International Journal of Life science and Pharma Research

# GENE PREDICTION IN HETEROGENEOUS CANCER TISSUES AND ESTABLISHMENT OF LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR MODEL FOR LUNG SQUAMOUS CELL CARCINOMA.

**ATEEQ MUHAMMED KHALIQ[1], SHARATHCHANDRA R G[1*] AND MEENAKSHI RAJAMOHAN[1]**

*[1]Department of Studies and Research in Biotechnology and Centre for Bioinformation, Tumkur University, Tumkur, Karnataka, India*

## ABSTRACT

This study is aimed to establish a Least Absolute Shrinkage and Selection Operator (LASSO) model based on tumor heterogeneity to predict the best features of LUSC in various cancer subtypes. 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 the 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. 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 up regulated and 1495 were down regulated in case of subtype-1 and 5016 genes were up regulated and 3224 were down regulated in case of subtype-2. LASSO model was established to predict the best features from each subtype, 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. This study suggests that 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.

**KEYWORDS:** ***Cancer Prediction, LASSO, Unsupervised clustering, Machine Learning, Cancer heterogeneity, Biomarkers***

**SHARATHCHANDRA R G ***

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

**Received on: 14-08-2019**
**Revised and Accepted on: 11-10-2019**
**DOI: http://dx.doi.org/10.22376/ijpbs/lpr.2019.9.4.L34-48**

# INTRODUCTION

Lung cancer is among the most deadly cancers[1]. Its 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 [2]. 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 [3,4]. 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 [5–8]. Previous studies indicate standard stepwise selection approaches which are not best for regression models with a very large number of covariates [9]. 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 [10]. 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, $\sum j=1k\beta j$, 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. 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/ .

### *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) [11] was used for the identification of subtypes on set 2 group. Silhouette width [12] 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[13] (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[14] (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[15] were generated by selecting the best predictors from individual subtypes. Kaplan-Meier curves were generated using the TRGAted [16](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[17] (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[18] (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) [19] was utilized to annotate genes to pathways, and Disease Ontologies.

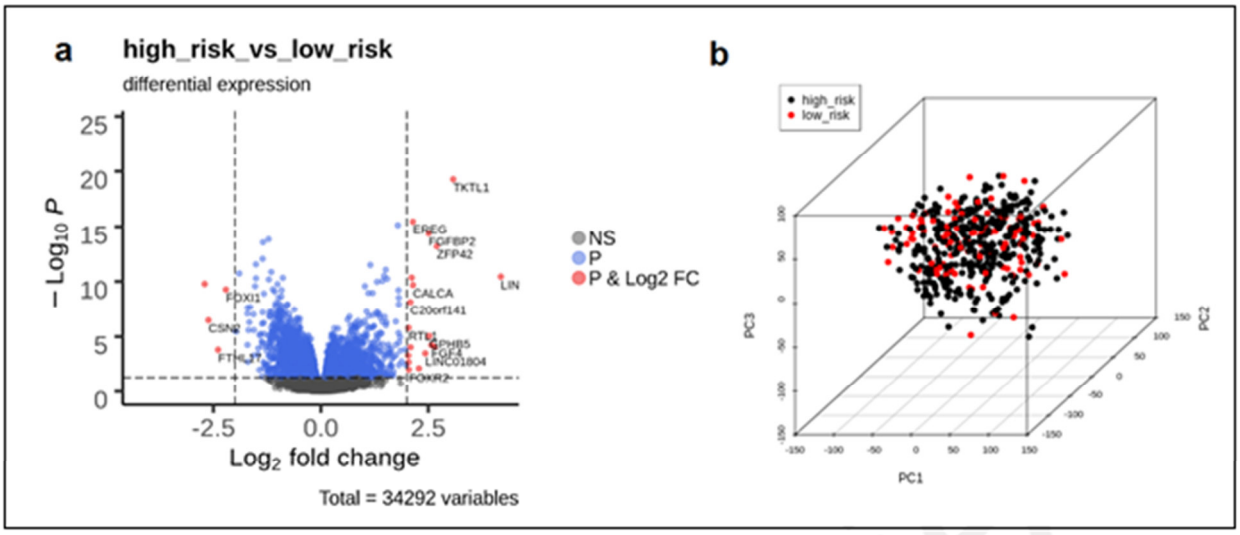


**Figure 1**

***Expression Pattern and PCA analysis: (a) Volcano plot of differentially expressed genes (DEGs) in High risk Vs. low risk samples. (b) Principal component analysis for High risk and Low risk samples***

