Investigating the genomic profile of inherited prostate cancer using whole exome sequencing data

Alexandre Sá Ferreira[1], Pedro Soares[2], and Andreia Brandão[3]

[1]Escola de Engenharia - Universidade do Minho
[2]Centre of Molecular and Environmental Biology (CBMA), Department of Biology, University of Minho
[3]Cancer Genetics Group, IPO Porto Research Center (CI-IPOP)/RISE@CI-IPOP (Health Research Network), Portuguese Oncology Institute of Porto (IPO Porto)/Porto Comprehensive Cancer Center, Porto, Portugal.

2025

Abstract

Prostate cancer (PrCa) is the second most common cancer among men and a leading cause of cancer-related death. While hereditary factors contribute to approximately 20% of cases, the role of rare and low-frequency genetic variants in PrCa predisposition remains underexplored. This project aims to investigate the mutational landscape of rare (0.1% < MAF < 1%) and low-frequency (1% < MAF < 5%) variants in 96 PrCa patients from 45 families using whole exome sequencing (WES). The study will identify significantly altered genes and enriched pathways associated with PrCa susceptibility, focusing on frameshift, nonsense, splicing, and deleterious missense variants. This research aims to contribute to early detection and targeted treatment strategies.

Keywords: hereditary prostate cancer; single nucleotide variations; whole exome sequencing.

1 Introduction

Repository location `https://github.com/Alexsf35/Projeto_Bioinf.git`

Context During their lifetime, men have a 40.1% chance of being diagnosed with any invasive cancer (slightly higher than women, 38.7%)[1, 2]. Among these diagnoses, approximately 14.2% are prostate cancers (PrCa). As of 2022, PrCa has been determined to be the second most prevalent cancer worldwide in terms of incidence and the fifth most prevalent in terms of mortality among the male population. It has been identified as the most common cancer among men in 118 different countries, including Portugal[3, 4].

A notable factor in PrCa is its hereditary component. The proportion of PCa attributable to hereditary factors has been estimated to be 5 to 15%, with other estimates stating that up to 20% of men diagnosed with PrCa have a family history of the disease within their paternal or fraternal lineage[5]. In the case of any affected family member, the relative risk is doubled, with an additional increase observed in relation to the number of relatives and their age (under 60 years of age) at the time of diagnosis[6]. In addition, ethnicity has been shown to exert a significant influence on the outcomes of PrCa cases, with African ancestry being recognized as a well-established risk factor[1, 7, 8].

The studies done on hereditary cancer syndromes such as Hereditary Breast and Ovarian Cancer (HBOC) and Lynch Syndrome (LS), have contributed to the identification of many pathogenic and "likely/potentially pathogenic" germline mutations in homologous recombination genes (*BRCA1/2*, *ATM*, *PALB2*, *CHEK2*) and mismatch repair genes (*MLH1*, *MSH2*, *MSH6*, and *PMS2*) respectively, that may also underlie hereditary prostate cancer (HPC)[5, 9, 10, 11]. It is estimated that these germline mutations may account for 7.4–21% of cases. However, it is noteworthy that 79–92% of cases are attributable to genes that have yet to be identified[5].

The etiology of PrCa remains to be fully explained; however, mounting evidence suggests that the aforementioned factors are significant indicators of genetic susceptibility to PrCa. This susceptibility is linked to the inheritance of a combination of rare germline variants (defined as having a minor allele frequency [MAF] $< 1\%$) in moderate to high-penetrance genes and common genetic alterations in low-risk genes[4, 11].

PrCa is thus characterized as a cancer with a multigenic etiology, that can be caused by single nucleotide variations (SNVs) (which can lead to truncating, missense or nonsense changes). While few such mutations may have low impact, the cumulative effects might be of particular interest in HPC studies[12].

