Received: 18 August 2023 | Revised: 27 September 2023 | Accepted: 8 October 2023 | Published online: 10 October 2023

RESEARCH ARTICLE

Medinformatics
2024, Vol. 1(1) 27–34
DOI: 10.47852/bonviewMEDIN32021554

BON VIEW PUBLISHING

Identification of Key Gene Modules and Novel Transcription Factors in Tetralogy of Fallot Using Machine Learning and Network Topological Features

Sona Charles[1,2] **and Jeyakumar Natarajan**[1,*]

[1]*Department of Bioinformatics, Bharathiar University, India*

[2]*Crop Improvement and Biotechnology, ICAR-Indian Institute of Spices Research, India*

Abstract: Tetralogy of Fallot (TOF) is a combinatorial congenital abnormality comprising ventricular septal defect, pulmonary valve stenosis, a misplaced aorta, and a thickened right ventricular wall. Biologically relevant module identification from transcriptome data may be considered as a binary classification problem. We utilized publicly accessible mRNA expression data to extract the differentially expressed genes and further weighted gene co-expression network analysis to identify ten modules in TOF. Network topological properties of modular and non-modular genes were considered as features for binary classification. We applied support vector machine, random forest (RF), decision trees (DTs), K-nearest neighbor, and Naïve Bayes algorithms to network features. RF and DT algorithms displayed an accuracy of 99.1% and 98%, respectively. All the methods, in combination, predicted 71 common genes, which were used to construct a gene regulatory network. The network was expanded to include 30 miRNAs targeting the genes. Interestingly, 39 out of 71 genes were transcription factors out of which ELN, SOX6, and FOXO3 genes are novel candidates in TOF. The work also provides a sub-module of genes and miRNAs supported by statistical models as prospective candidates to be biomarkers.

Keywords: Tetralogy of Fallot, machine learning, network features, transcription factor, gene modules, WGCNA

1. Introduction

Tetralogy of Fallot (TOF) is a cyanotic cardiac abnormality that is characterized by four conditions, namely ventricular septal defect, overriding aorta, right ventricular hypertrophy, and pulmonary stenosis. The combination of these defects constitutes 7–10% of malformations of heart (Bailliard & Anderson, 2009). Several mutations in genes such as GATA4, NKX2.5, JAG1, FOXC2, TBX5, HAND2, and TBX1 have been implicated in TOF (Morgenthau & Frishman, 2018). The involvement of variants in transcription factors (TFs) is noteworthy. With the emergence of advanced sequencing technologies, rare sequence variants have been identified (Manshaei et al., 2020).

The classification of genes in a network into modular and non-modular categories serves as a fundamental necessity in understanding the complex machinery of biological systems. A gene module is a group of genes that have similar expression profiles and tend to be functionally related. A module consists of a group of co-expressed genes that form strongly connected sub-networks and can be extracted by applying clustering algorithms on gene expression data (van Dam et al., 2018). Modular genes (MGs) are those that function within distinct, interchangeable modules or units, often contributing to multiple biological processes (Melo et al., 2016). On the other hand, non-modular genes (NMGs) operate as essential, standalone entities with limited functional flexibility. It is significant to unveil the crucial nodes that belong to modules in critical developmental processes. The characteristics of a network are reflected in the topological properties (Tornow & Mewes, 2003). The structural and functional properties of a gene network are rewired within the higher levels of cellular pathways and hence the classification of MG and NMG aids in deciphering the intricate network of molecular interactions within a cell. The MGs are highly committed to the genes of the same module than others. Hence, the topological properties of a modular network are different from that of a non-modular network. However, it is practically not feasible to experimentally validate and identify all the genes in a module. Computational methods come into play in this scenario. Recognizing the distinction is vital for unraveling the genetic basis of any diseases, as mutations in MGs can lead to widespread consequences across different cellular pathways, while alterations in NMGs may have more specific and predictable effects (Melo et al., 2016).

Machine learning approaches have been applied to biological data to retrieve meaningful classification, prediction, discovery, and significant inferences. Machine learning models, trained on network topological features, can offer predictive power, aiding in

*Corresponding author: Jeyakumar Natarajan, Department of Bioinformatics, Bharathiar University, India. Email: n.jeyakumar@yahoo.co.in

the discovery of key genes or regulatory hubs that play pivotal roles in controlling biological pathways. Machine learning in combination with network topological features has been used to predict essential genes and proteins (Zhang et al., 2016), genes in metabolism (Moore et al., 2019), and driver genes in diseases (Han et al., 2019). Machine learning algorithms have great potential for mining patterns in complex biological data (Xu & Jackson, 2019). Given the challenges in mining the modularity, we utilized machine learning approaches to expand the hidden knowledge in RNASeq data to reveal intrinsic sub-networks or modules.

In this study, we propose to classify MGs and NMGs in TOF using machine learning algorithms. We applied weighted gene co-expression network analysis (WGCNA) algorithm on RNASeq data to extract functional modules in TOF. The network topological properties of the MG and NMG network were extracted. The machine learning algorithm was trained using the network topological properties as features. We combined five prediction algorithms to obtain a subset of MGs. Differentially expressed miRNAs were also integrated into the MG network. We applied a wide range of bioinformatics tools and statistical tests to extract meaningful insights from the datasets, which are detailed in the methodology section. This work provides an insight into the selection of MGs based on network topological properties.

2. Method

The methodology adopted for this study is summarized in Figure 1.

