Informatics in Medicine Unlocked 43 (2023) 101410

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Unveiling genetic variant-level biomarkers for aggressive prostate cancer

Tasnimul Alam Taz [*,1], Suzan Arslanturk [1]

Department of Computer Science, Wayne State University, Detroit, MI, 48202, USA

ARTICLE INFO

Keywords:
Biomarker
Multiple comparison
Over-representation analysis
Decision tree

ABSTRACT

Prostate cancer (PCa) represents the second most frequently diagnosed malignancy among males in the United States and ranks fourth in terms of general cancer prevalence on a global scale. A critical assessment of existing literature indicates a notable deficiency in the identification of biomarkers that are uniquely associated with aggressive forms of PCa. The principal objective of this paper is to discover biomarkers at the genetic variant level by deploying statistical methodologies to determine associations between such variants and the aggressive and lethal form of the disease. Employing the multiple comparisons technique, we identified four variants that were statistically significant at the 5 % significance level. Furthermore, we utilized Over-representation analysis (ORA) to identify the biological pathways linked with these genetic variants. To validate our findings, we employed a decision tree algorithm on an independent dataset comparing the proposed biomarkers with random subsets of variants. Results have shown that the predictive accuracy of aggressive samples was 97 % for the proposed biomarkers, while this figure dropped to 67 % when randomly selected variants were considered.

1. Introduction

Prostate cancer (PCa) is the most prevalent malignancy in males and a prominent cause of cancer-related mortality. As of 2022, 1,414,259 new cases and 375,304 deaths have been reported from PCa worldwide [1]. In the United States, PCa is the primary cause of cancer incidence and the second highest cause of cancer death in males. Recent data indicate a 3 % annual increase in PCa incidence from 2014 to 2019. PCa is frequently non-aggressive, and treatment is often curative. Due to the harmful effects of over- and under-treatment, PCa is the primary cause of cancer-associated disability worldwide. Therefore, different types of PCa require distinct treatment alternatives. The field of aggressive prostate cancer treatment is rapidly evolving. Although the 5-year survival rate for indolent PCa is 99 %, aggressive PCa is typically considered incurable. This further underscores the crucial importance of early treatment for aggressive cases [2,3].

Biomarkers serve as essential indicators for the early detection of cancer. Through the identification of specific biomarkers associated with a particular disease, healthcare professionals can screen individuals who may be at risk or in the early stages of the disease. This enables early intervention and treatment, significantly improving patient survival [4]. Genetic variants are among the various categories of biomarkers that play a crucial role in different aspects of healthcare, including disease diagnosis, prognosis, treatment selection, and monitoring treatment response. Recent technological advancements in genome sequencing, particularly whole-genome sequencing (WGS), have provided valuable resources for comprehending cancer at the molecular level. These advancements have allowed for a focused investigation of genetic variants that contribute to the development and progression of pathogenic cancers [5].

A recent large-scale genetic study identified nine novel PCa risk variants (rs73923570, rs60985508, rs72960383, rs144842076, rs13172201, rs114053368, rs9895704, rs73991216, and rs150947563) contributing to our improved understanding of the disease. Furthermore, a comprehensive multiancestry polygenic risk score analysis was conducted, revealing these variants as potential biomarkers for aggressive PCa. Importantly, this analysis effectively distinguished between the risks associated with aggressive and non-aggressive forms of the disease. However, it remains to be determined whether these variants exert any influence on the expression or functionality of genes specifically associated with aggressive PCa cases [6]. Despite the utilization of factors such as Gleason score and tumor stage for prognosis determination, current treatment approaches do not differ for aggressive PCa patients. Consequently, a recent genome-wide association study aimed to explore genetic variants that may be associated with an increased risk of more aggressive PCa [7]. Within this investigation, a particular

[*] Corresponding author.
E-mail address: hj4298@wayne.edu (T.A. Taz).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.imu.2023.101410
Received 1 September 2023; Received in revised form 30 October 2023; Accepted 13 November 2023
Available online 19 November 2023

variant located on 15q13, denoted as rs6497287, exhibited a robust association with a higher risk of developing aggressive disease (p-value = 0.004) compared to less aggressive forms (p-value = 0.14). However, the association was not stronger for more aggressive disease. This finding may be attributed to the limitation of a small sample size.

In this paper, we initially gathered metastatic PCa data from the publicly accessible database cBioPortal. From the dataset, we specifically extracted genetic variant information, among various other data points. The genetic variant data were provided in the mutation annotation format (MAF), which primarily encompasses somatic mutations. It is worth noting that the variant call format (VCF) is more commonly utilized for the storage and exchange of genetic variant information. For this reason, we converted the dataset from MAF format to the VCF format. In order to identify aggressive and non-aggressive groups from the dataset, we filtered the 444 metastatic PCa samples according to their survival status. After filtering we found 19 patients whom we defined as aggressive PCa patients and 46 patients whom we termed as non-aggressive patients. Statistical analysis was then conducted to determine the probability of observing genetic variants exclusively in aggressive patients but not in non-aggressive patients. Based on the statistical analysis, we identified three significant variants: rs777215086 (adjusted p-value = 0.0012), rs5759167 (adjusted p-value = 0.0045), and rs864309495 (adjusted p-value = 0.0072) in aggressive PCa patients. Additionally, we discovered a novel variant (C/A) located on chromosome 15 at position 50904997, which also demonstrated statistical significance (adjusted p-value = 0.0034). Subsequently, we examined the impact of these variants on genes and biological pathways using over-representation analysis (ORA). Literature review along with Machine learning (ML) study confirmed that our findings are in accordance with the prognosis of aggressive PCa.

2. Dataset information

The dataset utilized in this study was obtained from cBioPortal (https://www.cbioportal.org/), a publicly available resource that provides access to comprehensive cancer genomic datasets. This source includes data from prominent consortium initiatives such as Therapeutically Applicable Research to Generate Effective Treatments (TARGET), along with individual laboratory publications. The dataset comprised 429 patients with metastatic castrate-resistant prostate cancer (mCRPC), encompassing 444 tumor/normal whole exome sequencing pairs (Name of the dataset: “Metastatic Prostate Adenocarcinoma (SU2C/PCF Dream Team, PNAS 2019”), Link: https://www.cbioportal.org/study/summary?id=prad_su2c_2019) The patients included in the dataset were undergoing a clinical trial involving the PARP inhibitor olaparib and the Aurora kinase A inhibitor alisertib, specifically targeted towards individuals with neuroendocrine features. The dataset contained overall survival information for 128 patients, which we used to define the phenotypic groups for our study. Based on discussions with oncologists and subject matter experts, patients who died within the first year of diagnosis were categorized as the aggressive group (19 patients), while those who survived for more than two years were classified as the non-aggressive group (46 patients). For the purpose of this study, which was to focus specifically on two phenotypes, we excluded the remaining patients. We then extracted the genetic variant information pertaining to a total of 65 patients. A comprehensive summary of these genetic variants is depicted in Fig. 1. It can be observed from the figure that the majority of the identified variants fall into the category of missense mutations. Furthermore, SNP (Single Nucleotide Polymorphism) is the predominant variant type in this dataset. Specifically, there is a higher prevalence of C-T SNPs compared to other types. For improved clarity, we have provided a detailed breakdown of the abbreviations used: T > G: Thymine (T) replaced by Guanine (G), T > A: Thymine (T) replaced by Adenine (A), T > C: Thymine (T) replaced by Cytosine (C), C > T: Cytosine (C) replaced by Thymine (T), C > G: Cytosine (C) replaced by Guanine (G), C > A: Cytosine (C) replaced by Adenine (A).

