Identification and Predictive Analysis of Differentially Expressed Genes in Lung Adenocarcinoma

Chalasani Namratha Shivani

INFO-B528 Computational Analysis of Hight-throughput Biomedical Data

Final project report

Abstract

Motivation: Lung Adenocarcinoma (LUAD) is the most common histological subtype of non-small cell lung cancer and a leading cause of cancer death. Due to the high heterogeneity of LUAD it became a major public health problem with poor survival rate. Comprehending how intrinsic factors and environmental stresses contribute to the transformation of cells, and how cancer cells manipulate signaling pathways in their cell of origin, can provide possibilities to customize treatments for different subsets of LUAD. Hence the aim of this study is performing predictive analysis, identifying differentially expressed genes (DEGs), and seeing survival in gene clusters is to gain insights into the molecular mechanisms that underlie cancer progression and to develop predictive models that can aid in the diagnosis and treatment of cancer.

Results: A total of 90 genes have been identified as highly differentially expressed in tumor and normal samples. However, co-expression network analysis did not reveal any significant relationship between gene expression and survival time, indicating that these highly differentially expressed genes may not be directly linked to patient survival. Predictive models built using these 90 genes and covariates such as age, sex, gender, survival time, and vital status showed that the logistic regression model was able to determine positive and negative results with 63% accuracy.

1 Introduction

Lung cancer is the third most common cancer with major causative factor being tobacco smoking. As a result of the highly heterogenous nature of lung cancer, it remains the leading cause of death worldwide and inflicts one of the lowest 5-year survival rate (1). According to GLOBALCON 2020 statistics, lung cancer is the second most diagnosed cancer in both sexes and the leading cause of death comprising 18% of all cancer deaths (2). In the United States (US) alone, lung cancer is expected to lead to an estimated 131,880 deaths in the year 2021, more than any other cancer (3). According to American cancer society, 238,340 people have been estimated to be diagnosed with lung cancer (4).

Based on the histology, lung cancer is divided into two types: Small Cell Lung cancer and Non-Small Cell Lung cancer. The latter accounting for more than 80% of all cases of lung cancer. Non-small cell lung cancer is further classified into four different types, out of which Lung Adenocarcinoma (LUAD) represents more than 50% of the incidence (1). LUAD usually evolves from the mucosal glands and is found to have strong association with smoking history. However, it is also the most common subtype seen in non-smokers (5).

Integrated genome and transcriptome analysis have shown that LUAD is characterized by high tumor mutational burden. Many studies on the microenvironment and progression of LUAD revealed that the prognosis of LUAD depends on the associations between the tumor oncogenes and the microenvironment from which the tumor originated (1). In this study, the objective was to obtain a better understanding of the which genes are involved in the progression of cancer and to create predictive models that can assist in the diagnosis and treatment of cancer. This is achieved by conducting predictive analyses, identifying genes that are expressed differently, and analyzing survival rates in gene clusters.

2 Methods

2.1 Data Collection and Preprocessing

The data for this study has been taken from TCGA-LUAD project. Clinical and Transcriptome data from 577 cases, 518 primary tumor and 59 normal have been downloaded using R TCGAbiolinks package (6). The missing values for stage have been imputed following the TMN cancer staging classification by using the T, M and N data for the samples in the clinical data table. The vital status column in the clinical data has been converted to 1, and 0 representing alive and dead respectively and an overall survival time has been added using the days to last follow-up and days to death columns in the clinical data. These two columns are later used for survival analysis.

2.2 Differential Gene Expression Analysis

Differential gene expression analysis of the datasets is done using the DESeq2 (7). DESeqDataSetFromMatrix function was used to create an DESeq object designed over the type of sample (tumor or normal) and covariates. Genes with counts greater than equal to 10 in more than 75% of the samples are retained while the others are removed. Genes with FDR less than 0.05 and |logFC >=0| are considered as differentially expressed genes. Gene Ontology analysis was performed on the differentially expressed genes using enrichGO function in clusterprofiler R package (8).

