Chapter 16
Microarray Differential Expression Analysis

17

Microarray Differential Expression Analysis

Gene expression is the process by which mRNA, and eventually protein, is synthesized from the DNA template of each gene. The level of gene expression can be measured by a particular technology, called microarray technology (Schena et al., 1995), in which we can measure the expression of thousands of different RNA molecules at a given time point in the life of an organism, a tissue, or a cell. Comparisons of the levels of RNA molecules can be used to decipher the thousands of processes going on simultaneously in living organisms. Also, comparing healthy and diseased cells can yield vital information on the causes of the disease. The microarray technology has been successfully applied to several biological problems and, as arrays become more easily accessible to researchers, the popularity of these kinds of experiments will increase. The demand for good statistical analysis regimens and tools tailored for microarray data analysis will increase as the popularity of microarrays grows. The future will likely bring many new microarray applications, each with its own demands for specialized statistical analysis. Starting from this chapter, we will learn some of the basic statistical methods for microarray data analysis.

17.1 Data preparation

First, we need to understand the format of microarray data. The data are usually arranged in a matrix form, with rows representing genes and the columns representing tissue samples. The expressing level of the ith gene in the jth sample is denoted by y_{ij},$\forall i = 1, ..., N; j = 1, ..., M$, where N is the number of genes and M is the number of tissue samples, as shown in the sample data (Table 17.1). Each data point is assumed to have been properly transformed and normalized. Data preparation includes such data transformation and normalization.

Table 17.1. A sample dataset of microarray gene expression.

Gene	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	...	Sample M
1	6.73293	7.02988	6.85878	6.94274	5.00395	...	7.80914
2	4.40305	5.06069	4.95442	4.83628	6.37724	...	5.76519
3	6.58147	6.66364	6.88531	6.90465	8.32814	...	5.38998
4	4.40183	5.14982	5.21494	5.11018	6.61258	...	4.18205
5	7.03632	6.55607	7.06851	6.79872	5.2776	...	5.48147
6	5.36317	5.90971	5.46848	5.76832	5.86363	...	5.36598
7	5.3303	5.22467	4.77238	5.06765	7.97398	...	5.18403
8	6.08199	6.32077	6.12796	6.11744	6.78004	...	4.53367
9	5.83802	5.45788	5.55721	5.8354	4.69592	...	5.38174
10	7.23807	7.10562	6.92991	7.03077	5.30231	...	5.94411
11	6.11014	5.56375	5.85536	6.09334	5.92345	...	5.76738
...	...	...	...	...	...	...	...
N	8.03291	7.84023	7.65354	8.10648	5.43285	...	6.05514

17.1.1 Data transformation

Most statistical methods require specific statistical models and distributions of the errors. Linear model is the most commonly used model for microarray data analysis. Normal distribution is most frequenctly assumed for the errors. However, data points in the original form usually cannot be described by a linear model and the errors may not follow a normal distribution. Therefore, some type of data transformation is usually recommended before the data analysis. Let w_{ij} be the original expression level of the ith gene in the jth condition. The transformed value can be expressed as

$$y^*_{ij} = f(w_{ij}) \tag{17.1}$$

where $f(.)$ represents any monotonic function. All the statistical methods to be described in the book are actually performed on the transformed data y^*_{ij}, not the raw data w_{ij}. There are many different ways to transform the data. We will list a few most commonly used ones as examples.

Logarithmic transformation

By far the most common transformation applied to microarray readings is the logarithmic transformation

$$y^*_{ij} = \log(w_{ij}) \tag{17.2}$$

The base of the logarithm may be 2, 10 or the natural logarithmic constant e. The choice of base is largely a matter of convenient interpretation. The log transformation tends to provide values that are approximately normally distributed and for which conventional linear regression and ANOVA models are appropriate.

Square root transformation

The square root transformation

$$y_{ij}^{*} = \sqrt{w_{ij}} \tag{17.3}$$

is a variance-stabilizing transformation. In other words, if the variance is proportional to the mean, the square root transformation will correct this to some degree so that the variance of the transformed values is independent of the mean of the transformed values.

Box-Cox transformation family

The two transformations described above are members of the Box-Cox family of transformations (Box and Cox, 1964). This family is defined as

$$y_{ij}^{*} = \frac{w_{ij}^{\delta} - 1}{\delta} \tag{17.4}$$

where δ is the parameter of transformation chosen by the investigator. The square root transformation corresponds to $\delta = \frac{1}{2}$. The logarithmic transformation corresponds to the limit of this equation when $\delta \longrightarrow 0$. The case where $\delta = 1$ corresponds to taking no transformation at all, except for a change in the origin. The Box-Cox family provides a range of transformations that may be examined to see which value of δ yields transformed values with the desired statistical properties.

17.1.2 Data normalization

The laboratory preparation of each biological specimen on a microarray slide introduces an arbitrary scale or dilution factor that is common to expression readings for all genes. We usually correct the readings for the scale factor and other variations using a process called normalization. The purpose of normalization is to minimize extraneous variation in the measured gene expression levels of hybridized mRNA samples so that biological differences (differential expression) can be more easily distinguished.

Practical experience has shown that, in addition to array effects, other extraneous sources of variation may be present that cloud differential gene expression if not taken into account. These sources of variations are called systematic errors, e.g., the effects of dye colors. These systematic errors will mask the true biological differences and should be removed before the data are analyzed.

Normalization across genes

The normalization is done by subtracting the mean expression of all genes from the individual gene expression. The normalized data point is

$$y_{ij} = y^*_{ij} - \overline{y}^*_{.j} \tag{17.5}$$

where $\overline{y}^*_{.j} = \frac{1}{N}\sum_{i=1}^{N} y^*_{ij}$ is the average expression of all genes in sample j. All subsequent statistical analysis should be performed on the normalized values y_{ij}.

