FLT3-ITD AML scRNA-seq Data Analysis

Brad Blaser

3/30/2020

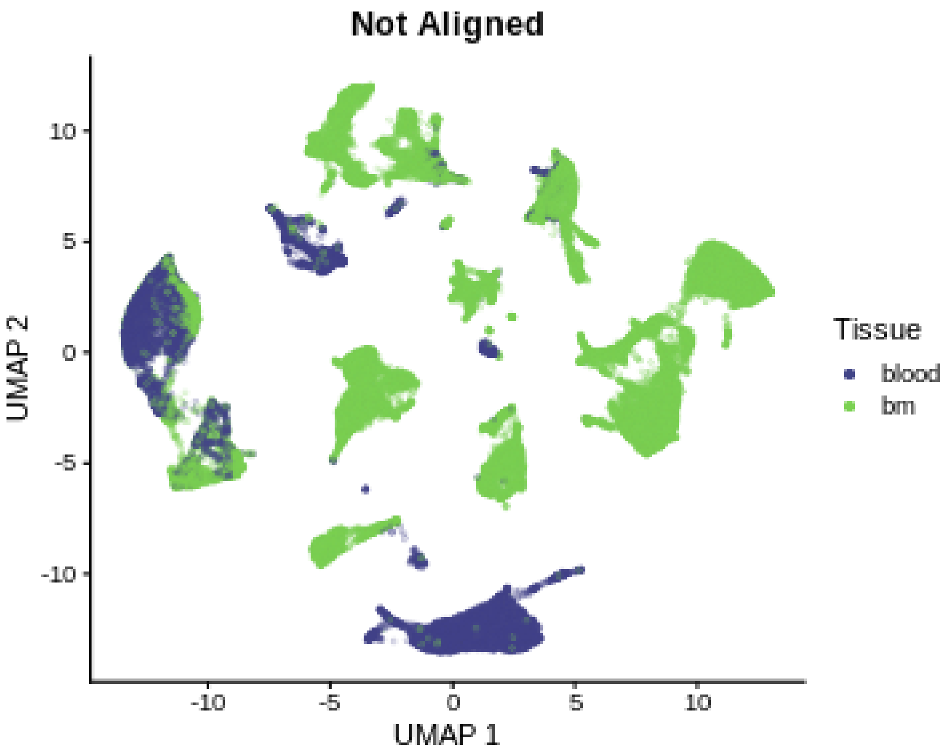
Sample Selection

We have 10 samples from 4 patients. There is a combination of different timepoints relative to treatment with Gilteritinb. There are different response, timepoints and tissues collected.

The first question is whether we can use the blood samples with or without batch correction.

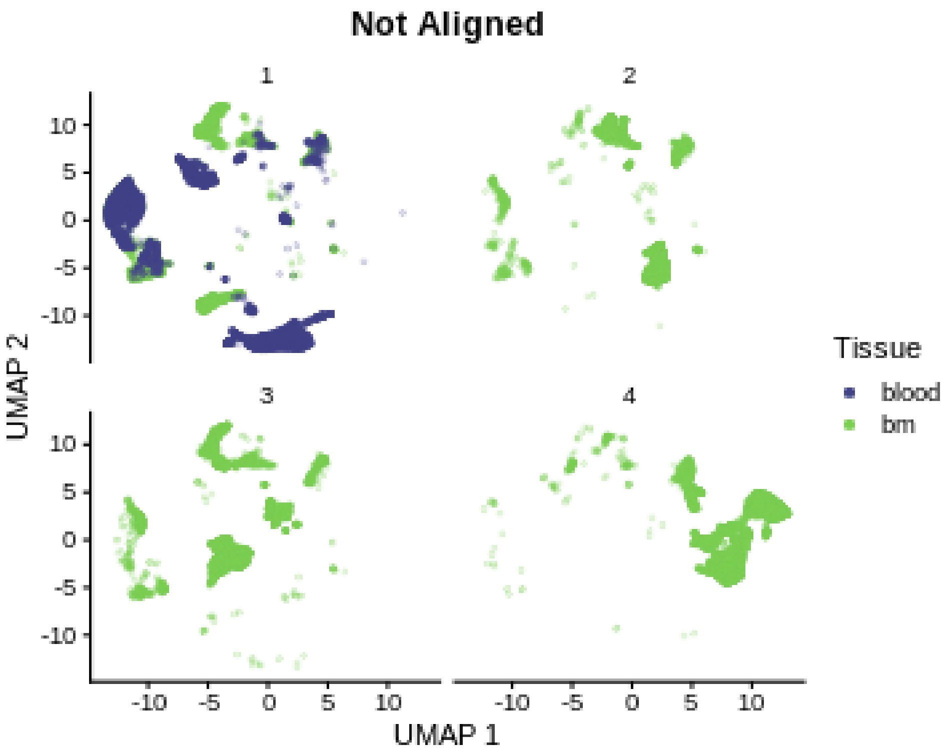
Here is the dimension-reduction (UMAP) plot for all cells colored by tissue:

```
plot(cds_trimmed_all_cvp)
```



Here it is split by patient:

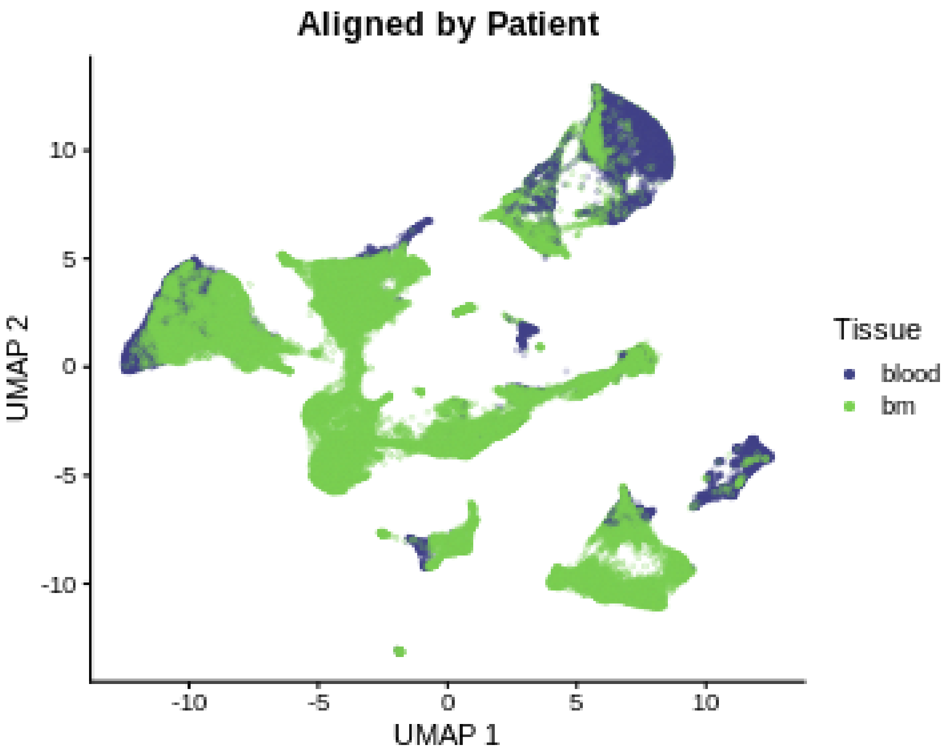
```
plot(cds_trimmed_all_cvp_faceted)
```



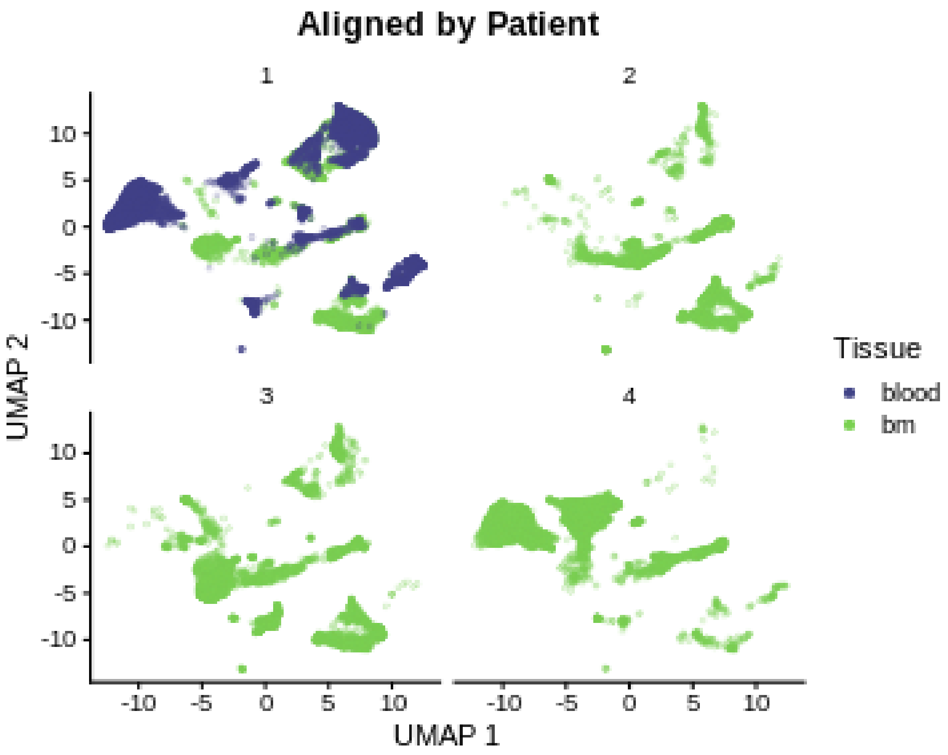
So it looks like there is a substantial difference between blood and BM.