Objectives Whilst the majority of studies on the genetic predisposition to PrCa have focused on common variants (MAF $> 5\%$) and very rare, high-risk variants (MAF $< 0.1\%$), the present project aims to bridge this gap by investigating the mutational landscape of rare ($0.1\% <$ MAF $<1\%$) and low-frequency ($1\% <$ MAF $< 5\%$) single nucleotide variants. The present study will analyse whole exome sequencing (WES) data from 96 PrCa patients across 45 families, this analysis would seek to characterize the low-frequency and rare variants that cause HPC. In addition, the study seeks to identify significantly impacted genes that play a role in predisposition to PrCa and identify enriched biological pathways that may be responsible for increased susceptibility. With all these objectives brought together, the project aims to provide a comprehensive picture of the enriched pathways implicated in PrCa susceptibility, which can eventually contribute to more effective early detection and targeted treatment strategies.

2 Epidemiology

PrCa is a major health issue worldwide with an average age of diagnosis being 66 years old[11]. The age-standardized incidence rate (ASR) in 2020 was of 31 per 100,000 males, with a lifetime risk of 3.9%. Still, these incidences vary tremendously by region.

For instance, incidences of as high as 83 per 100,000 are seen in Northern Europe, followed by Western Europe (78), the Caribbean (76), and Australia/New Zealand. Conversely, incidences as low as 6.3 per 100,000 have been observed in countries such as South-Central Asia, 14 for South-Eastern Asia, and 17 for Northern Africa. These variations are the reflection of the differences in screening efforts, access to healthcare and public awareness about the disease. As for mortality, in 2020 the global ASR was approximately 7.7 per 100,000 males. The less developed socioeconomic areas have much higher mortality rates, with the Caribbean at 28 per 100,000, Middle Africa at 25, and Southern Africa at 22. In contrast, Asian regions have significantly lower mortality rate, South-Central Asia at 3.1, Eastern Asia at 4.6, and South-Eastern Asia at 5.4 per 100,000[6].

In the future, projections are not too bright. It is estimated that new PrCa cases will nearly double, from 1.4 million in 2020 to approximately 2.9 million in 2040. Similarly, PrCa deaths are also estimated to rise by 85%, from 375,000 to nearly 700,000 in the same period[8]. This expected rise is partly due to world population aging and the varied effectiveness of screening and treatment initiatives around the world[1, 6].

PrCa incidence has been rising since the 1980s, largely as a result of widespread of prostate-specific antigen (PSA) testing and once again an increase in disease awareness. Which in turn results in early diagnosis and treatment, having reflected a decline in mortality rates, that have decreased by 52% from their peak in 1993, with recent patterns showing a moderate decline through 2020[1, 7, 8].

Lifestyle factors are also a significant in PrCa diagnosis. A meta-analysis in 2016 found that higher physical activity is correlated with a 38% reduction in PrCa-specific mortality, which suggests that an active lifestyle may be important in managing disease progression[13]. Other factors, such as diet and smoking habits, have also been shown to affect both incidence and mortality[6, 7, 8].

3 Genetic Etiology

To be diagnosed with hereditary prostate cancer (HPC) families must meet the Johns Hopkins criteria by either having three or more first-degree relatives diagnosed with PrCa; cases spanning three successive generations; or having at least two relatives diagnosed with early-onset PrCa (before the age of 56 years)[11].

Notably, any affected family member doubles PrCa risk and is increased by 2.5-fold if one first-degree relative is less than 60 years old at diagnosis (1.6 times if the first-degree relative is older than 60 years). If 2 or more relatives are diagnosed before 60 years old, the relative risk can escalate up to 5.7[6]. These elevated risks may be attributed to the existence of rare variants that are present in less than 1% of the population, which get pass down hereditarily and affect moderate- to high-penetrance genes. These genetic factors are frequently identified in families that also exhibit other hereditary cancer syndromes, which are among the most important risk factors compared to age, race, ethnicity and environmental factors for the development of PCa and this risk is estimated at 40% to 50 %.[5] They are:

HBOC, is most commonly associated with inherited pathogenic mutations in the *BRCA1* and *BRCA2* genes, though other genes (such as *PALB2*, *ATM*, and *CHEK2*) can also play a role. *PALB2*, for instance, links *BRCA1* and *BRCA2* forming the "BRCA complex" that repairs DNA double-strand breaks via HR. Furthermore, it has been established that they also exert control over centrosome dynamics, chromosome segregation and cytokinesis, thereby contributing to the temporally and spatially stabilised genome within the cell cycle[9]. The other genes (*ATM* and *CHEK2*) are also related to the correction of damaged genome; *CHEK2* encodes a checkpoint kinase that interacts with cell cycle regulators and DNA repair proteins, like the aforementioned *ATM* serine threonine kinase that recognises double stranded DNA breaks and initiates multiple aspects of the damage response cascade[14].

