Title

A Large-Scale Association Study Detects Novel Rare Variants, Risk Genes, Functional Elements, and Polygenic Architecture of Prostate Cancer Susceptibility

Journal

Cancer Research, 81(7)

ISSN

0008-5472

Authors

Emami, Nima C

Cavazos, Taylor B

Rashkin, Sara R

et al.

Publication Date

2021-04-01

DOI

10.1158/0008-5472.can-20-2635

Peer reviewed

Published in final edited form as:

Cancer Res. 2021 April 01; 81(7): 1695–1703. doi:10.1158/0008-5472.CAN-20-2635.

A large-scale association study detects novel rare variants, risk genes, functional elements, and polygenic architecture of prostate cancer susceptibility

Nima C. Emami[1,2,*], **Taylor B. Cavazos**[1,*], **Sara R. Rashkin**[2], **Clinton L. Cario**[1,2], **Rebecca E. Graff**[2], **Caroline G. Tai**[2], **Joel A. Mefford**[3], **Linda Kachuri**[2], **Eunice Wan**[4], **Simon Wong**[4], **David Aaronson**[5], **Joseph Presti**[5], **Laurel A. Habel**[6], **Jun Shan**[6], **Dilrini K. Ranatunga**[6], **Chun R. Chao**[7], **Nirupa R. Ghai**[7], **Eric Jorgenson**[6], **Lori C. Sakoda**[6], **Mark N. Kvale**[4], **Pui-Yan Kwok**[3,4], **Catherine Schaefer**[6], **Neil Risch**[1,2,3,4,6,9], **Thomas J. Hoffmann**[1,2,4], **Stephen K. Van Den Eeden**[6,8], **John S. Witte**[1,2,3,4,8,9,†]

[1]Program in Biological and Medical Informatics, University of California San Francisco, San Francisco, California, United States of America

[2]Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, California, United States of America

[3]Program in Pharmaceutical Sciences and Pharmacogenomics, University of California San Francisco, San Francisco, California, United States of America

[4]Institute for Human Genetics, University of California San Francisco, San Francisco, California, United States of America

[5]Department of Urology, Kaiser Oakland Medical Center, Oakland, CA

[6]Division of Research, Kaiser Permanente Northern California, Oakland, CA

[7]Department of Research and Evaluation, Kaiser Permanente Southern California, Pasadena CA

[8]Department of Urology, University of California San Francisco, San Francisco, California, United States of America

[9]Helen Diller Family Comprehensive Cancer Center, University of California San Francisco, San Francisco, California, United States of America

Abstract

To identify rare variants associated with prostate cancer (PrCa) susceptibility and better characterize the mechanisms and cumulative disease risk associated with common risk variants, we conducted an integrated study of PrCa genetic etiology in two cohorts using custom genotyping microarrays, large imputation reference panels, and functional annotation approaches. Specifically, 11,984 men (6,196 PrCa cases, 5,788 controls) of European ancestry from Northern California Kaiser Permanente were genotyped and meta-analyzed with 196,269 men of European ancestry (7,917 PrCa cases, 188,352 controls) from the UK Biobank. Three novel loci, including

[†]Corresponding Author. JWitte@ucsf.edu. 415-502-6882. 1450 3rd Street, HD-388, San Francisco, CA 94158.
[*]These authors contributed equally to this work.

two rare variants (European ancestry minor allele frequency < 0.01, at 3p21.31 and 8p12), were significant genome-wide in a meta-analysis. Gene-based rare variant tests implicated a known PrCa gene (*HOXB13*), as well as a novel candidate gene (*ILDR1*), which encodes a receptor highly expressed in prostate tissue and is related to the B7/CD28 family of T cell immune checkpoint markers. Haplotypic patterns of long-range linkage disequilibrium were observed for rare genetic variants at *HOXB13* and other loci, reflecting their evolutionary history. Additionally, a polygenic risk score (PRS) of 188 PrCa variants was strongly associated with risk (90th vs. 40–60th percentile OR = 2.62, P = $2.55*10^{-191}$). Many of the 188 variants exhibited functional signatures of gene expression regulation or transcription factor binding, including a six-fold difference in log-probability of Androgen Receptor binding at the variant rs2680708 (17q22). Rare variant and PRS associations, with concomitant functional interpretation of risk mechanisms can help clarify the full genetic architecture of PrCa and other complex traits.

INTRODUCTION

For a number of diseases, including prostate cancer (PrCa), there has been limited success in detecting associated rare genetic variants, some of which may have larger effect sizes than common variants [1]. This is in part due to the difficulty of typing or imputing rare variants in adequately powered studies. Still, some rare germline variants associated with PrCa have been detected, such as in the DNA damage repair gene *BRCA2* [2,3] and the developmental transcription factor *HOXB13* [4]. While relatively few rare variants have been discovered, in aggregate, they may comprise a substantial portion of PrCa risk heritability [5]. In contrast, genome-wide association studies (GWAS) of more common variants have identified nearly 200 independent genetic variants associated with PrCa [6]. Each variant is typically associated with only a modest increase in PrCa risk, and thus individually not of sufficient magnitude to be clinically significant. However, combining all associated variants together into a single polygenic risk score (PRS) may distinguish men with a meaningfully increased risk of PrCa.

To investigate the impact of rare and common variants on PrCa risk, we undertook a large scale genome-wide association study of over 200,000 male subjects from two large cohorts: Kaiser Permanente (KP) in California [7] and the UK Biobank (UKB) [8]. Genotype microarrays, including GWAS backbones and custom rare variant content, were assayed in both cohorts, and unmeasured genotypes were imputed using a reference panel of over 27,000 phased Haplotype Reference Consortium (HRC) genomes [9]. We evaluated associations between individual rare and common variants and PrCa risk and interpreted the evolutionary origin and functional mechanisms of novel findings using multi-omics data. We also performed PRS modeling and functional characterization of the known common risk variants.

MATERIALS AND METHODS

Study Populations

We studied two cohorts of PrCa cases and cancer-free controls of European ancestry: 1) KP subjects from the RPGEH Genetic Epidemiology Research on Adult Health and Aging

(GERA) cohort, the California Men's Health Study (CMHS) and the ProHealth Study; and 2) the UKB. The KP cohort included 6,196 male cases and 5,788 male controls of European ancestry (mean age at diagnosis for cases = 68.1 years, mean age at enrollment into GERA among controls = 71.5; Supplementary Table 1). Controls were selected to be roughly matching in age and based on the availability of DNA for genotyping. The UKB cohort included 7,917 cases and 188,352 controls of European ancestry (mean age at diagnosis = 64.1, mean age at study enrollment among controls = 57.1; Supplementary Table 1).

Custom Microarray Design and Genotyping of Kaiser Permanente Subjects

To directly assay or tag potentially functional rare variation and to expand genome-wide coverage of less common variation in samples from KP, we collaborated with Affymetrix Inc. on the design of a custom Axiom DNA microarray (Supplementary Figure 1a) that was complementary to a GWAS array previously used to genotype the RPGEH GERA population [10–12]. The algorithm used to select variants on the custom array (Supplementary Figure 1b) resulted in 416,047 variant probesets comprising 54 distinct modules, including missense and loss-of-function mutations, rare exonic mutations from The Cancer Genome Atlas (TCGA) and dbGaP PrCa tumor exomes [13,14], and variants to supplement the previously genotyped GWAS array [7,10–12] (Supplementary Table 2). Many modules and most of the design content overlapped with the probesets on the UKB Affymetrix Axiom array, for which the array design, sample processing, and genotyping have been detailed [8].

