**Online Methods**

Study subjects

A brief overview and study details for participating prostate cancer (PrCa) studies in the newly genotyped OncoArray project are provided in **Supplementary Table 1**1 for men of European ancestry. **Supplementary Table 2** summarizes the PrCa sample series of the Elucidating Loci Involved in Prostate Cancer Susceptibility (ELLIPSE) consortia contributing both newly obtained genotyping data for the OncoArray and previous genome-wide association studies (GWAS). The majority of the studies contributing to the OncoArray were case-control studies primarily based in either the United States or Europe. In total 52 new studies provided core data on disease status, age at diagnosis (age at observation or questionnaire for controls), family history of PrCa, and clinical factors for cases (*e.g.* PSA at diagnosis, Gleason score, etc.) for 48,455 PrCa cases and 28,321 disease-free controls. Previous GWAS contributed an additional 32,255 PrCa cases and 33,202 disease-free controls of European ancestry for the overall meta-analysis1. **Supplementary Table 3** provides quality control information by consortia (*i.e.* OncoArray project, UK GWAS, etc) for both samples and SNPs. After removing all overlapping samples the OncoArray contribution for newly genotyped samples was 46,939 PrCa cases and 27,910 disease-free controls.

Several strata-specific analyses were implemented to evaluate the impact of genetic variation in PrCa disease aggressiveness. **Supplementary Table 4** describes the analysis title, outcome and reference groups, and the statistical model used. Several classification schemes (*i.e.* low aggressiveness, intermediate aggressiveness, etc.) were implemented to better assess the spectrum of genetic involvement. All classification schemes incorporated the diagnostic clinical features PSA, tumor stage and Gleason score. In order to compare to previous PrCa aggressive analyses1 by our research group, we included the ‘Advanced (plus death due to PrCa)’ classification. Contributing study groups missing clinical features were excluded (**Supplementary Table 2**). Individuals with missing or granular clinical information were excluded. The strata-specific sample sizes by PrCa GWAS consortium are provided in **Supplementary Table 5**. Furthermore we analysed Gleason score as a continuous variable.

OncoArray SNP selection

The NCI Genetic Associations and Mechanisms in Oncology (GAME-ON) consortia (<http://epi.grants.cancer.gov/gameon/)> provided SNPs to be included on the Illumina OncoArray. Approximately 50% of the OncoArray was a compilation of SNP lists by the GAME-ON disease consortia of cancer (breast, colorectal, lung, ovarian, and prostate), a common set of variants for common risk regions, other related traits (*i.e.* BMI, age at menarche, etc), pharmacogenetics, and candidates2. The remaining content of the OncoArray was selected as a “GWAS backbone” (Illumina HumanCore), which aimed to provide high coverage for the majority of common variants through imputation. Approximately 79k SNPs were selected specifically for their relevance to PrCa, based on prior evidence of association with overall or subtype-specific disease, fine-mapping of known PrCa regions, and candidate submissions (*i.e.* survival, exome sequencing, etc). In order to maximize efficiency of the array, cancer-specific candidate lists were merged to remove redundant genetc variation2.

Genotype calling and quality control

Details of the genotype calling and quality control (QC) for the iCOGS and GWAS are described elsewhere3-19.

Of the 568,712 variants selected for genotyping on OncoArray, 533,631 were successfully manufactured on the array (including 778 duplicate probes). OncoArray genotyping of ELLIPSE studies was conducted at five sites (Cambridge [UK], CIDR, Copenhagen, USC, NCI). Details of the genotyping calling for the OncoArray are described in more detail elsewhere2. Briefly, we developed a single calling pipeline that was applied to more than 500,000 samples across the GAME-ON consortia. An initial cluster file was generated using 56,284 samples selected from all major genotyping centers and ethnicities, using the Gentrain2 algorithm. Variants likely to have problematic clusters were selected for manual inspection using the following criteria: call rate below 99%, minor allele frequency (MAF) <0.001, poor Illumina intensity and clustering metrics, deviation from the MAF observed in the 1000 Genomes Project (1KGP) using the criterion: , where *p*0 and *p*1 are the minor frequencies in the 1KGP and OncoArray datasets, respectively, and *C*=0.008. This resulted in manual adjustment of the cluster file for 3,964 variants, and the exclusion of 16,526 variants. The final cluster file was then applied to the full dataset.

