SUPPLEMENTARY NOTE

Study-specific GWAS quality control and imputation

UK Biobank (UKB)

Genotyping and imputation for the UK Biobank cohort have been previously described[1]. Briefly, participants were genotyped on the UKB Affymetrix Axiom array (89%) or the UK BiLEVE array (11%) with imputation performed using the Haplotype Reference Consortium (HRC) and the merged UK10K and 1000 Genomes (1000G) phase 3 reference panels. Genetic ancestry principal components (PCs) were computed using fastPCA based on a set of 407,219 unrelated samples and 147,604 genetic markers[1]. Association analyses in UKB were restricted to individuals of predominantly European ancestry based on self-report ("White") and after excluding samples with any of the first two genetic ancestry principal components (PCs) outside of 5 standard deviations (SD) of the population mean, as previously descibed[2]. We removed samples with discordant self-reported and genetic sex, as well as one sample from each pair of first-degree relatives identified using KING[3]. Using a subset of genotyped autosomal variants with (MAF)≥0.01 and call rate ≥97%, we filtered samples with call rates <97% or heterozygosity >5 SD from the mean. We excluded variants that were out of Hardy-Weinberg equilibrium in cancer-free individuals ($P_{HWE}<1\times10^{-5}$) or had low imputation quality (INFO<0.30). Analyses were restricted to variants with MAF≥0.005.

Genetic Epidemiology Research on Adult Health and Aging (GERA)

Genotyping, imputation and QC of the GERA cohort has been previously described[4-7]. Briefly, all men were genotyped for over 650,00 SNPs on four race/ethnicity-specific Affymetrix Axiom arrays that were optimized for individuals who self-reported non-Hispanic white, Latino, East Asian, and African American, respectively[6,7]. Genotype quality control (QC) procedures and imputation for the original GERA cohort assays were performed on an array-wise basis, as previously described[4,8]. Pre-phasing was done by SHAPEIT v2.5[9] and imputation with IMPUTE2 v2.3.1[10] using the 1000 Genomes Project October 2014 release with 2.504 samples. The top 10 genetic ancestry principal components from Eigenstrat v4.2[11], were included in the linear model as ancestry covariates. Analyses were conducted according to self-identified race/ethnicity groups, since participants genotyped on four race/ethnicity-specific Affymetrix Axiom arrays, restricting to variants with INFO≥0.3, MAF≥0.01, and Hardy-Weinberg equilibrium p-value $\geq10^{-5}$ in controls."

Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial

Extensive quality control filtering was performed for subsequent imputation and association analyses. Iterative 80% and 95% sample- and variant-level call rate filters were applied to remove poorly genotyped samples and variants. Samples with greater than 20% estimated contamination based on VerifyIDintensity[12] were also removed. Heterozygosity outliers were detected using absolute values from PLINK method-of-moments F coefficients greater than 0.2. Samples with discordant self-reported and genetically inferred sex were identified by comparing the reported sex with the observed sex based on X chromosome method-of-moments F coefficient from PLINK using 0.5 as the threshold (F coefficients are close to 0.0 for males and 1.0 for females). Discordant samples were further screened for sex chromosome aneuploidies by STR profiling using the Identifier assay. Pairwise genotype concordance for all subjects was assessed to discover unexpected replicates, defined as genotype concordance greater than 95% based on a set of LD-pruned SNPs. Pairwise genotype concordance for all subjects from different datasets within the same platform was also assessed. Cumulatively, samples were filtered to remove abnormal levels of heterozygosity (N=12), sex discordance (N=31), sex-chromosome abnormalities (N=47), unexpected duplicates (N=130), discordant expected duplicates (N=14) and unexpected duplicates (N=126). We also removed one individual in a set of third-degree relatives (N=291).

Genetic ancestry for PLCO Atlas participants, form which the subjects included in the present analysis were selected, was determined using GRAF[13] on a set of 10,000 pre-selected fingerprinting variants. GRAF assigned individuals into the following 9 ancestral groups: “African”, “African American”, “East Asian”, “European”, “Hispanic1”, “Hispanic2”, “Other”, “Other Asian”, and “South Asian”. Hispanic1 included individuals of Dominican or Puerto Rican ancestry whereas Hispanic2 included individuals of Mexican or Latin American ancestry. For parsimony we merged “African” and “African American” into a “African American (Combined)” and “East Asian” and “Other Asian” into a “East Asian (Combined)”. The largest ancestral sets in the PLCO Atlas included European (N=100,448), African American (Combined) (N=4,576) and East Asian (Combined) (N=3,528).

