Assignment 1

• The `golub` dataset implanted in the `multtest` package is gene expression data extracted from the leukemia microarray study of Golub et al. (1999). They were interested in identifying genes that are differentially expressed in patients with two types of leukemias, acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). Gene expression levels were measured using Affymetrix high-density oligonucleotide chips containing $p = 6817$ human genes. The dataset comprises $n = 38$ samples, 27 ALL cases and 11 AML cases. three preprocessing steps were applied to the normalized matrix of intensity values available on the website: (i) thresholding: floor of 100 and ceiling of 16,000; (ii) filtering: exclusion of genes with $\max / \min \leq 5$ or $(\max - \min) \leq 500$, where max and min refer respectively to the maximum and minimum intensities for a particular gene across mRNA samples; (iii) base 10 logarithmic transformation. The expression levels were standardized within arrays before combining data across samples. The data were then summarized by a 3051×38 matrix $X = (x_{ji})$, where x_{ji} denotes the expression level for gene j in sample i. The `golub` in `multtest` includes

- `golub`: a 3051×38 matrix of expression levels;
- `golub.gnames`: a 3051×3 matrix of gene identifiers;
- `golub.cl`: a vector of tumor class labels (0 for ALL, 1 for AML).

To load the leukemia dataset, use

```
> library(multtest)
> data(golub)
```

Perform the following analyses to `golub` dataset.

1. Construct the baseline arrays for ALL and AML by taking sample median of each gene. For example, the baseline array of ALL can be obtained via the following R codes:

   ```
   > ALL = golub[,which(golub.cl==0)]
   > base.ALL = apply(ALL,1,median)
   ```

   (Type `?apply` to view the useage of the function `apply`.) Use appropriate visualization tools to compare the expression levels of the two baseline arrays. Do the expression values differ for the two types of leukemia?

2. Identify genes that were differentially expressed in the two classes (ALL/AML). Conduct (1) two sample t-test, (2) two sample Wilcoxon test, and (3) permutation test and adjust their p-values to control FWER and FDR, respectively.

3. Collect the genes that were significant according to two sample Wilcoxon test at FDR = 0.01 based on BY algorithm. Use these genes to perform cluster analysis. You can pick any distance measure. Provide the clustering results of (1) one of the partitioning method with appropriate number of clusters (2) one setting of hierarchical method, and (3) SOM with grids of your choice.

4. Write a short paragraph to summerize your findings.