A Large Multiethnic Genome-Wide Association Study of Prostate Cancer Identifies Novel Risk Variants and Substantial Ethnic Differences

RESEARCH ARTICLE

ABSTRACT
A genome-wide association study (GWAS) of prostate cancer in Kaiser Permanente health plan members (7,783 cases, 38,595 controls; 80.3% non-Hispanic white, 4.9% African-American, 7.0% East Asian, and 7.8% Latino) revealed a new independent risk indel rs4646284 at the previously identified locus 6q25.3 that replicated in PEGASUS (N = 7,539) and the Multiethnic Cohort (N = 4,679) with an overall \( P = 1.0 \times 10^{-15} \) (OR, 1.18). Across the 6q25.3 locus, rs4646284 exhibited the strongest association with expression of SLC22A1 (\( P = 1.3 \times 10^{-23} \)) and SLC22A3 (\( P = 3.2 \times 10^{-32} \)). At the known 19q13.33 locus, rs2659124 (\( P = 1.3 \times 10^{-12} \); OR, 1.18) nominally replicated in PEGASUS. A risk score of 105 known risk SNPs was strongly associated with prostate cancer (\( P < 1.0 \times 10^{-9} \)). Comparing the highest to lowest risk score deciles, the OR was 6.22 for non-Hispanic whites, 5.82 for Latinos, 3.77 for African-Americans, and 3.38 for East Asians. In non-Hispanic whites, the 105 risk SNPs explained approximately 7.6% of disease heritability. The entire GWAS array explained approximately 33.4% of heritability, with a 4.3-fold enrichment within DNaseI hypersensitivity sites (\( P = 0.004 \)).

SIGNIFICANCE: Taken together, our findings of independent risk variants, ethnic variation in existing SNP replication, and remaining unexplained heritability have important implications for further clarifying the genetic risk of prostate cancer. Our findings also suggest that there may be much promise in evaluating understudied variation, such as indels and ethnically diverse populations. Cancer Discov; 5(8); 878–91. © 2015 AACR.

INTRODUCTION
Prostate cancer is the second most common cancer diagnosed in men worldwide. Family history of this disease greatly increases risk; roughly 10% to 15% of men with prostate cancer have an affected relative (1, 2), and familial risk increases roughly 2-fold for first-degree male relatives of affected individuals (3). In addition, twin studies indicate that prostate cancer is among the most heritable cancers (4–6). Thus, it is essential to identify genetic risk factors to fully characterize disease burden.

Eight years of genome-wide association studies (GWAS) have identified at least 105 risk variants for prostate cancer (7–28); most of these are common with modest effects, and alone explain little heritability. However, an overall genetic risk score combining these variants could be substantially more predictive of disease and explain a reasonable heritability proportion, although there remain undiscovered loci. How many loci remain and where best to search for them remain unclear.

To further characterize prostate cancer’s genetic basis, we undertook a GWAS using multiple study cohorts within Kaiser Permanente (KP), a fully integrated health care delivery system: the Research Program on Genes, Environment and Health (RPGEH) cohort [database of Genotypes and Phenotypes (dbGaP) phs000674.p1], the ProHealth Study, and the California Men’s Health Study (CMHS; ref. 29). This diverse population has not been included in any prior prostate cancer GWAS, allowing for both novel risk variant discovery and assessment of the replication, prediction, and heritability explained by previously reported risk variants across multiple ethnic populations. We first searched for novel prostate cancer risk SNPs by conducting genome-wide scans using high-throughput genotyping arrays optimized for four major race/ethnicity groups—non-Hispanic white, Latino, East Asian, and African-American (30, 31)—and meta-analyzing the corresponding results. We then tested the novel findings from this GWAS for independent replication in PEGASUS non-Hispanic whites, and African-Americans comprised mostly of the Multiethnic Cohort (MEC; ref. 14). Next, we looked at 105 previously identified risk variants to assess how well they replicate in KP overall and within ethnic subgroups, and
their ability to predict prostate cancer. Finally, we evaluated the heritability explained in the largest race/ethnicity group (non-Hispanic whites) by the known SNPs, how much heritability remains unaccounted for, and which genomic regions are most likely to contain additional risk SNPs.

RESULTS

Study Population

Table 1 presents descriptive information for the KP study by prostate cancer status (by study; Supplementary Table S1): 7,783 men diagnosed with prostate cancer and 38,595 men free from prostate cancer. Approximately 80% were non-Hispanic white, and this ethnic group had a higher case percentage than controls, as did the African-Americans. Most men were older than 60 years, with an increased percentage of cases diagnosed between ages 60 and 70 years. On average, prostate-specific antigen (PSA) levels were much higher in cases than in controls, as PSA was used for annual prostate cancer screening in KP. Most cases had Gleason scores of 6 or 7.

GWAS Findings

Our GWAS discovery stage detected 16 loci containing prostate cancer–associated variants at genome-wide significance (P < 5 × 10^{-8}; Fig. 1 and Supplementary Table S2). The genomic inflation factor was 1.052, suggesting that our findings were not due to systematic bias. Supplementary Fig. S1.1–1.36 show Manhattan and Q–Q plots for the initial analysis of each race/ethnicity group, in addition to further results from analyses conditioning on the genome-wide significant SNPs in KP. Conditional analyses at known loci identified 10 additional independent secondary genome-wide significant variants (Supplementary Table S2; conditioning round given in the first column). Of the 16 original plus 10 conditional genome-wide significant variants, 18 clearly replicated previous GWAS findings (discussed further below). Of the remaining eight variants, four were from our original GWAS: 2q22.3 (rs13016083), 6q25.3 (rs4646284), 14q23.1 (rs34582366), and 19q13.33 (rs2659124); and four were from the conditional analysis: 2p22.3 (rs36004513), 8q24.21 (rs77541621), 9p13.3 (b37.9.33975799), and 12p12.1 (b37.12.25430787; Supplementary Table S2). Although some of these loci contain previously reported prostate cancer risk SNPs, there was generally limited linkage disequilibrium (LD) between them and the eight significant SNPs that we detected. We assessed whether these eight SNPs were independent novel risk hits by undertaking analyses conditional on known risk SNPs at their corresponding loci and by replication in the PEGASUS and MEC studies.

The single base pair indel rs4646284 between SLC22A1 and SLC22A2, at the previously reported 6q25.3 locus, was associated with prostate cancer in the KP, PEGASUS, and MEC populations (overall meta OR, 1.18; P = 1.0 × 10^{-19}; Table 2A). Within KP, the smallest P value was observed among the non-Hispanic whites (P = 6.7 × 10^{-11}), as expected in light of their being the largest race/ethnic group. We observed nominal significance in African-Americans (P = 0.029), and although not statistically significant in Latinos, the estimated magnitude of rs4646284’s effect on prostate cancer was similar to that in non-Hispanic whites and African-Americans; rs4646284 did not exhibit any association among East Asians (Table 2B). This SNP replicated at genome-wide significance in the PEGASUS study of non-Hispanic whites (P = 1.4 × 10^{-8}) and nominally in the MEC study of African-Americans (P = 0.0094).