LS, is caused by mutations in DNA mismatch repair (MMR) genes, most commonly *MLH1*, *MSH2*, *MSH6*, and *PMS2* . Individuals with LS have a higher lifetime risk of developing colorectal and endometrial cancer and a range of other cancers such as ovarian, gastric and urinary tract[15].Its influence in PrCa susceptibility is still controversial and the evidence is not consistent, but overall risk studies indicate a moderately increased risk in men who possess an MMR gene mutation[10, 11].

Distinct from the broader hereditary cancer syndromes, *HOXB13* is the only gene being specifically and consistently associated with an increased risk of HPC, especially the G84E mutation[11, 16]. Even though the mutations can be found in both affected and healthy man, the carrier rate was found to be considerably elevated among affected males (194 out of 382; 51% in some studies) in comparison to their counterparts within these families[16]. The protein expressed by *HOXB13* plays a crucial role in the embryonic development of the prostate gland and continues to be expressed in normal prostate tissue throughout adulthood. Moreover, it has been demonstrated to interact with the androgen receptor, thereby impacting the proliferation and differentiation of both normal and cancerous prostate cells, which underlines its significance as a PrCa susceptibility gene[17].

4 Methods

4.1 Dataset

Forty-nine of the 462 families provided DNA samples from two or three affected relatives (including the proband). Of those, forty-five families, specifically, six with three available DNA samples and 39 with two available DNA samples, were selected for WES according to the following prioritization: firstly, families with three available DNA samples; secondly, families with more than two family members diagnosed with PrCa; thirdly, families with probands diagnosed before the age of 61; and, lastly, families with the lowest average age at onset. The genetic screening with whole-exome sequencing therefore included 96 patients from 45 families of the 462 families recruited.

4.2 Capture and sequencing

Approximately 1μg of genomic DNA was enriched for exonic regions using the SureSelectXT2 Human All Exon v5 kit (Agilent Technologies, Santa Clara, CA, USA), according to the SureSelectXT2 Target Enrichment System for Illumina Multiplexed Sequencing protocol, at the Carvajal-Carmona Laboratory, Genome Center & Department of Biochemistry and Molecular Medicine, University of California, Davis, CA, USA. Sequencing of the pooled enriched libraries was an outsourced service, provided by the Beijing Genomics Institute sequencing facility at the UC Davis campus in Sacramento, CA, USA, and was performed with 100bp paired-end reads using the Illumina HiSeq 2000 platform (Illumina Inc., San Diego, CA, USA).

4.3 Data processing

Raw sequence data was received in paired-end FASTQ format and processed at the Carvajal-Carmona Laboratory. Data were demultiplexed and converted to Sanger encoding using seqtk (https://github.com/lh3/seqtk). Fastq quality was assessed and checked with FastqQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Quality trimming was performed by SICKLE (https://github.com/najoshi/sickle). Sequence reads were trimmed and aligned to the Genome Analysis Toolkit (GATK) bundle human genome build GRCh37_decoy/hg19 using Burrows-Wheeler Alignment (BWA-mem, version 0.7.8-r455) [18, 19] and PCR duplicates were removed using Picard (1.118). GATK (v3.2-2) IndelRealigner and BaseRecalibrator were used for indel realignment and base quality score recalibration, and variants were called with five different callers: Freebayes [20], GATK HaplotypeCaller algorithm [21, 22], SAMtools [23], SNVer [24], and VarScan [25].

