medRxiv preprint doi: https://doi.org/10.1101/2023.10.27.23297676; this version posted August 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Genome-wide association study of prostate-specific antigen levels in 392,522 men identifies new loci and improves cross-ancestry prediction

Thomas J Hoffmann[1,2,*], Rebecca E Graff[2,*], Ravi K Madduri[3], Alex A Rodriguez[3], Clinton L Cario[4], Karen Feng[5], Yu Jiang[2,4], Anqi Wang[6,7], Robert J Klein[8], Brandon L Pierce[9,10,11], Scott Eggener[12,11,13], Lin Tong[9], William Blot[14], Jirong Long[14], Louisa B Goss[15], Burcu F Darst[15], Timothy Rebbeck[16], Joseph Lachance[17], Caroline Andrews[16], Akindele O Adebiyi[18], Ben Adusei[19], Oseremen I Aisuodionoe-Shadrach[20], Pedro W Fernandez[21], Mohamed Jalloh[22], Rohini Janivara[17], Wenlong C Chen[23], James E Mensah[24], Ilir Agalliu[25], Sonja I Berndt[26], John P Shelley[27], Kerry Schaffer[28], Mitchell J Machiela[26], Neal D Freedman[26], Wen-Yi Huang[26], Shengchao A Li[26], Phyllis J Goodman[29], Cathee Till[29], Ian Thompson[30], Hans Lilja[31,32], Dilrini K Ranatunga[33], Joseph Presti[34], Stephen K Van Den Eeden[34], Stephen J Chanock[26], Jonathan D Mosley[35,27], David V Conti[6,7], Christopher A Haiman[6,7], Amy C Justice[36,37], Linda Kachuri[4,38,**], John S Witte[4,38,5,39,**]

1. Institute for Human Genetics, University of California San Francisco, San Francisco, CA, USA
2. Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, CA, USA
3. Data Science and Learning Division, Argonne National Laboratory, Argonne, IL, USA
4. Department of Epidemiology and Population Health, Stanford University School of Medicine, Stanford, CA, USA
5. Department of Biomedical Data Science, Stanford University, Stanford, CA, USA
6. Center for Genetic Epidemiology, Department of Population and Preventive Health Sciences, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA
7. Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA
8. Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA
9. Department of Public Health Sciences, University of Chicago, Chicago, IL, USA
10. Department of Human Genetics, University of Chicago, Chicago, IL, USA
11. Comprehensive Cancer Center, University of Chicago, Chicago, IL, USA
12. Department of Urology, University of Chicago, Chicago, IL, USA
13. Department of Surgery, University of Chicago, Chicago, IL, USA
14. Division of Epidemiology, Vanderbilt University Medical Center, Nashville, TN, USA
15. Public Health Sciences, Fred Hutchinson Cancer Center, Seattle, WA, USA
16. Dana Farber Cancer Institute, Harvard School of Public Health, Boston, MA, USA
17. School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, USA
18. Department of Community Medicine, College of Medicine, University of Ibadan, Ibadan, Nigeria
19. 37 Military Hospital, Accra, Ghana

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2023.10.27.23297676; this version posted August 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

20. College of Health Sciences, University of Abuja, Abuja, Nigeria; Cancer Science Centre Abuja, Abuja, Nigeria; University of Abuja Teaching Hospital, Abuja, Nigeria
21. Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa
22. Hospital General Idrissa Pouye, Dakar, Senegal; Ecole Doctorale, Univesite Iba Der Thiam de Thies
23. Strengthening Oncology Services Research Unit, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa; Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa; National Cancer Registry, National Institute for Communicable Diseases a Division of the National Health Laboratory Service, Johannesburg, South Africa
24. Korle-Bu Teaching Hospital and University of Ghana Medical School, Accra, Ghana
25. Dept of Epidemiology and Population Health, and Dept of Urology, Albert Einstein College of Medicine, New York, NY, USA
26. Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA
27. Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA
28. Division of Hematology and Oncology, Vanderbilt University Medical Center, Nashville, TN, USA
29. SWOG Statistics and Data Management Center, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
30. CHRISTUS Santa Rosa Medical Center Hospital, San Antonio, TX, USA
31. Departments of Pathology and Laboratory Medicine, Surgery, Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA
32. Department of Translational Medicine, Lund University, Malmö, Sweden
33. Kaiser Permanente Research Bank, Oakland, CA, USA
34. Division of Research, Kaiser Permanente Northern California, Oakland, CA, USA
35. Department of Internal Medicine, Vanderbilt University Medical Center, Nashville, TN, USA
36. Veterans Administration Connecticut Healthcare System, West Haven, Connecticut, USA
37. Department of Internal Medicine and Yale University School of Public Health, Yale School of Medicine, New Haven, Connecticut, USA
38. Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA, USA
39. Department of Genetics (by courtesy), Stanford University, Stanford, CA, USA

*These authors contributed equally to this work.
**These authors jointly supervised this work.
Correspondence should be addressed to: J.S.W. (jswitte@stanford.edu) and L.K. (lkachuri@stanford.edu)

medRxiv preprint doi: https://doi.org/10.1101/2023.10.27.23297676; this version posted August 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Abstract

We conducted a multi-ancestry genome-wide association study of prostate-specific antigen (PSA) levels in 296,754 men (211,342 European ancestry; 58,236 African ancestry; 23,546 Hispanic/Latino; 3,630 Asian ancestry; 96.5% of participants were from the Million Veteran Program). We identified 318 independent genome-wide significant (p≤5e-8) variants, 184 of which were novel. Most demonstrated evidence of replication in an independent cohort (n=95,768). Meta-analyzing discovery and replication (n=392,522) identified 447 variants, of which a further 111 were novel. Out-of-sample variance in PSA explained by our genome-wide polygenic risk scores ranged from 11.6%-16.6% in European ancestry, 5.5%-9.5% in African ancestry, 13.5%-18.2% in Hispanic/Latino, and 8.6%-15.3% in Asian ancestry, and decreased with increasing age. Mid-life genetically-adjusted PSA levels were more strongly associated with overall and aggressive prostate cancer than unadjusted PSA. Our study highlights how including proportionally more participants from underrepresented populations improves genetic prediction of PSA levels, offering potential to personalize prostate cancer screening.

medRxiv preprint doi: https://doi.org/10.1101/2023.10.27.23297676; this version posted August 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Introduction

Prostate-specific antigen (PSA) is a protein encoded by the *KLK3* gene and secreted by the prostate gland[1–3]. PSA levels are often elevated in those with prostate cancer (PCa); however, elevated levels can also be caused by other factors, such as benign prostatic hyperplasia (BPH), local inflammation or infection, prostate volume, age, and germline genetics[4–8]. PSA screening for PCa was approved by the Food and Drug Administration in 1994, but it is unclear if the benefits for PCa-specific mortality reduction outweigh the harms from overdiagnosis and treatment of clinically insignificant disease[9–12]. Previous work estimates 20-60% of screen-detected PCas are overdiagnoses (i.e., PCa that would not otherwise clinically manifest, or result in PCa-related death[13]), other work has suggested that 229 individuals would need to be invited to screen and 9 would need to be diagnosed to prevent one death[14], and the United States[15], Canada[16], and the United Kingdom[17] recommend against universal population-based screening. If PSA levels could be adjusted for an individual's predisposition in the absence of PCa, the specificity (to reduce overdiagnosis) and sensitivity (to prevent more deaths) of screening could be improved.

Twin studies estimate PSA heritability at 40-45%[18,19], and genome-wide heritability estimates at 25%-30%[20] suggesting that incorporating genetic factors may improve screening. Recent work from our group based on 85,824 European ancestry and 9,944 non-European ancestry men found that genetically-adjusted PSA (i.e., the PSA measure inflated or deflated due to an individual's genetic variants) most improved the discrimination of PSA screening for aggressive tumors[20]. In that work, we identified 128 genome-wide significant variants that explained up to 7% of PSA variation in European ancestry, suggesting that many more PSA loci remain. Genome-wide polygenic risk scores (PRSs) explained up to 10% in European ancestry; however, the PRSs were less predictive in other groups, especially African ancestry (1-3%). Additional variant discovery with larger, more diverse cohorts could provide novel insights into the genetic architecture of PSA and further improve PCa screening.

Results

Composition of discovery and replication cohorts

Our discovery population consisted of 296,754 men without PCa from 9 cohorts: 211,342 European ancestry (71.2%), 58,236 African ancestry (19.6%), 23,546 Hispanic/Latino (7.9%), and 3,630 Asian ancestry (1.2%). None had been included in previous PSA GWASs. We present genotype platform details in **Table S1**, demographics in **Table S2**, and quality control (QC) metrics in **Table S3**. The pooled mean age at PSA measurement across the discovery cohorts was 57.4 (standard deviation [SD]=9.6), and the pooled median PSA was 0.84. The Million Veteran Program (MVP) comprised 96.5% of the discovery cohort. For replication, we utilized results from 95,768 independent individuals from previous work[20], including 85,824 European ancestry, 3,509 African ancestry, 3,098 Hispanic/Latino, and 3,337 Asian ancestry individuals (**Table S3**). **Figure 1** summarizes our analytical workflow and describes cohort ancestry compositions.

Discovery GWAS analysis of PSA-associated variants

medRxiv preprint doi: https://doi.org/10.1101/2023.10.27.23297676; this version posted August 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

In our discovery cohorts, we identified 318 independent genome-wide significant variants (264 European, 51 African, 17 Hispanic/Latino, and 2 Asian) in a multi-ancestry analysis of log-transformed PSA levels (**Figure 2**, **Figure S1**, **Table S4**, **Table S5**) that used multiple reference panels to account for different ancestries (see **Methods**). Among them, 184 independent variants selected by mJAM[21] were novel (see **Methods**). Of the novel variants, 57 replicated at a Bonferroni level ($p<0.05/184=0.00027$, same direction of effect on PSA), an additional 80 replicated at $p<0.05$ (and the same direction), 43 demonstrated the same effect direction (but $p>0.05$), and four showed no indication of replication (effect in the opposite direction). On average, compared to the non-replicated variants, the replicated variants had slightly larger effect sizes (mean $\beta=0.30$ versus 0.27) and were slightly more precise (mean standard error=0.0039 versus 0.0042).

Of the 184 variants that were novel in the multi-ancestry analysis, 112 were genome-wide significant in the European ancestry discovery cohort, eight were genome-wide significant in the African ancestry cohort, and none were genome-wide significant in Asian or Hispanics/Latino individuals (likely due to low sample size; **Figure S2**). Of the eight in African ancestry individuals, only two variants were frequent enough (see **Methods**) to be assessed in other ancestry groups: rs2071041 (*ITIH4*; $\beta_{African}=0.0237$, 95% CI=0.0152-0.0322, $p_{African}=4.9e-8$) that was also genome-wide significant in European ancestry individuals ($\beta_{European}=0.0180$, 95% CI=0.0124 to 0.0235, $p_{European}=2.6e-10$, minor allele frequency (MAF)=23.7%), and rs1203888 (*LINC00261*; $\beta_{African}=-0.0423$, 95% CI=-0.0539 to -0.0307, $p_{African}=8.7e-13$) that was not significant in European ancestry individuals ($p>.05$, MAF=0.8%). The latter variant showed similar magnitude of effect but was not Bonferroni significant in discovery Hispanic/Latino individuals ($\beta_{Hispanic/Latino}=-0.0748$, 95% CI=-0.120 to -0.0297, $p_{Hispanic/Latino}=0.0012$, MAF=3.1%) and was not significant in discovery Asian ancestry ($p>.05$, MAF=3.5%) or the replication cohorts ($p>.05$) (**Table S4**). The remaining six African ancestry variants were too rare to be assessed in European ancestry individuals. The variant rs184476359 (*AR*, multi-ancestry discovery $\beta=-0.0590$, 95% CI=-0.0774 to -0.0406, p=3.4e-10; replication $\beta=-0.0870$, 95% CI=-0.1370 to -0.0371, p=6.3e-4) was common in African ancestry individuals (MAF=17.7%), less common in Hispanic/Latinos (MAF=1.1%), and not adequately polymorphic to be imputed in Asian individuals. Three variants in genes that encode PSA (rs76151346, $\beta_{African}=0.0821$, 95% CI=0.0577 to 0.107, $p_{African}=4.6e-11$, *KLK3*; rs145428838, $\beta_{African}=0.224$, 95% CI=0.165 to 0.284, $p_{African}=1.4e-13$, *KLK3*; rs182464120 $\beta_{African}=-0.213$, 95% CI=0.278 to 0.147, $p_{African}=2.0e-10$, *KLK2*) exclusively imputed in African ancestry individuals (all MAF<5%, two <1%) did not exhibit strong evidence of replication in African ancestry individuals($p>0.05$). The remaining two variants identified in African ancestry (rs7125654, $\beta_{African}=-0.384$, 95% CI=-0.0489 to -0.0279, $p_{African}=7.0e-13$; rs4542679, $\beta_{African}=0.0422$, 95% CI=0.0288 to 0.0557, $p_{African}=7.9e-10$), were more common (MAF>5%) but also did not replicate ($p>0.05$). Further, rs7125654 (*TRPC6*) was less common in Latinos, but more common in Asian ancestry, and rs4542679 (*RP11-345M22.3*) was less common in Latinos and not adequately polymorphic in Asians.

