Cell Classification Benchmarking Using Data Reduction Methods

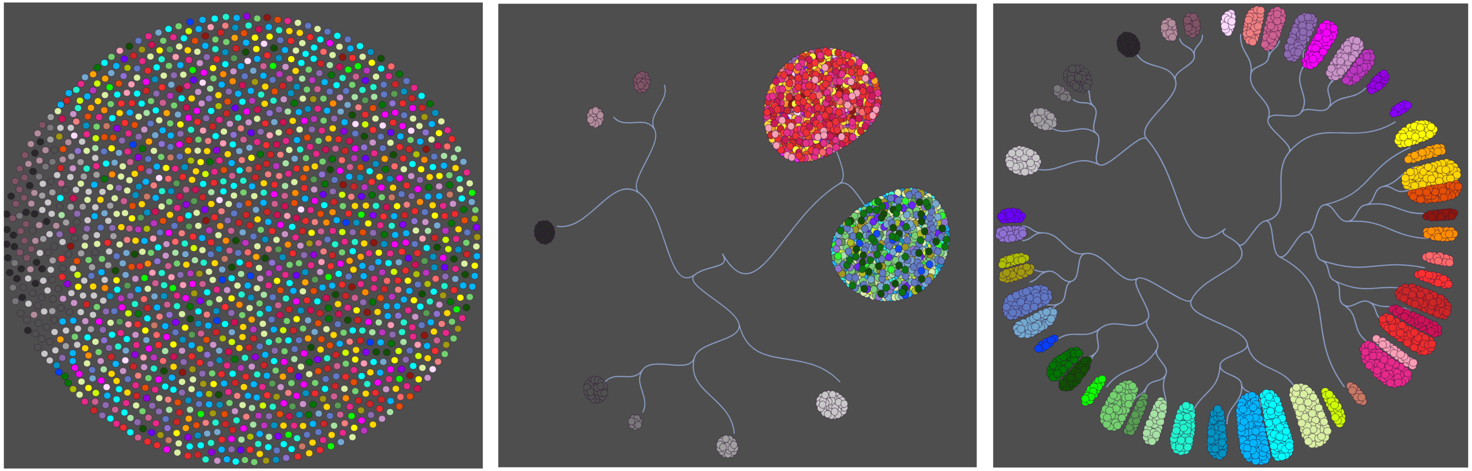
GW McElfresh

9/15/2020

Dataset and Motivation

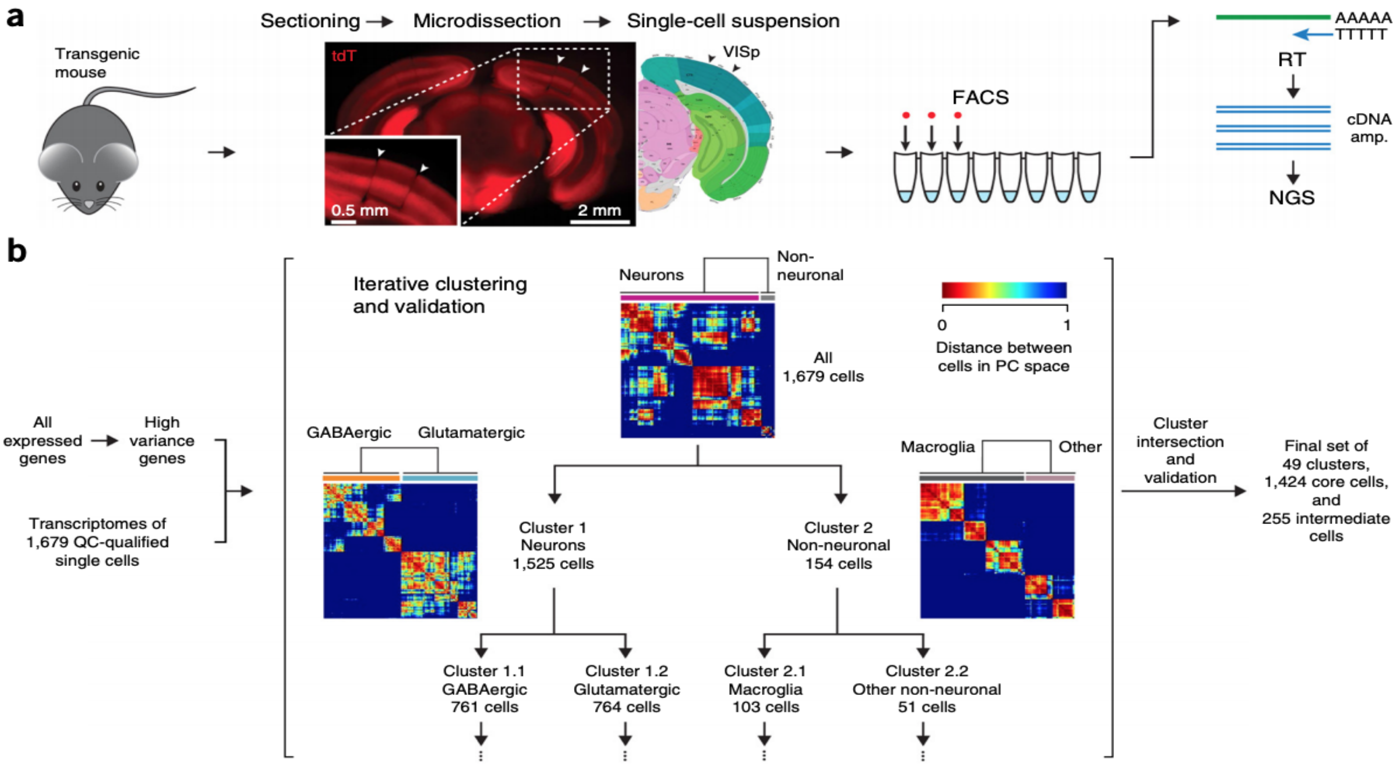
The Tasic Brain Data set is a set of scRNASeq expression profiles from ~1600 brain cells from adult mice.

The original analysis yielded an impressive cell taxonomy
(best viewed at casestudies.brain-map.org/celltax)



Dataset and Motivation

Tasic *et al.* used a novel method of determining cell type from scRNASeq expression profiles

