PROJECT 1.
Automatic cluster of mouse gene regarding human brain gene

Introduction

The major central question of biology is to systematically identify how the genotype influences the phenotype. To understand the mediation of biological processes requires comprehending knowledge, how transcription factors (TFs) function sequence-specifically. This specific sequence recognized by TFs is called motif, whereas transcription factor binding sites (TFBS) are characterized. These TFBS can be regulated either proximal or distal to a transcription start site and can be activated or repressed depending on its context. Furthermore, altered activity of TF can play a pivotal role, for example TF expression variability in sequence and expression can consequently alter the expression of target gene in coordinate fashion.

Processing data

The steps below encompass the standard pre/processing workflow for scRNA-seq data with a certain package from R. The selection and filtration of analysis is based on QC metrics, data normalization and scaling, and the detection of highly variable features.

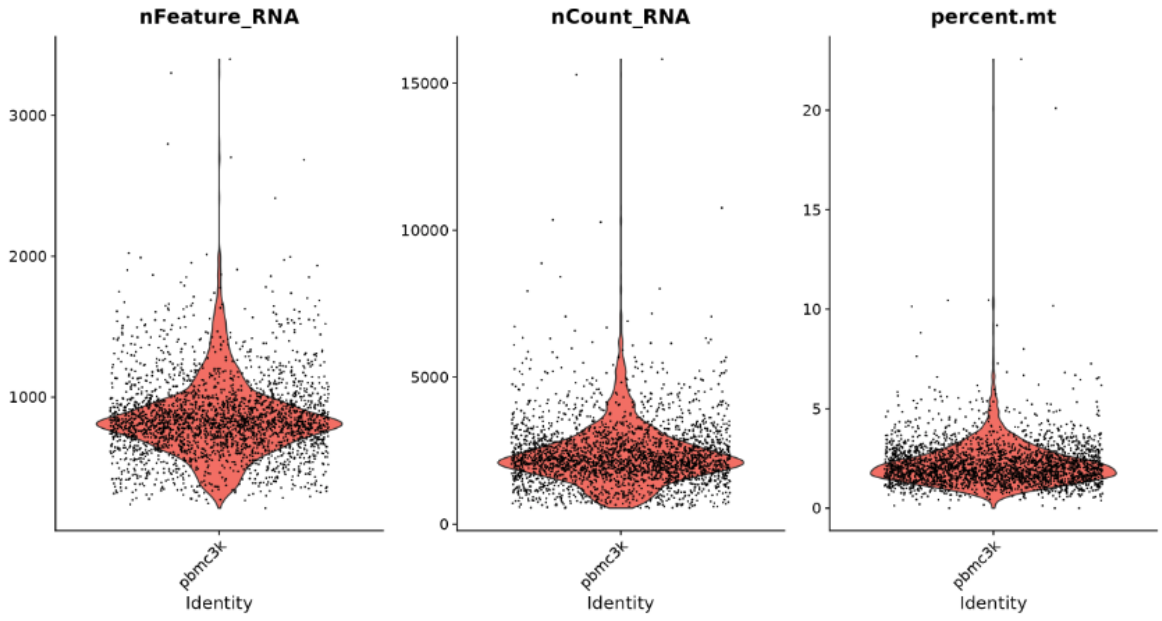


Fig1. Filtering the cells that have unique feature counts over 2,500 or less than 200 and cells that have >5% mitochondrial counts to exclude outliers and noises from contaminations and aberrantly exposures.

Identification of highly variable features

After removing unwanted cells from the dataset and global normalizing the data for clarity of

comparison, calculation of features is required to exhibit high cell-to-cell variation in the dataset. Focusing on certain genes helps to highlight biological signals in single-cell datasets.