Normalization across tissue samples

The normalization is done by subtracting the mean expression of all tissue samples from the individual gene expression. The normalized data point is

$$y_{ij} = y^*_{ij} - \overline{y}^*_{i.} \tag{17.6}$$

where $\overline{y}^*_{i.} = \frac{1}{M}\sum_{j=1}^{M} y^*_{ij}$ is the average expression of all samples for gene i. All subsequent statistical analysis should be performed on the normalized values y_{ij}.

Normalization across both genes and samples

The normalized data point is expressed as the deviation of the unnormalized value from the mean of all genes in the particular sample and the mean of all samples for the particular gene,

$$y_{ij} = y^*_{ij} - \overline{y}^*_{i.} - \overline{y}^*_{.j} + \overline{y}^*_{..} \tag{17.7}$$

where $\overline{y}_{..} = \frac{1}{MN}\sum_{i=1}^{N}\sum_{j=1}^{M} y^*_{ij}$ is the overall mean of gene expression across all genes and samples.

Normalization via analysis of covariance

Analysis of covariance is an approach to removing the influence of systematic errors on the effects of treatments in an ANOVA. If we ignore any systematic errors (effects), we may write a linear model to describe the expression of the ith gene in the jth treatment,

$$y^*_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk} \tag{17.8}$$

where subscript k represents the kth replicate of gene i in treatment j. In this model, μ is the grand mean, α_i is the effect of the ith gene, β_j is the effect of the jth treatment, ϵ_{ijk} is the error term, and γ_{ij} is the interaction effect between the ith gene and the jth treatment. The interaction term, γ_{ij}, actually reflects the differential expression of gene i. A sufficiently large γ_{ij} indicates that the ith gene expresses differently from one treatment level to another. It is γ_{ij} that we are interested in. By formulating the above linear (ANOVA) model, we can separate γ_{ij} (the interesting part of the model) from the other effects (the parts that are not interesting to us). If we estimate the non-interesting effects and remove them from y^*_{ijk}, we have

$$y_{ijk} = y^*_{ijk} - \widehat{\mu} - \widehat{\alpha}_i - \widehat{\beta}_j = \gamma_{ij} + \epsilon_{ijk} \tag{17.9}$$

The non-interesting effects can be estimated first and them removed from the model. This is a generalized normalization process. It has normalized the gene expression across both the genes and the treatments. The method of estimation for the non-interesting effects can be the usual least square method from a simple two-way ANOVA or the mixed model approach (Wolfinger et al., 2001) by treating the gene and treatment effects as fixed and the interaction effects as random. One advantage of using the mixed model approach for normalization over the previous method is that the method can handle unbalanced data.

More importantly, we can incorporate systematic effects into the ANOVA linear model as covariates and perform covariance analysis to remove the extraneous systematic errors. For example, the dye color asymmetry has led to the use of microarray study designs in which arrays are produced in pairs with the colors in one array reversed relative to the colors in the second array in order to compensate for the color differences. These are called reversed-color designs. In this kind of design, the color effects, which is not of our interest, should be included in the model. Therefore, the modified model incorporating the color effects appears

$$y^*_{ijlk} = \mu + \xi_l + \alpha_i + \beta_j + (\xi\alpha)_{li} + (\xi\beta)_{lj} + \gamma_{ij} + \epsilon_{ijlk} \tag{17.10}$$

where $\xi_l, \forall l = 1, 2$ is the color effect, $(\xi\alpha)_{li}$ is the interaction of the lth color with the ith gen, and $(\xi\beta)_{lj}$ is the interaction of the lth color with the jth treatment. If we do not include these effects in the model, they will be absorbed by the error term. With the above model, they can be estimated and removed from the analysis, as shown below,

$$y_{ijlk} = y^*_{ijlk} - \widehat{\mu} - \widehat{\xi}_l - \widehat{\alpha}_i - \widehat{\beta}_j - \widehat{(\xi\alpha)}_{li} - \widehat{(\xi\beta)}_{lj} = \gamma_{ij} + \epsilon_{ijlk}, \tag{17.11}$$

leaving only γ_{ij} in the model. Therefore, analysis of covariance is a generalized normalization approach to removing all non-interesting effects.

Once the data are properly transformed and normalized, they are ready to be analyzed using any of the statistical methods described in subsequent sections.

17.2 F-test and t-test

Assume that we collect tissues from M_1 mice affected by some particular disease and tissues from M_2 mice that have the same diseases but treated with a newly developed drug. The tissue of each mouse is microarrayed and the expression of N genes are measured. The purpose of the experiment is to find which genes have different levels of expression between the untreated and treated groups of mice, i.e., to find genes responding to the drug treatment. The data matrix has N rows and $M = M_1 + M_2$ columns. However, each data point (gene expression level) is better denoted by variable y with three subscripts, y_{ijk}, where $i = 1, \ldots, N$ indexes genes, $j = 1, \ldots, P$ indexes the level of treatment ($P = 2$ in the case of two levels of treatment), $k = 1, \ldots, M/P$ indexes the replication within each treatment group. The number of replicates within each treatment group is assumed to be $M_j = M/P$ for all $j = 1, \ldots, P$. If the data are not balanced, i.e., $M_1 \neq M_2$, the replication index k should be subscripted with j so that $k_j = 1, \ldots, M_j$. The t-test statistic of differential expression for the ith gene is

$$t_i = \frac{|\overline{y}_{i1.} - \overline{y}_{i2.}|}{s_{\overline{y}_{i1.} - \overline{y}_{i2.}}}, \tag{17.12}$$

where

$$\overline{y}_{ij.} = \frac{1}{M_j} \sum_{k=1}^{M_j} y_{ijk}, \quad \forall j = 1, 2 \tag{17.13}$$

$$s_{\overline{y}_{i1.} - \overline{y}_{i2.}} = \sqrt{s_{i1}^2/M_1 + s_{i2}^2/M_2} \tag{17.14}$$

and

$$s_{ij}^2 = \frac{1}{M_j - 1} \sum_{k=1}^{M_j} (y_{ijk} - \overline{y}_{ij.})^2, \ \forall j = 1, 2. \tag{17.15}$$

One can rank t_i across all $i = 1, ..., N$, and select all genes with $t_i > t_{df,1-\alpha}$ as differentially expressed genes, where $t_{df,1-\alpha}$ is the critical value chosen by the investigator, $df = M_1 + M_2 - 2$ is the degrees of freedom and $0 < \alpha < 1$ is a probability that controls the Type I error rate of the experiment (discussed later).

