Chapter 9

Challenges and Approaches to Statistical Design and Inference in High-Dimensional Investigations

Gary L. Gadbury, Karen A. Garrett, and David B. Allison

Abstract

Advances in modern technologies have facilitated high-dimensional experiments (HDEs) that generate tremendous amounts of genomic, proteomic, and other "omic" data. HDEs involving whole-genome sequences and polymorphisms, expression levels of genes, protein abundance measurements, and combinations thereof have become a vanguard for new analytic approaches to the analysis of HDE data. Such situations demand creative approaches to the processes of statistical inference, estimation, prediction, classification, and study design. The novel and challenging biological questions asked from HDE data have resulted in many specialized analytic techniques being developed. This chapter discusses some of the unique statistical challenges facing investigators studying high-dimensional biology and describes some approaches being developed by statistical scientists. We have included some focus on the increasing interest in questions involving testing multiple propositions simultaneously, appropriate inferential indicators for the types of questions biologists are interested in, and the need for replication of results across independent studies, investigators, and settings. A key consideration inherent throughout is the challenge in providing methods that a statistician judges to be sound and a biologist finds informative.

Key words: FDR, genomics, high-dimensional, microarray, multiple testing, statistics.

1. Introduction

The present genomic era (1) has ushered in new challenges in high-dimensional study design and analysis. Draft sequences of several genomes coupled with new technologies allow study of entire genomes rather than isolated single genes. Questions from such high-dimensional investigations involve multiplicity at unprecedented scales. These questions may involve thousands of

Dmitry A. Belostotsky (ed.), *Plant Systems Biology, vol. 553*

DOI 10.1007/978-1-60327-563-7_9 Springerprotocols.com

genetic polymorphisms, gene expression levels, protein measurements, genetic sequences, or any combination of these and their interactions.

This chapter is targeted to statisticians, biologists, and those whose expertise bridges the interface. Questions that biologists want to ask from high-dimensional experiment (HDE) data require novel analytic approaches. It is important that the statistical methods applied to HDE data are aimed at drawing inferences biologists are interested in and also that these analytic methods have sound statistical foundations. There is now a relatively large and quickly growing body of statistical literature on the design of HDEs and on the analysis of resulting data from HDEs. Here we summarize some key methodological developments from statisticians as they pertain to HDEs. Included is some review of statistical foundations related to the interpretation of statistical evidence (e.g., a *P*-value), sampling variability, and aspects of a study design.

In the next section, we discuss design, analysis, and inference in the context of a single gene (i.e., a single hypothesis test). An example of a microarray experiment is used to illustrate the ideas. In **Section 3** we extend the discussion to high-dimensional studies where many hypotheses are simultaneously investigated. **Section 4** focuses more on the false discovery rate (FDR) (2) and related quantities that have garnered increased interest when analyzing high-dimensional data. **Section 5** discusses some other topics related to HDEs.

2. Statistical Inference for a Single Gene

A variable, Y, will be used to denote the information of interest in an HDE. In a microarray experiment, Y will be a measure of genetic expression after perhaps background correction, normalization, or transformation. These latter pre-processing choices are generally determined by the technology used in the experiment, potential biases induced in the measurement due to factors involved in the experiment, and characteristics regarding the statistical distribution of Y. In the discussion that follows, Y will be the genetic expression after pre-processing and, in this section, we will consider the analysis for differential expression for a single gene. Later, the issues encountered in high-dimensional settings will be discussed. For some references on pre-processing of gene expression data, see (3–6).

2.1. Discussion of Designs

Consider an experiment where there is one treatment factor with T levels. The goal is to determine if a gene is differentially expressed across the levels of the treatment. In many earlier studies

$T = 2$ and the two levels were a test treatment versus a control treatment. Travers et al. (7), for example, compared gene expression in plants experiencing ambient precipitation patterns and plants that experienced precipitation altered following a pattern predicted by models of climate change in the United States Great Plains. More generally, T can be greater than 2, such as a set of T different precipitation patterns, a set of T plant genotypes, an infection by T plant pathogens or combinations of plant pathogens, or planting in a set of T different soil types.

