HUB GENES PREDICTION FOR CHRONIC OBSTRUCTIVE PULMONARY DISEASE USING MRNA EXPRESSION DATA

A Thesis submitted for
the partial fulfillment of the degree of Master of Engineering in
Computer Science and Engineering, Jadavpur University

by
Pabitra Kumar Mondal
Exam Roll No.- M4CSE22010
Reg. No. - 154134 of 2020-2021

under guidance of
Dr. Anasua Sarkar
Assistant Professor
Department of Computer Science and Engineering
Jadavpur University
Kolkata – 700032, India

Declaration

I, Pabitra Kumar Mondal having Examination Roll Number M4CSE22010 and Registration number 154134 of 2020-2021, do hereby declare that this thesis entitled ”HUB GENES PREDICTION FOR CHRONIC OBSTRUCTIVE PULMONARY DISEASE USING MRNA EXPRESSION DATA” contains literature survey and original research work done by the undersigned candidate as part of Post Graduate studies.

All information in this thesis have been obtained and presented in accordance with existing academic rules and ethical conduct. I declare that, as required by these rules and conduct, I have fully cited and referred all materials and results that are not original to this work.

Signature of Candidate:
Date :

Certificate of Recommendation

This is to certify that Pabitra Kumar Mondal having Examination Roll Number M4CSE22010 and Registration number 154134 of 2020-2021 has completed his thesis entitled "HUB GENES PREDICTION FOR CHRONIC OBSTRUCTIVE PULMONARY DISEASE USING MRNA EXPRESSION DATA", under the supervision and guidance of Prof. Dr. Anasua Sarkar, Jadavpur University, Kolkata. We are satisfied with his work, which is being presented for the partial fulfillment of the degree of Master of Engineering in Computer Science and Engineering, Jadavpur University, Kolkata.

____

Dr. Anasua Sarkar
Supervisor
Department of Computer Science & Engineering
Jadavpur University, Kolkata – 700032

____

Dr. Nandini Mukhopadhyay
Head of Department
Department of Computer Science & Engineering
Jadavpur University, Kolkata – 700032

____

Dr. Chandan Mazumdar
Dean
Faculty Council of Engineering and Technology
Jadavpur University, Kolkata – 700032

Certificate of Approval

The foregoing thesis is hereby approved as a creditable study of an engineering subject, carried out and presented in a manner satisfactory to warrant its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the thesis only for the purpose for which it has been submitted.

Final examination evaluation of
Thesis of

____
(Signature of Candidate)

______________________________ ______________________________
(Signature of the Examiner) (Signature of the Supervisor)

Acknowledgement

First of all, I would like to express my sincere gratitude to my supervisor, Professor Dr. Anasua Sarkar for her valuable guidance and insightful suggestions throughout my thesis work. I am very much thankful to her for giving me the opportunity to pursue my PG thesis work under her guidance. Without her precious knowledge in the field of bioinformatics and exceptional encouragement my thesis work would not have taken a meaningful shape. I will be benefited throughout my life from the knowledge, rightful guidance and unparallel professionalism that I acquired here.

I would like to thank our HOD and the coordinator of our course Dr. Nandini Mukhopadhyay for her contribution throughout the course. I would also like to thank all our professors for their valuable knowledge. Without them this two year journey cannot be so memorable. I would also like to thank my fellow classmates, without whom this journey may not be completed so smoothly.

Last but not the least; I would like to express my gratitude to my friends and family for their constant support throughout my life.

____

Pabitra Kumar Mondal
Exam Roll No.- M4CSE22010
Reg. No. - 154134 of 2020-2021
Master of Engineering
Computer Science & Engineering
Jadavpur University

Abstract

At the moment, chronic obstructive pulmonary disease (COPD) has a world-wide high death rate. And most of the deaths occur in low and middle income countries (LMICs). As the number of COPD patients increases every year, it is considered as a burden to the world. In the absence of a proper cure, we can just slow down the progression with medicines. Therefore researches need to be done in the direction of finding a cure for COPD. We have some positive results in this direction but more studies need to be done.

We collect some mRNA expression data of a mouse to study the genes that are responsible for the disease. Our data has some control samples and some Disease samples. Our study was on the over expressed (up-regulated) genes to identify possible hub genes of COPD.

We perform our task on python and Cytoscape. We cluster the data in some subgroups in python and then find the hub genes with the help of Cytoscape. For functional enrichment analysis like Gene Ontology (GO) and KEGG pathways we use Enrichr and DAVID.

From the clusters we got 80 top ranked genes with the help of cytoHubba plugin in cytoscape by applying cytoHubba on every individual cluster. Then from the outcome we mark **Rac2, Anxa5, Exosc10, Hif1a, Pik3r2, Tia1, Ctnnb1, Jak2, Bdnf, and Pten** as the final 10 possible hub genes.

Abbreviations

COPD - Chronic obstructive pulmonary disease
LMIC – Low and Middle Income Country
DAVID - Database for Annotation, Visualization and Integrated Discovery
GSEA - Gene Set Enrichment Analysis
OA - Osteoarthritis
OSCC - Oral Squamous Cell Carcinoma
GC - Gastric Cancer
DT - Drug Treated
C - Control
D - Disease
DBSCAN - Density Based Spatial Clustering of Application with Noise
PCA - Principal Component Analysis
ICA - Independent Component Analysis
LDA - Linear Discriminant Analysis
LLE - Locally Linear Embedding
MDS - Multidimensional Scaling
t-SNE - t-distributed Stochastic Neighbor Embedding
GO – Gene Ontology
KEGG - Kyoto Encyclopedia of Genes and Genomes
PPI – Protein-Protein Interaction
BP – Biological Process
MF – Molecular Function
CC – Cellular Component

List of Figures

Fig. 1 : Conditions applied to filter out up-regulated genes
Fig. 2 : (a) Difference between K-Means and DBSCAN clusters,
(b) Outliers in DBSCAN
Fig. 3 : Scatter plot using different manifold learning algorithms
(a) Isomap, (b) LLE, (c) Spectral Embedding, (d) MDS, (e) t-SNE
Fig. 4 : DAVID analysis tool window
Fig. 5 : PPI network of filtered up-regulated genes
Fig. 6 : Volcano plot
Fig. 7 : Correlation matrix
Fig. 8 : Heatmap of 50 genes
Fig. 9 : Heatmap with all filtered genes
Fig. 10: TSNEPLOT of the clusters
Fig. 11: Scatter plot after (a) 1 layer of t-SNE and (b) 2 layer of t-SNE
Fig. 12: GO Biological Process
Fig. 13: GO Molecular Functions
Fig. 14: GO Cellular Components
Fig. 15: PPI Network of top 80 genes
Fig. 16: PPI Network of top 80 genes with Degree Sorted Circle Layout
Fig. 17: PPI Network of top 20 genes with Degree Sorted Circle Layout

List of Tables

Table 1: GO terms for individual clusters
Table 2: KEGG pathway list of all genes
Table 3: Cluster 0 KEGG Pathways
Table 4: Cluster 1 KEGG Pathways
Table 5: Possible hub genes for individual clusters

TABLE OF CONTENTS

1. Introduction	11
2. Literature Survey and Existing Work	12-13
3. Methodology	
3.1. Data Collection	13
3.2. Preprocessing	13-14
3.3. Clustering	14-17
3.4. Functional Enrichment Analysis	17-18
3.5. PPI Network Analysis	19-20
4. Result and Discussion	
4.1. Volcano Plot	20
4.2. Correlation Matrix	21
4.3. Heatmap	21-22
4.4. Cluster	23
4.5. Gene Ontology	24
4.6. KEGG Pathways	25-26
4.7. PPI Networks	26-28
5. Conclusion	29
6. References	30-33

1. Introduction

Chronic obstructive pulmonary disease (COPD) is leading third in the table of the causes of death in the world [1]. As reported in 2019 [2], approx 3.23 million people were died in the year due to COPD. More than 90% of these deaths occur in low and middle income countries (LMICs) [3]. India is also one of the LMICs. COPD is also a serious cause of disability. The global burden of disease study for India [4] reported the appearance of COPD increased to 4.4% (55.3 million cases) in 2016 which was 3.3% (28.1 million cases) in 1990.

Currently, there is no cure for COPD [5], but medication is available to control the symptoms and slow the progression of COPD. Therefore, the search for the cure is still on and in the process of drug discovery, it is important to find those particular genes that are responsible in the progression of the disease. In most of the cases, drugs are first tested on mouse and after receiving desired output, it can go for a human trial.

We collect the mRNA expression data of a mouse from a drug (unknown) treatment process. Currently there are two main techniques that can be applied on gene expression data, DNA microarray and RNA-seq. A microarray is a laboratory tool that is used to detect gene expressions on a large scale. The number of genes can be thousands at a time. Microarray data allows determining a sample's gene expression level for many genes at once. In our case, the data gives some insight into which genes may be responsible for a disease and the drug should treat those over expressed genes. At the end of the study, we will get the possible target genes for the drug.

There are multiple microarray platforms [6-9] like Affymetrix, Illumina, Exiqon, Agilent etc. which is used to access microarray data. The data we collect is Agilent data. We processed the data and filter out the over expressed (up-regulated) genes. Then we cluster them in python. Then for Gene Ontology we get help from Enrichr[10-12], an enrichment analysis tool. The gene networks we provide in the study are generated using Cytoscape [13-15].

