Re-analysis of Marin-Carli et al sMEC Data

Dave Bridges

January 4, 2021

Contents

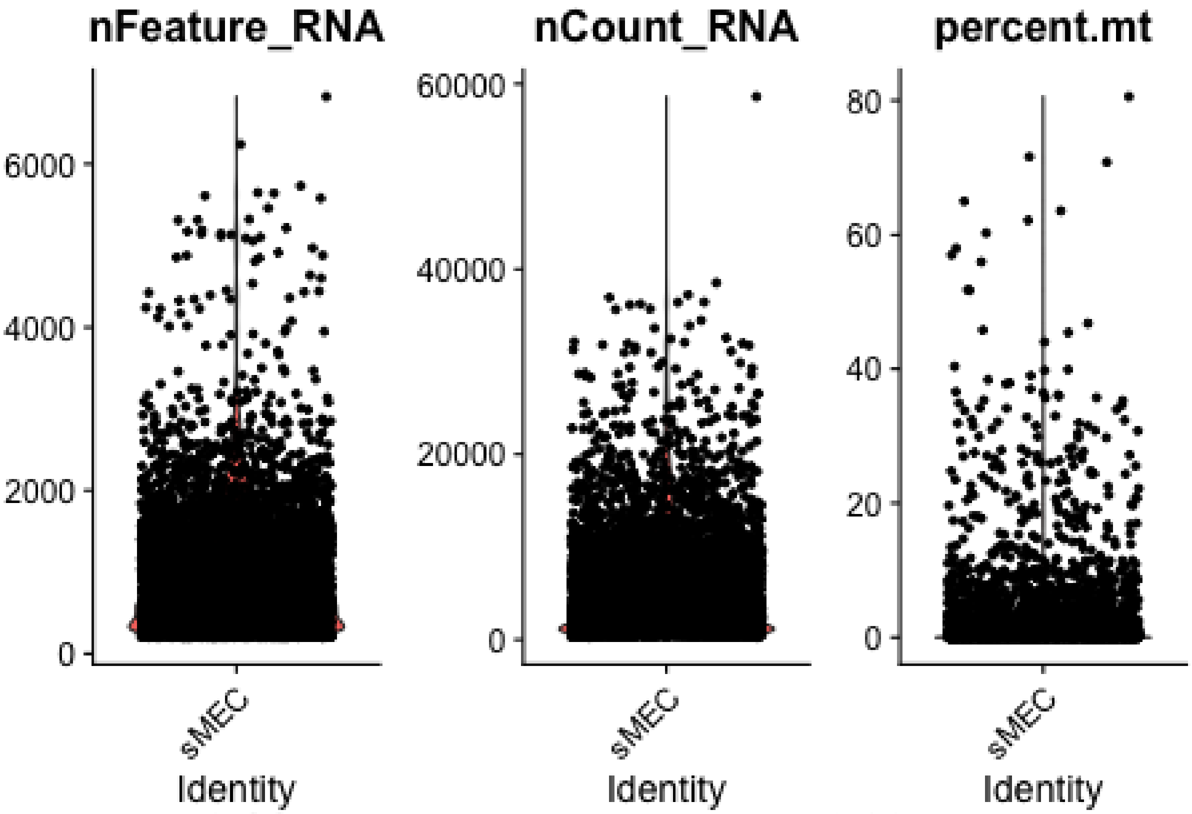
1 Purpose

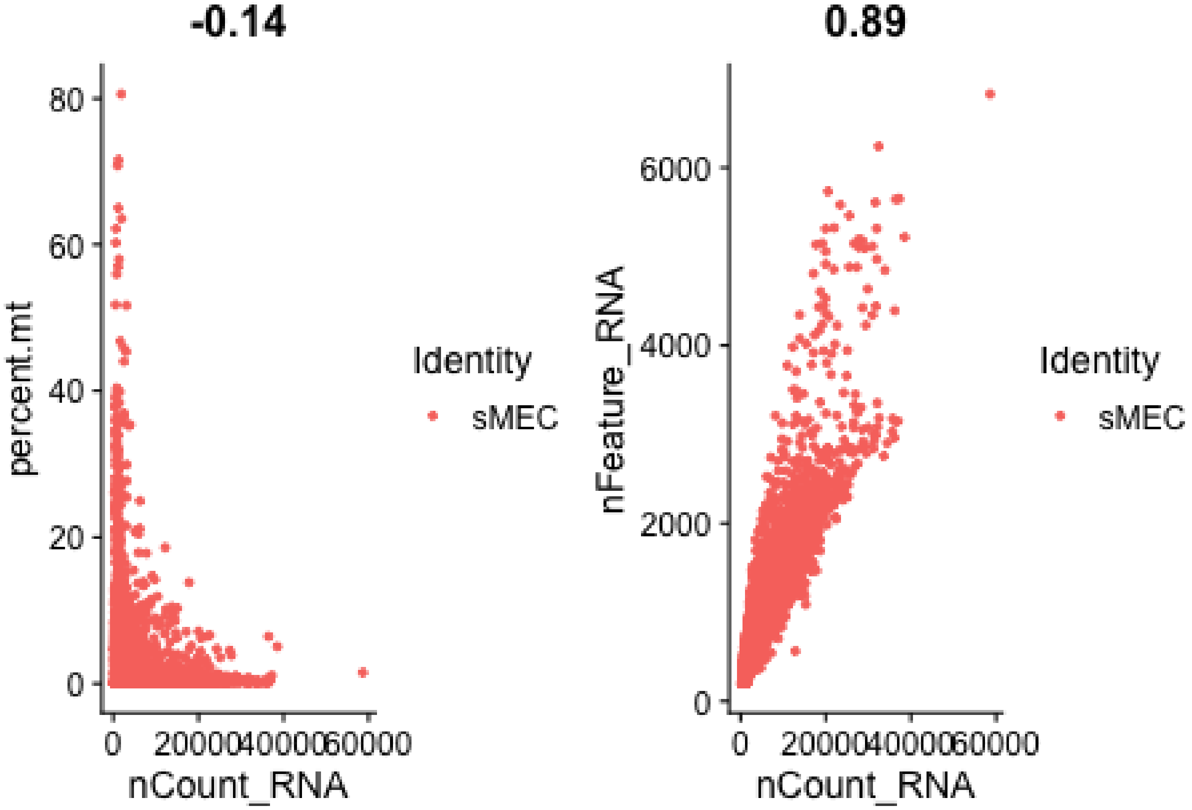
To re-analyse cell populations from the Martin-Carli *et al*'s scRNAseq study of lactating mammary glands. This work is described in Martin Carli et al. (2020). This follows the analysis flow suggested for Seurat 3.2 seen at https://satijalab.org/seurat/v3.2/pbmc3k_tutorial.html

2 Data Input

Downloaded the data from GSE15889 and removed prefixes from filenames.

2.1 Preprocessing



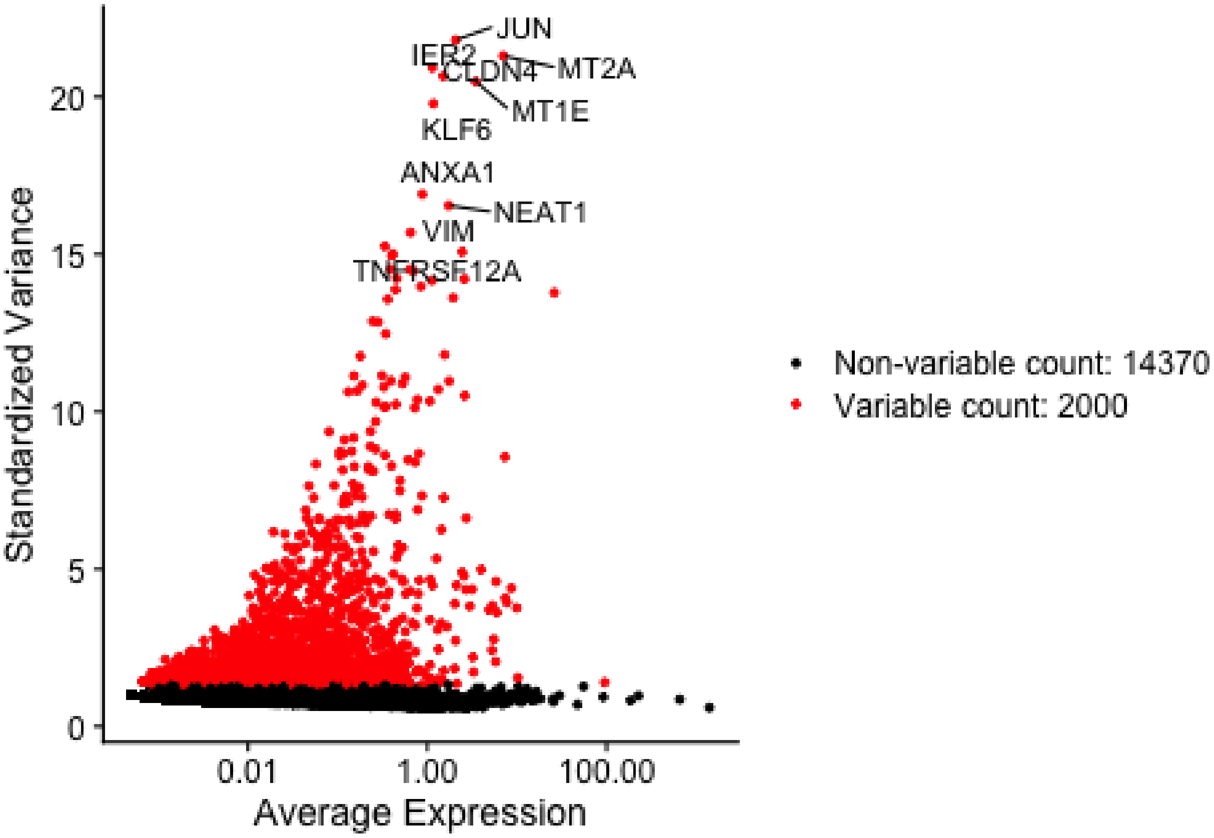


Normalizing Data

Normalizes the feature expression for each cell by the total expression x 1000, then log transforms