3. Results

3.1. Effect of the filtered genetic variants on genes

Given the limited sample size of this study, we have conducted an individualized examination of the genetic variants. Specifically, our analysis focused on the 19 samples with aggressive disease progression. We have specifically examined the impact of variants on genes that are found only in the aggressive samples. To accomplish this, we employed iVariantGuide (AdvaitaBio), which integrates SnpEff [8], enabling us to estimate the impact of each variant on the transcript. Fig. 2 presents the ratio of highly impacted genes for the 19 aggressive patients compared with non-aggressive subset of patients. Results show that the aggressive samples have overall higher percentages of impacted genes when compared with the non-aggressive group.

Note that, not all variants will have an equal contribution to the genes. The effects of genetic variants are divided into four categories by SnpEff: high, moderate, low, and modifier. Fig. 3 depicts top three patients from each group who exhibit the highest percentage of genes highly impacted by the variants. The variants present in the three aggressive-case patients exert a more significant impact on genes

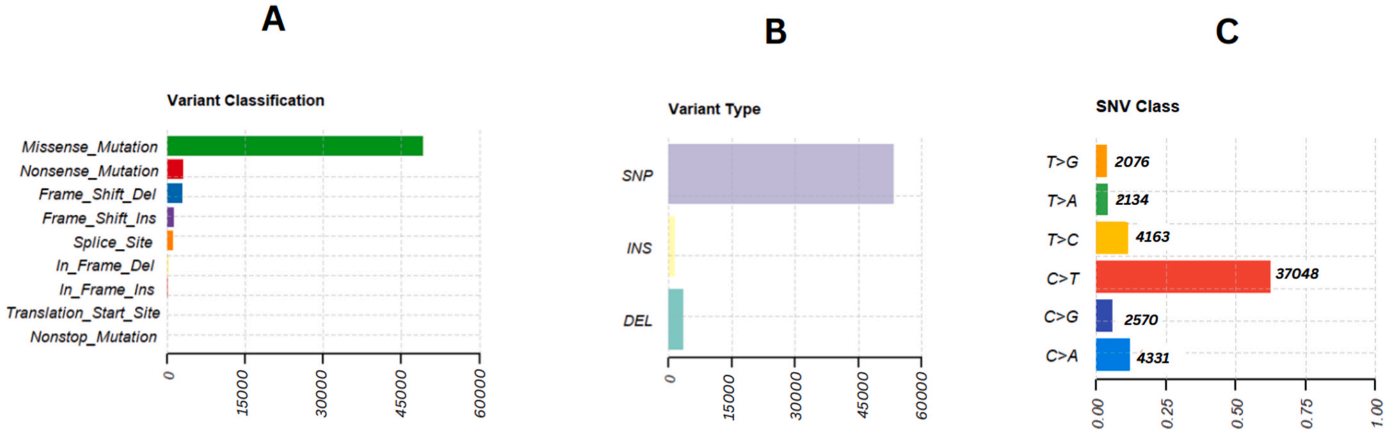


Fig. 1. Summary of the genetic variants observed in our dataset. (A) Distribution of variant classifications is presented, with the x-axis representing the number of variants and the y-axis representing the categories of variant types. (B) The distribution of nucleotide substitutions, commonly known as SNPs, and Indels (insertions and deletions). (C) The SNV class plot illustrates the distribution of variants based on their Minor Allele Frequency (MAF) values. The x-axis represents specific MAF value ranges, while the y-axis displays the count or frequency of variants falling within each MAF range.

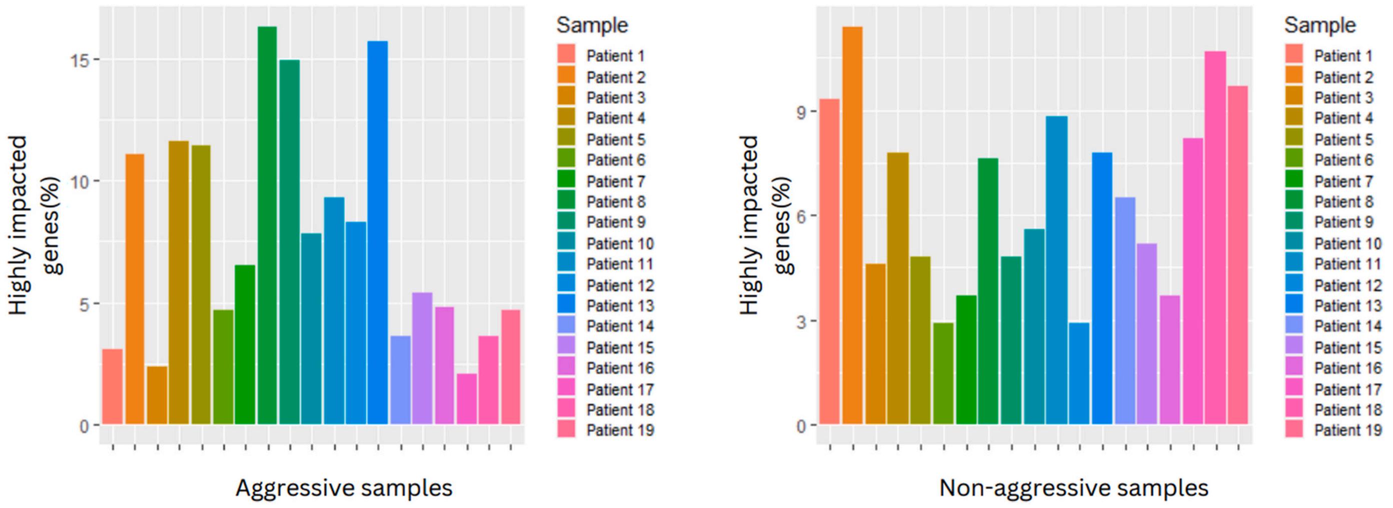


Fig. 2. Patient-specific comparison between two groups, in-terms of highly impacted genes those are affected by the group of variants found in aggressive and non-aggressive samples.

