Computational Identification of Ovarian Cancer Candidate Driver Genes with Mutated Protein Structures Caused by Missense Variants

Ian Hou[1] and Yongsheng Bai[2,3,*]

[1] The John Cooper School, The Woodlands, Houston, USA
[2] Next-Gen Intelligent Science Training, Ann Arbor, MI, USA
[3] Department of Biology, Eastern Michigan University, Ypsilanti, MI, USA
Email: ihou651088@gmail.com (I.H.); bioinformaticsresearchtomorrow@gmail.com (Y.B.)
*Corresponding author

***Abstract*—Ovarian cancer detection remains elusive due to a lack of screening tests and non-specific symptoms. A crucial factor in cancer development is DNA sequence mutations, particularly missense mutations that can alter protein structure, thereby potentially initiating carcinogenesis. Advances in sequencing technology have paved the way for detailed analysis of individual genetic profiles, spotlighting genes with missense mutations as prospective biomarkers. Such biomarkers are pivotal for personalizing cancer therapies, as they can guide medication choices, ensuring efficacy and minimizing detrimental effects. Despite tools like AlphaFold predicting 3D protein structures and Phyre2 assessing mutated amino acid impacts, no model concurrently predicts wild-type and mutated protein structures. Also, integrating structure changes with drug target identification remains under-explored. Analyzing the TCGA Ovarian Cancer transcriptome data, this research postulated that missense mutations in highly expressed genes significantly influence protein structure, earmarking these genes as potential therapeutic targets. Twelve genes were discerned to affect ovarian cancer patient survival rates. An original platform, MiSeVis, was introduced, offering insights into potential drug targets for specific genes, survival analysis, and 3D protein structure alterations. This comprehensive methodology, unifying transcriptome analysis, pinpointing genes with impactful missense mutations, and presenting a user-centric visualization tool, marks considerable progress in ovarian cancer treatment and precision medicine.**

***Keywords*—ovarian cancer, missense mutations, sequencing technology, biomarkers, transcriptome analysis, protein structure, survival analysis, personalized therapy**

I. Introduction

Ovarian cancer detection poses significant challenges, primarily due to several factors, such as a lack of a screening test and vague, non-specific symptoms, making it difficult to identify the disease in its early, more treatable stages. One of the key drivers of cancer development is mutations in the DNA sequence of the genome. A mutation is a change in the DNA sequence of the genome that may lead to the development of cancer. A missense mutation can cause a change in protein structure and its function, thus transforming normal cells into cancerous ones. These deleterious mutations have a higher probability of changing protein stability and pathogenicity, making them particularly valuable for early cancer detection [1]. Recent advancements in sequencing technology have ignited hope for the early detection of ovarian cancer. These breakthroughs enable researchers to delve into the genetic makeup of individuals and their tumors, shedding light on specific candidate genes with missense mutations. Identification of these genes is becoming increasingly vital [2], as they may serve as potential biomarkers.

In a previous study, the author adopted a large-scale protein-based model to predict functional and structurally disruptive variant effects and identified several Single Amino-acid Variants (SAVs) for gold-standard candidate genes [3]. These types of disruptive variants often affect protein stability, which will consequently alter protein structure and function. Existing cutting-edge tools that utilize machine learning methods such as AlphaFold can predict the 3D structures of proteins, including mutated ones with high confidence [4]. AlphaFold is an artificial intelligence-based system designed to predict 3-dimensional structures of proteins based on their amino acid sequence. The prediction confidence is quantified using a confidence score known as the predicted Local Difference Distance Test (pLDDT) score. Additionally, the tool Phyre2 builds 3D structures by assessing the impact of amino acid sequences including mutated sequences. Users can input protein sequences into the web server to generate these structural predictions [5].

Each cancer patient has a unique set of mutations, which can serve as specific biomarkers. These biomarkers are essential for tailoring cancer treatment strategies and realizing personalized medicine. For example, they may help identify drugs that are likely to be ineffective or even deadly to the patient, allowing doctors to administer specific drugs that may interact

Manuscript received October 23, 2023; revised December 20, 2023; accepted December 28, 2023; published May 29, 2024.

with or change protein structure function. Different clinical studies have also been conducted to evaluate the role of biomarkers in drug development. Finding prognostic biomarkers requires knowledge of tumor and host immune system interactions [6].

