

1 **A custom genotyping array reveals population-level heterogeneity for**  
2 **the genetic risks of prostate cancer and other cancers in Africa**

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

15  
16  
17 **Running title:** MADCaP Array and prostate cancer in Africa

18  
19 **Keywords:** Africa, genomic medicine, genotyping array, population genetics, prostate  
20 cancer

21  
22 **Financial support:** This work was supported by a large multi-site NIH/NCI grant  
23 (U01CA184374). Additional funding for this work includes startup funds from the School  
24 of Biological Sciences at Georgia Institute of Technology to J. Lachance and a seed  
25 grant from the Integrated Cancer Research Center at Georgia Institute of Technology.

26  
27 **Conflict of interest:** A. Mittal, C. Warren, M.H. Woehrmann are employed by  
28 ThermoFisher Scientific, the manufacturer of the MADCaP Array. No conflicts of  
29 interest were reported by other authors.

30 **\*Corresponding author:** joseph.lachance@biology.gatech.edu

31 Joseph Lachance

34 Office: 404-894-0794

32 950 Atlantic Dr.

35 Cell: 631-332-6112

33 Atlanta, GA 30332

36 Fax: 404-894-0519

37

### 38 **Author affiliations**

39 <sup>1</sup> School of Biological Sciences, Georgia Institute of Technology, Atlanta Georgia, USA

40 <sup>2</sup> Clark Atlanta University, Atlanta, Georgia, USA

41 <sup>3</sup> Centre for Proteomics and Genomics Research, Cape Town, South Africa

42 <sup>4</sup> ThermoFisher Scientific, Santa Clara, California, USA

43 <sup>5</sup> Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South  
44 Africa

45 <sup>6</sup> Stanford Cancer Institute, Stanford University, Stanford, California, USA

46 <sup>7</sup> University of Mauritius, Réduit, Mauritius

47 <sup>8</sup> Albert Einstein College of Medicine, Bronx, New York, USA

48 <sup>9</sup> Hôpital Général de Grand Yoff, Institut de Formation et de Recherche en Urologie et  
49 Santé Familiale, Dakar, Senegal

50 <sup>10</sup> 37 Military Hospital, Accra, Ghana

51 <sup>11</sup> Korle-Bu Teaching Hospital and University of Ghana, Accra, Ghana

52 <sup>12</sup> College of Medicine, University of Ibadan, Ibadan, Nigeria

53 <sup>13</sup> College of Health Sciences, University of Abuja and University of Abuja Teaching  
54 Hospital, Abuja, Nigeria

55 <sup>14</sup> Non-Communicable Diseases Research Division, Wits Health Consortium (PTY) Ltd,  
56 Johannesburg, South Africa

57 <sup>15</sup> MRC Developmental Pathways to Health Research Unit, Department of Paediatrics,  
58 Faculty of Health Sciences, University of Witwatersrand, Johannesburg, South Africa

59 <sup>16</sup> Division of Human Genetics, School of Pathology, Faculty of Health Sciences,  
60 University of the Witwatersrand, Johannesburg, South Africa

61 <sup>17</sup> National Cancer Registry, National Health Laboratory Service, Johannesburg, South  
62 Africa

- 63 <sup>18</sup> Herbert Irving Comprehensive Cancer Center, Columbia University, New York, USA
- 64 <sup>19</sup> Beth Israel Deaconess Medical Center and Harvard Medical School, Boston,  
65 Massachusetts, USA
- 66 <sup>20</sup> Center for Inherited Disease Research, Johns Hopkins University, Baltimore,  
67 Maryland, USA
- 68 <sup>21</sup> Dana-Farber Cancer Institute, Boston, Massachusetts, USA
- 69 <sup>22</sup> Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA

70 **Abstract**

71 Although prostate cancer is the leading cause of cancer mortality for African men, the  
72 vast majority of known disease associations have been detected in European study  
73 cohorts. Furthermore, most genome-wide association studies have used genotyping  
74 arrays that are hindered by SNP ascertainment bias. To overcome these disparities in  
75 genomic medicine, the Men of African Descent and Carcinoma of the Prostate  
76 (MADCaP) Network has developed a genotyping array that is optimized for African  
77 populations. The MADCaP Array contains more than 1.5 million markers and an  
78 imputation backbone that successfully tags over 94% of common genetic variants in  
79 African populations. This array also has a high density of markers in genomic regions  
80 associated with cancer susceptibility, including 8q24. We assessed the effectiveness of  
81 the MADCaP Array by genotyping 399 prostate cancer cases and 403 controls from  
82 seven urban study sites in sub-Saharan Africa. We find that samples from Ghana and  
83 Nigeria cluster together, while samples from Senegal and South Africa yield distinct  
84 ancestry clusters. Using the MADCaP array, we identified cancer-associated loci that  
85 have large allele frequency differences across African populations. Polygenic risk  
86 scores were also generated for each genome in the MADCaP pilot dataset, and we  
87 found that predicted risks of CaP are lower in Senegal and higher in Nigeria.

88

89 **Significance:** We have developed an Africa-specific genotyping array which enables  
90 investigators to identify novel disease associations and to fine-map genetic loci that are  
91 associated with prostate and other cancers.

## 92 Introduction

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

101 Cancer is considered a genetic disease, and CaP has a high heritability: on the  
102 order of 58% (6). Risks of CaP tend to run in families; the relative risk of men with  
103 affected fathers is 2.1-fold higher compared to men without a family history and the  
104 relative risk of men with affected brothers is 2.3-fold higher compared to men without  
105 family history (7). In recent years, multiple genome-wide association studies (GWAS)  
106 have detected genetic associations with CaP (8-11). Collectively, these studies have  
107 yielded over 200 independent CaP risk-associated loci, and one key genomic region  
108 that has been repeatedly implicated in CaP and other cancers is 8q24 (12-14). The  
109 most comprehensive CaP GWAS analyzed to date included more than 140,000 cases  
110 and controls of European ancestry (10). In this study, Schumacher et al. generated a  
111 polygenic risk score (PRS) that successfully classified individuals into high and low risk  
112 categories.

113 Unfortunately, most GWAS have not focused on populations from sub-Saharan  
114 Africa. As of 2016, 81% of all GWAS samples were of European ancestry and 14% of  
115 all GWAS samples were of East Asian ancestry (15). Additionally, existing genotyping  
116 arrays do not adequately capture genetic variation in diverse African populations. Both  
117 of the aforementioned issues limit what is known about cancer genetics in African  
118 populations. Thus, the current set of known disease-associated loci is enriched for  
119 alleles with an intermediate frequency in Europe or Asia, but not Africa. This lack of  
120 representation can exacerbate existing health disparities by not capturing relevant  
121 genetic risk associations in African populations (16,17). Disease associations do not

122 always replicate across populations and ancestries, and the directions of risk  
123 associations at cancer-associated loci may differ across study populations (18,19).  
124 Previous work also indicates that risks of CaP vary by genetic ancestry across the globe  
125 (20,21). For example, an earlier GWAS in Ghana showed that the most promising SNP  
126 in this African population has not been identified in other populations (22). Hence, there  
127 is a clear need for more studies that analyze the genetics of African populations (23,24).  
128 Commonly used genotyping arrays tend to use markers that were originally ascertained  
129 in European populations (25). This can cause polygenic risks of complex diseases,  
130 including CaP, to be wrongly estimated (26). For example, the OncoArray Consortium  
131 has developed an array with over 500,000 markers, half of which are in genomic  
132 regions that tag cancer susceptibility (28). However, the OncoArray is not enriched for  
133 African polymorphisms. By contrast, the H3Africa Consortium has developed an array  
134 that includes over two million markers (27), but the H3Africa Array was not specifically  
135 designed for cancer studies. Existing arrays may therefore be sub-optimal for detecting  
136 cancer associations in African populations.

137 To address this problem, the Men of African Descent and Carcinoma of the Prostate  
138 (MADCaP) Network (29) developed a customized genotyping array optimized for fine-  
139 mapping and detecting novel associations with CaP in African populations. The  
140 MADCaP array will ultimately be used in an African GWAS containing over 6000 cases  
141 and controls. Here, we analyze a pilot dataset of over 800 individuals from sub-  
142 Saharan Africa. In this paper, we compare multiple genotyping platforms and test the  
143 efficacy of the MADCaP Array by genotyping samples from seven African study sites.  
144 Using data derived from the MADCaP Array, we also infer population structure, identify  
145 cancer-associated loci that have large allele frequency differences across Africa, and  
146 quantify polygenic risks of CaP in urban African populations.

