The Plant Cell, Vol. 18, 2112–2121, September 2006, www.plantcell.org

CURRENT PERSPECTIVE ESSAY
SPECIAL SERIES ON LARGE-SCALE BIOLOGY

A Discussion of Statistical Methods for Design and Analysis of Microarray Experiments for Plant Scientists

There is much excitement among biologists and statisticians regarding new high-dimension data sets that have arisen from the application of microarray technology. In statistics, there has been a flurry of activity surrounding the development of new methods for the analysis of such data, and biologists are eager to extract as much information as possible from their substantial investments in microarray experiments. In this article, I offer statistical advice for plant biologists engaged in microarray research. My views are those of a statistician who has been working with scientists on the design and analysis of microarray experiments for the past 5 years. I will describe statistical concepts important for all researchers to understand and present data analysis strategies that I have found useful.

FUNDAMENTAL STATISTICAL CONCEPTS FOR SCIENTIFIC INFERENCE

In his text on experimental design (Fisher, 1951), the great statistician and quantitative geneticist R.A. Fisher wrote, ''My immediate point is that the questions involved can be dissociated from all that is strictly technical in the statistician's craft, and, when so detached, are questions only of the right use of human reasoning powers, with which all intelligent people, who hope to be intelligible, are equally concerned, and on which the statistician, as such, speaks with no special authority. The statistician cannot excuse himself from the duty of getting his head clear on the principles of scientific inference, but equally no other thinking man can avoid a like obligation.''

Fisher's principles of scientific inference include understanding how to design and analyze experiments to investigate the causal effects of treatments on a response variable of interest. Fisher's statement emphasizes that a clear understanding of this knowledge is the business of everyone engaged in science. Large-scale biology experiments differ from the experiments of Fisher's day in that we are able to simultaneously measure thousands of response variables (e.g., expression levels) for each experimental unit (e.g., a plant) rather than only one or a few. This complexity does not excuse us from paying attention to the basics of experimental design and data analysis. On the contrary, attention to the fundamentals of experimental design is more important now than ever before given the cost of experimentation and the immortalization of data sets in data repositories and databases. A poorly designed experiment can be costly to the individual investigator and can also hinder growing efforts to glean meaningful information via meta-analysis of multiple data sets.

www.plantcell.org/cgi/doi/10.1105/tpc.106.041616

FUNDAMENTAL PRINCIPLES OF EXPERIMENTAL DESIGN

Three fundamental experimental design principles attributed to Fisher are randomization, replication, and blocking. A scientist with a clear understanding of these three concepts will be well positioned to design effective experiments. Randomization is the practice of randomly assigning selected treatments to the experimental units available for use in an experiment. Replication involves applying a treatment independently to multiple experimental units and separately measuring responses for these experimental units. Blocking refers to the process of grouping similar experimental units together and assigning the treatments of interest (randomly) to the experimental units within such groups. There are many reasons for the use of randomization, replication, and blocking in experimental design. A complete discussion of these concepts is beyond the scope of this article, but I offer a brief nontechnical description of the importance of each concept below in the hopes that those who find their understanding lacking will make an effort to learn more.

Randomization, replication, and blocking are all motivated by the fact that there is variation among experimental units. No two experimental units will exhibit precisely the same response, even when treated identically. For example, microenvironmental variation causes differences even among genetically identical plants. Therefore, without knowledge of the degree of variation in the response among experimental units treated alike, it is impossible to judge whether differences in the response of experimental units treated differently are due to treatment differences or are instead simply a reflection of existing variation among experimental units. Replication enables us to assess the degree of existing variation in a response variable among experimental units treated alike. This enables us to recognize when differences between groups of experimental units treated differently are sufficiently large to suggest that changes in a response variable are due to treatment.

Randomization provides a mechanism for assigning treatments to experimental units that is free from intentional or unintentional biases that can be introduced by a researcher wishing to find evidence that a treatment causes changes in a response. Furthermore, the use of random assignment justifies the formal probability statements that play a central role in statistical inference. Perhaps the best way to grasp the importance

of randomization is to understand a method of inference due to Fisher known as a randomization test. Suppose, for example, that two treatments (A and B) are randomly assigned to eight plants, with four plants per treatment. Suppose that a quantitative response of interest is measured for each of the eight plants following treatment and that the resulting data are as in the first line of Table 1. Note that the average response for plants receiving treatment B is 2.0 units greater than the average response for plants receiving treatment A. A researcher might wish to conclude from this difference between averages that treatment B caused an increase in the response of interest relative to treatment A. There is, however, another explanation that must be considered before this conclusion can be drawn. There is clearly variation in the response of interest even among experimental units (plants) treated identically. Perhaps treatments A and B had no effect whatsoever on the response and the responses associated with these eight plants would have been exactly the same regardless of which treatment each received. The difference in averages could be simply a consequence of the random assignment of treatments to the eight experimental units; that is, it is possible that by chance the four plants that ultimately had the lowest four responses happened to be chosen to receive treatment A, while the four with the highest responses happened to be chosen to receive treatment B. Because treatments were randomly assigned to experimental units initially, we can compute the probability of such a coincidence.

Accepting for the moment the assumption that the treatments had no effect on the responses, there are 70 different data sets that could have resulted from our experiment. Each of these data sets corresponds to one of the 70 different ways that the eight experimental units could have been divided into two groups of four for treatments A and B. Nine of the 70 possibilities are presented explicitly in Table 1 along with the difference in treatment averages that would have resulted. The entire distribution of the 70 differences between averages is depicted in Figure 1. From this figure, we can see that most of the random assignments would have resulted in a difference between averages closer to 0 than the difference we observed in the actual experiment. Only two of the 70 random assignments (the first and last in Table 1) provide a difference in averages as far from 0 as the difference we observed. Thus, under the assumption of no treatment effect, the chance was 2/70 $\approx$ 0.0286 of seeing a difference in averages as far from 0 as the difference we observed. Because a difference in averages so far from 0 would be unlikely to occur if there were no treatment effect, we have good reason to believe that the treatments did indeed affect the response.