2.2 Highly Variable Features

Features with high cell to cell variability (highly expressed in some cells but not others)

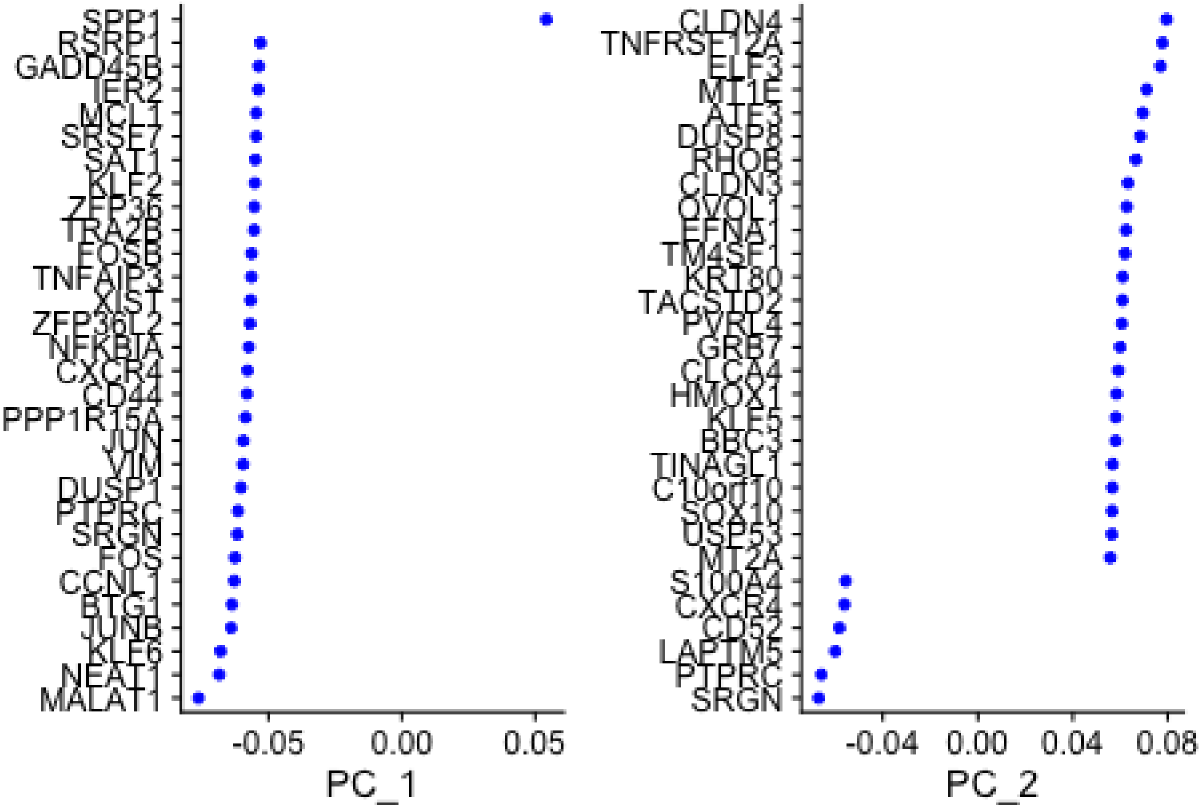


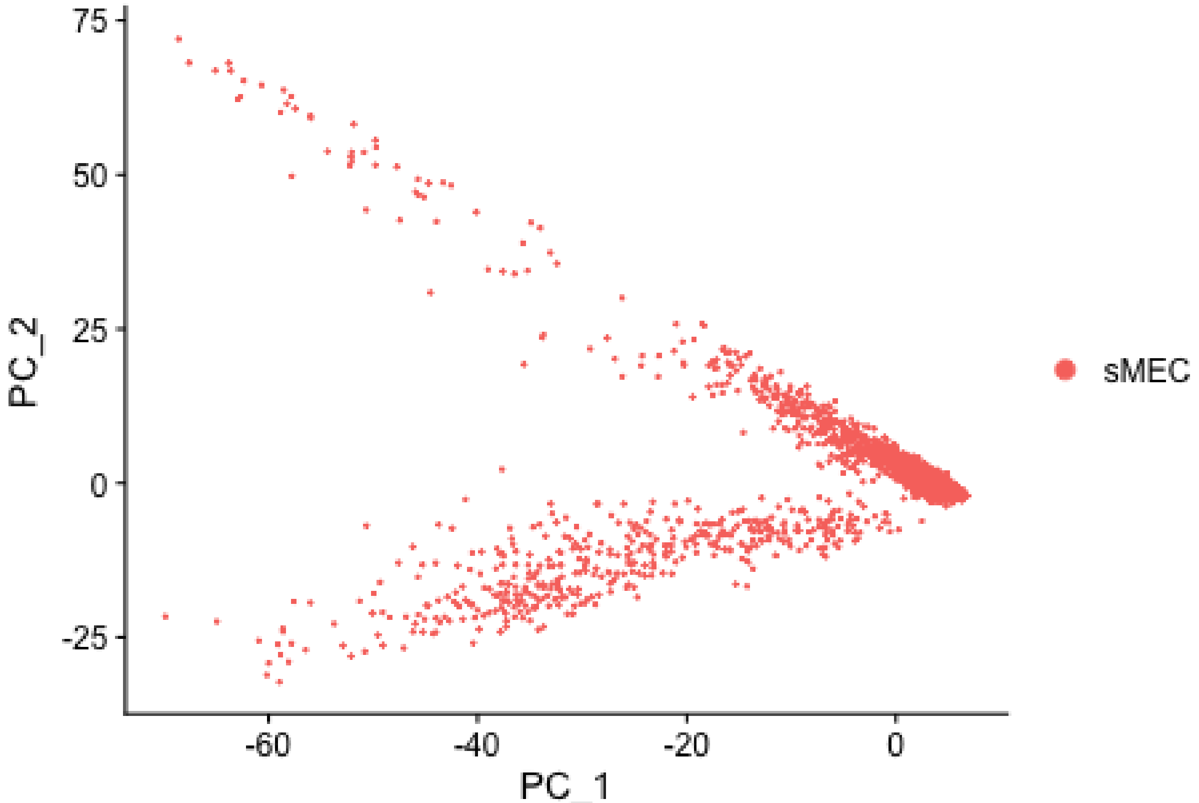
2.3 Scaling

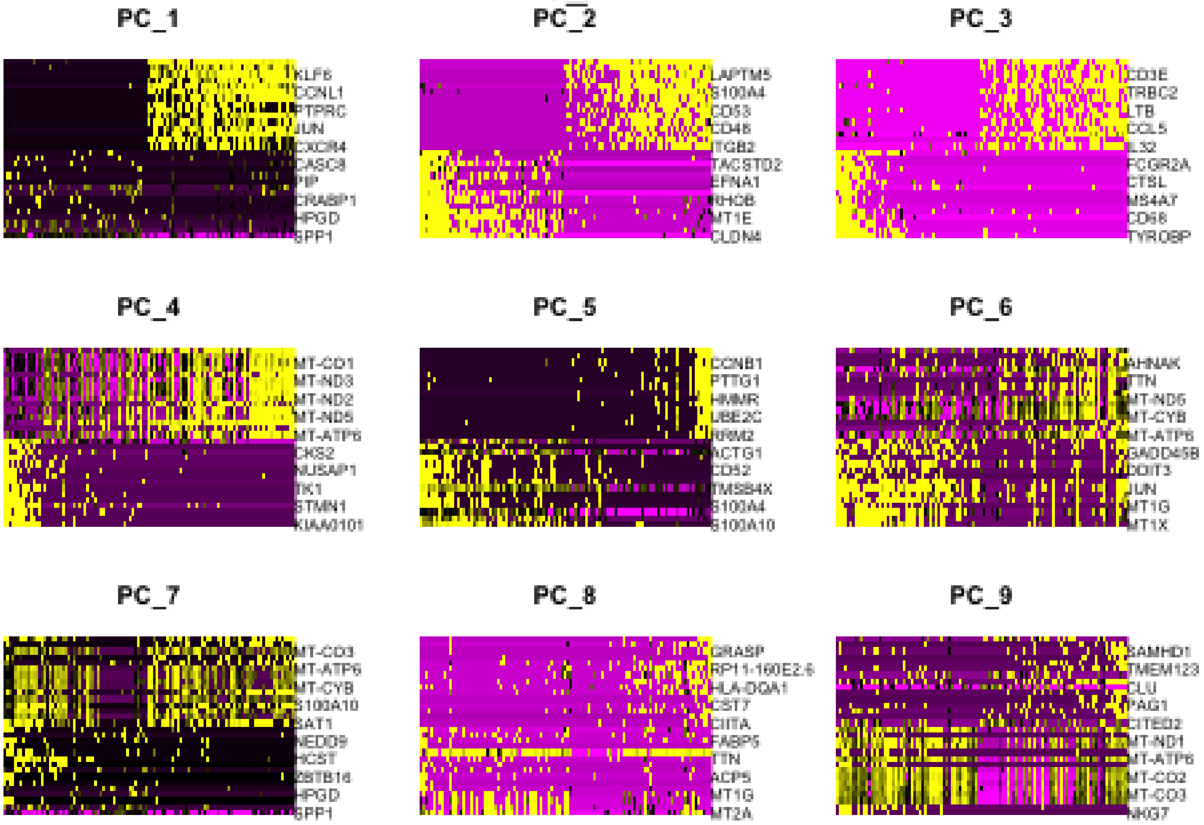
Shifts expression so that mean expression across cells is 0, and variance is 1. This reduces the impact of outliers on downstream analyses. We did not regress out specific sources of heterogeneity like mitochondrial contamination or cell cycle stage.

