MSCBIO 2070/02-710: Computational Genomics, Spring 2015

A3: Phylogeny, Sequence alignment, DE gene analysis

Due: March 16, 2015 by email to Marta Wells (mmw88@pitt.edu)
TA in charge: Marta Wells (Q1, Q2) & Silvia Liu (Q3, Q4)

Your goals in this assignment are to

1. Build a phylogenetic tree;
2. Implement sequence alignment algorithm (global and local);
3. Analysis of DE genes;
4. Write code to detect DE genes in the real data.

What to hand in. Write a short report addressing each of the questions below (either a hand-written report or a report submitted as a pdf file is acceptable). The report should be self-contained, we should not have to run your code to be convinced that your code is correct. Be sure to comment your code and use any programming language you like (if not specified). Also, include instructions on how to run your code. In your report you can assume that we know the context of the questions, so do not spend time repeating material in the hand-out or in class notes. Email a zip file containing the complete code (if any) and pdf file (if any) to: Marta Wells, *mmw88@pitt.edu* with the Subject: *2070S15 A3 YourName*.

1 [15 points] Phylogeny

We will investigate the evolutionary relationships between 14 species of mammals with UPGMA and Neighbor Joining algorithms. We will see that even though the two methods disagree to some extent, they also agree on some divisions. The DNA used in the tree construction is from the interphotoreceptor retinoid binding protein. Sequences for the 14 species were taken from Genbank, aligned using CLUSTALW and a 532 nucleotide ungapped sub alignment extracted by eye and used as input.

The 14 species are: (1) Marsupial Mole, (2) Wombat, (3) Rodent, (4) Elephant Shrew, (5) Elephant, (6) Whale, (7) Dolphin, (8) Pig, (9) Horse, (10) Bat, (11) Insectivore, (12) Human, (13) Sea Cow, (14) Hyrax.

The evolutionary distances between sequences were extracted under Felsenstein model (not discussed in class) and reported in the file: `evolDist.txt` (downloadable from course webpage).

Your task

1. (7 points) Download the distance matrix from the course webpage (`evolDist.txt`) Construct the phylogenetic tree using UPGMA.
2. (7 points) Use the same distance matrix to construct a phylogenetic tree using Neighbor Joining algorithm.
3. (1 points) Comment on the similarities and differences between the two trees.

Please see figure 1.

