Carson Lab/Brooke scRNA-seq

Jami Shaffer (10X Cell and Library Prep), Bradley W. Blaser (Data Analysis)

4/8/2020

Samples

8 samples from 4 patients at c1 dminus 7 and c1 dplus 1 uwere provided for scRNA sequencing. Single cell isolation was performed using the 10X single cell 5' gene expression and TCR profiling kit. Sequencing data were processed using cellranger v3.1.0 with subsequent analysis in Monocle 3 (Trapnell C. et. al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol. 32, 381-386 (2014)). The 10X qc reports showed the read count per cell and percent of reads in barcoded cells met or exceeded our minimum thresholds (20,000 reads per cell(one sample was slightly low at 18,340 reads per cell but not a big deal), 70% of reads in cells). So all 8 samples can be used for the analysis.

In addition to the 8 samples, publicly available data from a reference set of 8258 PBMC from a healthy donor were incorporated to aid cluster/partition assignment (more on that later).

We targeted recovery of 4000 cells per sample and obtained:

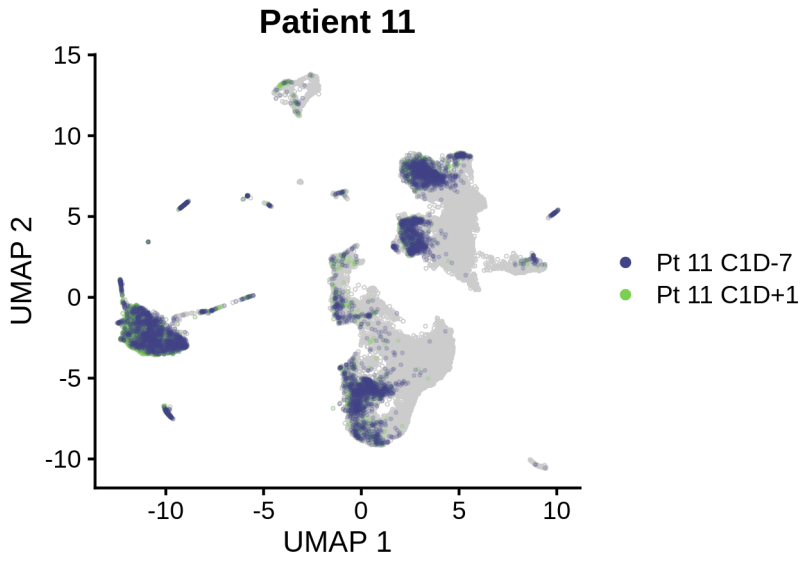
```
sample_counts %>% select(pt_cycle_day,n_cells,running_total)
```

```
## # A tibble: 8 x 3
##   pt_cycle_day n_cells running_total
##   <chr>          <int>         <int>
## 1 11_1_minus7     5164          5164
## 2 11_1_plus1      4251          9415
## 3 15_1_minus7     3535         12950
## 4 15_1_plus1      4765         17715
## 5 17_1_minus7     6827         24542
## 6 17_1_plus1      3368         27910
## 7 22_1_minus7     4957         32867
## 8 22_1_plus1      4972         37839
```

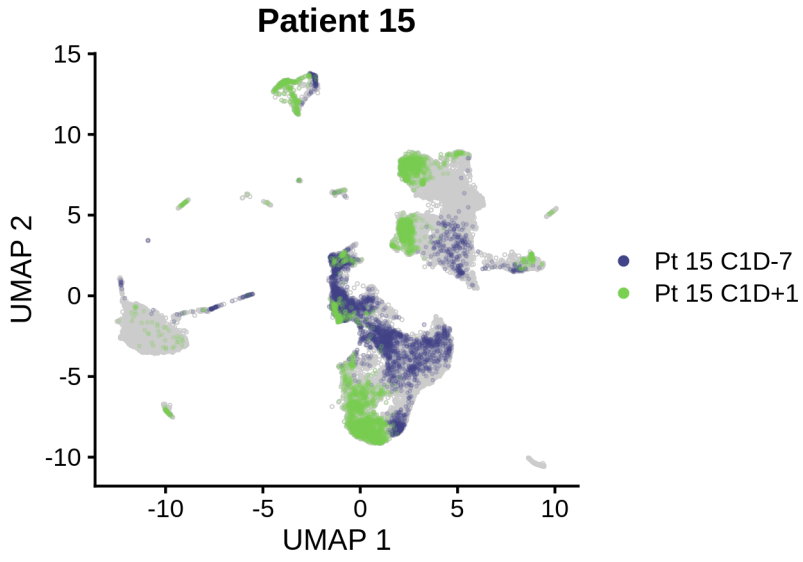
Patient to patient variability was reduced using BATCHelor and aligning coordinates by the "patient" variable (Haghverdi, L. et. al. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat. Biotechnol. 36, 421-427 (2018). Aligned UMAP coordinates were calculated for 37839 patient cells and plotted by patient (2 timepoints each):

```
pt_varplots
```

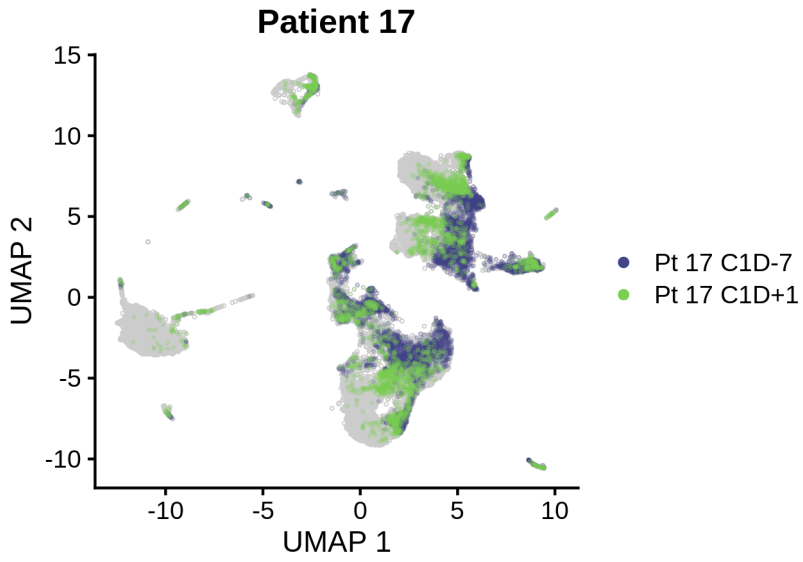
```
## [[1]]
```


Patient 11
UMAP 2
UMAP 1
15
10
5
0
-5
-10
-10
-5
0
5
10
Pt 11 C1D-7
Pt 11 C1D+1

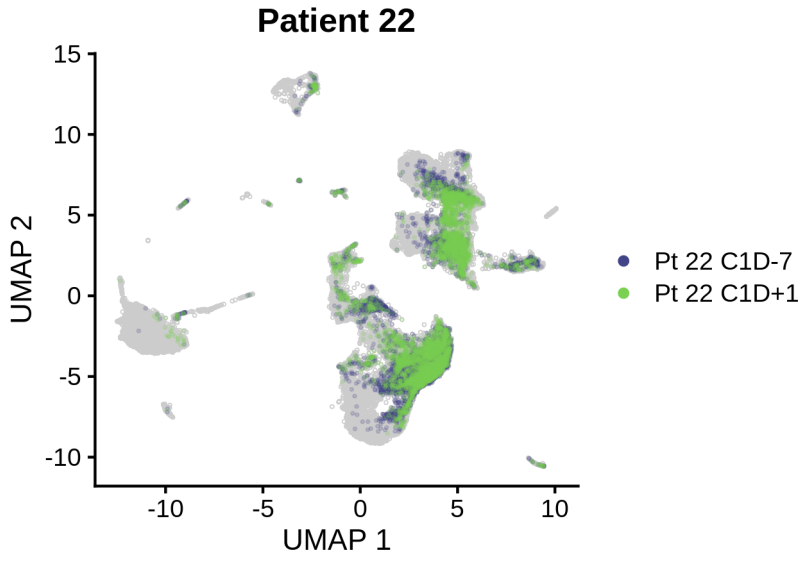
```
##
## [[2]]
```


Patient 15
UMAP 2
UMAP 1
15
10
5
0
-5
-10
-10
-5
0
5
10
Pt 15 C1D-7
Pt 15 C1D+1

```
##
## [[3]]
```



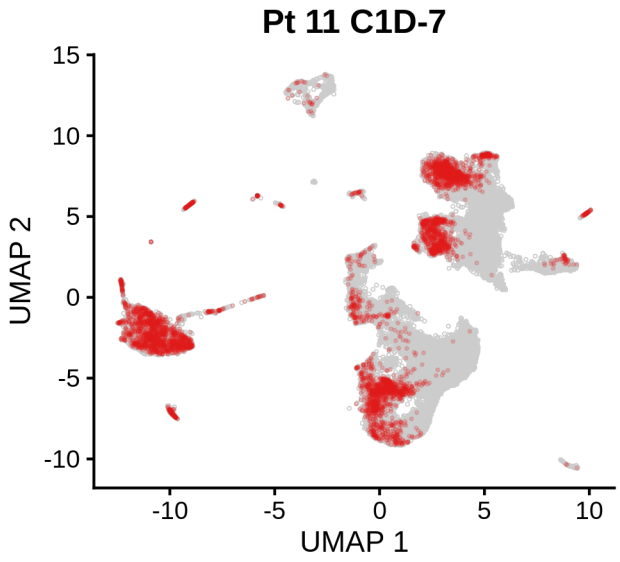

```
##
## [[4]]
```



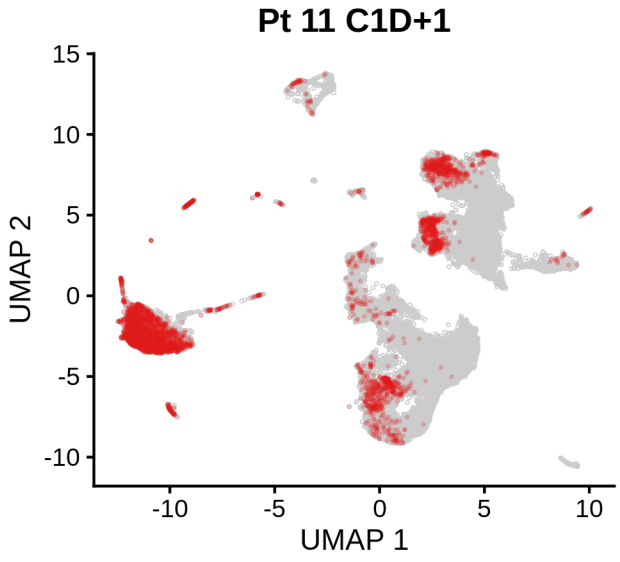
There is still some overlap of the samples so here they are plotted separately:

```
sample_varplots
```

