Large-scale fine mapping of the \textit{HNF1B} locus and prostate cancer risk

Sonja I. Berndt\textsuperscript{1,*}, Joshua Sampson\textsuperscript{1}, Meredith Yeager\textsuperscript{1,2}, Kevin B. Jacobs\textsuperscript{3}, Zhaoming Wang\textsuperscript{1,2}, Amy Hutchinson\textsuperscript{1,2}, Charles Chung\textsuperscript{1,2}, Nick Orr\textsuperscript{1}, Sholom Wacholder\textsuperscript{1}, Nilanjan Chatterjee\textsuperscript{1}, Kai Yu\textsuperscript{1}, Peter Kraft\textsuperscript{4}, Heather Spencer Feigelson\textsuperscript{5}, Michael J. Thun\textsuperscript{6}, W. Ryan Diver\textsuperscript{6}, Demetrius Albanes\textsuperscript{1}, Jarmo Virtamo\textsuperscript{7}, Stephanie Weinstein\textsuperscript{1}, Fredrick R. Schumacher\textsuperscript{8}, Geraldine Cancel-Tassin\textsuperscript{9}, Olivier Cussenot\textsuperscript{9}, Antoine Valeri\textsuperscript{9}, Gerald L. Andriole\textsuperscript{10}, E. David Crawford\textsuperscript{11}, Christopher Haiman\textsuperscript{6}, Brian Henderson\textsuperscript{8}, Laurence Kolonel\textsuperscript{12}, Loic Le Marchand\textsuperscript{12}, Afshan Siddiq\textsuperscript{13}, Elio Riboli\textsuperscript{13}, Ruth C. Travis\textsuperscript{14}, Rudolf Kaaks\textsuperscript{15}, William Isaacs\textsuperscript{16}, Sarah Isaacs\textsuperscript{16}, Kathleen E. Wiley\textsuperscript{16}, Henrik Gronberg\textsuperscript{17}, Fredrik Wiklund\textsuperscript{17}, Pär Stattin\textsuperscript{18}, Jianfeng Xu\textsuperscript{19}, S. Lilly Zheng\textsuperscript{19}, Jielin Sun\textsuperscript{19}, Lars J. Vatten\textsuperscript{20}, Kristian Hveem\textsuperscript{20}, Inger Njolstad\textsuperscript{21}, Daniela S. Gerhard\textsuperscript{22}, Margaret Tucker\textsuperscript{1}, Richard B. Hayes\textsuperscript{23}, Robert N. Hoover\textsuperscript{1}, Joseph F. Fraumeni Jr\textsuperscript{1}, David J. Hunter\textsuperscript{24}, Gilles Thomas\textsuperscript{1} and Stephen J. Chanock\textsuperscript{1}

