

Identification of multiple novel prostate cancer susceptibility loci by a genome-wide association study

Rosalind A. Eeles^{1,2}, Zsófia Kote-Jarai^{*1}, Graham G. Giles^{*3,4}, Ali Amin Al Olama^{*5}, Michelle Guy^{*1}, Sarah K. Jugurnauth⁺¹, Shani Mulholland⁺¹, Daniel A. Leongamornlert⁺¹, Stephen M. Edwards⁺¹, Jonathan Morrison⁺⁵, Helen I. Field⁺⁶, Melissa C. Southey⁺⁷, Gianluca Severi^{+3,4}, Jenny L. Donovan⁺⁸, Freddie C. Hamdy⁺⁹, David P. Dearnaley^{+1,2}, Kenneth R. Muir⁺¹⁰, Charmaine Smith⁺³, Melisa Bagnato⁺³, Audrey T. Ardern-Jones², Amanda L. Hall^{1,2}, Lynne T. O'Brien¹, Beatrice N. Gehr-Swain^{1,2}, Rosemary A. Wilkinson¹, Angie Cox⁹, Sarah Lewis⁸, Paul M. Brown¹¹, Sameer G. Jhavar¹, Malgorzata Tymrakiewicz¹, Artitaya Lophatananon¹⁰, Sarah L. Bryant¹, The UK Genetic Prostate Cancer Study Collaborators/British Association of Urological Surgeons' Section of Oncology[^], The UK ProtecT Study Collaborators^{^^}, Alan Horwich^{1,2}, Robert A. Huddart^{1,2}, Vincent S. Khoo^{2,1}, Christopher C. Parker^{1,2}, Christopher J. Woodhouse², Alan Thompson², Tim Christmas², Chris Ogden², Cyril Fisher², Charles Jamieson², Colin S. Cooper¹, Dallas R. English⁴, John L. Hopper⁴, David E. Neal^{++11,12}, Douglas F. Easton⁺⁺⁵

¹The Institute of Cancer Research, 15 Cotswold Road, Sutton, Surrey, SM2 5NG, UK

²The Royal Marsden NHS Foundation Trust, Downs Road, Sutton, Surrey, SM2 5PT, and Fulham Road, London SW3 6JJ, UK

³Cancer Epidemiology Centre, The Cancer Council Victoria, 1 Rathdowne street, Carlton VIC 3053, Australia

⁴Centre for Molecular, Environmental, Genetic and Analytic Epidemiology, The University of Melbourne, 723 Swanston street, Carlton VIC 3053, Australia

⁵CR-UK Genetic Epidemiology Unit, University of Cambridge, Strangeways Laboratory, Worts Causeway, Cambridge, CB1 8RN, UK

⁶Department of Oncology, University of Cambridge, Strangeways Laboratory, Worts Causeway, Cambridge, CB1 8RN, UK

⁷Genetic Epidemiology Laboratory, Department of Pathology, The University of Melbourne, Grattan street, Parkville VIC 3052, Australia

⁸Department of Social Medicine, University of Bristol, Canynge Hall, Whiteladies Road, Bristol BS8 2PR, UK

⁹Academic Urology Unit, University of Sheffield, S10 2JF, UK

¹⁰University of Nottingham Medical School, Queens Medical Centre, Nottingham, NG7 2UH, UK

¹¹Surgical Oncology (Uro-Oncology: S4), Departments of Oncology and Surgery, University of Cambridge, Box 279, Addenbrooke's Hospital, Hills Road, Cambridge, CB2 2QQ, UK

¹²Cancer Research UK Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge, CB2 0RE, UK

Corresponding author: R. Eeles

*equal authors at this position

+equal authors at this position

++ equal authors at this position

[^] UK Genetic Prostate Cancer Study Collaborators/British Association of Urological Surgeons' Section of Oncology - see appendix 1

^{^^}UK ProtecT Study Collaborators: Prasad Bollina, Sue Bonnington, Debbie Cooper, Michael Davis, Andrew Doble, Alan Doherty, Garrett Durkan, Emma Elliott, David

Gillatt, Pippa Herbert, Peter Holding, Joanne Howson, Mandy Jones, Roger Kockelbergh, Howard Kynaston, Athene Lane, Teresa Lennon, Norma Lyons, Hing Leung, Hilary Moody, Philip Powell, Stephen Prescott, Pauline Thompson, care of Surgical Oncology (Uro-Oncology: S4), University of Cambridge, Box 279, Addenbrooke's Hospital, Hills Road, Cambridge CB2 2QQ

Prostate cancer (PrCa) is the commonest male cancer in developed countries. It exhibits consistent evidence of familial aggregation, but the causes of this aggregation are mostly unknown. To identify common PrCa susceptibility alleles, we conducted a genome-wide association study (GWAS) using blood DNA samples from 1,854 PrCa cases with clinically detected PrCa diagnosed at ≤ 60 years or with a family history of disease, and 1,894 population screened controls with a low prostate specific antigen (PSA) of $< 0.5 \text{ ng/ml}$. These were analysed for 541,129 SNPs using the Illumina Infinium platform. Initial putative associations were confirmed using a further 3,268 cases and 3,366 controls. We identified seven novel PrCa susceptibility loci on chromosomes 3, 6, 7, 10, 11, 19 and X ($p=2.7 \times 10^{-8}$ to $p=8.7 \times 10^{-29}$). We confirmed previous reports of common PrCa susceptibility loci at 8q24 and 17q. Three of the novel loci contain candidate susceptibility genes: *MSMB*, *LMTK2*, and *KLK3*.