2. Literature Survey and Existing Work

There are many tools or softwares or websites to analyze the gene expressions data. For example, Cytoscape, it is an open source software for bioinformatics to integrate molecular interaction networks with high-throughput expression data. It is freely available to download as a java application from www.cytoscape.org. Then there is DAVID (Database for Annotation, Visualization and Integrated Discovery) [16-18], which is developed by the Laboratory of Immunopathogenesis and Bioinformatics. It was designed for Gene Set Enrichment analysis (GSEA) [19]. Subio Platform[20] is also a good platform to work with microarray data. ShinyGO [21] can be a good option for gene enrichment analysis. It has a good graphical representation of PPI network, KEGG pathways and Gene Ontology. It's almost an all in one tool. ToppGene [22] can also be helpful for functional enrichment of genes.

RNA sequencing technologies are broadly used in the study of various tumors, and have great significance in the development of new diagnostic and anti-tumor treatment methods.

Zhang et al. [23] analyze the T cell immune receptor of human colorectal cancer by single-cell sequencing technology, and revealed the tumor heterogeneity and gene expression of drug target gene of colorectal cancer T cells. Breast cancer is a diversified disease caused by genetically altered mammary epithelial cells.

The findings by Nguyen et al. [24-26] help to understand the primary source of breast cancer and provide the foundation for enhancing the early cancer detection and prevention of cancer progression.

In the research of acute leukemia, Ley et al. [27] found AML associated genes, and measured 9 mutations in each sample cell of tumor. The outcomes suggest that these diversities caused by genetic mutations may be associated to pathogenesis.

Single-cell sequencing technology can study numerous different ages of nerve cells, and draw a detailed single-cell map to understand and identify different kinds of neurons and their connecting molecules in the brain [28]. Luo et al. [29] differentiated human and mouse frontal cortical neuronal cells subtypes by highthroughput single-cell methylation sequencing.

Fan Liang et al. [30] finds CCNB1, CCNB2, CCNA2, BUB1, and BUB1B as the hub genes for Osteoarthritis (OA).

Guang Li et al. [31] finds some novel biomarkers for Oral Squamous Cell Carcinoma (OSCC).

Shijie Duan et al. [32] finds COL12A1, GSTA3, FGG and FGA as the biomarkers for Gastric Cancer (GC).

In a genome-wide study, Amit Katiyar et al. [33] finds UBC, VCP, ITGA4. HSP90AB1, and VCAM1 as potential biomarkers for multiple myeloma.

J Yang et al. [34] identifies CDKN1A and HDAC1 as the hub genes in COPD.

3. Methodology

3.1. Data Collection

We collect our data from a senior scientist of CSIR – Indian Institute of Chemical Biology. We got three excel files with the data of 8 samples containing signal values in multiple conditions, which was scanned by using Agilent Technologies. We also got some normalized data within the files containing p-values and fold-change values of the genes with expression values of different samples. We mainly work with the normalized data. It contains 10980 genes.

3.2. Preprocessing

Every data file with microarray data contains immense data and to make it understandable the data need to be processed. We processed the data using python. First, we read the excel files using pandas and exclude the comments from the data file. Then we made some changes in column names where required for better understanding. Then we save it as it is for future use with same name as the sheet name (p corr). The code is in makeFile.ipynb file. From now we will read only the p corr file.

We consider only the up-regulated genes to perform our tasks. For that, we need to filter out the up-regulated genes from the source file with 10980 genes. The dataset chosen is between normal (control) and disease samples. We applied few conditions (Fig. 1) to get desired output.

```
df = df[df['GeneSymbol'].notna()]
CvsD_up = df
CvsD_up = CvsD_up[CvsD_up['p ([D] vs [C])'] < 0.05]
CvsD_up = CvsD_up[CvsD_up['Log FC ([C] vs [D])'] > 2]
up_genes = list(CvsD_up.GeneSymbol)
```

Fig. 1: Conditions applied to filter out up-regulated genes

We save the names of up-regulated genes in a list for future use in enrichment analysis. After filtering the up-regulated genes, we save the data with few selective columns in a separate file named CvsD_up_genes.csv. The file includes columns named 'GeneSymbol', 'Description', 'GenbankAccession', 'EntrezGeneID', 'Log FC ([C] vs [D])', 'FC ([C] vs [D])', 'p ([D] vs [C])', 'sample1: gProcessedSignal', 'sample2: gProcessedSignal', 'sample3: gProcessedSignal', 'sample4: gProcessedSignal', 'sample5: gProcessedSignal', 'sample6: gProcessedSignal', 'sample7: gProcessedSignal', 'sample8: gProcessedSignal' from the source file.

Before clustering, we perform volcano plot with log fold-change value and p-value. For that, we use visuz module from bioinfokit package in python. We make a separate dataframe to perform the task.

3.3. Clustering

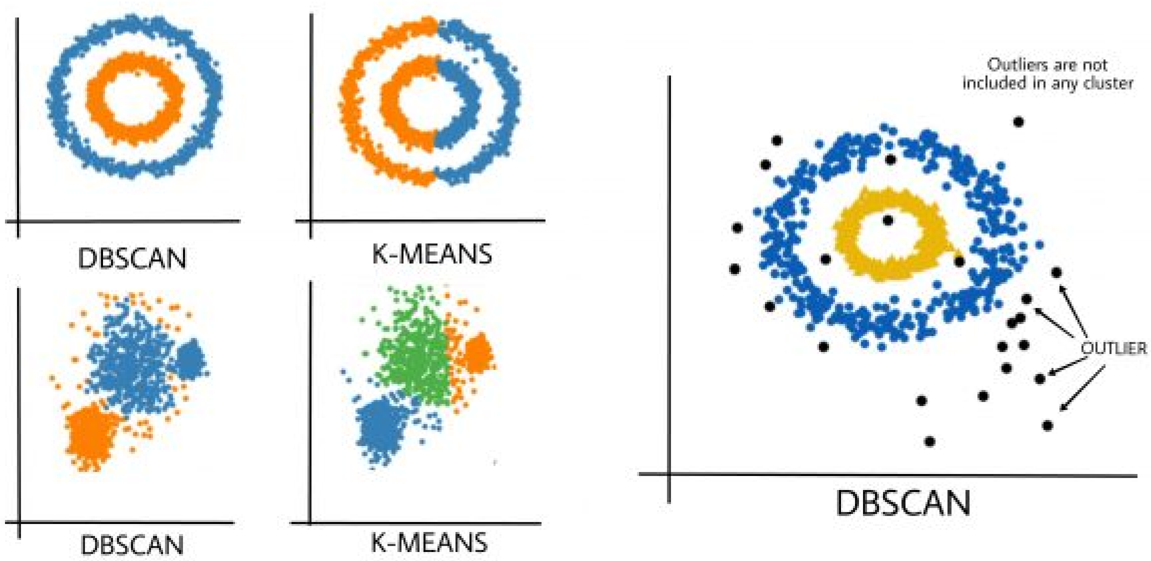
Genes can be clustered by their expression states in order to identify covarying genes.The microarray data we have, has 8 samples. Sample 1, sample 2, and sample 3 are Drug Treated (DT) samples. Sample 4, and sample 5 are Control (C) samples. Sample 6, sample 7, and sample 8 are Disease (D) samples. We do the clustering for C vs D samples. That's why we consider Log FC ([C] vs [D]) and p ([D] vs [C]) to filter out the up-regulated genes. We consider only C and D samples, i.e. 5 samples.

We read the data from CvsD_up_genes.csv file which was prepared before at the preprocessing part. After filtering with the given conditions, we get 2998 up-regulated genes from 10980 genes which is slightly higher in number. So, we decided to filter it again with log fold-change > 3 and p < 0.01 and we get 1454 genes. Then after setting the GeneSymbol column as index column the dimension of the dataframe was 1454x5.

Then again, by using the visuz module of bioinfokit package we construct the heatmap and correlation matrix of the samples.

One of the clustering methods used majorly is K-Means [35]. K-Means forms disjoint groups of data points. K-Means is an unsupervised learning algorithm where K is the number of clusters and it has to be predefined. It is a centroid based algorithm, where every cluster is belongs to a centroid. We want a minimum number of clusters but don't want it to be predifined.

Then we came to know about DBSCAN [36], a Density Based Spatial Clustering of Application with Noise. It finds some high density points and expands the cluster around the points. Also the numbers of clusters are not predefined. DBSCAN is not influenced by outliers (Fig. 2(b)). Though it is not faster than K-Means, but it fulfills our requirement. Some differences are shown in the Fig. 2(a).



