An integrative multi-omics analysis to identify candidate DNA methylation biomarkers related to prostate cancer risk

Lang Wu1,91, Yaohua Yang2,91, Xingyi Guo2, Xiao-Ou Shu2, Quiyin Cai2, Xiang Shu2, Bingshan Li3,4, Ran Tao4,5, Chong Wu6, Jason B. Nikas7, Yanfa Sun1,8, Jingjing Zhu1, Monique J. Roobol9, Graham G. Giles10,11, Hermann Brenner12,13,14, Esther M. John15, Judith Clements16,17, Eli Marie Grindedal18, Jong Y. Park19, Janet L. Stanford20,21, Zsofia Kote-Jarai22, Christopher A. Haiman23, Rosalind A. Eeles22, Wei Zheng2, Jirong Long2✉, The PRACTICAL consortium*, CRUK Consortium*, BPC3 Consortium*, CAPS Consortium* & PEGASUS Consortium*

It remains elusive whether some of the associations identified in genome-wide association studies of prostate cancer (PrCa) may be due to regulatory effects of genetic variants on CpG sites, which may further influence expression of PrCa target genes. To search for CpG sites associated with PrCa risk, here we establish genetic models to predict methylation (N = 1,595) and conduct association analyses with PrCa risk (79,194 cases and 61,112 controls). We identify 759 CpG sites showing an association, including 15 located at novel loci. Among those 759 CpG sites, methylation of 42 is associated with expression of 28 adjacent genes. Among 22 genes, 18 show an association with PrCa risk. Overall, 25 CpG sites show consistent association directions for the methylation-gene expression-PrCa pathway. We identify DNA methylation biomarkers associated with PrCa, and our findings suggest that specific CpG sites may influence PrCa via regulating expression of candidate PrCa target genes.

1 Cancer Epidemiology Division, Population Sciences in the Pacific Program, University of Hawaii Cancer Center, University of Hawaii at Manoa, Honolulu, HI, USA. 2 Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University Medical Center, Nashville, TN, USA. 3 Department of Molecular Physiology & Biophysics, Vanderbilt University, Nashville, TN, USA. 4 Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA. 5 Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, USA. 6 Department of Statistics, Florida State University, Tallahassee, FL, USA. 7 Research & Development, Genomix Inc, Minneapolis, MN, USA. 8 College of Life Science, Longyan University, Longyan, Fujian, P. R. China. 9 Department of Urology, Erasmus University Medical Center, Rotterdam, The Netherlands. 10 Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, University of Melbourne, 207 Bouverie St, Melbourne, VIC 3010, Australia. 11 Cancer Epidemiology & Intelligence Division, Cancer Council Victoria, 615 St Kilda Rd, Melbourne, VIC 3004, Australia. 12 Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany. 13 German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany. 14 Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany. 15 Department of Medicine (Oncology) and Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA, USA. 16 Australian Prostate Cancer Research Centre-QLD, Institute of Health and Biomedical Innovation and School of Biomedical Science, Queensland University of Technology, Brisbane, QLD, Australia. 17 Translational Research Institute, Brisbane, QLD, Australia. 18 Department of Medical Genetics, Oslo University Hospital, Oslo, Norway. 19 Department of Cancer Epidemiology, Moffitt Cancer Center, Tampa, FL, USA. 20 Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA. 21 Department of Epidemiology, School of Public Health, University of Washington, Seattle, WA, USA. 22 Division of Genetics and Epidemiology, The Institute of Cancer Research, and The Royal Marsden NHS Foundation Trust, London, UK. 23 Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA. 91 These authors contributed equally: Lang Wu, Yaohua Yang. ✉ Email: lwu@cc.hawaii.edu; jirong.long@vumc.org
Prostate cancer (PrCa) is the second most frequently diagnosed malignancy among men and the fifth leading cause of cancer death worldwide. Its survival rate is relatively high for localized stage disease, but decreases substantially for metastatic disease. Effective strategies are critical for risk assessment, screening, and early detection of PrCa, aimed at decreasing its public health burden. Although prostate-specific antigen (PSA) has demonstrated efficacy for detecting PrCa early, there lacks a clear cutoff point for PSA with high sensitivity and specificity. The benefits of PSA screening for reducing PrCa mortality remains controversial. Furthermore, there are adverse effects, such as overdiagnosis. Therefore, additional effective biomarkers are needed for risk assessment and early detection of PrCa.

Aligned with findings of a crucial role for DNA methylation in PrCa development, research has identified several methylation markers to be potentially associated with PrCa risk, such as methylation at GSTP1, CDKN2A, DNMT3B, SCGB3A1, and HIF3A. However, most prior studies have only assessed a couple of candidates. Recent emerging studies profiling genome-wide methylation usually included a relatively small number of subjects, resulting in inadequate power for the identification of associated methylation biomarkers. Besides these limitations, there are a number of biases commonly encountered in conventional epidemiologic studies, including selection bias, uncontrolled confounding, and reverse causation, that make it difficult to determine whether the identified associated markers are causally associated with PrCa.

One strategy to reduce some of these biases is to use genetic variants to develop an instrument to assess the association between DNA methylation and PrCa. Such an approach is based on the principle of the random assortment of alleles from parents to offspring during gamete formation, and thus a genetically determined proportion of DNA methylation levels should be less susceptible to selection bias and reverse causation in principal. Research has shown that a large portion of CpG sites have high heritability. Genome-wide association studies (GWAS) have also identified a large number of genetic loci associated with DNA methylation levels. Many of these genetic variants could potentially serve as strong instrumental variables for evaluating associations between DNA methylation and PrCa risk in an adequately powered study.

Besides a potential utility in improving PrCa risk assessment, the identification of promising DNA methylation markers using a design of genetic instruments may also contribute to understanding of the genetics and etiology of PrCa. Epidemiological research provides strong support for a genetic predisposition to PrCa. To date, GWAS have identified ~150 genetic loci for PrCa. However, together these variants explain <30% of the familial relative risk, and the underlying biological mechanisms for a majority of the identified loci remain unclear. Recently, we performed a large transcriptome-wide association study (TWAS) of PrCa, in which we identified multiple associations between genetically predicted gene expression and PrCa risk. Interestingly, many of the associated genes were identified to be candidate target genes of GWAS-identified risk SNP. Aligned with the recognized role of DNA methylation in regulating gene expression, we hypothesize that some GWAS-identified risk SNPs may regulate expression of their target genes through influencing DNA methylation levels. In this study, we perform a large integrative multi-omics analysis involving data of genomics, methylyomics, and transcriptomics aiming to uncover novel CpG sites and genes that may contribute to PrCa development.