There is strong evidence from family studies of genetic predisposition to PrCa: the relative risk of PrCa is increased about two-fold in first degree relatives of affected men, particularly when diagnosed at younger ages¹. Despite this, few susceptibility genes for PrCa have been identified. Linkage studies based on multiple case families have not identified reproducible susceptibility loci, suggesting that predisposition may be mediated through multiple common low penetrance alleles. Recent association studies have identified common alleles on 8q24 and on 17q associated with PrCa risk²⁻⁷. In addition, rare mutations in candidate genes (notably *BRCA2*), are associated with PrCa risk⁸. However, these mutations explain less than 10% of the familial relative risk of PrCa⁸.

Genome-wide association studies (GWASs) have emerged as a powerful new approach to identify common disease alleles without prior knowledge of position or function^{9,10}. Genotype frequencies are compared between cases and controls at large numbers of single nucleotide polymorphisms (SNPs), chosen to report on most known common variants in the genome. We conducted a two-stage GWAS to identify common PrCa susceptibility alleles. In the first stage, we studied 1,906 PrCa cases and 1,934 controls collected through national studies in the UK; the final number analysed after exclusions (see methods) was 1854 cases and 1894 controls (table 1). The cases were detected through clinical symptoms, rather than routine screening by PSA, because these cases have known clinical relevance. We further “enriched” the case series by including men diagnosed aged ≤ 60 years or with a family history of PrCa, since such cases are thought to be more likely to carry susceptibility alleles, thereby increasing statistical power. Controls were men aged ≥ 50 years identified through a national case-finding study (ProtecT), who had a baseline PSA of $< 0.5 \text{ ng/ml}$. Men with a low PSA level are known to be at a low risk for the subsequent development of clinically significant PrCa¹¹, potentially further improving power.

DNA samples from these individuals were evaluated for a set of 569,243 SNPs using the Illumina Infinium platform (see methods). This SNP set has been estimated to report on approximately 90% of the SNPs typed in HapMap, based on data on the Caucasian (CEU) samples, at an LD coefficient (r^2) > 0.80 . In this analysis, we utilised data on 541,129 SNPs that were genotyped on all subjects, passed quality control

(QC) criteria, and had a minor allele frequency of at least 1% in controls (see methods).

Figure 1 shows the Q-Q plot for the distribution of test statistics for comparison of genotype frequencies in cases versus controls (1 degree of freedom (df) Cochran-Armitage trend test). There was little evidence of any general inflation of the test statistics (estimated inflation factor $\lambda=1.05$ based on the bottom 90% of the distribution). This was as expected, since cases and controls were of European origin, and were broadly matched for region of residence. This is consistent with previous observations that population structure across the UK causes little inflation of the test statistics in association studies⁹. There was, however, a marked excess of “significant” associations. A total of 197 SNPs were significant at the $p<10^{-4}$ level, compared with the 54.1 that would have been expected by chance; of these, 53 were significant at the $p<10^{-6}$ level, compared with 0.5 expected by chance (supplementary table 1).

Of the 53 SNPs significant at the $p<10^{-6}$ level, 20 were on 8q24, a region previously shown to harbour PrCa susceptibility loci (supplementary tables 1 and 2). These occurred in three distinct LD blocks. The strongest associations in these regions were found with rs6983267 (per allele odds ratio (OR) 1.42, $p=9\times 10^{-14}$), rs1016343 (per allele OR 1.37, $p=1.6\times 10^{-8}$) and rs4242384 (per allele OR 1.88, $p=2.8\times 10^{-17}$).

Six SNPs on chromosome 17 reached $p<10^{-6}$. Four of these were located at 17q12, the strongest association being with rs7501939 in *TCF2* (per allele OR 0.71, $p=10^{-12}$). The other two SNPs were located at 17q24, the strongest association being with rs1859962 (per allele OR 1.26, $p=5.5\times 10^{-7}$). These results provide strong confirmation of previous observations by Gudmundsson et al⁷.

The remaining 27 SNPs were from eight genomic regions (supplementary tables 1 and 2). We conducted multiple logistic regression analyses, based on the SNPs in each of the regions, and identified eleven that appeared to be independently significant (supplementary table 3). To confirm these associations, SNPs were evaluated in a second stage of 3,268 PrCa cases and 3,366 controls from studies in the UK and Australia. For 8 of these 11 SNPs, from seven regions, there was confirmatory evidence (at least $p<.002$ and in the same direction as in stage 1) with a combined significance level over both stages of $p=2.7\times 10^{-8}$ or better. This provides strong evidence of association at a level of significance appropriate for a GWAS (table 2)^{9,12}. One SNP on chromosome 12 (rs902774) showed a strong association with disease in stage 1 ($p=2\times 10^{-7}$) but this was not replicated in stage 2, suggesting that the significance of the initial association may be a type I error. Of the three SNPs typed on chromosome 19, rs2735839, which showed the strongest association in stage 1 ($p=2.4\times 10^{-20}$) also gave evidence of association in stage 2 ($p=.0002$; $p=1.5\times 10^{-18}$ overall), but the other two SNPs did not replicate. Of the two SNPs tested on chromosome 10, both showed strong evidence of association. However, the association with rs10993994 was far stronger in both stages, and the association with rs7920517 was not significant after adjustment for rs10993994 in stage 2.

