Project log – week 8

*The housekeeping proteome: identifying proteins essential for human cells*

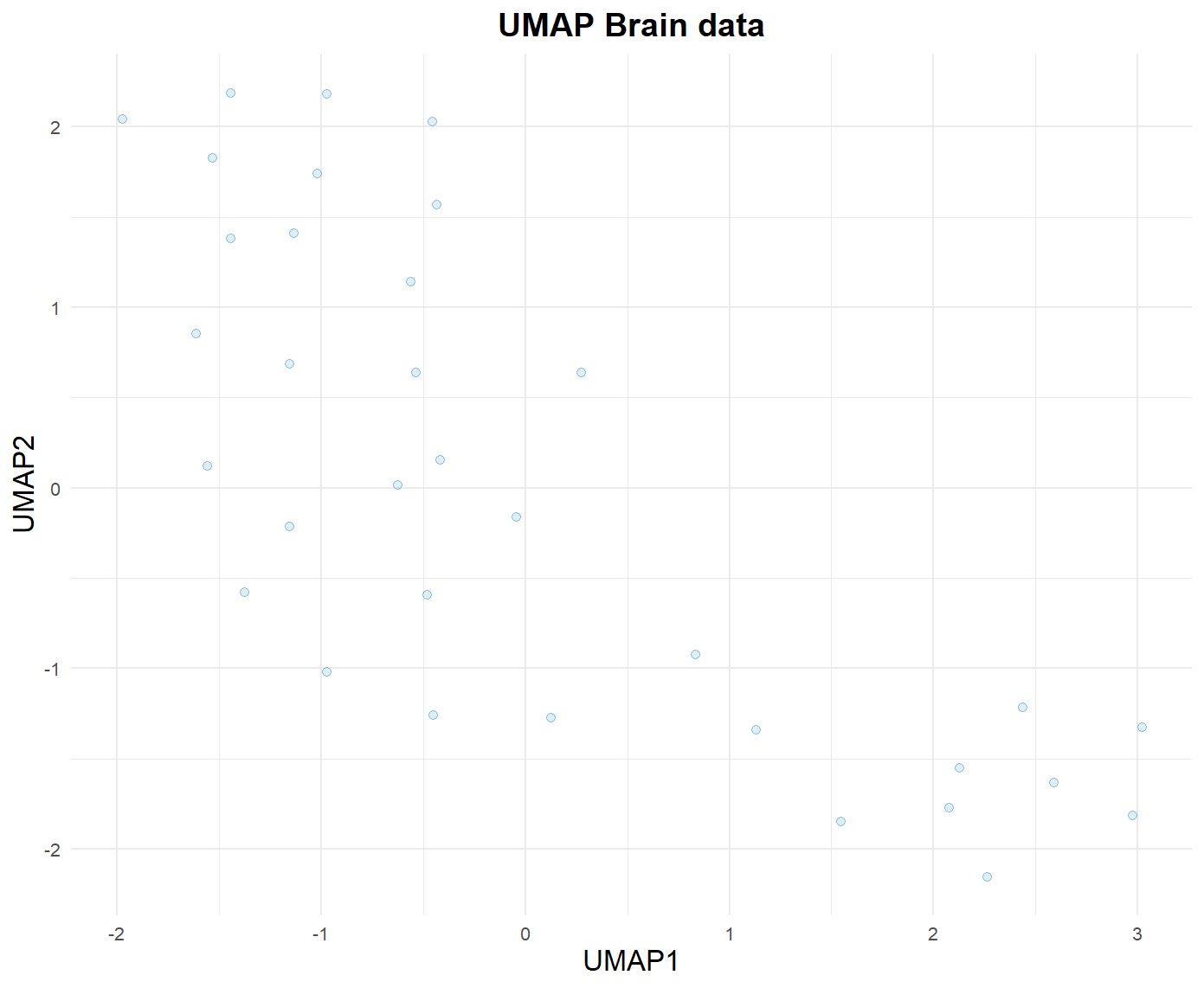
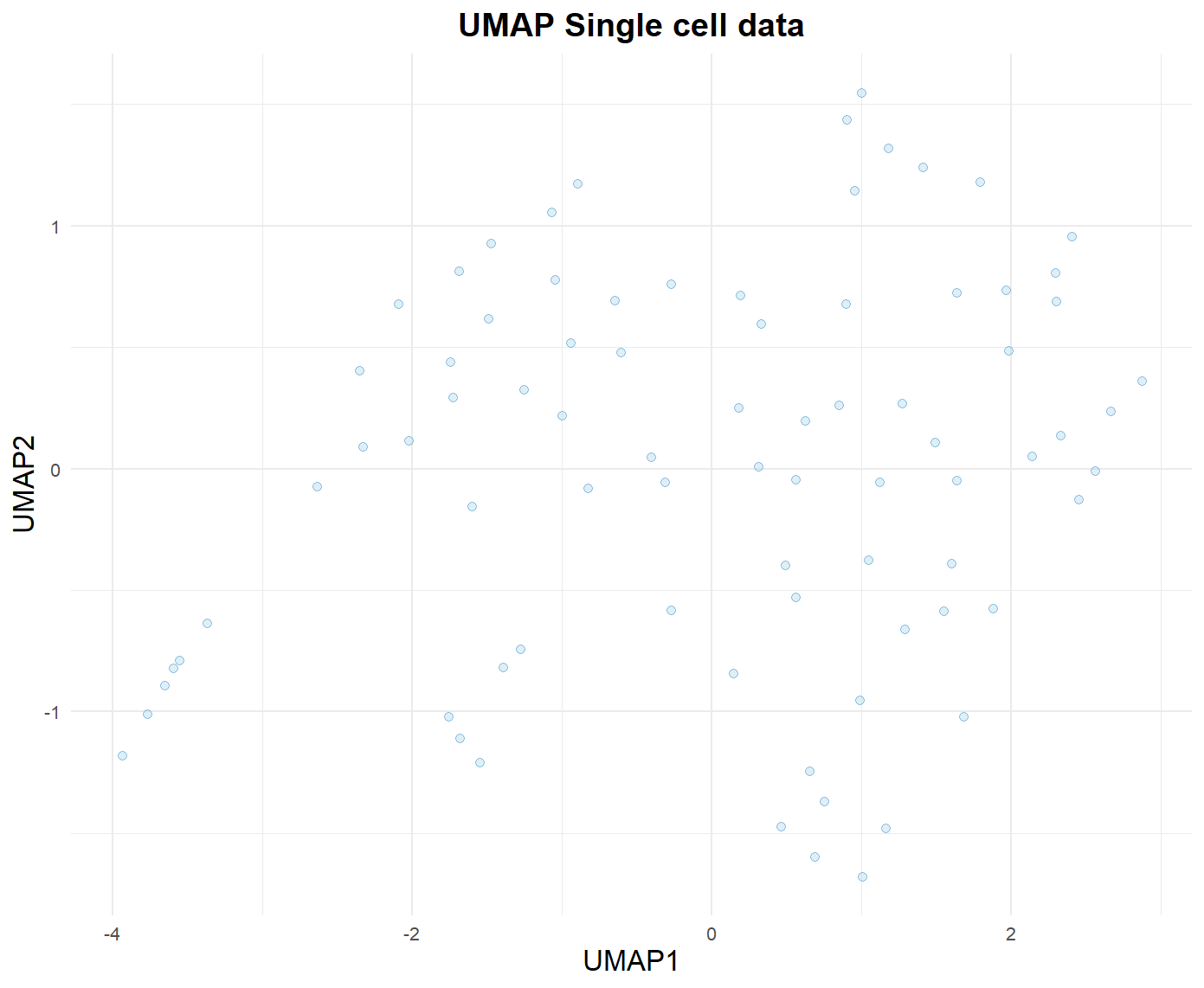
Overview

