SCREENING OF OVEREXPRESSED HUMAN ENZYMES IN TRIPLE NEGATIVE BREAST CANCER FOR IDENTIFYING POTENTIAL BIOMARKERS

Project Reference No.: 48S_BE_4559

College : BMS COLLEGE OF ENGINEERING
Branch : Department of Biotechnology
Guide(s) : Prof. Divijendranatha Reddy Sirigiri
Student(s): Ms. Aishwarya HS
Mr. Anjan P Kumar
Ms. Asna Zainab
Ms. Chetana V

Keywords:

Breast cancer, TNBC, Multi-omics, Molecular docking, STRING analysis, Biomarker, Gene Ontology, Metabolomics.

Introduction:

Triple-negative breast cancer (TNBC) is a highly aggressive subtype of breast cancer characterized by the absence of estrogen, progesterone, and HER2 receptors. This unique molecular profile makes TNBC resistant to conventional hormonal or targeted therapies, posing significant challenges for effective treatment. As a result, TNBC has become a critical area of focus in oncology research, necessitating innovative approaches to improve patient outcomes. The urgency to address TNBC stems from its poor prognosis, higher recurrence rates, and limited therapeutic options compared to other breast cancer subtypes.

This project aims to deepen the understanding of molecular mechanisms underlying TNBC by identifying and studying overexpressed enzymes associated with this cancer type. Enzymes play pivotal roles in cellular processes, and their dysregulation often drives cancer progression and survival. By targeting these enzymes as biomarkers, the research opens new avenues for precision medicine—designing drugs that inhibit enzyme activity to disrupt critical pathways involved in TNBC growth and proliferation. Such targeted interventions hold promise for reducing side effects while enhancing therapeutic efficacy.

Advanced bioinformatics tools, such as UALCAN, are integral to this study, enabling high-throughput screening of genetic and proteomic data. These tools facilitate the identification of specific enzymes linked to TNBC progression and highlight potential therapeutic targets. The integration of bioinformatics with molecular biology represents a significant advancement in cancer research, paving the way for systematic biomarker discovery and personalized treatment strategies. Furthermore, innovations in enzyme screening methodologies may extend to other cancers and diseases, amplifying the impact across biotechnology and medical fields. By addressing TNBC's challenges through cutting-edge research, this project contributes to the global effort to combat aggressive cancers and improve public health outcomes.

Objectives:

1. Conduct a comprehensive screening to identify human enzymes overexpressed in triple-negative breast cancer (TNBC) compared to normal breast cells at both protein and mRNA levels
2. Incorporating string analysis for protein-protein interaction mapping between the shortlisted genes.
3. Investigate the functional roles of these overexpressed enzymes in the progression, survival, and metastasis of TNBC cells.
4. Perform KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway analysis to elucidate the biological pathways associated with the identified enzymes and GO (Gene Ontology) analysis to categorize their molecular functions, biological processes, and cellular components.
5. Develop and validate diagnostic tools and biomarkers based on the overexpressed enzymes for early detection, prognosis, and monitoring of TNBC.

Methodology:

This project employs a multidisciplinary approach to identify and validate enzymes overexpressed in triple-negative breast cancer (TNBC) using bioinformatics, molecular biology, and computational tools. Below is a detailed description of the methods, techniques, and materials used:

1. Data Sources and Gene Annotation

1) HGNC: Data Source

The Human Gene Nomenclature Committee (HGNC) database was utilized to extract a comprehensive list of human enzymes. HGNC provides standardized gene symbols, annotations, and functional insights essential for understanding enzymatic roles in cellular and pathological contexts.

2) SynGo

SynGo was employed to convert gene symbols into full gene names for better interpretation of their biological functions. For example, the gene symbol "BRCA1" corresponds to "Breast Cancer 1," which provides detailed information about its role in TNBC.

2. Bioinformatics Analysis

1) UALCAN: Proteomics and Genomics Analysis

UALCAN, an interactive web portal, was used to screen over 3,500 enzymes for differential expression between TNBC tumor samples and normal tissues. The tool leverages data from The Cancer Genome Atlas (TCGA) to identify over- or under-expressed genes and assess their association with patient survival rates.

2) SurvExpress: Survival Analysis

SurvExpress was employed for multi-gene biomarker validation and survival analysis across TNBC datasets. It provides risk assessments based on gene expression signatures, enabling rapid identification of prognostic markers.

3) STRING: Protein-Protein Interaction Analysis

STRING database was used to analyze protein-protein interactions (PPIs) among identified enzymes. Functional enrichment analyses helped visualize interaction networks and pinpoint key proteins involved in TNBC progression.

