A Custom Genotyping Array Reveals Population-Level Heterogeneity for the Genetic Risks of Prostate Cancer and Other Cancers in Africa

Maxine Harlemon[1,2], Olabode Ajayi[3], Paidamoyo Kachambwa[3], Michelle S. Kim[1], Corinne N. Simonti[1], Melanie H. Quiver[1], Desiree C. Petersen[3], Anuradha Mittal[4], Pedro W. Fernandez[5], Ann W. Hsing[6], Shakuntala Baichoo[7], Ilir Agalliu[8], Mohamed Jalloh[9], Serigne M. Gueye[9], Nana Yaa F. Snyper[10], Ben Adusei[10], James E. Mensah[11], Afua O.D. Abrahams[11], Akindele O. Adebiyi[12], Akin T. Orunmuyi[12], Oseremen I. Aisuodionoe-Shadrach[13], Maxwell M. Nwegbu[13], Maureen Joffe[14,15], Wenlong C. Chen[16,17], Hayley Irusen[5], Alfred I. Neugut[18], Yuri Quintana[19], Moleboheng Seutloali[3], Mayowa B. Fadipe[3], Christopher Warren[4], Marcos H. Woehrmann[4], Peng Zhang[20], Chrissie M. Ongaco[20], Michelle Mawhinney[20], Jo McBride[3], Caroline V. Andrews[21], Marcia Adams[20], Elizabeth Pugh[20], Timothy R. Rebbeck[21,22], Lindsay N. Petersen[3], and Joseph Lachance[1]

ABSTRACT

Although prostate cancer is the leading cause of cancer mortality for African men, the vast majority of known disease associations have been detected in European study cohorts. Furthermore, most genome-wide association studies have used genotyping arrays that are hindered by SNP ascertainment bias. To overcome these disparities in genomic medicine, the Men of African Descent and Carcinoma of the Prostate (MADCaP) Network has developed a genotyping array that is optimized for African populations. The MADCaP Array contains more than 1.5 million markers and an imputation backbone that successfully tags over 94% of common genetic variants in African populations. This array also has a high density of markers in genomic regions associated with cancer susceptibility, including 8q24. We assessed the effectiveness of the MADCaP Array by genotyping 399 prostate cancer cases and 403 controls from seven urban study sites in sub-Saharan Africa. Samples from Ghana and Nigeria clustered together, whereas samples from Senegal and South Africa yielded distinct ancestry clusters. Using the MADCaP array, we identified cancer-associated loci that have large allele frequency differences across African populations. Polygenic risk scores for prostate cancer were higher in Nigeria than in Senegal. In summary, individual and population-level differences in prostate cancer risk were revealed using a novel genotyping array.

Significance: This study presents an Africa-specific genotyping array which enables investigators to identify novel disease associations and to fine-map genetic loci that are associated with prostate and other cancers.

Introduction

Prostate cancer is a complex disease that disproportionally affects men of African descent (1). Prostate cancer is the leading cause of cancer-related mortality in African men (2). In the United States, African American men have a higher risk of developing prostate cancer and an even higher increased risk of dying from it compared with men of European or Asian descent (3). In the United Kingdom, men of African descent have an increased risk of being diagnosed and dying from prostate cancer (4). Furthermore, the highest reported mortality rates of prostate cancer are found in Caribbean men of African descent (5). Multiple socioeconomic, environmental, and genetic factors contribute to this health inequity.

Cancer is considered a genetic disease, and prostate cancer has a heritability of 58% (6). Risks of prostate cancer run in families; the relative risk of men with affected fathers is 2.1-fold higher compared

[1]School of Biological Sciences, Georgia Institute of Technology, Atlanta, Georgia. [2]Clark Atlanta University, Atlanta, Georgia. [3]Centre for Proteomic and Genomic Research, Cape Town, South Africa. [4]Thermo Fisher Scientific, Santa Clara, California. [5]Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa. [6]Stanford Cancer Institute, Stanford University, Stanford, California. [7]University of Mauritius, Réduit, Mauritius. [8]Albert Einstein College of Medicine, Bronx, New York. [9]Hôpital Général de Grand Yoff, Institut de Formation et de Recherche en Urologie et Santé Familiale, Dakar, Senegal. [10]37 Military Hospital, Accra, Ghana. [11]Korle-Bu Teaching Hospital and University of Ghana, Accra, Ghana. [12]College of Medicine, University of Ibadan, Ibadan, Nigeria. [13]College of Health Sciences, University of Abuja and University of Abuja Teaching Hospital, Abuja, Nigeria. [14]Non-Communicable Diseases Research Division, Wits Health Consortium (PTY) Ltd, Johannesburg, South Africa. [15]MRC Developmental Pathways to Health Research Unit, Department of Paediatrics, Faculty of Health Sciences, University of Witwatersrand, Johannesburg, South Africa. [16]Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa. [17]National Cancer Registry, National Health Laboratory Service, Johannesburg, South Africa. [18]Herbert Irving Comprehensive Cancer Center, Columbia University, New York, New York. [19]Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, Massachusetts. [20]Center for Inherited Disease Research, Johns Hopkins University, Baltimore, Maryland. [21]Dana-Farber Cancer Institute, Boston, Massachusetts. [22]Harvard T.H. Chan School of Public Health, Boston, Massachusetts.

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

Corresponding Author: Joseph Lachance, Georgia Institute of Technology, 950 Atlantic Dr., Atlanta, GA 30332. Phone: 404-894-0794; Fax: 404-894-0519; E-mail: joseph.lachance@biology.gatech.edu

Cancer Res 2020;XX:XX–XX

doi: 10.1158/0008-5472.CAN-19-2165

with men without a family history (7). In recent years, multiple genome-wide association studies (GWAS) have detected genetic associations with prostate cancer (8–11). Collectively, these studies have yielded over 200 independent prostate cancer risk–associated loci, and one key genomic region that has been repeatedly implicated in prostate cancer and other cancers is 8q24 (12–14). The most comprehensive prostate cancer GWAS analyzed to date included over 140,000 cases and controls of European ancestry (10). In this study, Schumacher and colleagues generated a polygenic risk score (PRS) that successfully classified individuals into high- and low-risk categories.

Unfortunately, most GWAS have not focused on populations from sub-Saharan Africa. As of 2016, 81% of all GWAS samples were of European ancestry, and 14% of all GWAS samples were of East Asian ancestry (15). In addition, existing genotyping arrays do not adequately capture genetic variation in diverse African populations. Both of the aforementioned issues limit what is known about cancer genetics in African populations. Thus, the current set of known disease-associated loci is enriched for alleles with an intermediate frequency in Europe or Asia, but not Africa. This lack of representation can exacerbate existing health disparities by not capturing relevant genetic risk associations in African populations (16, 17). Disease associations do not always replicate across populations, and the directions of risk associations at cancer-associated loci may differ across study cohorts (18, 19). Previous work also indicates that risks of prostate cancer vary by genetic ancestry across the globe (20, 21). For example, a GWAS in Ghana showed that the most promising SNP in this African population has not been identified in other populations (22). Hence, there is a clear need for more studies that analyze the genetics of African populations (23, 24). Commonly used genotyping arrays tend to use markers that were originally ascertained in European populations (25). This can cause polygenic risks of complex diseases, including prostate cancer, to be wrongly estimated (26). For example, the OncoArray Consortium has developed an array with over 500,000 markers, half of which are in genomic regions that tag cancer susceptibility (27). However, the OncoArray is not enriched for African polymorphisms. By contrast, the H3Africa Consortium has developed an array that includes over 2 million markers (28), but the H3Africa Array was not specifically designed for cancer studies. Existing arrays may therefore be suboptimal for detecting cancer associations in African populations.

