1.a. Full Title: 8q24 Region Variants from GWAS and Exome Sequencing and Prostate Cancer Risk in the ARIC study

b. Abbreviated Title (Length 26 characters): 8q24 and Prostate Cancer Risk

2. Writing Group:
   Writing group members: Melody Swen, Adrienne Tin, Elizabeth A. Platz, Terri Beaty, Corrie E. Joshu, Eric Boerwinkle, and all other interested ARIC Investigators

I, the first author, confirm that all the coauthors have given their approval for this manuscript proposal. _MS_ [please confirm with your initials electronically or in writing]

First author: Melody Swen

   Address: 615 N. Wolfe Street
   Office W2510
   Baltimore, MD, 21205
   Phone: (949) 838-6789
   E-mail: mswen1@jhu.edu

ARIC author to be contacted if there are questions about the manuscript and the first author does not respond or cannot be located (this must be an ARIC investigator).

   Name: Elizabeth A. Platz
   Address: 615 N. Wolfe Street
   Office E6132
   Baltimore, MD 21205
   Phone: (410) 614-9674     Fax: (410) 614-2632
   E-mail: eplatz1@jhu.edu

3. Timeline:
   Data analysis on the exome sequencing data – 2 months
   Draft of the results from exome sequencing – 3 months

4. Rationale:
Prostate cancer is the second most common cancer in men in the United States with 161,360 estimated new cases in 2017 [1]. With an estimated 26,730 deaths in the U.S. attributed to the disease in 2017, prostate cancer is also the second leading cause of cancer death [1]. Prostate cancer is a complex disease, with a number of established risk factors such as older age, African ancestry, and a family history of prostate cancer [2].

While a family history of prostate cancer is suggestive of a genetic contribution, shared environment could also be explanatory. Studies of mono- and dizygotic twins indicate that approximately 42% to 58% of prostate cancer risk is due to inherited genetics [2, 3].

Previous genome-wide association studies (GWASs) have sought to identify common genetic determinants of prostate cancer. This approach has been successful at identifying about 90 variants associated with prostate cancer risk, albeit modestly, in populations of European ancestry [4]. A meta-analysis of >10 million SNPs in 43,303 prostate cancer cases and 43,737 controls in European, African, Japanese and Latino ancestral populations revealed 23 new susceptibility loci that explained 33% of familial risk for the disease in European-ancestry populations [5]. A follow-up study that fine-mapped 64 GWAS regions estimated that the prostate cancer associated loci explain 38.9% of the familial relative risk of prostate cancer [6]. This is a substantial inherited component, and high for a complex disease.

One of the regions most consistently associated with prostate cancer risk is the 8q24 region. A review performed by Goh et al. reported six independent signals in this region [10]. SNPs in the 8q24 region have been reported to be associated with worse pathological tumor stage and biochemical relapse post-prostatectomy [11]. Multiple GWASs of cancer have demonstrated an association at the 8q24 region, including prostate, breast, colon, ovarian, and bladder cancers and chronic lymphocytic leukemia [11]. The 8q24 region includes a number of oncogenes. The MYC gene is a proto-oncogene that is hypothesized to be a target of 8q24 enhancers [12]. In a study investigating the association between genetic variants in the 8q24 region and different diseases, most of the SNPs found to be associated with prostate cancer were located in the intronic region (rs101643, rs1447295, rs16902094, rs445114) [10]. Prior studies have not investigated rare exonic variants in this region, which may have larger effects on risk for prostate cancer than common variants [4].

To further interrogate genetic variants associated with prostate cancer in the 8q24 region, defined as chr8:127575690 - 130446154 kb, we will use imputed dosage from genome-wide array and exome sequencing data to capture genetic variants in this region. Common variants from the genome-wide array platform will allow us to confirm in the Atherosclerosis Risk in Communities (ARIC) study common and lower frequency variants previously found to be associated with prostate cancer risk in European (see Table 1) and African American men (see Table 2) [17]. Use of the exome sequencing data gives the potential to identify variation underlying complex traits and could identify rare causal variants that are associated with prostate cancer risk in men with African and European ancestry. Genome-wide array data are available for ~9000 European American participants and ~3000 African American participants. Exome sequencing data are available for ~7800 European American participants and ~2500 African American participants.