Saliva biospecimens from KP participants were processed for DNA extraction using a protocol previously reported [10]. DNA samples from KP were processed using Samasy [15], a sample management system providing a visual and machine interface to facilitate robot liquid handling automation from source plates to destination plates matched by age, case status, and ethnicity. The algorithm implemented for destination plate randomization is described in the Supplementary Materials. A total of 173 96-well destination plates were amplified to increase DNA yields, and 200 ng of input DNA per well were array hybridized for 48 hours at 48 °C and genotyped using an Affymetrix GeneTitan Multi-Channel instrument. All of the reagents used for the genotyping assay are from the Axiom 2.0 Reagent Kit (Thermo Fisher Scientific, Santa Clara, CA; details given at https://tinyurl.com/y3yfse29).

Quality Control and Imputation

Detailed descriptions of the sample and genotype quality control (QC) procedures are given in the Supplementary Materials. Briefly, for the KP samples, we excluded specimens with poor resolution fluorescent measurements (DQC < 0.75) or call rate < 0.95 (Supplementary Figure 2a). Based on heterozygosity rate, call rate, and plate call rate, samples were further stratified into three tiers that were used to guide genotype QC. Specifically, genotype calls and posterior cluster locations from higher tier samples (as a consequence of higher input DNA quantities) were prioritized and used as empirical priors for resolving genotypes of lower tier samples using the Affymetrix AxiomGT1 algorithm (Supplementary Figure 2b) [16]. Genotypes were also filtered based on batch differences across the RPGEH GERA, CMHS, and ProHealth, and based on the fold-difference in minor allele frequency (MAF)

relative to the HRC and 1000 Genomes Project reference panels. These genotypes were then merged with previously assayed GWAS genotypes for the KP subjects (dbGaP Study Accession: phs001221.v1.p1), whose QC was described in a prior publication [7].

The KP data were phased using Eagle v2.3 (cohort-based) [17], and single nucleotide variant calls were imputed using Minimac3 using a reference panel consisting of a subpopulation of 27,165 HRC genomes accessible via the European Genome Archive (EGAS00001001710, which includes the 1000 Genomes Project Phase III samples). Indel polymorphisms were imputed only using the 1000 Genomes Project Phase III reference panel (2,504 genomes) (indels were not included in first release of HRC due to additional difficulty with harmonization; Supplementary Figure 3). Variants with $r^2_{INFO} < 0.3$ or with MAF less than $1/N_{REF}$ (where N_{REF} represents the total number of samples in the respective reference panel) were removed from the imputed genotypes.

For the UKB data, pre-imputation QC protocols have been previously described [8]. Genotypes were imputed by the UKB using two reference panels: the complete HRC reference (32,488 genomes) [9], and the combined UK10K plus 1000 Genomes Project Phase III reference panels (9,746 haplotypes). We excluded poorly imputed ($r^2_{INFO} < 0.3$) and excessively rare (MAF < $3*10^{-5}$) genotypes from the UKB.

Association Analyses

Associations between variant genotypes and PrCa were evaluated separately for the UKB and KP cohorts, and later combined in a meta-analysis. Association testing used logistic regression, with adjustment for age (for PrCa cases, age at diagnosis, versus age at time of entry into the GERA study for controls), body mass index, principal components of ancestry (PCs), and for the KP cohort genotyping array (since a limited number of the GERA EUR samples were typed with a different array; Supplementary Figure 3). The KP models controlled for 20 ancestry PCs using PLINK v2.00 [18], and the UKB models were adjusted for 10 PCs. The KP and UKB results were combined by fixed-effect meta-analysis using Metasoft v2.0.0 [19]. Gene-based rare variant tests (observed MAF < 1%) were conducted with the Sequence Kernel Association Test (SKAT) using the rvtests package (v20171009) [20], and meta-analyzed by Fisher's method [21] using R v3.3.3. Exome-wide significance ($P < 2.27*10^{-6}$) was evaluated by Bonferroni correction for 21,984 genes, and suggestive genes were designated as $2.27*10^{-6} < P < 2.27*10^{-5}$.

Evolutionary History of Rare Variants

To estimate the recency in origin of rare PrCa risk variants and further quantitate their evolutionary selection, we examined the extended haplotype homozygosity (EHH), or the length of a haplotype on which a variant allele resides, using the reference panel of 27,165 phased HRC genomes and the selscan package [22]. We also quantified the integrative haplotype score (iHS), or log ratio between a variant's major and minor alleles of the area under the EHH curves for each allele [22], to reflect differences in allelic age or selective pressure between the derived and ancestral alleles. The iHS was computed using an EHH cutoff of 0.05, both upstream (iHS_L) and downstream (iHS_R) of the query position.

Polygenic Risk Score Analyses

Each individual's PRS was computed by multiplying the out-of-sample effect sizes [6,7] for each of the 188 previously reported PrCa risk loci (log ORs) by their risk allele dosages, and then summing the resulting 188 values together (Supplementary Table 3). The odds ratios (OR) and 95% confidence intervals (CI) for associations between standardized PRS values (mean = 0, standard deviation = 1) and PrCa case-control status were estimated using logistic regression with adjustment for the same covariates modeled in our association analyses, with the exception of genotyping array for the purpose of consistency between UKB and KP PRS values. Additionally, we developed a PRS using summary statistics generated from the fixed-effect meta-analysis of UKB and KP. PRS were constructed from significant ($P < 5*10^{-8}$ and $P < 1*10^{-6}$) and independent (LD $r^2 < 0.1$ within a 3 Mb window) variants. In sample weights were adjusted for bias through a lasso-type winner's curse correction [23].

Functional Annotation

To consider the functional relevance of the known PrCa risk variants, we integrated two different analyses and sources of data. We trained elastic net regression models of normal prostatic gene expression [24], with a linear combination of germline genotypes as the predictor, using GLMNet [25] and a dataset of 471 subjects with normal prostate tissue RNA expression and genotype data [26]. Among the 188 previously reported PrCa risk variants, as well as the novel genome-wide significant variants identified here, we reported those that were included—or were in linkage disequilibrium (LD $r^2 > 0.5$) with a variant—in our expression models. We used a chi-squared test to assess whether the number of such expression quantitative trait (eQTL) variants was statistically significantly different from expected, where the latter was estimated from a set of 1000 minor allele frequency matched variants randomly selected from the HRC. For the same set of previously reported and newly identified variants, allele-specific differential transcription factor binding affinity was also estimated using sTRAP transcription factor affinity prediction [27] with the major and minor alleles.

