ARTICLE

Fine-mapping of prostate cancer susceptibility loci in a large meta-analysis identifies candidate causal variants

Tokhir Dadaev, Edward J. Saunders et al.#

Prostate cancer is a polygenic disease with a large heritable component. A number of common, low-penetrance prostate cancer risk loci have been identified through GWAS. Here we apply the Bayesian multivariate variable selection algorithm JAM to fine-map 84 prostate cancer susceptibility loci, using summary data from a large European ancestry meta-analysis. We observe evidence for multiple independent signals at 12 regions and 99 risk signals overall. Only 15 original GWAS tag SNPs remain among the catalogue of candidate variants identified; the remainder are replaced by more likely candidates. Biological annotation of our credible set of variants indicates significant enrichment within promoter and enhancer elements, and transcription factor-binding sites, including AR, ERG and FOXA1. In 40 regions at least one variant is colocalised with an eQTL in prostate cancer tissue. The refined set of candidate variants substantially increase the proportion of familial relative risk explained by these known susceptibility regions, which highlights the importance of fine-mapping studies and has implications for clinical risk profiling.
Prostate cancer (PrCa) is the most common cancer among males in developed countries. As there is evidence for a large heritable component for PrCa, the identification of genetic variation that increases susceptibility may help to inform screening strategies and clinical management of patients in the future. Currently, only a handful of rare genetic variants with larger effect sizes have been reported that increase the risk of PrCa (e.g., BRCA2 and ATM)\(^1\). By comparison, genome-wide association studies (GWAS) have reported \(>100\) low-penetrance PrCa risk signals with small odds ratios (ORs)\(^3\). Individually, these GWAS loci only modestly influence risk. However, because the risk alleles are relatively common within the general population their cumulative impact is substantial.

When an initial GWAS identifies a susceptibility locus, any one (or more) of a large number of variants within the region may underlie the molecular mechanism that modulates risk. This includes correlated variants in linkage disequilibrium (LD) that may capture the same association signal and additional variants with independent associations. Genotyping a denser set of variants in the region facilitates characterisation of the underlying genetic architecture and makes subsequent imputation more precise and complete. Although forward stepwise selection is frequently used for fine-mapping, it has severe limitations, particularly the way LD can lead to misleading results. In this manuscript, we report the findings of a PrCa fine-mapping study in a European ancestry meta-analysis sample set that is the largest to date and utilise the well-established stochastic search and model selection framework, which more accurately represents the uncertainty in determining both the number of signals and the set of single-nucleotide polymorphisms (SNPs) that best describe the association in each region\(^4\). To leverage the large sample size from the overall meta-analysis, we use a novel multivariate Bayesian variable selection approach, which takes marginal SNP summary statistics as input and accounts for LD, to jointly analyse all SNPs in a region. We identify a catalogue of variants and further prioritise within this set through functional annotation, to assist identification of putative causal variants. This refined credible set of variants explains a substantially larger proportion of the estimated familial relative risk (FRR) of PrCa compared with the original GWAS tags.

**Results**

**Replication of reported associations prior to fine-mapping.** In this study, we examined 92 PrCa GWAS risk associations within 85 distinct genomic regions reported prior to the recent meta-analysis using the OncoArray experiment\(^6\); due to their complexity, two regions (Chr8q24 and Chr6p21-MHC) were excluded and are subject to separate studies. Some regions contained more than one signal due to close proximity between the reported index SNPs. Summary results from the large European ancestry meta-analysis comprising 82,591 PrCa cases and 61,213 controls from eight GWAS subcohorts (OncoArray, iCOGS, UK stage 1 and 2, CaPS 1 and 2, BPC3 and NCI PEGASUS), imputed to the 1000 Genomes phase 3 reference panel, were used for our fine-mapping analysis.

We first assessed whether all 92 original associations had replicated with at least one variant in the region at a genome-wide significant level (marginal \(P\)-value \(<5 \times 10^{-8}\)). Five regions had not replicated and were excluded from downstream fine-mapping analyses accordingly (Supplementary Table 1). An additional 3 associations previously reported in different ancestral populations also had not replicated in our European sample set; however, these original lead variants were each situated within the region boundary of another replicated GWAS association and therefore the expanded region boundary was retained during fine-mapping for logistical purposes, although only the associations replicated in Europeans were considered as index variants. Fine-mapping was therefore conducted for 84 replicated, previously reported GWAS signals, within 80 distinct regions (Fig. 1). This included the region encompassing the moderate penetrance risk SNP rs138213197 in HOXB13, which although originally identified through sequencing\(^7\) was included due to its relatively close proximity to the GWAS association rs11650494. The HOXB13 region therefore also served as a useful positive control during mapping, since the known causal variant exerts a relatively large effect size (OR 3.85) and has low minor allele frequency (MAF), but the signal is also detectable through a cluster of more common variants as a ‘synthetic association’\(^8\).

The eight signals that did not replicate in our European meta-analysis may remain risk loci for PrCa in other ancestral populations or specific disease phenotypes rather than overall PrCa risk, although we cannot completely exclude the possibility that some were false positives. Two of these variants were originally reported in a multi-ethnic meta-analysis (rs7153648 and rs12051443), one failed quality control (QC) due to strongly discordant MAF between individual sub-studies within the meta-analysis (rs6625711) and is also reported as having extremely discordant MAF between 1000 Genomes phase 1 and phase 3 cohorts (MAF in EUR 0.45 vs. 0.16), one was associated with young-onset disease only (rs636291), one only for aggressive PrCa (rs1571801) and the final three were reported in populations of Chinese (rs103294), Japanese (rs2055109) or African (rs7210100) ancestry and had not been confirmed in Europeans to date\(^11\). Multivariate fine-mapping from univariate summary statistics.

When an initial GWAS identifies a susceptibility locus, any one (or more) of a large number of variants within the region may underlie the molecular mechanism that modulates risk. This includes correlated variants in linkage disequilibrium (LD) that may capture the same association signal and additional variants with independent associations. Genotyping a denser set of variants in the region facilitates characterisation of the underlying genetic architecture and makes subsequent imputation more precise and complete. Although forward stepwise selection is frequently used for fine-mapping, it has severe limitations, particularly the way LD can lead to misleading results. In this manuscript, we report the findings of a PrCa fine-mapping study in a European ancestry meta-analysis sample set that is the largest to date and utilise the well-established stochastic search and model selection framework, which more accurately represents the uncertainty in determining both the number of signals and the set of single-nucleotide polymorphisms (SNPs) that best describe the association in each region\(^4\). To leverage the large sample size from the overall meta-analysis, we use a novel multivariate Bayesian variable selection approach, which takes marginal SNP summary statistics as input and accounts for LD, to jointly analyse all SNPs in a region. We identify a catalogue of variants and further prioritise within this set through functional annotation, to assist identification of putative causal variants. This refined credible set of variants explains a substantially larger proportion of the estimated familial relative risk (FRR) of PrCa compared with the original GWAS tags.
GWAS results, including a novel more strongly associated lead variant in 4 of the 5 regions (Supplementary Data 1). The majority of variants within the JAM credible set were common (Supplementary Fig. 1a), with only 2 variants having MAF < 1% and 48 variants MAF < 5%; lower MAF variants do however represent the most likely candidate causal variants within certain regions. We also observed a slight increase in the distribution of univariate ORs for the novel lead variants we have identified in comparison to the original GWAS tag SNPs (Supplementary Fig. 1b). Only 15 original GWAS tag SNPs remained within the catalogue of candidate variants, with all other signals being replaced by more likely candidates. As expected, fine-mapping performance varied by region, with 95% credible set sizes ranging from 1 to 606 variants. We did however observe strong refinement of variants within the majority of regions (median 24 variants per region overall and 21 for single-signal regions). Indeed, among the 63 single-signal regions, 30 returned a 95% credible set containing ≤ 20 variants, of which 20 comprised ≤ 10 variants and 4 returned a credible set containing a single variant. These represent the putative causal PrCa susceptibility variant within that locus and include the well-established HOXB13 causal variant rs138213197 at Chr17q219, as well as rs10993994 in the promoter of MSMB, which modulates gene expression in prostate tissue17–19. These two regions serve as proof of principle; our

Fig. 1 Overview of the fine-mapping workflow. Flowchart describing the procedure followed during fine-mapping, providing an overview of the outcomes at each stage and suggesting possible applications for the final catalogue of variants.
methodology selected the presumed causal variants and therefore the remaining two single candidate variants are very likely to be causal and are strong candidates to test in functional studies. These two variants are an intronic SNP in TBX1, and a low MAF frameshift insertion in the final exon of FAM111A; which confirms for the first time in Europeans the GWAS hit at this locus previously reported in Japanese, although the European and Japanese variants are not in LD. The 12 regions with multiple independent risk signals contained 31 independent signals in total, represented by a 95% credible set of 626 variants (median 33.5 variants per region, average 20.2 variants per association). Prioritisation also performed well in these complex regions. In the TERT region at Chr3p15 we observed the highest number of independent signals, 5, and the credible set comprised only 30 SNPs. Similarly, 3 regions each containing 3 signals (Chr2q37:FARP2/ANO7, Chr17q12:HNF1B and Chr19q13:KLK3) returned a combined credible set of 61 variants representing these 9 PrCa associations. Notably, we observed that the regions found to contain multiple independent signals generally had P-values and marginal ORs towards the upper end of the distribution of original GWAS hits in the univariate meta-GWAS (Supplementary Fig. 2).

Integration of annotation. We annotated variants for indicators of putative biological functionality using data from publically available databases. Intragenic variants were ascribed to genes relative to GENCODEv19, miRNA variants using MirBasev20 and variants situated within segments of the genome under evolutionary conservation were annotated using conserved element outputs generated by four algorithms (GERP++, SiPhy Omega, SiPhy Pi and Phastcons)20–22. For information derived from tissue-based experimental data sets, we focused primarily on those conducted in prostate cell lines; specifically DNase hypersensitivity sites in three prostate cell types from seven experiments in the ENCODE project, chromatin-state characterisations by ChromHMM from Taberlay et al.23, ChIP-seq experiments in the ENCODE project, chromatin-state characterisations by ChromHMM from Taberlay et al.23, and a low MAF frameshift insertion in the final exon of FAM111A; which confirms for the first time in Europeans the GWAS hit at this locus previously reported in Japanese, although the European and Japanese variants are not in LD. The 12 regions with multiple independent risk signals contained 31 independent signals in total, represented by a 95% credible set of 626 variants (median 33.5 variants per region, average 20.2 variants per association). Prioritisation also performed well in these complex regions. In the TERT region at Chr3p15 we observed the highest number of independent signals, 5, and the credible set comprised only 30 SNPs. Similarly, 3 regions each containing 3 signals (Chr2q37:FARP2/ANO7, Chr17q12:HNF1B and Chr19q13:KLK3) returned a combined credible set of 61 variants representing these 9 PrCa associations. Notably, we observed that the regions found to contain multiple independent signals generally had P-values and marginal ORs towards the upper end of the distribution of original GWAS hits in the univariate meta-GWAS (Supplementary Fig. 2).
Table 2 Overview of fine-mapping results by region for regions 28–54 of the 80 regions fine-mapped

