Pituitary corticotroph gene expression analysis

Ho et al 8week female scRNA seq dataset

Craig McDougall

October 2020

Introduction

This document details the analysis of the single RNA sequence data by Ho et al. available from NCBI.

Step 1. Call Libraries & Set Directories

Before the main body of the pipeline is run, all of the required libraries are called and

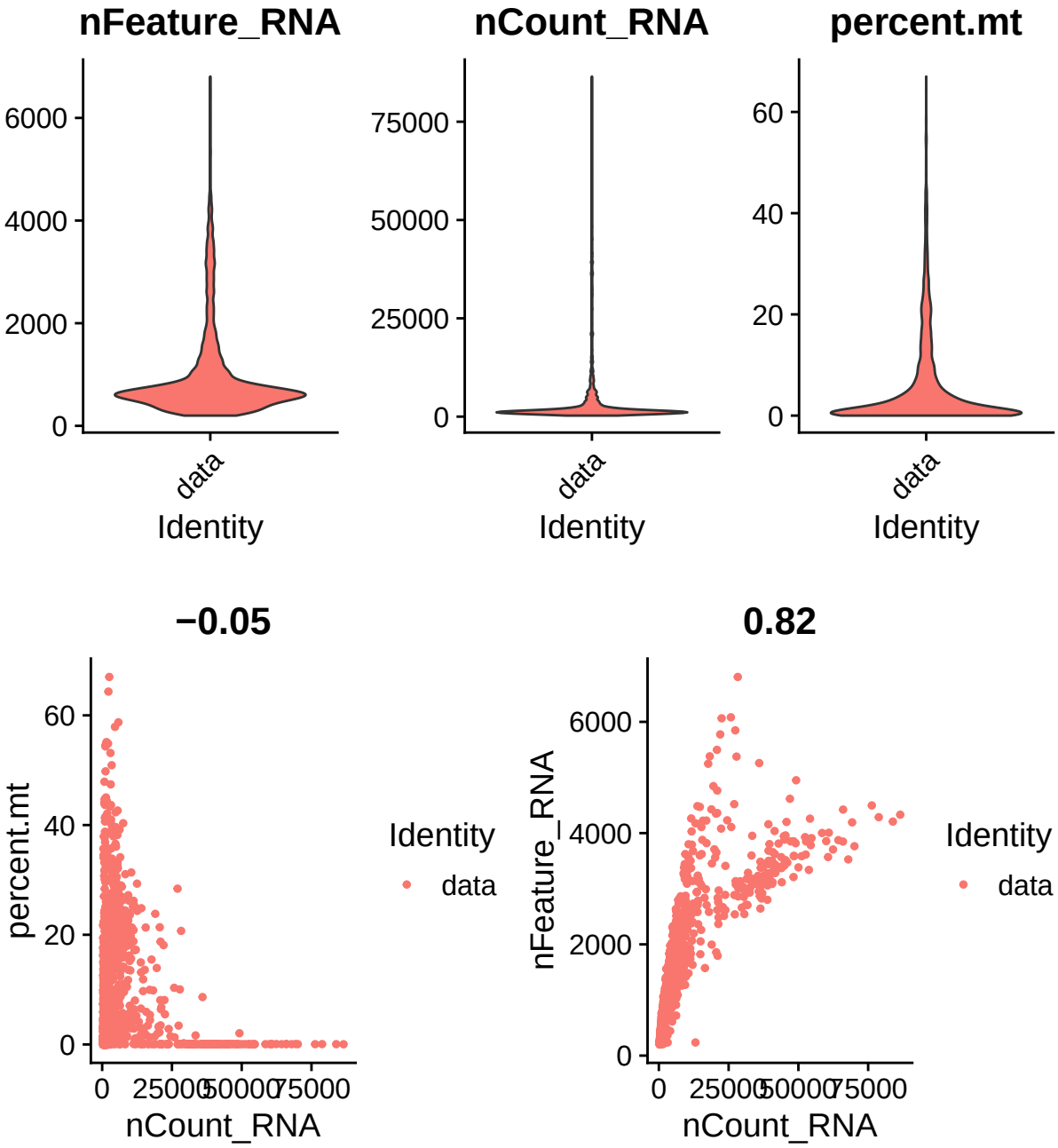
the working directories are then set.

The raw count and gene data is read in from the 10X sequence data files described in Ho et al and downloaded from Ho et al.

For the Ho analysis, the Ho folder is indicted in the path, with the correct sex chosen.

Step 2. Quality control plots

Before processing the data, some quality control plots are made to assess the quality of the data and explore confounding factors such as assessing the number of reads that map to mitochondrial genome.