The t-test only applies to situations where there are two levels of treatment. For multiple levels of treatment, an F-test must be used. Let P be the number of levels of treatment. For example, if there are four groups of mice with the first group being the untreated group and the remaining three groups being treated with three different doses of a particular drug, then $P = 4$. The F-test statistic for the ith gene is calculated from the ANOVA table (see Table 17.2), where

$$SS_T = \sum_{j=1}^{P} M_j (\overline{y}_{ij.} - \overline{y}_{i..})^2 \tag{17.16}$$

and

$$SS_E = \sum_{j=1}^{P} \sum_{k=1}^{M_j} (y_{ijk} - \overline{y}_{ij.})^2 \tag{17.17}$$

The F-test statistic for the ith gene is defined as

$$F_i = \frac{MS_T}{MS_E} \tag{17.18}$$

The critical value used to declare significance is $F_{df_T, df_E, 1-\alpha}$, which is the $1-\alpha$ quantile of the F distribution with degrees of freedom df_T and df_E. All genes with $F_i > F_{df_T, df_E, 1-\alpha}$ are called significant. Again, the value of α is chosen by the investigator (discussed later).

Table 17.2. Analysis of variance (ANOVA) table for differential gene expression analysis

Variation	df	SS	MS	$F-$test statistic
Treatment	$df_T = P - 1$	SS_T	$MS_T = \frac{SS_T}{df_T}$	$\frac{MS_T}{MS_E}$
Error	$df_E = \sum_{j=1}^{P}(M_j - 1)$	SS_E	$MS_E = \frac{SS_E}{df_E}$	

17.3 Type I error and false discovery rate

In the previous section, the critical value for a test statistic used to select the list of significant genes is denoted by $t_{df, 1-\alpha}$ for the t-test or $F_{df_T, df_E, 1-\alpha}$ for the F-test. The α value is called the Type I error. A small α means a large critical value and thus generates a short list of significant genes while a large α will produce a long list of significant genes. Therefore, Type I error determines the list of significant genes. Let H_0 denote the null hypothesis that a gene is not differentially expressed and H_1 denote the alternative hypothesis that the gene is differentially expressed. When we perform a statistical test on a particular gene, we may make errors if the sample size is small. There are two types of errors we can make. If a gene is not differentially expressed but our test statistic is greater than the chosen critical value, we will make the Type I error whose probability is denoted by α as mentioned before. On the other hand, if a gene is differentially expressed but our test statistic is less than the chosen critical value, we will make the Type II error whose probability is denoted by β. In other words, we will make the Type I error if H_0 is true but H_1 is accepted, and make the Type II error if H_1 is true but H_0 is accepted. The Type I and Type II errors are also called false positive and false negative errors, respectively. These two errors are negatively related, i.e., a high Type I error leads to a low Type II error.

The probability that H_1 is accepted while H_1 is indeed true is called the statistical power. Therefore, the statistical power is simply $\omega = 1 - \beta$. A gene can be differentially expressed or not differentially expressed. So, if we use an indicator variable h to denote the true status of the gene, we have $h = H_0$ or $h = H_1$. After the statistical test, the gene will have one of two outcomes, significant or not. Let $\hat{h}$ be the outcome of the statistical test of the gene, then $\hat{h} = H_0$ if H_0 is accepted and $\hat{h} = H_1$ if H_1 is accepted. We can now define the Type I error as

$$\alpha = \Pr(\hat{h} = H_1 | h = H_0) \tag{17.19}$$

and the Type II error as

$$\beta = \Pr(\hat{h} = H_0 | h = H_1). \tag{17.20}$$

The statistical power is

$$\omega = \Pr(\hat{h} = H_1 | h = H_1). \tag{17.21}$$

So far, we only discussed the Type I error rate for a single test. In microarray data analysis, we have to test the differential expression for every gene. Therefore, a microarray experiment with N genes involves N hypothesis tests. The Type I error for a single test needs to be adjusted to control the Type I error of the entire experiment. The Bonferroni correction introduced in section 8.5 of Chapter 8 applies here. If the experiment-wise Type I error is γ, the nominal Type I error rate after correcting for multiple tests should be

$$\alpha = \frac{\gamma}{N} \tag{17.22}$$

Assume that a microarray experiment involves $N = 1000$ genes collected from $M_1 = 10$ cases and $M_2 = 10$ controls. A t-test statistics has been calculated for each of the N genes. We want to find the critical value for the t-test to compare so the the experiment wise Type I error is controlled at $\gamma = 0.05$. First, we need to find out the nominal Type I error rate using the Bonferroni correction,

$$\alpha = \frac{\gamma}{N} = \frac{0.05}{1000} = 0.00005$$

If the p-values are already given in the differential expression analysis, one may simply compare the gene specific p-values to 0.00005, declaring significance if the p-value is less than 0.00005. If the p-values are not given, we need to find the critical value used to declare significance using

$$t_{df,1-\alpha/2} = t_{18,0.999975} = 5.2879056 \tag{17.23}$$

Any genes with t-test statistics greater than 5.2879 will be declared as significant. Bonferroni correction usually provides a very conservative result of tests, i.e., reports less genes than the actual number of significant genes. In practice, Bonferroni is rarely used because of the conservativeness; instead, people often use permutation test to draw an empirical critical value. This will be discussed in the next section.

Another related performance measure in multiple testing is the so called false discovery rate (FDR) developed by Benjamini and Hochberg (1995). The FDR measure looks at error controls from a different perspective. Instead of conditioning on the true but unknown state of whether a gene is differentially expressed or not, the FDR is defined as the probability that the test statistic indicates that the gene is differentially expressed but in fact it is not. This probability, denoted by δ, is

$$\delta = \Pr(h = H_0 | \hat{h} = H_1) \tag{17.24}$$

We can control the value of FDR (δ) and use the fixed FDR value to select the list of significant genes.

