Clustering Stage 1

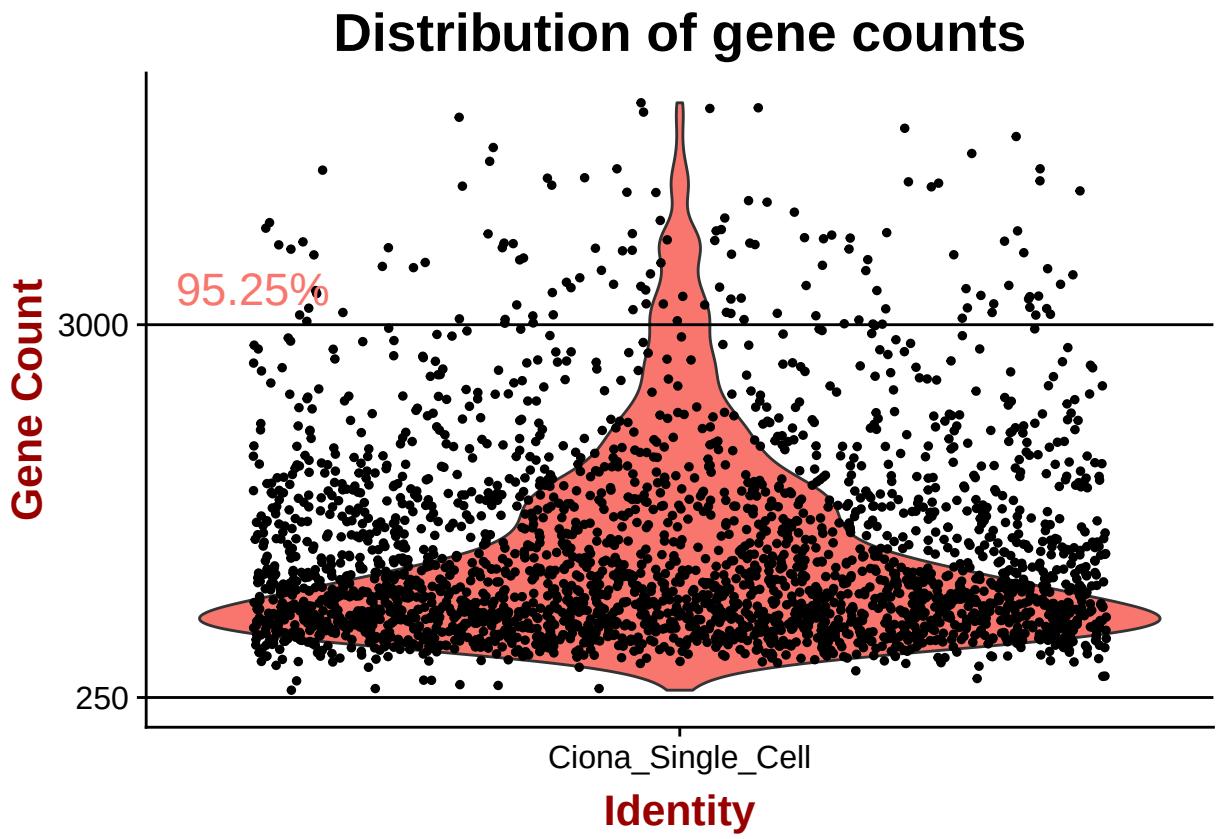
Sarthak Sharma

May 2, 2018

Preprocessing

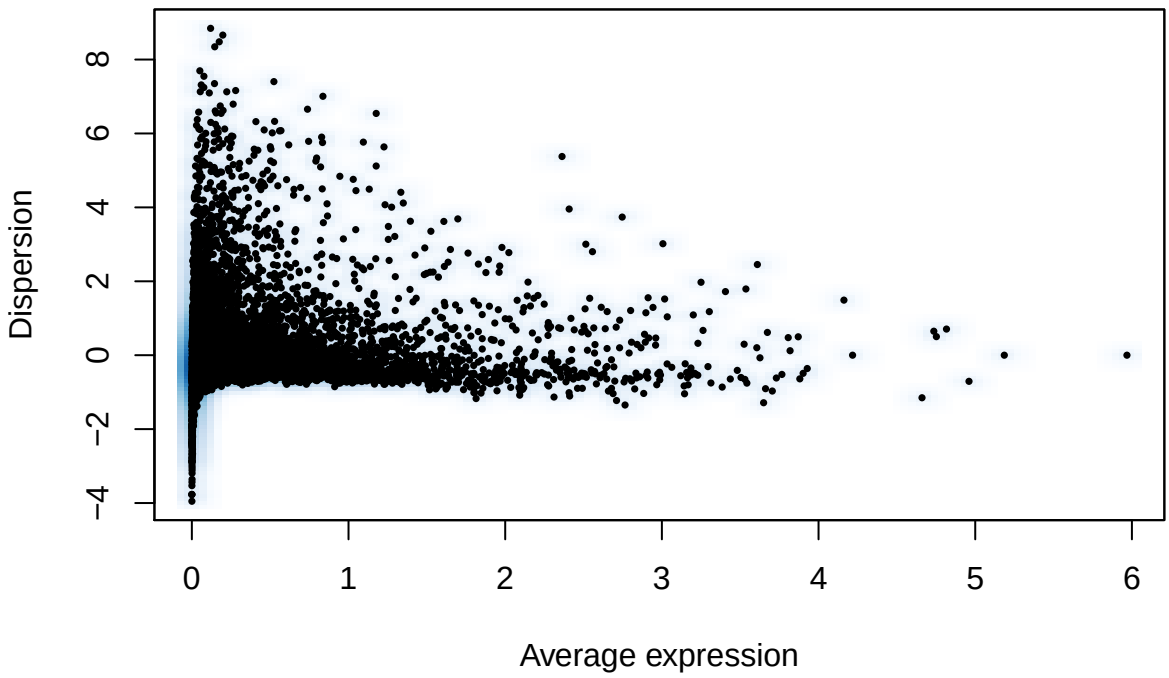
To filter out low quality single cell transcriptomes, we selected genes which were expressed in at least 3 cells and selected cells which expressed a minimum of 250 genes and a maximum of 3000 genes.

```
cb.data <- Read10X(data.dir = "~/KH2013/")
cbObject <- CreateSeuratObject(raw.data = cb.data, min.cells = 3, min.genes = 200, normalization.method
cb.filtered <- FilterCells(object = cbObject, subset.names = c("nGene"), low.thresholds = c(250), high.
```



Identification of Variable Genes

```
cb.filtered <- FindVariableGenes(object=cb.filtered, mean.function = ExpMean,
                                 dispersion.function = LogVMR, x.low.cutoff = 0.0125,
                                 x.high.cutoff = 3, y.cutoff = 0.5, do.plot = F)
```