3. Structural and Druggability Assessment

1) AlphaFold: 3D Structure Prediction

AlphaFold was used to predict the three-dimensional structures of overexpressed enzymes. This computational tool enhances the understanding of protein function and aids in drug design by identifying active sites.

2) **CanSAR: Druggability Assessment**

CanSAR integrated molecular profiling data to evaluate the druggability of identified enzymes. It provided insights into ligand-binding potential and chemical bioactivity for targeted drug discovery.

4. Ligand Discovery and Docking Studies

1) **ZINC15 Database**

ZINC15 facilitated the discovery of potential ligands by providing access to over 120 million purchasable "drug-like" compounds. Ligands were selected based on their compatibility with enzyme active sites.

2) **MTiAutoDock/MTiOpenScreen**

Docking studies were performed using MTiAutoDock and MTiOpenScreen tools to identify small molecules that bind effectively to enzyme targets. Binding poses and energies were analyzed for top-ranked ligands.

3) **CB-Dock: Validation**

CB-Dock validated docking results by predicting binding sites on enzymes using curvature-based cavity detection methods. Interactive visualizations enhanced understanding of ligand-binding modes.

5. Functional Annotation

1) **DAVID: Gene Ontology Analysis**

DAVID bioinformatics tool was employed for functional annotation of identified enzymes, focusing on their roles in biological pathways relevant to TNBC.

2) **KEGG Pathway Analysis**

KEGG database was used to map molecular interactions and pathways involving the overexpressed enzymes, providing insights into their roles in TNBC pathogenesis.

6. Experimental Validation

1) **RNA Isolation Protocol**

RNA isolation was performed using TRIzol reagent to extract high-quality RNA from TNBC cell lines. The protocol included phase separation, precipitation, washing, and resuspension steps to ensure RNA purity (260/280 absorbance ratio > 1.8).

2) cDNA Synthesis

Complementary DNA (cDNA) synthesis was carried out using reverse transcription methods with oligo(dT) primers. The cDNA served as a stable template for downstream applications like qPCR.

7. Computational Tools for Visualization

1) Google Colab

Google Colab was used for dynamic visualization of bioinformatics results, integrating Python libraries such as Matplotlib and Seaborn for clear graphical representation.

8. Literature Review

1) PubMed Database

PubMed was extensively searched for relevant literature on TNBC biomarkers, enzyme functions, and therapeutic strategies to contextualize findings within existing research.

9. Statistical Analysis

1) MetaboAnalyst Tool

MetaboAnalyst software facilitated statistical analysis of metabolomic data associated with TNBC biomarkers, enabling pathway mapping and identification of potential therapeutic targets. This comprehensive methodology integrates advanced computational tools with experimental validation techniques to identify enzymes involved in TNBC progression while exploring their therapeutic potential through drug discovery approaches. Visual aids such as pathway diagrams from KEGG or PPI networks from STRING can be included in reports or presentations for enhanced understanding.

Result and Conclusion:

1. **Identification of Overexpressed Genes**:

   A list of 3,600 enzyme-coding genes from the HGNC database was analyzed using TCGA data, identifying 30 genes significantly overexpressed in breast cancer samples (p-value < 0.0001).

   TNBC (Triple-Negative Breast Cancer) subtype exhibited the highest expression levels among breast cancer subtypes.