RESULTS

Genome-wide Association Study

Among 188 SNPs previously implicated in PrCa GWAS, our association meta-analysis found that: 135 replicated with nominal significance ($P < 0.05$) and effect estimates in the same direction as previously reported; 87 of these SNPs had $P < 0.05 / 188$ (Bonferroni correction); 4 had suggestive P-values ($5*10^{-7} > P > 5*10^{-8}$); and 42 replicated with genome-wide significance ($P < 5*10^{-8}$) (Figure 1, Supplementary Table 3). Genome-wide significant associations (meta-analysis $P < 5*10^{-8}$) were observed at three loci not previously associated in PrCa GWAS (>3 Mb away and LD $r^2 < 0.005$ in all 1000 Genomes Phase III populations, relative to 188 known PrCa loci). Among the three, noncoding lead variants at these loci, rs557046152 (Chr 8p12; proximal to *DUSP4* and *KIF13B*), rs555778703 (4q31.21; intronic to *TBC1D9*), and rs62262671 (3p21.31; intronic to *BSN*), two were imputed variants that are rare in European (non-Finnish) ancestry populations: rs557046152 (gnomAD v2.1.1 MAF = 0.001) and rs555778703 (gnomAD v2.1.1 MAF =

0.009). Three additional novel loci had genome-wide significant *P*-values in the meta-analysis (Table 1), but lacked nominal significance (rs80242938 and rs149892036) and directional consistency (rs139191981) in both cohorts.

Gene-based Rare Variant Analysis

An additional gene-based rare variant meta-analysis of KP and the UKB data, using the sequence kernel association test (SKAT) and rare variants with MAF less than 0.01, yielded an exome-wide significant association at *HOXB13* ($P = 1.72*10^{-7}$; Figure 2, Supplementary Table 4), a well-characterized PrCa risk locus harboring the rare yet highly penetrant missense founder mutation G84E rs138213197 [4]. The association at *HOXB13* was primarily driven by the rs138213197 risk SNP (Supplementary Figure 4), which was highly significant in the GWAS meta-analysis ($P = 2.96*10^{-43}$). The SNP within *HOXB13* with the second smallest *P*-value, rs116931900, was located in the C-terminal untranslated region of the *HOXB13* open reading frame ($P = 2.95*10^{-4}$). SKAT also identified a suggestive *P*-value for *ILDR1* ($P = 7.46*10^{-6}$), a gene primarily expressed in prostate tissue [28]. The *ILDR1* association appeared to be primarily driven by a variant with a suggestive association in in the KP cohort; this was not associated with PrCa in the UKB (Supplementary Figure 4, Supplementary Table 5).

Evolutionary Characterization

We observed atypically long-range LD for the *HOXB13* G84E rare variant rs138213197-T, extending beyond a 1Mb window from its chromosomal position (Supplementary Figure 5). This observation was substantiated by considerable extended haplotype homozygosity for the rare missense allele (Figure 3a). In particular, rs138213197 had an integrated haplotype score (iHS) equal to 2.87 (iHS_L: 3.53, iHS_R: 2.54) in our HRC haplotype data, greater than the |iHS| > 2.5 threshold corresponding to the most extreme 1% of values [29], and reflecting the recent origin and/or selective constraint at the rs138213197 locus. Likewise, for the novel rare variant rs555778703 in the intron of gene *TBC1D9*, the rare G risk allele (Figure 3b) had an iHS equal to 2.31 (iHS_L: 2.00, iHS_R: 2.79). For a proxy single nucleotide polymorphism (SNP) rs57029021 (LD $r^2 = 0.666$ in 1000 Genomes Project Phase III EUR) of the novel rare SNP rs557046152 (which was unmeasured in the EGA HRC reference genomes), the rare A allele had an iHS equal to 0.87 (iHS_L: 1.60, iHS_R: 0.77; Figure 3c).

Polygenic Risk Scores

For subjects in KP and the UKB, there was a strong association between being in the top decile versus the referent percentile (40–60%) of the PRS and PrCa risk (Supplementary Figure 6, Supplementary Table 6; Meta-analysis OR [95% CI] = 2.62 [2.46, 2.79], $P = 2.55*10^{-191}$; KP OR = 2.40 [2.06, 2.80], $P = 4.38*10^{-29}$; UKB OR = 2.71 [2.52, 2.93], $P = 2.25*10^{-150}$). There was also a significant decrease in risk as a result of being in the bottom decile (Supplementary Figure 6, Supplementary Table 6; Meta-analysis OR [95% CI] = 0.34 [0.30, 0.37], $P = 1.69*10^{-98}$; KP OR = 0.36 [0.31, 0.42], $P = 6.78*10^{-40}$; UKB OR = 0.33 [0.29, 0.38], $P = 1.31*10^{-54}$). No significant heterogeneity between KP and UKB cohorts was detected for any PRS decile association with PrCa risk ($I^2 < 4\%$; $p_{Cochran's\text{-}Q} > 0.84$). The top decile included 9.40% of all controls, but 18.3% of the cases, while the bottom decile 10.4% of controls and 4.02% of cases. When limiting our PRS to only the 87 PrCa

risk variants that replicated in our data ($P < 0.05 / 188$) we saw little impact to risk as a result of being in the top decile (Meta-analysis OR [95% CI] = 2.60 [2.44, 2.77]) or the bottom decile (Meta-analysis OR [95% CI] = 0.34 [0.31, 0.38]). However, when the three novel variants (rs557046152, rs555778703, and rs62262671) discovered in our study were incorporated into the PRS, following in-sample bias correction [30] of variant effects, there was an increase in the risk in the top decile (Meta-analysis OR [95% CI] = 2.74 [2.57, 2.92]) with risk in the bottom decile remaining unchanged (Meta-analysis OR [95% CI] = 0.34 [0.31, 0.38]).

We also generated PRS from our meta-analysis summary statistics using independent (LD $r^2 < 0.1$) and GWAS significant ($P < 5*10^{-8}$) or suggestive ($P < 1*10^{-6}$) variants and adjusted weights from our study where a lasso-type in-sample bias correction procedure [23] was applied to prevent overfitting due to winner's curse (Supplementary Table 7). The GWAS significant and suggestive PRS developed in our study had similar risk from being in the top decile (Meta-analysis OR [95% CI] = 2.34 [2.19, 2.49] and 2.55 [2.39, 2.72]; Supplementary Table 8) and bottom decile (Meta-analysis OR [95% CI] = 0.48 [0.44, 0.52] and 0.44 [0.40, 0.49]) as the PRS constructed from the 188 previously identified PrCa SNPs. Although we corrected our PRS variant effects, results may be inflated due to in-sample prediction and therefore should be interpreted with caution; further application to an independent cohort is needed to test the validity of this PRS.

Functional Interpretation

To characterize the functional consequences of common PrCa variants, we examined their effects on gene expression and transcription factor binding. Among the 188 previously reported PrCa risk variants and 3 novel risk variants identified, 80 were in linkage disequilibrium (LD $r^2 > 0.5$ in 1000 Genomes Project Phase III EUR) with an eQTL variant in our regularized models of normal prostatic expression levels, which was significantly higher than expected ($P = 3.07*10^{-5}$; Supplementary Table 9). This included one of the three novel variants implicated in our meta-analysis, rs62262671, which is predicted to alter expression of two genes, *RMB6* and *UBA7*.

