

Family-based association analysis of 42 hereditary prostate cancer families identifies the *Apolipoprotein L3* region on chromosome 22q12 as a risk locus

Bo Johanneson^{1,†}, Shannon K. McDonnell^{2,†}, Danielle M. Karyadi^{1,†}, Pascale Quignon¹, Laura McIntosh³, Shaun M. Riska², Liesel M. FitzGerald³, Gregory Johnson¹, Kerry Deutsch⁴, Gabrielle Williams¹, Lori S. Tillmans^{6,7}, Janet L. Stanford^{3,5}, Daniel J. Schaid², Stephen N. Thibodeau^{6,7} and Elaine A. Ostrander^{1,*}

¹National Human Genome Research Institute, National Institutes of Health, 50 South Drive, Building 50, Room 5351, Bethesda, MD 20892, USA, ²Department of Health Sciences Research, Mayo Clinic, Rochester, MN 55905, USA, ³Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, M4-B874, PO Box 19024, Seattle, WA 98109-1024, USA, ⁴Institute for Systems Biology, 1441 N 34th Street, Seattle, WA 98103-8904, USA, ⁵Department of Epidemiology, School of Public Health, University of Washington, PO Box 357236, Seattle, WA 98195, USA and ⁶Department of Laboratory Medicine and ⁷Department of Pathology, Mayo Clinic College of Medicine, 200 First Street Southwest, Rochester, MN 55905, USA

Received March 26, 2010; Revised and Accepted July 2, 2010

Multiple genome-wide scans for hereditary prostate cancer (HPC) have identified susceptibility loci on nearly every chromosome. However, few results have been replicated with statistical significance. One exception is chromosome 22q, for which five independent linkage studies yielded strong evidence for a susceptibility locus in HPC families. Previously, we refined this region to a 2.53 Mb interval, using recombination mapping in 42 linked pedigrees. We now refine this locus to a 15 kb interval, spanning *Apolipoprotein L3* (*APOL3*), using family-based association analyses of 150 total prostate cancer (PC) cases from two independent family collections with 506 unrelated population controls. Analysis of the two independent sets of PC cases highlighted single nucleotide polymorphisms (SNPs) within the *APOL3* locus showing the strongest associations with HPC risk, with the most robust results observed when all 150 cases were combined. Analysis of 15 tagSNPs across the 5' end of the locus identified six SNPs with P -values $\leq 2 \times 10^{-4}$. The two independent sets of HPC cases highlight the same 15 kb interval at the 5' end of the *APOL3* gene and provide strong evidence that SNPs within this 15 kb interval, or in strong linkage disequilibrium with it, contribute to HPC risk. Further analyses of this locus in an independent population-based, case-control study revealed an association between an SNP within the *APOL3* locus and PC risk, which was not confirmed in the Cancer Genetic Markers of Susceptibility data set. This study further characterizes the 22q locus in HPC risk and suggests that the role of this region in sporadic PC warrants additional studies.

INTRODUCTION

Prostate cancer (PC) is the most frequently diagnosed solid tumor among men in the US today, with an estimated 192 280 cases diagnosed in 2009 and ~27 360 deaths (1).

The disorder is classically divided into sporadic and hereditary forms, although clinically the two are virtually indistinguishable aside from the earlier average age at diagnosis of men with the inherited form. Sporadic cases probably develop

*To whom correspondence should be addressed. Tel: +1 3015945284; Fax: +1 3014800472; Email: eostrand@mail.nih.gov

[†]These authors contributed equally to this work.

because of accumulated somatic mutations in critical dividing cells, although genetic factors that initiate that process remain largely unknown. Hereditary prostate cancer (HPC) is believed to originate with one or more germline mutations that accelerate the oncogenic process, giving carriers an increased risk of disease.

Numerous genome-wide linkage studies of HPC families have been undertaken in an attempt to identify PC susceptibility loci, leading to the reporting of a large number of putative loci. However, the existence and significance for most of these is debatable (2–5). This is not surprising as multiple loci are believed to contribute to PC susceptibility and there are few, if any, criteria to distinguish genuinely hereditary cases from phenocopies, even within individual families. In addition, it is clear that the penetrance of disease alleles in the population is variable, making it difficult to accurately assign affection status, thus diluting any real signal in either genome-wide linkage or association studies.

One strategy for enhancing the utility of family-based data sets in confirmation and fine-mapping studies is to stratify pedigrees into comparatively more homogenous samples based on clinical or tumor characteristics. An excellent example of this is the chromosome 22q12.3 locus. By stratifying pedigrees based on number of affected men, aggressive disease and median age at diagnosis, we and others have independently identified and confirmed the existence of an HPC risk locus on chromosome 22q12.3 (6–11). The International Consortium for Prostate Cancer Genetics (ICPCG) has similarly confirmed these results in a combined data set of 269 HPC pedigrees, each with five or more PC cases, demonstrating an heterogeneity log of odds of 3.57 at 22q12.3 (12). This makes it the only region in this ICPCG analysis to be significantly linked with HPC. However, as with all linkage peaks, the 1-log of the odd (LOD)-support interval in the ICPCG study was large, spanning over 6 Mb and nearly 100 genes.

We and others have worked to further characterize the 22q12.3 locus. Camp *et al.* (13,14) in an analysis of 59 large Utah pedigrees, and subsequently, in a data set of 54 putatively linked families from the ICPCG, were able to reduce the interval to 2.18 Mb. In our own previously published study, 42 high-risk families from the *PROGRESS* and Mayo Clinic studies with evidence of linkage to 22q12.3 were used for recombination mapping using a large set of markers spanning the entire interval of interest (15). While no overlapping consensus interval could be detected for all families, an 8.7 Mb interval (26.00–34.74 Mb) defined by three recombination events on each side was identified in 35 families. A smaller consensus interval of 2.54 Mb (33.48–36.01 Mb) was identified in 12 of the 14 families, all of whom had five or more affected men. Our fine-mapping data, combined with that of Camp *et al.* (13,14), highlight a minimal shared consensus interval of ~1.36 Mb that spans only 16 genes (15).

Since the linkage information in the Mayo Clinic and *PROGRESS* family data sets was fully explored in our previous analyses, in this study we have elected to take advantage of family-based association methods in an effort to further refine the shared consensus interval. For these analyses, we utilized only cases from families with strong evidence of linkage to 22q12.3 and compared that to a set of 506 unrelated population controls. We hypothesized that our restricted selec-

tion of cases from linked families would reduce the misclassification arising from phenocopies as well as unaffected risk-variant carriers. In doing so, we identify a 15 kb interval spanning two linkage disequilibrium (LD) blocks within the *Apolipoprotein L3 (APOL3)* locus that appears to contain the risk variant. Analyses of an independent, population-based, case–control data set support this conclusion.

RESULTS

In this study, the 42 Caucasian pedigrees used in our previous recombination mapping study of 22q12.3 (15) were evaluated further using family-based association methods. The 42 pedigrees, 18 from *PROGRESS* and 24 from the Mayo Clinic, had each achieved a pedigree-LOD score of ≥ 0.58 within the ICPCG-defined LOD-1 support interval. The data set included a total of 213 men with PC, of whom 150 had DNA available for this study (Table 1). The average age at diagnosis of men in the study is 66.1 (SD = 5.5), with 21 pedigrees having an average age of diagnosis <66 years (Table 1), and 24 having five or more affected family members.

For this study, 168 tagSNPs were initially chosen from a larger data set of 668 chromosome 22 tagSNPs that had been selected to cover most of the genes on 22q, and analyzed in a case–control study performed by the Mayo Clinic (results not shown, see Materials and Methods). The 168 tagSNPs represent all markers that both showed nominal evidence of association with PC ($P < 0.05$) in the initial case–control study from the Mayo Clinic, and were located within the maximum shared consensus interval indicated by our previous recombination mapping study (26.00–36.01 Mb). From the data set of 168 tagSNPs, 145 were successfully genotyped in the 18 *PROGRESS* families and used as the initial screening set.