After batch correction, these are the plots:

```
plot(cds_aligned_all_cvp)
```



```
plot(cds_aligned_all_cvp_faceted)
```



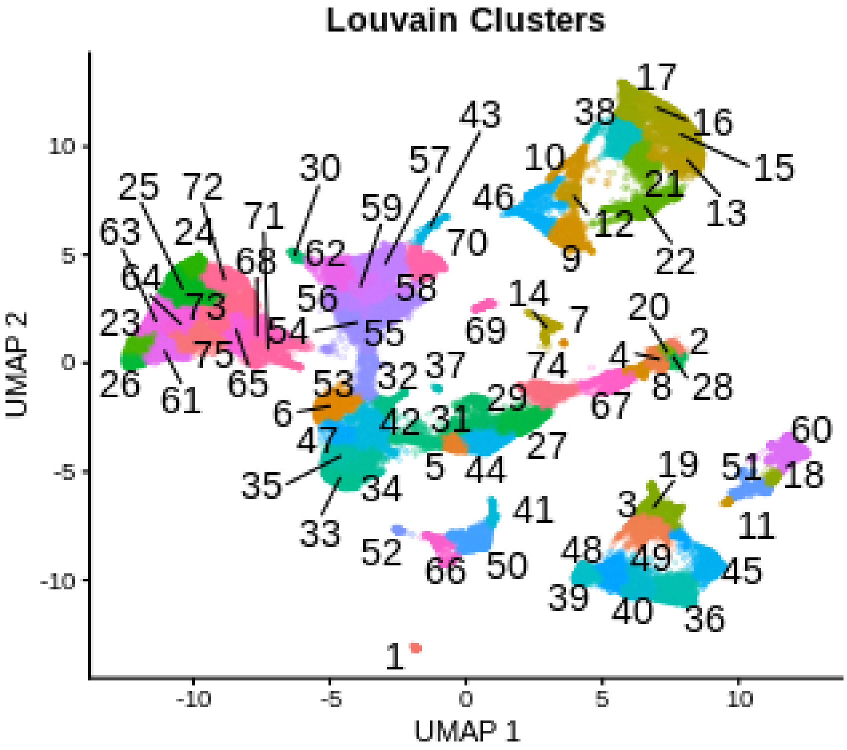
This looks pretty good. If you look at the faceted plot, you can see that the large population of PB blasts overlaps a fair number of BM cells. So probably OK to use. We may have to revisit this, however, and for sure we shouldn't drill down too deeply into differences across tissue types.

Unsupervised clustering

Cells were clustered using the Louvain and Partitioning methods. Different algorithms are used for these methods and partitions end up being much larger.

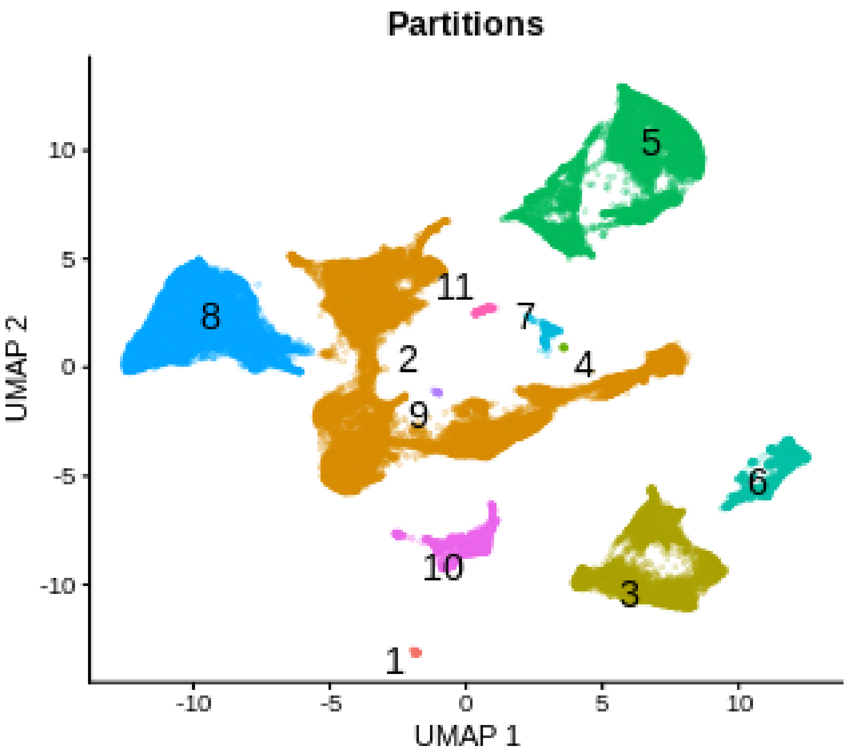
Louvain Clusters:

```r
plot(cluster_plot)
```



Partitions:

```r
plot(partition_plot)
```

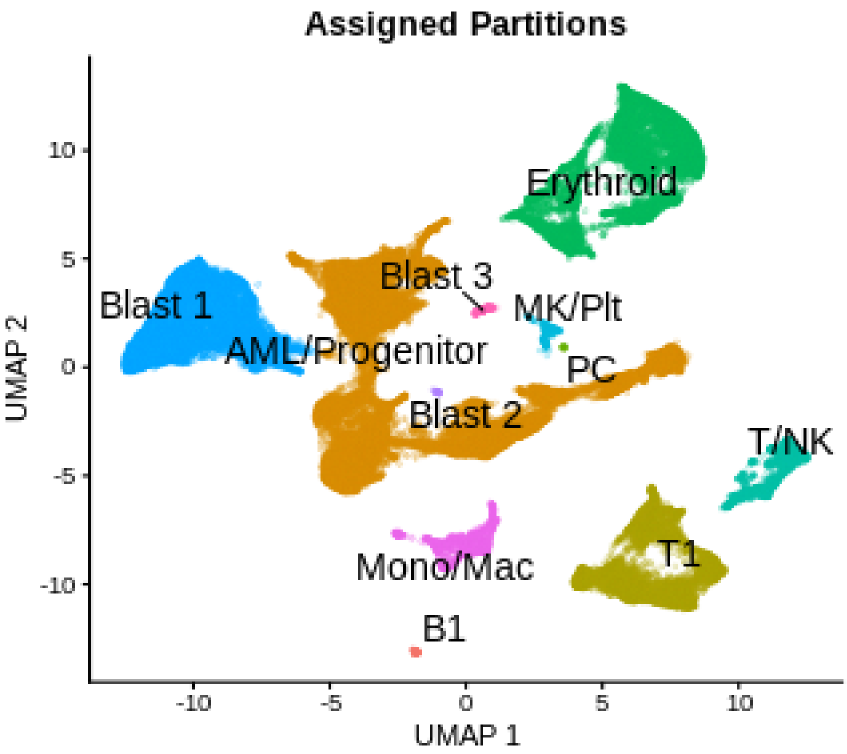


There are way too many clusters to start with at this point. If we find something interesting we may use these to drill down later.

The top 50 genes for each cluster and each partition have been generated and can be found in the data_out folder. Cell types were identified by manual inspection of the top genes in the partitions. Save the clusters for later.