In a completely randomized design comparing T levels of a treatment, a total of N samples are randomly divided into groups of size $n_1, n_2, \ldots, n_T$ with each group receiving one of the levels of the treatment. In such a design the number of possible treatment assignments is

$$C = \frac{N!}{n_1! n_2! \ldots n_T!}, \qquad [1]$$

where $N = \sum_i n_i$. Observed data are represented by Y_{ij}, i.e., the expression of a gene for the jth sample in the ith treatment group where $i = 1, \ldots, T, j = 1, \ldots, n_i$, and n_i is the number of samples assigned to the ith treatment group. The one-way analysis of variance (ANOVA) model that is often used to model the data is of the form, $Y_{ij} = \mu_i + \varepsilon_{ij}$, where μ_i is the population mean response of genetic expression for samples exposed to the ith treatment level and ε_{ij} is a random error term. In a hypothesis test to determine if there is any mean differential expression due to the treatment, the ANOVA model above is compared to a reduced null model $Y_{ij} = \mu + \varepsilon_{ij}$. The null hypothesis that all means are equal, Ho: $\mu_1 = \mu_2 = \cdots = \mu_T$, is tested against an alternative that there is at least one difference between means, stated as Ho: $\mu_i \neq \mu_{i'}$ for some $i \neq i'$. Tests of linear combinations of means may also be of interest. These are of the form, Ho: $\sum_{i=1}^{T} c_i \mu_i = 0$ versus an alternative Ha: $\sum_{i=1}^{T} c_i \mu_i \neq 0$ where $c_1, \ldots, c_T$ are constants chosen by the investigator, e.g., $c_1 = 1, c_2 = -1$ and all other constants equal to 0 would be a test of $\mu_1 - \mu_2$.

More complex designs may be needed in contexts such as ecological studies, where the design may be influenced by conditions in the field where samples are obtained. In the Travers et al. (7) example, the emphasis of the analysis was on comparing gene expression for plants in replicate plots experiencing ambient precipitation patterns and plants in replicate plots experiencing precipitation patterns altered to follow a climate change prediction. But this study was performed in a pre-existing field experiment that had its own experimental structure. The precipitation treatments were applied in the field in a randomized complete block design, where plants within a block were likely to be somewhat more similar genetically and to have somewhat more similar

interactions with other organisms. Time of day is also very important in determining levels of expression for some genes. Since collecting plant samples and preserving them for later processing is time-consuming, this model also included time of day as a predictor in a strip-plot design. This field study also included other treatment structures, such as a temperature treatment with two levels, ambient and increased, applied to subplots. Milliken et al. (8) address the issues involved in choosing among designs for pairing samples (in 2-dye experiments) collected from pre-existing split-plot experiments such as this one, where it may not be feasible to include comparisons of all treatment combinations and all dye combinations on the same microarrays. More discussion of microarray designs is given in (9).

Missing data can arise in microarray experiments and some designs are more robust to missing data than others. Field samples may be prone to missing data because of potentially more degraded tissue. Cross-species hybridization may also result in more missing data because of lower homology between species for genes that are less highly conserved. Consideration of potential sources of missing data in an HDE may aid in the choice of a design that minimizes some of the negative consequences of missing data while maintaining adequate statistical efficiency to detect effects that are of interest.

2.2. Statistical Tests

A statistical test involves a metric, say δ, that can be computed from observed data. This metric quantifies a departure from the null hypothesis and compares the size of this metric to what could have been observed by chance if the null hypothesis, *Ho*, were true. This assessment of chance is quantified by the *P*-value that is computed as a probability of observing a value of the metric as extreme (i.e., favoring the alternative *Ha*) or more extreme if *Ho* were true. Small *P*-values represent evidence in favor of *Ha*. *P*-values can be computed in two ways: under a random sampling framework or a random treatment assignment framework.

A random treatment assignment compares the observed metric with what would have been observed under different treatment assignments if *Ho* were true. Consider a two-sample completely randomized design ($T=2$) that is testing *Ho*: $\mu_1 - \mu_2 = 0$ versus a two-tailed alternative. One metric would be the usual estimate of $\mu_1 - \mu_2$, $\delta = \bar{y}_1. - \bar{y}_2$, where $\bar{y}_i. = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$, $i = 1, 2$. The statistical test is the usual Fisher randomization test (10). Under *Ho*, values of Y_{ij} are permuted across the assigned treatments resulting in C (equation [1], for $T=2$) values of $\bar{y}_1. - \bar{y}_2$. The proportion of these $\bar{y}_1. - \bar{y}_2$ that are further away from zero than the observed value of $\bar{y}_1. - \bar{y}_2$ is the randomization-based *P*-value. More details of this test, and the required assumptions, are in Mehta et al. (11).

