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], Clint 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], Timothy Rebbeck[15], Joseph Lachance[16], Caroline Andrews[15], Akindele O Adebiyi[17], Ben Adusei[18], Oseremen I Aisuodionoe-Shadrach[19], Pedro W Fernandez[20], Mohamed Jalloh[21], Rohini Janivara[16], Wenlong C Chen[22], James E Mensah[23], Ilir Agalliu[24], Sonja I Berndt[25], John P Shelley[26], Kerry Schaffer[27], Mitchell J Machiela[25], Neal D Freedman[25], Wen-Yi Huang[25], Shengchao A Li[25], Phyllis J Goodman[28], Cathee Till[28], Ian Thompson[29], Hans Lilja[30,31], Stephen K Van Den Eeden[32], Stephen J Chanock[25], Jonathan D Mosley[33,26], David V Conti[6,7], Christopher A Haiman[6,7], Amy C Justice[34,35], Linda Kachuri[4,36,**], John S Witte[4,36,5,37,**]

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. Dana Farber Cancer Institute, Harvard School of Public Health, Boston, MA, USA
16. School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, USA
17. Department of Community Medicine, College of Medicine, University of Ibadan, Ibadan, Nigeria
18. 37 Military Hospital, Accra, Ghana
19. College of Health Sciences, University of Abuja, Abuja, Nigeria; Cancer Science CentreAbuja, Abuja, Nigeria; University of Abuja Teaching Hospital, Abuja, Nigeria
20. Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa
21. Hospital General Idrissa Pouye, Dakar, Senegal; Ecole Doctorale, Univesite Iba Der Thiam de Thies
22. 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
23. Korle-Bu Teaching Hospital and University of Ghana Medical School, Accra, Ghana
24. Dept of Epidemiology and Population Health, and Dept of Urology, Albert Einstein College of Medicine, New York, NY, USA
25. Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA
26. Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA
27. Division of Hematology and Oncology, Vanderbilt University Medical Center, Nashville, TN, USA
28. SWOG Statistics and Data Management Center, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
29. CHRISTUS Santa Rosa Medical Center Hospital, San Antonio, TX, USA
30. Departments of Pathology and Laboratory Medicine, Surgery, Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA
31. Department of Translational Medicine, Lund University, Malmö, Sweden
32. Division of Research, Kaiser Permanente Northern California, Oakland, CA, USA
33. Department of Internal Medicine, Vanderbilt University Medical Center, Nashville, TN, USA
34. Veterans Administration Connecticut Healthcare System, West Haven, Connecticut, USA
35. Department of Internal Medicine and Yale University School of Public Health, Yale School of Medicine, New Haven, Connecticut, USA
36. Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA, USA
37. 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)

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 new polygenic risk score reached 16.9% (95% CI=16.1%-17.8%) in European ancestry, 9.5% (95% CI=7.0%-12.2%) in African ancestry, 18.6% (95% CI=15.8%-21.4%) in Hispanic/Latino, and 15.3% (95% CI=12.7%-18.1%) in Asian ancestry, and lower for higher age. Our study highlights how including proportionally more participants from underrepresented populations improves genetic prediction of PSA levels, with potential to personalize prostate cancer screening.

Introduction

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

Twin studies estimate PSA heritability at 40-45%[18,19], and genome-wide heritability has been estimated to be between 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 of an individual is inflated or deflated based on the genetic variants an individual has) 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. Additional genome-wide polygenic risk scores (PRSs) explained up to 10% in European ancestry; however, the PRSs were substantially less predictive in other groups, especially men of 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 prostate cancer screening.

Results

Composition of discovery and replication cohorts

Our discovery population consisted of 296,754 men without prostate cancer 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 of these men had been included in previous genome-wide association studies (GWAS) of PSA levels. We present genotype platform details in **Table S1**, demographics in **Table S2**, and quality control metrics in **Table S3**. The Million Veteran Program (MVP) made up 96.5% of the discovery cohort. For replication, we utilized results from 95,768 independent individuals who were described in our previous work[20], including 85,824 European ancestry, 3,509 African ancestry, 3,098 Hispanic/Latinos, 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

In our discovery cohorts, we identified 318 independent genome-wide significant variants in a multi-ancestry analysis of log-transformed PSA levels (Circos plot, **Figure 2**; overall and ancestry-specific Manhattan plots, **Figure S1**; numerical results, **Table S4**; ancestry-specific lead variants **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 (as defined in **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).

Of the 184 variants that were novel in the multi-ancestry analysis (**Figure 2, Figure S1, Table S4**), 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 Asians or Hispanics/Latinos (likely due to low sample size; overlap given in **Figure S2**). Of the eight in the African ancestry population, only two variants were frequent enough (see **Methods**) to be assessed in other ancestry groups: one with European ancestry minor allele frequency (MAF) 23.7% (rs2071041, *ITIH4*) that was also genome-wide significant in European ancestry individuals and the other (rs1203888, *LINC00261*) 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/Latinos (p=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 p=3.4e-10, replication 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 East Asian individuals. Three variants in genes that encode PSA (rs76151346 and rs145428838, *KLK3*; rs182464120, *KLK2*) exclusively imputed in African ancestry individuals (all MAF<5%, two <1%) did not exhibit strong evidence of replication (p>0.05). The remaining two variants identified in African ancestry (rs7125654, and rs4542679), were more common (MAF>5%) but also did not exhibit evidence of replication (p>0.05). For these, rs7125654 (*TRPC6*) was less common in Latinos, but more common in Asian ancestry and rs4542679 (*RP11-345M22.3*) was also less common in Latinos and not adequately polymorphic in East Asians.

