Germline *HOXB13* variant contributes to risk of prostate cancer in men of African ancestry

Supplemental Information

Supplemental Methods

Participants

Participants included men of African ancestry with genome-wide genotyping data, with N=9,464 from the African Ancestry Prostate Cancer GWAS Consortium (AAPC1M), N=8,184 from the ELLIPSE/PRACTICAL OncoArray Consortium (ONCO-AAPC), N=2,638 from the California Uganda Study (CA UG Study), N=1,274 from the Ghana Study (GPS), and N=801 from the Men of African Descent and Carcinoma of the Prostate (MADCaP) Network (**Supplemental Tables 1-2**)[1-4]. This study was conducted with the approval of the institutional review boards at each participating institution, and all subjects provided written informed consent to participate in the study.

Genotyping and Imputation

AAPC1M samples were genotyped on the Illumina Human 1M array and ELLIPSE/PRACTICAL ONCO-AAPC samples were genotyped on the Illumina OncoArray, while the CA UG Study was genotyped on the H3 Africa array, GPS on the HumanOmni array, and MADCaP on the custom MADCaP array. Genotype calling and quality control are described in detail elsewhere[1, 3-6].

The rs77179853 variant (TA>T) was imputed using Phase 3 of the 1000 Genomes Project (1KGP)[7] and version r2 of the Trans-Omics for Precision Medicine

(TOPMed) program[8] as the reference panels. In Phase 3 of 1KGP, 3 carriers were observed out of 2,501 participants (0.12% carrier frequency), while in Freeze 8 of TOPMed, 172 carriers were observed out of 132,345 participants (0.13% carrier frequency reported in Bravo: https://bravo.sph.umich.edu/freeze8/hg38/); based on the latter carrier frequency, we estimated that approximately 126 of the 97,256 TOPMed r2 participants used for the TOPMed imputation in the present study were carriers. Using the 1KGP reference panel, the info score for rs77179853 was 0.819 in the AAPC1M samples, 0.748 in the ONCO-AAPC samples, 0.684 in the CA UG Study, 0.753 in the GPS, and 0.819 in the MADCaP Consortium (**Supplementary Table 4**). Using the TOPMed reference panel, the info score for rs77179853 was 0.921 in the AAPC1M samples, 0.918 in the ONCO-AAPC samples, 0.949 in the CA UG Study, 0.967 in the GPS, and 0.941 in the MADCaP Consortium (**Supplementary Table 4**).

We genotyped rs77179853 using TaqMan in a subset of 1,555 men, including 124 carriers based on imputation with TOPMed (n=82) and 1KGP (n=42) from AAPC1M, ONCO-AAPC, and the Ghana Prostate Study. Also included were 1,431 non-carriers, 1,409 of which were Ugandan, based on TOPMed and 1KGP from ONCO-AAPC and the CA UG Study. Based on TaqMan genotyping, 81 of the 82 TOPMed identified carriers were confirmed to be carriers, whereas the 42 1KGP identified carriers were genotyped as non-carriers. All 1,431 non-carriers were all confirmed to be non-carriers based on TaqMan genotyping.

Pathogenic, Likely Pathogenic, and Deleterious Variant Definition

We identified pathogenic, likely pathogenic, and deleterious (P/LP/D) variants in *HOXB13* as previously described[9]. Briefly, variants had either a) a Variant Effect

Predictor (VEP) Impact score of “high”[10], representing variants with deleterious (protein truncating or splice altering) functional consequences, or b) a Pathogenic or Likely Pathogenic ClinVar classification[11] to identify known pathogenic variants, including non-synonymous substitutions. Variants identified are presented in **Supplementary Table 5**.

Statistical Analyses

To evaluate the association between germline variant rs77179853 genotype and prostate cancer risk, logistic regression models were used, adjusting for age at diagnosis for cases or at study visit for controls, study, and the first ten principal components (described below) to account for potential population stratification. Models were run separately for participants from AAPC1M, ONCO-AAPC, CA UG Study, GPS, and MADCaP, and the resulting summary statistics were meta-analyzed using METAL[12] or the R package “meta”. Analyses were repeated comparing controls to cases with low-risk disease (Gleason<7 tumors, stage T1/T2, and PSA<10 ng/ml), intermediate-risk disease (Gleason=7 tumors, stage T1/T2, and PSA=10-20 ng/ml), and high-risk disease (stage T3/T4, Gleason 8-10 tumors, PSA=20-100 ng/ml, metastatic disease, PSA>100 ng/ml, or PCa death). Analyses were also repeated within African ancestry men from the Americas and from West African countries (Ghana, Nigeria, and Senegal). Additive models were used to test the effect of the minor allele. P-values less than 0.05 were considered statistically significant. Sensitivity analyses were performed using dosages instead of genotypes to evaluate whether imputation dosage uncertainty impacted results; logistic regression models using dosages led to highly similar results (not shown).

Principal components (PCs) were calculated to account for potential population stratification using principal component analyses performed with KING[13], PC-AiR[14], and PC-Relate[15] or EIGENSTRAT[16]. Common (MAF≥1%) and independent genotyped autosomal SNPs were used to calculate PCs across all five studies (AAPC1M, ONCO-AAPC, CA UG Study, GPS, and MADCaP) and separately within each study. PCs calculated across studies were used to create PC plots (**Figure 1 and Supplementary Figure 2**), while PCs calculated within studies were included as covariates in regression analyses (**Table 1**), which were performed within each of the five studies and meta-analyzed across populations.

Ancestry proportions were calculated using ADMIXTURE[17] with 20,494 common and independent SNPs and an unsupervised K=4 approach. African and European ancestry individuals from 1KGP[7], as well as all MADCaP, Ghanaian, and Ugandan participants, were included as reference samples, and ancestry proportions were projected onto the remaining samples using the population structure learned from the reference panel. Resulting components corresponded to proportions of Eastern, Southern, and Western African and European ancestry. ADMIXTURE was similarly run with K=2 but on the full sample (without projections) to determine global African versus European ancestry.

Estimating Allelic Age

We used two approaches to estimate the allelic age of the *HOXB13* X285K variant. Note that because selection can reduce allelic age[18], the estimates calculated are upper bounds.

1. Estimating Allelic Age based on Genealogic Tree Construction

We estimated the age of the derived allele rs77179853 in two separate cohorts: 801 African individuals from MADCaP cohort and 1,760 individuals from Ghana (N=1,274) and Uganda (N=486). In total, 47,936 and 39,409 biallelic markers directly genotyped and imputed with R^2 >0.9 were available on chromosome 17 for the MADCaP and Ghana/Uganda cohorts, respectively. We reconstructed the genealogical tree sequence using RELATE[19] with the default parameters suggested in the user manual. The default genetic map in hg19 as supplied by EAGLE[20] was used for tree reconstruction. We report both the minimum and maximum allelic age of rs77179853 using the age_begin and age_end columns of the RELATE .mut output file, which is based on the most recent and ancient time estimates, respectively, of the branch leading to the clade of carrier haplotypes (**Supplementary Figure 5**). We assessed the uncertainties of the age estimates using jackknife standard errors computed by splitting each of the samples into 20 equally sized blocks. Age is estimated based on a generation time of 25 years.

2. Estimating Allelic Age based on Derived Allele Frequency

The age of the *HOXB13* X285K variant is estimated following the approach presented by Slatkin and Rannala[21] based on derived allele frequency. The cumulative distribution of allele age is given by:

$$P(t_1 \leq t) \simeq (1 - p)^{-1+n/(1+nt/2)}, \qquad (1)$$

where n is the sample[21], and the derivative of the function yields the probability density distribution of t_1.

We pooled the samples from Ghana ($n_G = 751$) and Nigeria ($n_N = 320$), yielding a sample size of 1,071 and a *HOXB13* X285K risk allele frequency of $\hat{p} = 0.37\%$. Given the limited sample size ($n = 1{,}071$) and the binomial sampling, it is desirable to account for uncertainty in the point estimate $\hat{p}$ of the allele frequency. Assuming that the sampled populations are in Hardy-Weinberg equilibrium, the allele frequency is binomially distributed with parameters $2n$ and $\hat{p}$. Thus, the variance of the observed allele frequency is $\sigma^2 = \frac{1}{2n}\hat{p}(1-\hat{p})$ [22]. Given the observed allele frequency and its associated variance, the confidence intervals for $\hat{p}$ are defined by:

$$\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{2n}}, \qquad (2)$$

where z is the quantile of a standard normal distribution[22]. For our pooled population, p is normally distributed with a mean of 0.0037 and a standard deviation of 0.0013 ($p \sim N(0.0037, 0.0013)$. Since allele age depends on the derived allele frequency, we define their joint distribution of $P(t_1, p)$.

To obtain the joint probability distribution of the allele age, the allele frequency is scaled in terms of twice the effective population size ($2N_e$), which is why estimates of N_e are required. To obtain estimates for N_e, we utilized the relationship between gene diversity (H) and scaled mutation rate (θ). Assuming Hardy-Weinberg equilibrium and a neutral mutation rate μ, H is given by:

$$H = \frac{\theta}{\theta+1}, \qquad (3)$$

where $\theta = 4N_e\mu$ [23]. Gene diversity can be estimated from sequence data, making it possible to obtain estimates for N_e. We estimated H in individuals from the Yoruba in Ibadan (YRI) and Esan (ESN) populations available in the 1000 Genomes Project

Phase 3 data[7] (excluding offspring and individuals with ambiguous pedigree) by determining the number of heterozygous sites. The number of heterozygous sites was counted using PLINK 2.0[24] and divided by the size of the human genome, which was assumed to be 3.1×10^9 base pairs (size of GRCh38.p13, https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.39, accessed June 2020). Solving for θ and presuming a neutral mutation rate μ of 1.2×10^{-8} per base pair per generation yields an estimate of the distribution of the effective population size ($N_e \sim N(21975,163)$).

By sampling 1,000,000 replicates from the joint distribution of allele age and allele frequency and the distribution of the effective population size, the probability distribution of allele age is obtained in terms of generations. The distribution accounts for allele age, allele frequency, and effective population size as sources of uncertainty.

Because the *HOXB13* X285K variant was not observed in the Bantu-speaking population from Kampala, Uganda ($\widehat{p_u} = 0, n_U = 677$), we hypothesized that the variant must have arisen after the Bantu migration. Thus, to refine the estimated probability distribution of allele age, we consider the probability of the variant arising before the Bantu migration but not being sampled in the Ugandan cohort by chance by down weighting the likelihood of older ages preceding the Bantu migration. Under the neutral model, the frequency of the allele at the time of the Bantu migration, conditioned on observing it in the present, is given by:

$$p_B = \frac{t_1 - t_B}{t_1} \times \hat{p}, \qquad (4)$$

where p_B is the allele frequency at the time of the Bantu migration, t_1 is the age of the allele in years, t_B is the time of the Bantu migration in years ago, $\hat{p}$ is the present

frequency of the allele sampled from the distribution defined in Equation 1, and $p_B = 0$ for $t_1 < t_B$. Hence, the probability of not observing the variant in the Ugandan cohort by chance is:

$$P(\widehat{p_U} = 0) = (1 - p_B)^{2n_U a}, \quad (5)$$

where a is the proportion of Bantu ancestry. We presumed that the Bantu migration occurred 3,000 years ago and that the Bantu ancestry proportion is uniformly distributed on the interval from 50-75%[25]. By sampling 500,000 Bantu ancestry proportions and allele frequencies, confidence intervals for the probability of not observing the variant in the Ugandan cohort given an allele age t_1 are obtained (with $t_1 \geq t_B$).