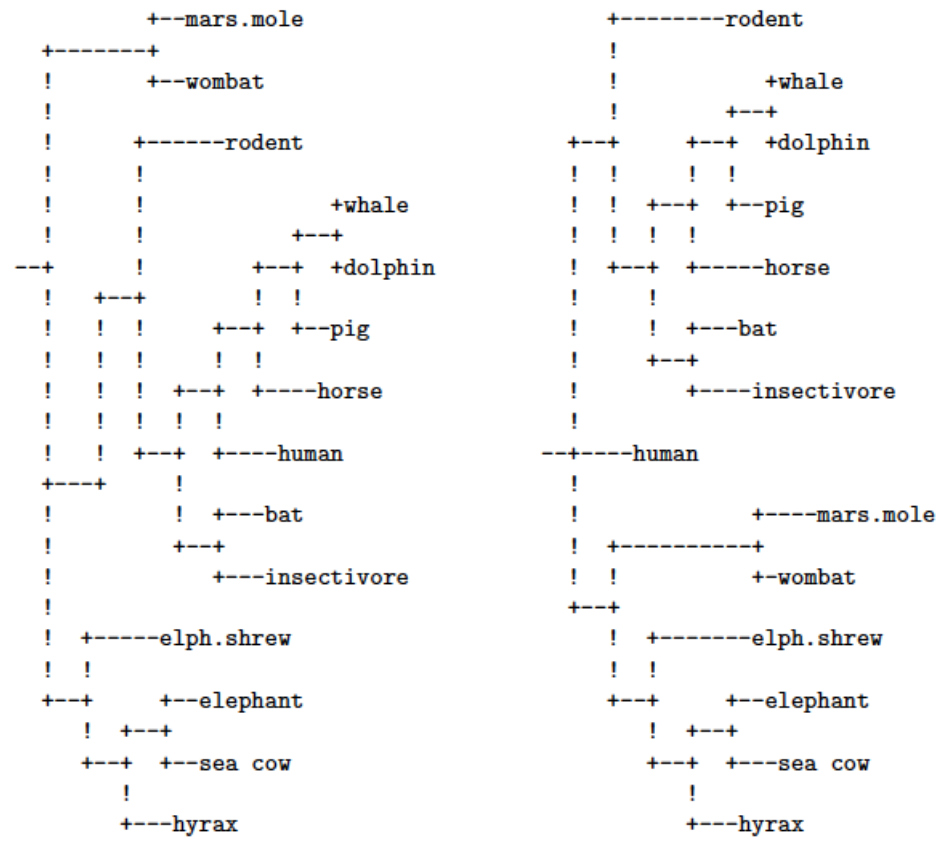

Figure 1: Phylogenetic tree

2 [30 points] Sequence alignment

You have been sent a gene to analyze. Someone labeled it as `Mus sapiens`. Having never met a `Mus sapiens`, you can only conclude that it is either from a mouse (`Mus musculus`) or human (`Homo sapiens`). The gene is in `mus_sapiens.fasta`.

Your task

1. (4 points) Which organism is it from? Explain why. What gene is it? What does it do?

   Blast will tell you that the gene is *Homo sapiens* PFKL, or phosphofructokinase which takes part in glycolysis in the liver.

2. (12 points) You would like to show every place that the gene differs from the version in the other organism. Grab the human and mouse sequences of this gene. Write a program to perform global alignment on them. Use a match score +5, mismatch penalty of -4 and a gap penalty of -5. Run your algorithm, report your alignment and its score.

   Depending on which file you downloaded either 6823, or 7315.

3. (6 points) Run the alignment using an affine gap penalty, with a gap open penalty of -10, and a gap addition penalty of -1. How do the alignments compare?

   This was pretty variable depending on if you did local or global, so I was pretty lenient. Doing global alignment I got around 7000 and local I got around 9000.

4. (8 points) You would like to compare the protein sequences of these genes. Write a program to find the best locally aligned region. Use `BLOSUM62` and set gap open/extend penalties as -5. Run your algorithm and report the alignment result.

   You should have found that the full protein sequence was the best alignment with a score of 3828.

3 [15 points] Differential Expression Analysis

Statistical tests were performed for identifying differentially expressed genes by some method. Assume that we know the true results of the test and we wish to compare it to the method we used. The following table contains the results of this comparison.

		Test Result		Total
		DE	Not DE	
True Result	DE	S=360	R=125	V=485
	Not DE	Q=550	P=8965	U=9515
Total		N=910	M=9090	T=10000

Figure 2: Number of DE and non-DE genes

Your task

1. (4 points) The false positive rate (FPR) is the proportion of negative instances that were erroneously reported as being positive. Compute the FPR for the above tests. How does decreasing the true number of not-differentially-expressed genes affect the false positive rate? Based on the FPR, is the method used for testing a good method? Why?

$$FPR = \frac{Q}{U} = \frac{550}{9515} = 5.78\%$$

Decreasing true not DE genes will not affect the FPR since the true not DE genes are always high. Hence, FPR is not a useful error measure.

2. (4 points) In multiple hypothesis testing, the False Discovery Rate (FDR) is defined as the expected proportion of incorrectly rejected null hypotheses (or expected proportion of false positives among the declared significant results). As discussed in class, we see that the FDR is a useful measure to control. For the above tests, compute the FDR, sensitivity, and specificity.

$$FDR = \frac{Q}{N} = \frac{550}{910} = 60.44\%$$
$$\text{sensitivity} = \frac{S}{V} = \frac{360}{485} = 74.22\%$$
$$\text{specificity} = \frac{P}{U} = \frac{8965}{9515} = 94.22\%$$

3. Controlling false positives is an important issue in the identification of differentially expressed genes. FDR is a useful measure to control, in order to control the false positives. In multiple hypothesis testing, another error measure is the family wise error rate (FWER): it is defined as the probability of making one or more false discoveries.