Family-based association testing was performed using Ped-Genie (16). Analysis was initially performed separately on cases from the 18 *PROGRESS* and 24 Mayo Clinic families. However, in each situation, data were compared with the same set of 506 unrelated Caucasian population controls genotyped by the Mayo Clinic. Analysis of 84 cases from the Mayo Clinic pedigrees versus population controls revealed two adjacent SNPs, rs2097465 and rs132656, each of which show compelling evidence of association ($P < 3.3 \times 10^{-4}$), based on 3000 simulations for all models except the dominant model (Fig. 1A). Both markers are located within the *APOL3* locus, separated by a distance of 5.8 kb ($D' = 0.706$, $r^2 = 0.354$). When these two markers were further evaluated using 200 000 simulations, rs2097465 yielded a P -value of 1×10^{-5} for all models except the dominant model. The risk-allele in the Mayo Clinic families appeared to be the T-allele, with a frequency of 0.30 in the controls. For rs132656, the strongest P -value in the Mayo Clinic families was observed under the recessive model ($P = 2 \times 10^{-4}$) with the risk-allele, C, having a frequency of 0.45 in the controls. The strongest association in the 66 *PROGRESS* cases compared with Mayo Clinic controls was also observed with SNP rs2097465 (Fig. 1B; $P = 0.0013$, additive trend model). When 3000 simulation analyses were performed on the 145 SNPs in the combined set of 150 Mayo Clinic and *PROGRESS*

Table 1. Characteristics of PROGRESS and Mayo Clinic HPC pedigrees and population controls

Pedigree characteristics	All pedigrees		PROGRESS pedigrees		Mayo pedigrees ^a		Controls ^b	
	# Peds	No. Affected	Avg. age at Dx (SD)	Avg. no. affected (range)	# Peds	No. affected	Avg. age at Dx (SD)	Avg. age at study (SD)
All families	42	213	66.1 (5.5)	4.8 (3-10)	24	126	66.8 (5.4)	65.7 (9.0)
All Typed	150	150	64.9 (5.7)	3.7 (3-7)	84	84	66.2 (5.7)	
Avg. age at Dx								
<66	21	100	62.0 (4.5)	4.8 (3-10)	9	42	62.0 (5.9)	
All Typed	73	73	61.6 (4.2)	3.8 (3-7)	27	27	61.3 (5.8)	
≥66	21	113	70.2 (2.4)	4.8 (4-8)	15	84	69.7 (2.2)	
All typed	77	77	68.3 (5.0)	3.3 (3-4)	57	57	69.2 (3.0)	
No. affected								
≤4	18	65	65.1 (6.8)	3.7 (3-4)	8	28	64.2 (8.0)	
All typed	56	56	64.7 (6.1)	3.2 (3-4)	24	24	64.3 (8.0)	
≥5	24	148	66.8 (4.2)	6.3 (5-10)	16	98	68.1 (3.0)	
All typed	94	94	65.1 (5.5)	4.3 (3-7)	60	60	67.2 (4.1)	

^aThe number of Mayo pedigrees reported is based on the total number of affected men in each pedigree compared with the number of affected men who are informative for linkage as reported previously (15).
^b506 unrelated population controls from the Mayo Clinic.

cases compared with the 506 Mayo controls, the associations became even stronger for both SNPs (Fig. 1C). Based on 200 000 simulations, rs2097465 yielded a P -value of $<5 \times 10^{-6}$ for all models except the dominant model, and rs132656 gave a P -value of 5×10^{-5} with a recessive model.

Encouraged by the overlapping signals at the *APOL3* locus, we tested 15 HapMap tagSNPs that spanned a 28 kb interval, inclusive of both rs2097465 and rs132656, in order to determine which LD-block(s) yielded the strongest association for HPC risk. The interval included three LD blocks that span the 5' end of the *APOL3* gene (Fig. 2D). Block one is defined by rs2017329 (34 879 112) and rs2105915 (34 882 285) and is ~3.2 kb in size. Block 2 is defined by rs132648 (34 882 348) and rs132665 (34 894 116) and is ~11.8 kb in size. Block 3 is defined by rs132671 (34 899 790) and rs11089782 (34 907 299) and is 7.5 kb in size. In the combined set of 150 cases and 497 controls (9 of the initial 506 controls were dropped in this analysis due to lack of DNA), 6 SNPs yielded P -values $\leq 2 \times 10^{-4}$ based on 200 000 simulations (rs2017329, rs132647, rs2097465, rs132649, rs132654 and rs132656; Fig. 2C). These six SNPs are all located within Blocks 1 and 2 (Fig. 2D). Both blocks are in strong LD with one another as evidenced by a multiallelic D' of 0.92. The strongest association with HPC risk was still with rs2097465 ($P < 1 \times 10^{-6}$ based on 1 000 000 simulations), which is within the 3.2 kb LD Block 1. No associations were detected for any of the five SNPs located in the telomeric 7.5 kb LD Block 3 (Fig. 2D).

We also analyzed our data using the LAMP program (17,18), which tests for association in the presence of linkage, and also tests whether any of the associated SNPs explain all or part of the original linkage signal. All SNPs were analyzed in the combined set of 42 PROGRESS and Mayo Clinic pedigrees and all 506 unrelated Mayo population controls. Consistent with the PedGenie results, SNPs within the *APOL3* locus were associated with HPC risk in the LAMP analyses. Two SNPs had P -values $< 1 \times 10^{-4}$; rs2017329 and rs2097465 ($P = 7.5 \times 10^{-5}$ and 6.7×10^{-5} , respectively). Both of these SNPs tag Block 1 at the 5' end of *APOL3* (Fig. 2D). As expected for HPC, neither SNP explains the entire linkage signal with a reduction in LOD score > 1 LOD unit for each SNP, assuming complete LD. The fact that cases selected and analyzed separately from the PROGRESS and Mayo Clinic family collections each highlight SNPs within a 15 kb interval at the *APOL3* locus as being associated with risk is compelling evidence that these SNPs, or those in LD with them, contributed to the initial 22q12.3 linkage peak.

To screen for causal mutations in *APOL3*, we sequenced a 28.9 kb region that included all exons and introns of the most common isoforms of *APOL3* (19) in both affected and unaffected individuals from the 18 PROGRESS pedigrees (Fig. 2A). We also sequenced a 14.3 kb upstream interval that included an alternative promoter (19) and a 7.4 kb conserved region located between the *APOL3* and *APOLA4* genes. In total, 42.1 kb (97.6%) of the targeted region was successfully sequenced (Fig. 2B). Analysis of the data revealed 235 variants, of which 219 were SNPs and 16 indels. To identify potential functional variants, an *in silico* analysis was performed and highlighted one SNP, rs132660, which is located