It is notable that, for each confirmed SNP, the estimated per allele OR is stronger for stage 1 than stage 2. This may reflect the fact that we have restricted attention to

highly significant loci (“winner’s curse”), but may also reflect, in part, the enriched nature of the cases and controls in stage 1. For these reasons we regard only the OR estimates from stage 2 as valid estimates of the relative risks in the general population.

We investigated the associations of the SNPs with PSA in a sample of 1646 UK controls in stage 2. Four of the SNPs, rs10993994, rs7920517, rs2659056 and rs2735839, were strongly associated with PSA level in the same direction as the association with PrCa risk (supplementary table 4). Of the two chromosome 10 SNPs, the association with rs7920517 was not significant after adjustment for rs10993994, consistent with the pattern for PrCa risk. A weaker association with PSA levels was also observed for rs5945619, again in the same direction as PrCa risk. It is notable that rs10993994 and rs2735839 showed a marked difference in allele frequency for controls between stages 1 and 2. These results indicate that the stronger associations seen in stage 1 for the chromosome 10 and 19 loci may partly reflect the selection of low PSA controls. However, the persistence of the associations in stage 2 (which selected controls from four population-based studies, all unselected for PSA levels) indicate that the association with disease is not solely due to selection on PSA levels.

There were no marked differences in allele frequencies between the controls from UK and Australia, or between either of the control groups used in each country, for any SNP (table 2 and supplementary table 5). We also note the control frequencies for all SNPs in stage 2 are very close to those for the 1958 Birth Cohort, a UK based cohort study used in other GWAS (<http://www.b58cgene.sgu.ac.uk>). No evidence of regional variation in genotype frequencies was observed in our study or the 1958 Birth Cohort. These observations provide reassurance that the control frequencies are robust and not subject to significant regional variation.

In stage 1, two of the SNPs (rs10993994 and rs7931342) showed stronger associations for familial cases ($p=.04$ and $p=.0002$ respectively). In stage 2, there was some evidence of a higher relative risk for cases diagnosed before age 60 years for rs266849 (supplementary table 6, $p=.01$). Larger case-control studies will, however, be required to estimate more precisely the age-specific relative risks associated with these SNPs.

We are able to compare our results with the publicly available results from the Cancer Genetic Markers of Susceptibility (CGEMS) study, a genome scan of 1,117 screen detected PC cases and 1,105 controls that used the same platform (<http://cgems.cancer.gov/>). There was some evidence of association for each of the SNPs on chromosomes 10 (rs10993994; $p=.009$), 11 (rs7931342; $p=.015$), 19 (rs2735839; $p=.004$) and X (rs5945619; $p=.0004$), in addition to the *TCF2* association (rs7501939; $p=.002$), but no evidence for the novel associations on chromosomes 3, 6 or 7. These differences may reflect the greater size of our study and the selection for clinically detected disease with genetic enrichment due to early age at onset and family history.

For four SNPs, the genotype-specific ORs showed some evidence of departure from a multiplicative (allele dosage) model. For rs2660753, rs6465657 and rs109933994 the rare homozygote ORs were greater than would be expected under this model ($p=.02$, $p=.05$ & $p=.05$ respectively), whilst for rs93364554, there was no apparent difference

in risk between rare homozygotes and heterozygotes (table 2 and supplementary table 7).

The seven novel susceptibility regions contain several strong plausible candidate genes (see figure 2). rs10993994 and rs7920517 lie within a ~100kb LD block on chromosome 10, containing the microseminoprotein beta gene, *MSMB*. The most strongly associated SNP, rs10993994, is 2bp upstream of the transcription start site of *MSMB*. Its location and the strength of the association raises the possibility that this SNP may be causally related to disease risk, but resequencing and further analyses will be needed clarify the functional basis of this association. Of note, the risk allele removes multiple predicted binding sites for transcription and splicing factors (<http://www.genomatix.de>). Putative androgen and estrogen receptor binding sites lie less than 50bp upstream of this SNP. *MSMB* encodes for PSP94, a member of the immunoglobulin binding factor family synthesized by epithelial cells of the prostate and secreted into seminal plasma. Loss of expression of PSP94 is associated with recurrence after radical prostatectomy¹³.

rs2735839 lies between *KLK2* (hK2) and *KLK3* (PSA). *KLK2* and *KLK3* kallikreins are a subgroup of serine proteases located in a cluster on chromosome 19. PSA is a serine protease which liquefies semen and as a serum marker is used in screening and disease monitoring; there is also evidence that hK2 may also be useful for screening and prognosis^{14,15}. Multiple SNPs in the promoter region have been associated with PSA levels^{16,17} and some have been suggested to be associated with PrCa risk^{17,18}. rs27358389 lies 3' of *KLK3* and shows a much stronger association with PSA levels than those previously reported, suggesting a novel functional effect.

rs6465657 is in intron 9 of *LMTK2* (*cprk*; *BREK*; Brain-Enriched Kinase)¹⁹ which encodes a neuronal kinase, cyclin-dependent kinase 5 (cdk5)/p35-regulated kinase (*cprk*). Somatic mutations in *LMTK2* have been found in a small proportion of cancers at other sites²⁰.

rs9364554 is in intron 5 of *SLC22A3*, one of the solute carrier family 22 (organic cation transporter; OCT) genes. OCTs are critical for elimination of some drugs and environmental toxins. rs5945619 is in a ~2MB LD block on Xp between *NUDT10* and *NUDT11* [nudix (nucleoside diphosphate linked moiety X)-type motif 11], about 2kb upstream of the latter. These genes encode isoforms of diphosphoinositol polyphosphate phosphohydrolase which determine the rate of phosphorylation in DNA repair, stress responses and apoptosis²¹. rs2660753 is in a gene-poor region on chromosome 3 and rs7931342 lies in an LD block of 70kb on chromosome 11 which is a gene desert.

