Hematopeotic Stem Cells- Seurat Analysis

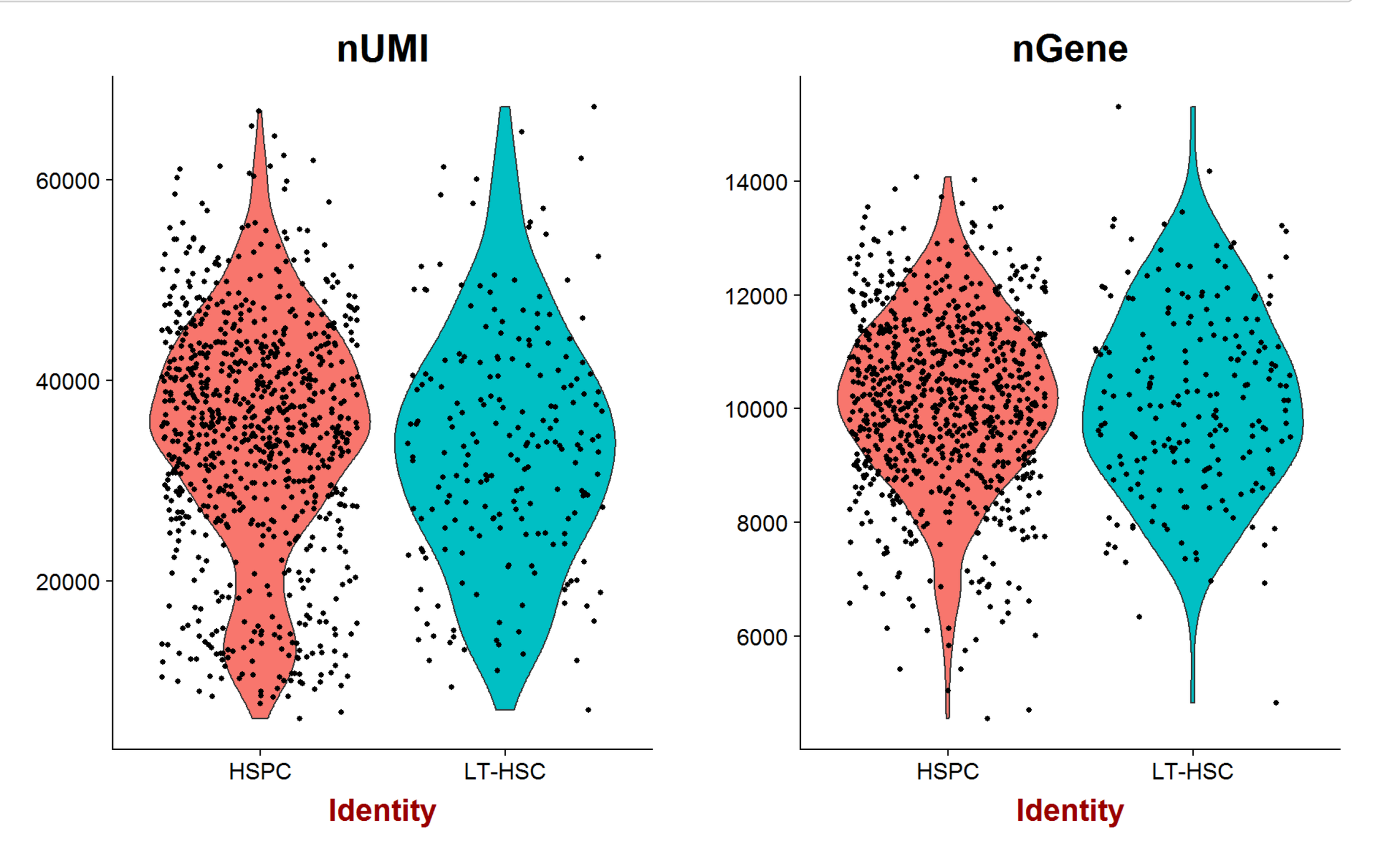
Nick Wawee

September 16, 2018

This document will analyze the HSC dataset published by Nestorowa et. al. with the Seurat package.

Plots are generated to see how many genes are mapped within each cell, and the number of genes expressed in each cell. Cells with number of readcounts >=4*10^5 are filtered out of the analysis because of potential doublets. Genes are filtered out if they contained 95% or more zeros. The HSPC and LT-HSC population is filtered in to focus on potential lymphoid progenitors. After all filtering, 22290 genes and 1044 cells undergo analyses. The expression levels are log2 transformed.