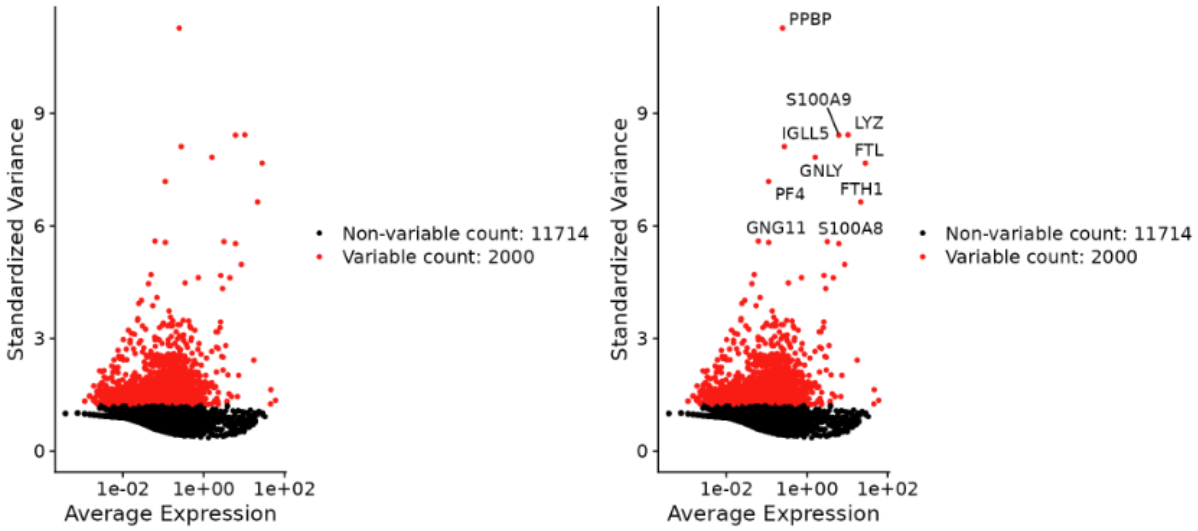
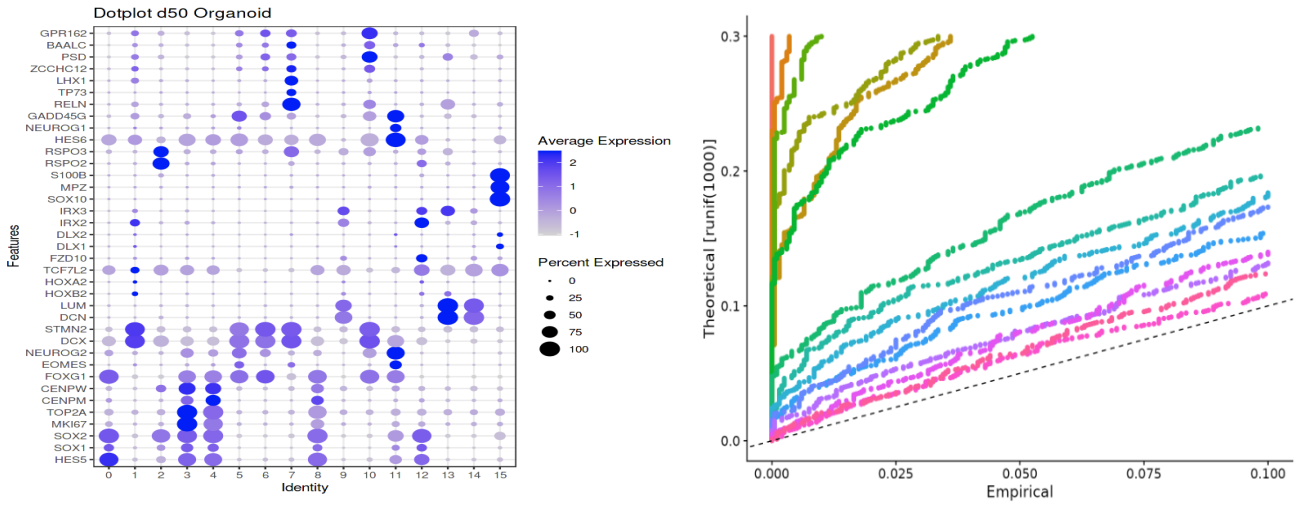


Fig.2 The highly expressed genes are marked on the right side with its gene names and modeled with the mean variance relationship inherent in single-cell data.

