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.

Conclusion:

- 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.

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.

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.

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.

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.