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; Carlson 2020; 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)

#save data
saveRDS(seurat, file = "haod.RDS")
```

scRNA-seq and CITE-seq Compilation

Following the data filtering, the CITE-seq data needs to be added as an additional assay in the Seurat Object (Satija et al. 2015; Stoeckius et al. 2017). Since we reduced the data size, the CITE-seq data needs to be filtered as well.

```
#read in data and CITE-seq
seurat <- readRDS("haod.RDS")
cb <- Read10X("GSE164378_CITE/")

#filter CITE-seq
adt_assay <- CreateAssayObject(counts =
                                 cb[,colnames(cb) %in% colnames(seurat)])
#add CITE-seq to SeuratObject
seurat[["ADT"]] <- adt_assay

#scale and normalize CITE-seq
seurat <- NormalizeData(seurat, assay = "ADT", normalization.method = "CLR")
```

```
## Normalizing across features
```

```
seurat <- ScaleData(seurat, assay = "ADT")
```

```
## Centering and scaling data matrix
```

Get Gene Sets and Run CAMML

In order to run CAMML, a gene set either needs to be built or loaded from the pre-built datasets. "GetGeneSets" will be pull the gene set data for 5 immune cell types in the following code. The data frame that is outputted can then be fed directly into CAMML to score each cell type.

```
gene.set.df <- GetGeneSets(data= "immune.cells")
seurat <- CAMML(seurat, gene.set.df)
```

```
## Computing VAM distances for 5 gene sets, 56775 cells and 17808 genes.

## Min set size: 9, median size: 24

## Warning: Feature names cannot have underscores ('_'), replacing with dashes
## ('-')

## Warning: Feature names cannot have underscores ('_'), replacing with dashes
## ('-')
```

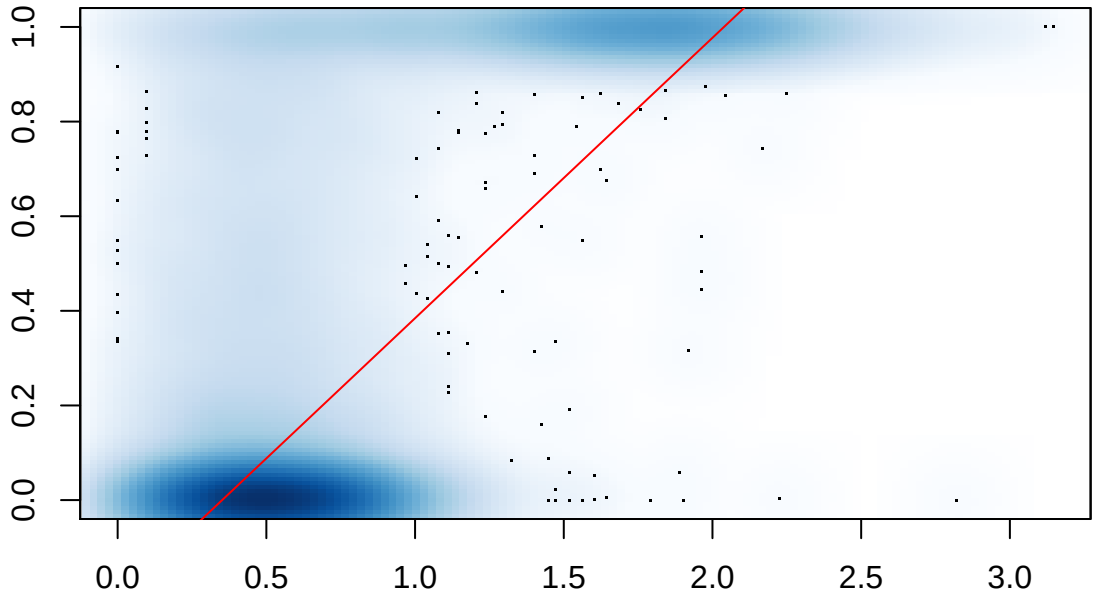
Visualize CAMML vs. CITE-seq

Following the running of CAMML, an assay for CAMML will be added to the Seurat Object with the scores for each cell type in each cell. The following code will save vectors for the CAMML scores for each cell type and the expression of their corresponding CITE-seq markers. These will then be visualized with SmoothScatter in R with a line of best fit for the linear model between the CAMML scores and the cell surface CITE-seq marker expression.

```
B <- c(as.matrix(seurat@assays$CAMML[which(rownames(seurat@assays$CAMML)=="B-cell"),]))
H <- c(as.matrix(seurat@assays$CAMML[which(rownames(seurat@assays$CAMML)=="HSC-CD34+"),]))
M <- c(as.matrix(seurat@assays$CAMML[which(rownames(seurat@assays$CAMML)=="Monocyte"),]))
N <- c(as.matrix(seurat@assays$CAMML[which(rownames(seurat@assays$CAMML)=="NK-cell"),]))
Tall <- c(as.matrix(seurat@assays$CAMML[which(rownames(seurat@assays$CAMML)=="T-cells"),]))


Bc <- c(as.matrix(seurat@assays$ADT[which(rownames(seurat@assays$ADT)=="CD19"),]))
Hc <- c(as.matrix(seurat@assays$ADT[which(rownames(seurat@assays$ADT)=="CD34"),]))
Mc <- c(as.matrix(seurat@assays$ADT[which(rownames(seurat@assays$ADT)=="CD14"),]))
Nc <- c(as.matrix(seurat@assays$ADT[which(rownames(seurat@assays$ADT)=="CD16"),]))
T8c <- c(as.matrix(seurat@assays$ADT[which(rownames(seurat@assays$ADT)=="CD8"),]))
T4c <- c(as.matrix(seurat@assays$ADT[which(rownames(seurat@assays$ADT)=="CD4-1"),]))

smoothScatter(Mc, M,xlab = "", ylab = "")
abline(lm(M ~ Mc), col="red")
```



References

Carlson, Marc. 2020. *Org.Hs.eg.db: Genome Wide Annotation for Human.*

Edgar, Ron, Michael Domrachev, and Alex E. Lash. 2002. "Gene Expression Omnibus: NCBI Gene Expression and Hybridization Array Data Repository." *Nucleic Acids Research* 30 (1): 207–10. https://doi.org/10.1093/nar/30.1.207.

Hao, Yuhan, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck, Shiwei Zheng, Andrew Butler, Maddie J. Lee, et al. 2021. "Integrated Analysis of Multimodal Single-Cell Data." *Cell* 184 (13): 3573–3587.e29. https://doi.org/10.1016/j.cell.2021.04.048.

Liberzon, Arthur, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P. Mesirov. 2011. "Molecular Signatures Database (MSigDB) 3.0." *Bioinformatics* 27 (12): 1739–40. https://doi.org/10.1093/bioinformatics/btr260.

Robinson, Mark D, Davis J McCarthy, and Gordon K Smyth. 2010. "edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics* 26 (1): 139–40. https://doi.org/10.1093/bioinformatics/btp616.

Satija, Rahul, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. 2015. "Spatial Reconstruction of Single-Cell Gene Expression Data." *Nature Biotechnology* 33 (5): 495–502. https://doi.org/10.1038/nbt.3192.

Schiebout, Courtney, and H. Robert Frost. 2022. "CAMML: Multi-Label Immune Cell-Typing and Stemness Analysis for Single-Cell RNA-Sequencing." *Pacific Symposium on Biocomputing.*

Stoeckius, Marlon, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. 2017. "Simultaneous Epitope and Transcriptome Measurement in Single Cells." *Nature Methods* 14 (9): 865–68. https://doi.org/10.1038/nmeth.4380.