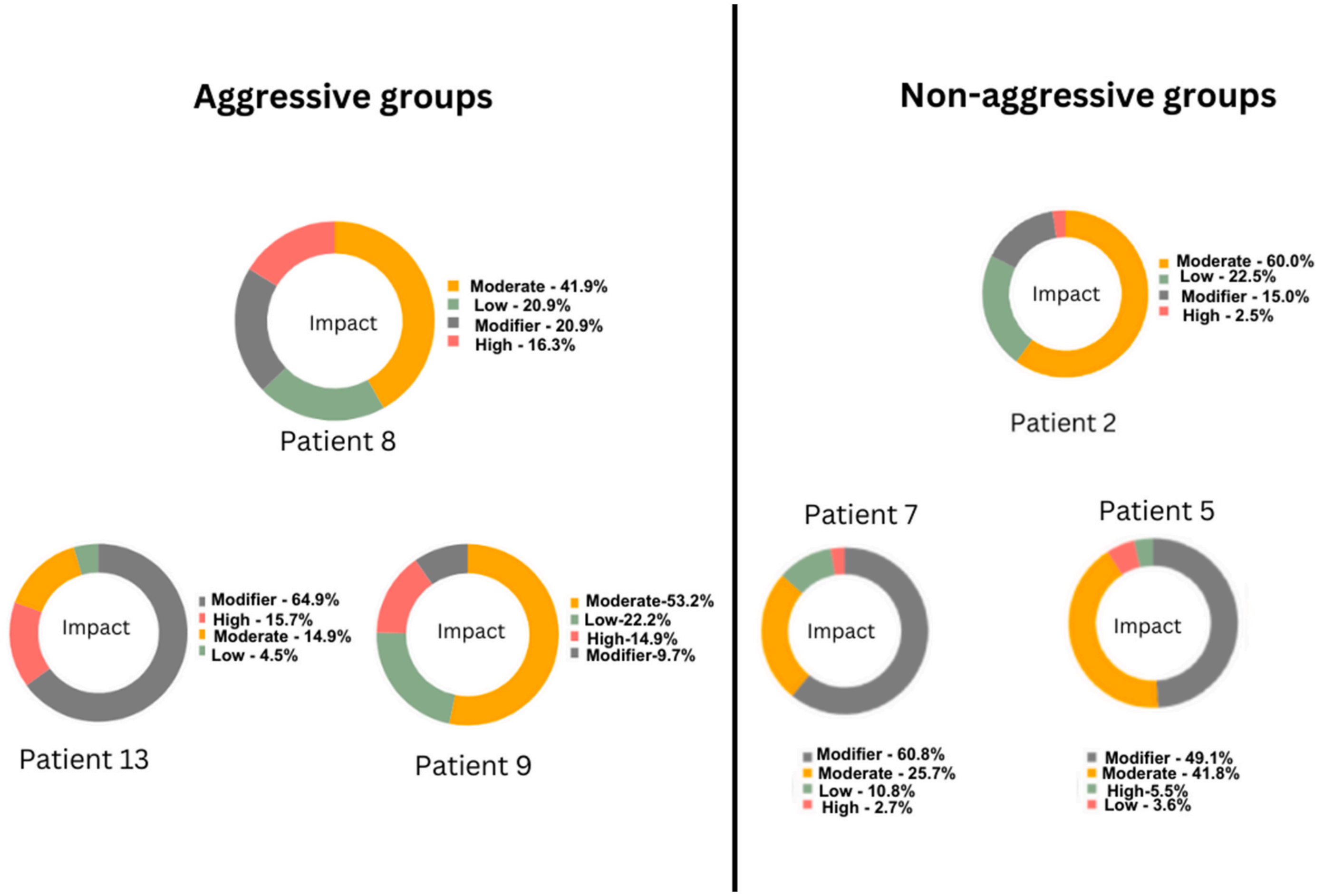


Fig. 3. Representation of three patients from each group exhibiting the highest percentage of highly impacted genes (colored in red) affected by the group of variants. The percentage of genes with varying impact levels is depicted using different colors: red for high impact, yellow for moderate impact, green for low impact, and grey for modifier impact. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

compared to their non-aggressive counterparts. The highly impacted gene percentages for patients 8, 13, and 9 in the aggressive group are 16.3 %, 15.7 %, and 14.9 %, respectively, which is higher than those observed in the indolent patients.

3.2. Statistical analysis results

There are 6645 variants out of a total of 65 samples in the data set (aggressive = 19, non-aggressive = 46). Based on the frequency distribution of these variants, it is evident that all aggressive patients share ten variants. In addition, 18 aggressive patients out of 19 aggressive patients possessed an additional five variants. In contrast, none of these variants were found in non-aggressive patients. The investigation of the variants' genotype information reveals that each variant is a single nucleotide polymorphism (SNP). All pairings contain heterozygous genotypes, denoted by '0/1'. The frequency distribution and genotypic information of the variants that are more prevalent in the aggressive samples are shown in Table 1.

Based on the results of hypothesis testing for multiple comparisons between aggressive and non-aggressive groups, we identified four variants as significant at a 5 % level of significance. Our hypothesis was to determine the probability of a variant that occurs frequently in aggressive cases but not in non-aggressive cases. The overall details of these variants are presented in Table 2.

Primarily, Janus kinase 1 (JAK1) serves as the host gene for the rs777215086 variant, which is identified as a frameshift variant. According to a recent study, recurrent frameshift mutations in JAK1 have been associated with increased mutation load and microsatellite instability in cases of prostate cancer (PCa) [9]. Moreover, a recent investigation revealed that JAK1 is regulated epigenetically in PCa patients [10].

Importantly, a unique variant was identified from the list of variants, on chromosome 15 at position 50904997, and the position represents the genomic coordinate at which the mutation occurred. If a variant is missing an rsID (Reference SNP cluster ID), it indicates that the variant does not have a unique identifier in the dbSNP (Single Nucleotide Polymorphism Database (dbSNP:https://www.ncbi.nlm.nih.gov/snp/), which is a widely used and comprehensive public database of genetic variations. It is possible that a variant without an rsid has not been previously discovered or reported in public databases, rendering it novel. This variant may be uncommon and unique to a specific population or individual, but it has not been extensively studied or included in public databases. The host gene for this unique variant is identified as Transient Receptor Potential Cation Channel Subfamily M Member 7 (TRPM7). Functioning as a Mg2+/Ca2+ permeable channel and a protein kinase, TRPM7 is involved in the regulation of various cellular mechanisms, including cell adhesion, migration, and survival, particularly in the context of metastatic PCa [11,12].

In addition, the variant rs5759167 was considered significant among the list of 6645 variants. A recent review that primarily focuses on genome-wide association studies which was conducted in metastatic PCa patients to identify genetic markers associated with PCa risk. This study identified rs5759167 (p-value = 3.29E-02) as significant among the PCa risk-associated SNPs [13,14]. The final significant variant associated with PCa in the list is rs864309495, which is hosted by the tumor protein 53 gene (TP53). Existing research indicates that structural variants within TP53 are primarily responsible for the aggressive manifestations of PCa [15].

Table 1
Frequency distribution of variants that existed in the aggressive samples and not found in the non-aggressive samples.

Chromosome	Position	Genotype (GT:AD: DP)	Frequency (%)
chr1	65325832	0/1:54,0:54	100
chr3	10146353	0/1:356,5:361	100
chr5	39126099	0/1:24,20:44	100
chr6	24556933	0/1:217,44:261	100
chr9	79002398	0/1:66,46:112	100
chr11	1.14E+08	0/1:268,27:295	100
chr15	50904997	0/1:15,0:15	100
chr17	7578212	0/1:181,0:181	100
chr21	47421171	0/1:249,5:254	100
chr22	43500212	0/1:29,132:161	100
chr7	47463715	0/1:334,5:339	94.74
chr11	33566639	0/1:285,196:481	94.74
chr18	14851528	0/1:36,13:49	94.74
chr18	25589727	0/1:51,39:90	94.74
chr19	50755932	0/1:75,51:126	94.74

Table 2
Variants that are found significant from the statistical analysis.

Chromosome	Position	ID	adj. p-value
chr1	65325832	rs777215086	0.0012
chr15	50904997	Novel variant	0.0034
chr22	43104206	rs5759167	0.0045
chr17	7578212	rs864309495	0.0072

3.3. Pathway analysis results

Originally designed for the analysis of gene expression data, pathway analysis has evolved into a robust analytical method for the comprehensive extraction of genome-wide genetic variants data. Furthermore, it facilitates the interpretation of genetic variants within the context of the biological processes involving the implicated genes and proteins.

