scRNA-Seq Lab

(solution)

2017-02-19

Introduction/Quality control

For this exercise, we will be analyzing the a single cell RNA-Seq dataset of Peripheral Blood Mononuclear Cells (PBMC) from the 10X Genomics platform. We will primarily be using the seurat package from the Satija Lab, which includes a vignette here. http://satijalab.org/seurat/pbmc-tutorial.html. The purpose of single cell RNA-Seq analysis is to uncover interesting biology that occurs at a granularity–the single cell–that isn't appreciated when these features become averaged in bulk. The goal of this analysis is to uncover heterogenity in PBMCs and understanding the analysis workflows for single cell technologies.

First, load the packages and the data object

```
library(Seurat)
library(Matrix)
library(dplyr)

pbmc <- readRDS("pbmc.rds")
dim(pbmc@raw.data)
```

```
## [1] 32643  2001
```

Note: to achieve this object, the counts matrix had to be determined using a standard alignment protocol similar to bulk RNA-Seq analyses. The **.rds** object contains a **seurat** object with 2001 samples and over 32,000 genes. This sample set includes roughly 1,000 PBMC samples from two different batches.

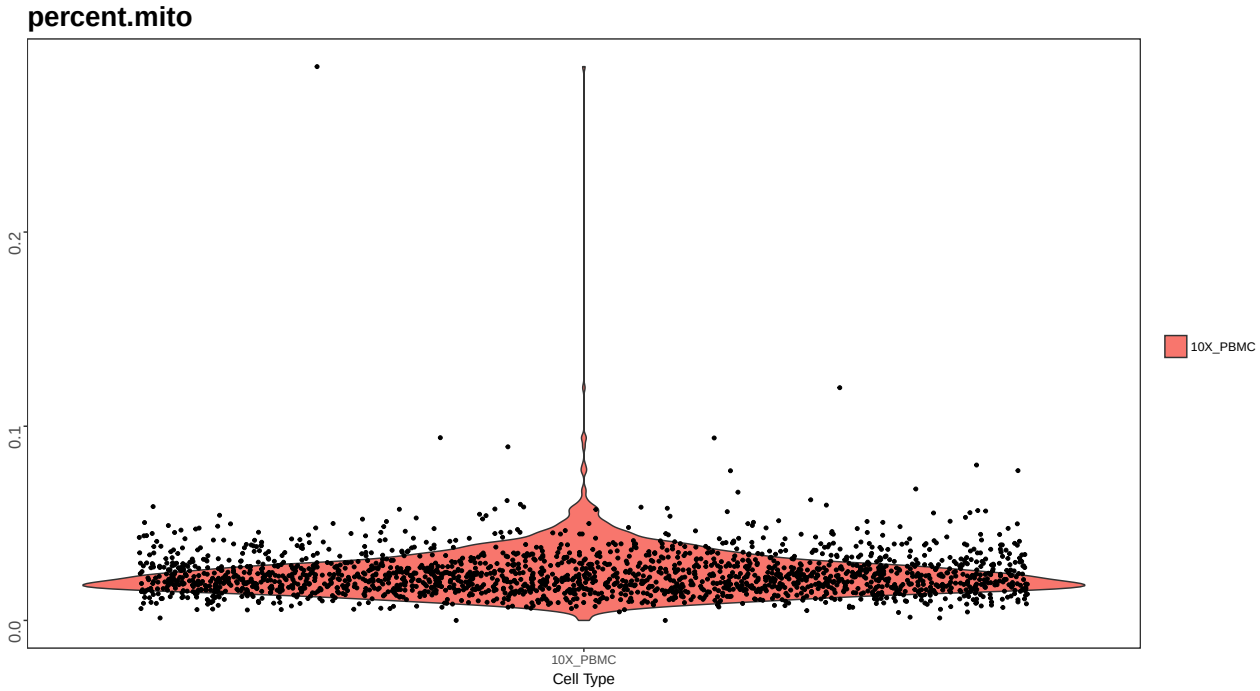
Analysis

(1) The substantial sparsity associated with scRNA-Seq data makes analysis a unique challenge. Use the 'Setup' command to filter lowly expressed genes and weakly detected cells in this raw dataset. How many samples and genes are filtered afterwards?

```
pbmc <- Setup(pbmc, min.cells = 3, min.genes = 200, do.logNormalize = TRUE,
              total.expr = 1e4, project = "10X_PBMC")
# use 'str' command to determine still 2,000 samples and 12,800 genes remain after filter
```

(2a) When trying to discover rare cell types, one has to be weary of technical confounders that imply heterogenity that are actually false. Two measures of technical confounders are the number of mitochondrial reads mapped as well as the number of unique genes mapped. In this dataset, how many mitochondrial genes are there? What is the distribution of the proportion of reads for these mitochondrial genes? How many samples express a number of genes that significantly deviates from the rest?

