AWARD NUMBER: W81XWH-18-1-0204

TITLE: Population-Based Identification of Prostate Cancer

PRINCIPAL INVESTIGATOR: L.A. Cannon-Albright, PhD

CONTRACTING ORGANIZATION: University of Utah, Salt Lake City, Utah

REPORT DATE: November 2022

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Development Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE
November 2022

2. REPORT TYPE
Final

3. DATES COVERED
01Aug2018-31Jul2022

4. TITLE AND SUBTITLE
Population-Based Identification of Prostate Cancer

5a. CONTRACT NUMBER
W81XWH-18-1-0204

5b. GRANT NUMBER
PC170413

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

LA Cannon-Albright

E-Mail:lisa.albright@utah.edu

5d. PROJECT NUMBER
PC170413

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Utah
201 Presidents Circle
Salt Lake City, UT 84112-9049

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Development Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Background: Prostate cancer (PrCa) is the most common cancer diagnosed in the US and one of the most familial. There is evidence for an inherited contribution to PrCa. Analysis of PrCa pedigrees led to discovery of genes that explain a small number of pedigrees (*ELAC2, RNASEL*, and *HOXB13*); and more than 100 common genetic variants have been reported to confer modest risk to PrCa. However, taken together, these recognized genetic risk factors explain few pedigrees. The likely genetic heterogeneity of PrCa and lack of success in gene identification suggests a different approach is needed to identify responsible predisposition genes. Analysis of related cases in extended high-risk cancer pedigrees is a powerful approach for identification of cancer predisposition genes. In Utah a resource combining the genealogy of the pioneer founders and their descendants with Utah cancer data allows identification of extended high-risk prostate cancer pedigrees. Analysis of the most clinically significant PrCa cases (those who die from their disease- lethal PrCa or LPrCa) in these pedigrees further enhances the power of this approach.

15. SUBJECT TERMS
None listed.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRDC
Unclassified	Unclassified	Unclassified	Unclassified	29	19b. TELEPHONE NUMBER

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

TABLE OF CONTENTS

	Page
1. Introduction	4
2. Keywords	4
3. Accomplishments	4
4. Impact	6
5. Changes/Problems	6
6. Products	7
7. Participants & Other Collaborating Organizations	7
8. Special Reporting Requirements	8
9. Appendices	8

1. INTRODUCTION

Prostate cancer (**PrCa**) is the most common cancer diagnosed in the US and one of the most familial. There is evidence for an inherited contribution to PrCa. Men with a first-degree relative with PrCa have a 2-3 fold risk, and those with ≥1 first-degree relative ≤ 50 years have ~4-fold risk (Albright et al., 2015). These risks are even higher for family history of death from PrCa (Albright et al., 2016). Analysis of PrCa pedigrees had led to discovery of a few genes that explain a small number of pedigrees (*ELAC2*, *RNASEL*, and *HOXB13*); and more than 100 common genetic variants have been reported to confer modest risk to PrCa. However, taken together, these recognized genetic risk factors explain few pedigrees. The likely genetic heterogeneity of PrCa and lack of success in gene identification suggests a different approach is needed to identify responsible predisposition genes.

Although it is a commonly held belief that complex diseases are likely due to many genes with small effects, and the common approach to identification of such predisposition genes is a genome-wide association study (GWAS) with spectacularly large sample sizes, it is clear that there exist extended high-risk prostate cancer pedigrees that are extremely likely to be due to rare variants with at least moderate effect size; such rare variants are not likely to be identified in a GWAS. Analysis of multi-generation families with an excess of disease has been shown to be a powerful approach to identification of predisposition genes. Our previous studies of Utah high-risk cancer pedigrees have identified predisposition genes for breast cancer (*BRCA1*-Miki 1994; *BRCA2*-Wooster 1994) and melanoma (*CDKN2A*-Kamb 1994); similar pedigree studies in other populations have more recently identified predisposition genes for other cancers (*HOXB13*-Ewing 2012; *ATM*-Roberts 2012).

We currently have several funded cancer predisposition gene identification grants using an innovative high-risk pedigree study design based on whole exome sequencing of affected cousin pairs (CA164138 - colon cancer; CA 195614 – melanoma; CA 205796 - small intestine carcinoid cancer). **Here we propose a similar approach using the more clinically significant phenotype of lethal PrCa and analyzing unique and informative extended high-risk Utah PrCa pedigrees to identify PrCa predisposition genes.**

2. KEYWORDS

Prostate cancer
Sequencing
Lethal prostate cancer
High-risk pedigree

3. ACCOMPLISHMENTS

• What were the major goals and objectives of the project?

We previously used the same high-risk pedigree approach proposed here to identify cancer predisposition genes in the Utah population (*BRCA1*- Miki, 1994; *BRCA2*- Wooster, 1994; *CDKN2A*- Kamb, 1994). We are currently funded by NCI to use this high-risk pedigree pair approach for colon cancer, melanoma, and small intestine carcinoid cancer in Utah pedigrees. We propose that WGS of pairs of related LPrCa cases from high-risk PrCa pedigrees will lead to identification of rare variants in genes that underlie predisposition to PrCa. Extended high-risk pedigrees are likely to evidence a strong role for genetic factors and the LPrCa case pairs to be sequenced are selected for clinical significance (death from PrCa) to ensure that they exhibit enhanced genetic contribution and limited genetic heterogeneity. This study relies on unique Utah resources and a powerful high-risk pedigree approach; it is complementary to previous efforts to identify prostate cancer predisposition genes (in pedigrees and GWAS), and it is not possible elsewhere.

SPECIFIC AIMS

Aim 1. Perform WGS on 50 related LPrCa case pairs (cousins) from the most informative set of 50 Utah high-risk PrCa pedigrees. Perform bioinformatics analysis to identify the top 4,000 rare candidate variants shared by a pair of LPrCa cases in a region shared Identical by Descent (IBD).

Subtask 1: Identify all sampled LPrCa and metastatic cancer cases with genealogy data; assemble all clusters including >=2 sampled cases; identify the subset of pedigrees that are high-risk PrCa; select a cousin pair from each of the most informative 50 high-risk pedigrees
Subtask 2: Perform QC on the selected samples; perform whole genome sequencing on the 100 LPrCa cases in the 50 pairs
Subtask 3: Perform bioinformatics analysis of WGS data and PC analysis of NDAR controls using the Genetic Epidemiology pipeline to identify variants of interest; filter for rare variant allele frequency and sharing in at least 1 cousin pair to identify the 4,000 candidate variants
Subtask 4: Share candidate variants with ICPCG
Local IRB Approval

Task 1. We updated data to identify all sampled LPrCa and metastatic cancer cases with genealogy data, assembled all high-risk pedigrees with at least 2 sampled cases, and selected a cousin pair from each.

Task 2. We performed QC on the selected samples, and found replacement samples/pedigrees where necessary. The first set of 50 samples was shipped to MedGenome for whole genome sequencing and the data were received 7/30/19. The second set of 50 samples has also been sent and data received in November 2019.

Task 3. We applied, and were approved, for space on the protected servers at the Center for High Performance Computing where all analysis was accomplished. Analysis is complete and ~4,000 candidate coding and non-coding variants were identified

Aim 2. Assay 4,000 rare shared candidate variants identified in Aim 1 in 1000 additional sampled Utah PrCa cases in the 50 high-risk pedigrees to establish evidence of segregation. Candidate *genes* with significant evidence for segregation will proceed to full candidate gene analysis in Aim 3.

Subtask 1: Design and order the Illumina Infinium iSelectHD assay for up to 5,000 candidate variants and order (minimum order = n=1,052 samples)
Subtask 2: Assemble, perform QC, and plate the ~1,000 additional sampled LPrCa and PrCa cases in the selected 50 high-risk pedigrees to deliver to the core
Subtask 3: Assay the 5,000 candidate variants; identify the carriers of the candidate variants
Subtask 4: Perform RVsharing analysis for each rare variant observed in the 1,000 cases
Subtask 5: Identify the ~400 candidate genes with significant evidence for segregation of at least 1 candidate variant in at least 1 high-risk pedigree

Task 1. Assay designed and orderd
Task 2. 1,195 prostate cancer samples plated for assay
Task 3. All ~4,000 candidate variants assayed in 1,195 cases
Task 4. Thousands of variants were observed in pedigrees. For this reason we removed variants occurring too frequently among the 1,195 cases, and we removed pedigrees that did not have a significant excess of prostate cancer.
Task 5. We have identified 571 candidate variants.

Aim 3. Test the ~400 candidate genes identified in Aim 3 with a targeted sequencing panel in 1,000 Caucasian and AA LPrCa and PrCa cases and compare to 1,000 cancer-free controls, to establish association with PrCa risk.
This Aim was revised to : Validate the ~500 candidate predisposition variants that gave evidence for segregation in Aim 2 with risk association in 2 large resources (UKBiobank with > 8,000 cases and >400,000 controls and PRACTICAL prostate cancer consortium with 79,194 cases and >61,000 controls.
And new tasks:
1: Get UKBiobank approval and import data for cases and controls
2: Utilize public shared PRACTICAL imputed genotypes for validation of candidates
3: perform case/ control association with PrCa for ~500 best candidate variants identified.
4: consolidate risk association validation results to identify significant predisposition variants
5. search for significant variants in existing sequence data for Utah high risk cancer pedigrees in Dr. Cannon-Albright's biorepository to extend high-risk pedigrees

~~Subtask 1: Design and order the Roche SeqCap EZ Choice targeted capture kit for the 400 candidate genes~~
~~Subtask 2: Assemble, perform QC, and plate the 174 AA cases and 926 Utah PrCa cases and deliver to Core~~
~~Subtask 3: Assay the 400 candidate genes; identify all rare variants in any candidate gene in cases~~
~~Subtask 4: Analyze the WGS data for 1,000 Utah controls and the ~3,000 NDAR matched Controls and perform case/control association analysis for validation testing of all variants in~~

~~all candidate genes.~~
~~Subtask 5: Share the candidate genes with ICPCG~~
Subtask 6: Publish validated candidates

Task 1. Received UKBiobank approval and data
Task 3 and 4. Performed UKBiobank case/control association and identified 875 candidate variants with data in UKBiobank and found 12 with significant association with prostate cancer.
Task 5. Analysis underway to identify additional carriers in existing sequence data for other studies.
Task 6. Manuscript is in preparation presenting the 571 candidate variants and some excellent example high-risk pedigrees.