These results provide strong evidence for seven new PrCa susceptibility loci. In addition, they provide strong confirmation for three loci on 8q24 and two on 17q. The fact that the previously reported PrCa susceptibility loci could be confirmed at genome-wide levels of significance after stage 1 reflects the size of this study and perhaps also the enrichment of cases due to early age at diagnosis or PrCa family history. Based on the effect size seen in stage 2 (that is, ignoring the effect of enrichment of the stage 1 set), the current study had approximately 52% power to detect the *MSMB* association, rising to close to 100% power based on the effect size seen stage 1. The results therefore suggest that few, if any, further common loci will

be detected with effect sizes that are stronger than *MSMB* or the strongest 8q loci, at least using the current platforms that provide comprehensive coverage based on HapMap. We have, however, only followed up a small number of the most significant associations in stage 1. It is likely that many other loci will be detectable by further follow-up of this and other scans, together with combined analyses of multiple scans.

Based on the estimated ORs in stage 2 of our study, the novel loci reported here would together explain approximately 6% of the familial risk of PrCa, with *MSMB* being the most significant (~2% of the familial risk, comparable to the two strongest 8q loci). Together with the previously reported loci, approximately 15% of familial risk in PrCa can now be explained.

The results of this study confirm that PrCa is genetically complex, and help clarify the genetic architecture of PrCa. The loci include plausible candidates, including a kinase gene, loci without obvious candidates and one gene desert, suggesting that diverse pathways are likely to be involved. Resequencing of these regions, further genotyping and functional analyses will be required to identify the causal SNPs.

These results may have a variety of clinical implications. The involvement of *MSMB* highlights a potential role for its product in PrCa screening, whilst *LMTK2* might provide a potential therapeutic target. There are also potential implications for risk counselling. As expected, the relative risks conferred by these loci are modest: the homozygote OR for rs10993994 at *MSMB* was 1.61 fold (95%CI 1.40-1.86). rs2660753 had the highest homozygote OR (2.09), but with a wide confidence interval. It is possible, however, that the associations we have found using tag SNPs may reflect stronger associations with the causal variants. Furthermore, the combined effect of these SNPs may be substantial, and as other SNPs are identified it may be possible to define genotypes that are sufficiently predictive of risk to be useful clinically.

Methods

Samples

PrCa cases for stage 1 were selected from the UK Genetic Prostate Cancer Study (UKGPCS)²². Cases were selected on the basis of either a diagnosis at age ≤ 60 years (n=1291) or a first or second degree family history of prostate cancer (n=726). We excluded men who reported to be non-white and men who were known to be diagnosed through asymptomatic PSA screening.

Controls for stage 1 were selected through the ProtecT study. ProtecT is a national study of community-based PSA testing and a randomised trial of subsequent prostate cancer treatment. Approximately 200,000 men between the ages of 50 and 69 years, ascertained through general practices in nine regions in the UK, are being recruited. For stage 1, we selected men aged ≥ 50 years with a PSA of < 0.5 ng/ml. Men known to be non-white were excluded. We then selected 2,001 controls to be frequency matched to the geographical distribution of the cases.

Stage 2 comprised PrCa cases and controls from the UK and Australia. The former were ascertained through the UK GPCS as above (n=332) and through a

systematically collected series from PrCa clinics in the Urology Unit at The Royal Marsden NHS Foundation Trust (n= 1680) over a 14 year period. UK controls were identified through two sources. Four hundred and forty nine controls were drawn from the UK GPCS study (Prostate Cancer Research Foundation Study component) and were geographically, ethnically and age matched to the UKGPCS young onset cases. They had no family or personal history of PrCa. The remaining controls (n=1712) were selected from men in the ProtecT study who had a PSA of <10ng/ml. Men with PSA >4ng/ml were excluded if they had a positive prostatic biopsy. As for stage 1, we excluded men known to be non-white.

The Australian cases were ascertained from three studies: (i) a population-based series of PrCa cases identified from the Victorian Cancer Registry since 1999, diagnosed at <56 years (Early Onset Prostate Cancer Study, EOPCFS; n=526); (ii) a population-based case control study based on cases diagnosed in Melbourne and Perth (Risk Factors for Prostate Cancer Study, RFPCS; n=594). Cases were identified from the population cancer registries, with histopathologically confirmed prostate cancer, excluding tumors with Gleason scores of less than 5, diagnosed at < 70 years with sampling stratified by age at diagnosis^{23,24,25}; (iii) a prospective cohort study of 17,154 men aged 40 to 69 years at recruitment in 1990-4 (Melbourne Collaborative Cohort Study, MCCS; n=190)^{26,27}. Controls were selected from the RFPCS study, in which they were identified through government electoral rolls and frequency matched to the age distribution of the RFPCS cases (n=509), together with a random sample from the MCCS cohort (n=760).

