Homework Assignment 5
Total Points: 115
(Due Thursday October 26, 2023 at 5PM)

Please email your answer (compiled pdf file from R markdown) and R code to Cenxiao (CENXIAO@email.sc.edu).

1. Read in the h5data.txt and weight data from the class website (command below),

```
>h5data.url<-"http://people.stat.sc.edu/hoyen/STAT588/Data/h5data.txt"
>weight.url<-"http://people.stat.sc.edu/hoyen/STAT588/Data/weight.txt"
>h5data<-read.table(file=h5data.url, header=T, sep="\t")
>map.url<-"http://people.stat.sc.edu/hoyen/STAT588/Data/SNPmap.txt"
>map<-read.table(map.url, header=T, sep="\t", stringsAsFactor=F)
>weight<-read.table(file=weight.url, header=T, sep="\t")$x
>rownames(h5data)<-map$name
```

(a) How many individual(s) have missing genotype > 80%? How many markers have missing > 80%? How many markers have p values < 0.05 from Hardy-Weinberg equilibrium test? How many markers have minor allele frequency (MAF) <5%? Remove the individual(s) or marker(s) data with missing genotype percentage >80%, or with p value < 0.05 from HWE test or MAF<5%. (10 points)

(b) Perform genome-wide association analysis one-snp-at-a-time, what is the most significant SNP marker? (10 points)

(c) Plot Manhattan plot (10 points)

(d) Plot QQ plot (5 points)

(e) Perform multiple comparison adjustment controlling false discovery rate at 0.05. Use both q value and Benjamini Hochberg approaches. (5 points)

(f) Produce a table with the top 10 most Significant SNP markers. The table should contain the following columns: Name of the SNP markers, Chromosome location, position, minor allele, minor allele frequency, p value adjusted p value (using Benjamini Hochberg). For example:

```
    SNP       CHR  Position      Minor Allele      MAF    p value        adj P
1   s15151.1   1    259949980          B         0.43     1.31e-05        0.49
2   s65168.1   1   132566495           B         0.27      5.02e-05       0.49
…
```

(10 points).

2. In this question, we will go through the typical microarray data analysis procedure using some popular Bioconductor packages. 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 show

that using Bioconductor packages these tasks can be easily achieved in a few lines of R codes.
The Bioconductor packages to be used in this homework are: oligo (for reading in data and normalization), limma and siggenes (for differential expression), pd.hg.u133.plus.2 (for annotation and generating reports).

Data:
We will use the microarray data provided by gene expression omnibus (GEO) database under accession number GSE18088. For details of the data please visit the webpage below.
http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE18088

The analysis results of this dataset were published earlier cited below.
Gröne J, Lenze D, Jurinovic V, Hummel M et al. Molecular profiles and clinical outcome of stage UICC II colon cancer patients. Int J Colorectal Dis 2011 Jul;26(7):847-58. PMID: 21465190

Steps:
1. Install following Bioconductor packages: affy, oligo, limma, siggenes, pd.hg.u133a, pd.hg.u133.plus.2.
2. Do following to download and decompress the raw data from the tar file.
a. Go to GEO database homepage at http://www.ncbi.nlm.nih.gov/geo/.
b. Put "GSE18088" in the Datasets query box and click "go".
c. Click the top query result link to go to the summary page for GSE18088. You can briefly read the descriptions.
d. Go down to the bottom of the page, under "Supplementary file" table, download GSE18088_RAW.tar (243.6Mb) by clicking the ftp or http link.
e. Decompress the tar file to your folder using a software. The individual files contained in the tar are gz files (compressed). You don't need to decompress them because the function in oligo can read in without decompressing.

Questions, 8 points each.

(1) What is the platform they used to generate the gene expression data? Follow the steps describe above to obtain the CEL files, and read the data into R. [Hint: read.celfiles()]
(2) Examine the distributions of raw intensities using density and box plots. Describe what you see in the distribution plots. Do you observe any sign of batch effect? (15 points)
(3) Make MA plots using raw intensities for all the samples [Hint:MAplot()]; describe what you see in the MA plots. (15 points)
(4) Perform normalization of the raw data (10 points)
(5) Make densities, box plots and MA plots for the normalized intensities for all samples. Describe what you see in the plots.

3. Short answer questions, 5 points each. Be creative in answering the questions.

(1) What does a gene expression microarray measure? Why is it important?

(2) What is hybridization? The amount of hybridization on different probes are extracted from the images and called “fluorescent intensities”. What do the fluorescent intensities of each probe represent?
(3) Why does one have to normalize the microarray data?
(4) What is quantile normalization? Can you think of any pitfall of the procedure?
(5) What is summarization of gene expression? Can you design a simple method to do the summarization?