Systems Biology Final Project

Professor: Dr. Najmeh Salehi
Presenter: Neda Esfehani
Spring 2023

Introduction

GCNF encodes a nuclear receptor, a member of the nuclear hormone receptor family. Its expression pattern suggests that it may play a role in neurogenesis and germ cell growth. This gene controls Oct4 expression levels during ES cell differentiation and early mouse embryonic development. GCNF was overexpressed in H9 human ES cells using the Tet-On Inducible system (a conditional gene expression system in which transcription is turned on or off in the presence of tetracycline or doxycycline).

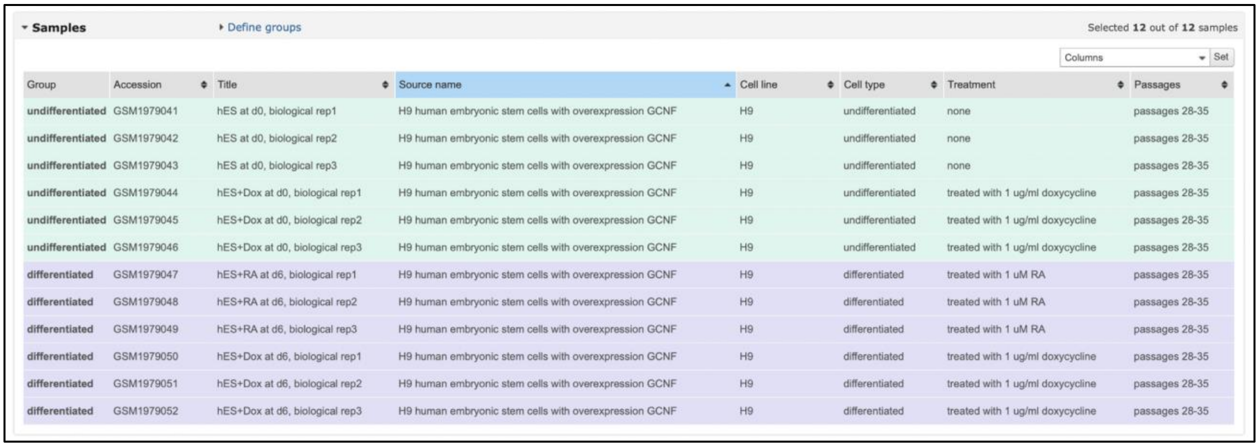
mRNA microarray analysis showed that overexpression of GCNF globally regulates gene expression in undifferentiated and differentiated hES cells. The aim of this study is to investigate the regulation of GCNF expression in differentiated and undifferentiated ES cells.

Analyze a microarray dataset from GEO database using R2GEO and answer the following questions and submit a complete report:

Dataset Title	Expression data from undifferentiated and differentiated human ES cells
GEO Accession	GSE76282
Platform	Affymetrix Human Genome U133 Plus 2.0 Array

1. What categorizations have you considered for this dataset? (At least two data groups should be considered, for example patient and healthy)

The selected dataset contains gene expression data from 12 samples of human ES cells with GCNF overexpression. The dataset is categorized into undifferentiated and differentiated groups (Figure1)



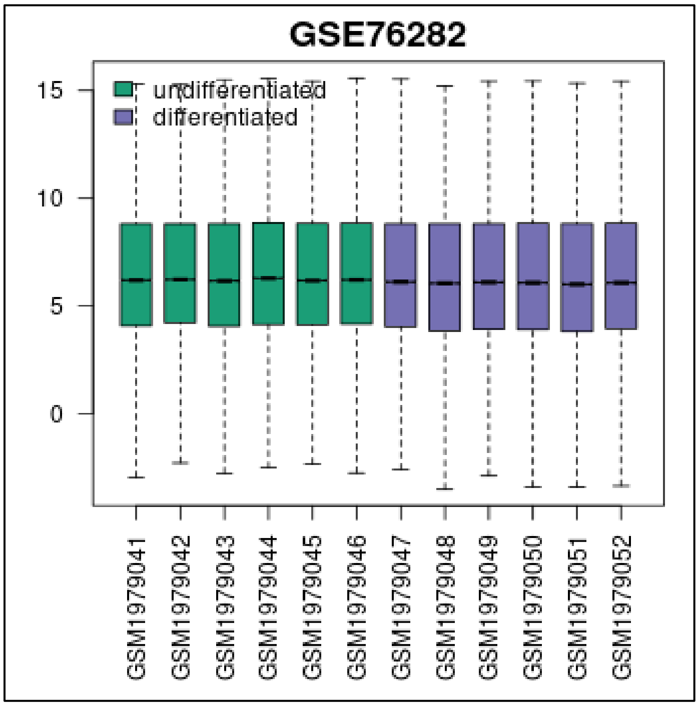
Group	Accession	Title	Source name	Cell line	Cell type	Treatment	Passages
undifferentiated	GSM1979041	hES at d0, biological rep1	H9 human embryonic stem cells with overexpression GCNF	H9	undifferentiated	none	passages 28-35
undifferentiated	GSM1979042	hES at d0, biological rep2	H9 human embryonic stem cells with overexpression GCNF	H9	undifferentiated	none	passages 28-35
undifferentiated	GSM1979043	hES at d0, biological rep3	H9 human embryonic stem cells with overexpression GCNF	H9	undifferentiated	none	passages 28-35
undifferentiated	GSM1979044	hES+Dox at d0, biological rep1	H9 human embryonic stem cells with overexpression GCNF	H9	undifferentiated	treated with 1 ug/ml doxycycline	passages 28-35
undifferentiated	GSM1979045	hES+Dox at d0, biological rep2	H9 human embryonic stem cells with overexpression GCNF	H9	undifferentiated	treated with 1 ug/ml doxycycline	passages 28-35
undifferentiated	GSM1979046	hES+Dox at d0, biological rep3	H9 human embryonic stem cells with overexpression GCNF	H9	undifferentiated	treated with 1 ug/ml doxycycline	passages 28-35
differentiated	GSM1979047	hES+RA at d6, biological rep1	H9 human embryonic stem cells with overexpression GCNF	H9	differentiated	treated with 1 uM RA	passages 28-35
differentiated	GSM1979048	hES+RA at d6, biological rep2	H9 human embryonic stem cells with overexpression GCNF	H9	differentiated	treated with 1 uM RA	passages 28-35
differentiated	GSM1979049	hES+RA at d6, biological rep3	H9 human embryonic stem cells with overexpression GCNF	H9	differentiated	treated with 1 uM RA	passages 28-35
differentiated	GSM1979050	hES+Dox at d6, biological rep1	H9 human embryonic stem cells with overexpression GCNF	H9	differentiated	treated with 1 ug/ml doxycycline	passages 28-35
differentiated	GSM1979051	hES+Dox at d6, biological rep2	H9 human embryonic stem cells with overexpression GCNF	H9	differentiated	treated with 1 ug/ml doxycycline	passages 28-35
differentiated	GSM1979052	hES+Dox at d6, biological rep3	H9 human embryonic stem cells with overexpression GCNF	H9	differentiated	treated with 1 ug/ml doxycycline	passages 28-35

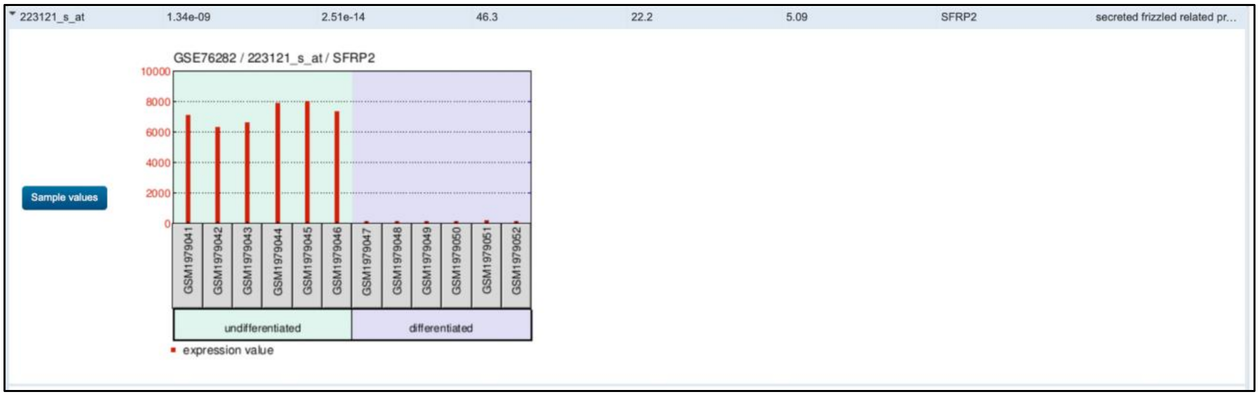
Figure 1: Categorized sample groups in the dataset.

2. Identify the Differentially Expressed Genes (DEG) between the groups. What cutoffs have you used to filter the data? After filtering, send the related file.

Boxplots are used to visualize the distribution of gene expression values in each sample. Since there are no outlier points above and below the maximum and minimum lines it seems we don't have outlier data. Outlier points may indicate extreme or unusual expression values that deviate markedly from the overall distribution. Also, the median expression seems to be roughly equal across all groups. The median in the center shows that the groups are comparable. The box length indicates the spread or diversity of gene expression in the sample. A wider box indicates greater diversity. This factor also does not differ much between groups.

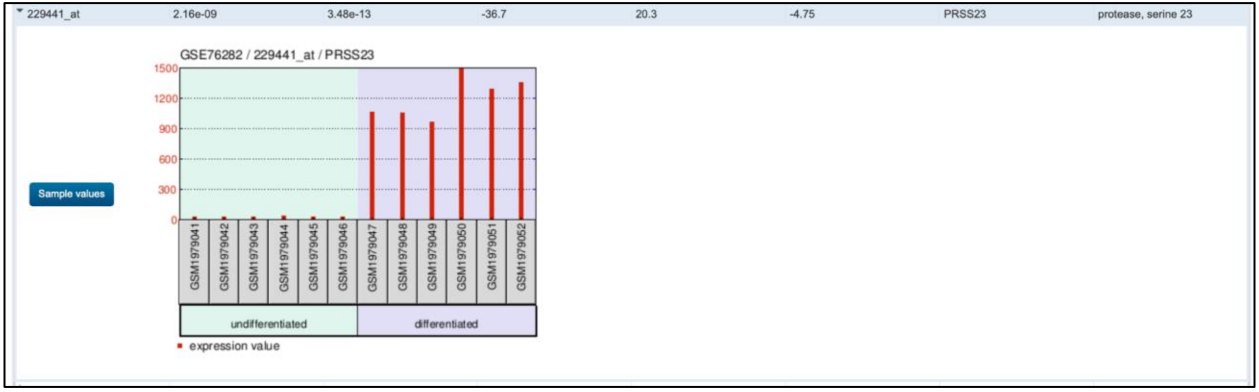
Since this plot is made for all the genes of interest in the samples, we do not expect to see stark differences in appearance. If we want to find differences in expression of different genes in differentiated vs undifferentiated samples, we can click on each individual gene to obtain this information (Figures 2 and 3).





223121_s_at 1.34e-09 2.51e-14 46.3 22.2 5.09 SFRP2 secreted frizzled related pr...

Figure 2: SFRP2 is an example of a gene that had higher expression in the undifferentiated ES cell group



229441_at 2.16e-09 3.48e-13 -36.7 20.3 -4.75 PRSS23 protease, serine 23

Figure 3: PRSS23 is an example of a gene that had higher expression in the differentiated cell group

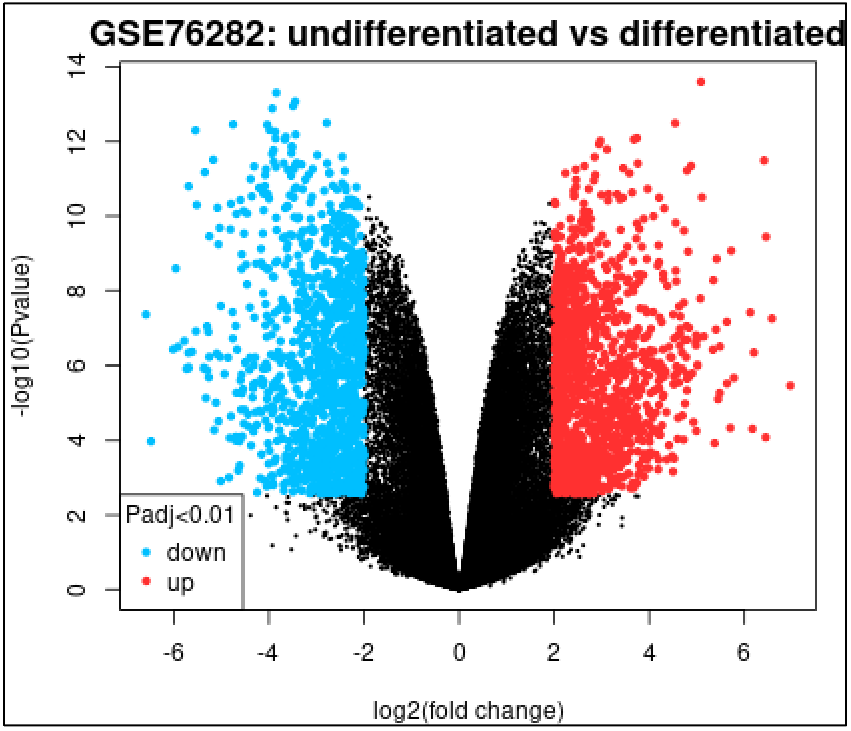


Figure 4 shows the volcano plot for the genes. (Refer to this link to examine each gene)

The p-value indicates the probability that the observed differences in gene expression between the two groups are due to chance. A smaller p-value indicates a higher level of statistical significance. For example, if $p - value = 0.001$ then:

$$-\log_{10} 0.001 = 3$$

But if $p - value = 0.1$ then:

$$-\log_{10} 0.1 = 1$$

So, the higher the significance, the gene will be higher up the y-axis. In the volcano plot, a higher LogFC indicates a greater difference in gene expression between the two groups being compared.

