



PREDICTION OF BEST FEATURES IN HETEROGENEOUS LUNG ADENOCARCINOMA SAMPLES USING LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR.

ATEEQ MUHAMMED KHALIQ¹, SHARATHCHANDRA R G^{1*} AND MEENAKSHI RAJAMOHAN¹

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

ABSTRACT

This study aims to create a tumor heterogeneity-based model for predicting the best features of lung adenocarcinoma (LUAD) in multiple cancer subtypes using the Least Absolute Shrinkage and Selection Operator (LASSO). The RNASeq data of 533 LUAD 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. 89 and 444 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, 2033 genes were upregulated and 505 were downregulated in case of subtype-1 and 5039 genes were upregulated and 1219 were downregulated in case of subtype-2. LASSO model was established to predict the best features from each subtypes, 40 and 43 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. The study suggests that unsupervised clustering and Machine learning methods such as LASSO model-based feature selection can be effectively used to predict relevant genes which might play an essential role in cancer diagnosis.

KEYWORDS: *LASSO, Feature Selection, Heterogeneity, Biomarkers, Survival analysis, Expression patterns*



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: 28-10-2019

DOI: <http://dx.doi.org/10.22376/ijpbs/lpr.2019.9.4.L59-67>

INTRODUCTION

Lung cancer is reported to be the most deadly cancer¹ it is showing the worst survival rate when compared with colon, breast, and pancreatic cancers combined. According to the American Cancer society, it estimates both small cell and non-small cell lung cancer as the second most common cancer in both men and women. About 13% of all cancers are lung cancer. Lung Squamous cell carcinoma (LUSC) and Adenocarcinoma (LUAD) account for 15% and 85% of all lung cancer respectively². Lung cancer is a highly heterogeneous disease and the identification of cancer subtypes is decisive 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 resistance³. Therefore, deciphering LUAD heterogeneity will have a significant impact on 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. Many computer-based diagnostic and predictive 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. Previous studies indicate standard stepwise selection approaches which are not best for regression models with a vast number of covariates⁴. Alternatively, the least absolute shrinkage and selection operator (LASSO), has received much attention for the identification and selection of best variables. LASSO was introduced by Robert Tibshirani in 1996⁵. Regularization and feature selection are the two critical tasks LASSO performs. 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. In this study, we selected the best features using LASSO established model. We downloaded the RNASeq data for LUAD 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 also 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 insights into tumor heterogeneity and its importance in sample classification for predicting biomarkers of LUAD.

MATERIALS AND METHODS

Data source

The RNASeq data of LUAD, including 533 LUAD samples, and 59 normal samples were downloaded from the TCGA database (<https://portal.gdc.cancer.gov/>) in June 2019. All the raw, preprocessed data, images, supplemental tables and supporting files can be accessed at <https://github.com/AteeqKhaliq/LUAD/>.

Data preprocessing and grouping

533 Primary solid Tumor samples and 59 Solid Tissue Normal samples were downloaded from the TCGA database. We calculated a variance stabilizing transformation (VST) from the raw count data and transformed the counts yielding a matrix of values approximately homoscedastic.

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 $\text{MAD} > 0.5$ were selected from set 2 groups. Consensus clustering (CC)⁶ was used for the identification of subtypes on tumor samples. Silhouette width⁷ was used to validate sample clustering to its identified subtype when compared to other subtypes

Differential gene expression analysis

Differential gene expression (DGE) was assessed by using the DESeq2 package⁸ (Version 1.24.0, <https://bioconductor.org/packages/release/bioc/html/DESeq2.html>) on Subtype-1 and Subtype-2 samples when compared with normal samples. 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 Package⁹ (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 curves¹⁰ were generated by selecting the best predictors from individual subtypes. Kaplan-Meier curves were generated using the TRGAteD¹¹ (<https://github.com/ncborcherding/TRGAteD>) package implemented in R.

Quantification of the abundance of the immune and stromal cell population in Cancer Subtypes

We estimated the abundance of tissue-infiltrating

immune and non-immune stromal cell populations in Subtype-1 and Subtype-2 samples. MCP-counter¹² (<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 in tumor samples.

