Deciphering the Polygenic Basis of Racial Disparities in Prostate Cancer By an Integrative Analysis of Genomic and Transcriptomic Data

Wensheng Zhang[1], Thea Nicholson[2], and Kun Zhang[1,3]

ABSTRACT

Prostate cancer prevalence in African Americans (AA) is over 1.5 times the prevalence in European Americans (EA). Among over a hundred index risk SNPs for prostate cancer, only a few can be verified using the available AAs' data. Their relevance to the prevalence inequality and other racial disparities has not been fully determined. We investigated this issue by an integrative analysis of five public datasets. We categorized the datasets into two classes. The training class consisted of the datasets generated by three genome-wide association studies. The test class contained the prostate adenocarcinoma data of The Cancer Genome Atlas and the data of African and European super-populations in the 1000-Genome project. The polygenic risk scores (PRS) of test samples for cancer occurrence were calculated according to the effects of genetic variants estimated from the training samples. We obtained the following findings. Africans' PRSs are higher than Europeans' scores ($P < 1 \times 10^{-6}$). AA patients' PRSs are higher than EA patients' scores ($P < 3 \times 10^{-9}$). The patients with tumors presenting fusion or abnormal expression in *ERG* and other E26 transformation-specific (ETS) family genes have lower PRSs than the patients without such aberrations ($P < 7 \times 10^{-5}$). Five tumor progression-related genes have the expression levels being significantly correlated with PRS (FDR < 0.01). Additional simulation analysis shows that the high prostate cancer prevalence in African populations makes it challenging to identify individual risk variants using African men's data. These results implicate that the index risk SNP-based PRS is compatible with the observed racial disparity in prostate cancer prevalence and ETS abnormal cancers may be less heritable compared with other subtypes.

Prevention Relevance: This study reveals the relevance of index risk SNP markers with racial disparities in prostate cancer. The findings also indicate that PRS can be used in prostate cancer subtype prediction.

Introduction

Prostate cancer, in which adenocarcinomas amount to 95% of the cases, is the most commonly diagnosed nonskin cancer and the second leading cause of cancer mortality in American men (1, 2). Many racial disparities in this cancer type were reported in the past years (3). For example, the prevalence in African Americans (AA) is over 1.5 times the prevalence in European Americans (EA; ref. 4), the mortality rate of AA patients is nearly two times of the rate of EA patients (5), and the portion of tumors with *TMPRSS2-ERG* fusion in EAs is about two times of the fraction in AAs (6). The etiology of these disparities could be multifactorial. While the contribution of environmental factors, such as diets, lifestyles, and socioeconomic issues, is still elusive (3), the roles of genetic factors are clearly suggested by the familial occurrence of the cancer (7–9). Therefore, understanding the underlying genetic basis is an essential step for eliminating racial disparities in the screening, diagnosis, and treatment of prostate cancer. This issue has attracted widespread interest from genetics, biomedical, and public health communities (10–12).

Previous studies show that prostate cancer cases attributed to the germline aberration of a major cancer gene such as *BRCA* is relatively rare (13). This implies that prostate cancer susceptibility could be largely considered as a complex trait determined by many minor genetic factors. While DNA alterations of other types, such as short sequence repeats and indels (14), have been reported to be associated with prostate cancer occurrence, former research was mostly focused on single nucleotide polymorphisms (SNPs). By 2015, 104 risk SNP markers (this marker set is used in our analysis. See Material and Methods section for an explanation) for prostate cancer susceptibility, aggressiveness, and mortality had been identified by genome-wide association studies (GWAS) on single cohort data (15). Recently, a transancestry genome-wide association meta-analysis brought the number loci for prostate cancer

[1]Bioinformatics Core of Xavier NIH RCMI Center of Cancer Research, Xavier University of Louisiana, New Orleans, LA 70125, USA. [2]Dual Degree Biomedical Engineering Program, Department of Biology, Xavier University of Louisiana, New Orleans, LA 70125, USA. [3]Department of Computer Science, Xavier University of Louisiana, New Orleans, LA 70125, USA.

Note: Supplementary data for this article are available at Cancer Prevention Research Online (http://cancerprevres.aacrjournals.org/).

Corresponding Author: Kun Zhang, Bioinformatics Core of Xavier NIH RCMI Center of Cancer Research, Xavier University of Louisiana, 1 Drexel Drive, New Orleans, LA 70125, USA. Phone: 504-520-6700; E-mail: kzhang@xula.edu

Cancer Prev Res 2022;15:161–72

doi: 10.1158/1940-6207.CAPR-21-0406

susceptibility to 269 (16). Among them, ten are located in the genomic region *8q24*, which also host several risk loci for other cancer types, such as breast cancer, and the well-known oncogene *MYC* (17, 18). For a risk marker, any SNP that is physically adjacent or close to it and is in strong linkage disequilibrium (LD) with it can save as its proxy risk marker. The index risk SNPs, i.e., risk SNP markers, can explain one third of familial prostate cancer risk and 5% to 10% susceptibility variance of the disease in a liability scale (19–21). Accordingly, multiple SNP markers-based polygenic risk score (PRS) was conceived as a useful metric for prostate cancer susceptibility and have found widespread applications.

PRS is defined as the sum of the effects of test alleles (usually specified with minor alleles) over SNP markers on the focused trait (22), which is prostate cancer occurrence in this study. A higher PRS mean indicates a greater genetic predisposition for the disease. The multiplicative model is usually used to estimate allele effects in case-control studies. It works on singular SNP and calculates the log odds (LOOD) value as the effect measure (23). A problem in the application is that a SNP may be overrepresented in the resulting PRS when it is in strong or moderate LD with other markers. An alternative model is the logistic multivariate regression model, which includes the dosage of test alleles of all considered SNPs as independent variables (24). While the collinearity due to LD may impact the estimation of the significance levels of effects, the overrepresentation of individual SNPs can be largely alleviated in the resulting PRS. Another method is the genomic relationship matrix-based linear mixed model (LMM), in which allele effects are assumed to be random over SNPs and fixed across samples (25, 26). Theoretically, the number of SNPs considered in LMM is not limited and even can be larger than the sample size. But in a practical application, the result could be subject to severe overfitting and the substantial noise introduced by the numerous independent variables.

Because most risk SNP markers for prostate cancer were initially identified in the EA and other European populations, and only a few (less than ten) of them showed a sufficiently-high statistical significance in the available AA data (27–30), their relevance, as a whole, to racial disparities has yet to be determined. With the hypothesis that a high-risk population could have a high PRS, we investigated this issue by an integrative analysis of a few large-scale representative public datasets. In the relevant analysis, we further assumed that the genetic architecture for prostate cancer predisposition is not distinguished between sexes, while the disease only occurs in males. Especially, before predicting the PRSs of test samples, we employed the aforementioned LLM and other two methods to estimate the effects of susceptibility SNPs. This is different from a recent publication that directly used the LOOD value of the risk allele of a locus for prostate cancer occurrence reported by the original GWAS study in the prediction (31). Moreover, by taking advantage of the comprehensive genomic and clinical data generated by The Cancer Genome Atlas (TCGA), we extend the analysis from the genetic basis of the disparity in occurrence to the relevance with somatic mutation-based cancer subtypes and gene expression traits, offering unique insight into racial disparities in prostate cancer.