• What opportunities for training and professional development did the project provide?
No formal opportunities yet.

• How were the results disseminated to communities of interest?
Manuscript submitted to Cancer Research for public dissemination

• What do you plan to do during the next reporting period to accomplish the goals and objectives?

4. IMPACT

Nothing to report yet.

What was the impact on the development of the principal discipline(s) of the project?

Large number of candidate predisposition variants presented to the scientific community to pursue

What was the impact on other disciplines?

The methods and tools developed have been extended to other phenotypes.

What was the impact on technology transfer?

Nothing to report.

What was the impact on society beyond science and technology?

Improved understanding of prostate cancer genetics

5. CHANGES/PROBLEMS

Nothing to report.

Changes in approach and reasons for change

Given our finding that validation of risk association is a more powerful next step for identification of predisposition variants than investigation of all possible variants in candidate genes that may not all be associated with risk, we amended Aim 3. We did not sequence the 400 candidate genes with segregation evidence, but rather sought validation of association with risk. This required development of new software but resulted in a strong set of candidate variants to be published and followed in future projects.

Actual or anticipated problems or delays and actions or plans to resolve them

Changes that had a significant impact on expenditures

None

Significant changes in use or care of human subjects, vertebrate animals, biohazards, and/or select agents

None

6. PRODUCTS

Nothing to report.

7. PARTICIPANTS & OTHER COLLABORATING ORGANIZATIONS

What individuals have worked on the project?

Name: **Lisa A. Albright**
Project Role: PI/PD
Researcher Identifier: 0000-0003-2602-3668
Nearest person month worked: 1.2
Contribution to Project: Dr. Cannon-Albright has directed the course of the research, selection of pedigrees/cases, and methods of analysis, as well as supervised the day to day activities of the research team.
Funding Support: N/A

Name: **Craig C. Teerlink**
Project Role: Co-Investigator
Researcher Identifier: 0000-0002-1992-2326
Nearest person month worked: 1.8
Contribution to Project: Dr. Teerlink has been responsible for oversight of data quality control and genetic analyses (including bioinformatics analysis of sequence data) and method development/testing
Funding Support:

Name: **Steven Backus**
Project Role: Computer Professional
Researcher Identifier: N/A
Nearest person month worked: 0.8
Contribution to Project: Mr. Backus has performed data management, data extraction, security, and data storage.

Funding Support: N/A

Name: **James Farnham**
Project Role: Applied Biostatistician
Researcher Identifier: 0000-0002-8213-949X
Nearest person month worked: 0.8
Contribution to Project: Mr. Farnham has been responsible for the generation and quality control of all data files for pedigree/case selection and tracking.
Funding Support: N/A

Name: **Kim Nguyen**
Project Role: Laboratory Specialist
Researcher Identifier: N/A
Nearest person month worked: 2.4
Contribution to Project: Mr. Nguyen has been responsible for testing the concentration and quality of DNA, preparing stored samples, performing quality controls for all samples, and the appropriate storage and inventory and shipping of all biospecimen samples.
Funding Support: N/A

Name: **Jeff Stevens**
Project Role: Laboratory Specialist
Researcher Identifier: N/A
Nearest person month worked: 3
Contribution to Project: Mr. Stevens is responsible for bioinformatic analysis of the sequence data.
Funding Support: N/A

Has there been a change in the active other support of the PD/PI(s) or senior/key personnel since the last reporting period?
Some grants have ended and others have been started; overall FTE on this project was not impacted. It was in an unfunded continuation so that we could finish analysis and our manuscript-no support has been requested for this.

What other organizations were involved as partners?

none

8. SPECIAL REPORTING REQUIREMENTS

none

9. APPENDICES

None

REFERENCES

Albright F, **Stephenson RA, Agarwal N**, & **Cannon-Albright LA.** (2016, Aug 16). Relative risks for lethal prostate cancer based on complete family history of prostate cancer death. *The Prostate*, doi: 10.1002/pros.23247. [Epub ahead of print]

Nelson Q, Agarwal N, Stephenson R, & **Cannon-Albright LA**. (2013). A population-based analysis of clustering identifies a strong genetic contribution to lethal prostate cancer. *Front Genet*, *4*, 152.

High-risk Pedigree Study Identifies *LRBA* (rs62346982) as a Likely Predisposition Variant for Prostate Cancer

Lisa A Cannon-Albright[1,2,3], Jeff Stevens[1], Julio C. Facelli[3], Craig C Teerlink[1,2], Kristina Allen-Brady[1], Neeraj Agarwal[4,5]

1 Genetic Epidemiology, Department of Internal Medicine, University of Utah School of Medicine

2 George E. Wahlen Department of Veterans Affairs Medical Center, Salt Lake City, Utah

3 Department of Biomedical Informatics and Clinical and Translational Science Institute, Spencer Fox Eccles School of Medicine, University of Utah

4 Huntsman Cancer Institute, Salt Lake City, Utah

5 Division of Oncology, Department of Internal Medicine, University of Utah School of Medicine

Keywords: prostate cancer, predisposition, high-risk pedigree, UPDB, *LRBA*

Running Title: *LRBA* is a Likely Predisposition Variant for Prostate Cancer

COI:

L.A.C-A., J.S., J.C.F., C.C.T., K.A.B. declare no potential conflicts of interest.

N. A. (lifetime disclosures): Consultancy to Astellas, Astra Zeneca, Aveo, Bayer, Bristol Myers Squibb, Calithera, Clovis, Eisai, Eli Lilly, EMD Serono, Exelixis, Foundation Medicine, Genentech, Gilead, Janssen, Merck, MEI Pharma, Nektar, Novartis, Pfizer, Pharmacyclics, and Seattle Genetics. Research funding to Neeraj Agarwal's institution: Arnivas, Astellas, Astra Zeneca, Bavarian Nordic , Bayer, Bristol Myers Squibb, Calithera, Celldex, Clovis, Crispr, Eisai, Eli Lilly, EMD Serono, Exelixis, Genentech, Gilead, Glaxo Smith Kline, Immunomedics, Janssen, Lava, Medivation, Merck, Nektar, Neoleukin, New Link Genetics, Novartis, Oric, Pfizer, Prometheus, Rexahn, Roche, Sanofi, Seattle Genetics, Takeda, and Tracon.

Corresponding Author:
Lisa A. Cannon-Albright, PhD
Genetic Epidemiology
Williams Building
295 Chipeta Way
Salt Lake City, UT 84108
Lisa.albright@utah.edu

Word count: 5312
Number of Figures: 8
Number of Tables: 1 (supplementary)

ABSTRACT

There is evidence for contribution of inherited factors to prostate cancer, and more specifically to lethal prostate cancer, but few responsible genes/variants have been identified.

We examined genetic sequence data for 51 affected cousin pairs who each died from prostate cancer and who were members of high-risk prostate cancer pedigrees in order to identify rare variants shared in the cousins as candidate predisposition variants. Candidate variants were tested for association with prostate cancer risk in UKBiobank data. Candidate variants were also assayed in 1,195 additional sampled Utah prostate cancer cases. We used 3D protein structure prediction methods to analyze structural changes and provide insights into mechanisms of pathogenicity.

Almost 4,000 rare (<0.005) variants were identified as shared in the 51 affected cousin pairs. One candidate variant was also significantly associated with prostate cancer risk among the 840 of the variants with data in UKBiobank, in the gene *LRBA* (p=3.2e-5; OR=2.09). The rare risk variant in *LRBA* was observed to segregate in 5 pedigrees. The overall predicted structures of the mutant protein do not show any significant overall changes upon mutation, but the mutated structure loses a helical structure for the two residues after the mutation.

This unique analysis of closely related individuals with lethal prostate cancer, who were members of high-risk prostate cancer pedigrees, has identified a strong set of candidate predisposition variants which should be pursued in independent studies. Validation data for a subset of the candidates identified is presented, with strong evidence for a rare variant in *LRBA*.

INTRODUCTION

There is evidence for an inherited contribution to predisposition to prostate cancer, and stronger evidence for lethal prostate cancer (Cannon et al., 1982; Pritchard et al., 2016; Rantapero et al., 2020; Albright et al., 2015; Albright et al., 2017). While hundreds of common variants in low-risk genes have been associated with risk for prostate cancer (e.g. Schumacher et al., 2018), few rare variants in moderate to high penetrance genes have been identified, and they explain little of familial prostate cancer (Castro and Eeles, 2012; Xu et al., 2013). The genes most consistently recognized to affect prostate cancer risk include *ATM, BRCA1, BRCA2, CHEK2, HOXB13, MLH1, MSH2, MSH6, PMS2*, and *PALB2* (Das et al., 2019).

