Identification of new biomarkers in Prostate Cancer

Lorena Magraner Pardo

Master´s Degree in Bioinformatics and Computational Biology

MÁSTERES DE LA UAM 2018 - 2019

Escuela Politécnica Superior

UAM Universidad Autónoma de Madrid

UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

Identification of new biomarkers in Prostate Cancer

Master´s Degree in Bioinformatics and Computational Biology

Author: Magraner Pardo, Lorena

Co-Director: Elena Castro, MD, PhD (CNIO)

Co-Director: José Ramón Valverde, MD, PhD (CNB-CSIC)

Ponent: Gonzalo Martínez, PhD (UAM)

June, 2019

Identification of new biomarkers in Prostate Cancer

Author: Magraner Pardo, Lorena

Co-Director: Elena Castro, MD, PhD (CNIO)

Co-Director: José Ramón Valverde, MD, PhD (CNB-CSIC)

Ponent: Gonzalo Martínez, PhD (UAM)

June, 2019

Summary

Background

Prostate cancer (PC) is one of the most heritable tumours as 57% of the interindividual risk is attributed to genetic factors. Germline mutations in DNA damage and repair (DDR) genes have been found in up to 11.8% of men with metastatic castration resistant prostate cancer (mCRPC), but the spectrum and prevalence of these mutations in mCRPC Spanish patients has not been established yet.

Methods

The germline DNA from two cohorts of mCRPC Spanish patients (419 and 93 from PROREPAIR-B and PRORADIUM) were screened for mutations in 107 and 55 DDR genes, respectively. We also analysed germline variants from the general population included in Exome Aggregation Consortium (ExAC) and CIBERER Spanish Variant Server (CSVS). Computational tools and databases such as ANNOVAR, ClinVar and dbNSFP were used to associate the phenotypic effect of the identified variants, followed by their classification by using the guidelines of the American College of Medical Genetics and Genomics. We also used Pfam and Interactome3D databases to analyse the distribution of pathogenic variants across protein sequences and three-dimensional structures.

Results

A total of 72 germline DDR (gDDR) pathogenic mutations were identified in 68 mCRPC patients (16.2%) in PROREPAIR-B, and 13 pathogenic mutations in 14 mCRPC patients (19.4%) in the PRORADIUM cohorts. The most recurrent mutated gene in PROREPAIR-B and PRORADIUM was BRCA2, with a significantly higher prevalence in Spanish cohorts than in the general non-cancer population (CSVS $P < .001$ and ExAC $P < .001$). Structural analysis in some of the DDR protein-coding genes affected revels clustering of the pathogenic variants in hotspots for different tumour types.

Conclusions

The incidence of germline mutations in genes related to the DDR processes in Spanish mCRPC patients is higher than in general non-cancer population, with BRCA2 as the most recurrent mutated gene. We also suggested that BRCA2 could be a key factor for the advance disease.

Key words

Prostate Cancer / DNA damage repair genes / Targeted exome sequencing / Bioinformatics / Functional annotation of variant

Index

Acronyms

ACMG/AMP	The American College of Medical Genetics and Genomics and the Association for Molecular Pathology
ADT	Androgen Deprivation Treatment
BWA	Burrows Wheeler Algorithm
CI	Confidence interval
CNV	Copy-number variant
CSVS	CIBERER Spanish Variant Server
DDR	DNA damage and repair
DGV	The Database of Genomic Variants
DNA	Desoxiribonucleic Acid
ESP	NHLBI GO Exome Sequencing Project
ExAC	Exome Agregation Consortium
GATK	Genome Analysis Tool Kit
gDDR	Germline DDR
gnomAD	The Genome Aggregation Database
HGP	Human genome project
INDELS	Small insertion and deletion variants
mCRPC	Metastatic Castration Resistant Prostate Cancer
NGS	Next-generation sequencing
NS	Not statistically significant
OR	Odds ratio
PARP	Poly (ADP-ribose) polymerase
PC	Prostate cancer
PCR	Polymerase chain reaction
PSA	Prostate specific antigen
SNP	Single-nucleotide polymorphisms
SNV	Single-nucleotide variant
VCF	variant call format
VUS	Variant of unknown significance
3D	Three-dimensional

1. Introduction

1.1. Landscape of clinical and pathological biomarkers to predict prognosis in patients with prostate cancer

Prostate cancer (PC) is the second most common cancer in men worldwide, with over 1 million newly diagnosed cases every year. Nonetheless, mortality has recently decreased as a result of early detection initiatives and screening programmes (Bray et al., 2018). A patient's risk to suffer from PC is currently assessed by combining three biomarkers: prostate specific antigen (PSA) levels, tumour staging and Gleason score (Mottet et al., 2017). However, these prognostic tools have some limitations (i.e., low specificity of PSA, intra and interoperator-dependency of histopathological reading). Although the majority of patients benefit from the current classification and undergo successful radical treatment with surgery (prostatectomy), radiotherapy or brachytherapy (Mottet et al., 2017), about a quarter of patients (20-25%) with localized disease relapse and progress to lethal metastatic castration resistant prostate cancer (mCRPC) (Chamberlain et al., 1997; Graham et al., 2008; Han et al., 2001; Kuban et al., 2005).

Limited clinical options are available to treat patients with mCRPC that are characterized by a very poor clinical outcome (survival at 5 years: 28.5%) (Surveillance, 2015). The increased number of lethal cases, partially related to the higher incidence of PC, leads to the need of identify and validate new biomarkers that could improve the specificity of the current ones, and at the same time to distinguish between aggressive and non-aggressive PC. Given the variability of this disease, the new biomarkers could improve the prediction of biochemical recurrence and metastatic progression, and could eventually result in prolonging patients' survival and their quality of life.

1.2. The role of DNA damage and repair genes in mCRPC

It has been established that the family history of PC patients is a risk factor for the disease, in addition to age and race (Parker et al., 2015; Mucci et al., 2016), positioning the PC as a one of the most hereditable tumour types. Moreover, inherited mutations associated with genes involved in DNA damage and repair (DDR) mechanisms are highly frequent in some hormone-driven cancers such as breast, ovarian and also PC. The high number of alterations in germline DDR (gDDR) genes detected in mCRPC patients (between 15 and 30% according to the

clinical setting and population background) compared with the prevalence of this alterations in localized PC (5%) or in the general population (3%) suggest an association with aggressiveness in PC (Pritchard et al., 2016).

DDR family member BRCA2 often appears as the most recurrent mutated gene with a significantly higher prevalence than other DDR genes across different mCRPC cohorts: 6% in 150 patients (Robinson et al., 2015), 5.35% in 692 patients (Pritchard et al., 2016), 3.51% in 313 patients (Na et al., 2017), 3,60% in 139 patients (Mijuskovic et al., 2018), 3,47% in 202 patients (Annala et al., 2018), and 2.91% in 172 patients (Antonarakis et al., 2018). It suggests that BRCA2 is an independent poor prognostic factor for PC with shorter metastasis-free survival and cause-specific survival (Castro et al., 2013; Castro et al., 2015; Kote-Jarai et al., 2015; Pritchard et al., 2016; Conteduca et al., 2018; Castro et al., 2019). In addition, mutation carriers in both germline and somatic alterations in DDR genes have a significant likelihood of developing aggressive/metastatic cancer, and these alterations could be biomarkers for some therapies like chemotherapy and PARP inhibitors (Mateo et al., 2015; Karzai et al., 2018; Cheng et al., 2016; Kaufman et al., 2015).

1.3. The potential use of Next-generation sequencing technologies for clinical genetic screening of patients with cancer

The public availability of the human genome sequence represents a very important impact for the biomedical research community. In 2003 the human genome project (HGP) released about 3.3 billion of bases from all of the 20.000-25.000 genes in the genome. Since this date, the increasing efficiency and the decreasing cost of Next-generation sequencing (NGS) analyses allows this technology to be rapidly introduced into many fields, and revolutionized biomedical research and the clinical practice of medicine (Mark et al., 2019).

Automation is critical for the routine use of the NGS technology in clinical and public health laboratory practices. Sequencing and bioinformatics technologies are rapidly evolving, so the integration of sequencing and data analysis in an efficient workflow is a challenge nowadays. Despite of that, NGS have been widely used in genomics research, in particular in the field of cancer.

The most comprehensive discovery, from the PanCancer project, suggested 299 driver genes across 9,423 tumour exomes. In addition, the putative number of missense driver mutations identified by the PanCancer project is larger than 3,400 (Bailey et al, 2018). Another important aspect emerging from the PanCancer project is that oncogenesis can be summarized in the following three facets: 1) somatic driver mutations and germline pathogenic variants, 2) influence of the tumour genome and epigenome on transcriptome and proteome, and 3) relationship between tumour and the microenvironment (Ding et al., 2018). Findings in the PanCancer project are in line with those obtained by NGS techniques in mCRPC cohorts (Hovelson and Tomlins, 2016; Barata et al. 2018).

Some of the discoveries in different mCRPC cohorts have been described before in section 1.2.

Taking into account the lack of knowledge or understanding to detect and better characterize PC patients with potentially lethal form of the mCRPC disease, and the potential use of NGS technologies in the genetic screening, we develop this research project for the Master's Degree in Bioinformatics and Computational Biology. We carry out the investigation about the genetic alterations in germline DDR (gDDR) genes in two mCRPC cohorts, PROREPAIR-B and PRORADIUM, in collaboration with more than 30 Spanish Hospital Centres. In the present work, we will focus on the Bioinformatics analysis and the interpretation of NGS data for more than 500 patients, in order to discover new information or to reach a new understanding in mCRPC molecular mechanisms that allow us a better classification of mCRPC patients.

2. Objectives

This research project aims to evaluate the prevalence and effect of gDDR mutations in mCRPC patients. We have analysed two prospective cohorts of mCRPC patients, PROREPAIR-B (Castro et al., 2019) and PRORADIUM (an ongoing research), enrolled in more than 30 different institutions.

The main objective in this work for the Master's Degree is to produce and automatize a bioinformatics pipeline to study targeted-sequencing data with a custom-designed panel of DDR genes. This pipeline allows us:

1) To evaluate the prevalence and effect of gDDR mutations in the PROREPAIR-B and PRORADIUM cohorts of mCRPC Spanish patients.
2) To compare the mutational status in PROREPAIR-B and PRORADIUM cohorts with other published studies in mCRPC patients.
3) To do a comparative study of the distribution of pathogenic mutations in DDR genes with non-cancer general population data in ExAC and CSVS datasets.
4) To study the distribution of the identified mutations in the protein sequences and 3D-structures.
5) To discover other kind of DNA alterations (i.e. copy number variations) in the whole gene sequence of BRCA1, BRCA2, ATM, PALB2 and MSH2 from the PRORADIUM cohort.

3. Material and Methods

3.1. Patients enrolled

Patients with histologically confirmed PC and unknown mutational status were enrolled at the time of metastatic castration-resistant diagnosis and observed until death. All patients provided informed consent at study entry and passed the eligibility criteria. The complete list of inclusion and exclusion criteria for the study is provided in the appendix section of Castro et al., 2019.

PROREPAIR-B cohort

The 419 patients enrolled in the project –from 38 Spanish Hospital Centres– came from the same "at risk" population, characterized by histological confirmation of mCRPC and unknown mutational status. This study did not dictate any treatment, and all the patients were treated at the discretion of the physicians (Castro et al., 2019).

PRORADIUM cohort

The 93 patients enrolled in the project –from 33 Spanish Hospital Centres– came from the same "at risk" population, characterized by CRPC and metastasis in the bones, but not to other parts of the body. They are resistant to Androgen Deprivation Treatment (ADT) or other surgical treatment that reduce testosterone levels. All these patients have received the standard treatment Radium223.

3.2. Experimental processing of the samples

The experimental processing of the samples was the same for the two cohorts PROREPAIR-B and PRORADIUM, and was done entirely by the Prostate Cancer Clinical Unit members at CNIO. All the experimental details are provided in Castro et al., 2019. In brief, germline DNA was extracted from 5-ml blood samples and purified. Library preparation was done using a custom NimbleGen SeqCap XL Target Enrichment (Roche, Pleasanton, CA) panel. The different steps that includes DNA fragmentation, adapter ligation with a barcode to identify each patient, fragments selection and amplification using a primer from the extreme of the

adaptor, and finally, DNA purification and quantification are summarized in the **Figure 1**.