2.3 Co-Expression Network Construction

The co-expression network was constructed using the weighted gene co-expression network analysis (WGCNA) package (9) in R on the 518 tumor samples. To reduce noise while constructing the weighted adjacency matrix, a power was selected using the pickSoftThreshold() function. Using the adjacency matrix, a topological matrix (TOM) was created that estimates the connectedness of two nodes (genes). The genes were then hierarchically clustered based on the TOM-based dissimilarity (1-TOM) using the flashClust library. The clusters were divided into modules using the dynamic tree-cutting method with a cut height of 0.99, minimum cluster size of 300, and deep split of 4. The deep split with a small number of large modules was selected for further analysis. WGCNA with a minimum cluster size of 30 was performed using only the top 2000 genes sorted on adjusted p-values to check for variations in the analysis. This difference in cluster size was chosen to include as many of the 2000 genes as possible in the modules for analysis.

2.4 Calculating Module Eigengene and Module Membership

Succeeding the selection of modules, the average gene expression levels of genes in each module, known as module eigengene (ME), and the degree of correlation between modules and genes, known as module membership (MM) were calculated. If MM is close to 0 that means the gene is not a part of the module, conversely if the MM is close to 1 or -1 then the gene is highly related to the module.

2.5 Survival Analysis

For each module, the expression data of the genes in the module were extracted. Using the Survival package (10) in R, Cox proportional hazards regression was used to model the relationship between gene expression and survival. The samples were also clustered into groups using KMeans clustering. The samples in each cluster were divided into high and low expression groups based on the average DEGs expression. The survminer library (11) was used to plot the Kaplan-Meier curve.

2.6 Prediction Model

Employing python scikit learn library (12) Recursive feature elimination (RFE) method was applied on differential expressed genes expression values and covariates like age, vital status, overall survival, race, sex, and smoking history to identify the highly significantly related factors for staging of cancer. The selected features are then made use of to develop a predictive model. The data was split into train and test sets in two ways. The first way involved manually using 75% of the data in each stage as the train sample. The second way used the stratifiedKfold method with 5 splits to test different train-test splits of the data. Various machine learning models are employed and the Area Under the Curve (AUC) is plotted for all the splits.

3 Results

3.1 Differential Expression

Overall, 17427 genes were retained from 60660 after the filtering on gene counts. Further differential expression analysis was performed over these 17427 genes and 90 DEGs were identified with FDR<0.05 [Figure 1]. Out of the 90 genes, 49 were upregulated and 41 were downregulated. The GO analysis showed that these genes belong to developmental and cellular processes. [Supplemental Fig 1].

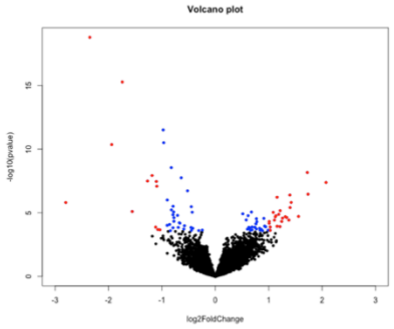


Figure 1: Volcano Plot depicting the DEGs.

3.2 Co-Expression Network

A soft threshold of 10 was selected as the power for adjacency matrix calculation since scale-free topology R^2 reached a value above range of 0.8 [Figure 2]. The hierarchical clustering tree with modules was plotted, and it was observed that deepsplit 4 resulted in the genes being clustered into 4 modules: blue, brown, grey, and turquoise [Figure 3]. Since grey module is the default color for unassigned genes, it was not used for analysis. There were 569, 435, and 779 genes mapped to blue, brown, and turquoise modules, respectively.

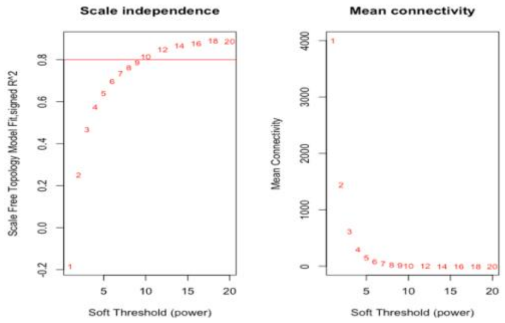


Figure 2: Scale Free Network Plot depicting the different powers.

Out of the three blue module is associated with ATP synthesis, brown with immune response and turquoise with RNA and chromatin regulation [Figure 4].

