Application of VAM to 10x PBMC 3k scRNA-seq data using the SCTransform method.

H. Robert Frost

1 Load the VAM package

Loading VAM will also load the required packages MASS and Matrix. Seurat is referenced via suggests so must be directly loaded to enable access to Seurat functions. Note that the sctransform package is also required in this case.

```
> library(VAM)
> library(Seurat)
```

2 Load and process the 10x PBMC scRNA-seq data

This example uses the same 10x PBMC scRNA-seq data set that is used in the Seurat Guided Clustering vignette
(https://satijalab.org/seurat/v3.1/pbmc3k_tutorial.html). The Cell Ranger files for this data set can be downloaded from
https://s3-us-west-2.amazonaws.com/10x.files/samples/cell/pbmc3k/pbmc3k_filtered_gene_bc_matrices.tar.gz.
Except for the use of SCTransform instead of log normalization, the logic here is identical to that found in the Seurat Guided Clustering vignette.

```
> # update the data.dir argument to reflect the local location of the PBMC data
> pbmc.data = Read10X(data.dir = "./filtered_gene_bc_matrices/hg19/")
> pbmc = CreateSeuratObject(counts = pbmc.data, project = "pbmc3k", min.cells = 3,
+         min.features = 200)
> pbmc[["percent.mt"]] = PercentageFeatureSet(pbmc, pattern = "^MT-")
> pbmc = subset(pbmc, subset = nFeature_RNA > 200 & nFeature_RNA < 2500 & percent.mt < 5)
> pbmc = SCTransform(pbmc, verbose=F)
> pbmc = RunPCA(pbmc, npcs=10)
> pbmc = RunUMAP(pbmc, dims = 1:10)
> pbmc

An object of class Seurat
26233 features across 2638 samples within 2 assays
Active assay: SCT (12519 features, 3000 variable features)
 3 layers present: counts, data, scale.data
 1 other assay present: RNA
 2 dimensional reductions calculated: pca, umap
```

3 Load Ensembl IDs

The Ensembl IDs and gene names must be read in from the genes.tsv file and filtered to match genes left after the quality control steps performed in the prior section.

```
> feature.data = read.delim("./filtered_gene_bc_matrices/hg19/genes.tsv",
+        header = FALSE, stringsAsFactors = FALSE)
> ensembl.ids = feature.data[,1]
> gene.names = feature.data[,2]
> genes.after.QC = rownames(pbmc)
> indices.to.keep = unlist(sapply(genes.after.QC, function(x) {which(gene.names == x)[1]}))
> ensembl.ids = ensembl.ids[indices.to.keep]
> gene.names = gene.names[indices.to.keep]
```

