

Pituitary corticotroph gene expression analysis

Cheung et al scRNA seq dataset

Craig McDougall

October 2020

Introduction

This document details the analysis of the single RNA sequence data by Chung et.al. 2018. available from the NCBI gene expression omnibus.

Step 1. Call Libraries & Set Directories

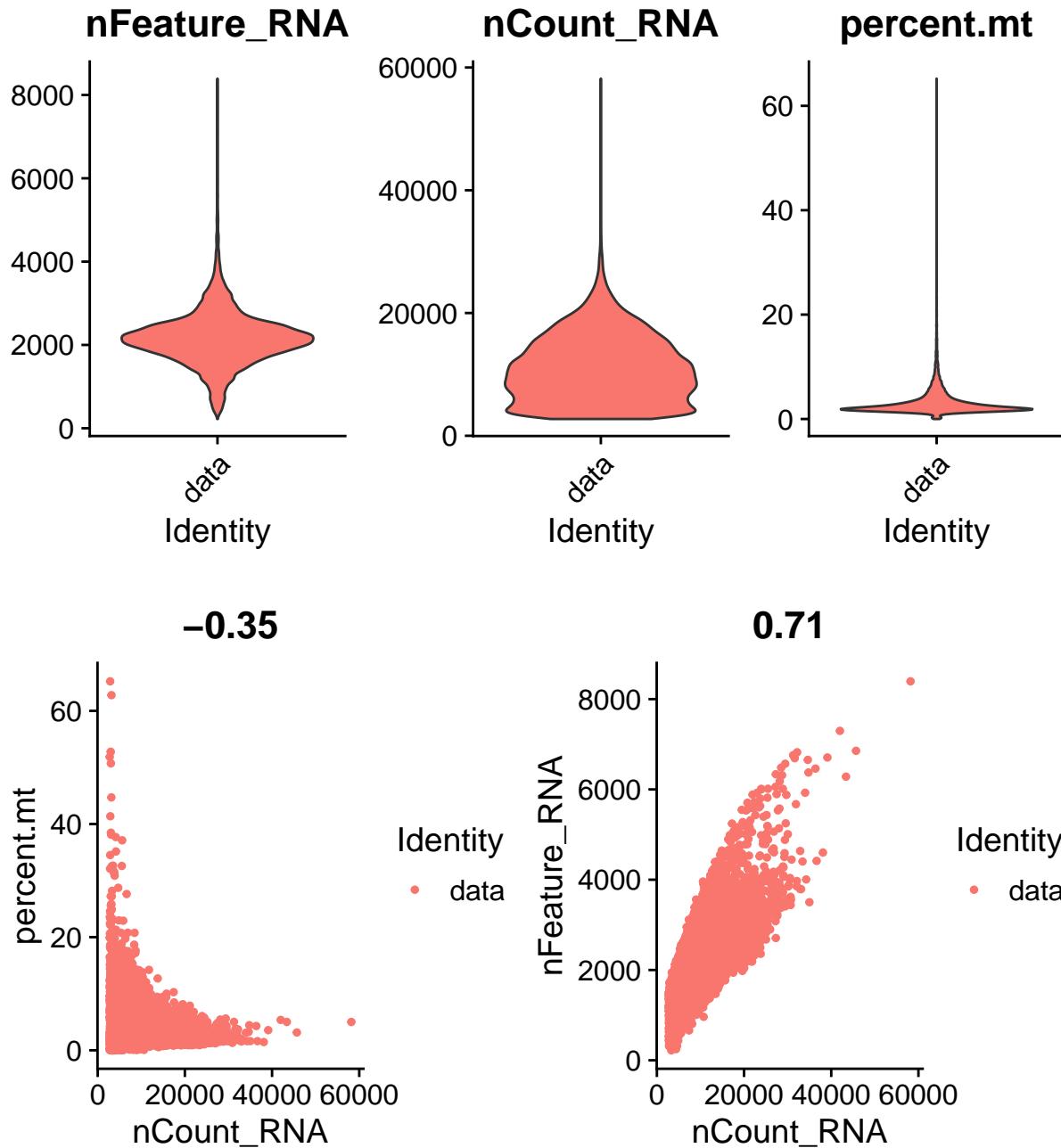
Before the main body of the pipeline is run, all of the required libraries are called and the working directories are then set.

