SYSTEMS BIOLOGY

PRACTICE-1

GENOMIC TOOLS FOR ANALYZING TRANSCRIPTIONAL REGULATORY NETWORKS

Aim: Overview of genomic tools used for analysing transcriptional regulatory networks. These tools facilitate the identification, characterization, and interpretation of regulatory interactions between transcription factors (TFs) and their target genes.

Introduction

Transcriptional regulatory networks (TRNs) are complex systems that govern gene expression in cells. These networks are composed of transcription factors, which bind to specific DNA sequences to regulate the transcription of target genes. Understanding TRNs is critical to deciphering cellular functions, development, and responses to environmental changes. Genomic tools allow researchers to analyse TRNs by identifying binding sites, interactions, and regulatory motifs, providing insights into the hierarchical structure of gene regulation and the dynamics of cellular processes.

High-throughput technologies, such as Chromatin Immunoprecipitation followed by sequencing (ChIP-seq) and RNA sequencing (RNA-Seq), produce large volumes of data that can be analysed using specialized bioinformatics tools. These tools facilitate mapping of TF binding sites, prediction of target genes, and reconstruction of regulatory networks, enhancing our understanding of gene regulation at a systems level.

Tools for Analysing Transcriptional Regulatory Networks

1. **ChIP-Seq Analysis Tools:** ChIP-Seq (Chromatin Immunoprecipitation followed by sequencing) is a powerful method for identifying binding sites of TFs across the genome. Several tools help analyse ChIP-seq data to determine where and how TFs interact with DNA.

- **MACS (Model-based Analysis of ChIP-Seq)**: MACS is widely used for peak calling in ChIP-Seq data, identifying significant TF binding sites across the genome. It uses a model to predict the background distribution of noise and identifies enriched regions, helping researchers pinpoint regulatory elements accurately.
  Description: MACS normalizes data across control and treatment samples to identify peaks and assess their significance. It is highly effective in analysing TF binding patterns and enhancing our understanding of DNA-protein interactions in regulatory networks.
- **HOMER (Hypergeometric Optimization of Motif EnRichment)**: HOMER is a comprehensive toolkit for motif discovery and functional annotation of regulatory regions in ChIP-Seq data. It can identify both known and novel motifs in DNA sequences bound by TFs.
  Description: HOMER offers tools for peak calling, motif analysis, and gene ontology (GO) enrichment, making it valuable for comprehensive TRN analysis. Its motif discovery

function identifies binding motifs within ChIP-Seq peaks, elucidating TF-DNA binding specificities.

2. **RNA-Seq Analysis Tools:** RNA-Seq is a high-throughput technique for profiling gene expression. Several tools are available for analysing RNA-Seq data to determine the effect of TFs on gene expression and reconstruct TRNs.

- **DESeq2**: DESeq2 is a statistical tool for analysing differential gene expression in RNA-Seq data. It can identify genes that are upregulated or downregulated under specific conditions or in response to TF activity.
  Description: DESeq2 models the data based on a negative binomial distribution, adjusting for variability across samples. It helps identify genes regulated by specific TFs, providing insights into their roles in the TRN.
- **EdgeR**: EdgeR is another differential expression analysis tool for RNA-Seq data, useful for identifying genes that change expression levels in response to TF modulation.
  Description: Like DESeq2, EdgeR uses a negative binomial distribution model and is efficient for large RNA-Seq datasets. It aids in understanding gene expression changes influenced by TFs, revealing their regulatory impact on target genes.

3. **Motif Discovery Tools:** Motif discovery is essential in identifying binding sites of TFs within promoter or enhancer regions of target genes. Various tools facilitate motif analysis for TRN studies.

- **MEME (Multiple EM for Motif Elicitation)**: MEME is a widely used tool for discovering motifs in unaligned DNA sequences. It identifies recurring patterns that may represent TF binding sites.
  Description: MEME uses the Expectation-Maximization algorithm to identify motifs, providing insights into binding preferences of TFs and revealing conserved motifs across regulatory regions. It is useful for predicting novel TF binding motifs that contribute to TRN dynamics.
- **FIMO (Find Individual Motif Occurrences)**: FIMO, part of the MEME suite, scans a given sequence for occurrences of known motifs.
  Description: FIMO calculates the statistical significance of motif occurrences, allowing researchers to pinpoint specific TF binding sites within genomic sequences. It enhances the accuracy of regulatory network models by identifying exact binding sites based on motif data.

4. **Network Analysis Tools:** Network analysis tools are used to reconstruct and analyse the structure of TRNs, allowing researchers to visualize interactions between TFs and target genes.

- **Cytoscape**: Cytoscape is a network visualization tool that helps researchers visualize and analyse complex networks, including TRNs.
  Description: Cytoscape enables the integration of diverse data types and supports various plugins for enrichment analysis, network clustering, and pathway annotation. It helps in

mapping the hierarchical structure of TRNs, enabling insights into key regulators and target gene interactions.

- **ARACNe (Algorithm for the Reconstruction of Accurate Cellular Networks)**: ARACNe is used for inferring regulatory networks by identifying significant interactions between TFs and target genes.
  Description: ARACNe uses mutual information to detect TF-target interactions, filtering out indirect associations. It is highly effective in reconstructing TRNs by identifying direct regulatory relationships, contributing to understanding TF influence on cellular functions.

5. **Dynamic Network Analysis Tools:** Dynamic network analysis tools provide insights into changes in TRNs under different conditions or over time.

- **GENIE3 (GEne Network Inference with Ensemble of Trees)**: GENIE3 uses machine learning to infer regulatory networks from gene expression data by building an ensemble of regression trees.
  Description: GENIE3 predicts interactions based on gene expression patterns and is useful in dynamic analysis by capturing changes in regulatory networks across different conditions. It helps in identifying potential regulatory changes linked to environmental or experimental variations.
- **Inferelator**: Inferelator is a regression-based tool for inferring dynamic regulatory networks. It models TF-gene relationships by using time-series or perturbation data to predict how TRNs change in response to external stimuli.
  Description: Inferelator applies stability selection and regularization methods to reduce noise in data, making it effective for analysing large-scale dynamic regulatory networks. It provides insights into temporal regulatory changes and helps in identifying responsive elements within TRNs.

Conclusion

Analysing transcriptional regulatory networks requires a combination of tools that can handle diverse data types, including ChIP-Seq and RNA-Seq. Tools such as MACS and HOMER facilitate the identification of TF binding sites, while DESeq2 and EdgeR help profile gene expression influenced by TFs. For motif discovery, MEME and FIMO provide insights into TF binding preferences. Network analysis tools like Cytoscape and ARACNe aid in visualizing and reconstructing TRNs, while dynamic tools such as GENIE3 and Inferelator capture changes over time or in response to stimuli. Together, these tools provide a comprehensive framework for understanding the complex regulatory networks that control gene expression in cells, enhancing our understanding of cellular functions, developmental processes, and responses to environmental changes.

PRACTICE-2

REPORT ON ANALYSING THE INSULIN NETWORK USING THE STRING DATABASE

Aim: To analyse the insulin signalling network using the STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) database.

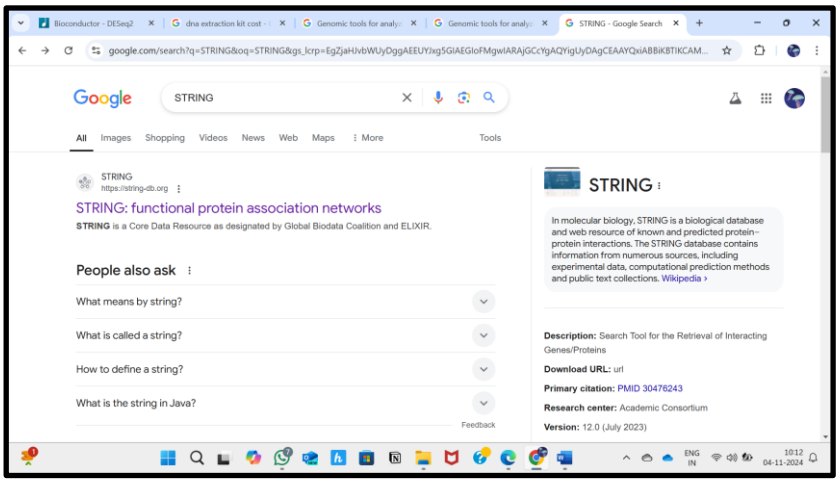
Introduction

The insulin signalling pathway is critical for regulating glucose metabolism, cell growth, and energy balance. Insulin binds to its receptor on target cells, triggering a cascade of downstream effects involving multiple proteins and signalling molecules. Disruptions in the insulin pathway are associated with various diseases, including diabetes and metabolic syndrome.

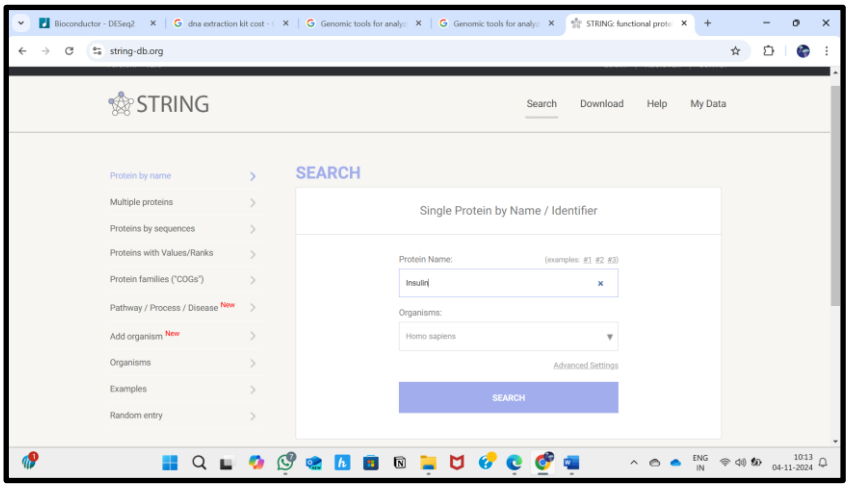
STRING is an online database designed for exploring PPIs, helping researchers study interactions at a systems level. It integrates data from various sources, including high-throughput experiments, computational predictions, and text mining, to create an extensive PPI network. By using STRING, we can visualize the insulin network and identify key proteins involved in insulin-mediated signalling pathways.

Procedure: Step-by-Step Analysis of the Insulin Network Using STRING

1. **Access the STRING Database**
   - Open a web browser and go to the STRING database website at https://string-db.org.

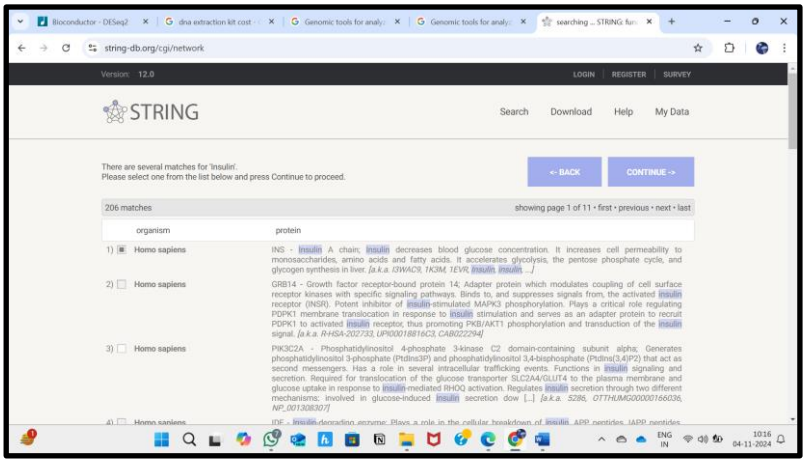


2. **Search for the Insulin Protein**
   - In the search bar on the STRING homepage, enter “insulin” (or "INS" if using the human protein symbol).
   - Select the appropriate species, such as *Homo sapiens* (human), from the drop-down list to ensure that only human-related interactions are shown.
   - Click "Search" to proceed.



3. **Explore the Insulin Protein Page**

   - After the search, you'll be directed to the insulin protein's page, which provides general information about insulin, including its function and known interactions.
   - STRING offers a variety of data types, including experimental data, curated interactions, and predictions, each contributing to the interaction score.



4. **Adjust Interaction Score Thresholds**

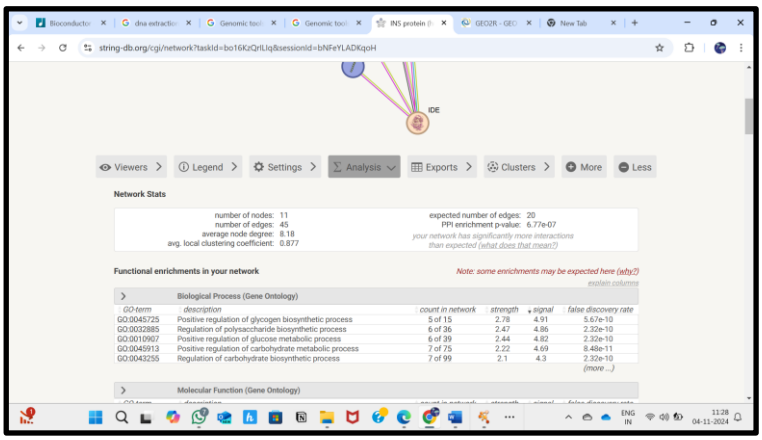
   - To refine the interaction network, you can adjust the confidence score threshold in the "Settings" panel on the right.
   - For a high-confidence network, set the score to "high confidence" (usually 0.7 or above). This filters out low-confidence interactions, displaying only reliable connections.

5. **Visualize the Protein Interaction Network**

   - STRING generates a network visualization displaying the interactions of insulin with other proteins.
   - Each node represents a protein, and the lines (edges) connecting them represent interactions.

- The edges vary in colour and thickness, indicating the interaction type and confidence level.

6. **Analyse the Interaction Types**
   - STRING provides various types of interactions, which you can view by clicking on the "Network" tab.
   - Interaction types include experimental interactions, co-expression, database annotations, text mining, and homology.
   - Hovering over each line reveals more details about the interaction source and confidence score.
7. **Expand the Insulin Network**
   - To explore additional layers of the network, use the "Expand network" option.
   - Adding more proteins provides insights into secondary interactions, allowing you to observe indirect connections and pathways that insulin may influence.
8. **Identify Key Pathway Components**
   - Examine high-confidence interactors that are part of known signalling pathways, such as the insulin receptor substrate (IRS), phosphoinositide 3-kinase (PI3K), and AKT proteins, which play essential roles in insulin signalling.
   - STRING also offers pathway enrichment analysis, identifying pathways enriched in the network, such as the PI3K-AKT pathway and the MAPK pathway, which are crucial for insulin's regulatory functions.
9. **Explore Functional Enrichment Analysis**
   - STRING includes tools for functional enrichment analysis to identify Gene Ontology (GO) terms, pathways, and protein families overrepresented in the network.
   - Go to the "Analysis" tab and select "Functional Enrichment" to view enriched biological processes, molecular functions, and cellular components associated with insulin and its interactors.

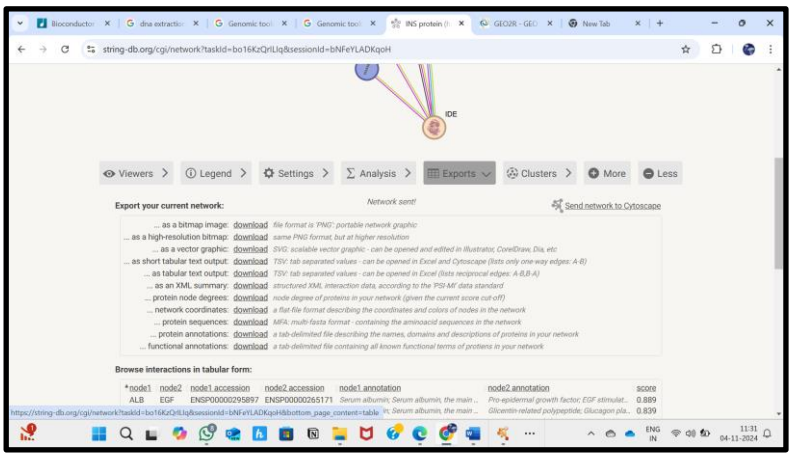


10. **Interpret the Network**

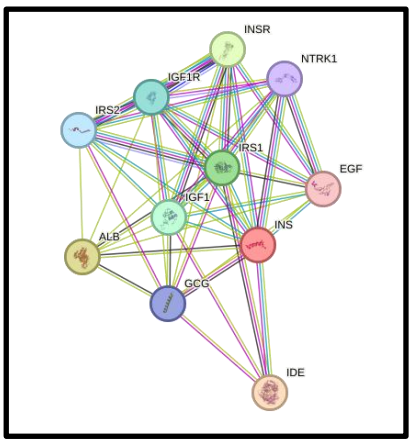
- Use the generated interaction map to interpret how insulin may influence various cellular processes through its protein interactions.
- Identify core interactors that may play a role in metabolic regulation, glucose transport, and cell growth, all of which are essential functions mediated by insulin signalling.

11. **Save or Export the Results**

- STRING provides options to download the network data and visualizations in various formats.
- Use the "Export" option to save your network as an image or download the interaction data for further analysis in external tools like Cytoscape.



Result



Explanation of some of the key components:

1. **INS (Insulin)** - Central protein in this network, involved in glucose uptake and metabolism regulation.
2. **IGF1 and IGF1R (Insulin-like Growth Factor 1 and its Receptor)** - Related to cell growth and development, these proteins interact closely with insulin signalling components.

3. **INSR (Insulin Receptor)** - This receptor binds insulin and initiates downstream signalling, leading to various metabolic effects.
4. **IRS1 and IRS2 (Insulin Receptor Substrate 1 and 2)** - These are adaptor proteins that mediate signalling from INSR to downstream effectors.
5. **EGF (Epidermal Growth Factor)** and **EGFR (Epidermal Growth Factor Receptor)** - These are involved in cell growth and differentiation and also interact with insulin signalling.
6. **IDE (Insulin-Degrading Enzyme)** - Breaks down insulin and modulates insulin signalling duration.
7. **NTRK1** - This receptor interacts with neurotrophins, potentially linking insulin signalling to neuronal growth.
8. **ALB (Albumin)** and **GCG (Glucagon)** - These proteins indirectly relate to glucose metabolism and insulin signalling.

The coloured lines connecting the proteins represent different types of interactions, such as:

- **Green lines** for activation
- **Blue lines** for binding
- **Pink lines** for post-translational modifications
- **Yellow lines** for co-expression relationships
- **Black lines** for reaction links

This network highlights how insulin and growth factor signalling pathways are interconnected, with multiple points of interaction across metabolic and growth-related processes.

Conclusion

The STRING database enables a comprehensive analysis of the insulin network, providing insights into its direct and indirect interactions with other proteins. By adjusting interaction scores and exploring pathway enrichment, researchers can gain a better understanding of the insulin signalling pathway's components and the proteins involved in its regulatory functions. Such analyses offer valuable information for studying the pathophysiology of metabolic diseases and identifying potential therapeutic targets within the insulin network.

PRACTICE-3

Analysing the Network of Breast Cancer (GSE80751) Using STRING

Aim: To conduct a comprehensive analysis of the gene expression profiles associated with breast cancer using the GEO dataset GSE80751. This includes identifying differentially expressed genes (DEGs) between control and diseased samples, extracting gene-related information, and analysing the interaction network of the top 100 DEGs using the STRING database.

Introduction:

Breast cancer, one of the most common cancers in women globally, results from a complex mix of genetic, environmental, and lifestyle factors. Understanding molecular changes in breast cancer is essential for exploring its development and creating targeted treatments. High-throughput techniques like microarray expression profiling help researchers analyse large-scale gene expression changes in breast cancer. The GEO dataset GSE80751 provides gene expression data from breast cancer and normal tissues, enabling comparative analysis. Using GEO2R for differential expression analysis, researchers can identify significantly upregulated or downregulated genes. Additionally, constructing a gene interaction network via STRING reveals relationships among these genes, highlighting pathways potentially involved in cancer progression.

METHODOLOGY:

1. Data Acquisition and Preprocessing: The gene expression dataset GSE80751 was accessed from the Gene Expression Omnibus (GEO). The dataset comprises samples from breast cancer patients, including both diseased and control groups.

2. Differential Expression Analysis Using GEO2R

- **Selection of Samples:** A total of 14 samples were selected from GSE80751, consisting of 7 control samples and 7 breast cancer diseased samples.
- **Analysis Tool:** GEO2R was utilized to identify differentially expressed genes (DEGs) between the control and diseased samples.
- **Statistical Method:** The analysis involved calculating p-values and fold changes to determine the significance of the gene expression differences.