Dimensional Reduction Analysis

The data was scaled to remove unwanted sources of variation. The variable genes were used as input to the PCA. The PCA Scores were then projected onto the rest of the genes.

```
cb.filtered <- ScaleData(object = cb.filtered, vars.to.regress = c("nUMI"))
```

```
cb.filtered <- RunPCA(cb.filtered, pc.genes = cb.filtered@var.genes, do.print = T, pcs.print = 1:5, gene
```

```
## [1] "PC1"
## [1] "KH2013:KH.C6.197"  "KH2013:KH.C13.119" "KH2013:KH.C1.212"
## [4] "KH2013:KH.C13.16"  "KH2013:KH.C2.1055"
## [1] ""
## [1] "KH2013:KH.C9.234"  "KH2013:KH.C8.400"  "KH2013:KH.C10.332"
## [4] "KH2013:KH.C12.16"  "KH2013:KH.C1.118"
## [1] ""
## [1] ""
## [1] "PC2"
## [1] "KH2013:KH.C1.312"  "KH2013:KH.C4.577"  "KH2013:KH.C1.266"
## [4] "KH2013:KH.L132.11" "KH2013:KH.L18.135"
## [1] ""
## [1] "KH2013:KH.C8.717"  "KH2013:KH.C1.118"  "KH2013:KH.C8.400"
## [4] "KH2013:KH.C2.101"  "KH2013:KH.C10.332"
## [1] ""
## [1] ""
## [1] "PC3"
```

```
## [1] "KH2013:KH.S1182.1" "KH2013:KH.C4.577"  "KH2013:KH.C1.266"
## [4] "KH2013:KH.L18.125" "KH2013:KH.C1.312"
## [1] ""
## [1] "KH2013:KH.C5.449" "KH2013:KH.C1.706" "KH2013:KH.C1.799"
## [4] "KH2013:KH.C2.848" "KH2013:KH.L41.41"
## [1] ""
## [1] ""
## [1] "PC4"
## [1] "KH2013:KH.C2.101"  "KH2013:KH.S1104.4" "KH2013:KH.C8.717"
## [4] "KH2013:KH.C12.16"  "KH2013:KH.C4.127"
## [1] ""
## [1] "KH2013:KH.C10.609" "KH2013:KH.C12.424" "KH2013:KH.C3.585"
## [4] "KH2013:KH.S390.1"  "KH2013:KH.L116.49"
## [1] ""
## [1] ""
## [1] "PC5"
## [1] "KH2013:KH.L36.7"  "KH2013:KH.C7.391" "KH2013:KH.C2.910"
## [4] "KH2013:KH.C11.48" "KH2013:KH.C4.428"
## [1] ""
## [1] "KH2013:KH.C4.73"   "KH2013:KH.C5.369"  "KH2013:KH.C1.802"
## [4] "KH2013:KH.C14.293" "KH2013:KH.L140.5"
## [1] ""
## [1] ""
```

```
cb.filtered <- ProjectPCA(cb.filtered, do.print = F)
```

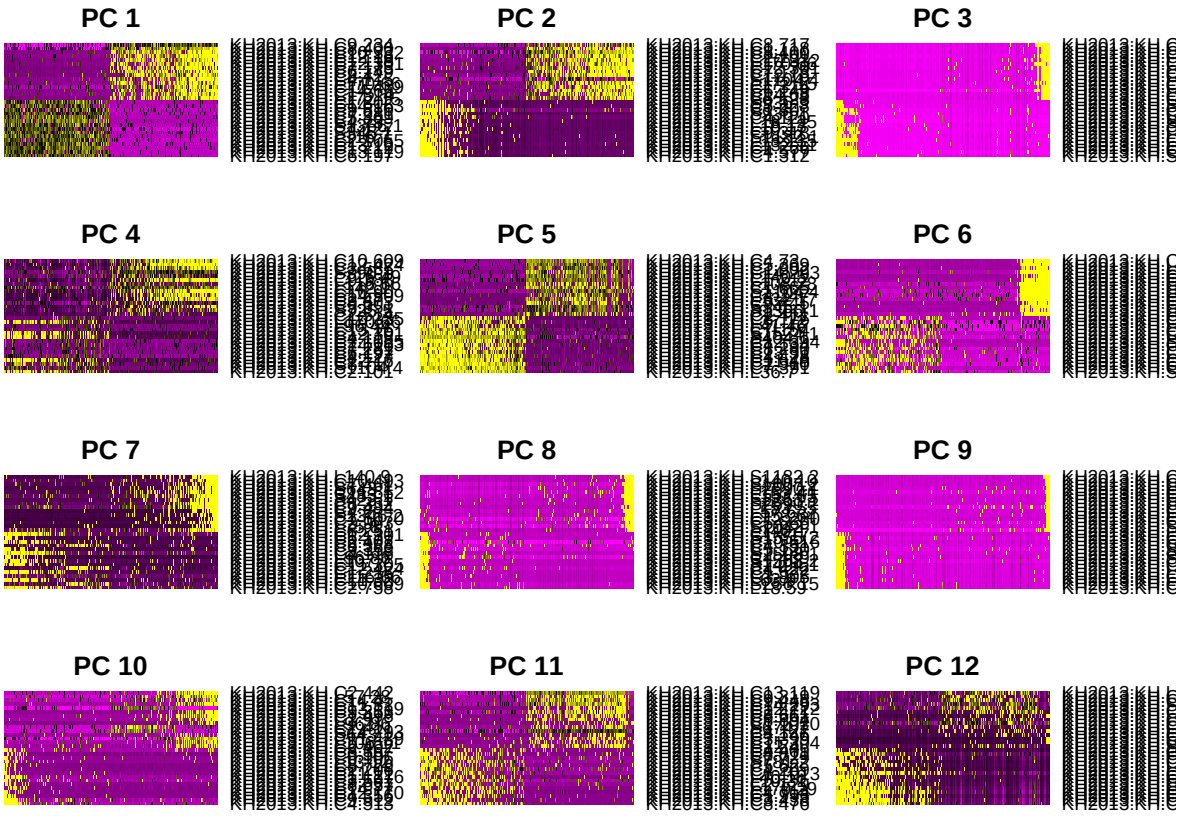
Determination of statistically significant PCs

