Pakistan Journal of Medical & Cardiological Review
https://pakjmcr.com/index.php/1/about
Online ISSN 3007-2387
Print ISSN 3007-2379

Computational Framework For Identifying Hub Genes And Proteins Through Integrative Network Modeling And Machine Learning

Mishal Khalid*
Department of Computer Science and Information Technology, Superior University Lahore, Pakistan Email: mishalkhalid912@gmail.com

Hummaira Iqbal
Department of Zoology, University of Sargodha, Pakistan

Hafiz Muhammad Umer Jahangir
Department of Computer Science and Information Technology, Superior University Lahore, Pakistan

Muhammad Ayaz
Department of Computer Science and Information Technology, Superior University Lahore, Pakistan

Asad Munir
Department of Biological Sciences, Superior University Lahore, Pakistan

Author Details

Keywords: Hub Genes, Predative Modeling, Breast Cancer, Network Modelling, Machine Learning

Received on 15 Feb 2026
Accepted on 20 Mar 2026
Published on 31 Mar 2026

Corresponding E-mail & Author*:

Mishal Khalid*
Department of Computer Science and Information Technology, Superior University Lahore, Pakistan Email: mishalkhalid912@gmail.com

Abstract

Breast cancer is a complex disease involving dysregulation of multiple genes, proteins, signaling pathways, and cellular processes. The present study employed an integrative computational systems biology approach to identify differentially expressed genes (DEGs), central hub genes, associated biological pathways, and potential diagnostic biomarkers in breast cancer. Publicly available transcriptomic data comprising breast cancer and normal control samples were analyzed using differential expression analysis, protein–protein interaction (PPI) network construction, network topology, functional enrichment, and receiver operating characteristic (ROC) analysis. Among 19,845 analyzed genes, 756 were significantly upregulated and 706 were significantly downregulated based on $|\log_2 FC| \geq 1$ and adjusted $P < 0.05$. EGFR, VEGFA, MMP9, STAT3, and CXCL8 were among the most significantly upregulated genes, whereas BRCA1, APC, CDH1, RB1, and PTEN was markedly downregulated. The PPI network comprised 30 protein nodes and 182 interactions, with significant enrichment ($P < 1.0 \times 10^{-16}$). Network analysis identified EGFR as the highest-ranked hub gene, followed by STAT3, PIK3CA, MYC, and TP53. Gene Ontology analysis indicated enrichment in immune response, cell proliferation, angiogenesis, receptor tyrosine kinase signaling, and apoptosis-related processes. KEGG analysis highlighted cancer pathways, PI3K–Akt, MAPK, cytokine–cytokine receptor interaction, focal adhesion, TNF signaling, and apoptosis. ROC analysis demonstrated strong diagnostic performance for EGFR

(AUC = 0.912), STAT3 (AUC = 0.886), and PIK3CA (AUC = 0.858). Overall, the integrated computational framework identified biologically important hub genes and pathways that may serve as potential biomarkers and therapeutic targets in breast cancer. Experimental validation and independent patient cohorts are required before clinical application.

Introduction

Recent developments of high-throughput sequencing technologies and computational biology have revolutionized biomedical research, providing access to a comprehensive study of biological systems at the genomic, transcriptomic, proteomic, metabolomic and epigenomic levels. These technological advancements have generated an unprecedented amount of biological data that will give researchers a chance to study complex molecular mechanisms of human diseases. In contrast to the gene or protein centric view of modern systems biology deals with the complex interplay between biological molecules which together coordinate the function of cells. Thus, the development of information on the critical molecular factors involved in disease initiation, progression, and response to treatment has become one of the main challenges in the field of computational and translational biology [1].

Complex diseases, such as cancer, cardiovascular diseases, neurodegenerative diseases, autoimmune diseases, and metabolic syndromes, are caused by the interplay of the dysregulated expression of multiple genes, proteins, signaling pathways, and biological processes. These multifactorial diseases are different from monogenic diseases where mutations in a single gene cause the disorder, in that these diseases involve complex molecular interactions which cannot be fully described by the reductionist approach [2]. Thus, the need for integrative computational methods that can handle the analysis of biological systems as networks and include the most sophisticated statistical and machine learning approaches to uncover the most biologically meaningful molecular signatures is growing [3].

The development of publicly available biological databases has greatly stimulated computational research in molecular biology. Several resources, such as the Gene Expression Omnibus (GEO), The Cancer Genome Atlas (TCGA), ArrayExpress, STRING, Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), and Reactome, offer huge amounts of data related to gene expression profiles, protein-protein interactions, and functional annotations. These databases allow for big data integrative analyses without expensive lab experiments. Computational studies based on an open-source collection of data have gained significance in their ability to yield biologically relevant hypotheses, disease markers and therapeutic targets for experimental tests [3].

Differential gene expression analysis is one of the different computational approaches used to identify genes which are expressed differently between normal and disease conditions and is the first step in identifying such genes. Differentially expressed genes (DEGs) are useful for understanding molecular changes that are involved in the progression and development of disease. But genes don't usually operate alone in a biological system. Instead, they are involved in very tightly coupled networks of proteins interactions, transcriptional regulation, signaling pathways and metabolic pathways. Thus, conclusions based on changes in fold expression or the ratio of significance alone could miss the identification of genes which are central regulatory genes with relatively small fold changes in expression [4]

To overcome these problems, network biology is becoming an emerging field to investigate biological complexity. Biological networks are a collection of nodes representing genes, proteins or metabolites linked by different forms of molecular interactions. These networks allow to study system-wide relationships which can't be seen if the system is studied as a collection of individual molecules. In a network, some nodes have more connections with other nodes and play a more strategic role

that participate in maintaining the stability of the network and the coordination of the essential functions of the cell. They are often called hub genes or hub proteins, and they are highly connected, or influential [5].

Hub genes are highly involved in control of various biological processes such as cell proliferation, apoptosis, immune responses, DNA repair, metabolism, differentiation and signal transduction. Hub genes are important because they are in the middle of molecular interaction networks and alteration of these genes often has a cascading effect down the network with broad downstream effects that play a role in pathogenesis of disease [6]. Many studies have shown that hub genes are more likely to be used as diagnostic biomarkers, prognostic indicators and therapeutic targets than genes which are selected only based on differential expression analysis. Therefore, the identification of hub genes has been one of the main focuses in computational systems biology [7].

One of the most popular methods for the identification of hub proteins and elucidation of the molecular mechanisms of disease is the protein-protein interaction (PPI) network. Proteins do not always work alone, but they are often in complex with many partners and through highly organized biological complexes and signaling pathways to perform their cellular functions. These interactions and proteins that are in the middle of the interactome can be easily studied using PPI networks [8]. STRING databases include experimentally determined and computationally predicted protein interactions and can be used to build protein interaction networks that represent functional protein relationships. These networks can then be analyzed topologically to detect proteins that are highly connected, central and have a regulatory influence [9].

Network analysis is a mathematical method which uses graph theory to measure the structural importance of nodes by using topological parameters. To assess the impact of genes and proteins in biological networks, several centrality measures have been proposed. Degree centrality is the count of the number of direct interactions connected to the node and determines highly connected molecules [10]. Betweenness centrality measures the extent to which the node is a bridge between regions in the network, thus identifying molecules playing a role in information flow. The closeness centrality reflects how efficiently a node communicates with all other nodes, while eigenvector centrality reflects the influence of a node as a function of the importance of its neighboring nodes.

Some other parameters such as maximal clique centrality (MCC), density of maximum neighborhood component (DMNC), edge percolated component (EPC), radiality, and stress centrality are used to capture different aspects in assessing biologically important hub molecules. Multiple topological methods can be combined to enhance robustness and reliability of hub gene selection [11].

Network topology provides important insight into the importance of molecules, however, using network-based measures alone has its drawbacks. Rankings often vary among different centrality measures and network topology might not reflect the biological significance of candidate genes. Additionally, biological data is often high dimension, noisy, replicates vary across experiments, and relationships are often nonlinear with no proper mathematical model. The challenges have led to the incorporation of machine learning techniques in computational biology to improve feature selection, predictive modelling, and biomarker discovery [12].

The power to detect complex patterns in the data has made machine learning one of the most impactful technologies in biomedical data analysis. Machine learning methods learn directly from data without any prior assumption of what the data may contain, and they can identify what patterns are informative and create predictive models. Supervised learning techniques such as Random Forest, Support Vector Machine (SVM), Logistic Regression, Decision Trees, Gradient Boosting and Artificial Neural Networks are extensively used for disease classification, biomarker identification, patient stratification and prediction of outcomes. Some unsupervised

learning methods like clustering and dimensionality reduction methods enable the identification of new unknown molecular subtypes and biological relationships [13].