17.4 Selection of differentially expressed genes

Data transformation is to ensure that the residual errors of gene expressions follow a normal distribution. A normal distribution will make the t-test or F-test statistic follow the expected t- or F-distribution under the null model. For most microarray data, transformation can only improve the normality and rarely make the residual errors perfectly normal. Therefore, the critical value for a test statistic drawn from the expected t- or F-distribution is problematic. In addition, the multiple test adjustment using the Bonferroni correction is far too conservative when N is very large. Therefore, the optimal way of finding the critical value perhaps relies on some empirical methods that are data dependent. In other words, different data sets should have different critical values, reflecting the particular natures of the data.

17.4.1 Permutation test

Permutation test is a way to generate the distribution of the test statistic under the null model. In differential expression analysis, the null model is that no genes are differentially expressed. Consider M tissue samples with M_1 samples being the control and M_2 being the treatment for $M_1 + M_2 = M$. Each sample is labeled as 0 for the control and 1 for the treatment. If the tissue samples and the labels are randomly shuffled, the association between the gene expressions and the labels will be destroyed. The distribution of the test statistic will mimic the distribution under the null model.

The total number of ways of shuffling is

$$T = \frac{M!}{M_1! M_2!}, \tag{17.25}$$

equivalent to the number of ways of randomly sampling M_1 or M_2 items from a total of M items. When M is large, T may be extremely large, making the permutation analysis very difficult. In practice, one may only use a proportion of the reshuffled samples to draw the null distribution of the test statistic.

This type of random shuffling may not generate the null distribution accurately because there is a chance that all M_1 tissue samples in a reshuffled dataset are actually from the control and M_2 samples from the treatment. If some genes are indeed differentially expressed, then the test statistics of these genes are not drawn from the null distribution. Tusher et al. (2001) proposed a balanced shuffling approach that can avoid this problem. In the balanced random shuffling, each group in the reshuffled dataset contains the samples from the original groups in proportion. Take the case-control experiment for example, in the reshuffled dataset, the M_1 controlled group should contain $\frac{M_1}{M_1+M_2}M_1$ samples from the original control group and $\frac{M_2}{M_1+M_2}M_1$ samples from the original treatment group. Similarly, the M_2 treatment group in the reshuffled dataset should contain $\frac{M_1}{M_1+M_2}M_2$ samples from the original control groups and $\frac{M_2}{M_1+M_2}M_2$ from the original treatment group. Under this restriction, we guarantee that the test statistics for all genes are sampled from the null distribution. The balanced shuffling is not easy to conduct if any one of $\frac{M_1}{M_1+M_2}M_1$, $\frac{M_2}{M_1+M_2}M_1$,$\frac{M_1}{M_1+M_2}M_2$ and $\frac{M_2}{M_1+M_2}M_2$ is not an integer. It is convenient to conduct the balanced shuffling if M_1 is an even number and $M_1 = M_2$. In this case, the total number of reshuffled datasets will be

$$T = \frac{M_1!}{(\frac{1}{2}M_1)!(\frac{1}{2}M_1)!}\frac{M_2!}{(\frac{1}{2}M_2)!(\frac{1}{2}M_2)!}. \tag{17.26}$$

For example, assume that $M = 8$ and $M_1 = M_2 = 4$, the total number of reshuffled datasets without restriction is $\frac{8!}{4!4!} = 70$ while the number of reshuffled datasets with the balance restriction is $\frac{4!}{2!2!}\frac{4!}{2!2!} = 36$.

For the kth reshuffled dataset, for $k = 1, \ldots, T$, the F-test statistics for the N genes are ranked in descending order so that

$$F^k_{(1)} > F^k_{(2)} > \cdots > F^k_{(N)}, \tag{17.27}$$

where $F^k_{(j)}$ is the jth largest F-test statistic of the kth reshuffled dataset. Let $\bar{F}_{(i)} = \frac{1}{T}\sum_{k=1}^{T} F^k_{(j)}$ be the average of the ith largest F-test statistic across the reshuffled datasets so that

$$\bar{F}_{(1)} > \bar{F}_{(2)} > \cdots > \bar{F}_{(N)}. \tag{17.28}$$

The empirical critical value drawn from this permutation test is $\bar{F}_{(C)}$, where C is chosen such that $\frac{C}{N} = \gamma$ and γ is a preset experiment-wise Type I error rate. Let

$$F_{(1)} > F(2) > \cdots > F_{(N)} \tag{17.29}$$

be the list of ranked F-test statistics calculated from the original dataset. All genes with a ranked $F_{(i)} > \bar{F}_{(C)}$ are selected as significant genes.

Table 17.3 gives an example of 20 hypothetical genes and their test statistics (ranked). The table also provides the ranked average test statistics obtained from a permutation test.

Table 17.3. The rank of 20 hypothetical genes.

Gene	Ranking(i)	$F_{(i)}$	$\bar{F}_{(i)}$
6	1	43.6478	5.9448
4	2	17.3289	5.1476
14	3	9.8718	4.7502
9	4	6.7659	4.4627
12	5	5.9085	4.2743
1	6	5.4049	4.1099
13	7	5.1551	3.9715
15	8	4.8471	3.8872
11	9	4.4245	3.7793
7	10	4.0889	3.6800
10	11	4.0834	3.5990
3	12	4.0557	3.5371
8	13	3.9786	3.4796
17	14	3.9667	3.4209
18	15	3.9480	3.3653
5	16	3.9219	3.3272
16	17	3.9102	3.2711
19	18	3.9101	3.2242
20	19	3.8748	3.1811
2	20	3.8736	3.1357

Assume that we want to control the experimental Type I error rate at $\gamma = 2/20 = 0.10$. Therefore, $C = 2$ and $\bar{F}_{(2)} = 5.1476$ is the critical value. Since $F_{(i)} > \bar{F}_{(2)}$ for $i = 1, \ldots, 7$, seven genes are selected as differentially expressed. The list of the significant genes is $\{6, 4, 14, 9, 12, 1, 13\}$. The permutation test also allows us to estimate the empirical FDR. Since seven genes are detected, among which two genes are expected to be false positive, the empirical FDR is $\delta = 2/7 = 0.2857$.