Our quality control pipeline for ELLIPSE excluded SNPs with a call rate <95% by study, not in Hardy-Weinberg equilibrium (*P*<10-7 in controls, or *P*<10-12 in cases) or with concordance <98% among 11,260 duplicate pairs. In order to minimize imputation errors, we additionally excluded SNPs with a MAF<1% and a call rate <98% in any study, SNPs that could not be linked to the 1KGP reference, those with MAF for Europeans that differed from that for the 1KGP and a further 16,526 SNPs where the cluster plot was judged to be not ideal. Of the 533,631 manufactured SNPs on the OncoArray, we retained 498,417 SNPs among our samples of European ancestry following QC.

We excluded duplicate samples and first-degree relatives within each study, duplicates across studies, samples with a call rate <95%, and samples with extreme heterozygosity (>4.9 standard deviations from the mean for the reported ethnicity). We excluded duplicated samples as well as first-degree relatives across the GWAS studies CAPS1, CAPS2, UK Stage 1, UK Stage 2, and iCOGS. Duplicate and first-degree related samples were assessed across the BPC3 and Pegasus GWAS studies as well. Ancestry was computed using a principal component analysis using 2,318 informative markers on a subset of ~47,000 samples and projected onto the complete OncoArray dataset. The current analysis was restricted to men of European ancestry, defined as individuals with an estimated proportion of European ancestry >0.8, with reference to the HapMap populations, based on the first two principal components. Of the 78,182 samples genotyped (regardless of race/ethnicity), the final dataset consisted of 74,849 samples, of which 46,939 PrCa cases and 27,910 disease-free controls (**Supplementary Table 3**) after excluding overlap samples, were meta-analysed with previous studies.

Imputation

Genotypes for ~70M SNPs were imputed for all samples using the October 2014 (Phase 3) release of the 1KGP data as the reference panel. The OncoArray and GWAS datasets were imputed using a two-stage imputation approach, using SHAPEIT20 for phasing and IMPUTEv221 for imputation. The imputation was performed in 5Mb non-overlapping intervals. All subjects were split into subsets of ~10,000 samples, with subjects from the same group in the subset. We imputed genotypes for all SNPs that were polymorphic (MAF>0.1%) in European samples. We excluded data for all monomorphic SNPs and those with an imputation r2<0.3 leaving a total of 20,370,935 SNP across chromosomes 1-22 and chromosome X. Of the SNPs imputed, 49.3% had a MAF<1%, 15.2% had a MAF ranging between 1-5%, and 35.5% had a MAF≥5%.

Statistical analyses

Per-allele odds ratios and standard errors were generated for the OncoArray and each GWAS, adjusting for principal components 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 principal components in our analysis of individuals of European ancestry, as additional components did not further reduce inflation in the test statistics.

Odds ratio (OR) estimates were derived using either SNPTEST (<https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html>) or an in-house C++ program (**Supplementary Table 3**). OR estimates and standard errors were combined by a fixed effects inverse variance meta-analysis using METAL22. All statistical tests conducted were two-sided.

Definition of new hits

To search for novel loci, we assessed all SNPs excluding those within a known PrCa locus, defined by current fine-mapping assessments (**Supplementary Table 7**). SNPs that were associated with disease risk at *P*<5x10-8 in the meta-analysis (GWAS and OncoArray) were considered novel. The SNP with the lowest p-value in a region was considered the lead SNP. Imputation quality assessed by IMPUTE2 imputation r2 in the OncoArray dataset (**Supplementary Table 8**).