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 Cheung analysis, the Cheung folder is indicated 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 Cheung data, all cells below 200 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;

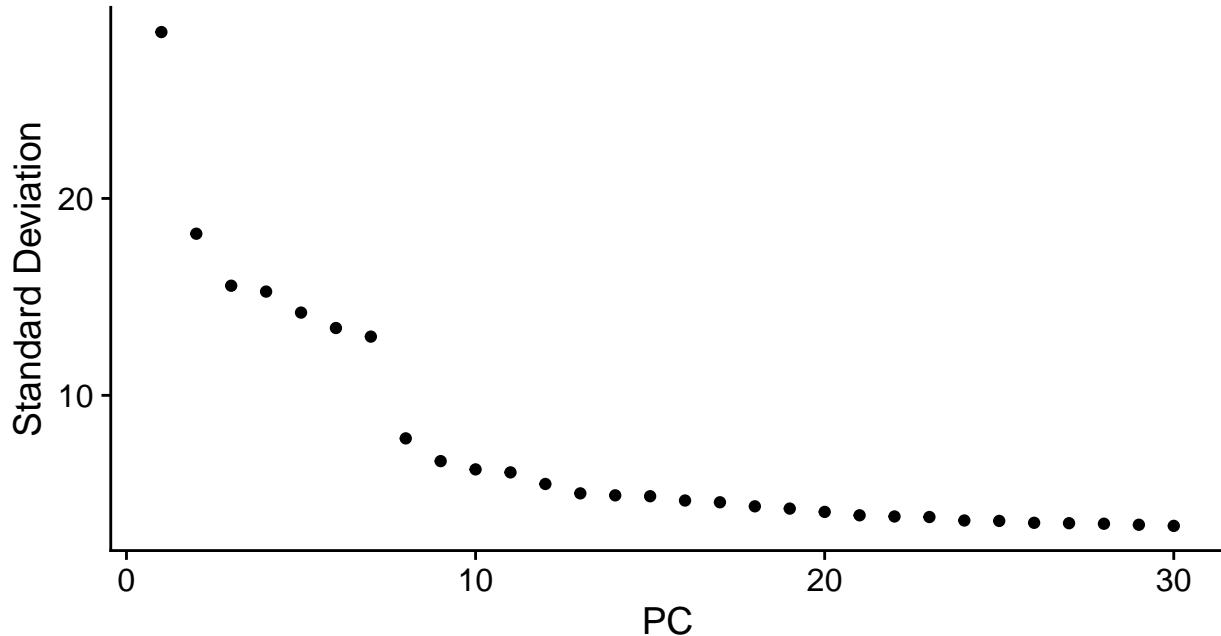
- Normalised to ensure all genes carry equal weight going forward and prevent heavily expressing genes from being over represented.
- Regressed to remove confounding factors such as areas mapped to the mitochondrial genome.
- Assessed for variable genes.

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 20 Principal components, so the last 10 are dropped during the next stage.

