Receptor Abundance Estimation using SPECK: An Application to the GSE164378 Peripheral Blood Mononuclear Cells (PBMC) data

Azka Javaid and H. Robert Frost

Load the SPECK package

SPECK depends on functionalities from the Seurat, rsvd and Ckmeans.1d.dp packages. Loading SPECK will load these required packages. Seurat is also individually loaded in this vignette to facilitate access to downstream data visualization capabilities. The ggplot2 and gridExtra packages are loaded to enable graphical arrangement of downstream imputed abundance profiles.

```
library(SPECK)
library(Seurat)
library(ggplot2)
library(gridExtra)
```

Load a subset of the GSE164378 single cell RNA-sequencing (scRNA-seq) data

A subset of 1000 samples from a human peripheral blood mononuclear cells (PBMC) scRNA-seq dataset, accessible using the Gene Expression Omnibus (GEO) at accession number GSE164378 and DOI: 10.1016/j.cell.2021.04.048 (Hao et al. 2021), is subsequently loaded. The transformation from the raw data files to this subset can be accessed at the `SPECK/data-raw/pbmc_data_processing.R` script.

```
data("pbmc.rna.mat")
dim(pbmc.rna.mat)
#> [1]  1000 33538
```

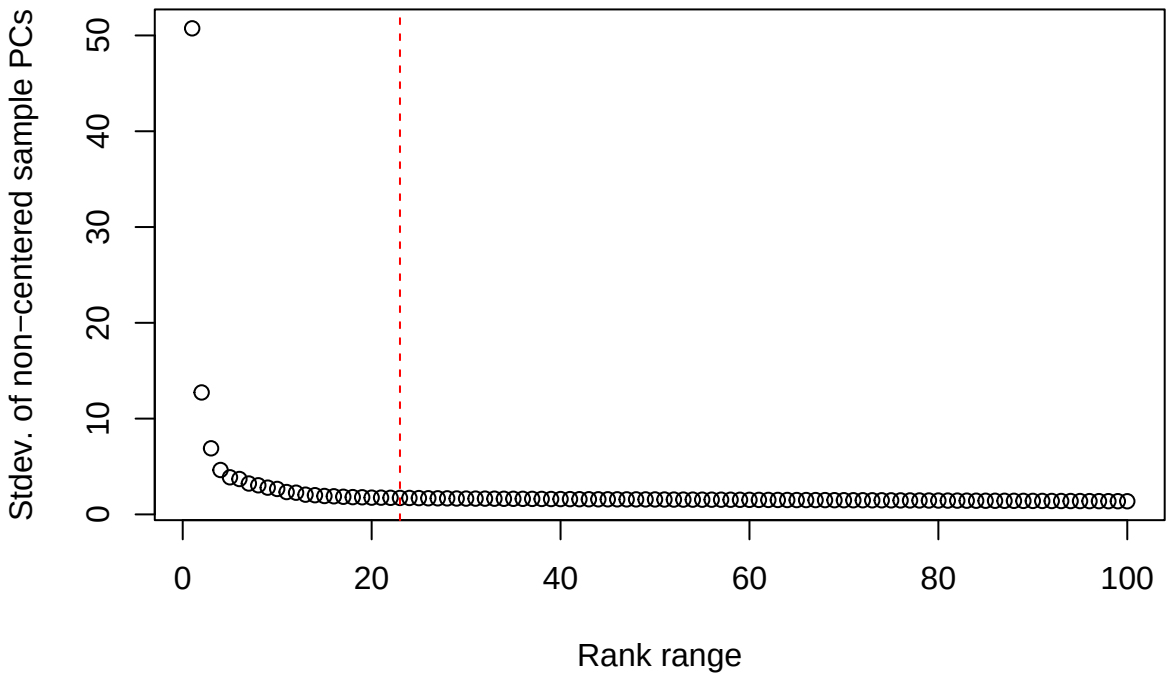
Execute the SPECK method

SPECK method is executed on the subset, `pbmc.rna.seurat` data, using the `SPECK()` function. The thresholded and reduced rank reconstructed output, which is a $m \times n$ matrix consisting of m (i.e. 1000) samples and n genes, can be accessed via the `thresholded.mat` component of the returned list. This matrix is then set to SPECK assay in the `pbmc.rna.seurat` object.

```
speck.full <- speck(counts.matrix = pbmc.rna.mat, rank.range.end = 100,
                    min.consec.diff = 0.01, rep.consec.diff = 2,
                    manual.rank = NULL, max.num.clusters = 4,
                    seed.rsvd = 1, seed.ckmeans = 2)

speck.rank <- speck.full$rrr.rank
paste("Rank: ", speck.rank, sep = "")
#> [1] "Rank: 23"

plot(speck.full$component.stdev, ylab = "Stdev. of non-centered sample PCs",
xlab = "Rank range", main = paste("Selected rank (k=", speck.rank, ")", sep=""))
abline(v = speck.rank, lty = 2, col = "red")
```




```