Table 1. Randomization: Example Data Set

Random Assignment	Treatment A	Treatment B	Difference between Averages
1	3.4 3.6 3.9 4.3	4.6 5.9 6.0 6.7	2.00
2	3.4 3.6 3.9 4.6	4.3 5.9 6.0 6.7	1.85
3	3.4 3.6 4.3 4.6	3.9 5.9 6.0 6.7	1.65
–	–	–	–
–	–	–	–
–	–	–	–
34	3.4 4.3 4.6 6.7	3.6 3.9 5.9 6.0	0.10
35	3.4 3.9 5.9 6.0	3.6 4.3 4.6 6.7	0.00
36	3.6 4.3 4.6 6.7	3.4 3.9 5.9 6.0	0.00
37	3.6 3.9 5.9 6.0	3.4 4.3 4.6 6.7	−0.10
–	–	–	–
–	–	–	–
–	–	–	–
69	4.3 5.9 6.0 6.7	3.4 3.6 3.9 4.6	−1.85
70	4.6 5.9 6.0 6.7	3.4 3.6 3.9 4.3	−2.00

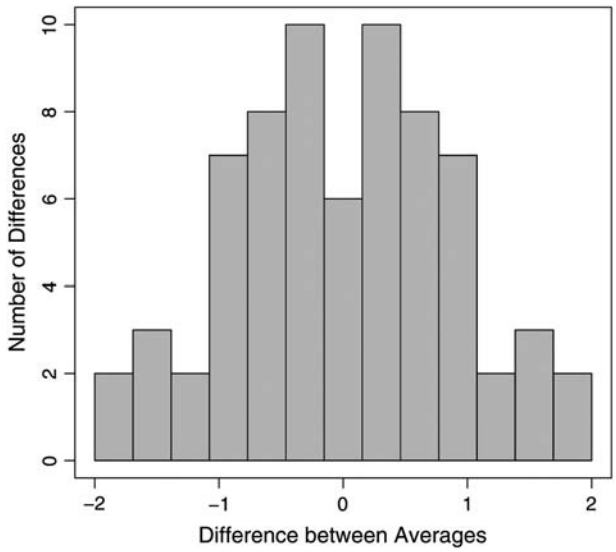

Figure 1. Randomization: Distribution of 70 Possible Differences between Averages for the Example Data in Table 1.

The quantity 2/70 is an example of a P value. The use of P values for detecting differentially expressed genes is discussed in a subsequent section of this article. For now, note that we were able to compute a P value in this example without making any distributional assumptions about the data. (For example, we did not need to assume the data were normally distributed, as is done when conducting a standard two-sample *t* test.) The key point for computing the P value was that all 70 random assignments were equally likely to have occurred due to our initial random assignment of treatments to the experimental units. Without this random assignment, the argument for a treatment effect breaks down, and in this way, randomization plays a crucial role in establishing that a treatment causes changes in a response. Note that replication was also essential in our argument. Had there been only one experimental unit (plant) for each treatment, any difference between the plants could have been attributed to natural variation between experimental units rather than a treatment effect.

CURRENT PERSPECTIVE ESSAY

Blocking, the third of Fisher's fundamental design principles, is used when it is recognized before the beginning of an experiment that certain groups of experimental units are expected to be more homogenous than experimental units in general. For example, if plants serving as experimental units have been grown on different shelves in a growth chamber, it may be appropriate to consider each group of plants sharing a single shelf as a block. Ideally, all treatments of interest would be randomly assigned to plants in each block. This strategy permits a relatively precise comparison of treatments among experimental units in which microenvironmental variation has been minimized and avoids the possibility of partial or complete confounding between the effects of blocks and the effects of treatments. Such confounding would occur, for example, in the extreme case in which all the plants receiving a particular treatment were on one shelf, while all the plants receiving another treatment were on a different shelf. In this case, there would be no way to distinguish differences in the response caused by the treatments from differences due to shelf effects. Blocking can also be used effectively when the workload associated with treating and measuring experimental units requires that the experiment be divided over multiple time periods as is often the case in microarray experimentation. The experimental units processed during any single time period (e.g., a day) can form a block in which all treatments of interest will be assessed (in a random order). Replication can then by achieved by repeating the process over multiple time periods with a newly randomized processing order for each time period.

Blocking, randomization, and replication are the fundamental components of experimental design. These concepts can be applied in a variety of ways to create designs ranging from very simple to extremely complex. All researchers engaged in experimentation should have a solid understanding of these principles before wrestling with specific issues that arise when designing microarray experiments. Such issues have been discussed by several authors, including Kerr and Churchill (2001a, 2001b), Churchill (2002), Yang and Speed (2002), Dobbin et al. (2003), Kendziorski et al. (2003b, 2005), Kerr (2003), Glonek and Solomon (2004), and Altman (2005), among many others. While it is important to learn the special issues that arise when designing microarray experiments, such knowledge is no substitute for knowing the basics of experimental design.

Allison et al. (2006) provide a recent summary of some key points that have emerged from the microarray-specific experimental design literature. One point involving replication is worth repeating here. Two types of replication, biological and technical, are often discussed in the context of microarray experimental design. Technical replication involves measuring a given experimental unit multiple times. Biological replication is the replication referred to in my remarks above in which multiple independent experimental units (for example, plants or separate pools of plants) are measured individually for each treatment. Biological replication is essential for attributing observed changes in expression to the effects of treatment. Technical replication is not. For a given number of microarray slides or chips, power for detecting a treatment effect will be maximized by measuring each experimental unit only once. For a fixed number of experimental units, power for detecting differences can be improved to some extent by measuring experimental units multiple times because an average of many measurements is less variable than a single measurement. However, the power for detecting differential expression will always be limited by the number of biological replications regardless of how many measurements are obtained for each experimental unit. Technical replication is useful for separating variability associated with the measurement process from biological variation across experimental units, but it is not necessary to separately estimate the variation from these sources when the primary goal is to determine whether a treatment causes a change in expression. Thus, when the cost of microarray slides or chips is the limiting factor governing the size of an experiment, measuring each experimental unit only once is recommended to maximize biological replication and, thus, power to detect expression differences. Dobbin et al. (2003) provide formal statistical arguments to support these claims in the context of two-treatment two-color microarray experiments. Straightforward statistical arguments can be used to extend these ideas to more complex experimental designs and single-channel platforms.