\textsuperscript{1}Division of Cancer Epidemiology and Genetics, Department of Health and Human Services, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA, \textsuperscript{2}Core Genotyping Facility, Advanced Technology Program, SAIC-Frederick Inc., NCI-Frederick, Frederick, MD, USA, \textsuperscript{3}Bioinformed Consulting Services, Gaithersburg, MD, USA, \textsuperscript{4}Program in Molecular and Genetic Epidemiology, Department of Epidemiology, Harvard School of Public Health, Boston, MA, USA, \textsuperscript{5}Institute for Health Research, Kaiser Permanente, Denver, CO, USA, \textsuperscript{6}Department of Epidemiology and Surveillance Research, American Cancer Society, Atlanta, GA, USA, \textsuperscript{7}Department of Chronic Disease Prevention, National Institute for Health and Welfare, Helsinki, Finland, \textsuperscript{8}Department of Preventive Medicine, Keck School of Medicine, Los Angeles, CA, USA, \textsuperscript{9}Centre de Recherche pour les Pathologies Prostatiques, Hôpital Tenon, Assistance Publique-Hôpitaux de Paris, 75970 Paris, France, \textsuperscript{10}Division of Urologic Surgery, Washington University School of Medicine, St Louis, MO, USA, \textsuperscript{11}Department of Surgery, University of Colorado at Denver and Health Sciences Center, Denver, CO, USA, \textsuperscript{12}Program in Molecular and Genetic Epidemiology, Department of Epidemiology, Harvard School of Public Health, Boston, MA, USA, \textsuperscript{13}Department of Epidemiology, Public Health and Primary Care, Imperial College London, London, UK, \textsuperscript{14}Cancer Epidemiology Unit, Nuffield Department of Clinical Medicine, University of Oxford, Oxford, UK, \textsuperscript{15}Division of Cancer Epidemiology, German Cancer Research Centre (DKFZ), Heidelberg, Germany, \textsuperscript{16}Department of Urology, Johns Hopkins Medical Institutions, Baltimore, MD, USA, \textsuperscript{17}Department of Medical Epidemiology and Biostatistics, CLINTEC, Karolinska Institute, Stockholm, Sweden, \textsuperscript{18}Department of Surgical and Perioperative Sciences, Urology and Andrology, Umeå University, Umeå, Sweden, \textsuperscript{19}Center for Cancer Genomics, Wake Forest University School of Medicine, Winston-Salem, NC, USA, \textsuperscript{20}Department of Public Health and General Practice, Norwegian University of Science and Technology, Trondheim, Norway, \textsuperscript{21}Institute of Community Medicine, University of Tromsø, Tromsø, Norway, \textsuperscript{22}Office of Cancer Genomics, Department of Health and Human Services, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA, \textsuperscript{23}Department of Environmental Medicine, New York University Medical Center, New York, NY, USA, \textsuperscript{24}Channing Laboratory, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA, USA

Received February 1, 2011; Revised April 14, 2011; Accepted May 9, 2011

\textsuperscript{*}To whom correspondence should be addressed at: Sonja Berndt, Division of Cancer Epidemiology and Genetics, National Cancer Institute, 6120 Executive Blvd, EPS 8116, MSC 7240, Bethesda, MD 20892-7240. Tel: +1 3015947898; Fax: +1 3014021819; Email: berndts@mail.nih.gov

Published by Oxford University Press 2011.
Previous genome-wide association studies have identified two independent variants in HNF1B as susceptibility loci for prostate cancer risk. To fine-map common genetic variation in this region, we genotyped 79 single nucleotide polymorphisms (SNPs) in the 17q12 region harboring HNF1B in 10,272 prostate cancer cases and 9,123 controls of European ancestry from 10 case–control studies as part of the Cancer Genetic Markers of Susceptibility (CGEMS) initiative. Ten SNPs were significantly related to prostate cancer risk at a genome-wide significance level of $P < 5 \times 10^{-8}$ with the most significant association with rs4430796 ($P = 1.62 \times 10^{-24}$). However, risk within this first locus was not entirely explained by rs4430796. Although modestly correlated ($r^2 = 0.64$), rs7405696 was also associated with risk ($P = 9.35 \times 10^{-23}$) even after adjustment for rs4430769 ($P = 0.007$). As expected, rs11649743 was related to prostate cancer risk ($P = 3.54 \times 10^{-8}$); however, the association within this second locus was stronger for rs4794758 ($P = 4.95 \times 10^{-19}$), which explained all of the risk observed with rs11649743 when both SNPs were included in the same model ($P = 0.32$ for rs11649743; $P = 0.002$ for rs4794758). Sequential conditional analyses indicated that five SNPs (rs4430796, rs7405696, rs4794758, rs1016990 and rs3094509) together comprise the best model for risk in this region. This study demonstrates a complex relationship between variants in the HNF1B region and prostate cancer risk. Further studies are needed to investigate the biological basis of the association of variants in 17q12 with prostate cancer.

