Title
Whole-exome sequencing of over 4100 men of African ancestry and prostate cancer risk

Permalink
https://escholarship.org/uc/item/9vm9s6h6

Journal
Human Molecular Genetics, 25(2)

ISSN
0964-6906

Authors
Rand, Kristin A
Rohland, Nadin
Tandon, Arti
et al.

Publication Date
2016-01-15

DOI
10.1093/hmg/ddv462

Peer reviewed

OXFORD

Human Molecular Genetics, 2016, Vol. 25, No. 2 371–381

doi: 10.1093/hmg/ddv462
Advance Access Publication Date: 24 November 2015
Association Studies Article

ASSOCIATION STUDIES ARTICLE

Whole-exome sequencing of over 4100 men of African ancestry and prostate cancer risk

Kristin A. Rand[1,2,†], Nadin Rohland[3,4,†], Arti Tandon[3,4], Alex Stram[1], Xin Sheng[1], Ron Do[3,4], Bogdan Pasaniuc[5,6,7], Alex Allen[3,4], Dominique Quinque[3,4], Swapan Mallick[3,4,8], Loic Le Marchand[9], Sam Kaggwa[10], Alex Lubwama[11], The African Ancestry Prostate Cancer GWAS Consortium[‡], The ELLIPSE/GAME-ON Consortium[‡], Daniel O. Stram[1,2], Stephen Watya[11,12], Brian E. Henderson[1,2,§], David V. Conti[1,2], David Reich[3,4,8,†] and Christopher A. Haiman[1,2,†,*]

[1]Department of Preventive Medicine, Keck School of Medicine, [2]Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA 90033, USA, [3]Department of Genetics, Harvard Medical School, Harvard University, Boston, MA 02115, USA, [4]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA, [5]Bioinformatics Interdepartmental Program, [6]Department of Human Genetics, David Geffen School of Medicine, [7]Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA, [8]Howard Hughes Medical Institute, Harvard Medical School, Boston, MA 02115, USA, [9]Epidemiology Program, Cancer Research Center, University of Hawaii, Honolulu, HI 96813, USA, [10]Department of Surgery, [11]School of Public Health, Makerere University College of Health Sciences, Kampala, Uganda and [12]Uro Care, Kampala, Uganda

*To whom correspondence should be addressed at: Harlyne Norris Research Tower, 1450 Biggy Street, Room 1504, Los Angeles, CA 90033, USA. Tel: +1 3234427755; Fax: +1 3234427749; Email: haiman@usc.edu

Abstract

Prostate cancer is the most common non-skin cancer in males, with a ~1.5–2-fold higher incidence in African American men when compared with whites. Epidemiologic evidence supports a large heritable contribution to prostate cancer, with over 100 susceptibility loci identified to date that can explain ~33% of the familial risk. To explore the contribution of both rare and common variation in coding regions to prostate cancer risk, we sequenced the exomes of 2165 prostate cancer cases and 2034 controls of African ancestry at a mean coverage of 10.1×. We identified 395 220 coding variants down to 0.05% frequency [57% non-synonymous (NS), 42% synonymous and 1% gain or loss of stop codon or splice site variant] in 16 751 genes with the strongest associations observed in *SPARCL1* on 4q22.1 (rs13051, *Ala49Asp*, OR = 0.78, $P = 1.8 \times 10^{-6}$) and *PTPRR* on 12q15 (rs73341069, *Val239Ile*, OR = 1.62, $P = 2.5 \times 10^{-5}$). In gene-level testing, the two most significant genes were *C1orf100* ($P = 2.2 \times 10^{-4}$) and *GORAB* ($P = 2.3 \times 10^{-4}$). We did not observe exome-wide significant associations (after correcting for multiple hypothesis testing) in single variant or gene-level testing in the overall case–control or case–case analyses of disease aggressiveness. In this first whole-exome

[†] These authors contributed equally to this work. The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors and the last two senior authors who jointly directed this work should be regarded as joint Last Authors.

[‡] Membership of The African Ancestry Prostate Cancer GWAS Consortium and The ELLIPSE/GAME-ON Consortium is provided in the text.

[§] In Memoriam.

Received: July 11, 2015. **Revised:** September 14, 2015. **Accepted:** November 6, 2015

sequencing study of prostate cancer, our findings do not provide strong support for the hypothesis that NS coding variants down to 0.5–1.0% frequency have large effects on prostate cancer risk in men of African ancestry. Higher-coverage sequencing efforts in larger samples will be needed to study rarer variants with smaller effect sizes associated with prostate cancer risk.

Introduction

Prostate cancer is the most common non-skin cancer in males in the USA, with an estimated 220 800 new cases diagnosed in 2015 (www.cancer.gov). This disease disproportionately affects men of African ancestry, with the incidence being 1.5–2-fold greater in men of African ancestry compared with men in other racial/ ethnic populations (1). Epidemiologic evidence suggests a strong heritable contribution to prostate cancer (2,3), and previous genome-wide association studies (GWAS) have been successful in identifying over 100 common genetic variants associated with risk (4–20). These common risk alleles (frequencies > 5%) were primarily discovered in European and Asian populations, have modest effect sizes (relative risks < 1.3) and are estimated to explain ∼33% of familial risk (20). A recent study examining 82 risk variants in ∼4800 prostate cancer cases and ∼4700 controls of African ancestry found 83% of variants to have directionally consistent effect estimates, suggesting that the majority of GWAS-identified loci harbor risk alleles that are common and shared across populations (21).

An unexplored hypothesis is that 'missing heritability' of complex diseases such as prostate cancer may be attributed to rare variants. GWAS have been limited in their ability to adequately assess the contribution of risk from rare variants [minor allele frequency (MAF) <1%], as current genotyping array technology inadequately captures this spectrum of variation (22,23). Sequencing in families with a history of breast and ovarian cancers have revealed rare deletions in *BRCA2* that are associated with a ∼5-fold increase in risk of developing prostate cancer, with risk increasing to ∼7-fold for early-onset prostate cancer (age < 65 years) (24). More recently, a rare non-synonymous (NS) variant *Gly84Glu* (rs138213197) in *HOXB13* has been found to be associated with risk of both hereditary (ORs = 4.5–9.0) and sporadic prostate cancer (ORs = 2.5–4.5) (25–28). This variant is only found in men of European ancestry and is a founder mutation in the Nordic population where the population frequency is ∼1%, whereas the frequency of the risk allele is reported to be ≤0.2% in other European ancestry populations (26,27). While the evidence supporting rare coding variation in prostate cancer is limited, these examples suggest that rare coding variation may contribute to prostate cancer susceptibility, with the allelic effect being larger than loci revealed through GWAS.

To further explore the contribution of rare coding variation in prostate cancer, we performed whole-exome sequencing in 2165 prostate cancer cases and 2034 controls of African ancestry to identify and directly test rare variants in protein-coding sequence that may be important and/or unique to this high-risk population. In addition to association testing of single variants, we performed gene-level tests to investigate the aggregate effects of rare coding variants within genes and in specific candidate pathways that have been implicated in the pathogenesis of prostate cancer.

Results

We targeted 51 Mb to capture 20 965 genes and 334 278 exons and were able to confidently call variants down to 0.05% (observed at least 4 times in >8000 chromosomes, see Materials and Methods). The mean coverage of the targeted regions before quality control filtering was 7× (range of coverage across all samples: <1–30.9×; 80% of samples had a mean coverage of >3.5×), with 91% of reads mapped to target regions. After removing poor-performing samples and variants ($n = 423$, $n = 332\,042$ of 727 262 variants, respectively; see Materials and Methods) and excluding intronic regions, the overall mean depth was 10.1× in 1938 cases and 1838 controls (Supplementary Material, Table S1). Overall, 57% of the variants in gene-coding sequence were NS, with 42% synonymous, and 1% exonic splice sites or stop codon loss or gain; distributions that are comparable to those observed in an African American sample ($n = 2203$) in the Exome Sequencing Project (ESP: 58% NS, 38% synonymous, 4% splice sites). Of the 148 866 variants with a MAF of ≤0.1%, 12.6% were reported in the AFR population of the 1000 Genomes Project (1KGP) and 37.9% were reported by the ESP. Of the 163 783 variants with a MAF between 0.1% and 0.5%, 36.8% were observed in 1KGP, whereas 66.9% were reported in the ESP. The overlap substantially increased in the 19 995 variants with a MAF between 0.5% and 1%, where 87.2% were in 1KGP and 92.5% were reported by the ESP. As expected, there was very high overlap in the 61 790 common variants (MAF > 1%), with 96.5% overlap with 1KGP and 94.8% overlap with ESP (Table 1). Over 60% of the variants in our data with a MAF of <0.1% are not in ESP, indicating that a large fraction of coding variation has yet to be discovered or tested in association with prostate cancer risk in this population. However, a limitation of the low-to-moderate coverage sequencing approach is that we missed ∼20% of coding variants with frequencies between 0.5 and 50% that were found by the ESP (Supplementary Material, Table S2). These are variants that we had more than adequate samples and coverage to observe; however, they were located in regions that were removed during quality control filtering as a consequence of the low-coverage approach. We were able to test 94% of these variants through imputation as described later (and see Materials and Methods). We were also unable to study insertion or deletion (indels) variants, which require high-coverage sequence data to call accurately.

