Research Article

Cancer Epidemiology, Biomarkers & Prevention

A Genome-Wide Survey over the ChIP-On-Chip Identified Androgen Receptor-Binding Genomic Regions Identifies a Novel Prostate Cancer Susceptibility Locus at 12q13.13

Junjie Feng[1,2], Jielin Sun[1,2], Seong-Tae Kim[1,2], Yizhen Lu[3], Zhong Wang[1,2], Zheng Zhang[1,2], Henrik Gronberg[4], William B. Isaacs[5], S. Lilly Zheng[1,2], and Jianfeng Xu[1,2,3,6]

Abstract

Background: The molecular mechanisms for the genome-wide association studies (GWAS)-identified prostate cancer (PCa) risk-associated single-nucleotide polymorphisms (SNP) remain largely unexplained. One recent finding that the PCa risk SNPs are enriched in genomic regions containing androgen receptor (AR)-binding sites has suggested altered AR signaling as a potentially important mechanism.

Methods: To explore novel associations by leveraging this knowledge, we utilized a meta-analysis previously done over SNPs harbored in ChIP-on-chip identified AR-binding genomic regions using the GWAS data from the Johns Hopkins Hospital (JHH) and the Cancer Genetic Markers of Susceptibility (CGEMS) study, and subsequently evaluated the top associations in a third population from the CAncer of the Prostate in Sweden (CAPS) study.

Results: One SNP (rs4919743: G>A), located at the *KRT8* locus at 12q13.13 which encodes a keratin protein (K8) long used as a prostate epithelial malignancy marker and implicated in the tumorigenesis of several cancer types, was identified to be associated with PCa risk. The frequency of its minor "A" allele was consistently higher in PCa cases than in controls in all three study populations, with a combined OR of 1.22 (95% CI: 1.13–1.32) and an overall P value of 4.50×10^{-7} (Bonferroni corrected, $P = 0.006$).

Conclusion: We have identified a novel genetic locus that is associated with PCa risk.

Impact: This study illustrated the great potential of prior biological knowledge in facilitating the search for novel disease-associated genetic loci. This finding warrants further replication in other studies. *Cancer Epidemiol Biomarkers Prev; 20(11); 2396–403. ©2011 AACR.*

Introduction

Prostate cancer (PCa) is the most prevalent nonskin malignancy in men in Western countries, and is one of the leading causes of cancer mortality (1). PCa represents one of the most heritable types of cancers, with inherited genetic factors making a significant contribution to its susceptibility (2). Tremendous efforts have been invested to identify these PCa risk-associated genetic factors. To date, more than 30 single-nucleotide polymorphisms (SNPs) have been identified to be reproducibly associated with PCa predisposition, thanks primarily to the successful application of genome-wide association studies (GWAS; refs. 3–16).

It is of note that among all established PCa risk SNPs, most are located within noncoding genomic regions, leaving the question wide open as to which molecular mechanisms account for these gene desert-localized genetic variants that are associated with PCa susceptibility. Recently, through a genome-wide survey over the SNPs located on the 22,447 androgen receptor (AR)-binding sites (MIM# 313700), previously identified by chromatin immunoprecipitation combined with tiled oligonucleotide microarrays (ChIP-on-chip; ref. 17), our group has suggested the AR signaling pathway as a potentially important mechanism explaining how PCa-associated SNPs may act to influence PCa risk (18, 19). In these reports, we showed that the PCa risk-associated SNPs are significantly enriched in the AR-binding genomic regions, and that as many as one-third of the consensual PCa risk SNPs (11 of 33, as before December, 2009) lie in regions containing AR-binding sites, among which notably include 5 SNPs (rs16901979, rs620861, rs1447295, rs1859962, and

Authors' Affiliations: [1]Center for Cancer Genomics, [2]Center for Genomics and Personalized Medicine Research, Wake Forest School of Medicine, Winston-Salem, North Carolina; [3]Fudan-VARI Center for Genetic Epidemiology, Fudan University, Shanghai, China; [4]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; [5]Department of Urology, Johns Hopkins Medical Institutions, Baltimore, Maryland; and [6]Van Andel Research Institute, Grand Rapids, Michigan

Note: Supplementary data for this article are available at Cancer Epidemiology, Biomarkers & Prevention Online (http://cebp.aacrjournals.org/).

J. Feng and J. Sun contributed equally to this work.

Corresponding Author: Jianfeng Xu, Center for Cancer Genomics, Medical Center Blvd., Winston-Salem, NC 27157. Phone: 336-713-7500; Fax: 336-713-7566; E-mail: jxu@wfubmc.edu

doi: 10.1158/1055-9965.EPI-11-0523

rs9623117) that are located in the gene desert regions such as 8q24, 17q24.3, and 22q13.

There have been mounting evidences that AR plays a crucial role in the PCa initiation and progression (20). A member of the steroid receptor subclass of nuclear receptors, AR functions primarily as a ligand-dependent transcription factor. Upon binding of androgens, which include testosterone, and its more potent metabolite 5α-dihydrotestosterone (DHT), AR is activated and the activated AR dimerizes, translocates into nuclei, binds DNA at specific sequences, and ultimately results in transcriptional up- and downregulation of its downstream genes. The critical role of AR signaling in the oncogenesis of PCa is evidenced by the observation that PCa is generally androgen dependent (21), and is further supported by the ∼25% reduction in PCa risk seen in 2 large clinical trials testing agents which block androgen activation (22, 23), and by the successful application of androgen ablation therapies as the mainstay of treatment for advanced diseases that are hormone dependent (24). Although androgen-dependent PCa almost invariably progresses into an androgen-independent late stage, accumulating evidences support a model that this process relies on the reactivation of AR activity (24, 25). Combining these findings with ours, it seems a plausible hypothesis that risk SNPs located in putative AR-binding sites might change the affinity of the androgen-AR complexes to their binding sequences, which in turn may result in changes of expression of its downstream genes, ultimately leading to modification of PCa risk.