Materials and Methods

Data

The used data are categorized into two classes (**Table 1**). The training class consists of the datasets from three genome-wide association studies, and the test class contains the datasets of TCGA prostate carcinoma samples and the datasets of the African (AFR) and European (EUR) samples of the 1000-Genome project (32).

GWAS data

The three datasets were generated by The Cancer Genetic Markers of Susceptibility (CGEMS; refs. 18, 34), The Breast and Prostate Cancer Cohort Consortium (BPC3; ref 33), and Multiethnic Cohort Study (MEC; ref. 19). These projects/experiments adopted the nested case-control design. Of MEC data, except for special notes, only AA samples (cohort) are considered and discussed in this study. A more detailed description of these datasets is presented in Supplementary Materials and Methods 1.

TCGA data

The TCGA prostate project includes over 550 patients with prostate adenocarcinomas. Among them, 333 subjects were analyzed in detail by the TCGA's paper published in the

Table 1. Summary of the used datasets.

Dataset category	Dataset[a]	No. of case samples	No. of control samples	Genotype platform	No. of SNPs (million)
Training class	CGEMS (EUR)	1,147	1,098	Illumina HumanHap300v1.1 & -250Sv1.0	0.56
	BPC3 (EA)	2,758	4,482	Illumina Human660W-Quad_v1_A	0.58
	MEC (AA)	2,306	2,463	Illumina Human1M-Duov3_B	1.15
Test class	TCGA (mixed)	496		Affymetrix Genome-Wide Human SNP 6.0	0.91
	1000G (AFR)	661		Illumina sequencing	84.4
	1000G (EUR)	503		Illumina sequencing	84.4

Abbreviation: 1000G, 1000 Genomes project.
[a]In the parentheses are the racial population or super-population codes of samples.

Cell journal (35). Since the subtype identification (mainly based on somatic mutation profiles) and the resulting sample stratification presented in the article are well recognized by the biomedical community, we used the data of the 333 patients in this study. The SNP and gene expression data were generated using blood specimens and solid prostate tumor tissue samples, respectively. TCGA quantified and normalized the digital gene expression levels using RNASeq by Expectation Maximization (RSEM) method to generate the level-3 data (36). We further performed $\log_2$ transformation of the expression levels before the statistical analysis.

1000 Genome

"Beagle" is a software package for phasing genotypes and imputing ungenotyped markers (37). From the Beagle 5.1 website (http://faculty.washington.edu/browning/beagle/beagle. html/) maintained by the University of Washington (Seattle, WA), we downloaded (Oct 28, 2018) 1000 Genomes Project Phase III data release (version 5a) (https://www.internationalgenome.org/) in VCF format for use with Beagle version 4.x. The genotypes of all individuals, both males and females, in EUR and AFR super-populations are used in this study.

Data integration

Data manipulation

All the five datasets were preprocessed by their authors. In our study, no additional quality control operation is performed. The PLINK tool (version 1.9; refs. 38, 39; http://www.nitrc.org/projects/plink/) is employed for the integration and refining (filtering out microsatellites and SNPs on mitochondria genomes) of the CGEMS, BPC3, and MEC data. In these datasets, the alleles of genotyped SNPs are encoded with the DNA bases on the forward strands. However, TCGA and 1000 Genome datasets include many SNPs encoded on reverse strands. Using a lab-owned R program, we edit the latter datasets by flipping the allele codes of the SNPs whose strands are inconsistent with the corresponding ones in SNP Build 130 (downloaded from UCSC Genome Browser on Nov 14, 2018; http://genome.ucsc.edu/). Then, we compare the edited allele codes with those in CGEMS/BPC3/MEC datasets to filter out the incompatible SNPs, which account for approximately 1% of the total.

Risk SNP markers and proxy risk markers

In this study, PRS calculation is based on a work marker set. It consists of 93 index risk SNP (r-SM) or proxy-risk markers (p-r-SM) for prostate cancer. Here, the term p-r-SM indicates an SNP in the same linkage block with an r-SM ($R^2 > 0.5$). Both the r-SMs and p-r-SMs were collected or identified by Chen and colleagues (15). The 93 SNPs are selected by a priority procedure (Supplementary Table S1; Supplementary Materials and Methods 2). Of the 93 SNPs in the working marker set, six are located on the *X* chromosome and their test allele effects are not estimated when the LMM method is used. Among others (located on autosomes), one or two SNPs have the allele frequencies (MAF) less than 0.01 in a training dataset and are excluded in effect estimation and PRS calculation. Hereafter, we do not distinguish p-r-SMs from r-SMs in the statistical analysis and result presentation. Here, it should be especially noted that the bigger marker set ($N = 269$) redefined in the most recent transancestry genome-wide association meta-analysis (16) is not used in order to avoid the noise that could be introduced in the imputation of the genotypes of the marker SNPs (258 of 269) unmeasured in our training set.

Statistical analysis

Outline

Regarding a specific training dataset, we specify the minor allele of a specific SNP as "test allele" and the alternative allele as "reference". The effects of the test alleles of SNP markers on cancer occurrence are estimated by three methods, i.e., multiplicative model, multivariate logistic regression, and linear mixed model, using the three training sets separately. The PRS of a subject in the test set is accordingly calculated by summarizing the weighted test allele dosages over SNP markers. Besides those specially-noted software applications, the R package "stats" and lab-owned R functions are used to conduct other statistical computations. For example, the logistic regression is performed by implementing the R function *glm()*.

Multiplicative model

All the three model models used in this study assume that the test allele of a SNP exerts a dosage-dependent effect on cancer susceptibility. Multiplicative model (MPM) measures the effect of individual allele on a relative-risk scale. We use the PLINK software to calculate the per-allele log-odds ratio as the effect quantity. The formula for predicting the PRS of the i^{th} test sample is $PRS_i = \sum_{j=1}^{m} \hat{\beta}_j x_{ij}$, where $\hat{\beta}_j$ is the log-odds ratio for the j^{th} SNP estimated using a training dataset, m is the number of the SNP markers, and x_{ij} is the dosage of the test allele of the j^{th} SNP.

Multivariate logistic regression

With multivariate logistic regression (MLR), the effect of a test allele is estimated after the effects of variants on other loci and additional fixed factors are adjusted. Let $p_i = P(Y_i = 1)$ indicate the probability of a case. MLR can be expressed with the following mathematical form.

$$\log(\boldsymbol{p}/(1-\boldsymbol{p})) = 1\mu + X_1\alpha + X_2\beta \qquad \text{(A)}$$

In Eq. (A), μ is the intercept, X_1 is the $n \times k$ design matrix of n subjects for k additional factors besides SNP markers, X_2 is the $n \times m$ matrix for the dosages of test alleles over the m SNPs, $\boldsymbol{p} = (p_1, p_2, \ldots p_n)'$, $\alpha = (\alpha_1, \alpha_2, \ldots \alpha_k)'$, and $\beta = (\beta_1, \beta_2, \ldots \beta_m)'$ are the corresponding regression coefficient vectors, respectively. In particular, the ages of subjects (at the dates of diagnosis for cases and the cohort entry dates for controls) and the first ten principal components (representing population stratification) of the between-subjects genomic relationship matrix (40), which is calculated based on

the genotypes of approximately 150K SNPs common in the CGEMS, BPC3, MEC, and TCGA data, are included in the matrix X_1. The formula for predicting the PRS of a test sample is similar to that described in the MPM paragraph except that a constant, that is, the estimate of μ, is added to the score of each sample to make the results obtained from different training datasets comparable.