The joint distribution of the allele age and the probability of not observing the variant in Uganda by chance if it arose prior to the Bantu migration yields the final probability distribution of the allele age with a mode of 2,290 years, a median of 3,035 years, a mean of 13,095, and 95% CI from 325 -79,115 years (assuming a generation time of 25 years). Code used for this analysis can be found here: https://github.com/AaronRuben/allele_age

Absolute Risk

Absolute risks of prostate cancer were estimated by rs77179853 carrier status using the odds ratios for carriers combined with mortality and incidence rates for African American men, while accounting for competing causes of death. Absolute risks by age *t* were calculated using age-specific prostate cancer incidence, $\mu(t)$, from the Surveillance, Epidemiology, and End Results (SEER) Program (1999-2013)[26] and age-specific mortality rates, $\mu_D(t)$, from the National Center for Health Statistics, CDC (1999-

2013)[27]. The approach constrains the risk category-specific absolute risks for a given age to be equivalent to the age-specific incidences for the entire population[28-31]. In other words, age-specific incidence rates are calculated to increase or decrease based on the estimated carrier risk and the proportion of the population within the carrier status. The calculation also accounts for competing causes of death.

Specifically, for a given carrier status *k*, the absolute risk by age *t* is computed as: $AR_k(t) = \sum_0^t P_{ND}(t)\, S_k(t) I_k(t)$. This calculation consists of three components:

(1) $P_{ND}(t)$ is the probability of not dying from another cause of death by age *t* using age-specific mortality rates, $\mu_D(t)$: $P_{ND}(t) = \exp[-\sum_0^t \mu_D(t-1)]$. Age-specific mortality rates are provided from a reference cohort.

(2) $S_k(t)$ is the probability of surviving prostate cancer by age *t* in the risk category *k* and uses the prostate cancer incidence by age *t* for category *k*: $S_k(t) = \exp[-\sum_0^t I_k(t-1)]$.

(3) The prostate cancer incidence by age *t* for risk category *k* is $I_k(t)$ and is calculated by multiplying the population prostate cancer incidence for the reference category, $I_0(t)$ and the corresponding risk ratio for category *k,* as estimated from the odds ratio obtained from the population-specific individual-level analysis: $I_k(t) = I_0(t)\exp(\beta_k)$.

To complete the calculations, the prostate cancer incidence for age *t* for the reference category, $I_0(t)$*,* is obtained by constraining the weighted average of the population cancer incidences for carriers to the population age-specific prostate cancer incidence, $\mu(t)$.

$I_0(t) = \mu(t)\frac{\sum_K f_k S_k(t-1)}{\sum_K f_k S_k(t-1)\exp(\beta_k)}$. f_k is the frequency of the risk category *k* with $f_k = 0.1$ for all non-reference categories.

By leveraging the definition that $S_k(t=0) = 1$, for all *k*, the absolute risks were

calculated iteratively by first getting $I_0(t=1)$, then $I_k(t=1)$, then $S_k(t=1)$ and finally $AR_k(t=1)$. Subsequent values were then calculated recursively for all t. Confidence intervals for absolute risk estimates were obtained via a parametric bootstrap repeating the above calculations for 1,000 bootstraps with the β_k's sampled from their corresponding estimated distributions using the standard error of the estimate.

References

[1] Conti DV, Wang K, Sheng X, Bensen JT, Hazelett DJ, Cook MB, et al. Two Novel Susceptibility Loci for Prostate Cancer in Men of African Ancestry. J Natl Cancer Inst. 2017;109.

[2] Han Y, Rand KA, Hazelett DJ, Ingles SA, Kittles RA, Strom SS, et al. Prostate Cancer Susceptibility in Men of African Ancestry at 8q24. J Natl Cancer Inst. 2016;108.

[3] Conti DV, Darst BF, Moss LC, Saunders EJ, Sheng X, Chou A, et al. Trans-ancestry genome-wide association meta-analysis of prostate cancer identifies new susceptibility loci and informs genetic risk prediction. Nat Genet. 2021;53:65-75.

[4] Harlemon M, Ajayi O, Kachambwa P, Kim MS, Simonti CN, Quiver MH, et al. A Custom Genotyping Array Reveals Population-Level Heterogeneity for the Genetic Risks of Prostate Cancer and Other Cancers in Africa. Cancer Res. 2020;80:2956-66.

[5] Al Olama AA, Kote-Jarai Z, Berndt SI, Conti DV, Schumacher F, Han Y, et al. A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. Nat Genet. 2014;46:1103-9.

[6] Amos CI, Dennis J, Wang Z, Byun J, Schumacher FR, Gayther SA, et al. The OncoArray Consortium: A Network for Understanding the Genetic Architecture of Common Cancers. Cancer Epidemiol Biomarkers Prev. 2017;26:126-35.

[7] Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526:68-74.

[8] Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature. 2021;590:290-9.

[9] Darst BF, Dadaev T, Saunders E, Sheng X, Wan P, Pooler L, et al. Germline sequencing DNA repair genes in 5,545 men with aggressive and non-aggressive prostate cancer. J Natl Cancer Inst. 2020.

[10] McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17:122.

[11] Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42:D980-5.

[12] Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26:2190-1.

[13] Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26:2867-73.

[14] Conomos MP, Miller MB, Thornton TA. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. Genet Epidemiol. 2015;39:276-93.

[15] Conomos MP, Reiner AP, Weir BS, Thornton TA. Model-free Estimation of Recent Genetic Relatedness. Am J Hum Genet. 2016;98:127-48.

[16] Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38:904-9.

[17] Alexander DH, Lange K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics. 2011;12:246.

[18] Maruyama T. The age of an allele in a finite population. Genet Res. 1974;23:137-43.

[19] Speidel L, Forest M, Shi S, Myers SR. A method for genome-wide genealogy estimation for thousands of samples. Nat Genet. 2019;51:1321-9.

[20] Loh PR, Danecek P, Palamara PF, Fuchsberger C, Y AR, H KF, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet. 2016;48:1443-8.

[21] Slatkin M, Rannala B. Estimating allele age. Annu Rev Genomics Hum Genet. 2000;1:225-49.

[22] Weir BS. Genetic Data Analysis II: Methods for Discrete Population Genetic Data: Sinauer Associates, Inc. Publishers; 1996.

[23] Charlesworth B, Charlesworth D. Elements of Evolutionary Genetics. Greenwood Village, Colorado: Roberts and Company Publishers; 2010.

[24] Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.

[25] Choudhury A, Aron S, Botigue LR, Sengupta D, Botha G, Bensellak T, et al. High-depth African genomes inform human migration and health. Nature. 2020;586:741-8.

[26] Surveillance Research Program, National Cancer Institute SEER*Stat software.

[27] Centers for Disease Control and Prevention, National Center for Health Statistics. Underlying Cause of Death 1999-2013 on CDC WONDER Online Database, released in 2019. Data are from the Multiple Cause of Death Files, 1999-2013, as compiled from

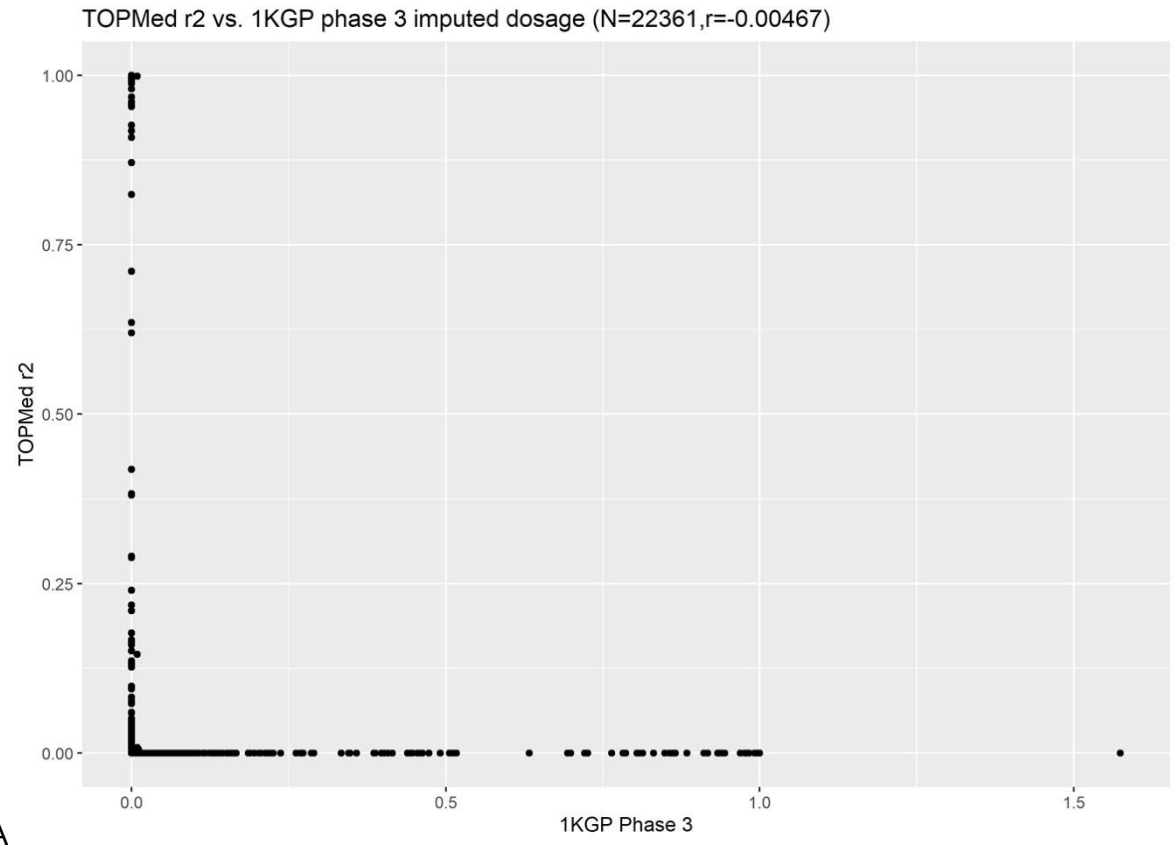
data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program.

[28] Antoniou AC, Beesley J, McGuffog L, Sinilnikova OM, Healey S, Neuhausen SL, et al. Common breast cancer susceptibility alleles and the risk of breast cancer for BRCA1 and BRCA2 mutation carriers: implications for risk prediction. Cancer Res. 2010;70:9742-54.