3. Extraction of Gene Information

- **Excel Sheet Extraction:** The results from GEO2R were exported to an Excel sheet, containing multiple gene IDs, gene symbols, and their corresponding descriptions.
- **Data Format:** The sheet included columns for gene ID, gene symbol, and a brief description of each gene.

4. Network Analysis Using STRING:

- **Top 100 Genes:** The top 100 differentially expressed genes were selected based on adjusted p-values and fold changes.

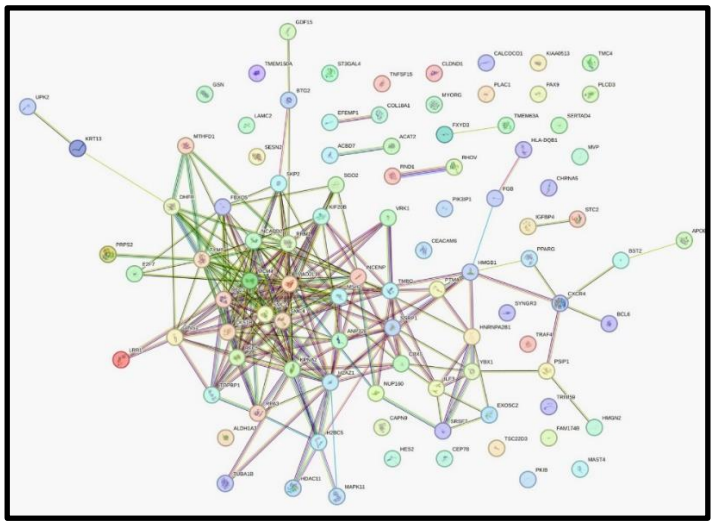
- **STRING Database:** The STRING database (Search Tool for the Retrieval of Interacting Genes/Proteins) was used to analyse the interaction network of these genes.
- **Network Construction:** The network was constructed to visualize interactions and associations between the selected genes, which can provide insights into the biological pathways involved in breast cancer.

5. Differentially Expressed Genes: The GEO2R analysis revealed a list of differentially expressed genes (DEGs) between control and diseased samples. The top 100 genes were identified based on the criteria set.

6. STRING Network Analysis:

- **Network Visualization:** The interaction network for the top 100 genes was visualized using the STRING database. The network highlights the relationships between genes, including direct and indirect interactions.
- **Key Findings:** Important pathways and clusters were identified within the network, which can help elucidate the molecular mechanisms underlying breast cancer.

Results:



- **Clustered Subnetworks:** Distinct clusters or modules are observed, indicating groups of proteins that are densely interconnected. These clusters likely represent functional groups or complexes involved in related biological pathways, such as DNA repair, cell cycle regulation, or transcriptional control.
- **Peripheral Proteins:** On the outskirts, proteins like UBP2, KRT18, CXCR6, and APOE exhibit fewer connections, suggesting specialized roles or transient interactions. These peripheral proteins may participate in specific signalling pathways or function as receptors or ligands in cellular communication.
- **Types of Interactions:** The presence of various edge colours typically signifies different interaction types—experimental, predicted, and co-expression data—indicating a well-validated network. Heavily interconnected regions, characterized by multiple colours, suggest that these interactions are supported by various data sources, enhancing their reliability.

PRACTICE-4

Analysis of Breast Cancer Genes Using Cytoscape and CytoHubba

Aim: To analyse breast cancer-related genes using Cytoscape and CytoHubba, identifying key genes with high centrality values—degree, closeness, betweenness, and clustering coefficient—to uncover potential hub genes important for breast cancer progression and possible treatment targets.

Introduction:

Breast cancer is one of the most prevalent cancers worldwide, with a complex molecular basis involving numerous genetic alterations. Network-based analysis provides insights into the functional interactions between proteins, highlighting key regulators in cancer-related pathways.

Cytoscape is a powerful tool for visualizing complex networks and analysing biological pathways. In this study, we use the CytoHubba plugin within Cytoscape, which enables centrality analysis to identify hub genes. Centrality measures such as degree, closeness, and betweenness are used to rank genes based on their importance within the network. These measures can indicate genes that play significant roles in regulating other genes in cancer pathways, thus identifying potential biomarkers for breast cancer.

3. METHODOLOGY:

Step 1: Data Preparation

1. **Download Gene Interaction Data**: Obtain a breast cancer-related gene dataset from the STRING database. Export the file as a TSV (tab-separated values) file.
2. **Convert TSV to CSV**: Open the TSV file in a spreadsheet editor (like Excel) and save it as a CSV file for easier importing into Cytoscape.

Step 2: Network Construction in Cytoscape

1. **Import CSV File**:
   - Open Cytoscape, go to File > Import > Network from Table (Text File) and upload the CSV file to visualize the gene interaction network.
   - **Nodes:** represent ten biological entities (e.g.: genes or proteins)
   - **Edges:** Define interactions between these nodes based on hypothetical relationships.

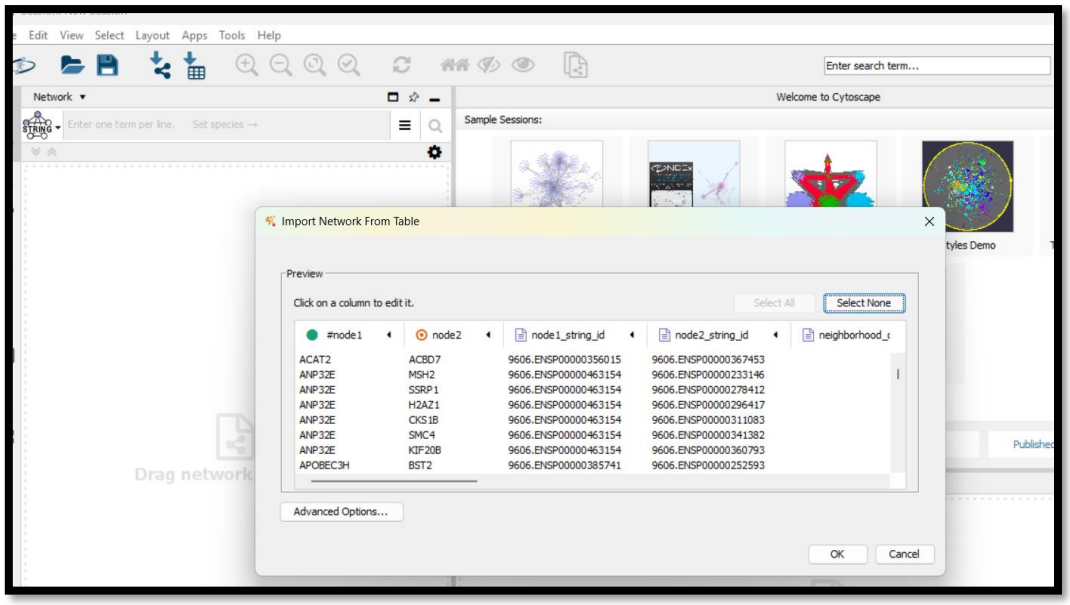


Fig 1: Imported the network from a table, displaying nodes and their targets.

2. **Visualize the Network**: Customize the network appearance for clear visualization by adjusting node and edge colours based on attributes (e.g., gene expression levels, interaction confidence).

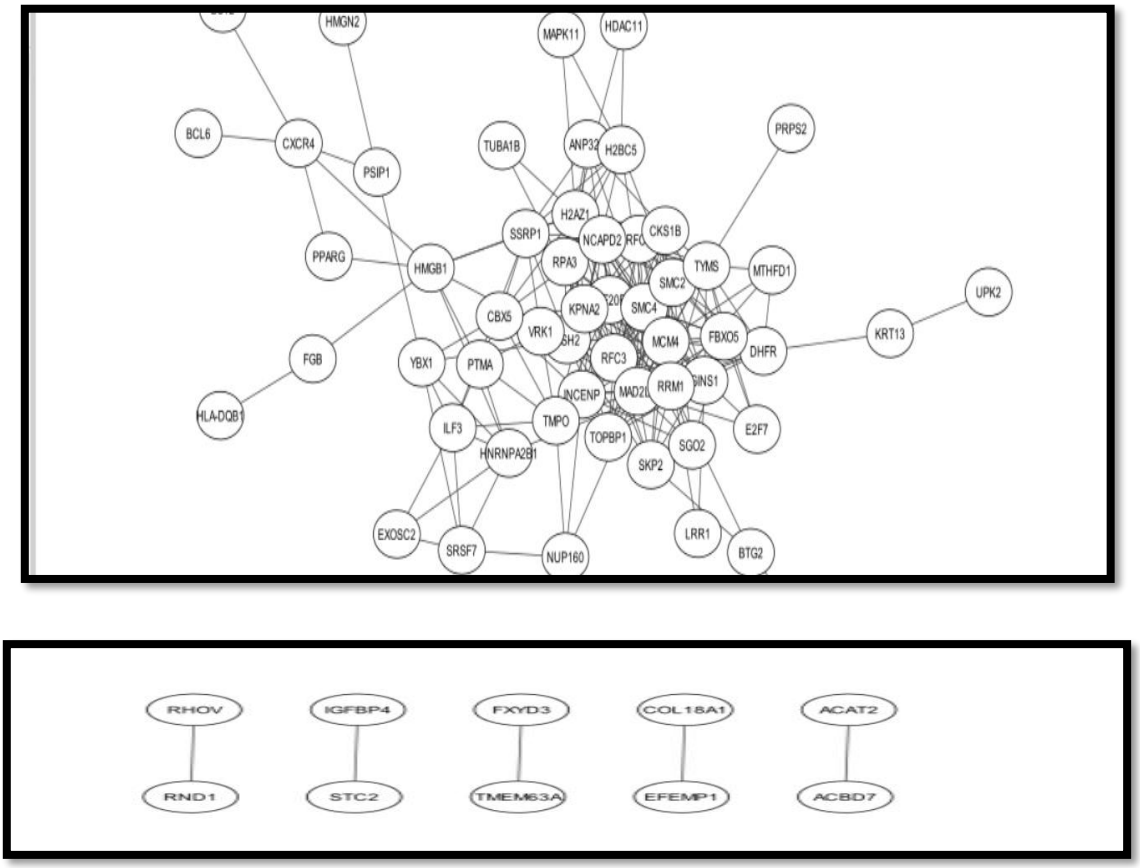


Fig 2: Network Construction and their interactions of nodes and edges

Step 3: CytoHubba Analysis

1. **Load CytoHubba Plugin**: Go to Apps > CytoHubba in Cytoscape to launch the plugin.

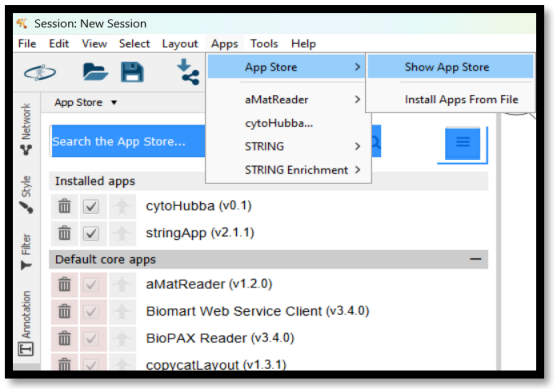


Fig3: Launching the CytoHubba app for network analysis.

2. **Choose Centrality Methods**:

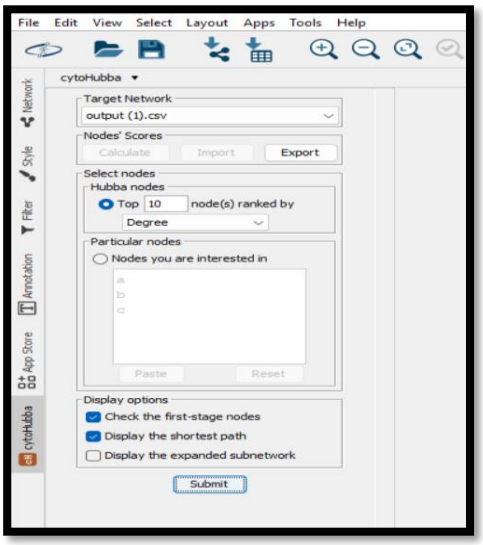


Fig4: Selection of the top 10 nodes from a complex network using a chosen network analysis method.

- Use CytoHubba to apply different centrality methods:
  - **Degree Centrality**: Identifies highly connected genes, potentially critical for information flow.
  - **Closeness Centrality**: Measures how quickly a gene can interact with others, indicating its central role in pathways.
  - **Betweenness Centrality**: Highlights genes that control information passage between clusters.
  - **Clustering Coefficient**: Measures the likelihood of connected genes forming clusters, pointing to dense interaction regions.

3. **Run the Analysis**:
   - Select each method individually and run the analysis to generate centrality scores for each gene.
   - Save the results for each method as a CSV file.

Step 4: Interpretation and Ranking of Hub Genes

1. **Identify Key Genes**:
   - Based on the centrality scores, identify genes with high scores across multiple measures as potential hub genes.
   - Highlight genes with high clustering coefficients, as these might indicate genes involved in dense interaction clusters relevant to breast cancer pathways.
2. **Visualization of Hub Genes**:
   - Highlight or filter out the top-ranked genes in Cytoscape's network view to emphasize their roles.

Results:

The results section should summarize the findings of your CytoHubba analysis. Use tables and graphs to present the centrality measures for each gene:

1. **Centrality Measures**: Summarize the key findings for each centrality measure (degree, closeness, betweenness, clustering coefficient) using tables. Highlight the top-ranked genes for each measure. Centrality measures help rank genes based on their roles in the network, allowing us to identify key "hub" genes that may influence cancer progression.
2. **Identification of Hub Genes**: List the hub genes with the highest centrality scores across multiple measures. These genes could represent potential biomarkers or therapeutic targets in breast cancer.

Degree:

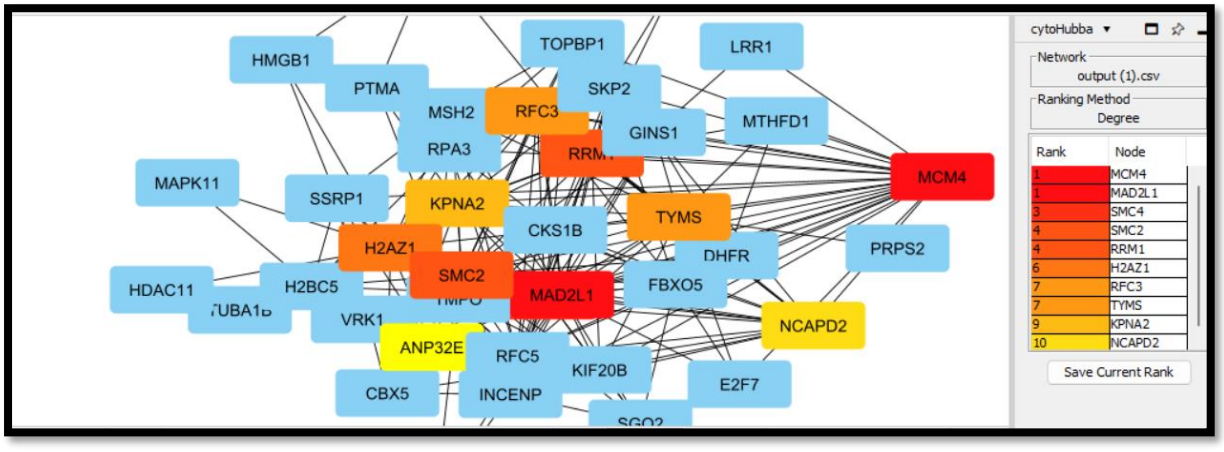


Fig 5: Top 10 genes identified by degree centrality.

- Using the degree centrality method in CytoHubba, we identified the top 10 hub genes in the breast cancer network based on the number of direct connections each gene has. The higher the degree centrality, the more connections a gene has, suggesting a critical role in network stability and potentially significant influence over other genes in the network.
- These genes, highlighted in the network visualization, display varying levels of connectivity, with **MCM4** showing the highest degree centrality. This indicates that **MCM4** may act as a crucial hub in the breast cancer network, potentially playing a significant role in regulating gene interactions related to cancer progression.

Closeness Centrality

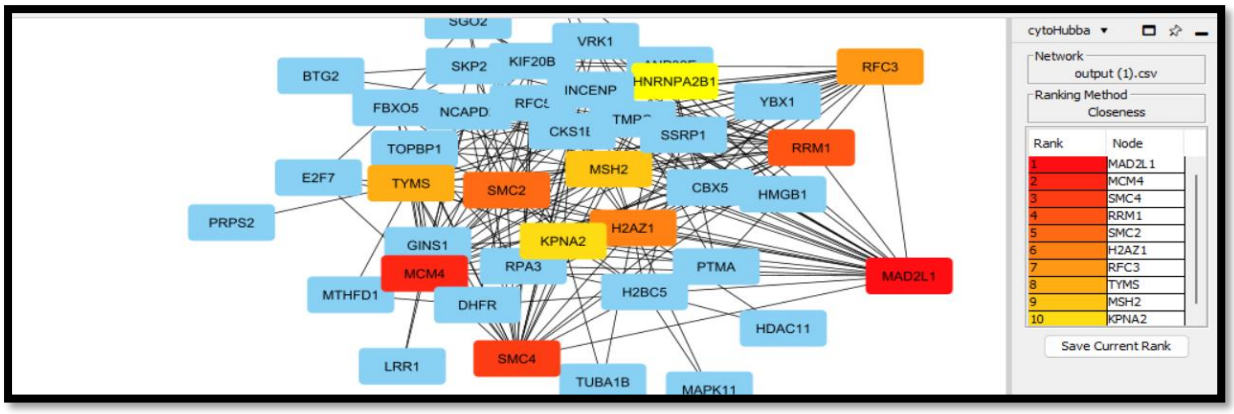


Fig 6: Top 10 genes identified by Closeness centrality.

- This indicates that **MAD 2L 1** may act as a crucial hub in the breast cancer network, potentially playing a significant role in regulating gene interactions related to cancer progression.
- Other genes, such as **MCM4** and **SMC4**, also show high connectivity and may serve as essential regulators or biomarkers in breast cancer pathways.

Betweenness Centrality:

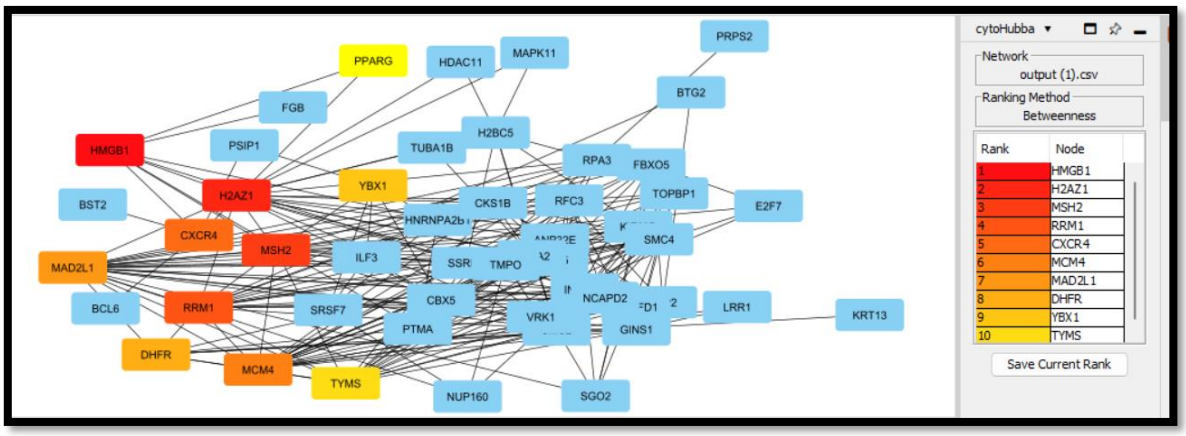


Fig 7: Top 10 genes identified by Betweenness Centrality

- This indicates that **HMGB1** may act as a crucial hub in the breast cancer network, potentially playing a significant role in regulating gene interactions related to cancer progression.
- Other genes, such as **H2AZ1** and **MSH2**, also show high connectivity and may serve as essential regulators or biomarkers in breast cancer pathways.