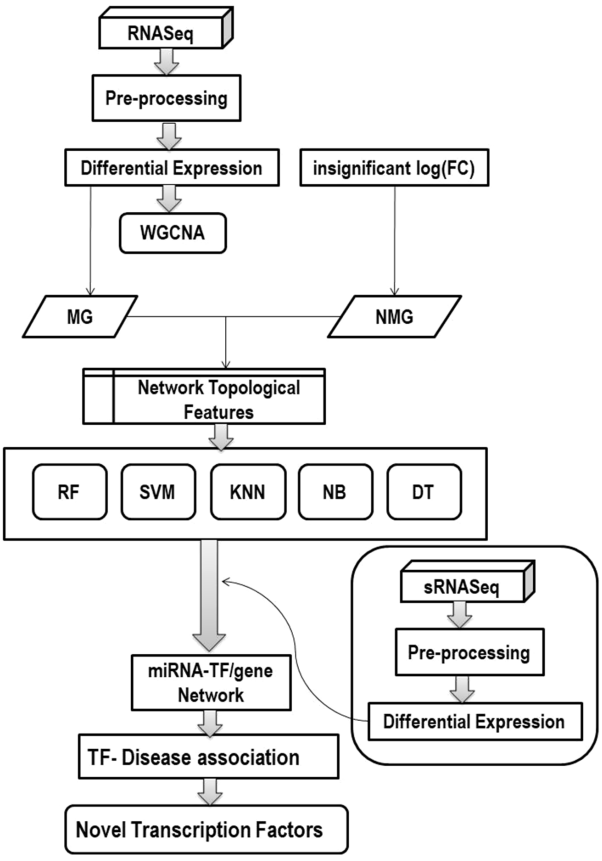


Figure 1. Study design and pipeline

2.1. Data collection and pre-processing

Publicly available raw gene expression reads from sample accession GSE36761 were downloaded from the Sequence Read Archive SRA database. The meta-analysis was performed with 22 TOF samples from right ventricles and 8 samples (4 right ventricular and 4 left ventricular, respectively) from normal unaffected hearts. The selection criterion was based on the fact that RNAseq data had corresponding sRNAseq that is quantified from the same patient samples. sRNAseq is used in the downstream analysis to include miRNAs in the study. Furthermore, the selected datasets were free from treatment. The gene/miRNA counts were extracted for further analysis. To evaluate the expression pattern in samples, principal component analysis (PCA) was performed to cluster sample with respect to its phenotype. PCA is done to identify and correct batch effects, ensuring that the observed differences are biologically meaningful rather than artifacts of the experimental process. Differential expression screening was performed using DeSeq2 with |log2 fold change| of >1 and FDR cutoff of 0.1 (Love et al., 2014). DESeq is specifically designed to identify genes that are differentially expressed between different conditions or groups in an RNA-Seq experiment by employing a negative binomial distribution to model the count data.

2.2. WGCNA and module identification

The genes that have mild levels of expression are uninformative and were filtered as a part of differential expression during DeSeq2 analysis. The expression of co-expressed genes is similar in samples under the same condition and involves closely related biochemical signaling pathways leading to common biological outcomes. In order to identify the co-expressed genes in TOF, we used the WGCNA package in R Bioconductor (Langfelder & Horvath, 2008). The differentially expressed genes (DEGs) with |log2 fold change| of >1 and p-value < 0.05 were considered for WGCNA. WGCNA calculates the correlation between all the genes which is then converted into an adjacency matrix through soft thresholding (β value). A scale-free network consists of small number of genes with high connectivity (Khanin & Wit, 2006). Low connectivity genes will be abundant in a scale-free network. Soft threshold value of $\beta = 5$ was considered as a cutoff for obtaining a scale-free topology of the network. The dendrogram with corresponding modules was derived for further analysis.

2.3. Feature identification

The MGs identified from WGCNA were used to train a model to predict prioritized genes in the network. We mapped the co-expressed MGs to STRING database into a PPI network. NMGs were identified from the DEGs with |log (FC)| values between 0.10 and −0.108o9. We considered the fact that NMGs will have insignificant differential expression between normal and TOF samples. Similarly the NMGs were mapped into another PPI network. The interactome was analyzed using Network Analyzer app of Cytoscape (Shannon et al., 2003). The topological properties of MG and NMG networks were calculated using Network Analyzer (Assenov et al., 2008). We shortlisted average shortest path, clustering coefficient, closeness centrality, eccentricity, stress, degree, betweenness centrality, neighborhood connectivity, radiality, and topological coefficient as the features. Feature selection improves the performance of the model by removing those features that do not contribute to the prediction variable. The feature importance was evaluated using Boruta algorithm (Chen et al., 2020). The algorithm permutes and

extends the initial data with artificial features known as shadows to assess the importance of actual features in high-dimensional datasets. The algorithm then performs a feature selection process by training a classifier on both the original dataset and the shadow dataset. Features that consistently perform better than their random counterparts are deemed relevant, while those that do not are considered uninformative. By comparing the actual features' importance scores to those of the shadow features, Boruta provides a principled way to identify the most relevant attributes, mitigating the risk of overfitting and enhancing the model's generalization performance in machine learning tasks.

2.4. Training the model

Identification of MG and NMG can be considered as a binary classification problem. We considered only the genes in modules enriched with GO terms. We obtained a total of 585 annotated genes as positive dataset. In order to create balanced data, we extracted 585 NMGs, which are those genes not present in any module and do not show significant differential expression (|logFC| between 1 and −1). We divided the data into training set and test set in the ratio 70:30 with two classes, modular and non-modular. We implemented five popular machine learning classification algorithms K-nearest neighbors (KNN), decision tree (DT), support vector machine (SVM), Naive Bayes (NB), and random forest (RF) using Sklearn python library (Pedregosa et al., 2011) to evaluate the binary classification model. In order to prevent bias due to the small size of the dataset, 10-fold cross-validation was performed in which 10 subsets of the dataset were randomly selected and one subset was fixed as the testing set (one fold) while the others were considered as training set data (9 folds). The average performance was calculated over 10 folds.

2.5. Classification algorithms

2.5.1. Support vector machines

The SVC class from Scikit-learn Python library (Pedregosa et al., 2011) was used for implementing SVM. We applied SVM using linear, polynomial, and RBF kernel. RBF kernel was identified to be the best kernel for our data.

2.5.2. Random forest

The RF model was developed using RandomForestClassifier (Breiman, 2001) of Scikit Learn. RF is the ensemble of classification or regression trees and has a wide array of applications in biological studies. It supports possible interactions among variables of the datasets.

2.5.3. K-nearest neighbor

KNN is a parametric supervised machine learning algorithm. KNeighborsClassifier (Fix & Hodges, 1989) with $k = 3$ (obtained after hyperparameter tuning) was used to generate the KNN model.

2.5.4. Naïve Bayes

NB is a supervised learning algorithm, which assumes that every pair of features is conditionally independent with respect to the class variable. The algorithm is based on Bayes' theorem. We implemented Gaussian NB algorithm Gaussian NB (Zhao et al., 2018) to our dataset.

2.5.5. Decision trees

DTs are decision rule-based supervised algorithms. DecisionTreeClassifier (Quinlan, 1986) was applied to the dataset to generate the model.

2.6. Performance evaluation

We used four performance evaluation metrices to evaluate the models: accuracy, precision, recall, and F-1 score. They were calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\ score = \frac{2\ (Precision\ x\ Recall)}{Precision + Recall}$$

where TP is the true positives, TN is the true negatives, FP is the false positives, and FN is the false negatives.

2.7. Integration of miRNAs and TFs in network

We combined the MGs selected by five prediction algorithms in this study. The MGs selected by all the five classifiers were integrated into an interaction network. The miRNAs targeting the MGs were identified using the combined target prediction tool miRDip (Tokar et al., 2018). The cutoff was set to 'very high' with a minimum overlap of five target prediction tools. The miRNAs were then cross-validated for their expression pattern in TOF by comparing with the differentially expressed miRNAs in the dataset (GSE36759). The miRNAs were included in the prioritized gene interaction network. Experimental datasets of TF2DNA database were utilized to identify the TFs in the network (Pujato et al., 2014). The enrichment analysis of genes was performed using gene set enrichment analysis (GSEA) (Subramanian et al., 2005).

2.8. Corroboration with experimental evidences

We attempted to map the results obtained on to the existing literature and validated candidate genes in TOF. The TFs were queries against DisGeNet database (Piñero et al., 2015) to identify associations with other cardiovascular diseases. TF-variant association was also performed to collect the variants commonly present in the TFs, which might play a role in TOF.