3 Dimensionality Reductions

```
## PC_ 1
## Positive:  SPP1, CLU, SCGB3A1, HPGD, MYBPC1
## Negative:  MALAT1, NEAT1, KLF6, JUNB, BTG1
## PC_ 2
## Positive:  CLDN4, TNFRSF12A, ELF3, MT1E, ATF3
## Negative:  SRGN, PTPRC, LAPTM5, CD52, CXCR4
## PC_ 3
## Positive:  TYROBP, FCER1G, SPI1, CD68, FABP5
## Negative:  ETS1, TRAC, CD3E, CD2, FYN
## PC_ 4
## Positive:  KIAA0101, TYMS, PTTG1, STMN1, CCNB1
## Negative:  NEAT1, MT-CO3, MT-CO1, MT-ND4, XIST
## PC_ 5
## Positive:  S100A10, S100A6, FTL, S100A4, HCST
## Negative:  KIAA0101, BIRC5, CCNB1, TYMS, TOP2A
```

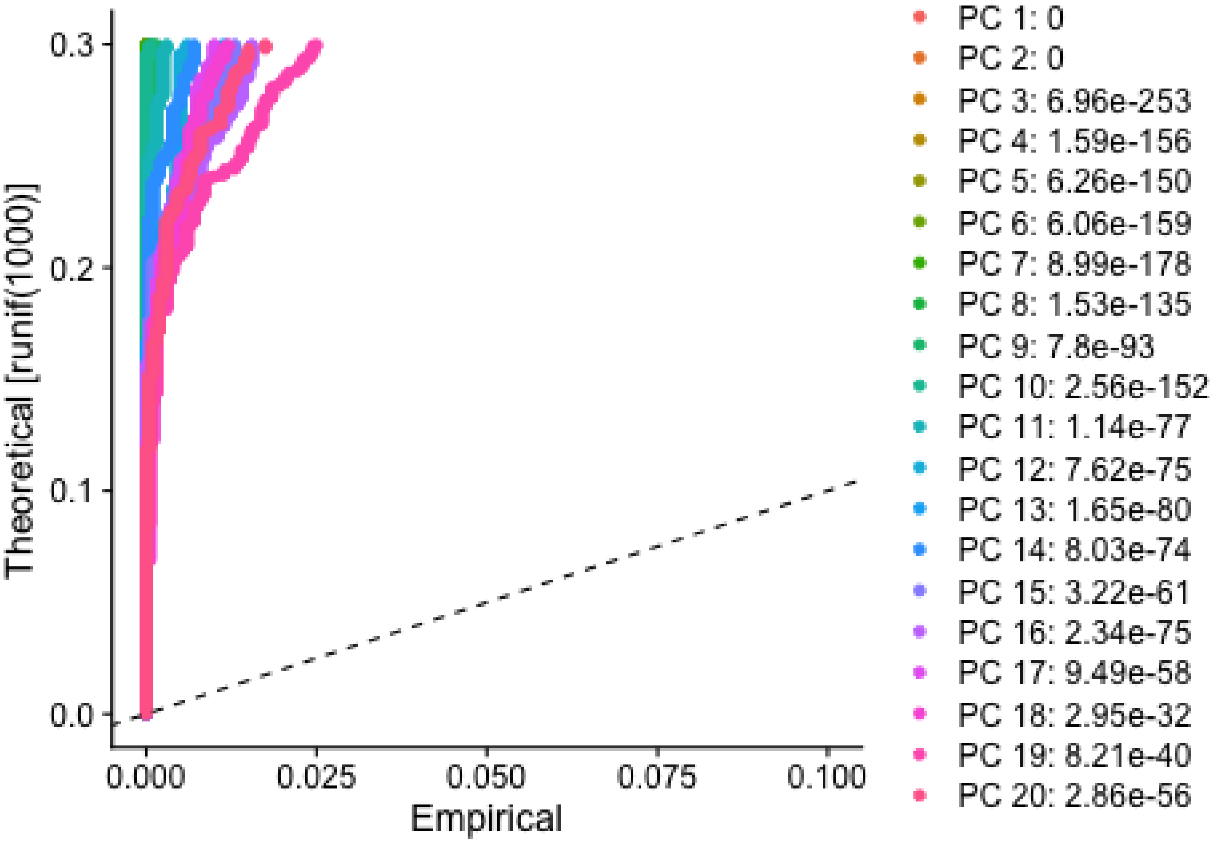


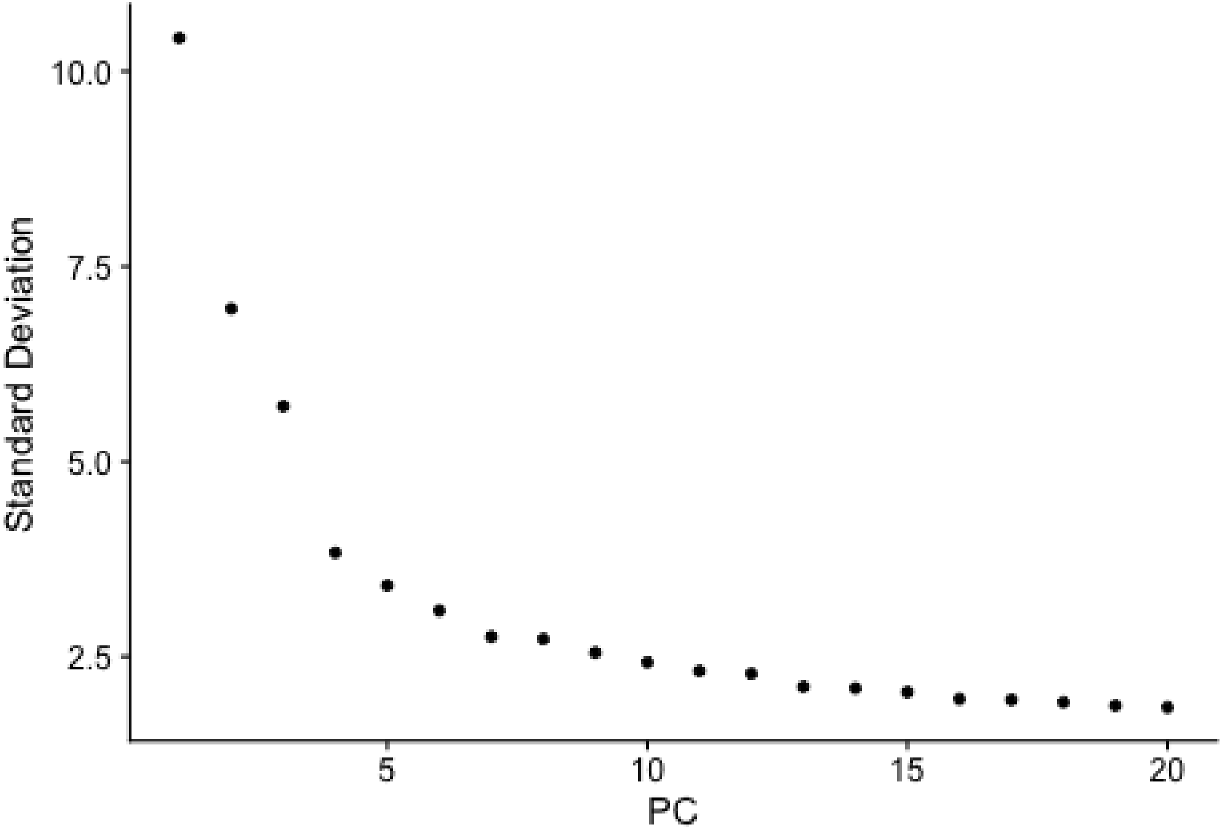

75
50
25
0
-25
PC_2
-60
-40
-20
0
PC_1
sMEC


PC_1
KLF6
CCNL1
PTPRC
JUN
CXCR4
CASC3
PIP
CRABP1
HPGD
SPP1
PC_2
LAPTM5
S100A4
CD53
CD48
ITGB2
TACSTD2
EFNA1
RHOB
MT1E
CLDN4
PC_3
CD3E
TRBC2
LTB
CCL5
IL32
FCGR2A
CTSL
MS4A7
CD68
TYROBP
PC_4
MT-CO1
MT-ND3
MT-ND2
MT-ND5
MT-ATP6
CKS2
NUSAP1
TK1
STMN1
KIAA0101
PC_5
CCNB1
PTTG1
HMMR
UBE2C
RRM2
ACTG1
CD52
TMSB4X
S100A4
S100A10
PC_6
AHNAK
TTN
MT-ND5
MT-CYB
MT-ATP6
GADD45B
DDIT3
JUN
MT1G
MT1X
PC_7
MT-CO3
MT-ATP6
MT-CYB
S100A10
SAT1
NEDD9
HCST
ZBTB16
HPGD
SPP1
PC_8
GRASP
RP11-160E2.6
HLA-DQA1
CST7
CIITA
FABP5
TTN
ACP5
MT1G
MT2A
PC_9
SAMHD1
TMEM123
CLU
PAG1
CITED2
MT-ND1
MT-ATP6
MT-CO2
MT-CO3
NKG7

4 Determination of Number of Clusters

Used the JackStraw procedure in Macosko et al. (2015), sampling 1% of the data re-running the PCA and constructing a null distribution of feature scores, then repeating. This identified ‘significant’ PCs. We also did an elbow plot.





Based on this we decided to use 13 PCs to cluster the cells.

5 Clustering Cell Types

Seurat 3.2 uses a K-nearest neighbor approach then tries to partition this into commmunities of cell types.