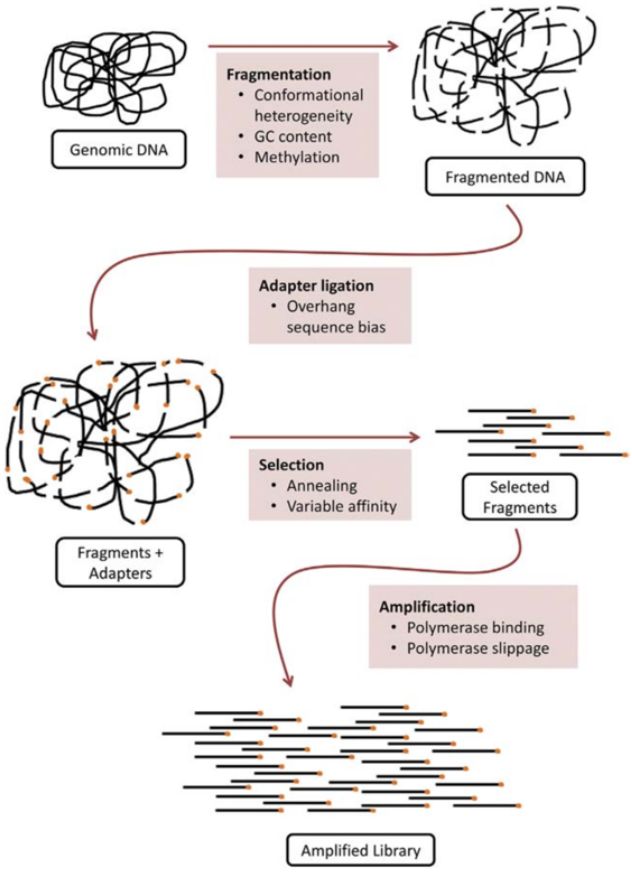


***Figure 1**. General overview of the molecular steps involved in a typical library preparation protocol (figure taken from Bohannan et al., 2019)*

3.3. Target sequencing

The DNA libraries were read using Illumina NexSEquation 500 (Illumina, San Diego, CA) at IdiPAZ (INGEM, Madrid, Spain) and Illumina HiSeqX in Genohub Inc. (Austin, TX, USA) sequencing platforms, for PROREPAIR-B and PRORADIUM studies, respectively. In chronological order, the PROEPAIR-B cohort was analysed first. The study panel was composed of 107 DDR genes (coding regions) in order to include the largest possible number of DDR genes related with cancer predisposition, and the analysis showed that only 28 genes were pathogenic mutations carriers. After that, we studied the PRORADIUM cohort, and we decided to dispose of the genes in which neither we nor other groups had found pathogenic alterations. In this case we studied a panel compose of 55 DDR genes associated with cancer predisposition syndrome and in which we found mutations in the PROREPAIR-B cohort.

In addition, given the high frequency of the alterations in ATM, BRCA1, BRCA2, MSH2 and PALB2, it was decided to sequence these whole genes and not only the coding regions, in order to analyse other types of DNA alterations (i.e. copy number variants). The complete list of genes is shown in **Table 1**.

***Table 1**. Full list of the DDR genes included in custom panels and screened in the PROREPAIR-B (panel A) and PRORADIUM (panel B) cohorts. In bold the genes that were subject to whole sequencing in PRORADIUM.*

A)

APEX1	DDB1	FAM175B	GTF2H3	MSH4	ARP2	RAD51D	XAB2
APEX2	DMC1	FANCA	GTF2H4	MSH5	PARP3	RAD52	XPA
APLF	EME1	FANCB	GTF2H5	MSH6	PMS1	RAD54B	XPC
ATM	EME2	FANCC	KIAA0415	MUS81	PMS2	RAD54L	XRCC1
ATR	EPCAM	FANCD2	LIG4	MUTYH	PNKP	RBBP8	XRCC2
BARD1	ERCC1	FANCE	MBD4	NBN	PRKDC	RPA1	XRCC3
BRCA1	ERCC2	FANCF	MLH1	NEIL1	RAD9A	RPA2	XRCC4
BRCA2	ERCC3	FANCG	MLH3	NEIL2	RAD17	RPA3	XRCC5
BRIP1	ERCC4	FANCI	MMS19	NEIL3	RAD23A	SLX1A	XRCC6
CDK7	ERCC5	FANCL	MNAT1	NHEJ1	RAD23B	SLX1B	
CDK12	ERCC6	FANCM	MPG	NTHL1	RAD50	SLX4	
CHEK1	ERCC8	GEN1	MRE11A	OGG1	RAD51	SMUG1	
CHEK2	FAAP20	GTF2H1	MSH2	PALB2	RAD51B	TDG	
DCLRE1C	FAAP24	GTF2H2	MSH3	PARP1	RAD51C	UNG	
Panel gene PROREPAIR-B							

B)

ATM	ERCC3	FANCM	RAD51C
ATR	ERCC5	GEN1	RAD51D
BAP1	ERCC6	MLH1	SLX4
BARD1	FAM175A	MRE11A	XRCC2
BLM	FAM175B	**MSH2**	
BRCA1	FAN1	MSH3	
BRCA2	FANCA	MSH6	
BRIP1	FANCC	MUTYH	
CDK7	FANCD2	NBN	
CHEK1	FANCE	**PALB2**	
CHEK2	FANCF	PMS2	
DCLRE1C	FANCG	POLD1	
EPCAM	FANCI	POLE	
ERCC2	FANCL	RAD51B	
Panel gene PRORADIUM			

3.4. Bioinformatics pipeline for the analysis and prioritization of mCRPC-related variants

In general, a Bioinformatics pipeline for the analysis of single-nucleotide variants (SNV) and copy-number variants (CNV) is shown in **Figure 2**. The first steps in "Data Processing" (i.e., evaluate the reads quality, alignment against a reference genome, remove duplicate reads, sort and indexing) are well established and are common in different types of NGS analyses. The remaining steps "Variant Discovery", "Variant Annotation", and "Variant Prioritization" are more diverse in the use of a particular computational tool. Nevertheless, for these steps some consensus approaches or guidelines exist, such as the GATK workflow at Broad Institute, as we will discuss in the sections below.

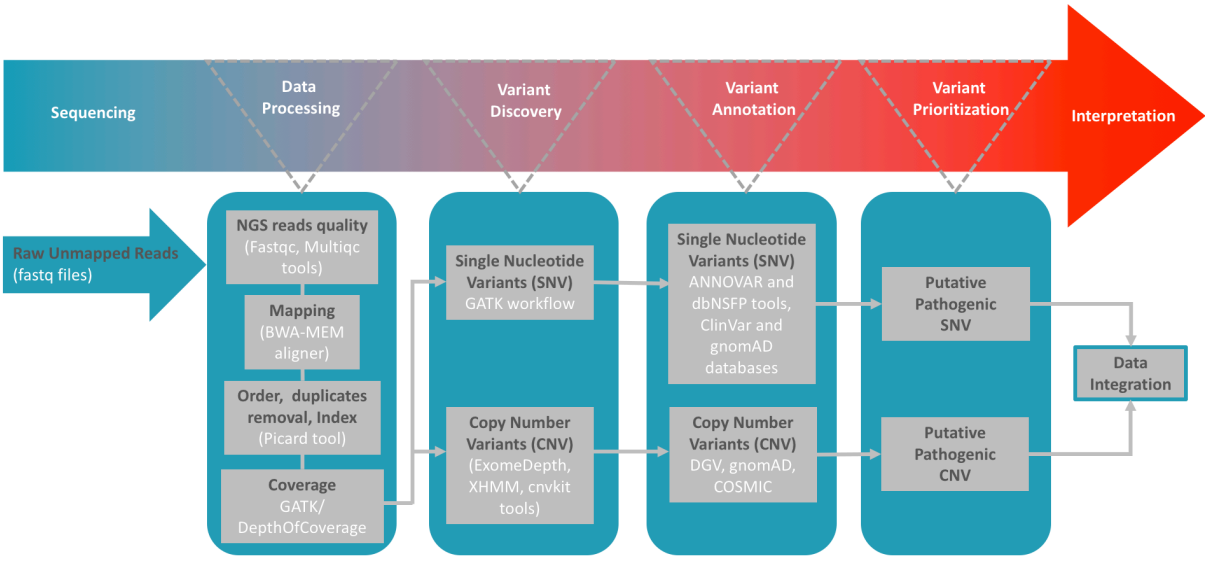


***Figure 2**. Bioinformatics pipeline for the analysis of single-nucleotide variants (SNV) and copy-number variants (CNV).*

3.4.1. Data Processing

Quality control:

The most widely used software for this task are FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and multiQC (https://github.com/MultiQC). This combination of software provides us with different metrics for a qualitative and quantitative checking of the sequencing process. The main objective is to use high-quality reads to get accurate Bioinformatics analysis.

Some of the metrics we used are listed below. For more details, see **Supplementary information S1 to S6** which includes quality metrics in the PRORADIUM cohort.

- The general quality of the reads, and quality of each base according to the phred quality score (green area represents high quality).
- The coverage per sample.
- The CG content.
- The number of bases not specified (content of 'N' in the reads).
- The adapter content.
- Allele Frequencies Ratio

Alignment:

The paired-end sequencing reads in the fastq files were mapped to the human genome reference sequence (GRCh37) using the Burrows-Wheeler aligner (BWA-MEM) v.0.7.15 (http://bio-bwa.sourceforge.net), which is recommended by Illumina for mapping low-divergent sequences, for high-quality queries, and also, it is faster and more accurate. The heuristic of local alignment is based on generating initially seeds an alignment with supermaximal exact matches (SMEMs) by picking out best partial alignments first and prune out alignments considered sub-optimal. Moreover, BWA is the aligner recommended by the Genome Analysis Toll Kit (GATK) workflow (https://software.broadinstitute.org/gatk/) that we will use in following steps. This step is crucial for the rest of the experiment, because sequencing depth (or coverage) is directly affected by the accuracy of genome alignment algorithms, and also, is correlated with the accuracy of the allele calls.

Order reads, marks duplicate and removal, indexing:

The output files in BAM/SAM format, containing the alignment results, were transformed with Picard tools v.2.1.0 (http://picard.Sourceforget.net). Sort, mark and exclusion of duplicates, and building indexed files were done using the Picard tools options SortSam, MarkDuplicates, RemoveDuplicate and BuildBamIndex, respectively. All programs were executed with default parameters.

Coverage:

In order to get coverage metrics from the samples, we used two different tools: CollectWgsMetrics (from Picard tools) and genomeCoverageBed (from SAM Tools; https://github.com/samtools/samtools). CollectAlignmentSummaryMetrics (from Picard tools) was also used to get alignment metrics.

3.4.2. Variant Discovery

Single nucleotide variants calling:

The following steps adhere to the best practices developed by the GATK team for high-quality SNV/INDELS calls. The GATK workflow implements the following functionalities:

- "IndelRealigner" performs the local realignment of the INDELS regions to reduce the artefacts in the Single Nucleotide Polymorfisms (SNP) caller.
- "BaseRecalibrator" allows minimizing the systematic errors. The variant detection is based on the quality scores assigned to each base during the sequencing process. Systematic errors could affect the scores, and recalibration of values to adjust the qualities is necessary.
- "HaplotypeCaller" allows performing the variant calling *per se*. This function provides one VCF file per patient, which contain SNPs and INDELS simultaneously.

The VCF format is a tab delimited text file that contains meta-information lines, a header line, and then, data lines each containing information about the variants identified in the cohort.

