Manual Annotation of Pooled Cells

Ariel Hippen

2022-06-14

Contents

R Markdown

This is the code I used to cluster the pooled data and identify markers that eventually turned into cell types. Most of the heavy lifting was done with internet searches of the markers findMarkers() recovered. Notes from that search are stored in a google doc in my drive; I'm currently trying to decide the best way to preserve that info for reproducibility.

```
suppressPackageStartupMessages({
  library(data.table)
  library(scater)
  library(scran)
  library(batchelor)
  library(igraph)
  library(WebGestaltR)
})
```

Load and prep data

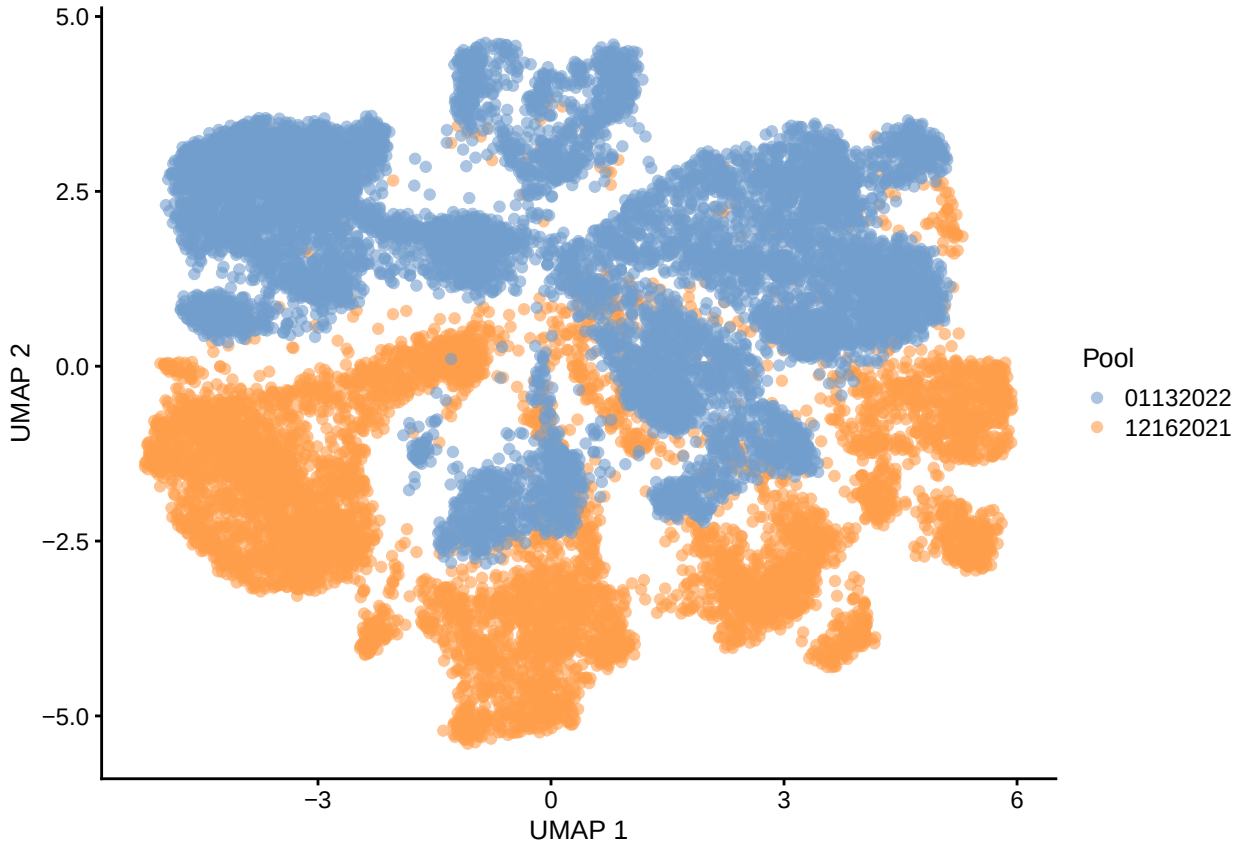
```
# Load and combine pooled data objects
pool1 <- readRDS("../../data/sce_objects/12162021.rds")
pool1$Pool <- "12162021"
pool2 <- readRDS("../../data/sce_objects/01132022.rds")
```

```
pool2$Pool <- "01132022"

sce <- cbind(pool1, pool2)
```

Plot and check for pool-specific structure suggesting batch effects

```
set.seed(531)
sce <- runUMAP(sce,
               BNPARAM = BiocNeighbors::AnnoyParam(),
               BPPARAM = BiocParallel::MulticoreParam(),
               min_dist = 0.5,  repulsion_strength = 0.25,
               spread = 0.7,
               n_neighbors = 15)
plotUMAP(sce, colour_by = "Pool")
```



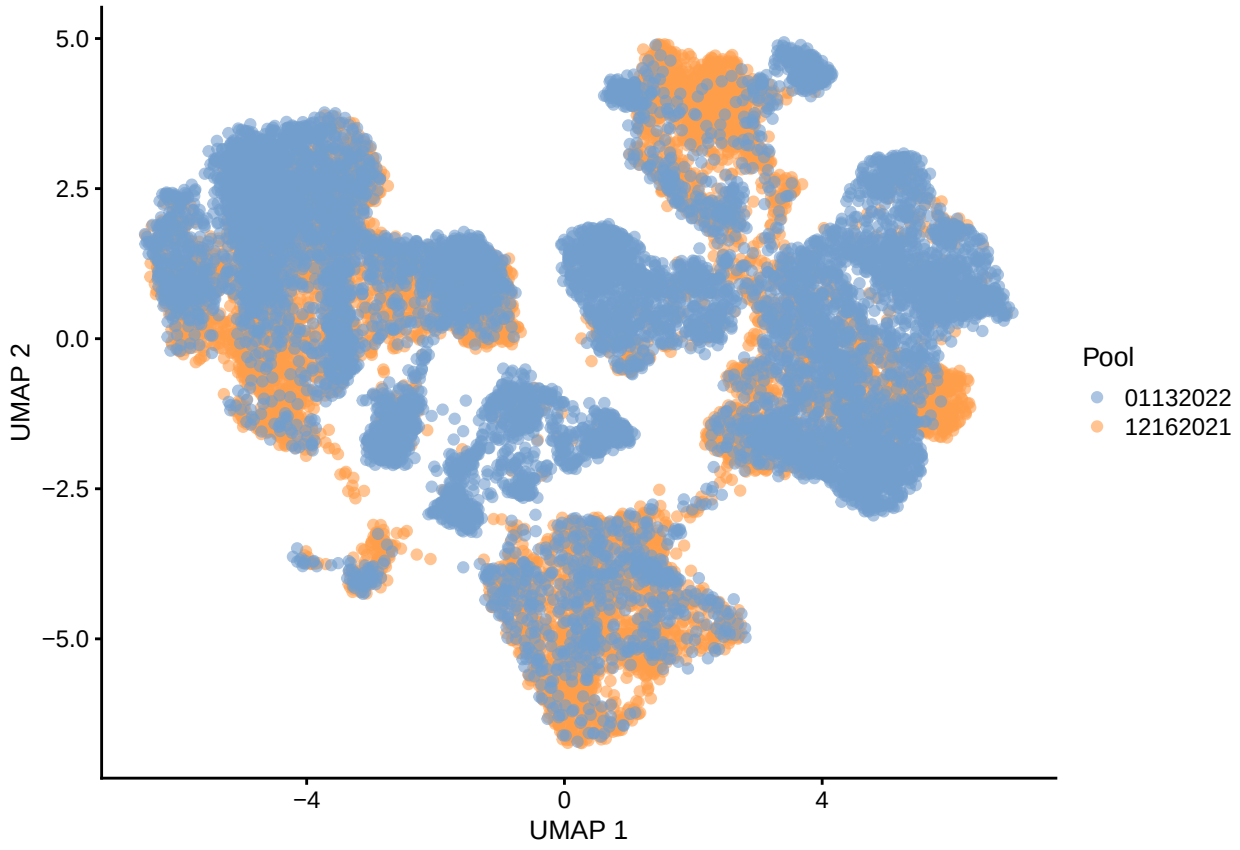
Run batchelor to correct for batch effects

```
set.seed(531)
mnn <- fastMNN(sce,  batch = sce$Pool,
               BSPARAM = BiocSingular::IrlbaParam(deferred = TRUE),
               BNPARAM = BiocNeighbors::AnnoyParam(),
               BPPARAM = BiocParallel::MulticoreParam())

reducedDim(sce,  "MNN") <- reducedDim(mnn,  "corrected")
rm(mnn)
```