We next tested for heterogeneity (i.e., 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 that was significant at a Bonferroni level (p<0.05/184=0.00027). The variant had a strong discovery effect in European ancestry individuals (β=0.0327, p=1.2e-15, MAF=0.10), but was not significant in other groups (African ancestry β=0.0131, p=0.42, MAF=0.021; Asian β=-0.176, p=0.027, MAF=0.021; Hispanic/Latino β=-0.0102, p=0.37, MAF=0.120). In our replication cohort, the variant nominally (i.e., p<0.05) replicated (p=0.0065, European ancestry p=0.003) and showed no statistically significant evidence of heterogeneity across ancestry groups (I^2=0.0, p=0.44), although our sample sizes to detect heterogeneity 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%) additional were eQTLs in other tissues (**Table S4**). Five novel variants were missense and predicted to be deleterious, with >20 Combined Annotation Dependent Depletion (CADD) scores (**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 based on 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, and 32 additional (25.0%) were in the same direction, and the remaining 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 both the discovery and replication cohorts (Manhattan plot of p<5e-8, **Figure 3**; additional Manhattan plots, **Figure S1**; numerical results, **Table S7**), we identified 447 independent variants. Among the 111 variants that were further novel in this analysis, none showed evidence of heterogeneity (p>0.05/111=0.00045). A total of 56 (50.4%) were genome-wide significant in European ancestry individuals, but none of the novel variants were genome-wide significant in a non-European ancestry group (**Table S8**). The allele frequencies and effect sizes of the newly discovered variants largely followed those expected by power curves (**Figure 4**).

In the joint meta-analysis, 12 (10.8%) of the novel variants were prostate tissue eQTLs, and 50 (45.0%) additional were eQTLs for other tissues. Two of the novel variants 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.

Out-of-sample PSA variance explained by PRS

First, we evaluated different strategies for constructing polygenic risk scores (PRSs) for PSA levels first using results from our discovery cohort (see **Methods**). Here, four cohorts of men without prostate cancer were out-of-sample: the 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 the All of Us (AOU)[25] cohorts.

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 the 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 American 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 trained using the discovery GWAS ($PRS_{CSx\text{-}disc}$; includes more than genome-wide significant variants; comprising 1,070,230 SNPs; see **Methods**) 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 American ancestry, with variance explained 8.5% (95% CI=6.1%-11.0%).

Second, we developed genetic scores 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 would therefore not be considered out-of-sample .

Focusing first on the independent genome-wide significant PRSs, in SELECT European ancestry (n=22,173), PRS_{318} explained 9.5% (8.8%-10.3%) of variation in baseline PSA levels, while PRS_{447} (from the 447 conditional independent genome-wide significant variants identified in the joint meta-analysis) explained 10.9% (10.2%-11.8%) of the variance, which exceeded the variance explained of 8.5% (95% CI=7.8%-9.2%) by PRS_{128} (from the 128 independent variants described in our prior GWAS of 95,768 men[20]). PCPT European ancestry (n=5,725) was estimated slightly lower, and AOU European ancestry (n=11,922) slightly higher, with PRS_{128} explaining 8.6% (95% CI=7.7%-9.6%), while PRS_{318} explained 9.6% (95% CI=8.6%-10.6%), and PRS_{447} explained 11.3% (95% CI 10.2%-12.4%). We further assessed whether the presence of benign prostatic hyperplasia (BPH), a condition known to influence PSA levels, could be responsible for any of this difference. We did not observe an appreciable change after removing BPH individuals, although variance explained was estimated to be slightly higher in all populations (<0.5% higher), but always with overlapping CIs (**Table S9**).

Among SELECT African ancestry (n=1,173), PRS_{128} explained 3.4% (95% CI=1.6%-5.8%), while PRS_{318} explained 6.5% (95% CI=4.0%-9.5%), and PRS_{447} explained 7.0% (95% CI=4.5%-10.1%); the newer estimates proposed here more than doubled previous GWAS significant variant PRSs. AOU African ancestry (n=2,471) estimates were all 1-2% smaller.

Expanding to a genome-wide PRS-CSx ($PRS_{CSx\text{-}joint}$ based on the joint analysis resulted in a modest increase compared to from $PRS_{CSx\text{-}disc}$ by about 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%). We note that $PRS_{CSx\text{-}joint}$ also improved ~3% upon the Kachuri et al.[20] PRS-CSx ($PRS_{CSx\text{-}Kachuri}$) previously reported estimates of 8.60% in PCPT and 10.94% in SELECT. 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 reported $PRS_{CSx\text{-}Kachuri}$ estimates of 1.64% in SELECT, although still under half of that in European ancestry.

Third, we also examined how the variance in PSA levels explained by the PRS varied across age groups. 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 PRS 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 <50 years old, and this decreased to 8.7% (95% CI 7.0%-10.5%) for men older than 80 years.

Finally, we note that for PRS constructed from genome-wide significant independent variants, the variance explained was almost always equal to or higher when using weights corresponding to the effect sizes estimated by the multi-ancestry meta-analysis compared to using ancestry-specific weights for the same variants (**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 generally wide confidence intervals around the estimate (i.e., the smallest sample sizes likely have unstable estimates).

Relationship of PSA PRS with prostate cancer aggressiveness using Gleason score

In GERA, we performed a case-only analysis to examine the association between PSA $PRS_{CSx,disc}$ (the PRS from an out-of-sample with the highest variance explained) and Gleason score. Our results were consistent with previous work which 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]; namely, we found that in European ancestry cases, an SD increase in $PRS_{CSx\text{-}disc}$ was inversely associated with a Gleason of 7 (OR=0.78, 95% CI=0.73 to 0.84, p=1.2e-13) and ≥8 (OR=0.71, 95% CI=0.64 to 0.79, p=6.2e-10) compared to a Gleason score ≤6 (reference). Other ancestry groups had similar estimated ORs though not always statistically significant likely owing to sample size (**Table S11**), e.g., African ancestry Gleason of 7 (OR=0.88, 95% CI=0.67 to 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=2401), 11.0% 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, 3.7% were reclassified to biopsy eligible, resulting in a control net reclassification improvement (NRI) of 7.4% (95% CI=6.3% to 8.4%; **Figure 6A**, **Table S12**). In individuals with positive biopsies (i.e., cases; n=3,568), 3.7% were re-classified to eligible, while 6.9% were re-classified to ineligible, resulting in a case NRI of -3.2% (95% CI=-3.8% to -2.6%). Of cases who became ineligible, 67.6% had Gleason scores ≤7, as compared to 55.3% who remained eligible (although we note that some of these men may have had biopsies for reasons other than their PSA level, e.g., abnormal digital rectal exam, strong family history). In African American controls (n=110), 10.0% were reclassified to ineligible, while 4.5% were reclassified to eligible, resulting in an NRI of 5.5% (95% CI=1.2% to 9.7%; **Figure 6B**). In African American cases (n=390), 1.8% were reclassified to eligible and 2.8% were reclassified to ineligible, resulting in an NRI of -1.0% (95% CI=-2.0% to -0.0%). Other groups are shown in **Figure S3** with details in **Table S12**.