Single variant associations

Under the assumption that rare variants will have a large effect (ORs > 5), we performed a power calculation to determine a lower allele frequency threshold for single-variant tests and determined that with our current sample size, we have 65% power to detect an OR = 6 and >99% power to detect an OR = 10 down to 0.2% frequency with an α-level = 3.75×10^{-7}. Consequently, we removed 261 853 variants with an allele frequency of <0.2% from all analyses and any variant without at least one count in either cases or controls in each sub-analysis. We report ORs and 95% confidence intervals (CI) from a logistic regression model and *P*-values from a likelihood ratio test.

Overall prostate cancer risk

After filtering, association testing was performed for 133 367 variants available for analysis. We observed only one variant at $P < 10^{-5}$ (one expected) and nine variants at $P < 10^{-4}$ (13 expected) whereas the QQ plot showed no evidence for systematic error (lambda = 1.03, Supplementary Material, Fig. S1). The two most significant associations were with NS variants in *SPARCL1* on 4q22.1 (rs13051: control freq 0.25, *Ala49Asp*, OR = 0.78, $P = 1.8 \times$

Table 1. Annotation of exonic data from 3776 men of African ancestry (MAC ≥ 4)

Minor allele frequency	Total	Splicing	Non-synonymous	Synonymous	Stoploss, Stopgain	% in 1KGP	% in ESP
≤0.1%	148866	154	90341	56009	2362	12.6	37.9
>0.1, ≤0.5%	163783	250	94914	66246	2373	36.8	66.9
>0.5, ≤1%	19995	25	10742	9098	130	87.2	92.5
>1%	61790	71	29079	32356	284	96.5	94.8
Total	394434[a]	500	225076	163709	5149	40.1	62.2

[a]786 variants could not be annotated.

Table 2. Overall single variant association results (1938 cases/1838 controls)

Variant	Chromosome, base pair	Amino acid change	Gene	Risk/ref allele	Case–control[a]	Afr/Eur[b]	OR (95% CI)[c]	P-value[d]
rs13051	4:88416188	Ala49Asp	*SPARCL1*	G/T	0.19/0.25	0.12/0.64	0.78 (0.70–0.86)	1.8×10^{-6}
rs73341069	12:71147994	Val239Ile	*PTPRR*	T/C	0.05/0.03	0.08/–	1.62 (1.29–2.04)	2.5×10^{-5}
rs148679475	7:141954999	Leu104Leu	*PRSS58*	C/T	0.002/0.01	0.01/–	0.28 (0.14–0.55)	2.7×10^{-5}
rs735320	3:42915878	Pro477Pro	*CYP8B1*	T/C	0.19/0.16	0.17/0.16	1.27 (1.13–1.43)	3.6×10^{-5}
rs2041388	12:6562836	Pro173Pro	*TAPBPL*	A/G	0.06/0.05	0.01/0.29	1.52 (1.24–1.85)	4.0×10^{-5}
rs12999160	2:186661567	Cys3324Tyr	*FSIP2*	A/G	0.02/0.01	0.01/0.09	2.13 (1.46–3.10)	4.3×10^{-5}
rs6003217	22:43870800	Ser197Ser	*MPPED1*	A/G	0.28/0.24	0.30/0.01	1.24 (1.11–1.37)	5.9×10^{-5}
rs3735319	7:149152770	Val115Ala	*ZNF777*	A/G	0.46/0.41	0.44/0.44	1.20 (1.09–1.31)	7.2×10^{-5}
rs62246603	3:42781276	Leu338Leu	*CCDC13*	T/G	0.17/0.14	0.13/0.27	1.27 (1.13–1.43)	8.9×10^{-5}
rs112002818	16:57935442	Ala961Val	*CNGB1*	A/G	0.02/0.008	0.01/0.07	2.29 (1.48–3.54)	9.0×10^{-5}

[a]Risk allele frequencies for cases and controls.
[b]Risk allele frequencies for African and European populations from the 1000 Genomes Project.
[c]ORs and 95% CIs are presented from a logistic regression model adjusted for age, study and PC1-10.
[d]P-values are presented from a likelihood ratio test adjusted for age, study and PC1-10.

10^{-6}) and *PTPRR* on 12q15 (rs73341069: control freq 0.03, *Val239Ile*, OR = 1.62, $P = 2.5 \times 10^{-5}$). The 10 most significantly associated variants are listed in Table 2. Of note, two of the variants are polymorphic in populations of African ancestry and monomorphic in European populations (rs73341069, *Val239Ile*; rs148679475, *Leu104Leu*), and one variant has a minor allele frequency of 30% in African populations, compared with 1% in European populations (rs6003217, *Ser197Ser*). Overall, none of the single variant associations reached exome-wide significance ($P < 3.75 \times 10^{-7}$) after adjustment for multiple comparisons.

In an attempt to replicate these findings, we analyzed the top 10 significant variants and overall prostate cancer risk in a replication set of 3069 cases and 2850 controls of African ancestry from the African Ancestry Prostate Cancer GWAS Consortium (AAPC; see Materials and Methods); however, none of the top 10 significantly associated variants were replicated at $P < 0.05$ in this replication set (Supplementary Material, Table S3a). Six of the 10 variants were genotyped in AAPC, 3 variants with a MAF of >1% had imputation quality scores of >0.80 and 1 rare variant was imputed with a quality score of 0.63 (Supplementary Material, Table S3a). To examine the 20% of variants that were observed in the ESP but removed from our data owing to post-calling quality control filters, we imputed missing data down to 0.5% frequency with a linkage disequilibrium (LD)-aware caller (29). Imputation allowed us to recover 94% of the variants observed in the ESP; however, none were significantly associated with prostate cancer risk.

Case-only analysis

In case–case analyses (611 aggressive, 1054 non-aggressive cases), the two most significant associations were with a synonymous variant in *TRMT1* on 19p13.2 (rs140145761: *Leu83Leu*, OR = 13.55, $P = 5.5 \times 10^{-6}$) and a NS variant in *SNTN* on 3p14 (rs73111385: *Lys52Arg*, OR = 7.45, $P = 1.3 \times 10^{-5}$). Of the top 10 associations from the case–case analysis, 9 were significantly associated ($P < 0.05$) with aggressive disease (versus controls), whereas 4 SNPs were associated with non-aggressive disease (Table 3). One of these associations was replicated, albeit weakly, in the AAPC case–case analysis of 528 aggressive and 2541 non-aggressive cases (rs118023699, OR = 3.81, $P = 0.03$) but was not significantly associated with aggressive or non-aggressive disease at a $P < 0.05$ (Supplementary Material, Table S3b). Two of the 10 variants were genotyped in AAPC, 6 variants had imputation quality scores of ≥0.80, 1 rare variant was imputed with a quality score of 0.31, and 1 variant was monomorphic and was not analyzed (Supplementary Material, Table S3b).

Young onset disease: case–control analysis

In the young onset disease analysis (154 cases age ≤ 55 and 1625 controls), there were two variants in *HYLS*, that just surpassed exome-wide significance on 11q24 (rs78786765, *Lys91Asn*, OR = 0.48, $P = 3.0 \times 10^{-7}$ and rs12274443, *Asn9Asn*, OR = 0.47, $P = 3.1 \times 10^{-7}$, Supplementary Material, Table S4). These variants are correlated in 1KGP AFR ($r^2 = 1.0$) and are less common in young onset cases (frequency = 0.003) than controls (frequency = 0.03). These associations were not replicated in the AAPC young onset disease analysis of 659 cases and 2850 controls. Three of the 10 variants were genotyped in AAPC and 7 variants had imputation quality scores of ≥0.93 (Supplementary Material, Table S3c).