Clustering Coefficient:

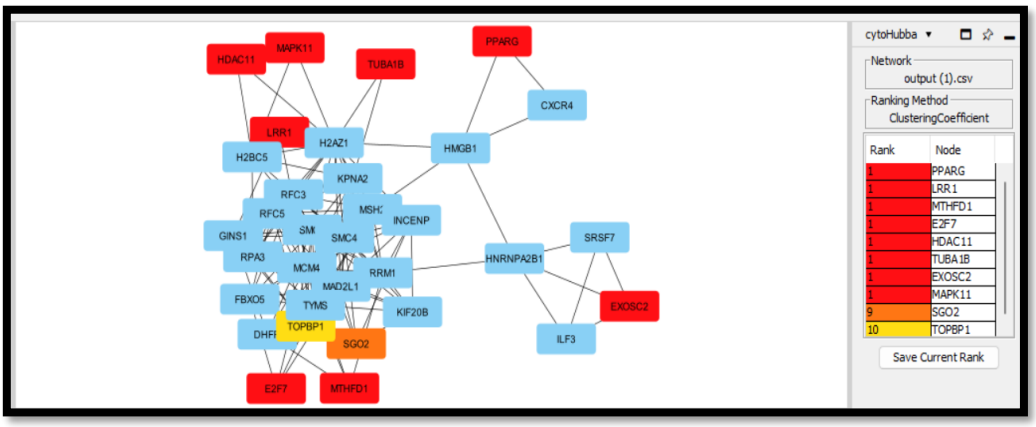


Fig 8: Top 10 genes identified by Clustering Coefficient

- This indicates that **PPARG** may act as a crucial hub in the breast cancer network, potentially playing a significant role in regulating gene interactions related to cancer progression. Other genes, such as **LRR1** and **IMTHFD 1,** also show high connectivity and may serve as essential regulators or biomarkers in breast cancer pathways.

<u>Conclusion:</u>

- This study utilized the degree centrality method in CytoHubba to identify key hub genes within the breast cancer network. By ranking genes based on the number of direct connections, we highlighted potential regulatory hubs that may play significant roles in cancer progression and gene interactions. Key genes such as **MCM4**, **MAD2L1**, **HMGB1**, and **PPARG** emerged as central hubs, suggesting that they may contribute to network stability and influence various cancer-related pathways.
- Other genes, including **SMC4**, **H2AZ1**, **MSH2**, **LRR1**, and **MTHFD1**, also demonstrated high connectivity, reinforcing their potential as critical regulators or biomarkers for breast cancer. The varying levels of connectivity among these genes highlight their diverse roles in breast cancer pathways, underscoring the complexity of gene interactions in cancer biology.
- Overall, these findings provide a foundation for further research into the functional significance of these hub genes. The identification of highly connected genes through network analysis could guide future studies on their roles in cancer progression, potentially leading to novel therapeutic targets and biomarkers for breast cancer diagnosis and treatment.

Practice-5

Identification and Analysis of Cancer-Related Genes Using Gene Cards Database And performing Gene Ontology (GO) using PANTHER Tool

Aim: To utilize the GeneCards database to identify key cancer-related genes, with a focus on understanding the roles of BRCA1 and BRCA2 in cancer susceptibility and DNA repair mechanisms followed by performing Gene Ontology using PANTHER Tool.

Description:

GeneCards is a comprehensive database that provides detailed information on human genes, including their functions, associated disorders, pathways, and molecular data. Widely used in biomedical research, GeneCards integrates data from multiple sources, making it a valuable tool for understanding gene roles in health and disease. In cancer research, GeneCards allows researchers to explore genes involved in cancer development, progression, and treatment by providing access to gene-specific information such as protein function, molecular interactions, and clinical relevance.

The Gene Ontology (GO) is a standardized vocabulary used to describe the functions of genes and gene products. GO annotations enable researchers to consistently categorize and interpret the functions of genes and proteins, supporting comparative and functional genomics. To facilitate these analyses, the PANTHER (Protein ANalysis THrough Evolutionary Relationships) tool is frequently used alongside GO for functional classification and enrichment analysis.

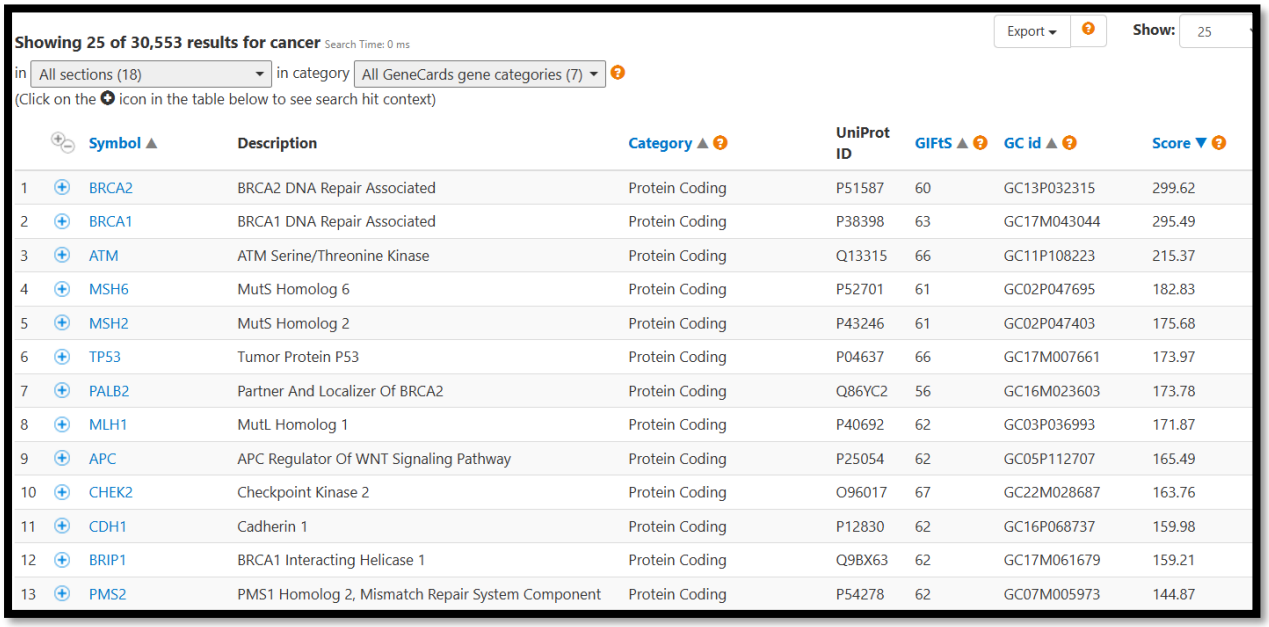
Methodology:

GeneCards Search:

- Open GeneCards and use "cancer" as the search term.

Fig : Searching for cancer-related genes in the query box

- Review the search results, which should display a list of genes associated with cancer.
- Identify and copy the top 10 genes from the GeneCards search results. These genes are likely the most researched or relevant to cancer biology.



Showing 25 of 30,553 results for cancer Search Time: 0 ms

Export Show: 25

in All sections (18) in category All GeneCards gene categories (7)

(Click on the icon in the table below to see search hit context)

	Symbol	Description	Category	UniProt ID	GIFtS	GC id	Score
1	BRCA2	BRCA2 DNA Repair Associated	Protein Coding	P51587	60	GC13P032315	299.62
2	BRCA1	BRCA1 DNA Repair Associated	Protein Coding	P38398	63	GC17M043044	295.49
3	ATM	ATM Serine/Threonine Kinase	Protein Coding	Q13315	66	GC11P108223	215.37
4	MSH6	MutS Homolog 6	Protein Coding	P52701	61	GC02P047695	182.83
5	MSH2	MutS Homolog 2	Protein Coding	P43246	61	GC02P047403	175.68
6	TP53	Tumor Protein P53	Protein Coding	P04637	66	GC17M007661	173.97
7	PALB2	Partner And Localizer Of BRCA2	Protein Coding	Q86YC2	56	GC16M023603	173.78
8	MLH1	MutL Homolog 1	Protein Coding	P40692	62	GC03P036993	171.87
9	APC	APC Regulator Of WNT Signaling Pathway	Protein Coding	P25054	62	GC05P112707	165.49
10	CHEK2	Checkpoint Kinase 2	Protein Coding	O96017	67	GC22M028687	163.76
11	CDH1	Cadherin 1	Protein Coding	P12830	62	GC16P068737	159.98
12	BRIP1	BRCA1 Interacting Helicase 1	Protein Coding	Q9BX63	62	GC17M061679	159.21
13	PMS2	PMS1 Homolog 2, Mismatch Repair System Component	Protein Coding	P54278	62	GC07M005973	144.87

Fig: Cancer-related genes with their IDs and descriptions

BRCA1

- **Description**: BRCA1 DNA Repair Associated
- **Category**: Protein Coding
- **UniProt ID**: P38398
- **GIFtS**: 63
- **GC ID**: GC17M043044
- **Score**: 295.49

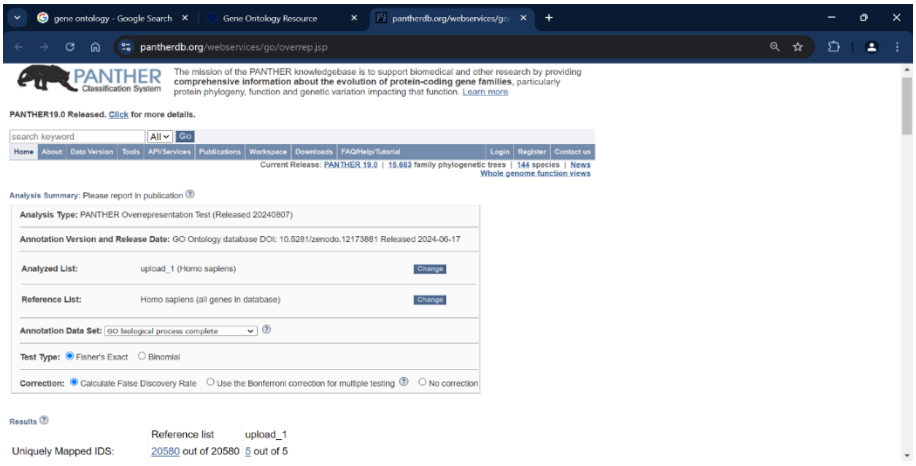
BRCA2

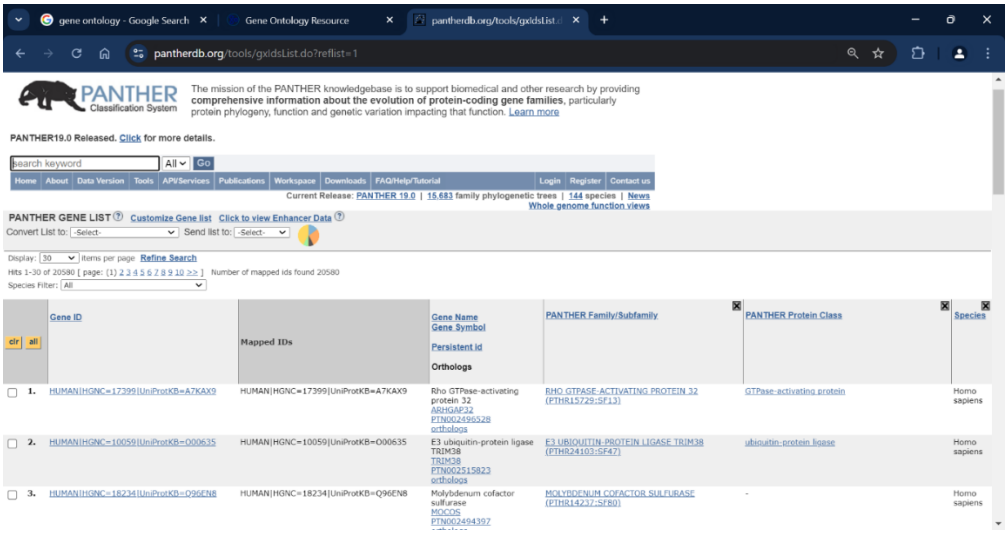
- **Description**: BRCA2 DNA Repair Associated
- **Category**: Protein Coding
- **UniProt ID**: P51587
- **GIFtS**: 60
- **GC ID**: GC13P032315
- **Score**: 299.62

Gene Ontology

- **Access the GO website:** Go to http://geneontology.org/ or search for "Gene Ontology" in your web browser.
- **Search for GO terms or gene products:** Use the search bar at the top of the page to search for specific GO terms or gene products. You can enter terms or use the dropdown menus to filter your search results.

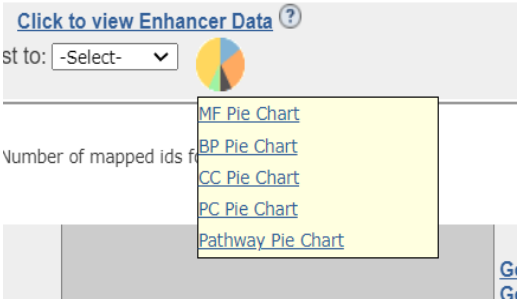
- **Paste the gene list:** Enter the list of genes you want to analyse into the designated field.
- **Select "uniquely mapped IDs":** This option will focus on genes that were successfully mapped to GO terms, providing more accurate enrichment analysis results.
- **View the results:** The results page will display a list of enriched GO terms or protein domains, along with statistical significance values.

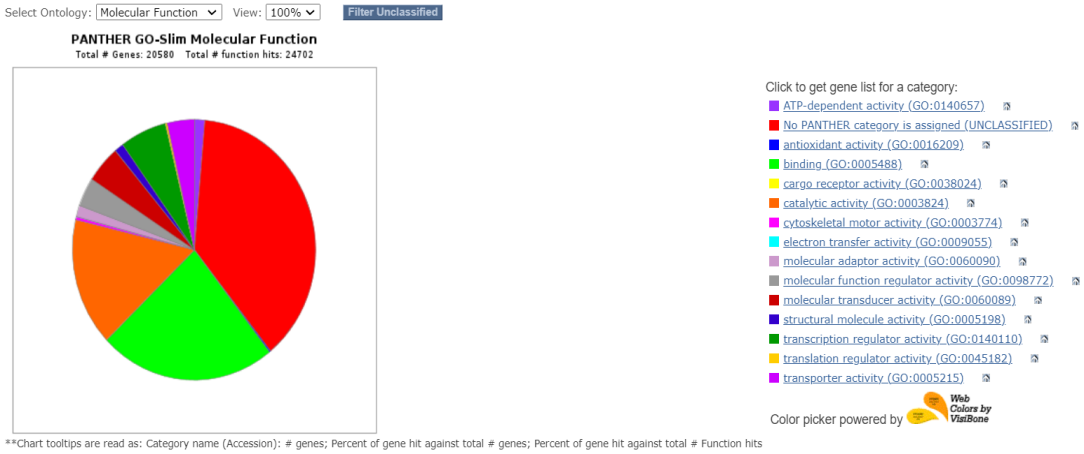


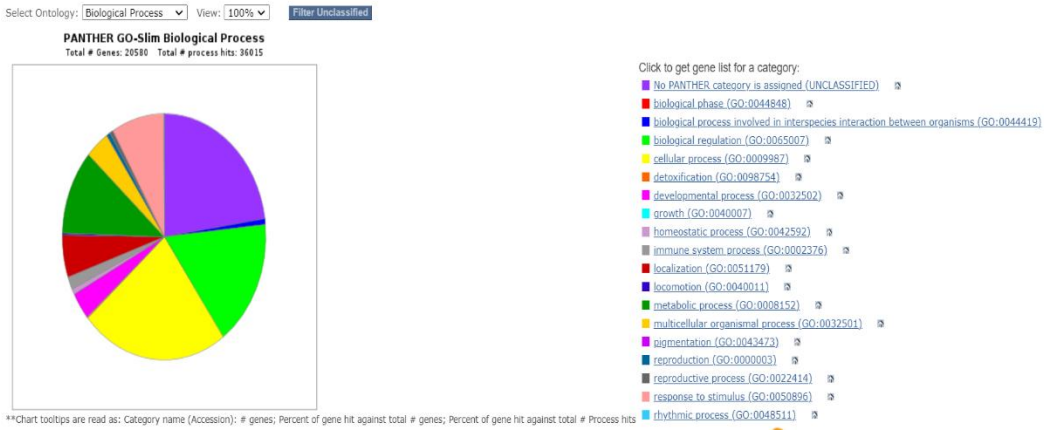


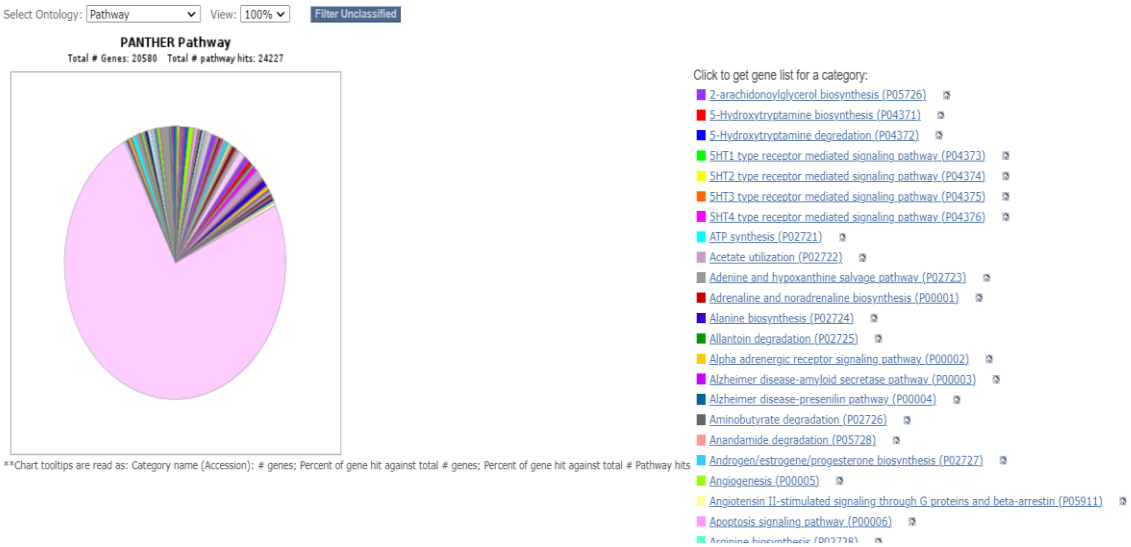
- **Select the chart type:** Click on the dropdown menu and choose the desired chart type. The available options are:

  - **MF Pie Chart:** This chart shows the distribution of genes based on their molecular functions.
  - **BP Pie Chart:** This chart shows the distribution of genes based on their biological processes.
  - **CC Pie Chart:** This chart shows the distribution of genes based on their cellular components.
  - **PC Pie Chart:** This chart shows the distribution of genes based on their protein classes.
  - **Pathway Pie Chart:** This chart shows the distribution of genes based on the pathways they are involved in.

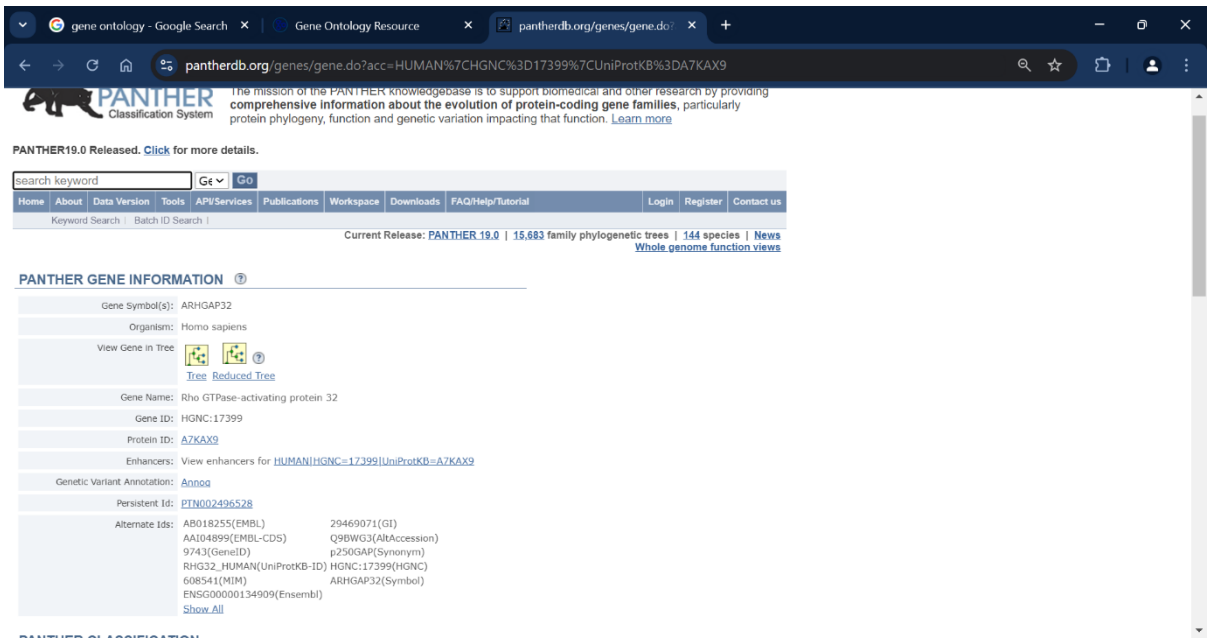








- Analyse its function, pathways, interactions, expression patterns, genetic variants, and evolutionary history.