4.4 Variant annotation and prioritization

Variants for all callers were combined and filtered according to the following filters: coverage ≥10; variant counts ≥5; variant frequency ≥10%; average single nucleotide variants (SNV) base quality ≥22; and ≥10% of variant reads on both strands. For secondary variant filtering, intronic variants at more than 2-bp away from exon-intron boundaries, synonymous, UTR variants, and variants present in more than 10% of the samples were excluded. Variant annotation will be carried out with ANNOVAR software [26], a Perl-based software tool that enables functional gene-based, region-based and filter-based annotation of variants derived from next-generation sequencing. ANNOVAR gathers information for MAF in non-Finnish European (NFE) and Iberic (IBS) populations, from the Genome Aggregation Database (gnomAD) and the 1000 Genomes Project [1000G, Phase 3 data], and for the pathogenicity predictors available at dbNSFP (v3.5), which compiles results from multiple prediction tools.

Variant Filtering and Annotation

- Frameshift, nonsense, or splicing variants with MAF $\leq$ 1% and 1% $<$ MAF $<$ 5% in both the NFE and IBS populations

- Missense variants with MAF $\leq$ 1% and 1% $<$ MAF $<$ 5% that are predicted to be damaging or deleterious by at least 9 out of 11 pathogenicity predictors and supported by at least 3 out of 4 conservation predictors. These will be considered "potentially pathogenic".

Identification of significantly altered genes

Identification of significantly altered genes will be performed using tools like MutSigCV[27] or OncodriveFML[28]. MutSigCV evaluates mutation significance after adjusting for background mutation rate, gene length, and mutational clustering patterns. Requiring input files in Mutation Annotation Format (MAF), along with coverage and covariate files generated from sequencing data. OncodriveFML complements this by assessing the functional bias of mutations, identifying genes enriched for functional alterations. Both studies will help to identify genes whose mutational patterns are unlikely to occur by chance and may be indicative of involvement in PrCa susceptibility. Multiple testing adjustment will be done using False Discovery Rate (FDR) in order to offer statistical robustness.

Following identification of significantly mutated genes, gene set enrichment analysis (GSEA) a method used to determine whether a group of genes (a *gene set*) has statistically significant differences, will be conducted to identify enriched or depleted biological processes and pathways potentially implicated in PrCa susceptibility. Tools such as DAVID[29], Enrichr[30], and g:Profiler[31] will be used to assess the overrepresentation of gene sets within curated databases, including KEGG, Reactome, and Gene Ontology (GO).

References

[1] Rebecca L. Siegel, Kimberly D. Miller, and Ahmedin Jemal. "Cancer statistics, 2020". In: *CA: A Cancer Journal for Clinicians* 70.1 (2020), pp. 7–30. DOI: `https://doi.org/10.3322/caac.21590`. eprint: `https://acsjournals.onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21590`. URL: `https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21590`.

[2] Freddie Bray et al. "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". In: *CA: a cancer journal for clinicians* 74.3 (2024), pp. 229–263.

[3] International Agency for Research on Cancer. *Prostate Cancer Fact Sheet*. PDF available online: `https://gco.iarc.who.int/media/globocan/factsheets/cancers/27-prostate-fact-sheet.pdf`. Accessed: March 20, 2025.

[4] Marta Cardoso et al. "Exome sequencing of affected duos and trios uncovers PRUNE2 as a novel prostate cancer predisposition gene". In: *British Journal of Cancer* 128.6 (2023), pp. 1077–1085.

[5] Maria Teresa Vietri et al. "Hereditary prostate cancer: genes related, target therapy and prevention". In: *International journal of molecular sciences* 22.7 (2021), p. 3753.

[6] Giorgio Gandaglia et al. "Epidemiology and prevention of prostate cancer". In: *European urology oncology* 4.6 (2021), pp. 877–892.

[7] Hyuna Sung et al. "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". In: *CA: a cancer journal for clinicians* 71.3 (2021), pp. 209–249.

[8] Nicholas D James et al. "The Lancet Commission on prostate cancer: planning for the surge in cases". In: *The Lancet* 403.10437 (2024), pp. 1683–1722.

[9] Reiko Yoshida. "Hereditary breast and ovarian cancer (HBOC): review of its molecular characteristics, screening, treatment, and prognosis". In: *Breast Cancer* 28.6 (2021), pp. 1167–1180.

