et al. 2007). Exons of the four genes were also re-sequenced for all 40 prostate cancer 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. Unique priming sites were identified flanking non-unique regions and amplified using the Expand Long Template PCR System

Table 1 Study population characteristics

	Training		Test		Combined	
	Controls	Cases	Controls	Cases	Controls	Cases
No.	292	292	215	215	507	507
Mean age ^a (years)	63.4	61.3	60.7	60.6	62.3	61.0
Median PSA ^a	0.95	5.7	0.92	5.6	0.92	5.7
Median Gleason sum	–	6	–	6	–	6
Gleason sum \leq 6, No.	–	145	–	114	–	259
Gleason sum \geq 7, No.	–	130	–	96	–	226
Affected in pedigree, No. ^b	0	292	215	–	507	–
	2	–	–	184	–	326
	\geq 3	–	–	108	–	181

^a At diagnosis for cases, at screen for controls

^b Proband plus first and second degree affected relatives

(Roche Diagnostics, Indianapolis, IN). Amplimer primers for long range PCR are provided in Electronic supplementary Table 1. 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. Nested amplimers were designed in overlapping fashion along a given long range amplimer, enabling PCR-based assays. These primers are provided in Electronic supplementary Table 2. High pairwise LD between SNPs within a given non-unique region and flanking unique SNPs supported correct non-unique copy assay (visible in Fig. 1). All nested assays within long-range amplimers yielded one allele per subject, concordant with a unique X-chromosomal region for a male.

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 (Barrett et al. 2005). LDSelect was used for tagging SNP selection, specifying a MAF threshold of 0.05 and an r^2 threshold of 0.9 (Carlson et al. 2004). 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 were 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 Electronic supplementary Table 1.

Statistical analyses

A sliding window approach tested a haplotype window of N markers, sliding the window along the map in single marker increments (Fallin et al. 2001; Mathias et al. 2006; Yaspan et al. 2007). 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 2×2 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 prostate cancer 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 pseudo case status was permuted among cases and controls of the test group. 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 regression model was adjusted for the matching variable: age (age at diagnosis for cases, age at negative screen for controls). Permutation testing was employed to assign significance.

Results and discussion

Our allele discovery and characterization was done within a screening set of 40 familial prostate cancer probands from the training group. 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 Fig. 1).

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 prostate cancer 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 prostate cancer risk. These four candidate risk haplotypes are numbered 1–4 in Table 2.

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 2. 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 3, haplotype 1 ($\chi^2 = 5.24$, $P = 0.023$) and haplotype 3 ($\chi^2 = 5.08$, $P = 0.020$)). We evaluated evidence

Table 2 Sliding window risk haplotypes at Xq27—training subjects

Location	Significant Haplotype Windows ^a	Allele	Cases	Controls	P^b
Haplotype 1 (<i>hRPL44</i>)	rs11095852	A	92 (39.8)	63 (27.3)	0.003
	rs5907823	C			
	rs7880499	G			
	rs1016824	T			
	rs12156848	G			
	rs7885649	A			
	rs5953563	A			
Haplotype 2 (<i>RBMX2P1</i>)	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			

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

^b P for haplotype designated in black, with corresponding numbers of cases and controls, and haplotype frequencies (%)

Table 3 Tagged risk haplotypes at Xq27—training and test subjects

Location	htSNP	Allele	Training			Test		
			Cases	Controls	<i>P</i>	Cases	Controls	<i>P</i>
Haplotype 1 (<i>hRPL44</i>)	rs5907823	C						
	rs7880499	G						
	rs1016824	T	95	71	0.023	73	79	0.536
	rs12156848	G	(38.6)	(28.9)		(39.3)	(42.5)	
	rs7885649	A						
Haplotype 2 (<i>RBMX2P1</i>)	rs845150	A	19	9	0.062			
	rs861508	C						
Haplotype 3 (ChrX: 140190766- 140213636)	rs845165	A						
	rs845190	T						
	rs845187	C	18	7	0.020	13	5	0.040
	rs845186	A	(6.8)	(2.6)		(6.6)	(2.5)	
	rs1493189	A						
Haplotype 4 (ChrX: 140266943- 140295222)	rs844963	A	22	14	0.168			
	rs844956	C	(7.7)	(4.9)				

