Identification of blood-derived key hub gene biomarkers associated with chronic obstructive pulmonary disease using a comprehensive bioinformatics and machine learning approach

Mohanraj Mani, Mohanapriya Arumugam

Department of Biotechnology, School of Biosciences and Technology,

Vellore Institute of Technology, Vellore-632014

Email: mohanraj.m2023@vitstudent.ac.in

Abstract

Chronic obstructive pulmonary disease (COPD) is a heterogeneous lung disease with chronic respiratory symptoms and non-reversible airflow limitation, currently the third leading cause of death. While COPD is not curable, it can be effectively managed. Validated biomarkers are required for patient assessment, risk prediction, treatment guidance, and response evaluation, given the disease's complexity. This study aims to identify key blood-derived hub gene biomarkers associated with COPD using bioinformatics and machine learning. Whole-blood sample datasets were retrieved via the Gene Expression Omnibus (GEO) database. The dataset was downloaded with GSE100153 (19 COPD and 24 control samples) as the training set and GSE146560 (8 COPD and 8 control samples) as the validation set. The training dataset included a Differentially Expressed Genes (DEGs) analysis to compare gene expression levels between COPD and control samples. Weighted Gene Co-Expression Network Analysis (WGCNA) was used to screen for COPD-related module genes. Functional Enrichment Analysis including of Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis performed to candidate genes. The key hub gene biomarkers were further verified using machine learning (ML) algorithms. Finally, validating the key hub genes by immune cell infiltration analysis and receiver operating characteristic (ROC) curve analysis with GSE100153 (training set) and GSE146560 (validation set).

Keywords: Chronic obstructive pulmonary disease (COPD), biomarkers, machine learning (ML)