5.1 Identification and Assignment of Clusters

```
## Modularity Optimizer version 1.3.0 by Ludo Waltman and Nees Jan van Eck
##
## Number of nodes: 5917
## Number of edges: 202546
##
## Running Louvain algorithm...
## Maximum modularity in 10 random starts: 0.8391
## Number of communities: 9
## Elapsed time: 0 seconds
```

5.2 Non-Linear Dimensionality Reduction

Did both UMAP and t-SNE plots using 13 clusters

5.2.1 UMAP

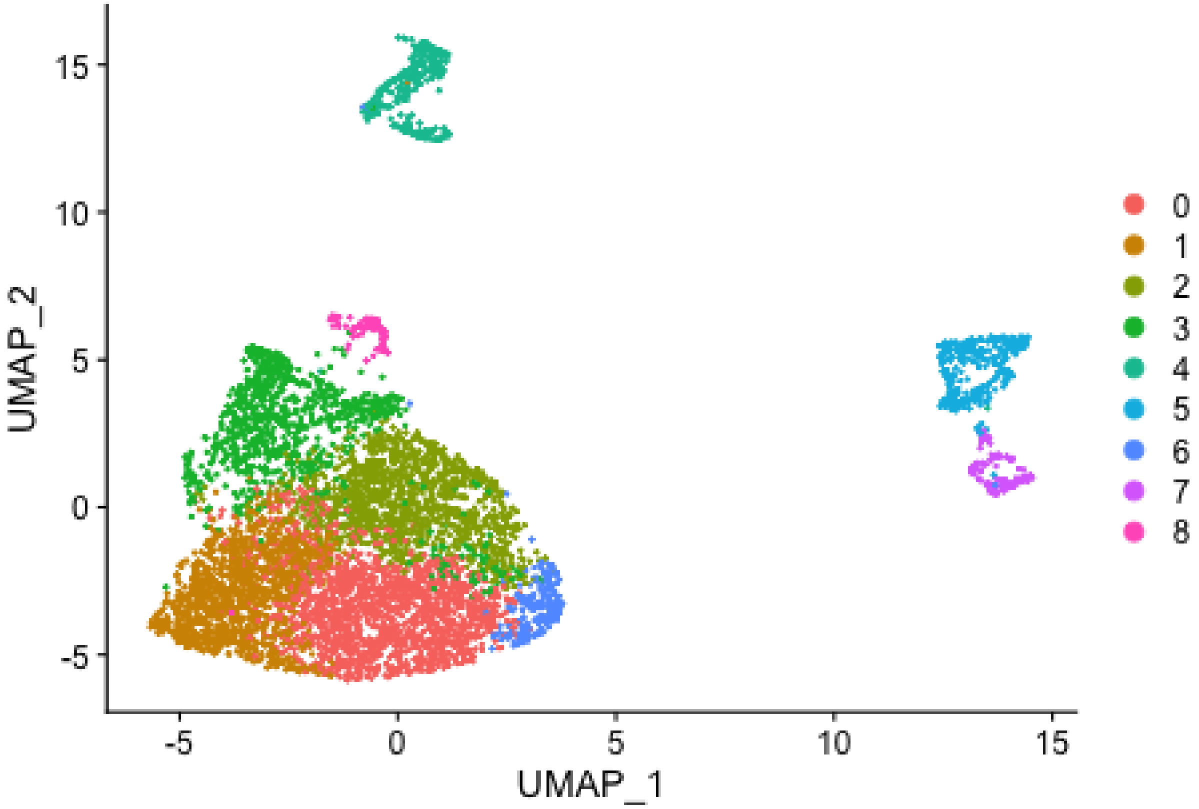


Figure 1: UMAP plot of secreted MECs

5.2.2 t-SNE Plots

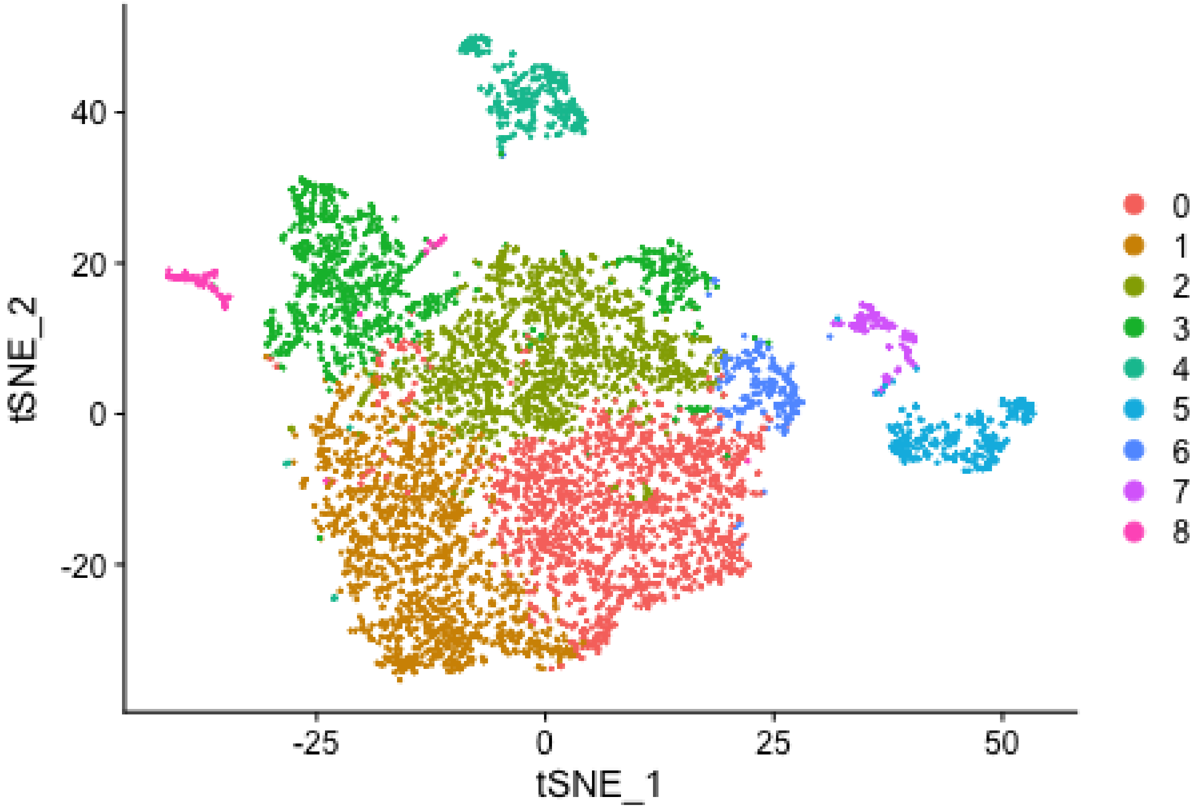


Figure 2: t-SNE plot of secreted MECs

Table 1: Cell Specific Markers (All)

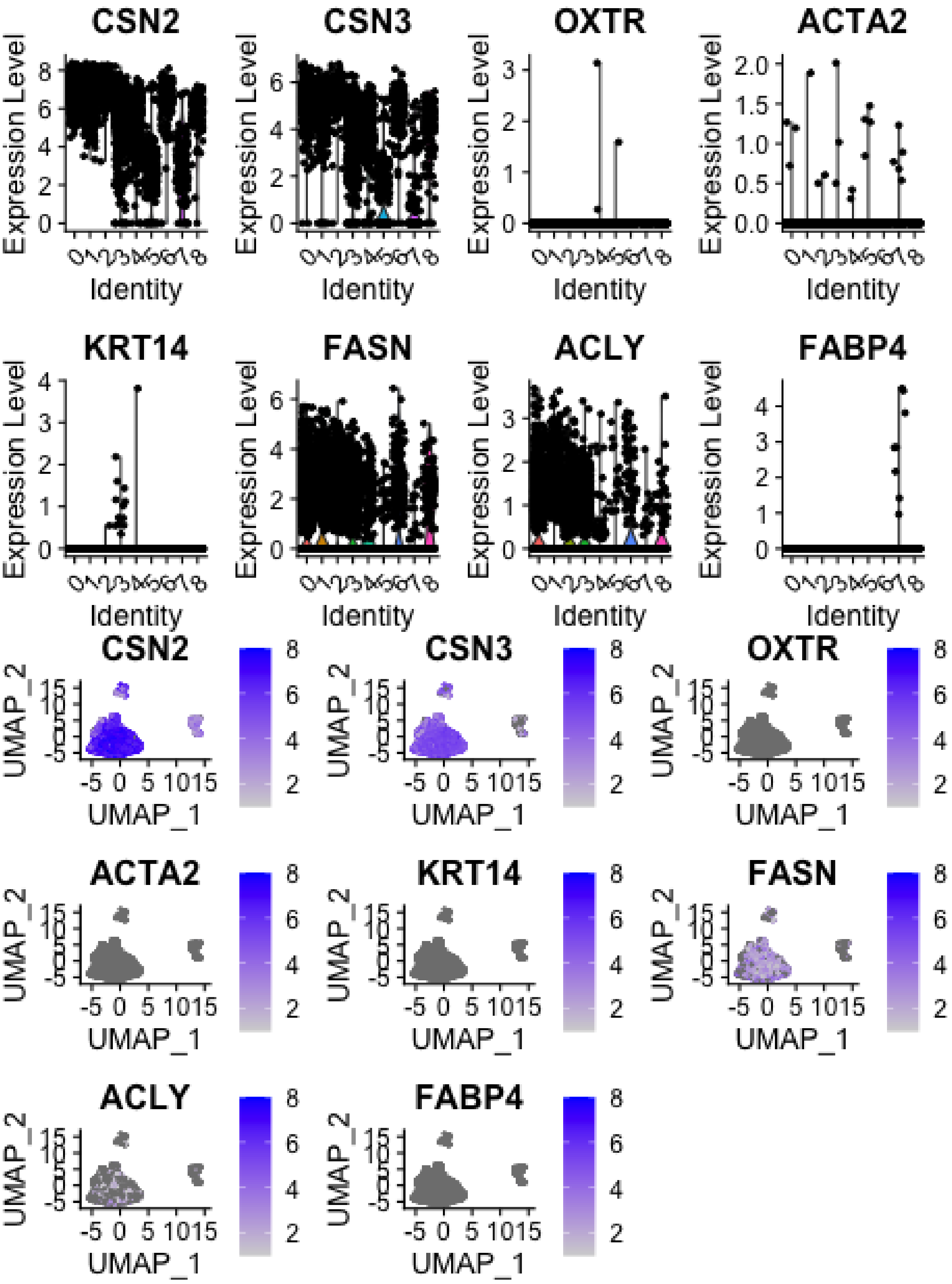
p_val	avg_logFC	pct.1	pct.2	p_val_adj	cluster	gene
0	0.773	0.635	0.077	0	2	C4BPA
0	0.510	0.595	0.055	0	2	HLA-DRB5
0	1.537	0.940	0.574	0	3	CLU
0	1.409	0.600	0.165	0	3	KRT15
0	4.451	0.675	0.042	0	4	MT1E
0	4.749	0.773	0.103	0	4	MT2A
0	3.261	0.997	0.136	0	5	MALAT1
0	2.832	0.994	0.545	0	5	MT-ATP6
0	1.743	1.000	0.666	0	6	MT-CO1
0	1.606	1.000	0.651	0	6	MT-CO3
0	3.163	0.852	0.422	0	7	CD74
0	3.184	0.911	0.834	0	7	FTL
0	1.985	0.717	0.083	0	8	STMN1
0	1.962	0.808	0.319	0	8	H2AFZ