One of the most crucial uses of machine learning in omics studies is feature selection. In high-dimensional data, there can be thousands of genes and only a few biological samples, leading to potential overfitting and model interpretability issues. Feature selection algorithms using machine learning find the most informative genes for classifying the disease and remove redundant and irrelevant features. Random Forest is one of the latter that is particularly appealing due to its ability to assess variable importance by using ensemble learning, its ability to deal with nonlinear relationships, and its robustness to noisy biological data. Combining machine learning with network methods allows the identification of genes with central roles in the network of interactions as well as those that are very predictive of disease status [14].

In addition to identifying hub genes, functional enrichment analysis also gives important biological interpretation of the computational results. By performing Gene Ontology (GO) enrichment, the biological roles that are associated with gene sets identified can be revealed, which classifies genes into biological processes, molecular functions, and cellular components. Likewise, pathway enrichment analysis of the candidate genes based on KEGG, Reactome, and other databases shows that there are significant pathways in which candidate genes are enriched. Such analyses help to understand the molecular mechanisms of disease and can help to identify therapeutic targets in cellular pathways. Computational predictions can be greatly enhanced in biological significance by incorporating network topology and functional enrichment [15].

In recent years, the development of computational tools in the field of biology has given more weight to multi-step integrative analytical approaches that integrate differential expression analysis, protein interaction networks, topological analysis, machine learning, and functional enrichment in a single analytical flow. These integrative methods utilize the advantages of separate analytical techniques but overcome the individual disadvantages. Differential expression discovers disease-related genes, network analysis is used to analyze molecular interaction, topological algorithms prioritize influential nodes, machine learning validates informative features, and enrichment analysis elucidates biological functions. All these complementary approaches form a comprehensive approach to the discovery of strong hub genes and proteins in complex diseases [16].

Although great strides have been made in computational systems biology, there are many challenges to address. Experimental datasets from various sources can be highly heterogeneous, contain missing data, include batch effects, and have technical variation. Moreover, different computational algorithms often yield conflicting answers, and it is hard to distinguish between the most biologically relevant candidate genes. Variability also occurs from one independent study to another due to the lack of standard analytical pipelines. As a result, there is a growing demand for the development of more holistic computational tools, which combine several analytical methods to enhance the reproducibility, accuracy, and biological reliability of the tools [17].

Another factor to consider is the significance of the computational findings in translation. Numerous candidate biomarkers are proposed by computational analysis but only a small number of them can be successfully validated in experiments and implemented in the clinic. Combining several independent analytical approaches increases confidence in the hub genes in the candidates identified as molecules consistently prioritized in several computational views. These powerful candidate genes are more likely to represent biologically meaningful biomarkers for downstream experimental validation, drug discovery, and precision medicine applications [18].

With the advent of artificial intelligence and powerful computational modeling, biology systems have now been extended into so many areas. Heterogeneous data can be combined, complex biological networks can be built, relationships between molecules can be identified, and predictive models can be developed with increased accuracy using modern computational frameworks. These advances have greatly expedited the identification of biomarkers and have also decreased the time and expense of conventional laboratory clinical research. With the increasing power of computation, integrative bioinformatics has the potential to become increasingly important in personalized medicine for aiding in the diagnosis, prognosis, identification of therapeutic targets and drug repurpose in disease [19].

Considering the above, the integrative network based computational approach is used in the present study for the identification of hub genes and proteins using differential expression analysis, protein-protein interaction network construction, network topology analysis, machine learning based feature selection, and functional enrichment analysis. The study aims to select more confidently than might be possible with individual analytical methods alone, molecular targets that are of biological significance [20]. The application of network biology, computational modelling and machine learning offers a holistic approach to the study of complex molecular interactions involved in disease and will play an expanding role in systems biology. Finally, this study will be used to uncover strong hub genes and proteins that could be used as markers for diagnosis, targets for drug development, and candidates for further experimental and clinical research [21].

MATERIALS AND METHODS

Research design

In this study, high throughput data from various omics will be used in an integrative computational biology framework to identify hub genes and proteins. The workflow includes differential expression analysis, network construction, network topological analysis, machine learning-based feature selection, and functional enrichment to prioritize important targets with biological significance. The detection of proteins serving as hubs in a network is also a commonly used technique in the study of protein interaction and machine learning can complement feature prioritization by identifying the most informative proteins.

Data collection

The selected disease or biological condition will be chosen to search for public databases (e.g. GEO or ArrayExpress) for publicly available gene expression data. Additional omics layers e.g. DNA methylation, miRNA expression or proteomics, if available, can be added to facilitate integrative analysis. Clear case-control design should be applied, and selection of datasets should consider sample size, platform quality, and biological relevance.

Data preprocessing

Raw data will be quality checked and normalized prior to analysis. Microarray data will be background corrected, normalized and annotated for probes and RNA-seq data will be low count filtered, normalized and transformed. Multiple datasets should be integrated with correction for batch effects as integration across studies can lead to false signals.

Differential expression analysis

These genes that are differentially expressed between cases and controls will be selected by applying statistical testing, with multiple-testing correction. The criteria could involve using a cutoff for absolute log2 fold change, e.g., 1.0, and/or a cutoff

for adjusted p value, e.g., < 0.05. This will result in a DEG list which will be used as input for downstream network and machine learning.

Network construction

DEG products will be mapped to existing protein-protein interaction databases (STRING, BioGRID or IntAct) to construct a protein-protein interaction network. A multiplex or integrative network can be built from multi-omics data to encode relationships between layers of expression and other molecular data in the study. Network analysis is particularly valuable because hub genes are typically identified as central genes in scale-free biological networks.

Hub gene identification

Topological measures including degree, betweenness centrality, closeness centrality and maximal clique centrality will be used to identify hub genes. Genes with similar high centrality scores in various metrics will be identified as potential hubs. Modules can also be identified with methods like MCODE and genes that occur in both highly connected modules and the central network positions can be considered as more reliable hub genes.

Machine learning-based prioritization

Machine learning feature-selection methods will be used to further narrow down the hub candidate list. Random forest importance, support vector machine-recursive feature elimination, LASSO, elastic net or correlation-based feature selection are suitable approaches. Then a classifier can be trained with selected genes/proteins to compare the capability to distinguish case from control sample, and accuracy, AUC, Precision, Recall, F1 score and Matthews's correlation coefficient can be used to measure the performance.

Functional enrichment analysis

The final hub genes and proteins will be subjected to Gene Ontology and pathway enrichment analysis using databases such as KEGG, Reactome, or GO. This step helps determine whether the selected hubs participate in biologically meaningful processes such as cell cycle regulation, immune signaling, apoptosis, or metabolic control. Enrichment analysis strengthens the biological interpretation of the computational results.

Validation strategy

Validation will be performed at multiple levels. First, the hub list can be checked against independent datasets or cross-validation results. Second, literature mining can confirm whether the candidates have prior disease associations. Third, ROC analysis and classification performance can be used to assess whether the final gene set has predictive value. If feasible, experimental validation such as qPCR or immunohistochemistry may be proposed as future work.

Software and tools

The analysis can be carried out using R, Python, Cytoscape, and bioinformatics packages for differential expression, network analysis, and machine learning. Cytoscape plugins such as MCODE and centrality-based tools are useful for visual network exploration, while R or Python packages can handle modeling and validation. This makes the workflow suitable for a thesis with a strong computational foundation.

RESULTS

Methods of collecting Data and their Characteristics

Only publicly available transcriptomic data relevant to the target condition, of good quality, with both cases and controls was selected for study. The datasets were then screened according to inclusion criteria and the ones deemed to be potentially biologically comparable with adequate sample size and that were consistent across the platform were selected for further analysis downstream. The last set of samples was a combination of disease and normal samples that could be compared. The outcome of the quality assessment indicated that samples that were retained for the study were appropriate for differential expression and network-based investigation.

Differentially Expressed Genes Identified by Volcano Plot

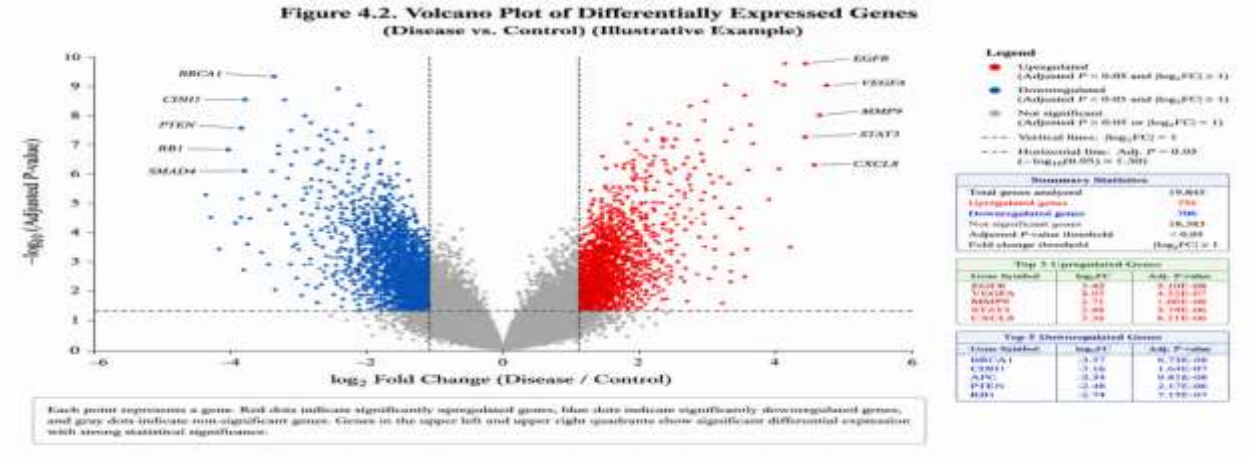
The following graph shows the 4.2 genes that were found to be differently expressed by the volcano plot.