To determine statistically significant and suitable PCs, we performed a supervised analysis as described by the 'Seurat' Documentation (http://satijalab.org/seurat/).

Heatmap and pairwise comparison

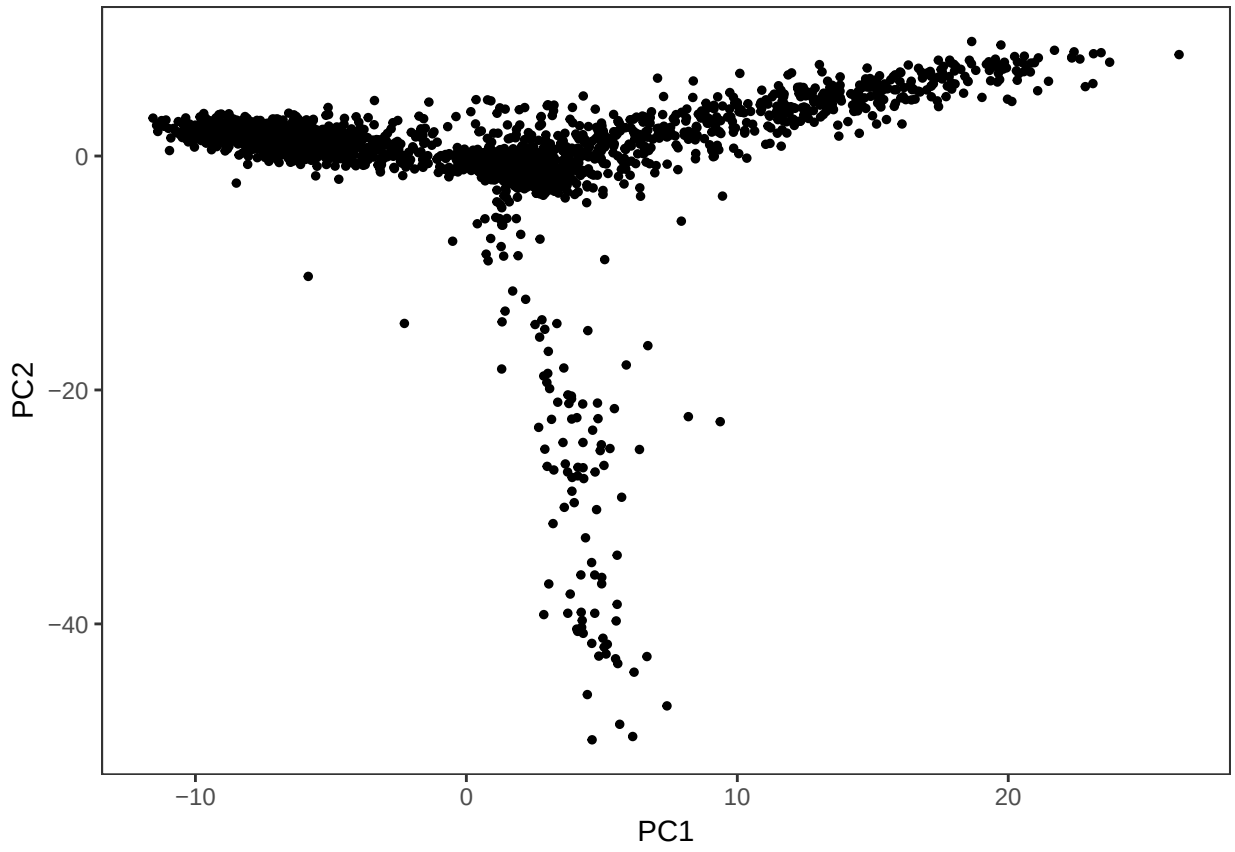
Heatmap

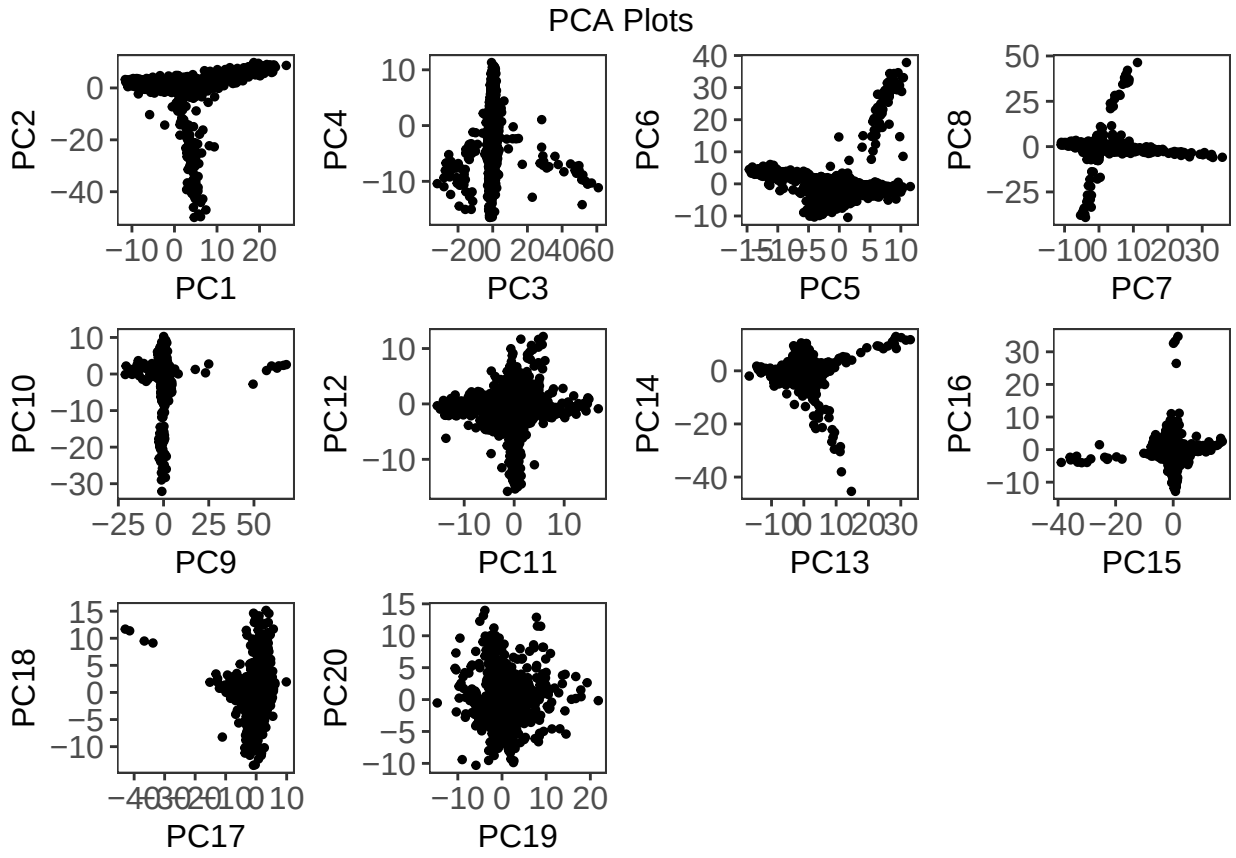
```
PCHeatmap(cb.filtered, pc.use = 1:12,cells.use = 500,do.balanced = T, label.columns = F, use.full = F, l
```



