Quick 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 (Satija et al. 2015; Robinson, McCarthy, and Smyth 2010; Carlson 2023; Liberzon et al. 2011). Packages will also load additional libraries they depend on.

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

Get Gene Set

Cell type gene sets can be loaded with the GetGeneSet function. In this case, we will load "immune.cells" which calls data for 5 immune cell types: T cells, B cells, NK cells, Monocytes, and Hematopoietic Stem Cells (HSCs).

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

Load Data

For this quick example, we will use "pbmc_small" from Seurat, which will provide a Seurat Object of 80 peripheral blood mononuclear cells (Satija et al. 2015).

```
seurat <- SeuratObject::pbmc_small
seurat <- RunPCA(seurat)
```

```
## Warning in irlba(A = t(x = object), nv = npcs, ...): You're computing too large
## a percentage of total singular values, use a standard svd instead.
```

```
## Warning in irlba(A = t(x = object), nv = npcs, ...): did not converge--results
## might be invalid!; try increasing work or maxit
```

```
## Warning: Requested number is larger than the number of available items (20).
## Setting to 20.
```

```
## Warning: Requested number is larger than the number of available items (20).
## Setting to 20.
```

```
## Warning: Requested number is larger than the number of available items (20).
## Setting to 20.
```

```
## Warning: Requested number is larger than the number of available items (20).
## Setting to 20.

## Warning: Requested number is larger than the number of available items (20).
## Setting to 20.

## PC_ 1
## Positive:  SDPR, PF4, PPBP, TUBB1, CA2, TREML1, MYL9, PGRMC1, RUFY1, PARVB
## Negative:  HLA-DPB1, HLA-DQA1, S100A9, S100A8, GNLY, RP11-290F20.3, CD1C, AKR1C3, IGLL5, VDAC3
## PC_ 2
## Positive:  HLA-DPB1, HLA-DQA1, S100A8, S100A9, CD1C, RP11-290F20.3, PARVB, IGLL5, MYL9, SDPR
## Negative:  GNLY, AKR1C3, VDAC3, PGRMC1, TUBB1, PF4, TREML1, RUFY1, CA2, PPBP
## PC_ 3
## Positive:  S100A9, S100A8, RP11-290F20.3, AKR1C3, PARVB, GNLY, PPBP, PGRMC1, MYL9, TUBB1
## Negative:  HLA-DQA1, CD1C, IGLL5, HLA-DPB1, RUFY1, PF4, VDAC3, SDPR, TREML1, CA2
## PC_ 4
## Positive:  IGLL5, RP11-290F20.3, VDAC3, PPBP, TUBB1, TREML1, PF4, CA2, PARVB, MYL9
## Negative:  CD1C, AKR1C3, S100A8, GNLY, HLA-DPB1, HLA-DQA1, S100A9, PGRMC1, RUFY1, SDPR
## PC_ 5
## Positive:  MYL9, PARVB, IGLL5, TREML1, AKR1C3, PGRMC1, HLA-DPB1, S100A9, TUBB1, PF4
## Negative:  VDAC3, RP11-290F20.3, RUFY1, CD1C, HLA-DQA1, CA2, S100A8, PPBP, GNLY, SDPR
```

```
seurat <- RunUMAP(seurat, dims = 1:10)
```

```
## Warning: The default method for RunUMAP has changed from calling Python UMAP via reticulate to the R-
## To use Python UMAP via reticulate, set umap.method to 'umap-learn' and metric to 'correlation'
## This message will be shown once per session

## 15:08:13 UMAP embedding parameters a = 0.9922 b = 1.112

## Found more than one class "dist" in cache; using the first, from namespace 'spam'

## Also defined by 'BiocGenerics'

## 15:08:13 Read 80 rows and found 10 numeric columns

## 15:08:13 Using Annoy for neighbor search, n_neighbors = 30

## Found more than one class "dist" in cache; using the first, from namespace 'spam'

## Also defined by 'BiocGenerics'

## 15:08:13 Building Annoy index with metric = cosine, n_trees = 50

## 0%   10   20   30   40   50   60   70   80   90   100%

## [----|----|----|----|----|----|----|----|----|----|

## **************************************************|
## 15:08:13 Writing NN index file to temp file /var/folders/wv/9lqlnjl571q8w6tn77wg10pr0000gp/T//RtmpTf
## 15:08:13 Searching Annoy index using 1 thread, search_k = 3000
## 15:08:13 Annoy recall = 100%
## 15:08:13 Commencing smooth kNN distance calibration using 1 thread with target n_neighbors = 30
## 15:08:13 7 smooth knn distance failures
## 15:08:14 Initializing from normalized Laplacian + noise (using RSpectra)
## 15:08:14 Commencing optimization for 500 epochs, with 2410 positive edges
## 15:08:14 Optimization finished
```