MIXED LINEAR MODEL ANALYSIS OF MICROARRAY EXPERIMENTS

Many methods have been proposed for the analysis of microarray experiments. I make no attempt to provide a comprehensive review here. Instead, I will simply describe a general analysis strategy that I have found effective in a wide variety of circumstances. To clarify important concepts, it will be helpful to have in mind a hypothetical example experiment described as follows. Suppose researchers are interested in studying the effect of soil moisture and a viral infection on gene expression. Three soil moisture levels (low, medium, and high) are randomly assigned to 15 individually potted plants, such that five plants are treated with each moisture level. Suppose that two leaves of comparable developmental stage are identified for each plant. One of the leaves on each plant is randomly selected for infection with a plant virus. The other leaf receives the same treatment (i.e., infiltration, injection, or topical application) without the virus to serve as an uninfected control. Suppose that after the soil moisture and viral treatments have been applied for a relevant length of time, sufficient RNA can be extracted from each leaf to obtain a measure of expression for a gene of interest or perhaps for thousands of genes using microarray technology.

For a given gene, many potentially interesting questions can be posed. For example, does varying moisture level affect the expression level of the gene either in infected leaves or in uninfected control leaves? Are any changes due to varying moisture level the same in infected and in uninfected control leaves? Do expression levels differ between infected and

uninfected leaves at low, medium, or high moisture levels? Are any differences that may exist between infected and uninfected leaves the same for each soil moisture level? For each gene, measures of expression in 30 leaves are available. How can such data be used to address these questions?

Generally speaking, data analysis should be matched with experimental design. The use of linear or mixed-effects linear modeling strategies provides a general framework for data analysis that naturally incorporates experimental design. For the simplest of experiments (for example, a completely randomized design with two treatments), a linear model analysis of normalized log-scale expression measures for a single gene would amount to the two-sample *t* test that students are taught in an introductory undergraduate statistics course. More complex experiments, like the one described above, may use approximate *t* tests or *F*-tests as part of a mixed-effects linear model analysis conducted separately for each gene. Mixed-effects linear models (also known as mixed linear models) are ''mixed'' in that they include both fixed and random effects. The fixed effects specify the mean of the response variable as a function of treatment conditions of interest. The random effects specify the correlation structure among observations of the response variable that might arise due to the structure of the experimental design. Typically, the scientific questions of interest are addressed by testing hypotheses regarding the fixed-effects parameters. The correlation structure specified by the random effects is taken into account when judging the statistical significance of an observed test statistic.

To illustrate these concepts, consider a mixed linear model for a single gene from our example experiment. There are six treatment conditions specified by the six combinations of soil moisture level (L = low, M = medium, and H = high) and virus exposure (I = infected and U = uninfected). Denote these six treatments as follows: 1 = LI, 2 = MI, 3 = HI, 4 = LU, 5 = MU, and 6 = HU. Associated with these six treatments are six underlying mean (log-scale) expression levels denoted μ_1, μ_2, μ_3, μ_4, μ_5, and μ_6. Scientific questions of interest can be addressed by answering questions about these means. For example, a test of $\mu_1 = \mu_2 = \mu_3$ addresses whether varying moisture level affected the expression level of the gene in infected leaves; a test of $\mu_1 = \mu_4$ addresses whether expression levels in infected and uninfected leaves differed in plants with low soil moisture; and a test of $\mu_1 - \mu_4 = \mu_2 - \mu_5 = \mu_3 - \mu_6$ addresses whether expression level differences between infected and uninfected leaves were the same for each soil moisture level. Although the means are unknown, they can be estimated from the data simply by averaging the five log-scale expression measurements obtained from the five leaves associated with each combination of soil moisture level and virus exposure. Let a_1, a_2, a_3, a_4, a_5, and a_6 denote these averages.

To determine, for example, whether $\mu_1 = \mu_4$ seems plausible based on our observed data, we would examine $a_1 - a_4$ as an estimator of $\mu_1 - \mu_4$. Due to natural variation among leaves, microarray chips, hybridization conditions, etc., $a_1 - a_4$ will not equal $\mu_1 - \mu_4$. We can use variation in the observed data to estimate the variation of $a_1 - a_4$ as an estimator of $\mu_1 - \mu_4$. The estimator $a_1 - a_4$ divided by the square root of its estimated variation (standard error) serves as a test statistic for testing whether $\mu_1 - \mu_4 = 0$. Values close to 0 suggest that $\mu_1 = \mu_4$ is plausible, while values far from 0 provide evidence that μ_1 and μ_4 differ. Results of tests are typically summarized by P values that are discussed at length in the next section. Note that even large values of $a_1 - a_4$ may not provide evidence that μ_1 and μ_4 differ if the standard error of $a_1 - a_4$ is large. When the data suggest great uncertainty in our estimates, even large differences do not provide compelling evidence of a treatment effect, and for this reason, point estimates alone (e.g., expression fold changes) are not effective for identifying differential expression.