TIMER2.0, a widely used bioinformatics database, has been developed to elaborate tumor and host interactions in the tumor microenvironment through The Cancer Genome Atlas (TCGA) database analysis [7]. The existing studies suggested that personalized immunotherapeutic treatment strategies are important in targeting anatomical sites and biomarkers due to unique reaction mechanisms for different immune cell types [8].

However, there is a lack of generalized models that can simultaneously predict the structure of both wild-type and mutated proteins due to missense variants. In addition, identifying drug targets based on candidate biomarkers has not been explored in the context of the structure changes. It's worth noting that predicting the structure of larger mutated proteins requires higher processing power and subsequently longer queue time. Furthermore, analysis of databases such as cBioPortal and Phyre2 requires expertise in identifying genomic markers and other uses for each database. A stand-alone downloadable visualization platform that allows users to access these results easily is an ideal solution.

In this project, the author initially analyzed TCGA Ovarian Cancer transcriptome sequencing data to select highly expressed genes with their missense mutations. The hypothesis was that missense mutations occurring in highly expressed genes are likely to affect the protein's structure, making the host genes potential therapeutic targets. This study has identified 12 candidate genes highly expressed in ovarian cancer patients that either lower or raise the survival rate of these patients. MiSeVis, an innovative R-Shiny-based platform developed in this project, enables users to report potential drug targets for a given gene, output a survival analysis graph, and observe changes in 3D protein structure, providing invaluable insights into the potential functional consequences of these genetic variations. This integrated approach through combining transcriptome analysis, identification of highly expressed genes with missense mutations, and the development of a user-friendly visualization platform, represents a significant step forward in the quest for effective ovarian cancer therapies and precision medicine.

The MiSeVis software platform is freely available on the web at https://github.com/IHou594/MiSeVis.

II. Materials and Methods

A. Candidate Gene Selection and Mutated Sequence Generation

The author first explored the National Cancer Institutes Database — The Cancer Genome Atlas and identified 44 genes with high expression in ovarian cancer, as revealed by the TCGA ovarian cancer dataset. After searching the Ensembl and Alphafold databases to retrieve the amino acid sequences of these genes, the ovarian cancer datasets and the Pan-Cancer analysis of the whole genome dataset from NCBI cBioPortal were downloaded as well. By inserting the mutation information from cBioPortal using a custom script the author wrote to generate novel mutated sequences, both the wild-type and mutated sequences were fed into Phyre2 for 3D structure analysis. The genes that showed a protein structure change due to missense mutations were kept for the next step of the analysis.

B. Drug Target Identification, Immune Infiltration, and Survival Analysis of Candidate Genes

The author searched DrugBank to investigate available drug targets of these 44 genes with the predicted protein structure changes by Phyre2, and those with drug targets were recorded. The R package survival (version 3.5-7) was used to create the Kaplan Meier survival plot to identify candidate genes that affect patient survival. The gene immune infiltration analysis was conducted for the 44 candidate genes using TIMER2.0 to pinpoint genes correlated with OV-infiltrating Immune Cell Types. Specifically, correlation scores were calculated for various in vitro immune cell types of ovarian cancer. The author subsequently selected immune infiltrates such as T Cell CD8, T Cell CD4, dendritic cell, B cell, neutrophil, and macrophage to check each candidate gene and analyzed their association between gene expression and immune infiltration level.

C. Development of the MiSeVis platform

To facilitate the broader scientific community's usage, The author wrote a custom script and developed a convenient platform named MiSeVis using RShiny. The complete workflow of the analysis is shown in Fig. 1.

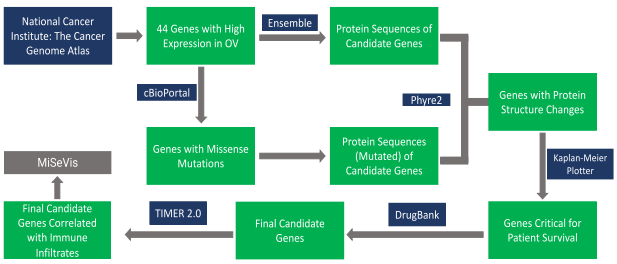


Fig. 1. Workflow of identifying ovarian cancer driving genes.

