Germline DNA repair gene mutations in young onset prostate cancer cases in the UK: evidence for a more extensive genetic panel

D.A. Leongamornlert[1], E.J. Saunders[1], S. Wakerell[1], I. Whitmore[1], T. Dadaev[1], C. Cieza-Borrella[1], Sarah Benafif[1], M.N. Brook[1], J. Donovan[2], F. Hamdy[3], D. Neal[4], K. Muir[5], K. Govindasami[1], D.V. Conti[6], Z. Kote-Jarai[1,&*], R.A. Eeles[1,7*]

[1]Oncogenetics, Division of Genetics and Epidemiology, The Institute of Cancer Research, 123 Old Brompton Road, London, SW7 3RP, UK.

[2]School of Social and Community Medicine, University of Bristol, Canynge Hall, 39 Whatley Road, Bristol, BS8 2PS, UK.

[3]Nuffield Department of Surgical Sciences, University of Oxford, Oxford, UK, Faculty of Medical Science, University of Oxford, John Radcliffe Hospital, Oxford, OX1 2JD, UK.

[4]University of Cambridge, Department of Oncology, Addenbrooke's Hospital, Cambridge, CB2 0QQ, UK. and Cancer Research UK Cambridge Research Institute, Li Ka Shing Centre, Cambridge, CB2 0RE, UK

[5]Institute of Population Health, University of Manchester, Manchester, M13 9PL, UK.

[6]Department of Preventive Medicine, Keck School of Medicine, University of Southern California/Norris Comprehensive Cancer Center, Los Angeles, CA 90015, USA.

[7]The Royal Marsden NHS Foundation Trust, Fulham Road, London, SW3 6JJ, UK

[*]Joint last authors

[&] Corresponding author

Word count:

Abstract – 308

Main – 3414

Abstract

Background

Rare germline mutations in DNA repair genes are associated with prostate cancer (PCa) predisposition and prognosis.

Objective

To quantify the frequency of germline DNA repair gene mutations in UK PCa cases and controls, in order to more comprehensively evaluate the contribution of individual genes to overall PCa risk and likelihood of aggressive disease.

Design, Setting, and Participants

We sequenced 167 DNA repair and 8 PCa candidate genes in a UK based cohort of 1,281 young onset PCa cases (diagnosed at ≤60yrs) and 1,160 selected controls.

Outcome Measurements and Statistical Analysis

Gene-level SKAT-O and gene-set adaptive combination of *P* values (ADA) analyses were performed separately for cases versus controls, and aggressive (Gleason ≥8, n=201) versus non-aggressive (Gleason ≤7, n=1,048) cases.

Results and Limitations

We identified 233 unique protein truncating variants (PTVs) with MAF <0.5% in controls in 97 genes. The total proportion of PTV carriers was higher in cases than controls (15% vs. 12%, OR = 1.29, 95% CI 1.01-1.64, *P* = 0.036). Gene-level analyses selected *NBN* ($P_{SKAT\text{-}O}$ = $2.4\text{x}10^{-4}$) for overall risk and *XPC* ($P_{SKAT\text{-}O}$ = $1.6\text{x}10^{-4}$) for aggressive disease, both at candidate level significance (*P* <$3.1\text{x}10^{-4}$ & *P* <$3.4\text{x}10^{-4}$ respectively). Gene-set analysis identified a subset of 20 genes associated with increased PCa risk (OR = 3.2, 95% CI 2.1-4.8, P_{ADA} = $4.1\text{x}10^{-3}$), and 4 genes that increased risk of aggressive disease (OR = 11.2, 95% CI 4.6-27.7, P_{ADA} = $5.6\text{x}10^{-3}$), 3 of which overlap the predisposition gene set.

Conclusions

The union of the gene and gene-set level analyses identified 23 unique DNA repair genes associated with PCa predisposition or risk of aggressive disease. These findings will help facilitate the development of a PCa specific sequencing panel with both predictive and prognostic potential.

Patient summary

This large sequencing study assessed the rate of inherited DNA repair gene mutations between PCa patients and disease free men. A panel of 23 genes was identified that may improve risk prediction or treatment pathways in future clinical practice.

Introduction

Prostate cancer (PCa) is the most common solid tumour in men living in the developed world besides non-melanoma skin cancer and responsible for over 300,000 deaths per year worldwide [1], although the majority of PCa cases are diagnosed with low or intermediate risk disease. Family history (FH) is a strong risk factor for PCa, and twin studies demonstrate a large contribution by heritable genetic factors [2]. Increasing evidence indicates that both common and rare germline variation contribute to PCa predisposition [3, 4]. Rare loss of function (LoF) germline mutations in *BRCA2* have been convincingly implicated as contributing to both FH of PCa and increased likelihood of aggressive disease with poor prognosis, whilst lower mutational frequencies or less consistent evidence have also been presented for a small subset of additional DNA repair genes including *ATM*, *BRCA1*, *BRIP1*, *CHEK2*, *GEN1*, *MSH2*, *NBN*, *PALB2*, *RAD51D* and *RNASEL* [5-7].

In this study, we performed screening of 167 genes from DNA damage response and repair pathways within a large UK based case-control cohort with long follow-up, to further investigate the role of germline DNA repair gene mutations in PCa predisposition, clinical outcome and survival. To maximise the power in this study we utilised young onset cases (diagnosed at ≤60 years) and control samples screened for either no PCa FH or low PSA (<0.5ng/ml). These results should help to inform the composition of future gene panels for clinical screening and risk profiling.

Methods and materials

Study population

Self-reported European ancestry PCa cases were randomly selected from the young-onset (diagnosed at ≤60yrs) sub-cohort of the UK Genetic Prostate Cancer Study (UKGPCS) [8]. Control men with no FH of PCa were recruited from GP practices participating in UKGPCS, or with PSA <0.5 ng/ml from the Prostate Testing for Cancer and Treatment (ProtecT) trial [9]. Cases and controls were matched for genetic ancestry, with ethnicity confirmed for all samples by Principal Component Analysis and analyses restricted to genetically European ancestry individuals (Supplementary Methods, Supplementary Figures 6 & 7). No formal matching by age was performed, although the age profiles of the case cohort and control men with known age at recruitment were broadly similar (Table1). All studies were approved by the appropriate ethics committees (UKGPCS 848). All participants gave written informed consent.

Analyses were performed comparing all post-QC PCa cases (n=1,281) versus controls (n=1,160), and for case-case comparisons of aggressive (Gleason ≥8, n=201) versus non-aggressive (Gleason ≤7, n=1,048) cases (Table 1).

Target genes

We constructed a 175 gene sequencing panel after a literature review of DNA repair, damage response and cell cycle pathways and databases (Supplementary Methods). The panel comprised 107 genes in DNA repair pathways, 60 DNA damage response and cell cycle regulation genes and 8 other candidate PCa predisposition genes (*HOXB13, MSR1, RNASEL, AR, ESR1, ESR2*, *NKX3-1* and *SPOP*) (Table 2, Supplementary Table 1).