Results
DNA methylation prediction models. Using FHS data, we were able to build DNA methylation prediction models for 223,959 CpG sites, of which 81,432 showed a prediction performance \( R^2 \) of at least 0.01 (±10% correlation between predicted and measured DNA methylation levels). For 77,243 of those CpG sites, there were no SNPs within the binding site. Interestingly, there tended to be positive weak correlations between methylation prediction model performance and number of input variants within the 2-MB window of each CpG site (Pearson correlation coefficient 0.03, \( P = 1.60 \times 10^{-13} \); Spearman correlation coefficient 0.02, \( P = 1.43 \times 10^{-8} \)). We further applied these 77,243 models to the genetic data in WHI and evaluated their performance by comparing predicted methylation levels with measured levels. Overall, DNA methylation that could be predicted well in FHS also tended to be predicted well in WHI (a correlation coefficient of 0.96 for \( R^2 \) in two datasets; Supplementary Fig. 1). These 77,243 CpG sites were selected for analyses for their associations between predicted DNA methylation and PrCa risk.

Associations of genetically predicted methylation with PrCa. Of the 77,243 CpG sites tested, genetically predicted DNA methylation of 759 located at 82 genomic loci were associated with PrCa risk after Bonferroni correction (\( P < 6.47 \times 10^{-7} \)) (Table 1; Supplementary Table 1 and Supplementary Data 1; Manhattan plot in Fig. 1). This included 15 located at 10 genomic loci that were more than 500 kb away from any PrCa risk variant identified in GWAS or fine-mapping studies (Table 1). An association between a higher DNA methylation level and increased PrCa risk was detected for cg18800143, cg07645299, cg12627844, cg16397176, cg11562153, cg13866093, cg00444740, cg20100049, cg22370235, cg04739953, cg0175842, and cg23397578. Conversely, an inverse association between methylation level and PrCa risk was identified for cg24388424, cg06836406, and cg13230424. Of these 15 CpG sites at novel loci, after conditioning on the near PrCa risk variant, the associations of genetically predicted DNA methylation levels for four CpG sites (cg18800143, cg16397176, cg06836406, and cg13230424) remained at \( P < 6.47 \times 10^{-7} \) (Table 1).

For the remaining 744 CpG sites located at known PrCa risk loci (Supplementary Table 1 and Supplementary Data 1), after conditioning on the adjacent PrCa risk SNP, an association at \( P < 6.47 \times 10^{-7} \) persisted for 63 CpG sites (Supplementary Table 1). This suggests that the associations of these 63 CpG sites with PrCa risk are potentially independent of the PrCa risk SNPs identified in GWAS or fine-mapping studies (Supplementary Table 1). For the other 681 CpG sites, their associations with PrCa risk became weaker, if not completely attenuated, after conditioning on the PrCa risk SNP (Supplementary Data 1). These are potentially due to (1) the previously identified associations of risk SNPs with PrCa at these loci may be mediated through the DNA methylation of these CpG sites identified in the current study, or (2) confounding effects (Supplementary Data 1). We estimated that the 15 CpG sites at novel loci and the 63 CpG sites independent of PrCa risk SNPs could explain 0.69% of familiar risk of PrCa (methods in Supplementary Information).

Based on annotation using ANNOVAR, there were substantial inflations of the “exonic” and “ncRNA exonic” regions for the identified PrCa-associated CpG sites when compared with the overall tested 77,243 CpG sites (chi-square tests: 15.28% versus 7.44%, \( P = 6.36 \times 10^{-10} \); 5.53% versus 2.19%, \( P = 6.37 \times 10^{-8} \) (Supplementary Table 2). Also, a substantial decreased proportion of the “intergenic” region was observed (chi-square test: 15.42% versus 20.10%, \( P = 1.13 \times 10^{-9} \) (Supplementary Table 2).

Through an annotation of the 759 PrCa-associated CpG sites using eFORGE v1.2, there tends to be an overlap of their positions with regions containing lysine 4 mono-methylated H3
Table 1 Fifteen novel methylation-prostate cancer associations for CpG sites located at genomic loci at least 500 kb away from any known prostate cancer risk varianta.

| CpG site | Chr Position (build37) | Classification | OR (95% CI) | P valuec | risk SNP | Distance to the risk SNP (kb) | P value after adjusting for risk SPNd |
|----------|------------------------|----------------|-------------|-----------|----------|-----------------------------|------------------------------------------|
| cg18800143 1 | 16393791 | Intrinsic | 0.10 | 1.12 (1.07–1.17) | 7.56 × 10⁻⁸ | rs636291 | 5837.7 | 7.07 × 10⁻⁹ |
| cg07645299 2 | 63991864 | Intergenic | 0.01 | 1.49 (1.30–1.71) | 1.58 × 10⁻⁸ | rs58235267 | 714.0 | 0.80 |
| cg12627844 2 | 64245000 | Intrinsic | 0.03 | 1.38 (1.28–1.50) | 1.98 × 10⁻¹⁰ | rs58235267 | 967.2 | 0.61 |
| cg16397176 5 | 11089931 | ncRNA_intronic | 0.05 | 1.15 (1.09–1.22) | 6.42 × 10⁻⁷ | rs10793821 | 22936.9 | 6.25 × 10⁻⁷ |
| cg11526153 6 | 28493500 | Upstream | 0.04 | 1.22 (1.13–1.31) | 1.57 × 10⁻⁷ | rs776188 | 1580.3 | 1.56 × 10⁻⁴ |
| cg13666093 6 | 28502727 | UTR3 | 0.05 | 1.14 (1.09–1.20) | 2.09 × 10⁻⁷ | rs776188 | 1571.0 | 3.26 × 10⁻⁵ |
| cg24388424 6 | 28565403 | Intrinsic | 0.01 | 0.78 (0.71–0.86) | 3.31 × 10⁻⁷ | rs776188 | 1508.4 | 1.08 × 10⁻⁵ |
| cg00444740 8 | 129162178 | Upstream | 0.02 | 1.21 (1.13–1.30) | 1.55 × 10⁻⁷ | rs7837688 | 622.8 | 1.01 × 10⁻³ |
| cg06836406 9 | 130461544 | Intergenic | 0.02 | 0.79 (0.72–0.86) | 3.55 × 10⁻⁷ | rs1182 | 2114.5 | 1.74 × 10⁻⁷ |
| cg20100049 11 | 67979188 | Intrinsic | 0.02 | 1.30 (1.22–1.39) | 2.79 × 10⁻⁷ | rs11228565 | 999.4 | 2.44 × 10⁻⁴ |
| cg22370225 11 | 68451852 | Upstream | 0.02 | 1.29 (1.17–1.41) | 1.50 × 10⁻⁷ | rs11228565 | 526.7 | 0.37 |
| cg04739953 11 | 68451852 | Upstream | 0.01 | 1.62 (1.41–1.87) | 2.06 × 10⁻¹¹ | rs11228565 | 526.7 | 0.15 |
| cg07645299 7 | 15587682 | Upstream | 0.47 | 1.05 (1.03–1.07) | 2.95 × 10⁻⁷ | rs199737822 | 2866.7 | NA |
| cg13203424 17 | 45930033 | Intrinsic | 0.05 | 0.87 (0.82–0.91) | 3.16 × 10⁻⁷ | rs138213197 | 875.7 | 5.74 × 10⁻⁸ |
| cg23397578 19 | 37742925 | ncRNA_exonic | 0.01 | 1.40 (1.24–1.57) | 1.81 × 10⁻⁸ | rs8102476 | 992.7 | 1.57 × 10⁻³ |