147

148

## 149 **Materials and Methods**

150

## 151 **Inclusion criteria for markers on the MADCaP Array**

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

174

## 175 **Assessment of Imputation performance**

176 Imputation performance of the MADCaP Array was computed using the African  
177 Genome Resource reference panel, comprising of data from 4,956 individuals of African  
178 descent (33). African polymorphisms with MAF  $> 0.05$  were classified as common  
179 SNPs and African polymorphisms with a MAF between 0.01 and 0.05 were classified as  
180 rare SNPs. Imputation was performed with IMPUTE2 (v2.3.2) software using 10-fold

181 cross validation (34). Coverage in each population was calculated as the proportion of  
182 polymorphisms in the African Genome Resource reference panel in high LD ( $r^2 \geq 0.8$ )  
183 with markers on the MADCaP Array.

184

### 185 **Biospecimen and DNA quantification**

186 Biospecimens were obtained with informed consent using protocols approved from  
187 each study site's Institutional Review Board/Ethics Review Board. Blood samples were  
188 collected in EDTA vacutainer tubes and stored at either -20°C or -80°C dependent on  
189 the timeframe for DNA extraction. DNA was isolated using QIAamp DNA Blood kits. A  
190 total of 1.8 to 3.0  $\mu\text{g}$  high purity DNA at a concentration of 30 to 50  $\text{ng}/\mu\text{l}$  per sample  
191 was submitted for genotyping. DNA was transferred from study sites to genotyping  
192 laboratories using BioMatrica DNASTable 2D barcoded plates. Samples were then re-  
193 arrayed into plates using a BioMicroLab XL20 at a minimum concentration of 10  $\text{ng}/\mu\text{l}$  in  
194 50 $\mu\text{l}$ . All samples were run on the Infinium QC array and the MADCaP Array. Plate  
195 maps used a randomized block design to control for study site and case vs. control  
196 status.

197

### 198 **SNP calling, QC, and data curation**

199 Standard quality control (QC) procedures for Axiom genotyping data analysis were  
200 performed (35,36). Sample pre-processing was performed according to guidelines  
201 provided in the Thermo Fisher Scientific Axiom Genotyping Solution Data Analysis  
202 Guide (36). The custom MADCaP Array is based on a two-peg design. Peg 1 contains  
203 a total of 852,610 probe sets, covering 801,275 markers. Peg 2 contains a total of  
204 790,524 probe sets, covering 790,170 markers. 1,902 probe sets overlap both pegs.  
205 Raw data CEL files, representing more than 802 samples, as well as 28 technical  
206 replicates and additional controls, were imported into the Axiom Analysis Suite (AxAS)  
207 version 4.0.3.3 for filtering of sample call rate and clustering of SNP genotype calls.  
208 Samples with DishQC  $\geq 0.82$  and a QC call rate  $>97\%$  were included for downstream  
209 genotyping analysis. Additional SNP metrics included tests of Hardy-Weinberg  
210 proportions, reproducibility of genotyping calls, and identifying Mendelian



211 inconsistencies. The Centre for Proteomics and Genomics Research (CPGR) in Cape  
212 Town, South Africa and the Center for Inherited Disease Research (CIDR) at Johns  
213 Hopkins University independently assessed QC metrics for each probe set.

214 Data from both pegs of the MADCaP Array were merged and PLINK was used to  
215 remove related samples (identity-by-state > 0.5). Multi-allelic SNPs were excluded from  
216 downstream analyses. After filtering markers with low call rates and excluding poorly  
217 called and related samples, 1,513,172 markers and 802 samples were used in  
218 subsequent population genetic analyses. This MADCaP pilot dataset contains 399 CaP  
219 cases and 403 controls. Details of MADCaP case and control recruitment have been  
220 previously reported (29).

221

## 222 **Array comparisons**

223 We compared markers on arrays developed by the MADCaP Network (29), the  
224 OncoArray Consortium (28), and the H3Africa Consortium (27). Genomic positions  
225 from the MADCaP Array, Infinium Oncoarray and H3Africa Array were intersected to  
226 determine overlapping markers between arrays. The *LiftOver* bioinformatics tool was  
227 used to ensure that all genomic positions used build GRCh38/hg38 of the human  
228 reference genome. Mean derived allele frequencies (DAF) for each array were  
229 calculated as described previously (26). This involved obtaining allele frequencies from  
230 each of the five continental regions using the 1000 Genomes Project (31): Africa (AFR),  
231 Americas (AMR), East Asia (EAS), Europe (EUR), and South Asia (SAS). Calculations  
232 of mean derived allele frequencies used 450,000 markers that were randomly selected,  
233 without replacement, from each array. The joint allele frequency distribution of all  
234 1,513,172 markers on the MADCaP Array was found by comparing African and pooled  
235 non-African data from the 1000 Genomes Project.

236

## 237 **MDS and ADMIXTURE**

238 Using PLINK, an LD pruned subset of 25,000 autosomal SNPs for each of 802  
239 samples was obtained (MAF > 0.05,  $r^2 < 0.8$ ). The same subset of SNPs was used for  
240 Multidimensional Scaling (MDS) and ADMIXTURE analyses. Two dimensional MDS

241 plots were generated using PLINK and R. ADMIXTURE software (37) was run for K = 2  
242 through K = 5. Cross-validation was performed to determine the optimal K value.

243

### 244 **Runs of homozygosity and LD decay**

245 As per Schlebusch et al. (38), runs of homozygosity were identified using PLINK for  
246 homozygous lengths between 500kb and 1000kb. This analysis was repeated for all  
247 802 samples in the MADCaP dataset. Individual runs of homozygosity were summed to  
248 yield the cumulative runs of homozygosity (cROH) for each sample. For each MADCaP  
249 study site, PLINK v1.90b6.9 was used to calculate LD between all variants with a MAF >  
250 0.10. These calculations were made for all pairs of markers within 100kb and 100  
251 marker windows. Distances between genetic variants were used to place pairs of  
252 variants into 1kb bins. For each study site, the mean  $r^2$  between pairs of genetic  
253 variants was calculated for each 1kb bin.

254

### 255 **Identification of divergent loci via PBS calculations**

256 Population branch statistics (PBS) were calculated as per Yi et al. (39). Data from  
257 multiple MADCaP study sites were pooled to yield allele frequencies for three  
258 populations: Senegal (HOGGY), Ghana & Nigeria (37 Military, KBTH, UATH, and UCH)  
259 and South Africa (WITS and SU). Genetic distances between pairs of populations were  
260 calculated using Weir and Cockerham's  $F_{st}$  (40). The following equations were then  
261 used to calculate PBS scores for three different evolutionary branches:

262

$$263 \quad PBS_{Senegal} = \frac{-\ln(1-F_{ST,sn-za})-\ln(1-F_{ST,sn-gh\&ng})+\ln(1-F_{ST,za-gn\&ng})}{2} \quad (1)$$

264

$$265 \quad PBS_{Ghana \& Nigeria} = \frac{-\ln(1-F_{ST,sn-gh\&ng})-\ln(1-F_{ST,za-gh\&ng})+\ln(1-F_{ST,sn-za})}{2} \quad (2)$$

266

$$267 \quad PBS_{South Africa} = \frac{-\ln(1-F_{ST,sn-za})-\ln(1-F_{ST,za-gh\&ng})+\ln(1-F_{ST,sn-gh\&ng})}{2} \quad (3)$$

268

269 Subscripts in the above equations refer to country codes: *sn* for Senegal, *gh* for  
270 Ghana, *ng* for Nigeria, and *za* for South Africa. Undefined and negative values of Weir  
271 and Cockerham's  $F_{st}$  were treated as zero in the above equations, and undefined or  
272 negative PBS scores were also treated as zero. PBS scores were calculated for 2,477  
273 unique markers from the NHGRI-EBI GWAS Catalog (30) that yield 5,337 cancer and  
274 cancer-related associations.

275

### 276 **Calculation of polygenic risk scores (PRS)**

277 Polygenic risk scores (PRS) were built using a curated set of 141 CaP-associated  
278 loci. Schumacher et al. previously developed a 147 loci PRS for CaP (10), and 119 of  
279 these 147 markers are on the MADCaP Array. Proxies were found for 22 of the  
280 remaining 27 markers, by identifying markers on the MADCaP array in LD with loci from  
281 the Schumacher PRS ( $r^2 > 0.4$ ). LDlink (41) was then used to select alleles that tag  
282 increased CaP risk at proxy markers. Table S2 lists markers that were used to  
283 generate the PRS described here.

284 As per Schumacher et al. (10), effect size information was incorporated into PRS  
285 calculations. For each locus, we counted whether an individual has 0, 1, or 2 copies of  
286 the risk-increasing allele (i.e. the allele dose  $g_{i,j}$  for locus  $i$  in individual  $j$ ). Here we  
287 used adjusted effect sizes:  $\beta_i = \ln(OR_i) \times r_i^2$ , where effect sizes from Schumacher et  
288 al. (10) are scaled by how well proxy markers tag each disease-associated locus.  
289 Doses of risk-increasing alleles were weighted by adjusted effect sizes and summed  
290 across all 141 loci to a raw PRS for each individual.

