Figure Vignette

Courtney Schiebout, H. Robert Frost

Load Libraries

Libraries "CAMML" (Schiebout and Frost 2022) and "Seurat" (Satija et al. 2015) need to be loaded to carry out this vignette, in addition to several other libraries for data processing and gene set development (Robinson, McCarthy, and Smyth 2010; **orghs?**; Liberzon et al. 2011). Packages will also load additional libraries they depend on.

```
library(CAMML)
library(Seurat)
library(edgeR)
library(org.Hs.eg.db)
library(msigdbr)
```

Data Processing

The following code outlines how the scRNA-seq data from Hao, et al. (2021), held in Gene Expression Omnibus (GEO) at GSE164378 was reduced to ~57,000 cells for this analysis using Seurat (Satija et al. 2015; Hao et al. 2021; Edgar, Domrachev, and Lash 2002). Note: this code requires substantial computing power.

```
#read in data
seurat <- Read10X(data.dir = c("GSE164378_RAW/"))

#create object
seurat <- CreateSeuratObject(counts = seurat, project = "GSE164378",
                             min.cells = 100 , min.features = 500,
                             num.var.features = 2000)

#filtering steps
seurat[["percent.mt"]] <- PercentageFeatureSet(seurat, pattern = "^MT-")
seurat = subset(seurat, subset = nFeature_RNA > 200 & nFeature_RNA < 5000
                & percent.mt < 5)

#normalize, scale, and cluster data
seurat <- NormalizeData(seurat)
seurat <- FindVariableFeatures(seurat, selection.method = "vst", nfeatures = 2000)

all.genes <- rownames(seurat)
seurat <- ScaleData(seurat)

seurat <- RunPCA(seurat)
seurat <- FindNeighbors(seurat, dims = 1:30)
seurat <- FindClusters(seurat, resolution = 0.25)

seurat <- RunUMAP(seurat, dims = 1:30)
```