We next tested for effect size differences across ancestry groups for the 184 novel variants. Only one variant, rs12700027 (*BRAT1*/*LFNG*, $I^2=84.8$, p=0.00019), demonstrated heterogeneity

medRxiv preprint doi: https://doi.org/10.1101/2023.10.27.23297676; this version posted August 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

that was Bonferroni-significant (p<0.05/184=0.00027). The variant had a strong discovery effect in European ancestry individuals (β=0.0327, 95% CI=0.0247 to 0.0407, p=1.2e-15, MAF=0.10), but was not significant in other groups (African ancestry β=0.0131, 95% CI=-0.0190 to 0.0452, p=0.42, MAF=0.021; Asian β=-0.176, 95% CI=-0.0203 to 0.0797, p=0.027, MAF=0.021; Hispanic/Latino β=-0.0102, 95% CI=-0.0326 to 0.0121, p=0.37, MAF=0.120). In our replication cohort, the variant nominally (i.e., p<0.05) replicated (p=0.0065, β=0.0175, 95% CI=0.00491 to 0.0302; European ancestry p=0.003, β=0.0327, 95% CI=0.0247 to 0.0407) and showed no statistically significant evidence of differences across ancestry groups (I^2=0.0, p=0.44), although sample sizes for detecting differences were smaller.

In-silico assessment of potential functional features revealed that 20 of the novel variants (10.8%) were prostate tissue expression quantitative trait loci (eQTLs), and another 65 (35.3%) were eQTLs in other tissues (**Table S4**). Five novel variants were missense and predicted to be deleterious, with Combined Annotation Dependent Depletion (CADD) scores >20 (**Table S4**): rs11556924 in *ZC3HC1*, which regulates cell division onset; rs74920406 in *ELAPOR1*, a transmembrane protein; rs2229774 in *RARG*, a gene in the hormone receptor family; rs113993960 (delta508) in *CFTR*, a causal mutation for cystic fibrosis[22]; and rs2991716 upstream of LOC101927871. An additional 11 variants were predicted to have high pathogenicity (CADD scores >15;**Table S4**).

Replication analysis of previously-reported variants in the discovery cohort

When we tested 128 previously identified variants[20] in our discovery cohort, 106 (82.8%) replicated at a genome-wide significance level, an additional 15 replicated (11.7%) at a Bonferroni level (p<0.05/128=0.00039), an additional 6 replicated at p<0.05 (4.7%), and one variant flipped effect direction (**Table S6**). Replication was highest for European ancestry, likely due to sample size, with 94 variants (73%) reaching genome-wide significance, an additional 22 variants (17.2%) meeting a Bonferroni-corrected level, and 8 (6.3%) additional variants meeting p<0.05 (**Table S6**). Replication rates within African ancestry, our next largest group, were lower: 16 (12.5%) were genome-wide significant, 26 others (20.3%) met a Bonferroni level, an additional 39 (30.5%) had p<0.05, 32 additional (25.0%) were in the same direction, and 15 (11.7%) were in the opposite direction. Estimated rates were similar for Hispanic/Latino and lowest for Asian populations. Lastly, 16 of the 128 known variants showed heterogeneity across the four groups (Bonferonni corrected p<0.05/128=0.00039).

Joint meta-analysis of discovery and replication cohorts

In the multi-ancestry analysis including the discovery and replication cohorts, we identified 447 independent variants (409 European, 56 African, 22 Hispanic/Latino, 6 Asian, including 46 in >1 group; **Figure 3, S1**; **Table S7**, **S8**). Among the 111 variants that were novel even relative to discovery alone, none showed evidence of ancestry effect size differences (p>0.05/111=0.00045). Fifty-six (50.4%) of the 111 were genome-wide significant in European ancestry individuals, but none were genome-wide significant in a non-European ancestry group (**Table S8**). Allele frequencies and effect sizes of the novel variants largely followed those expected by power curves (**Figure 4**).

medRxiv preprint doi: https://doi.org/10.1101/2023.10.27.23297676; this version posted August 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

In the joint meta-analysis, 12 (10.8%) novel variants were prostate tissue eQTLs, and 50 (45.0%) additional were eQTLs for other tissues. Two were missense substitutions (**Table S7**): rs1049742 in *AOC1* and rs74543584 in *MPZL2*. Three additional novel variants had CADD scores>15: rs1978060, an eQTL for *TBX1* in prostate tissue; rs339331 an eQTL for *FAM162B* in adipose tissue; and rs57580158, an intergenic variant with evidence of conservation.

Medication sensitivity analysis

The UK Biobank (UKB) had medication information available to conduct a sensitivity analysis to see whether excluding individuals taking medications that could affect PSA levels (i.e., 5-alpha reductase inhibitors and testosterone) would impact our results. For PSA-associated variants, our primary results in the UK Biobank (UKB) were highly correlated (R=0.93, **Figure S3**) with the sensitivity analyses, suggesting these medications did not impact our results.

Out-of-sample PSA variance explained by PRSs

We evaluated different strategies for constructing PRSs for PSA levels first using results from our discovery cohort (see **Methods**). For testing these PRSs, four cohorts of men without PCa were out-of-sample: Kaiser Permanente's Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort, the Selenium and Vitamin E Cancer Prevention Trial (SELECT),[23] the Prostate Cancer Prevention Trial (PCPT),[24] and All of Us (AOU)[25].

In GERA, PRS_{318}, constructed from the 318 conditionally-independent genome-wide significant variants in the multi-ancestry meta-analysis, generally had higher variance explained when using longitudinal measurements, rather than earliest PSA value, with 13.9% (95% CI=13.1%-14.6%) in European ancestry (n=35,322), 13.1% (95% CI=10.6%-15.6%) in Hispanics/Latinos (n=2,716), 9.3% (95% CI=6.8%-12.0%) in African ancestry (n=1,585), and 9.0% (95% CI=7.0%-11.4%) in East Asian ancestry (n=2,518). The variance explained in the other three cohorts was ~3-6% lower depending on the group (**Table S9**).

Expanding to a genome-wide approach, PRS-CSx ($PRS_{CSx\text{-}disc}$; included more than genome-wide significant variants (1,070,230 SNPs; see **Methods**) and resulted in improved predictive performance. The variance explained increased to 16.6% (95% CI=15.9%-17.5%) in men of European ancestry and 18.2% (95% CI=15.4%-20.8%) in Hispanic/Latino men (**Figure 5A**, **Table S9**). The relative increase was largest in East Asian ancestry, with variance explained reaching 15.3% (95% CI=12.7%-18.1%), and smallest in African ancestry, with variance explained 8.5% (95% CI=6.1%-11.0%).

Second, we developed PRSs for PSA using the results from the joint GWAS meta-analysis (n=392,522), which combined the discovery meta-analysis with previously published results from Kachuri et al[20]. These scores were validated in PCPT, SELECT, and AOU, but not GERA, which was included in the previously published meta-analysis and was therefore not out-of-sample .

medRxiv preprint doi: https://doi.org/10.1101/2023.10.27.23297676; this version posted August 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

For the independent genome-wide significant PRSs, PRS_{318} explained 9.5% (8.8%-10.3%) of variation in baseline PSA levels in SELECT European ancestry (n=22,173), while PRS_{447} (from the 447 conditionally independent genome-wide significant variants identified in the joint meta-analysis) explained 10.9% (10.2%-11.8%) of the variance, which exceeded the 8.5% (95% CI=7.8%-9.2%) variance explained by PRS_{128} (from the 128 independent variants described in our prior GWAS of 95,768 men[20]). Variance explained in PCPT European ancestry (n=5,725) was slightly lower. In AOU European ancestry (n=11,922), variance explained was slightly higher, with PRS_{128} explaining 8.6% (95% CI=7.7%-9.6%), PRS_{318} explaining 9.6% (95% CI=8.6%-10.6%), and PRS_{447} explaining 11.3% (95% CI 10.2%-12.4%). We did not observe an appreciable change in differences across cohorts after removing individuals with BPH, a condition known to influence PSA levels. However, variance explained was slightly higher in all populations (<0.5% higher), albeit with overlapping CIs (**Table S9**).

Among SELECT African ancestry (n=1,173), PRS_{128} explained 3.4% (95% CI=1.6%-5.8%), PRS_{318} explained 6.5% (95% CI=4.0%-9.5%), and PRS_{447} explained 7.0% (95% CI=4.5%-10.1%); PRS_{447} more than doubled variance explained by PRS_{128}. AOU African ancestry (n=2,471) estimates were 1-2% smaller.

A genome-wide PRS-CSx ($PRS_{CSx\text{-}joint}$) compared to $PRS_{CSx\text{-}disc}$ modestly increased variance explained by roughly 1-1.5% in European ancestry in PCPT (11.6%, 95% CI=10.0%-13.1%), SELECT (13.9%, 95% CI=13.1%-14.9%), and AOU (14.7%, 95% CI=13.5%-16.0%). $PRS_{CSx\text{-}joint}$ also improved ~3% upon the previously published PRS from the Kachuri et al.[20] paper estimated here to be 8.6% in PCPT and 10.4% in SELECT for PRS-CSx ($PRS_{CSx\text{-}Kachuri}$, **Table S9**). Among men of African ancestry in SELECT, $PRS_{CSx\text{-}joint}$ showed no improvement (7.2%, 95% CI=4.6%-10.0%) over $PRS_{CSx\text{-}disc}$, while variance explained in AOU increased by 0.3% (5.8%, 95% CI=4.1%-7.8%). Notably, $PRS_{CSx\text{-}joint}$ yielded a substantial improvement upon the previously published $PRS_{CSx\text{-}Kachuri}$ estimates[20] of 1.64% in SELECT, although still under half of that in European ancestry.

In SELECT European ancestry individuals, PRS-CSx explained 13.9% of variation, while PRS_{447} explained 10.9% of the variation. Assuming that variation explained is nested between these approaches, we estimate 78.4% (=10.9%/13.9%) of PRS-CSx variation may be explained by PRS_{447}. This is expected since information across the different PRSs overlaps, and the initial genome-wide significant SNPs from our large-scale GWAS are the most informative for explaining variation in PSA levels.

Third, we examined how the variance in PSA levels explained by the PRSs varied by age. These analyses were performed in GERA to have a large enough sample size in each age group and used $PRS_{CSx\text{-}disc}$ to provide out-of-sample estimates. The estimated variance explained by the PRSs decreased with increasing age in all GERA ancestry groups, albeit with somewhat wide confidence intervals (**Figure 5B, Table S10**). For example, $PRS_{CSx\text{-}disc}$ explained 16.4% (95% CI 14.6%-18.5%) of variation in PSA levels among European ancestry individuals <50y, and this decreased to 8.7% (95% CI 7.0%-10.5%) for men ≥80y.

medRxiv preprint doi: https://doi.org/10.1101/2023.10.27.23297676; this version posted August 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Finally, for PRSs constructed from genome-wide significant independent variants, the variance explained when using weights corresponding to effect sizes from the multi-ancestry meta-analysis was almost always equal to or higher than variance explained when using ancestry-specific weights (**Table S10**). This was observed both for the discovery (PRS_{318}) and joint meta-analysis (PRS_{447}). The few instances where the variance explained was estimated lower almost always had <1% difference and wide confidence intervals around the estimate (i.e., the smallest sample sizes likely had unstable estimates).

Relationship of PSA PRSs with PCa aggressiveness using Gleason score

In GERA, we performed a case-only analysis to examine the association between PSA $PRS_{CSx,disc}$ (the out-of-sample PRSs with the highest variance explained) and Gleason score. Results were consistent with previous work that suggested that screening bias decreases the likelihood of identifying high-grade disease, whereby men with higher PRS values (indicating a genetic predisposition to higher constitutive PSA levels) are more likely to be biopsied, but less likely to have high grade disease[20]; we found that in European ancestry cases, an SD increase in $PRS_{CSx\text{-}disc}$ was inversely associated with Gleason 7 (OR=0.78, 95% CI=0.73-0.84, p=1.2e-13) and ≥8 (OR=0.71, 95% CI=0.64-0.79, p=6.2e-10) compared to Gleason ≤6. Other ancestry groups had similar estimated ORs, though they were not always statistically significant, likely owing to sample size (**Table S11**; e.g., African ancestry Gleason 7 (OR=0.88, 95% CI=0.67-1.17, p=0.39) and ≥8 (OR=0.65, 95% CI=0.43-0.99, p=0.043)).