We sequenced germline DNA from 51 pairs of cousins who both died from prostate cancer who were also members of pedigrees with a significant excess of prostate cancer cases. We identified the set of rare coding and non-coding variants shared in these high-risk, lethal prostate cancer-affected cousin pairs as strong candidate variants for predisposition to prostate cancer. Confirmation of significant association with prostate cancer risk in an independent population, and observation of segregation of variants with prostate cancer in multiple high-risk pedigrees provided additional validation for a subset of the candidates considered. We used 3D protein structure prediction methods to analyze structural changes in one outstanding candidate variant that may provide insights on mechanisms of pathogenicity (Hernandez and Facelli, 2021).

DATA/METHODS

Utah Population Data Base (UPDB) The UPDB consists of the genealogy of the majority of the Utah population, from its founders in the mid 1800s to their modern-day descendants. The genealogy has been linked to the Utah Cancer Registry (UCR), and to Utah death certificates coded with International Classification of Disease causes of death from 1904. The statewide UCR was created in 1966 and has been an NCI Surveillance, Epidemiology, and End-Results (SEER) registry since 1973. It records and tracks all independent, primary cancers diagnosed or treated in Utah. Of the over 3 million individuals with genealogy of at least 3 generations linking to Utah founders, there were 7,727 individuals whose linked Utah death certificate indicated prostate cancer as a cause of death; 6,328 of these individuals also had a UCR record confirming a prostate cancer diagnosis.

Using the combined UPDB genealogy and UCR data, the Genetic Epidemiology Program at the University of Utah has previously identified and sampled ~2,500 prostate cancer cases and ~7,500 relatives who belonged to approximately 500 Utah pedigrees that each had an excess of prostate cancer cases among the descendants (high-risk pedigrees). 422 of these sampled prostate cancer cases were since identified to have died from their prostate cancer based on the presence of prostate cancer as a cause of death on their linked Utah death certificate. These men were termed lethal prostate cancer (LPrCa) cases. All genetic relationships among these cases were identified in the UPDB genealogy to identify descending pedigrees.

Affected LPrCa cousin pairs in high-risk pedigrees All independent descending clusters (pedigrees) including two or more of the 422 sampled men who died of their prostate cancer (LPrCa cases) were identified in the UPDB. No pedigrees were completely overlapping, but an LPrCa case could belong to more than one pedigree through different ancestors. Each pedigree that included at least two sampled LPrCa cases was tested for a significant excess of prostate cancer cases as follows. All individuals with biological sex of male and at least 3 generations of genealogy linking to Utah founders were assigned to a cohort based on five-year birthyear, and birth state (Utah or not). Prostate cancer cohort-specific rates were estimated as the number of prostate cancer cases in each cohort divided by the number of males with genealogy data in the cohort. Each pedigree was tested for an excess of prostate cancer cases by comparing the observed number of prostate cancer cases among the descendants to the expected number of prostate cancer cases among the descendants. The expected number of cases among the descendants was estimated by summing the cohort-specific rate of prostate cancer

for all male descendants. A pedigree was termed high-risk for prostate cancer in the presence of a significant excess of observed cases ($p<0.05$). For each of these high-risk prostate cancer pedigrees identified, which also contained at least 2 sampled LPrCa cases, we selected those related LPrCa cases who were related as first or second cousins for sequencing. Each of the 102 LPrCa cases in the 51 cousin pairs identified had a stored sample of germline DNA extracted from whole blood available for whole genome sequencing.

Whole Genome Sequencing/Identification of candidate predisposition variants /Assay development

The 102 samples of high molecular weight DNA (>30Kb in length) were whole genome sequenced at MedGenome, Foster City, Ca. (http://www.medgenome.com) utilizing 10X Genomics (https://www.10xgenomics.com/) long read sequence technology. Genomes were aligned to GRCh37 with 10X Genomics LongRanger.2.2.2 software. VCF's were merged with BCFTOOLS. Merged VCF was annotated with ANNOVAR (Wang, K. 2010; https://academic.oup.com/nar/article/38/16/e164/1749458).

A total of 114,513 rare (MAF < 0.005 in GnomAD 2.1) coding (nonsynonymous, frameshift, startloss, startgain, stoploss) variants were identified in the 102 LPrCa cases. Of these, 17,859 variants had allele counts >1; and 3,251 of these rarecoding variants (in 1,762 genes) were shared within at least one LPrCa cousin pair. Non-coding and UTR variants were prioritized after RegulomeDB scoring. The RegulomeDB score represents evidence that each variant functions in a regulatory role (ranging from 1, indicating strong evidence, to 6, indicating weak evidence). A total of 546 rare, non-coding variants with RegulomeDB scores ranging between 2a-4, and shared in at least one cousin pair, were identified.

These 3,797 (3,251 coding + 546 non-coding) rare shared candidate predisposition variants were submitted for assay design using the Illumina iSelectHD Custom Genotyping BeadChips (https://www.illumina.com/products/by-type/informatics-products/designstudio.html). 200 base pairs of genomic sequence surrounding each variant was used for the design process. Assays were manufactured for 3,559 of these candidate variants (238 failed to design). Illumina iSelectHD custom beadchips were processed at the University of Utah HSC Genomics Core.

Segregation of candidate predisposition variants in pedigrees

1,298 additional previously sampled Utah prostate cancer cases were identified for assay with the candidate predisposition variants to determine segregation of the candidate variants; these assayed prostate cancer cases included the original affected cousin pairs (n=102); all other sampled prostate cancer cases whose death certificate indicated prostate cancer as

a cause of death (n=320 LPRCA cases); all sampled prostate cancer cases who were first-, second-, or third-degree relatives of the affected cousin pair cases (n=168) and all sampled prostate cancer cases who were first-, second-, and third-degree relatives (n=307) of these 168 cases; and all metastatic prostate cancer cases recruited by author N.A. in the Huntsman Cancer Institute Urology Clinic (n=401). DNA for 1,195 of these 1,298 sampled prostate cancer cases passed quality control and was assayed for the candidate variants to test for segregation and to identify additional carriers.

Case/Control Risk Association in UKBiobank

These rare, shared candidate predisposition variants that were identified in the sequencing experiment and had data in UKbiobank were analyzed for association with prostate cancer risk in a set of 7,764 Caucasian prostate cancer cases and 1:1 ancestrally matched controls from among the UKBiobank's 488,377 total subjects genotyped on the Illumina OmniExpress SNP array (Sudlow 2015). UKBiobank case and control subjects were matched via principal components (**PC**) using ~27K independent markers that excluded several genomic regions known to adversely affect PC analysis (Abraham 2017). FLASHPCA2 software was used to generate eigenvectors for control selection (Abraham 2017). Controls were selected from among 64,284 Caucasian UKBiobank subjects who were male, over age 70 years of age, and had no cancer diagnosis. One control, representing the nearest neighbor based on Euclidean distance of the first two PCs, was selected for each case. 129 outlier cases and controls were removed, leaving 7635 cases and controls.

UKBiobank Imputation The selected UKBiobank case and control subjects were imputed to ~40M SNP markers using the Haplotype Reference Consortium's (HRC) 67K background genomes (Haplotype Reference Consortium, 2016). Beginning with 784,256 observed SNP genotypes, pre-imputation quality control using PLINK software (Purcell 2007) required sample genotyping >98% (no subjects removed); a total of 353,578 markers were removed by filtering for genotyping call rate <98%, HWE $p<1e^{-5}$, MAF<0.005, duplicated position in the HRC's reference genome, or site not included in the HRC's reference genome. The remaining 430,678 SNPs were converted to human genome B37 forward strand orientation using GenotypeHarmonizer software (Deelen 2014) and served as the basis for imputation. Imputation was performed with EAGLE v2.3 software for phasing (Loh 2016) and MINIMAC3 software for imputation (Das 2016). Post-imputation quality control included removing markers with imputation information score (INFO-r^2) < 0.7 [Ziv et al., 2015; Schumacher et al., 2018; Huyghe et al., 2019].

Protein Structure Prediction

LRBA (Lipopolysaccharide-Responsive, Beige-like Anchor Protein) is a very large protein with 2863 amino acids that is not tractable using existing 3D structure prediction methods. The only experimental structure available for LRBA is an X-ray structure for positions 2076-2489 (Lopez-Herrera et al, 2012; Gebauer et al., 2004). There are two isoforms; here we consider the isoform 1 as the canonical sequence. For this sequence the single point mutation considered here is T2533P. The protein has multiple domains, and the mutation in question, T2533P, is localized between the second BEACH (2200-2489) domain and the second WD2 (2591-2633) domain. To perform 3D structure prediction with I-TASSER structure prediction software using full homology modeling (Zhang 2008, Zhang 2009, Roy 2010, Yang 2015) we constructed a model including the PH domain (2073-2181), the BEACH domain (2200-2489) and the WD2 – WD6 domains (2591-2858). For this model, the amino acid substitution under consideration is T461P. The I-TASSER predictions resulted in a set of three candidate structures with C-scores of -3.86, -3.85, and -3.93, for the wild-type sequence, and two with C-scores of -3.74 and -3.80 for the mutant sequence. All the structures obtained were visualized and analyzed using Chimera (Pettersen 2004).

This project was approved by the University of Utah Institutional Review Board and has approval from the Resource for Genetic Epidemiology which has oversight for the UPDB. Informed consent was obtained from each subject.

Data Availability: All study subjects have relatives within the study set, which may allow identification from genetic data. Use of UPDB relationship data requires separate project application and approval. Interested researchers can contact the authors and data will be made available to those who obtain appropriate approvals.

RESULTS

Case/Control Risk Association

Only 840 of the 3,797 candidate predisposition variants had some data present in UKBiobank data, likely due to the very low frequency (<0.005). Only one of the 840 variants was significantly associated with prostate cancer risk after correction for multiple testing (p<0.05/840 or p<5.95E-05), in the gene *LRBA* (p=3.2e-5; OR=2.09).