The indel rs4646284 is in weak LD with two previously reported prostate cancer risk SNPs at 6q25.3: rs9364554 and rs651164 (pairwise r^2 of these three SNPs was less than 0.20 in all race/ethnicity groups, except r^2 = 0.55 in East Asians; see Table 2A). To verify that rs4646284 is associated with prostate cancer independently of these two SNPs, we fit joint (conditional) models containing all three variants. Here, rs4646284 remained strongly associated with prostate cancer overall (OR, 1.16; P = 5.4 × 10^{-10}; Table 2A). The conditional analysis slightly weakened the association for rs9364554 (overall P value went from 6.3 × 10^{-12} to 1.9 × 10^{-5} after conditioning), and completely attenuated the result for rs651165 (overall P value decreased from 1.9 × 10^{-4} to 0.89 with conditioning). A 6q25.3 regional plot of association for rs9364554 and rs651165 for KP shows that rs4646284 was a single hit, with very limited surrounding LD and no other strongly associated risk variants nearby (Fig. 2A; each race/ethnicity group Supplementary Fig. S2.1–2.24).

Table 1. Descriptive factors for KP study population used in genome-wide association study of prostate cancer

<table>
<thead>
<tr>
<th>Race/ethnicity</th>
<th>Cases, N(%)</th>
<th>Controls, N(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Hispanic white</td>
<td>6,406 (82.3%)</td>
<td>30,866 (80.0%)</td>
</tr>
<tr>
<td>African-American</td>
<td>601 (7.7%)</td>
<td>1,650 (4.3%)</td>
</tr>
<tr>
<td>Asian</td>
<td>288 (3.7%)</td>
<td>2,938 (7.6%)</td>
</tr>
<tr>
<td>Latino</td>
<td>488 (6.3%)</td>
<td>3,141 (8.1%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age, y</th>
<th>Cases, N(%)</th>
<th>Controls, N(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age &lt; 50</td>
<td>86 (1.1%)</td>
<td>4,619 (12.0%)</td>
</tr>
<tr>
<td>50 ≤ age &lt; 60</td>
<td>1,224 (15.7%)</td>
<td>7,536 (19.5%)</td>
</tr>
<tr>
<td>60 ≤ age &lt; 70</td>
<td>3,604 (46.3%)</td>
<td>12,240 (31.7%)</td>
</tr>
<tr>
<td>70 ≤ age</td>
<td>2,869 (36.9%)</td>
<td>14,200 (36.8%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PSA*</th>
<th>Cases, N(%)</th>
<th>Controls, N(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (SD)</td>
<td>11.6 (5.4)</td>
<td>2.5 (5.5)</td>
</tr>
<tr>
<td>Median (MAD)</td>
<td>6.4 (3.0)</td>
<td>1.5 (1.2)</td>
</tr>
<tr>
<td>Interquartile range</td>
<td>4.8–9.7</td>
<td>0.8–2.8</td>
</tr>
</tbody>
</table>

Gleason ≤5                     | 167 (2.3%)  | —              |
| 6                             | 3,067 (33.8%)| —              |
| 7                             | 1,641 (33.1%)| —              |
| 8                             | 292 (5.8%)   | —              |
| 9                             | 174 (3.5%)   | —              |
| 10                            | 27 (0.5%)    | —              |

Abbreviation: MAD, median absolute deviation.

*PSA is given as the latest measure before diagnosis for cases, and the latest measure available to us from electronic medical records for controls.
GWAS of Prostate Cancer

Figure 1. Results from a GWAS of prostate cancer in the KP population (8,399 cases and 38,745 controls), highlighting key chromosomal regions. P values are for variant associations with prostate cancer from transethnic fixed-effects meta-analysis of four race/ethnicity GWAS (non-Hispanic white, Latino, African-American, East Asian), each adjusted for age and ancestry principal components. Horizontal dashed lines indicate genome-wide statistical significance ($P < 5 \times 10^{-8}$), Bonferroni-corrected significance for 105 known prostate cancer risk variants ($P < 0.05/105$), and nominal significance ($P < 0.05$). Red points denote our findings for the 105 known risk SNPs, and pink indicates results for SNPs within a 0.5 Mb window around these SNPs. The new 6q25.3 indel rs4646284 is colored magenta, and the 19q13.33 prostate cancer or PSA SNP rs2659124 is noted in green. Those loci containing previously reported variants that we replicated at genome-wide significance are noted in black text. Loci with variants that were novel in the KP GWAS but failed to replicate are noted in gray text.

Table 2. Prostate cancer genome-wide association study results for new risk indel rs4646284 at 6q25.3 and for risk SNP rs2659124 at 19q13.33

| Variant | Risk allele | Ref. allele | Group       | Risk AF | OR (95% CI) | P        | OR (95% CI) | P        | $r^2_{\text{info}}$
|---------|-------------|-------------|-------------|---------|-------------|----------|-------------|----------|------------------
| Indel: rs4646284 | TG (indel) | T           | KP NH white | 0.298   | 1.17 (1.12–1.23) | 6.7 x 10^{-11} | 1.16 (1.10–1.23) | 3.3 x 10^{-7} | 0.91 |
| KP Latino      | 0.250       | 1.13 (0.94–1.36) | 0.20 | 1.06 (0.85–1.31) | 0.61 | 0.89 |
| KP Asian       | 0.324       | 1.03 (0.84–1.27) | 0.76 | 0.91 (0.68–1.23) | 0.56 | 0.89 |
| KP Af-Amer     | 0.365       | 1.18 (1.02–1.37) | 0.029 | 1.21 (1.03–1.42) | 0.020 | 0.85 |
| KP Meta FE     | —           | 1.17 (1.12–1.22) | 5.5 x 10^{-12} | 1.15 (1.09–1.21) | 8.5 x 10^{-8} | — |
| KP Meta RE     | —           | 1.17 (1.12–1.22) | 5.5 x 10^{-12} | 1.15 (1.07–1.22) | 7.7 x 10^{-5} | — |
| MEC            | 0.361       | 1.13 (1.03–1.25) | 0.0094 | 1.13 (1.03–1.25) | 0.014 | 0.81 |
| PEGASUS        | 0.278       | 1.27 (1.17–1.37) | 1.4 x 10^{-8} | 1.20 (1.09–1.32) | 0.00024 | 0.80 |
| Overall Meta FE | —           | 1.18 (1.14–1.22) | 1.0 x 10^{-9} | 1.16 (1.11–1.21) | 5.4 x 10^{-12} | — |
| Overall Meta RE | —           | 1.18 (1.13–1.23) | 9.8 x 10^{-17} | 1.16 (1.11–1.21) | 5.4 x 10^{-12} | — |

(Continued on next page)
## Table 2. Prostate cancer genome-wide association study results for new risk indel rs4646284 at 6q25.3 and for risk SNP rs2659124 at 19q13.33 (Continued)