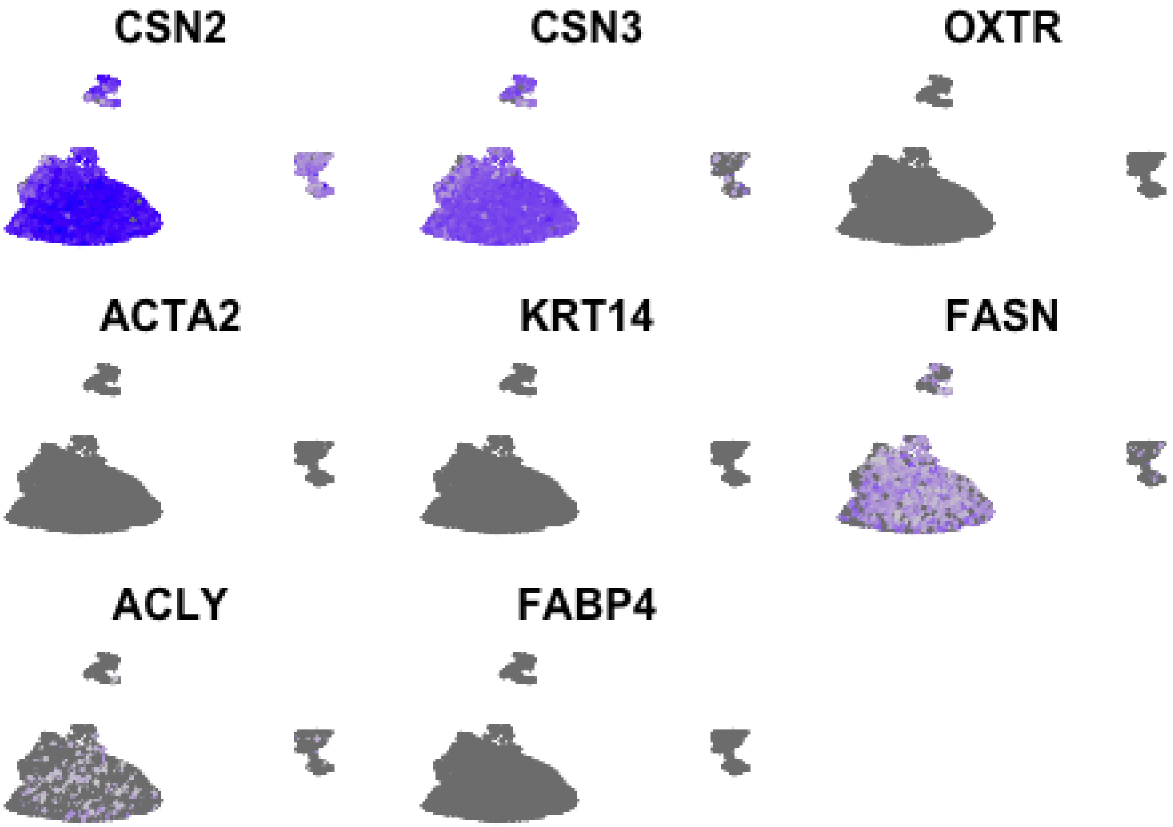
6 Feature Analysis

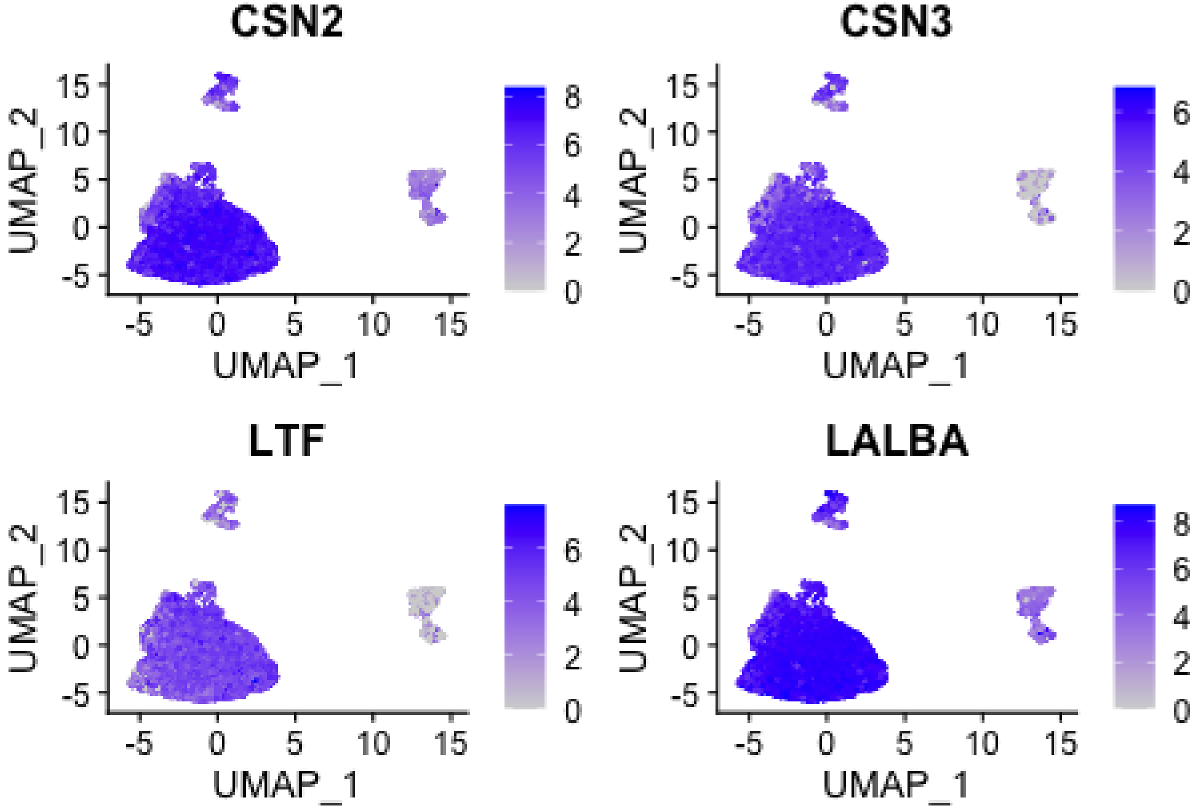
6.1 Analysis of clusters

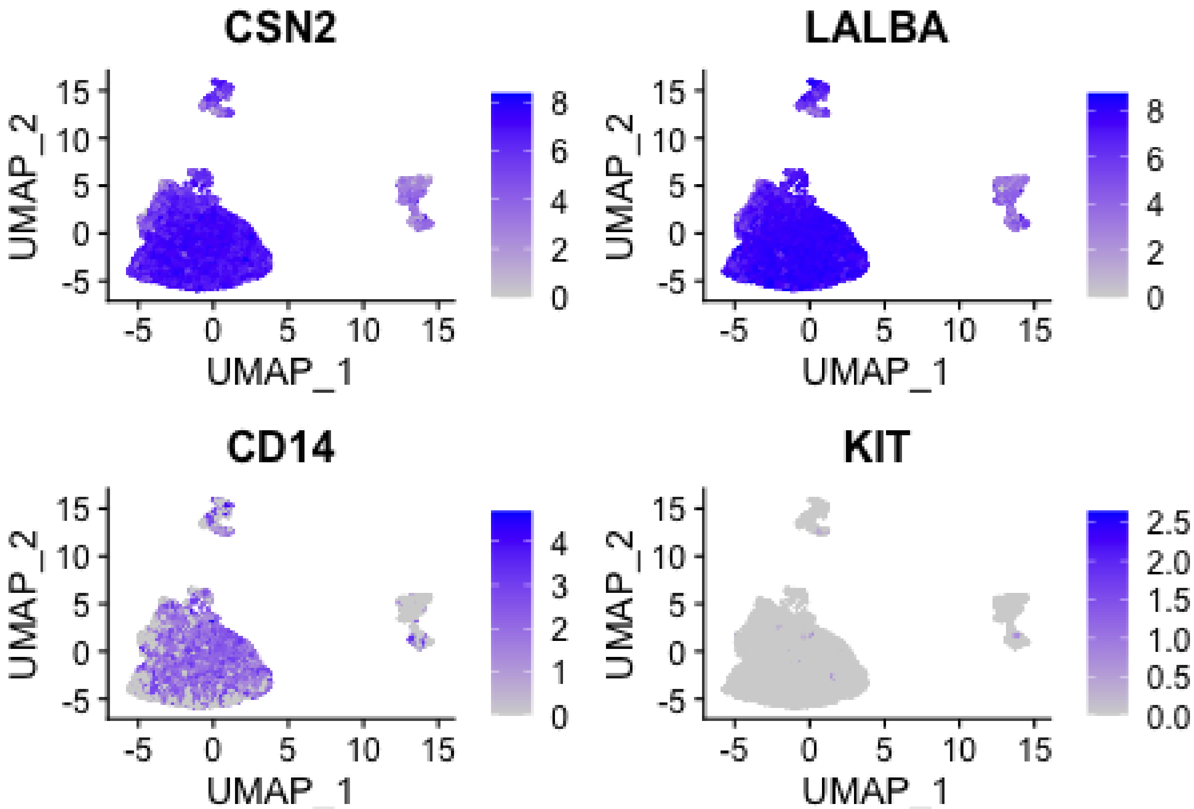
Table 2: Clusters of genes of interest

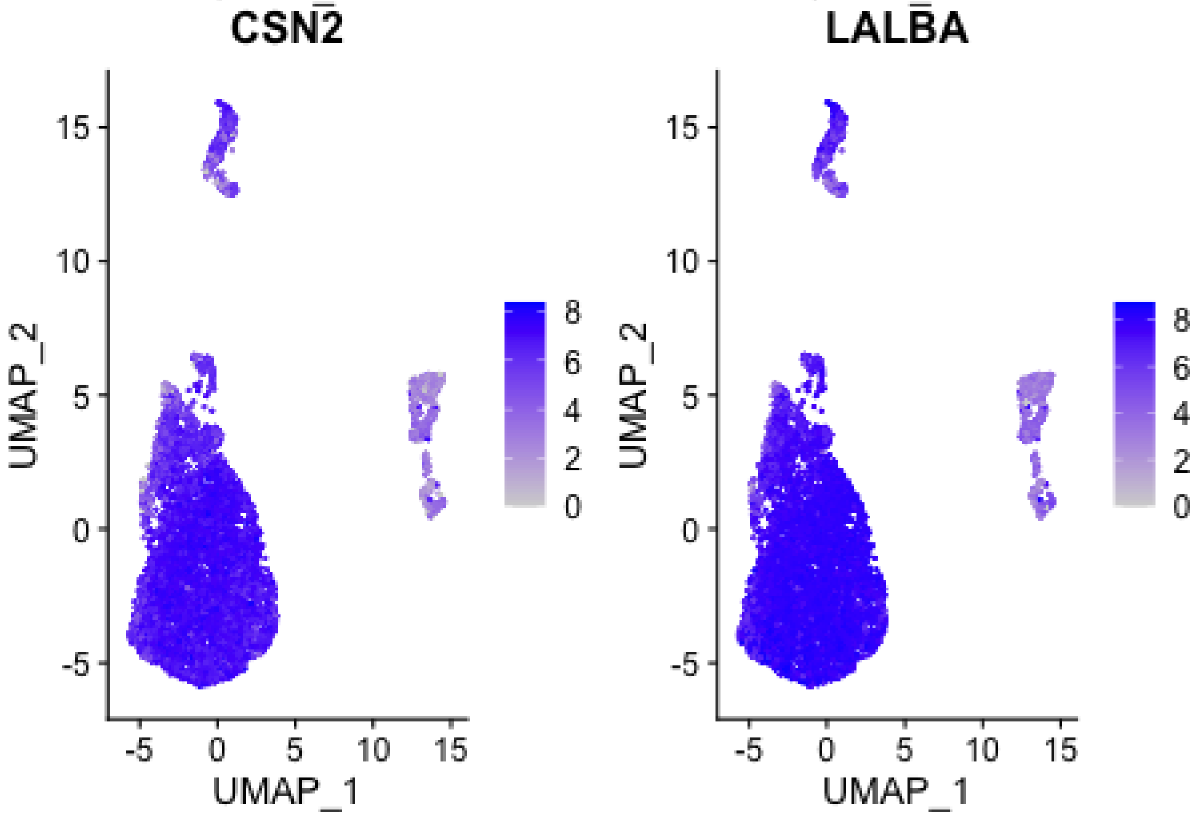
	p_val	avg_logFC	pct.1	pct.2	p_val_adj	cluster	gene
CSN2	0	0.504	1.000	0.988	0.000	2	CSN2