To properly estimate the variation in our estimators from the observed data, we must recognize that all 30 observations are not independent of one another. The 30 leaves were obtained by sampling two leaves from each of 15 plants, and this should be accounted for in our model for the data. One way for our model to capture this structure in our data is to include a random effect for each plant. The observations from the two leaves on a single plant will share that plant's random effect. This will account for the natural variation in expression from plant to plant that is unrelated to soil moisture and virus exposure. The observations from the two leaves on a single plant will be positively correlated because of their shared random effect. This positive correlation implies that when the measure of expression in one leaf is above (below) the average for its treatment condition, the other leaf is more likely than not to be above (below) the average expression for its treatment group. This is consistent with the idea that some plants will tend to have higher (lower) levels of expression than others aside from the effects of treatment. We model the plant effects in this experiment as random effects because we are not interested specifically in the 15 plants used in this experiment but are rather interested in making inferences to a larger population of plants from which the 15 at hand can be considered to be like a random sample. We consider the effects associated with soil moisture level and virus exposure to be fixed rather than random because our attention is fixed on these particular treatment combinations whose effects are not assumed to be like a random sample from a larger population of effects.

Linear and mixed-effects linear models have a long history of use in science dating back to Fisher's well known analysis of variance (ANOVA). The opening chapter of McCulloch and Searle (2001) provides a more thorough introduction to the topic than I have been able to provide here. Kerr et al. (2000) were among the first to recommend the use of linear model analysis for microarray data; Wolfinger et al. (2001) recommended the use of mixed linear models shortly thereafter. There are many different analysis strategies that could correctly be described as linear or mixed linear model approaches. For example, current use of linear models in microarray analysis often differs from the original proposal of Kerr et al. (2000) in that a separate analysis is conducted for each gene rather than attempting to fit one large

CURRENT PERSPECTIVE ESSAY

linear model to all genes simultaneously. Different genes tend to exhibit different levels of variation in expression, and this heterogeneity is difficult to address with one large linear model for all genes. Baldi and Long (2001), Wright and Simon (2003), Smyth (2004), and Cui et al. (2005) offer linear model approaches that can be viewed as compromises between global and gene-specific modeling of expression variability. These methods take advantage of data from a large number of genes to outperform individual gene analyses in simulations, particularly when within-gene samples sizes are low.

USING P VALUES TO IDENTIFY DIFFERENTIALLY EXPRESSED GENES

Regardless of the details behind linear or mixed linear model analysis strategies, a P value for a test of interest is typically obtained for each of thousands of genes. In the simplest case, each P value might correspond to a test whose null hypothesis claims that a particular gene is not differentially expressed across two or more conditions. The P value from a single test is often misunderstood to be the probability that the null hypothesis is true. Though such a probability would be quite useful in decision making, it is not the correct interpretation of a P value. The P value is the probability, computed under the assumption that the null hypothesis is true, of obtaining a data set that provides as much or more evidence of differential expression than the data observed in the experiment at hand. Thus, the P value makes a statement about the probability of data under an assumption about the true state of nature rather than a probability statement about the true state of nature, given the observed data.

Many researchers errantly believe that if a gene is not differentially expressed, it will tend to have a large P value when tested for differential expression. In fact, when a true null hypothesis is tested using an appropriate continuous test statistic, the P value for the test will be uniformly distributed on the interval 0 to 1. This means that the P value is equally likely to fall anywhere between 0 and 1, and a small P value is just as likely to occur as a large one. Conversely, if a gene is differentially expressed, a P value for the test of its differential expression is more likely to be small than large, though large P values are still quite possible. This situation is illustrated in Figure 2, which shows the P value distribution for a two-sample *t* test for various states of nature. Considered are three different levels of fold change: a fold change of 1, indicating no differential expression, a fold change of 1.5, indicating a 50% increase in expression level, and a fold change of 2, indicating a doubling of expression level. The variance of expression on the log scale has been fixed at 1, and two treatment groups with five experimental units in each group have been assumed. The area under a curve in any subinterval between 0 and 1 represents the probability that the P value will fall in that subinterval under the specified conditions. As the degree of differential expression increases, smaller P values become more likely, but it is important to understand that the full range of P values is possible in all three situations. As the number of experimental units per treatment group is increased, the curves for fold changes of 1.5 and 2.0 will shift probability from the high P values to low P values; however, the P value distribution for a fold change of 1.0 will always remain uniform. It is valuable to keep these facts in mind when interpreting thousands of P values that result from the analysis of a microarray experiment.

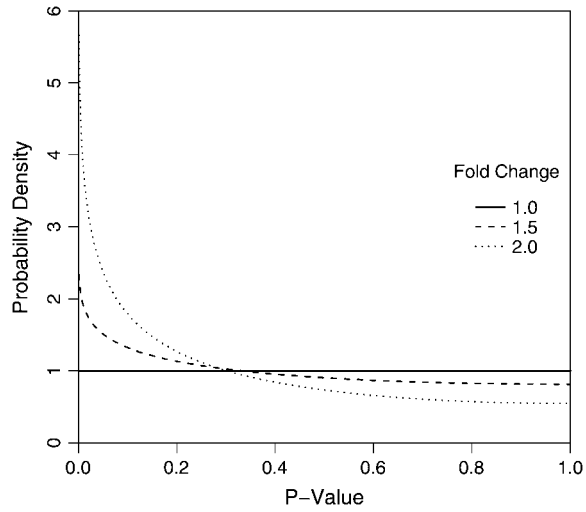

Figure 2. Example of P Value Distributions for the Two-Sample *t* Test.

Figure 3 shows a histogram of 10,000 P values computed from *t* tests on data simulated to mimic a simple microarray experiment. Note that the P value distribution appears to be a mixture of uniformly distributed P values and P values that tend to be smaller than uniform, indicating a mixture of nondifferentially and differentially expressed genes. There are many statistical tools that can be used to identify differentially expressed genes from such a P value distribution. I will demonstrate one approach by applying it to the simulated P values in Figure 3. Because the P values were simulated, we know the truth about differential expression for each gene and thus will be able to evaluate the performance of the method for this example.