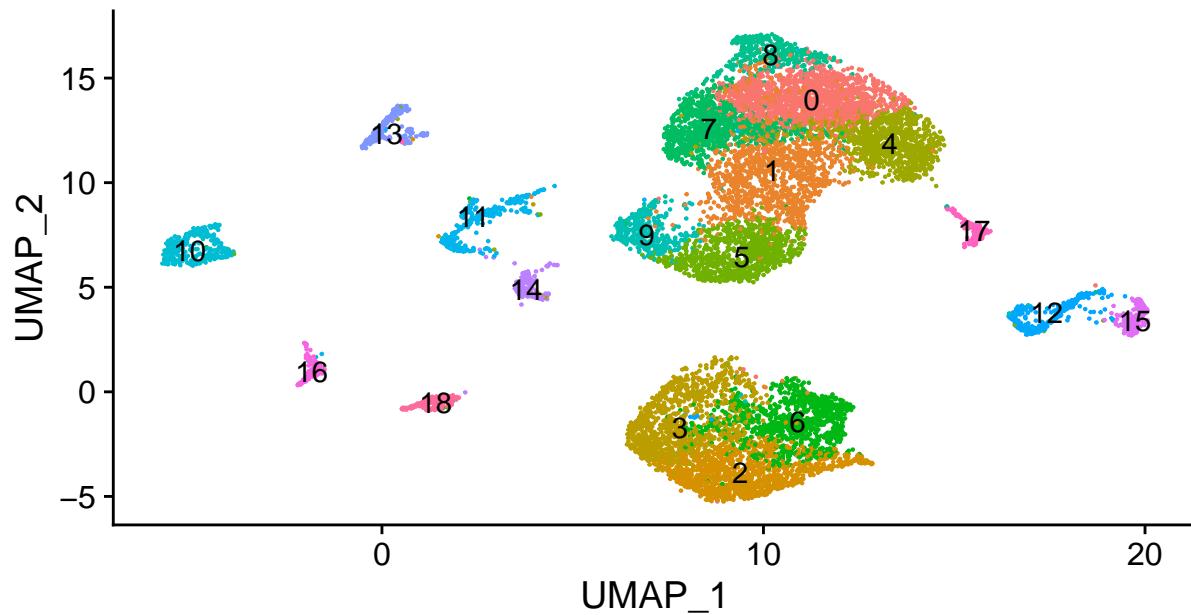
Uniform Manifold Approximation and Projection (UMAP) is an algorithm for dimensional reduction and is applied on the first 20 principal components established in the first step.

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 can 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

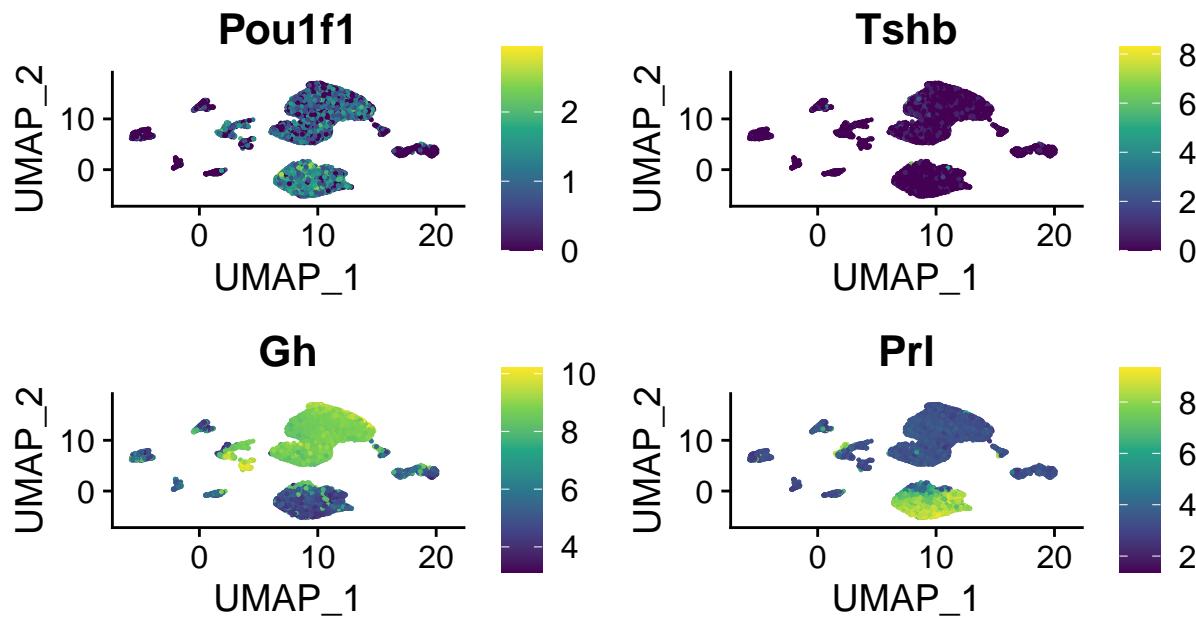
The main question now, is how well do the clusters approximate cell type? It is difficult to make absolute claims when attributing cell types to empirical cluster data based purely on an *in silico* analysis. However we may approximate the cell type of each cluster based on known canonical expression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

