This homework concerns the analysis of RNA-seq data, in particular the identification of differentially expressed (DE) genes and isoforms. EBSeq is an empirical Bayes approach available at BioConductor and at http://www.biostat.wisc.edu/~kendzior/EBSEQ/

The data sets you will consider were obtained from a study of human islet done in Dr. Alan Attie’s lab here at UW-Madison. The mRNA’s are from 20 separate samples: 10 with an over-expressed gene of interest (Asf1b) and 10 control. The data was preprocessed via Bowtie and RSEM. Gene level expression estimates, isoform level expression estimates, and the file indicating uncertainty group assignment for each isoform were obtained using RSEM-1.2.6 (gene.rsemdata.txt, isoform.rsemdata.txt and NgV.csv, respectively).

1. Use EBSeq to identify genes that are DE using gene.rsemdata.txt. (a) Report the number of genes that are DE with the false discovery rate (FDR) controlled at 5%. (b) What is the FDR associated with the top 50 most DE genes (meaning those with highest posterior probability of being DE). (c) Check diagnostics. Does the model fit well? Are the results reliable? (d) Generate a plot that considers at least 10 DE genes to see if they appear to be DE (this can be a heat map, a series of bar charts, or some other plot(s) of your choice). The goal is to see if there is visual evidence in favor of the DE calls. Note that if you generate a heat map, you will likely want to consider more than 10 genes.

2. Repeat 1, for the isoform data (isoform.rsemdata.txt).