All studies were approved by the appropriate ethics committees.

Genotyping

Stage 1 genotypes were generated using the Illumina Infinium HumanHap550 array. We utilised only samples which called on at least 97% of SNPs at a confidence score of ≥ 0.25 . Owing to a re-synthesis of the beadset between the stages, the marker sets (versions 1 and 3) are slightly different: 534,446 SNPs were common to both sets, 14,356 markers were unique to version 1 and 20,441 markers were unique to version 3. We utilised data on 3840 individuals (1906 cases, 1934 controls): 323 typed on version 1 and 3525 on version 3 (including 8 duplicates typed on both versions). All the SNPs re-evaluated in stage 2 were from the common set of SNPs, and for simplicity the QQ plot and summary results utilise this SNP set.

SNPs were selected for evaluation in stage 2 on the basis of a significance level of $p < 10^{-6}$ based on a 1df trend test. We excluded SNPs from the previously reported regions of association on 8q24 and 17q. We then conducted multiple logistic regression using the set of SNPs in each of the remaining regions, to define SNPs that showed evidence of independent association at $p < .05$.

Genotyping in stage 2 was performed by 5' nuclease assay (Taqman™) using the ABI Prism 7900HT sequence detection system according to the manufacturer's instructions. Primers and probes were supplied directly by Applied Biosystems (<http://www.appliedbiosystems.com/>) as Assays-By-Design™. All assays were carried out in 384-well format. Each plate included at least 2 negative controls and 2 duplicates.

Statistical Methods

To identify close relatives we computed identity-by-state (IBS) probabilities for all pairs. We identified 27 cryptic duplicate samples (or MZ twins) and 3 pairs of probable brothers (IBS >0.86). In each case we excluded the individual with the lower call rate. By computing IBS scores between participants and individuals in HapMap and using multi-dimensional scaling, we identified 59 individuals who appeared to have significant Asian or African ancestry (approximately 10% non-European ancestry). We removed five cases of apparent Klinefelter's syndrome. After these exclusions, 1854 cases and 1894 controls were used in the final analysis of stage 1.

We filtered out all SNPs with a call rate <95%, a minor allele frequency in controls of <1%, or whose genotype frequency in controls departed from Hardy-Weinberg equilibrium at $p < .00001$. After these exclusions, we analyzed 541,129 SNPs, of which 509,295 were in both versions and common to all samples. Duplicate concordance was 98.8%.

In stage 2, we excluded 123 samples that failed on four or more of the assays used. The call rates were at least 0.97 for each SNP in each population. Genotype distributions in each control population for each SNP were consistent with Hardy-Weinberg equilibrium.

We assessed associations between each SNP and disease at stage 1 using a 1df Cochran-Armitage trend test and a general 2df chi-squared test. Inflation in the chi-squared statistic was assessed using the genomic control approach: we derived an inflation factor (λ) by dividing the median of the lowest 90% of the 1df statistics by the 45% percentile of a 1df chi-squared distribution (0.357). This cutoff was used to avoid inclusion of SNPs likely to be associated with risk. We chose to present p-values uncorrected for λ since the estimated λ (1.05) was very close to 1, making little difference to the significance levels, and to preserve consistency with the stage 2 analysis.

After stage 2, stratified 1df and 2df tests were performed, stratifying by stage and country. Odds ratios and confidence limits were estimated from the stage 2 data using unconditional logistic regression, stratified by country. Tests of homogeneity of the odds ratios across strata were assessed using likelihood ratio tests. Geographical variation in allele frequencies within the UK was assessed by classifying individuals into nine regions. Modification of the odds ratios by age was assessed using a case-only analysis, assessing the effect of age on SNP genotype in the cases using polytomous regression. The effects of SNP genotypes on PSA level were assessed using linear regression, after log-transformation of PSA level to correct for skewness. Analyses were performed in R (principally using SNPMatrix²⁸) and Stata. Binding site predictions were examined using MatInspector from Genomatix (<http://www.genomatix.de/>).

Acknowledgements

We should like to thank all the patients and control men who took part in this study. This work was supported by Cancer Research UK Grant C5047/A3354. DFE is a Principal Research Fellow of Cancer Research UK. John Hopper is an Australia Fellow of the NHMRC. We would also like to thank the following for funding support: The Institute of Cancer Research and The Everyman Campaign, The Prostate Cancer Research Foundation, Prostate Research Campaign UK, The National Cancer Research Network UK, The National Cancer Research Institute (NCRI) UK, grants from the National Health and Medical Research Council, Australia (209057, 251533, 450104), VicHealth, The Cancer Council Victoria, The Whitten Foundation, and Tattersall's. We are grateful to the staff at Illumina: Karine Viaud, Celeste McBride, Josh Bernd, Robert Morey, and the rest of the Illumina Genotyping Service Laboratory, Sandy McBean, Karen Cook, Fahim Amini, Marc Laurent, Mark Gibbs and those at Tepnel Life Sciences: Sheila Doyle, Amanda Priest, Alison Barlow, Sarah Howe and Louise Holliday, for their help with this study. We would like to acknowledge the help of Nathan Gauge, Charlotte Bamber, Sandra Barrett and Penelope Kelham for sample collection and retrieval and Pat Hamilton for assistance with preparing the manuscript. The ProtecT study is ongoing and is funded by the Health Technology Assessment Programme (projects 96/20/06, 96/20/99). The authors would like to acknowledge the tremendous contribution of all members of the ProtecT study research group, especially those listed. The ProtecT trial and its linked ProMPT and CAP (Comparison Arm for ProtecT) studies are supported by Department of Health, England; Cancer Research UK grant number C522/A8649, Medical Research Council of England grant number G0500966, ID 75466 and The NCRI, UK. The epidemiological data for ProtecT were generated through funding from the Southwest National Health Service Research and Development. The views and opinions expressed therein are those of the authors and do not necessarily reflect those of the Department of Health of England.