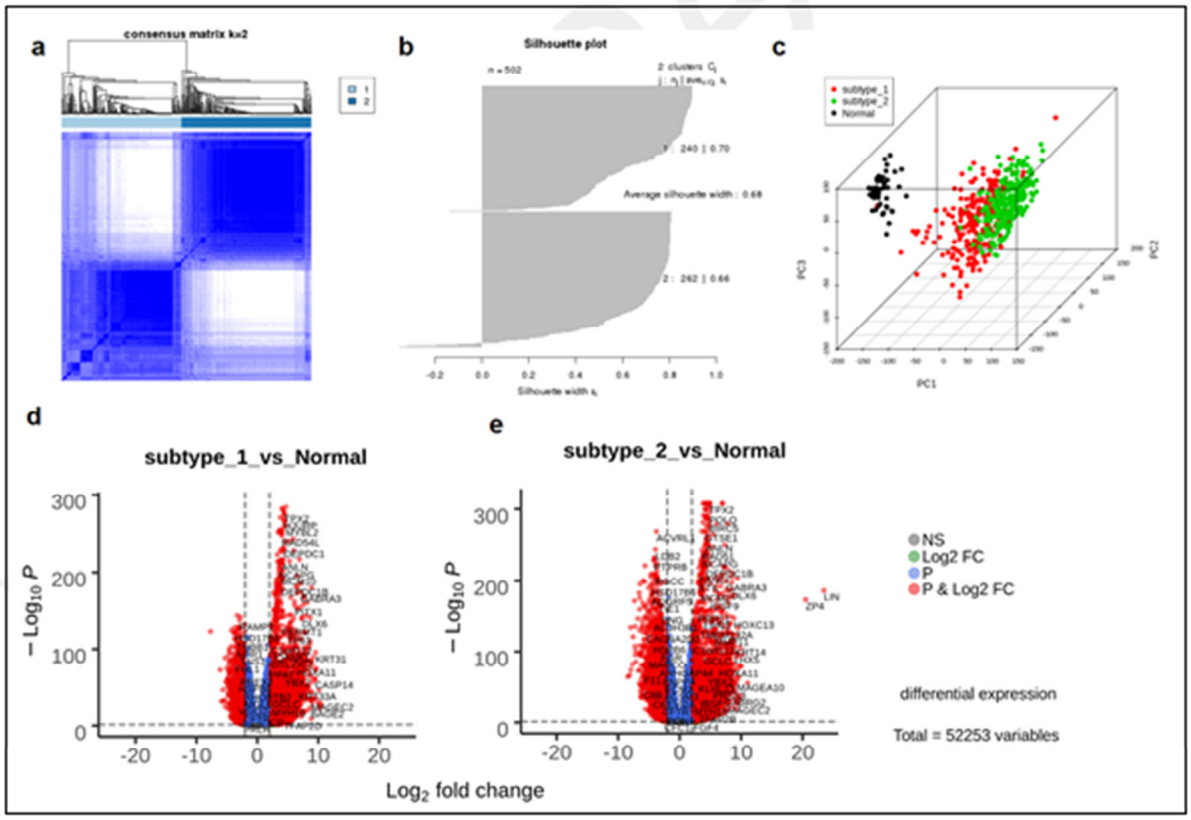


**Figure 2**

***Cluster identification and validation. Fig 2.a. Consensus clustering results for LUSC samples. (b)Silhouette width for Cancer subtype Validation. (c) Principal component analysis for LUSC samples. (d) Differential gene expression in subtype 1 Vs. Normal samples. (e) Differential gene expression in subtype 2 Vs. Normal samples***