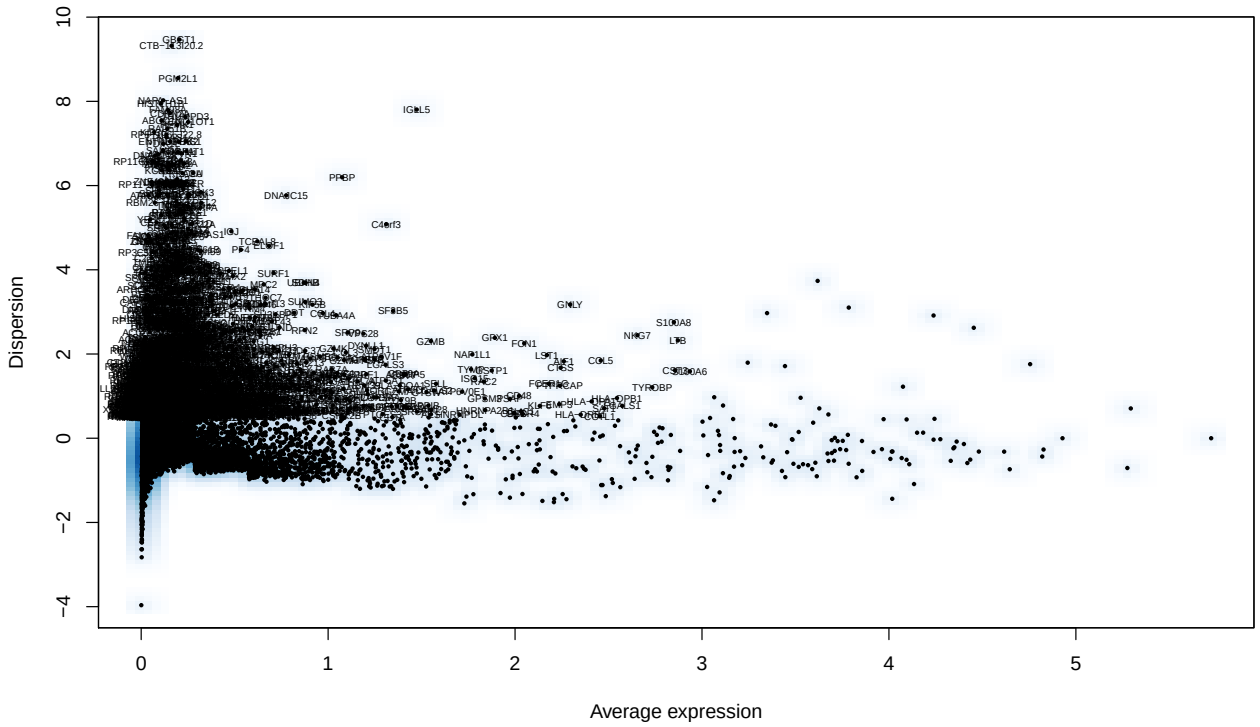
```
mito.genes <- grep("^MT-", rownames(pbmc@data), value = TRUE)
percent.mito <- colSums(expm1(pbmc@data[mito.genes, ]))/colSums(expm1(pbmc@data))

#AddMetaData adds columns to object@data.info, and is a great place to stash QC stats
pbmc <- AddMetaData(pbmc, percent.mito, "percent.mito")
VlnPlot(pbmc, c("percent.mito"), nCol = 1)
```




```
pbmc <- SubsetData(pbmc, subset.name = "nGene", accept.high = 2500)
pbmc <- SubsetData(pbmc, subset.name = "percent.mito", accept.high = 0.05)
```

(3) In scRNA-Seq, why is specifying the mean-variance relationship important? Create a Mean-Variance plot to assess this dataset. Comment on its appearance. How many variable genes using reasonable cutoffs remain?

```
pbmc <- MeanVarPlot(pbmc ,fxn.x = expMean, fxn.y = logVarDivMean,
                    x.low.cutoff = 0.0125, x.high.cutoff = 3, y.cutoff = 0.5, do.contour = FALSE)
```




```
length(pbmc@var.genes)
```

```
## [1] 1793
```

Linear/Non-linear dimensional reduction

(4a) Rather than focusing on specific differentially expressed genes, a staple in scRNA-Seq analyses involves dimension reduction. Compute the top principal components using the variable genes and determine which genes contribute most to these PCs. Find some other ways to display and interpret the results of the dimensionality reduction.

```
pbmc <- PCA(pbmc, pc.genes = pbmc@var.genes, do.print = FALSE, pcs.print = 5, genes.print = 5)
```

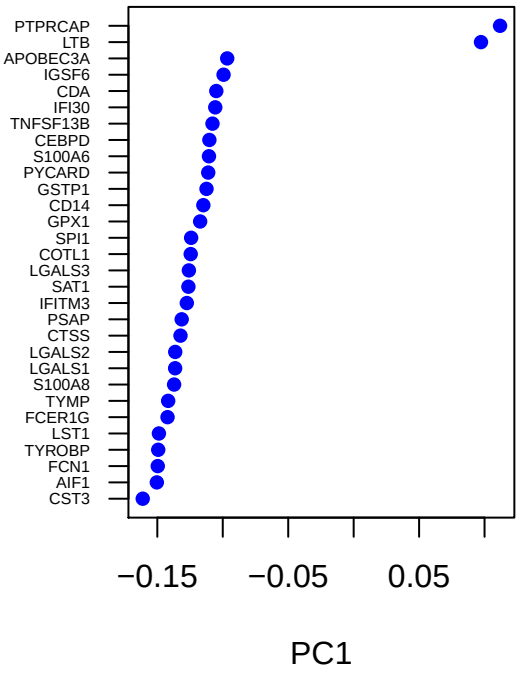
Seurat provides several useful ways of visualizing both cells and genes that define the PCA, including **PrintPCA()**, **VizPCA()**, **PCAPlot()**, and **PCHeatmap()**

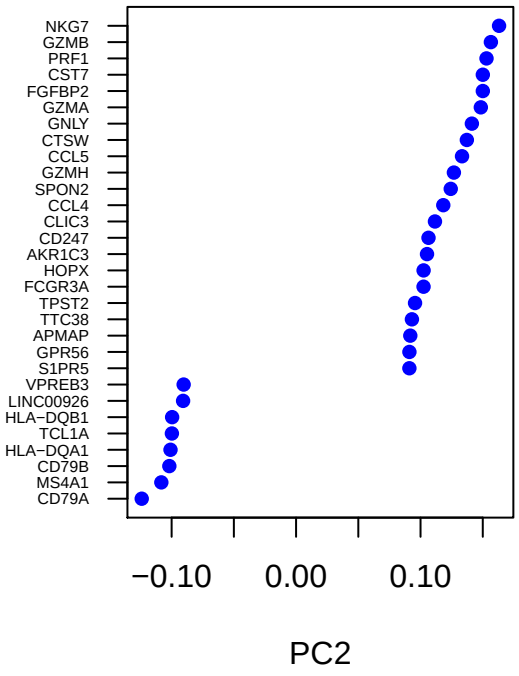
```
# Examine  and visualize PCA results a few different ways
PrintPCA(pbmc, pcs.print = 1:5, genes.print = 5, use.full = FALSE)
```

```
## [1] "PC1"
## [1] "CST3"    "AIF1"    "FCN1"    "TYROBP" "LST1"
## [1] ""
## [1] "PTPRCAP" "LTB"     "CD3D"    "CXCR4"   "AES"
## [1] ""
## [1] ""
## [1] "PC2"
## [1] "CD79A"    "MS4A1"    "CD79B"    "HLA-DQA1" "TCL1A"
## [1] ""
## [1] "NKG7"   "GZMB"   "PRF1"   "CST7"   "FGFBP2"
```