3. Results and Discussion

3.1. PCA and differential expression analysis

The normal and TOF samples were well separated with a variance of 28% and 14% across PC1 and PC2. The significant genes that fulfill the cutoff criteria for differential expression were extracted by DeSeq2 (Figure 2). The DEGs and fold change are provided in Supplementary Table S1. The numbers of upregulated and downregulated genes in TOF are 927 and 1246, respectively (Supplementary Table S2 and S3).

DEGs in a condition tend to be a part of cluster or module of genes, which act in synchrony to bring about a phenotypic change.

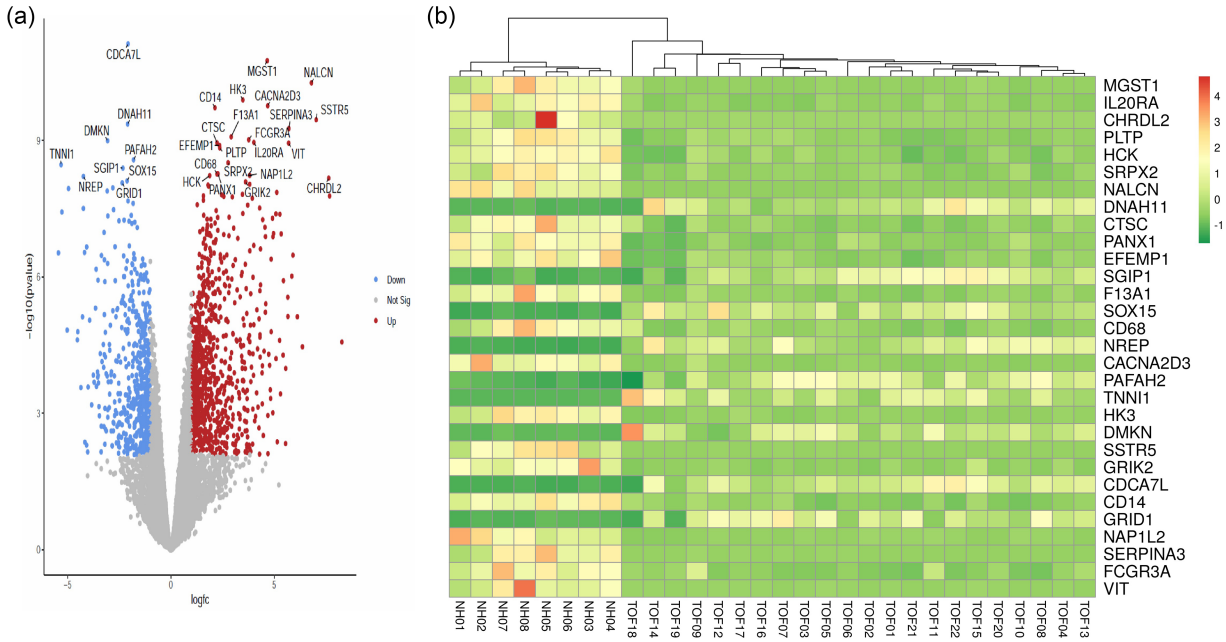


Figure 2. (a) Volcano plot showing differential expression pattern, (b) Heatmap of top 30 DEGs

3.2. Module extraction

WGCNA is a powerful bioinformatics tool used extensively to uncover MGs within complex biological systems. This approach relies on the concept that genes with similar expression patterns across various samples are likely to be functionally related and involved in similar biological processes. WGCNA constructs a network of gene co-expression relationships, where nodes represent genes, and edges signify the strength of their associations (Langfelder & Horvath, 2008).

WGCNA was performed with hierarchical tree clustering algorithm. The soft threshold value $\beta = 5$ had the highest agreement toward a scale-free topology in the network with fit index value of 0.90 (Figure 3). With the above parameters, WGCNA identified 10 modules with a total of 1712 genes (Figure 4). The GO enrichment as well as pathway analysis of each module was performed and summarized in Supplementary Tables S3–S8. The turquoise module consisted of 444 genes (highest), while the purple module consisted of only 21 genes. In WGCNA, the genes are clustered into modules combining the adjacency of two genes and the strength of interaction of these two genes with the neighboring genes. Hence, the genes within modules have high connectivity among themselves than genes in other modules.

The number of genes in each module is given in Table 1.

Functional annotation by gene ontology analysis was performed to study the biological significance of genes in each module. Out of the 10 modules, only 5 modules (turquoise, brown, yellow, red, and purple)

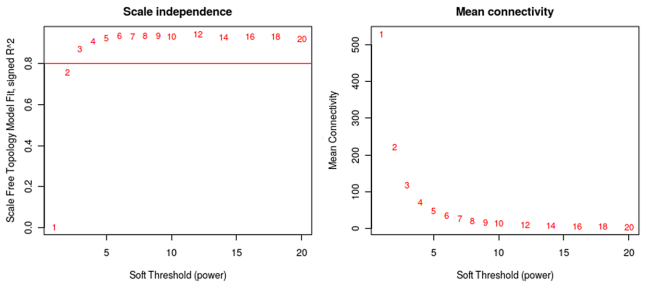


Figure 3. Network topology analysis for varying soft thresholds based on scale fit index and mean connectivity

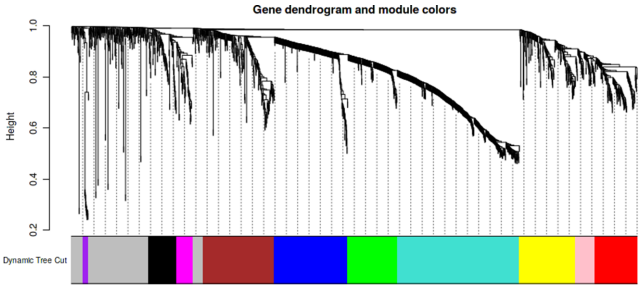


Figure 4. Clustering dendrogram of genes obtained from WGCNA with assigned module colors

Table 1. Module colors and number of genes present in each module

Sl no.	Module color	Number of genes
Module 1	Turquoise	433
Module 2	Blue	260
Module 3	Brown	252
Module 4	Yellow	198
Module 5	Green	174
Module 6	Red	149
Module 7	Black	98
Module 8	Pink	69
Module 9	Magenta	59
Module 10	Purple	20

were mapped to gene ontology terms. Only two modules were mapped to KEGG pathways. Turquoise module was mapped to KEGG pathways such as cardiac muscle contraction and dilated cardiomyopathy. Surprisingly, 39 genes were enriched in the term coronavirus disease, which suggest a possibility of altering genes in TOF condition. Red module was enriched in pathways related to infection such as complement and coagulation cascades, phagosome, systemic lupus erythematosus, and Staphylococcus aureus infection. Biological process category was mostly related to terms in association with morphogenesis, differentiation, translation, and immune response.

Functional annotation by gene ontology analysis was performed to study the biological significance of genes in each module. Out of the 10 modules, only 5 modules (turquoise, brown, yellow, red, and purple) were mapped to gene ontology terms (Supplementary Figures 1-5 represent the GO category enrichment in each module).

3.3. Training and performance of models

For developing a computational model, the most important step is the construction of a dataset that efficiently reflects the pattern of the class variable. We used 1170 genes as the training dataset. We selected 585 non-redundant MGs (positive set) and 585 NMGs (negative set) as the training set. To prevent the class imbalance, we included an equal number of positive and negative data in our training set. The topological features were extracted.