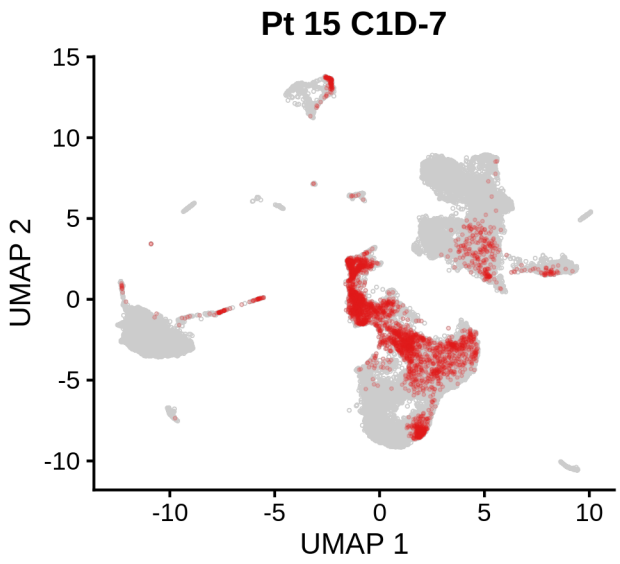
```
## [[1]]
```


Pt 11 C1D-7
UMAP 2
UMAP 1
15
10
5
0
-5
-10
-10
-5
0
5
10

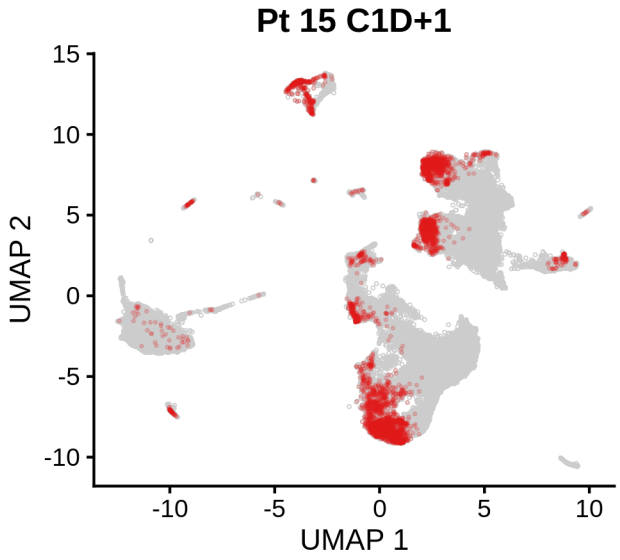
```
##
## [[2]]
```


Pt 11 C1D+1
UMAP 2
UMAP 1
15
10
5
0
-5
-10
-10
-5
0
5
10

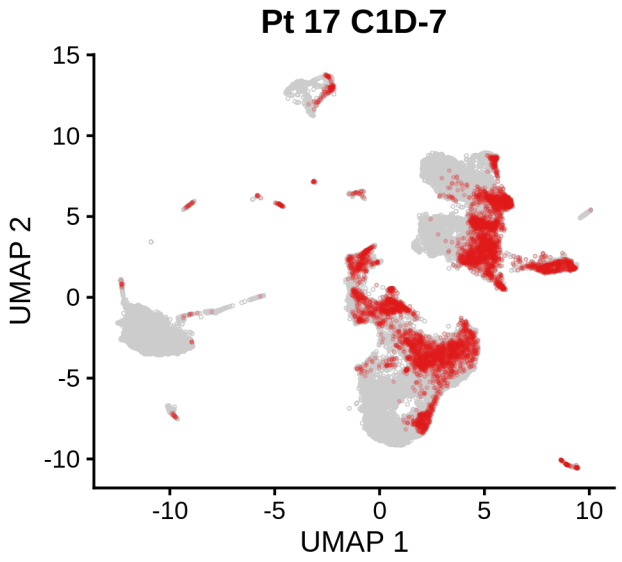
```
##
## [[3]]
```


Pt 15 C1D-7
UMAP 2
UMAP 1
15
10
5
0
-5
-10
-10
-5
0
5
10

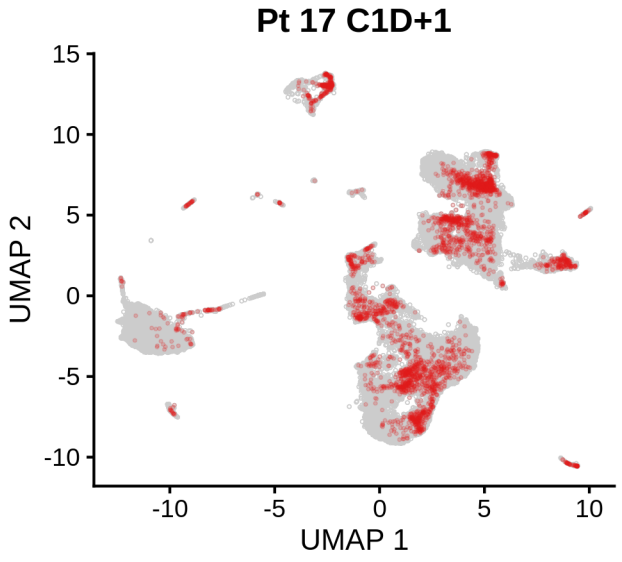
```
##
## [[4]]
```


Pt 15 C1D+1
UMAP 2
UMAP 1
15
10
5
0
-5
-10
-10
-5
0
5
10

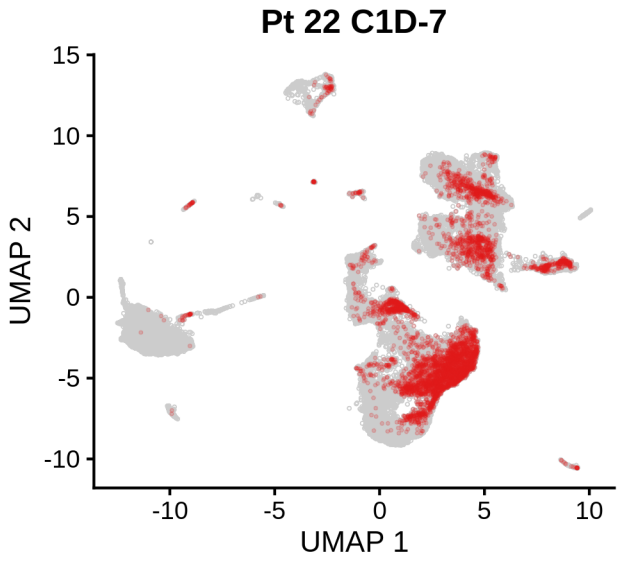
```
##
## [[5]]
```


Pt 17 C1D-7
UMAP 2
UMAP 1
15
10
5
0
-5
-10
-10
-5
0
5
10

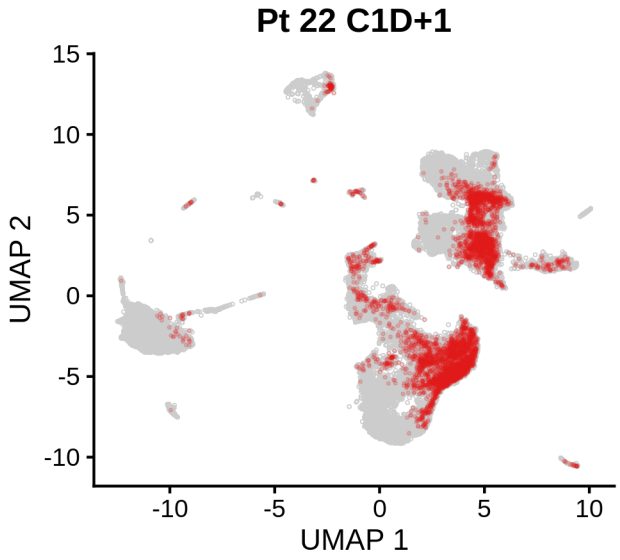
```
##
## [[6]]
```


Pt 17 C1D+1
UMAP 2
UMAP 1
15
10
5
0
-5
-10
-10
-5
0
5
10

```
##
## [[7]]
```

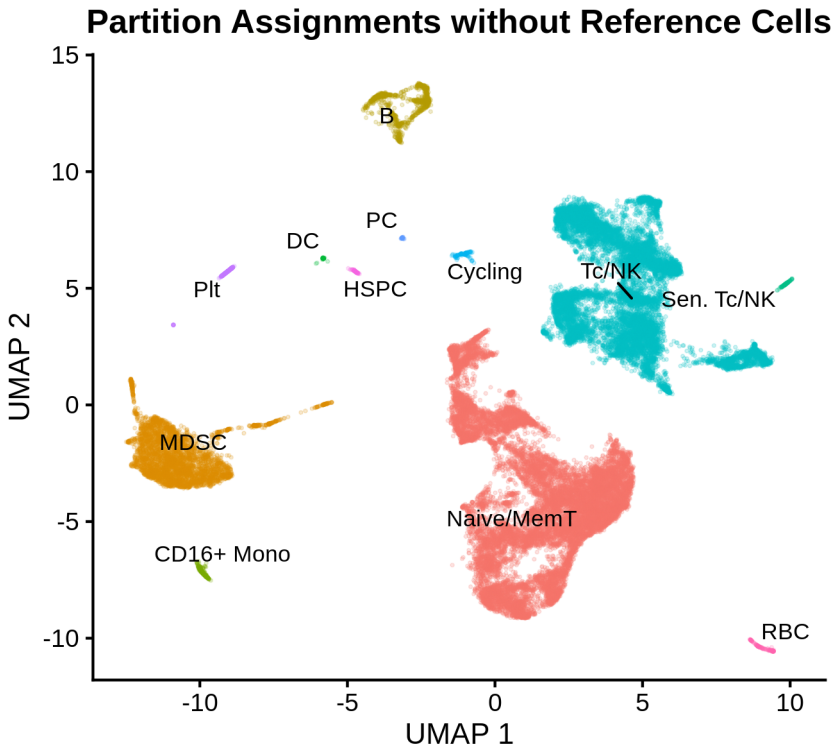



```
##
## [[8]]
```



It would help to know what these cells are, so we can perform clustering using 2 methods. The one we will talk about here is called "partitioning". This is a higher-level clustering method so you have fewer clusters to work with at the beginning and can drill down later if you like. 12 partitions were identified and their top specific markers were identified using Jensen-Shannon specificities and logistic regression (see "data_out/cluster_markers.csv"). Based on these markers we can make the following partition assignments (note this plot excludes the reference cells):