Copy number variants detection:

To detect CNV alterations in ATM, BRCA1, BRCA2, MSH2 and PALB2 whole-gene sequenced in the PRORADIUM cohort, we used read-depth (or depth of sequencing) calculations and also background read-count distributions. For the CNV analysis we used three different tools to reach consensus results:

- ExomeDepth (https://rdrr.io/cran/ExomeDepth/)
- XHMM (https://github.com/RRafiee/XHMM)
- CNVkit (https://cnvkit.readthedocs.io/en/stable/)

These computational tools cover the analysis of whole-exome as well as targeted whole-genome sequence data, and were executed with the default parameters proposed by authors.

3.4.3. Variant Annotation

Single nucleotide variants (SNV):

To enrich the preliminary annotations of the variants obtained after the variant calling, we used the following repositories and computational tools: ANNOVAR (http://annovar.openbioinformatics.org/en/latest/) and dbNSFP (https://sites.google.com/site/jpopgen/dbNSFP). These tools aggregate information from multiple sources, and provide us annotations with three different levels: gene-based, region-based and filter-based:

- Gene-based annotation: Provides information about the relationship and functional impact on the gene, identifying whether the alteration is affecting the protein-coding region, and the amino acid affected.
- Region-based annotation: Provides information about variants that localize within specific conserved regions, such as predicted transcription factor binding sites or ChIP-seq peaks, transcripts overexpressed in RNA-seq experiments, among other annotations on genomic intervals.
- Filter-based annotation: Provides information about documented variants. For example, the population frequency reported by different projects (i.e. 1000 Genome Project, The Exome Aggregation Consortium (ExAC), The Genome Aggregation Database (gnomAD), and NHLBI GO Exome Sequencing Project (ESP)). Also, collect prediction scores from SIFT (https://sift.bii.a-star.edu.sg/), PolyPhen2 (http://genetics.bwh.harvard.edu/pph2/), LRT (http://www.genetics.wustl.edu/jflab/lrt_query.html), MutationTaster (http://www.mutationtaster.org/), MutationAssessor (http://mutationassessor.org/r3/), FATHMM (http://fathmm.biocompute.org.uk/), REVEL (https://sites.google.com/site/revelgenomics/).

We also used the ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/) database that provides assertions about clinical significance of germline variants, information about the submitter, and other complimentary data.

Copy Number Variants (CNV):

In-house scripts were developed to annotate CNV calls by incorporating orthogonal data from the Genome Aggregation Database (gnomAD; https://gnomad.broadinstitute.org/) the Database of Genomic Variants (DGV; http://dgv.tcag.ca/dgv/app/home), and gene annotations using GRCh37 (http://hgdownload.cse.ucsc.edu/). The scripts were written in R language (https://www.r-project.org/). At the moment, we are developing a set of filtering criteria to build a CNV dataset of high confidence. This part of the project is under development.

3.4.4. Variant Prioritization

The pathogenicity of germline variants was predicted following the consensus criteria provided by the guidelines of The American College of Medical Genetics and Genomics, and the Association for Molecular Pathology (ACMG/AMP) (Richards et al., 2015). The guidelines provide the basic goals and updated rules to standardize and reduce annotation inconsistencies between laboratories, and also, for an appropriate clinical interpretation of genetic variants. All the predicted deleterious variants were manually curated against the published literature and public databases, including ClinVar. In addition, we discarded polymorphisms with a minor allele frequency of 1% or higher (Brookes et al., 1999) according to 1000G and ExAC databases.

3.5. General non-cancer population databases

To compare the prevalence of the prioritized variants in PRORADIUM and PROREPAIR cohorts with their frequency in the general non-cancer population, we used two different databases: The Exome Aggregation Consortium (ExAC; http://exac.broadinstitute.org/) non-cancer data with N = 53,105, and CIBERER Spanish Variant Server (CSVS; http://csvs.babelomics.org/) with N = 1,551. The two general cohorts ExAC and CSVS were annotated, filtered and reviewed using the same criteria than in PROREPAIR-B and PRORADIUM cohorts.

4. Results and Discussion

4.1. Prevalence and effect of the gDDR mutations in PROREPAIR-B and PRORADIUM cohorts

An exhaustive quality control of the sequenced samples was assessed by a range of markers, and those samples that didn´t pass the criteria were discarded. Some of the markers studied were: the general quality of the reads and quality of each base according to the phred quality score, the coverage per sample, the CG content and the distribution of the allele frequencies (**Supplementary S1-S6**).

In the PROREPAIR-B cohort we identified 68 carriers (16.2%) out of 419 eligible patients with predicted pathogenic mutations in 28 genes (**Figure 3** and **Supplementary S7**). Five of the patients contain 2 pathogenic mutations in two different genes (RAD17 and FANCL in patient1, MUTYH and FANCG in patient2, FANCD2 and ERCC3 in patient3, FANCL and EME2 in patient4, DCLRE1C and MUTYH in patient5). The vast majority of the deleterious variants (77%) are frame-shifts INDELS or stop-gains, which alters the protein sequence as well as the 3D-structure. Besides, the most recurrent mutated genes were BRCA2, MUTYH, ATM and BRCA1 that account for 39 (54.16%) out of 72 mutations.

In the PRORADIUM cohort we have expanded the perspective of the Bioinformatics study by analysing not only the pathogenic mutations, but also, evaluating the number of germline mutational per patients. We identified 14 carriers (19.4%) out of 93 eligible patients with pathogenic mutations (depicted with red circle in **Figure 4** and **Supplementary S8**). One of the patients contains 2 pathogenic mutations in two different genes (BRCA1 and CHEK2). Moreover, this patient shows the highest number of germline mutational in the PRORADIUM cohort. As shown in the legend of **Figure 4 panel C**, the vast majority of the carriers (12 out of 14) with pathogenic mutation are located in the first quadrant on the heat map. This quadrant boxed with a red line also indicates the most recurrent mutated genes and patients. In the PRORADIUM cohort we also predicted 55.2% of the alterations as variants of unknown significance (VUS).

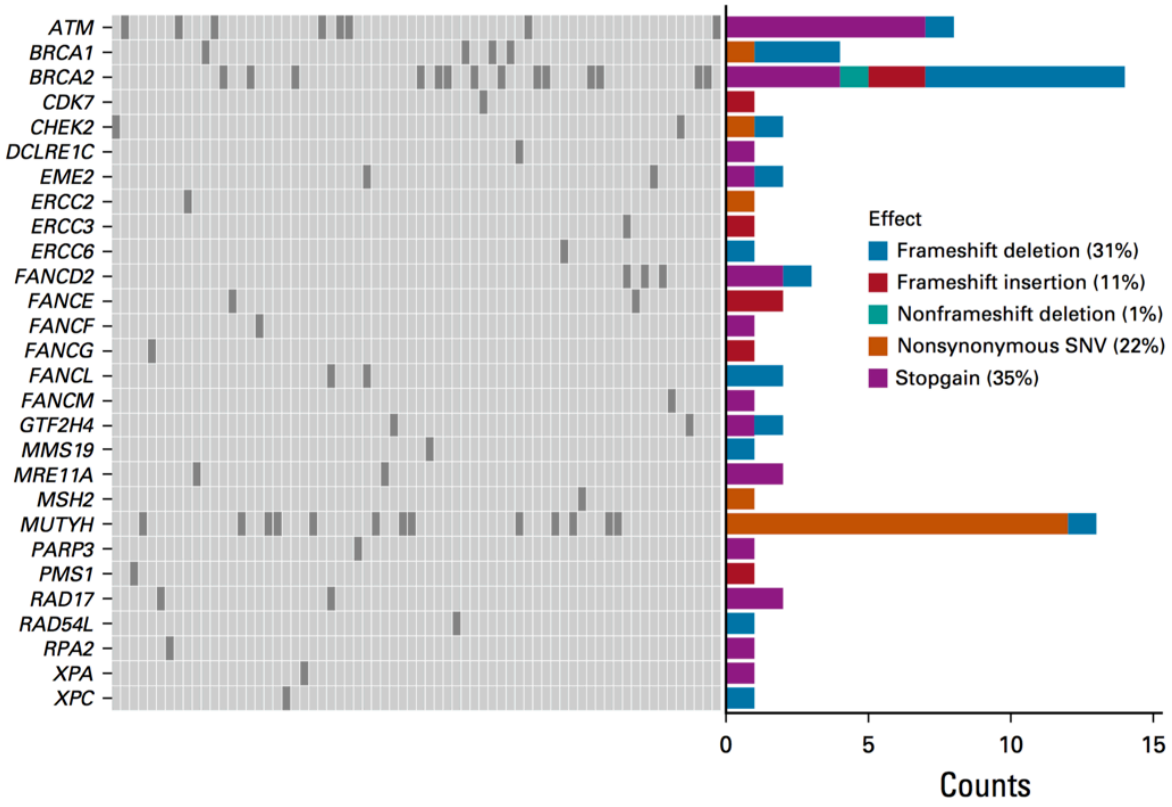


Figure 3. *Distribution of the pathogenic and likely pathogenic mutations identified in PROREPAIR-B cohort.*

As previously observed in the PROREPAIR-B cohort, the vast majority of the deleterious variants in PRORADIUM (**Supplementary S8**) (64.28%) are frame-shifts and stop-gains. Also, the most recurrent mutated genes with pathogenic variants were BRCA2, MUTYH and ATM that account for 78.57% of the mutations (12 out of 14 mutations). Moreover, the median number of genomic alterations per patient (discarding patient with two pathogenic mutations) was 3 (range 1 to 9) (depicted in horizontal discontinuous red line in **Figure 4 panel B**). All these results concur with previously published data by Ikeda et al. (2018), and at the same time expands the current knowledge about the contribution of germline variants to the molecular mechanisms in mCRPC patients (see sections below).

According to the six classes of base substitution in the PRORADIUM cohort (**Figure 4 panel D**), we observed three overrepresented base transitions: C-to-T (28.9%), G-to-A (21.6%), and A-to-G (18.5%). From the literature we know that enrichment in C-to-T transitions is observed in the mutational signatures 1A/B, 6, 7, 11, 15, and 19 described by Alexandrov and colleagues (Alexandrov et al., 2013). Remarkably, mutational signatures 1A and 6 have been validated for PC, and signature 6 is associated with DNA MMR deficiency. On the opposite site, transversions, which are not enriched in PRORADIUM, are the most abundant class in smoking-associated cancers (e.g., lung, liver, head and neck).

We review the landscape of variants identify in PROREPAIR-B and PRORADIUM in Pfam domains, proteins 3D-structure and hotspots regions in section 4.4.