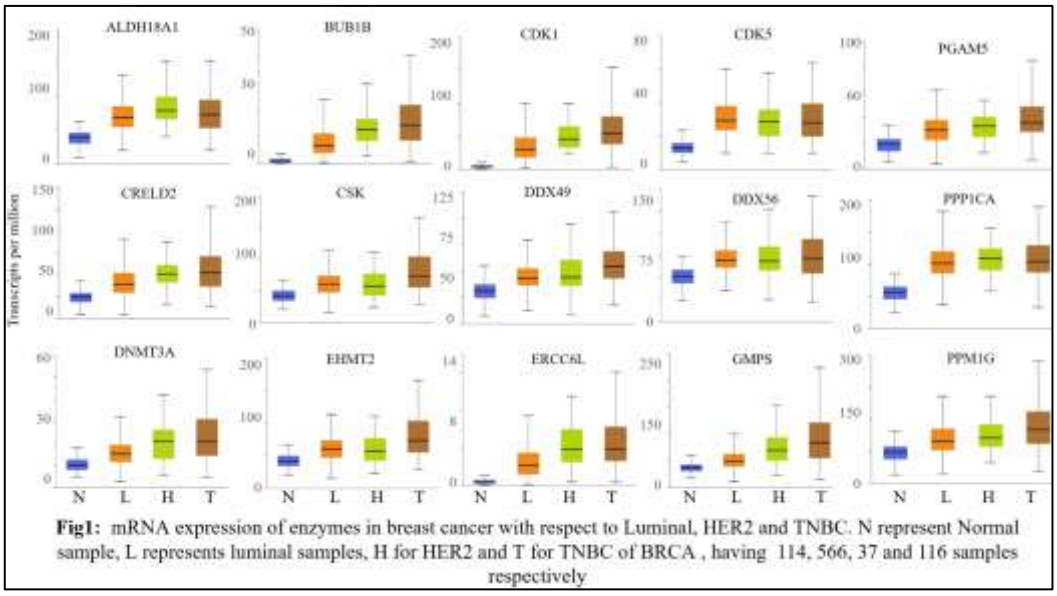


Fig1: mRNA expression of enzymes in breast cancer with respect to Luminal, HER2 and TNBC. N represent Normal sample, L represents luminal samples, H for HER2 and T for TNBC of BRCA , having 114, 566, 37 and 116 samples respectively

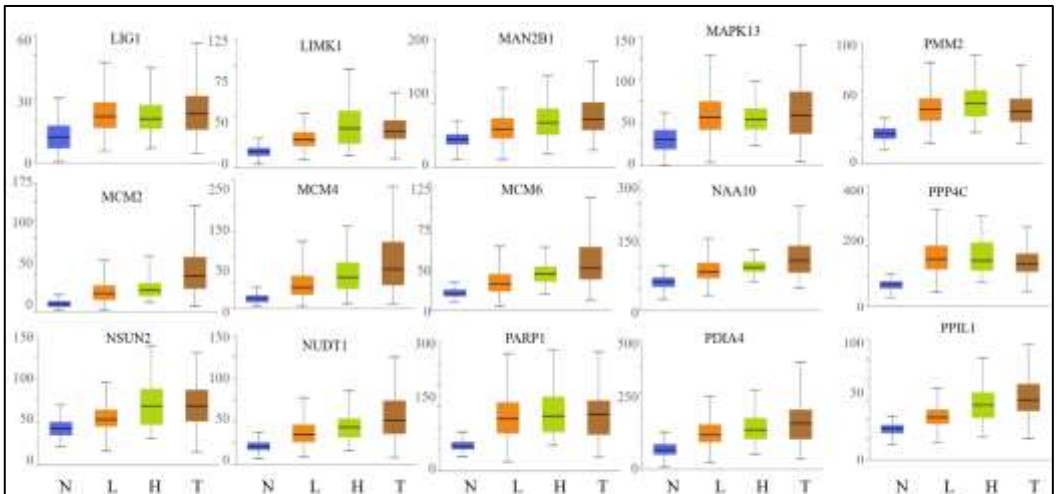


Fig1a: mRNA expression of enzymes in breast cancer with respect to Luminal, HER2 and TNBC. N represent Normal sample, L represents luminal samples, H for HER2 and T for TNBC of BRCA , having 114, 566, 37 and 116 samples respectively

2. **Proteomics Validation**:

   Proteomic analysis confirmed that mRNA overexpression translated to higher protein levels, validating the significance of these genes.

