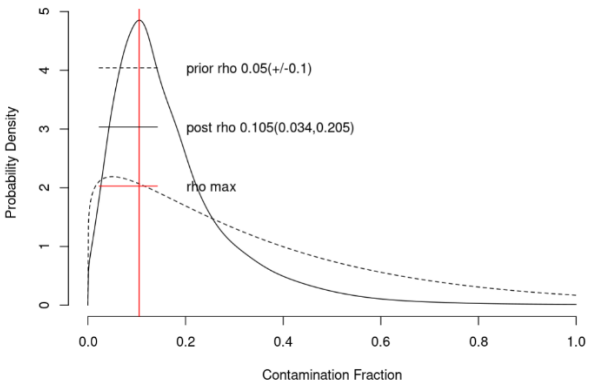
Mikhail Kouzminov

Perrimon Lab Pedro Project report

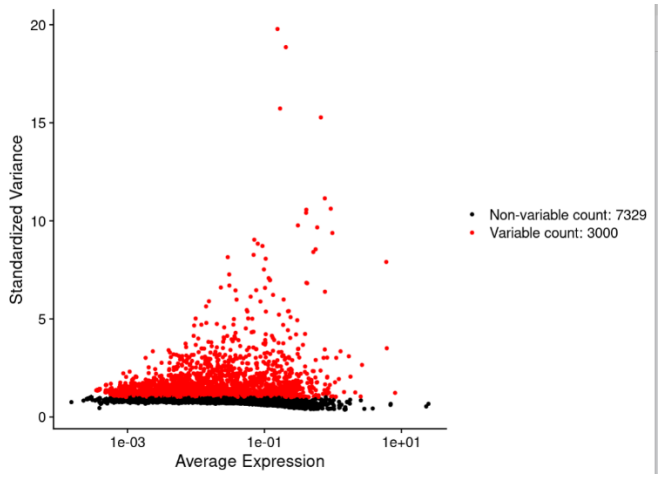
This is a report on example of my analysis fly single cell data for a study by The Perrimon Lab. Sequencing was done for a project and paper led by a postdoc named Pedro Saavedra, so samples in this initial analysis were named after him. Samles Pedro3 and Pedro4 came from flies with inhibited translation of the foxo gene and Samples Pedro 5 and 6 had come from flies with inhibited translation for the reptor gene using RNA interference.

Data was sequenced using Illumina's platform and cellranger. In this case, my analysis largely involved preprocessing the data.

Specifically we were interested in the expression of gene networks and the differences in their expression between various treatments that inhibited the foxo and reptor genes. This particular sample was suspected to have a heavy amount of cell-free contamination so some additional preprocessing had to be done with SoupX. SoupX predicts the contamination fraction to be about 10 % of data given, which was then cleaned.

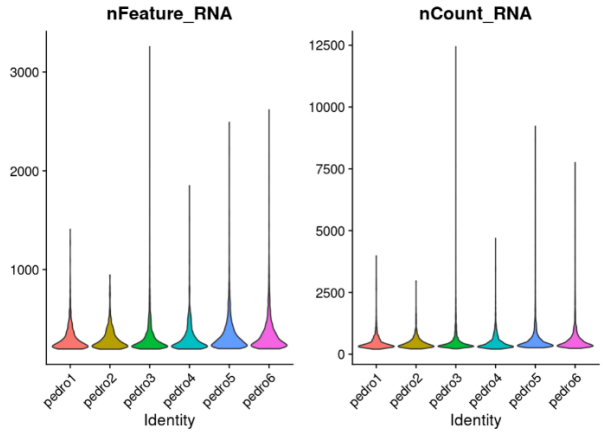


Data was then processed to find cells with highly variable genes.

