The Solution Lab Test

Guojing Wu

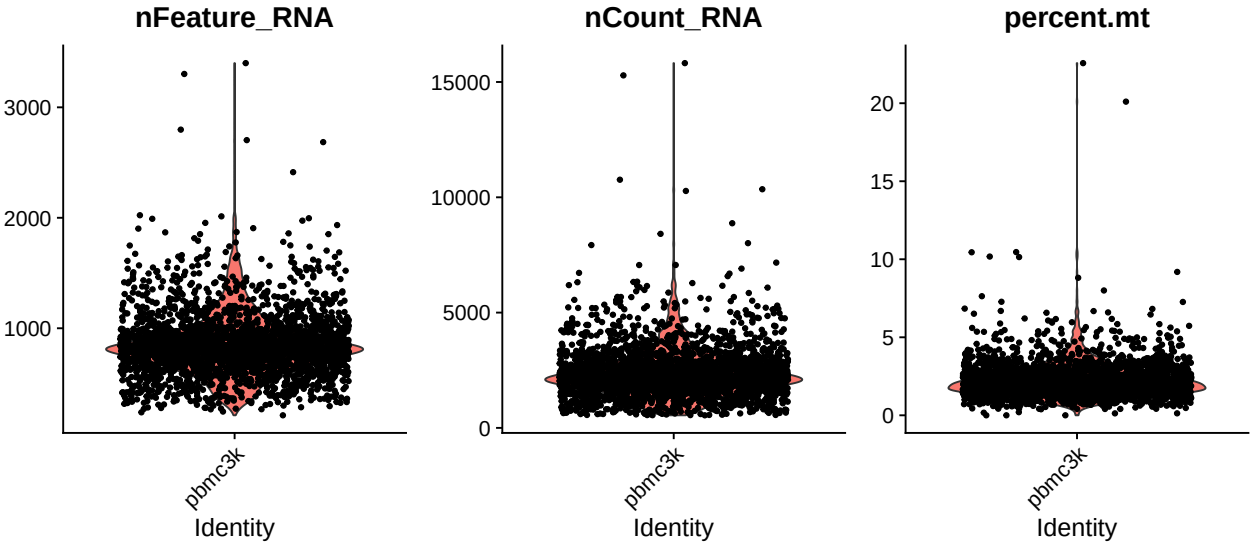
5/22/2020

Load and Preprocess

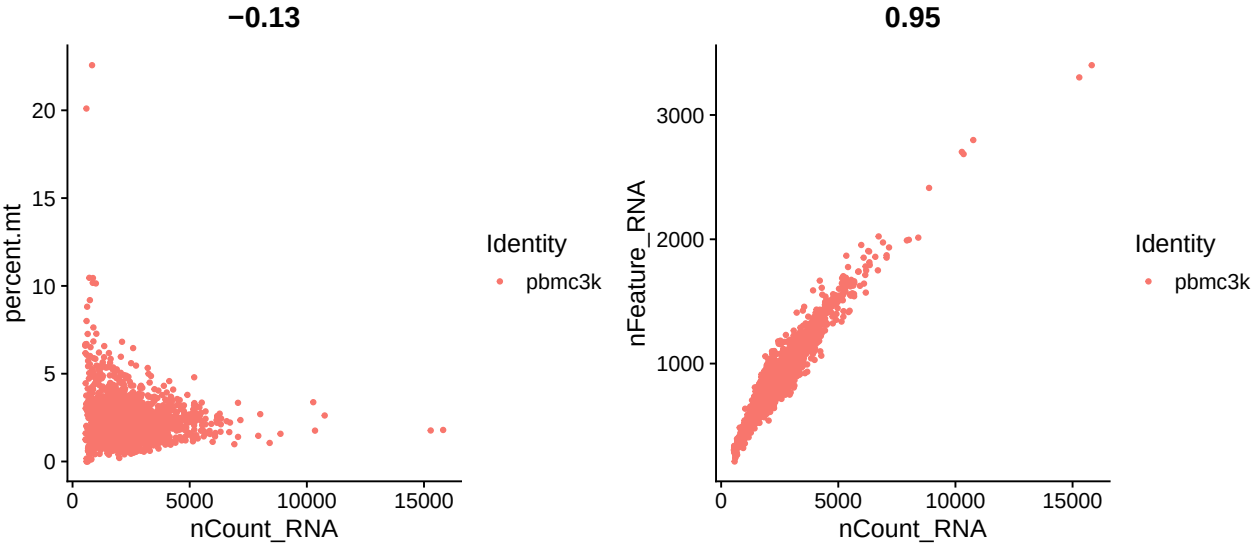
```
An object of class Seurat
13714 features across 2700 samples within 1 assay
Active assay: RNA (13714 features, 0 variable features)
```

visualize QC metrics, and use these to filter cells:

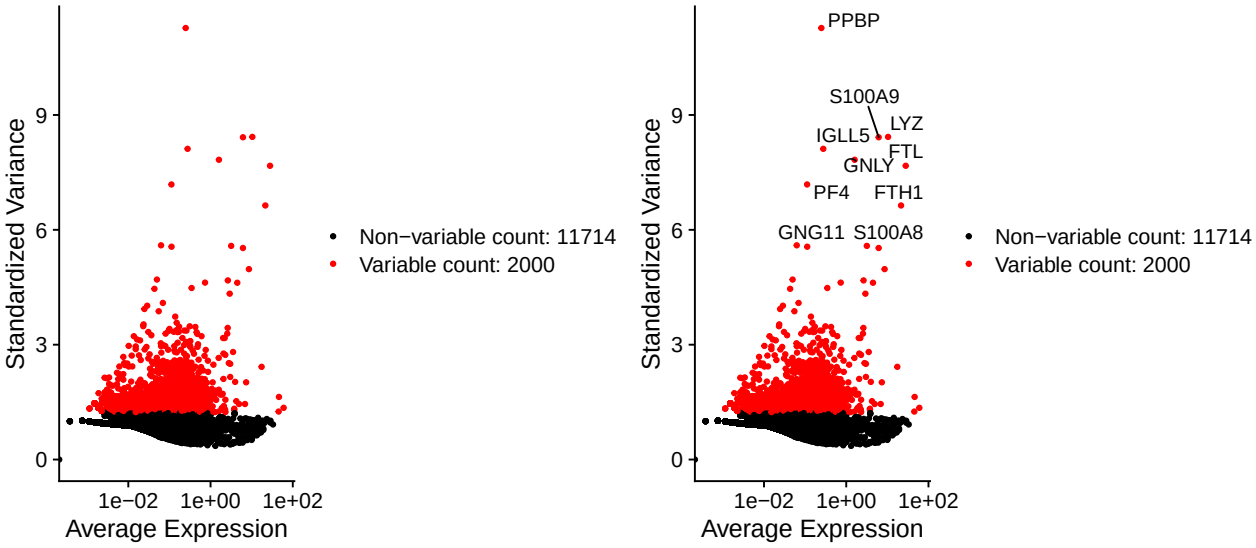
- We filter cells that have unique feature counts over 2,500 or less than 200
- We filter cells that have >5% mitochondrial counts