Associations with previously-reported prostate cancer variants

In our discovery cohort, 20 of our 184 novel PSA-associated variants (10.8%) were genome-wide significantly associated with prostate cancer in the PRACTICAL consortium's European ancestry GWAS[26] (**Tables S3-S4**). An additional 19 variants (10.3%) were associated with prostate cancer 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)[20,27], 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 prostate cancer, and an additional 11 (9.8%) were significant at a Bonferroni level (p<0.00045). With bias correction,

this was again reduced, with 5 (4.5%) genome-wide significant, and an additional 4 (3.6%) Bonferroni significant.

Associations with previously reported BPH variants

In our discovery cohort, one variant (rs1379553) was genome-wide significantly associated with benign prostatic hyperplasia (BPH), out of the 137 variants available in a large GWAS in European ancestry in the UKB[28]. An additional 8 more 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 ($p<0.045/96=0.00052$).

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 PRS was up to 16.9% in men of European ancestry, 9.5% in men of African ancestry, 18.6% in Hispanics/Latinos, and 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 prostate cancer.

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 men of European ancestry, and 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 meeting replication. AR signaling is required for normal prostate development and function but is hijacked during carcinogenesis.[29] Because prostate tumor growth and progression depend on AR signaling, androgen deprivation therapy remains a frontline treatment for progressing prostate cancer the inhibition of *AR* activity may delay progression.[30]

A total of 10.8% of the novel discovery and replication variants were found to be prostate tissue eQTLs, and another 49.7% were eQTLs in other tissues. In addition, 16 discovery variants and five meta-analysis variants were predicted to have deleterious regulatory effects. Putative deleterious genes included: *AOC1*, which regulates histamine metabolism and sensitivity to non-steroidal anti-inflammatory drugs[31,32]; *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[33,34]. 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,[35] and has been linked to obstructive azoospermia (ClinVar[36] accession SCV001860325). We detected another signal with possible links to male fertility, rs372203682,in *LMTK2*, a gene implicated in spermatogenesis[37] that also interacts with the androgen receptor and inhibits its transcriptional activity[38].

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 both SELECT and PCPT was substantially less than that in GERA, even though we evaluated only the variants from our discovery cohort that 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 only at most 0.5% higher than SELECT, and thus also substantially lower than GERA. For African ancestry men in AOU, variance explained was 2-3% lower than 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 PRS performance. The estimated variance explained was <0.5% higher when excluding men with a BPH diagnosis. 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.

With respect to the performance of different PRS methods, for PRS constructed from fine-mapped variant weights derived from the multi-ancestry meta-analysis typically surpassed or at least matched the performance of ancestry-specific weights. As expected, genome-wide PRS-CSx generally achieved 1-6% higher explained variance than the PRS limited to 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, in whom explained variance reached or exceeded estimates in European ancestry men, followed by Asian ancestry men. This is also the first time we were able to assess out of sample PRS performance 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. PRS-CSx uses a single hyperparameter to couple posterior effect sizes across ancestry groups, which may not be sufficient to capture different correlation structures among populations. In addition, HapMap3 variants used by PRS-CSx do not tag genetic variation equally well across non-European ancestries. Fine-mapping PRS methods do not limit to this set of tagging SNPs and may be more likely to capture population-specific variants. The choice of 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 making use of only 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 subset of GERA participants. 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.

Our investigation had several limitations. The replication sample sizes were somewhat small, especially for variants identified in individuals of African ancestry. Nevertheless, for African ancestry, 43% of variants met a nominal replication threshold of $p<0.05$, many more than the 5% that would be expected by chance. We also suspect that we 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 was 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 prostate cancer-free individuals, some individuals likely had undetected prostate cancer[39]. However, most novel PSA-associated variants were not associated with prostate cancer, 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 from others work on UKB European ancestry individuals[28], 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[40]. Finally, we note that our GWAS and resulting PRS 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] and substantially improved our understanding of the genetic basis of PSA levels and the value of PSA testing for prostate cancer screening. Using an ancestrally diverse study population, we detected hundreds of novel variants associated with PSA levels that were largely independent of prostate cancer or BPH. These findings explain additional variation in PSA levels, especially among men of African ancestry, who suffer the highest morbidity and mortality due to prostate cancer, 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 moves us closer to leveraging genetic information to personalize and improve PSA screening for prostate cancer across diverse populations.

Methods

Discovery Participants and Phenotype Measurements

Our primary analyses included 296,754 men from 7 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 prostate cancer[41], participants were restricted to men with no history of prostate cancer 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, meta-analysis of PSA medians from each cohort was done with the weighted median of medians method in the R v4.2.3[42] package metamediation v1.0.0[43]. 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 prostate cancer phenotyping.[26]

Mount Sinai BioMe® Biobank (BioMe). BioMe is a longitudinal cohort linked to Epic EHR[44]. Individuals were of European ancestry, Hispanic, and African ancestry.

Chicago Multiethnic Prevention and Surveillance Study (COMPASS). COMPASS is a longitudinal study of Chicagoans with currently >11,000 participants enrolled (82% African American).[45] PSA data has been described previously[46].

Men of African Descent and Carcinoma of the Prostate (MADCaP). MADCaP is a consortium of epidemiologic studies addressing the high prostate cancer burden in African ancestry men.[47,48]

Multiethnic Cohort (MEC). MEC is a prospective cohort study that enrolled >215,000 Hawaii/Los Angeles residents ages 45-75 years between 1993-1996.[49,50]

Million Veteran Program (MVP). MVP is a multi-ancestry cohort recruited nationwide. Information is obtained from electronic health records (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[51]. Groups (European, African, Hispanic, and Asian) were created using the harmonized ancestry and race/ethnicity (HARE) method.[52]

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.[53]

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, genome-wide significant variants previously-identified[20] were tested for replication in our independent discovery cohort. All statistical tests here and throughout were two-sided.