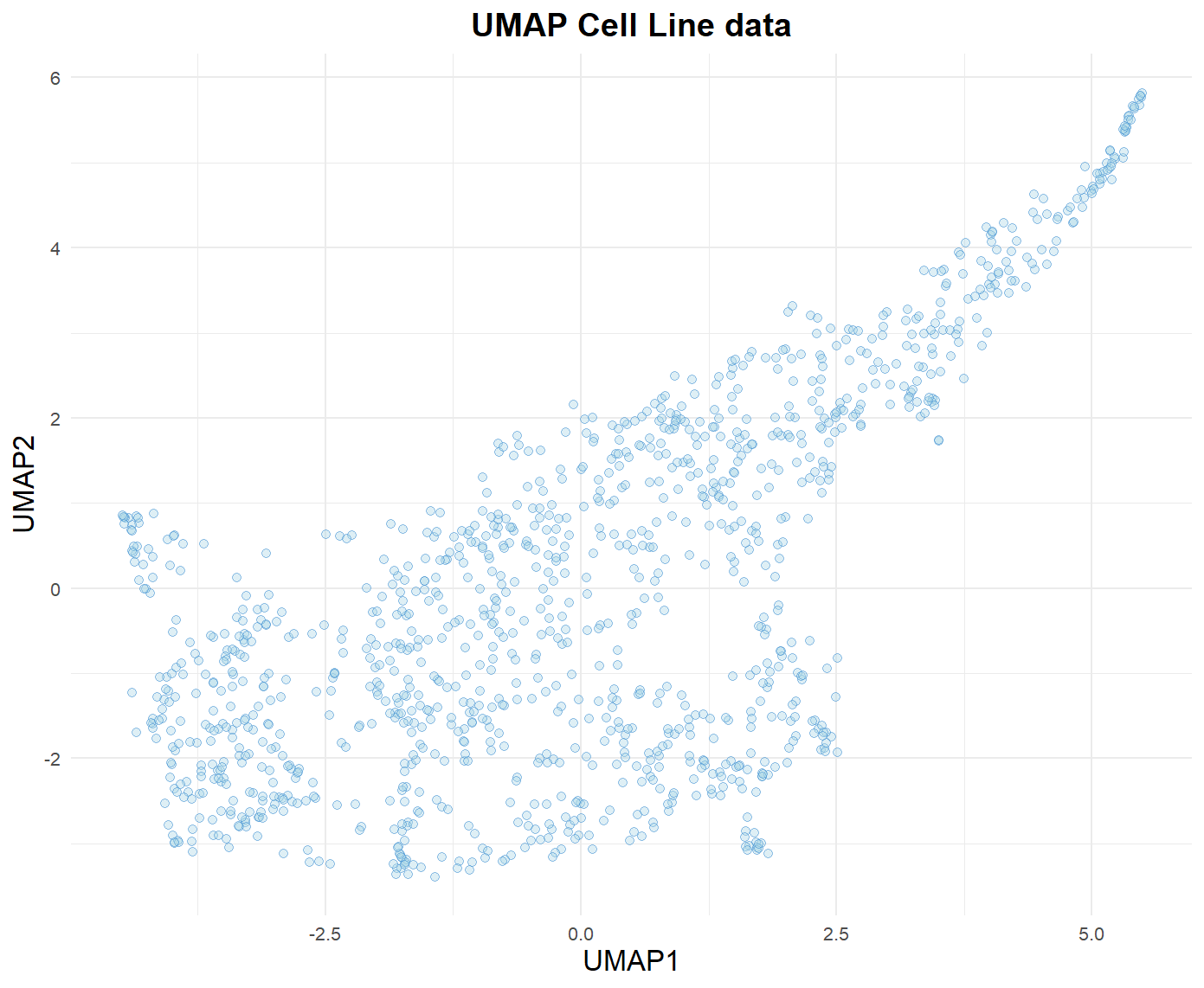
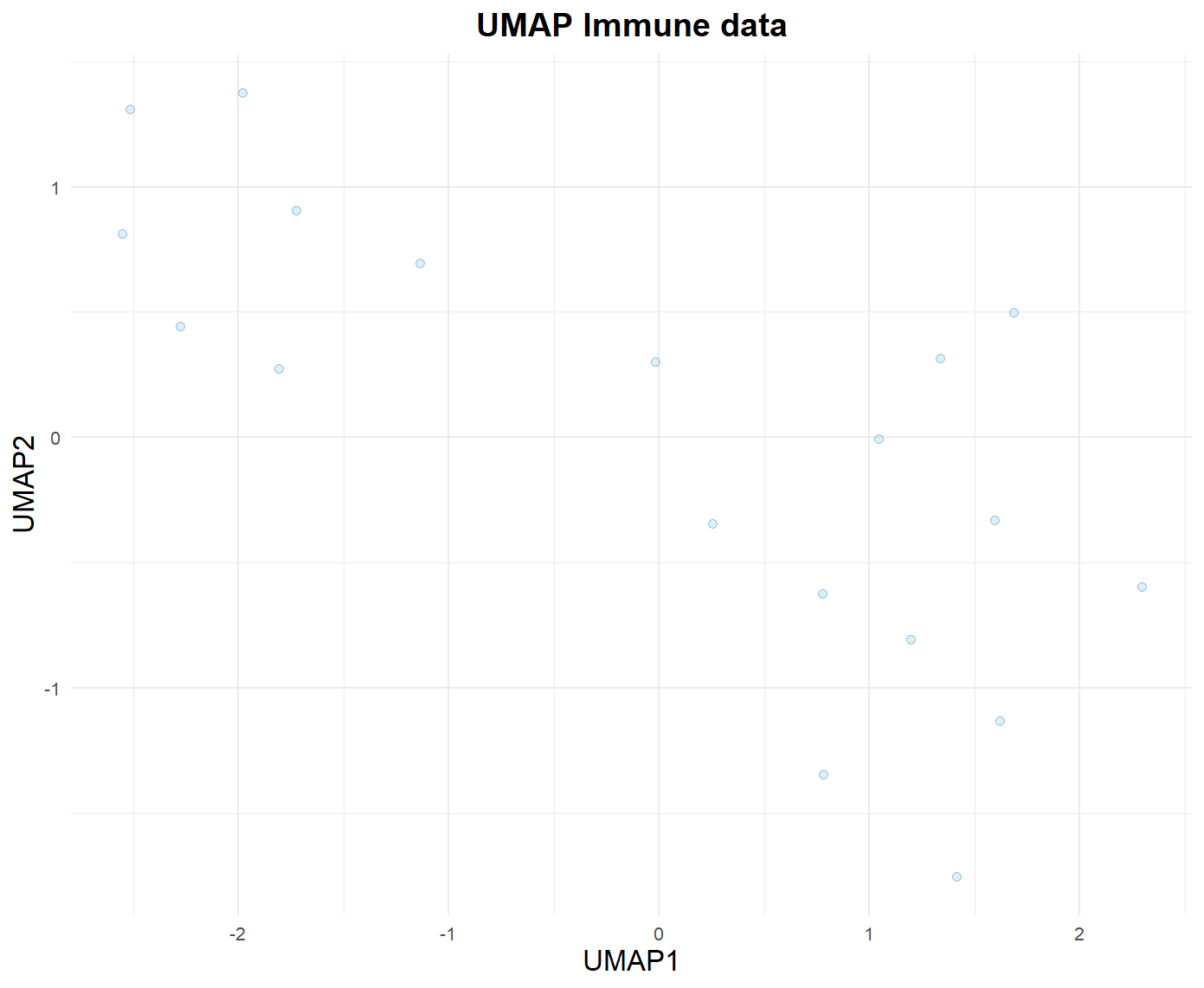
I have updated the UMAP of the samples’ profile distribution. Different thresholds for the “always expressed” gene have been applied. Only the single-cell data has distinct clusters of “always expressed” genes.

