Homework Assignment 8
(Due Wed, November 18, 2015 before class)

Please hand in a print-out of your answer and R code, and also email your R code to Zhiyuan (Jason) Xu <xuxx0284@umn.edu>.

1. This is in continuation of Question 1 in Homework Assignment 7. The workflow of microarray data analysis usually follows the steps of (1) reading in data (often from binary files), (2) normalization, (3) differential expression detection and (4) generate report. We will focus on 3 and 4 in this question. We will continue to use the following packages from Bioconductor: oligo (for reading in data and normalization), limma and siggenes (for differential expression), pd.hg.u133.plus.2 (for annotation and generating reports). Refer to Homework 7 to download the microarray data provided through gene expression omnibus under accession number GSE18088.
   Questions:
   (1) How many patients developed relapse events? (5 points)
   (2) In order to identify the differential expression genes between patients with relapse events and patients without relapse, what is the design matrix for this comparison? (20 points)
   (3) Use limma to detect differentially expressed genes between patients with relapse events and patients without relapse. (35 points)
   (4) How many genes are differentially expressed under FDR < 0.05 in U133Plus2 platform? How many genes with p value < 0.05? What are the top 30 differentially genes among them? [Hint: use **hgu133plus2SYMBOL** to convert Affymetrix probe id to Entrez gene symbols] (40 points)

2. Do these significant genes come from certain pathways? Please perform the following analyses to find out.

(1) Perform gene set enrichment analyses for the genes with p value < 0.05 using hypergeometric test. Use the gene sets defined in KEGG. [Hint: Useful R code in Lab18 and Chapter 14 in Bioconductor Case Studies] (40 points)
(2) What might be potential problems using the approach in (1)? (20 points)
(3) Perform gene set enrichment analyses for the differentially expressed genes identified through U133Plus2 platform using alternative approach. How many KEGG pathways have p value < 0.05 using gene set enrichment tests? What are the top 10 enriched KEGG pathways? [Hint: use **Category** package and some useful R code in **Chapter 13** in Bioconductor Case Studies] (40 points)