III. Result and Discussion

A. Association Analysis between Immune Infiltrates and Gene Expression in TCGA

Out of the 44 genes that were obtained from the Cancer Genome Atlas, 12 candidate genes have been demonstrated to affect cancer probability through analysis of multiple databases, including Phyre2 and DrugBank. Among these candidates, five drug targets (DB01593, DB09130, DB14487, DB14533, and DB14548) were found to be recurring three times each (Table I).

The author obtained four oncogenes (*COL3A1*, *FN1*, *CLU*, *FTL*) and 6 tumor suppressor genes (*GNAS*, *UBC*, *EEF2*, *PSAP*, *TUBA1B*, *HSP90AB1*) based on the TCGA

patient RNA-seq data. The remaining genes do not have a clear tumorigenesis classification.

TABLE I. PREDICTED 3D STRUCTURES FOR THE MUTATED CANDIDATE GENES USING PHYRE2 AND POTENTIAL DRUGBANK IDS

Gene Name	Phyre2 Prediction	Structural Change	Drug Bank Target
ACTG1	Available	Y	DB09130; DB11638
EEF1A1	Available	Y	DB01593; DB04315; DB09130; DB11638; DB14487; DB14533; DB14548
FTL	Available	Y	DB00893; DB02285; DB09147; DB09517; DB13995
COL3A1	Available	Y	DB00048
FN1	Available	Y	DB01593; DB06245; DB08888; DB14487; DB14533; DB14548
CLU	Available	Y	DB01593; DB09130; DB14487; DB14533; DB14548
PSAP	Available	Y	DB01966
TUBA1B	Available	Y	DB01873; DB03010; DB05147; DB07574
HSP90AB1	Available	Y	DB02424; DB02754; DB03758; DB05134; DB06070; DB07594; DB07877; DB08045; DB08153; DB08292; DB08293; DB08346; DB08464; DB08465; DB09221
GNAS	Available	Y	DB02587; DB06843
UBC	Available	Y	DB04464
EEF2	Available	Y	DB02059; DB03223; DB04315; DB08348; DB11823; DB12688

TABLE II. SURVIVAL ANALYSIS RESULT AND IMMUNE CELL TYPE CORRELATION ANALYSIS FOR THE CANDIDATE GENES

Gene Name	UALCAN Result	T Cell CD8	Neutrophil
ACTG1	Not Clear	−0.003	0.117
EEF1A1	Not Clear	−0.17	−0.193
FTL	Oncogene	0.116	0.13
COL3A1	Oncogene	0.191	0.177
FN1	Oncogene	0.166	0.296
CLU	Oncogene	0.167	−0.086
PSAP	Tumor Surpressant	0.256	0.368
TUBA1B	Tumor Surpressant	0.005	0.212
HSP90AB1	Tumor Surpressant	−0.122	0.013
GNAS	Tumor Surpressant	0.028	−0.061
UBC	Tumor Surpressant	0.222	0.205

Note: Red numbers indicate statistically significant positively correlated values while blue numbers represent statistically significant negative correlations.

The immune association analysis revealed that most oncogenes exhibit a positive correlation with immune infiltrate CD8 and neutrophil immune cell type, whereas tumor suppressor genes don't generally exhibit such trends (Table II). These trends are shown on TIMER2.0 as a heat map table representing variations in immune infiltration levels between tumors with mutations in the input gene and tumors without mutations in the input gene. Additionally, the results indicated differential expression of oncogenes in the immune infiltrate CD8 and neutrophil when compared to other immune cell types (data not shown). Furthermore, these results were cross-examined with those from The University of Alabama at Birmingham Cancer Data Analysis Portal (UALCAN), a web source for accessing publicly available cancer data, performing gene expression analysis, etc.

B. Survival Analysis of Cancer Biomarkers

Kaplan-Meier survival analysis was conducted to further evaluate the clinical implications of the candidate genes. These analysis modules have been integrated into the bioinformatics software that was developed, MiSeVis. An illustrative example of oncogene classification is shown in Fig. 2.