Partition plot with assignments:

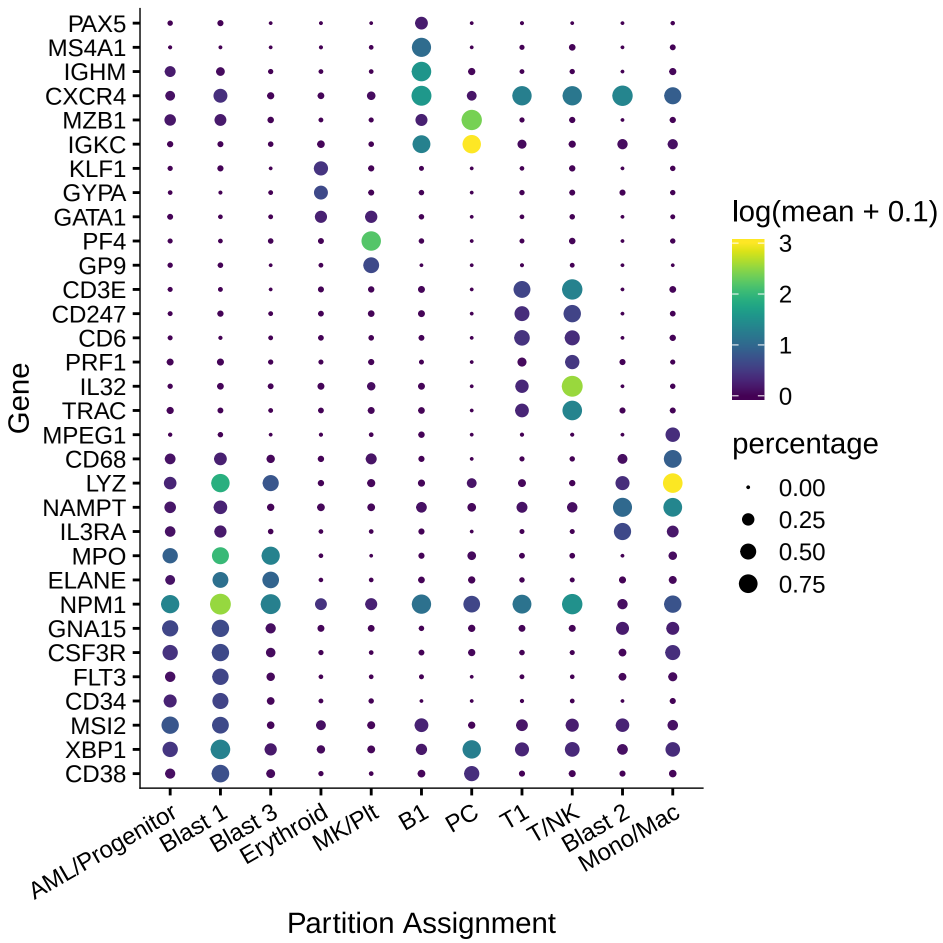
```
plot(partition_assignment_plot)
```



One way to visualize the top genes for each cluster is in "bubble plots". In these plots, the size of the circle corresponds to the percent of cells in the cluster expressing the marker, and the color scale corresponds to normalized expression.

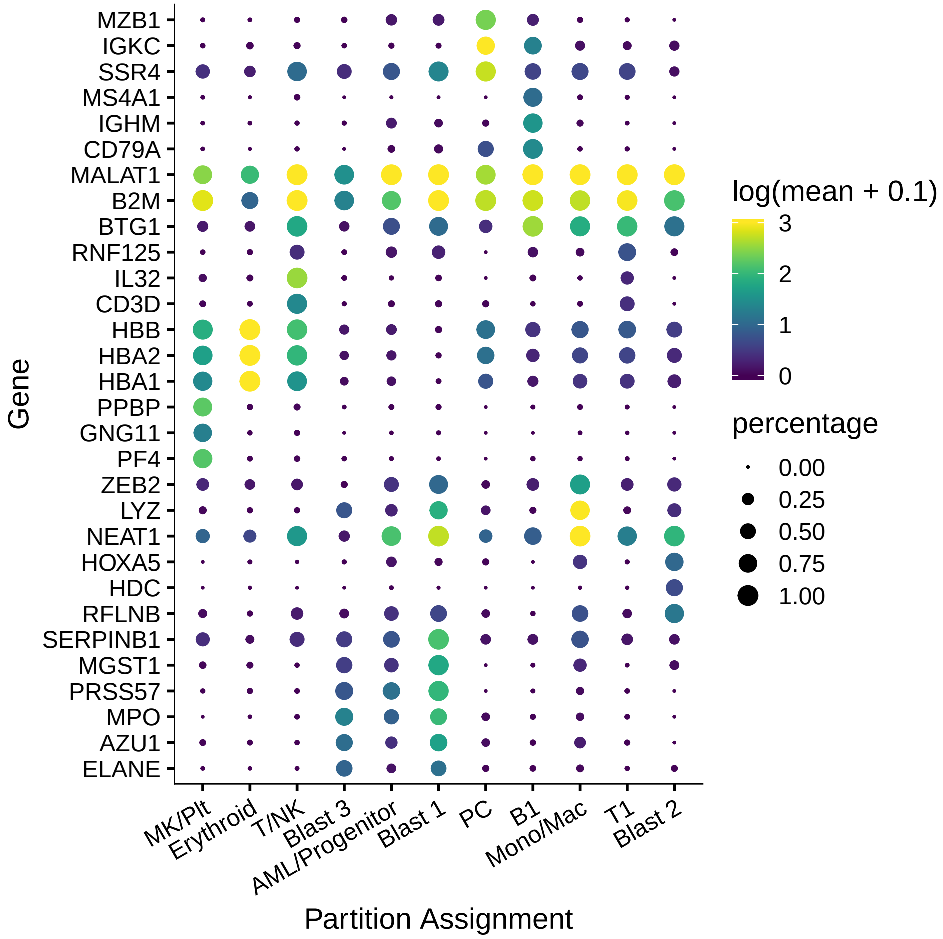
Here is a selection of genes that stood out to me when reviewing the lists:

```
plot(gene_dot_plot1)
```



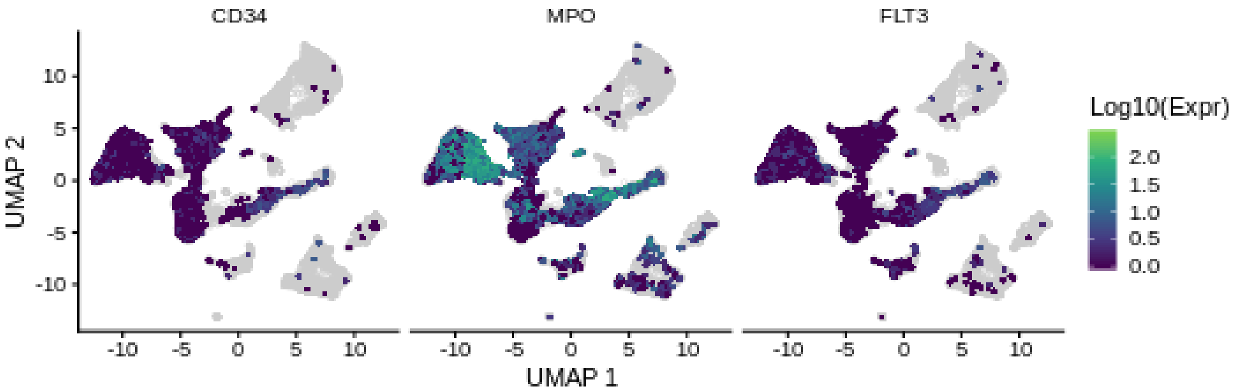
We can also calculate the top specific genes for each cluster. One statistic to measure this is pseudo-R2. This is the logistic regression correlate to the R2 value from linear regression. Here are the top 3 genes in each group by this metric:

```
plot(gene_dot_plot2)
```



We can also print out specific genes and look at them on at the single cell level:

```r
plot(three_gene)
```