aNA not available. Bold values represent that these association p values remain largely unchanged after adjusting for risk SNP.
bRisk SNPs identified in previous GWAS or fine-mapping studies.
cOR (odds ratio) and CI (confidence interval) per one standard deviation increase in genetically predicted DNA methylation.
dP value: derived from association analyses of 79,194 cases and 61,112 controls (two-sided); associations with P ≤ 6.47 × 10⁻¹⁰ based on Bonferroni correction of 77,243 tests (0.05/77,243) are shown.
eUsing COO method.

Further evaluated their associations using independent UK Biobank data. In this analysis with far fewer PrCa cases, 554 CpG sites (73%) also showed an association at P < 0.05 with the same direction of effect (Supplementary Data 2). These suggested that the CpG-PrCa risk associations identified in the main analyses using data of the PRACTICAL, CRUK, CAPS, BPC3, and PEGASUS consortia were quite robust. We performed downstream analyses focusing on these 759 CpG sites.

Potential target genes of the PrCa-associated CpG sites. Of the 759 PrCa-associated CpG sites, association analyses were performed for 689 pairs of CpG site-gene, including 613 CpG sites with 244 flanking genes. Overall, associations at a false discovery rate (FDR) < 0.05 were observed for methylation levels of 42 CpG sites with expression of 28 neighbor genes in blood tissue (Supplementary Table 3). Interestingly, we also observed several associations between DNA methylation and expression of genes encoding transcription factors at P < 0.05 (Supplementary Table 4). In the TCGA dataset of tumor-adjacent normal prostate tissue, albeit with a quite limited sample size (n = 34), we observed that 26 of the 37 associations that could be assessed showed the same direction of effect compared with that in the blood tissue (Supplementary Table 5). Among them, 11 showed statistical significance at P < 0.05 in this small dataset (Supplementary Table 5).

Associations of potential target genes with PrCa risk. Of the 28 potential target genes of the identified CpG sites based on blood tissue analyses, blood tissue gene expression prediction models were built for 22 genes, and prostate tissue prediction models were built for 14 genes with a prediction performance (R²) of at least 0.01 (≥10% correlation). Using the S-PrediXcan method, we evaluated associations between the genetically predicted expression of these genes and PrCa risk. Of the 22 genes with blood tissue prediction models built, 18 demonstrated an association at FDR < 0.05 (Table 2). For 12 of them with prostate tissue prediction models built as well, nine showed an association at P < 0.05 (Table 2). For all of the nine genes except for VPS53, the direction of associations was consistent for the predicted expression in blood versus prostate tissue. Of two other genes with models built

![Fig. 1 A Manhattan plot of the association results from the prostate cancer methylene-wide association study using S-PrediXcan. The red line represents P = 6.47 × 10⁻⁷ (Bonferroni correction of 77,243 tests (0.05/77,243)). Each dot represents the genetically predicted DNA methylation of one specific CpG site. The x axis represents the genomic position of the corresponding CpG site, and the y axis represents the negative logarithm of the association P value. CpG sites at novel loci were highlighted with green color. Two-sided test was conducted.](https://example.com/fig1.png)
for prostate tissue only, HLA-DOB showed a significant association with PrCa risk (beta = 0.068, P = 2.65 × 10^{-4}), and C11orf21 did not show a significant association (P = 0.21).

**Associations showing consistent direction of effect.** There were 25 CpG sites and 14 genes with consistent directions of association for the DNA methylation–gene expression–PrCa pathway (Table 3). For example, the CpG site cg20240347 located upstream of MDM4, and its DNA methylation level was positively associated with expression of MDM4 (coefficient 0.21; P = 1.69 × 10^{-14}). There was an inverse association between genetically predicted expression of MDM4 and PrCa risk (OR = 0.36; P = 1.55 × 10^{-19}). There was also evidence supporting the genetically predicted DNA methylation of cg20240347 to be associated with a decreased PrCa risk (OR = 0.93; P = 2.61 × 10^{-19}). Interestingly, MDM4 has been previously implicated as a potential target gene that is responsible for the identified association signal of index SNP rs4245739 in GWAS, and in our recent TWAS study. Our results highlight a possible role of the CpG site cg20240347 in the underlying biological mechanism of the link between MDM4 and PrCa. Whether the DNA methylation of these CpG sites at the corresponding loci of the genes in Table 3 may play a role in PrCa etiology through the regulation of expression of these genes warrants further investigation. Ingenuity pathway analysis (IPA) suggested potential enrichment of cancer-related functions for the 14 implicated genes (Supplementary Table 6). The top canonical pathways identified included cell cycle (P = 0.033) and cancer drug resistance (P = 0.039). It is worth noting that based on the predicted DNA methylation–PrCa risk, DNA methylation–gene expression, and predicted gene expression–PrCa risk results, we also observed six CpG sites and four genes (VAMP8, C4B, BAIAP2L1, and NCOA4) with inconsistent directions of association for the DNA methylation–gene expression–PrCa pathway (Supplementary Table 7). Of these genes, NCOA4, BAIAP2L1, and VAMP8 are candidate PrCa susceptibility genes identified in earlier TWAS and future work is needed to better understand these associations.

**Discussion**

This is the first large-scale study to comprehensively evaluate associations of genetically predicted DNA methylation levels with PrCa risk. We identified 759 CpG sites whose predicted DNA methylation levels demonstrated an association after Bonferroni correction, including 15 located at novel loci. The 744 CpG sites located at known PrCa risk loci, 63 showed an association, even after conditioning on adjacent PrCa risk SNPs. In additional analyses involving gene expression, we observed some evidence suggesting that 25 CpG sites may influence PrCa risk via regulating expression of 14 candidate PrCa target genes. Our study provided substantial information to improve the understanding of genetics and etiology for PrCa, and it also generated multiple CpG sites as potential biomarkers for risk assessment of PrCa, the most common male malignancy globally.