Linear mixed model.

Linear Mixed Model (LMM) assumes that cancer occurrence is a threshold trait and individual SNPs exert random effects on its liability. For sample i, the binary case identity Y_i and the underlying liability λ_i have the following relationship.

$$Y_i = \begin{cases} 1 \text{ for } \lambda_i \geq 0 \\ 0 \text{ for } \lambda_i < 0 \end{cases} \quad (B)$$

With liability as the dependent variable, LMM can be expressed with the following mathematical form.

$$\lambda = \mathbf{1}\mu + X_1\alpha + X_2\beta + E. \quad (C)$$

In Eq. (B), $\lambda = (\lambda_1, \lambda_2, \ldots \lambda_n)'$ is a vector for the liability quantities of the n included samples. In Eq. (C), the first and second right-side terms are similarly defined with those in the Eq. (A). X_2 is the matrix for normalized doses of test alleles over the m SNP markers. The random test allele effects have the distribution $\beta \sim N(0,\ I_{m\times m}\sigma_b^2)$. $\boldsymbol{E}$ is a vector for random noise with $E \sim N(0,\ I_{n\times n}\sigma_E^2)$, subject to $\sigma_e^2 = 1 - m\sigma_b^2$. The trait heritability $(h^2_{l(m)})$ attributed to the SNP markers can be estimated by $m\sigma_b^2$.

The implementation of the liability threshold LMM needs running a Bayesian Markov chain Monte Carlo (MCMC) algorithm (41), which is time-consuming. In the Genome-wide Complex Trait Analysis (GCTA) software used in this study, the left-side term of the Eq. (C) is replaced with the 0/1 valued case identity vector Y such that it can be efficiently implemented using the equivalence model technique and Restricted Maximum Likelihood (REML) method (25, 42).

Methods for simulation study

The simulation study is conducted to show the different efficiencies in identifying risk SNP markers from AA and EA datasets. Here, the statistical model for simulating prostate cancer susceptibility is presented. The model implementation and other methods used in the relevant analysis are included in the last subsection of the Results.

For an individual, denoted by i, in a population ("AA" or "EA"), we simulate his prostate cancer susceptibility (P_i) and actual occurrence (O_i) using the following model.

$$P_i = min(B_i + g_i\gamma, 1),\ B_i \sim n(B, c^2 \times B^2),\ p(O_i = 1) = P_i \quad (D)$$

In Eq. (D), B is a probability corresponding to the "background" prostate cancer prevalence determined by all non-genetic and genetic factors except for the focused genetic marker (SNP) G whose risk allele dosage is denoted by g_i. γ is the effect of the risk allele compared with the reference allele. c is a constant heuristically assigned with a numeric value between 0.05 to 0.15. The variability coefficient of a metric trait in humans and animals usually falls within the interval.

Data availability

The genotype datasets of the three GWAS projects and TCGA are deposited in The database of Genotypes and Phenotypes (dbGaP; https://www.ncbi.nlm.nih.gov/gap/) at phs000207.v1.p1, phs000812.v1.p1, phs000306.v3.p1, and phs000178.v10.p8. Access to these collections is controlled by the data access committees in the NIH. The gene expression data of the TCGA samples is publically available in Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/). The 1000-Genome data is currently available from http://aws.amazon.com/1000genomes/.

Results

Estimating PRS is the key step to derive the main results of this study. While three methods are used in the calculations, here we first focus on the outputs when LMM is used. The model is favored because it is based on the classic quantitative genetics theory, in which the primary assumption is that a trait is determined by minor-effect factors such as germline SNPs existing in normal cells. Then, the outputs' consistency from different methods is presented in a separated subsection. In addition, a simulation experiment is performed to justify the results of real data analysis.

Africans' higher PRSs compared with Europeans

We investigate the polygenic basis for racial disparity in cancer occurrence by comparing the PRSs of Africans and Europeans. The used test samples are the individuals of the EUR and AFR super-populations in the 1000-Genome project. This design is based on the assumption that each man has an inherent prostate cancer susceptibility (or liability), which can be estimated by the PRS, regardless of the actual life-span diagnosis for the disease. The hypothesis to be tested is that a high-risk population, such as AA, has a higher average PRS than a (relatively) low-risk population.

As shown in **Fig. 1**, the racial disparity patterns in PRS are clear and robust to the training datasets (i.e., CGEMS, BPC3, and MEC) that are used to estimate variant effects. For each test sample, we calculate an aggregated PRS by averaging the three scores that are based on the three training sets. Two t tests are performed on the aggregated PRSs, with a person or country (region) based population as the measurement unit, respectively. Both the tests show that the differences between Africans and Europeans are extremely significant ($P < 1 \times 10^{-168}$ or $P < 1 \times 10^{-6}$).

It should be noted that the term "aggregated PRSs" referred to afterwards has the same meaning as defined in this subsection.