| Fine-mapping region boundary | Original index SNPs mapped | Pruning $r^2$ threshold | SNPs (tags) analysed | Number of signals | Credible set SNPs (tags) | Credible set eQTL SNPs (tags) | Credible set SNPs P < 0.05 in AAsa | Region contribution to overall FRR of PrCab |
|-----------------------------|---------------------------|-------------------------|---------------------|-------------------|------------------------|-------------------------------|--------------------------------------|----------------------------------------|
| chr6:116666036-117710052    | rs339331                   | 0.9                     | 2981 (433)          | 1                 | 102 (3)                | 0 (0)                         | 101                                  | 0.18 (0.16, 0.20)                      |
| chr6:152923566-153941079    | rs1933488                  | 0.9                     | 3636 (599)          | 1                 | 86 (6)                 | 45 (6)                        | 20                                  | 0.12 (0.10, 0.14)                      |
| chr6:160081543-161382029    | rs9364554                  | 0.9                     | 4101 (737)          | 3                 | 151 (15)               | 65 (10)                       | 7                                    | 1.03 (0.91, 1.19)                      |
| chr6:29573776-30573776      | rs7761788                  | 0.75                    | 7085 (464)          | 1                 | 606 (22)               | 372 (16)                      | 13                                  | 0.07 (0.03, 0.11)                      |
| chr6:41036427-42043793      | rs1983891                  | 0.9                     | 2840 (779)          | 3                 | 33 (2)                 | 9 (2)                         | 33                                  | 0.18 (0.16, 0.21)                      |
| chr7:5759882-76995982       | rs9443389                  | 0.6                     | 1966 (72)           | 2                 | 0 (0)                  | 0 (0)                         | 0                                   | 0.06 (0.05, 0.07)                      |
| chr7:20449491-21496953      | rs1215572                  | 0.9                     | 3170 (782)          | 1                 | 4 (1)                  | 0 (0)                         | 2                                   | 0.16 (0.15, 0.19)                      |
| chr7:27091215-30466563      | rs10486657                 | 0.9                     | 3372 (691)          | 1                 | 11 (2)                 | 1 (1)                         | 3                                    | 0.34 (0.31, 0.39)                      |
| chr7:46937244-47937244      | rs56223056                 | 0.9                     | 2803 (473)          | 1                 | 53 (6)                 | 0 (0)                         | 34                                  | 0.08 (0.04, 0.13)                      |
| chr7:97307882-98316327      | rs64665567                 | 0.9                     | 2892 (411)          | 1                 | 31 (1)                 | 11 (1)                        | 0                                   | 0.27 (0.24, 0.31)                      |
| chr8:22399287-24028511      | rs1512268                  | 0.9                     | 3507 (755)          | 2                 | 74 (3)                 | 1 (1)                         | 16                                  | 0.77 (0.68, 0.87)                      |

Published GWAS SNPs for which the signal or region replicated in our EUR meta-analysis are indicated, alongside the region co-ordinates assigned for fine-mapping analyses (GRCh37/hg19 assembly). The fine-priority pruner thresholds used and numbers of variants and priority pruner tags included in the analysis are shown. Summaries of the fine-mapping analysis results for each region contain the number of independent PrCa risk signals identified within each region, the size of the credible set of variants identified by JAM and the number of variants within the credible set that were also significantly associated eQTLs in TCGA PRAD data. As an additional category to assist variant prioritisation, the number of variants in the credible set that achieved a nominally significant P-value threshold ($P < 0.05$) in an unconnected African Ancestry GWAS is indicated. The estimated contribution of each GWAS region to the overall familial relative risk of PrCa after fine-mapping is also provided. These results are a continuation from the regions displayed in Table 1 and results for all remaining regions fine-mapped are provided in Table 3.

a | AAs African Ancestry population PrCa meta-analysis.

37 863 pruned tags against annotation features using a conditional quantile regression (QR) analysis. quantitative at multiple quantiles (99.2, 99.4, 99.6, 99.8 and 99.9%). These correspond to posterior probability ranges from 0.01 to 0.99, with the exact values conditional on the linear combination of the annotations. At each quantile, we used the fitted model to calculate a predicted posterior probability given the SNPs annotation features. A single expected posterior probability was then calculated from a weighted average of these quantile-specific expected posterior probabilities with the weight reflecting both the fit (i.e., a function of the likelihood) and variance of the predicted values from the quantile-specific model to the data. We selected a single data set for each annotation category for the QR analysis to minimise correlation between variables. Whilst the majority of tag probabilities were not notably adjusted during QR, an appreciable subset of variants were up- or downgraded based upon their annotations ($\Delta$Posterior probability$_{QR}$ ranged between $-0.304$ and $0.254$; 63 of the 37,863 tags had a $\Delta$Posterior probability$_{QR}$ of magnitude $\pm 0.005$ or greater) (Supplementary Fig. 3). The conditional QR also facilitates identification of the annotations that demonstrate an association across the extreme quantiles of the posterior probabilities. Specifically, several annotations (eQTLs within TCGA PrCa tissue, AR and GATA2 transcription factor-binding sites, LNCaP DNASel, H3K27Ac and H3K4Me3 histone marks, enhancer and repressed chromatin states by ChromHMM, conservation according to GERP++, higher CADD scores and protein altering variants) had statistically significant associations ($P < 1.0 \times 10^{-5}$) for at least one quantile (Supplementary Data 2). That is, the upper quantiles of the posterior probability distribution for variants with any of these annotations were larger when compared with SNPs without those annotations.

For comparison to the conditional QR approach, we also used Fisher’s exact test to examine the representation of individual annotation features across variants included in the 95% credible set of prospective PrCa causal variants relative to variants not selected. Independent tests were conducted for each annotation upon the set of 37,863 tag variants analysed by JAM, of which 343 tags represented the 95% credible set of 3700 SNPs and annotations for all proxy SNPs were inherited by the tag variant. We observed significant enrichment of a number of annotations among variants in the credible set (Fig. 2, Supplementary Data 2). In particular, enrichment was found for eQTLs in the TCGA data set ($P = 1.15 \times 10^{-23}$); intragenic variants within protein-coding genes ($P = 8.15 \times 10^{-11}$; $P = 6.03 \times 10^{-5}$ for protein altering variants exclusively) but not non-coding transcripts ($P = 0.29$); promoter ($P = 1.66 \times 10^{-5}$), enhancer ($P = 3.42 \times 10^{-6}$) and transcribed ($P = 3.07 \times 10^{-7}$) ChromHMM states in prostate epithelial cells; DNASel hypersensitivity sites from all seven ENCODE prostate data sets ($P = 1.28 \times 10^{-7}$ to $7.61 \times 10^{-17}$); for
Table 3 Overview of fine-mapping results by region for regions 55–80 of the 80 regions fine-mapped, and summary results across all 80 regions

| Fine-mapping region boundary | Original index SNPs mapped | Pruning \( r^2 \) threshold | SNPs (tags) analysed | Number of signals | Credible set SNPs (tags) | Credible set eQTL SNPs (tags) | Credible set SNPs \( P < 0.05 \) in AAsa | Region contribution to overall FRR of PrCa\(^b\) |
|-----------------------------|---------------------------|-----------------------------|---------------------|-------------------|-------------------------|-----------------------------|---------------------------------|-----------------------------------|
| chr12:5773904-53816821      | rs902774                  | 0.9                         | 3182 (553)           | 1                 | 28 (1)                  | 0 (0)                      | 10                              | 0.32 (0.28, 0.36)                 |
| chr13:73228139-74486916     | rs9600079                 | 0.9                         | 3995 (888)           | 1                 | 14 (5)                  | 0 (0)                      | 10                              | 0.13 (0.11, 0.14)                 |
| chr14:2872330-53898699      | rs8008270                 | 0.9                         | 2588 (410)           | 1                 | 12 (2)                  | 0 (0)                      | 0                               | 0.11 (0.10, 0.13)                 |
| chr14:68502988-69626744     | rs7141529                 | 0.9                         | 3015 (822)           | 1                 | 72 (17)                 | 1 (1)                      | 4                               | 0.07 (0.02, 0.12)                 |
| chr14:70592256-71592256     | rs8014671                 | 0.9                         | 2671 (139)           | 1                 | 0 (0)                   | 0 (0)                      | 0                               | 0.05 (0.02, 0.10)                 |
| chr17:118965-11991422       | rs684232                  | 0.9                         | 3015 (848)           | 1                 | 11 (4)                  | 0 (0)                      | 11                              | 0.21 (0.19, 0.24)                 |
| chr17:35547276-36603565     | rs11649743;               | 0.9                         | 1803 (444)           | 1                 | 26 (10)                 | 0 (0)                      | 12                              | 1.24 (1.10, 1.42)                 |
| chr17:46302314-47211374     | rs138213197               | 0.9                         | 2338 (521)           | 1                 | 1 (1)                   | 0 (0)                      | 0                               | 6.87 (4.24, 10.41)                |
| chr17:47211375-47952263     | rs11650494;               | 0.9                         | 1319 (378)           | 1                 | 83 (3)                  | 0 (0)                      | 24                              | 0.07 (0.02, 0.14)                 |

Published GWAS SNPs for which the signal or region replicated in our EUR meta-analysis are indicated, alongside the region co-ordinates assigned for fine-mapping analyses (GRCh37/hg19 assembly). The final priori pryuner threshold used and numbers of variants and priority pruner tags included in the analysis are shown. Summaries of the fine-mapping results for each region contain the number of independent PrCa risk signals identified within each region, the size of the set of variants identified by JAM and the number of variants within the credible set that were also significantly associated eQTLs in TCGA PRAD data. As an additional category to assist variant prioritisation, the number of variants in the credible set that achieved a nominally significant \( P \) value (\( P < 0.05 \)) in an unconnected African Ancestry meta-analysis is indicated. The estimated contribution of each GWAS region to the overall familial relative risk of PrCa after fine-mapping is also provided. These results are a continuation from the regions displayed in Tables 1 and 2. Aggregated summary results across all of the 80 regions fine-mapped presented across Tables 1–3 are displayed in the final row of this table (in bold).

a AAs: African Ancestry population PrCa meta-analysis.

b 84 of the 95 original GWAS signals identified in fine-mapping replicated in our EUR meta-analysis and were used when performing calculation of Familial Relative Risk of PrCa. rs2055109, rs7210100, and rs66275711 did not replicate in EUR but are situated within the regions defined in Tables 1 and 2. Aggregated results across 80 or 37,863 signals were not included prior to fine-mapping. For five previously reported variants (rs753648, rs2051443, rs636291, rs1571801, and rs103294)), no variant within the region boundary replicated in the meta-analysis, and these regions were excluded prior to Bayesian analysis.

AR (\( P = 2.33 \times 10^{-15} \) to \( 2.86 \times 10^{-12} \)), ERG (\( P = 5.33 \times 10^{-12} \) to \( 1.00 \times 10^{-20} \)), FOXA1 (\( P = 9.18 \times 10^{-18} \) to \( 1.14 \times 10^{-18} \)), GABPA (\( P = 4.39 \times 10^{-12} \)), GATA2 (\( P = 1.24 \times 10^{-12} \)), HOXB13 (\( P = 8.25 \times 10^{-9} \)) and NKX3.1 (\( P = 9.44 \times 10^{-8} \) to \( 1.43 \times 10^{-15} \)) transcription factor-binding sites from one or more experimental data set; for H3K27Ac (\( P = 5.34 \times 10^{-19} \) to \( 1.39 \times 10^{-21} \)) and H3K4Me3 (\( P = 1.30 \times 10^{-39} \) to \( 8.27 \times 10^{-14} \)) histone marks; and conserved elements within the human genome according to all four variants (\( P = 1.89 \times 10^{-7} \) to \( 4.04 \times 10^{-11} \)). Of particular interest, in over half of the regions fine-mapped, at least one variant within our credible set intersected a significantly associated eQTL with a colocalisation score >0.9 (overlap between eQTL and GWAS signal) in the TCGA PrCa data set. In all, 40 of the 75 regions contained an eQTL variant among the credible set, with 91 distinct genes represented (Tables 1–3, Supplementary Data 3). In total, 127 of the 343 tags representing the credible set included an eQTL annotation (37%), compared with 5711 of the total 37,863 tags within these regions (17.8%). This corresponds to 1027 prostate eQTL variants among the 3700 credible set variants represented by the 343 JAM tags (27.8%), compared with 37,331 eQTLs from the 203,211 total variants within these 75 regions (18.4%).