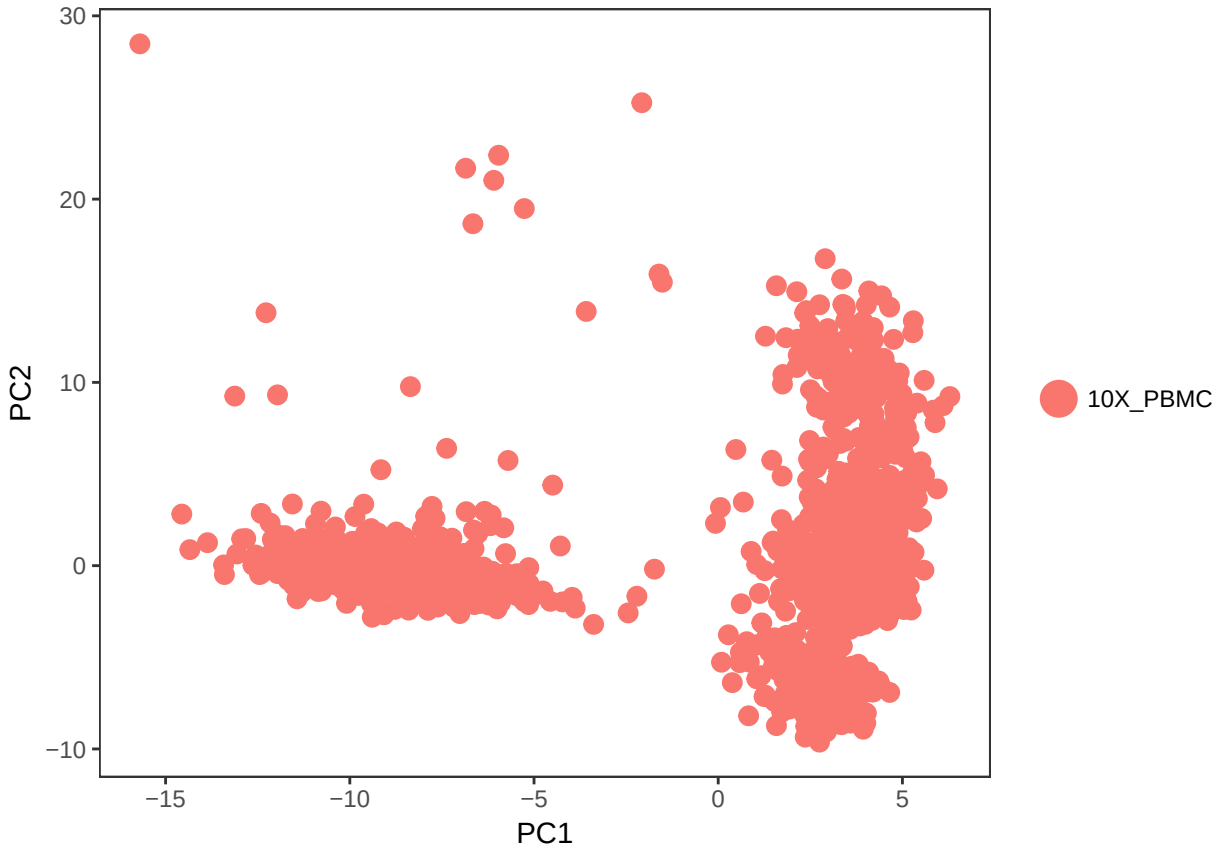
```
## [1] ""
## [1] ""
## [1] "PC3"
## [1] "NKG7"   "PRF1"   "CST7"   "FGFBP2" "GZMB"
## [1] ""
## [1] "PPBP"      "PF4"       "SDPR"      "GNG11"     "HIST1H2AC"
## [1] ""
## [1] ""
## [1] "PC4"
## [1] "CD3D"   "FYB"    "RGCC"   "CD27"   "NDFIP1"
## [1] ""
## [1] "CD79A"    "HLA-DQA1" "CD79B"    "HLA-DQB1" "HLA-DPB1"
## [1] ""
## [1] ""
## [1] "PC5"
## [1] "NAP1L1"    "LTB"       "HNRNPA2B1" "ABI3"      "PTGES3"
## [1] ""
## [1] "S100A8"  "S100A12" "CD14"    "CLEC4E"  "GZMB"
## [1] ""
## [1] ""
```

```
VizPCA(pbmc, 1:2)
```