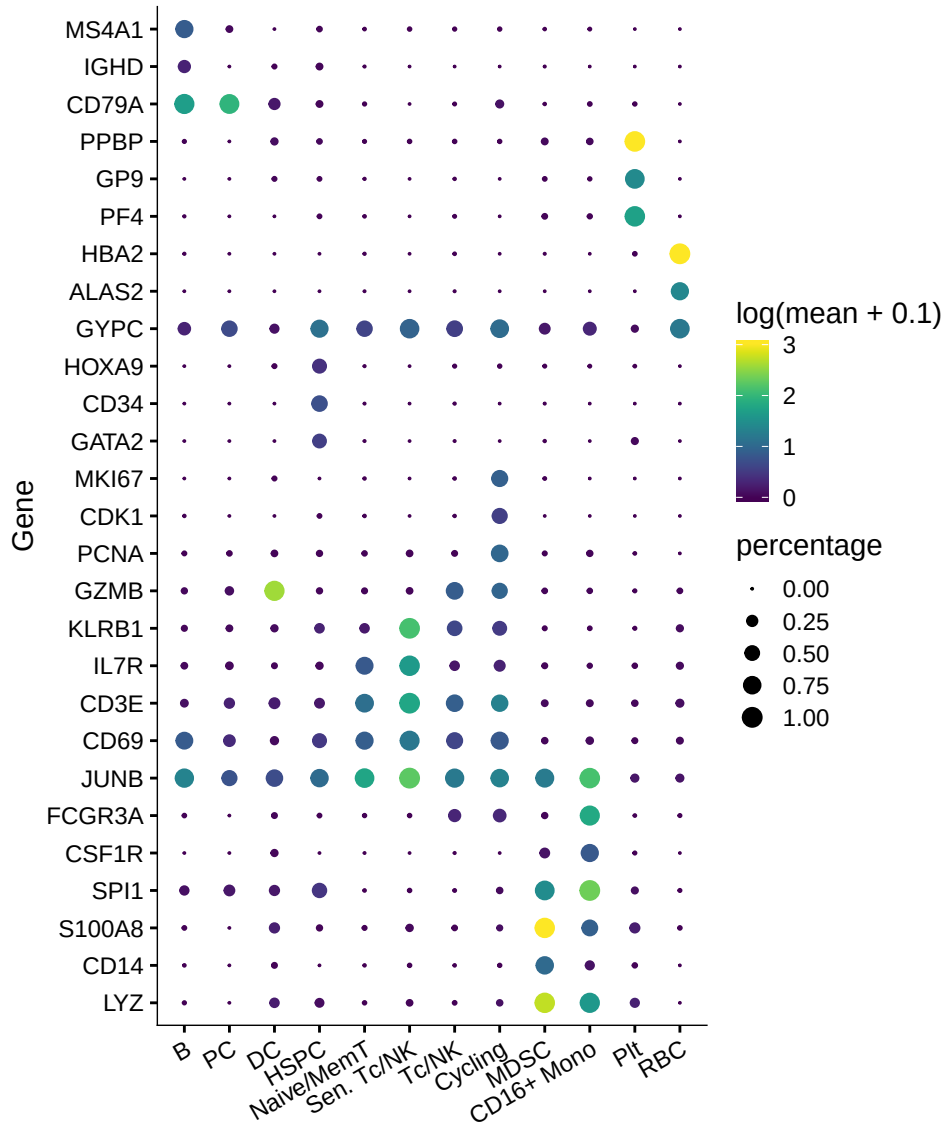
```
plot(partition_assignment_plot_noref)
```



Sen. stands for senescent. Cycling cells have things like cyclins, PCNA, Ki-67 as top markers. There was an interesting cluster of cells that included things like CD34, GATA2 and HOXA9 which I could label nothing other than HSPCs.

To visualize expression of some of these important markers and validate the partition assignments, I generated a plot of several marker genes where the size of the circle represents the percent of cells in a partition expressing the marker and the color scale represents log10 expression of the marker (this plot excludes reference cells):

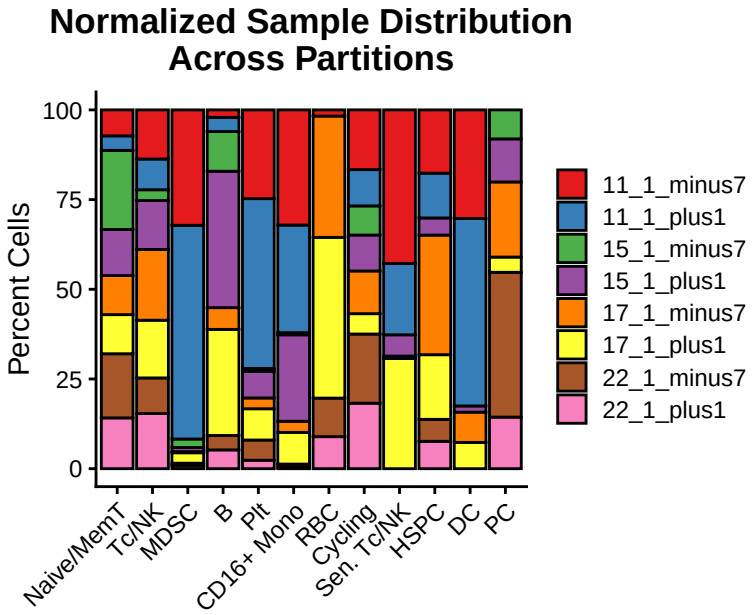
```
plot(marker_gene_dotplot2)
```



MDSC

This brings us to kind of a tough spot in the analysis. Almost all of the MDSCs were identified in patient 11. Here is the distribution of cells in each partition according to sample. (Cell numbers have been normalized here to account for the different number of cells recovered from each sample.)

```
plot(partition_distribution)
```

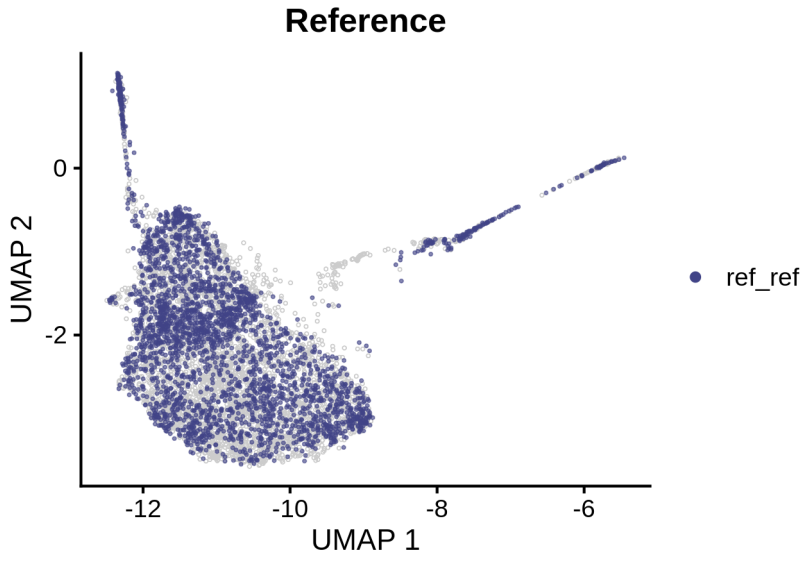


If you look at the third column from the left you can see that probably 95% of the mdsc come from patient 11. Worse yet it looks like there are more after ibrutinib treatment.

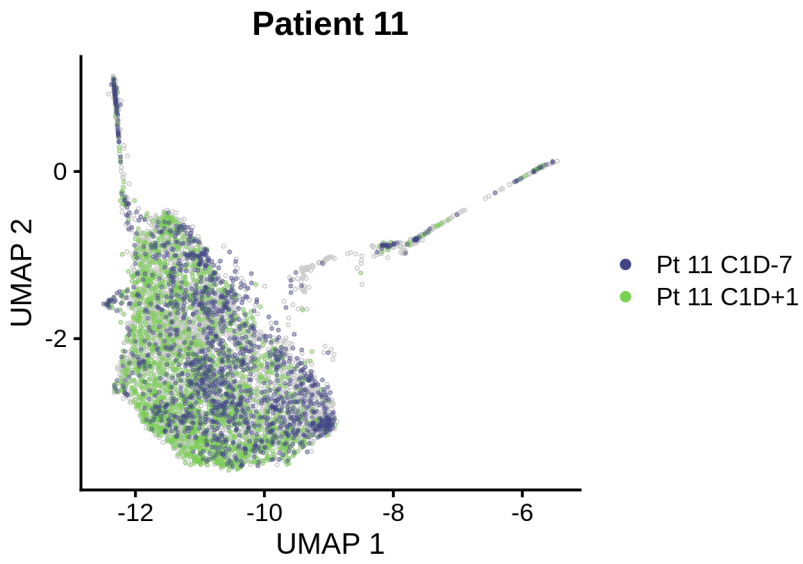
Here are the individual patient plots focusing on MDSC:

```
mdsc_plots
```

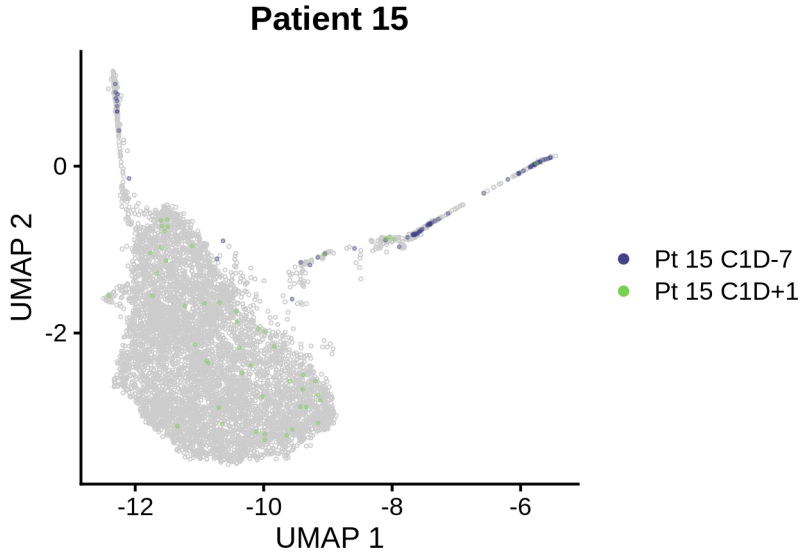
```
## [[1]]
```



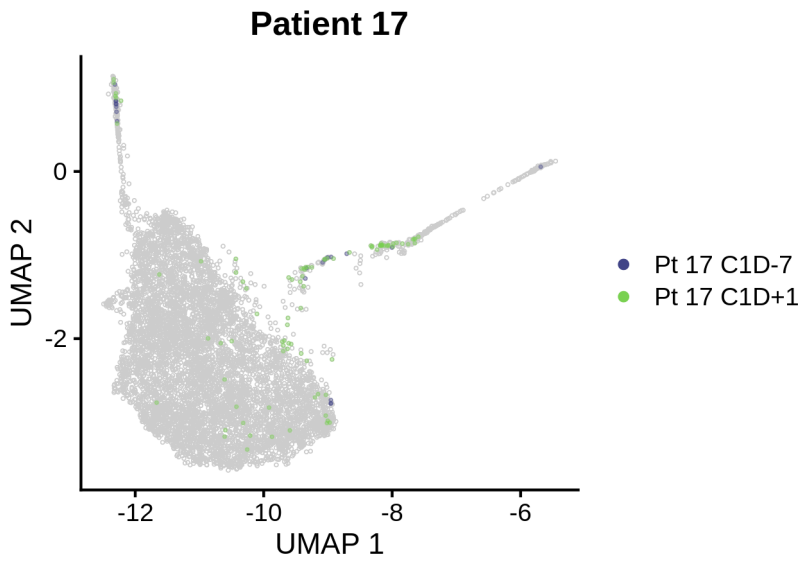

```
##
## [[2]]
```


Patient 11
UMAP 2
0
-2
-12
-10
-8
-6
UMAP 1
Pt 11 C1D-7
Pt 11 C1D+1

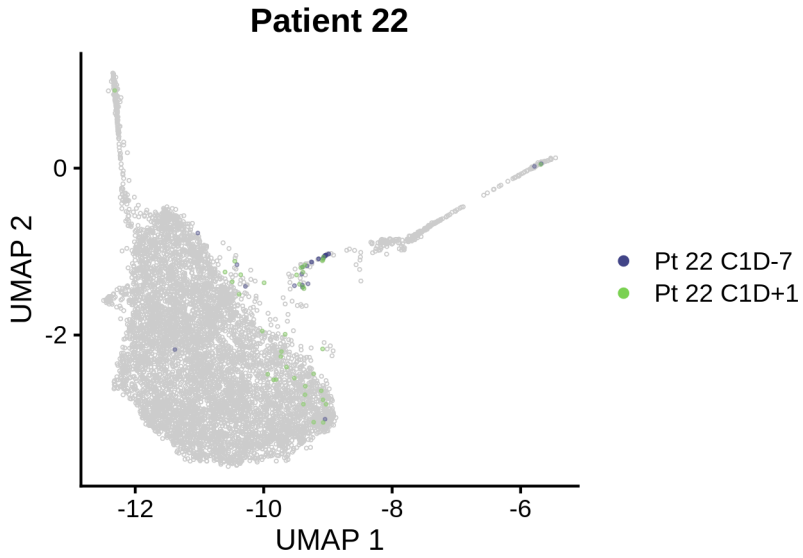
```
##
## [[3]]
```


Patient 15
UMAP 2
0
-2
-12
-10
-8
-6
UMAP 1
Pt 15 C1D-7
Pt 15 C1D+1

```
##
## [[4]]
```