Now we can replot to confirm that the pool-specific structure looks better

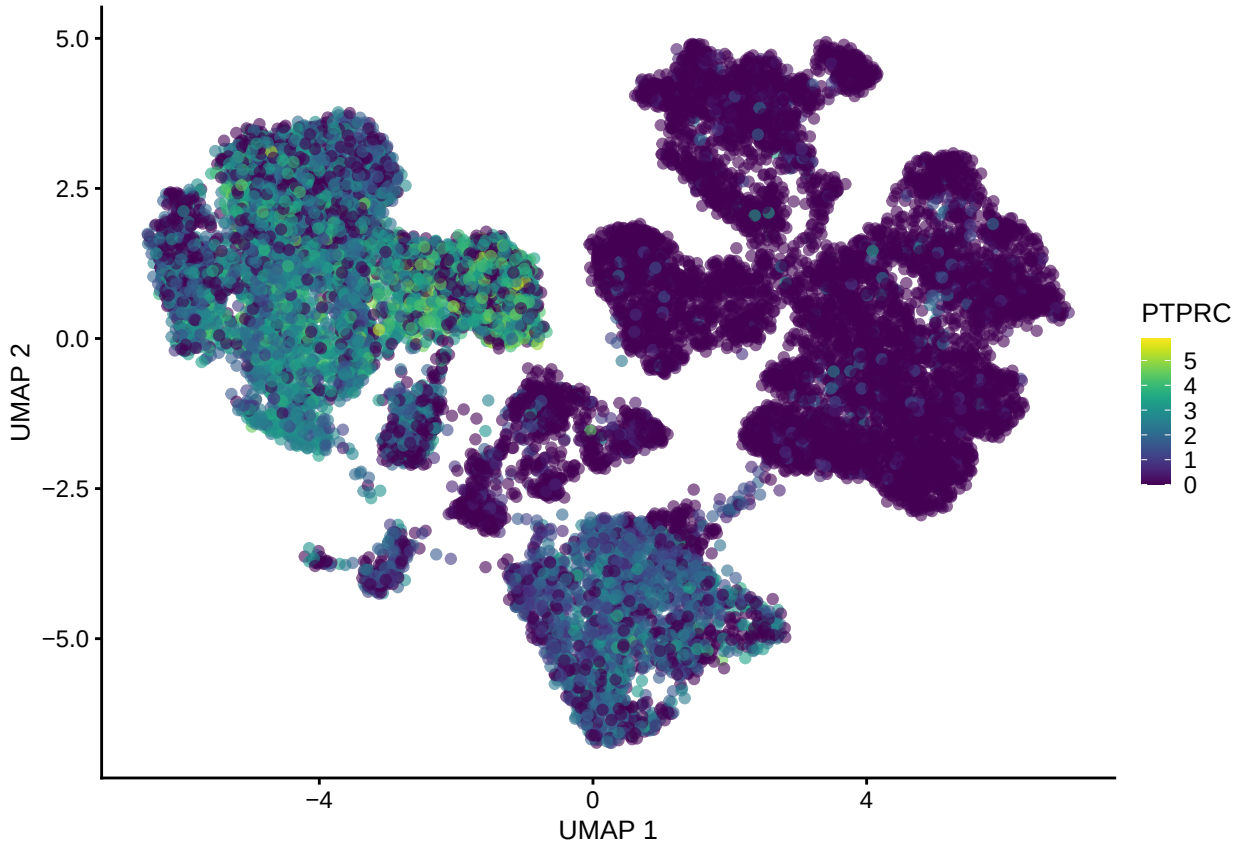
```
set.seed(531)
sce <- runUMAP(sce,
               dimred = "MNN",
               BNPARAM = BiocNeighbors::AnnoyParam(),
               BPPARAM = BiocParallel::MulticoreParam(),
               min_dist = 0.5,  repulsion_strength = 0.25,
               spread = 0.7,
               n_neighbors = 15)
plotUMAP(sce, colour_by = "Pool")
```



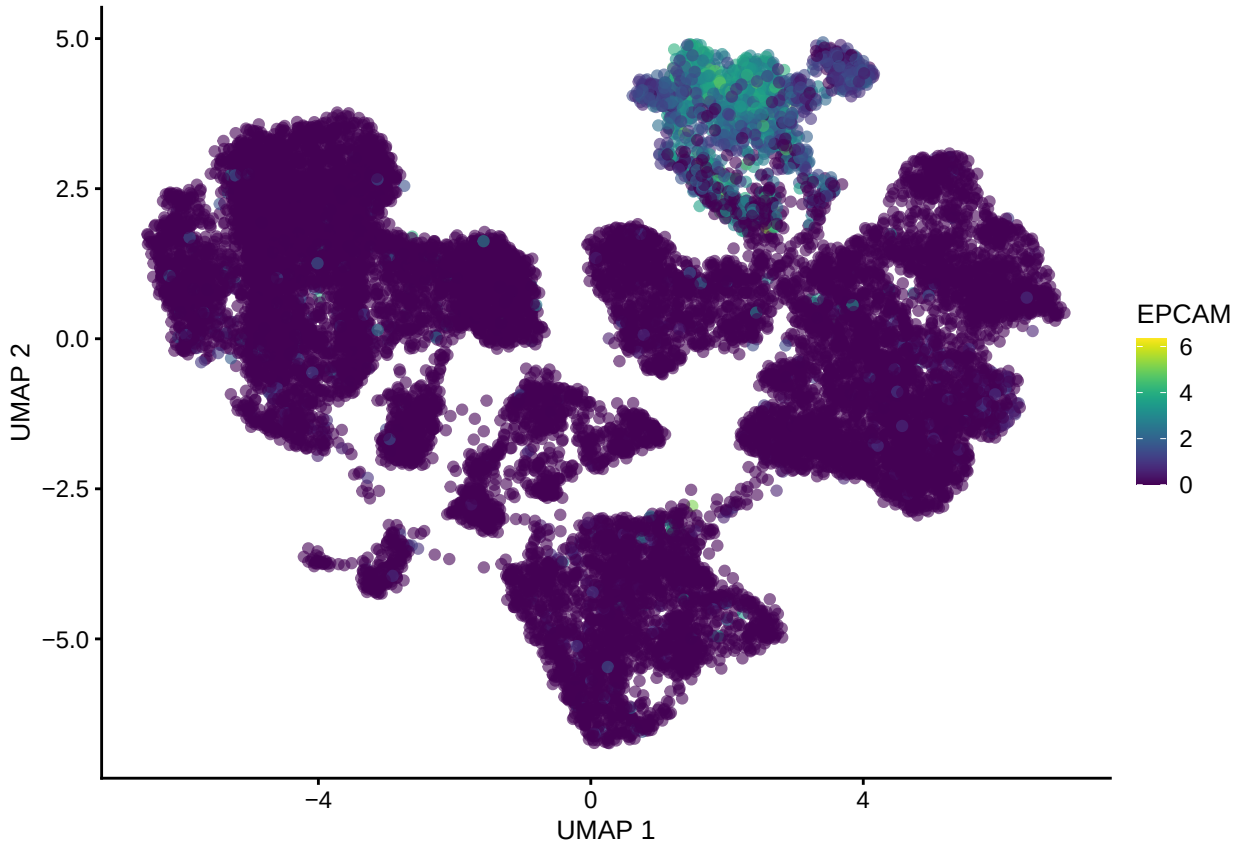
Check for basic markers

Look for basic cell type markers, PTPRC (CD45) for immune, ACTA2 (smooth muscle actin) for fibroblasts, and EPCAM and PAX8 for epithelial/cancer cells.

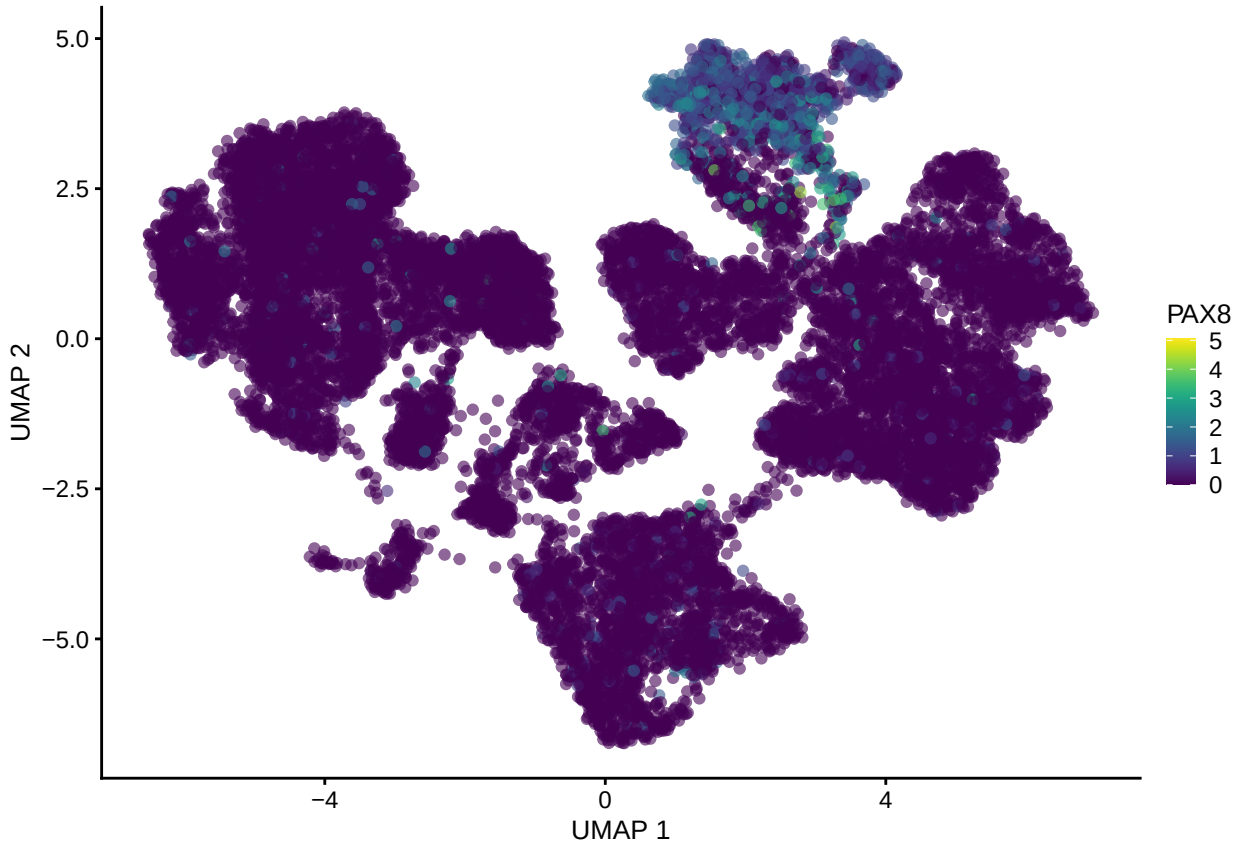
```
plotUMAP(sce, colour_by = "PTPRC")
```