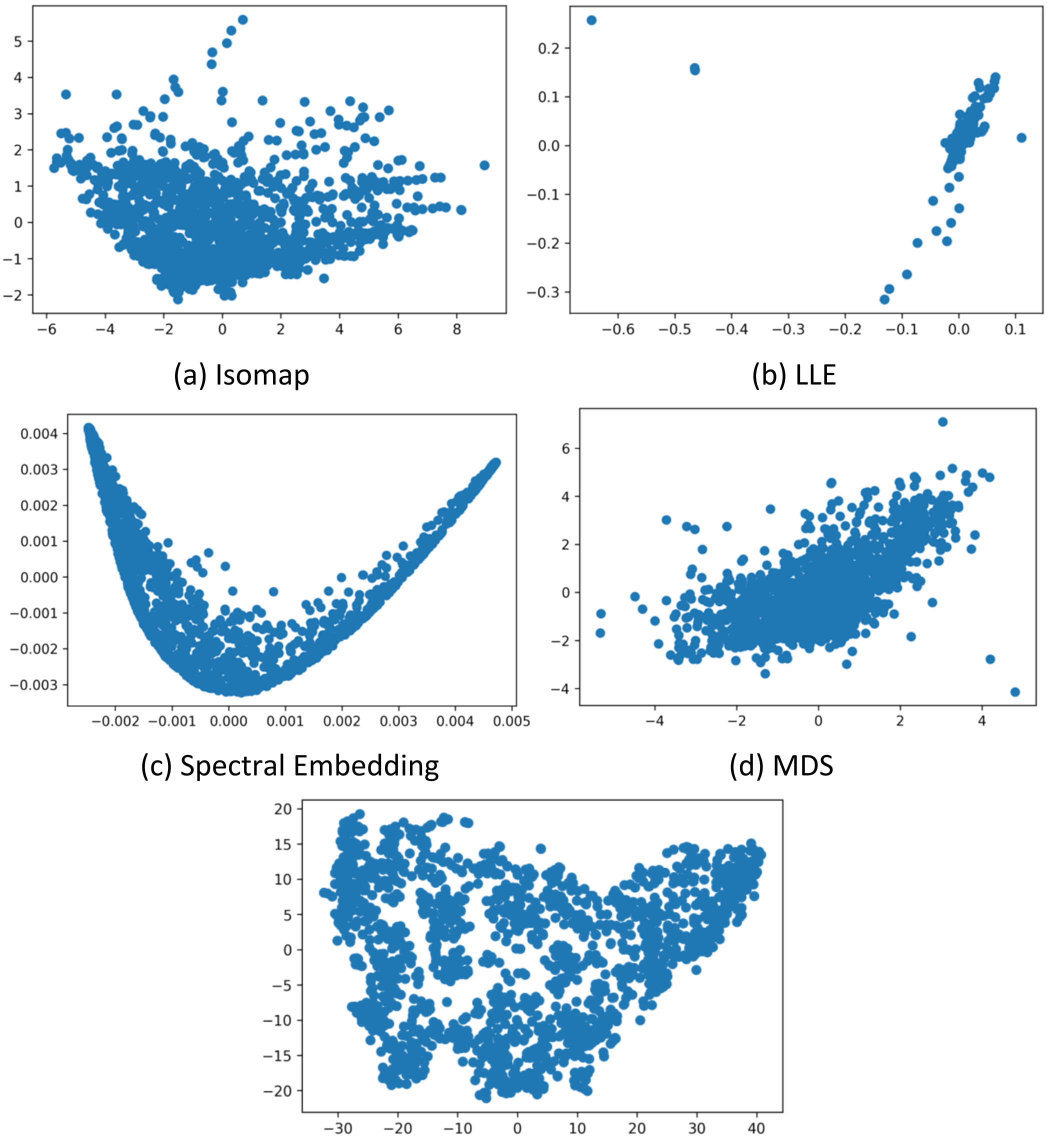
(a) (b)

Fig. 2 : (a) Difference between K-Means and DBSCAN clusters,
(b) Outliers in DBSCAN [Source : geekforgeeks]

Now, to cluster our data points we use DBSCAN algorithm with eps = 3 and min_samples = 22. For that, we get help from scikit-learn package. eps is the most important parameter to choose. The optimization of eps is essential for well defined clusters. After multiple run with different values, it is seen that 3 is a good choice for eps.

As mention earlier, the dimension of our data was 1454x5. And high dimensional data are very hard to visualize. To visualize our data on a 2D plane we need to apply a dimensionality reduction algorithm. There are several frameworks designed to reduce the dimensionality like Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Linear Discriminant Analysis (LDA). These methods are strong enough but sometimes miss important non-linear structure in data.

Here comes Manifold Learning [37], it is an attempt to generalize linear framework like PCA to be sensitive to non-linear structure in data. There are multiple implementations of manifold learning algorithm such as Isomap algorithm [38], Locally Linear Embedding (LLE) [39], Spectral Embedding [40], Multidimensional Scaling (MDS) [41], and t-distributed Stochastic Neighbor Embedding (t-SNE) [42-43].



(a) Isomap

(b) LLE

(c) Spectral Embedding

(d) MDS

(e) TSNE

Fig. 3 : Scatter plot using different manifold learning algorithms

We plot the data points after perform dimensionality reduction by manifold learning using different algorithms (Fig. 3) and see that in Fig. 3(b) data points are not well distributed. And in Fig. 3(a), Fig. 3(c) and Fig. 3(d) we cannot see well distinguish patches that we can say a possible cluster. So we decided to use t-SNE manifold learning to reduce the dimensionality. The main purpose of t-SNE is to visualize the high dimensional data. Hence, it works best when the data is embedded in 2D or 3D. The algorithm is available under scikit-learn as a function. There are five parameters which control the optimization of t-SNE.

- perplexity
- early exaggeration factor
- learning rate
- number of iteration
- angle

We use two layer of t-SNE for better visualization. First time with perplexity = 60 and other parameters with their default values. And in second time we reduce the perplexity to 50. There is a problem with t-SNE that it cannot preserved its global structure, but this can be overcome by initializing it with PCA using init = 'pca' as parameter.

After applying t-SNE to our data, we pass the t-SNE output to DBSCAN as input. And the clustering was performed on the t-SNE data. We got nine clusters with some outliers and plot them using tsneplot() with colours. tsneplot() can be found under cluster module of bioinfokit.visuz package. We add three new data columns in our data to store the t-SNE values and the cluster group. We save the data in a separate file named clusters.csv.

3.4. Functional Enrichment Analysis

We use Enrichr (https://maayanlab.cloud/Enrichr/) for Gene Ontology (GO) analysis and DAVID bioinformatics database (https://david.ncifcrf.gov/tools.jsp) for Kyoto En cyclopedia of Genes and Genomes (KEGG) pathway analysis. For Enrichr, we create a link (https://amp.pharm.mssm.edu/Enrichr/enrich?dataset=94a0604324df717cfc4cb941ba42ae75) with th e help of json and request packages containing the gene list we prepared in the prep rocessing part. We also do the same for every cluster to identify their functions.

Next for the KEGG pathways, we directly paste the GenBank_Accession of those 1454 genes in the textbox of the DAVID analysis tool. Then after selecting the identifier as GENBANK_ACCESSION and the list type as Gene List, we need to hit the submit button. And we will get the window (Fig. 4) where we need to select for Functional Annotation Tool.

Fig. 4 : DAVID analysis tool window

After selecting Functional Annotation Tool, we will get the below options.

- Disease
- Functional Annotations
- Gene Ontology
- General Annotations
- Interactions
- Literature
- Pathways
- Protein Domains
- Tissue Expression

From the KEGG_PATHWAY option listed in Pathways we get a list of pathways.

3.5. PPI Network Analysis

We use Cytoscape for protein-protein interaction (PPI) Network analysis. In Cytoscape, we first install some apps like stringApp, Largest Subnetwork, Legend Creator, and cytoHubba which we used for the analysis.

- stringApp - it is used to retrieve PPI networks from the STRING database.
- Largest Subnetwork - it is used to select the largest subnetwork in the current network.
- Legend Creator - it is used to create legends in a better way for a graph.
- cytoHubba - it is used to predict important nodes and subnetwork in a network using several topological algorithms.

At first we paste the GeneSymbols of 1454 up-regulated genes in the network search bar after selecting STRING protein query from the drop-down. Then in the options panel we change the species from Homo sapiens to Mus musculus and keep other options as it is. Then click on the search icon. And got the below result (Fig. 5).

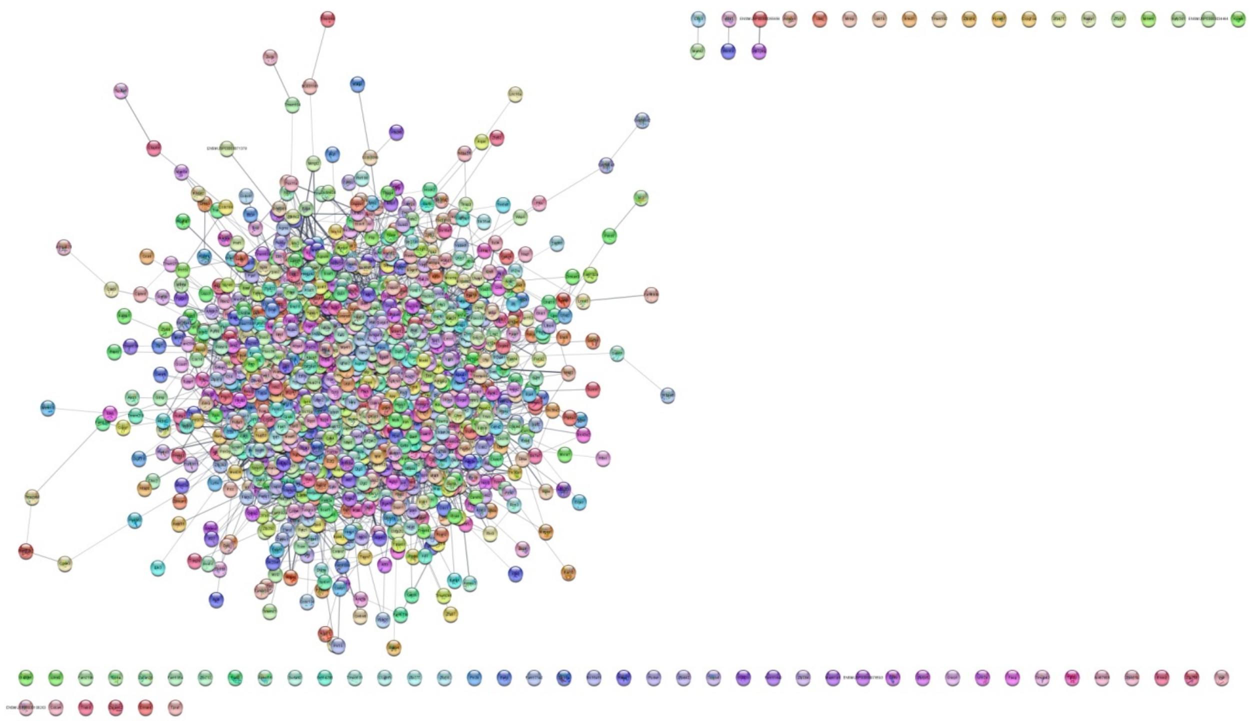


Fig. 5 : PPI network of filtered up-regulated genes

Now here comes the role of Largest Subnetwork App. We select the largest subnetwork from the whole network. Then we import the cluster.csv file into the node table to construct separate subnetwork for every cluster using the filter option. In the filter option we add a new column filter using the cluster column and give the values 0 in both places. And thus we select the nodes from cluster 0. Then we create a new network from the selected nodes. We do the same for all the clusters.