In this research, our objective was to address the issue of single-SNP analysis in genetic association studies through the utilization of pathway analysis. Single-SNP analysis in genetic association studies involves investigating the correlation between individual single nucleotide polymorphisms (SNPs) and a specific trait or disease. To minimize false positives, these analyses typically apply rigorous statistical criteria, consequently identifying only those SNPs with highly significant associations as potentially relevant. In contrast, pathway analysis offers a more macroscopic perspective, classifying SNPs into biologically pertinent pathways for a broader and more comprehensive interpretation of the genetic findings across two phenotypes. Fig. 4 illustrates the significant pathways (KEGG) determined by impact analysis for the two categories under consideration. The p-values, as indicated on the x-axis, represent a combination of enrichment and perturbation p-values that have been subsequently adjusted using the false discovery rate (FDR) method. All experiments were performed using the iVariantGuide (AdvaitaBio).

Moreover, two variants from our statistically significant list - rs777215086 and a novel variant (chromosome - 15, position - 50904997, Reference allele - C, Alternate allele - CT) - exert a considerable impact on these two pathways. The rs777215086 variant has a substantial influence on the JAK1 gene, which serves as its host. This variant notably affects the JAK1 gene's interaction with the PI3K-Akt signaling pathway [41–43], as depicted in Fig. 5. Furthermore, several other genes, including PTEN, TSC2, and GYS, also demonstrate a high impact within this pathway. The pathway representation uses nodes to illustrate genes. If a gene within a particular node is affected or "impacted" by a set of variants, that node will be colored to indicate the level of impact. The coloring system is designed to highlight the severity of the impact of the variants on the genes within a node. The color red, in particular, is used to indicate nodes (and, by extension, the genes within them) that are impacted by variants with a high predicted effect. In other words, if there's at least one gene in a node that has a variant with a predicted "high impact" (based on predictions from a tool named SnpEff [8]), that node is colored red. This visual cue allows for a quick and easy identification of nodes with genes that might be of significant concern or interest due to the presence of these high impact variants. We also performed this experiment in iVariantGuide (AdvaitaBio).

3.4. Validation of significant variants

In order to validate the significance of the identified variants in our research, we employed a Decision Tree algorithm on an independent dataset obtained from The Cancer Genome Atlas (TCGA). The primary objective was to assess the algorithm's ability to differentiate between aggressive and non-aggressive cases. This independent dataset comprised 494 patients, among whom 30 were identified as aggressive (deceased within one year) and 67 were identified as non-aggressive

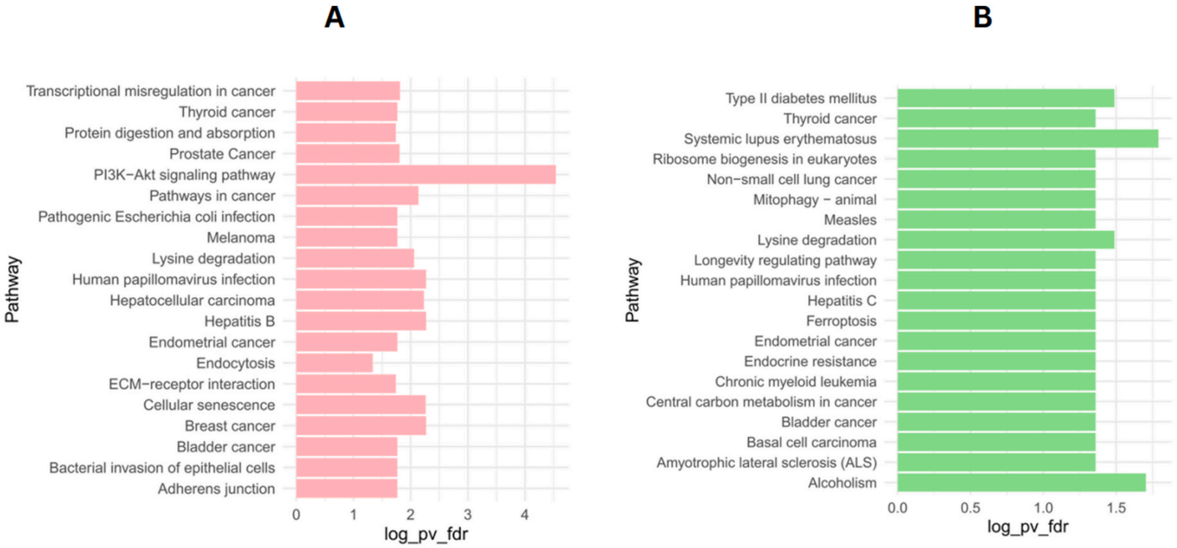


Fig. 4. Comparison of significantly impacted pathways by variants in aggressive cases (A) and non-aggressive samples (B). The X-axis represents the contribution score, while the Y-axis displays the list of pathways.

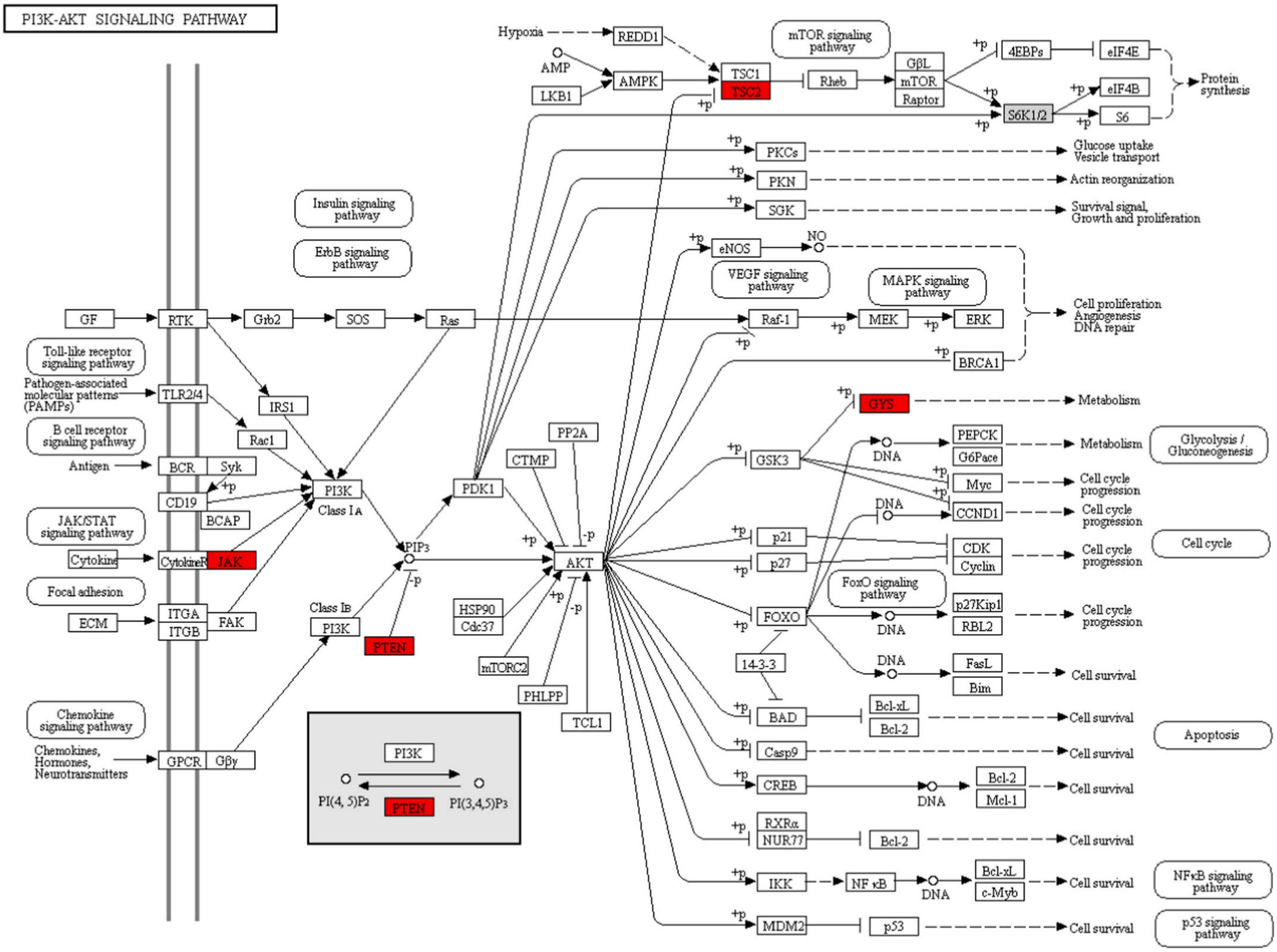


Fig. 5. Representation of the PI3K-Akt signaling pathway. Red nodes indicate genes within the pathway that are highly impacted by the rs77721508 variant. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

(surviving past two years).

