

## **Gene prediction in heterogeneous cancer tissues and establishment of Least Absolute Shrinking and Selection Operator model of lung squamous cell carcinoma.**

Ateeq Muhammed Khaliq<sup>1</sup>, SharathChandra RG<sup>1</sup>, Meenakshi Rajamohan<sup>1</sup>

<sup>1</sup> Department of Studies and Research in Biotechnology and Centre for Bioinformation, Tumkur University, Tumkur, Karnataka, India

Corresponding Author:

Ateeq Muhammed Khaliq<sup>1</sup>

Department of Studies and Research in Biotechnology and Centre for Bioinformation, Tumkur University, Tumkur, Karnataka, India

Email address: atheet.khaliq@gmail.com

## **Abstract**

**Background.** This study is aimed to establish a Least Absolute Shrinking and Selection Operator (LASSO) model based on tumor heterogeneity to predict the best features of LUSC in various cancer subtypes.

**Methods.** The RNASeq data of 505 LUSC cancer samples were downloaded from the TCGA database. Subsequent to the identification of differentially expressed genes (DEGs), the samples were divided into two subtypes based on the consensus clustering method. The subtypes were estimated with the abundance of immune and non-immune stromal cell populations which infiltrated tissue. LASSO model was established to predict each subtype's best genes. Enrichment pathway analysis was then carried out. Finally, the validity of the LUSC model for identifying features was established by the survival analysis.

**Results.** 240 and 262 samples were clustered in Subtype-1 and Subtype-2 groups respectively. DEG analysis was performed on each subtype. A standard cutoff was applied and in total, 4586 genes were upregulated and 1495 were downregulated in case of subtype-1 and 5016 genes were upregulated and 3224 were downregulated in case of subtype-2. LASSO model was established to predict the best features from each subtypes, 49 and 34 most relevant genes were selected in subtype-1 and subtype-2. The abundance of tissue-infiltrates analysis distinguished the subtypes based on the expression pattern of immune infiltrates. Survival analysis showed that this model could effectively predict the best and distinct features in cancer subtypes.

**Discussion.** This study suggests that the unsupervised clustering and LASSO model-based feature selection can be effectively used to predict relevant genes which might play an important role in cancer diagnosis.

## Introduction

Lung cancer is among the most deadly cancers (Siegel, 2017). It shows the worst survival rate when compared with colon, breast, and pancreatic cancers combined. Lung cancer is classified as non-small-cell lung cancer (NSCLC) and small-cell lung cancer (SCLC). NSCLCs are generally subcategorized into adenocarcinoma (LUAD), squamous cell carcinoma (LUSC), and large cell carcinoma. LUSC and LUAD account for 15% and 85% of all lung cancer, respectively (Inamura, 2017). Lung cancer is a highly heterogeneous disease and identification of cancer subtypes is pivotal for clinicians. Genetic mutations, cancer microenvironment, immune, and therapeutic selection pressures all dynamically contribute to tumor heterogeneity. Heterogeneity may lead to cells with a differential molecular signature within single tumor tissue and in some cases, it may contribute to therapy resistance (Beca F, 2016; Bolck et al., 2019). Therefore, deciphering LUSC cancer heterogeneity will have a major impact in designing precision medicine strategy. Heterogeneous data suffers from a large number of covariates, and identification of variable selection is necessary to obtain more accurate predictions with a large number of covariates.

Over the past decades, many computer-aided diagnostic models have been used for predicting the risk of a variety of cancers, such as logistic regression, Cox proportional hazard model, Artificial neural networks, decision trees and Support vector machines (Bartfay, Mackillop & Pater, 2006; Ayer et al., 2010; Jiang & Ching, 2012; ZHU et al., 2013). Previous studies indicate standard stepwise selection approaches which are not best for regression models with a very large number of covariates (Houssami et al., 2004). Alternatively, least absolute shrinkage and selection operator (LASSO), has received much attention for identification and selection of best

variables. LASSO was first formulated by Robert Tibshirani in 1996 (Tibshirani, 1996). It is a powerful method that performs two main tasks: regularization and feature selection. LASSO estimates the regression coefficients by maximizing the log-likelihood function with the constraint that the sum of the absolute values of the regression coefficients,  $\sum_{j=1}^k \beta_j$ , is less than or equal to a positive constant  $s$ .

In this study, we downloaded the RNASeq data for LUSC cancer samples from The Cancer Genome Atlas (TCGA) database. We differentiated the samples based on clusters into two subtypes to study the tumor heterogeneity. Differentially expressed genes (DEGs) were identified between two subtypes and normal groups, followed by predicting relevant variables that are associated with the response variable using the LASSO model and validating the variables using survival analysis. We estimated the population abundance of tissue-infiltrating immune and stromal cell populations in each subtype to decipher the inflammatory, antigenic, and desmoplastic reactions occurring in cancer tissue. Our study provides new insight into tumor heterogeneity and its importance in sample classification for predicting of biomarkers of LUSC cancer.

## Materials & Methods

**Data source.** The RNASeq data of Lung Squamous cell cancer, including 505 LUSC samples, and 49 normal samples were downloaded from the TCGA database

(<https://portal.gdc.cancer.gov/>) in May 2019. All the raw, preprocessed data and supporting files can be accessed at [https://bitbucket.org/lusc\\_data/supporting\\_data/src/master/](https://bitbucket.org/lusc_data/supporting_data/src/master/).

**Data preprocessing and grouping.** Based on the clinical data, the LUSC cancer samples downloaded from TCGA database were divided into two sets, the first set was divided into 114 low-risk samples and 390 high-risk sample groups according to the AJCC Cancer Staging (<https://cancerstaging.org/>). The second set (set2) was divided into 505 Primary solid Tumor samples and 49 Solid Tissue Normal samples. We calculated a variance stabilizing transformation (VST) from the data and transformed the counts yielding a matrix of values approximately homoskedastic.

**Molecular subtyping analysis.** Feature dimension reduction was needed to remove irrelevant features and to reduce noises, we used median absolute deviation (MAD) method and the features with  $MAD > 0.5$  were selected from set 2 groups. Consensus clustering (CC) (Monti et al., 2003) was used for the identification of subtypes on set 2 group. Silhouette width (Rousseeuw, 1987) was used to validate sample clustering to its identified subtype compared to other subtypes.

**Differential gene expression analysis.** Differential gene expression was assessed by using the DESeq2 package (Love MI, Huber W, 2014) (Version 1.24.0, <https://bioconductor.org/packages/release/bioc/html/DESeq2.html>) on set1 (High-Risk samples Vs. Low-risk samples) and set2 (Subtype-1 vs. Normal and Subtype-2 vs. Normal). Log2 fold change  $> 2$  and P-value  $< 0.05$  were used as the cut-off values to identify the DEGs.

**Construction of the LASSO Model.** Glmnet Package (Friedman J, Hastie T, 2010) (Version 2.0-18, <https://cran.r-project.org/web/packages/glmnet/index.html>) was used to fit a generalized

linear model via penalized maximum likelihood, LASSO model was established (Least Absolute Shrinkage and Selection Operator) on the DEGs from individual Subtype-1 and Subtype-2 cancer samples. We built a single pass (single fold) lasso-penalized model and performed 10-fold cross-validation to identify the best predictor.

**Survival Analysis.** To find clinically or biologically meaningful biomarkers Kaplan-Meier survival curves (E. L. Kaplan & Paul Meier, 1958) were generated by selecting the best predictors from individual subtypes. Kaplan-Meier curves were generated using the TRGAted (Nicholas Borcharding , Nicholas L. Bormann, Andrew P. Voigt, 2018)(<https://github.com/ncborcherding/TRGAted>) package implemented in R.

### **Quantification of the absolute abundance of eight immune and two stromal cell**

**populations.** We estimated the abundance of tissue-infiltrating immune and non-immune stromal cell populations in Subtype-1 and Subtype-2 samples. MCP-counter (Etienne Becht, Nicolas A. Giraldo, Laetitia Lacroix, Bénédicte Buttard, Nabila Elarouci, Florent Petitprez, Janick Selves, Pierre Laurent-Puig, Catherine Sautès-Fridman, 2016) (<https://github.com/ebecht/MCPcounter>) Package was used to estimate the Microenvironment Cell Populations. VST normalized gene expression matrix was used for the estimation of an immune and stromal cell population.

**Gene classification and enrichment analyses.** clusterProfiler (Yu G, Wang L, Han Y, 2012) (Version 3.12.0, <http://bioconductor.org/packages/release/bioc/html/clusterProfiler.html>) was used to annotate the DEGs from Subtype-1 and Subtype-2 groups to biological processes,

molecular functions, and cellular components in a directed acyclic graph structure with a q-value cutoff of 0.2, Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa M, Goto S, Furumichi M, Tanabe M, 2010) was utilized to annotate genes to pathways, and Disease Ontologies.

## **Results**

### **Identification of DEGs in High-risk LUSC tumors.**

The genes with p-value cutoff  $< 0.05$  and log<sub>2</sub> fold change  $> 2$  were considered to be differentially expressed. A total of 22 genes were differentially expressed between high risk and low-risk samples, which includes 4 downregulated and 18 upregulated genes. *Figure 1* displays the heat map of the risk-related DEGs. Which is suggestive of similar gene expression pattern in both groups, which makes it difficult to classify the samples on the gene expression pattern. PC analysis shows the homogeneity of the data between the High and Low-risk group (*Fig 2*).

### **Molecular cancer subtype identification in LUSC Tumor samples and validation of clusters.**

We used an unsupervised clustering method Consensus clustering (CC) (Monti et al., 2003). CC method is most widely used for subtype discovery in high dimensional datasets. We used settings of the agglomerative hierarchical clustering algorithm using Pearson correlation distance. Two distinct clusters were discovered in our datasets, 240 and 262 samples were clustered in Subtype-1 and Subtype-2 groups respectively. (*Table S1 and S2*) We have validated consistency within clusters of data using Silhouette Plot (Rousseeuw, 1987). The Average Silhouette width for our generated clusters is 0.68 (*Fig 3, 4 and 5*).