<table>
<thead>
<tr>
<th>Variant</th>
<th>Risk allele</th>
<th>Ref. allele</th>
<th>Group</th>
<th>OR (95% CI)</th>
<th>P</th>
<th>OR (95% CI)</th>
<th>P</th>
<th>r² info</th>
</tr>
</thead>
<tbody>
<tr>
<td>Previous SNP:</td>
<td>T</td>
<td>C</td>
<td>KP NH white</td>
<td>0.279</td>
<td>1.10</td>
<td>1.05 (1.00–1.10)</td>
<td>0.070</td>
<td>1.00</td>
</tr>
<tr>
<td>rs9364554</td>
<td></td>
<td></td>
<td>KP Latino</td>
<td>0.227</td>
<td>1.16</td>
<td>1.13 (0.93–1.36)</td>
<td>0.21</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>KP Asian</td>
<td>0.319</td>
<td>1.06</td>
<td>1.05 (0.86–1.28)</td>
<td>0.63</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>KP AF-Amer</td>
<td>0.079</td>
<td>1.14</td>
<td>1.11 (0.85–1.43)</td>
<td>0.45</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>KP Meta FE</td>
<td>—</td>
<td>1.10</td>
<td>1.06 (1.01–1.11)</td>
<td>0.019</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>KP Meta RE</td>
<td>—</td>
<td>1.10</td>
<td>1.06 (1.01–1.11)</td>
<td>0.019</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MEC</td>
<td>0.074</td>
<td>1.21</td>
<td>1.19 (1.00–1.40)</td>
<td>0.041</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PEGASUS</td>
<td>0.293</td>
<td>1.22</td>
<td>1.16 (1.07–1.25)</td>
<td>0.00022</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Overall Meta FE</td>
<td>—</td>
<td>1.14</td>
<td>1.09 (1.05–1.13)</td>
<td>1.9 × 10⁻⁵</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Overall Meta RE</td>
<td>—</td>
<td>1.15</td>
<td>1.10 (1.05–1.15)</td>
<td>0.00014</td>
<td>—</td>
</tr>
<tr>
<td>B</td>
<td>A</td>
<td>G</td>
<td>KP NH white</td>
<td>0.308</td>
<td>0.94</td>
<td>1.01 (0.96–1.06)</td>
<td>0.79</td>
<td>1.00</td>
</tr>
<tr>
<td>rs651164</td>
<td></td>
<td></td>
<td>KP Latino</td>
<td>0.380</td>
<td>0.88</td>
<td>0.91 (0.76–1.09)</td>
<td>0.29</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>KP Asian</td>
<td>0.515</td>
<td>0.91</td>
<td>0.85 (0.65–1.11)</td>
<td>0.24</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>KP AF-Amer</td>
<td>0.264</td>
<td>1.02</td>
<td>1.08 (0.93–1.23)</td>
<td>0.30</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>KP Meta FE</td>
<td>—</td>
<td>0.94</td>
<td>1.00 (0.93–1.05)</td>
<td>0.96</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>KP Meta RE</td>
<td>—</td>
<td>0.94</td>
<td>1.00 (0.93–1.05)</td>
<td>0.96</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MEC</td>
<td>0.307</td>
<td>0.98</td>
<td>1.03 (0.93–1.13)</td>
<td>0.60</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PEGASUS</td>
<td>0.306</td>
<td>0.91</td>
<td>0.99 (0.92–1.07)</td>
<td>0.82</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Overall Meta FE</td>
<td>—</td>
<td>0.94</td>
<td>1.00 (0.97–1.04)</td>
<td>0.89</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Overall Meta RE</td>
<td>—</td>
<td>0.94</td>
<td>1.00 (0.97–1.04)</td>
<td>0.89</td>
<td>—</td>
</tr>
</tbody>
</table>

### A

Effective risk allele frequency (AF): KP NH white, 0.854; KP Latino, 0.809; KP Asian, 0.587; KP AF-Amer, 0.847; KP Meta FE, 0.307; KP Meta RE, 0.307; MEC, 0.839; PEGASUS, 0.868; Overall Meta FE, 0.827; Overall Meta RE, 0.827.

### B

Effective risk allele frequency (AF): KP NH white, 0.150; KP Latino, 0.214; KP Asian, 0.419; KP AF-Amer, 0.302; KP Meta FE, 0.311; KP Meta RE, 0.140; MEC, 0.880; PEGASUS, 0.880; Overall Meta FE, 0.880; Overall Meta RE, 0.880.

**NOTE:** These variants were detected in the KP study population and tested for replication in the PEGASUS and MEC cohorts.

A. Single variant and conditional analyses for 6q25.3 indel rs4646284 and the previous prostate cancer risk SNPs rs9364554 and rs651164. Meta-analysis is given for the fixed effects (FE) and random effects (RE) analysis. Chromosome positions are given in b37. The risk allele frequency (AF) is the pooled frequency across both cases and controls. Pairwise correlations of these three SNPs in non-Hispanic whites, Latinos, East Asians, and African-Americans, respectively (KP data) r² info = 0.50, 0.58, 0.85, and 0.055.

B. Single variant and conditional analyses for 19q13.33 SNP rs2659124 detected here and the previous prostate cancer risk SNP rs2735839. Pairwise correlations of these two SNPs in non-Hispanic whites, Latinos, East Asians, and African-Americans, respectively (KP data) r² info = 0.50, 0.58, 0.85, and 0.055.

Abbreviation: CI, confidence interval.
GWAS of Prostate Cancer

Figure 2. Regional fixed-effects meta-analysis plots from GWAS in KP population of the two risk variants that replicated: (A) 6q25.3 (rs4646284), (B) 19q13.33 (rs2659124). The color code for the points represents the $r^2$ of each SNP with the risk variant (ranges defined in the legend). The dotted vertical lines pass through the risk variants and the other significant SNPs in the region, on which analyses were conditioned.
Our cis-expression quantitative trait loci (eQTL) analysis detected an association between rs4646284 and decreased expression of SLC22A1, SLC22A2, and SLC22A3 in prostate tissue (Supplementary Table S3). In data from the Mayo Clinic, we observed extremely significant associations between the rs4646284 insertion and lower expression of SLC22A1 (effect size = -0.42; \( P = 1.3 \times 10^{-10} \)) and SLC22A3 (coefficient = -0.68; \( P = 3.2 \times 10^{-12} \)). Regional eQTL analysis of all variants within 1.1 Mb of these two genes—including previously reported prostate cancer risk SNPs—indicated that the rs4646284 indel was clearly the strongest predictor of expression at SLC22A1 and SLC22A3 in the Mayo Clinic samples, with eQTL \( P \) values for surrounding SNPs orders of magnitude larger (Supplementary Fig. S1–3.2). Our replication analyses in the Physicians’ Health Study plus Health Professionals Follow-up Study (PHS+HPFS) normal/tumor tissues and The Cancer Genome Atlas (TCGA) tumor tissues also showed reduced expression in SLC22A1 (overall coefficient = -0.07; one-sided \( P = 0.046 \)) and even more so in SLC22A3 (overall coefficient = -0.32; one-sided \( P = 0.0012 \)). The PHS+HPFS normal tissue did not perform as well as SLC22A1 (90% expression mark = 3.4, where >5 is desirable). This allele may also be associated with lower expression of lipoprotein(A)-like 2 (LPAL2), a pseudogene structurally similar to the gene that encodes lipoprotein(A) (LPA), but produces mRNA with a premature stop codon (32). Taken together, these results indicate rs4646284 is an independent risk indel for prostate cancer that improves upon the previously reported findings for the 6q25.3 locus.