LogFC is calculated as the log2 ratio of a gene's expression level in one condition (e.g. undifferentiated stem cells) to its expression level in another condition (e.g. differentiated stem cells). A positive LogFC indicates the gene was upregulated in the first group relative to the second group, while a negative LogFC indicates downregulation.

For identifying differentially expressed genes, a higher LogFC represents a greater change in gene expression. Genes with higher LogFC values are often selected for further analysis and investigation. However, it is important to consider statistical significance (p-value) along with LogFC to ensure that observed differences are not due to random fluctuations.

To examine the genes and limit their number as well as increase confidence in the analyses, the cutoffs used were LogFC>2 (genes with LogFC greater than 2 and less than -2 are important) and p-adj<0.01 (genes with significance below 0.01 are important). As can also be seen in the plot, values of x between 2 and -2 as well as

P-adj < 0.01
log2(0.01) < -2
-log2(0.01) > 2
Y > 2

have been considered as genes that did not have expression differences in the two groups.

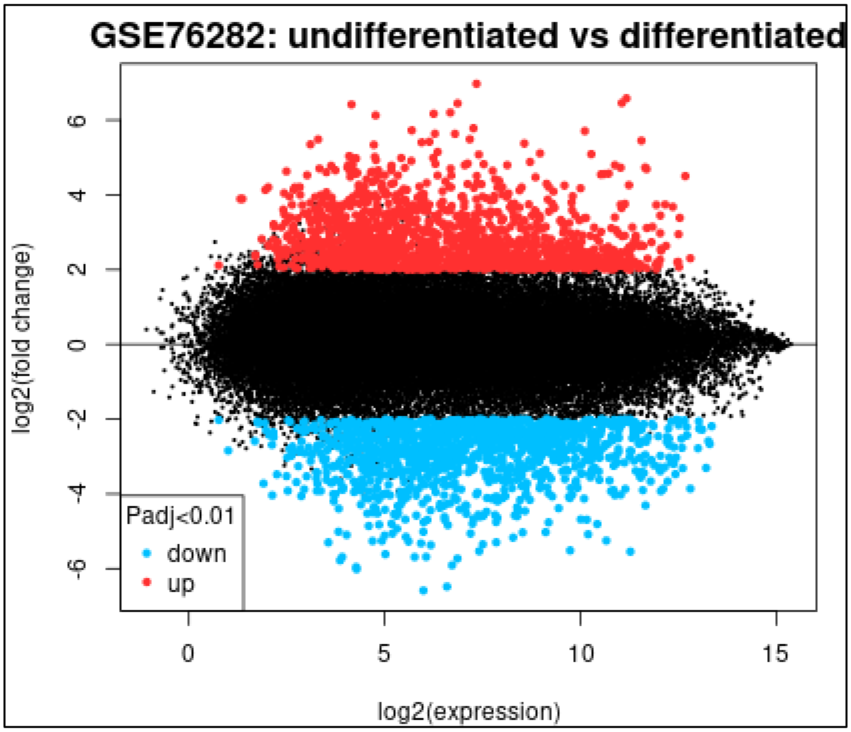


Figure 5 shows the gene difference plot-mean. (Refer to this link to examine each gene)

Genes clustered around the center line (LogFC = 0) have smaller differences in expression between groups, while genes farther from the center line have larger changes. The width of the distribution can indicate the overall magnitude of gene expression changes observed in the dataset; here the magnitude is between -6 and 6.

The x-axis shows the mean expression level of a gene, which can help identify genes with low or high expression levels under the conditions studied. Genes with very high or very low expression levels are found at the far right and left ends of the x-axis.

Looking at the black points, there is a tendency for the points to shift towards the right side of the axis; this indicates that genes with low expression have greater diversity in LogFC while genes with higher expression typically shift towards (LogFC = 0).

The MA plot can also be useful for evaluating data quality and normalization method effectiveness. Even distribution of data points around the center line indicates good normalization and minimal bias. The data appears to be evenly distributed so no need for force normalization.

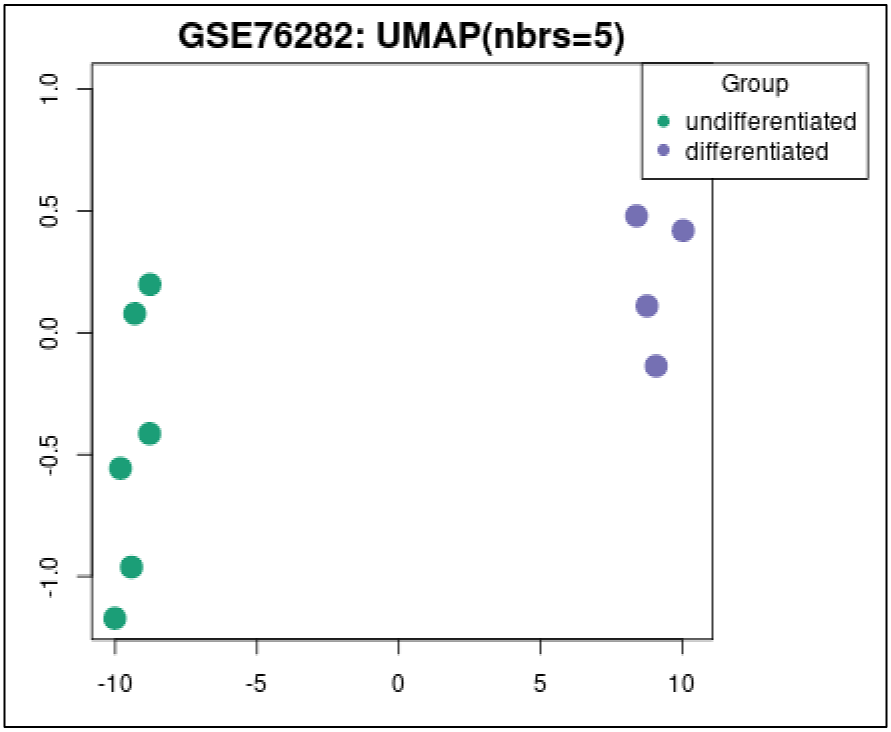


Figure 6 shows the gene UMAP plot.

UMAP (Uniform Manifold Approximation and Projection) is a dimension reduction technique commonly used to visualize and analyze high-dimensional gene expression data from DEGs. UMAP can provide insights into relationships, clusters, and patterns within the DEGs.

UMAP can reveal clusters or groups of DEGs with similar expression patterns. By representing high-dimensional expression data in a lower dimensional space, UMAP can identify regions where DEGs with similar expression profiles are more prevalent. Clusters in the above plots may indicate functional modules or pathways that are affected under the conditions studied.

It can be seen that the DEGs are nicely clustered into distinct groups associated with the grouping of interest. Each point in the plot represents genes with similar expressions that are grouped based on a nearest neighbors' parameter when considering local structure. The neighbors balance local vs global structure in the UMAP.

In the above plot there are no outlier points. Outlier points in UMAP plots can indicate genes with unique or aberrant expression patterns compared to other DEGs. These outlier points may highlight genes with specialized functions. Note that UMAP is a nonlinear dimension reduction technique that aims to preserve local and global data structures. It creates a low dimensional representation in which data points close in the high dimensional space are modeled to be positioned close together in the UMAP plot. Therefore, the x and y values in UMAP plots do not have direct biological interpretation.

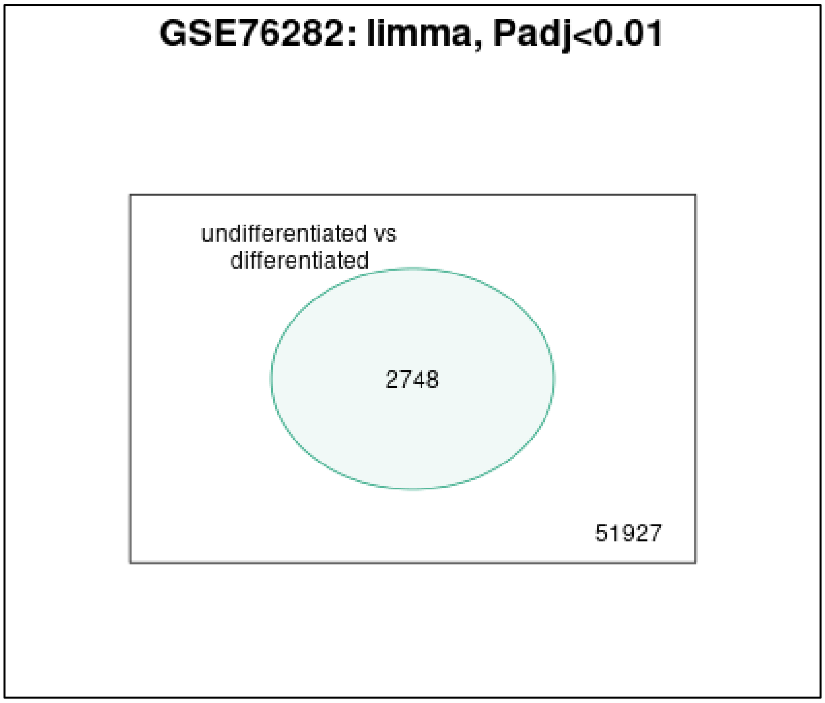

GSE76282: limma, Padj<0.01
undifferentiated vs differentiated
2748
51927

Figure 7 shows the gene Venn diagram. (Refer to this link to examine each gene)

The Venn diagram is a graphical representation commonly used to show the overlap or intersection between different sets of differentially expressed genes (DEGs) across different groupings. It can be used to identify DEGs that are specifically upregulated or downregulated in one group vs another. Or it can be used to compare ratios of DEGs across different groups.

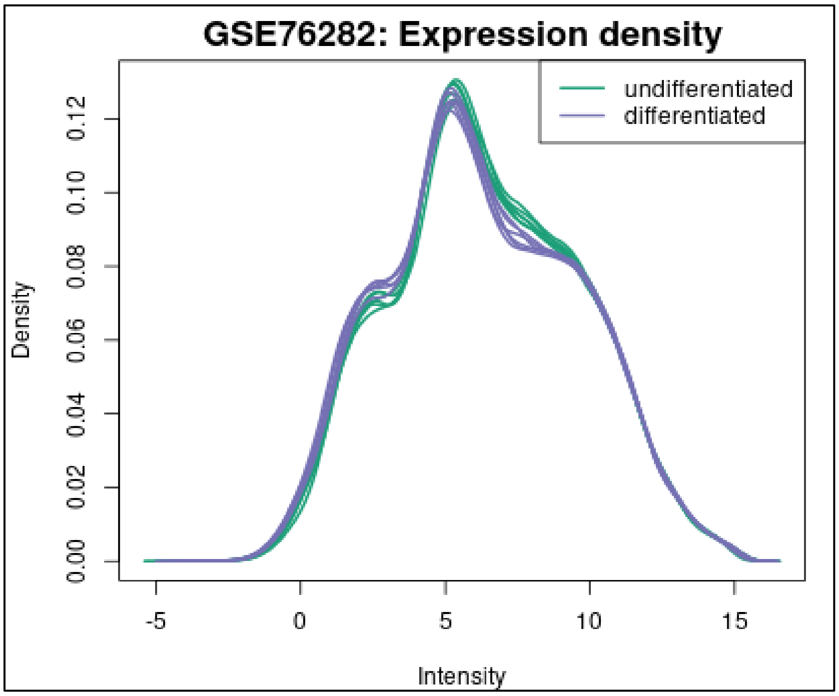


Figure 8: Expression density plot

In the expression density graph, the x-axis represents gene expression values and the y-axis represents the density or frequency of occurrence of those expression values. The x-axis corresponds to the range of gene expression levels observed in the set of differentially expressed genes (DEGs). The y-axis represents the density or frequency of occurrence of gene expression values. It reflects the probability of finding a particular expression value in DEGs. The y-axis values are non-negative and represent the relative density or abundance of expression values at different points along the x-axis.

By looking at this graph, it seems that the peaks and patterns are almost the same between the groups. It can be said that the gene expression patterns are similar between the compared groups.

Consistent peaks and patterns indicate that genes have the same or similar regulatory mechanisms between groups. This could indicate the existence of common transcription factors, signaling pathways, or other regulatory factors that affect gene expression in a coordinated manner across groups. It can also be concluded that probably the underlying biological processes or pathways are common in both groups or have similar activity levels. Therefore, the compared groups may have similar functional characteristics or biological responses. These similar biological processes may be due to the presence of conserved molecules.

We expect outliers to appear as individual peaks that deviate from the general pattern of the density curve. According to the graph, it seems that there is no outlier expression.