Assay of candidate variants in 1,195 prostate cancer cases

102 of the 1,195 prostate cancer cases assayed for the candidate variants were the original target pair LPrCa cases; these cases were assayed in order to validate performance of the assays. Of the 3,559 variants assayed, 2,035 failed to identify any heterozygous (het) carriers with the assay. These variants were considered to have failed performance, since the assay should have at least identified the original cousin pair carriers. For 562 of the variants, although some het carriers were identified, the assay did not identify the original cousin pair as carriers, so these variants were also considered to have failed the assay. For 78 of the variants, the only carriers identified in the assay were the original cousin pair; these variant assays were considered to have performed appropriately, but they did not identify any additional carriers of the candidate variants. This finding was not unexpected given the rare nature of the variants, but these variants did not provide any additional evidence for segregation and were not further considered here. For the remaining 884 of the variant assays, all original cousin pair carriers were identified, as well as additional het carriers. These 884 candidate variants were considered the best rare candidate lethal prostate cancer predisposition variants. These variants were examined further in the 1,195 prostate cancer cases assayed, to determine if segregation to other related prostate cancer cases was observed.

All het carriers for each of these 884 rare candidate variants were analyzed by identifying all genealogical relationships among the het carriers using the UPDB linked genealogy. This analysis identified many more clusters than the original 51 pedigrees analyzed since all genetic relationships among all cases were considered; 1,959 total clusters (pedigrees) that included 3 or more related het carriers for at least one of these 884 variants were identified. Although none of these pedigrees overlapped completely, a case could belong to more than one.

These 884 candidate variants were all originally selected for their low frequency (<0.005) based on review of GnomAD. However, some of the variants were observed at much higher frequency than 0.005 in the assay of 1,195 prostate cancer cases. It is possible that these variants might actually represent truly rare variants that were observed in much higher frequencies in the assay of prostate cancer cases because they are, in fact, strongly associated with prostate cancer risk. However, these results might also be due to inaccuracies in reporting of frequencies for what were, in fact, more common variants, or due to assay performance failure. We could not confirm the lack of a data problem for these more commonly observed variants, so we elected to exclude consideration of any of the 1,959 variant-segregating pedigrees for any variant for which more than 10% (n>110 carriers) het carriers were observed in the assay of 1,195 prostate cancer cases.

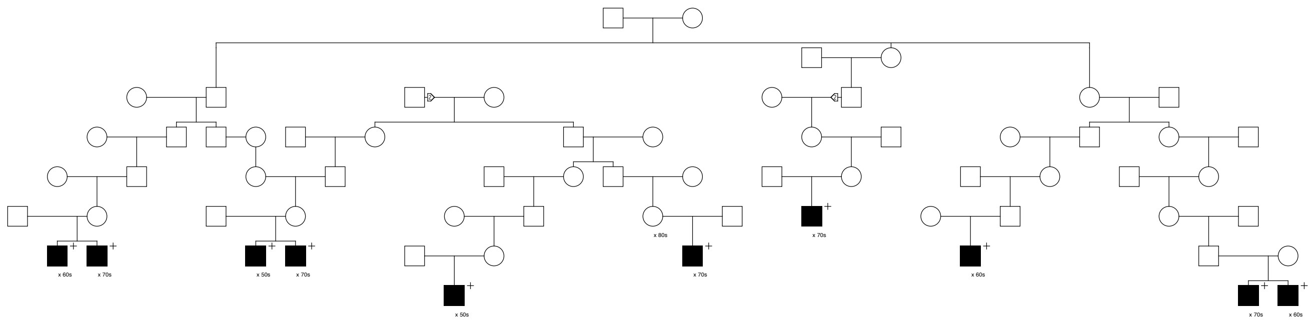
This left 1,070 pedigrees of interest that included at least 3 related het carriers of one of the candidate variants. Because some of these clusters of het carriers might represent random clusters, rather than high-risk prostate cancer pedigrees, we additionally excluded those variant-sharing pedigrees that did not exhibit a significant excess of prostate cancer cases ($p<0.05$) among the descendants of the common founding ancestor of all the related carriers. This left 814 high-risk prostate cancer pedigrees including 559 different rare candidate variants in 398 genes that showed evidence of segregation of the candidate variant with prostate cancer. The list of these 559 rare candidate predisposition variants is provided in Supplemental Table 1, which includes the UKBiobank association test results for those variants with data, and the ClinVar diagnosis and interpretation data, as well as identifying those variants associated with a prostate cancer pathway in Ingenuity.

Supplemental Table 1 Legend:

Table 1. Summary data for the 559 candidate prostate cancer predisposition variants with evidence of segregation.

Only one of these 559 candidate variants was significantly validated for risk association with prostate cancer (*LRBA*). This *LRBA* variant (rs62346982) is classified in ClinVar with "conflicting interpretations of pathogenicity". The *LRBA* variant was observed to segregate with prostate cancer in a total of 5 high-risk prostate cancer pedigrees. Figure 1 shows segregation of the rare *LRBA* variant in the largest *LRBA*-segregating pedigree identified. The founder of this pedigree was born in the late 1700s in Vermont and has almost 32,000 descendants in UPDB, with a total of 230 prostate cancer cases observed and 174.9 expected ($p=4.0\ e^{-5}$) among all descendants (not shown); variant-carrying cases are shown. A multiple marriage in the third generation is noted with a triangle on the two horizontal marriage lines, variant carriers are noted by "+", and the decade of prostate cancer diagnosis is noted below each case. It must be noted that complete prostate cancer case diagnosis data for Utah is only available from 1973, which explains why cases are only observed in the bottom, most recent, few generations of the high-risk pedigrees shown.

Figure 1. One of the high-risk prostate cancer pedigrees segregating rare *LRBA* candidate prostate cancer predisposition variant.



Protein Structure Prediction

Lipopolysaccharide-responsive and beige-like anchor protein (*LRBA*), the single candidate predisposition variant identified in a cousin pair and also confirmed with significant association to prostate cancer in the UKBiobank case/control analysis, is involved in coupling signal transduction and vesicle trafficking to enable polarized secretion and/or membrane deposition of immune effector molecules (this function was assigned by similarity). It is involved in phagophore growth during mitophagy by regulating ATG9A trafficking to mitochondria. A mutation of this protein has been associated with CVID8, an autosomal recessive immunologic disorder associated with defective B-cell differentiation.

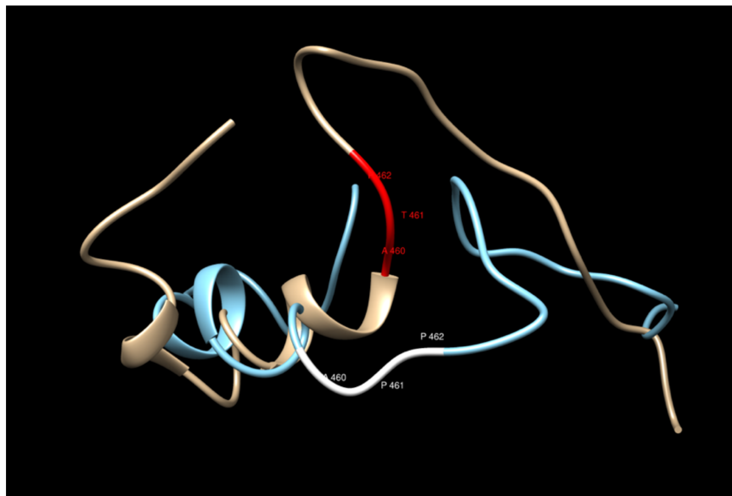
The variant considered here corresponds to rs62346982. PolyPhen-2 predictor predicts this variant as benign with a score of 0.167 (sensitivity 0.92 and specificity 0.87) (Adzhubei et al 2010). The overall predicted sequences do not show any significant changes upon the amino acid replacement T461P in the model considered here, but careful inspection of the region close to the replacement (440-480) shows that the mutated structure loses a helical structure two residues after the replacement. (See Figures 2 and 3). The water accessibility does not change upon mutation, keeping this region buried.

Figure 2. Secondary structure comparison in the region of the T461P replacement.

LRBA-WT-PH-WD6	CCCCCCCCCSSSSCCCCCCCCCCHHHCCCCCCSSSSCCCC
LRBA-MT-PH-WD6	CCCCCCCCCCSSCCCCCCCCCCCCSCCCCCCSSSCCCCCC

440 480

Figure 3. 3D Structure Comparison in the region of the T461P replacement. The bronze/red structure corresponds to the WT and the blue/white one to the MT.

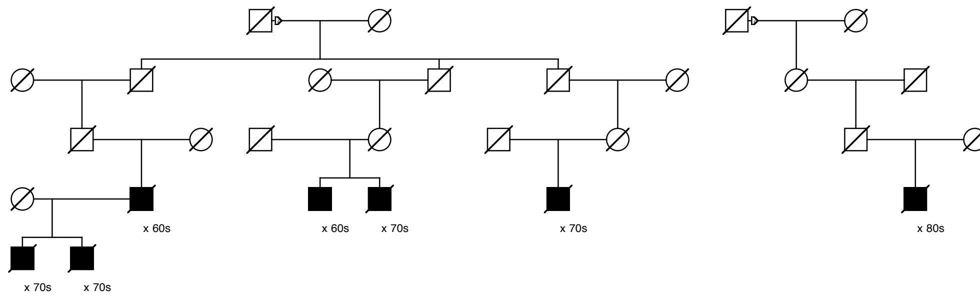


Consideration of other likely candidate predisposition variants identified

In addition to the significant evidence identifying the rare *LRBA* variant as a strong candidate prostate cancer predisposition variant, this study identified a large set of other rare candidate predisposition variants, many of which are worthy of further consideration. We have used various criteria to select a subset of these LPrCa candidate predisposition variants for presentation in more detail below. These criteria include: i) known cancer pathogenic variants, ii) variants in genes recognized to be implicated in breast cancer (BROCA genes), and iii) non-coding variants with suggestive Regulome DB scores.