To address this problem, the Men of African Descent and Carcinoma of the Prostate (MADCaP) Network (29) developed a customized genotyping array optimized for fine-mapping and detecting novel associations with prostate cancer in African populations. The MADCaP array will ultimately be used in an African GWAS containing over 6,000 cases and controls. Here, we analyze a pilot dataset of over 800 individuals from sub-Saharan Africa. In this article, we compare multiple genotyping platforms and test the efficacy of the MADCaP Array by genotyping samples from seven African study sites. Using data derived from the MADCaP Array, we also infer population structure, identify cancer-associated loci that have large allele frequency differences across Africa, and quantify polygenic risks of prostate cancer in African populations.

Materials and Methods

Inclusion criteria for markers

The MADCaP Array was developed using the Applied Biosystems Axiom genotyping solution from Affymetrix/Thermo Fisher Scientific. This array consists of a two-peg design. Multiple inclusion criteria were used for markers on the MADCaP Array, including: enrichment for GWAS loci, markers near cancer susceptibility loci, prostate expression quantitative trait loci (eQTL), markers found on other arrays, and markers tagging African polymorphisms. Note that 38,649 unique markers that are associated with traits and diseases from the NHGRI-EBI GWAS Catalog are included on the MADCaP Array (30). Using 1000 Genomes Project (31) data, we included every SNP with an African minor allele frequency (MAF) >0.05 that was located within 50 kb of a known prostate cancer hit or within 5 kb of other cancer associations. We used the Genotype-Tissue Expression (GTEx V7) project (32) to identify SNPs that modify gene expression in the prostate (i.e., prostate eQTLs). The MADCaP Array contains 24,595 prostate eQTLs (P value cutoff for inclusion: 10^{-9}). Markers were also preferentially included if they overlapped the OncoArray or H3Africa Array. Working with Thermo Fisher Scientific, a GWAS backbone was built using Applied Biosystems Axiom genotyping array technology by iteratively selecting markers that maximized the ability to impute African genetic variation. When possible, we used probes that had a prior track record of working on existing genotyping arrays. Multiple probes per marker were included for prostate cancer loci and unvalidated markers. An overlapping set of more than 1000 markers was chosen to be on both pegs, with priority given to prostate cancer loci and markers satisfying multiple inclusion criteria. Supplementary Table S1 lists successfully called markers on the MADCaP Array.

Assessment of imputation performance

Imputation performance of the MADCaP Array was computed using the African Genome Resource reference panel, comprising of whole-genome sequence data from 4,956 individuals and 11 populations of African descent (33). We classified African polymorphisms with an MAF >0.05 as common SNPs and African polymorphisms with an MAF between 0.01 and 0.05 as rare SNPs. Imputation was performed with IMPUTE2 (v2.3.2) software using 10-fold cross-validation (34). Coverage in each population was calculated as the proportion of polymorphisms in the African Genome Resource reference panel in high LD ($r^2 \geq 0.8$) with markers on the MADCaP Array.

Biospecimen and DNA quantification

Biospecimens were obtained with informed consent using protocols approved from each study site's Institutional Review Board/Ethics Review Board. Written-informed consent was obtained from patients, and studies were conducted in concordance with recognized ethical guidelines (the Declaration of Helsinki and the U.S. Common Rule). Blood samples were collected in EDTA vacutainer tubes and stored at either −20°C or −80°C. DNA was isolated using QIAamp DNA Blood kits. A total of 1.8 to 3.0 μg high-purity DNA at a concentration of 30 to 50 ng/μL per sample was submitted for genotyping. DNA was transferred from study sites to genotyping laboratories using BioMatrica DNAStable 2D barcoded plates. Samples were then rearrayed into plates using a BioMicroLab XL20 at a minimum concentration of 10 ng/μL in 50 μL. All samples were run on the Infinium QC array and the MADCaP Array. Plate maps used a randomized block design to control for study site and case versus control status.

SNP calling, QC, and data curation

Standard quality control (QC) procedures for Axiom genotyping data analysis were performed (35, 36). Sample preprocessing was performed according to guidelines provided in the Thermo Fisher Scientific Axiom Genotyping Solution Data Analysis Guide (36). The custom MADCaP Array is based on a two-peg design. Peg 1 contains

852,610 probe sets, covering 801,275 markers. Peg 2 contains 790,524 probe sets, covering 790,170 markers. Note that 1,902 probe sets overlap both pegs. Raw data CEL files, representing more than 802 samples, as well as 28 technical replicates and additional controls, were imported into the Axiom Analysis Suite (AxAS) version 4.0.3.3 for filtering of sample call rate and clustering of SNP genotype calls. Samples with DishQC ≥0.82 and a QC call rate >97% were included for downstream genotyping analysis. A full list of Axiom QC thresholds is included in Supplementary Table S2. The Centre for Proteomics and Genomics Research in Cape Town, South Africa, and the Center for Inherited Disease Research at Johns Hopkins University independently assessed QC metrics for each probe set.

Data from both pegs of the MADCaP Array were merged. PLINK was used to remove markers with low call rates (marker missingness > 5%). PLINK was also used to exclude samples that were poorly called (sample missingness > 2%) or related (identity-by-descent > 0.5). Multi-allelic SNPs were excluded from downstream analyses. After filtering, 1,513,172 markers and 802 samples were used in subsequent analyses. This MADCaP pilot dataset contains 399 prostate cancer cases and 403 controls. Details of MADCaP case and control recruitment have been previously reported (29).

Array comparisons

We compared markers on arrays developed by the MADCaP Network, the OncoArray Consortium (27), and the H3Africa Consortium (28). Genomic positions from the MADCaP Array, Infinium Oncoarray, and H3Africa Array were intersected to determine overlapping markers. The *liftOver* bioinformatics tool was used to convert all genomic positions to build GRCh38/hg38 of the human reference genome. Derived allele frequencies (DAF) for each array were calculated as described previously (26). This involved obtaining allele frequencies from the five continental populations of the 1000 Genomes Project (31): Africa (AFR), Americas (AMR), East Asia (EAS), Europe (EUR), and South Asia (SAS). Calculations used 450,000 markers that were selected without replacement from each array (only markers with 1000 Genomes Project data were considered here). The joint allele frequency distribution of markers on the MADCaP Array was found by comparing African and pooled non-African data.

MDS and ADMIXTURE

Using PLINK, we obtained an LD-pruned subset of 25,000 autosomal SNPs for 802 samples (MAF > 0.05, $r^2 < 0.8$). The same subset of SNPs was used for multidimensional scaling (MDS) and ADMIXTURE analyses. Two-dimensional MDS plots were generated using PLINK and R. ADMIXTURE software (37) was run for $K = 2$ through $K = 5$. Cross-validation was performed to determine the optimal K value.

Runs of homozygosity and LD decay

As per Schlebusch and colleagues (38), runs of homozygosity were identified using PLINK for homozygous lengths between 500 kb and 1,000 kb. This analysis was repeated for all 802 samples in the MADCaP dataset. Individual runs of homozygosity were summed to yield the cumulative runs of homozygosity (cROH) for each sample. For each study site, PLINK v1.90b6.9 was used to calculate LD between all variants with an MAF >0.10. These calculations were made for all pairs of markers within 100 kb and 100 marker windows. For each study site, the mean r^2 between pairs of genetic variants was calculated for 1 kb bins.

Identification of divergent loci via population branch statistics calculations

Population branch statistics (PBS) were calculated as per Yi and colleagues (39). Data from multiple MADCaP study sites were pooled to yield allele frequencies for three populations: Senegal (HOGGY), Ghana & Nigeria (37 Military, KBTH, UATH, and UCH), and South Africa (WITS and SU). Genetic distances between pairs of populations were calculated using Weir and Cockerham's F_{st} (40). We then calculated PBS scores for three different branches:

$$PBS_{Senegal} = \frac{-\ln\left(1 - F_{ST,sn-za}\right) - \ln\left(1 - F_{ST,sn-gh\&ng}\right) + \ln\left(1 - F_{ST,za-gn\&ng}\right)}{2} \quad (A)$$

$$PBS_{Ghana\ \&\ Nigeria} = \frac{-\ln\left(1 - F_{ST,sn-gh\&ng}\right) - \ln\left(1 - F_{ST,za-gh\&ng}\right) + \ln\left(1 - F_{ST,sn-za}\right)}{2} \quad (B)$$

$$PBS_{South\ Africa} = \frac{-\ln\left(1 - F_{ST,sn-za}\right) - \ln\left(1 - F_{ST,za-gh\&ng}\right) + \ln\left(1 - F_{ST,sn-gh\&ng}\right)}{2} \quad (C)$$

Subscripts refer to country codes: *sn* for Senegal, *gh* for Ghana, *ng* for Nigeria, and *za* for South Africa. Undefined and negative values of Weir and Cockerham's F_{st} were treated as zero, and undefined or negative PBS scores were also treated as zero. PBS scores were calculated for 2,477 unique markers from the NHGRI-EBI GWAS Catalog (30) that yield 3,557 cancer and cancer-related associations.