In this study, we reported our discovery of a novel PCa susceptibility locus at 12q13.13, through our further exploration of SNPs that are located within the genomic regions containing AR-binding sites. A meta-analysis over the SNPs harbored in AR-binding regions was done using the GWAS data from the Johns Hopkins Hospital (JHH) study and the Cancer Genetic Markers of Susceptibility (CGEMS). The top associated SNPs were subsequently tested for their associations with PCa risk in an independent population from the Cancer of the Prostate in Sweden (CAPS) study.

Materials and Methods

Study populations

The 2 primary GWAS populations were from the PCa GWAS study carried out by JHH and from Stage 1 of the National Cancer Institute CGEMS study. The JHH population included 1,964 PCa cases, whose clinical characteristics have been described previously (18), and 3,172 control subjects, which were from an independent Illumina iControlDB (iControls) dataset (26). The CGEMS population included 1,172 PCa cases and 1,157 control subjects (5). The genotype and phenotype data of this study are publicly available and our use of the data was approved by CGEMS. Our confirmation population was from the CAPS study, which included 2,899 PCa cases and 1,722 control subjects and has been described in great detail elsewhere (27). The research ethics committees at Wake Forest University School of Medicine and the Karolinska Institute approved the study. All of the study subjects were of European ancestry.

GWAS genotyping data, imputation, and quality control

Genotyping of SNPs in the JHH case population and in the iControls population was done using the Illumina 610K chip, and Illumina Hap300/Hap550 Chips respectively. For all 3 study populations, imputation of all the known SNPs that are catalogued in HapMap Phase II (28) was done by the IMPUTE computer program (29) with a posterior probability of 0.9 as a threshold to call genotypes. The following quality control criteria were used to filter SNPs: Minor allele frequency (MAF) < 0.01, Hardy–Weinberg equilibrium < 0.001 and call rate < 0.90.

Genotyping for the confirmation study

A subset of SNPs was genotyped using the MassArray System from Sequenom. PCR and extension primers for these SNPs were designed using the MassARRAY Assay Design 3.0 software. PCR and extension reactions were performed according to the manufacturer's instructions, and extension product sizes were determined by mass spectrometry using the Sequenom iPLEX system. Duplicates and water samples, to which the technician was blind, were included in each 96-well plate as PCR negative controls. The genotype call rates of these SNPs were > 98% and the average concordance rate between samples was > 99%.

SNPs within ChIP-on-chip detected AR-binding site regions

The data on the 22,447 putative AR-binding genomic regions across the genome discovered by ChIP-on-chip analysis in 2 PCa cell lines are publically available (17). The experimentally determined AR-binding regions range from 299 to 5,554 base pairs (bp), with a median size of 911 bp. The search for all known SNPs harbored in these AR-binding regions was based on the HapMap database (Build 36).

Statistical analysis

Allele frequency differences between case patients and control subjects were tested for each of the SNPs under investigation using a χ^2 test with 1 degree of freedom. Allelic OR and 95%CI were estimated based on a multiplicative model. Due to potential population stratification in the JHH study, the logistic regression analysis was carried out to reduce the spurious association results by adjusting for the top 5 eigenvectors that were estimated using EIGENSOFT software (30). Results from multiple case–control populations were combined using a Mantel–Haenszel model in which the populations were allowed to have different population frequencies for alleles but were assumed to have a common OR. The homogeneity of ORs among different study populations was tested using

Breslow–Day χ^2 test. SNP-SNP interactions were tested by including both SNPs and an interaction term (product of 2 SNPs) in a logistic regression model. An additive genetic model was used in the analysis.

Reported PCa risk-associated SNPs by GWAS

From PCa GWAS reported before December 2009, a total of 33 PCa risk-associated SNPs were selected for comparison in this study, based on the selection criterion that the associations with PCa exceeded genome-wide significance levels in their initial reports ($P < 10^{-7}$) which had also been replicated in independent study populations. The linkage disequilibrium (LD) blocks for these risk SNPs were also inferred to define PCa risk-associated genomic regions, based on the CEU genotype data from HapMap release#27 (Phase II + Phase III). A LD block was defined as a set of SNPs within a genomic region of 1,000 kb with pairwise r^2 value ≥ 0.5, and was estimated using the CLUMP function of the PLINK software (31). After exclusion of one SNP (rs16902094), the data of which is unavailable in the HapMap database, a total of 32 PCa susceptibility associated LD blocks were identified. The SNP list and pairwise r^2 values of the PCa risk-associated SNPs have been described (18).

Results

A genome-wide survey over the 22,447 ChIP-on-chip detected AR-binding site regions revealed a total of 18,401 SNPs that are located on AR-binding regions, based on the HapMap database (Build 36), among which there were 13,899 SNPs, found to reside in 8,189 AR-binding regions, that were commonly found from both the CGEMS and JHH populations to be directly genotyped or indirectly imputed and successfully passed our quality control criteria (Supplementary Table S1). The allele frequency differences between PCa cases and controls were tested for each of these SNPs in each individual and their combined population. A total of 46 AR-binding region harboring SNPs were found to be associated with PCa risk at $P < 0.001$ in the combined GWAS data from JHH and CGEMS studies (Table 1). Given that some of these SNPs are in LD, SNPs are grouped into LD blocks ($r^2 \geq 0.5$). Among the 29 LD blocks in which these 46 SNPs are located, there are notably 6 (20.7%, 6/29) blocks, containing 15 (30.6%, 15/46) SNPs, that overlap with the 32 PCa risk-associated LD blocks derived from the consensual PCa associated SNPs (Table 1). These observations are similar to those reported in our previous article (18) although different quality control criteria were used.