Intuitively, some degree of correlation between the annotation features we examined would be expected, since regulatory regions of DNA may be indicated through various experimental techniques. Although annotations were jointly modelled in QR, any partial correlation could potentially inflate the extent of enrichment observed during independent Fisher’s tests. To preclude this outcome, we examined the level of correlation between separate annotations. Correlation between replicate data sets representing the same annotation category was usually moderate to high as would be expected, with more modest levels
of correlation observed between different markers and information types (Supplementary Fig. 4). The level of correlation increased slightly when individual SNP annotations were collapsed onto tags, as the tag variants can inherit different annotations from separate SNPs. We performed logistic regression of the annotations used in the QR analysis in a single model, to evaluate their informativeness after adjustment for other annotation categories. In this regression, the TCGA eQTL coding transcript and ERG transcription factor annotations were all highly significant after adjusting for multiple testing, whilst the AR transcription factor annotation was also nominally significant (Supplementary Fig. 5). The remaining annotations were not significant after adjustment for other annotations; however, within the range of information types selected, separate data sets represent broader or greater resolution functional information relative to one another and therefore may partially overlap with other markers whilst remaining instructive individually.

**Fine-mapping resolution.** At several regions our catalogue of variants highlighted putative biological mechanisms that may be responsible for the differential risk of PrCa development, as well as credible sets sufficiently small to enable subsequent laboratory follow-up. One example is the Chr2q37 region described by rs3771570 in the original publication. The original lead variant is intronic in FAR2P2, but multiple genes are located within the region. During fine-mapping, we observed evidence for three independent signals, one more than we previously detected. These signals are represented by a credible set of 14 variants from 7 tags, demonstrating highly successful refinement of the original signal (Fig. 3a, Tables 1–3, Supplementary Data 1). The majority of these prospective causal variants are centred on the ANO7 gene, approximately 100 kb centromeric of FAR2P2. ANO7 is expressed predominantly in the prostate (http://www.proteinatlas.org/ENSG00000146205-ANO7/tissue), unlike FAR2P2, which is ubiquitously expressed across tissue types. Within the credible set 3 tags are selected with particularly high confidence (posterior probabilities 0.72–1); all 3 represent only themselves with no additional proxy variants to consider, and are therefore the most likely causal variants underlying the 3 signals detected. Two of these 3 candidate causal variants (rs77559646 and rs77482050) are non-synonymous SNPs in ANO7 that are uncommon among European ancestry populations, whilst the third (rs62187431) is intronic in ANO7. The 11 remaining variants in the credible set include one more missense SNP within ANO7 (rs76832527), 2 intronic variants in ANO7 (rs111770284 and rs6091437), a synonymous variant in ANO7 (rs2074840) and 7 variants that are all intronic within other genes (FARP2, PPP1R7, HDLBP and SEPT2). Our fine-mapping results therefore strongly implicate the ANO7 gene as a prospective biological effector modulating susceptibility for PrCa.

The region at Chr6q22 described by rs339331 in the original publication presents a good example of how variant annotations can assist further prioritisation of the most likely candidate variants even within regions where the credible set remains comparatively large after fine-mapping (Fig. 3b, Tables 1–3, Supplementary Data 1). rs339331 is intronic in RF6X, a member of the regulatory factor X transcription factor family. We observed a single signal during fine-mapping, but due to high LD between variants the credible set comprises 102 variants from 3 tags (the top tag with posterior probability 0.76 tagging 35 proxy SNPs, another with posterior probability 0.15 tagging 40 SNPs and the last with posterior probability 0.08 tagging 27 SNPs). Only 14 of these variants demonstrate any plausible biological evidence however, therefore the credible set can be filtered to prioritise this subset of variants. Four of these are proxies of the tag with the greatest statistical evidence, including the variant that demonstrates the greatest biological evidence for functionality; the original index SNP rs339331, which resides within a DNaseI peak, intersects binding sites for multiple transcription factors, including AR, FOXA1, GATA2, HOXB13 and NKX3.1, and is situated within a conserved element. rs339331 would therefore be ranked highest for follow-up based on combined statistical information and biological annotations, and has been demonstrated to alter HOXB13 transcription factor binding and RXF6 transcription during a previous functional investigation of this region.

At the TMPRSS2 region on Chr21q22, we detected a single PrCa risk signal with a credible set of 31 SNPs from 8 tags, all of which are situated within the promoter region or first intron of TMPRSS2 (Fig. 3c, Tables 1–3, Supplementary Data 1). In all, 20 of these variants are eQTLs for TMPRSS2 in prostate tissue, whilst 2 variants intersect transcription factor-binding sites in multiple data sets, including for AR, ERG, FOXA1, GABPA, GATA2, HOXB13 and NKX3.1. In this region, the tag selected by JAM with the highest posterior probability is substantially downweighted after QR (ΔPosterior probability = 0.18) due to lack of overlap with informative biological annotations, therefore it and its proxies may not in fact represent the most likely candidate causal variants. An early and common event in prostate tumour development involves a translocation that forms a TMPRSS2: ERG fusion, bringing the ERG transcription factor under transcriptional control of the more active TMPRSS2 promoter. Our fine-mapping results and biological annotations therefore allude to the possibility that subtle, heritable differences in TMPRSS2 expression could potentially operate in conjunction with a common somatic alteration to influence development of PrCa. Intriguingly, we also observed significant enrichment for variants intersecting ERG transcription factor-binding sites among our combined credible set of candidate variants across all regions using Fisher’s exact test (Supplementary Data 2, Fig. 2).

**Comparison with African Ancestry meta-analysis results.** Since LD patterns and allele frequencies of variants frequently differ among ancestral populations, as an additional prioritisation strategy we cross-checked meta-analysis results for variants in our 95% credible set against data from a meta-analysis of 10,202 cases and 10,810 controls with African Ancestry (AA). A total of 3633 of the 3700 SNPs in our credible set were available in the AA cohort, 1155 (31.8%) of which were nominally significant at P < 0.05 in the AA meta-analysis. In addition, of the 175 variants that reached genome-wide significance within the five regions in which JAM did not resolve candidate variants, 111 were nominally significant in the AA data. We would hypothesise that variants demonstrating no evidence of association in the AA data set would generally represent less likely candidate causal variants than any nominally significant variants within their region specific credible set and should be assigned lower priority when considering variants for functional confirmation studies. This extra prioritisation step does not enable us to formally exclude any variants from our credible set however, as the AA analysis may be underpowered to detect association with PrCa at specific SNPs, and additional variants within the regions fine-mapped in Europeans but not included in our credible set were not examined for association in AA data.

**Estimating the GWAS loci contribution to FRR of PrCa.** The proportion of FRR of PrCa explained by these risk loci before and after fine-mapping were calculated using conditional effect estimates and standard errors derived from the OncoArray sample...
The post fine-mapping calculation was performed separately for the full set of 99 signals identified and a restricted subset of 84 variants (matching the number of original associations), in order to investigate the relative importance between replacement of GWAS tag SNPs and addition of extra novel signals. Single lead variants representing the independent signals were selected for this calculation. In regions containing a single signal, the JAM tag in the credible set with the highest Bayes factor was designated as the new lead variant, or for the signal, the JAM tag in the credible set with the highest Bayes factor. Our FRR calculations use conditional posterior support by JAM. Our FRR calculations use conditional risk estimates incorporating uncertainty for each variant, plus a correction for potential bias due to risk estimation in the same sample as discovery and uncertainty in the specification of the FRR. This novel but more conservative method of risk calculation estimated that: (1) inclusion of only single ‘best’ replacement variants for each tag SNP contributes 26.5% (95% credible interval, CI, 22.7–31.5) of the known FRR of PrCa compared to 23.2% (95% CI 19.4–27.9) for the 84 previously known GWAS tag SNPs; and (2) inclusion of lead SNPs representing all of the 99 independent signals contributes 30.3% (95% CI 26.0–35.9) (Supplementary Data 4). This substantial enhancement demonstrates that the variant catalogue identified through fine-mapping explains a greater proportion of the FRR of PrCa compared to the original GWAS index SNPs, with replacement of the 84 original GWAS tag SNPs conferring a similar magnitude of increase as addition of the 15 novel independent signals we identified. We additionally calculated the contribution to FRR of PrCa for each region individually, to highlight regions that make the greatest contributions towards PrCa susceptibility (Tables 1–3). Whilst the majority of the fine-mapped GWAS loci individually contribute a small proportion towards the FRR, six regions confer in excess of 1% each. These include the moderate penetrance HOXB13 rs138213197 variant, which demonstrated the greatest contribution at 6.87%, and the multi-signal TERT locus, which explained the next highest level at 2.57%. Each of the remaining regions of higher FRR contribution contained multiple independent signals, with the exception of the single-signal MSMB locus. The magnitude of increase in proportion of FRR explained by each locus after fine-mapping was also generally greater for regions where additional independent signals were identified; for example, the ANO7 region increased 6.5 fold (from 0.1% for the original GWAS tag SNP to 0.65% after fine-mapping) and the KLK region 1.9 fold (from 0.45 to 0.86%), partly due the identification of 2 novel signals within each.
with those selected in the credible set. Names of the representative variants for each independent signal used in the familial relative risk calculation are shown in black and the original GWAS tag SNP marked in red. Only variants selected in the credible set are shown on the positions of tags included in the 95% credible set marked as dashed lines and positions of all their respective proxy SNPs indicated as coloured circles. Posterior probability of association for priority pruner tags (PostProb panel). Triangles and circles on the meta-analysis plot denote variants directly underrepresent the uncertainty in the analysis and yield false approaches, such as forward stepwise selection, which tend to combinations of variants within each region, as determined by a set of posterior probabilities of association for each variant and combinations of variants within each region, as determined by a set of best models. This framework is preferred over alternative approaches, such as forward stepwise selection, which tend to underrepresent the uncertainty in the analysis and yield false levels of confidence for the final set of SNPs and number of signals represented by the single ‘best’ model. JAM also has advantages over similar Bayesian variable selection algorithms as it incorporates an extremely computationally efficient formal reversible jump Markov Chain Monte Carlo (MCMC) stochastic model search, which allows application to very large regions and does not require a prior assumption on the maximum number of causal SNPs within each region, making it more applicable to regions with larger or unknown numbers of causal variants. Linear model-based summary data methods such as JAM represent the current state of the art and have demonstrated good performance when applied to transformed logistic ORs from binary traits as opposed to linear effects for continuous traits. The effectiveness of logistic/linear mapping will however vary between different genomic architectures and is dependent on factors including the number of variants and correlation structure between them within each region. In general however, the approximation should work well provided no individual variants exert large effects, as expected for GWAS loci. For 5 of the 80 regions that had replicated at genome-wide significance, JAM was unable to fit a model to the summary data and consequently we could not resolve candidate variants beyond the catalogue of genome-wide significant variants within these regions. Four of these regions were not densely genotyped on the OncoArray genotyping chip, as their discovery in a multi-ethnic meta-analysis occurred only late during chip design. In addition, the top hit within these 5 regions ranked towards the weaker end of the $P$-value and effect size distributions in the univariate meta-analysis.
prior to fine-mapping. The inability of JAM to resolve candidate causal variants within these regions therefore most likely results from mismatch between the reference correlation structure and meta-GWAS effect patterns, issues with the logistic/linear mapping in the presence of complex correlation structure, or possibly simply low signal to noise ratio within the data.