**INTRODUCTION**

Of all cancers, prostate cancer is one of the most heritable with genetic factors estimated to account for 42% of the risk (1). Genome-wide association studies (GWAS) have been highly successful in discovering susceptibility loci for prostate cancer and at least 30 loci have been identified to date (2–15). One of the earliest loci to be discovered for prostate cancer was a variant, rs4430796, in HNF1B at chromosome 17q12 in men of European background (6). In a subsequent GWAS in Japanese men, the same locus was identified (15). A second independent variant, rs11649743, located at chromosome 17q12 and separated by a recombination hotspot from the first variant, was subsequently found to be associated with risk (10). The HNF1B locus as well as two other prostate cancer susceptibility loci, chromosome 7p15.2 (JAZF1) and chromosome 2p21 (THADA), have also been shown to be associated with diabetes risk (6,16). Although epidemiologic studies have shown that diabetes is inversely associated with risk after adjustment for rs4430796 ($P = 3.54 \times 10^{-8}$); however, the association within this second locus was stronger for rs4794758 ($P = 4.95 \times 10^{-19}$), which explained all of the risk observed with rs11649743 when both SNPs were included in the same model ($P = 0.32$ for rs11649743; $P = 0.002$ for rs4794758). Sequential conditional analyses indicated that five SNPs (rs4430796, rs7405696, rs4794758, rs1016990 and rs3094509) together comprise the best model for risk in this region. This study demonstrates a complex relationship between variants in the HNF1B region and prostate cancer risk. Further studies are needed to investigate the biological basis of the association of variants in 17q12 with prostate cancer.

**RESULTS**

We analyzed 79 SNPs located in a 249 kb region surrounding HNF1B (chromosome 17: 33,010,707–33,259,778) in 10,272 prostate cancer cases and 9,123 controls from 10 studies (Supplementary Material, Table S1). Ten SNPs were significantly associated with prostate cancer risk below the threshold of genome-wide significance ($P < 5 \times 10^{-8}$); the most significant association observed was the previously identified SNP rs4430796 ($P = 1.62 \times 10^{-24}$) (Table 1, Fig. 1A). Eight of the 10 significant SNPs were located in the first HNF1B region associated with prostate cancer, and all eight were highly correlated in controls with $D^2 \geq 0.6$ and pairwise $r^2$ values between 0.13 and 0.94. However, risk within this first locus was not entirely explained by rs4430796. Although modestly correlated ($r^2 = 0.64$), rs7405696 was also associated with prostate cancer risk ($P = 9.35 \times 10^{-23}$). After conditioning on rs4430796, rs7405696 retained an association with risk ($P = 0.007$). None of the other SNPs in region 1 remained associated with risk after adjustment for rs4430796 (Supplementary Material, Table S2).

Two SNPs located in the second identified region by Sun et al. (10) were also associated with the prostate cancer risk at a significance level of $5 \times 10^{-8}$. As expected, the previously identified SNP in the second region, rs11649743, was associated with prostate cancer risk ($P = 3.54 \times 10^{-8}$); however, the association within this second locus was stronger for rs4794758 ($P = 4.95 \times 10^{-19}$). The two SNPs were correlated in controls ($r^2 = 0.61$), but when rs11649743 and rs4794758 were included in the same model, only rs4794758 remained associated with risk ($P = 0.002$ for rs4794758; $P = 0.32$ for rs11649743). Interestingly, it accounted for the risk associated with rs11649743.
Table 1. SNPs in the HNF1B region that were associated with prostate cancer at genome-wide significance levels ($P < 5 \times 10^{-8}$).

