

OPEN

# An exome-wide rare variant analysis of Korean men identifies three novel genes predisposing to prostate cancer

Jong Jin Oh<sup>1</sup>, Manu Shivakumar<sup>2</sup>, Jason Miller<sup>3</sup>, Shefali Verma<sup>3</sup>, Hakmin Lee<sup>1</sup>, Sung Kyu Hong<sup>1</sup>, Sang Eun Lee<sup>1</sup>, Younghee Lee<sup>4</sup>, Soo Ji Lee<sup>4</sup>, Jooheon Sung<sup>5</sup>, Dokyoon Kim<sup>2,6\*</sup> & Seok-Soo Byun<sup>1\*</sup>

Since prostate cancer is highly heritable, common variants associated with prostate cancer have been studied in various populations, including those in Korea. However, rare and low-frequency variants have a significant influence on the heritability of the disease. The contributions of rare variants to prostate cancer susceptibility have not yet been systematically evaluated in a Korean population. In this work, we present a large-scale exome-wide rare variant analysis of 7,258 individuals (985 cases with prostate cancer and 6,273 controls). In total, 19 rare variant loci spanning 7 genes contributed to an association with prostate cancer susceptibility. In addition to replicating previously known susceptibility genes (e.g., *CDYL2*, *MST1R*, *GPER1*, and *PARD3B*), 3 novel genes were identified ( $FDR\ q < 0.05$ ), including the non-coding RNAs *ENTPD3-AS1*, *LOC102724438*, and protein-coding gene *SPATA3*. Additionally, 6 pathways were identified based on identified variants and genes, including estrogen signaling pathway, signaling by *MST1*, *IL-15* production, *MSP-RON* signaling pathway, and *IL-12* signaling and production in macrophages, which are known to be associated with prostate cancer. In summary, we report novel genes and rare variants that potentially play a role in prostate cancer susceptibility in the Korean population. These observations demonstrated a path towards one of the fundamental goals of precision medicine, which is to identify biomarkers for a subset of the population with a greater risk of disease than others.

Prostate cancer is a common malignancy of a gland in the male reproductive system. It is the fifth leading cancer diagnosed and the seventh leading cause of cancer deaths in Korean men<sup>1</sup>. The overall mortality rate in Korean men due to cancer was 188.7 per 100,000 and 6.6 per 100,000 for prostate cancer in 2014<sup>2</sup>. Additionally, the prevalence, incidence, and mortality of prostate cancer in Korean men have increased significantly in the past few years<sup>1</sup>. Heritable genetic factors contribute to the susceptibility of various cancers and the genetic attribution of the incidence of prostate cancer is more than any other cancer type<sup>3</sup>. Twin studies have shown that genetic factors contribute to 42% of the incidence of prostate cancer<sup>3</sup>. Furthermore, first degree relatives are known to have two to three-fold increased risk of developing prostate cancer<sup>4</sup>. These observations indicate that germline variants contribute to prostate cancer. However, the identification of germline factors involved in prostate cancer has been limited in scope, and it is still a subject of ongoing research.

In recent years, there has been a push to discover common and rare variants associated with prostate cancer among different ethnicities, as growing evidence suggests that germline factors associated with prostate cancer susceptibility may differ among different ethnicities<sup>5</sup>. Prostate cancer is the second leading cause of cancer mortality in American men<sup>6</sup>. The prevalence and mortality rates of prostate cancer differ across European, African,

<sup>1</sup>Department of Urology, Seoul National University College of Medicine, Seoul National University Bundang Hospital, Seongnam, Korea. <sup>2</sup>Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. <sup>3</sup>Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. <sup>4</sup>Department of Biomedical Informatics, University of Utah, University of Utah School of Medicine, Salt Lake City, UT, USA. <sup>5</sup>Complex Diseases and Genome Epidemiology Laboratory, Department of Public Health, Graduate School of Public Health, Seoul National University, Seoul, Korea. <sup>6</sup>Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA, USA. \*email: [Dokyoon.Kim@penmedicine.upenn.edu](mailto:Dokyoon.Kim@penmedicine.upenn.edu); [ssbyun@snubh.org](mailto:ssbyun@snubh.org)

	Case (n = 1,008)	Control (n = 6,438)
<b>Age (years)</b>		
Mean	67.43	54.39
Sd	7.23	9.32
Median	68	54
<b>BMI (kg/m<sup>2</sup>)</b>		
Mean	24.47	24.16
Sd	8.23	3.12
Median	24.31	24.20

**Table 1.** Characteristics for study population showing mean, median and standard deviation of Age and Body Mass Index (BMI).

and Asian ethnic groups<sup>6</sup>. In particular, the incidence rates are particularly high in African American men and substantially low in Japanese and mainland Chinese population<sup>6</sup>. However, the incidence is higher among immigrant Japanese in the United States as compared Japanese living in Japan, but it is still about half that of American European population<sup>6</sup>. Thus, it is evident that germline factors associated with prostate cancer differ among ethnicities; therefore, in this study, we aimed to find the germline variants specific to Korean population that are associated with prostate cancer.

With the advent of high throughput genotyping technologies in the past few years, it has become easier to sequence thousands of samples. Many genome-wide association studies (GWAS) have been conducted to identify common variants associated with prostate cancer<sup>7–9</sup>. However, European populations are better characterized relative to Asian populations<sup>8</sup>. For instance, more than 100 loci have been identified in GWAS studies using European cohorts<sup>7,8</sup>, while studies that have performed GWAS using participants of Chinese and Japanese ethnicities have identified only 12 significant loci associated with prostate cancer<sup>8</sup>. Another GWAS study using common variants identified 5 significant loci associated with prostate cancer in a Korean population<sup>9</sup>. However, the common variants discovered to date explain only a small portion of heritability of prostate cancer, thus leaving the majority of genetic risk unexplained<sup>10</sup>.