Gene-level analyses

We performed gene-level tests using a gene-sum test, which assumes all variants have the same direction of effect, and the sequence kernel association test (SKAT), which allows for variants to either be protective or confer risk (30). We have limited the

Table 3. Single variant association results for case–case analysis (611 aggressive cases/1054 non-aggressive cases)

Variant	Chromosome, base pair	Amino acid change	Gene	Risk/ref allele	Agg/non-agg[a]	Afr/Eur[b]	Case–case[c] OR (95% CI), *P*-value	Agg versus Ctrl[d] *P*-value	Non versus Ctrl[d] *P*-value
rs140145761	19:13220803	Leu83Leu	*TRMT1*	G/C	0.01/0.001	0.01/–	>10[e] 5.5×10^{-6}	1.4×10^{-2}	4.9×10^{-3}
rs73111385	3:63645410	Lys52Arg	*SNTN*	G/A	0.02/0.002	0.01/0.03	7.45 (2.51–22), 1.3×10^{-5}	5.7×10^{-3}	1.5×10^{-2}
rs2305772	19:52033742	Pro246Ser	*SIGLEC6*	G/A	0.39/0.46	0.46/0.41	0.74 (0.64–0.85), 1.8×10^{-5}	8.8×10^{-4}	1.4×10^{-1}
rs2164808	2:25377176	Tyr807Tyr	*EFR3B*	T/C	0.13/0.20	0.11/0.40	0.67 (0.56–0.81), 2.0×10^{-5}	1.1×10^{-3}	1.1×10^{-1}
rs138602074	1:151006691	Ala448Val	*PRUNE*	T/C	0.01/0.0005	0.01/0.01	>10[e], 2.1×10^{-5}	1.4×10^{-1}	9.5×10^{-4}
rs10841611	12:20903757	His531His	*SLCO1C1*	C/T	0.21/0.28	0.20/0.50	0.71 (0.60–0.84), 3.2×10^{-5}	2.5×10^{-3}	1.2×10^{-1}
rs17767238	14:65207819	Ala472Ala	*PLEKHG3*	T/C	0.03/0.01	0.01/0.06	2.98 (1.74–5.11), 3.2×10^{-5}	8.1×10^{-4}	2.3×10^{-1}
rs118023699	8:144812633	Ser40Ser	*FAM83H*	T/C	0.01/0.002	–/0.02	8.36 (2.41–29), 4.7×10^{-5}	2.1×10^{-2}	1.1×10^{-2}
rs201921601	2:242695307	Arg261Gln	*D2HGDH*	A/G	0.009/0.0005	0.01/–	>10[e], 5.1×10^{-5}	1.5×10^{-4}	3.8×10^{-1}
rs377195382	11:34173968	Arg1015His	*ABTB2*	T/C	0.01/0.0005	–/–	>10[e], 6.0×10^{-5}	1.3×10^{-3}	1.1×10^{-1}

[a] Risk allele frequencies for aggressive and non-aggressive cases, respectively.

[b] Risk allele frequencies for African and European populations from the 1000 Genomes Project.

[c] ORs and 95% CIs are presented from a logistic regression model; *P*-values are reported from a likelihood ratio test; analyses are adjusted for age and PC1-10.

[d] Likelihood ratio test *P*-values are reported for aggressive cases compared with controls and non-aggressive cases compared with controls.

[e] Unable to estimate stable effects and 95% CIs because of the very small allele frequency in non-aggressive cases.

Table 4. Top associations for gene-sum test and the respective *P*-value in the SKAT analysis in all cases and controls (1938 cases/1838 controls)

Gene	Count[a]	Gene-sum freq (ca/ctrl)[b]	OR (95% CI)[c]	*P*-value[d]	P-SKAT
C1orf100	12	0.0007/0.0017	0.49 (0.34–0.72)	2.2×10^{-4}	9.0×10^{-1}
GORAB	12	0.0018/0.0008	2.02 (1.38–2.94)	2.3×10^{-4}	7.6×10^{-2}
DIDO1	75	0.0019/0.0014	1.27 (1.12–1.43)	2.5×10^{-4}	7.8×10^{-2}
NR4A2	3	0.0014/0.00009	12.14 (1.64–90)	3.4×10^{-4}	6.6×10^{-1}
C11orf35	10	0.0016/0.0030	0.58 (0.43–0.78)	4.3×10^{-4}	6.3×10^{-1}
THAP9	16	0.0021/0.0012	1.68 (1.26–2.25)	4.6×10^{-4}	9.3×10^{-2}
CCDC33	18	0.0021/0.0031	0.70 (0.57–0.85)	4.8×10^{-4}	9.2×10^{-1}
SYTL3	15	0.0032/0.0019	1.49 (1.19–1.86)	5.1×10^{-4}	7.4×10^{-1}
TEX9	5	0.0032/0.0012	2.13 (1.36–3.33)	5.4×10^{-4}	4.7×10^{-2}
REG3A	3	0.0003/0.0016	0.17 (0.05–0.57)	7.6×10^{-4}	7.7×10^{-1}

[a] The count of variants included in the gene-level tests.

[b] The frequency of all variants contributing to the gene-sum score in cases and controls.

[c] OR and 95% CIs presented from a logistic regression model.

[d] *P*-value from a likelihood ratio test.

gene-level tests to NS, stoploss; gain or splicing variants with a MAF of <0.01 within each gene and present results from the gene-sum test. We have also included the *P*-value from SKAT for the top 10 associations from the gene-sum test results (Table 4), and the top 100 associations for the overall case–control analysis from the SKAT test are provided in Supplementary Material, Table S5. The lambda for the gene-sum test was 1.04 whereas the SKAT test showed significant over-dispersion (lambda = 1.59). All gene-level testing was corrected for population structure by applying genomic control (see Discussion).

Overall prostate cancer risk

In the gene-sum test, we tested 16 751 genes and observed no gene with a *P*-value of $<10^{-4}$ (two expected) and 19 genes with a *P*-value of $<10^{-3}$ (17 expected). The two most significant genes were *C1orf100* ($P = 2.2 \times 10^{-4}$) and *GORAB* ($P = 2.3 \times 10^{-4}$, Table 4). Neither gene was significant in the SKAT analysis (Table 4).

Case-only analysis

The two most significant associations in the case–case analysis were observed with *SNTN*, a gene involved in calcium ion binding ($P = 2.0 \times 10^{-5}$) and *ZBTB46*, a zinc finger gene that encodes a zinc finger protein ($P = 2.3 \times 10^{-4}$, Table 5). The top nine most significant genes in the case–case analysis were also marginally significant in the analysis of aggressive cases versus controls; however, only four genes were marginally significant in non-aggressive disease, but with different directions of the ORs (Table 5).

High-risk genes

We examined 30 candidate prostate cancer genes, which consist mainly of DNA repair genes (31) as well as 8 additional genes previously implicated in prostate cancer (*HOXB13, KLK2, KLK3, MSMB, MYH6, RAD51D, RNASEL, TEP1*). In the overall case–control and aggressive analyses, there were no significant findings in the gene-sum test. In the case–case analysis, the most significant associations were observed with *MLH1* and *ATM* (P = 0.02 and P = 0.03, respectively, Supplementary Material, Table S6).

Genes near known risk loci

We examined genes nearest to the 100 known loci for prostate cancer risk and in the overall analysis the strongest association was with *ARMC2* (P = 0.006). As other genes of interest could be located within an LD block of a risk variant, we also examined 940 genes within 1 MB of the 100 known loci for prostate cancer risk (Supplementary Material, Table S7). In the overall analysis, the most significant association was observed with *DIDO1*

Table 5. Top 10 gene associations in the case–case gene-sum test (611 aggressive cases/1054 non-aggressive cases) and the respective results for aggressive and non-aggressive disease compared with controls ($N = 1{,}625$ controls)

