**Introduction**

Interstitial lung disease (ILD) is a disorder characterized by progressive scarring and inflammation of the lung. ILD may be caused by environmental triggers such as prolonged exposure to asbestos, an autoimmune disorder such as rheumatoid arthritis (RA), or it may be idiopathic. Currently, the gold standard for ILD diagnosis and subtype classification is high resolution CT (HRCT) scan. ILD subtypes are often classified by the various different appearances (or phenotypes) that their corresponding imaging scans may take. However, appropriately reading CT scans, identifying ILD and classifying ILD subtypes requires specialist expertise and training. Even among pulmonary specialists, many physicians are ill equipped to recognize and interpret subtle differences on imaging scans between ILD phenotypes, often leading to misdiagnosis and inappropriate treatment.

Development of an accurate serum biomarker assay to differentiate between ILD phenotypes could improve diagnostic accuracy for ILD and lead to better treatment, particularly for populations and in geographic areas where ILD specialists are inaccessible. The objective of this project is to identify protein signatures that differentiate ILD phenotypes.

**Study Design**

In this study, the investigators enrolled 160 patients with either idiopathic pulmonary fibrosis (IPF) or rheumatoid arthritis ILD (RA ILD), including patients with fibrotic nonspecific interstitial pneumonia (NSIP), cellular NSIP, and chronic organizing pneumonia (COP) patterns. The investigators also enrolled healthy volunteers and subjects with rheumatoid arthritis, but no ILD for comparison. Blood samples were taken from each patient, and the transcription levels of 1321 proteins with known targets were measured. Demographic and body size variables were also collected for some subjects. Some subjects also have lung function data available (FEV1, FVC, FEV1/FCV ratio, and TLC).

**Key Questions**

The goal of this project is to identify sets of proteins that are differentially expressed between the following ILD phenotypes: RA ILD vs. RA no ILD, IPF vs. healthy, RA ILD vs. IPF, RA fibrotic vs RA cellular. (You do not have to look at all of these comparisons, but these are the most clinically meaningful comparisons that you can make). Please give consideration (but do not limit yourself to) the following questions:

* What proteins are differentially expressed between ILD phenotypes?
* How many of these differentially expressed proteins would you expect to be false positives (think of your significance level)? Can you reduce the false positive rate? (hint: consider corrections for multiple hypothesis testing)
* Are there differences in demographic, body size, or lung function variables between groups?
* Does adjustment for relevant demographic, body size, or lung function variables change your protein signature profiles?

*Some useful hints:*

* You will need to find ways to scale your computations so that you are running your analyses on *all* of your proteins at the same time. Think of mutate\_at() and map() functions.
* Check out the `limma` package which was designed for this kind of analysis!