Pairwise comparison of PCs

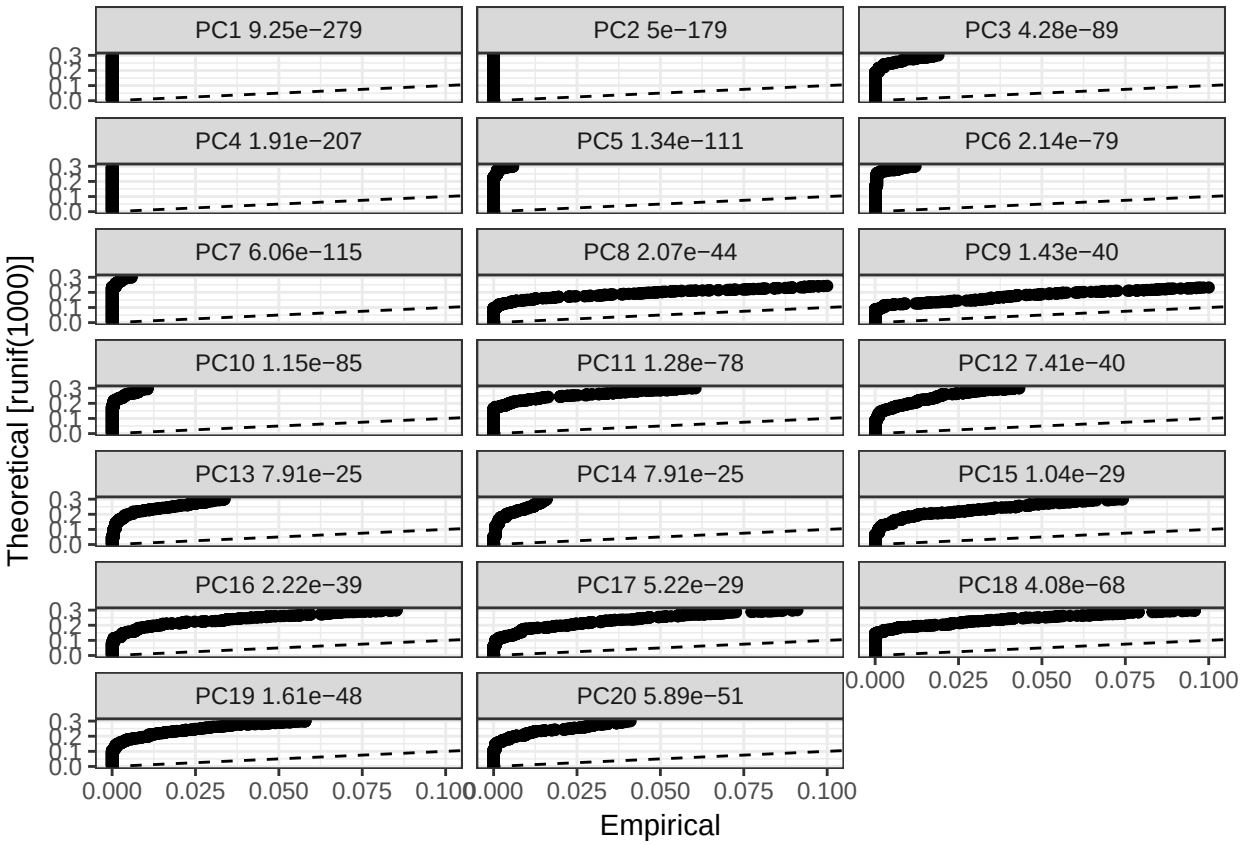
```
# Removing the colors use cols.use
PCAPlot(cb.filtered,dim.1 = 1, dim.2 = 2,do.return = F, no.legend = T,
                    cols.use = rep(c("black"),times=10))
```