Now after constructing the networks, we run cytoHubba to predict the ranks and create a hubba table with the scores for every algorithm. We consider the Degree algorithm to rank top 10 genes from the network. We repeat the task for every cluster. Thus we get top 10 genes from every cluster.

4. Result and Discussion

4.1. Volcano Plot

The volcano plot (Fig. 6) is constructed by using visuz.GeneExpression.volcano() function. The function accepts a dataframe having log fold-change value and p-value. The visuz module is present in the bioinfokit package. The green dots are indicating up-regulated genes and the red dots are indicating down-regulated genes from the datasets. We use threshold 3 and –2 for up and down regulated genes (lfc_thr=(3,2)). And for p-value we use the threshold of 0.01 (pv_thr=(0.01,0.01)).

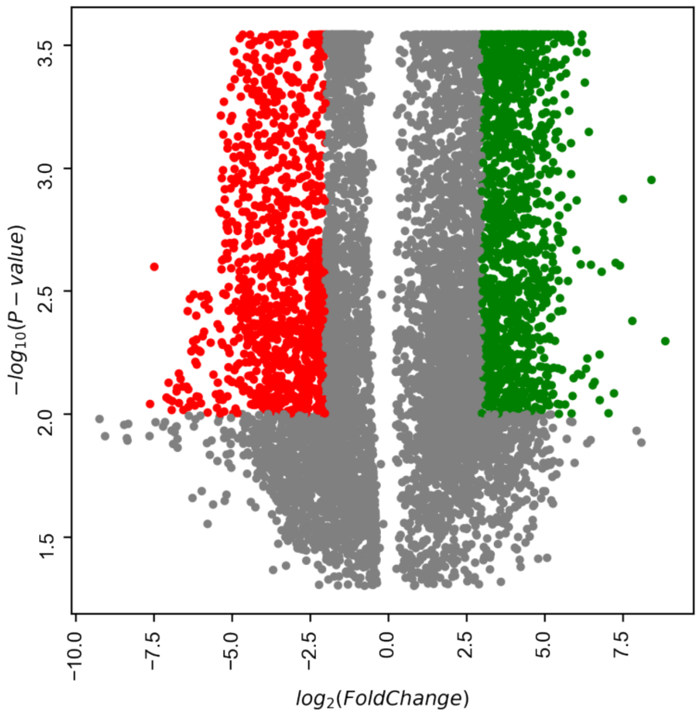


Fig. 6 : Volcano plot

In the above figure, we see that the green dots are well scattered. It indicates how differentially they expressed.

4.2. Correlation Matrix

Then we plot the correlation matrix (Fig. 7) of the expression values of different samples using visuz.stat.corr_mat() function. Sample 4 and 5 are the C samples and sample 6, 7, and 8 are D samples.

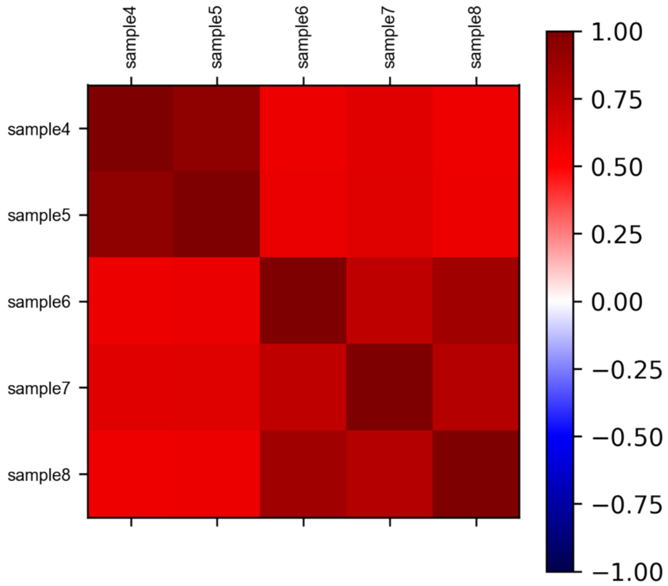


Fig. 7 : Correlation matrix

We can see that C samples and D samples are positively related to each other. Sample 6 and sample 8 are highly related to each other compared to their relation with sample 7. Same can be seen in the upper tree of the heatmap of the samples.

4.3. Heatmap

This heatmap is generated by using visuz.GeneExpression.hmap() function from visuz module. The relationships between D samples can also be noticed here. We generate the heatmap (Fig. 8) only for 50 genes to demonstrate the changes and at the same time to show the names of the genes. We can also see the hierarchical clustering in the left side of the heatmap. But it is not for the complete gene list. The complete heatmap (Fig. 9) with all the genes are also provided below (The names of the genes may not so clear for the resolution and space).

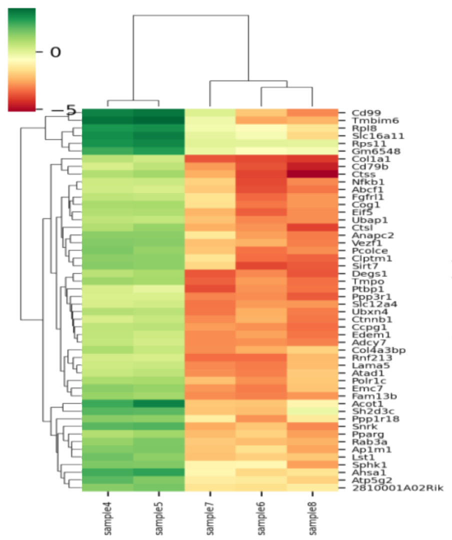


Fig. 8 : Heatmap of 50 genes

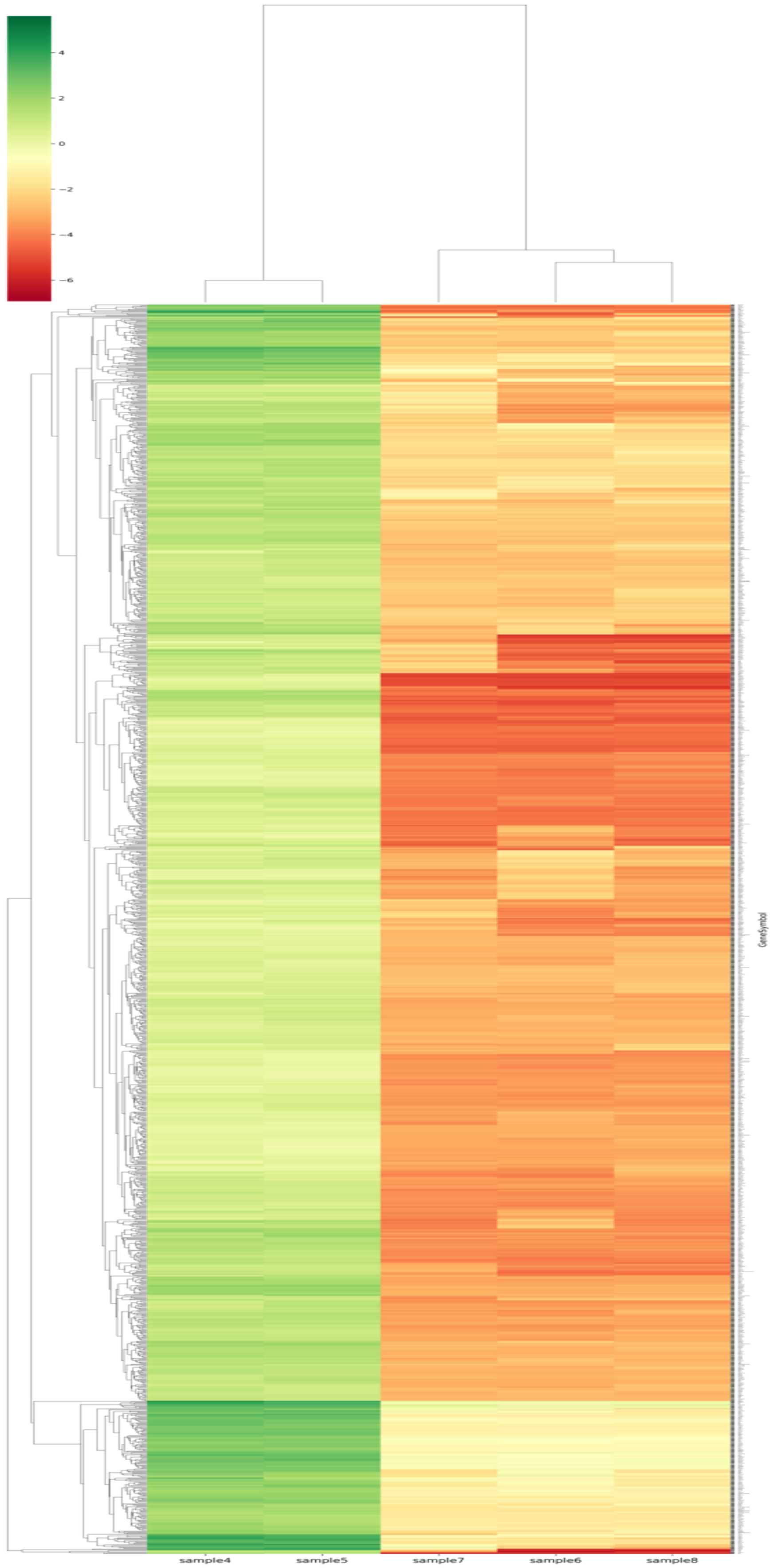


Fig. 9 : Heatmap with all filtered genes

4.4. Cluster

After performing DBSCAN on the data which we get after 2 layers of tsne embedding on the expression values of the genes, we get the below output (Fig. 10). There are nine clusters.