Calculation of PRS

PRS were built using a curated set of 139 prostate cancer–associated loci. Schumacher and colleagues previously developed a 147-loci PRS for prostate cancer (10), and 116 of these 147 markers are on the MADCaP Array. Proxies were found for 23 of the remaining 31 markers by identifying markers on the MADCaP array in LD with loci from the Schumacher PRS ($r^2 > 0.4$). Alleles at proxy markers that tag increased prostate cancer risk were inferred using LDlink (41). Supplementary Table S3 lists markers used to generate the PRS described here.

As per Schumacher and colleagues (10), effect size information was incorporated into PRS calculations. For each locus, we counted whether an individual has 0, 1, or 2 copies of the risk-increasing allele (i.e., the allele dose $g_{i,j}$ for locus i in individual j). Here, we used adjusted effect sizes: $\beta_i = \ln(OR_i) \times {r_i}^2$, where effect sizes from Schumacher and colleagues (10) are scaled by how well proxy markers tag each disease-associated locus. Doses of risk-increasing alleles were weighted by adjusted effect sizes and summed across all 139 loci to generate a raw PRS for each individual.

$$PRS_j = \sum_{i=1}^{139} g_{i,j}\beta_i \quad (D)$$

PRS were calculated for 802 MADCaP samples and 240 male European samples from 1000 Genomes Project (31). Standardized

PRS values were then generated for all 1,042 individuals by scaling raw PRS values to have a mean of zero and SD of one.

Statistical methods

Wilcoxon rank-sum tests were used to compare DAF distributions, ADMIXTURE proportions, runs of homozygosity, and PRS distributions. We corrected for multiple statistical tests using the Benjamini–Hochberg approach (42). An FDR of 0.05 was used to generate adjusted *P* values shown in Supplementary Table S2. Two sample Z-tests were used to compare allele frequencies for different African countries.

Results

Imputation using the MADCaP Array

Using whole-genome sequences, we quantified the extent to which the MADCaP array tags African genetic variation (**Fig. 1**). Depending on the population, 94% to 99% of common African SNPs were successfully tagged by the MADCaP Array ($r^2 \geq 0.8$, MAF > 0.05). The MADCaP Array also tagged 63% to 97% of rare African SNPs ($r^2 \geq 0.8$, MAF between 0.01 and 0.05). Note that 98% of common variants and 79% to 83% of rare variants in admixed African-Caribbean and African-American genomes are in high-LD ($r^2 \geq 0.8$) with markers on the MADCaP Array. A larger fraction of Ugandan genetic variation than Ethiopian and KhoeSan variation is captured by this array. Regardless of population, the MADCaP Array successfully tags a large fraction of African genetic variation.

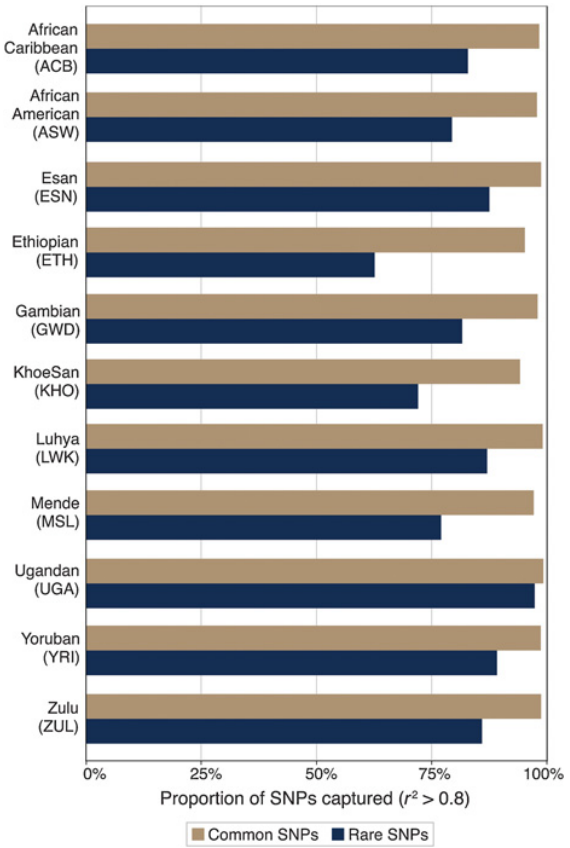

Figure 1.
Successful tagging of African SNPs using the MADCaP Array. Proportions of SNPs in 11 populations of African ancestry that are successfully tagged by markers on the MADCaP Array are shown ($r^2 \geq 0.8$). Here, common SNPs have an MAF >0.05, and rare SNPs have an MAF between 0.01 and 0.05.

Comparisons with other arrays

Many of the markers on the MADCaP Array are shared with the Infinium OncoArray and the H3Africa Array (**Fig. 2A**). Overall, 73,019 markers are included on all three arrays. Note that 131,469 markers are shared between the MADCaP Array and the OncoArray, and 398,460 markers are shared between the MADCaP Array and the H3Africa Array. This overlap will facilitate data harmonization and the ability to combine genotype information from different arrays into the same study.

We compared the DAF of markers found on the MADCaP, OncoArray, and H3Africa arrays, using continental allele frequencies from the 1000 Genomes Project (**Fig. 2B**). The null expectation here is that the mean DAF should be the nearly identical for each population because all humans are evolutionarily equidistant to other primates. Mean DAFs of markers on the MADCaP Array are similar for each continental population (**Fig. 2B**), suggesting that the MADCaP Array is relatively unbiased with respect to SNP selection. By contrast, mean DAFs of markers on the OncoArray and H3Africa Array are notably lower for African populations than non-African populations—a pattern that is indicative of SNP ascertainment bias (26). Differences between populations are statistically significant (Benjamini–Hochberg adjusted *P* values $>2.2 \times 10^{-16}$, Wilcoxon rank-sum tests), and mean and median DAFs are listed in Supplementary Table S2. Examining the joint site frequency spectrum of non-African and African populations, we find that the MADCaP Array is enriched for markers that are polymorphic in Africa but monomorphic outside of Africa, but not vice versa (**Fig. 2C**).