Furthermore, 32 variants were predicted to significantly alter transcription factor binding site (TFBS) affinities (Supplementary Table 10), with greater than three-fold predicted differences in TFBS log-probabilities. rs2680708 (17q22) showed the greatest fold change in predicted binding affinity ($P = 3.91\times10^{-7}$) of any variant-TF pair analyzed (Supplementary Table 10), and was predicted to abrogate binding of the androgen receptor (AR) transcription factor, a sentinel of prostatic gene expression [31]. In keeping with this AR-mediated link between genetic variation and PrCa risk, a pattern of androgen-mediated influence, and/or PrCa-related expression, also emerged among the remaining variant-TF pairings, including DBP (rs2680708) [32], HMGIY (rs5799921) [33], AP1 (rs2660753) [32,34], DELTAEF1 (also known as ZEB1; rs7210100) [35], STAT5A (rs742134) [36], HNF1B (rs742134) [37], and MAFB (rs9625483) [38], among others. Moreover, the rs62262671 risk variant associated in the meta-analysis was predicted to impact binding of OCT1 ($P = 5.04\times10^{-3}$), a TF closely intertwined with AR signaling in prostate cancer cell lines [39–42].

DISCUSSION

We undertook a large-scale association analysis of genotyped and imputed common and rare variants, and implicated three novel loci, including two rare variants rs557046152 (Chr 8p12; proximal to *DUSP4* and *KIF13B* at 8p12) and rs555778703 (4q31.21; intronic to *TBC1D9*), and one common variant rs62262671 (3p21.31; intronic to *BSN*). Gene-based rare variant tests, in our meta-analysis of 14,113 PrCa cases and 201,722 controls across the KP and UKB cohorts, also revealed significant and suggestive associations, respectively, for *HOXB13*, a well-studied PrCa risk gene, and a novel candidate gene *ILDR1*, which encodes a B7-like receptor protein related to the immune checkpoint blockade immuno-oncology pathway. Evolutionary analysis revealed patterns of haplotypic natural selection for both the novel rare variants associated in our study and also a known PrCa rare variant rs138213197 (*HOXB13* G84E). Functional analyses suggested novel mechanistic hypotheses of transcription factor binding and gene expression modulation for dozens of PrCa risk loci, including previously reported variants and the novel variant rs62262671. Finally, a polygenic risk score analysis of 188 PrCa variants indicated that men in the highest PRS decile have a substantially increased risk that may be of clinical importance.

The two rare variants associated in our meta-analysis, rs557046152 (Meta OR [95% CI] = 1.79 [1.47, 2.17]) and rs555778703 (Meta OR [95% CI] = 1.82 [1.48, 2.24]), both exhibited OR effect sizes greater than what is most often seen for common GWAS variants (OR > 1.2), in keeping with their rare allele frequencies. Because rare allele frequencies with large effect sizes can be explained by purifying selection in the human genome [43], we examined the haplotypic patterns of rare variants associated in our study for hallmarks of natural selection. Indeed, evolutionary analyses revealed signs of selection at rs555778703 (iHS = 2.31), suggested the possibility of selection at an LD proxy for rs557046152 (rs57029021 iHS = 0.87), and confirmed the presence of selective forces at a known PrCa rare variant rs138213197 (*HOXB13* G84E; iHS = 2.87). While previous studies have identified the haplotype for the G84E variant [44] and estimated its recent date of origin in the 18th century [45], our study is the first demonstration, to our knowledge, of evolutionary forces of natural selective pressure at the *HOXB13* G84E haplotype.

Among the two genes associated in our gene-based rare variant tests, *ILDR1* (Immunoglobulin Like Domain Containing Receptor 1) encodes a cell-surface receptor protein ILDR1 that is a gene highly expressed in prostate tissue [28], and which localizes at tricellular tight junctions while sealing extracellular regions, along with its closely related homologs ILDR2 and ILDR3 [46]. Recent studies have revealed that ILDR2 is a related to the B7 family of proteins [47], which includes the PD-L1 (B7-H1) and PD-L2 (B7-H2) immune checkpoint ligands [48]. The success of Anti-PD-1 and Anti-PD-L1 immunotherapies has recently inspired preclinical testing of a novel immune checkpoint inhibitor targeting ILDR2 [49]. Our finding of a rare variant association in *ILDR1*, as well as the sequence similarity between ILDR1 and ILDR2 (39% overall sequence identity) and structural similarity of ILDR2 to the B7 protein family [47], motivates study of the potential involvement of *ILDR1* in PrCa and cancer immunology.

Our constructed PRS for European ancestry men exhibited a large magnitude of effect. Comparing men in the top 1% of the PRS distribution to the median (40–60%) yielded an OR of 4.12 [95% CI 3.60–4.72], which is only of somewhat lower magnitude as high risk genes such as *BRCA1* for breast cancer, where comparing mutation carriers to non-carriers have an OR = 5.91 [95% CI 5.25, 6.67] [50]. Although the PRS effect is of relatively large magnitude, the scores may not be fully transferable to individuals of non-European descent [7] and need to be examined in other ethnic groups [51,52]. In spite of these limitations, over a decade of GWAS efforts [53] has advanced the genetic characterization of PrCa considerably. Our construction of a PRS model for PrCa with high discrimination demonstrates this remarkable progress and the predictive power of aggregating PrCa risk loci. Interestingly, our meta-analysis did not replicate many of the previously associated SNPs included in the PRS and removing the non-replicated SNPs had very little impact on the effect of the PRS. This may reflect winner's curse for the original findings or differences in populations studied. Nevertheless, the PRS remained strongly associated with PrCa even when including all of the SNPs.

Using eQTL models of prostatic gene expression, we developed functional annotations for germline variants implicated in our analysis, or previously implicated and present in our polygenic risk score. Among the genes putatively targeted by 80 such PrCa risk variants and their LD proxies, many have been previously described as prostate cancer risk genes in transcriptome wide association studies (TWAS) [24,54,55]: *AGAP7*, *BHLHA15*, *C2orf43*, *C9orf78*, *CTBP2*, *EHBP1*, *FAM57A*, *FOXP4*, *GEMIN4*, *IRX4*, *MMP7*, *MSMB*, *NCOA4*, *PPP1R14A*, *RAB7L1*, *RGS17*, *SLC22A3*, *UHRF1BP1*. Among the remaining fraction not directly implicated by PrCa TWAS, several additional genes have been genetically or transcriptionally linked to prostate cancer pathogenesis, including *ACVR2A* [56], *NUDT11* [57], and *PPFIBP2* [58].

Integration of gene expression and transcription factor binding site affinity data suggested novel mechanisms for many of the common PrCa variants previously reported. One example is a highly significant change in transcription factor binding affinity at rs2680708. This finding is especially interesting given that the risk allele rs2680708-G abrogates a binding site for AR, a master regulator of prostatic gene expression. The magnitude of the rs2680708 TFBS effect we predicted exceeds, by several fold, the magnitude of effect for TFBS variants that have been previously reported and validated by in-vitro functional assays, including rs8134378, rs11084033, and rs2659051 [24,59–61]. While our eQTL analyses did not nominate any gene whose expression may be affected as a consequence of eliminating this particular binding site, further study may reveal the effect of rs2680708 on the dysregulation of gene expression or additional molecular processes that potentially link its reported effect on PrCa risk with its putative influence on AR binding.