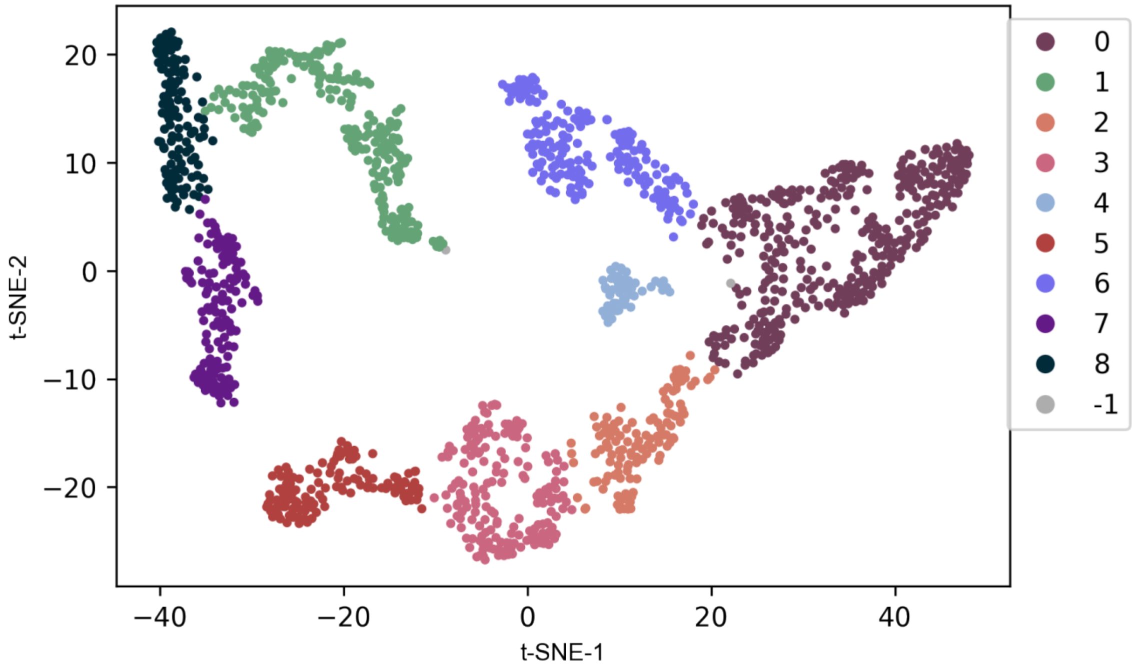
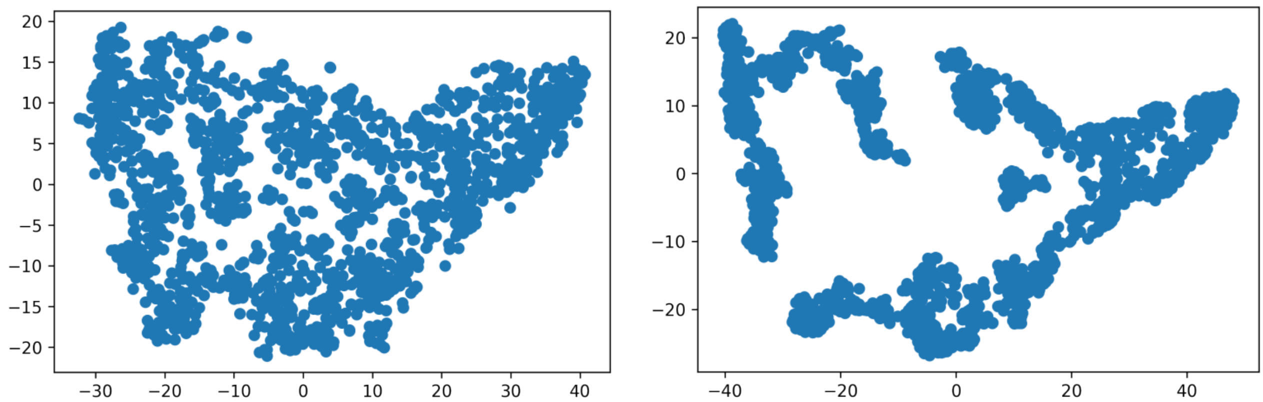


Fig. 10 : TSNEPLOT of the clusters

The difference we get after applying 2 layers of tsne embedding can easily be compared by the scatter plot (Fig. 11) of the tsne values while applying the layers.



(a) (b)

Fig. 11 : Scatter plot after (a) 1 layer of t-SNE and (b) 2 layer of t-SNE

Next we perform functional enrichment analysis on every individual cluster using Enrichr and find the KEGG pathway using DAVID.

4.5. Gene Ontology

As mentioned earlier, we use Enrichr analysis tool for Gene Ontology. And we get the following results for the complete up-regulated gene list.

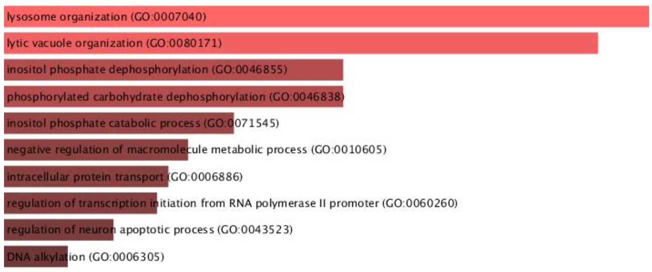


Fig. 12 : GO Biological Process

In the GO Biological Process (BP) enrichment analysis (Fig. 12), we got two main biological process as GO:0007040 and GO:0080171.

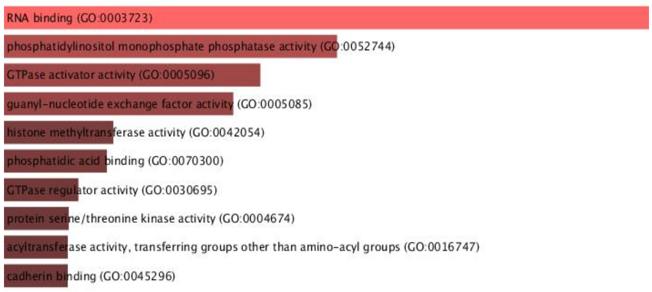


Fig. 13 : GO Molecular Function

In the GO Molecular Function (MF) enrichment analysis (Fig. 13), we got RNA binding (GO:0003723) as the main function.

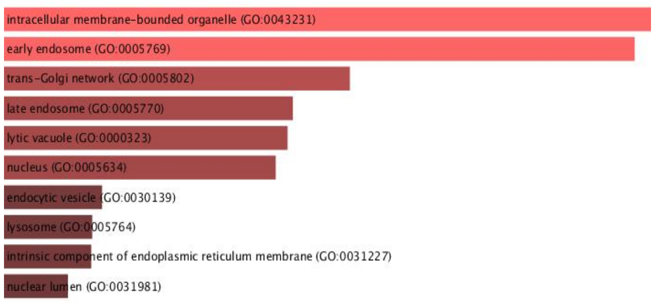


Fig. 14 : GO Cellular Component

In the GO Cellular Component (CC) enrichment analysis (Fig. 14), we got two main component, intracellular membrane-bounded organelle (GO:0043231) and early endosome (GO:0005769).

We do the same analysis for individual cluster to identify the functions of each cluster and made the following table (Table 1) with GO terms.

Cluster	GO terms
Cluster 0	GO:0016236, GO:0032511, GO:0005769, GO:0004364
Cluster 1	GO:0097067, GO:0097066, GO:0000323, GO:0046966
Cluster 2	GO:0048681, GO:0000083, GO:0070571, GO:0005658, GO:0031415, GO:0042583, GO:0045252, GO:0003899, GO:0030695, GO:0035198, GO:0030291
Cluster 3	GO:0010256, GO:0034470, GO:0031981, GO:0005666, GO:0005730, GO:0070300
Cluster 4	GO:0010458, GO:0005680, GO:0016836
Cluster 5	GO:1903749, GO:0005634, GO:0043175
Cluster 6	GO:0018146, GO:0090161, GO:0042339, GO:0000139, GO:0045296, GO:0005384, GO:1990050
Cluster 7	GO:1901566, GO:1904705, GO:0031306, GO:0004674, GO:0046915
Cluster 8	GO:1905168, GO:2000779, GO:0071782, GO:0032541, GO:0005634, GO:0043231, GO:0043130

Table 1 : GO terms for individual clusters

4.6. KEGG Pathways

When we analyze the genes for KEGG pathways in DAVID we find Metabolic pathways (mmu01100), Pathways of neurodegeneration - multiple diseases (mmu05022), Pathways in cancer (mmu05200), Amyotrophic lateral sclerosis (mmu05014), Human T-cell leukemia virus 1 infection (mmu05166), MAPK signaling pathway (mmu04010) as the important KEGG pathways (Table 2). Metabolic pathways (mmu01100) has the most number of genes from the gene list.