   - (4 points) Show that the FWER equals the FDR when all the null hypotheses are true.

$$FWER = P(Q \geq 1)$$

Assume null hypothesis (gene is not DE) is true. If $S = 0$, hypothesis rejected will be falsely rejected. Therefore $N = Q + S = Q$. If test is good, we would like $Q = 0$.

Thus we can get $FWER = FDR = 0$. But if $Q > 0$, then all null hyp rejected will be falsely rejected. Thus we can get $FWER = FDR = 1$.

- (3 points) Show that any procedure that controls the FWER also controls the FDR. *[Hint:* $E(X) = E(X|A = 0)P(X|A = 0) + E(X|A \geq 1)P(X|A \geq 1)$*]*

  Assume $X = Q/N$.

$$FDR = E[X] = E[X|Q = 0]P(Q = 0) + E[X|Q \geq 1]P(Q \geq 1)$$

  Suppose at least one null hypothesis is false. This implies that $Q \geq 1$).

$$FDR = Q/N \leq 1 \quad \rightarrow \quad E[X|Q \geq 1] \leq 1$$

$$\rightarrow E[X|Q \geq 1]P(Q \geq 1) \leq P(Q \geq 1) = FWER$$

  So that any procedure that controls the FWER also controls the FDR.

4 [40 points] DE Genes Detection: Application in Microarray Data

Following Question 4 in PS2, we will continue to detect deferentially expressed genes from microarray data step by step. You can select any programming language you like. MATLAB or R are encouraged and relative ready-made functions will be provided in the hint. Please insert the plots into your write-up and submit your script file(s) separately.

4.1 [2 points] Load the workspace of Q4 in PS2

Your task

We will keep on using the data *GeneExpres.txt* in PS2. Following the previous steps in Q4 of PS2, please impute the missing value by KNN method and normalize the data by quantile normalization. If you saved the working image in the last homework, you can load it directly into your workspace.

4.2 [15 points] DE analysis: p-value calculation

After data pre-processing, we can determine differentially expressed genes between the two conditions. We will use the quantile normalized data to implement two DE tests.

Method 1: log likelihood ratio test

The log likelihood ratio test is a statistical test to compare null models and alternative models. The test statistic D can be written as,

$$D = -2 \ln \frac{L(0)}{L(1)}$$

where $L(0)$ is the likelihood for null model and $L(1)$ is the likelihood for alternative model.

In this question, for each gene, we have the following hypothesis,

$$\begin{aligned} H_0: &\quad y_A \sim N(\mu_0, \sigma^2) \quad y_B \sim N(\mu_0, \sigma^2) \\ H_A: &\quad y_A \sim N(\mu_A, \sigma^2) \quad y_B \sim N(\mu_B, \sigma^2) \end{aligned}$$

where σ^2 is the variance of each gene. We can calculate empirical σ^2 from the sample.

Your task

1. (1 point) Write down the likelihood formulas under the null and alternative model.

$$\begin{aligned} L_i(0) &= \prod_{j \in A} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i^j - \mu_0)^2}{2\sigma^2}} \prod_{j \in B} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i^j - \mu_0)^2}{2\sigma^2}} \\ L_i(1) &= \prod_{j \in A} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i^j - \mu_A)^2}{2\sigma^2}} \prod_{j \in B} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i^j - \mu_B)^2}{2\sigma^2}} \end{aligned}$$

2. (1 point) Write down the test statistic D.

$$
\begin{aligned}
D_i &= -2 \ln \frac{L_i(0)}{L_i(1)} \\
&= -2 \ln \frac{\prod_{j \in A} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i^j - \mu_0)^2}{2\sigma^2}} \prod_{j \in B} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i^j - \mu_0)^2}{2\sigma^2}}}{\prod_{j \in A} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i^j - \mu_A)^2}{2\sigma^2}} \prod_{j \in B} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i^j - \mu_B)^2}{2\sigma^2}}} \\
&= \frac{1}{\sigma^2} \left(\sum_{j \in A} (y_i^j - \mu_0)^2 + \sum_{j \in B} (y_i^j - \mu_0)^2 - \sum_{j \in A} (y_i^j - \mu_A)^2 - \sum_{j \in B} (y_i^j - \mu_B)^2 \right)
\end{aligned}
$$

3. (1 point) D follows χ^2 distribution. What's the degree of freedom in this case? Why?