291

$$293 \quad PRS_j = \sum_{i=1}^{141} g_{i,j} \beta_i \quad (4)$$

292  
294

295 PRS were calculated for 802 MADCaP samples and 240 men of European ancestry  
296 from 1000 Genomes Project (31). Standardized PRS values were then generated for all

297 1,042 individuals by scaling raw PRS values to have a mean of zero and standard  
298 deviation of one.

299

300

## 301 **Results**

302

### 303 **Imputation using the MADCaP Array**

304 Using whole genome sequences, we quantified the extent to which the MADCaP  
305 array tags African genetic variation (Fig. 1). Depending on the population, 94% to 99%  
306 of common African SNPs were successfully tagged by the MADCaP Array ( $r^2 \geq 0.8$ ,  
307  $MAF > 0.05$ ). The MADCaP Array also tagged 63% to 97% of rare African SNPs ( $r^2 \geq$   
308  $0.8$ ,  $MAF$  between  $0.01$  and  $0.05$ ). It captures a larger fraction of Ugandan genetic  
309 variation than Ethiopian and KhoeSan variation. Fig. 1 also shows that the MADCaP  
310 Array successfully tags variation in admixed African-Caribbean and African-American  
311 genomes. Regardless of population, the MADCaP Array successfully captures a large  
312 fraction of African genetic variation.

313

### 314 **Comparisons with other arrays**

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

321 We compared the mean derived allele frequencies (DAF) of markers found on the  
322 MADCaP, OncoArray, and H3Africa arrays, using Continental allele frequencies from  
323 the 1000 Genomes Project (Fig. 2B). The null expectation here is that the mean DAF  
324 should be the same for each population since all humans are evolutionarily equidistant  
325 to other primates. Mean DAFs of markers on the MADCaP Array were similar for each

326 continental population, suggesting that the MADCaP Array is relatively unbiased with  
327 respect to SNP selection. By contrast, the mean DAFs of markers on the OncoArray  
328 and H3Africa Array were lower for African populations than non-African populations  
329 (Fig. 2B). These DAF differences are indicative of SNP ascertainment bias (26).  
330 Examining the joint site frequency spectrum of non-African and African populations,  
331 similar counts of MADCaP markers are found above and below the diagonal in Fig. 2C.  
332 One exception to this pattern is that the MADCaP Array is enriched for markers that are  
333 polymorphic in Africa but monomorphic outside of Africa, but not vice-versa.

334 Densities of markers that are found in different genomic regions vary by genotyping  
335 array. Here, we focus on 8q24, a cancer-associated genomic region that contains  
336 *PCAT2*, *CCAT2*, and the proto-oncogene *c-myc*. Numbers of markers per 100kb are  
337 shown for three different arrays in Fig. 2D. The MADCaP Array contains a moderately  
338 high density of markers across the genome, with peaks near known cancer-associated  
339 loci. Neighboring markers on the MADCaP Array have a median distance of 856bp and  
340 a mean distance of 2082bp. The Infinium OncoArray has high marker densities near  
341 cancer-associated loci, but a low density of markers for other parts of the genome. In  
342 contrast, the H3Africa Array has a moderately even density of markers across the entire  
343 genome.

344

### 345 **Efficacy of the MADCaP Array**

346 We tested the efficacy of the MADCaP Array by genotyping over 800 African  
347 individuals from seven MADCaP study sites (Fig. 3A): the Hôpital Général de Grand  
348 Yoff/Institut de Formation et de Recherche en Urologie in Dakar, Senegal (HOGGY), 37  
349 Military Hospital in Accra, Ghana (37 Military), Korle-Bu Teaching Hospital in Accra,  
350 Ghana (KBTH), University College Hospital in Ibadan, Nigeria (UCH), University of  
351 Abuja Teaching Hospital in Abuja, Nigeria (UATH), WITS Health Consortium/National  
352 Health Laboratory Services in Johannesburg, South Africa (WITS), and Stellenbosch  
353 University in Cape Town, South Africa (SU). Sample accrual was restricted to  
354 individuals with sub-Saharan African ancestry; admixed individuals with European

355 ancestry from Cape Town were excluded. Of the MADCaP samples analyzed herein,  
356 399 are CaP cases and 403 are controls (Fig. 3B).

357 Up to 94.9% of the markers on peg 1 and 95.9% of the markers on peg 2 passed  
358 QC filtering. We note that probe sets from the MADCaP Array were 2.4 times less likely  
359 to fail than probe sets from the OncoArray Array (28). For both peg 1 and peg 2, mean  
360 call rates, reproducibility, and concordance all exceeded 99.5%, and only a small subset  
361 of markers had Mendelian inconsistencies (Table 1). Overall, we find that the MADCaP  
362 Array is an effective genotyping platform.

363

### 364 **Population structure and genetic admixture**

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

378 ADMIXTURE plots reveal shared ancestry among MADCaP samples (Fig. 3E). In  
379 these plots, individuals are linear mixtures of multiple genetic ancestries – indicated by  
380 different colors. Cross-validation error is minimized at  $K=3$ , i.e. the best fit to the data  
381 occurs for three ancestry colors (Fig. S1). At  $K=2$ , we are able to distinguish between  
382 West African and South African populations. Setting  $K=3$  reveals three major ancestry  
383 clusters: gold in Senegal, green for Ghana and Nigeria, and blue in South Africa. At  $K=$   
384 4 ancestry patterns match each country. Intriguingly, individuals from Ibadan, Nigeria

385 (UCH) share ancestry with samples from Ghana, i.e. they contain moderate amounts of  
386 light green ancestry at  $K = 4$ . Similarly, individuals from Johannesburg, South Africa  
387 contain traces of genetic ancestry that are primarily found in Nigeria (dark green),  
388 perhaps due to the Bantu expansion during the last 5,000 years (42).  $K = 5$  reveals  
389 evidence of population structure within South Africa, with greater proportions of light  
390 blue ancestry found in Johannesburg (WITS) compared to Cape Town (SU). Both study  
391 sites from Accra, Ghana (37 Military and KBTH) have similar genetic ancestry profiles.  
392 Finally, we note that cases and controls for each study site are ancestry-matched. On a  
393 genome-wide scale, individuals in the MADCaP study with CaP have similar ancestry  
394 proportions compared to healthy MADCaP controls.

395

### 396 **Runs of homozygosity and linkage disequilibrium**

397 Genotyping arrays can be used to identify runs of homozygosity – stretches of DNA  
398 where maternally and paternally inherited haplotypes are identical. Using the MADCaP  
399 array, we quantified cumulative runs of homozygosity (cROH) in each genome (Fig. 4A).  
400 Although there is heterogeneity within each study site, cROH are smaller for South  
401 African genomes than Senegalese, Ghanaian, or Nigerian genomes analyzed in this  
402 study ( $p$ -value =  $3.31 \times 10^{-9}$ , Wilcoxon rank-sum tests). This lower homozygosity can  
403 either be due to large historical population sizes or due to admixture.

404 To distinguish whether this lower homozygosity is due to large historical population  
405 sizes or admixture, we calculated LD decay curves for each of the seven MADCaP  
406 study sites (Fig. 4B). Populations with small effective population sizes have more LD  
407 than populations with large effective population sizes (43). Admixture also increases  
408 the amount of LD (44). In general, we observed less LD for South African sites than  
409 other study sites (WITS and SU in Fig. 4B). These differences in LD decay curves do  
410 not appear to be due to admixture, since the South African populations studied here  
411 have similar levels of admixture to other African populations (Fig. 3E). Overall, the data  
412 in Fig. 4 support the idea that historic population sizes were larger in South Africa than  
413 West Africa. One implication of the smaller haplotype blocks that are found in genomes  
414 from Johannesburg and Cape Town is that GWAS using these samples will require



415 arrays with high densities of markers, a characteristic that is shared by the MADCaP  
416 Array.

417

### 418 **Divergent allele frequencies at cancer-associated loci**

419 Risk allele frequencies at cancer-associated loci can vary across the African  
420 continent. To identify genetic variants that have large allele frequency differences  
421 across MADCaP populations, population branch statistic (PBS) scores were calculated  
422 for all observed cancer-associated loci in the NHGRI-EBI GWAS Catalog that have  
423 markers on the MADCaP Array, as well as loci associated with other cancer-related  
424 traits (e.g., skin pigmentation and smoking). These scores were calculated for three  
425 different evolutionary branches: Senegal (Fig. 5A), Ghana & Nigeria (Fig. 5B), and  
426 South Africa (Fig. 5C). Here, CaP hits used in PRS calculations are represented by  
427 black points, while gray and colored points indicate other cancer-associated loci. Table  
428 S2 contains PBS scores and allele frequencies for 141 CaP markers used in MADCaP  
429 PRS calculations. Table S3 contains PBS scores and allele frequencies for 2,477  
430 markers that are associated with cancer and cancer-related traits.