### **Identification of DEGs in Subtype-1 and Subtype-2 LUSC Tumors.**

We compared the subtype-1 and subtype-2 with the normal samples and based on the p-value cutoff  $< 0.05$  and  $\log_2$  fold change  $> 2$  we identified significant DEGs. 4586 genes were upregulated and 1495 were downregulated in case of subtype-1 (*Fig 6*) and 5016 genes were upregulated and 3224 were downregulated in case of subtype-2 (*Fig 7*) shows differential expression pattern in subtype-1 and subtype-2. The DEGs in both subtypes were used for building LASSO predictive model and for the identification of best predictor genes.

### **LASSO model for identification of best predictive genes.**

LASSO (Least Absolute Shrinkage and Selection Operator) was first formulated by Robert Tibshirani in 1996 (Tibshirani, 1996). It is a powerful method that performs two main tasks: regularization and feature selection. RNASeq datasets are high dimensional datasets, with smaller sample size and a large number of features also called small-n-large-p datasets ( $p \gg n$ ). High dimensional data will be sparse and only a few features affect the response variable and LASSO is known to identify the best features that affect the response variable. We deal with a  $p \gg n$  situation for feature selection in our Subtype-1 and Subtype-2 datasets, thus probably not all DEGs are relevant for the identification of features which affect the response variable. The purpose of our analysis is to identify the feature selection task and underline which genes are more relevant to predict and to classify them as biomarkers, to do so we have used the LASSO method.

The result shows the trends of the 49 and 34 most relevant features selected by our model in subtype-1 and subtype-2 LUSC cancer respectively (*Fig S1 and S2*). The next step would be to



find the most appropriate values for  $\lambda$  for our LASSO model. We analyzed the  $\lambda$  value using the 10 fold cross-validation (*Fig S3 and S4*), between  $\lambda$  min that gives minimum mean cross-validated error or  $\lambda_{1se}$  that gives a model within one standard error of the minimum. Using this analysis we obtained the most relevant genes which are unique to subtype-1 and subtype-2 in the detection of a LUSC cancer. A list of best-predicted genes available for each cancer subtypes is shown in *Table 1* and *Table 2*.

### **Analysis of the microenvironment of Subtype-1 and Subtype-2 LUSC cancer.**

The abundance of tissue-infiltrating immune and non-immune stromal cell populations is highly informative. It has been shown that the extent of infiltrating immune cells is associated with disease prognosis (Becht et al., 2016). T-cell infiltrates, endothelial cells and fibroblasts are associated with a favorable outcome and also poor prognosis in some cancer types (Pagès F, Berger A, Camus M, Sanchez-Cabo F, Costes A, Molidor R, Mlecnik B, Kirilovsky A, Nilsson M, Damotte D, Meatchi T, Bruneval P, Cugnenc P, Trajanoski Z, Fridman W, 2005; Galon J, Costes A, Sanchez-Cabo F, Kirilovsky A, Mlecnik B, Lagorce-Pagès C, Tosolini M, Camus M, Berger A, Wind P, Zinzindohoué F, Bruneval P, Cugnenc P, Trajanoski Z, Fridman W, 2006; Fridman WH, Pagès F, Sautès-Fridman C, 2012; Giraldo NA, Becht E, Pagès F, Skliris G, Verkarre V, Vano Y, Mejean A, Saint-Aubert N, Lacroix L, Natario I, Lupo A, Alifano M, Damotte D, Cazes A, Triebel F, Freeman GJ, Dieu-Nosjean M, Oudard S, Fridman WH, 2015). To understand the immunological microenvironment in our expression subset-1 and subset-2 we used MCP-counter method as described by Becht et al (Etienne Becht, Nicolas A. Giraldo, Laetitia Lacroix, Bénédicte Buttard, Nabila Elarouci, Florent Petitprez, Janick Selves, Pierre Laurent-Puig, Catherine Sautès-Fridman, 2016). The estimations consist of single sample scores

which are computed on each sample independently in two subtypes. The heatmap shown in *Figure 8* clearly distinguish our subtype-1 and subtype-2 into two different categories based on tissue-infiltrating immune and non-immune stromal cell populations. Subtype-1 shows clear increase in CD8 T cells, Cytotoxic lymphocytes and Natural killer cells and Subtype-2 shows decreased levels of T-cells, macrophages, B cells, and natural killer (NK) cells, as well as endothelial cells and fibroblasts. Our study clearly distinguishes LUSC subtypes based on their inflammatory and stromal profiles and Subtype-1 LUSC samples show increased expression of immunological markers than Subtype 2 LUSC samples.

### **Disease pathway analysis.**

KEGG pathway analysis for Subtype-1 and Subtype-2 revealed many significant cancer pathways, including genes involved in Focal adhesion, mTOR signaling pathway, Axon guidance pathway, Cellular senescence pathway, ErbB signaling pathway, Longevity regulating pathway, HIF-1 signaling pathway, AMPK signaling pathway, Pancreatic cancer, Chronic myeloid leukemia, Colorectal cancer, TGFbeta signaling pathway etc. (*Table 3 and Table 4*). Genes such as EIF4EBP1, FOXA2, PECAM1, TGFBR2, TNNC1, ACSM2B and ABCA8 picked up by our model plays a pivotal role in cancer regulatory pathways (*Fig 9 and 10*). Both Subtype-1 and Subtype-2 groups showed distinct biological processes and cellular components after GO Enrichment Analysis of the gene sets from Subtype-1 and Subtype-2 groups. (*Fig. S5, S6, S7 and S8 and Table S3, S4, S5 and S6*).

### **Validation with survival analysis.**

The LUSC Tumor samples were classified into two subtypes based on the consensus clustering method. Using the survival data of these samples, survival analysis for the most predictive genes identified by our model in Subtype-1 and Subtype-2 groups was conducted. The genes predicted by our LASSO model accurately predicted the outcome of a patient's survival using gene expression data. Genes such as GAL, TFAP2A, AFF3, TNNC1, TGBR2, HELT, and SFTA1P yielded accurate predictions for the risk of LUSC cancer and can be used in cancer prediction. Survival plots and its p-value is shown in (*Fig 11 and 12, Supplemental Figure S9*).

## **Discussion**

In this study, we developed a LASSO based model for accurate feature selection in LUSC cancer. Our model removed variables that are redundant and removed features which do not add any valuable information in disease prediction. Analysis using the survival data for the predicted genes showed that the model could effectively predict genes responsible for disease prognosis in high dimensional datasets. Deciphering cancer heterogeneity is very critical in understanding cancer dynamics and also for the development of personalized cancer treatment (R. Rosenthal, N. McGranahan, J. Herrero, 2017; Dagogo-Jack & Shaw, 2018) We used Consensus clustering method to determine the number of clusters in our samples, and we clustered the samples into two groups which produced optimal silhouette width for the determined clusters. Differential gene expression analysis showed distinct expression patterns in Subtype-1 and Subtype-2.

The number of differentially expressed genes were very high and in these situations, it is difficult to predict the relevant variables. LASSO model was built around 6081 and 8240 DE genes in Subtype-1 and Subtype-2 respectively. Not all the expressed genes were relevant, our model predicted the most relevant genes which were involved in disease progression. Decreased expression of AFF3, TNNC1, TGFBR2, FFAR4 and HELT in Subtype-1 and GGTL3,

GUCA2A, HBM, SFTA1P and SYNE1 in Subtype-2 showed worse overall survival in LUSC cancer samples. Whereas increased expression of genes such as PPIAP45, CAGE1, TFAP2A, CENPF in Subtype-1 decreased overall survival in LUSC cancer samples.

Long intervening noncoding RNAs (lncRNAs) are known to be key regulators of numerous biological processes, and substantial evidence supports that lncRNA expression plays a significant role in tumorigenesis and tumor progression (Ming-Chun Jiang, Jiao-Jiao Ni, Wen-Yu Cui, Bo-Ya Wang, 2019). Increased expression of LINC01977, LINC01572 in Subtype-1 samples correlates with worse survival in LUSC cancer subtypes. Whereas, decreased expression of LINC02058 in Subtype-1 and LINC00670 in Subtype-2 showed worse survival in LUSC samples. The LASSO method predicts the most relevant and distinct genes from Subtype-1 and Subtype-2 samples which might be important factor in cancer diagnosis.

The best predictors for subtype 1 and subtype 2 from the LASSO model were found to be involved in several regulatory pathways. The genes such as TGFBR2, EIF4EBP1, and ROR1 which are predicted only in case of Subtype-1 are found to be involved in several cell cycle and growth regulatory pathways and thereby having a strong correlation with cancer. The gene gp9 plays an important role in ECM-receptor interactions, which is critical in disease progression and malignant cell behavior (Walker, Mojares & Del Río Hernández, 2018). The neuroactive ligand-receptor interaction signaling pathway is a collection of receptors and ligands on the plasma membrane that are associated with intracellular and extracellular signaling pathways. It is found to be associated with prostate cancer, bladder cancer, and renal cell carcinoma (He et al., 2018). In our study, the gene RXFP2 that is predicted only in Subtype-2 is found to be involved in

neuroactive ligand-receptor interaction. RXFP2 is also found to be involved in Relaxin signaling which induces cell invasion and is reported in several cancers (Fue et al., 2018). The modulators of ABC transporters is reported to have the potential to augment the efficacy of anticancer drugs (Chen et al., 2012). ABCA8 is one such gene and it was predicted only in Subtype-2.