The newly implicated rs62262671 risk variant (3p21) was also predicted to have an impact on binding affinity for OCT1, a TF with a known impact on PrCa and AR signaling [39–42]. Given that rs62262671 was also identified as an eQTL affecting the expression of *RBM6* and *UBA7*, these findings suggest that OCT1 may be involved in the regulation of the expression of these two genes, and provides a hypothesis for future functional follow-up regarding the involvement of these genes in PrCa development. A potential limitation of *in*

silico TFBS prediction using the sTRAP algorithm is that, for certain transcription factor profiles, very similar and significant *P*-values are sometimes observed for different TF's at the same SNP. While this observation could simply reflect similarity between particular transcription factor binding profile models, it may also reflect limitations in the models that force highly significant differences to converge to the same *P*-value upon reaching the tail of an empirical background distribution. Nevertheless, our approach provides a convenient and valid method for mechanistic interpretation and hypothesis generation in research involving allele-specific TF binding at genetic polymorphisms.

To improve detection of rare variant associations our study design prioritized directly genotyping variants of putative functional significance, rare variants from trait-specific whole exome sequencing (WES) cohorts, and rare variants with proximity to trait-associated loci, all on a custom microarray. While our use of two large, population-based cohorts empowered univariate association testing of PrCa loci driven by rare genetic variants genome-wide, multivariate testing of rare variants colocalizing at particular genes remains an effective means of revealing loci driven by rare variants which lack marginal, univariate significance, but which significantly contribute to disease risk when considered in aggregate. Furthermore, due in part to limited statistical power, the mechanisms through which the rare, noncoding variants we identified are associated with PrCa remain somewhat unclear, with a lack of precise functional evidence regarding mechanism of action or close proximity to genes or known risk loci in cis. We used biophysical models of transcription factor binding to identify variant alleles that may introduce or interrupt a TFBS that are independent of allele frequency. Nevertheless, there remains a challenge of not only detecting—but also interpreting—how noncoding rare variants impact the genetic etiology of complex traits using existing gene-based methodologies and functional genomic datasets.

By undertaking a GWAS in the large KP and UKB population-based cohorts, we detected several novel PrCa risk loci, a novel risk gene, and a polygenic signature of PrCa risk. Functional characterization of PrCa risk variants using gene expression and transcription factor binding affinity data revealed putative mechanisms of disease risk. However, further study is needed to more fully illuminate the biological mechanisms that underpin the influence of PrCa risk loci, particularly for rare variants.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

The authors thank Elad Ziv, Nadav Ahituv, and Ryan Hernandez for their guidance and feedback. We are also grateful to 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 California Men's Health Study. This research has been conducted using the UK Biobank Resource under Application Number 14105, and has been supported by the Robert Wood Johnson Foundation (J.S. Witte), the National Institutes of Health R01CA088164, R01CA201358, R25CA112355, T32GM067547, and U01CA127298 (J.S. Witte), the UCSF Goldberg-Benioff Program in Cancer Translational Biology (J.S. Witte), the UCSF Discovery Fellows program (N.C. Emami, T.B. Cavazos), the Microsoft Azure for Research program (C.L. Cario, N.C. Emami), and the Amazon AWS Cloud Credits for Research program (C.L. Cario, N.C. Emami). Support for participant enrollment, survey completion, and biospecimen collection for the RPGEH was provided by the Robert Wood Johnson Foundation, the Wayne and Gladys Valley Foundation, the Ellison Medical Foundation, and Kaiser Permanente

national and regional community benefit programs. Genotyping of the GERA cohort was funded by a grant from the National Institute on Aging, the National Institute of Mental Health, and the NIH Common Fund (RC2 AG036607). The sponsors played no role in the study.

Financial disclosures: John S. Witte certifies that all conflicts of interest, including specific financial interests and relationships and affiliations relevant to the subject matter or materials discussed in the manuscript (e.g. employment/affiliation, grants or funding, consultancies, honoraria, stock ownership or options, expert testimony, royalties, or patents filed, received, or pending), are the following: None.

REFERENCES

1. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 2010;11:446–50 [PubMed: 20479774]
2. Ostrander EA, Udler MS. The role of the BRCA2 gene in susceptibility to prostate cancer revisited. Cancer Epidemiol Biomarkers Prev 2008;17:1843–8 [PubMed: 18708369]
3. Oh M, Alkhushaym N, Fallatah S, Althagafi A, Aljadeed R, Alsowaida Y, et al. The association of BRCA1 and BRCA2 mutations with prostate cancer risk, frequency, and mortality: A meta-analysis. Prostate 2019;79:880–95 [PubMed: 30900310]
4. Ewing CM, Ray AM, Lange EM, Zuhlke KA, Robbins CM, Tembe WD, et al. Germline mutations in HOXB13 and prostate-cancer risk. N Engl J Med 2012;366:141–9 [PubMed: 22236224]
5. Mancuso N, Rohland N, Rand KA, Tandon A, Allen A, Quinque D, et al. The contribution of rare variation to prostate cancer heritability. Nat Genet 2016;48:30–5 [PubMed: 26569126]
6. Schumacher FR, Al Olama AA, Berndt SI, Benlloch S, Ahmed M, Saunders EJ, et al. Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. Nat Genet 2018;50:928–36 [PubMed: 29892016]
7. Hoffmann TJ, Van Den Eeden SK, Sakoda LC, Jorgenson E, Habel LA, Graff RE, et al. A large multiethnic genome-wide association study of prostate cancer identifies novel risk variants and substantial ethnic differences. Cancer Discov 2015;5:878–91 [PubMed: 26034056]
8. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature 2018;562:203–9 [PubMed: 30305743]
9. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet 2016;48:1279–83 [PubMed: 27548312]
10. Kvale MN, Hesselson S, Hoffmann TJ, Cao Y, Chan D, Connell S, et al. Genotyping Informatics and Quality Control for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. Genetics 2015;200:1051–60 [PubMed: 26092718]
11. Hoffmann TJ, Kvale MN, Hesselson SE, Zhan Y, Aquino C, Cao Y, et al. Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array. Genomics 2011;98:79–89 [PubMed: 21565264]
12. Hoffmann TJ, Zhan Y, Kvale MN, Hesselson SE, Gollub J, Iribarren C, et al. Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. Genomics 2011;98:422–30 [PubMed: 21903159]
13. Cancer Genome Atlas Research N. The Molecular Taxonomy of Primary Prostate Cancer. Cell 2015;163:1011–25 [PubMed: 26544944]
14. Kumar A, White TA, MacKenzie AP, Clegg N, Lee C, Dumpit RF, et al. Exome sequencing identifies a spectrum of mutation frequencies in advanced and lethal prostate cancers. Proc Natl Acad Sci U S A 2011;108:17087–92 [PubMed: 21949389]
15. Cario CL, Witte JS. Samasy: a Sample Management System. BioTechniques 2018
16. Inc. TFS. Axiom Genotyping Solution Data Analysis Guide.
17. Loh PR, Danecek P, Palamara PF, Fuchsberger C, Y AR, H KF, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet 2016;48:1443–8 [PubMed: 27694958]
18. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 2015;4:7 [PubMed: 25722852]