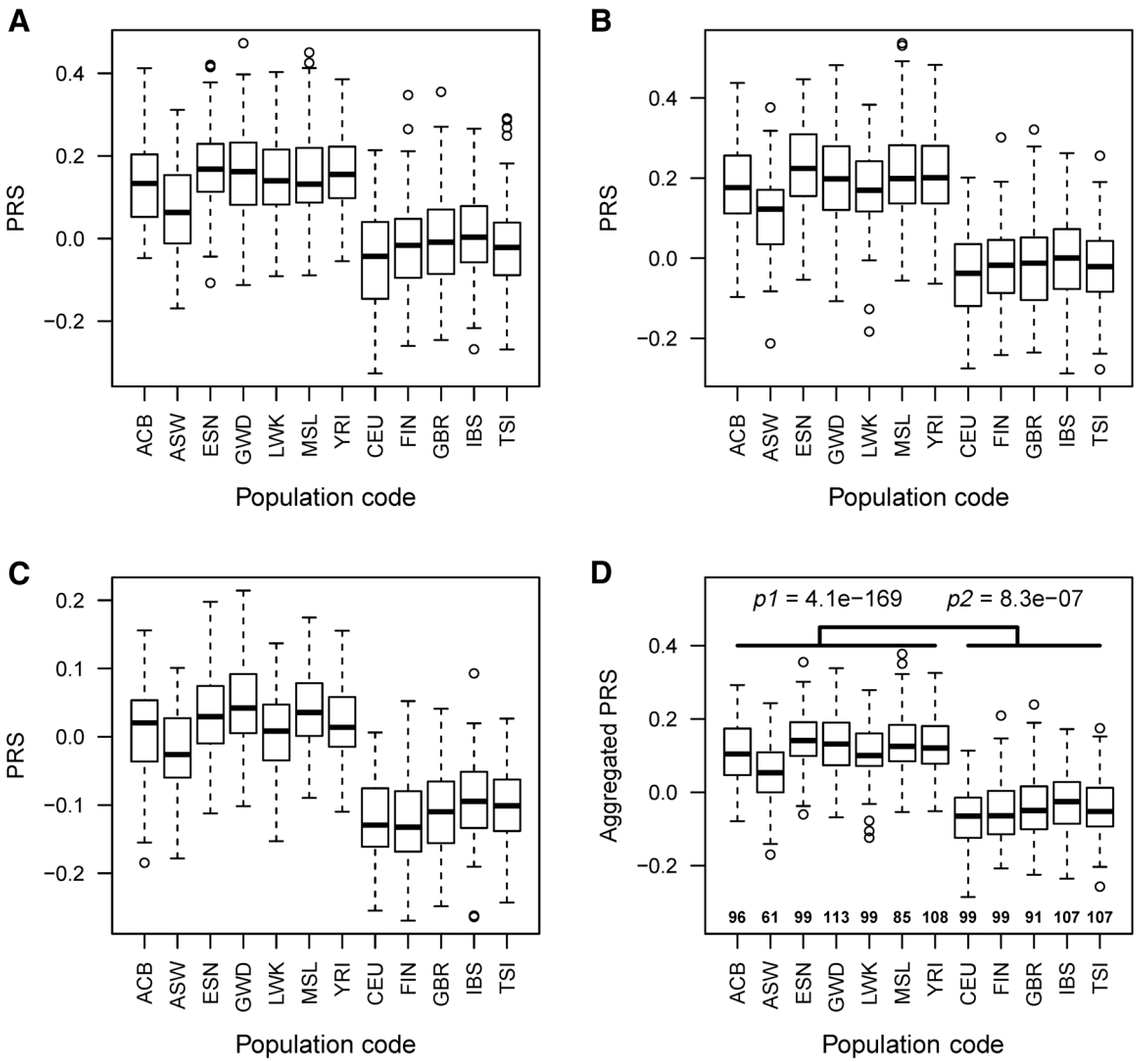

Figure 1.

PRS prediction of the 1000-Genome samples. **A–C,** PRSs are calculated according to the test allele effects that are estimated using the CGEMS, BPC3, and MEC datasets, respectively. **D,** Aggregated PRS of a subject is the average of the PRSs used in **A–C,** and the p1 and p2 are the *P* values of two t tests with a person or a population as the measurement unit, respectively. Sample categories (*x*-axis) are denoted by population codes. On the left side are seven African-ancestry populations, including ACB, ASW, ESN, GWD, LMK, MSL, and YRI. On the right are five European-ancestry populations, including CEU, FIN, GBR, IBS, and TSI. The full names of the populations denoted by the codes are available in ref. 32.

PRS stratification in patients with prostate cancer

The aforementioned result inspires us to propose another hypothesis. That is, genetic factors contribute more to prostate cancer cases in a high-risk population than to those in a low-risk population. We test this hypothesis using the TCGA data. As demonstrated in **Fig. 2A**, the aggregated PRS in AA patients is significantly higher than the score in EA patients ($P < 3 \times 10^{-9}$). The PRS distribution profile of the "unknown group" is almost the same as that of the EA group. This is logical because the TCGA data are predominated with EA samples. The PRS profile of Asian patients is unexpected, not compatible with the quite low prostate cancer prevalence in Asians. However, because there are only six samples in the group, it is insufficient to investigate the underlying mechanisms further.

Association between PRS and cancer subtypes

The most common alterations in prostate cancer genomes are fusions of promoters of androgen-regulated genes with the gene bodies of transcription factors in the E26 transformation-specific (ETS) family. The genes involved in the fusions mainly include *TMPRSS2* and four ETS members (*EGR, ETV1,*

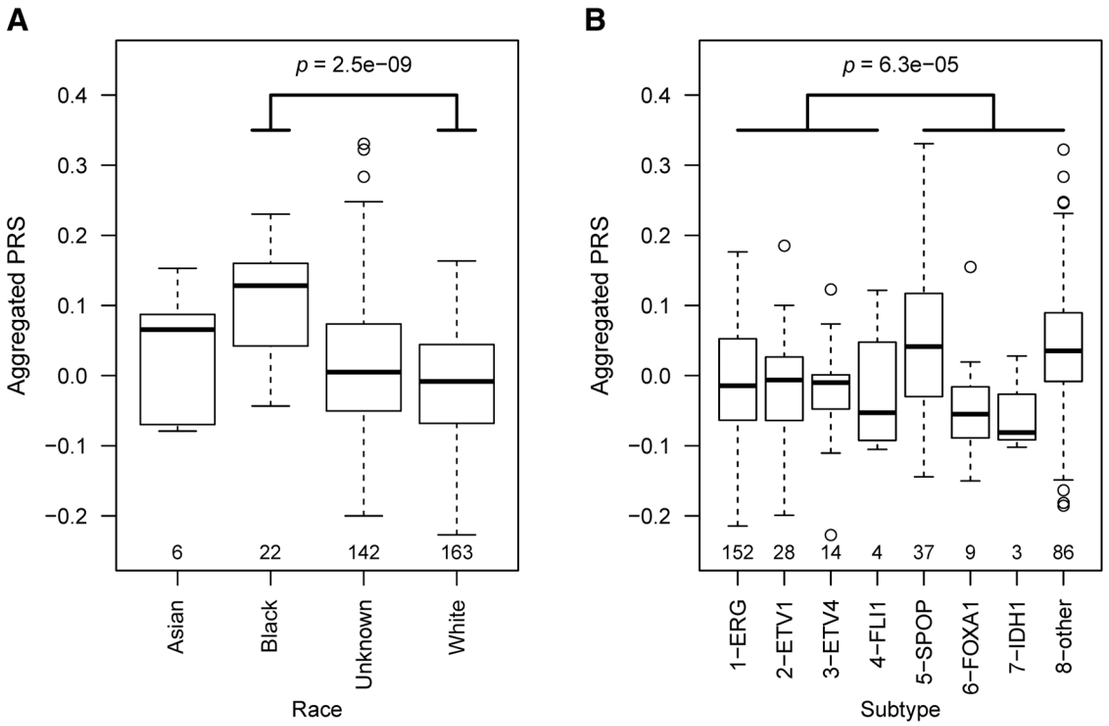

Figure 2.

The stratification of aggregated PRSs across racial groups (**A**) and somatic mutation-based cancer subtypes (**B**) of the patients in the TCGA prostate adenocarcinoma data. In **B**, the four categories on the left side constitute the ETS^+ subtype cluster and the four categories on the right side constitute the ETS^- subtype cluster. *P* values are estimated using a *t* test.

ETV1, FLI1), with *TMPRSS2-ERG* as the predominant fusion type (43). Racial disparity in the fraction of *TMPRSS2-ERG* carcinomas among prostate cancer cases is well known. We wonder if there is a relationship between such a clinical disparity and the germline alterations that predispose prostate cancer. In the molecular taxonomy established by TCGA researchers, the focused 333 tumors each falls into one of eight categories, of which seven are subtypes defined by specific gene fusions (*ERG, ETV1, ETV4*, and *FLI1*) or mutations (*SPOP, FOXA1*, and *IDH1*; ref. 35). Considering the limited sample size, we combine the eight categories (subtypes) into two subtype clusters, i.e., ETS-fusion–positive (ETS^+) and ETS-fusion–negative (ETS^-), to facilitate statistical analysis (**Fig. 2B**).

The result from a *t* test shows that the aggregated PRSs of ETS^+ cluster are significant lower than the scores of ETS^- cluster ($P < 1 \times 10^{-4}$). We further find that this association is still significant ($P < 0.001$) after PRS is adjusted for races by a linear model analysis, where PRS is the dependent variable and race and ETS-fusion status are independent variables. Moreover, an ROC analysis focusing on EA patients demonstrates that the race-adjusted PRS has substantial classification strength in distinguishing the men with ETS^+ tumors from those with ETS^- tumors. The resulting AUC is 0.61.