Run CAMML

Once a gene set data frame and Seurat Object are defined, CAMML can simply be run by inputting both in the CAMML function. Labels can be defined for each cell using GetCAMMLLabels and designating the preferred label types.

```
seurat <- CAMML(seurat, gene.set.df)
```

```
## Computing VAM distances for 4 gene sets, 80 cells and 230 genes.
```

```
## Min set size: 1, median size: 6
```

```
## Warning in vamForCollection(gene.expr = Matrix::t(normalized.counts),
## gene.set.collection = gene.set.collection, : Gene set 3 has just a single
## member!
```

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

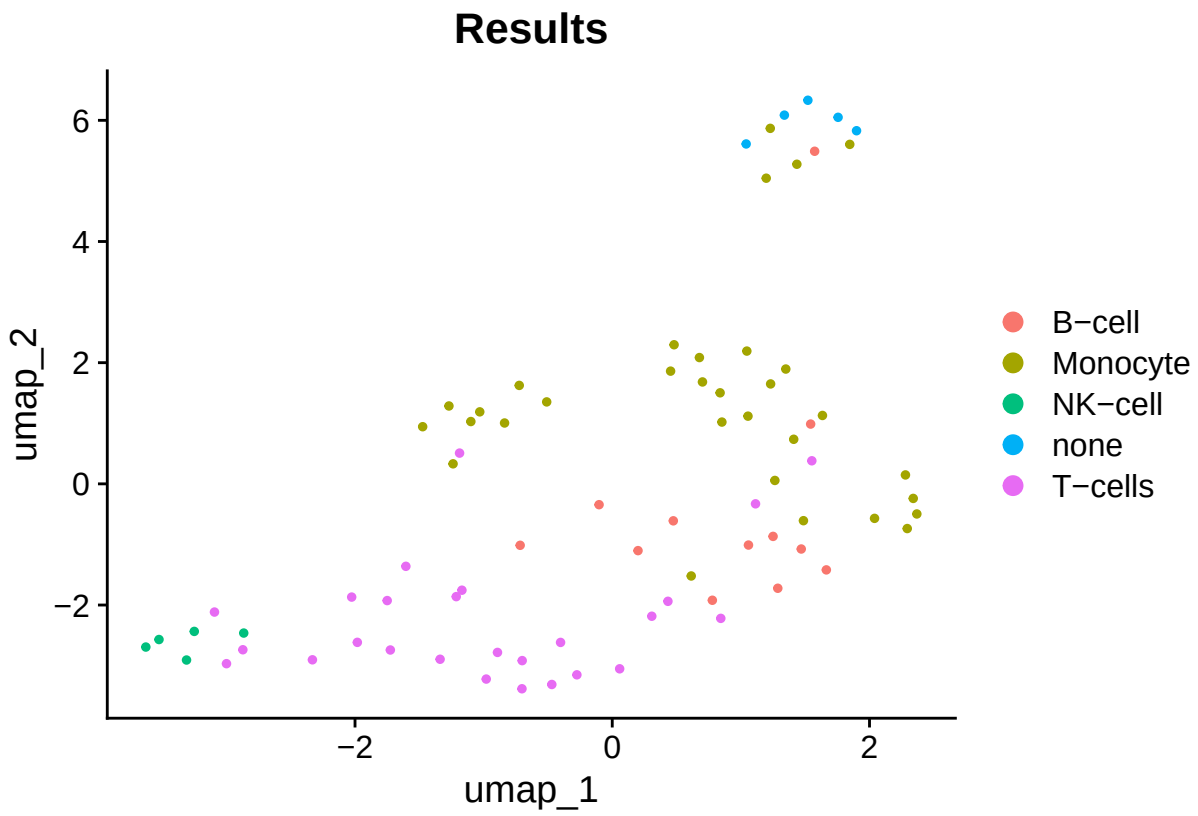
```
## Warning: Feature names cannot have underscores ('_'), replacing with dashes
## ('-')
```

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

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

```
## Warning: Key 'vamcdf_' taken, using 'camml_' instead
```

```
results <- GetCAMMLLabels(seurat, labels = "top1")
seurat$Results <- unlist(results)
UMAPPlot(object = seurat, group = "Results")
```



References

Carlson, Marc. 2023. *Org.hs.eg.db: Genome Wide Annotation for Human.*

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." In *Pacific Symposium on Biocomputing*. Waimea, HI: World Scientific Publishing.