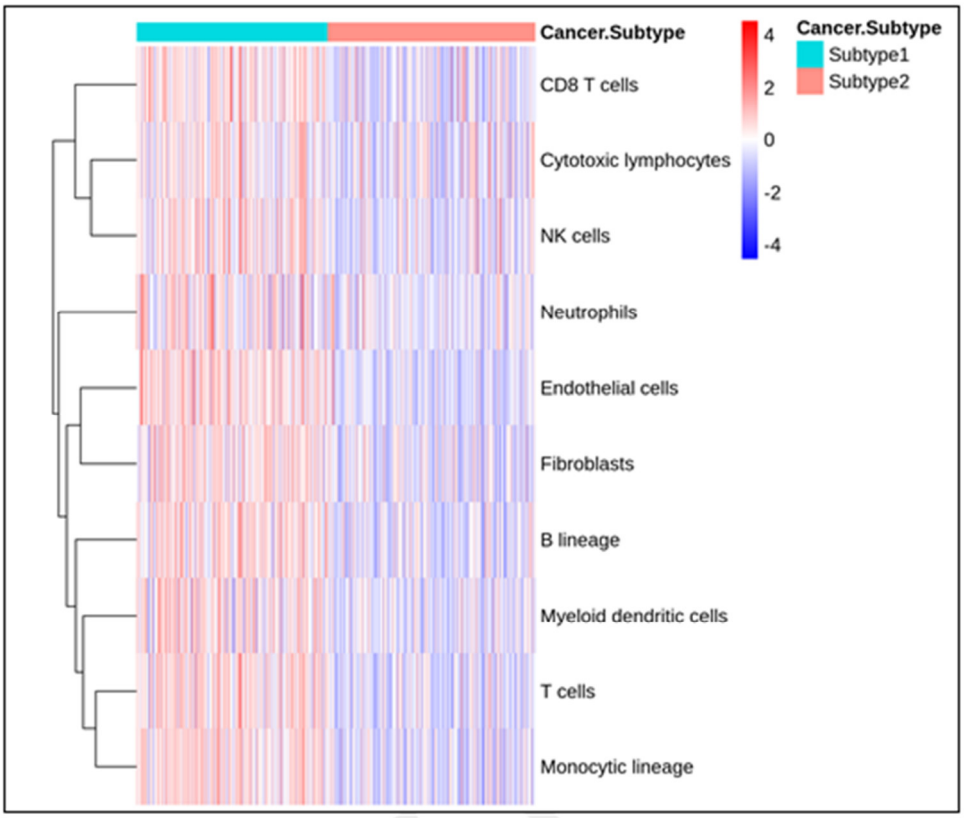


**Figure 3**

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

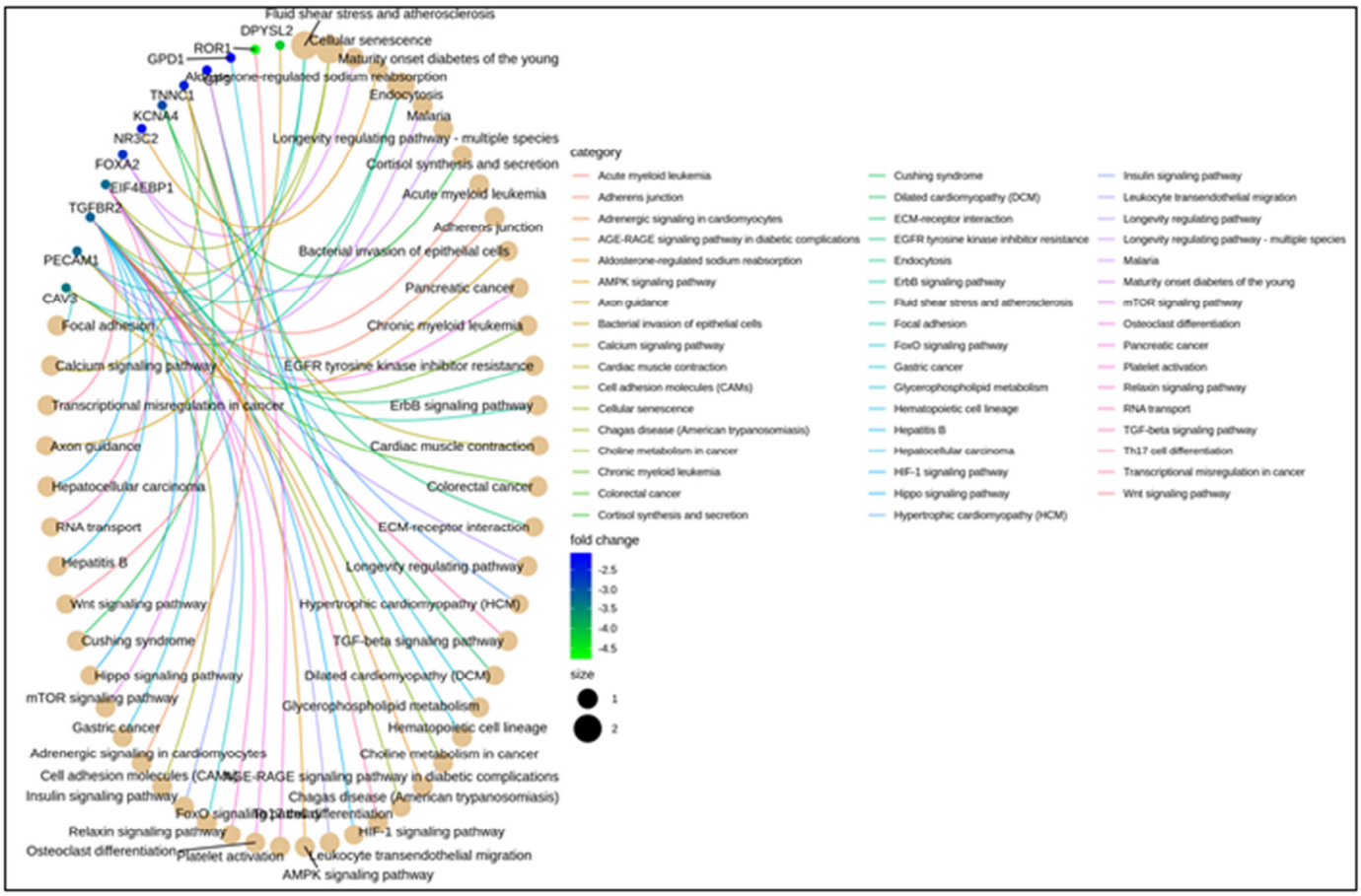


**Figure 4 a**

***Cancer regulatory pathways in (a) Subtype-1 samples and (b) Subtype-2 samples***

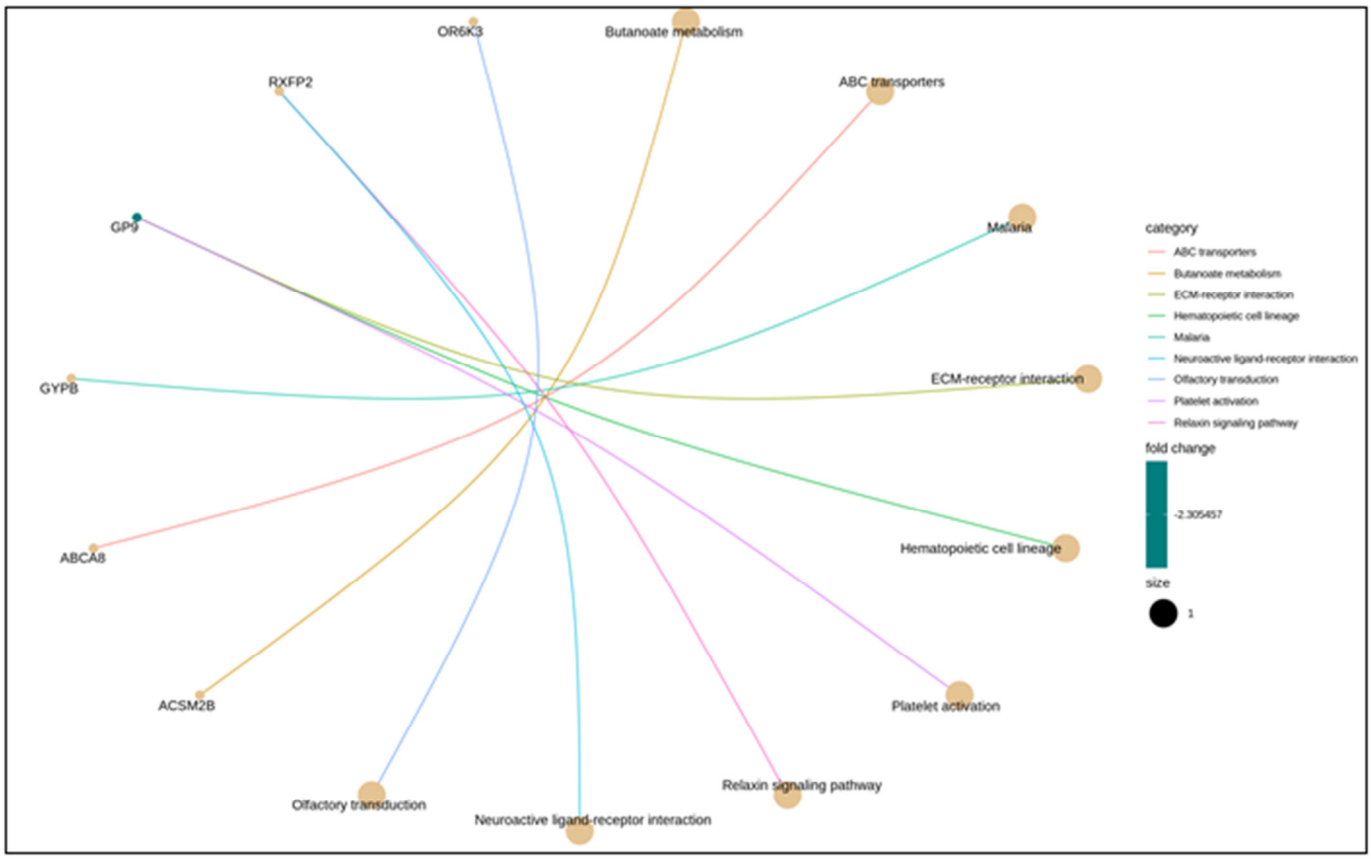


**Figure 4 b**

***Cancer regulatory pathways in Subtype-2 samples***

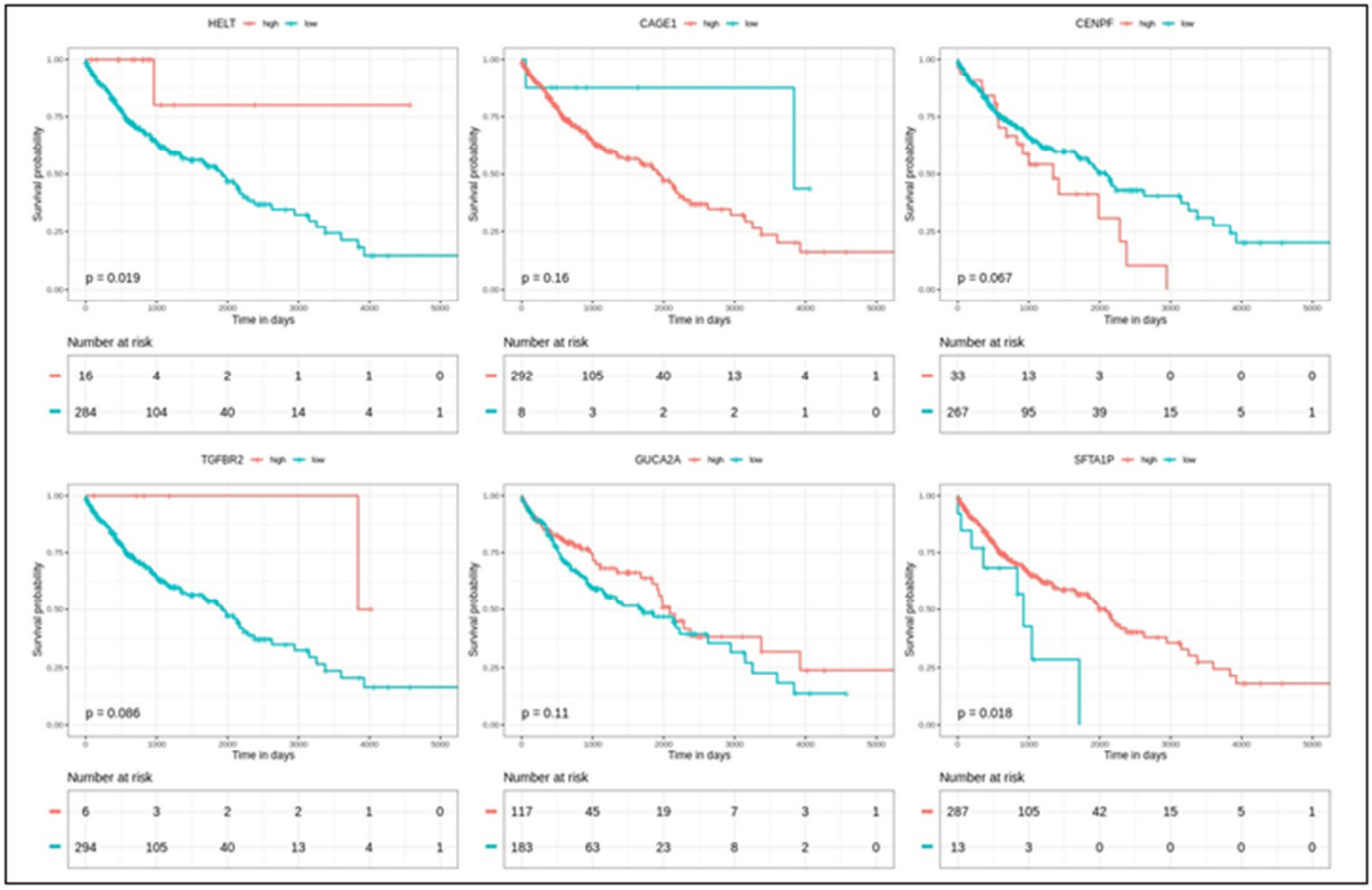