The QC plots to inform which cells should be excluded on quality basis (low quality, double counts). There appear to be separate features in the feature data, so all cells below 200 counts and above 2000 are dropped and cells with >5% mitochondrial DNA are also dropped.

Step 3. Transform the data

With the data trimmed to exclued lower quality cells, it must be;

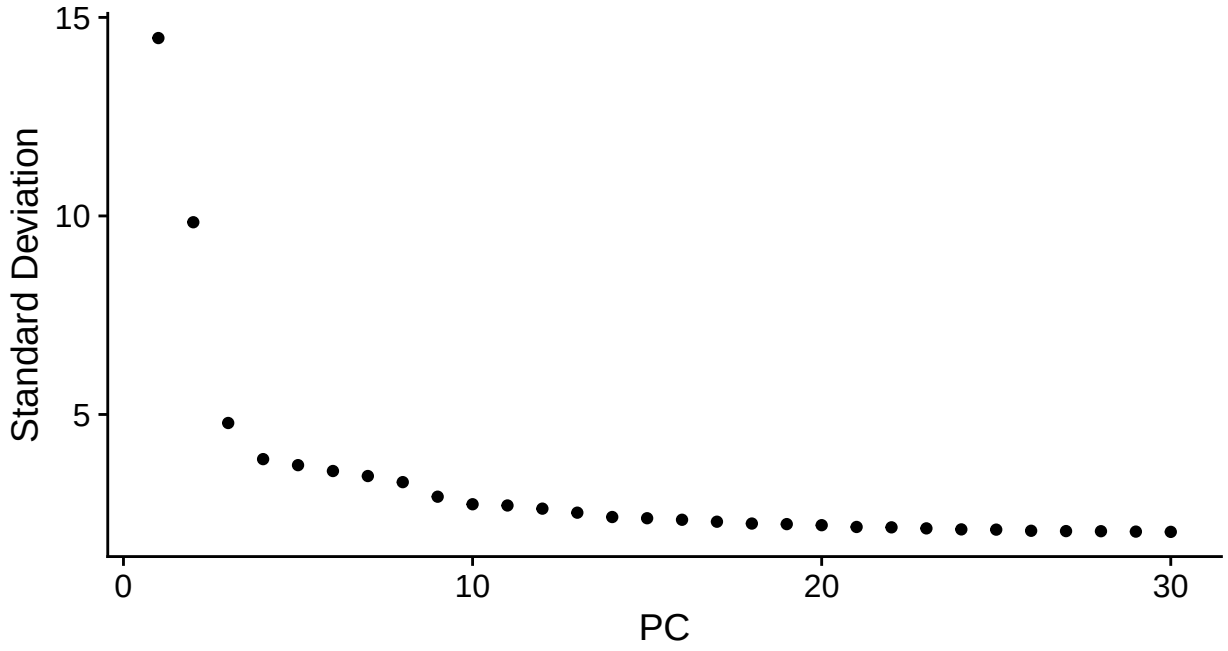
- Normalised to ensure all genes carry equal weight going forward and prevent heavily expressing genes from being over represented.
- Regressed to remove confounding factors such as areas mapped to the mitochondrial genome.
- Assessed for variable genes.

The SCTransform function performs all of these actions and greatly simplifies the code, negating the use of four separate functions.

Step 4. Dimensional reduction

The first step in dimensional reduction is to conduct principal component analysis (PCA) on the scaled data, using the variable features as the input. PCA transforms the from the table into new features known as principal components, which capture the information of the dataset in a new way.

An elbow plot can be used to assess which principal components contain the majority of the information and identify which may be removed from the dataset;



The elbow plot shows that the majority of the information is captured within the first 14 Principal components, so the last 16 are dropped during the next stage.