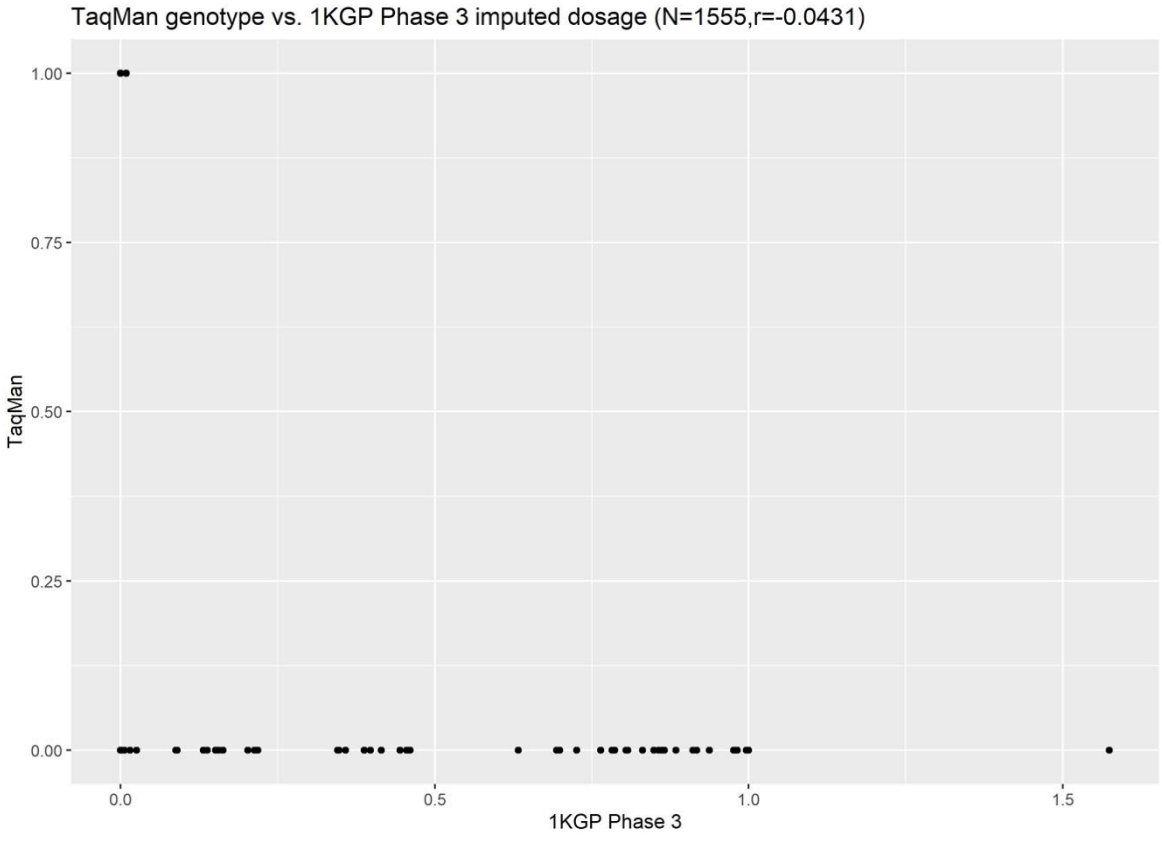
[29] Antoniou AC, Pharoah PD, McMullan G, Day NE, Ponder BA, Easton D. Evidence for further breast cancer susceptibility genes in addition to BRCA1 and BRCA2 in a population-based study. Genetic epidemiology. 2001;21:1-18.

[30] Amin Al Olama A, Benlloch S, Antoniou AC, Giles GG, Severi G, Neal DE, et al. Risk Analysis of Prostate Cancer in PRACTICAL, a Multinational Consortium, Using 25 Known Prostate Cancer Susceptibility Loci. Cancer Epidemiol Biomarkers Prev. 2015;24:1121-9.

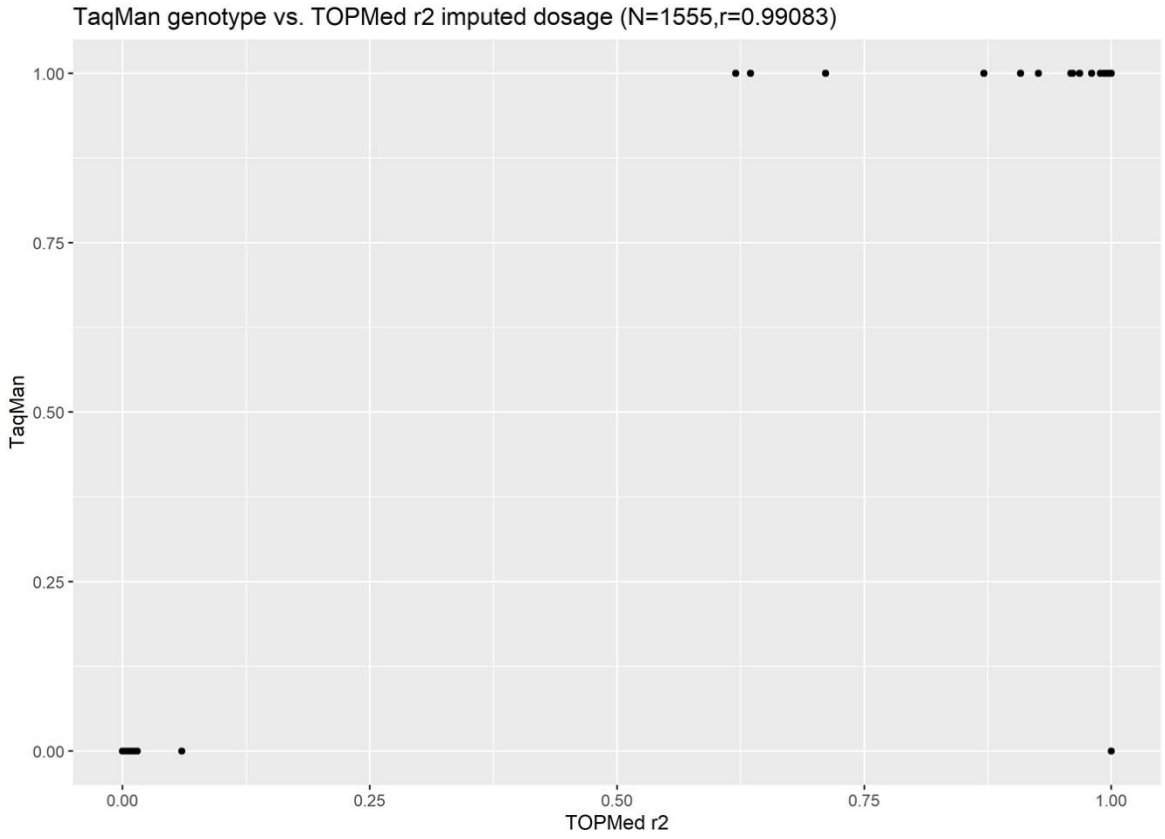
[31] Kuchenbaecker KB, McGuffog L, Barrowdale D, Lee A, Soucy P, Dennis J, et al. Evaluation of Polygenic Risk Scores for Breast and Ovarian Cancer Risk Prediction in BRCA1 and BRCA2 Mutation Carriers. J Natl Cancer Inst. 2017;109.


TOPMed r2 vs. 1KGP phase 3 imputed dosage (N=22361,r=-0.00467)
TOPMed r2
1.00
0.75
0.50
0.25
0.00
0.0
0.5
1.0
1.5
1KGP Phase 3

A

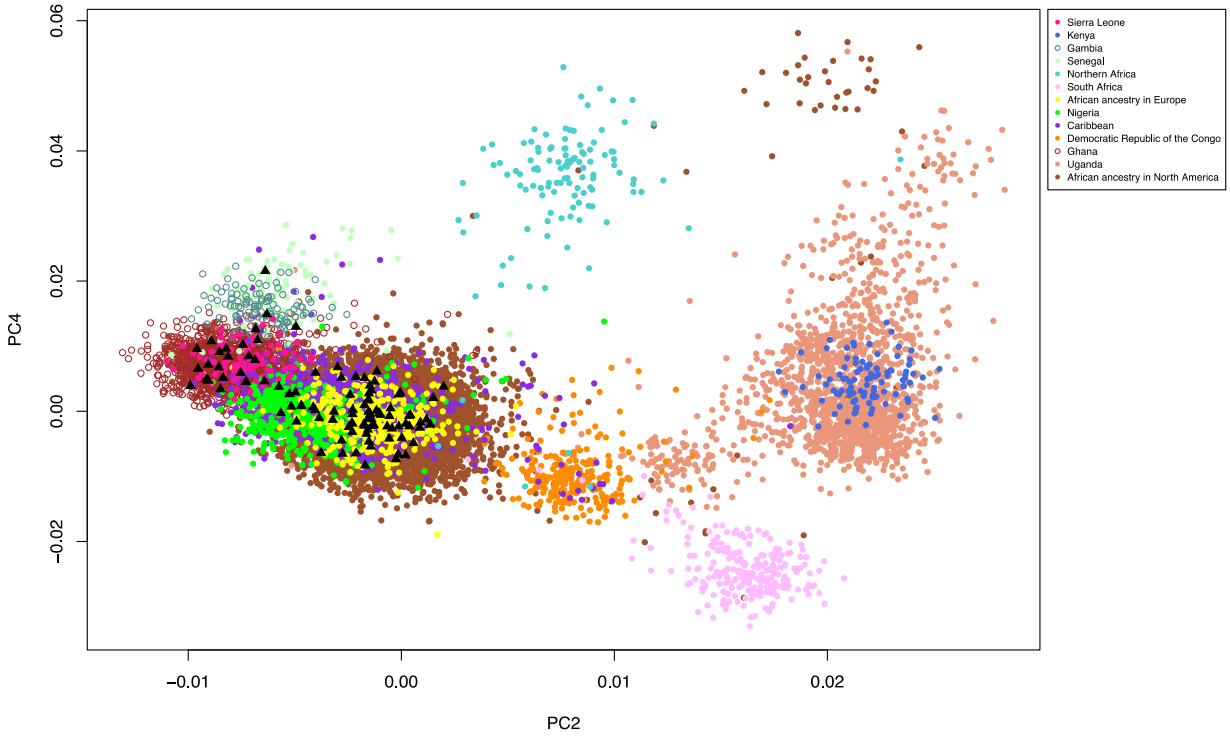

TaqMan genotype vs. 1KGP Phase 3 imputed dosage (N=1555,r=-0.0431)
TaqMan
1.00
0.75
0.50
0.25
0.00
0.0
0.5
1.0
1.5
1KGP Phase 3