<table>
<thead>
<tr>
<th>SNP</th>
<th>Base pair position</th>
<th>Minor allele</th>
<th>Other allele</th>
<th>Minor allele frequency: cases</th>
<th>OR (95% CI)$^a$</th>
<th>Other allele frequency: controls</th>
<th>OR (95% CI)$^a$</th>
<th>$P$-value</th>
<th>Heterogeneity P</th>
<th>$\Delta$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs4430796</td>
<td>33172153</td>
<td>G</td>
<td>A</td>
<td>0.42</td>
<td>0.81 (0.77–0.84)</td>
<td>0.44</td>
<td>0.81 (0.77–0.84)</td>
<td>0.0003</td>
<td>0.64</td>
<td>0</td>
</tr>
<tr>
<td>rs2005705</td>
<td>33170413</td>
<td>T</td>
<td>C</td>
<td>0.43</td>
<td>0.81 (0.78–0.85)</td>
<td>0.43</td>
<td>0.81 (0.78–0.85)</td>
<td>0.0003</td>
<td>0.64</td>
<td>0</td>
</tr>
<tr>
<td>rs7405696</td>
<td>33176148</td>
<td>C</td>
<td>G</td>
<td>0.43</td>
<td>0.81 (0.78–0.85)</td>
<td>0.43</td>
<td>0.81 (0.78–0.85)</td>
<td>0.0003</td>
<td>0.64</td>
<td>0</td>
</tr>
<tr>
<td>rs7501939</td>
<td>33175269</td>
<td>T</td>
<td>C</td>
<td>0.43</td>
<td>0.81 (0.78–0.85)</td>
<td>0.43</td>
<td>0.81 (0.78–0.85)</td>
<td>0.0003</td>
<td>0.64</td>
<td>0</td>
</tr>
<tr>
<td>rs4239217</td>
<td>33173100</td>
<td>G</td>
<td>A</td>
<td>0.43</td>
<td>0.81 (0.78–0.85)</td>
<td>0.43</td>
<td>0.81 (0.78–0.85)</td>
<td>0.0003</td>
<td>0.64</td>
<td>0</td>
</tr>
<tr>
<td>rs757210</td>
<td>33170628</td>
<td>A</td>
<td>G</td>
<td>0.43</td>
<td>0.81 (0.78–0.85)</td>
<td>0.43</td>
<td>0.81 (0.78–0.85)</td>
<td>0.0003</td>
<td>0.64</td>
<td>0</td>
</tr>
<tr>
<td>rs3760511</td>
<td>33176148</td>
<td>C</td>
<td>A</td>
<td>0.43</td>
<td>0.81 (0.78–0.85)</td>
<td>0.43</td>
<td>0.81 (0.78–0.85)</td>
<td>0.0003</td>
<td>0.64</td>
<td>0</td>
</tr>
<tr>
<td>rs4794758</td>
<td>33154541</td>
<td>T</td>
<td>C</td>
<td>0.43</td>
<td>0.81 (0.78–0.85)</td>
<td>0.43</td>
<td>0.81 (0.78–0.85)</td>
<td>0.0003</td>
<td>0.64</td>
<td>0</td>
</tr>
</tbody>
</table>

$^a$Odds ratio per minor allele.

To examine the interdependence of the signals observed on chromosome 17q12, we first conducted a set of sequential conditional analyses, conditioning on the most significant SNP from the unconditional analysis and each conditional analysis sequentially until no SNPs remain nominally associated with risk ($P < 0.05$) (Supplementary Material, Table S2). After conditioning on the most significant SNP (rs4430796 in region 1), six SNPs remained nominally associated with risk ($P < 0.05$) with the most significant SNP being rs4794758 in region 2 (Fig. 1B, Supplementary Material, Table S2). Although rs4430796 and rs4794758 were separated by a modest recombination hotspot, there was some correlation between them ($r^2 = 0.04$) and consequently the $P$-value for rs4794758 was attenuated after conditioning on rs4430796 ($P = 3.45 \times 10^{-5}$). After conditioning on both rs4430796 and rs4794758, four SNPs were nominally associated with risk ($P < 0.05$) with the most significant SNP being rs1016990 ($P = 0.009$). This SNP was only nominally associated with risk in the unconditional model ($P = 0.0002$) and not significantly associated with risk after conditioning on rs4430796 only. Further sequential conditional analyses yielded rs7405696 ($P = 0.01$) as the most significant SNP after conditioning on rs4430796, rs4794758 and rs1016990, followed by rs3094509 ($P = 0.02$) after conditioning on rs4430796, rs4794758, rs1016990 and rs7405696. No other SNPs remained nominally associated with prostate cancer risk after conditioning on these five SNPs, suggesting that these SNPs (i.e. rs4430796, rs4794758, rs1016990, rs7405696 and rs3094509) capture the risk in this region.