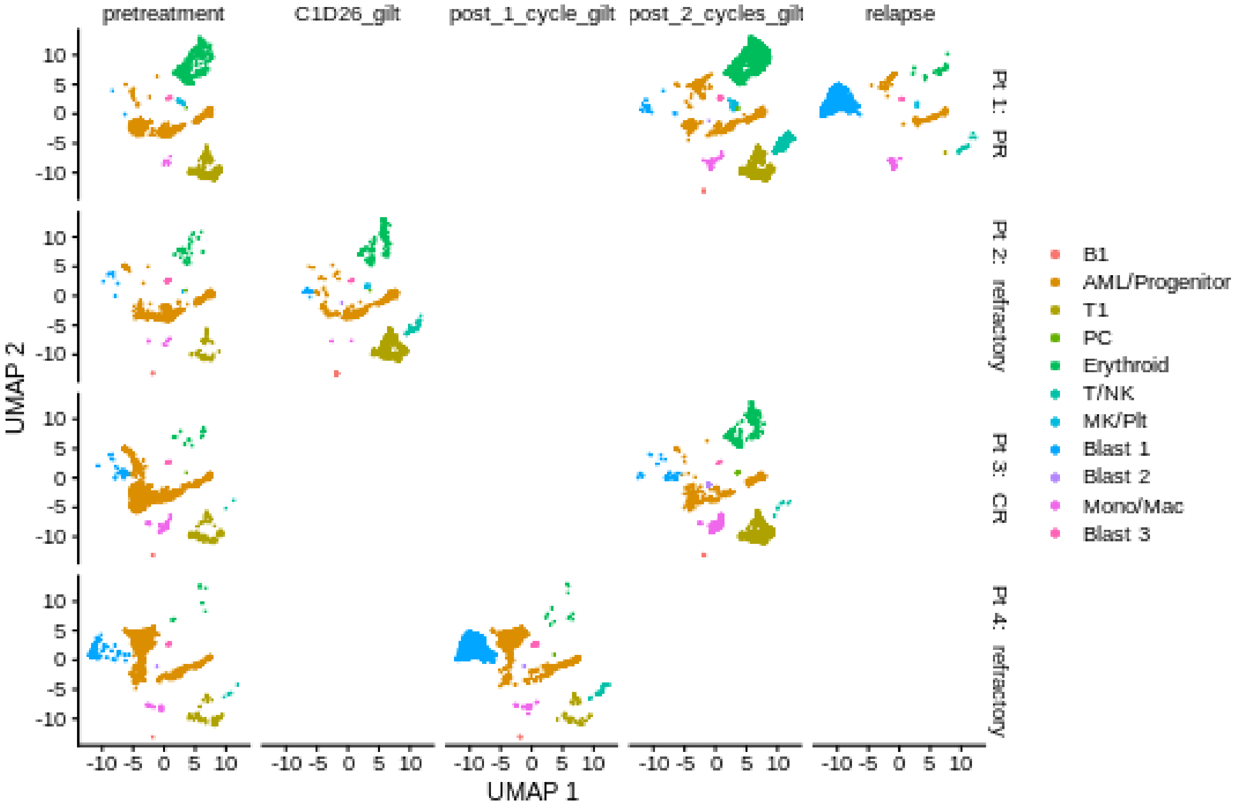


The counts for each patient/timepoint/partition combination are in the data_outs folder (partition_assignment_counts.csv).

Look for patterns in the cell clusters

Having clustered all of the cells into the same 2-dimensional space and identified groups of cells (partitions), we can restratify them by patient and timepoint (basically by sample).

```
plot(pa_pt_response_timepoint)
```

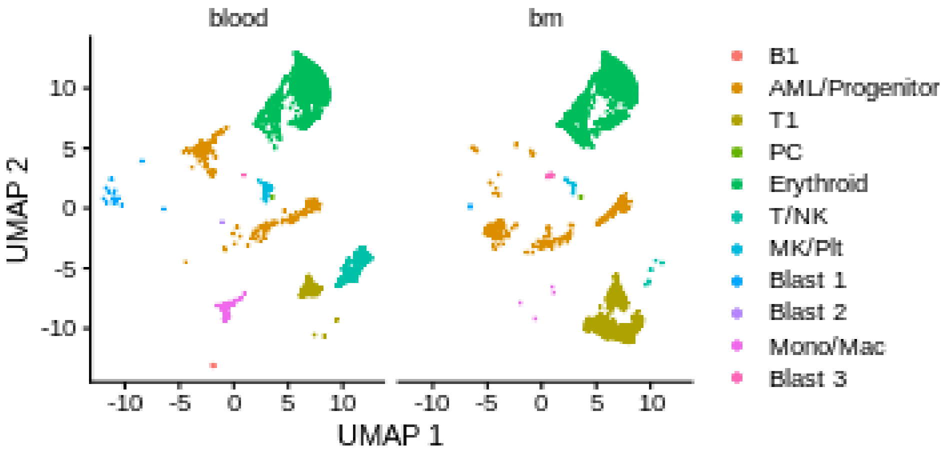


This shows a couple of interesting things.

For patient 1, the concern was that the blood was not going to be useful. Relapse is blood and post 2 cycles is blood and marrow. The concern is that the green "erythrocytic" cluster is RBC contaminants or doublets. Seems less likely because this cluster goes way down at relapse (also a blood sample) and the fact that there are lots of these cells in other samples. Erythrocyte doublets generally look different and express fewer stem cell genes than these cells express (you can take a look at partition_markers.csv).

To further make this point, we can look just at the post 2 cycles samples for Pt1 and stratify by tissue:

```
plot(pt1_post2_tissue)
```



Not identical, but very close. The big green erythroid population is the same for sure.

In light of this, we can feel more confident that those are blasts with erythroid differentiation, rather than reads contaminated by rbc or free globin RNA. We can also observe that in the post cycle 2 sample you start to see some blue cells showing up in the blast cluster. So this might be a pattern of emergent drug resistance.

Pt 2 is something of a mystery. This patient is supposed to have 80% or so blasts. They must have a very progenitor-like phenotype, since we can see both pretreatment and C1D26 look similar to the remission patient.

The remission patient seems to have expansion of the AML/progenitor (orange) cluster. This is relatively reduced in remission. And we see expansion of the erythroid cluster, which in this case are probably normal erythroid lineage cells.

Patient 4 seems to have blasts in the AML/progenitor and blast 1 partitions. It looks like the blast 1 population grew out on therapy, similar to the pattern seen in patient 1.

The blast 2 and blast 3 populations are really small but seem to be present in most of the samples. Would be interesting to know if they had the ITD or not. Maybe they are normal myeloblasts or maybe they are LSCs.

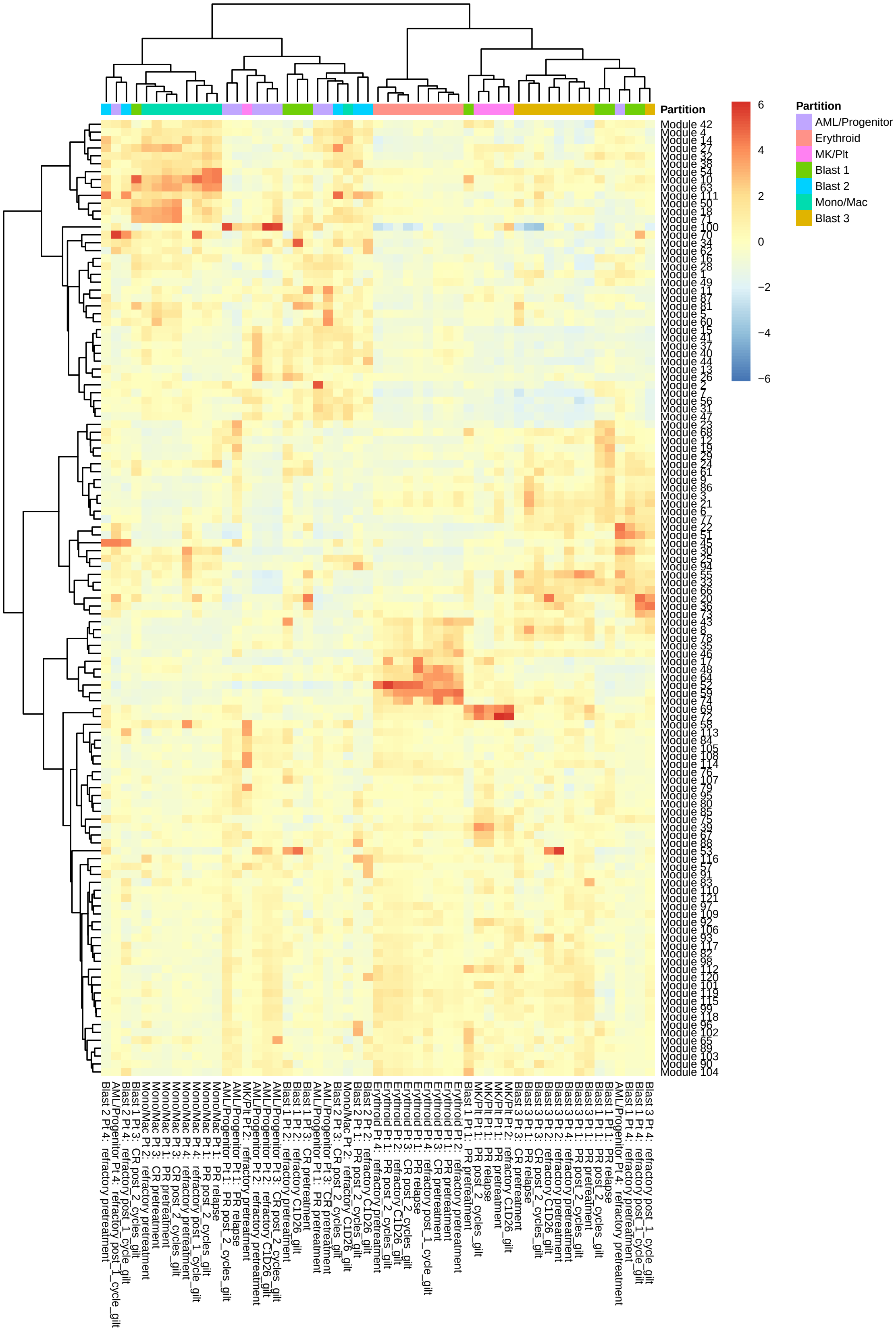
It is also interesting that the remission patient had some cells in the blue blast 1 partition, even with 0% blasts by morphology. So I wonder if there is some substructure within this population. One way to look at this question is to use gene modules.