   There are two free parameters for null hypothesis (μ_0, σ) and three parameters for the alternative hypothesis (μ_A, μ_B, σ). Thus the difference is 1. So the degree of freedom for log likelihood ratio test is 1.

4. (4 points) Implement the log likelihood ratio test for the quantile normalized data and calculate the one-sided p-values for each gene. Plot the histogram of the p-values.

   See figure 3.

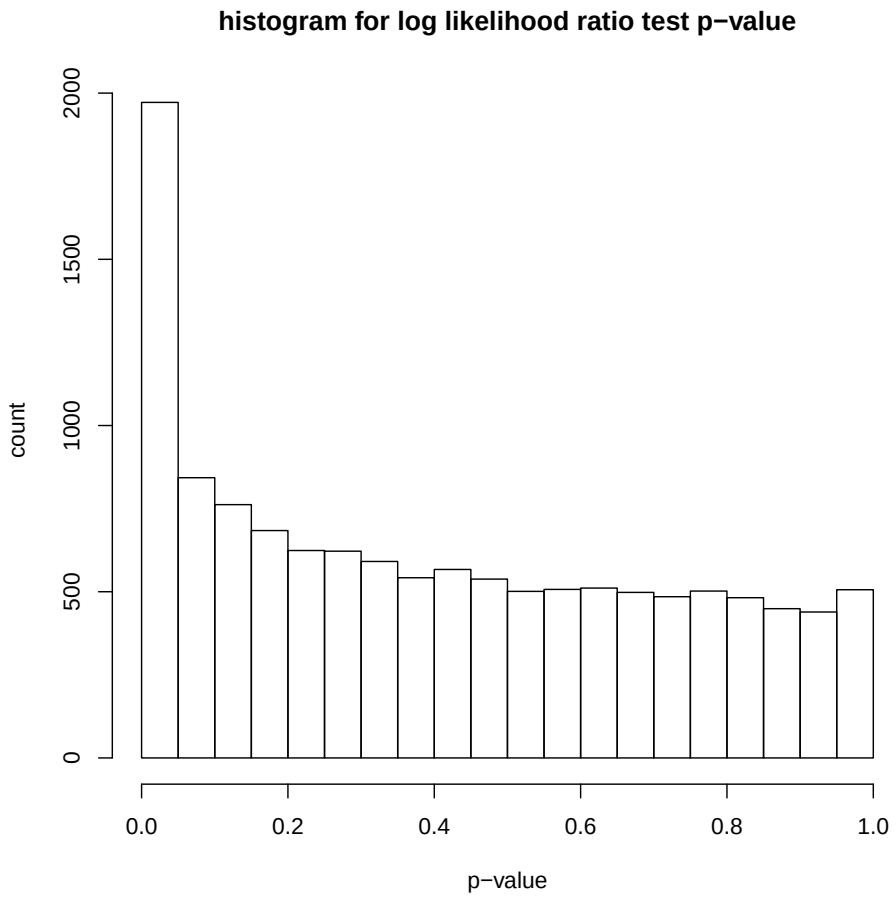

Figure 3: Histogram for log likelihood ratio test p-values

Method 2: Student t-test

A standard statistical test is student t-test. Here we will implement *two-sample unpaired equal-variance t test* to each gene to test the difference between the two conditions.

Define $\mu_A(i)$ and $\mu_B(i)$ the mean expression intensity for gene i in group A and B. Statistically, we have the following hypothesis formula,

$$\begin{aligned} H_0: &\quad \mu_A(i) = \mu_B(i) \\ H_A: &\quad \mu_A(i) \neq \mu_B(i) \end{aligned}$$

Your task

1. (1 points) What's the assumption for this t test?

   There are three assumptions for the two-sample unpaired equal variance t-test:

   (a) The gene expression of the patients are normally distributed around some unknown means for each population, μ_A and μ_B, and unknown common variance σ^2.

   (b) The gene expression of all sampled patients are independent of one another.