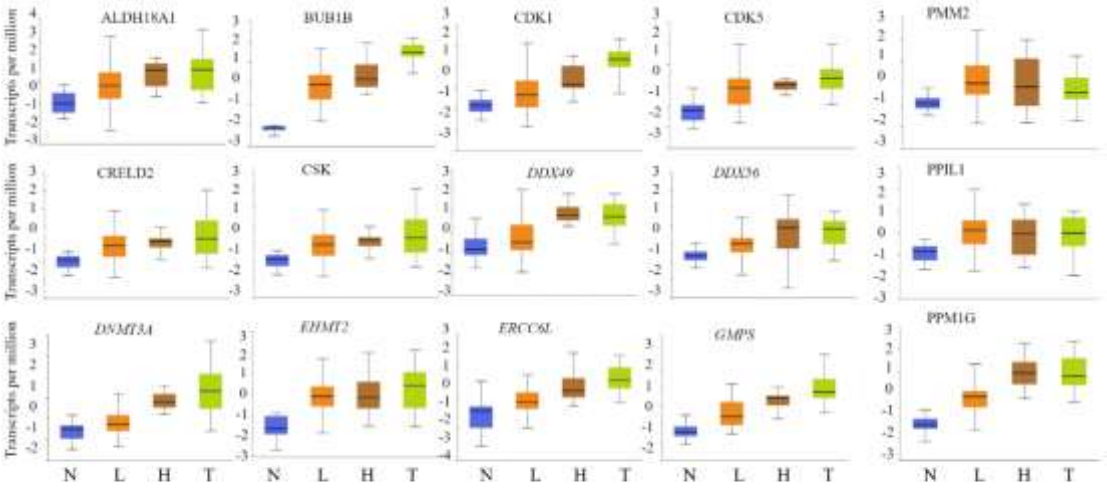


Fig2: Protein expression of enzymes in breast cancer with respect to Luminal, HER2 and TNBC. N represent Normal sample, L represents luminal samples, H for HER2 and T for TNBC of BRCA , having 18, 64, 10 and 16 samples respectively

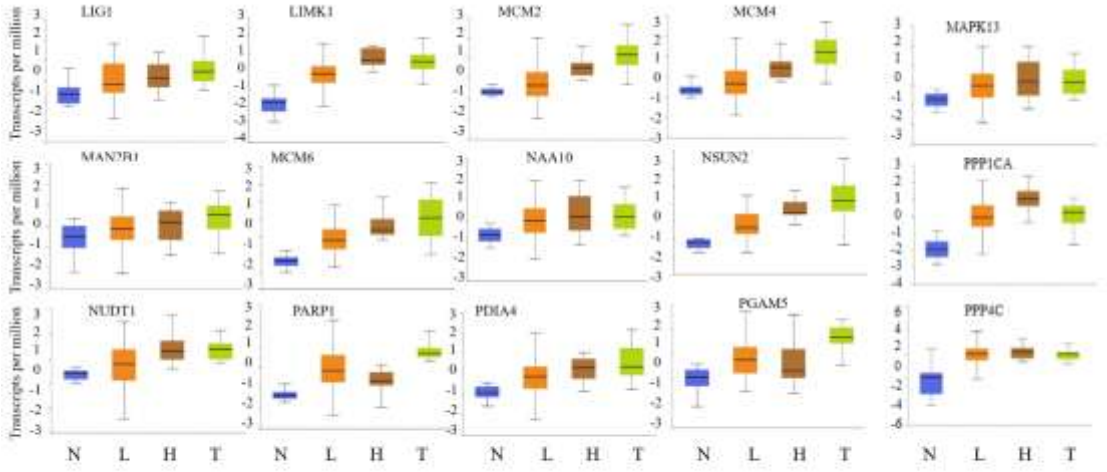


Fig2a: Protein expression of enzymes in breast cancer with respect to Luminal, HER2 and TNBC. N represent Normal sample, L represents luminal samples, H for HER2 and T for TNBC of BRCA , having 18, 64, 10 and 16 samples respectively

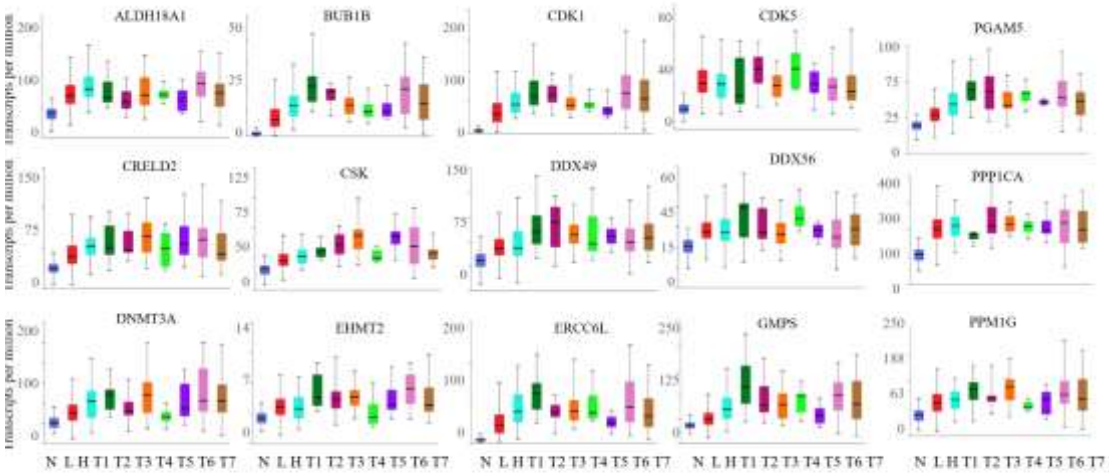


Fig3: mRNA expression of enzymes in breast cancer with respect to TNBC and its substages, N represent Normal sample, L represents luminal samples, H for HER2 and T1-T7 for subtypes of TNBC , having 114, 566, 37, 13, 11, 20, 8, 8, 29 and 27 samples respectively