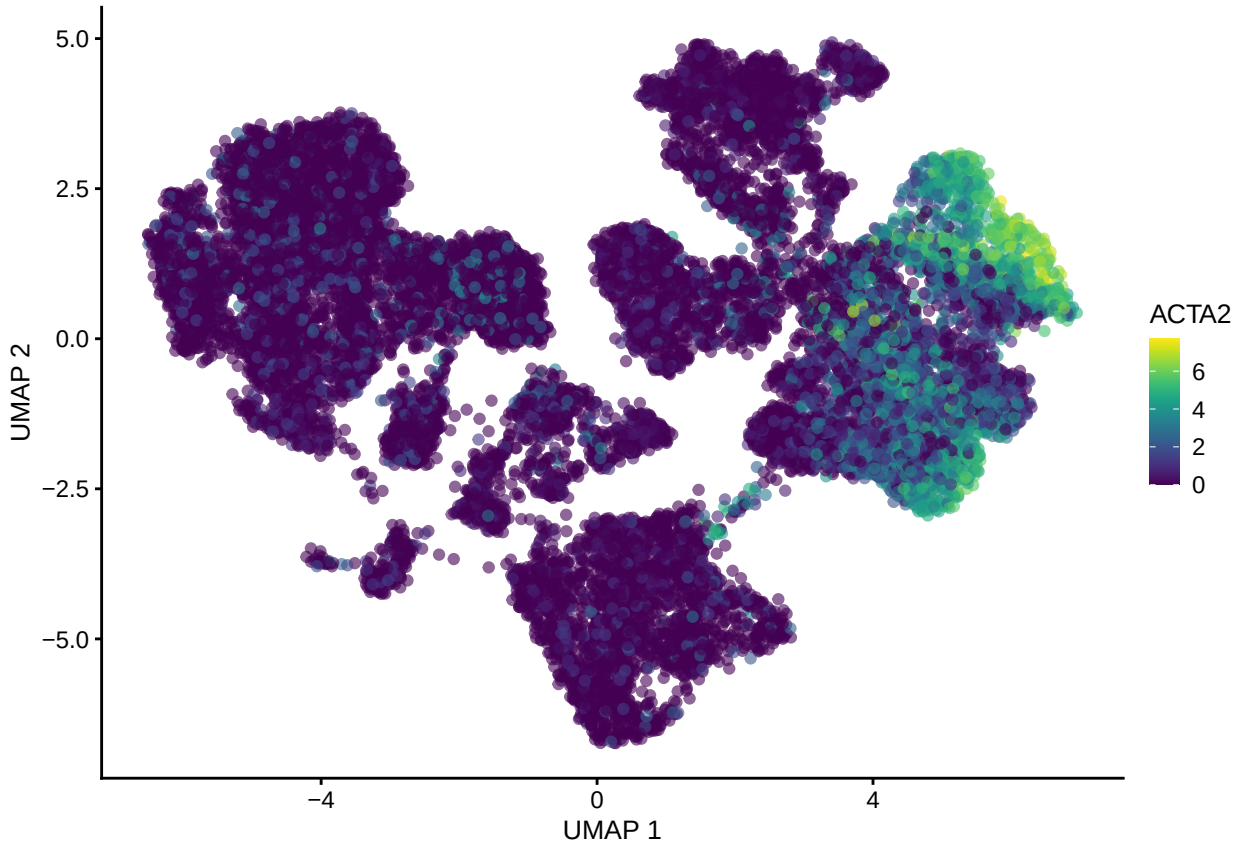

```
plotUMAP(sce, colour_by = "EPCAM")
```




```
plotUMAP(sce, colour_by = "PAX8")
```




```
plotUMAP(sce, colour_by = "ACTA2")
```

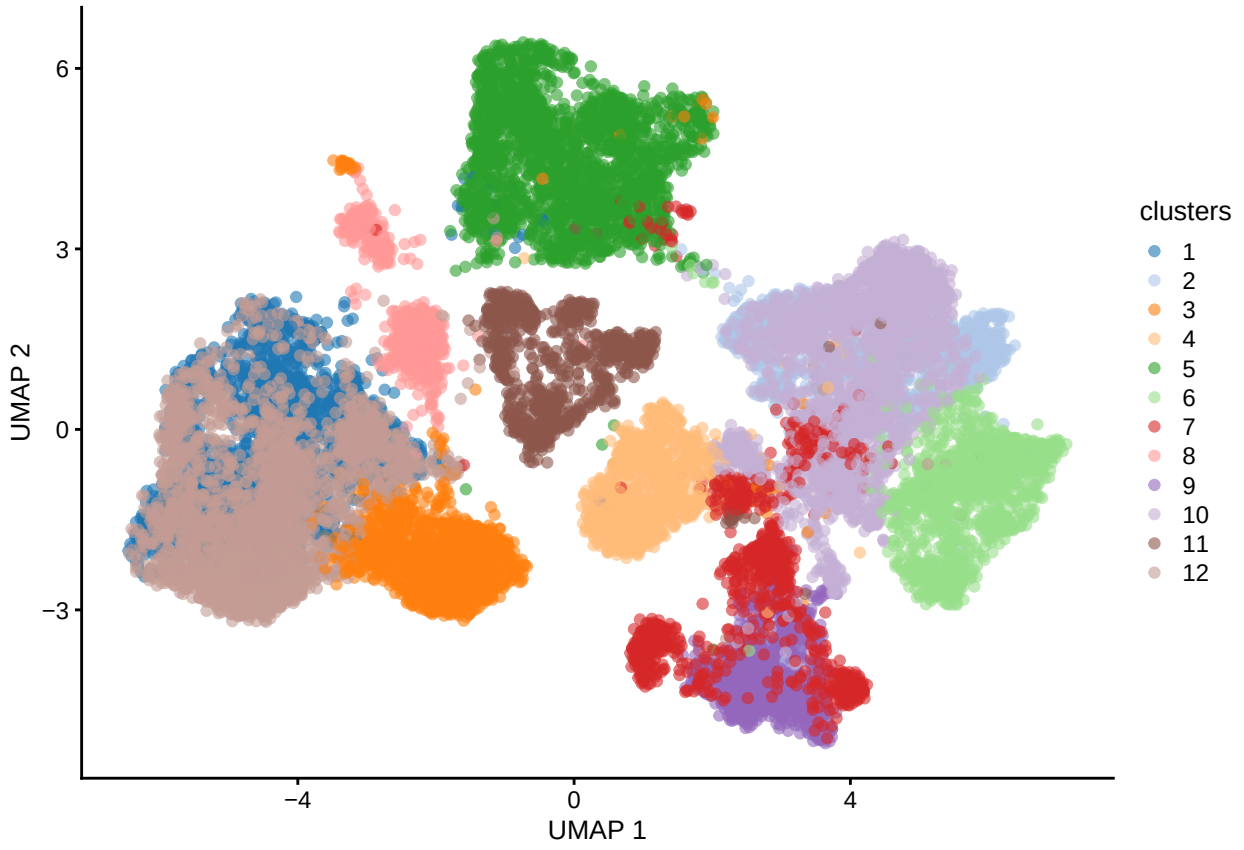


Cluster cells

Creating the SNN graph takes >1hr, so I will leave the code in to generate the clusters but after the first time just load in the existing RDS object. Also, I initially ran this several times with k=30, 50, and 80, but decided to go with the resolution at 50.

```
#g <- buildSNNGraph(sce,
#                    k = 50,    # higher = bigger clusters
#                    BNPARAM = BiocNeighbors::AnnoyParam(),
#                    BPPARAM = BiocParallel::MulticoreParam())
#clusters <- as.factor(igraph::cluster_louvain(g)$membership)
#sce$clusters <- clusters
#saveRDS(sce,  file = "../../data/sce_objects/pooled_clustered_50.rds")
sce <- readRDS("../../data/sce_objects/pooled_clustered_50.rds")

plotUMAP(sce, colour_by = "clusters")
```



