ACB Assignment Report

Meghal Dani MT17144

November 11, 2017

1 Procedure

Downloaded single cell recount data in .tsv format from recount2 database (SRP057196). Performed Normalisation using TMM and later log transformation on the normalised data.
Applied 3 Clustering algorithms (k-means, fuzzy clustering and DBScan) and applied PCA to generate plots and analysed them.

2 Dataset

The samples in study contains adult and fetal human brain at whole transcriptome level.

3 Details of tools and algorithms

edgeR is a library in R to perform gene expression using count data.It can be installed using command : >source("http://www.bioconductor.org/biocLite.R") >biocLite("edgeR").

PCA is Principal component Analysis is done on the data when we have large dataset and we need to reduce its dimensions and without the loss of any important information in the data.This plays a very crucial role in data analysis.

K-means is a clustering method which partitions n observations in data into r clusters using the command :
>kmeans(data,5,nstart=20) Where data is the normalised and log transformed, next parameter is the number of clusters to be generated.We can see in the plot generated that there are 5 coloured clusters out of which green and blue show overlapping of data which show there is some correlation between the variables.The points in black i noise i.e., they are the outliers which don't fall into any cluster.

Fuzzy Clustering, fanny algorithm is used for clustering of a dataset using function fanny() and clusters the data with respect to euclidean distance,manhattan distance,etc.It is mainly used to check id data belongs to more than one cluster or not so as to define its properties.

DBScan is Density-based spatial clustering of applications with noise.It discovers clusters of arbitrary shape.It take 2 parameters eps : maximum radius of neighbourhood and MinPts : minimum no of points in eps neighbourhood of a point.The output can be seen divided as core,border and outlier wherein core point is dense neighbourhood , border point in cluster but neighbourhood is not dense and outlier is not in the cluster.

4 Interpretation

kmeans generated 5 clusters of different colors with 2 concentrated clusters due to overlapping of data which can be easily verified by Fuzzy clustering that there are datasets belonging to more than one cluster. DBScan plot shows core point, border point and outliers with only one cluster due to concentrated data.
Thus we can conclude that the dataset is a lot correlated.