List of UKGPCS collaborators

See appendix 1.

Competing Financial Interests

None

Author contributions

R.A.E. and D.F.E. designed the study and obtained financial support. Z.K-J. directed the genotyping of stage 2. D.F.E. and A.A. conducted the statistical analysis. J.M. and H.F. provided bioinformatics support. M.G. co-ordinated the UKGPCS. G.G.G., J.L.H. and D.R.E. directed the Australian studies. M.C.S. and G.S. co-ordinated the Australian studies. D.E.N., J.L.D. and F.C.H. directed the ProtecT study. The remaining authors collated samples and performed laboratory analyses. R.A.E. and D.F.E. drafted the manuscript.

References

References

1. Edwards, S. M. & Eeles, R. A. Unravelling the genetics of prostate cancer. *American Journal of Medical Genetics Part C-Seminars in Medical Genetics* **129C**, 65-73 (2004).
2. Amundadottir, L. T. *et al.* A common variant associated with prostate cancer in European and African populations. *Nature Genetics* **38**, 652-658 (2006).
3. Gudmundsson, J. *et al.* Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nature Genetics* **39**, 631-637 (2007).
4. Yeager, M. *et al.* Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nature Genetics* **39**, 645-649 (2007).
5. Freedman, M. L. *et al.* Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 14068-14073 (2006).
6. Haiman, C. A. *et al.* Multiple regions within 8q24 independently affect risk for prostate cancer. *Nature Genetics* **39**, 638-644 (2007).
7. Gudmundsson, J. *et al.* Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. *Nature Genetics* **39**, 977-983 (2007).
8. Edwards, S. M. *et al.* Two percent of men with early-onset prostate cancer harbor germline mutations in the BRCA2 gene. *American Journal of Human Genetics* **72**, 1-12 (2003).
9. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-678 (2007).
10. Easton DF *et al.* Genome-wide association study identifies breast cancer susceptibility loci. *Nature* **447**, 1087-1093 (2007).
11. Lilja, H. *et al.* Long-term prediction of prostate cancer up to 25 years before diagnosis of prostate cancer using prostate kallikreins measured at age 44 to 50 years. *Journal of Clinical Oncology* **25**, 431-436 (2007).
12. Thomas, D. C., Haile, R. W., & Duggan, D. Recent developments in genomewide association scans: A workshop summary and review. *American Journal of Human Genetics* **77**, 337-345 (2005).
13. Reeves, J. R., Dulude, H., Panchal, C., Daigneault, L., & Ramnani, D. M. Prognostic value of prostate secretory protein of 94 amino acids and its binding protein after radical prostatectomy. *Clinical Cancer Research* **12**, 6018-6022 (2006).
14. Steuber, T., Helo, P., & Lilja, H. Circulating biomarkers for prostate cancer. *World Journal of Urology* **25**, 111-119 (2007).
15. Steuber, T. *et al.* Risk assessment for biochemical recurrence prior to radical prostatectomy: significant enhancement contributed by human glandular kallikrein 2 (hk2) and free prostate specific antigen (PSA) in men with moderate PSA-elevation in serum. *International Journal of Cancer* **118**, 1234-1240 (2006).

16. Cramer, S. D. *et al.* Association between genetic polymorphisms in the prostate-specific antigen gene promoter and serum prostate-specific antigen levels. *Journal of the National Cancer Institute* **95**, 1044-1053 (2003).
17. Lai, J. *et al.* PSA/KLK3 ARE1 promoter polymorphism alters androgen receptor binding and is associated with prostate cancer susceptibility. *Carcinogenesis* **28**, 1032-1039 (2007).
18. Severi, G. *et al.* Variants in the prostate-specific antigen (PSA) gene and prostate cancer risk, survival, and circulating PSA. *Cancer Epidemiology Biomarkers & Prevention* **15**, 1142-1147 (2006).
19. Kawa, S., Fujimoto, J., Tezuka, T., Nakazawa, T., & Yamamoto, T. Involvement of BREK, a serine/threonine kinase enriched in brain, in NGF signalling. *Genes to Cells* **9**, 219-232 (2004).
20. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153-158 (2007).
21. Hidaka, K. *et al.* An adjacent pair of human NUDT genes on chromosome X are preferentially expressed in testis and encode two new isoforms of diphosphoinositol polyphosphate phosphohydrolase. *Journal of Biological Chemistry* **277**, 32730-32738 (2002).
22. Eeles, R. A. Genetic predisposition to prostate cancer. *Prostate Cancer and Prostatic Diseases* **2**, 9-15 (1999).
23. Giles, G. G. *et al.* Smoking and prostate cancer: Findings from an Australian case-control study. *Annals of Oncology* **12**, 761-765 (2001).
24. Giles, G. G. *et al.* Androgenetic alopecia and prostate cancer: Findings from an Australian case-control study. *Cancer Epidemiology Biomarkers & Prevention* **11**, 549-553 (2002).
25. Severi, G. *et al.* ELAC2/HPC2 polymorphisms, prostate-specific antigen levels, and prostate cancer. *Journal of the National Cancer Institute* **95**, 818-824 (2003).
26. MacInnis, R. J., English, D. R., Gertig, D. M., Hopper, J. L., & Giles, G. G. Body size and composition and prostate cancer risk. *Cancer Epidemiology Biomarkers & Prevention* **12**, 1417-1421 (2003).
27. Severi, G. *et al.* Circulating steroid hormones and the risk of prostate cancer. *Cancer Epidemiology Biomarkers & Prevention* **15**, 86-91 (2006).
28. Clayton, D. & Leung, H. T. An R package for analysis of whole-genome association studies. *Human Heredity* **64**, 45-51 (2007).