Gene	Count[a]	Case–case analysis		Aggressive versus controls		Non-aggressive versus controls		Frequencies		
		OR (95% CI)[b]	P-value[c]	OR (95% CI)[b]	P-value[c]	OR (95% CI)[b]	P-value[c]	Agg_case[d]	Non-agg_case[d]	Control[d]
SNTN	2	6.53 (2.43–17.57)	2.0×10^{-5}	2.11 (1.2–3.71)	1.3×10^{-2}	0.34 (0.13–0.86)	1.1×10^{-2}	0.0082	0.0012	0.0042
ZBTB46	2	[e]	2.3×10^{-4}	3.09 (1.10–8.66)	3.3×10^{-2}	[e]	1.6×10^{-2}	0.0033	–	0.0009
ARV1	2	[e]	3.6×10^{-4}	3.03 (1.21–7.61)	1.9×10^{-2}	[e]	1.1×10^{-2}	0.0037	–	0.0011
CYFIP1	25	0.51 (0.35–0.75)	4.1×10^{-4}	0.53 (0.37–0.77)	4.0×10^{-4}	1.06 (0.85–1.33)	6.2×10^{-1}	0.0011	0.0022	0.0020
FBXW9	13	2.34 (1.47–3.72)	4.7×10^{-4}	1.94 (1.31–2.87)	1.6×10^{-3}	0.80 (0.52–1.22)	3.1×10^{-1}	0.0026	0.0011	0.0014
MYH7	16	0.29 (0.14–0.63)	5.2×10^{-4}	0.37 (0.18–0.77)	3.1×10^{-3}	1.17 (0.80–1.71)	4.3×10^{-1}	0.0004	0.0014	0.0012
C9orf47	6	0.09 (0.01–0.64)	5.2×10^{-4}	0.14 (0.02–1.01)	5.2×10^{-3}	1.19 (0.71–1.99)	5.3×10^{-1}	0.0001	0.0017	0.0014
SERPINB9	9	2.76 (1.54–4.94)	5.5×10^{-4}	2.06 (1.28–3.32)	3.9×10^{-3}	0.68 (0.38–1.20)	1.9×10^{-1}	0.0028	0.0009	0.0013
ASH1L	32	0.48 (0.31–0.75)	6.4×10^{-4}	0.56 (0.36–0.87)	7.1×10^{-3}	1.21 (0.94–1.57)	1.6×10^{-1}	0.0006	0.0014	0.0011
ZSWIM6	16	0.37 (0.20–0.69)	6.7×10^{-4}	0.64 (0.34–1.18)	1.4×10^{-1}	1.59 (1.10–2.30)	1.7×10^{-2}	0.0006	0.0017	0.0012

[a] The count of variants included in the gene-level tests.
[b] OR and 95% CIs presented from a logistic regression model.
[c] P-value from a likelihood ratio test.
[d] The frequency of all variants contributing to the gene-sum score in aggressive cases, non-aggressive cases and controls.
[e] Unable to estimate stable effects and 95% CIs because the variant was not observed in non-aggressive cases.

($P = 3.0 \times 10^{-4}$), which is 493 Kb away from rs2427345, a known risk SNP for prostate cancer.

Discussion

In this first whole-exome sequencing study of prostate cancer, we examined the hypothesis that genetic variation in protein-coding sequence may have appreciable effects on disease risk in men of African ancestry. More specifically, with 1938 cases and 1838 controls, we were well powered (>80%) to detect effects of 3.0 and 4.0 for alleles of 1 and 0.5%, respectively. While these effect sizes are large, they are similar to those observed for the *HOXB13* mutation *Gly84Glu* found in men of European ancestry, which has an effect size of 2.5–4.5 for sporadic disease (26). In this study, we did not identify any single variant or a combination of rare alleles within a gene to be associated with such large effects for prostate cancer. These findings are consistent with our initial multiethnic study of 4376 unselected (i.e. sporadic) cases and 7545 controls in which we failed to identify any strong associations with single variants or aggregate effects of rare coding variants in genes from the Exome chip, with the content selected primarily from populations of European ancestry (32).

In this study, we sequenced a large number of individuals to increase the probability of ascertaining rare alleles, rather than sequencing a smaller sample to high-coverage where the minor allele of low-frequency variants would most likely not be observed (33). For a fixed cost, such a design has been demonstrated (via simulations) to have greater statistical power than higher coverage in a smaller sample (34). We recognize that this strategy results in a reduced rate of detection of the rarest variants (singletons and doubletons especially) versus a high-coverage design. There is also a tradeoff with this approach in that variants (regardless of frequency) in regions with very low coverage have a higher probability of being excluded as a result of poor call rate. We applied a conservative call rate filter and removed low-quality variants, which resulted in ~20% of detectable variation >0.5% observed in the ESP being excluded from our analysis. However, we were able to impute missing calls and low-quality variants to recover 94% of these variants down to 0.5% frequency. Despite this limitation, we were able to identify a large fraction of variants that had not been reported previously (~60% with frequencies of ≤0.1%), which highlights the importance and tradeoff of sample size versus high-coverage in rare variant discovery.

Eight of the 10 most statistically significantly associated variants we observed in the overall case–control analysis had MAFs of >1% in African and European populations, which is clearly the spectrum of variation that we had the greatest statistical power to examine. One might expect that the large ORs observed with such variants, if real, to have been identified previously in many of the large-scale prostate cancer GWAS in European ancestry populations, which employed imputation to HapMap or 1KGP. We attempted to replicate our top associations in 14 160 prostate cancer cases and 12 712 controls of European ancestry from the ELLIPSE/GAME-ON Consortium, which consists of 5 independent studies/consortia (see description in Materials and Methods, detailed description of participating studies in the Supplementary Material, Note). Eight of the 10 coding variants were available for replication (rs73341069 and rs148679475 are monomorphic in European populations); however, we did not replicate any association at $P < 0.05$. We also attempted to replicate these findings in individuals of African ancestry from the AAPC replication set and no variant replicated at a $P < 0.05$. SNPs with a MAF of >1% were well-imputed (quality > 0.80, mean quality = 0.93); however, there were rare variants (rs112002818 and rs201921601) with low-quality scores (0.63 and 0.31, respectively), which potentially decreased the power to replicate the associations.

One aspect of rare variant discovery within individuals of African ancestry that warrants discussion is the potential effect of fine-scale population stratification. As African populations carry a larger number of rare variants as compared with European populations, fine-scale population stratification and admixture can have a greater influence on association tests for rare variation because a rare allele may be limited to small-scale groups of related individuals (i.e. those containing a local ancestral haplotype), which may not be captured by global ancestry estimation via principal components (PC) (35–37). While this is an active area of research, it has been shown that PCs are still an acceptable way to control for confounding owing to population stratification from both common and rare variation at the level of global ancestry (38). To address this in our data, we calculated PCs in two ways: with only common variants included (MAF ≥ 5%) and again with all variants included down to 0.2% in attempt to capture more fine-scale variability in population substructure. The association results were similar for each set of PCs. We do not believe that residual confounding by fine-scale population structure is an issue for the single variant tests as we have filtered out all variants with a MAF of <0.2%; however, in the SKAT analysis, we found an over-dispersion of significant genes (lambda = 1.59) which we believe could be due to this issue. Previous research via simulations of subtle population geographic/ancestral differences has shown inflation in *P*-values from joint tests (i.e. SKAT) that allow variants to have effects in opposite direction (39). Given this potential, we have corrected the results for all gene-level tests (gene-sum and SKAT analyses) by applying genomic control (40,41).

The higher rate of prostate cancer in men of African ancestry may be due in part to alleles that are found only in this population. We attempted to address this question by examining coding variants that are only found in men of African ancestry. Using data from ESP, we identified ∼20 000 rare coding variants that were found in the ESP African American sample (*n* = 2203) as well as in our study but were not polymorphic in the ESP European Ancestry sample (*n* = 4300). In examining these African-specific variants in our study, we found no evidence of an over-representation of more significant associations than expected. However, much larger sample sizes will be needed to examine the contribution of very rare (<0.2%) population-specific alleles to differences in risk across racial/ethnic populations.

With respect to understanding disease heritability, ORs between 2 and 6 would be expected if rare coding variants (0.1–1%) make a similar contribution as the 100 common variants identified to date (32). The inability to identify coding variants with such effect sizes suggests that the contribution of coding variants to overall prostate cancer heritability may be minor. Rare variants in the protein-coding sequence could still be important in disease risk, but with more moderate-to-small effect sizes thus requiring substantially larger samples sizes to detect in single variant or gene-level tests. For example, Zuk *et al.* describes a rare variant association study will require 25 000 cases for the discovery set, with a large independent replication set to provide 90% power to detect modest effects through burden testing of genes with ORs as low as 2 (42).

Another limitation of this study is that we did not investigate indels that could result in protein truncation mutations, which may be pathogenic and have been shown to be important in prostate cancer. For example, 2% of men of European ancestry with early-onset prostate cancer have been found to carry protein truncation mutations in *BRCA2* (43), and more recent studies find deletions or frameshift mutations in *BRCA2* to be associated with a more aggressive phenotype (44). Recently, Leongamornlert *et al.* reported 14 putative loss of function mutations in DNA repair genes associated with familial and aggressive disease in 191 men with three or more prostate cancer cases in their family (31). We did not observe striking evidence of associations in individual variant associations or gene-level tests with these genes, although as stated earlier, we were unable to study variants that occurred in only one individual owing to our low depth of coverage and we did not examine loss of function protein truncation mutations. Accurately calling indels presents a technological challenge in next-generation sequencing and is further complicated in whole-exome sequencing, where an additional hybridization step can lead to a reference strand bias, which results in less efficient coverage of the non-reference read (45). It is known that loss-of-function mutations are enriched for false positives (46,47), and the ability to accurately call indels decreases with lower coverage. O'Rawe *et al.* compared three variant calling software tools and found 28.6% indel concordance across the three callers with a validation rate ranging from 44.6 to 78.1% (48). Future work is needed to call indels in this sample.