For processing DNA methylation data for genetic model building, we performed quartile normalization for subjects followed by rank normalization for methylation levels, a standard approach widely used in the community for DNA methylation analyses. We acknowledge, however, that such an approach could be suboptimal for CpG sites whose distribution of methylation does not resemble standard normal. Future endeavors for developing more sophisticated methods to deal with this are needed to pick up additional relevant signals. In this study, we identified 759 associated CpG sites, of which 42 were observed to be associated with expression of 28 flanking genes that were annotated by ANNOVAR, based on positions. For the other identified CpG sites, it is possible that genes that are not the most proximal ones could be target genes for local or distal regulation. However, to determine the exact target genes of these CpG sites involves additional lines of evidence besides statistical association, which is beyond the scope of this study. We observed 25 CpG sites with consistent directions of association for the DNA methylation–gene expression–PrCa pathway. Of the 14 linked genes, 10 (MDM4, NUCKS1, PM20D1, VAMP5, GPR160, PDK1, UHRF1BP1, MCAT, LY6G5C, and VPS53) demonstrated an association with PrCa.

### Table 2 Associations between genetically predicted mRNA expression levels of candidate target genes of identified CpG sites and prostate cancer risk.

| Gene         | Blood tissue prediction model | Prostate tissue prediction model |
|--------------|-------------------------------|----------------------------------|
|              | R²a  | OR (95% CI)b  | P valuec  | R²a  | OR (95% CI)b  | P valuec  |
| NCOA4        | 0.14 | 3.80 (2.91-4.96) | 1.39 × 10^{-22} | 0.18 | 1.41 (0.67-2.96) | 0.36 |
| MDM4         | 0.06 | 0.36 (0.29-0.45) | 1.55 × 10^{-19} | NA  | NA  | NA  |
| BAIAP2L1     | 0.03 | 2.21 (1.84-2.67) | 5.86 × 10^{-17} | NA  | NA  | NA  |
| GPR160       | 0.46 | 0.78 (0.73-0.83) | 2.03 × 10^{-16} | NA  | NA  | NA  |
| PDK1         | 0.09 | 1.86 (1.56-2.22) | 8.81 × 10^{-12} | NA  | NA  | NA  |
| TRIM26       | 0.09 | 0.43 (0.34-0.55) | 1.19 × 10^{-11} | 0.03 | 1.97 (0.53-1.78) | 0.93 |
| UHRF1BP1     | 0.40 | 1.11 (1.07-1.15) | 1.99 × 10^{-8}  | 0.21 | 1.18 (1.11-1.25) | 3.24 × 10^{-8} |
| MCAT         | 0.03 | 0.71 (0.62-0.80) | 2.13 × 10^{-8}  | NA  | NA  | NA  |
| NUCKS1       | 0.05 | 3.20 (2.12-4.83) | 2.81 × 10^{-8}  | 0.09 | 1.35 (1.17-1.55) | 3.59 × 10^{-5} |
| C4B          | 0.22 | 0.92 (0.89-0.95) | 3.65 × 10^{-8}  | 0.06 | 0.79 (0.69-0.89) | 2.18 × 10^{-4} |
| PM20D1       | 0.44 | 1.07 (1.04-1.10) | 2.40 × 10^{-7}  | 0.15 | 1.10 (1.06-1.14) | 5.61 × 10^{-7} |
| CFAP44       | 0.04 | 1.25 (1.14-1.36) | 7.44 × 10^{-14} | 0.03 | 1.91 (1.61-2.26) | 9.11 × 10^{-14} |
| LY6G5C       | 0.48 | 1.06 (1.03-1.10) | 9.52 × 10^{-3}  | 0.17 | 1.11 (1.04-1.18) | 1.16 × 10^{-3} |
| MICB         | 0.37 | 0.94 (0.90-0.97) | 8.86 × 10^{-4}  | 0.18 | 0.89 (0.85-0.94) | 3.32 × 10^{-6} |
| VAMP8        | 0.01 | 0.66 (0.51-0.85) | 1.37 × 10^{-3}  | 0.09 | 1.08 (0.99-1.18) | 0.08 |
| ZC3HC7       | 0.10 | 0.80 (0.69-0.92) | 2.52 × 10^{-3}  | 0.15 | 0.83 (0.77-0.89) | 3.78 × 10^{-7} |
| VAMP5        | 0.10 | 1.19 (1.05-1.34) | 5.01 × 10^{-3}  | NA  | NA  | NA  |
| VPS53        | 0.63 | 1.03 (1.01-1.06) | 9.02 × 10^{-3}  | 0.45 | 0.95 (0.92-0.98) | 2.86 × 10^{-3} |

* R²: mRNA expression prediction model performance (R²) derived using GTEx data.
* OR (odds ratio) and CI (confidence interval) per one standard deviation increase in genetically predicted mRNA expression levels.
* P value: derived from association analyses (two-sided); associations of genetically predicted DNA methylation levels with FDR < 0.05 are shown.
* NA: no prostate tissue prediction model was built.
Table 3 Associations showing consistent direction of effect for the methylation–gene expression–prostate cancer risk pathway.