Table 1. Prostate cancer risk variants at 8q24 in men of European ancestry
Table 2. Prostate cancer risk variants at 8q24 in men of African ancestry

<table>
<thead>
<tr>
<th>8q24 SNP [17]</th>
<th>Position (build 37)</th>
<th>OR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs7816007</td>
<td>128012359</td>
<td>1.21 (1.12 to 1.30)</td>
</tr>
<tr>
<td>rs114798100</td>
<td>128085434</td>
<td>2.32 (2.02 to 2.66)</td>
</tr>
<tr>
<td>rs111906932</td>
<td>128086204</td>
<td>1.72 (1.45 to 2.03)</td>
</tr>
<tr>
<td>rs72725879</td>
<td>128103969</td>
<td>1.38 (1.29 to 1.47)</td>
</tr>
<tr>
<td>rs2445605</td>
<td>128161944</td>
<td>1.30 (1.18 to 1.44)</td>
</tr>
<tr>
<td>rs7824868</td>
<td>128524414</td>
<td>1.43 (1.25 to 1.64)</td>
</tr>
<tr>
<td>rs11784480</td>
<td>1286259</td>
<td>1.18 (1.09 to 1.28)</td>
</tr>
</tbody>
</table>

5. Main Hypothesis/Study Questions:

1. Evaluate the association between common and low frequency single-nucleotide variants from GWAS (observed and imputed) in the 8q24 region and prostate cancer risk, overall and in men of African and European ancestry. This evaluation will be performed to confirm prior findings (Tables 1 and 2).

2. Evaluate the association between putative functional variants in the 8q24 region from exome sequencing and prostate cancer risk. Putative functional variants are defined as missense, frameshift, stopgain, stoploss, and splice variants. If variants with large effect are identified, these variants may contribute to risk stratification in prostate cancer screen.

6. Design and analysis (study design, inclusion/exclusion, outcome and other variables of interest with specific reference to the time of their collection, summary of data analysis, and any anticipated methodologic limitations or challenges if present).

**Study design:** Prospective cohort

**Inclusion/exclusion**
All male ARIC participants with adequate genotyping data from Affymetrix 6.0 and exome sequencing and who are at risk for a first primary prostate cancer will be included in this study. We will exclude all participants who did not provide informed consent to DNA usage for future studies and/or to studies of other chronic diseases like cancer.
Exposure: We will define the 8q24 region as chr8:127575690 - 130446154 kb, which encompasses 500 kb on either side of the SNPs found in the GWAS Catalogue [18] and in Oh et al. 2017 in the 8q24 region as being associated with prostate cancer. In this region, based on the common variants in genome-wide array platform, we will capture genotyped and imputed common and low frequency SNPs, including MYC and POU5F1B. We have exome sequencing data available for 64% of men in the ARIC prostate cancer file (which includes men at risk and prostate cancer cases). In this region, exome sequencing has captured variants in MYC and POU5F1B. SNPs that are associated with prostate cancer risk in the 8q24 region, thus far, are intergenic. We will exclude synonymous SNPs in favor of a more focused analysis.

Outcome: First primary prostate cancer

We will use the ARIC prostate cancer case file 1987-2012. Prostate cancer incidence data was obtained from cancer registries, hospital records, and death certificates.

Covariates: Age, prostate cancer family history (from Visit 3), genetic principal components to control for population stratification.

We will use the following factors to characterize the men and for considering effect modification: smoking status at baseline (never, former quit >10 years ago, current/quit within 10 years), body mass index, physical activity, diabetes, height, aspirin use, and statin drug use.

Statistical analysis

Statistical model: Cox proportional hazards model

Genetic model: For genotyped and imputed 8q24 variants from the GWAS, we use an additive genetic model, coded 0 to 2 to represent the estimated number of alleles. For the variants identifiable from exome sequencing, we will also use the additive genetic model.

We will analyze European American and African Americans separately to reduce population stratification. If we can confidently control for population stratification, we will then perform meta-analysis to combine the results from the two race groups.

Genome-wide array and exome sequencing data

In ARIC, genotyping was conducted using Affymetrix 6.0 and subsequently imputed based on 1000 Genome Reference Panel Integrated variant set release (v3) in NCBI build 37 (hg19) with high quality SNPs using IMPUTE2.

Exome sequencing was conducted using Illumina HiSeq 2000 or the HiSeq 2500 platform using the HGSC Mercury analysis pipeline (https://www.hgsc.bcm.edu/content/mercury). Variants were filtered by a set of quality control criteria, including low posterior probability (< 0.95), low read count (<3), and low total coverage (<10 fold). Samples were excluded for missingness > 20%, singleton count, heterozygote to homozygote ratio, or Ti/Tv ratio > 6 SD.
**Association analysis**

We will use single-variant tests to investigate the associations of common variants with prostate cancer risk. Because single variant tests do not have sufficient power to detect low frequency variants, we will combine exonic variants and use aggregation-based tests of association. The genomic region for aggregation will be at the gene level. Aggregation tests will be conducted using the optimal sequence kernel association tests (SKAT-O). The SKAT-O will compute a p-value after evaluating the association from a burden test, which assumes all alleles influencing the association in the same direction, or a SKAT (variance component) test that accounts for risk and protective alleles to maximize power [6]. The SKAT-O procedure is more powerful than running the burden test and SKAT test separately, as the SKAT-O approximates the test by using an optimal value of $p$ estimated. Taking the minimum p-value of either the burden test or SKAT test can lead to an inflated type I error rate, and we can correct this by using an omnibus test such as the SKAT-O that analytically calculates the optimal p-value [16]. Only putative functional variants as defined above with MAF < 5% will be included in SKAT-O tests, as putative functional variants are more likely to be causal. We will use the seqMeta package, which compute summary statistics for each variant accounting for their correlation at the gene level.