4 Define gene set collection

A gene set collection containing just the BIOCARTA_BLYMPHOCYTE_PATHWAY from the
MSigDB C2.CP.BIOCARTA collection will be used for this example. To create a version of this gene set
that can be used with the PBMC scRNA-seq data, the Entrez IDs from MSigDB were first mapped to
Ensembl IDs using the Bioconductor org.Hs.egENSEMBL package and a gene set collection list object
required by vamForSeurat() was created using the createGeneSetCollection() helper function. This helper
function filters the original 14 genes down to the 11 that were also contained in the PBMC scRNA-seq
data and generates a list whose elements are vectors of gene indices in the scRNA-seq data. To use the
VAM method with an entire MSigDB gene set collection (or other collection of pathways), similar logic
would be needed to filter genes, determine Ensembl IDs and map these IDs to gene position in the Seurat
Assay.

```
> gene.set.name = "BIOCARTA-BLYMPHOCYTE-PATHWAY"
> blymphocyte.gene.ids = c("ENSG00000121594", "ENSG00000005844", "ENSG00000203710",
+ "ENSG00000160255", "ENSG00000117322", "ENSG00000101017", "ENSG00000204287",
+ "ENSG00000198502", "ENSG00000090339", "ENSG00000072694", "ENSG00000081237",
+ "ENSG00000196126", "ENSG00000231021", "ENSG00000230463")
> # Create a collection list for this gene set based on the Ensembl IDs
> gene.set.id.list = list()
> gene.set.id.list[[1]] = blymphocyte.gene.ids
> names(gene.set.id.list)[1] = gene.set.name
> gene.set.id.list

$`BIOCARTA-BLYMPHOCYTE-PATHWAY`
 [1] "ENSG00000121594" "ENSG00000005844" "ENSG00000203710" "ENSG00000160255"
 [5] "ENSG00000117322" "ENSG00000101017" "ENSG00000204287" "ENSG00000198502"
 [9] "ENSG00000090339" "ENSG00000072694" "ENSG00000081237" "ENSG00000196126"
[13] "ENSG00000231021" "ENSG00000230463"

> # Create the list of gene indices required by vamForSeurat()
> (gene.set.collection = createGeneSetCollection(gene.ids=ensembl.ids,
+        gene.set.collection=gene.set.id.list))

$`BIOCARTA-BLYMPHOCYTE-PATHWAY`
ENSG00000005844 ENSG00000203710 ENSG00000160255 ENSG00000117322 ENSG00000101017
           9477            1122           12482            1121           10961
ENSG00000204287 ENSG00000198502 ENSG00000090339 ENSG00000072694 ENSG00000081237
           3985            3986           11280             944            1062
ENSG00000196126
           3987
```

```
> blymphocyte.gene.indices = gene.set.collection[[1]]
> (blymphocyte.gene.names = gene.names[blymphocyte.gene.indices])

 [1] "ITGAL"     "CR1"       "ITGB2"     "CR2"       "CD40"      "HLA-DRA"
 [7] "HLA-DRB5"  "ICAM1"     "FCGR2B"    "PTPRC"     "HLA-DRB1"
```

5 Execute VAM method

Since the scRNA-seq data has been processed using Seurat, we execute VAM using the vamForSeurat()
function. We have set return.dist=T so that the squared adjusted Mahalanobis distances will be returned
in a "VAMdist" Assay.

```
> pbmc = vamForSeurat(seurat.data=pbmc,
+     gene.set.collection=gene.set.collection,
+     center=F, gamma=T, sample.cov=F, return.dist=T)
```

Look at the first few entries in the "VAMdist" and "VAMcdf" Assays.

```
> pbmc@assays$VAMdist@data[1,1:10]

AAACATACAACCAC-1 AAACATTGAGCTAC-1 AAACATTGATCAGC-1 AAACCGTGCTTCCG-1
       9.044914       171.277651         7.190289       210.518543
AAACCGTGTATGCG-1 AAACGCACTGGTAC-1 AAACGCTGACCAGT-1 AAACGCTGGTTCTT-1
      60.349052         3.534960        24.697057         0.000000
AAACGCTGTAGCCA-1 AAACGCTGTTTCTG-1
      41.312718        15.789831


> pbmc@assays$VAMcdf@data[1,1:10]

AAACATACAACCAC-1 AAACATTGAGCTAC-1 AAACATTGATCAGC-1 AAACCGTGCTTCCG-1
     0.12233048       0.95423467       0.09633896       0.97816364
AAACCGTGTATGCG-1 AAACGCACTGGTAC-1 AAACGCTGACCAGT-1 AAACGCTGGTTCTT-1
     0.64128566       0.04507302       0.32410303       0.00000000
AAACGCTGTAGCCA-1 AAACGCTGTTTCTG-1
     0.49493757       0.21381777
```

6 Visualize VAM scores

Visualize VAM scores using Seurat FeaturePlot(). The default Assay must first be changed to "VAMcdf".

```
> DefaultAssay(object = pbmc) = "VAMcdf"
> FeaturePlot(pbmc, reduction="umap", features=gene.set.name)
```

BIOCARTA−BLYMPHOCYTE−PATHWAY