Impact of genetically-adjusted PSA on prostate biopsy eligibility

We examined how $PRS_{CSx,disc}$ would have changed biopsy recommendations for cases and controls, according to age-specific thresholds in GERA (see **Methods**). In European ancestry individuals who had negative biopsies (i.e., controls, n=2378), 16.0% with unadjusted PSA levels that exceeded age-specific thresholds for biopsy were reclassified to ineligible for biopsy. Among controls with PSA levels that did not indicate biopsy, 2.4% were reclassified to biopsy eligible, resulting in a control net reclassification improvement (NRI) of 13.6% (95% CI=12.2%-15.0%; **Figure 6A**, **Table S12**). In individuals with positive biopsies (i.e., cases; n=2,358), 3.9% were reclassified to eligible, while 13.1% were reclassified to ineligible, resulting in a case NRI of -9.2% (95% CI=-10.3% to -8.0%). Of cases who became ineligible, 71.1% had Gleason scores ≤7, as compared to 56.5% who remained eligible (although we note that some of these men may have had biopsies for reasons other than their PSA level, such as abnormal digital rectal exam or strong family history). In African ancestry controls (n=110), 16.0% were reclassified to ineligible, while 2.4% were reclassified to eligible, resulting in an NRI of 3.6% (95% CI=0.1% to 7.1%; **Figure 6B**). In African ancestry cases (n=310), 5.2% were reclassified to eligible and 6.8% were reclassified to ineligible, resulting in an NRI of -1.6% (95% CI=3.0% to -0.2%). Other groups are shown in **Figure S4** with details in **Table S12**. We additionally obtained eight years of additional follow up on the 78 controls in all groups now classified as eligible; three were later diagnosed with PCa.

medRxiv preprint doi: https://doi.org/10.1101/2023.10.27.23297676; this version posted August 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

To assess potential variability in genetic adjustment across PSA levels, we compared measured versus genetically adjusted PSA levels across a range of values in GERA men. Plotting measured versus adjusted PSA indicated consistent relative adjustment on the log scale (**Figure S5**). For example, at a measured PSA of 2.5, 6.5, and 10.0 ng/ml, the genetically adjusted PSA' IQR ranged from 2.6-3.7, 4.6-6.8, and 9.1-12.5, respectively. These results suggest genetic adjustment is applicable to PSA values <20 displayed in the figure, although the implications are most profound around values where clinical decisions are made (e.g., near age-specific PSA thresholds).

Impact of genetically-adjusted PSA on overall and aggressive PCa

Previous work has suggested that midlife PSA predicts lethal PCa[26]. In GERA European ancestry, genetically-adjusted midlife log PSA had a larger estimated magnitude of association with overall PCa (OR=4.57, 95% CI=4.27-4.88) than measured PSA (OR=4.30, 95% CI=4.04-4.58)), though the CIs overlapped. The difference between associations was even larger for aggressive disease, with OR=3.92 (95% CI=3.54-4.35) for adjusted vs. OR=3.46 (95% CI=3.15-3.81) for measured, though again CIs overlapped. African ancestry showed similar trends, with the genetically-adjusted association with PCa OR=5.85 (95% CI=4.73-7.23) vs. measured OR=4.72 (95% CI=3.56-6.27), and the aggressive genetically-adjusted OR=5.39 (95% CI=3.95-7.35) vs. measured OR=4.72 (3.56-6.27). Estimates in Hispanic/Latino were also similar, but East Asian ancestry showed no difference (**Table S13**). Cross-validated AUC estimates also showed essentially no difference between adjusted and measured PSA, with estimates ranging from 0.7-0.8 in the different groups (**Table S13**).

Associations with previously-reported PCa variants

In our discovery cohort, 20 of our 184 novel PSA-associated variants (10.8%) were genome-wide significantly associated with PCa in the PRACTICAL consortium's European ancestry GWAS[27] (**Tables S4, S6**). An additional 19 variants (10.3%) were associated with PCa at a Bonferroni level (p<0.05/184=0.00027). With correction for bias related to more frequent screening in men with higher constitutive PSA levels (see Methods)[20,28], this count was reduced to 13 (7.0%) significant at the genome-wide level, and an additional 14 (7.6%) at the Bonferroni-corrected level. Out of the 111 novel PSA-associated variants from the meta-analysis, 8 (7.1%) were genome-wide significantly associated with PCa, and an additional 11 (9.8%) were significant at a Bonferroni level (p<0.00045). With bias correction, 5 (4.5%) were genome-wide significant, and an additional 4 (3.6%) were Bonferroni-significant.

Associations with previously reported BPH variants

In our discovery cohort, one variant (rs1379553) was genome-wide significantly associated with BPH out of 137 variants available in a GWAS in UKB European ancestry[29]. An additional eight met a Bonferroni level (p<0.05/137=0.00036). Out of the 96 available variants identified from the meta-analysis, one variant was genome-wide significant (rs627320), and 6 more met a Bonferroni level (p<0.045/96=0.00052).

medRxiv preprint doi: https://doi.org/10.1101/2023.10.27.23297676; this version posted August 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Associations with urinary symptom variants

In variants identified in our discovery cohort, rs12573077 (p=8.4e-5) met a Bonferroni level (p<0.05/177=0.00028) for a test for an association with urinary symptoms in the GERA cohort (**Table S4, S6, S14**). In variants identified in the joint meta-analysis, none met a Bonferroni level of significance (p<0.05/110=0.00045).

PSA variant associations with prostate volume

We found that 31 of the 407 PSA variants tested demonstrated some evidence of an association with prostate volume in the PASS cohort. rs182464120 on chromosome 19 was strongly associated (p=2.0e-11), rs12344353 on chromosome 9 was associated at a Bonferroni level (p=5.3e-5 <0.05/407=0.00012), and 29 other variants were associated at a nominal significance level (p<0.05) (**Table S15**).

Associations with KLK3 plasma pQTL

We annotated the 447 variants from the joint GWAS meta-analysis using recently published plasma pQTL association results for *KLK3* from the UKB Pharma Proteomics Project using the Olink Explore platform;[30] 409 had corresponding pQTL associations in the UKB. In European ancestry individuals (n=46,214), GWAS and *KLK3* pQTL effects were highly correlated (R=0.85, p=3.7e-117) (**Figure S6**). Eleven variants were associated with relative *KLK3* abundance at p<0.05/409, and, as expected, the strongest two associations were in *KLK3* (rs17632542, rs61752561) (**Table S16**). Among African ancestry individuals (n=1,065), we observed an attenuated correlation with effects on *KLK3* abundance (R=0.14, p=0.0034), although none of the individual pQTL associations reached statistical significance.

Associations with eGenes

In our sc-RNA-seq data analysis, we found that each of the eGenes for PSA-associated variants is expressed across prostate cells. This is especially true in prostate luminal epithelial cells (which produce PSA) and is expected if the genes modify PSA levels (**Table S17**). **Figure S7** shows expression of the eQTL genes across multiple prostate tissue cell types, including luminal cells of the prostate epithelium and its precursor cells (e.g., basal epithelial cells of prostate; expression sorted by *KLK3*). Percentile expression of the eQTL-associated genes was significantly higher in luminal cells than all other prostate cell types (P=0.0006), suggesting these genes are more active in this cell type than other prostate cells (**Figure S8**) and supporting the hypothesis that these eQTL genes are involved in PSA expression.

Discussion

Our GWAS detected 448 genome-wide significant variants associated with PSA levels, of which 295 were novel (184 in discovery and 111 in a meta-analysis), nearly quadrupling the total number of associated variants. The variance explained by genome-wide PRSs ranged from 11.6%-16.9% in men of European ancestry, 5.5%-9.5% in men of African ancestry, 13.5%-18.6% in Hispanics/Latinos, and 8.6%-15.3% in the East Asian ancestry group. We also observed a decline in PRS predictive performance with increasing age, particularly at the oldest ages. The majority of newly identified variants were uniquely associated with PSA and not PCa.

medRxiv preprint doi: https://doi.org/10.1101/2023.10.27.23297676; this version posted August 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Our discovery cohort included more African ancestry individuals than any prior study of PSA genetics. Of the eight genome-wide significant variants that were identified in the discovery phase in African ancestry, only two were sufficiently common to be assessed in European ancestry. Of those two, the association between rs1203888 (*LINC00261*) and PSA levels was unique to the African ancestry population. These eight variants generally failed to meet Bonferroni significance in our replication cohort, although the sample size was small (3,509 individuals of African ancestry); the variant rs18447639 in the *AR* gene was closest to replicating. Androgen receptor (AR) signaling is required for normal prostate development and function but is hijacked during carcinogenesis.[31] Because prostate tumor growth and progression depend on AR signaling, androgen deprivation therapy remains a frontline treatment for progressing PCa, and the inhibition of *AR* activity may delay progression.[32]

A total of 10.8% of novel discovery and replication variants were prostate tissue eQTLs, and another 49.7% were eQTLs in other tissues. In addition, 16 discovery variants and five meta-analysis variants had predicted deleterious regulatory effects. Putative deleterious genes included: *AOC1*, which regulates histamine metabolism and sensitivity to non-steroidal anti-inflammatory drugs[33,34]; *MPZL2*, which is involved in thymus development and T-cell maturation; and *ZC3HC1*, a regulator of cell cycle progression and established susceptibility locus for coronary artery disease[35,36]. We also observed an association with PSA levels for the deltaF508 mutation in *CFTR* that causes cystic fibrosis, which is accompanied by infertility in 97% of affected males[37] and has been linked to obstructive azoospermia (ClinVar[38] accession SCV001860325). We detected another signal with possible links to male fertility, rs372203682 in *LMTK2*, a gene implicated in spermatogenesis[39] that interacts with AR and inhibits its transcriptional activity[40].

In SELECT, the variance in PSA levels explained by our independently associated GWAS variants was ~1% larger than previously explained[20] in European and ~3% higher in African ancestry individuals. The variance explained in SELECT and PCPT was substantially less than that in GERA, even though we evaluated only variants from our discovery cohort (which did not include GERA). This may be due in part to the studies' selection criteria, as individuals in SELECT and PCPT were required to have PSA≤3 ng/mL[23] and ≤4 ng/mL[24], respectively, at baseline. However, in AOU, which did not have this selection criteria, variance explained for European ancestry men was at most 0.5% higher than SELECT, and thus also lower than GERA. For African ancestry men in AOU, variance explained was 2-3% lower than in SELECT, suggesting that differences in performance may be attributed to factors other than preferential selection for low baseline PSA. We also investigated whether BPH may contribute to variability in PRSs performance. The estimated variance explained was <0.5% higher when excluding men with a BPH diagnosis. By BPH, we mean a clinical diagnosis rather than presence of BPH. Most patients evaluated for potential PCa have evidence of BPH, which can result in elevated PSA levels[41,42]. These findings highlight the need to evaluate genetically adjusted PSA in a wider range of clinical settings, as well as the challenges with curating out-of-sample cohorts with clinical data sufficient for such evaluations.

medRxiv preprint doi: https://doi.org/10.1101/2023.10.27.23297676; this version posted August 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

The performance of PRS constructed using weights derived from the multi-ancestry meta-analysis typically matched or surpassed the performance of PRSs based on ancestry-specific weights. As expected, genome-wide PRS-CSx generally achieved 1-6% higher explained variance than the PRSs limited to mJAM genome-wide significant variants. However, the improvement in performance observed for PRS-CSx was not equal across populations. The largest increase was observed for Hispanic/Latino men, followed by European ancestry men (generally similar to Hispanic/Latino men), followed by Asian ancestry men. To our knowledge, this is the first time out of sample PRS performance was assessed in a Hispanic/Latino population. Relative to the fine-mapped PRS, the degree of improvement was smallest for African ancestry. This may be due to a number of factors. First, PRS-CSx uses a single hyperparameter to couple posterior effect sizes across ancestry groups, which may be insufficient to capture different correlation structures among populations. Second, HapMap3 variants used by PRS-CSx do not tag genetic variation equally well across ancestries. Fine-mapping PRS methods do not limit to this set of tagging SNPs and may be more likely to capture population-specific variants. Third, the choice of linkage disequilibrium (LD) reference panels has slightly different implications for the two PRS approaches. PRS-CSx relies on LD reference panels for estimating joint SNP effect sizes, while fine-mapping requires LD information for identifying independent variants from summary statistics. mJAM advances other fine-mapping approaches by incorporating population-specific LD, which is more accurate than using a single population as the LD reference[21] or only making use of the largest ancestry group. While PRS-CSx provides more flexibility to accommodate different genetic architectures, it may be more sensitive to the choice of LD reference panels and mismatches in LD structure between PRS training and testing populations, especially without a separate dataset for parameter tuning.