Another genome-wide significant SNP in KP, rs2659124 at 19q13.33, replicated in the PEGASUS study. This SNP is near the 5’ untranslated region (UTR) terminus of kallikrein-related peptidase 3 (KLK3), which encodes PSA. The KP meta-analysis yielded \( P = 1.9 \times 10^{-15} \), with similar ORs observed in the non-Hispanic whites, African-Americans, and Latinos (Table 2B). rs2659124 replicated in PEGASUS with \( P = 0.0027 \) (overall pooled \( P = 1.3 \times 10^{-15} \); OR, 1.18). The rs2659124 association was slightly attenuated but remained suggestive (pooled \( P = 4.3 \times 10^{-10} \)) after adjusting for the previously known risk SNP at 19q12.33 (rs2735839). In contrast, the conditional analysis completely attenuated the rs2735839 association with prostate cancer (Table 2B). Figure 2B gives a regional plot for this locus, showing that rs2659124 is the strongest risk SNP.

The remaining six genome-wide significant SNPs in our cohort did not clearly replicate in PEGASUS or MEC (Supplementary Table S2). With regard to their associations with prostate cancer, rs13016083 in 2q22.3 had KP meta-analysis OR = 1.13, but PEGASUS OR = 1.03 and MEC OR = 0.99. For the SNP 9:33975799, the KP OR = 0.86 but PEGASUS OR = 1.05 and MEC OR = 0.94. rs34582366 had KP OR = 1.32, but PEGASUS and MEC had ORs in the opposite direction (0.90 and 0.96, respectively). The SNP 12:25430787 had KP OR = 0.74, but the effect was in the opposite direction in PEGASUS, OR = 1.10; this SNP had too low an allele frequency among African-Americans to be tested in MEC. rs360054513 had KP OR = 0.58 but MEC OR = 0.98, and was too rare in non-Hispanic whites to test in PEGASUS. Finally, rs77541621 remained significant after conditioning on all 12 8q24 loci in KP non-Hispanic whites (OR, 0.61; it was very rare in African-Americans and Asians). In the PEGASUS cohort, it was significant genome-wide before conditioning (OR, 0.51; \( P = 8.6 \times 10^{-10} \)), and was nominally significant after adjusting for the 12 8q24 loci (OR, 0.68; \( P = 0.02 \)). However, because we conditioned on 12 SNPs here, some of which were imputed, we suspect this result may reflect incomplete tagging of existing 8q24 loci.

**Replication of Known GWAS Results**

The remaining 18 of 26 genome-wide significant associations were clear replications of previously reported findings at 2p11.2, 3p11.2, 4q24, 6q22.1, 7p15.2, 8p21.2, 8q24.21, 10q11.23, 11q13.3, 12q13.12, 12q13.13, 17q12, 17q24.3, 19q13.33, and 22q13.2 (Fig. 1; refs. 7–9, 9–16, 24–28). Our 18 lead risk SNPs did not appear to improve upon or exhibit independence from those known. Supplementary Table S2 gives the correlations between these SNPs in non-Hispanic whites.

Supplementary Table S4 presents ORs and \( P \) values for the associations between prostate cancer and these 18 SNPs plus the known variants (105 SNPs total) from the previous reports (16) and from KP. These 105 variants exhibited high replication based on the magnitude and direction of their associations with prostate cancer. In particular, we observed clear agreement between the ORs for variant associations from previous reports and those from the KP GWAS (meta-analysis in Fig. 3; each race/ethnicity group in Supplementary Fig. S4). A large majority of the ORs were larger in previous studies than observed in ours; nevertheless, we saw generally high agreement (Fig. 3). Moreover, within ethnic groups, 99 of 103 of non-Hispanic white, 67 of 99 of East Asian, 79 of 102 of Latino, and 78 of 105 of African-American ORs were in the same direction as previously found (i.e., ORs > 1.0). Only four of the variants with ORs < 1.0 were statistically significant: three for Asians and one for African-Americans.

The meta-analysis \( P \) values were less than 0.05 for 66 of the 105 SNPs (62.9%) and less than the Bonferroni-corrected \( \alpha \) level of 0.00048 = 0.05/105 for 33 (31.4%). When stratified by ethnic group, most SNPs that had \( P \) values less than these \( \alpha \) levels were associated in non-Hispanic whites, as expected because this population has been most commonly used for discovery efforts and is the largest ancestry group in the KP study. In particular, the known risk variants had \( P < 0.05 \) for 62 SNPs (59.0%) in non-Hispanic whites, 11 (10.5%) in Latinos, 13 (12.4%) in East Asians, and 13 (12.3%) in African-Americans. The known risk variants had \( P < 0.00048 \) for 30 SNPs in non-Hispanic whites, zero in Latinos, one (1.0%) in East Asians, and two (1.9%) in African-Americans.

**Prediction with GWAS SNPs**

We used risk profile scoring to assess the predictive value of the 105 known GWAS risk SNPs for prostate cancer in KP (which is independent of the populations in which these risk SNPs were discovered). We combined the 105 SNPs into a single score by applying the previously estimated ORs to our genotype data (details in Methods). The risk score was highly statistically significant for all four major ethnic groups: non-Hispanic white \( P = 1.0 \times 10^{-211} \), Latino \( P = 3.5 \times 10^{-13} \), East Asian \( P = 1.0 \times 10^{-10} \); and African-American \( P = 1.1 \times 10^{-15} \). To see how the magnitude of the associations with risk scores varied across race/ethnic groups, we calculated prostate cancer ORs corresponding to increasing deciles of
GWAS of Prostate Cancer

**Figure 3.** Comparison of variant associations from previous reports and results from the KP GWAS meta-analysis for 105 known prostate cancer risk SNPs. Plotted values are the log odds ratios from the previous and current studies. Error lines denote the 95% CIs for the respective studies, colored by chromosome. Blue line is the diagonal, showing generally high correlation between the previous and new odds ratios. The SNP rs116041037 appears to be an African-American–specific SNP and is an outlier so we left it off of the plot to more easily view the other points; the SNP had a previous OR = 2.45 (95% CI, 1.65–3.62), and KP OR = 2.67 (1.96, 3.64).

Risk scores, using the lowest decile as the referent (Fig. 4). All four race/ethnic populations exhibited clear trends of increasing risk score ORs across increasing deciles, but the non-Hispanic white and Latino groups had substantially higher ORs than the African-Americans, and the East Asians always had the lowest ORs (Fig. 4). Comparing the highest to lowest risk score deciles, the association with prostate cancer for non-Hispanic whites gave an OR = 6.22 [95% confidence interval (CI), 5.38–7.19]; for Latinos the OR = 5.82 (95% CI, 3.36–10.1); for African-Americans the OR = 3.77 (95% CI, 2.34–6.08); and for East Asians the OR = 3.38 (95% CI, 1.91–5.97). Although the risk score was highly predictive across...
groups, it was much less predictive for the African-Americans and East Asians. This may, in part, reflect lower LDs or different allele frequencies in these ethnic groups.

