Prostate cancer meta-analysis of more than 140,000 men identifies 63 novel prostate cancer susceptibility loci

Currently genome-wide association studies (GWAS) and fine-mapping efforts have identified over 100 prostate cancer (PrCa) susceptibility loci, yet over two thirds of the PrCa familial relative risk (FRR) remains unexplained. Therefore, we used a custom high-density array comprised of ~533K SNPs, where ~50% provided a GWAS backbone and the remainder included top association signals from several cancer-specific meta-analyses. We combined newly genotyped data from 46,939 PrCa cases and 27,910 controls of European ancestry with previously genotyped data for 32,255 PrCa cases and 33,202 controls of European ancestry. After excluding previously reported loci, our combined meta-analysis of 79,194 PrCa cases and 61,112 controls identified 62 novel susceptibility loci associated ($P<5.0x10^{-8}$) with overall PrCa, and a locus significantly associated with early-onset PrCa (age at diagnosis ≤ 55 years). Among our novel findings, we identified significant associations with four missense variants including rs1800057 (OR=1.16; $P=8.2x10^{-9}$; G>C [Pro1054Arg]) in *ATM* and rs2066827 (OR=1.06; $P=2.3x10^{-9}$; T>G [Val109Gly]) in *CDKN1B*. The combination of our novel 62 PrCa loci with previously known loci captures 28.4% of the PrCa FRR. Using a polygenic risk score based on previous and newly identified PrCa loci, we estimate elevated PrCa risks for men in the top 90-99%-ile (RR=2.69; 95%CI: 2.55-2.82) and 1%-ile (RR=5.71; 95%CI: 5.04-6.48) risk stratum compared to the population average. These findings demonstrate the utility of high-density arrays, genotyping new sample series, and collaborative efforts for novel discovery. These results will moderately improve risk prediction, enhance fine-mapping efforts, and provide insight into the underlying biology of PrCa susceptibility[1].

Although prostate cancer (PrCa) is the most common non-cutaneous cancer among men in the Western world and 1 in 7 men will be diagnosed during their lifetime[2], very few modifiable risk factors have been established[3]. In addition to age, epidemiological studies have identified a positive family history and race/ethnicity as the most prominent risk factors for PrCa[4-7]. PrCa incidence is highest among men of African ancestry, followed by men of European ancestry and then those of Asian ancestry. These ancestral differences of PrCa risk, in conjunction with studies demonstrating the

influence of family history[8,9], highlight the contribution of genetic risk to PrCa etiology[10]. Our previous work, utilizing a multiplicative model, estimated that over 1,800 common single nucleotide polymorphisms (SNPs), independently contribute to PrCa risk among populations of European ancestry[11]. Genome-wide association studies (GWAS) have reported over 100 of these PrCa variants across multi-ethnic populations, with the vast majority being identified in populations of European ancestry[12-29].

In order to facilitate additional novel discovery of PrCa genetic risk factors we developed a custom high-density genotyping array, the OncoArray, including a 260K SNP backbone designed to adequately tag most common genetic variants (MAF>5% in Europeans), and 310K SNPs from the meta-analysis of five cancers (breast, colorectal, lung, ovarian, and prostate)[30]. Approximately 80,000 PrCa-specific markers derived from our previous multi-ethnic meta-analysis[12] (including populations of European, African American, Japanese, and Latino ancestry), fine-mapping of known PrCa loci, and candidate SNPs nominated by scientific collaborators were included on the OncoArray. We assembled a new PrCa sample series from 52 studies and genotyped them with the OncoArray (**Supplementary Tables 1 & 2**). After applying rigorous quality control criteria and removing overlapping samples from previous genotyping arrays, our OncoArray sample yielded 46,939 PrCa cases and 27,910 controls without a known diagnosis of PrCa of European ancestry for analysis (see **Online Methods, Supplementary Table 3**). Genotypes were phased and imputed to the cosmopolitan panel of the 1000 Genomes Project (1KGP; 2014 June release) using the SHAPEIT[31] and IMPUTEv2[32] software (**Online Methods**, **Supplementary Table 3**). We performed a fixed-effects meta-analysis combining the summary statistics from our new case-control series, the OncoArray, and seven previous PrCa GWAS or high-density SNP panels of European ancestry imputed to 1KGP: BPC3 (2,068 cases/2,993 controls); CAPS 1 (474 cases/482 controls) and CAPS 2 (1,458 cases/512 controls); iCOGS (20,219 cases/ 20,440 controls); NCI PEGASUS (4,600 cases/2,941 controls); UK stage 1 (1,854 cases/1,894 controls); and UK stage 2 (3,650 cases/3,940 controls). The final meta-analysis was based on 79,194 PrCa cases and 61,112 controls without a known diagnosis of PrCa (**Figure 1**).