B

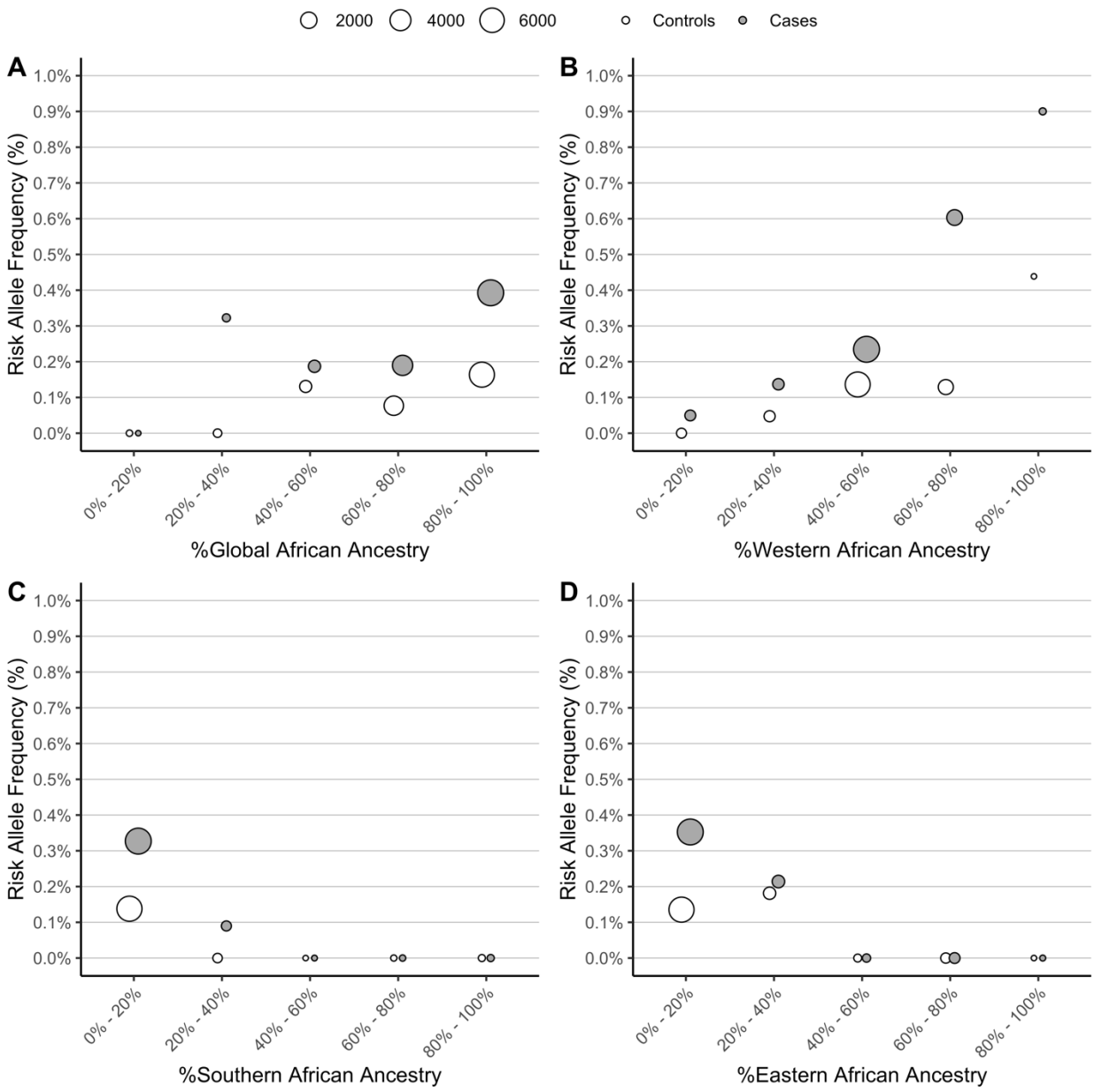


C

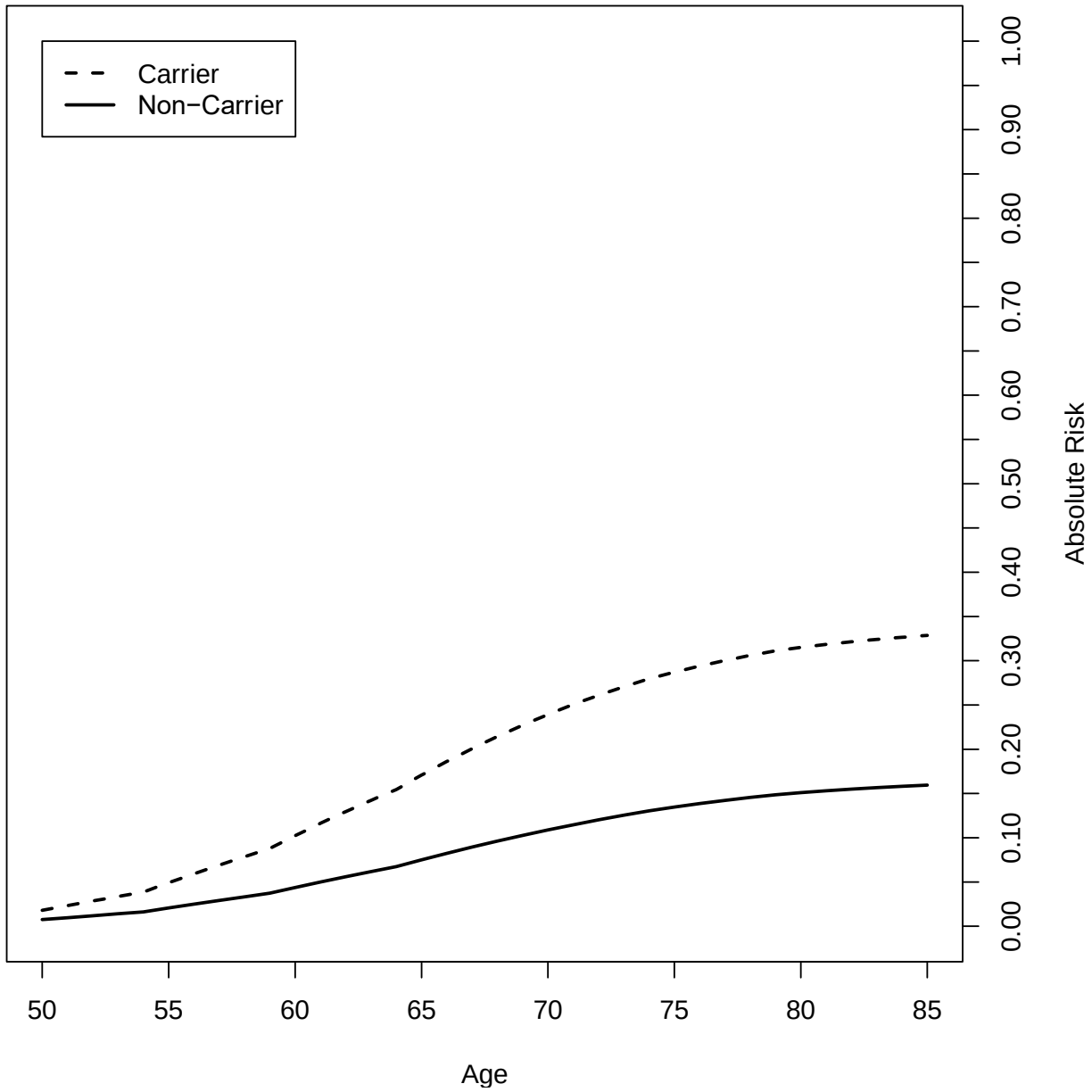
Supplemental Figure 1. Correlation between *HOXB13* rs77179853 imputed dosages and TaqMan genotyping. A) Correlation between rs77179853 imputation using 1000 Genomes Phase 3 and TOPMed r2 reference panels in 23,361 men; B) Correlation between rs77179853 imputation using 1000 Genomes Phase 3 reference panel and TaqMan genotyping in 1,555 men; C) Correlation between rs77179853 imputation using TOPMed r2 reference panel and TaqMan genotyping in 1,555 men.



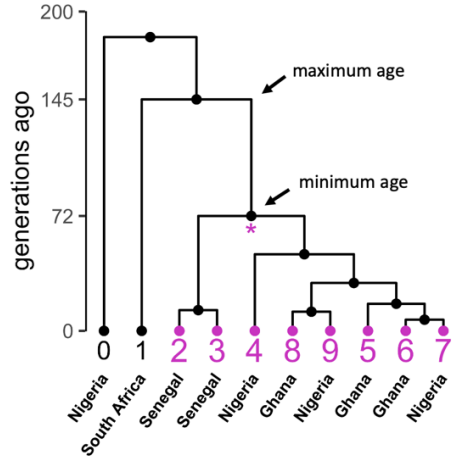
Supplemental Figure 2. Distribution of *HOXB13* rs77179853 by genetic ancestry comparing principal components 2 and 4 calculated in our sample of 22,361 African ancestry men. Men carrying the rs77179853 delA risk allele are highlighted by black triangles.



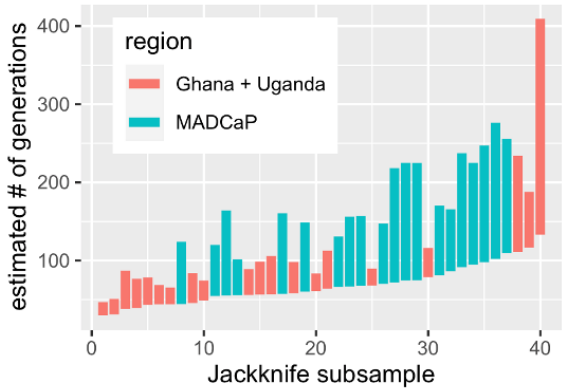
Supplemental Figure 3. Risk allele frequency of the *HOXB13* rs77179853 delA risk allele in 22,361 men by percentage of A) Global African ancestry (Global European ancestry=1-Global African ancestry), B) Western African ancestry, C) Southern African ancestry, and D) Eastern African ancestry. The size of the circle corresponds to sample size while color corresponds to prostate cancer status.



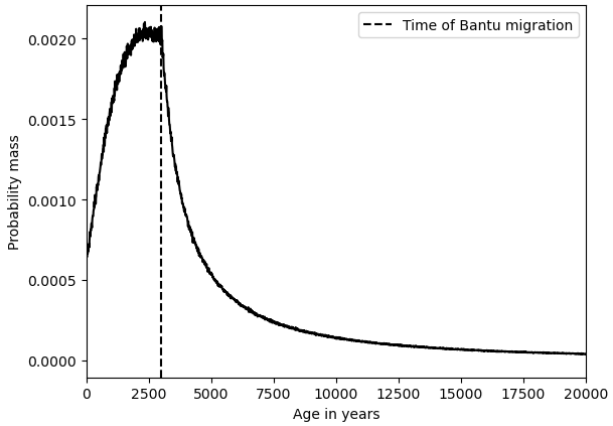
Supplemental Figure 4. Absolute risk of prostate cancer by *HOXB13* rs77179853 carrier status and age.



A



B



C

Method	Age Estimate	Ghana + Uganda	MADCaP
Genealogical estimates	**Minimum**	1,530 (837.5 - 2,222.5)	1,855 (1,417.5 - 2,292.5)
Genealogical estimates	**Maximum**	2,820 (815 - 4,825)	4,570 (3,327.5 - 5,812.5)
Method	**Age Estimate**	**Ghana + Nigeria**	
Derived Allele Frequencies	**Median**	3,035 (325 - 79,115)	

D

Supplemental Figure 5. Estimated allelic age of the *HOXB13* X285K (rs77179853) variant. A) Genealogic subtrees constructed using the Ghana + Uganda (N=1,760) cohort, estimating the local genealogy spanning the *HOXB13* X285K variant. Only subtrees of the haplotypes carrying the derived allele of rs77179853 (haplotypes 2-7) and two most closely related non-carrier haplotypes (haplotypes 0-1) are shown. This panel illustrates that conceptually, the minimum and maximum ages correspond to the recent and ancient time estimates of the branch leading to the most recent common ancestor node (marked by asterisks) of carriers of the derived allele for rs77179853. B) the minimum and maximum age estimates from each of the 40 jackknife samples used to estimate variability in allelic ages in the Ghana + Uganda (red) and MADCaP (blue) cohorts. C) The joint distribution of allele age and the probability of not observing the variant in Uganda by chance if it arose prior to the Bantu migration (indicated by the vertical dotted line) based on allele frequencies from N=1,071 Ghana and Nigeria samples. D) Allelic age estimates for each approach in years with 95% confidence intervals indicated. Age is estimated based on a generation time of 25 years.