To confirm the above association results, we genotyped 23 candidate SNPs in the CAPS study population, which represent the tagging SNPs for the 23 of the aforementioned 29 AR-binding regions containing LD blocks. The SNPs in the remaining 6 LD blocks that contain at least one of the established PCa risk-associated SNPs were reported in CAPS population in the previous study (27) and thus were not examined further in this study. Except for one SNP (rs8087095), which failed our quality control criteria, the allelic test was performed to examine the association of each of the remaining 22 SNPs with the risk of PCa. As shown in Table 2, one SNP (rs4919743: G>A at 12q13) was significantly associated with PCa susceptibility ($P = 7.17 \times 10^{-4}$), with its minor "A" allele more frequently found in PCa cases than in control subjects (OR = 1.22, 95% CI = 1.09–1.37) in the CAPS population. The association remained significant after Bonferroni correction for multiple comparisons ($P = 0.016$). The direction of association was also consistent with that in the other 2 GWAS populations (OR = 1.21, 95% CI = 1.06–1.37 for JHH, and OR = 1.22, 95% CI = 1.02–1.46 for CGEMS, Table 1). Age adjustment did not seem to alter the significance of the associations across the 3 populations (data not shown). It is noteworthy that when all 3 populations were combined, the association of rs4919743 with PCa risk ($P = 4.50 \times 10^{-7}$, OR = 1.22, 95% CI = 1.13–1.32, Table 3) remained significant after Bonferroni correction ($P = 0.006$). No heterogeneity was detected across the 3 study populations ($P = 0.9974$ for Breslow-Day χ^2 test). Except for rs4919743, we failed to find any significant associations for the other SNPs in this study population.

We further evaluated whether rs4919743 interacts with any of the other 21SNPs. In the CAPS population, the most significant interaction ($P_{interaction} = 0.008$, Table 4) was observed between rs4919743: G>A and rs4741304: G>T, which interestingly was also top ranked in the univariate analysis (Table 2). Further examination indicated that the interaction between these 2 SNPs was also nominally significant in the CGEMS population ($P_{interaction} = 0.05$), but not so in the JHH population ($P_{interaction} = 0.37$). The analysis over the combined population revealed that interaction between these 2 SNPs was significant ($P = 0.002$, Bonferroni corrected, $P = 0.042$; Table 4). No interaction effect was confirmed for other pairs in either CGEMS or JHH population (data not shown).

Discussion

Through a genome-wide survey over the SNPs that are located on the AR-binding genomic regions detected by the ChIP-on-chip technology, we identified a novel genetic locus at 12q13.13 (rs4919743) that was highly associated with PCa risk ($P = 4.50 \times 10^{-7}$) and remained significant after correction for multiple comparisons. The minor "A" allele of this SNP was consistently shown to confer higher risk for PCa across 3 independent study populations with a combined OR of 1.22 (95% CI: 1.13–1.32).

Rs4919743 is located within the keratin gene cluster at 12q13.13 which encodes ~30 keratin proteins that constitute the cytoskeletonal proteins of the intermediate filament (IF). This SNP is in a 60-Kbp LD block spanning the whole gene region of *KRT8* (MIM# 148060, Supplementary Fig. S1), which encodes a keratin protein (K8) that typically dimerizes with K18 (encoded by *KRT18*; MIM# 148070) to form IFs in simple single-layered epithelial cells. Due to its cell type and differentiation/functional

Table 1. AR-binding SNPs associated with prostate cancer risk from the 2 GWAS datasets (CGEMS and JHH) at $P \leq 0.001$[a]