Numbers of subjects and tagged haplotype frequencies (%) are indicated

of association between these two htSNP haplotypes and prostate cancer in a second, independent study group of 215 familial prostate cancer 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 3, $\chi^2 = 3.73$, $P = 0.040$). 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 prostate cancer, 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 prostate cancer 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 number of cases with aggressive prostate cancer among the test subjects was too few to provide a significant estimate of effect size. However, the 226 cases with aggressive prostate cancer (Gleason score ≥ 7) among the full study population yielded a more marked 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 (Chang et al. 2005).

The location of the haplotype found to be significantly associated with prostate cancer in this study coincides with that described through prior high-density simple tandem repeat mapping within a Finnish study population (Baffoe-Bonnie et al. 2005a, b). Among the simple tandem repeats,

bG82i1.1 was most significantly associated with prostate cancer in the prior study. The peak associated haplotype 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 Fig. 1. The most centromeric SNP of the associated haplotype (rs5907859) is 4.0 kb downstream from bG82i1.1. The most telomeric SNP of the associated haplotype (rs1493189) is 1.9 kb downstream from bG82i1.0. The same genomic region of interest 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 prostate cancer recently published by Thomas et al. rs845189 has a Whole Genome Rank of 1,135 out of 527,869 SNPs assessed, with a significance of $P = 0.002$ (Thomas et al. 2008). 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 LD 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 (Kouprina et al. 2007). That study also found no evidence to support an association between *SPANXC* alleles and risk of prostate cancer. 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 prostate cancer at Xq27.

The haplotype significantly associated with prostate cancer 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. 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 had an $r^2 > 0.8$ with the respective sub-haplotypes. A matrix depiction of pairwise

r^2 values between these is illustrated in Electronic supplementary Fig. 1. The T30S (ss78456788) variant of *SPANXC* (in LD Block A) 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 group 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. With only one exception, χ^2 tests of association for these block A SNP–block B SNP pair haplotypes were each associated with prostate cancer in our training group (P range 0.0008–0.030). These SNP pairs and the original sliding window haplotype spanning the LD break each detect the association with prostate cancer with varying efficiencies.

Our study sought to identify the genetic variant predisposing to familial prostate cancer at Xq27, a locus initially identified by linkage study of American, Swedish, and Finnish hereditary prostate cancer pedigrees, and subsequently refined by linkage disequilibrium analysis of the Finnish familial prostate cancer cases. After a comprehensive effort in the present study, we identified a single candidate haplotype that was associated with familial prostate cancer 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 (Hunter et al. 2007; Tang et al. 2005). We believe that this haplotype represents the best candidate within the region for further investigation within additional study populations. If confirmed, these findings begin to clarify the X-linked heritable component of prostate cancer risk.

Acknowledgments 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.

References

Baffoe-Bonnie AB, Smith JR, Stephan DA, Schleutker J, Carpten JD, Kainu T, Gillanders EM, Matikainen M, Teslovich TM, Tammela