To ascertain whether the same SNPs were identified using other statistical approaches, we performed forward stepwise regression adding SNPs below a significance level of 0.05. This method resulted in the inclusion of four SNPs: rs4430796, rs4794758 and rs1016990 and rs74056969 in the model. Using lasso, five SNPs (rs4430796, rs4794758, rs1016990, rs74056969 and rs3094509) capture the risk in this region. These SNPs (i.e. rs4430796, rs4794758, rs1016990, rs74056969 and rs3094509) capture the risk in this region.
Table 3). Men with eight or more risk alleles had a 1.88-fold increased risk of prostate cancer compared with men with zero to two risk alleles (95% CI: 1.52–2.33, \( P = 4.29 \times 10^{-2} \)). In comparison, when only the most significant SNP from region 1 (rs4430796) and region 2 (rs4794758) were examined, men with four or more risk alleles had a 1.69-fold risk compared with men with no risk alleles (95% CI: 1.40–2.04, \( P = 6.66 \times 10^{-2} \)) with \( P_{\text{trend}} = 1.80 \times 10^{-25} \). The variance explained by the five SNPs was 0.9% compared with 0.7% for the two best SNPs from region 1 and region 2.

Finally, we conducted stratified analyses by family history of prostate cancer and did observe a qualitative interaction for rs2107131 (\( P_{\text{interaction}} = 4.29 \times 10^{-2} \)) that remained statistically significant after adjustment for multiple testing (\( P_{\text{adj}} = 0.003 \)) (Supplementary Material, Table S5). Among men with a family history of prostate cancer, the T allele at rs2107131 was associated with a reduced risk of prostate cancer (OR = 0.85; 95% CI: 0.74–0.97), whereas among men without a family history, it was associated with an
increased risk of prostate cancer (OR = 1.13; 95% CI: 1.08–1.19). Stratified analyses were also performed for prostate cancer aggressiveness (Gleason $7$ and Stage A/B versus Gleason $\geq 7$ or Stage C/D); however, there were no significant differences beyond what would be expected by chance (Supplementary Material, Table S6). We did not observe any significant heterogeneity between studies beyond what would be expected by chance (Supplementary Material, Table S1), except for rs1058166 ($P_{\text{heterogeneity}} = 0.0002$).

**DISCUSSION**

Our fine mapping study of a region on 17q12 associated with prostate cancer confirmed the previously established signals (6,7,10) and found evidence that additional variants contribute to the risk of prostate cancer. Although rs4430796 was the most significant SNP associated with risk, rs7405696 was also significant at a genome-wide significance level and explained part of the risk associated with the first HNF1B locus, suggesting a more complex genetic architecture for common variants in this region. Since this study used SNP markers, further work is needed to investigate the biological basis of the association with common variants in 17q12, which may regulate HNF1B, or perhaps another gene in prostate cancer. It is plausible that multiple variants are directly associated with prostate cancer susceptibility.

In the second HNF1B locus, we found that rs4794758 was more strongly associated with risk than the previously identified variant, rs11649743. When both variants were included in the same model, rs4794758 explained all of the risk associated with rs11649743, indicating that this variant more aptly captured the risk attributable to this locus. Although the first and second HNF1B loci are separated by recombination hotspot (10), the two loci are not completely independent and the risk associated with rs4794758 was attenuated after conditioning on the most significant SNP in the first locus, rs4430796.

In our study, the best model for prostate cancer risk in the HNF1B region included five SNPs (rs4430796, rs7405696, rs4794758, rs1016990 and rs3094509). Although three of these SNPs reached genome-wide significance in unconditional models, rs1016990 was only nominally associated with risk ($P = 0.0002$) and rs3094509 was not associated with risk ($P = 0.80$) in unconditional models. Together,
these five variants provided a better model for the data than other combinations, suggesting that these SNPs together capture more of the risk associated with this region. Furthermore, haplotype analysis revealed that risk was not explained by a single haplotype, suggesting a role for multiple variants or combinations of variants. Imputation using data from the 1000 Genomes Project yielded similar results; however, additional variants discovered in future next generation sequencing may also contribute to risk.

Interestingly, we observed a significant interaction between a SNP located in the second HNF1B locus, rs2107131, and family history of prostate cancer. Family history was only captured at baseline for participants and could have changed over time, and the number of subjects with a positive family history of prostate cancer was limited (N = 2182 men), so this finding should be interpreted with caution. However, it is possible that the interaction reflects haplotype segregation with uncommon disease-causing alleles associated with familial cases and common disease-causing alleles associated with sporadic cases.