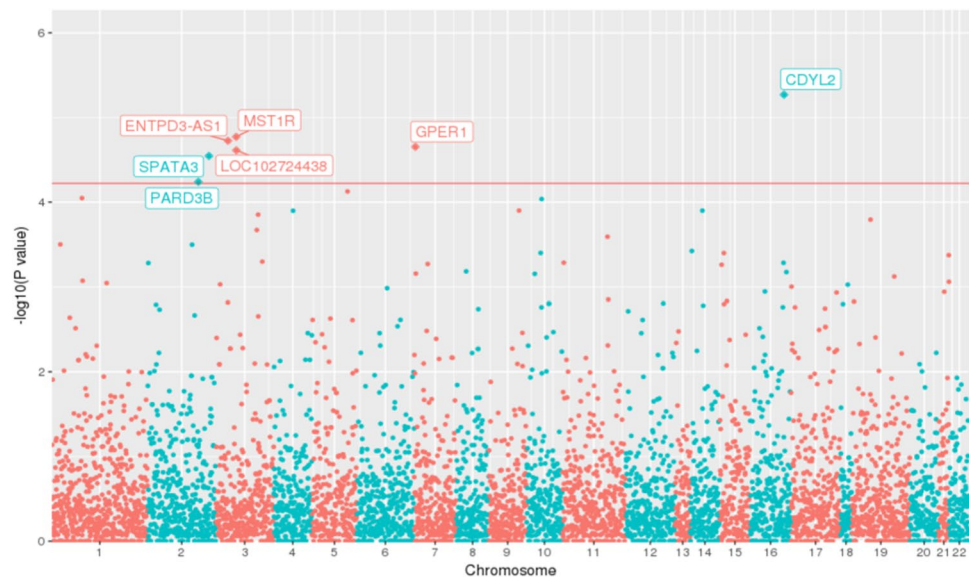
Most association studies on prostate cancer to date have focused mainly on common variants. However, a proportion of the missing heritability in prostate cancer could be further explained through low-frequency and rare variants. Rare variants play a key role in the contribution to heritability among different cancer types<sup>10,11</sup>. Many exome-wide studies have shown rare variants associated with susceptibility genes in colorectal cancer<sup>12</sup>, breast cancer<sup>13</sup>, prostate cancer<sup>14</sup>, and endometrial cancer<sup>15</sup>. However, very few rare variant studies have been conducted in Asian populations and even fewer in the Korean population. The variants discovered using common and rare variants are rarely replicated among European, Chinese, Japanese, and Korean population studies<sup>7–9,16</sup>. Thus, it is important to conduct separate association studies on ethnic populations to improve our understanding of the heritability of prostate cancer among population subsets. To this end, we analyzed whole-exome array data of germline samples from Korean individuals with and without prostate cancer. Rare variants were collapsed into genes and pathways across the genome and tested for association with prostate cancer. The significant variants in the gene or pathway were further annotated and evaluated for their association with prostate cancer.

## Results

**Study design and quality control.** Case population comprising 1,008 patients and a control group of 6,438 patients were obtained from the Korean Association Resource (KARE) study, which is a part of the Korean Genome and Epidemiology Study (KoGES)<sup>17</sup>. A summary of patient demographics for cases and controls is described in Table 1. All patients were men and of Korean ethnicity. The average ages of case and control patients were 67.43 and 54.39 years, respectively. The average body mass indices (BMI) of cases and controls were similar.

The quality control (QC) was performed to filter out bad samples and markers. All samples with marker call rate < 99%, sample call rate < 99%, and samples which were closely related (based on identity by descent (IBD) cutoff of 0.125) were dropped from the further analysis. After QC, 7,258 samples included 985 cases and 6,273 controls. The detailed steps involved in quality control are shown in Supplementary Figs 1 and 2. Using 71,270 variants that passed QC filters, rare variants with minor allele frequency (MAF) < 0.05 were binned into genes and association tests were performed. Further, the genomic inflation rate was calculated. A high genomic inflation rate usually indicates population substructure in data and spurious associations in the results. We found the genomic inflation factor  $\lambda_{1000}$  to be 1.196<sup>18</sup> for an equivalent study of 1,000 cases and 1,000 controls (Supplementary Fig. 3). Finally, 7,258 samples with 71,270 variants that passed QC were used for subsequent statistical analysis, as described in the Methods section.

**Identification of significant genes associated with prostate cancer.** The gene-based rare variant analysis was performed to identify associations between rare variants and prostate cancer. After binning variants into 5,830 gene bins, seven genes were identified to be associated with prostate cancer (FDR < 0.05 (Fig. 1 and Table 2)). In total, 19 rare variant loci spanning across 7 genes contributed to an association with prostate cancer susceptibility (Supplementary Table 1). Three genes, including *MST1R*, *GPER1*, and *PARD3B*, have been previously implicated in prostate cancer<sup>19–23</sup> and *CDYL2* in breast cancer<sup>24</sup>. Supplementary Fig. 4 shows regional plots for the genes with variants and exon span. Further, the distribution of MAF for the significant variants in this study population; Northeast Asians reference panel (NARD)<sup>25</sup>, which consists of the population from Korea



**Figure 1.** Manhattan plot showing the results of rare variant analysis. The 7 significant genes identified are shown in the Manhattan plot above the red line. The x axis highlights the chromosomes and y axis is  $p$ -value from the dispersion test (SKAT-O). The horizontal red line indicates genome wide significance level of FDR  $q = 0.05$ .