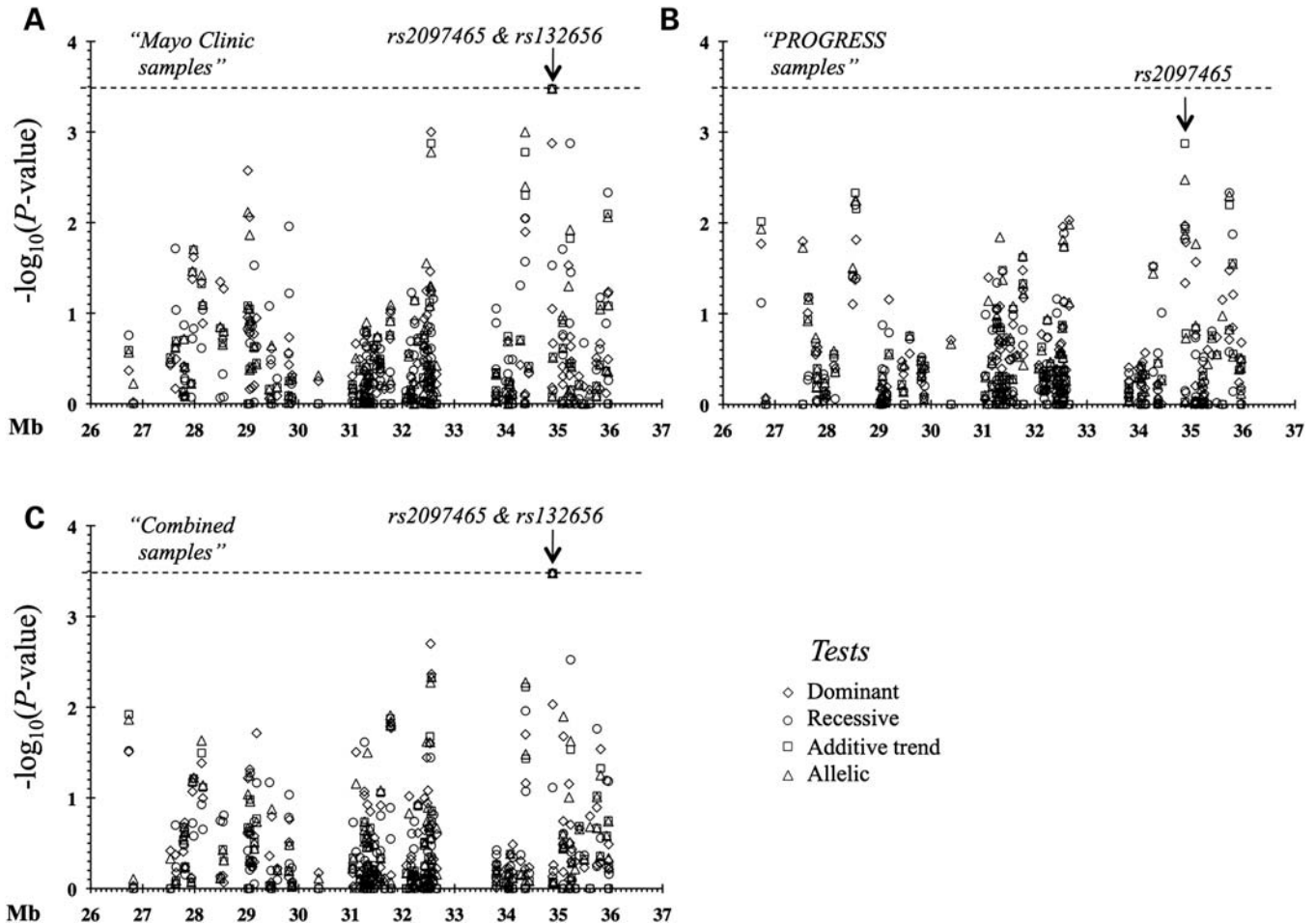


Figure 1. Results from a family-based association analysis of 145 SNPs at 22q12.3 (26–36 Mb) using PedGenie. (A) Association analysis of 84 cases from 24 Mayo Clinic pedigrees and 506 unrelated population controls. (B) Association analysis of 66 cases from 18 *PROGRESS* pedigrees and the 506 controls. (C) Combined association analysis of 145 SNPs on 150 cases from the 42 combined pedigrees and the 506 controls. Dashed line at $-\log_{10}(p) = 3.5$, indicates the maximum $-\log_{10}(p)$ given 3000 simulations ($P < 3.3 \times 10^{-4}$). SNPs giving the strongest association in both sets of pedigrees are indicated with arrows. The dominant test is indicated by diamonds, recessive test by circles, additive trend test by squares and allelic test by triangles.

within the core motif of a putative TATAA-box, 11 bp upstream of Exon 1a in the alternative promoter (19). Since rs132660 was the only variant with a potential effect on function, we genotyped this SNP in the Mayo Clinic families and controls and analyzed the data with PedGenie in the combined set of cases and controls. The resulting association for rs132660 was $P = 5 \times 10^{-6}$ based on 200 000 simulations (Fig. 2C; risk-allele frequency: C = 0.43 in controls). This is the second strongest result behind rs2097465 in the PedGenie analyses. Not surprisingly, in the Mayo controls, SNPs rs2097465 and rs132660 are in LD with one another ($D' = 0.72$, $r^2 = 0.32$).

To evaluate the potential association of rs132660 and rs2097465 with sporadic prostate cancer risk, both SNPs were genotyped in a combined data set (1320 cases and 1266 controls) of Caucasian men from one of two population-based, case–control studies of prostate cancer conducted in Western Washington (20,21). Genotype distributions for both rs132660 and rs2097465 were consistent with Hardy–Weinberg equilibrium in the control population. Of the two SNPs, only the

minor C-allele of rs132660 was found to be significantly associated with PC risk ($p_{\text{trend}} = 0.015$ for allele dosage; Table 2). Further analyses indicated that the risk estimate for rs132660 did not differ substantially by family history of prostate cancer, Gleason score (≤ 7 , 3 + 4 versus ≥ 7 , 4 + 3) or a composite measure of aggressive disease based on Gleason score, stage and diagnostic PSA level.

We also tested both markers in the Cancer Genetic Markers of Susceptibility (CGEMS) prostate cancer study (22). Raw genotype data for SNP rs2097465 was available for 1176 cases (688 aggressive cases and 488 nonaggressive cases) and 1105 controls. Marker rs132660 was not directly genotyped in CGEMS, and thus had to be imputed using HapMap data (see Materials and Methods). The imputation quality for marker rs132660 was very good ($R^2 = 0.9688$). However, neither of the two SNPs showed an association with prostate cancer risk in the CGEMS data set. The OR for rs132660 was 1.06 when all cases were considered (95% CI 0.96, 1.20), and was 1.04 for all cases (95% C.I. 0.92, 1.17) for rs2097465. The results did not change appreciably