After QC exclusions, the PLCO Atlas genotyping effort resulted in a total of 112,065 DNA samples genotyped across 110,562 unique individuals on high-density Illumina genotyping arrays. For participants genotyped on multiple genotyping arrays (N=1,192), only genotype data from one array was included following the prioritization of Global Screening Array (GSA) > Oncoarray > Omni2.5M > OmniExpress (OmniX) to ensure non-redundant subject-level genotyping data. The predominant genotyping array was the GSA (N=84,731), followed by the OncoArray (N=16,893), Omni2.5M (N=7,211) and OmniX (N=1,727).

Prior to submitting data to the Michigan Imputation Server, we removed variants with minor allele frequency ≤ 0.01, variant-level missingness ≥ 0.05, and Hardy Weinberg equilibrium exact test p-value $\leq 1\times10^{-6}$. Data from each genotyping platform were then analyzed using a community-recommended script for aligning data to reference datasets (HRC-1000G-check-bim.pl, from https://www.well.ox.ac.uk/~wrayner/tools/). The script was modified to support TOPMed 5b as a reference panel using a pre-existing test imputation with 1000 Genomes subjects versus the TOPMed 5b reference panel. Data were uploaded to the MIS in GRCh37/hg19 and lifted over by the MIS. Pre-phasing using phased reference data from TOPMed release 5b was conducted using EAGLE 2.4. Imputation was conducted against the same reference panel using minimac4.

Following imputation, raw data were partitioned into ancestry subpopulations based on GRAF groupings to estimate ancestry-specific imputation quality. After partitioning by ancestry and recomputing imputation quality Rsq values, each platform and ancestry pair was cleaned according to the filtering method described by Kowalski et al[14]. Briefly, all variants with Rsq <0.3 were removed to be consistent with traditional quality filters. Remaining variants were partitioned into MAF bins and each bin was filtered, starting at the variant with the lowest Rsq, until the average Rsq of remaining variants within the MAF bin was at least 0.9.

BioVU

Participants were identified using Vanderbilt University Medical Center’s (VUMC) BioVU resource, a DNA biobank comprising ~270,000 individuals and linked to a de-identified electronic health record (EHR)[15]. All participants (n=8,074) were genotyped on Illumina’s Expanded Multi-Ethnic Genotyping Array (MEGA^{EX}) platform. Genetic ancestries were assigned by running principal component analysis using SNPRelate[16] on a set of pruned SNPs (Rsq<0.5, MAF≥0.1). Subjects were classified as being of European ancestry if their first two PCs were within 4 SDs of the median for the subjects reporting “White” as their race. Subjects were classified as being of African ancestry if their first two PCs were within 4 SDs of the median for subjects reporting their race as “Black”. All quality control procedures were performed using PLINK version 1.90. We removed samples with indeterminate genetic sex or discordant self-reported and genetic sex. We removed one randomly selected sample out of each pair of related individuals (pi-hat≥0.2) identified using identity-by-descent. Among subjects passing these initial steps, we removed all palindromic variants and variants with a call rate<3% or MAF<1%. We then excluded subjects with SNP missingness >3% or heterozygosity >5 SD from the mean. Prior to imputation, data were pre-processed using the HRC-1000G-check tool v4.2.5 (http://www.well.ox.ac.uk/~wrayner/tools/) and pre-phased using Eagle v2.4.128. Genetic data was imputed on the Michigan Imputation Server using 1000 Genomes phase 3 version 5 as the reference panel. Final association analyses were performed on imputed variants with MAF>0.005, INFO>0.3, and HWE v-value>$1x10^{-5}$.

Malmö Diet and Cancer Study (MDCS)

4069 individuals from the MDCS were genotyped on various versions of Illumina Omni chips at different times, for a total of seven different batches. Data from all seven batches were merged. We noted that some SNPs appeared more than once under different names on the same Illumina chip. For these cases, we kept the SNP with the higher genotyping rate. Ambiguous SNPs (e.g. A/T or C/G alleles), SNPs that were not bimorphic, and variants that were not single nucleotide polymorphisms, were removed. From this, only SNPs that we could unambiguously map to release 1 of the 1000 Genomes project were kept; the final files from this were aligned to build 37 of the human genome. Individuals with more than 10% missingness were removed. Next, SNPs with a missingness rate greater than 10% or deviation from Hardy-Weinberg equilibrium ($p<0.001$) were removed.