Reliability of Imputation

Novel SNPs with an IMPUTE2 r2<0.80 among the OncoArray sample series (**Supplementary Table 8**) were flagged for further investigation to minimize the probability of a false positive. First, we examined linkage disequilibrium (LD) plots (<http://locuszoom.org/>) for poorly imputed SNPs (+/-500kb) including only genotyped SNPs within the region. The imputed index SNP was included in the plot to determine the strength of LD with nearby signals and assess a pattern of association. Furthermore, we performed an imputation experiment using the 2,504 1KGP Phase 3 samples. We split this sample into two parts: a random sample of 259 individuals of European ancestry (excluding the Finnish) and a mixed-population reference panel of 2,245 individuals. The random sample of 259 individuals of European ancestry was filtered to include only the genetic variants available from the OncoArray following QC. This ensured the same imputation input used in the overall imputation. The 259 individuals were imputed using 2,245 individuals as the reference panel. A 5 MB segment of the genome was selected based on the target SNP (+/- 250 MB). SHAPEIT2 was used for pre-phasing and IMPUTE2 for imputation. Customized imputation settings included an effect size of 20,000, allowance of large region imputation and a random seed of 12345. A weighted linear Kappa statistic was calculated to determine correlation of the imputation with the true genotypes.

We evaluated four SNPs where the IMPUTE2 r2 was less than 0.80 in the OncoArray sample series: rs527510716 (Chr 7), rs6602880 (Chr 10), rs533722308 (Chr 18) and rs144166867 (Chr X). **Supplementary Figure 3** includes the LD plots for three of the poorly imputed SNPs. The variant rs144166867 (Chr X) could not be plotted given no genotype SNPs were available +/-500 KB on the OncoArray. Both LD plots for markers rs527510716 (Chr 7) and rs533722308 (Chr 18) showed significant associations (*P*<1x10-3) for several genotype markers with moderate LD of the index SNP. The Kappa coefficient for markers rs527510716 (Chr 7) and rs533722308 (Chr 18) was 0.911 and 0.931, respectively (**Supplementary Table 9**). The marker rs6602880 (Chr 10) had a Kappa coefficient of 0.812 and was the only significant variant in the LD plot. The Kappa coefficient for marker rs144166867 (Chr X) was 0.665 (**Supplementary Table 9**). The markers rs6602880 (Chr 10) and rs144166867 (Chr X) are most likely false positives due to poor imputation for these regions.

Proportion of familial risk explained

The contribution of the known SNPs to the familial risk of PrCa, under a multiplicative model, was computed using the formula



where is the observed familial risk to first degree relatives of PrCa cases23,24, assumed to be 2.5, and is the familial relative risk due to locus k, given by:



whereis the frequency of the risk allele for locus k,  and is the estimated per-allele odds ratio.

Based on the assumption of a log-additive model, we constructed a polygenic risk score (PRS) from the summed risk allelic dosages weighted by the per-allele log-odds ratios. Thus for each individual *j* we derived:



Where:

: Number of SNPs

: Allele dose at SNPi for individual j

: Per-allele log-odds ratio of SNPi

The risk of PrCa was estimated for the percentile of the distribution of the PRS (<1%, 1-10%, 10-25%, 25-75%, 75-90%, 90-99%, >99% and <10%, 10-25%, 25-75%, 75-90%, >90%) where cumulative score thresholds were determined by the observed distribution among controls. We applied effect sizes and allele frequencies obtained from the overall meta-analysis of Europeans to estimate risk scores for individuals of European ancestry in the OncoArray study25. A standardized PRS score was calculated by dividing the observed PRS score by the standard deviation of the PRS score among controls. A logistic regression framework was used to evaluate the percentile comparisons and determine the risk estimate. The models were adjusted for the first seven principal components to account for population stratification and stratified by country.

The FRR and PRS risk estimation was limited to the variants where our overall meta-analysis observed a statistically significant association. In total, we included 147 PrCa index SNPs in our risk score modelling, including 85 previously published associations and the 62 novel findings reported here. To correct for potential bias in effect estimation of newly discovered variants, we implemented a fully Bayesian version of a weighted correction given in Zhong and Prentice, Eq 3.426. Specifically, we place a normal prior distribution on MLE effect estimates of the form . Here, *βm* is the log odds ratio from the overall meta-analysis; is the bias corrected estimate calculated using the expectation-adjusted estimator from Eq 3.1 in Zhong and Prentice; and τ is a pre-specified variance of the effect distribution reflecting the bias and is defined as .