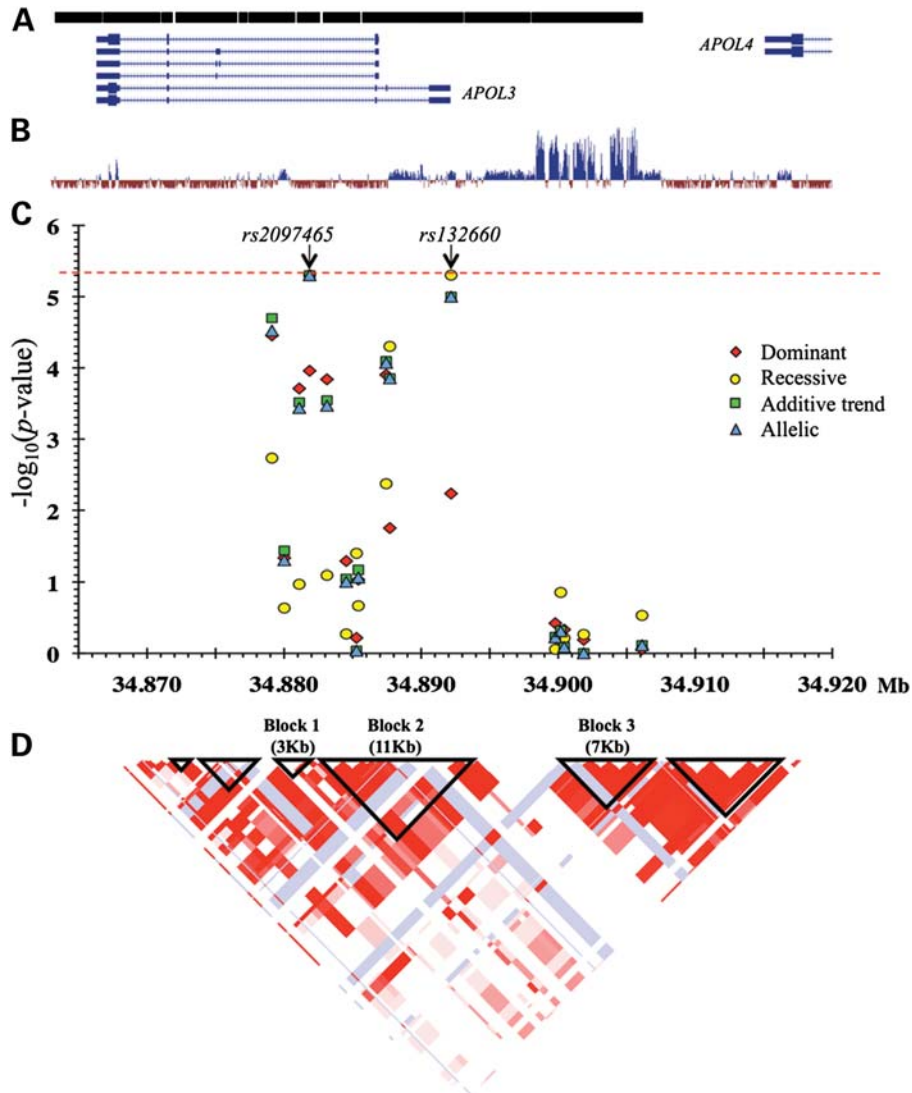


Figure 2. Summary of associations between SNPs in *APOL3* and HPC. (A) Position of exons in *APOL3* and *APOLA* defined by the most common transcripts. Black bar indicates area sequenced. (B) Position of a 7 kb evolutionary conserved region upstream of *APOL3* (based on 44 vertebrate species in UCSC ‘conservation track’). (C) Results from combined PedGenie family-based association analysis of 15 HapMap tagSNPs, using 150 cases from the 42 combined pedigrees and 497 unrelated population controls. Dashed line at $-\log_{10}(p) = 5.3$, indicates the maximum $-\log_{10}(p)$ given 200 000 simulations ($P < 5 \times 10^{-6}$). Dominant test is indicated by red diamonds, recessive test by yellow circles, additive trend test by green squares and allelic test by blue triangles. Arrows indicate SNP rs2097465 which has the strongest association, and rs132660 which was analyzed separately and is located in a putative TATAA-box. (D) LD block structure over the region calculated from Caucasian population HapMap data. Reverse triangles indicate localization of LD blocks using solid spine of LD method in HaploView. The three LD blocks tagged by the 15 tagSNPs are indicated by their size in kilobases.

when the data were analyzed by aggression status (aggressive and nonaggressive), versus controls. However, as with the Western Washington data set, we note that the minor C allele (for rs132660) is over-represented in cases versus controls (C allele frequency = 45.3% in cases and 43.8% in controls in CGEMS, and 45.8% in cases and 42.4% in controls in the Western Washington data set).

DISCUSSION

We have identified a 15 kb interval spanning the region at 22q12.3 from 34 879 112 to 34 894 116 bp (March 2006, NCBI build 36/hg18) at the 5' end of the *APOL3* gene that

accounts, at least in part, for the linkage signal that we and others (6–8,10,11,14) have reported in HPC families at 22q12.3.

A meta-analysis conducted by the ICPCG resolved the locus to a 12 cM interval (12), which was subsequently reduced to 2.18 Mb (13). In our previous work (15), we performed a recombination analysis in a set of 42 HPC pedigrees, 18 from the *PROGRESS* data set and 24 from the Mayo Clinic that showed evidence of linkage to the 1-LOD support interval defined by the ICPCG at 22q12.3. Our work (15), combined with that of Camp *et al.* (13,14), defined two distinct but overlapping consensus intervals that were shared by the majority of the pedigrees. The goal of the present study was to perform family-based association analyses within the consensus interval

Table 2. Association results for rs132660 and rs2097465 and prostate cancer risk in Caucasians

SNP	Position	Cases (%) ^a (n = 1320)	Controls (%) ^a (n = 1266)	Genotype	OR	95% CI	P-value ^b
rs132660	34 892 182	343 (27.7)	399 (33.1)	AA	ref		
		653 (52.8)	592 (49.0)	AC	1.28	1.07–1.54	0.015
		240 (19.4)	216 (17.9)	CC	1.29	1.02–1.63	
				AA versus AC + CC	1.28	1.08–1.53	
rs2097465	34 881 862	561 (43.4)	568 (45.7)	CC	ref		
		583 (45.1)	555 (44.6)	CT	1.05	0.90–1.25	0.16
		148 (11.5)	121 (9.7)	TT	1.22	0.94–1.60	
				CC versus CT + TT	1.08	0.93–1.27	

^aNumber of cases and controls varies due to missing genotype information.

^bP-value for trend.

using our 42 families, to further characterize the locus at 22q12.3.

HPC susceptibility studies are often hindered by disease heterogeneity, locus heterogeneity and risk alleles with weak to moderate penetrance. Since it is impossible to clinically distinguish true hereditary cases from sporadic cases or true controls from unaffected risk allele carriers, we designed a family-based association analysis that we hypothesized would reduce the impact of these sources of misclassification. We anticipated several potential challenges including the likely existence of unaffected risk allele carriers. This includes men who, although they carry the risk alleles, are too young to have been diagnosed with PC. Finally, some men will be phenocopies, i.e. their PC is not due to mutations at 22q12.3.

To address the high phenocopy rate, we utilized only the 42 previously described HPC families that showed strong evidence for linkage (pedigree LOD > 0.58) to 22q12.3 in our previous recombination mapping study (15). The presence of phenocopies in this restricted data set is expected to be considerably less than if cases from all HPC pedigrees available had been used. To reduce misclassification introduced by unaffected risk allele carriers, we utilized a set of unrelated population controls rather than internal family-based controls. While these two strategies might raise concerns about population stratification, the overall approach should increase the power of the analysis.

Since only linked families were analyzed, we expect that the association evidence from PedGenie as well as the LAMP LOD scores will be inflated and, as such, these approaches are not appropriate for reporting the initial findings of association or linkage. They are useful, however, for prioritizing regions or LD blocks with the strongest HPC risk association and highlighting the best areas within an already described linkage peak to pursue screening for causal variant(s), as we have done here.

The 168 SNPs selected for this study were from a set of 668 tagSNPs spanning the q arm of chromosome 22 that had been used by Mayo investigators for a previously unpublished case-control study of PC risk. All 168 SNPs reached a nominal P-value of <0.05 in the previous study and are located within the maximum shared consensus interval. Our initial family-based association analysis identified rs2097465, which is within the *APOL3* locus, as having the strongest association with HPC risk when the *PROGRESS* and Mayo Clinic families were analyzed independently, com-

paring to the same set of Mayo Clinic controls. The association at rs2097465 became stronger when the *PROGRESS* and Mayo Clinic data sets were combined.

In the analysis of tagSNPs across three LD blocks at this locus (Fig. 2D), six SNPs, tagging Blocks 1 and 2 and covering a 15 kb interval within the *APOL3* region, were associated with HPC risk. An association was also observed in the LAMP analysis, in which two SNPs, rs2097365 and rs2017329, demonstrated the strongest associations with prostate cancer risk. Both SNPs are located within Block 1. One of the two SNPs, rs2097465, showed the strongest association in the PedGenie analysis. Block 3, which is telomeric to Blocks 1 and 2, is unlikely to be relevant, as five tagSNPs within Block 3 demonstrated no associations with HPC risk.

Two additional LD blocks, located between 34 871 671 and 34 878 352, and centromeric to Block 1, are in strong LD with both Blocks 1 and 2. The multiallelic *D'* between Block 1 and the two additional blocks is 1.0 and 0.9, respectively. Inclusion of these two additional blocks would increase the region to 22.4 kb and the resulting interval would include both *APOL3* promoters, the two alternative exons (1a and 1b) and exons 1, 2 and 3, ending just 78 bp before exon 4 within the *APOL3* gene. Since the centromeric side of the associated interval is not well defined, the entire 22.4 kb interval must be considered as potentially carrying a causal variant(s).

As we would predict, none of the tagSNPs tested in the LAMP analysis fully explained the initial linkage signal, as PC is a complex and multifactorial disease. This suggests two likely possibilities. Either there are other independently associated risk variants within the same region, or the associated SNPs are in strong LD with, but are not themselves, the causal variant. Since it is highly unlikely that all linked families have precisely the same risk variant, we hypothesize that both possibilities may be true.