Supplemental Table 1. Description and study design of the studies included.

Study Name	Study Abbreviation	Group	No. of Cases	No. of Controls	No. of Cases in analysis	No. of Controls in analysis	Design, location	Source of cases	Source of controls	Study Reference
Multiethnic Cohort, African Americans	MEC	AAPC GWAS	1841	1758	1766	1648	Case-control in cohort, HI and CA, U.S.	MEC	MEC	PMID: 10695593
Southern Community Cohort Study	SCCS	AAPC GWAS	263	523	250	513	Case-control in cohort, Southeastern U.S.	SCCS	SCCS	PMID: 16080667
The Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial	PLCO	AAPC GWAS	286	269	231	240	Case-control in screening trial, U.S.	PLCO	PLCO	PMID: 11189683
The Cancer Prevention Study II Nutrition Cohort	CPS-II	AAPC GWAS	76	152	64	112	Case-control in cohort, U.S.	CPS-II	CPS-II	PMID: 12015775
Prostate Cancer Case-Control Studies at MD Anderson	MDA	AAPC GWAS	543	474	527	437	Case-control, Houston, TX, U.S.	Houston Medical Center	Random-digit-dialing or hospital visitors	PMID: 15264247
Identifying Prostate Cancer Genes	IPCG	AAPC GWAS	368	172	353	157	Case-control, Maryland, U.S.	Johns Hopkins Hospital and Sidney Kimmel Cancer Center	Men undergoing screening for prostate cancer at the same institutions	PMID: 17401366
The Los Angeles Study of Aggressive Prostate Cancer	LAAPC	AAPC GWAS	296	303	286	285	Case-control, Los Angeles County, CA, U.S.	Los Angeles County Cancer Surveillance Program	Los Angeles County, neighborhood walk algorithm and the MEC	PMID: 20364112
Prostate Cancer Genetics Study	CaP Genes	AAPC GWAS	75	85	71	85	Case-control, Cleveland, OH, U.S.	Medical institutions in Cleveland, Ohio	Screened men at same medical institutions	PMID: 16931544
Case-Control Study of Prostate Cancer among African Americans in Washington, DC	DCPC	AAPC GWAS	292	359	263	339	Case-control, Washington, DC, U.S.	Howard University Hospital (HUH)	Men undergoing screening for prostate cancer at HUH	PMID: 19902474
King County (Washington) Prostate Cancer Studies	KCPCS	AAPC GWAS	145	81	141	75	Case-control, King County, WA, U.S.	Seattle-Puget Sound SEER cancer registry	Random-digit-dialing	PMID: 10548316
The Gene-Environment Interaction in Prostate Cancer Study	GECAP	AAPC GWAS	234	92	224	89	Case-control, Detroit, MI, U.S.	The Henry Ford Health System (HFHS)	HFHS population base	PMID: 17067754

North Carolina Prostate Cancer Study	NCPCS	AAPC GWAS	216	249	203	231	Case-control, NC, U.S.	North Carolina Central Cancer Registry	Friend referral, same county	PMID: 19117981
Selenium and Vitamin E Cancer Prevention Trial	SELECT	AAPC GWAS	223	224	212	208	Case-control in clinical trial, U.S.	SELECT	SELECT	PMID: 19066370
Prostate Cancer in a Black Population	PCBP	AAPC GWAS	238	231	231	223	Case-control, Barbados	All newly diagnosed cases in Barbados	Selected from a national database	PMID: 22402288
Vanderbilt Bio VU	BioVU	ELLIPSE/ OncoArray	213	0	204	0	Opt-out clinical biobank linked to de-identified electronic health records, Nashville, TN, USA	Patients who had an outpatient visit at VUMC with a blood draw ordered for clinical care who did not opt-out of the VUMC biobank (BioVU) and who were 18 years of age or older at the time his or her electronic health record was accessed for prostate cancer case status (in early 2014).	N/A (no matching controls)	PMID: 18500243 PMID: 23424142
Center for Prostate Disease Research	CPDR	ELLIPSE/ OncoArray	145	44	134	41	Retrospective cohort study; Greater Washington DC Metro Area, USA	Patients enrolled at Walter Reed National Military Medical Center with biopsy-confirmed prostate cancer who underwent radical prostatectomy	Patients enrolled at Walter Reed National Military Medical Center who had a negative DRE and PSA <2.0 ng/mL	PMID: 20056617
EPIdemiology of Prostate CAncer	EPICAP	ELLIPSE/ OncoArray	64	63	20	9	Case-control, France	North African origins living in the France Metropolitan, Cancer registry	Population-based	PMID: 24552491
Karuprostate	Karuprostate	ELLIPSE/ OncoArray	384	411	363	386	Population-based case-control in Guadeloupe and hospital-based case-control in DR Congo	Incident cases from Guadeloupe (Afro-Caribbean) and the DR Congo (African)	Free health screening program open to the general population (Guadeloupe); Men attending for prostate cancer screening or benign prostatic hyperplasia (DR Congo)	PMID: 20566993
Multiethnic Cohort Study	MEC	ELLIPSE/ OncoArray	489	529	462	499	Case-control in cohort, HI and CA, U.S.	MEC	MEC	PMID: 10695593
Moffitt Prostate Cancer Study	MOFFITT	ELLIPSE/ OncoArray	106	93	100	91	Case-control at Moffitt Cancer Center	Moffitt Cancer Center	Non-cancer visitors	PMID: 21802122
Nashville Men's Health Study	NMHS	ELLIPSE/ OncoArray	188	201	175	188	Case-control, Nashville, TN	Men seeking a prostate biopsy in all urology clinics in Nashville, TN	Men without PC at biopsy from these urology clinics.	PMID: 23079532

Prostate Cancer Prevention Trial	PCPT	ELLIPSE/ OncoArray	44	129	43	113	Case-control drawn from a randomized clinical trial; US and Canada	PCPT	PCPT	PMID: 12824459
The North Carolina-Louisiana Prostate Cancer Project	PCaP	ELLIPSE/ OncoArray	1022	0	958	0	Population-based Case-only	North Carolina Central Cancer Registry for NC cases and LSUHSC Cancer (SEER) Registry for LA cases	NA	PMID: 16676364
The Prostate Cancer and Environment Study	PROtEuS	ELLIPSE/ OncoArray	72	58	70	57	Case-control, Montreal, Canada	New incident cases across Montreal hospitals	Electoral list, from same residential areas as cases	PMID: 26385727
CerePP French Prostate Cancer Case-Control Study	ProGene	ELLIPSE/ OncoArray	107	105	101	85	Case-control, France	North Africa, Africa or Caribbean origins, living in France Metropolitan	Controls were recruited as participating in a systematic health screening program	PMID: 18264096
Southern Community Cohort Study	SCCS	ELLIPSE/ OncoArray	301	1557	286	1468	Case-control in cohort, Southeastern U.S.	SCCS	SCCS	PMID: 16080667
South Carolina Prostate Cancer Study	SCPCS	ELLIPSE/ OncoArray	64	39	57	32	Case-control, South Carolina, U.S.	South Carolina Central Cancer Registry	Health Care Financing Administration Medicare Beneficiary File	PMID: 15280622
Selenium and Vitamin E Cancer Prevention Trial	SELECT	ELLIPSE/ OncoArray	30	173	27	166	Case-control in clinical trial, U.S.	SELECT	SELECT	PMID: 19066370
San Francisco Prostate Cancer Study	SFPCS	ELLIPSE/ OncoArray	86	37	79	36	Case-control in Bay Area, CA	Non-Hispanic African-American men ages 40-79 years diagnosed with advanced prostate cancer from 1997-2000. Cases were identified through the Greater Bay Area Cancer Registry.	Non-Hispanic African-American men ages 40-79 years without a history of prostate cancer	PMID: 1595859]
A Case Control Study in Uganda	UGANDA	ELLIPSE/ OncoArray	571	485	560	480	Case-control in Kampala, Uganda	Incident cases from Mulago Hospital	Patients in other clinics at Mulago	PMID: 29356057
UK Prostate Cancer Study	UKGPCS	ELLIPSE/ OncoArray	375	0	365	0	Cases from the UK	Cases identified through clinics at the Royal Marsden hospital and nationwide NCRN hospitals	NA	http://www.icr.ac.uk/research/team_leaders/Eeles_Rosalind/Eeles_Rosalind_RES/index.shtml
San Antonio Biomarkers of Risk	SABOR	ELLIPSE/ OncoArray	106	106	103	105	Case-control from SA, TX	Incident and Prevalent cases from SABOR	SABOR	PMID: 20086112

Wake Forest Prostate Cancer Study	WFPCS	ELLIPSE/ OncoArray	59	66	54	47	Case-control, Winston-Salem, NC	Incident cases from Wake Forest Baptist Health Urology Clinic	Men with normal PSA/DRE from the same clinic	PMID: 15342424
Washington University Prostate Cancer Study	WUGS	ELLIPSE/ OncoArray	75	153	70	150	Case Control from St. Louis MO	Incident and Prevalent cases from Barnes Jewish Hospital	St. Louis MO	PMID: 21602798
California and Uganda Prostate Cancer Study	CA UG Study	H3	1,590	1,048	1,590	1,048	Case-control from Los Angeles, California and Kampala, Uganda	Cases from Los Angeles, CA through SEER registry and Incident cases from Mulago Hospital in Kampala, Uganda	Cancer-free controls were from the African American Eye Disease Study and patients in other clinics at Mulago	PMID: 29580111
Ghana Prostate Study	GPS	HumanOmni 5, Human Omni5Exome	642	636	640	634	Case-control, Greater Accra, Ghana	Patients from a local teaching hospital and cases identified from the population-based, probability sample that underwent screening for prostate cancer	Population-based, probability sample designed using the 2000 Ghana Population	PMID:24185611
The Men of African Descent and Carcinoma of the Prostate Consortium	MADCaP	MADCaP	397	401	397	401	Clinic-based case-control from 7 urban study sites in Senegal, Ghana, Nigeria, and South Africa	Incident cases diagnosed at one of 7 sub-Saharan African centers within 6 months before study contact were eligible	Controls were frequency matched to cases by age and center	PMID: 32393663

AAPC GWAS = African Ancestry Prostate Cancer Genome-Wide Association Study; ELLIPSE = Elucidating Loci Involved in Prostate Cancer Susceptibility; NA = not available; PMID = identifier number used in PubMed

Supplemental Table 2. Study participant characteristics.