431 All three branches contain multiple loci with PBS scores that are located in the  
432 MHC/HLA region on chromosome 6. For example, rs3817963 has the top PBS score  
433 for the Senegalese branch (Fig. 5A). This SNP at 6p21.32 has been associated with  
434 lung adenocarcinoma (45). The risk-increasing allele at rs3817963 has an allele  
435 frequency of 33.9% in Senegal, 12.9% in Ghana, 10.4% in Nigeria, and 8.4% in South  
436 Africa ( $p$ -values  $< 0.0001$  for pairwise comparisons between Senegal and other  
437 countries, two sample Z-test). Another cancer-associated variant that has large allele  
438 frequency differences between African populations is rs2294008, located at 8q24.3.  
439 This SNP has the second highest PBS score for the South African branch and it has  
440 previously been associated with bladder and gastric cancer (46,47). The risk-increasing  
441 allele at rs2294008 has an allele frequency of 28.7% in Senegal, 35.7% in Ghana,  
442 28.8% in Nigeria, and 54.8% in South Africa ( $p$ -values  $< 0.0001$  for pairwise  
443 comparisons between South Africa and other countries, two sample Z-test).



444 Focusing on CaP-associated loci, we identify loci with large allele frequency  
445 differences between populations, as well as loci that have similar allele frequencies for  
446 each African population. For example, rs5919432 is a CaP-associated SNP that is  
447 located 71kb from the *androgen receptor* gene at Xq12 (48). This SNP has the highest  
448 X-linked PBS score in Fig. 5C. Compared to other study sites, we found that South  
449 Africans have elevated frequencies of the risk allele at rs5919432 (Fig. 5D). These  
450 allele frequency differences contribute to population-level differences in CaP-risk. We  
451 found that the risk-increasing T allele at rs5919432 is more common in MADCaP cases  
452 than controls (34.2% vs. 32.1%). Although the association between rs5919432 and  
453 CaP was originally discovered in European men, this data reveals that rs5919432 has a  
454 similar effect in African populations. The genomic region 8q24.21 contains multiple loci  
455 that have been associated with CaP in European men, including rs6983267 (10).  
456 Although rs6983267 at 8q24.21 is associated with CaP, it does not have large allele  
457 frequency differences between African populations (Fig. 5E). We also found that the  
458 risk-increasing G allele at rs6983267 is more common in MADCaP cases than controls  
459 (98.2% vs. 97.9%). Note that the protective allele at rs6983267 is rare in Africa but  
460 moderately common in Europe (the T allele is found at 50.0% in EUR, 1000 Genomes  
461 Project data). This pattern suggests that while rs6983267 contributes to continental-  
462 level differences in CaP risk, it has only a minimal effect on population-level differences  
463 in CaP risk within sub-Saharan Africa.

464

### 465 **Predicted risks of prostate cancer (CaP) in urban African populations**

466 Using the MADCaP Array we tested whether polygenic risks of CaP vary by  
467 population. For each individual, risk scores were calculated by counting risk alleles at  
468 141 CaP-associated loci and weighting by effect size. Higher PRS values indicate that  
469 an individual has a higher predicted risk of CaP. Fig. 6 compares PRS distributions for  
470 seven African study sites as well as European men from the 1000 Genomes Project,  
471 and mean PRS values for each population are indicated by filled rectangles. Overall,  
472 we find that predicted risks of CaP are much greater for urban African genomes than  
473 European genomes ( $p$ -value  $< 2.2 \times 10^{-16}$ , Wilcoxon rank-sum test). This continental-

474 level pattern is consistent with public health data (2). Note that differences in predicted  
475 CaP risks between European and African populations exceed differences in predicted  
476 risk within Africa. Focusing on MADCaP study sites, there is a substantial amount of  
477 overlap in the polygenic risk score distributions of different African populations. Despite  
478 this similarity, we observe within-continent heterogeneity for the predicted risk of CaP.  
479 The rank order of MADCaP study sites from lowest to highest predicted risk of CaP is:  
480 HOGGY, KBTH, SU, WITS, 37 MILITARY, UCH, UATH. Individuals from Dakar,  
481 Senegal (gold in Fig. 6) have lower predicted risks of CaP than other African study sites.  
482 Conversely, individuals from Abuja, Nigeria (dark green in Fig. 6) have higher predicted  
483 risks of CaP than other African study sites. Some of these differences are statistically  
484 significant: p-values are  $\leq 0.046$  for HOGGY vs 37 Military, UATH, UCH, and WITS, and  
485 p-values are  $\leq 0.031$  for UATH vs. HOGGY, KBTH, WITS, and SU (pairwise Wilcoxon  
486 rank-sum tests). Rare genetic variants with large effect sizes (e.g. rs183373024 and  
487 rs1447295) contribute to the wide tails of each PRS distribution in Fig. 6. Taken  
488 together, these results suggest that allele frequency differences at common disease-  
489 associated loci can contribute to population-level differences in CaP risk.

490

491

## 492 **Discussion**

493 Using the Axiom genotyping solution, the MADCaP Network has developed a two-  
494 peg array that is optimized for studying the genetic basis of CaP in men of African  
495 descent. This array successfully tags common and rare variation in African genomes  
496 (Fig. 1). The MADCaP Array combines the strengths of the Infinium OncoArray and the  
497 H3Africa Array, while maintaining excellent genotyping metrics for diverse African  
498 samples. Markers on the MADCaP Array will enable novel disease associations to be  
499 discovered and existing cancer associations to be fine-mapped. The 1.5 million  
500 markers described in Table S1 are also likely to be of use to researchers developing  
501 their own custom genotyping arrays. Applying the MADCaP Array to over 800 African  
502 samples, we are able to infer details of population structure, identify loci that contribute

503 to population-level differences in cancer susceptibility, and generate personalized  
504 predictions of CaP risk. These findings demonstrate that the MADCaP Array is an  
505 effective technology for inferring the population genetics of cancer risks in sub-Saharan  
506 Africa.

507 Sub-Saharan Africa contains substantial amounts of genetic diversity (33,38,49),  
508 and this contributes to population-level heterogeneity in cancer risks. For the urban  
509 study sites analyzed here, we found that genomes tend to fall into three distinct clusters  
510 (Fig. 3C). These clusters broadly match geography: samples from Senegal display  
511 similar genetic profiles, samples from Ghana and Nigeria cluster together, and samples  
512 from different locations in South Africa cluster together. We also found evidence that  
513 the genomes of African individuals contain mixtures of divergent genetic ancestries (Fig.  
514 3E) and that South African study sites have larger effective population sizes than West  
515 African study sites (Fig. 4). Clearly, a one-size-fits-all approach is suboptimal when it  
516 comes to the genetics of African populations. The genetic heterogeneity of African  
517 populations calls for genotyping arrays that accurately capture African polymorphisms.

518 Genetic risks of cancer have changed during recent human history (50), and our  
519 analysis of urban African genomes found many cancer-associated loci that have  
520 divergent allele frequencies (Fig. 5, Table S2, and Table S3). There are multiple  
521 evolutionary reasons why allele frequencies at cancer-associated loci can differ across  
522 human populations. These evolutionary causes include neutral processes like genetic  
523 drift and population bottlenecks. Natural selection can also contribute to large allele  
524 frequency differences between populations, either directly or indirectly via genetic  
525 hitchhiking (20). Regardless of the specific evolutionary cause, differences in allele  
526 frequencies at cancer-associated loci can lead to population-level differences in disease  
527 risks, as observed in Fig. 6. As SNP-based heritability is a function of allele frequency,  
528 loci that are important to disease risks in one population need not contribute much to  
529 SNP-based heritability in other populations. Africa is not monomorphic when it comes  
530 to the genetic risk of CaP and there is a clear need to conduct studies that cover a  
531 broad range of populations.

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

538

## 539 **Disclosures of Potential Conflicts of Interest**

540 A. Mittal, C. Warren, M.H. Woehrmann are employed by ThermoFisher Scientific,  
541 the manufacturer of the MADCaP Array. No conflicts of interest were reported by other  
542 authors.