Gene	Chr: Build 38 position	# variants	cMAF	cMAF Case	cMAF Control	SKAT-O p-value	FDR q-value
CDYL2	16: 80597899-80805043	2	0.00055	0.00050	0.00056	5.39E-06	0.0277
MST1R	3: 49886471-49903637	5	0.01707	0.02066	0.01651	1.70E-05	0.0277
ENTPD3-AS1	3: 40390951-40453308	2	0.00034	0.00101	0.00024	1.89E-05	0.0277
GPER1	7: 1086807-1093815	2	0.02891	0.03175	0.02846	2.23E-05	0.0277
LOC102724438	3: 49899191-49925038	2	0.00296	0.00554	0.00255	2.45E-05	0.0277
SPATA3	2: 230996124-231019939	2	0.00372	0.00706	0.00319	2.85E-05	0.0277
PARD3B	2: 204545793-205620162	6	0.04163	0.04234	0.04154	5.76E-05	0.0480

**Table 2.** Gene-based rare variant analysis results using SKAT-O. Chr: Chromosome; cMAF: cumulative minor allele frequency of all variants included in the gene/bin; Seven genes significantly associated with prostate cancer after adjusting the SKAT-O  $p$ -value for multiple tests.

( $N = 850$ ), Mongolia ( $N = 386$ ), Japan, China, and Hong Kong; and gnomAD<sup>26</sup> are shown in Supplementary Table 2. The samples that had more than one rare variant among significant locus are shown in Supplementary Fig. 5.

Further, the rare association tests were rerun by removing one variant at a time from the gene/bin to elucidate the significance of that particular variant in the gene/bin (Supplementary Table 1). A decrease in the significance of the gene (increase in  $p$ -value) represents a significant contribution of the variant while maintaining a significant association suggests that the contribution of the variant is insignificant. In case of *MST1R* and *LOC102724438*, when the stop gained variant, rs200626206 was removed and the gene was insignificant ( $p_{rm} > 0.05$ ), indicating that it is the most significant variant in the gene/bin, as most of the signal detected in the genes is attributed to rs200626206 loci (Supplementary Table 1). Nonsense mutations often produce nonfunctional protein products due to premature termination of translation, thus the significant contribution of rs200626206 loci to the signal was expected. The variants rs181756759 and rs201829385 in genes *GPER1* and *PARD3B*, respectively, have insignificant  $p_{rm}$  and are predicted to be ‘probably damaging’ with Polyphen score  $> 0.95$  (Supplementary Table 1). Thus, there is a high possibility that these variants result in partial or complete loss of protein function<sup>27,28</sup>.

**Variant annotation.** To characterize the clinical significance, effect of variants on the protein, and implications in human inherited diseases, the variants within significantly associated genes were annotated using ClinVar and Variant Effect Predictor (VEP). None of the significant variants binned were found in ClinVar. VEP annotated the variants concerning their effect on the coding region (Supplementary Table 1). VEP also annotates variants by their potential influence on protein sequence (e.g., high, moderate, and low). High impact variants have a disruptive effect on proteins, such as protein truncation and loss of function. Moderate effect variants are non-disruptive but can change protein sequence while low effect variants are unlikely to change protein behavior.

Database	Pathway/Ontologies	p-value
WikiPathways (2019)	Estrogen signaling pathway WP712	0.008023
KEGG (2019)	Estrogen signaling pathway	0.04698
Reactome (2016)	Signaling by MST1_Homo sapiens_R-HSA-8852405	0.001749
Ingenuity canonical pathway	IL-15 Production	0.004
	MSP-ROn Signaling Pathway	0.01
	Sperm Motility	0.018
	IL-12 Signaling and Production in Macrophages	0.022
GO molecular function (2018)	MAP kinase activity (GO:0004709)	0.02733
	Transmembrane receptor protein tyrosine kinase activity (GO:0004714)	0.02150
	Transmembrane receptor protein kinase activity (GO:0019199)	0.02185
	Mitogen-activated protein kinase kinase binding (GO:0031434)	0.02288
	Phosphatidylinositol binding (GO:0035091)	0.03482

**Table 3.** Pathways/Ontologies discovered by Ingenuity pathway analysis, KEGG, Reactome, WikiPathways and GO molecular function ontology. The pathways discovered by Ingenuity pathway analysis using the 19 rare variants present in the 7 significant genes associated with prostate cancer. The WikiPathways, KEGG, Reactome, GO molecular function were generated using gene set enrichment analysis web server Enrichr<sup>29,30</sup>.

Here, VEP categorized 16 variants as a moderate effect, 2 variants as high effect, and 1 variant as a modifier (Supplementary Table 1). Therefore, most rare variants within the coding regions of genes associated with prostate cancer possibly influence protein function.

**Identification of significant pathways associated with prostate cancer.** Pathway analysis was performed using variants and genes discovered by gene-based rare variant association test. The biological pathways were derived using Ingenuity Pathway Analysis (IPA, QIAGEN Redwood City, [www.qiagen.com/ingenuity](http://www.qiagen.com/ingenuity)), and 4 canonical pathways were identified, including IL-15 Production, MSP-ROn Signaling Pathway, Sperm Motility and IL-12 Signaling, and Production in Macrophages (Table 3). IL-15 Production, MSP-ROn Signaling. Further, pathway enrichment test was also run using EnrichR, which identified estrogen signaling pathway, signaling by MST1 pathway, and four Gene Ontology (GO) molecular function ontologies<sup>29,30</sup>.

## Discussion

In this work, an exome-wide rare variant analysis study was performed, and 19 novel low-frequency variants in 7 genes were identified to be associated with prostate cancer in Korean men. Since common variants alone do not completely explain the heritability of prostate cancer, integrative analyses of rare variants across the genome can provide us with a new understanding of prostate cancer heritability. Though common variants have been studied previously in the Korean population, this is the first exome-wide study of rare variants associated with prostate cancer in the Korean population. The genes and variants discovered in this study can potentially help in early diagnostic and understanding of carcinogenesis in prostate cancer.