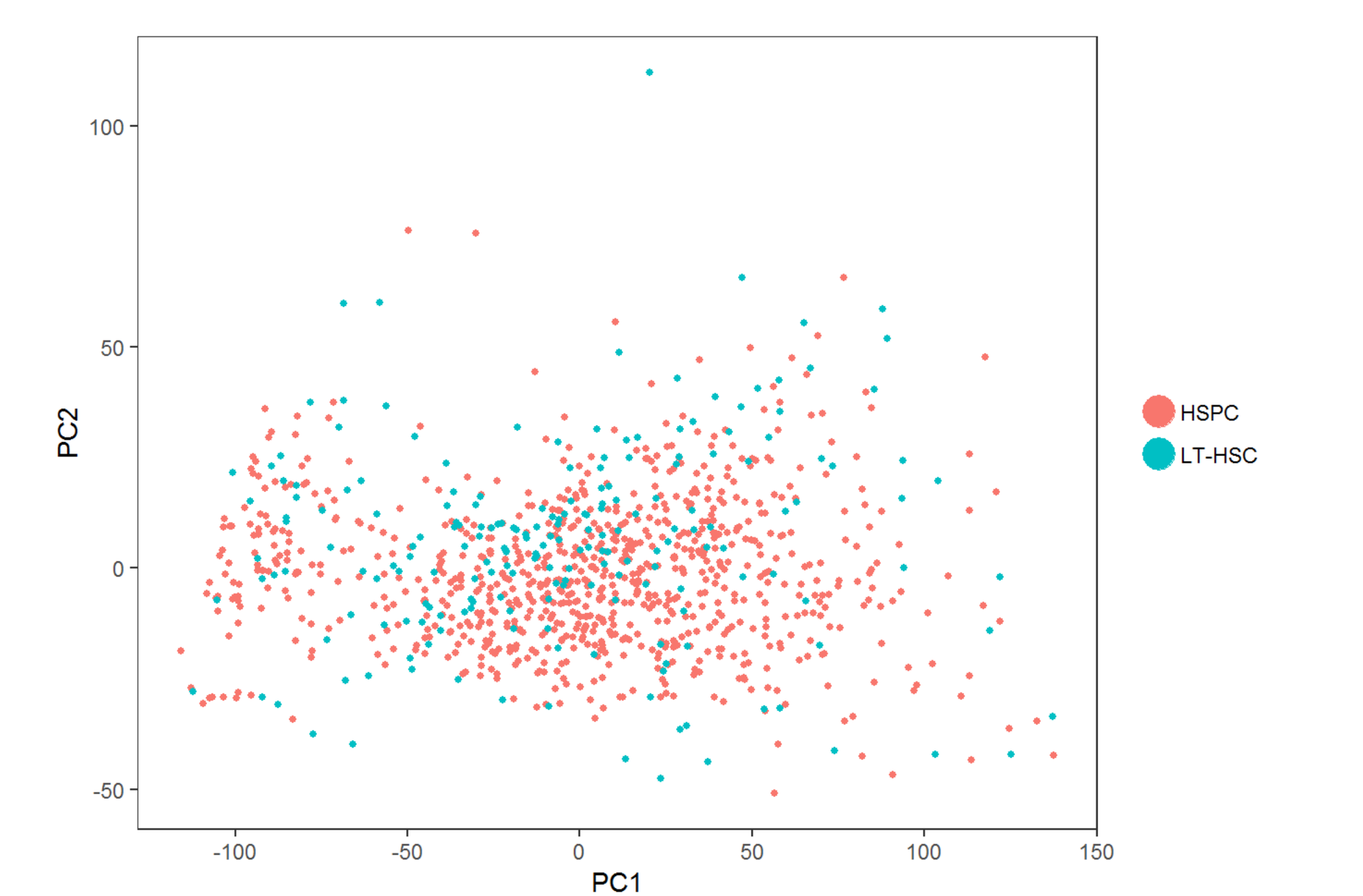
```
bigcds<- CreateSeuratObject(raw.data = nummat)
mito.genes <- grep(pattern = "^Mt-", x = rownames(x = bigcds@data), value = TRUE)
mito.fraction <- Matrix::colSums(bigcds@raw.data[mito.genes,])/Matrix::colSums(bigcds@raw.data)
bigcds<- AddMetaData(object =bigcds, metadata = mito.fraction, col.name = "mito.fraction")
VlnPlot(object = bigcds, features.plot = c("nUMI", "nGene"), nCol= 2)
```