Table 1. Characteristics of the study sets used in the final analysis of genotypes.

Country	Study Population	Number of Prostate cancer cases	Number of controls	Mean age (range) in years
Stage 1				
UK	UK GPCS Total	1854	1894	
	Early Onset Cases	1171		54 (36-60)
	Familial Cases	683		62 (39-88)
	ProtecT		1894	61 (50-71)
Stage 2				
UK	Total	1960	2104	
	UK GPCS Early onset and familial	303	453	57 (36-77)
	RMH Series	1657		68 (24-89)
	ProtecT		1651	59 (50-70)
Australia	Total	1308	1262	
	EOPCFS	526	0	52 (38-55)
	MCCS	190	756	56 (38-80)
	RFPCS	592	506	61 (40-69)

Table 2. Summary results for 15 SNPs selected for genotyping in stage 2⁸.

Marker ¹	Position ²	Population	Maf ³		Per allele ⁴ OR (95%CI)	Het OR ^{4,5} (95%CI)	Hom OR ^{4,6} (95%CI)	P-value ⁷	
			Controls	Cases				Stage	Combined
rs2660753 C/T	3 87193364	Stage 1 UK	0.081	0.118	1.52 (1.30-1.77)	1.48 (1.25-1.76)	2.88 (1.38-5.99)	9.5x10 ⁻⁸	
		Stage 2 UK	0.105	0.124	1.20 (1.05-1.38)	1.10 (0.94-1.28)	2.50 (1.45-4.30)		
		Australia	0.118	0.133	1.15 (0.97-1.35)	1.11 (0.92-1.34)	1.62 (0.86-3.04)		
		All stage 2	0.110	0.128	1.18 (1.06-1.31)	1.10 (0.98-1.24)	2.09 (1.39-3.15)	.0018	2.7x10 ⁻⁸
rs9364554 C/T	6 160753654	Stage 1	0.285	0.338	1.28 (1.16-1.41)	1.17 (1.03-1.34)	1.82 (1.45-2.29)	9.3x10 ⁻⁷	
		Stage 2 UK	0.293	0.322	1.15 (1.04-1.26)	1.20 (1.05-1.37)	1.25 (1.00-1.56)		
		Australia	0.288	0.325	1.20 (1.06-1.35)	1.36 (1.15-1.60)	1.21 (0.92-1.60)		
		All stage 2	0.291	0.323	1.17 (1.08-1.26)	1.26 (1.14-1.39)	1.24 (1.04-1.47)	4.8x10 ⁻⁵	5.5x10 ⁻¹⁰
rs6465657 T/C	7 97654263	Stage 1	0.443	0.508	1.30 (1.19-1.43)	1.19 (1.02-1.38)	1.72 (1.43-2.07)	1.2x10 ⁻⁸	
		Stage 2 UK	0.467	0.497	1.13 (1.04-1.24)	1.07 (0.92-1.24)	1.29 (1.08-1.54)		
		Australia	0.458	0.484	1.11 (0.99-1.24)	0.96 (0.80-1.15)	1.26 (1.01-1.57)		
		All stage 2	0.463	0.492	1.12 (1.05-1.20)	1.03 (0.91-1.15)	1.27 (1.11-1.46)	.001	1.1x10 ⁻⁹
rs7920517 A/G	10 51202627	Stage 1	0.427	0.511	1.39 (1.27-1.53)	1.15 (0.99-1.34)	1.99 (1.66-2.39)	7.2x10 ⁻¹³	
		Stage 2 UK	0.476	0.525	1.21 (1.11-1.32)	1.18 (1.01-1.37)	1.47 (1.23-1.75)		
		Australia	0.474	0.528	1.23 (1.11-1.37)	1.10 (0.91-1.32)	1.52 (1.22-1.90)		
		All stage 2	0.476	0.526	1.22 (1.14-1.31)	1.15 (1.02-1.30)	1.49 (1.30-1.71)	9.3x10 ⁻⁹	5.4x10 ⁻¹⁹
rs10993994 C/T	10 51219502	Stage 1	0.344	0.459	1.62 (1.47-1.78)	1.42 (1.24-1.64)	2.80 (2.30-3.42)	8.0x10 ⁻²⁴	
		Stage 2 UK	0.403	0.451	1.21 (1.11-1.32)	1.12 (0.97-1.29)	1.51 (1.27-1.81)		
		Australia	0.402	0.468	1.31 (1.18-1.47)	1.20 (1.01-1.43)	1.79 (1.42-2.25)		
		All stage 2	0.402	0.458	1.25 (1.17-1.34)	1.15 (1.03-1.28)	1.61 (1.40-1.86)	1.5x10 ⁻¹⁰	8.7x10 ⁻²⁹