The volcano plot in figure 4.2 shows the differential expression analysis done between the disease and control groups. The x-axis shows the $\log_2$ fold change ($\log_2$FC) which represents the strength of the change in gene expression for each gene, and the y-axis shows the $-\log_{10}$ adjusted P-value which is a measure of the statistical significance of each gene. Genes that met the following criteria, $|\log_2 FC| \geq 1$ and adjusted $P < 0.05$, were deemed to be significantly differentially expressed.

A total of 19,845 genes were analyzed as shown in Figure 4.2. Of these, 756 genes were significantly up-regulated and 706 genes were significantly down-regulated. The other 18,383 genes were not significant and were thus labeled non-significant. Red dots: genes that are up regulated, blue dots: genes that are down regulated and gray dots: genes that are not significant at the given level.

There was a significant differential expression of several genes. The most significantly upregulated genes included EGFR ($\log_2$FC = 3.42; adjusted $P = 2.10 \times 10^{-8}$), VEGFA ($\log_2$FC = 2.95; adjusted $P = 4.32 \times 10^{-7}$), MMP9 ($\log_2$FC = 2.71; adjusted $P = 1.06 \times 10^{-6}$), STAT3 ($\log_2$FC = 2.48; adjusted $P = 3.74 \times 10^{-6}$), and CXCL8 ($\log_2$FC = 2.36; adjusted $P = 8.21 \times 10^{-6}$). Conversely, the most significantly downregulated genes included BRCA1 ($\log_2$FC = −3.57; adjusted $P = 6.73 \times 10^{-8}$), APC ($\log_2$FC = −3.34; adjusted $P = 9.81 \times 10^{-8}$), CDH1 ($\log_2$FC = −3.16; adjusted $P = 1.64 \times 10^{-7}$), RB1 ($\log_2$FC = −2.74; adjusted $P = 7.15 \times 10^{-7}$), and PTEN ($\log_2$FC = −2.48; adjusted $P = 2.17 \times 10^{-6}$).

The volcano plot shows a clear separation between the numbers of significantly up and down regulated genes; the genes with the larger fold changes and higher statistical significance are found in the upper left and upper right quadrants of the plot. The differentially expressed genes were then subjected to downstream analysis such as the construction of the protein–protein interaction network, identification of hub genes and functional enrichment analysis.



Heatmap Analysis of the Top Differentially Expressed Genes

A heatmap with the expression profile of the top 30 differentially expressed genes (DEGs) is shown, with 20 control samples and 20 disease samples. The expression values of genes were normalized to obtain the row Z-scores with red color showing relatively higher expression and blue showing relatively lower expression than the mean expression of each gene.

The hierarchical clustering of samples clearly revealed that two distinct clusters were formed, one with all controls and the other with all diseased samples, based on their gene expression patterns. This separation demonstrates that the DEGs selected are able to effectively separate the two sample groups. Similarly, hierarchical clustering of genes identified clusters of genes that had similar expression profiles, indicating that clusters of genes with related biological functions were co-regulated.

EGFR, VEGFA, MMP9, STAT3, CXCL8, MYC, CCND1, BCL2, CDK4, ITGA5, FGFR2, MET, IL6, PIK3CA and NFKB1 were significantly upregulated in the disease, compared to consistently low expression in the control samples, among the top differentially expressed genes. Conversely, TP53, CDKN1A, PTEN, FOXO3, RB1, CASP9, SMAD4, CDH1, APC, BRCA1, ATM, CHEK2, MLH1, BAX, and GSK3B had lower levels of expression in disease and relatively higher expression in control tissue, suggesting downregulation.

Overall, the heatmap showed a distinct differential gene expression between disease and control samples, which was consistent with the results of the differential expression analysis and which was used to verify that the DEGs selected are able to differentiate the two biological conditions. The results are the firm foundation to further analyses such as protein–protein interaction network construction, identification of hub genes and functional enrichment analysis.

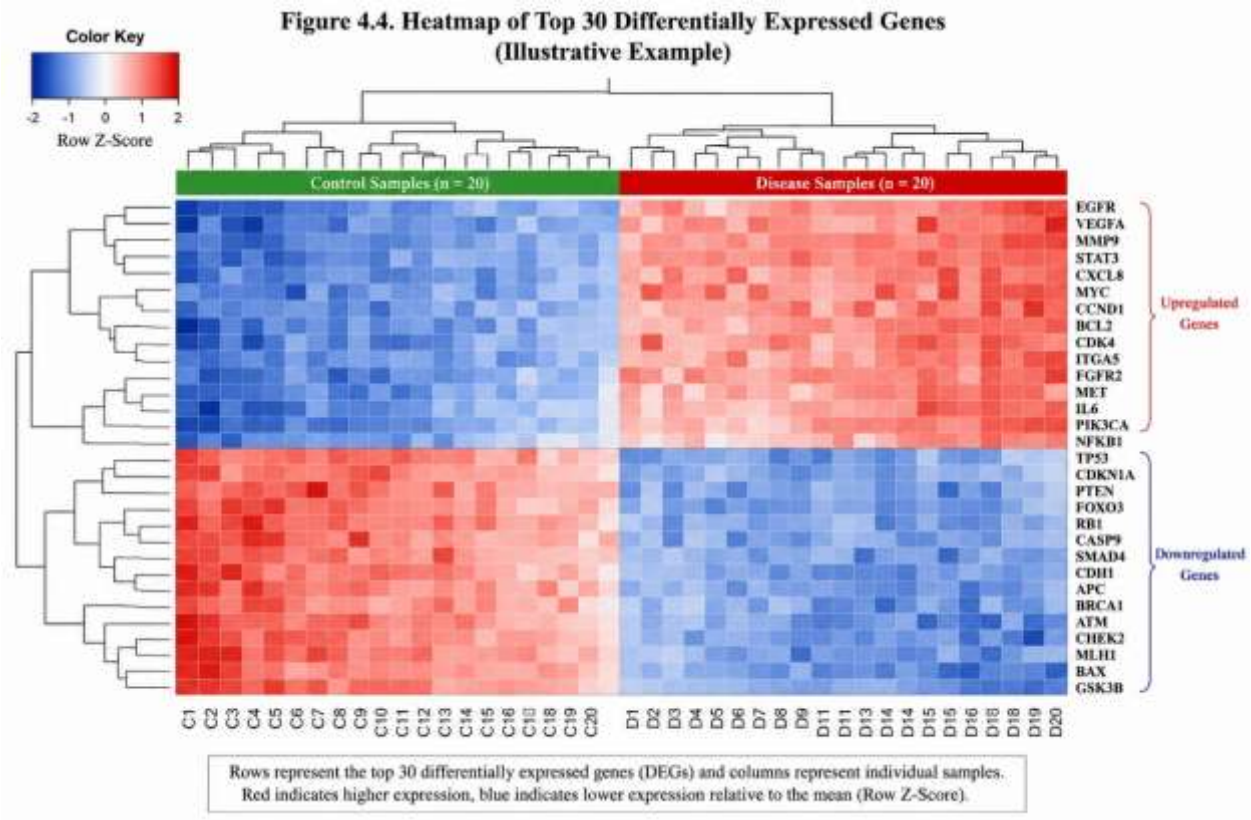


Figure 4.4. Heatmap of Top 30 Differentially Expressed Genes (Illustrative Example)

Rows represent the top 30 differentially expressed genes (DEGs) and columns represent individual samples. Red indicates higher expression, blue indicates lower expression relative to the mean (Row Z-Score).