19. Han B, Eskin E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. Am J Hum Genet 2011;88:586–98 [PubMed: 21565292]
20. Zhan X, Hu Y, Li B, Abecasis GR, Liu DJ. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. Bioinformatics 2016;32:1423–6 [PubMed: 27153000]
21. Lee S, Teslovich TM, Boehnke M, Lin X. General framework for meta-analysis of rare variants in sequencing association studies. Am J Hum Genet 2013;93:42–53 [PubMed: 23768515]
22. Szpiech ZA, Hernandez RD. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. Mol Biol Evol 2014;31:2824–7 [PubMed: 25015648]
23. Shi J, Park JH, Duan J, Berndt ST, Moy W, Yu K, et al. Winner's Curse Correction and Variable Thresholding Improve Performance of Polygenic Risk Modeling Based on Genome-Wide Association Study Summary-Level Data. PLoS Genet 2016;12:e1006493 [PubMed: 28036406]
24. Emami NC, Kachuri L, Meyers TJ, Das R, Hoffman JD, Hoffmann TJ, et al. Association of imputed prostate cancer transcriptome with disease risk reveals novel mechanisms. Nat Commun 2019;10:3107 [PubMed: 31308362]
25. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 2010;33:1–22 [PubMed: 20808728]
26. Thibodeau SN, French AJ, McDonnell SK, Cheville J, Middha S, Tillmans L, et al. Identification of candidate genes for prostate cancer-risk SNPs utilizing a normal prostate tissue eQTL data set. Nat Commun 2015;6:8653 [PubMed: 26611117]
27. Manke T, Roider HG, Vingron M. Statistical modeling of transcription factor binding affinities predicts regulatory interactions. PLoS Comput Biol 2008;4:e1000039 [PubMed: 18369429]
28. Hauge H, Patzke S, Delabie J, Aasheim HC. Characterization of a novel immunoglobulin-like domain containing receptor. Biochem Biophys Res Commun 2004;323:970–8 [PubMed: 15381095]
29. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol 2006;4:e72 [PubMed: 16494531]
30. Zhong H, Prentice RL. Correcting "winner's curse" in odds ratios from genomewide association findings for major complex human diseases. Genet Epidemiol 2010;34:78–91 [PubMed: 19639606]
31. Levina E, Ji H, Chen M, Baig M, Oliver D, Ohouo P, et al. Identification of novel genes that regulate androgen receptor signaling and growth of androgen-deprived prostate cancer cells. Oncotarget 2015;6:13088–104 [PubMed: 26036626]
32. Wang G, Wang Y, Feng W, Wang X, Yang JY, Zhao Y, et al. Transcription factor and microRNA regulation in androgen-dependent and -independent prostate cancer cells. BMC Genomics 2008;9 Suppl 2:S22
33. Qi C, Bin L, Yang Y, Yang Y, Li J, Zhou Q, et al. Glipizide suppresses prostate cancer progression in the TRAMP model by inhibiting angiogenesis. Sci Rep 2016;6:27819 [PubMed: 27292155]
34. Kavya K, Kumar MN, Patil RH, Hegde SM, Kiran Kumar KM, Nagesh R, et al. Differential expression of AP-1 transcription factors in human prostate LNCaP and PC-3 cells: role of Fra-1 in transition to CRPC status. Mol Cell Biochem 2017;433:13–26 [PubMed: 28386843]
35. Orellana-Serradell O, Herrera D, Castellon EA, Contreras HR. The transcription factor ZEB1 promotes an aggressive phenotype in prostate cancer cell lines. Asian J Androl 2018;20:294–9 [PubMed: 29271397]
36. Haddad BR, Gu L, Mirtti T, Dagvadorj A, Vogiatzi P, Hoang DT, et al. STAT5A/B gene locus undergoes amplification during human prostate cancer progression. Am J Pathol 2013;182:2264–75 [PubMed: 23660011]
37. Hu YL, Zhong D, Pang F, Ning QY, Zhang YY, Li G, et al. HNF1b is involved in prostate cancer risk via modulating androgenic hormone effects and coordination with other genes. Genet Mol Res 2013;12:1327–35 [PubMed: 23661456]
38. Matsushita S, Suzuki K, Ogino Y, Hino S, Sato T, Suyama M, et al. Androgen Regulates Mafb Expression Through its 3'UTR During Mouse Urethral Masculinization. Endocrinology 2016;157:844–57 [PubMed: 26636186]

39. Obinata D, Takayama K, Fujiwara K, Suzuki T, Tsutsumi S, Fukuda N, et al. Targeting Oct1 genomic function inhibits androgen receptor signaling and castration-resistant prostate cancer growth. Oncogene 2016;35:6350–8 [PubMed: 27270436]
40. Obinata D, Takayama K, Urano T, Murata T, Kumagai J, Fujimura T, et al. Oct1 regulates cell growth of LNCaP cells and is a prognostic factor for prostate cancer. Int J Cancer 2012;130:1021–8 [PubMed: 21387309]
41. Takayama KI, Suzuki Y, Yamamoto S, Obinata D, Takahashi S, Inoue S. Integrative Genomic Analysis of OCT1 Reveals Coordinated Regulation of Androgen Receptor in Advanced Prostate Cancer. Endocrinology 2019;160:463–72 [PubMed: 30649323]
42. Yamamoto S, Takayama KI, Obinata D, Fujiwara K, Ashikari D, Takahashi S, et al. Identification of new octamer transcription factor 1-target genes upregulated in castration-resistant prostate cancer. Cancer Sci 2019;110:3476–85 [PubMed: 31454442]
43. Uricchio LH, Zaitlen NA, Ye CJ, Witte JS, Hernandez RD. Selection and explosive growth alter genetic architecture and hamper the detection of causal rare variants. Genome Res 2016;26:863–73 [PubMed: 27197206]
44. Hoffmann TJ, Sakoda LC, Shen L, Jorgenson E, Habel LA, Liu J, et al. Imputation of the rare HOXB13 G84E mutation and cancer risk in a large population-based cohort. PLoS Genet 2015;11:e1004930 [PubMed: 25629170]
45. Chen Z, Greenwood C, Isaacs WB, Foulkes WD, Sun J, Zheng SL, et al. The G84E mutation of HOXB13 is associated with increased risk for prostate cancer: results from the REDUCE trial. Carcinogenesis 2013;34:1260–4 [PubMed: 23393222]
46. Higashi T, Tokuda S, Kitajiri S, Masuda S, Nakamura H, Oda Y, et al. Analysis of the 'angulin' proteins LSR, ILDR1 and ILDR2--tricellulin recruitment, epithelial barrier function and implication in deafness pathogenesis. J Cell Sci 2013;126:966–77 [PubMed: 23239027]
47. Hecht I, Toporik A, Podojil JR, Vaknin I, Cojocaru G, Oren A, et al. ILDR2 Is a Novel B7-like Protein That Negatively Regulates T Cell Responses. J Immunol 2018;200:2025–37 [PubMed: 29431694]
48. Havel JJ, Chowell D, Chan TA. The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy. Nat Rev Cancer 2019;19:133–50 [PubMed: 30755690]
49. Huetter J, Gritzan U, Gutcher I, Golfier S, Doecke W-D, Luetke-Eversloh MV, et al. Abstract 2778: Discovery and preclinical characterization of BAY 1905254 a novel immune checkpoint inhibitor for cancer immunotherapy targeting the immunoglobulin-like domain containing receptor 2 (ILDR2). Cancer Research 2018;78:2778–
50. Kurian AW, Hughes E, Handorf EA, Gutin A, Allen B, Hartman A-R, et al. Breast and Ovarian Cancer Penetrance Estimates Derived From Germline Multiple-Gene Sequencing Results in Women. JCO Precision Oncology 2017:1–12
51. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, et al. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am J Hum Genet 2017;100:635–49 [PubMed: 28366442]
52. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet 2019;51:584–91 [PubMed: 30926966]
53. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet 2017;101:5–22 [PubMed: 28686856]
54. Mancuso N, Gayther S, Gusev A, Zheng W, Penney KL, Kote-Jarai Z, et al. Large-scale transcriptome-wide association study identifies new prostate cancer risk regions. Nat Commun 2018;9:4079 [PubMed: 30287866]
55. Wu L, Wang J, Cai Q, Cavazos TB, Emami NC, Long J, et al. Identification of Novel Susceptibility Loci and Genes for Prostate Cancer Risk: A Transcriptome-Wide Association Study in Over 140,000 European Descendants. Cancer Res 2019;79:3192–204 [PubMed: 31101764]
56. Rossi MR, Ionov Y, Bakin AV, Cowell JK. Truncating mutations in the ACVR2 gene attenuates activin signaling in prostate cancer cells. Cancer Genet Cytogenet 2005;163:123–9 [PubMed: 16337854]