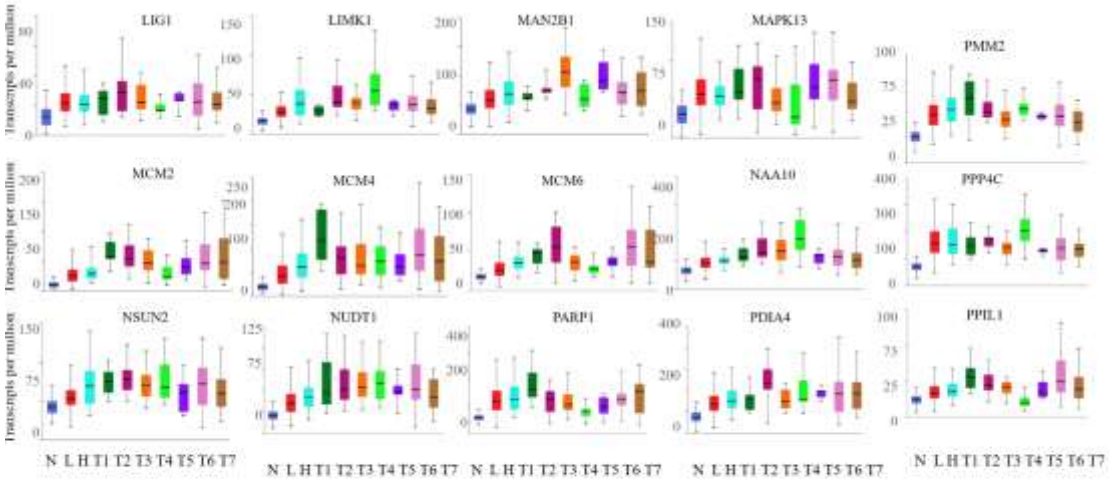


Fig3a: mRNA expression of enzymes in breast cancer with respect to TNBC and its substages, N represent Normal sample, L represents luminal samples, H for HER2 and T1-T7 for subtypes of TNBC , having 114, 566, 37, 13, 11, 20, 8, 8, 29 and 27 samples respectively

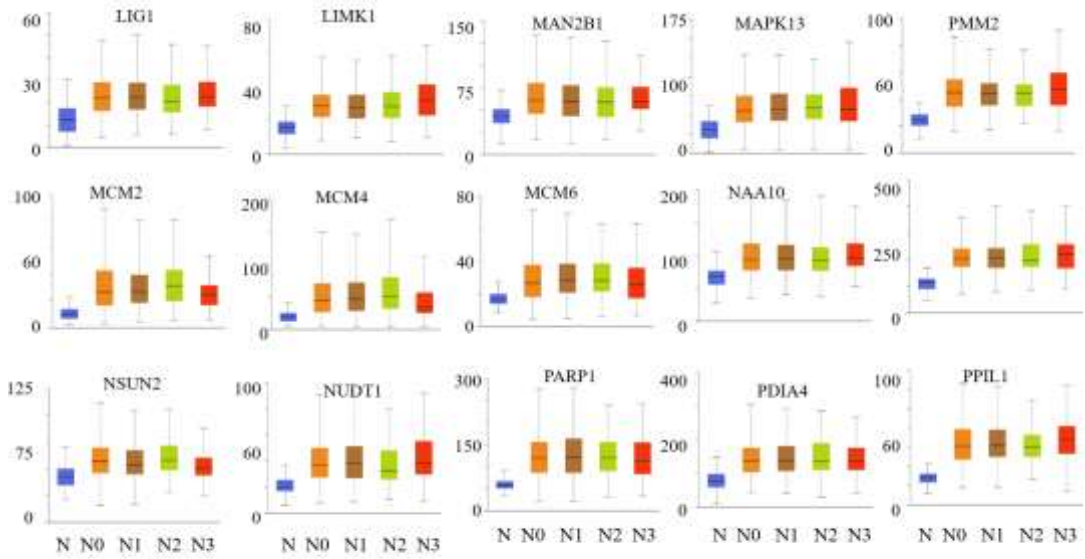


Fig4a mRNA expression of enzymes in breast cancer with respect to Luminal, HER2 and TNBC. N represent Normal sample, N0-N3 represents different nodal stages of BRCA , having 114, 516, 362, 120 and 77 samples respectively