The metric δ need not be a mean difference. It could be a standardized mean difference, i.e., the usual *t*-statistic or a modified *t*-statistic as described in (12). Pepe et al. (13) proposed a metric derived from receiver operating characteristic (ROC) curves. The Wilcoxon rank-sum test is a randomization test based on the ranks of gene expression. A limitation of randomization-based *P*-values is their discreteness in small samples (14, 15). For example, if $N=6$ and $n_1 = n_2 = 3$, there are only 10 possible two-tailed *P*-values resulting from the randomization test. This discreteness limits follow-up work that may involve ranking the most promising results in genetic expression studies and/or estimation of FDR.

Randomization tests can be extended for T treatment groups in a completely randomized design. The metric must quantify how different the group means are from an overall grand mean. One possibility is $\delta = \sum_{i=1}^{T} n_i(\bar{y}_i. - \bar{y}..)^2$ where $\bar{y}..$ is the mean of all N observations. The value of δ computed from observed data is then compared with the C possible values obtained by permuting the observed data across treatment groups. Other metrics are possible and the computation for a randomization test becomes more complex. Mielke and Berry (16) have details regarding permutation-type tests for more involved designs. When sample sizes are small, the discreteness issue that was present for two treatment groups is still present. When samples are larger, the number of possible randomizations becomes extremely large and computation of a *P*-value may require Monte Carlo approximation. Parametric tests can also be used to approximate a randomization-based *P*-value. The common parametric tests are the two-sample *t*-test for two treatment groups or the ANOVA-based *F*-test for multiple treatment groups, or two-way and higher order ANOVA designs when multiple treatments or blocking variables are used.

In a random sampling framework for comparing T levels of a treatment, the data for the expression of a particular gene are assumed to be a random sample from a larger population. For example, in an ecological study involving big bluestem plants, some samples might be drawn from a population of diseased big bluestem plants while other samples may be drawn from a population of non-diseased big bluestem plants. While big bluestem plants may be distributed through a large part of the United States, it may only be realistic to sample from a single state or even from a single prairie and assume that the sample is roughly representative of a larger target population that is of interest in the study. The resulting data are then assumed to be obtained from a statistical population model, that is, for the ith treatment group, $\Upsilon_{i1}, \Upsilon_{i2}, \ldots, \Upsilon_{in_i} \sim F_\Upsilon(y; \mu_i, \sigma_i)$ where F_Υ is some population

distribution that is used as a model for the response variable. If F_Y is a normal distribution or if sample sizes are large enough, then the appropriate statistical tests are the usual *F*-tests for ANOVA models and, in the case of just two treatments, the usual two-sample *t*-test.

For more complicated designs, mixed effects models for gene expression data (17) can include random effects. Blocks may be extremely important in biological studies, in the greenhouse, growth chamber, or in the field, and are often reasonably included as random effects. Many factors influencing gene expression are not yet understood but may be dealt with to some extent by blocking. Blocking may be done across space, across time, and across individual scientists working with samples. Other treatments of more direct interest may also be included as random effects. For example, if expression in multiple genotypes is compared and the genotypes are randomly selected, it would be reasonable to treat genotype as a random effect.

Regardless of the design and model used to analyze resulting data, ultimately some hypothesis will be of interest such as determining if a gene is differentially expressed across two or more treatment conditions, or testing a contrast in a more complex model. If required assumptions are met regarding the distribution of data, the appropriate test results in a "valid" *P*-value.

