Pituitary corticotroph gene expression analysis

Mayran et al scRNA seq dataset

Craig McDougall

October 2020

Introduction

This document details the analysis of the single RNA sequence data by Mayran 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

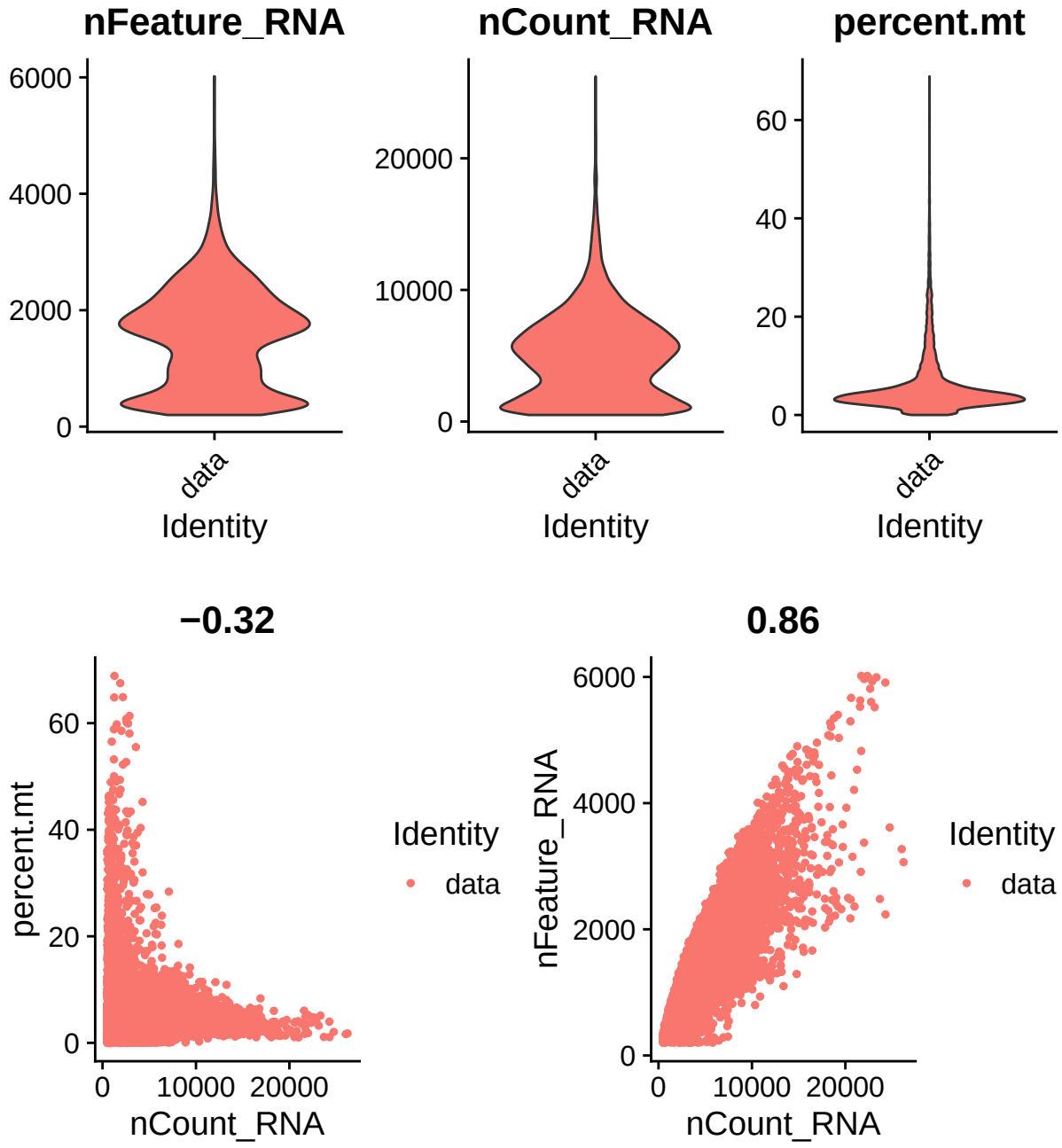
the working directories are then set.

The raw count and gene data is read in from the 10X sequence data files described in Chung et.al. 2018 and downloaded from the NCBI gene expression omnibus.