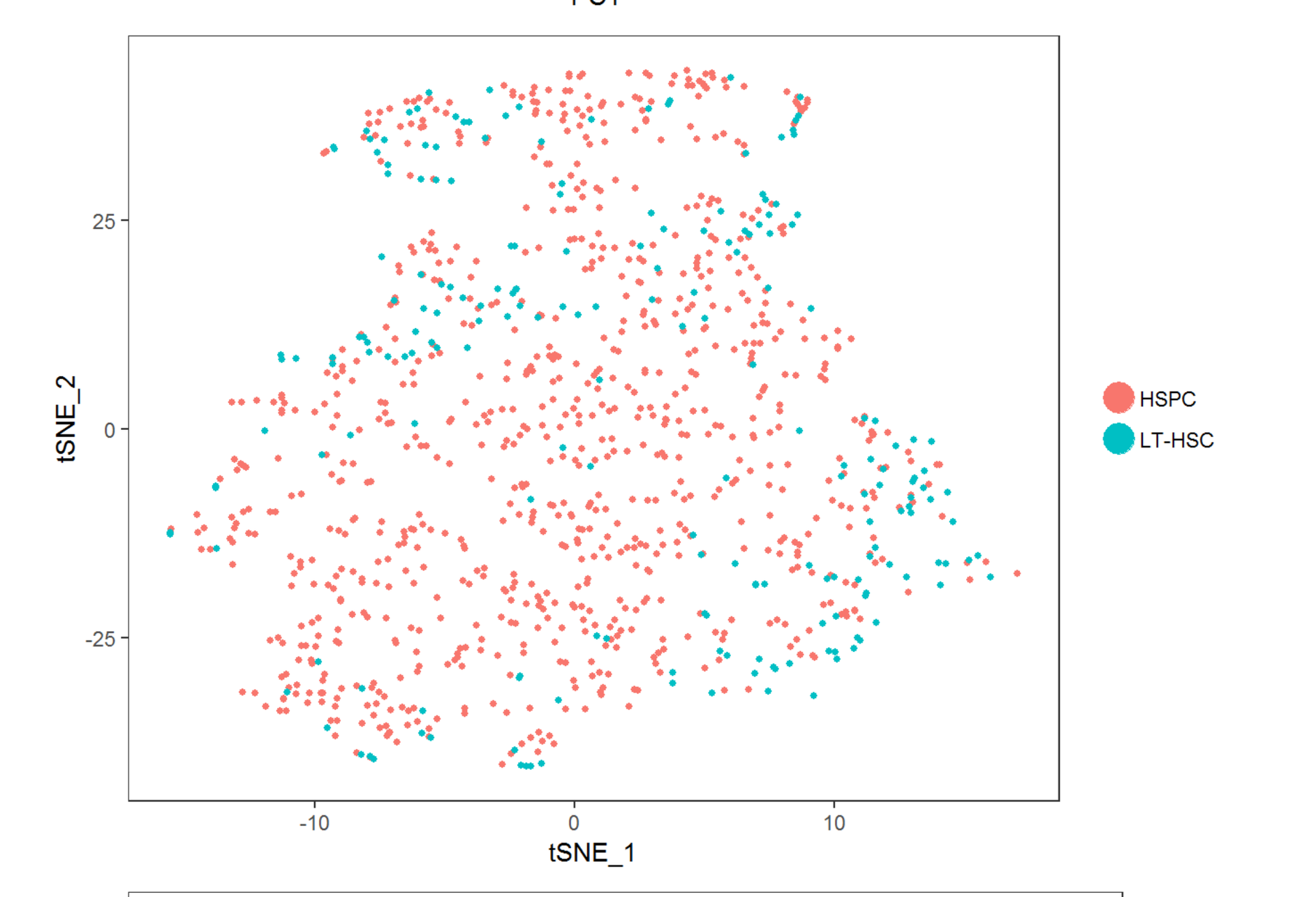


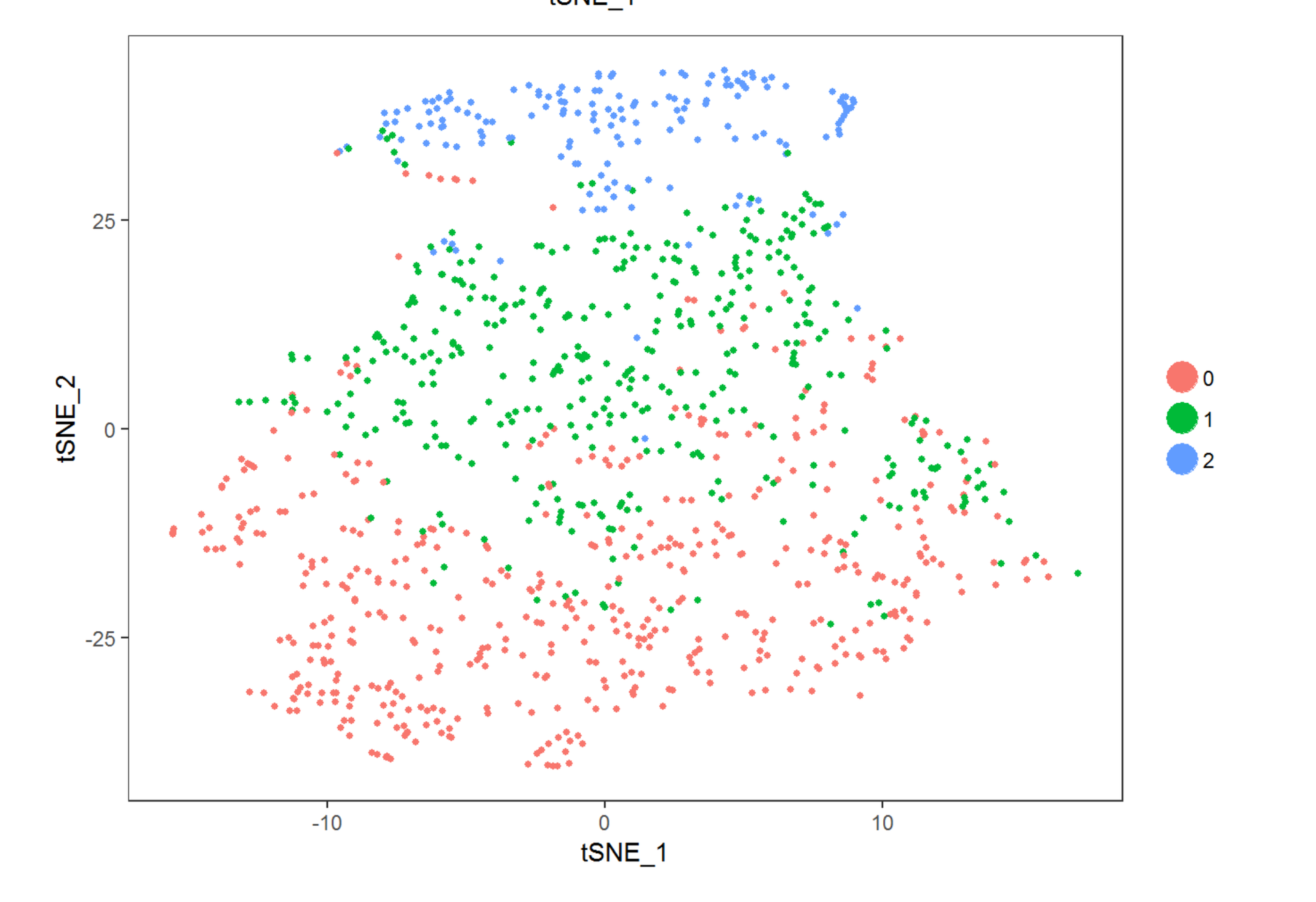
Data is normalized prior to PCA and tSNE.

```
#bigcds<- NormalizeData(object = bigcds)
bigcds<- ScaleData(object = bigcds, check.for.norm = FALSE)
bigcds<- FindVariableGenes(object = bigcds, do.plot = FALSE,mean.function = ExpMean, dispersion.function = LogVMR
)
```