Densities of markers found in different genomic regions vary by genotyping array. Here, we focus on 8q24, a cancer-associated genomic region that contains *PCAT2, CCAT2,* and the proto-oncogene *MYC.* Numbers of markers per 100 kb are shown for three different arrays in **Fig. 2D**. The MADCaP Array contains a moderately high density of markers across the genome, with peaks near known cancer-associated loci. Neighboring markers on the MADCaP Array have a median distance of 856 bp and a mean distance of 2,082 bp. The OncoArray has higher marker densities near cancer-associated loci compared with other parts of the genome. By contrast, the H3Africa Array has a moderately even density of markers across the entire genome. A total of 3,082 markers on the MADCaP Array, 2,349 markers on the H3Africa Array, and 1,057 markers on the OncoArray overlap known cancer associations from the NHGRI-EBI GWAS Catalog (accessed September 11, 2019). Focusing on 100 kb genomic windows centered around 147 known prostate cancer loci (10), we find that the MADCaP array contains 28,422 markers flanking prostate cancer loci, as opposed to 14,959 flanking markers on the H3Africa Array and 11,290 flanking markers on the OncoArray (Supplementary Table S3)

Efficacy of the MADCaP Array

We tested the efficacy of the MADCaP Array by genotyping 802 individuals from seven MADCaP study sites (**Fig. 3A**): the Hôpital Général de Grand Yoff/Institut de Formation et de Recherche en Urologie in Dakar, Senegal (HOGGY), 37 Military Hospital in Accra, Ghana (37 Military), Korle-Bu Teaching Hospital in Accra, Ghana (KBTH), University College Hospital in Ibadan, Nigeria (UCH), University of Abuja Teaching Hospital in Abuja, Nigeria (UATH),

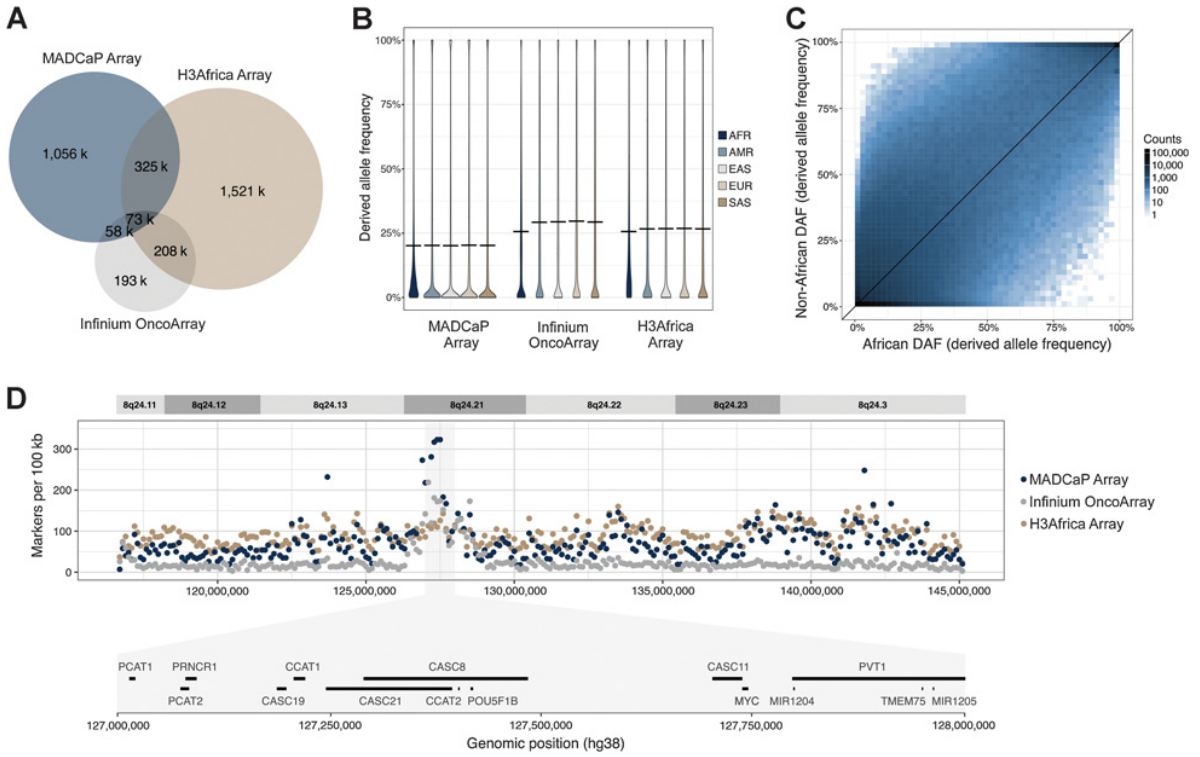

Figure 2.

Comparisons between the MADCaP Array, Infinium OncoArray, and H3Africa Array. **A,** Venn diagram showing overlap between markers on each array. Sizes of circles are proportional to the number of markers on each array. **B,** Violin plots indicate DAF distributions of markers on the MADCaP Array, Infinium OncoArray, and H3Africa Array. Continental allele frequencies from the 1000 Genomes Project are shown here. Horizontal black lines indicate the mean DAF for each array and population combination. **C,** Joint site frequency spectrum of markers on the MADCaP Array. African and pooled non-African allele frequencies from 1000 Genomes Project are shown here. Shading indicates the number of markers on the MADCaP Array that are in each bin. **D,** Density of markers per nonoverlapping 100 kb window at 8q24. Genes in the zoomed-in region are shown.

WITS Health Consortium/National Health Laboratory Services in Johannesburg, South Africa (WITS), and Stellenbosch University in Cape Town, South Africa (SU). Sample accrual was restricted to individuals with sub-Saharan African ancestry; admixed individuals with European ancestry from Cape Town were excluded. Of the MADCaP samples analyzed herein, 399 are prostate cancer cases and 403 are controls (**Table 1**).

Note that 94.9% of the markers on peg 1 and 95.9% of the markers on peg 2 passed QC filtering. By contrast, approximately 533,00 out of 600,000 (88.8%) markers on the OncoArray were successfully manufactured (27). For both peg 1 and peg 2, mean call rates, reproducibility, and concordance all exceeded 99.5%, and only a small subset of markers had Mendelian inconsistencies (Supplementary Table S2). These genotyping metrics indicate that the MADCaP Array is an effective genotyping platform.

Population structure and genetic admixture

We used two-dimensional MDS plots to detect population structure among MADCaP samples and study sites. Individuals with similar genomes are located close to one another in MDS space. MADCaP samples fall into three broad clusters in **Fig. 3B**: Senegalese individuals (gold) are found in the bottom left, Ghanaian and Nigerian individuals (green) are found in the top left, and South African individuals (blue) are found in the top right. Nigerians from Ibadan (UCH, light green) are closer in MDS space to Ghanaian individuals than Nigerians from Abuja (UATH, dark green). The right-to-left gradient of blue points in MDS space suggests that some individuals from South Africa share a fraction of their genetic ancestry with present-day Nigerians. Rotating the MDS plot 85 degrees clockwise reveals that genes mirror geography, at least for the African populations analyzed in our study (**Fig. 3C**). Samples from geographically close locations tend to share greater amounts of genetic similarity.

ADMIXTURE plots reveal shared ancestry among MADCaP samples (**Fig. 3D**). In these plots, individuals are linear mixtures of multiple genetic ancestries—indicated by different colors. Cross-validation error is minimized at K = 3, i.e., the best fit to the data occurs for three ancestry colors (Supplementary Fig. S1). At K = 2, we are able to distinguish between West African and South African populations. Setting K = 3 reveals three major ancestry clusters: gold in Senegal, green in Ghana and Nigeria, and blue in South Africa. At K = 4 ancestry patterns match each country. Intriguingly, individuals from Ibadan, Nigeria (UCH), share ancestry with samples from Ghana, i.e., they contain moderate amounts of light green ancestry at K = 4. Similarly, individuals from Johannesburg, South Africa, contain traces of genetic ancestry that are primarily found in Nigeria (dark green), perhaps due to the Bantu expansion during the last 5,000 years (43). K = 5 reveals evidence of population structure within South Africa, with greater proportions of light blue ancestry found