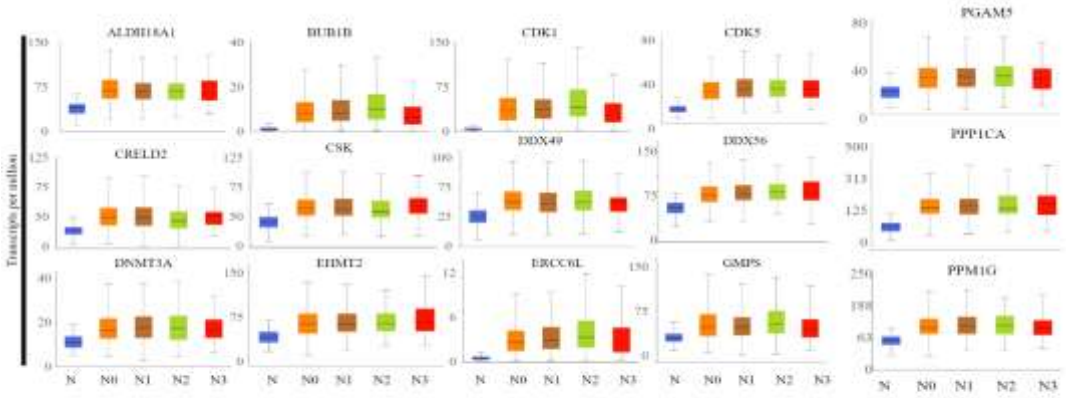


Fig4 mRNA expression of enzymes in breast cancer with respect to Luminal, HER2 and TNBC. N represent Normal sample, N0-N3 represents different nodal stages of BRCA , having 114, 516, 362, 120 and 77 samples respectively

1) **Survival Analysis**:

(a) Genes like PARP1 and CDK1 showed strong statistical significance, correlating higher expression with shorter survival times.

(b) Genes such as ERCC6L and MCM4 also demonstrated potential links to poor survival outcomes.

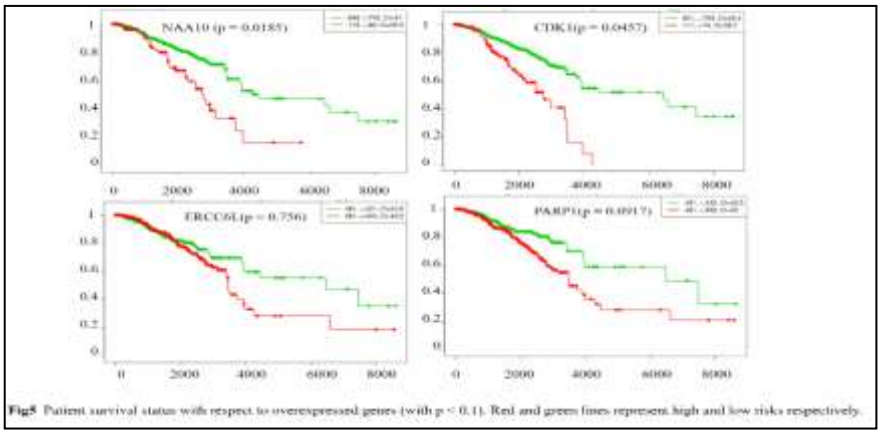


Fig5 Patient survival status with respect to overexpressed genes (with $p < 0.1$). Red and green lines represent high and low risks respectively.

2) **Protein-Protein Interaction Analysis**:

(a) High connectivity was observed among proteins like MCM2, MCM4, and MCM6, suggesting their involvement in related biological processes.

(b) Independent proteins like MAPK13 and CDK5 may function separately from other identified genes.