PCA and tSNE







Gene expression levels are displayed in the tSNE 2-d respresentation. 3 distinct clusters are found within the data.

Markers are found between each cell cluster to identify sub populations and markers are uploaded from gene ontology GO:0002320 in lymphoid cell differentiation. All markers found that were differentially expressed were matched to the genes in the GO to identify the lymphoid cell cluster.

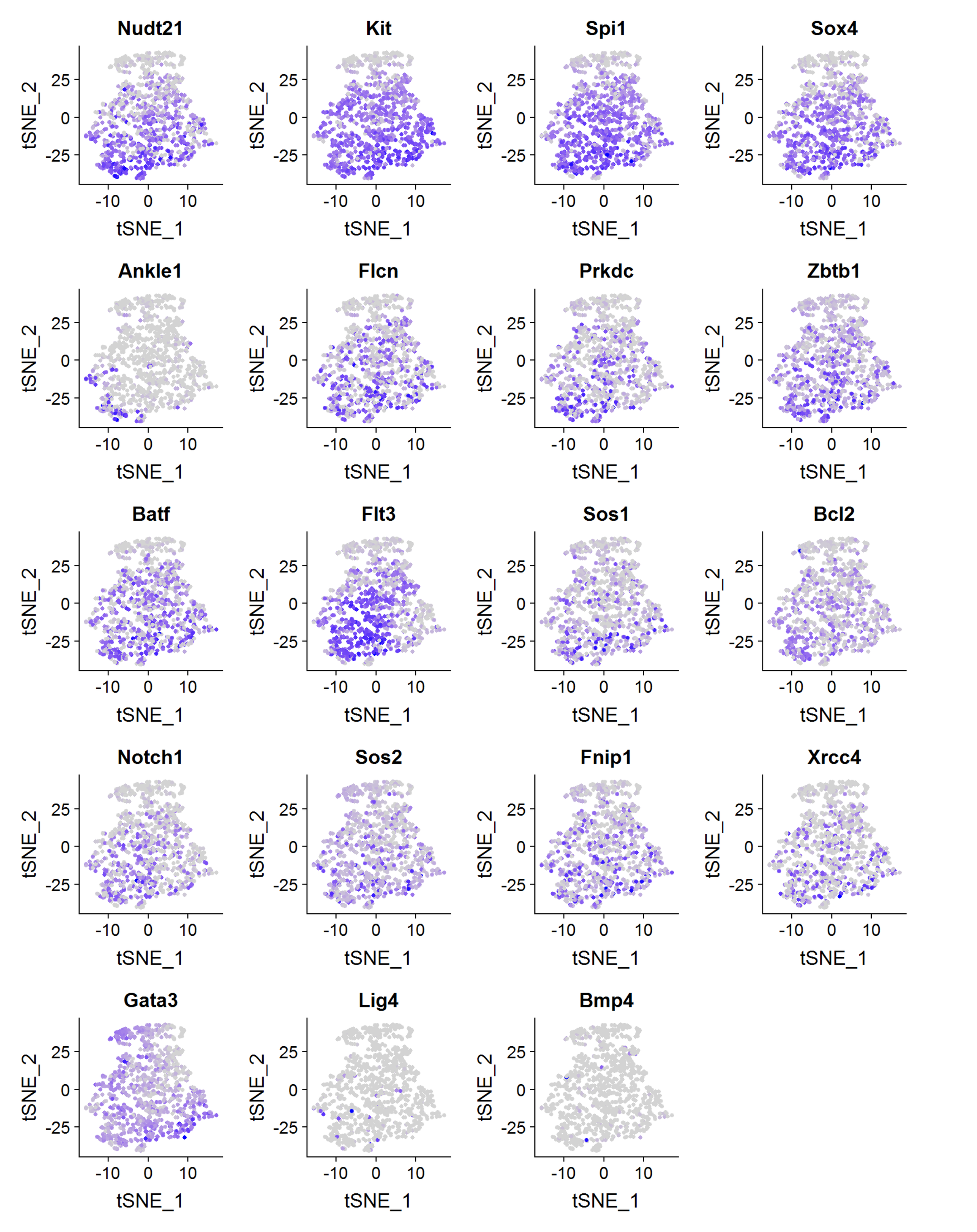
```
diffmarkers<-FindAllMarkers(bigcds)

GO2320xlsx<-read.xlsx('GO_term_summary_20180922_140331.xlsx', sheetIndex = 1)
matchind<-match(GO2320xlsx$Symbol, diffmarkers$gene)
matchind<-matchind[-which(is.na(matchind))]
matchedgenes<-diffmarkers[matchind,]
matchedgenes<-arrange(matchedgenes, p_val)
```

```
## Warning: package 'bindrcpp' was built under R version 3.5.1
```