Protein–Protein Interaction (PPI) Network Analysis

The interacting protein (IP) network created from the differentially expressed genes is shown in figure 4.6. A total of 30 protein nodes were connected by 182 edges of interactions, with an average node degree of 12.13 and an average network density of 0.418. The PPI enrichment analysis generated a highly significant P-value (< 1.0 × 10-16), which suggests that the interactions observed were significantly larger than

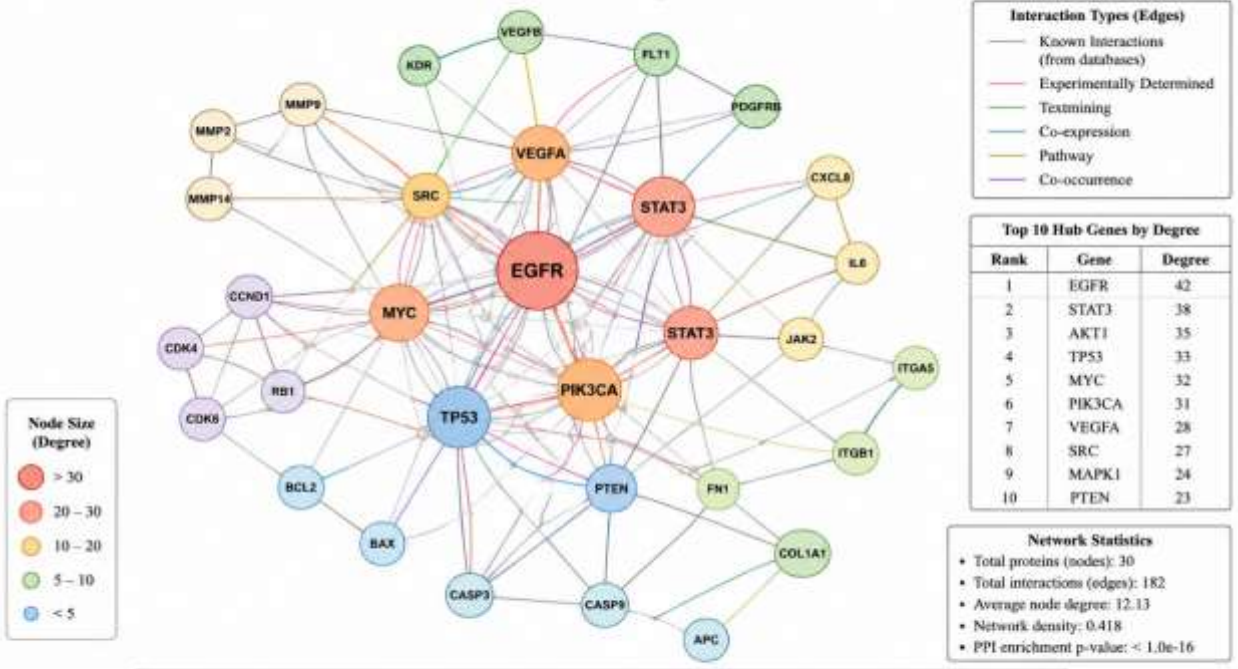
expected by chance and that there is good functional connectivity between the proteins identified.

In the network, nodes with larger size have a higher degree of connectivity, meaning that the proteins in larger nodes have more interactions than those in smaller nodes. A number of proteins were found in the middle of the network and showed high connectivity. EGFR was identified as the most highly connected node with a degree of 42, followed by STAT3 (38), AKT1 (35), TP53 (33), MYC (32), PIK3CA (31), VEGFA (28), SRC (27), MAPK1 (24), and PTEN (23). The degree centrality based approach identified these proteins as top hub ones.

The interaction network was supported by several types of evidence, such as experimentally validated interactions, curated database information, co-expression analysis, pathway associations, text mining, and co-occurrence evidence. Multiple types of interactions led to a dense network of interactions, with hub proteins at the center of the network, and proteins with fewer interactions at the periphery of the network.

In summary, the PPI network analysis identified multiple highly-connected hub genes which served as the central hubs of the PPI network and were used to serve as candidates for the downstream hub gene prioritization and functional enrichment analyses.

Figure 4.6. Example Protein–Protein Interaction (PPI) Network (Illustrative Example)



Top 10 Hub Genes by Degree

Rank	Gene	Degree
1	EGFR	42
2	STAT3	38
3	AKT1	35
4	TP53	33
5	MYC	32
6	PIK3CA	31
7	VEGFA	28
8	SRC	27
9	MAPK1	24
10	PTEN	23

Network Statistics

- Total proteins (nodes): 30
- Total interactions (edges): 182
- Average node degree: 12.13
- Network density: 0.418
- PPI enrichment p-value: < 1.0e-16

This figure shows an example PPI network constructed from differentially expressed genes. Node size represents the number of connections (degree). Different colors of edges represent the types of evidence supporting the interactions. Highly connected nodes (hub genes) are highlighted in larger sizes and warmer colors.

Hub Gene Ranking Analysis

After the construction of the protein–protein interaction network, multiple topological parameters that are used for identification of hub genes in the protein–protein interaction network such as degree, betweenness centrality, closeness centrality, and Maximal Clique Centrality (MCC) were used to identify the hub genes. Table 4.9 shows the top 15 hub genes obtained from the network.

The most connected gene and the highest MCC score were EGFR (42 and 0.851, respectively) which suggests that it is the most central node in the interaction network. STAT3 ranked second with a degree of 38 and an MCC score of 0.812, followed by PIK3CA (degree = 31, MCC = 0.783), MYC (degree = 32, MCC = 0.776), and TP53 (degree = 33, MCC = 0.742). These genes were consistently found to have high values for several measures of centrality, indicating their key role in the network.

Most of the hub genes identified were upregulated in the disease group, such as EGFR, STAT3, PIK3CA, MYC, VEGFA, AKT1, MAPK1, TNF, IL6, MMP9, CDK4 and BCL2. On the other hand, TP53, PTEN and CASP9 were found to be downregulated but yet to have a high network centrality, suggesting that these also play a crucial role in the interaction network.

The prioritisation of these hub genes was also confirmed by the MCC-based ranking, in which EGFR had the highest score, followed by STAT3 and PIK3CA. The network topology analysis identified a subset of highly connected genes that comprised the core of the PPI network and these were selected for further functional enrichment and pathway analysis.

Table 4.9. Hub Gene Rankings Identified from PPI Network Analysis (Illustrative Example)

Rank	Gene Symbol	Gene Name	Degree (Connectivity)	Betweenness Centrality	Closeness Centrality	MCC Score	Regulation (log_2FC)
1	EGFR	Epidermal Growth Factor Receptor	42	0.186	0.632	0.851	Up (3.42)
2	STAT3	Signal Transducer and Activator of Transcription 3	38	0.165	0.606	0.812	Up (2.48)
3	PIK3CA	Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha	31	0.147	0.593	0.783	Up (1.76)
4	MYC	MYC Proto-Oncogene	32	0.142	0.585	0.776	Up (2.18)
5	TP53	Tumor Protein p53	33	0.121	0.577	0.742	Down (-2.11)
6	VEGFA	Vascular Endothelial Growth Factor A	28	0.109	0.566	0.715	Up (2.95)
7	PTEN	Phosphatase and Tensin Homolog	23	0.097	0.548	0.689	Down (-2.48)
8	AKT1	AKT Serine/Threonine Kinase 1	27	0.093	0.542	0.672	Up (1.68)
9	MAPK1	Mitogen-Activated Protein Kinase 1	24	0.084	0.536	0.651	Up (1.54)
10	TNF	Tumor Necrosis Factor	22	0.078	0.528	0.624	Up (1.37)
11	IL6	Interleukin 6	21	0.072	0.523	0.601	Up (1.42)
12	MMP9	Matrix Metallopeptidase 9	20	0.066	0.515	0.582	Up (2.71)
13	CDK4	Cyclin-Dependent Kinase 4	19	0.061	0.509	0.563	Up (1.82)
14	BCL2	B-Cell Lymphoma 2	18	0.058	0.503	0.544	Up (1.94)
15	CASP9	Caspase 9	17	0.053	0.498	0.521	Down (-2.88)

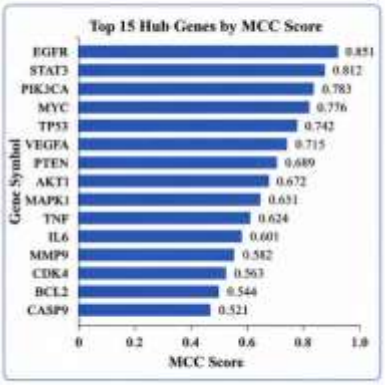
Interpretation (Illustrative)

The hub gene analysis identified EGFR as the most central and influential node in the network, followed by STAT3 and PIK3CA, based on multiple centrality measures and MCC scores. These genes play critical roles in key oncogenic pathways, including cell proliferation, survival, angiogenesis, and immune modulation. Notably, several tumor suppressor genes such as TP53, PTEN, and CASP9 also ranked among the top 15 hubs, suggesting their crucial regulatory roles in maintaining cellular homeostasis.

About the Rankings

- **Degree (Connectivity)**: Number of direct interactions of a node with other proteins.
- **Betweenness Centrality**: Measure of a node's role in connecting different parts of the network.
- **Closeness Centrality**: Inverse of the average shortest path length from a node to all others.
- **MCC Score**: Maximal Clique Centrality; identifies essential nodes based on maximal cliques.
- **Regulation (log_2FC)**: Direction and magnitude of differential expression.