Benjamini and Hochberg (1995) introduced the concept of false discovery rate (FDR) for inference from multiple P values. Storey and Tibshirani (2003) presented a method for identifying differentially expressed genes that provides approximate control of the FDR. To understand the meaning of the FDR associated with a method for identifying differentially expressed genes, it is useful to imagine a scientist who conducts an infinite number of microarray experiments. For each experiment, the scientist uses a certain method for producing a list of genes declared to be differentially expressed. Each list may contain some false positive results, that is, genes that are not truly differentially expressed but nonetheless were declared to be differentially expressed. Consider, for each list, the fraction of false positive results given by the number of false positive results

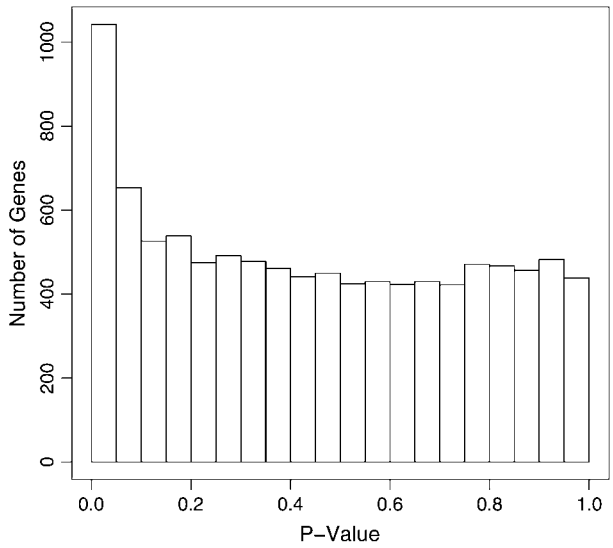

Figure 3. Histogram of 10,000 P Values from a Simulated Data Set.

on the list divided by the number of genes on the list, defining the fraction to be zero for a list that contains no genes at all (that is, a list for an experiment where no differences in expression were declared). The average of these fractions is the FDR for the method used to make the gene lists. Thus, a method that controls the FDR at 5%, for example, provides an average false positive fraction of 5% in the long run. The false positive fractions will vary from list to list. A method that controls the FDR at 5% does not guarantee that the false positive fraction will be no larger than 5% for a given experiment; instead, the 5% figure makes a statement about the average performance of the method in the long run.

Storey and Tibshirani (2003) convert the P values from a given experiment to *q*-values. The *q*-values are convenient for producing a gene list for any desired FDR. For example, if a researcher wishes to control FDR at 5%, he or she may declare all genes with *q*-values <0.05 to be differentially expressed. Some of these declarations of differential expression are likely to be in error, but the *q*-values are designed so that this strategy for producing a gene list will approximately control FDR at 5%.

As described by Storey and Tibshirani (2003), each *q*-value is a nontrivial function of all P values observed in the experiment. By construction, the *q*-values have the same order as the P values. Thus, a gene list consisting of genes with the smallest *q*-values will always include the genes with the smallest P values. Despite this connection between P values and *q*-values, the *q*-values are more directly useful than P values when conducting many tests because the *q*-values are easy to interpret in terms of FDR as described above. In contrast with *q*-values, P values are naturally interpreted in terms of the expected proportion of all nondifferentially expressed genes that are errantly declared to be differentially expressed. For example, if all genes with P values <0.01 are declared to be differentially expressed and placed on a list, we would expect $\sim$1% of all nondifferentially expressed genes to appear on the list. Because 1% is a proportion of all nondifferentially expressed genes rather than a proportion of genes on the list, using the P values directly is less practical than using the *q*-values. For example, a list of 100 genes with P values <0.01 may or may not be useful depending on the total number of nondifferentially expressed genes in the experiment. If only 1000 genes are nondifferentially expressed, then the list of 100 genes with P values <0.01 would be expected to contain $\sim 0.01 \times 1000 = 10$ false positives and would thus be potentially useful. However, the list would be expected to consist almost entirely of false positives in an experiment in which 10,000 genes were nondifferentially expressed ($0.01 \times 10{,}000 = 100$). Thus, the utility of a list generated from a P value threshold for significance cannot be properly assessed without additional information and calculation; by contrast, the interpretation of a list generated from a *q*-value threshold for significance is relatively straightforward.

When the P values in Figure 3 are converted to *q*-values using the method of Storey and Tibshirani (2003), FDR control at 5, 10, 15, and 20% yields lists of 3, 23, 51, and 156 genes, respectively. Because the data were simulated, we are able to compute the actual false positive fractions for each of these gene lists as $0/3 = 0\%$, $3/23 \approx 13\%$, $6/51 \approx 12\%$, and $26/156 \approx 17\%$, respectively. Note that the actual false positive fractions were no larger than the nominal FDR levels for the 5, 15, and 20% lists, although, as discussed above, the method does not guarantee this type of control. The observed false positive fraction of $\sim$13% for the 10% FDR list illustrates that FDR control does not guarantee control of the false positive fraction in any particular experiment; rather, it aims to control the average of such fractions over repeated experimentation.

A researcher could choose any of the four gene lists for follow-up research. The 5% FDR list is perfect in the sense that it contains no false positive results, but it identified only three differentially expressed genes. At the other extreme, the 20% FDR gene list identified 156 genes of which 26 were false positives. The greater level of discovery may be worth the higher error rate depending upon the purpose of the experiment and the way in which the gene list will be used in follow-up research. There is no one FDR level that is appropriate for all experiments. FDR control at 5% might yield a list of thousands of genes in a different experiment. Thus, FDR levels lower than 5% might sometimes be desired to keep the number of identified genes at a manageable level.