57. Grisanzio C, Werner L, Takeda D, Awoyemi BC, Pomerantz MM, Yamada H, et al. Genetic and functional analyses implicate the NUDT11, HNF1B, and SLC22A3 genes in prostate cancer pathogenesis. Proc Natl Acad Sci U S A 2012;109:11252–7 [PubMed: 22730461]

58. Wu Y, Yu H, Zheng SL, Feng B, Kapron AL, Na R, et al. Germline mutations in PPFIBP2 are associated with lethal prostate cancer. Prostate 2018;78:1222–8 [PubMed: 30043417]

59. Clinckemalie L, Spans L, Dubois V, Laurent M, Helsen C, Joniau S, et al. Androgen regulation of the TMPRSS2 gene and the effect of a SNP in an androgen response element. Mol Endocrinol 2013;27:2028–40 [PubMed: 24109594]

60. O'Mara TA, Nagle CM, Batra J, Kedda MA, Clements JA, Spurdle AB. Kallikrein-related peptidase 3 (KLK3/PSA) single nucleotide polymorphisms and ovarian cancer survival. Twin Res Hum Genet 2011;14:323–7 [PubMed: 21787114]

61. Jin HJ, Jung S, DebRoy AR, Davuluri RV. Identification and validation of regulatory SNPs that modulate transcription factor chromatin binding and gene expression in prostate cancer. Oncotarget 2016;7:54616–26 [PubMed: 27409348]

STATEMENT OF SIGNIFICANCE

This study maps the biological relationships between diverse risk factors for prostate cancer, integrating different functional datasets to interpret and model genome-wide data from over 200,000 men with and without prostate cancer.

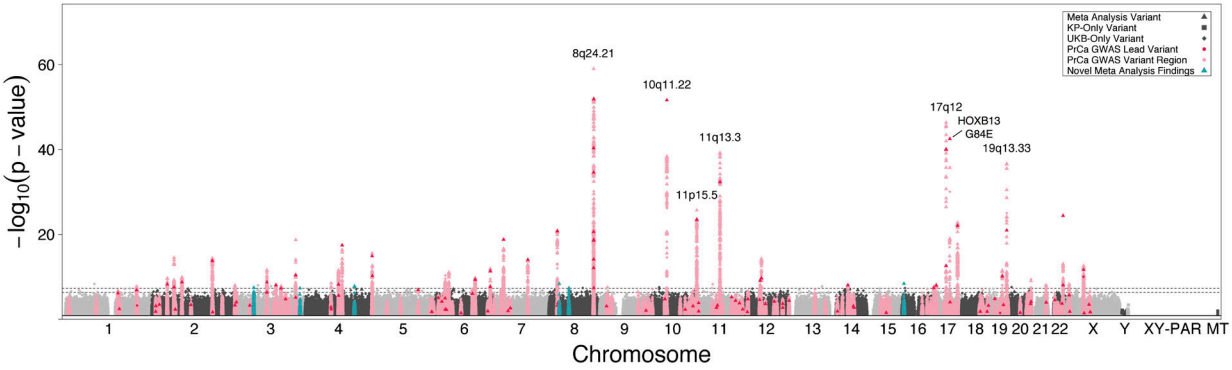

Figure 1 : "Prostate Cancer Risk Meta-Analysis Manhattan Plot for Kaiser Permanente and UK Biobank European Ancestry Subjects"

Manhattan plot depicting the results of a meta-analysis of male European ancestry subjects from the Kaiser Permanente (KP; N = 6,196 PrCa cases, 5,788 controls) and UK Biobank (UKB; N = 7,917 PrCa cases, 188,352 controls) cohort genome-wide associations with prostate cancer (PrCa) risk. The associations ($-\log_{10}$(*P*-value), Y-axis) are plotted against the chromosome (1–22, X, Y, XY-pseudoautosomal region XY-PAR, and mitochondrial chromosome MT) and position (X-axis) of the genotyped or imputed genetic variants, with thresholds for significant ($P < 5.0*10^{-8}$) and suggestive ($5.0*10^{-7} < P < 5.0*10^{-8}$) associations illustrated by dashed grey lines. Other variants on odd and even chromosomes are colored in alternating shades, and all variants with $P > 0.05$ are excluded from the plot. Triangular data points illustrate variants that were meta-analyzed between KP and UKB, while squares and circles indicate variants present exclusively in the KP or UKB summary statistics, respectively. Previously discovered PrCa loci are highlighted in pink for a 2 Mb window around the reported lead variant, which is highlighted in red, and previously unreported loci reaching genome-wide significance in our meta-analysis are colored in teal.

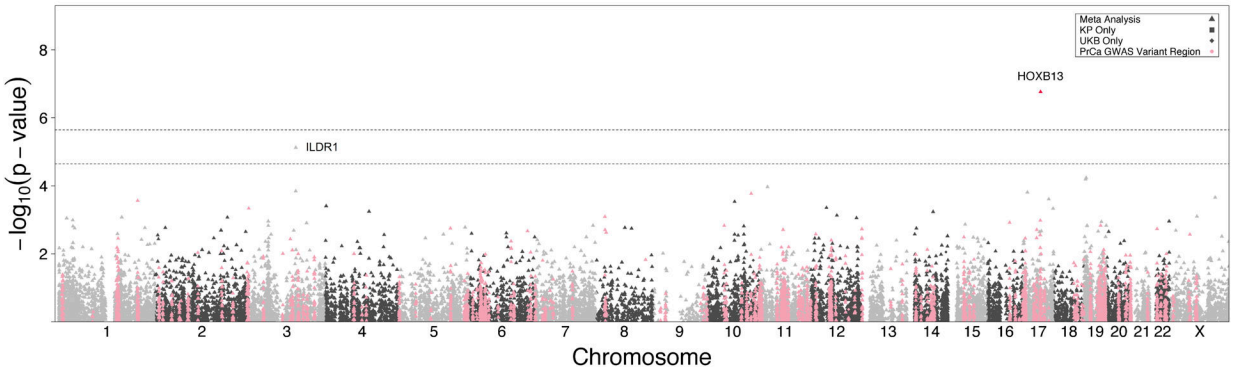

Figure 2 : "SKAT Gene-Based Rare Variant (MAF < 1%) Prostate Cancer Meta-Analysis of KP and UKB Subjects"