There is heterogeneity in both the ETS^+ and ETS^- clusters. The difference in PRS between them is due to the high scores in the two major subtypes, i.e., *SPOP* and "others", of the ETS^- cluster. While the result of a statistical test is subject to group sizes, the linear model analysis, as mentioned in the last paragraph, shows that the PRS differences for *ERG* versus *SPOP* and *ERG* versus "others" are still significant, with the *P* values being 0.019 and 0.002, respectively. When the *ERG* subtype is considered as the canonical $TRMPSS2^+$ and the aggregate of the other seven subtypes is considered as $TRMPSS2^-$, the PSR of the former is significantly lower than the latter ($P < 0.03$). The *FOXA1* and *IDH1* subtypes of the ETS^- cluster have low scores. However, because there are only 9 and 3 samples, respectively, in the two groups, it is difficult to obtain meaningful results by comparing them with other subtypes.

Logically, the risk SNP markers do not equally contribute to the association between the ETS-fusion status and PRS. In this regard, we scan the effects of genotypes of the 93 SNPs in the working marker set on the ETS-fusion status using a logistic model, identifying three SNPs (rs10934853, rs339331, rs10505483) with FDR less than 0.05 (**Table 2**). These SNPs are located on chromosome-3, -6, and -8, respectively. Among the SNPs that do not meet the FDR cut-off, 11 (rs1391438, rs1983891, rs1512268, rs4242382, rs723338, rs385894,

Table 2. Risk SNPs associated with the ETS-fusion status[a].

	rs10934853	rs339331	rs10505483
Genome	Chr 3	Chr 6	Chr 8
Physical position in hg38	129521063	117316745	128194377
Allele-A[b]	A	C	A
Allele-B	C	T	G
Standardized regression coefficient[c]	−3.68	3.18	−3.69
Allele-A frequency in TCGA samples	0.34	0.26	0.12
Allele-A frequency in CGEMS cases	0.31	0.28	0.04
Allele-A frequency in CGEMS controls	0.27	0.31	0.03
Allele-A frequency in BPC3 cases	0.30	0.28	0.05
Allele-A frequency in BPC3 controls	0.28	0.31	0.03
Allele-A frequency in MEC cases	0.71	0.22	0.47
Allele-A frequency in MEC controls	0.69	0.25	0.40
Odds ratio in CGEMS	1.22	0.90	1.24
Odds ratio in BPC3	1.10	0.88	1.37
Odds ratio in MEC	1.10	0.80	1.30

[a]For each of the three SNPs, the germline genotypes of patients are significantly associated with ETS-fusion status (ETS^+or ETS^-) of their tumor samples. rs10505483 is a proxy risk marker that is in strong linkage disequilibrium with the index risk SNP rs16901979 ($R^2 = 1$).
[b]Allele-A and Allele-B are temporarily treated as the test allele and reference allele in calculating the odds ratios shown in this table.
[c]The standardized regression coefficient is estimated using a logistic model, in which the ETS-fusion status, indicated with 1 and 0, is the response variable and the dosage of the test allele of a SNP is the independent variable.

rs2191139, rs6091236 rs5945619, rs276698, and rs11568818) have the ordinary P value < 0.05. Intuitively, this subset of SNPs also can contribute to the differentiation of *ETS/TMPRSS2* status on PRS.

Association between PRS and gene expression level in tumor samples

Using the information of TCGA samples, we scan the Pearson correlations between PRSs and the expression levels of approximately 20,000 genes and test their significance. The P values are adjusted using the Benjamin—Hochberg method. With the cut-off of adjusted P (FDR) < 0.01, 10 significant genes are identified (**Table 3**). Except for *DHRS4-AS1*, *KRTAP4-3* and *C3orf33*, all the other eight genes are differentially expressed between ETS^+ and ETS^- patients (ordinary P value < 0.01). As demonstrated in literature, a fraction of these genes play roles in prostate cancer. For example, SET domain bifurcated (*SETDB1*), coding a histone methyltransferase, was proposed to be an oncogene in prostate cancer, which is required for the proliferation, migration, and invasion of cancer cells (44). Glutathione S-Transferase Mu 3 (*GSTM3*) is a player in driving tumor progression, being dysregulated in prostate cancer (45). Forkhead box O1 (*FOXO1*) inhibits Runx2 transcriptional activity and prostate cancer cell migration and invasion (46).

Table 3. Significant genes for the association between gene expression level and PRS[a].

Symbol	Genome	r[b]	p1[b]	Diff[c]	p2[c]
ACAT2	chr6	0.27	2.9×10^{-9}	0.38	2.1×10^{-6}
DHRS4-AS1	chr14	−0.23	4.9×10^{-7}	0.01	8.3×10^{-1}
C3orf33	chr3	0.24	1.4×10^{-7}	0.11	1.3×10^{-2}
FOXK1	chr7	−0.21	4.6×10^{-6}	−0.31	4.1×10^{-5}
GSTM3	chr1	0.21	4.7×10^{-6}	0.54	4.5×10^{-7}
KRTAP4-3	chr17	0.21	3.7×10^{-6}	0.14	4.7×10^{-1}
LOC90784	chr2	−0.24	9.3×10^{-8}	−0.24	6.2×10^{-3}
POGZ	chr1	−0.21	3.7×10^{-6}	−0.24	1.0×10^{-10}
SETDB1	chr1	−0.21	2.4×10^{-6}	−0.2	1.4×10^{-9}
STK35	chr20	−0.22	9.2×10^{-7}	−0.26	1.3×10^{-9}

[a]The gene expression levels of TCGA prostate tissue samples and the patients' aggregated PRSs are used in the analysis.
[b]r and p1 are the Pearson correlation values between gene expression level and PRS and the corresponding P value.
[c]Diff and p2 are the differences of gene expression levels between ETS^+ and ETS^- tumor samples and the corresponding P value.

Robustness of PRS to statistical methods

We calculate TCGA samples' PRSs using different statistical models (MPM, MLR, and LMM) and different training datasets (CGEMS, BPC3, and MEC), and depict these PRS estimates with nine scatter plots. Pearson correlations for the paired PRS vectors are calculated to evaluate the consistency. As shown in **Fig. 3**, the estimates from MLR are highly congruent with the estimates from LMM ($r \geq 0.93$). The consistency between the results of MPM and the results of the other two methods is relatively poorer, with the correlation coefficients ranging from 0.82 to 0.93. Nevertheless, the aforementioned associations between PRSs and tumor subtypes, as well as the associations between PRSs and patient races, can be repeated when MPM and MLR are used in the PRS calculation. However, the significant levels, evaluated by P values, are somewhat lower.