In this study, we focused on the 1–2% of the genome comprised of protein-coding sequence with a strong prior for having variation that might have a more serious impact on disease biology. However, based on what we have learned from GWAS where the vast majority of risk alleles are in non-protein-coding sequence, it is equally likely that rare variation in non-coding sequence could also have an important role in cancer susceptibility (22,23). One such example is a non-coding variant at a known susceptibility locus on 8q24, which is rare in populations of European ancestry (rs183373024, MAF = 0.5%) and has a sizeable effect (OR = 2.9) (49). This variant maps to transcription-factor binding sites of the androgen receptor and FoxA1, and binding specificity is altered by the risk allele (50). A second example at 8q24 is with rs116041037, a non-coding variant that is polymorphic in African Americans only (MAF = 2%) and has a large effect on prostate cancer risk (OR = 2.5) (4). High-coverage whole-genome sequencing will be required to better understand the contribution of rare variation (MAF < 1%) in non-coding regions of the genome.

In summary, in the first whole-exome sequencing study of prostate cancer in men of African ancestry, our results do not support the hypothesis that there are NS variants of ≥0.5% in frequency with large odds ratios. These data provide an invaluable resource that has already contributed population-specific content for custom array design (the Illumina MEGA SNP Chip). Future sequencing efforts in much larger sample sizes will be needed to elucidate the role of rare variation in prostate cancer susceptibility.

Materials and Methods

Ethics statement

All work has been performed under national and international guidelines. Written consent was obtained for all participants at the time of blood/saliva collection. The Institutional Review Board at the University of Southern California and at Makerere University approved the study protocol.

Study population

The men in this study were from the Multiethnic Cohort and the Uganda Prostate Case Control Study. There were also additional studies used for quality control assessment and as replication sets. These studies are described later.

The Multiethnic Cohort

The Multiethnic Cohort (MEC) is comprised of over 215 000 men and women recruited from Hawaii and the Los Angeles area between 1993 and 1996 and has been described elsewhere in detail (51). Participants are primarily of Native Hawaiian, Japanese, European American, African American, or Latino ancestry, and were between the ages of 45 and 75 at baseline at which time they completed a detailed questionnaire to collect information on demographics and lifestyle factors, including diet and medical conditions. Between 1995 and 2006, over 65 000 blood samples were collected from participants for genetic analyses. To identify incident cancer cases, the MEC was cross-linked with the population-based Surveillance, Epidemiology and End Results (SEER) registries in California and Hawaii, and unaffected cohort participants with blood samples were selected as controls. Information on stage and grade of disease were also obtained through SEER. Cases and controls were identified through 2012, and the case–control study of prostate cancer in African American men included 1833 incident cases and 1799 controls.

Uganda Prostate Cancer Study

The Uganda Prostate Cancer Study (UGPCS) is a case–control study of prostate cancer in Kampala Uganda that was initiated in 2011. Men with prostate cancer were enrolled from the Urology unit at Mulago Hospital and men without prostate cancer (i.e. controls) were enrolled from other clinics (i.e. surgery) at the hospital. All patients meeting the inclusion criteria (cases: ≥39 years of age; controls: ≥39 years of age, PSA level < 4 ng/ml to rule out undiagnosed prostate cancer) and willing to give consent were recruited into the study. Written consent is obtained and two identical informed consent forms translated into Luganda are provided to each participant for them to read or to be read to them, sign or thumb print. After enrollment, each study participant was interviewed using a standardized questionnaire to collect descriptive and prostate cancer risk factor information. A biospecimen was collected using the Oragene saliva collection kit. As of 31 December 2012, UGPCS included 332 cases and 235 controls which were included in this study.

Exome SNP chip in the MEC

The Illumina HumanExome SNP array was used as part of a previous multiethnic study of breast and prostate cancer in the MEC (32). After quality control measures were implemented, 191 032 common and rare variants were analyzed in 4376 prostate cancer cases and 7545 controls and 2984 breast cancer cases and 7545 controls. There were 1117 cases and 2146 controls of African ancestry included in the prostate cancer analysis, 2100 of which were also sequenced as part of the current study. Concordance between sequence and genotype data was evaluated in this sample to set QC metrics (i.e. filtering criteria). Details of the QC measures employed in the Exome SNP chip analysis have been previously described (32).

African ancestry replication set

The AAPC GWAS Consortium, which consists of 14 independent studies (Supplementary Material, Note) and has been described elsewhere in detail (52,53), was utilized as the replication sample. Samples were genotyped using the Illumina Infinium 1M-Duo bead array, and imputation was performed using IMPUTE2 (v2.2, https://mathgen.stats.ox.ac.uk/impute/impute_v2.html), using the October 2014 release of 1KGP as the reference set. All samples and SNPs with a call rate of <95% were removed. After samples from the MEC were removed for the purpose of the replication analysis, there were 3069 cases and 2850 controls from 13 studies available for replication.

The ELLIPSE/GAME-ON Consortium

The ELLIPSE Consortium is focused on the identification and functional follow-up of prostate cancer susceptibility loci within the GAME-ON initiative. The replication set used for this analysis consisted of 14 160 cases and 12 712 controls of European ancestry from five major studies/consortia (described in detail in the Supplementary Material, Note). All genotyping and imputation quality metrics have been described elsewhere (20).

Exome capture and sequencing

Library preparation and enrichment

We utilized a method developed by Rohland and Reich (54), which creates cost-effective DNA-sequencing libraries suitable for multiplexed target capture. This high-throughput method parallelizes the library preparation in 96-well plates and attaches internal barcodes directly to fragmented DNA from a sample to allow for multiplexed sample pooling for target enrichment via hybridization without a substantial loss in capture efficiency. We processed plates of samples that were randomized with respect to case–control status. Pools of eight libraries each were prepared in equimolar concentrations and enriched using the Agilent SureSelect All Exon kit version 4, targeting a 51 Mb region designed to capture 20 965 genes and 334 278 exons. Sequencing was conducted at Illumina (San Diego, CA, USA) using HiSeq 2000 instruments for 100 cycles paired end sequencing.

We aimed to sequence to an average coverage of 10 × of the 51 Mb targeted regions. While these exomes are low in coverage, most of the information relevant to disease gene mapping comes from the first few-fold coverage of samples, and higher-coverage data are more redundant per sequencing rate.

Alignment and genotype calling

Sequences were aligned to the human genome reference sequence (hg19) using BWA version 0.6.1 (55). Variants were called using the GATK best practices workflow (56), including mapping the raw reads to the human genome reference sequence (hg19), base recalibration and compression, and joint calling and variant recalibration. The only change implemented was to keep 2 base pairs (bp) around the target region instead of the standard 50 bp recommended, as the sequence data outside the targeted region did not yield high-quality calls.

Sample and variant filtering

There were 727 262 variants identified in coding regions before any post-variant calling quality control or allele-count filtering. Variants with a call rate of <85% (n = 166 527), and individuals with a call rate of <80% (n = 307) were removed. To determine appropriate quality control filters, genotypes from a subset of individuals (n = 2100) genotyped on the Illumina Human Exome BeadChip (described earlier) (32) were compared with the sequence variant calls. Assuming the array data as the gold standard, concordance was calculated (n = 94 796) across various quality control measures and cut points to better understand filters that should applied with the highest sensitivity and specificity. Filtering the data using a QUAL score of >20 and a minor allele count (MAC) of four or more (MAF ∼0.05%) removed 162 645 variants and retained the most accurate data, with sample concordance of 99.7%. The resulting variants (n = 398 090) were annotated using ANNOVAR (57) to identify exonic, splicing, and stop-loss; gain variants. There were 2870 variants that could not be annotated or mapped to multiple positions in the genome and were removed. There were 59 samples that failed sequencing. Twenty-five unintended replicates (UGPCS) and 32 samples that did not have data available to calculate PCs (discussed later) were removed from the analysis. Following quality control filtering, 395 220 variants in coding regions and 3 776 individuals (1938 cases and 1838 controls) sequenced at a mean coverage of 10.1× were available for analysis.