Like most methods in statistics, FDR focuses on the control of type 1 errors, which, in this case, are errors in declaring genes to be differentially expressed when in fact they are not. Type 2 errors involve failing to identify a differentially expressed gene as such. When the number of replications in a microarray experiment is low and many genes have only small changes in expression, the number of type 2 errors may be quite large. For the simulated data of Figure 3, 1500 genes were simulated to be differentially expressed. Thus, even when controlling the FDR at

CURRENT PERSPECTIVE ESSAY

20%, only a small fraction of the truly differentially expressed genes were detected. It is not possible to construct a gene list based on P values (or *q*-values) that captures all the differentially expressed genes without also including many nondifferentially expressed genes. It is easy to see why this is so by reexamining Figure 2. Many nondifferentially expressed genes will have small P values when thousands of genes are not differentially expressed. Also, some differentially expressed genes may have relatively large P values. Thus, the mixing of P values between nondifferentially and differentially expressed genes is inevitable. The degree of mixing can be reduced by increasing the number of replications used in an experiment because, as mentioned previously, the P value distributions for differentially expressed genes shift toward small P values as the number of replications increases.

The high number of type 2 errors in the simulated example data set is most likely not atypical for current microarray experiments. A high number of type 2 errors does not mean that such gene lists are not useful, but it is important for researchers to understand that the true number of differentially expressed genes will often be much larger than the number of genes that can be declared to be differentially expressed when controlling FDR or other error measures. There are methods for estimating the number of differentially expressed genes that could be used to approximate the extent of type 2 error for a given gene list. For example, the method of Langaas et al. (2005) estimates the number of differentially expressed genes to be 1110 for the P values in Figure 3. Although this is an underestimate of the actual 1500 differentially expressed genes, it correctly indicates that even the 20% FDR list of 156 genes will result in several hundred type 2 errors.

TESTING FOR INTERACTION

Researchers often wish to compare or contrast multiple gene lists to find genes of interest. For example, consider an experiment in which plants of two genotypes (e.g., wild type and mutant) are exposed to nonstress and stress conditions. Researchers may be interested in finding genes that change expression in response to stress in one genotype but not the other. It is natural to produce a list of differentially expressed genes for each genotype and to search for genes that appear on one list but not the other. Venn diagrams are often used to display the results of such an analysis. One major problem with this approach is that, as illustrated in the previous section, each gene list may contain only a fraction of the truly differentially expressed genes. The absence of a gene on one list should not be taken to mean that the gene does not change expression in response to stress in that particular genotype. It simply means that there was not sufficient evidence to declare the gene differentially expressed when trying to control FDR or a similar error measure.

As an alternative to comparing gene lists, a test for interaction can be used to directly search for genes whose expression change in one genotype differs from its expression change in the other. Interaction can occur in experiments with multiple factors, where each factor has multiple levels. A treatment is defined by a combination of one level from each factor. In the example experiment of this section, there are two factors (genotype and stress), each with two levels (wild type versus mutant for the genotype factor and absent versus present for the stress factor). Thus, we have four treatments: 1 = wild type, stress absent; 2 = wild type, stress present; 3 = mutant, stress absent; and 4 = mutant, stress present. If we let μ_i denote the mean log-scale expression level for a given gene under the i^{th} treatment, the null hypothesis for the interaction test is H_0: $\mu_1 - \mu_2 = \mu_3 - \mu_4$. If this null hypothesis is true, the effect of stress on the gene's expression is the same within both genotypes. If the null hypothesis is false, the effect of stress differs for the wild type and mutant plants. Genes that exhibit significant interaction are perhaps of greatest scientific interest if the goal is to understand how the mutation affects the plant's ability to cope with stress at the molecular level.

In experiments involving time as a factor and another generic factor, say condition, testing for time-by-condition interaction can be used to identify genes whose expression difference across conditions at an initial time point differs from the difference across conditions at a later time point. Thus, the test for interaction can identify many genes of interest, including those that do not differ across conditions initially but develop differences across conditions as time unfolds.

In general, interaction is present when the effects of one factor on the response vary across levels of a second factor. Interaction is often the most interesting type of differential expression in multifactor experiments. Linear modeling provides a natural framework for producing P values for tests of interaction that can be used to find genes of greatest interest. This approach is likely to be more meaningful than comparing gene lists that result from separate analyses.

EXAMPLES FROM THE LITERATURE

In this section, I describe a few articles from the plant microarray literature that make use of valid statistical tools for microarray data analysis. This section is, of course, not meant to contain an exhaustive accounting of such articles. Also, although it might be argued that none of the studies is perfect in every detail, I withhold all minor criticisms and instead focus on positive aspects of these examples with the intent of helping readers better appreciate the value of statistical methods in plant microarray research.

Vuylsteke et al. (2005) conducted mixed linear model analyses of expression variation across seven *Arabidopsis* genotypes consisting of three inbred lines and a subset of possible reciprocal crosses among these inbreds. They measured expression using two-color cDNA microarrays and a loop design with two biological replications for each of the seven genotypes. The loop design was arranged so that each sample was measured an equal number of times with each dye and so that genotype pairs

representing comparisons of greatest interest were hybridized together on individual slides. Their analyses illustrate the flexibility of mixed linear modeling for directly addressing a variety of scientific questions of interest using multiple contrasts of estimated means. For example, they use contrasts of estimated means (1) to identify genes whose expression differs between any pair of inbreds, (2) to identify genes whose expression differs between reciprocal crosses of a given pair of inbred lines, and (3) to identify genes whose expression in a hybrid differs from the average expression of the two parental lines. The authors used *q*-values associated with the contrasts to identify differentially expressed genes and displayed results using volcano plots in their Figure 2. Volcano plots have been used by many authors to illustrate the relationship between estimated fold change and measures of statistical significance, such as P values or *q*-values.