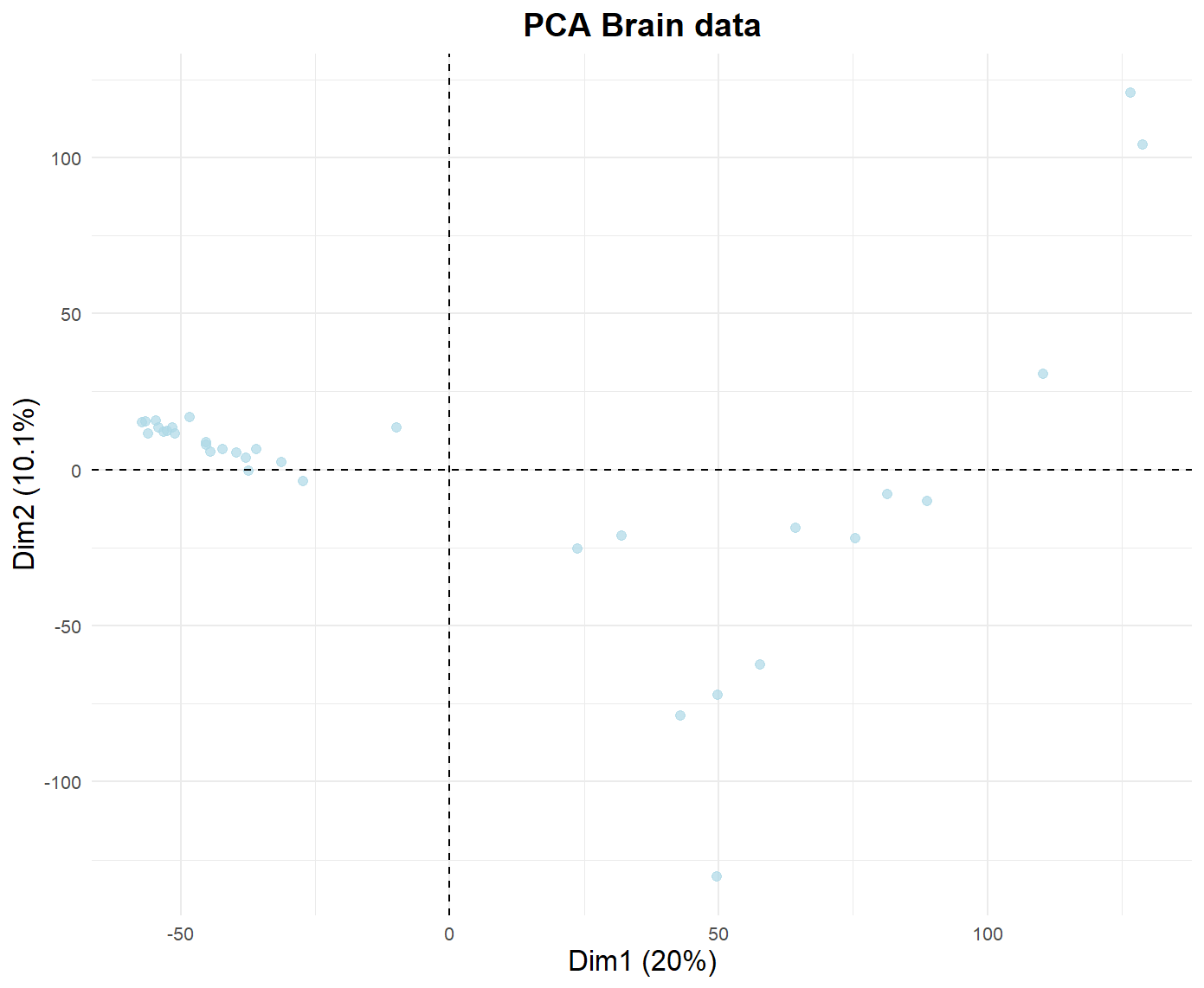
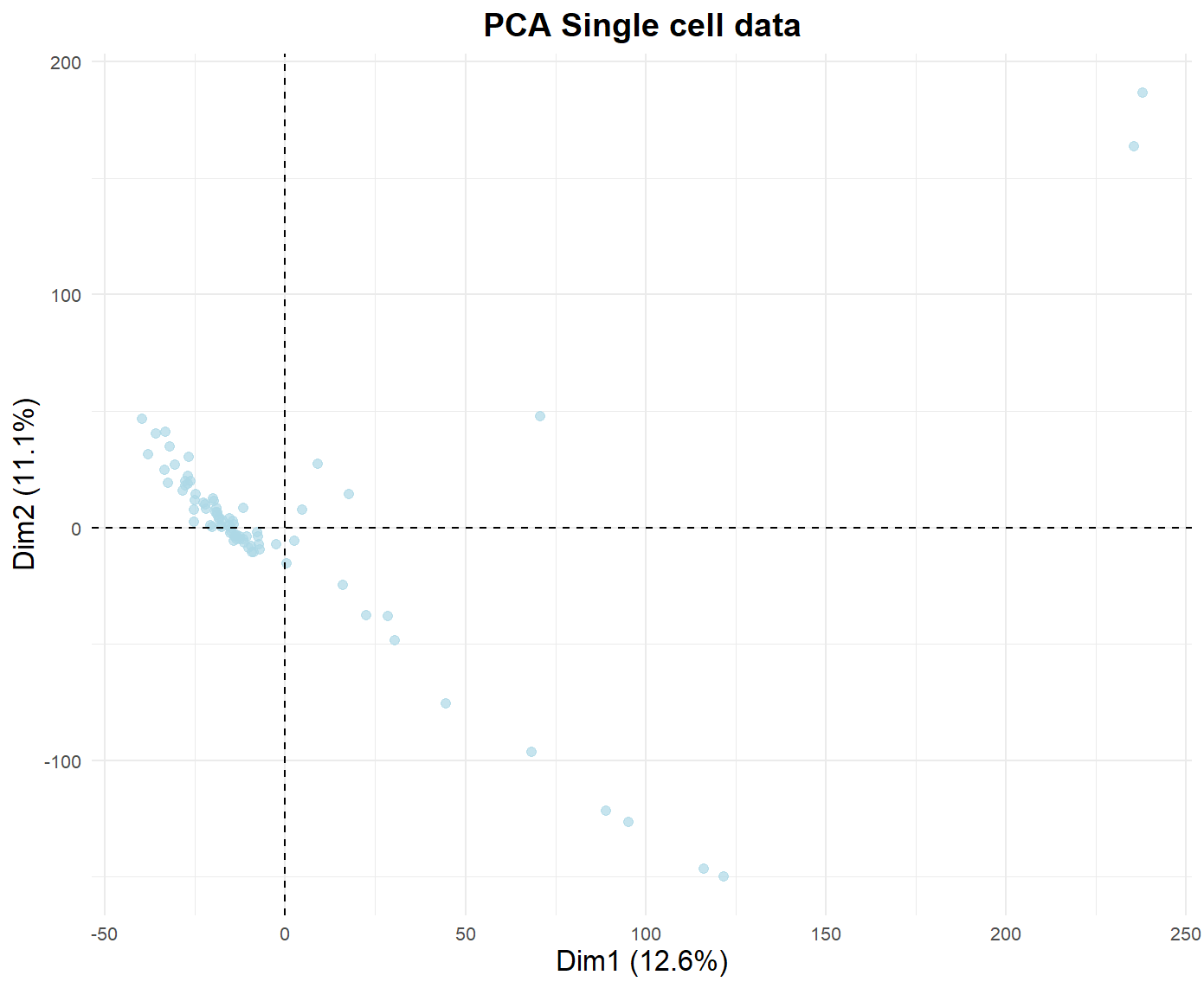
Summary of results