We found that genetically adjusting PSA levels reduced unnecessary biopsies in controls, albeit less so than in previous work[20] in the same GERA population. It is likely that our previous study overestimated reclassification in controls because there was partial overlap between the GWAS meta-analysis used to train the PRS used for adjustment - GERA was included - and the population in which we undertook genetic adjustment. In the present study, we performed genetic adjustment using a PRS trained on a large GWAS that did not include GERA. We also saw an increase in magnitude in genetically-adjusted midlife PSA association with PCa in most GERA groups, although CIs overlapped for all, and while our previous study[20] did not see any benefit in African ancestry, here we saw an increase.

Our investigation had several limitations. Relative to prior studies of PSA genetics, the discovery and replication cohorts included here substantially increased the number of men from diverse populations. While both were very large (N~300K and ~100K men total), the replication cohorts had disproportionately smaller African ancestry (discovery N ~58K and replication N ~3.5K) and Hispanic/Latino populations (N~24K and ~3K). Going forward, our PSA Consortium will continue to seek new study populations with both genotypic and phenotypic data that represent diverse participants. Nevertheless, for African ancestry, 43% of variants met a nominal replication threshold, many more than the 5% that would be expected by chance. We also suspect that we

medRxiv preprint doi: https://doi.org/10.1101/2023.10.27.23297676; this version posted August 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

had limited power to detect effect size heterogeneity, especially since variants that exhibited significant heterogeneity were mostly known variants in strongly associated regions. Another limitation is that GERA biopsy reclassification may have been specific to Kaiser Permanente clinical guidelines, as previously discussed[20]. In addition, while we did our best to restrict relevant analyses to PCa-free individuals, some individuals likely had undetected PCa[43]. However, the number was unlikely to be large enough to materially impact our results because our study population was relatively young; the average age among men in the MVP (which comprised a large majority of our discovery study population, N~286K) was 58 years for men of European ancestry, 52 years for men of African ancestry, and 54 years for Hispanic/Latino men or Asian ancestry men. Further, the PSA PRS_{CSx} explained increasing variability in PSA levels for individuals of younger ages. Most novel PSA-associated variants were not associated with PCa, and those that were may have been due to screening bias, as previously shown[20]. The lack of BPH information in most of our cohorts was an additional limitation, but most novel variants associated with PSA levels were not associated with BPH in others' work on UKB European ancestry individuals[29], and the variance explained by PRSs in SELECT was affected by <0.5% in participants with BPH. We were unable to account for prostate volume, a strong predictor of PSA levels[44]. Finally, we note that our GWAS and resulting PRSs were developed for total PSA. Future work should work toward capturing genetic factors that are specific to constituents of total PSA.

In summary, we undertook a large-scale, multi-ancestry study with over three times the sample size of previous work,[20] expanded our understanding of the genetic basis of PSA levels and our potential to improve the accuracy of PSA genetic adjustment across ancestries. Using an ancestrally diverse study population, we detected hundreds of novel variants associated with PSA levels that were largely independent of PCa and BPH. These findings explain additional variation in PSA levels, especially among men of African ancestry, who suffer the highest morbidity and mortality due to PCa, as well as among Hispanic/Latino men. This highlights the importance of studying diverse populations to enable novel discoveries and construct PRS that will perform equally across ancestry groups. Taken together, our work moved us closer to leveraging genetic information to personalize and substantially improved our understanding of the genetic basis of PSA levels and of genetic adjustment of PSA levels across individuals of diverse ancestries.

Methods

Discovery participants and phenotype measurements

Our primary analyses included 296,754 men from seven cohorts that had not previously been analyzed in studies of PSA genetics. These are described briefly below; additional details, including array, ancestry, imputation reference panels, sample sizes, number of variants, and standard filters applied are described in **Tables S1-S3**. To ensure participants had a functional prostate unaffected by surgery or radiation and to exclude individuals at a high risk of undiagnosed PCa[45], participants were restricted to men with no history of PCa or surgical resections of the prostate, and at least one PSA measurement between 0.01 and 10ng/mL. Analyses were based on each individual's earliest recorded PSA level. For descriptive statistics,

medRxiv preprint doi: https://doi.org/10.1101/2023.10.27.23297676; this version posted August 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

meta-analysis of PSA medians from each cohort was done with the weighted median of medians method in the R v4.2.3[46] package metamediation v1.0.0[47]. Sub-populations were defined by self-identified race/ethnicity and/or genetically-inferred ancestry, depending on the cohort.

African American Prostate Consortium (AAPC). The AAPC is comprised of African ancestry studies with PCa phenotyping.[27]

Mount Sinai BioMe® Biobank (BioMe). BioMe is a longitudinal cohort linked to Epic electronic health records (EHRs)[48]. Individuals were European ancestry, Hispanic/Latino, and African ancestry.

Chicago Multiethnic Prevention and Surveillance Study (COMPASS). COMPASS is a longitudinal study of Chicagoans with >11,000 participants currently enrolled (82% African American).[49] PSA data has been described previously[50].

Men of African Descent and Carcinoma of the Prostate (MADCaP). MADCaP is a consortium of epidemiologic studies addressing the high PCa burden in African ancestry men.[51,52]

Multiethnic Cohort (MEC). MEC is a prospective cohort study that enrolled >215,000 Hawaii/Los Angeles residents ages 45-75 years between 1993-1996.[53,54]

Million Veteran Program (MVP). MVP is a multi-ancestry cohort recruited nationwide. Information is obtained from EHRs, including inpatient International Classification of Diseases (ICD)-9 codes, Current Procedural Terminology (CPT) procedure codes, clinical laboratory measurements, and reports of diagnostic imaging modalities[55]. Sub-populations were created using the harmonized ancestry and race/ethnicity (HARE) method.[56]

Southern Community Cohort Study (SCCS). SCCS is a prospective cohort study that recruited 85,000 predominantly African ancestry adults from community health centers in the southeastern United States. This study included only men of African ancestry.[57]

Replication cohorts

Genome-wide significant variants identified in the discovery cohort were tested for replication in the previous largest GWAS of PSA levels, which included 95,768 men (85,824 European ancestry, 89.6%)[20], using a Bonferroni corrected α level. In addition, previously-identified genome-wide significant variants [20] were tested for replication in our independent discovery cohort. Statistical tests throughout were two-sided.

Additional PRS evaluation cohorts

For our discovery cohort results, we evaluated PSA PRS performance and reclassification in individuals from the GERA cohort (also in the replication cohort, out-of-sample for the discovery

medRxiv preprint doi: https://doi.org/10.1101/2023.10.27.23297676; this version posted August 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

cohort (n=35,322; 28,503 European; 2,716 Latino; 2,518 East Asian; and 1,585 African American)).

Additional out-of-sample cohorts for (both the discovery analysis and the joint meta-analysis of discovery and replication) PRS assessment was done in genotyped individuals from the PCPT[24] (n=5,725 European) and SELECT[23] (n=25,366; 22,173 European; 1,763 African American/European; 1,173 African American; and 257 East Asian) and All of Us (AOU; n=17,512; 11,922 European; 2,469 African American; 1,783 other; 1,336 Hispanic/Latino)[25], which have been previously described. Briefly, PCPT and SELECT began as randomized, placebo-controlled, double-blind clinical trials of finasteride and selenium and vitamin E, respectively, and both enrolled men ≥55y. Individuals in SELECT and PCPT were required to have PSA≤3 ng/mL[23] and ≤4 ng/mL[24], respectively, at baseline. The National Institutes of Health (NIH) AOU cohort is committed to including groups that have been historically underrepresented in research[25]. From AOU, we selected individuals with PSA>0.01 between the ages of 40 and 90, with short-read whole-genome sequencing (WGS) data and no survey or EHR conditions/observations reflecting a history of PCa. The median PSA measurement we used was required to be ≤10 ng/mL. PRSs were calculated with the WGS data subset to variants with population-specific allele frequency ≥1% or a population-specific allele count >100 for any genetic ancestry. Genetic ancestry was determined using a random forest classifier trained on the principal component (PC) space of the Human Genome Diversity Project and 1000 Genomes Project (KGP)[58].

Ethical considerations

Informed consent was obtained from all study participants. AAPC was approved by their IRB. The ethics review board of the Program for the Protection of Human Subjects of Mount Sinai School of Medicine approved BioMe (#HSD09-00030, #07-0529 0001 02 ME). The University of Chicago Biological Sciences Division IRB Committee A (#IRB12-1660) approved COMPASS. Local and national IRBs approved MadCAP. MEC was approved by their IRB. The VA Central institutional review board (IRB) approved the MVP. The IRBs at Vanderbilt University and Meharry Medical College approved SCCS. Vanderbilt University Medical Center IRB approved BioVU. GERA was approved by the Kaiser Permanente Northern California IRB and the University of California, San Francisco. A local ethics committee approved the Malmo Diet and Cancer Study (MDCS). The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial was approved by the IRBs at each participating center and the National Cancer Institute, and the informed consent document allows data use for cancer and other adult disease investigations; we used publicly posted summary statistics, for which no IRB is required. The research was conducted with approved access to UKB data (#14105).

Genotype quality control and imputation

medRxiv preprint doi: https://doi.org/10.1101/2023.10.27.23297676; this version posted August 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Study subjects were genotyped using conventional GWAS arrays (**Table S1**). Genotypes were then imputed using imputation servers[59], Minimac3[60], or IMPUTE2[61]. The vast majority of studies imputed to the KGP phase 3 reference panel[62], with one sub-study imputing to KGP phase 1 just for the X chromosome[63] and another imputing to the TOPMed reference panel[59]. Since all but two studies (>95% of participants) used genome build 37, welifted over the assembly of those from build 38 using triple-liftOver[64] v133 (2022-05-20), an extension of LiftOver[65] that accounts for regions that are inverted between builds.

Standard genotype and individual-level QC procedures were implemented in each ancestry group in each participating study. Specific study protocols are delineated in **Table S1**, with additional QC steps and details in **Table S2**. Unless information was unavailable or a filter did not make sense for a particular group, variants were retained if their imputation quality score was ≥0.3, their MAF was ≥0.5% if the sample size was ≥1000 and ≥5% otherwise, their Hardy-Weinberg equilibrium (HWE) was ≥1e-8, they were mapped in build 37, and they had an MAF difference ≤0.2 compared to KGP populations (full details in **Table S3**). For the cohorts that meta-analyzed sub-cohorts (e.g., the three small African ancestry sub-cohorts within the SCCS African ancestry group; **Table S2**), we also required that variants be present in all sub-cohorts (necessary for multi-ancestry analysis method limitations, although this removed only a very small number of variants, **Table S3**). Finally, we excluded variants if they were present in only one study with n<2,000.

Association analyses

GWAS within each ancestry group in each study were undertaken using linear regression of log PSA on additive genotypes, and, when using multiple measurements, the long-term average residual by individual[66]. The minimum set of covariates included age at PSA measurement and genetic ancestry PCs. If available, GWAS also adjusted for batch/array, body mass index, and smoking status (**Table S1**). Meta-analyses of each ancestry group and across the overall discovery cohort were conducted using inverse-variance weighted fixed effects models using a custom patched version of METAL v2011-03-25 that prevents numerical precision loss (lines 633 and 635 of “Main.cpp” modified to the number 15 to output 15 digits precision)[67]. We also assessed heterogeneity with Cochran’s Q across the four ancestry groups.

To identify independently associated genome-wide significant (p≤5e-8) variants with computational efficiency, we first formed clumps of genome-wide significant variants such that all clumps were ≥10Mb apart and independent of one another; specifically, the top variant was chosen, genome-wide significant variants ≤10Mb from any variant in the clump were added to the clump, the process was iterated until a final clump was formed, and then the process was repeated to form more clumps (i.e., clumps were created such that there was no additional

medRxiv preprint doi: https://doi.org/10.1101/2023.10.27.23297676; this version posted August 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

genome-wide significant variant ≤10Mb). Within each clump, we used mJAM v2022-08-05[21], which uses population-specific LD reference panels for each contributing cohort and ancestry group to model the correlation among variants, with an r^2<0.01 threshold in all ancestry groups. Genotypes utilizing the appropriate GERA cohort group (European, Hispanic/Latino, African American, and East Asian) served as references[68].

To maximize discovery efforts, we combined our discovery cohort (n=296,754) with our replication cohort (n=95,768), for a total of 392,522 individuals.

Associations were considered novel if they had low LD from all previously-reported variants[20]. Specifically, we required r^2<0.01 in all four ancestry groups, again using GERA as LD reference.