Target capture and sequencing

A custom SureSelect XT bait library (Agilent Technologies, Santa Clara, CA, US) was designed for coding regions of the 175 target genes. DNA libraries were prepared using an automated in-house sample preparation protocol (Supplementary Methods) and captured libraries sequenced using Illumina HiSeq 2000 v4 chemistry (Illumina, San Diego, CA).

Sequence data analysis, variant annotation and QC

Raw sequencing reads were aligned to GRCh37 using BWA 0.5.8 [10]. Samples reaching ≥80% of the target at ≥10x read depth as defined by Picard v.1.52 (http://broadinstitute.github.io/picard/) and contamination <3% as estimated by verifyBamID v1.1.1 (https://github.com/statgen/verifyBamID/releases) were genotyped using GATK v2.8-1 [11]. Per gene coverage levels were assessed using the GATK tool "DiagnoseTargets", with a per-base coverage QC threshold set at ≥8 reads at base quality ≥20. Low quality genotypes were removed according to established thresholds (Supplementary Methods) [12-14]. Standard QC procedures were applied to remove poorly performing samples and variants [15]. These include variant-level filters such as heterozygosity and missingness and sample-level filters including relatedness and divergent ancestry (Supplementary Methods). Due to the targeted nature of the sequencing data, ancestry QC was augmented with additional QC data from the OncoArray platform [16].

Variants were annotated by wANNOVAR [17] using RefSeq Gene definitions [18] and variant consequence was checked using Variant Effect Predictor (VEP; release 84, March 2016) [19]. Protein truncating variants (PTVs; frameshift indels, stop gain and splice variants) were also annotated with the VEP plugin Loss-Of-Function Transcript Effect Estimator (LOFTEE; https://github.com/konradjk/loftee/) and INDELs in splice sites were manually reviewed for consequence. For further analysis, variants were categorised into two groups; 1) Tier 1 contained all high confidence PTVs according to LOFTEE and manual splice-site review, 2) Tier 2 all remaining variants with Combined Annotation Dependent Depletion (CADD) v1.3 score >20 [20].

PCa Susceptibility Gene Identification

Comparisons of rare PTV frequencies between our cohort and previous publications were restricted to Tier 1 mutations with MAF <0.5% in our controls. For novel gene discovery tests, due to the low frequencies of individual variants in this study, we performed two distinct aggregate statistical tests for each study phenotype: (A) a gene-level SNP-set association test over all genes containing ≥2 Tier1 or 2 variants and (B) a gene-set level association test where Tier 1 mutations with MAF <0.5% in controls were collapsed per gene.

To identify associated genes (A) we used SKAT-O, a unified test able to tolerate the inclusion of neutral variants or variants with opposing direction of effect, which finds the optimal combination between burden and kernel tests for the tested data [21]. We tested only genes containing ≥2 variants (Tier 1 or Tier 2), with statistical significance set at a Bonferroni adjusted *P* value of $\alpha = 0.05$ / number of genes; *P* $<3.1\text{x}10^{-4}$ for case/control analysis (159 genes) and *P* $<3.4\text{x}10^{-4}$ for aggressive phenotype analysis (146 genes). To further investigate gene-level SKAT-O association signals we used Adaptive Combination of *P*-values (ADA), a "combination of *P*-values" method which adaptively truncates *P*-values with an optimal threshold for the tested data set, removing neutral variants and identifying the likely underlying variant-level components of the gene-level signal [22]. Gene-level ADA for genes identified by SKAT-O was run using all Tier 1 and 2 variants within these genes and default settings (corresponding to *P* value truncation thresholds of 0.1 to 0.2 considered in 0.01 increments) except increasing to 10,000 permutations and using the mid *P*-value setting [23].

We subsequently performed an additional gene discovery analysis (B) in which ADA was used to identify a candidate gene set rather than individual variants, by collapsing Tier 1 mutations with MAF <0.5% in controls on a per gene rather than variant level basis (except for *CHEK2* where 1100delC was a separated from all other *CHEK2* PTVs due to its relatively higher frequency), under the assumption that rare Tier 1 variants are more likely to confer a homogenous effect within each gene. For each phenotype, gene-set level ADA was run with default settings except mode = “dominant”, twoSided = F, midp = TRUE and 10,000 permutations. We report both the permuted *P*-value (P_{ADA}) and the truncation threshold (opt.t). To display the resulting gene set selected by ADA, Forest plots were constructed showing gene-level adjusted ORs calculated from the collapsed Tier 1 MAF <0.5% variant count using unadjusted Firth’s regression.

Survival Analysis

Survival analyses were performed within the PCa case cohort to examine the effect of gene set’s carrier status on patient outcome. The follow-up period was based on date of diagnosis, date of consent into the UKGPCS, and date of last follow-up. Cases were diagnosed and came under observation at date of consent. Survival time was calculated as the difference in time between age of diagnosis and last recorded follow-up or date of death.

Kaplan-Meier survival analysis and univariable Cox regression models, adjusted for age, were performed. Log-rank tests were performed to investigate the equality of survivor functions across gene sets. Multivariable Cox regression models of gene set carrier status were constructed, adjusted for age and all covariates significant at $P<0.05$ under Cox univariate regression. All survival analyses were performed in Stata 14.2 [24].

Results

Sequencing and sample summary

After QC, variant data were available for 1,281 PCa cases and 1,160 control samples. Of 175 genes targeted, three (*GTF2H2, SLX1A* & *SLX1B*) were excluded due to low coverage resulting from segmental duplication and one (*PRKDC*) removed as wANNOVAR was unable to annotate coding consequences due to an incomplete RefSeq gene definition (Supplementary Figure 1 & Supplementary Table 2). From the 171 tractable target genes, we classified 2,078 variants in 164 genes as Tier 1 or 2 (Supplementary Table 3).

Known Gene set enrichment

A total of 233 PTVs with MAF <0.5% in controls were identified in 97 of the genes passing QC. Overall PTV carrier burden was significantly enriched in PCa cases compared to controls (15% vs. 12%; *P* = 0.036). This enrichment was greater within the BROCA panel of cancer predisposition genes, primarily focussed on hormone-driven breast and ovarian cancers (http://web.labmed.washington.edu/tests/genetics/BROCA_VERSIONS) [25]. For the original 22 gene BROCA panel 57 PTVs were identified in 15 genes (4.5% in cases vs. 2.2% in controls; $P = 2.5\text{x}10^{-3}$), whilst for the current BROCAv7 containing 66 genes, 80 PTVs were identified in 23 genes (5.5% in cases vs. 3.5% in controls; *P* = 0.020). The greatest enrichment was for the top five genes reported by Pritchard et al. [7] (*ATM*, *BRCA1*, *BRCA2*, *CHEK2, GEN1*), with 38 total PTVs identified across all 5 genes (3.8% vs. 1.4%; $P = 2.1\text{x}10^{-4}$).