| CpG site      | Chr | Position | Associated gene | Classification | DNA methylation and prostate cancer risk | DNA methylation and gene expression | Gene expression and prostate cancer risk |
|---------------|-----|----------|-----------------|----------------|------------------------------------------|-------------------------------------|------------------------------------------|
| cg02403474    | 1   | 204465584| MDM4            | Upstream       | 0.93, 2.61 × 10^-19                       | 0.21, 1.69 × 10^-14                  | 0.36, 1.55 × 10^-19                     |
| cg15999181    | 1   | 205670604| NUCK3I          | Upstream       | 0.94, 5.10 × 10^-9                        | −0.08, 2.18 × 10^-3                  | 3.20, 2.81 × 10^-8                      |
| cg14857161    | 3   | 205819251| PIWDD1          | UTR5           | 0.97, 1.11 × 10^-7                        | −0.08, 2.70 × 10^-3                  | 1.07, 2.40 × 10^-7                      |
| cg01768772    | 1   | 205819463| PAK4            | Upstream       | 0.97, 1.47 × 10^-7                        | −0.08, 1.83 × 10^-3                  | 0.08, 2.78 × 10^-3                      |
| cg24503407    | 1   | 205819492| PRPSA           | Upstream       | 0.97, 1.27 × 10^-7                        | −0.08, 2.12 × 10^-4                  | 1.07, 2.40 × 10^-7                      |
| cg07357837    | 3   | 205819609| PANK2           | Upstream       | 0.93, 6.31 × 10^-7                        | −0.08, 8.76 × 10^-9                  | 1.07, 2.40 × 10^-7                      |
| cg02655297    | 2   | 85811292 | VAMP5           | Upstream       | 0.93, 6.31 × 10^-7                        | −0.08, 8.76 × 10^-9                  | 1.07, 2.40 × 10^-7                      |
| cg10165864    | 2   | 173419899| PDK1            | Upstream       | 0.89, 6.02 × 10^-14                       | −0.14, 9.34 × 10^-4                  | 1.07, 2.40 × 10^-7                      |
| cg16797009    | 2   | 173427347| Downstream      | 0.90, 2.31 × 10^-10 | 0.14, 3.52 × 10^-10 | 1.07, 2.40 × 10^-7 |
| cg25053018    | 2   | 173477995| Downstream      | 0.93, 4.42 × 10^-10 | 0.14, 3.52 × 10^-10 | 1.07, 2.40 × 10^-7 |
| cg07128416    | 3   | 113160490| CAFAP4          | Upstream       | 1.25, 9.81 × 10^-11                       | 0.09, 6.67 × 10^-4                   | 1.25, 7.44 × 10^-7                      |
| cg07054641    | 3   | 113160554| UHRF1BP1        | Upstream       | 1.22, 6.46 × 10^-11                       | 0.09, 6.47 × 10^-4                   | 1.22, 6.46 × 10^-11                      |
| cg20138861    | 3   | 169755992| GPR160          | Intronic       | 1.17, 3.70 × 10^-14                       | −0.11, 5.97 × 10^-5                  | 0.78, 2.03 × 10^-16                     |
| cg24064041    | 6   | 30165027 | TRIM26          | Intronic       | 0.91, 3.36 × 10^-9                        | 0.13, 8.69 × 10^-7                   | 0.43, 1.19 × 10^-11                     |
| cg00266604    | 6   | 30178334 | MICB            | Intronic       | 1.21, 2.05 × 10^-12                       | −0.10, 3.84 × 10^-4                  | 0.94, 8.86 × 10^-4                      |
| cg20010709    | 6   | 31466798 | LYG6SC          | Intronic       | 0.96, 4.25 × 10^-9                        | −0.12, 4.42 × 10^-6                  | 1.06, 9.52 × 10^-5                      |
| cg13892322    | 6   | 31685664 | LY6G6C          | Upstream       | 0.88, 5.48 × 10^-5                        | −0.08, 2.49 × 10^-3                  | 0.08, 2.49 × 10^-3                      |
| cg22786465    | 6   | 31694902 | Downstream      | 1.23, 7.28 × 10^-10 | 0.08, 1.05 × 10^-5 | 0.08, 2.49 × 10^-3 |
| cg01372338    | 6   | 31695196 | Downstream      | 1.27, 7.26 × 10^-10 | 0.08, 1.05 × 10^-5 | 0.08, 2.49 × 10^-3 |
| cg25769566    | 6   | 31652785 | Downstream      | 1.05, 5.09 × 10^-10 | 0.26, 1.20 × 10^-16 | 0.26, 1.20 × 10^-16 |
| cg24520979    | 6   | 31653162 | Downstream      | 1.15, 6.87 × 10^-10 | 0.10, 2.37 × 10^-4 | 0.10, 2.37 × 10^-4 |
| cg07306190    | 6   | 34760872 | UHRF1BP1        | Intronic       | 0.95, 2.36 × 10^-8                        | −0.33, 2.00 × 10^-16                 | 1.11, 1.99 × 10^-8                      |
| cg01751842    | 16  | 85045600  | ZDHHC7          | Upstream       | 1.05, 2.93 × 10^-6                        | −0.09, 6.69 × 10^-4                  | 0.80, 2.52 × 10^-3                      |
| cg01799818    | 17  | 594735   | VPS33           | Intronic       | 1.10, 7.40 × 10^-19                       | 0.09, 4.81 × 10^-4                   | 1.03, 9.02 × 10^-3                      |
| cg10288850    | 22  | 43595885 | MCAT            | Upstream       | 2.18, 6.23 × 10^-7                        | −0.09, 8.52 × 10^-4                  | 0.71, 2.13 × 10^-8                      |

Due to the relatively small sample size available. In this study, for assessing DNA methylation–gene expression associations to determine potential target genes of identified CpG sites, besides using data from blood tissue (Supplementary Table 3), we also leveraged data from tumor-adjacent normal prostate tissue in TCGA. Despite a small sample size, we observed evidence supporting many of the associations identified using blood tissue data (Supplementary Table 5). For evaluating predicted gene expression–PrCa risk associations, our analyses using prostate tissue gene expression prediction models also support many of the associations identified using blood tissue prediction models (Table 2).

In the current work, a large number of subjects (N = 1595) in the reference FHS dataset was used for the DNA methylation prediction model building. Aligned with the huge sample size for our main association analyses for PrCa risk (79,194 cases and 61,112 controls), our study provides an unparalleled opportunity to detect the DNA methylation–PrCa associations. The use of genetic instruments rendered our study as potentially less susceptible to several limitations commonly encountered in conventional epidemiological studies, such as selection bias and reverse causation. On the other hand, it is worth noting that similar to TWAS, the associations observed in our analyses focusing on CpG sites are also vulnerable to confounding due to pleiotropy and co-localization of genetic signals. For instance, it would be difficult to distinguish a situation in which one causal methylation quantitative trait locus (mQTL) regulates the methylation of two CpG sites from a scenario in which two CpG sites have two causal mQTLs that are in linkage disequilibrium (LD) with each other. Correlated total methylation levels across CpG sites, correlated predicted DNA methylation across CpG sites, as well as shared genetic variants between DNA methylation genetic prediction models and gene expression prediction models, could all lead to spurious associations in our analyses. When faced with two correlated predictors, regularized regression models like elastic net will randomly down weight one of them, which may be the true causal variant. Despite these potential limitations, our study generated a list of promising PrCa-associated CpG sites that warrant further investigation. By integrating the relationship between DNA methylation, gene expression, and prostate cancer risk pathways, we hope to provide valuable insights into the potential mechanisms underlying prostate cancer risk.
expression, and PrCa risk using multi-omics data from different sources, we were able to identify consistent associations of the DNA methylation–gene expression–PrCa risk pathway. This supports a very interesting hypothesis that methylation at selected CpG sites could influence PrCa risk through the regulation of expression of adjacent target genes, which warrants further investigation. The current work generates a list of promising CpG sites showing an association with PrCa, which can be investigated further in future studies that directly measure levels of these CpG sites. Identification of circulating DNA methylation biomarkers could be useful for PrCa risk assessment.

In conclusion, in a large-scale study to evaluate associations between genetically predicted DNA methylation levels and PrCa risk, we identified 759 CpG sites that showed an association, including 15 at novel loci, and an additional 63 that represent association signals independent of known risk variants. We also observed that specific CpG sites may influence PrCa risk via regulating expression of candidate PrCa target genes. Further investigation of these findings will provide additional insight into the biology and genetics of PrCa, as well as facilitate risk assessment of PrCa.