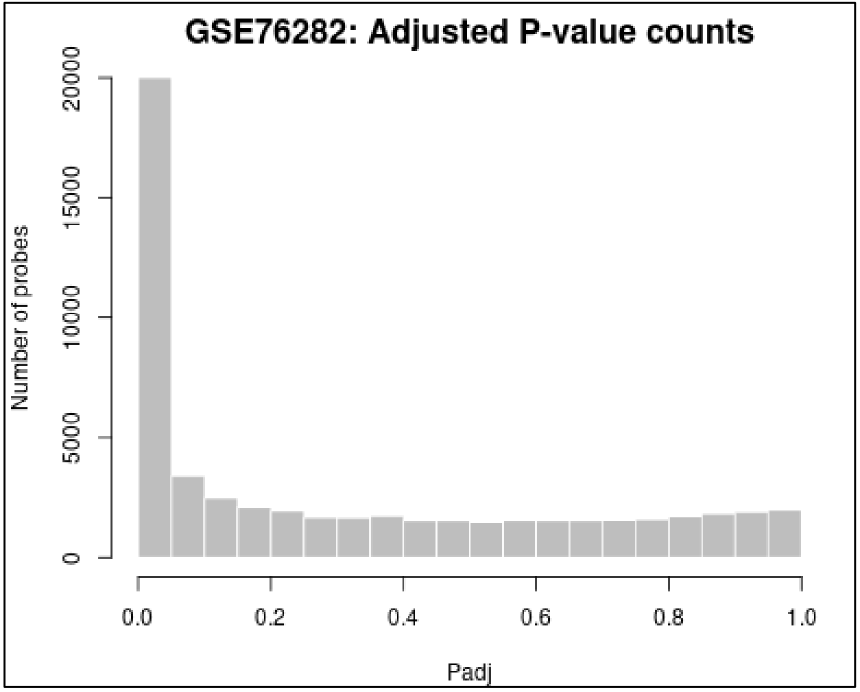


Figure 9: Adjusted p-value count plot

The x-axis is the adjusted p-value. This value is the smallest level of significance and false discovery rate that exists in multiple comparison tests. In other words, in examining the difference between different genes in different groups, the p-value is adjusted.

The accumulation of smaller p-values can be seen on the left side of the x-axis. This indicates a greater number of findings or meaningful evidence against the null hypothesis. In other words, there are many p-values that show statistically significant results, indicating a strong difference between expression in differentiated and undifferentiated cell groups. Therefore, it is expected that the uniform distribution of p values in the entire range of the x-axis confirms the null hypothesis.

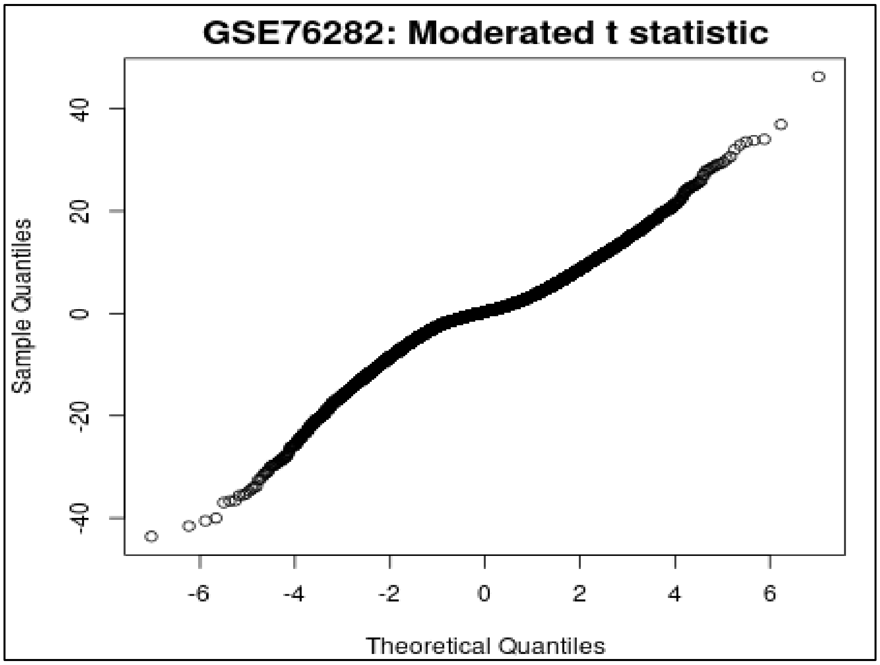


Figure 10: Moderated t statistic plot

The t-test assumes that gene expression levels approximately follow a normal distribution. With the assumption of normality, it is possible to use established statistical theory, p-value calculation, confidence intervals and hypothesis testing. These statistical measures help to determine the significance of DEG between groups. In the graph, the observed t values can be seen with the expected values assuming a normal distribution. If the t statistic exactly follows a diagonal line on the graph, it indicates a normal distribution. Therefore, it can be said that our dataset and the expression of genes almost follows the normal state.

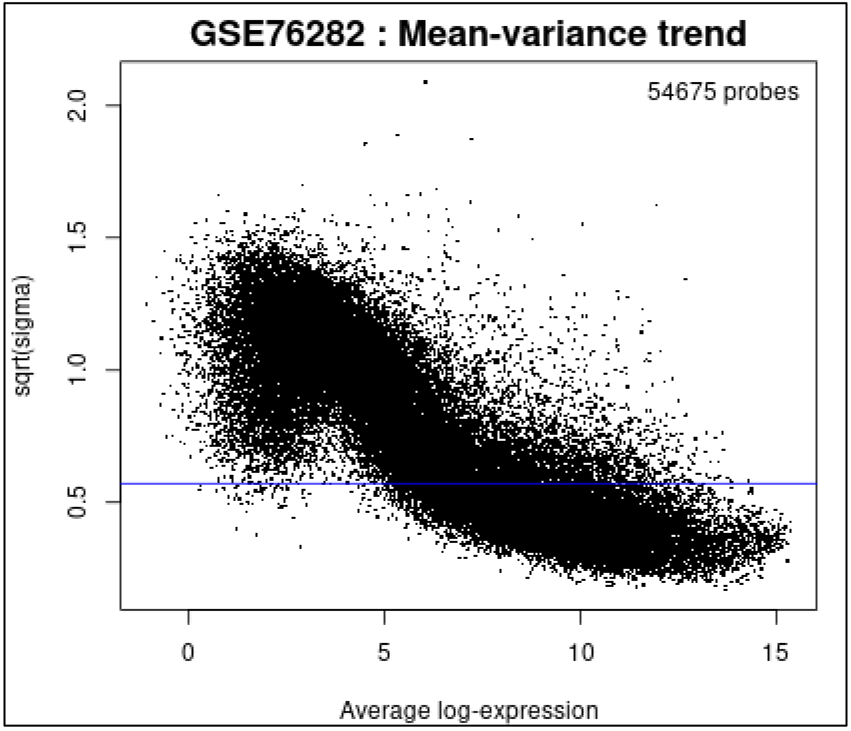


Figure 11: Mean-variance trend plot

Values that deviate significantly from the diagonal line in the QQ plot may indicate the presence of outliers. There seem to be no outliers.

In the mean-variance diagram, each point represents a gene. After fitting a linear model is used to check the mean variance relationship of the expression data, so it does not need to select a group and it shows the diversity in the data. This chart can help in evaluating whether or not to use the precision weights option in data analysis.

The precision weights option in DEG analysis aims to consider the trend of the average variance observed in the M-A chart. In this method, specific weights are given to specific genes based on their expression levels and diversity. In situations where the trend of the variance and the mean are interdependent and have a clear relationship (the variance increases or decreases continuously as the mean increases), this method gives more weight to genes with significantly different expression patterns and can improve the accuracy of DEG detection. This work helps to give more importance to genes with different and significant expressions in the analysis.

It seems that there is no clear and reliable trend between the variance and the mean, so giving weight to the genes and selecting the precision weights option does not help much to detect DEGs.

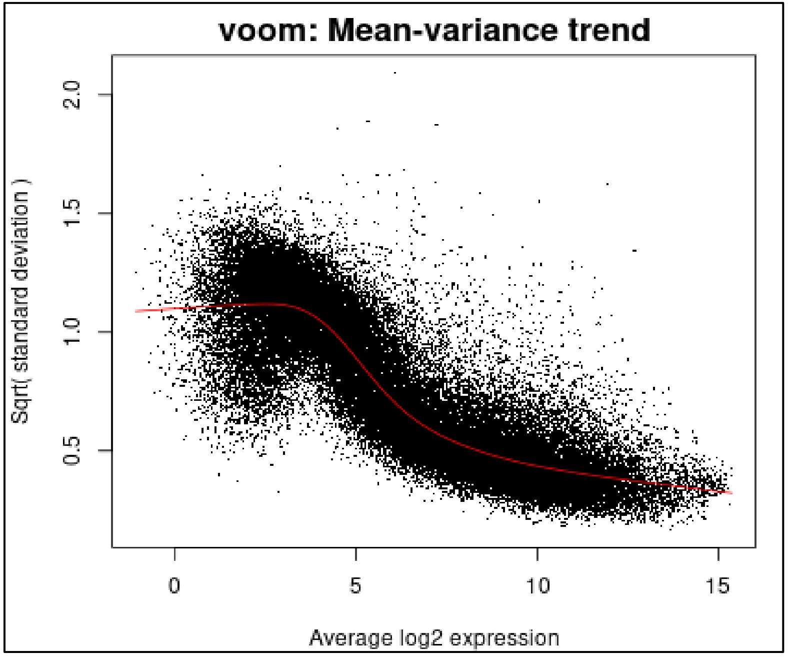
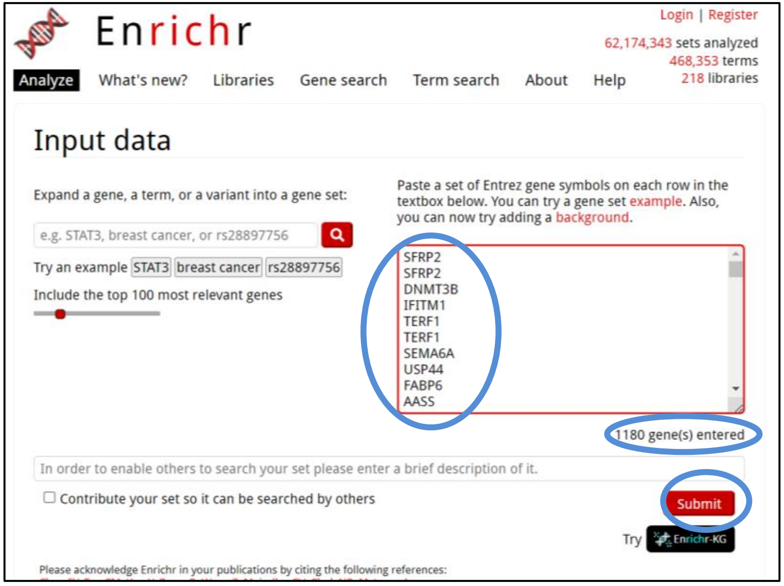


Figure 12: Mean-variance trend plot

3. Do enrichment analysis for DEGs and show in which Biological Processes, Molecular Functions, and Pathways they play a role.



62,174,343 sets analyzed
468,353 terms
218 libraries

Input data

Expand a gene, a term, or a variant into a gene set:

e.g. STAT3, breast cancer, or rs28897756

Try an example STAT3 breast cancer rs28897756

Include the top 100 most relevant genes

Paste a set of Entrez gene symbols on each row in the textbox below. You can try a gene set example. Also, you can now try adding a background.

SFRP2
SFRP2
DNMT3B
IFITM1
TERF1
TERF1
SEMA6A
USP44
FABP6
AASS

1180 gene(s) entered

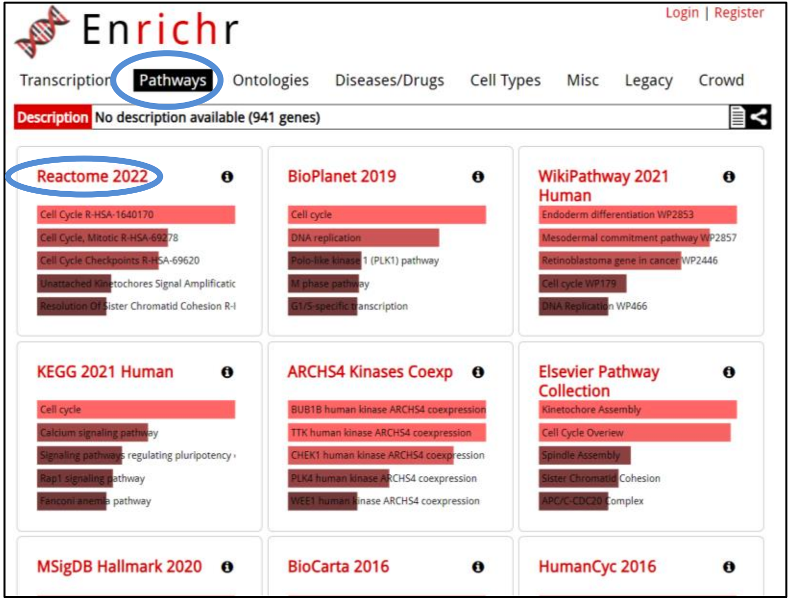
In order to enable others to search your set please enter a brief description of it.

☐ Contribute your set so it can be searched by others

Please acknowledge Enrichr in your publications by citing the following references:

Figure 13: Input to Enrichr site

In this section, we input the filtered genes based on adjusted p-value (values less than 0.01) and LogFC (values greater than 2) into Enrichr (1180 genes). Also, genes whose names were not mentioned in the downloaded file from GEO or had multiple genes or miRNAs written were removed. After submitting the data, the following results are obtained:



Description No description available (941 genes)

Reactome 2022
Cell Cycle R-HSA-1640170
Cell Cycle, Mitotic R-HSA-69278
Cell Cycle Checkpoints R-HSA-69620
Unattached Kinetochores Signal Amplificatic
Resolution Of Sister Chromatid Cohesion R-I

BioPlanet 2019
Cell cycle
DNA replication
Polo-like kinase 1 (PLK1) pathway
M phase pathway
G1/S-specific transcription

WikiPathway 2021 Human
Endoderm differentiation WP2853
Mesodermal commitment pathway WP2857
Retinoblastoma gene in cancer WP2446
Cell cycle WP179
DNA Replication WP466

KEGG 2021 Human
Cell cycle
Calcium signaling pathway
Signaling pathways regulating pluripotency
Rap1 signaling pathway
Fanconi anemia pathway

ARCHS4 Kinases Coexp
BUB1B human kinase ARCHS4 coexpression
TTK human kinase ARCHS4 coexpression
CHEK1 human kinase ARCHS4 coexpression
PLK4 human kinase ARCHS4 coexpression
WEE1 human kinase ARCHS4 coexpression

Elsevier Pathway Collection
Kinetochore Assembly
Cell Cycle Overiew
Spindle Assembly
Sister Chromatid Cohesion
APC/C-CDC20 Complex

MSigDB Hallmark 2020

BioCarta 2016

HumanCyc 2016

Figure 14: In the Pathways section, a number of biological pathways in which these genes are involved are shown (Click here to access this page)

If we click on the first result, which is for the 2022 Reactome database, the types of pathways in which the input genes are involved are sorted based on p-value.

