AI4ALL 2021 Project: Neuroimmune phenotype in COVID-19 patients

Lead TA: Jiapei Chen

Elevator pitch

Use supervised machine learning models to predict neuroimmune conditions from cerebrospinal fluid patient samples, and explore unsupervised learning in dimension reduction and clustering to gain in-depth understanding of the neuroimmune landscape in COVID-19 patients.

Abstract

Up to 57% of patients suffering from COVID-19 can develop neurological symptoms, such as headache, neuroinflammatory or cerebrovascular disease. These conditions are more frequent in patients with severe COVID-19. Despite the prevalence and severity of these neurological conditions, the underlying mechanisms remain obscure. (ref: Neurological associations of COVID-19)

Questions to address:

1. Which cell type best predicts different neuroimmune conditions (including COVID-19) in our dataset?
2. Which supervised machine learning model has the best performance in predicting neuroimmune conditions based on gene expression data? For the best performed model, is its decision making process transparent? If so, can we find which gene expression values helped the model to predict?
3. Using unsupervised clustering, how are cell type compositions different among the neuroimmune conditions? Select a few genes you've examined in question 2, and examine how their expression pattern looks across different cell types and neuroimmune conditions - is this what you expect to see?

Database

We'll be using a published dataset from Heming et al, "Neurological Manifestations of COVID-19 Feature T Cell Exhaustion and Dedifferentiated Monocytes in Cerebrospinal Fluid" (Paper from *Immunity* 2021)

- Raw data can be downloaded at: GEO Accession viewer
    - Raw data consist of cell barcodes, a list of features (i.e. genes), a count matrix, and annotations (patient ID, diagnosis, cell type) from single-cell RNA sequencing of cerebrospinal fluid samples
    - Samples come from 8 Neuro-COVID patients, 9 idiopathic intracranial hypertension (IIH) patients, 9 multiple sclerosis (MS) patients, and 5 viral encephalitis (VE) patients
- I've done the following preprocessing steps from raw data:
    - Pair each example (i.e. cell) with its corresponding label (i.e. annotation, like diagnosis)
    - Generate normalized gene expression values for monocytes (mono2), T regulatory cells (treg) and CD4+ T cells (cd4). This data can be found in my GitHub repository here.

Key Learning Opportunities

Students will be exposed to concepts of supervised and unsupervised learning, disease prediction, dimensionality reduction, and cell type classification. Practically, students will use different machine learning packages from Python's sklearn (e.g. decision tree, logistic regression, random forest etc) to train and predict neuroimmune diseases based on gene expression data. In addition, students will implement the Seurat package on R, a different but extremely useful coding language (!), and explore unsupervised learning.

Deliverables

- Train, test, and evaluate different supervised machine learning models on disease prediction (COVID, IIH, MS, or VE) from single-cell RNA sequencing data
    - Bonus: use models to predict COVID disease severity (which 3 patients had severe symptoms?)
- Unsupervised clustering using PCA-based methods (UMAP or tSNE) to identify different cell types
- Explore differences in cell type composition and gene expressions among neuroimmune conditions
    - Bonus: cell type classification (CelliD) using established marker gene sets