Annotation

Variants were annotated using FUMA[69]. We first prioritized genes that included a significant prostate eQTL from GTeX v8 (www.gtexportal.org). We then prioritized other significant eQTLs and finally by closest gene. Deleteriousness of mutations was determined by CADD scores; a recommended cutoff to identify potentially pathogenic variants of scores ≥15 has been suggested (the median of splice site changes and non-synonymous variants from CADD v1.0; corresponds to the top 3.2% of variants)[70]. Gene functions were characterized with RefSeq[71]. Circos plots were generated using Circos v0.69-6[72].

Medication sensitivity analysis

Some of our study participants may have taken medications that can affect PSA levels. In particular, 5-alpha reductase inhibitors and testosterone can impact PSA levels[73,74]. We assessed the use of these medications among 26,669 men of European ancestry in the UKB with at least one PSA measurement. Men with a prescription for at least one of the two medications prior to PSA measurement were considered users. Ten percent of the men were prescribed 5-alpha reductase inhibitors and 0.56% testosterone. We also controlled for potential confounding by alpha blocker use.

Out-of-sample PRS variance explained

We calculated PRSs to assess the overall PSA variance explained by genetics, and to adjust PSA measurements for PSA genetics. All PRS results are shown only in independent cohorts (i.e., training dataset completely independent of testing dataset), such that assessments of performance are unbiased. Nonparametric bootstrap percentile CIs for variance explained were calculated using 1000 replicates.

We used two sets of individuals to construct the PRSs. First, we constructed PRSs from our discovery cohorts to allow assessment in GERA, PCPT, SELECT, and AoU. Second, we constructed PRSs from the meta-analysis of discovery and replication cohorts (which included GERA), with assessment in PCPT, SELECT, and AoU only. For GERA we included results

medRxiv preprint doi: https://doi.org/10.1101/2023.10.27.23297676; this version posted August 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

using first and multiple measurements; for PCPT and SELECT we include results using the first measurement.

We also used two sets of variants to calculate the PRSs in each of the two sets of individuals. We first utilized the independent genome-wide significant variants discovered in our analyses (one for discovery and one for the meta-analysis of discovery and replication). Second, we constructed a genome-wide score using PRS-CSx v2023-08-10[75], which was implemented utilizing GWAS summary statistics, the 1,287,078 HapMap3 variants as an LD reference that had an imputation quality >0.9 in SELECT, using a global shrinkage parameter of ϕ=0.0001 (which performed well in our previous work[20]), and variants with imputation quality ≥0.9. Since PRS-CSx only considers autosomes, independent genome-wide significant X chromosome variants were included (and produced a negligible increase in performance). The final scores were calculated by summing the effect size times the (probabilistic) number of alleles at each locus with PLINK v2.00a3.7LM[76].

We also assessed the variance explained by the discovery PRS-CSx within age intervals in GERA; we looked only in GERA to have an out-of-sample estimate from discovery and a large enough sample size at each age. An individual could be in multiple bins, but using just the first measurement of that individual per age bin.

Fur PSA for PCa screening in GERA

We adjusted PSA levels as has been described previously[20]. Briefly, PSA values for individual i were adjusted by $PSA_i^{adj} = PSA_i / a_i$, where a_i is a personalized adjustment factor derived from our PRS, as: $a_i = \exp(PRS_i) / \exp(\text{mean}(PRS))$. Here we estimated the mean(PRS) value within each group in the GERA cohort. We then evaluated the potential utility to alter biopsy referrals using age-specific PSA thresholds used within the Kaiser system (40-49y=2.5, 50-59y=3.5, 60-69=4.5, and 70-79=6.5 ng/ml[77]), evaluating net reclassification in cases and controls[20].

We also tested for associations of our PSA^{adj} with Gleason score (≤6, 7, and ≥8) using multinomial logistic regression with the R package nnet v7.3.18[78].

To assess whether there was variability in PSA adjustment across PSA levels, we first binned PSA values (with smaller ranges in lower values where there was more data). Within each bin, we computed PSA - $PSA_{adjusted}$, and then computed the median and IQR of these values. The median and IQR were then plotted at the center point of each bin by adding them to the identity line.

Impact of genetically-adjusted midlife PSA on overall and aggressive PCa prediction

We next investigated the impact of genetically adjusting PSA on the prediction of overall and aggressive PCa in the GERA cohort (3540 cases [1,028 aggressive, Gleason ≥7], 21,702

medRxiv preprint doi: https://doi.org/10.1101/2023.10.27.23297676; this version posted August 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

controls). We constructed a midlife PSA[26] based on each participant's median PSA between 50-60y, with cases restricted to measurements ≥1y before diagnosis. Genetic PSA adjustment was performed as in the previous section. Associations between PSA or genetically adjusted log PSA and PCa risk were assessed using logistic regression for overall PCa cases vs controls and for aggressive cases vs. controls, adjusting for covariates in Table S1. AUC was estimated using 10-fold repeated cross-validation (10 repeats) with caret v6.0.90[79].

Bias-corrected PCa estimates

PCa associations in individuals with European ancestry in the PRACTICAL consortium[27] were adjusted for screening bias[28], using estimates previously derived[20]: $\beta'_{Cancer} = \beta_{Cancer} - b\beta_{PSA}$, $SE'_{Cancer}=(SE_{Cancer}^2 + b^2SE_{PSA}^2 + SE_b^2\beta_{PSA} + SE_b^2SE_{PSA}^2)$, where SE is the standard error, and estimates were b=1.144, and SE_b=2.909e-4.

Associations with urinary symptom variants

We evaluated whether the novel PSA variants were associated with urinary symptoms in the GERA cohort. Individuals in this cohort completed the first 6 (of 7) questions from the American Urological Association Symptom Index (AUA-SI)[80] with 5-point Likert scale responses. The questions asked about incomplete emptying, frequency, intermittency, urgency, weak stream, and straining (**Table S13**). The one missing question from the AUA-SI regarded nocturia. We calculated total scores as the sum of the questions, giving each individual a value ranging from 6-30. The score was dichotomized at <7, ≥7 to differentiate men with little or no BPH (N=12,846) from those with moderate or severe BPH (N=15,480). We then assessed the association between the PSA variants and the urinary symptom score.

Prostate volume analysis

We evaluated associations between PSA variants and prostate volume in patients on active surveillance (AS) enrolled in the Canary Prostate Active Surveillance Study (PASS). Between 2008-2017, PASS prospectively enrolled 1455 patients with clinically localized PCa (cT1-cT2 and Gleason Grade 1-2) to undergo AS at one of 10 national sites[81]. Prostate volume was measured at diagnosis, with a median measurement of 43.0cc (IQR=31.0-57.5). The median age at diagnosis was 63 years (IQR=58-67), and 85% of the PASS cohort self-reported as European ancestry. Genotyping was conducted in 1,220 participants[82]. We assessed potential associations between the 407 PSA variants that we successfully imputed in PASS and prostate volume using mixed models with fixed effects for genetic variants, age at diagnosis, and 10 PCs, and a random effect for a genetic relationship matrix.

Associations with eGenes

medRxiv preprint doi: https://doi.org/10.1101/2023.10.27.23297676; this version posted August 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

We used single cell RNA sequencing (scRNA-seq) data to assess whether the eGenes for PSA-associated variants (Table S7) are expressed in secretory prostate cell types, particularly luminal epithelial cells, more so than other prostate cell types. For these analyses, we used data from the Chan Zuckerberg Cell by Gene census v2023-12-15[83].

Data availability

Summary statistics (from the discovery analysis and the final meta-analysis) will be made available in the NHGRI-EBI GWAS Catalog (https://www.ebi.ac.uk/gwas/downloads/summary-statistics). PRS weights (PRS_{318}, PRS_{447}, $PRS_{CSx,disc}$, $PRS_{CSx,joint}$) will be made available in the PGS catalog (https://www.pgscatalog.org/). To protect individuals' privacy, complete GERA data are available upon approved applications to the Kaiser Permanente Research Bank Portal (https://researchbank.kaiserpermanente.org/for-researchers). UKB data are available in the UKB cloud-based Research Analysis Platform (https://www.ukbiobank.ac.uk). GTEx data were obtained from the GTEx portal (www.gtexportal.org) and can be obtained from dbGaP Accession phs000424.v8.p2.

Code availability

Genome-wide association analysis were conducted using PLINK v2.0a3.7LM (http://www.cog-genomics.org/plink/2.0/). Meta-analysis were conducted with a custom-patched METAL v2011-03-25 (https://genome.sph.umich.edu/wiki/METAL_Documentation) that prevents numerical precision loss (lines 633 and 635 of "Main.cpp" modified to the number 15 to output 15 digits precision"), and with MJAM v2022-08-05 (https://github.com/USCbiostats/hJAM/R). Imputation was done via imputation servers (https://imputationserver.sph.umich.edu, https://imputation.biodatacatalyst.nhlbi.nih.gov), Minimac3 (https://genome.sph.umich.edu/wiki/Minimac3), and IMPUTE2 (https://mathgen.stats.ox.ac.uk/impute/impute_v2.html). Analysis were also conducted in R, including v4.2.0 (https://cran.r-project.org/). FUMA was used for annotation (https://fuma.ctglab.nl). Circos plots were generated using Circos v0.69-6 (https://fuma.ctglab.nl). The genome-wide PRS was conducted with PRS-CSx v2023-08-10 (https://github.com/getian107/PRScsx).

Acknowledgements

We thank the participants who generously agreed to participate in each cohort. This research is based in part on data from the Million Veteran Program (MVP), Office of Research and Development, Veterans Health Administration, and was supported by the MVP017 Exemplar Cancer Project. This research was conducted using the UKB resource under application number 14105. In addition, the All of Us Research Program would not be possible without the partnership of its participants.

Funding

The Precision PSA study is supported by funding from the National Institutes of Health (NIH) National Cancer Institute (NCI) under award number R01CA241410 (PI: JSW) and U01CA261339 (MPI: JSW). LK is supported by funding from the National Cancer Institute

medRxiv preprint doi: https://doi.org/10.1101/2023.10.27.23297676; this version posted August 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

(R00CA246076). REG is supported by a Young Investigator Award from the Prostate Cancer Foundation. HL is supported in part by NIH/NCI by a Cancer Center Support Grant to Memorial Sloan Kettering Cancer Center (P30 CA008748, PI: Vickers, S), U01-CA266535 (PI: Carlsson, S), R01-CA244948 (PI: RJK), and Swedish Cancer Society (Cancerfonden 20 1354 PJF; PI: HL). This work was supported by research grants from the NIH National Institute of General Medical Sciences (NIGMS) under award number R01GM130791 (PI: JDM); the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai; the Office of Research Infrastructure of the National Institutes of Health under award number S10OD026880 and NIH/NCI funding (R01CA175491, R01CA244948; PI: RJK); the National Cancer Institute of the National Institutes of Health (UM1CA182883, PI: CM Tangen/IM Thompson; U10CA37429, PI: CD Blanke). MADCaP was supported by U01CA184374 (PI: TR). COMPASS was supported by P30CA014599. Support for GERA participant enrollment, survey completion, and biospecimen collection for 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 benefit programs. GERA genotyping was funded by National Institute on Aging and NIH Common Fund (grant RC2 AG-036607 to Cathy Schaeffer and Neil Risch). The All of Us Research Program is supported by the National Institutes of Health, Office of the Director: Regional Medical Centers: 1 OT2 OD026549; 1 OT2 OD026554; 1 OT2 OD026557; 1 OT2 OD026556; 1 OT2 OD026550; 1 OT2 OD 026552; 1 OT2 OD026553; 1 OT2 OD026548; 1 OT2 OD026551; 1 OT2 OD026555; IAA #: AOD 16037; Federally Qualified Health Centers: HHSN 263201600085U; Data and Research Center: 5 U2C OD023196; Biobank: 1 U24 OD023121; The Participant Center: U24 OD023176; Participant Technology Systems Center: 1 U24 OD023163; Communications and Engagement: 3 OT2 OD023205; 3 OT2 OD023206; and Community Partners: 1 OT2 OD025277; 3 OT2 OD025315; 1 OT2 OD025337; 1 OT2 OD025276.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders had no role in study design, data collection and analysis, the decision to publish, or preparation of the manuscript.