Study- and consortia-specific meta-analyses were performed to identify novel PrCa loci. We established a *P*-value threshold of 5.0×10^{-8} to determine genome-wide significance.

Our large sample size enabled several stratified meta-analyses focusing on key clinical and biological parameters. These included: 1) high vs low aggressive PrCa; 2) high vs low/intermediate aggressive PrCa; 3) advanced vs non-advanced PrCa; 4) advanced PrCa vs controls; 5) early-onset PrCa (≤55 yrs) vs controls; and 6) Gleason score (**Supplementary Tables 4 & 5**). We defined low aggressive as tumor stage ≤T1 and Gleason ≤6 and PSA <10 ng/mL, intermediate aggressive as tumor stage T2 or Gleason=7 or PSA 10-20 ng/mL, high aggressive as tumor stage T3/T4 or N1 or M1 or Gleason ≥8 or PSA >20 ng/mL, and advanced as either metastatic disease, Gleason ≥8, PSA>100 or PrCa-related deaths (**Supplementary Table 4**). All analyses used a likelihood ratio test to minimize bias from rare variants. A logistic regression framework was used for all analyses, except for the Gleason score stratified analysis where linear regression was utilized. The genotype dosages were incorporated in an allelic genetic model. We report the inflation statistic calibrated to a sample size of 1000 cases and 1000 controls[33], λ_{1000} (**Supplementary Table 6**). The average λ_{1000} across the eight GWAS studies was 1.02 (range: 0.98-1.09) and 1.00 for the overall meta-analysis. Our novel findings excluded variants within previously defined fine-mapped regions of previously reported PrCa loci (**Supplementary Table 7**). For ten regions where the newly identified locus was near a previously known region, we reported a novel association if the pairwise r^2 between the new and the previously known SNP was less than 0.2. For novel PrCa associations where the variant was imputed in the OncoArray study samples series and had an imputed quality score less than 0.70, we assessed the quality of the imputation by masking the variant in a subset of the 1KGP European sample and calculating the concordance following re-imputation in the remaining 1KGP samples (**Online Methods**).

Of the 104 previously reported loci for either overall or stratum-specific (*i.e.* aggressive, early-onset, etc.) PrCa risk, we observed associations at $P<5.0x10^{-8}$ for 84 markers and $P<0.05$ for 97 markers in the overall meta-analysis (**Supplementary Table 7**). The effect sizes with the OncoArray were similar to those previously observed. Of the twenty markers failing to replicate at the genome-wide significance level ($P<5.0x10^{-8}$) nine were initially identified in non-European ancestral populations (8 Asian and 1 African American) and an additional four loci were identified in a multiethnic meta-analysis including European, African, Japanese and Latino populations[12]. An index variant, rs636291, previously reported to be associated with early-onset failed to replicate with

overall PrCa (P=5.87x10^{-7}) or the updated early-onset meta-analysis (P=3.57x10^{-6})[12]. A similar null association was observed for rs1571801, a previously reported aggressive prostate cancer marker (P=3.78x10^{-3})[28]. One previously reported PrCa marker (rs662571 on chromosome X[12]) failed to replicate (*P=0.55*; **Supplementary Table 7**), highlighting the likelihood of a false-positive result. The minor allele frequency (MAF) was initially reported as 41% among Europeans for rs662571 whereas our European OncoArray subset observed a MAF of 18%[12], likely indicating a genotyping array design error. Limiting the association results to the OnocArray only, we replicated 75 of the 104 known PrCa loci (P=0.05/104=4.8x10^{-4}; **Supplementary Table 7**). Of the 29 loci failing to replicate, 16 were discovered in previous GWAS of individuals of European ancestry, 4 were reported in multi-ethnic meta-analyses, and 9 were discovered in populations of non-European ancestry. Fine-mapping is needed to determine if correlated markers within these regions can clarify this apparent inconsistency. These fine-mapping efforts are underway to address this discrepancy.