Methods

Study design. The overall study design is shown in Fig. 2. First, we built comprehensive genetic prediction models for DNA methylation levels by using data of the Framingham Heart Study (FHS). After external validation, we selected methylation models with satisfactory prediction performance for association analyses of genetically predicted methylation levels with PrCa risk, by using data of the PRACTICAL consortia which involves 79,194 cases and 61,112 controls. For CpG sites showing an association with PrCa risk, we assessed associations of their methylation with expression of adjacent genes (FHS, N = 1367), to identify potential target genes of these CpG sites. For the suggested candidate target genes, we further assessed associations of their genetically predicted expression with PrCa risk.

Building of DNA methylation prediction models. We obtained the individual level genome-wide genotyping and white blood cell DNA methylation data from the FHS Offspring Cohort (dbGaP accession numbers: phs000342 and phs000724). The details of the FHS Offspring Cohort have been described elsewhere. In brief, DNA was genotyped using the Affymetrix 500 K array, and DNA methylation was profiled using the Illumina HumanMethylation450 BeadChip. The genotype data were imputed to the HaploTyper Reference Consortium reference panel. SNPs with high imputation quality (R² > 0.8), minor allele frequency ≥ 0.05, included in the HapMap Phase 2 version, and those that were not strand ambiguous were used to build DNA methylation prediction models. For DNA methylation data, the “minfi” package was used to filter out low-quality samples, exclude low-quality methylation probes, estimate cell-type composition, and calculate methylation beta values. We performed quantile normalization to bring the methylation profile of each sample to the same scale, and rank normalization for each CpG site to map expression-PrCa risk pathway.

Genetic prediction models

Technical validation. To further evaluate the validity of the built methylation prediction models, we performed technical validation using data from 883 unrelated healthy female participants of European descent included in The Women’s Health Initiative (WHI) (dbGaP accession numbers: phs000315, phs000675, and phs001335). Genotype data and white blood cell DNA methylation data were processed using a similar approach, as described above. The predicted DNA methylation for each CpG site was calculated using the models that were established using FHS data, and then compared with the measured level using Spearman’s correlation.

Associations between predicted methylation and PrCa. Considering that our model external validation dataset WHI included females only, and that there is a high concordance of the model performance (R²) in FHS and WHI, we included DNA methylation prediction models (1) with a R² ≥ 0.01 (<10% correlation between predicted and measured methylation levels) in FHS, a standard criterion used in TWAS for gene expression, and (2) for probes with no SNPs within the probe-binding site, considering that the measurement of DNA methylation levels for such probes tends to be unbiased. Overall, we evaluated associations between genetically predicted methylation levels of 77,243 CpG sites with PrCa risk.
We estimated the association between genetically predicted DNA methylation levels and PrCa risk using S-PrediXcan, which has been described elsewhere57 (Supplementary Software 1). We used the summary statistics data for the association of genetic variants with PrCa risk that had been generated from 79,194 PrCa cases and 61,112 controls of European ancestry in the PRACTICAL, CRUK, CAPS, BPC3, and PEGASUS consortia26–28. In brief, 46,939 PrCa cases and 27,910 controls were genotyped using OncoArray, which included 570,000 SNPs (http://epigaphs.meta-analysis.org/). Also included were data from several previous PrCa GWAS of European ancestry: UK stage 1 and stage 2, CaPS 1 and CaPS 2, BPC3, NCI PEGASUS, and iCOGS. These genotype data were imputed using the June 2014 release of the 1000 Genomes Project data as reference. Logistic regression summary statistics were then meta-analyzed using an inverse variance fixed effect approach.

A Bonferroni-corrected threshold of \( P < 6.47 \times 10^{-7} \) (0.05/77,243) was used to determine a statistically significant association. For CpG sites showing a significant association between genetically predicted methylation levels with PrCa risk, we further evaluated whether the observed associations were independent of nearby PrCa risk variants using the genetic data from the controls included in the pancreatic cancer GWAS data \( (N = 3655) \) to generate the predicted methylation of these CpGs for each of the participants. Detailed information for this dataset, quality control, and imputation has been described elsewhere35. We further used the formula \( Z^2 = 1 + (N_1 + N_2)/\bar{N}^2 \) to estimate the heritability explained by these CpG sites. Here for each CpG, \( Z \) represents the Z score of the association between the predicted methylation and PrCa risk; \( N_1 \) and \( N_2 \) represent the number of individuals in the cases and the controls, respectively; \( \bar{N} \) represents the mean of \( N_1 \) and \( N_2 \). The estimated heritability of PrCa risk that could be explained by the predicted methylation of the CpGs of interest, along with the standard error and \( P \) value, were estimated. Given that the heritability of PrCa was estimated to be 57%36, the familial relative risk of PrCa that could be explained by predicted methylation levels of these CpGs was calculated as \( \sqrt{\text{heritability}}/\sqrt{0.57} \).

Validation of identified CpG sites using the UK Biobank. Individual level data of the UK Biobank were used to validate the identified associated CpG sites. The UK Biobank released GWAS data on ~500,000 individuals. PrCa cases were determined by combining Hospital Episode Statistics (HES) data and self-reported data. Specifically, hospital admission, type of cancer, or cause of death due to ICD-9 185.9 or ICD-10 C61 or a self-reported cancer code. We calculated associations between genetically predicted DNA methylation of the identified CpG sites with PrCa risk, adjusting for age, sex, and top 20 PCs provided by the UK Biobank. As the number of cases in the UK Biobank is substantially smaller than that in the PRACTICAL, CRUK, CAPS, BPC3, and PEGASUS consortia, we used results from the UK Biobank to confirm the validity of the CpG sites identified in analyses of the consortia data, instead of using their results to filter out CpG sites.

Functional annotation of PrCa-associated CpG sites. We annotated the position and genomic region information of the identified PrCa-associated CpG sites through ANNOVAR54. The CpG sites were annotated into one of 13 functional categories, including exonic, intronic, intergenic, upstream, 3′-UTR, 5′-UTR, ncRNA exon, ncRNA intron, splicing, downstream, ncRNA upstream/downstream, 5′-UTR/3′-UTR, and exonic/splicing. We used eFORGE55 v1.2 to assess whether the identified CpG sites were enriched in DNase I hypersensitive sites (DHSs) and loci overlapping with various histone modification types, such as H3K27me3, H3K36me3, H3K4me3, H3K9me3, and H3K4me1, across different tissues and cell lines available in the Roadmap Epigenomics Project56, the Encyclopedia of DNA Elements (ENCODE)57 and the BLUPRINT Epigenome58. For each CpG site set of interest, eFORGE performs an overlap analysis against the functional elements for each tissue or cell line separately, and then counts the number of overlaps. A background distribution of the expected overlap counts for the CpG site set of interest was obtained by shuffling sets of CpG sites with the same number as the test set, matched for gene relationship and CpG island relationship annotation. The matched background sets are then overlapped with the functional elements and the background distribution of overlaps are determined. 1000 matched sets are used. The enrichment value for the test set is expressed as the \(-\log_{10}(\text{binomial } P \text{ value})\) . Enrichments outside the nominal 95th and 99th percentile of the binomial distribution (after Benjamini–Yekutieli multiple testing correction) are considered significant. We also evaluated whether the associated CpG sites were enriched in loci of genes encoding transcription factors39.