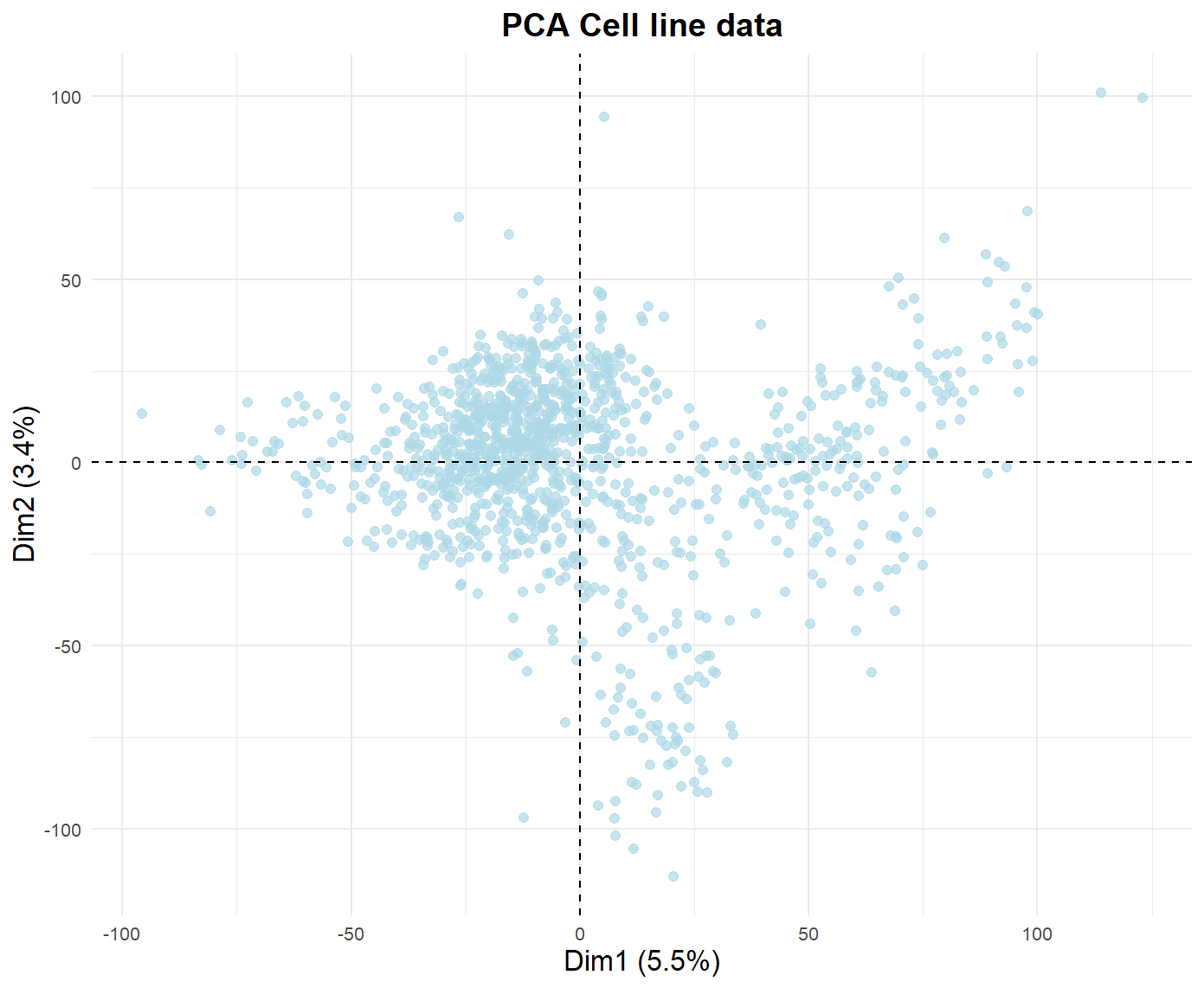
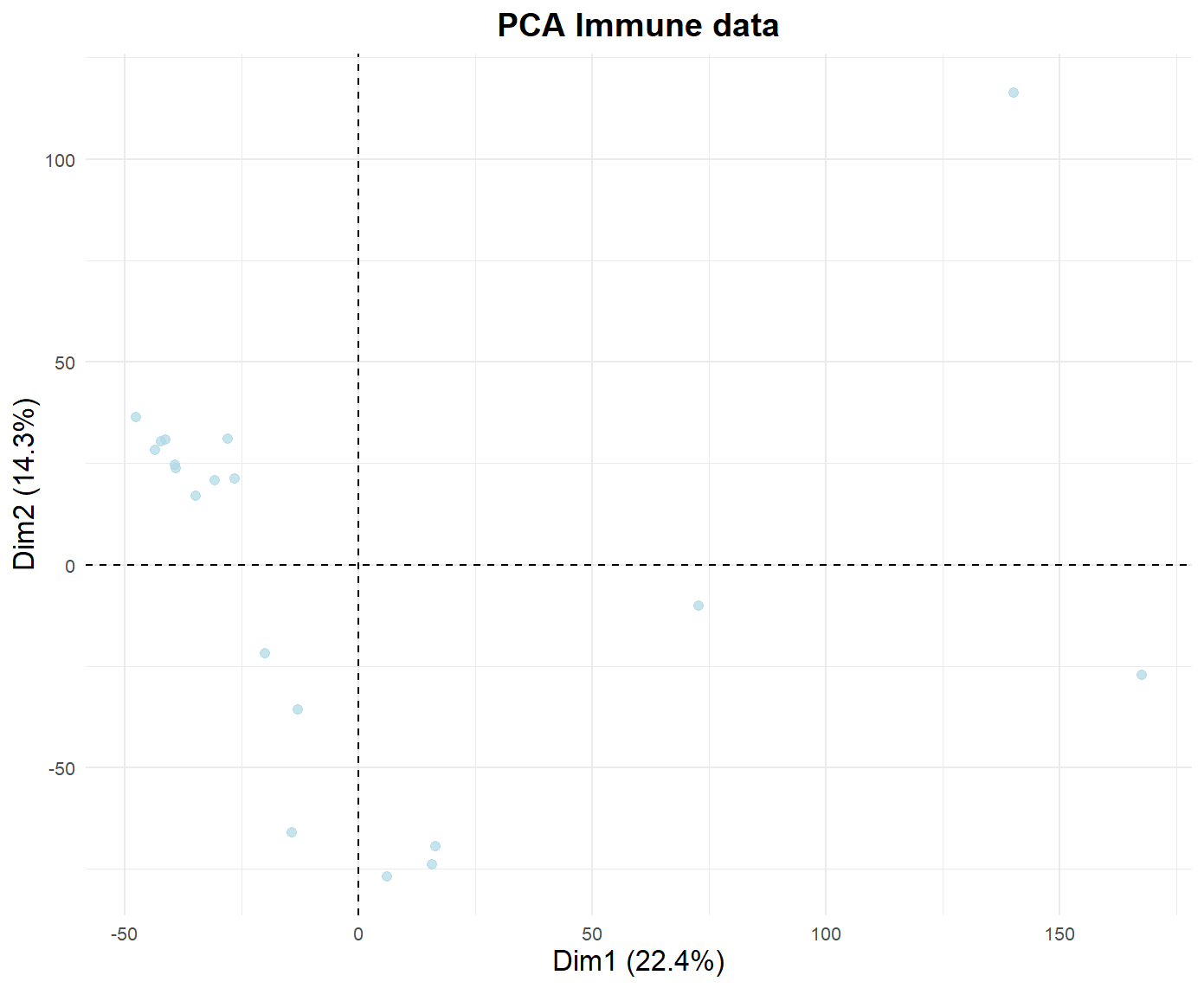
1. Data pre-process