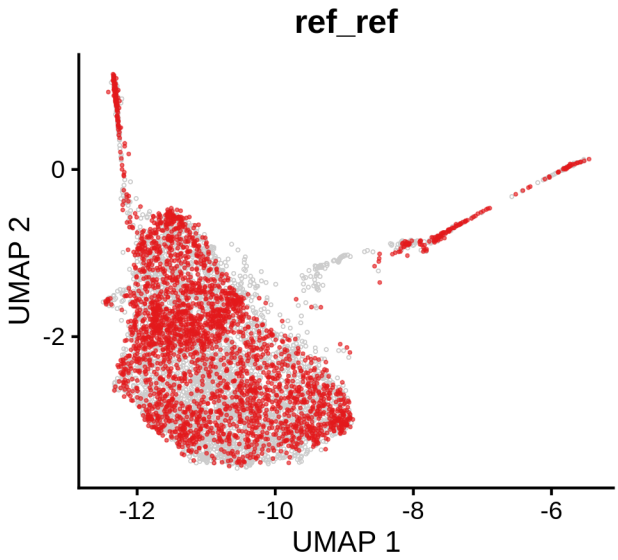

```
##
## [[5]]
```



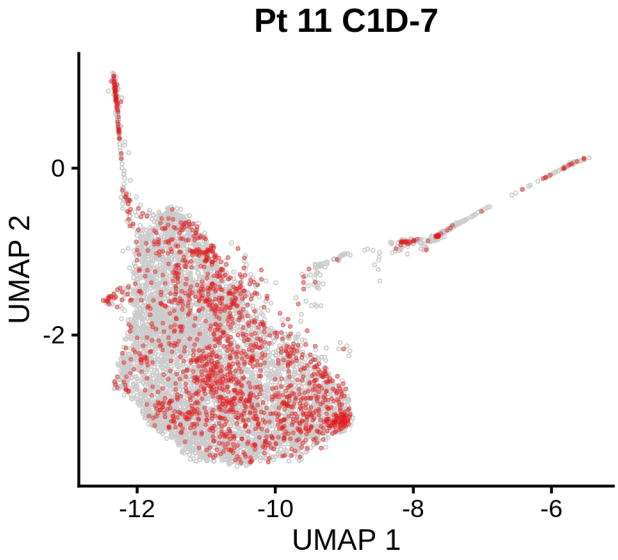
Plotted singly:

```
mdsc_varplots
```

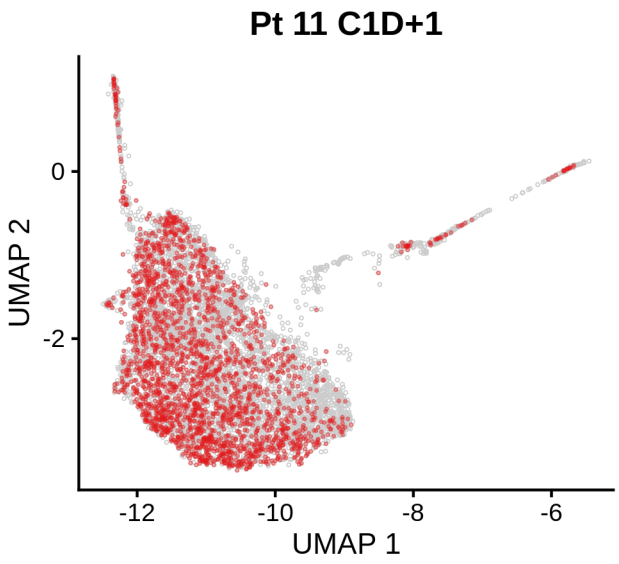
```
## [[1]]
```


ref_ref
0
-2
UMAP 2
-12
-10
-8
-6
UMAP 1

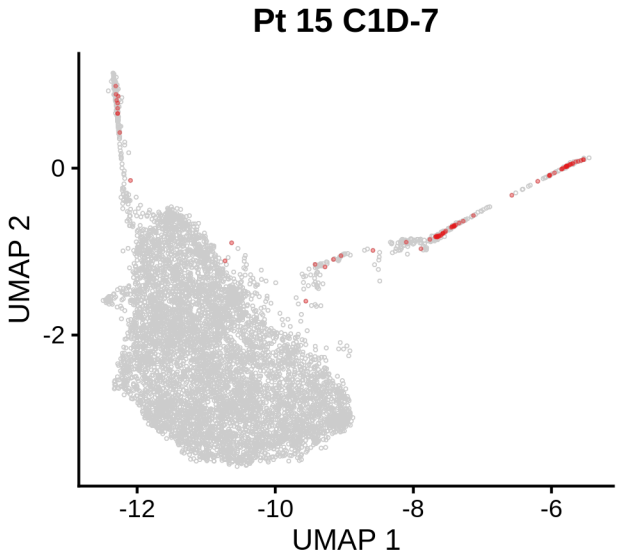
```
##
## [[2]]
```


Pt 11 C1D-7
0
-2
UMAP 2
-12
-10
-8
-6
UMAP 1

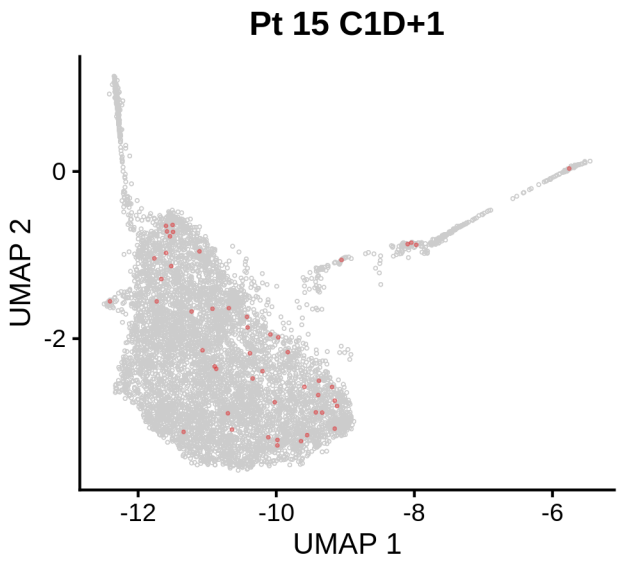
```
##
## [[3]]
```


Pt 11 C1D+1
UMAP 2
0
-2
-12
-10
-8
-6
UMAP 1

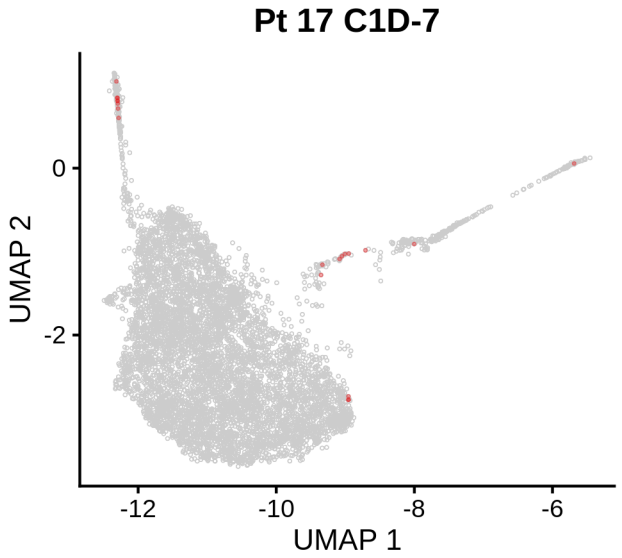
```
##
## [[4]]
```


Pt 15 C1D-7
UMAP 2
0
-2
-12
-10
-8
-6
UMAP 1

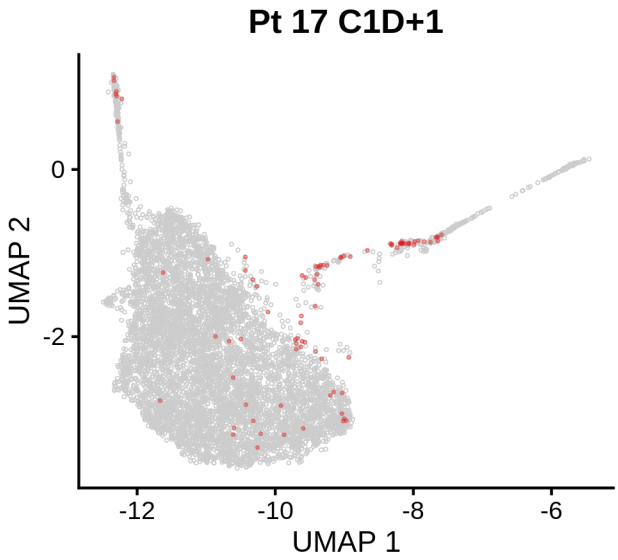
```
##
## [[5]]
```


Pt 15 C1D+1
UMAP 2
0
-2
-12
-10
-8
-6
UMAP 1

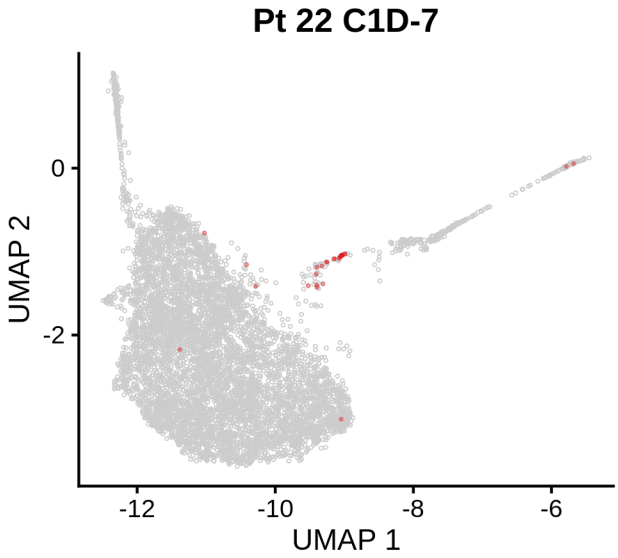
```
##
## [[6]]
```


Pt 17 C1D-7
UMAP 2
0
-2
-12
-10
-8
-6
UMAP 1

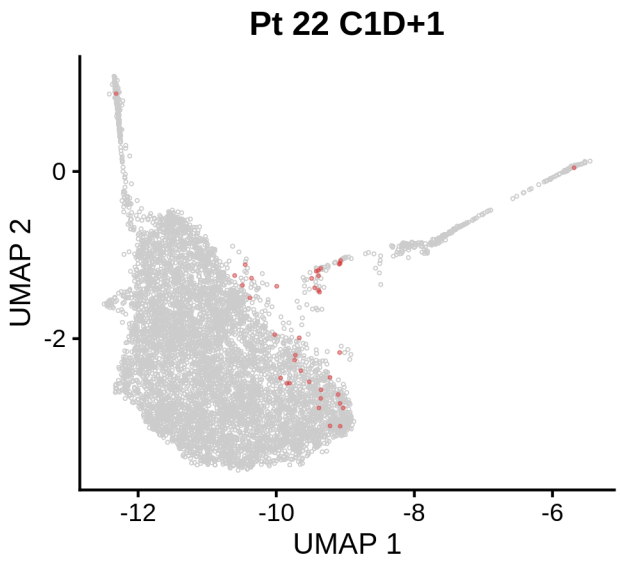
```
##
## [[7]]
```


Pt 17 C1D+1
UMAP 2
0
-2
-12
-10
-8
-6
UMAP 1

```
##
## [[8]]
```


Pt 22 C1D-7
UMAP 2
0
-2
-12
-10
-8
-6
UMAP 1

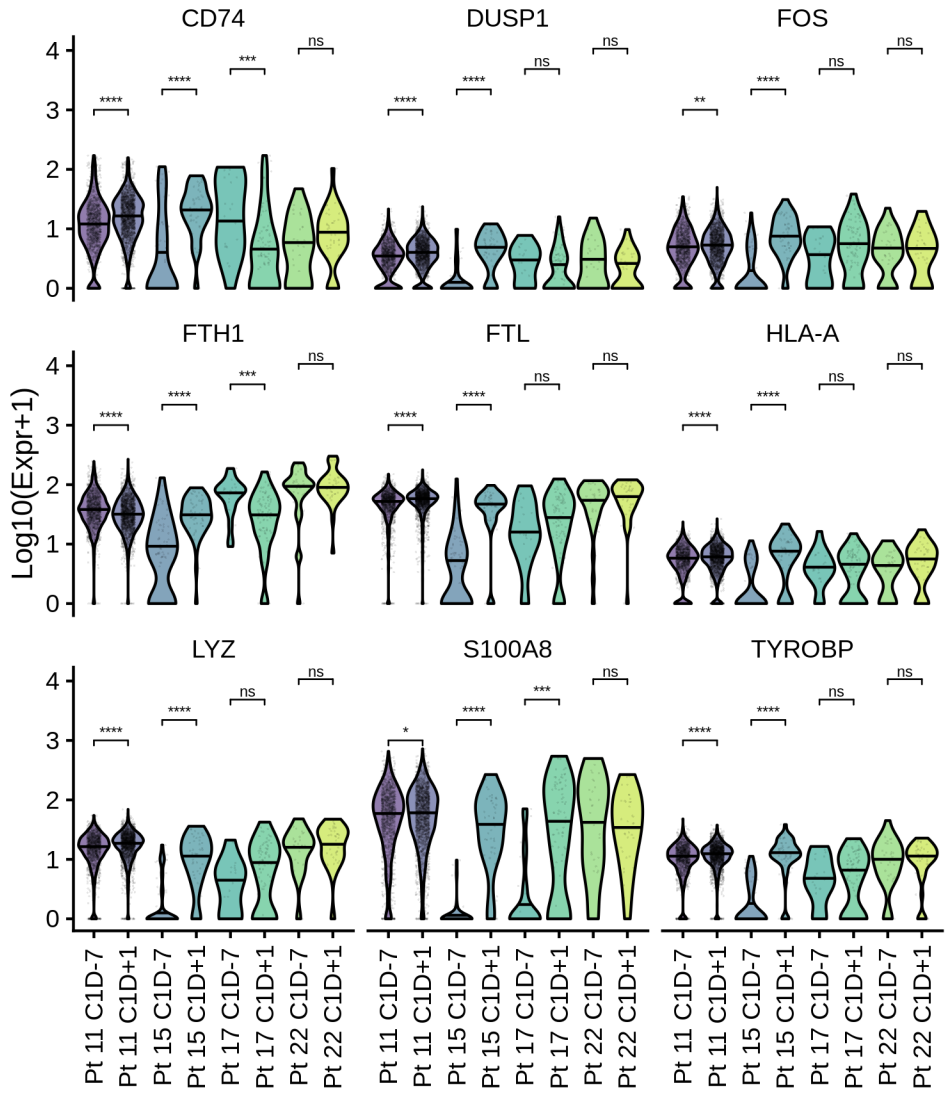
```
##
## [[9]]
```



One way we might be able to salvage something is to notice that if you look at it in the right way, in all of the patients the cells seem to go from right to left from D-7 to D+1. So maybe there are some consistent gene expression changes that drive this shift.