Sequencing of the *APOL3* gene, including introns, exons, known upstream and downstream promoters and regulatory elements, did not reveal any obvious disease-associated variants, i.e. rare in the general population and segregating with PC in the linked families. However, the SNP rs132660, located inside a TATAA-box sequence in an alternative promoter region upstream of exon 1a, was of potential functional interest. This A/C SNP is at the third A-nucleotide that makes up the core motif (5'-TATAAA-3'). Multiple studies have shown that an A to C substitution at this position is likely to completely inhibit the binding of a TATAA-binding protein

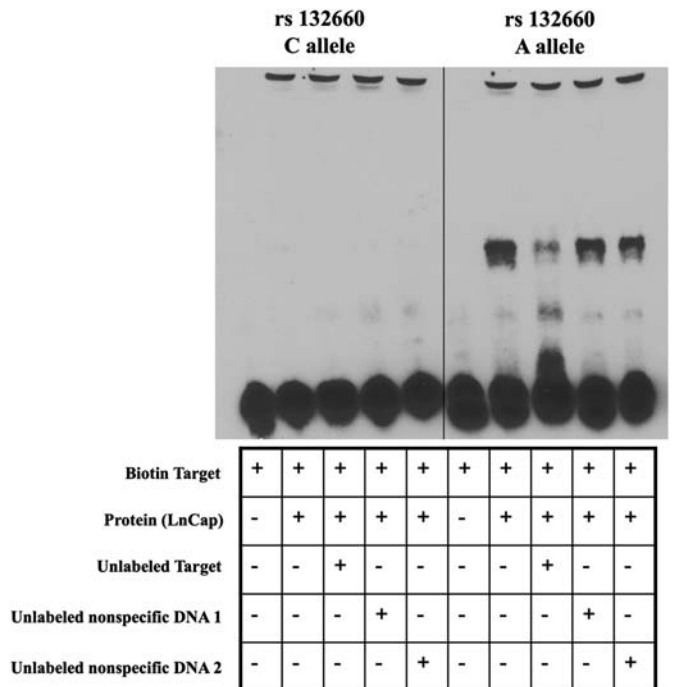


Figure 3. Band shift assay for the TATAA box. LNCaP nuclear extract was incubated with biotin-labeled oligonucleotides specific for each allele (A and C) of rs132660. A gel shift was detected for the A-allele specific oligonucleotide (lane 7) but not for the C-allele (lane 2). The A-allele specific complex was disrupted with an excess of unlabeled specific oligonucleotide (lane 8) but not by an excess of two different non-specific oligonucleotides (lanes 9 and 10).

(23). Although we do not know whether this variant in the TATAA box is functional, a band shift assay clearly indicates that the A allele, not the C allele, binds the protein (Fig. 3).

A TATAA box's optimal range from the transcription start site (TSS) is normally between -29 and -32 bp, and is generally nonfunctional if located closer to the TSS (24). However, it is possible that this TATAA box regulates a previously unknown transcript that uses a TSS further downstream. More work is required to find out whether this putative TATAA box is regulatory and whether the associated SNP is functional.

Although the involvement of the *APOL3* gene itself in PC has yet to be established, it clearly is an interesting candidate gene. *APOL3* is one of the six apolipoprotein-L gene family members (25,26) that are located in a 6219 kb interval at 22q12.3 (27). The encoded proteins are thought to be involved in lipid transport and metabolism (25). In addition, they appear to be an important link between programmed cell death of host cells (28,29) and host immunity to various pathogens (26), and play a role in inflammatory response (30,31).

None of the major genome-wide association studies reported to date yield a result in precisely this region, but other parts of chromosome 22 have been implicated. Specifically, Eeles *et al.* (32) reported an association with marker, rs5759167 at 41 830 156 (6.9 Mb telomeric of rs2097465). In addition, Sun *et al.* (33) report a marker, rs9623117 at 38 782 065 (3.9 Mb telomeric of rs2097465).

In considering our results it is worth noting that while the 22.4 kb interval at the 5' end of *APOL3* represents the most likely interval where one of the causative risk variant(s) is located, it is not known whether any of the already identified variants within the region contributes specifically to HPC susceptibility. Also, it is possible that the HPC risk alleles identified here may not involve the *APOL3* gene itself, rather, the associated interval may include key regulatory elements that affect either local genes, or one or more genes located some distance away. Additional studies are needed to address this issue and to determine the precise role of the implicated variants.

Because we saw over-representation of the risk allele in cases in both the Western Washington and CGEMS data, it is possible that the 22q locus may play a role in sporadic prostate cancer risk. However, the CGEMS data were not significant and additional studies are therefore needed to determine the overall importance of this locus with regard to prostate cancer risk.

The work presented here provides a new paradigm for overcoming some of the common problems associated with reducing megabase-sized chromosomal segments discovered in linkage analysis of complex traits to kilobase-sized intervals, suitable for mutation scanning. While functional studies will ultimately be needed to illuminate the mechanism by which specific risk-associated variants act, the work reported here delineates a critical 22.4 kb region of association. We thus demonstrate that family-based association methods, when applied to selected families showing preliminary evidence of linkage, are useful mechanisms for reducing a region of linkage by orders of magnitude.

MATERIALS AND METHODS

HPC pedigrees and unrelated population controls for family-based association analysis

The results reported here include 254 of the extended HPC pedigrees from the Seattle-based *PROGRESS* study, which include 929 sampled affected men and 1176 relatives. Families met at least one of the following criteria for inclusion: ≥ 3 affected first-degree relatives, PC in three successive generations or two affected with a mean age at diagnosis of < 65 years or who were African-American (7). Pedigrees from the Mayo Clinic include 189 Caucasian HPC families with 498 affected men (6). Mayo Clinic HPC families were required to have at least three men with PC in the family, of whom two or more were alive for recruitment. The unrelated population controls were part of an ongoing study at Mayo Clinic of men from Olmsted County, MN, sampled using a scheme provided by the Rochester Epidemiology Project (34). The disease status of controls was last updated in 2008. Samples from 506 Caucasian controls were available for this study. *PROGRESS* study forms and protocols were approved by the Institutional Review Board (IRB) of the Fred Hutchinson Cancer Research Center. Additionally, genotyping protocols were approved by the IRB of the National Human Genome Research Institute. Mayo Clinic study materials and protocols were approved by the Mayo Clinic Human Subjects IRB.

Methods to select the 42 HPC pedigrees have been described previously (15). In brief, our analysis focused only on HPC families showing the greatest evidence for linkage within the ICPCG 6 Mb 1-LOD support interval defined by the ICPCG (12). To accomplish this, 443 *PROGRESS* and Mayo Clinic HPC pedigrees were reanalyzed using marker sets previously genotyped in earlier linkage scans. Only pedigrees with individual family-based LOD scores >0.58 were eligible for the fine-mapping effort (15). In addition, we required that all affected men within each family selected for fine-mapping shared a chromosomal segment within the ICPCG 1-LOD interval. After applying these strict criteria, 42 Caucasian pedigrees (24 Mayo Clinic and 18 *PROGRESS*) were selected for fine-mapping.

Family-based association analysis, marker selection and genotyping

Previously, a panel of 738 tagSNPs were selected for genotyping in all Mayo Clinic families and controls (15). The interval covers 216 genes located in a 19 cM region on chromosome 22q that broadly spanned the region of interest. The tagSNPs were selected from the HapMap Consortium (v. 2, October 2005) and Perlegen Sciences using the algorithm implemented in ldSelect. TagSNPs were identified such that each tagSNP exceeded an r^2 of ≥ 0.8 threshold with all other SNPs in the bin. Using a minimum LD coverage threshold of 70%, we were able to successfully identify tagSNPs for 183 (85%) of 216 genes with an average coverage of 87%. Details were provided previously (15).

The initial genotyping was done on 498 familial cases from 178 families and 533 population controls from the Mayo Clinic at the Center for Inherited Disease Research (CIDR) using the Illumina Platform. Of the 738 tagSNPs, 680 were successfully genotyped and 668 passed quality control (QC; seven monomorphic SNPs and five SNPs with call rate <0.90 were excluded). Of those, 168 had uncorrected P -values of <0.05 when evaluated for an association with risk of PC, and were located within the 22q recombination consensus interval from 26.00 to 36.01 Mb that we defined previously (15).

