ACB Assignment Report

Meghal Dani MT17144

November 12, 2017

1 Data Significance

To study the data of brain and gut gene expression in adult male head.

2 Procedure

Downloaded 2 FASTQ files for study SRR6238873 and SRR6238874 from the ncbi website and performed a quality check on the files using the command : fastqc SRR6238873.fastq fastqc SRR6238874.fastq The results are given in html format.

Tophat: for the human data hg19 is used tophat2 –bowtie1 hg19 SRR6238873.fastq tophat2 –bowtie1 hg19 SRR6238874.fastq .bam files are generated which will be used for raw count generation.The .bam files are sorted using samtools command :
$Samtools sort accepted_hits.bam -o accepted_hits_n sort.bam$

HTSeq is used to produce raw counts.

Now the read count data is loaded in R and normalisation is performed using TMM.This data will be used for Differential Expression analysis. And the normalised data is log transformed which is used for clustering and PCA. The clustering algorithm used is k-means and fanny(fuzzy clustering) algorithm and plots are generated after applying PCA.

Differential expression analysis is done using library DESeq which generates pvalue and log fold change and then most important genes are selected. The genes with p-value < 0.001 are taken to differentially expressed and a heatmap is plot to visualize this.

3 Details of tools and algorithms

DESeq2 is the R library used for Differential Expression Analysis based on negative binomial distribution.It is installed using the command :
source("https://bioconductor.org/biocLite.R")
biocLite("DESeq2")