Additional PRS evaluation cohorts

For our discovery cohort results, we evaluated the PSA PRS performance and reclassification in individuals from the GERA cohort (also in the replication cohort, out-of-sample for the discovery 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) cohorts[25], which have been previously described. Briefly, the PCPT and SELECT cohorts 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 Institute of Health (NIH) AOU cohort is committed to including groups that have been historically underrepresented in research[25]. From the AOU cohort we selected individuals with PSA>0.01 between the ages of 40 and 90, with short-read whole-genome sequencing (WGS) data, and with no survey or EHR conditions/observations reflecting a history of prostate cancer. The median PSA measurement, which was used, was required to be ≤10 ng/mL. PRS were calculated with the WGS data subset to variants with population-specific allele frequency ≥1% or a population-specific allele count greater than 100 for any genetic ancestry. Genetic ancestry was determined using a random forest classifier trained on the principal component space of the Human Genome Diversity Project and 1000 Genomes Project[54].

Ethical considerations

Informed consent was obtained from all study participants. The VA Central IRB approved the MVP. The institutional review boards at Vanderbilt University and Meharry Medical College approved SCCS. Institutional review boards approved the MEC. GERA was approved by the Kaiser Permanente Northern California institutional review board and the University of California, San Francisco. The research was conducted with approved access to UKB data (#14105). The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial was approved by the institutional review board 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. Vanderbilt University Medical Center institutional review board approved BioVU. A local ethics committee approved the Malmo Diet and Cancer Study (MDCS). The University of Chicago Biological Sciences Division Institutional Review Board Committee A (#IRB12-1660) approved COMPASS. Local and national institutional review boards approved MadCAP. 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).

Genotype quality control and imputation

Study subjects were genotyped using conventional GWAS arrays (**Table S1**). Genotypes were then imputed using imputation servers[55], Minimac3[56], or IMPUTE2[57]. The vast majority of studies imputed to the 1000 Genomes Project (KGP) phase 3 reference panel[58], with one sub-study imputing to KGP phase 1 just for the X chromosome[59] and another imputing to the TOPMed reference panel[55]. Since all but two studies (>95% of participants) used genome build 37, we used build 37 here, lifting over the assembly of those from build 38 using triple-liftOver[60] v133 (2022-05-20), an extension of LiftOver[61] that accounts for regions that are inverted between builds.

Standard genotype and individual-level quality control (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, 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[62]. The minimum set of covariates adjusted for included age at PSA measurement and genetic ancestry principal components (PCs). If available, GWAS also adjusted for batch/array, body mass index (BMI), 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)[63]. We also assessed heterogeneity with Cochran’s Q across the four ancestry groups.

To identify independently associated variants (genome-wide significant, p≤5e-8) 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 genome-wide significant variant ≤10Mb). Within each clump, we used mJAM v2022-08-05[21], which uses population-specific linkage disequilibrium (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[64].

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[65]. We first prioritized genes that included a significant prostate expression quantitative trait locus (eQTL) from GTeX v8 (www.gtexportal.org). We then prioritized other significant eQTLs and finally by the 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)[66]. Gene functions were characterized with RefSeq[67]. Circos plots were generated using Circos v0.69-6[68].

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 the assessments of performance are unbiased.

We used two sets of individuals to construct the PRSs. First, we constructed PRSs from our discovery cohorts to allow assessment in GERA, PCPT, and SELECT. Second, we constructed PRSs from the meta-analysis of discovery and replication cohorts (which included GERA), with assessment in PCPT and SELECT only. For GERA we include results 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[69], 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, and a variation 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 also included (and produced a negligible increase in performance). The final scores were calculated by summing up the effect size times the (probabilistic) number of alleles at each locus with PLINK v2.00a3.7LM[70].

We also assessed the variance explained of the PRS-CSx discovery PRS within age interval bins 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. Here an individual could be in multiple bins, but using just the first measurement of that individual per age bin.

Genetic adjustment of PSA for prostate cancer 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 within 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), evaluating net reclassification in cases and controls[20].

We also tested for associations of our PSA^{adj} with a trichotomized Gleason score (≤6, 7, and ≥8) using multinomial logistic regression with the R package nnet v7.3.18[71].

Bias-corrected prostate cancer estimates

Prostate cancer associations in individuals with European ancestry in the PRACTICAL consortium[26] were adjusted for screening bias[27], using estimates previously derived[20], specifically, $\beta'_{Cancer} = \beta_{Cancer} - b\beta_{PSA}$, $SE'_{Cancer}=(SE_{Cancer}^2 + b^2SE_{PSA}^2 + SE_b^2\beta_{PSA} + SE_b^2SE_{PSA}^2)$, where SE is the standard error, and estimates were b=1.144, and SE_b=2.909e-4.

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), and PRS weights (PRS_{318}, PRS_{447}, $PRS_{CSx,disc}$, $PRS_{CSx,joint}$) 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). UK Biobank data are publicly available by request from https://www.ukbiobank.ac.uk. GTEx data was 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 (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.

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.