2.3. Discussion of P-values

Exploiting the properties of a *P*-value, as a random variable (18), has recently become popular in methods that analyze high-dimensional data (19–21), though the idea goes back further (22). A key feature of a valid *P*-value is that its distribution, when the null hypothesis is true, is uniform on the interval from 0 to 1. This leads to a convenient interpretation of a *P*-value as a measure of the making of a type I error when rejecting a null hypothesis (i.e., rejecting a true null hypothesis). What a *P*-value is and what it is not are best illustrated with simple probability statements. Let $\{Ho\}$ be the event that the null hypothesis is true and $\{\overline{Ho}\}$ the event that it is false, and suppose that a null hypothesis will be rejected if a *P*-value is less than some threshold, τ. Denote a *P*-value, as a random variable, as *P*. Then $\Pr[P \leq \tau | \{Ho\}] = \tau$, that is, the probability of rejecting a true null hypothesis is equal to the threshold at which it was rejected. This is a probability of committing a type I error. Replacing the threshold with the actual observed *P*-value from a test then allows the *P*-value to be interpreted as the chance of a type I error.

A small *P*-value is often interpreted as evidence against the null hypothesis and is, thus, often misinterpreted. Berger and Sellke (23) discuss this and show some examples where a *P*-value is equal to 0.05, but the probability that the null hypothesis is false, given the data, is closer to 0.50. However, it is this latter probability that

is more intuitive to investigators (11). Stated as a probability this is $\Pr[\{Ho\}|P \leq \tau]$, the probability the null hypothesis is true given that a *P*-value falls below a given threshold. Interpretation in a high-dimensional setting is as follows: if null hypotheses are to be rejected when the corresponding *P*-values are less than or equal to τ, $\Pr[\{Ho\}|P \leq \tau]$ is an expected proportion of those "discoveries" that are false. An issue in computing this probability is that a prior probability, $\Pr[\{Ho\}]$, is needed. Computing this prior probability from *P*-values obtained from multiple hypothesis tests has been the focus of many methods that analyze high-dimensional data (cf., 20, 24, 25).

3. Statistical Inference in High-Dimensional Experiments

3.1. Multiple Test Statistics and Multiple P-Values

In a high-dimensional experiment there are, say K, observations per sampled unit and data from a completely randomized design comparing T levels of a treatment are of the form $\Upsilon_{ij} = \left(\Upsilon_{1ij}, \ldots, \Upsilon_{Kij}\right)'$ for the jth sample in the ith treatment group. Randomization-based inference follows as discussed in **Section 2** except the entire vector of observations for the jth unit is permuted across treatment conditions, $i = 1, \ldots, T$. In the random sampling framework the sample for the ith treatment group is $\Upsilon_{i1}, \ldots, \Upsilon_{in_i} \sim F_\Upsilon(y; \mu_i, \Sigma_i)$, where μ_i is a K-dimensional vector of the mean expression for each gene represented on the arrays, and Σ_i is a $K \times K$ variance–covariance matrix. The earlier discussions for testing a single gene for differential expression across the T treatment conditions would, with this multivariate structure, now be a test on a marginal distribution for a single gene; there are K marginal distributions, one for each gene. The result from gene-specific tests is a distribution of K test statistics or *P*-values. The most "interesting" genes are determined by a ranking procedure or assignment of a posterior probability of being differentially expressed, i.e., using the notation from **Section 2** this would be $\Pr\left[\left\{\overline{Ho}\right\}|P \leq \tau\right]$ as defined in (20).

Figure 9.1 illustrates two distributions of *P*-values obtained from an experiment that evaluated the effect of two treatments, drought stress and infection by a rust fungus, in a factorial design (26). The drought treatment levels consisted of the presence or absence of drought stress. The pathogen treatment levels consisted of presence or absence of rust infection. The distribution of *P*-values for a drought effect shows a stronger "signal" than that for a rust effect because more *P*-values seem to be clustering toward zero than would be expected under a global null hypothesis, i.e., a "global null hypothesis" that there were no genes