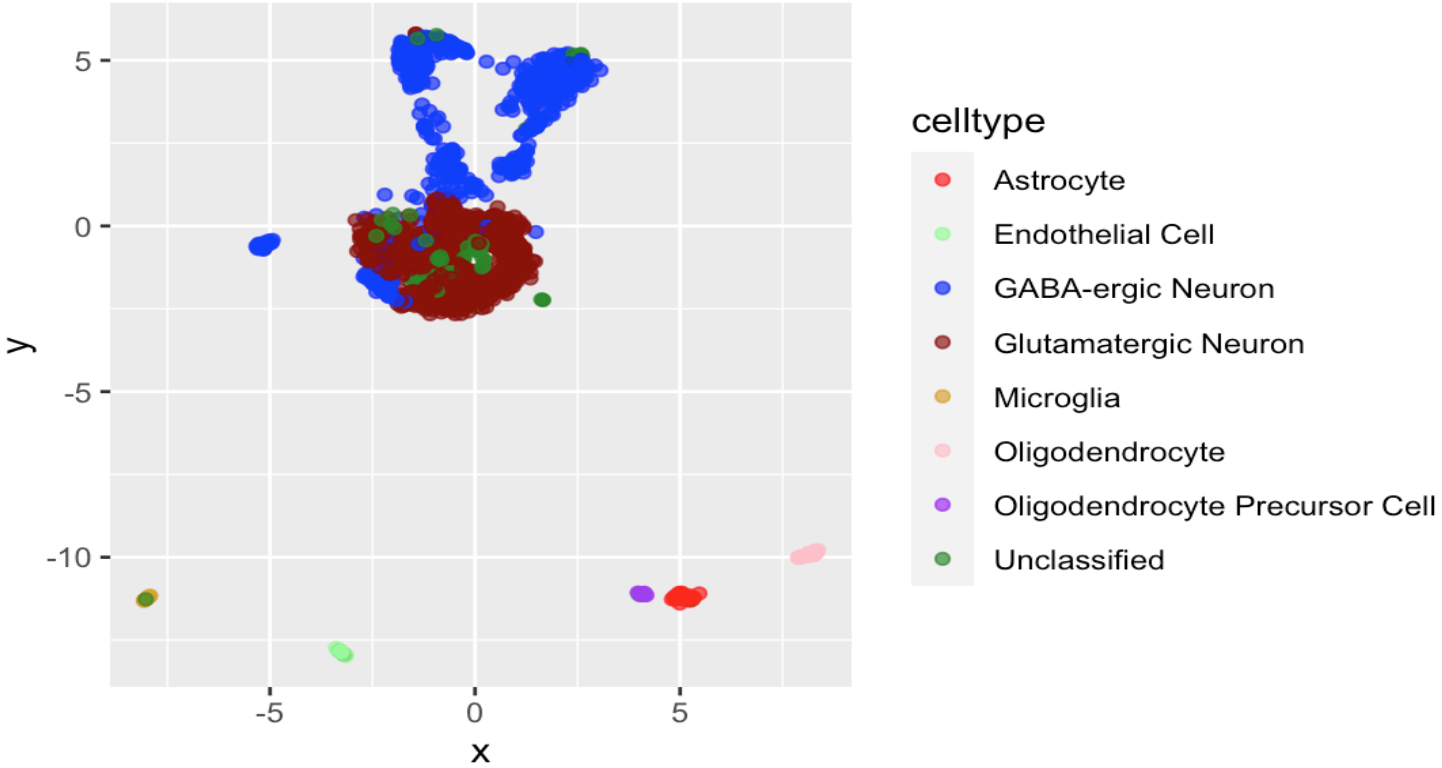

a
Transgenic mouse
Sectioning
Microdissection
Single-cell suspension
tdT
0.5 mm
2 mm
VISp
FACS
AAAAA
TTTTT
RT
cDNA amp.
NGS
b
All expressed genes
High variance genes
Transcriptomes of 1,679 QC-qualified single cells
Iterative clustering and validation
Neurons
Non-neuronal
All 1,679 cells
0
1
Distance between cells in PC space
GABAergic
Glutamatergic
Cluster 1 Neurons 1,525 cells
Cluster 2 Non-neuronal 154 cells
Macroglia
Other
Cluster 1.1 GABAergic 761 cells
Cluster 1.2 Glutamatergic 764 cells
Cluster 2.1 Macroglia 103 cells
Cluster 2.2 Other non-neuronal 51 cells
Cluster intersection and validation
Final set of 49 clusters, 1,424 core cells, and 255 intermediate cells

Tasic *et al.* 2016

The Goal

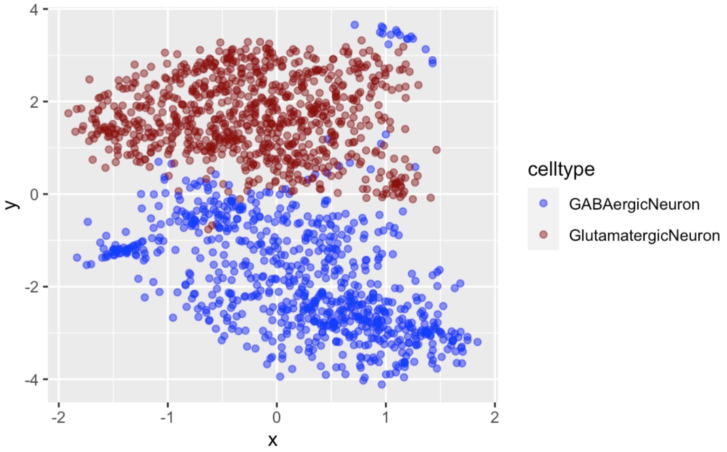
The original analysis was able to determine 49 unique transcriptomic profiles, but the methodology was unable to cluster 82 of their cortex cells.