Annotate clusters

For the annotation, we'll be using scran's findMarkers function. It returns results for all genes regardless of significance, so for clarity we'll pull out the ones with a FDR of <=0.05.

```
getOnlySignificantGenes <- function(mylist) {
  for (i in 1:length(mylist)) {
    x <- mylist[[i]]
    x <- subset(x, select = c("p.value", "FDR"))
    y <- subset(x, x$FDR <= 0.05)
    if (nrow(y) > 0) {
      y <- y[order(y$FDR), ]
    }
    mylist[[i]] <- y
  }
  as.list(mylist)
}
```

Let's start by seeing if we can pin down any particular clusters with genes that are upregulated compared to all other clusters

```
markers <- findMarkers(sce,  groups = sce$clusters,
                       block = sce$Pool,  # use to get within-donor DE
                       direction = "up",  lfc = 1.5,
```

```
                      pval.type = "all",
                      BPPARAM = BiocParallel::MulticoreParam())

sig_genes <- getOnlySignificantGenes(markers)
sig_genes
```

```
## $`1`
## DataFrame with 0 rows and 2 columns
##
## $`2`
## DataFrame with 0 rows and 2 columns
##
## $`3`
## DataFrame with 0 rows and 2 columns
##
## $`4`
## DataFrame with 2 rows and 2 columns
##           p.value         FDR
##         <numeric>   <numeric>
## VWF   3.01196e-12 1.10241e-07
## GNG11 4.79774e-10 8.78010e-06
##
## $`5`
## DataFrame with 10 rows and 2 columns
##             p.value          FDR
##           <numeric>    <numeric>
## LYZ    2.09699e-110 7.67520e-106
## AIF1    5.25783e-76  9.62209e-72
## IFI30   2.93063e-55  3.57547e-51
## TYROBP  6.55316e-47  5.99630e-43
## FCER1G  4.24529e-43  3.10764e-39
## SAT1    4.00316e-31  2.44199e-27
## IL1B    3.77846e-16  1.97565e-12
## NLRP3   2.07608e-13  9.49834e-10
## CXCL8   5.83537e-11  2.37312e-07
## SPP1    5.73430e-08  2.09881e-04
##
## $`6`
## DataFrame with 1 row and 2 columns
##          p.value        FDR
##        <numeric>  <numeric>
## RGS5 1.25699e-12 4.6007e-08
##
## $`7`
## DataFrame with 0 rows and 2 columns
##
## $`8`
## DataFrame with 0 rows and 2 columns
##
## $`9`
## DataFrame with 4 rows and 2 columns
##           p.value         FDR
##         <numeric>   <numeric>
```

```
## CLDN3 3.13653e-11 8.74154e-07
## CLDN4 4.77666e-11 8.74154e-07
## WFDC2 6.01223e-09 7.33512e-05
## SLPI  1.44113e-07 1.31867e-03
##
## $`10`
## DataFrame with 0 rows and 2 columns
##
## $`11`
## DataFrame with 0 rows and 2 columns
##
## $`12`
## DataFrame with 0 rows and 2 columns
```

Still looking for global markers, but with a relaxed lfc

```
markers <- findMarkers(sce,  groups = sce$clusters,
                       block = sce$Pool,  # use to get within-donor DE
                       direction = "up",  lfc = 1.3,
                       pval.type = "all",
                       BPPARAM = BiocParallel::MulticoreParam())

sig_genes <- getOnlySignificantGenes(markers)
sig_genes
```

```
## $`1`
## DataFrame with 0 rows and 2 columns
##
## $`2`
## DataFrame with 1 row and 2 columns
##        p.value       FDR
##      <numeric> <numeric>
## C7 1.28489e-06 0.0470284
##
## $`3`
## DataFrame with 0 rows and 2 columns
##
## $`4`
## DataFrame with 5 rows and 2 columns
##            p.value         FDR
##          <numeric>   <numeric>
## GNG11  4.94782e-16 1.81095e-11
## VWF    3.09048e-14 5.65574e-10
## PECAM1 1.12801e-07 1.37621e-03
## CLDN5  1.32872e-06 1.21581e-02
## RAMP2  3.44463e-06 2.52153e-02
##
## $`5`
## DataFrame with 12 rows and 2 columns
##             p.value          FDR
##           <numeric>    <numeric>
## LYZ    1.25906e-125 4.60827e-121
## AIF1    2.17703e-99  3.98408e-95
```

```
## IFI30   8.22939e-79  1.00401e-74
## TYROBP  1.31926e-65  1.20716e-61
## FCER1G  2.38071e-65  1.74273e-61
## ...              ...          ...
## NLRP3   1.10940e-25  5.07565e-22
## CXCL8   5.53661e-18  2.25162e-14
## SPP1    4.58077e-13  1.67661e-09
## S100A9  6.08069e-07  2.02327e-03
## CD14    7.65627e-06  2.33523e-02
##
## $`6`
## DataFrame with 2 rows and 2 columns
##            p.value         FDR
##          <numeric>   <numeric>
## RGS5   1.38976e-14 5.08665e-10
## COL4A1 6.41900e-07 1.17471e-02
##
## $`7`
## DataFrame with 0 rows and 2 columns
##
## $`8`
## DataFrame with 0 rows and 2 columns
##
## $`9`
## DataFrame with 5 rows and 2 columns
##           p.value         FDR
##         <numeric>   <numeric>
## CLDN4 4.09833e-15 1.50003e-10
## CLDN3 1.14360e-12 2.09284e-08
## WFDC2 1.73676e-11 2.11891e-07
## SLPI  1.05436e-10 9.64767e-07
## ELF3  2.52918e-06 1.85141e-02
##
## $`10`
## DataFrame with 0 rows and 2 columns
##
## $`11`
## DataFrame with 0 rows and 2 columns
##
## $`12`
## DataFrame with 0 rows and 2 columns
```