Gene classification and enrichment analyses

Cluster Profiler¹³ (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)¹⁴ was utilized to annotate genes to pathways, and Disease Ontologies.

Table description

Table 1
List of best-predicted genes by LASSO Model for cancer Subtype-1 and Subtype-2

LASSO best predicted genes in LUAD	
Subtype 1	Subtype-2
LYVE1	LINC01785
LINC00211	LINC02570
LAD1	HTR3C
IL22RA2	LINC02416
LINC00862	PLAC9P1
MED28P8	LINC01081
BCL9P1	LINC01996
TMPRSS4	HELT
LINC01028	NCAPGP2
NECTIN4	STARD13.AS
VPS9D1.AS1	LGR4
AWAT2	TESC
ST13P20	HBM
SLC25A48	LINC01506
PRR9	FGF10.AS1
CDC37P2	GPT2
DUXA	PACRG.AS3
S100A12	TOP2A
CD5L	OTC
TMEM51.AS1	EFNA3
DCST1	ZNF695
	STX11
	LGI3
	TEK

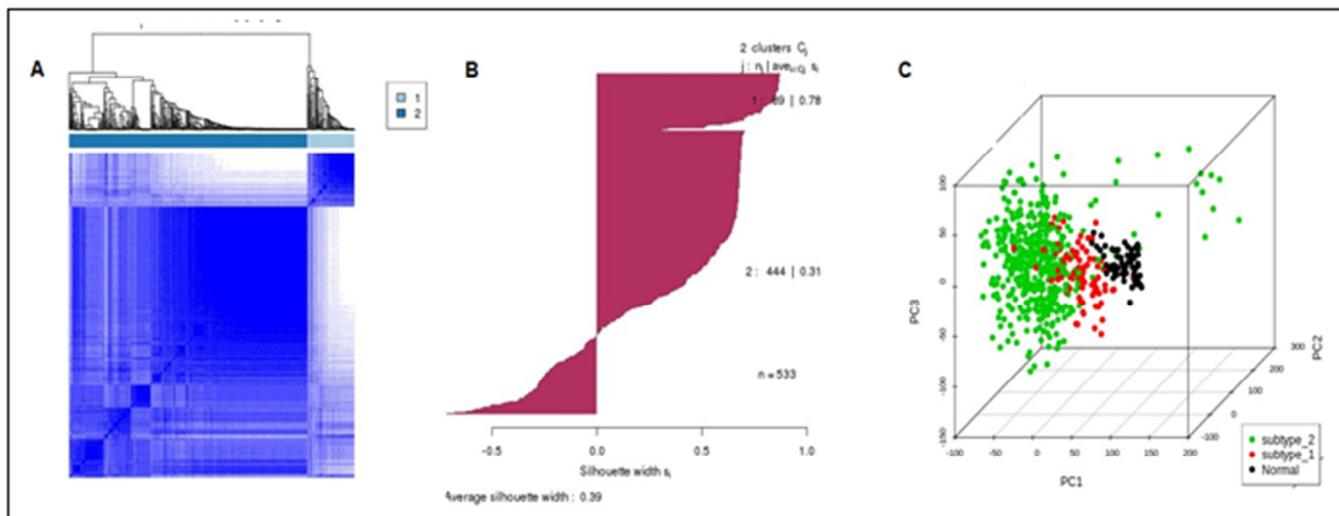


Figure 1

LUAD sample clustering. (A) CC plot shows the clustering of samples into two distinct subtypes. (B) Silhouette plot for validating the sample clustering. (C) PCA plot indicates distinct sample groups.