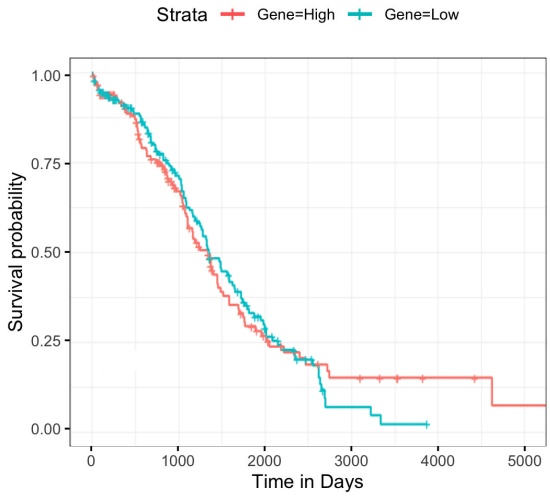


Fig. 2. Kaplan-Meier survival analysis of tumor suppressor gene PSAP expression (high vs. low) on overall survival in ovarian cancer.

Table III shows that some oncogenes are recurrently mutated and have the same amino acid positions. Although tumor suppressor genes also mutated at the same amino acid positions, there are also mutations through protein truncating alterations. This result is consistent with previous published studies [9].

TABLE III. PROTEIN MISSENSE MUTATION INFORMATION FOR ONCOGENES AND TUMOR SUPPRESSOR GENES

Gene Name	Mutation Location
COL3A1	G243E; G294E; V529F; G1041V; G1128C; A1203T; K1407R
FN1	M119T; R222C; C258Y; Q792E; S390L; S1340N; P1584A; T2163S; T2254S; D2331
CLU	W110C; R198W; P234L
PSAP	P294S; G480R; C482F
EEF1A1	W58*; H95Qfs*15
GNAS	T90A; Y163*; F273L; GNAS-IGF2 FUSION
UBC	G35Dfs*15; I36Sfs*38; D52N; D52Nfs*34; E64K; G111Dfs*15; L112Sfs*15; G162SL208V; T387Hfs*9
EEF2	G31C; L315V; I451S

C. MiSeVis: An R-Shiny Based Application of Modeling 3D Protein Structures and Predicting Drug Targets of Cancer Biomarkers

Using previous research results, the author created a convenient platform called MiSeVis using R-Shiny by combining all analysis steps. MiSeVis features a protein visualization function wherein users can choose a PDB file and visualize the 3D representation of the protein's structure. Users have the flexibility to choose between mutated and non-mutated files and scrutinize the exact mutation locations using various imaging options. Additionally, the platform has a function that allows users to view the contents of both FASTA and PDB files and input their own files as well. As an illustrative case study, the author visualized a gene Prosaposin (PSAP) with a mutation at amino acid position 469. The software clearly showed how the protein structure is altered and reported its drug targets.

1) MiSeVis RShiny tool implementation

This pipeline not only outputs the contents of an inputted sequence and structure file but also enables users to visualize 3D models of a PDB file (Fig. 3). Users can manipulate the model's parameters, analyze individual amino acids through the generated model, and search for drug matches in the DrugBank database based on their selection of the PDB visualization file.

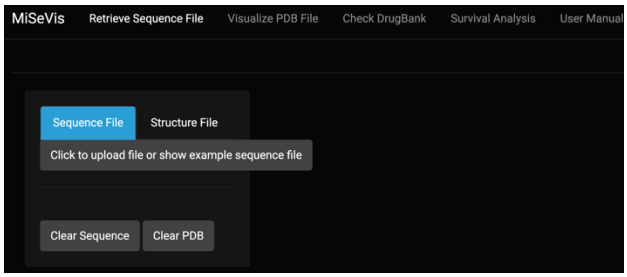


Fig. 3. Overview of RShiny GUI for MiSeVis.

2) Retrieve sequence file

The "Structure File" tab operates like the "Sequence File" tab, except the required file type (Fig. 4). In the "Structure File" tab, users are required to upload a Protein Data Bank (PDB) file, rather than a FASTA file, to ensure accurate processing and display of structural information. On the "Retrieve Sequence" Tab, there are two options for file input: A Sequence file and a Structure file. In the Sequence File section, there is a button accessible via the "Click to Show Upload Option" button, which allows users to upload a file from their local computer. It's important to note that the uploaded file must be in FASTA format to ensure proper rendering of the inputted FASTA file. Users may click on the "Clear Sequence" button to clear the sequence shown or the "Clear PDB" button to get rid of the output of the Structure File tab. The "Structure File" tab works very similarly to the "Sequence File" tab except the required file type. Users are required to upload a Protein Data Bank file instead of a FASTA file.