The biological mechanism by which 17q12 variants may alter the regulation or splicing of a plausible candidate gene, such as HNF1B, is not clear. Down-regulation of HNF1B expression has been associated with renal cell cancer progression (24), and differential expression of HNF1B has been associated with prostate cancer recurrence (23). HNF1B is a transcription factor that encodes three isoforms in humans. Isoforms HNF1B(A) and HNF1B(B) appear to be transcriptional activators, whereas isoform HNF1B(C) is a transcriptional repressor (25). Differences in isoform-specific HNF1B expression have also been observed between normal and malignant prostate tissue, with prostate tumors displaying greater isoform HNF1B(B) expression but less isoform HNF1B(C) expression than normal prostate tissue (26). It is possible that variants in HNF1B contribute to the altered expression of HNF1B isoforms in prostate cancer. It is also plausible that variants in this region of 17q12 could also influence the regulation or expression of other genes at a distance.

In conclusion, this large-scale fine mapping study revealed that the association between variants in the 17q12 region and prostate cancer risk is more complex than earlier studies have indicated. Additional sequencing and functional studies are needed to pinpoint the variants in this region that are directly associated with prostate cancer risk and the biological mechanisms involved.

MATERIALS AND METHODS

Study subjects

As described previously (11), prostate cancer cases and controls of European ancestry were drawn from 10 studies in the USA and Europe: Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial (972 cases/927 controls); Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study (ATBC) (906 cases/868 controls); American Cancer Society Cancer Prevention Study II (CPSII) (1643 cases/1640 controls); Health Professionals Follow-up Study (HPFS) (595 cases/589 controls); CeRePP French Prostate Case-Control Study (FPCC) (998 cases/952 controls); Multiethnic Cohort Study (MEC) (676 cases/682 controls); European Prospective Investigation into Cancer and Nutrition (EPIC) (682 cases/990 controls); Cohort of Norway (CONOR) (606 cases/662 controls); Cancer of the Prostate in Sweden (CAPS) (2213 cases/1362 controls); and a hospital-based case–control from the Johns Hopkins Hospital (JHH) (990 cases/451 controls). A total of 10,272 prostate cancer cases and 9,123 controls were available for this study. Aggressive prostate cancer was defined at Gleason score ≥7 or stage C/D. Among the cases with information on disease aggressiveness, 4,824 cases were defined as aggressive and 4,337 were non-aggressive (Gleason score <7 and stage A/B). Family history of prostate cancer was defined by self-report from the participants and available for 8 of the 10 studies. A positive family history of prostate cancer was defined as a first degree relative with prostate cancer. Each study obtained an informed consent from study participants and approval from their respective institutional review boards for this study.

SNP selection

A total of 88 SNPs within a 249,071 bp region encompassing HNF1B were selected for fine mapping. The SNPs were chosen based on the 0.2 cM HapMap recombination region flanking the most significant SNP (rs4430796) in the HNF1B locus from the second stage of the CGEMS GWAS (7). The entire region was tagged to capture SNPs with a minor allele frequency ≥5% at a D’ > 0.6 based on the HapMap CEU population (Build 26). All three SNPs with a P < 10⁻⁸ from the second stage of CGEMS in this region (i.e. rs4430796, rs7501939 and rs11649743) were selected as obligated SNPs to be included. The final tag SNPs were selected if they were found to be correlated with an r² of ≥0.8 in the HapMap CEU, YRI or JPT + CHB populations with the obligated SNPs. The tag SNP selection was performed using the GLU software package (http://code.google.com/p/glugenetics/). The chosen SNPs were primarily located in first (chromosome 17: 33.161–33.205 Mb, NCBI Build 36) and second (chromosome 17: 33.116–33.161 Mb, NCBI Build 36) HNF1B regions identified to be associated with prostate cancer risk; however, other SNPs outside these regions were also included.