FASN	0	0.735	0.848	0.756	0.007	6	FASN

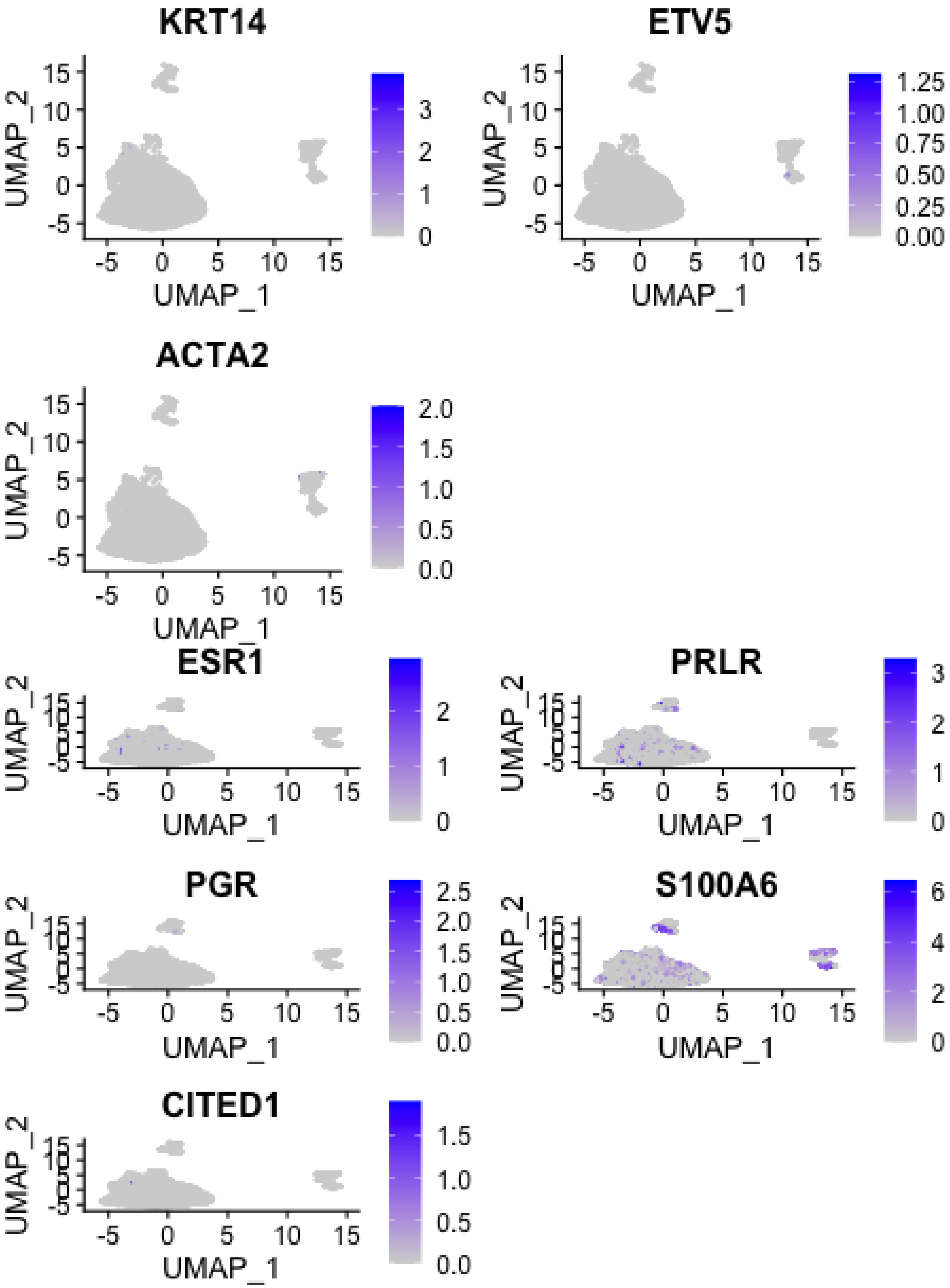

CSN2
CSN3
OXTR
ACTA2
KRT14
FASN
ACLY
FABP4
Expression Level
Identity
0 1 2 3 4 5 6 7 8
0 2 4 6 8
0 2 4 6
0 1 2 3
0.0 0.5 1.0 1.5 2.0
0 1 2 3 4
0 2 4 6
0 1 2 3
0 1 2 3 4
CSN2
CSN3
OXTR
ACTA2
KRT14
FASN
ACLY
FABP4
UMAP_2
UMAP_1
-5 0 5 10 15
-5 0 5 10 15
2 4 6 8