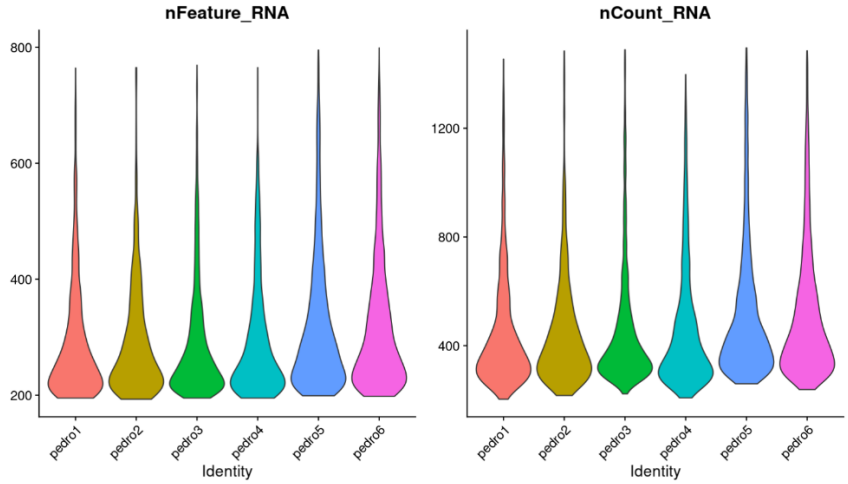


In order to account for damaged cells, I filtered those with an extraordinary amount of UMI barcodes and an extraordinary amount of genes found.

Feature plot before filtering:

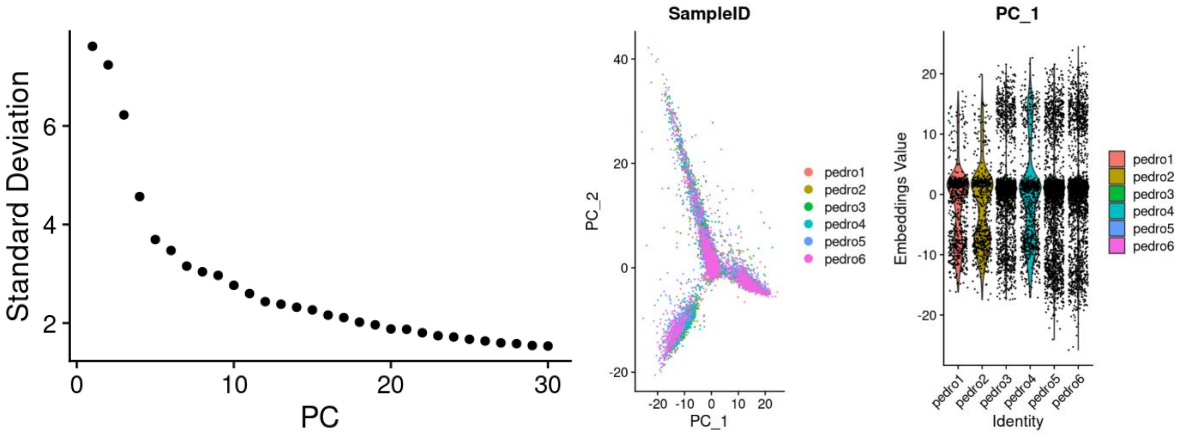


Feature plot after filtering:

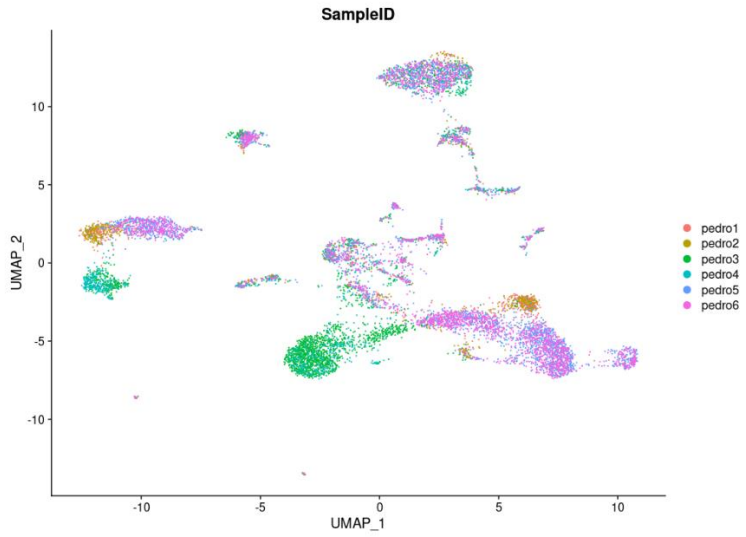


In order to check that the principal components accurately represented all samples, I manually checked the top principal components and found that cells tended to be relatively widely spread across all combined samples in this study.