17.4.2 Selecting genes by controlling FDR

The permutation test given in the above example shows that when we set $\gamma = 0.10$, the empirical FDR is $\delta = 0.2857$. This suggests a way to select significant genes by controlling the FDR rather than the Type I error rate. Let $\gamma_{(i)} = i/N$, for $i = 1, \ldots, N$, denote the Type I error rate when $\bar{F}_{(i)}$ is used as the critical value. The list of significant genes includes all genes with $F_{(i')} > \bar{F}_{(i)}$, for $i' = 1, \ldots, S_{(i)}$, where $S_{(i)}$ is the largest i' such that $F_{(i')} > \bar{F}_{(i)}$. The number of significant genes under $\gamma_{(i)}$ is then $S_{(i)}$. Therefore, the empirical FDR under $\gamma_{(i)}$ is $\delta_{(i)} = i/S_{(i)}$. The $\gamma_{(i)}$ and $\delta_{(i)}$ values for the 20 hypothetical genes are listed in Table 17.4.

Table 17.4. Empirical FDR of the 20 hypothetical genes.

Gene	Ranking (i)	$F_{(i)}$	$\bar{F}_{(i)}$	$\gamma_{(i)}$	$S_{(i)}$	$\delta_{(i)}$
6	1	43.6478	5.9448	0.05	4	0.2500
4	2	17.3289	5.1476	0.10	7	0.2857
14	3	9.8718	4.7502	0.15	8	0.3750
9	4	6.7659	4.4627	0.20	8	0.5000
12	5	5.9085	4.2743	0.25	9	0.5556
1	6	5.4049	4.1099	0.30	9	0.6667
13	7	5.1551	3.9715	0.35	13	0.5833
15	8	4.8471	3.8872	0.40	18	0.4444
11	9	4.4245	3.7793	0.45	20	0.4500
7	10	4.0889	3.6800	0.50	20	0.5000
10	11	4.0834	3.5990	0.55	20	0.5500
3	12	4.0557	3.5371	0.60	20	0.6000
8	13	3.9786	3.4796	0.65	20	0.6500
17	14	3.9667	3.4209	0.70	20	0.7000
18	15	3.9480	3.3653	0.75	20	0.7500
5	16	3.9219	3.3272	0.80	20	0.8000
16	17	3.9102	3.2711	0.85	20	0.8500
19	18	3.9101	3.2242	0.90	20	0.9000
20	19	3.8748	3.1811	0.95	20	0.9500
2	20	3.8736	3.1357	1.00	20	1.0000

If we want to set the FDR at $\delta = 0.375$, we will select 8 significant genes. The Type I error corresponding to this FDR is $\gamma = 0.15$ with a critical value of $\bar{F}_3 = 4.7502$. There are 8 genes with test statistics larger than 4.7502. The list of the 8 significant genes is $\{6, 4, 14, 9, 12, 1, 13, 15\}$.

The empirical method for selecting significant genes by controlling the FDR may not work for some datasets. The problem is that the relationship between $\gamma_{(i)}$ and $\delta_{(i)}$ may not be monotonic, leading to multiple values of γ corresponding to the same δ value. For example, both $\gamma_{(4)}$ and $\gamma_{(10)}$ correspond to $\delta = 0.50$. Therefore, the empirical method by controlling FDR is not recommended.

Although it is hard to select genes using the exact FDR control, Benjamini and Hochberg (1995) suggests to control FDR, not at δ but at $< \delta$. This approach does not require permutation test. It only requires calculation of the p-values corresponding to the test statistics. Genes are then ranked in ascending order based on their p-values. Significant genes under FDR $< \delta$ are selected in the following steps.

1. Rank genes based on the p-values in ascending order,

$$p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(N)}$$

Let $g_{(i)}$ be the gene corresponding to p-value $p_{(i)}$.

2. Let $i_{\max}$ be the largest i for which

$$p_{(i)} \leq \frac{i}{N}\delta$$

3. Declare significance for gene $g_{(i)}$ $\forall i = 1, ..., i_{\max}$. The nominal Type I error rate is $\alpha = p_{(i_{\max})}$.

Table 17.5 shows an example with $N = 20$ genes under FDR $< \delta = 0.05$. It shows that $i_{\max} = 2$. Therefore, two genes ($g_{(1)} = 1$ and $g_{(2)} = 7$) are selected as significant with a nominal Type I error of $\alpha = 0.00172817$.

Table 17.5. False discovery rate of the 20 hypothetical genes ($\delta = 0.05$).

Gene	Ranking	$p_{(i)}$	$\frac{i}{N}$	$\frac{i}{N}\delta$
1	1	0.00036084	0.05	0.0025
7	2	0.00172817	0.10	0.0050
9	3	0.01443942	0.15	0.0075
15	4	0.02210477	0.20	0.0100
4	5	0.02354999	0.25	0.0125
3	6	0.03488643	0.30	0.0150
6	7	0.03488643	0.35	0.0175
5	8	0.03906270	0.40	0.0200
8	9	0.03927667	0.45	0.0225
10	10	0.04195647	0.50	0.0250
2	11	0.04934124	0.55	0.0275
13	12	0.10373490	0.60	0.0300
14	13	0.17910006	0.65	0.0325
20	14	0.28077512	0.70	0.0350
16	15	0.34277900	0.75	0.0375
19	16	0.35010501	0.80	0.0400
12	17	0.36536933	0.85	0.0425
11	18	0.40129684	0.90	0.0450
18	19	0.47260844	0.95	0.0475
17	20	0.48037560	1.00	0.0500

Note that the FDR control is not a new statistical method for parameter estimation; rather, it is simply a different way provided by statisticians for biologists to decide the cutoff point for the “significant” genes.

