Feature Selection and Classification Reveal Key lncRNAs for Multiple Cancers

Abdullah Al Mamun and Ananda Mohan Mondal
Florida International University, Miami, FL
Email: {mmamu009, amondal}@fiu.edu

Abstract

In the present study, a computational framework is developed to identify cancer specific key long non-coding RNAs (lncRNAs) using the lncRNA expression of cancer patients only. The framework consists of two state-of-the-art feature selection techniques - Recursive Feature Elimination (RFE) and Least Absolute Shrinkage and Selection Operator (LASSO); and five machine learning models - Naive Bayes, K-Nearest Neighbor, Random Forest, Support Vector Machine, and Deep Neural Network. For experiment, expression values of lncRNAs for 8 cancers from TCGA are used. Capability of these key lncRNAs in classifying 8 different cancers is checked by the performance of five classification models. This study identified 37 key lncRNAs that can classify 8 different cancer types with an accuracy ranging from 94% to 97%. Finally, survival analysis supports that the discovered key lncRNAs are capable of differentiating between high-risk and low-risk patients.

Introduction

Recent studies indicate that several cancer risk loci are transcribed into lncRNAs and these transcripts play key roles in tumorigenesis [1-3]. For example

- ***ANRIL*** for remodeling of chromatin
- ***H19*** for transcriptional co-activation and co-repression
- ***TERRA*** for protein inhibition
- ***MALAT1*** for post-transcriptional modifications
- ***PTENP1*** for decoy

Research on cancer classification and biomarker identification are rarely found due to the high dimensionality of the data. Many computational methods fail to identify a small number of important features, rather increase learning costs and deteriorates performance. To overcome this issue, researchers used feature selection algorithm for dimension reduction such as RFE (Recursive Feature Elimination) LASSO are used in various studies as a feature selection method.

We proposed a computational framework using feature selection and classification methods that can identify key lncRNAs and classify different cancers based on the expression value of those key lncRNAs. Important features or lncRNAs are selected in two steps: First, number of feature is reduced using a cutoff on expression values and then using a combination of two feature selection algorithms RFE and LASSO. This study discovered 37 key lncRNAs for eight different cancers. Then the capability of identified lncRNAs in classifying 8 different cancers is checked by the performance of five classification models. Finally, survival analysis is conducted to check whether the discovered lncRNAs are really capable of differentiating between high-risk and low-risk patients.

Data and Preprocessing

To validate the idea, we downloaded (April, 2019) RNA-seq FPKM normalized expression data from UCSC xena. Selection of this eight cancers is based on the number of samples (ranges from 309 to 585) to have a bal- anced dataset as shown in Table I. Combined dataset consists of 3656 patients with respect to 60483 RNAs representing 8 tumor types. The row represents the RNA Ids while column represents the sample Ids. Values of each cell represents the normalized read counts of a sample for that specific RNA sequence. In this study, we used lncRNA expression as classification features. The database TANRIC provides 12727 lncRNA Ids that has been used to extract expression of lncRNAs from the combined RNA expression data. Now we have 12309 common lncRNA Ids with expression data for all selected cancers. The combined number of expressed lncRNA applying this threshold is 4786.