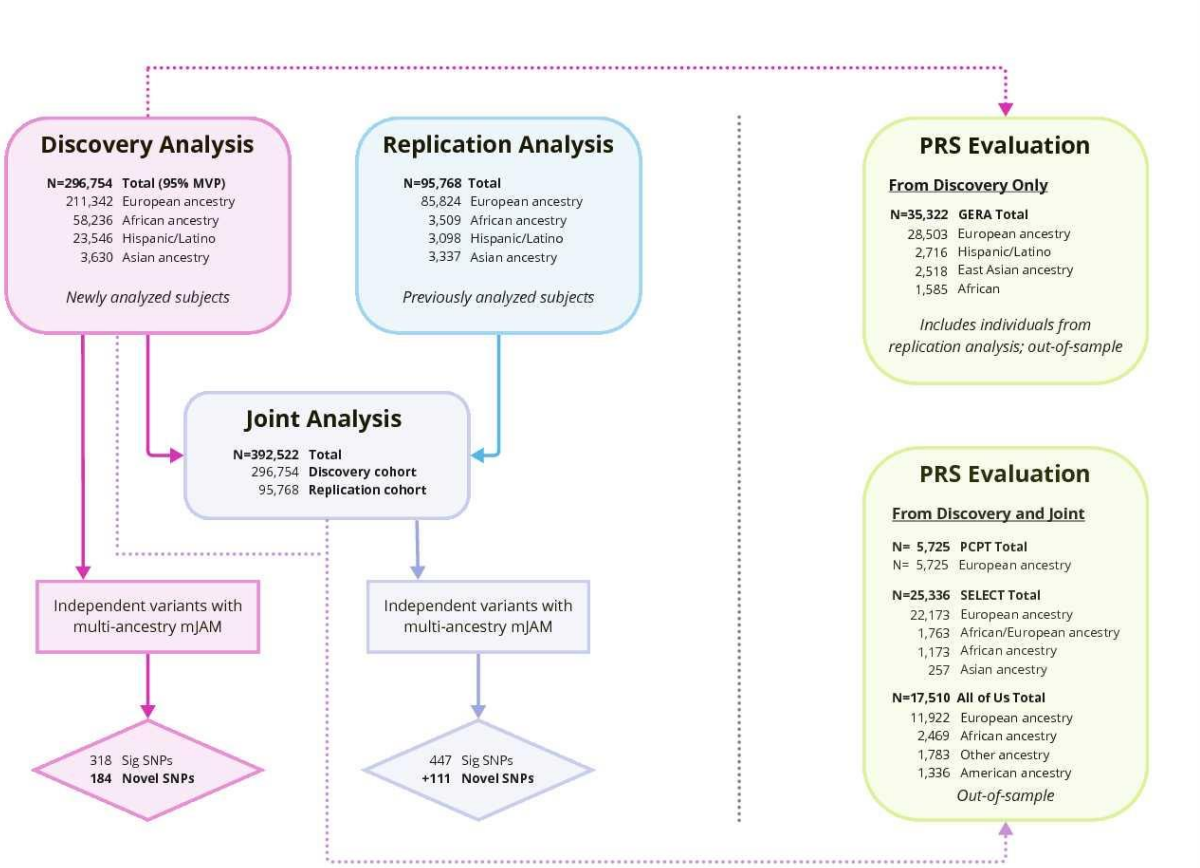


Figure 2. Circos plot showing PSA GWAS hits by chromosome from the discovery cohort. Concentric tracks are colored based on results from individual ancestries, with gray indicating results from the overall discovery meta-analysis. The top 100,000 GWAS SNPs (with the smallest p-values) per ancestry are 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.

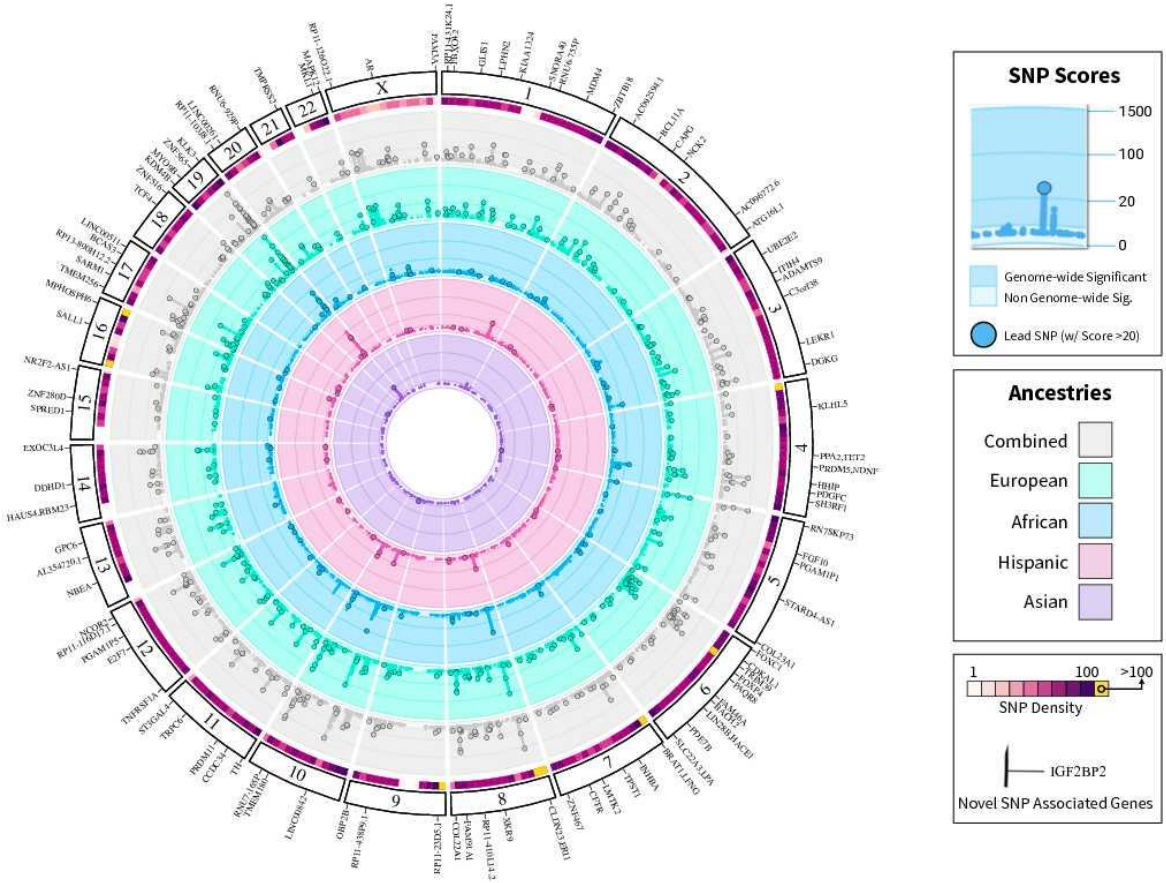


Figure 3. Manhattan plot showing results from the joint multi-ancestry meta-analysis of the discovery (n=296,754) and replication (n=95,768) studies. Only genome-wide significant associations (p<5e-8) are plotted. The joint analysis detected 447 independent genome-wide significant PSA-associated SNPs. 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.