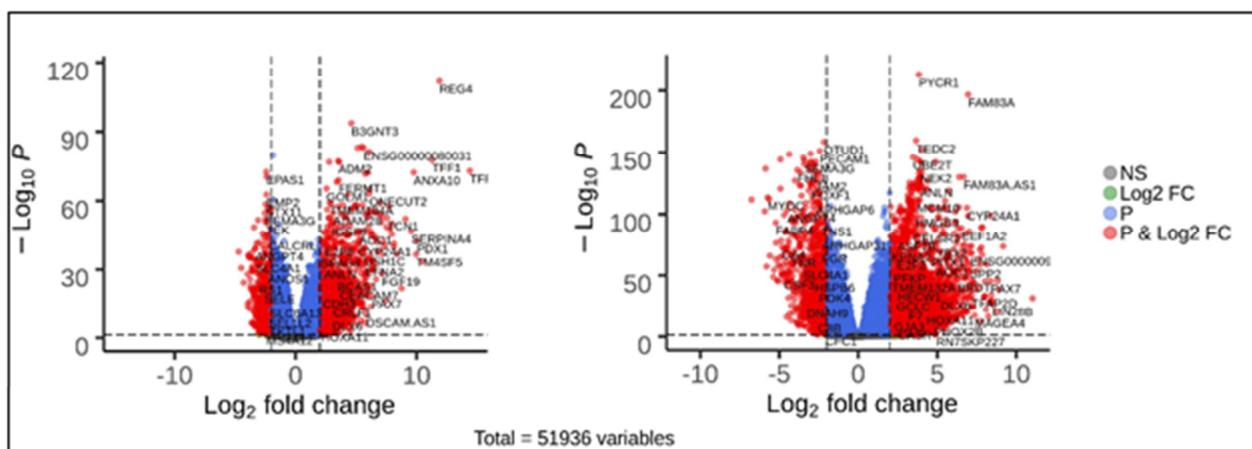


Figure 2

DGE analysis at standard cutoff of ± 2 fold change at p value of <0.05 . (A) DGE volcano plot for Subtype-1 samples. (B) DGE volcano plot for Subtype-2 samples.

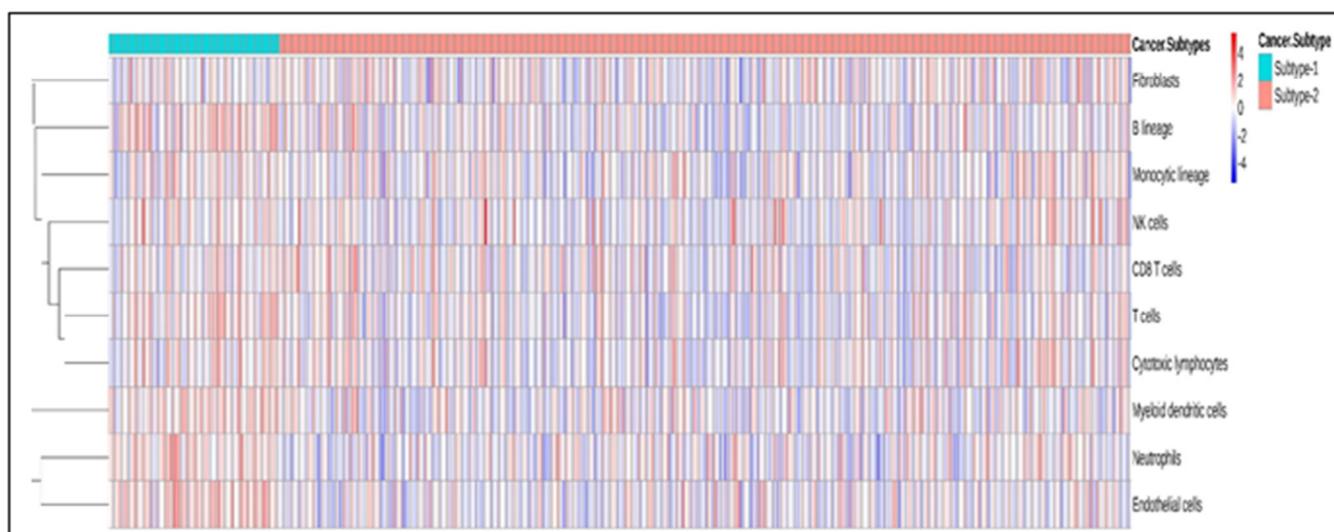


Figure 3

The heat map clearly differentiates subtype-1 and subtype-2 into two different categories based on tissue-infiltrating immune and non-immune stromal cell population.