					JHH				CGEMS				Combined				
					MAF				MAF								
SNP	CHR	BP[b]	Alleles	Minor allele	Case	Ctrl	OR (95% CI)	*P*	Case	Ctrl	OR (95% CI)	*P*	OR (95% CI)	*P*	LD	Known risk SNPs[c]	Conf[d]
rs6703670	1p36.23	7,953,443	A/G	A	0.01	0.02	0.63 (0.45–0.87)	5.16E-03	0.01	0.02	0.66 (0.41–1.04)	0.07	0.64 (0.49–0.83)	9.09E-04	1		Yes
rs3765227	1q24.2	167,703,777	A/C	A	0.05	0.06	0.75 (0.62–0.90)	2.29E-03	0.03	0.05	0.70 (0.52–0.94)	0.02	0.73 (0.63–0.86)	1.26E-04	2		Yes
rs3737683	1q24.2	167,704,804	G/A	G	0.05	0.06	0.75 (0.62–0.90)	2.31E-03	0.03	0.05	0.70 (0.52–0.94)	0.02	0.73 (0.63–0.86)	1.27E-04	2		
rs6712527	2q33.3	208,588,496	C/T	C	0.26	0.25	1.09 (0.99–1.21)	0.08	0.30	0.25	1.27 (1.11–1.44)	4.02E-04	1.15 (1.07–1.25)	3.76E-04	3		Yes
rs7569918	2q34	213,592,762	T/C	T	0.29	0.26	1.13 (1.03–1.24)	0.01	0.29	0.26	1.16 (1.01–1.32)	0.03	1.14 (1.05–1.23)	8.85E-04	4		Yes
rs7640145	3q22.3	140,216,904	G/A	G	0.15	0.17	0.86 (0.77–0.97)	0.01	0.15	0.19	0.76 (0.65–0.89)	6.48E-04	0.83 (0.76 - 0.91)	4.55E-05	5		Yes
rs13134869	4q31.21	144,922,904	G/A	G	0.32	0.35	0.85 (0.78–0.93)	2.51E-04	0.33	0.34	0.96 (0.85–1.09)	0.53	0.88 (0.82–0.95)	7.78E-04	6		Yes
rs2940919	5p12	42,506,487	G/A	G	0.21	0.19	1.13 (1.02–1.26)	0.02	0.21	0.18	1.24 (1.07–1.44)	4.52E-03	1.17 (1.07 - 1.27)	3.77E-04	7		
rs2940920	5p12	42,507,356	C/T	C	0.21	0.19	1.13 (1.02–1.26)	0.02	0.21	0.18	1.24 (1.07–1.44)	4.52E-03	1.17 (1.07–1.27)	3.76E-04	7		Yes
rs11955681	5q31.3	142,920,885	A/T	A	0.16	0.13	1.28 (1.14–1.44)	6.05E-05	0.17	0.16	1.05 (0.89–1.23)	0.59	1.19 (1.08 - 1.31)	3.95E-04	8		Yes
rs12213703	6p22.3	22,929,929	A/G	A	0.11	0.09	1.31 (1.14–1.50)	9.50E-05	0.10	0.09	1.06 (0.87–1.29)	0.58	1.22 (1.09–1.37)	4.14E-04	9		Yes
rs17156041	7p15.2	27,942,270	T/C	T	0.20	0.22	0.85 (0.77–0.94)	1.89E-03	0.19	0.21	0.87 (0.76–1.01)	0.07	0.86 (0.79 - 0.93)	3.19E-04	10	rs10486567	
rs10486567	7p15.2	27,943,088	A/G	A	0.21	0.24	0.86 (0.78–0.94)	2.01E-03	0.21	0.23	0.87 (0.76–1.00)	0.06	0.86 (0.79–0.93)	2.83E-04	10	rs10486567	
rs1398240	8p21.2	23,557,551	C/A	C	0.43	0.39	1.19 (1.10–1.30)	3.14E-05	0.42	0.41	1.08 (0.95–1.22)	0.23	1.15 (1.08 - 1.24)	4.00E-05	11	rs1512268	
rs1160267	8p21.2	23,585,466	G/A	G	0.46	0.42	1.22 (1.12–1.32)	3.28E-06	0.44	0.43	1.06 (0.94–1.19)	0.37	1.16 (1.09–1.24)	1.34E-05	11	rs1512268	
rs13259131	8q24.13	124,106,137	T/C	T	0.43	0.39	1.19 (1.09–1.30)	6.08E-05	0.41	0.39	1.09 (0.96–1.22)	0.19	1.15 (1.08–1.24)	5.25E-05	12		Yes
rs7824451	8q24.21	128,183,643	G/C	G	0.04	0.03	1.66 (1.33–2.07)	8.44E-06	0.04	0.03	1.23 (0.91–1.66)	0.19	1.49 (1.25–1.79)	1.22E-05	13	rs16901979	
rs7812429	8q24.21	128,589,355	A/G	A	0.12	0.09	1.36 (1.19–1.54)	4.12E-06	0.14	0.10	1.52 (1.27–1.82)	6.38E-06	1.41 (1.27 - 1.56)	1.87E-10	14	rs1447295	
rs7812894	8q24.21	128,589,661	A/T	A	0.12	0.09	1.36 (1.19–1.54)	4.12E-06	0.14	0.10	1.52 (1.27–1.82)	6.38E-06	1.41 (1.27–1.56)	1.87E-10	14	rs1447295	
rs10099413	8q24.21	128,591,245	T/C	T	0.15	0.12	1.26 (1.12–1.42)	9.28E-05	0.18	0.13	1.44 (1.22–1.70)	1.41E-05	1.32 (1.20 - 1.45)	1.20E-08	14	rs1447295	
rs7814837	8q24.21	128,591,384	T/G	T	0.12	0.09	1.36 (1.19–1.54)	4.12E-06	0.14	0.10	1.52 (1.27–1.82)	6.38E-06	1.41 (1.27–1.56)	1.87E-10	14	rs1447295	
rs2472537	8q24.22	133,929,827	G/A	G	0.24	0.21	1.19 (1.08–1.31)	6.59E-04	0.23	0.21	1.12 (0.97–1.29)	0.11	1.16 (1.07 - 1.26)	2.12E-04	15		Yes
rs4741304	9p23	13,478,608	G/T	G	0.14	0.12	1.19 (1.05–1.34)	5.55E-03	0.14	0.11	1.24 (1.03–1.48)	0.02	1.20 (1.09–1.33)	3.26E-04	16		Yes
rs7075697	10q11.23	51,217,377	C/G	C	0.50	0.45	1.25 (1.14–1.36)	6.80E-07	0.46	0.41	1.23 (1.10–1.39)	5.55E-04	1.24 (1.16 - 1.33)	1.46E-09	17	rs10993994	
rs12223916	11q22.1	99,969,463	G/C	G	0.10	0.12	0.87 (0.76–0.99)	0.04	0.10	0.12	0.77 (0.64–0.93)	6.19E-03	0.83 (0.75–0.93)	9.94E-04	18		
rs11224335	11q22.1	100,008,564	T/C	T	0.09	0.11	0.84 (0.73–0.97)	0.01	0.09	0.12	0.75 (0.62–0.91)	3.90E-03	0.81 (0.73 - 0.91)	2.58E-04	18		Yes
rs11224336	11q22.1	100,008,742	A/G	A	0.09	0.11	0.84 (0.73–0.97)	0.01	0.09	0.12	0.75 (0.62–0.92)	4.42E-03	0.81 (0.73–0.91)	2.74E-04	18		
rs7122442	11q22.3	107,983,350	G/A	G	0.03	0.04	0.73 (0.58–0.94)	0.01	0.02	0.03	0.57 (0.38–0.87)	8.78E-03	0.69 (0.56 - 0.85)	5.17E-04	19		Yes
rs4919743	12q13.13	51,595,851	A/G	A	0.13	0.11	1.21 (1.06–1.37)	4.34E-03	0.14	0.12	1.22 (1.02–1.46)	0.03	1.21 (1.09–1.34)	3.64E-04	20		Yes
rs9595100	13q14.11	43,783,474	G/A	G	0.23	0.25	0.90 (0.82–0.99)	0.03	0.22	0.26	0.83 (0.72–0.95)	7.49E-03	0.88 (0.81–0.95)	9.35E-04	21		Yes
rs3813546	14q24.3	75,689,714	G/A	G	0.21	0.19	1.14 (1.03–1.27)	0.01	0.20	0.17	1.20 (1.03–1.39)	0.02	1.16 (1.06–1.26)	7.42E-04	22		Yes
rs16956801	15q13.3	29,579,883	A/C	A	0.16	0.13	1.28 (1.14–1.44)	3.72E-05	0.15	0.14	1.09 (0.92–1.29)	0.32	1.21 (1.10–1.34)	7.65E-05	23		Yes
rs12918539	16p12.3	17,386,896	T/C	T	0.37	0.39	0.91 (0.84–1.00)	0.04	0.34	0.38	0.83 (0.73–0.94)	3.00E-03	0.89 (0.83–0.95)	7.38E-04	24		Yes
rs8071558	17q24.3	66,619,268	G/C	G	0.48	0.52	0.85 (0.79–0.93)	1.48E-04	0.47	0.52	0.84 (0.75–0.94)	2.45E-03	0.85 (0.79–0.91)	1.25E-06	25	rs1859962	
rs8072254	17q24.3	66,619,411	G/A	G	0.48	0.52	0.85 (0.79–0.93)	1.48E-04	0.47	0.52	0.84 (0.75–0.94)	2.45E-03	0.85 (0.79–0.91)	1.25E-06	25	rs1859962	
rs984434	17q24.3	66,619,722	C/T	C	0.48	0.52	0.85 (0.79–0.93)	1.48E-04	0.47	0.52	0.84 (0.74–0.94)	2.21E-03	0.85 (0.79–0.91)	1.15E-06	25	rs1859962	