Use Gene Modules to find a resistance signature

Gene modules are like cell clusters in reverse. The computer takes all 35,000 genes in the reference genome and asks how they vary in expression across the 77,000 cells in the dataset. The computer calculates UMAP coordinates and clusters genes, just like it does for cells. Gene clusters are called Gene Modules in monocle. They are defined based on the heterogeneity of gene expression across all cells but what is cool is that you can basically generate a module score for an arbitrary subset of cells.

We can ask the computer to calculate a module score for each partition in each sample at each timepoint. Then we can cluster these values together to find relationships in the data and extract an interesting set of genes.

```r
plot(modules_granular_noTBNKPC)
```



This is a busy heatmap. I've excluded the T, B, NK and plasma cell populations because they are largely invariant and not completely relevant to the current questions.

Each column is a composite variable of partition, patient (with response), and timepoint. Along the top I've added an annotation to show only the partition. So the columns that are interesting are the ones where the composite variable doesn't all cluster together. In other words, this means that gene expression within the partition varies according to patient and/or time. So it looks like Blast1 and AML/progenitor is where most of the action is.

Each row is a gene module. So what you can do is look for modules that go up or down as you go across patients or timepoints within partitions. For example module 10 is super-hot in Patient 3 (CR) at remission.

We can take the genes from module 10 and ask which GO terms they are associated with. There are many tools to do this, but I like gorilla. http://cbl-gorilla.cs.technion.ac.il/

All gene modules are in 1 csv file: data_out/gene_modules.csv. At the gorilla website…

Step 1: Choose Homo sapiens;

Step 2: Choose two unranked lists of genes;

Step 3: Copy/paste module 10 into the box for target set and copy/paste all genes in to the box for background set.

Step 4: Choose all.

Step 5: Click Search Enriched GO Terms.

Output shows you all of the enriched go terms. In this case you get a bunch of things related to myeloid cell activation, which seems to suggest maturation arrest may have been lifted. That's one hypothesis anyways.

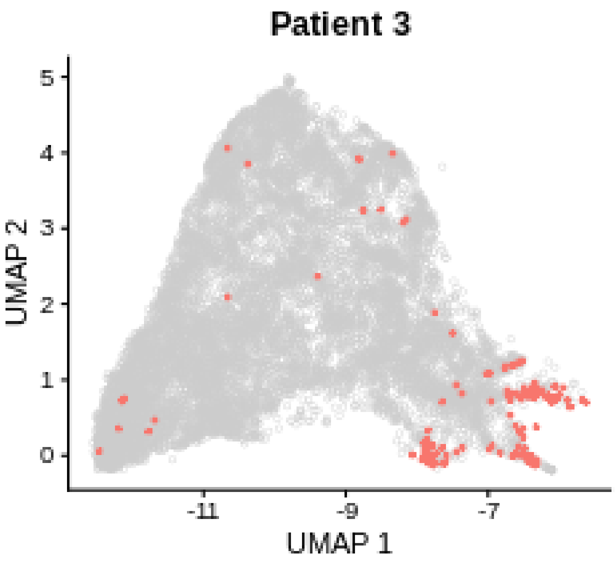
So I think this is where you can spend a little time exploring the data. Find the modules that look interesting and plug them into gorilla like I did.

You can explore Revigo (on the gorilla home page) for different ways to summarize redudndant GO terms and for visualization.

Go back to clusters focus on a few interesting ones

Now that our attention is on the Blast 1 partition, we can notice that the cells from patient 3 (CR) in that partition mostly lie in the lower right corner of the population whereas the rest seems to be coming from patients 1 and 4 (PR and refractory).

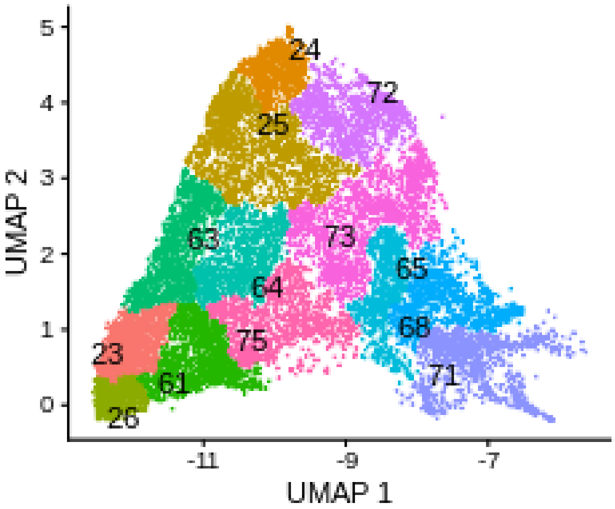
```
plot(pt3_blast1)
```



So is there some substructure within this partition?

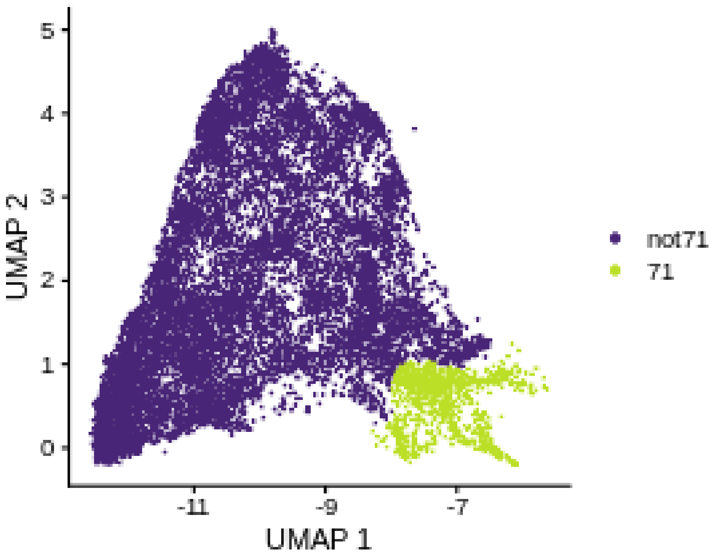
To answer this, we refer to the initial cluster plot, it looks like clusters 71 might correspond to the CR patient and clusters 23,24,25,26,61,63,64,65,68,72,73,75 to the PR and refractory patient.

```
plot(blast1_clusters)
```