After the exclusion of all previously identified susceptibility regions (fine-mapping coordinates provided in **Supplementary Table 7**) we identified 65 loci associated with overall PrCa susceptibility and one with early-onset (P<5.0x10^{-8}) in the meta-analysis (**Supplementary Figure 1**). Of the 65 markers, 52 were imputed and 12 were genotyped within the OncoArray study samples. The cluster plots for the genotyped makers are presented in **Supplementary Figure 2**. Although a majority of the imputed markers were of high quality with an average imputed r^2 greater than 0.80 for 61 of the 65 loci across all contributing GWAS (**Supplementary Table 8**), we closely examined four loci with an imputation quality score of r^2 <0.80 within the OncoArray sample series by inspecting linkage disequilibrium (LD) plots including only genotyped SNPs from the OncoArray and performing an imputation quality control assessment (**Online Methods**). After reviewing the LD plots and the imputation QC, we determined loci rs6602880 and rs144166867 are most likely false positives due to imputation artifacts (**Supplementary Figure 3; Supplementary Table 9**). Overall, we identified 62 novel loci associated with overall PrCa risk and one novel loci associated with early-onset (**Table 1**). The consortia specific associations were consistent across the eight contributing GWAS studies (**Supplementary Table 10**) as indicated by the effect estimates and the test for heterogeneity. Although several markers showed evidence for heterogeneity (*P*-

values$_{het}$<0.05), these failed to reach statistical significance after correcting for multiple comparisons (P=0.05/63=7.9x10^{-4}).

We performed several sub-analyses defined by clinical and population parameters to determine novel effects by age and disease aggressiveness. We detected a novel variant, rs138004030, significantly associated with early-onset disease (**Table 1**), but only nominally significant for overall PrCa risk (*P=0.02*). In addition, we detected four markers significantly associated (P<5x10^{-8}) with advanced PrCa and two markers associated with early-onset PrCa (**Supplementary Table 11**). Although significant markers from the advanced PrCa analysis (rs11859370, rs2788524, rs17321482 and rs56366063) were significantly associated with overall PrCa susceptibility, the case-only analyses indicated marginally statistically significant evidence (*P<1.0x10^{-3}*) for advanced disease. Additionally, these markers were in linkage disequilibrium with nearby index markers associated with overall PrCa and were not significantly associated with overall aggressive disease after adjusting for the index marker (**Supplementary Table 11**). A similar association pattern was observed for rs111599055, which was in linkage disequilibrium with marker rs7295014 (r^2=0.54) associated with overall disease. The early-onset marker rs77777548 is independent of novel and known PrCa loci. However, the marker is relatively rare (EAF<0.02), indicated as monomorphic in 1KGP, and has a moderate imputation quality score (average r^2=0.57) hence we did not include it in further analyses.

Among the 63 novel associations, 38 variants are located within gene-rich regions (**Supplementary Table 12**): intronic (32 SNPs), missense (four SNPs), and 3'-UTR (two SNPs). An eQTL analysis of the TCGA database identified statistically significant associations (*P<0.05*; **Supplementary Table 12**) in normal PrCa tissue for 17 of the novel associations, including both 3'UTR SNPs and 11 of the 32 intronic SNPs. *Cis* eQTL associations were identified for 3'UTR variant rs1048169 with *HAUS6* (3'UTR) and intronic variants rs182314334 with *MBNL1*, rs4976790 with *COL23A1*, rs9469899 with *UHRF1BP1*, rs878987 with *B3GAT1*, rs11629412 with *PAX9*, and rs11666569 with *MYO9B*. The eQTL associations are consistent with the observed PrCa-SNP associations, given we assessed colocalization between the GWAS and eQTL SNPs. The TCGA data analysis failed to identify an eQTL association with any of the four missense SNPs.