0
−20
−40
PC2
−10
0
10
20
PC1



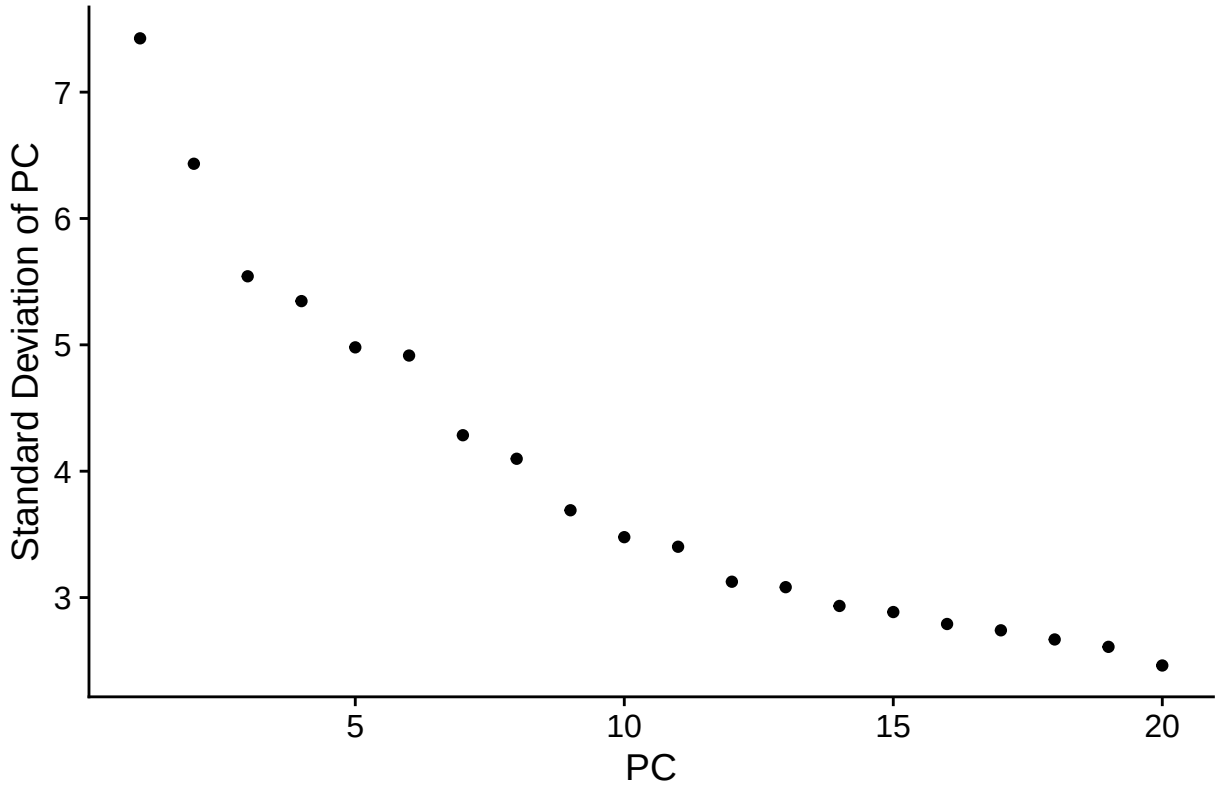
JackStraw resampling test with 100 replications

```
cb.filtered <- JackStraw(cb.filtered, num.replicate = 100, do.print = F)
```



Plot of Standard Deviations for each of the PCs

```
PCElbowPlot(cb.filtered)
```



Single Cell Clustering and differential gene expression

Find and Validate Clusters

```
cb.filtered <- FindClusters(cb.filtered, reduction.type = "pca",
                            dims.use = 1:7, resolution = 0.6,
                            print.output = 0, save.SNN = T, force.recalc = T)

ValidateClusters(cb.filtered,pc.use = 1:7,top.genes = 30,
                 min.connectivity = 0.001,acc.cutoff = 0.9,verbose = T)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'kernlab'
```

```
## The following object is masked from 'package:ggplot2':
##
##     alpha
```