From the list of four significant variants, we found that three of them were present in the independent dataset. To evaluate their predictive capabilities in distinguishing aggressiveness from indolence, we utilized the genotype information of these three variants as input for the decision tree model. The obtained classification report, as provided in Table 3, demonstrates that the model performed effectively in distinguishing between the aggressive and non-aggressive classes.

To further investigate and clarify the matter, we randomly selected three variants from the dataset and repeated this process 50 times. Each time, we trained and tested the model using these randomly chosen subset of variants. The results were compared with the performance of the model using the three proposed significant variants (i.e., biomarkers). Based on the findings of our study, the accuracy of the model employing the three significant variants was 97 % outperforming the median accuracy obtained from the randomly selected three variants at 65 %. These findings strongly suggest that the model trained on the important variants exhibits superior predictive capabilities regarding the association with aggressive samples, in contrast to the model trained on randomly selected variants. In the course of conducting the experiment 50 times, we observed varying accuracies for each iteration. Fig. 6 visualizes the comparison of accuracies using two subsets of variants - a random subset selection of variants vs. the three proposed biomarkers. The variant employing the proposed biomarkers exhibits no variance, indicative of the outcome being based on a single run.

4. Methodology

4.1. Data preprocessing

As the initial stage in data preprocessing, the MAF file format was converted to the VCF file format. While both VCF and MAF have their applications in genomic data analysis, the preference between them depends on the specific research goal, objectives, and requirements of the study [16–18]. In this study, we converted MAF files to VCF files using vcf2maf tools [19] specifically employing the maf2vcf.pl script. While the MAF format contains extensive annotations for each variant, including information on biological significance, effects, known phenotype associations, and more, the conversion to VCF by maf2vcf.pl generally retains essential variant information, such as genomic position, reference allele, and alternate allele. However, it may not preserve all the rich annotations found in the MAF due to differences in the purposes and structures of the two formats. From an initial set of 64,566 variants, we refined the data by applying a minimum read depth of 10 and a minimum genotype quality of 90. This filtration process ultimately yielded a final count of 6645 variants.

4.2. Statistical analysis

The aim of this statistical analysis is to identify variants that are present in aggressive samples and absent in non-aggressive ones. Accordingly, our null hypothesis posits that no such variant exists solely in aggressive cases and is absent in non-aggressive samples. Given that we have two distinct groups (Aggressive and Non-aggressive), hypothesis testing will be conducted between multiple groups. When performing multiple tests (each variant is individually subjected to the null hypothesis test), each test carries the potential to yield a false positive. This increases the probability of encountering at least one false positive. In fact, the likelihood of obtaining at least one false positive escalates with the increase in the number of tests performed. To correct for multiple comparisons, we employed the permutation correction approach. This is a widely used method for adjusting p-values, taking into account potential correlations.

Table 3
Classification performance metrics of the Decision tree model for distinguishing aggressive and non-Aggressive cases.

	Precision	Recall	F1-score
Aggressive	1	0.9	0.95
Non-aggressive	0.95	1	0.98
Accuracy			0.97

Fig. 7 visually represents the permutation correction approach through a demo example. Here, we have five variants and eight samples, and we have the true label for all the samples. If mutation exists in any sample for any variant, we have labeled it as 1, otherwise 0. The procedure of permutation correction starts by changing the measurements randomly between the aggressive and non-aggressive groups. Alternatively, the same result can be achieved by randomly assigning the “aggressive” and “non-aggressive” labels to the various measurements. Fig. 8, represents the first round of such permutation. For such permutation, we calculated the p-value for each variant. For example, from Fig. 7, we can see that for samples 1 and 3, the first variant is present in aggressive samples and for samples 5 and 8, this variant is not present in non-aggressive samples totaling a sum of $1 + 1 = 2$. We have permuted the labels 10000 times and for each permutation we have performed the hypothesis testing. The p-value for each variant at each permutation is corrected using the Holm’s step-down method [20]. In this study, this method orders the variants in increasing order based on their p-value and make successive smaller adjustments.

Suppose we have a set of m variants. Each variant is classified into one of two categories: aggressive and non-aggressive. The null hypothesis (H_0) for a given variant is that there is no variant that exists in aggressive or in non-aggressive cases. The subsequent procedure for Holm’s step-wise correction is as follows:

1. Compute the p-values, P_1, …., P_m for the m null hypotheses H_{01}, …., H_{0m}.
2. Order the m number of p-values, so that

$$P_1 \le P_2 \le .... \le P_m$$

3. Compare the p-values of each variant with a threshold based on the variant’s position in the ordered list of values.

$$L = p_{(j)} \le \frac{\alpha}{m+1-j} \tag{1}$$

4 Reject all null hypotheses H_{0j} for which

$$p_{(j)} < p_{(L)}$$

Once the calculation of first permutation is completed, a new permutation is formed and new p-values resulting from this permutation is calculated. This entire procedure (random labeling and testing) is repeated tens of thousands of times. The p-value for variant (i) is the proportion of times that the value of t calculated for the real labels t_i is less than or equal to the value of t calculated for random permutations.

$$p_i = \frac{\text{Number of permutations for which} u_j{}^{(b)} \ge t_i}{\text{Total number of permutations}} \tag{2}$$

where $u_j{}^{(b)}$ are the values corrected as in Holm’s step-down method for permutation b.

4.3. Pathway analysis

The primary objective of Pathway analysis is to identify pathways that experience significant impacts from genetic variants. This process involves assigning scores to pathways based on the enrichment of genes affected by at least one preset variant. The scoring method used is known as Over Representation Analysis (ORA), which generates a unique p-value, denoted as pORA, for each pathway and set of variants.