Genotyping

All SNPs for this study were genotyped on a custom Illumina iSelect assay panel as part of the third stage of CGEMS as described previously (11). In brief, a total of 6,652 SNPs, including 1,400 SNPs to monitor population stratification, were attempted in 22,057 samples, including quality control duplicates. Duplicate samples yielded a 99.97% concordance rate. Subjects with <90% completion (n = 1,350), missing covariate data or sparse group (n = 104), likely an intra-study duplicate based on genotype concordance ≥99% (n = 18), or non-European ancestry defined by <0.80 European admixture as estimated using STRUCTURE (27) (n = 372) were excluded, leaving 10,272 cases and 9,123 controls for analysis. SNPs with <80% completion within a study were removed from analysis for that study. SNPs that failed to provide genotypes for more than six studies (n = 1), had a minor allele...
count \( \leq 10 \) \((n = 7)\) or had genotypes that were inconsistent with Hardy–Weinberg proportions among controls \((P < 0.001)\) \( (n = 1)\) were excluded from analysis, leaving 79 SNPs for analysis.

**Imputation**

To explore the possibility that other variants in the region could be related to risk, we imputed the 249,071 bp region encompassing our genotyped SNPs using IMPUTE2 and the 1000 Genomes Project data (June 2010 release). A total of 653 SNPs were imputed from this region, but only SNPs with a quality score (i.e. info) \( > 0.40\) were considered for analysis \((N = 353)\). All imputed SNPs were analyzed using SNPTEST v.2.2.0 accounting for imputation uncertainty.

**Statistical analysis**

Principal components analysis was conducted using EIGENSTRAT \((28)\) with 1399 SNPs that were genotyped, passed quality control criteria and selected for population stratification; these SNPs were chosen because they had minimal correlation \((r^2 < 0.004)\) \((29)\). The Wilcoxon rank test was used to test the association between the top five eigenvectors and case–control status. Four eigenvectors displayed a significant or borderline significant association with prostate cancer \((P < 0.08)\) and were included in the analysis. The association between each SNP and prostate cancer risk was estimated using logistic regression adjusting for age \((<50, 50–59, 60–69, 70–79, \geq 80)\), study (including country for EPIC) and significant principal components. Stratified analyses were conducted to examine differences by disease aggressiveness, family history and study. Heterogeneity between aggressive and non-aggressive disease was assessed in a case-only analysis using logistic regression. Heterogeneity between studies and risk modification by family history was assessed using a likelihood ratio test comparing the model with and without the cross-product term(s). \(P\)-values for interactions tests were adjusted for the false discovery rate \((30)\).

To explore the interdependence of the associations observed, three separate approaches were taken: (i) sequential conditional analyses, (ii) stepwise regression, and (iii) lasso \((31)\). Sequential conditional analyses were conducted by including the most significant SNP in the unconditional logistic regression model followed by sequential inclusion of the most significant SNP in each conditional model and examining the association with each of the remaining SNPs independently. Forward stepwise regression was performed using the SNPs that reached a nominal significance level in the unconditional model. SNPs were sequentially included in the model based on a minimal \( P < 0.05\). Lasso was conducted in R using all SNPs without missing data and an unweighted penalty function.

Linkage disequilibrium measures \((D'\) and \(r^2)\) were estimated in the controls using Haploview. Haplotypes were estimated using an expectation-maximization algorithm and analyzed using the generalized linear regression model implemented in HaploStats \((32)\). Unless otherwise indicated, all analyses were conducted using PLINK or STATA.

**SUPPLEMENTARY MATERIAL**

Supplementary Material is available at HMG online.

**ACKNOWLEDGEMENTS**

This study was supported by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health (NIH). The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services nor does mention of trade names, commercial products or organization indicate endorsement by the US Government. The authors thank Drs Christine Berg and Philip Prorok, Division of Cancer Prevention, NCI, the screening center investigators and staff of the PLCO Cancer Screening Trial, Mr Thomas Riley and staff at Information Management Services, Inc., and Ms Barbara O’Brien and staff at Westat, Inc. for their contributions to the PLCO. Finally, we are grateful to the study participants for donating their time and making this study possible.

**Conflict of Interest statement.** None declared.

**FUNDING**

This study was supported by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health and in part by National Cancer Institute, National Institutes of Health (CA129684 to J.X.) and contract number HHSN261200800001E.

**REFERENCES**