To find these gene expression changes, I calculated the top markers that are different between the D-7 and D+1 mdsc and plotted some of the more interesting ones here. Most of the interesting ones are immune function-related, which may not be a big surprise.

```
plot(mdsc_violins)
```



So it looks like processes like antigen presentation and macrophage killing might be turned on, although we have to accept the caveat that the numbers of MDSC in 3/4 patients are very small.

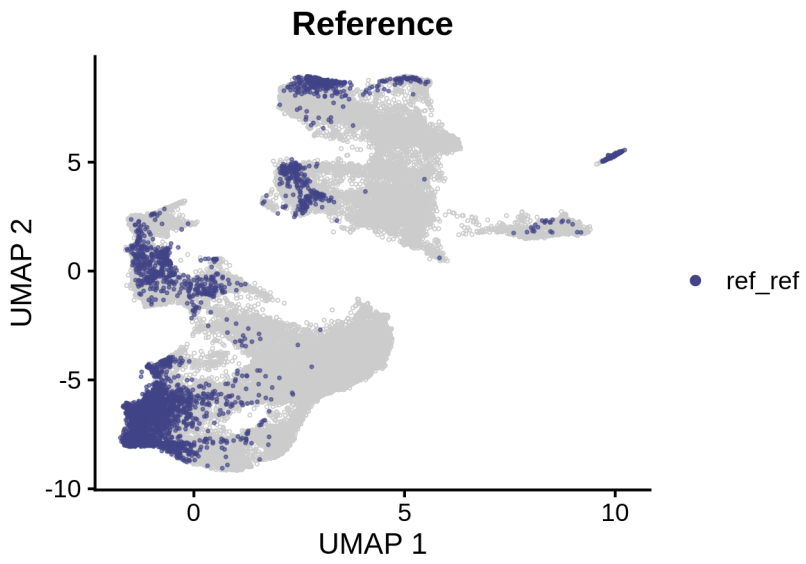
T cell analysis

There were two major T cell populations identified: Tc/NK and Memory/Naive T. These are the major divisions we see typically in pbmc datasets. Cytotoxic T cells and NK cells are very similar transcriptionally, so they tend to fall into the same clusters. They typically express granzymes, perforin, tbet, and cd161/klrb1. The naive and memory T cells typically express high levels of IL7R. I made a large series of plots of T/NK markers which you can find in "plots_out/t_tnk_genes".

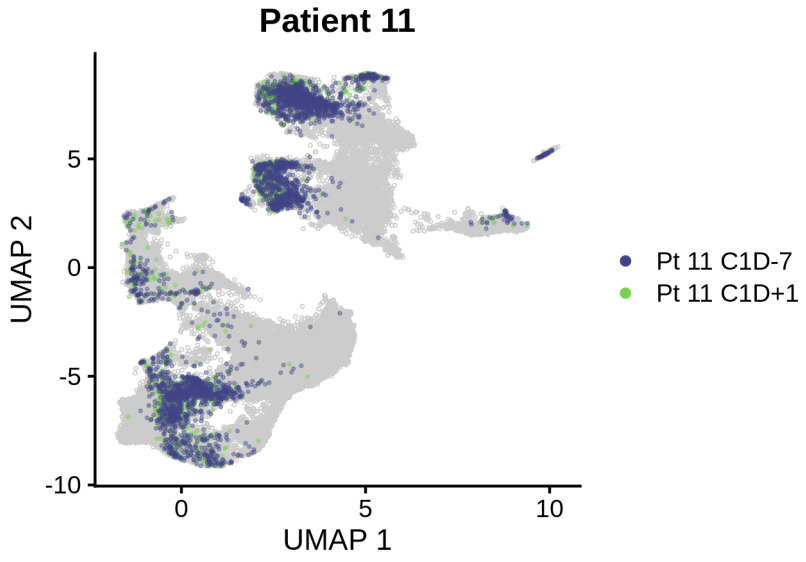
Global transcription patterns of the T cells were not consistently changing from d-7 to d+1.

```
TNK_plots
```