We assessed the association of our newly discovered loci with prostate-specific antigen (PSA) levels using a series of disease-free controls (N=9,090; see **Online Methods**). Among the 48 available loci we observed a significant association for rs8093601 ($P=5.0x10^{-4}$; **Supplementary Table 13**) after correcting for multiple testing ($P=0.05/48=1.0x10^{-3}$). This marker lies near *MBD2*, methyl-CpG binding domain protein 2, and has not been previously associated with either PrCa risk or PSA levels. The effect sizes by disease aggressiveness or age of onset did not differ with overall PrCa effect estimates and we did not observe any association with Gleason score (**Supplementary Table 14**). Linkage disequilibrium plots incorporating several functional annotation features for each of the 63 novel markers is presented in **Supplementary Figure 4**.

Several strong candidate genes were identified among the PrCa susceptibility loci, including *ATM,* a key gene within the DNA damage response pathway, in which truncating variants are known to contribute towards PrCa susceptibility and progression, particularly aggressive PrCa[34,35]. The index variant within this region is the missense variant rs1800057, exerting a modest increased risk of PrCa (OR=1.16; $P=8.15x10^{-9}$; G>C [Pro1054Arg]; **Figure 2, Panel A**). Although rs1800057 is designated 'benign' by ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/), it was previously suggested to be associated with a two-fold increased risk of early-onset PrCa in a small clinical series and was unassociated with morbidity following treatment[36]. An intronic variant within the region, rs141379009, is also in high LD with rs1800057 (r^2=1.0 in Europeans). Fine-mapping will be needed to comprehensively evaluate the *ATM* region. In addition to the *ATM* region, we identified missense variants in three separate loci: rs2066827 within the cyclin-dependent kinase inhibitor *CDKN1B*, which controls cell cycle progression; rs33984059 within the transcription factor *RFX7*; and rs2277283 within *INCENP*, which encodes a centromere-interacting protein.

rs1048169 at 9p22 is located in the 3'UTR of *HAUS6* (**Figure 2, Panel B**), a gene that encodes a subunit of augmin, a protein complex required for proper microtubule formation and chromosome segregation during cell division[37]. rs1048169 is also an eQTL for *HAUS6* expression. Interestingly, an additional novel lead SNP identified in this study, rs11666569 at 19p13, was also an eQTL for two genes including *HAUS8*, which is

another member of the augmin complex. These discoveries may implicate a potential role for augmin in PrCa susceptibility.

Variant rs7968403 (OR=1.06; $P=3.38x10^{-12}$; **Figure 2, Panel C**) is situated withing the first intron of *RASSF3.* Members of the Ras association domain family (RASSF) are putative tumor suppressors implicated in a range of biological processes[38]. *RASSF3* is ubiquitously expressed across tissue types and has been observed to arrest the cell cycle in the G1 phase and induce apoptosis through the p53 pathway[39]. A previously identified PrCa risk locus, ~100kb away, within the *RASSF6* family member was previously identified[11]. However, rs7968403 was also an eQTL for the more distant *WIF1* (WNT inhibitory factor 1) gene (**Figure 2, Panel C**). *WIF1* inhbits Wnt signaling and is frequently downregulated in prostate cancer[40], whilst aberrant activation of Wnt signaling is a common event in many solid tumor types. Restoration of *WIF1* expression has also been demonstrated to decrease cell motility and invasiveness in a metastatic PrCa cell-line and reduce tumor growth in a mouse xenograft model[41]. Both *RASSF3* and *WIF1* therefore represent plausible mechanisms for the modulation of PrCa risk at this locus.

rs28441558 at 17p13 is the lead variant for a cluster of highly correlated SNPs centered on the *CHD3* gene (**Figure 2, Panel D**). *CHD3* is an ATPase that forms a component of the NuRD (nucleosome remodeling and deacetylase) histone deacetylase complex, which is involved in chromatin remodeling. NuRD plays an important role in regulating gene expression, both as a silencer and an activator of transcription, in addition to maintenance of genomic integrity and the DNA damage response[42]. Alterations to NuRD function have been implicated in several cancer types in a highly complex manner[43,44]. Additionally however, rs28441558 was observed to be an eQTL for three genes; *LOC284023*, a currently uncharacterized non-coding RNA transcript, *GUCY2D*, a guanylate cyclase enzyme expressed predominantly in the retina and *ALOX15B*, a member of the lipoxygenase family of enzymes that produce fatty acid hyperoxides. Whilst *CHD3* therefore appears to represent the most biologically plausible candidate genes for this locus, we cannot exclude a role for any of these genes.