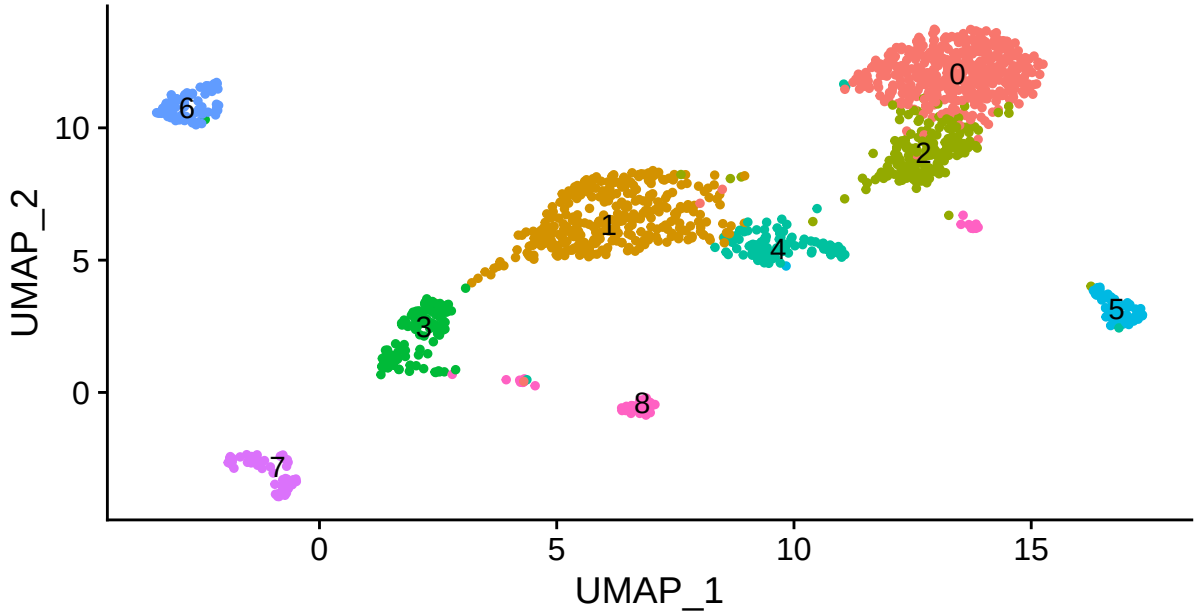
Uniform Manifold Approximation and Projection (UMAP) is an algorithm for dimensional reduction and is applied on the first 14 principal components established in the first step.

Step 5. Clustering

The first stage in clustering the cells is to establish a shared nearest neighbour (SNN) information for the data by calculating the overlap between each cell based on its k.parameters using a Seurat function.

The clusters cn then be identified based on the SNN using a clustering algorithm.

Once the clusters have been established, they can be projected into two lower dimensional space using a dimensional reduction plot for visualisation.



Step 6. Cell Type assessment

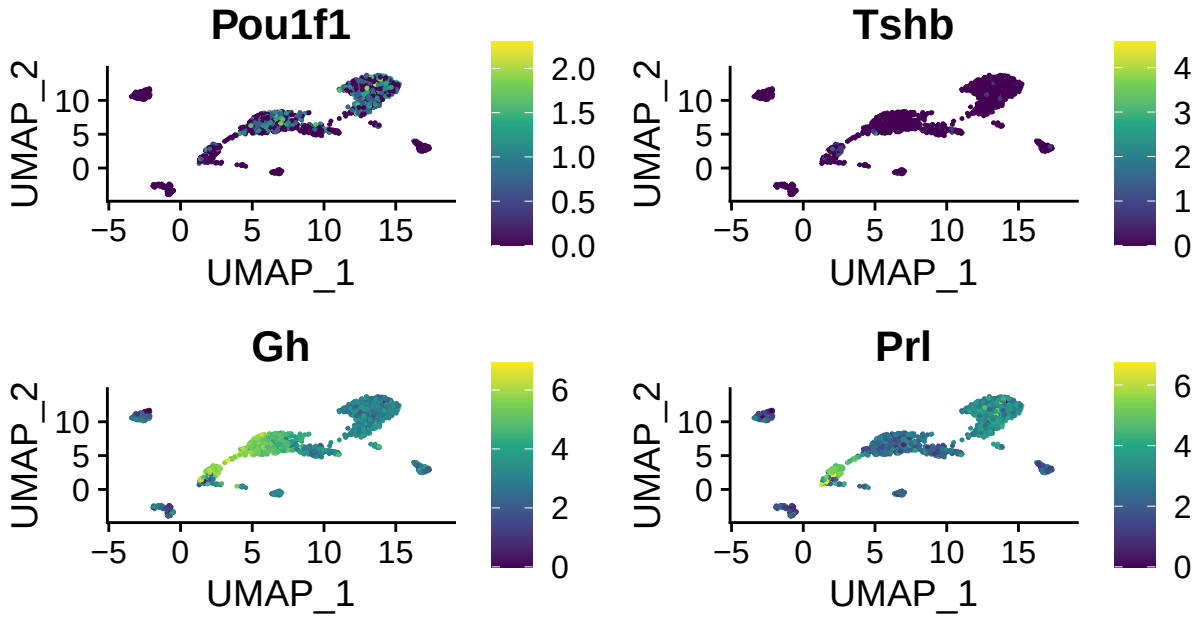
The main question now, is how well do the clusters approximate cell type? It is difficult to make absolute claims when attributing cell types to empirical cluster data based purely on an *in silico* analysis. However we may approximate the cell type of each cluster based on known canonical expression markers [Fletcher et.al 2019] [Chung et.al 2018].