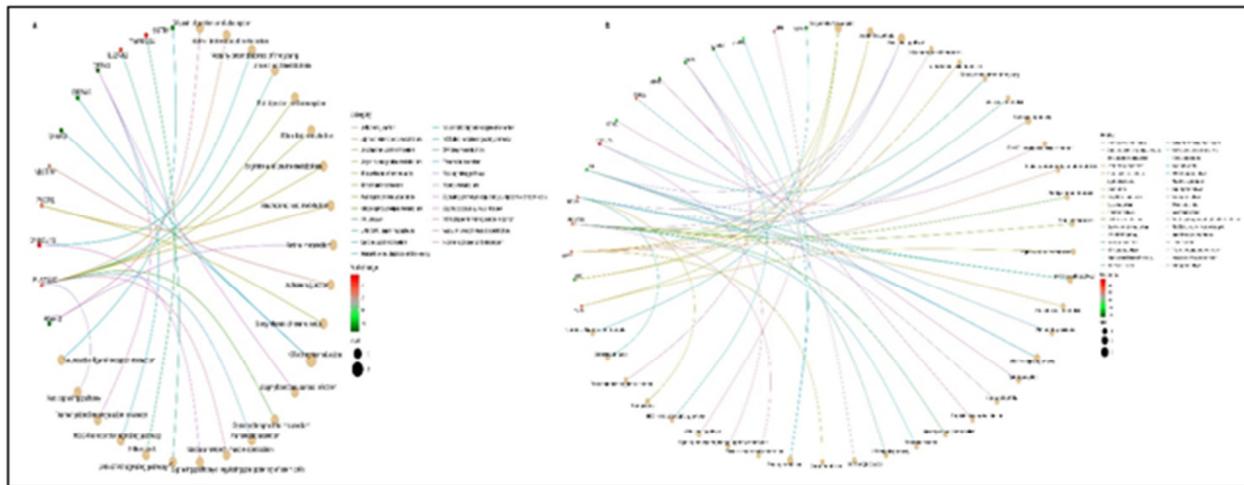


Figure 4

Pathway analysis for the best predicted genes by the LASSO model. (A) Subtype-1 LASSO predicted genes pathway analysis. (B) Subtype-2 LASSO predicted genes pathway analysis.

RESULTS

Cancer subtype identification in LUAD samples

We used the Consensus clustering (CC) method, an unsupervised clustering method for grouping subtypes in LUAD. CC method is the most widely used method 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, 89 and 444 samples were clustered in Subtype-1 and Subtype-2 respectively (Fig.1.A). We have validated consistency within clusters of data using the Silhouette width (Fig.1.B).

Identification of DEGs in Subtype-1 and Subtype-2 LUAD samples

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. 2033 genes were upregulated, and 505 were downregulated in the case of subtype-1 (Fig.2.A), and 5309 genes were upregulated, and 1219 were downregulated in the case of subtype-2 (Fig.2.B) shows differential expression pattern in subtype-1 and subtype-2. The DEGs in both subtypes were used for building the LASSO predictive model and for the identification of best predictor genes in LUAD heterogeneous cancer data.

LASSO model for identification of best predictive genes.

LASSO was first described by Robert Tibshirani in 1996. Regularization and feature selection are the two crucial tasks performed by LASSO. RNASeq

datasets are high dimensional datasets, with a 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 model. The result shows the trends of the 40 and 43 most relevant features selected by our model in subtype-1 and subtype-2 LUAD, respectively. The next step would be to find the most appropriate values for λ for our LASSO model. We analyzed the λ value using the 10 fold cross-validation, between λ_{\min} that gives minimum mean cross-validated error or $\lambda_{1\text{se}}$, that 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 LUAD. A list of best-predicted genes available for each cancer a subtype is shown in Table 1.