head(speck.full$clust.num); table(speck.full$clust.num)
#>              number.clusters
#> MIR1302-2HG                2
#> FAM138A                    1
#> OR4F5                      1
#> AL627309.1                 1
#> AL627309.3                 1
#> AL627309.2                 3
#>
#>     1     2     3     4
#> 25656  5883  1471   528
head(speck.full$clust.max.prop)
#>              proportion.max.clust
#> MIR1302-2HG                    0
#> FAM138A                        0
#> OR4F5                          0
#> AL627309.1                     0
#> AL627309.3                     0
#> AL627309.2                     0

speck.output <- speck.full$thresholded.mat
paste("# of samples in RRR object:", dim(speck.output)[1])
#> [1] "# of samples in RRR object: 1000"
paste("# of genes in RRR object:", dim(speck.output)[2])
#> [1] "# of genes in RRR object: 33538"
SPECK_assay <- CreateAssayObject(counts = t(speck.output))
pbmc.rna.seurat <- CreateSeuratObject(counts = t(as.matrix(pbmc.rna.mat)))
pbmc.rna.seurat[["SPECK"]] <- SPECK_assay
```

Visualize estimated abundance profiles

Seurat's preprocessing workflow is next applied to the `pbmc.rna.seurat` object to enable downstream visualization.

```
DefaultAssay(pbmc.rna.seurat) <- "RNA"
pbmc.rna.seurat <- NormalizeData(pbmc.rna.seurat)
pbmc.rna.seurat <- FindVariableFeatures(pbmc.rna.seurat,
                                        selection.method = "vst",
                                        nfeatures = 2000)
all.genes <- rownames(pbmc.rna.seurat)
pbmc.rna.seurat <- ScaleData(pbmc.rna.seurat, features = all.genes)
pbmc.rna.seurat <- RunPCA(pbmc.rna.seurat,
                          features = VariableFeatures(object = pbmc.rna.seurat))
pbmc.rna.seurat <- FindNeighbors(pbmc.rna.seurat, dims = 1:10)
pbmc.rna.seurat <- FindClusters(pbmc.rna.seurat, resolution = 0.5)
#> Modularity Optimizer version 1.3.0 by Ludo Waltman and Nees Jan van Eck
#>
#> Number of nodes: 1000
#> Number of edges: 33928
#>
#> Running Louvain algorithm...
#> Maximum modularity in 10 random starts: 0.8700
#> Number of communities: 8
#> Elapsed time: 0 seconds
pbmc.rna.seurat <- RunUMAP(pbmc.rna.seurat, dims = 1:10)
```