Feature-feature relationships:

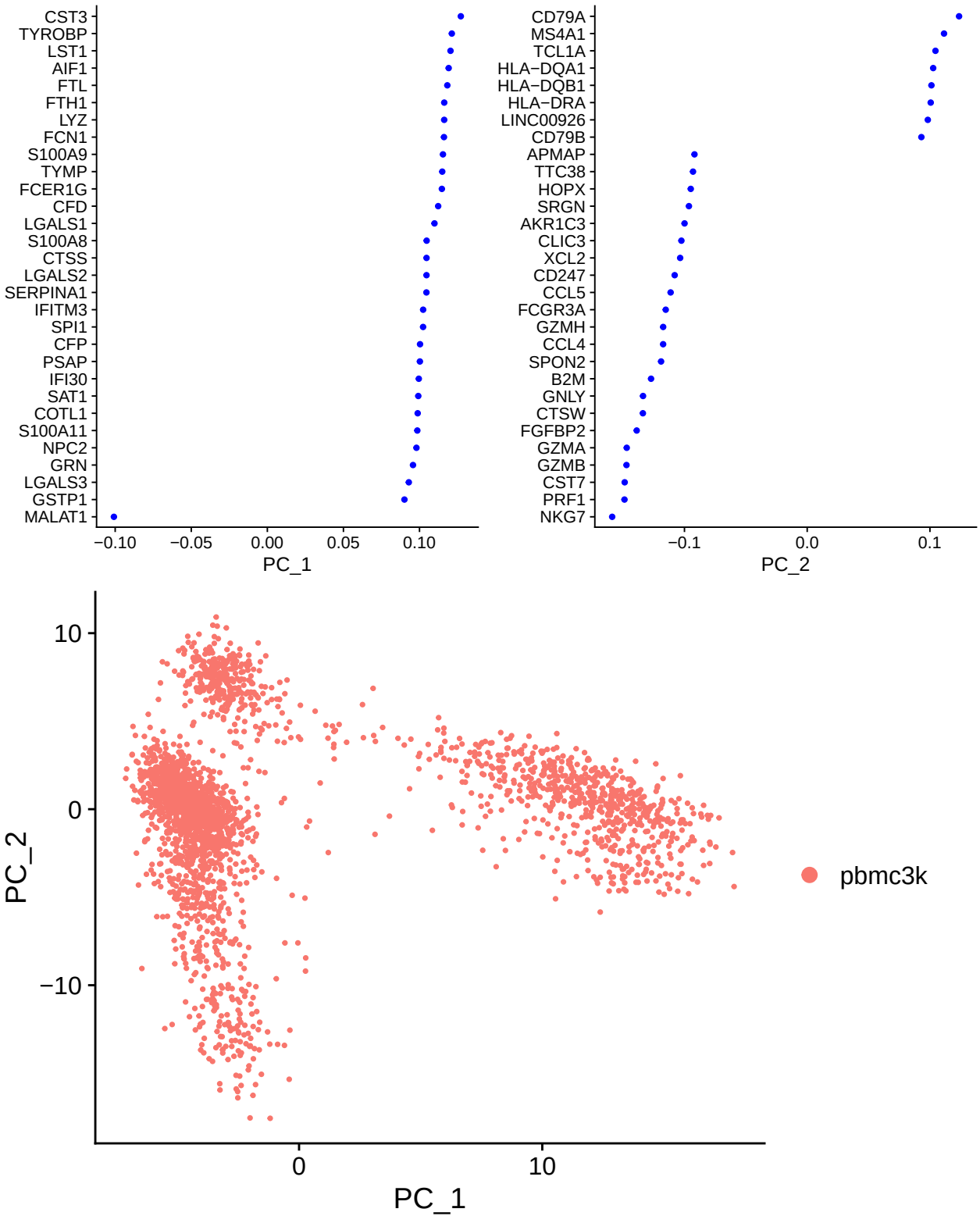


Feature selection (2000 features):