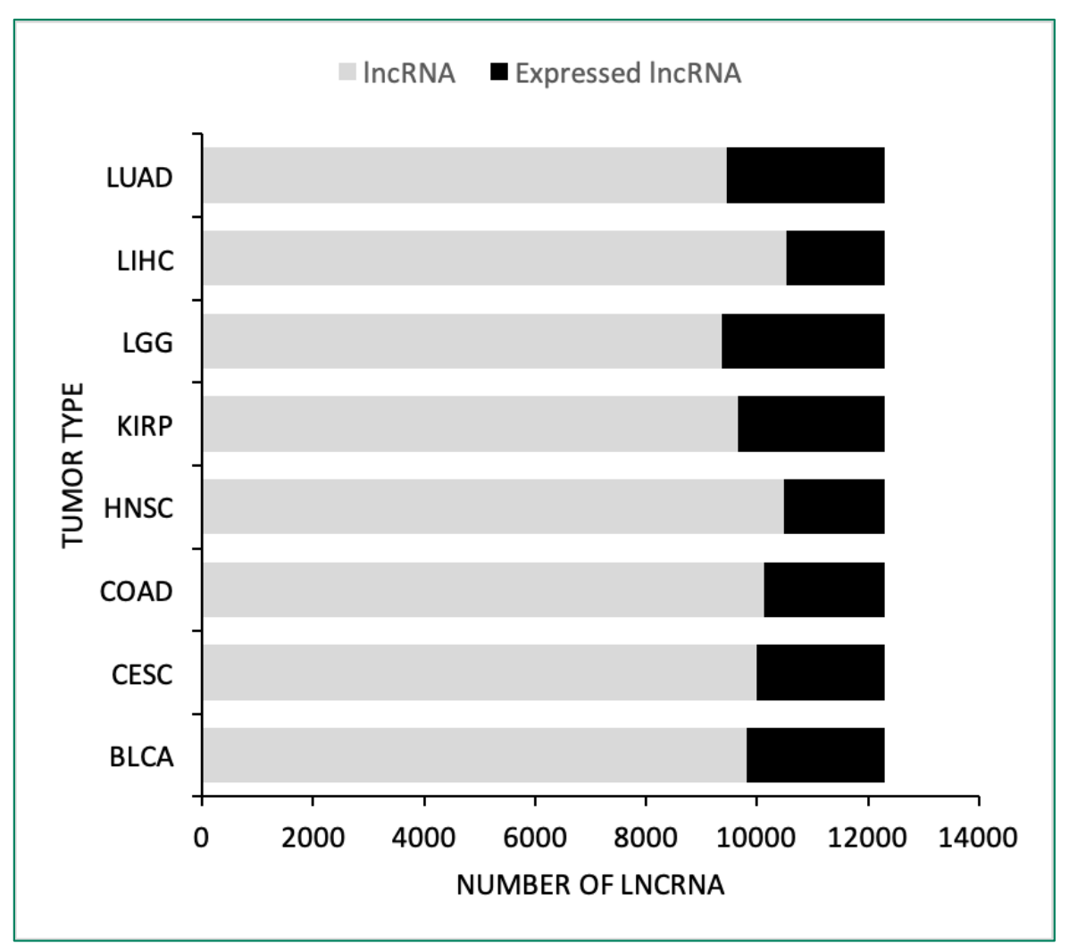


Fig. 1: Summary of TCGA RNA-seq data sets used in this study.

Methodology

Feature Selection

The lncRNAs those have more contribution towards the classification of cancer types are more likely to be the key lncRNA for cancer diagnosis and prognosis. Feature selection methods can reduce the number of irrelevant and noisy lncRNAs and select the most related lncRNAs to improve the classification results, which decrease the computational costs and improve the cancer classification performance