Analysis of the microenvironment of Subtype-1 and Subtype-2 LUAD samples

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. T-cell infiltrates, endothelial cells and fibroblasts are associated with a favorable outcome

and also poor prognosis in some cancer types¹⁵. To understand the immunological microenvironment in our expression subset-1 and subset-2 we used the MCP-counter method as described by Becht et al.¹². The estimations consist of single sample scores which are computed on each sample independently in two subtypes. The heat map shown in Figure 3 clearly distinguishes subtype-1 and subtype-2 into two different categories based on tissue-infiltrating immune and non-immune stromal cell populations. Subtype-1 shows an apparent increase in B lineage cells, monocytic cells, Cytotoxic lymphocytes, Natural killer cells, and CB 8 T 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 LUAD subtypes based on their inflammatory and stromal profiles, and Subtype-1 LUAD samples show increased expression of immunological markers than Subtype-2 samples.

Pathway analysis

Subtype-1 and Subtype-2 showed distinct and unique features which are involved in cancer progression. Genes such as IL22RA2, PLA2G2C identified in subtype-1 are up-regulated and found to be involved in canonical cancer regulatory pathways such as the JAK-STAT signaling pathway and RAS signaling pathway (Fig. 4). The PLA2G2C gene is associated with alpha-Linolenic acid as well as ether lipid metabolism, and both are known to play a role in cancer progression¹⁶. Our model identified the ONECUT1 gene in subtype1, which is associated with regulating pluripotency of stem cells and the DEFA3 gene which is associated with Transcriptional misregulation in cancer. The gene AWAT2 identified in subtype1 is found to be involved in Retinol metabolism, and studies suggest that retinoid signaling triggers tumor development¹⁷. The NECTIN4 gene highlighted in our study is associated with Adherens junction, which plays an essential role in cancer initiation and progression¹⁸. PYCR1 gene involved in proline metabolism which is highly correlated with cancer is also found to be differentially expressed in subtype1. EFNA3 is up-regulated in Subtype-2, and previous studies show that it contributes to the metastatic spread of breast cancer¹⁹. Our model predicted that increased expression of LGR4, TESC, TOP2A, and ZNF695 is suggestive of increased invasive and metastatic activity in Subtype-2²⁰. Decreased expression of FGF10 is seen in Subtype-2 and predicted by our model is

suggestive of dysplasia in LUAD subtype-2 samples²¹.

Validation with survival analysis

The LUAD samples were classified into two subtypes based on the consensus clustering method. Overall 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 LAD1, NECTIN4, SLC25A48, LYVE1, EFNA3, FGF10, HELT, HTR3C, and LGI3 yielded accurate predictions for the risk of LUAD and can be used in cancer prediction. Survival plots and its p-value is shown in Supplemental Figures.

DISCUSSION

In this study, we developed a LASSO based model for accurate feature selection in LUAD. Our model removed variables that are redundant and removed features which do not add any valuable information in disease prediction. Moreover in our study we are highlighting the Heterogeneity and cancer prediction in Lung adenocarcinoma and Squamous cell carcinoma. 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 treatment²². We used the 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 both Subtype-1 and Subtype-2. The number of differentially expressed genes was very high, and in these situations, it is difficult to predict the relevant variables. LASSO model was established around DE genes in Subtype-1 and Subtype-2 groups. Not all the expressed genes were relevant, and our model predicted the most relevant genes which were involved in disease progression. Decreased expression of LYVE1 and MED28P8 in Subtype-1 and FGF10, HELT, HTR3C, LGI3, PACRG, PLAC9P1, and STX11 in Subtype-2 showed worse overall survival in LUAD samples. Whereas increased expression of genes such as CDC37P2, DCST1, IL22RA2, LAD1, NECTIN4, SLC25A48, TMEM51.AS1, TMPRSS4, VPS9D1-AS1,