Look for markers that are upregulated compared to at least some clusters

```
markers <- findMarkers(sce,  groups = sce$clusters,
                       block = sce$Pool,  # use to get within-donor DE
                       direction = "up",  lfc = 1.5,
                       pval.type = "some",
                       BPPARAM = BiocParallel::MulticoreParam())

sig_genes <- getOnlySignificantGenes(markers)
sig_genes
```

```
## $`1`
```

```
## DataFrame with 25 rows and 2 columns
##            p.value         FDR
##          <numeric>   <numeric>
## CCL5   2.63973e-57 9.66169e-53
## HCST   9.46207e-52 1.73161e-47
## CD52   4.81951e-44 5.87997e-40
## CD69   7.60552e-34 6.95924e-30
## IL32   4.38617e-32 3.21076e-28
## ...            ...         ...
## BTG1   1.10745e-07 0.000193018
## RUNX3  1.13344e-06 0.001885684
## DUSP2  1.52848e-06 0.002432342
## STK17B 1.76313e-06 0.002688844
## CXCR4  2.00116e-05 0.029297784
##
## $`2`
## DataFrame with 24 rows and 2 columns
##             p.value          FDR
##           <numeric>    <numeric>
## DCN    8.49384e-115 3.10883e-110
## COL3A1  6.01375e-62  1.10055e-57
## COL1A2  1.49971e-55  1.82970e-51
## COL1A1  2.98542e-46  2.73173e-42
## COL6A2  1.59059e-43  1.16435e-39
## ...              ...          ...
## CCDC80  2.99446e-12  5.48002e-09
## IGFBP5  1.57052e-11  2.73727e-08
## MT1E    1.48844e-10  2.47629e-07
## TIMP1   2.02627e-10  3.22450e-07
## IGFBP2  4.74961e-08  7.24336e-05
##
## $`3`
## DataFrame with 3 rows and 2 columns
##             p.value         FDR
##           <numeric>   <numeric>
## BTG1    1.47175e-31 5.38676e-27
## PTPRC   9.21383e-20 1.68618e-15
## TNFAIP3 9.19089e-17 1.12132e-12
##
## $`4`
## DataFrame with 18 rows and 2 columns
##              p.value          FDR
##            <numeric>    <numeric>
## SPARCL1 1.82340e-117 6.67381e-113
## VWF      1.68442e-62  3.08258e-58
## TM4SF1   3.78045e-62  4.61228e-58
## EMP1     1.78050e-61  1.62920e-57
## CCL14    2.43741e-61  1.78423e-57
## ...              ...          ...
## RAMP2    9.46412e-18  2.47426e-14
## PLVAP    1.32086e-09  3.22299e-06
## ENG      2.26259e-08  5.17580e-05
## SPRY1    1.30117e-07  2.80143e-04
## THBD     3.90829e-07  7.94708e-04
```

```
##
## $`5`
## DataFrame with 37 rows and 2 columns
##              p.value          FDR
##            <numeric>    <numeric>
## TYROBP  1.07277e-188 3.92644e-184
## HLA-DRA 3.18279e-185 5.82467e-181
## LYZ     2.51258e-168 3.06543e-164
## CD74    1.22738e-140 1.12309e-136
## FCER1G  2.64038e-132 1.93281e-128
## ...               ...          ...
## CYBA     1.64558e-06   0.00182514
## CD14     3.11043e-06   0.00334838
## LGALS1   4.10182e-06   0.00428945
## ZEB2     5.28672e-06   0.00532547
## BCL2A1   5.38353e-06   0.00532547
##
## $`6`
## DataFrame with 36 rows and 2 columns
##             p.value          FDR
##           <numeric>    <numeric>
## CALD1  4.75121e-183 1.73899e-178
## COL4A1 7.09067e-132 1.29763e-127
## COL4A2 1.14298e-109 1.39447e-105
## IGFBP7  4.76291e-97  4.35819e-93
## MYL9    1.47598e-85  1.08045e-81
## ...              ...          ...
## CRIP1   1.37189e-06   0.00156914
## CAV1    2.74813e-06   0.00304801
## IGFBP4  3.29386e-06   0.00354584
## LHFPL6  4.21186e-06   0.00440452
## MT1E    3.11137e-05   0.03163313
##
## $`7`
## DataFrame with 4 rows and 2 columns
##           p.value         FDR
##         <numeric>   <numeric>
## WFDC2 9.23319e-70 3.37944e-65
## SLPI  2.55461e-39 4.67507e-35
## KRT19 5.24629e-08 6.40065e-04
## KRT8  3.24646e-06 2.97059e-02
##
## $`8`
## DataFrame with 9 rows and 2 columns
##               p.value          FDR
##             <numeric>    <numeric>
## CD74     3.35593e-137 1.22831e-132
## HLA-DRA   6.18838e-74  1.13250e-69
## HLA-DPA1  2.10112e-39  2.56343e-35
## HLA-DPB1  5.54857e-35  5.07708e-31
## HLA-DRB1  7.13825e-18  5.22534e-14
## RPS27     9.30265e-10  5.67477e-06
## CD37      1.09252e-09  5.71247e-06
## LAPTM5    4.79383e-09  2.19324e-05
```

```
## IRF8      2.71306e-07  1.10334e-03
##
## $`9`
## DataFrame with 29 rows and 2 columns
##              p.value          FDR
##            <numeric>    <numeric>
## EPCAM   2.09909e-145 7.68287e-141
## ELF3    3.85834e-144 7.06096e-140
## CLDN3   8.50395e-114 1.03751e-109
## SLPI     1.16172e-96  1.06301e-92
## WFDC2    8.76814e-79  6.41846e-75
## ...              ...          ...
## NR2F6    2.54590e-07   0.00037273
## CCN1     2.63319e-06   0.00370682
## SNHG29   6.62834e-06   0.00898533
## S100A2   1.09618e-05   0.01432905
## S100A11  1.40291e-05   0.01770620
##
## $`10`
## DataFrame with 17 rows and 2 columns
##               p.value          FDR
##             <numeric>    <numeric>
## COL1A1   5.64050e-165 2.06448e-160
## DCN      4.70248e-137 8.60577e-133
## COL3A1   4.28452e-123 5.22726e-119
## COL1A2   2.37113e-104 2.16964e-100
## LUM       1.52001e-87  1.11267e-83
## ...               ...          ...
## C11orf96  4.10480e-16  1.15569e-12
## SFRP2     2.13391e-08  5.57880e-05
## MGP       4.08556e-07  9.96904e-04
## TAGLN     1.23722e-06  2.83021e-03
## VCAN      1.01917e-05  2.19428e-02
##
## $`11`
## DataFrame with 7 rows and 2 columns
##             p.value          FDR
##           <numeric>    <numeric>
## MZB1   8.84459e-176 3.23721e-171
## IGHG1  2.90841e-164 5.32254e-160
## IGHG3  2.48325e-113 3.02965e-109
## IGKC    3.90419e-61  3.57243e-57
## SSR4    3.71535e-53  2.71971e-49
## IGHG4   3.12442e-32  1.90595e-28
## JCHAIN  2.71966e-27  1.42203e-23
##
## $`12`
## DataFrame with 6 rows and 2 columns
##              p.value          FDR
##            <numeric>    <numeric>
## CCL5    8.74371e-112 3.20029e-107
## ZFP36L2  8.07355e-20  1.47750e-15
## BTG1     2.05800e-08  2.51083e-04
## TXNIP    1.34801e-07  1.23347e-03
```

```
## DUSP2    5.35001e-07  3.91631e-03
## TNFAIP3  4.74124e-06  2.89223e-02
```