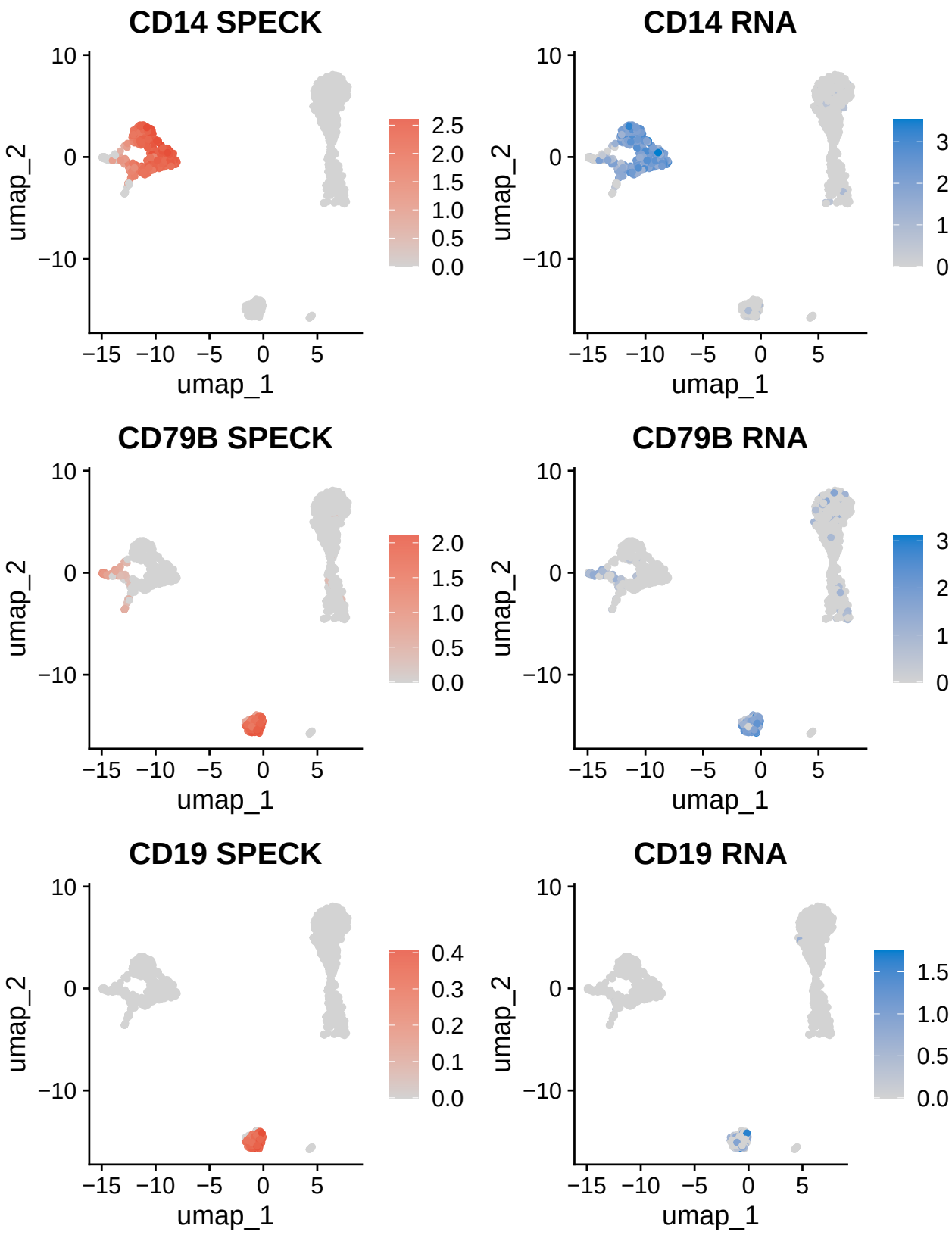
Estimated abundance profiles are visualized on clustered data using the `FeaturePlot()` function.

```
DefaultAssay(pbmc.rna.seurat) <- "RNA"
p1 <- FeaturePlot(pbmc.rna.seurat, "CD14", cols = c("lightgrey", "#007ece")) +
  ggtitle("CD14 RNA")
DefaultAssay(pbmc.rna.seurat) <- "SPECK"
p2 <- FeaturePlot(pbmc.rna.seurat, "CD14", cols=c("lightgrey", "#E64B35CC")) +
  ggtitle("CD14 SPECK")

DefaultAssay(pbmc.rna.seurat) <- "RNA"
p3 <- FeaturePlot(pbmc.rna.seurat, "CD79B", cols = c("lightgrey", "#007ece")) +
  ggtitle("CD79B RNA")
DefaultAssay(pbmc.rna.seurat) <- "SPECK"
p4 <- FeaturePlot(pbmc.rna.seurat, "CD79B", cols=c("lightgrey", "#E64B35CC")) +
  ggtitle("CD79B SPECK")

DefaultAssay(pbmc.rna.seurat) <- "RNA"
p5 <- FeaturePlot(pbmc.rna.seurat, "CD19", cols = c("lightgrey", "#007ece")) +
  ggtitle("CD19 RNA")
DefaultAssay(pbmc.rna.seurat) <- "SPECK"
p6 <- FeaturePlot(pbmc.rna.seurat, "CD19", cols=c("lightgrey", "#E64B35CC")) +
  ggtitle("CD19 SPECK")

grid.arrange(p2, p1,
             p4, p3,
             p6, p5,
             nrow = 3)
```


CD14 SPECK
CD14 RNA
CD79B SPECK
CD79B RNA
CD19 SPECK
CD19 RNA
umap_1
umap_2
10
0
−10
−15
−10
−5
0
5
2.5
2.0
1.5
1.0
0.5
0.0
3
2
1
0
0.4
0.3
0.2
0.1
0.0
1.5
1.0
0.5
0.0