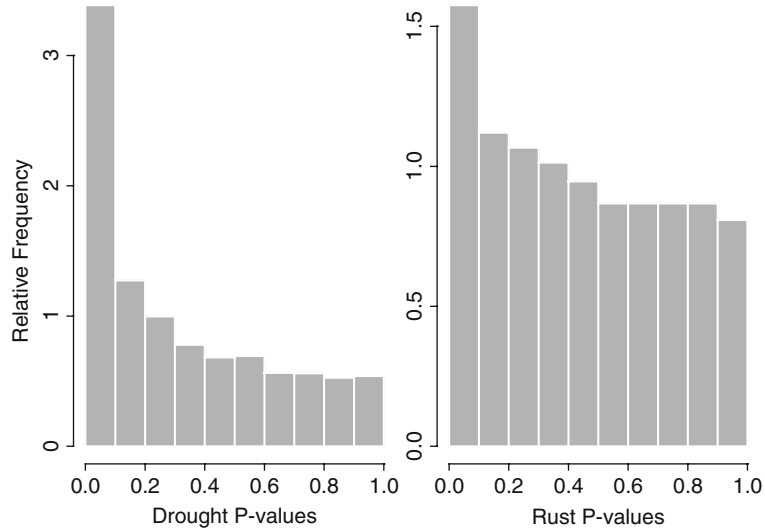

Fig. 9.1. Distribution of *P*-values from tests for differential expression due to a drought effect *(left panel)* and a rust effect *(right panel)*.

differentially expressed. If a global null hypothesis were true, one would expect the histogram of *P*-values to be relatively flat on the interval from 0 to 1. If genes were to be declared "statistically significant" if a *P*-value is less than the threshold $\tau = 0.01$, then an estimate of the "true-positive" probability $\Pr\left[\left\{\overline{Ho}\right\} | P \leq \tau\right]$ is 0.952 for a drought effect and 0.673 for a rust effect. These probabilities were called true-positive probabilities in (20), and the method reported there was used in the computations here. Subtracting this probability from 1 is related to the false discovery rate to be discussed in a later section. So of these genes that are declared significant for a drought effect, most are expected to be true discoveries, but only a little over two-thirds are expected to be true discoveries when looking at a rust effect. Lowering the threshold will increase this true-positive probability but at a cost of a smaller set of genes with *P*-values below the threshold.

3.2. Sampling Variability and Replication

Sampling variability in HDEs can arise from multiple sources (27, 28). A figure in Gadbury et al. (29, p. 81) and accompanying discussion illustrated sources of variability affecting a distribution of *P*-values. Technical variability of pixel effects on a spot was discussed by Brody et al. (30), and other design issues affecting technical variability have been considered by others, for example (31). Whether to spot one gene multiple times on a single microarray or to have repeated microarrays for a single tissue sample are aspects of assessing technical variability within and across arrays (32). Ultimately, statistical inference generally is targeted to some defined population of organisms and it is biological variability that

is of primary interest and is, in fact, essential for drawing valid inferences to a larger population of organisms (33). If the cost of obtaining replicate biological samples is not large versus that of obtaining a measurement (i.e., running a microarray), then there are design advantages of obtaining biological replicates versus expending resources on repeated measurements (9). Moreover, increasing sample size (number of biological samples or replicates) can increase the true-positive probability discussed above and increase the chances of discovering true results in an HDE, i.e., a higher expected discovery rate (34).

Hereafter, replication in an HDE will refer to biological replication, i.e., distinct tissue samples that are appropriately considered replicates in the context of the experiment being performed (35). In some microarray experiments, a tissue sample will correspond to a microarray or to a dye on a microarray in the case of dye swap experiments. In randomization tests or resampling procedures such as the bootstrap (36), the biological tissue or the microarray represented by the data vector $\Upsilon_{ij} = \left(\Upsilon_{1ij}, \ldots, \Upsilon_{Kij}\right)'$ is the unit of randomization or resampling. As mentioned earlier, randomization tests can produce coarse distributions of test statistics (and, hence, *P*-values), making it impossible to identify a list of the most promising candidate genes. It is tempting, therefore, to take advantage of the large number of genes and permute or resample genes themselves. However, genes are not exchangeable and variance of gene expression values is not homogeneous across genes. Methods involving mixed effects models and/or empirical Bayesian methods involving variance shrinkage have been proposed to address inferential issues associated with unequal variances across genes (17, 37, 38).

Gene-specific hypothesis tests are often carried out for each gene and variance estimates are computed for each gene. Correlation structure among measurements on a tissue sample (e.g., for co-regulation of certain sets of genes in a microarray experiment) leads to correlated *P*-values from multiple hypothesis tests, and this correlation structure cannot be estimated from observed data due to the high dimensionality. Yet this correlation can increase sampling variability leading to increased variance of estimates obtained from an HDE such as the true-positive probability defined in **Section 3.1**.