AWAT2, BCL9P1 and CD5L in Subtype-1 and EFNA3, LGR4, TESC, TOP2A and ZNF695 in subtype-2 showed decreased overall survival in LUAD samples. Long intervening noncoding RNAs (lncRNAs) are known to be critical regulators of numerous biological processes, and substantial evidence supports that lncRNA expression plays a significant role in tumorigenesis and tumor progression. Increased expression of LINC00862 in Subtype-1 samples correlates with worse survival in LUAD subtypes. Whereas, decreased expression of LINC00211 in Subtype-1 and LINC01506, LINC01785 and LINC01996 in Subtype-2 showed worse survival in LUAD samples. Our LASSO model predicts the most relevant and distinct genes from Subtype-1 and Subtype-2 samples which might be an important factor in cancer diagnosis and management. The best predictors for subtype-1 and subtype-2 from the LASSO model were found to be involved in several regulatory pathways. Subtype-1 gene such as S100A12 is a vital serum inflammatory marker and has been illustrated in several cancer types such as oropharyngeal squamous cell carcinoma and gastric cancers²³. Subtype-1 samples show increased expression of CD5L and TMPRSS4, which induces cancer stem cell-like properties and promotes malignant transformation by limiting lung epithelial cell apoptosis and promoting immune escape in NSCLC patients²⁴. Long noncoding RNA VPS9D1-AS1 overexpression in subtype-1 predicts poor prognosis and serves as a biomarker to predict the prognosis of NSCLC²⁵. Overexpression of nectin-4 oncoprotein, LYVE-1/PCAB, PROM2, and LAD1 are associated with poor overall survival in subtype-1 samples and can be considered as candidate serum and tissue biomarker as well as therapeutic target²⁶. Overexpression of Lgr4 in Subtype-2 samples promotes tumor aggressiveness may potentially become a novel target for cancer therapy²⁷. Subtype-2 samples show up regulation of TOP2A and ZNF695 is associated with a worse prognosis and induces overrepresentation of growth and proliferation pathways and can act as prognostic and predictive markers²⁸. Hypoxia-inducible oncogene EFNA3 is overexpressed in subtype-2 samples may play a critical role in the focal adhesion kinase (FAK) signaling and VEGF-associated tumor angiogenesis pathway¹⁹. Down regulation of tumor suppressor gene STX11 in

Subtype-2 samples predicts poor prognosis. Various studies indicate the role of STX11 expression in suppressing the proliferation of T-cell²⁹. Our model predicts lncRNAs such as LINC01506, LINC01785, LINC01996, LINC00862, and LINC02014 were expressed in subtype-1 and subtype-2 samples which can be considered as potential biomarkers and shows poor overall survival in LUAD. LncRNAs might be used as biomarkers and drug targets for early diagnosis, prognosis and personalized treatment of LUAD patients. Our study suggests that Consensus Clustering methods and LASSO combined will help us to develop a model with the most appropriate characteristics. Consistent with these findings, different subtypes showed distinct unique features which underscore the importance of sample grouping and assessment. Furthermore, Survival analyses validate that the survival time of the predicted genes correlates with gene expression pattern, which is recognizably different in both Subtypes, indicating that the LASSO model could effectively be used to overcome the feature selection problem and can be used for accurate prediction of risk in LUAD.

CONCLUSION

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 Lung Adenocarcinoma. With this model, we can identify the prognostic biomarkers of LUAD 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.