(i) Cancer pathogenic variants Only one candidate variant that was classified as "pathogenic/likely pathogenic" in ClinVar was identified, in *MUTYH* (seq_1_45797228_C-T_MUTYH). This variant was identified in 29/1,195 of the assayed cases and segregated in 3 high-risk prostate cancer pedigrees including 6, 3, and 3 sampled, related cases, respectively. Figure 4 below shows the largest of these 3 pedigrees, with full shading for prostate cancer cases; the 6 assayed het carriers are identified with "+". One inferred case carrier (not sampled) is also shown; decade of age at diagnosis of prostate cancer is shown below each case. The founder of this pedigree was born in New York in the early 1800s, with a total of almost 16,000 descendants included in UPDB; only the descending lines to the 6 identified het carriers are shown; in the UPDB a total of 114 prostate cancer cases were observed among all descendants with 68.3 cases expected, p = $2.7e^{-7}$. The founding male had 2 marriages, indicated with a small triangle on the top horizontal marriage lines.

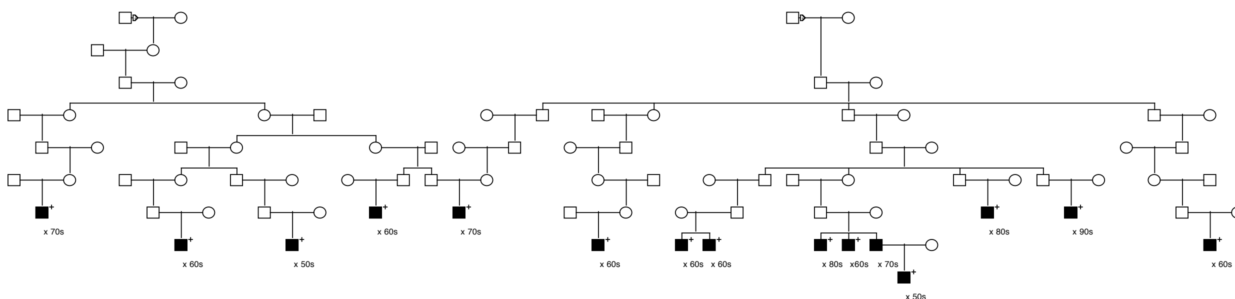
Figure 4. High-risk prostate cancer pedigree segregating rare pathogenic *MUTYH* variant.



Eleven candidate variants found to segregate in high-risk pedigrees were classified with "conflicting interpretations of pathogenicity" in ClinVar, these included variants in *DUOX2, LRBA (discussed previously), MYO3A, EVC, CP, FBN1, MYH11, KCND3, ACNT2, ADAM9* and *PTCH1.* The largest pedigree observed for the "conflicting" variant identified for *DUOX2* (n=14 case carriers) is shown in Figure 5, and the largest pedigree observed for *LRBA* (n=10) was shown in Figure 1. *PTCH1* was also selected as a BROCA gene, and a pedigree is shown later. A variant in *MSH6* was the only variant of 14 that were classified in ClinVar as of "uncertain significance", that also reported a cancer-related clinical diagnosis; it was also selected as a BROCA variant and is shown later.

Figure 5 shows the segregation of the rare *DUOX2* variant in a high-risk prostate cancer pedigree. The founder of this pedigree was born in the mid 1700s in North Carolina and has almost 153,000 descendants in UPDB; 897 prostate cancer cases were observed among all descendants, with 811.5 expected, p=0.0016). The male founder of this pedigree had 2 spouses, shown with a triangle on the top horizontal marriage lines. All 14 prostate cancer case variant carriers are shown with "+", one inferred case (father of an affected carrier who was not sampled) is also shown.

Figure 5. High-risk prostate cancer pedigree segregating a rare *DUOX2* variant.



ii) Variants in recognized BROCA genes

Seven of the rare candidate predisposition variants found to segregate in prostate cancer cases who were members of high-risk prostate cancer pedigrees were in known BROCA genes, including the *MUTYH* variant already discussed, as well as variants in *APC* (ClinVar: conflicting interpretations of pathogenicity), *BRCA1* (benign), *MSH6* (uncertain significance), *PTCH1* (conflicting), *SDHC* (not classified), and *SLX4* (benign/likely benign). The pedigrees segregating the rare variants in *BRCA1, MSH6* and *PTCH1* are shown in Figures 6-8.

Figure 6 shows the single high-risk prostate cancer pedigree segregating a *BRCA1* variant (seq_17_41228587_T-G_BRCA1). This variant has been classified as “benign” in ClinVar, based on the breast cancer phenotype. The founder of this pedigree was born in the early 1800s in Canada and has a total of over 2,300 descendants in the UPDB; a total of 25 prostate cancer cases were observed among all descendants, with 12.1 expected, $p=7.6e^{-4}$. The 3 identified variant carriers are shown with “+”, the unsampled pedigree member father of one of the affected pair cases was also diagnosed with prostate cancer is shown. This pedigree also has a significant excess of breast cancer (18 observed, 11.3 expected, p=0.04. There are no identified breast-cancer-affected descendants of the first son of the founder pair shown in Figure 6, but the second son of the founder has 8 descendants diagnosed with breast cancer (breast cancer cases not shown).

Figure 6. High-risk prostate cancer pedigree segregating rare *BRCA1* variant.

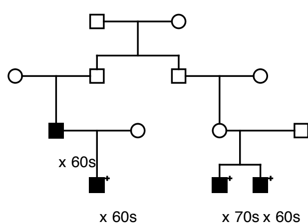


Figure 7 shows the single high-risk pedigree segregating a rare *MSH6* variant (seq_2_48026979_A-C_MSH6). This variant is classified “uncertain” in ClinVar. The founder of this pedigree was born in the early 1800s in Massachusetts and has almost 23,000 descendants in UPDB; 166 prostate cancer cases were observed among all descendants, with 113.0 expected ($p=1.8e^{-6}$); 129 breast cancers were observed with 109.8 expected (p=0.04); 23 ovarian cancers were observed with 13.9 expected (p=0.016). Neither colorectal nor endometrial cancers were observed in excess among the descendants in this pedigree. Prostate cancer case variant carriers are shown with “+”; one unsampled father of a carrier was also

diagnosed with prostate cancer as shown. Two of the prostate cancer case carriers had an additional cancer diagnosis (one kidney and one chronic lymphocytic leukemia), not shown.

Figure 7. High-risk prostate cancer pedigree segregating a rare *MSH6* variant.

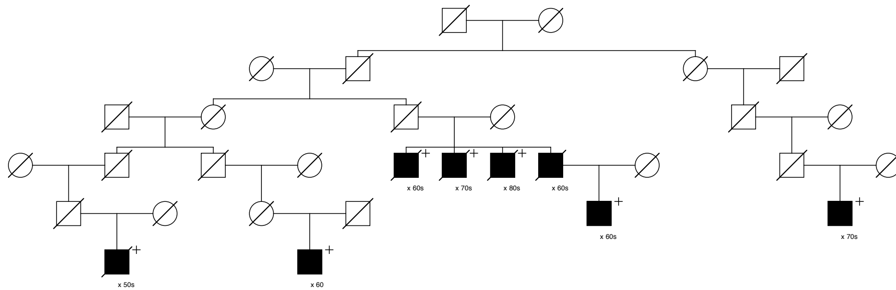
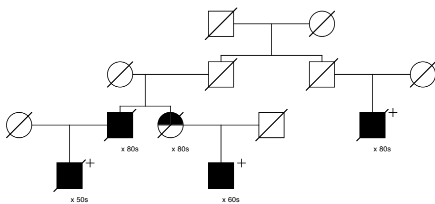


Figure 8 shows the single high-risk pedigree segregating a rare *PTCH1* variant (seq_9_98240378_C-T_PTCH1). This variant is classified as "conflicting" in ClinVar. The founder of this pedigree was born in the mid 1800s with no recorded birthplace and has almost 4,600 descendants in UPDB; a total of 33 prostate cancers were observed among all descendants, with 20.0 expected (p=0.0047). This pedigree also exhibits a significant excess of breast cancer; a total of 37 breast cancers were observed in the entire descending pedigree, with 23.1 expected, p=0.0045. One of these breast cancer cases was the mother of a carrier, shown with half shading.

Figure 8. High-risk prostate cancer pedigree segregating rare *PTCH1* variant.



<u>iii) Non-coding variants with suggestive Regulome DB scores</u>

Regulome DB (RGDB) scores ranging from "2a" to "4" were observed among the rare candidate variants. The two variants with the most evidence for a regulatory role based on RGDB score were in *PAX6* and *ZWILCH*, both were scored "2a". Two pedigrees with the *PAX6* variant were observed with 3 het carriers each, and one pedigree with the *ZWILCH* variant also included 3 het carriers; these pedigrees are not shown.

Multiple variants with a RGDB score of “2b” were identified, but pedigrees are not shown here. The founder of the largest pedigree, with 11 het carriers of an *APCDD1L-DT* variant with score “2b”, was born in the late 1700s in Massachusetts and has over 36,000 total descendants with 268 prostate cancer cases observed among all descendants, and 167.8 expected ($p=6.6e^{-13}$).