❑ **LASSO**

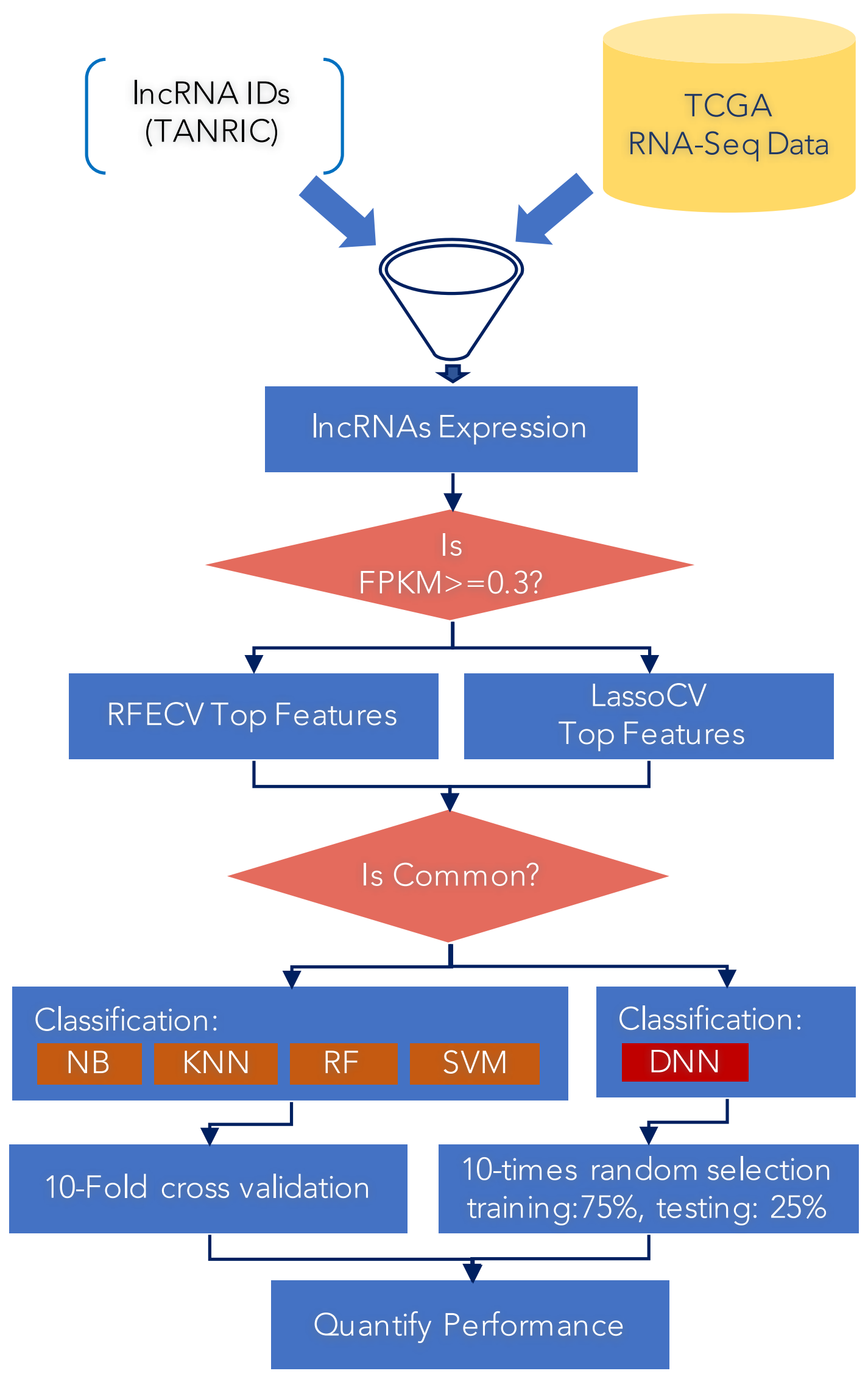
The Least Absolute Shrinkage and Selection Operator method applies a regularization (shrinking) process where it penalizes the coefficients of the regression variables and shrinks these to zero. The optimized λ = 0.0036 is calculated by 5-fold cross validation which is able to picked 765 important features in 62 secs with 96% accuracy.

❑ **RVECV**

Recursive Feature Elimination RFE algorithm constructs a ranking coefficient according to the weight vector *w* generated by an estimator e.g. linear regression during training. 786 respectively from 4786 features.

Classification

We used scikit-learn, a python library, for machine learning models. For KNN model, *k* was set to 7. In SVM, linear kernel is used. For the RF model, the number of estimator is 10 with entropy ensembling. Finally, Gaussian NB algorithm is used for Naive Bayes model. All models are executed on a CPU Intel core i7 with 16GB RAM.


lncRNA IDs (TANRIC)

TCGA RNA-Seq Data

lncRNAs Expression

Is FPKM>=0.3?

RFECV Top Features

LassoCV Top Features

Is Common?