Gene-level association

Gene-level analyses were restricted to genes containing ≥2 Tier 1 and 2 variants. In the case/control analysis (159 genes tested) *NBN* reached significance ($P = 2.4\text{x}10^{-4}$; *P* = 0.18 for aggressiveness), as did *XPC* for the aggressive phenotype (146 genes tested) (P = 1.6x10^{-4}; P = 0.90 for overall PCa) (Figure 1, Supplementary Figures 2 & 3). In addition, *HOXB13* (P = 1.1x10^{-3}; P = 0.12 for aggressiveness) and *POLL* (P = 9.1x10^{-4}; P = 0.11 for aggressiveness) demonstrated nominal significance (P <0.05) in the case/control analysis.

To further investigate these SKAT-O association signals, we used ADA to interrogate the combination of variants contributing to the association (*HOXB13* and *POLL* were also included due to the well characterised role of *HOXB13* in PCa predisposition). For both *NBN* and *HOXB13*, ADA identified a single recurrent heterozygous non-synonymous variant enriched among PCa cases as responsible for the gene-level signal, whilst for *POLL* 4 of the 15 tested variants were identified as potentially modulating risk (3 protective and 1 pathogenic). For *XPC*, ADA selected 6 singleton heterozygous variants from the 9 variants tested as contributing to the aggressive phenotype, all of which were observed in different individuals (Table 3).

Candidate gene set discovery

For the case/control phenotype, ADA selected 20 distinct genes containing rare heterozygous protein truncating variants from a panel of 57 genes (both categories of *CHEK2* PTV selected). These genes were significantly enriched among PCa cases compared to controls (8.5% vs. 2.8%, OR = 3.2, 95% CI 2.1-4.8, P_{ADA} = 4.1x10^{-3}, opt.t = 0.2, Figure 2A) and eight patients were carriers of a more than one PTV (Supplementary Table 4). Only five of these genes overlap the BROCA 22 gene set (*ATM, BRCA1, BRCA2, CHEK2* and *MSH2*). In the aggressive phenotype analysis, out of 35 genes, ADA selected four that were significantly enriched in Gleason ≥8 cases in comparison to patients with Gleason ≤7 (8.0% vs. 0.8%, OR = 11.2, 95% CI 4.6-27.7, P_{ADA} = 5.6x10^{-3}, opt.t = 0.1, Figure 2B). Three of these genes overlap with the case/control gene set (*BRCA2, CHEK2* and *MSH2*), while *ERCC2* is unique to the aggressive set. In contrast to other *CHEK2* PTVs, the *CHEK2* 1100delC variant was not enriched among aggressive cases.

The combined set of 21 genes identified in these analyses demonstrated a continuum of aggressive phenotype risk (Supplementary Figure 4), with the upper tail defining predisposition genes with lower risk of aggressive disease and the lower tail the converse. We partitioned the gene set into non-overlapping sets of 18 genes in the predisposition panel (Predis18) and four in the aggressive (Agg4), with *CHEK2* split (1100delC in Predis18 and other PTVs in Agg4; Figure 2C). As would be expected given the phenotype criteria, Agg4 carriers showed significant enrichment for several clinical indicators of aggressive disease (higher PSA, Gleason, tumour stage and nodal spread). Predis18 carriers showed no association with any clinical variable (Table 4). A modest increase in PCa FH rate was observed among Predis18 carriers compared to non-carriers, whilst PCa FH rates were lower among Agg4 carriers; however both these trends were non-significant. Suggestive but non-significant increases in rates of breast and pancreatic cancer FH were also observed for carriers of the Agg4 gene set (Supplementary Table 5). Kaplan-Meier survival analysis showed a significant global difference across gene set carriers (Agg4, Predis18 and non-carriers) for both all-cause and PCa-specific mortality (log-rank test, $P_{\text{all-cause}} = 9.8\text{x}10^{-8}$, $P_{\text{PCa-specific}} = 4.1\text{x}10^{-6}$). This is attributable to Agg4 carriers demonstrating significantly worse survival than non-carriers, as survival between Predis18 carriers and non-carriers was very similar. For all-cause survival (Figure 3A), the five-year survival rate for Agg4 was 60% (95% CI 34-79%), Predis18 93% (95% CI 85%-97%), and non-carriers 89% (95% CI 87-91%). The hazard ratio for Agg4 carriers compared to non-carriers was 2.69 (95% CI 1.32-5.50; Figure 3C). A similar pattern was observed when considering only PCa-specific survival (Figure 3B), though hazard ratios were not statistically significant, possibly due to the reduction in the number of events (282 cf. 212). Five-year survival rate for Agg4 was 60% (95% CI 34-79%), Predis18 94% (95% CI 86-98%), and non-carriers 91% (95% CI 89-92%). The hazard ratio for Agg4 carriers compared to non-carriers was 1.83 (95% CI 0.77-4.39; Figure 3D).

Discussion

Direct sequencing approaches are required to investigate the effect of rarer germline variants in complex disease predisposition; however, to date these studies in PCa have generally been smaller in size, considered only a handful of candidate genes, or lacked control cohorts. In this study, we investigated the role of DNA repair and damage response genes in predisposition to PCa and aggressive disease in a case/control cohort. We focused on protein truncating (Tier1) and predicted conserved (Tier 2) variants using both gene-level SKAT-O and gene-set level ADA analyses.

Gene-level analysis of Tier 1 and 2 variants identified significant associations in *NBN* for PCa predisposition and *XPC* for disease aggressiveness. The *NBN* signal was refined by ADA to rs61753720, a G>T SNV resulting in a D95N substitution. A previous study by the ICPCG consortium found this variant at low frequency in both unselected (1/613) and familial (1/121) Finnish PCa cohorts and absent (0/440) in controls [26]. For the association between the *XPC* gene and higher Gleason score, ADA selected multiple singleton SNVs across the gene. Both *POLL* and *HOXB13* were also marginally associated with PCa predisposition in the case/control analysis. Since the role of *HOXB13* rs138213197 in PCa risk has been well established, sample size may have been a limiting factor in achieving Bonferroni corrected significance, suggesting that *POLL* may also warrant additional follow-up in larger cohorts or meta-analyses of individual studies.

Gene-set level analysis identified 20 genes in which PTVs were associated with PCa predisposition. These included the established *BRCA1/2* genes, a handful of additional genes that have been indicated previously as prospective PCa candidates (*ATM, CHEK2*, *GEN1*, *MSH2* and *RNASEL)*, and several novel genes for which limited substantive evidence for a role in PCa predisposition has been presented to date (*BLM*, *CDC25C*, *ERCC3*, *LIG4*, *MSH5*, *NEIL2*, *NHEJ1*, *PARP2*, *POLD1*, *POLE*, *POLM*, *RECQL4* and *TDP1*). We furthermore identified four genes associated with more aggressive PCa phenotype, three of which overlapped the 20 gene PCa predisposition set. These include *BRCA2*, for which association with a more aggressive phenotype has been reliably demonstrated [6, 7, 27, 28], whilst we also present evidence that carriers of PTVs in *MSH2, CHEK2* (excluding 1100delC) and *ERCC2* also have substantially higher likelihood of developing aggressive disease.