It appears that clusters 0 and 1 have the most genes that match. Expression values are plotted to visualize expression within each cell.

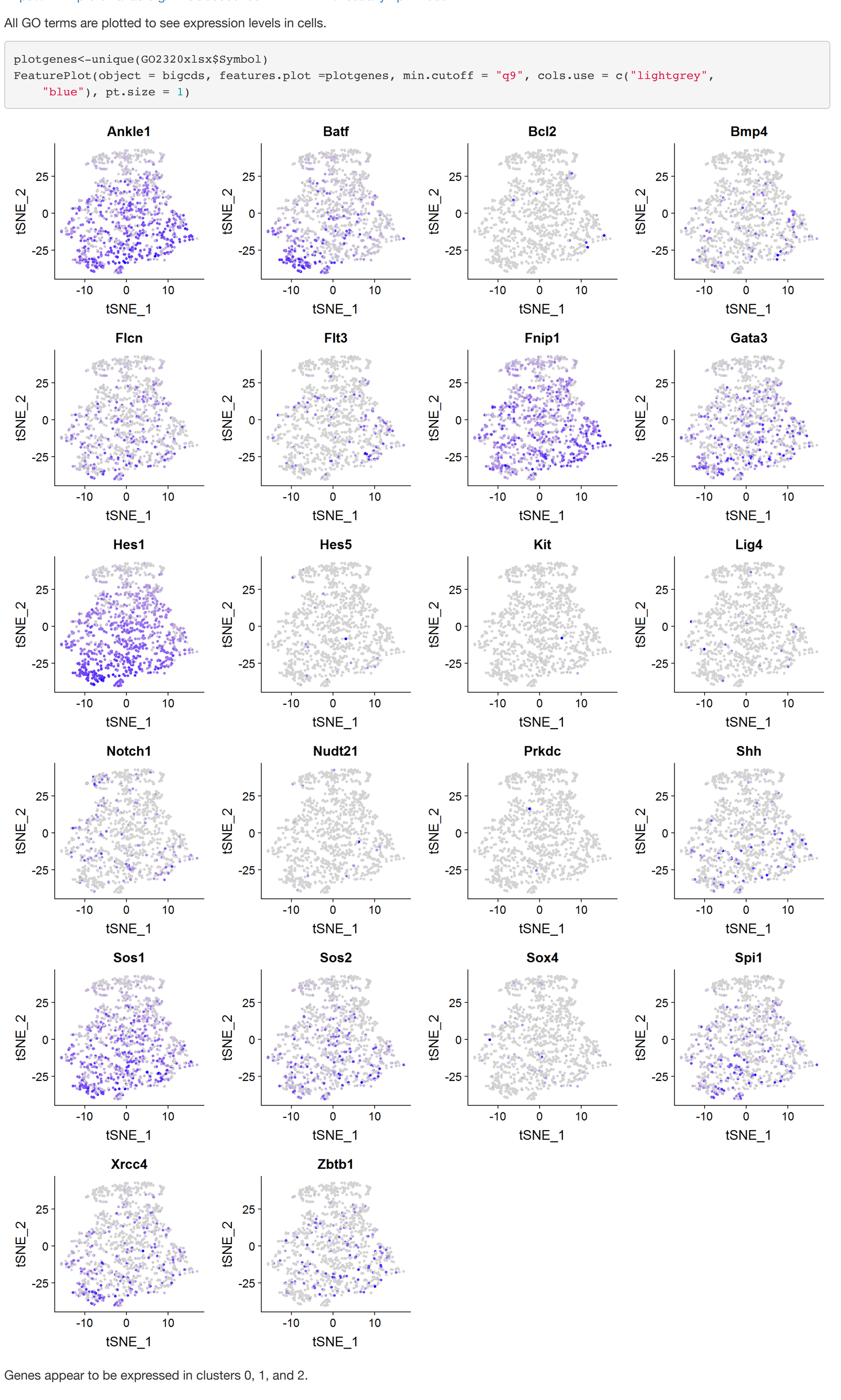
```
plotgenes<-unique(matchedgenes$gene)
FeaturePlot(object = bigcds, features.plot =plotgenes, min.cutoff = "q9", cols.use = c("lightgrey",
    "blue"), pt.size = 1)
```



Expression values appear to be throughout clusters 0 and 1 . Ankle 1 codes for a protein that is highly expressed in lymphocytes https://www.proteinatlas.org/ENSG00000160117-ANKLE1/tissue/lymph+node

All GO terms are plotted to see expression levels in cells.