* Data sets: Single cell type(81), single nuclei – (34), immune - flow sorted cells (19), cell lines (1206)
* Pre-process: I did UMAP and PCA on the sample profile for each data set
  + Sample profile means combining all the gene expression of one sample
  + **Note:** I used the wrong data last time (gene expression profile instead of sample profile, so the result is different from the last meeting)
  + **Discussion:** The data distribution seems normal



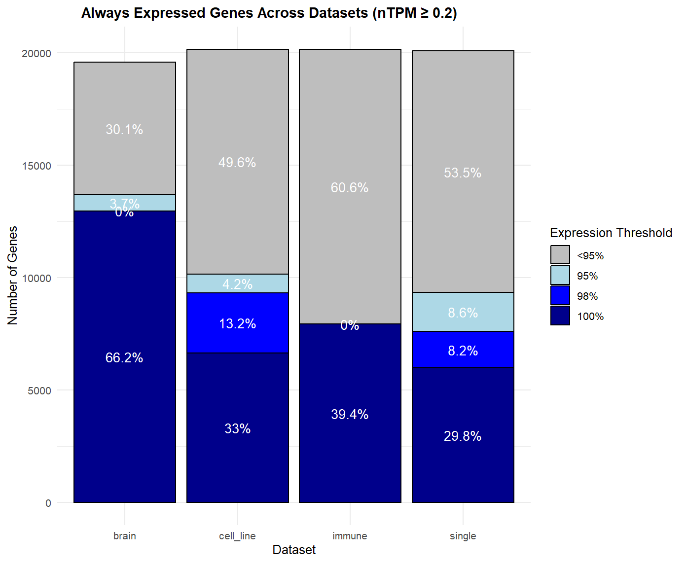
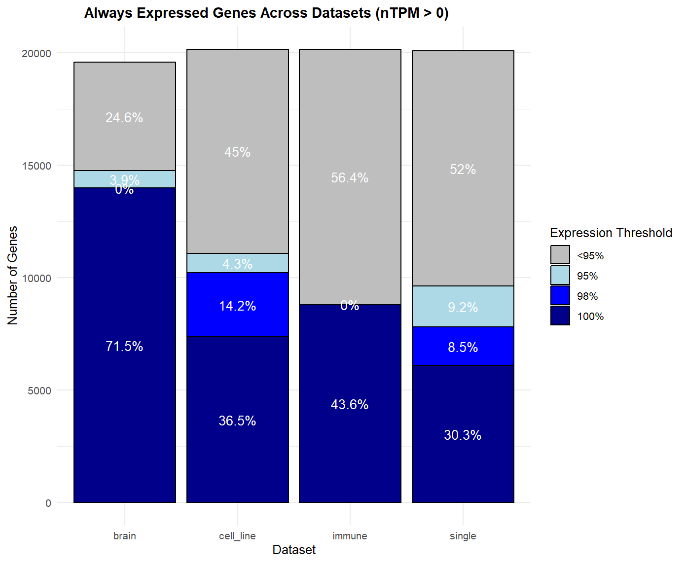


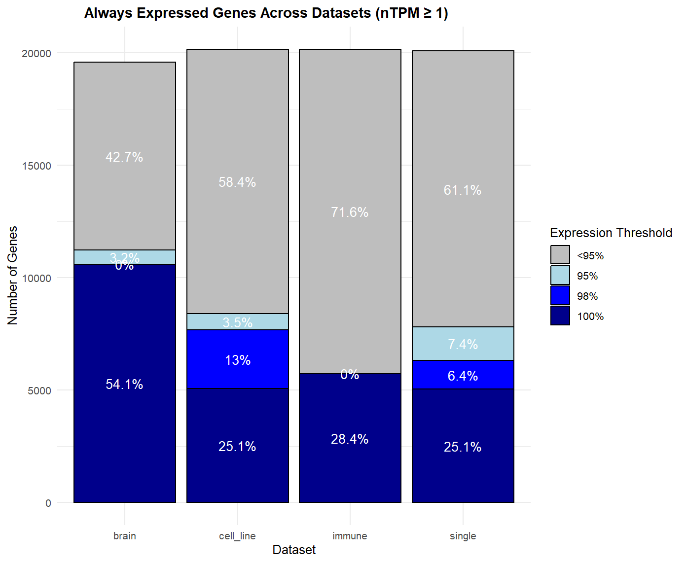
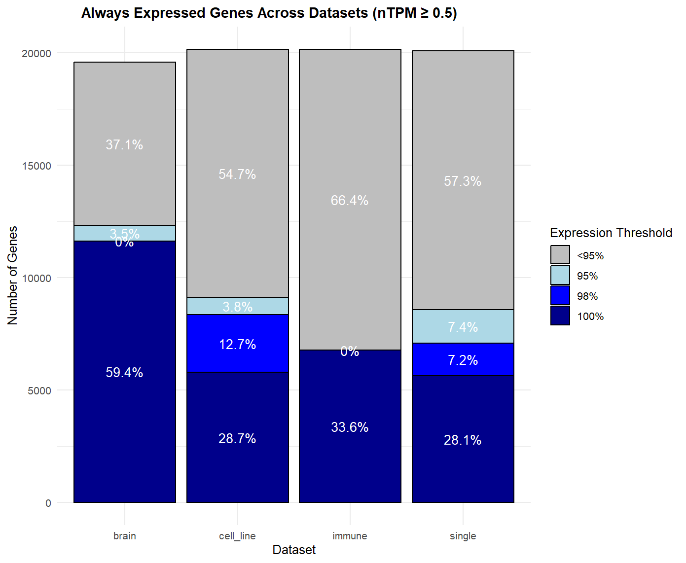