Feature selection methods reveal the optimal features in a dataset since the performance of the classifier tends to be crucially dependent on the attributes of the genes. Boruta feature selection algorithm evaluated and reported all features as non-redundant and significant in the classification problem (Figure 5). Radiality was identified as the most important feature.

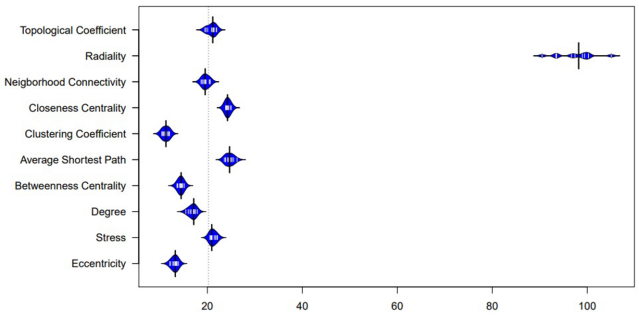


Figure 5. Feature significance evaluation of Boruta algorithm on network topological features

The performance of five models was compared in terms of accuracy, precision, recall, and F-1 score (Figure 6). Results of all the models RF and DTs gave the best classification accuracy of 99.1% and 98%, respectively, followed by NB, SVM, and KNN (Supplementary Table S9). In total, 71 genes were commonly identified as modular in all the methods. In order to validate the model, we used GSE217772 as an independent dataset. GSE217772 consists of samples from 5 TOF-affected children and 5 healthy individuals. RF model performed at an accuracy of 98.9% on the dataset.

Any classification problem is expected to have false positives and false negative results due to the noise in target data. However, RF and DTs performed well on the data. Since modularity is a property of the genes in a network, the topological properties of the network are the best features that can predict whether a gene is modular or not. Several studies have exploited network topological properties for predictions such as infectious disease genes, essential genes, and disease–gene

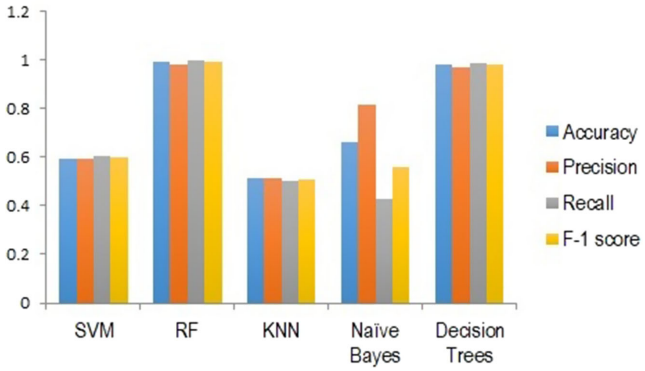


Figure 6. Comparison of performance measures of SVM, RF, KNN, Naïve Bayes, and decision trees

relationship (Azhagesan et al., 2018; Barman et al., 2019; Ramadan et al., 2016). Since topological structure of a network defines the modularity of genes within the network, they are efficient features to classify MGs and NMGs. Moreover, we selected the MGs in the positive training dataset from WGCNA, which is a highly robust approach in identifying gene modules. Our negative data in the training set are selected based on two criteria a) not being a part of the module selected by WGCNA and b) not significantly differentially expressed. Boruta algorithm also predicted important features and identified no features to be omitted for the classification.

3.4. Network construction and enrichment analysis

The MGs identified from all the prediction algorithms were reconstructed into an interaction network using String database version 11.0. String database contains experimentally validated as well as predicted results for interactions between proteins. In order to provide more insights into the regulation and modularity of the network, we included miRNAs that are targets of the genes. From dataset GSE36759, 30 miRNAs whose expression was quantified were included in the network (Figure 7). Further, we performed a GSEA of the 71 genes, which identified the overrepresented

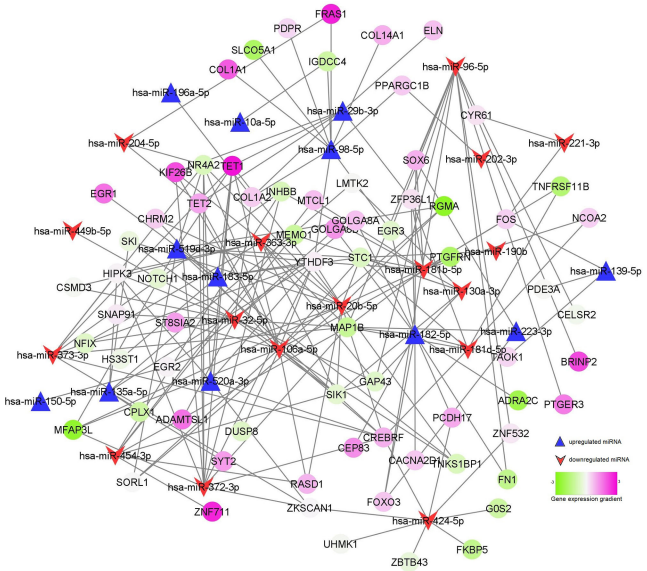


Figure 7. Sub-module of genes integrated with differentially expressed miRNAs

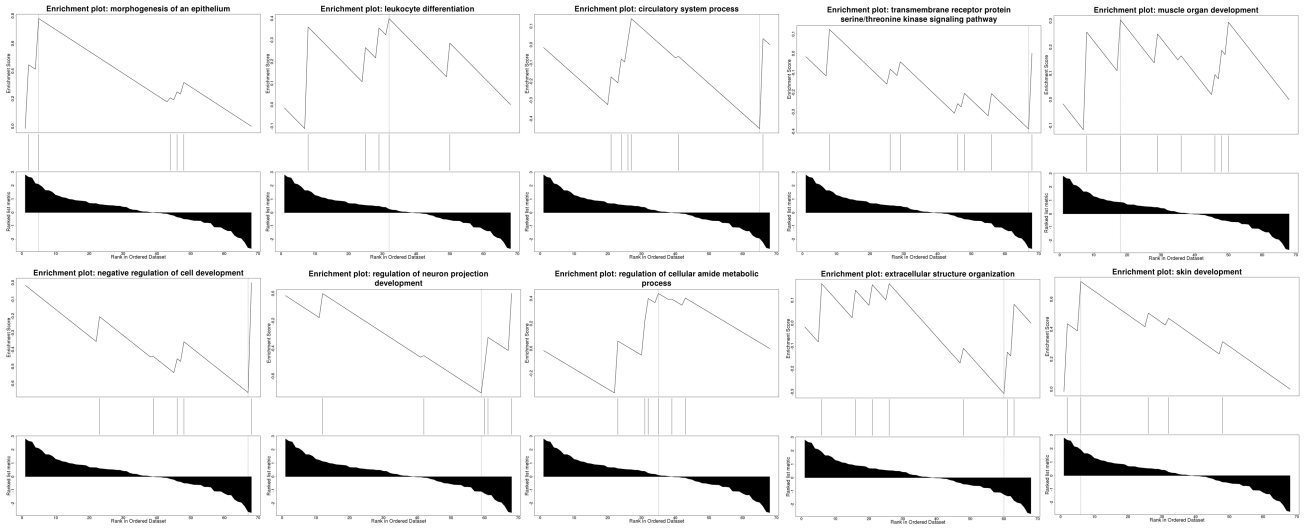


Figure 8. Top 10 results of gene set enrichment analysis of sub-modular genes

pathways. TF2DNA analysis revealed 39 out of 71 genes to be TFs (Supplementary Table S10). Among 71 genes, 59 genes were assigned to functional categories by GSEA after 1000 permutations. The top 10 functionally enriched categories are given in Figure 8.