rs7931342 G/T	11 68751073	Stage 1	0.498	0.438	0.79 (0.72-0.86)	0.80 (0.68-0.92)	0.62 (0.52-0.74)	2.4x10 ⁻⁷	
		Stage 2 UK	0.476	0.438	0.86 (0.78-0.94)	0.90 (0.78-1.04)	0.73 (0.61-0.87)		
		Australia	0.503	0.452	0.82 (0.73-0.91)	0.74 (0.61-0.89)	0.67 (0.54-0.84)		
		All stage 2	0.486	0.444	0.84 (0.79-0.90)	0.84 (0.75-0.94)	0.71 (0.62-0.81)	1.1x10 ⁻⁶	1.7x10 ⁻¹²
rs902774 G/A	12 51560171	Stage 1	0.139	0.183	1.39 (1.23-1.57)	1.34 (1.16-1.55)	2.29 (1.50-3.50)	2.0x10 ⁻⁷	
		Stage 2 UK	0.154	0.155	1.00 (0.89-1.13)	0.96 (0.84-1.11)	1.16 (0.80-1.69)		
		Australia	0.132	0.144	1.10 (0.94-1.29)	1.10 (0.91-1.32)	1.25 (0.71-2.20)		
		All stage 2	0.146	0.150	1.03 (0.94-1.14)	1.01 (0.90-1.13)	1.19 (0.87-1.63)	.486	0.00015
rs2659056 A/G	19 56027755	Stage 1	0.213	0.265	1.33 (1.20-1.49)	1.27 (1.11-1.45)	2.01 (1.50-2.68)	1.2x10 ⁻⁷	
		Stage 2 UK	0.259	0.247	0.94 (0.85-1.04)	0.97 (0.85-1.10)	0.83 (0.64-1.07)		
		Australia	0.257	0.260	1.02 (0.89-1.15)	1.11 (0.94-1.30)	0.86 (0.62-1.20)		
		All stage 2	0.258	0.252	0.97 (0.89-1.05)	1.02 (0.92-1.13)	0.84 (0.68-1.03)	.424	.0125
rs266849 A/G	19 56040902	Stage 1	0.249	0.170	0.62 (0.55-0.69)	0.60 (0.52-0.69)	0.42 (0.30-0.58)	1.0x10 ⁻¹⁶	
		Stage 2 UK	0.190	0.187	0.98 (0.87-1.10)	1.03 (0.90-1.18)	0.81 (0.57-1.15)		
		Australia	0.205	0.187	0.90 (0.78-1.03)	0.81 (0.68-0.96)	1.08 (0.73-1.61)		
		All stage 2	0.196	0.187	0.95 (0.87-1.03)	0.94 (0.84-1.04)	0.92 (0.71-1.20)	.228	9.9x10 ⁻¹⁰
rs2735839 G/A	19 56056435	Stage 1	0.211	0.129	0.56 (0.50-0.64)	0.58 (0.50-0.67)	0.29 (0.19-0.43)	2.4x10 ⁻²⁰	
		Stage 2 UK	0.152	0.131	0.84 (0.74-0.95)	0.78 (0.68-0.91)	1.01 (0.65-1.57)		
		Australia	0.151	0.126	0.81 (0.69-0.95)	0.82 (0.68-0.98)	0.63 (0.34-1.15)		
		All stage 2	0.152	0.129	0.83 (0.75-0.91)	0.80 (0.71-0.89)	0.85 (0.60-1.22)	.0002	1.5x10 ⁻¹⁸
rs5945619 T/C	X 51258412	Stage 1	0.353	0.442	1.46 (1.28-1.66)	N/A	1.46 (1.28-1.66)	2.2x10 ⁻⁸	
		Stage 2 UK	0.356	0.391	1.16 (1.02-1.32)	N/A	1.16 (1.02-1.32)		
		Australia	0.361	0.410	1.23 (1.05-1.44)	N/A	1.23 (1.05-1.44)		

		All stage 2	0.358	0.398	1.19 (1.07-1.31)	N/A	1.19 (1.07-1.31)	.0008	1.5x10 ⁻⁹
--	--	-------------	-------	-------	---------------------	-----	---------------------	-------	----------------------

¹ dbSNP rs number and major/minor alleles, based on the frequencies in stage 1 controls.

² Chromosome and build 36 position.

³ Minor allele frequency.

⁴ OR = odds ratio.

⁵ OR in heterozygotes, relative to common homozygotes.

⁶ OR in homozygotes, relative to common homozygotes.

⁷ Cochran-Armitage test for trend.

⁸ Genotype counts are given in supplementary table 8.

Figure1. Quantile-quantile plot for the test statistics (Cochran-Armitage 1df chisquared trend tests) for stage 1. The continuous line gives the expected distribution assuming no inflation of the test statistics. Inflation was assessed using the lowest 90% of the test statistics (expected values less than 2.71).

Figure 2. Summary of results. Dx=age at diagnosis, chrom.=chromosome.