Conclusion:

The GeneCards tool helps in extracting the significant genes responsible for the specific diseases, which provides insights for further studies. Using the PANTHER Tool for Gene Ontology (GO) analysis enabled us to identify key biological processes, molecular functions, and cellular components associated with our gene set. Enrichment analysis highlighted significant pathways, offering insights into the functional roles and biological relevance of these genes. PANTHER's evolutionary framework further enhanced our understanding, providing a solid foundation for future research and experimental validation.

Practice-6

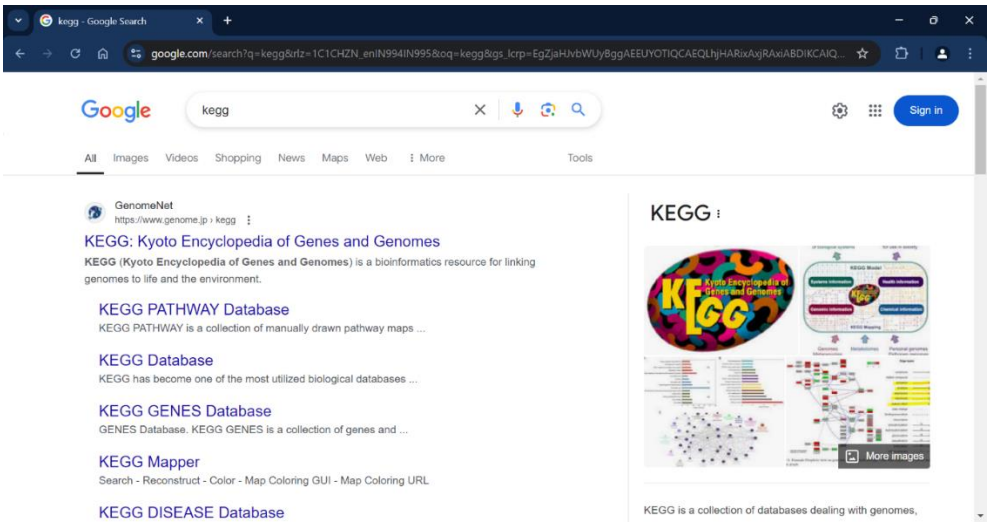
KEGG

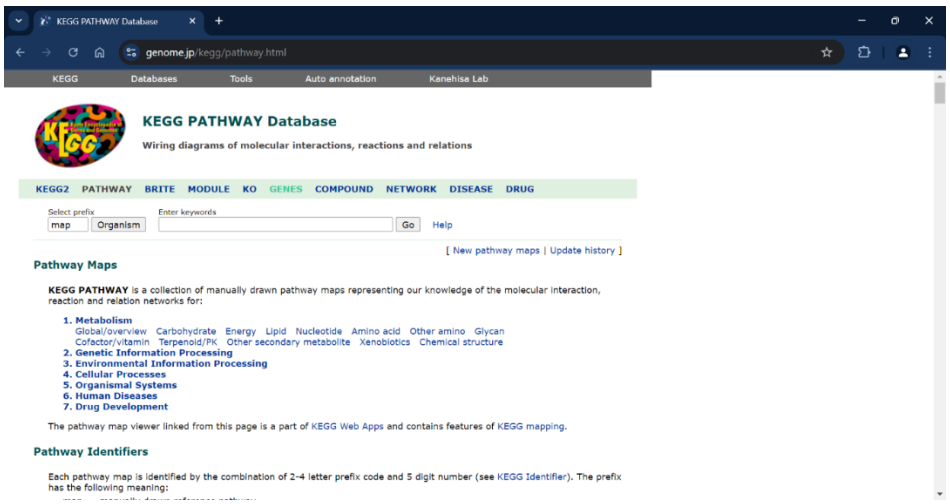
Aim: To conduct pathway analysis for the genes using KEGG Pathway Database.

Description: Kyoto Encyclopaedia of Genes and Genomes (KEGG) is a collection of online databases dealing with genomes, enzymatic pathways, and biological chemicals. KEGG is a comprehensive, high-quality database and resource widely used in bioinformatics and systems biology for understanding biological systems, molecular networks, and gene functions. Developed by the Kanehisa Laboratories in Japan, KEGG integrates information from genomics, chemical, and health-related data to create a knowledge base that facilitates the interpretation of large-scale datasets.

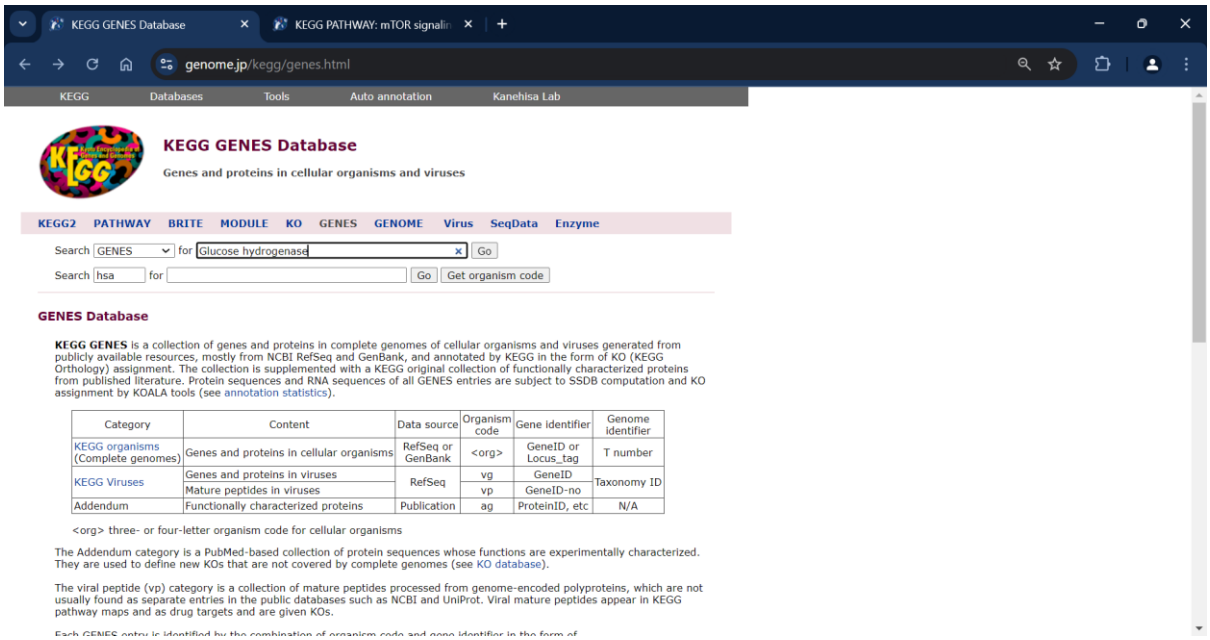
Methodology:

- Go to the KEGG website: https://www.genome.jp/kegg/
- **DATABASE OF KEGG** maintain 6 main databases: KEGG Pathway (Atlas) | KEGG Genes | KEGG Genome | KEGG Ligand | KEGG BRITE | KEGG Cancer
- Select "**GENES**" from the top menu.

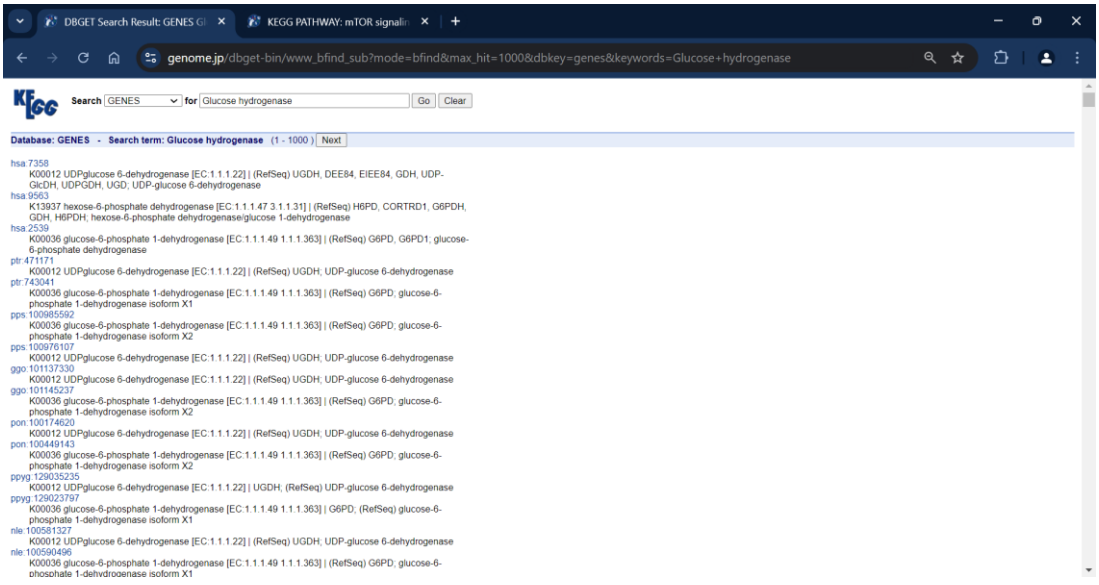




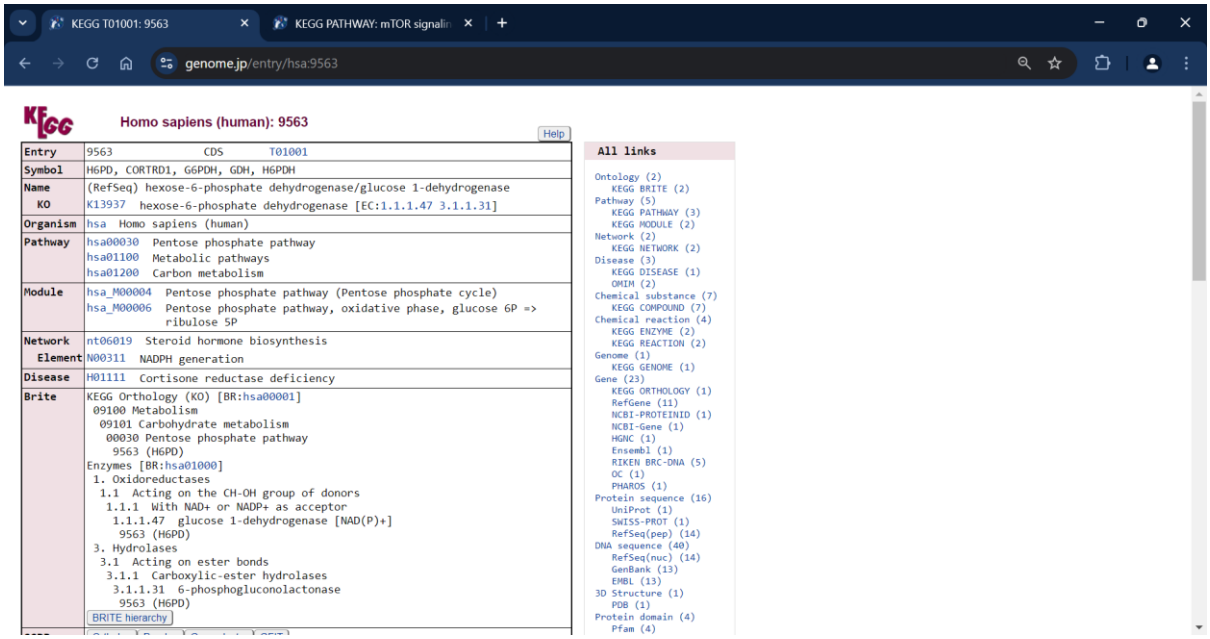
- **Search for genes:** Use the search bar to enter the organism's name or keyword you want to search for. Click "Go" to initiate the search.



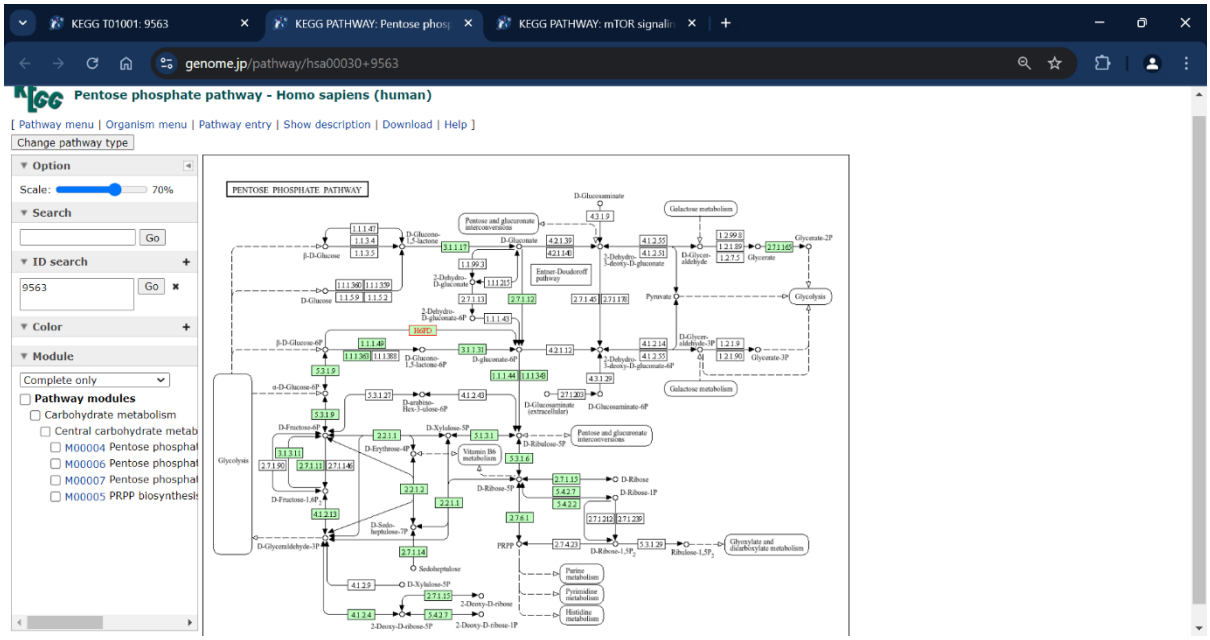
- **Select a specific gene:** Choose one of the genes listed in the search results. You can click on the gene's name to view its details.
- The gene details page will provide information about the gene's function, location, and related pathways.



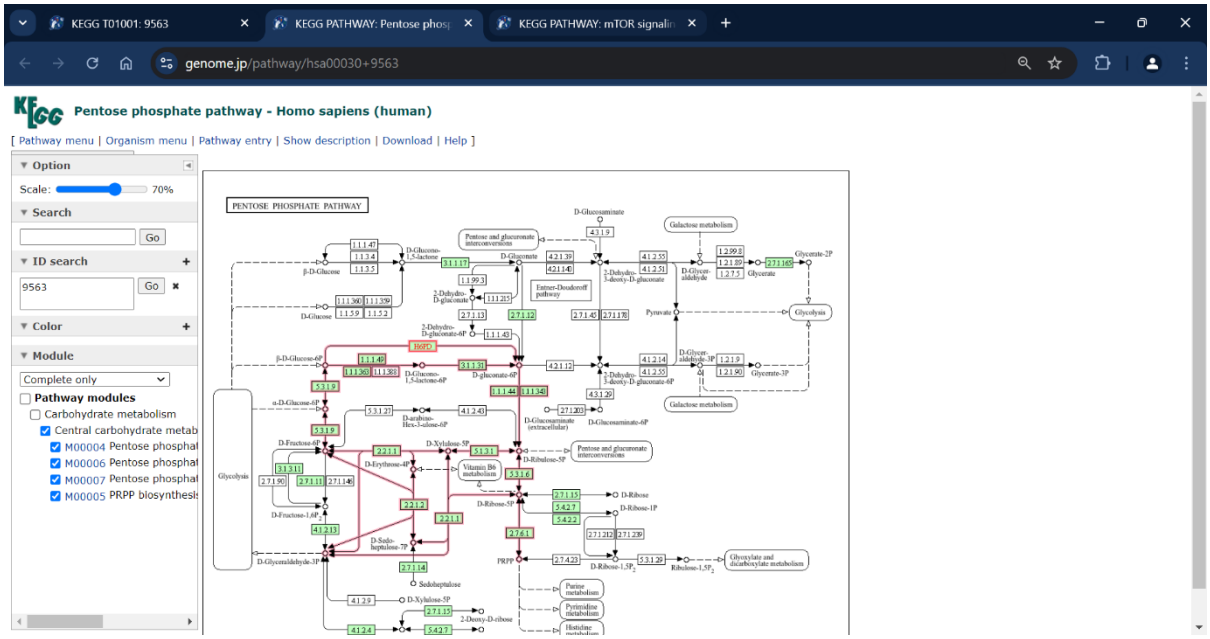
- Once you select a gene, examine its name, EC number, and function to understand its role.
- Use the "Pathway" section to identify the pathways the gene is involved in, revealing its broader biological context.
- Select a pathway from the given results.



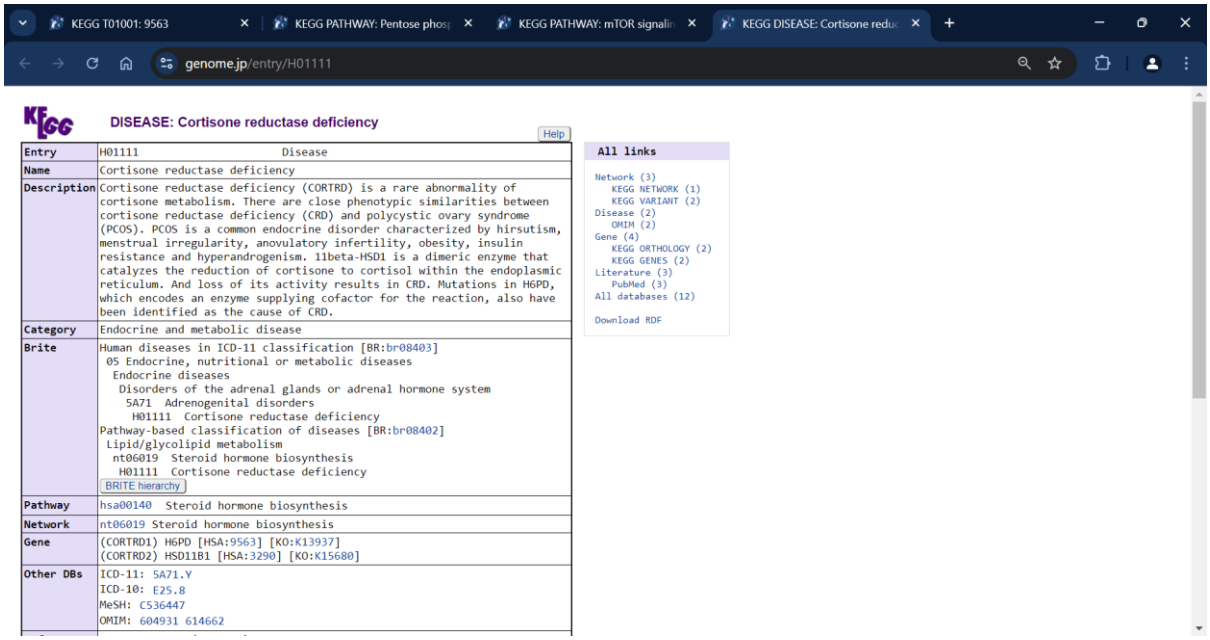
- The selected pathway with be displayed and information will appear on the right side of the screen, showing its scale, ID, modules, and other relevant details.