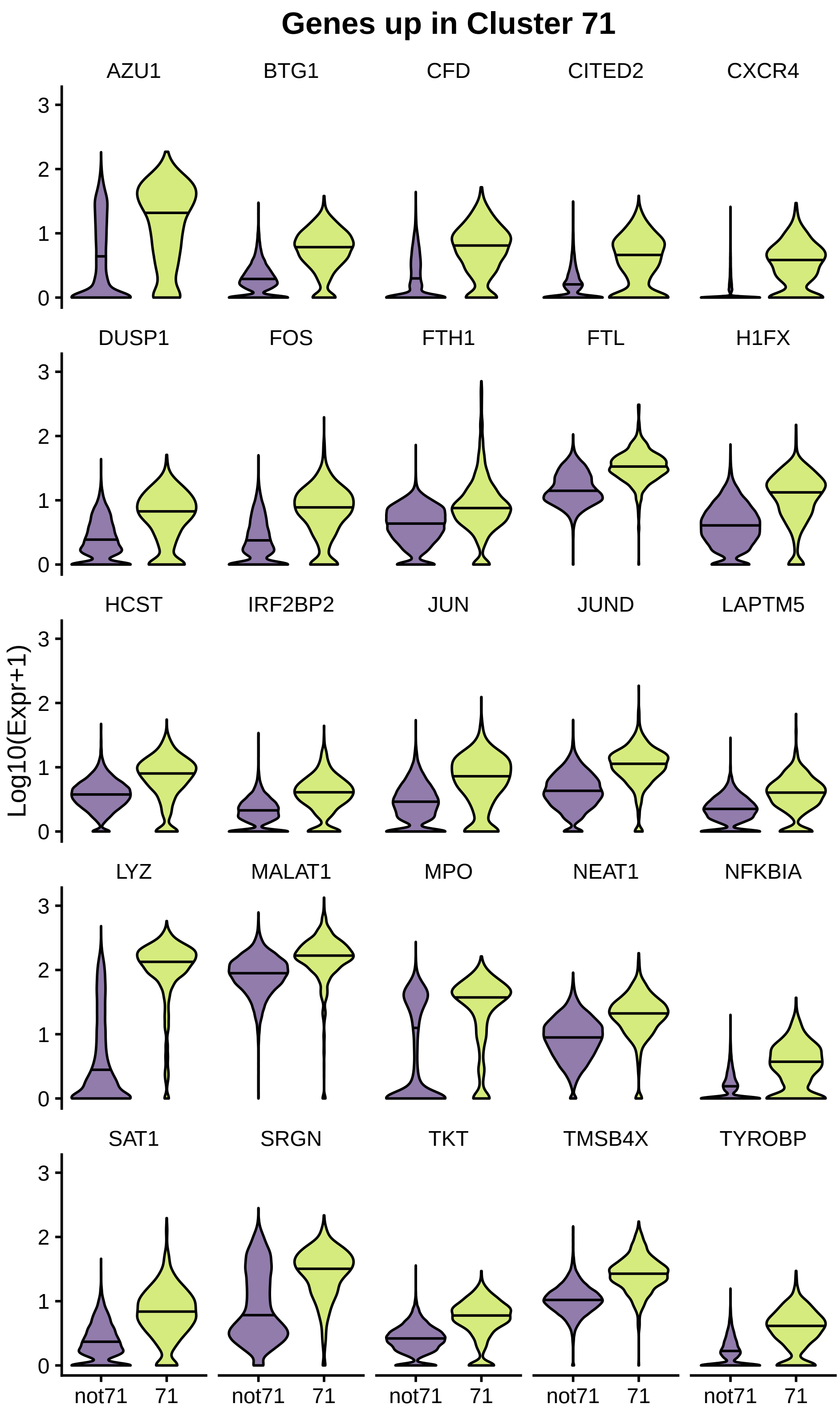
We bin the clusters into "71" and "not71":

```
plot(blast1_clusters_binary)
```

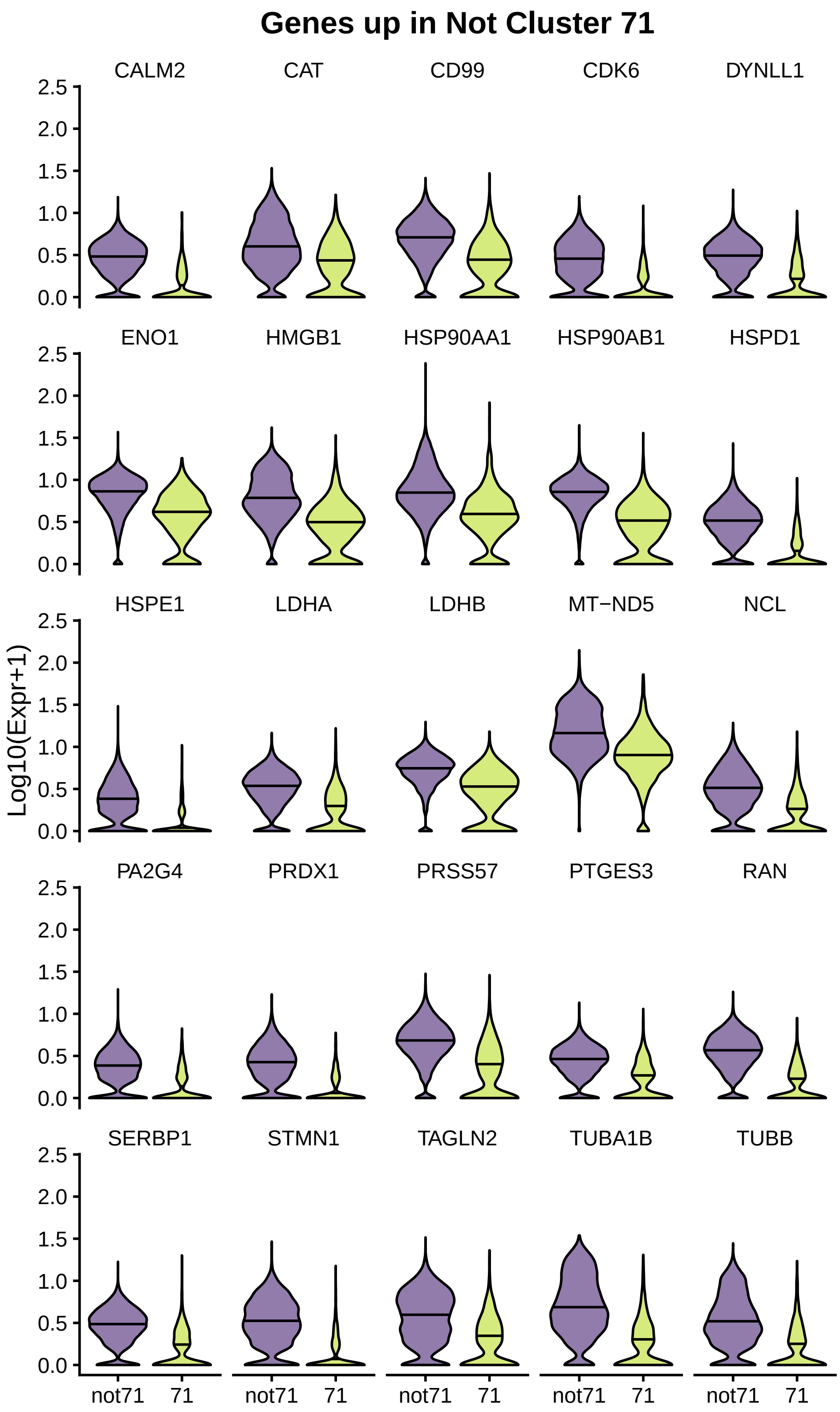


Then we can calculate the top 25 or more markers for cluster 71 and not cluster 71, using the same method described above. These are all highly significant.

```
plot(violin_up_in_71)
```

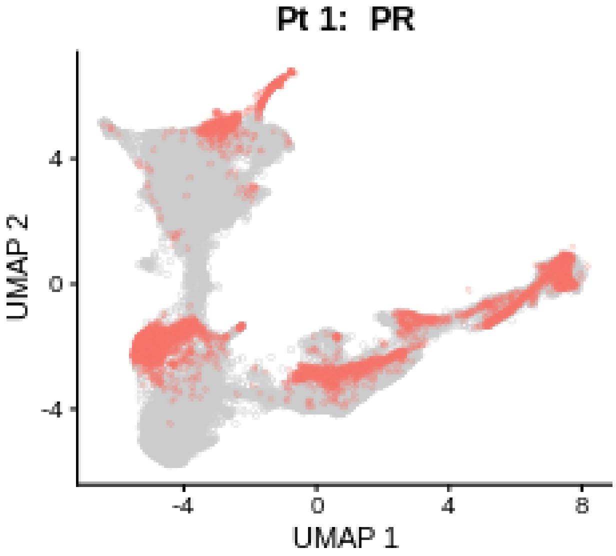

Genes up in Cluster 71
AZU1
BTG1
CFD
CITED2
CXCR4
DUSP1
FOS
FTH1
FTL
H1FX
HCST
IRF2BP2
JUN
JUND
LAPTM5
LYZ
MALAT1
MPO
NEAT1
NFKBIA
SAT1
SRGN
TKT
TMSB4X
TYROBP
Log10(Expr+1)
0
1
2
3
not71
71

```r
plot(violin_up_in_not71)
```

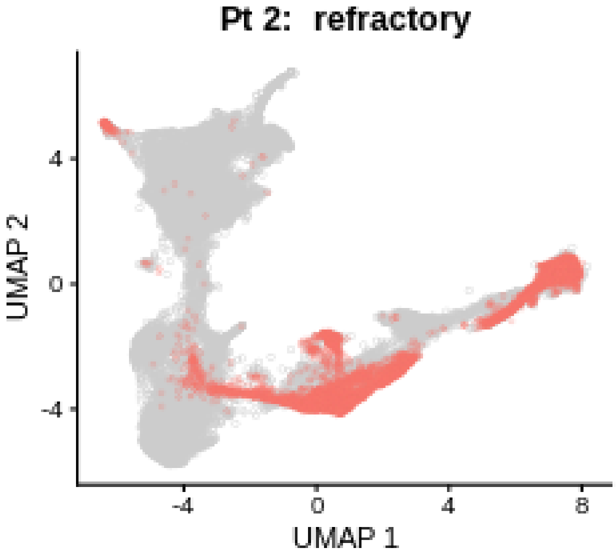


Now let's repeat this analysis within the AML/Progenitor partition. Breaking this partition down by patient:

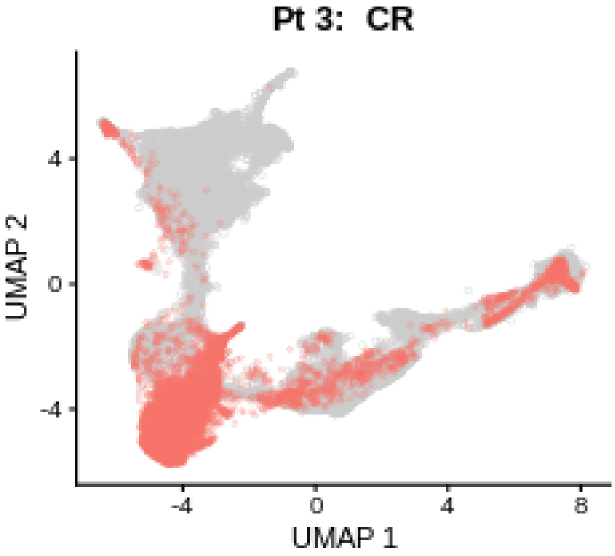
```
plot(amlprog_by_pt_list[[1]])
```



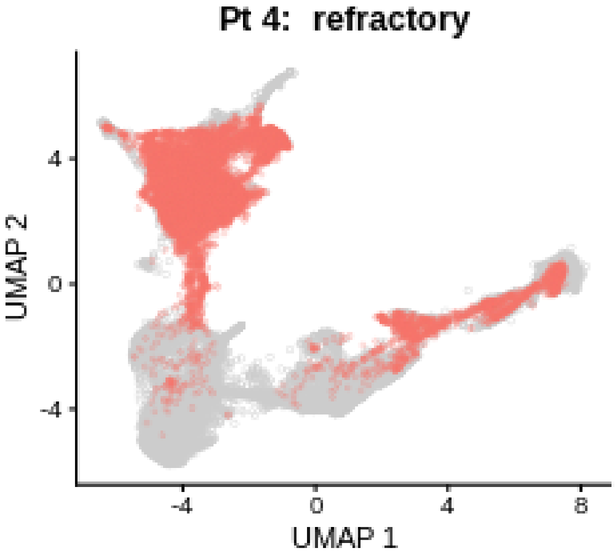
```
plot(amlprog_by_pt_list[[2]])
```



```
plot(amlprog_by_pt_list[[3]])
```

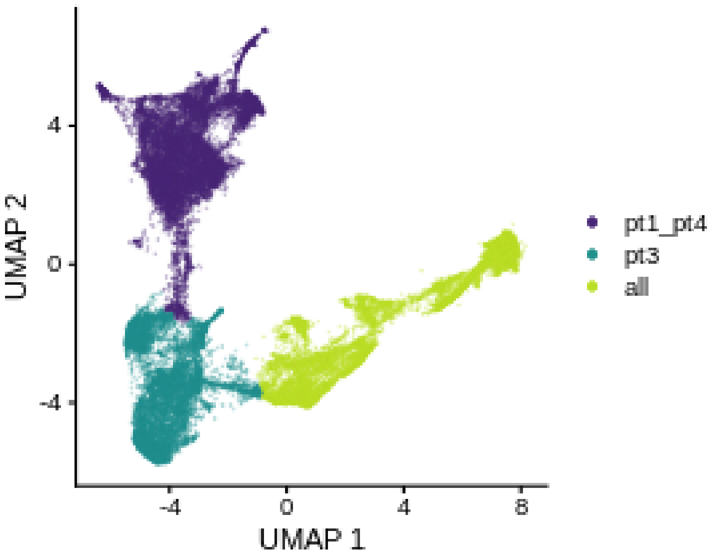


```
plot(amlprog_by_pt_list[[4]])
```



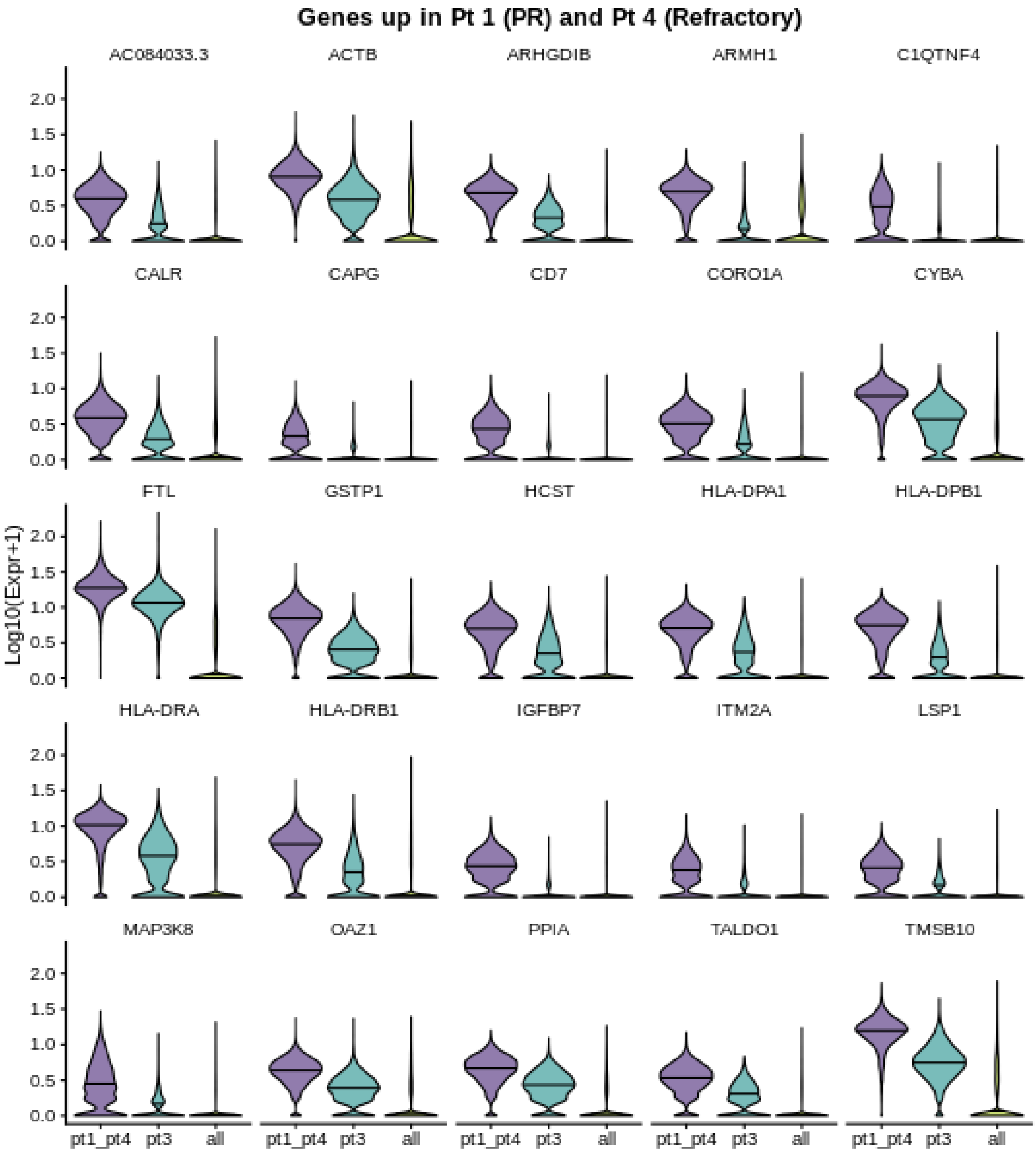
Patient 2 isn't so helpful. But it looks like Patient 1 and Patient 4 (PR and refractory) have a different pattern compared to Patient 3. I binned the clusters accordingly:

```
plot(amlprog_clusters_trinary)
```