Statistical analysis

Association tests for single variants

For each variant, analyses were conducted using a likelihood ratio test adjusting for age, study, and 10 PCs, assuming a log-additive model. We tested associations in all cases and controls, aggressive cases and controls, non-aggressive cases and controls, in a case–case analysis, and in young onset cases (≤55) compared with all controls. We report ORs and 95% CIs from logistic regression and P-values calculated from a likelihood ratio test. In this study, aggressive disease was defined as metastatic disease (stage = 4), a Gleason score of ≥8, PSA of >100, or death from prostate cancer ($n = 611$). Non-aggressive disease was defined as non-metastatic disease (stage = 1–3) and a Gleason score of <8 ($n =$ 1054). We also examined a more stringent non-aggressive phenotype defined as localized disease (stage = 1) and Gleason of <8 ($n = 866$). Using this more stringent definition also did not reveal any single variant or gene-level test reaching exome-wide significance. PCs were calculated using SNPs from a parallel sequencing effort of ~70 known prostate cancer risk loci in these same individuals (58). All SNPs in LD with any known risk SNP were removed ($r^2 > 0.2$) as were SNPs with a call rate of <99.5%. LD-pruning (if $r^2 > 0.2$) resulted in 12 494 independent SNPs with a MAF of ≥0.2% for use in calculating PCs (59). Only MEC participants were included in the analyses by aggressiveness and young onset disease, as stage and Gleason grade was not available for UGPCS. All coding variants were analyzed, which included NS, synonymous, stop-loss or stop-gain and splicing site variants, and the α-level for genome-wide statistical significance was 3.75×10^{-7} after applying a Bonferroni correction for testing 133 367 variants. All statistical analyses were conducted using PLINK v1.07 (60) and the R statistical computing platform. Results for the 100 most significant associations are provided in Supplementary Material, Table S4. To explore the role of variants filtered in the quality control process, we also performed LD-aware genotype calling starting from the genotype likelihoods estimated by GATK using Beagle (29). Single variant analyses were performed using the imputed dosages.

Gene-level testing

The cumulative effects of rare putatively functional variants (NS, stop or splice variants with a MAF of ≤1%) within each gene were tested using a gene-sum test, where minor alleles were summed across genes in each individual and analyzed as the independent variable in a case–control analysis. This model assumes that each variant affects the phenotype in a similar direction. Gene-level testing was also performed using SKAT (30), a variance components test that does not assume each variant influences the phenotype in the same direction; however, results are discussed within the context of the most significant findings from the gene-sum test. In total, we tested 16 751 genes, and used an α-level of 3.0×10^{-6} to determine global significance after applying a Bonferroni correction and genomic control corrections were applied to each gene-level test. All variants with a MAC of >4 (in all samples) were included in gene-level tests ($n = 395\,220$). Gene-level analyses were performed in all cases and controls, by disease aggressiveness (in aggressive cases compared with controls and non-aggressive cases compared with controls), and in a case–case analysis (aggressive versus non-aggressive disease). Gene-sum tests were calculated using a likelihood ratio test, adjusted for age, study (overall gene-level tests) and PC1-10. All statistical analyses for gene-sum testing and SKAT were conducted using the R statistical computing platform. The top 100 most significant genes for all gene-level analyses are provided in Supplementary Material, Table S8.

Data access

The data reported in this study are available at the database of Genotypes and Phenotypes (dbGaP) under data accession phs000306.

Supplementary Material

Supplementary Material is available at *HMG* online.

Acknowledgements

We are forever indebted to Dr Brian Henderson, who passed away before this paper was published. Without his efforts in co-founding the MEC, this work would not have been possible. We would also like to acknowledge all of the men who participated in these studies.

Conflict of Interest statement. None declared.

Funding

This work was supported by NIH grants R01 CA 165862, UM1 CA 164973 and RC2 CA 148085. K.A.R. gratefully acknowledges Gretchen Ponty Smith and is supported, in part, by the Margaret Kersten Ponty postdoctoral fellowship endowment, Achievement Rewards for College Scientists (ARCS) Foundation, Los Angeles Founder Chapter.

The African Ancestry Prostate Cancer GWAS Consortium is supported by NIH grants CA1326792, CA148085 and HG004726.

The MEC was supported by NIH grants CA164973, CA63464 and CA54281.

Genotyping of the PLCO samples was funded by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics, NCI, NIH.

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

KCPCS was supported by NIH grants CA056678, CA082664 and CA092579, with additional support from the Fred Hutchinson Cancer Research Center and the Intramural Program of the National Human Genome Research Institute.

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

GECAP was supported by NIH grant ES011126.

CaP Genes was supported by CA88164 and CA127298.

IPCG was support by DOD grant W81XWH-07-1-0122.

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

CPS-II is supported by the American Cancer Society.

SELECT is funded by Public Health Service cooperative Agreement grant CA37429 awarded by the National Cancer Institute, National Institutes of Health.

SCCS is funded by NIH grant CA092447. SCCS sample preparation was conducted at the Epidemiology Biospecimen Core Lab that is supported in part by the Vanderbilt Ingram Cancer Center (CA68485). Data on SCCS cancer cases used in this publication were provided by the Alabama Statewide Cancer Registry; Kentucky Cancer Registry; Tennessee Department of Health, Office of Cancer Surveillance; Florida Cancer Data System; North Carolina Central Cancer Registry, North Carolina Division of Public Health; Georgia Comprehensive Cancer Registry; Louisiana Tumor Registry; Mississippi Cancer Registry; South Carolina Central Cancer Registry; Virginia Department of Health, Virginia Cancer Registry; Arkansas Department of Health, Cancer Registry. The Arkansas Central Cancer Registry is fully funded by a grant from National Program of Cancer Registries, Centers for Disease Control and Prevention (CDC). Data on SCCS cancer cases from Mississippi were collected by the Mississippi Cancer Registry which participates in the National Program of Cancer Registries (NPCR) of the Centers for Disease Control and Prevention (CDC). The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the CDC or the Mississippi Cancer Registry.

The ELLIPSE (Elucidating Loci in Prostate Cancer Susceptibility)/GAME-ON Consortium was supported by NCI U19 grant CA148537.

UKGPCS is supported by the Institute of Cancer Research and The Everyman Campaign, Cancer Research UK C5047/A8384, Prostate Cancer Research Foundation (now Prostate Cancer UK), The Orchid Cancer Appeal, The National Cancer Research Network UK, The National Cancer Research Institute (NCRI) UK. We are grateful for support of NIHR funding to the NIHR Biomedical Research Centre at The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust. UKGPCS should also like to acknowledge the NCRN nurses, data managers and Consultants for their work in the UKGPCS study.

CAPS: The Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm, Sweden was supported by the Cancer Risk Prediction Center (CRisP; www.crispcenter.org), a Linneus Centre (Contract ID 70867902) financed by the Swedish Research Council, Swedish Research Council (grant no K2010-70X-20430-04-3, 10-3674), the Swedish Cancer Foundation (grant no 09-0677, 11-484, 12-823), the Hedlund Foundation, the Söderberg Foundation, the Enqvist Foundation, ALF funds from the Stockholm County Council. Stiftelsen Johanna Hagstrand och Sigfrid Linnér's Minne, Karlsson's Fund for urological and surgical research.

The BPC3 was supported by the U.S. National Institutes of Health, National Cancer Institute (cooperative agreements U01-CA98233, U01-CA98710, U01-CA98216 and U01-CA98758, and Intramural Research Program of NIH/National Cancer Institute, Division of Cancer Epidemiology and Genetics). The ATBC study was supported in part by the Intramural Research Program of the NIH and the National Cancer Institute. Additionally, this research was supported by U.S. Public Health Service contracts N01-CN-45165, N01-RC-45035, N01-RC-37004 and HHSN261201000006C from the National Cancer Institute, Department of Health and Human Services.

Genotyping in PEGASUS/PLCO was supported and funded by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics, NCI, NIH.

African Ancestry Prostate Cancer Consortium (AAPC)

Sara S. Strom, Rick A. Kittles, Benjamin A. Rybicki, Janet L. Stanford, Phyllis J. Goodman, Sonja I. Berndt, John Carpten, Graham Casey, Lisa Chu, Ryan W. Diver, Anselm JM Hennis, Eric A. Klein, Suzanne Kolb, Loic Le Marchand, M. Cristina Leske, Adam B. Murphy, Christine Neslund-Dudas, Jong Y. Park, Esther M. John, Adam S. Kibel, Curtis Pettaway, Susan M. Gapstur, S. Lilly Zheng, Suh-Yuh Wu, John S. Witte, Jianfeng Xu, William Isaacs, Sue A. Ingles, Ann Hsing, Barbara Nemesure, William J. Blot, Brian E. Henderson, Christopher A. Haiman.