Our criteria to stratify cases for the aggressive phenotype analysis (Gleason score ≤7 versus ≥8) were chosen to maximise the homogeneity and risk of the aggressive group. Within the Gleason 7 category however, Gleason 4+3 patients have poorer prognosis than those with 3+4, with these two subgroups categorised separately according to the prognostic grade grouping (PGG) method [29]. We therefore compared the results of our aggressive analysis to those with Gleason 4+3 cases re-classified as aggressive; equivalent to Grade group ≤2 versus ≥3 (n = 924 vs. 324) instead of Grade group ≤3 versus ≥4 used for our primary analysis. Under this classification, ADA selected the Agg4 gene set alongside three additional genes (*ESR2*, *GTF2H4* and *SETMAR*; $P_{ADA} = 8.1\times10^{-3}$, opt.t = 0.105). Additional comparisons between Gleason ≥8 cases versus controls selects the same Agg4 genes as our primary aggressiveness analysis (P_{ADA} = 0.014, opt.t = 0.115), whereas analysis of Gleason ≤7 cases versus controls selects 12 genes overlapping the Predis18 gene set identified in the case/control analysis (*ATM*, *BRCA1*, *CDC25C*, *CHEK2* 1100delC, *GEN1, LIG4, NEIL2, PARP2, POLD1*, *POLM*, *RECQL4, TDP1*; P_{ADA} = 0.029, opt.t = 0.12).

The overall 23 gene panel represented by the union of our gene and gene set level results for PCa susceptibility and disease aggressiveness represents a range of primary DNA repair pathways (Supplementary Table 1); with homologous recombination, mismatch repair, base excision repair, nucleotide excision repair, non-homologous end joining and DNA damage response all represented through multiple genes. Although Gleason Score was used to stratify aggressive and non-aggressive disease and is correlated to other features indicative of poor prognosis, among carriers of mutations in the Agg4 gene set we nevertheless observed substantial enrichment over non-carriers for nodal invasion (38% vs. 9.5%), metastatic disease (18% vs. 11%) and reduced survival (PCa-specific five-year survival rate 60% vs. 91%), suggesting that these genes could potentially demonstrate clinical utility for the identification of individuals at higher risk of advanced disease prior to progression. The absence of *BRCA1* and *ATM* from our aggressive gene set is however notable, as PTVs in these genes have been implicated in increased risk of metastatic and lethal PCa cancer previously [6, 7, 30]. This discrepancy may in part reflect our use of Gleason Score to define aggressive disease due to the modest proportion of patients with metastatic disease in our unselected cohort (7.2% of overall cohort, 11% excluding unknown status) in comparison to the more stringent metastatic or lethality indicators employed elsewhere in cohorts enriched for these outcomes, or alternatively that these genes confer lower influence upon aggressiveness in younger patients. It is also noteworthy that while *CHEK2* was associated with PCa predisposition for both 1100delC and other PTVs, only the non-1100delC *CHEK2* variants were found to contribute towards aggressive disease in our study. This observation however contrasts with a recent report in which only the 1100delC variant and not overall *CHEK2* mutations were enriched in lethal PCa patients [31], therefore requires further validation in independent cohorts. These combined reports could however potentially indicate that the downstream functional consequence of the 1100delC founder mutation may partly differ from those of other *CHEK2* PTVs in prostate tissue.

Whilst the novel genes we have identified represent exciting candidate moderate penetrance PCa risk genes, these findings nonetheless require additional validation in independent cohorts. In particular, we note that the optimal *P* Value truncation thresholds used by ADA are tuned towards greater sensitivity than specificity to maximise power for rare variant discovery in sequencing study sample sizes, and no suitable replication set was available for confirmation of our findings. Furthermore, even though this is the largest DNA repair gene germline sequencing study for PCa to date, our power to detect rare associations with moderate effect sizes remained modest.

While our strategy of using screened controls (no PCa FH or PSA <0.5ng/ml) potentially increased our power to detect associations, this also has the potential to introduce bias in our case/control analyses. We therefore cannot completely exclude the possibility that the use of PSA or FH in our control selection criteria led to an observed depletion of LoF variants among controls; although this would imply a uniform direction and comparatively high penetrance of effects across a wide range of DNA repair genes and pathways were these associations driven exclusively by extraneous variables such as low PSA levels independently of PCa.

Conclusion

In this study, we confirmed previous PCa predisposition gene reports and also present evidence for additional novel genes. Our combined gene and gene set level analyses provide evidence for a prospective screening panel of 23 genes that may facilitate identification of individuals at higher PCa risk prior to disease onset, who would warrant enhanced screening. In addition, PCa patients that are carriers of mutations in these genes could potentially benefit from personalised treatment pathways [27, 32]. We believe that these genes warrant evaluation by the wider scientific and clinical communities in larger prospective studies or meta-analyses. There is also a need to formally test the ability of these genes to predict survival in an independent cohort within aggressiveness strata.

Acknowledgements

We would like to acknowledge the NCRN nurses and consultants for their work in the UKGPCS study. We also wish to thank all the patients who took part in this study.

We would like to thank the following for funding support: Cancer Research UK Grants C5047/A17528, The Prostate Cancer Research Foundation (now Prostate Cancer UK), Prostate Research Campaign UK (now Prostate Cancer UK).

We also thank The Institute of Cancer Research, The Everyman Campaign, The National Cancer Research Network UK and The National Cancer Research Institute (NCRI) UK.

We are grateful for support of NIHR funding to the NIHR Biomedical Research Centre at The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust.

References

[1] Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer. 2015;136:E359-86.

[2] Mucci LA, Hjelmborg JB, Harris JR, Czene K, Havelick DJ, Scheike T, et al. Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. JAMA. 2016;315:68-76.

[3] Benafif S, Kote-Jarai Z, Eeles RA. A review of prostate cancer genome wide association studies (GWAS). Cancer Epidemiol Biomarkers Prev. 2018.

[4] Mancuso N, Rohland N, Rand KA, Tandon A, Allen A, Quinque D, et al. The contribution of rare variation to prostate cancer heritability. Nat Genet. 2016;48:30-5.

[5] Leongamornlert D, Saunders E, Dadaev T, Tymrakiewicz M, Goh C, Jugurnauth-Little S, et al. Frequent germline deleterious mutations in DNA repair genes in familial prostate cancer cases are associated with advanced disease. Br J Cancer. 2014;110:1663-72.