DISCUSSION

It is estimated that 10-20% of prostate cancer cases occur in a familial context, (Cannon et al., 1982; Carter et al., 1993). Genetic predisposition to prostate cancer development has been associated with both rare variants in moderate to high penetrance genes, as well as with common genetic alterations in low-risk genes (e.g., Castro and Eeles 2012; Schumacher et al., 2018). Almost 200 common variants have been identified in large case-control cohorts, but other evidence for their direct association with prostate cancer is lacking, and only a small fraction of familial prostate cancer is associated with known rare predisposition variants.

High-risk pedigree studies remain a powerful mechanism for identification of predisposition genes and variants (Manolio et al., 2009; Wijsman, 2012, Ott et al., 2015, Terwilliger and Goring, 2009). This has proven true for prostate cancer (Ewing et al., 2012; Cannon-Albright et al., 2021), although such high-risk prostate cancer pedigrees remain infrequently presented. Here we have taken advantage of unique Utah resources, combined with an unusual and powerful study design that includes sampled affected cousin pairs, to generate, and begin to evaluate, a strong list of candidate predisposition variants for prostate cancer.

Extensive linked genealogic and disease registries existing in Utah have been used to identify and study thousands of Utah high-risk pedigrees (Cannon-Albright, 2008). We have previously used this same sequencing approach in affected cousins belonging to high-risk pedigrees to identify multiple candidate predisposition variants for several different phenotypes, including *GOLM1* for melanoma, *ERF* for bladder cancer (Cannon-Albright et al., 2021); *FANCM* for colorectal cancer (Cannon-Albright et al., 2020), *MEGF* for osteoporosis (Teerlink et al., 2020), *HOXC4* for Chiari Malformations (Brockmeyer et al., 2022), and multiple candidates for Alzheimer’s (Teerlink et al., 2022) and exceptional longevity (Miller et al., 2020).

Using the combined genealogy and cancer registry resource in the UPDB, the strongest evidence for an inherited contribution to prostate cancer has been shown for the subset of lethal prostate cancer cases (Nelson et al., 2012). The present analysis of available germline DNA for 51 pairs of cousins who each died of prostate cancer and belonged to a pedigree with a significant excess of prostate cancer among the descendants has identified thousands of rare, shared coding and non-coding variants in those cousins, each of which represents a candidate predisposition variant for prostate cancer which can be explored further. Since validation of segregation and risk association was based on the prostate cancer phenotype, rather than on the more rare (and remaining unknown for most cases) phenotype of lethal prostate cancer, conclusions may therefore be limited to prostate cancer, rather than lethal prostate cancer, pending further study.

One outstanding candidate, in *LRBA*, was the only candidate predisposition variant to be further validated in the independent UKBiobank population for association with prostate cancer risk; this variant was found to segregate in five independent Utah high-risk pedigrees. While protein prediction modeling for this variant did not show any significant changes, the mutated structure does lose a helical structure two residues after the replacement. (Figures 2 and 3). The COACH results did not find any binding site in the region considered for this protein, nor there is any ligand information from experimental studies, therefore it is not clear how to relate the additional helix structure with possible increase of loss of function of the protein (Yang et al., 2013). Further study of the specific Utah high-risk pedigrees identified here, as well as analysis in other independent populations of prostate cancer cases can further clarify whether, and how, this specific variant is specifically associated with increased risk for prostate cancer, or lethal prostate cancer.

While not all of the thousands of rare, shared candidate prostate cancer predisposition variants identified in this study have been reviewed in detail due to limitations of time and space, several specific subsets of variants were presented in more detail. Subsets considered included known cancer pathogenic variants, BROCA genes, and non-coding variants with suggestive Regulome DB scores. The identification of multiple strong candidate variants in these subsets, some of which represented more than one subset (e.g., *PTCH1* or *MSH6*) demonstrates the power of this approach, and serves as validation of this approach to identify both known, and new predisposition variants.

The strengths of this study include the large number of affected cousin pairs analyzed, leading to a large number of rare, shared candidate predisposition variants identified. Strength also comes from the analysis of the phenotype of lethal prostate cancer, which was confirmed both by a linked Utah death certificate for prostate cancer, and validation of the cancer diagnosis within the linked Utah Cancer Registry. The available stored germline DNA from thousands of Utah

prostate cancer cases, many of whom were related to the sequenced, affected cousin pairs, allowed further validation of candidate variants by the confirmation of segregation of the variants. Limitations of this study include the censoring of some death, cancer, and genealogy data which may have occurred through errors of reporting or record linking, or through lack of records. Also of note is that the founding population of Utah is largely Northern European (Cannon-Albright et al., 2005), and while immigration has been significant, these results may only specifically apply to this limited population, and will require confirmation in other, independent, and more diverse populations. Confirmation of these results for the lethal prostate cancer phenotype will have to be made when more data for this more extreme phenotype is available for this, and other independent populations.

Analysis of high-risk cancer pedigrees identified in this powerful Utah resource previously provided the identification of *BRCA1* and *BRCA2* (Miki et al., 1994; Tavtigian et al., 1996), and *CDKN2A* (Kamb et al., 1994), which remain the most common cancer predisposition genes to be identified. Large scale case and pedigree studies strongly suggest that other common cancer predisposition genes may not remain to be identified, and that, rather, most familial cancer predispositions might be the result of many, varied, rare predisposition genes and variants (Terwilliger and Goring, 2009). Whether or not this is the case, studies such as this one, that analyze related affected individuals within a large number of high-risk pedigrees, have shown the strong potential to identify many candidate predisposition genes and variants for many different phenotypes. Such studies should be pursued, and the candidate predisposition variants identified are worthy of further exploration.

AUTHOR CONTRIBUTIONS

Conceptualization, Lisa A. Cannon-Albright; Data curation, Jeff Stevens, Julio Facelli and Craig Teerlink; Formal analysis, Lisa A. Cannon-Albright, Jeff Stevens, Julio Facelli and Craig Teerlink; Funding acquisition, Lisa A. Cannon-Albright and Neeraj Agarwal; Investigation, Kristina Allen-Brady; Methodology, Lisa A. Cannon-Albright, Jeff Stevens, Julio Facelli and Craig Teerlink; Project administration, Lisa A. Cannon-Albright and Kristina Allen-Brady; Resources, Lisa A. Cannon-Albright and Neeraj Agarwal; Software, Jeff Stevens, Julio Facelli and Craig Teerlink; Validation, Lisa A. Cannon-Albright, Craig Teerlink and Kristina Allen-Brady; Visualization, Julio Facelli; Writing – original draft, Lisa A. Cannon-Albright, Jeff Stevens, Julio Facelli and Craig Teerlink; Writing – review & editing, Lisa A. Cannon-Albright, Jeff Stevens, Julio Facelli, Craig Teerlink, Kristina Allen-Brady and Neeraj Agarwal.

ACKNOWLEDGEMENTS

This research was supported by the U.S. Department of Defense Prostate Cancer Research Program of the Office of the Congressionally Directed Medical Research Programs, Grant Number DOD PC170413, Lethal Prostate Cancer Gene Identification, awarded to Lisa Cannon-Albright. This research was conducted using the UK Biobank Resource under Application Number 43460. This work was partially supported by the Utah Center for Clinical and Translational Science funded by NCATS award 1ULTR002538. Computer resources were provided by the University of Utah Center for High-Performance Computing, which has been partially funded by the NIH Shared Instrumentation Grant 1S10OD02164401A1.

This research was supported by the Utah Cancer Registry, which is funded by the National Cancer Institute's SEER Program, Contract No. HHSN261201800016I, the US Center for Disease Control and Prevention's National Program of Cancer Registries, Cooperative Agreement No. NU58DP007131, with additional support from the University of Utah and Huntsman Cancer Foundation. Partial support for all datasets within the Utah Population Database is provided by the University of Utah, Huntsman Cancer Institute and the Huntsman Cancer Institute Cancer Center Support grant, P30 CA42014 from the National Cancer Institute. L.A.C.A. acknowledges partial support from the Huntsman Cancer Institute Cancer Center Support grant, P30 CA42014 from the National Cancer Institute.

REFERENCES

Abraham, G., Qiu, Y., & Inouye, M. (2017). FlashPCA2: principal component analysis of Biobank-scale genotype datasets. Bioinformatics, 33, 2776-2778. doi: 10.1093/bioinformatics/btx299

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. Nat Methods **7**(4):248-249 (2010

Albright FS, Stephenson RA, Agarwal N, Teerlink CC, Lowrance WT, Farnham JM, Cannon-Albright LA. Prostate cancer risk prediction based on complete prostate cancer family history. Prostate, 2015 75(4):390-8.

Albright FS, Stephenson RA, Agarwal N, Cannon-Albright LA. Relative Risks for lethal prostate cancer based on complete family history of prostate cancer death. Prostate. 2017 Jan;77(1):41-48. doi: 10.1002/pros.23247. Epub 2016 Aug 16. PMID: 27527734

Brandt A, Sundquist J, Hemminki K. Risk for incident and fatal prostate cancer in men with a family history of any incident and fatal cancer. Ann Oncol 2012, 23, 251-256.

Brockmeyer DL, Cheshier SH, Stevens J, Facelli JC, Rowe K, Heiss JD, Musolf A, Viskochil DH, Allen-Brady KL, Cannon-Albright LA. A likely HOXC4 predisposition variant for Chiari malformations. J Neurosurg 2022 Nov 25:1-9.

Cannon-Albright LA. (2008). Utah family-based analysis: past, present and future. *Hum Hered*, *65*(4), 209-20.

Cannon L, Bishop DT, Skolnick MH, Hunt S, Lyon JL, Smart C. (1982). Genetic epidemiology of prostate cancer in the Utah Mormon genealogy. *Cancer Surv*, *1*(1), 48-69.