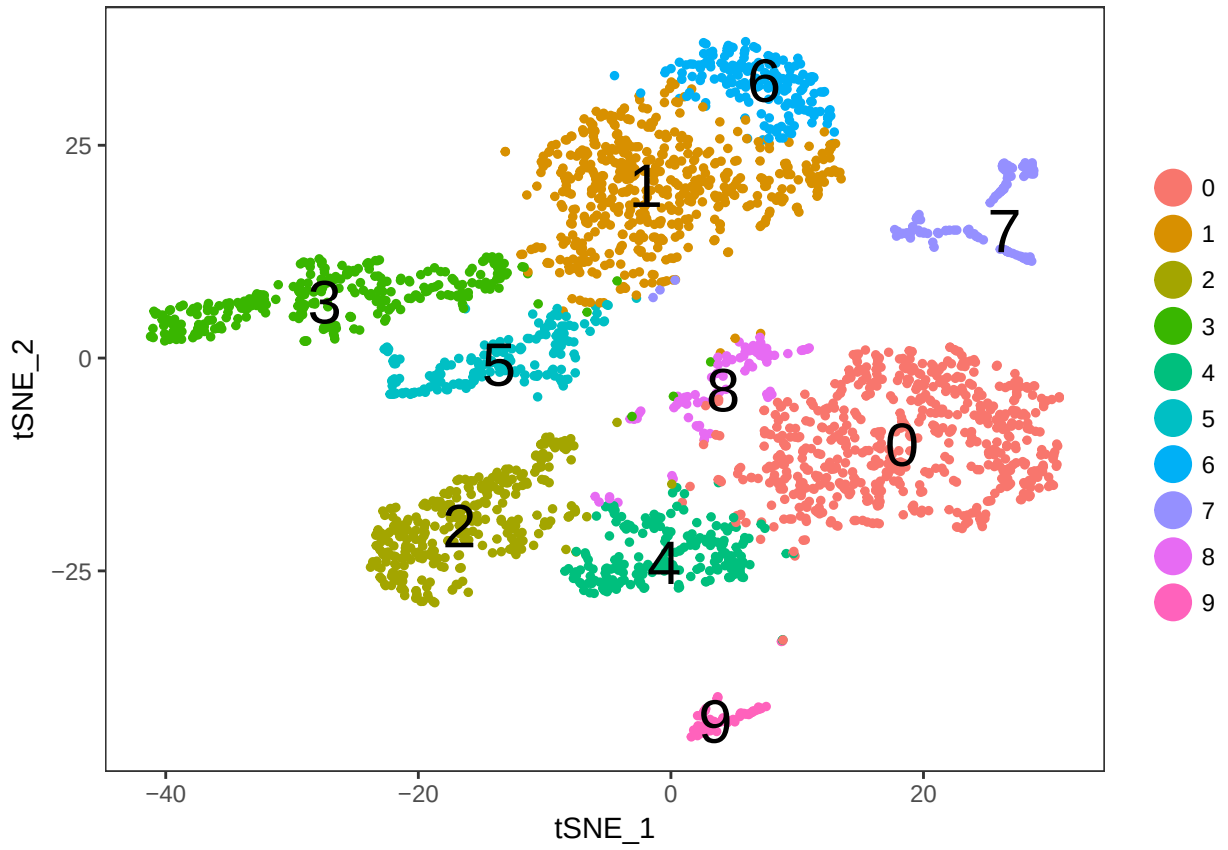
```
## [1] "  0% complete --- Last 5 cluster comparisons failed to merge, still checking possible merges ..
## [1] "100% complete --- started with 10 clusters, 10 clusters remaining"
```

```
## An object of class seurat in project Ciona_Single_Cell
##  13435 genes across 2607 samples.
```

TSNE Plot

Project the clusters onto 2-D space

```
cb.filtered <- RunTSNE(cb.filtered, dims.use = 1:7, do.fast = T)
TSNEPlot(cb.filtered,do.label = T, label.size = 8,pt.size = 1)
```



Find Differentially Expressed Gene Markers

```
cb.filtered.allMarkers <- FindAllMarkers(cb.filtered)
```

We can now use the obtained markers to output them to csv files or use them in other analyses.

Save the R object

```
saveRDS(cb.filtered, file = "~/cb_filtered.rds")
```