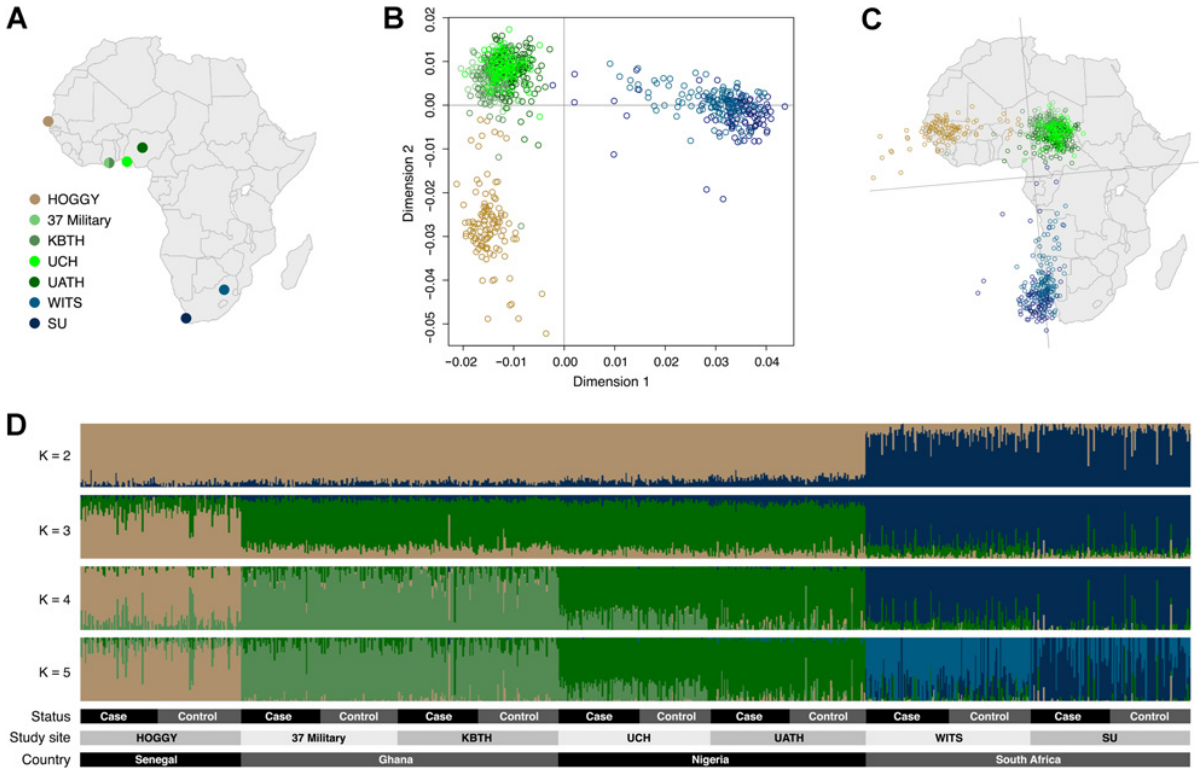

Figure 3.
The MADCaP Array reveals population structure and shared genetic ancestries among urban African study sites. **A,** Geographic locations of each MADCaP study site. **B,** Two-dimensional MDS plot of 802 MADCaP samples. Senegalese samples are represented by gold circles, Ghanaian and Nigerian samples are represented by green circles, and South African circles are represented by blue circles. **C,** Genes mirror geography when the two-dimensional MDS plot is rotated 85° clockwise. **D,** ADMIXTURE plot of 802 MADCaP samples. The best fit to genetic data occurs at K = 3.

in Johannesburg (WITS) compared with Cape Town (SU). Both study sites from Accra, Ghana (37 Military and KBTH), have similar genetic ancestry profiles. Finally, we note that cases and controls for each study site are ancestry-matched (Benjamini–Hochberg adjusted P values > 0.674, Wilcoxon rank-sum tests, Supplementary Table S2). On a genome-wide scale, individuals in the MADCaP study with prostate cancer have similar ancestry proportions compared with healthy MADCaP controls.

Runs of homozygosity and linkage disequilibrium

Runs of homozygosity are stretches of DNA where maternally and paternally inherited haplotypes are identical. Using the MADCaP array, we quantified cROH in each genome (**Fig. 4A**). Although there is heterogeneity within each study site, cROH are smaller for South African genomes than Senegalese, Ghanaian, or Nigerian genomes analyzed in this study (Benjamini–Hochberg adjusted P values < 0.0132, Wilcoxon rank-sum tests, Supplementary Table S2). This lower homozygosity can either be due to large historical population sizes or admixture.

To distinguish between each of these hypotheses, we calculated LD decay curves for each of the seven MADCaP study sites (**Fig. 4B**). Populations with small effective population sizes have more LD than populations with large effective population sizes (44). Admixture also increases the amount of LD (45). In general, we

Table 1. Numbers of cases and controls from each study site.

Study site	Location	Cases	Controls
Hôpital Général de Grand Yoff (HOGGY)	Dakar, Senegal	56	59
37 Military Hospital (37 Military)	Accra, Ghana	59	59
Korle-Bu Teaching Hospital (KBTH)	Accra, Ghana	53	58
University College Hospital (UCH)	Ibadan, Nigeria	56	56
University of Abuja Teaching Hospital (UATH)	Abuja, Nigeria	56	57
WITS Health Consortium (WITS)	Johannesburg, South Africa	61	61
Stellenbosch University (SU)	Cape Town, South Africa	58	53

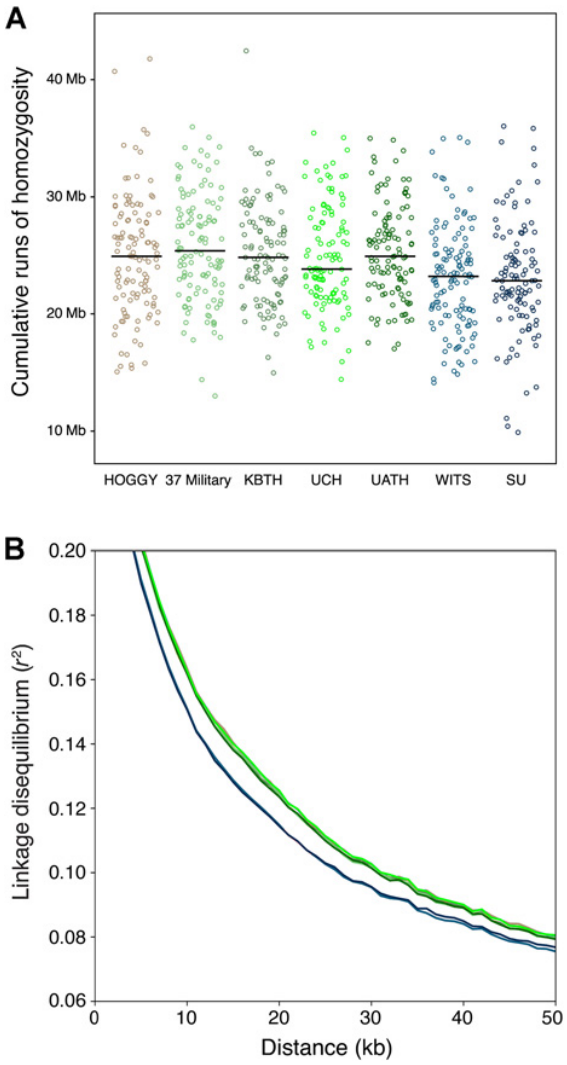

Figure 4.

Runs of homozygosity and LD decay curves vary by African study site. **A,** cROH 500 kb to 1,000 kb in length for each MADCaP sample, labeled by study site. Median cROH for each study site are indicated by horizontal lines in this jitter plot. **B,** LD decay curves for each study site. Gold indicates Senegalese data, green indicates Ghanaian and Nigerian data, and blue indicates South African data. South African study sites have less LD than West African study sites (Senegalese data overlaps Ghanaian and Nigerian data).

observed less LD for South African sites than other study sites (WITS and SU in **Fig. 4B**). These differences in LD decay curves do not appear to be due to admixture, because the South African populations studied here have similar levels of admixture to other African populations (**Fig. 3D**). Overall, the data in **Fig. 4** support the idea that historic population sizes were larger in South Africa than West Africa. One implication of the smaller haplotype blocks that are found in genomes from Johannesburg and Cape Town is that GWAS using these samples will require arrays with high densities of markers, a characteristic that is shared by the MADCaP Array.