(Continued on the following page)

Table 1. AR-binding SNPs associated with prostate cancer risk from the 2 GWAS datasets (CGEMS and JHH) at $P \leq 0.001$[a] (Cont'd)

					JHH				CGEMS				Combined				
					MAF				MAF								
SNP	CHR	BP[b]	Alleles	Minor allele	Case	Ctrl	OR (95% CI)	*P*	Case	Ctrl	OR (95% CI)	*P*	OR (95% CI)	*P*	LD	Known risk SNPs[c]	Conf[d]
rs1859962	17q24.3	66,620,348	T/G	T	0.48	0.52	0.86 (0.79–0.93)	2.61E-04	0.47	0.51	0.84 (0.75–0.94)	2.42E-03	0.85 (0.80–0.91)	2.28E-06	25	rs1859962	
rs8077906	17q24.3	66,623,828	A/G	A	0.31	0.33	0.91 (0.84–1.00)	0.04	0.32	0.36	0.83 (0.74–0.94)	3.39E-03	0.88 (0.82–0.95)	7.45E-04	25	rs1859962	
rs8087095	18q21.2	49,551,741	A/G	A	0.20	0.23	0.86 (0.78–0.95)	3.49E-03	0.23	0.25	0.89 (0.77–1.02)	0.10	0.87 (0.80–0.94)	8.25E-04	26		Yes
rs6036007	20p11.22	21,922,590	C/T	C	0.43	0.46	0.91 (0.84–0.99)	0.03	0.42	0.47	0.84 (0.75–0.94)	3.49E-03	0.89 (0.83–0.95)	4.67E-04	27		
rs201548	20p11.22	21,922,835	C/T	C	0.43	0.46	0.91 (0.84–0.99)	0.03	0.42	0.47	0.84 (0.75–0.94)	3.49E-03	0.89 (0.83–0.95)	4.67E-04	27		Yes
rs6106438	20p11.22	21,923,026	C/T	C	0.43	0.46	0.91 (0.84–0.99)	0.03	0.42	0.47	0.84 (0.75–0.95)	4.30E-03	0.89 (0.83–0.95)	5.38E-04	27		
rs6022257	20q11.23	36,052,672	T/C	T	0.16	0.19	0.82 (0.74–0.92)	4.96E-04	0.17	0.19	0.87 (0.74–1.01)	0.07	0.84 (0.77–0.92)	1.06E-04	28		Yes
rs6022259	20q11.23	36,052,895	T/C	T	0.16	0.19	0.83 (0.75–0.93)	1.02E-03	0.17	0.19	0.87 (0.74–1.01)	0.07	0.85 (0.77–0.92)	2.02E-04	28		
rs2899334	22q13.1	38,881,796	T/C	T	0.40	0.44	0.86 (0.79–0.93)	2.61E-04	0.40	0.41	0.96 (0.85–1.08)	0.45	0.89 (0.83–0.95)	6.39E-04	29		
rs6001831	22q13.1	38,921,204	T/G	T	0.40	0.44	0.86 (0.79–0.93)	2.66E-04	0.40	0.41	0.95 (0.85–1.07)	0.43	0.89 (0.83–0.95)	6.16E-04	29		Yes

[a]The *P* values and OR for each SNP were calculated based on multiplicative models. SNPs were listed according to their chromosomal positions and organized based on the LD blocks ($r^2 > = 0.5$) where they are located.

[b]The base position (BP) of each SNP was based on NCBI build 36.