For the Mayran analysis, the Mayran folder is indicted in the path.

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). In the case of the Mayran data,there are clearly two separate features in the nfeature data, so all cells below 1200 counts and above 4000 counts are dropped and cells with >5% mitochondrial DNA are also dropped.

Step 3. Transform the data

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

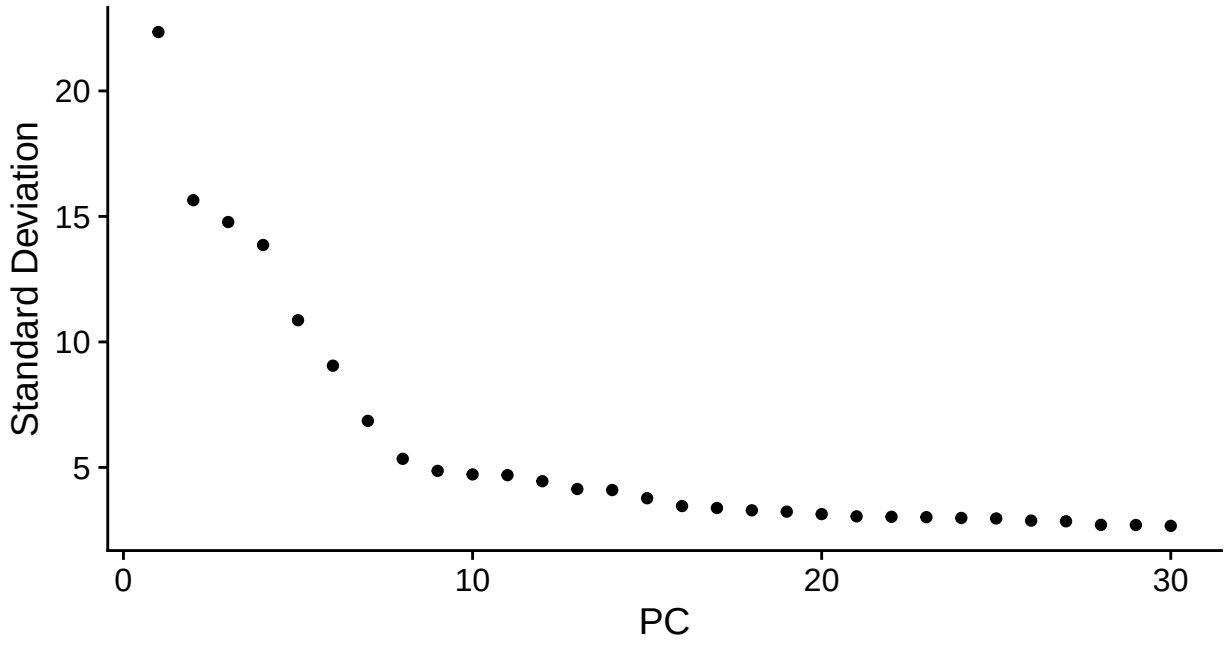
- Normalised to ensure all genes carry equal weight going forward and prevent heavily expressing genes from being over represented.
- Regressed to remove confounding factors such as areas mapped to the mitochondrial genome.
- Assessed for variable genes.

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 17 Principal components, so the last 13 are dropped during the next stage.