The rare variants were binned into genes and association tests were performed across the genes. Binning of rare variants increases statistical power to detect rare variant associations<sup>31</sup> and helps interpret the effect of rare variants on the prognosis and progression of prostate cancer. Several genes identified in this study were previously implicated with prostate cancer and other cancers. *In-vivo* studies have shown an association of *MST1R*, *GPER1*, and *PARD3B* with prostate cancer<sup>19–23</sup>. Another gene we found, *CDYL2*, has common variants that are implicated in breast cancer<sup>24</sup>. Since three of the genes discovered in this study have been previously validated to be associated with prostate cancer, it would appear that our analysis predicted true associations.

Angiogenesis, cell survival, and cell proliferation are hallmarks of cancer<sup>32</sup>. One of the genes that we found was significantly associated with prostate cancer, *MST1R* (Macrophage Stimulating 1 Receptor) or *RON*, is over-expressed in prostate cancer and various other cancers<sup>33</sup>. *RON* is known to be overexpressed in breast cancer and bladder cancer and is associated with poor clinical outcome<sup>33</sup>. *In vivo* study using genetically engineered mouse model has shown that the *RON* receptor plays a functional role in prostate tumor and that deficient *Ron* receptor signaling is associated with smaller tumor size<sup>34</sup>. Another study on stromal cells of the prostate tumor using mouse model showed that loss of *Ron* in tumor-associated macrophages inhibits cancer cell growth<sup>19</sup>. Another gene significantly associated with prostate cancer, *GPER1* (G Protein-Coupled Estrogen Receptor 1) or *GPR30*, is known to regulate cell growth by non-genomic signaling of estrogen<sup>35</sup>. *GPER1* is also known to stimulate cell proliferation in breast, endometrial, ovarian, and thyroid cancer cells by rapid but transient activation of *Erk1/2*<sup>35</sup>. Besides, in the case of prostate cancer, *GPER1* is known to control cancer cell growth through *GPER1* mediated pathways<sup>35</sup>. The presence of alternate allele in rs11544331, one of the rare variant loci binned in *GPER1*, is known to result in the expression of P16L variant of *GPER1*<sup>36</sup>. The substitution of proline with leucine at position 16 of the *GPER* protein sequence blocks *GPER* from being glycosylated and causes it to localize to the nucleus, although typically it should localize outside of nucleus<sup>36</sup>. The P16L in the nucleus may also regulate transcription of cancer-relevant genes and migration of carcinoma cells<sup>36</sup>.

Of the genes identified in this study, *PARD3B* (Partitioning defective 3 homolog) plays an essential role in asymmetric cell division, polarized growth, and maintenance of cell-polarity<sup>37</sup>. Mutational inactivation of its homolog gene *PARD3* is known to cause carcinogenesis in prostate cancer<sup>38</sup>. *In vivo* studies have shown that



downregulation of Par3 in breast cancer induces cell invasion and metastasis by decreasing cell-cell cohesion in a Tiam1/Rac-GTP pathway-dependent manner<sup>23</sup>. Higher expression of *PAR3B* is associated with colorectal cancer malignancy and poor survival, as *PAR3B* inhibits Lkb1/AMPK signaling pathway and its knockout induces apoptosis and reduces proliferation, supporting its role in colorectal cancer cell survival<sup>39,40</sup>. Additionally, a previous genome-wide association study found rs2335704, which resides in *PAR3B*, to be associated with tuberculosis<sup>41</sup>. Another gene, *CDYL2* (Chromodomain Y Like 2 or Prostate Cancer Candidate Protein 1), is involved in catalytic activity, protein binding, and methylated histone binding<sup>42</sup>. Genome-wide studies using common variants have identified loci in *CDYL2* associated with breast cancer<sup>43</sup>.

*ENTPD3-AS1* is a long non-coding RNA that we found was significantly associated with prostate cancer. A locus (rs193921050) in *ENTPD3-AS1* has been reported for 'Malignant tumor of prostate' in ClinVar with uncertain clinical significance and review status of 0/4<sup>44</sup>. The mutation on the locus was discovered in somatic tissue but was not found to be mutated at a significantly higher rate relative to the background mutation rate<sup>45</sup>.

In addition to the genes, the estrogen signaling pathway was significantly enriched using WikiPathways and KEGG pathway. Evidence suggests that prostate carcinogenesis and progression involves local estrogen signaling mechanisms<sup>46,47</sup>. Further, signaling by MST1 pathway was enriched using the Reactome pathway database. An *in vitro* study showed that MST1 suppressed prostate cancer growth<sup>48</sup>. Moreover, MST1 is the key kinase component of the Hippo-YAP pathway, which restricts prostate cancer progression by interacting with multiple molecular pathways<sup>49</sup>. The Ingenuity pathway analysis revealed three significant pathways. IL-15 production pathway, one of the significant pathways, is known to be associated with prostate cancer. The expression of IL-15 is known to decrease the migration, invasion, and angiogenesis but increase tumor volume by increasing lipid deposition and inflammation in prostate cancer<sup>50</sup>. IL-15 also alters the expression of genes involved in cell death and immune response<sup>50</sup>. Vaccinations using IL-15 are effective in up-regulating immune responses, reducing invasion, and improving survival<sup>51</sup>. Another pathway, MSP-RON signaling pathway, has been previously known to be associated with many cancer types, including prostate cancer, and has been extensively studied *in vivo* and *in vitro*. The MSP-RON signaling generates oncogenic variants and activates downstream pathways, resulting in tumorigenesis, proliferation, angiogenesis, invasion, and resistance to chemotherapy<sup>52</sup>. Loss of RON in myeloid cells has been shown to reduce prostate cancer growth in mice models<sup>52</sup>. IL-12 signaling is anti-carcinogenic, and IL-12 deficiency in mice is known to induce the development of spontaneous tumors and promote their growth<sup>53</sup>. Further, the GO molecular function ontology enrichment indicated four significantly enriched molecular function ontologies, including Mitogen-activated protein kinase (MAPK) activity, Transmembrane receptor protein tyrosine activity, Transmembrane receptor protein kinase activity, and Phosphatidylinositol binding. All the ontologies discovered are known to be associated with prostate cancer<sup>33,54,55</sup>.