**Figure 5 a**

***Survival analysis for LASSO predicted genes in Subtype-1 samples***

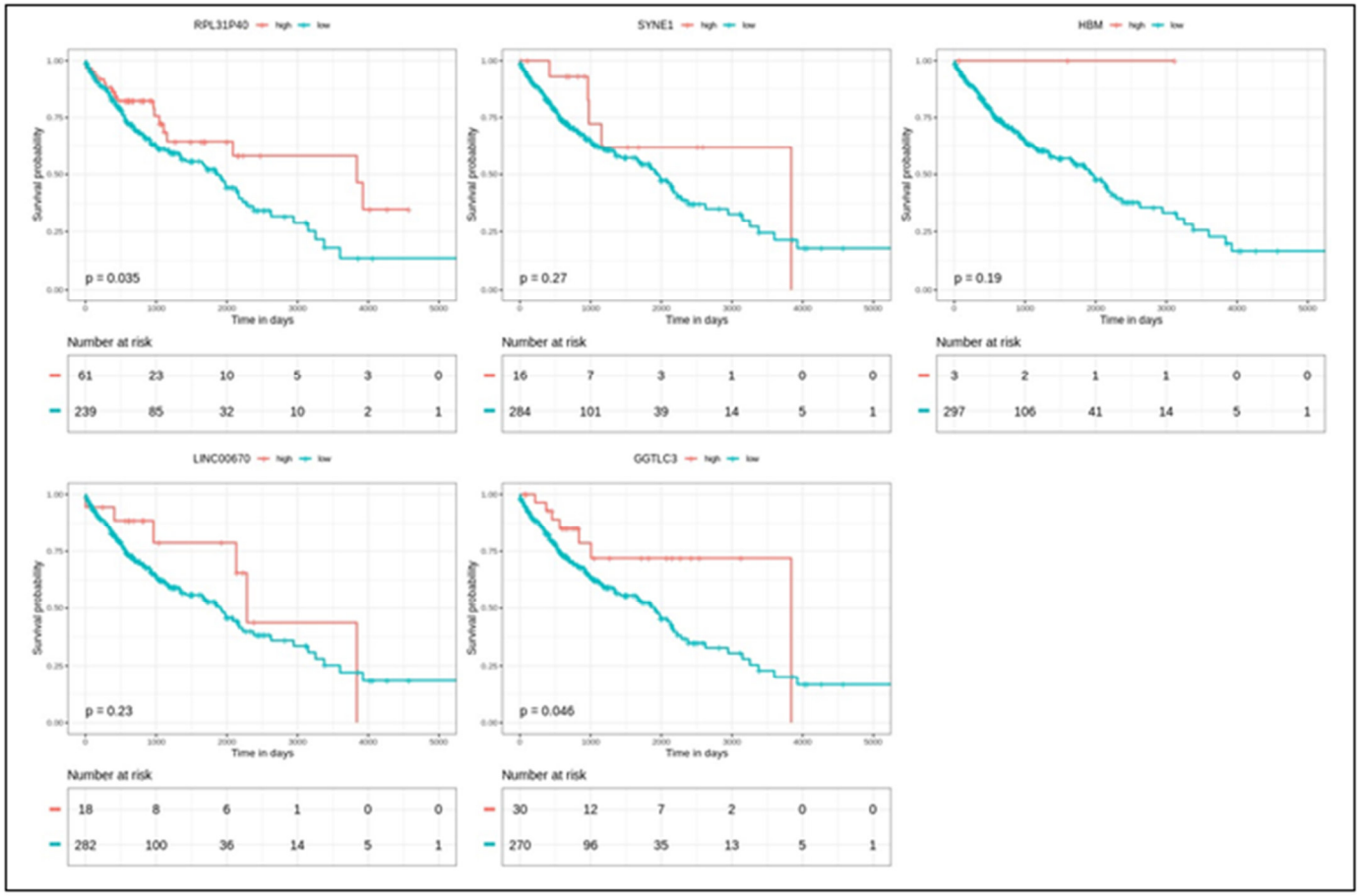


**Figure 5 b**

***Survival analysis for LASSO predicted genes in Subtype-2 samples***

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

| Genes | Coefficients |
|---|---|
| GAL | 0.003457 |
| MYCT1 | 0.003006 |
| IMPDH 1P8 | 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 |
| TM M249 | 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**
***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***
***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, | AMPK signaling pathway, Insulin signaling pathway, mTOR signaling pathway, RNA |