3.3. An Illustration of Correlated Tests

In an HDE, there are quantities of interest to the investigator that summarize results over thousands of tests. One quantity is the number of *P*-values below a threshold given that the global null hypothesis is true. Another is the proportion of all hypotheses tests for which the null hypothesis is true. Yet another is the true-positive probability discussed above or the analogous quantity, the false discovery rate. Many methods that estimate these

quantities may perform well, on average, but some estimates that are produced can have high variance when there is correlation of gene expression values leading to correlated *P*-values (39–41).

Figure 9.2 shows the effect of correlation on the sampling variability in a distribution of 10,000 *P*-values when the global null hypothesis is true. The data that would have produced the distributions of *P*-values would correspond to a situation where there was no mean difference in expression across two treatment groups for any genes. The test statistics that produced the *P*-values were standard normal in all four plots. However, all test statistics were independent in **Fig. 9.2(A)** but were correlated in **Fig. 9.2(B)–(D)** by a correlation matrix, Σ. In **Fig. 9.2(A)**, Σ was the identity matrix meaning that all tests were independent and that resulting *P*-values were uncorrelated. The histogram of *P*-values is nearly uniform, as would be expected. Repeated sampling from this model and computing a distribution of *P*-values will result in plots resembling that in **Fig. 9.2(A)**. In the other parts of **Fig. 9.2**, Σ was block diagonal where 20 blocks of size 500 were used. Correlation between all pairs of genes within a block was set to 0.5 and correlation of genes in different blocks was 0.

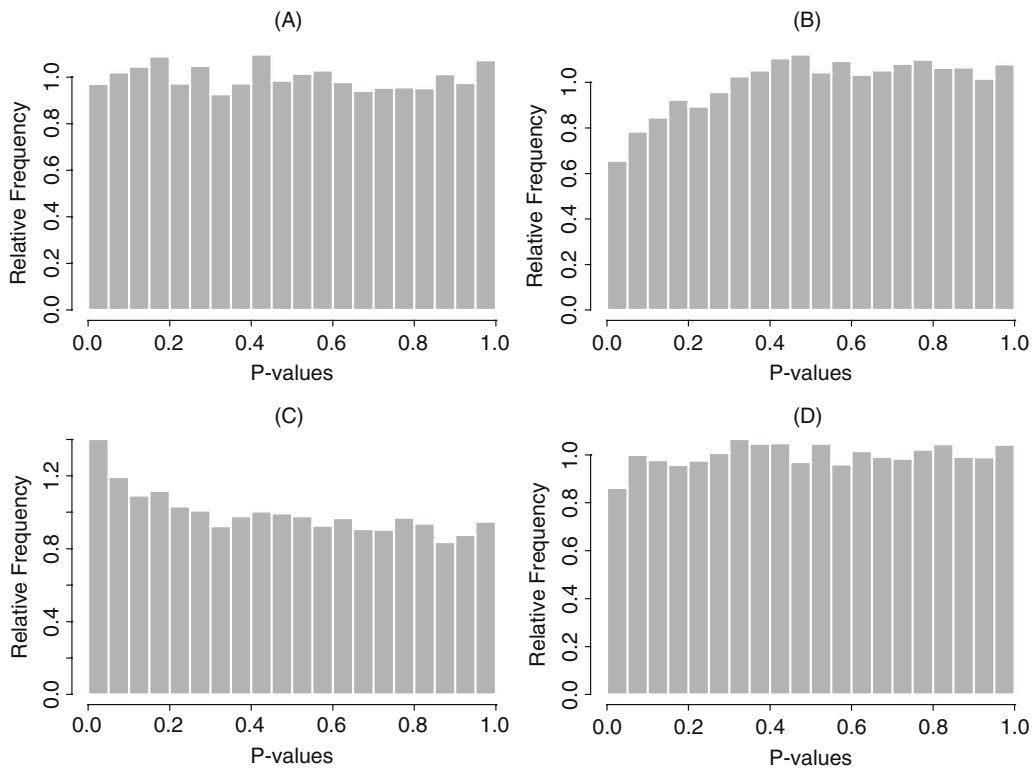

Fig. 9.2. 10,000 *P*-values under the global null hypothesis. *P*-values are uncorrelated in (**A**) but correlated in (**B**)–(**D**) using 20 blocks of size 500 equicorrelation matrices where the common correlation is 0.5.