Determine genes associated with identified CpG sites. For CpG sites with genetically predicted DNA methylation levels significantly associated with PrCa risk, we evaluated associations between methylation and expression levels of genes flanking their loci by using data from the FHS Offspring Cohort (dbGaP accession numbers: phs000363 and phs000724) and The Cancer Genome Atlas (TCGA). Details of the FHS Offspring Cohort, DNA methylation, and gene expression data have been described elsewhere26,61. Overall, DNA methylation and gene expression data were available for 1367 unrelated individuals. For the CpG sites showing a significant association with PrCa risk, associations between the normalized methylation levels in beta values and normalized expression levels of genes flanking the CpG sites were estimated, after adjusting for age, sex, top PCs, and estimated cell-type compositions based on methylation data. We further assessed significant methylation–gene expression associations identified in blood tissue analyses in adjacent normal prostate tissue of PrCa patients in the TCGA (\( N = 34 \)). The processing of DNA methylation and gene expression data has been described elsewhere26,63.

Associations of potential target genes with PrCa risk. For genes whose expression levels were associated with DNA methylation levels, we assessed whether the genetically predicted expression levels of these genes in blood and prostate tissue were also associated with PrCa risk44,45. We used prediction models developed using the PrediXcan method (Elastic Net) and leveraging data from the v8 version of the Genotype-Tissue Expression dataset (GTEx) project (http://gtexportal.org/). Details of the methods of building gene expression prediction models using SNPs have been described elsewhere60,62,63. The prediction models were used to estimate the associations between genetically predicted gene expression levels and PrCa risk in the PRACTICAL, CRUK, CAPS, BPC3, and PEGASUS consortia using S-PrediXcan57.

Associations showing a consistent direction of effect. We assessed the associations between genetically predicted DNA methylation levels and PrCa risk, associations between DNA methylation and gene expression levels, and the associations between genetically predicted gene expression and PrCa risk to assess associations showing consistent direction of effect for the DNA methylation–gene expression–PrCa risk pathway. This could indicate the possibility that genetically predicted DNA methylation might putatively influence PrCa risk through the regulation of expression of flanking target genes.

Functional enrichment analysis. We performed functional enrichment analysis for the identified genes consistent with the DNA methylation–gene expression–PrCa risk pathway. Canonical pathways, top associated diseases and biofunctions, and top networks associated with these genes were estimated using IPA software64.

Data availability

The OncoArray genotype data and relevant covariate information (i.e., ethnicity, country, principal components, etc.) for prostate cancer study are available in dbGaP (Accession no.: phs001391.v1.p1). In total, 47 of the 52 OncoArray studies, encompassing ~90% of the individual samples, are available. The previous meta-analysis summary results and genotype data are currently available in dbGaP (Accession no: phs001081.v1.p1). The datasets of FHS Offspring Cohort and WHI are publicly available via dbGaP (www.ncbi.nlm.nih.gov/gap): dbGaP Study Accession: phs000342 and phs000724 for FHS, and phs000135, phs000675, and phs001335 for WHI. TCGA data can be accessed through the Genomic Data Commons Data Portal.

Code availability

The relevant codes are available in the Supplementary Software 1.

Received: 11 December 2019; Accepted: 28 June 2020; Published online: 06 August 2020

References

1. Torre, L. A. et al. Global cancer statistics, 2012. CA: Cancer J. Clin. 65, 87–108 (2015).
1. Massie, C. E., Mills, I. G. & Lynch, A. G. The importance of DNA methylation in prostate cancer. *J. Natl. Cancer Inst.* 101, 374–389 (2013).

2. Massie, C. E., Mills, I. G. & Lynch, A. G. The importance of DNA methylation in prostate cancer development. *J. Steroid Biochem. Mol. Biol.* 166, 1–15 (2017).

3. Catalona, W. J., Smith, D. S., Ratliff, T. L. & Basler, J. W. Detection of organ-confined prostate cancer is increased through prostate-specific antigen-based screening. *J. Am. Med. Assoc.* 270, 948–954 (1993).

4. Schroder, F. H. et al. Screening and prostate-cancer mortality in a randomized European study. *N. Engl. J. Med.* 360, 1320–1328 (2009).

5. Andrieu, G. L. et al. Mortality results from a randomized prostate-cancer screening trial. *N. Engl. J. Med.* 360, 1310–1319 (2009).

6. Draisma, G. et al. Lead time and overdiagnosis in prostate-specific antigen-based prostate cancer detection in a longitudinal screening study. *J. Urol.* 178, 90–93 (2004).

7. Thompson, I. M. et al. Operating characteristics of prostate-specific antigen in men with an initial PSA level of 3.0 mg/ml or lower. *J. Am. Med. Assoc.* 294, 66–70 (2005).

8. Parizkova, D., Ankerst, D. P., Troyer, D., Srivastava, S. & Thompson, I. M. Biomarkers for prostate cancer detection. *J. Urol.* 178, 2252–2259 (2007).

9. Bell, J. T. et al. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol.* 21, 1017–1027 (2010).

10. FitzGerald, L. M. et al. Genome-wide measures of peripheral blood DNA methylation profiles identify 63 new prostate cancer susceptibility loci. *Nat. Genet.* 50, 959–968 (2018).

11. Kannel, W. B., Feinleib, M., McNamara, P. M., Garrison, R. J. & Castelli, W. P. An investigation of coronary heart disease in families: the Framingham Offspring Study. *Am. J. Epidemiol.* 110, 281–290 (1979).

12. Wu, L. et al. A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. *Nat. Genet.* 50, 968–978 (2018).

13. Yang, Y. et al. Genetically predicted levels of DNA methylation biomarkers and breast cancer risk: data from 228 951 women of European descent. *J. Natl Cancer Inst.* 112, 295–304 (2020).

14. Hannon, E., Weedon, M., Bray, N., O’Donovan, M. & Mortensen, J. H. Genetic associations of DNA methylation with genetic ancestry and breast cancer risk: data from 228,951 women of European descent. *Hum. Genet.* 130, 1373–1387 (2017).

15. Kramer, A., Green, J., Pollard, J. & Tugendreich, S. Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics* 30, 523–530 (2014).