[6] Na R, Zheng SL, Han M, Yu H, Jiang D, Shah S, et al. Germline Mutations in ATM and BRCA1/2 Distinguish Risk for Lethal and Indolent Prostate Cancer and are Associated with Early Age at Death. Eur Urol. 2017;71:740-7.

[7] Pritchard CC, Mateo J, Walsh MF, De Sarkar N, Abida W, Beltran H, et al. Inherited DNA-Repair Gene Mutations in Men with Metastatic Prostate Cancer. N Engl J Med. 2016;375:443-53.

[8] Eeles RA, Dearnaley DP, Ardern-Jones A, Shearer RJ, Easton DF, Ford D, et al. Familial prostate cancer: the evidence and the Cancer Research Campaign/British Prostate Group (CRC/BPG) UK Familial Prostate Cancer Study. Br J Urol. 1997;79 Suppl 1:8-14.

[9] Lane JA, Donovan JL, Davis M, Walsh E, Dedman D, Down L, et al. Active monitoring, radical prostatectomy, or radiotherapy for localised prostate cancer: study design and diagnostic and baseline results of the ProtecT randomised phase 3 trial. Lancet Oncol. 2014;15:1109-18.

[10] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754-60.

[11] DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491-8.

[12] Carson AR, Smith EN, Matsui H, Braekkan SK, Jepsen K, Hansen JB, et al. Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. BMC Bioinformatics. 2014;15:125.

[13] Garner C. Confounded by sequencing depth in association studies of rare alleles. Genet Epidemiol. 2011;35:261-8.

[14] Lim ET, Wurtz P, Havulinna AS, Palta P, Tukiainen T, Rehnstrom K, et al. Distribution and medical impact of loss-of-function variants in the Finnish founder population. PLoS Genet. 2014;10:e1004494.

[15] Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. Nat Protoc. 2010;5:1564-73.

[16] Schumacher FR, Al Olama AA, Berndt SI, Benlloch S, Ahmed M, Saunders EJ, et al. Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. Nat Genet. 2018;50:928-36.

[17] Chang X, Wang K. wANNOVAR: annotating genetic variants for personal genomes via the web. J Med Genet. 2012;49:433-6.

[18] Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, et al. The UCSC Genome Browser database: 2015 update. Nucleic Acids Res. 2015;43:D670-81.

[19] McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17:122.

[20] Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46:310-5.

[21] Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, 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. 2012;91:224-37.

[22] Lin WY. Beyond Rare-Variant Association Testing: Pinpointing Rare Causal Variants in Case-Control Sequencing Study. Sci Rep. 2016;6:21824.

[23] Lin WY, Lou XY, Gao G, Liu N. Rare variant association testing by adaptive combination of P-values. PLoS One. 2014;9:e85728.

[24] StataCorp. Stata Statistical Software: Release 14. College Station, TX: StataCorp LP. 2015.

[25] Walsh T, Lee MK, Casadei S, Thornton AM, Stray SM, Pennil C, et al. Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing. Proc Natl Acad Sci U S A. 2010;107:12629-33.

[26] Hebbring SJ, Fredriksson H, White KA, Maier C, Ewing C, McDonnell SK, et al. Role of the Nijmegen breakage syndrome 1 gene in familial and sporadic prostate cancer. Cancer Epidemiol Biomarkers Prev. 2006;15:935-8.

[27] Castro E, Goh C, Leongamornlert D, Saunders E, Tymrakiewicz M, Dadaev T, et al. Effect of BRCA Mutations on Metastatic Relapse and Cause-specific Survival After Radical Treatment for Localised Prostate Cancer. Eur Urol. 2015;68:186-93.

[28] Castro E, Goh C, Olmos D, Saunders E, Leongamornlert D, Tymrakiewicz M, et al. Germline BRCA mutations are associated with higher risk of nodal involvement, distant metastasis, and poor survival outcomes in prostate cancer. J Clin Oncol. 2013;31:1748-57.

[29] Epstein JI, Zelefsky MJ, Sjoberg DD, Nelson JB, Egevad L, Magi-Galluzzi C, et al. A Contemporary Prostate Cancer Grading System: A Validated Alternative to the Gleason Score. Eur Urol. 2016;69:428-35.

[30] Mijuskovic M, Saunders EJ, Leongamornlert DA, Wakerell S, Whitmore I, Dadaev T, et al. Rare germline variants in DNA repair genes and the angiogenesis pathway predispose prostate cancer patients to develop metastatic disease. Br J Cancer. 2018.

[31] Wu Y, Yu H, Zheng SL, Na R, Mamawala M, Landis T, et al. A comprehensive evaluation of CHEK2 germline mutations in men with prostate cancer. Prostate. 2018;78:607-15.

[32] Antonarakis ES, Lu C, Luber B, Liang C, Wang H, Chen Y, et al. Germline DNA-repair Gene Mutations and Outcomes in Men with Metastatic Castration-resistant Prostate Cancer Receiving First-line Abiraterone and Enzalutamide. Eur Urol. 2018;74:218-25.

Table 1: Summary of study cohort characteristics

Clinical variable		Cases (n=1281)	Controls (n=1160)
Age of Diagnosis (Cases) or Blood Draw (Controls)	**Median**	57	56
	Quartiles	54-58	53-59
	Range	38-60	44-67
	Unknown (Count)	0 (0%)	637 (55%)
Ethnicity	**European Ancestry**	1281 (100%)	1160 (100%)
Diagnosis Method	**Clinical Symptoms**	739 (58%)	-
	Screen Detected	403 (31%)	-
	Unknown	139 (11%)	-
PCa Family History	**0**	973 (76%)	510 (44%)
	1	207 (16%)	17 (1.5%)
	2	40 (3.1%))	1 (0.1%)
	3+	5 (0.4%)	-
	Unknown	56 (4.4%)	632 (54%)
PSA at Diagnosis (ng/mL)	**Median**	8.4	-
	Quartiles	5.6-18.3	-
	Range	0.04-9020	-
	Unknown (Count)	43 (3.4%)	-
Gleason Score (Highest Recorded)	**≤6**	576	-
	7	472	-
	≥8	201	-
	Unknown	32	-
Primary Tumour Stage at Diagnosis	**T1**	365 (28%)	-
	T2	524 (41%)	-
	T3	295 (23%)	-
	T4	63 (4.9%)	-
	T_X	34 (2.7%)	-
Lymph Node Status at Diagnosis	**N0**	787 (61%)	-
	N1	89 (6.9%)	-
	N_X	405 (32%)	-
Distant Metastases at Diagnosis	**M0**	757 (59%)	-
	M1	92 (7.2%)	-
	M_X	432 (34%)	-

Table 2: Summary of gene panel composition by primary DNA repair pathway

Consensus pathway	Total Number of Genes
Direct reversal repair (DRR)	3
Base excision repair (BER)	25
Mismatch repair (MMR)	12
Nucleotide excision repair (NER)	30
Homologous recombination (HR)	26
Non-homologous end joining (NHEJ)	11
Fanconi Anaemia (FA)	19
DNA damage response (DDR)	22
Cell cycle regulation	19
PCa candidates	8
Total	**175**

Table 3. Variant level investigation of genes nominally significant in the SKAT-O gene-level analysis of Tier 1 and 2 variants

The number of unique variants per gene tested, individual variants selected by ADA, case and control variant counts, variant CADD v1.3 score, minor allele frequency in ExAC non-Finnish Europeans (NFE) and variant level *P*-values (using unadjusted Firth's logistic regression) are shown for each variant selected by ADA.