Our model identified cancer/testis antigen gene CAGE-1 which is overexpressed in Subtype-1 and might act as a plasma biomarker for lung cancer early detection. Previous studies showed that CAGE-1 provides an important addition to the armamentarium the clinician to aid early detection of lung cancer in high-risk individuals (Park et al., 2003; Kunze, Wendt & Schlott, 2006; Parmigiani et al., 2006; Kunze & Schlott, 2007; Chapman et al., 2011; Kim et al., 2013). GUCA2A was down-regulated in Subtype-2 samples, many studies on GUCA2A indicates its role as a biomarker in diagnosing cancer. Aberrantly expressed GUCA2A can be a candidate marker of poor prognosis in patients with LUSC and Colorectal cancers, which may be a therapeutic target for precision medicine (Kulaksiz H, Rehberg E, Stremmel W, 2002; Chen YB, Zhu YP, Feng HY, Liu Y, Qian J, Fan YT, 2009; Hui Zhang, Yuanyuan Du, Zhuo Wang, Rui Lou, Jianzhong Wu, 2019) Under expression of TGFBR2 in Subtype-1 samples is associated with poor prognosis, and TGFBR2 is also associated with poor prognosis in cervical cancer (Yang et al., 2017; Yokouchi et al., 2017). CENP-F, a cell cycle-regulated centromere protein, has been shown to affect numerous tumorigenic processes, increased expression of CENP-F in subtype-1 correlates with poor survival. Previous studies demonstrate that CENP-F may serve as a valuable molecular marker for predicting the prognosis of esophageal squamous cell carcinoma patients and nasopharyngeal carcinoma progression (J.-Y. et al., 2010; Chen et al., 2011; Mi et al., 2013; Yang et al., 2017). Downregulation of SFTA1P in Subtype-2 correlates with poor survival, previous studies suggest SFTA1P regulates both oncogene and tumor suppressor genes

during the carcinogenesis of lung squamous cell carcinoma(Zhao, Luo & Jiao, 2014; Huang et al., 2017; Zhang et al., 2017; Ma et al., 2018) which can be used as a prognostic biomarker. Furthermore, Consensus clustering and LASSO helps us to choose a model with the most relevant features. Consistent with this finding, the clustered samples into two different subtypes showed distinct features, highlighting the better sample grouping and risk assessment. Moreover, the results of survival analysis validates that the survival time of the predicted genes correlates with gene expression pattern, which is recognizably different in both the Subtypes, indicating that this model could effectively distinguish the samples with different expression pattern by overcoming the feature selection problem and was accurate for predicting the risk of LUSC cancer.

## **Conclusions**

In conclusion, this study suggests that the unsupervised method such as Consensus clustering and LASSO model-based feature selection could be used to evaluate prediction and prognosis of LUSC cancer. With this model, we can identify the prognostic biomarkers of LUSC cancer, and the model-predicted genes would be helpful for clinicians in the management of cancer patients.

## **Acknowledgements**

We wish to acknowledge Sangram Keshari Sahu from Bioinformatics lab of the IISER Mohali, India for his respective technical and scientific expertise. The efforts of Sangram in Gene Expression analysis and the expression project for Oncology are greatly acknowledged.

## References

- Ayer T, Chhatwal J, Alagoz O, Kahn CE, Woods RW, Burnside ES. 2010. Comparison of Logistic Regression and Artificial Neural Network Models in Breast Cancer Risk Estimation. *RadioGraphics*. DOI: 10.1148/rg.301095057.
- Bartfay E, Mackillop WJ, Pater JL. 2006. Comparing the predictive value of neural network models to logistic regression models on the risk of death for small-cell lung cancer patients. *European Journal of Cancer Care*. DOI: 10.1111/j.1365-2354.2005.00638.x.
- Beca F PK. 2016. Intratumor Heterogeneity in Breast Cancer. *Adv Exp Med Biol* 882:169–89. DOI: 10.1007/978-3-319-22909-6\_7.
- Becht E, Giraldo NA, Germain C, de Reyniès A, Laurent-Puig P, Zucman-Rossi J, Dieu-Nosjean MC, Sautès-Fridman C, Fridman WH. 2016. Immune Contexture, Immunoscore, and Malignant Cell Molecular Subgroups for Prognostic and Theranostic Classifications of Cancers. In: *Advances in Immunology*. DOI: 10.1016/bs.ai.2015.12.002.
- Bolck HA, Corrà C, Kahraman A, von Teichman A, Toussaint NC, Kuipers J, Chiovaro F, Koelzer VH, Pauli C, Moritz W, Bode PK, Rechsteiner M, Beerenwinkel N, Schraml P, Moch H. 2019. Tracing Clonal Dynamics Reveals that Two- and Three-dimensional Patient-derived Cell Models Capture Tumor Heterogeneity of Clear Cell Renal Cell Carcinoma. *European Urology Focus*. DOI: 10.1016/j.euf.2019.06.009.
- Chapman CJ, Thorpe AJ, Murray A, Parsy-Kowalska CB, Allen J, Stafford KM, Chauhan AS, Kite TA, Maddison P, Robertson JFR. 2011. Immunobiomarkers in small cell lung cancer: Potential early cancer signals. *Clinical Cancer Research*. DOI: 10.1158/1078-0432.CCR-10-1363.
- Chen W Bin, Cheng X Bin, Ding W, Wang YJ, Chen D, Wang JH, Fei RS. 2011. Centromere

- protein F and survivin are associated with high risk and a poor prognosis in colorectal gastrointestinal stromal tumours. *Journal of Clinical Pathology*. DOI: 10.1136/jcp.2011.089631.
- Chen ZS, Sun YL, Patel A, Kumar P. 2012. Role of ABC transporters in cancer chemotherapy. *Chinese Journal of Cancer* 31:51–57. DOI: 10.5732/cjc.011.10.
- Chen YB, Zhu YP, Feng HY, Liu Y, Qian J, Fan YT LD. 2009. Differential expression of guanylin in colorectal cancer. *chinese journal of gastrointistinal surgery* 5:515–7.
- Dagogo-Jack I, Shaw AT. 2018. Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology*. DOI: 10.1038/nrclinonc.2017.166.
- E. L. Kaplan & Paul Meier. 1958. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association* 53.
- Etienne Becht, Nicolas A. Giraldo, Laetitia Lacroix, Bénédicte Buttard, Nabila Elarouci, Florent Petitprez, Janick Selves, Pierre Laurent-Puig, Catherine Sautès-Fridman WHF& A de R. 2016. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biology* volume 17:249. DOI: 10.1186/s13059-016-1070-5.
- Fridman WH, Pagès F, Sautès-Fridman C GJ. 2012. The immune contexture in human tumours: impact on clinical outcome. *Nat Rev Cancer* 12:298–306. DOI: 10.1038/nrc3245.
- Friedman J, Hastie T TR. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* 33:1–22. DOI: 2010;33(1):1-22.
- Fue M, Miki Y, Takagi K, Hashimoto C, Yaegashi N, Suzuki T, Ito K. 2018. Relaxin 2/RXFP1 signaling induces cell invasion via the  $\beta$ -catenin pathway in endometrial cancer. *International Journal of Molecular Sciences* 19. DOI: 10.3390/ijms19082438.



- Galon J, Costes A, Sanchez-Cabo F, Kirilovsky A, Mlecnik B, Lagorce-Pagès C, Tosolini M, Camus M, Berger A, Wind P, Zinzindohoué F, Bruneval P, Cugnenc P, Trajanoski Z, Fridman W PF. 2006. Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science* 313:1960–4. DOI: 10.1126/science.1129139.
- Giraldo NA, Becht E, Pagès F, Skliris G, Verkarre V, Vano Y, Mejean A, Saint-Aubert N, Lacroix L, Natario I, Lupo A, Alifano M, Damotte D, Cazes A, Triebel F, Freeman GJ, Dieu-Nosjean M, Oudard S, Fridman WH S-FC. 2015. Orchestration and prognostic significance of immune checkpoints in the microenvironment of primary and metastatic renal cell cancer. *Clin Cancer Res* 21:3031–4. DOI: 10.1158/1078-0432.CCR-14-2926.
- He Z, Tang F, Lu Z, Huang Y, Lei H, Li Z, Zeng G. 2018. Analysis of differentially expressed genes, clinical value and biological pathways in prostate cancer. *American Journal of Translational Research* 10:1444–1456.
- Houssami N, Irwig L, Simpson JM, McKessar M, Blome S, Noakes J. 2004. The influence of clinical information on the accuracy of diagnostic mammography. *Breast Cancer Research and Treatment*. DOI: 10.1023/B:BREA.0000025416.66632.84.
- Huang GQ, Ke ZP, Hu HB, Gu B. 2017. Co-expression network analysis of long noncoding RNAs (lncRNAs) and cancer genes reveals SFTA1P and CASC2 abnormalities in lung squamous cell carcinoma. *Cancer Biology and Therapy*. DOI: 10.1080/15384047.2017.1281494.
- Hui Zhang, Yuanyuan Du, Zhuo Wang, Rui Lou, Jianzhong Wu and JF. 2019. Integrated Analysis of Oncogenic Networks in Colorectal Cancer Identifies GUCA2A as a Molecular Marker. *Biochemistry Research International*. DOI: 10.1155/2019/6469420.
- Inamura K. 2017. Lung Cancer: Understanding Its Molecular Pathology and the 2015 WHO

- Classification. *Frontiers in Oncology*. DOI: 10.3389/fonc.2017.00193.
- J.-Y. C, L. L, S. C, X. Z, Y. M, Z. L, M. L, H. Z, C. Q, J. S, L. F, Y. X, M. Z. 2010. Prognostic significance and therapeutic implications of centromere protein F expression in human nasopharyngeal carcinoma. *Molecular Cancer*. DOI: 10.1186/1476-4598-9-237.
- Jiang H, Ching WK. 2012. Correlation kernels for support vector machines classification with applications in cancer data. *Computational and Mathematical Methods in Medicine*. DOI: 10.1155/2012/205025.
- Kanehisa M, Goto S, Furumichi M, Tanabe M HM. 2010. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res*. DOI: 10.1093/nar/gkp896.
- Kim Y, Park D, Kim H, Choi M, Lee H, Lee YS, Choe J, Kim YM, Jeoung D. 2013. MiR-200b and cancer/testis antigen cage form a feedback loop to regulate the invasion and tumorigenic and angiogenic responses of a cancer cell line to microtubule-Targeting drugs. *Journal of Biological Chemistry* 288:36502–36518. DOI: 10.1074/jbc.M113.502047.
- Kulaksiz H, Rehberg E, Stremmel W CY. 2002. Guanylin and functional coupling proteins in the human salivary glands and gland tumors: expression, cellular localization, and target membrane domains. *Am J Pathol* 161:655–64. DOI: 10.1016/S0002-9440(10)64221-6.
- Kunze E, Schlott T. 2007. High frequency of promoter methylation of the 14-3-3  $\sigma$  and CAGE-1 genes, but lack of hypermethylation of the caveolin-1 gene, in primary adenocarcinomas and signet ring cell carcinomas of the urinary bladder. *International Journal of Molecular Medicine*.
- Kunze E, Wendt M, Schlott T. 2006. Promoter hypermethylation of the 14-3-3  $\sigma$ , SYK and CAGE-1 genes is related to the various phenotypes of urinary bladder carcinomas and

associated with progression of transitional cell carcinomas. *International Journal of Molecular Medicine*.