Our pathway analysis based on mapping each SNP to the nearest gene (**see Online Methods**) using the meta-analysis summary association statistic identified several pathways implicated in PrCa susceptibility. The top 53 pathways detected (enrichment

score, ES>0.50) are provided (**Supplementary Table 15**). The most significant pathway detected was PD-1 signaling (ID: 389948), ES=0.74, as defined by the REACTOME database (**Supplementary Figure 5**). This pathway encompasses 22 genes including several with either intronic or closely proximal PrCa susceptibility variants particularly the *HLA* and *CD4* regions. The PD-1 signaling pathway is intriguing given the therapeutic potential of several checkpoint inhibitors focusing on the PD-1 signaling pathway to enhance immune responses[45].

In summary, we have identified 63 novel PrCa susceptibility variants, including strong candidate loci highlighting the DNA repair and cell cycle pathways. Previous studies likely overestimated the effect estimates of PrCa loci due to the "winner's curse", thus yielding a biased FRR and polygenic risk score (PRS). Here, we apply a weighted Bayesian correction approach and demonstrate the impact of our large sample size minimizing the "winner's curse" bias (**Online Methods; Supplementary Figure 6**)[46]. Further support demonstrating the impact of a large sample size can be seen by the λ_{1000} inflation statistics (**Supplementary Table 6**). We applied the beta estimates calculated in our overall meta-analysis to the OncoArray sample set to calculate the FRR and PRS risk models (**Supplementary Table 16**). In total our prediction models included 85 previously reported PrCa loci that replicated in our overall meta-analysis as well as our 62 novel loci associated with overall PrCa risk (excluding the early-onset locus). Assuming a familial risk estimate of 2.5 for PrCa[47,48], we demonstrate our 147 loci captures 28.4% of the FRR (**Supplementary Table 17**). The newly 62 identified PrCa loci increase the FRR by 4.4%. On the assumption of a log-additive model, the estimated RR for PrCa relative to men in the 25-75% PRS percentile (baseline group), was 5.71 (95%CI: 5.04-6.48) for men in the top 1% of the polygenic risk score (PRS) distribution and 2.69 (95%CI: 2.55-2.82) for individuals in the 90-99%-ile of the PRS (**Table 2**). We observed a positive association of our PRS score for overall prostate cancer risk compared to all controls (OR=1.86; 95%CI: 1.83-1.89). When comparing to all controls, the PRS effect estimates increased in strata of positive family history (OR=1.95; 95%CI: 1.85-2.05) and age of onset ≤55 (OR=2.29; 95%CI: 2.20-2.39) compared to negative family history (OR=1.84; 95%CI: 1.80-1.88) and age of onset >55 (OR=1.84; 95%CI: 1.81-1.87) strata, respectively (**Supplementary Table 18**). Our strata specific effects for PRS by disease aggressiveness were similar. This is not surprising given the PrCa loci were identified for overall disease and the case-only analyses were

null (**Supplementary Table 14**). Our novel associations highlight several biological pathways that suggest further investigation is warranted. The increased PRS can be used to improve the identification of men at high risk of PrCa and therefore inform PSA guidelines for screening and management to reduce the burden of over testing.

Data Availability

The OncoArray genotype data and relevant covariate information (*i.e.* ethnicity, country, principal components, etc.) generated during this study are currently being deposited into dbGAP for access (Accession #: phs001391.v1.p1). In total 47 of the 52 OncoArray studies, encompassing nearly 90% of the individual samples, will be available (**Supplementary Table 19**). The previous meta-analysis summary results and genotype data currently[12] are available in dbGAP (Accession #: phs001081.v1.p1). The complete meta-analysis summary associations statistics is publicly available at the PRACTICAL website (http://practical.icr.ac.uk/blog/).