eQTL analyses

Genotype and gene expression data were downloaded from The Cancer Genome Atlas (TCGA) for 494 samples with PrCa (<https://gdc-portal.nci.nih.gov>). Quality Control (QC) was performed on both these datasets as follows: on the genotype, we filtered out samples with high heterozygosity (mean heterozygosity +/- 2 standard deviation) and missing genotypes, duplicated or related samples. We then performed Principal Component Analysis on the 494 samples plus 2,506 samples from 1KGP to infer the ancestry of the TCGA samples; samples of non-European ancestry were removed. We also filtered out variants with missing call rate > 5%. For the expression data, samples from two plates had, on average, much higher expression values than the remaining samples, and these were excluded. We also filtered genes with mean expression across samples <= 6 counts. Finally, expression values were quantile-normalized by samples and rank-transformed by genes. After QC we used the data from 359 samples. For the eQTL analysis, 35 PEER factors from the top 10,000 expressed genes were used as covariates, plus three genotyping PCs (which explained 18% of total variation). eQTL analysis was performed using FastQTL with 1,000 permutations over the 85 regions. We used a window of 1 Megabases (upstream/downstream) from the transcription start site (TSS) of each gene.

Gene Set Enrichment Analyses

The file Human\_GOBP\_AllPathways\_no\_GO\_iea\_September\_01\_2016\_symbol.gmt (<http://baderlab.org/EM_GeneSets>), from the GeneSets database27, was used for all analyses. This database contains pathways from Reactome28, NCI Pathway Interaction Database29, GO (Gene Ontology) biological process30, HumanCyc31, MSigdb32, NetPath33 and Panther34. We manually corrected several pathways where the *PDPK1* gene was entered as *PDK1.* GO pathways inferred from electronic annotation terms were excluded. The same pathway (e.g. apoptosis) may be defined in two or more databases with potentially different sets of genes, and all versions of these duplicate/overlapping pathways were included. Pathway size was determined by the total number of genes in the pathway to which SNPs in the imputed GWAS dataset could be mapped. To provide more biologically meaningful results, and reduce false positives, only pathways that contained between 10 and 200 geneswere considered.

Gene information (hg19) was downloaded from the ANNOVAR35 website (http://www.openbioinformatics.org/annovar/). SNPs were mapped to the nearest gene within 500kb window; those that were further away from any gene were excluded. Gene significance was calculated by assigning the lowest p-value observed across all SNPs assigned to a gene36,37, based on the combined European meta-analysis (previous GWAS and OncoArray).

The gene set enrichment analysis (GSEA)27 algorithm, as implemented in the GenGen package (http://gengen.openbioinformatics.org/en/latest/)37,38 was used to perform pathway analysis. Briefly, the algorithm calculates an enrichment score (ES) for each pathway based on a weighted Kolmogorov-Smirnov statistic38. To calculate the ES we performed 100 permutations and averaged the final score. Pathways that have most of their genes at the top of the ranked list of genes obtain higher ES values. Only pathways with positive ES and at least one gene with P<5x10-8 were retained for subsequent analysis. An enrichment map was created using the Enrichment Map (EM) v 2.1.0 app27 in Cytoscape v3.4039, applying force directed layout, weighted mode. We restricted our pathway analysis those with an ES≥0.50 to ensure a true positive rate > 0.20 and a false positive rate < 0.15.

**References**

1. Al Olama, A.A. *et al.* A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. *Nat Genet* **46**, 1103-9 (2014).

2. Amos, C.I. *et al.* The OncoArray Consortium: a Network for Understanding the Genetic Architecture of Common Cancers. *Cancer Epidemiol Biomarkers Prev* (in press).

3. Al Olama, A.A. *et al.* Multiple loci on 8q24 associated with prostate cancer susceptibility. *Nat Genet* **41**, 1058-60 (2009).

4. Amundadottir, L.T. *et al.* A common variant associated with prostate cancer in European and African populations. *Nat Genet* **38**, 652-8 (2006).

