Overview

Single cell RNA sequencing (scRNA-seq) is a recently developed technology platform used to measure gene expression from thousands of individual cells. scRNA-seq is transforming biology and medicine as it enables studies of cell-heterogeneity, identification of cell types and subpopulations in complex tissues and measuring cell transformations during development. However, scRNA-seq also has a number of analysis challenges commonly encountered in data science. One of the main challenges is that the read count matrix (genes x cells) is sparse in that it contains many zeros due to technological limitations (e.g. only a thousands genes have count data from an individual cell even though many more genes are expressed in that cell). For an overview of how scRNA-seq data is obtained and best practices for scRNA-seq analysis see Huemos et al. (2023, PMC10066026).

The goal of this hackathon is to identify doublets in scRNA-seq data. Doublets or multiplets are cases where gene expression count data is generated from two or more cells rather than a single cell. This occurs when a unique molecular identifier (UMI) is associated with two cells in a liquid droplet rather than one, resulting in a mixture of expression from both cells. There are a variety of ways in which doublets can be detected, most often by cells that appear to show a mixture of expression patterns from two different cell types. Doublets involving the same cell type are much more difficult to detect. One's ability to detect doublets depends on how one deals with sparse data, normalization and cell type clustering. For a review of methods see Xi and Li (2021, PMC7897250). Accurate detection of doublets remains a significant challenge and no single method outperforms all others.