References

1. Goh, C.L. *et al.* Genetic variants associated with predisposition to prostate cancer and potential clinical implications. *J Intern Med* **271**, 353-65 (2012).
2. Siegel, R.L., Miller, K.D. & Jemal, A. Cancer statistics, 2016. *CA Cancer J Clin* **66**, 7-30 (2016).
3. Cuzick, J. *et al.* Prevention and early detection of prostate cancer. *Lancet Oncol* **15**, e484-92 (2014).
4. Altekruse, S.F. *et al.* Spatial patterns of localized-stage prostate cancer incidence among white and black men in the southeastern United States, 1999-2001. *Cancer Epidemiol Biomarkers Prev* **19**, 1460-7 (2010).
5. Stanford, J.L. & Ostrander, E.A. Familial prostate cancer. *Epidemiol Rev* **23**, 19-23 (2001).
6. Bunker, C.H. *et al.* High prevalence of screening-detected prostate cancer among Afro-Caribbeans: the Tobago Prostate Cancer Survey. *Cancer Epidemiol Biomarkers Prev* **11**, 726-9 (2002).
7. Ghadirian, P., Howe, G.R., Hislop, T.G. & Maisonneuve, P. Family history of prostate cancer: a multi-center case-control study in Canada. *Int J Cancer* **70**, 679-81 (1997).
8. Gronberg, H., Damber, L. & Damber, J.E. Familial prostate cancer in Sweden. A nationwide register cohort study. *Cancer* **77**, 138-43 (1996).
9. Matikaine, M.P. *et al.* Relatives of prostate cancer patients have an increased risk of prostate and stomach cancers: a population-based, cancer registry study in Finland. *Cancer Causes Control* **12**, 223-30 (2001).
10. Eeles, R. *et al.* The genetic epidemiology of prostate cancer and its clinical implications. *Nat Rev Urol* **11**, 18-31 (2014).
11. 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).
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**, 1103-9 (2014).
13. Al Olama, A.A. *et al.* Multiple loci on 8q24 associated with prostate cancer susceptibility. *Nat Genet* **41**, 1058-60 (2009).
14. Amundadottir, L.T. *et al.* A common variant associated with prostate cancer in European and African populations. *Nat Genet* **38**, 652-8 (2006).
15. 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).
16. Eeles, R.A. *et al.* Multiple newly identified loci associated with prostate cancer susceptibility. *Nat Genet* **40**, 316-21 (2008).
17. Gudmundsson, J. *et al.* Genome-wide association and replication studies identify four variants associated with prostate cancer susceptibility. *Nat Genet* **41**, 1122-6 (2009).
18. Gudmundsson, J. *et al.* Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat Genet* **39**, 631-7 (2007).
19. Gudmundsson, J. *et al.* Common sequence variants on 2p15 and Xp11.22 confer susceptibility to prostate cancer. *Nat Genet* **40**, 281-3 (2008).