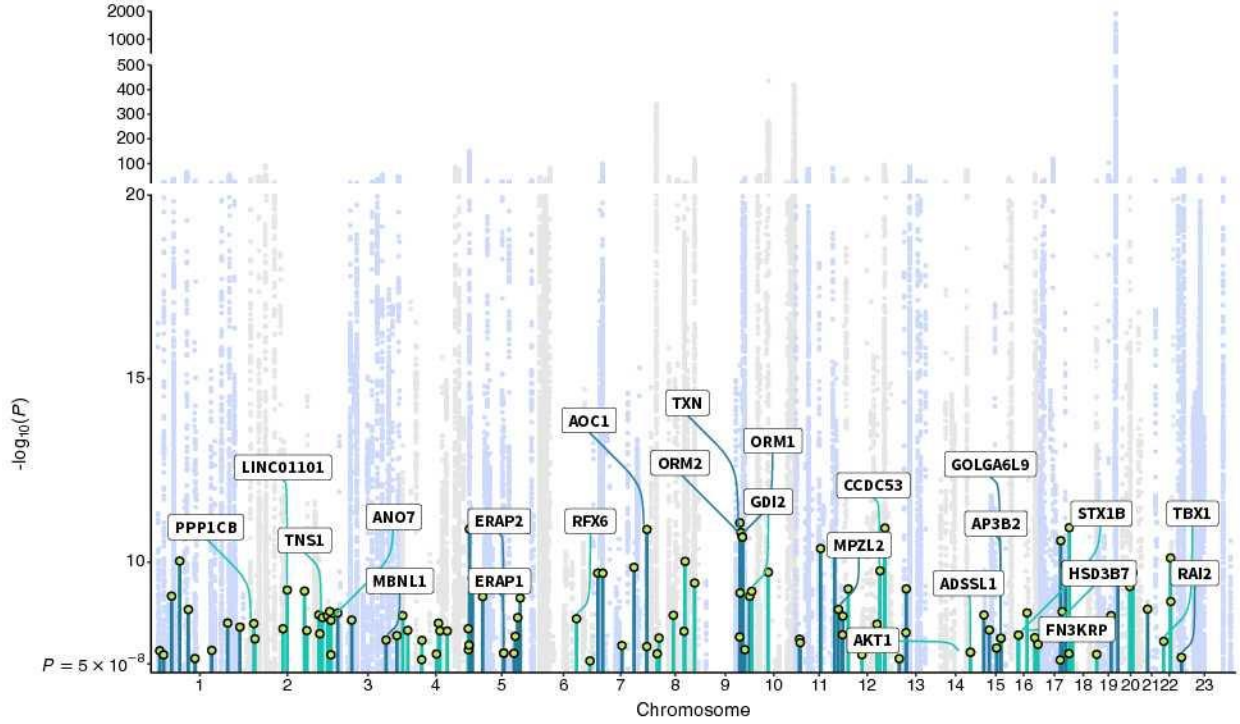


Figure 4. Plot of the relationship between minor allele frequency and estimated effect sizes for PSA GWAS hits. Each point represents one of the 447 independent genome-wide significant SNPs identified in our mJAM multi-ancestry GWAS joint meta-analysis. The SNP effect sizes are expressed in ln(PSA) per 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).

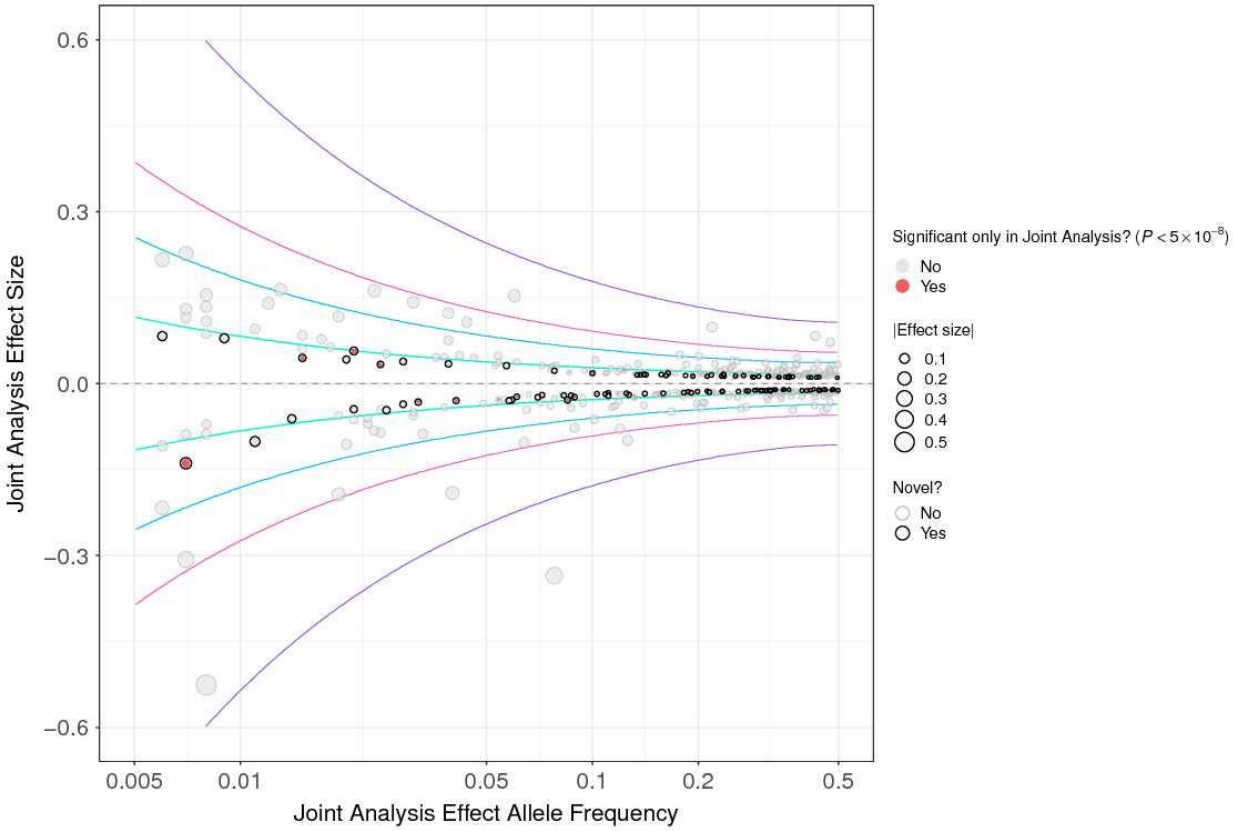


Figure 5. Variance in PSA levels explained by PRS. We constructed PRSs for PSA from our discovery cohorts to allow assessment in GERA and our PRS validation cohorts (PCPT and SELECT). We also constructed PRSs from the joint meta-analysis of discovery and replication cohorts, with assessment in our validation 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 PRS **(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.