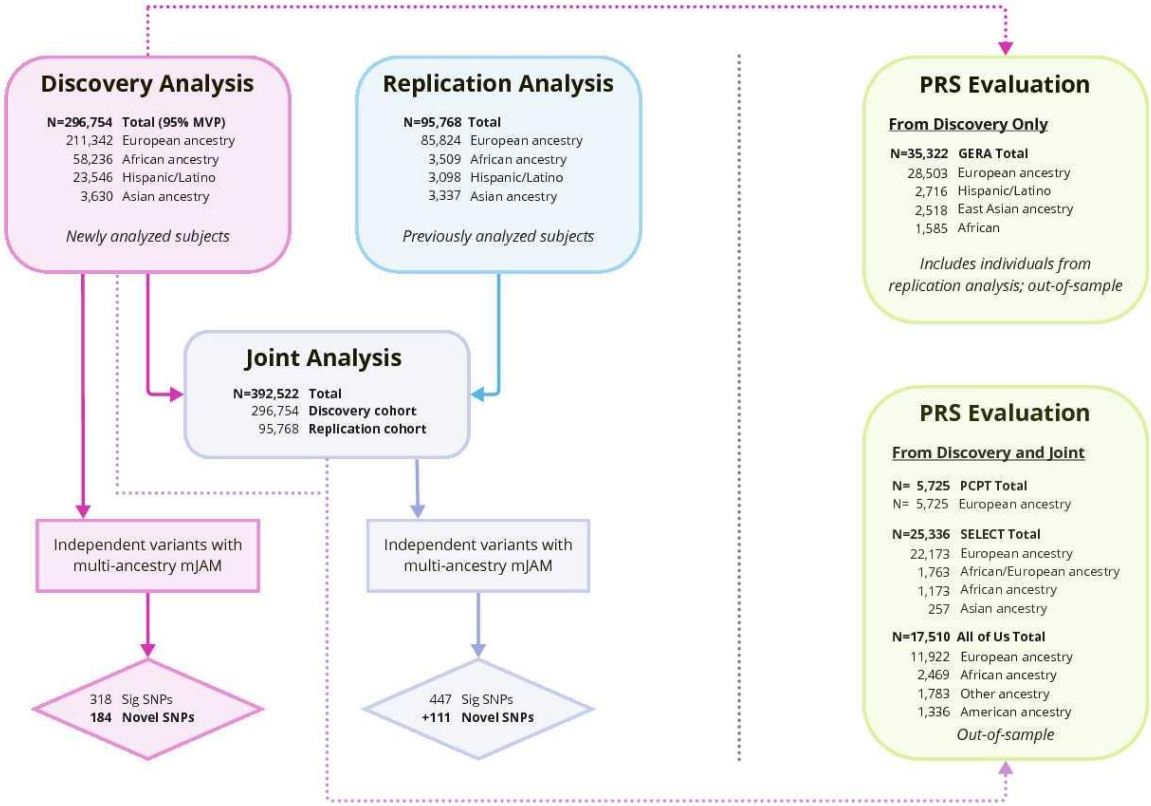
Conflict of Interest

JSW and CLC are non-employee co-founders of Avail Bio. HL is named on a patent for intact PSA assays and a patent for a statistical method to detect prostate cancer that is licensed to and commercialized by OPKO Health. HL receives royalties from sales of the test and has stock in OPKO Health. REG consults for Hunton Andrews Kurth LLP on subject matter unrelated to this study.

medRxiv preprint doi: https://doi.org/10.1101/2023.10.27.23297676; this version posted August 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

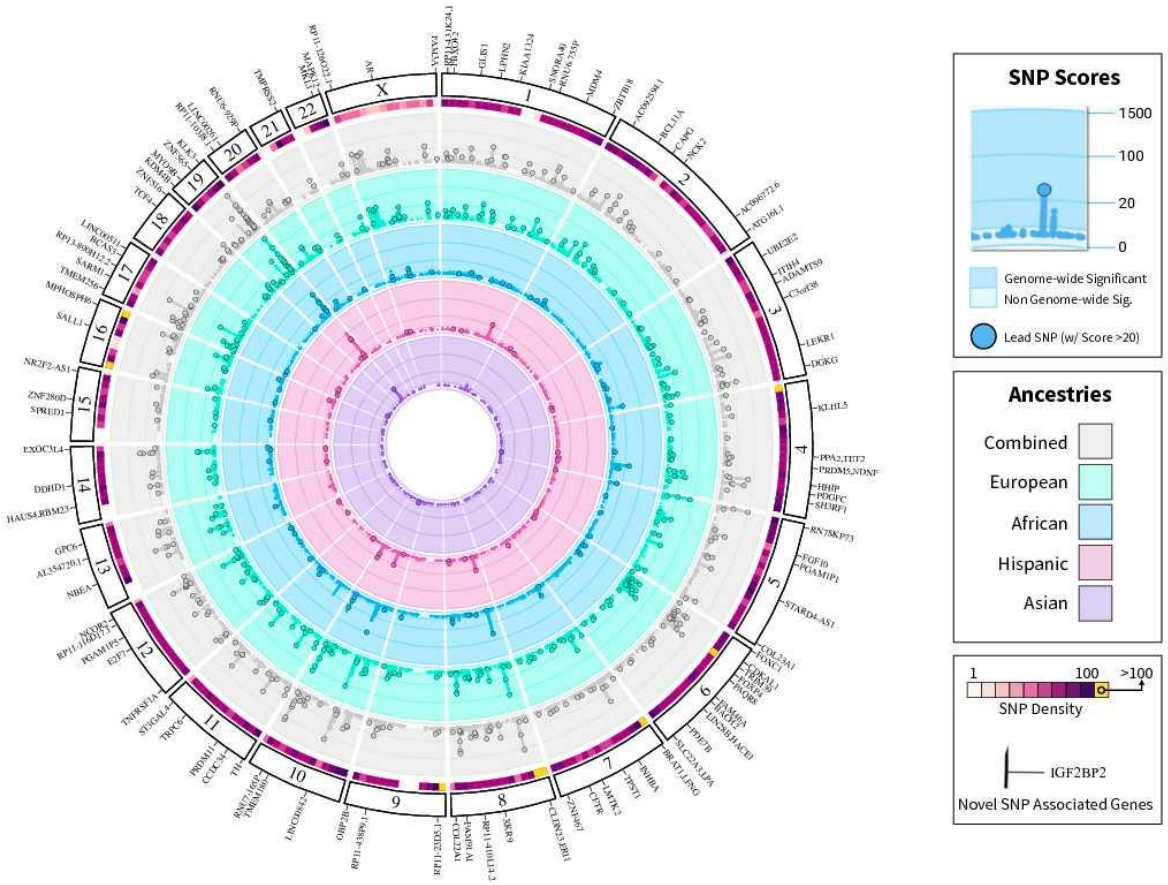
Figures and Tables

Figure 1. Flowchart describing the Precision PSA project analysis workflow and ancestry compositions of the discovery, replication, and joint meta-analysis cohorts. The discovery GWAS analysis revealed 318 genome-wide significant (p<5e-8) SNPs associated with PSA levels, of which 184 were novel. The joint analysis (consisting of the discovery and replication cohorts) revealed 447 genome-wide significant SNPs associated with PSA levels, of which an additional 111 were novel. Both discovery and joint GWAS results were used to develop PRSs for PSA, which were then evaluated in GERA (when out-of-sample), PCPT, and SELECT.



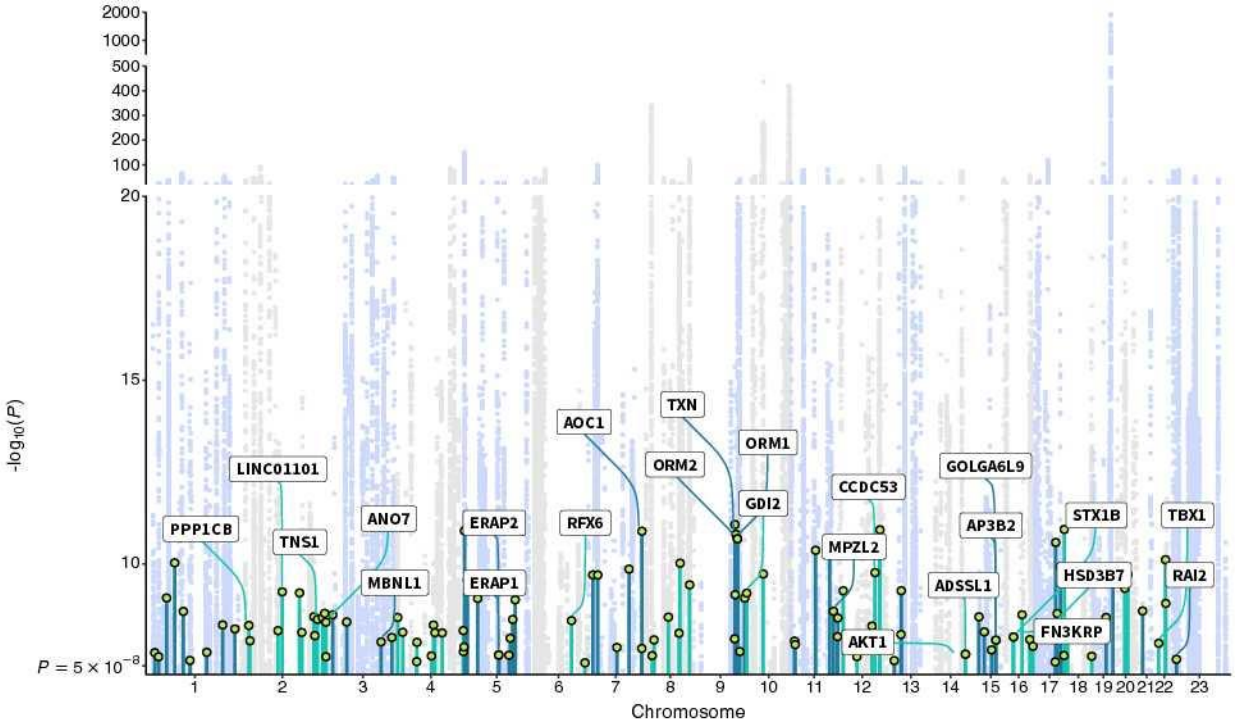
medRxiv preprint doi: https://doi.org/10.1101/2023.10.27.23297676; this version posted August 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Figure 2. Circos plot showing PSA GWAS significant variants by chromosome from the discovery cohort. Concentric tracks are colored based on results from individual ancestries, with gray indicating results from the overall discovery meta-analysis. The top 100,000 GWAS SNPs (with the smallest p-values) per ancestry are shown as points; larger circled points indicate the 318 genome-wide significant variants (p<5e-8; 184 of which were novel) from the overall discovery analysis in all ancestries. SNP density in 10Mbp bins from the overall analysis is shown as a heatmap above the overall track. The outermost ring displays genes associated with novel discovery PSA SNPs.



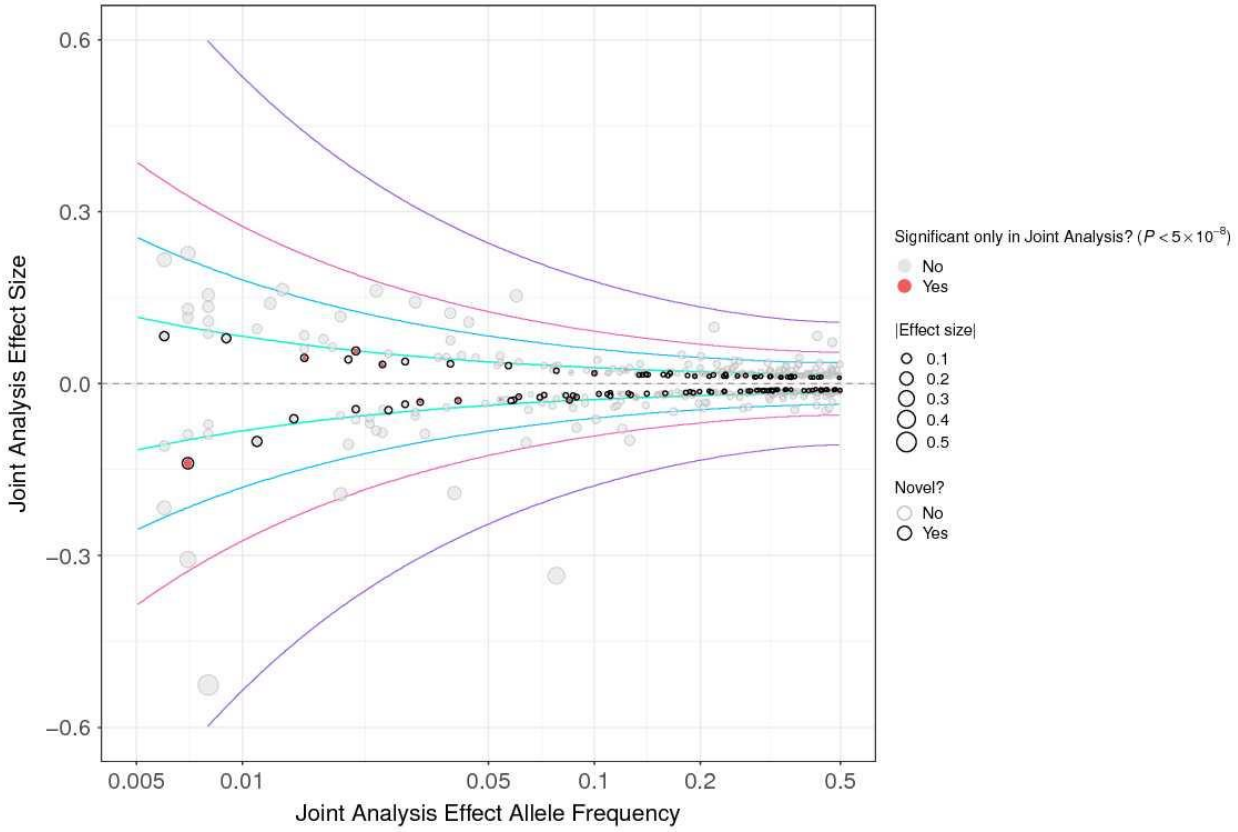
medRxiv preprint doi: https://doi.org/10.1101/2023.10.27.23297676; this version posted August 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Figure 3. Manhattan plot showing results from the joint multi-ancestry meta-analysis of the discovery (n=296,754) and replication (n=95,768) studies. Only genome-wide significant associations (p<5e-8) are plotted. The joint analysis detected 447 independent genome-wide significant PSA-associated SNPs. These included 111 novel variants that were conditionally independent from previous findings and the discovery only analyses in each study alone (indicated by the circles). Gene labels are given for variants with CADD>15 and/or variants that are prostate tissue eQTLs.