Classification: NB KNN RF SVM

Classification: DNN

10-Fold cross validation

10-times random selection training:75%, testing: 25%

Quantify Performance

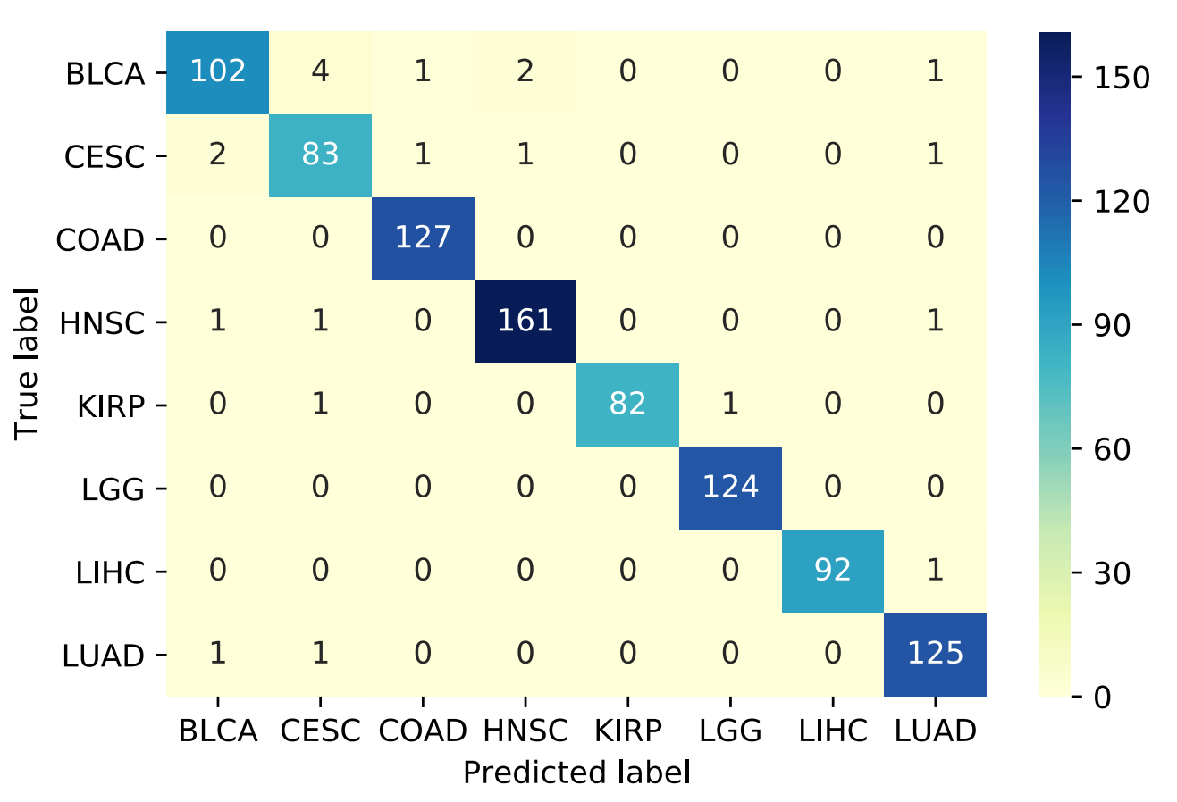
Fig. 2: Overall Process for Data Preparation and Methodology

Results

37 key lncRNA are identified and used for classification and prediction. The results obtained, 37 key lncRNAs, are validated using t-SNE plot and survival analysis.

Table 1: Key lncRNAs (*n* = 37)

AC000111.6, AC005082.12, AC005355.2, AC009299.3, AL450992.2, AP001626.1, BBOX1-AS1, CTA-384D8.31, EMX2OS, FAM182A, FENDRR, GATA3-AS1, H19, HAGLR, HOXA10-AS, HOXA11-AS, HOXD-AS2, KIZ, LINC00857, LINC00958, LINC01082, LINC01158, MIR205HG, NKX2-1-AS1, RP11-157J24.2, RP11-30K9.5, RP11-373D23.2, RP11-435O5.6, RP11-445O3.2, RP11-535M15.1, RP11-76C10.5, SFTA1P, TBX5-AS1, TMEM51-AS1, TP53TG1, UCA1, XIST