[10] Shae Ryan, Mark A Jenkins, and Aung Ko Win. "Risk of prostate cancer in Lynch syndrome: a systematic review and meta-analysis". In: *Cancer epidemiology, biomarkers & prevention* 23.3 (2014), pp. 437–449.

[11] Andreia Brandão, Paula Paulo, and Manuel R Teixeira. "Hereditary predisposition to prostate cancer: from genetics to clinical implications". In: *International journal of molecular sciences* 21.14 (2020), p. 5036.

[12] Charles M Ewing et al. "Germline mutations in HOXB13 and prostate-cancer risk". In: *New England Journal of Medicine* 366.2 (2012), pp. 141–149.

[13] ANNE McTiernan et al. "Physical activity in cancer prevention and survival: a systematic review". In: *Medicine and science in sports and exercise* 51.6 (2019), p. 1252.

[14] Brennan Decker et al. "Rare, protein-truncating variants in ATM, CHEK2 and PALB2, but not XRCC2, are associated with increased breast cancer risks". In: *Journal of medical genetics* 54.11 (2017), pp. 732–741.

[15] Päivi Peltomäki. "Lynch syndrome genes". In: *Familial cancer* 4 (2005), pp. 227–232.

[16] Jianfeng Xu et al. "HOXB13 is a susceptibility gene for prostate cancer: results from the International Consortium for Prostate Cancer Genetics (ICPCG)". In: *Human genetics* 132 (2013), pp. 5–14.

[17] Jennifer L Beebe-Dimmer et al. "The HOXB13 G84E mutation is associated with an increased risk for prostate cancer and other malignancies". In: *Cancer epidemiology, biomarkers & prevention* 24.9 (2015), pp. 1366–1372.

[18] Heng Li and Richard Durbin. "Fast and accurate short read alignment with Burrows–Wheeler transform". In: *bioinformatics* 25.14 (2009), pp. 1754–1760.

[19] Heng Li. "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM". In: *arXiv preprint arXiv:1303.3997* (2013). Accessed: April 07, 2025. URL: http://github.com/lh3/bwa.

[20] Erik Garrison and Gabor Marth. "Haplotype-based variant detection from short-read sequencing". In: *arXiv preprint arXiv:1207.3907* (2012). Accessed: April 07, 2025. URL: https://arxiv.org/abs/1207.3907v2.

[21] Aaron McKenna et al. "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data". In: *Genome research* 20.9 (2010), pp. 1297–1303.

[22] Mark A DePristo et al. "A framework for variation discovery and genotyping using next-generation DNA sequencing data". In: *Nature genetics* 43.5 (2011), pp. 491–498.

[23] Heng Li et al. "The sequence alignment/map format and SAMtools". In: *bioinformatics* 25.16 (2009), pp. 2078–2079.

[24] Zhi Wei et al. "SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data". In: *Nucleic acids research* 39.19 (2011), e132–e132. URL: https://doi.org/10.1093/NAR/GKR599.

[25] Daniel C Koboldt et al. "VarScan: variant detection in massively parallel sequencing of individual and pooled samples". In: *Bioinformatics* 25.17 (2009), pp. 2283–2285.

[26] Kai Wang, Mingyao Li, and Hakon Hakonarson. “ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data”. In: *Nucleic acids research* 38.16 (2010), e164–e164. URL: `https://doi.org/10.1093/NAR/GKQ603`.

[27] Michael S Lawrence et al. “Mutational heterogeneity in cancer and the search for new cancer-associated genes”. In: *Nature* 499.7457 (2013), pp. 214–218.

[28] Loris Mularoni et al. “OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations”. In: *Genome biology* 17 (2016), pp. 1–13.

[29] Brad T Sherman et al. “DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update)”. In: *Nucleic acids research* 50.W1 (2022), W216–W221. URL: `https://david-d.ncifcrf.gov/`.

[30] Edward Y Chen et al. “Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool”. In: *BMC bioinformatics* 14 (2013), pp. 1–14. URL: `http://amp.pharm.mssm.edu/Enrichr`.

[31] Liis Kolberg et al. “g: Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update)”. In: *Nucleic acids research* 51.W1 (2023), W207–W212. URL: `https://biit.cs.ut.ee/gprofiler`.