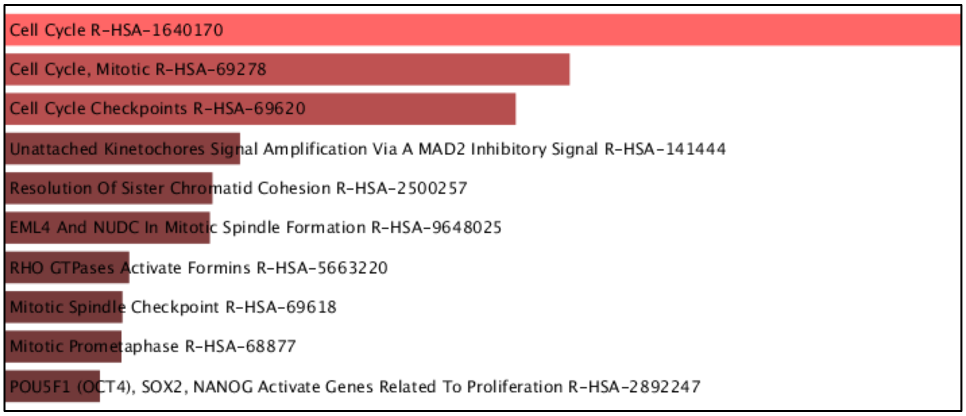
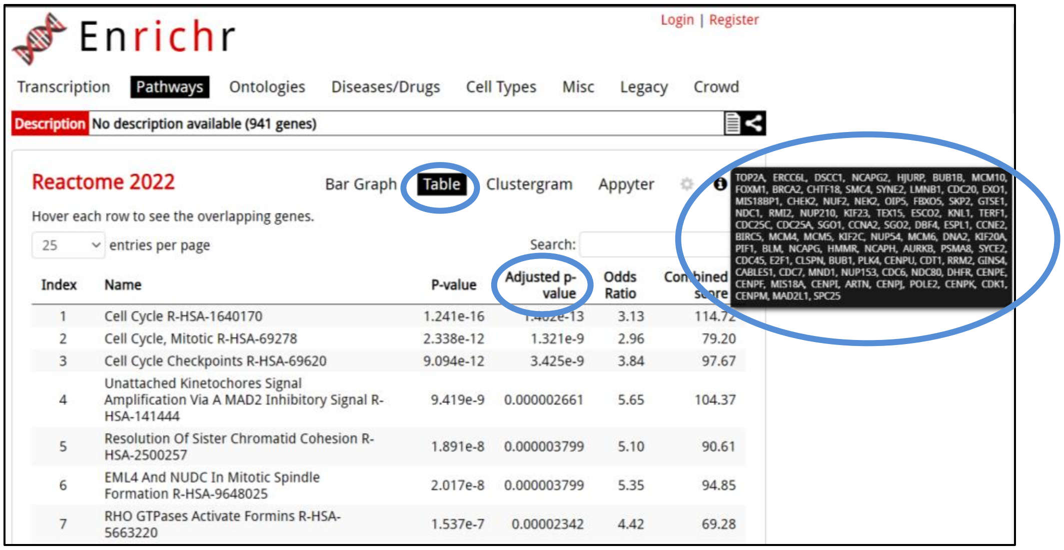


Figure 15: Graph of biological pathways based on input genes in Enrichr site

The results can also be viewed in a table format.



Description No description available (941 genes)

Reactome 2022

Hover each row to see the overlapping genes.

25 entries per page

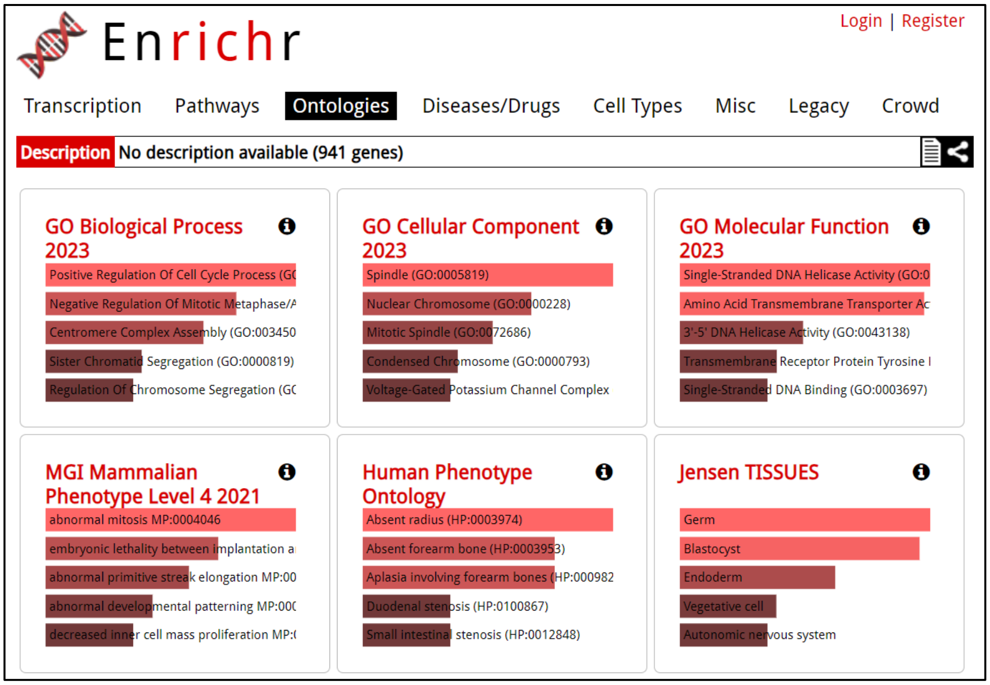
Search:

TOP2A, ERCC6L, DSCC1, NCAPG2, HJURP, BUB1B, MCM10, FOXM1, BRCA2, CHTF18, SMC4, SYNE2, LMNB1, CDC20, EXO1, MIS18BP1, CHEK2, NUF2, NEK2, OIP5, FBXO5, SKP2, GTSE1, NDC1, RMI2, NUP210, KIF23, TEX15, ESCO2, KNL1, TERF1, CDC25C, CDC25A, SGO1, CCNA2, SGO2, DBF4, ESPL1, CCNE2, BIRC5, MCM4, MCM5, KIF2C, NUP54, MCM6, DNA2, KIF20A, PIF1, BLM, NCAPG, HMMR, NCAPH, AURKB, PSMA8, SYCE2, CDC45, E2F1, CLSPN, BUB1, PLK4, CENPU, CDT1, RRM2, GINS4, CABLES1, CDC7, MND1, NUP153, CDC6, NDC80, DHFR, CENPE, CENPF, MIS18A, CENPI, ARTN, CENPJ, POLE2, CENPK, CDK1, CENPM, MAD2L1, SPC25

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	Cell Cycle R-HSA-1640170	1.241e-16	1.462e-13	3.13	114.72
2	Cell Cycle, Mitotic R-HSA-69278	2.338e-12	1.321e-9	2.96	79.20
3	Cell Cycle Checkpoints R-HSA-69620	9.094e-12	3.425e-9	3.84	97.67
4	Unattached Kinetochores Signal Amplification Via A MAD2 Inhibitory Signal R-HSA-141444	9.419e-9	0.000002661	5.65	104.37
5	Resolution Of Sister Chromatid Cohesion R-HSA-2500257	1.891e-8	0.000003799	5.10	90.61
6	EML4 And NUDC In Mitotic Spindle Formation R-HSA-9648025	2.017e-8	0.000003799	5.35	94.85
7	RHO GTPases Activate Formins R-HSA-5663220	1.537e-7	0.00002342	4.42	69.28

Figure 16: Table related to biological pathways of input genes (Click here to access the full table data)

As can be seen in the image, different cell pathways are sorted based on p-value. The lower the pathway p-value, the more significant data we have, so more genes from the input list to the site are involved in these pathways. By hovering over each of the pathways, it also displays the genes involved.



Enrichr

Login | Register

Transcription Pathways Ontologies Diseases/Drugs Cell Types Misc Legacy Crowd

Description No description available (941 genes)

GO Biological Process 2023
- Positive Regulation Of Cell Cycle Process (GC
- Negative Regulation Of Mitotic Metaphase/A
- Centromere Complex Assembly (GO:003450
- Sister Chromatid Segregation (GO:0000819)
- Regulation Of Chromosome Segregation (GC

GO Cellular Component 2023
- Spindle (GO:0005819)
- Nuclear Chromosome (GO:0000228)
- Mitotic Spindle (GO:0072686)
- Condensed Chromosome (GO:0000793)
- Voltage-Gated Potassium Channel Complex

GO Molecular Function 2023
- Single-Stranded DNA Helicase Activity (GO:0
- Amino Acid Transmembrane Transporter Ac
- 3'-5' DNA Helicase Activity (GO:0043138)
- Transmembrane Receptor Protein Tyrosine I
- Single-Stranded DNA Binding (GO:0003697)

MGI Mammalian Phenotype Level 4 2021
- abnormal mitosis MP:0004046
- embryonic lethality between implantation a
- abnormal primitive streak elongation MP:00
- abnormal developmental patterning MP:00(
- decreased inner cell mass proliferation MP:(

Human Phenotype Ontology
- Absent radius (HP:0003974)
- Absent forearm bone (HP:0003953)
- Aplasia involving forearm bones (HP:000982
- Duodenal stenosis (HP:0100867)
- Small intestinal stenosis (HP:0012848)

Jensen TISSUES
- Germ
- Blastocyst
- Endoderm
- Vegetative cell
- Autonomic nervous system

Figure 17: Types of data related to Ontologies section

To find the Biological Processes and Molecular Function of the input genes, we refer to the Ontology section. In this section, as can be seen in the image above, the data related to Biological Process and Molecular Functions for 2023 have been classified from the Gene Ontology database, and more details can be obtained by clicking on each one.

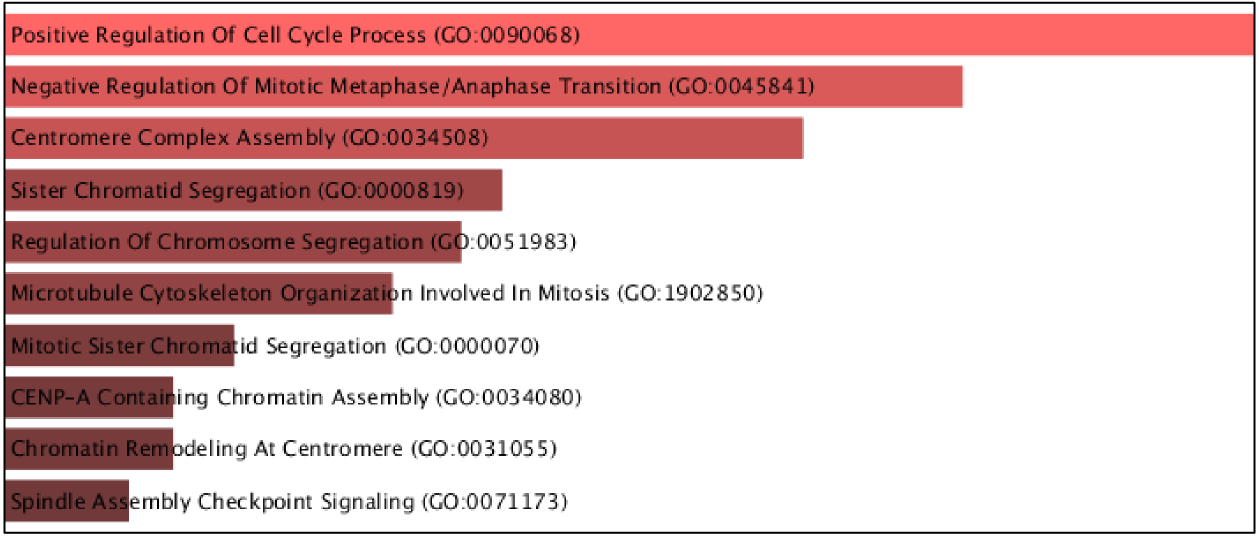
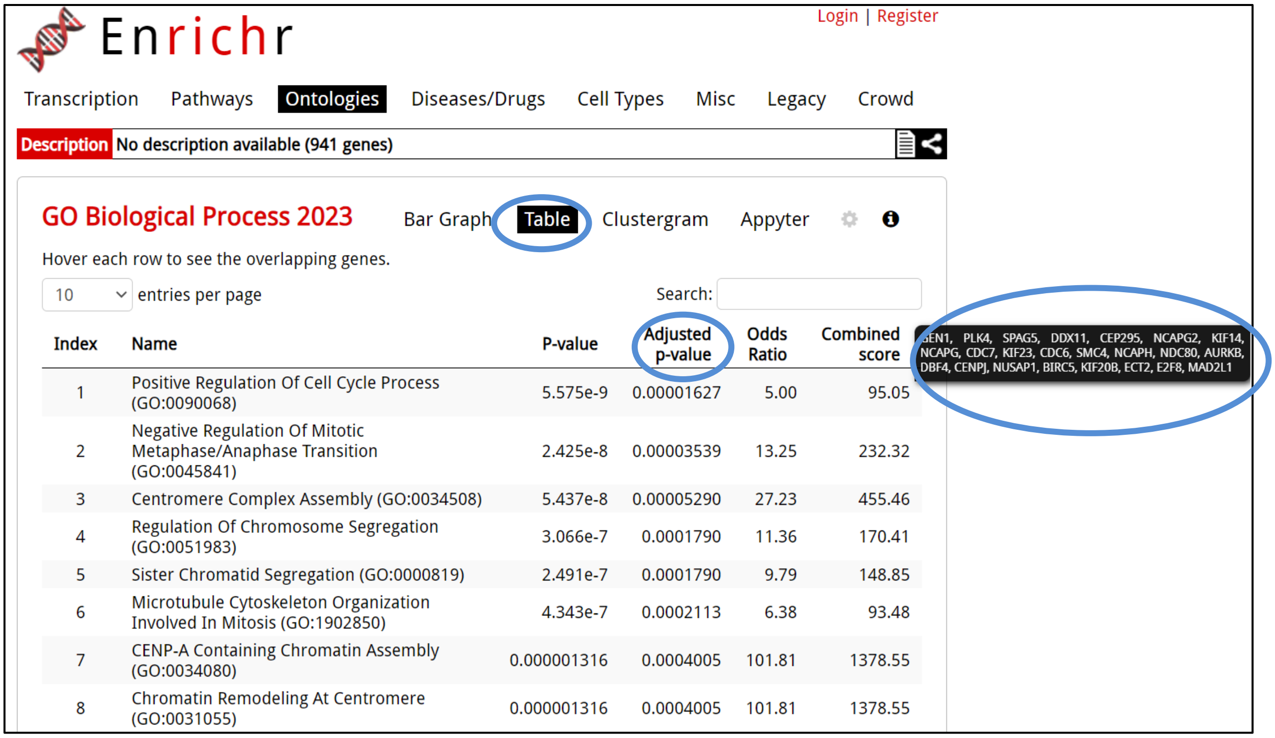