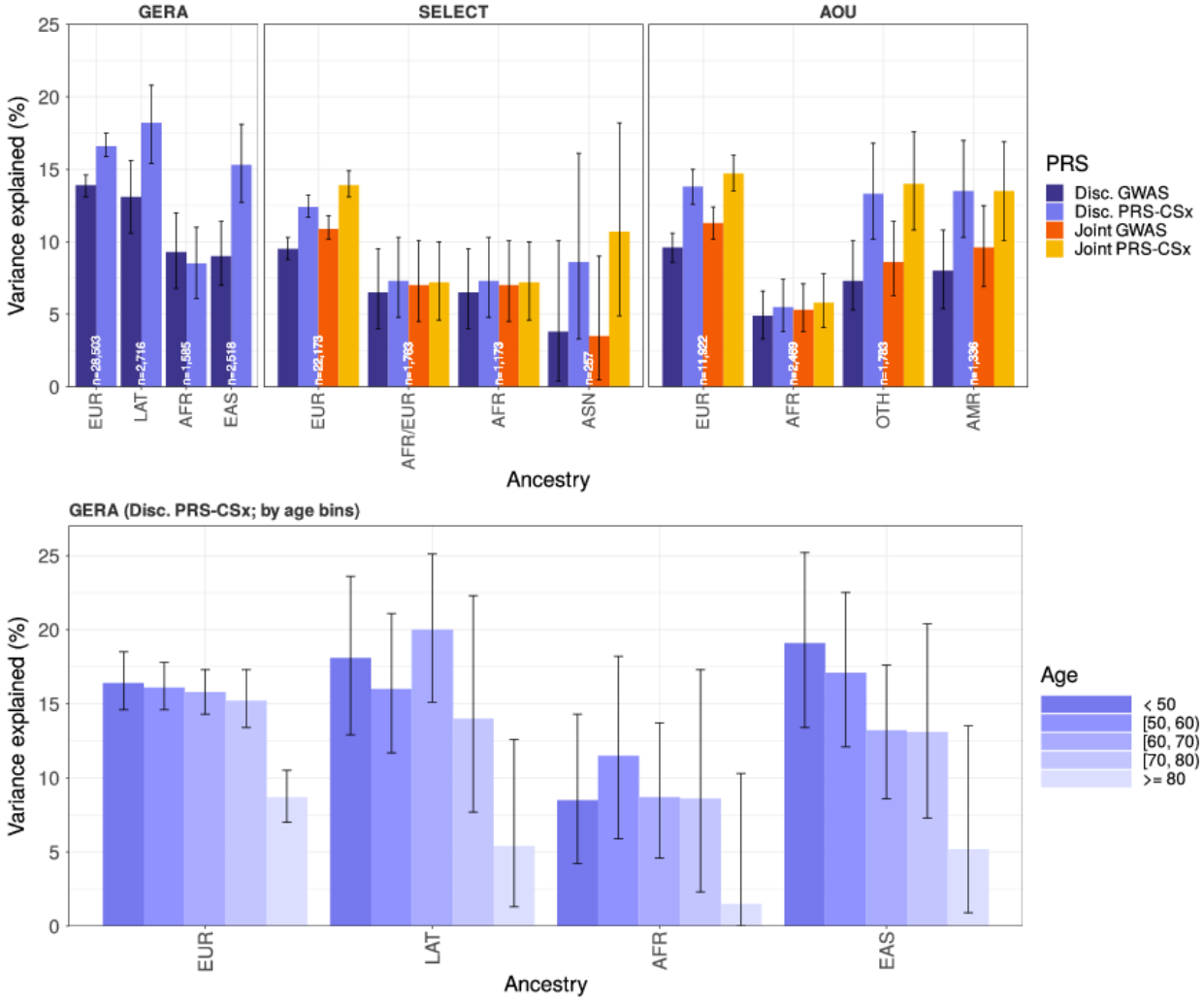
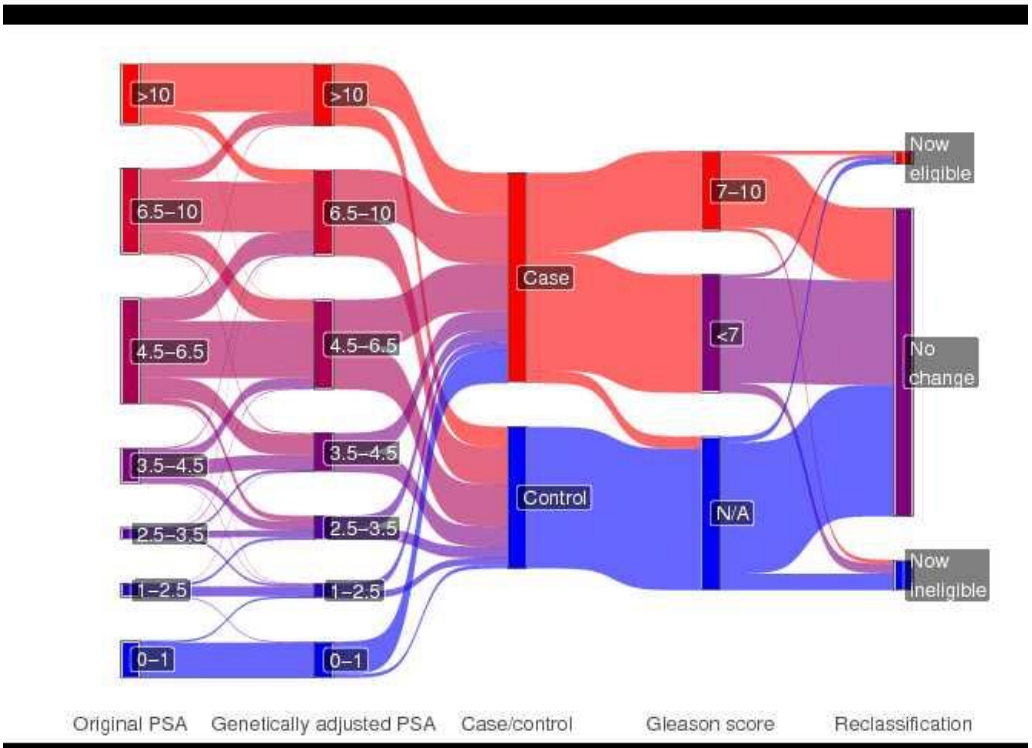
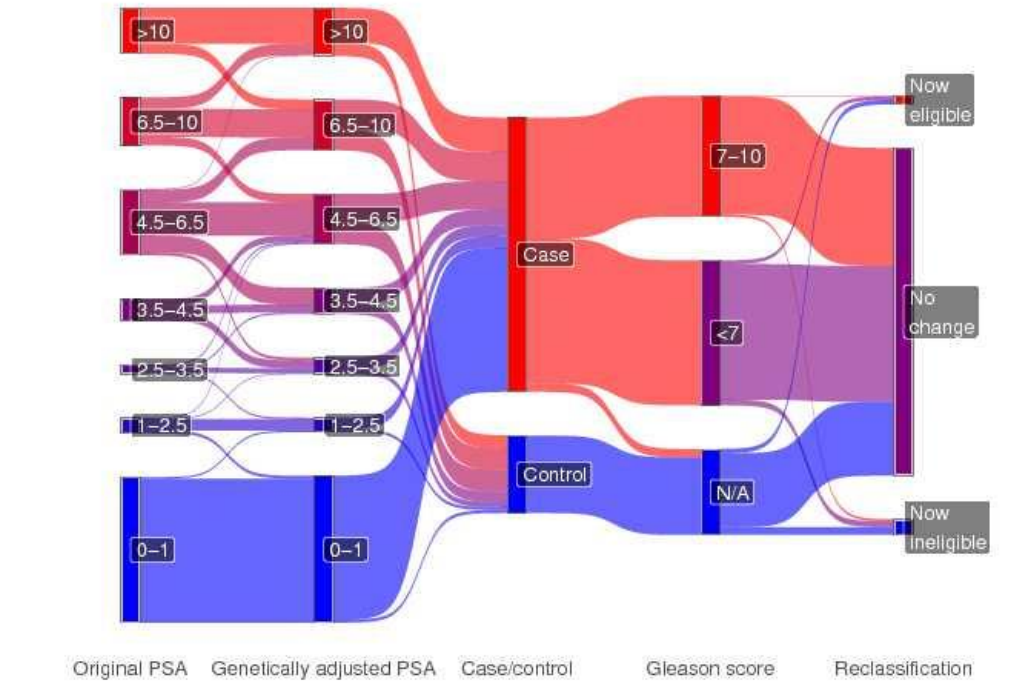