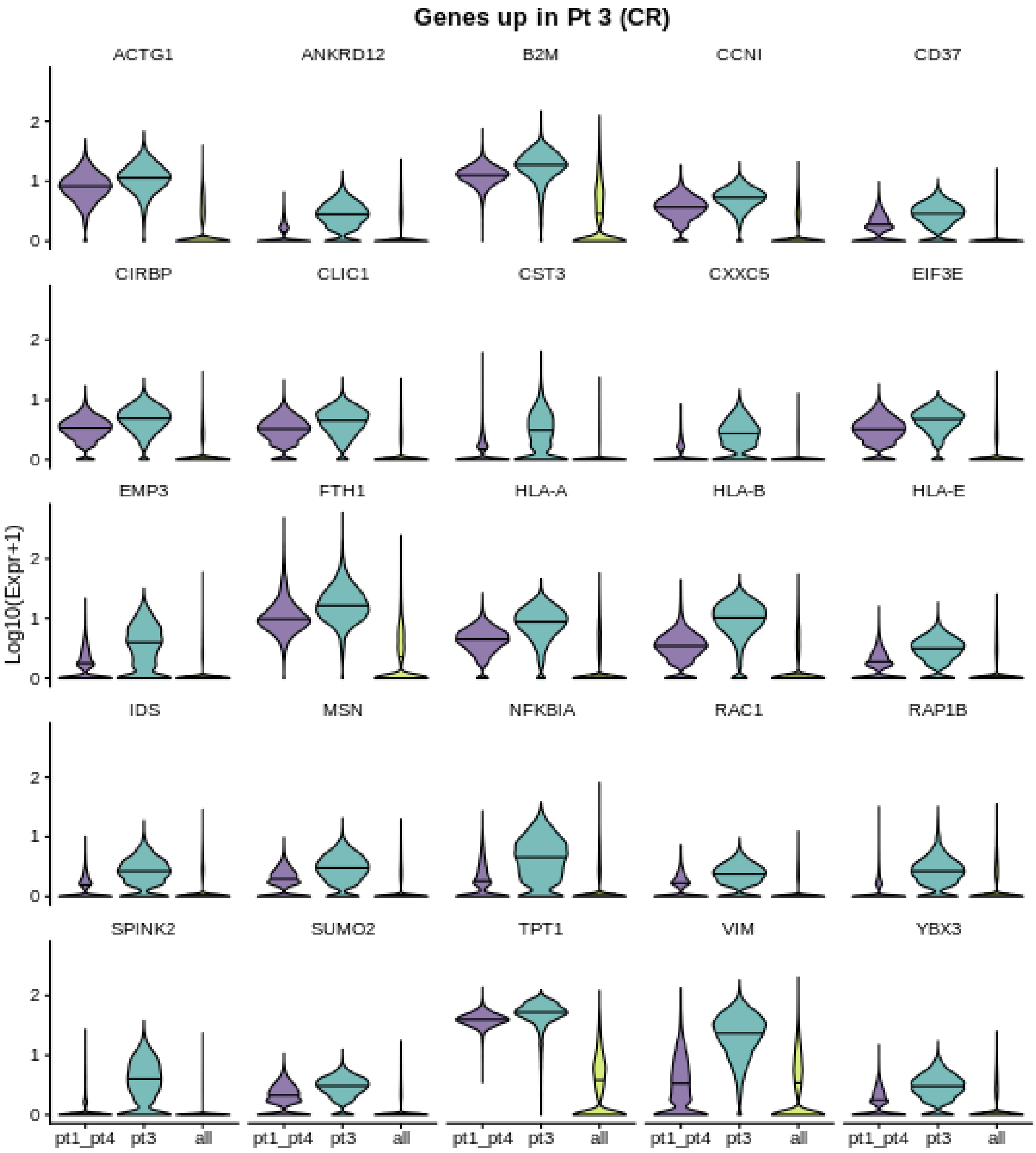


Then we can calculate the top markers in each sub-partition (or meta-cluster if you prefer):

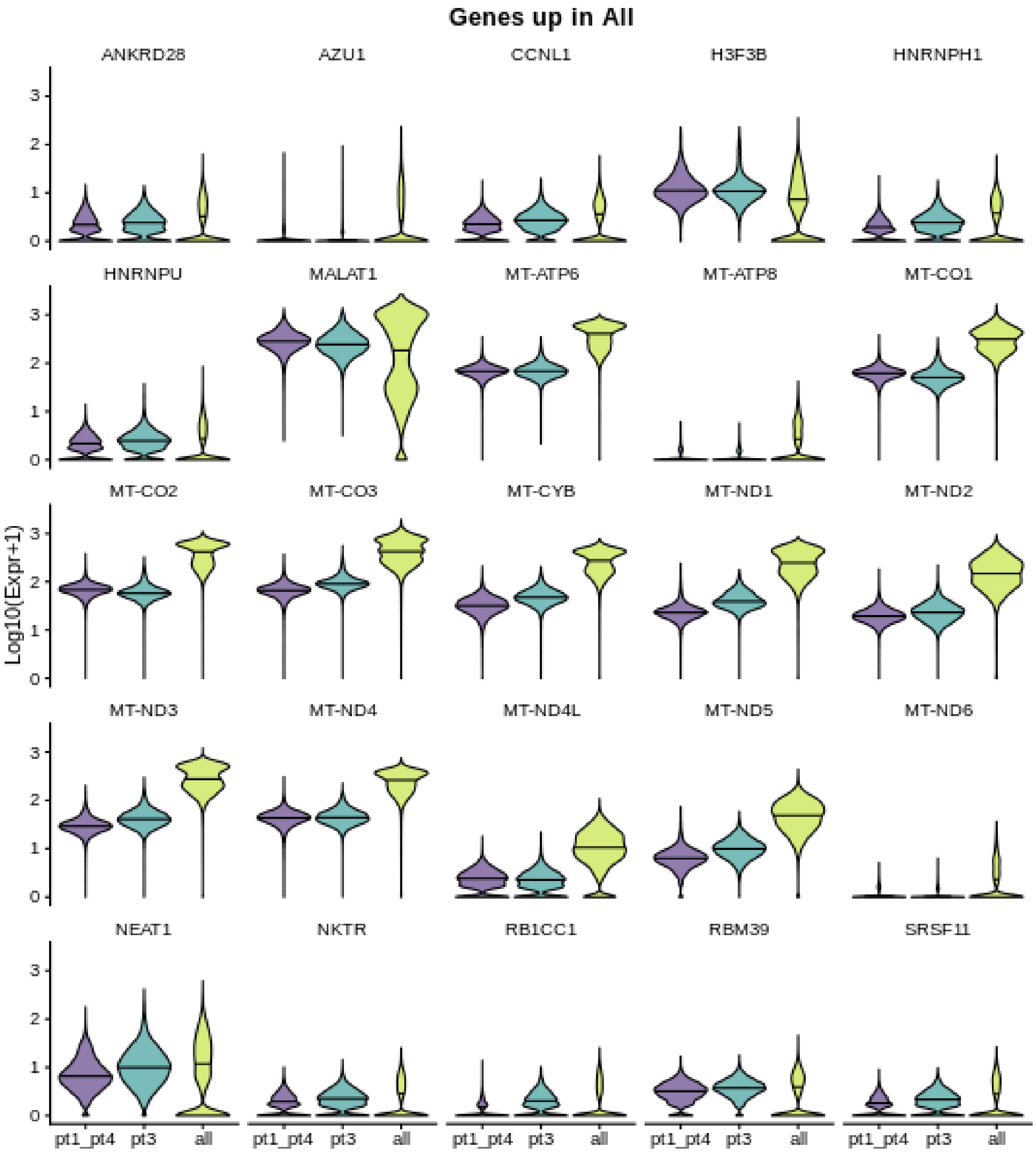
```
plot(violin_up_in_pt1_pt4)
```



```r
plot(violin_up_in_pt3)
```



```r
plot(violin_up_in_all_amlprog)
```



One obvious thing to notice is that Class I is relatively down in Pt 1 and 4 and relatively up in Pt 3. Not sure if you were expecting immune evasion, but maybe this is one thing that is happening. Another thing is that the "Up in All" plot shows lots of MT-genes. This means these are poor quality cells and we should ignore.

I've calculated the top 100 markers for each of the sub-partitions in Blast 1 and AML/Progenitor. They are in the data outs folder.

Summary

There are a lot of cells here. So as with many of these analyses as we have now done, it is easy to find certain signficant differences between patients, but since we only have a few patients (4 is still on the low end for this type of study), the major challenge is finding a consistent theme that applies to more than 1 patient. This analysis incorporates some of the more useful stratification techniques I've come across. Machine learning techniques are now being applied to these kind of data; I'm not sure how much more there is to discover with these techniques beyond sitting and spending time with the data. My impression after a few days is that there are some signatures of immune/myeloid cell development and immune evasion that covary with the treatment response. Once you have a chance to look at things, let me know what you think, how it fits into the rest of the project and I can help you go from there.

Once we have the ITD data it could provide another way to identify the malignant clone. Unfortnately it will be little while with the hiatus we are currently on.

Feel free to contact me any time to discuss.