Love MI, Huber W AS. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. DOI: 2014;15(12):550.

Ma H, Ma T, Chen M, Zou Z, Zhang Z. 2018. The pseudogene-derived long non-coding RNA SFTA1P suppresses cell proliferation, migration, and invasion in gastric cancer. *Bioscience Reports* 38:BSR20171193. DOI: 10.1042/bsr20171193.

Mi YJ, Gao J, Xie JD, Cao JY, Cui SX, Gao HJ, Yao SP, Liu T, Zhang YY, Guo CH, Qiu GQ, Chen YQ. 2013. Prognostic relevance and therapeutic implications of centromere protein F expression in patients with esophageal squamous cell carcinoma. *Diseases of the Esophagus*. DOI: 10.1111/dote.12002.

Ming-Chun Jiang, Jiao-Jiao Ni, Wen-Yu Cui, Bo-Ya Wang and WZ. 2019. Emerging roles of lncRNA in cancer and therapeutic opportunities. *Am J Cancer Res* 9:1354–1366. DOI: 9(7):1354-1366.

Monti S, Tamayo P, Mesirov J, Golub T. 2003. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*. DOI: 10.1023/A:1023949509487.

Nicholas Borcharding , Nicholas L. Bormann, Andrew P. Voigt WZ. 2018. TRGAted: A web tool for survival analysis using protein data in the Cancer Genome Atlas. *F1000Research* 7. DOI: 10.12688/f1000research.15789.2. eCollection 2018.

Pagès F, Berger A, Camus M, Sanchez-Cabo F, Costes A, Molitor R, Mlecnik B, Kirilovsky A, Nilsson M, Damotte D, Meatchi T, Bruneval P, Cugnenc P, Trajanoski Z, Fridman W GJ. 2005. Effector memory T cells, early metastasis, and survival in colorectal cancer. *N Engl J*

*Med* 353:2654–66. DOI: 10.1056/NEJMoa051424.

Park S, Lim Y, Lee D, Cho B, Bang YJ, Sung S, Kim HY, Kim DK, Lee YS, Song YW, Jeoung

D II. 2003. Identification and characterization of a novel cancer/testis antigen gene CAGE-

1. *Biochimica et Biophysica Acta - Gene Structure and Expression* 1625:173–182. DOI:

10.1016/S0167-4781(02)00620-6.

Parmigiani RB, Bettoni F, Vibranovski MD, Lopes MH, Martins WK, Cunha IW, Soares FA,

Simpson AJG, de Souza SJ, Camargo AA. 2006. Characterization of a cancer/testis (CT)

antigen gene family capable of eliciting humoral response in cancer patients. *Proceedings of*

*the National Academy of Sciences* 103:18066–18071. DOI: 10.1073/pnas.0608853103.

R. Rosenthal, N. McGranahan, J. Herrero and CS. 2017. Deciphering genetic intratumor

heterogeneity and its impact on cancer evolution. *Annual Review of Cancer Biology* 1:223–

240. DOI: 10.1158/2159-8290.CD-15-0412.

Rousseeuw PJ. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster

analysis. *Journal of Computational and Applied Mathematics*. DOI: 10.1016/0377-

0427(87)90125-7.

Siegel R. 2017. Cáncer Statistics. *Ca Cáncer Journal*. DOI: 10.3322/caac.21387.

Tibshirani R. 1996. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal*

*Statistical Society: Series B (Methodological)*. DOI: 10.1111/j.2517-6161.1996.tb02080.x.

Walker C, Mojares E, Del Río Hernández A. 2018. Role of extracellular matrix in development

and cancer progression. *International Journal of Molecular Sciences*. DOI:

10.3390/ijms19103028.

Yang H, Zhang H, Zhong Y, Wang Q, Yang L, Kang H, Gao X, Yu H, Xie C, Zhou F, Zhou Y.

2017. Concomitant underexpression of TGFBR2 and overexpression of hTERT are

associated with poor prognosis in cervical cancer. *Scientific Reports*. DOI: 10.1038/srep41670.

Yokouchi H, Nishihara H, Harada T, Ishida T, Yamazaki S, Kikuchi H, Oizumi S, Uramoto H, Tanaka F, Harada M, Akie K, Sugaya F, Fujita Y, Takamura K, Kojima T, Higuchi M, Honjo O, Minami Y, Watanabe N, Goto A, Suzuki H, Dosaka-Akita H, Isobe H, Nishimura M, Munakata M. 2017. Immunohistochemical profiling of receptor tyrosine kinases, MED12, and TGF- $\beta$ 2/RII of surgically resected small cell lung cancer, and the potential of c-kit as a prognostic marker. *Oncotarget*. DOI: 10.18632/oncotarget.14410.

Yu G, Wang L, Han Y HQ. 2012. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology* 16:284–287. DOI: 10.1089/omi.2011.0118.

Zhang H, Xiong Y, Xia R, Wei C, Shi X, Nie F. 2017. The pseudogene-derived long noncoding RNA SFTA1P is down-regulated and suppresses cell migration and invasion in lung adenocarcinoma. *Tumor Biology*. DOI: 10.1177/1010428317691418.

Zhao W, Luo J, Jiao S. 2014. Comprehensive characterization of cancer subtype associated long non-coding RNAs and their clinical implications. *Scientific Reports*. DOI: 10.1038/srep06591.

ZHU L, LUO W, SU M, WEI H, WEI J, ZHANG X, ZOU C. 2013. Comparison between artificial neural network and Cox regression model in predicting the survival rate of gastric cancer patients. *Biomedical Reports*. DOI: 10.3892/br.2013.140.

**Figure 1**

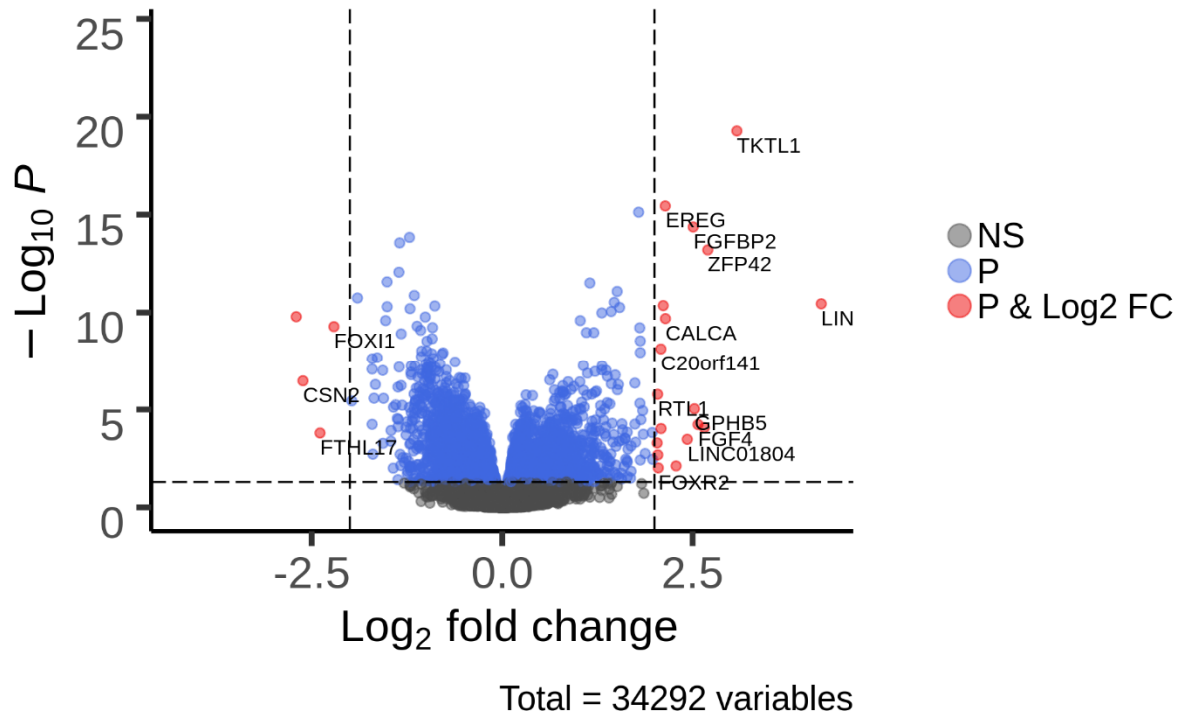
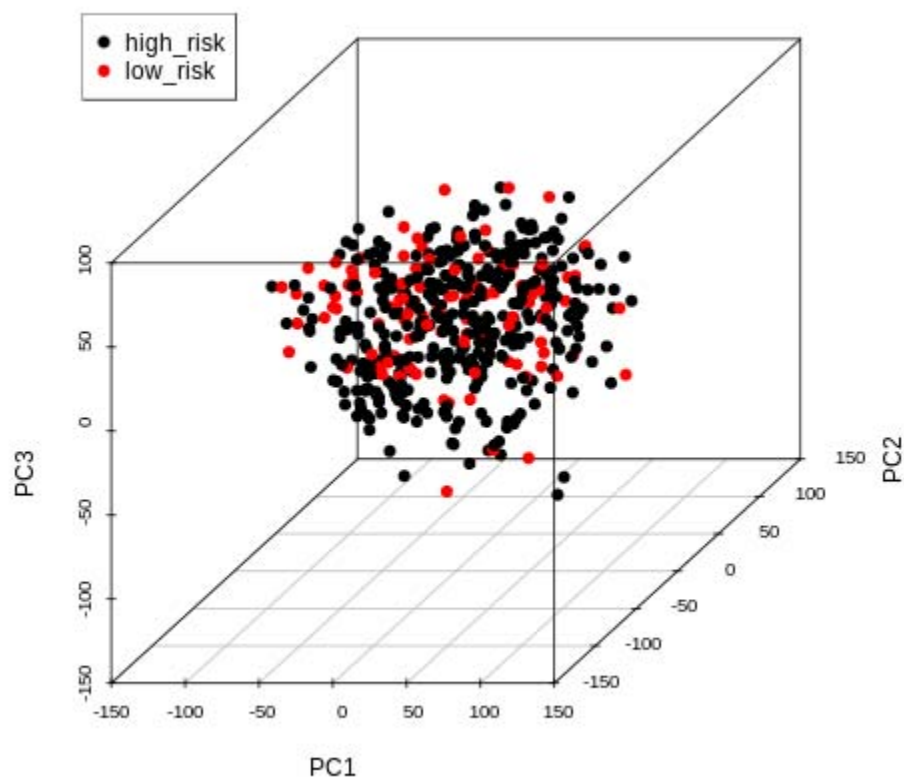