Even though some of the genes we found have already been implicated, we found 19 novel variants and 3 novel genes that are associated with prostate cancer. Moreover, all the pathways that were found to be associated have been well studied and have been found to play a key role in cancer. All the variants discovered were missense and stop gained, except one variant in an intron, as categorized by VEP. Variation in amino acid sequence could potentially affect stability, conformational dynamics, drug response, and other protein properties that could result in a pathological condition and increased susceptibility to disease<sup>56</sup>. Many variants were also filtered out when re-clustering and filtering were performed using CHARGE criteria, a more relaxed approach shown in Park *et al.*, which could be used to preserve more variants while maintaining the genotyping accuracy of common and rare variants<sup>57</sup>. Further studies would be required to validate our findings, as variants discovered in this study were not discovered in the European population studies<sup>58</sup>. This is probably due to different genetic factors affecting prostate cancer susceptibility among different ethnic groups<sup>5</sup>. Additional exploration of the molecular mechanisms is required to understand the exact role of the variants in prostate cancer. Further studies are also required to elucidate the role of lifestyle/environment, especially dietary factors, in the Korean population, as they have been previously shown to be associated with prostate cancer<sup>4</sup>. In conclusion, we found novel genes and rare variants that are associated with prostate cancer in the Korean population, revealing potential biomarkers for prostate cancer that are unique to Korean ethnicity. They could also help us explain the missing heritability in prostate cancer, which could be applied in the field of precision medicine.

## Methods

**Samples and data set.** Between November 2003 and July 2013, we prospectively recruited 1,008 prostate cancer patients from a single tertiary hospital, Seoul National University Bundang Hospital, and conducted a case-control study that included 6,438 age-matched controls from the Korean Association Resource (KARE) study, which was a part of the Korean Genome and Epidemiology Study (KoGES)<sup>17</sup>. Blood specimens were prospectively collected throughout the course of the study from all of the prostate cancer patients. The automatic firing mechanism was used to perform transrectal ultrasound-guided multi-core ( $\geq 12$ ) biopsies bilaterally near the base, mid-gland, and apex, with at least six biopsies per side. A total of 12 baseline biopsy cores were taken from all of the men, and additional biopsies of suspicious lesions were obtained if needed. Further, 820 patients among the study population who had prostate cancer were treated with radical prostatectomy (RP) in the same hospital. The genotyping was done using the blood samples collected.

**Exome chip.** The Illumina HumanExome BeadChip 12v1-1 system provides 242,901 variants selected over 12,000 individual human exome and whole-genome sequences representing diverse populations and ethnicities. The chip focuses on protein-altering variants. A more detailed explanation is available at [http://genome.sph.umich.edu/wiki/Exome\\_Chip\\_Design](http://genome.sph.umich.edu/wiki/Exome_Chip_Design).

**Genotyping and quality control.** The datasets were generated using Illumina HumanExome BeadChip 12v1-1. We used Illumina's GenTrain version 2.0 clustering algorithm with the GenomeStudio software

(V2011.1) for genotype calling. The genotype calling for the exome chip was performed following the best practices defined in Grove *et al.*<sup>59</sup>. We performed manual re-clustering and visual inspection using CHARGE clustering method<sup>59</sup> to improve the accuracy of variant calling (Supplementary Fig. 6). A separate study on the quality of variants showed that re-clustering using CHARGE criteria on KoGES dataset with more relaxed cutoffs has 99.9% concordance rate for rare variants with whole-exome sequencing data, which indicates that rare variant calls are robust<sup>57</sup>. Quality control filters were applied to both case and control datasets separately. Since the number of variants differed in the datasets, only common variants between cases and controls were selected for the analysis. The datasets were merged after QC, and filters were again applied to the merged dataset. As a part of the quality control, sample call rates, marker call rates, and sample relatedness were checked<sup>60</sup>. The palindromic SNPs and SNPs with indels were removed. The samples and markers with a call rate of less than 99% were removed. Identity by descent (IBD) was calculated using plink, and IBD threshold of 0.125 was used to remove related samples. The detailed quality control steps are shown in Supplementary Figs 1 and 2.

The final merged dataset had 7,258 samples with 985 cases and 6,273 controls and 71,270 variants. The dataset was checked for batch effects. Since datasets from different sources were merged, population stratification in the data could have occurred<sup>60</sup>. Principal Component Analysis was performed using SMARTPCA<sup>61</sup> on the dataset after the LD pruning using plink option '-indep-pairwise 50 5 0.2' and removing all SNPs with MAF < 0.05. PCA was performed to check case and control sample clusters (Supplementary Fig. 7). PCA was also performed by projecting onto 1,000 genomes data. The case and controls clustered together around the South Asian population, as shown in Supplementary Fig. 8. Further, quantile-quantile (Q-Q) plot was drawn using SKAT-O p-values to check for inflation (Supplementary Fig. 3).

**Rare variant gene-based association test.** The rare variant analysis was performed using BioBin (<https://ritchielab.psu.edu/software/biobindownload>), a tool that can be used to perform rare variant burden tests<sup>31</sup>. BioBin bins all variants into gene bins and variants outside genes into intergenic region bins. Subsequently, SKAT-O was used to test for statistical significance of associations<sup>62</sup>. SKAT-O increases statistical power by optimally combining burden and dispersion (SKAT) tests and adaptively applying them<sup>62</sup>. Since rare variants are statistically underpowered for the association test, binning of rare variants by biologically informed units, such as gene or pathway, increases statistical power to detect rare variant associations by increasing the composite allele frequency and reducing the degrees of freedom<sup>31</sup>.