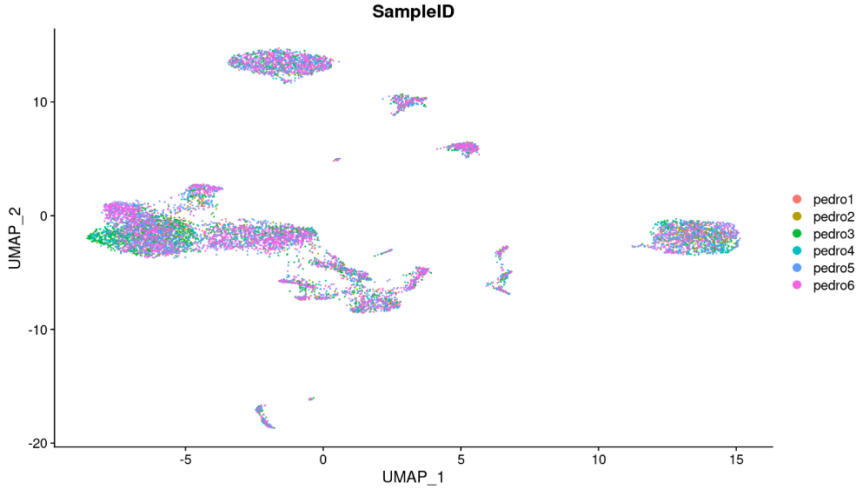
The L bend where an increase in principal components was judged to safely be around 15. Additionally, several of the most representative components were checked to accurately represent the data across all samples.



UMAPS of the data were then created, checking first that samples with similar conditions were clustered together:

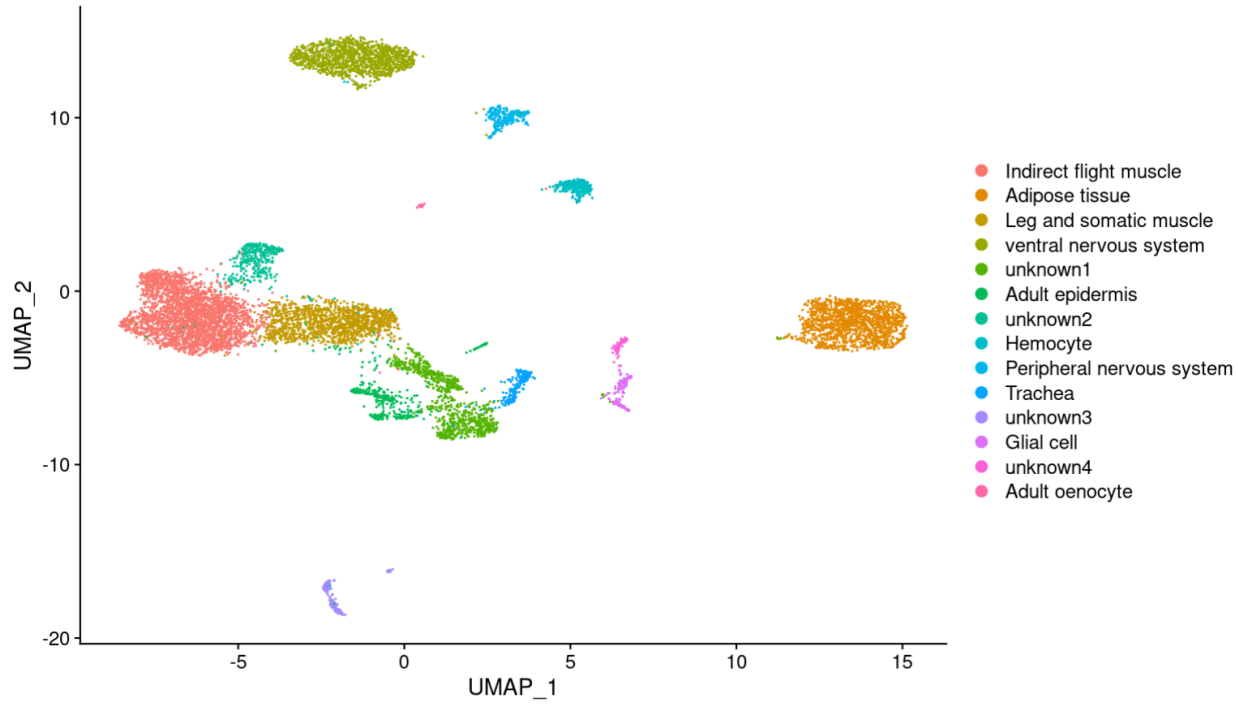


And then using the HARMONY algorithm to integrate the data across various conditions.