3.5. Validation of key genes

The network constructed by intersecting MGs selected by all algorithms can be considered as central network of MGs in TOF. Examination of the gene set in this network revealed processes/ pathways related to developmental processes, which is very much significant in TOF. Furthermore, we can consider this network as a core sub-module with integrated miRNAs. The miRNAs targeting the genes were selected in such a way that they express a differential expression of $|\log FC| > 2$. The network is highly enriched in TFs and is highly robust. Given the critical role that TFs play in TOF, it is reasonable to anticipate that the TFs identified in this study are instrumental in shaping the intricate gene regulatory network underpinning the disease. In order to ensure the validity of the above findings, we further delved into these results and conducted a literature survey to identify the novel TFs that play a pivotal role in TOF. Rare variants affecting the function of TFs are commonly observed in TOF (Töpf et al., 2014). Hence, we focused on identifying disease relationships of TFs using DisGeNet database. Upon querying, we obtained 51 cardiovascular diseases associated with 18 TFs (Supplementary Table S11). The network of TF diseases is depicted in Figure 9.

We observed that ELN is involved in the highest number of diseases related to the heart. Elastin (ELN) is found abundantly in the arteries and is associated with 7q11.23 CNV, which is a loss of function mutation consequently leading to supravalvar aortic stenosis (Richards & Garg, 2010). Furthermore, we also studied the variants in the TFs (Supplementary Table S12). SOX6 and FOX03 were found to have intronic mutations, rs10832571 and rs12212067, respectively, in relation to cardiovascular diseases. We also checked the expression patterns of these genes in the independent dataset GSE217772 and found that the pattern of expression (up/down regulation) of 13 genes was same in both cases (Figure 10).

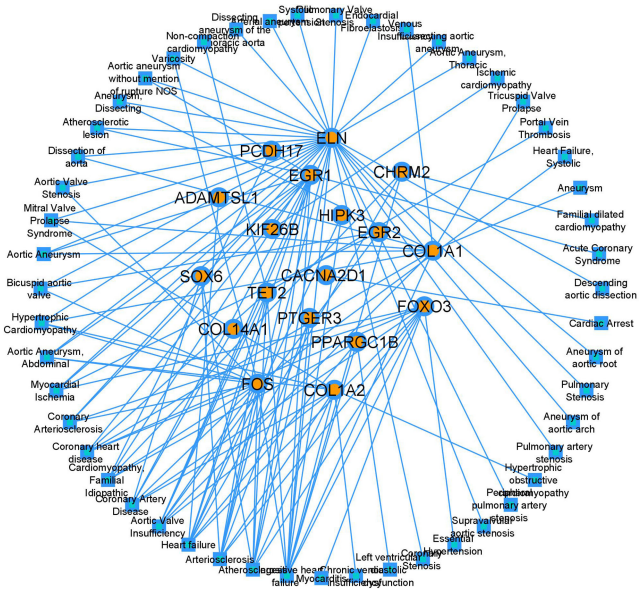


Figure 9. Transcription factor-disease association network

In summary, we constructed the sub-module of 71 DEGs and 30 miRNA from RNAseq and sRNASeq datasets in Tetralogy of Fallout. Further enrichment analysis reveals following enriched pathways such as morphogenesis of epithelium, leukocyte differentiation, circulatory system process, serine/threonine kinase signaling pathway, and muscle development. The enriched terms are related to development, differentiation, and morphogenesis, which rightly points out the validity of selected MGs in a developmental disorder. The identified TFs may be considered for experimental validation in TOF cases since they have high potential of being candidate genes or markers in TOF. Our study highlights the significance of module detection-based studies of diseases to identify potential biomarker candidates.

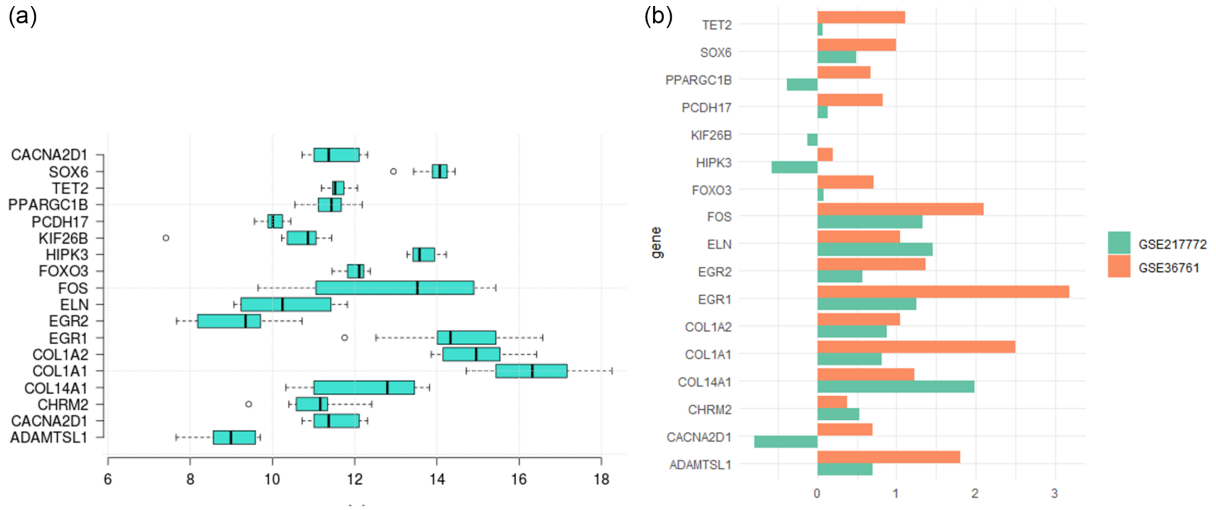


Figure 10. (a) The expression pattern of 18 TFs in independent validation dataset GSE217772, (b) Differential expression pattern of 18 TFs of GSE36761 with respect to GSE217772

4. Conclusion

In conclusion, our in-silico analysis demonstrates the promising potential of network features in effectively discerning between MG and NMG. Additionally, through the utilization of machine learning techniques, we have identified a specific subset of MGs that could play a pivotal role within a critical sub-module relevant to TOF. By integrating these MGs with microRNAs, we have unveiled a regulatory network enriched with highly significant TFs. Notably, we have further examined these TFs for variants that hold significance in the context of cardiovascular diseases. Our findings also strongly suggest a noteworthy involvement of ELN, SOX6, and FOXO3 TFs in the pathogenesis of TOF. This comprehensive analysis underscores the potential for a deeper understanding of the molecular mechanisms underlying TOF and presents avenues for future research and therapeutic exploration. Our present in-silico study has potential for experimental validation to detect biomarkers that may exist within the sub-module. By harnessing the power of computational techniques and data-driven insights, novel key genetic factors that contribute to the disease's pathogenesis can be identified, which aids in early diagnosis risk assessment and the way for personalized treatment strategies.