Figure 18: Biological Process bar graph for 2023 from GO database

In this diagram, each of the biological pathways are arranged according to the p-value, so that the positive regulation pathway of the cell cycle is the most significant data and contains the largest number of genes.



Description No description available (941 genes)

GO Biological Process 2023

Hover each row to see the overlapping genes.

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	Positive Regulation Of Cell Cycle Process (GO:0090068)	5.575e-9	0.00001627	5.00	95.05
2	Negative Regulation Of Mitotic Metaphase/Anaphase Transition (GO:0045841)	2.425e-8	0.00003539	13.25	232.32
3	Centromere Complex Assembly (GO:0034508)	5.437e-8	0.00005290	27.23	455.46
4	Regulation Of Chromosome Segregation (GO:0051983)	3.066e-7	0.0001790	11.36	170.41
5	Sister Chromatid Segregation (GO:0000819)	2.491e-7	0.0001790	9.79	148.85
6	Microtubule Cytoskeleton Organization Involved In Mitosis (GO:1902850)	4.343e-7	0.0002113	6.38	93.48
7	CENP-A Containing Chromatin Assembly (GO:0034080)	0.000001316	0.0004005	101.81	1378.55
8	Chromatin Remodeling At Centromere (GO:0031055)	0.000001316	0.0004005	101.81	1378.55

GEN1, PLK4, SPAG5, DDX11, CEP295, NCAPG2, KIF14, NCAPG, CDC7, KIF23, CDC6, SMC4, NCAPH, NDC80, AURKB, DBF4, CENPJ, NUSAP1, BIRC5, KIF20B, ECT2, E2F8, MAD2L1

Figure 19: Table related to the Biological Process of input genes (click here to access the complete data of the table)

Similar to the previous part, you can see the data in the form of a table. If we refer to each line of the table, it shows the genes involved in this biological pathway. In this table, the data are arranged according to p-value.

For the Biological Process section, we do the same as the Biological Process.

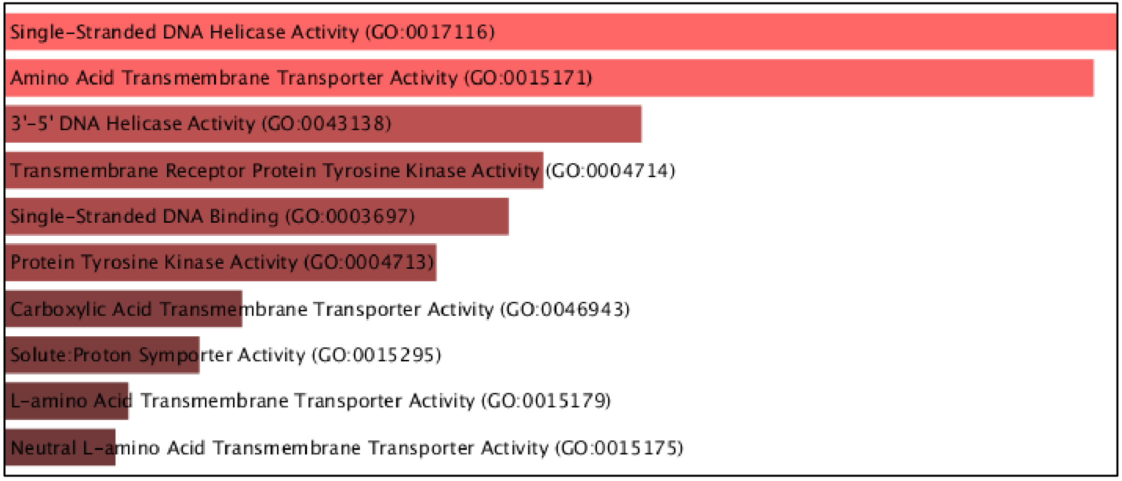
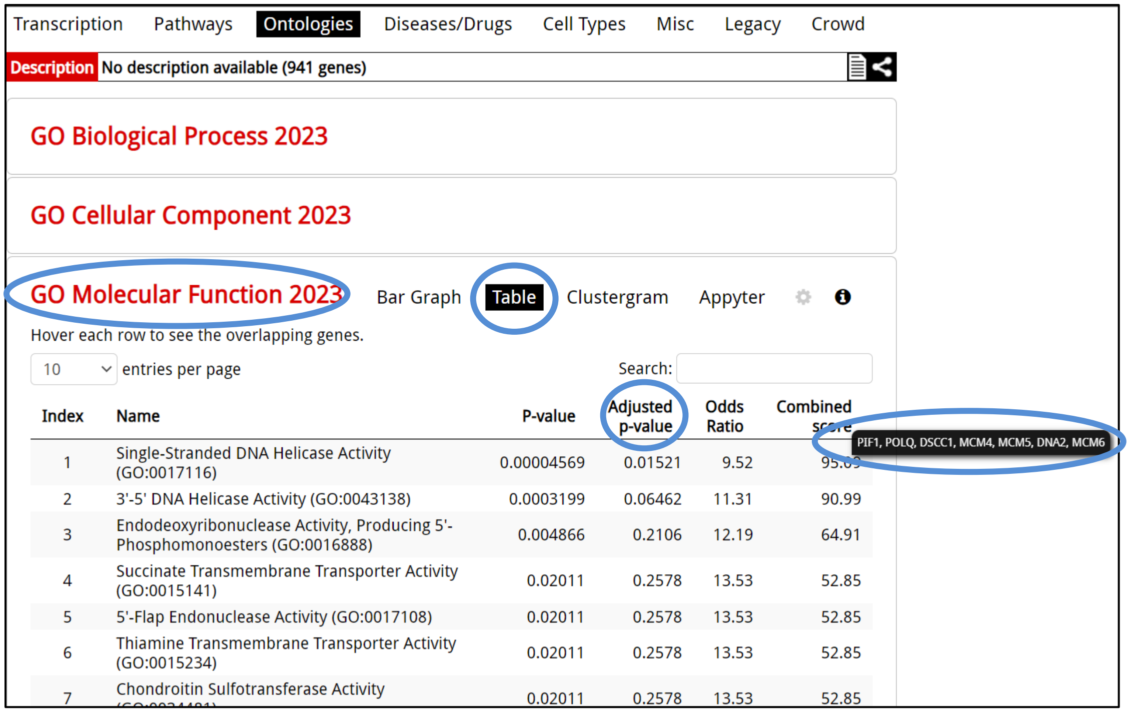


Figure 20: Bar graph of Molecular Function for 2023 from GO database



Description No description available (941 genes)

GO Biological Process 2023

GO Cellular Component 2023

GO Molecular Function 2023

Hover each row to see the overlapping genes.

10 entries per page

Search:

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	Single-Stranded DNA Helicase Activity (GO:0017116)	0.00004569	0.01521	9.52	95.09
2	3'-5' DNA Helicase Activity (GO:0043138)	0.0003199	0.06462	11.31	90.99
3	Endodeoxyribonuclease Activity, Producing 5'-Phosphomonoesters (GO:0016888)	0.004866	0.2106	12.19	64.91
4	Succinate Transmembrane Transporter Activity (GO:0015141)	0.02011	0.2578	13.53	52.85
5	5'-Flap Endonuclease Activity (GO:0017108)	0.02011	0.2578	13.53	52.85
6	Thiamine Transmembrane Transporter Activity (GO:0015234)	0.02011	0.2578	13.53	52.85
7	Chondroitin Sulfotransferase Activity (GO:0034481)	0.02011	0.2578	13.53	52.85

PIF1, POLQ, DSCC1, MCM4, MCM5, DNA2, MCM6

Figure 21: Table related to Molecular Function for 2023 from GO database

As can also be seen in the table above, the most important function of these genes is their effect on the activity of single-stranded DNA helicases.

The table related to the data of each section of this question is also available at this link.

4. For the upregulated genes in one of the groups, download and plot the protein-protein interaction and gene regulatory network from relevant databases using Cytoscape. Please color the nodes based on degree and discuss the genes with the highest degree and Betweenness Centrality values in the networks.

Genes with positive LogFC indicate higher expression of these genes in the undifferentiated group. Therefore, by applying a filter in the related table, we can find the upregulated genes in this group. Due to the large number of genes and easier analysis of data, genes with LogFC> 2 were filtered. (List of filtered genes is available via this link) We enter the obtained gene list in the Search section of the String website and using STRING: protein query we draw the Protein-Protein Interaction network.

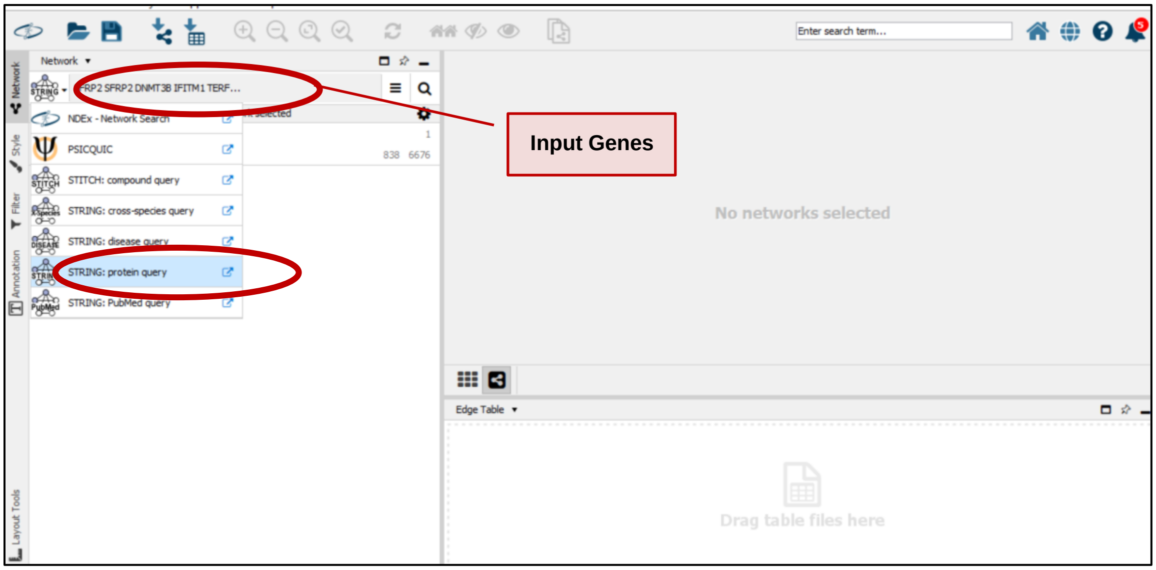


Figure 22: Input of filtered genes to String site using Protein query from STRING database

After receiving the network related to these genes, we color the nodes based on degree using the Style window. Thus, nodes related to genes with higher degrees are displayed more brightly than other genes. To examine Betweenness, we need a connected network. For various reasons, including lack of sufficient information about some genes in the STRING database, some nodes in the resulting network had zero degrees. By removing these nodes, a connected network was finally obtained that is suitable for examining Betweenness.