Cannon-Albright LA, Teerlink CC, Stevens J, Huang FW, Sipeky C, Schleutker J, Hernandez R, Facelli J, Agarwal N, Trump DL. A Rare Variant in ERF (rs144812092) Predisposes to Prostate and Bladder Cancers in an Extended Pedigree. 2021 Cancers, 13(10), 2399.

Cannon-Albright LA, Farnham JM, Thomas A, Camp NJ. (2005). Identification and study of Utah pseudo-isolate populations-prospects for gene identification. *Am J Med Genet A*, *137*(3), 269-75.

Cannon-Albright LA, Teerlink CC, Stevens J, Snow AK, Thompson BA, Bell R, Nguyen KN, Sargent NR, Kohlmann W, Neklason DW, Tavtigian SV. FANCM c5791CT stopgain mutation (rs144567652) is a familial colorectal cancer risk factor. Molecular Genetics & Genomic Medicine; 2020 Dec;8(12).

Cannon-Albright LA, Teerlink CC, Stevens J, Huang FW, Sipeky C, Schleutker J, Hernandez R, Facelli J, Agarwal N, Trump DL. A Rare Variant in ERF (rs144812092) Predisposes to Prostate and Bladder Cancers in an Extended Pedigree. 2021 Cancers, 13(10), 2399.

Carter BS, Bova GS, Beaty TH, Steinberg GD, Childs B, Issacs WB, Walsh PC. Hereditary prostate cancer: epidemiologic and clinical features. J Urol 1993, 150,797-802.

Castro E, Eeles R. The role of BRCA1 and BRCA2 in prostate cancer. Asian J Androl 2012, 14, 409-14.

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M.,Schlessinger, D., Stambolian, D., Loh, P., Iacono, W. G., Swaroop, A., Scott, L. J., Cucca, F., Kronenberg, F., Boehnke, M., Abecasis, G. R., & Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. Nat Genet, 48, 1284–1287. doi: 10.1038/ng.3656

Das S, Salami SS, Spratt DE, Kaffenberger SD, Jacobs MF, Morgan TM. Bringing prostate cancer germline genetics into clinical practice. J Urol 2019 Aug; 202(2):223-230.

Deelen, P., Bonder, M. J., van der Velde, K. J., Westra, H., Winder, E., Hendriksen, D., Franke, L., & Swertz, M. A. (2014). Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. BMC Research Notes, 7, 901. doi: 10.1186/1756-0500-7-901

Ewing CM, Ray AM, Lange EM, Zuhlke KA, Robbins CM, Tembe WD, Wiley KE, Isaacs SD, Johng D, Wang Y, Bizon C, Yan G, Gielzak M, Partin AW, Shanmugam V, Izatt T, Sinari S, Craig DW, Zheng SL, Walsh PC, Montie JE, Xu J, Carpten JD, Isaacs WB, Cooney KA. Germline mutations in HOXB13 and prostate cancer risk. NEJM 2012 Jan12;366(2):141-9.

Gebauer D, Li J, Jogl G, Shen Y, Myszka DG, Tong L. Crystal structure of the PH-BEACH domains of human LRBA/BGL. *Biochemistry* **43** 14873-80 (2004) PMID: 15554694

The Haplotype Reference Consortium. (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet, 48, 1279-1283. doi: 10.1038/ng.3643

Hernandez R, Facelli JC. Understanding protein structural changes for oncogenic missense variants. Heliyon. 2021 Jan 27;7(1):e06013. doi: 10.1016/j.heliyon.2021.e06013. PMID: 33553733; PMCID: PMC7846930.

Huyghe, J. R., Blen, S. A., Harrison, T. A., Kang, H. M., Chen, S., Schmit, S. L., Conti, D. V., Qu, C., Jeon, J., Edlund, C. K., Greenside, P., Wainberg, M., Schumacher, F. R., Smith, J.D., Levine, D. M., Nelson, S. C., Sinnott-Armstrong, N. A., Albanes, D., Alonso, M. H., Anderson, K., Arnau-Collell, C., Arndt, V., Bamia, C., Banbury, B. L., Baron, J. A., Berndt, S. I., Bézieau, S., Bishop, D. T., Boehm, J., Boeing, H., Brenner, H., Brezina, S., Buch, S., Buchanan, D.D., Burnett-Hartman, A., Butterbach, K., Caan, B. J., Campbell, P. T., Carlson, C. S., Castellví-Bel, S., Chan, A. T., Chang-Claude, J., Chanock, S. J., Chirlaque, M. D., Cho, S. H., Connolly, C. M., Cross, A. J., Cuk, K., Curtis, K. R., de la Chapelle, A.,

Doheny, K. F., Duggan, D., Easton, D. F., Elias, S. G., Elliott, F., English, D. R., Feskens, E. J. M., Figueiredo, J. C., Fischer, R., FitzGerald, L. M., Forman, D., Gala, M., Gallinger, S., Gauderman, W. J., Giles, G. G., Gillanders, E., Gong, J., Goodman, P. J., Grady, W. M., Grove, J. S., Gsur, A., Gunter, M. J., Haile, R. W., Hampe, J., Hampel, H., Harlid, S., Hayes, R. B., Hofer, P., Hoffmeister, M., Hopper, J. L., Hsu, W. L., Huang, W. Y., Hudson, T. J., Hunter, D. J., Ibañez-Sanz, G., Idos, G. E., Ingersoll, R., Jackson, R. D., Jacobs, E. J., Jenkins, M. A., Joshi, A. D., Joshu, C. E., Keku, T. O., Key, T. J., Kim, H. R., Kobayashi, E., Kolonel, L. N., Kooperberg, C., Kühn, T., Küry, S., Kweon, S. S., Larsson, S. C., Laurie, C. A., Le Marchand, L., Leal, S. M., Lee, S.C., Lejbkowicz, F., Lemire, M., Li, C. I., Li, L., Lieb, W., Lin, Y., Lindblom, A., Lindor, N. M., Ling, H., Louie, T. L., Männistö, S., Markowitz, S. D., Martín, V., Masala, G., McNeil, C. E., Melas, M., Milne, R. L., Moreno, L., Murphy, N., Myte, R., Naccarati, A., Newcomb, P. A., Offit, K., Ogino, S., Onland-Moret, N. C., Pardini, B., Parfrey, P. S., Pearlman, R., Perduca, V., Pharoah, P. D. P., Pinchev, M., Platz, E. A., Prentice, R. L., Pugh, E., Raskin, L., Rennert, G., Rennert, H. S., Riboli, E., Rodríguez-Barranco, M., Romm, J., Sakoda, L. C., Schafmayer, C., Schoen, R. E., Seminara, D., Shah, M., Shelford, T., Shin, M. H., Shulman, K., Sieri, S., Slattery, M. L., Southey, M. C., Stadler, Z. K., Stegmaier, C., Su, Y. R., Tangen, C. M., Thibodeau, S. N., Thomas, D. C., Thomas, S. S., Toland, A. E., Trichopoulou, A., Ulrich, C. M., Van Den Berg, D. J., van Duijnhoven, F. J. B., Van Guelpen, B., van Kranen, H., Vijai, J., Visvanathan, K., Vodicka, P., Vodickova, L., Vymetalkova, V., Weigl, K., Weinstein, S. J., White, E., Win, A. K., Wolf, C. R., Wolk, A., Woods, M. O., Wu, A. H., Zaidi, S. H., Zanke, B. W., Zhang, Q., Zheng, W., Scacheri, P. C., Potter, J. D., Bassik, M. C., Kundaje, A., Casey, G., Moreno, V., Abecasis, G. R., Nickerson, D. A., Gruber, S. B., Hsu, L., & Peters, U. (2019). Discovery of common and rare genetic risk variants for colorectal cancer. Nat Genet, 51, 76-87. doi: 10.1038/s41588-018-0286-6

Loh, P. R., Palamara, P. F., & Price, A. L. (2016). Fast and accurate long-range phasing in a UK Biobank cohort. Nat Genet, 48, 811-816. doi: 10.1038/ng.3571.

Lopez-Herrera G, Tampella G, Pan-Hammarstrom Q, Herholz P, Trujillo-Vargas CM, Phadwal K, Simon AK, Moutschen M, Etzioni A, Mory A, Srugo I, Melamed D, Hultenby K, Liu C, Baronio M, Vitali M, Philippet P, Dideberg V, Grimbacher B. Deleterious mutations in LRBA are associated with a syndrome of immune deficiency and autoimmunity._Am J Hum genet 90:986-1001 (2012)

Manolio, T.A.; Collins, F.S.; Cox, N.J.; Goldstein, D.B.; Hindorff, L.A.; Hunter, D.J.; McCarthy, M.I.; Ramos, E.M.; Cardon, L.R.; Chakravarti, A.; et al. Finding the missing heritability of complex diseases. *Nat. Cell Biol.* **2009**, *461*, 747–753.

Miller JB, Ward E, Staley LA, Stevens J, Teerlink CC, Tavana JP, Cloward M, Page M, Dayton L, Alzheimer's Disease Genetics Consortium, Cannon-Albright LA, Kauwe JSK. Identification and genomic analysis of pedigrees with exceptional longevity identifies candidate rare variants. Neurobio Dis 143, Sept 2020.

Nelson Q, Agarwal N, Stephenson R, **Cannon-Albright LA**. (2013). A population-based analysis of clustering identifies a strong genetic contribution to lethal prostate cancer. *Front Genet*, *4*, 152.

Ott J, Wang J, Leal SM Genetic linkage analysis in the age of whole-genome sequencing. *Nat. Rev. Genet.* **2015**, *16*, 275–284.

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera—A visualization system for exploratory research and analysis. Journal of computational chemistry. 2004;25:1605-12.