Moreover, the user has the option to clear the displayed sequence by clicking the "Clear Sequence" button or eliminate the output from the "Structure File" tab by clicking the "Clear PDB" button.

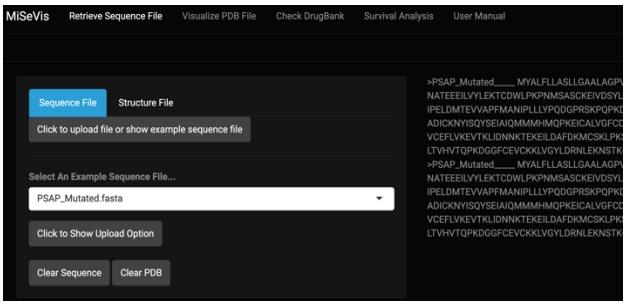
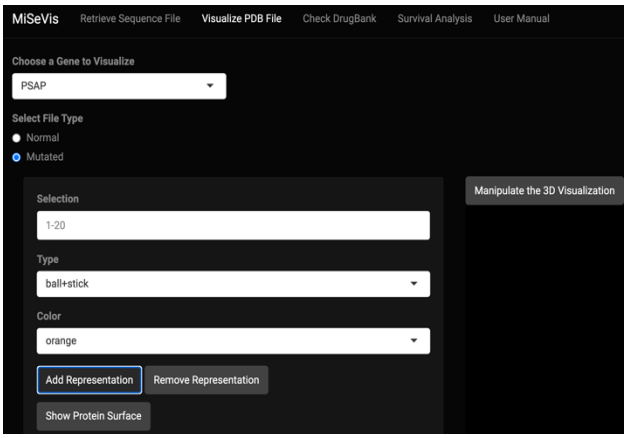


Fig. 4. Example of sequence retrieval in MiSeVis.

3) Visualize PDB file

The "Visualize PDB" tab allows users to examine and modify the 3D model of the inputted gene (Fig. 5). To begin, users must select a gene in the "Choose a Gene to Visualize" dropdown menu. Then, they need to specify whether they would like to view the normal or mutated structure of the gene. The 3D visualization will then appear automatically, enabling users to navigate the model using their cursor to explore the different sides of the model. Furthermore, since the model is based on the specified quaternary structure of the PDB file, users can hover over individual segments of the protein and observe specific amino acids at precise locations in the amino acid sequence. Users also have the option to highlight certain sections of the protein by clicking the "Manipulate the 3D Visualization" button. The first parameter is the amino acid selection, which allows the user to define the range of highlighted amino acids. The second parameter is the representation type for the highlighted amino acids. The third parameter specifies the color of the highlighted regions in the protein. The "Add" button will add the specified parameters while the "Remove" button will remove them. Finally, users can visualize the surface of the entire 3D model by clicking the "Click to Show Protein Surface" button. However, it's essential to have the type parameter set to "surface" for this button to function properly.



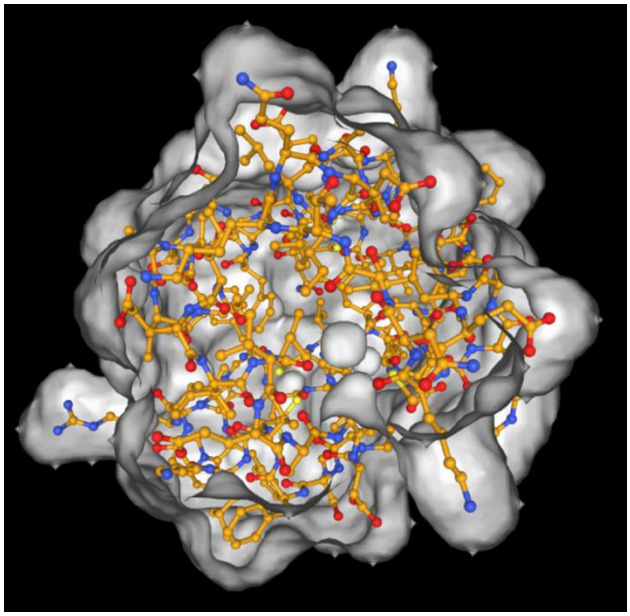


Fig. 5. Example 3D protein structure generation from PDB file.