Let's try running WebGestaltR, which runs overrepresentation analysis

```
reference_genes <- rownames(sce)
for (i in 1:length(unique(sce$clusters))) {
  cluster_genes <- sig_genes[[i]]
  pathways <- WebGestaltR(enrichDatabase = "geneontology_Biological_Process",
                                  interestGene = rownames(cluster_genes),
                                  interestGeneType = "genesymbol",
                                  referenceGene = reference_genes,
                                  referenceGeneType = "genesymbol",
                                  isOutput = FALSE)
  pathways$userId <- NULL; pathways$overlapId <- NULL

  print(head(pathways))
}
```

```
## Loading the functional categories...
## Loading the ID list...
## Loading the reference list...
## Performing the enrichment analysis...
##      geneSet                                description
## 1 GO:0042110                           T cell activation
## 2 GO:0051249         regulation of lymphocyte activation
## 3 GO:0002694          regulation of leukocyte activation
## 4 GO:0030217                      T cell differentiation
## 5 GO:0045059           positive thymic T cell selection
## 6 GO:0002696 positive regulation of leukocyte activation
##                                                  link size overlap     expect
## 1 http://amigo.geneontology.org/amigo/term/GO:0042110  450       9 0.60138501
## 2 http://amigo.geneontology.org/amigo/term/GO:0051249  400       8 0.53456445
## 3 http://amigo.geneontology.org/amigo/term/GO:0002694  477       8 0.63746811
## 4 http://amigo.geneontology.org/amigo/term/GO:0030217  230       6 0.30737456
## 5 http://amigo.geneontology.org/amigo/term/GO:0045059   13       3 0.01737334
## 6 http://amigo.geneontology.org/amigo/term/GO:0002696  294       6 0.39290487
##   enrichmentRatio       pValue          FDR
## 1        14.96545 2.854819e-09 1.792826e-05
## 2        14.96545 2.689369e-08 8.444618e-05
## 3        12.54965 1.048295e-07 2.194432e-04
## 4        19.52016 4.312206e-07 6.770163e-04
## 5       172.67832 5.873647e-07 7.377300e-04
## 6        15.27087 1.808369e-06 1.591249e-03
## Loading the functional categories...
## Loading the ID list...
## Loading the reference list...
## Performing the enrichment analysis...
##      geneSet                                                    description
## 1 GO:0030198                                 extracellular matrix organization
## 2 GO:0043062                               extracellular structure organization
## 3 GO:0070208                                       protein heterotrimerization
## 4 GO:0001101                                         response to acid chemical
```

```
## 5 GO:0051216                                                cartilage development
## 6 GO:0060351 cartilage development involved in endochondral bone morphogenesis
##                                                 link size overlap     expect
## 1 http://amigo.geneontology.org/amigo/term/GO:0030198  347      10 0.46373466
## 2 http://amigo.geneontology.org/amigo/term/GO:0043062  400      10 0.53456445
## 3 http://amigo.geneontology.org/amigo/term/GO:0070208   13       4 0.01737334
## 4 http://amigo.geneontology.org/amigo/term/GO:0001101  332       7 0.44368849
## 5 http://amigo.geneontology.org/amigo/term/GO:0051216  203       6 0.27129146
## 6 http://amigo.geneontology.org/amigo/term/GO:0060351   45       4 0.06013850
##   enrichmentRatio       pValue          FDR
## 1        21.56406 7.862822e-12 4.937852e-08
## 2        18.70682 3.198330e-11 1.004276e-07
## 3       230.23776 1.696438e-09 3.551210e-06
## 4        15.77683 1.674731e-07 2.596063e-04
## 5        22.11644 2.066929e-07 2.596063e-04
## 6        66.51313 3.437435e-07 3.597849e-04
## Loading the functional categories...
## Loading the ID list...
## Loading the reference list...
## Performing the enrichment analysis...
##      geneSet                                    description
## 1 GO:0032703 negative regulation of interleukin-2 production
##                                                 link size overlap      expect
## 1 http://amigo.geneontology.org/amigo/term/GO:0032703   23       2 0.004191471
##   enrichmentRatio       pValue        FDR
## 1        477.1594 5.597104e-06 0.03514982
## Loading the functional categories...
## Loading the ID list...
## Loading the reference list...
## Performing the enrichment analysis...
## NULL
## Loading the functional categories...
## Loading the ID list...
## Loading the reference list...
## Performing the enrichment analysis...
##      geneSet
## 1 GO:0036230
## 2 GO:0019886
## 3 GO:0002495
## 4 GO:0002504
## 5 GO:0042119
## 6 GO:0002478
##                                                                                description
## 1                                                                     granulocyte activation
## 2         antigen processing and presentation of exogenous peptide antigen via MHC class II
## 3                   antigen processing and presentation of peptide antigen via MHC class II
## 4 antigen processing and presentation of peptide or polysaccharide antigen via MHC class II
## 5                                                                       neutrophil activation
## 6                          antigen processing and presentation of exogenous peptide antigen
##                                                 link size overlap    expect
## 1 http://amigo.geneontology.org/amigo/term/GO:0036230  497      15 1.1170575
## 2 http://amigo.geneontology.org/amigo/term/GO:0019886   95       9 0.2135221
## 3 http://amigo.geneontology.org/amigo/term/GO:0002495   98       9 0.2202649
## 4 http://amigo.geneontology.org/amigo/term/GO:0002504   99       9 0.2225125
```

```
## 5 http://amigo.geneontology.org/amigo/term/GO:0042119  491      14 1.1035719
## 6 http://amigo.geneontology.org/amigo/term/GO:0002478  119       9 0.2674645
##   enrichmentRatio       pValue          FDR
## 1        13.42814 6.505907e-14 4.085710e-10
## 2        42.15021 5.246914e-13 1.206191e-09
## 3        40.85990 6.994405e-13 1.206191e-09
## 4        40.44717 7.682743e-13 1.206191e-09
## 5        12.68608 1.195377e-12 1.501394e-09
## 6        33.64933 4.162448e-12 4.356696e-09
## Loading the functional categories...
## Loading the ID list...
## Loading the reference list...
## Performing the enrichment analysis...
##      geneSet                              description
## 1 GO:0030198        extracellular matrix organization
## 2 GO:0043062     extracellular structure organization
## 3 GO:0070208              protein heterotrimerization
## 4 GO:0001101                response to acid chemical
## 5 GO:0071230 cellular response to amino acid stimulus
## 6 GO:0061448            connective tissue development
##                                                 link size overlap     expect
## 1 http://amigo.geneontology.org/amigo/term/GO:0030198  347      11 0.71668084
## 2 http://amigo.geneontology.org/amigo/term/GO:0043062  400      11 0.82614506
## 3 http://amigo.geneontology.org/amigo/term/GO:0070208   13       4 0.02684971
## 4 http://amigo.geneontology.org/amigo/term/GO:0001101  332       9 0.68570040
## 5 http://amigo.geneontology.org/amigo/term/GO:0071230   67       5 0.13837930
## 6 http://amigo.geneontology.org/amigo/term/GO:0061448  262       7 0.54112502
##   enrichmentRatio       pValue          FDR
## 1        15.34853 5.782597e-11 3.631471e-07
## 2        13.31485 2.632867e-10 8.267203e-07
## 3       148.97738 1.069879e-08 2.239614e-05
## 4        13.12527 1.666724e-08 2.616757e-05
## 5        36.13257 2.436502e-07 3.060246e-04
## 6        12.93601 8.896126e-07 8.580650e-04
## Loading the functional categories...
## Loading the ID list...
## Loading the reference list...
## Performing the enrichment analysis...
## NULL
## Loading the functional categories...
## Loading the ID list...
## Loading the reference list...
## Performing the enrichment analysis...
##      geneSet
## 1 GO:0060333
## 2 GO:0019886
## 3 GO:0002495
## 4 GO:0002504
## 5 GO:0002478
## 6 GO:0019884
##                                                                                  description
## 1                                                 interferon-gamma-mediated signaling pathway
## 2         antigen processing and presentation of exogenous peptide antigen via MHC class II
## 3                   antigen processing and presentation of peptide antigen via MHC class II
```

```
## 4 antigen processing and presentation of peptide or polysaccharide antigen via MHC class II
## 5                          antigen processing and presentation of exogenous peptide antigen
## 6                                 antigen processing and presentation of exogenous antigen
##                                                 link size overlap     expect
## 1 http://amigo.geneontology.org/amigo/term/GO:0060333   84       5 0.04082129
## 2 http://amigo.geneontology.org/amigo/term/GO:0019886   95       5 0.04616693
## 3 http://amigo.geneontology.org/amigo/term/GO:0002495   98       5 0.04762483
## 4 http://amigo.geneontology.org/amigo/term/GO:0002504   99       5 0.04811080
## 5 http://amigo.geneontology.org/amigo/term/GO:0002478  119       5 0.05783015
## 6 http://amigo.geneontology.org/amigo/term/GO:0019884  127       5 0.06171790
##   enrichmentRatio       pValue          FDR
## 1       122.48512 1.696536e-10 6.156735e-07
## 2       108.30263 3.178716e-10 6.156735e-07
## 3       104.98724 3.724001e-10 6.156735e-07
## 4       103.92677 3.921488e-10 6.156735e-07
## 5        86.46008 9.983150e-10 1.253884e-06
## 6        81.01378 1.387961e-09 1.452732e-06
## Loading the functional categories...
## Loading the ID list...
## Loading the reference list...
## Performing the enrichment analysis...
##      geneSet                                                        description
## 1 GO:0070268                                                      cornification
## 2 GO:0043588                                                  skin development
## 3 GO:0008544                                               epidermis development
## 4 GO:0030216                                         keratinocyte differentiation
## 5 GO:0060706 cell differentiation involved in embryonic placenta development
## 6 GO:0001892                                        embryonic placenta development
##                                                   link size overlap     expect
## 1 http://amigo.geneontology.org/amigo/term/GO:0070268  110       5 0.17373345
## 2 http://amigo.geneontology.org/amigo/term/GO:0043588  407       7 0.64281375
## 3 http://amigo.geneontology.org/amigo/term/GO:0008544  453       7 0.71546592
## 4 http://amigo.geneontology.org/amigo/term/GO:0030216  293       6 0.46276273
## 5 http://amigo.geneontology.org/amigo/term/GO:0060706   26       3 0.04106427
## 6 http://amigo.geneontology.org/amigo/term/GO:0001892   89       4 0.14056615
##   enrichmentRatio       pValue         FDR
## 1       28.779720 7.150010e-07 0.004490206
## 2       10.889624 2.349518e-06 0.007377485
## 3        9.783834 4.767484e-06 0.008089192
## 4       12.965608 5.152351e-06 0.008089192
## 5       73.056213 8.876821e-06 0.011149288
## 6       28.456353 1.089564e-05 0.011404100
## Loading the functional categories...
## Loading the ID list...
## Loading the reference list...
## Performing the enrichment analysis...
##      geneSet                           description
## 1 GO:0043062  extracellular structure organization
## 2 GO:0030198     extracellular matrix organization
## 3 GO:0030199          collagen fibril organization
## 4 GO:0051216                 cartilage development
## 5 GO:0061448         connective tissue development
## 6 GO:0030208 dermatan sulfate biosynthetic process
##                                                 link size overlap     expect
```

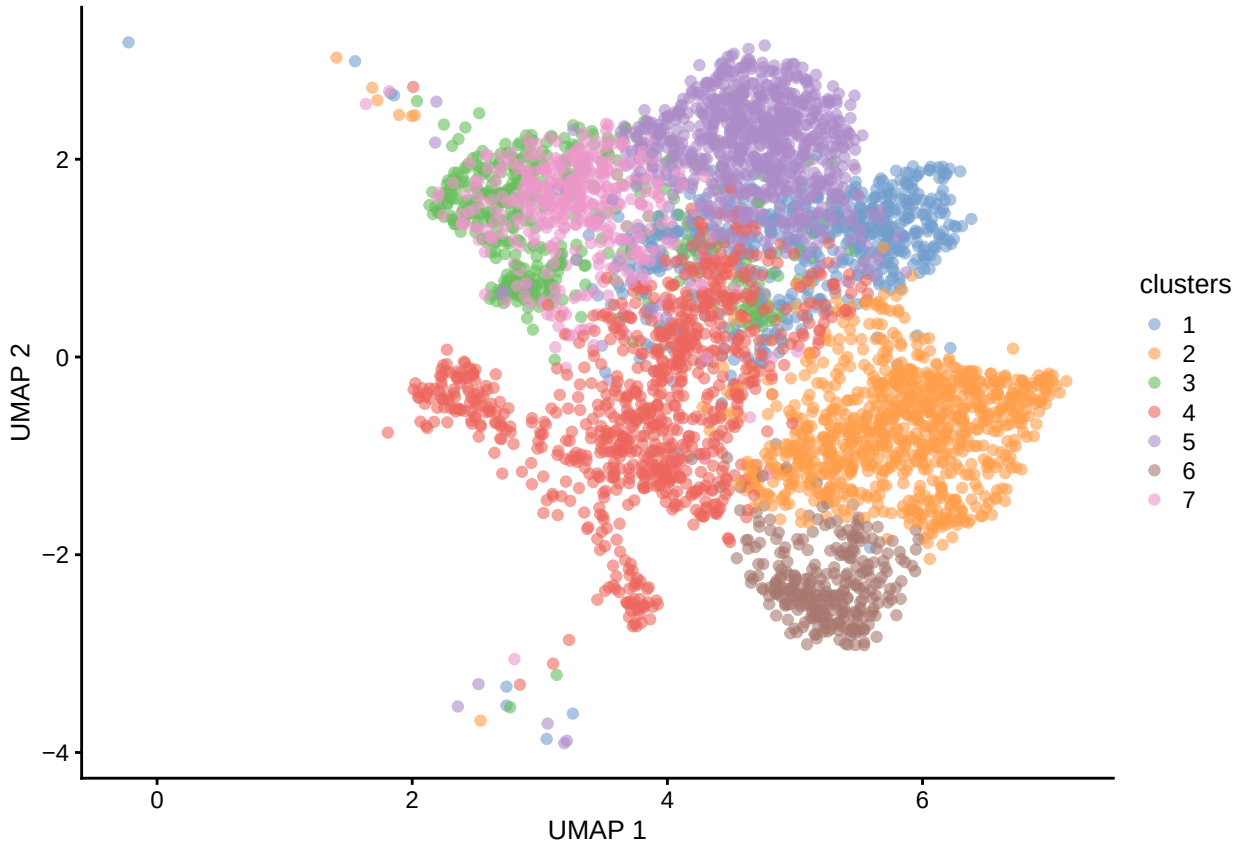
```
## 1 http://amigo.geneontology.org/amigo/term/GO:0043062  400      12 0.38877415
## 2 http://amigo.geneontology.org/amigo/term/GO:0030198  347      12 0.33726157
## 3 http://amigo.geneontology.org/amigo/term/GO:0030199   49       6 0.04762483
## 4 http://amigo.geneontology.org/amigo/term/GO:0051216  203       6 0.19730288
## 5 http://amigo.geneontology.org/amigo/term/GO:0061448  262       6 0.25464707
## 6 http://amigo.geneontology.org/amigo/term/GO:0030208   12       3 0.01166322
##   enrichmentRatio       pValue          FDR
## 1        30.86625 0.000000e+00 0.000000e+00
## 2        35.58069 0.000000e+00 0.000000e+00
## 3       125.98469 3.965051e-12 8.300172e-09
## 4        30.41010 2.359693e-08 3.704718e-05
## 5        23.56202 1.075359e-07 1.350651e-04
## 6       257.21875 1.648455e-07 1.494446e-04
## Loading the functional categories...
## Loading the ID list...
## Loading the reference list...
## Performing the enrichment analysis...
## NULL
## Loading the functional categories...
## Loading the ID list...
## Loading the reference list...
## Performing the enrichment analysis...
## NULL
```