16. Schumacher, F. R. et al. Association analyses of more than 140,000 men identify 83 new prostate cancer susceptibility loci. *Nat. Genet.* 50, 911–926 (2018).

17. Wu, L. et al. Identification of novel susceptibility loci and genes for prostate cancer risk: a transcriptome-wide association study in over 140,000 European descendants. *Cancer Res.* 79, 3192–3204 (2019).

18. Kramer, A., Green, J., Pollard, J. & Tugendreich, S. Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics* 30, 523–530 (2014).

19. Emami, N. C. et al. Association of imputed prostate cancer transcriptome with disease risk reveals novel mechanisms. *Nat. Commun.* 10, 3107 (2019).

20. Mancuso, N. et al. Large-scale transcriptome-wide association study identifies new prostate cancer risk regions. *Nat. Commun.* 9, 4079 (2018).

21. Hukriede, A. N., Thibodeau, S. N. et al. Identification of candidate genes for prostate cancer risk SNPs utilizing a normal prostate tissue eQTL data set. *Nat. Commun.* 6, 6852 (2015).

22. Li, W. et al. CD44 regulates prostate cancer proliferation, invasion and migration via PDK1 and PFKFB4. *Oncotarget* 8, 65143–65151 (2017).

23. Wu, L. et al. A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. *Nat. Genet.* 50, 968–978 (2018).

24. Wainberg, M. et al. Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.* 51, 592–599 (2019).

25. Adams, D. et al. BLUEPRINT to decode the epigenetic signature written in human genomes. *Nature* 571, 530–535 (2019).

26. Nikas, J. B., Mitanis, N. T. & Nikas, E. G. Whole exome and transcriptome sequencing model for the diagnosis of prostate cancer. *ACS Omega* 5, 481–486 (2020).

27. Wu, L. et al. A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. *Nat. Genet.* 50, 968–978 (2018).

28. Yang, Y. et al. Genetically predicted levels of DNA methylation biomarkers and breast cancer risk: data from 228 951 women of European descent. *J. Natl Cancer Inst.* 112, 295–304 (2020).

29. Wu, L. et al. A transcriptome-wide association study of 97,898 women identifies new candidate susceptibility genes for epithelial ovarian cancer risk. *Cancer Res.* 78, 5419–5430 (2018).

30. Wu, L. et al. Analysis of Over 140,000 European descendants identifies genetically predicted blood protein biomarkers associated with prostate cancer risk. *Cancer Res.* 79, 4592–4598 (2019).

31. Yang, J. et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* 44, 369–375 (2012).
The PRACTICAL consortium

Rosalind A. Eeles22, Brian E. Henderson23,92, Christopher A. Haiman23, Zsofia Kote-Jarai22, Fredrick R. Schumacher24,25, Douglas Easton26, Sara Benlloch22,26, Ali Amin Al Olama26,27, Kenneth Muir28,29, Sonja I. Berndt30, David V. Conti23, Fredrik Wiklund31, Stephen Chanock30, Susan M. Gapstur32, Victoria L. Stevens32, Catherine M. Tangen33, Jytotsna Batra16,34, Judith Clements16,17, Henrik Gronberg31, Nora Pashayan35,36, Johanna Schleutker37,38, Demetrius Albanes30, Stephanie Weinstein30, Alicja Wolk39,40, Catharine West41, Lorelei Mucci42, Géraldine Cancel-Tassin43,44, Stella Koutros30, Karina Dalsgaard Sorensen45,46, Eli Marie Grindedal18, David E. Neal47,48, Freddie C. Hamdy49, Jenny L. Donovan50, Ruth C. Travis51, Robert J. Hamilton52, Sue Ann Ingles23, Barry S. Rosenstein53,54, Monique J. Roobol9, Kim De Ruyck71, Azad Razack72, Lisa F. Newcomb20,73, Marija Gamulin74, Kathryn L. Penney62, Jong Y. Park19, Janet L. Stanford20,21, Cezary Cybulski63, Børge G. Nordestgaard64,65, Hermann Brenner12,13,14, Christiane Maier66, Jeri Kim67, Esther M. John15, Manuel R. Teixeira68,69, Susan L. Neuhausen70, Yong-Jie Lu55, Graham G. Giles10,11, Adam S. Kibel56, Ana Vega57, Manolis Kogevinas58,59,60,61, Nan Li60,61, Manik Rag substrates, and co-workers. The original studies. All authors have reviewed and approved the final manuscript.

Author contributions

J.L. and W.Z. conceived the study. L.W. and Y.Y. contributed to the study design. L.W. performed statistical analyses and wrote the paper, with significant contributions from Y.Y. and J.L. X.G. contributed to study discussion, C.W., I.B.N., Y.S., and I.Z. contributed to statistical analyses. X.-O.S., Q.C., X.S., B.L., R.T., M.J.R., G.G.G., H.B., E.M.J., J.C., E.M.G., J.Y.P., J.L.S., Z.K.-J., C.A.H., R.A.E., and W.Z. contributed to paper revision and/or PRACTICAL data management. The PRACTICAL, CRUK, BPC3, CAPS, and PEGASUS consortia are included in the Supplementary Note. Lang Wu is supported by the University of Hawaii Cancer Center Seed Grant. Yanfa Sun is partially supported by the Department of Education of Fujian Province, P R China.

Competing interests

R.A.E. has received speakers bureau honoraria and has provided expert testimony for GU-ASCO, BMH FR MTG, and the University of Chicago. The remaining authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-17673-9.

Correspondence

and requests for materials should be addressed to L.W. or J.L.

Peer review information

Nature Communications thanks Francesca Demichelis and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information

is available at http://www.nature.com/reprints

Publisher’s note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
CRUK Consortium

Rosalind A. Eeles22, Zsofia Kote-Jarai22, Catharine West41, David E. Neal47,48, Freddie C. Hamdy49, Jenny L. Donovan50, Ruth C. Travis51 & Elio Riboli90

BPC3 Consortium

Brian E. Henderson23,92, Christopher A. Haiman23, Fredrick R. Schumacher24,25, Sonja I. Berndt30, Stephen Chanock30, Susan M. Gapstur32, Victoria L. Stevens32, Demetrius Albanes30, Stephanie Weinstein30, Lorelei Mucci42, Stella Koutros30, Ruth C. Travis51, Kathryn L. Penney62, David J. Hunter88 & Elio Riboli90
CAPS Consortium
Fredrik Wiklund\textsuperscript{31} & Henrik Gronberg\textsuperscript{31}

PEGASUS Consortium
Sonja I. Berndt\textsuperscript{30}, Stephen Chanock\textsuperscript{30}, Demetrius Albanes\textsuperscript{30}, Stephanie Weinstein\textsuperscript{30} & Stella Koutros\textsuperscript{30}