Application of VAM to 10x PBMC 3k scRNA-seq data using Seurat log normalization.

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.

```
> 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. This data is loaded and processed using the same Seurat logic found in the Guided Clustering vignette. In particular, the Seurat log normalization method implemented by NormalizeData() is used with variable genes determined by FindVariableFeatures(). This method for variable feature determination decomposes the measured variance for each gene into biological and technical components and provides the values of technical variance input to the VAM algorithm.

```
> # 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 = NormalizeData(pbmc)
> pbmc = FindVariableFeatures(pbmc, selection.method = "vst", nfeatures = 2000)
> pbmc = ScaleData(pbmc, features = rownames(pbmc))
> pbmc = RunPCA(pbmc, features = VariableFeatures(object = pbmc))
> pbmc = RunUMAP(pbmc, dims = 1:10)
> pbmc
```

```
An object of class Seurat
13714 features across 2638 samples within 1 assay
Active assay: RNA (13714 features, 2000 variable features)
 3 layers present: counts, data, scale.data
 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
          10383            1257           13675            1256           12011
```

```
ENSG00000204287 ENSG00000198502 ENSG00000090339 ENSG00000072694 ENSG00000081237
           4410            4411           12356            1057            1188
ENSG00000196126
           4412
```

```
> 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 
        6.687974        60.048298         3.159264        80.796323 
AAACCGTGTATGCG-1 AAACGCACTGGTAC-1 AAACGCTGACCAGT-1 AAACGCTGGTTCTT-1 
       33.611489         4.484255        20.982912         0.000000 
AAACGCTGTAGCCA-1 AAACGCTGTTTCTG-1 
       17.465160        11.849158 
```

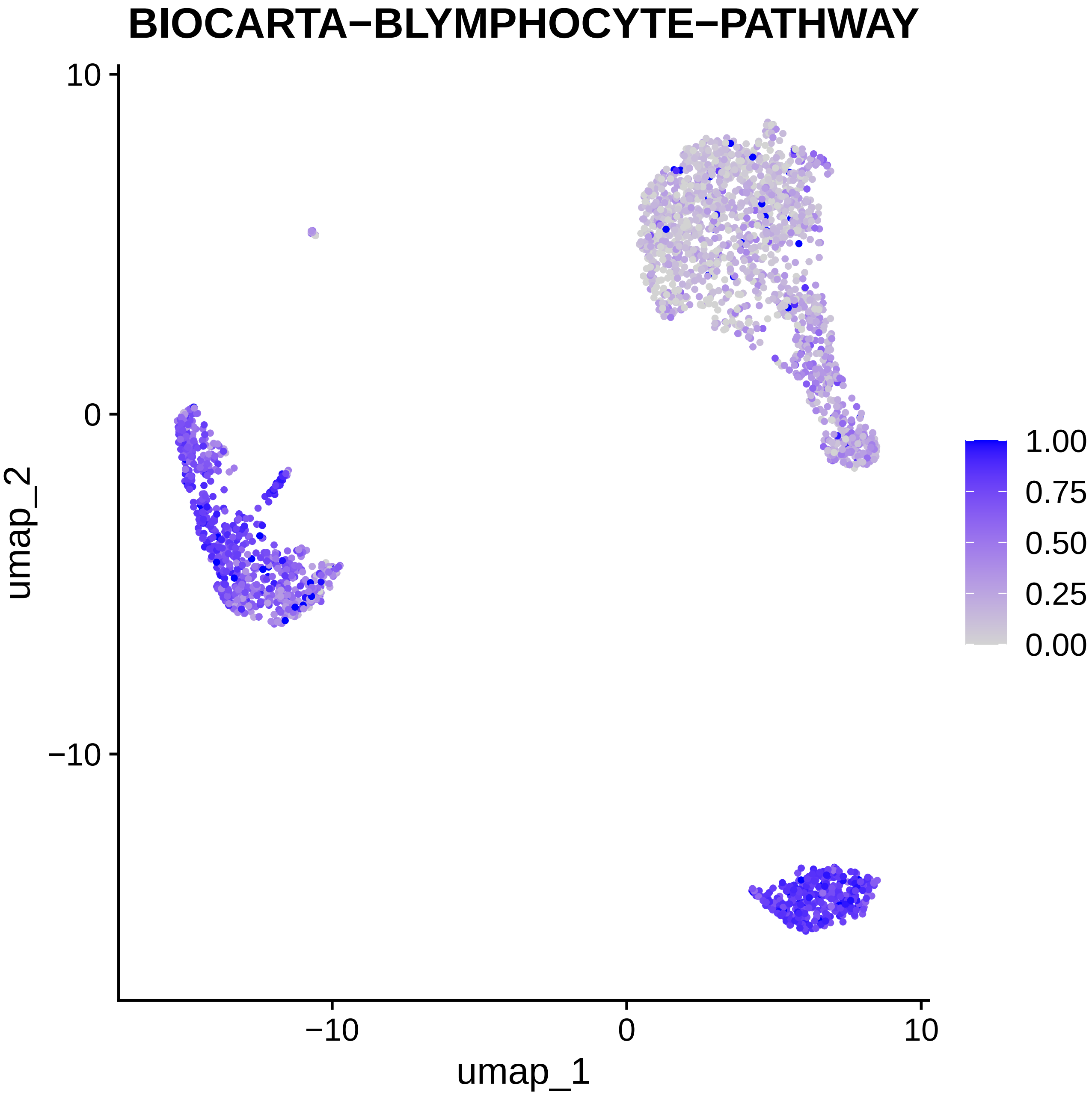
```
> pbmc@assays$VAMcdf@data[1,1:10]
```

```
AAACATACAACCAC-1 AAACATTGAGCTAC-1 AAACATTGATCAGC-1 AAACCGTGCTTCCG-1 
      0.17209578       0.78889014       0.08844222       0.87451969 
AAACCGTGTATGCG-1 AAACGCACTGGTAC-1 AAACGCTGACCAGT-1 AAACGCTGGTTCTT-1 
      0.58812094       0.12113573       0.43081777       0.00000000 
AAACGCTGTAGCCA-1 AAACGCTGTTTCTG-1 
      0.37658687       0.27814101 
```

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
10
0
−10
umap_2
−10
0
10
umap_1
1.00
0.75
0.50
0.25
0.00