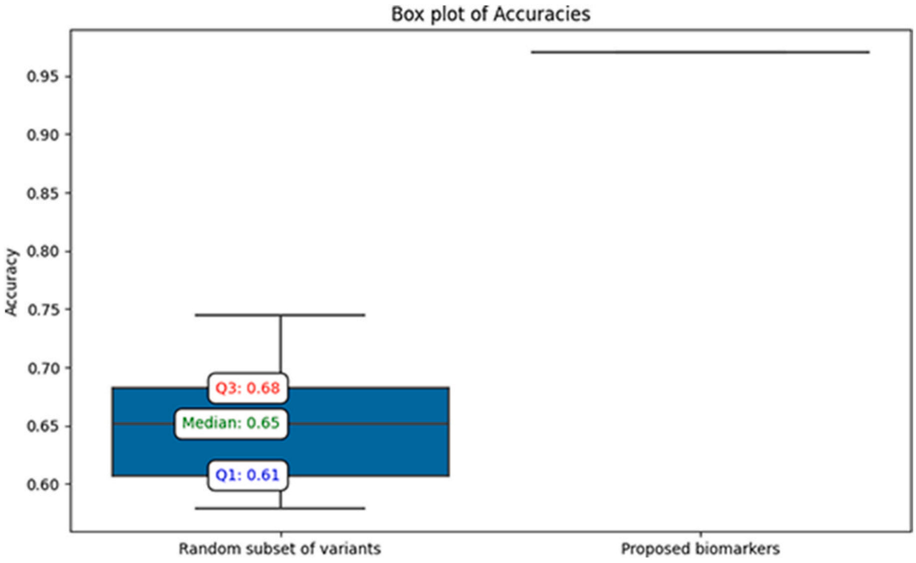
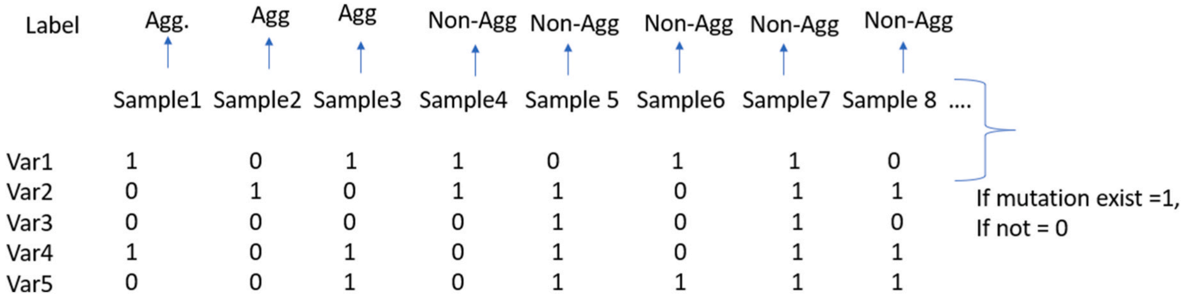


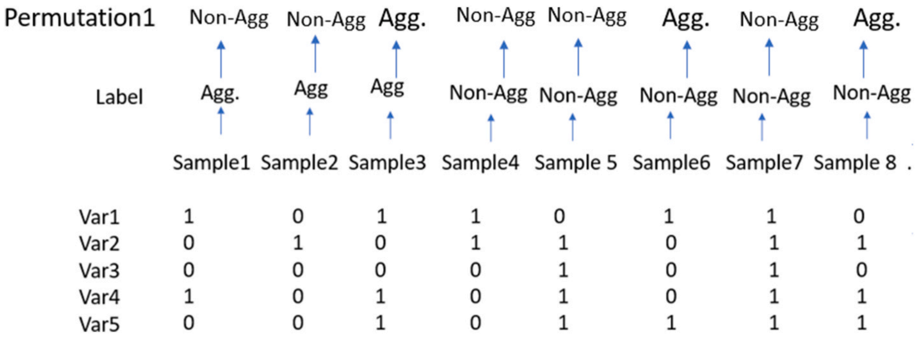
Fig. 6. Comparative analysis of model accuracies. The box plot presents the distribution of accuracies for two distinct models: a random subset of variants (50 runs) and a proposed biomarker set.



Label	Agg.	Agg	Agg	Non-Agg	Non-Agg	Non-Agg	Non-Agg	Non-Agg
	Sample1	Sample2	Sample3	Sample4	Sample 5	Sample6	Sample7	Sample 8 ....
Var1	1	0	1	1	0	1	1	0
Var2	0	1	0	1	1	0	1	1
Var3	0	0	0	0	1	0	1	0
Var4	1	0	1	0	1	0	1	1
Var5	0	0	1	0	1	1	1	1

If mutation exist =1,
If not = 0

Fig. 7. Visual representation of the permutation correction approach employed to analyze multiple groups. Each variant in the samples is assigned a label of 1 if a mutation is present, and 0 if not. These labels serve as the true labels for all the samples.



Permutation1	Non-Agg	Non-Agg	Agg.	Non-Agg	Non-Agg	Agg.	Non-Agg	Agg.
Label	Agg.	Agg	Agg	Non-Agg	Non-Agg	Non-Agg	Non-Agg	Non-Agg
	Sample1	Sample2	Sample3	Sample4	Sample 5	Sample6	Sample7	Sample 8 .
Var1	1	0	1	1	0	1	1	0
Var2	0	1	0	1	1	0	1	1
Var3	0	0	0	0	1	0	1	0
Var4	1	0	1	0	1	0	1	1
Var5	0	0	1	0	1	1	1	1

Fig. 8. Illustration of the initial iteration of permutation testing, serving as a demonstration of the first phase in a series of 10,000 label permutations.

To obtain the pathway composition, encompassing all genes associated with a specific pathway, we referred to the KEGG database [21]. The pORA value represents the probability of observing an equal or greater number of impacted genes in a given pathway, purely by chance [22, 23].

Suppose we have N genes measured in the experiment, and out of these, M genes are associated with the specific pathway under investigation. Through a priori selection of impacted genes using Preset Variants, K out of the M pathway-associated genes were identified as impacted. The significance of the pathway is determined based on an assessment of whether the number of impacted genes observed is unexpectedly high. To evaluate the improbability of observing K or a greater number of impacted genes on the pathway, we calculate the probability of randomly selecting K or more out of the M genes measured within the pathway.

For any number x, the probability of observing exactly x impacted genes on the given pathway is computed based on the hypergeometric distribution:

$$P(X = x|N, M, K) = \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}} \tag{3}$$

Since the hypergeometric distribution is a discrete probability distribution, we can calculate the probabilty of observing fewer than x genes affected on a specific pathway just by chance. This can be achieved by summing the probabilities of randomly observing 1, 2, …, up to x-1 impacted genes on that pathway:

$$p_u(x-1) = P(X=1) + P(X=2) + \ldots + P(X=x-1) = \sum_{i=0}^{x-1} \frac{\binom{M}{i}\binom{N-M}{K-i}}{\binom{N}{K}} \tag{4}$$

Then we calculated the probability of randomly observing a number of impacted genes on the given pathway that is greater than or equal to the number of impacted genes obtained from data, by computing the over-representation p-value: pORA = p_o(x) = 1-p_u(x-1):

$$p_o(x) = 1 - \sum_{i=0}^{x-1} \frac{\binom{M}{i}\binom{N-M}{K-i}}{\binom{N}{K}} \tag{5}$$

5. Discussion

In this study, our objective was to identify genetic variants as potential biomarkers for aggressive PCa patients at the genetic variant level. To establish these variants as biomarkers, we employed a rigorous statistical analysis process to identify variants that were present in aggressive PCa samples but absent in non-aggressive PCa samples. However, it is important to note that while these variants may be statistically significant, not all of them may have an effect on the differential expression of genes. Therefore, to investigate this further, we examined the genes that were highly impacted by the identified set of significant variants as determined by our statistical analysis. Additionally, we explored the pathways associated with these highly impacted genes, as they carry the variants of interest.

For our analysis, we converted the MAF formatted dataset to VCF files. This conversion was necessary because VCF files provide information on all transcripts affected by a mutation, whereas MAF files only report on the most significantly impacted ones [24,25]. At the initiation of the study, the dataset consisted of 64,566 variants. We proceeded to extract genotype information for each variant, applying specific filters [26,27].