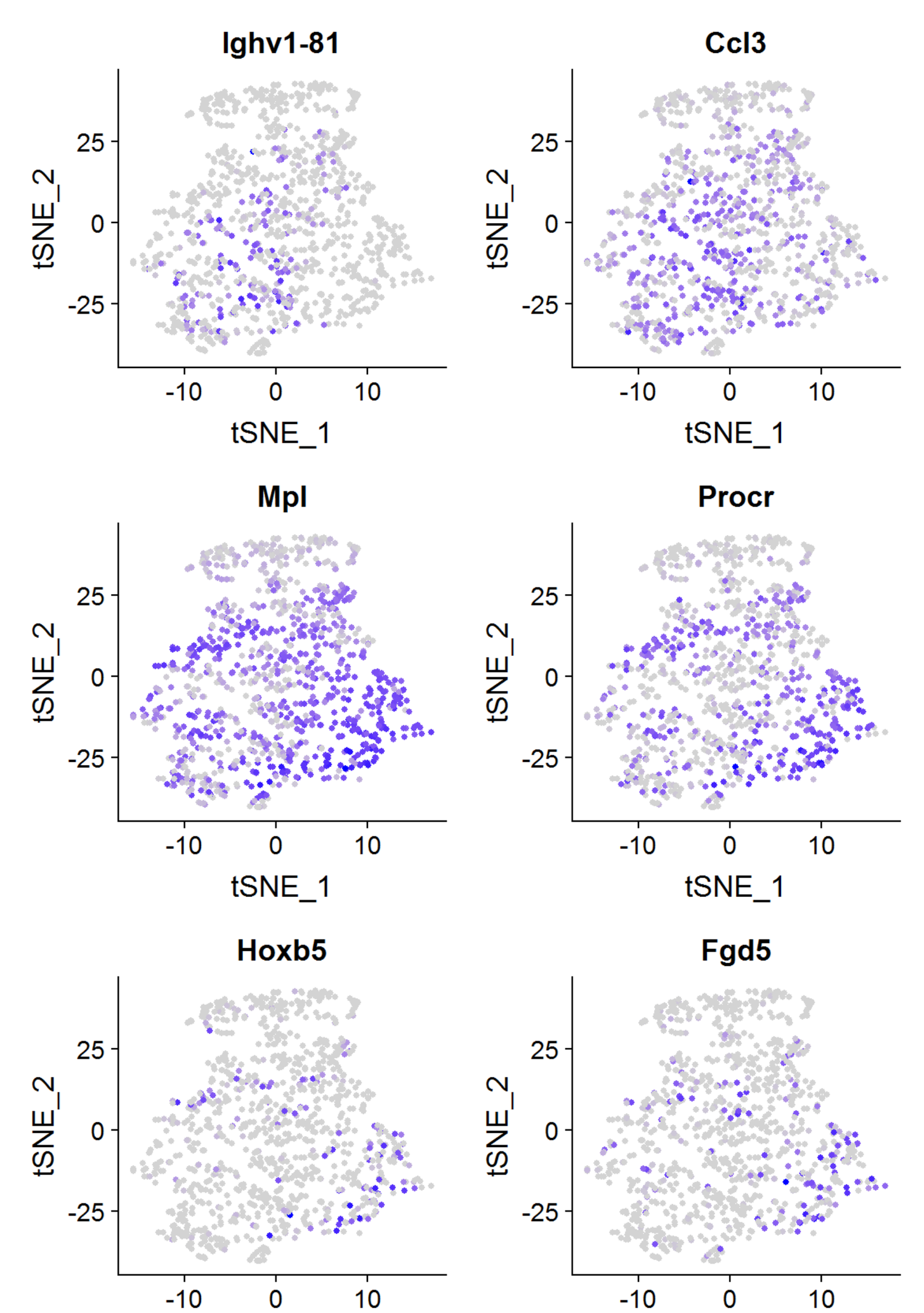
```
plotgenes<-unique(GO2320xlsx$Symbol)
FeaturePlot(object = bigcds, features.plot =plotgenes, min.cutoff = "q9", cols.use = c("lightgrey",
    "blue"), pt.size = 1)
```



Genes appear to be expressed in clusters 0, 1, and 2.