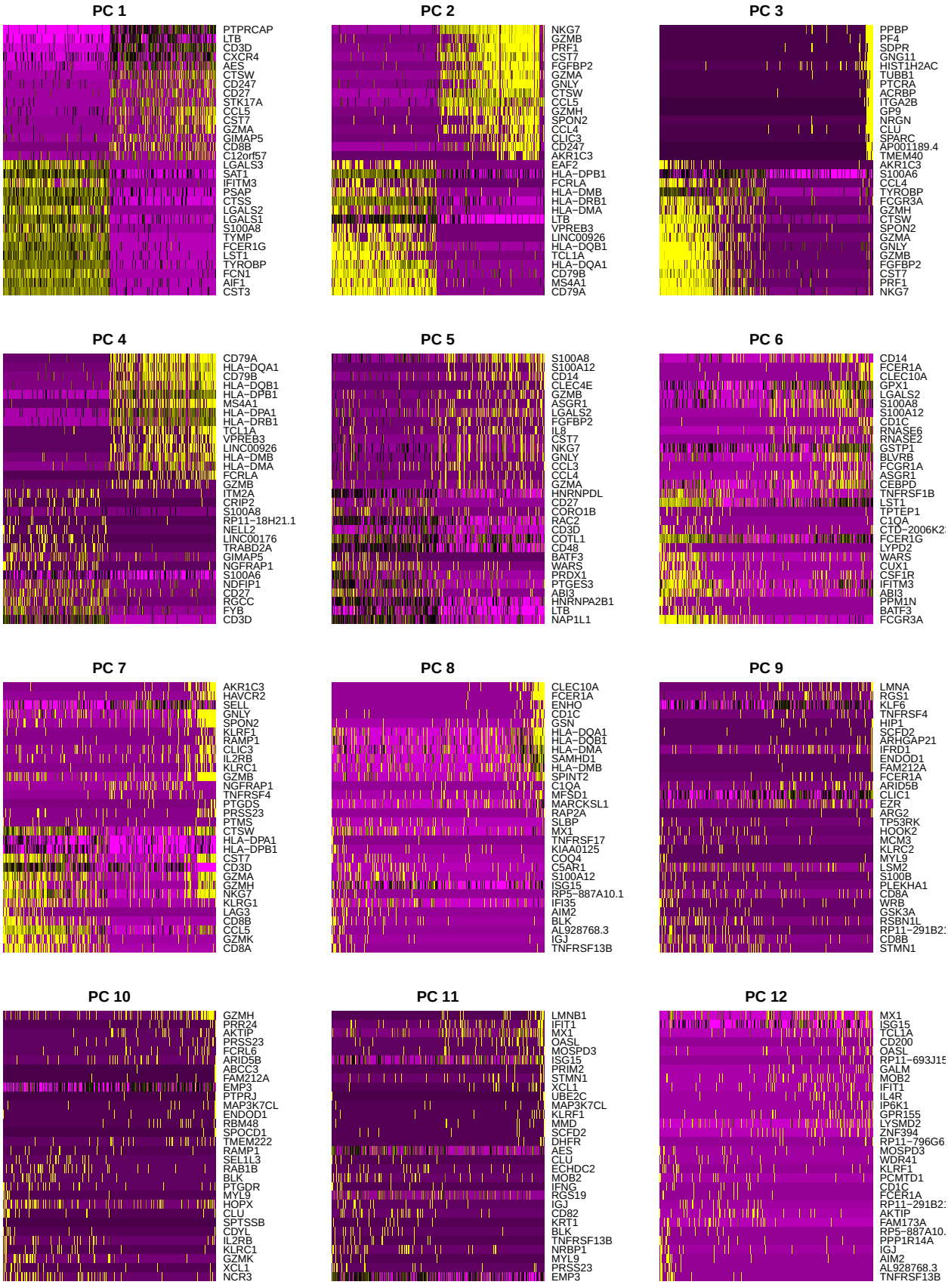





```
PCAPlot(pbmc, 1, 2)
```

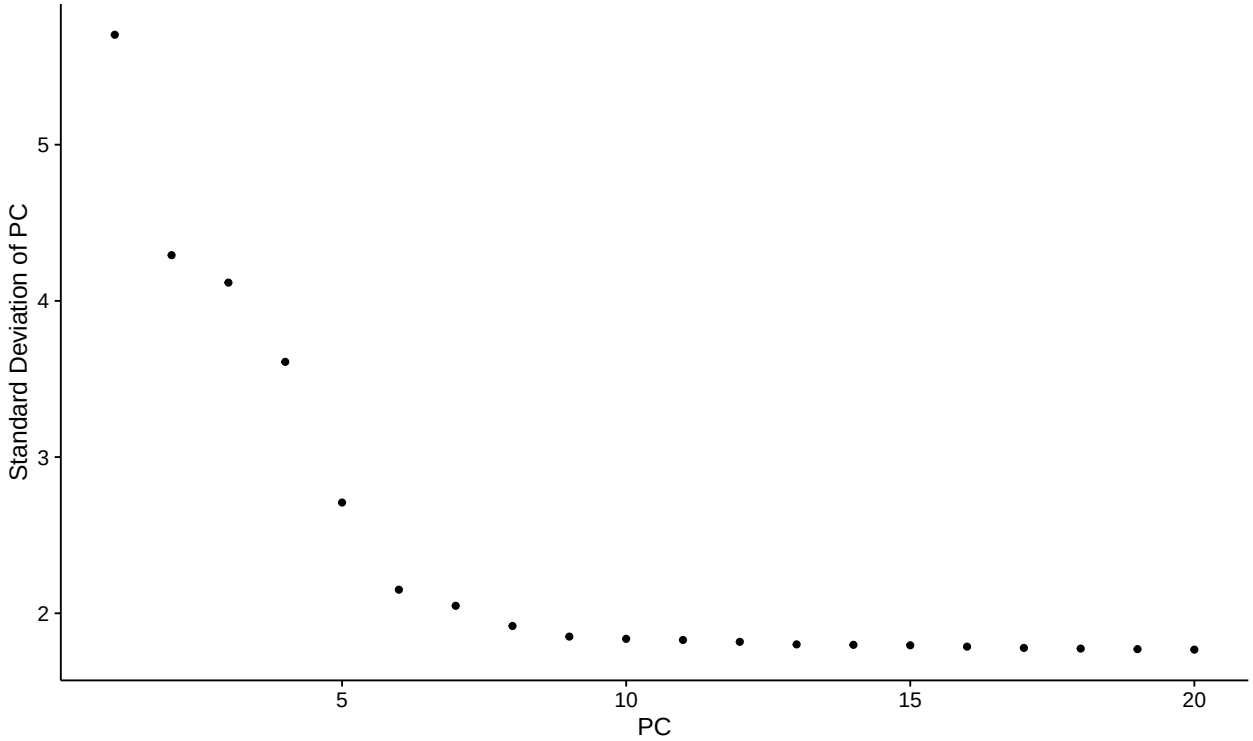



```
PCHeatmap(pbmc, pc.use = 1:12, cells.use = 500, do.balanced = TRUE,
          label.columns = FALSE, use.full = FALSE)
```