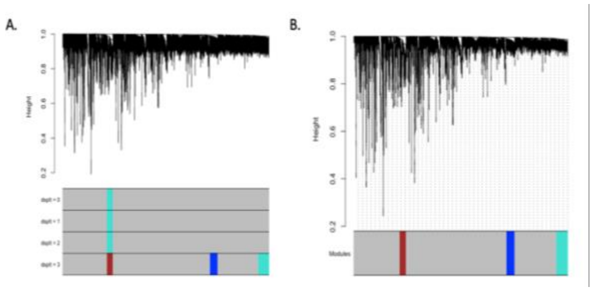


Figure 3: Hierarchical Clustering of Genes into Modules
A) All the deepsplits, B) Selected deepsplit

While using the top 2000 genes, a soft threshold of 6 was applied to create a network. The hierarchical clustering tree was generated, and it was observed that deepsplit 2 resulted in a small number of large modules compared to other splits. Therefore, it was selected, and out of the 6 modules obtained, turquoise module was found to be the largest among them [Supplemental Fig 2].

3.3 Survival Analysis

Survival analysis done using Cox proportional hazards regression model showed that there is no significant change in survival risk among different modules [Supplemental Fig 3]. This result was consistent across the modules of obtained using top 2000 genes as well [Supplemental Fig 4].

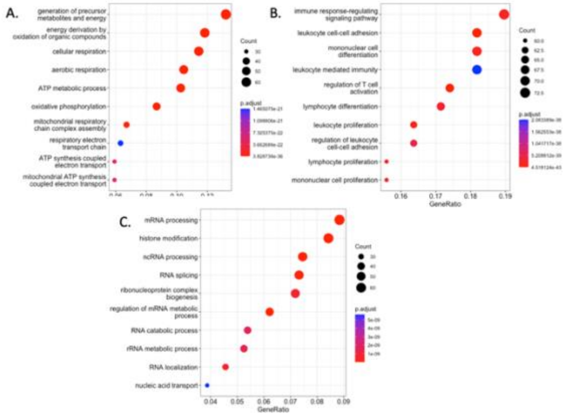
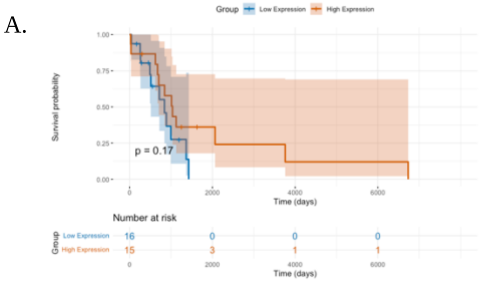


Figure 4: Gene Ontology Analysis. A) Module Blue, B) Module Brown, and C) Module Turquoise

However, in each sample cluster, the survival plots showed in cluster 4 and cluster 1 the group with low expression values of the DEGs had less survival time than high expression group [Figure 5]. Nevertheless, the P values are not found to be significant in any cluster suggesting that this association is not strong enough.



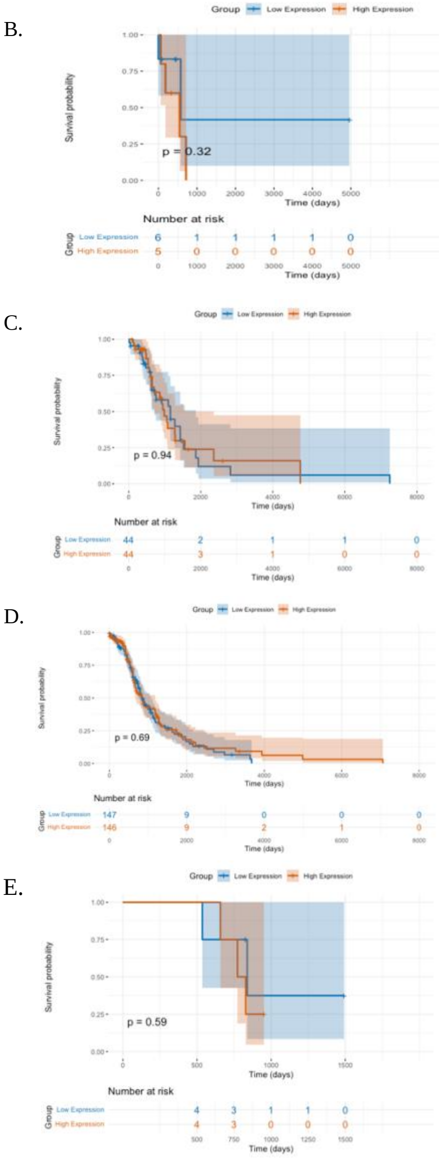


Figure 5: Kaplan-Meier curves of the average DEGs expression in Sample clusters A) Cluster 1 B) Cluster 2
C) Cluster 3 D) Cluster 4 E) Cluster 5