At this stage, the principal components of ancestry were computed. Individuals for whom the inferred sex based on X chromosome heterozygosity was not male, or for whom there were more than two genetic mismatches with 40 SNPs we had previously genotyped in these samples with targeted genotyping[17] were excluded. To identify genetic ancestry, the SNP chip data was combined with data from HapMap phase 3 for all SNPs in common across all genotyping batches. We then filtered for SNPs with less than 0.01% missingness, and further filtered for SNPs in LD using the –indep-parwise 50 5 0.05 command in Plink v1.07. We then ran SMARTPCA on the resulting 18,299 SNPs and kept the top 10 principal components. Analyses were restricted to individuals of European ancestry based on clustering with HapMap reference populations and exclusion of outliers with a Z-score on PC1 and PC2 greater than 5.

To impute these individuals using the TopMed imputation server, a custom Python script using Spark was used to convert the master tped/tfam file to a VCF file for each chromosome. SNPs with missingness greater than 5% or Hardy-Weinberg equilibrium P<1e-05 were removed. The input Plink file was first aligned to the build37 reference genome on the basis of chromosome, position and alleles so that the resulting VCF file can properly report genotypes based on the reference and alternate alleles. A total of 847,133 SNPs that passed pre-imputation QC were uploaded to the TopMed imputation server for imputation. As the server converts data to build 38, the resulting files were analyzed in build 38.

From the resulting imputed files, only individuals not known to be a case by December 31, 2014, with individual missingness less than 3%, a Z-score less than 5.0 for heterozygosity, and PSA between 0.01 and 10 ng/mL were included. A linear regression model in which the covariates (Supplementary Table 21) were used to predict log(PSA) was fit using robust linear regression with Tukey biweights using the RLM function in the statsmodels Python package, and the residuals were extracted for testing with the individual SNPs.

Prostate Cancer Prevention Trail (PCPT) and Selenium and Vitamin E Cancer Prevention Trial (SELECT)

This analysis includes a subset of subjects from PCPT and SELECT who were genotyped on the Illumina Global Screening Array (GSA) 24v2-0. Genotyping calling and quality control were performed at the Center for Inherited Disease Research (CIDR) at Johns Hopkins University using GenomeStudio version 2011.1, genotyping module version 1.9.4, and GenTrain version 1.0. After removal of samples that failed to produce valid output during initial processing and clustering, the completion rate was 0.9951 and 0.9959 in PCPT and SELECT, respectively. A two-stage filter by completion rate threshold of 0.8 for samples and 0.8 for vairants, followed by 0.95 for samples and 0.95 for variants was performed. In PCPT no samples were excluded by the initial filter and the subsequent filter excluded 36 samples and 4076 variants, resulting in 6325 samples at 701,852 variants. In SELECT the first stage excluded 4 samples, and 6276 variants. The subsequent filter excluded 81 samples and 3776 variants, resulting in 27,608 samples at 702,137 variants.

Samples that are contaminated during extraction or staging activities can be identified and excluded prior to genotyping via use of an STR fingerprinting (IdentifilerTM) assay. For samples that became contaminated during array processing, or for those samples not screened by an STR assay prior to array processing, sample contamination is predicted by running the tool VerifyIDintensity[12] on each sample that passes all completion rate filters or has a median raw intensity > 6000. The tool is run one sample at a time and uses a

population frequency file created with the 1000G all population frequencies for the GSA SNPs. Samples with predicted contamination > 10% were excluded (n=1 in PCPT and n=2 in SELECT).

Sex of each remaining subject is verified by comparing the reported sex with the observed sex based on X chromosome method-of-moments F coefficient, expected to be close to 0.0 and 1.0 for females and males respectively. Subjects with F coefficients of 0.5 were excluded. Pairwise genotype concordance for all subjects was assessed to discover unexpected replicates. Subjects with genotype concordance > 95% at a set of LD-pruned SNPs were considered replicates. Identity-by-descent (IBD) for all subject pairs were using PLINK and close (1st and 2nd degree) relatives were identified based on a threshold of 0.20. One randomly selected sample from pair of relatives was retained.