Acknowledgments

The authors are grateful to Mr. Subeesh A at the ICAR-Central Institute of Agricultural Engineering, Bhopal, India for the assistance in data analysis.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support this work are available upon reasonable request to the corresponding author.

References

Assenov, Y., Ramírez, F., Schelhorn, S. E., Lengauer, T., & Albrecht, M. (2008). Computing topological parameters of biological networks. *Bioinformatics*, *24*(2), 282–284. https://doi.org/10.1093/bioinformatics/btm554

Azhagesan, K., Ravindran, B., & Raman, K. (2018). Network-based features enable prediction of essential genes across diverse organisms. *PLoS ONE*, *13*(12), e0208722. https://doi.org/10.1371/journal.pone.0208722

Bailliard, F., & Anderson, R. H. (2009). Tetralogy of Fallot. *Orphanet Journal of Rare Diseases*, *4*, 2. https://doi.org/10.1186/1750-1172-4-2

Barman, R. K., Mukhopadhyay, A., Maulik, U., & Das, S. (2019). Identification of infectious disease-associated host genes using machine learning techniques. *BMC Bioinformatics*, *20*, 736. https://doi.org/10.1186/s12859-019-3317-0

Chen, R. C., Dewi, C., Huang, S. W., & Caraka, R. E. (2020). Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, *7*(1), 52. https://doi.org/10.1186/s40537-020-00327-4

Fix, E., & Hodges, J. L. (1989). Discriminatory analysis. Nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique*, *57*(3), 238–247. https://doi.org/10.2307/1403797

Han, Y., Yang, J., Qian, X., Cheng, W. C., Liu, S. H., Hua, X., . . . , & Lu, Y. (2019). DriverML: A machine learning algorithm for identifying driver genes in cancer sequencing studies. *Nucleic Acids Research*, *47*(8), e45. https://doi.org/10.1093/nar/gkz096

Khanin, R., & Wit, E. (2006). How scale-free are biological networks. *Journal of Computational Biology*, *13*(3), 810–818. https://doi.org/10.1089/cmb.2006.13.810

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*(1), 559. https://doi.org/10.1186/1471-2105-9-559

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.

Genome Biology, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Manshaei, R., Merico, D., Reuter, M. S., Engchuan, W., Mojarad, B. A., Chaturvedi, R., . . . , & Bassett, A. S. (2020). Genes and pathways implicated in tetralogy of Fallot revealed by ultra-rare variant burden analysis in 231 genome sequences. *Frontiers in Genetics*, *11*, 957. https://doi.org/10.3389/fgene.2020.00957

Melo, D., Porto, A., Cheverud, J. M., & Marroig, G. (2016). Modularity: Genes, development, and evolution. *Annual Review of Ecology, Evolution, and Systematics*, *47*, 463–486. https://doi.org/10.1146/annurev-ecolsys-121415-032409

Moore, B. M., Wang, P., Fan, P., Leong, B., Schenck, C. A., Lloyd, J. P., . . . , & Shiu, S. H. (2019). Robust predictions of specialized metabolism genes through machine learning. *Proceedings of the National Academy of Sciences*, *116*(6), 2344–2353. https://doi.org/10.1073/pnas.1817074116

Morgenthau, A., & Frishman, W. H. (2018). Genetic origins of tetralogy of Fallot. *Cardiology in Review*, *26*(2), 86–92. https://doi.org/10.1097/CRD.0000000000000170

Pedregosa, F., et al. (2011) Scikit-Learn Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Piñero, J., Queralt-Rosinach, N., Bravo, A., Deu-Pons, J., Bauer-Mehren, A., Baron, M., . . . , & Furlong, L. I. (2015). DisGeNET: A discovery platform for the dynamical exploration of human diseases and their genes. *Database*, *2015*, bav028. https://doi.org/10.1093/database/bav028

Pujato, M., Kieken, F., Skiles, A. A., Tapinos, N., & Fiser, A. (2014). Prediction of DNA binding motifs from 3D models of transcription factors; identifying TLX3 regulated genes. *Nucleic Acids Research*, *42*(22), 13500–13512. https://doi.org/10.1093/nar/gku1228

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*, 81–106. https://doi.org/10.1007/BF00116251

Ramadan, E., Alinsaif, S., & Hassan, M. R. (2016). Network topology measures for identifying disease-gene association in breast cancer. *BMC Bioinformatics*, *17*, 274. https://doi.org/10.1186/s12859-016-1095-5

Richards, A., & Garg, V. (2010). Genetics of congenital heart disease. *Current Cardiology Reviews*, *6*(2), 91–97. https://doi.org/10.2174/157340310791162703

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., . . . , & Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*, 2498–2504. http://www.genome.org/cgi/doi/10.1101/gr.1239303.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., . . . , & Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, *102*(43), 15545–15550. https://doi.org/10.1073/pnas.0506580102

Tokar, T., Pastrello, C., Rossos, A. E., Abovsky, M., Hauschild, A. C., Tsay, M., . . . , & Jurisica, I. (2018). mirDIP 4.1—Integrative database of human microRNA target predictions. *Nucleic Acids Research*, *46*(D1), D360–D370. https://doi.org/10.1093/nar/gkx1144

Töpf, A., Griffin, H. R., Glen, E., Soemedi, R., Brown, D. L., Hall, D., . . . , & Goodship, J. A. (2014). Functionally significant, rare transcription factor variants in tetralogy of Fallot. *PLoS ONE*, *9*(8), e95453. https://doi.org/10.1371/journal.pone.0095453

Tornow, S., & Mewes, H. W. (2003). Functional modules by relating protein interaction networks and gene expression. *Nucleic Acids Research*, *31*(21), 6283–6289. https://doi.org/10.1093/nar/gkg838

van Dam, S., Võsa, U., van der Graaf, A., Franke, L., & de Magalhães, J. P. (2018). Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in Bioinformatics*, *19*(4), 575–592. https://doi.org/10.1093/bib/bbw139