Since genotyping was complete on the Mayo Clinic data set, genotyping was then done on the 18 *PROGRESS* families using the multiplex MassArray spectrometry (iPLEX) genotyping system (Sequenom, San Diego, CA, USA). All PCR and iPLEX reactions were performed using standard conditions (35). Genotypes were called using the iPLEX MassArray Typer v3.4 software (Sequenom). To ensure genotyping accuracy, five blind duplicates were included in the data set, generating 1061 duplicated genotypes, only two of which were discrepant, for an error rate of $<0.2\%$. Markers with a call rate of $<75\%$ were excluded. Minor allele frequency (MAF) was compared between the sets and the HapMap Caucasian population. No major differences could be found, except for two SNPs, which were mono-allelic in one set, and were thus excluded. Thus, 145 of the 168 SNPs derived from the Mayo Clinic data set passed QC in the *PROGRESS* data set. The genetic position for all markers was determined using the UCSC Genome Browser (March 2006, NCBI Build 36.1). This set of markers was then analyzed in both the Mayo and *PROGRESS* data sets.

Family-based association analyses

Single-marker association tests were performed using PedGeniev. 2.4.2 (16). Four genetic models were tested (dominant, recessive, additive-trend and allelic tests). One advantage of PedGenie is that it does not trim extended pedigrees and permits any combination of pedigrees and/or cases and controls. The ability to use population controls makes PedGenie preferable for our data set over other analysis methods since the Mayo Clinic pedigrees have no unaffected relatives and few pedigrees with parental genotypes. PedGenie employs a Markov–Chain Monte–Carlo permutation test to correct for relatedness between individuals and calculates empirical P -values. Simulations are used to build multiple null genotypic configurations with a test statistic calculated for each. This is done by using Mendelian ‘gene dropping’ on founder individuals from the original pedigree structure, based on allele frequencies from the unrelated population controls. The permutation P -value is equal to the percentage of times a simulated test statistic is more extreme than the observed statistic. In the first round of analysis, the 145 SNPs that passed QC in the *PROGRESS* and Mayo Clinic data sets were analyzed. In the second round of analysis, the 15 HapMap tagSNPs in the region of greatest association (see below) were analyzed.

While PedGenie corrects for relatedness, this program does not condition on the fact that these families are linked to the 22q12.3 region. To overcome this limitation, LAMP software was used to test for association in the presence of linkage, and to test whether the associated SNPs could explain the initial linkage result, either partially or completely (17,18). The program quantifies the degree of LD between the candidate SNP and the putative disease locus through four maximum likelihood models. These models are used to construct three likelihood ratios to assess whether the candidate SNP and disease locus are linked (i.e. complete LD) or associated (i.e. partial LD) or whether there are other variants that can explain the linkage signal.

The analysis was performed with the combined set of 42 pedigrees from *PROGRESS* and Mayo Clinic, including a set of 506 unrelated population controls from Mayo. The controls, which largely overlap those used in the initial marker selection described above, were used to estimate allele frequencies and LD with the underlying causal variant. The *PROGRESS* pedigrees were too large to analyze in the LAMP program and so the pedigrees were trimmed to reduce the bit-size. However, no genotyped affected men were trimmed. Because the SNPs were in partial LD with one another, the LAMP analysis required that we first select 30 tagSNPs using an r^2 threshold of <0.1 . These SNPs make up the linkage framework map for the LAMP analyses. Each of the 145 SNPs and 11 additional SNPs that were within the broad linkage interval (25.7–37.4 Mb) initially defined by the Mayo Clinic (15) were then analyzed to test for both linkage and association. If any candidate SNP was in LD ($r^2 > 0.4$) with any SNP in the framework map, the framework SNP was removed when analyzing that candidate to ensure that residual LD did not influence the results. A disease prevalence of 0.15 was used in these analyses.

LD analysis

To target the region of greatest association identified in the analysis of the 145 markers, the LD block structure over the *APOL3* gene was determined in Haploview (v.4.0) (36), using all SNPs between 34 863 700 and 34 920 000 bp that have been genotyped in the HapMap Caucasian population (HapMap Data Rel 24/phase II Nov08, on NCBI B36 assembly, dbSNP b126) (37). Nineteen tagSNPs describe three LD-blocks at the 5' end of *APOL3* using default parameters and the 'solid spine of LD' method in Haploview. These tagSNPs were genotyped in the 18 *PROGRESS* pedigrees using direct Sanger sequencing and in the 24 Mayo Clinic pedigrees and 497 of the 506 unrelated population controls from Mayo using the ABI SNPlex Genotyping System. Nine of the 506 population controls were unavailable due to insufficient DNA. The 19 SNPs reduced to 15 as 1 SNP did not amplify in the SNPlex and 3 had an MAF of <0.05. Thus, 15 tagSNPs were available for analysis of the *APOL3* LD block structure.

Mutation screening of the *APOL3* gene

The genetic position of all exons was taken from the UCSC Genome Browser (NCBI Build 36.1). All exons and introns from the six most common transcript variants of *APOL3* were sequenced in all individuals from the 18 *PROGRESS* pedigrees (66 affected males, 68 unaffected males and 40 women) using direct Sanger sequencing. We also included a 14 kb upstream interval that spanned both putative promoters (19), and a 7 kb region of high conservation. Primer sequences were designed using Primer3 (38). Amplification was performed using standard PCR conditions with TaqGold Polymerase [Applied Biosystem (ABI), Foster City, CA, USA]. Amplicons were sequenced using the Big-Dye Terminator Cycle Sequencing Kit (v.3.1) (ABI). Sequences were collected on an ABI 3730xl DNA analyzer and analyzed using phred-Phrap, polyPhred and Consed softwares (39–42). All genotypes were confirmed using both forward and reverse sequencing data and tested for Mendelian inconsistency in each family. Primer sequences, the name and positions of all SNPs in this analysis, are available upon request.

Population-based prostate cancer case–control data set

To assess the significance of the putative risk SNPs, we looked for association in a population-based case–control data set. The study population consists of participants from one of two population-based case–control studies of PC risk factors in residents of King County, Washington (Study I and Study II), which have been described previously (20,21). Briefly, subjects diagnosed with histologically confirmed PC were ascertained from the Seattle-Puget Sound SEER cancer registry. In Study I, cases were diagnosed between 1 January 1993 and 31 December 1996 and were 40–64 years of age at diagnosis. In Study II, cases were diagnosed between 1 January 2002 and 31 December 2005 and were 35–74 years of age at diagnosis. Overall, 2244 eligible PC patients were identified, 1754 (78.2%) were interviewed and blood samples yielding sufficient DNA for genotyping were drawn from 1457

(83.1%) interviewed cases. A comparison group of controls without a self-reported physician's diagnosis of PC was identified using random digit telephone dialing. Controls were frequency matched to cases by 5 year age groups and recruited evenly throughout each ascertainment period for cases. A total of 2448 men were identified who met the eligibility criteria, 1645 (67.2%) were interviewed and blood samples were drawn and DNA prepared from 1352 men (82.2%), using standard protocols. For the current analyses, only Caucasian participants with DNA available were included (1320 cases and 1266 controls).

SNP genotyping of population-based case–control samples

Genotyping for rs132660 and rs2097465 was performed at the National Human Genome Research Institute using the SNPlex Genotyping System (Applied Biosystems, Inc., Foster City, CA, USA) according to the manufacturer's protocol. The details of this assay have been described previously (43,44). The GeneMapper software package (Applied Biosystems) was used to assign genotypes for each SNP. Replicate samples ($n = 143$) were interspersed throughout all genotyping batches, and the concordance levels for blind duplicate samples were 99.2% for rs132660 and 100% for rs2097465. All genotyping scores, including QC data, were re-checked by different laboratory personnel and the accuracy of each assay was confirmed.