BioBin is configured by default to bin all variants with minor allele frequency (MAF) below 5%. Library of Knowledge Integration (LOKI) is a database of genomic locations of SNPs and genes as well as known relationships among genes and proteins, such as interaction pairs, pathways, and ontological categories integrated from various disparate data sources<sup>31</sup>. LOKI provides prior knowledge to BioBin<sup>31</sup>. All variants with MAF < 0.05% were removed. Variants with MAF > 0.05% in case population or control population were included, and only genes with at least 2 variants were tested. Age and first 5 principal components were incorporated as covariates to adjust for age and population stratification. The first 5 principal components that defined maximum variance were selected, as shown in Supplementary Fig. 9. The weight-loci argument was used to add Madsen & Browning weights to each locus<sup>63</sup>. BioBin creates bin based on gene regions when the bin-regions argument is set using gene information from LOKI. The false discovery rate (FDR) correction was applied to adjust for multiple testing. Any FDR adjusted *q-value* < 0.05 was considered significant. The rare association tests were run again by removing one variant at a time from the significant bins to ascertain the significance of the variant. The higher the SKAT-O *p-value* ( $p_{rm}$ ), the more significant is the contribution of the variant in the bin.

**Variant annotation.** ClinVar is a public archive that connects human variation to phenotypes, clinical significance, relationship to human health, and other supporting data obtained through submissions by various groups<sup>44</sup>. These are aggregated to reflect both consensus and conflicting assertions<sup>44</sup>. Variant effect predictor (VEP) provides information about the variants' location, gene/transcript affected by variants, types of mutation (i.e., stop gained, missense, stop-lost, and frameshift), and protein change scores, which indicate possible partial/complete loss of function of the protein due to amino acid substitution. All the variants in significantly associated genes were annotated using ClinVar and VEP.

**Pathway enrichment analysis.** The significant variants and genes were used for pathway enrichment analysis using Ingenuity pathway analysis and Enrichr. Enrichr is a web-based enrichment analysis tool which contains pathway and ontology libraries from various sources<sup>29,30</sup>. The WikiPathways (2019), KEGG (2019), Reactome (2016), and GO molecular function (2018) libraries were used as part of Enrichr.

**Ethics statement.** This study was approved by our institutional review board (Seoul National University Bundang Hospital Institutional review board; IRB number, B-1312/232-302) and followed the rules stated in the Declaration of Helsinki. All participants provided written informed consent.

Received: 15 November 2018; Accepted: 25 October 2019;

Published online: 20 November 2019

## References

1. Han, H. H. *et al.* Epidemiology of prostate cancer in South Korea. *Prostate Int* 3, 99–102, <https://doi.org/10.1016/j.prmil.2015.06.003> (2015).
2. Shin, H.-Y. *et al.* Cause-of-death statistics in the Republic of Korea, 2014. *J Korean Med Assoc* 59, 221–232 (2016).