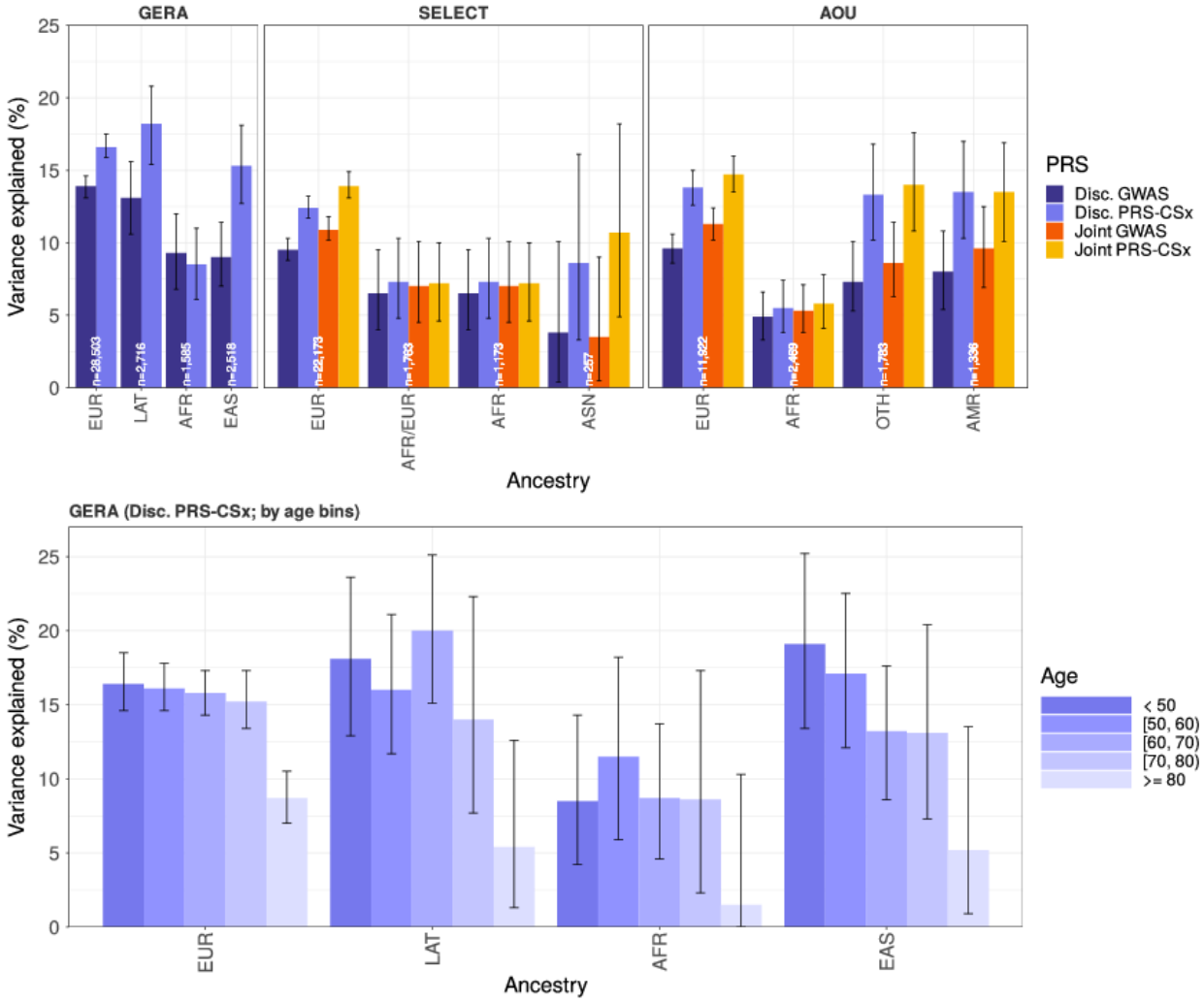
medRxiv preprint doi: https://doi.org/10.1101/2023.10.27.23297676; this version posted August 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Figure 4. Plot of the relationship between minor allele frequency and estimated effect sizes for PSA GWAS significant variants. Each point represents one of the 447 independent genome-wide significant SNPs identified in our mJAM multi-ancestry GWAS joint meta-analysis. The SNP effect sizes are expressed in ln(PSA) per minor allele. The curves indicate the hypothetical detectable SNP effect sizes for a given minor allele frequency, assuming statistical power of 80%, α=5e-8 (genome-wide significant), and the sample size of each of our populations here (297,166 European ancestry, N=61,745 African ancestry, 6,967 Asian ancestry, 26,644 Hispanic).



medRxiv preprint doi: https://doi.org/10.1101/2023.10.27.23297676; this version posted August 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

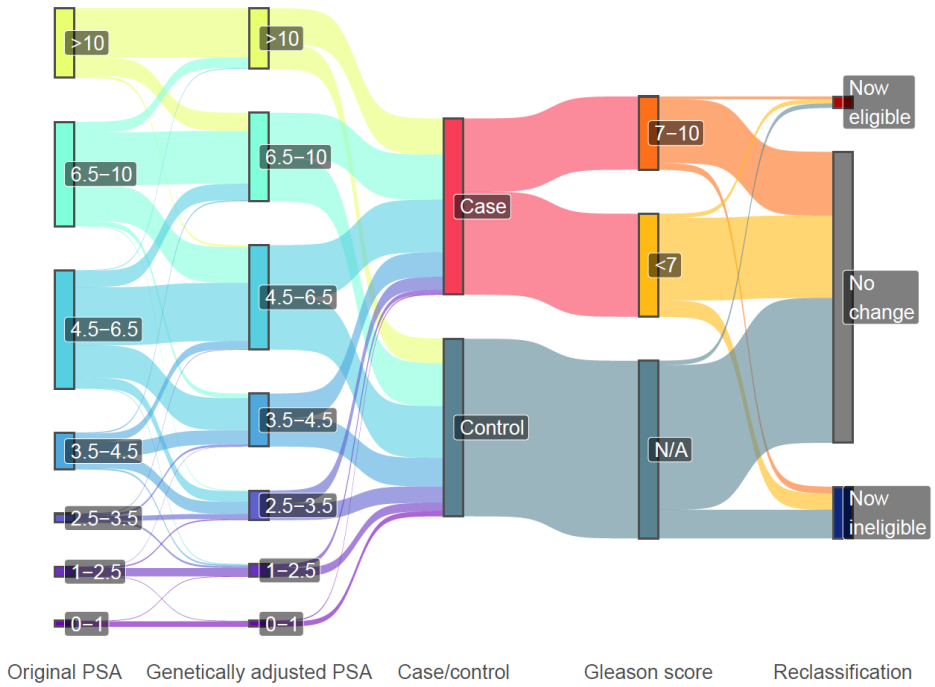
Figure 5. Variance in PSA levels explained by PRSs. We constructed PRSs for PSA from our discovery cohorts to allow assessment in GERA and our PRS validation cohorts (PCPT and SELECT). We also constructed PRSs from the joint meta-analysis of discovery and replication cohorts, with assessment in our validation cohorts. The PRSs were based on the multi-ancestry identified conditionally independent genome-wide significant variants using mJAM and on a multi-ancestry genome-wide score using PRS-CSx. The genome-wide score generally performed better than the genome-wide significant score. The variance explained by genome-wide PRSs **(A)** was up to 16.9% in Europeans, 18.6% in Hispanics/Latinos, 9.5% in Africans, and 15.3% in East Asians, and **(B)** decreased as age increased.



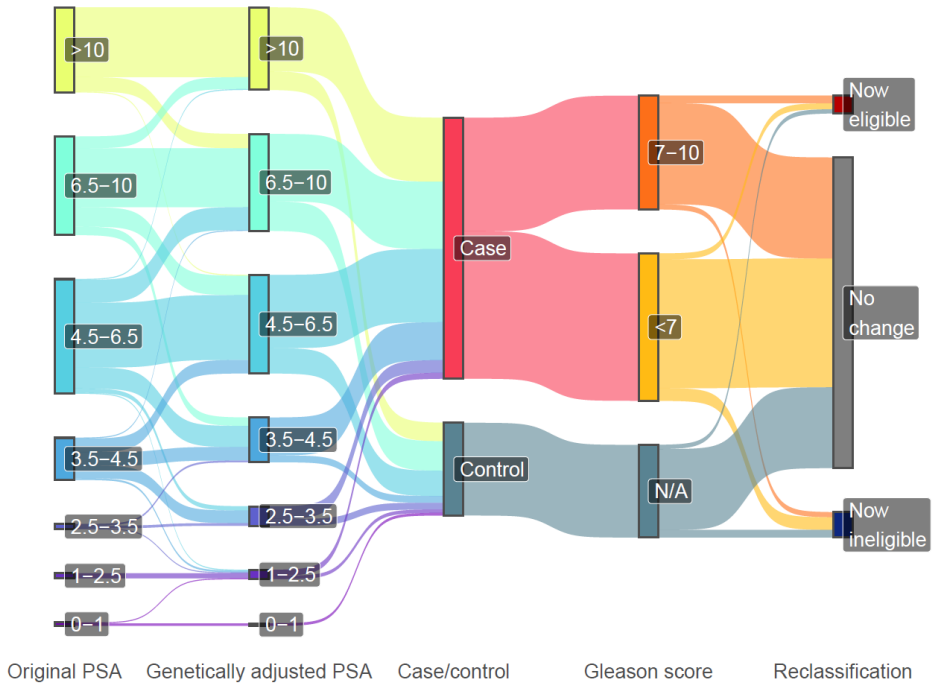
medRxiv preprint doi: https://doi.org/10.1101/2023.10.27.23297676; this version posted August 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Figure 6. Biopsy reclassification with genetically-adjusted PSA. PSA levels were adjusted using the PRS-CSx estimate from the out-of-sample discovery cohort, assessed in GERA using age-specific cutoffs in **(A)** Europeans and **(B)** African Americans (see **Methods**). GERA Latinos and East Asians are shown in **Figure S3**.