The ELLIPSE/GAME-ON Consortium

Rosalind A. Eeles, Douglas Easton, Zsofia Kote-Jarai, Kenneth Muir, Ali Amin Al Olama, Fredrik Wiklund, Henrik Grönberg, Peter Kraft, Susan Gapstur, Elio Riboli, David Hunter, Loic Le Marchand, Christopher A. Haiman, Brian E. Henderson, Victoria Stevens, Sonja I. Berndt, Stephen J. Chanock.

References

1. Kolonel, L.N., Altshuler, D. and Henderson, B.E. (2004) The multiethnic cohort study: exploring genes, lifestyle and cancer risk. *Nat. Rev. Cancer*, **4**, 519–527.
2. Hemminki, K. (2012) Familial risk and familial survival in prostate cancer. *World J. Urol.*, **30**, 143–148.
3. Hjelmborg, J.B., Scheike, T., Holst, K., Skytthe, A., Penney, K.L., Graff, R.E., Pukkala, E., Christensen, K., Adami, H.O., Holm, N. V. *et al.* (2014) The heritability of prostate cancer in the nordic twin study of cancer. *Cancer Epidemiol. Biomarkers Prev.*, **23**, 2303–2310.
4. Haiman, C.A., Patterson, N., Freedman, M.L., Myers, S.R., Pike, M.C., Waliszewska, A., Neubauer, J., Tandon, A., Schirmer, C., McDonald, G.J. *et al.* (2007) Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat. Genet.*, **39**, 638–644.
5. Gudmundsson, J., Sulem, P., Rafnar, T., Bergthorsson, J.T., Manolescu, A., Gudbjartsson, D., Agnarsson, B.A., Sigurdsson, A., Benediktsdottir, K.R., Blondal, T. *et al.* (2008) Common sequence variants on 2p15 and Xp11.22 confer susceptibility to prostate cancer. *Nat. Genet.*, **40**, 281–283.
6. Al Olama, A.A., Kote-Jarai, Z., Giles, G.G., Guy, M., Morrison, J., Severi, G., Leongamornlert, D.A., Tymrakiewicz, M., Jhavar, S., Saunders, E. *et al.* (2009) Multiple loci on 8q24 associated with prostate cancer susceptibility. *Nat. Genet.*, **41**, 1058–1060.
7. Eeles, R.A., Kote-Jarai, Z., Al Olama, A.A., Giles, G.G., Guy, M., Severi, G., Muir, K., Hopper, J.L., Henderson, B.E., Haiman, C.A. *et al.* (2009) Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. *Nat. Genet.*, **41**, 1116–1121.
8. Gudmundsson, J., Sulem, P., Gudbjartsson, D.F., Blondal, T., Gylfason, A., Agnarsson, B.A., Benediktsdottir, K.R., Magnusdottir, D.N., Orlygsdottir, G., Jakobsdottir, M. *et al.* (2009) Genome-wide association and replication studies identify four variants associated with prostate cancer susceptibility. *Nat. Genet.*, **41**, 1122–1126.
9. Jia, L., Landan, G., Pomerantz, M., Jaschek, R., Herman, P., Reich, D., Yan, C., Khalid, O., Kantoff, P., Oh, W. *et al.* (2009) Functional enhancers at the gene-poor 8q24 cancer-linked locus. *PLoS Genet.*, **5**, e1000597.
10. Sun, J., Zheng, S.L., Wiklund, F., Isaacs, S.D., Li, G., Wiley, K.E., Kim, S.T., Zhu, Y., Zhang, Z., Hsu, F.C. *et al.* (2009) Sequence variants at 22q13 are associated with prostate cancer risk. *Cancer Res.*, **69**, 10–15.
11. Xu, J., Mo, Z., Ye, D., Wang, M., Liu, F., Jin, G., Xu, C., Wang, X., Shao, Q., Chen, Z. *et al.* (2012) Genome-wide association study in Chinese men identifies two new prostate cancer risk loci at 9q31.2 and 19q13.4. *Nat. Genet.*, **44**, 1231–1235.
12. Kote-Jarai, Z., Olama, A.A., Giles, G.G., Severi, G., Schleutker, J., Weischer, M., Campa, D., Riboli, E., Key, T., Gronberg, H. *et al.* (2011) Seven prostate cancer susceptibility loci identified

by a multi-stage genome-wide association study. *Nat. Genet.*, **43**, 785–791.