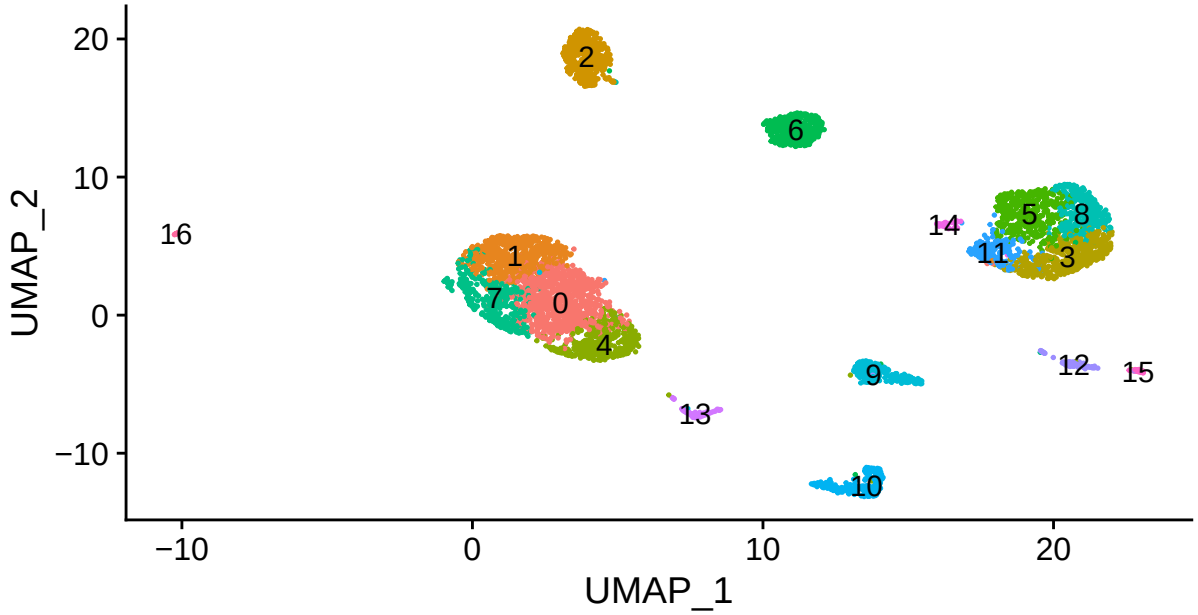
Uniform Manifold Approximation and Projection (UMAP) is an algorithm for dimensional reduction and is applied on the first 17 principal components established in the first step.

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

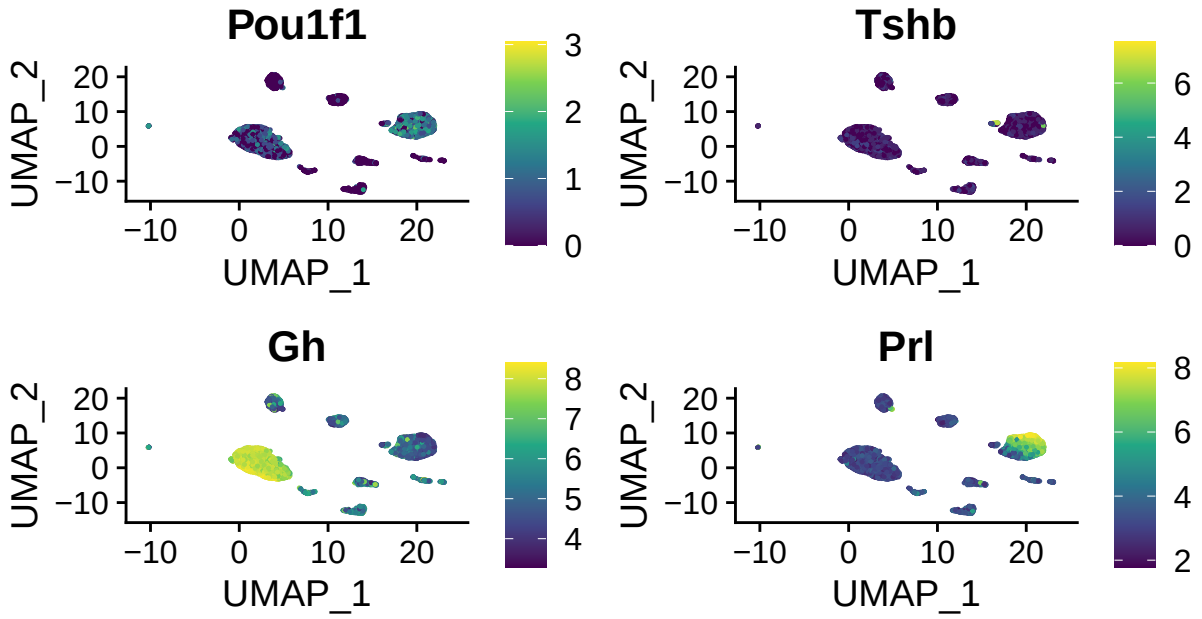
The main question now, is how well do the clusters approximate cell type? It is difficult to make absolute claims when attributing cell types to empirical cluster data based purely on an *in silico* analysis. However we may approximate the cell type of each cluster based on known canonical expression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