   (c) The patients we have sampled are random samples from each population.

2. (1 points) Based on the hypothesis setting, which kind of test do we need, two-sided or one-sided t-test? Why?

   We will have two-sided test, because the alternative hypothesis is $\mu_A \neq \mu_B$. That is, we want to find the differentially expressed genes between these two groups.

   If the alternative hypothesis is $\mu_A > \mu_B$ (gene expression in group A is higher than B) or $\mu_A < \mu_B$ (gene expression in group A is lower than B), we will use one-sided t-test.

3. (1 points)Write down the formula for the test statistics (T). Please well annotate your symbols in the formula.

$$\begin{aligned} S^2_{pooled} &= \frac{\text{pooled sum of squared deviations from the 2 samples}}{\text{pooled degrees of freedom from 2 samples}} \\ &= \frac{(n_A - 1)S^2_A + (n_B - 1)S^2_B}{(n_A - 1) + (n_B - 1)} \\ T &= \frac{(\overline{X}_A - \overline{X}_B) - (\mu_A - \mu_B)}{\sqrt{S^2_{pooled}/n_A + S^2_{pooled}/n_B}} \\ &= \frac{(\overline{X}_A - \overline{X}_B)}{\sqrt{S^2_{pooled}/n_A + S^2_{pooled}/n_B}} \end{aligned}$$

Where

n_A – –the number of patients in group A (say BCR/ABL group)
n_B – –the number of patients in group B (say ALL1/AF4 group)
S_A^2 – –sample variance of group A
S_B^2 – –sample variance of group B
$\overline{X}_A$ – –sample mean of group A
$\overline{X}_B$ – –sample mean of group B

4. (1 points) T follows student t-distribution. What's the degree of freedom?

$$df = (n_A - 1) + (n_B - 1) = n_A + n_B - 2$$

5. (4 points) Implement t-test for the quantile normalized data and calculate the p-value for each gene. Plot the histogram of the p-values. *[Hint: In MATLAB, you can use the function* mattest. *In R, you can use function* mt.teststat *from biocLite package* multtest *to get the test statistics and use function* pt *to calculate the p-value. Alternatively, in R, you can use function* t.test *to get test statistic and p-value together.]*

The t-test p-value histogram is shown in figure 4.

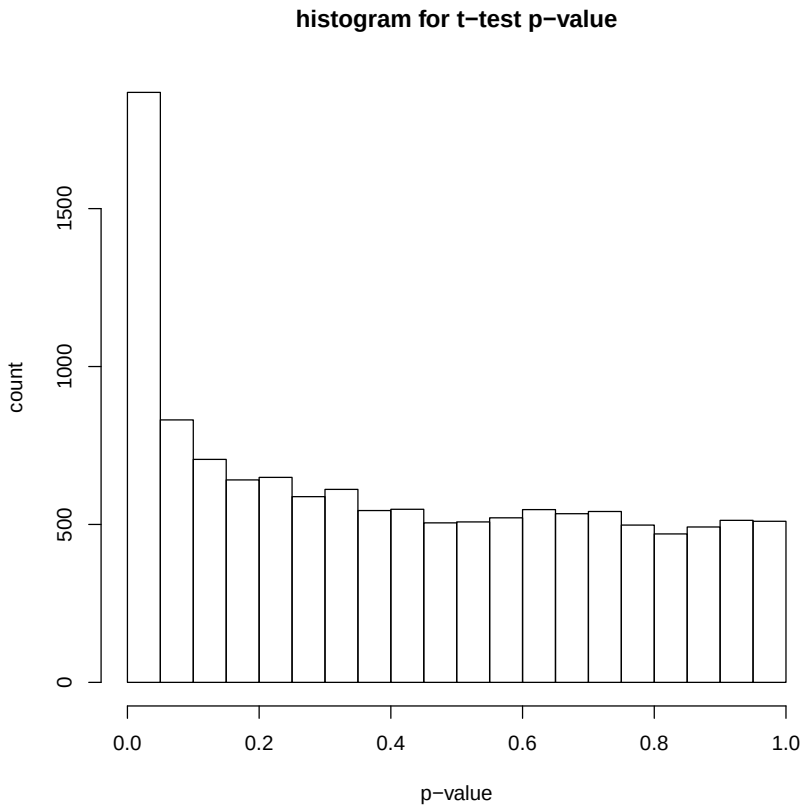

Figure 4: Histogram for the t-test p-values.

