Oncology 2025
The Cancer Control Challenge:
Better Understanding It to Better Face It

APPLICATION OF RANDOM FOREST AND LASSO ALGORITHMS FOR BIOMARKER DISCOVERY IN NON-SMALL CELL LUNG CANCER SUBTYPES

Camila Tavares Carvalho Uchôa[1], Lorena Duarte Fernandes[1], Sérgio Augusto Antunes Ramos[1], Diego Pereira[1], Valéria Cristiane Santos da Silva[1], Jessica Manoelli Costa da Silva[1], Ronald Matheus da Silva Mourão[1], Samir Mansour Casseb[1], Paulo Pimentel Assumpção[1], Fabiano Cordeiro Moreira[1]

[1]Universidade Federal do Pará (UFPA)

Introduction: Lung cancer is the leading cause of cancer-related deaths worldwide and represents a significant public health challenge. In Brazil, particularly in the northern region, lung cancer ranks as the second or third most common tumor among men. An increasing incidence of lung adenocarcinoma (LUAD) compared to lung squamous cell carcinoma (LUSC) has been observed, altering the epidemiological profile of the disease. These histological subtypes are classified as non-small cell lung cancer (NSCLC) and require distinct therapeutic strategies, highlighting the critical need for accurate differential diagnosis. **Objectives:** This study aimed to apply machine learning algorithms to identify potential gene expression biomarkers capable of distinguishing LUAD from LUSC, with a view toward future clinical applications. **Methods:** The UFPA-Cohort utilized a dataset comprising 18 NSCLC samples, including 6 LUSC and 12 LUAD samples, obtained from surgical procedures performed on patients at Hospital Universitário João de Barros Barreto (HUJBB). The Research Ethics Committee approved this study under protocol number CAAE: 41667021.4.3001.0017. The samples were sequenced on an Illumina NextSeq 500/550 platform using mRNA-seq for gene expression analysis. RNA-seq data processing was performed using the nf-core/rnaseq v3.1.4 pipeline, referencing human genome coding transcripts from GENCODE v43. Additionally, the TCGA-NSCLC dataset was used for transcriptome analysis to validate the results of the UFPA-Cohort. Differential gene expression analysis was performed on the TCGA-NSCLC and UFPA-Cohort datasets using the DESeq2 package in R. Genes were classified as upregulated or downregulated based on the threshold $|\log_2$ fold-change (FC)| > 2.0 and FDR < 0.05. After data preprocessing and normalization of gene expression levels, feature selection was performed using Random Forest and LASSO logistic regression models, combined with cross-validation strategies to enhance result robustness. **Results:** RNA-seq data analysis of the UFPA-Cohort revealed 535 differentially expressed genes between LUAD and LUSC tumor samples. A total of 233 genes were upregulated and 302 genes were downregulated. TCGA-NSCLC data identified 1229 differentially expressed genes, with 453 upregulated and 776 downregulated genes. Principal component analysis (PCA) highlighted a distinct separation between histological subtypes in the UFPA-Cohort and TCGA-NSCLC dataset. The heatmap demonstrated differences in the gene expression patterns between the histological subtypes. Nineteen genes were selected by training

the TCGA-NSCLC dataset using the Random Forest and LASSO algorithms and subsequently tested using Random Forest on the UFPA-Cohort data. When evaluated individually, *CALML3* and *IRF6* exhibited the highest importance scores according to the Random Forest analysis and were validated by ROC curve analysis (AUC > 0.85). **Conclusion:** Our findings suggest that *CALML3* and *IRF6* are promising diagnostic biomarkers for distinguishing LUSC from LUAD. Furthermore, the results highlight the potential of machine-learning-based biomarker discovery to improve the differential diagnosis of NSCLC.

Keywords: Lung cancer; random forest; lasso