Case-control phenotype

Gene (Variants Tested)	ADA Selected Variants	rsID	Tier	Case (n=1281)	Control (n=1160)	CADD	ExAC NFE	Variant *P* Value
NBN **(4)**	8:90993640_C/T	rs61753720	2	18	2	26.3	0.0030	4.3x10^{-4}
POLL **(15)**	10:103339221_G/A	rs555309980	2	3	0	34	0.000047	0.13
	10:103339487_C/T	rs200705693	2	0	2	22.3	0.000091	0.20
	10:103342648_C/T	rs139871590	2	1	5	34	0.0015	0.09
	10:103343423_G/A	rs142726673	2	0	10	23.7	0.00080	4.7x10^{-4}
HOXB13 **(9)**	17:46805705_C/T	rs138213197	2	20	3	29.6	0.0031	5.9x10^{-4}

Aggressive phenotype

Gene (Variants Tested)	ADA Selected Variants	rsID	Tier	Case (n=1281)	Control (n=1160)	CADD	ExAC NFE	Variant *P* Value
XPC (9)	3:14187577_G/A	-	2	1	0	23.5	0.000015	0.07
	3:14193884_G/A	rs3731152	2	1	0	31	0.000033	0.07
	3:14199634_C/G	-	2	1	0	26.8	-	0.07
	3:14208716_T/C	rs200485886	2	1	0	24.7	0.000078	0.07
	3:14209787_G/A	rs188716339	2	1	0	24.2	0.000031	0.07
	3:14214457_G/A	-	2	1	0	22.8	-	0.07

Table 4. Clinical characteristics of Predis18 and Agg4 carrier and non-carrier cases

Age and PSA at diagnosis, Gleason Score, tumour grade, nodal spread and metastatic statuses are shown for carrier and non-carrier PCa cases of each gene set. Tests for enrichment between carriers and non-carriers were performed using Mann-Whitney U test (Age and PSA), Mantel-Haenszel test for linear-trend (Tumour stage) or Fisher's Exact test (nodal and metastatic spread).

Clinical variable		Agg4			Predis18		
		Carriers (n=24)	Non-carriers (n=1257)	Trend	Carriers (n=87)	Non-carriers (n=1194)	Trend
Age at diagnosis (yrs.)	Median	58	57	*P* = 0.14 U = 12470	57	57	*P* = 0.50 U = 54198
	Quantiles	54-59	54-58		54-58	54-58	
	Range	47-60	38-60		43-60	38-60	
PSA at diagnosis (ng/ml)	Median	29.6	8.3	*P* = 9.5x10^{-4} U = 8836	9.1	8.4	*P* = 0.57 U = 45811
	Quantiles	10.5-99.5	5.5-18		6-16.1	5.5-18.5	
	Range	0.41-399	0.04-9020		1.1-1151	0.04-9020	
	Unknown	0	43		5	38	
Gleason Score (Highest Recorded)	≤6	6	570		40	536	
	7	2	470		35	437	
	≥8	16	185		6	195	
	Unknown	0	32		6	26	
Primary Tumour Stage at Diagnosis	T1	1	364		18	347	
	T2	6	518	*P* = 1.1x10^{-5}	40	484	*P* = 0.40
	T3	9	286	M^2 = 19	22	273	M^2 = 0.70
	T4	5	58		3	60	
	T$_X$	3	31		4	30	
Lymph Node Status at Diagnosis	N0	13	774	-	54	733	
	N1	8	81	*P* = 5.6x10^{-4}	8	81	*P* = 0.51
	N$_X$	3	402		25	380	
Distant Metastases at Diagnosis	M0	18	739		52	705	
	M1	4	88	*P* = 0.29	3	89	*P* = 0.26
	M$_X$	2	430		32	400	