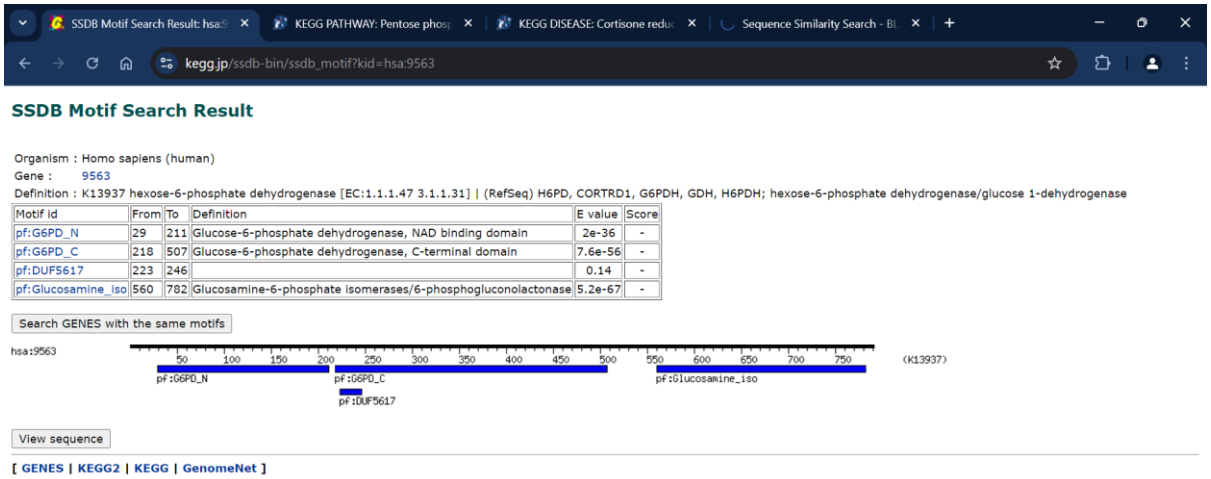
- **Select a module:** Click on any module in the pathway map. The module will be highlighted, and its details will be displayed in the information panel.
- **Explore other modules:** You can continue to select different modules to highlight them and explore their relationships within the pathway.



- **Identify associated diseases:** Check the "Disease" section to see if the gene is linked to any human health conditions.



- **Identify motifs:** The motifs found in the particular gene, which are protein domains associated with specific functions.
- **Explore related genes:** The "Search GENES with the same motifs" section allows you to find other genes that contain these motifs, suggesting potential functional similarities.
- Amino Acid Sequence and "DB search" indicate a Blast search.



AA seq 791 aa [AA seq] [DB search]

```
MWNMLIVAMCLALLGCLQAQELQGHVSIILLGATGDLAKKYLWQGLFQLYLDEAGRGHSF
SFHGAALTAPKQGQELMAKALESLSCPKDMAPSHCAEHKDQFLQLSQYRQLKTAEDYQAL
NKDIEAQLQHAGLREAGRIFYFSVPPFAYEDIARNINSSCRPGPGAWLRVVLEKPFGHDH
FSAQQLATELGTFFQEEEMYRVDHYLGKQAVAQILPFRDQNRKALDGLWNRHHVERVEII
MKETVDAEGRTSFYEEYGVIRDVLQNHLTEVLTLVAMELPHNVSSAEAVLRHKLQVFQAL
RGLQRGSAVVGQYQSYSEQVRRELQKPDSFHSLTPTFAAVLVHIDNLRWEGVPFILMSGK
ALDERVGYARILFKNQACCVQSEKHWAAAQSQCLPRQLVFHIGHGDLGSPAVLVSRNLFR
PSLPSSWKEMEGPPGLRLFGSPLSDYYAYSPVRERDAHSVLLSHIFHGRKNFFITTENLL
ASWNFWTPLLESLAHKAPRLYPGGAENGRLLDFEFSSGRLFFSQQQPEQLVPGPGPAPMP
SDFQVLRAKYRESPLVSAWSEELISKLANDIEATAVRAVRRFGQFHLALSGGSSPVALFQ
QLATAHYGFPWAHTHLWLVDERCVPLSDPESNFQGLQAHLLQHVRIPYYNIHPMPVHLQQ
RLCAEEDQGAQIYAREISALVANSSFDLVLLGMGADGHTASLFPQSPTGLDGEQLVVLTT
SPSQPHRRMSLSLPLINRAKKVAVLVMGRMKREITTLVSRVGHEPKKWPISGVLPHSGQL
VWYMDYDAFLG
```

Conclusion:

In conclusion, this pathway analysis using the KEGG Pathway Database provided valuable insights into the functional roles and interactions of the analysed genes. The identification of enriched pathways has shed light on the cellular processes these genes are involved in, facilitating a better understanding of their roles in health and disease contexts. These findings can aid in prioritizing pathways for further study, potentially leading to the identification of novel therapeutic targets.

Practice-7

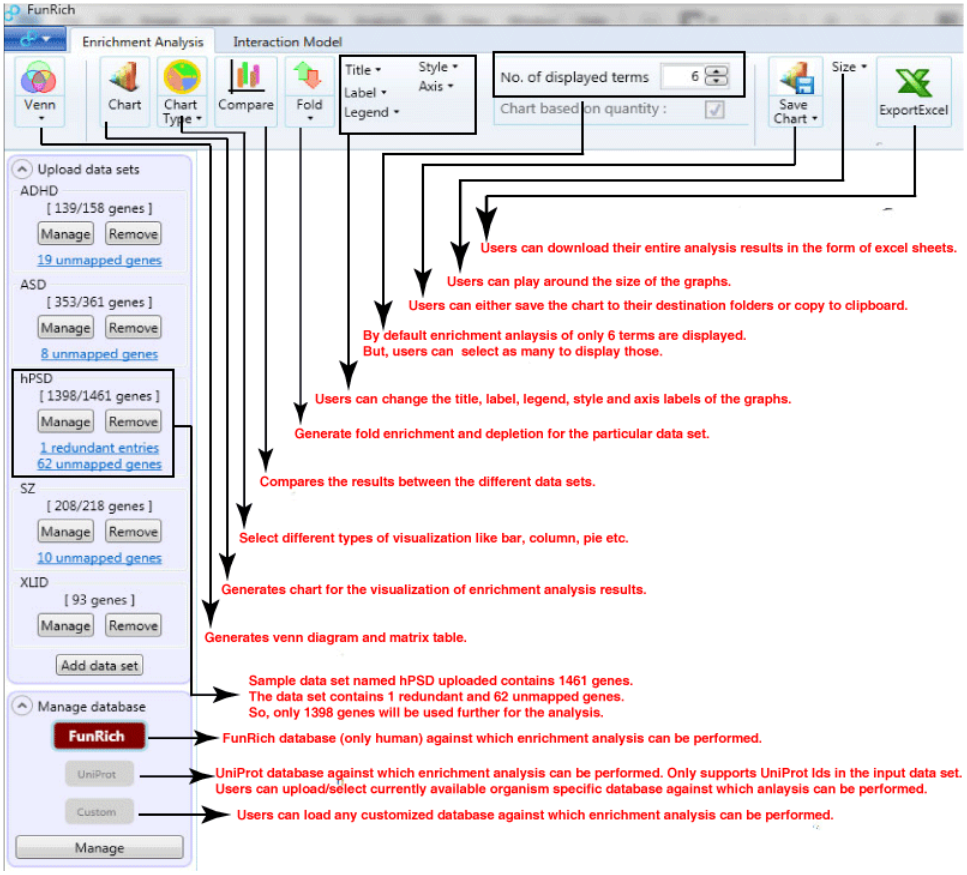
FUNRICH SOFTWARE

Aim: To perform functional annotation studies using Funrich Software

Highlights of FunRich:

- The results of the enrichment analysis can be visualized in a variety of graphical formats like column and bar graphs, pie chart, Venn diagram, heatmap and doughnut charts.
- The user can customize the graphs by changing/editing the legends, axis labels, colors, styles, fonts and title of the graphs in FunRich.
- The number of enriched terms displayed on the graphs can be selected, controlled and sorted.
- Users can perform enrichment analysis against custom background database irrespective of species.
- Interaction network analysis can be performed by using FunRich and custom database options.

Overview of features in FunRich



Type of query/input list

FunRich accepts following type of genes and protein ids separated by new lines for analysis.

1. Official gene symbols (for example: BTK, BRCA1, EGFR)

2. Entrez Ids (for example: 695, 672, 1956) 3. UniProt Ids (for example: Q06187, P00533)
3. UniProt accessions (for example: BTK_HUMAN, BRCA1_HUMAN)
4. RefSeq Ids (for example: NP_958441.1, NP_958440.1)

Enrichment analysis

Enrichment analysis can be performed by selecting different background database options implemented in FunRich.

- Cellular process (CC)
- Biological process (BP)
- Molecular function (MF)
- Protein domains
- Site of expression
- Biological pathways
- Transcription factors
- Clinical phenotypes

Background database

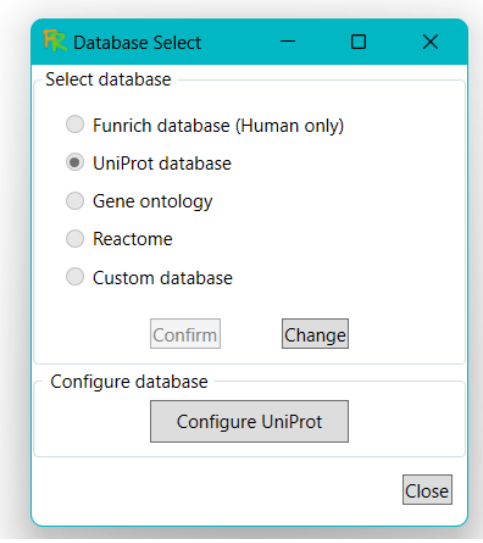
1. FunRich database
2. UniProt database
3. Gene Ontology
4. Reactome
5. Custom database

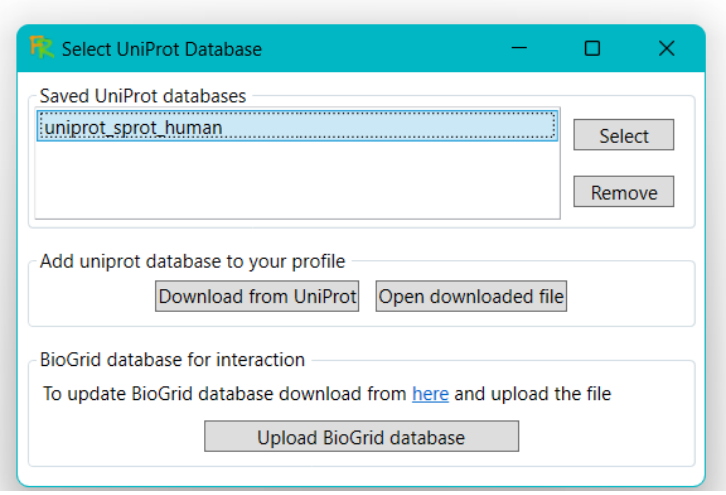
Methodology:

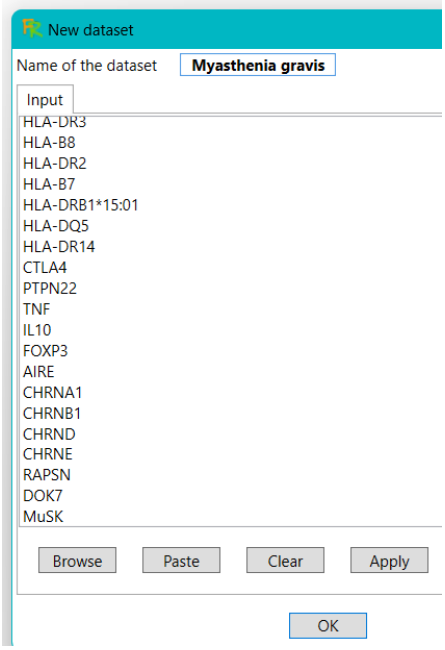
A. Adding data set and selecting database:

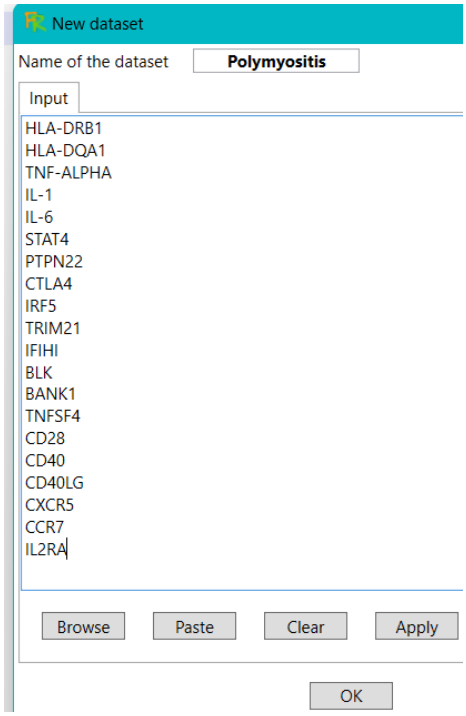
1. Click on the 'add data set' under the category 'upload data sets'.
2. Click on 'Browse' button to upload the query gene/protein lists. Users can also simply 'copy' the genes/protein list and click on the 'apply' the button. Users can submit all types of genes/protein list as explained in the 'type of query/input list' under the section 'Overview of FunRich tool features'.
3. Enter the name of the data set. A short name is advisable.
4. In order to select the background FunRich database, click on the 'manage' button under the category 'Manage database'. Select 'change' button to select the respective database.
5. To select UniProt database, click on the 'manage' button under the category 'Manage database'. Further, click on the 'change' button and select 'UniProt database' options and confirm the same.
6. Changing the database to UniProt from the other database will result in deleting the entire user submitted query list. The same message will be displayed as popup message. Please select 'Yes' to proceed further.
7. Next click on the 'Configure UniProt' button under the category 'Configure database', a new window will appear asking the user either to 'Add new UniProt' database or 'Select' one from the already existing 'Saved UniProt database' list.

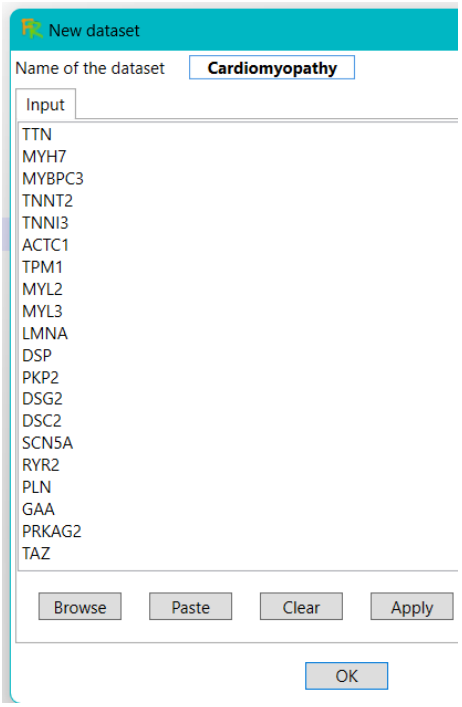
8. If the user decides to add new UniProt database, then click on the 'Add new UniProt' button. A new window appears asking the user, either to select one from the list of organism to download the corresponding organism specific UniProt database. Currently, the list contains ~11 different taxonomical level against which enrichment analysis can be performed. Within the same window, users can either upload the UniProt database stored in their local system or can directly click on the URL link provided to download the most updated database from UniProt directly and can further upload the same.

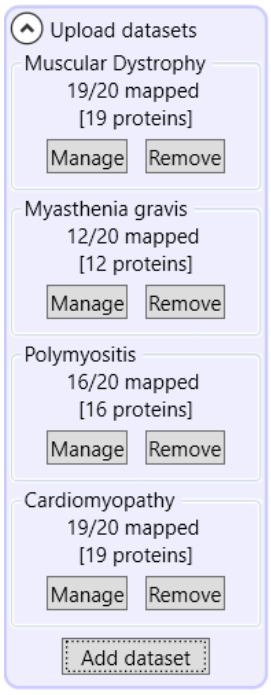


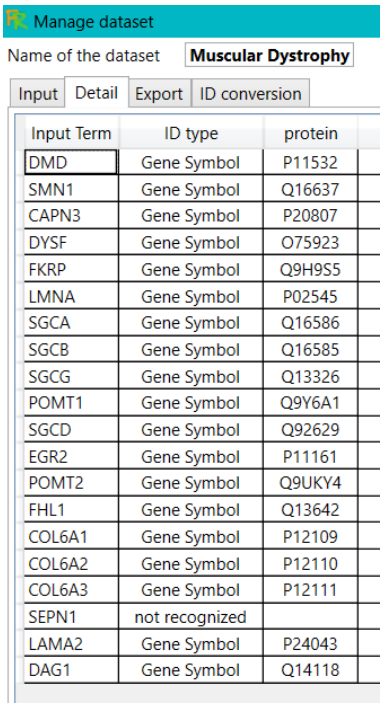




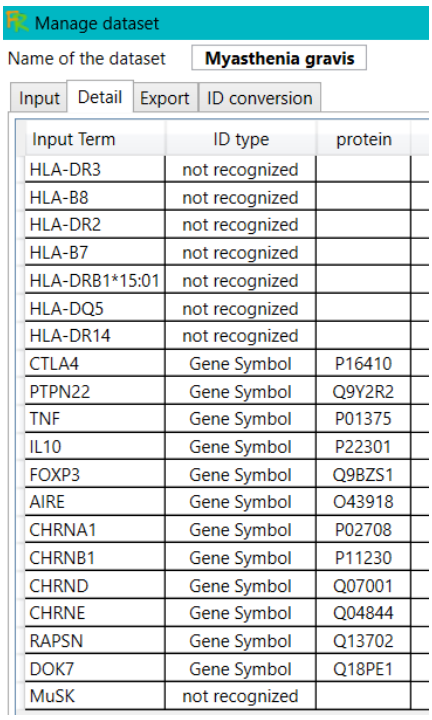




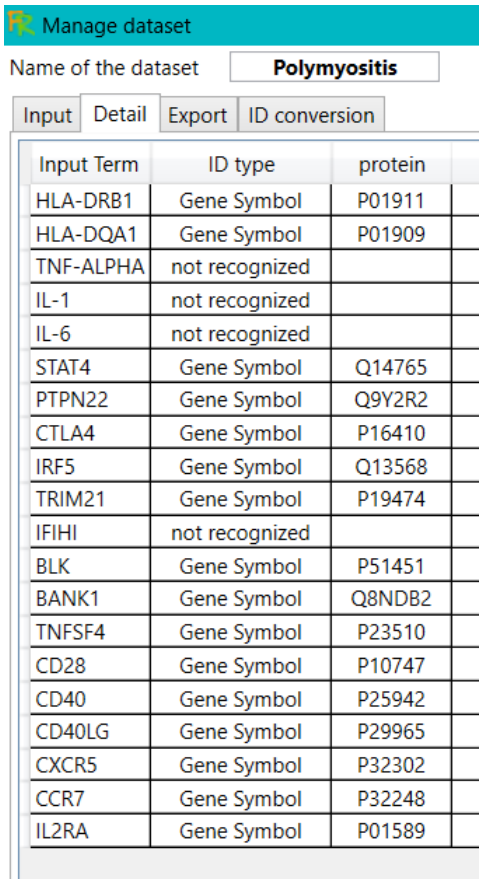




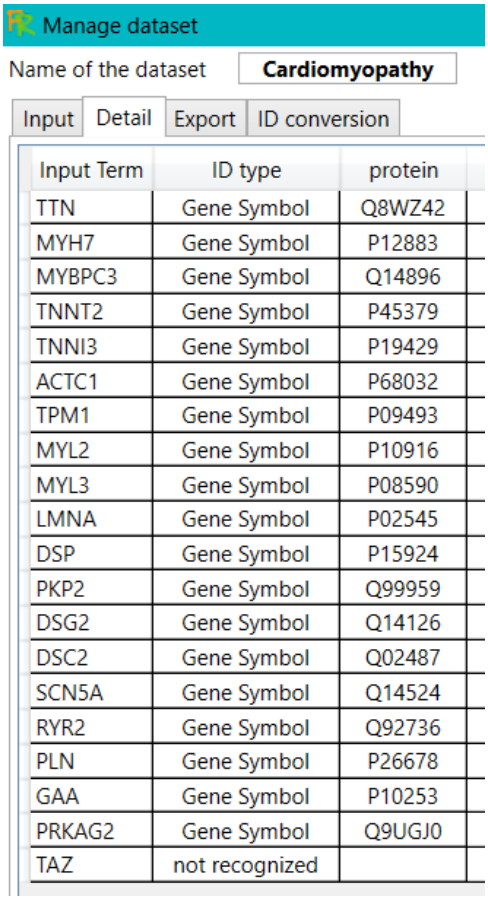
Input Term	ID type	protein
DMD	Gene Symbol	P11532
SMN1	Gene Symbol	Q16637
CAPN3	Gene Symbol	P20807
DYSF	Gene Symbol	O75923
FKRP	Gene Symbol	Q9H9S5
LMNA	Gene Symbol	P02545
SGCA	Gene Symbol	Q16586
SGCB	Gene Symbol	Q16585
SGCG	Gene Symbol	Q13326
POMT1	Gene Symbol	Q9Y6A1
SGCD	Gene Symbol	Q92629
EGR2	Gene Symbol	P11161
POMT2	Gene Symbol	Q9UKY4
FHL1	Gene Symbol	Q13642
COL6A1	Gene Symbol	P12109
COL6A2	Gene Symbol	P12110
COL6A3	Gene Symbol	P12111
SEPN1	not recognized	
LAMA2	Gene Symbol	P24043
DAG1	Gene Symbol	Q14118