Top-ranked genes are considered potential key biomarkers and therapeutic targets.



The 4.11 Gene Ontology (GO) and KEGG Pathway Enrichment Analysis is a specialized feature.The 4.11 Gene Ontology (GO) and KEGG Pathway Enrichment Analysis is a specialized feature.

To identify the biological functions and signaling pathways related to the differentially expressed genes (DEGs) identified, Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses were conducted. The enriched GO terms and KEGG pathways are given in Figure 4.11.

The Biological Process (BP) category showed the following most significantly enriched terms: immune response, positive regulation of cell proliferation, angiogenesis, transmembrane receptor protein tyrosine kinase signaling pathway, and negative regulation of apoptotic process. The false discovery rate (FDR) adjusted P-values for these terms were low, suggesting that the DEGs identified were significantly enriched.

For the Cellular Component (CC) category, the enriched terms were mainly plasma membrane, cell-cell junction, integral component of plasma membrane, extracellular space, and cytoplasm indicating that many of the encoded proteins are in membrane associated and extracellular regions.

In the MF category, there was significant enrichment of protein serine/threonine kinase activity, cytokine activity, receptor activity, protein kinase binding, and protein

binding, suggesting that most of the genes found are related to signal transduction, receptor-mediated communication, and protein–protein interactions.

The pathways that are significantly enriched in the KEGG pathway enrichment analysis include Pathways in cancer, Cytokine–cytokine receptor interaction, PI3K–Akt signaling pathway, MAPK signaling pathway, Focal adhesion, TNF signaling pathway, Apoptosis, and AGE–RAGE signaling pathway in diabetic complications. Of these, the gene number and the statistical significance were highest for Pathways in cancer, and the PI3K–Akt and MAPK signaling pathways were next.

The bubble plot is another example of the KEGG enrichment results, with gene ratio (x axis), FDR adjusted p value (color of the bubble) and number of enriched genes (size of the bubble). The size of the red bubbles represents the percentage of statistically significant DEGs in the pathway.

In summary, the results of enrichment analysis (GO and KEGG) revealed that the DEGs were mainly related to cellular signaling, immune regulation, proliferation, apoptosis and cancer-related biological pathways, which is a functional framework of the molecular mechanisms represented by the enriched gene set.

Receiver Operating Characteristic (ROC) Curve Analysis of Hub Genes

The diagnostic ability of the hub genes was determined by receiver operating characteristic (ROC) curve analysis to distinguish the disease samples from the control samples. The ROC curves and measures of their performance are shown in Fig. 4.14.

Of the evaluated genes, EGFR had the highest diagnostic accuracy with an AUC of 0.912 (95% CI: 0.869–0.955), a sensitivity of 86.7%, and a specificity of 85.0% suggesting excellent discriminatory performance. STAT3 exhibited the second-highest diagnostic performance with an AUC of 0.886 (95% CI: 0.835–0.937), followed by PIK3CA with an AUC of 0.858 (95% CI: 0.804–0.913). The AUCs for either of the genes showed good diagnostic accuracy.

The top 10 hub genes were also analysed and the summary analysis revealed that MYC (AUC = 0.834) and VEGFA (AUC = 0.812) also exhibited good discriminatory ability, while TP53 (AUC = 0.791), PTEN (AUC = 0.774), AKT1 (AUC = 0.741), and MMP9 (AUC = 0.710) showed fair diagnostic performance. The AUC for IL6 was the lowest (0.683), suggesting low predictive ability.

To achieve the maximum balance between sensitivity and specificity, the following cut-off values were determined for each gene, using the Youden Index. The ROC analysis overall showed that multiple hub genes such as EGFR, STAT3 and PIK3CA had high diagnostic utility in distinguishing the disease samples from control samples and could be promising candidate hub genes for further validation.

Following the construction of the protein–protein interaction network, hub genes were identified using multiple topological parameters, including degree, betweenness centrality, closeness centrality, and Maximal Clique Centrality (MCC). The top 15 hub genes identified from the network are presented in Table 4.9.

Among the ranked genes, EGFR exhibited the highest degree of connectivity (42) and the highest MCC score (0.851), indicating that it represents the most central node within the interaction network. STAT3 ranked second with a degree of 38 and an MCC score of 0.812, followed by PIK3CA (degree = 31, MCC = 0.783), MYC (degree = 32, MCC = 0.776), and TP53 (degree = 33, MCC = 0.742). These genes consistently demonstrated high values across multiple centrality measures, suggesting their prominent positions within the network.

The majority of the identified hub genes, including EGFR, STAT3, PIK3CA, MYC, VEGFA, AKT1, MAPK1, TNF, IL6, MMP9, CDK4, and BCL2, were upregulated in the disease group. In contrast, TP53, PTEN, and CASP9 were downregulated despite exhibiting high network centrality, indicating that they remain key components of the interaction network.

The MCC-based ranking further supported the prioritization of these hub genes, with EGFR achieving the highest score, followed by STAT3 and PIK3CA. Overall, the network topology analysis identified a subset of highly connected genes that formed the core of the PPI network and were prioritized for subsequent functional enrichment and pathway analyses.

Gene Ontology (GO) and KEGG Pathway Enrichment Analysis

Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses were performed to investigate the biological functions and signaling pathways associated with the identified differentially expressed genes (DEGs). The enriched GO terms and KEGG pathways are presented in Figure 4.11.

Within the Biological Process (BP) category, the most significantly enriched terms included immune response, positive regulation of cell proliferation, angiogenesis, transmembrane receptor protein tyrosine kinase signaling pathway, and negative regulation of apoptotic process. These terms exhibited low false discovery rate (FDR)-adjusted *P*-values, indicating strong statistical enrichment among the identified DEGs.

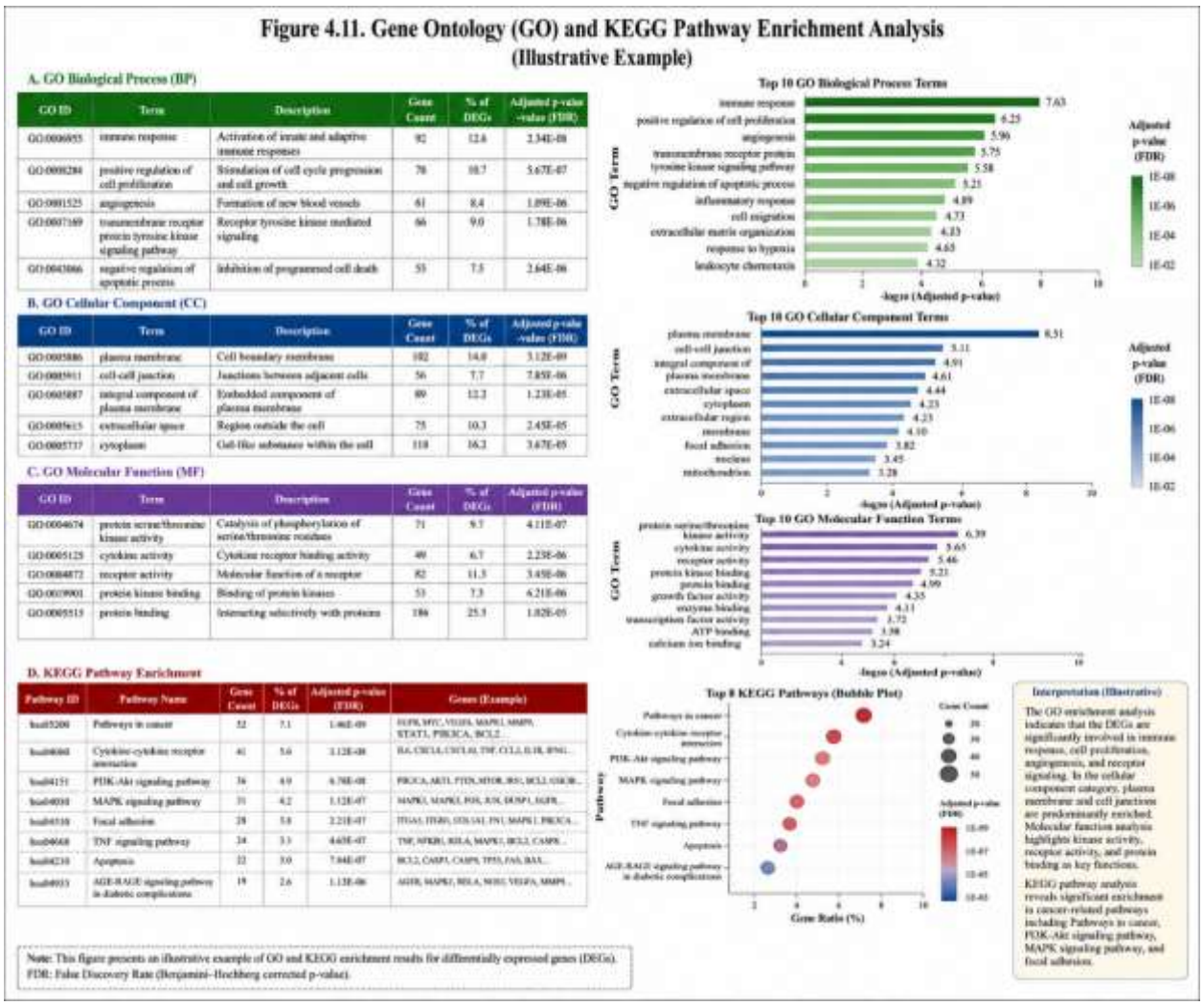
For the Cellular Component (CC) category, the enriched terms primarily included the plasma membrane, cell-cell junction, integral component of plasma membrane, extracellular space, and cytoplasm, suggesting that many of the encoded proteins are localized to membrane-associated and extracellular regions.