A



B



medRxiv preprint doi: https://doi.org/10.1101/2023.10.27.23297676; this version posted August 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

References

1. Lilja, H. A kallikrein-like serine protease in prostatic fluid cleaves the predominant seminal vesicle protein. *J Clin Invest* **76**, 1899–1903 (1985).
2. Lilja, H., Ulmert, D. & Vickers, A. J. Prostate-specific antigen and prostate cancer: prediction, detection and monitoring. *Nat Rev Cancer* **8**, 268–278 (2008).
3. Balk, S. P., Ko, Y.-J. & Bubley, G. J. Biology of prostate-specific antigen. *J Clin Oncol* **21**, 383–391 (2003).
4. Lilja, H., Ulmert, D. & Vickers, A. J. Prostate-specific antigen and prostate cancer: prediction, detection and monitoring. *Nat Rev Cancer* **8**, 268–278 (2008).
5. Pinsky, P. F. *et al.* Prostate volume and prostate-specific antigen levels in men enrolled in a large screening trial. *Urology* **68**, 352–356 (2006).
6. Lee, S. E. *et al.* Relationship of prostate-specific antigen and prostate volume in Korean men with biopsy-proven benign prostatic hyperplasia. *Urology* **71**, 395–398 (2008).
7. Grubb, R. L. *et al.* Serum prostate-specific antigen hemodilution among obese men undergoing screening in the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial. *Cancer Epidemiol Biomarkers Prev* **18**, 748–751 (2009).
8. Harrison, S. *et al.* Systematic review and meta-analysis of the associations between body mass index, prostate cancer, advanced prostate cancer, and prostate-specific antigen. *Cancer Causes Control* **31**, 431–449 (2020).
9. Jemal, A. *et al.* Prostate Cancer Incidence and PSA Testing Patterns in Relation to USPSTF Screening Recommendations. *JAMA* **314**, 2054–2061 (2015).
10. Gulati, R., Inoue, L. Y. T., Gore, J. L., Katcher, J. & Etzioni, R. Individualized estimates of overdiagnosis in screen-detected prostate cancer. *J Natl Cancer Inst* **106**, djt367 (2014).
11. Sammon, J. D. *et al.* Prostate-Specific Antigen Screening After 2012 US Preventive Services Task Force Recommendations. *JAMA* **314**, 2077–2079 (2015).
12. Hugosson, J. *et al.* A 16-yr Follow-up of the European Randomized study of Screening for Prostate Cancer. *Eur Urol* **76**, 43–51 (2019).
13. Sandhu, G. S. & Andriole, G. L. Overdiagnosis of Prostate Cancer. *J Natl Cancer Inst Monogr* **2012**, 146–151 (2012).
14. Frånlund, M. *et al.* Results from 22 years of Followup in the Göteborg Randomized Population-Based Prostate Cancer Screening Trial. *J Urol* **208**, 292–300 (2022).
15. US Preventive Services Task Force. Screening for Prostate Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA* **319**, 1901–1913 (2018).
16. Bell, N. *et al.* Recommendations on screening for prostate cancer with the prostate-specific antigen test. *CMAJ* **186**, 1225–1234 (2014).
17. Tikkinen, K. A. O. *et al.* Prostate cancer screening with prostate-specific antigen (PSA) test: a clinical practice guideline. *BMJ* **362**, k3581 (2018).
18. Bansal, A. *et al.* Heritability of prostate-specific antigen and relationship with zonal prostate volumes in aging twins. *J Clin Endocrinol Metab* **85**, 1272–1276 (2000).
19. Pilia, G. *et al.* Heritability of Cardiovascular and Personality Traits in 6,148 Sardinians. *PLoS Genet* **2**, e132 (2006).
20. Kachuri, L. *et al.* Genetically adjusted PSA levels for prostate cancer screening. *Nat Med* **29**, 1412–1423 (2023).
21. Shen, J. *et al.* Fine-mapping and credible set construction using a multi-population joint analysis of marginal summary statistics. doi:https://doi.org/10.1101/2022.12.22.521659.
22. Kerem, B. S. *et al.* DNA marker haplotype association with pancreatic sufficiency in cystic fibrosis. *Am J Hum Genet* **44**, 827–834 (1989).
23. Lippman, S. M. *et al.* Effect of selenium and vitamin E on risk of prostate cancer and other cancers: the Selenium and Vitamin E Cancer Prevention Trial (SELECT). *JAMA* **301**, 39–51

medRxiv preprint doi: https://doi.org/10.1101/2023.10.27.23297676; this version posted August 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

(2009).

24. Thompson, I. M. *et al.* Assessing prostate cancer risk: results from the Prostate Cancer Prevention Trial. *J Natl Cancer Inst* **98**, 529–534 (2006).
25. Mayo, K. R. *et al.* The All of Us Data and Research Center: Creating a Secure, Scalable, and Sustainable Ecosystem for Biomedical Research. *Annu Rev Biomed Data Sci* **6**, 443–464 (2023).
26. Preston, M. A. *et al.* Baseline Prostate-Specific Antigen Levels in Midlife Predict Lethal Prostate Cancer. *J Clin Oncol* **34**, 2705–2711 (2016).
27. Conti, D. V. *et al.* Trans-ancestry genome-wide association meta-analysis of prostate cancer identifies new susceptibility loci and informs genetic risk prediction. *Nat Genet* **53**, 65–75 (2021).
28. Dudbridge, F. *et al.* Adjustment for index event bias in genome-wide association studies of subsequent events. *Nat Commun* **10**, 1561 (2019).
29. Jiang, L., Zheng, Z., Fang, H. & Yang, J. A generalized linear mixed model association tool for biobank-scale data. *Nat Genet* **53**, 1616–1621 (2021).
30. Eldjarn, G. H. *et al.* Large-scale plasma proteomics comparisons through genetics and disease associations. *Nature* **622**, 348–358 (2023).
31. Kim, J. & Coetzee, G. A. Prostate specific antigen gene regulation by androgen receptor. *J Cell Biochem* **93**, 233–241 (2004).
32. Heinlein, C. A. & Chang, C. Androgen Receptor in Prostate Cancer. *Endocrine Reviews* **25**, 276–308 (2004).
33. Agúndez, J. A. G. *et al.* The diamine oxidase gene is associated with hypersensitivity response to non-steroidal anti-inflammatory drugs. *PLoS One* **7**, e47571 (2012).
34. Amo, G. *et al.* FCERI and Histamine Metabolism Gene Variability in Selective Responders to NSAIDS. *Front Pharmacol* **7**, 353 (2016).
35. Miller, C. L. *et al.* Integrative functional genomics identifies regulatory mechanisms at coronary artery disease loci. *Nat Commun* **7**, 12092 (2016).
36. Linseman, T. *et al.* Functional Validation of a Common Nonsynonymous Coding Variant in ZC3HC1 Associated With Protection From Coronary Artery Disease. *Circ Cardiovasc Genet* **10**, e001498 (2017).
37. Cuppens, H. & Cassiman, J.-J. CFTR mutations and polymorphisms in male infertility. *Int J Androl* **27**, 251–256 (2004).
38. Landrum, M. J. *et al.* ClinVar: improvements to accessing data. *Nucleic Acids Research* **48**, D835–D844 (2020).
39. Kawa, S. *et al.* Azoospermia in mice with targeted disruption of the Brek/Lmtk2 (brain-enriched kinase/lemur tyrosine kinase 2) gene. *Proc Natl Acad Sci U S A* **103**, 19344–19349 (2006).
40. Cruz, D. F., Farinha, C. M. & Swiatecka-Urban, A. Unraveling the Function of Lemur Tyrosine Kinase 2 Network. *Frontiers in Pharmacology* **10**, (2019).
41. Bostwick, D. G. *et al.* The association of benign prostatic hyperplasia and cancer of the prostate. *Cancer* **70**, 291–301 (1992).
42. Ørsted, D. D. & Bojesen, S. E. The link between benign prostatic hyperplasia and prostate cancer. *Nat Rev Urol* **10**, 49–54 (2013).
43. Thompson, I. M. *et al.* Prevalence of prostate cancer among men with a prostate-specific antigen level < or =4.0 ng per milliliter. *N Engl J Med* **350**, 2239–2246 (2004).
44. Coric, J., Mujic, J., Kucukalic, E. & Ler, D. Prostate-Specific Antigen (PSA) and Prostate Volume: Better Predictor of Prostate Cancer for Bosnian and Herzegovina Men. *Open Biochem J* **9**, 34–36 (2015).
45. D'Amico, A. V. Risk-based management of prostate cancer. *N Engl J Med* **365**, 169–171 (2011).

medRxiv preprint doi: https://doi.org/10.1101/2023.10.27.23297676; this version posted August 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

46. R Core Team. R: A language and environment for statistical computing. (2012).
47. McGrath, S., Zhao, X., Qin, Z. Z., Steele, R. & Benedetti, A. One-sample aggregate data meta-analysis of medians. *Statistics in Medicine* **38**, 969–984 (2019).
48. Tayo, B. O. *et al.* Genetic background of patients from a university medical center in Manhattan: implications for personalized medicine. *PLoS One* **6**, e19166 (2011).
49. Aschebrook-Kilfoy, B. *et al.* Cohort profile: the ChicagO Multiethnic Prevention and Surveillance Study (COMPASS). *BMJ Open* **10**, e038481 (2020).
50. Press, D. J. *et al.* Tobacco and marijuana use and their association with serum prostate-specific antigen levels among African American men in Chicago. *Prev Med Rep* **20**, 101174 (2020).
51. Andrews, C. *et al.* Development, Evaluation, and Implementation of a Pan-African Cancer Research Network: Men of African Descent and Carcinoma of the Prostate. *J Glob Oncol* **4**, 1–14 (2018).
52. Harlemon, M. *et al.* A Custom Genotyping Array Reveals Population-Level Heterogeneity for the Genetic Risks of Prostate Cancer and Other Cancers in Africa. *Cancer Res* **80**, 2956–2966 (2020).
53. Kolonel, L. N., Altshuler, D. & Henderson, B. E. The multiethnic cohort study: exploring genes, lifestyle and cancer risk. *Nat Rev Cancer* **4**, 519–527 (2004).
54. Kolonel, L. N. *et al.* A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. *Am. J. Epidemiol.* **151**, 346–357 (2000).
55. Gaziano, J. M. *et al.* Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol* **70**, 214–223 (2016).
56. Fang, H. *et al.* Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity in Genome-wide Association Studies. *The American Journal of Human Genetics* **105**, 763–772 (2019).
57. Signorello, L. B. *et al.* Southern community cohort study: establishing a cohort to investigate health disparities. *J Natl Med Assoc* **97**, 972–979 (2005).
58. Venner, E. *et al.* The frequency of pathogenic variation in the All of Us cohort reveals ancestry-driven disparitie. Preprint at https://doi.org/10.1101/2022.12.19.22283658 (2022).
59. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program | Nature. https://www.nature.com/articles/s41586-021-03205-y.
60. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat Genet* **48**, 1284–1287 (2016).
61. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* **44**, 955–959 (2012).
62. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
63. Consortium, T. 1000 G. P. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
64. Sheng, X. *et al.* Inverted genomic regions between reference genome builds in humans impact imputation accuracy and decrease the power of association testing. *HGG Adv* **4**, 100159 (2023).
65. Karolchik, D. *et al.* The UCSC Genome Browser database: 2014 update. *Nucl. Acids Res.* **42**, D764–D770 (2014).
66. Ganesh, S. K. *et al.* Effects of Long-Term Averaging of Quantitative Blood Pressure Traits on the Detection of Genetic Associations. *The American Journal of Human Genetics* **95**, 49–65 (2014).
67. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
68. Hoffmann, T. J. *et al.* Genome-wide association study of prostate-specific antigen levels

medRxiv preprint doi: https://doi.org/10.1101/2023.10.27.23297676; this version posted August 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

identifies novel loci independent of prostate cancer. *Nat Commun* **8**, 14248 (2017).

69. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* **8**, 1826 (2017).
70. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310–315 (2014).
71. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733-745 (2016).
72. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639–1645 (2009).
73. Gerstenbluth, R. E., Maniam, P. N., Corty, E. W. & Seftel, A. D. Prostate-specific antigen changes in hypogonadal men treated with testosterone replacement. *J Androl* **23**, 922–926 (2002).
74. Guess, H. A., Heyse, J. F. & Gormley, G. J. The effect of finasteride on prostate-specific antigen in men with benign prostatic hyperplasia. *Prostate* **22**, 31–37 (1993).
75. Ruan, Y. *et al.* Improving polygenic prediction in ancestrally diverse populations. *Nat Genet* **54**, 573–580 (2022).
76. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, (2015).
77. Oesterling, J. E. *et al.* Serum prostate-specific antigen in a community-based population of healthy men. Establishment of age-specific reference ranges. *JAMA* **270**, 860–864 (1993).
78. W. N. Ripley & B. D. Ripley. *Modern Applied Statistics with S*. (Springer, New York, NY, USA, 2002).
79. Kuhn, M. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* **28**, 1–26 (2008).
80. Barry, M. J. *et al.* The American Urological Association Symptom Index for Benign Prostatic Hyperplasia. *J Urol* **197**, S189–S197 (2017).
81. Newcomb, L. F. *et al.* Canary Prostate Active Surveillance Study: design of a multi-institutional active surveillance cohort and biorepository. *Urology* **75**, 407–413 (2010).
82. Jiang, Y. *et al.* Genetic Factors Associated with Prostate Cancer Conversion from Active Surveillance to Treatment. *HGG Adv* **3**, 100070 (2022).
83. Program, C. S.-C. B. *et al.* CZ CELL×GENE Discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. 2023.10.30.563174 Preprint at https://doi.org/10.1101/2023.10.30.563174 (2023).