PC 1
PTPRCAP
LTB
CD3D
CXCR4
AES
CTSW
CD247
CD27
STK17A
CCL5
CST7
GZMA
GIMAP5
CD8B
C12orf57
LGALS3
SAT1
IFITM3
PSAP
CTSS
LGALS2
LGALS1
S100A8
TYMP
FCER1G
LST1
TYROBP
FCN1
AIF1
CST3
PC 2
NKG7
GZMB
PRF1
CST7
FGFBP2
GZMA
GNLY
CTSW
CCL5
GZMH
SPON2
CCL4
CLIC3
CD247
AKR1C3
EAF2
HLA-DPB1
FCRLA
HLA-DMB
HLA-DRB1
HLA-DMA
LTB
VPREB3
LINC00926
HLA-DQB1
TCL1A
HLA-DQA1
CD79B
MS4A1
CD79A
PC 3
PPBP
PF4
SDPR
GNG11
HIST1H2AC
TUBB1
PTCRA
ACRBP
ITGA2B
GP9
NRGN
CLU
SPARC
AP001189.4
TMEM40
AKR1C3
S100A6
CCL4
TYROBP
FCGR3A
GZMH
CTSW
SPON2
GZMA
GNLY
GZMB
FGFBP2
CST7
PRF1
NKG7
PC 4
CD79A
HLA-DQA1
CD79B
HLA-DQB1
HLA-DPB1
MS4A1
HLA-DPA1
HLA-DRB1
TCL1A
VPREB3
LINC00926
HLA-DMB
HLA-DMA
FCRLA
GZMB
ITM2A
CRIP2
S100A8
RP11-18H21.1
NELL2
LINC00176
TRABD2A
GIMAP5
NGFRAP1
S100A6
NDFIP1
CD27
RGCC
FYB
CD3D
PC 5
S100A8
S100A12
CD14
CLEC4E
GZMB
ASGR1
LGALS2
FGFBP2
IL8
CST7
NKG7
GNLY
CCL3
CCL4
GZMA
HNRNPDL
CD27
CORO1B
RAC2
CD3D
COTL1
CD48
BATF3
WARS
PRDX1
PTGES3
ABI3
HNRNPA2B1
LTB
NAP1L1
PC 6
CD14
FCER1A
CLEC10A
GPX1
LGALS2
S100A8
S100A12
CD1C
RNASE6
RNASE2
GSTP1
BLVRB
FCGR1A
ASGR1
CEBPD
TNFRSF1B
LST1
TPTEP1
C1QA
CTD-2006K2
FCER1G
LYPD2
WARS
CUX1
CSF1R
IFITM3
ABI3
PPM1N
BATF3
FCGR3A
PC 7
AKR1C3
HAVCR2
SELL
GNLY
SPON2
KLRF1
RAMP1
CLIC3
IL2RB
KLRC1
GZMB
NGFRAP1
TNFRSF4
PTGDS
PRSS23
PTMS
CTSW
HLA-DPA1
HLA-DPB1
CST7
CD3D
GZMA
GZMH
NKG7
KLRG1
LAG3
CD8B
CCL5
GZMK
CD8A
PC 8
CLEC10A
FCER1A
ENHO
CD1C
GSN
HLA-DQA1
HLA-DQB1
HLA-DMA
SAMHD1
HLA-DMB
SPINT2
C1QA
MFSD1
MARCKSL1
RAP2A
SLBP
MX1
TNFRSF17
KIAA0125
COQ4
C5AR1
S100A12
ISG15
RP5-887A10.1
IFI35
AIM2
BLK
AL928768.3
IGJ
TNFRSF13B
PC 9
LMNA
RGS1
KLF6
TNFRSF4
HIP1
SCFD2
ARHGAP21
IFRD1
ENDOD1
FAM212A
FCER1A
ARID5B
CLIC1
EZR
ARG2
TP53RK
HOPX
MCM3
KLRC2
MYL9
LSM2
S100B
PLEKHA1
CD8A
WRB
GSK3A
RSBN1L
RP11-291B2
CD8B
STMN1
PC 10
GZMH
PRR24
AKTIP
PRSS23
FCRL6
ARID5B
ABCC3
FAM212A
EMP3
PTPRJ
MAP3K7CL
ENDOD1
RBM48
SPOCD1
TMEM222
RAMP1
SEL1L3
RAB1B
BLK
PTGDR
MYL9
HOPX
CLU
SPTSSB
CDYL
IL2RB
KLRC1
GZMK
XCL1
NCR3
PC 11
LMNB1
IFIT1
MX1
OASL
MOSPD3
ISG15
PRIM2
STMN1
XCL1
UBE2C
MAP3K7CL
KLRF1
MMD
SCFD2
DHFR
AES
CLU
ECHDC2
MOB2
IFNG
RGS19
IGJ
CD82
KRT1
BLK
TNFRSF13B
NRBP1
MYL9
PRSS23
EMP3
PC 12
MX1
ISG15
TCL1A
CD200
OASL
RP11-693J15
GALM
MOB2
IFIT1
IL4R
IP6K1
GPR155
LYSMD2
ZNF394
RP11-796G6
MOSPD3
WDR41
KLF1
PCMTD1
CD1C
FCER1A
RP11-291B2
AKTIP
FAM173A
RP5-887A10.
PPP1R14A
IGJ
AIM2
AL928768.3
TNFRSF13B

(5) Which principal components are statistically significant? Comment on one or more approaches to determine this.

```
PCElbowPlot(pbmc)
```



A more ad hoc method for determining which PCs to use is to look at a plot of the standard deviations of the principle components and draw your cutoff where there is a clear elbow in the graph. This can be done with **PCElbowPlot()**. In this example, it looks like the elbow would fall around PC 9.

(6) Use the FindClusters command to determine sample modules in the PBMC data. Comment on the type of clustering performed. Is it supervised or unsupervised?

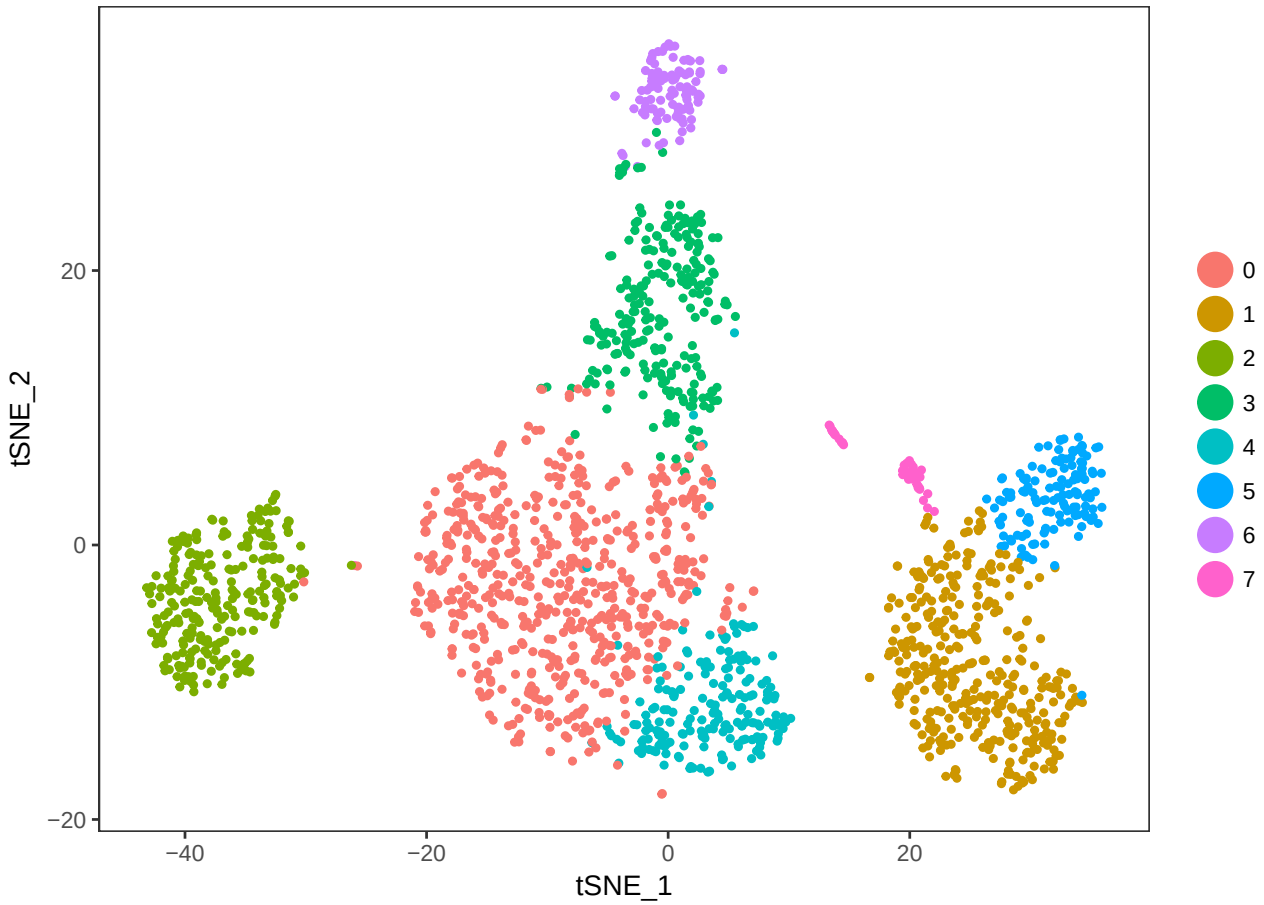
```
pbmc <- FindClusters(pbmc, pc.use = 1:10, resolution = 0.6, print.output = 0, save.SNN = T)
```