Study	Group	#Cases	#Controls	Median Age (IQR) in Ca	Median Age (IQR) in Co	FH+/FH- in Ca, n (%)	FH+/FH- in Co, n (%)	Low Risk, n (%)	Intermediate Risk, n (%)	High Risk, n (%)	Lethal, n (%)	Median %Global AFR in Ca (IQR)	Median %Global AFR in Co (IQR)
MEC	AAPC GWAS	1766	1648	67 (12)	69 (11)	328 (19) / 1296 (73)	179 (11) / 1280 (78)	341 (19)	730 (41)	391 (22)	195 (11)	81 (17)	80 (20)
SCCS	AAPC GWAS	250	513	62 (9)	59 (11)	20 (8) / 203 (81)	32 (6.2) / 422 (82)	75 (30)	90 (36)	46 (18)	23 (9.2)	88 (9.8)	88 (9.4)
PLCO	AAPC GWAS	231	240	68 (9)	63 (8.2)	19 (8.2) / 203 (88)	24 (10) / 210 (88)	115 (50)	56 (24)	37 (16)	15 (6.5)	82 (13)	83 (17)
CPS-II	AAPC GWAS	64	112	70 (8)	70 (9)	5 (7.8) / 59 (92)	3 (2.7) / 109 (97)	24 (38)	15 (23)	8 (12)	7 (11)	77 (21)	74 (21)
MDA	AAPC GWAS	527	437	60 (12)	58 (14)	132 (25) / 384 (73)	62 (14) / 372 (85)	113 (21)	183 (35)	170 (32)	16 (3)	84 (13)	85 (13)
IPCG	AAPC GWAS	353	157	57 (10)	52 (22)	80 (23) / 206 (58)	3 (1.9) / 3 (1.9)	133 (38)	77 (22)	122 (35)	0 (0)	82 (14)	85 (12)
LAAPC	AAPC GWAS	286	285	63 (12)	64 (11)	63 (22) / 223 (78)	24 (8.4) / 243 (85)	132 (46)	0 (0)	114 (40)	22 (7.7)	82 (14)	80 (20)
CaP Genes	AAPC GWAS	71	85	67 (10)	66 (10)	16 (23) / 55 (77)	8 (9.4) / 77 (91)	0 (0)	35 (49)	15 (21)	3 (4.2)	82 (12)	86 (14)
DCPC	AAPC GWAS	263	339	64 (14)	58 (14)	33 (13) / 122 (46)	27 (8) / 140 (41)	44 (17)	9 (3.4)	23 (8.7)	24 (9.1)	86 (18)	86 (19)
KCPCS	AAPC GWAS	141	75	59 (10)	53 (9.5)	27 (19) / 114 (81)	8 (11) / 67 (89)	47 (33)	40 (28)	32 (23)	6 (4.3)	82 (16)	80 (14)
GECAP	AAPC GWAS	224	89	62 (11)	62 (11)	50 (22) / 162 (72)	15 (17) / 69 (78)	78 (35)	69 (31)	52 (23)	6 (2.7)	83 (12)	84 (13)
NCPCS	AAPC GWAS	203	231	61 (9)	52 (14)	61 (30) / 142 (70)	5 (2.2) / 16 (6.9)	30 (15)	36 (18)	19 (9.4)	0 (0)	84 (13)	85 (13)
SELECT	AAPC GWAS	212	208	64 (11)	64 (10)	60 (28) / 133 (63)	31 (15) / 161 (77)	109 (51)	47 (22)	14 (6.6)	3 (1.4)	84 (14)	80 (18)
PCBP	AAPC GWAS	231	223	66 (14)	66 (13)	27 (12) / 135 (58)	15 (6.7) / 140 (63)	0 (0)	0 (0)	0 (0)	11 (4.8)	91 (7.4)	91 (9.9)
BioVU	ELLIPSE/ OncoArray	204	0	61 (11)	--	0 (0) / 0 (0)	--	1 (0.49)	54 (26)	94 (46)	2 (0.98)	81 (17)	--
CPDR	ELLIPSE/ OncoArray	134	41	56 (11)	65 (2.9)	43 (32) / 69 (51)	4 (9.8) / 37 (90)	55 (41)	23 (17)	35 (26)	0 (0)	82 (14)	80 (18)
EPICAP	ELLIPSE/ OncoArray	20	9	65 (6.5)	62 (8)	6 (30) / 13 (65)	0 (0) / 8 (89)	0 (0)	0 (0)	1 (5)	3 (15)	21 (13)	18 (6.5)
Karuprostate	ELLIPSE/ OncoArray	363	386	67 (11)	60 (12)	140 (39) / 216 (60)	70 (18) / 309 (80)	110 (30)	127 (35)	109 (30)	0 (0)	94 (19)	93 (22)

MEC	ELLIPSE/ OncoArray	462	499	66 (12)	69 (10)	127 (27) / 295 (64)	38 (7.6) / 412 (83)	118 (26)	158 (34)	110 (24)	40 (8.7)	80 (19)	79 (21)
MOFFITT	ELLIPSE/ OncoArray	100	91	62 (11)	56 (10)	30 (30) / 70 (70)	6 (6.6) / 83 (91)	49 (49)	34 (34)	13 (13)	1 (1)	85 (15)	86 (10)
NMHS	ELLIPSE/ OncoArray	175	188	64 (12)	62 (10)	27 (15) / 148 (85)	32 (17) / 156 (83)	63 (36)	16 (9.1)	24 (14)	1 (0.57)	81 (14)	81 (14)
PCPT	ELLIPSE/ OncoArray	43	113	67 (7)	67 (7)	3 (7) / 40 (93)	17 (15) / 96 (85)	26 (60)	11 (26)	3 (7)	0 (0)	78 (20)	78 (18)
PCaP	ELLIPSE/ OncoArray	958	0	62 (12)	--	239 (25) / 719 (75)	--	446 (47)	242 (25)	94 (9.8)	72 (7.5)	84 (14)	--
PROtEuS	ELLIPSE/ OncoArray	70	57	63 (9)	64 (10)	7 (10) / 63 (90)	7 (12) / 50 (88)	19 (27)	11 (16)	10 (14)	1 (1.4)	92 (12)	92 (17)
ProGene	ELLIPSE/ OncoArray	101	85	63 (11)	62 (13)	13 (13) / 81 (80)	8 (9.4) / 77 (91)	38 (38)	33 (33)	27 (27)	3 (3)	75 (80)	27 (78)
SCCS	ELLIPSE/ OncoArray	286	1468	58 (11)	61 (14)	61 (21) / 202 (71)	95 (6.5) / 1276 (87)	25 (8.7)	90 (31)	47 (16)	16 (5.6)	86 (12)	86 (11)
SCPCS	ELLIPSE/ OncoArray	57	32	71 (7)	68 (5.2)	14 (25) / 43 (75)	6 (19) / 26 (81)	27 (47)	12 (21)	14 (25)	0 (0)	89 (15)	89 (10)
SELECT	ELLIPSE/ OncoArray	27	166	64 (11)	60 (12)	6 (22) / 21 (78)	23 (14) / 139 (84)	9 (33)	9 (33)	2 (7.4)	0 (0)	84 (14)	82 (16)
SFPCS	ELLIPSE/ OncoArray	79	36	62 (11)	62 (7.5)	21 (27) / 58 (73)	6 (17) / 30 (83)	0 (0)	0 (0)	63 (80)	16 (20)	83 (13)	82 (12)
UGANDA	ELLIPSE/ OncoArray	560	480	70 (13)	65 (10)	54 (9.6) / 351 (63)	11 (2.3) / 437 (91)	43 (7.7)	50 (8.9)	229 (41)	167 (30)	99 (3)	99 (4)
UKGPCS	ELLIPSE/ OncoArray	365	0	63 (11)	--	58 (16) / 212 (58)	--	93 (25)	59 (16)	80 (22)	16 (4.4)	94 (14)	--
SABOR	ELLIPSE/ OncoArray	103	105	63 (14)	64 (15)	0 (0) / 0 (0)	0 (0) / 0 (0)	32 (31)	18 (17)	14 (14)	0 (0)	83 (13)	83 (12)
WFPCS	ELLIPSE/ OncoArray	54	47	60 (11)	57 (13)	13 (24) / 40 (74)	2 (4.3) / 45 (96)	17 (31)	10 (19)	9 (17)	2 (3.7)	80 (15)	81 (13)
WUGS	ELLIPSE/ OncoArray	70	150	63 (14)	69 (5)	12 (17) / 57 (81)	15 (10) / 135 (90)	3 (4.3)	6 (8.6)	20 (29)	21 (30)	84 (9.7)	81 (15)
CA UG Study	H3	1590	1048	62 (13)	62 (15)	0 (0) / 0 (0)	0 (0) / 0 (0)	414 (26)	459 (29)	341 (21)	126 (8)	81 (17)	81 (17)
GPS	GPS	640	634	70 (11)	59 (11)	--	--	47 (7.3)	44 (6.9)	150 (23)	167 (26)	98 (0.8)	98 (0.8)
MADCaP	MADCaP	405	396	67 (10)	67 (10)	50 (12) / 170 (42)	24 (6) / 197 (50)	6 (1.5)	11 (2.7)	144 (36)	190 (47)	97 (2.6)	97 (2.9)

AAPC GWAS = African Ancestry Prostate Cancer Genome-Wide Association Study; ELLIPSE = Elucidating Loci Involved in Prostate Cancer Susceptibility; IQR=Interquartile range; FH+/FH-=Family history positive/negative; RAF=Risk allele frequency

Supplemental Table 3. Imputation quality scores for *HOXB13* rs77179853 across African ancestry studies.

	1KGP Phase 3			TOPMed r2		
Array	**Info**	**Control Freq**	**Case Freq**	**Info**	**Control Freq**	**Case Freq**
AAPC1M	0.819	0.24%	0.15%	0.921	0.13%	0.17%
ONCO-AAPC	0.748	0.20%	0.21%	0.918	0.11%	0.34%
H3 (CA UG)	0.684	0.15%	0.13%	0.949	0.15%	0.23%
HumanOmni (GPS)	0.753	0.16%	0.10%	0.967	0.49%	1.15%
MADCaP	0.819	0.25%	0.19%	0.941	0.12%	0.88%

Supplemental Table 4. Pathogenic/Likely Pathogenic/Deleterious *HOXB13* variants identified. The highlighted variant is the variant under study in this investigation.

					Imputation Info Scores				MAF	
Position	rsid	Consequence	Impact	ClinVar	AAPC	ONCO	Ghana	CA UG	Controls	Cases
48726791	rs77179853	frameshift_variant,stop_lost	HIGH	Uncertain significance	0.921	0.918	0.967	0.949	0.001	0.003
48726984	rs1351160874	frameshift_variant	HIGH	-	0.002	0.004	1E-05	6E-04	0	0
48727992	rs763590684	splice_donor_variant	HIGH	-	0	0.853	0	6E-04	5E-05	4E-05
48728006	rs771483373	frameshift_variant	HIGH	-	0.005	6E-04	9E-05	7E-04	0	0
48728241	rs1306259595	frameshift_variant	HIGH	-	0.021	4E-05	0	5E-05	0	0
48728267	rs749101324	stop_gained	HIGH	-	3E-04	2E-05	0	1E-05	0	0
48728343	rs138213197	missense_variant	MODERATE	pathogenic/likely pathogenic, risk factor, pathogenic, likely pathogenic	0.990	0.887	1E-04	0.951	4E-04	4E-04
48728383	rs762197066	stop_gained	HIGH	-	0.024	9E-04	1E-04	7E-05	0	0
48728491	rs1382962811	frameshift_variant	HIGH	-	0.053	0.002	4E-05	1E-04	0	0
48728584	rs931621182	frameshift_variant	HIGH	-	0.938	0.017	0.009	0.018	5E-05	0

MAF: Minor allele frequency

Supplemental Table 5. Risk allele frequency of *HOXB13* rs77179853 by African ancestry population. Study acronym is provided in parentheses or in the footnote.