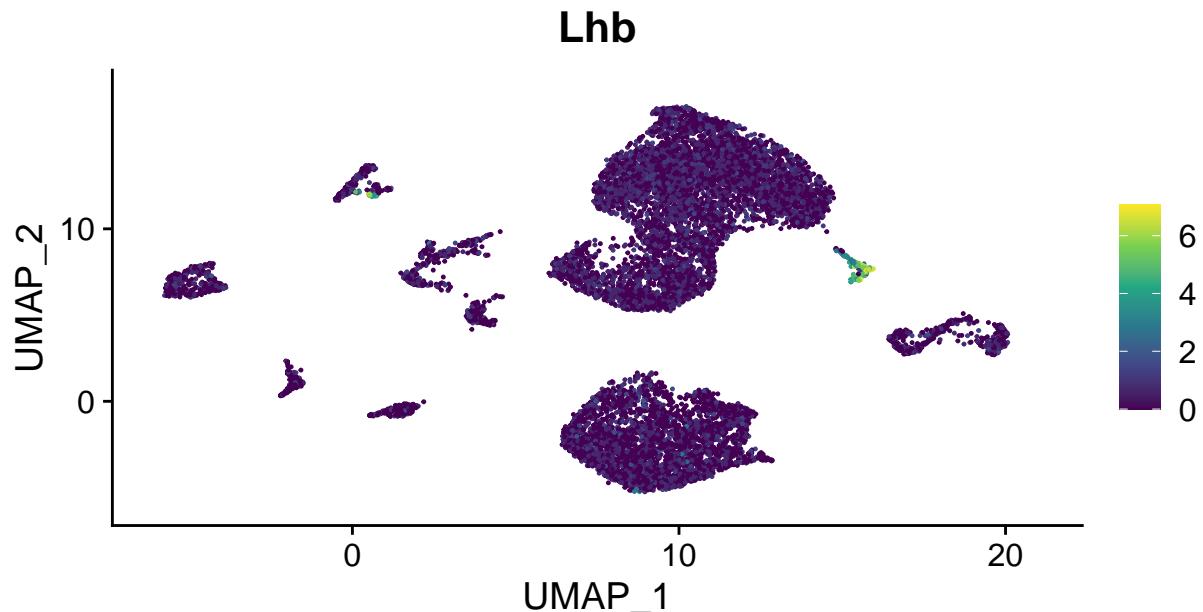
- Thyrotrophs (Pou1f1 + Tshb),
- Somatotrophs(Pou1f1 + Gh)
- Lactotrophs (Pou1f1 + Prl)
- Gonadotrophs (Lhb)
- Melanotrophs (Pomc + Pcsk2 + Pax7)
- Corticotrophs (Pomc + Crhr1 + Avpr1b + Gpc5 - Pcsk2 - Pax7)

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.