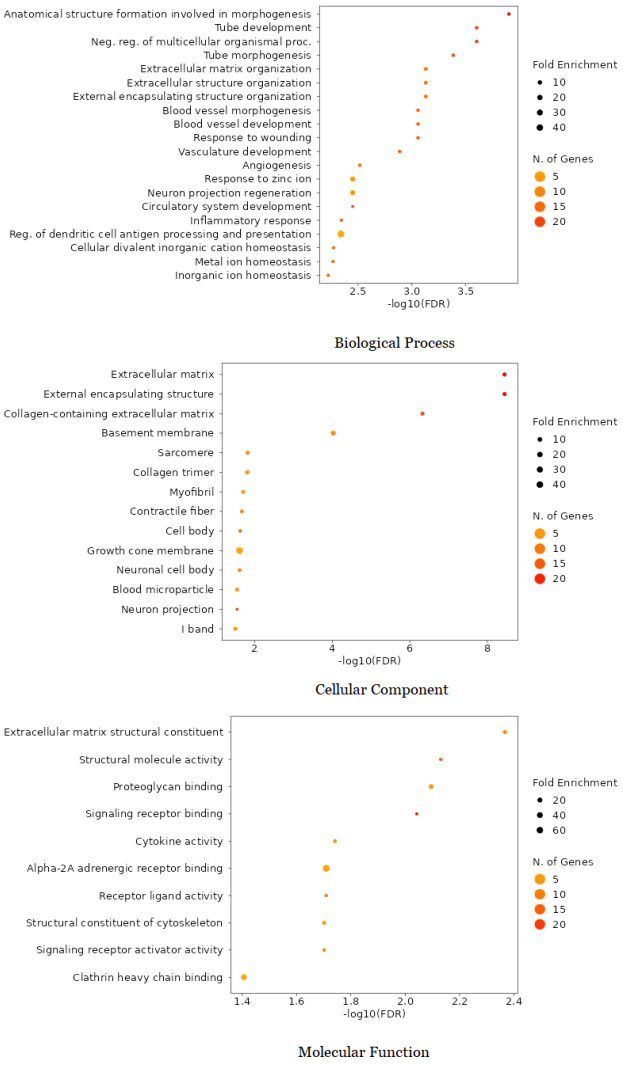
Xu, C., & Jackson, S. A. (2019). Machine learning and complex biological data. *Genome Biology*, *20*, 76. https://doi.org/10.1186/s13059-019-1689-0

Zhang, X., Acencio, M. L., & Lemke, N. (2016). Predicting essential genes and proteins based on machine learning and network topological features: A comprehensive review. *Frontiers in Physiology*, *7*, 75. https://doi.org/10.3389/fphys.2016.00075

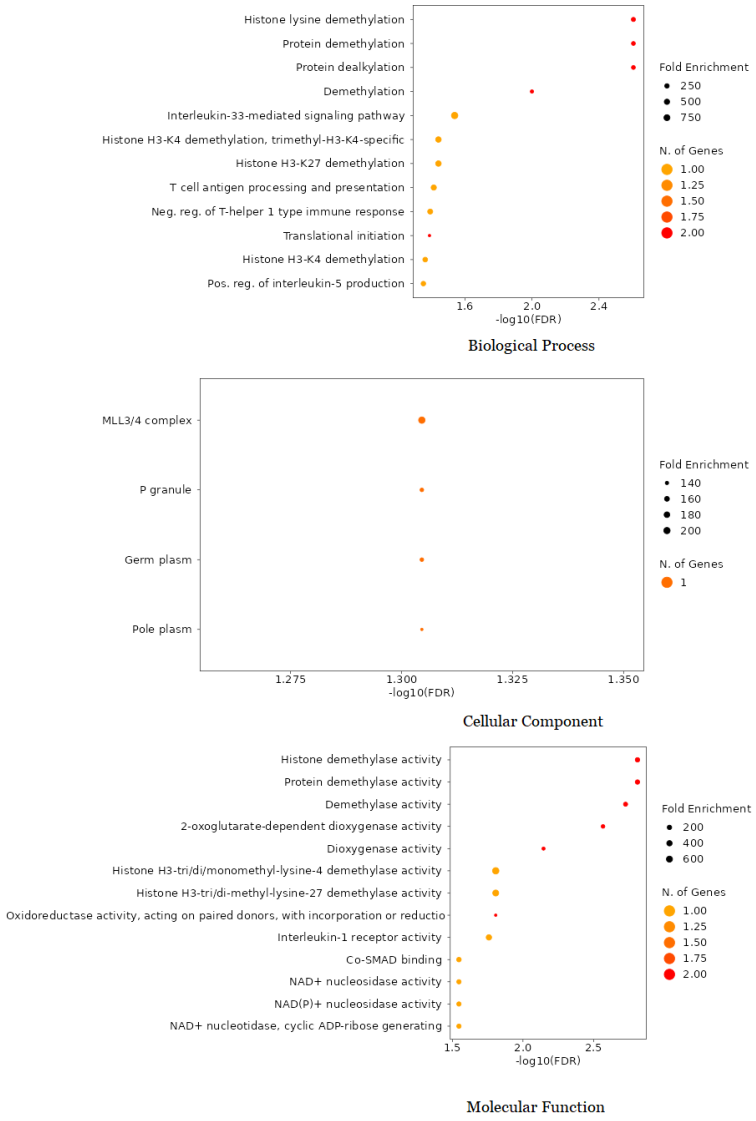
Zhao, T., Deng, L., Wang, W., Elson, D. S., & Su, L. (2018). Bayes' theorem-based binary algorithm for fast reference-less calibration of a multimode fiber. *Optics Express*, *26*(16), 20368–20378. https://doi.org/10.1364/OE.26.020368

How to Cite: Charles, S., & Natarajan, J. (2024). Identification of Key Gene Modules and Novel Transcription Factors in Tetralogy of Fallot Using Machine Learning and Network Topological Features. *Medinformatics*, *1*(1), 27–34. https://doi.org/10.47852/bonviewMEDIN32021554

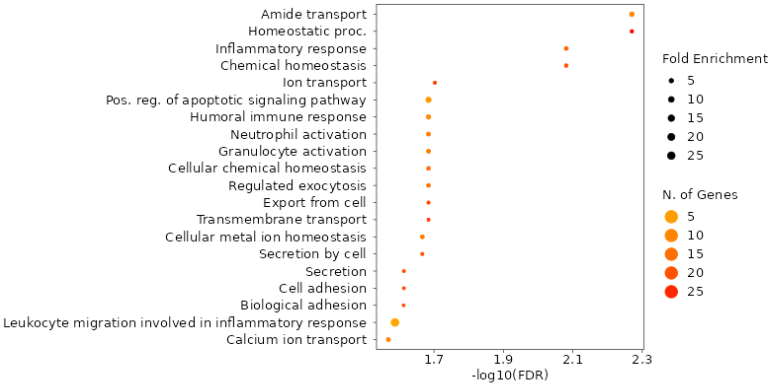
Supplementary Figures



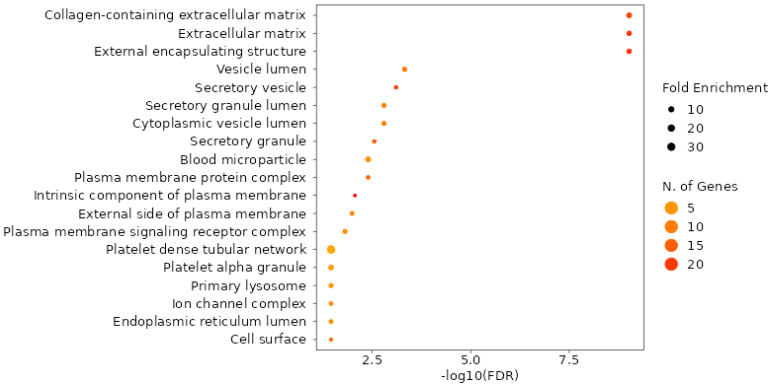
Supplementary Figure 1. GO yellow module