Use of multivariate models prioritised a 95% credible set of 3700 candidate variants from the 203,211 variants analysed within the 75 regions in which candidate variants were resolved; thereby markedly reducing the number of variants for further consideration. In addition, previous reports of multiple independent signals at several PrCa risk loci were confirmed, with evidence for multiple signals at 12 regions; of which 7 regions contained 2 signals, 4 demonstrated evidence for 3 signals and 5 signals were observed at the Chr5p15 TERT gene locus, which is known to contain susceptibility variants for many cancer types. We observed no consistent pattern of LD relationship between the original GWAS tag SNPs and the independent signals identified through fine-mapping in the regions containing multiple independent signals (Supplementary Fig. 6). For example, at the ANO7 locus, the original index SNP (rs3771570) is not selected in the credible set and correlated with only 1 of the 3 independent signals detected (rs62187431, $r^2 = 0.61$). In contrast, at the TERT region, the original index SNP (rs2242652) is in moderate or modest LD ($r^2 = 0.08–0.43$) with 4 of the variants selected by JAM as representative of the 5 independent signals. Previous smaller fine-mapping studies using stepwise selection approaches had also identified evidence for independent association signals within several regions. However, these are potentially more sensitive towards subjective measures such as the P-value threshold chosen for secondary signal inclusion and LD level used to define the final list of candidate variants represented by the selected marker(s). Due to our substantially larger sample size and variant density available and the well-established superiority of Bayesian search procedures over stepwise selection in high-dimensional settings, we therefore consider this the most detailed fine-mapping study to date for variant prioritisation. Comparing our results to the previous iCOGS fine-mapping study, in which refinement of 64 GWAS loci was attempted in a smaller European ancestry cohort of 25,723 PrCa cases and 26,274 controls, 48 regions corresponding to 52 original index SNPs replicated at genome-wide significance in both studies, of which only 21 regions had been densely genotyped on the iCOGS chip (Supplementary Data 5). Within these comparable regions, 70% of the ‘best candidate SNPs’ established using the iCOGS sample set were also included in the credible set we have identified in this study. This indicates broad stability of the results from fine-mapping studies conducted in the same ancestral population. The additional power and more dense genotyping across all regions in this study has however facilitated further refinement of potential candidate variants, identification of additional candidate variants within several regions and refinement for the first time of a number of regions in which fine-mapping had not previously been performed or had been unsuccessful. We have confirmed the existence of multiple independent risk signals at 10 loci previously reported, including identifying extra signals at the TERT (Chr5p15), ANO7 (Chr2q37) and SLC22A3 (Chr6q25) loci, and identified multiple independent association signals for the first time at two further loci, including KLK3 (Chr19q13). Eight regions demonstrating evidence for multiple independent signals in the iCOGS fine-mapping study were however not corroborated in this larger study. Notably, the conditional P-values for these secondary signals in the iCOGS fine-mapping study were below genome-wide significance in all but one of these regions. This may suggest that contrary to general assumptions that a lower burden of evidence is valid for uncorrelated variants in loci for which a genome-wide significant association has previously been observed, instead equally stringent significance thresholds should be applied for both secondary signals and initial primary signals. It is also notable that in this well-powered study, the vast majority of regions containing multiple independent signals were first reported as associated with PrCa in early GWAS using relatively modest sample sizes. This may indicate that regions with lower effect sizes and weaker evidence for association, which require larger sample sizes for their detection, are less likely to contain additional independent risk variants. Alternatively however, it could reflect lower power for the detection of additional independently associated variants within the regions that contain weaker signals, despite the large sample cohort utilised in this study.

As would be expected, refinement of putative causal risk variants varied between regions, with credible sets ranging from a single variant or handful of variants to >100 variants for a small number of regions. The regions retaining large credible set sizes appear to result primarily from large numbers of variants in high LD with the actual causal variant as opposed to low power within the region however, rendering further refinement of these signals to facilitate functional validation studies more complicated. One approach to further prioritise candidate variants could be to leverage the different LD patterns among different ancestral populations, provided that the underlying causal variants are shared and present at sufficient frequency between populations. Cross-referencing the 3700 variants within our 95% credible set with data for an African American PrCa meta-analysis from the African Ancestry Prostate Cancer GWAS Consortium highlighted a subset of 1155 variants with nominal or genome-wide significant evidence for association in this additional population. An alternative prioritisation approach is to consider pre-existing biological information, as we have described for the RFX6 (Chr6q22) region. We annotated variants against a number of publically available data sets, observing enrichment of several plausible markers of biological function active in prostate cell lines within our credible set, including intersection with prospective promoter and enhancer elements, DNaseI hypersensitivity sites, histone modification or transcription factor-binding peaks, and variants residing within protein-coding transcripts and conserved regions of the genome. Of particular interest, more than a quarter of the variants within our credible set were also eQTLs within the TCGA prostate adenocarcinoma data set. Given their statistical selection independent of this annotation and demonstrated effects upon gene expression, these eQTL variants should be considered high priority when selecting candidate causal variants for functional confirmation, alongside variants that modify the coding sequence of genes, or appear to reside within reliably annotated promoters or enhancers. Another important discovery of this study is that an appreciable number of highly ranked variants within the credible set are non-synonymous SNPs. This provides evidence that subtle alterations to structure and activity of specific proteins may give rise to the functional mechanisms behind a proportion of GWAS associations.

Some alternative fine-mapping algorithms integrate functional annotations during the statistical analysis when considering evidence for causality for each variant. These methods can prove useful for enhancing variant prioritisation, provided that the annotation information is reliably indicative of causal variants. We preferred to perform statistical analysis separately from annotation and compare statistical and functional evidence for causality afterwards using conditional QR. We believe this more clearly allows the most informative annotations, and the variants that are characterised by those annotations to be highlighted within the data set, whilst also reducing the potential for...
penalisation of strong candidate variants due to localised artefacts or cell line-specific effects within the whole-genome biological data sets used for annotation. Our conditional QR analysis resulted in adjustment of posterior probability for a small proportion of variants and may further assist prioritisation of the most likely functional variants among the credible set selected for each region.

Fine-mapping studies are important to reveal information on the biological mechanisms underlying disease predisposition by pinpointing potential candidate genes, signalling pathways and networks that account for differences in disease risk between individuals. In addition, these studies may help to refine the contribution of GWAS loci to PrCa risk by incorporating more likely candidate variants. This study evaluated almost all previously reported PrCa GWAS regions, apart from the highly complex Chr8q24 and major histocompatibility complex (MHC) regions and associations that did not replicate in the largest European meta-analysis to date. We then subsequently re-evaluated the contribution to FRR of these known PrCa risk loci using an enhanced method in which the overall FRR of PrCa was revised upwards from 2.0 to 2.5 to reflect the most recent estimates and also accounted for uncertainty of various estimates that can introduce bias in these calculations. Our approach therefore provides more conservative estimates than in previous publications. We demonstrated a substantial increase in the proportion of FRR explained through fine-mapping these GWAS regions (from 23.2 to 30.3%), with detailed investigation showing that a similar proportion of this enhancement was conferred by replacement of the original tag SNPs and discovery of secondary signals. It is also noteworthy that the 7.1% magnitude of increase in FRR explained after fine-mapping known loci is substantially greater than the 4.4% increase achieved through identification of 62 novel PrCa loci. This highlights the invaluable importance of fine-mapping studies for risk prediction and their potential utility in helping to inform clinical screening studies.

Fine-mapping of GWAS loci requires comprehensive examination of variation within the region. Logistical constraints generally preclude resequencing of disease-associated loci to achieve complete variant coverage in large sample cohorts and instead mandate the use of genotype array data followed by imputation, in order to achieve sufficient sample sizes. To ensure the accuracy of downstream fine-mapping analyses, stringent variant QC must be applied to imputed data, to exclude low-quality variants that may be indicative of imputation artefacts. In this study, initial pre-imputation QC of the meta-analysis data set was first performed to exclude potential genotyping errors, followed by post-imputation QC in which variants with low MAF or imputation information score, or divergent MAF consistency between dosage and ‘best guess’-derived MAF estimates were excluded. The MAF estimate consistency check was performed to highlight additional variants for which reliability of imputation may be reduced and evaluation of variants excluded in this step revealed that the majority were situated within segments of the genome flagged as repetitive or otherwise ambiguous. Whilst we cannot guarantee that no causal variants at GWAS loci would be located within repetitive elements, we believe that the high proportion of variants filtered during QC that are located within potentially difficult to impute segments indicates an appropriate balance between controlling against both type I and II errors during the subsequent fine-mapping analyses. The inability to directly interrogate this category of variants during this study could however reflect a potential limitation.

The multivariate fine-mapping strategy we employed enabled identification of small numbers of prospective causal variants amenable to functional follow-up at many known PrCa susceptibility regions. Within this credible set of variants, we found evidence of enrichment for a number of biologically plausible mechanisms through which PrCa risk could potentially be modulated. We observed multiple independent PrCa associations at 15% of the loci fine-mapped, and several candidate genes were indicated for consideration through functional annotation. As rare variants with MAF < 0.005 were not included in our analyses, we cannot exclude a contribution of rare casual variants exerting a greater effect size giving rise to synthetic associations at any GWAS loci, although our findings indicate that these are unlikely to be widespread. Importantly, replacement of the original GWAS tag SNPs with more likely candidate variants and identification of additional independent signals resulted in a substantial increase in the proportion of the FRR of PrCa explained by these loci. This finding accounts for a portion of the ‘missing heritability’ of PrCa and has important implications for clinical risk profiling and management of patients.

Methods
Identification of PrCa risk loci to fine-map. We identified 101 independent PrCa GWAS risk associations within the literature that had been reported at genomewide significance prior to the start of this study, the majority of which had previously been replicated within a European ancestry population. Six of these lead variants were located within the Chr8q24 region that is associated with multiple cancer types in a highly complex manner, and three within the MHC Chr6p21 linkage region. Due to the large numbers of variants, high levels of correlation and greater complexity within these regions, they are the subject of separate fine-mapping and risk stratification studies and were excluded from consideration in this analysis; the remaining 92 previously reported GWAS SNPs were selected for fine-mapping in this study. For 5 of these originally reported GWAS SNPs, no variant within ±500 kb replicated at genome-wide significance in our larger European meta-analysis and these loci were subsequently excluded from downstream Bayesian analyses and FRR calculations. An additional 2 GWAS SNPs originally reported in non-European ancestral populations and 1 reported in a previous meta-analysis did not replicate, but were situated <500 kb from an independent, replicated European risk association and were therefore still considered within the region boundaries of signals that were fine-mapped.