| | | |
|---|---|---|
| | hsa04152, hsa04910, hsa04150, hsa03013, hsa05163, hsa04151, hsa05168 | 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, hsa05202, hsa05166, hsa04060, hsa04010 | 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 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***
***KEGG pathway analysis for Subtype-2 group***

| Genes | ID | Description | P value | p. adjust | Q value |
|---|---|---|---|---|---|
| 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 |

# RESULT

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

The genes with p-value cutoff < 0.05 and log2 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.a* displays the heat map of the risk-related DEGs. It 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 1b.*).

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

We used an unsupervised clustering method Consensus clustering (CC) [11]. 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 [12]. The Average Silhouette width for our generated clusters is 0.68 (*Fig 2a, 2b. and 2c.*).

***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 log2 fold change > 2 we identified significant DEGs. 4586 genes were upregulated and 1495 were downregulated in case of subtype-1 (*Fig 2d.*) and 5016 genes were upregulated and 3224 were downregulated in case of subtype-2 (*Fig 2e*) 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***

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 >> 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 >> 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 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 λ for our LASSO model. We analyzed the λ value using 10 fold cross-validation *(Fig S3 and S4),* between λ min that gives minimum mean cross-validated error or λ1se, which gives a model such that error is 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 subtype 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 [20]. T-cell infiltrates, endothelial cells and fibroblasts are associated with a favorable outcome and also poor prognosis in some cancer types [21–24]. To understand the immunological microenvironment in our expression subset-1 and subset-2 we used MCP-counter method as described by Becht et al [17]. The estimations consist of single sample scores which are computed on each sample independently in two subtypes. The heatmap shown in *Figure 3* 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 4a. and 4b.*). 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 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 5a. and 5b* , *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 [25,26]. 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 numbers 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 GGTLC3, GUCA2A, HBM, SFTA1P and SYNE1 in Subtype-2 showed worse overall survival in LUSC cancer samples. Whereas increased expression of genes such as PPIA P45, 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[27]. 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 an 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[28]. 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 [29]. 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 [30]. The modulators of ABC transporters is reported to have the potential to augment the efficacy of anticancer drugs [31]. 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 [32–37]. 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 [38–40] Under expression of TGFBR2 in Subtype-1 samples is associated with poor prognosis, and TGFBR2 is also associated with poor prognosis in cervical cancer [41,42]. 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 [42-45]. Down regulation of SFTA1P in Subtype-2 correlates with poor survival, previous studies suggest SFTA1P regulates both oncogene and tumor suppressor genes during carcinogenesis of lung squamous cell carcinoma [46-49] 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 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 the 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.

# AUTHORS CONTRIBUTION STATEMENT

Ateeq Muhammed Khaliq., Meenakshi R, and Dr. Sharathchandra R.G equally contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript.

# 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.

# CONFLICT OF INTEREST

Conflict of interest declared none.

# REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. CA Cancer J Clin. 2015;65(1):5–29. DOI: 10.3322/caac.21254
2. Inamura K. Lung Cancer: Understanding Its Molecular Pathology and the 2015 WHO Classification. Front Oncol. 2017;7. DOI: 10.3389/fonc.2017.00193
3. Beca F, Polyak K. Intratumor Heterogeneity in Breast Cancer. Advances in Experimental Medicine and Biology. Springer International Publishing; 2016. p. 169–89. DOI: 10.1007/978-3-319-22909-6_7
4. Bolck HA, Corrò C, Kahraman A, von Teichman A, Toussaint NC, Kuipers J, et al. Tracing Clonal Dynamics Reveals that Two- and Three-dimensional Patient-derived Cell Models Capture Tumor Heterogeneity of Clear Cell Renal Cell Carcinoma. Eur Urol Focus. 2019; DOI: 10.1016/j.euf.2019.06.009
5. Ayer T, Chhatwal J, Alagoz O, Kahn CE, Woods RW, Burnside ES. Comparison of Logistic Regression and Artificial Neural Network Models in Breast Cancer Risk Estimation. RadioGraphics. 2010;30(1):13–22. DOI: 10.1148/rg.301095057
6. Zhu L, Luo W, Su M, Wei H, Wei J, Zhang X. Comparison between artificial neural network and Cox regression model in predicting the survival rate of gastric cancer patients. Biomed Reports. 2013;1(5):757–60. DOI: 10.3892/br.2013.140