20. 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).
21. 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).
22. Kote-Jarai, Z. *et al.* Seven prostate cancer susceptibility loci identified by a multi-stage genome-wide association study. *Nat Genet* **43**, 785-91.
23. Schumacher, F.R. *et al.* Genome-wide association study identifies new prostate cancer susceptibility loci. *Hum Mol Genet* **20**, 3867-75.
24. 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).
25. 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).
26. Thomas, G. *et al.* Multiple loci identified in a genome-wide association study of prostate cancer. *Nat Genet* **40**, 310-5 (2008).
27. Yeager, M. *et al.* Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet* **39**, 645-9 (2007).
28. 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).
29. Amin Al Olama, A. *et al.* A meta-analysis of genome-wide association studies to identify prostate cancer susceptibility loci associated with aggressive and non-aggressive disease. *Hum Mol Genet* **22**, 408-15 (2013).
30. Amos, C.I. *et al.* The OncoArray Consortium: a Network for Understanding the Genetic Architecture of Common Cancers. *Cancer Epidemiol Biomarkers Prev* (2016).
31. Delaneau, O., Marchini, J. & Zagury, J.F. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**, 179-81 (2012).
32. 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).
33. de Bakker, P.I. *et al.* Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet* **17**, R122-8 (2008).
34. Leongamornlert, D. *et al.* Frequent germline deleterious mutations in DNA repair genes in familial prostate cancer cases are associated with advanced disease. *Br J Cancer* **110**, 1663-72 (2014).
35. Mateo, J. *et al.* DNA-Repair Defects and Olaparib in Metastatic Prostate Cancer. *N Engl J Med* **373**, 1697-708 (2015).
36. Meyer, A. *et al.* ATM missense variant P1054R predisposes to prostate cancer. *Radiother Oncol* **83**, 283-8 (2007).
37. Sanchez-Huertas, C. & Luders, J. The augmin connection in the geometry of microtubule networks. *Curr Biol* **25**, R294-9 (2015).

38. Volodko, N., Gordon, M., Salla, M., Ghazaleh, H.A. & Baksh, S. RASSF tumor suppressor gene family: biological functions and regulation. *FEBS Lett* **588**, 2671-84 (2014).
39. Kudo, T. *et al.* The RASSF3 candidate tumor suppressor induces apoptosis and G1-S cell-cycle arrest via p53. *Cancer Res* **72**, 2901-11 (2012).
40. Wissmann, C. *et al.* WIF1, a component of the Wnt pathway, is down-regulated in prostate, breast, lung, and bladder cancer. *J Pathol* **201**, 204-12 (2003).
41. Yee, D.S. *et al.* The Wnt inhibitory factor 1 restoration in prostate cancer cells was associated with reduced tumor growth, decreased capacity of cell migration and invasion and a reversal of epithelial to mesenchymal transition. *Mol Cancer* **9**, 162 (2010).
42. Allen, H.F., Wade, P.A. & Kutateladze, T.G. The NuRD architecture. *Cell Mol Life Sci* **70**, 3513-24 (2013).
43. Lai, A.Y. & Wade, P.A. Cancer biology and NuRD: a multifaceted chromatin remodelling complex. *Nat Rev Cancer* **11**, 588-96 (2011).
44. Basta, J. & Rauchman, M. The nucleosome remodeling and deacetylase complex in development and disease. *Transl Res* **165**, 36-47 (2015).
45. McDermott, D.F. & Atkins, M.B. PD-1 as a potential target in cancer therapy. *Cancer Med* **2**, 662-73 (2013).
46. Zhong, H. & Prentice, R.L. Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics* **9**, 621-34 (2008).
47. 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).
48. Albright, F. *et al.* Prostate cancer risk prediction based on complete prostate cancer family history. *Prostate* **75**, 390-8 (2015).

Figure Legends

Figure 1. ELLIPSE/PRACTICAL study design of prostate cancer (PrCa) GWAS meta-analysis. The top section describes the PrCa GWAS meta-analysis published in 2014 (AA Olama *et al*, Nature Genetics 2014[12]) where 23 novel variants were identified. The current PrCa GWAS meta-analysis incorporates an additional 46,939 PrCa cases and 27,910 controls independent of the meta-analyses. The current analysis discovered 62 novel variants associated with overall PrCa and 1 novel variant associated with early-onset PrCa.