3.4 Prediction Model

Differentially gene expression values and covariates were merged, and Features that were ranked high by RFE were selected for further analysis. Forty-seven features were selected, and seven machine learning models were trained on these features. In all machine models, the manual train-test split showed better AUC values than stratified K-fold splits. The AdaBoost classifier and logistic regression had AUC of 0.68 and 0.63, respectively, using the manual split method. However, on average, the performance of the models to predict the stage of LUAD in different K folds remained low with logistic regression having the highest average [Figure 6].

4 DISCUSSION

In this study, 90 differentially expressed genes are identified that belong to developmental and cellular processes. A co-expression network was built on the tumor samples to cluster genes into modules based on the similarity of their expression. For each module, survival analysis was performed based on the module eigengene expression. Survival analysis was also performed on sample clusters to observe variations in different clusters. The average gene expression in each sample of the cluster was calculated, and how survival changes in each sample with respect to the expression was verified. Finally, using the 90 DEGs and covariates, feature selection and machine learning predictive models were built, with the target group being the stage of cancer. The data was divided into different test-train splits, and the performance of the 7 ML models was checked across each split.

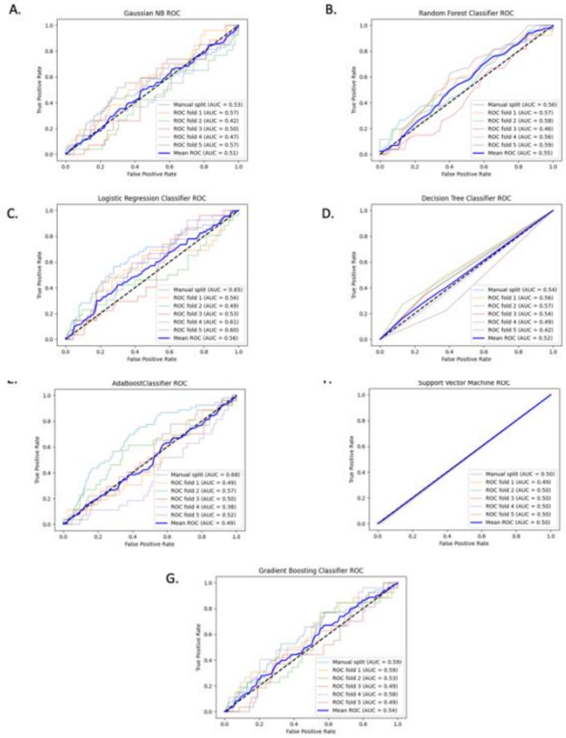


Figure 6: ROC curves depicting the AUC values for different test-train splits A) Gaussian Naïve Bayes B) Random Forest classifier C) Logistic Regression Classifier D) Decision Tree Classifier E) AdaBoost Classifier F) Support Vector Machine G) Gradient Boosting Classifier

The co-expression network analysis resulted in four modules of genes that were subsequently used to examine changes in survival time. However, further analysis of these modules showed no significant differences in survival among different gene clusters, as demonstrated in Supplemental Fig 3. The modules generated using the top 2000 genes also showed similar results, with none of the 6 modules having a significant change compared to the others, as seen in Supplemental Fig 4. Furthermore, analysis on sample groups revealed that the low expression level of DEGs was associated with low survival time. Nevertheless, none of the P-values indicated a significant correlation between gene expression levels and survival time.

In predictive analysis, the RFE model identified 47 features that were significantly related to the stage of cancer. The data was then split into training and test sets, with a manual split performed by dividing 75% of samples in each stage into the train and test sets separately. It was observed that in all the models, AdaBoost classifier and logistic regression performed better at demarcating the positive and negative classes. However, when using the k-fold cross-validation strategy, none of the classifiers had a mean AUC above 0.6. In both conditions, the logistic regression model seems to perform better. Nonetheless, the values of the AUC are still too low for the classifier to be considered the best. It is possible that performing hyperparameter tuning may improve the performance of the models, but at this point, it is still unclear whether the issue for the low AUC lies with the dataset or the model splits. Additional investigation is needed to determine the underlying cause of the low performance.