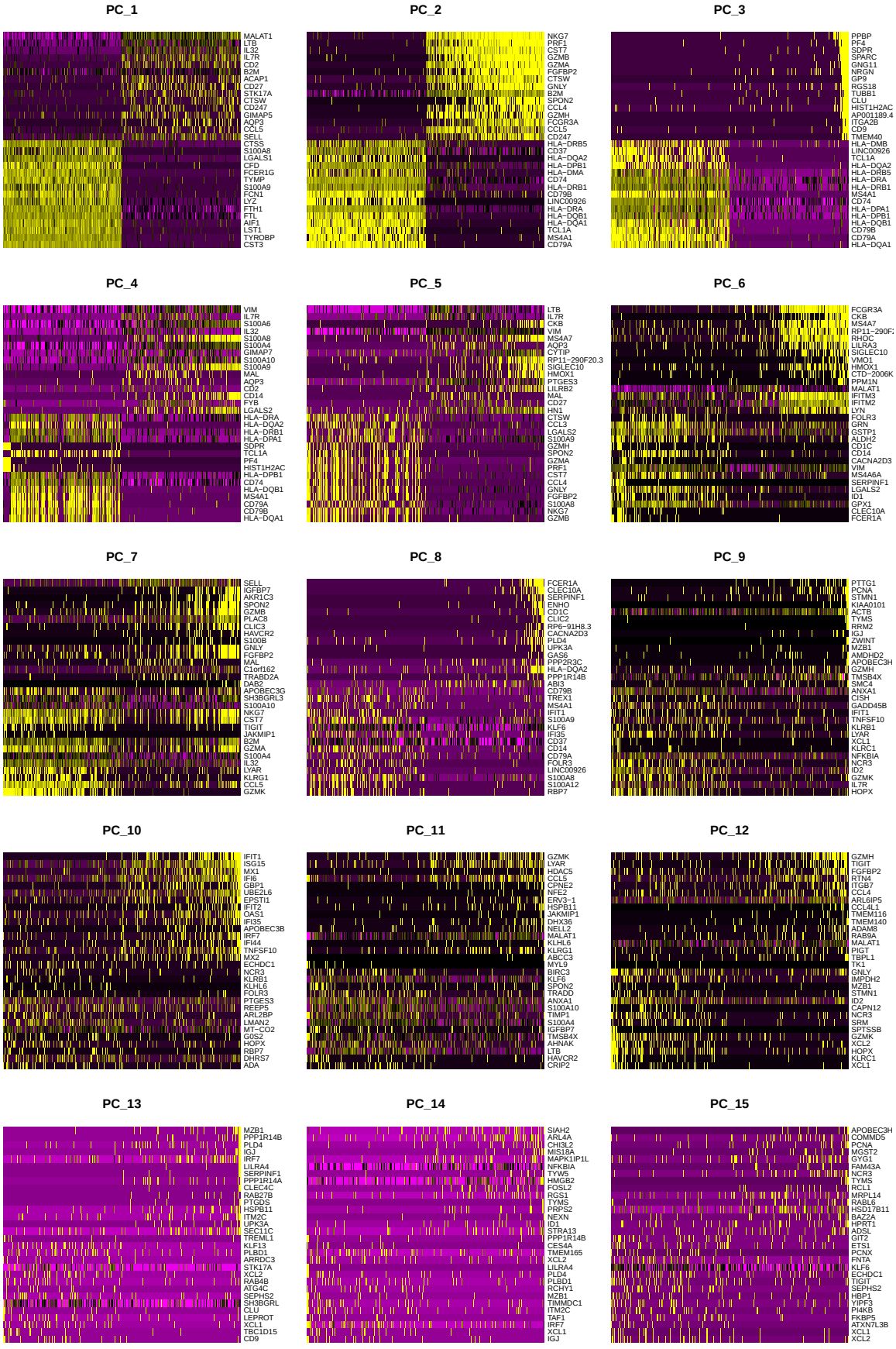


Linear Dimensional Reduction: PCA

```
PC_ 1
Positive:  CST3, TYROBP, LST1, AIF1, FTL
Negative:  MALAT1, LTB, IL32, IL7R, CD2
PC_ 2
Positive:  CD79A, MS4A1, TCL1A, HLA-DQA1, HLA-DQB1
Negative:  NKG7, PRF1, CST7, GZMB, GZMA
PC_ 3
Positive:  HLA-DQA1, CD79A, CD79B, HLA-DQB1, HLA-DPB1
Negative:  PPBP, PF4, SDPR, SPARC, GNG11
PC_ 4
Positive:  HLA-DQA1, CD79B, CD79A, MS4A1, HLA-DQB1
Negative:  VIM, IL7R, S100A6, IL32, S100A8
PC_ 5
Positive:  GZMB, NKG7, S100A8, FGFBP2, GNLY
Negative:  LTB, IL7R, CKB, VIM, MS4A7
```

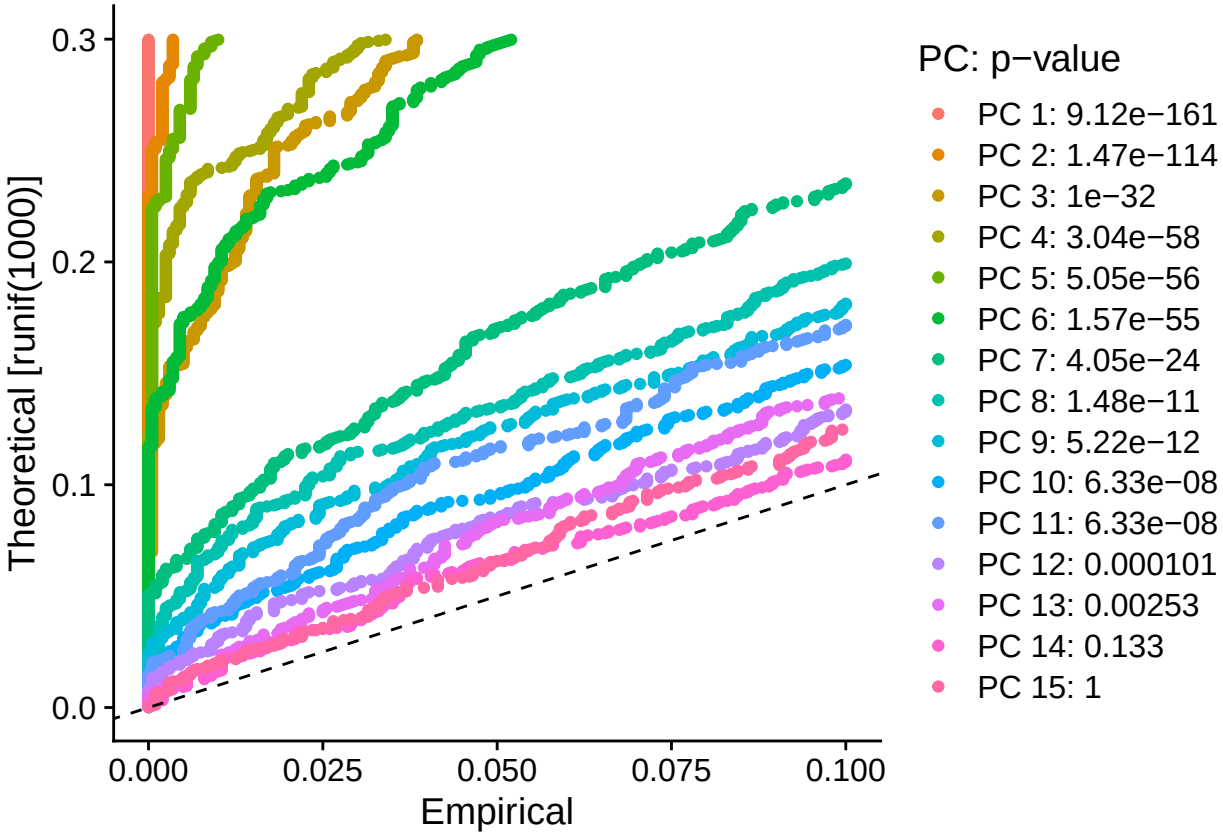


Heatmap based on PCs:

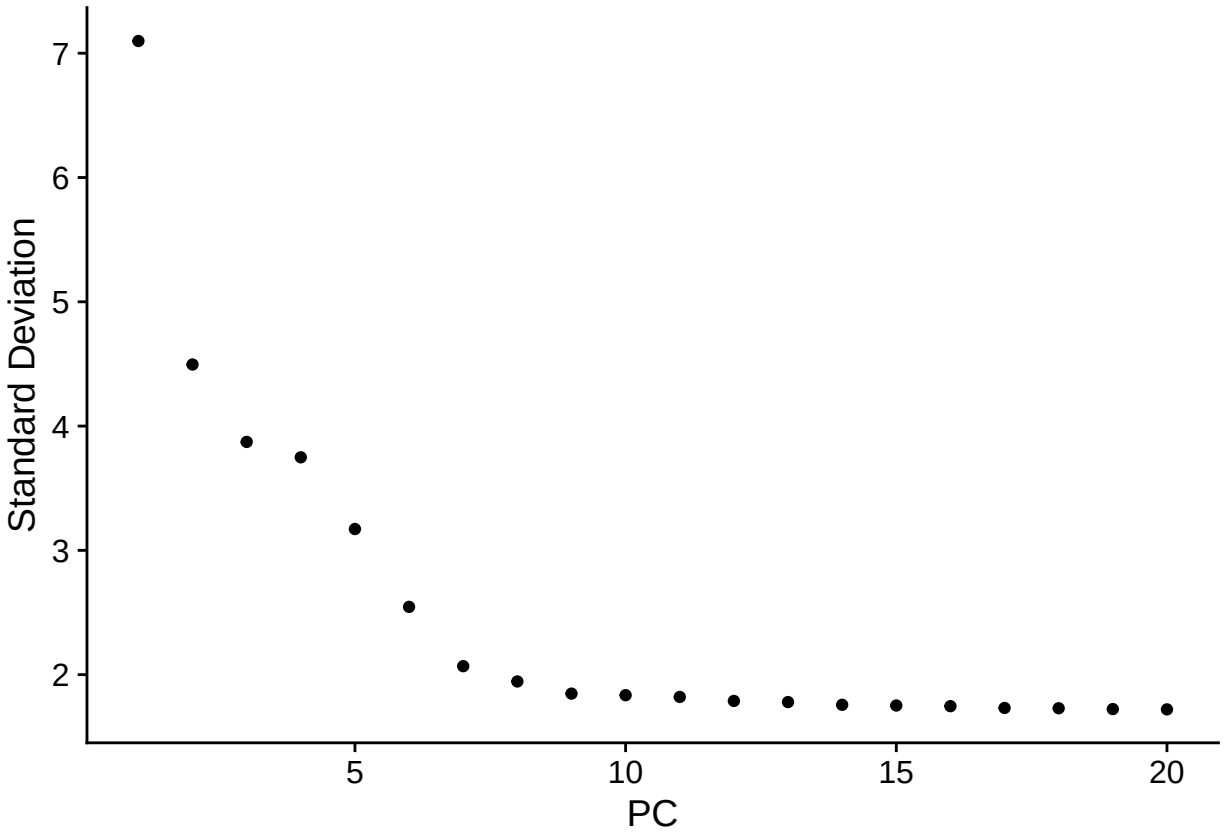

PC_1
MALAT1 LTB IL32 IL7R CD2 B2M ACAP1 CD27 STK17A CTSW CD247 GIMAP5 AQP3 CCL5 SELL CTSS S100A8 LGALS1 CFD FCER1G TYMP S100A9 FCN1 LYZ FTH1 FTL AIF1 LST1 TYROBP CST3
PC_2
NKG7 PRF1 CST7 GZMB GZMA FGFBP2 CTSW GNLY B2M SPON2 CCL4 GZMH FCGR3A CCL5 CD247 HLA−DRB5 CD37 HLA−DQA2 HLA−DPB1 HLA−DMA CD74 HLA−DRB1 CD79B LINC00926 HLA−DRA HLA−DQB1 HLA−DQA1 TCL1A MS4A1 CD79A
PC_3
PPBP PF4 SDPR SPARC GNG11 NRGN GP9 RGS18 TUBB1 CLU HIST1H2AC AP001189.4 ITGA2B CD9 TMEM40 HLA−DMB LINC00926 TCL1A HLA−DQA2 HLA−DRB5 HLA−DRA HLA−DRB1 MS4A1 CD74 HLA−DPA1 HLA−DPB1 HLA−DQB1 CD79B CD79A HLA−DQA1
PC_4
VIM IL7R S100A6 IL32 S100A8 S100A4 GIMAP7 S100A10 S100A9 MAL AQP3 CD2 CD14 FYB LGALS2 HLA−DRA HLA−DQA2 HLA−DRB1 HLA−DPA1 SDPR TCL1A PF4 HIST1H2AC HLA−DPB1 CD74 HLA−DQB1 MS4A1 CD79A CD79B HLA−DQA1
PC_5
LTB IL7R CKB VIM MS4A7 AQP3 CYTIP RP11−290F20.3 SIGLEC10 HMOX1 PTGES3 LILRB2 MAL CD27 HN1 CTSW CCL3 LGALS2 S100A9 GZMH SPON2 GZMA PRF1 CST7 CCL4 GNLY FGFBP2 S100A8 NKG7 GZMB
PC_6
FCGR3A CKB MS4A7 RP11−290F RHOC LILRA3 SIGLEC10 VMO1 HMOX1 CTD−2006K PPM1N MALAT1 IFITM3 IFITM2 LYN FOLR3 GRN GSTP1 ALDH2 CD1C CD14 CACNA2D3 VIM MS4A6A SERPINF1 LGALS2 ID1 GPX1 CLEC10A FCER1A
PC_7
SELL IGFBP7 AKR1C3 SPON2 GZMB PLAC8 CLIC3 HAVCR2 S100B GNLY FGFBP2 MAL C1orf162 TRABD2A DAB2 APOBEC3G SH3BGRL3 S100A10 NKG7 CST7 TIGIT JAKMIP1 B2M GZMA S100A4 IL32 LYAR KLRG1 CCL5 GZMK
PC_8
FCER1A CLEC10A SERPINF1 ENHO CD1C CLIC2 RP6−91H8.3 CACNA2D3 PLD4 UPK3A GAS6 PPP2R3C HLA−DQA2 PPP1R14B ABI3 CD79B TREX1 MS4A1 IFIT1 S100A9 KLF6 IFI35 CD37 CD14 CD79A FOLR3 LINC00926 S100A8 S100A12 RBP7
PC_9
PTTG1 PCNA STMN1 KIAA0101 ACTB TYMS RRM2 IGJ ZWINT MZB1 AMDHD2 APOBEC3H GZMH TMSB4X SMC4 ANXA1 CISH GADD45B IFIT1 TNFSF10 KLRB1 LYAR XCL1 KLRC1 NFKBIA NCR3 ID2 GZMK IL7R HOPX
PC_10
IFIT1 ISG15 MX1 IFI6 GBP1 UBE2L6 EPSTI1 IFIT2 OAS1 IFI35 APOBEC3B IRF7 IFI44 TNFSF10 MX2 ECHDC1 NCR3 KLRB1 KLHL6 FOLR3 PTGES3 REEP5 ARL2BP LMAN2 MT−CO2 G0S2 HOPX RBP7 DHRS7 ADA
PC_11
GZMK LYAR HDAC5 CCL5 CPNE2 NFE2 ERV3−1 HSPB11 JAKMIP1 DHX36 NELL2 MALAT1 KLHL6 KLRG1 ABCC3 MYL9 BIRC3 KLF6 SPON2 TRADD ANXA1 S100A10 TIMP1 S100A4 IGFBP7 TMSB4X AHNAK LTB HAVCR2 CRIP2
PC_12
GZMH TIGIT FGFBP2 RTN4 ITGB7 CCL4 ARL6IP5 CCL4L1 TMEM116 TMEM140 ADAM8 RAB9A MALAT1 PIGT TBPL1 TK1 GNLY IMPDH2 MZB1 STMN1 ID2 CAPN12 NCR3 SRM SPTSSB GZMK XCL2 HOPX KLRC1 XCL1
PC_13
MZB1 PPP1R14B PLD4 IGJ IRF7 LILRA4 SERPINF1 PPP1R14A CLEC4C RAB27B PTGDS HSPB11 ITM2C UPK3A SEC11C TREML1 KLF13 PLBD1 ARRDC3 STK17A XCL2 RAB4B ATG4C SEPHS2 SH3BGRL CLU LEPROT XCL1 TBC1D15 CD9
PC_14
SIAH2 ARL4A CH3L2 MIS18A MAPK1IP1L NFKBIA TYW5 HMGB2 FOSL2 RGS1 TYMS PRPS2 NEXN ID1 STRA13 PPP1R14B CES4A TMEM165 XCL2 LILRA4 PLD4 PLBD1 RCHY1 MZB1 TIMMDC1 ITM2C TAF1 IRF7 XCL1 IGJ
PC_15
APOBEC3H COMMD5 PCNA MGST2 GYG1 FAM43A NCR3 TYMS RCL1 MRPL14 RABL6 HSD17B11 BAZ2A HPRT1 ADSL GIT2 ETS1 PCNX FNTA KLF6 ECHDC1 TIGIT SEPHS2 HBP1 YIPF3 PI4KB FKBP5 ATXN7L3B XCL1 XCL2