5. Duggan, D. *et al.* Two genome-wide association studies of aggressive prostate cancer implicate putative prostate tumor suppressor gene DAB2IP. *J Natl Cancer Inst* **99**, 1836-44 (2007).

6. Eeles, R.A. *et al.* Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. *Nat Genet* **41**, 1116-21 (2009).

7. Eeles, R.A. *et al.* Multiple newly identified loci associated with prostate cancer susceptibility. *Nat Genet* **40**, 316-21 (2008).

8. Eeles, R.A. *et al.* Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. *Nat Genet* **45**, 385-91, 391e1-2 (2013).

9. Gudmundsson, J. *et al.* Genome-wide association and replication studies identify four variants associated with prostate cancer susceptibility. *Nat Genet* **41**, 1122-6 (2009).

10. Gudmundsson, J. *et al.* Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat Genet* **39**, 631-7 (2007).

11. Gudmundsson, J. *et al.* Common sequence variants on 2p15 and Xp11.22 confer susceptibility to prostate cancer. *Nat Genet* **40**, 281-3 (2008).

12. Gudmundsson, J. *et al.* Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. *Nat Genet* **39**, 977-83 (2007).

13. 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**, 570-3 (2011).

14. Kote-Jarai, Z. *et al.* Seven prostate cancer susceptibility loci identified by a multi-stage genome-wide association study. *Nat Genet* **43**, 785-91.

15. Schumacher, F.R. *et al.* Genome-wide association study identifies new prostate cancer susceptibility loci. *Hum Mol Genet* **20**, 3867-75.

16. Sun, J. *et al.* Evidence for two independent prostate cancer risk-associated loci in the HNF1B gene at 17q12. *Nat Genet* **40**, 1153-5 (2008).

17. Takata, R. *et al.* Genome-wide association study identifies five new susceptibility loci for prostate cancer in the Japanese population. *Nat Genet* **42**, 751-4 (2010).

18. Thomas, G. *et al.* Multiple loci identified in a genome-wide association study of prostate cancer. *Nat Genet* **40**, 310-5 (2008).

19. Yeager, M. *et al.* Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet* **39**, 645-9 (2007).

20. Delaneau, O., Marchini, J. & Zagury, J.F. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**, 179-81 (2012).

21. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).

22. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-1 (2010).

23. Kicinski, M., Vangronsveld, J. & Nawrot, T.S. An epidemiological reappraisal of the familial aggregation of prostate cancer: a meta-analysis. *PLoS One* **6**, e27130 (2011).

24. Albright, F. *et al.* Prostate cancer risk prediction based on complete prostate cancer family history. *Prostate* **75**, 390-8 (2015).

25. Amin Al Olama, A. *et al.* Risk Analysis of Prostate Cancer in PRACTICAL, a Multinational Consortium, Using 25 Known Prostate Cancer Susceptibility Loci. *Cancer Epidemiol Biomarkers Prev* **24**, 1121-9 (2015).

26. Zhong, H. & Prentice, R.L. Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics* **9**, 621-34 (2008).

27. Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G.D. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS One* **5**, e13984 (2010).

28. Joshi-Tope, G. *et al.* Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res* **33**, D428-32 (2005).

29. Schaefer, C.F. *et al.* PID: the Pathway Interaction Database. *Nucleic Acids Res* **37**, D674-9 (2009).

30. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-9 (2000).

31. Romero, P. *et al.* Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol* **6**, R2 (2005).

32. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-50 (2005).

33. Kandasamy, K. *et al.* NetPath: a public resource of curated signal transduction pathways. *Genome Biol* **11**, R3 (2010).

34. Thomas, P.D. *et al.* PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res* **13**, 2129-41 (2003).

35. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).

36. Wang, L., Jia, P., Wolfinger, R.D., Chen, X. & Zhao, Z. Gene set analysis of genome-wide association studies: methodological issues and perspectives. *Genomics* **98**, 1-8 (2011).

37. Wang, K., Li, M. & Hakonarson, H. Analysing biological pathways in genome-wide association studies. *Nat Rev Genet* **11**, 843-54 (2010).

38. Wang, K., Li, M. & Bucan, M. Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet* **81**, 1278-83 (2007).

39. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-504 (2003).