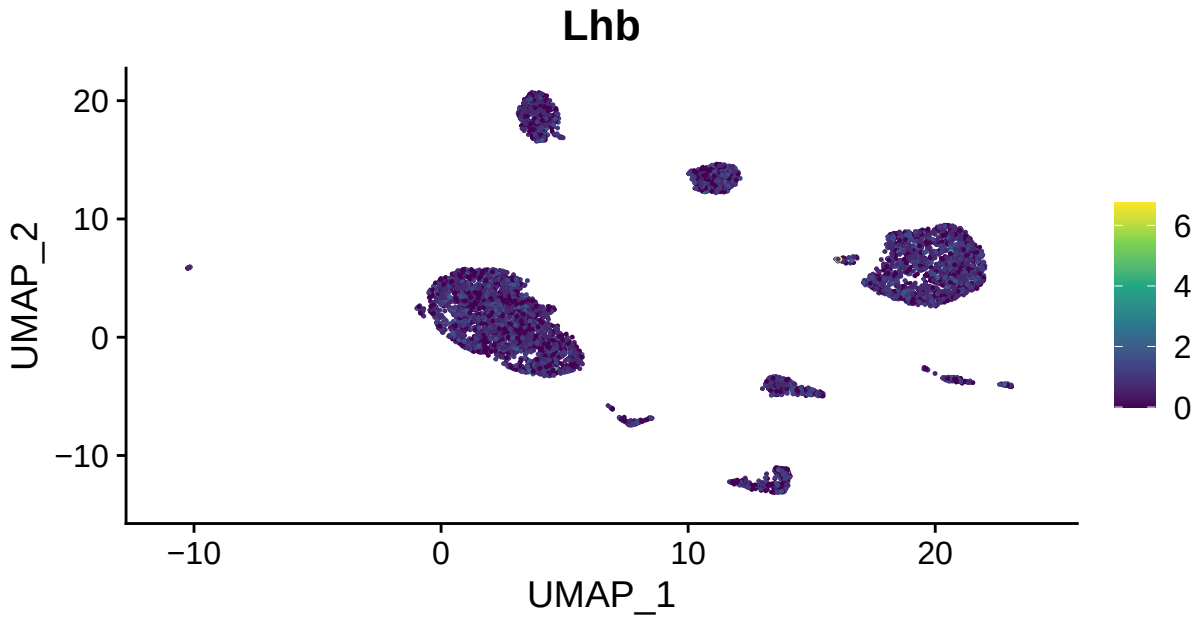
- Thyrotrophs (Pou1f1 + Tshb),
- Somatotrophs(Pou1f1 + Gh)
- Lactotrophs (Pou1f1 + Prl)
- Gonadotrophs (Lhb)
- Melanotrophs (Pomc + Pcsk2 + Pax7)
- Corticotrophs (Pomc + Crhr1 + Avpr1b + Gpc5 - Pcsk2 - Pax7)

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].

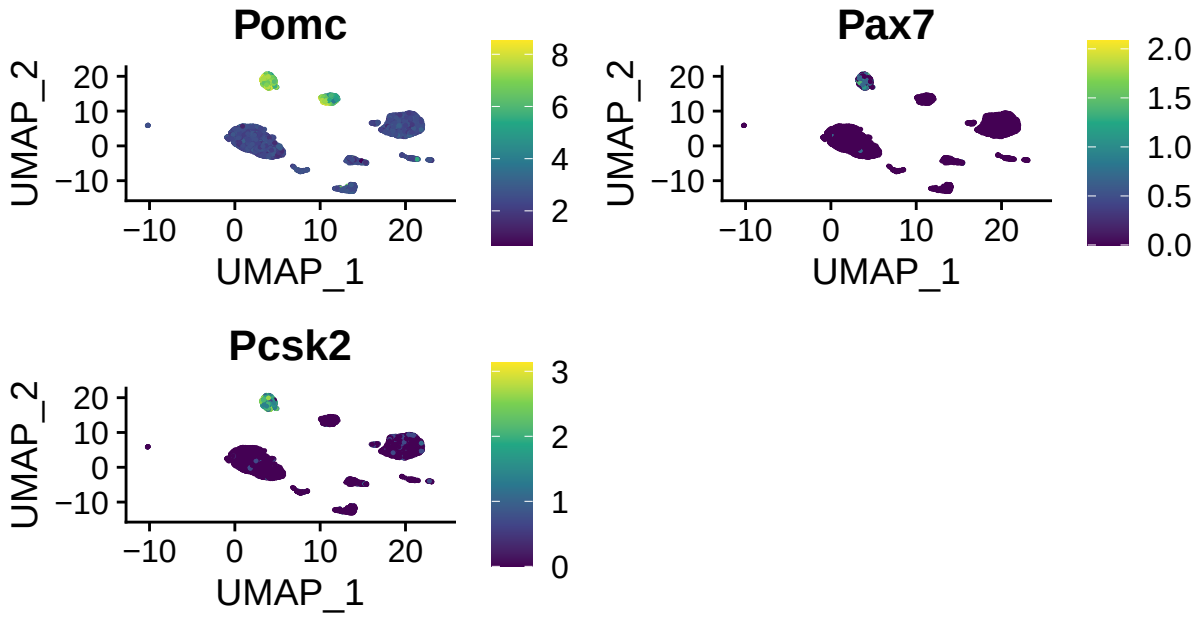
First the thyrotophs (Pou1f1 + Tshb), somatotrophs (Pou1f1 + Gh) and lactotrops (Pou1f1 + Prl) may be excluded based on gene expression:



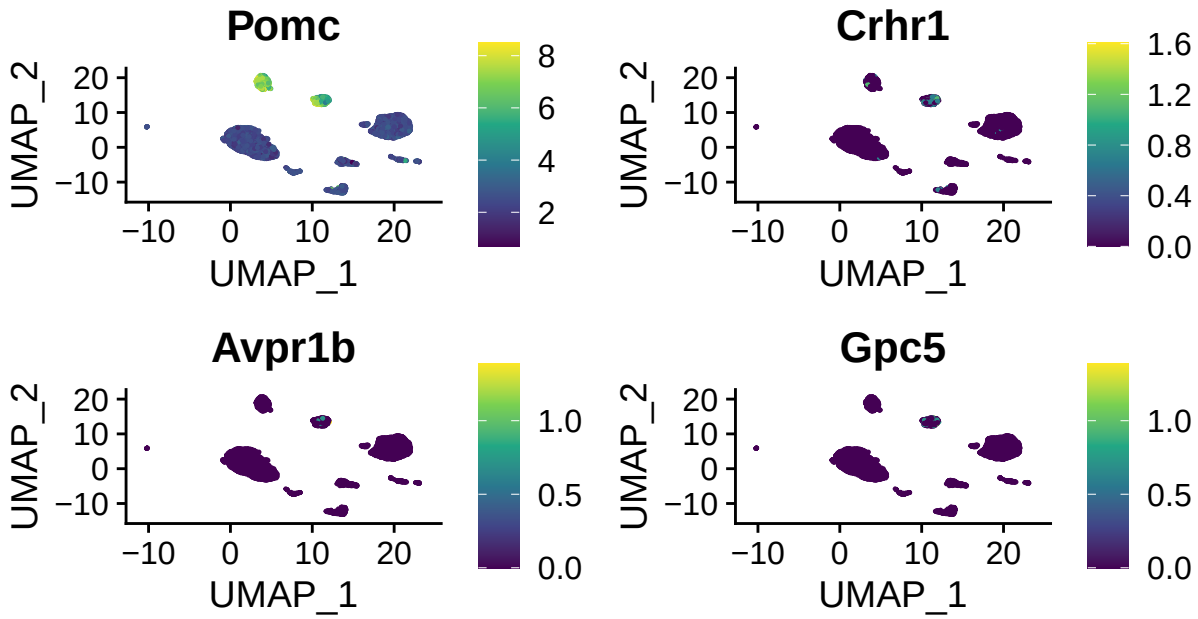
Then the Gonadotoprhs may be excluded by Lhb expression;