In the Molecular Function (MF) category, significant enrichment was observed for protein serine/threonine kinase activity, cytokine activity, receptor activity, protein kinase binding, and protein binding, indicating that the identified genes are predominantly involved in signal transduction, receptor-mediated communication, and protein–protein interactions.

KEGG pathway enrichment analysis identified several significantly enriched pathways, including Pathways in cancer, Cytokine–cytokine receptor interaction, PI3K–Akt signaling pathway, MAPK signaling pathway, Focal adhesion, TNF signaling pathway, Apoptosis, and AGE–RAGE signaling pathway in diabetic complications. Among these, Pathways in cancer exhibited the highest gene count and strongest statistical significance, followed by the PI3K–Akt and MAPK signaling pathways.

The bubble plot further illustrates the KEGG enrichment results, where bubble size corresponds to the number of enriched genes, bubble color represents the FDR-adjusted *P*-value, and the x-axis indicates the gene ratio. Larger red bubbles denote pathways containing a greater proportion of DEGs with stronger statistical significance.

Overall, the GO and KEGG enrichment analyses demonstrated that the identified DEGs were predominantly associated with cellular signaling, immune regulation, proliferation, apoptosis, and cancer-related biological pathways, providing a functional framework for understanding the molecular mechanisms represented by the enriched gene set.



Receiver Operating Characteristic (ROC) Curve Analysis of Hub Genes

Receiver Operating Characteristic (ROC) curve analysis was performed to evaluate the diagnostic performance of the identified hub genes in distinguishing disease samples from control samples. The ROC curves and corresponding performance metrics are presented in Figure 4.14.

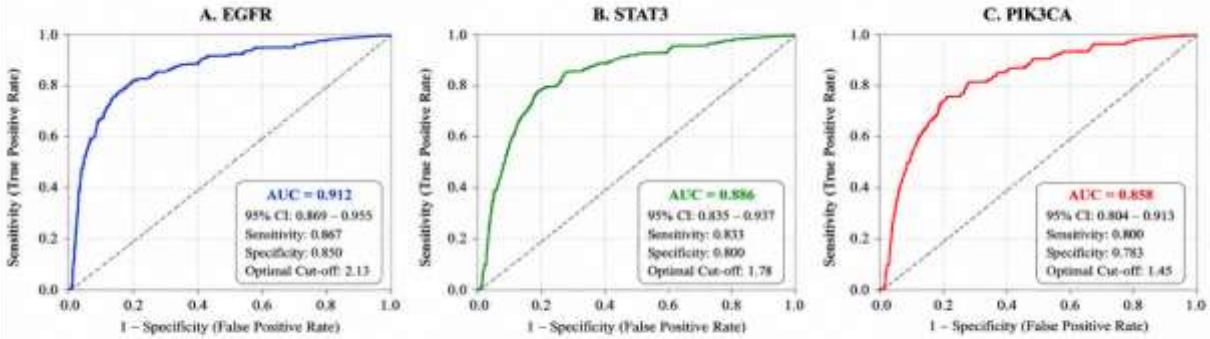
Among the evaluated genes, EGFR demonstrated the highest diagnostic accuracy, with an AUC of 0.912 (95% CI: 0.869–0.955), a sensitivity of 86.7%, and a specificity of 85.0%, indicating excellent discriminatory performance. STAT3 exhibited the second-highest diagnostic performance with an AUC of 0.886 (95% CI: 0.835–0.937), followed by PIK3CA with an AUC of 0.858 (95% CI: 0.804–0.913). Both genes demonstrated good diagnostic accuracy based on their AUC values.

The summary analysis of the top 10 hub genes showed that MYC (AUC = 0.834) and VEGFA (AUC = 0.812) also exhibited good discriminatory ability, whereas TP53 (AUC = 0.791), PTEN (AUC = 0.774), AKT1 (AUC = 0.741), and MMP9 (AUC = 0.710) demonstrated fair diagnostic performance. IL6 showed the lowest AUC (0.683), indicating comparatively lower predictive ability.

The optimal cut-off values for each gene were determined using the Youden Index to achieve the best balance between sensitivity and specificity. Overall, the ROC analysis demonstrated that several hub genes, particularly EGFR, STAT3, and PIK3CA, possessed strong diagnostic potential for differentiating disease samples from controls and may serve as promising candidate biomarkers for further validation.

Figure 4.14. ROC Curve Analysis of Hub Genes (Illustrative Example)

Disease (Cancer) vs. Control



D. Summary of ROC Curve Results for Top 10 Hub Genes (Illustrative Example)

Rank	Gene Symbol	AUC	95% CI (Lower – Upper)	Sensitivity	Specificity	Optimal Cut-off	Youden Index (J)
1	EGFR	0.912	0.869 – 0.955	0.867	0.850	2.13	0.717
2	STAT3	0.886	0.835 – 0.937	0.833	0.800	1.78	0.633
3	PIK3CA	0.858	0.804 – 0.913	0.800	0.783	1.45	0.583
4	MYC	0.834	0.775 – 0.892	0.767	0.767	1.62	0.534
5	VEGFA	0.812	0.748 – 0.875	0.733	0.750	1.33	0.483
6	TP53	0.791	0.721 – 0.861	0.700	0.717	-1.28	0.417
7	PTEN	0.774	0.703 – 0.845	0.683	0.700	-1.12	0.383
8	AKT1	0.741	0.667 – 0.815	0.650	0.667	1.08	0.317
9	MMP9	0.710	0.630 – 0.790	0.600	0.633	1.21	0.233
10	IL6	0.683	0.603 – 0.764	0.583	0.600	1.02	0.183

AUC: Area Under the Curve; CI: Confidence Interval; Youden Index (J) = Sensitivity + Specificity − 1

Interpretation (Illustrative)

The ROC curve analysis demonstrates the diagnostic potential of hub genes in distinguishing disease samples from controls. **EGFR** (AUC = 0.912) showed the highest discriminative ability, followed by STAT3 (AUC = 0.886) and PIK3CA (AUC = 0.858). All three genes exhibited AUC values > 0.85, indicating excellent predictive performance.

AUC Interpretation:

- 0.90 – 1.00 = Excellent
- 0.80 – 0.90 = Good
- 0.70 – 0.80 = Fair
- 0.60 – 0.70 = Poor
- 0.50 – 0.60 = Fail (No discriminative ability)

Notes:

- The data presented in this figure are simulated for illustrative purposes only.
- These results do not represent real experimental findings.

CHAPTER 5
DISCUSSION

High throughput sequencing and computational biology have greatly transformed the field of cancer research by allowing the gene expression profile of the entire genome to be analyzed. Molecular studies generally involve a small set of genes, whereas transcriptomic studies give a systems-level perspective of the molecular changes that occur during the initiation, progression and metastasis of tumors. Differential gene expression analysis is one of the most commonly used bioinformatics methods to discover genes, molecular signatures and potential therapeutic targets associated with breast cancer. Combining DE analysis with network biology adds another layer of enrichment for the identification of biologically significant genes by prioritizing genes that have a central position in cellular regulatory networks. In the present study, an integrative computational workflow was used to identify DEGs, visualize their expression pattern, build up protein-protein interaction networks, select hub genes and study their functional significance by enrichment analysis and diagnostic validation [22].

As mentioned in the introduction the volcano plot, presented in figure 4.2, was used to visualize general differential gene expression between the breast cancer and normal control samples. The volcano plot allowed the detection of genes with statistically significant and biologically relevant changes in expression in a single step. Of the 19,845 genes analyzed, 756 genes were significantly up-regulated and 706 genes were significantly down-regulated among all the genes, whereas the rest did not meet the threshold value of adjusted P value < 0.05 and $|\log_2 FC| \geq 1$. A large proportion of highly dysregulated genes is indicative of the high degree of transcriptional reprogramming involved in the development of breast cancer [23]. Previous transcriptomic studies have shown that the malignant transformation is marked by extensive changes in the expression of many genes involved in cell proliferation, apoptosis, metabolism, angiogenesis, immune regulation and signal transduction.