In the corticotroph study, hormone secreting pituitary cell types were identified based on the following canonical gene expression markers:

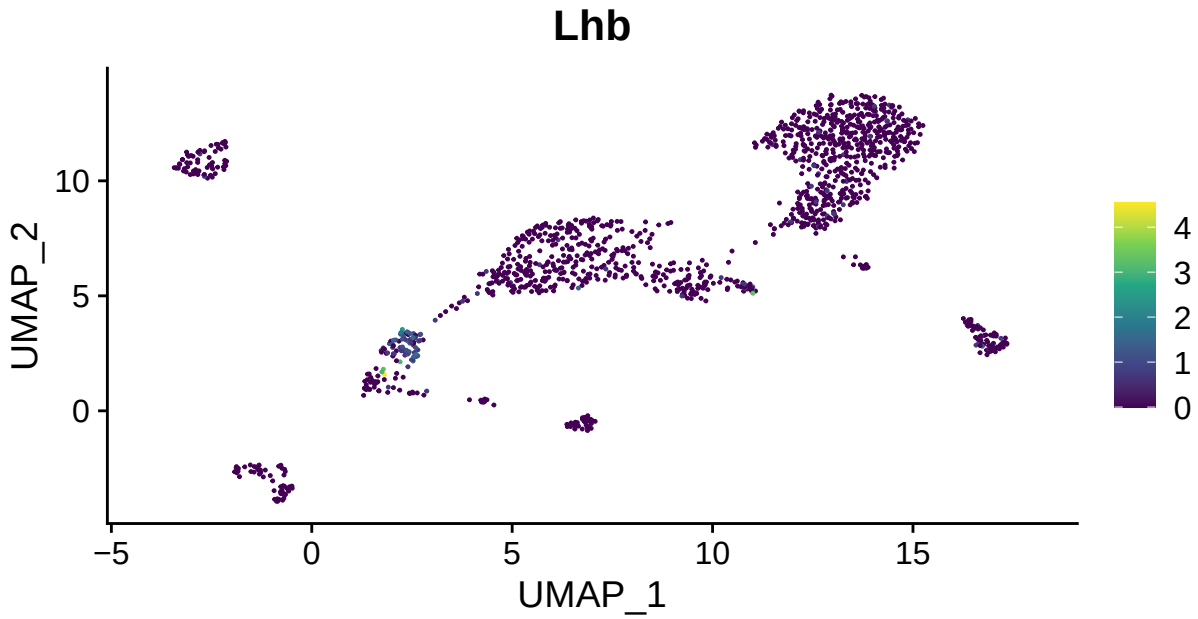
- Thyrotrophs (Pou1f1 + Tshb),
- Somatotrophs(Pou1f1 + Gh)
- Lactotrophs (Pou1f1 + Prl)
- Gonadotrophs (Lhb)
- Melanotrophs (Pomc + Pcsk2 + Pax7)
- Corticotrophs (Pomc + Crhr1 + Avpr1b + Gpc5 - Pcsk2 - Pax7)

Feature plots were used to visually assess gene expression levels of single cells. Cell type was assigned based on visual inspection of canonical markers [Fletcher et al, Chung et al].

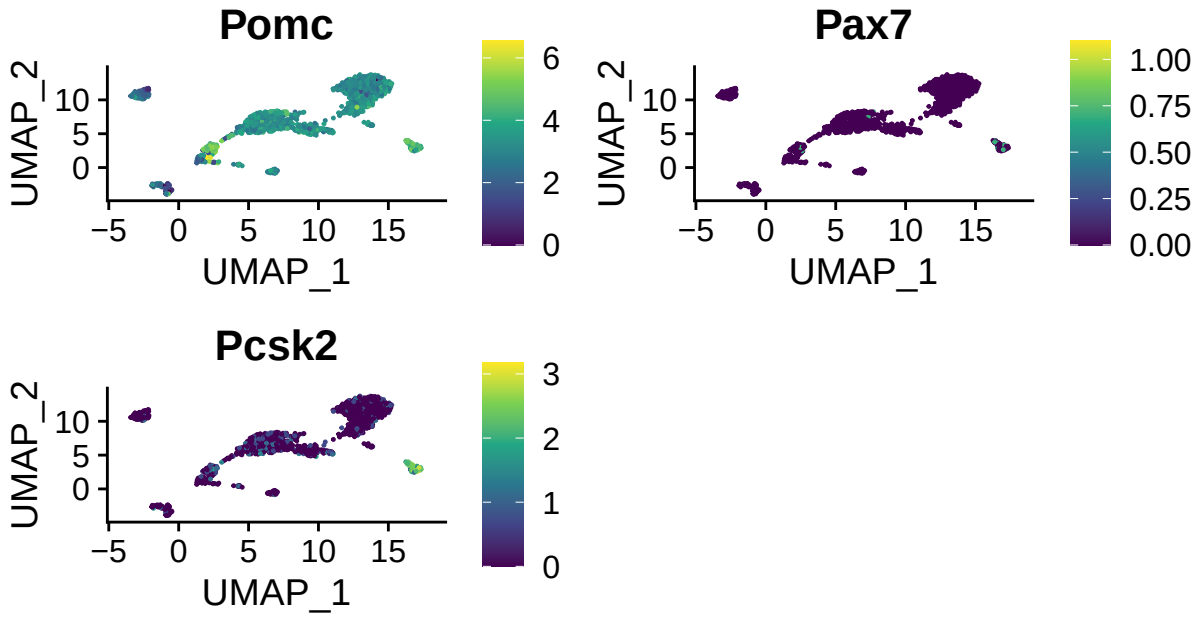
First the thyrotophs (Pou1f1 + Tshb), somatotrophs (Pou1f1 + Gh) and lactotrops (Pou1f1 + Prl) may be excluded based on gene expression:



Then the Gonadotoprhs may be excluded by Lhb expression;



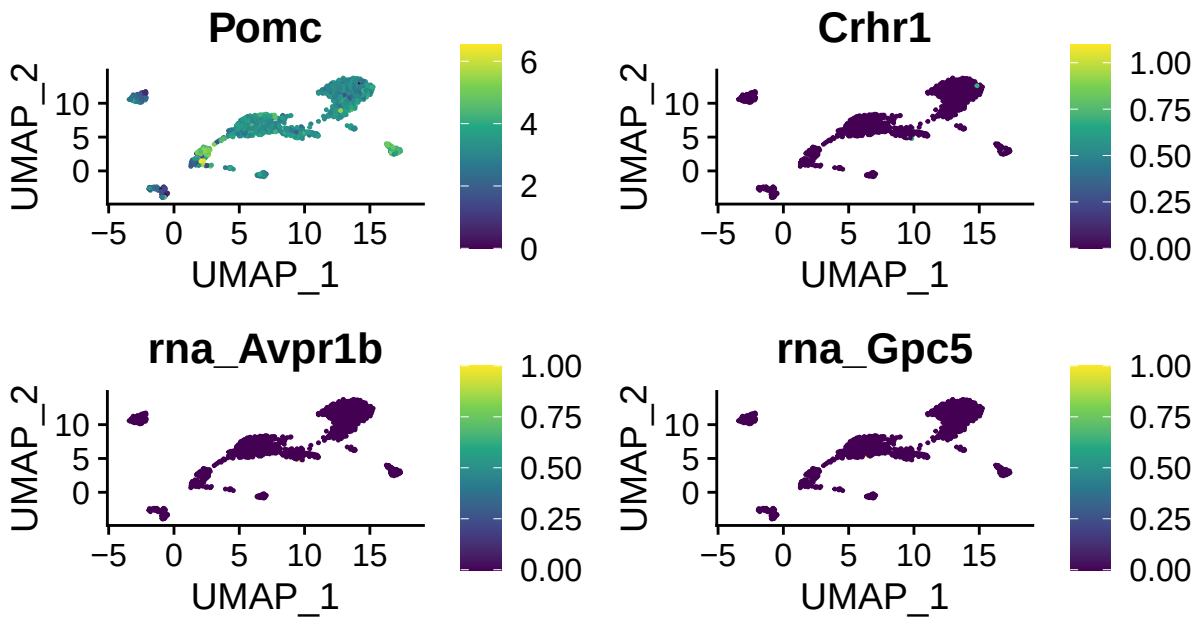
Then the melanotrophs may be excluded based on Pomc, Pcsk2 and Pax7 expression:



Then confirm the identification of the corticotrophs by Pomc, Crh1, Avpr1b and Gpc5.