Selection of SNPs for fine-mapping on the OncoArray. A total of 78 PrCa risk associations that had been reported prior to the design of the OncoArray genotyping platform were densely genotyped within the OncoArray sample cohort. Region boundaries for dense genotyping were defined as the greater of ±500 kb from the index SNPs or the maximum distance of any variant with r2 > 0.3 to the index SNP in 1KG (phase 1 version 3, March 2012 release). All SNPs within these regions with MAF > 0.01 in any ancestral population were extracted and then we obtained Illumina Design Scores for all variants from the 1000 Genomes Project (phase 1 version 3, March 2012 release). From designable variants with a Design Score > 0.8, we used Snagger to select (a) all variants correlated to the known hits at r2 > 0.6 and P < 0.05 in the iCOGS study, (b) all variants from lists of potentially functional variants, defined through ENCODE and RegulomeDB and (c) a set of SNPs to tag all remaining variants at r2 > 0.9. The 23 risk loci reported in a recent multi-ethnic meta-analysis study were not densely genotyped as these loci were reported after the OncoArray design; however, these regions were also fine-mapped in this study.

Meta-analysis and imputation. Genotype data for a combined 82,591 PrCa cases and 61,213 controls of European ancestry from eight GWAS (OncoArray, iCOGS, UK stage 1 and 2, CaPS 1 and 2, BPC3 and NCI PEGASUS) were used for the meta-analysis. Per allele ORs and standard errors were generated for the OncoArray and each GWAS, adjusting for principal components (PCs) and study relevant covariates using logistic regression. The OncoArray and iCOGS analyses were additionally stratified by country and study, respectively. We used the first seven PCs for OncoArray and first eight PCs for iCOGS samples, as additional components did not further reduce inflation in the test statistics. OR estimates were derived using either SNPTEST (http://mathgen.stats.ox.ac.uk/genetics_software/snpset/nptest.html) or an in-house software C++ programme. OR estimates and standard errors were combined by a fixed effects inverse variance meta-analysis using METAL. All statistical tests conducted were two-sided.

IMPUTE2 was used to impute non-genotyped SNPs within a boundary flanked by ±500 kb of the maximum distance of any variant with r2 > 0.3 to the index SNP in 1KG phase 1 from the originally reported GWAS index SNP in the meta-analysis cohort. For the OncoArray data, un-phased imputation was carried out for all the fine-mapping regions. Where the boundaries of adjacent associations to fine-map overlapped, these were merged for imputation; therefore, imputation was performed on 82 discrete chunks. Within 3 of these chunks the separate signals to analyse were sufficiently dispersed to enable clear demarcation of the individual signals and retention of an appropriate flank distance; these 3 imputation chunks
were therefore split prior to statistical analysis and the 92 original index SNPs analysed were fine-mapped as 85 separate regions.

We conducted a two-stage post-imputation QC process. During basic QC, imputed genotype data were filtered to retain variants with INFO ≥ 0.4 and MAF ≥ 0.005. We subsequently instituted an additional QC measure to remove imputed variants with greater genotype uncertainty in which separate MAFs were calculated based on the minor allele frequency (MAF) estimates for a variant would indicate unreliable imputation performance; variants for which these differed by ≥10% were excluded from analysis. An additional benefit of this methodology is that inherently applies progressively greater stringency of QC filtering the rarer a variant is within the study population. During the post-imputation QC process, 288,033 rare variants were excluded, whilst a further 146,088 variants were removed due to low INFO score or divergent MAF consistency. This resulted in a final post-QC set for analysis of 213,728 SNPs within the 80 fine-map regions that had replicated in the initial meta-GWAS, with a minimum variant INFO score within the final data set of 0.63, and the vast majority of variants having INFO > 0.9 (Supplementary Fig. 7).

As an additional safeguard, we investigated the proportion of common variants (MAF ≥ 0.05) in 1000 Genomes European samples that were retained or excluded during our QC procedure. In total, 186,907 of 227,793 common 1000 Genomes based on variants with greater genotype uncertainty in which separate MAFs were calculated all adjusted for one another, and standard model above through by the transpose of the genotype matrix, JAM

\[ y \sim N(X\beta, \sigma^2) \]

where \( \sigma^2 \) represents the residual variance, \( \beta \) represents a vector of effects, which are all adjusted for one another, and \( I \) is the \( N \times N \) identity matrix. Multiplying the standard model above through by the transpose of the genotype matrix, JAM makes inference under the resulting multivariate normal (MVN) model:

\[ y \sim N(X\beta, \Sigma) \]

The motivation for using the model in (2) rather than (1) is that individual-level data are no longer required; \( Xy \) can be derived from one-at-a-time univariate effect estimates of each variant\(^{67,68} \) and \( XX \) from an estimate of the genetic correlation matrix. Note that in the case of the PrCa summary statistics, we derive \( Xy \) after first mapping the univariate log ORs to approximate linear effects via their z-scores, a strategy adopted for binary traits in other linear model-based summary statistic frameworks\(^{7,48} \). Consequently, the model residuals have the same interpretation as in a linear regression of a binary outcome; they cannot exceed 1 and, under the null model, their variance \( \sigma^2 \) equals the trait variance, \( \mathbb{P}(1 - p) \) where \( p \) is the proportion of cases. Since each region is unlikely to explain much heritability individually, we specify an inverse gamma \( (\Gamma \sim 1^{-1}) \) prior that loosely targets the PrCa variance in the meta-GWAS:

\[ \sigma^2 \sim \Gamma(1, 0.24) \]

This corresponds to a prior expectation for \( \sigma^2 \) equal to the PrCa variance in the meta-GWAS, 0.24, and 95% weight over the range (0.05, 0.69). The JAM model is completed by specifying a so-called ‘g-prior’ over the genetic effects, \( \beta \):

\[ \beta \sim MVN(0, \sigma^2(X'X)^{-1}) \]

The conjugate g-prior supports effects inversely proportional to the corresponding genetic co-variances and variances, as estimated from the reference matrix, \( XX \), and has been shown to help to regularise models by a factor of 10-100\(^{69} \). There is a substantial literature on choices for the hyper-parameter, \( \gamma \); we follow recommendations to set a value equal to the maximum of \( N \) and \( P^2 \), where \( P \) is the number of variants in the region\(^{60,62} \).

Crucially, both (1) and (2) are parameterised by the same vector of multivariate \( \beta \); correlation adjusted effect estimates. JAM is therefore able to utilise information from a multivariate analysis of individual-level data. Optimal performance is achieved when the correlation structure \( XX \) is taken from the original GWAS population, rather than an external reference population. We applied JAM to summary statistics from the meta-analysis data set using LD pruning (38,745 selected tags). The data was processed using wANNOVAR in relation to GENCODEv19 transcripts\(^{58} \). Variants residing within mRNA transcripts were subsequently added in relation to miRBase release 20 (ftp://mirbase.org/pub/mirbase/20/genes/hsa.gtf)\(^{70} \). Annotation of variants that reside within genomic elements demonstrating evidence for evolutionary constraint was performed against conserved element peak outputs from four algorithms; GERP++ (http://med Mendel.stanford.edu/SidowLab/downloads/gerp)\(^\text{35} \), SiPhy_Omega, SiPhy_Pi (https://www.broadinstitute.org/science/community/science/projects/mammals-models/29-mammals-project-supplementary-info)\(^{21} \) and PhastCons (ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/phastCons100way/)\(^{22} \). Annotation of variants that lie within conserved elements was used to identify predicted likely consequence or relevance of the variant resulting from its primary genomic context, or the proximity to annotated regulatory features within cell lines derived from normal prostate or PrCa tissues.

Gene-based annotation of variants was performed using wANNOVAR in relation to GENCODEv19 transcripts\(^{58} \). Variants residing within mRNA transcripts were subsequently added in relation to miRBase release 20 (ftp://mirbase.org/pub/mirbase/20/genes/hsa.gtf)\(^{70} \). Annotation of variants that reside within genomic elements demonstrating evidence for evolutionary constraint was performed against conserved element peak outputs from four computational genomics analyses by four algorithms; GERP++ (http://med Mendel.stanford.edu/SidowLab/downloads/gerp)\(^\text{35} \), SiPhy_Omega, SiPhy_Pi (https://www.broadinstitute.org/science/community/science/projects/mammals-models/29-mammals-project-supplementary-info)\(^{21} \) and PhastCons (ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/phastCons100way/)\(^{22} \). Annotation of variants that lie within conserved elements was used to identify predicted likely consequence or relevance of the variant resulting from its primary genomic context, or the proximity to annotated regulatory features within cell lines derived from normal prostate or PrCa tissues.

Similar to other Bayesian stochastic variable selection approaches, JAM models a latent vector of binary indicators, \( y \), for whether each variant should be included (\( y = 1 \) if variant \( v \) is associated and included in the model, or \( 0 \) otherwise). Any specific configuration of indicators then specifies a specific model, \( M \). Using a Bayesian stochastic search, specifically a Reversible Jump MCMC (RJMCMC) algorithm\(^{34,16} \), JAM searches over different possible models. By specifying a prior on the probability of including a variant, \( \phi \), a prior for the probability of the model space, \( \psi \), and a prior that the model space, \( \psi \), and a prior parameter, \( \psi \), that indicates the probability of including any combination of variants, we induce a prior over all possible models according to the beta-binomial distribution. Since these are previously discovered regions, this is far more generous than our prior belief would be that a random region of the genome is associated with PrCa but is more conservative for the regions in this analysis, where we estimate the false discovery rate is <10%. The marginal prior odds of any particular SNP being selected is 1/1, and decreases with the total number of variants in the region, providing an intrinsic multiplicity correction as a function of region size\(^{35-36} \). The prior probabilities for \( \psi \) range from 0.25 to 0.12, respectively (Supplementary Table 1). More detail on the JAM16 and RJMCMC algorithm can be found in the original paper\(^{16} \). For this analysis, each region was analysed independently, and by running two independent JAM seeds for 10 million iterations each. The JAM output provides posterior probabilities for each variant, \( Pr(y = \text{in data}) \), and for each combination of variants, \( Pr(M = \text{model}) \). To determine statistical significance for individual variants, combinations of variants and for the possible number of independent signals we use Bayes factors\(^{57} \), the ratio of the posterior odds to the prior odds. Specifically, we used the inference of the minimum of independent signals in the model at a regional Bayes factor threshold of 3 to define the evidence for multiple signals.

Pruning thresholds were 0.01 (http://prcancer-genome-sequencing-project.org) for 10 million iterations each. The JAM output provides posterior probabilities for each variant, \( Pr(y = \text{in data}) \), and for each combination of variants, \( Pr(M = \text{model}) \). To determine statistical significance for individual variants, combinations of variants and for the possible number of independent signals we use Bayes factors\(^{57} \), the ratio of the posterior odds to the prior odds. Specifically, we used the inference of the minimum of independent signals in the model at a regional Bayes factor threshold of 3 to define the evidence for multiple signals.

Pruning thresholds were 0.01 (http://prcancer-genome-sequencing-project.org) for 10 million iterations each. The JAM output provides posterior probabilities for each variant, \( Pr(y = \text{in data}) \), and for each combination of variants, \( Pr(M = \text{model}) \). To determine statistical significance for individual variants, combinations of variants and for the possible number of independent signals we use Bayes factors\(^{57} \), the ratio of the posterior odds to the prior odds. Specifically, we used the inference of the minimum of independent signals in the model at a regional Bayes factor threshold of 3 to define the evidence for multiple signals.