3. Lichtenstein, P. *et al.* Environmental and Heritable Factors in the Causation of Cancer — Analyses of Cohorts of Twins from Sweden, Denmark, and Finland. *New England Journal of Medicine* **343**, 78–85, <https://doi.org/10.1056/NEJM200007133430201> (2000).
4. Grönberg, H. Prostate cancer epidemiology. *The Lancet* **361**, 859–864, [https://doi.org/10.1016/S0140-6736\(03\)12713-4](https://doi.org/10.1016/S0140-6736(03)12713-4) (2003).
5. Nakagawa, H. Prostate cancer genomics by high-throughput technologies: genome-wide association study and sequencing analysis. *Endocr Relat Cancer* **20**, R171–R181, <https://doi.org/10.1530/ERC-13-0113> (2013).
6. Judd Moul, T. Z., Andrew A & Joseph L. *Prostate Cancer* (<http://www.cancernetwork.com/cancer-management/prostate-cancer>, 2015).
7. Al Olama, A. A. *et al.* A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. *Nat Genet* **46**, 1103–1109, <https://doi.org/10.1038/ng.3094> (2014).
8. Wang, M. *et al.* Large-scale association analysis in Asians identifies new susceptibility loci for prostate cancer. *Nature Communications* **6**, <https://doi.org/10.1038/ncomms9469> (2015).
9. Oh, J. J. *et al.* Exome-based genome-wide association study and risk assessment using genetic risk score to prostate cancer in the Korean population. *Oncotarget* **5**, <https://doi.org/10.18632/oncotarget.16540> (2017).
10. Mancuso, N. *et al.* The contribution of rare variation to prostate cancer heritability. *Nat Genet* **48**, 30–35, <https://doi.org/10.1038/ng.3446> (2016).
11. Wang, Y. *et al.* Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. *Nat Genet* **46**, 736–741, <https://doi.org/10.1038/ng.3002> (2014).
12. Gylfe, A. E. *et al.* Eleven Candidate Susceptibility Genes for Common Familial Colorectal Cancer. *PLOS Genetics* **9**, e1003876, <https://doi.org/10.1371/journal.pgen.1003876> (2013).
13. Thompson, E. R. *et al.* Exome Sequencing Identifies Rare Deleterious Mutations in DNA Repair Genes FANCC and BLM as Potential Breast Cancer Susceptibility Alleles. *PLOS Genetics* **8**, <https://doi.org/10.1371/journal.pgen.1002894> (2012).
14. FitzGerald, L. M. *et al.* Germline Missense Variants in the BTNL2 Gene Are Associated with Prostate Cancer Susceptibility. *Cancer Epidemiol Biomarkers Prev* **22**, 1520–1528, <https://doi.org/10.1158/1055-9965.EPI-13-0345> (2013).
15. Chen, M. M. *et al.* Exome-Wide Association Study of Endometrial Cancer in a Multiethnic Population. *Plos One* **9**, e97045, <https://doi.org/10.1371/journal.pone.0097045> (2014).
16. Ren, X.-Q. *et al.* Variants on 8q24 and prostate cancer risk in Chinese population: a meta-analysis. *Int J Clin Exp Med* **8**, 8561–8570 (2015).
17. Kim, Y. & Han, B.-G. Cohort Profile: The Korean Genome and Epidemiology Study (KoGES) Consortium. *Int J Epidemiol*, <https://doi.org/10.1093/ije/dyv316>.
18. Freedman, M. L. *et al.* Assessing the impact of population stratification on genetic association studies. *Nat Genet* **36**, 388–393, <https://doi.org/10.1038/ng1333> (2004).
19. Gurusamy, D. *et al.* Myeloid-Specific Expression of Ron Receptor Kinase Promotes Prostate Tumor Growth. *Cancer Res* **73**, 1752–1763, <https://doi.org/10.1158/0008-5472.CAN-12-2474> (2013).
20. Lee, H.-R., Hwang, K.-A. & Choi, K.-C. The estrogen receptor signaling pathway activated by phthalates is linked with transforming growth factor- $\beta$  in the progression of LNCaP prostate cancer models. *International Journal of Oncology* **45**, 595–602 (2014).
21. Nelles, J. L., Hu, W.-Y. & Prins, G. S. Estrogen action and prostate cancer. *Expert Rev Endocrinol Metab* **6**, 437–451, <https://doi.org/10.1586/eem.11.20> (2011).
22. Zhang, K. *et al.* Shp2 promotes metastasis of prostate cancer by attenuating the PAR3/PAR6/aPKC polarity protein complex and enhancing epithelial-to-mesenchymal transition. *Oncogene* **35**, 1271, <https://doi.org/10.1038/ncb2663>, <https://www.nature.com/articles/ncb2663#supplementary-information> (2015).
23. Xue, B., Krishnamurthy, K., Allred, D. C. & Muthuswamy, S. K. Loss of Par3 promotes breast cancer metastasis by compromising cell-cell cohesion. *Nat Cell Biol* **15**, 189–200, <https://doi.org/10.1038/ncb2663> (2013).
24. Ghossaini, M., Pharoah, P. D. P. & Easton, D. F. Inherited Genetic Susceptibility to Breast Cancer: The Beginning of the End or the End of the Beginning? *The American Journal of Pathology* **183**, 1038–1051, <https://doi.org/10.1016/j.ajpath.2013.07.003> (2013).
25. Yoo, S.-K. *et al.* Whole-genome reference panel of 1,781 Northeast Asians improves imputation accuracy of rare and low-frequency variants. *bioRxiv*, 600353, <https://doi.org/10.1101/600353> (2019).
26. Karczewski, K. J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv*, 531210, <https://doi.org/10.1101/531210> (2019).
27. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248–249, <https://doi.org/10.1038/nmeth0410-248> (2010).
28. Ramensky, V., Bork, P. & Sunyaev, S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res* **30**, 3894–3900, <https://doi.org/10.1093/nar/gkf493> (2002).
29. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128–128, <https://doi.org/10.1186/1471-2105-14-128> (2013).
30. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* **44**, W90–W97, <https://doi.org/10.1093/nar/gkw377> (2016).
31. Moore, C. B., Wallace, J. R., Frase, A. T., Pendergrass, S. A. & Ritchie, M. D. BioBin: a bioinformatics tool for automating the binning of rare variants using publicly available biological knowledge. *BMC Med Genomics* **6**, S6, <https://doi.org/10.1186/1755-8794-6-S2-S6> (2013).
32. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646–674, <https://doi.org/10.1016/j.cell.2011.02.013> (2011).
33. Leonis, M. A., Thobe, M. N. & Waltz, S. E. Ron-receptor tyrosine kinase in tumorigenesis and metastasis. *Future Oncol* **3**, 441–448, <https://doi.org/10.2217/14796694.3.4.441> (2007).
34. Thobe, M. N. *et al.* The RON receptor promotes prostate tumor growth in tramp mouse model. *Oncogene* **30**, 4990–4998, <https://doi.org/10.1038/ncb2011.205> (2011).
35. Chan, Q. K. Y. *et al.* Activation of GPR30 inhibits the growth of prostate cancer cells through sustained activation of Erk1/2, c-jun/c-fos-dependent upregulation of p21, and induction of G(2) cell-cycle arrest. *Cell Death Differ* **17**, 1511–1523, <https://doi.org/10.1038/cdd.2010.20> (2010).
36. Pupo, M. *et al.* A genetic polymorphism repurposes the G-protein coupled and membrane-associated estrogen receptor GPER to a transcription factor-like molecule promoting paracrine signaling between stroma and breast carcinoma cells. *Oncotarget* **8**, 46728–46744, <https://doi.org/10.18632/oncotarget.18156> (2017).
37. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **45**, D158–D169, <https://doi.org/10.1093/nar/gkw1099> (2017).
38. Kunnev, D., Ivanov, I. & Ionov, Y. Par-3 partitioning defective 3 homolog (C. elegans) and androgen-induced prostate proliferative shutoff associated protein genes are mutationally inactivated in prostate cancer cells. *BMC Cancer* **9**, 318, <https://doi.org/10.1186/1471-2407-9-318> (2009).
39. Li, T., Liu, X., Jiang, Q., Lei, X. & Liu, D. High expression of partitioning defective 3-like protein is associated with malignancy in colorectal cancer. *Tumor Biology* **39**, 1010428317698393, <https://doi.org/10.1177/1010428317698393> (2017).
40. Li, T., Liu, D., Lei, X. & Jiang, Q. Par3L enhances colorectal cancer cell survival by inhibiting Lkb1/AMPK signaling pathway. *Biochemical and Biophysical Research Communications* **482**, 1037–1041, <https://doi.org/10.1016/j.bbrc.2016.11.154> (2017).
41. Thye, T. *et al.* Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11.2. *Nat Genet* **42**, 739–741, <https://doi.org/10.1038/ng.639> (2010).

42. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat Genet* **25**, 25–29 (2000).
43. Michailidou, K. *et al.* Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet* **45**, 353–361e3612, <https://doi.org/10.1038/ng.2563> (2013).
44. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* **42**, D980–D985, <https://doi.org/10.1093/nar/gkt1113> (2014).
45. Lindberg, J. *et al.* The Mitochondrial and Autosomal Mutation Landscapes of Prostate Cancer. *European Urology* **63**, 702–708, <https://doi.org/10.1016/j.eururo.2012.11.053> (2013).
46. Bonkhoff, H. Estrogen receptor signaling in prostate cancer: Implications for carcinogenesis and tumor progression. *The Prostate* **78**, 2–10, <https://doi.org/10.1002/pros.23446> (2018).
47. Misawa, A. & Inoue, S. Estrogen-Related Receptors in Breast Cancer and Prostate Cancer. *Front Endocrinol (Lausanne)* **6**, <https://doi.org/10.3389/fendo.2015.00083> (2015).
48. Cinar, B. *et al.* MST1 is a multifunctional caspase-independent inhibitor of androgenic signaling. *Cancer Res* **71**, 4303–4313, <https://doi.org/10.1158/0008-5472.CAN-10-4532> (2011).
49. Ready, D. *et al.* Mapping the STK4/Hippo signaling network in prostate cancer cell. *Plos One* **12**, e0184590–e0184590, <https://doi.org/10.1371/journal.pone.0184590> (2017).
50. Rohena-Rivera, K. *et al.* IL-15 regulates migration, invasion, angiogenesis and genes associated with lipid metabolism and inflammation in prostate cancer. *Plos One* **12**, e0172786–e0172786, <https://doi.org/10.1371/journal.pone.0172786> (2017).
51. Morris, J. C. *et al.* Vaccination with tumor cells expressing IL-15 and IL-15R $\alpha$  inhibits murine breast and prostate cancer. *Gene Ther* **21**, 393–401, <https://doi.org/10.1038/gt.2014.10> (2014).
52. Yao, H.-P., Zhou, Y.-Q., Zhang, R. & Wang, M.-H. MSP–RON signalling in cancer: pathogenesis and therapeutic potential. *Nat Rev Cancer* **13**, 466, <https://doi.org/10.1038/nrc3545> (2013).
53. Tugues, S. *et al.* New insights into IL-12-mediated tumor suppression. *Cell Death Differ* **22**, 237–246, <https://doi.org/10.1038/cdd.2014.134> (2015).
54. Huang, Q., Shen, H. M. & Ong, C. N. Emodin inhibits tumor cell migration through suppression of the phosphatidylinositol 3-kinase-Cdc42/Rac1 pathway. *CMLS, Cell. Mol. Life Sci.* **62**, 1167–1175, <https://doi.org/10.1007/s00018-005-5050-2> (2005).
55. Zhang, Z. *et al.* Phosphatidylinositol 3-kinase  $\beta$  and  $\delta$  isoforms play key roles in metastasis of prostate cancer DU145 cells. *The FASEB Journal* **32**, 5967–5975, <https://doi.org/10.1096/fj.201800183R> (2018).
56. Stefl, S., Nishi, H., Petukh, M., Panchenko, A. R. & Alexov, E. Molecular mechanisms of disease-causing missense mutations. *J Mol Biol* **425**, 3919–3936, <https://doi.org/10.1016/j.jmb.2013.07.014> (2013).
57. Park, T.-J. *et al.* Practical Calling Approach for Exome Array-Based Genome-Wide Association Studies in Korean Population. *Int J Genomics* **2015**, 421715–421715, <https://doi.org/10.1155/2015/421715> (2015).
58. Gudmundsson, J. *et al.* A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. *Nat Genet* **44**, 1326–1329, <https://doi.org/10.1038/ng.2437> (2012).
59. Grove, M. L. *et al.* Best Practices and Joint Calling of the Human Exome BeadChip: The CHARGE Consortium. *Plos One* **8**, <https://doi.org/10.1371/journal.pone.0068095> (2013).
60. Turner, S. *et al.* Quality Control Procedures for Genome Wide Association Studies. *Curr Protoc Hum Genet* CHAPTER, Unit1.19, <https://doi.org/10.1002/0471142905.hg0119s68> (2011).
61. Patterson, N., Price, A. L. & Reich, D. Population Structure and Eigenanalysis. *PLOS Genetics* **2**, e190, <https://doi.org/10.1371/journal.pgen.0020190> (2006).
62. Lee, S. *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* **91**, 224–237, <https://doi.org/10.1016/j.ajhg.2012.06.007> (2012).
63. Madsen, B. E. & Browning, S. R. A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. *PLOS Genetics* **5**, <https://doi.org/10.1371/journal.pgen.1000384> (2009).

## Acknowledgements

This work was supported by NLM R01 NL012535. This project is also funded, in part, under a grant with the Pennsylvania Department of Health (#SAP 4100070267). The Department specifically disclaims responsibility for any analyses, interpretations or conclusions. This work was also supported by 1) Korean Urologic Oncology Society research fund (KUOS17-1), 2) SNUBH Research fund (No. 13-2015-009 & 02-2017-009) and 3) Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2017R1D1A1A09000743).

## Author contributions

J.J.O., M.S., D.K. and S.S.B. conceived the project. J.J.O., H.L., S.K.H., S.E.L. and S.S.B. curated data. J.J.O., M.S., J.M., S.V., Y.L., S.J.L., J.S. performed the formal analysis along with visualization. D.K. and S.S.B. provided resources. J.J.O., M.S., J.M., D.K. and S.S.B. supervised all aspects of this work. The original draft was written by J.J.O. and M.S. All authors reviewed and edited the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41598-019-53445-2>.

**Correspondence** and requests for materials should be addressed to D.K. or S.-S.B.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

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





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

© The Author(s) 2019