Statistical analysis for population-based case–control data set

Departure from Hardy–Weinberg equilibrium was assessed for each SNP separately in controls. Unconditional logistic regression models were used to estimate odds ratios (ORs) and 95% confidence intervals (95% CI) to measure the association between individual SNP genotypes and prostate cancer risk (45), with age at reference date included in the models. Log-additive (trend), dominant and co-dominant models were considered for each SNP. Differences in risk estimates by first-degree family history of PC (yes versus no) were tested by including an interaction term in the regression model and comparing the $-2 \log$ likelihoods for the full (main effects plus the interaction term) and reduced (main effects only) models. Polytomous regression models were used to generate ORs and 95% CIs for the association between SNP genotypes and cases stratified by disease aggressiveness (less versus more) and Gleason score [≤ 7 (3 + 4) versus ≥ 7 (4 + 3)] compared with controls. More aggressive cases were those with either a Gleason score of ≥ 7 (4 + 3), regional or distant stage disease, or a PSA level ≥ 20 ng/ml at diagnosis. A χ^2 test was used to test for significant differences in risk estimates between more and less aggressive cases and between lower and higher Gleason scores. Analyses were performed using SAS version 9.1.3.

Genotyping and analysis of SNPs in the CGEMS study

Raw genotypes for marker rs2097465 were downloaded from 1176 affected (688 aggressive and 488 nonaggressive) cases and 1101 controls from the CGEMS prostate cancer study,

phase-1 (<http://cgems.cancer.gov/>) (22). Because rs132660 was not genotyped in the CGEMS study, genotypes were imputed using a hidden Markov model programmed in MACH (v1.0.14) (46). Imputation was performed using HapMap CEU phased haplotypes from 120 chromosomes from 60 founders (phase II, release 22) (37). For additional details see (MaCH; Markov Chain Haplotyping Package, <http://www.sph.umich.edu/csg/abecasis/MaCH>). Logistic regression was used to test the association between the two SNPs and prostate cancer risk, estimating genotype ORs and 95% confidence interval. Data were also stratified by disease aggressiveness (aggressive versus nonaggressive cases) for comparison to controls.

Electrophoretic mobility shift assay

The electrophoretic mobility shift assay (EMSA) was performed using the Lightshift Chemiluminescent EMSA kit (Pierce, Rockford, IL, USA) essentially as described by the manufacturer. All oligonucleotides were purchased from Integrated DNA Technologies, Inc. (Coralville, IA, USA). The target oligonucleotide—5'-TTTACCCATCATT[TATAA/CA]GAAAAGCCACTCTGGG (SNP rs132660 in bold and the transcription binding sequence for the TATAA-binding protein in box)—was 5' end labeled with biotin. Unlabeled nonspecific oligonucleotides had the following sequences: unlabeled nonspecific DNA1, 5'-CGCCCGAAGCCCGACCCGC and unlabeled nonspecific DNA2, 5'-GGATGCCTGCTCTCCACACATCCTTCAAAC. The LNCaP nuclear extract was purchased from Abcam (Cambridge, MA, USA). The binding reactions (20 μ l) included 1X binding buffer, 1 μ g poly dI-dC, 1 μ g of nuclear extract, 500-fold excess of unlabeled oligo for the competitive assay (with a 20 min room temperature pre-incubation step) and 20 fmol of 5'-labeled oligo (with a 20 min room temperature pre-incubation step). Reaction mixtures were subjected to electrophoresis using 6% DNA Retardation Gels (Invitrogen, Carlsbad, CA, USA) and then transferred onto a nylon membrane (Pierce). The binding interactions were detected using the Chemiluminescent Nucleic Acid Detection Module Kit (Pierce) and were visualized by autoradiography.

ACKNOWLEDGEMENT

We are grateful for the participation of the many men and women who contributed time and information to the *PROGRESS* and Mayo Clinic studies.

Conflict of Interest statement. None declared.

FUNDING

This work was supported by the US Public Health Service, National Institutes of Health grants [RO1 CA080122, RO1 CA056678, RO1 CA092579 to J.L.S.], [P50 CA097186 supports L.M.F.] and [RO1 CA72818 to S.N.T. and D.J.S.]. *PROGRESS* investigators also acknowledge additional support from the Prostate Cancer Foundation and the Fred Hutchinson Cancer Research Center. Mayo Clinic investigators [S.N.T.,

S.K.M., S.M.R., D.J.S.] acknowledge additional support from the Ralph C. Wilson Medical Research Foundation. E.A.O., B.J., D.M.K., P.Q., G.J. and G.W. acknowledge support from the Intramural Program of the National Institutes of Health. Genotyping services were provided by the Center for Inherited Disease Research (CIDR). CIDR is fully funded through a federal contract from the National Institutes of Health to The Johns Hopkins University, Contract Number N01-HG-65403.

REFERENCES

- Jemal, A., Siegel, R., Ward, E., Hao, Y., Xu, J. and Thun, M.J. (2009) Cancer statistics, 2009. *CA Cancer J. Clin.*, **59**, 225.
- Easton, D.F., Schaid, D.J., Whittemore, A.S. and Isaacs, W.J. (2003) Where are the prostate cancer genes?—A summary of eight genome wide searches. *Prostate*, **57**, 261–269.
- Ostrander, E.A. and Friedrichsen, D.M. (2004) Genetic factors: finding cancer susceptibility genes. In: Abeloff, M.D., Armitage, J.O., Niederhuber, J.E., Kastan, M.B. and McKenna, W.G. (eds), *Clinical Oncology*, 3rd edn. Elsevier Science, Philadelphia, PA, pp. 253–267.
- Schaid, D.J. (2004) The complex genetic epidemiology of prostate cancer. *Hum. Mol. Genet.*, **13** (Spec. 1), R103–R121.
- Ostrander, E.A. and Johannesson, B. (2008) Prostate cancer susceptibility: finding the genes. *Adv. Exp. Med. Biol.*, **617**, 179–190.
- Cunningham, J.M., McDonnell, S.K., Marks, A., Hebbing, S., Anderson, S.A., Peterson, B.J., Slager, S., French, A., Blute, M.L., Schaid, D.J. *et al.* (2003) Genome linkage screen for prostate cancer susceptibility loci: results from the Mayo Clinic Familial Prostate Cancer Study. *Prostate*, **57**, 335–346.
- Janer, M., Friedrichsen, D., Stanford, J.L., Badzioch, M.D., Kolb, S., Deutsch, K., Peters, M.A., Goode, E.L., Welti, R., DeFrance, H.B. *et al.* (2003) A genomic scan of 254 hereditary prostate cancer families. *Prostate*, **57**, 309–319.
- Lange, E.M., Gillanders, E.M., Davis, C.C., Brown, W.M., Campbell, J.K., Jones, M., Gildea, D., Riedesel, E., Albertus, J., Freas-Lutz, D. *et al.* (2003) 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**, 326–334.
- Camp, N.J., Farnham, J.M. and Cannon-Albright, L.A. (2005) Genomic search for prostate cancer predisposition loci in Utah pedigrees. *Prostate*, **65**, 365–374.
- Chang, B.L., Isaacs, S.D., Wiley, K.E., Gillanders, E.M., Zheng, S.L., Meyers, D.A., Walsh, P.C., Trent, J.M., Xu, J. and Isaacs, W.B. (2005) Genome-wide screen for prostate cancer susceptibility genes in men with clinically significant disease. *Prostate*, **64**, 356–361.
- Stanford, J.L., McDonnell, S.K., Friedrichsen, D.M., Carlson, E.E., Kolb, S., Deutsch, K., Janer, M., Hood, L., Ostrander, E.A. and Schaid, D.J. (2006) Prostate cancer and genetic susceptibility: a genome scan incorporating disease aggressiveness. *Prostate*, **66**, 317–325.
- Xu, J., Dimitrov, L., Chang, B.L., Adams, T.S., Turner, A.R., Meyers, D.A., Eeles, R.A., Easton, D.F., Foulkes, W.D., Simard, J. *et al.* (2005) A combined genomewide linkage scan of 1,233 families for prostate cancer-susceptibility genes conducted by the international consortium for prostate cancer genetics. *Am. J. Hum. Genet.*, **77**, 219–229.
- Camp, N.J., Cannon-Albright, L.A., Farnham, J.M., Baffoe-Bonnie, A.B., George, A., Powell, I., Bailey-Wilson, J.E., Carpten, J.D., Giles, G.G., Hopper, J.L. *et al.* (2007) Compelling evidence for a prostate cancer gene at 22q12.3 by the International Consortium for Prostate Cancer Genetics. *Hum. Mol. Genet.*, **16**, 1271–1278.
- Camp, N.J., Farnham, J.M. and Cannon-Albright, L.A. (2006) Localization of a prostate cancer predisposition gene to an 880-kb region on chromosome 22q12.3 in Utah high-risk pedigrees. *Cancer Res.*, **66**, 10205–10212.
- Johannesson, B., McDonnell, S.K., Karyadi, D.M., Hebbing, S.J., Wang, L., Deutsch, K., McIntosh, L., Kwon, E.M., Suuriniemi, M., Stanford, J.L. *et al.* (2008) Fine mapping of familial prostate cancer families narrows the interval for a susceptibility locus on chromosome 22q12.3 to 1.36 Mb. *Hum. Genet.*, **123**, 65–75.