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	KEGG_PATHWAY	Metabolic pathways	RT		132	9.4	1.2E-2	3.6E-1
	KEGG_PATHWAY	Pathways of neurodegeneration - multiple diseases	RT		47	3.4	6.9E-3	3.6E-1
	KEGG_PATHWAY	Pathways in cancer	RT		47	3.4	6.4E-2	4.7E-1
	KEGG_PATHWAY	Amyotrophic lateral sclerosis	RT		37	2.6	1.6E-2	3.6E-1
	KEGG_PATHWAY	Human T-cell leukemia virus 1 infection	RT		29	2.1	5.6E-3	3.6E-1
	KEGG_PATHWAY	MAPK signaling pathway	RT		29	2.1	4.1E-2	4.2E-1
	KEGG_PATHWAY	Endocytosis	RT		28	2.0	2.8E-2	3.9E-1
	KEGG_PATHWAY	Human cytomegalovirus infection	RT		26	1.9	4.0E-2	4.2E-1
	KEGG_PATHWAY	Kaposi sarcoma-associated herpesvirus infection	RT		24	1.7	2.9E-2	3.9E-1
	KEGG_PATHWAY	Protein processing in endoplasmic reticulum	RT		21	1.5	1.2E-2	3.6E-1
	KEGG_PATHWAY	Hepatocellular carcinoma	RT		21	1.5	1.4E-2	3.6E-1
	KEGG_PATHWAY	Focal adhesion	RT		21	1.5	5.2E-2	4.5E-1
	KEGG_PATHWAY	Lysosome	RT		20	1.4	1.8E-3	3.6E-1
	KEGG_PATHWAY	Cellular senescence	RT		20	1.4	4.2E-2	4.2E-1
	KEGG_PATHWAY	Tuberculosis	RT		19	1.4	6.1E-2	4.7E-1

Table 2 : KEGG pathway list of all genes

As we do for Gene Ontology, we perform the same for KEGG pathways. We analyze it for every individual cluster.

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	KEGG_PATHWAY	Metabolic pathways	RT		38	11.2	3.2E-2	1.0E0
	KEGG_PATHWAY	Pathways of neurodegeneration - multiple diseases	RT		19	5.6	1.0E-3	1.1E-1
	KEGG_PATHWAY	Amyotrophic lateral sclerosis	RT		14	4.1	1.0E-2	5.0E-1
	KEGG_PATHWAY	Huntington disease	RT		13	3.8	5.3E-3	3.3E-1
	KEGG_PATHWAY	Alzheimer disease	RT		12	3.5	6.2E-2	1.0E0
	KEGG_PATHWAY	Endocytosis	RT		10	2.9	4.2E-2	1.0E0
	KEGG_PATHWAY	Glutathione metabolism	RT		7	2.1	1.4E-3	1.1E-1
	KEGG_PATHWAY	Spinocerebellar ataxia	RT		6	1.8	9.2E-2	1.0E0
	KEGG_PATHWAY	Fatty acid elongation	RT		5	1.5	1.4E-3	1.1E-1
	KEGG_PATHWAY	Ovarian steroidogenesis	RT		5	1.5	2.2E-2	7.9E-1
	KEGG_PATHWAY	Platinum drug resistance	RT		5	1.5	4.7E-2	1.0E0
	KEGG_PATHWAY	Biosynthesis of unsaturated fatty acids	RT		4	1.2	2.0E-2	7.9E-1
	KEGG_PATHWAY	Vasopressin-regulated water reabsorption	RT		4	1.2	3.8E-2	1.0E0
	KEGG_PATHWAY	Fatty acid metabolism	RT		4	1.2	8.8E-2	1.0E0

Table 3 : Cluster 0 KEGG pathways

From the above table (Table 3) we see that Metabolic pathways is the main pathways for cluster 0.

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
☐	KEGG_PATHWAY	Pathways in cancer	RT		11	5.1	6.3E-2	8.6E-1
☐	KEGG_PATHWAY	Lysosome	RT		9	4.2	9.2E-5	2.0E-2
☐	KEGG_PATHWAY	Apoptosis	RT		8	3.7	6.0E-4	6.7E-2
☐	KEGG_PATHWAY	cAMP signaling pathway	RT		7	3.3	3.0E-2	8.4E-1
☐	KEGG_PATHWAY	Thyroid hormone signaling pathway	RT		6	2.8	9.1E-3	6.0E-1
☐	KEGG_PATHWAY	Hippo signaling pathway	RT		6	2.8	2.6E-2	8.4E-1
☐	KEGG_PATHWAY	Proteoglycans in cancer	RT		6	2.8	6.8E-2	8.6E-1
☐	KEGG_PATHWAY	Neutrophil extracellular trap formation	RT		6	2.8	7.0E-2	8.6E-1
☐	KEGG_PATHWAY	B cell receptor signaling pathway	RT		5	2.3	1.1E-2	6.0E-1
☐	KEGG_PATHWAY	HIF-1 signaling pathway	RT		5	2.3	3.3E-2	8.4E-1
☐	KEGG_PATHWAY	Natural killer cell mediated cytotoxicity	RT		5	2.3	3.4E-2	8.4E-1
☐	KEGG_PATHWAY	AMPK signaling pathway	RT		5	2.3	4.6E-2	8.6E-1
☐	KEGG_PATHWAY	Yersinia infection	RT		5	2.3	5.5E-2	8.6E-1
☐	KEGG_PATHWAY	Apelin signaling pathway	RT		5	2.3	5.9E-2	8.6E-1
☐	KEGG_PATHWAY	Insulin signaling pathway	RT		5	2.3	6.1E-2	8.6E-1
☐	KEGG_PATHWAY	Autophagy - animal	RT		5	2.3	6.5E-2	8.6E-1
☐	KEGG_PATHWAY	Fc epsilon RI signaling pathway	RT		4	1.9	3.3E-2	8.4E-1
☐	KEGG_PATHWAY	Colorectal cancer	RT		4	1.9	6.8E-2	8.6E-1

Table 4 : Cluster 1 KEGG pathways

For cluster 1 we find *Pathways in cancer* as the main pathways (Table 4). Similarly for cluster 2 we get a single pathway as *Carbon metabolism*, for cluster 3 we get *Aminoacyl-tRNA biosynthesis*. For cluster 4 and 6 we get *Metabolic pathways* again. For cluster 5 we get *Human T-cell leukemia virus 1 infection*. For cluster 6 we get *Metabolic pathways* and *Endocytosis*. For cluster 7 we get *Cytokine-cytokine receptor interaction*. And for cluster 8 we get *MAPK signaling pathway*.

4.7. PPI Networks

We construct the PPI networks in cytoscape. We consider the largest network for every analysis we do. By applying cytoHubba we get top 10 genes from every cluster. While constructing network for Cluster 4, we didn't get much edges and that can affect the cytoHubba score. So we discard cluster 4 while picking up possible hub genes. And the PPI network for those top 80 genes is given below in Fig. 15.

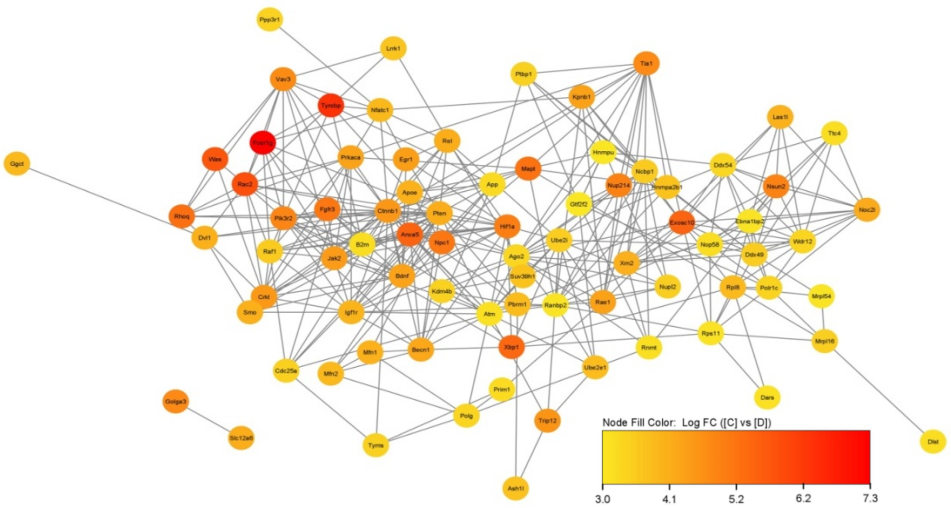


Figure 15 : PPI Network of top 80 genes

We mention those possible hub genes in the below table with their degree score for their respective clusters.