After some grueling googling, I have a tentative identification for most clusters. I'm going to try breaking the larger groups up by these definitions and seeing what happens.

Fibroblasts

```
#fibroblasts <- sce[, sce$clusters %in% c(2, 6, 10)]
#g <- buildSNNGraph(fibroblasts,
#                   k = 50,   # higher = bigger clusters
#                   BNPARAM = BiocNeighbors::AnnoyParam(),
#                   BPPARAM = BiocParallel::MulticoreParam())
#clusters <- as.factor(igraph::cluster_louvain(g)$membership)
#fibroblasts$clusters <- clusters
#saveRDS(fibroblasts,  file = "../../data/sce_objects/fibroblasts.rds")
fibroblasts <- readRDS("../../data/sce_objects/fibroblasts.rds")

plotUMAP(fibroblasts, colour_by = "clusters")
```




```
markers <- findMarkers(fibroblasts,  groups = fibroblasts$clusters,
                       block = fibroblasts$Pool,  # use to get within-donor DE
                       direction = "up",  lfc = 1.5,
                       pval.type = "some",
                       BPPARAM = BiocParallel::MulticoreParam())

sig_genes <- getOnlySignificantGenes(markers)
sig_genes
```

```
## $`1`
## DataFrame with 3 rows and 2 columns
##             p.value         FDR
##           <numeric>   <numeric>
## COLEC11 8.74195e-19 3.19964e-14
## C7      2.15308e-14 3.94024e-10
## RBP1    3.40226e-08 4.15087e-04
##
## $`2`
## DataFrame with 9 rows and 2 columns
##               p.value         FDR
##             <numeric>   <numeric>
## COL4A1   2.37254e-100 8.68372e-96
## COL4A2    6.62114e-68 1.21170e-63
## RGS5      6.13691e-49 7.48723e-45
## NDUFA4L2  3.38894e-21 3.10097e-17
```

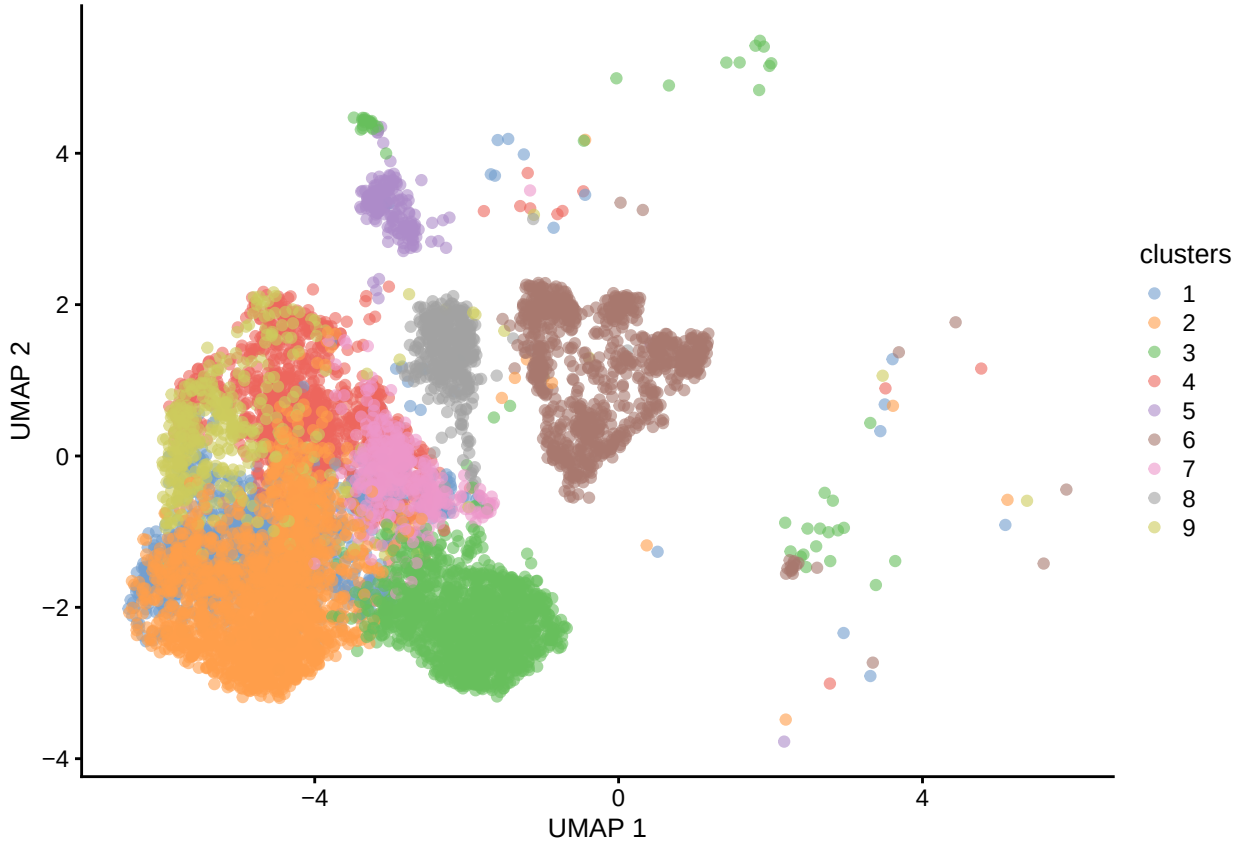
```
## COL18A1   4.98053e-21 3.64585e-17
## CRIP1     6.26339e-13 3.82077e-09
## PPP1R14A  8.53192e-10 4.46110e-06
## MCAM      1.25916e-08 5.76083e-05
## CCDC102B  8.50051e-07 3.45697e-03
##
## $`3`
## DataFrame with 11 rows and 2 columns
##             p.value         FDR
##           <numeric>   <numeric>
## MT1X    8.69440e-37 3.18224e-32
## SOD2    2.53891e-17 4.64633e-13
## IGFBP6  6.25056e-17 7.62590e-13
## C3      4.57632e-15 4.18745e-11
## MT1M    3.36786e-14 2.46534e-10
## DCN     1.82187e-12 1.11137e-08
## MT1E    1.44288e-10 7.54440e-07
## S100A10 9.44412e-10 4.32080e-06
## MT2A    4.74693e-08 1.93047e-04
## CCDC80  4.09392e-07 1.49841e-03
## FBLN1   5.43818e-06 1.80948e-02
##
## $`4`
## DataFrame with 1 row and 2 columns
##           p.value        FDR
##         <numeric>  <numeric>
## NEAT1 6.57008e-08 0.00240471
##
## $`5`
## DataFrame with 3 rows and 2 columns
##             p.value         FDR
##           <numeric>   <numeric>
## CTHRC1  2.35884e-18 8.63360e-14
## RARRES2 1.18089e-14 2.16108e-10
## COL3A1  1.28519e-09 1.56797e-05
##
## $`6`
## DataFrame with 16 rows and 2 columns
##             p.value         FDR
##           <numeric>   <numeric>
## MT1M   1.37886e-101 5.04677e-97
## ADIRF   3.10126e-77 5.67546e-73
## MT2A    1.41018e-48 1.72047e-44
## MUSTN1  2.81540e-38 2.57617e-34
## CRIP1   2.36361e-37 1.73021e-33
## ...              ...         ...
## MYL9    5.24246e-13 1.59899e-09
## TAGLN   1.75203e-11 4.93277e-08
## TPM2    1.58899e-06 4.15420e-03
## ACTA2   2.24858e-06 5.48669e-03
## DSTN    4.83676e-06 1.10644e-02
##
## $`7`
## DataFrame with 3 rows and 2 columns
```

```
##         p.value         FDR
##       <numeric>   <numeric>
## C7  2.12234e-30 7.76796e-26
## DCN 3.73127e-22 6.82841e-18
## MGP 4.24642e-17 5.18078e-13
```