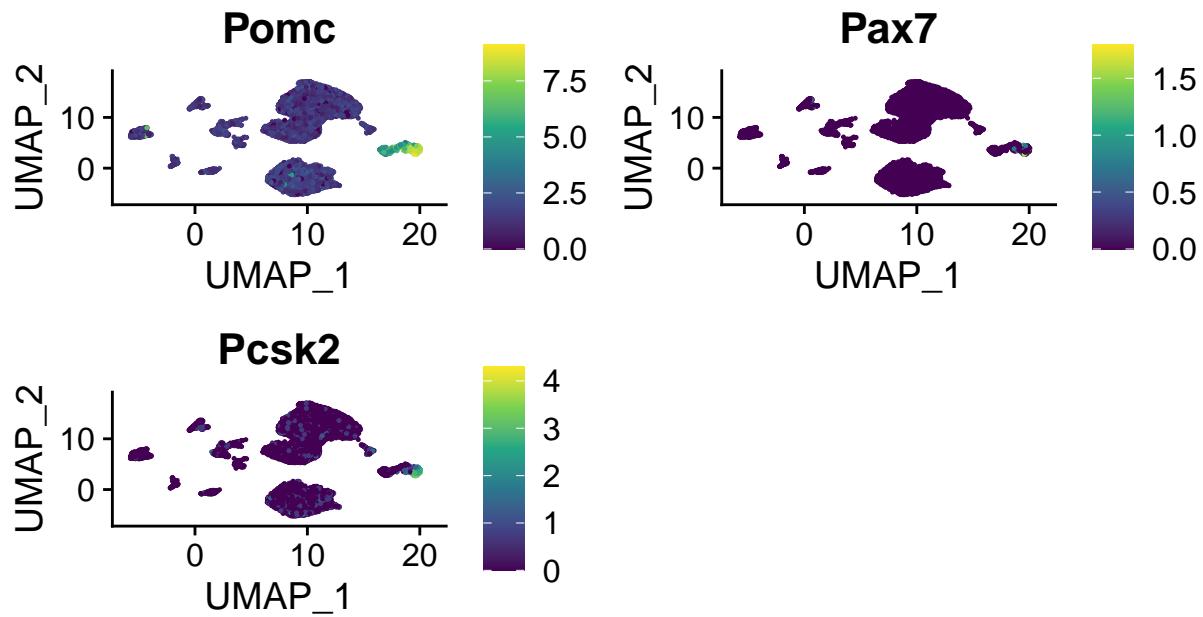
First the thyrotrophs (Pou1f1 + Tshb), somatotrophs (Pou1f1 + Gh) and lactotrophs (Pou1f1 + Prl) may be excluded based on gene expression:



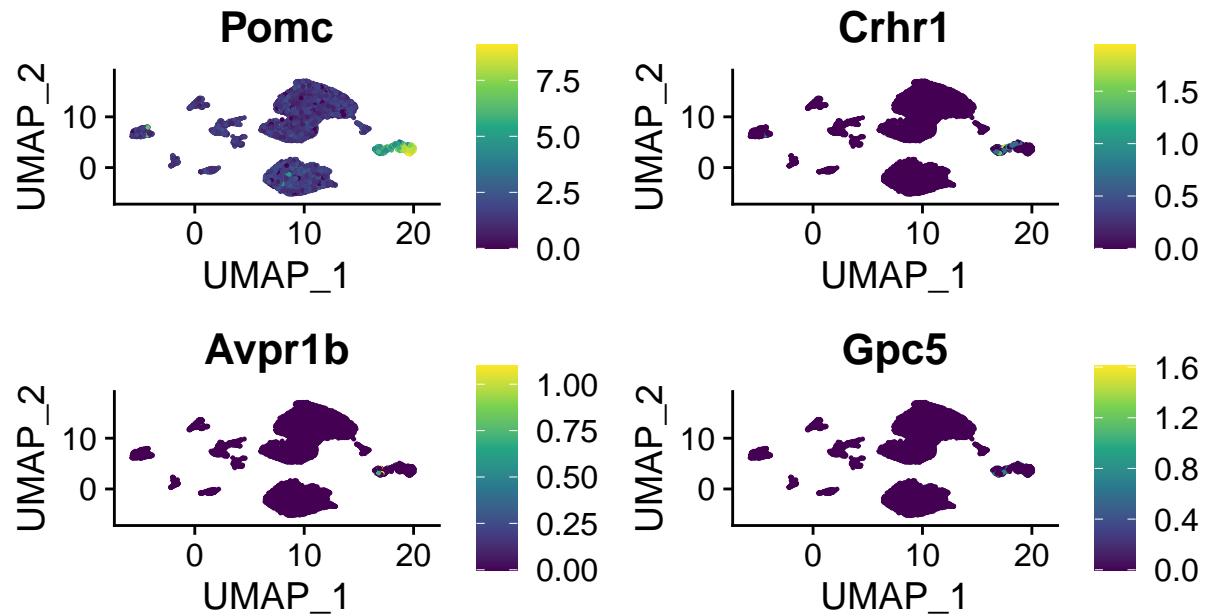
Then the Gonadotrophs may be excluded by Lhb expression;



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



Then confirm the identification of the corticotrophs by *Pomc*, *Crhr1*, *Avpr1b* and *Gpc5*.