Manhattan plot of associations for a gene-based meta-analysis between the Kaiser Permanente and UK Biobank. The associations ($-\log_{10}$(*P*-value), Y-axis) are plotted against the chromosome (1–22, X) and position (X-axis) of the modeled genes, with thresholds for Bonferroni-significant ($P < 2.27*10^{-6}$) and suggestive ($2.27*10^{-5} < P < 2.27*10^{-6}$) associations illustrated by dashed grey lines. Non-significant genes on odd and even chromosomes are colored in alternating shades. Triangular data points illustrate variants that were meta-analyzed between KP and UKB, while squares and circles indicate genes present exclusively in the KP or UKB summary statistics, respectively. Previously discovered PrCa loci are highlighted in pink for a 2 Mb window around the reported lead variant.

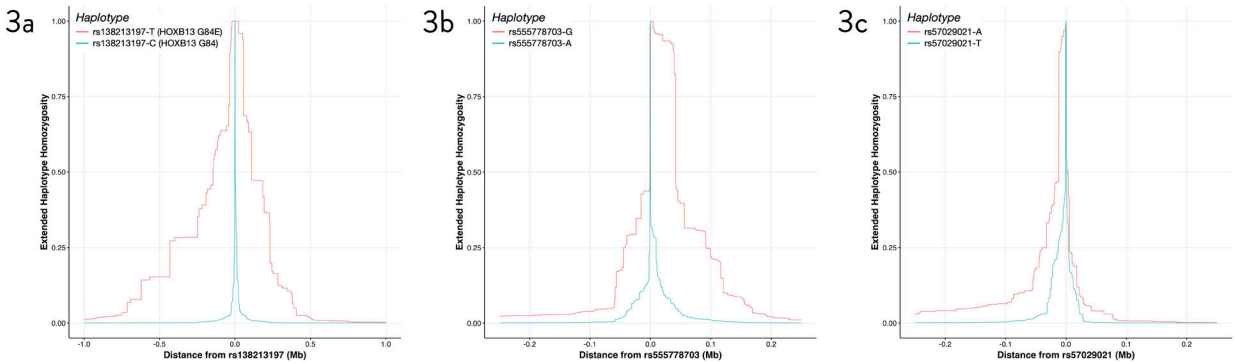

Figure 3 : "Extended Haplotype Homozygosity of Prostate Cancer Associated Rare Variants"

Haplotype Lengths for Rare PrCa Risk Variants. Extended haplotype homozygosity (EHH) plots illustrating the decay in non-recombinant linkage (Y-axis) with increasing distance along the length of the haplotypes centered at two alleles of a "core" query variant (X-axis). Differences in EHH, iHH (the area under the EHH curve), and iHS (the log-ratio between the iHH for the derived and ancestral allele) may reflect a difference in allelic age between the derived and ancestral alleles, or alternatively the selective pressure to retain a particular allele with preference to the alternative. 3a. EHH curves for the rare *HOXB13* G84E missense variant and Northern European founder mutation rs138213197, for which the iHS value of 2.87 (iHS_L: 3.53, iHS_R: 2.54) reflects the more recent origin of the derived G84E allele rs138213197-T. 3b. EHH curves for the novel rare variant association rs555778703, with an iHS value of 2.31 (iHS_L: 2.00, iHS_R: 2.79). 3c. EHH curves for rs57029021, an LD proxy variant for the novel rare indel association rs557046152 (LD r^2 = 0.666 in 1000 Genomes Project Phase III EUR) with an iHS value of 0.87 (iHS_L: 1.60, iHS_R: 0.77).

Table 1.

Novel Prostate Cancer Susceptibility Associations from a Meta-Analysis of Subjects from Kaiser Permanente and UK Biobank

Risk Variant dbSNP rsid Genomic Locus gnomAD MAF Risk Allele (Ref)	Kaiser Permanente (KP) (6,196 cases, 5,788 controls)			UK Biobank (UKB) (7,917 cases, 188,352 controls)			Meta-Analysis KP + UKB Subjects	
	Odds Ratio [95% CI]	*P*-value	KP MAF (r^2_{INFO})	Odds Ratio [95% CI]	*P*-value	UKB MAF (r^2_{INFO})	Odds Ratio [95% CI]	*P*-value
rs557046152 Locus: 8p12 MAF: 0.001 * G (GTT)	2.26 [1.72, 2.96]	$3.70 * 10^{-9}$	0.015 (0.94)	1.40 [1.06, 1.85]	0.019	0.0037 (0.85)	1.79 [1.47, 2.17]	$4.50 * 10^{-9}$
rs555778703 Locus: 4q31.21 MAF: 0.009 G (A)	1.54 [1.08, 2.17]	0.016	0.0046 (0.50)	2.00 [1.54, 2.58]	$1.64 * 10^{-7}$	0.0044 (0.74)	1.82 [1.48, 2.24]	$1.65 * 10^{-8}$
rs62262671 Locus: 3p21.31 MAF: 0.08 G (A)	1.18 [1.09, 1.27]	$3.47 * 10^{-5}$	0.062 (0.98)	1.10 [1.05, 1.15]	$7.56 * 10^{-5}$	0.14 (1.0)	1.12 [1.07, 1.16]	$3.55 * 10^{-8}$
Significantly Associated Variants in Meta-Analysis, Absent Nominal Significance in Both Cohorts								
rs80242938 Locus: 16p13.3 MAF: 0.0003 G (A)	7.10 [$9.5 * 10^{-5}$, $5.3 * 10^{6}$]	0.73	0.0067 (0.67)	11.7 [5.17, 26.7]	$4.18 * 10^{-9}$	$8.9 * 10^{-5}$ (0.80)	11.7 [5.16, 26.6]	$3.95 * 10^{-9}$
rs149892036 Locus: 8q12.1 MAF: 0.001 T (C)	1.53 [0.81, 2.88]	0.19	0.0033 (0.80)	2.31 [1.71, 3.12]	$5.37 * 10^{-8}$	0.0015 (0.85)	2.14 [1.63, 2.81]	$4.32 * 10^{-8}$
rs139191981 Locus: 3q26.33 MAF: 0.0005 A (G)	0.88 [0.19, 4.09]	0.88	0.013 (0.90)	7.62 [3.93, 14.8]	$1.87 * 10^{-9}$	0.00015 (0.92)	5.43 [2.95, 9.97]	$4.96 * 10^{-8}$

* rs557046152 (merged into rs78795568 in dbSNP build 151) minor allele frequency from 1000 Genomes Project Phase III EUR (not present in gnomAD). Remaining gnomAD minor allele frequencies derive from the European (non-Finnish) subpopulation frequencies.