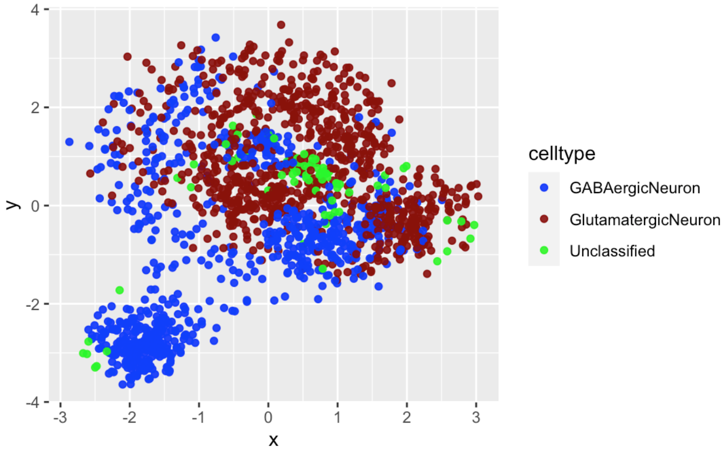
Looking at the data, we see that the unclassified data fall into established clusters, but may be difficult to classify correctly.



The Goal

We can see this more clearly when we look at just the neuron data, where including the unclassified data smears the clusters together.





Is there a way to classify these cells?

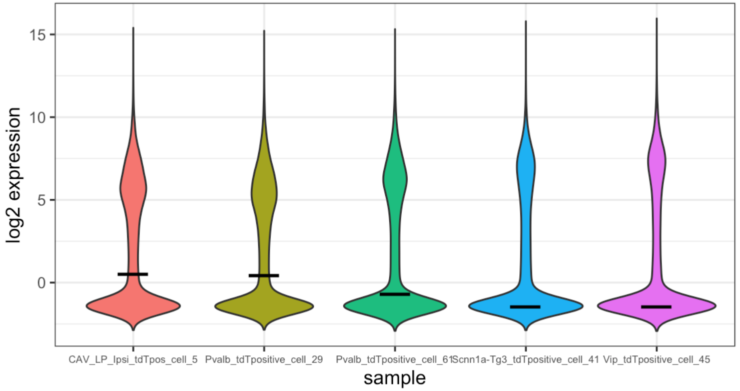
Hypotheses

The unclassified cells may be able to be re-classified by post-hoc machine learning from the classified dataset

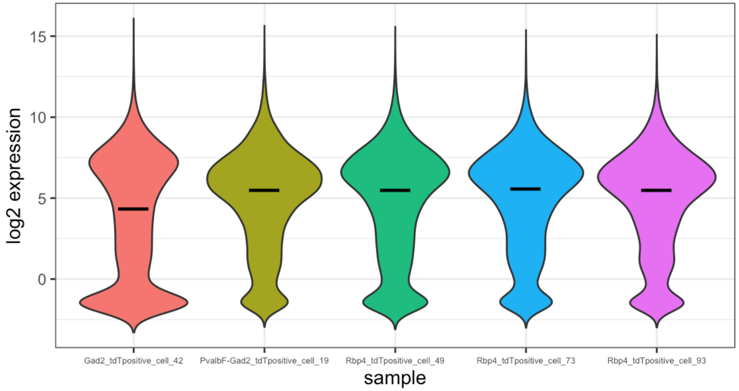
- Newer clustering methods (UMAP) may be more robust than PCA
- Gene subsets (differential expression) may be able to discriminate the data more successfully