Divergent allele frequencies at cancer-associated loci

Risk allele frequencies at cancer-associated loci can vary across populations. Here, the MADCaP Array was used to identify previously published loci that have large allele frequency differences across Africa. PBS scores were calculated for all observed cancer-associated loci in the NHGRI-EBI GWAS Catalog that have markers on the MADCaP Array, as well as loci associated with other cancer-related traits (e.g., skin pigmentation and smoking). These scores were calculated for three different evolutionary branches: Senegal (**Fig. 5A**), Ghana & Nigeria (**Fig. 5B**), and South Africa (**Fig. 5C**). Here, prostate cancer hits used in PRS calculations are represented by black points, whereas gray and colored points indicate other cancer-associated loci. Note that *Y* axes in these Manhattan plots quantify evolutionary branch lengths, not statistical strengths of association. Also note that 99th percentiles of PBS scores for all markers on the MADCaP array are represented by dashed lines. Supplementary Table S3 contains PBS scores and allele frequencies for 139 prostate cancer markers used in MADCaP PRS calculations. Supplementary Table S4 contains PBS scores and allele frequencies for 2,477 markers that are associated with cancer and cancer-related traits.

All three branches contain loci with high PBS scores in the MHC/HLA region on chromosome 6. For example, rs3817963 has the top PBS score for the Senegalese branch (**Fig. 5A**). This SNP at 6p21.32 has been associated with lung adenocarcinoma (46). The risk-increasing allele at rs3817963 has an allele frequency of 33.9% in Senegal, 12.9% in Ghana, 10.4% in Nigeria, and 8.4% in South Africa (P values <0.0001 for pairwise comparisons between Senegal and other countries, two sample Z-test). Another cancer-associated variant that has large allele frequency differences between African populations is rs2294008, located at 8q24.3. This SNP has the second highest PBS score for the South African branch, and it has previously been associated with bladder and gastric cancer (47). The risk-increasing allele at rs2294008 has an allele frequency of 28.7% in Senegal, 35.7% in Ghana, 28.8% in Nigeria, and 54.8% in South Africa (P values <0.0001 for pairwise comparisons between South Africa and other countries, two sample Z-test).

Some previously known prostate cancer–associated loci have large allele frequency differences between African populations, whereas other loci have allele frequencies that vary little across the continent. For example, rs5919432 is a prostate cancer–associated SNP that is located 71 kb from the *androgen receptor* gene at Xq12 (48). This SNP has the highest X-linked PBS score in **Fig. 5C**. The risk-increasing T allele at rs5919432 is more common in MADCaP cases than controls (34.2% vs. 32.1%), and South Africans have elevated risk allele frequencies at this SNP (**Fig. 5D**). 8q24.21 contains multiple loci that have been associated with prostate cancer in European men, including rs6983267 (10). Although the risk-increasing G allele at rs6983267 is more common in MADCaP cases than controls (98.2% vs. 97.9%), there are only minimal allele frequency differences between African populations at this SNP (**Fig. 5E**). The risk allele at rs6983267 is found at 50.0% in Europe (1000 Genomes Project data). This pattern suggests that although rs6983267 contributes to continental-level differences in prostate cancer risk, it has only a minimal effect on population-level differences in prostate cancer risk within sub-Saharan Africa.

The MADCaP pilot dataset also yields novel prostate cancer associations to be followed-up in subsequent studies. The SNP with the largest allele frequency difference between cases and controls is rs7063314 (located near the SPANX family of spermatogenesis genes at Xq27.2). MADCaP cases have elevated frequencies of the C allele at rs7063314, and this allele is associated with reduced expression of

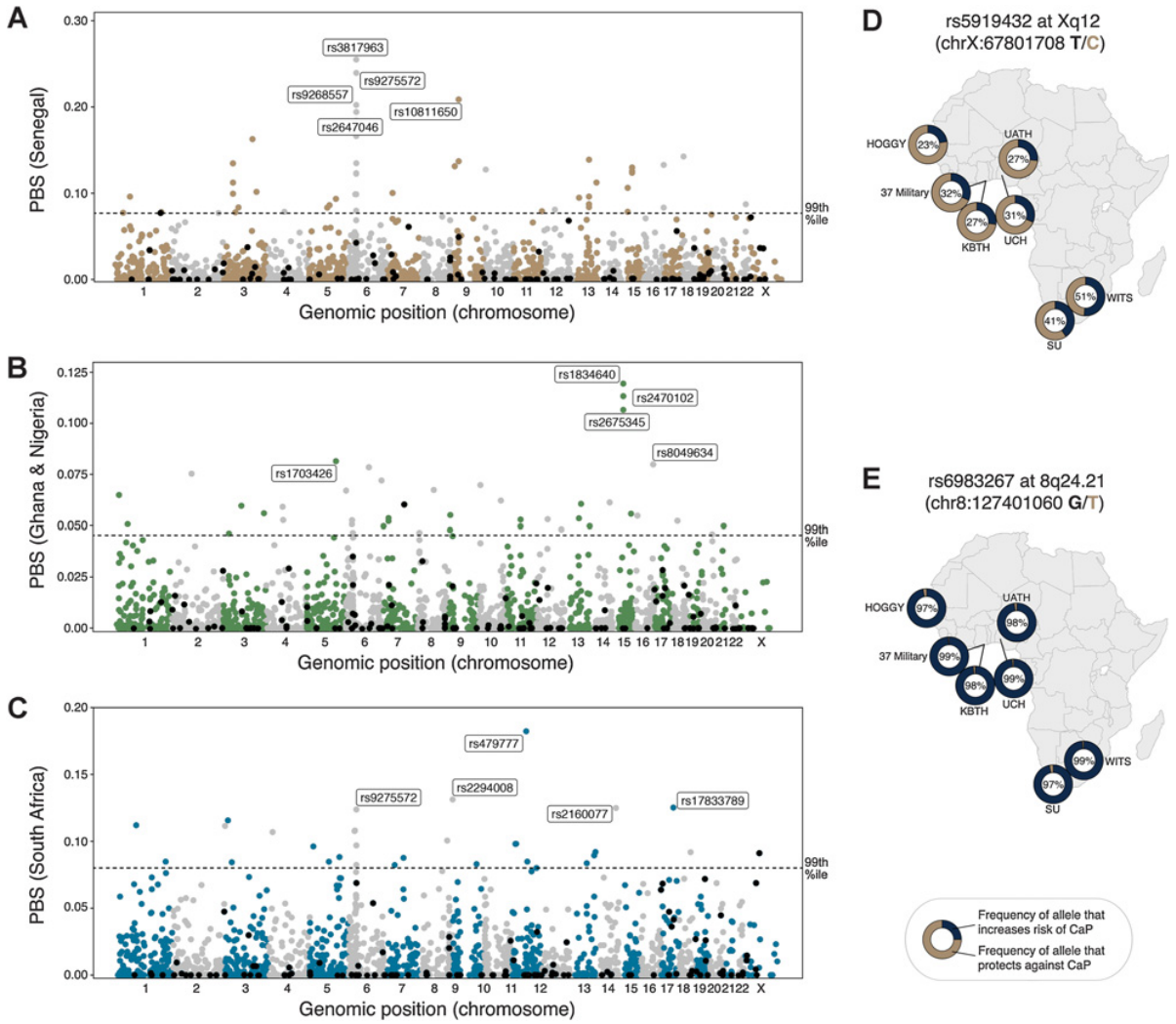

Figure 5.

PBS scores identify previously published cancer-associated loci that have large allele frequencies in Africa. Prostate cancer (CaP) associations from the Schumacher and colleagues GWAS (10) are represented by black points, and other cancer-associated loci are represented by gray and colored points. *Y* axes in each Manhattan plot quantify evolutionary branch lengths, not statistical strengths of association. Note that 99th percentiles of PBS scores for all markers on the MADCaP array are represented by dashed lines. **A,** PBS scores for the Senegal branch. **B,** PBS scores for the Ghana & Nigeria branch. **C,** PBS scores for the South African branch. **D,** Allele frequencies at the prostate cancer–associated SNP rs5919432 vary greatly across Africa. **E,** Allele frequencies at the prostate cancer–associated SNP rs6983267 are similar across Africa.

SPANXC and *SPANXA2-OT1* genes in testis (32). Other outlier loci to be followed-up in future studies include rs2188886, rs5980062, rs5977191, rs73119086, and rs5755469.