Cluster	Possible Hub genes with degree
Cluster 0	App(18), Rps11(12), Rpl8(12), B2m(12), Igf1r(11), Anxa5(11), Apoe(11), Xbp1(10), Becn1(10), Ube2i(9)
Cluster 1	Rac2(19), Hnrnpa2b1(11), Tyrobp(10), Pik3r2(10), Fcer1g (10), Rhoq(9), Mapt(9), Hif1a(8), Npc1(8), Vav3(8)
Cluster 2	Polr1c(4), Mrpl16(4), Rnmt(3), Tyms(3), Dlst(3), Polg(3), Gtf2f2(3), Prim1(2), Ebna1bp2(2), Mrpl54(1)
Cluster 3	Ranbp2(7), Ddx54(7), Lrrk1(6), Wdr12(6), Atm(6), Nop58(6), Ddx49(5), Kdm4b(5), Dars(5), Pbrm1(5)
Cluster 5	Ncbp1(8), Suv39h1(8), Ago2(6), Cdc25a(5), Ttc4(4), Hnrnpu(4), Nupl2(4), Ppp3r1(4), Ptbp1(4), Ash1l(4)
Cluster 6	Ctnnb1(13), Bdnf(7), Xrn2(6), Jak2(5), Nsun2(5), Noc2l(5), Golga3(4), Las1l(4), Fgfr3(4), Slc12a6(3)
Cluster 7	Pten(10), Raf1(4), Smo(4), Crkl(4), Dvl1(4), Mfn2(4), Prkaca(3), Ube2e1(3), Mfn1(3), Ggct(3)
Cluster 8	Rae1(7), Egr1(7), Exosc10(6), Trip12(6), Was(6), Tia1(6), Nfatc1(6), Kpnb1(5), Nup214(5), Rel(5)

Table 5 : Possible hub genes for individual clusters

From the above 80 genes we construct the PPI network with the Degree Sorted Circle Layout (Fig. 16). And the colour mapping for the border paint was done using the log fold-change column.

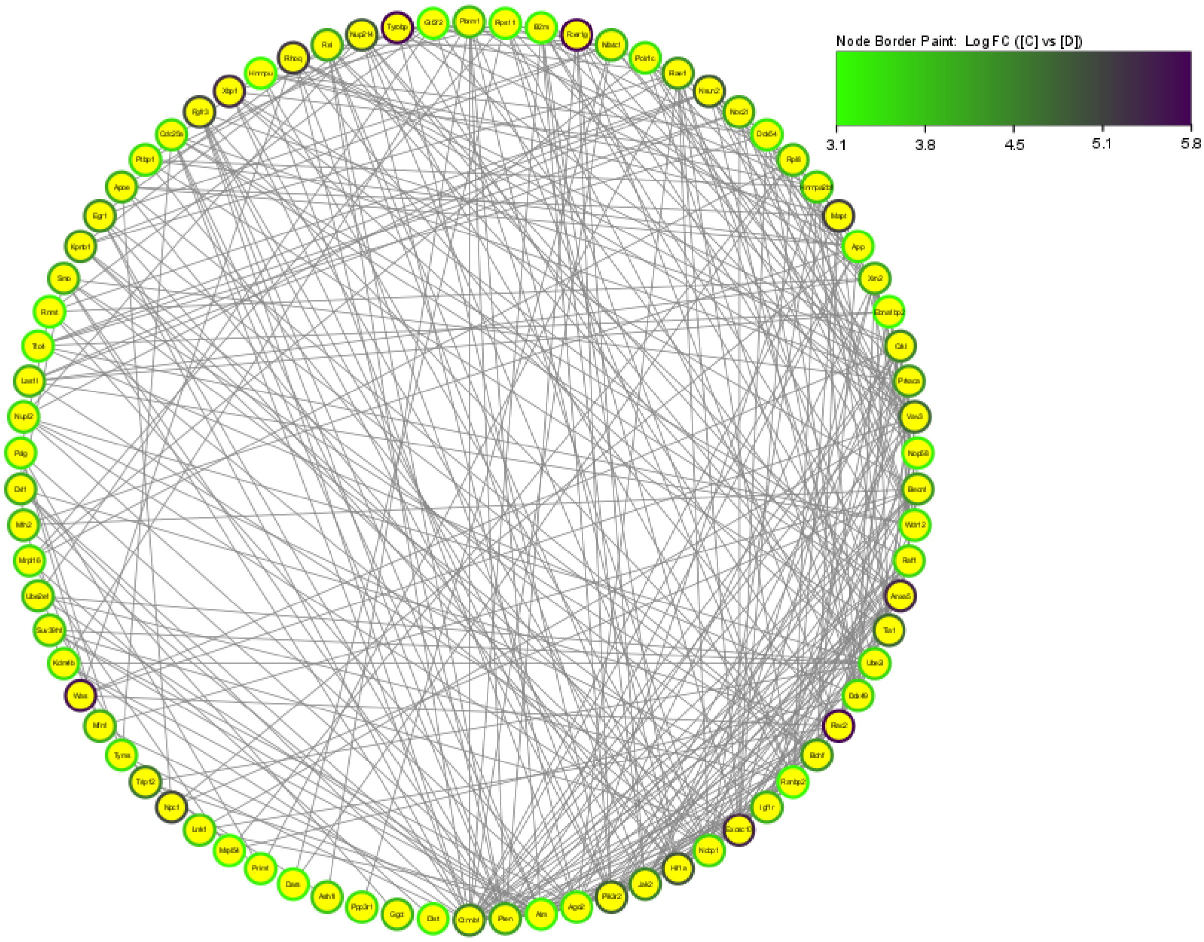


Fig. 16 : PPI Network of top 80 genes with Degree Sorted Circle Layout

Now on those 80 genes we again apply cytoHubba and construct PPI network for top 20 genes (Fig. 17).

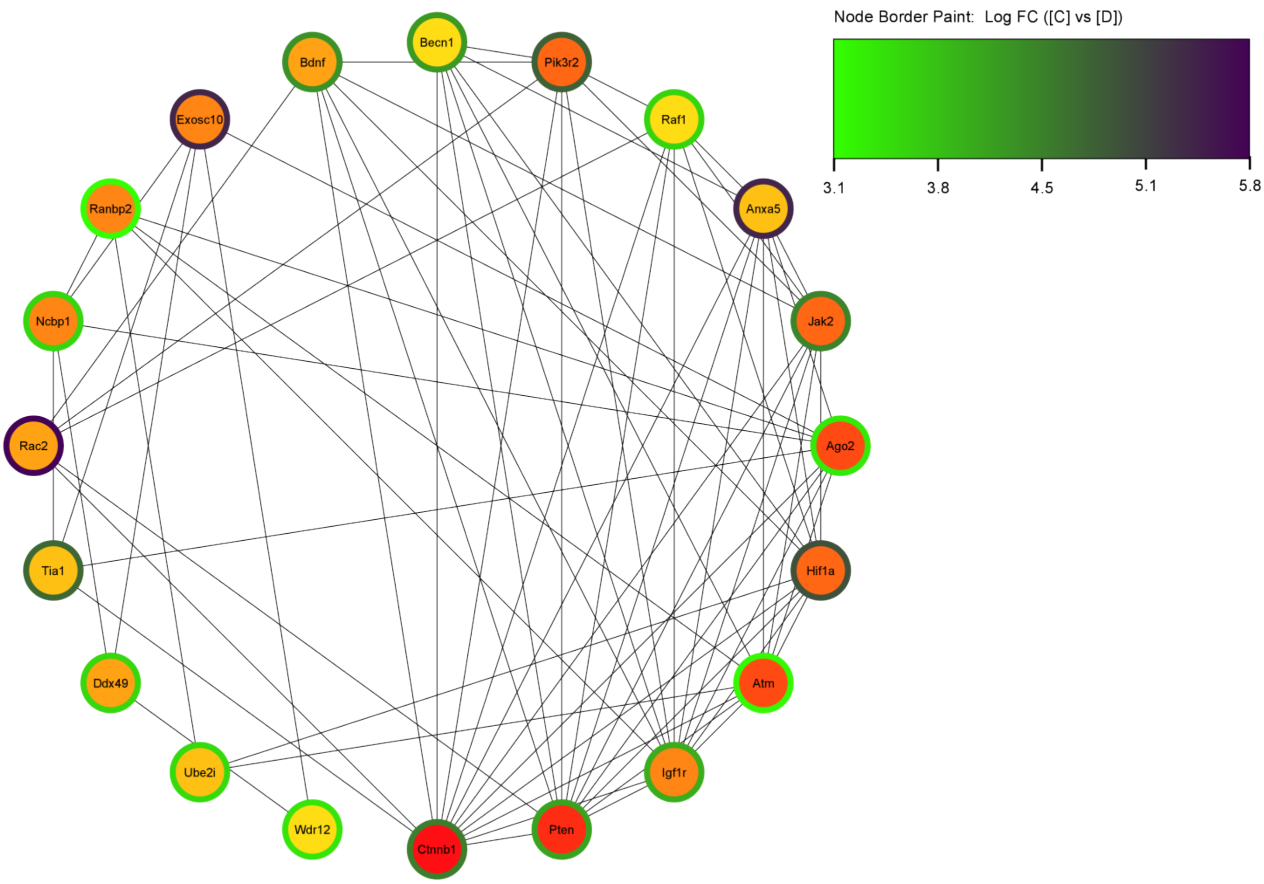


Fig. 17 : PPI Network of top 20 genes with Degree Sorted Circle Layout

Now from those top 20 genes we select top 10 genes based on the log fold-change value. The border colour of the nodes in the network depicts the value of log fold-change. We use a continuous colour mapping from green to purple. So we select nodes with dark border. We get Rac2, Anxa5, Exosc10, Hif1a, Pik3r2, Tia1, Ctnnb1, Jak2, Bdnf, and Pten as the final 10 possible hub genes.

5. Conclusion

Based on our analysis we mark Rac2, Anxa5, Exosc10, Hif1a, Pik3r2, Tia1, Ctnnb1, Jak2, Bdnf, and Pten as the final 10 hub genes (biomarker) for COPD. Ctnnb1 was the best performer based on degree score but the log fold-change value was greater in case of Rac2, Anxa5, Exosc10, Hif1a, Pik3r2, and Tia1. We perform the cluster using t-SNE and DBSCAN. We found a single use of t-SNE was not performing well for our datasets, so we decide to perform 2 layer of t-SNE. There is always having chance for improvement. Some better algorithm may give better result. Further study can be done to find the changes on the top genes while a drug is being applied on them. These genes can be the target genes for the drug.