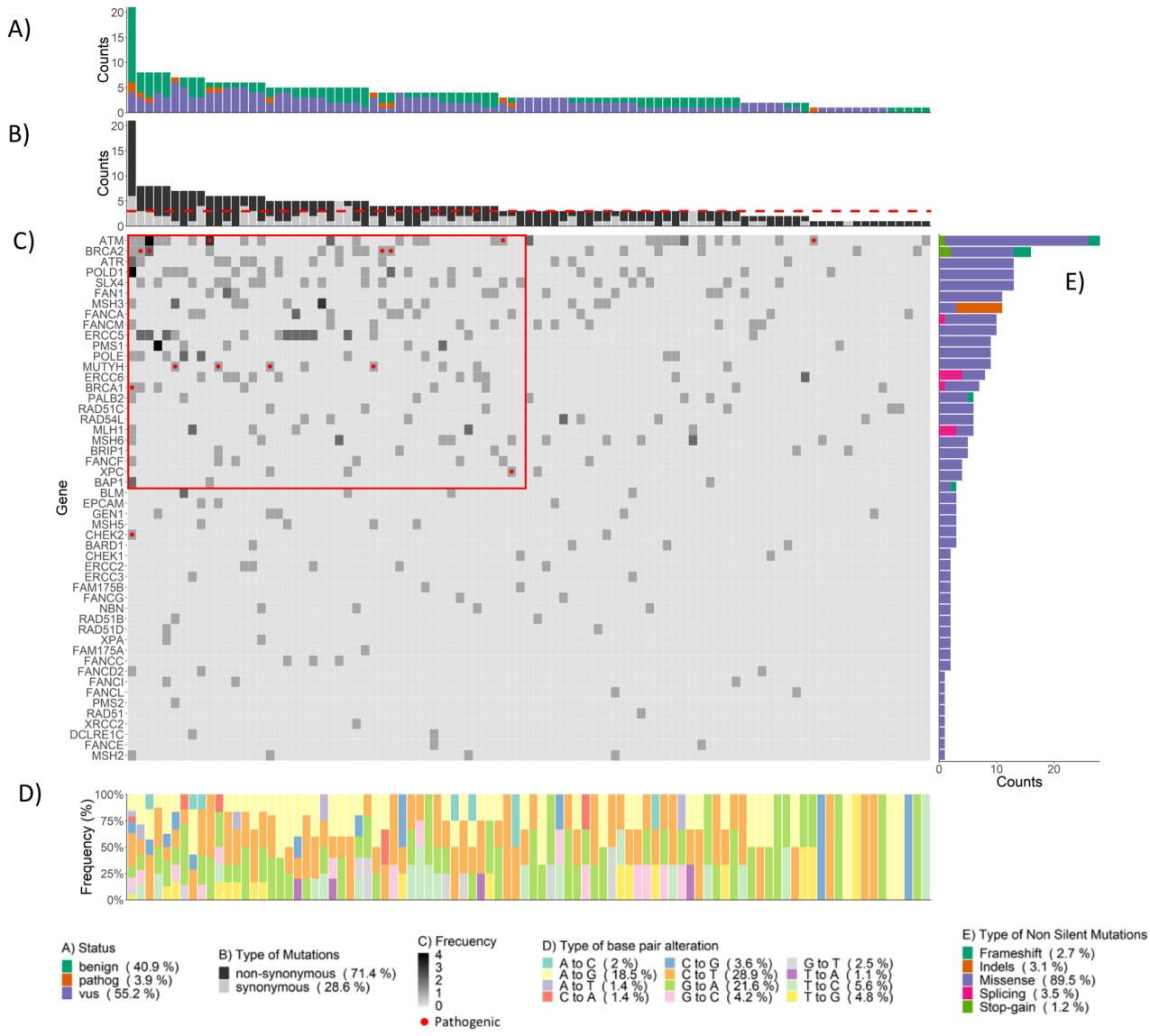


Figure 4. *Overview of the mutations identified in PRORADIUM cohort. A) Bar graph representing the frequency of mutational status described as a benign, pathogenic and variants of unknown significance (VUS) B) Bar graph representing the frequency of synonymous (black) and non-synonymous (grey) genetic alterations C) Heat map showing the number of germline mutations per patient (ordered by decreasing number of mutations, from left to right) in each gene (ordered by decreasing number of mutations, from top to bottom). Dark grey represents more alterations, and red circle represents pathogenic mutations D) The type of base pair alteration within the samples is plotted above the heat map E) The frequency of the type of non silent mutation is described on the right bar graph.*

4.2. Comparison of the mutational status in PROREPAIR-B and PRORADIUM cohorts with other published studies.

Since, mutation rates may differ significantly among groups of PC patients, in this study, we also analysed small and large cohorts with different genetic backgrounds. The results in **Table 2** indicates that despite a different genetic background of the ethnic groups, BRCA2 is the most affected gene. In addition, the number of pathogenic germline variants identified in BRCA2 are in the same order of magnitude in the Spanish cohorts PROREPAIR-B and PRORADIUM (3.34% and 4.30%, respectively), that those published using different mCRPC cohorts.

***Table 2**. Number of pathogenic germline variants identified in ATM, BRCA1 and BRCA2 in different mCRPC publish cohorts, and also, in the PROREPAIR-B and PRORADIUM Spanish cohorts.*

	ATM	BRCA1	BRCA2	total	N population	%BRCA2 mutated
Robinson et al., 2015	8	4	9	21	150	6.00
Pritchard et al., 2016	11	6	37	54	692	5.35
Na et al., 2017	6	2	11	19	313	3.51
Mijuskovic et al., 2018	4	0	5	9	139	3.60
Annala et al., 2018	1	0	7	8	202	3.47
Antonarakis et al., 2018	3	1	5	9	172	2.91
PROREPAIR-B	8	4	14	26	419	3.34
PRORADIUM	3	1	4	8	93	4.30

Based on these results compared to others available in the literature, we claimed in Castro et al., 2019 that BRCA2 is the most frequently altered DDR gene in unselected patients with mCRPC and germline mutations in BRCA2 have a deleterious impact on mCRPC outcomes.

4.3. Comparison of the prevalence of gDDR mutations in PROREPAIR-B and PRORADIUM cohorts with general non-cancer populations datasets ExAC and CSVS

The prevalence of pathogenic mutations identified in the mCRPC Spanish cohorts PROREPAIR-B and PRORADIUM, and the gDDR mutations in general non-cancer population datasets ExAC and CSVS is shown in **Table 3** and **Table 4**. The odds ratio (OR) for carrier status in mCRPC compared with non-cancer populations were substantially increased for several DDR genes, including ATM, BRCA1 and BRCA2.

The prevalence of ATM, BRCA2 and MUTYH mutations was significantly higher in mCRPC (PROREPAIR-B and PRORADIUM) than in CSVS and ExAC non-cancer populations. Complete data about statistical analyses are described in **Table 3** and **Table 4.** Briefly, all *P* values were two sided. The comparison of the prevalence of mutations between mCRPC patients and the general non-cancer populations were tested using the $\chi 2$, Fisher's exact, or Mann-Whitney U test, as appropriate (Altman et al., 1991). The SPSS version 19 (SPSS, Chicago, IL) and R version 3.3.3 (https://www.r-project.org/) programs were used for the statistical analysis.

***Table 3**. Prevalence of deleterious germline variants in the DDR genes analysed in PROREPAIR-B compared with non-cancer populations –the Exome Aggregation Consortium (ExAC) and the CIBERER Spanish Variant Server (CSVS) population*

	PROREPAIR (n = 419)		ExAC (n = 53,105)		CSVS (n = 1,551)		PROREPAIR v ExAC			PROREPAIR v CSVS		
Gene	**No.**	**%**	**No.**	**%**	**No.**	**%**	**OR**	**(95% CI)**	***P***	**OR**	**(95% CI)**	***P***
ATM	8	1.91	139	0.26	3	0.19	7.4	(3.6 to 15.2)	<.001	10	(2.7 to 38.0)	<.001
BRCA1	4	0.95	97	0.18	2	0.13	5.3	(1.9 to 14.4)	<.001	7.5	(1.4 to 40.9)	0.021
BRCA2	14	3.34	159	0.30	4	0.26	11.50	(6.6 to 20.1)	<.001	13.4	(4.4 to 40.8)	<.001
CDK7	1	0.24	11	0.02	1	0.06	11.50	(1.5 to 89.6)	<.001	3.70	(0.2 to 59.4)	NS
CHEK2	2	0.48	273	0.51	1	0.06	0.9	(0.2 to 3.7)	NS	7.40	(0.7 to 82.2)	NS
DCLRE1C	1	0.24	60	0.11	1	0.06	2.1	(0.3 to 15.3)	NS	3.70	(0.2 to 59.4)	NS
EME2	2	0.48	89	0.17	1	0.06	2.90	(0.7 to 11.6)	NS	7.40	(0.7 to 82.2)	NS
ERCC2	1	0.24	230	0.43	0	0.00	0.50	(0.1 to 3.9)	NS			NS
ERCC3	1	0.24	102	0.19	1	0.06	1.20	(0.2 to 8.9)	NS	3.7	(0.2 to 59.4)	NS
ERCC6	1	0.24	62	0.12	1	0.06	2	(0.3 to 14.8)	NS	3.7	(0.2 to 59.4)	NS
FANCD2	3	0.72	77	0.14	1	0.06	5.00	(1.6 to 15.8)	0.0251	11.3	(1.0 to 107.7)	NS
FANCE	2	0.48	34	0.06	0	0.00	7.5	(1.8 to 31)	0.0323			0.0452
FANCF	1	0.24	178	0.34	0	0.00	0.7	(0.1 to 5.1)	NS			NS
FANCG	1	0.24	25	0.05	1	0.06	5.10	(0.7 to 37.6)	NS	3.70	(0.2 to 59.4)	NS
FANCL	2	0.48	15	0.03	0	0.00	17.00	(3.9 to 74.5)	0.0077			0.0452
FANCM	1	0.24	178	0.34	1	0.06	0.7	(0.1 to 5.7)	NS			NS
GTF2H4	2	0.48	4	0.01	0	0.00	63.7	(11.6 to 348)	<.001			0.0452
MMS19	1	0.24	12	0.02	0	0.00	10.6	(1.4 to 81.6)	<.001			NS
MRE11A	2	0.48	32	0.06	0	0.00	8	(1.9 to 33.3)	0.0291			0.0452
MSH2	1	0.24	21	0.04	4	0.26	6.00	(1.0 to 45.0)	NS	0.9	(0.1 to 8.3)	NS
MUTYH	13	3.10	729	1.37	14	0.90	2.3	(1.3 to 4.0)	0.0091	3.5	(1.6 to 7.5)	0.0016
PARP3	1	0.24	195	0.37	5	0.32	0.6	(0.1 to 4.6)	NS	0.70	(0.1 to 6.3)	NS
PMS1	1	0.24	71	0.13	3	0.19	1.8	(0.2 to 12.9)	NS	1.20	(0.1 to 11.9)	NS
RAD17	2	0.48	17	0.03	3	0.19	15	(3.4 to 65.0)	0.0096	2.5	(0.4 to 14.9)	NS
RAD54L	1	0.24	48	0.09	0	0.00	2.6	(0.4 to 19.2)	NS			NS
RPA2	1	0.24	10	0.02	1	0.06	12.70	(1.6 to 99.5)	NS	3.70	(0.2 to 59.4)	NS
XPA	1	0.24	40	0.08	0	0.00	3.20	(0.4 to 23.1)	NS			NS
XPC	1	0.24	338	0.64	8	0.52	0.40	(0.1 to 2.7)	NS	0.5	(0.1 to 3.7)	NS

Table 4*. Prevalence of deleterious germline variants in the DDR genes analysed in PRORADIUM compared with non-cancer populations –the Exome Aggregation Consortium (ExAC) and the CIBERER Spanish Variant Server (CSVS) population.*

	PRORADIUM		**ExAC**		**CSVS**							
	(n = 93)		**(n = 53,105)**		**(n = 1,551)**		**PRORADIUM v ExAC**			**PRORADIUM v CSVS**		
Gene	**No.**	**%**	**No.**	**%**	**No.**	**%**	**OR**	**(95% CI)**	***P***	**OR**	**(95% CI)**	***P***
ATM	3	3.23	139	0.26	3	0.19	12.70	(3.97 to 40.61)	0.002202	17.20	(3.42 to 86.43)	0.0033
BRCA1	1	1.08	97	0.18	2	0.13	5.94	(0.82 to 43.05)	NS	8.42	(0.76 to 93.70)	NS
BRCA2	4	4.30	159	0.30	4	0.26	14.97	(5.43 to 41.25)	<.001	17.38	(4.28 to 70.65)	<.001
CHEK2	1	1.08	273	0.51	1	0.06	2.10	(0.29 to 15.15)	NS	16.85	(1.05 to 271.52)	NS
MUTYH	4	4.30	729	1.37	14	0.90	3.23	(1.18 to 8.82)	0.044	4.93	(1.59 to 15.30)	0.01772
XPC	1	1.08	338	0.64	8	0.52	1.70	(0.24 to 12.21)	NS	2.10	(0.26 to 16.94)	NS

4.4. Distribution of the gDDR mutations in the protein sequence and 3D-structure

Distribution of pathogenic gDDR mutations across functional domains

The distribution of the pathogenic variants in DDR protein-coding genes across functional domains annotated in the Pfam database (https://pfam.xfam.org/) is shown in **Figure 5**. The complete list of pathogenic gDDR mutations in PROREPAIR-B and PRORADIUM are provided in **Supplementary S7** and **S8**.