Immune cells

```
# Looking at the somewhat enigmatic immune clusters (5 was clearly macrophages)
#immune <- sce[, sce$clusters %in% c(1, 3, 8, 11, 12)]
#g <- buildSNNGraph(immune,
#                   k = 50,   # higher = bigger clusters
#                   BNPARAM = BiocNeighbors::AnnoyParam(),
#                   BPPARAM = BiocParallel::MulticoreParam())
#clusters <- as.factor(igraph::cluster_louvain(g)$membership)
#immune$clusters <- clusters
#saveRDS(immune,  file = "../../data/sce_objects/immune.rds")
immune <- readRDS("../../data/sce_objects/immune.rds")

plotUMAP(immune, colour_by = "clusters")
```




```
markers <- findMarkers(immune,  groups = immune$clusters,
                       block = immune$Pool,  # use to get within-donor DE
                       direction = "up",  lfc = 1.5,
```

```
                      pval.type = "some",
                      BPPARAM = BiocParallel::MulticoreParam())

sig_genes <- getOnlySignificantGenes(markers)
sig_genes
```

```
## $`1`
## DataFrame with 0 rows and 2 columns
##
## $`2`
## DataFrame with 1 row and 2 columns
##          p.value         FDR
##        <numeric>   <numeric>
## CCL5 2.62811e-44 9.61915e-40
##
## $`3`
## DataFrame with 0 rows and 2 columns
##
## $`4`
## DataFrame with 8 rows and 2 columns
##            p.value         FDR
##          <numeric>   <numeric>
## GZMA   4.26431e-18 1.56078e-13
## IL32   1.29107e-14 2.36272e-10
## GZMB   3.78832e-13 4.62187e-09
## CCL5   5.96590e-10 5.45895e-06
## NKG7   4.37510e-09 3.20266e-05
## CD3D   8.33477e-09 5.08435e-05
## GNLY   1.15012e-07 6.01365e-04
## TMSB4X 3.38064e-06 1.54669e-02
##
## $`5`
## DataFrame with 25 rows and 2 columns
##                p.value         FDR
##              <numeric>   <numeric>
## GZMB       3.45984e-59 1.26634e-54
## CD74       5.01912e-47 9.18524e-43
## IRF4       1.62593e-46 1.98369e-42
## PPP1R14B   1.18850e-30 1.08751e-26
## HLA-DRA    1.65231e-29 1.20953e-25
## ...                ...         ...
## HLA-DPA1   5.23774e-06  0.00912888
## AC239799.2 7.41812e-06  0.01234139
## SERPINF1   1.39923e-05  0.02226660
## C12orf75   1.89952e-05  0.02896843
## PLAC8      2.84813e-05  0.04169774
##
## $`6`
## DataFrame with 7 rows and 2 columns
##             p.value          FDR
##           <numeric>    <numeric>
## IGHG1  6.73640e-116 2.46559e-111
## MZB1   2.53546e-100  4.64002e-96
```

```
## IGHG3   7.67538e-85  9.36422e-81
## IGKC    1.93783e-55  1.77317e-51
## SSR4    7.08594e-30  5.18705e-26
## IGHG4   5.10438e-22  3.11376e-18
## JCHAIN  2.97353e-18  1.55477e-14
##
## $`7`
## DataFrame with 7 rows and 2 columns
##            p.value         FDR
##          <numeric>   <numeric>
## GNLY   6.21018e-63 2.27299e-58
## AREG   5.50735e-22 1.00787e-17
## TYROBP 1.64916e-17 2.01203e-13
## CTSW   1.52240e-14 1.39304e-10
## CCL5   2.23730e-13 1.63775e-09
## XCL1   3.16699e-09 1.93192e-05
## NKG7   4.86510e-07 2.54382e-03
##
## $`8`
## DataFrame with 4 rows and 2 columns
##              p.value         FDR
##            <numeric>   <numeric>
## HLA-DRA  1.82607e-78 6.68358e-74
## CD74     9.64164e-59 1.76447e-54
## HLA-DPB1 2.49199e-11 3.04031e-07
## HLA-DPA1 1.37252e-08 1.25589e-04
##
## $`9`
## DataFrame with 3 rows and 2 columns
##          p.value         FDR
##        <numeric>   <numeric>
## IL32 6.48344e-43 2.37300e-38
## CD3D 3.31435e-23 6.06543e-19
## TRAC 1.11243e-06 1.35720e-02
```

APCs

```
apcs <- immune[, immune$clusters %in% c(5, 8)]
markers <- findMarkers(apcs,  groups = apcs$clusters,
                       block = apcs$Pool,  # use to get within-donor DE
                       direction = "up",  lfc = 1.5,
                       pval.type = "all",
                       BPPARAM = BiocParallel::MulticoreParam())

sig_genes <- getOnlySignificantGenes(markers)
sig_genes
```

```
## $`1`
## DataFrame with 0 rows and 2 columns
##
## $`2`
## DataFrame with 0 rows and 2 columns
```

```
##
## $`3`
## DataFrame with 0 rows and 2 columns
##
## $`4`
## DataFrame with 0 rows and 2 columns
##
## $`5`
## DataFrame with 21 rows and 2 columns
##                p.value         FDR
##              <numeric>   <numeric>
## GZMB       4.41406e-86 1.61559e-81
## IRF4       3.31392e-54 6.06464e-50
## SEC61B     1.67945e-37 2.04899e-33
## PPP1R14B   2.94534e-29 2.69506e-25
## CST3       9.09055e-29 6.65446e-25
## ...                ...         ...
## ALOX5AP    9.76827e-07  0.00210311
## MCL1       2.29066e-06  0.00462270
## AC007952.4 2.39969e-06  0.00462270
## SERPINF1   2.79846e-06  0.00512132
## PLD4       7.71531e-06  0.01344705
##
## $`6`
## DataFrame with 0 rows and 2 columns
##
## $`7`
## DataFrame with 0 rows and 2 columns
##
## $`8`
## DataFrame with 4 rows and 2 columns
##           p.value         FDR
##         <numeric>   <numeric>
## TXNIP 1.42341e-30 5.20982e-26
## KLF2  4.02113e-13 7.35886e-09
## BTG1  3.27843e-10 3.99979e-06
## MS4A1 3.82683e-09 3.50165e-05
##
## $`9`
## DataFrame with 0 rows and 2 columns
```

T cells

```
tcells <- sce[, sce$clusters %in% c(1, 3, 12)]
markers <- findMarkers(tcells,  groups = tcells$clusters,
                       block = tcells$Pool,  # use to get within-donor DE
                       direction = "up",  lfc = 1.5,
                       pval.type = "some",
                       BPPARAM = BiocParallel::MulticoreParam())
sig_genes <- getOnlySignificantGenes(markers)
sig_genes
```

```
## $`1`
## DataFrame with 105 rows and 2 columns
##            p.value         FDR
##          <numeric>   <numeric>
## RPS15A 1.21737e-80 4.45568e-76
## RPL39  2.29503e-78 4.20002e-74
## RPS14  2.56910e-77 3.06761e-73
## HCST   3.35249e-77 3.06761e-73
## FAU    5.19321e-73 3.80153e-69
## ...             ...         ...
## RACK1  1.64308e-07 5.95428e-05
## ACTB   1.11065e-06 3.98537e-04
## EIF1   1.14461e-06 4.06737e-04
## MYL12A 6.37933e-06 2.24509e-03
## IFITM2 4.07988e-05 1.42217e-02
##
## $`2`
## DataFrame with 0 rows and 2 columns
##
## $`3`
## DataFrame with 0 rows and 2 columns
##
## $`4`
## DataFrame with 0 rows and 2 columns
##
## $`5`
## DataFrame with 0 rows and 2 columns
##
## $`6`
## DataFrame with 0 rows and 2 columns
##
## $`7`
## DataFrame with 0 rows and 2 columns
##
## $`8`
## DataFrame with 0 rows and 2 columns
##
## $`9`
## DataFrame with 0 rows and 2 columns
##
## $`10`
## DataFrame with 0 rows and 2 columns
##
## $`11`
## DataFrame with 0 rows and 2 columns
##
## $`12`
## DataFrame with 72 rows and 2 columns
##            p.value         FDR
##          <numeric>   <numeric>
## RPS15A 1.30953e-80 4.79299e-76
## RPS14  1.22030e-74 2.23321e-70
## RPS3   1.68192e-68 2.05199e-64
## RPS27  3.57607e-67 3.27219e-63
```

```
## RPS12  1.36037e-66 9.95821e-63
## ...            ...         ...
## RPL8   2.18571e-07 0.000117646
## RPL37A 7.16660e-07 0.000380152
## RPS20  5.23538e-06 0.002737430
## CCL5   6.23696e-06 0.003215196
## RPL27  1.71389e-05 0.008712511
```