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



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



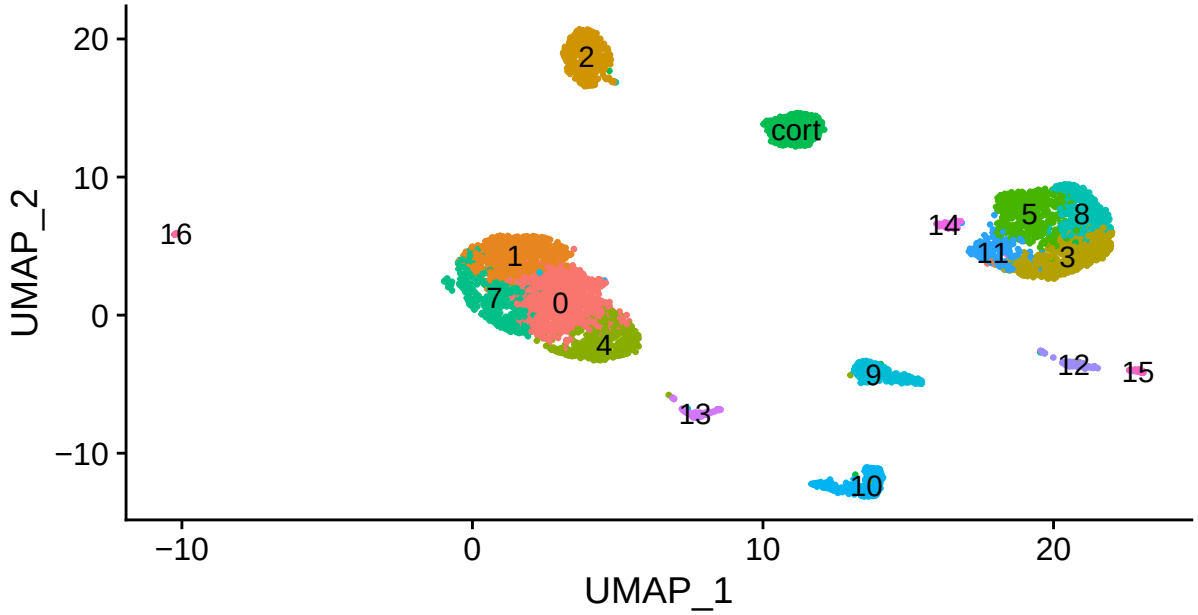
These plots would suggest cluster 6 is the most likely corticotroph cluster.

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;



Find the biomarkers for the corticotroph cluster compared to all other cells, what gene expression makes it different from the others? For a gene to be included in the differential expression for the cluster vs all other clusters, it must be expressed in at least 25% of the cells in the cluster `min.pct=0.25`. `Pct.1` is the percentage of cells in the cluster where the gene is detected, `pct.2` is the percentage of cells on average in all the other clusters where the gene is detected.

```
##                 p_val avg_logFC pct.1 pct.2     p_val_adj
## AW551984  0.000000e+00 1.1950312 0.872 0.102  0.000000e+00
## Tnnt1     0.000000e+00 1.0192818 0.494 0.018  0.000000e+00
## Atp1a2    0.000000e+00 0.9320393 0.689 0.035  0.000000e+00
## Tbx19     0.000000e+00 0.8148467 0.691 0.068  0.000000e+00
## Rprml     0.000000e+00 0.5541229 0.509 0.027  0.000000e+00
## Crhr1     0.000000e+00 0.4883848 0.472 0.007  0.000000e+00
## Obsl1    1.672596e-281 0.9196774 0.620 0.068 2.792232e-277
## Tekt1    7.135331e-233 0.6343119 0.568 0.068 1.191172e-228
## Sparcl1  3.631972e-229 0.5984997 0.642 0.087 6.063214e-225
## Mcub     8.870009e-222 1.0614584 0.891 0.261 1.480759e-217
```