4) Check DrugBank

The Check DrugBank tab is used to identify drug targets for specific genes using information from the DrugBank database (Fig. 6). By inputting a gene name, users can output the DrugBank IDs of different drugs by typing in the gene name and then clicking on the "Output Drug Targets" button. This tab allows users to find drug targets for each candidate gene, giving them information on the main panel that will display the drug IDs that are linked to the queried gene.

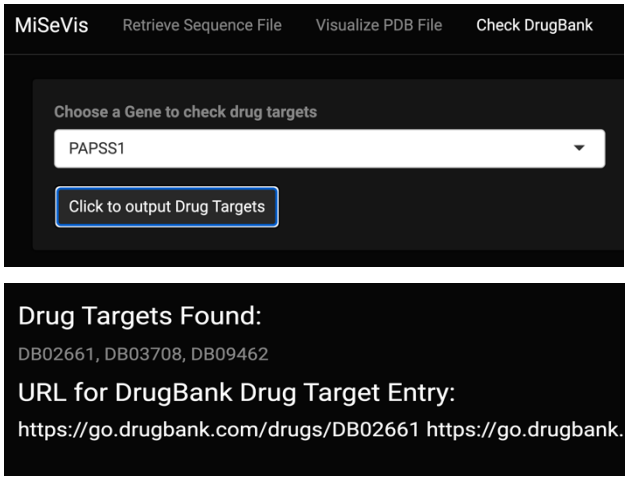


Fig. 6. Drug target retrieval using DrugBank.

5) Survival analysis

The survival analysis tab is designed to assess the effect of gene expression levels on survival rates. Users can perform this analysis simply by providing the gene name. The main panel will then display the corresponding survival plot for the selected gene. The survival plots generated will always include survival probability and time, as well as different lines for gene expression if available. These plots are automatically generated using the R survival library. For instance, Fig. 7 below depicts the overall survival probability over time for ovarian cancer patients with high and low PSAP gene expression.

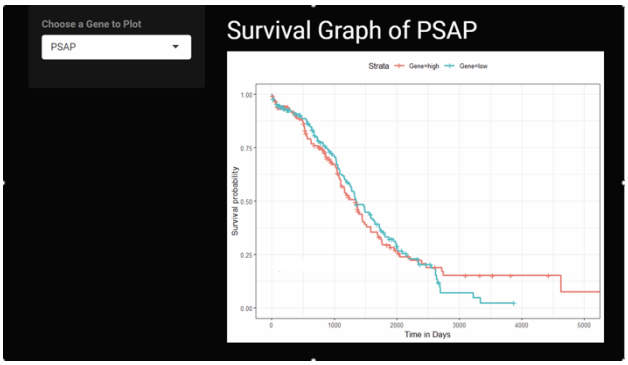


Fig. 7. Survival analysis of ovarian cancer patients based on PSAP gene expression.

6) Missense mutation analysis

Fig. 8 below shows missense mutations portrayed by the PDB file. The specified amino acid location, which is taken from cBioPortal, is found in the fifth column of the PDB file, and the two files show a missense mutation in that position. This missense mutation is reflected in the generated 3D protein structure graph, as the segment of the protein highlighted in the diagram is different between the wild-type and mutated protein. This difference also reflects a possible biomarker, as many patient samples from the ovarian cancer dataset of cBioPortal have this mutation.

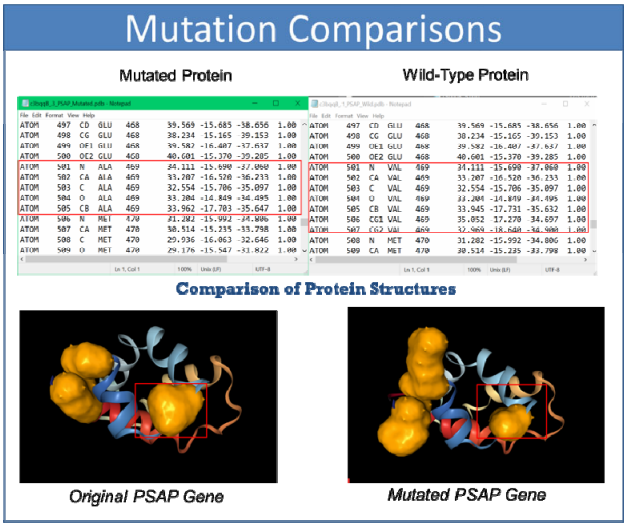


Fig. 8. Mutation comparison from PDB files and 3D protein structure for PSAP.