Details are in the Seurat source code as well as several paragraphs in the vignette. It is supervised.

(7a) A popular method for displaying scRNA-Seq data is by creating two dimensions using tSNE. Run and visualize tSNE for this data. Comment on how this approach is different than PCA.

```
pbmc <- RunTSNE(pbmc, dims.use = 1:10, do.fast = T)
```

```
TSNEPlot(pbmc)
```



Note: tSNE is, by definition, a stochastic process so be sure to cache your data at this point or save the file image before re-running later steps! Main difference is linear/non linear effects; tSNE in this case is using the PCs as input

---

Differentially expressed genes (cluster biomarkers)

(8) Now that we've defined data-driven clusters, we'd like to identify markers that define clusters via differential expression. What markers distinguish cluster 2? What markers distinguish cluster 2 from cluster 4? Every cluster from all others.

```
# find all markers of cluster 2
cluster2.markers <- FindMarkers(pbmc, ident.1 = 2, min.pct = 0.25)
print(head(cluster2.markers, 5))
```

```
##                 p_val avg_diff pct.1 pct.2
## CD74    4.626575e-276 2.003223 1.000 0.754
## CD79A   3.774364e-266 2.989068 0.933 0.038
## CD79B   6.083228e-243 2.474387 0.922 0.131
## HLA-DRA 1.383353e-201 1.839170 0.996 0.464
## MS4A1   5.769042e-193 2.333413 0.816 0.045
```

```
cluster24.markers <- FindMarkers(pbmc, 2, 4, min.pct = 0.25)
print(head(cluster24.markers, 5))
```

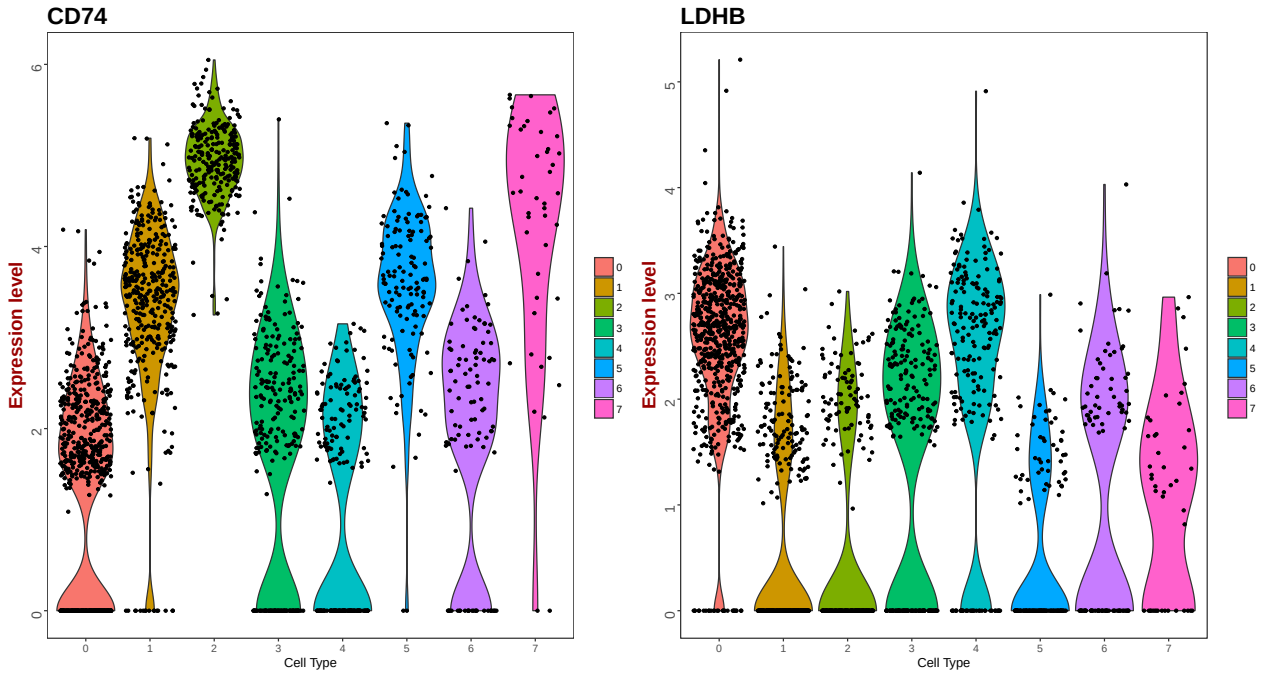
```
##                  p_val avg_diff pct.1 pct.2
## CD74     2.101969e-220 3.291178 1.000 0.520
## HLA-DRA  5.246923e-167 3.923024 0.996 0.133
## HLA-DPB1 7.946604e-123 3.046278 0.984 0.117
## HLA-DRB1 2.527228e-116 3.279697 0.973 0.026
## CD79A    1.155224e-103 3.064796 0.933 0.010
```

```
pbmc.markers <- FindAllMarkers(pbmc, only.pos = TRUE, min.pct = 0.25, thresh.use = 0.25)
pbmc.markers %>% group_by(cluster) %>% top_n(2, avg_diff)
```