543

## 544 **Author's Contributions**

545 **Conception and design:** T.R. Rebbeck, J. Lachance

546 **Development of methodology:** J. Lachance

547 **Acquisition of data (provided animals, acquired and managed patients, provided**

548 **facilities, etc.):** P. Fernandez, M. Jalloh, S.M. Gueye, N.Y. Snyder, B. Adusei, J.E.

549 Mensah, A.O.D. Abrahams, A.O. Adebijoyi, A. Orunmuyi, O.I. Aisuodionoe-Shadrach,

550 M.M. Nwegbu, M. Joffe, W.C. Chen, H. Irusen

551 **Analysis and interpretation of data (e.g., statistical analysis, biostatistics,**

552 **computational analysis):** M. Harlemon, O. Ajayi, P. Kachambwa, M.S. Kim, C.N.

553 Simonti, M.H. Quiver, A. Mittal, C. Warren, M.H. Woehrmann, P. Zhang, C. Ongaco, E

554 Pugh

555 **Writing, review, and/or revision of the manuscript:** M. Harlemon, O. Ajayi, P.

556 Kachambwa, D.C. Petersen, A.W. Hsing, I. Agalliu, S. Baichoo, A.O. Adebijoyi, A.

557 Orunmuyi, O.I., T.R. Rebbeck, L. Petersen, J. Lachance

558 **Administrative, technical, or material support (i.e., reporting or organizing data,**

559 **constructing databases):** A.I. Neugut, A.W. Hsing, Y. Quintana, M. Mawhinney, C.

560 Andrews, J. McBride, M. Adams  
561 **Study supervision:** L. Petersen, J. Lachance  
562 **Other (performed genotyping):** M. Seutloali, M Fadipe, J. McBride

563

## 564 **Acknowledgments**

565 This work is a product of the MADCaP network (Men of African Descent, Carcinoma  
566 of the Prostate, <https://www.madcapnetwork.org/>). This work was supported by a large  
567 multi-site NIH/NCI grant (U01CA184374). Additional funding for this work includes  
568 startup funds from the School of Biological Sciences at Georgia Institute of Technology  
569 to J. Lachance and a seed grant from the Integrated Cancer Research Center at  
570 Georgia Institute of Technology.

571

572

## 573 **References**

- 574 1. Rebbeck TR. Prostate Cancer Genetics: Variation by Race, Ethnicity, and  
575 Geography. *Semin Radiat Oncol* **2017**;27:3-10
- 576 2. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer  
577 statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for  
578 36 cancers in 185 countries. *CA Cancer J Clin* **2018**;68:394-424
- 579 3. Powell IJ. Epidemiology and pathophysiology of prostate cancer in African-  
580 American men. *J Urol* **2007**;177:444-9
- 581 4. Lloyd T, Hounsome L, Mehay A, Mee S, Verne J, Cooper A. Lifetime risk of being  
582 diagnosed with, or dying from, prostate cancer by major ethnic group in England  
583 2008–2010. *BMC medicine* **2015**;13:171
- 584 5. Center MM, Jemal A, Lortet-Tieulent J, Ward E, Ferlay J, Brawley O, *et al.*  
585 International variation in prostate cancer incidence and mortality rates. *Eur Urol*  
586 **2012**;61:1079-92
- 587 6. Wu X, Gu J. Heritability of prostate cancer: a tale of rare variants and common  
588 single nucleotide polymorphisms. *Ann Transl Med* **2016**;4:206

- 589 7. Frank C, Fallah M, Sundquist J, Hemminki A, Hemminki K. Population  
590 Landscape of Familial Cancer. *Sci Rep* **2015**;5:12891
- 591 8. Al Olama AA, Kote-Jarai Z, Berndt SI, Conti DV, Schumacher F, Han Y, *et al.* A  
592 meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for  
593 prostate cancer. *Nat Genet* **2014**;46:1103-9
- 594 9. Eeles RA, Kote-Jarai Z, Giles GG, Olama AA, Guy M, Jugurnauth SK, *et al.*  
595 Multiple newly identified loci associated with prostate cancer susceptibility. *Nat*  
596 *Genet* **2008**;40:316-21
- 597 10. Schumacher FR, Al Olama AA, Berndt SI, Benlloch S, Ahmed M, Saunders EJ,  
598 *et al.* Association analyses of more than 140,000 men identify 63 new prostate  
599 cancer susceptibility loci. *Nat Genet* **2018**;50:928-36
- 600 11. Haiman CA, Chen GK, Blot WJ, Strom SS, Berndt SI, Kittles RA, *et al.* Genome-  
601 wide association study of prostate cancer in men of African ancestry identifies a  
602 susceptibility locus at 17q21. *Nat Genet* **2011**;43:570-3
- 603 12. Freedman ML, Haiman CA, Patterson N, McDonald GJ, Tandon A, Waliszewska  
604 A, *et al.* Admixture mapping identifies 8q24 as a prostate cancer risk locus in  
605 African-American men. *Proceedings of the National Academy of Sciences*  
606 **2006**;103:14068-73
- 607 13. Fernandez P, Salie M, du Toit D, van der Merwe A. Analysis of prostate cancer  
608 susceptibility variants in South African men: replicating associations on  
609 chromosomes 8q24 and 10q11. *Prostate cancer* **2015**;2015
- 610 14. Murphy AB, Ukoli F, Freeman V, Bennett F, Aiken W, Tulloch T, *et al.* 8q24 risk  
611 alleles in West African and Caribbean men. *Prostate* **2012**;72:1366-73
- 612 15. Popejoy AB, Fullerton SM. Genomics is failing on diversity. *Nature*  
613 **2016**;538:161-4
- 614 16. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, *et al.*  
615 Human Demographic History Impacts Genetic Risk Prediction across Diverse  
616 Populations. *Am J Hum Genet* **2017**;100:635-49

- 617 17. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of  
618 current polygenic risk scores may exacerbate health disparities. *Nature genetics*  
619 **2019**;51:584
- 620 18. Wang S, Qian F, Zheng Y, Ogundiran T, Ojengbede O, Zheng W, *et al.* Genetic  
621 variants demonstrating flip-flop phenomenon and breast cancer risk prediction  
622 among women of African ancestry. *Breast Cancer Res Treat* **2018**;168:703-12
- 623 19. Du Z, Lubmawa A, Gundell S, Wan P, Nalukenge C, Muwanga P, *et al.* Genetic  
624 risk of prostate cancer in Ugandan men. *Prostate* **2018**;78:370-6
- 625 20. Lachance J, Berens AJ, Hansen MEB, Teng AK, Tishkoff SA, Rebbeck TR.  
626 Genetic Hitchhiking and Population Bottlenecks Contribute to Prostate Cancer  
627 Disparities in Men of African Descent. *Cancer Res* **2018**;78:2432-43
- 628 21. Petersen DC, Jaratlerdsiri W, van Wyk A, Chan EKF, Fernandez P, Lyons RJ, *et*  
629 *al.* African KhoeSan ancestry linked to high-risk prostate cancer. *BMC Med*  
630 *Genomics* **2019**;12:82
- 631 22. Cook MB, Wang Z, Yeboah ED, Tettey Y, Biritwum RB, Adjei AA, *et al.* A  
632 genome-wide association study of prostate cancer in West African men. *Hum*  
633 *Genet* **2014**;133:509-21
- 634 23. Sirugo G, Williams SM, Tishkoff SA. The Missing Diversity in Human Genetic  
635 Studies. *Cell* **2019**;177:1080
- 636 24. Hindorff LA, Bonham VL, Brody LC, Ginoza MEC, Hutter CM, Manolio TA, *et al.*  
637 Prioritizing diversity in human genomics research. *Nat Rev Genet* **2018**;19:175-  
638 85
- 639 25. Lachance J, Tishkoff SA. SNP ascertainment bias in population genetic  
640 analyses: why it is important, and how to correct it. *Bioessays* **2013**;35:780-6
- 641 26. Kim MS, Patel KP, Teng AK, Berens AJ, Lachance J. Genetic disease risks can  
642 be misestimated across global populations. *Genome Biol* **2018**;19:179
- 643 27. Mulder N, Abimiku A, Adebamowo SN, de Vries J, Matimba A, Olowoyo P, *et al.*  
644 H3Africa: current perspectives. *Pharmgenomics Pers Med* **2018**;11:59-66



