Identifying miR-331-3p as a Unique Blood-Based Biomarker for Lung Adenocarcinoma through Random Forest Classification

Moore, Madden (School: Academies of Loudoun)

Though early cancer detection is vital to patient survival, most lung cancers are diagnosed late due to the lack of a routine screening test. While blood miRNAs hold potential for this purpose, combined miRNA expressions in blood including those from all organs poses a challenge. To minimize signals from other organs, ideal blood miRNA biomarkers would be those also found in lung tumors. As differential expression analysis alone has not successfully found miRNA biomarkers common to lung tumors and blood, I used machine learning to identify them. Random forest models were used to extract highly predictive microRNAs for lung adenocarcinoma in both tissue and blood samples. Using all miRNAs as inputs, 51 miRNAs in tissue and 22 miRNAs in blood were identified as highly predictive across all 10 random forest models, of which 3 miRNAs were found in both. Using the 51 and 22 miRNAs as random forest inputs in tissue and blood, respectively, the average AUC-ROC increased compared to using all miRNAs. While many of the target genes of these overlapping miRNAs have already been linked to lung adenocarcinoma, many miR-331-3p targets were identified in enriched KEGG pathways of non-small cell lung cancer. miR-331-3p was not statistically significant in differential expression ($p = 0.067$) and has not been linked to lung adenocarcinoma before but its high predictive power in random forests for tissue and blood samples make it a unique, blood-based biomarker for lung adenocarcinoma. These results show the viability of random forests for identification of new microRNA diagnostic targets for early cancer detection.