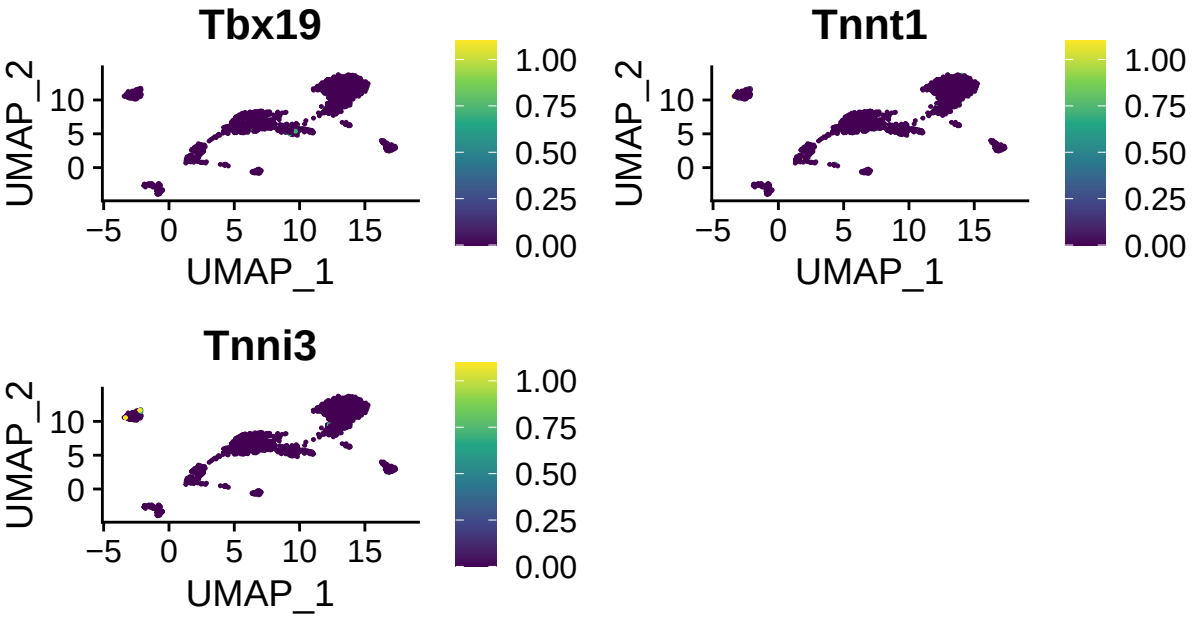
```
## Warning: Could not find Avpr1b in the default search locations, found in RNA
## assay instead
```

```
## Warning: Could not find Gpc5 in the default search locations, found in RNA assay
## instead
```



Still dififcult to isolte the corticotroph cluster, seek more canonical genes (from Cheung wt al paper). Note that cluster 3 appears to hold high levels of all genes, could this possibly be a cluster of double reads?

```
## Warning in FetchData(object = object, vars = c(dims, "ident", features), : The
## following requested variables were not found: Gm15543
```



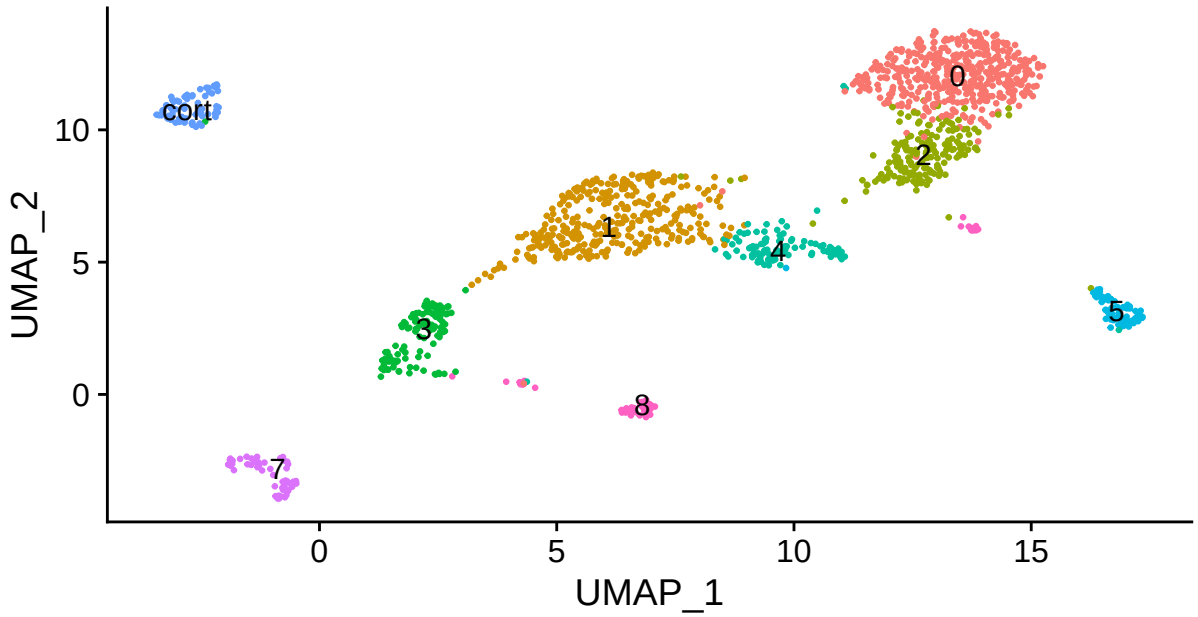
These plots would suggest cluster 6 is the most likely corticotroph cluster.

Step 7. Isolate desired cell cluster

To examine cell homogeneity within the corticotrophs, the cluster identified in the previous step was pulled out as a subset.