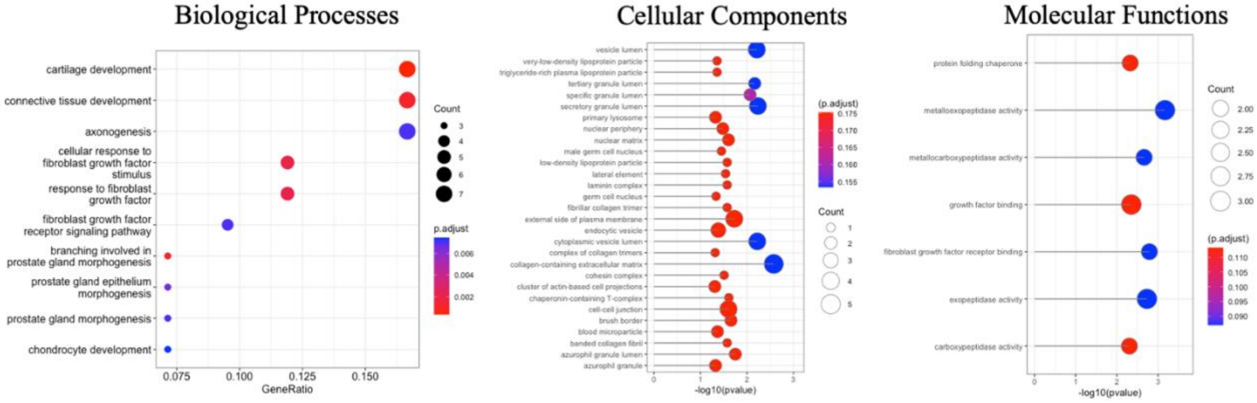
5. Conclusion

The study identified 90 differentially expressed genes related to developmental and cellular processes in LUAD. However, co-expression network and survival analysis did not reveal significant changes in survival based on gene clusters. Predictive analysis using machine learning models showed limited performance in predicting LUAD stages, with logistic regression being the best-performing model. Further investigation is required to enhance the model’s ability to predict by employing methods like hyperparameter tuning.