13. Lindstrom, S., Schumacher, F., Siddiq, A., Travis, R.C., Campa, D., Berndt, S.I., Diver, W.R., Severi, G., Allen, N., Andriole, G. *et al.* (2011) Characterizing associations and SNP-environment interactions for GWAS-identified prostate cancer risk markers–results from BPC3. *PLoS One*, **6**, e17142.
14. Schumacher, F.R., Berndt, S.I., Siddiq, A., Jacobs, K.B., Wang, Z., Lindstrom, S., Stevens, V.L., Chen, C., Mondul, A.M., Travis, R.C. *et al.* (2011) Genome-wide association study identifies new prostate cancer susceptibility loci. *Hum. Mol. Genet.*, **20**, 3867–3875.
15. Akamatsu, S., Takata, R., Haiman, C.A., Takahashi, A., Inoue, T., Kubo, M., Furihata, M., Kamatani, N., Inazawa, J., Chen, G. K. *et al.* (2012) Common variants at 11q12, 10q26 and 3p11.2 are associated with prostate cancer susceptibility in Japanese. *Nat. Genet.*, **44**, 426–429, S421.
16. Amin Al Olama, A., Kote-Jarai, Z., Schumacher, F.R., Wiklund, F., Berndt, S.I., Benlloch, S., Giles, G.G., Severi, G., Neal, D.E., Hamdy, F.C. *et al.* (2013) A meta-analysis of genome-wide association studies to identify prostate cancer susceptibility loci associated with aggressive and non-aggressive disease. *Hum. Mol. Genet.*, **22**, 408–415.
17. Lindstrom, S., Schumacher, F.R., Campa, D., Albanes, D., Andriole, G., Berndt, S.I., Bueno-de-Mesquita, H.B., Chanock, S.J., Diver, W.R., Ganziano, J.M. *et al.* (2012) Replication of five prostate cancer loci identified in an Asian population–results from the NCI Breast and Prostate Cancer Cohort Consortium (BPC3). *Cancer Epidemiol. Biomarkers Prev.*, **21**, 212–216.
18. Xu, J., Zheng, S.L., Isaacs, S.D., Wiley, K.E., Wiklund, F., Sun, J., Kader, A.K., Li, G., Purcell, L.D., Kim, S.T. *et al.* (2010) Inherited genetic variant predisposes to aggressive but not indolent prostate cancer. *Proc. Natl Acad. Sci. USA*, **107**, 2136–2140.
19. Eeles, R.A., Olama, A.A., Benlloch, S., Saunders, E.J., Leongamornlert, D.A., Tymrakiewicz, M., Ghoussaini, M., Luccarini, C., Dennis, J., Jugurnauth-Little, S. *et al.* (2013) Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. *Nat. Genet.*, **45**, 385–391, 391e381–382.
20. Al Olama, A.A., Kote-Jarai, Z., Berndt, S.I., Conti, D.V., Schumacher, F., Han, Y., Benlloch, S., Hazelett, D.J., Wang, Z., Saunders, E. *et al.* (2014) A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. *Nat. Genet.*, **46**, 1103–1109.
21. Han, Y., Signorello, L.B., Strom, S.S., Kittles, R.A., Rybicki, B.A., Stanford, J.L., Goodman, P.J., Berndt, S.I., Carpten, J., Casey, G. *et al.* (2015) Generalizability of established prostate cancer risk variants in men of African ancestry. *Int. J. Cancer*, **136**, 1210–1217.
22. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.
23. Eeles, R., Goh, C., Castro, E., Bancroft, E., Guy, M., Al Olama, A.A., Easton, D. and Kote-Jarai, Z. (2014) The genetic epidemiology of prostate cancer and its clinical implications. *Nat. Rev. Urol.*, **11**, 18–31.
24. Gayther, S.A., de Foy, K.A., Harrington, P., Pharoah, P., Dunsmuir, W.D., Edwards, S.M., Gillett, C., Ardern-Jones, A., Dearnaley, D.P., Easton, D.F. *et al.* (2000) The frequency of germ-line mutations in the breast cancer predisposition genes BRCA1 and BRCA2 in familial prostate cancer. The Cancer Research Campaign/British Prostate Group United Kingdom Familial Prostate Cancer Study Collaborators. *Cancer Res.*, **60**, 4513–4518.
25. Huang, H. and Cai, B. (2014) G84E mutation in HOXB13 is firmly associated with prostate cancer risk: a meta-analysis. *Tumour Biol.*, **35**, 1177–1182.
26. Laitinen, V.H., Wahlfors, T., Saaristo, L., Rantapero, T., Pelttari, L.M., Kilpivaara, O., Laasanen, S.L., Kallioniemi, A., Nevanlinna, H., Aaltonen, L. *et al.* (2013) HOXB13 G84E mutation in Finland: population-based analysis of prostate, breast, and colorectal cancer risk. *Cancer Epidemiol. Biomarkers Prev.*, **22**, 452–460.
27. Ewing, C.M., Ray, A.M., Lange, E.M., Zuhlke, K.A., Robbins, C.M., Tembe, W.D., Wiley, K.E., Isaacs, S.D., Johng, D., Wang, Y. *et al.* (2012) Germline mutations in HOXB13 and prostate-cancer risk. *N. Engl. J. Med.*, **366**, 141–149.
28. Xu, J., Lange, E.M., Lu, L., Zheng, S.L., Wang, Z., Thibodeau, S.N., Cannon-Albright, L.A., Teerlink, C.C., Camp, N.J., Johnson, A.M. *et al.* (2013) HOXB13 is a susceptibility gene for prostate cancer: results from the International Consortium for Prostate Cancer Genetics (ICPCG). *Hum. Genet.*, **132**, 5–14.
29. Browning, B.L. and Browning, S.R. (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.*, **84**, 210–223.
30. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M. and Lin, X. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, **89**, 82–93.
31. Leongamornlert, D., Saunders, E., Dadaev, T., Tymrakiewicz, M., Goh, C., Jugurnauth-Little, S., Kozarewa, I., Fenwick, K., Assiotis, I., Barrowdale, D. *et al.* (2014) Frequent germline deleterious mutations in DNA repair genes in familial prostate cancer cases are associated with advanced disease. *Br. J. Cancer*, **110**, 1663–1672.
32. Haiman, C.A., Han, Y., Feng, Y., Xia, L., Hsu, C., Sheng, X., Pooler, L.C., Patel, Y., Kolonel, L.N., Carter, E. *et al.* (2013) Genome-wide testing of putative functional exonic variants in relationship with breast and prostate cancer risk in a multiethnic population. *PLoS Genet.*, **9**, e1003419.
33. Li, Y., Sidore, C., Kang, H.M., Boehnke, M. and Abecasis, G.R. (2011) Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res.*, **21**, 940–951.
34. Flannick, J., Korn, J.M., Fontanillas, P., Grant, G.B., Banks, E., Depristo, M.A. and Altshuler, D. (2012) Efficiency and power as a function of sequence coverage, SNP array density, and imputation. *PLoS Comput. Biol.*, **8**, e1002604.
35. Gravel, S., Henn, B.M., Gutenkunst, R.N., Indap, A.R., Marth, G.T., Clark, A.G., Yu, F., Gibbs, R.A. and Bustamante, C.D. (2011) Demographic history and rare allele sharing among human populations. *Proc. Natl Acad. Sci. USA*, **108**, 11983–11988.
36. Fu, W., O'Connor, T.D., Jun, G., Kang, H.M., Abecasis, G., Leal, S.M., Gabriel, S., Rieder, M.J., Altshuler, D., Shendure, J. *et al.* (2013) Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, **493**, 216–220.
37. Mathieson, I. and McVean, G. (2012) Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.*, **44**, 243–246.
38. Zhang, Y., Guan, W. and Pan, W. (2013) Adjustment for population stratification via principal components in association analysis of rare variants. *Genet. Epidemiol.*, **37**, 99–109.
39. Zawistowski, M., Reppell, M., Wegmann, D., St Jean, P.L., Ehm, M.G., Nelson, M.R., Novembre, J. and Zollner, S. (2014) Analysis of rare variant population structure in Europeans explains differential stratification of gene-based tests. *Eur. J. Hum. Genet.*, **22**, 1137–1144.
40. Devlin, B. and Roeder, K. (1999) Genomic control for association studies. *Biometrics*, **55**, 997–1004.

41. Devlin, B., Bacanu, S.A. and Roeder, K. (2004) Genomic control to the extreme. *Nat. Genet.*, **36**, 1129–1130; author reply 1131.
42. Zuk, O., Schaffner, S.F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M.J., Neale, B.M., Sunyaev, S.R. and Lander, E.S. (2014) Searching for missing heritability: designing rare variant association studies. *Proc. Natl Acad. Sci. USA*, **111**, E455–E464.
43. Edwards, S.M., Kote-Jarai, Z., Meitz, J., Hamoudi, R., Hope, Q., Osin, P., Jackson, R., Southgate, C., Singh, R., Falconer, A. *et al.* (2003) Two percent of men with early-onset prostate cancer harbor germline mutations in the BRCA2 gene. *Am. J. Hum. Genet.*, **72**, 1–12.
44. Castro, E., Goh, C., Olmos, D., Saunders, E., Leongamornlert, D., Tymrakiewicz, M., Mahmud, N., Dadaev, T., Govindasami, K., Guy, M. *et al.* (2013) Germline BRCA mutations are associated with higher risk of nodal involvement, distant metastasis, and poor survival outcomes in prostate cancer. *J. Clin. Oncol.*, **31**, 1748–1757.
45. Shao, H., Bellos, E., Yin, H., Liu, X., Zou, J., Li, Y., Wang, J. and Coin, L.J. (2013) A population model for genotyping indels from next-generation sequence data. *Nucl.* Acids Res., **41**, e46.
46. Lescai, F., Bonfiglio, S., Bacchelli, C., Chanudet, E., Waters, A., Sisodiya, S.M., Kasperaviciute, D., Williams, J., Harold, D., Hardy, J. *et al.* (2012) Characterisation and validation of insertions and deletions in 173 patient exomes. *PLoS One*, **7**, e51292.
47. MacArthur, D.G. and Tyler-Smith, C. (2010) Loss-of-function variants in the genomes of healthy humans. *Hum. Mol. Genet.*, **19**, R125–R130.
48. O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., Bodily, P., Tian, L., Hakonarson, H., Johnson, W.E. *et al.* (2013) Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med.*, **5**, 28.
49. Gudmundsson, J., Sulem, P., Gudbjartsson, D.F., Masson, G., Agnarsson, B.A., Benediktsdottir, K.R., Sigurdsson, A., Magnusson, O.T., Gudjonsson, S.A., Magnusdottir, D.N. *et al.* (2012) A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. *Nat. Genet.*, **44**, 1326–1329.
50. Hazelett, D.J., Coetzee, S.G. and Coetzee, G.A. (2013) A rare variant, which destroys a FoxA1 site at 8q24, is associated with prostate cancer risk. *Cell Cycle*, **12**, 379–380.
51. Kolonel, L.N., Henderson, B.E., Hankin, J.H., Nomura, A.M., Wilkens, L.R., Pike, M.C., Stram, D.O., Monroe, K.R., Earle, M.E. and Nagamine, F.S. (2000) A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. *Am. J. Epidemiol.*, **151**, 346–357.
52. Haiman, C.A., Chen, G.K., Blot, W.J., Strom, S.S., Berndt, S.I., Kittles, R.A., Rybicki, B.A., Isaacs, W.B., Ingles, S.A., Stanford, J.L. *et al.* (2011) Genome-wide association study of prostate cancer in men of African ancestry identifies a susceptibility locus at 17q21. *Nat. Genet.*, **43**, 570–573.
53. Haiman, C.A., Chen, G.K., Blot, W.J., Strom, S.S., Berndt, S.I., Kittles, R.A., Rybicki, B.A., Isaacs, W.B., Ingles, S.A., Stanford, J.L. *et al.* (2011) Characterizing genetic risk at known prostate cancer susceptibility loci in African Americans. *PLoS Genetics*, **7**, e1001387.
54. Rohland, N. and Reich, D. (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res.*, **22**, 939–946.
55. Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
56. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491–498.
57. Wang, K., Li, M. and Hakonarson, H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucl.* Acids Res., **38**, e164.
58. Mancuso, N., Rohland, N., Rand, K.A., Tandon, A., Allen, A., Quinque, D., Mallick, S., Li, H., Stram, A., Sheng, X. *et al.* (2015) The contribution of rare variation to prostate cancer heritability. *Nat. Genet.*, in press.
59. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.
60. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.