Input Term	ID type	protein
HLA-DR3	not recognized	
HLA-B8	not recognized	
HLA-DR2	not recognized	
HLA-B7	not recognized	
HLA-DRB1*15:01	not recognized	
HLA-DQ5	not recognized	
HLA-DR14	not recognized	
CTLA4	Gene Symbol	P16410
PTPN22	Gene Symbol	Q9Y2R2
TNF	Gene Symbol	P01375
IL10	Gene Symbol	P22301
FOXP3	Gene Symbol	Q9BZS1
AIRE	Gene Symbol	O43918
CHRNA1	Gene Symbol	P02708
CHRNB1	Gene Symbol	P11230
CHRND	Gene Symbol	Q07001
CHRNE	Gene Symbol	Q04844
RAPSN	Gene Symbol	Q13702
DOK7	Gene Symbol	Q18PE1
MuSK	not recognized	



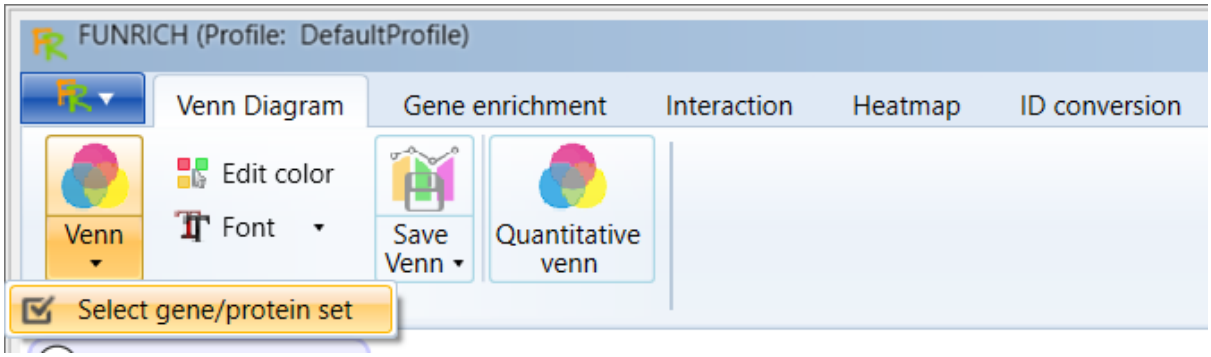
Input Term	ID type	protein
HLA-DRB1	Gene Symbol	P01911
HLA-DQA1	Gene Symbol	P01909
TNF-ALPHA	not recognized	
IL-1	not recognized	
IL-6	not recognized	
STAT4	Gene Symbol	Q14765
PTPN22	Gene Symbol	Q9Y2R2
CTLA4	Gene Symbol	P16410
IRF5	Gene Symbol	Q13568
TRIM21	Gene Symbol	P19474
IFIHI	not recognized	
BLK	Gene Symbol	P51451
BANK1	Gene Symbol	Q8NDB2
TNFSF4	Gene Symbol	P23510
CD28	Gene Symbol	P10747
CD40	Gene Symbol	P25942
CD40LG	Gene Symbol	P29965
CXCR5	Gene Symbol	P32302
CCR7	Gene Symbol	P32248
IL2RA	Gene Symbol	P01589

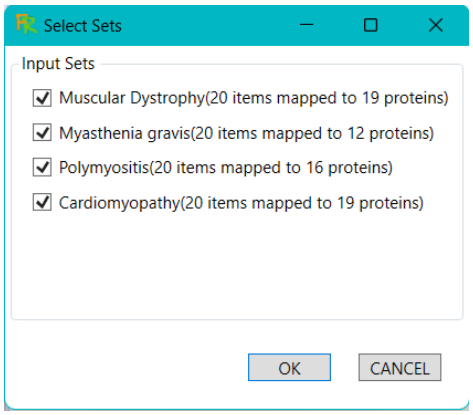


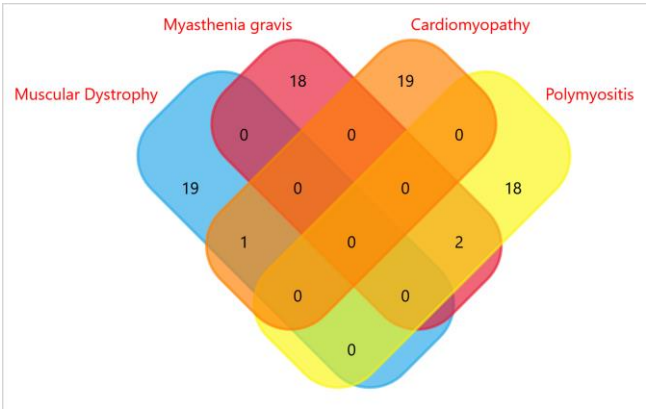
Input Term	ID type	protein
TTN	Gene Symbol	Q8WZ42
MYH7	Gene Symbol	P12883
MYBPC3	Gene Symbol	Q14896
TNNT2	Gene Symbol	P45379
TNNI3	Gene Symbol	P19429
ACTC1	Gene Symbol	P68032
TPM1	Gene Symbol	P09493
MYL2	Gene Symbol	P10916
MYL3	Gene Symbol	P08590
LMNA	Gene Symbol	P02545
DSP	Gene Symbol	P15924
PKP2	Gene Symbol	Q99959
DSG2	Gene Symbol	Q14126
DSC2	Gene Symbol	Q02487
SCN5A	Gene Symbol	Q14524
RYR2	Gene Symbol	Q92736
PLN	Gene Symbol	P26678
GAA	Gene Symbol	P10253
PRKAG2	Gene Symbol	Q9UGJ0
TAZ	not recognized	

B. Generate Venn diagram

1. Upload the input/query gene lists into the FunRich tool using the 'Upload datasets as explained above.
2. Click on the icon named 'Venn' to generate Venn diagram between these three datasets. By default, scalable Venn diagram for these datasets is generated.
3. Users can also right click on those numbers to further select those genes for further analysis.
4. Load input/query gene list and automatically, Venn diagram between these four datasets will be generated .
5. Click on the 'Venn' icon options and then select gene/protein list. Here, uncheck the dataset to revert back to the above Scalable-Venn diagram.
6. Click on this icon just located above the 'Venn' icon. Select 'Option' and uncheck the option named 'Area to be proportional to the number of genes' and then click on the 'Apply' button. This will generate unscalable-Venn diagram .
7. To generate matrix table also known as pairwise comparison, again repeat the above step but check the option named 'Show pair-comparison only' and then click on the 'Apply' button.





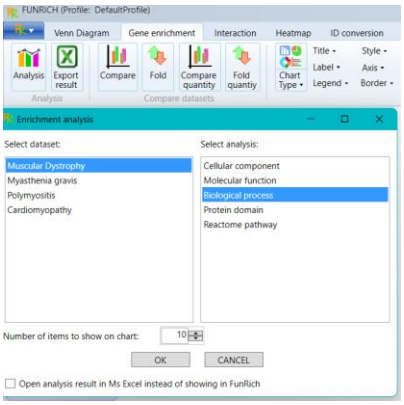


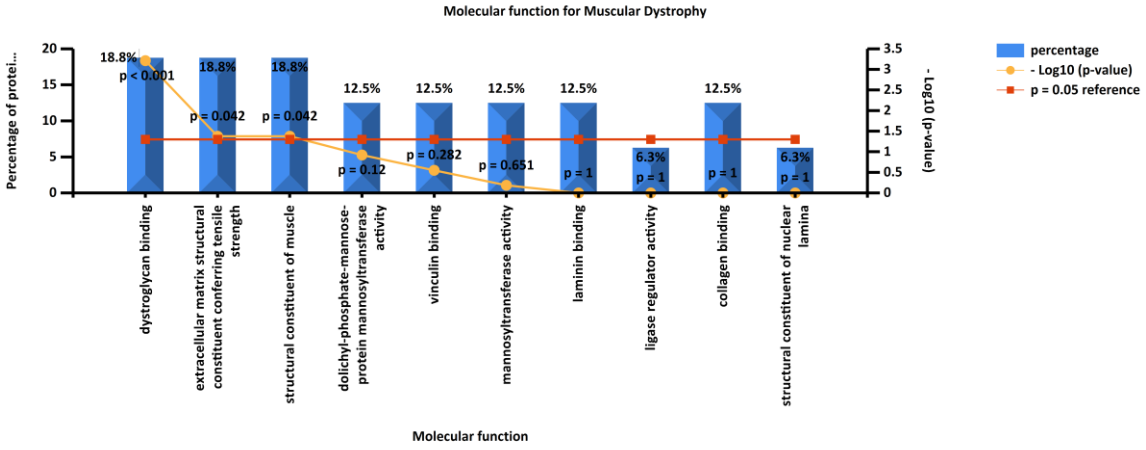
C. Enrichment Analysis:

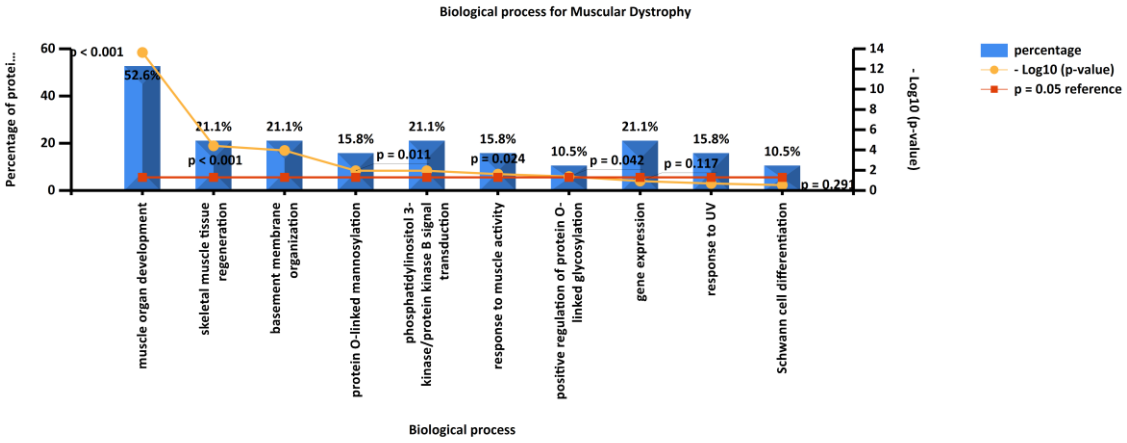
1. Upload the downloaded sample datasets into the FunRich tool as explained in the above steps.
2. Click on the icon named 'Chart' and select required dataset as input dataset and click 'OK' to display the enrichment results against Uniprot database.
3. By default, 'Column graph' depicting the enrichment analysis of gene list for first 10 'Cellular Component (CC)' terms is displayed based on the p-value.
4. The size of the graph displayed can be anytime increased or decreased by clicking on this icon.
5. Click on the 'Axis' button and select the 'X-axis label rotation' option to specific angle for clear visualization of the terms on the 'X-axis' .
6. On the right side of the graph, a table containing all those eight CC terms is displayed. Users can uncheck any of those terms for not to be shown on the graph.
7. Users can also play around the other graph options like title, style, label and legends to make it more customized.
8. To further analyze same dataset for the enrichment of Molecular Function (MF), Biological Process (BP), Biological Pathway (BP), Protein Domain, Site of Expression, Transcription factor and Clinical Phenotype click on the tabs displayed on the bottom of the graph respectively.
9. Click on the icon 'Compare' and select dataset from the set A and select other dataset from the set B and click 'OK' button.
10. Click on the icon 'Chart Type' and select 'Doughnut chart' and select on the tab named 'Clinical phenotype' on the bottom of the graph. Click on the 'Style' icon and select 'SoftEdge' as one of the Style option.

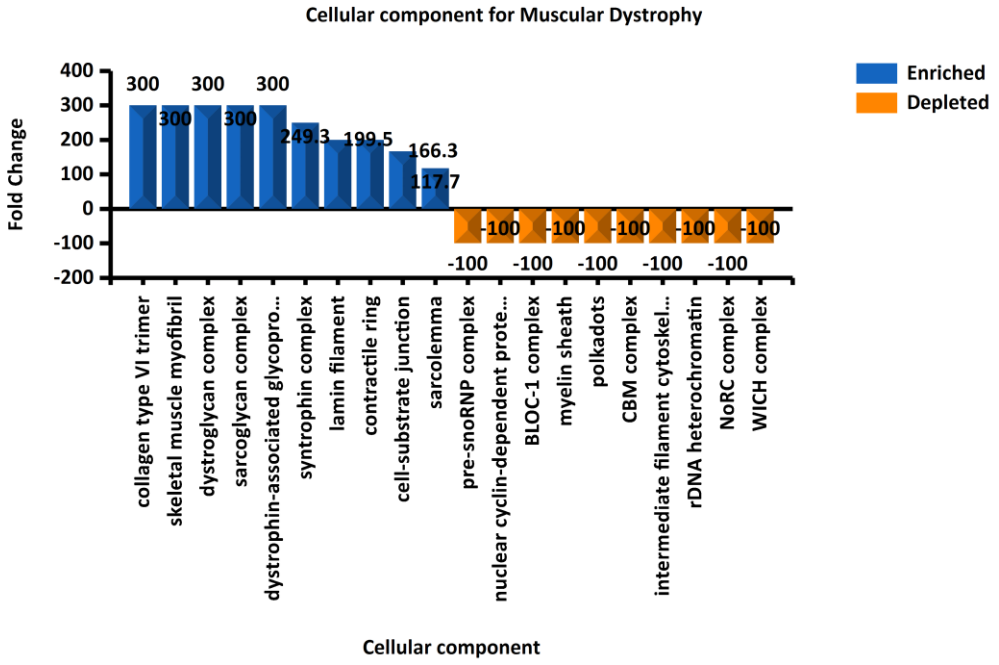
11. Click on the ‘Legend’ icon and reduce the height of the legends by selecting the option ‘Reduce Height’ under the ‘Adjust Legend Size’ option.

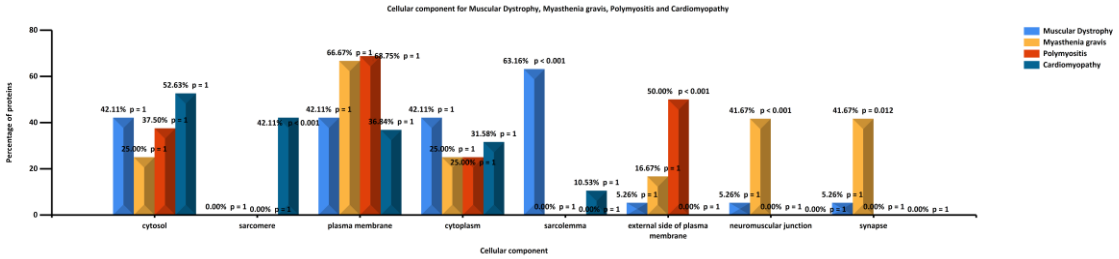
12. To save the chart click on the icon ‘Save Chart’. The FunRich tool generated ‘Doughnut chart’ with ‘SoftEdge’ style depicting the ‘Clinical Phenotype’ enrichment results of the two datasets one as on the outer circle and other on the inner circle will be saved in your local system.







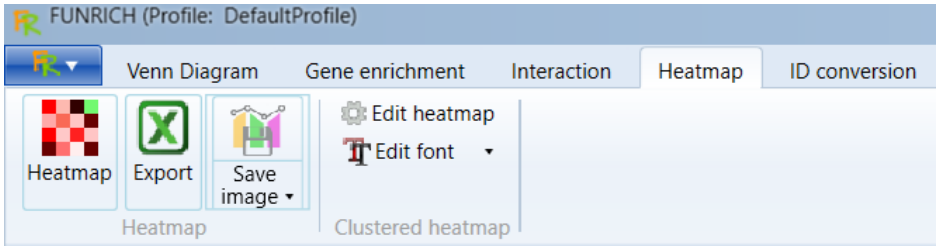


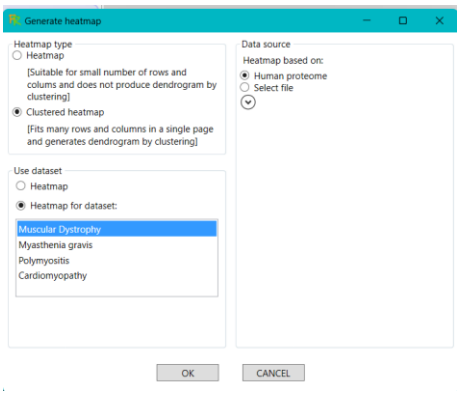


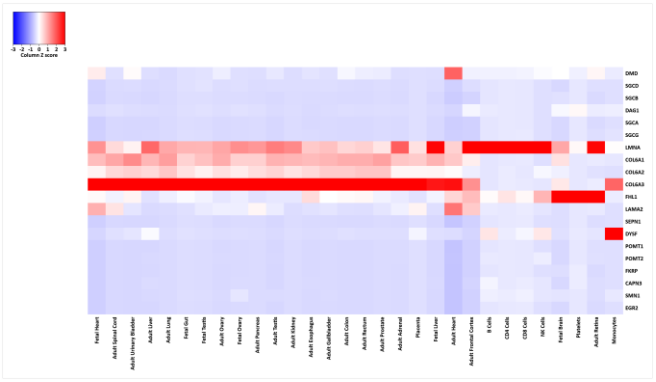
D. Generating Heatmap

Currently, Heatmap for a list of human query/input gene list is generated using Human proteome map quantitative dataset as background. Heatmap works independently of 'FunRich', 'UniProt' and 'Custom' database.

1. Load sample input gene list from the sample dataset file as explained in the above steps.
2. Click on the icon named 'Heatmap' under the tab named 'Heatmap'. Automatically, Heatmap for the input sample gene list will be generated using quantitative Human proteome map expression database as background.
3. In order to visualize the normalized expression values, check the option button named 'Normalize each gene'. For each gene, its relative expression across all those tissues is calculated and higher expression, by default, will have intense red color relative to the lower expression values.
4. Click on the icon named 'Save Chart' and select 'Save' to save the Heatmap image in your local system folder. Users can further change the color scale of the Heatmap accordingly.





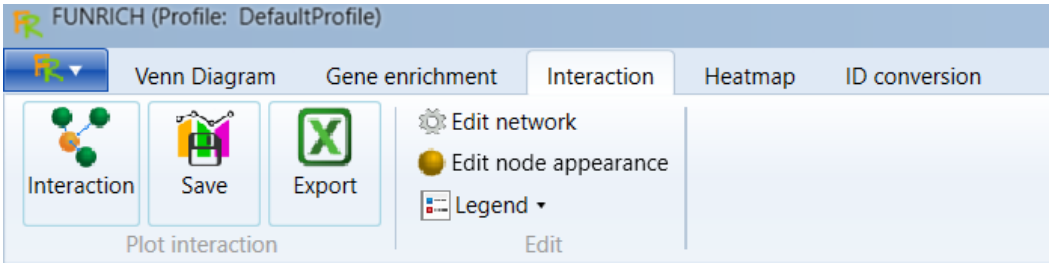


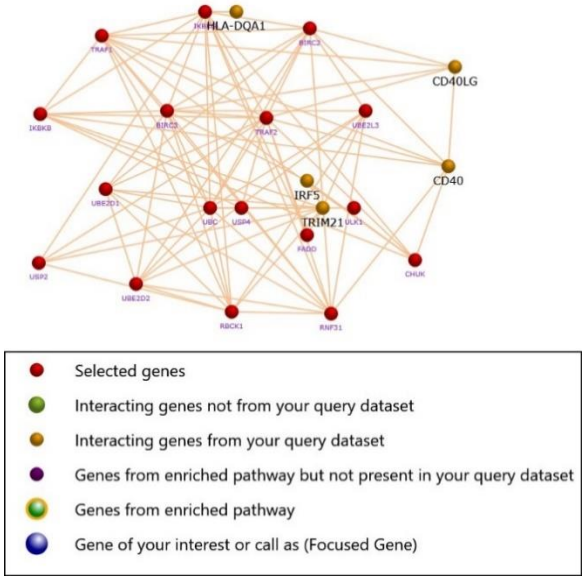
E. Generate interaction networks

Interaction networks of query genes list can be depicted by using both 'FunRich' and 'Custom' database as background.