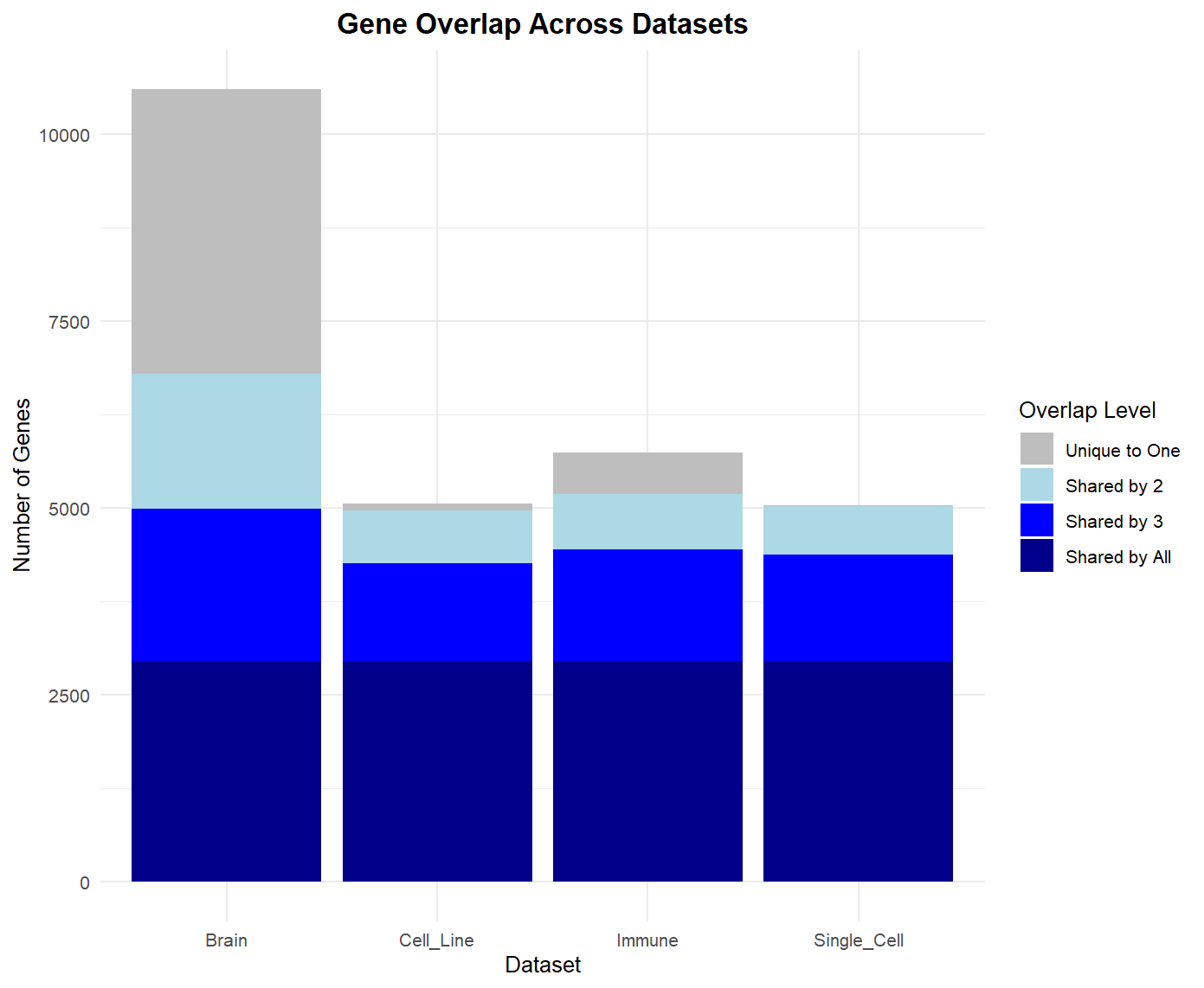
1. “Always expressed” gene identification

* “Always expressed” genes being identified by two thresholds:
  + nTPM min: >0, >= 0.2, >= 0,5, >= 1
  + The percentage of samples qualified the nTPM threshold: 100%, 98%, 95%
* The stacked bar plots illustrate the proportion of the percentage of samples qualified of the same nTPM threshold:
  + The 100% percentage is the percentage number at the bottom of the stacked bar
  + The 98% percentage is the sum of the second bottom of the stacked bar and the percentage of the 100% percentage. Since the 98% category is the 100% category and some additional genes qualified (e.g., expression passed nTPM threshold in 99% samples), only its additional part is illustrated in the stacked bar plot
  + Vice versa for the 95%

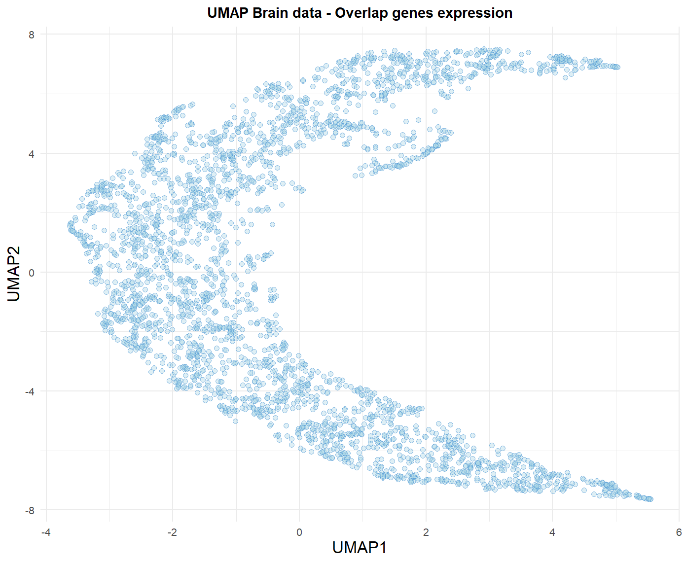
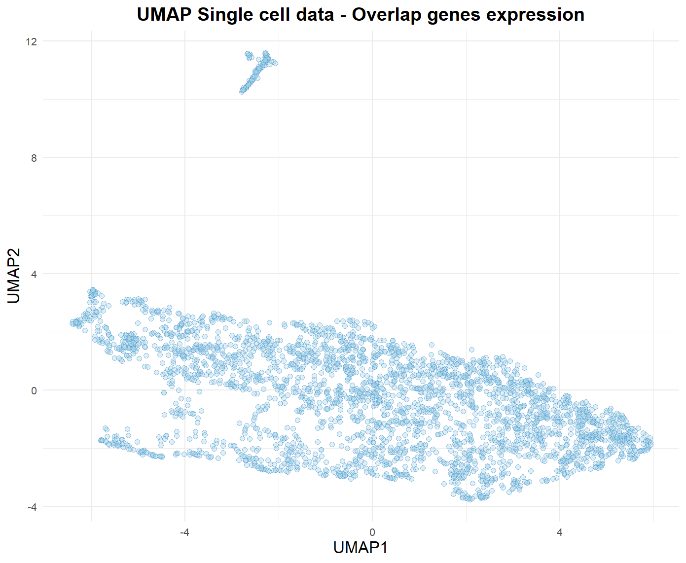




* The overlaps of the “always expressed” genes found in all four data sets:
  + The overlap was computed on the strictest thresholds: nTPM >= 1 found in 100% sample
  + There are about 2600 “always expressed” genes found in all data sets