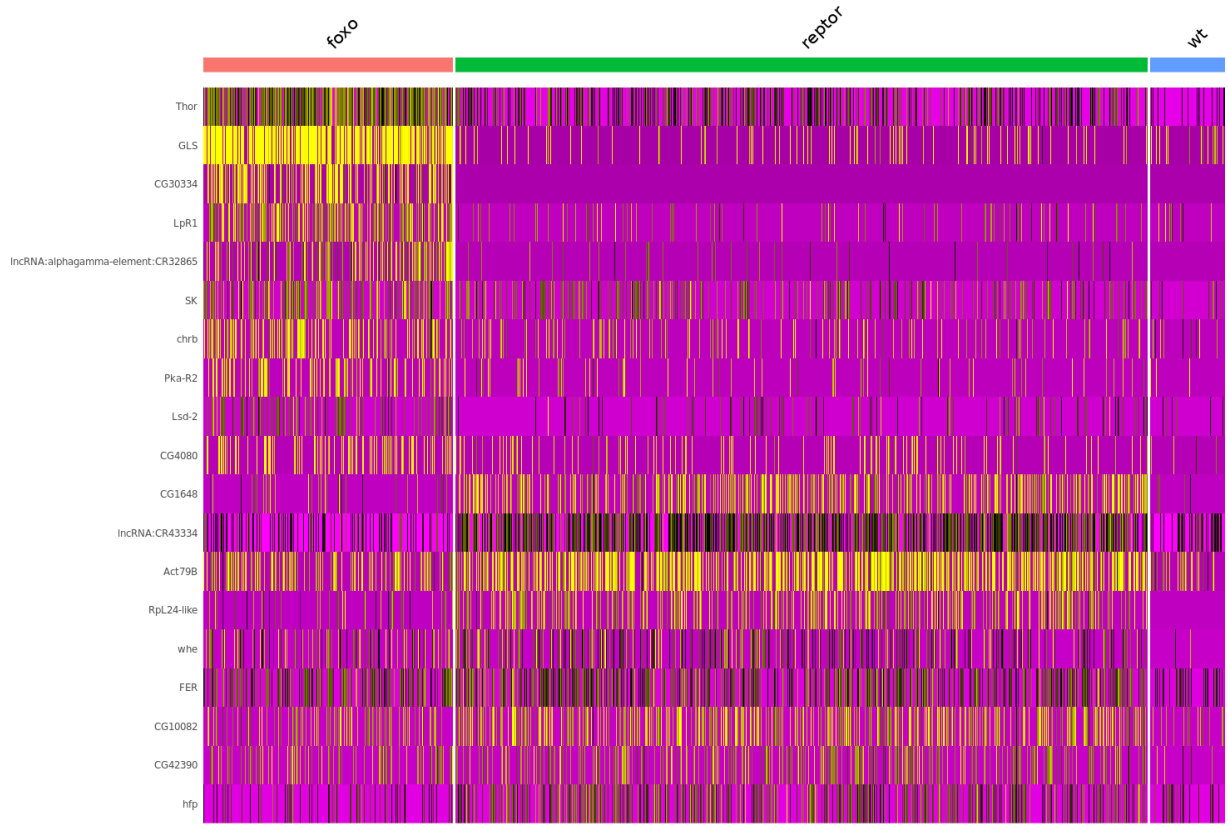


Sample clusters were then identified using marker genes:

Genes of interest were in large part identified by Pedro, the postdoc in charge of the experiment, who was more familiar with specific genes of interest. I looked up some genes in a preliminary analysis, but he did a large amount of the identification.



Of particular interest was the Thor gene which we thought would be highly expressed as a sign of particular stress, but wasn't in this case possibly due to the contamination that we attempted to use SoupX to account for. Note that this diagram is specifically a heatmap of cluster 2: leg and somatic tissue, which was considered of interest. Heatmaps of other clusters are available.



If this study had continued beyond this point, I would likely have used SCENIC and GENIE3 to analyze related gene networks, as I did for similar studies – as included in my code sample.

I also used Monocle 3 for pseudotime analysis for some projects.