We observed that pathogenic gDDR mutations are frequently located outside the functional protein domains. Moreover, 56 out of 72 variants (77%) identified in PROREPAIR-B are frameshifts and stop-gain, whereas 9 out of 14 variants (64.28%) in PRORADIUM. This type of mutations truncates the protein sequence and consequently impacts their three-dimensional structure and function.

The accumulation of frameshifts and stop-gain mutations outside functional protein domains is somehow expected because of inactivation of gene products. We hypothesize that this type of mutations may increase protein degradation of DDR regulators.

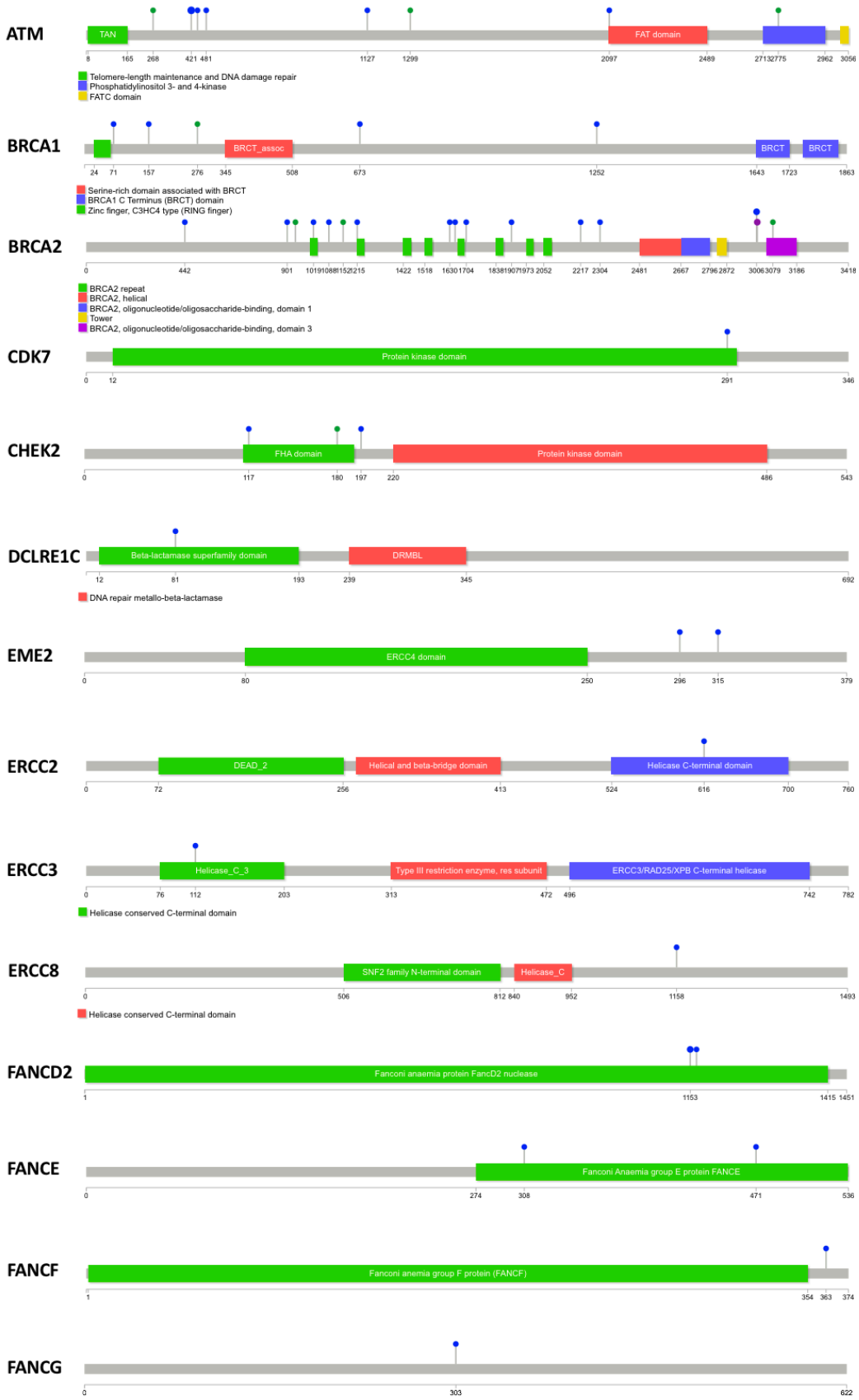


Figure 5. *Lolliplots showing the distribution of the presumed pathogenic mutations identified in PROREPAIR (blue), PRORADIUM (green), both projects (purple).*

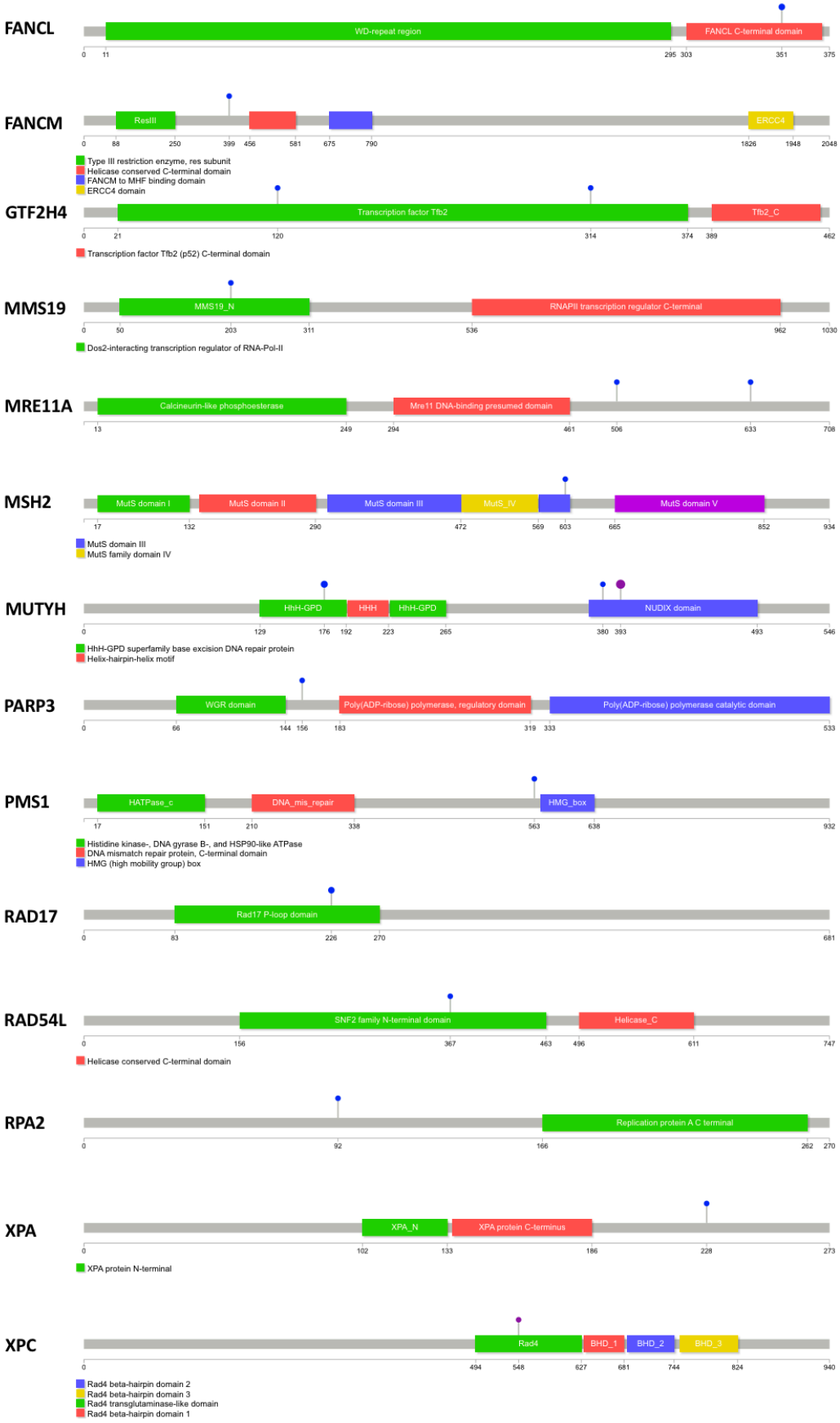


Figure 5. *(Continued)*

Distribution of pathogenic gDDR mutations across protein 3D-Structure

Only 2 of the 28 affected proteins-coding genes in the two mCRPC Spanish cohorts had three-dimensional (3D) structures available at PDB (http://www.rcsb.org/). Therefore, we focused the structural analysis on these two proteins CHEK2 (PDB ID: 1GXC, X-ray diffraction at 2.7Å resolution) and MUTYH (PDB ID: 1X51, NMR solution). After mapping the pathogenic gDDR mutations onto 3D-structure (**Figure 6**), we found that CHEK2 variant p.R117G (PROREPAIR-B) is localized in the phosphopeptide binding site, so it could impair substrate recognition. Interestingly, two other variants identified in this study, p.K197fs (PROREPAIR-B) and p.R180C (PRORADIUM), are clustered in the protein regions where variants for other tumour types have already been described in COSMIC. For example, the variant p.K197fs (PROREPAIR-B) is in the close vicinity of p.N196I (mouth) and p.V198L (rectum and breast) previously observed in COSMIC, and p.R180C (PRORADIUM) is located in the same region than p.R180H (in colon, bladder and PC), p.R181L (lung) and p.R182S (in colon and melanoma), also described in COSMIC.

A similar scenario was observed for MUTYH, where p.L380fs (PROREPAIR-B) is located in the same region than the variant p.G381V (stomach) described in COSMIC, and p.G393D (PRORADIUM) are also located in a hotspot site with neighbouring variants described for different tumour type in COSMIC, such as p.Q388* (colon), p.L398P (bladder), and p.W399S (colon).

These findings suggest that accumulation of variants in these regions impairs protein-protein interaction interfaces, and also, the biological function of the protein. The co-localizing and co-clustering of somatic mutations and germline variants onto protein 3D-structure have been applied to link rare predisposition variants to functional consequence (Huang et al., 2018). Therefore, to study the distribution of the remaining pathogenic variants in the 3D-struture of DDR affected proteins, may provide key insights in understanding pathogenicity of these alterations, and how these alterations may have an effect in protein stability and function.