Pruning thresholds were 0.01 (http://prcancer-genome-sequencing-project.org) for 10 million iterations each. The JAM output provides posterior probabilities for each variant, \( Pr(y = \text{in data}) \), and for each combination of variants, \( Pr(M = \text{model}) \). To determine statistical significance for individual variants, combinations of variants and for the possible number of independent signals we use Bayes factors\(^{57} \), the ratio of the posterior odds to the prior odds. Specifically, we used the inference of the minimum of independent signals in the model at a regional Bayes factor threshold of 3 to define the evidence for multiple signals.

Pruning thresholds were 0.01 (http://prcancer-genome-sequencing-project.org) for 10 million iterations each. The JAM output provides posterior probabilities for each variant, \( Pr(y = \text{in data}) \), and for each combination of variants, \( Pr(M = \text{model}) \). To determine statistical significance for individual variants, combinations of variants and for the possible number of independent signals we use Bayes factors\(^{57} \), the ratio of the posterior odds to the prior odds. Specifically, we used the inference of the minimum of independent signals in the model at a regional Bayes factor threshold of 3 to define the evidence for multiple signals.

Pruning thresholds were 0.01 (http://prcancer-genome-sequencing-project.org) for 10 million iterations each. The JAM output provides posterior probabilities for each variant, \( Pr(y = \text{in data}) \), and for each combination of variants, \( Pr(M = \text{model}) \). To determine statistical significance for individual variants, combinations of variants and for the possible number of independent signals we use Bayes factors\(^{57} \), the ratio of the posterior odds to the prior odds. Specifically, we used the inference of the minimum of independent signals in the model at a regional Bayes factor threshold of 3 to define the evidence for multiple signals.
For annotation against prospective regulatory elements within the genome, which frequently operate in a tissue-specific context, these data sets were primarily retrieved from experiments using prostate-derived cell lines. We annotated variants that intersected DNaseI peaks data in seven individual ENCODE prostate data sets from three cell lines (LNCaP, PC3, and VCaP) with human prostate tumour tissue was downloaded from the Cistrome Data Browser (http://cistrome.org/db/); a resource that accumulates publically available ChIP-seq data sets and re-analyses their raw data through a standardised pipeline and QC procedure.25 Downloaded Cistrome data were converted from GRCm38 to GRCm38/hg19 reference assembly co-ordinates for compatibility with our variant data set using the UCSC Genome Browser LiftOver tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver). Transcript factor-binding site data were obtained for the Androgen Receptor (GSM1236922, GSM1328945 and GSM1328967), CTCE (GSM1236927, GSM1328967 and GSM1328978), FOXA1 (GSM1608136, GSM1724873 and GSM1717662), GABPA (GSM1193960), GATA1 (GSM1600544), HOX11B3 (GSM1717663 and GSM1717664) and Hign31.1 (GSM699633 and GSM699640). Histone modification data were obtained for H3K27Ac (GSM1249447 and GSM1249448), H3K27me3 (GSM1383866 and GSM1383872) and H3K4me3 (GSM1358374 and GSM1358376). Finally, to facilitate deeper categorisation of the genomic context of variants within prospective regulatory features, they were annotated with their chromatin state categorisations by ChromHMM from two prostate cell lines (PC3 and PCE; GSM574984, and three ENCODE tier 182 cell lines (GSM12878, H1HESC and HUVEC) to enable comparison of tissue specificity for prospective regulatory elements.26,27,28

eQTL analysis. Genotype and gene expression data for 494 samples with PrCa were downloaded from TCGA (https://gdc-portal.nci.nih.gov). For the genotype data set, QC was performed according to the protocol suggested by Anderson et al.49, removing samples with heterozygosity >2 standard deviations from the mean, individuals with low genotype call rate (<95%), non-male samples and related or duplicate samples (individuals with identity-by-descent >0.185). Variants with call rate <95% were also excluded from analysis. PC analysis was performed to correct for genotypic effects. Genes with mean expression across samples of ≤6 counts or with expression variance = 0 were also excluded (4123 and 370 genes removed, respectively). Finally, expression values were quantile-normalised by samples and rank-transformed by genes. In total, 16,038 genes passed QC out of the initial 20,531.

For the expression data set, we observed that samples from two plates (A31K and A30D) exhibited values that were significantly lower than the expression levels of the other samples, and which were also significantly different from the expected expression levels (Finnish samples removed from the analysis). In total, 108 samples and 106 SNPs in total were excluded (4123 and 370 genes removed, respectively). Finally, expression values were quantile-normalised by samples and rank-transformed by genes. In total, 16,038 genes passed QC out of the initial 20,531.

The eQTL and imputed SNP are the same variant. Second, we ranked the expression values for each pair of eQTL and imputed SNPs as:

$$v = \sum_i \text{ALD}(|ZB|\lambda_i, \tau_i)$$

Notice $v$ is affected by both $P$ and $ZB$, and it suggests a weighted average of $P$ and fitted regression quantiles can approximate $v$. The density function of ALD distribution is

$$f(v|ZB, \lambda, \tau) = \frac{1}{\lambda} \exp\left(-\frac{v - ZB}{\lambda}\right)$$

For $\lambda$, the maximum is achieved at

$$\lambda^* = \frac{\sum_i \rho_i \{v_i - ZB_i\}}{N}$$

We fix $\lambda$ to be

$$\lambda = \frac{\sum_i \rho_i \{v_i - ZB_i\}}{N}$$

where $\theta$ is coefficient estimates from classical conditional QR. With $\lambda$ fixed, only the exponential parts of the Gaussian distribution and ALD involve $v$, to which we assign weight to $P$ and $ZB$. Specifically classical QR yields a prediction for each SNP, i.e. $P_i = ZB_i$ at $\tau_i$; the larger the penalty $\rho_i \{P_i - \tilde{P}_i\}$ and $\tilde{P}_i$, the less influence $P_i$ should have on $ZB_i$. We normalise the weight for $P_i$ to be 1. For $P_i$, we assign weight

$$w_i = \exp\left(-\frac{\rho_i \{P_i - \tilde{P}_i\}}{\lambda} - \frac{\{P_i - \tilde{P}_i\}^2}{2\sigma^2}\right)$$

which is approximately the penalty at $v = (P_i + \tilde{P}_i)/2$. Our approximate value for $v$ is then

$$\hat{v} = \frac{P + \sum_i w_i y_i}{1 + \sum_i w_i}$$

Proportion of familial risk explained. The contribution and comparison of the newly identified SNPs and the previously known variants to the familial risk, under
a multiplicative model, was computed using the formula
\[ \lambda_m = \frac{p_a \lambda_m + q_a}{(p_a + q_a)^2} \]
where \( \lambda_m \) is the observed FRR to first degree relatives of cases and \( \lambda_m \) is the FRR due to locus \( m \), calculated assuming a per-allele effect:
\[ \lambda_m = \frac{p_a \lambda_m + q_a}{(p_a + q_a)^2} \]
where \( p_a \) is the frequency of the risk allele for locus \( m \), \( q_a = 1 - p_a \) and \( r_m \) is the estimated per-allele OR.

This calculation was performed using a Bayesian framework, which allows us to attenuate any ‘winners curse’ bias, and incorporates the uncertainty in estimating the variant-specific per-allele OR, \( r_m \) and the value of the observed familial risk, \( \lambda_m \).

To correct for potential bias in effect estimation from using the same sample to determine the credible set of SNPs (the so-called ‘winners curse’), we implemented a hierarchical model similar in spirit to Zhong and Prentice69 by placing a normal prior distribution on effect estimates of the form \( \beta_m \sim N(0, \tau^2) \), which places a 95% posterior probability in the range \([2.22, 2.78]\) on the FRR of \( \tau \).

The meta-analysis summary data used in this paper was obtained from the OncoArray sub-cohort of 53,449 cases and 36,225 controls, for which individual-level data were available.

The meta-analysis summary data used in this fine-mapping project are available from the PRACTICAL Consortium (http://practical.icr.ac.uk) or GitHub (https://github.com/oncogenetics/LocusExplorer/tree/master/Data/ProstateData). Results from the fine-mapping analyses may be explored interactively through Locus Explorer71 (http://www.oncogenetics.icr.ac.uk/LocusExplorer/).