- T, Sood R, Balslem AM, Scarborough SD, Xu J, Isaacs WB, Trent JM, Kallioniemi OP, Bailey-Wilson JE (2005a) A major locus for hereditary prostate cancer in Finland: localization by linkage disequilibrium of a haplotype in the HPCX region. *Hum Genet* 117:307–316
- Baffoe-Bonnie AB, Smith JR, Stephan DA, Schleutker J, Carpten JD, Kainu T, Gillanders EM, Matikainen M, Teslovich TM, Tammela T, Sood R, Balslem AM, Scarborough SD, Xu J, Isaacs WB, Trent JM, Kallioniemi OP, Bailey-Wilson JE (2005b) A major locus for hereditary prostate cancer in Finland: localization by linkage disequilibrium of a haplotype in the HPCX region (Erratum). *Hum Genet* 118:307
- Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21:263–265
- Bochum S, Paiss T, Vogel W, Herkommer K, Hautmann R, Haeussler J (2002) Confirmation of the prostate cancer susceptibility locus HPCX in a set of 104 German prostate cancer families. *Prostate* 52:12–19
- Brown WM, Lange EM, Chen H, Zheng SL, Chang B, Wiley KE, Isaacs SD, Walsh PC, Isaacs WB, Xu J, Cooney KA (2004) Hereditary prostate cancer in African American families: linkage analysis using markers that map to five candidate susceptibility loci. *Br J Cancer* 90:510–514
- Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* 74:106–120
- Chang BL, Isaacs SD, Wiley KE, Gillanders EM, Zheng SL, Meyers DA, Walsh PC, Trent JM, Xu J, Isaacs WB (2005) Genome-wide screen for prostate cancer susceptibility genes in men with clinically significant disease. *Prostate* 64:356–361
- Cunningham JM, McDonnell SK, Marks A, Hebbing S, Anderson SA, Peterson BJ, Slager S, French A, Blute ML, Schaid DJ, Thibodeau SN (2003) Genome linkage screen for prostate cancer susceptibility loci: results from the Mayo Clinic Familial Prostate Cancer Study. *Prostate* 57:335–346
- Fallin D, Cohen A, Essioux L, Chumakov I, Blumenfeld M, Cohen D, Schork NJ (2001) Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. *Genome Res* 11:143–151
- Farnham JM, Camp NJ, Swensen J, Tavtigian SV, Albright LA (2005) Confirmation of the HPCX prostate cancer predisposition locus in large Utah prostate cancer pedigrees. *Hum Genet* 116:179–185
- Gillanders EM, Xu J, Chang BL, Lange EM, Wiklund F, Bailey-Wilson JE, Baffoe-Bonnie A, Jones M, Gildea D, Riedesel E, Albertus J, Isaacs SD, Wiley KE, Mohai CE, Matikainen MP, Tammela TL, Zheng SL, Brown WM, Rokman A, Carpten JD, Meyers DA, Walsh PC, Schleutker J, Gronberg H, Cooney KA, Isaacs WB, Trent JM (2004) Combined genome-wide scan for prostate cancer susceptibility genes. *J Natl Cancer Inst* 96:1240–1247
- Gudmundsson J, Sulem P, Rafnar T, Berghthorsson JT, Manolescu A, Gudbjartsson D, Agnarsson BA, Sigurdsson A, Benediktsson KR, Blondal T, Jakobsdottir M, Stacey SN, Kosti J, Kristinsson KT, Birgisdottir B, Ghosh S, Magnusdottir DN, Thorlacius S, Thorleifsson G, Zheng SL, Sun J, Chang BL, Elmore JB, Breyer JP, McReynolds KM, Bradley KM, Yaspan BL, Wiklund F, Stattin P, Lindstrom S, Adami HO, McDonnell SK, Schaid DJ, Cunningham JM, Wang L, Cerhan JR, St Sauver JL, Isaacs SD, Wiley KE, Partin AW, Walsh PC, Polo S, Ruiz-Echarri M, Navarrete S, Fuertes F, Saez B, Godino J, Weijerman PC, Swinkels DW, Aben KK, Witjes JA, Suarez BK, Helfand BT, Frigge ML, Kristjansson K, Ober C, Jonsson E, Einarsson GV, Xu J, Gronberg H, Smith JR, Thibodeau SN, Isaacs WB, Catalona WJ, Mayordomo JI, Kiemeny LA, Barkardottir RB, Gulcher JR, Thorsteinsdottir U, Kong A, Stefansson K (2008) Common sequence variants on