[c]Known risk SNPs were selected from consensual PCa risk-associated SNPs which are in LD ($r^2 > = 0.5$) with the SNPs under investigation (3–16).

[d]SNPs that were selected for confirmation ("Conf") were labeled as "Yes." Abbreviations: CHR, chromosome; Ctrl, control subjects.

Table 2. Confirmation of AR-binding SNPs in the CAPS study population[a]

					MAF			
SNP	CHR	BP[b]	Alleles	Minor allele	Cases	Controls	*P*	OR (95% CI)
rs4919743	12q13.13	51595851	A/G	A	0.17	0.15	7.17E-04	1.22 (1.09–1.37)
rs4741304	9p23	13478608	G/T	G	0.13	0.11	0.06	1.13 (1.00–1.29)
rs3813546	14q24.3	75689714	G/A	G	0.17	0.15	0.19	1.08 (0.96–1.21)
rs2472537	8q24.22	133929827	G/A	G	0.23	0.24	0.21	0.94 (0.85–1.04)
rs6703670	1p36.23	7953443	A/G	A	0.06	0.06	0.27	0.90 (0.76–1.08)
rs13134869	4q31.21	144922904	G/A	G	0.34	0.35	0.27	0.95 (0.87–1.04)
rs2940920	5p12	42507356	C/T	C	0.16	0.16	0.40	0.95 (0.85–1.07)
rs12918539	16p12.3	17386896	T/C	T	0.33	0.33	0.43	1.04 (0.95–1.13)
rs16956801	15q13.3	29579883	A/C	A	0.14	0.15	0.48	0.96 (0.85–1.08)
rs6001831	22q13.1	38921204	T/G	T	0.39	0.39	0.56	0.97 (0.89–1.06)
rs3765227	1q24.2	167703777	A/C	A	0.03	0.03	0.61	1.07 (0.82–1.39)
rs201548	20p11.22	21922835	C/T	C	0.39	0.39	0.67	0.98 (0.90–1.07)
rs11224335	11q22.1	100008564	T/C	T	0.10	0.10	0.73	1.03 (0.89–1.18)
rs6712527	2q33.3	208588496	C/T	C	0.26	0.25	0.74	1.02 (0.92–1.12)
rs13259131	8q24.13	124106137	T/C	T	0.42	0.42	0.75	1.01 (0.93–1.11)
rs7640145	3q22.3	140216904	G/A	G	0.15	0.15	0.75	1.02 (0.91–1.15)
rs6022257	20q11.23	36052672	T/C	T	0.17	0.17	0.78	0.98 (0.88–1.10)
rs9595100	13q14.11	43783474	G/A	G	0.26	0.26	0.85	0.99 (0.90–1.09)
rs11955681	5q31.3	142920885	A/T	A	0.19	0.19	0.87	0.99 (0.88–1.11)
rs7569918	2q34	213592762	T/C	T	0.29	0.29	0.91	1.01 (0.92–1.10)
rs12213703	6p22.3	22929929	A/G	A	0.10	0.10	0.94	1.01 (0.87–1.16)
rs7122442	11q22.3	107983350	G/A	G	0.03	0.03	0.99	1.00 (0.79–1.27)

[a]The *P* values and ORs for each SNP were calculated based on multiplicative models. SNPs were ranked based on their *P* values in the association test;

[b]The BP of each SNP was based on NCBI build 36.

Abbreviation: CHR, chromosome.

status-specific expression in the epithelial cells, K8 has long been used as a diagnostic marker for a variety of epithelial malignancies including PCa (reviewed in ref. 32). The functional implication of *KRT8* in epithelial tumorigenesis is suggested by the findings that K8 deficiency or overexpression can induce colorectal hyperplasia (33) and pancreatic neoplasia (34), respectively, and that the altered K8 expression or phosphorylation may be causally related to the invasion and metastasis of several cancer types (35–37). Given these revelations, it is thus possible that rs4919743, or another SNP in LD with it, may cause the change of structure, expression, or regulation of the K8 protein such that it confers susceptibility to PCa. It is noteworthy that 2 nonsynonymous SNPs (rs11170164 and rs641615) within *KRT5* (MIM# 148040), which encodes another keratin protein (K5) expressed in the basal epithelial cells, have been associated with predisposition for basal cell carcinoma in a recent GWAS (38).

Table 3. The association results for rs4919743 in the 3 study populations

			Genotype counts (AA, GA, GG)						Risk allele frequency			
Population	Alleles	Risk allele	Cases			Controls			Case	Controls	OR (95% CI)[a]	*P*[a]
JHH	A/G	A	45	366	1363	35	550	2298	0.13	0.11	1.21 (1.06–1.37)	4.34E-03
CGEMS	A/G	A	18	277	825	17	213	817	0.14	0.12	1.22 (1.02–1.46)	0.03
CAPS	A/G	A	81	823	1958	33	439	1266	0.17	0.15	1.22 (1.09–1.37)	7.17E-04
Combined	A/G	A									1.22 (1.13–1.32)	4.50E-07

[a]The *P* values and ORs were calculated based on multiplicative models. The combined *P* values and ORs were calculated based on Mantel–Haenszel test, assuming a fixed effect model.

Table 4. Interaction analysis between rs4919743 and rs4741304[a]

	rs4919743 (A/G)[b]	rs4741304 (G/T)[b]	Interaction		
Population	*P*	*P*	OR (95% CI)	$P_{interaction}$	β
JHH	0.004	0.002	0.89 (0.68–1.16)	0.37	−0.12
CGEMS	0.004	5.00E-04	0.66 (0.44–1.00)	0.05	−0.41
CAPS	2.41E-05	0.002	0.72 (0.56–0.92)	0.008	−0.33
Combined	7.37E-14	5.94E-07	0.78 (0.66–0.92)	0.002	−0.25

[a]The additive genetic model was used for the interaction analyses over the 3 individual (CGEMS, JHH, and CAPS) and their combined populations. [b]The parenthesis indicates minor/major alleles. The minor alleles were risk alleles in the additive model. β, coefficient for the multiplicative interaction term.