The desired cluster must be manually entered upon visual inspection of the gene expression feature plots in the previous step.

As a sanity check that the correct cluster was selected, a dimensional reduction plot;



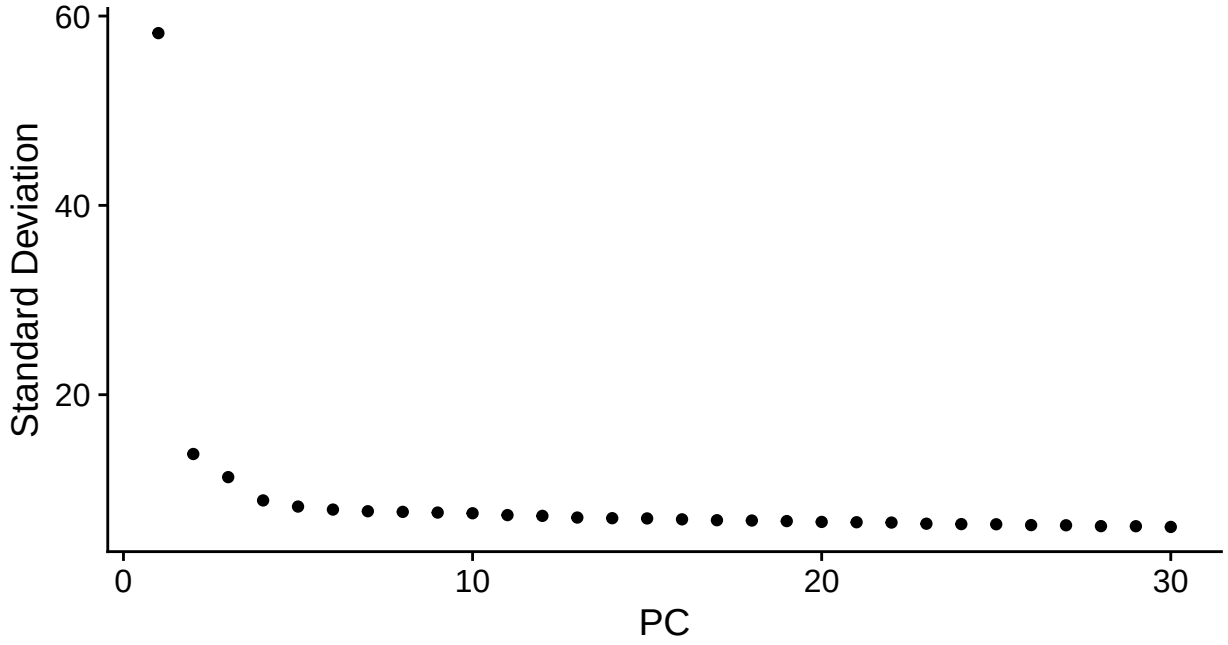
And the data for the cluster of interest can be extracted for further study.

Step 8. Re-cluster selected cell cluster

To investigate cell homogeneity within the cell cluster of interest, the same procedure may be repeated;

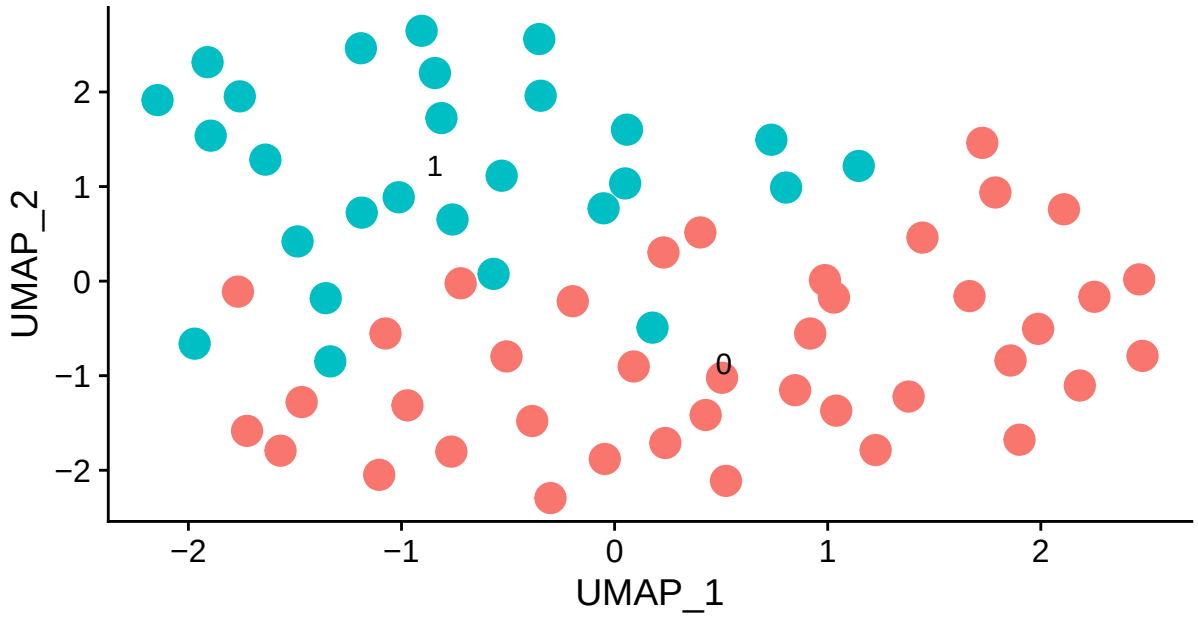
Data transformation, identification of principal components (PCA) Dimensional reduction (UMAP) *Identification of nearest neighbours (SNN)* Clustering using SNN.