We also wanted to see how well we could predict prostate cancer using the genetic risk score versus other covariates [i.e., body mass index (BMI), age, ancestry principal components (PC)]. For each race/ethnicity group, we split the cohort in half for training/testing to estimate the non-genetic covariate coefficients. Results for different combinations of genetic risk score and covariates are shown in Supplementary Fig. S5. Relative to the other covariates alone, we generally observed an increase in AUC of approximately 5% with the genetic risk score. Supplementary Table S5 presents the variance in prostate cancer explained by BMI, age, and ancestry covariates, compared with that explained by also including the risk score. For non-Hispanic whites, including the risk score increased the variance explained from 0.077 to 0.127. The increase was similar in the other groups, though the overall variance explained was lowest in African-Americans. Ignoring the covariates and restricting the risk score to only include those SNPs with nominal (P < 0.05), replication-wide (P < 0.00048), and genome-wide (P < 5 × 10⁻⁸) significance level in the non-Hispanic whites gave variance explained of 0.122, 0.122, and 0.114, respectively.

**Heritability: GWAS Array and Functional Categories**

We calculated the narrow sense heritability (i.e., the additive genetic component of the phenotypic variance) explained by variants typed and tagged by the GWAS array in the non-Hispanic whites. The estimated heritability for the genotyped SNPs was $h^2 = 0.201$ (standard error, SE = 0.041). When including the imputed SNPs, this substantially increased to $h^2 = 0.334$ (SE = 0.060). When we partitioned the data into the 105 previously known genotyped SNPs versus the rest of the genome, the heritability estimates were 0.076 (SE = 0.012) and 0.215 (SE = 0.058), respectively. These approximately sum to the entire study’s heritability, and are close to the increase in variability explained.

We then calculated the narrow sense heritability explained by variants within functional regions of the genome, and estimated whether certain regions explained a disproportionate amount of this heritability in comparison with their size (33). We found that the DNaseI hypersensitivity sites (DHS) exhibited 4.3-fold increased enrichment ($P = 0.0039$) and the intergenic regions had a 0.2-fold decreased enrichment ($P = 0.0058$; Table 3). We also found nonsignificant enrichment in coding (4.6-fold), UTR (3.7-fold), and intronic (0.4-fold) regions (Table 3). These results suggest that the SNPs underlying prostate cancer are more likely to be located in the coding, DHS, and UTR regions, less likely to be located in the promoter regions, and least likely to be located in the intronic and intergenic regions (Table 3). Looking at the previously reported risk SNPs (102 autosomal), we found a similar but slightly weaker pattern of enrichment: coding, 2.4-fold; UTR, 0.7-fold; DHS, 2.5-fold; intron, 0.7-fold; and intergenic, 0.7-fold. For the two key variants reported here, rs4646284 is in a DHS and rs2659124 is intergenic. Note that the smaller percentage of the genome in the promoter category observed here, in contrast with ref. 33, reflects our use of a more recent promoter definition from the Eukaryotic Promoter Database new v003 (34).

**DISCUSSION**

In this large, ethnically diverse, and previously unstudied cohort, we detected two risk variants for prostate cancer: an indel (rs4646284 at 6q25.3) and a SNP that may be involved with PSA (rs2659124 at 19q13.33). We also replicated a large majority of the known risk SNPs, which taken together as a risk score in a polygenic model were very strongly associated with prostate cancer, albeit with substantial variation across ethnic groups. In addition, we estimated that approximately 65% of the heritability assayed by the GWAS array remains unexplained by the 105 known risk SNPs, and that heritability is overall enriched in the DHS regions. This indicates that substantial genetic variation in prostate cancer remains to be uncovered.

The novel intergenic indel we identified (rs4646284) is in a recombination hotspot, and is positioned roughly 1.8 kb downstream of the 3′-UTR terminus of SLC22A1 and 56 kb upstream from SLC22A2. Previous work identified other prostate cancer susceptibility variants at 6q25.3, including rs9364554 within intron 5 of SLC22A1 (9) and, 250 kb upstream, rs651164 outside

---

**Table 3. Array heritability enrichment of the coding, UTR, promoter, DHS, intronic, and intergenic partitions for non-Hispanic whites using the imputed data**

<table>
<thead>
<tr>
<th>Region</th>
<th>Univariate $h^2$ (SE)</th>
<th>Multivariate fit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$h^2$ (SE)</td>
<td>% $h^2$</td>
</tr>
<tr>
<td>Coding</td>
<td>0.029 (0.014)</td>
<td>0.013 (0.014)</td>
</tr>
<tr>
<td>UTR</td>
<td>0.048 (0.019)</td>
<td>0.016 (0.020)</td>
</tr>
<tr>
<td>Promoter</td>
<td>0.001 (0.003)</td>
<td>0.001 (0.003)</td>
</tr>
<tr>
<td>DHS</td>
<td>0.299 (0.049)</td>
<td>0.243 (0.065)</td>
</tr>
<tr>
<td>Intron</td>
<td>0.193 (0.045)</td>
<td>0.039 (0.055)</td>
</tr>
<tr>
<td>Intergenic</td>
<td>0.199 (0.051)</td>
<td>0.030 (0.060)</td>
</tr>
<tr>
<td>Total</td>
<td>0.334 (0.060)</td>
<td>0.342 (0.060)</td>
</tr>
</tbody>
</table>

Note: Results are from fitting each partition separately (univariate), and then subsequent results are from the joint multivariate fit, Kaiser Permanente Study Population.
GWAS of Prostate Cancer

we relaxed the SNP inclusion criteria and incorporated larger numbers of variants: we found that the rs4646284 SNP was more strongly associated with decreased expression of SLC22A1 and SLC22A3, substantially more than any other variant at the 6q25 locus, including the two previous GWAS-identified prostate cancer risk SNPs. This rs4646284 SNP is within binding site signals from ENCODE chromatin immunoprecipitation sequencing data for several transcription factors, and appears closer than rs651164 to a local binding site peak for c-MYC across several cell lines and to peaks for CCCTC-binding factor (CTCF) in both androgen-treated and androgen untreated prostate adenocarcinoma (LNCaP) cells. Although rs4646284 is a common (insertion frequency ~30% in non-Hispanic whites) and highly statistically significant risk variant, previous work may not have detected this indel because it is not in HapMap and is in low LD with neighboring SNPs, making it difficult to cover with early-generation genotyping arrays. Our finding highlights the importance of studying indels in GWAS and that these can be imputed with high confidence with appropriate reference panels.