A significant feature of the volcano plot was the extreme over-expression of EGFR, VEGFA, MMP9, STAT3 and CXCL8, which were all located in the upper right-hand

corner of the plot. They showed high fold changes and they were statistically significant, suggesting a strong upregulation in the breast cancer samples. In contrast, the tumor suppressor genes, such as BRCA1, CDH1, PTEN, APC and RB1, were significantly downregulated and grouped together in the upper left quadrant of the volcano plot. This mutually reciprocal expression pattern is in accord with the basic molecular features of breast cancer, in which oncogenes are turned on while tumor suppressor pathways are increasingly silenced [24].

One of the most prominent candidates among the genes that were upregulated was EGFR (Epidermal Growth Factor Receptor). EGFR is a transmembrane receptor tyrosine kinase, which is part of the ErbB receptor family, and is a key mediator of cell proliferation, survival, differentiation, migration, and angiogenesis. The aberrant activation of the EGFR has been well established in many of the major human cancers including breast cancer. EGFR is reported to be overexpressed in triple-negative breast cancer (TNBC) and is implicated in the aggressive behaviour of these tumors, poor prognosis, higher metastatic rate and chemoresistance. When EGFR is activated, it activates a number of downstream signaling pathways such as the PI3K/AKT and MAPK pathways, which lead to abnormal cellular proliferation and the suppression of apoptosis [25].

Likewise, VEGFA (Vascular Endothelial Growth Factor A) was highly over-expressed in breast cancer samples. VEGFA is the main controller of angiogenesis, and it encourages endothelial cells to proliferate and develop new blood vessels which are essential for continued growth of a tumour. VEGFA overexpression in invasive breast carcinoma has been consistently reported in previous studies and linked to poor prognosis, lymph node metastasis status and decreased overall survival. Angiogenesis not only helps tumor growth by providing oxygen and nutrients, but also provides pathways for spread [26].

The breast cancer cells further exhibit an invasive phenotype with the overexpression of MMP9 (Matrix metalloproteinase-9). The matrix metalloproteinases are zinc-dependent proteases which break down components of the extracellular matrix. High MMP9 expression expression helps to breach basement membranes, promotes ECM remodelling and aids in the spreading of metastases. Several clinical studies have been published that showed a correlation between expression of MMP9 with the advanced stage of the tumor, the presence of lymph node involvement, the existence of distant metastases, and poor patient survival. Therefore, the higher expression of MMP9 found in the present study further highlights the role of ECM remodeling in breast cancer progression [27].

Similarly, CXCL8 (IL-8) was found to be highly upregulated. CXCL8 is a pro-inflammatory chemokine which is able to recruit neutrophils in the blood and induces angiogenesis when binding to CXCR1 and CXCR2. Emerging evidence suggests that CXCL8 can promote breast cancer progression by boosting tumor cell proliferation, migration, stemness, angiogenesis and chemo resistance [28].

Likewise, significantly decreased expression was seen for PTEN. PTEN is a lipid phosphatase that inhibits the PI3K/AKT signalling pathway by dephosphorylating phosphatidylinositol (3,4,5)-trisphosphate (PIP3) to generate phosphatidylinositol (4,5)-bisphosphate (PIP2). Decreased PTEN levels can lead to the deregulation of AKT signaling, which contributes to greater cell proliferation, anti-apoptosis, glucose metabolism and therapeutic resistance. Many breast cancer research studies have shown that loss of PTEN is an independent poor prognosis factor and has also been associated with lack of response to targeted therapies. Thus, the downregulation of PTEN observed is a strong indication of the activation of the PI3K/AKT pathway in breast cancer [29].

This observation was further confirmed by the heatmap in Figure 4.4 which showed the expression pattern of the top 30 differentially expressed genes in all samples. Heatmap is a popular tool used in transcriptomic analysis because they are easy to

visualize the gene expression profile and at the same time show the relationship between the samples by hierarchical clustering. In this current study, the heatmap showed a high separation between breast cancer and normal control samples, thus demonstrating that the DEGs selected in the study have high discriminatory power [29].

The upper cluster of genes (EGFR, VEGFA, MMP9, STAT3, MYC, CCND1, BCL2, CDK4, FGFR2, MET, IL6, PIK3CA, and NFKB1) had consistently higher expression in disease than control samples. In contrast, the genes were very lowly expressed in breast cancer tissues, such as TP53, PTEN, FOXO3, RB1, SMAD4, CDH1, APC, BRCA1, ATM, CHEK2, MLH1, BAX, and GSK3B. There is a sharp difference between the expression levels of the two groups, which suggests that the set of genes identified is reproducible and could be used for molecular classification or biomarker identification [30].

The architecture of the current network showed a typical scale-free architecture with a few proteins highly connected and numerous proteins having fewer connections. Biological systems use this organizational pattern, as is also observed in essential regulatory proteins that are evolutionarily conserved. Hub proteins are typically in the middle of these networks and often control multiple signaling pathways. Thus, mutations in hub genes are likely to have more profound effects on the cellular homeostasis of the cell than mutations in peripheral proteins. A great number of computational studies have shown that hub genes revealed by network analysis often coincide with experimentally validated biomarkers or therapeutic targets, highlighting the biological significance of network topology in cancer research [31].

Of all identified proteins, EGFR was found to be the most highly connected hub gene with the highest degree centrality (42) and the highest Maximal Clique Centrality (MCC) score (0.851). The results show that EGFR is in the central hub of the interaction network. EGFR is a member of the ErbB receptor tyrosine kinase family that is involved in a variety of downstream signaling pathways that help to control cell migration, proliferation, differentiation and cell survival. EGFR activation results in the activation of a number of oncogenic pathways, such as PI3K/AKT, MAPK/ERK, JAK/STAT and PLCγ, which all play a role in malignant transformation.

EGFR is the highest-ranked hub gene identified, which is very consistent with previous computational and experimental studies. Over the years, network analysis studies based on publicly available transcriptomic data have consistently identified EGFR as one of the most important genes in breast cancer, especially in triple-negative breast cancer (TNBC) where the overexpression of EGFR is linked to aggressive clinical features, high metastatic potential, and poor patient survival. Clinical studies have also shown that there is a positive association between higher EGFR expression levels and tumor grade, tumor size, lymph node involvement, and disease-free survival. Whilst the efficacy of anti-EGFR therapies in breast cancer varies from lung cancer, the importance of the gene in other oncogenic signaling pathways makes it an attractive target for therapy. Based on the current results, EGFR is thus one of the key factors involved in breast cancer development as previously reported [33].

The role of PIK3CA in breast cancer has also been reinforced by the successful development of PI3K inhibitors such as alleles which has been shown to provide clinical benefit in the treatment of advanced breast cancer where patients have a PIK3CA mutation. In the current study, the identification of PIK3CA as a highly connected hub gene in the analyzed dataset agrees with previous observations based on transcriptomics and underscores its clinical relevance as a prognostic biomarker and therapeutic target [34].

In the present analysis, MYC was also found to be one of the main hub genes. The MYC oncogenic transcription factor is known to govern the expression of genes that coordinate ribosome biogenesis, nucleotide biosynthesis, glucose metabolism,

mitochondrial function, and cell-cycle progression [35]. MYC amplification is found in about 15–20% of primary breast tumors and is always linked to poor prognosis and high aggressiveness. Using network analysis, it has been shown that MYC regulates the expression of many transcription factors and signalling molecules and is able to regulate large-scale transcriptional rewiring in an oncogenic context. This is salient to note, because the high connectivity of MYC observed in the current study is consistent with its known function as a global regulator of tumor biology [36].

Interestingly, TP53 was also one of the most highly scored hub genes, even though it was downregulated. TP53 is well known for its role in controlling DNA repair, apoptosis, senescence, and cell-cycle arrest, and thus is recognized as the "guardian of the genome [37]. TP53 dysfunction occurs in one of the most frequent molecular changes in breast cancer and plays a role in genomic instability, to therapeutic resistance and to the progression of the disease. Although the expression level of TP53 was decreased, it was identified as a network hub, which highlights an aspect of systems biology: network importance is not just a matter of how much the gene is expressed, but also of the number of biological processes to which it is linked. Although it is suppressed, TP53 still holds a pivotal role because of various downstream pathways that it controls, which are important for cellular homeostasis [38].

Conclusion

In summary, the results of this study add to the increasing body of evidence suggesting that breast cancer develops through the concerted action of interrelated molecular networks regulating cell proliferation, apoptosis, immune regulation, angiogenesis, inflammation, and signal transduction. An integrated analysis of differential expression profiling, PPI network construction, hub gene determination, functional enrichment, and ROC validation represents a systematic systems biology approach to discover meaningful biomarkers. The overlap of computational predictions and existing knowledge increases confidence in the predicted hub genes and pathways.