Figures

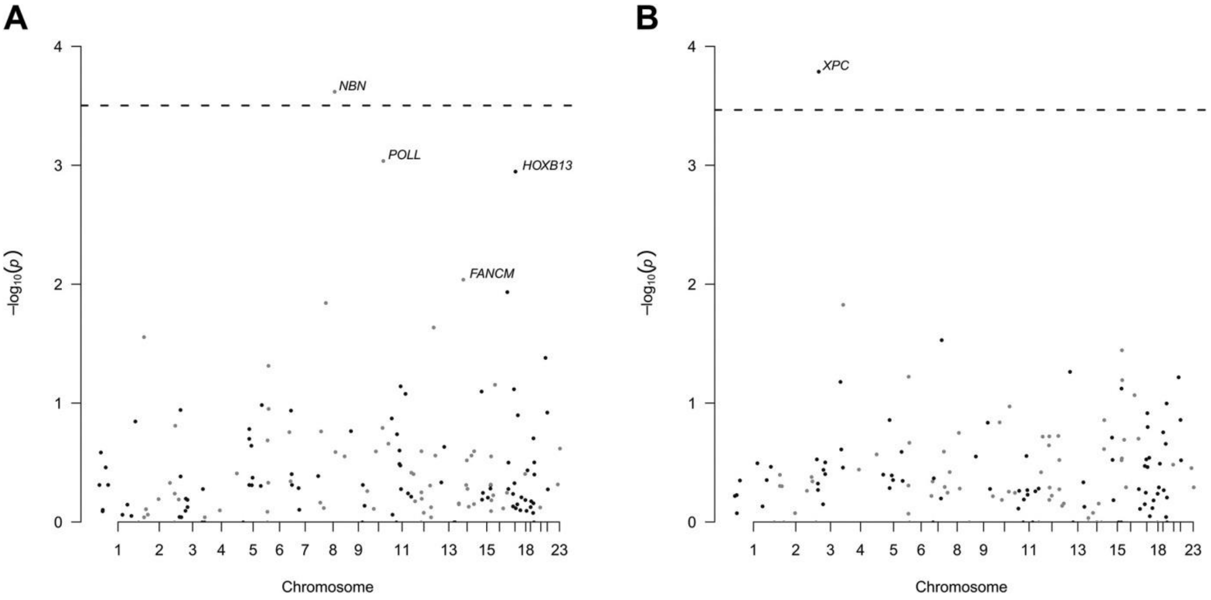
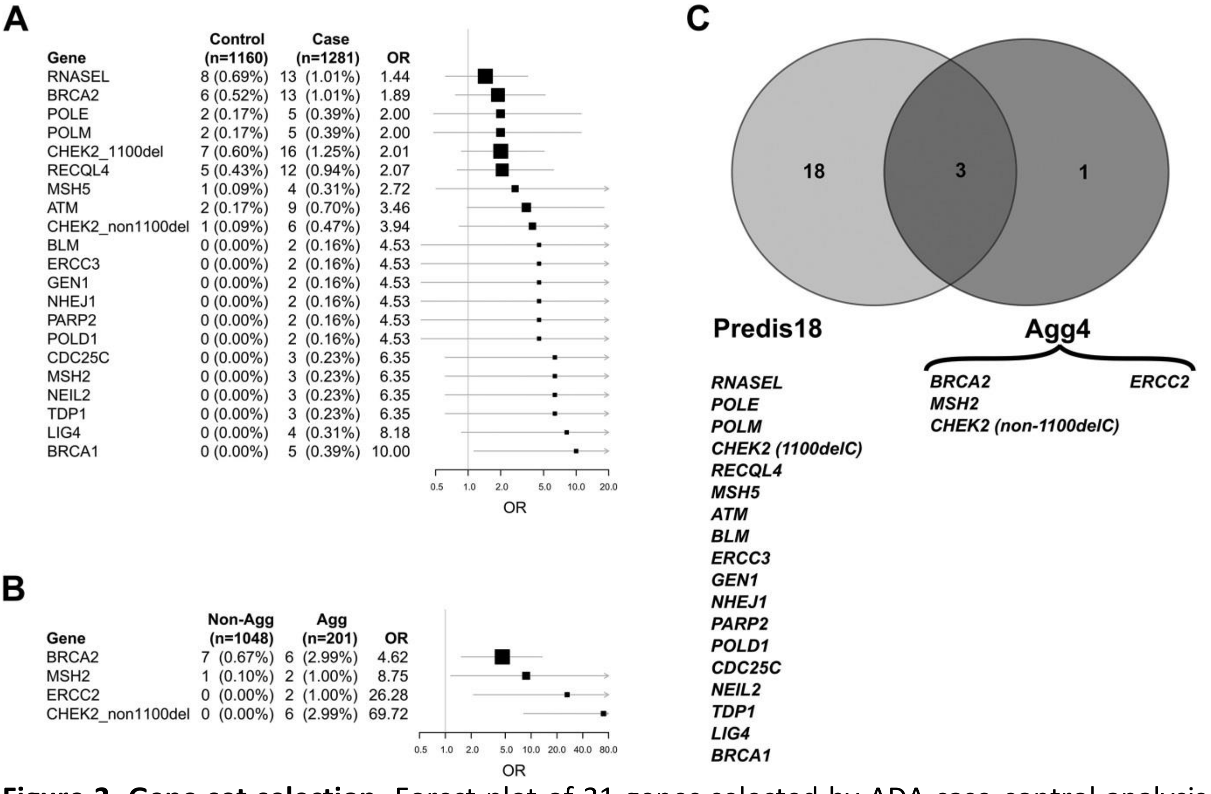


Figure 1. SKAT-O results for case control (A) and aggressive phenotypes (B). The blue line denotes the Bonferroni corrected candidate-level significance threshold for each phenotype, according to the number of genes containing ≥2 Tier 1 & 2 PTVs included in the analysis (159 and 146 respectively, genes are labelled at *P* < 0.05).