- 645 28. Amos CI, Dennis J, Wang Z, Byun J, Schumacher FR, Gayther SA, *et al.* The  
646 OncoArray Consortium: A Network for Understanding the Genetic Architecture of  
647 Common Cancers. *Cancer Epidemiol Biomarkers Prev* **2017**;26:126-35
- 648 29. Andrews C, Fortier B, Hayward A, Lederman R, Petersen L, McBride J, *et al.*  
649 Development, Evaluation, and Implementation of a Pan-African Cancer Research  
650 Network: Men of African Descent and Carcinoma of the Prostate. *J Glob Oncol*  
651 **2018**;4:1-14
- 652 30. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, *et al.* The new  
653 NHGRI-EBI Catalog of published genome-wide association studies (GWAS  
654 Catalog). *Nucleic Acids Res* **2017**;45:D896-D901
- 655 31. 1000 Genomes Project Consortium. A global reference for human genetic  
656 variation. *Nature* **2015**;526:68-74
- 657 32. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis:  
658 multitissue gene regulation in humans. *Science* **2015**;348:648-60
- 659 33. Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I,  
660 Hatzikotoulas K, *et al.* The African Genome Variation Project shapes medical  
661 genetics in Africa. *Nature* **2015**;517:327-32
- 662 34. Howie B, Marchini J, Stephens M. Genotype imputation with thousands of  
663 genomes. *G3 (Bethesda)* **2011**;1:457-70
- 664 35. Reed E, Nunez S, Kulp D, Qian J, Reilly MP, Foulkes AS. A guide to genome-  
665 wide association analysis and post-analytic interrogation. *Stat Med*  
666 **2015**;34:3769-92
- 667 36. ThermoFisher Scientific. Axiom® genotyping solution data analysis guide. 2017.
- 668 37. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry  
669 in unrelated individuals. *Genome Res* **2009**;19:1655-64
- 670 38. Schlebusch CM, Malmstrom H, Gunther T, Sjodin P, Coutinho A, Edlund H, *et al.*  
671 Southern African ancient genomes estimate modern human divergence to  
672 350,000 to 260,000 years ago. *Science* **2017**;358:652-5
- 673 39. Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZX, Pool JE, *et al.* Sequencing of  
674 50 human exomes reveals adaptation to high altitude. *Science* **2010**;329:75-8



- 675 40. Weir BS, Cockerham CC. Estimating F-Statistics for the Analysis of Population  
676 Structure. *Evolution* **1984**;38:1358-70
- 677 41. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring  
678 population-specific haplotype structure and linking correlated alleles of possible  
679 functional variants. *Bioinformatics* **2015**;31:3555-7
- 680 42. Li S, Schlebusch C, Jakobsson M. Genetic variation reveals large-scale  
681 population expansion and migration during the expansion of Bantu-speaking  
682 peoples. *Proc Biol Sci* **2014**;281
- 683 43. Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM, Goddard ME, *et al.*  
684 Recent human effective population size estimated from linkage disequilibrium.  
685 *Genome Res* **2007**;17:520-6
- 686 44. Loh PR, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, *et al.* Inferring  
687 admixture histories of human populations using linkage disequilibrium. *Genetics*  
688 **2013**;193:1233-54
- 689 45. Shiraishi K, Kunitoh H, Daigo Y, Takahashi A, Goto K, Sakamoto H, *et al.* A  
690 genome-wide association study identifies two new susceptibility loci for lung  
691 adenocarcinoma in the Japanese population. *Nat Genet* **2012**;44:900-3
- 692 46. Wu X, Ye Y, Kiemeny LA, Sulem P, Rafnar T, Matullo G, *et al.* Genetic variation  
693 in the prostate stem cell antigen gene PSCA confers susceptibility to urinary  
694 bladder cancer. *Nat Genet* **2009**;41:991-5
- 695 47. Study Group of Millennium Genome Project for Cancer, Sakamoto H, Yoshimura  
696 K, Saeki N, Katai H, Shimoda T, *et al.* Genetic variation in PSCA is associated  
697 with susceptibility to diffuse-type gastric cancer. *Nat Genet* **2008**;40:730-40
- 698 48. Kote-Jarai Z, Olama AA, Giles GG, Severi G, Schleutker J, Weischer M, *et al.*  
699 Seven prostate cancer susceptibility loci identified by a multi-stage genome-wide  
700 association study. *Nat Genet* **2011**;43:785-91
- 701 49. Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, *et al.*  
702 The genetic structure and history of Africans and African Americans. *Science*  
703 **2009**;324:1035-44

- 704 50. Berens AJ, Cooper TL, Lachance J. The Genomic Health of Ancient Hominins.  
705 Hum Biol **2017**;89:7-19  
706

707 **Tables**

708

709 **Table 1.** Genotyping metrics for peg 1 and peg 2 of the MADCaP Array.

710

<b>QC metric</b>	<b>MADCaP Array (Peg 1)</b>	<b>MADCaP Array (Peg 2)</b>
Mean call rate (proportion markers called)	99.55%	99.63%
Mean reproducibility (Same calls for each probe)	99.85%	99.90%
Mean concordance (Same calls in replicates)	99.53%	99.56%
Error rate (Mendelian inconsistencies)	0.051%	0.032%

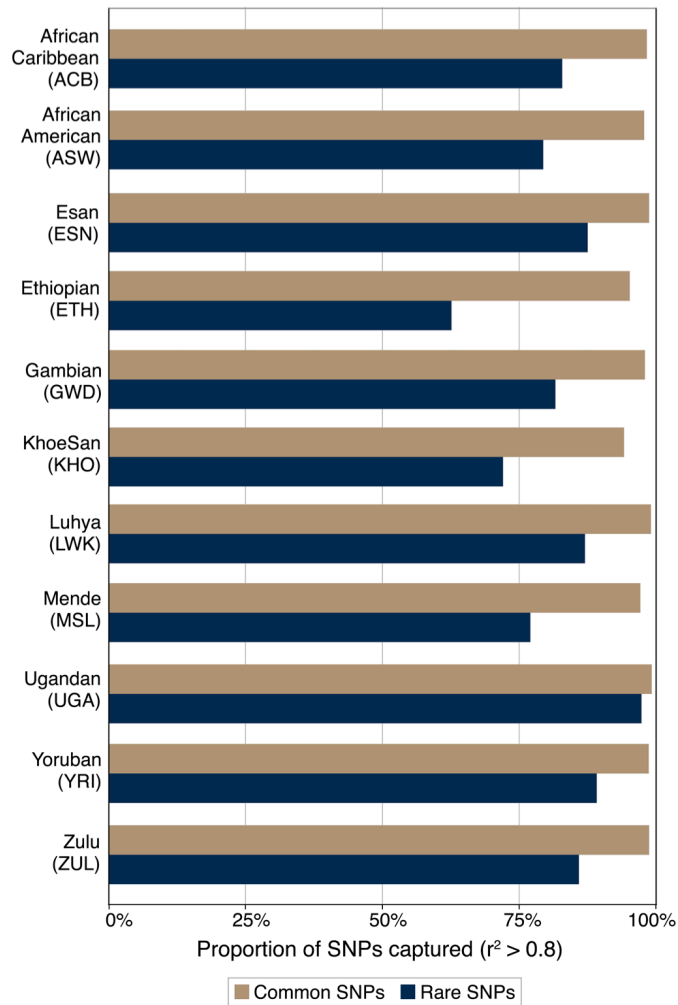
711

712 **Figures**

713

714 **Figure 1**

715



716

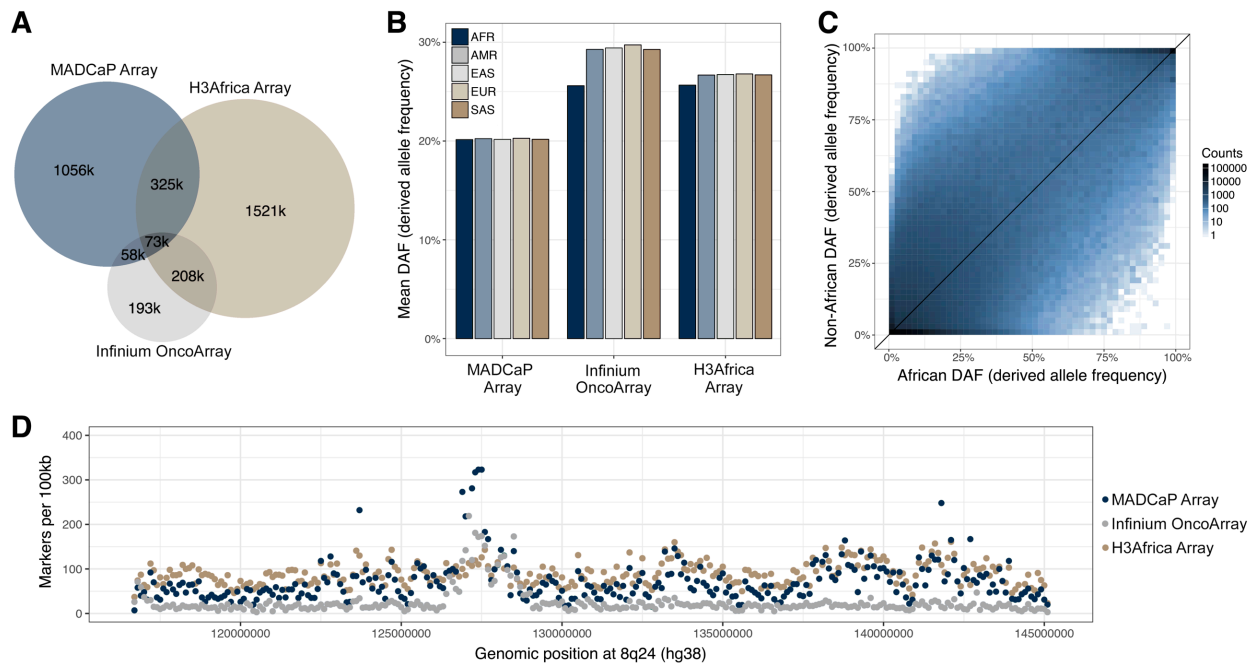
717

718 **Figure 1**

719 African SNPs are accurately imputed using MADCaP Array. Proportions of SNPs in 11  
720 populations of African ancestry that are successfully tagged by markers on the  
721 MADCaP Array are shown ( $r^2 \geq 0.8$ ). Here, common SNPs have a MAF > 0.05 and rare  
722 SNPs have a MAF between 0.01 and 0.05.