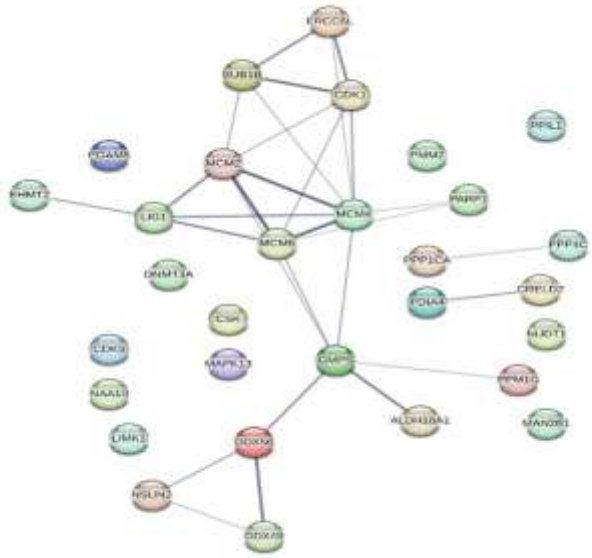
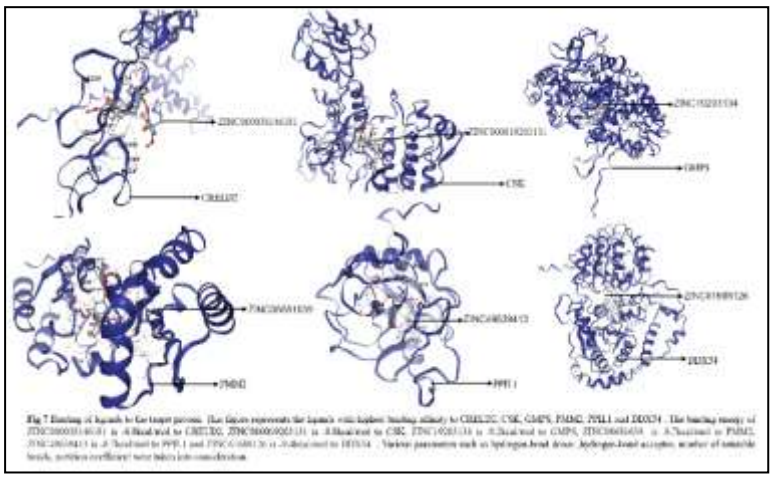


Fig6 Protein-protein Interactions between the shortlisted enzymes

3) **Molecular Docking**:

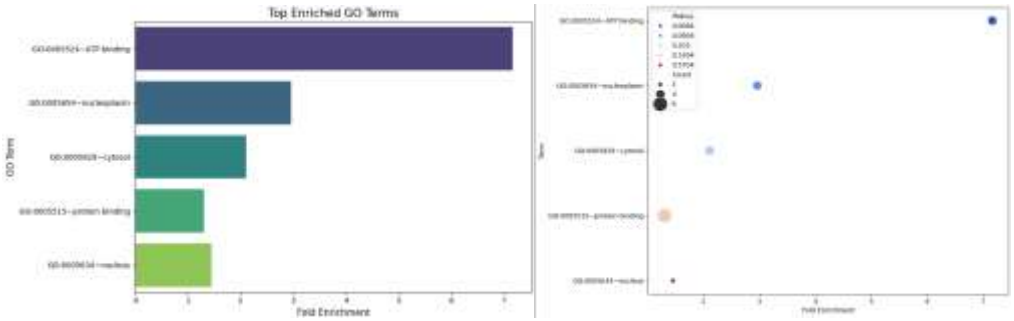
(a) Novel druggable targets (e.g., CRELD2, CSK, GMPS) were analyzed with 136 natural ligands. Promising candidates were identified based on binding energy and molecular properties.



4) **Gene Ontology (GO) and KEGG Pathway Analysis**:

(a) GO annotations highlighted roles in ATP binding, nucleoplasm localization, and protein binding.

Category	Term	Count	%	PValue	Fold Enrichment
GOTERM_MF_DIRECT	GO:0005524~ATP binding	4	57.14%	0.0084	7.16
GOTERM_CC_DIRECT	GO:0005654~nucleoplasm	4	57.14%	0.0904	2.96
GOTERM_CC_DIRECT	GO:0005829~cytosol	4	57.14%	0.203	2.11
GOTERM_MF_DIRECT	GO:0005515~protein binding	6	85.71%	0.3394	1.3
GOTERM_CC_DIRECT	GO:0005634~nucleus	3	42.86%	0.5704	1.45