Received: 28 April 2017 Accepted: 5 April 2018
Published online: 11 June 2018

References
1. Kote-Jarai, Z. et al. BRCA2 is a moderate penetrance gene contributing to young-onset prostate cancer: implications for genetic testing in prostate cancer patients. Br. J. Cancer 105, 1230–1234 (2011).
2. Pritchard, C. C. et al. Inherited DNA-repair gene mutations in men with metastatic prostate cancer. N. Engl. J. Med. 375, 433–435 (2016).
3. Mikropoulos, C., Goh, C., Leongamornlert, D., Kote-Jarai, Z. & Eeles, R. Translating genetic risk factors for prostate cancer to the clinic: 2013 and beyond. Future Oncol. 10, 1679–1694 (2014).
4. Conti, D. V. & Gauderman, W. J. SNPs, haplotypes, and model selection in a candidate gene region: the SIMPLE analysis for multicollinear data. Genet. Epidemiol. 27, 429–441 (2004).
5. Fedor, B. L. Bayesian variable and model selection methods for genetic association studies. Genet. Epidemiol. 33, 27–37 (2009).
6. Viarengo, V., Raftery, A. E. & Richardson, S. Variable selection and Bayesian model averaging in case-control studies. Stat. Med. 20, 3215–3230 (2001).
7. Wallace, C. et al. Dissection of a complex disease susceptibility region using a Bayesian stochastic search approach to fine mapping. PLoS Genet. 11, e1005272 (2015).
8. Schumacher, F. R. et al. Prostate cancer meta-analysis from more than 140,000 men identifies 63 novel prostate cancer susceptibility loci. Nat. Genet. https://doi.org/10.1038/s41588-018-0124-y (2018).
9. Ewig, C. M. et al. Germline mutations in HOXB13 and prostate-cancer risk. N. Engl. J. Med. 366, 141–149 (2012).
10. Saunders, E. J. et al. Fine-mapping the HOXB region detects common variants tagging a rare coding allele: evidence for synthetic association in prostate cancer. PLoS Genet. 10, e1004129 (2014).
11. Akamatsu, S. et al. Common variants at 11q12, 10q26 and 3p11.2 are associated with prostate cancer susceptibility in Japanese. Nat. Genet. 44, 426–429 (2012).
12. Al Olama, A. A. et al. A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. Nat. Genet. 46, 1103–1109 (2014).
13. Duggirala, R. et al. Two genome-wide association studies of aggressive prostate cancer implicate putative prostate tumor suppressor gene DARZP1. J. Natl. Cancer Inst. 99, 1836–1844 (2007).
14. Haiman, C. A. et al. Genome-wide association study of prostate cancer in men of African ancestry identifies a susceptibility locus at 17q21. Nat. Genet. 43, 570–573 (2011).
15. Xu, J. et al. Genome-wide association study in Chinese men identifies a new prostate cancer risk locus at 9q31.2 and 19q13.4. Nat. Genet. 44, 1231–1235 (2012).
16. Newcombe, P. J., Conti, D. V. & Richardson, S. JAM: a scalable Bayesian framework for joint analysis of marginal SNP effects. Genet. Epidemiol. 40, 188–201 (2016).
17. FitzGerald, L. M. et al. Investigation of the relationship between prostate cancer and MSMB and NCOA4 genetic variants and protein expression. Hum. Mutat. 34, 149–156 (2013).
18. Pomerantz, M. M. et al. Analysis of the 10q11 cancer risk locus implicates MSMB and NCOA4 in human prostate tumorigenesis. PLoS Genet. 6, e1001204 (2010).
19. Whitaker, K. H. et al. The rs10993994 risk allele for prostate cancer results in clinical relevance in microsemionprotein-beta expression in tissue and urine. PLoS ONE 5, e13360 (2010).
20. Davydov, E. V. et al. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput. Biol. 6, e1001025 (2010).
21. Garber, M. et al. Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics 25, i45–i49 (2009).
22. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15, 1034–1050 (2005).
23. Taberlay, P. C., Statham, A. L., Kelly, T. K., Clark, S. J. & Jones, P. A. Reconfiguration of nucleosome-depleted regions at distal regulatory elements accompanies DNA methylation of enhancers and insulators in cancer. Genome Res. 24, 1421–1432 (2014).
24. Mei, S. et al. Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. Nucleic Acids Res. 45, D658–D662 (2017).
25. Koerner, R. Quantile Regression (Cambridge University Press, New York, 2005).
26. Kozumi, H. & Kobayashi, G. Gibbs sampling methods for Bayesian quantile regression. J. Stat. Comput. Simul. 81, 1565–1578 (2011).
27. Edes, R. A. et al. Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. Nat. Genet. 45, 385–391 (2013).
28. Amin Al Olama, A. et al. Multiple novel prostate cancer susceptibility signals identified by fine-mapping of known risk loci among Europeans. Hum. Mol. Genet. 24, 5589–5602 (2015).
29. Takata, R. et al. Genome-wide association study identifies five new susceptibility loci for prostate cancer in the Japanese population. Nat. Genet. 42, 751–754 (2010).
30. Spiska, S. et al. CAUSE: an epigenome- and genome-editing pipeline for establishing functionality of noncoding GWAS variants. Nat. Med. 21, 1357–1363 (2015).
31. Conti, D. V. et al. Two novel susceptibility loci for prostate cancer in men of African ancestry. J. Natl. Cancer Inst. 109, djx084 (2017).
32. Chung, C. C. et al. Fine mapping of a region of chromosome 11q13 reveals multiple independent loci associated with risk of prostate cancer. Hum. Mol. Genet. 20, 2869–2877 (2011).
33. Han, Y. et al. Integration of multiplex fine-mapping and genomic annotation to prioritize candidate functional SNPs at prostate cancer susceptibility regions. Hum. Mol. Genet. 24, 5603–5618 (2015).
34. Kote-Jarai, Z. et al. Identification of a novel prostate cancer susceptibility variant in the KLK3 gene transcript. Hum. Genet. 129, 687–694 (2011).
35. Kote-Jarai, Z. et al. Fine-mapping identifies multiple prostate cancer risk loci at 5p15, one of which associates with TERT expression. Hum. Mol. Genet. 22, 2520–2528 (2013).
36. Bennett, C. et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. Bioinformatics 32, 1493–1501 (2016).
37. Chen, W. et al. Fine mapping causal variants with an approximate Bayesian method using marginal test statistics. Genetica 200, 719–736 (2015).
Statistical Computing (DSC 2003), Vienna, Austria (eds Hornik, K. et al.) 3–10 (2003).

71. Dadacz, T., Leongamornlert, D. A., Saunders, E. I., Elees, R. & Kote-Jarai, Z. LocustExplorer: a user-friendly tool for integrated visualization of human genetic association data and biological annotations. Bioinformatics 32, 949–951 (2016).

Acknowledgements

We would particularly like to thank all the patients and control men who took part in all the studies involved in this work, as well as all the researchers, clinicians, technicians and administrative staff who have enabled this work to be carried out, and the collaborators in the PRACTICAL consortium. We wish to thank all GWAS studies contributing to the meta-analysis data set from which these fine-mapping analyses were conducted: BPC3 (Breast and Prostate Cancer Cohort Consortium); CAPS (Cancer of the Prostate in Sweden); PEGASUS (Prostate Cancer Genome-wide Association Study of Uncommon Susceptibility Loci); The PRACTICAL (Prostate Cancer Association Group to Investigate Cancer-Associated Alterations in the Genome) Consortium; and The GAME-ON/ELLIPSE Consortium. Detailed acknowledgements and funding information for all GWAS study groups and from all the individual studies involved in the PRACTICAL Consortium are included in Supplementary Note 1.

Author contributions

The contributions of each author to this study are as follows. Performance of fine-map analyses: T.D., E.J.S., P.J.N., E.A., D.A.L., M.N.B., C.C.B., M.M., S.Wa., Z.Z., D.V.C. and Z.K.J. Performance of variant annotation: E.J.S., E.A., M.M., C.C.B. and S.W. OncoChip design (PoCA content): T.D., E.J.S., D.A.L., A.A.O., F.R.S., X.S., P.K., F. Wu, Ste., B.E.H., D.F.E., C.A.Hai., R.A.E. and Z.K.J. Provision of DNA samples and/or phenotypic data: T.D., E.J.S., D.A.L., M.N.B., A.L., V.L.S., M.G., B.D., C.M.T., P. G., I.M.T., J.B., S.U., L.M., J.C., L.M., H.T., W.G., H.G., M.A., T.N., P.B., N.S., J.T., L.T., T.S., A.A., D.A., S.W., A.W., N.H., C.W., A.M.D., N.B., L.M., E.G., A.O.C., G. C., S., K., L.B., K.S., T.F.O., M.B., L.M., E.M.G., D.E.N., J.L.D., F.C.H., R.M.M., R. C.T., T.K., J.B., N.E.F., A.F., S.A.L., M.C.S., B.R., S.K., H.O., Y.L.I., H.W.Z., N.F., X. M., X.G., W.Z., G.G., X.G., M.C.S., R.I.M., L.F.K., A.S.K., B.D., A.V., A.G.C., L.F., R. S., M.E., K., J.L.I., G.C.V., K.L.P., M.S., J.T., A.S., H.Y., J.L., L.S., C.C.Y., D.W., L.I., E. A.O., M.S.G., B.G.N., S.F.W., R.B., M.A., P.I., H.K., B.C., H., C.M., M.L., T.S., J. K., C.J.L., E.M.I., M.K.T., P.P., M.C., S.L., N.S., L.Y., C.D., K.R.D.G.M., P.O., A.R., J.L., S.H.T., D.W.L., L.F.K., D.L., M.G., T.K., R.K., N.U., C.S., V.M., M.F., S.S., F.C.L., S.J., T. V., S.A., P.A.T., C.A.H. Y., M.G.D., J.E.C., J.M., M.I.B., G.I., R.H.W., F.M., T.T., Y.A. K., J.X., K.T.K., L.C., H.P., A.M., K., S.N.T., S.K.M., D.I.S., L.S.I., C.T., J.M., D.I.H.E., E.R., A.S., F.G., L.N.K., L.L.M., R.H., M.I.M., P.K., The PRACTICAL Consortium, F.W. and Z.K.J. Data management for meta-analysis: S.B. and S.S. Data QC for meta-analysis: T. D., E.J.S., A.A.O., F.R.S., S.I.B. and T.T. Imputation for meta-analysis: A.A.O., F.R.S., S.I.B. and L.F. Coordination of meta-analysis project: S.B., M.A., M.F., C.E., B.E.H., D.F.E., C.A.Hai., R.A.E. and Z.K.J. Provision of genotype data for individual GWAS sub-studies in meta-analysis: P.K., F.W., Ste.C., B.E.H., D.F.E., C.A.Hai., R.A.E. and Z.K.J. Writing of manuscript: E.J.S., T.D., P.J.N., D.V.C. and Z.K.J.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-04109-8.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

*Open Access* This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a notice to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2018
of California, Fremont, CA 94538, USA. 96)Department of Health Research & Policy (Epidemiology) and Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA 94305-5101, USA. 97)Department of Genetics, Portuguese Oncology Institute of Porto, 4200-072 Porto, Portugal. 98)Biomedical Sciences Institute (ICBAS), University of Porto, 4050-313 Porto, Portugal. 99)Department of Population Sciences, Beckman Research Institute of the City of Hope, Duarte, CA 91010, USA. 100)Ghent University, Faculty of Medicine and Health Sciences, Basic Medical Sciences, B-9000 Gent, Belgium. 101)Department of Radiotherapy, Ghent University Hospital, B-9000 Gent, Belgium. 102)Department of Surgery, Faculty of Medicine, University of Malaya, 50603 Kuala Lumpur, Malaysia. 103)Cancer Research Malaysia (CRM), Outpatient Centre, Subang Jaya Medical Centre, 47500 Subang Jaya, Selangor, Malaysia. 104)Department of Urology, University of Washington, Seattle, WA 98195, USA. 105)Institute of Human Genetics, University Medical Center Hamburg-Eppendorf, D-20246 Hamburg, Germany. 106)Division of Medical Oncology, Urogenital Unit, Department of Oncology at the University Hospital Centre Zagreb, Šalata 2, 10000 Zagreb, Croatia. 107)Department of Urology, University Hospital Center Zagreb, University of Zagreb School of Medicine, Šalata 2, 10000 Zagreb, Croatia. 108)Molecular Medicine Center, Department of Medical Chemistry and Biochemistry, Medical University of Sofia, 1431 Sofia, Bulgaria. 109)Department of Oncology, Cross Cancer Institute, University of Alberta, Edmonton, AB T6G 1Z2, Canada. 110)Division of Radiation Oncology, Cross Cancer Institute, Edmonton, AB T6G 1Z2, Canada. 111)Department of Urology and Alexandria Hospital University Medical School, University of Sofia, 1431 Sofia, Bulgaria. 112)Molecular Endocrinology Laboratory, Department of Cellular and Molecular Medicine, KU Leuven, BE-3000 Leuven, Belgium. 113)Department of Urology, University Hospitals Leuven, BE-3000 Leuven, Belgium. 114)Southampton General Hospital, The University of Southampton, Southampton SO16 6YD, UK. 115)Manchester Cancer Research Centre, Faculty of Biology Medicine & Health, Manchester Academic Health Science Centre, NIHR Manchester Biomedical Research Centre, Health Innovation Manchester, University of Manchester, Manchester M13 9WL, UK. 116)The University of Surrey, Guildford, Surrey GU2 7XH, UK. 117)Genomic Medicine Group, Galician Foundation of Genomic Medicine, Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS), Complejo Hospitalario Universitario de Santiago, Servicio Galego de Saúde, SERGAS, 15706 Santiago de Compostela, Spain. 118)Moores Cancer Center, University of California San Diego, La Jolla, CA 92037, USA. 119)Genetic Oncology Unit, CHUVI Hospital, Complexo Hospitalario Universitario de Vigo, Instituto de Investigación Biomédica Galicia Sur (IBIS), 36204 Vigo (Pontevedra), Spain. 120)Moores Cancer Center, Department of Family Medicine and Public Health, University of California San Diego, La Jolla, CA 92037-0012, USA. 121)Department of Urology, Erasmus University Medical Center, 3015 CE Rotterdam, The Netherlands. 122)Department of Clinical Chemistry, Erasmus University Medical Center, 3015 CE Rotterdam, The Netherlands. 123)Cancer & Environment Group, Center for Research in Epidemiology and Population Health (CESP), INSERM, University Paris-Sud, University Paris-Saclay, 94807 Villejuif Cédex, France. 124)Program for Personalized Cancer Care, NorthShore University HealthSystem, Evanston, IL 60201, USA. 125)Clinical Gerontology Unit, University of Cambridge, Cambridge CB2 2QK, UK. 126)Division of Genetic Epidemiology, Department of Medicine, University of Utah School of Medicine, Salt Lake City, UT 84112, USA. 127)George E. Wahlen Department of Veterans Affairs Medical Center, Salt Lake City, UT 84148, USA. 128)Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN 55905, USA. 129)Division of Biomedical Statistics & Informatics, Mayo Clinic, Rochester, MN 55905, USA. 130)Department of Epidemiology, University of Washington, Seattle, WA 98195, USA. 131)Program in Genetic Epidemiology and Statistical Genetics, Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA. 132)Department of Epidemiology and Biostatistics, School of Public Health, Imperial College, London SW7 2AZ, UK. 133)Genomics England, Queen Mary University of London, Dawson Hall, Charterhouse Square, London EC1M 6BJ, UK. 134)National Cancer Epidemiology Group, German Cancer Research Center (DKFZ), D-69120 Heidelberg, Germany. 135)Epiblogm Program, University of Hawaii Cancer Center, Honolulu, HI 96813, USA. 136)Dana-Farber Cancer Institute, Boston, MA 02215, USA. 137)Royal Marsden NHS Foundation Trust, London SW3 6JJ, UK. These authors contributed equally: Tokhir Dadaev, Edward J. Saunders. These authors jointly supervised this work: Christopher A. Haiman, Rosalind A. Eeles, David V. Conti, Zsofia Kote-Jarai. Deceased: Brian E. Henderson. A full list of consortium members appears at the end of the paper.