Thus our findings, combined with this one, points to potentially important implications of keratin variants in the cancer etiologies. Despite reports that expression of *KRT8* may be correlated with increased invasiveness of cancers, our analysis on rs4919743 failed to identify any associations of this SNP with the aggressiveness of PCa (Supplementary Table S2).

This hypothesis aside, the fact that rs4919743 is harbored within a ChIP-on-chip identified AR-binding site region suggests a potential involvement of the AR signaling in the risk-conferring mechanism. Recent advances reveal that many of the sites whereby AR acts on the androgen-responsive genes localize outside of the classic promoter regions; they instead predominantly lie in distal enhancer regions (39). It is thus possible that the genetic variant of rs4919743 or other SNPs harbored in the AR-binding site region (chromosome 12: from 51,595,612 to 51,596,236 bp, based on NCBI Build 36.1) identified by Wang and colleagues (17), may affect the binding/signaling of AR at this very region which in turn could influence the expression of certain AR-target genes that are causally related to PCa risk, although not necessarily in close proximity to the regulatory element. It should be noted that under this alternative hypothesis the possibility still holds that the causal event is channeled through the altered expression of *KRT8* gene, as suggested by a report that *KRT8* expression in prostate epithelia might be AR repressive (40). Additional studies are warranted to replicate this association and investigate the precise molecular mechanism conferring PCa susceptibility.

An evidence of epistasis between rs4919743 and rs4741304 at 9p23 was suggested statistically in 2 of the study populations we examined. Because no genes are located within the ~200 kbp window upstream and downstream of rs4741304, it is difficult to make inferences about the biological mechanism. Given the fact that these 2 SNPs are both harbored in AR-binding regions, it seems possible that the yet-to-be-identified genes regulated by these 2 likely AR enhancers may be involved in a shared biological process. Alternatively, there may exist certain physical interactions between these 2 loci, mediated possibly by AR and its transcriptional regulators. More studies are required to elucidate the underlying mechanism.

This study has illustrated the great potential of prior biological knowledge in facilitating the search for novel disease-associated genetic variants. With more than 1 million genome-wide SNP markers being tested, current GWAS usually requires a stringent multiple-testing adjustment to control for false positive associations. Such a harsh penalty causes that only the most significant associations are established, with the majority of plausible associations still remaining buried within the statistical "noise" that is inherent to this approach (41). Here, by incorporating the experimentally determined AR-binding knowledge into the currently available PCa GWAS data, we have constrained multiple testing and detected a novel PCa risk-associated locus which was able to be confirmed in an independent study population but had not previously been detected. It should be mentioned that this association ($P = 4.50 \times 10^{-7}$), albeit did not reach the most conservative genome-wide significance level ($P < 5 \times 10^{-8}$), was strictly statistically significant given the much smaller number of SNPs (13,899) being tested in this study that are located in the AR-binding regions (Bonferroni corrected, $P = 0.006$.). It is noteworthy that the currently widely applied "pathway analysis," that intends to better identify disease associated genetic loci by leveraging the available GWAS data, is essentially based on this very same concept. Although the pathway analysis utilizes previously curated pathway knowledge (KEGG, NCI, and Biocarta, etc.) and focuses on protein-encoding genes, the approach used in this study is based on experimentally obtained knowledge on genomic regions that are of relevance. Not only does his new approach widens the scope of pathway analysis by extending to noncoding genomic regions, but it also is more reliable because it is based on carefully controlled experiments, which has an advantage of being immune to selection biases in the pathway analysis. We expect more and more novel-yet-hidden associations to be uncovered by this approach.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Grant Support

The study is partially supported by the National Cancer Institute (R01CA129684 to J. Xu).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received June 14, 2011; revised September 12, 2011; accepted September 16, 2011; published OnlineFirst September 28, 2011.