Ancestry was estimated using a set of LD-pruned markers and running SNPWEIGHTS[18] with the reference panel provided containing the following populations: European, West African, and East Asian, with a threshold of 0.8 used for imputed ancestry designation. Subjects were assigned to a single ancestry group if the ancestry score was equal to or above 0.80 for just one group. Individuals were assigned to an admixed cluster if their ancestry score was greater than 0.20 and less than 0.80 for only one group (eg: ADMIXED_AFR where AFR=0.75, EUR=0.17, EAS=8). Intermediate ancestry clusters included individuals with ancestry scores matching those criteria in multiple groups: 0.20<AFR_EUR<0.80 (eg: AFR=0.65, EUR=0.33) and 0.20<EAS_EUR<0.80 (eg: EUR=0.55, EAS=0.43). Autosomal heterozygosity was assessed using the method-of-moments F coefficient ([observed hom. count] - [expected count]) / ([total observations] - [expected count]) calculated within each ancestry cluster. Heterozygosity outliers were identified and excluded using a threshold of 0.10. To resolve more detailed population substructure, principal components analysis was performed with SMARTPCA in EIGENSOFT (https://github.com/chrchang/eigensoft) on a set of LD-pruned markers after splitting by ancestry cluster, excluding duplicates, and related subjects. Genetic ancestry principal components (PC's) were not computed for small clusters (n<50) or individuals who failed other QC filters. For validation of PGS_{PSA} in PCPT and SELECT we combined ADMIXED_AFR and AFR_EUR and treated this as a single group with intermediate AFR and EUR ancestry proportions (AFR/EUR). ADMIXED_EAS and EAS_EUR were also combined into a single group with intermediate EAS and EUR ancestry (EAS/EUR).

Pre-imputation variant-filtering criteria included: minor allele frequency (MAF) > 0.1%, Hardy-Weinberg equilibrium p-value >10^{-6}, and SNP call rate > 98%. A total of 491,015 genotyped variants remained after these quality control (QC) steps. To prepare the genotype data for the Trans-Omics for Precision Medicine (TOPMed) Imputation, we checked the QC'd plink files against the TOPMed reference SNP list in advance of imputation. For PCPT, a total of 6055 samples and 474,046 variants were submitted to the imputation server. A total 26,365 samples were genotyped on the GSAMD-24v2-0 Illumina genotyping array. The pre-imputation sample-filtering and variant-filtering process followed the same steps as for the PCPT study. After QC, a total of 402,993 variants of 26,297 subjects were input to the TOPMed Imputation Server for imputation.

References

1. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).
2. Rashkin, S.R. *et al.* Pan-cancer study detects genetic risk variants and shared genetic basis in two large cohorts. *Nat Commun* **11**, 4423 (2020).
3. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867-73 (2010).
4. 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).
5. Hoffmann, T.J. *et al.* A large multiethnic genome-wide association study of prostate cancer identifies novel risk variants and substantial ethnic differences. *Cancer Discov* **5**, 878-91 (2015).
6. Hoffmann, T.J. *et al.* Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. *Genomics* **98**, 422-30 (2011).
7. Hoffmann, T.J. *et al.* Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array. *Genomics* **98**, 79-89 (2011).
8. Kvale, M.N. *et al.* Genotyping Informatics and Quality Control for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. *Genetics* **200**, 1051-60 (2015).
9. Delaneau, O., Marchini, J. & Zagury, J.F. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**, 179-81 (2011).
10. 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**, 955-9 (2012).
11. Banda, Y. *et al.* Characterizing Race/Ethnicity and Genetic Ancestry for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. *Genetics* **200**, 1285-95 (2015).
12. Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet* **91**, 839-48 (2012).
13. Jin, Y., Schaffer, A.A., Sherry, S.T. & Feolo, M. Quickly identifying identical and closely related subjects in large databases using genotype data. *PLoS One* **12**, e0179106 (2017).
14. Kowalski, M.H. *et al.* Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. *PLoS Genet* **15**, e1008500 (2019).
15. Roden, D.M. *et al.* Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin Pharmacol Ther* **84**, 362-9 (2008).
16. Zheng, X. *et al.* A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326-8 (2012).
17. Klein, R.J. *et al.* Evaluation of multiple risk-associated single nucleotide polymorphisms versus prostate-specific antigen at baseline to predict prostate cancer in unscreened men. *Eur Urol* **61**, 471-7 (2012).
18. Chen, C.Y. *et al.* Improved ancestry inference using weights from external reference panels. *Bioinformatics* **29**, 1399-406 (2013).