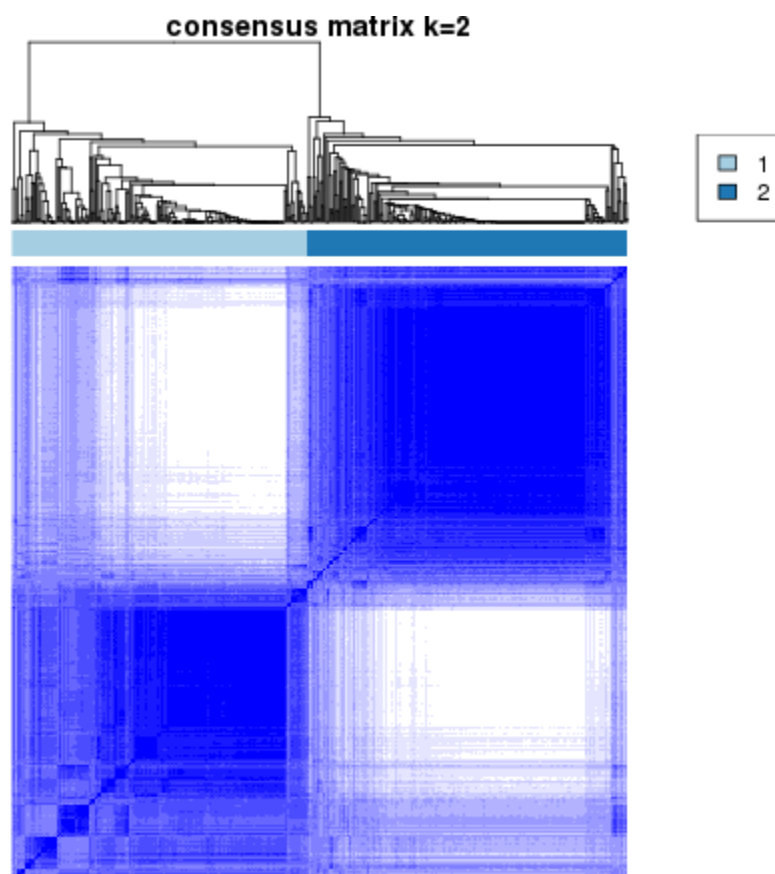
Figure 1: Volcano plot of differentially expressed genes (DEGs) in High risk Vs. low risk samples