Filtering The Data

Filtering for lowly expressed genes yields violin plots that *might* work to discriminate the data, but filtering and selecting just the neuron cells is more promising



Randomly Sampled CPMs (All Cell Types) (log2)
filtered, normalized (TMM)



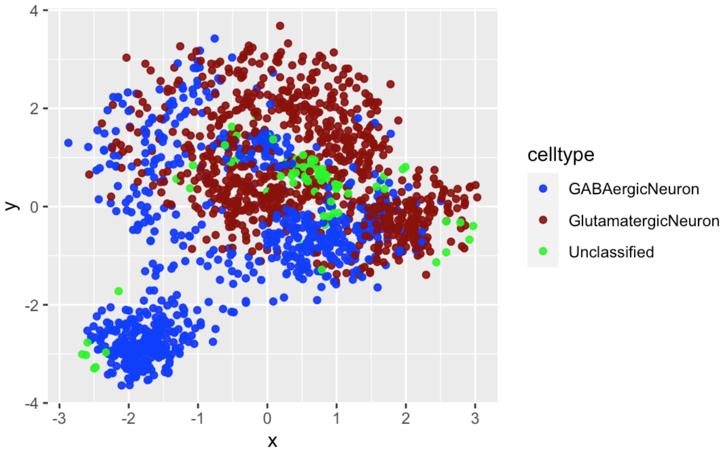
Randomly Sampled CPMs (Neurons Only) (log2)
filtered, normalized (TMM)

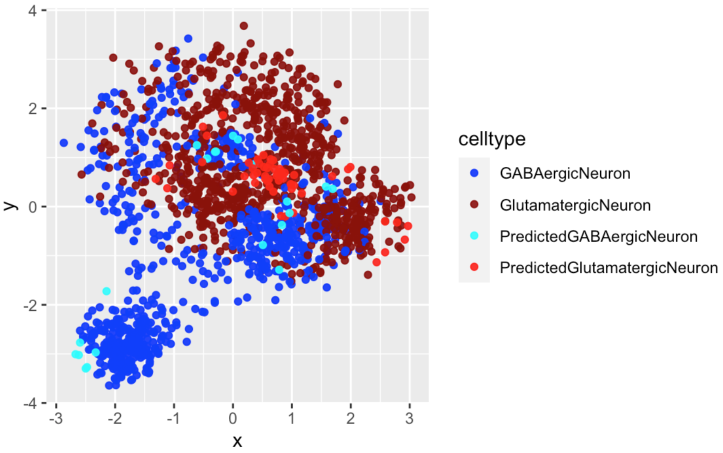
Filter criterion:

- average CPM > 1 per sample

Hypothesis 1: Training a Classifier on UMAP Projection

Now that the expression matricies are less sparse, let's train a random forest and see if the filtering and normalization are enough to cluster the data

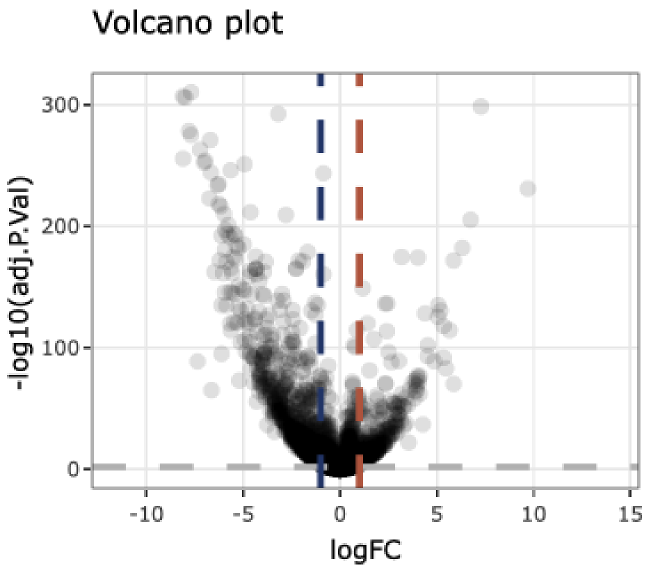




	GABAergicNeuron	GlutamatergicNeuron
GABAergicNeuron	138	19
GlutamatergicNeuron	21	137