The PRACTICAL (Prostate Cancer Association Group to Investigate Cancer-Associated Alterations in the Genome) Consortium

Margaret Cook4, Alison Thwaites1, Michelle Guy1, Ian Whitmore1, Angela Morgan1, Cyril Fisher1, Steve Hazel1, Naomi Livni7, Amanda Spurdle138, Srilakshmi Srinivasan15,16, Mary-Anne Kedda15,16, Joanne Aitken17,18, Robert Gardiner139,140, Vanessa Hayes141, Lisa Butler142, Renea Taylor143, Trina Yeadon15,16, Allison Eckert15,16, Pamela Saunders144, Anne-Maree Haynes20,141, Melissa Papargiris143, Paula Kujala145, Kirsi Talala146, Teemu Murtola31,147, Kimmo Taari148, David Dearnaley137, Gill Barnett35, Søren Bentzen149, Rebecca Elliott34, Hardeep Ranu36, Belynda Hicks150, Aurelie Vogt150, Amy Hutchinson151, Angela Cox152, Michael Davis47, Paul Brown52, Anne George153, Gemma Marsden46,48, Athene Lane47, Sarah J. Lewis47, Clare Berry132, Girish S. Kulkarni52, Ants Toi154, Andrew Evans155, Andrei Zlotter152, Theodorus H. van der Kwast155, Takashi Imai158, Guoping Ren159, Yongwei Yu160, Yudong Wu161, Ji Wu162, Bo Zhou163, Yangling Zhang159, Jie Li60, Weiyang He60, Jianming Guo62, John Pedersen164, John L. Hopper64, Roger Milne63,64, Christa Stegmaier91, Walther Vogt92, Kathleen Herkommer171, Philipp Bohnert93, Sofia Maia97, Maria P. Silva97, Sofie De Langhe100, Hubert Thierens100, Meng H. Tan102, Aik T. Ong102, Zeljko Kastelan107, Elenko Popov111, Julio Pow-Sang170, Hyun Park78, Babu Zachariah78, Wojciech Kluzniak82, Suzanne Kolb80, Peter Klarskov86, Miguel Aguado68, Jose Manuel Ruiz-Dominguez216, Lluís Cecchini72, Esther Gracia-Lavedan72,73,74,75, Trinidad Dierssen-Sotos73,76, Ines Gomez-Acebo73,76, Mariona Bustamante72, Robert Gardiner139,140, Vanessa Hayes141, Lisa Butler142, Renea Taylor143, Trina Yeadon15,16, Allison Eckert15,16, Pamela Saunders144, Anne-Maree Haynes20,141, Melissa Papargiris143, Paula Kujala145, Kirsi Talala146, Teemu Murtola31,147, Kimmo Taari148, David Dearnaley137, Gill Barnett35, Søren Bentzen149, Rebecca Elliott34, Hardeep Ranu36, Belynda Hicks150, Aurelie Vogt150, Amy Hutchinson151, Angela Cox152, Michael Davis47, Paul Brown52, Anne George153, Gemma Marsden46,48, Athene Lane47, Sarah J. Lewis47, Clare Berry132, Girish S. Kulkarni52, Ants Toi154, Andrew Evans155, Andrei Zlotter152, Theodorus H. van der Kwast155, Takashi Imai158, Guoping Ren159, Yongwei Yu160, Yudong Wu161, Ji Wu162, Bo Zhou163, Yangling Zhang159, Jie Li60, Weiyang He60, Jianming Guo62, John Pedersen164, John L. Hopper64, Roger Milne63,64, Christa Stegmaier91, Walther Vogt92, Kathleen Herkommer171, Philipp Bohnert93, Sofia Maia97, Maria P. Silva97, Sofie De Langhe100, Hubert Thierens100, Meng H. Tan102, Aik T. Ong102, Zeljko Kastelan107, Elenko Popov111,
Darina Kachakova, Atanaska Mitkova, Aleksandra Vlahova, Tihomir Dakov, Svetlana Christova, Angel Carracedo, Christopher Bangma, F.H. Schroder, Sylvie Cenee, Brigitte Tretarre, Xavier Rebillard, Claire Mulot, Marie Sanchez, Jan Adolfsson, Par Stattin, Jan-Erik Johansson, Carin Cavalli-Bjerkeman, Ami Karlsson, Michael Broms, Huihai Wu, Lori Tillmans & Shaun Riska

138. Molecular Cancer Epidemiology Laboratory, QIMR Berghofer Institute of Medical Research, Herston, QLD 4006, Australia. 139. School of Medicine, University of Queensland, Herston, QLD 4006, Australia. 140. Royal Brisbane & Women’s Hospital, Herston, QLD 4029, Australia. 141. The Kinghorn Cancer Centre (TKCC), Victoria, NSW 2010, Australia. 142. Prostate Cancer Research Group, South Australian Health & Medical Research Institute, Adelaide, SA 5000, Australia. 143. Department of Physiology, Biomedicine Discovery Institute, Cancer Program, Monash University, Melbourne, VIC 3800, Australia. 144. University of Adelaide, North Terrace, Adelaide, SA 5005, Australia. 145. FinLab Laboratories, Tampere University Hospital, FI-35320 Tampere, Finland. 146. Finnish Cancer Registry, FI-00130 Helsinki, Finland. 147. Faculty of Medicine and Life Sciences, University of Tampere, FI-33014 Tampere, Finland. 148. Department of Urology, Helsinki University Central Hospital and University of Helsinki, FI-00101 Helsinki, Finland. 149. Division of Biostatistics and Bioinformatics, University of Maryland Greenebaum Cancer Center, and Department of Epidemiology and Public Health, University of Maryland School of Medicine, Baltimore, MD 21201, USA. 150. Cancer Genomics Research Laboratory (CGR), Division of Cancer Epidemiology and Genetics, FNLCR Leidos Biomedical Research, National Cancer Institute, Frederick, MD 21701, USA. 151. DNA Extraction and Staging Laboratory (DESL), Cancer Genomics Research Laboratory (CGR), Division of Cancer Epidemiology and Genetics, FNLCR Leidos Biomedical Research, National Cancer Institute, Frederick, MD 21701, USA. 152. Sheffield Institute for Nuclear Acids, University of Sheffield, Sheffield S10 2TN, UK. 153. Cambridge Cancer Trials Centre, Cambridge Clinical Trials Unit - Cancer Theme, Cambridge University Hospitals NHS Foundation Trust, Cambridge CB2 0QQ, UK. 154. Department of Medical Imaging, University Health Network, Toronto, ON M5G 2C4, Canada. 155. Department of Pathology, University Health Network, Toronto, ON M5G 2C4, Canada. 156. Advanced Radiation Biology Research Program, Research Center for Charged Particle Therapy, National Institute of Radiological Sciences, Chiba 263-8555, Japan. 157. Department of Urology, National Hospital Organization Tokyo Medical Center, Tokyo 152-8902, Japan. 158. Department of Urology, Nanfang Hospital, Southern Medical University, 510515 Guangzhou, China. 159. Department of Pathology, The First Affiliated Hospital, Zhejiang University Medical College, 310009 Hangzhou, China. 160. Department of Pathology, Changbai Hospital, The Second Military Medical University, 200433 Shanghai, China. 161. Department of Urology, First Affiliated Hospital, Medical College, Zhengzhou University, 450003 Zhengzhou, China. 162. Department of Urology, North Sichuan Medical College, 637000 Nanchong, China. 163. Department of Nutrition Science, Shenyang Medical College, 110034 Shenyang, China. 164. Tissupath Pty Ltd., Melbourne, VIC 3122, Australia. 165. Department of Medical Physics, Complexa Hospitalario Universitario de Santiago, SERGAS, 15706 Santiago de Compostela, Spain. 166. Urology Department, Hospital Germans Trias i Pujol, 08916 Barcelona, Spain. 167. Laboratory and Department of Urology, Hospital Clínico, Institut d’Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Universitat de Barcelona, 08036 Barcelona, Spain. 168. Centre de Recerca Biomèdica CELEX, 08036 Barcelona, Spain. 169. Department and Laboratory of Urology, Hospital Clínico, Institut d’Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Universitat de Barcelona, 08036 Barcelona, Spain. 170. Genitourinary Program, Moffitt Cancer Center, Tampa, FL 33612, USA. 171. Department of Urology, Klinikum rechts der Isar der Technischen Universität Muenchen, 81675 Munich, Germany. 172. Department of General and Clinical Pathology and Alexandrovska University Hospital, Medical University, 1431 Sofia, Bulgaria. 173. Center of Excellence in Genomic Medicine Research, King Abdulaziz University, Jeddah 22523270, Saudi Arabia. 174. Grupo de Medicina Xenomica, CIBERER, CIBM, Universidad de Santiago de Compostela, Avenida de Barcelona, 15782 Santiago de Compostela, Spain. 175. Paris-Sud University, UMRS 1018, Cedex, 94807 Villejuif, France. 176. Hérault Cancer Registry, Montpellier cedex 5, Montpellier 34298, France. 177. Urology Department, Clinique Beau Soleil, 34070 Montpellier, France. 178. INSERM U1147, 75013 Paris, France. 179. Department of Clinical Science, Intervention and Technology, Karolinska Institute, SE-171 77 Stockholm, Sweden. 180. Swedish Agency for Health Technology Assessment and Assessment of Social Services, SE-102 33 Stockholm, Sweden. 181. Department of Surgical and Perioperative Sciences, Urology and Andrology, Örebro University, SE-701 85 Örebro, Sweden. 182. Department of Urology, Faculy of Medicine and Health, Örebro University, SE-701 82 Örebro, Sweden.