Predicted risks of prostate cancer in African populations

Using the MADCaP Array, we tested whether polygenic risks of prostate cancer vary by population. For each individual, PRS were calculated by counting risk alleles at 139 prostate cancer–associated loci and weighting by effect size. Higher PRS values indicate that an individual has a higher predicted risk of prostate cancer. **Fig. 6** compares PRS distributions for seven African study sites as well as European men from the 1000 Genomes Project, and median PRS values for each population are indicated by filled rectangles. Overall, we find that predicted risks of prostate cancer are much greater for African genomes than European genomes (Benjamini–Hochberg adjusted P values $<1.7 \times 10^{-6}$, Wilcoxon rank-sum test). This continental-level pattern is consistent with public health data (2). Differences in predicted prostate cancer risks between European and African populations exceed differences in predicted risk within Africa. There is a substantial amount of overlap in the PRS distributions of different African populations. Despite this similarity, we observe within-continent heterogeneity for the predicted risk of prostate cancer. The rank order of MADCaP study sites from lowest to highest predicted risk of prostate cancer is: HOGGY, KBTH, WITS, SU, UCH, 37 MILITARY, UATH. Individuals from Dakar, Senegal (gold in **Fig. 6**), have lower predicted risks of prostate cancer than other

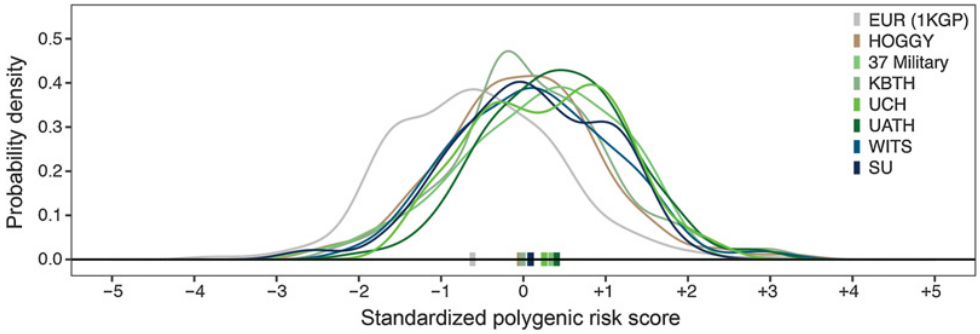

Figure 6.

PRS for prostate cancer differ for European and African genomes. Distributions of the genetic risk of prostate cancer are shown for Europeans from the 1000 Genomes Project and Africans from MADCaP study sites. Median PRS values for each study site are represented by colored rectangles. Markers used in PRS calculations are listed in Supplementary Table S3.

African study sites. Conversely, individuals from Abuja, Nigeria (dark green in **Fig. 6**), have higher predicted risks of prostate cancer than other African study sites. Some of these differences are statistically significant: Benjamini–Hochberg adjusted *P* values are 3.25×10^{-2} for HOGGY versus UCH, 1.14×10^{-3} for HOGGY versus UATH, 3.25×10^{-2} for KBTH versus UATH, 4.32×10^{-2} for WITS versus UATH, and 3.70×10^{-2} for SU versus UATH, (pairwise Wilcoxon rank-sum tests, Supplementary Table S2). Rare genetic variants with large effect sizes (e.g., rs183373024 and rs1447295) contribute to the wide tails of each PRS distribution in **Fig. 6**. Taken together, these results suggest that allele frequency differences at common disease-associated loci can contribute to population-level differences in prostate cancer risk.

Discussion

Using the Axiom genotyping solution, the MADCaP Network has developed a two-peg array that is optimized for studying the genetic basis of prostate cancer in men of African descent. This array successfully tags common and rare variation in African genomes (**Fig. 1**). Genomes of northeast African populations contain admixture with non-African populations (49), and this may contribute to less effective capture of Ethiopian genetic variation. The MADCaP Array combines the strengths of the Infinium OncoArray and the H3Africa Array, while maintaining excellent genotyping metrics for diverse African samples. The MADCaP Array will enable novel disease associations to be discovered and existing cancer associations to be fine-mapped. The 1.5 million markers described in Supplementary Table S1 are also likely to be of use to researchers developing their own custom genotyping arrays. Applying the MADCaP Array to over 800 African samples, we infer details of population structure, identify loci that contribute to population-level differences in cancer susceptibility, and generate personalized predictions of prostate cancer risk. These findings demonstrate that the MADCaP Array is an effective technology for inferring the population genetics of cancer risks in sub-Saharan Africa.

Sub-Saharan Africa contains substantial amounts of genetic diversity (33, 38, 49), and this contributes to population-level heterogeneity in cancer risks. For the study sites analyzed here, we found that genomes tend to fall into three distinct clusters (**Fig. 3B**). These clusters broadly match geography: samples from Senegal display similar genetic profiles, samples from Ghana and Nigeria cluster together, and samples from different locations in South Africa cluster together. We also found evidence that the genomes of African individuals contain mixtures of divergent genetic ancestries (**Fig. 3D**) and that South African study sites have larger effective population sizes than West African study sites (**Fig. 4**). Clearly, a one-size-fits-all approach is suboptimal when it comes to the genetics of African populations. The genetic heterogeneity of African populations calls for genotyping arrays that accurately capture African polymorphisms.

Genetic risks of cancer have changed during human history (50), and our analysis identified many cancer-associated loci with large allele frequency differences between African populations (**Fig. 5**; Supplementary Tables S3 and S4). There are multiple evolutionary reasons why allele frequencies at cancer-associated loci can differ across human populations. These causes include neutral processes like genetic drift and population bottlenecks. Natural selection can also contribute to large allele frequency differences between populations, either directly or indirectly via genetic hitchhiking (20). Regardless of the specific cause, differences in allele frequencies at cancer-associated loci can lead to population-level differences in disease risks, as observed in **Fig. 6**. However, we note that differences in PRS distributions between populations can either be due to real differences in risk or due to SNP ascertainment bias (26). As SNP-based heritability is a function of allele frequency, loci that are important to disease risks in one population need not contribute much to SNP-based heritability in other populations. Africa is not monomorphic when it comes to the genetic risk of prostate cancer, and there is a clear need to conduct studies that cover a broad range of populations.

Genotyping tools such as the MADCaP Array will enable novel cancer associations to be discovered in historically understudied African populations. Smaller LD blocks in African populations will also aid in fine-mapping of disease associations. Only by genotyping diverse study cohorts can researchers assess how well polygenic predictions of cancer risks are able to be generalized from large European study cohorts to the rest of the world.

Disclosure of Potential Conflicts of Interest

A. Mittal and M.H. Woehrmann are Sr. Staff Scientist at Thermo Fisher Scientific. A.I. Neugut is Consultant at Otsuka, United Biosource Corp, Hospira, and Eisai; is scientific advisory board member for EHE Intl; reports receiving Commercial Research Grant from Otsuka; and has an expert testimony from various organization. Christopher Warren is Senior Bioinformatic Scientist at Thermo Fisher Scientific. No potential conflicts of interest were disclosed by the other authors.