4.3 [8 points] Adjust p-value to detect DE genes

When performing DE analysis, two types of errors can occur: a false positive in which we declare that a gene is differentially expressed when it is not, and a false negative when the test fails to identify a truly differentially expressed gene. In multiple hypothesis testing, which simultaneously tests the null hypothesis of thousands of genes, each test has a specific false positive rate, or a false discovery rate (FDR). Instead of controlling the chance of any false positives (as Bonferroni methods do), FDR controls the expected proportion of false positives. An FDR threshold is then determined from a p-value distribution; hence, FDR is adaptive to the amount of signal in your data. Here we will implement two p-value correction methods for the t-test p-values you obtained above. The p-value cut-off is set to 0.05.

Method 1: Bonferroni correction
The motivation of Bonferroni correction is,

$$p(\text{specific } T_i \text{ passes}|H_0) < \frac{\alpha}{n}$$
$$p(\text{some } T_i \text{ passes}|H_0) < \alpha$$

where α is the p-value cut-off and n is the total number of p-values to be corrected.

Your task
(4 points) Implement Bonferroni correction for the p-values got from the t-test. Given the p-value cut-off $\alpha = 0.05$, how many DE genes do you find? List the top 10 DE genes (genes with smallest p-values) in the table, showing their probe names, original p-values and adjusted p-values. *[Hint: In MATLAB you can use function* mafdr. *In R, you can use function* p.adjust *and set the parameter* method='bonferroni'.*]*

(Different outputs are allowed.) There are 151 DE genes will adjusted p-values smaller than the cut-off. The top 10 DE genes are listed below:

gene	p-value	adjusted p-value
1914_at	$1.16E-29$	$1.47E-25$
37809_at	$2.28E-23$	$2.88E-19$
36873_at	$2.69E-23$	$3.40E-19$
40763_at	$7.93E-21$	$1.00E-16$
34210_at	$3.57E-19$	$4.51E-15$
41448_at	$1.33E-18$	$1.68E-14$
33358_at	$1.87E-16$	$2.36E-12$
37978_at	$1.09E-13$	$1.38E-09$
40480_s_at	$3.25E-13$	$4.10E-09$
1307_at	$4.27E-13$	$5.39E-09$

Method 2: Benjamini-Hochberg (BH) correction
The BH is defined as:
Let $p_1 \leq \cdots \leq p_n$ be ordered p-values. Define

$$k = i : p_n \leq \frac{iq}{n}$$

and reject $H_0^{(1)}, \ldots, H_0^{(k)}$. If no such i exists, reject null hypothesis.

Your task

(4 points) Implement BH correction for the p-values got from t-test. Given the cut-off 0.05, how many DE genes can you find from your adjusted p-values? *[Hint: In MATLAB, you can use function* mafdr, *setting parameters* METHOD *to be* BHFDR *and* BHFDRValue *to be* true*]. In R, you can use function* p.adjust, *setting* method='BH'*.]*

Given cut-off 0.05, totally 712 DE genes can be detected. Answers may vary a little bit.

4.4 [15 points] Significance analysis of microarrays, SAM

SAM is one of the commonly used algorithm for DE gene detection. For each gene, the test statistic is defined to be the relative difference of the two groups. For each gene i, relative difference $d(i)$ has the formula,

$$d(i) = \frac{\mu_A(i) - \mu_B(i)}{s(i) + s_0}$$

where s_0 is the exchangeability factor. $s(i)$ is the gene specific scatter,

$$s(i) = \sqrt{a\left\{\sum_{j\in A}\left[y_i^j - \mu_A(i)\right]^2 + \sum_{j\in B}\left[y_i^j - \mu_B(i)\right]^2\right\}}$$

where $a = \frac{1/n_A + 1/n_B}{n_A + n_B - 2}$. The relative difference can be sorted with decreasing order.