Figure 2. Locus Explorer plots depicting the statistical association with PrCa and biological context of variants from four of the novel prostate cancer loci identified. For each panel (**a-d**), top panels depict Manhattan plots of variant -log10 *P* values (y-axis), with the index SNP labeled. Variants that were directly genotyped by the OncoArray are represented as triangles and imputed variants are represented as circles. Variants in linkage disequilibrium with the index SNP are denoted by color (red = r^2 >0.8, orange = r^2 0.6-0.8, yellow = r^2 0.4-0.6, green = r^2 0.2-0.4, blue = r^2 ≤0.2). Middle panels depict the relative locations of selected biological annotations; histone marks within 7 cell lines from the ENCODE project; genes for which the index SNP is an eQTL in the TCGA prostate adenocarcinoma dataset; chromatin state annotation by ChromHMM in PrEC cells; conserved elements within the genome and DNAseI hypersensitivity sites in ENCODE prostate cell lines. The lower panel denotes the position of genes within the region, with genes on the positive and negative strands marked in green and purple, respectively. The horizontal axis represents genomic co-ordinates in the hg19 reference genome. **(a)** rs1800057 (chr11:107643000-108644000) - The index variant is a non-synonymous SNP in the *ATM* gene. **(b)** rs1048160 (chr9:18556000-19557000) - The index variant is located within the 3'UTR of the *HAUS6* gene and is an eQTL for *HAUS6*. **(c)** rs7968403 (chr12:64513000-65514000) - The signal is centered on the *RASSF3* gene, with the index variant located within the first intron. This SNP is also situated within a region annotated for multiple regulatory markers and is an eQTL for the more distant *WIF1* gene. **(d)** rs28441558 (chr17:7303000-8304000) - The signal implicates a cluster of highly correlated variants centered upon the *CHD3* gene. The index SNP is also an eQTL for three other more distantly located genes.

Supplementary Figure Legends

Supplementary Figure 1. Manhattan plot of known (blue) and novel (mageneta) prostate cancer loci (N=63) from the meta-analysis of overall prostate cancer risk. Manhattan plot of genotyped and imputed results from the European-ancestry meta-analysis of overall prostate cancer risk. All SNPs within 500 kb of known and novel PrCa index SNPs are highlighted (known=blue; novel=magenta). The red line and blue line represents $P = 5x10^{-8}$ and $P = 1x10^{-5}$. The y-axis has been limited to $P>1x10^{-20}$.

Supplementary Figure 2. Cluster plots for twelve of the novel index SNPs. Panels a-l represent the cluster plots for all cancer samples genotyped using the OncoArray platform in the GAME-ON consortium. Twelve of the 63 novel prostate cancer susceptibility loci were genotyped and represented. The respective colored clusters represent the genotype, i.e. homozygote wild-type, heterozygote, homozygote variant-type. The genotype counts are listed in the upper right corner of each cluster plot.

Supplementary Figure 3. Linkage disequilibrium plots of novel prostate cancer variants with poor imputation quality scores (r^2<0.80). The index signal (purple diamond) is imputed whereas the remaining points are the genotyped markers from the OncoArray. The strength of the linkage disequilibrium (LD) between the imputed index signal and the genotyped markers is represented by the colors. The x-axis represents the genomic position +/-500 KB of the index signal. The y-axis is the –log*P* of the association with prostate cancer risk from the OncoArray analysis only. The bottom panel lists the known genes within the region. The OncoArray for the locus rs144166867 (Chr X) did not contain any genotyped SNPs. **(a)** LD region for rs527510716 (Chr 7); **(b)** LD region for rs6602880 (Chr 10); and **(c)** LD region for rs533722308 (Chr 18).

Supplementary Figure 4. LocusExplorer plots of the 63 novel prostate cancer loci. The index SNPs are labeled on the top portion of the plot. Clusters of correlated variants for each signal are distinguished using different colors in the plot and on the panel below. Stronger shading indicates greater correlation with the index SNP, with variants not correlated at $r^2 \geq 0.5$ with any index SNP uncolored. Log_{10} *P*-values are shown on the *Y*-axis of the plot. The position of genes within the region and the genomic coordinates of the plot are shown on the lower panel, with genes on the positive strand in green and the negative strand in purple.

Supplementary Figure 5. Enrichment map of the PD-1 Signaling Pathway. Shaded circles represent pathways (darker red indicates higher enrichment score) and larger size denotes a greater number of genes in the pathway), as labelled. Edges connect similar pathways, as determined by gene set overlap. Green lines connect those that are most similar in terms of gene set overlap (>70%) thicker lines denote greater similarity).

Supplementary Figure 6. Effect estimation bias of the overall meta-analysis. A correlation plot of the known PrCa (black) and novel PrCa (red) loci where the x-axis is the log odds ratio from the overall meta-analysis and the y-axis is the Bayesian corrected estimate.