Authors' Contributions

Conception and design: M. Jalloh, S.M. Gueye, T.R. Rebbeck, J. Lachance
Development of methodology: T.R. Rebbeck, J. Lachance
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): P.W. Fernandez, A.W. Hsing, N.Y.F. Snyper, B. Adusei, J.E. Mensah, A.O.D. Abrahams, A.O. Adebiyi, A.T. Orunmuyi, O.I. Aisuodionoe-Shadrach, M.M. Nwegbu, M. Joffe, W.C. Chen, H. Irusen, A.I. Neugut, C.M. Ongaco, C.V. Andrews, T.R. Rebbeck
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): M. Harlemon, O. Ajayi, P. Kachambwa, M.S. Kim, C.N. Simonti, M.H. Quiver, A. Mittal, I. Agalliu, S.M. Gueye, H. Irusen, A.I. Neugut, C. Warren, M.H. Woehrmann, P. Zhang, E. Pugh, T.R. Rebbeck
Writing, review, and/or revision of the manuscript: M. Harlemon, O. Ajayi, P. Kachambwa, D.C. Petersen, A.W. Hsing, S. Baichoo, I. Agalliu, M. Jalloh, J.E. Mensah, A.O. Adebiyi, A.T. Orunmuyi, M. Joffe, H. Irusen, A.I. Neugut, C.M. Ongaco, E. Pugh, T.R. Rebbeck, L.N. Petersen, J. Lachance
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): J.E. Mensah, A.T. Orunmuyi, H. Irusen, Y. Quintana, M. Seutloali, M.B. Fadipe, M. Mawhinney, J. McBride, C.V. Andrews, M. Adams, T.R. Rebbeck
Study supervision: J.E. Mensah, L.N. Petersen, J. Lachance
Others (references): A.T. Orunmuyi

Acknowledgments

This work is a product of the MADCaP network (https://www.madcapnetwork.org/). This work was supported by a large multisite NIH/NCI grant (U01CA184374). Additional funding for this work includes startup funds from the School of Biological Sciences at Georgia Institute of Technology to J. Lachance and a seed grant from the Integrated Cancer Research Center at Georgia Institute of Technology.

Received July 15, 2019; revised October 3, 2019; accepted May 6, 2020; published first xx xx, xxxx.

References

1. Rebbeck TR. Prostate cancer genetics: variation by race, ethnicity, and geography. Semin Radiat Oncol 2017;27:3–10.
2. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68:394–424.
3. Powell IJ. Epidemiology and pathophysiology of prostate cancer in African-American men. J Urol 2007;177:444–9.
4. Lloyd T, Hounsome L, Mehay A, Mee S, Verne J, Cooper A. Lifetime risk of being diagnosed with, or dying from, prostate cancer by major ethnic group in England 2008–2010. BMC Med 2015;13:171.
5. Center MM, Jemal A, Lortet-Tieulent J, Ward E, Ferlay J, Brawley O, et al. International variation in prostate cancer incidence and mortality rates. Eur Urol 2012;61:1079–92.
6. Wu X, Gu J. Heritability of prostate cancer: a tale of rare variants and common single nucleotide polymorphisms. Ann Transl Med 2016;4:206.
7. Frank C, Fallah M, Sundquist J, Hemminki A, Hemminki K. Population landscape of familial cancer. Sci Rep 2015;5:12891.
8. 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.
9. Eeles RA, Kote-Jarai Z, Giles GG, Olama AA, Guy M, Jugurnauth SK, et al. Multiple newly identified loci associated with prostate cancer susceptibility. Nat Genet 2008;40:316–21.
10. 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.
11. 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.
12. Freedman ML, Haiman CA, Patterson N, McDonald GJ, Tandon A, Waliszewska A, et al. Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. Proc Natl Acad Sci U S A 2006;103:14068–73.
13. Fernandez P, Salie M, du Toit D, van der Merwe A. Analysis of prostate cancer susceptibility variants in South African men: replicating associations on chromosomes 8q24 and 10q11. Prostate Cancer 2015;2015:465184.
14. Murphy AB, Ukoli F, Freeman V, Bennett F, Aiken W, Tulloch T, et al. 8q24 risk alleles in west African and Caribbean men. Prostate 2012;72:1366–73.
15. Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature 2016;538:161–4.
16. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, et al. Human demographic history impacts genetic risk prediction across diverse populations. Am J Hum Genet 2017;100:635–49.
17. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet 2019;51:584.
18. Wang S, Qian F, Zheng Y, Ogundiran T, Ojengbede O, Zheng W, et al. Genetic variants demonstrating flip-flop phenomenon and breast cancer risk prediction among women of African ancestry. Breast Cancer Res Treat 2018;168:703–12.
19. Du Z, Lubmawa A, Gundell S, Wan P, Nalukenge C, Muwanga P, et al. Genetic risk of prostate cancer in Ugandan men. Prostate 2018;78:370–6.
20. 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.
21. Petersen DC, Jaratlerdsiri W, van Wyk A, Chan EKF, Fernandez P, Lyons RJ, et al. African KhoeSan ancestry linked to high-risk prostate cancer. BMC Med Genomics 2019;12:82.
22. Cook MB, Wang Z, Yeboah ED, Tettey Y, Biritwum RB, Adjei AA, et al. A genome-wide association study of prostate cancer in West African men. Hum Genet 2014;133:509–21.
23. Sirugo G, Williams SM, Tishkoff SA. The missing diversity in human genetic studies. Cell 2019;177:1080.
24. Hindorff LA, Bonham VL, Brody LC, Ginoza MEC, Hutter CM, Manolio TA, et al. Prioritizing diversity in human genomics research. Nat Rev Genet 2018;19:175–85.
25. Lachance J, Tishkoff SA. SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. Bioessays 2013;35:780–6.
26. Kim MS, Patel KP, Teng AK, Berens AJ, Lachance J. Genetic disease risks can be misestimated across global populations. Genome Biol 2018;19:179.
27. 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.
28. Mulder N, Abimiku A, Adebamowo SN, de Vries J, Matimba A, Olowoyo P, et al. H3Africa: current perspectives. Pharmgenomics Pers Med 2018;11:59–66.
29. Andrews C, Fortier B, Hayward A, Lederman R, Petersen L, McBride J, et al. Development, evaluation, and implementation of a Pan-African Cancer Research Network: Men of African Descent and Carcinoma of the Prostate. J Glob Oncol 2018;4:1–14.
30. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI catalog of published genome-wide association studies (GWAS catalog). Nucleic Acids Res 2017;45:D896–901.
31. 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature 2015;526:68–74.
32. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 2015;348:648–60.
33. Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, et al. The african genome variation project shapes medical genetics in Africa. Nature 2015;517:327–32.
34. Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. G3 (Bethesda) 2011;1:457–70.
35. Reed E, Nunez S, Kulp D, Qian J, Reilly MP, Foulkes AS. A guide to genome-wide association analysis and post-analytic interrogation. Stat Med 2015;34:3769–92.
36. ThermoFisher Scientific. Axiom® genotyping solution data analysis guide. 2017.
37. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res 2009;19:1655–64.
38. Schlebusch CM, Malmstrom H, Gunther T, Sjodin P, Coutinho A, Edlund H, et al. Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. Science 2017;358:652–5.

39. Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZX, Pool JE, et al. Sequencing of 50 human exomes reveals adaptation to high altitude. Science 2010;329:75–8.
40. Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. Evolution 1984;38:1358–70.
41. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics 2015;31:3555–7.
42. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stati Soc B 1995;57:289–300.
43. Li S, Schlebusch C, Jakobsson M. Genetic variation reveals large-scale population expansion and migration during the expansion of Bantu-speaking peoples. Proc Biol Sci 2014;281.pii: 20141448.
44. Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM, Goddard ME, et al. Recent human effective population size estimated from linkage disequilibrium. Genome Res 2007;17:520–6.
45. Loh PR, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, et al. Inferring admixture histories of human populations using linkage disequilibrium. Genetics 2013;193:1233–54.
46. Shiraishi K, Kunitoh H, Daigo Y, Takahashi A, Goto K, Sakamoto H, et al. A genome-wide association study identifies two new susceptibility loci for lung adenocarcinoma in the Japanese population. Nat Genet 2012;44:900–3.
47. Wu X, Ye Y, Kiemeney LA, Sulem P, Rafnar T, Matullo G, et al. Genetic variation in the prostate stem cell antigen gene PSCA confers susceptibility to urinary bladder cancer. Nat Genet 2009;41:991–5.
48. Kote-Jarai Z, Olama AA, Giles GG, Severi G, Schleutker J, Weischer M, et al. Seven prostate cancer susceptibility loci identified by a multi-stage genome-wide association study. Nat Genet 2011;43:785–91.
49. Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, et al. The genetic structure and history of Africans and African Americans. Science 2009;324:1035–44.
50. Berens AJ, Cooper TL, Lachance J. The genomic health of ancient hominins. Hum Biol 2017;89:7–19.