The suggestive SNP rs2659124 is 3.3 kb from the 5′-UTR terminus of KLK3, which encodes PSA and is one of 15 kal-likrein genes located sequentially on 19q13.33-41. rs2659124 is 10 kb away from rs2735839, a previously reported risk SNP for both prostate cancer (9) and PSA levels (35, 36). rs2659124 and rs2735839 are on the opposite sides of KLK3, and the latter SNP is <1 kb from the 3′-UTR terminus (r² = 0.50 between these SNPs in KP non-Hispanic whites). There is evidence that the association between rs2735839 and prostate cancer risk may depend on PSA levels (37) or disease aggressiveness (35, 38). rs2659124 was previously reported as part of a fine mapping study of the KLK3 locus conducted among men from the Cancer Genetic Markers of Susceptibility Project (CGEMS)—which is included in the PEGASUS sample—with nominal statistical significance (P = 0.02; ref. 35).

We replicated a large proportion of the 105 known risk SNPs, especially when considering the ancestry group in which the SNPs were discovered. The “winner’s curse,” that effect sizes are often larger in the populations in which they are discovered, may be one reason why some SNPs failed to replicate, and why ORs were generally smaller in our cohort than previously found (39). It is also possible that control misclassification yielded more conservative estimates at these and all SNPs. Nevertheless, the aggregate risk scores combining information across all 105 known SNPs were highly significant in all ethnic groups, although the magnitude of this association varied substantially by ethnicity. The variance in prostate cancer risk explained by SNPs in KP increased as the SNP inclusion criteria and incorporated larger numbers of variants: r² = 0.112 for the risk SNPs exhibiting genome-wide significance; r² = 0.122 for those replicating; and r² = 0.127 for all 105 previously reported risk SNPs. We also showed an increase of roughly 5% in the AUC when adding a genetic risk score to BMI, age, and PCs of ancestry. Finally, we showed that these aggregate risk scores had large ORs when comparing the upper to lower deciles. In these results, East Asians and African-Americans had much lower ORs for the highest decile than the non-Hispanic whites and Latinos. This may be due to different allele frequencies, LD patterns, or causal risk alleles in these populations, because the previous discovery cohorts were largely of European ancestry (Supplementary Table S4). For example, African-Americans and East Asians have higher average absolute risk allele frequency differences from non-Hispanic whites (0.16 and 0.15, respectively) than Latinos (0.05). As more information becomes available for ethnic-specific prostate cancer risk SNPs, ethnic-specific risk scores should improve prediction.

As expected, our estimated heritability among non-Hispanic whites (0.334; SE = 0.060) was lower than that from twin studies (0.58; ref. 5), but higher than previous array heritability estimates (0.204; SE = 0.056) that used a slightly smaller prostate cancer lifetime risk of 0.09 (40). The heritability estimate for the 105 known risk SNPs was 0.077 (SE = 0.058); there is still an estimated array heritability of 0.215 (SE = 0.058) that remains unexplained. Our known risk SNP heritability estimate is unbiased from the winner’s curse because these SNPs were previously discovered. This heritability appears enriched within the DHS regions, and possibly in the coding, UTR, and promoter regions. This enrichment result pattern is roughly similar to those found for the non-autoimmune phenotypes in ref. 33. Further refining these regions may be a promising area of future research.

In summary, we were able to detect new risk variants for prostate cancer, confirm many previously reported associations at various levels across ethnic groups, and provide independent evidence that additional risk SNPs are still to be found. Because a large amount of narrow-sense genetic array heritability remains to be explained, larger analyses or meta-analyses may uncover further genetic variants associated with disease. Additional advances may be possible by applying to existing data novel analytic approaches such as Bayesian models that incorporate local heritability estimates or prior biologic knowledge, or by undertaking scans for pleiotropic effects that leverage data across multiple cancers. In addition, we showed that the existing GWAS results are robust and predictive of prostate cancer risk, which may have important implications for using risk SNPs to guide individualized screening for this common but complex disease.

METHODS
Participants, Phenotype, and Genotyping

Our primary analyses used cases and controls from three KP studies: RPGEH, ProHealth, and CMHS. RPGEH participants included men in the RPGEH Genetic Epidemiology Research on Aging (GERA) cohort (dbGaP phs000674.v1.p1), as well as prostate cancer cases with a DNA sample in the RPGEH biorepository but who were not part of GERA. These studies have been previously described (dbGaP phs000674.v1.p1; refs. 29–31, 41). The ProHealth study focused on ascertaining KP Northern California African-American prostate cancer cases. The CMHS is a prospective cohort study of KP California men. Prostate cancer cases in these cohorts were identified from the KP Northern California Cancer Registry (KPNCCR), the KP Southern California Cancer Registry (KPSCCR), or through review of clinical electronic health records through the end of 2012. The cancer registries capture data on all prostate cancer cases newly diagnosed or treated at KP facilities. The cancer registries conform to standards...
of the North American Association of Central Cancer Registries and the National Cancer Institute’s Surveillance, Epidemiology and End Results (SEER) program. Controls were all men in GERA without prostate cancer diagnosis, who could have had other cancers. Our analyses included 7,783 cases and 38,595 controls after exclusions described below. These men were genotyped at over 650,000 SNPs on four race/ethnicity–specific Affymetrix Axiom arrays optimized for individuals of European (EUR), African-American (AFR), East Asian (EAS), and Latino (LAT) race/ethnicity. Specific details of array designs, including estimated genome-wide coverage, have been previously published (30, 31). Briefly, the proportion of common [minor allele frequency (MAF) > 0.05] 1000 Genomes SNPs (Interim Phase I release, 1,092 subjects) covered by the genome-wide array with r² > 0.8 equaled 0.93, 0.89, 0.93, and 0.95 for non-Hispanic whites, African-Americans, Asians, and Latinos, respectively; the proportion of less common (0.01 < MAF ≤ 0.05) variants covered was 0.73, 0.65, 0.80, and 0.80, respectively. The arrays were designed using human genome build 36, but the probesets were remapped to build 37, which was used for all SNP locations reported here. For the EUR and EAS arrays, we used dbSNP v130, and for the LAT and AFR arrays, we used dbSNP v132. All of the EUR and most of the EUR arrays (more details below) were processed using the Axiom v1 reagent kit, and the other arrays were processed on the Axiom v2 reagent kit. We note that a small number of participants in CMHS from Southern California overlap with MEC and were genotyped with the AFR array through a GWAS of the association between the true and imputed genotype (48). Our GWAS analysis used 10,109,774; 9,283,528; 10,776,138; and 17,141,436 SNPs with r² ≥ 0.3 and MAF≥ 0.01 for non-Hispanic white, East Asian, Latino, and African-American men, respectively (19,977,088 unique SNPs).

GWAS Preimputation Quality Control
Genotype quality control (QC) procedures for the original GERA cohort assays were performed on an array-wise basis, as described previously (dbGaP phs000674.v1.p1; ref. 43). This process included ProHealth, CMHS, and additional genotyped individuals not in the originally genotyped GERA cohort. Because we genotyped an additional set of individuals who were much more enriched for cases than the original GERA cohort, we also controlled for potential batch effects by removing SNPs with MAF < 0.01, call rate < 95%, or Hardy–Weinberg equilibrium (HWE) P value among homogeneous control groups < 1 × 10−5. This completed QC for EUR and LAT, resulting in 653,943 and 678,790 SNPs, respectively. EUR individuals were genotyped using the same reagent kit, as were LAT individuals. In addition, the cases and controls were randomly distributed among the genotype packages to control for any potential batch effects.