African Population	Controls			Cases		
	n	n Carrier	RAF	n	n Carrier	RAF
West Africa						
Ghana	751	6	0.40%	752	18	1.20%
Greater Accra	634	6	0.47%	640	15	1.17%
Accra [1]	117	0	0%	112	3	1.39%
Nigeria	320	2	0.31%	112	2	0.89%
Esan (ESN) [3]	99	0	0%	--	--	--
Yoruba in Ibadan (YRI) [3]	108	1	0.50%	--	--	--
Ibadan [1]	56	1	0.89%	56	2	1.67%
Abuja [1]	57	0	0%	56	0	0%
Senegal (Dekar) [1]	59	0	0%	56	2	2.01%
Sierra Leone (Mende, MSL) [3]	85	0	0%	--	--	--
Gambia (Western Division, GWD) [3]	113	0	0%	--	--	--
Central, East, and South Africa						
Uganda (Kampala) [2]	677	0	0%	849	0	0%
Kenya (Luhya in Webuye, LWK) [3]	99	0	0%	--	--	--
Democratic Republic of the Congo (KARUPROSTATE)	127	0	0%	138	0	0%
South Africa [m]	114	0	0%	119	0	0%
Cape Town [1]	53	0	0%	58	0	0%
Johannesburg [1]	61	0	0%	61	0	0%
North America						
Canada (Montreal, PROtEuS)	57	0	0%	70	1	0.71%
United States[4]	7,428	20	0.13%	8,067	40	0.25%
Mid-Atlantic	537	0	0%	750	2	0.13%
Southern/Southeastern	2,570	10	0.19%	2,287	15	0.33%
South-Central	542	2	0.18%	630	5	0.40%
African Ancestry in Southwest US (ASW) [3]	61	0	0%	--	--	--
Western	3,394	8	0.12%	4,035	15	0.19%
Midwest	324	0	0%	365	3	0.41%
Caribbean Islands	578	3	0.26%	456	3	0.33%
Barbados (PCBP)	223	0	0%	231	1	0.22%
African Caribbean in Barbados (ACB) [3]	96	2	1.00%	--	--	--
Guadeloupe (KARUPROSTATE)	259	1	0.19%	225	2	0.44%
Europe						
United Kingdom	--	--	--	365	7	0.96%
France (France/Caribbean/N. African)	94	0	0%	121	0	0%

RAF: Risk allele frequency

[1]MADCaP Population

[2]Ugandans from Kampala are from a PSA-screened population, with all controls having PSA<4 ng/mL, which may contribute to the lower frequency in this population.

[3]Risk allele frequencies are based on 1000 Genomes[7].

[4]Studies included in the United States: WUGS, WFPCS, SABOR, SFPCS, SELECT, SCPCS, SCCS, PCPT, NMHS, MOFFITT, MEC, CPDR, SELECT, NCPCS, GECAP, KCPCS, DCPC, CaP Genes, LAAPC, IPCG, MDA, CPS-II, PLCO. Studies included in the Mid-Atlantic US: CPDR, DCPC, IPCG. Studies included in Southern/Southeastern US: WFPCS, SCPCS, SCCS, NMHS, MOFFITT, NCPCS. Studies included in South-Central US: SABOR, MDA. Studies included in Western US: SFPCS, MEC, NMPC, KCPCS, LAAPC. Studies included in Midwest US: WUGS, GECAP, CaP Genes.

Supplemental Table 6. Distribution of *HOXB13* rs77179853 delA carriers by study and disease aggressiveness.

Study	Population	#Cases	#Controls	#Low Risk	#Int Risk	#High Risk	#Case Carriers	#Control Carriers	#Low Risk Carriers	#Int Risk Carriers	#High Risk Carriers	#Unknown Carriers
MEC	Los Angeles, CA, USA	2227	2147	459	888	736	4	5	1	1	2	0
SCCS	Southeastern, USA	536	1981	100	180	132	3	9	0	2	1	0
MDA	Houston, TX, USA	527	437	113	183	186	4	2	0	0	4	0
LAAPC	Los Angeles, CA, USA	286	285	132	0	136	3	0	1	0	2	0
DCPC	Washington DC, USA	263	339	44	9	47	1	0	0	0	0	1
GECAP	Detroit, MI, USA	224	89	78	69	58	2	0	0	2	0	0
PCBP	Barbados (Caribbean)	231	223	0	0	11	1	0	0	0	0	1
BioVU	Nashville, TN, USA	204	0	1	54	96	2	0	0	0	1	1
CPDR	Washington, DC, USA	134	41	55	23	35	1	0	0	0	1	0
Karuprostate	DR Congo	138	127	5	49	77	0	0	0	0	0	0
Karuprostate	Guadeloupe (Caribbean)	225	259	105	78	32	2	1	2	0	0	0
MOFFITT	Tampa, FL, USA	100	91	49	34	14	1	1	0	1	0	0
NMHS	Nashville, TN, USA	175	188	63	16	25	1	0	1	0	0	0
PCaP	North Carolina and Louisiana, USA	958	0	446	242	166	8	0	3	1	3	1
PROtEuS	Montreal, Canada	70	57	19	11	11	1	0	1	0	0	1
SFPCS	Bay Area, CA, USA	79	36	0	0	79	1	0	0	0	1	0
UKGPCS	United Kingdom	365	0	93	59	96	7	0	1	0	3	3
SABOR	San Antonio, TX, USA	105	103	32	18	14	1	0	0	0	0	1

WUGS	St Louis, MO, USA	70	150	3	6	41	1	0	0	0	0	1
CA UG Study	Kampala, Uganda	289	197	5	4	182	0	0	0	0	0	0
CA UG Study	Los Angeles, CA, USA (NMPC)	1301	0	409	455	285	7	0	1	4	1	1
CA UG Study	Los Angeles, CA, USA (AFEDS)	0	851	0	0	0	0	3	0	0	0	0
GPS	Greater Accra, Ghana	640	634	47	44	317	15	6	1	1	9	4
MADCaP	Dakar, Senegal	56	59	1	2	48	2	0	0	0	0	2
MADCaP	Accra, Ghana	112	115	0	0	91	3	0	0	0	3	0
MADCaP	Ibadan, Nigeria	55	56	1	0	107	2	1	0	0	2	0
--	TOTAL	9370	8465	2260	2424	3022	73	28	12	12	33	17
--	Non-Carrier Studies	2318	2208	622	440	829	--	--	--	--	--	--

Low-risk disease: Gleason <7, stage T1/T2, and PSA<10 ng/ml; Intermediate-risk disease: Gleason=7, stage T1/T2, and PSA=10–20 ng/ml; High-risk disease: Gleason 8–10, stage T3/T4, PSA>20 ng/ml, or died of prostate cancer

Supplemental Table 7. Association between rs77179853 and age at prostate cancer diagnosis in studies where the variant was observed (10,477 cases).

Study	n Cases	n Case Carriers	Mean Age in Carriers	Mean Age in Non-Carriers	Beta (95% CI)	P value
AAPC1M	4,822	16	59.6	64.0	-3.1 (-7.1, 0.9)	0.13
ONCO-AAPC	3,434	28	62.5	62.5	0.6 (-2.4, 3.7)	0.7
CA UG Study	1,301	7	62.0	60.9	1.3 (-4.6, 7.2)	0.7
GPS	640	15	67.3	69.7	-2.3 (-6.7, 2.2)	0.3
MADCaP	280	7	69.0	68.2	0.2 (-6.4, 6.9)	0.9
Meta-analysis	10,477	73	63.5	63.6	-0.7 (-2.7, 1.2)	0.5

Supplemental Table 8. Frequency of rs77179853 by age at prostate cancer diagnosis in studies where the variant was observed (10,477 cases).

Age Category	n Cases	n Case Carriers	RAF (95% CI)
≤50	687	6	0.44% (0.09%-0.79%)
51-55	1,249	9	0.36% (0.13%-0.60%)
56-60	1,947	11	0.28% (0.12%-0.45%)
61-65	2,178	16	0.37% (0.19%-0.55%)
65-70	2,108	16	0.38% (0.19%-0.57%)
71-75	1,396	5	0.18% (0.02%-0.34%)
76-80	645	7	0.54% (0.14%-0.94%)
>80	267	3	0.56% (0%-1.20%)

RAF: Risk allele frequency

Supplemental Table 9. Association between rs77179853 and family history of prostate cancer among cases in studies where the variant was observed.

	Family History Positive			Family History Negative			OR (95% CI)	P value
Study	n Cases	Carriers	Carrier Frequency	n Cases	Carriers	Carrier Frequency		
AAPC1M	921	4	0.43%	3,438	11	0.32%	1.04 (0.85-1.28)	0.7
ONCO-AAPC	758	6	0.79%	2,179	15	0.69%	1.04 (0.86-1.25)	0.7
MADCaP	35	0	0%	114	4	3.51%	0.73 (0.46-1.16)	0.18
Overall	1,714	10	0.58%	5,731	30	0.52%	1.01 (0.88-1.15)	0.9

Supplemental Table 10. Association between rs77179853 and PSA in prostate cancer cases and controls in studies where the variant was observed.

	Study	n	Carriers	Mean PSA in Carriers (ng/ml)	Mean PSA in Non-Carriers (ng/ml)	Beta (95% CI)	P value
Cases	AAPC1M	2,273	7	33.7	21.0	-1.3 (-69.9, 67.2)	>0.9
	ONCO-AAPC	1,313	10	8.0	30.2	-8.5 (-115, 98.5)	0.9
	CA UG Study	1,227	6	6.0	11.5	-5.4 (-18.3, 7.4)	0.4
	GPS	607	15	124	313	-174 (-754, 405)	0.6
	MADCaP	276	7	823	1,266	-665 (-4479, 3149)	0.7
	Meta-analysis	5,696	45	177	111	-5.4 (-17.9, 7.1)	0.4
Controls	AAPC1M	1,945	2	1.2	2.5	-0.4 (-10.4, 9.7)	0.9
	ONCO-AAPC	344	0	--	5.7	--	--
	GPS	634	6	1.3	2.0	-0.7 (-4.5, 3.0)	0.7
	MADCaP	36	0	--	2.8	--	--
	Meta-analysis	2,959	8	1.3	2.8	-0.7 (-4.2, 2.8)	0.7

PSA: Prostate-specific antigen

Supplemental Note

This work was support by NIH grants U19CA214253, R01CA165862, U19 CA148537, and K99CA246063. This work is also supported by the Achievement Rewards for College Scientists Foundation Los Angeles Founder Chapter

The AAPC studies were supported as follows:

MEC: The MEC and the genotyping in this study were supported by NIH grants CA63464, CA54281, CA1326792, CA148085 and HG004726. Cancer incidence data for the MEC and LAAPC studies have been collected by the Los Angeles Cancer Surveillance Program of the University of Southern California with Federal funds from the NCI, NIH, Department of Health and Human Services, under Contract No. N01-PC-35139, and the California Department of Health Services as part of the state-wide cancer reporting program mandated by California Health and Safety Code Section 103885, and grant number 1U58DP000807-3 from the Centers for Disease Control and Prevention.