1. Upload the sample gene list from the sample datasets.
2. Click on the icon named 'Interaction' which is under the tab named 'Interaction Model' to generate protein-protein interaction (PPI) networks of sample input list. By default, 'Planetary' layout is displayed and users can right click on the interaction network layout space and select the 'packed' layout .
3. Adjacent to PPI networks, the sample genes enriched for particular biological pathways are also shown along with their p-values.
4. Click on the pathway term and all those nodes known to be significantly enriched in that particular pathway will be highlighted.
5. For further analysis on this subset of nodes, right click on that particular pathway and select 'Plot to new network'. A new window pops up displaying three check box options using which further analysis can be carried out.
   a. First check box option named 'Add direct neighbors from dataset' represent adding interacting partners which directly interacts with those nodes.
   b. Second check box option named 'Add direct neighbors outside dataset' represent adding interacting partners which are known to directly interact with those nodes but not the part of interactors interacting with sample input list.
   c. Third check box option named 'Include other genes involved in the pathway' represent adding all those genes involved in that particular pathway as interacting partners which are known to interact with those nodes.
6. Select the first check box and click 'OK' button to visualize all the direct interactors of nodes separately and change the layout to 'Planetary'.

7. Click on the icon named ‘Save' to save the interaction network graph in your local system folder. Users can further change the color, size, thickness and font of the nodes etc. to make it more customized.





Conclusion:

Using FunRich software, functional annotation studies on four diseases: Muscular Dystrophy, Myasthenia Gravis, Polymyositis, and Cardiomyopathy. Each disease was mapped to specific proteins: Muscular Dystrophy (20 items mapped to 19 proteins), Myasthenia Gravis (20 items mapped to 12 proteins), Polymyositis (20 items mapped to 16 proteins), and Cardiomyopathy (20 items mapped to 19 proteins) was conducted. This analysis allowed us to explore the shared and distinct functional pathways, biological processes, and molecular functions associated with each disease. The findings from this functional annotation study highlight the potential involvement of specific proteins in these disease pathways, offering insights into their roles in disease progression and possible therapeutic targets. These insights provide a foundation for further experimental research to validate and explore these protein functions in detail.

Practice-8

TCGA Data Download Using R

Aim: To download and analyse multi-omics data from The Cancer Genome Atlas (TCGA) database using R packages, focusing on gene expression, DNA methylation, and mutation data for different cancer types, specifically breast cancer (BRCA) and glioblastoma (GBM).

Tools and Packages Required:

- **R Programming Language**: Used for data retrieval, analysis, and visualization.
- **TCGAbiolinks**: Allows querying, downloading, and preparing TCGA data.
- **tidyverse**: A collection of packages for data manipulation and visualization.
- **maftools**: Specialized for mutation data analysis and visualization.
- **pheatmap**: Creates heatmaps, useful for visualizing DNA methylation differences.
- **SummarizedExperiment**: Manages large experimental data in a structured format.

Introduction: The Cancer Genome Atlas (TCGA) is a comprehensive database that provides large-scale genomic data, allowing researchers to study various types of cancer. It includes multi-omics data like gene expression profiles, DNA methylation, and mutation data, which are essential for understanding cancer mechanisms and identifying biomarkers. This report demonstrates how to retrieve TCGA data using R packages, such as TCGAbiolinks and maftools, and how to preprocess and visualize the data. The primary focus is on breast cancer (TCGA-BRCA) and glioblastoma (TCGA-GBM), with an emphasis on RNA sequencing, DNA methylation, and mutation analysis.

Methodology:

Note: Before running the analysis, ensure that the required R packages are installed. To install any missing packages, use the following code in your R environment.

BiocManager::install("Package")

After installing the package, Make sure to load the package using

library(Package)

Step 1: Retrieve Available TCGA Projects and Summary

- A list of TCGA projects is retrieved using getGDCprojects().
- A project summary for TCGA-BRCA (breast cancer) is obtained to understand available datasets.

Step 2: Query and Download Gene Expression Data (RNA-Seq)

- A query is built to download RNA-Seq data from TCGA-BRCA, focusing on specific samples.
- Data is downloaded with GDCdownload, and GDCprepare prepares it for analysis.

- A matrix of gene expression values is extracted using assay(), specifically the fpkm_unstrand measurement.

Step 3: Query and Download DNA Methylation Data

- A query is created for DNA methylation data for glioblastoma (TCGA-GBM) using the Illumina Human Methylation 27 platform.
- The downloaded data is prepared and visualized by selecting probes with significant variability.
- A heatmap is generated using pheatmap to illustrate beta value differences between samples.

Step 4: Query and Download Mutation Data

- A query is set up for mutation data from TCGA-BRCA, selecting specific samples.
- Mutation data is visualized using maftools, displaying a summary and oncoprint for the top 10 mutated genes.

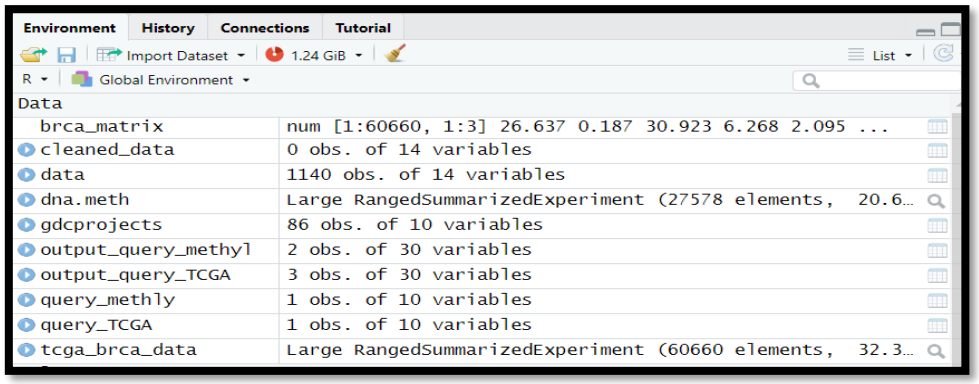


Fig: Uploading data into the R environment.

R Script:

```
# script to download data from TCGA using TCGAbiolinks

# setwd("~/Desktop/demo/TCGAbiolinks")

library(TCGAbiolinks)

library(tidyverse)

library(maftools)

library(pheatmap)

library(SummarizedExperiment)

# get a list of projects

gdcprojects <- getGDCprojects()
```

```
getProjectSummary('TCGA-BRCA')
# building a query
query_TCGA <- GDCquery(project = 'TCGA-BRCA',
        data.category = 'Transcriptome Profiling')
output_query_TCGA <- getResults(query_TCGA)
# build a query to retrieve gene expression data ----------
query_TCGA <- GDCquery(project = 'TCGA-BRCA',
                data.category = 'Transcriptome Profiling',
                experimental.strategy = 'RNA-Seq',
                workflow.type = 'STAR - Counts',
                access = 'open',
                barcode = c('TCGA-LL-A73Y-01A-11R-A33J-07', 'TCGA-E2-A1IU-01A-
11R-A14D-07','TCGA-AO-A03U-01B-21R-A10J-07'))
output_query_TCGA <- getResults(query_TCGA)
getResults(query_TCGA)
# download data - GDCdownload
GDCdownload(query_TCGA)
# prepare data
tcga_brca_data <- GDCprepare(query_TCGA, summarizedExperiment = TRUE)
brca_matrix <- assay(tcga_brca_data, 'fpkm_unstrand')
# build a query to retrieve DNA methylation data ----------
query_methly <- GDCquery(project = 'TCGA-GBM',
        data.category = 'DNA Methylation',
        platform = 'Illumina Human Methylation 27',
        access = 'open',
        data.type = 'Methylation Beta Value',
        barcode = c('TCGA-19-0962-01B-01D-0521-05', 'TCGA-06-0137-01A-01D-0218-05'))
output_query_methyl <- getResults(query_methly)
GDCdownload(query_methly)
```

```
# plot probes showing differences in beta values between samples
dna.meth <- GDCprepare(query_methly, summarizedExperiment = TRUE)
assay(dna.meth)
# plot probes showing differences in beta values between samples
dna.meth <- GDCprepare(query_methly, summarizedExperiment = TRUE)
assay(dna.meth)
idx <- dna.meth %>%
  assay %>%
  rowVars() %>%
  order(decreasing = TRUE) %>%
  head(10)
# plot
pheatmap(assay(dna.meth)[idx,])
# download and visualize mutation data from TCGA ----------------------
query_mutation <- GDCquery(project = 'TCGA-BRCA',
        data.category = 'Simple Nucleotide Variation',
        access = 'open',
        barcode = c('TCGA-LL-A73Y-01A-11D-A33E-09,TCGA-LL-A73Y-10B-01D-A33H-
09',
                    'TCGA-E9-A1NH-01A-11D-A14G-09,TCGA-E9-A1NH-11A-33D-A14G-09'))
output_query_mutation <- getResults(query_mutation)
GDCdownload(query_mutation)
maf <- GDCprepare(query_mutation, summarizedExperiment = TRUE)
maf <- GDCprepare(query_mutation, summarizedExperiment = TRUE)
# maftools utils to read and create dashboard
maftools.input <- read.maf(maf)
plotmafSummary(maf = maftools.input,
               addStat = 'median',
               rmOutlier = TRUE,
```

```
        dashboard = TRUE)
# oncoprint
oncoplot(maf = maftools.input,
        top = 10,
        removeNonMutated = TRUE)
```

RESULTS:

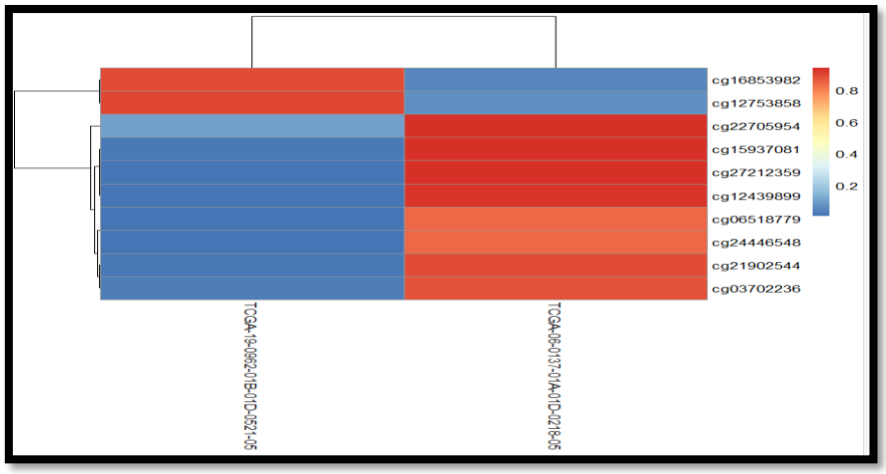


Fig: pheatmap

The heatmap illustrates differential intensity patterns for features (e.g., methylation sites) across two samples. One sample (TCGA-06-0137-01A-01D-0218-05) shows higher intensity values (in red) for most features, while the other (TCGA-19-0962-01B-01D-0521-05) shows lower values (in blue). Clustering on the left groups features with similar patterns, highlighting distinct profiles between the samples, which may indicate biologically relevant differences.

Mutation Profile:

- This displays various aspects of genetic mutations across samples.
- The Variant Classification panel indicates that missense mutations are predominant, followed by a few nonsense and splice site mutations.
- The Variant Type panel shows that SNPs are the most common mutation type, while the SNV Class panel reveals that C>T transitions are the most frequent single nucleotide variant.
- The Variants per Sample panel suggests a median of 59 mutations per sample, with missense mutations as the main type.
- Finally, the Top 10 Mutated Genes panel lists frequently mutated genes, each mutated in 50% of samples, suggesting potential relevance to the study.

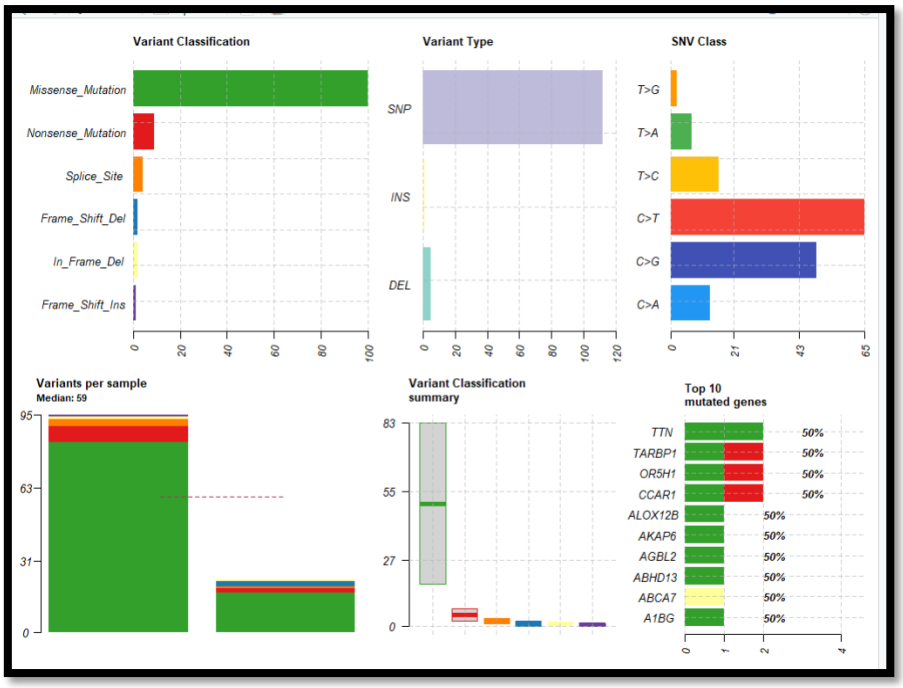


Fig: Mutation Profile maf plot

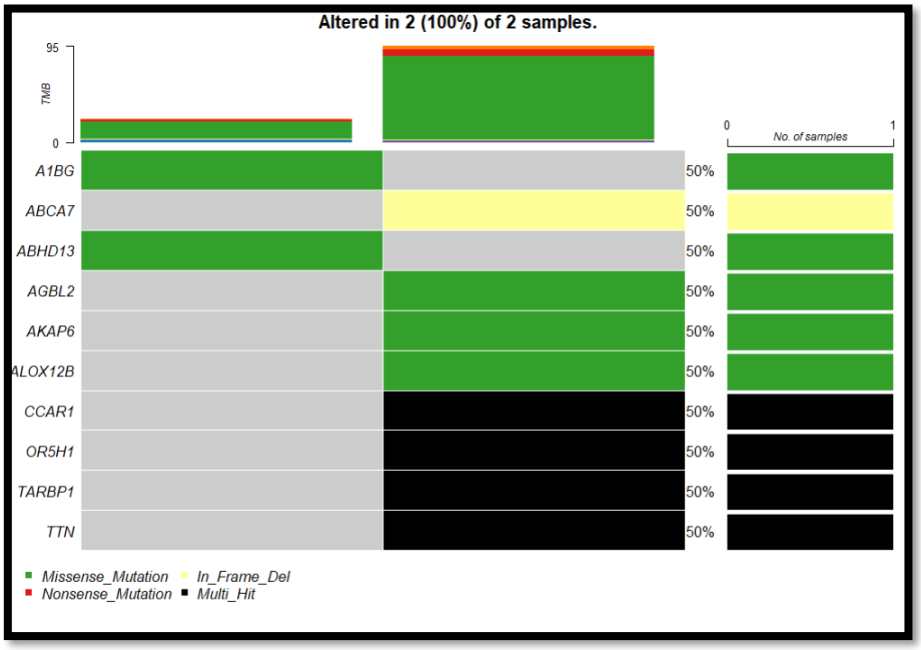


Fig: Oncoprint plot

- The Oncoprint plot shows various mutation types (e.g., Missense, In_Frame_Del, Nonsense, Multi_Hit) across specific genes, all altered in both samples. Colors represent mutation types, with the left bar indicating tumour mutation burden (TMB) by frequency. This visualization highlights key genes with high mutation rates across samples.

- The plot reveals that all genes listed are altered in 100% of the two samples analysed, indicating a high mutation prevalence across these genes.
- Missense mutations (green) are the most frequent type, while a few genes, like *ABCA7* and *A1BG*, also show in-frame deletions (yellow) and multi-hits (red).
- Genes such as *CCAR1*, *OR5H1*, *TARBP1*, and *TTN* have black bars, likely representing a lack of specific mutation type classification.

Conclusion

The analysis of genetic mutation and methylation data reveals significant insights into the underlying biological processes associated with the observed phenotype. The heatmap visualization indicates distinct methylation profiles across samples, with one sample exhibiting higher methylation levels, suggesting potential biological relevance. The Mutation Profile Summary highlights that missense mutations are the most prevalent, with SNPs being the dominant variant type and C>T transitions occurring frequently. The identification of the Top 10 mutated genes, with mutations found in 50% of the samples, underscores their potential role in the phenotype. The Oncoprint further confirms the widespread alteration of key genes, particularly through missense mutations, in both samples, indicating their likely involvement in the phenotype. The additional mutation types, such as in-frame deletions and multi-hits, in some genes emphasize their complexity and potential significance in the observed genetic landscape. These findings point to a set of critical genetic changes that could inform future research and therapeutic strategies.

Practice-9

MarkerDB

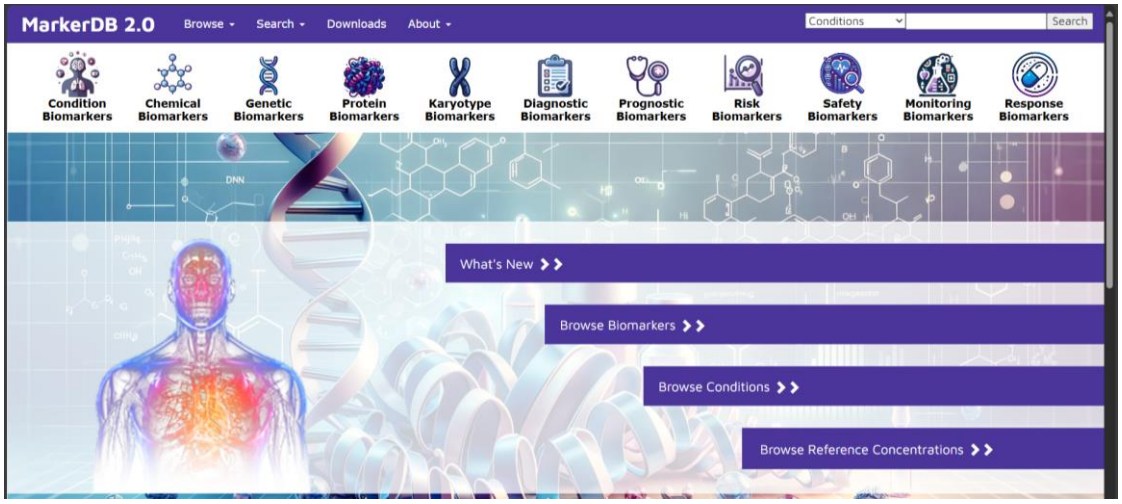
Aim: To explore the MarkerDB database for Biomarker Identification

Description:

MarkerDB is a freely available electronic database that attempts to consolidate information on all known clinical and a selected set of pre-clinical molecular biomarkers into a single resource. The database includes four major types of molecular biomarkers (chemical, protein, DNA [genetic] and karyotypic) and four biomarker categories (diagnostic, predictive, prognostic and exposure). MarkerDB provides information such as: biomarker names and synonyms, associated conditions or pathologies, detailed disease descriptions, detailed biomarker descriptions, biomarker specificity, sensitivity and ROC curves, standard reference values (for protein and chemical markers), variants (for SNP or genetic markers), sequence information (for genetic and protein markers), molecular structures (for protein and chemical markers), tissue or biofluid sources (for protein and chemical markers), chromosomal location and structure (for genetic and karyotype markers), clinical approval status and relevant literature references. Users can browse the data by conditions, condition categories, biomarker types, biomarker categories or search by sequence similarity through the advanced search function.