Figure 2c genes

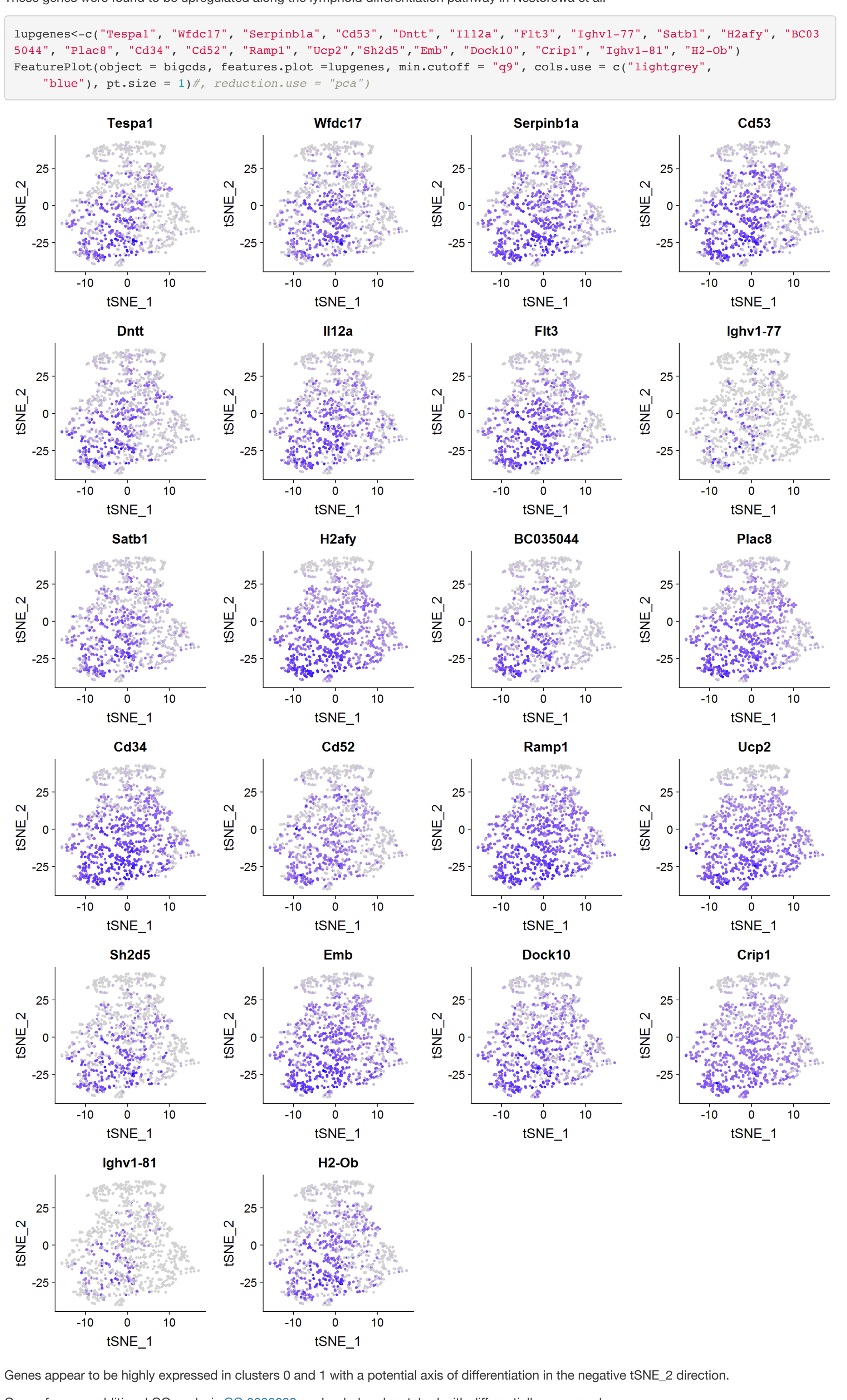
```
FeaturePlot(object = bigcds, features.plot =cgenes, min.cutoff = "q9", cols.use = c("lightgrey",
    "blue"), pt.size = 1)
```



Differentially expressed genes found for lymphocytes in the Nestorowa analysis appear to be expressed in clusters 0 and 1.

These genes were found to be upregulated along the lymphoid differentiation pathway in Nestorowa et al.

```
lupgenes<-c("Tespa1", "Wfdc17", "Serpinb1a", "Cd53", "Dntt", "Il12a", "Flt3", "Ighv1-77", "Satb1", "H2afy", "BC03
5044", "Plac8", "Cd34", "Cd52", "Ramp1", "Ucp2","Sh2d5","Emb", "Dock10", "Crip1", "Ighv1-81", "H2-Ob")
FeaturePlot(object = bigcds, features.plot =lupgenes, min.cutoff = "q9", cols.use = c("lightgrey",
    "blue"), pt.size = 1)#, reduction.use = "pca")
```