Figure 6. Biopsy reclassification with genetically-adjusted PSA. PSA levels were adjusted using the PRS-CSx PRS estimate from the out-of-sample discovery cohort, assessed in GERA using age-specific cutoffs in **(A)** Europeans and **(B)** African Americans (see **Methods**). GERA Latinos and East Asians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





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

(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. 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).
27. Dudbridge, F. *et al.* Adjustment for index event bias in genome-wide association studies of subsequent events. *Nat Commun* **10**, 1561 (2019).
28. 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).
29. Kim, J. & Coetzee, G. A. Prostate specific antigen gene regulation by androgen receptor. *J Cell Biochem* **93**, 233–241 (2004).
30. Heinlein, C. A. & Chang, C. Androgen Receptor in Prostate Cancer. *Endocrine Reviews* **25**, 276–308 (2004).
31. 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).
32. Amo, G. *et al.* FCERI and Histamine Metabolism Gene Variability in Selective Responders to NSAIDS. *Front Pharmacol* **7**, 353 (2016).
33. Miller, C. L. *et al.* Integrative functional genomics identifies regulatory mechanisms at coronary artery disease loci. *Nat Commun* **7**, 12092 (2016).
34. 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).
35. Cuppens, H. & Cassiman, J.-J. CFTR mutations and polymorphisms in male infertility. *Int J Androl* **27**, 251–256 (2004).
36. Landrum, M. J. *et al.* ClinVar: improvements to accessing data. *Nucleic Acids Research* **48**, D835–D844 (2020).
37. 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).
38. Cruz, D. F., Farinha, C. M. & Swiatecka-Urban, A. Unraveling the Function of Lemur Tyrosine Kinase 2 Network. *Frontiers in Pharmacology* **10**, (2019).
39. 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).
40. 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).
41. D'Amico, A. V. Risk-based management of prostate cancer. *N Engl J Med* **365**, 169–171 (2011).
42. R Core Team. R: A language and environment for statistical computing. (2012).
43. 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).
44. 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).
45. Aschebrook-Kilfoy, B. *et al.* Cohort profile: the ChicagO Multiethnic Prevention and Surveillance Study (COMPASS). *BMJ Open* **10**, e038481 (2020).
46. 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).
47. 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).
48. 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).
49. 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).
50. Kolonel, L. N. *et al.* A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. *Am. J. Epidemiol.* **151**, 346–357 (2000).
51. 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).
52. 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).
53. Signorello, L. B. *et al.* Southern community cohort study: establishing a cohort to investigate health disparities. *J Natl Med Assoc* **97**, 972–979 (2005).
54. 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).
55. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program | Nature. https://www.nature.com/articles/s41586-021-03205-y.
56. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat Genet* **48**, 1284–1287 (2016).
57. 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).
58. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
59. Consortium, T. 1000 G. P. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
60. 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).
61. Karolchik, D. *et al.* The UCSC Genome Browser database: 2014 update. *Nucl. Acids Res.* **42**, D764–D770 (2014).
62. 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).
63. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
64. Hoffmann, T. J. *et al.* Genome-wide association study of prostate-specific antigen levels identifies novel loci independent of prostate cancer. *Nat Commun* **8**, 14248 (2017).
65. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* **8**, 1826 (2017).
66. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310–315 (2014).
67. 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).
68. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome*

Res **19**, 1639–1645 (2009).
69. Ruan, Y. *et al.* Improving polygenic prediction in ancestrally diverse populations. *Nat Genet* **54**, 573–580 (2022).
70. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, (2015).
71. W. N. Ripley & B. D. Ripley. *Modern applied statistics with S*. (Springer, 2002).