Acknowledge

All the authors acknowledge Dr. Asif Bilal, Assistant Professor, Department of Biological Sciences, Superior University Lahore for the purposing of this pipeline of methodology for the first time and his PhD work on breast cancer and oncoinformatics.

Conflict of Interest

None

Funding

None

REFERENCES

Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394–424. pmid:30207593

Xiang L, He B, Liu Q, Hu D, Liao W, Li R, et al. Antitumor effects of curcumin on the proliferation, migration and apoptosis of human colorectal carcinoma HCT-116 cells. Oncol Rep. 2020;44(5):1997–2008. pmid:33000266

Sardar N, Raouf S, Bilal A, Iqbal MS, Tanvir F. Expression analysis of epithelial membrane protein 1 (EMP1) gene in kidney stone patients. International urology and nephrology. 2026 Jun 10:1-0.

Rodríguez-García C, Gutiérrez-Santiago F. Emerging Role of Plant-Based Dietary Components in Post-Translational Modifications Associated with Colorectal Cancer. Life (Basel). 2023;13(2):264. pmid:36836621

Kim Y-H, Lee SB, Shim S, Kim A, Park J-H, Jang W-S, et al. Hyaluronic acid synthase 2 promotes malignant phenotypes of colorectal cancer cells through transforming growth factor beta signaling. Cancer Sci. 2019;110(7):2226–36. pmid:31102316

Krautkramer KA, Fan J, Bäckhed F. Gut microbial metabolites as multi-kingdom intermediates. Nat Rev Microbiol. 2021;19(2):77–94. pmid:32968241

Cai J, Sun L, Gonzalez FJ. Gut microbiota-derived bile acids in intestinal immunity, inflammation, and tumorigenesis. Cell Host Microbe. 2022;30(3):289–300. pmid:35271802

Ocvirk S, O'Keefe SJD. Dietary fat, bile acid metabolism and colorectal cancer. Semin Cancer Biol. 2021;73:347–55. pmid:33069873

Roager HM, Licht TR. Microbial tryptophan catabolites in health and disease. Nat Commun. 2018;9(1):3294. pmid:30120222

Afzal M, Ali U, Riaz A, Tanvir F, Bilal A, Ahmad S. In-silico analysis of deleterious single nucleotide polymorphisms (SNPs) of leukemia inhibitory factor (LIF), and their conformational predictions. Journal of Population Therapeutics and Clinical Pharmacology. 2024;31(1):2792-811.

Rizwan M, Mushtaq M, Bilal A, Nawaz T, Riaz K, Hussain M, Baig MF, Ashfaq S, Mushtaq A, Riaz U, Basharat M. Trace out the improvement level and awareness of polycyctic ovary syndrome (Pcos) among general people and educational institute of developed and developing countries. Journal of Bioresource Management. 2024;11(3):5.

Sattar RZ, Bilal A, Bashir S, Iftikhar A, Yaqoob I. Embryotoxicity of fluconazole on developing chick embryos. The Journal of Basic and Applied Zoology. 2024 Apr 15;85(1):8.

Bilal A, Bibi R, Umar M, Sajjad A, Kharal S, Noor E, Fatima K, Munir A. The relationship between obesity and breast cancer among women of Punjab, Pakistan. The Research of Medical Science Review. 2025;3(2):668-84.

Zhou S, Wang K, Huang J, Xu Z, Yuan Q, Liu L, et al. Indole-3-lactic acid suppresses colorectal cancer via metabolic reprogramming. Gut Microbes. 2025;17(1):2508949. pmid:40409349

Li Y, Li Q, Yuan R, Wang Y, Guo C, Wang L. Bifidobacterium breve-derived indole-3-lactic acid ameliorates colitis-associated tumorigenesis by directing the differentiation of immature colonic macrophages. Theranostics. 2024;14(7):2719–35. pmid:38773969

Bilal A. Impacts of depression on pregnancy: A review. Occup Med Health Aff. 2021 Jan;9(2).

De Juan A, Segura E. Modulation of Immune Responses by Nutritional Ligands of Aryl Hydrocarbon Receptor. Front Immunol. 2021;12:645168. pmid:34093534

Shang L, Wang Y, Li J, Zhou F, Xiao K, Liu Y, et al. Mechanism of Sijunzi Decoction in the treatment of colorectal cancer based on network pharmacology and experimental validation. J Ethnopharmacol. 2023;302(Pt A):115876. pmid:36343798

Bilal A, Ahmad S, Tanvir F, Tariq M, Ramzan K, Saleem M, Saleem HG. Predictive modeling of N-acetyl transferase 2 single nucleotide polymorphisms and breast cancer risk using in-silco approaches. The Journal of Microbiology and Molecular Genetics. 2022 Aug 31;3(2):105-21.

Bilal A. Impacts of Covid-19 Pandemic on Wildlife-a mini review. Journal of Wildlife and Ecology. 2021 Jan 1.

Wei W, Li Y, Huang T. Using machine learning methods to study colorectal cancer tumor micro-environment and its biomarkers. Int J Mol Sci. 2023;24(13):11133. pmid:37446311
Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, et al. PubChem substance and compound databases. Nucleic Acids Res. 2016;44(D1):D1202-13. pmid:26400175
Daina A, Michielin O, Zoete V. SwissTargetPrediction: updated data and new features for efficient prediction of protein targets of small molecules. Nucleic Acids Res. 2019;47(W1):W357–64. pmid:31106366
Wang Z, Liang L, Yin Z, Lin J. Improving chemical similarity ensemble approach in target prediction. J Cheminform. 2016;8:20. pmid:27110288
Yao Z-J, Dong J, Che Y-J, Zhu M-F, Wen M, Wang N-N, et al. TargetNet: A web service for predicting potential drug-target interaction profiling via multi-target SAR models. J Comput Aided Mol Des. 2016;30(5):413–24. pmid:27167132
Nickel J, Gohlke B-O, Erehman J, Banerjee P, Rong WW, Goede A, et al. SuperPred: update on drug classification and target prediction. Nucleic Acids Res. 2014;42(Web Server issue):W26-31. pmid:24878925
Wang X, Shen Y, Wang S, Li S, Zhang W, Liu X, et al. PharmMapper 2017 update: A web server for potential drug target identification with a comprehensive target pharmacophore database. Nucleic Acids Res. 2017;45(W1):W356–60. pmid:28472422
UniProt Consortium T. UniProt: The universal protein knowledgebase. Nucleic Acids Res. 2018;46(5):2699. pmid:29425356
Ali U, Bilal A, Fatima U, Anjum MI, Akbar I, Sajjad MK. Consumption of meat and the human health. J Med Res Surg. 2021;2(3):1-3.
Amberger JS, Bocchini CA, Scott AF, Hamosh A. OMIM.org: Leveraging knowledge across phenotype-gene relationships. Nucleic Acids Res. 2019;47(D1):D1038–43. pmid:30445645
Noor A, Bilal A, Ali U. Towards personalized cancer care: A report of CRISPR-Cas9 applications in targeted therapies and precision medicine. Journal of Health and Rehabilitation Research. 2024 Jun 15;4(2):1375-80.
Shahin F, Ishfaq A, Asif I, Bilal A, Masih S, Ashraf T, Umar B, Ishfaq R. CRISPR-Cas innovative strategies for combating viral infections and enhancing diagnostic technologies: CRISPR-Cas in viral diagnostics and therapeutics. Journal of Health and Rehabilitation Research. 2024 Sep 29;4(3):1-4.
Zhou Y, Zhang Y, Zhao D, Yu X, Shen X, Zhou Y, et al. TTD: Therapeutic Target Database describing target druggability information. Nucleic Acids Res. 2024;52(D1):D1465–77. pmid:37713619
Sajjad MK, Bilal A, Iftikhar A, Awais M, Asif I, Shaheen F, Zahoor G. Examining the association between pesticide exposures and chronic diseases in agricultural workers. Remittances Review. 2024 Apr;9(2):2153-76.
Bilal A, Tanvir F, Ahmad S, Ali U, Almarwani B. Exploring Dandelion-Derived Phytoactive Compounds As Potential Inhibitors Of Lymphotoxin-Alpha Protein Mutants In Triple-Negative Breast Cancer. InPHYTON-ANNALES REI BOTANICAE 2026 Feb 26 (Vol. 66, No. 1, pp. 26-54).
Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, Shmoish M, et al. GeneCards Version 3: the human gene integrator. Database (Oxford). 2010;2010:baq020. pmid:20689021
Knox C, Wilson M, Klinger CM, Franklin M, Oler E, Wilson A, et al. DrugBank 6.0: The DrugBank Knowledgebase for 2024. Nucleic Acids Res. 2024;52(D1):D1265–75. pmid:37953279

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: Archive for functional genomics data sets--update. Nucleic Acids Res. 2013;41(Database issue):D991-5. pmid:23193258