References

1. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Thun MJ. Cancer statistics, 2009. CA Cancer J Clin 2009;59:225–49.
2. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, et al. Environmental and heritable factors in the causation of cancer–analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med 2000;343:78–85.
3. Amundadottir LT, Sulem P, Gudmundsson J, Helgason A, Baker A, Agnarsson BA, et al. A common variant associated with prostate cancer in European and African populations. Nat Genet 2006;38:652–8.
4. Gudmundsson J, Sulem P, Manolescu A, Amundadottir LT, Gudbjartsson D, Helgason A, et al. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. Nat Genet 2007;39:631–7.
5. Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, Wacholder S, et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat Genet 2007;39:645–9.
6. Gudmundsson J, Sulem P, Steinthorsdottir V, Bergthorsson JT, Thorleifsson G, Manolescu A, et al. Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. Nat Genet 2007;39:977–83.
7. Duggan D, Zheng SL, Knowlton M, Benitez D, Dimitrov L, Wiklund F, et al. Two genome-wide association studies of aggressive prostate cancer implicate putative prostate tumor suppressor gene DAB2IP. J Natl Cancer Inst 2007;99:1836–44.
8. Thomas G, Jacobs KB, Yeager M, Kraft P, Wacholder S, Orr N, et al. Multiple loci identified in a genome-wide association study of prostate cancer. Nat Genet 2008;40:310–5.
9. Gudmundsson J, Sulem P, Rafnar T, Bergthorsson JT, Manolescu A, Gudbjartsson D, et al. Common sequence variants on 2p15 and Xp11.22 confer susceptibility to prostate cancer. Nat Genet 2008;40:281–3.
10. Eeles RA, Kote-Jarai Z, Giles GG, Olama AA, Guy M, Jugurnauth SK, et al. Multiple newly identified loci associated with prostate cancer susceptibility. Nat Genet 2008;40:316–21.
11. Yeager M, Chatterjee N, Ciampa J, Jacobs KB, Gonzalez-Bosquet J, Hayes RB, et al. Identification of a new prostate cancer susceptibility locus on chromosome 8q24. Nat Genet 2009;41:1055–7.
12. Gudmundsson J, Sulem P, Gudbjartsson DF, Blondal T, Gylfason A, Agnarsson BA, et al. Genome-wide association and replication studies identify four variants associated with prostate cancer susceptibility. Nat Genet 2009;41:1122–6.
13. Eeles RA, Kote-Jarai Z, Al Olama AA, Giles GG, Guy M, Severi G, et al. Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. Nat Genet 2009;41:1116–21.
14. Sun J, Zheng SL, Wiklund F, Isaacs SD, Purcell LD, Gao Z, et al. Evidence for two independent prostate cancer risk-associated loci in the HNF1B gene at 17q12. Nat Genet 2008;40:1153–5.
15. Hsu FC, Sun J, Wiklund F, Isaacs SD, Wiley KE, Purcell LD, et al. A novel prostate cancer susceptibility locus at 19q13. Cancer Res 2009;69:2720–3.
16. Sun J, Zheng SL, Wiklund F, Isaacs SD, Li G, Wiley KE, et al. Sequence variants at 22q13 are associated with prostate cancer risk. Cancer Res 2009;69:10–5.
17. Wang Q, Li W, Zhang Y, Yuan X, Xu K, Yu J, et al. Androgen receptor regulates a distinct transcription program in androgen-independent prostate cancer. Cell 2009;138:245–56.
18. Lu Y, Sun J, Kader AK, Kim ST, Kim JW, Liu W, et al. Association of prostate cancer risk with SNPs in regions containing androgen receptor binding sites captured by ChIP-on-chip analyses. Prostate (in press).
19. Lu Y, Zhang Z, Yu H, Zheng SL, Isaacs WB, Xu J, et al. Functional annotation of risk loci identified through genome-wide association studies for prostate cancer. Prostate 2011;71:955–63.
20. Heinlein CA, Chang C. Androgen receptor in prostate cancer. Endocr Rev 2004;25:276–308.
21. Huggins C, Hodges CV. Studies on prostatic cancer. I. The effect of castration, of estrogen and androgen injection on serum phosphatases in metastatic carcinoma of the prostate. CA Cancer J Clin 1972;22:232–40.
22. Thompson IM, Goodman PJ, Tangen CM, Lucia MS, Miller GJ, Ford LG, et al. The influence of finasteride on the development of prostate cancer. N Engl J Med 2003;349:215–24.
23. Andriole GL, Bostwick DG, Brawley OW, Gomella LG, Marberger M, Montorsi F, et al. Effect of dutasteride on the risk of prostate cancer. N Engl J Med;362:1192–202.
24. Knudsen KE, Scher HI. Starving the addiction: new opportunities for durable suppression of AR signaling in prostate cancer. Clin Cancer Res 2009;15:4792–8.
25. Yuan X, Balk SP. Mechanisms mediating androgen receptor reactivation after castration. Urol Oncol 2009;27:36–41.
26. Available from: http://www.illumina.com/science/icontroldb.ilmn.
27. Sun J, Kader AK, Hsu FC, Kim ST, Zhu Y, Turner AR, et al. Inherited genetic markers discovered to date are able to identify a significant number of men at considerably elevated risk for prostate cancer. Prostate.
28. Available from: http://www.hapmap.org.
29. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38:904–9.
30. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 2007;39:906–13.
31. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–75.
32. Moll R, Divo M, Langbein L. The human keratins: biology and pathology. Histochem Cell Biol 2008;129:705–33.
33. Baribault H, Price J, Miyai K, Oshima RG. Mid-gestational lethality in mice lacking keratin 8. Genes Dev 1993;7:1191–202.
34. Casanova ML, Bravo A, Ramirez A, Morreale de Escobar G, Were F, Merlino G, et al. Exocrine pancreatic disorders in transsgenic mice expressing human keratin 8. J Clin Invest 1999;103:1587–95.
35. Chu YW, Seftor EA, Romer LH, Hendrix MJ. Experimental coexpression of vimentin and keratin intermediate filaments in human melanoma cells augments motility. Am J Pathol 1996;148:63–9.
36. Hendrix MJ, Seftor EA, Seftor RE, Trevor KT. Experimental co-expression of vimentin and keratin intermediate filaments in human breast cancer cells results in phenotypic interconversion and increased invasive behavior. Am J Pathol 1997;150:483–95.
37. Mizuuchi E, Semba S, Kodama Y, Yokozaki H. Down-modulation of keratin 8 phosphorylation levels by PRL-3 contributes to colorectal carcinoma progression. Int J Cancer 2009;124:1802–10.
38. Stacey SN, Sulem P, Masson G, Gudjonsson SA, Thorleifsson G, Jakobsdottir M, et al. New common variants affecting susceptibility to basal cell carcinoma. Nat Genet 2009;41:909–14.
39. He HH, Meyer CA, Shin H, Bailey ST, Wei G, Wang Q, et al. Nucleosome dynamics define transcriptional enhancers. Nat Genet;42:343–7.
40. Wang XH, Hsieh JT. [Localization and regulation of cytokeratin 8 mRNA expression in rat prostate epithelia]. Shi Yan Sheng Wu Xue Bao 1994;27:447–55.
41. Kruglyak L. The road to genome-wide association studies. Nat Rev Genet 2008;9:314–8.