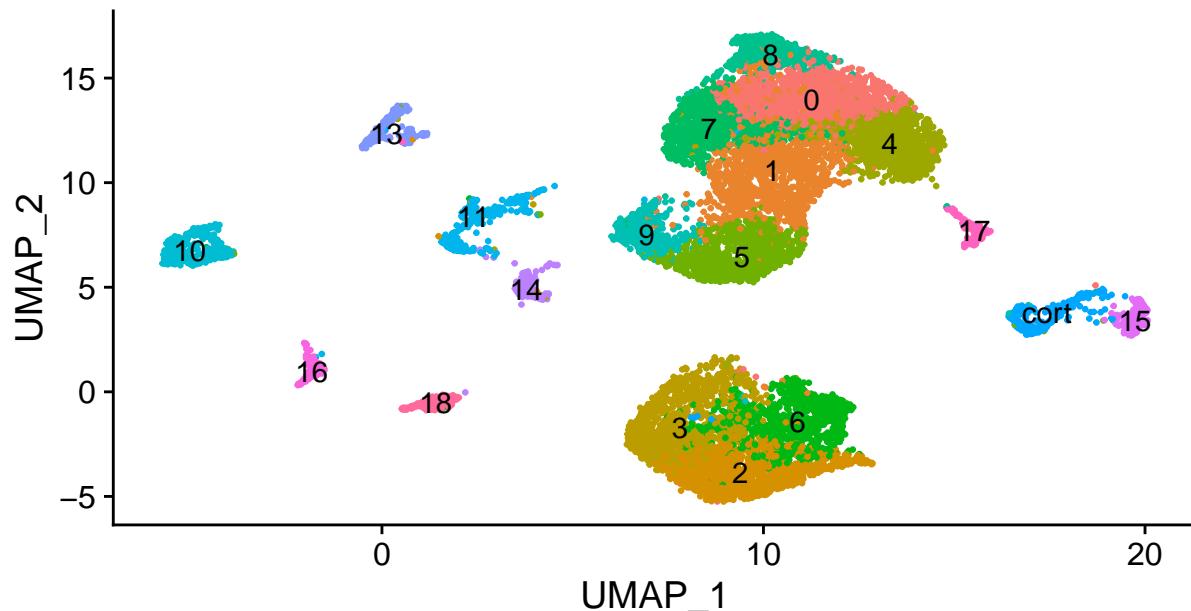
These plots would suggest cluster 12 is the most likely corticotroph cluster.