Hypothesis 2: Training a Classifier on Differential Expression Data

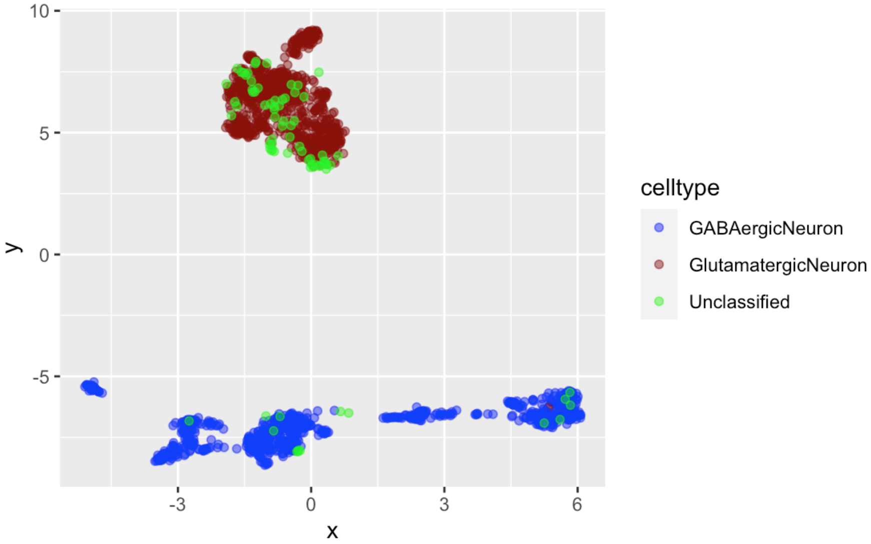
Differentially expressed genes are generally considered to be indicative of different cell types, so that may sufficient to discriminate between the clusters.

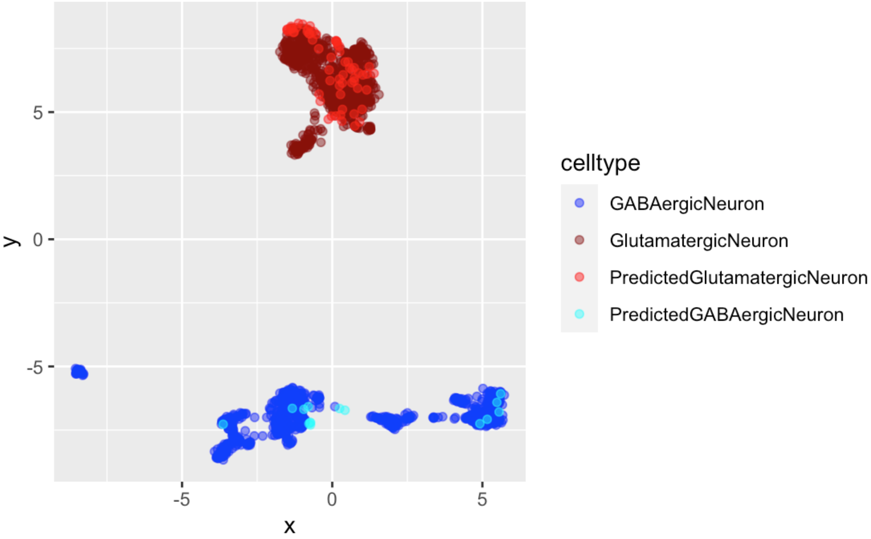


- By performing differential expression on the different neuron types, we drastically reduce the number of genes
- The reduced gene set of ~600 genes - which should be large enough for machine learning