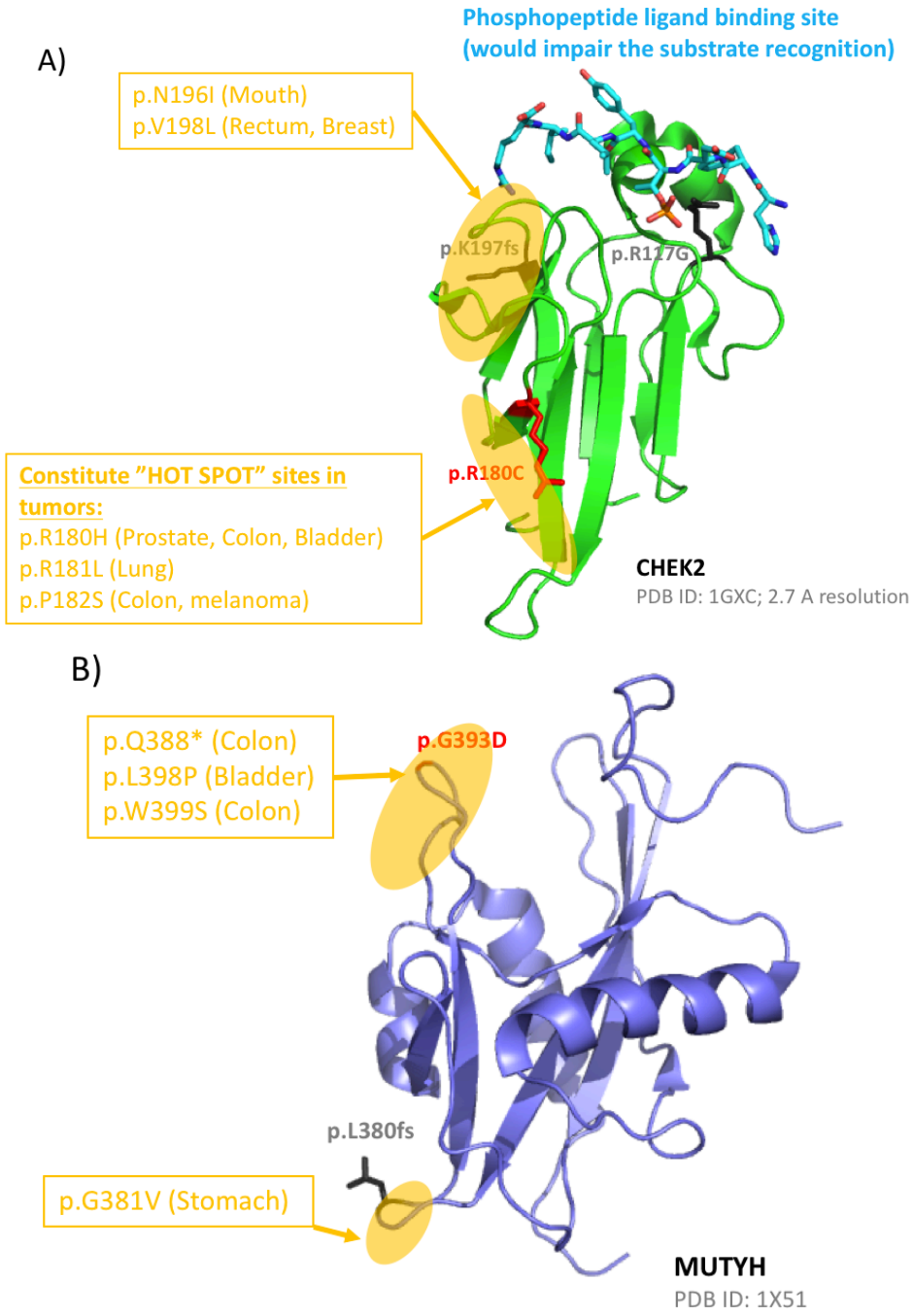


Figure 6. *Localization of presumed pathogenic mutations in the 3D-structure of proteins. A) Located the pathogenic mutations detected in CHEK2 protein B) Located the pathogenic mutations detected in MUTYH protein. Depicted as PROREPAIR-B (grey), PRORADIUM (red), pathogenic mutations documented in COSMIC (yellow)*

4.5. Exploring copy number alterations in the PRORADIUM cohort for the DDR genes BRCA1, BRCA2, ATM, PALB2 and MSH2

As we commented before in the materials and methods section, this analysis is under development. Currently, we are working on how to integrate results from different computational tools, and also, we have to distinguish between pathogenic CNVs and over-represented regions or polymorphisms. For now, we are implementing 46,533 annotations from ClinvarCnv, based on GRCh37/hg19 assembly.

5. Conclusions and future plan

Conclusions

In the present work we produced and automatized a bioinformatics pipeline to study targeted-sequencing data, with a custom-designed panel of DDR genes, in more than 500 mCRPC Spanish patients. The application of the pipeline to experimental data obtained by the Prostate Cancer Clinical Unit at CNIO-Carlos III, allows us to conclude that:

1) The most recurrent mutated genes with pathogenic gDDR variants in PROREPAIR-B and PRORADIUM cohorts were BRCA2, MUTYH and ATM.
2) The number of pathogenic variants identified in BRCA2 in PROREPAIR-B and PRORADIUM Spanish cohorts is in the same order of magnitude than in other mCRPC published cohorts.
3) The most recurrent mutated gene BRCA2 shows a significantly higher prevalence in Spanish cohorts than in the general non-cancer population (CSVS $P< 0.001$ and ExAC $P < 0.001$).
4) The gDDR pathogenic variants are frequently located outside protein domains but in some cases clustering in hotspots for different tumour types in the protein-coding genes studied.
5) The CNV analysis is time-consuming and use of different methods will be needed for a reliable prediction.

Part of the results obtained in this work have been published in Castro et al., 2019 and presented in the 42nd ESMO Congress, September 8-12, 2017, Madrid.

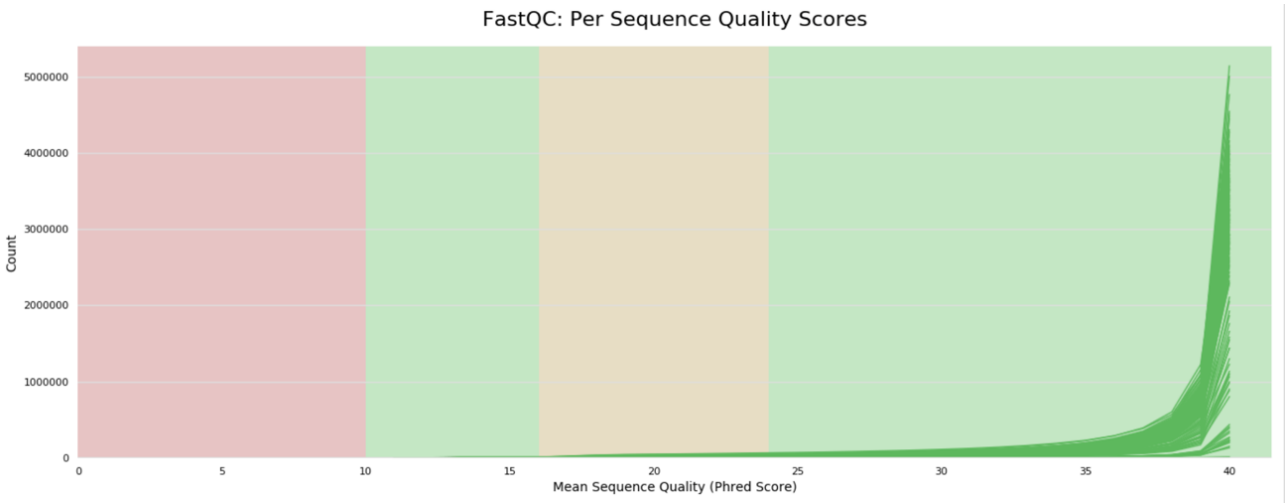
Future plan

The work presented for the Master's Degree in Bioinformatics and Computational Biology is part of my PhD research project, which investigates the role of DDR genes in mCRPC Spanish cohorts. Once the CNV analysis is complete, we will integrate SNV and CNV analyses in the PRORADIUM cohort. To study the DDR signalling pathways affected by CNV and/or SNV, in the integrative analysis, we also planned to study the distribution of the pathogenic variants, not studied here, identified in the PROREPAIR-B and PRORADIUM cohorts through the 3D-structure analysis of modelled proteins and protein-protein complexes.