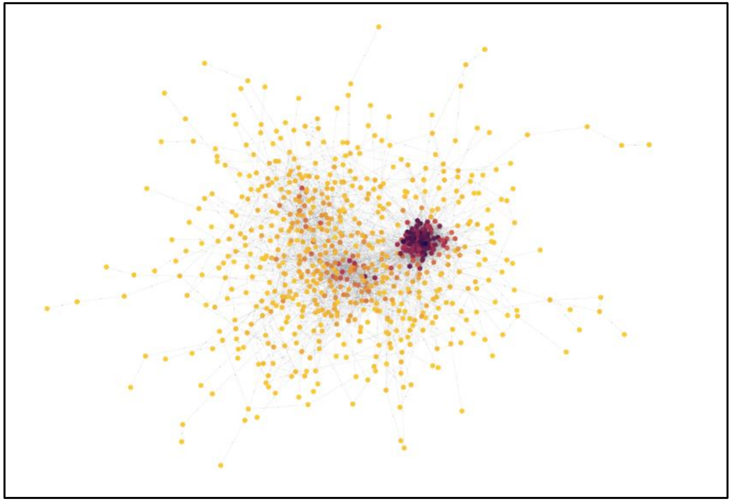
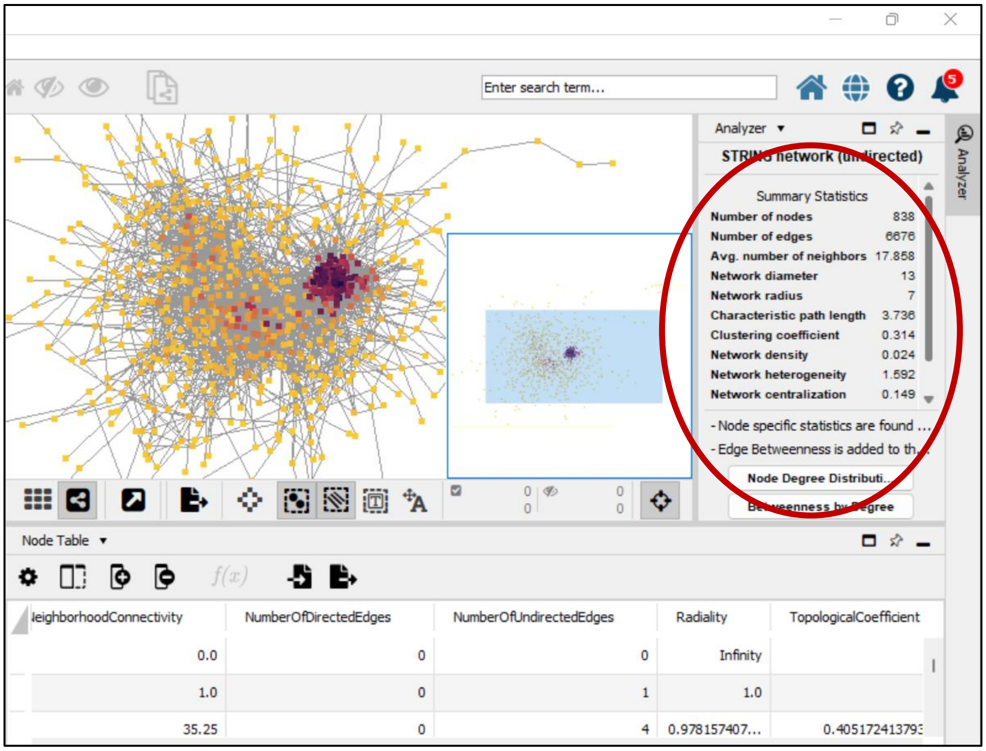


Figure 23: The drawn ppi network in which nodes are colored based on degree. (Higher degree, more colorful)

The drawn network has 838 nodes and 6676 edges. Other information is displayed after network analysis as follows. Also, to access the table of nodes of this network after analysis, click on this link.



STRING network (undirected)

Summary Statistics

Number of nodes	838
Number of edges	6676
Avg. number of neighbors	17.858
Network diameter	13
Network radius	7
Characteristic path length	3.736
Clustering coefficient	0.314
Network density	0.024
Network heterogeneity	1.592
Network centralization	0.149

- Node specific statistics are found ...
- Edge Betweenness is added to th...

leighborhoodConnectivity	NumberOfDirectedEdges	NumberOfUndirectedEdges	Radiality	TopologicalCoefficient
0.0	0	0	Infinity	
1.0	0	1	1.0	
35.25	0	4	0.978157407...	0.40517241379?

Figure 24: Summary of ppi network data analysis

In order to discuss the genes with the highest degree and compare them using the Betweenness Centrality criterion, it is necessary to plot the Betweenness Centrality versus Degree.

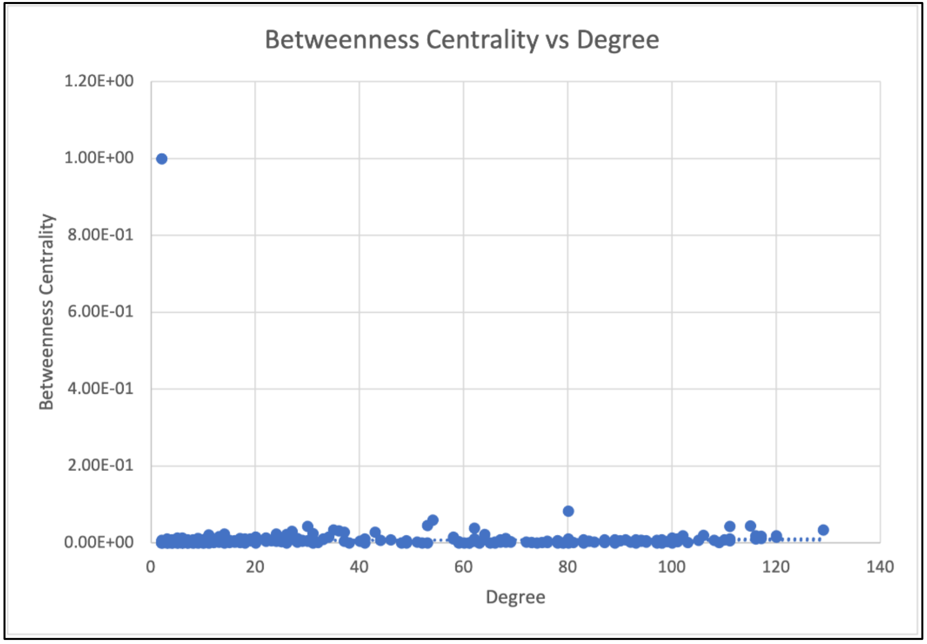


Figure 25: Betweenness Centrality vs Degree plot

The above table does not give us much information because the scales of these two criteria are not coordinated. One of them includes values greater than 100 and the other criterion includes values less than 1. Therefore, to better understand the relationship between these two criteria in the resulting network, we plot Betweenness Centrality based on the logarithm base 10 so that the following figure is obtained.

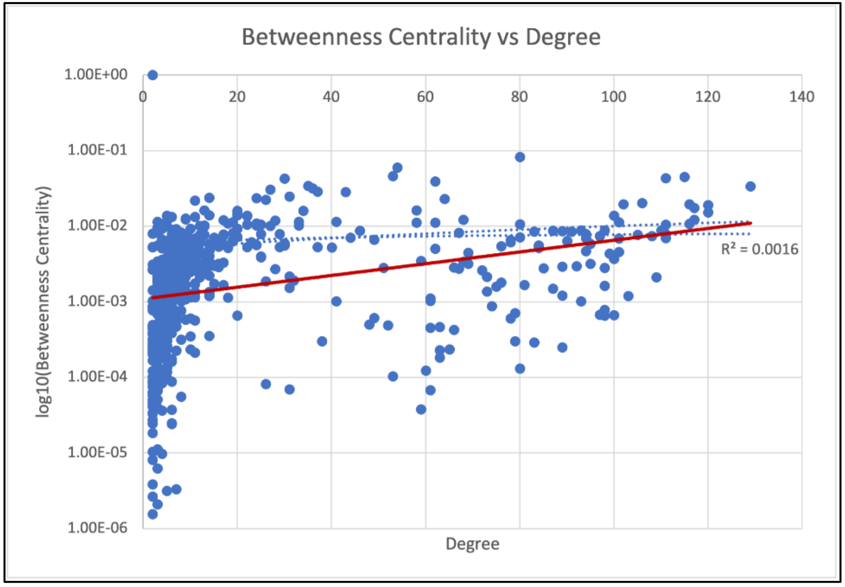


Figure 26: Log10(Betweenness Centrality) vs Degree plot

According to the figure above, no particular relationship can be considered between these two criteria; Although in general it can be said that with increasing degree, the value of Betweenness centrality increases.

It is also possible to color the nodes of the network based on Betweenness centrality and compare it with the previous coloration based on degree to analyze the relationship between these two criteria.

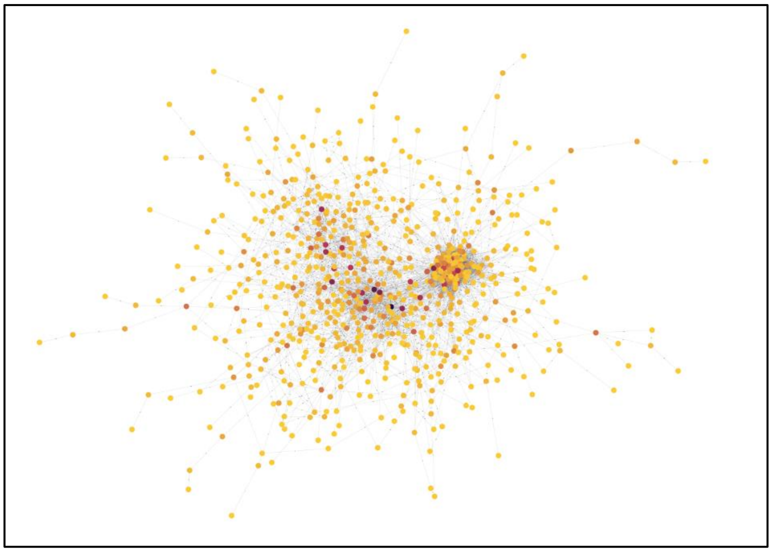


Figure 27: Network node coloring based on Betweenness Centrality value

As shown in the related charts and images of network coloring, there is no clear relationship between the two criteria of Degree and Betweenness Centrality.

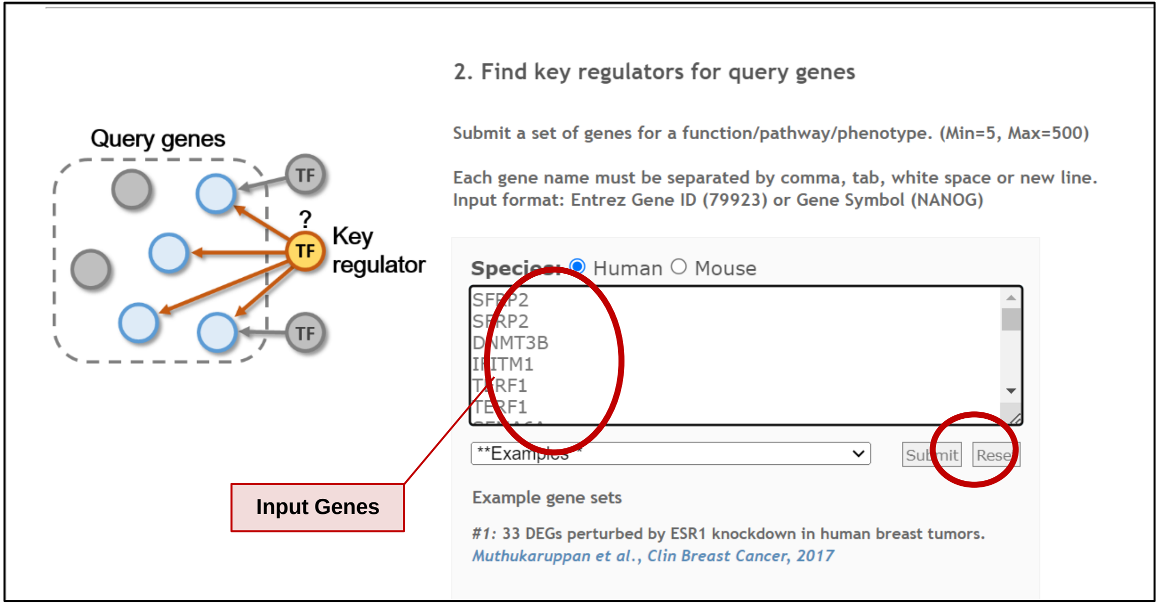
If Betweenness Centrality and Degree are not correlated, it indicates that the position of nodes in terms of network connectivity and penetration is not directly related. Here there are several possible interpretations:

1. Structural role: Nodes with high degree have many direct connections indicating their importance in terms of local connectivity. On the other hand, nodes with high Betweenness Centrality act as bridges or intermediaries in the network and facilitate the flow of information between other nodes. When these two metrics are uncorrelated, it shows that nodes with high degree centrality may not necessarily lie on the shortest paths between other nodes or have a significant impact on information flow.

2. Network modularity: In some networks, nodes may exhibit modular or clustered structures. In such cases, degree measures connectivity within modules, while Betweenness Centrality indicates connectivity between modules. If the network has distinct modules with limited inter-module connections, high degree nodes within modules may have low Betweenness Centrality as they do not bridge different modules.

3. Network dynamics: The correlation between Degree and Betweenness Centrality can be influenced by the dynamic nature of the network. In evolving networks or networks with changing connections over time, the relationship between these two metrics may be different. Nodes that have high degree in one network snapshot, may not retain high Betweenness Centrality as the network structure evolves.

To draw the Gene Regulatory network, we also use the TRRUST database. This database takes as input a list of genes and gives us a set of key regulators along with their target genes, which we can use to draw the gene regulatory network in Cytoscape.