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 1.4201545 0.900 0.157 0.000000e+00
## Tnnt1     0.000000e+00 1.1761745 0.478 0.016 0.000000e+00
## Atp1a2    0.000000e+00 0.7917767 0.572 0.027 0.000000e+00
## Sparcl1   0.000000e+00 0.7565430 0.609 0.031 0.000000e+00
## Tbx19     0.000000e+00 0.7156152 0.544 0.012 0.000000e+00
## Crhr1     0.000000e+00 0.6010419 0.503 0.004 0.000000e+00
## Neurod1   0.000000e+00 0.2809145 0.253 0.007 0.000000e+00
## Baiap3    1.736077e-264 0.6049906 0.534 0.053 3.009836e-260
## Cngb1     1.433219e-211 0.3158180 0.297 0.018 2.484772e-207
## Acbd7     1.730225e-204 0.3871606 0.338 0.026 2.999692e-200
```

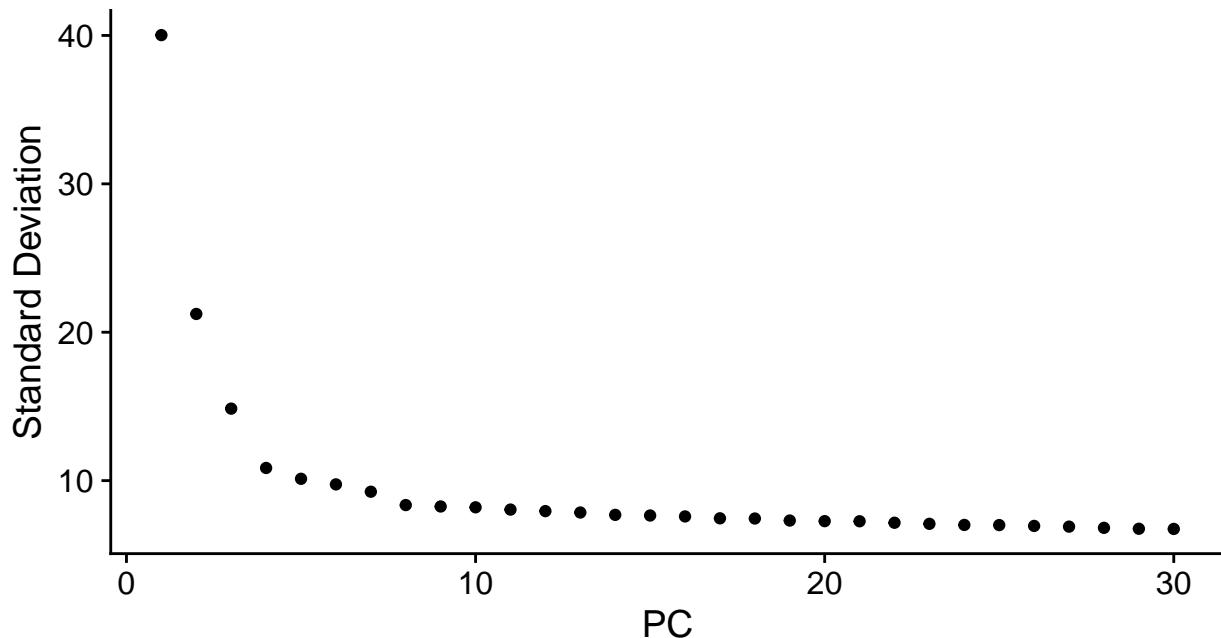
The data for the corticotroph cluster is extracted for further study using `subset`.

Step 8. Re-cluster corticotroph cluster and study cluster homogeneity

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

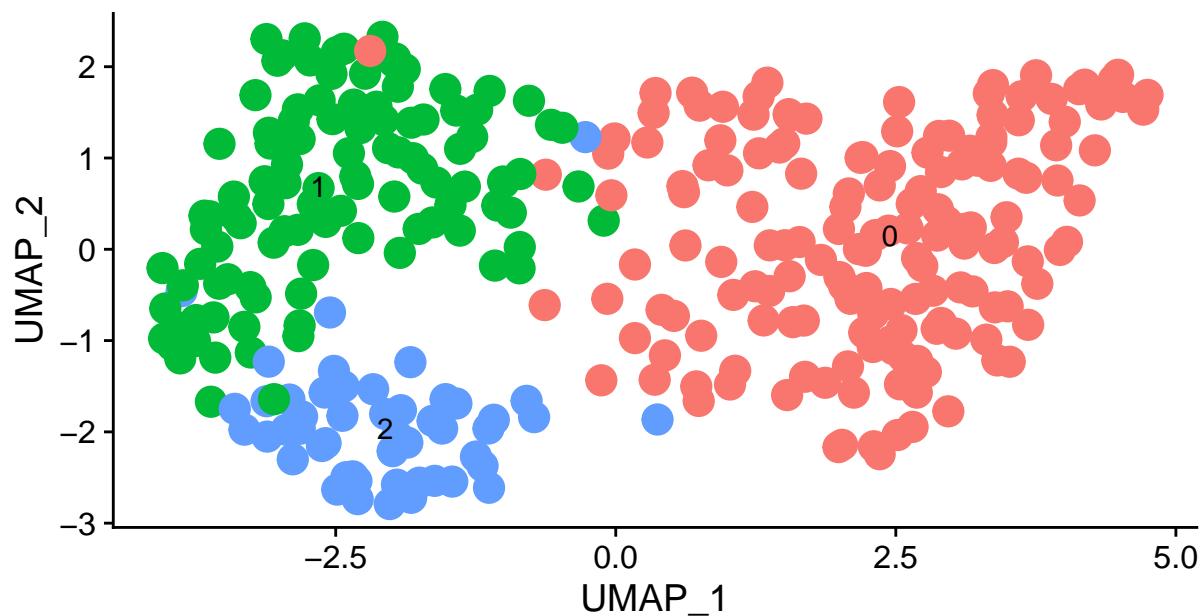
Data transformation, identification of principal components (PCA) Dimensional reduction (UMAP) Identification of nearest neighbours (SNN) Clustering using SNN.

Elbow plot to establish dimensionality;

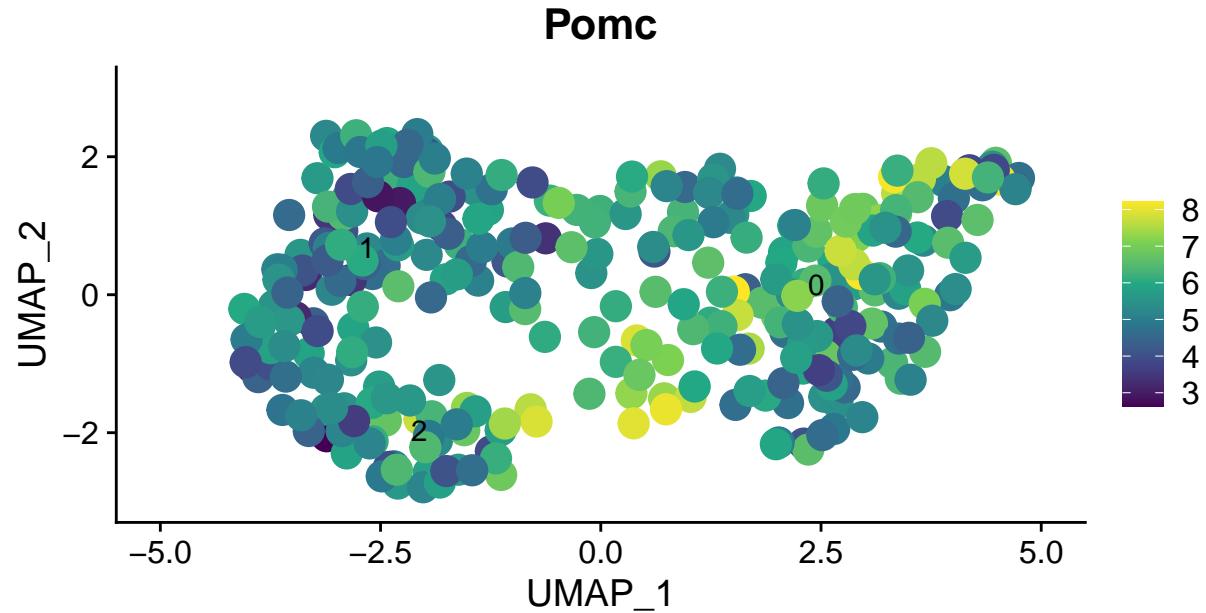


Identify most of information retained in 12 PCs

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 differentially expressed genes for each cluster are;