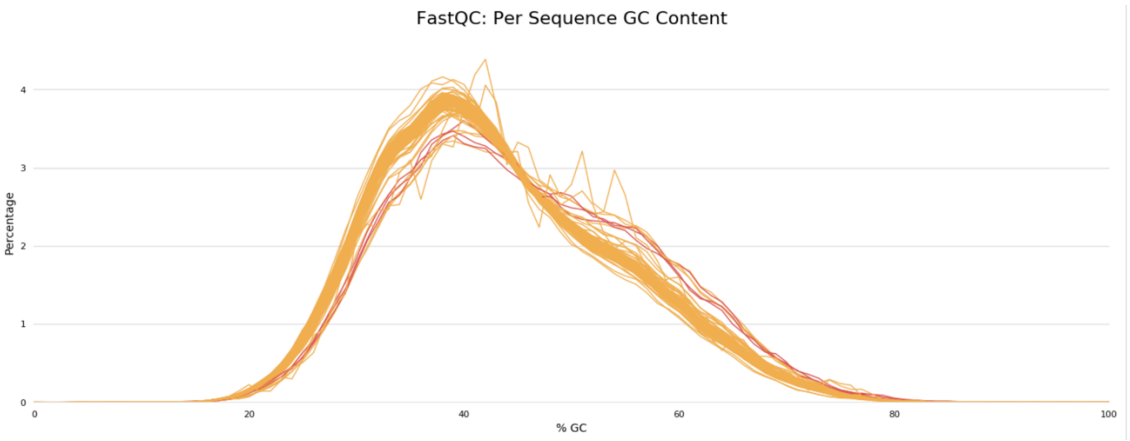
Supplementary

S1. The general quality of the base

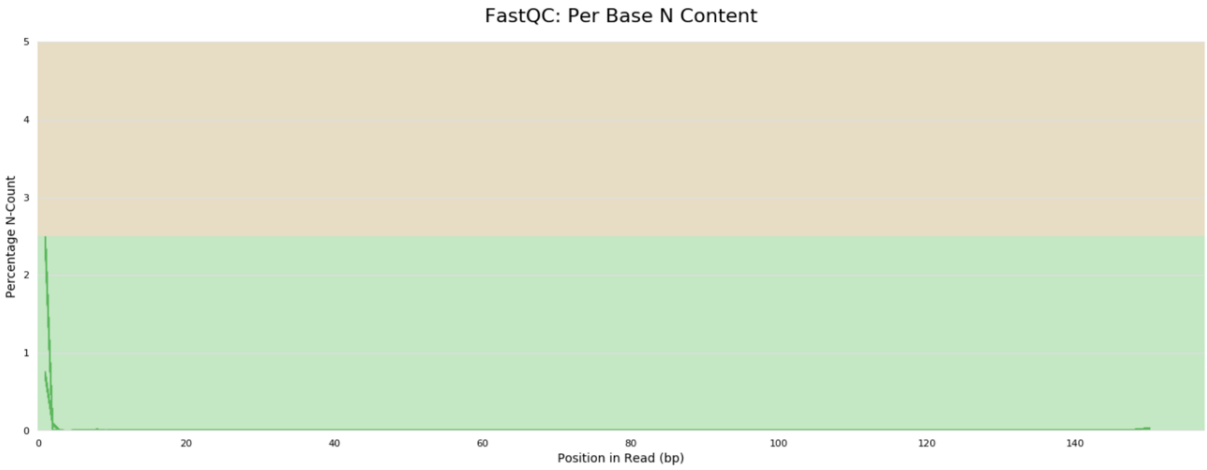
S2. The general quality of the reads



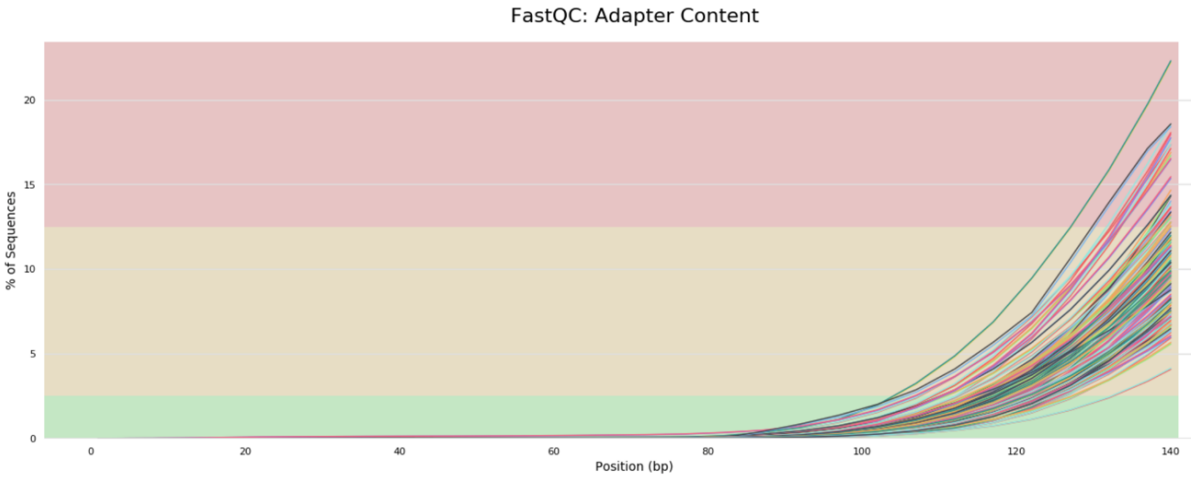
S3. The CG content



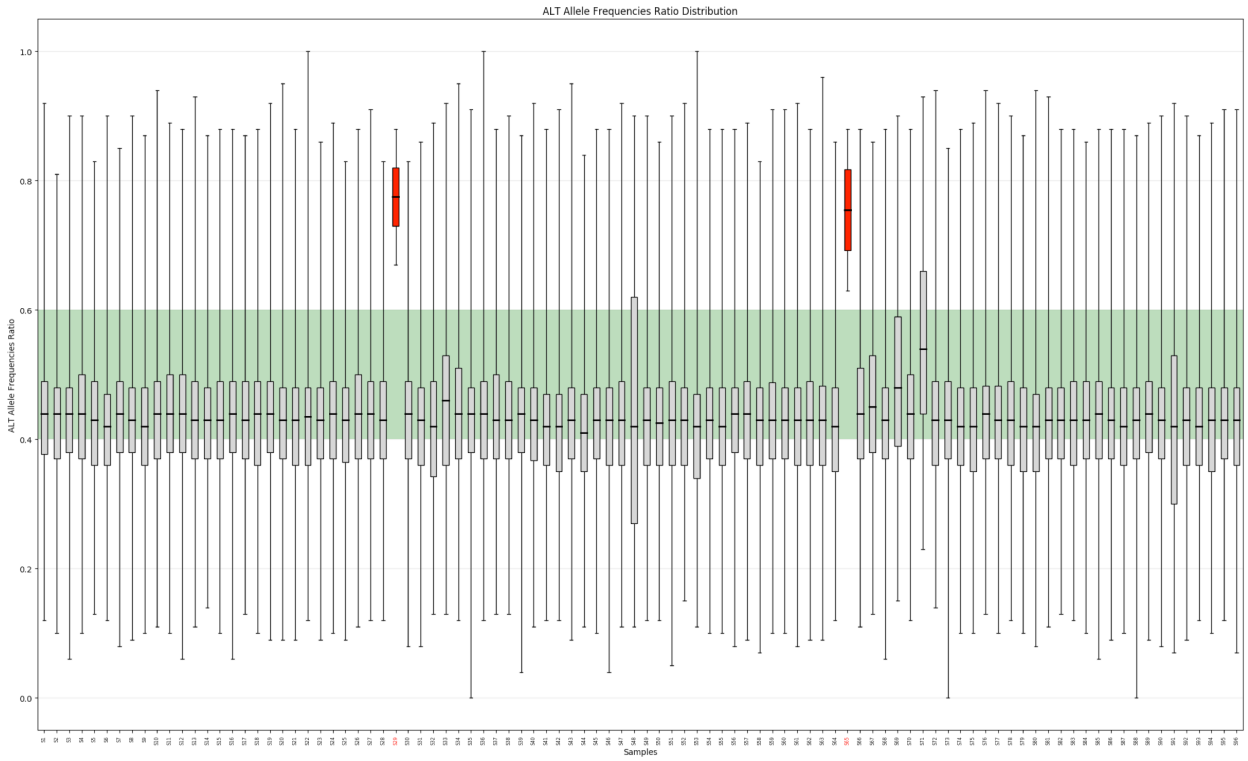
S4. Number of bases not specified (content of "N" in the reads)



S5. The adapter content



S6. Allele Frequencies Ratio



S7. List of Pathogenic/Likely Pathogenic Mutations Identified in the PROREPAIR Study

ID	Gene	Nucleotide Change	Amino Acid Change	Effect
B20018	ATM	c.3381_3384del	p.A1127fs	frameshift deletion
B07020	ATM	c.1442T>G	p.L481*	stopgain
B01019	ATM	c.1262C>A	p.S421*	stopgain
B58005	ATM	c.1262C>A	p.S421*	stopgain
B13040	ATM	c.1262C>A	p.S421*	stopgain
B14001	ATM	c.1262C>A	p.S421*	stopgain
B07003	ATM	c.1336C>T	p.Q446*	stopgain
B16005	ATM	c.6289G>T	p.E2097*	stopgain
B07013	BRCA1	c.2017_2023del	p.673fs	frameshift deletion
B47007	BRCA1	c.470_471del	p.S157fs	frameshift deletion
B13033	BRCA1	c.3756_3759del	p.L1252fs	frameshift deletion
B21052	BRCA1	c.A211G	p.R71G	nonsynonymous SNV
B14014	BRCA2	c.6911T>G	p.L2304*	stopgain
B48011	BRCA2	c.3264dupT	p.P1088fs	frameshift insertion
B41010	BRCA2	c.3648dupT	p.G1215fs	frameshift insertion
B46021	BRCA2	c.3059_3060del	p.L1019fs	frameshift deletion
B46013	BRCA2	c.9018C>A	p.Y3006*	stopgain

B42003	BRCA2	c.4963delT	p.C1654fs	frameshift deletion
B41035	BRCA2	c.2701delC	p.L901fs	frameshift deletion
B01009	BRCA2	c.5116_5119del	p.R1704fs	frameshift deletion
B21080	BRCA2	c.9018C>A	p.Y3006*	stopgain
B21066	BRCA2	c.1326_1329del	p.S442fs	frameshift deletion
B08001	BRCA2	c.4889c>G	p.Ser1630*	stopgain
B13029	BRCA2	c.6650_6654del	p.K2217fs	frameshift deletion
B13024	BRCA2	c.5720_5723del	p.S1907fs	frameshift deletion
B01006	BRCA2	c.9026_9030del	p.Y3009fs	frameshift deletion
B48002	CDK7	c.874_875dup	p.K291fs	frameshift insertion
B01008	CHEK2	c.349A>G	p.R117G	nonsynonymous SNV
B46005	CHEK2	c.591delA	p.K197fs	frameshift deletion
B13019	DCLRE1C	c.241C>T	p.R81*	stopgain
B26005	EME2	c.886C>T	p.Q296*	stopgain
B41031	EME2	c.949_953del	p.F315fs	frameshift deletion
B07021	ERCC2	c.1847G>C	p.R616P	nonsynonymous SNV
B21009	ERCC3	c.335dupA	p.H112fs	frameshift insertion
B21060	ERCC6	c.3474_3477del	p.E1158fs	frameshift deletion
B21091	FANCD2	c.3457G>T	p.E1153*	stopgain
B46023	FANCD2	c.3457G>T	p.E1153*	stopgain
B21009	FANCD2	c.3496delG	p.R1165fs	frameshift deletion
B09004	FANCE	c.1413_1414dup	p.V471fs	frameshift insertion
B21023	FANCE	c.929dupC	p.A308fs	frameshift insertion
B13021	FANCF	c.1087C>T	p.Q363*	stopgain
B02002	FANCG	c.907_908dup	p.L303fs	frameshift insertion
B26005	FANCL	c.1051_1052del	p.S351fs	frameshift deletion
B16001	FANCL	c.1051_1052del	p.S351fs	frameshift deletion
B46017	FANCM	c.1196C>G	p.S399*	stopgain
B36002	GTF2H4	c.358_374del	p.I120fs	frameshift deletion
B46019	GTF2H4	c.940C>T	p.R314*	stopgain
B41005	MMS19	c.607delA	p.R203fs	frameshift deletion
B33002	MRE11A	c.1516G>T	p.E506*	stopgain
B07011	MRE11A	c.1897C>T	p.R633*	stopgain
B21034	MSH2	c.1808A>G	p.D603G	nonsynonymous SNV
B36001	MUTYH	c.1178G>A	p.G393D	nonsynonymous SNV
B29001	MUTYH	c.1138delC	p.L380fs	frameshift deletion
B41027	MUTYH	c.1178G>A	p.G393D	nonsynonymous SNV
B02002	MUTYH	c.527A>G	p.Y176C	nonsynonymous SNV
B13013	MUTYH	c.1178G>A	p.G393D	nonsynonymous SNV
B09003	MUTYH	c.1178G>A	p.G393D	nonsynonymous SNV
B13019	MUTYH	c.1178G>A	p.G393D	nonsynonymous SNV
B14020	MUTYH	c.1178G>A	p.G393D	nonsynonymous SNV

B21001	MUTYH	c.527A>G	p.Y176C	nonsynonymous SNV
B21059	MUTYH	c.527A>G	p.Y176C	nonsynonymous SNV
B21045	MUTYH	c.1178G>A	p.G393D	nonsynonymous SNV
B21086	MUTYH	c.1178G>A	p.G393D	nonsynonymous SNV
B01035	MUTYH	c.1178G>A	p.G393D	nonsynonymous SNV
B20005	PARP3	c.466C>T	p.Q156*	stopgain
B01012	PMS1	c.1690dup	p.Y563fs	frameshift insertion
B06006	RAD17	c.676G>T	p.E226*	stopgain
B16001	RAD17	c.676G>T	p.E226*	stopgain
B47006	RAD54L	c.1099delG	p.A367fs	frameshift deletion
B06008	RPA2	c.276C>A	p.Y92*	stopgain
B14018	XPA	c.682C>T	p.R228*	stopgain
B13020	XPC	c.1643_1644del	p.V548fs	frameshift deletion

S8. List of Pathogenic/Likely Pathogenic Mutations Identified in the PRORADIUM Study

ID	Gene	Nucleotide Change	Amino Acid Change	Effect
5026	ATM	c.802C>T	p.Gln268*	stopgain
1210	ATM	c.3890_3891insT	p.Ala1299fs	frameshift insertion
IT002	ATM	c.8317_8318insCTGTC	p.Pro2775fs	frameshift insertion
5082	BRCA1	c.815_824dupAGCCATGTGG	p.Thr276fs	frameshift insertion
5069	BRCA2	c.9235delG	p.Val3079fs	frameshift deletion
5062	BRCA2	c.9025_9029delTATCA	p.Tyr3009fs	frameshift deletion
5079	BRCA2	c.3455T>G	p.Leu1152*	stopgain
5051	BRCA2	c.2806_2809delAAAC	p.Ala938fs	frameshift deletion
5082	CHEK2	c.538C>T	p.Arg180Cys	nonsynonymous SNV
5068	MUTYH	c.1106G>A	p.Gly393Asp	nonsynonymous SNV
5060	MUTYH	c.1106G>A	p.Gly393Asp	nonsynonymous SNV
4019	MUTYH	c.1106G>A	p.Gly393Asp	nonsynonymous SNV
IT018	MUTYH	c.1106G>A	p.Gly393Asp	nonsynonymous SNV
4065	XPC	c.1643_1644delTG	p.Val548fs	frameshift deletion

Bibliography

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Børresen-Dale AL, Boyault S, Burkhardt B, Butler AP, Caldas C, Davies HR, Desmedt C, Eils R, Eyfjörd JE, Foekens JA, Greaves M, Hosoda F, Hutter B, Ilicic T, Imbeaud S, Imielinski M, Jäger N, Jones DT, Jones D, Knappskog S, Kool M, Lakhani SR, López-Otín C, Martin S, Munshi NC, Nakamura H, Northcott PA, Pajic M, Papaemmanuil E, Paradiso A, Pearson JV, Puente XS, Raine K, Ramakrishna M, Richardson AL, Richter J, Rosenstiel P, Schlesner M, Schumacher TN, Span PN, Teague JW, Totoki Y, Tutt AN, Valdés-Mas R, van Buuren MM, van 't Veer L, Vincent-Salomon A, Waddell N, Yates LR; Australian Pancreatic Cancer Genome Initiative; ICGC Breast Cancer Consortium; ICGC MMML-Seq Consortium; ICGC PedBrain, Zucman-Rossi J, Futreal PA, McDermott U, Lichter P, Meyerson M, Grimmond SM, Siebert R, Campo E, Shibata T, Pfister SM, Campbell PJ, Stratton MR. (2013) Signatures of mutational processes in human cancer. Nature 500(7463):415-21.