2. Find key regulators for query genes

Submit a set of genes for a function/pathway/phenotype. (Min=5, Max=500)

Each gene name must be separated by comma, tab, white space or new line.
Input format: Entrez Gene ID (79923) or Gene Symbol (NANOG)

Species: ● Human ○ Mouse

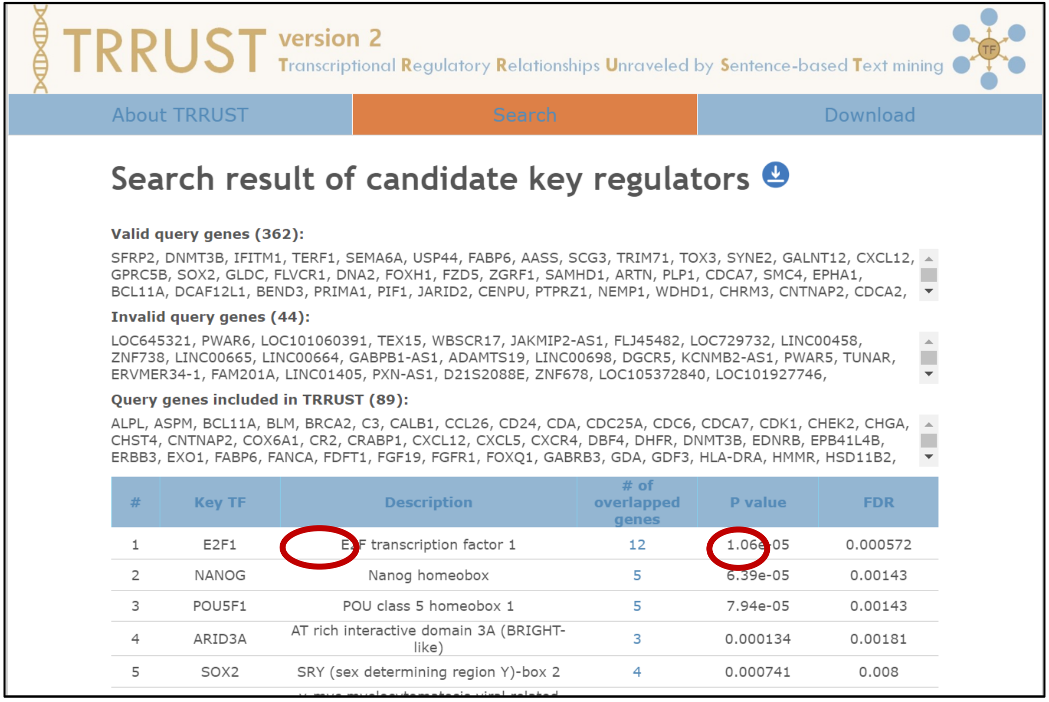
SFRP2
SFRP2
DNMT3B
IFITM1
TERF1
TERF1

Example gene sets

#1: 33 DEGs perturbed by ESR1 knockdown in human breast tumors.
Muthukaruppan et al., Clin Breast Cancer, 2017

Figure 28: Input to TRRUST database to find Key Regulators

After submitting the input genes, this database outputs a list of key regulatory transcription factors. Since our total number of genes is 1180 and this database cannot take more than 500 genes as input, we enter the first 500 upregulated genes.


TRRUST version 2
Transcriptional Regulatory Relationships Unraveled by Sentence-based Text mining
TF
About TRRUST
Search
Download

Search result of candidate key regulators

Valid query genes (362):

SFRP2, DNMT3B, IFITM1, TERF1, SEMA6A, USP44, FABP6, AASS, SCG3, TRIM71, TOX3, SYNE2, GALNT12, CXCL12, GPRC5B, SOX2, GLDC, FLVCR1, DNA2, FOXH1, FZD5, ZGRF1, SAMHD1, ARTN, PLP1, CDCA7, SMC4, EPHA1, BCL11A, DCAF12L1, BEND3, PRIMA1, PIF1, JARID2, CENPU, PTPRZ1, NEMP1, WDHD1, CHRM3, CNTNAP2, CDCA2,

Invalid query genes (44):

LOC645321, PWAR6, LOC101060391, TEX15, WBSCR17, JAKMIP2-AS1, FLJ45482, LOC729732, LINC00458, ZNF738, LINC00665, LINC00664, GABPB1-AS1, ADAMTS19, LINC00698, DGCR5, KCNMB2-AS1, PWAR5, TUNAR, ERVMER34-1, FAM201A, LINC01405, PXN-AS1, D21S2088E, ZNF678, LOC105372840, LOC101927746,

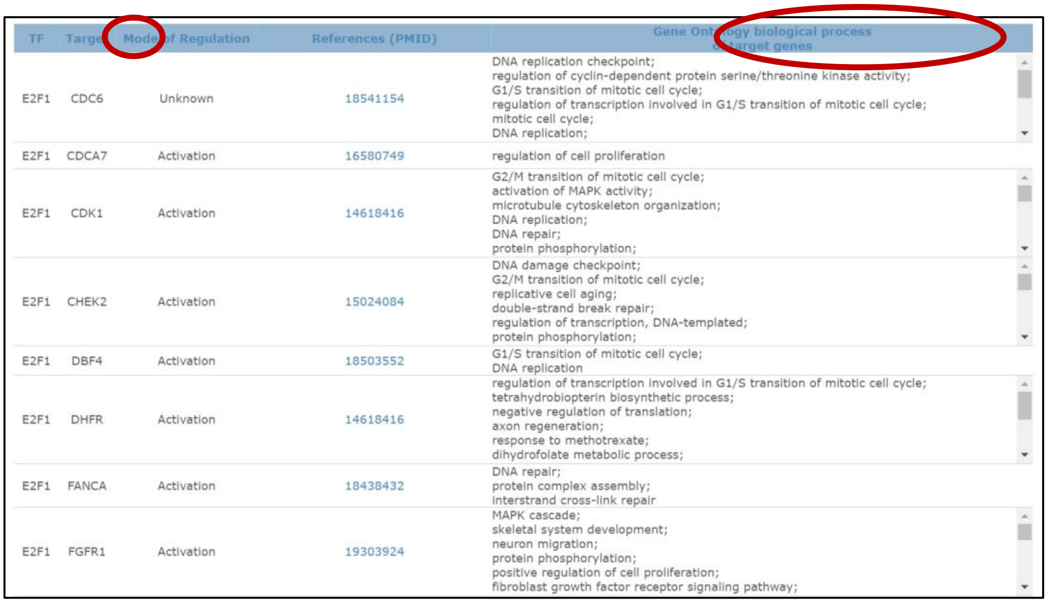
Query genes included in TRRUST (89):

ALPL, ASPM, BCL11A, BLM, BRCA2, C3, CALB1, CCL26, CD24, CDA, CDC25A, CDC6, CDCA7, CDK1, CHEK2, CHGA, CHST4, CNTNAP2, COX6A1, CR2, CRABP1, CXCL12, CXCL5, CXCR4, DBF4, DHFR, DNMT3B, EDNRB, EPB41L4B, ERBB3, EXO1, FABP6, FANCA, FDFT1, FGF19, FGFR1, FOXQ1, GABRB3, GDA, GDF3, HLA-DRA, HMMR, HSD11B2,

#	Key TF	Description	# of overlapped genes	P value	FDR
1	E2F1	E2F transcription factor 1	12	1.06e-05	0.000572
2	NANOG	Nanog homeobox	5	6.39e-05	0.00143
3	POU5F1	POU class 5 homeobox 1	5	7.94e-05	0.00143
4	ARID3A	AT rich interactive domain 3A (BRIGHT-like)	3	0.000134	0.00181
5	SOX2	SRY (sex determining region Y)-box 2	4	0.000741	0.008

Figure 29: Output regulator table (Full list of Key Regulators can be viewed via this link)

If we click on any of these regulators, it will display a list of the target genes of this regulator along with the type of regulation and also the Gene Ontology of the target genes.



TF	Target	Mode of Regulation	References (PMID)	Gene Ontology biological process of target genes

E2F1	CDC6	Unknown	18541154	DNA replication checkpoint;
				regulation of cyclin-dependent protein serine/threonine kinase activity;
				G1/S transition of mitotic cell cycle;
				regulation of transcription involved in G1/S transition of mitotic cell cycle;
				mitotic cell cycle;
				DNA replication;

E2F1	CDCA7	Activation	16580749	regulation of cell proliferation

E2F1	CDK1	Activation	14618416	G2/M transition of mitotic cell cycle;
				activation of MAPK activity;
				microtubule cytoskeleton organization;
				DNA replication;
				DNA repair;
				protein phosphorylation;

E2F1	CHEK2	Activation	15024084	DNA damage checkpoint;
				G2/M transition of mitotic cell cycle;
				replicative cell aging;
				double-strand break repair;
				regulation of transcription, DNA-templated;
				protein phosphorylation;

E2F1	DBF4	Activation	18503552	G1/S transition of mitotic cell cycle; DNA replication

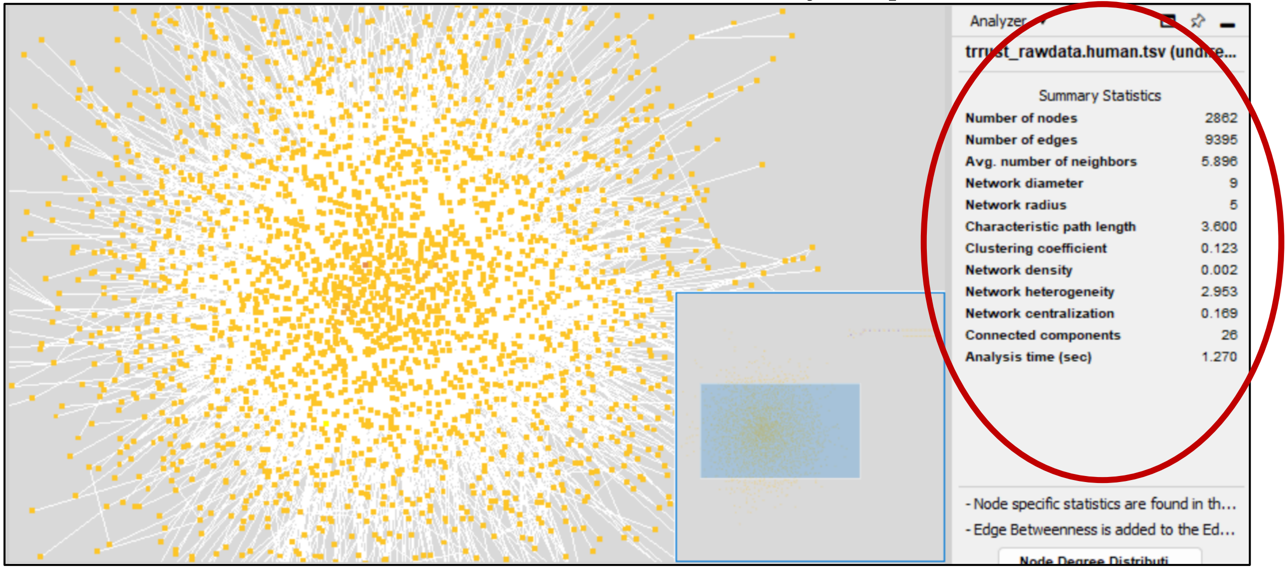
E2F1	DHFR	Activation	14618416	regulation of transcription involved in G1/S transition of mitotic cell cycle;
				tetrahydrobiopterin biosynthetic process;
				negative regulation of translation;
				axon regeneration;
				response to methotrexate;
				dihydrofolate metabolic process;

E2F1	FANCA	Activation	18438432	DNA repair;
				protein complex assembly;
				interstrand cross-link repair

E2F1	FGFR1	Activation	19303924	MAPK cascade;
				skeletal system development;
				neuron migration;
				protein phosphorylation;
				positive regulation of cell proliferation;
				fibroblast growth factor receptor signaling pathway;

Figure 30: Table related to target genes of regulator E2F1

To draw the gene regulation network, first, since we do not have access to the Regnetwork database, we download the complete network of Transcription Factors and their target genes in humans from the TRRUST database and draw its network in Cytoscape.



trrust_rawdata.human.tsv (undire...

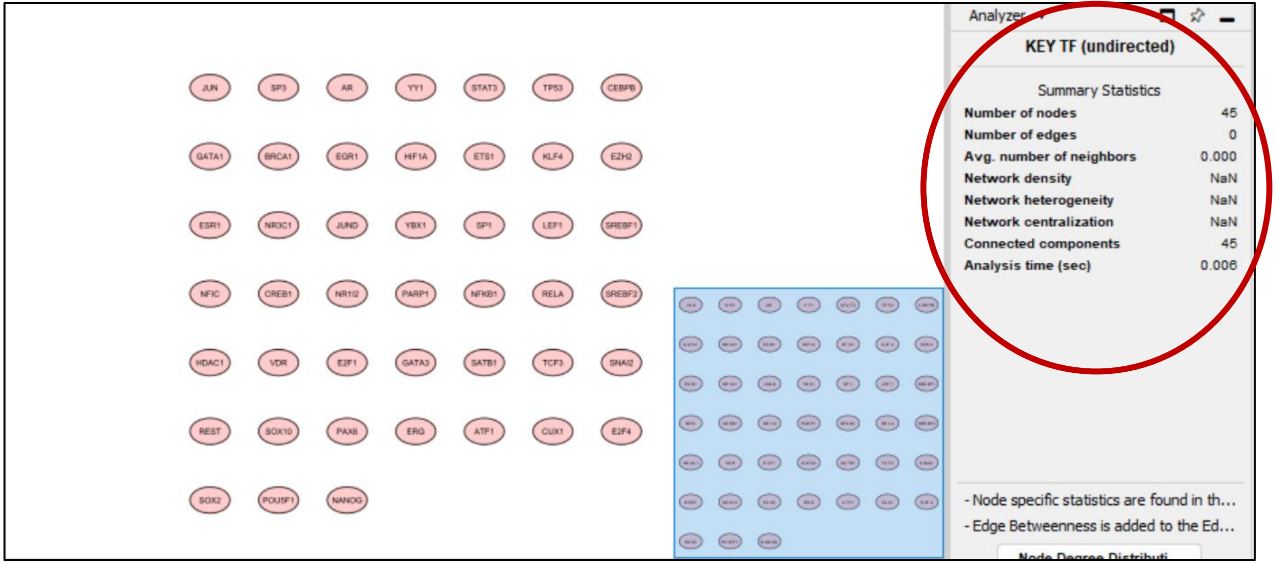
Summary Statistics

Number of nodes	2862
Number of edges	9395
Avg. number of neighbors	5.896
Network diameter	9
Network radius	5
Characteristic path length	3.600
Clustering coefficient	0.123
Network density	0.002
Network heterogeneity	2.953
Network centralization	0.169
Connected components	26
Analysis time (sec)	1.270

- Node specific statistics are found in th...
- Edge Betweenness is added to the Ed...