7. Bartfay E, Mackillop Wj, Pater Jl. Comparing the predictive value of neural network models to logistic regression models on the risk of death for small-cell lung cancer patients. Eur J Cancer Care (Engl). 2006;15(2):115–24. DOI: 10.1111/j.1365-2354.2005.00638.x
8. Jiang H, Ching W-K. Correlation Kernels for Support Vector Machines Classification with Applications in Cancer Data. Comput Math Methods Med. 2012;2012:1–7. DOI: 10.1155/2012/205025
9. Houssami N, Irwig L, Simpson JM, McKessar M, Blome S, Noakes J. The influence of clinical information on the accuracy of diagnostic mammography. Breast Cancer Res Treat. 2004;85(3):223–8. DOI: 10.1023/b:brea.0000025416.66632.84
10. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. J R Stat Soc Ser B. 1996;58(1):267–88. DOI: 10.1111/j.2517-6161.1996.tb02080.x
11. Monti S, Tamayo P, Mesirov J GT. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. Mach Learn. 2003;52(1–2):91–118. Available from: https://link.springer.com/article/10.1023/A:1023949509487
12. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53–65. DOI: 10.1016/0377-0427(87)90125-7
13. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12). DOI: 10.1186/s13059-014-0550-8
14. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010;33(1). DOI: 10.18637/jss.v033.i01
15. Kaplan EL, Meier P. Nonparametric Estimation from Incomplete Observations. J Am Stat Assoc 1958;53(282):457–81. DOI: 10.1080/01621459.1958.10501452
16. Borcherding N, Bormann NL, Voigt AP, Zhang W. TRGAted: A web tool for survival analysis using protein data in the Cancer Genome Atlas. F1000Research. 2018;7:1235. DOI: 10.12688/f1000research.15789.2
17. Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, et al. Erratum to: Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biol. 2016;17(1). DOI: 10.1186/s13059-016-1113-y
18. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. Omi A J Integr Biol. 2012;16(5):284–7. DOI: 10.1089/omi.2011.0118
19. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 2009;38(suppl_1):D355–60. DOI: 10.1093/nar/gkp896
20. Becht E, Giraldo NA, Germain C, de Reyniès A, Laurent-Puig P, Zucman-Rossi J. Immune Contexture, Immunoscore, and Malignant Cell Molecular Subgroups for Prognostic and Theranostic Classifications of Cancers. Advances in Immunology. Elsevier; 2016. p. 95–190. DOI: 10.1016/bs.ai.2015.12.002
21. Pagès F, Berger A, Camus M, Sanchez-Cabo F, Costes A, Molidor R. Effector Memory T Cells, Early Metastasis, and Survival in Colorectal Cancer. N Engl J Med. 2005;353(25):2654–66. DOI: 10.1056/nejmoa051424
22. Galon J. Type, Density, and Location of Immune Cells Within Human Colorectal Tumors Predict Clinical Outcome. Science (80- ). 2006;313(5795):1960–4. DOI: 10.1126/science.1129139
23. Fridman WH, Pagès F, Sautès-Fridman C, Galon J. The immune contexture in human tumours: impact on clinical outcome. Nat Rev Cancer. 2012;12(4):298–306. DOI: 10.1038/nrc3245
24. Giraldo NA, Becht E, Pages F, Skliris G, Verkarre V, Vano Y. Orchestration and Prognostic Significance of Immune Checkpoints in the Microenvironment of Primary and Metastatic Renal Cell Cancer. Clin Cancer Res. 2015;21(13):3031–40. DOI: 10.1158/1078-0432.ccr-14-2926
25. Dagogo-Jack I, Shaw AT. Tumour heterogeneity and resistance to cancer therapies. Nat Rev Clin Oncol. 2017;15(2):81–94. DOI: 10.1038/nrclinonc.2017.166
26. Murugaesu N, Wilson GA, Birkbak NJ, Watkins TBK, McGranahan N, Kumar S. Tracking the Genomic Evolution of Esophageal Adenocarcinoma through

Neoadjuvant Chemotherapy. Cancer Discov. 2015;5(8):821–31.
DOI: 10.1158/2159-8290.cd-15-0412