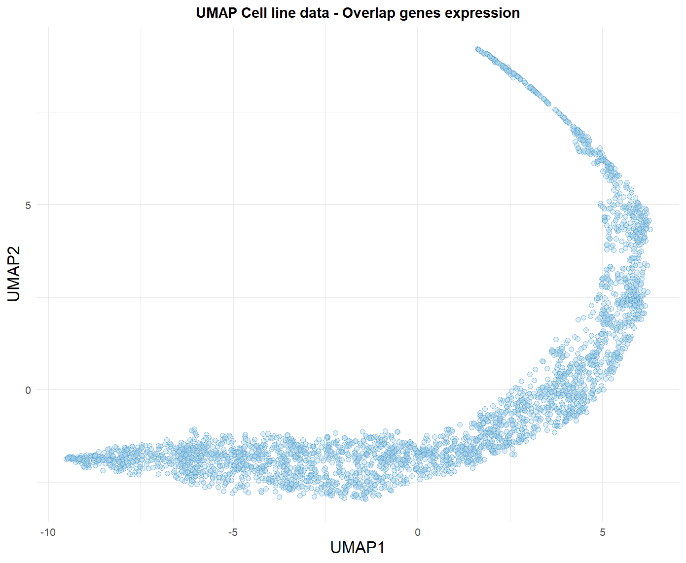
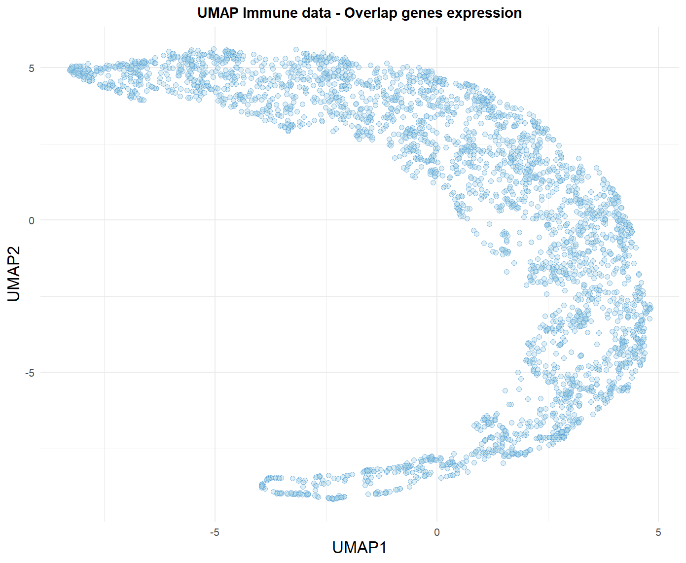


* For the “always expressed” genes found in 4 data sets, I did UMAP to check if they might share same patterns
  + I extracted the gene expression profile (the gene expression level across samples) for each “always expressed” gene in each data set
  + Their UMAP revealed if they had distinct expression trends
  + **Discussion:** only single-cell type data had two distinct clusters

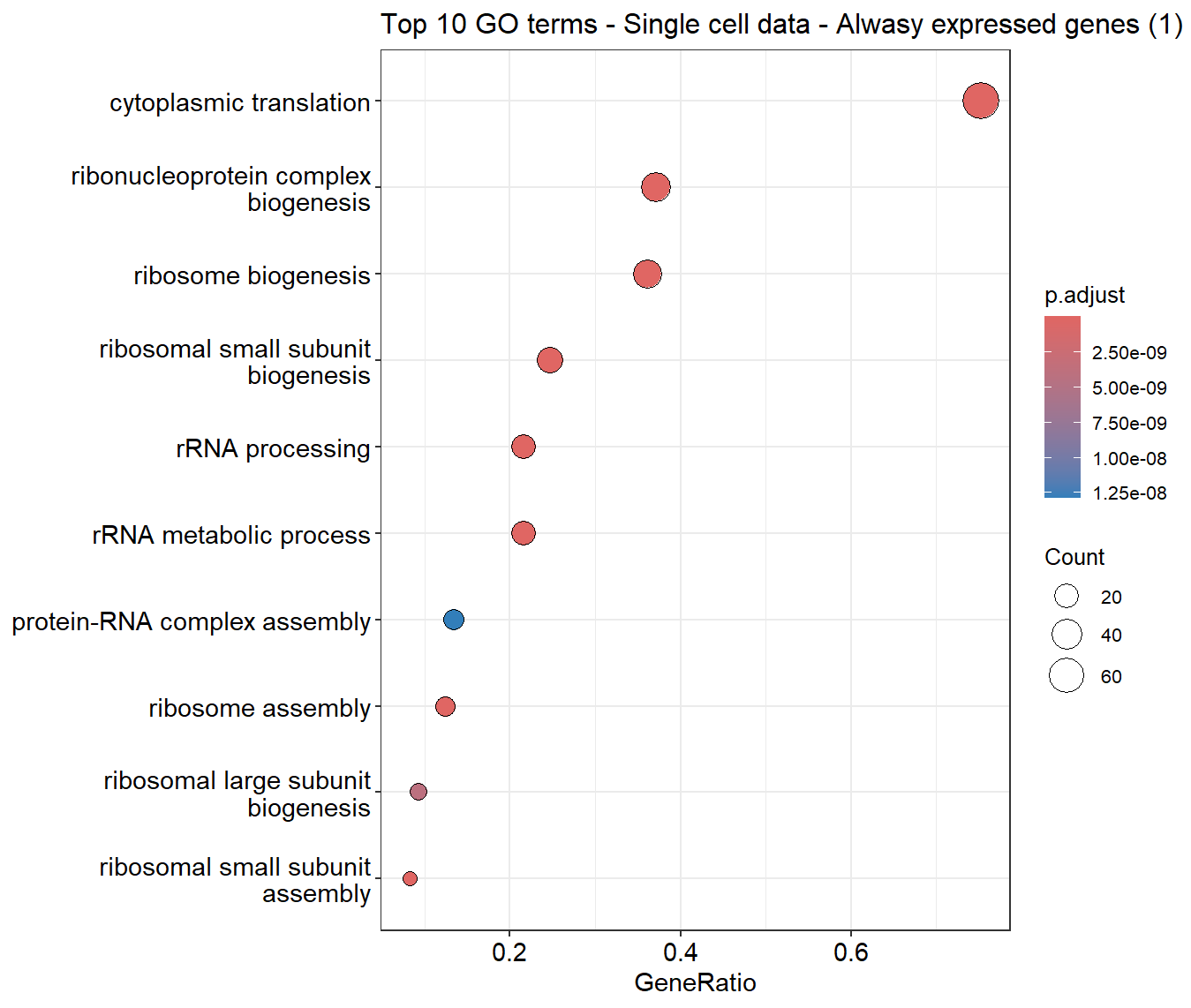


**2**

**1**



* I did GO enrichment of clusters 1 and 2 in the single-cell type data
  + Cluster 1 related more to protein synthesis
  + Cluster 2 related more to RNA maturation and modification



----------------------------------------------------------------------------------------------------------------