Determine the dimensionality:

- JackStraw plot:



- Elbow plot:



Based on the plot, we roughly choose the first 10 PCs

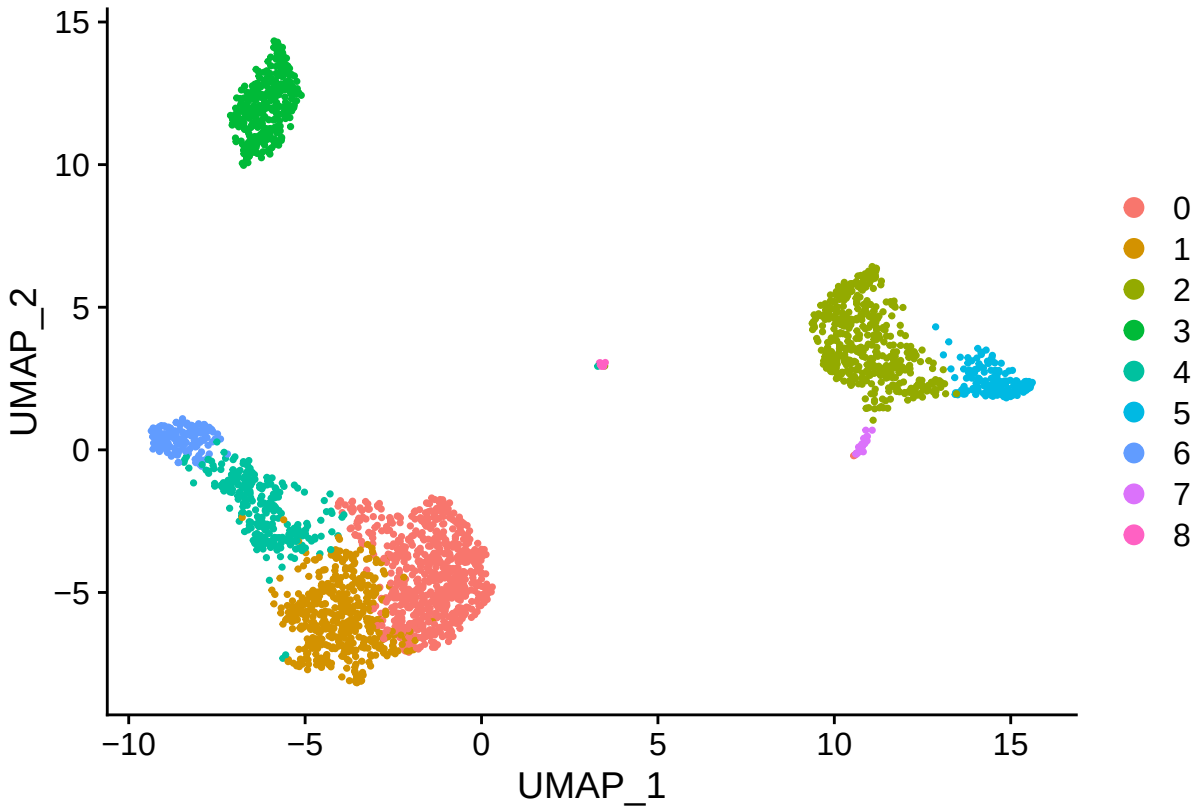
Clustering

Visualizing use UMAP:

```
Modularity Optimizer version 1.3.0 by Ludo Waltman and Nees Jan van Eck

Number of nodes: 2638
Number of edges: 96033

Running Louvain algorithm...
Maximum modularity in 10 random starts: 0.8720
Number of communities: 9
Elapsed time: 0 seconds
```