27. Jiang M-C, Ni J-J, Cui W-Y, Wang B-Y ZW. Emerging roles of lncRNA in cancer and therapeutic opportunities. Am J Cancer Res. 2019;9(7):1354–66. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6682721/
28. Walker C, Mojares E, del Río Hernández A. Role of Extracellular Matrix in Development and Cancer Progression. Int J Mol Sci. 2018;19(10):3028.
DOI: 10.3390/ijms19103028
29. He Z, Tang F, Lu Z, Huang Y, Lei H LZ. Analysis of differentially expressed genes, clinical value and biological pathways in prostate cancer. Am J Transl Res. 2018;10(5):1444–56.
30. Fue M, Miki Y, Takagi K, Hashimoto C, Yaegashi N, Suzuki T. Relaxin 2/RXFP1 Signaling Induces Cell Invasion via the β-Catenin Pathway in Endometrial Cancer. Int J Mol Sci. 2018;19(8):2438.
DOI: 10.3390/ijms19082438
31. Sun Y-L, Patel A, Kumar P, Chen Z-S. Role of ABC transporters in cancer chemotherapy. Chin J Cancer. 2012;31(2):51–7.
DOI: 10.5732/cjc.011.10466
32. Chapman CJ, Thorpe AJ, Murray A, Parsy-Kowalska CB, Allen J, Stafford KM. Immunobiomarkers in Small Cell Lung Cancer: Potential Early Cancer Signals. Clin Cancer Res. 2010;17(6):1474–80.
DOI: 10.1158/1078-0432.ccr-10-1363
33. Park S, Lim Y, Lee D, Cho B, Bang Y-J, Sung S. Identification and characterization of a novel cancer/testis antigen gene CAGE-1. Biochim Biophys Acta - Gene Struct Expr. 2003;1625(2):173–82.
DOI: 10.1016/s0167-4781(02)00620-6
34. Kim Y, Park D, Kim H, Choi M, Lee H, Lee YS. 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. J Biol Chem. 2013;288(51):36502–18.
DOI: 10.1074/jbc.m113.502047
35. Kunze E, Schlott T. High frequency of promoter methylation of the 14-3-3 σ 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. Int J Mol Med. 2007; Available from: http://dx.doi.org/10.3892/ijmm.20.4.557
36. Parmigiani RB, Bettoni F, Vibranovski MD, Lopes MH, Martins WK, Cunha IW. Characterization of a cancer/testis (CT) antigen gene family capable of eliciting humoral response in cancer patients. Proc Natl Acad Sci. 2006;103(48):18066–71. DOI: 10.1073/pnas.0608853103
37. Kunze E, Wendt M, Schlott T. Promoter hypermethylation of the 14-3-3 σ, SYK and CAGE-1 genes is related to the various phenotypes of urinary bladder carcinomas and associated with progression of transitional cell carcinomas. Int J Mol Med. 2006; DOI: 10.3892/ijmm.18.4.547
38. Zhang H, Du Y, Wang Z, Lou R, Wu J, Feng J. Integrated Analysis of Oncogenic Networks in Colorectal Cancer Identifies GUCA2A as a Molecular Marker. Biochem Res Int. 2019;2019:1–13.
DOI: 10.1155/2019/6469420
39. Chen Y, Zhu Y, Feng H, Liu Y, Qian J FY. Differential expression of guanylin in colorectal cancer. Zhonghua Wei Chang Wai Ke Za Zhi. 2009;12(5):515–7.
Available from: https://europepmc.org/abstract/med/19742348
40. Kulaksiz H, Rehberg E, Stremmel W, Cetin Y. Guanylin and Functional Coupling Proteins in the Human Salivary Glands and Gland Tumors. Am J Pathol. 2002;161(2):655–64.
DOI: 10.1016/s0002-9440(10)64221-6
41. Yokouchi H, Nishihara H, Harada T, Ishida T, Yamazaki S, Kikuchi H. Immunohistochemical profiling of receptor tyrosine kinases, MED12, and TGF-βRII of surgically resected small cell lung cancer, and the potential of c-kit as a prognostic marker. Oncotarget. 2016;8(24).
DOI: 10.18632/oncotarget.14410
42. Yang H, Zhang H, Zhong Y, Wang Q, Yang L, Kang H. Concomitant underexpression of TGFBR2 and overexpression of hTERT are associated with poor prognosis in cervical cancer. Sci Rep. 2017;7(1).
DOI: 10.1038/srep41670
43. Mi Y-J, Gao J, Xie J-D, Cao J-Y, Cui S-X, Gao H-J. Prognostic relevance and therapeutic implications of centromere protein F expression in patients with esophageal squamous cell carcinoma. Dis Esophagus. 2012;26(6):636–43.
DOI: 10.1111/dote.12002

44. Chen W-B, Cheng X-B, Ding W, Wang Y-J, Chen D, Wang J-H. Centromere protein F and survivin are associated with high risk and a poor prognosis in colorectal gastrointestinal stromal tumours. J Clin Pathol. 2011;64(9):751–5.
DOI: 10.1136/jcp.2011.089631
45. Cao J-Y, Liu L, Chen S-P, Zhang X, Mi Y-J, Liu Z-G. Prognostic significance and therapeutic implications of centromere protein F expression in human nasopharyngeal carcinoma. Mol Cancer. 2010;9(1):237.
DOI: 10.1186/1476-4598-9-237
46. Huang G-Q, Ke Z-P, Hu H-B, Gu B. Co-expression network analysis of long noncoding RNAs (IncRNAs) and cancer genes revealsSFTA1P and CASC2abnormalities in lung squamous cell carcinoma. Cancer Biol Ther. 2017;18(2):115–22.
DOI: 10.1080/15384047.2017.1281494
47. Zhang H, Xiong Y, Xia R, Wei C, Shi X, Nie F. The pseudogene-derived long noncoding RNA SFTA1P is down-regulated and suppresses cell migration and invasion in lung adenocarcinoma. Tumor Biol. 2017;39(2):101042831769141.
DOI: 10.1177/1010428317691418
48. Ma H, Ma T, Chen M, Zou Z, Zhang Z. The pseudogene-derived long non-coding RNA SFTA1P suppresses cell proliferation, migration, and invasion in gastric cancer. Biosci Rep. 2018;38(2):BSR20171193.
DOI: 10.1042/bsr20171193
49. Zhao W, Luo J, Jiao S. Comprehensive characterization of cancer subtype associated long non-coding RNAs and their clinical implications. Sci Rep. 2014;4(1).
DOI: 10.1038/srep06591