17.4.3 Problems of the previous methods

The simple t-test or F-test described above is not optimal for differential expression analysis when the sample size is small. The current microarray technology is still not sufficiently effective to allow investigators to microarray a large number of tissue samples in a single microarray experiment. Therefore, microarray data are in general conducted with a very small sample size, although many genes can be measured from each tissue sample. When the sample size is small, the t-test or F-test statistics are not stable. Although both the numerator (the estimated difference between the control and the treatment) and the denominator (the estimated standard error of the difference) are subject to large estimation errors, the error in the denominator is more sensitive to the small sample size. One solution is to modify the estimation of the denominator to make it less sensitive to the small sample size. This can be done by sharing information between different genes when estimating the denominator of the t-test statistics. The genes are measured simultaneously within the same tissues. Therefore, the test statistics of the genes are correlated. This information has not been incorporated into the estimation of the standard error of the expression difference between the control and the treatment in the simple t- or F-test statistic.

17.4.4 Regularized t-test

Baldi and Long (2001) developed a Bayesian method to test differentially expressed genes. Instead of using the observed standard error of the control-treatment difference as the denominator for the t-test, they used a Bayesian method to estimate this standard error. The Bayesian estimate for the error variance is a weighted average of the observed variance and a prior variance set by the investigator. This modified t-test is called the regularized t-test. Similar idea has been proposed by Efron et al. (2001) and Tusher et al. (2001), who added a constant to the observed standard error as a new denominator to modify the calculated t-test statistic. The modified t-test in Efron et al. (2001) andTusher et al. (2001) is

$$t_i = \frac{|\overline{y}_{i1.} - \overline{y}_{i2.}|}{s_{\overline{y}_{i1.} - \overline{y}_{i2.}} + s_0}, \tag{17.30}$$

where s_0 is a constant added to the denominators of the t-tests for all genes. This constant is analogous to the prior standard deviation of Baldi and Long (2001). The constant s_0 is often chosen in such a way that it depends on the entire data of the microarray experiment. Because the value of s_0 depends on the entire data, information sharing occurs between genes. Efron et al. (2001) ranked the observed standard deviation across genes and select the 95 percentile of this empirical distribution as the value of s_0.

The regularized t-test of Baldi and Long (2001) has been implemented in a software package called Cyber-T (www.genomics.uci.edu/software.html). The significance analysis of microarrays of Tusher et al. (2001) has been implemented in a software called SAM (http://www-stat-class.standford.edu/SAM/SAMSevervlet)

17.5 General linear model

The t-test or the regularized t-test method only applies to two levels of a treatment, i.e., the case-control study. When the treatment has more levels, only two levels are considered at a time. In this section, we introduce a general linear model that can handle differential expression analysis with an arbitrary number of treatment levels. In addition, all genes are analyzed simultaneously under a single general linear model so that information of data is shared among genes. This will automatically generate a more accurate estimation of the error variance for each gene. The method was developed by Smyth (2004) who also provided a program named Limma (Linear Models for Microarray Data) along with the method.

Under the general linear model framework, all equations are written in matrix forms. It is more convenient to denote the microarray dataset by a matrix with rows and columns flipped (from the original dataset) so that each row represents a sample and each column represents a gene. This transposed dataset is now an $M \times N$ matrix, where M is still the number of samples and N still represents the number of genes. Note that the data matrix is assumed to have been properly transformed and normalized prior to the data analysis. Let $y_j = \{y_{1j}, \ldots, y_{Mj}\}^T$ denote the jth column of the data matrix, i.e., it stores the expressions of the jth gene for all the M samples. We now use the following linear model to describe y_j,

$$y_j = \beta + Z\gamma_j + \epsilon_j, \forall j = 1, \ldots, N \tag{17.31}$$

where β is an $M \times 1$ vector for the mean expressions across all genes, γ_j is a $P \times 1$ vector ($P < M$) of latent variables for P different groups of tissue samples, i.e., the effects for P groups, and ϵ_j is an $M \times 1$ vector for the residual errors. Finally, Z is an $M \times P$ design matrix. For example, if there are $M = 6$ tissue samples and $P = 3$ types of tissues and assume that the first two samples are from the first type of tissue, the second two samples from the second type of tissue and the last two samples from the third type of tissue, then a class variable is denoted by a vector $G = \{1, 1, 2, 2, 3, 3\}$ and the design matrix $Z = \text{design}(G)$ has the form of

$$Z = \begin{bmatrix} 1\,0\,0 \\ 1\,0\,0 \\ 0\,1\,0 \\ 0\,1\,0 \\ 0\,0\,1 \\ 0\,0\,1 \end{bmatrix} \tag{17.32}$$

Detailed view of the linear model is

$$\begin{bmatrix} y_{1j} \\ y_{2j} \\ y_{3j} \\ y_{4j} \\ y_{5j} \\ y_{6j} \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{bmatrix} + \begin{bmatrix} 1\,0\,0 \\ 1\,0\,0 \\ 0\,1\,0 \\ 0\,1\,0 \\ 0\,0\,1 \\ 0\,0\,1 \end{bmatrix} \begin{bmatrix} \gamma_{1j} \\ \gamma_{2j} \\ \gamma_{3j} \end{bmatrix} + \begin{bmatrix} \epsilon_{1j} \\ \epsilon_{2j} \\ \epsilon_{3j} \\ \epsilon_{4j} \\ \epsilon_{5j} \\ \epsilon_{6j} \end{bmatrix} \tag{17.33}$$

Note that β_i for $i = 1, ..., M$ is the mean of expression levels for all genes measured in the ith tissue sample. Assume that ϵ_j has a multivariate $N(0, I\sigma_j^2)$ distribution. We now present two methods for the general linear model (GLM) analysis.