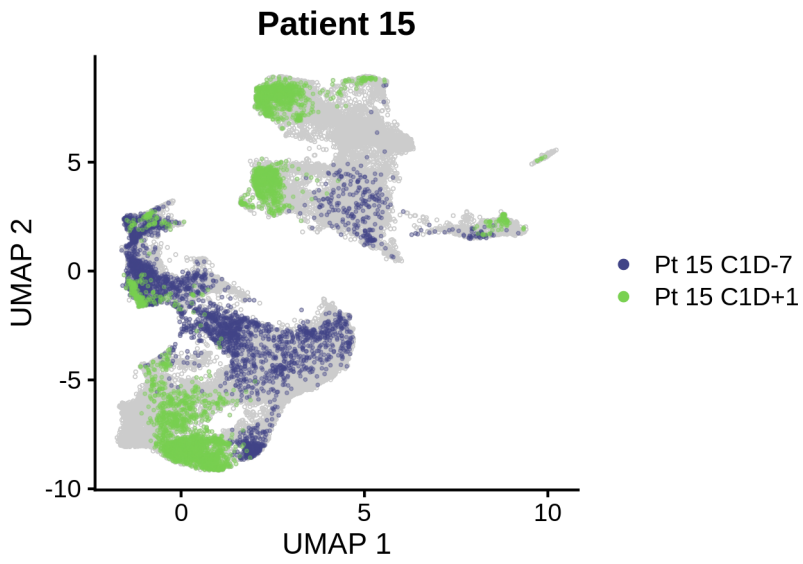
```
## [[1]]
```


Reference
UMAP 2
UMAP 1
5
0
-5
-10
0
5
10
ref_ref

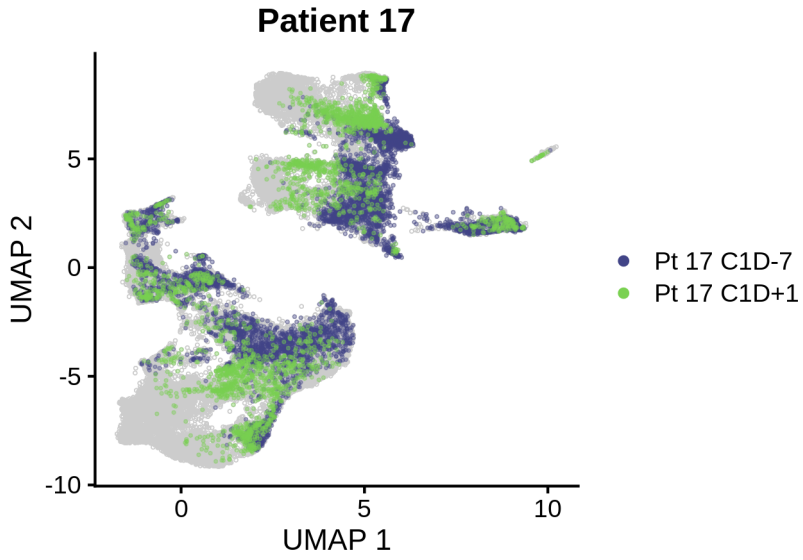
```
##
## [[2]]
```


Patient 11
UMAP 2
UMAP 1
5
0
-5
-10
0
5
10
Pt 11 C1D-7
Pt 11 C1D+1

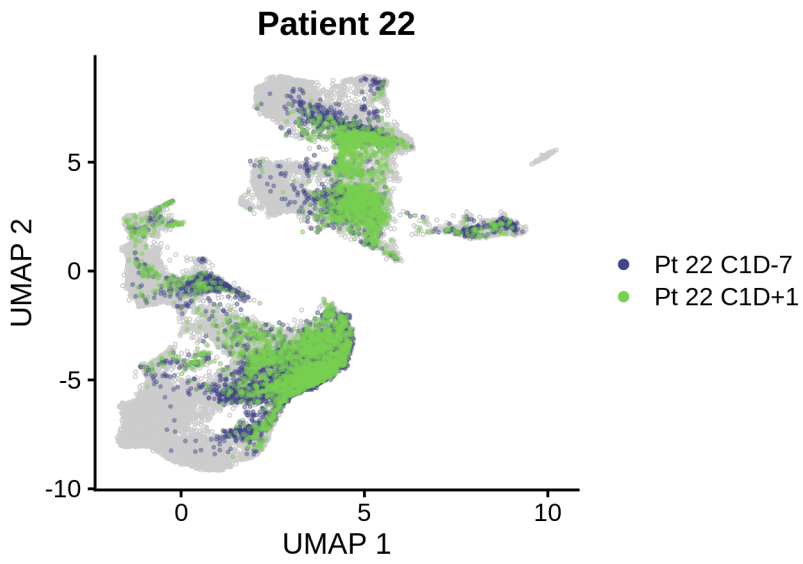
```
##
## [[3]]
```


Patient 15
UMAP 2
UMAP 1
5
0
-5
-10
0
5
10
Pt 15 C1D-7
Pt 15 C1D+1

```
##
## [[4]]
```