Figure 31: An image of the complete human gene regulation network along with network analysis (Node coloring is based on Betweenness Centrality)

Then we draw the network related to the connection of key Transcription Factors that we got output from TRRUST using Cytoscape.



KEY TF (undirected)

Summary Statistics

Number of nodes	45
Number of edges	0
Avg. number of neighbors	0.000
Network density	NaN
Network heterogeneity	NaN
Network centralization	NaN
Connected components	45
Analysis time (sec)	0.006

- Node specific statistics are found in th...
- Edge Betweenness is added to the Ed...

Figure 32: Image of the Key Transcription Factors association network along with network analysis

After drawing these two networks, we Merge the complete human gene regulation network with the network of key TFs to examine the more precise relationship between these key regulators with each other. As we saw in the complete regulatory network we observed for humans, there are three types of edges in these regulatory networks:

1. Edges ending in a bow indicate the **activating effect** of the two nodes on each other

2. Edges ending in a straight line indicate **the repressive effect** of the two nodes on each other
3. Edges ending in a semicircle have **unknown** effect/activity

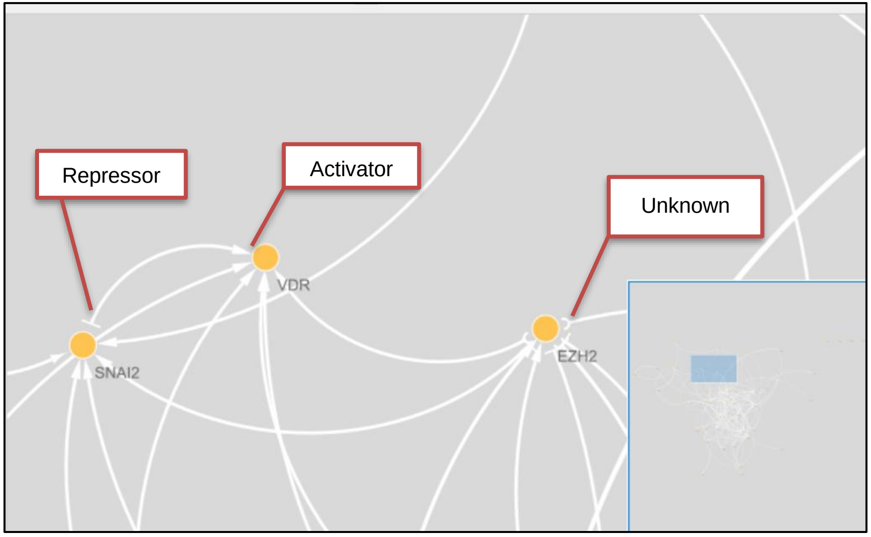
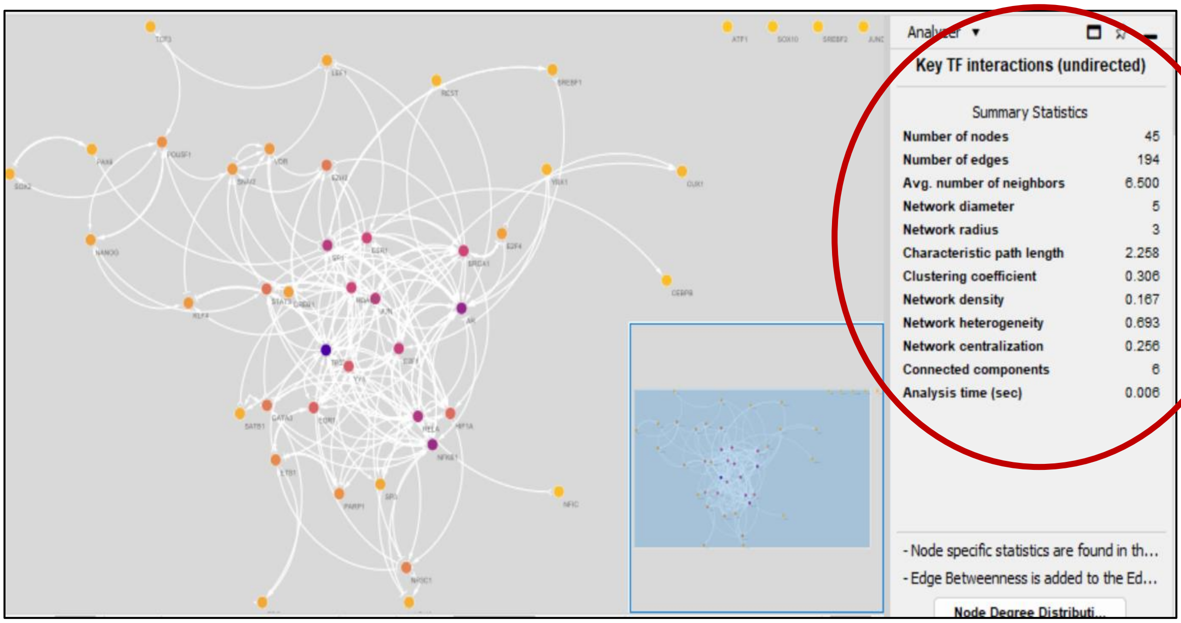


Figure 33: Image of part of the merged network containing all three types of edges.



Key TF interactions (undirected)

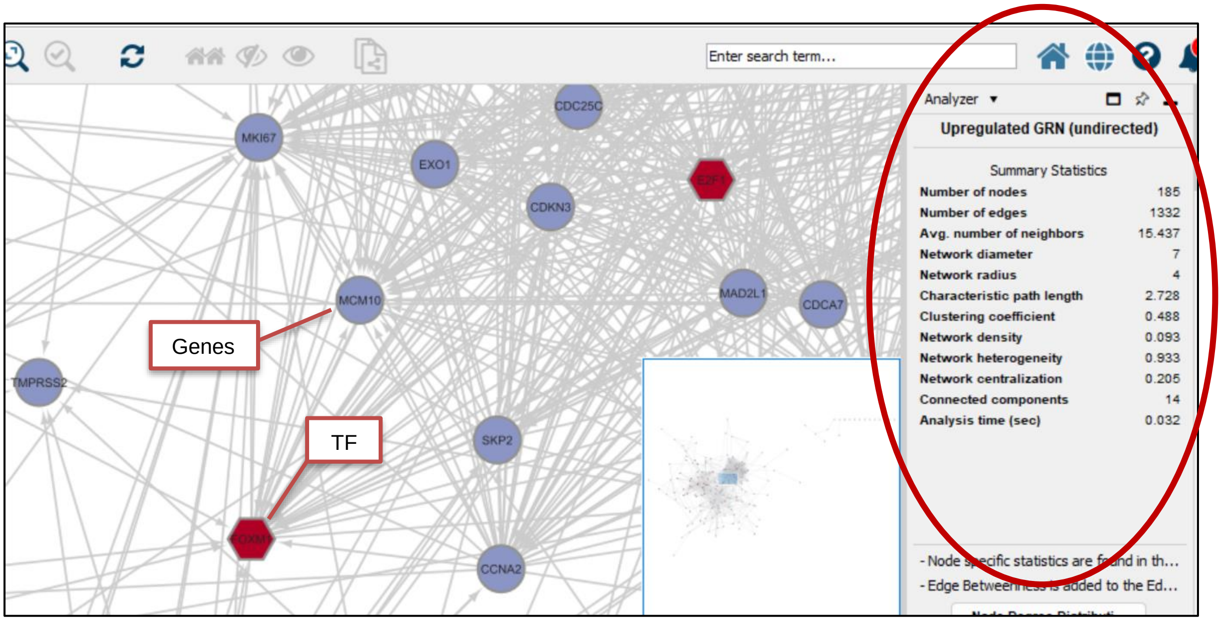
Summary Statistics

Number of nodes	45
Number of edges	194
Avg. number of neighbors	6.500
Network diameter	5
Network radius	3
Characteristic path length	2.258
Clustering coefficient	0.306
Network density	0.167
Network heterogeneity	0.693
Network centralization	0.256
Connected components	6
Analysis time (sec)	0.006

- Node specific statistics are found in th...
- Edge Betweenness is added to the Ed...

Figure 34: Network of TF relationships with each other resulting from merging the human gene regulation network and the list of Key TFs from the TRRUST database. Nodes in this network are colored based on degree, and network analysis is also visible.

In the next step, to draw the GRN of the ppi network, we merge the upregulated genes that we drew earlier with this obtained network to get the relationship between these TFs and their target genes. The noteworthy point is that in this network, not only the relationship between TF and TF on their target genes is depicted, but also the regulatory relationship between the products of these genes has been shown, so that a complete network of gene regulation in connection with this dataset is obtained.



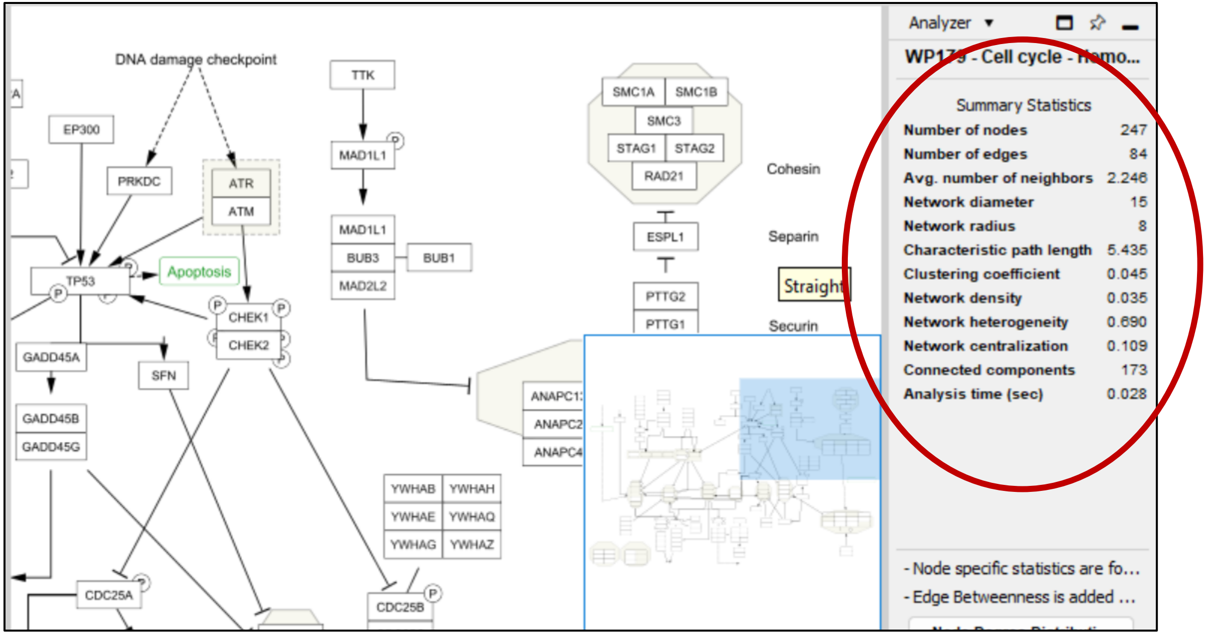
Upregulated GRN (undirected)

Summary Statistics

Number of nodes	185
Number of edges	1332
Avg. number of neighbors	15.437
Network diameter	7
Network radius	4
Characteristic path length	2.728
Clustering coefficient	0.488
Network density	0.093
Network heterogeneity	0.933
Network centralization	0.205
Connected components	14
Analysis time (sec)	0.032

Figure 35: Close-up image of the final gene regulation network obtained from the studied dataset along with network analysis. In this network, TFs are separated from genes with different shapes and colors. The association between genes and TFs is also characterized by three types of ridges.

Another way to understand the regulatory relationship between Transcription Factors and their target genes is to use ready-made networks available on the Cytoscape site itself. For this, we first search E2F1, the first TF in the Key regulators table, with its target genes in Network Search on the Cytoscape site. From among the available networks, we Import the network related to Homo Sapiens Cell Cycle. According to the functions we extracted from the Enrichr database in the previous question, as well as the table related to Key Regulators and the activity of their genes, it can be understood that the majority of genes involved in this dataset control activities related to the cell cycle and cell division. Therefore, a Regulatory Network related to the cell cycle in humans can provide good information about the dataset of interest.



WP179 - Cell cycle - Homo...

Summary Statistics

Number of nodes	247
Number of edges	84
Avg. number of neighbors	2.246
Network diameter	15
Network radius	8
Characteristic path length	5.435
Clustering coefficient	0.045
Network density	0.035
Network heterogeneity	0.690
Network centralization	0.109
Connected components	173
Analysis time (sec)	0.028

Figure 36: Close-up view of the human cell-cycle regulatory network along with network analysis

The complete set of regulatory networks used in this question is available via this link.