Pritchard CC, Mateo J, Walsh MF, De Sarkar N, Abida W, Beltran H, Garofalo A, Gulati R, Carreira S, Eeles R, Elemento O, Rubin MA, Robinson D, Lonigro R, Hussain M, Chinnaiyan A, Vinson J, Filipenko J, Garraway L, Taplin ME, AlDubayan S, Han GC, Beightol M, Morrissey C, Nghiem B, Cheng HH, Montgomery B, Walsh T, Casadei S, Berger M, Zhang L, Zehir A, Vijai J, Scher HI, Sawyers C, Schultz N, Kantoff PW, Solit D, Robson M, Van Allen EM, Offit K, de Bono J, Nelson PS. Inherited DNA-repair gene mutations in men with metastatic prostate cancer. NEJM 2016 Aug 4:375(5):443-53

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham. P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet, 81, 559-575. doi: 10.1086/519795

Rantapero T, Wahlfors T, Kähler A, Hultman C, Lindberg J, Tammela TL, Nykter M, Schleutker J, Wiklund F. Inherited DNA repair gene mutations in men with lethal prostate cancer. Genes (Basel). 2020 Mar 14; 11(3):314.

Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protocols. 2010;5(4):725-38.

Schumacher, F. R., Al Olama, A. A., Berndt, S. I., Benlloch, S., Ahmed, M., Saunders, E. J., Dadaev, T., Leongamornlert, D., Anokian, E., Cieza-Borrella, C., Goh, C., Brook, M. N., Sheng, X., Fachal, L., Dennis, J., Tyrer, J., Muir, K., Lophatananon, A., Stevens, V. L., Gapstur, S. M., Carter, B. D., Tangen, C. M., Goodman, P. J., Thompson, I. M., Batra, J., Chambers, S., Moya, L., Clements, J., Horvath, L., Tilley, W., Risbridger, G. P., Gronberg, H., Aly, M., Nordström, T., Pharoah, P., Pashayan, N., Schleutker, J., Tammela, T. L.J., Sipeky, C., Auvinen, A., Albanes, D., Weinstein, S., Wolk, A., Håkansson, N., West, C. M. L., Dunning, A. M., Burnet, N., Mucci, L. A., Giovannucci, E., Andriole, G. L., Cussenot, O., Cancel-Tassin, G., Koutros, S., Beane Freeman, L. E., Sorensen, K. D., Orntoft, T. F., Borre, M., Maehle, L., Grindedal, E. M., Neal, D. E., Donovan, J. L., Hamdy, F. C., Martin, R. M., Travis, R. C., Key, T. J., Hamilton, R. J., Fleshner, N. E., Finelli, A., Ingles, S. A., Stern, M. C., Rosenstein, B. S., Kerns, S. L., Ostrer, H., Lu, Y. J., Zhang, H. W., Feng, N., Mao, X., Guo, X., Wang, G., Sun, Z., Giles, G. G., Southey, M. C., MacInnis, R. J., FitzGerald, L. M., Kibel, A. S., Drake, B. F., Vega, A., Gómez-Caamaño, A., Szulkin, R., Eklund, M., Kogevinas, M., Llorca, J., Castaño-Vinyals, G., Penney, K. L., Stampfer, M., Park, J. Y., Sellers, T. A., Lin, H. Y., Stanford, J. L., Cybulski, C., Wokolorczyk, D., Lubinski, J., Ostrander, E. A., Geybels, M. S., Nordestgaard, B. G., Nielsen, S. F., Weischer, M., Bisbjerg, R., Røder, M. A., Iversen, P., Brenner, H., Cuk, K., Holleczek, B., Maier, C., Luedeke, M., Schnoeller, T., Kim, J., Logothetis, C. J., John, E. M., Teixeira, M. R., Paulo, P., Cardoso, M., Neuhausen, S. L., Steele, L., Ding, Y. C., De Ruyck, K., De Meerleer, G., Ost, P., Razack, A., Lim, J., Teo, S. H., Lin, D. W., Newcomb, L. F., Lessel, D., Gamulin, M., Kulis, T., Kaneva, R., Usmani, N., Singhal, S., Slavov, C., Mitev, V., Parliament, M., Claessens, F., Joniau, S., Van den Broeck, T., Larkin, S., Townsend, P. A., Aukim-Hastie, C., Gago-Dominguez, M., Castelao, J. E., Martinez, M. E., Roobol, M. J., Jenster, G., van Schaik, R. H. N., Menegaux, F., Truong, T., Koudou, Y. A., Xu, J., Khaw, K. T., Cannon-Albright, L., Pandha, H., Michael, A., Thibodeau, S. N., McDonnell, S. K., Schaid, D. J., Lindstrom, S., Turman, C., Ma, J., Hunter, D. J., Riboli, E., Siddiq, A,, Canzian, F., Kolonel, L. N., Le Marchand, L., Hoover, R. N., Machiela, M. J., Cui, Z., Kraft, P., Amos, C. I., Conti, D. V., Easton, D. F., Wiklund, F., Chanock, S. J., Henderson, B. E., Kote-Jarai, Z., Haiman, C. A., Eeles, R. A.; Profile Study; Australian Prostate Cancer BioResource (APCB); IMPACT Study; Canary PASS Investigators; Breast and Prostate Cancer Cohort Consortium (BPC3); PRACTICAL (Prostate Cancer Association Group to Investigate Cancer-Associated Alterations in the Genome) Consortium; Cancer of the Prostate in Sweden (CAPS); Prostate Cancer Genome-wide Association Study of Uncommon Susceptibility Loci (PEGASUS); & Genetic Associations and Mechanisms in Oncology (GAME-ON)/Elucidating Loci Involved in Prostate Cancer Susceptibility (ELLIPSE) Consortium. (2018). Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. Nat Genet, 50, 928-936. doi: 10.1038/s41588-018-0142-8

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins R. (2015). UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med, 12, e1001779. doi: 10.1371/journal.pmed.1001779

Teerlink CC, Miller JB, Vance EL, Staley LA, Stevens J, Tavana JP, Cloward ME, Page ML, Dayton L; Alzheimer's Disease Genetics Consortium, Cannon-Albright LA, Kauwe JSK. Analysis of high-risk pedigrees identifies 11 candidate variants for Alzheimer's disease. Alzheimer's Dement. 2022 Feb;18(2):307-317.

Teerlink CC, Stevens J, Hernandez R, Facelli JC, Cannon-Albright LA. An intronic variant in the CELF4 gene is associated with risk for colorectal cancer. Cancer Epidemiology 72, June 2021

Teerlink CC. Jurynec MJ, Hernande R, Stevens J, Hughes DC, Brunker CP, Rowe K, Grunwald DJ, Facelli JC, Cannon-Albright LA. A role for the MEGF6 gene in predisposition to osteoporosis. Annals of Hum Genet 2020; 1-15.

TerwilligerJD, Goring HH. Gene mapping in the 20th and 21st centuries: statistical methods, data analysis, and experimental design. Hum Biol 2009 Dec;81(5-6):663-728.

Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010 Sep;38(16):e164

Wijsman, E.M. The role of large pedigrees in an era of high-throughput sequencing. *Qual. Life Res.* **2012**, *131*, 1555–1563.

Xu J, Lange EM, Lu L, Zheng SL, Wang Z, Thibodeau SN, **Cannon-Albright LA**, Teerlink CC, Camp NJ, Johnson AM, Zuhlke KA, Stanford JL, Ostrander EA, Wiley KE, Isaacs SD, Walsh PC, Maier C, Luedeke M, Vogel W, Schleutker J,

Wahlfors T, Tammela T, Schaid D, McDonnell SK, DeRycke MS, Cancel-Tassin G, Cussenot O, Wiklund F, Grönberg H, Eeles R, Easton D, Kote-Jarai Z, Whittemore AS, Hsieh CL, Giles GG, Hopper JL, Severi G, Catalona WJ, Mandal D, Ledet E, Foulkes WD, Hamel N, Mahle L, Moller P, Powell I, Bailey-Wilson JE, Carpten JD, Seminara D, Cooney KA, Isaacs WB, International Consortium for Prostate Cancer Genetics. (2013). HOXB13 is a susceptibility gene for prostate cancer: results from the International Consortium for Prostate Cancer Genetics (ICPCG). *Hum Genet*, *132*(1), 5-14.

Yang J, Roy A, Zhang Y. *Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment*, Bioinformatics, 29:2588-2595 (2013a).

Yang J, Roy A, and Zhang Y. *BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions*, Nucleic Acids Research, 41: D1096-D1103 (2013b).

Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. Nat Meth. 2015;12:7-8.

Zhang Y. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics. 2008;9:40.

Zhang Y. I-TASSER: Fully automated protein structure prediction in CASP8. Proteins: Structure, Function, and Bioinformatics. 2009;77(S9):100-13.

Ziv, E., Dean, E., Hu, D., Martino, A., Serie, D., Curtin, K., Campa, D., Aftab, B., Bracci, P., Buda, G., Zhao, Y., Caswell-Jin, J., Diasio, R., Dumontet, C., Dudziński, M., Fejerman, L., Greenberg, A., Huntsman, S., Jamroziak, K., Jurczyszyn, A., Kumar, S., Atanackovic, D., Glenn, M., Cannon-Albright, L. A., Jones, B., Lee, A., Marques, H., Martin, T., Martinez-Lopez, J., Rajkumar, V., Sainz, J., Vangsted, A. J., Wątek, M., Wolf, J., Slager, S., Camp, N. J., Canzian, F., & Vachon, C. (2015). Genome-wide association study identifies variants at 16p13 associated with survival in multiple myeloma patients. Nat Communications, 6, 7539. doi: 10.1038/ncomms8539