IV. Conclusion

This study implemented a user-friendly tool designed to identify drug targets for queried genes. The tool offers the capability to access the contents of both FASTA and PDB files, as well as the flexibility to upload the user's files as well. Additionally, this software includes a protein visualization function with which the user can choose a PDB file and visualize the 3D protein structure, which also shows some missense mutations that impact

the protein structure. Users can also choose between mutated and non-mutated files, exploring the exact locations of the mutations with various imaging options. Kaplan-Meier survival analysis was also conducted to evaluate the clinical implications of the candidate genes. All these analysis modules have been seamlessly integrated into the bioinformatics software called MiSeVis that was developed for this study.

Preliminary results have identified 26 genes with potential drug targets. Among them, several genes exhibited a significant impact on the survival rate of ovarian cancer patients when expressed at high levels. This indicates that the highly expressed genes with altered protein structures due to missense mutations may serve as valuable biomarkers for ovarian cancer treatment. Further functional annotation studies are underway to validate the association between these prioritized genes and ovarian cancer. Future work will involve comparing prediction results across different genders and racial groups. These findings have the potential to guide medical researchers in prioritizing drug targets and advancing treatment strategies for ovarian cancer. This gene prioritization approach applies to other types of cancer as well, and the pipeline will be expanded and enhanced in the future by adding more functionalities.

In summary, this study successfully identified genes with pathogenic implications in ovarian cancers through comprehensive bioinformatics analysis. Additionally, a user-friendly R-Shiny tool called MiSeVis was developed, enabling users to visualize 3D protein structure changes and identify potential therapeutic drug targets for specific genes. MisSeVis has the unique capability to visualize both wild-type and mutant protein structures concurrently, making it a valuable asset for identifying candidate genes resulting from variant-causing structural alterations. This tool holds the potential to guide medical researchers in determining the most effective treatment based on the precise alterations in the protein structures that make up a person's particular collection of biomarkers. Scientists can also use these specific biomarkers to develop drugs that specifically target them.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

IH performed bioinformatics data analysis and drafted the paper; YB developed the concept, supervised the research, reviewed, and provided critical revisions of the paper; both authors have revised and approved the final version of the article.

Acknowledgment

The authors wish to thank Dr. Yongsheng Bai for invaluable mentorship and unwavering guidance throughout this project. The findings presented in this paper are in whole or partly based upon data generated by the TCGA Research Network.

References

[1] M. Petrosino, L. Novak, A. Pasquo, *et al.*, "Analysis and interpretation of the impact of missense variants in cancer," *International Journal of Molecular Sciences*, vol. 22, no. 11, 5416, May 2021.

[2] T. Guo, X. Dong, S. Xie, *et al.*, "Cellular mechanism of gene mutations and potential therapeutic targets in ovarian cancer," *Cancer Management and Research*, vol. 13, pp. 3081–3100, Apr. 2021.

[3] C. Li, I. Hou, M. Ma, *et al.*, "Orthogonal analysis of variants in APOE gene using in-silico approaches reveals novel disrupting variants," *Frontiers in Bioinformatics*, vol. 3, 1122559, Apr. 2023.

[4] J. Jumper, R. Evans, A. Pritzel, *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, pp. 583–589, July 2021.

[5] L. Kelley, S. Mezulis, C. Yates, *et al.*, "The Phyre2 web portal for protein modeling, prediction and analysis," *Nature Protocols*, vol. 10, pp. 845–858, May 2015.

[6] K. Strimbu and J. A. Jorge, "What are biomarkers?" *Current Opinion in HIV and AIDS*, vol. 5, p. 463, Nov. 2010.

[7] B. Li, E. Severson, J. C. Pignon, *et al.*, "Comprehensive analyses of tumor immunity: Implications for cancer immunotherapy," *Genome Biology*, vol. 17, pp. 1–16, Aug. 2016.

[8] I. Vázquez-García, F. Uhlitz, N. Ceglia, *et al.*, "Ovarian cancer mutational processes drive site-specific immune evasion," *Nature*, vol. 612, 7941, Dec. 2022.

[9] B. Vogelstein, N. Papadopoulos, and V. E. Velculescu, "Cancer genome landscapes," *Science*, vol. 339, pp. 1546–1558, Mar. 2013.