Linear dimensional analysis

With the help of PCA, the analysis allowed easy exploration of the primary sources of heterogeneity in a dataset, and chose carefully with its dimension for further downstream



analysis.

Fig.3 On the right side, it displays which genes are highly variable in certain identity dimensions. These genes are primarily interesting because of their representative of human brain genes. On the left side show the jackstrawplot to compare the distribution of p-values for each metafeatures of PCs. Significant PCs will show strong enrichment of features with low p-values.

Clustering the cells

The goal of these algorithms was to learn the underlying manifold of the data in order to place similar cells together in low dimensional space. One technique used here is UMAP, which can feature the clusters and co-localized.

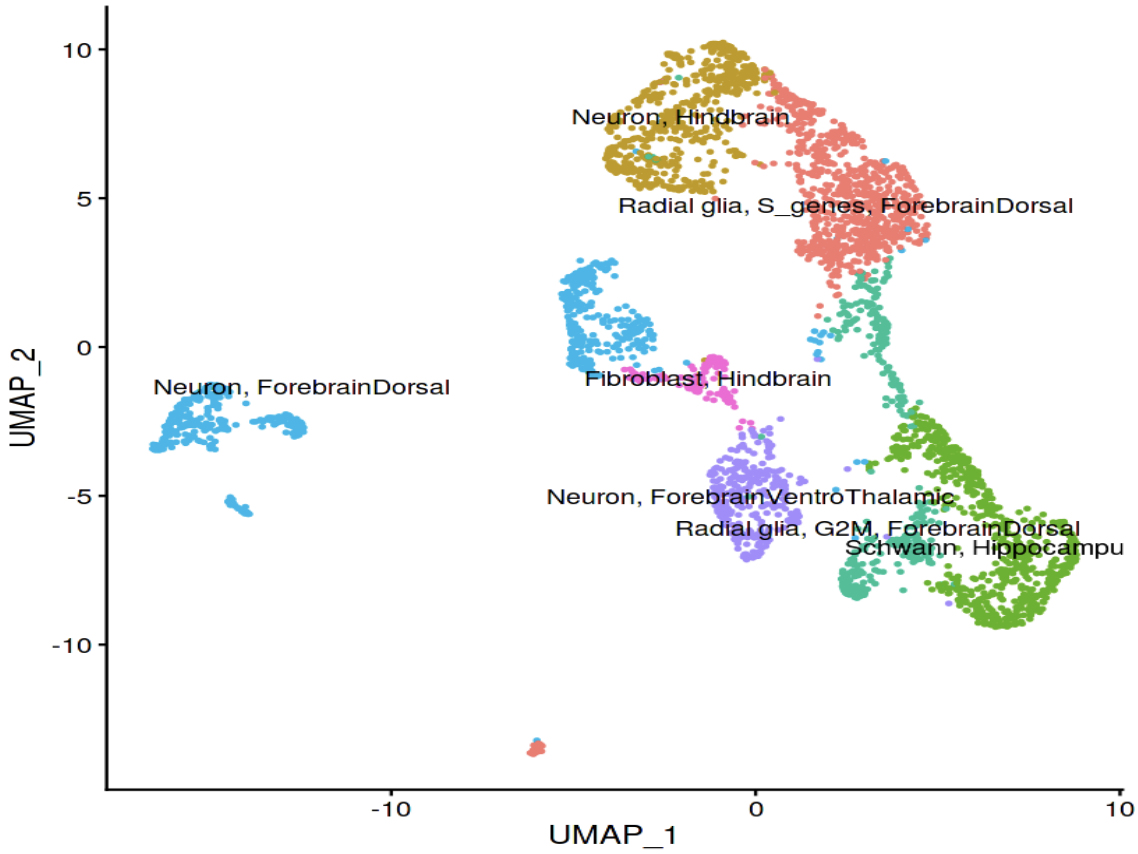


Fig.3 Each cells, which are having relevantly close expression levels, are clustered and the clusters are marked and named differently. Distinguishing clusters was performed by comparing every cluster to all remaining cells and reporting only positive ones.