**Figure 2**



*Figure 2: Principal component analysis for High risk and Low risk samples*

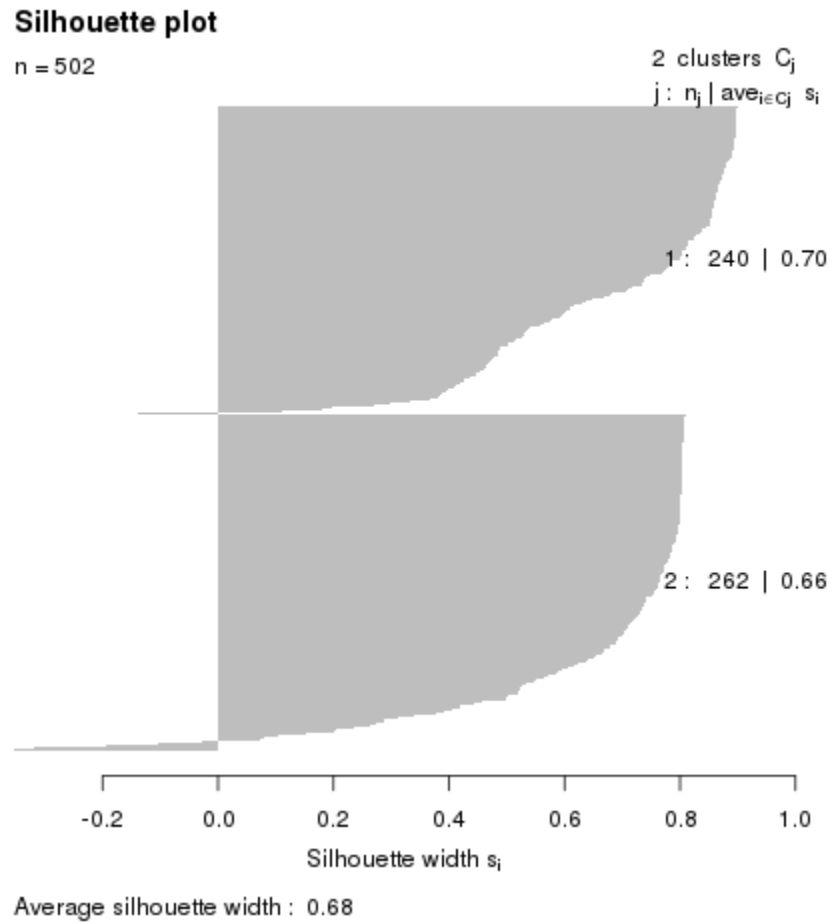
### Figure 3



*Figure 3: Consensus clustering results for LUSC samples*



**Figure 4**



*Figure 4: Silhouette width for Cancer subtype Validation*

## Figure 5

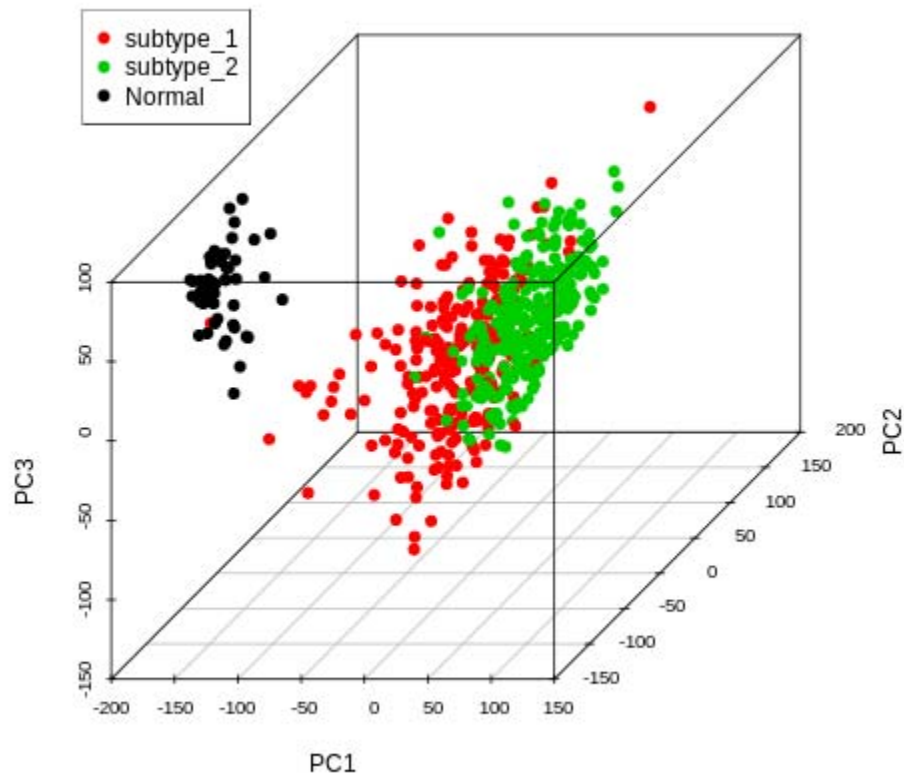


Figure 5 : Principal component analysis for LUSC samples

Figure 6

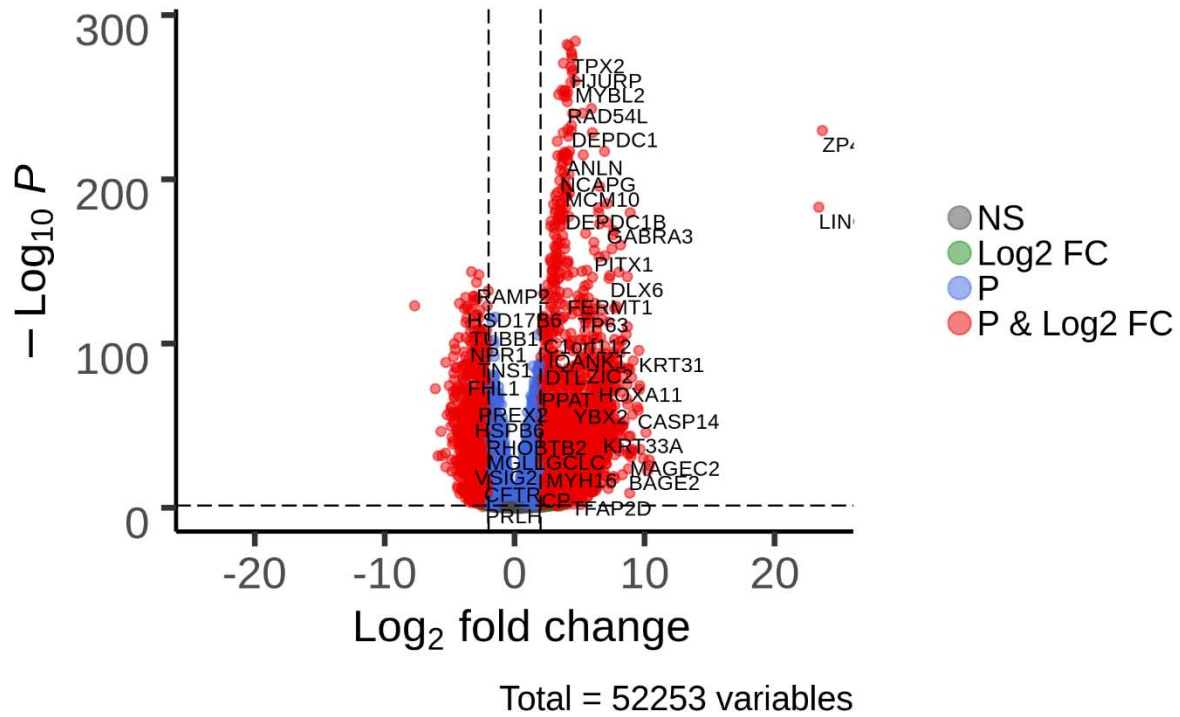


Figure 6: Differential gene expression in subtype 1 Vs. Normal samples

Figure 7

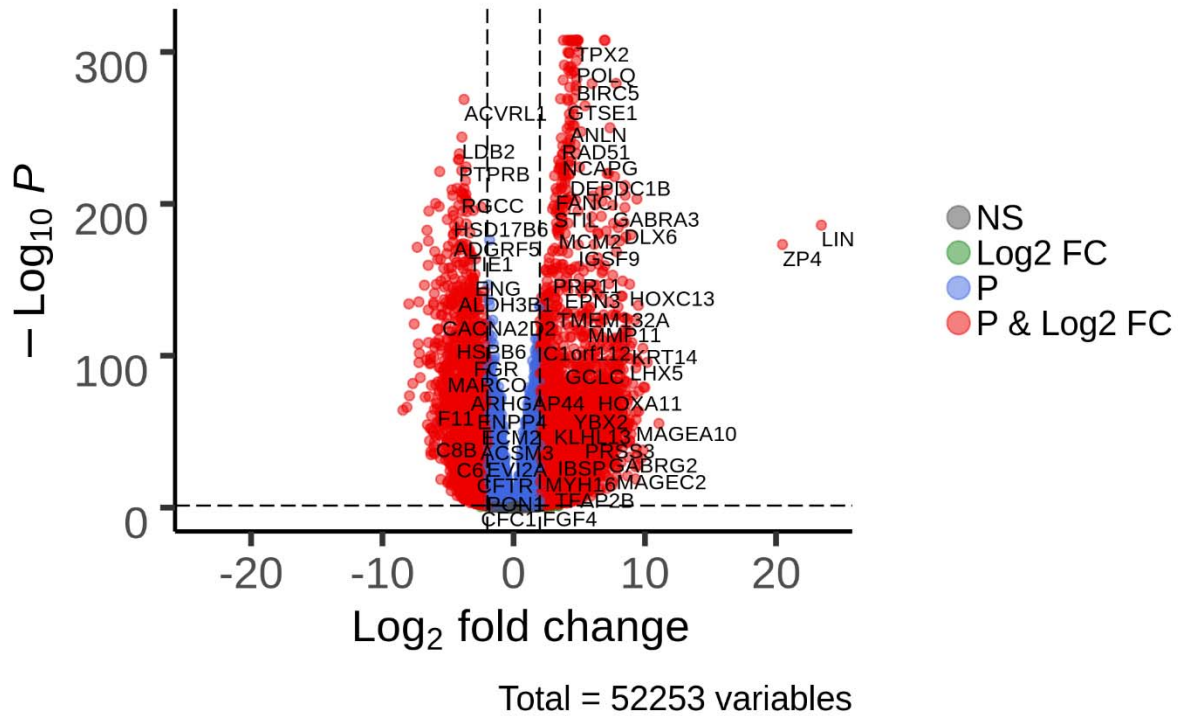
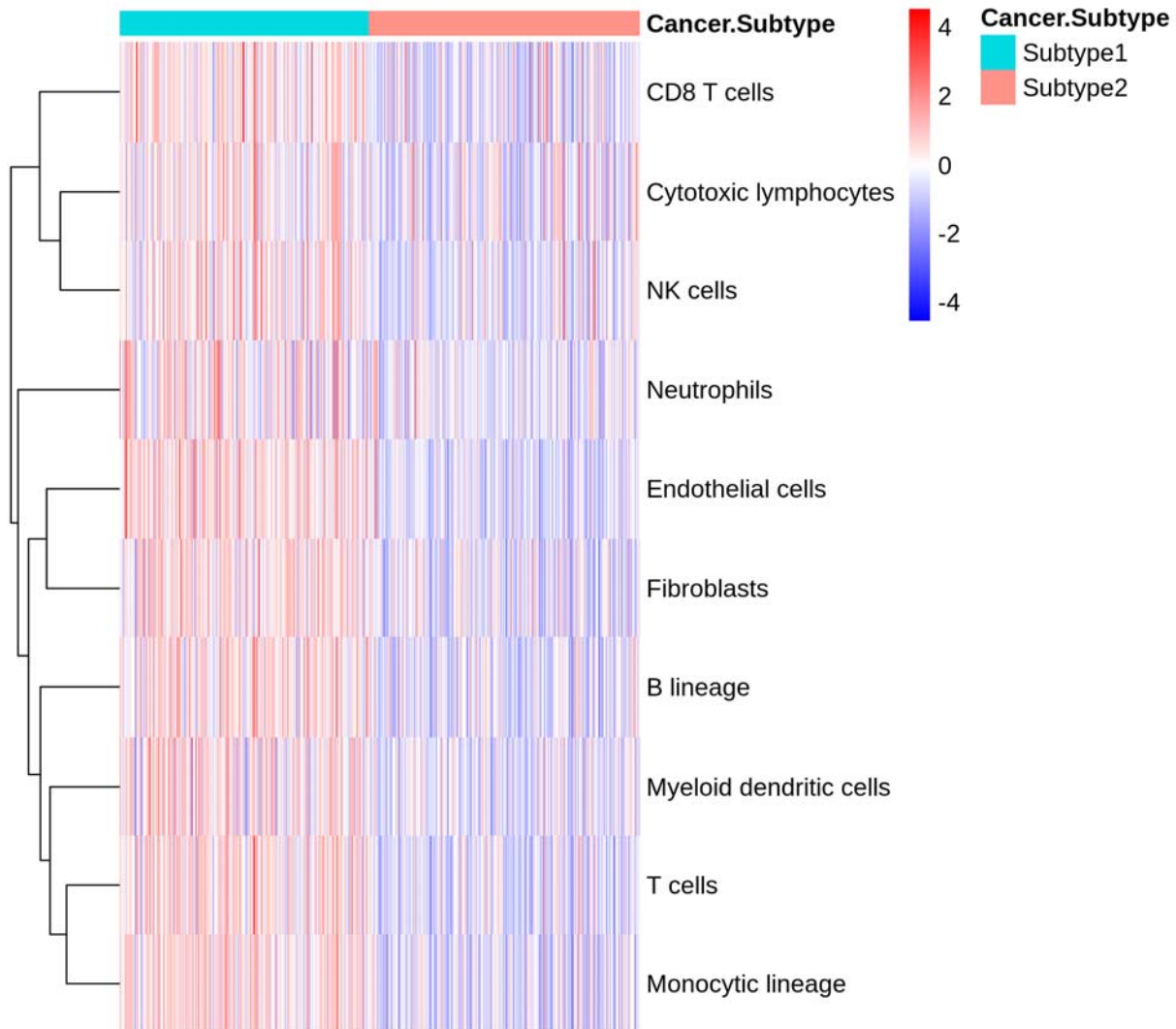


Figure 7: Differential gene expression in subtype 2 Vs. Normal samples

**Figure 8**



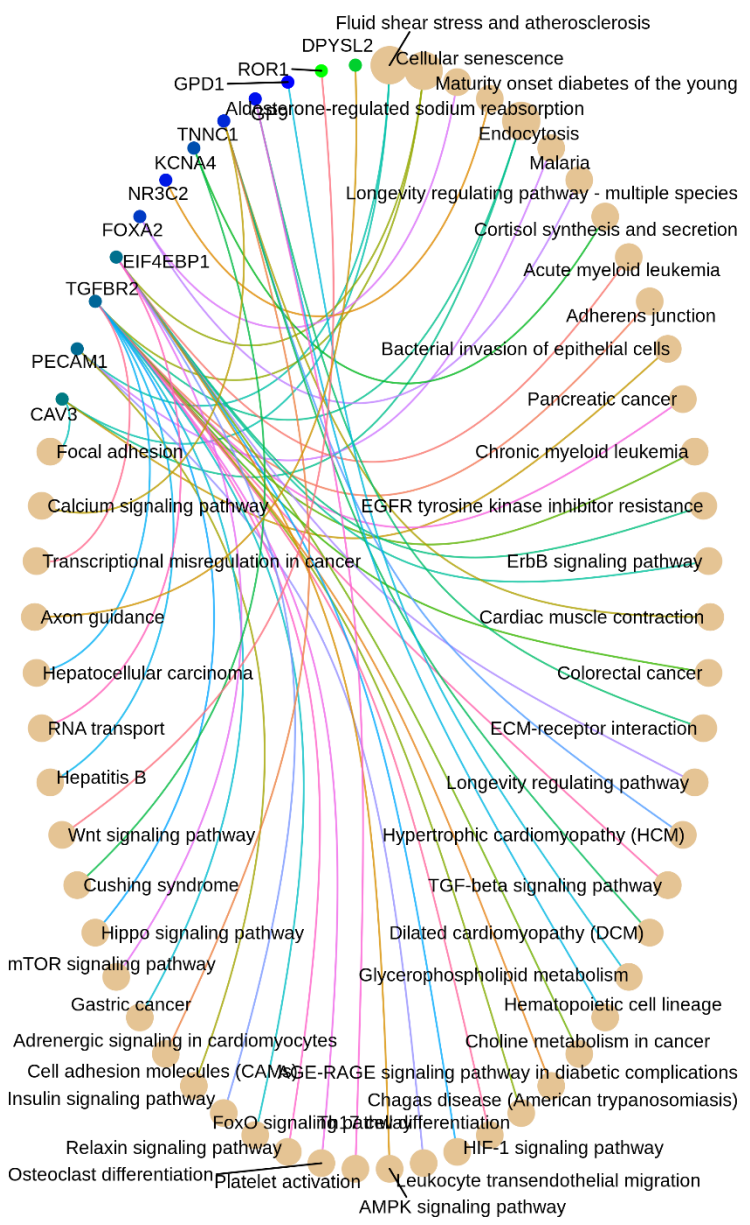
*Figure 8: Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression in Subtype-1 and Subtype-2 samples*