17.5.1 Fixed model approach

The fixed model approach of the GLM is simply an alternative way to perform the F-test (described early), except that the GLM provides more flexible way for the F-test. Under the fixed model framework, information sharing only occurs with the mean expressions (vector β). In fact, vector β may be removed before the analysis in the normalization step. In that case the model is simply represented by $y_j = Z\gamma_j + \epsilon_j$, where vector y_j has been normalized. The GLM analysis is so general that we can estimate β simultaneously along with γ_j and other parameters. Here we emphasize simultaneous analysis, and thus all genes are included in the same model. The log likelihood function for gene j is

$$L_j(\beta, \gamma_j, \sigma_j^2) = -\frac{1}{2}\ln(\sigma_j^2) - \frac{1}{2\sigma_j^2}(y_j - \beta - Z\gamma_j)^T(y_j - \beta - Z\gamma_j) \tag{17.34}$$

Assume that all genes are independent (this assumption may often be violated), the overall log likelihood function is

$$L(\beta, \gamma, \psi) = -\frac{1}{2}\sum_{j=1}^{N}\ln(\sigma_j^2) - \sum_{j=1}^{N}\frac{1}{2\sigma_j^2}(y_j - \beta - Z\gamma_j)^T(y_j - \beta - Z\gamma_j) \tag{17.35}$$

where $\gamma = \{\gamma_j\}_{j=1}^N$ and $\psi = \{\sigma_j^2\}_{j=1}^N$. The maximum likelihood estimates of β and γ can be obtained explicitly. However, using the following iterative approach simplifies the estimation,

$$\begin{aligned} \beta &= \frac{1}{N}\sum_{j=1}^{N}(y_j - Z\gamma_j) \\ \gamma_j &= (Z^TZ)^{-1}Z^T(y_j - \beta), \forall j = 1, \cdots, N \end{aligned} \tag{17.36}$$

The iteration process converges quickly to produce the MLE of β and γ, denoted by $\hat{\beta}$ and $\hat{\gamma}$. The MLE of σ_j^2 is

$$\hat{\sigma}_j^2 = \frac{1}{M}(y_j - \hat{\beta} - Z\hat{\gamma}_j)^T(y_j - \hat{\beta} - Z\hat{\gamma}_j), \forall j = 1, \cdots, N \tag{17.37}$$

The variance of the MLE of γ_j can be approximated using

$$\text{var}(\hat{\gamma}_j) = (Z^T Z)^{-1}\hat{\sigma}_j^2, \forall j = 1, \cdots, N \tag{17.38}$$

The next step is to perform statistical tests for differentially expressed genes. For three treatment groups, there are two orthogonal linear contrasts, which can be expressed as

$$\alpha_j = L^T\gamma_j = \begin{bmatrix} 1 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \gamma_{1j} \\ \gamma_{2j} \\ \gamma_{3j} \end{bmatrix} = \begin{bmatrix} \gamma_{1j} - \frac{1}{2}\gamma_{2j} - \frac{1}{2}\gamma_{3j} \\ \gamma_{2j} - \gamma_{3j} \end{bmatrix} \tag{17.39}$$

where

$$L = \begin{bmatrix} 1 & 0 \\ -\frac{1}{2} & 1 \\ -\frac{1}{2} & -1 \end{bmatrix} \tag{17.40}$$

The null hypothesis for $H_0 : \alpha_j = 0$ can be tested using

$$F_j = \frac{1}{r}\hat{\alpha}_j^T \text{var}^{-1}(\hat{\alpha}_j)\hat{\alpha}_j \tag{17.41}$$

where $r = r(L) = 2$ is the rank of matrix L and

$$\text{var}(\hat{\alpha}_j) = \text{var}(L^T\hat{\gamma}_j) = L^T\text{var}(\hat{\gamma}_j)L = L^T(Z^TZ)^{-1}L\hat{\sigma}_j^2 \tag{17.42}$$

Therefore,

$$F_j = \frac{1}{r}\hat{\alpha}_j^T \text{var}^{-1}(\hat{\alpha}_j)\hat{\alpha}_j = \frac{1}{r\hat{\sigma}_j^2}\hat{\gamma}_j^T L \left(L^T(Z^TZ)^{-1}L\right)^{-1} L^T\hat{\gamma}_j \tag{17.43}$$

There is not much advantage of this GLM generated F-test over the F-test described earlier, except that one can choose different linear contrast matrix L to perform different biologically meaningful tests.

As mentioned earlier that information sharing is not obvious for the fixed model approach of F-test. The problem for the sensitivity of $\hat{\sigma}_j^2$ to small sample size remains unsolved. A slight modification can resolve this problem. Let us now make the assumption of

$$\epsilon_j \sim N(0, I\sigma^2), \forall j = 1, \cdots, N \tag{17.44}$$

which states that all genes share a common residual error variance. The MLE estimate of σ^2 becomes

$$\hat{\sigma}^2 = \frac{1}{MN}\sum_{j=1}^{N}(y_j - \hat{\beta} - Z\hat{\gamma}_j)^T(y_j - \hat{\beta} - Z\hat{\gamma}_j) \tag{17.45}$$

After this slight modification, the F-test statistic is

$$F_j = \frac{1}{r\hat{\sigma}^2}\hat{\gamma}_j^T L \left(L^T(Z^TZ)^{-1}L\right)^{-1} L^T\hat{\gamma}_j \tag{17.46}$$

This is a much more robust test statistics. However, it is questionable for the validity of the assumption of common residual error variance across all genes.