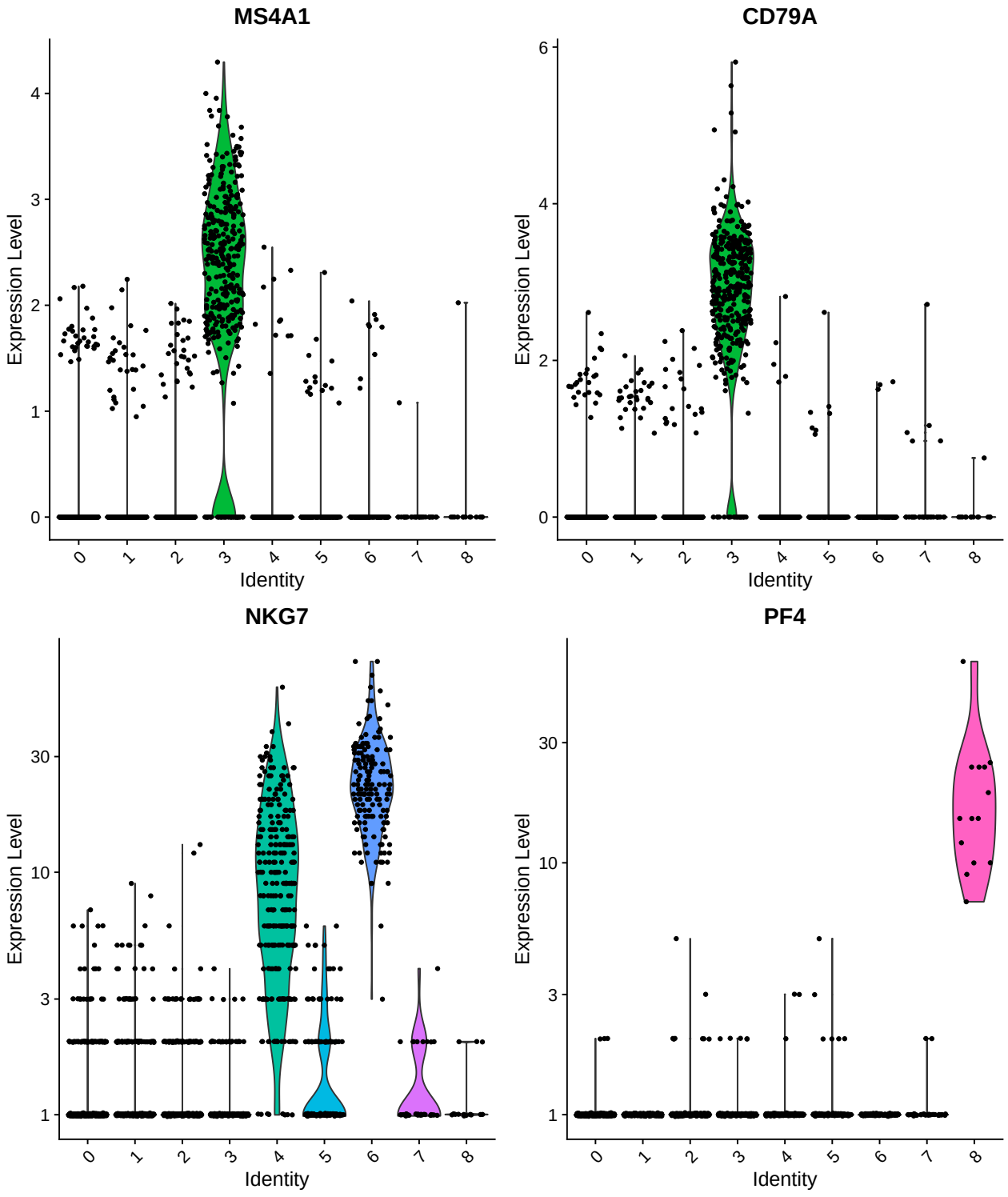


Finding differentially expressed features

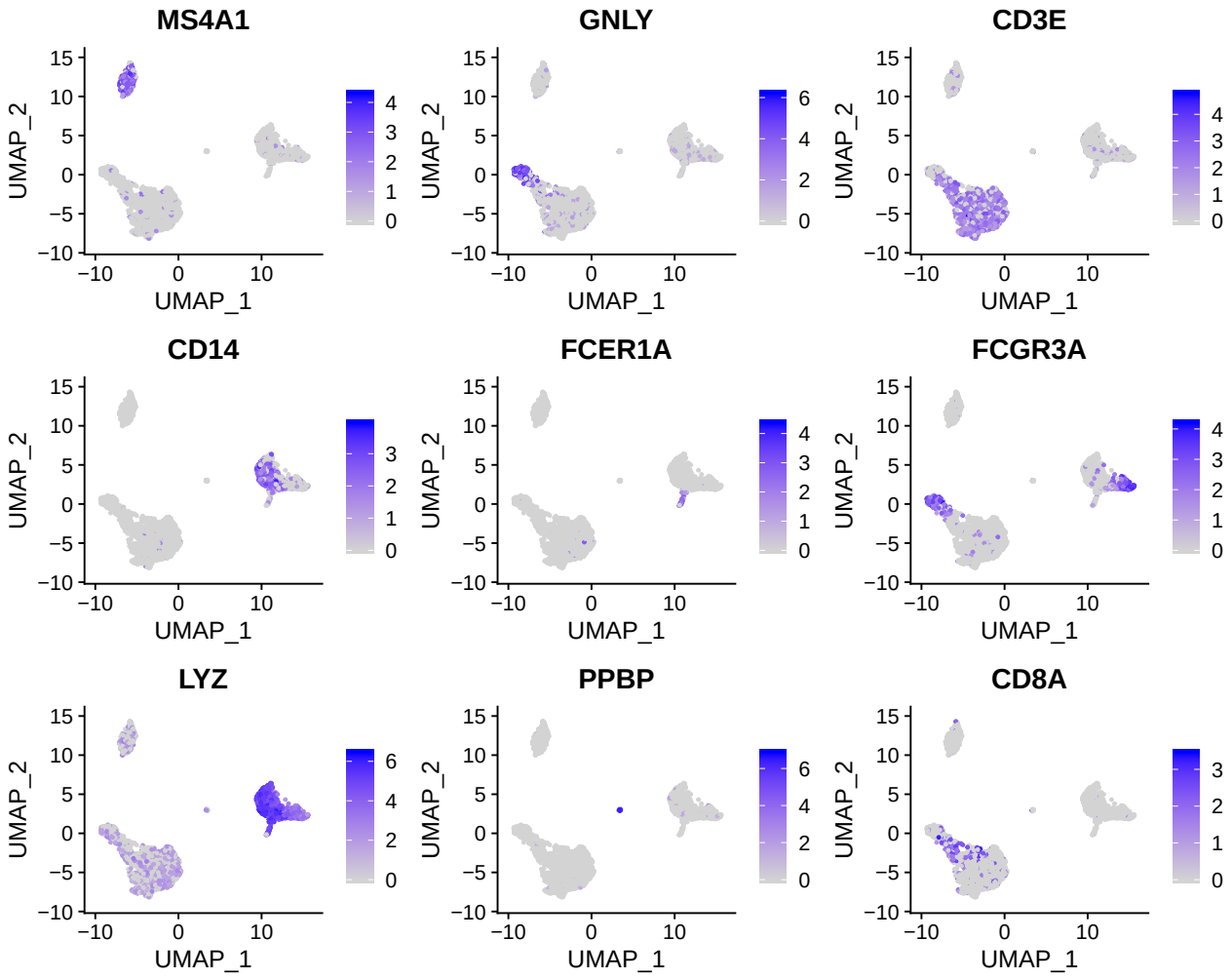
Markers for every cluster compared to all remaining cells:

p_val	avg_logFC	pct.1	pct.2	p_val_adj	cluster	gene
0	0.7300635	0.901	0.594	0	0	LDHB
0	0.9219135	0.436	0.110	0	0	CCR7
0	0.8921170	0.981	0.642	0	1	LTB
0	0.8586034	0.422	0.110	0	1	AQP3
0	3.8608733	0.996	0.215	0	2	S100A9
0	3.7966403	0.975	0.121	0	2	S100A8
0	2.9875833	0.936	0.041	0	3	CD79A
0	2.4894932	0.622	0.022	0	3	TCL1A
0	2.1220555	0.985	0.240	0	4	CCL5
0	2.0461687	0.587	0.059	0	4	GZMK
0	2.2954931	0.975	0.134	0	5	FCGR3A
0	2.1388125	1.000	0.315	0	5	LST1
0	3.3462278	0.961	0.068	0	6	GZMB
0	3.6898996	0.961	0.131	0	6	GNLY
0	2.6832771	0.812	0.011	0	7	FCER1A
0	1.9924275	1.000	0.513	0	7	HLA-DPB1
0	5.0207262	1.000	0.010	0	8	PF4
0	5.9443347	1.000	0.024	0	8	PPBP

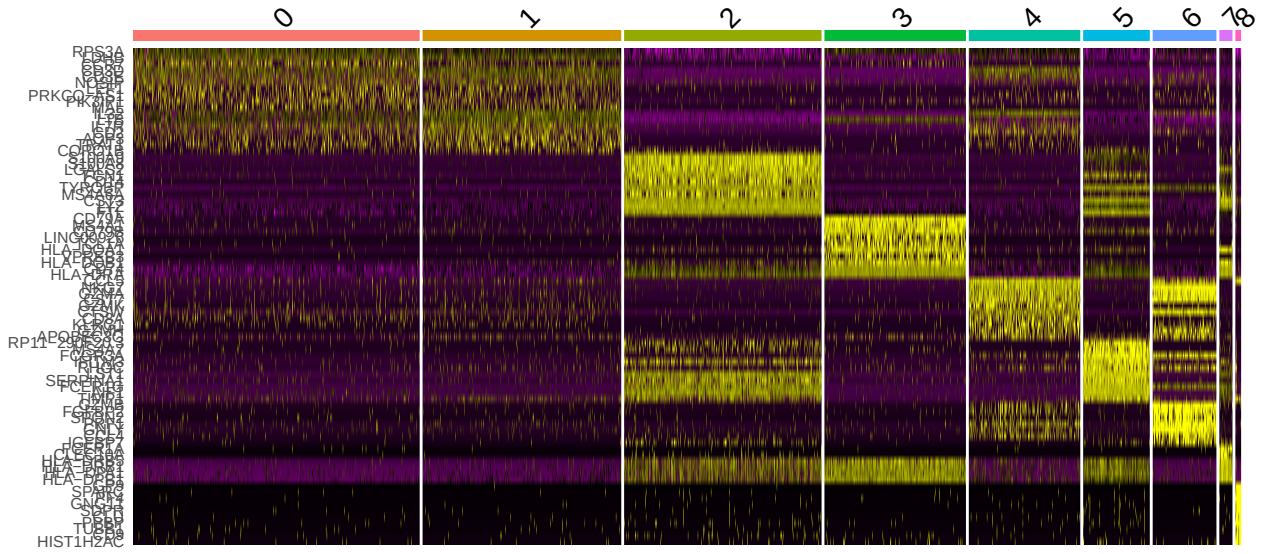
Visualize using VlnPlot:



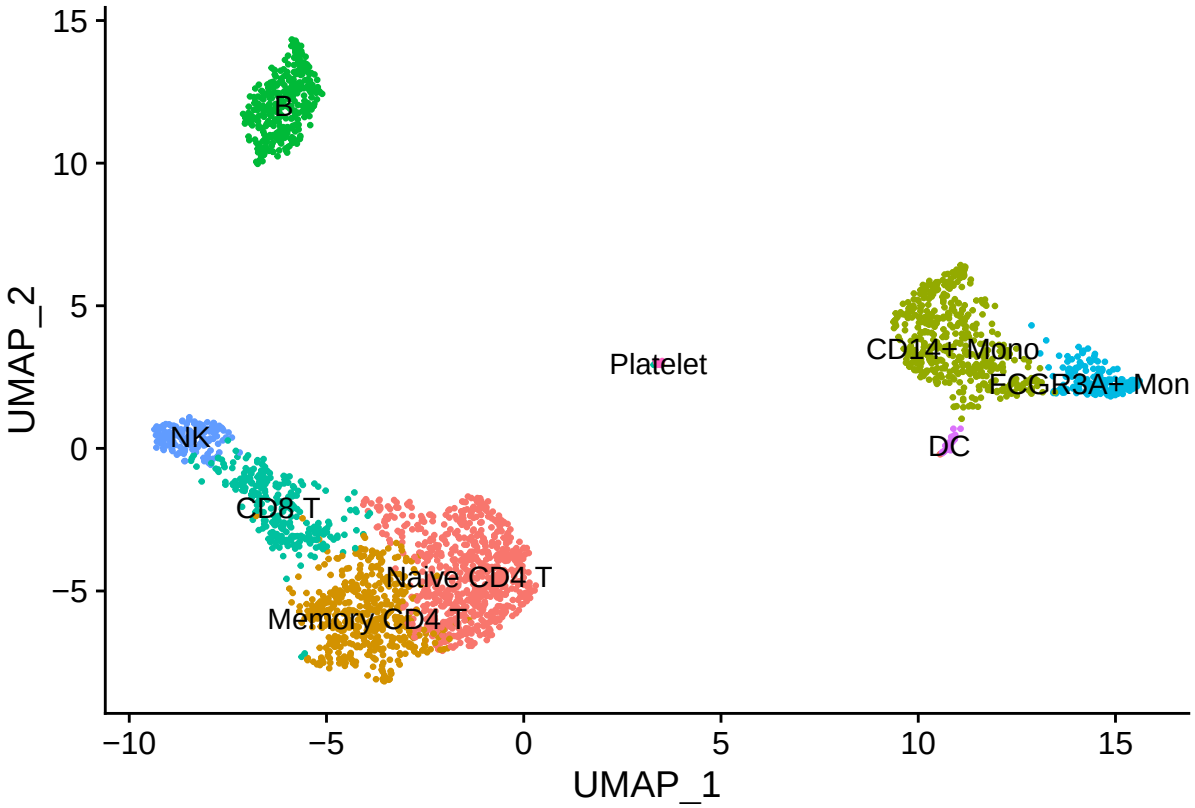
Visualize using FeaturePlot:



Plotting the top 20 markers (or all markers if less than 20) for each cluster use DoHeatmap:



Assigning cell type identity to clusters



Appendix A

```
knitr::opts_chunk$set(echo = FALSE, message = FALSE, warning = FALSE, comment = "")
library(dplyr)
library(Seurat)
library(patchwork)

# Load the PBMC dataset
pbmc.data <- Read10X(data.dir = "filtered_gene_bc_matrices/hg19/")

# Initialize the Seurat object with the raw (non-normalized data).
pbmc <- CreateSeuratObject(counts = pbmc.data, project = "pbmc3k", min.cells = 3, min.features = 200)
pbmc

# The [[ operator can add columns to object metadata. This is a great place to stash QC stats
pbmc[["percent.mt"]] <- PercentageFeatureSet(pbmc, pattern = "^MT-")

# Visualize QC metrics as a violin plot
VlnPlot(pbmc, features = c("nFeature_RNA", "nCount_RNA", "percent.mt"), ncol = 3)

# FeatureScatter is typically used to visualize feature-feature relationships, but can be used
# for anything calculated by the object, i.e. columns in object metadata, PC scores etc.
plot1 <- FeatureScatter(pbmc, feature1 = "nCount_RNA", feature2 = "percent.mt")
plot2 <- FeatureScatter(pbmc, feature1 = "nCount_RNA", feature2 = "nFeature_RNA")
```

```
plot1 + plot2

# filtering and normalizing
pbmc <- subset(pbmc, subset = nFeature_RNA > 200 & nFeature_RNA < 2500 & percent.mt < 5)
pbmc <- NormalizeData(pbmc, normalization.method = "LogNormalize", scale.factor = 10000)

# feature selection
pbmc <- FindVariableFeatures(pbmc, selection.method = "vst", nfeatures = 2000)
# Identify the 10 most highly variable genes
top10 <- head(VariableFeatures(pbmc), 10)
# plot variable features with and without labels
plot1 <- VariableFeaturePlot(pbmc)
plot2 <- LabelPoints(plot = plot1, points = top10, repel = TRUE)
plot1 + plot2

# scaling
all.genes <- rownames(pbmc)
pbmc <- ScaleData(pbmc, features = all.genes)

pbmc <- RunPCA(pbmc, features = VariableFeatures(object = pbmc))
# Examine and visualize PCA results a few different ways
print(pbmc[["pca"]], dims = 1:5, nfeatures = 5)

VizDimLoadings(pbmc, dims = 1:2, reduction = "pca")

DimPlot(pbmc, reduction = "pca")

# heatmap
DimHeatmap(pbmc, dims = 1:15, cells = 500, balanced = TRUE)

# JackStraw
pbmc <- JackStraw(pbmc, num.replicate = 100)
pbmc <- ScoreJackStraw(pbmc, dims = 1:20)
JackStrawPlot(pbmc, dims = 1:15)

ElbowPlot(pbmc)
# clustering
pbmc <- FindNeighbors(pbmc, dims = 1:10)
pbmc <- FindClusters(pbmc, resolution = 0.5)

# visualize result use UMAP
pbmc <- RunUMAP(pbmc, dims = 1:10)
DimPlot(pbmc, reduction = "umap")
# save files
saveRDS(pbmc, file = "pbmc_tutorial.rds")

# find markers for every cluster compared to all remaining cells, report only the positive ones
pbmc.markers = FindAllMarkers(pbmc, only.pos = TRUE, min.pct = 0.25, logfc.threshold = 0.25)
pbmc.markers %>%
  group_by(cluster) %>%
  top_n(n = 2, wt = avg_logFC) %>%
  knitr::kable()
```

```r
# Visualize using VlnPlot
VlnPlot(pbmc, features = c("MS4A1", "CD79A"))

# plot raw counts as well
VlnPlot(pbmc, features = c("NKG7", "PF4"), slot = "counts", log = TRUE)

FeaturePlot(pbmc, features = c("MS4A1", "GNLY", "CD3E", "CD14", "FCER1A", "FCGR3A", "LYZ", "PPBP",
    "CD8A"))
# plotting the top 20 markers (or all markers if less than 20) for each cluster
top10 = pbmc.markers %>% group_by(cluster) %>% top_n(n = 10, wt = avg_logFC)
DoHeatmap(pbmc, features = top10$gene) + NoLegend()

# assigning cell type identity to clusters
new.cluster.ids = c("Naive CD4 T", "Memory CD4 T", "CD14+ Mono", "B", "CD8 T", "FCGR3A+ Mono",
    "NK", "DC", "Platelet")
names(new.cluster.ids) = levels(pbmc)
pbmc = RenameIdents(pbmc, new.cluster.ids)
DimPlot(pbmc, reduction = "umap", label = TRUE, pt.size = 0.5) + NoLegend()

saveRDS(pbmc, file = "pbmc3k_final.rds")
```