Challenge for the identification of risk SNP markers in the AA (or other AFR) population

The aforementioned observations clearly indicate the relevance of the indexed risk SNP markers to racial disparity in

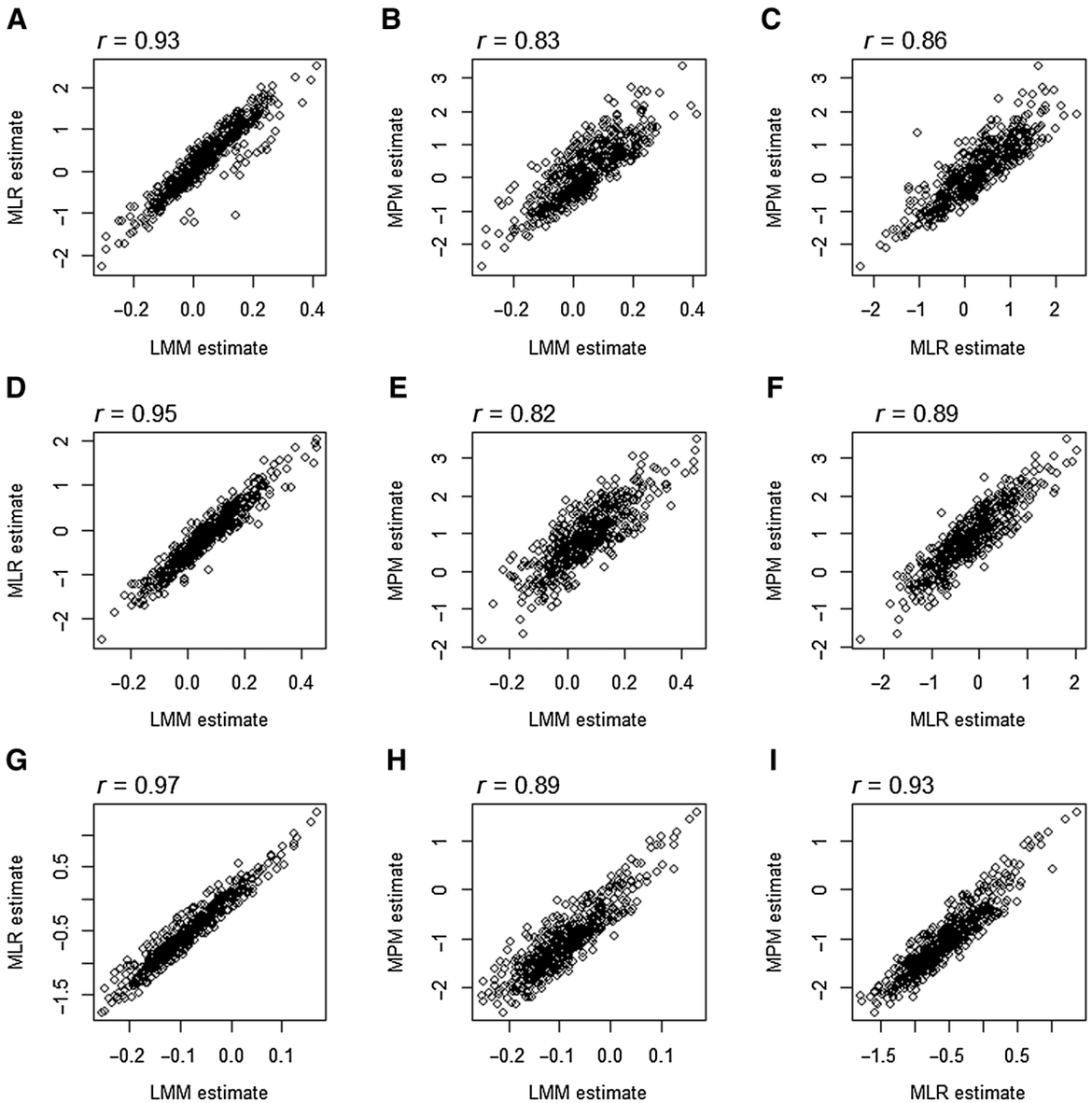

Figure 3.

Consistency evaluation of PRSs calculated using different statistical methods and training data. Data points in the plots represent test samples, i.e., the patients in TCGA data. The axis labels indicate the methods by which PRSs are calculated. In top row (**A–C**), middle row (**D–F**), and bottom row (**G–I**), the CGEMS, BPC3, and MEC datasets are used as the training set to estimate allele effects, respectively, from which PRSs are calculated.

prostate cancer. That is, Africans have higher PRSs, i.e., greater genetic predisposition for the disease, than Europeans. Here, we present the result of a simulation study, which shows that the high prostate cancer prevalence in African populations makes it challenging to identify risk SNP markers using African males' data. This finding can help explain the dilemma that the risk SNPs initially identified in EA cohorts rarely meet the primary significant criterion, such as 1×10^{-6} or 1×10^{-7}, when their association with prostate cancer occurrence is tested using an African/African American cohort dataset.

Prostate cancer susceptibility is simulated using the model i.e., Eq. (D), described in the subsection of "Methods for simulation study". For an "EA" population in the Hardy–Weinberg equilibrium, we specify B and c with 0.1 and 0.075, respectively. Considering the widely existing overdiagnosis, the assigned B value is lower (0.1 versus 0.15) than the prostate cancer prevalence derived from the Surveillance Epidemiology and End Results (SEER) database. We simulate 36 source data sets, each of which contains 2000,000 men and corresponds to a combination (scenario) of a risk allele

frequency (RAF = 0.01, 0.02, 0.04, 0.08, 0.16, 0.32, or 0.50) and an allele effect ($\gamma = 0.001, 0.002, 0.004, 0.008, 0.016$, or 0.032) for the genetic marker. From each source dataset, we sample 10,000 working datasets (replicates), each of which contains 4,000 cases (whose O_i is 1) and 4,000 controls (whose O_i is 0). Using each working dataset, we perform a logistic regression analysis, in which O_i represents the response variable and g_i represents the independent variable. We summarize the results by a table which presents, for each combination of RAF and α, the fraction of the working datasets from which the variant effect can be identified by the criterion of P value $< 1 \times 10^{-3}$, 1×10^{-4}, 1×10^{-5}, 1×10^{-6}, or 1×10^{-7}.

The data of an "AA" population is similarly simulated and analyzed as the "EA" population, except that B and c are specified with 0.15 and 0.05, respectively. Compared with the "EA" population, the c value is one-third lower such that the same susceptibility variance is maintained in "EA" and "AA" populations.

As shown in Supplementary Table S2 and S3, it is difficult to identify a risk marker with the usually used golden cut-off of $P<1 \times 10^{-7}$ when the effect β is less than 0.01. When the effect is at the level of 0.016 and the RAF is higher than 0.08, the significance level can reach, in at least 5% of replicates, the golden cut-off in the EA population but not in the AA population. The only scenarios in which a marker can be identified with the golden cut-off in the AA populations are the combinations of $\gamma = 0.032$ and $RAF > 0.08$. Given such a large effect, the marker can be identified in the EA population even when the RAF is as low as 0.04.

Discussion

Based on the aforementioned integrative analysis of five public genomic datasets, we conclude that the risk SNP marker-based PRS is compatible with the racial disparity in prostate cancer occurrence. The conclusion is supported by the following observations. First, high-risk (genetic) populations for prostate cancer, i.e., those in the AFR super-population of the 1000-Genome project, have higher average PRSs, is consistent with the reference (31). Second, susceptibility loci indexed by the risk SNP markers contribute more to the prostate cancer cases in a high-risk population than to those in a (relatively) low-risk population. These results are robust to the training datasets and statistical methods used to estimate the effects of individual variants. The importance of risk SNP markers to cancer disparity research is further highlighted by one of our findings. That is, the men with ETS^+ tumors have higher PRSs than those with ETS^- tumors. Several issues related to our results and meriting further investigation are discussed as follows.