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;7:193. DOI: 10.3389/fonc.2017.00193
3. Bolck HA, Corrò C, Kahraman A, von Teichman A, Toussaint NC, Kuipers J. 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
4. 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
5. 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
6. 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>
7. 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
8. 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
9. 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
10. Kaplan EL, Meier P. Nonparametric Estimation from Incomplete Observations. *J Am Stat Assoc.* 1958;53(282):457–81. DOI: 10.1080/01621459.1958.10501452
11. Borcherding N, Bormann NL, Voigt AP, Zhang W. TRGAte: A web tool for survival analysis using protein data in the Cancer Genome Atlas. *F1000Research.* 2018;7:1235. DOI: 10.12688/f1000research.15789.2
12. 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.* 2016;130:95–190. DOI: 10.1016/bs.ai.2015.12.002
13. 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
14. 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
15. 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
16. Chamberland JP, Moon H-S. Down-regulation of malignant potential by alpha linolenic acid in human and mouse colon cancer cells. *Fam Cancer.* 2014;14(1):25–30. DOI: 10.1007/s10689-014-9762-z
17. Tang X-H, Gudas LJ. Retinoids, Retinoic Acid Receptors, and Cancer. *Annu Rev Pathol Mech Dis.* 2011;6(1):345–64. DOI: 10.1146/annurev-pathol-011110-130303
18. Vasioukhin V. Adherens Junctions and Cancer. *Subcellular Biochemistry.* Springer Netherlands; 2012. p. 379–414. DOI: 10.1007/978-94-007-4186-7_16
19. Gómez-Maldonado L, Tiana M, Roche O, Prado-Cabrero A, Jensen L, Fernandez-Barral A. EFNA3 long noncoding RNAs induced by hypoxia promote metastatic dissemination. *Oncogene.* 2014;34(20):2609–20. DOI: 10.1038/onc.2014.200
20. An X, Xu F, Luo R, Zheng Q, Lu J, Yang Y.

The prognostic significance of topoisomerase II alpha protein in early stage luminal breast cancer. *BMC Cancer.* 2018;18(1).
DOI: 10.1186/s12885-018-4170-7

21. El Agha E, Bellusci S. Walking along the Fibroblast Growth Factor 10 Route: A Key Pathway to Understand the Control and Regulation of Epithelial and Mesenchymal Cell-Lineage Formation during Lung Development and Repair after Injury. *Scientifica (Cairo).* 2014;2014:1–20.
DOI: 10.1155/2014/538379

22. 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

23. Huang G-Q, Ke Z-P, Hu H-B, Gu B. Co-expression network analysis of long noncoding RNAs (lncRNAs) and cancer genes reveals SFTA1P and CASC2 abnormalities in lung squamous cell carcinoma. *Cancer Biol Ther.* 2017;18(2):115–22.
DOI: 15384047.2017.1281494

24. de Aberasturi AL, Redrado M, Villalba M, Larzabal L, Pajares MJ GJ. TMPRSS4 induces cancer stem cell-like properties in lung cancer cells and correlates with ALDH expression in NSCLC patients. *Cancer Lett.* 2016;370(2):165–76.

Available from:
<https://www.sciencedirect.com/science/article/pii/S0304383515006436>

25. Tan J, Yang L. Long noncoding RNA VPS9D1-AS1 overexpression predicts a poor prognosis in non-small cell lung cancer. *Biomed Pharmacother.* 2018;106:1600–6.
DOI: 10.1016/j.biopha.2018.07.113

26. Huber S, Gagliani N, Zenewicz LA, Huber FJ, Bosurgi L, Hu B. IL-22BP is regulated by the inflammasome and modulates tumorigenesis in the intestine. *Nature.* 2012;491(7423):259–63.
DOI: 10.1038/nature11535

27. Tan B, Shi X, Zhang J, Qin J, Zhang N, Ren H. Inhibition of Rspo-Lgr4 Facilitates Checkpoint Blockade Therapy by Switching Macrophage Polarization. *Cancer Res.* 2018;78(17):4929–42.
DOI: 10.1158/0008-5472.can-18-0152

28. Hou G-X, Liu P, Yang J, Wen S. Mining expression and prognosis of topoisomerase isoforms in non-small-cell lung cancer by using Oncomine and Kaplan–Meier plotter. *PLoS One.* 2017;12(3):e0174515.
DOI: 10.1371/journal.pone.0174515

29. Yoshida N, Tsuzuki S, Karube K, Takahara T, Suguro M, Miyoshi H. STX11 functions as a novel tumor suppressor gene in peripheral T-cell lymphomas. *Cancer Sci.* 2015;106(10):1455–62.
DOI: 10.1111/cas.12742