Next, we generate the null distribution by P-time permutations. In each permutation, we need to randomly permute the sample label (BCR/ABL and ALL1/AF4, in this case) and calculate a new test statistic for each row. The statistics will be ranked so that $d_p(i)$ represents the i^{th} largest relative difference for permutation p. The expected relative difference, $d_E(i)$, was calculated by the average of the P permutations, $d_E(i) = \frac{1}{P}\sum_{p=1}^{P} d_p(i)$.

You can refer the following paper for more details.

Virginia Goss Tusher, et al. Significance analysis of microarrays applied to the ionizing radiation response. PNAS, 2001 (98): 5116-5121.

Your task

Implement SAM for the quantile normalized data to calculate observed relative difference $d(i)$ and expected relative difference $d_E(i)$ (using 100 permutations). To simplify the code, use $s_0 = 0.031$. *[Hint: In R, there is a biocLite package called* samr*. You can code yourself or use this package directly.]*

1. (5 points) Scatter plot sorted $d(i)$ over $d_E(i)$. Draw a solid line $d(i) = d_E(i)$ on the plot. Also draw two dash lines $|d(i) - d_E(i)| = \Delta$. Plot two figures with $\Delta = 3$ and $\Delta = 2$ respectively. Compare the two figures and explain the difference. *[Hint: Δ is the threshold cut-off of the distance between the solid line and dash line. What are the meanings of the solid line and dash lines?]*

   Please see figure 5. Larger Δ allows larger difference between the test statistics ($d(i)$) and the expected one ($d_E(i)$).

2. (5 points) Scatter plot sorted $d(i)$ over $s(i)$. Explain the figure and what can you learn from it?

   Please see figure 6. The discussion is an open question.

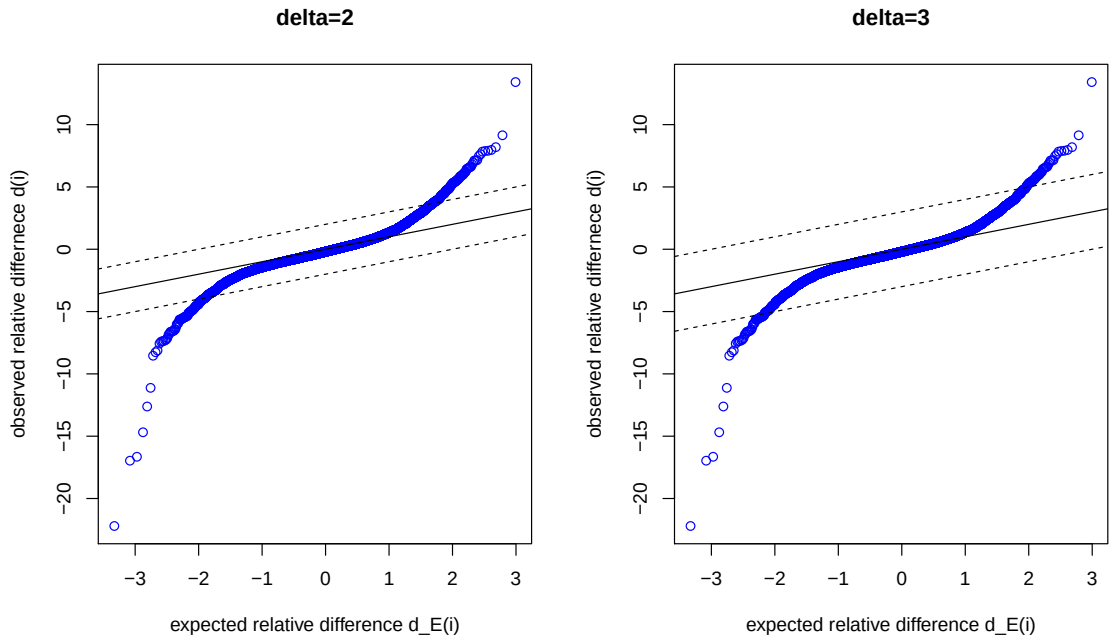

Figure 5: Scatter plot for $d(i)$ over $d_e(i)$.