Cell Type Classification on Differentially Expressed Genes

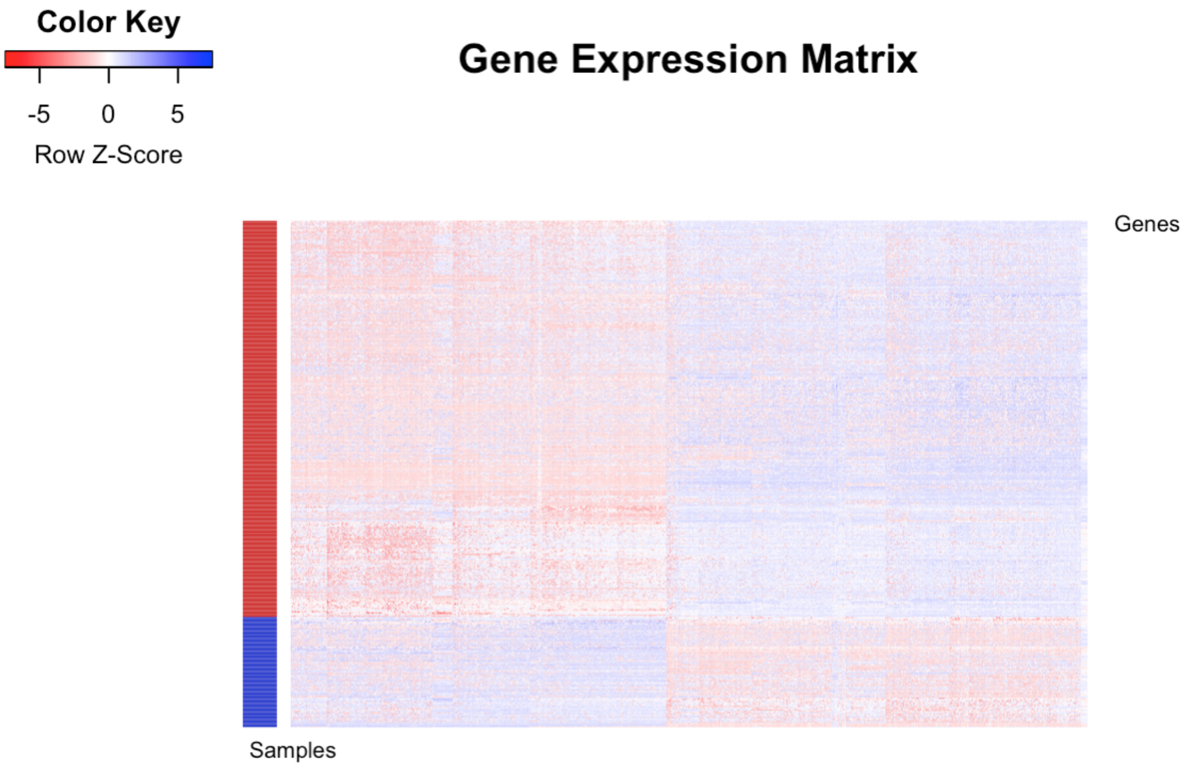
	GABAergicNeuron	GlutamatergicNeuron
GABAergicNeuron	253	0
GlutamatergicNeuron	0	272





Further Analysis

The differential expression analysis shows that the cells form two modules (upregulated genes and downregulated genes)



Further Analysis

Classification based on GO term enrichment is a promising avenue

- If the differentially expressed genes do not cluster the data well, GO terms might!
    - For these, linear regression (weights learned from GO term abundances) could be a solution

GSEA is also a regress-able data source, but this is likely to train an incredibly sensitive classifier suitible for hyper-specialized sub-cell types

Conclusions

- Differential gene expression may work to determine intermediate cell types where whole transcriptome profiles are unclear
- For more homogeneous data, higher order analyses (GO term enrichment, GSEA) may be required, but at the cost of machine learning sophistication