Vanneste et al. (2005) used two-factor ANOVA and an interesting clustering approach to identify 913 *Arabidopsis* genes as candidates for encoding regulatory proteins required for lateral root initiation. The ATH1 Affymetrix array was used to measure expression in root segments from the wild type and the dominant auxin signaling mutant *solitary root1* (*slr1*) at three time points during the early events of lateral root initiation. Two biological replications were used for each combination of genotype and time point. Gene-specific two-factor ANOVA identified 3110 genes with a significant (P value < 0.001) genotype main effect, time main effect, or genotype-by-time interaction. The P values were converted to *q*-values to ensure that the 0.001 P value threshold for significance would correspond to a low FDR (*q*-value < 0.05). The 3110 identified genes yielded 6220 estimated expression patterns across the three time points by estimating separate patterns in the wild-type and *slr1* genotypes for each gene. These 6220 patterns were separated into 14 clusters, and the cluster memberships of the two patterns associated with each gene were noted. Genes whose wild type cluster suggested a greater induction over time than that suggested by the cluster of the mutant pattern were identified as lateral root initiation (LRI) genes. Most of these LRI genes (815 out of 913) exhibited significant interaction between genotype and time (*q*-value < 0.10), which provides formal evidence that most of the LRI genes exhibited patterns of expression during the early stages of LRI that differed significantly between wild-type and *slr1* plants. Subsequent analysis of the functional annotations of the LRI genes led to a variety of biological insights, including the presentation of a model of the auxin-dependent regulatory network influencing LRI (Vanneste et al., 2005).

DeCook et al. (2006) used measures of gene expression and molecular marker genotypes to identify expression quantitative trait loci (eQTL) during the process of shoot formation in *Arabidopsis*. These eQTL are genomic locations associated with the expression of one or more genes. For each of 30 recombinant inbred lines, root explants from several hundred seedlings were treated with a shoot induction medium and pooled for RNA extraction and hybridization to an Affymetrix GeneChip. Expression levels of >20,000 genes were tested for association with each of 288 molecular markers spaced evenly throughout the genome. These tests for association yielded ~6 million P values. A permutation-based approach was used to approximate the FDR for various significance thresholds. Several thousand significant associations between marker loci and gene expression levels were identified. Discovering such relationships is a first step toward understanding the molecular genetic mechanisms underlying quantitative variation in shoot formation. Two loci previously identified by Lall et al. (2004) as shoot development QTL were shown by DeCook et al. (2006) to be associated with the expression of many genes. Studying the functions of such genes can provide insight into the mechanisms by which these QTL influence shoot formation.

Caldo et al. (2004) presented mixed linear model analyses aimed at identifying barley (*Hordeum vulgare*) genes involved in distinguishing compatible from incompatible plant–pathogen interactions. The Affymetrix Barley1 GeneChip was used to measure expression in three near-isogenic barley lines at six time points following exposure to two isolates of *Blumeria graminis* f sp *hordei*, the fungal pathogen that causes powdery mildew disease in compatible interactions. Three independent biological replications, each consisting of a pool of 15 2-week-old seedlings, were separately measured for all 36 combinations of barley genotype, fungal isolate, and time following inoculation. A mixed linear model analysis was conducted separately for the 108 data points obtained for each gene. A contrast of estimated means was used as part of each mixed linear model analysis to identify genes whose average pattern of expression following fungal inoculation differed between compatible and incompatible interactions. FDR considerations led to the identification of 22 genes whose expression patterns in compatible interactions were suppressed relative to the expression patterns in incompatible interactions in the latter half of the profiled time course. This period of suppression coincided with the establishment of membrane-to-membrane contact between fungal haustoria and host epidermal cells. The discovery is consistent with the hypothesis that host-specific resistance evolved from the recognition and prevention of the pathogen's suppression of plant basal defense. This work provides an example where many of the concepts discussed in this essay (blocking, randomization, replication, mixed linear model analysis, tests for interaction, and FDR estimation) were used to obtain unique biological insights that may have been difficult to uncover using other approaches.

CONCLUDING REMARKS

I have described some statistical methods that I have found useful for microarray experimental design and data analysis. There are many other statistical approaches that have merit for addressing many of the same problems that I have discussed. Nonparametric resampling based approaches (e.g., significance analysis of microarrays proposed by Tusher et al. [2001]) and

CURRENT PERSPECTIVE ESSAY

Bayesian or empirical Bayesian approaches (e.g., Kendziorski et al., 2003a) can be effective when experimental design, data structure, and questions of interest facilitate their use. Linear and mixed linear model approaches currently have advantages over these and other existing methods whenever tests for interaction are of primary interest or when experimental design complexity suggests the need for multiple random effects to account for multiple sources of variation.

The P values computed from linear or mixed linear model analyses depend on assumptions of normality and within-gene constant variance that are never precisely satisfied in practice. The famous statistician George E.P. Box is credited with the quote, ''All models are wrong. Some models are useful.'' My experience suggests that linear and mixed linear models are useful for the analysis of microarray data, though better methods for checking model assumptions and evaluating the impact of departures from these assumptions are needed. Many such tools are available for the analysis of individual data sets, but most of the standard methods do not extend effectively to the simultaneous analysis of thousands of dependent genes. This issue is not unique to linear and mixed linear model approaches, as all methods for microarray data analysis are based on some assumptions about the data. More statistical research in this area is warranted.

This essay has focused on gene-specific analyses of microarray data, but there are many other aspects of microarray data analysis worthy of discussion. For example, clustering of genes or samples based on expression profiles is routinely used in conjunction with gene-specific analysis to visualize, organize, and interpret results from a broader perspective. In addition, there are now several methods for identifying sets of functionally related genes that have jointly undergone significant changes in expression (see, for example, Barry et al., 2005; Subramanian et al., 2005; and the discussion of related approaches in Allison et al., 2006). This recent work provides an example of how available biological information on gene function can be incorporated formally in statistical analysis to achieve greater insights than would be otherwise possible.

The main point of my article is to encourage greater attention to statistical thinking in plant microarray work rather than to dictate specifically how microarray experiments should be designed and analyzed. One hallmark of a statistical approach is a clear recognition of uncertainty. Statistical methods provide not only an answer (e.g., a list of differentially expressed genes) but also an assessment of the uncertainty associated with that answer (e.g., an estimated FDR). While it can be challenging to deal with that uncertainty, its recognition plays an important role in scientific endeavors. Much of my essay has focused on elementary concepts in statistics that many readers will undoubtedly know well. However, my own interactions with scientists and aspiring scientists and my reading of the microarray literature suggest that a greater attention to fundamental statistical concepts is warranted. Some researchers seem to regard the statistical aspects of research as a necessary evil that perhaps can be alleviated by a good software package. From my perspective, statistical issues are very much at the heart of science, and I hope that statistical thinking will play a more prominent role in the data-rich science of modern plant biology.