723 **Figure 2**

724



725

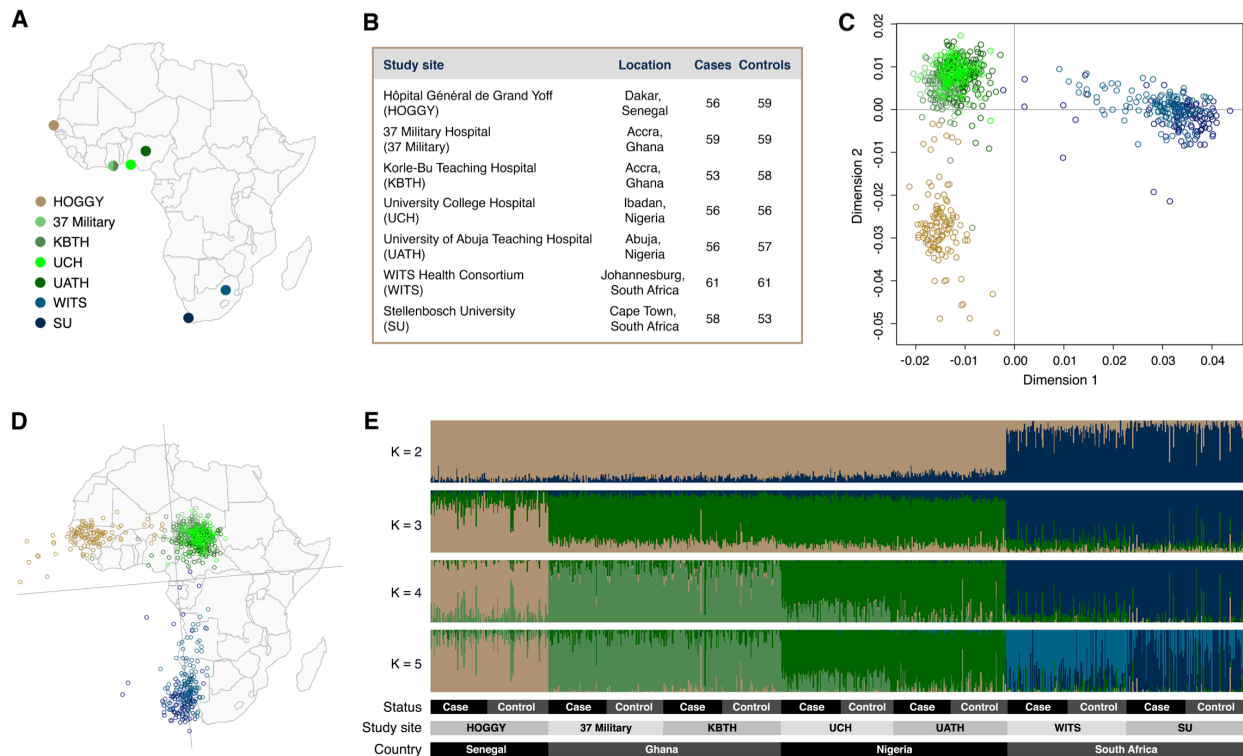
726

727 **Figure 2**

728 Comparisons between the MADCaP Array, Infinium OncoArray, and H3Africa Array. **A**,  
729 Venn Diagram showing overlap between markers on each array. Sizes of circles are  
730 proportional to the number of markers on each array. **B**, Mean DAF of markers on the  
731 MADCaP Array, Infinium OncoArray, and H3Africa Array. Continental allele frequencies  
732 from the 1000 Genomes Project are shown here. **C**, Joint site frequency spectrum of  
733 markers on the MADCaP Array. African and pooled non-African allele frequencies from  
734 1000 Genomes Project are shown here. Shading indicates the number of markers on  
735 the MADCaP array that are in each bin. **D**, Density of markers per non-overlapping  
736 100kb window. The 8q24 genomic region is shown here.

737 **Figure 3**

738



739

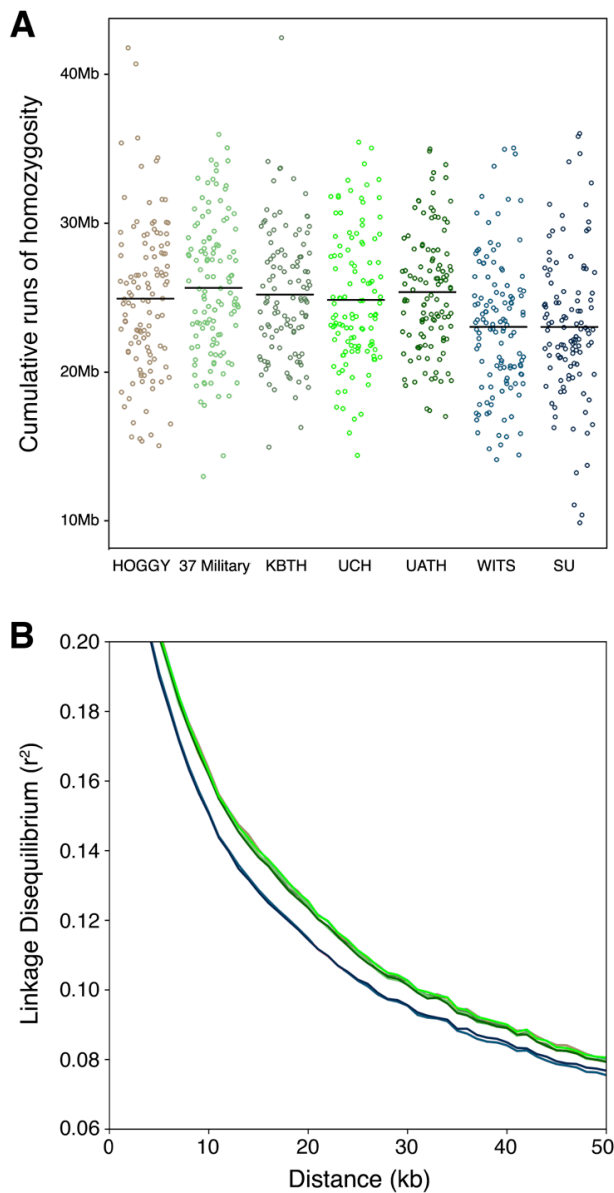
740

741 **Figure 3**

742 The MADCaP Array reveals population structure and shared genetic ancestries among  
 743 urban African study sites. **A**, Geographic locations of each MADCaP study site. **B**,  
 744 Numbers of cases and controls from each study site. **C**, Two-dimensional MDS plot of  
 745 802 MADCaP samples. Senegalese samples are represented by gold circles,  
 746 Ghanaian and Nigerian samples are represented by green circles, and South African  
 747 circles are represented by blue circles. **D**, Genes mirror geography when the two-  
 748 dimensional MDS plot is rotated clockwise. **E**, ADMIXTURE plot of 802 MADCaP  
 749 samples. The best to genetic data occurs at K = 3.