Elbow plot to establish dimensionality

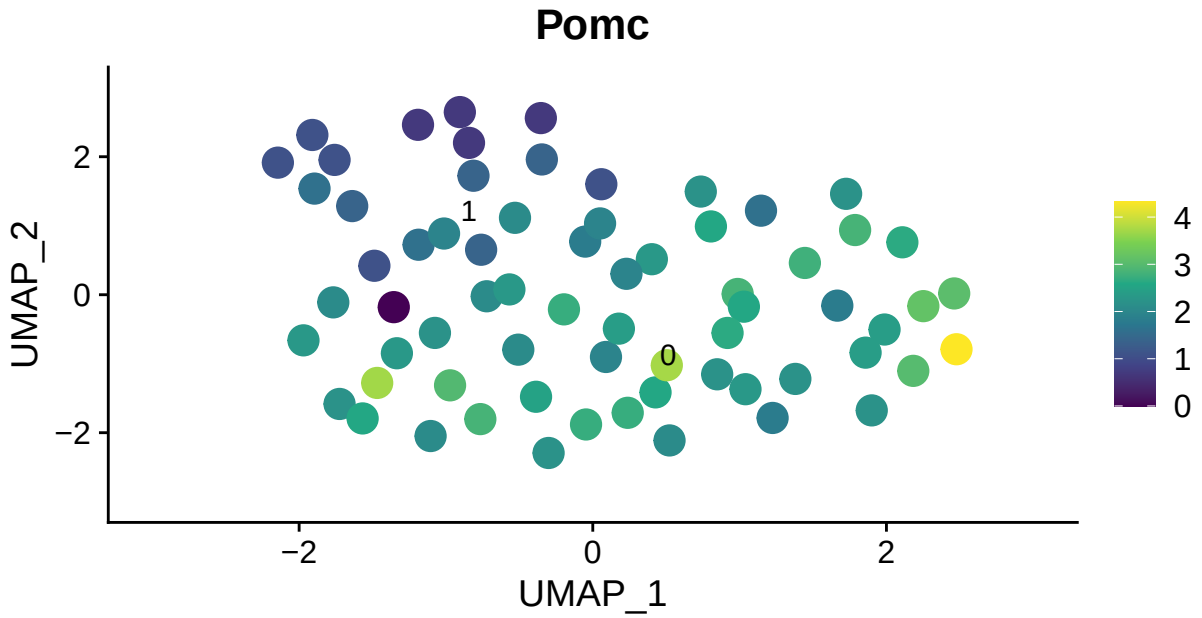


Identify most of information retained in 6 PCs

A dimensional reduction plot can then reveal any heterogeneity in the cluster;



A featureplot can be used to examine gene expression levels across these clusters within the corticotroph data;



Differential expression analysis can reveal differences between the clusters, for example the two top deferentially expressed genes for each cluster are;

```
## # A tibble: 4 x 7
## # Groups:   cluster [2]
##      p_val avg_logFC pct.1 pct.2  p_val_adj cluster gene
##      <dbl>     <dbl> <dbl> <dbl>      <dbl> <fct>   <chr>
## 1 6.23e-10      1.35  0.9  0.074 0.00000802 0       Meg3
## 2 5.67e- 4      1.54  0.95 0.963 1          0       Malat1
## 3 1.75e- 9      1.22  1    0.9   0.0000225  1       Rps18-ps3
## 4 3.15e- 8      1.25  1    0.725 0.000406   1       Mif
```

9. Summary

This analysis based on 14 of 30 possible principal components shows cluster 6 to be the corticotophs. When re-clustered, the corticotrophs demonstrate cell heterogeneity with two visible sub clusters.A differential gene expression analysis of these three sub-clusters shows the top differentiated genes in each cluster to be Meg3 & Malat1 (not statistically significant) in cluster zero and Rps18-ps3 & Mif in cluster one.