ACKNOWLEDGMENTS

I thank the many plant scientists whom I have had the pleasure to work with over the past several years. I also thank several anonymous reviewers for their comments that have helped to improve this essay. My work has been supported by USDA Grant 2002-35300-12619 and by National Science Foundation Plant Genome Grant 05-00461.

Dan Nettleton
Department of Statistics
Iowa State University
Ames, IA 50011-1210
dnett@iastate.edu

REFERENCES

Allison, D.B., Cui, X., Page, G.P., and Sabripour, M. (2006). Microarray data analysis: From disarray to consolidation and consensus. Nat. Rev. Genet. **7,** 55–65.

Altman, N.S. (2005). Replication, variation and normalization in microarray experiments. Appl. Bioinformatics **4,** 33–44.

Baldi, P., and Long, A. (2001). A Bayesian framework for the analysis inferences of gene changes. Bioinformatics **17,** 509–519.

Barry, W.T., Nobel, A.B., and Wright, F.A. (2005). Significance analysis of functional categories in gene expression studies: A structured permutation approach. Bioinformatics **21,** 1943–1949.

Benjamini, Y., and Hochberg, Y. (1995). Controlling false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc. Series B **57,** 289–300.

Caldo, R.A., Nettleton, D., and Wise, R.P. (2004). Interaction-dependent gene expression in *Mla*-specified response to barley powdery mildew. Plant Cell **16,** 2514–2528.

Churchill, G.A. (2002). Fundamentals of experimental design for cDNA microarrays. Nat. Genet. **32** (suppl.), 490–495.

Cui, X., Hwang, J.T.G., Qiu, J., Blades, N.J., and Churchill, G.A. (2005). Improved statistical tests for differential gene expression by shrinking variance component estimates. Biostatistics **6,** 59–75.

DeCook, R., Lall, S., Nettleton, D., and Howell, S.H. (2006). Genetic regulation of gene expression during shoot development in *Arabidopsis*. Genetics **172,** 1155–1164.

Dobbin, K., Shih, J.H., and Simon, R. (2003). Statistical design of reverse dye microarrays. Bioinformatics **19,** 803–810.

Fisher, R.A. (1951). Design of Experiments, 6th ed. (Edinburgh, UK: Oliver and Boyd).

Glonek, G.F.V., and Solomon, P.J. (2004). Factorial and time course designs for cDNA microarray experiments. Biostatistics **5,** 89–111.

Kendziorski, C., Irizarry, R.A., Chen, K.-S., Haag, J.D., and Gould, M.N. (2005). On the utility of pooling biological samples in microarray experiments. Proc. Natl. Acad. Sci. USA **102,** 4252–4257.

Kendziorski, C.M., Newton, M.A., Lan, H., and Gould, M.N. (2003a). On parametric empirical Bayes methods for comparing multiple

groups using replicated gene expression profiles. Stat. Med. **22,** 3899–3914.

Kendziorski, C.M., Zhang, Y., Lan, H., and Attie, A.D. (2003b). The efficiency of mRNA pooling in microarray experiments. Biostatistics **4,** 465–477.

Kerr, M.K. (2003). Design considerations for efficient and effective microarray studies. Biometrics **59,** 822–828.

Kerr, M.K., and Churchill, G.A. (2001a). Experimental design for gene expression microarrays. Biostatistics **2,** 183–201.

Kerr, M.K., and Churchill, G.A. (2001b). Statistical design and the analysis of gene expression microarrays. Genet. Res. **77,** 123–128.

Kerr, M.K., Martin, M., and Churchill, G.A. (2000). Analysis of variance for gene expression microarray data. J. Comput. Biol. **7,** 819–837.

Lall, S., Nettleton, D., DeCook, R., Che, P., and Howell, S.H. (2004). QTLs associated with adventitious shoot formation in tissue culture and the program of shoot development in *Arabidopsis*. Genetics **167,** 1883–1892.

Langaas, M., Lindqvist, B., and Ferkingstad, E. (2005). Estimating the proportion of true null hypotheses, with application to DNA microarray data. J. R. Stat. Soc. Series B **67,** 555–572.

McCulloch, C.E., and Searle, S.R. (2001). Generalized, Linear, and Mixed Models. (New York: John Wiley & Sons).

Smyth, G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol. **3,** Article3.

Storey, J.D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. USA **100,** 9440–9445.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA **102,** 15545–15550.

Tusher, V.G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. USA **98,** 5116–5121.

Vanneste, S., et al. (2005). Cell cycle progression in the pericycle is not sufficient for SOLITARY ROOT/IAA14-mediated lateral root initiation in *Arabidopsis thaliana*. Plant Cell **17,** 3035–3050.

Vuylsteke, M., van Eeuwijk, F., Van Hummelen, P., Kuiper, M., and Zabeau, M. (2005). Genetic analysis of variation in gene expression in *Arabidopsis thaliana*. Genetics **171,** 1267–1275.

Wolfinger, R.D., Gibson, G., Wolfinger, E.D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., and Paules, R.S. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. J. Comput. Biol. **8,** 625–637.

Wright, G.W., and Simon, R.M. (2003). A random variance model for detection of differential gene expression in small microarray experiments. Bioinformatics **19,** 2448–2455.

Yang, Y.H., and Speed, T.P. (2002). Design issues for cDNA microarray experiments. Nat. Rev. Genet. **3,** 579–588.

A Discussion of Statistical Methods for Design and Analysis of Microarray Experiments for Plant Scientists

Dan Nettleton

Plant Cell 2006;18;2112-2121
DOI 10.1105/tpc.106.041616