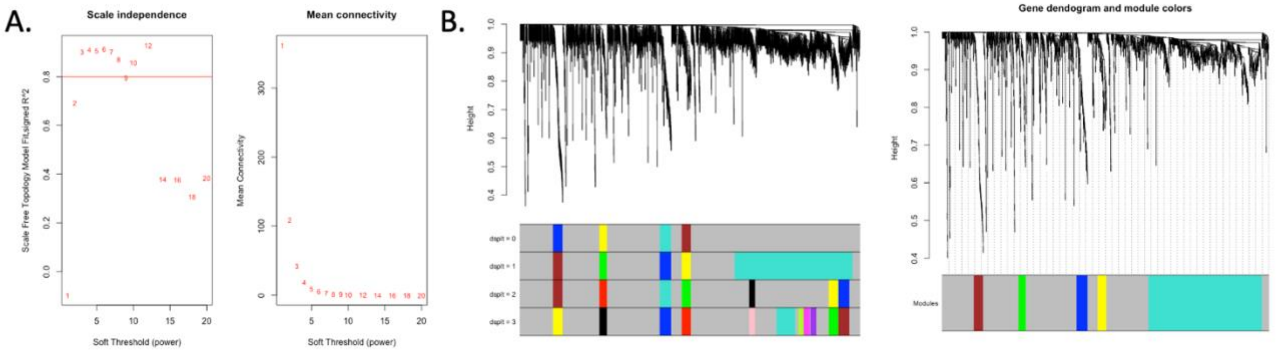
References

1. Seguin L, Durandy M, Feral CC. Lung adenocarcinoma tumor origin: a guide for personalized medicine. Cancers. 2022 Mar 30;14(7):1759.
2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2021 May;71(3):209-49.
3. Jaradeh M, Vigneswaran WT. Epidemiology of lung cancer and the gender differences in risk. Journal of Men's Health. 2022 Mar 2;18(3):73.
4. Cokkinides V, Albano J, Samuels A, Ward M, Thum J. American cancer society: Cancer facts and figures. Atlanta: American Cancer Society. 2005;2017.
5. Myers DJ, Wallen JM. Lung adenocarcinoma. InStatPearls [Internet] 2022 Jun 21. StatPearls Publishing.
6. Colaprico A, Silva TC, Olsen C, Garofano L, Garolini D, Cava C, Sabedot T, Malta T, Pagnotta SM, Castiglioni I, Ceccarelli M. Package ‘TCGAbiolinks’.
7. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology. 2014 Dec;15(12):1-21.
8. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. Omics: a journal of integrative biology. 2012 May 1;16(5):284-7.
9. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics. 2008 Dec;9(1):1-3.
10. Therneau T. A package for survival analysis in S. R package version. 2015 Mar 25;2(7).
11. Kassambara A, Kosinski M, Biecek P, Fabian S. Package ‘survminer’. Drawing Survival Curves using ‘ggplot2’(R package version 03 1). 2017 Mar 21.
12. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011 Nov 1;12:2825-30.

6. SUPPLIMENTAL FIGURES



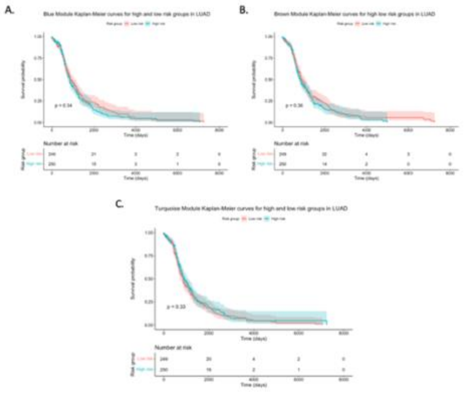
Supplemental Fig 1 : Figure depicts the Gene Ontology analysis on the differentially expressed genes. The biological processes show that most of the genes belong to developmental and cellular processes. The molecular functions show that most of the genes belong to growth factor activity regulation.



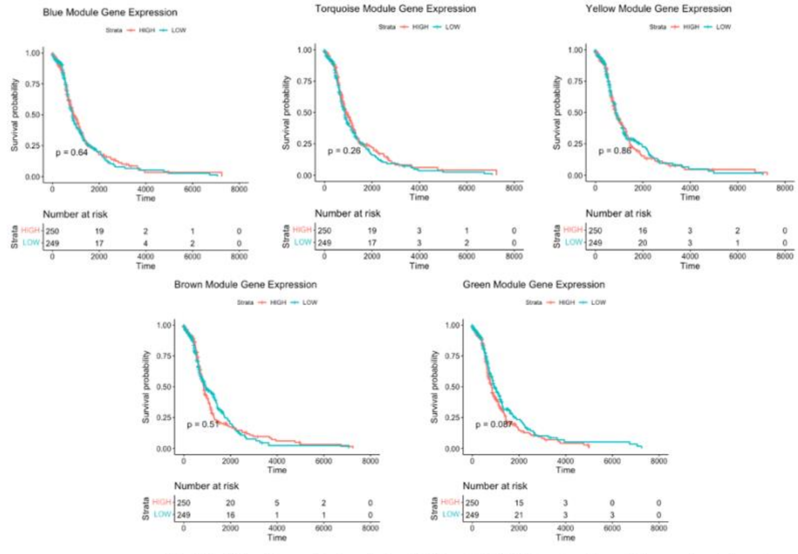
Supplemental Fig 2 : Figure depicts the WGCNA network construction on top 2000 differentially expressed genes.

Fig A : depicts Soft Power Threshold plot from which 6 was selected as the power since at that power the network starts to be scale free and the R^2 is above 0.9

Fig B : depicts four deepsplit modules of genes and out of the four 2nd deepsplit is selected as it contains small number of large modules compare to other deepsplits.



Supplemental Fig 4 :Plot depicting the survival analysis of different WGCNA network modules constructed on 17437 genes. It is observed that there is no change in survival times among clusters.



Supplemental Fig 3 :Plot depicting the survival analysis of different WGCNA network modules constructed on top 2000 DEGs. It is observed that there is no change in survival times among clusters.