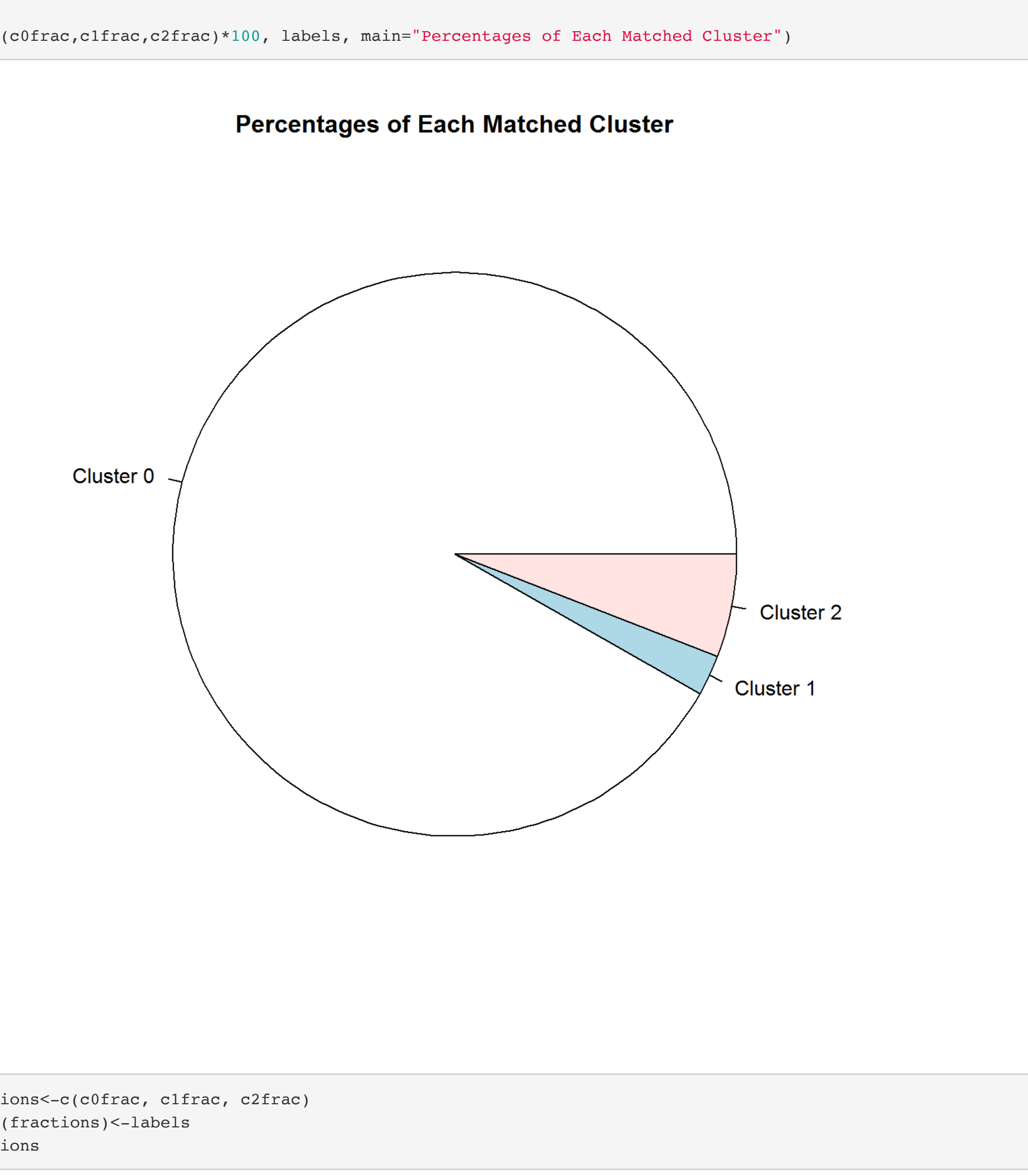


Genes appear to be highly expressed in clusters 0 and 1 with a potential axis of differentiation in the negative tSNE_2 direction.

Genes from an additional GO analysis GO:0030098 are loaded and matched with differentially expressed genes.

```
loadgo<-read.xlsx('GO_term_summary_20180920_140528.xlsx', sheetIndex = 1)
lymphoidsymbols<-as.character(loadgo$Symbol)
lymphoidsymbols<-unique(lymphoidsymbols)
matchind<-match(lymphoidsymbols, diffmarkers$gene)
matchind<-matchind[-which(is.na(matchind))]
matchedgenes<-diffmarkers[matchind,]
matchedgenes<-arrange(matchedgenes, p_val)
c0frac<-length(which(matchedgenes$cluster==0))/length(matchedgenes$cluster)
c1frac<-length(which(matchedgenes$cluster==1))/length(matchedgenes$cluster)
c2frac<-length(which(matchedgenes$cluster==2))/length(matchedgenes$cluster)
labels<-c("Cluster 0", "Cluster 1", "Cluster 2")

pie(c(c0frac,c1frac,c2frac)*100, labels, main="Percentages of Each Matched Cluster")
```




```
fractions<-c(c0frac, c1frac, c2frac)
names(fractions)<-labels
fractions
```

```
##  Cluster 0  Cluster 1  Cluster 2
## 0.91764706 0.02352941 0.05882353
```

It appears that 94% of the matched genes for lymphocyte differentiation are in clusters 0 and 1. After these analyses, it would be fair to say lymphocytes are present in clusters 0 and 1. Cells names with cluster designations of 0 and 1 are saved and Monocle is used to sort these cells. The 22290 genes are also exported for monocle analyses.

```
#Cells
clusters<-bigcds@meta.data$res.1
names(clusters)<-colnames(bigcds@data)
lymphoidcells<-clusters[which(clusters==0 | clusters==1)]
#Genes
goodgenes<-rownames(bigcds@data)

write.xlsx(goodgenes,'lymphoidgenes.xlsx')
write.xlsx(lymphoidcells,'lymphoidcells.xlsx')
```