## Figure 9



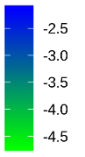




category

- Acute myeloid leukemia
- Adherens junction
- Adrenergic signaling in cardiomyocytes
- AGE-RAGE signaling pathway in diabetic complications
- Aldosterone-regulated sodium reabsorption
- AMPK signaling pathway
- Axon guidance
- Bacterial invasion of epithelial cells
- Calcium signaling pathway
- Cardiac muscle contraction
- Cell adhesion molecules (CAMs)
- Cellular senescence
- Chagas disease (American trypanosomiasis)
- Choline metabolism in cancer
- Chronic myeloid leukemia
- Colorectal cancer
- Cortisol synthesis and secretion
- Cushing syndrome
- Dilated cardiomyopathy (DCM)
- ECM-receptor interaction
- EGFR tyrosine kinase inhibitor resistance
- Endocytosis
- ErbB signaling pathway
- Fluid shear stress and atherosclerosis
- Focal adhesion
- FoxO signaling pathway
- Gastric cancer
- Glycerophospholipid metabolism
- Hematopoietic cell lineage
- Hepatitis B
- Hepatocellular carcinoma
- HIF-1 signaling pathway
- Hippo signaling pathway
- Hypertrophic cardiomyopathy (HCM)
- Insulin signaling pathway
- Leukocyte transendothelial migration
- Longevity regulating pathway
- Longevity regulating pathway - multiple species
- Malaria
- Maturity onset diabetes of the young
- mTOR signaling pathway
- Osteoclast differentiation
- Pancreatic cancer
- Platelet activation
- Relaxin signaling pathway
- RNA transport
- TGF-beta signaling pathway
- Th17 cell differentiation
- Transcriptional misregulation in cancer
- Wnt signaling pathway