Following the completion of the statistical analysis, we identified four significant variants out of the initial pool of 6645 variants. These four variants exhibited statistical significance at a 5 % significance level. Notably, within this set of significant variants, we discovered one particular variant that has been previously reported in the literature to be associated with an increased risk of aggressive prostate cancer (PCa). The association of rs5759167 with an increased risk of prostate cancer (PCa) has been consistently revealed through genome-wide association studies [28,29]. This variant has been identified in numerous studies involving key genes such as BRCA1, BRCA2, MMR, HOXB13, CHEK2, and NBS1. The collective findings from these studies indicate that rs5759167 is associated with moderate risks of PCa development and may contribute to a more aggressive disease phenotype.

The most noteworthy variant identified in our study, rs777215086 (p-value = 0.0012), exhibits a potential impact on the JAK1 gene. In a recent investigation, researchers explored JAK1 expression in prostate cancer (PCa) using RNA-sequencing data from The Cancer Genome Atlas (TCGA). Their findings revealed a significant decrease in JAK1 expression in PCa compared to adjacent normal tissues [30,31]. In our study, we have identified a novel variant in the TRPM7 gene associated with prostate cancer. We thoroughly examined publicly available databases, but found no previous mention of this specific variant. To determine the potential impact of this novel variant, we employed the Over-representation Analysis (ORA) method using iVariantGuide [32]. ORA is a method for objectively assessing whether a set of biologically relevant variables, such as a gene set or pathway, occurs more frequently in a set of variables of interest than expected by chance [33]. Our analysis revealed that the novel variant exerts a substantial influence on the TRPM7 gene Interestingly, recent research has explored the expression levels of TRPM7 in various types of PCa tumors [34]. The final variant that demonstrated significance is rs864309495. Through the application of the ORA analysis method, we have determined that this variant exerts a notable impact on the TP53 gene. A study conducted by Yong et al. [35] revealed that TP53 loss-of-function is associated with elevated levels of autophagy-related proteins in aggressive PCa. In a separate investigation by Thorsten et al. [36], the clinical significance of p53 alterations in surgically treated PCa patients was explored.

By utilizing the iVariantGuide tool, which incorporates the Over-representation Analysis (ORA), we have not only investigated the genes profoundly influenced by our variants of interest but also examined the pathways that exhibited substantial significance between aggressive and non-aggressive groups. Our findings indicate that the significant variants identified in our study exerted a significant influence on the most prominent pathways associated with aggressive PCa. The most significant pathway in aggressive PCa group was PI3K-Akt signaling pathway. Paul et al. [37] expressed their anticipation, in a comprehensive review, that gaining a deeper understanding of the biology of the PI3K/Akt pathway in PCa would facilitate the identification of relevant biomarkers and enable the development of rational combination therapies. Moreover, the activation of the PI3K/Akt pathway is a common characteristic observed in many cases of aggressive PCa. As PCa progresses towards a resistant and metastatic state, the activation of this pathway becomes even more prevalent. Signaling cascades emanating from the PI3K/Akt pathway stimulate a multitude of survival, growth, metabolic, and metastatic functions, all of which are hallmarks of aggressive cancer [38,39]. In a study conducted by Taylor et al. [40], an integrative genomic profiling of human prostate cancer (PCa) was performed. The research demonstrates a compelling interest in utilizing the PI3K/Akt pathway as a biomarker to distinguish highly significant, aggressive prostate cancer cases from less aggressive forms of the disease. However, it is crucial to acknowledge that the utilization of this pathway as a biomarker faces significant challenges, primarily due to the intricate nature of the biology in advanced PCa and the presence of tumor heterogeneity.

The overall discussion illustrates the effect of significant variants on genes and biological pathways in order to comprehend how our list of significant variants can affect the expression or function of genes and biological pathways or processes that have a remarkable effect on aggressive PCa. Based on the findings from our comprehensive study, it is evident that the variants identified in our research, along with the genes and pathways they impact, play a significant role in driving the progression of aggressive PCa. However, our research was conducted with a relatively limited dataset, primarily due to the inherent challenge in acquiring datasets for aggressive Prostate Cancer patients. Given that our approach is data-driven, expanding the study population would indeed be instrumental for future endeavors. The methodology employed in our study is based on a hypothesis testing paradigm, facilitating the identification of significant variants. However, each variant has a comprehensive suite of attributes (e.g., location, conservation, epigenetics) which can provide crucial insights regarding its association with a phenotype. Thus, integrating these features alongside the hypothesis can offer a more robust prediction on the relevance of specific variants to particular traits. These results align with the current state of the art and further emphasize the importance of understanding the genetic factors and biological mechanisms contributing to the development and advancement of aggressive PCa.

6. Conclusion

The primary objective of this research was to identify genetic variants that could serve as biomarkers in prostate cancer. To achieve this goal, we employed a statistical approach aimed at identifying variants present in aggressive PCa samples but absent in non-aggressive PCa samples. Based on this hypothesis, we successfully identified four variants that exhibited statistical significance at a 5 % level. Furthermore, through the utilization of Over-representation Analysis (ORA), we investigated the specific genes and biological pathways influenced by these significant variants. The ORA analysis shed light on the impact of these variants at the molecular level. Lastly, a comprehensive literature review corroborated the significance of our identified outcomes, highlighting their pivotal role in driving the progression of aggressive PCa.

CRediT authorship contribution statement

Tasnimul Alam Taz: Performed the experiments, Analyzed the data and results, Wrote the paper. Suzan Arslanturk: Conceived & designed the project, Analyzed the data & results, Wrote the paper.

Data availability

The results published here are in whole or part based upon data generated by The Cancer Genome Atlas managed by the NCI and NHGRI and CBioPortal. Information about TCGA can be found at http://cancergenome.nih.gov. Information about CBioPortal can be found at https://www.cbioportal.org.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was funded by the National Science Foundation (NSF: #1948338), and the Department of Defense (DoD: #W81XWH-21-1-0570). Both authors have read and agreed to the published version of the manuscript.

References

[1] Esteva A, Feng J, van der Wal D, Huang SC, Simko JP, DeVries S, Chen E, Schaeffer EM, Morgan TM, Sun Y, Ghorbani A. Prostate cancer therapy personalization via multi-modal deep learning on randomized phase III clinical trials. NPJ Digital Medicine 2022;5(1):71.

[2] Ottman R, Ganapathy K, Lin HY, Osterman CD, Dutil J, Matta J, Ruiz-Deya G, Wang L, Yamoah K, Berglund A, Chakrabarti R. Differential expression of miRNAs contributes to tumor aggressiveness and racial disparity in african American men with prostate cancer. Cancers 2023;15(8):2331.

[3] Swami U, McFarland TR, Nussenzveig R, Agarwal N. Advanced prostate cancer: treatment advances and future directions. Trends in cancer 2020;6(8):702–15.

[4] Li J, Guan X, Fan Z, Ching LM, Li Y, Wang X, Cao WM, Liu DX. Non-invasive biomarkers for early detection of breast cancer. Cancers 2020;12(10):2767.

[5] Montel RA, Gregory M, Chu T, Cottrell J, Bitasktsis C, Chang SL. Genetic variants as biomarkers for progression and resistance in multiple myeloma. Cancer genetics 2021;252:1–5.

[6] Chen F, Madduri RK, Rodriguez AA, Darst BF, Chou A, Sheng X, et al. Evidence of novel susceptibility variants for prostate cancer and a multiancestry polygenic risk score associated with aggressive disease in men of african ancestry. Eur Urol 2023;84(1):13–21.

[7] Allemailem KS, Almatroudi A, Alrumaihi F, Almansour NM, Aldakheel FM, Rather RA, Rah B. Single nucleotide polymorphisms (SNPs) in prostate cancer: its implications in diagnostics and therapeutics. Am J Tourism Res 2021;13(4):3868.