Methodology:

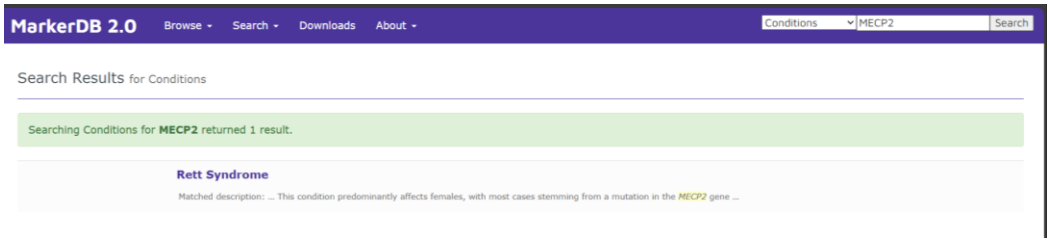
Step 1: Access MarkerDB: Open a web browser and go to https://markerdb.ca.

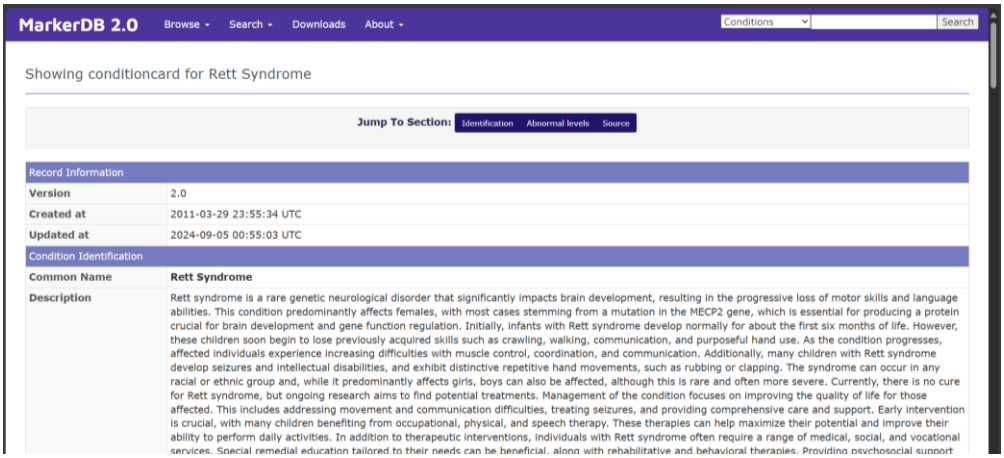


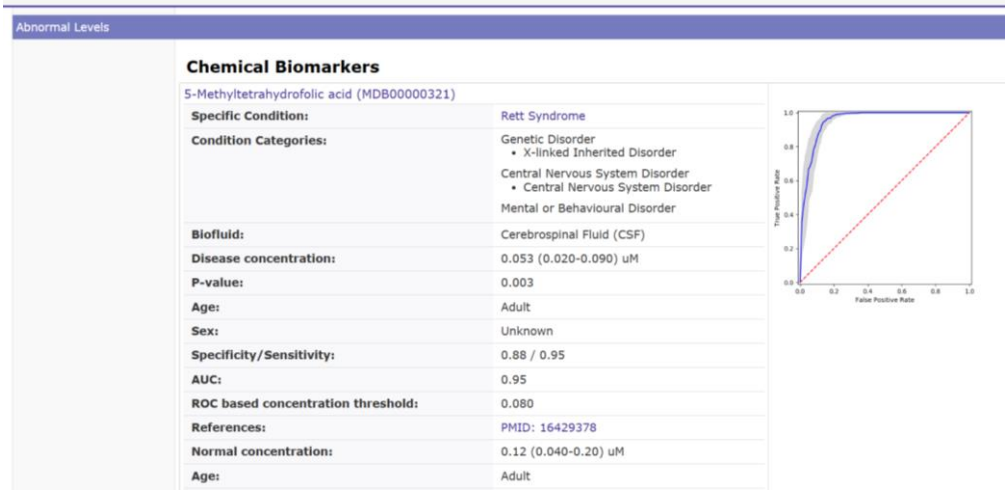
Step 2: Search Biomarkers Using the Basic Search Bar

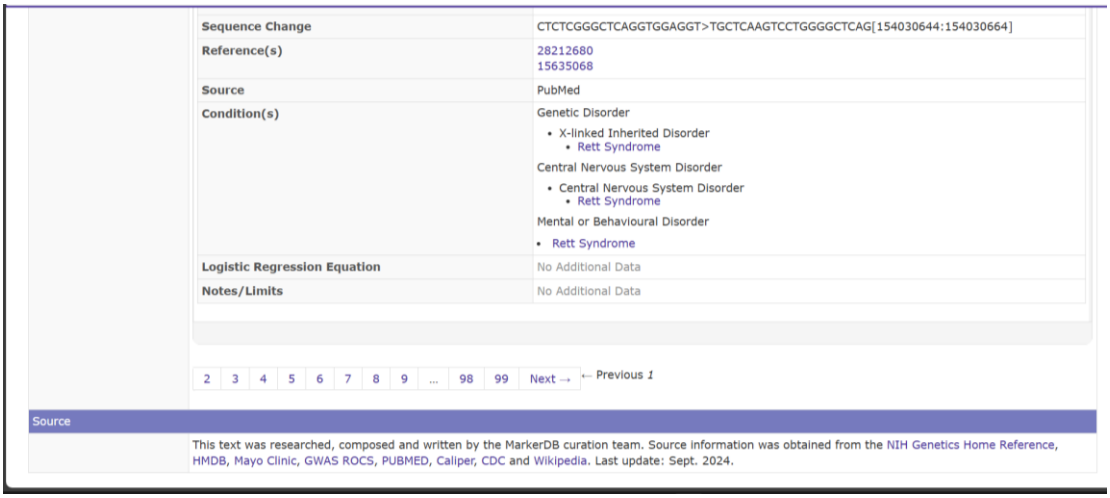
- **Locate the Search Bar**: Find the basic search bar on the homepage.
- **Search for a Specific Gene**: Enter the **gene name** of interest into the search bar. **Select the Category** relevant to your study. Categories may include genetic, chemical, condition biomarkers, etc.
- **Review the Search Results**: Open and explore the generated results. For each result, examine the following details as per the biomarker type:

- **Conditions**: Record Information, Condition Identification, Abnormal Levels, Source
- **Genetic Markers**: Record Information, Molecular Information, Gene Identification, Sequence Variants, GWAS-ROCS Study (Genome-Wide Association Studies or ROC Curve studies), Sources







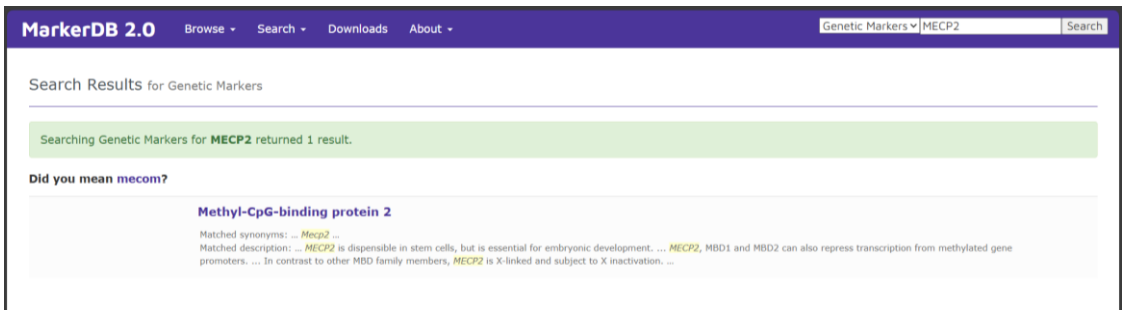


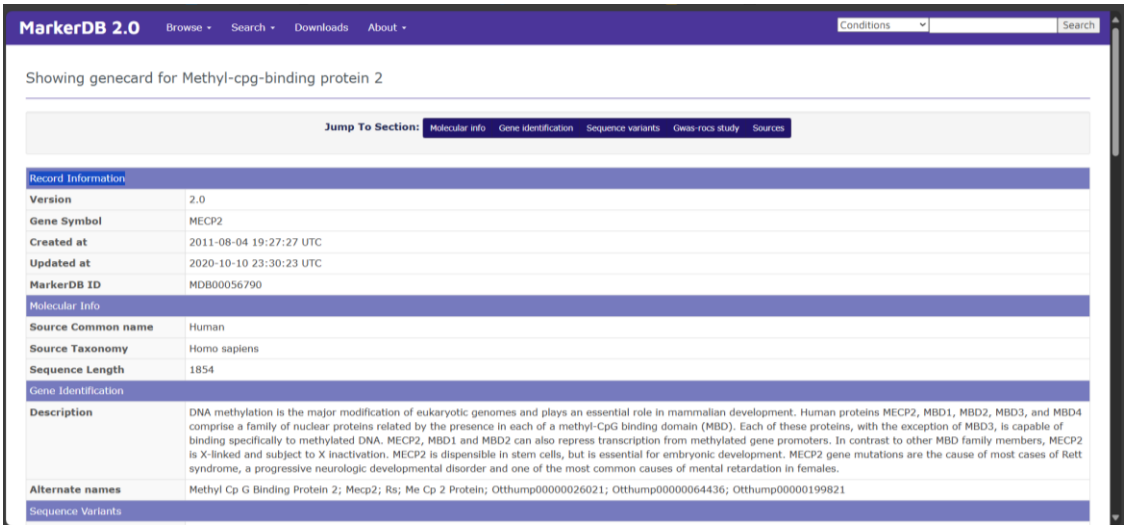
Sequence Change	CTCTCGGGCTCAGGTGGAGGT>TGCTCAAGTCCTGGGGCTCAG[154030644:154030664]
Reference(s)	28212680 15635068
Source	PubMed
Condition(s)	Genetic Disorder • X-linked Inherited Disorder • Rett Syndrome Central Nervous System Disorder • Central Nervous System Disorder • Rett Syndrome Mental or Behavioural Disorder • Rett Syndrome
Logistic Regression Equation	No Additional Data
Notes/Limits	No Additional Data

2 3 4 5 6 7 8 9 ... 98 99 Next → ← Previous 1

Source

This text was researched, composed and written by the MarkerDB curation team. Source information was obtained from the NIH Genetics Home Reference, HMDB, Mayo Clinic, GWAS ROCS, PUBMED, Caliper, CDC and Wikipedia. Last update: Sept. 2024.



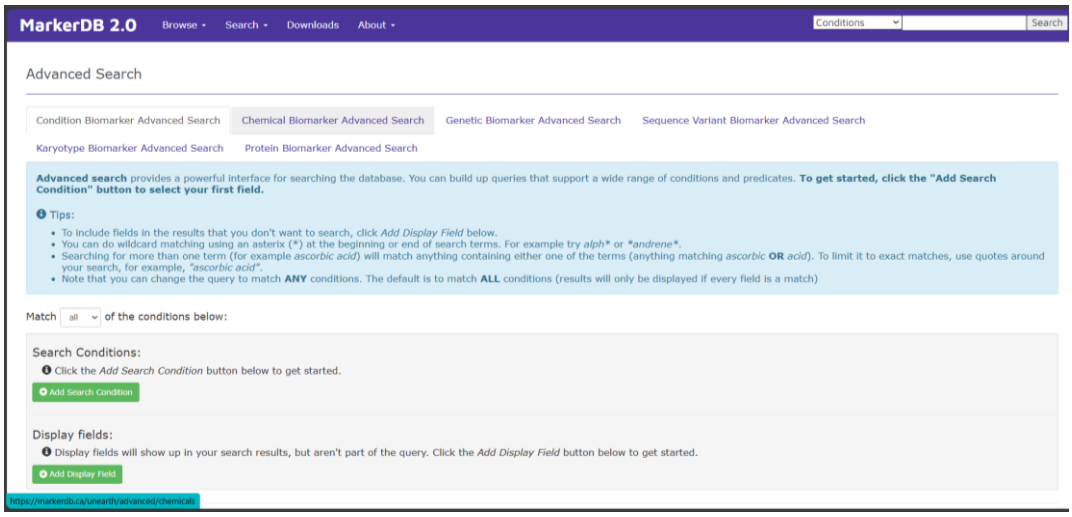


Step 3: Advanced Search: MarkerDB offers an **Advanced Search** option with targeted criteria for biomarker types. Follow these steps:

- **Select Advanced Search**: Navigate to the **Advanced Search** section.
- **Choose the Biomarker Type**: Select from the following types to refine your search:
  - Condition Biomarker Advanced Search
  - Chemical Biomarker Advanced Search
  - Genetic Biomarker Advanced Search

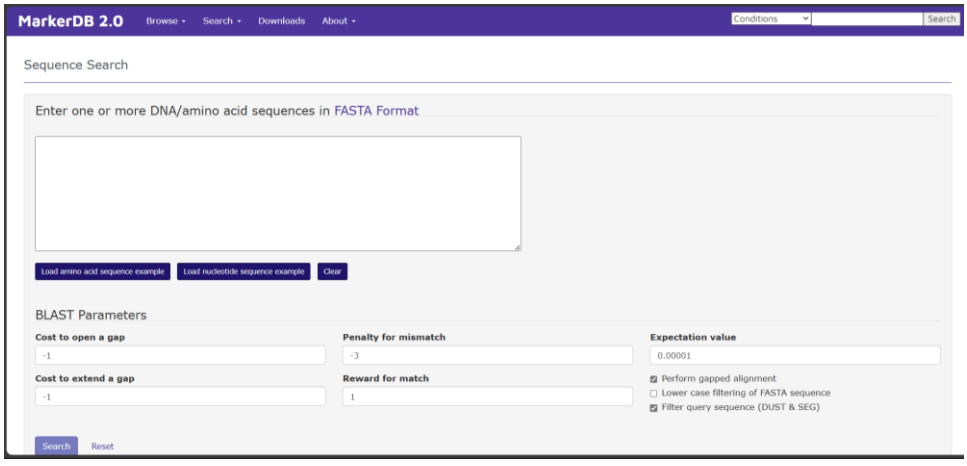
- Sequence Variant Biomarker Advanced Search
- Karyotype Biomarker Advanced Search
- Protein Biomarker Advanced Search

- **Refine Criteria**: Enter relevant parameters specific to the biomarker type selected to narrow down your search.



Step 4: Sequence Search (For DNA/Protein Analysis)

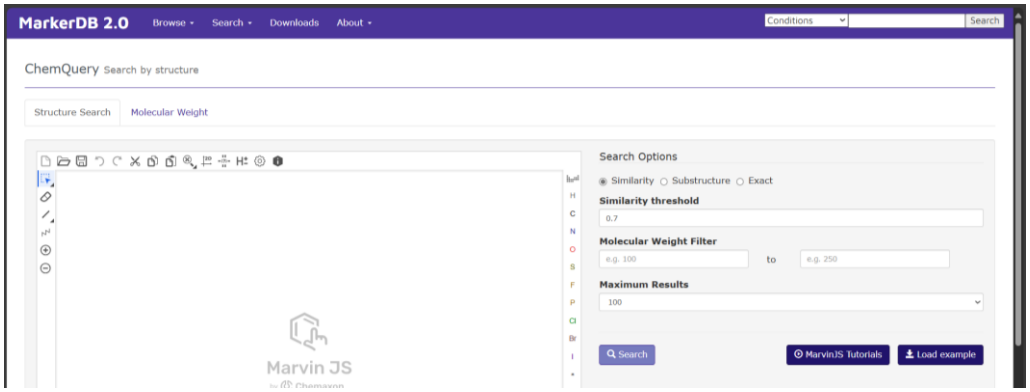
- **Access Sequence Search**: Open the **Sequence Search** tool from the MarkerDB homepage or search section.
- **Enter Sequence Data**: Input DNA or amino acid sequences in **FASTA format**.
- **Specify BLAST Parameters**: Customize BLAST (Basic Local Alignment Search Tool) parameters as required for sequence alignment.

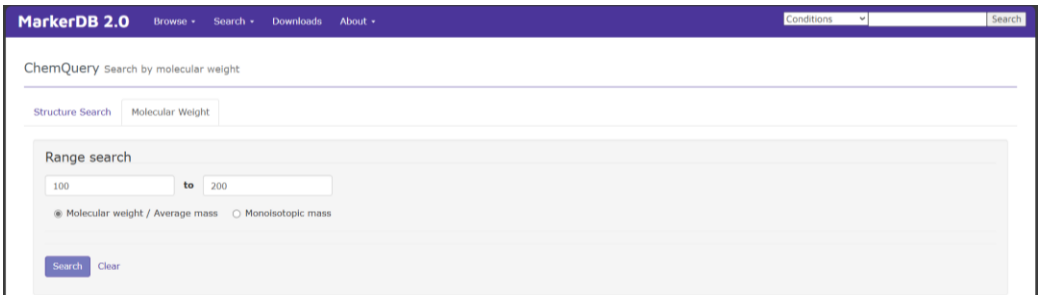


Step 5: ChemQuery Search (For Structural or Chemical Biomarkers)

- **Locate ChemQuery Search**: Use the **ChemQuery Search** option for searches based on molecular structure.

- **Choose Search Parameters**: Perform searches by **structure**, **molecular weight**, or **3D structure**.





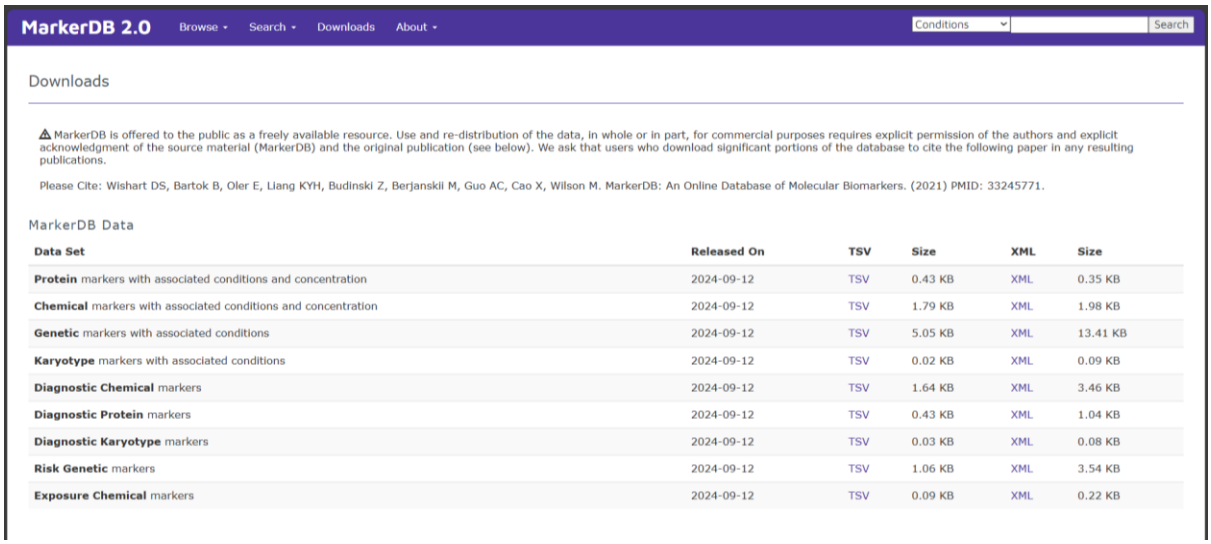
Step 6: Browse Options: MarkerDB provides an extensive browsing option to navigate biomarkers across various categories:

- **Open the Browse Section**: Click on the **Browse** option in the main menu.
- **Select Biomarker Type**: Choose from the following categories:
  - All Biomarkers
  - Condition-Specific Biomarkers
  - Chemical Biomarkers
  - Genetic Biomarkers
  - Protein Biomarkers
  - Karyotype Biomarkers
  - Diagnostic, Prognostic, Risk, Safety, Monitoring, or Response Biomarkers
  - Reference Concentrations
- **Review Data**: For each selected biomarker, examine details provided by MarkerDB, as outlined in **Step 2** for conditions and genetic markers.

Step 7: Download Data:

- **Access Download Options**: Locate the **Downloads** section on the website.

- MarkerDB data is freely available for non-commercial use.
- For commercial purposes, obtain explicit permission from the authors and ensure appropriate citation.
- **Cite MarkerDB**: If using downloaded data in publications, cite MarkerDB and the original publication, as specified by the database.



Conclusion:

In conclusion, MarkerDB serves as a comprehensive and user-friendly resource for identifying and analysing various types of biomarkers, providing valuable insights across genetic, chemical, and condition-specific markers. By using its diverse search and browsing functionalities, users can efficiently retrieve biomarker data, assess genetic variations, and explore condition-specific abnormalities. The availability of both simple and advanced search options, alongside tools like sequence alignment and ChemQuery, facilitates tailored searches that meet specific research needs. Furthermore, the ability to download data for further analysis expands MarkerDB's utility in supporting large-scale studies and data-driven research.