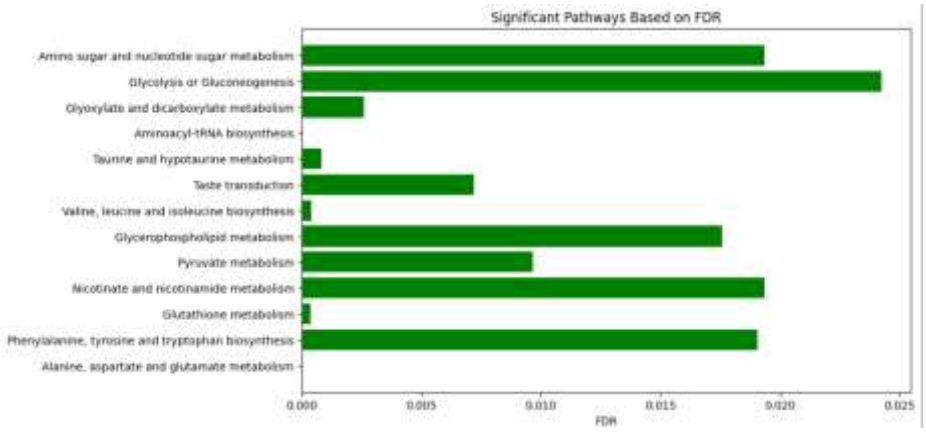


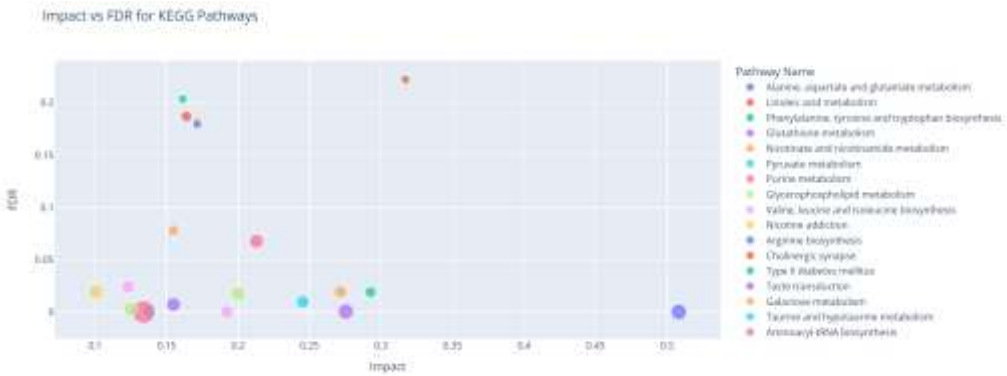
(b) KEGG pathway analysis revealed critical involvement in metabolic pathways (e.g., purine metabolism, nucleotide sugar biosynthesis), emphasizing their importance in TNBC cell proliferation.

Pathway Name	KEGG Pathway ID	Gene	Role/Function
Metabolic pathways	hsa01100	PMM2, GMPS	Central metabolic processes
Nucleotide metabolism	hsa01232	GMPS	Purine and pyrimidine synthesis
Biosynthesis of nucleotide sugars	hsa01250	PMM2	Formation of nucleotide sugar precursors
Biosynthesis of cofactors	hsa01240	PMM2	Cofactor synthesis pathways
Fructose and mannose metabolism	hsa00051	PMM2	Carbohydrate metabolism
Amino sugar and nucleotide sugar metabolism	hsa00520	PMM2	Synthesis of sugar derivatives
Purine metabolism	hsa00230	GMPS	Purine nucleotide synthesis and recycling
Drug metabolism - other enzymes	hsa00983	GMPS	Metabolism of xenobiotics
Spliceosome	hsa03040	PPIL1	RNA splicing and processing
Epithelial cell signaling in H. pylori	hsa05120	CSK	Signal transduction related to cell regulation

5) **Metabolomics Studies**:

(a) Pathway analysis identified significant pathways like amino sugar metabolism and purine metabolism, highlighting PMM2 and GMPS as key contributors to cancer cell growth and survival.





In conclusion, the study successfully identified 30 overexpressed genes in breast cancer, particularly TNBC subtype, which are involved in critical metabolic and signaling pathways. Proteomic validation strengthens their role as potential biomarkers

or therapeutic targets. Survival analysis highlights PARP1 and CDK1 as key predictors of patient prognosis. Molecular docking and pathway enrichment analyses suggest promising drug targets for TNBC treatment. CRISPR analysis provides insights into gene-editing strategies for functional studies or therapeutic applications.

Future Scope:

The future scope of this project includes:

1. Perform CRISPR Analysis on the shortlisted genes.
2. Identify the primers of the shortlisted genes.
3. Check for the expression levels in cancer cell lines

   RT-qPCR helps in quantify the expression levels of genes involved in the pathways identified during computational analysis. Western Blotting can be done to confirm changes in protein expression for key targets, providing direct evidence of pathway involvement.
4. Biomarker Verification: Immunohistochemistry (IHC): Detect and quantify biomarkers in biological samples, validating their potential as diagnostic or therapeutic targets.
5. Functional and Phenotypic Studies: Conduct enzyme assays to measure the activity of enzymes implicated in your study, verifying their role in the observed biological processes.