```
## # A tibble: 6 x 7
## # Groups:   cluster [3]
##       p_val avg_logFC pct.1 pct.2 p_val_adj cluster gene
##       <dbl>     <dbl> <dbl> <dbl>      <dbl> <fct> <chr>
## 1 1.67e-18     0.570 0.963 0.829  2.90e-14 0      Ndufa4
## 2 6.49e- 7     0.655 1        1      1.12e- 2 0      Pomc
## 3 2.18e-37     2.02  0.991 0.81   3.78e-33 1      Dnajb1
## 4 1.29e-26     1.75  0.927 0.701  2.24e-22 1      Hspb1
## 5 6.50e-16     1.35  1        0.985  1.13e-11 2      Mt1
## 6 1.91e-12     1.32  0.98   0.756  3.31e- 8 2      Mt2
```

9. Summary

This analysis based on 20 of 30 possible principal components identifies cluster 12 to be the corticotrophs based on expression of canonical markers. Differential expression analysis showed the top differentially expressed gene found in 90% of corticodroh cells (and 16% of all other cells) was AW551984.

When re-clustered, the corticotrophs demonstrate cell heterogeneity with three visible sub clusters. A differential gene expression analysis of these three sub-clusters shows the top differentiated genes in each cluster to be Ndufa4 & Pomp in cluster zero, Dnajb1 & Hspb1 in cluster one and Mt1 & Mt2 in cluster three.