On the EUR array, 3,843 men were run on a different reagent kit (Axiom v2 versus v1). To adjust for any potential kit effects, we undertook GWAS of the association between each SNP and reagent kit separately among cases and controls, adjusting for PCs. We removed kit-associated SNPs (P < 1 × 10−5). We also genotyped each of the new prostate cancer sample plates (those not genotyped with the original set of GERA individuals) with 12 other plates from the originally genotyped GERA cohort, and removed SNPs with >13/12,628 (1%) mismatches. This resulted in 604,255 SNPs. Similarly, we addressed potential plate batch issues for 1,308 men genotyped with the AFR array through a GWAS of the association between SNPs and whether a subject was typed in the originally genotyped GERA cohort versus later in additional prostate cancer batches separately among cases and controls, adjusting for PCs. Here, we removed batch-associated SNPs at P < 0.05. We used a stronger batch filter threshold on the AFR array than on the EUR array because fewer individuals were analyzed on the AFR array, resulting in lower power to detect batch effects. As before, we also genotyped the individuals in packages with just the new plates, re-genotyped them with some of the previous plates, and removed SNPs with >2/78 (2.6%) mismatches. This resulted in 568,496 SNPs.

We were able to accurately impute (see below) many SNPs removed by the QC steps. Specifically, of those genotyped SNPs that failed QC (MAF > 0.01), we imputed with accuracy r² ≥ 0.3 (and r² ≥ 0.8) a total of 58,333/63,865 or 91.3% (82.3%), 44,390/51,180 or 87.5% (77.6%), 30,273/35,765 or 84.6% (58.2%), and 305,303/312,202 or 97.8% (91.5%) of the SNPs among non-Hispanic whites, Latinos, East Asians, and African-Americans, respectively. The larger decrease in coverage with higher r² values for East Asians may reflect having designed the array with only a greedy SNP selection as opposed to the hybrid greedy/imputation-based approach of the AFR and LAT arrays. Although the EUR array was also designed with exclusively greedy SNP selection, this population may have higher r² than EUR because of a larger sample size (giving more accurate phasing) and stronger LD.

GWAS Genomic Imputation
Imputation was also performed on an array-wise basis. First genotypes were pre-phased with SHAPEIT v2.5 (44), including cryptic relatives to improve phasing. Vectors were then imputed from the 1000 Genomes Project October 2014 release with 2,504 samples with singletons removed (which impute poorly) as a cosmopolitan reference panel with IMPUTE2 v2.3.1 (45–47). The estimated QC r² value metric given here is from IMPUTE2, which estimates the correlation between the true and imputed genotype (48). Our GWAS analysis used 10,109,774; 9,283,528; 10,776,138; and 17,141,436 SNPs with r² ≥ 0.3 and MAF≥ 0.01 for non-Hispanic white, East Asian, Latino, and African-American men, respectively (19,977,088 unique SNPs).

GWAS Analysis and Covariate Adjustment
We first analyzed each of the four race/ethnicity groups (non-Hispanic white, Latino, East Asian, and African-American) separately. Within these groups, each SNP was modeled using additive dosages to account for imputation uncertainty, which works well in practice (49). Each SNP association with prostate cancer was tested via a logistic regression model adjusting for age, body mass index, and ancestry (described below). Age is given at diagnosis for cases, and at last PSA measured for controls. For computational efficiency, we initially regressed the phenotype on all covariates excluding the SNP. We then computed the sum of the estimated β coefficients times the original covariates to create a single covariate, and tested each SNP in a logistic regression model with this single covariate.

To adjust for genetic ancestry, we performed PC analysis using Eigenstrat v4.2 (50) on each of the four race/ethnicity subgroups. We used a subset of 28,174 SNPs with CR > 99% common to all arrays (dbGaP phs000674.v1.p1; ref. 51). For the largest race/ethnicity group (non-Hispanic whites), we performed the PC analysis on 20,000 random individuals, projecting the remaining individuals into the same space, as has been shown to work very well in practice (dbGaP phs000674.v1.p1; ref. 51). The PC analyses are nearly identical to that previously shown for GERA (dbGaP phs000674.v1.p1; ref. 51). The top 10 eigenvectors were included in each logistic regression model. The genomic inflation factor (52) was very modest for all GWAS analyses (all ≤ 1.065; exact values given for each analysis in Supplementary Fig. S1.1–1.36).

We undertook both random effects (RE) and fixed effects (FE) meta-analysis to combine the results of the four race/ethnicity groups using Metasoft (53). Then we assessed whether conditioning on our observed genome-wide significant results highlighted additional significant findings. We used results from the meta-analysis to group clumps of genome-wide significant SNPs (P < 5 × 10−8) within 1 Mb of another GWAS significant SNP. We chose the most significant SNP in each clump, and completely reran the full genome-wide analysis,
GWAS of Prostate Cancer

adjusting for these SNPs. We iterated these conditional analyses until no additional genome-wide significant SNPs were found. We also looked for additional independent SNPs at loci that were previously known to be associated with prostate cancer. We searched in a 1-Mb window around known SNPs at 87 loci, constituting 2.9% of the genome, and so adjusted for multiple comparisons by \(5.8 \times 10^{-8} \times 2.9\% = 2.1 \times 10^{-8}\), but no additional loci were found.

Replication of SNPs in PEGASUS and MEC Cohorts

To determine whether any of the eight new genome-wide significant prostate cancer associations from KP replicated, we evaluated them using independent data from PEGASUS and what we refer to as MEC, which consists mostly of MEC men plus men from other African-American studies (ref. 14; dbGaP phs000306.v3.p1). PEGASUS included 4,599 prostate cancer cases and 2,940 controls of non-Hispanic white race/ethnicity, genotyped using Illumina HumanOmni2.5 and imputed using 1000 Genomes Project Phase I data (1,092 individuals). The PEGASUS replication analyses adjusted for statistically significant ancestry PCs. MEC included 2,265 cases and 2,414 controls of African-American race/ethnicity, genotyped using the Illumina Infinium 1M Duo (dbGap phs000306.v3.p1) and imputed using the same 1000 Genomes reference panel used by KP. The MEC replication analyses adjusted for the first 10 ancestry PCs. All individuals from these replication cohorts were independent of KP.