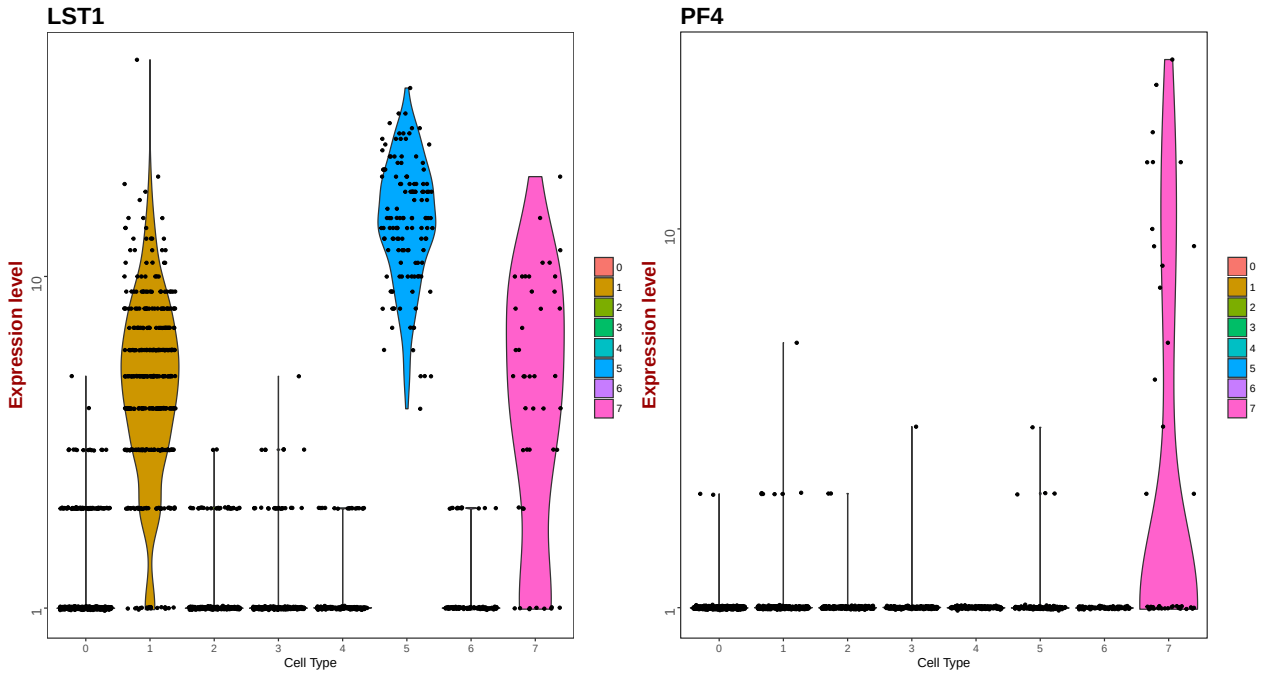
```
## Source: local data frame [16 x 6]
## Groups: cluster [8]
##
##            p_val  avg_diff pct.1 pct.2 cluster   gene
##            <dbl>     <dbl> <dbl> <dbl>  <fctr>  <chr>
## 1  4.565200e-146 1.0133146 0.956 0.506       0   LDHB
## 2  8.547608e-137 0.8729406 0.884 0.304       0   CD3D
## 3   0.000000e+00 3.8247836 0.984 0.109       1 S100A8
## 4   0.000000e+00 3.7477009 0.978 0.205       1 S100A9
## 5  3.774364e-266 2.9890678 0.933 0.038       2  CD79A
## 6  3.053285e-139 2.5657555 0.588 0.024       2  TCL1A
## 7  1.142654e-192 2.3393136 0.964 0.192       3   CCL5
## 8   3.050984e-95 2.2246331 0.543 0.033       3   GZMK
## 9   6.368285e-14 0.7227104 0.357 0.166       4   CCR7
## 10  6.434473e-12 0.7038178 0.291 0.130       4   LEF1
## 11 2.098630e-125 1.9500161 1.000 0.309       5   LST1
## 12 1.043530e-104 2.1994271 0.940 0.121       5 FCGR3A
## 13 2.401586e-141 3.5404225 1.000 0.122       6   GNLY
## 14 2.330636e-138 3.2910876 1.000 0.076       6   GZMB
## 15  1.410915e-27 4.2288101 0.311 0.020       7   PPBP
## 16  8.260548e-26 3.3515661 0.356 0.008       7    PF4
```

(9) Using the biomarkers identified above, select a few markers to distinguish the various subgroups. Try plotting different measurements, including raw and normalized counts on/not on the log scale.

```
VlnPlot(pbmc, c("CD74","LDHB"))
```




```
VlnPlot(pbmc, c("LST1","PF4"), use.raw = TRUE, y.log = TRUE)
```