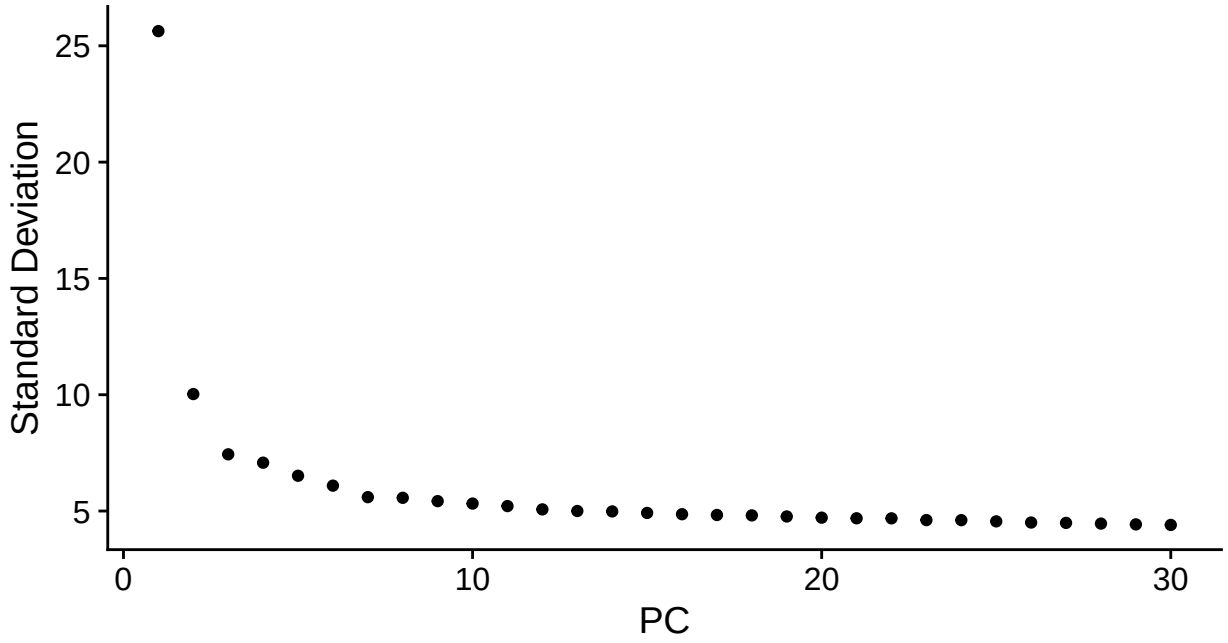
The data for the corticotroph cluster is extracted for further study using `subset`.

Step 8. Re-cluster selected cell cluster

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

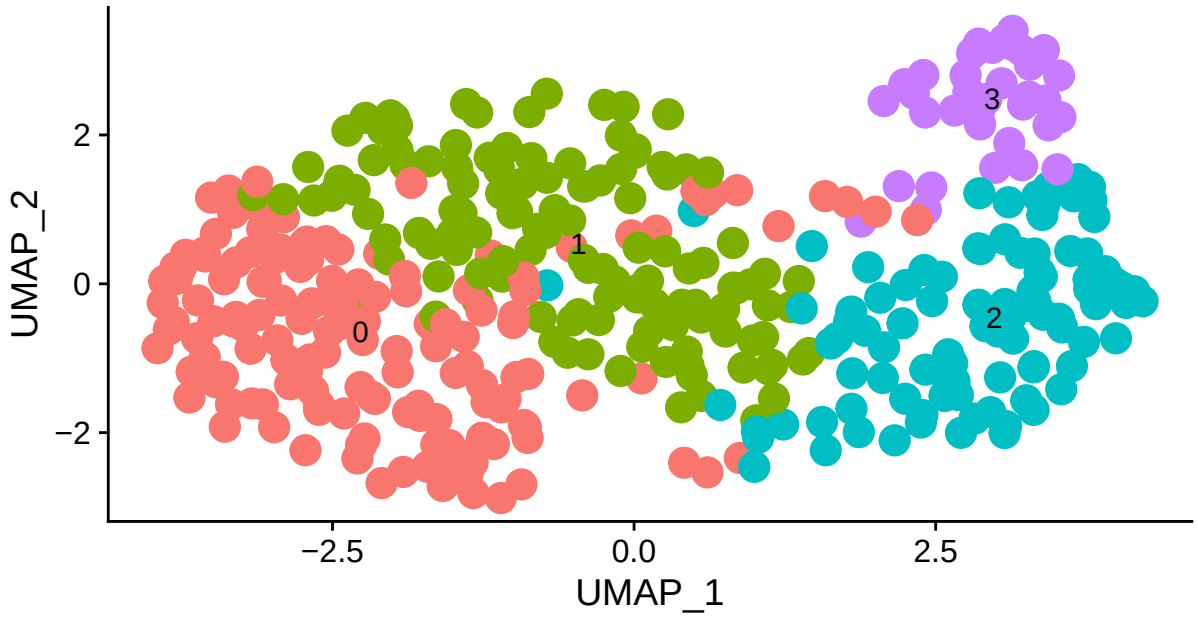
Data transformation, identification of principal components (PCA) Dimensional reduction (UMAP) *Identification of nearest neighbours (SNN)* Clustering using SNN.