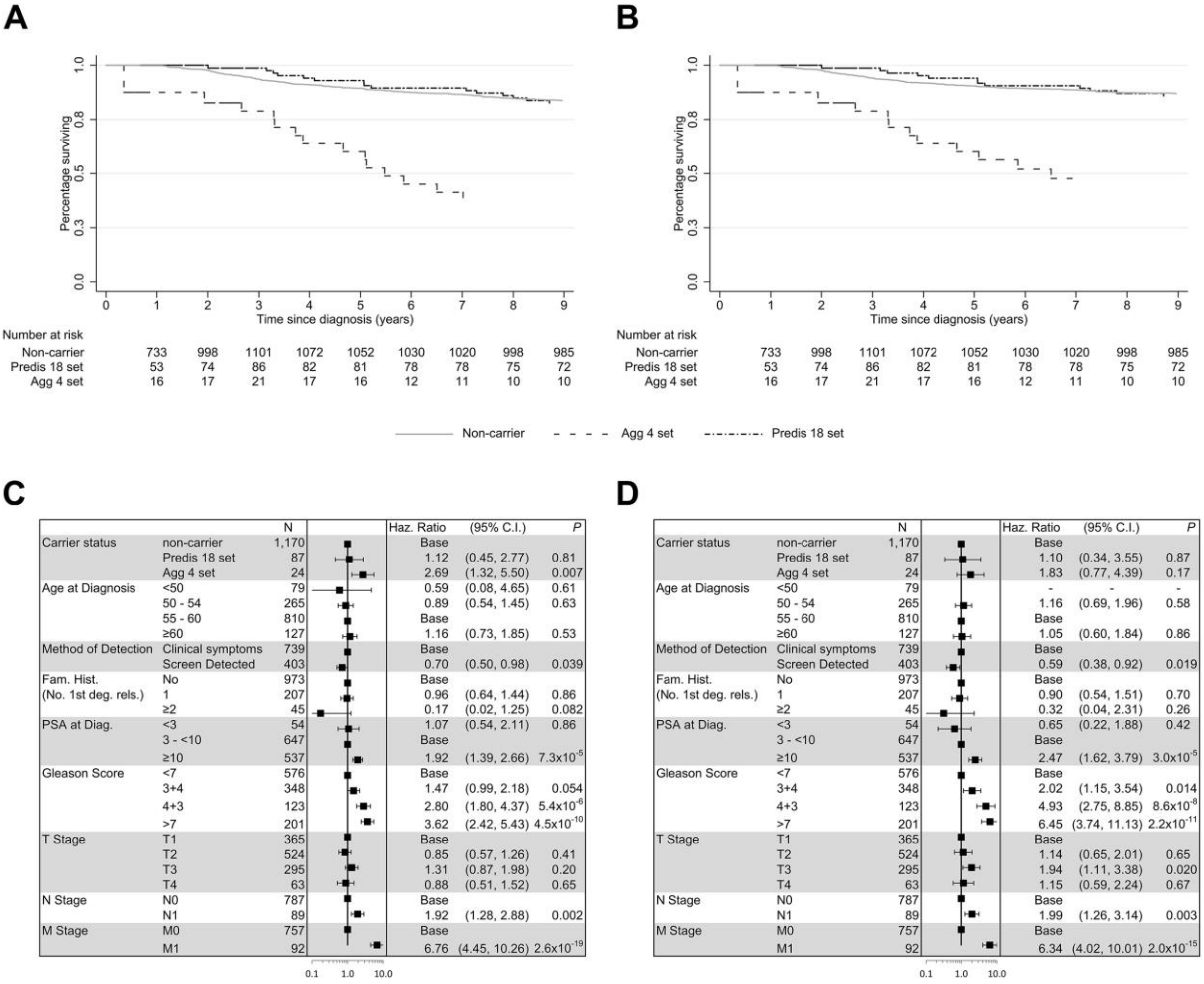


Gene	Control (n=1160)	Case (n=1281)	OR
RNASEL	8 (0.69%)	13 (1.01%)	1.44
BRCA2	6 (0.52%)	13 (1.01%)	1.89
POLE	2 (0.17%)	5 (0.39%)	2.00
POLM	2 (0.17%)	5 (0.39%)	2.00
CHEK2_1100del	7 (0.60%)	16 (1.25%)	2.01
RECQL4	5 (0.43%)	12 (0.94%)	2.07
MSH5	1 (0.09%)	4 (0.31%)	2.72
ATM	2 (0.17%)	9 (0.70%)	3.46
CHEK2_non1100del	1 (0.09%)	6 (0.47%)	3.94
BLM	0 (0.00%)	2 (0.16%)	4.53
ERCC3	0 (0.00%)	2 (0.16%)	4.53
GEN1	0 (0.00%)	2 (0.16%)	4.53
NHEJ1	0 (0.00%)	2 (0.16%)	4.53
PARP2	0 (0.00%)	2 (0.16%)	4.53
POLD1	0 (0.00%)	2 (0.16%)	4.53
CDC25C	0 (0.00%)	3 (0.23%)	6.35
MSH2	0 (0.00%)	3 (0.23%)	6.35
NEIL2	0 (0.00%)	3 (0.23%)	6.35
TDP1	0 (0.00%)	3 (0.23%)	6.35
LIG4	0 (0.00%)	4 (0.31%)	8.18
BRCA1	0 (0.00%)	5 (0.39%)	10.00

Gene	Non-Agg (n=1048)	Agg (n=201)	OR
BRCA2	7 (0.67%)	6 (2.99%)	4.62
MSH2	1 (0.10%)	2 (1.00%)	8.75
ERCC2	0 (0.00%)	2 (1.00%)	26.28
CHEK2_non1100del	0 (0.00%)	6 (2.99%)	69.72

Figure 2. Gene set selection. Forest plot of 21 genes selected by ADA case-control analysis (A) and 4 genes selected by ADA aggressive phenotype analysis (B). Odds Ratios (OR) were estimated from the collapsed Tier 1 MAF <0.5% variant count using unadjusted Firth's

logistic regression with 0.5 added to each count to provide estimates for genes with no carriers in one cohort. Overlap of gene sets between Predis18 and Agg4 candidate gene sets is also shown (C).



A

Number at risk									
Non-carrier	733	998	1101	1072	1052	1030	1020	998	985
Predis 18 set	53	74	86	82	81	78	78	75	72
Agg 4 set	16	17	21	17	16	12	11	10	10

B

Number at risk									
Non-carrier	733	998	1101	1072	1052	1030	1020	998	985
Predis 18 set	53	74	86	82	81	78	78	75	72
Agg 4 set	16	17	21	17	16	12	11	10	10

C

		N	Haz. Ratio	(95% C.I.)	*P*
Carrier status	non-carrier	1,170	Base		
	Predis 18 set	87	1.12	(0.45, 2.77)	0.81
	Agg 4 set	24	2.69	(1.32, 5.50)	0.007
Age at Diagnosis	<50	79	0.59	(0.08, 4.65)	0.61
	50 - 54	265	0.89	(0.54, 1.45)	0.63
	55 - 60	810	Base		
	≥60	127	1.16	(0.73, 1.85)	0.53
Method of Detection	Clinical symptoms	739	Base		
	Screen Detected	403	0.70	(0.50, 0.98)	0.039
Fam. Hist.	No	973	Base		
(No. 1st deg. rels.)	1	207	0.96	(0.64, 1.44)	0.86
	≥2	45	0.17	(0.02, 1.25)	0.082
PSA at Diag.	<3	54	1.07	(0.54, 2.11)	0.86
	3 - <10	647	Base		
	≥10	537	1.92	(1.39, 2.66)	$7.3x10^{-5}$
Gleason Score	<7	576	Base		
	3+4	348	1.47	(0.99, 2.18)	0.054
	4+3	123	2.80	(1.80, 4.37)	$5.4x10^{-6}$
	>7	201	3.62	(2.42, 5.43)	$4.5x10^{-10}$
T Stage	T1	365	Base		
	T2	524	0.85	(0.57, 1.26)	0.41
	T3	295	1.31	(0.87, 1.98)	0.20
	T4	63	0.88	(0.51, 1.52)	0.65
N Stage	N0	787	Base		
	N1	89	1.92	(1.28, 2.88)	0.002
M Stage	M0	757	Base		
	M1	92	6.76	(4.45, 10.26)	$2.6x10^{-19}$

D

		N	Haz. Ratio	(95% C.I.)	*P*
Carrier status	non-carrier	1,170	Base		
	Predis 18 set	87	1.10	(0.34, 3.55)	0.87
	Agg 4 set	24	1.83	(0.77, 4.39)	0.17
Age at Diagnosis	<50	79	-	-	-
	50 - 54	265	1.16	(0.69, 1.96)	0.58
	55 - 60	810	Base		
	≥60	127	1.05	(0.60, 1.84)	0.86
Method of Detection	Clinical symptoms	739	Base		
	Screen Detected	403	0.59	(0.38, 0.92)	0.019
Fam. Hist.	No	973	Base		
(No. 1st deg. rels.)	1	207	0.90	(0.54, 1.51)	0.70
	≥2	45	0.32	(0.04, 2.31)	0.26
PSA at Diag.	<3	54	0.65	(0.22, 1.88)	0.42
	3 - <10	647	Base		
	≥10	537	2.47	(1.62, 3.79)	$3.0x10^{-5}$
Gleason Score	<7	576	Base		
	3+4	348	2.02	(1.15, 3.54)	0.014
	4+3	123	4.93	(2.75, 8.85)	$8.6x10^{-8}$
	>7	201	6.45	(3.74, 11.13)	$2.2x10^{-11}$
T Stage	T1	365	Base		
	T2	524	1.14	(0.65, 2.01)	0.65
	T3	295	1.94	(1.11, 3.38)	0.020
	T4	63	1.15	(0.59, 2.24)	0.67
N Stage	N0	787	Base		
	N1	89	1.99	(1.26, 3.14)	0.003
M Stage	M0	757	Base		
	M1	92	6.34	(4.02, 10.01)	$2.0x10^{-15}$

Figure 3. Gene set survival. Kaplan-Meier survival plots depicting Overall Survival (A) and Cause-Specific Survival (B). Multivariate Cox regression analysis of phenotypic features and gene set carrier status are shown for Overall Survival (C) and Cause-Specific Survival (D). Analyses were conducted using PCa cases only.