Title: Gene expression and metadata based identification of key genes for lung cancer, COPD, and IPF using machine learning and statistical models

Author names: Mst. Farjana Yasmin ,Md. Faruk Hosen,Md. Abul Basar ,Anichur Rahman ,Mahedi Hasan,Fahmid Al Farid ,Hezerul Abdul Karim ,Abu Saleh Musa Miah

Abstract: Lung cancer (LC) is one of the most prevalent and deadly cancers globally, presenting a major public health challenge. Patients with chronic obstructive pulmonary disease (COPD) and idiopathic pulmonary fibrosis (IPF) are at a significantly higher risk of developing lung cancer. Despite developments in research, the primary molecular pathways of many disorders remain poorly understood. The current study aimed to identify potential therapeutic genes for lung cancer (LC), chronic obstructive pulmonary disease (COPD), and idiopathic pulmonary fibrosis (IPF) through machine learning (ML) and bioinformatics methodologies. The differentially expressed genes (DEGs) were identified across three datasets utilising DESeq2 and limma, and the common genes among the DEGs from these datasets were subsequently selected. The protein-protein interaction (PPI) networks were generated utilising STRING, and major hub genes were discerned via topological analysis. The Key hub genes, such as ETS1, MSH2, RORA, and PMAIP1, were detected. The pathways named KEGG and cancer pathway studies were conducted to evaluate their contributions to disease processes. The research included network-based methodologies, including transcription factors, GO keywords, gene–miRNA relationships, and survival data analyses, to further narrow the list of differential genes linked to LC, COPD, and IPF. The metadata for hub genes was aggregated from prior studies to integrate earlier discoveries. In the end, four key candidate genes (ETS1, MSH2, RORA, and PMAIP1) were found by intersecting the common differentially expressed genes, hub genes, major module genes, and meta-hub genes. The outcomes present a solid framework for subsequent research and therapy strategies for LC, COPD, and IPF. The potential drug compounds targeting the identified key genes are proposed, offering new avenues for the development of treatment.

- https://doi.org/10.1371/journal.pone.0344666