Altman DG. Practical statistics for medical research. London: Chapman and Hall, 1991

Annala M, Vandekerkhove G, Khalaf D, Taavitsainen S, Beja K, Warner E, Sunderland K, Kollmannsberger C, Eigl B, Finch D, Oja C, Vergidis J, Zulfiqar M, Azad A, Nykter M, Gleave M, Wyatt A, Chi K. Circulating tumor DNA genomics correlate with resistance to abiraterone and enzalutamide in prostate cancer. Cancer Discov 2018; 8: 444–57

Antonarakis ES, Lu C, Luber B, Liang C, Wang H, Chen Y, Silberstein JL, Piana D, Lai Z, Chen Y, Isaacs WB, Lou J. Germline DNA-repair gene mutations and outcomes in men with metastatic castration-resistant prostate cancer receiving first-line abiraterone and enzalutamide. Eur Urol 2018; 74: 218–25

Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, Colaprico A, Wendl MC, Kim J, Reardon B, Ng PK, Jeong KJ, Cao S, Wang Z, Gao J, Gao Q, Wang F, Liu EM, Mularoni L, Rubio-Perez C, Nagarajan N, Cortés-Ciriano I, Zhou DC, Liang WW, Hess JM, Yellapantula VD, Tamborero D, Gonzalez-Perez A, Suphavilai C, Ko JY, Khurana E, Park PJ, Van Allen EM, Liang H; MC3 Working Group; Cancer Genome Atlas Research Network, Lawrence MS, Godzik A, Lopez-Bigas N, Stuart J, Wheeler D, Getz G, Chen K, Lazar AJ, Mills GB, Karchin R, Ding L. (2018) Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell 173(2):371-385.e18.

Barata PC, Mendiratta P, Heald B, Klek S, Grivas P, Sohal DPS, Garcia JA.Targeted

Next-Generation Sequencing in Men with Metastatic Prostate Cancer: a Pilot Study. Target Oncol. 2018 Aug;13(4):495-500

Bohannan ZS, Mitrofanova A. Calling Variants in the Clinic: Informed Variant Calling Decisions Based on Biological, Clinical, and Laboratory Variables. Comput Struct Biotechnol J. 2019 Apr 8; 17:561-569

Brookes AJ. The essence of SNPs. Gene. 1999 Jul 8;234(2):177-86.

Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018; 68(6):394-424.

Castro E, Goh C, Olmos D, Saunders E, Leongamornlert D, Tymrakiewicz M, Mahmud N, Dadaev T, Govindasami K, Guy M, Sawyer E, Wilkinson R, Ardern-Jones A, Ellis S, Frost D, Peock S, Evans DG, Tischkowitz M, Cole T, Davidson R, Eccles D, Brewer C, Douglas F, Porteous ME, Donaldson A, Dorkins H, Izatt L, Cook J, Hodgson S, Kennedy MJ, Side LE, Eason J, Murray A, Antoniou AC, Easton DF, Kote-Jarai Z, Eeles R. Germline BRCA mutations are associated with higher risk of nodal involvement, distant metastasis, and poor survival outcomes in prostate cancer. J Clin Oncol. 2013 May 10;31(14):1748-57.

Castro E, Goh C, Leongamornlert D, Saunders E, Tymrakiewicz M, Dadaev T, Govindasami K, Guy M, Ellis S, Frost D, Bancroft E, Cole T, Tischkowitz M, Kennedy MJ, Eason J, Brewer C, Evans DG, Davidson R, Eccles D, Porteous ME, Douglas F, Adlard J, Donaldson A, Antoniou AC, Kote-Jarai Z, Easton DF, Olmos D, Eeles R. Effect of BRCA Mutations on Metastatic Relapse and Cause-specific Survival After Radical Treatment for Localised Prostate Cancer. Eur Urol. 2015 Aug;68(2):186-93

Castro E, Romero-Laorden N, Del Pozo A, Lozano R, Medina A, Puente J, Piulats JM, Lorente D, Saez MI, Morales-Barrera R, Gonzalez-Billalabeitia E, Cendón Y, García-Carbonero I, Borrega P, Mendez Vidal MJ, Montesa A, Nombela P, Fernández-Parra E, Gonzalez Del Alba A, Villa-Guzmán JC, Ibáñez K, Rodriguez-Vida A, Magraner-Pardo L, Perez-Valderrama B, Vallespín E, Gallardo E, Vazquez S, Pritchard CC, Lapunzina P, Olmos D. PROREPAIR-B: A Prospective Cohort Study of the Impact of Germline DNA Repair Mutations on the Outcomes of Patients With Metastatic Castration-Resistant Prostate Cancer. J Clin Oncol. 2019 Feb 20;37(6):490-503.

Chamberlain J, Melia J, Moss S, et al: The diagnosis, management, treatment and costs of prostate cancer in England and Wales. Health Technol Assess 1:i-vi, 1-53, 1997

Cheng HH, Pritchard CC, Boyd T, Nelson PS, Montgomery B. Biallelic inactivation of BRCA2 in platinum-sensitive metastatic castration-resistant prostate cancer. Eur Urol. 2016;69:992-995.

Conteduca V, Sigouros M, Sboner A, Pritchard CC, Beltran H. BRCA2-Associated Prostate Cancer in a Patient With Spinal and Bulbar Muscular Atrophy. JCO Precis Oncol. 2018;2.

Ding L, Bailey MH, Porta-Pardo E, Thorsson V, Colaprico A, Bertrand D, Gibbs DL, Weerasinghe A, Huang KL, Tokheim C, Cortés-Ciriano I, Jayasinghe R, Chen F, Yu L, Sun S, Olsen C, Kim J, Taylor AM, Cherniack AD, Akbani R, Suphavilai C, Nagarajan N, Stuart JM, Mills GB, Wyczalkowski MA, Vincent BG, Hutter CM, Zenklusen JC, Hoadley KA, Wendl MC, Shmulevich L, Lazar AJ, Wheeler DA, Getz G; Cancer Genome Atlas Research Network. (2018) Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. Cell 173(2):305-320.e10.

Graham J, Baker M, Macbeth F, et al: Diagnosis and treatment of prostate cancer: summary of NICE guidance. BMJ 336:610-2, 2008

Han M, Partin AW, Piantadosi S, et al: Era specific biochemical recurrence-free survival following radical prostatectomy for clinically localized prostate cancer. J Urol 166:416-9, 2001

Hovelson DH, Tomlins SA. The Role of Next-Generation Sequencing in Castration-Resistant Prostate Cancer Treatment. Cancer J. 2016 Sep/Oct;22(5):357-361.

Huang KL, Mashl RJ, Wu Y, Ritter DI, Wang J, Oh C, Paczkowska M, Reynolds S, Wyczalkowski MA, Oak N, Scott AD, Krassowski M, Cherniack AD, Houlahan KE, Jayasinghe R, Wang LB, Zhou DC, Liu D, Cao S, Kim YW, Koire A, McMichael JF, Hucthagowder V, Kim TB, Hahn A, Wang C, McLellan MD, Al-Mulla F, Johnson KJ; Cancer Genome Atlas Research Network, Lichtarge O, Boutros PC, Raphael B, Lazar AJ, Zhang W, Wendl MC, Govindan R, Jain S, Wheeler D, Kulkarni S, Dipersio JF, Reimand J, Meric-Bernstam F, Chen K, Shmulevich I, Plon SE, Chen F, Ding L. (2018) Pathogenic Germline Variantsin 10,389 Adult Cancers. Cell 173(2):355-370.e14

Ikeda S, Elkin SK, Tomson BN, Carter JL, Kurzrock R. Next-generation sequencing of prostate cancer: genomic and pathway alterations, potential actionability patterns, and relative rate of use of clinical-grade testing. Cancer Biol Ther. 2018 Oct 19:1-8.

Karzai F, VanderWeele D, Madan RA, Owens H, Cordes LM, Hankin A, Couvillon A, Nichols E, Bilusic M, Beshiri ML, Kelly K, Krishnasamy V, Lee S, Lee MJ, Yuno A, Trepel JB, Merino MJ, Dittamore R, Marté J, Donahue RN, Schlom J, Killian KJ, Meltzer PS, Steinberg SM, Gulley JL, Lee JM, Dahut WL. Activity of durvalumab plus olaparib in metastatic castration-resistant prostate cancer in men with and without DNA damage repair mutations. J Immunother Cancer. 2018 Dec 4;6(1):141

Kaufman B, Shapira-Frommer R, Schmutzler RK, et al. Olaparib monotherapy in patients with advanced cancer and a germline BRCA1/2 mutation. J Clin Oncol. 2015;33:244-250.

Kote-Jarai Z, Mikropoulos C, Leongamornlert DA, Dadaev T, Tymrakiewicz M, Saunders EJ, Jones M, Jugurnauth-Little S, Govindasami K, Guy M, Hamdy FC, Donovan JL, Neal DE, Lane JA, Dearnaley D, Wilkinson RA, Sawyer EJ, Morgan A, Antoniou AC, Eeles RA; UK Genetic Prostate Cancer Study Collaborators, and ProtecT Study Group. Prevalence of the HOXB13 G84E

germline mutation in British men and correlation with prostate cancer risk, tumour characteristics and clinical outcomes. Ann Oncol. 2015 Apr;26(4):756-61

Kuban D, Thames H, Levy L, et al: Failure definition-dependent differences in outcome following radiation for localized prostate cancer: can one size fit all? Int J Radiat Oncol Biol Phys 61:409-14, 2005

Mark J, Kucherov V, Pintauro M, Leong JY, Mann M, Trabulsi EJ, Lallas C, Chandrasekar T, Handley N, Kelly WK, Gomella LG. (2019) What is the clinical utility of next generation sequencing (NGS) in advanced urologic malignancies? Journal of Clinical Oncology 37, suppl (March 1):285-285.

Mateo J, Carreira S, Sandhu S, et al. DNA-repair defects and ola- parib in metastatic prostate cancer. N Engl J Med. 2015;373:1697- 1708.

Mijuskovic M, Saunders EH, Leongamornlert AD. Rare germline variants in DNA repair genes and the angiogenesis pathway predispose prostate cancer patients to develop metastatic disease. Br J Cancer 2018; 119: 96–104

Mottet N, De Santis M, Briers E, Bourke L, Gillessen S, Grummet JP, Lam TB, van der Poel HG, Rouvière O, van den Bergh RCN, Cornford P. Updated Guidelines for Metastatic Hormone-sensitive Prostate Cancer: Abiraterone Acetate Combined with Castration Is Another Standard. Eur Urol. 2017. pii: S0302-2838(17)30839-4.

Mucci LA, Hjelmborg JB, Harris JR, Czene K, Havelick DJ, Scheike T, Graff RE, Holst K, Möller S, Unger RH, McIntosh C, Nuttall E, Brandt I, Penney KL, Hartman M, Kraft P, Parmigiani G, Christensen K, Koskenvuo M, Holm NV, Heikkilä K, Pukkala E, Skytthe A, Adami HO, Kaprio J; Nordic Twin Study of Cancer (NorTwinCan) Collaboration. Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. JAMA. 2016 Jan 5;315(1):68-76.

Na R, Zheng SL, Han M,…, Brendler CB, Ding Q, Xu J, Isaas WB. Germline Mutations in ATM and BRCA1/ 2 Distinguish Risk for Lethal and Indolent Prostate Cancer and are Associated with Early Age at Death. Eur Urol 2017; 71: 740–7

Parker C, Gillessen S, Heidenreich A, Horwich A; ESMO Guidelines Committee. Cancer of the prostate: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2015;26 Suppl 5:v69-77

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. N Engl J Med. 2016 Aug 4;375(5):443-53

Richards S, Aziz N, Bale S, et al: Standards and guidelines for the interpretation of

sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 17:405-424, 2015

Robinson D, Van Allen EM, Wu YM, Schultz N, Lonigro RJ, Mosquera JM, Montgomery B, Taplin ME, Pritchard CC, Attard G, Beltran H, Abida W, Bradley RK, Vinson J, Cao X, Vats P, Kunju LP,… Sawyers CL, Chinnaiyan AM. Integrative Clinical Genomics of Advanced Prostate Cancer. Cell. 2015; 16:162(2):454

Surveillance E; End Results Program (SEER) Surveillance, Epidemiology, and Ends Results Program. Fast Stats. 2015. Available at: http://seer.cancer.gov/faststats/selections.php.