[BIOINFO 2024]

Development of AI model for classifying immune phenotypes in lung cancer based on a novel approach utilizing whole transcriptomics

Ki Wook Lee[1], Hyun Woo Park[1], Hye Jung Min1 Ji Min Seo[1], Na Young Kwon[1], Ji Min Seo[1],
Heeje cho[1], Balachandran Manavalan[1], Sehhoon Park[2] and Young-Jun Jeon[1]*

[1] *Department of Integrative Biotechnology, Sungkyunkwan University*
[2] *Department of Medicine, Samsung Medical Center*
**Corresponding author: jeon2020@skku.edu*

Despite the widespread application of immune checkpoint inhibitors (ICIs) as a standard therapy for advanced non-small-cell lung cancer (NSCLC), a reliable method for predicting patient response prior to treatment initiation remains elusive. Given their critical role in antitumor immunity, the infiltration of tumor-infiltrating lymphocytes (TILs) within the tumor microenvironment has emerged as a promising biomarker. However, current approaches that quantify TILs using hematoxylin and eosin (H&E)-stained whole-slide images (WSI) are labor-intensive and subject to observer bias. Moreover, these image-based immune phenotyping techniques are limited by the spatial constraints of tumor sampling, which impedes the ability to capture the full immune signature of the tumor.

In this study, we developed an artificial intelligence (AI)-based immune phenotyping model utilizing transcriptomic data. Leveraging 449 lung adenocarcinoma (LUAD) transcriptomic profiles from The Cancer Genome Atlas (TCGA) and immune phenotypes classified via LUNIT Scope, we partitioned the data into training and test sets with an 80:20 ratio. In contrast to the conventional approach of limiting the number of genes using the DEG method, we identified AI-based important genes (AIGs) through six tree-based machine learning classifiers applied to the entire gene expression dataset. Based on these AIGs, we developed immune phenotyping models to classify samples as either 'immune infiltrated' (IF) or 'non-infiltrated' (non-IF) using 15 different machine learning (ML) and deep learning (DL) algorithms, ultimately refining the model through an ensemble strategy. This novel approach produced a final model that achieved an AUC of 1.0 in the training set and 0.86 in an independent external test set.