- 2p15 and Xp11.22 confer susceptibility to prostate cancer. *Nat Genet* 40:281–283
- Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang Z, Welch R, Hutchinson A, Wang J, Yu K, Chatterjee N, Orr N, Willett WC, Colditz GA, Ziegler RG, Berg CD, Buys SS, McCarty CA, Feigelson HS, Calle EE, Thun MJ, Hayes RB, Tucker M, Gerhard DS, Fraumeni JF Jr, Hoover RN, Thomas G, Chanock SJ (2007) A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet* 39:870–874
- Kouprina N, Noskov VN, Solomon G, Otstot J, Isaacs W, Xu J, Schleutker J, Larionov V (2007) Mutational analysis of SPANX genes in families with X-linked prostate cancer. *Prostate* 67:820–828
- Lange EM, Chen H, Brierley K, Perrone EE, Bock CH, Gillanders E, Ray ME, Cooney KA (1999) Linkage analysis of 153 prostate cancer families over a 30-cM region containing the putative susceptibility locus HPCX. *Clin Cancer Res* 5:4013–4020
- Mathias RA, Gao P, Goldstein JL, Wilson AF, Pugh EW, Furbert-Harris P, Dunston GM, Malveaux FJ, Togiias A, Barnes KC, Beaty TH, Huang SK (2006) A graphical assessment of *p*-values from sliding window haplotype tests of association to identify asthma susceptibility loci on chromosome 11q. *BMC Genet* 7:38
- Schleutker J, Matikainen M, Smith J, Koivisto P, Baffoe-Bonnie A, Kainu T, Gillanders E, Sankila R, Pukkala E, Carpten J, Stephan D, Tammela T, Brownstein M, Bailey-Wilson J, Trent J, Kallioniemi OP (2000) A genetic epidemiological study of hereditary prostate cancer (HPC) in Finland: frequent HPCX linkage in families with late-onset disease. *Clin Cancer Res* 6:4810–4815
- Tang H, Quertermous T, Rodriguez B, Kardia SL, Zhu X, Brown A, Pankow JS, Province MA, Hunt SC, Boerwinkle E, Schork NJ, Risch NJ (2005) Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies. *Am J Hum Genet* 76:268–275
- Thomas G, Jacobs KB, Yeager M, Kraft P, Wacholder S, Orr N, Yu K, Chatterjee N, Welch R, Hutchinson A, Crenshaw A, Cancel-Tassin G, Staats BJ, Wang Z, Gonzalez-Bosquet J, Fang J, Deng X, Berndt SI, Calle EE, Feigelson HS, Thun MJ, Rodriguez C, Albanes D, Virtamo J, Weinstein S, Schumacher FR, Giovannucci E, Willett WC, Cussenot O, Valeri A, Andriole GL, Crawford ED, Tucker M, Gerhard DS, Fraumeni JF Jr, Hoover R, Hayes RB, Hunter DJ, Chanock SJ (2008) Multiple loci identified in a genome-wide association study of prostate cancer. *Nat Genet* 40:310–315
- Woolf CM (1960) An investigation of the familial aspects of carcinoma of the prostate. *Cancer* 13:739–744
- Xu J, Meyers D, Freije D, Isaacs S, Wiley K, Nusskern D, Ewing C, Wilkens E, Bujnovszky P, Bova GS, Walsh P, Isaacs W, Schleutker J, Matikainen M, Tammela T, Visakorpi T, Kallioniemi OP, Berry R, Schaid D, French A, McDonnell S, Schroeder J, Blute M, Thibodeau S, Gronberg H, Emanuelsson M, Damber JE, Bergh A, Jonsson BA, Smith J, Bailey-Wilson J, Carpten J, Stephan D, Gillanders E, Amundson I, Kainu T, Freas-Lutz D, Baffoe-Bonnie A, Van Aucken A, Sood R, Collins F, Brownstein M, Trent J (1998) Evidence for a prostate cancer susceptibility locus on the X chromosome. *Nat Genet* 20:175–179
- Yaspan BL, Breyer JP, Cai Q, Dai Q, Elmore JB, Amundson I, Bradley KM, Shu XO, Gao YT, Dupont WD, Zheng W, Smith JR (2007) Haplotype analysis of CYP11A1 identifies promoter variants associated with breast cancer risk. *Cancer Res* 67:5673–5682