Patient 17
UMAP 2
UMAP 1
5
0
-5
-10
0
5
10
Pt 17 C1D-7
Pt 17 C1D+1

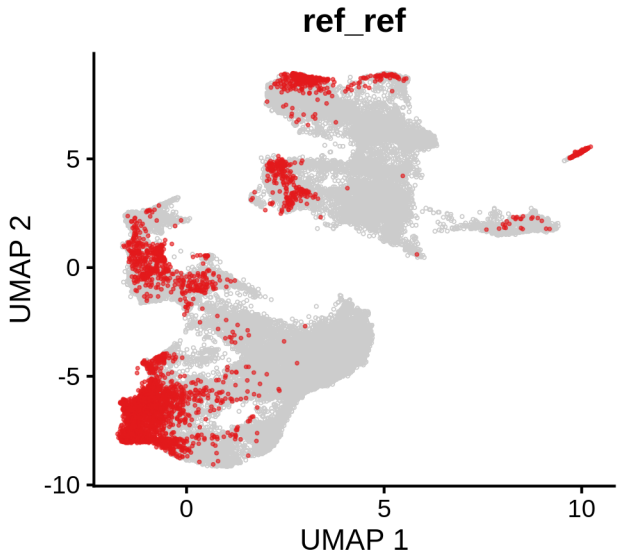
```
##
## [[5]]
```



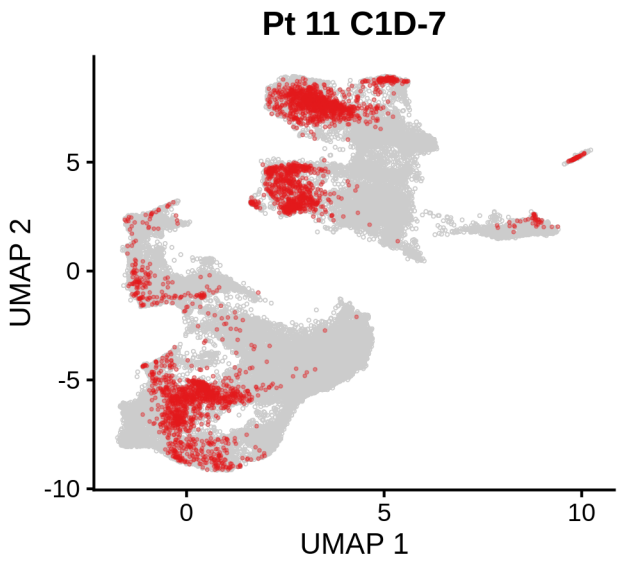
Plotted singly:

```
TNK_varplots
```

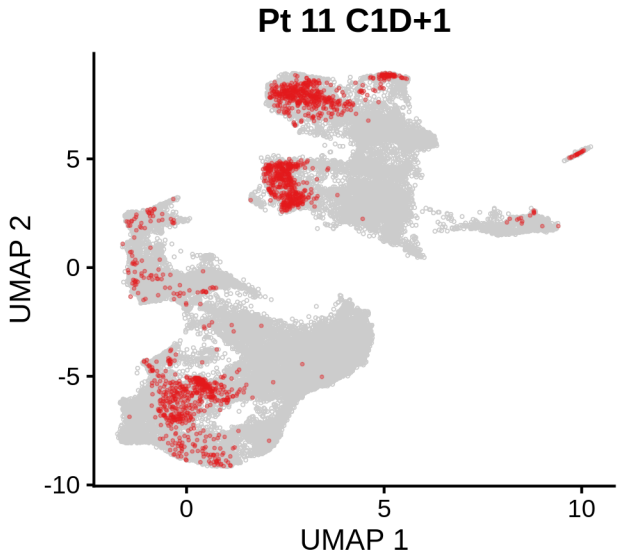
```
## [[1]]
```



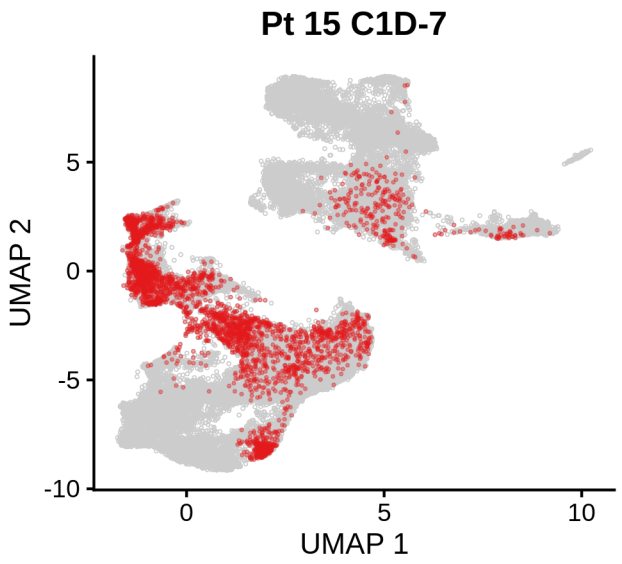

```
##
## [[2]]
```


Pt 11 C1D-7
UMAP 2
UMAP 1
5
0
-5
-10
0
5
10

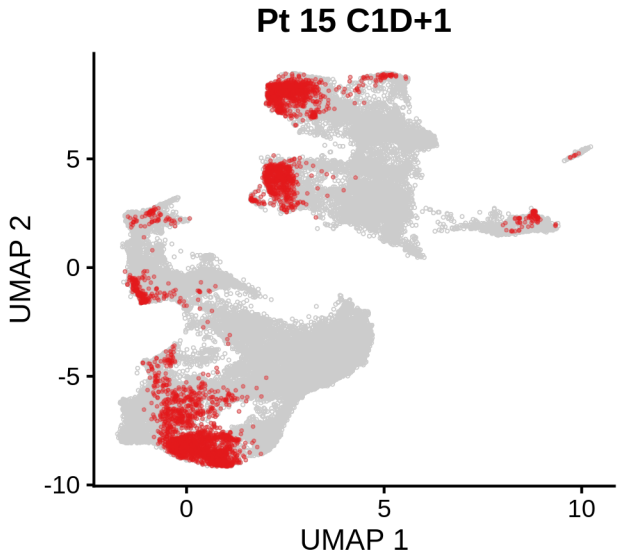
```
##
## [[3]]
```


Pt 11 C1D+1
UMAP 2
UMAP 1
5
0
-5
-10
0
5
10

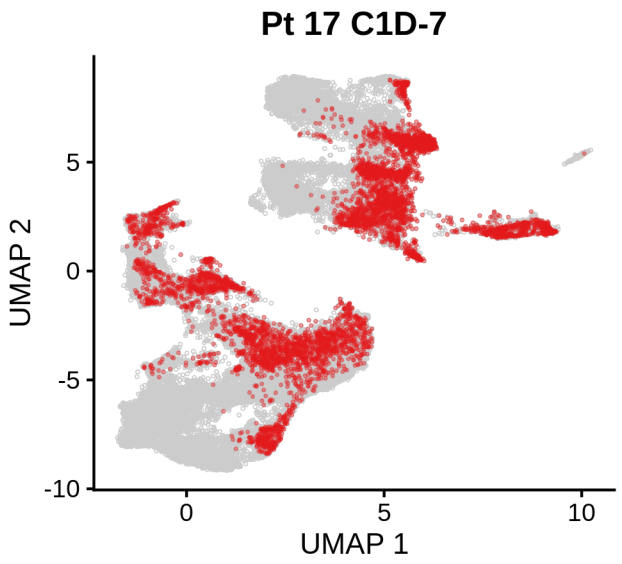
```
##
## [[4]]
```


Pt 15 C1D-7
UMAP 2
5
0
-5
-10
0
5
10
UMAP 1

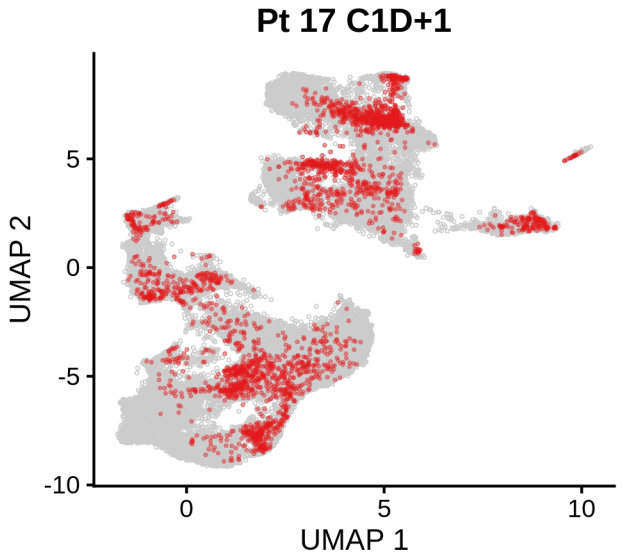
```
##
## [[5]]
```


Pt 15 C1D+1
UMAP 2
5
0
-5
-10
0
5
10
UMAP 1

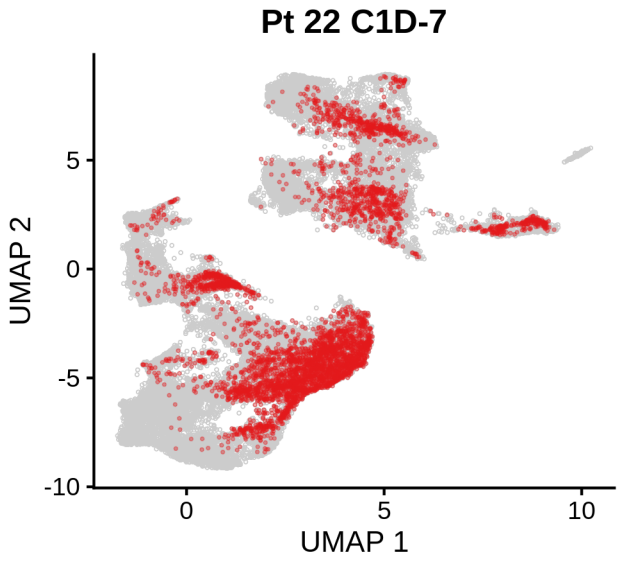
```
##
## [[6]]
```


Pt 17 C1D-7
UMAP 2
UMAP 1
5
0
-5
-10
0
5
10

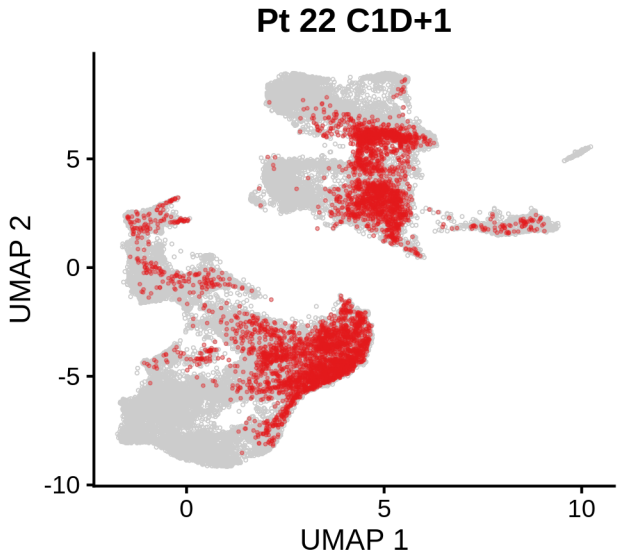
```
##
## [[7]]
```


Pt 17 C1D+1
UMAP 2
UMAP 1
5
0
-5
-10
0
5
10

```
##
## [[8]]
```




```
##
## [[9]]
```

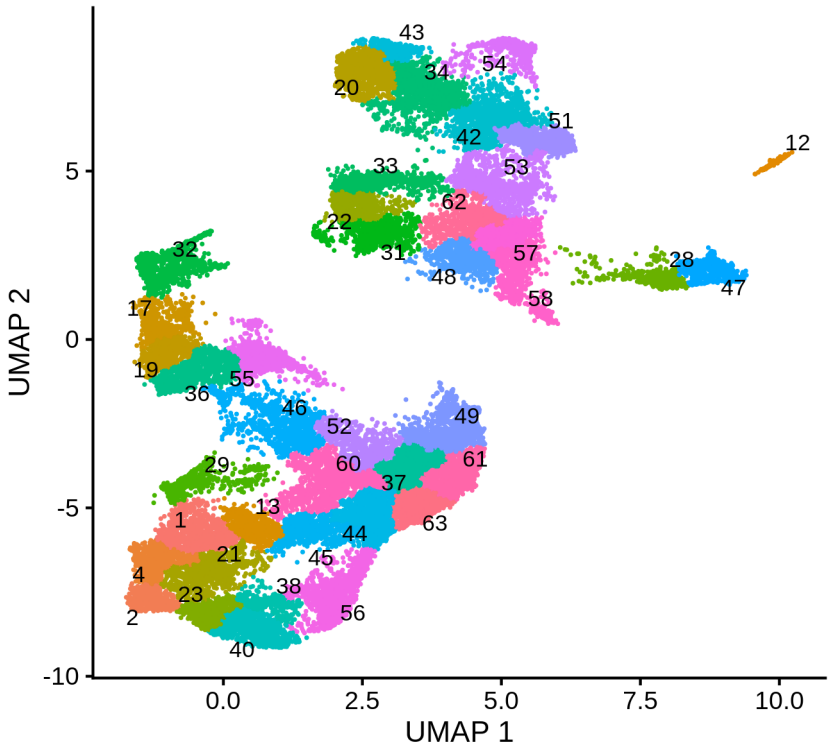


To me it looks like the reference cells and some of the samples like pt11, pt15 C1D1, and maybe a small degree pt17 C1D1 are off to the left side whereas the others are off to the right.

So what we can do is look at hyperfractionated clusters (Louvain method, in contrast to the larger partitions above) and combine these into sensible groups "Tc/NK ref-like", "Tc/NK not ref-like" and "Naive/Mem T ref-like" and "Naive/Mem T not ref-like".

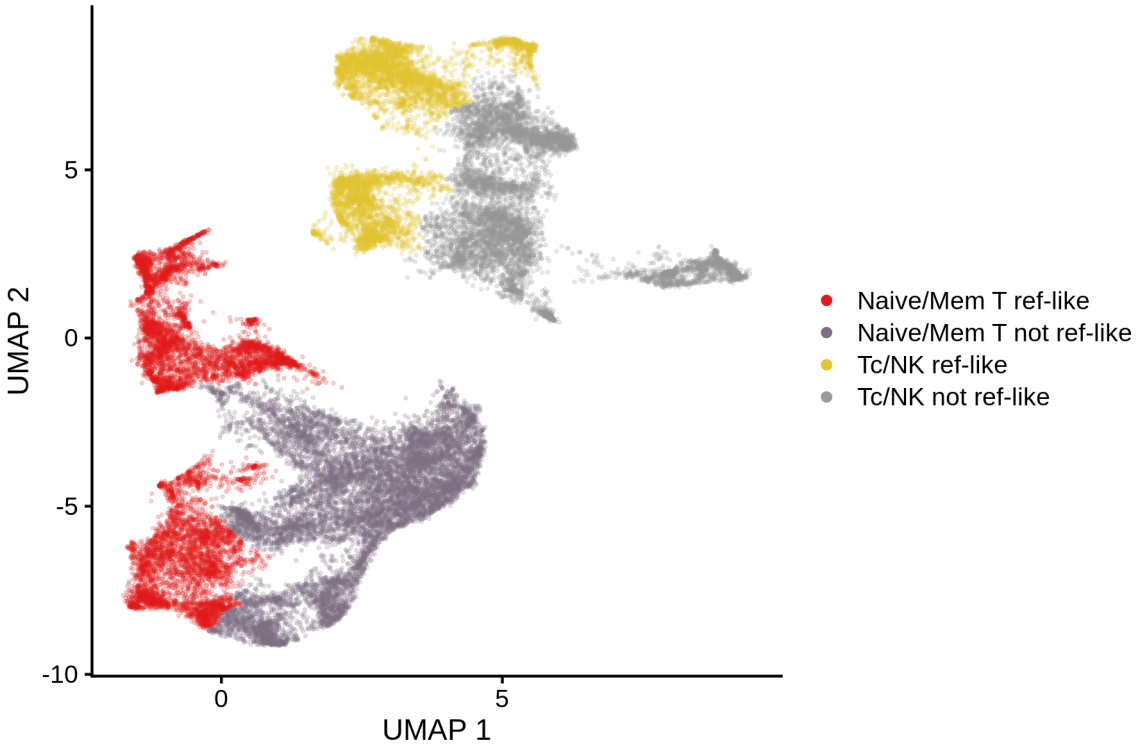
First the hyperfractionated clusters including reference:

```
plot(TNK_louvain)
```



Now reassign as described and also throw out the senescent cells because it loks like they are about 50/50.