16. Allen-Brady, K., Wong, J. and Camp, N. (2006) PedGenie: an analysis approach for genetic association testing in extended pedigrees and genealogies of arbitrary size. *BMC Bioinformatics*, **7**, 209.
17. Li, M., Boehnke, M. and Abecasis, G.R. (2005) Joint modeling of linkage and association: identifying SNPs responsible for a linkage signal. *Am. J. Hum. Genet.*, **76**, 934–949.
18. Li, M., Boehnke, M. and Abecasis, G.R. (2006) Efficient study designs for test of genetic association using sibship data and unrelated cases and controls. *Am. J. Hum. Genet.*, **78**, 778–792.
19. Duchateau, P.N., Pullinger, C.R., Cho, M.H., Eng, C. and Kane, J.P. (2001) Apolipoprotein L gene family: tissue-specific expression, splicing, promoter regions; discovery of a new gene. *J. Lipid Res.*, **42**, 620–630.
20. Stanford, J.L., Wiklund, K.G., McKnight, B., Daling, J.R. and Brawer, M.K. (1999) Vasectomy and risk of prostate cancer. *Cancer Epidemiol. Biomarkers Prev.*, **8**, 881–886.
21. Agalliu, I., Salinas, C.A., Hansten, P.D., Ostrander, E.A. and Stanford, J.L. (2008) Statin use and risk of prostate cancer: results from a population-based epidemiologic study. *Am. J. Epidemiol.*, **168**, 250–260.
22. Yeager, M., Orr, N., Hayes, R.B., Jacobs, K.B., Kraft, P., Wacholder, S., Minichiello, M.J., Fearnhead, P., Yu, K., Chatterjee, N. *et al.* (2007) Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat. Genet.*, **39**, 645–649.
23. Bucher, P. (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.*, **212**, 563–578.
24. Ponjavic, J., Lenhard, B., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y. and Sandelin, A. (2006) Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters %U. *Genome Biol.*, **7**, R78. <http://genomebiology.com/2006/7/8/R78>.
25. Vanhollenbeke, B. and Payes, E. (2006) The function of apolipoproteins L. *Cell. Mol. Life Sci.*, **63**, 1937–1944.
26. Smith, E.E. and Malik, H.S. (2009) The apolipoprotein L family of programmed cell death and immunity genes rapidly evolved in primates at discrete sites of host–pathogen interactions. *Genome Res.*, **19**, 850–858.
27. Page, N.M., Butlin, D.J., Lomthaisong, K. and Lowry, P.J. (2001) The human apolipoprotein L gene cluster: identification, classification, and sites of distribution. *Genomics*, **74**, 71–78.
28. Liu, Z., Lu, H., Jiang, Z., Pastuszyn, A. and Hu, C.A. (2005) Apolipoprotein I6, a novel proapoptotic Bcl-2 homology 3-only protein, induces mitochondria-mediated apoptosis in cancer cells. *Mol. Cancer Res.*, **3**, 21–31.
29. Vanhollenbeke, B. and Pays, E. (2006) The function of apolipoproteins L. *Cell. Mol. Life Sci.*, **63**, 1937–1944.
30. Horrevoets, A.J., Fontijn, R.D., van Zonneveld, A.J., de Vries, C.J., ten Cate, J.W. and Pannekoek, H. (1999) Vascular endothelial genes that are responsive to tumor necrosis factor-alpha *in vitro* are expressed in atherosclerotic lesions, including inhibitor of apoptosis protein-1, stannin, and two novel genes. *Blood*, **93**, 3418–3431.
31. Matsuda, A., Suzuki, Y., Honda, G., Muramatsu, S., Matsuzaki, O., Nagano, Y., Doi, T., Shimotohno, K., Harada, T., Nishida, E. *et al.* (2003) Large-scale identification and characterization of human genes that activate NF-kappaB and MAPK signaling pathways. *Oncogene*, **22**, 3307–3318.
32. Eeles, R.A., Kote-Jarai, Z., Al Olama, A.A., Giles, G.G., Guy, M., Severi, G., Muir, K., Hopper, J.L., Henderson, B.E., Haiman, C.A. *et al.* (2009) Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. *Nat. Genet.*, **41**, 1116–1121.
33. Sun, J., Zheng, S.L., Wiklund, F., Isaacs, S.D., Li, G., Wiley, K.E., Kim, S.-T., Zhu, Y., Zhang, Z., Hsu, F.-C. *et al.* (2009) Sequence variants at 22q13 are associated with prostate cancer risk. *Cancer Res.*, **69**, 10–15.
34. Chute, C.G., Panser, L.A., Girman, C.J., Oesterling, J.E., Guess, H.A., Jacobsen, S.J. and Lieber, M.M. (1993) The prevalence of prostatism: a population-based survey of urinary symptoms. *J. Urol.*, **150**, 85–89.
35. Scott, L.J., Mohlke, K.L., Bonnycastle, L.L., Willer, C.J., Li, Y., Duren, W.L., Erdos, M.R., Stringham, H.M., Chines, P.S., Jackson, A.U. *et al.* (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science*, **316**, 1341–1345.
36. Barrett, J.C., Fry, B., Maller, J. and Daly, M.J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.
37. IHC (2003) The International HapMap Project. *Nature*, **426**, 789–796.
38. Rozen, S. and Skaletsky, H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.*, **132**, 365–386.
39. Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.*, **8**, 186–194.
40. Ewing, B., Hillier, L., Wendl, M.C. and Green, P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.*, **8**, 175–185.
41. Gordon, D., Abajian, C. and Green, P. (1998) Consed: a graphical tool for sequence finishing. *Genome Res.*, **8**, 195–202.
42. Nickerson, D.A., Tobe, V.O. and Taylor, S.L. (1997) PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res.*, **25**, 2745–2751.
43. Tobler, A.R., Short, S., Andersen, M.R. *et al.* (2005) The SNPlex genotyping system: a flexible and scalable platform for SNP genotyping. *J. Biomol. Tech.*, **16**, 398–406.
44. De la Vega, F.M., Lazaruk, K.D., Rhodes, M.D. and Wenz, M.H. (2005) Assessment of two flexible and compatible SNP genotyping platforms: TaqMan SNP genotyping assays and the SNPlex genotyping system. *Mutat. Res.*, **573**, 111–135.
45. Breslow, N.E. and Day, N.E. (1980) *Statistical Methods in Cancer Research. Volume I. The Analysis of Case-control Studies*. Lyon: IARC Scientific Publications.
46. Li, Y., Willer, C.J., Sanna, S. and Abecasis, G.R. (2009) Genotype imputation. *Annu. Rev. Genomics Hum. Genet.*, **10**, 387–406.