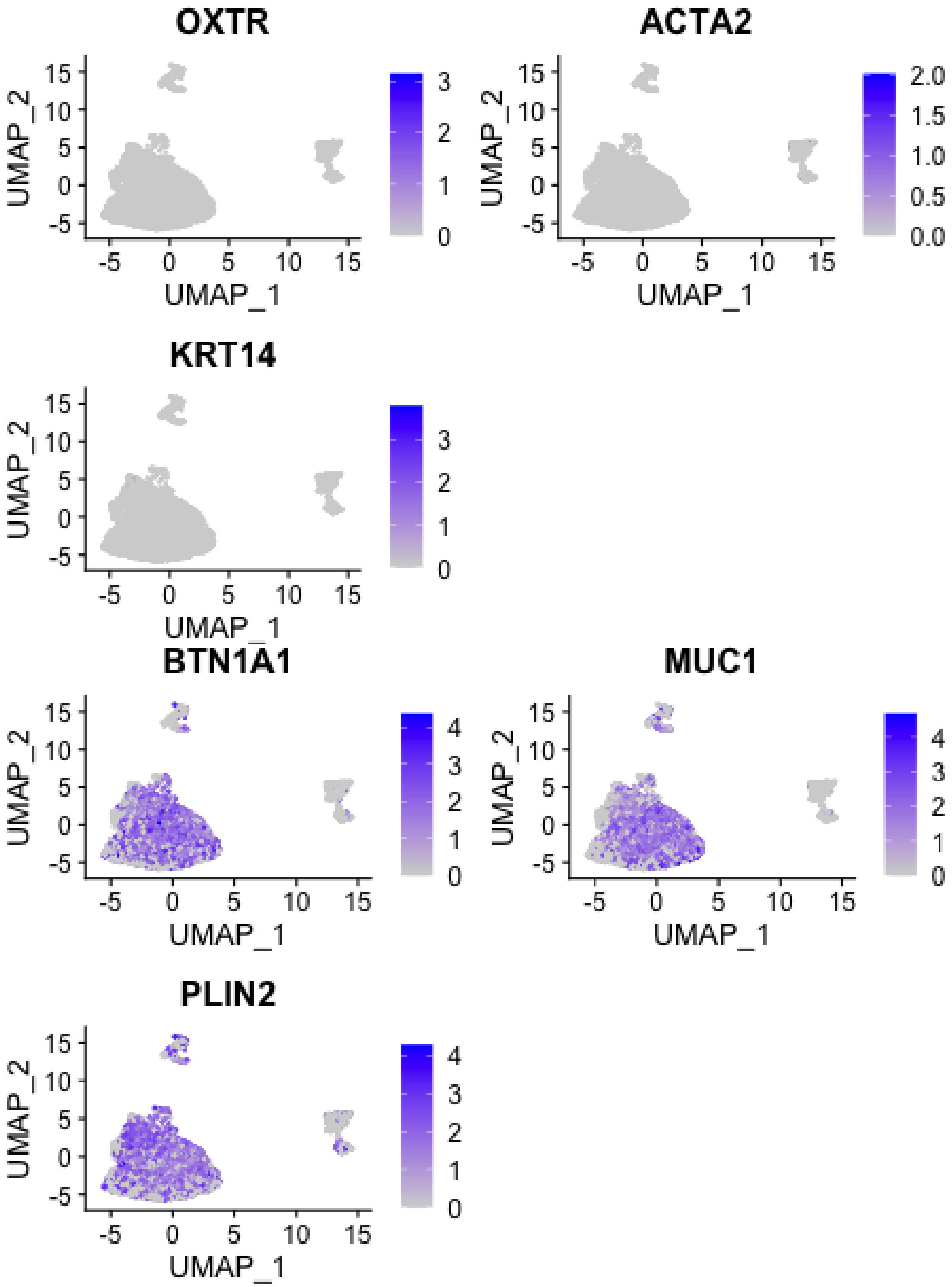

CSN2
CSN3
OXTR
ACTA2
KRT14
FASN
ACLY
FABP4

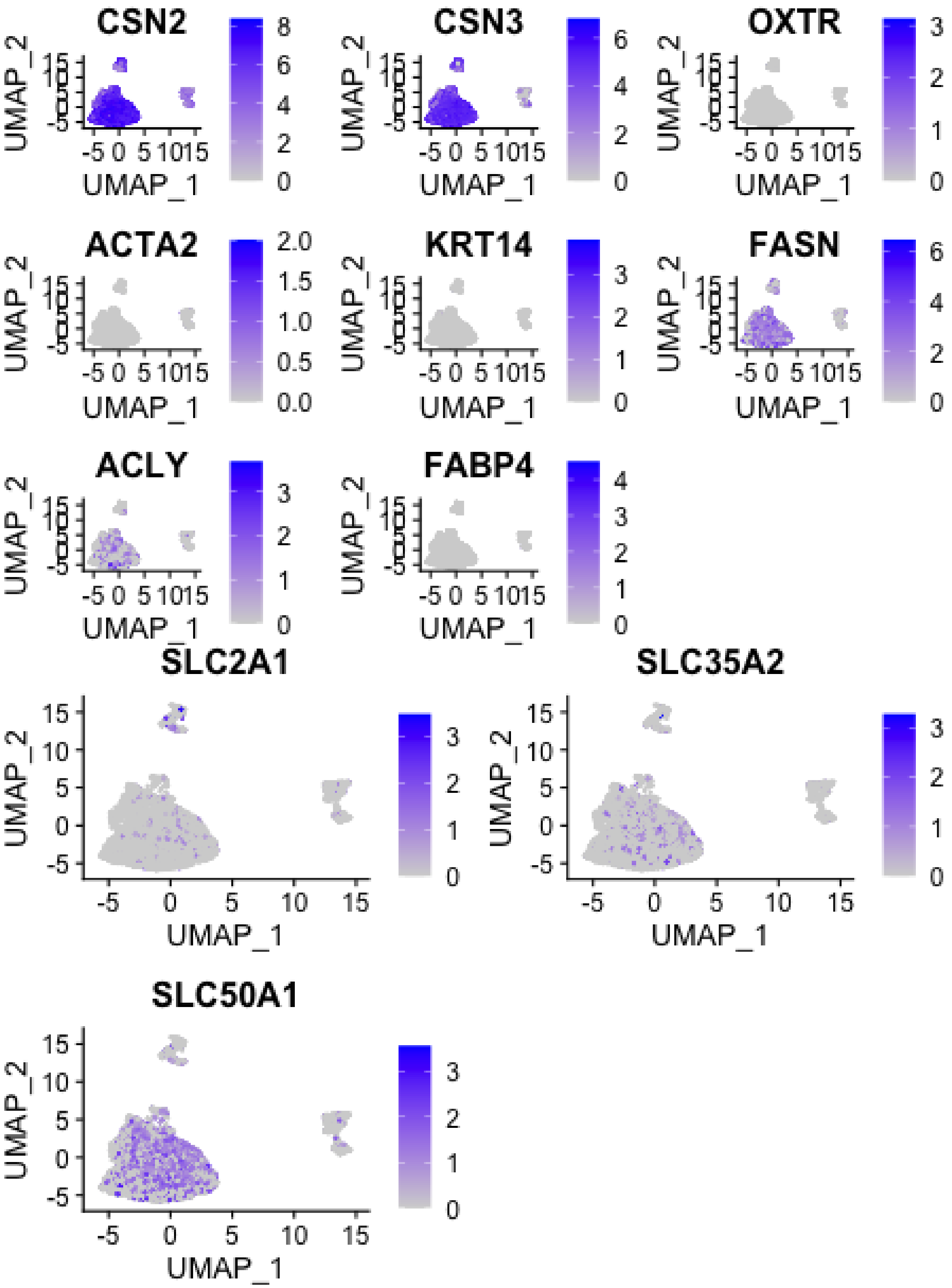

CSN2
CSN3
LTF
LALBA
UMAP_2
UMAP_1
15
10
5
0
-5
-5
0
5
10
15
8
6
4
2
0

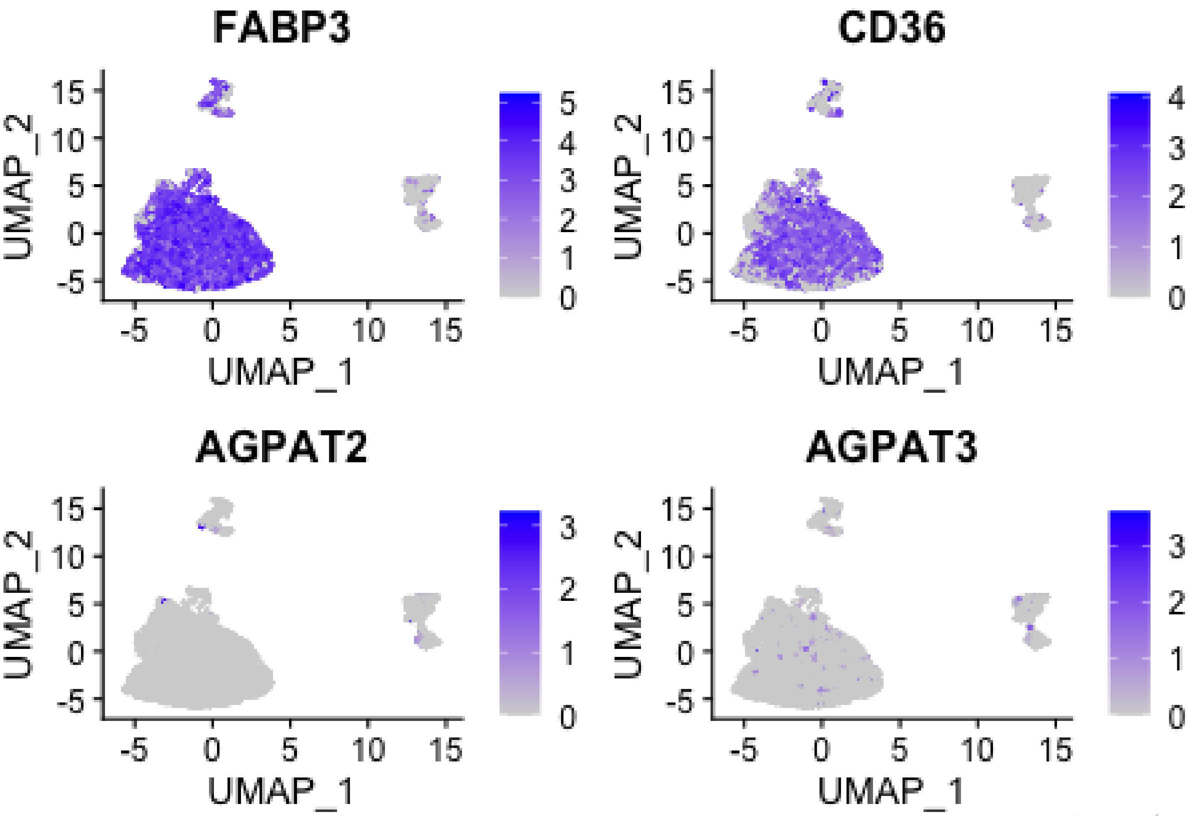

CSN2
LALBA
CD14
KIT
UMAP_1
UMAP_2
-5 0 5 10 15
15 10 5 0 -5
8 6 4 2 0
4 3 2 1 0
2.5 2.0 1.5 1.0 0.5 0.0

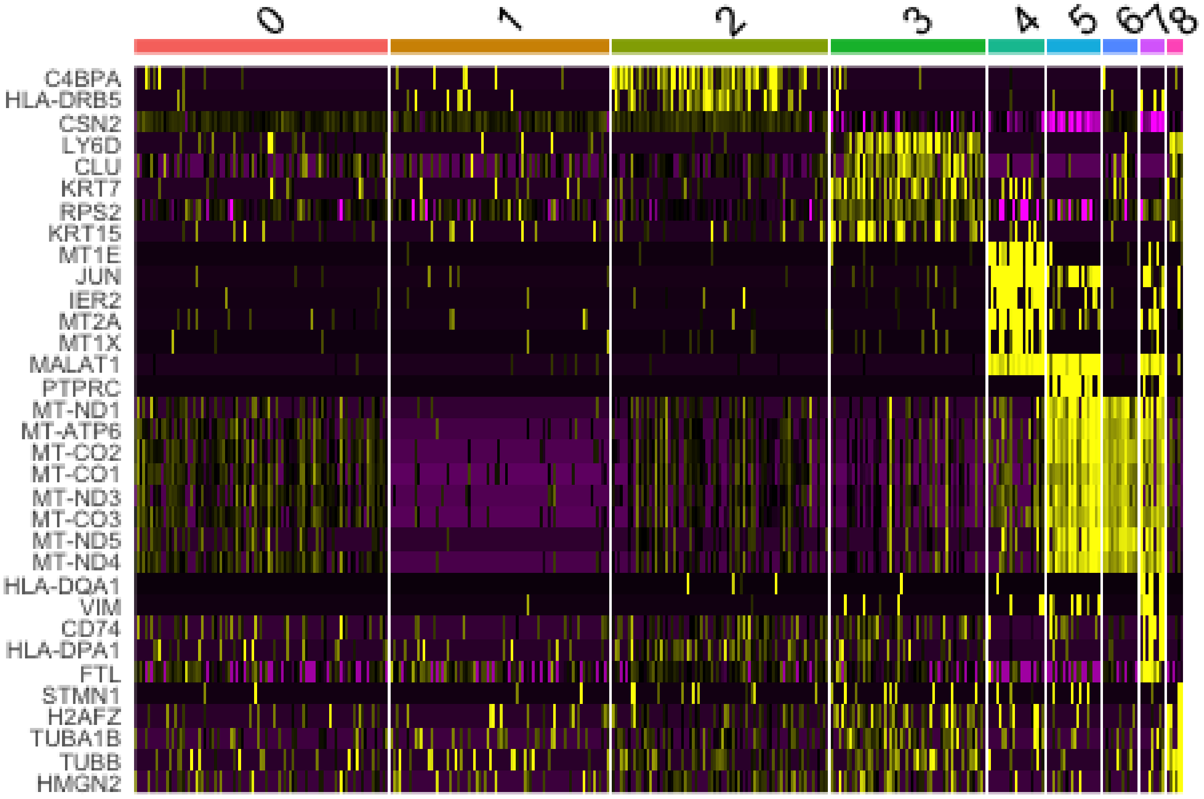

CSN2
LALBA
UMAP_1
UMAP_2
-5 0 5 10 15
15 10 5 0 -5
8 6 4 2 0


KRT14
UMAP_2
15
10
5
0
-5
-5
0
5
10
15
UMAP_1
3
2
1
0
ETV5
UMAP_2
15
10
5
0
-5
-5
0
5
10
15
UMAP_1
1.25
1.00
0.75
0.50
0.25
0.00
ACTA2
UMAP_2
15
10
5
0
-5
-5
0
5
10
15
UMAP_1
2.0
1.5
1.0
0.5
0.0
ESR1
UMAP_2
15
10
5
0
-5
-5
0
5
10
15
UMAP_1
2
1
0
PRLR
UMAP_2
15
10
5
0
-5
-5
0
5
10
15
UMAP_1
3
2
1
0
PGR
UMAP_2
15
10
5
0
-5
-5
0
5
10
15
UMAP_1
2.5
2.0
1.5
1.0
0.5
0.0
S100A6
UMAP_2
15
10
5
0
-5
-5
0
5
10
15
UMAP_1
6
4
2
0
CITED1
UMAP_2
15
10
5
0
-5
-5
0
5
10
15
UMAP_1
1.5
1.0
0.5
0.0


OXTR
ACTA2
KRT14
BTN1A1
MUC1
PLIN2
UMAP_1
UMAP_2
-5
0
5
10
15
0
1
2
3
4
0.0
0.5
1.0
1.5
2.0


CSN2
CSN3
OXTR
ACTA2
KRT14
FASN
ACLY
FABP4
SLC2A1
SLC35A2
SLC50A1
UMAP_1
UMAP_2
-5 0 5 10 15
8 6 4 2 0
6 4 2 0
3 2 1 0
2.0 1.5 1.0 0.5 0.0
3 2 1 0
6 4 2 0
3 2 1 0
4 3 2 1 0
3 2 1 0
3 2 1 0
3 2 1 0


FABP3
CD36
AGPAT2
AGPAT3
UMAP_2
UMAP_1
15
10
5
0
-5
-5
0
5
10
15
5
4
3
2
1
0


0
1
2
3
4
5
6
7
8
C4BPA
HLA-DRB5
CSN2
LY6D
CLU
KRT7
RPS2
KRT15
MT1E
JUN
IER2
MT2A
MT1X
MALAT1
PTPRC
MT-ND1
MT-ATP6
MT-CO2
MT-CO1
MT-ND3
MT-CO3
MT-ND5
MT-ND4
HLA-DQA1
VIM
CD74
HLA-DPA1
FTL
STMN1
H2AFZ
TUBA1B
TUBB
HMGN2

7 Session Information

```
sessionInfo()
```

```
## R version 4.0.2 (2020-06-22)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS  10.16
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRblas.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] ggplot2_3.3.2 Seurat_3.2.3  dplyr_1.0.2   tidyr_1.1.2   knitr_1.30
##
## loaded via a namespace (and not attached):
##   [1] nlme_3.1-151          matrixStats_0.57.0    RcppAnnoy_0.0.18
##   [4] RColorBrewer_1.1-2    httr_1.4.2            sctransform_0.3.2
##   [7] tools_4.0.2           R6_2.5.0              irlba_2.3.3
##  [10] rpart_4.1-15          KernSmooth_2.23-18    uwot_0.1.10
##  [13] mgcv_1.8-33           lazyeval_0.2.2        colorspace_2.0-0
##  [16] withr_2.3.0           tidyselect_1.1.0      gridExtra_2.3
##  [19] compiler_4.0.2        plotly_4.9.2.2        labeling_0.4.2
##  [22] scales_1.1.1          lmtest_0.9-38         spatstat.data_1.7-0
##  [25] ggridges_0.5.2        pbapply_1.4-3         