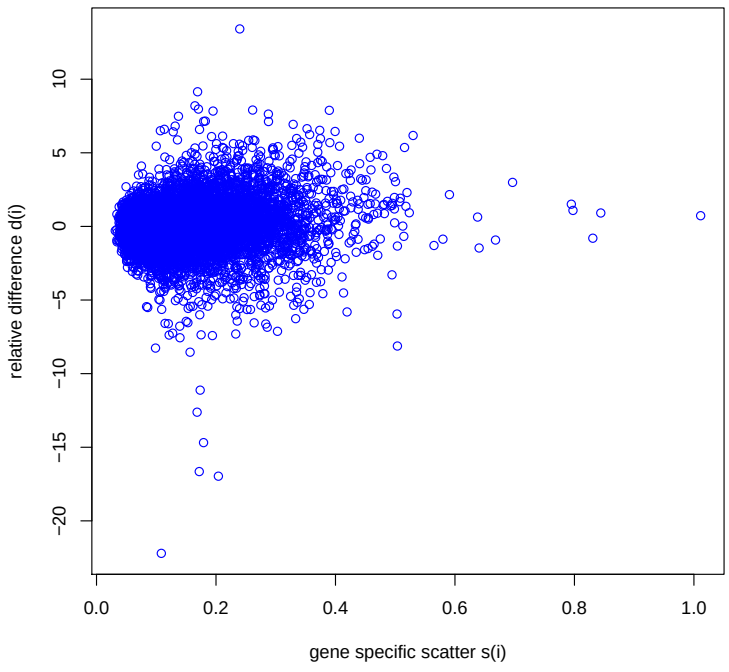

Figure 6: Scatter plot for $d(i)$ over $s(i)$.

3. (5 points) Scatter plot μ_A over μ_B. Specify the condition (BCR/ABL and ALL1/AF4) in x and y labels. Explain the figure and what can you learn from it?

   Please see figure 7. The discussion is an open question.

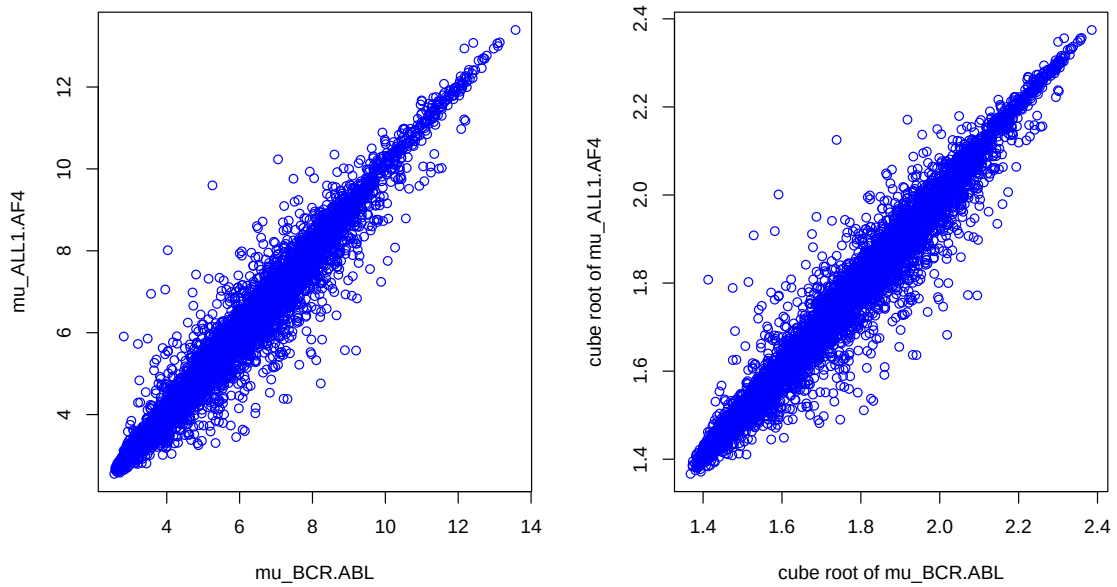

Figure 7: Scatter plot for μ_A over μ_B.