[8] Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly 2012;6(2):80–92.

[9] Albacker LA, Wu J, Smith P, Warmuth M, Stephens PJ, Zhu P, Yu L, Chmielecki J. Loss of function JAK1 mutations occur at high frequency in cancers with microsatellite instability and are suggestive of immune evasion. PLoS One 2017;12(11):e0176181.

[10] Danziger O, Shai B, Sabo Y, Bacharach E, Ehrlich M. Combined genetic and epigenetic interferences with interferon signaling expose prostate cancer cells to viral infection. Oncotarget 2016;7(32):52115.

[11] Chen L, Cao R, Wang G, Yuan L, Qian G, Guo Z, Wu CL, Wang X, Xiao Y. Downregulation of TRPM7 suppressed migration and invasion by regulating epithelial–mesenchymal transition in prostate cancer cells. Med Oncol 2017;34:1–11.

[12] Sun Y, Selvaraj S, Varma A, Derry S, Sahmoun AE, Singh BB. Increase in serum Ca2+/Mg2+ ratio promotes proliferation of prostate cancer cells by activating TRPM7 channels. J Biol Chem 2013;288(1):255–63.

[13] Chung BH, Horie S, Chiong E. The incidence, mortality, and risk factors of prostate cancer in Asian men. Prostate international 2019;7(1):1–8.

[14] Na R, Liu F, Zhang P, Ye D, Xu C, Shao Q, Qi J, Wang X, Chen Z, Wang M, He D. Evaluation of reported prostate cancer risk-associated SNPs from genome-wide association studies of various racial populations in Chinese men. Prostate 2013;73(15):1623–35.

[15] Sirohi D, Devine P, Grenert JP, van Ziffle J, Simko JP, Stohr BA. TP53 structural variants in metastatic prostatic carcinoma. PLoS One 2019;14(6):e0218618.

[16] Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G. The variant call format and VCFtools. Bioinformatics 2011;27(15):2156–8.

[17] Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 2013;31(3):213–9.

[18] Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, Rynes E, Maurano MT, Vierstra J, Thomas S, Sandstrom R. BEDOPS: high-performance genomic feature operations. Bioinformatics 2012;28(14):1919–20.

[19] Kalmár A, Galamb O, Szabó G, Pipek O, Medgyes-Horváth A, Barták BK, Nagy ZB, Szigeti KA, Zsigrai S, Csabai I, Igaz P. Patterns of somatic variants in colorectal adenoma and carcinoma tissue and matched plasma samples from the Hungarian oncogenome program. Cancers 2023;15(3):907.

[20] Huang Y, Hsu JC. Hochberg's step-up method: cutting corners off Holm's step-down method. Biometrika 2007;94(4):965–75.

[21] Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res 2014;42(D1):D199–205.

[22] Draghici S, Khatri P, Bhavsar P, Shah A, Krawetz SA, Tainsky MA. Onto-tools, the toolkit of the modern biologist: onto-express, onto-compare, onto-design and onto-translate. Nucleic Acids Res 2003;31(13):3775–81.

[23] Draghici S. Statistics and data analysis for microarrays using R and bioconductor. CRC Press; 2016.

[24] Fan Y, Xi L, Hughes DS, Zhang J, Zhang J, Futreal PA, Wheeler DA, Wang W. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. Genome Biol 2016;17(1):1–11.

[25] Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 2013;31(3):213–9.

[26] Wang H, Avillach P. Diagnostic classification and prognostic prediction using common genetic variants in autism spectrum disorder: genotype-based deep learning. JMIR medical informatics 2021;9(4):e24754.

[27] Doan RN, Lim ET, De Rubeis S, Betancur C, Cutler DJ, Chiocchetti AG, Overman LM, Soucy A, Goetze S, Autism Sequencing Consortium and Freitag, C.M.. Recessive gene disruptions in autism spectrum disorder. Nat Genet 2019;51(7):1092–8.

[28] Dias A, Kote-Jarai Z, Mikropoulos C, Eeles R. Prostate cancer germline variations and implications for screening and treatment. Cold Spring Harbor perspectives in medicine 2018;8(9).

[29] Lange EM, Johnson AM, Wang Y, Zuhlke KA, Lu Y, Ribado JV, Keele GR, Li J, Duan Q, Li G, Gao Z. Genome-wide association scan for variants associated with early-onset prostate cancer. PLoS One 2014;9(4):e93436.

[30] Chen B, Lai J, Dai D, Chen R, Li X, Liao N. JAK1 as a prognostic marker and its correlation with immune infiltrates in breast cancer. Aging (Albany NY) 2019;11(23):11124.

[31] Rossi MR, Hawthorn L, Platt J, Burkhardt T, Cowell JK, Ionov Y. Identification of inactivating mutations in the JAK1, SYNJ2, and CLPTM1 genes in prostate cancer cells using inhibition of nonsense-mediated decay and microarray analysis. Cancer Genet Cytogenet 2005;161(2):97–103.

[32] Chaudhry SR, Tainsky MA. Utilizing iVariantGuide for variant assessment of next-generation sequencing. Current Protocols in Bioinformatics 2019;65(1):e73.

[33] Pomyen Y, Segura M, Ebbels TM, Keun HC. Over-representation of correlation analysis (ORCA): a method for identifying associations between variable sets. Bioinformatics 2015;31(1):102–8.

[34] Lee EH, Lee JN, Park S, Chun SY, Yoon BH, Chung JW, Choi SH, Kim BS, Kim HT, Kim TH, Yoo ES. Inhibition of TRPM7 suppresses migration and invasion of prostate cancer cells via inactivation of ERK1/2, Src and Akt pathway signaling. J Mens Health 2022;18(7):144.

[35] Zhang Y, Song XL, Yu B, Foong LC, Shu Y, Mai CW, Hu J, Dong B, Xue W, Chua CW. TP53 loss-of-function causes vulnerability to autophagy inhibition in aggressive prostate cancer. Int J Urol 2022;29(9):1085–94.

[36] Schlomm T, Iwers L, Kirstein P, Jessen B, Köllermann J, Minner S, Passow-Drolet A, Mirlacher M, Milde-Langosch K, Graefen M, Haese A. Clinical significance of p53 alterations in surgically treated prostate cancers. Mod Pathol 2008;21(11):1371–8.

[37] Toren P, Zoubeidi A. Targeting the PI3K/Akt pathway in prostate cancer: challenges and opportunities. Int J Oncol 2014;45(5):1793–801.

[38] Scher HI, Fizazi K, Saad F, Taplin ME, Sternberg CN, Miller K, De Wit R, Mulders P, Chi KN, Shore ND, Armstrong AJ. Increased survival with enzalutamide in prostate cancer after chemotherapy. N Engl J Med 2012;367(13):1187–97.

[39] Ryan CJ, Smith MR, De Bono JS, Molina A, Logothetis CJ, De Souza P, Fizazi K, Mainwaring P, Piulats JM, Ng S, Carles J. Abiraterone in metastatic prostate cancer without previous chemotherapy. N Engl J Med 2013;368(2):138–48.

[40] Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, Carver BS, Arora VK, Kaushik P, Cerami E, Reva B, Antipin Y. Integrative genomic profiling of human prostate cancer. Cancer Cell 2010;18(1):11–22.

[41] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28:27–30.

[42] Kanehisa M. Toward understanding the origin and evolution of cellular organisms. Protein Sci 2019;28:1947–51.

[43] Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res 2023;51:D587–92.