Fig, 3: Confusion Matrix of DNN Model

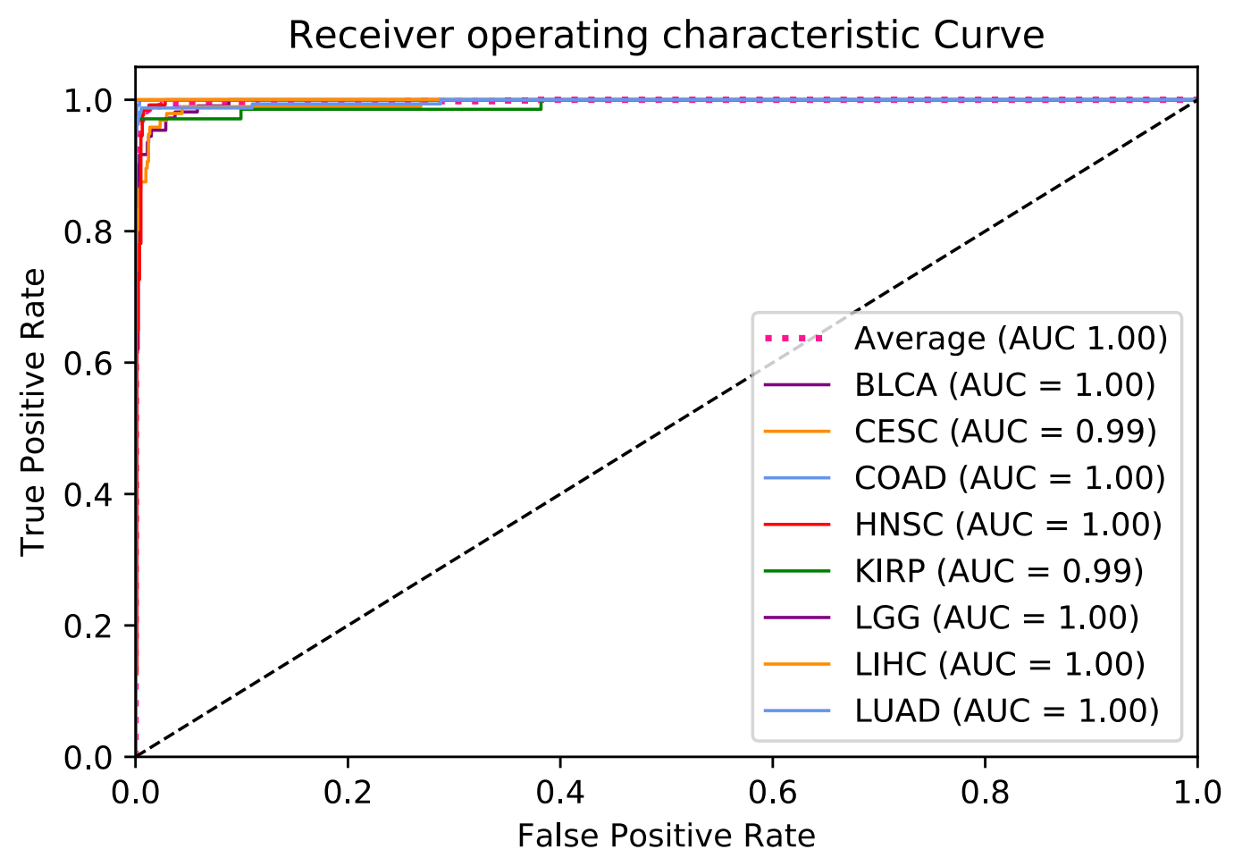


Fig. 4: ROC curve and AUC scores of different classes

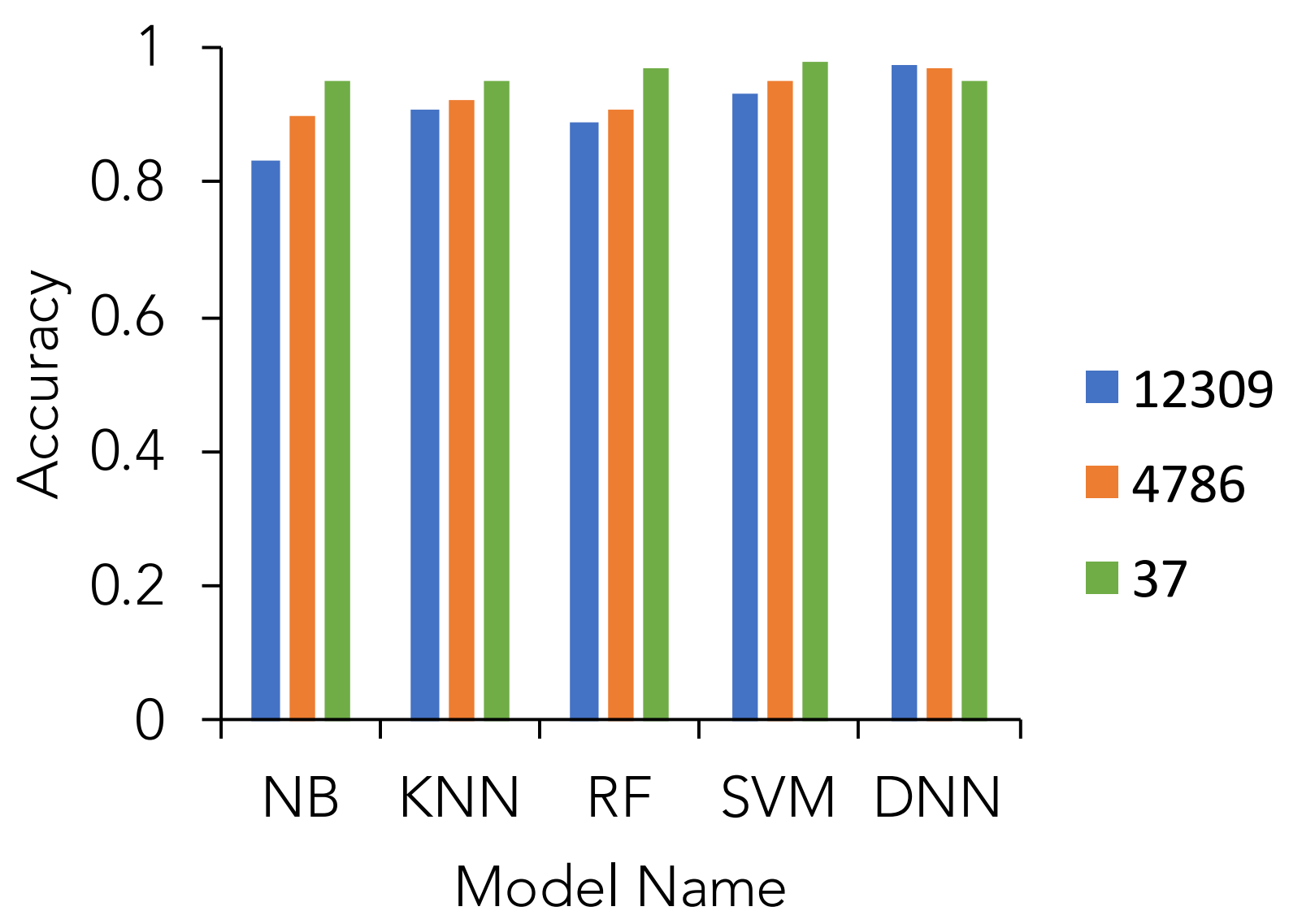


Fig. 5: Accuracy of different models with different number of features

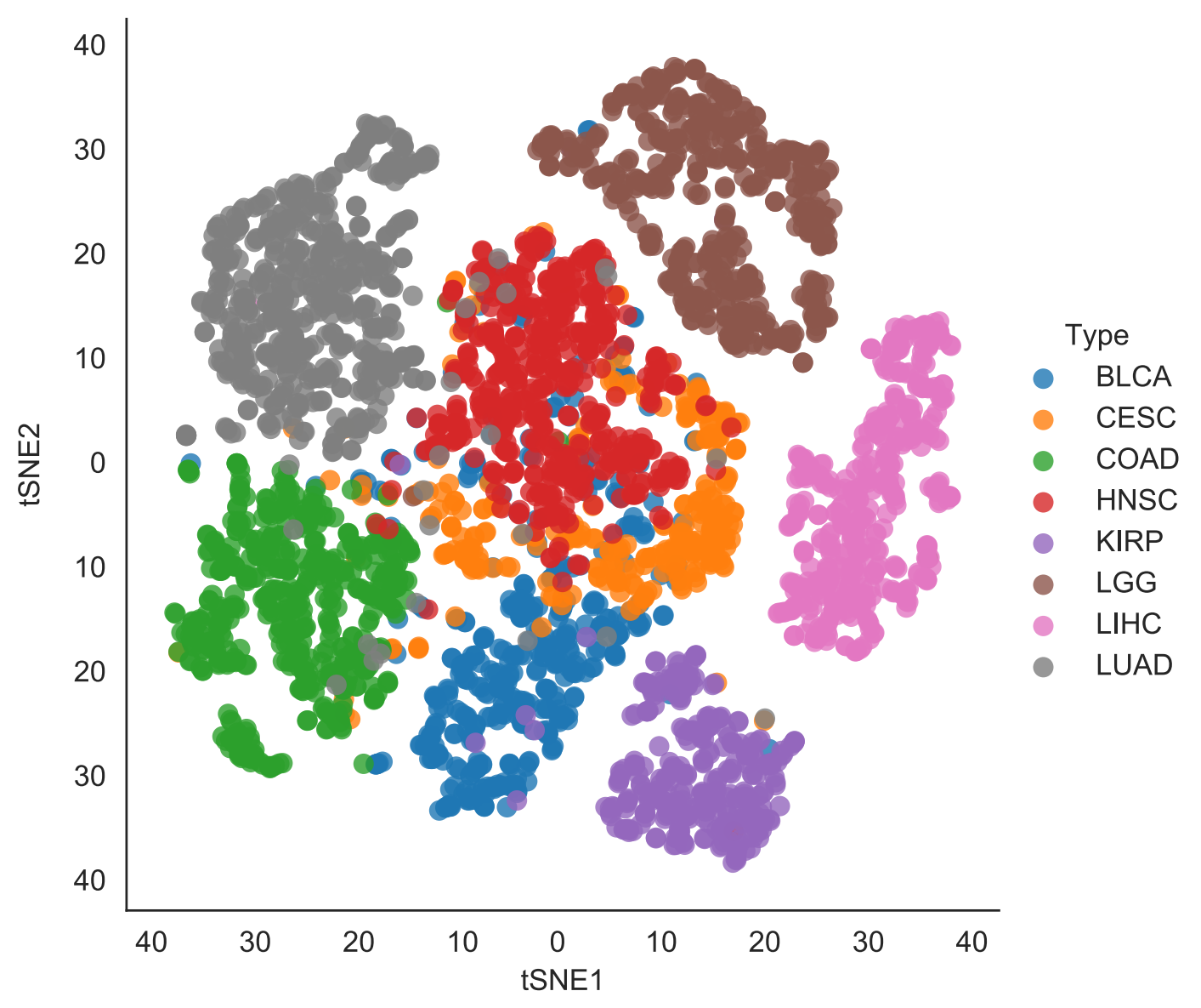


Fig. 6: tSNE representation of 37 lncRNA expressions for eight tumor types

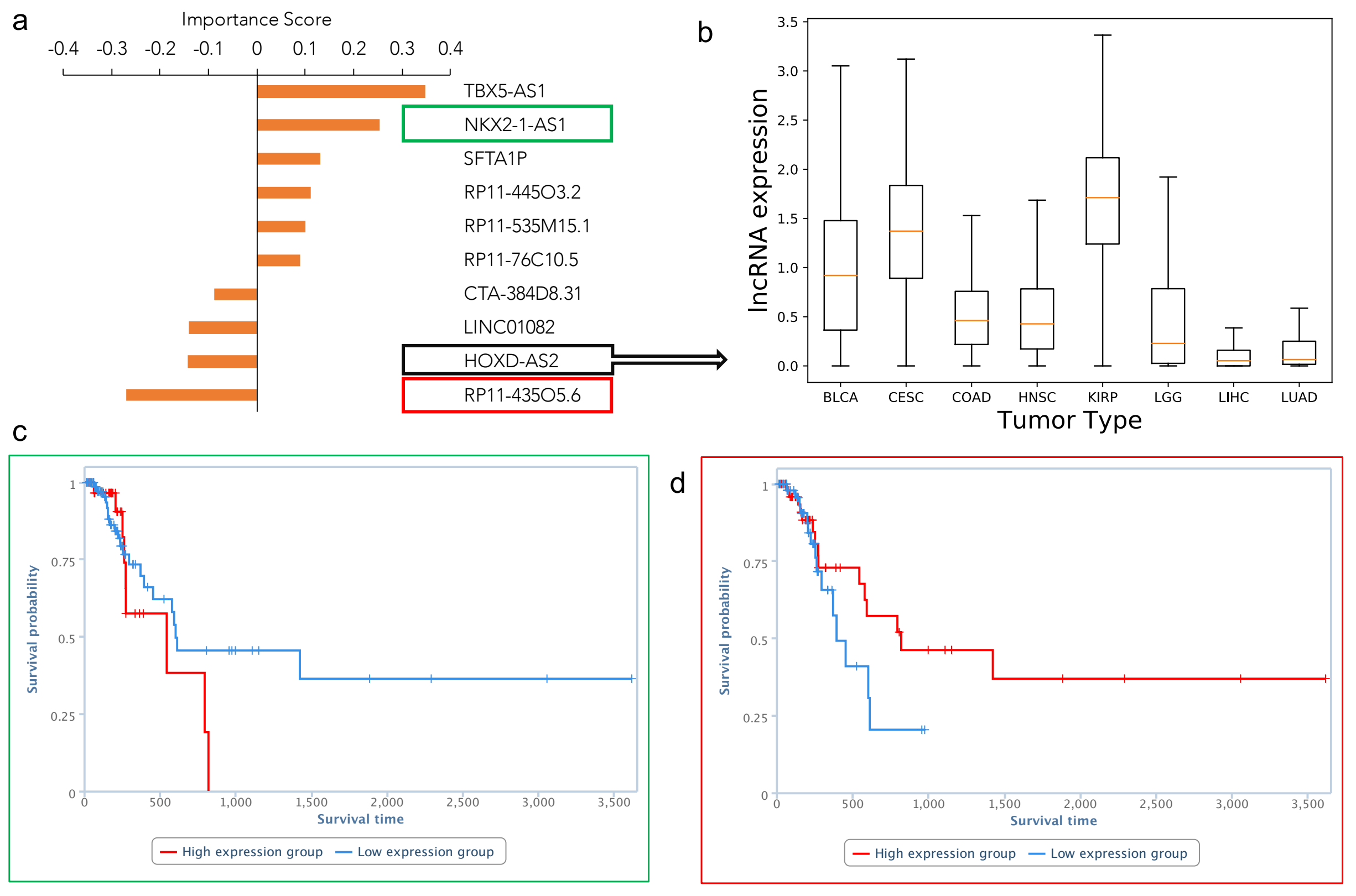


Fig. 7: Validation of discovered key lncRNAs. a) Top-10 lncRNAs with importance score by LASSO b) Box plot of expression values of lncRNA HOXD-AS2 for different cancers, c) Survival