750

751 **Figure 4**



752

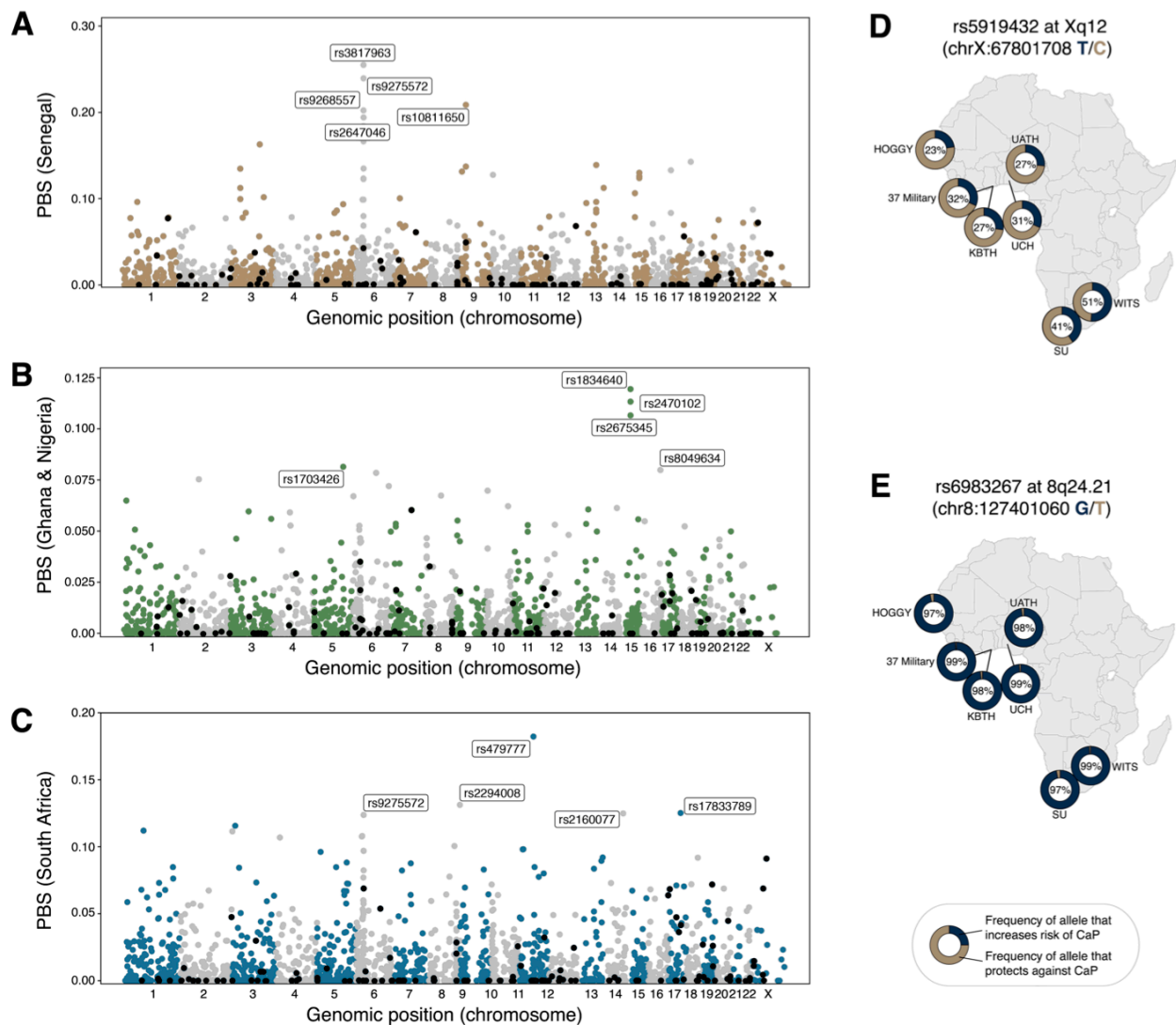
753

754 **Figure 4**

755 Runs of homozygosity and LD decay curves vary by African study site. **A**, Cumulative  
756 runs of homozygosity (cROH) 500kb to 1000kb in length for each MADCaP sample,  
757 labelled by study site. **B**, LD decay curves for each study site. Gold indicates  
758 Senegalese data, green indicates Ghanaian and Nigerian data, and blue indicates  
759 South African data. South African study sites have less LD than West African study  
760 sites (Senegalese data overlaps Ghanaian and Nigerian data).

761 **Figure 5**

762



763

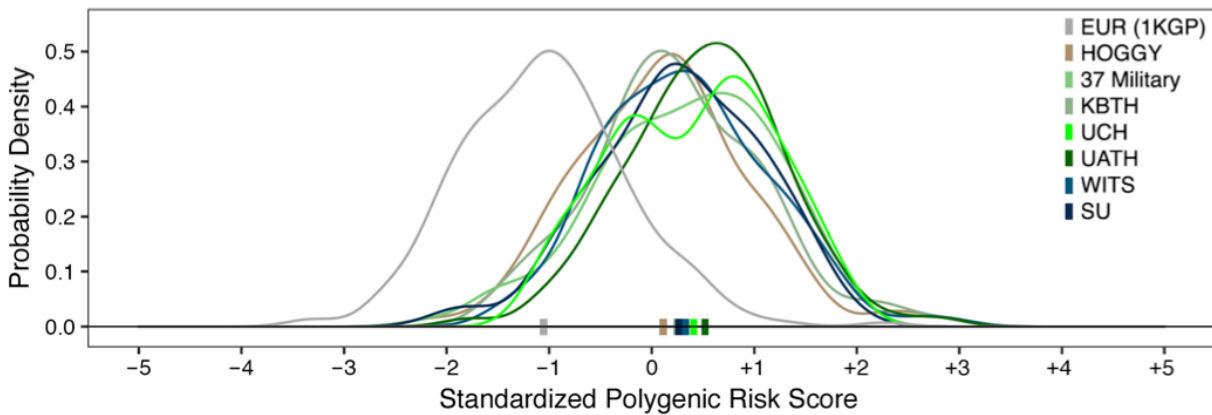
764 **Figure 5**

765 PBS scores identify divergent loci in Africa that are associated with cancer risks. CaP  
 766 associations from the Schumacher et al. 2018 GWAS (10) are represented by black  
 767 points, and other cancer-associated loci are represented by gray and colored points. **A**,  
 768 PBS scores for the Senegal branch. **B**, PBS scores for the Ghana & Nigeria branch. **C**,  
 769 PBS scores for the South African branch. **D**, Allele frequencies at the CaP-associated  
 770 SNP rs5919432 vary greatly across Africa. **E**, Allele frequencies at the CaP-associated  
 771 SNP rs6983267 are similar across Africa.



772 **Figure 6**

773



774

775

776 **Figure 6**

777 Polygenic risk scores for prostate cancer differ for European and African genomes.

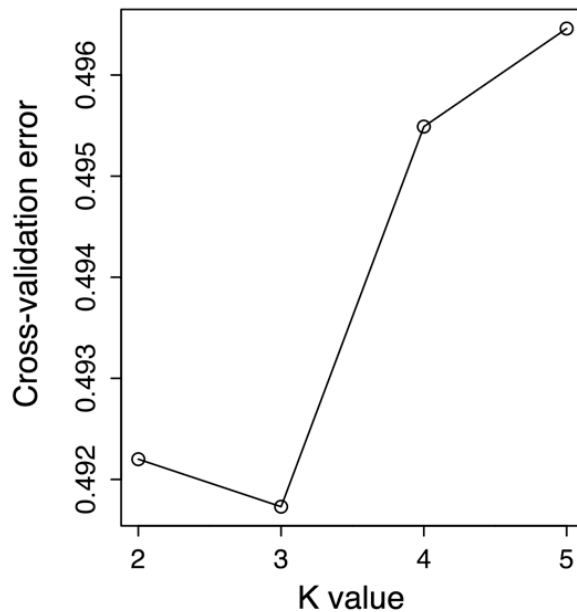
778 Distributions of the genetic risk of CaP are shown for Europeans from the 1000

779 Genomes Project and Africans from MADCaP study sites. Mean PRS values for each

780 study site are represented by colored rectangles. Markers used in PRS calculations are

781 listed in Table S2.

## 782 Supplemental Data



783

784 **Figure S1. Cross-validation error in ADMIXTURE analyses.** The best fit to data  
785 occurs at K = 3.

786

787 **Table S1. Successfully called markers on the MADCaP Array.** Tab-delimited file  
788 that includes genomic positions, inclusion criteria, and overlap with other arrays.

789

790 **Table S2. Markers used in PRS calculations.** This list includes all markers from the  
791 Schumacher et al. 2018 CaP PRS (10) as well as proxy markers that are found on the  
792 MADCaP Array. Chromosome and positions (build hg38) listed here are for MADCaP  
793 markers used in PRS calculations. Allele frequencies and PBS scores are also included  
794 for the 141 CaP markers used in MADCaP PRS calculations.

795

796 **Table S3. Comprehensive list of cancer-associated loci with African allele  
797 frequencies and PBS scores.** This dataset includes a total of 2,477 unique markers  
798 yielding 5,337 disease or trait associations.