PRS and susceptibility heritability

Our preliminary analysis as well as previous publications show that the heritability estimates range from 0.25 to 0.78 when the whole array polymorphisms (500K–1,000K SNPs) are considered or range from 0.04 from 0.10 when only the (proxy) SNP markers are considered (19, 47–49). This implies that, besides the index risk SNPs, a large set of SNPs have relevance to prostate cancer but have yet been identified. A naïve method to pick up the effects of these SNPs in PRS calculation is to include the whole array of variants in the mixed model. However, our preliminary analysis using the approximately 150K SNPs shared by the five datasets shows that, only a training set with a large sample size, such as the BPC3 data or the whole MEC data which combines all the AA, Japanese American (JA), and Latino American (LA) samples, can lead to a result being compatible with the racial disparity in prostate cancer occurrence.

PSR and subtype-specific cancer prevalence

According to SEER data, the prostate cancer prevalence in the United States ranges from 0.14 to 0.16 in the EA population and from 0.19 to 0.24 in the AA population across the survey years (4). The cohort data used in more-focused experimental studies show that *TMPRSS2-ERG*–positive tumors, i.e., those with *TMPRSS2-ERG* fusions, amount to 45% to 50% and 25% to 30% of the prostate cancer samples in the EA and AA (or other AFR) populations, respectively (6, 50–52). With these statistics, we can get an estimate of the prevalence of *TMPRSS2-ERG*–positive cancer. The estimate has a similar range in EAs (0.06–0.07) and AAs (0.05–0.07), implicating that racial disparity exists in the prevalence of *TMPRSS2-ERG*–negative prostate cancer but not in the prevalence of *TMPRSS2-ERG*–positive prostate cancer. Because *TMPRSS2-ERG* tumors are predominant (>80%) in tumors with ETS fusions, we could further perceive that racial disparity exist in the prevalence of ETS^- subtype clusters but not in the prevalence of ETS^+ subtype cluster. These arguments, together with the observed negative association between ETS fusion status and PRS, inspire us to propose a novel hypothesis. That is, prostate cancers could be partitioned into two categories, i.e., "hereditary" and "sporadic", and racial disparity in prevalence is limited to the first one.

PRS and gene expression levels

In this study, we identified 10 genes with their expression levels being associated with PRSs (FDR < 0.05). Of them, in addition to *SETDB1, GSTM3* and *FOXO1* whose roles in prostate cancer progression have been reported (see Results), other two genes are also involved in cancer types. A recent publication reports that the downregulated long noncoding RNA *DHRS4-AS1* is protumoral and associated with the prognosis of clear-cell renal cell carcinoma (53). A bioinformatics analysis shows that the coding gene *POGZ* is disrupted in many of the cancer datasets and has an unusually large number of predicted targets involved in the cancer pathways (54). The associations between PRS and gene expression level in tumor samples suggest that the germline mutations exert impact on cancer cells, implicating the active roles in tumor progression. Here, we wonder

whether and/or how PRS is relevant to the initialization of prostate tumors. In order to get insight into this issue, the analysis of gene expression profiles of the normal prostate tissue samples of young or middle-age adult men is required. At present, it is still difficult to get sufficient data for performing such research.

Unbalance of index risk SNPs across racial populations

Regarding prostate cancer susceptibility loci, the index SNPs identified in EUR populations are predominant over those identified in others, including AFR populations. Park and colleagues (2017) found that this unbalance mirrors the pro-European distribution of samples included in the discovery phase of cancer GWAS to date (55). In particular, of the nearly 100,000 prostate cancer cases and controls included in the discovery stage of GWAS, the ratio of EUR subjects to AFR subjects is 8:1, approximating the ratio between the number of index risk SNPs identified in Europeans and the number for Africans. In our study, the result of the simulation experiment provides a complementary explanation to the disparity in index risk SNPs. That is, the high prostate cancer prevalence in African populations makes it challenging to identify risk variants using African/African men's data. As such, considering that the heritability of prostate cancer susceptibility estimated using Africans' data is much lower than the estimates from Europeans' data (19), we conceive an issue meriting further investigation. Does the racial disparity in prostate cancer prevalence also exert influence on the estimation of the genetic parameter?

Authors' Disclosures

No disclosures were reported.

Authors' Contributions

W. Zhang: Conceptualization, data curation, software, formal analysis, validation, investigation, visualization, methodology, writing–original draft, project administration, writing–review and editing. **T. Nicholson:** Writing–original draft, writing–review and editing. **K. Zhang:** Conceptualization, resources, formal analysis, supervision, funding acquisition, investigation, methodology, writing–original draft, project administration, writing–review and editing.

Acknowledgments

This research is supported by the NIH grant 5U54MD007595 (all authors). The authors are grateful to the two reviewers for their constructive comments which have significantly improved this paper.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received August 24, 2021; revised November 22, 2021; accepted December 22, 2021; published first December 28, 2021.

References

1. Siegel R, Ma J, Zou Z, Jemal A. Cancer statistics, 2014. CA Cancer J Clin 2014;64:9–29.
2. PDQ Adult Treatment Editorial Board. Prostate Cancer Treatment (PDQ®)–Health Professional Version. In: PDQ Cancer Information Summaries. Bethesda (MD): National Cancer Institute.
3. Wu I, Modlin CS. Disparities in prostate cancer in African American men: what primary care physicians can do. Cleve Clin J Med 2012;79:313–20.
4. Noone AM, Howlader N, Krapcho M, Miller D, Brest A, Yu M, et al. (eds). SEER Cancer Statistics Review, 1975–2015. Bethesda, MD: National Cancer Institute; 2021. Available from: https://seer.cancer.gov/csr/1975_2015/.
5. Shenoy D, Packianathan S, Chen AM, Vijayakumar S. Do African-American men need separate prostate cancer screening guidelines? BMC Urol 2016;16:19.
6. Zhou CK, Young D, Yeboah ED, Coburn SB, Tettey Y, Biritwum RB, et al. TMPRSS2:ERG gene fusions in prostate cancer of West African men and a meta-analysis of racial differences. Am J Epidemiol 2017;186:1352–61.
7. Powell IJ, Meyskens FL Jr. African American men and hereditary/familial prostate cancer: Intermediate-risk populations for chemoprevention trials. Urology 2001;57:178–81.
8. Eldon BJ, Jonsson E, Tomasson J, Tryggvadottir L, Tulinius H. Familial risk of prostate cancer in Iceland. BJU Int 2003;92:915–9.
9. Johns LE, Houlston RS. A systematic review and meta-analysis of familial prostate cancer risk. BJU Int 2003;91:789–94.
10. Lynch HT, Kosoko-Lasaki O, Leslie SW, Rendell M, Shaw T, Snyder C, et al. Screening for familial and hereditary prostate cancer. Int J Cancer 2016;138:2579–91.
11. Haiman CA, Chen GK, Blot WJ, Strom SS, Berndt SI, Kittles RA, et al. Characterizing genetic risk at known prostate cancer susceptibility loci in African Americans. PLoS Genet 2011;7:e1001387.
12. Petrovics G, Li H, Stumpel T, Tan SH, Young D, Katta S, et al. A novel genomic alteration of LSAMP associates with aggressive prostate cancer in African American men. EBioMedicine 2015;2:1957–64.
13. Huang KL, Mashl RJ, Wu Y, Ritter DI, Wang J, Oh C, et al. Pathogenic germline variants in 10,389 adult cancers. Cell 2018;173:355–70.
14. Giovannucci E, Stampfer MJ, Krithivas K, Brown M, Dahl D, Brufsky A, et al. The CAG repeat within the androgen receptor gene and its relationship to prostate cancer. Proc Natl Acad Sci U S A 1997;94:3320–3.
15. Chen H, Yu H, Wang J, Zhang Z, Gao Z, Chen Z, et al. Systematic enrichment analysis of potentially functional regions for 103 prostate cancer risk-associated loci. Prostate 2015;75:1264–76.
16. 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.
17. Schumacher FR, Al Olama AA, Berndt SI, Benlloch S, Ahmed M, Saunders EJ, et al. Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. Nat Genet 2018;50:928–36.
18. Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, Wacholder S, et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat Genet 2007;39:645–9.
19. Mancuso N, Rohland N, Rand KA, Tandon A, Allen A, Quinque D, et al. The contribution of rare variation to prostate cancer heritability. Nat Genet 2016;48:30–5.

20. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, et al. Environmental and heritable factors in the causation of cancer–analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med 2000;343:78–85.
21. Eeles RA, Olama AA, Benlloch S, Saunders EJ, Leongamornlert DA, Tymrakiewicz M, et al. Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. Nat Genet 2013;45:385–91.
22. Dudbridge F. Power and predictive accuracy of polygenic risk scores. PLoS Genet 2013;9:e1003348.
23. Pashayan N, Pharoah PD, Schleutker J, Talala K, Tammela T, Maattanen L, et al. Reducing overdiagnosis by polygenic risk-stratified screening: findings from the Finnish section of the ERSPC. Br J Cancer 2015;113:1086–93.
24. Malovini A, Bellazzi R, Napolitano C, Guffanti G. Multivariate methods for genetic variants selection and risk prediction in cardiovascular diseases. Front Cardiovasc Med 2016;3:17.
25. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 2011;88: 76–82.
26. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci 2008;91:4414–23.
27. 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:djx084.
28. Haiman CA, Chen GK, Blot WJ, Strom SS, Berndt SI, Kittles RA, et al. Genome-wide association study of prostate cancer in men of African ancestry identifies a susceptibility locus at 17q21. Nat Genet 2011;43: 570–3.
29. 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:djv431.
30. Irizarry-Ramirez M, Kittles RA, Wang X, Salgado-Montilla J, Nogueras-Gonzalez GM, Sanchez-Ortiz R, et al. Genetic ancestry and prostate cancer susceptibility SNPs in Puerto Rican and African American men. Prostate 2017;77:1118–27.
31. Lachance J, Berens AJ, Hansen MEB, Teng AK, Tishkoff SA, Rebbeck TR. Genetic hitchhiking and population bottlenecks contribute to prostate cancer disparities in men of African Descent. Cancer Res 2018;78:2432–43.
32. 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.
33. Schumacher FR, Berndt SI, Siddiq A, Jacobs KB, Wang Z, Lindstrom S, et al. Genome-wide association study identifies new prostate cancer susceptibility loci. Hum Mol Genet 2011;20:3867–75.
34. Gohagan JK, Prorok PC, Hayes RB, Kramer BS, Prostate LC., Ovarian Cancer Screening Trial Project T. The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial of the National Cancer Institute: history, organization, and status. Control Clin Trials 2000; 21:251S–72S.
35. Cancer Genome Atlas Research Network. The molecular taxonomy of primary prostate cancer. Cell 2015;163:1011–25.
36. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinf 2011;12:323.
37. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 2007;81: 1084–97.
38. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81: 559–75.
39. 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.
40. Goddard ME, Hayes BJ, Meuwissen TH. Using the genomic relationship matrix to predict the accuracy of genomic selection. J Anim Breed Genet 2011;128:409–21.
41. Sorensen D, Andersen S, Gianola D, Korsgaard I. Bayesian inference in threshold models using Gibbs sampling. Genetics Selection Evolution 1995;27:229–49.
42. Lee SH, Wray NR, Goddard ME, Visscher PM. Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet 2011;88:294–305.
43. Rubin MA. ETS rearrangements in prostate cancer. Asian J Androl 2012;14:393–9.
44. Wu M, Fan B, Guo Q, Li Y, Chen R, Lv N, et al. Knockdown of SETDB1 inhibits breast cancer progression by miR-381–3p-related regulation. Biol Res 2018;51:39.
45. Checa-Rojas A, Delgadillo-Silva LF, Velasco-Herrera MDC, Andrade-Dominguez A, Gil J, Santillan O, et al. GSTM3 and GSTP1: novel players driving tumor progression in cervical cancer. Oncotarget 2018; 9:21696–714.
46. Zhang H, Pan Y, Zheng L, Choe C, Lindgren B, Jensen ED, et al. FOXO1 inhibits Runx2 transcriptional activity and prostate cancer cell migration and invasion. Cancer Res 2011;71:3257–67.
47. Machiela MJ, Chen CY, Chen C, Chanock SJ, Hunter DJ, Kraft P. Evaluation of polygenic risk scores for predicting breast and prostate cancer risk. Genet Epidemiol 2011;35:506–14.
48. Gusev A, Shi H, Kichaev G, Pomerantz M, Li F, Long HW, et al. Atlas of prostate cancer heritability in European and African-American men pinpoints tissue-specific regulation. Nat Commun 2016;7:10979.
49. Zhang W, Dong Y, Sartor O, Zhang K. Comprehensive analysis of multiple cohort datasets deciphers the utility of germline single-nucleotide polymorphisms in prostate cancer diagnosis. Cancer Prev Res 2021;14:741–52.
50. Khani F, Mosquera JM, Park K, Blattner M, O'Reilly C, MacDonald TY, et al. Evidence for molecular differences in prostate cancer between African American and Caucasian men. Clin Cancer Res 2014;20:4925–34.
51. Magi-Galluzzi C, Tsusuki T, Elson P, Simmerman K, LaFargue C, Esgueva R, et al. TMPRSS2-ERG gene fusion prevalence and class are significantly different in prostate cancer of Caucasian, African-American and Japanese patients. Prostate 2011;71:489–97.
52. Rosen P, Pfister D, Young D, Petrovics G, Chen Y, Cullen J, et al. Differences in frequency of ERG oncoprotein expression between index tumors of Caucasian and African American patients with prostate cancer. Urology 2012;80:749–53.
53. Wang C, Wang G, Zhang Z, Wang Z, Ren M, Wang X, et al. The downregulated long noncoding RNA DHRS4-AS1 is protumoral and associated with the prognosis of clear cell renal cell carcinoma. Onco Targets Ther 2018;11:5631–46.
54. Newton R, Wernisch L. A meta-analysis of multiple matched aCGH/expression cancer datasets reveals regulatory relationships and pathway enrichment of potential oncogenes. PLoS One 2019;14:e0213221.
55. Park SL, Cheng I, Haiman CA. Genome-wide association studies of cancer in diverse populations. Cancer Epidemiol Biomarkers Prev 2018;27:405–17.