Elbow plot to establish dimensionality;

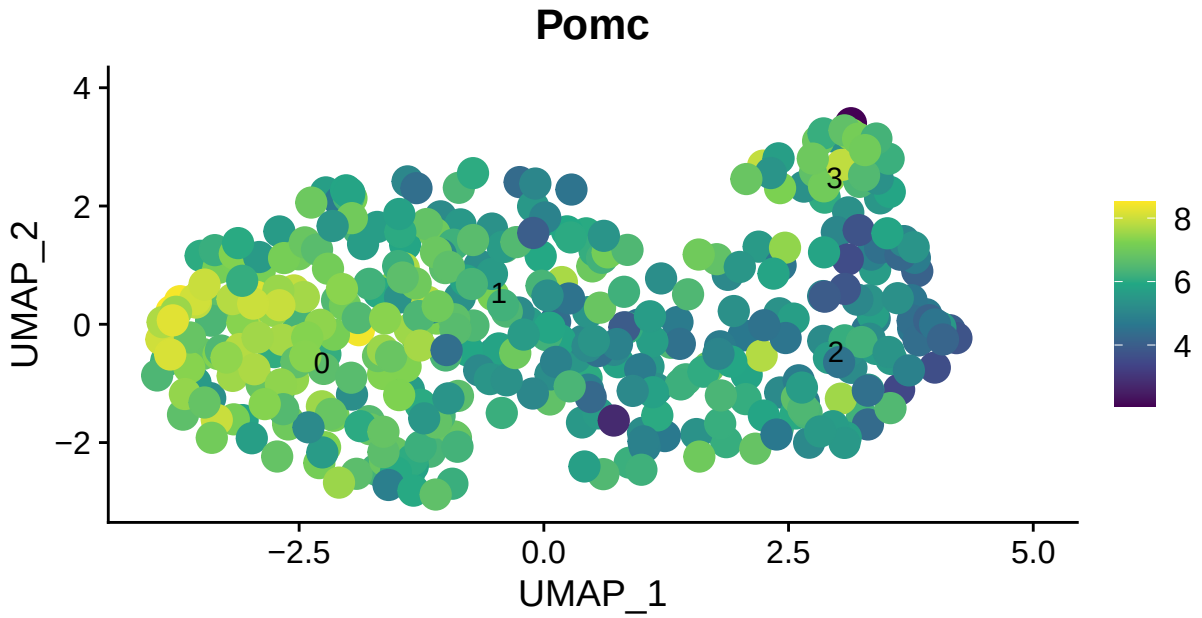


Identify most of information retained in 11 PCs

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



A featureplot can be used to examine gene expression levels across these sub-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: 8 x 7
## # Groups:   cluster [4]
##      p_val avg_logFC pct.1 pct.2 p_val_adj cluster gene
##      <dbl>     <dbl> <dbl> <dbl>     <dbl> <fct>   <chr>
## 1 1.23e-25     0.985 1     1      2.05e-21 0       Pomc
## 2 2.34e-17     0.996 0.503 0.135  3.90e-13 0       Nmb
## 3 3.93e-22     0.761 1     0.96   6.56e-18 1       Mt2
## 4 3.61e-21     0.955 0.915 0.659  6.03e-17 1       Srxn1
## 5 2.37e- 8     0.575 0.938 0.761  3.95e- 4 2       Jun
## 6 9.85e- 6     0.553 0.646 0.447  1.64e- 1 2       Tnnt1
## 7 3.71e- 6     0.918 0.829 0.716  6.19e- 2 3       Chga
## 8 1.38e- 3     1.00  0.6   0.357  1.00e+ 0 3       Gadd45g
```

9. Summary

This analysis based on 17 of 30 possible principal components suggests cluster 6 is the corticotophs.Differential expression analysis showed the top differentially expressed gene found in 87% of corticodroh cells (and 10% of all other cells) was AW551984 (similar to the Cheung et al. data analysis).

When re-clustered, the corticotrophs demonstrate cell heterogeneity with five visible sub clusters.A differential analysis of these three clusters shows the top differentiated genes in each cluster to be Pomc & Nmb in cluster zero, Mt2 & Srxn1 in cluster one, Jun & Tnnt1 (not statistically significant) in cluster 2 and Chga (not statistically significant) & Gadd45g (not statistically significant) in cluster 3.