analysis using positively co-related lncRNA NKX2-1-AS1 in BLCA, and d) Survival analysis using negatively co-related lncRNA RP11-435O5.6 in BLCA. Survival Analysis is done using *TANRIC*.

Conclusion

A computational framework is developed to identify key lncRNAs for multiple cancers employing two feature selection and five classification methods using lncRNA expression of cancer samples only. This study identified 37 key lncRNAs that can classify 8 cancers with an accuracy ranging from 94% to 97%. t-SNE plot and survival analysis support that the discovered 37 lncRNAs are capable of differentiating 8 cancers as well as differentiating between high-risk and low- risk patients. Thus, the discovered lncRNAs can be used as diagnostic and prognostic features for 8 cancers considered in this study. In the extended version of the paper, we plan to compare the discovered list of lncRNA with the existing literature if available. Another extension of this paper could be inclusion of lncRNA expression of corresponding normal samples.

Acknowledgements

This research is partially funded by NSF CAREER award #1651917 (transferred to #1901628) to AMM.

References

[1] SW Cheetham, F Gruhl, JS Mattick, and ME Dinger. Long noncoding rnas and the genetics of cancer. *British journal of cancer*, 108(12):2419, 2013.
[2] Yiwen Fang and Melissa J Fullwood. Roles, functions, and mechanisms of long non-coding rnas in cancer. *Genomics, proteomics & bioinfor- matics*, 14(1):42–54, 2016.
[3] YKotake,TNakagawa,KKitagawa,SSuzuki,NLiu,MKitagawa,and Y Xiong. Long non-coding rna anril is required for the prc2 recruitment to and silencing of p15p15 ink4b tumor suppressor gene. *Oncogene*, 30(16):1956, 2011.

Research Day, School of Computing and Information Sciences, FIU, Miami, FL, October 25, 2019