Confirmation of Imputed rs4646284 Indel

We used two approaches to validate that we correctly imputed the rs4646284 variant. First, we TaqMan genotyped (Life Technologies) 352 individuals who were also genotyped on our EUR array and imputed with our EUR individuals. We computed the correlation with the imputed genotype, and a bias-corrected and accelerated bootstrap CI (99,999 iterations). The genotyped indel showed high agreement with the imputed indel \(r^2 = 0.81\; 95\%\; CI, 0.75-0.86\). Second, we subsetted the 1000 Genomes Project data to the SNPs on the EUR array, and imputed the rs4646284 indel in a leave-one-out manner as described in ref. 31, using the 1000 Genomes Project data as a reference. We then computed the correlation between what was genotyped in 1000 Genomes and this imputed value. This also exhibited high agreement between the actual and imputed indel genotypes \(r^2 = 0.84\).

eQTL Analysis of Indel and 6q25 Locus

We evaluated the potential effect of the novel risk indel rs4646284 on expression of neighboring genes and pseudogenes (IGF2R, LOC729603, SLC22A1, SLC22A2, SLC22A3, LPAL2, and LPA4) in normal and cancerous prostate tissue. This eQTL analysis was undertaken in three studies. First, the Mayo Clinic study included normal prostate tissue from 471 men with Gleason \(\leq 7\) disease undergoing radical prostatectomy or cystoprostatectomy (54). Surgical hematoxylin and eosin (H&E) sections from fresh-frozen materials were reviewed to identify normal (necancerous) tissue samples, and RNA sequencing data were obtained with an Illumina HiSeq 2000. The second eQTL analysis included prostate tumor tissue samples from 99 men and normal prostate tissue from 56 men with incident prostate cancer who participated in the PHS and HPFS (NCBI GEO GSE62872; ref. 55) using genotype data from the Breast and Prostate Cancer Cohort Consortium (BPC3) aggressive prostate cancer GWAS (23). Prostate tissue was collected from archival transurethral resection or prostatectomy specimens (formalin-fixed, paraffin-embedded). mRNA expression was assayed using the Affymetrix GeneChip Human Gene 1.0 ST microarray. The third study included prostate tumor tissue from 128 participants in the TCGA (NCBI GEO GSE21032; ref. 56) using Illumina HiSeq 2000 mRNA expression data and genotypes from matched normal using the Affymetrix SNP 6.0 array (57). All three studies successfully imputed the rs4646284 indel for the eQTL analyses (Mayo \(r^2\) assoc = 0.71; PHS \(r^2\) info = 0.87; HPFS \(r^2\) assoc = 0.89; TGCA \(r^2\) assoc = 0.84), and used linear regression on the dosage.

Replication Analysis of Previously Detected SNPs

To determine whether the 105 previously reported variants associated with prostate cancer replicated in our cohort, we used a nominal significance level (0.05) and a Bonferroni-corrected \(\alpha\) level of 0.05/105 = 0.00048.

Risk Scores

We constructed a risk score for each man in the KP study by summing up the additive coding of the SNPs previously associated with prostate cancer weighted by the previously reported log(OR) from ref. 16. In the non-Hispanic white, Latino, and African-American race/ethnicity groups, no two SNPs had \(r^2 > 0.3\), so all were used. In the East Asian race/ethnicity group, rs16041037 and rs7210100 had \(r^2 = 0.87\), and rs1016343 and rs6983562 had \(r^2 = 0.55\); the latter SNP was removed in both circumstances when computing the risk score. To estimate the variance explained by these risk scores, we report the Nagelkerke pseudo-\(R^2\) estimate (58).

GWAS Array Heritability

We estimated the additive array heritability using Genome-wide Complex Trait Analysis (GCTA) v1.24 (59). Array heritability estimates can be more sensitive to artifacts than GWAS results (59). Thus, to limit any potential batch effects, we restricted this analysis to the homogeneous group of 30,598 non-Hispanic white men (3,605 cases and 26,993 controls) genotyped with the EUR array with Asniom v1 reagent kit. We also used a stronger set of filters than used in the GWAS (59). Specifically, in addition to the filters noted above, we excluded SNPs with \(HWE P < 0.05\) (in controls); significant differences in case-control missigness \(P < 0.05\); and absolute MAF differences \(> 0.15\) compared with 1000 Genomes Project European ancestry individuals. We also removed 22 outlier individuals who were outside of five standard deviations of the first two PCs, and eliminated individuals such that there were no pairwise relationships with estimated kinship \(> 0.05\). We used only the autosomes, as is commonly done for estimating heritability. Finally, we LD-filtered the SNPs such that no two SNPs had a pairwise \(r^2 > 0.8\). This resulted in 26,226 individuals (3,143 cases and 23,083 controls), 402,748 genotyped SNPs, and 2,184,083 imputed SNPs with \(r^2\) assoc \(\geq 0.3\) and MAF \(\geq 0.01\) used in the analysis. We assumed a population prevalence of prostate cancer = 0.12 in the GCTA liability threshold model.

We further partitioned the genome into several sets using a joint variance components fit in GCTA (60). We first tested the previously known hits versus the rest of the genome, and then partitioned the functional categories, prioritized similarly to ref. 33: coding, UTR, and intron, and intergenic. SNPs that happened to fall in overlapping regions were assigned to the highest priority category. The coding, UTR, and intron regions were determined from the UCSC Genome Browser known gene database (61). Unlike Gusev and colleagues (33), who defined the promoters as ±2 kb of a transcription start site, we defined the promoters from the Eukaryotic Promoter Database new v003 (34). The DEGs were determined from ref. 33. For this partitioning, we defined enrichment for each category as the percentage of heritability explained divided by the percentage of genome. CIs and \(P\) values were determined by 100 bootstrap iterations.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors’ Contributions

Study supervision: P.-Y. Kwok, C. Schaefer, J.S. Witte
Other (substantial intellectual input): D. Seminara

Disclaimer
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments
The authors thank the Kaiser Permanente Northern California members who have generously agreed to participate in the Kaiser Permanente Research Program on Genes, Environment, and Health, the ProHealth Study, and the CMHS. The authors also thank Peter Kraft, David Hunter, Sara Lindstrom, and Constance Chen for the contribution of PHS and HPFS genetic data for the eQTL analysis through the BPC3.

Grant Support
This work was supported by NIH grants CA127298, CA088164, and CA112355 (to J.S. Witte); AG046616 and DCO13300 (to T.J. Hoffmann); R01 GM107427 (to M.L. Freedman); CA136578, CA141298, and CA131945 (to K.L. Penney and L.A. Mucci); and CA151254 and CA157881 (to S.N. Thibodeau). This work was also supported by the UCSF Goldberg-Benioff Program in Cancer Translational Biology (to J.S. Witte), Prostate Cancer Foundation Young Investigator Awards (to K.L. Penney and L.A. Mucci), and Department of Defense grant W81XWH-11-1-0261 (to S.N. Thibodeau). The MEC and the genotyping (to K.L. Penney and L.A. Mucci), and Department of Defense grant W81XWH-11-1-0261 (to S.N. Thibodeau). The work was also supported by the NIH Common Fund (RC2 AG036607 to C. Schaefer and N. Risch). PEGASUS was supported by the Intramural Research Program, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH.

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 March 16, 2015; revised May 18, 2015; accepted May 26, 2015; published OnlineFirst June 1, 2015.

REFERENCES
GWAS of Prostate Cancer