6. References

[1] P Barnes, P Burney, E Silverman, et al., "Chronic obstructive pulmonary disease", Nat Rev Dis Primers 1, 15076 (2015). https://doi.org/10.1038/nrdp.2015.76

[2] World Health Organization, "Chronic obstructive pulmonary disease (COPD)", (20 May, 2022), available at https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd)

[3] World Health Global burden of disease study 2017. Lancet, 2017, 1–7.

[4] Salvi S, Kumar GA, Dhaliwal RS, et al.. The burden of chronic respiratory diseases and their heterogeneity across the states of India: the Global Burden of Disease Study 1990–2016. Lancet Glob Health 2018; 6(12):e1363–74.

[5] The National Health Service, Treatment -Chronic obstructive pulmonary disease (COPD), 2019, available at https://www.nhs.uk/conditions/chronic-obstructive-pulmonary-disease-copd/treatment

[6] D Petersen, GVR Chandramouli, J Geoghegan et al., "Three microarray platforms: an analysis of their concordance in profiling gene expression", BMC Genomics **6,** 63, 2005. https://doi.org/10.1186/1471-2164-6-63

[7] M Callari, M Dugo, V Musella, E Marchesi, G Chiorino et al., "Comparison of Microarray Platforms for Measuring Differential MicroRNA Expression in Paired Normal/Cancer Colon Tissues", PLoS ONE 7(9): e45105, 2012. doi:10.1371/journal.pone.0045105

[8] S Maouche, O Poirier, T Godefroy et al., "Performance comparison of two microarray platforms to assess differential gene expression in human monocyte and macrophage cells", BMC Genomics **9,** 302, 2008. https://doi.org/10.1186/1471-2164-9-302

[9] Carole L. Yauk, M. Lynn Berndt, Andrew Williams, and George R. Douglas, "Comprehensive comparison of six microarray technologies", Nucleic Acids Research, Volume 32, Issue 15, August 2004, Page e124, https://doi.org/10.1093/nar/gnh123

[10] Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A, "Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool", BMC Bioinformatics. 2013; 128(14).

[11] Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, McDermott MG, Monteiro CD, Gundersen GW, Ma'ayan A, “Enrichr: a comprehensive gene set enrichment analysis web server 2016 update”, Nucleic Acids Res. 2016 Jul 8; 44(W1):W90-7. doi: 10.1093/nar/gkw377.

[12] Xie Z, Bailey A, Kuleshov MV, Clarke DJB., Evangelista JE, Jenkins SL, Lachmann A, Wojciechowicz ML, Kropiwnicki E, Jagodnik KM, Jeon M, & Ma’ayan A, “Gene set knowledge discovery with Enrichr”, Current Protocols, 1, e90. 2021. doi: 10.1002/cpz1.90

[13] Paul Shannon, Andrew Markiel, Owen Ozier et al., "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks", Genome Research, CSH Press, 2003. doi:10.1101/gr.1239303

[14] S Killcoyne, GW Carter, J Smith, and J Boyle, "Cytoscape: a community-based framework for network modeling", Methods Mol Biol., 2009; 563:219-39. doi: 10.1007/978-1-60761-175-2_12

[15] NT Doncheva, JH Morris, J Gorodkin, and LJ Jensen, "Cytoscape StringApp: Network Analysis and Visualization of Proteomics Data", J Proteome Res., 2019 Feb 1; 18(2):623-632. doi: 10.1021/acs.jproteome.8b00702

[16] G Dennis Jr, BT Sherman, DA Hosack *et al.,* "DAVID: Database for Annotation, Visualization, and Integrated Discovery", *Genome Biol* **4,** P3 (2003). https://doi.org/10.1186/gb-2003-4-5-p3

[17] DW Huang, BT Sherman, Q Tan et al., "DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists", Nucleic Acids Research, Volume 35, Issue suppl_2, 1 July 2007, https://doi.org/10.1093/nar/gkm415

[18] DW Huang, BT Sherman, R Stephens et al., "DAVID gene ID conversion tool", Bioinformation. 2008 Jul 30;2(10):428-30. doi: 10.6026/97320630002428.

[19] Jui-Hung Hung, Tun-Hsiang Yang, Zhenjun Hu, Zhiping Weng, Charles DeLisi, “Gene set enrichment analysis: performance evaluation and usage guidelines”, Briefings in Bioinformatics, Volume 13, Issue 3, May 2012, Pages 281–291, https://doi.org/10.1093/bib/bbr049

[20] Aiko I. Klingler, Whitney W. Stevens, Bruce K. Tan, Anju T. Peters, Julie A. Poposki, Leslie C. Grammer, Kevin C. Welch, Stephanie S. Smith, David B. Conley, Robert C. Kern, Robert P. Schleimer, Atsushi Kato, “Mechanisms and biomarkers of inflammatory endotypes in chronic rhinosinusitis without nasal polyps”, Journal of Allergy and Clinical Immunology, Volume 147, Issue 4, 2021.

[21] Steven Xijin Ge, Dongmin Jung, and Runan Yao, “ShinyGO: a graphical gene-set enrichment tool for animals and plants”, Bioinformatics, Volume 36, Issue 8, 15 April 2020, Pages 2628–2629, https://doi.org/10.1093/bioinformatics/btz931

[22] Jing Chen, Eric E. Bardes, Bruce J. Aronow, and Anil G. Jegga, “ToppGene Suite for gene list enrichment analysis and candidate gene prioritization”, Nucleic Acids Research, Volume 37, Issue suppl_2, 1 July 2009, Pages W305–W311, https://doi.org/10.1093/nar/gkp427

[23] Zhang L, Yu X, Zheng L et al, “Lineage tracking reveals dynamic relationships of T cells in colorectal cancer”, Nature, 564, 2018.

[24] Nguyen QH, Pervolarakis N, Nee K, Kessenbrock K, ” Experimental considerations for single-cell RNA sequencing approaches”, Frontiers in Cell and Developmental Biology, 2018.

[25] Nguyen A, Khoo WH, Moran I, Croucher PI, Phan TG, “Single cell RNA sequencing of rare immune cell populations”, Frontiers in Immunology, 2018.

[26] Nguyen QH, Pervolarakis N, Blake K, Ma D, Davis RT, James N, et al, “Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity”, Nature Communications, 2018.

[27] Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, et al, “DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome”, Nature, 2008.

[28] Li H, Horns F, Wu B, Xie Q, Li J, Li T et al, “Classifying drosophila olfactory projection neuron subtypes by single-cell RNA sequencing”, Cell, 2017.

[29] Luo C, Keown CL, Kurihara L, Zhou J, He Y, Li J et al, “Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex”, Science, 2017.

[30] F Liang, L Peng, Y Ma, W Hu, W Zhang, M Deng, and Y Li, "Bioinformatics analysis and experimental validation of differentially expressed genes in mouse articular chondrocytes treated with IL-1β using microarray data", Exp Ther Med 23: 6, 2022

[31] G Li, X Li, M Yang *et al,* "Prediction of biomarkers of oral squamous cell carcinoma using microarray technology", *Sci Rep* **7,** 42105 (2017). https://doi.org/10.1038/srep42105

[32] S Duan, B Gong, P Wang, H Huang, L Luo, and F Liu, "Novel prognostic biomarkers of gastric cancer based on gene expression microarray: COL12A1, GSTA3, FGA and FGG", Molecular Medicine Reports 18, no. 4 (2018): 3727-3736. https://doi.org/10.3892/mmr.2018.9368

[33] A Katiyar, G Kaur, L Rani *et al.,* "Genome-wide identification of potential biomarkers in multiple myeloma using meta-analysis of mRNA and miRNA expression data", *Sci Rep* **11,** 10957 (2021). https://doi.org/10.1038/s41598-021-90424-y

[34] J Yang, MY Zhang, YM Du, XL Ji, and YQ Qu, "Identification and Validation of CDKN1A and HDAC1 as Senescence-Related Hub Genes in Chronic Obstructive Pulmonary Disease", Int J Chron Obstruct Pulmon Dis. 2022;17:1811-1825. https://doi.org/10.2147/COPD.S374684

[35] S. Na, L. Xumin and G. Yong, "Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm," 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, 2010, pp. 63-67, doi: 10.1109/IITSI.2010.74.

[36] Yewang Chen, Lida Zhou, Nizar Bouguila, Cheng Wang, Yi Chen, and Jixiang Du, "BLOCK-DBSCAN: Fast clustering for large scale data", Pattern Recognition, Volume 109, 2021

[37] A. Talwalkar, S. Kumar and H. Rowley, "Large-scale manifold learning," 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1-8, doi: 10.1109/CVPR.2008.4587670.

[38] JB Tenenbaum,VD Silva, and JC Langford,"A global geometric framework for nonlinear dimensionality reduction", Science 290 (5500), 2000

[39] S Roweis and L Saul, "Nonlinear dimensionality reduction by locally linear embedding", Science 290 (5500), 2000

[40] M Belkin and P Niyogi, "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation", Neural Computation, June 2003; 15 (6):1373-1396

[41] Ingwer Borg and Patrick Groenen, "Modern Multidimensional Scaling - Theory and Applications", Springer Series in Statistics, 1997

[42] Laurens van der Maaten, Geoffrey Hinton, "Visualizing High-Dimensional Data Using t-SNE", Journal of Machine Learning Research (2008)

[43] Belkina, A.C., Ciccolella, C.O., Anno, R. et al., "Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets", Nat Commun **10,** 5415 (2019). https://doi.org/10.1038/s41467-019-13055-y