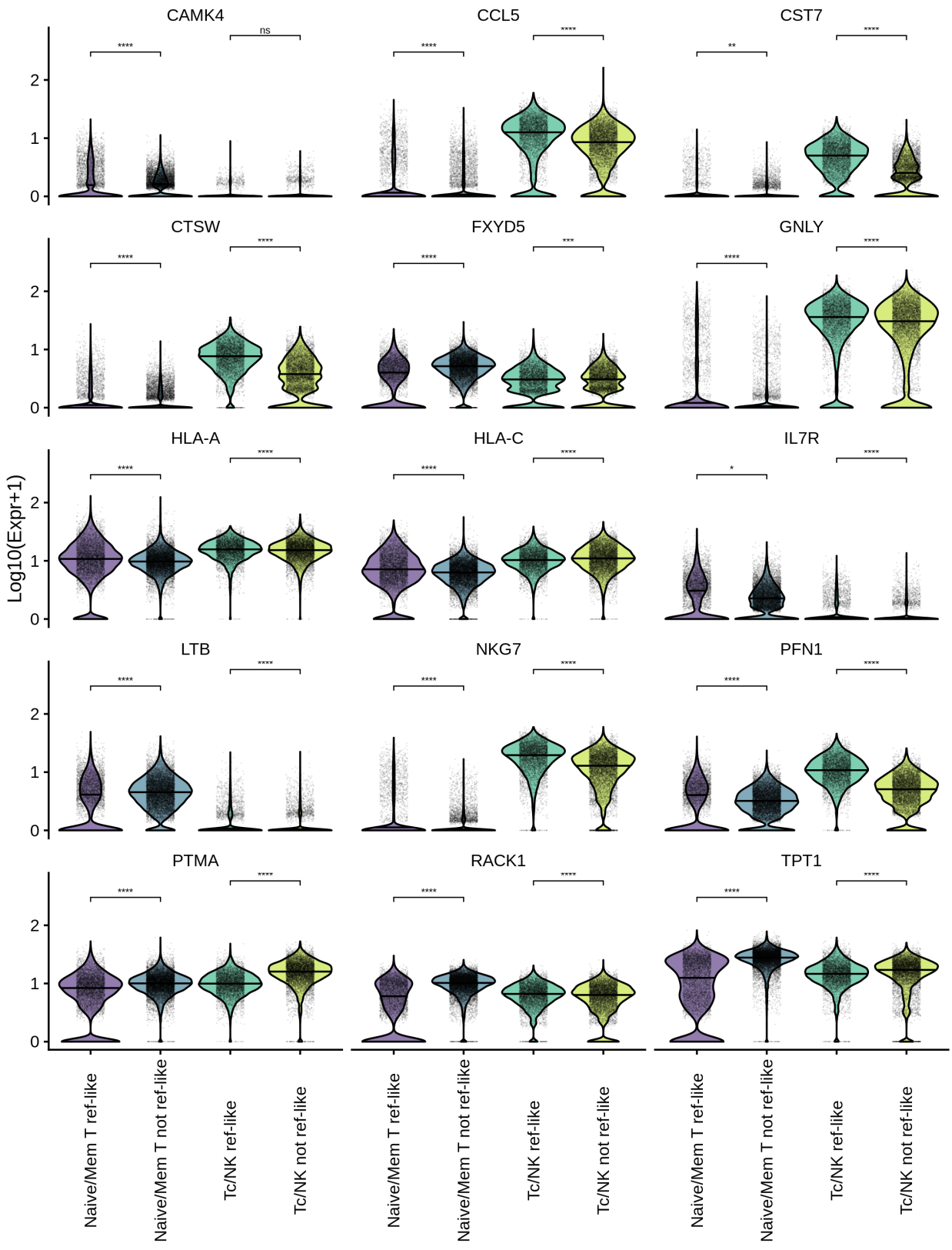
```
plot(TNK_recluster)
```



I calculated the top specific markers for these four groups, now excluding the reference cells because we are done with them. See "data_out/TM_TNK_reclust.csv".

Here are a few select genes:

```
plot(TNK_violin)
```



Brooke you should take a look at these genes and also the genes in the TM_TNK_reclust.csv file and let

me know if anything makes sense with the rest of your story.

TCR analysis

We were able to amplify and sequence the TCR-alpha and TCR-beta CDR3 regions on somewhere between 500 and 2500 T cells per sample.

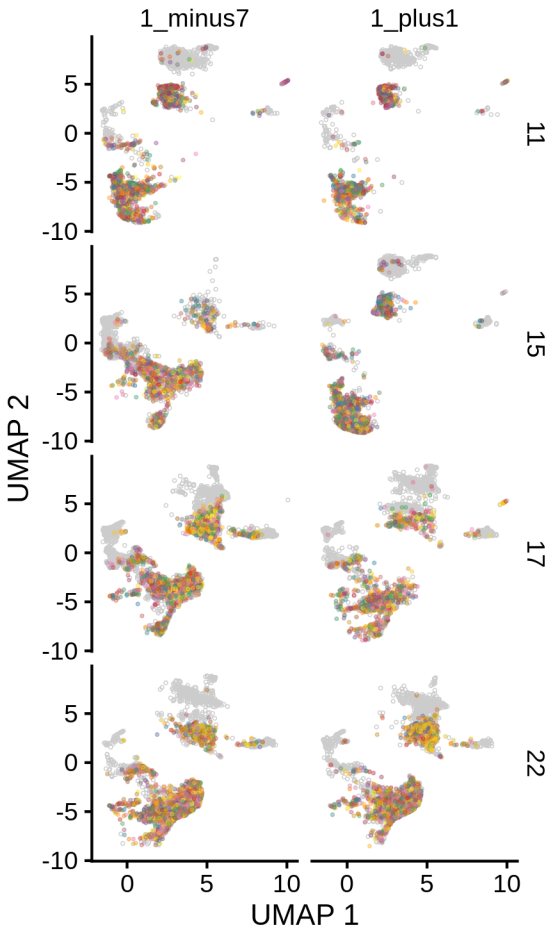
Most T cells express 1 unique alpha and 1 unique beta chain. Some express more than 1 (up to 6 in these datasets). Some of these are probably genuine and some are probably erroneous reads, but there aren't enough to really worry about. The 10X software identifies each combination of TCR chains uniquely as a clonotype. The number of unique clonotypes in each sample was

```
sample_clonotype_numbers
```

```
## # A tibble: 8 x 2
##   pt_cycle_day clonotype_id
##   <chr>               <dbl>
## 1 11_1_minus7          1392
## 2 11_1_plus1            670
## 3 15_1_minus7          1049
## 4 15_1_plus1           2154
## 5 17_1_minus7          1612
## 6 17_1_plus1            941
## 7 22_1_minus7          2575
## 8 22_1_plus1           1982
```

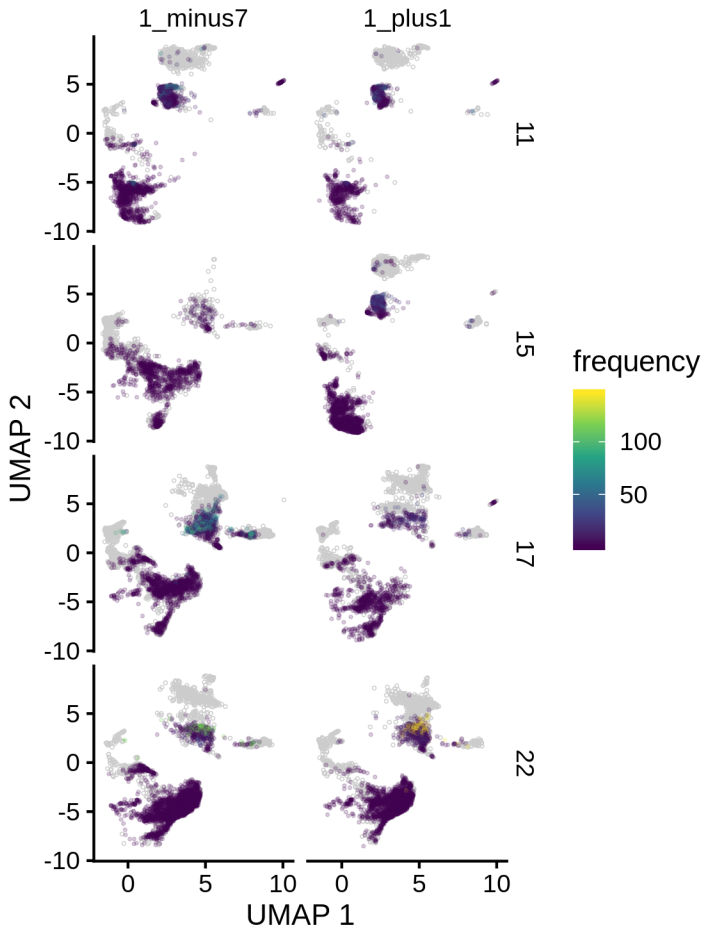
We can plot them in UMAP space with each color representing a unique clonotype.

```
plot(tcr_scatter_facet)
```



In this plot, there are so many different clonotypes it is hard to see if one is expanded or not. So we can also plot the frequency of each individual clonotype on a color scale with yellow int this case representing clones with the most copies:

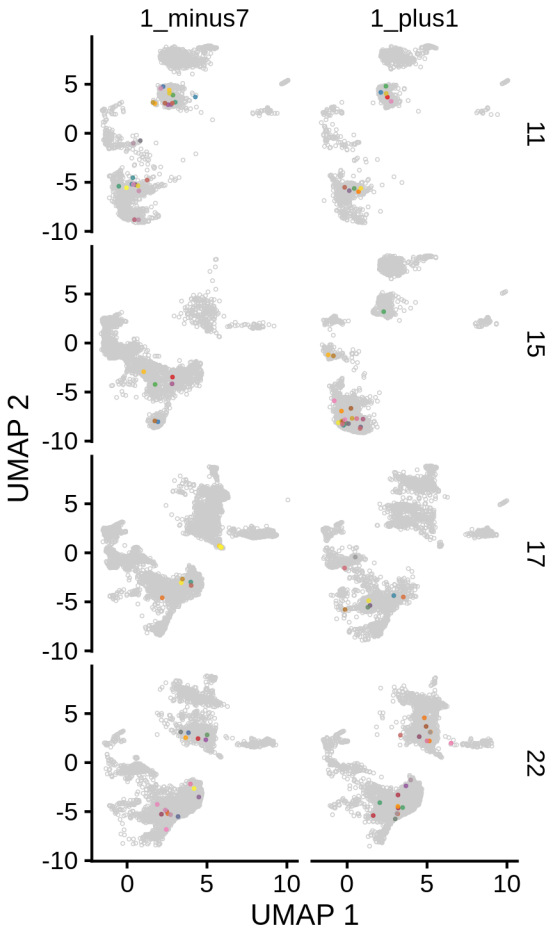
```
plot(frequency_scatter_facet)
```



It looks like patient 22 has a clone that expands on the study treatment and pt 17 has a clone that shrinks. Other than that, not much to say.

We can use a database to annotate the TCR-a and TCR-b sequences with assoications with pathological conditions. See http://friedmanlab.weizmann.ac.il/McPAS-TCR/. The annotation is very sparse, meaning that most TCR sequences have no known patholgic associations. Most that do are specific for EBV, CMV, and influenza, as you might expect. We can join this entire database onto our list of T cell clonotypes ("data_out/TRA_TRB_anno.csv"), then filter for clonotypes containing cancer-associated TCRs and then plot these:

```
plot(cancer_barcodes_facet)
```



I don't think there are enough there to draw any major conclusions. If you look at the TRA_TRB_anno.csv file, you will see that there are melanoma-specific T cells, but I'm not sure if there are more there than you would expect by chance alone. If you want I can give you the epitopes for these annotated TCRs.

Summary

Overall there were some challenges with this data set. I tried pretty hard to extract some meaningful data. The concern is that with just a couple of patients in each group and especially since only 1 had a subtantial number of MDSCs, it can be hard to justify any conclusions on these data alone (of course it's rare that a single experiment will tell the whole story). Hopefully these data support other lines of evidence in your paper. Certainly the data are good for generating new hypotheses if that is something you are interested in.

For the analysis of the MDSCs that we could find, it may be that upregulation of CD74 (class II associated molecule), HLA-A, LYZ, S100A8 and TYROBP represent a shift in immunologic function. This was mostly seen in pt 11, who was a responder.

For the T cell analysis, using the reference cells allowed us to define a presumably normal T cell state. Because there were many patient cells that shared this state and many that did not, it was easy to find many genes that are significantly different. For example CCL5 (RANTES) and NKG7 in the Tc/NK cells. You have to be careful interpreting some of these. For example, it says that HLA-A is significantly different (because we sampled so many cells) but the magnitude of the difference is small. Still it looks like the patient cells with non-reference like phenotype have suppressed expression of cytotoxic markers.

I think it is reasonable to add more samples to this study. But I think they should be prescreened to be sure there are MDSC present. Alternatively, if there are things in here that fully support major lines of evidence in your paper, we also don't have to repeat. No piece of data needs to stand on its own in my opinion.

Happy to discuss any questions you may have.