spatstat_1.64-1
##  [28] goftest_1.2-2         stringr_1.4.0         digest_0.6.27
##  [31] spatstat.utils_1.17-0 rmarkdown_2.6         pkgconfig_2.0.3
##  [34] htmltools_0.5.0       parallelly_1.22.0     limma_3.44.3
##  [37] highr_0.8             fastmap_1.0.1         htmlwidgets_1.5.3
##  [40] rlang_0.4.10          shiny_1.5.0           farver_2.0.3
##  [43] generics_0.1.0        zoo_1.8-8             jsonlite_1.7.2
##  [46] ica_1.0-2             magrittr_2.0.1        patchwork_1.1.1
##  [49] Matrix_1.2-18         Rcpp_1.0.5            munsell_0.5.0
##  [52] abind_1.4-5           reticulate_1.18       lifecycle_0.2.0
##  [55] stringi_1.5.3         yaml_2.2.1            MASS_7.3-53
##  [58] Rtsne_0.15            plyr_1.8.6            grid_4.0.2
##  [61] parallel_4.0.2        listenv_0.8.0         promises_1.1.1
##  [64] ggrepel_0.9.0         crayon_1.3.4          deldir_0.2-3
##  [67] miniUI_0.1.1.1        lattice_0.20-41       cowplot_1.1.0
##  [70] splines_4.0.2         tensor_1.5            magick_2.5.2
##  [73] pillar_1.4.7          igraph_1.2.6          future.apply_1.6.0
##  [76] reshape2_1.4.4        codetools_0.2-18      leiden_0.3.6
##  [79] glue_1.4.2            evaluate_0.14         data.table_1.13.4
##  [82] vctrs_0.3.6           png_0.1-7             httpuv_1.5.4
##  [85] gtable_0.3.0          RANN_2.6.1            purrr_0.3.4
##  [88] polyclip_1.10-0       scattermore_0.7       future_1.21.0
##  [91] xfun_0.19             rsvd_1.0.3            mime_0.9
##  [94] xtable_1.8-4         RSpectra_0.16-0      later_1.1.0.1
##  [97] survival_3.2-7       viridisLite_0.3.0    tibble_3.0.4
## [100] cluster_2.1.0        globals_0.14.0       fitdistrplus_1.1-3
## [103] ellipsis_0.3.1       ROCR_1.0-11
```

Macosko, Evan Z., Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, et al. 2015. "Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets." *Cell* 161 (5). Elsevier: 1202–14. https://doi.org/10.1016/j.cell.2015.05.002.

Martin Carli, Jayne F., G. Devon Trahan, Kenneth L. Jones, Nicole Hirsch, Kristy P. Rolloff, Emily Z. Dunn, Jacob E. Friedman, et al. 2020. "Single Cell RNA Sequencing of Human Milk-Derived Cells Reveals Sub-Populations of Mammary Epithelial Cells with Molecular Signatures of Progenitor and Mature States: a Novel, Non-invasive Framework for Investigating Human Lactation Physiology." *Journal of Mammary Gland Biology and Neoplasia*. Journal of Mammary Gland Biology; Neoplasia. https://doi.org/10.1007/s10911-020-09466-z.