**Follow-up analysis**

For significant genes or regions found within the scope of our region chr8:127575690 – 130446154 from the SKAT-O test, we will further perform:

1) Burden and SKAT tests to evaluate whether the associations are mainly in one direction
2) Additional analysis to determine whether the associations are driven by one or two variants or by a large number of variants within the gene.
3) Haplotype testing by estimating haplotype frequencies after imputing a likelihood for the observed genotypes. We will calculate the maximum likelihood for the combined sample, and then those with African ancestry separately from those with European ancestry.

If significant novel variants are identified, we will determine whether the association is independent of common variants that are known to be associated with prostate cancer risk.

**Statistical significance threshold**

We will account for multiple testing using the Bonferroni method to adjust our significance cut-off at $\alpha/n$:

- Single variant test: 0.05 divided by the number of common variants tested.
- Aggregation-based test: 0.05 divided by the number of tests.

**Power analysis**

Based on the inclusion and exclusion criteria, for this analysis, we have 382 prostate cancer cases among men of European ancestry and 214 prostate cancer cases among men of African ancestry. We used stpower cox in Stata to calculate power for single variant test assuming an alpha level of 0.05 and MAF of 5%, 10%, or 25% (common SNP genotyped or imputed). The minimum
detectable hazard ratios (HR) ranged from 1.88 for MAF of 5% to 1.27 for MAF of 25% for men of European ancestry and 2.30 for MAF of 5% to 1.37 for MAF of 25% for men of African ancestry (Table 3). These minimum detectable HRs are within the range of the odds ratios associated prostate cancer from known genetic variants (from 1.18 to 2.90, Tables 1 and 2). Rare exonic variants may have larger effect size [4].

Table 3. Minimum detectable hazard ratio assuming an alpha of 0.05 and 80% power.

<table>
<thead>
<tr>
<th>MAF</th>
<th>EA (cases=382)</th>
<th>AA (cases=214)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5%</td>
<td>1.88</td>
<td>2.30</td>
</tr>
<tr>
<td>10%</td>
<td>1.50</td>
<td>1.71</td>
</tr>
<tr>
<td>25%</td>
<td>1.27</td>
<td>1.37</td>
</tr>
</tbody>
</table>

All statistical analyses will be performed using R and STATA statistical software.

7.a. Will the data be used for non-CVD analysis in this manuscript?

X Yes  No

b. If Yes, is the author aware that the file ICTDER03 must be used to exclude persons with a value RES_OTH = “CVD Research” for non-DNA analysis, and for DNA analysis RES_DNA = “CVD Research” would be used?

X Yes  No
(This file ICTDER has been distributed to ARIC PIs, and contains the responses to consent updates related to stored sample use for research.)

8.a. Will the DNA data be used in this manuscript?

X Yes  No

8.b. If yes, is the author aware that either DNA data distributed by the Coordinating Center must be used, or the file ICTDER03 must be used to exclude those with value RES_DNA = “No use/storage DNA”?

X Yes  No

9. The lead author of this manuscript proposal has reviewed the list of existing ARIC Study manuscript proposals and has found no overlap between this proposal and previously approved manuscript proposals either published or still in active status. ARIC Investigators have access to the publications lists under the Study Members Area of the web site at: http://www.csc.unc.edu/ARIC/search.php

X Yes  No

10. What are the most related manuscript proposals in ARIC (authors are encouraged to contact lead authors of these proposals for comments on the new proposal or collaboration)?
11.a. Is this manuscript proposal associated with any ARIC ancillary studies or use any ancillary study data?

_X__ Yes  ___ No

11.b. If yes, is the proposal

_X__ A. primarily the result of an ancillary study (list number* 1995.04, 2011.07)

___ B. primarily based on ARIC data with ancillary data playing a minor role (usually control variables; list number(s)* __________ __________ __________)

*ancillary studies are listed by number at http://www.csc.c.unc.edu/aric/forms/

12a. Manuscript preparation is expected to be completed in one to three years. If a manuscript is not submitted for ARIC review at the end of the 3-years from the date of the approval, the manuscript proposal will expire.

12b. The NIH instituted a Public Access Policy in April, 2008 which ensures that the public has access to the published results of NIH funded research. It is your responsibility to upload manuscripts to PUBMED Central whenever the journal does not and be in compliance with this policy. Four files about the public access policy from http://publicaccess.nih.gov/ are posted in http://www.csc.c.unc.edu/aric/index.php, under Publications, Policies & Forms. http://publicaccess.nih.gov/submit_process_journals.htm shows you which journals automatically upload articles to Pubmed central.

REFERENCES