fold change



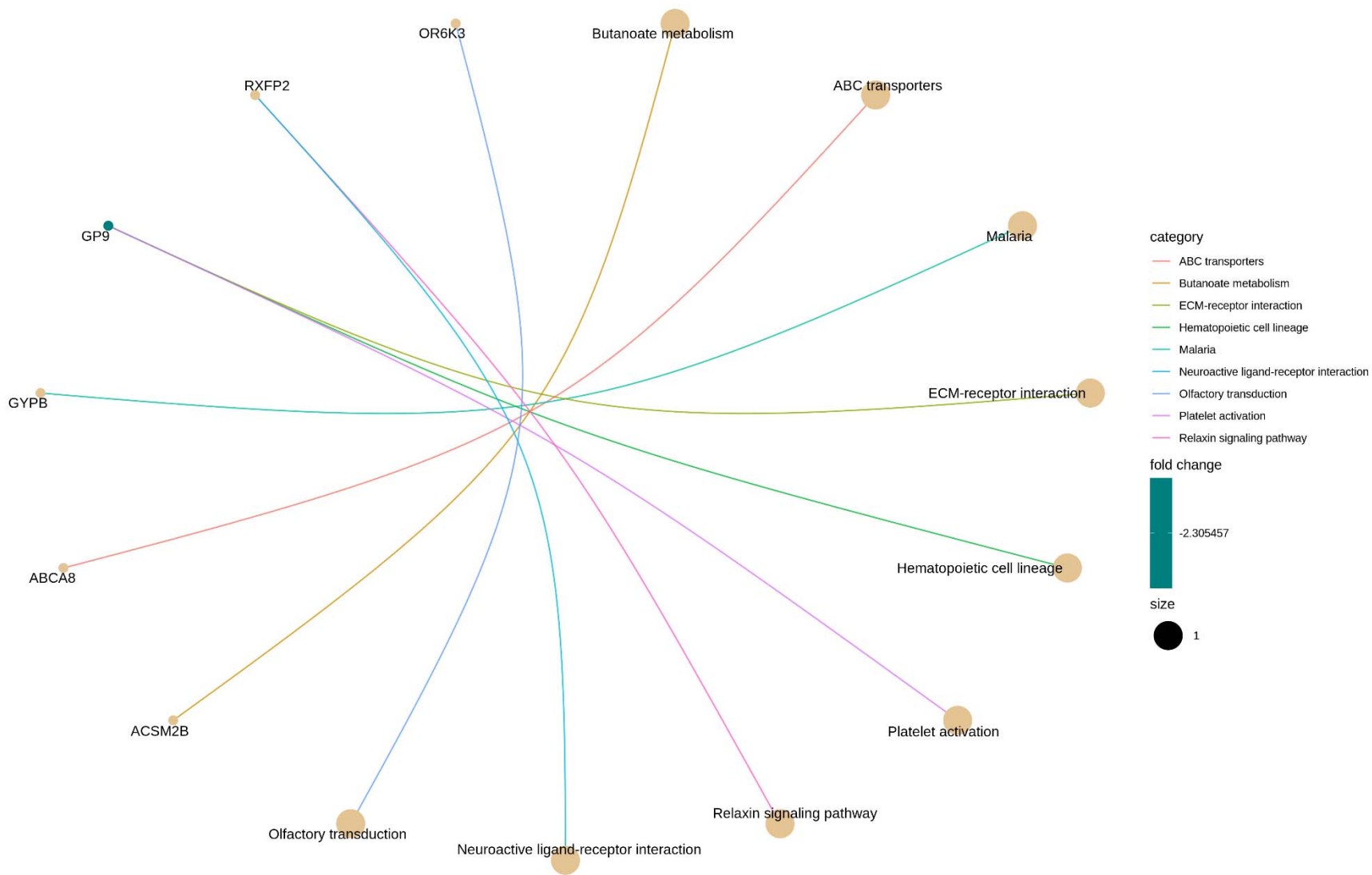
size



Figure 9: cancer regulatory pathways in Subtype-1 samples



Figure 10



*Figure 10: cancer regulatory pathways in Subtype-2 samples*

## Figure 11

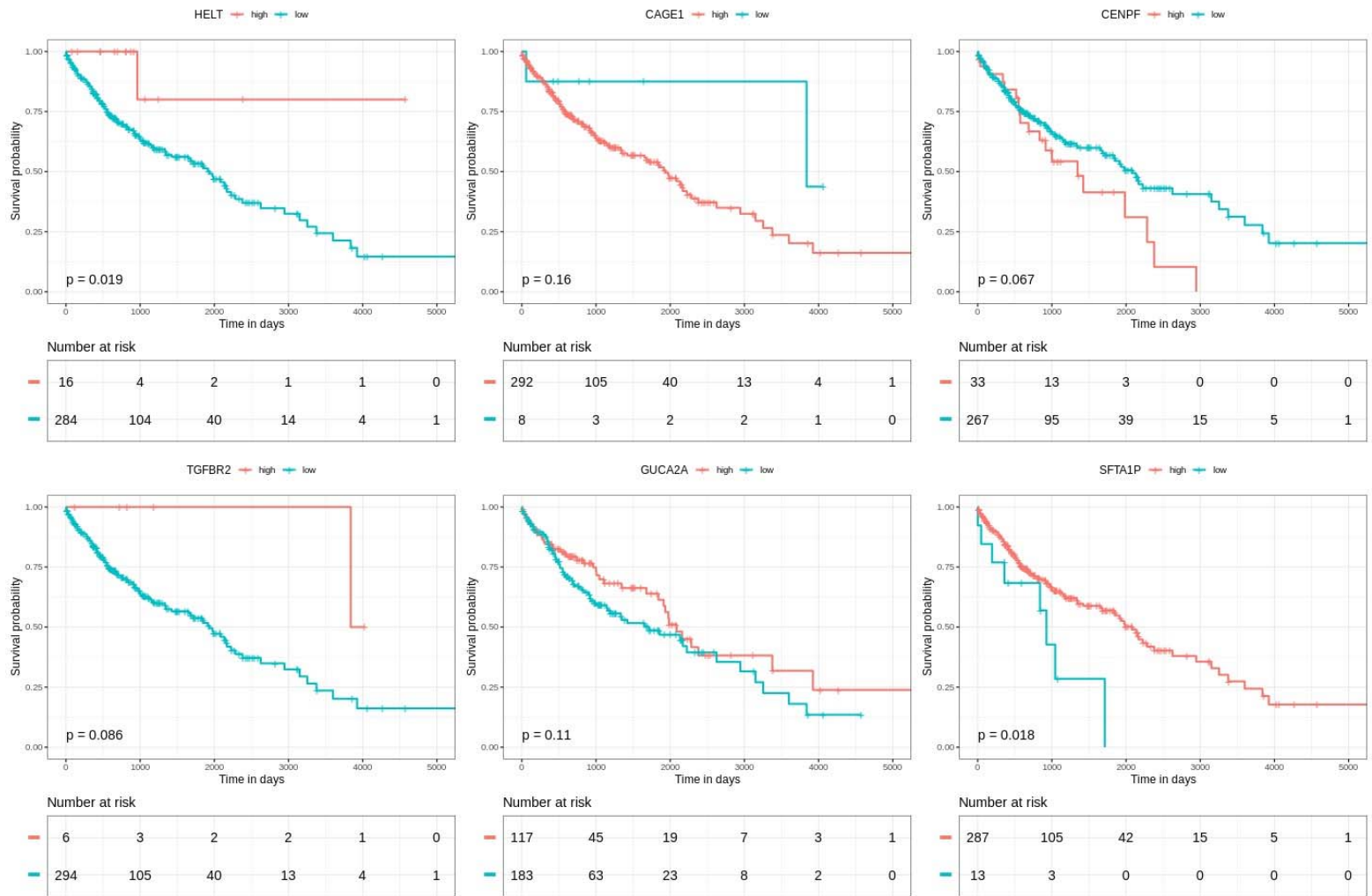


Figure 11: Survival analysis for LASSO predicted genes in Subtype-1 samples

## Figure 12



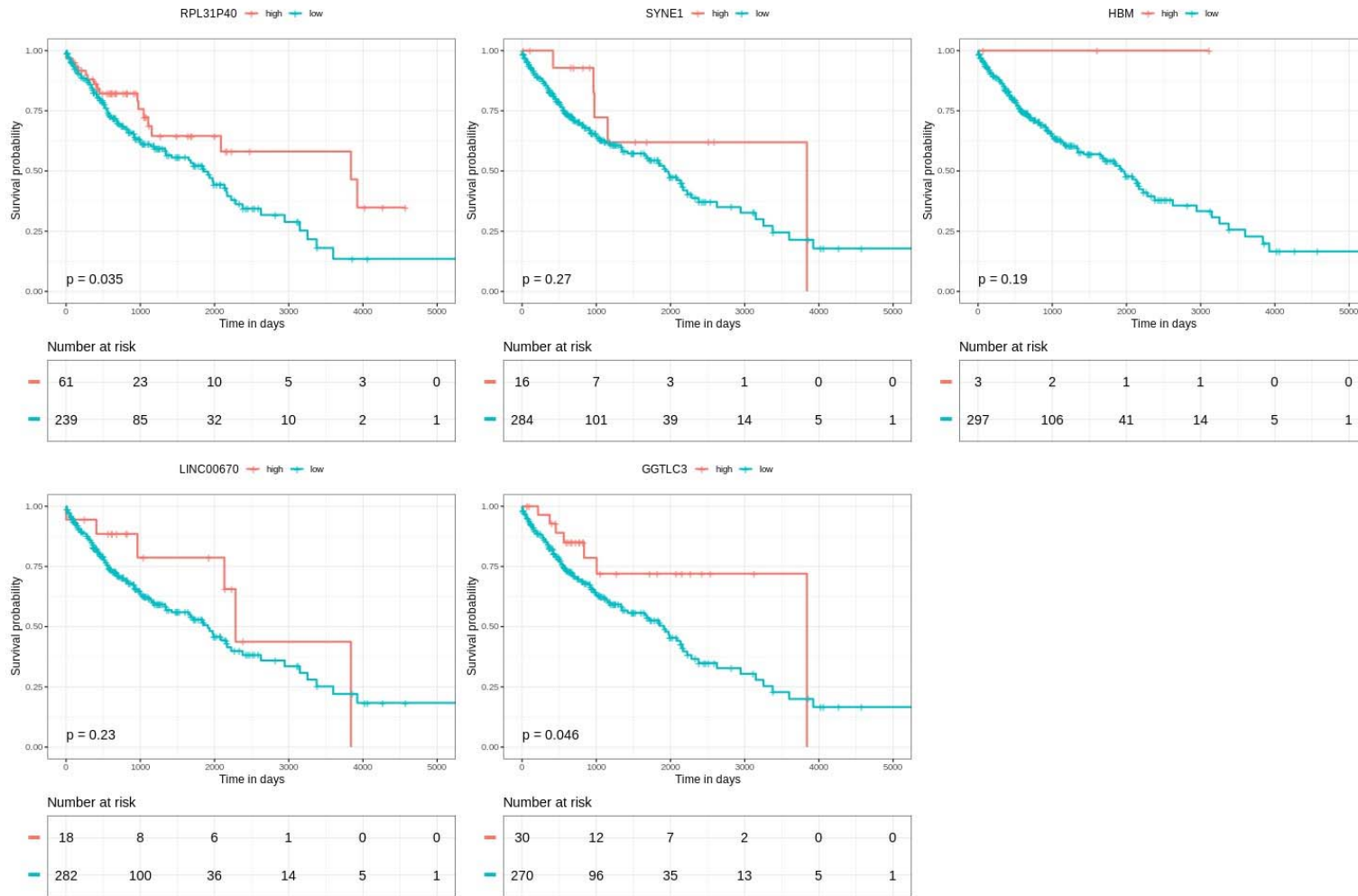


Figure 12: Survival analysis for LASSO predicted genes in Subtype-2 samples

## Table 1

*Table 1: Most relevant genes in established by LASSO model in Subtype-1 of LUSC*

<b>Genes</b>	<b>coefficients</b>
GAL	0.003457
MYCT1	0.003006
IMPDH1P8	0.001834
TCF21	0.001811
FOXA2	0.001753
ROR1	0.001641
NR3C2	0.001566
GPD1	0.001236
PPIAP45	0.001113
LINC01977	0.000995
GPR19	0.000891
RTBDN	0.000838
PGM5	0.000726
AFF3	0.000712
LINC01572	0.000662
TMEM249	0.000385
TNNC1	0.000216
CAGE1	0.000165
TFAP2A	2.63E-05
PGM5	-2.29E-05
HRCT1	-3.03E-05
C1orf87	-8.42E-05
DPYSL2	-0.00011
NEK5	-0.00014
NKAPL	-0.00016
RBMY1KP	-0.00017
TGFBR2	-0.00024
EIF4EBP1	-0.00051
CENPF	-0.0006
RPL31P40	-0.0006
FFAR4	-0.00061

LINC01863	-0.0007
GAS2L2	-0.0007
KCNA4	-0.00079
DUSP27	-0.00079
LINC00670	-0.0009
CAV3	-0.00101
MIR4717	-0.00114
PECAM1	-0.00139
LINC00891	-0.00154
HBM	-0.00157
GP9	-0.0016
LINC02016	-0.00169
HELT	-0.0017
OR6N1	-0.0021
OR6K4P	-0.00346
CELF2	-0.0037
LINC02058	-0.00486
LINC00710	-0.00499

## Table 2

*Table 2: Most relevant genes in established by LASSO model in Subtype-2 of LUSC*

Genes	Coefficient
MIR3131	-0.01287
LINC00844	-0.01122
RXFP2	-0.00708
PDZRN3.AS1	-0.00704
GYPB	-0.00534
KHDRBS2.OT	-0.00515
DPPA3P2	-0.00496
HBM	-0.00425
RPL31P40	-0.00424
LINC00710	-0.00408
SYNE1.AS1	-0.00382
GP9	-0.00296
PGM5.AS1	-0.00277
MIR6071	-0.00207
LINC01070	-0.00201
LINC01985	-0.00184
LINC02435	-0.00152
KCNA10	-0.00119
OR6K3	-0.00113
MIR4717	-0.00094
LINC02016	-0.00077
LINC00670	-0.00066
NCAPGP2	-0.00063
GGTLC3	-0.00055
GUCA2A	-0.00038
ART1	-0.0003
ACSM2B	-0.00021
GPIHBP1	2.05E-05
ATOH8	8.75E-05
TCF21	0.000133
ADGRD1	0.000279
MAMDC2	0.000348
ABCA8	0.000531

SFTA1P	0.001316
--------	----------

## TABLE 3

Table 3: KEGG pathway analysis for Subtype-1 group

Genes	KEGG ID	Description
CAV3	hsa05100,hsa04510,hsa05205	Bacterial invasion of epithelial cells, Focal adhesion, Proteoglycans in cancer
CAV3/PECAM1	hsa05418	Fluid shear stress and atherosclerosis
DPYSL2	hsa04360	Axon guidance
EIF4EBP1	hsa05221	Acute myeloid leukemia, EGFR tyrosine kinase inhibitor resistance, ErbB signaling pathway, Longevity regulating pathway, Choline metabolism in cancer, HIF-1 signaling pathway
EIF4EBP1	hsa05221, hsa01521, hsa04012, hsa04211, hsa05231, hsa04066, hsa04152, hsa04910, hsa04150, hsa03013, hsa05163, hsa04151, hsa05168	AMPK signaling pathway, Insulin signaling pathway, mTOR signaling pathway, RNA transport, Human cytomegalovirus infection, Human papillomavirus infection, PI3K-Akt signaling pathway,
FOXA2	hsa04950, hsa04213	Maturity onset diabetes of the young, Longevity regulating pathway - multiple species
GAL	hsa04080	Neuroactive ligand-receptor interaction
GP9	hsa04512, hsa04640, hsa04611	ECM-receptor interaction, Hematopoietic cell lineage, Platelet activation
GPD1	hsa00564	Glycerophospholipid metabolism
KCNA4	hsa04927, hsa04934	Cortisol synthesis and secretion, Cushing syndrome
NR3C2	hsa04960	Aldosterone-regulated sodium reabsorption
OR6N1	hsa04740	Olfactory transduction
PECAM1	hsa05144, hsa04670,hsa04514	Malaria, Leukocyte transendothelial migration, Cell adhesion molecules (CAMs)
ROR1	hsa04310	Wnt signaling pathway
TGFBR2	hsa04520	Adherens junction
TGFBR2	hsa05212, hsa05220, hsa05210, hsa04350, hsa04933, hsa05142, hsa04659, hsa04380, hsa04926, hsa04068, hsa05226, hsa04390, hsa05161, hsa05225,	Pancreatic cancer, Chronic myeloid leukemia, Colorectal cancer, TGF-beta signaling pathway, AGE-RAGE signaling pathway in diabetic complications, Chagas disease (American trypanosomiasis), Th17 cell differentiation, Osteoclast differentiation, Relaxin signaling pathway, FoxO signaling

	hsa05202, hsa05166, hsa04060, hsa04010	pathway, Gastric cancer, Hippo signaling pathway, Hepatitis B, Hepatocellular carcinoma, Transcriptional misregulation in cancer, Human T-cell leukemia virus 1 infection, Cytokine-cytokine receptor interaction, MAPK signaling pathway
TGFBR2/CAV3	hsa04144	Endocytosis
TGFBR2/EIF4EBP1	hsa04218	Cellular senescence
TNNC1	hsa04260, hsa05410, hsa05414, hsa04261, hsa04020	Cardiac muscle contraction, Hypertrophic cardiomyopathy (HCM), Dilated cardiomyopathy (DCM), Adrenergic signaling in cardiomyocytes, Calcium signaling pathway

## TABLE 4

*Table 4: KEGG pathway analysis for Subtype-2 group*

<b>Genes</b>	<b>ID</b>	<b>Description</b>	<b>pvalue</b>	<b>p.adjust</b>	<b>qvalue</b>
ACSM2B	hsa00650	Butanoate metabolism	0.021127	0.110183	0.077321
ABCA8	hsa02010	ABC transporters	0.033772	0.110183	0.077321
GYPB	hsa05144	Malaria	0.036728	0.110183	0.077321
GP9	hsa04512	ECM-receptor interaction	0.06515	0.121208	0.085058
GP9	hsa04640	Hematopoietic cell lineage	0.071609	0.121208	0.085058
GP9	hsa04611	Platelet activation	0.090762	0.121208	0.085058
RXFP2	hsa04926	Relaxin signaling pathway	0.094273	0.121208	0.085058
RXFP2	hsa04080	Neuroactive ligand-receptor interaction	0.231252	0.260158	0.182567
OR6K3	hsa04740	Olfactory transduction	0.296107	0.296107	0.207794