PLCO: Genotyping of the **PLCO** samples was funded by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics, NCI, NIH. The authors thank Drs. Christine Berg and Philip Prorok, Division of Cancer Prevention at the National Cancer Institute, the screening center investigators and staff of the PLCO Cancer Screening Trial for their contributions to the PLCO Cancer Screening Trial. We thank Mr. Thomas Riley, Mr. Craig Williams, Mr. Matthew Moore, and Ms. Shannon Merkle at Information Management Services, Inc., for their management of the data and

Ms. Barbara O'Brien and staff at Westat, Inc. for their contributions to the PLCO Cancer Screening Trial. We also thank the PLCO study participants for their contributions to making this study possible.

MDA: MDA was support by grants, CA68578, ES007784, DAMD W81XWH-07-1-0645, and CA140388.

LAAPC was funded by grant 99-00524V-10258 from the Cancer Research Fund, under Interagency Agreement #97-12013 (University of California contract #98-00924V) with the Department of Health Services Cancer Research Program. Cancer incidence data for the MEC and LAAPC studies have been collected by the Los Angeles Cancer Surveillance Program of the University of Southern California with Federal funds from the NCI, NIH, Department of Health and Human Services, under Contract No. N01-PC-35139, and the California Department of Health Services as part of the state-wide cancer reporting program mandated by California Health and Safety Code Section 103885, and grant number 1U58DP000807-3 from the Centers for Disease Control and Prevention

KCPCS was supported by NIH grants CA056678, CA082664 and CA092579, with additional support from the Fred Hutchinson Cancer Research Center. We thank the participants in these studies, and Ms. Suzanne Kolb for help with study management.

GECAP was supported by NIH grant ES011126.

CaP Genes was supported by CA88164 and CA127298.

IPCG was supported by the generous support from donors to The Patrick C. Walsh Hereditary Prostate Cancer Research Program at The Brady Urological Institute.

DCPC was supported by NIH grant S06GM08016 and DOD grants DAMD W81XWH-07-1-0203, DAMD W81XWH-06-1-0066 and DOD W81XWH-10-1-0532.

CPS-II is supported by the American Cancer Society.

PCBP: PCBP was supported by NHGRI contract N01HG25487 and NCI grant R01CA114379

SCPCS: SCPCS is funded by CDC grant S1135-19/19, and SCPCS sample preparation was conducted at the Epidemiology Biospecimen Core Lab that is supported in part by the Vanderbilt-Ingram Cancer Center (P30 CA68485).

SELECT is funded by Public Health Service cooperative Agreement grant CA37429 awarded by the National Cancer Institute as well as 5UM1CA182883 from the Office of Dietary Supplements at the National Institutes of Health. The authors thank the site investigators and staff and, most importantly, the participants from SELECT who donated their time to this trial.

SCCS is funded by NIH grant CA092447. SCCS sample preparation was conducted at the Epidemiology Biospecimen Core Lab that is supported in part by the Vanderbilt Ingram Cancer Center (CA68485). Data on SCCS cancer cases used in this publication were provided by the Alabama Statewide Cancer Registry; Kentucky Cancer Registry; Tennessee Department of Health, Office of Cancer Surveillance; Florida Cancer Data System; North Carolina Central Cancer Registry, North Carolina Division of Public Health; Georgia Comprehensive Cancer Registry; Louisiana Tumor Registry; Mississippi Cancer Registry; South Carolina Central Cancer Registry; Virginia Department of Health, Virginia Cancer Registry; Arkansas Department of Health, Cancer Registry. The Arkansas Central Cancer Registry is fully funded by a grant from

National Program of Cancer Registries, Centers for Disease Control and Prevention (CDC). Data on SCCS cancer cases from Mississippi were collected by the Mississippi Cancer Registry which participates in the National Program of Cancer Registries (NPCR) of the Centers for Disease Control and Prevention (CDC). The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the CDC or the Mississippi Cancer Registry.

The ELLIPSE/PRACTICAL OncoArray studies in men of African ancestry were supported as follows:

CeRePP: The authors thank Cecile Gaffory for technical assistance.

Karuprostate: The Karuprostate study was supported by the French National Health Directorate and by the Association pour la Recherche sur les Tumeurs de la Prostate. Séverine Ferdinand, Marc Romana.

MEC: The MEC was support by NIH grants CA63464, CA54281 and CA098758.

MIAMI (WFPCS): WFPCS was supported by a grant from the American Cancer Society (No. CNE-101119 to J.J.H.), a pilot grant from the Comprehensive Cancer Center of Wake Forest University (CA12197 to J.J.H.) and a grant from the National Research Foundation to the Wake Forest University's General Clinical Research Center (M01-RR07122). The authors are grateful to study participants. We also want to acknowledge the contributions of Frank M.Torti, MD; Robert Lee, MD; Charles J.Rosser, MD; Dean G.Assimos, MD; Elizabeth Albertson, MD; Dominck J.Carbone, MD; William Rice, MD;

Francis O'Brien, MD; Ray Morrow, MD; Franklyn Millman, MD; Nadine Shelton, Joel Anderson, Shirley Cothren, Eunkyung Chang, the General Clinical Research Center, the Urology Clinic and the Internal Medicine Clinic.

MOFFITT: The Moffitt group was supported by the US National Cancer Institute (R01CA128813, PI: J.Y. Park).

NMHS: Funding for the Nashville Men's Health Study (NMHS) was provided by the National Institutes of Health Grant numbers: RO1CA121060

PLCO: Genotyping of the PLCO samples was funded by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics, NCI, NIH. The authors thank Drs. Christine Berg and Philip Prorok, Division of Cancer Prevention at the National Cancer Institute, the screening center investigators and staff of the PLCO Cancer Screening Trial for their contributions to the PLCO Cancer Screening Trial. We thank Mr. Thomas Riley, Mr. Craig Williams, Mr. Matthew Moore, and Ms. Shannon Merkle at Information Management Services, Inc., for their management of the data and Ms. Barbara O'Brien and staff at Westat, Inc. for their contributions to the PLCO Cancer Screening Trial. We also thank the PLCO study participants for their contributions to making this study possible.

SFPCS: SFPCS was funded by grant 99-00527V-10182 from the California Cancer Research Fund under Interagency Agreement #97-12013 (University of California contract #98-00924V) with the Department of Health Services Cancer Research Program. Cancer incidence data used in this study have been collected by the Greater Bay Area Cancer Registry of the Cancer Prevention Institute of California under contract N01-PC-35136 with the National Cancer Institute and with support of the California

Cancer Registry, a project of the Cancer Surveillance Section, California Department of Health and Human Services, under subcontract 1006128 with the Public Health Institute.

PCPT: PCPT is funded by Public Health Service grants U10CA37429 and 5UM1CA182883 from the National Cancer Institute. The authors thank the site investigators and staff and, most importantly, the participants from PCPT who donated their time to this trial.

SELECT: SELECT is funded by Public Health Service Cooperative Agreement grant CA37429 awarded by the National Cancer Institute, National Institutes of Health. The authors thank the site investigators and staff and, most importantly, the participants from SELECT who donated their time to this trial.

WUGS/WUPCS: WUGS would like to thank the following for funding support: The Anthony DeNovi Fund, the Donald C. McGraw Foundation, and the St. Louis Men's Group Against Cancer.

UGANDA: The UGANDA study was support by R01CA165862.

SABOR: The SABOR research is supported by NIH/NCI Early Detection Research Network, grant U01 CA0866402-12. Also supported by the Cancer Center Support Grant to the Cancer Therapy and Research Center from the National Cancer Institute (US) P30 CA054174.

PCaP: The North Carolina - Louisiana Prostate Cancer Project (PCaP) is carried out as a collaborative study supported by the Department of Defense contract DAMD 17-03-2-0052. The authors thank the staff, advisory committees and research subjects participating in the PCaP study for their important contributions. We would also like to

acknowledge the UNC BioSpecimen Facility and the LSUHSC Pathology Lab for our DNA extractions, blood processing, storage and sample disbursement (https://genome.unc.edu/bsp).

EPICAP: The EPICAP study was supported by grants from Ligue Nationale Contre le Cancer, Ligue départementale du Val de Marne; Fondation de France; Agence Nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail (ANSES). The EPICAP study group would like to thank all urologists, Antoinette Anger and Hasina Randrianasolo (study monitors), Anne-Laure Astolfi, Coline Bernard, Oriane Noyer, Marie-Hélène De Campo, Sandrine Margaroline, Louise N'Diaye, Sabine Perrier-Bonnet (Clinical Research nurses).

BioVU: The dataset(s) used for the analyses described were obtained from Vanderbilt University Medical Center's BioVU which is supported by institutional funding and by the National Center for Research Resources, Grant UL1 RR024975-01 (which is now at the National Center for Advancing Translational Sciences, Grant 2 UL1 TR000445-06).

PROtEuS: PROtEuS was supported financially through grants from the Canadian Cancer Society [13149, 19500, 19864, 19865] and the Cancer Research Society, in partnership with the Ministère de l'enseignement supérieur, de la recherche, de la science et de la technologie du Québec, and the Fonds de la recherche du Québec - Santé. PROtEuS would like to thank its collaborators, research personnel, and urologists. We also wish to acknowledge the special contribution made by Drs. Ann Hsing and Anand Chokkalingam to the conception of the genetic component of the study.

CPDR: This research was supported by funds from the Center for Prostate Disease Research, Uniformed Services University Program, HU0001-10-2-0002 (PI: COL Inger L. Rosner, MD).

UKGPCS would also like to thank the following for funding support: The Institute of Cancer Research and The Everyman Campaign, The Prostate Cancer Research Foundation, Prostate Research Campaign UK (now Prostate Action), The Orchid Cancer Appeal, The National Cancer Research Network UK, The National Cancer Research Institute (NCRI) UK. We are grateful for support of NIHR funding to the NIHR Biomedical Research Centre at The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust. UKGPCS should also like to acknowledge the NCRN nurses, data managers and Consultants for their work in the UKGPCS study.

Other studies were supported as follows:

CA UG Study: The African American Eye Disease Study was supported by a National Institutes of Health Grant U10 (EY023575) and an unrestricted grant from the Research to Prevent Blindness, New York, New York. The UGANDA study was support by R01CA165862.

Ghana Prostate Study (GPS): The Ghana Prostate Study was funded by the Intramural Program of the National Cancer Institute, National Institutes of Health, Department of Health and Human Services including Contract No. HHSN261200800001E.

MADCaP: We thank all MADCaP study participants. This work is a product of the MADCaP network (https://www.madcapnetwork.org/). This work was supported by a large multisite NIH/NCI grant to Timothy Rebbeck (U01CA184374) and an NIGMS MIRA grant to Joseph Lachance (R35GM133727). Additional funding for this work includes a seed grant from the Integrated Cancer Research Center at Georgia Institute of Technology.