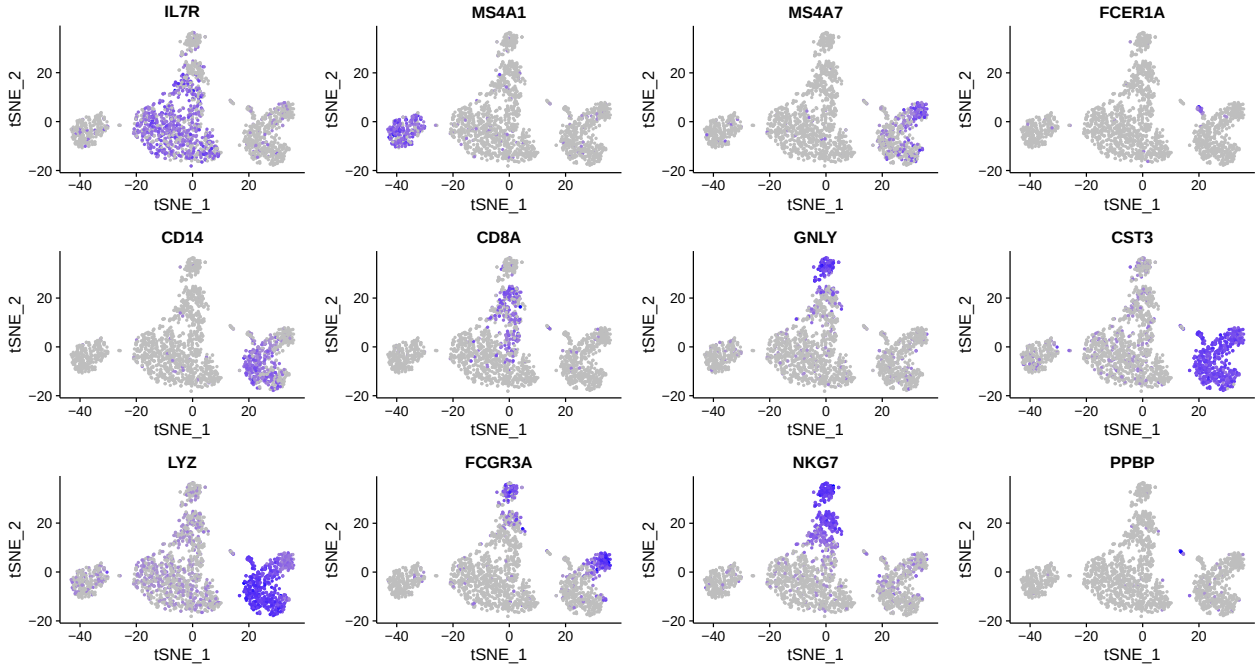


(10a) Using the table below, identify which clusters correspond to which cell subtypes in your tSNE projection. Do you observe any rare populations or mixed populations? Explore some other markers to characterize the behavior of these populations.

Cluster ID	Markers	Cell Type
?	IL7R	CD4 T cells
?	CD14, LYZ	CD14+ Monocytes
?	MS4A1	B cells
?	CD8A	CD8 T cells

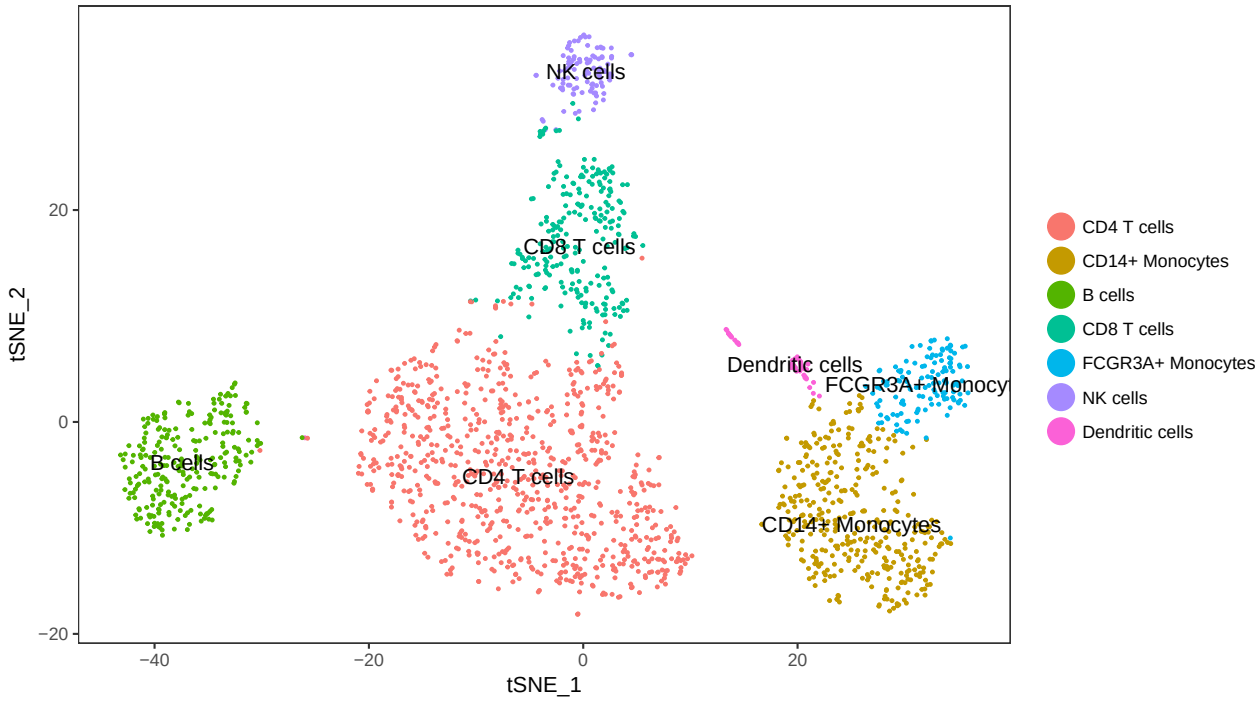
Cluster ID	Markers	Cell Type
?	FCGR3A, MS4A7	FCGR3A+ Monocytes
?	GNLY, NKG7	NK cells
?	FCER1A, CST3	Dendritic Cells
?	PPBP	Megakaryocytes

```
factors <- c("IL7R", "CD14", "LYZ", "MS4A1", "CD8A",
             "FCGR3A", "MS4A7", "GNLY", "NKG7", "FCER1A", "CST3", "PPBP")
FeaturePlot(pbmc,factors , cols.use = c("grey","blue"))
```



(10b) Using the inference above, annotate your tSNE with the cell type names.

```
current.cluster.ids <- c(0, 1, 2, 3, 4, 5, 6, 7)
new.cluster.ids <- c("CD4 T cells", "CD14+ Monocytes", "B cells",
                     "CD8 T cells", "CD4 T cells", "FCGR3A+ Monocytes", "NK cells",
                     "Dendritic cells")
pbmc@ident <- plyr::mapvalues(pbmc@ident, from = current.cluster.ids, to = new.cluster.ids)
TSNEPlot(pbmc, do.label = TRUE, pt.size = 0.5)
```


NK cells
CD8 T cells
Dendritic cells
FCGR3A+ Monocy
B cells
CD4 T cells
CD14+ Monocytes
20
0
−20
−40
−20
0
20
tSNE_2
tSNE_1
CD4 T cells
CD14+ Monocytes
B cells
CD8 T cells
FCGR3A+ Monocytes
NK cells
Dendritic cells