17.5.2 Random model approach

We now modify the model to allow all other non-interesting effects to be included in the model, which is

$$y_j = X\beta + Z\gamma_j + \epsilon_j \tag{17.47}$$

where X is an $M \times q$ design matrix and β is a $q \times 1$ vector. Under the fixed model approach, we simply assume that $X = I$ (an identity matrix). Now this assumption has been relaxed, although the fixed model approach can handle $X \neq I$ equally well. We now still assume $\epsilon_j \sim N(0, I\sigma_j^2), \forall j = 1, \cdots, N$. In addition, we make another assumption, $\gamma_j \sim N(0, \Pi), \forall j = 1, \cdots, N$, where Π is a $p \times p$ positive definite matrix, an unknown matrix subject to estimation. With this assumption, the model becomes a random model or mixed model considering β being fixed effects. The variance-covariance matrix Π is shared by all genes, which is what we call the information sharing among genes. Under the mixed model framework, we estimate parameters $\theta = \{\beta, \Pi, \psi\}$ using the ML method while $\gamma = \{\gamma_j\}$ are predicted rather than estimated because they are no longer called parameters. We may use a two-step approach to predict γ . The first step is to estimate the parameters and the second step is to predict γ given the estimated parameters and the data. Vector y_j now follows a multivariate normal distribution

$$y_j \sim N(X\beta, Z\Pi Z^T + I\sigma_j^2) \tag{17.48}$$

The log likelihood function for gene j is

$$\begin{aligned} L_j(\theta) = &-\frac{1}{2}\ln|Z\Pi Z^T + I\sigma_j^2| \\ &-\frac{1}{2}(y_j - X\beta)^T(Z\Pi Z^T + I\sigma_j^2)^{-1}(y_j - X\beta) \end{aligned} \tag{17.49}$$

which leads to an overall log likelihood function of

$$L(\theta) = \sum_{j=1}^{N} L_j(\theta) \tag{17.50}$$

The two-step approach can be combined into a single step, but with an iterative mechanism for the solution. The EM algorithm is such an algorithm with explicit expression of the solution in each iterative cycle. In the E-step, we predict γ_j conditional on the parameters and the data,

$$\hat{\gamma}_j = E(\gamma_j) = \left(\mathit{\Pi}^{-1}\sigma_j^2 + Z^T Z\right)^{-1} Z^T (y_j - X\beta) \tag{17.51}$$

and

$$\mathit{\Sigma}_j = \mathrm{var}(\hat{\gamma}_j) = \left(\mathit{\Pi}^{-1}\sigma_j^2 + Z^T Z\right)^{-1} \sigma_j^2 \tag{17.52}$$

These allow us to calculate

$$E(\gamma_j \gamma_j^T) = \mathrm{var}(\gamma_j) + E(\gamma_j)E(\gamma_j^T) = \mathit{\Sigma}_j + \hat{\gamma}_j\hat{\gamma}_j^T \tag{17.53}$$

which is required in the M-step. In the M-step, we calculate the parameter values using the quantities obtained in the E-step,

$$\begin{aligned}
\beta &= \frac{1}{N}(X^T X)^{-1} X^T \sum_{j=1}^{N} (y_j - Z\hat{\gamma}_j) \\
\mathit{\Pi} &= \frac{1}{N}\sum_{j=1}^{N} E(\gamma_j\gamma_j^T) = \frac{1}{N}\sum_{j=1}^{N} (\mathit{\Sigma}_j + \hat{\gamma}_j\hat{\gamma}_j^T) \\
\sigma_j^2 &= \frac{1}{M} E\left[(y_j - X\beta - Z\gamma_j)^T (y_j - X\beta - Z\gamma_j)\right]
\end{aligned} \tag{17.54}$$

where the residual variance can be further expressed as

$$\sigma_j^2 = \frac{1}{M} y_j^T (y_j - X\beta - Z\hat{\gamma}_j) \tag{17.55}$$

The E-step and M-step are alternated until a certain criterion of convergence is reached. Once the EM iteration converges, the predicted γ_j and the variance of the prediction are obtained as shown below,

$$\hat{\gamma}_j = \left(\hat{\mathit{\Pi}}^{-1}\hat{\sigma}_j^2 + Z^T Z\right)^{-1} Z^T (y_j - X\hat{\beta}) \tag{17.56}$$

and

$$\mathit{\Sigma}_j = \left(\hat{\mathit{\Pi}}^{-1}\hat{\sigma}_j^2 + Z^T Z\right)^{-1} \hat{\sigma}_j^2 \tag{17.57}$$

The F-test for $H_0 : L^T\gamma = 0$ is given by

$$\begin{aligned}
F_j =& \frac{1}{r}\hat{\gamma}_j^T L \left(L^T \mathit{\Sigma}_j L\right)^{-1} L^T \hat{\gamma}_j \\
=& \frac{1}{r\hat{\sigma}_j^2}\hat{\gamma}_j^T L \left[L^T \left(\hat{\mathit{\Pi}}^{-1}\hat{\sigma}_j^2 + Z^T Z\right)^{-1} L\right]^{-1} L^T \hat{\gamma}_j
\end{aligned} \tag{17.58}$$

We now compare the F-tests under the random model and that under the fixed model,

$$\begin{aligned} F_j(\text{random model}) &= \frac{1}{r\hat{\sigma}_j^2}\hat{\gamma}_j^T L \left[L^T \left(\hat{\Pi}^{-1}\hat{\sigma}_j^2 + Z^T Z \right)^{-1} L \right]^{-1} L^T \hat{\gamma}_j \\ F_j(\text{fixed model}) &= \frac{1}{r\hat{\sigma}_j^2}\hat{\gamma}_j^T L \left(L^T (Z^T Z)^{-1} L \right)^{-1} L^T \hat{\gamma}_j \end{aligned} \qquad (17.59)$$

The difference is obvious, an extra term $\hat{\Pi}^{-1}\hat{\sigma}_j^2$ occurs in the random model F-test. This illustrates the information sharing across genes. Further manipulation on the variance-covariance matrix of the predicted γ_j, we get

$$\Sigma_j = \left(\hat{\Pi}^{-1}\hat{\sigma}_j^2 + Z^T Z \right)^{-1} \hat{\sigma}_j^2 = \left(\hat{\Pi}^{-1} + \frac{1}{\hat{\sigma}_j^2} Z^T Z \right)^{-1} \qquad (17.60)$$

The first term $\hat{\Pi}^{-1}$ is shared for all genes and the second term $Z^T Z/\hat{\sigma}_j^2$ is gene specific. This idea is the same as the regularized t-test except that the shrinkage factor $\hat{\Pi}^{-1}$ is a matrix and it is estimated from the data rather than chosen by the investigator *a priori*.