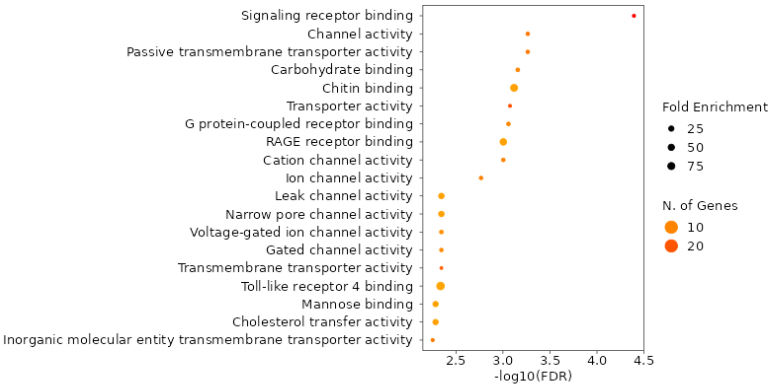
Supplementary Figure 2. Purple module



Biological Process

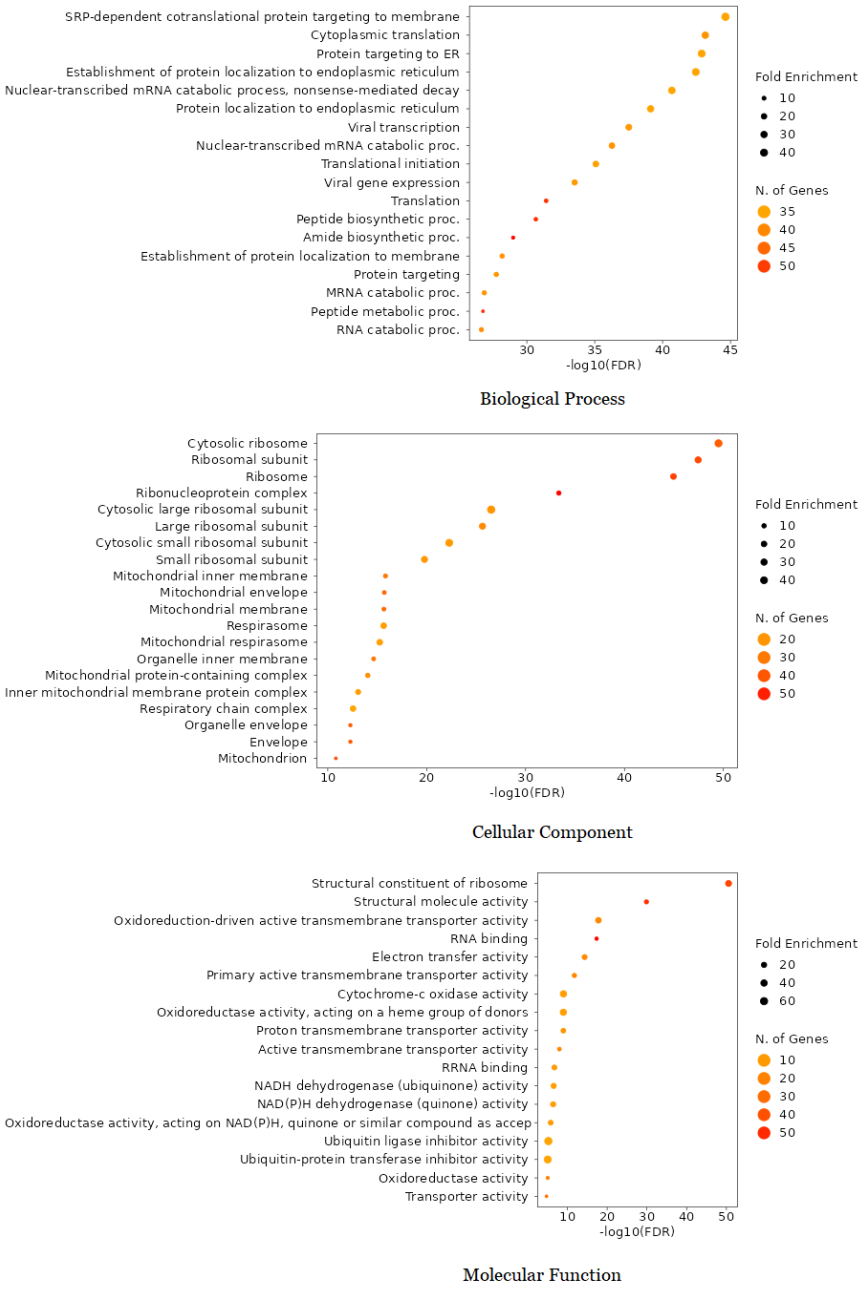


Cellular Component

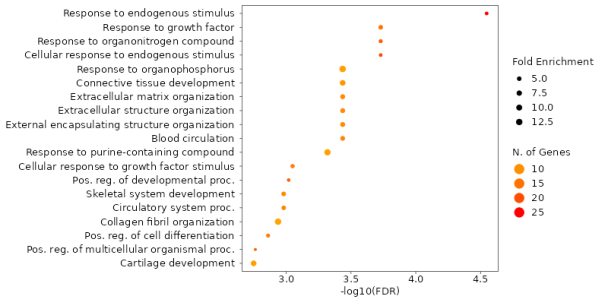


Molecular Function

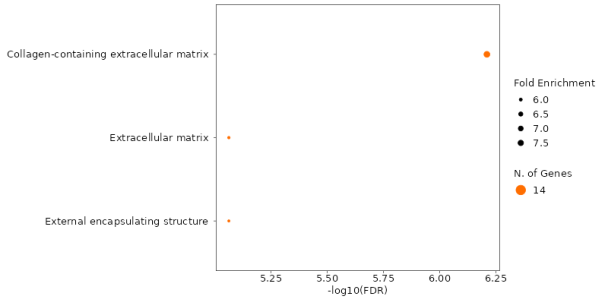
Supplementary Figure 3. Red module



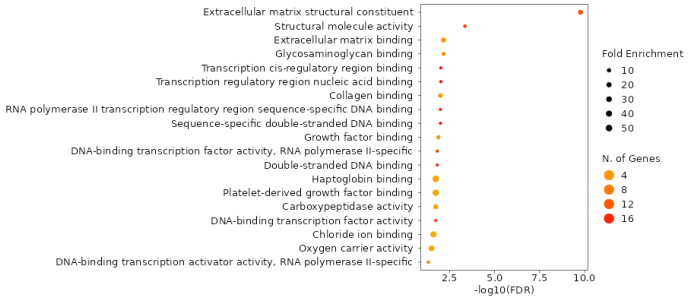
Supplementary Figure 4. Turquoise module



Biological Process



Cellular Component



Molecular Function

Supplementary Figure 5. Brown module