Figure 9.2(B)–(D) shows three different distributions of *P*-values computed from three simulations of data from this model. **Figure 9.2(D)** looks similar to **Fig. 9.2(A)** where genes are independent. However, **Fig. 9.2(B)** and **(C)** shows how patterns in the distribution can arise due to sampling variability, even though the global null hypothesis is true.

A simple statistic based on a distribution of *P*-values is the number of *P*-values below a given threshold, say 0.01. Let $N(p = 0.01)$ represent this number. Since there are 10,000 *P*-values and the global null hypothesis is true, we expect $N(p = 0.01)$ to be, on average, 100 regardless of the correlation structure. In the individual samples in **Fig. 9.2**, it is equal to 90, 36, 145, and 86 in parts **(A)–(D)**, respectively. $N(p = 0.01)$ is expected to be 100 in all of the plots but the standard deviation is 9.95 in part (A) and 67.48 in the other plots. A technique for deriving the standard deviation for this statistic was outlined in (22) with further details given in (42). The standard deviation of $N(p = 0.01)$, as a statistic, increases by a factor of 6 due to the correlation structure. What correlation structure is reasonable in a genetic expression study may depend on the organism and application. The one illustrated here might be rather extreme. The standard deviation of $N(p = 0.01)$ will decrease as the block size for correlated data decreases and/or as the value of the correlation decreases toward zero.

The performance of statistical methods for estimating quantities of interest in HDEs has typically been evaluating using simulations that include simulating data with various correlation structures (14, 20, 43). The key point here is that a weak signal in a distribution of *P*-values may be due to some genes differentially expressed or that the effect of correlation on sampling variability is producing the observed signal. It is unlikely that any correlation could produce a strong signal like that in **Fig. 9.1** for a drought effect.

Recent papers have appeared that deal in more detail on correlation structure among genes and the effect that it might have on conclusions from a study (40, 44). A topic discussed later and one of active research interest (33) is gene class testing where certain classes of genes, rather than individual genes, are tested for differential expression. The methodological issues that arise with these tests and their ability to overcome some of the issues associated with testing single genes are beginning to be investigated (45).

3.4. Multiple Testing in High-Dimensional Experiments

One of the topics given the most attention has been the issue of multiple testing in HDE settings. The multiple comparisons problem becomes especially acute when thousands of tests are being conducted simultaneously and one wants to guard against type I errors, i.e., rejecting a true null hypothesis. For example, as discussed in the illustration above using **Fig. 9.2** where there are 10,000 tests for which the null hypothesis is true, one would

expect to find 100 "statistically significant" results at a threshold of 0.01 and 500 at the threshold of 0.05. These numbers are the number of type I errors that would be committed if declaring a test significant at the two respective thresholds. One obvious technique to control the number of type I errors is to lower the threshold at which significance is declared. Development of statistical methods to control for the number or proportion of type I errors in multiple testing situations is its own area of research with entire texts devoted to the topic, for example (46).

A common method that controls for the probability of a single type I error is the Bonferroni adjustment. Suppose that only one test were to be conducted and statistical significance is set to be at a level 0.05, so that a *P*-value below this number is significant. When there are K tests being conducted simultaneously, a single test is declared significant, using a Bonferroni adjustment, at a *P*-value below $0.05/K$. The probability of one or more type I errors among all K tests is then less than or equal to 0.05. In each of the *P*-value distributions in **Fig. 9.2** where the global null hypothesis was true, K =10,000 and in all four cases there were no *P*-values below $0.05/K$, so no type I errors would be committed using a Bonferroni adjustment. Thus this adjustment did what it was supposed to do.

Now consider the two distributions of *P*-values shown in **Fig. 9.1** where there appears to be a signal in each. There are $K = 7550$ *P*-values representing a drought effect and 7471 representing a rust effect. For the drought effect, there are only 14 *P*-values below $0.05/K$ and zero below this for the rust effect. With a Bonferroni adjustment one would find 14 statistically significant results for a drought effect and none for a rust effect. The adjustment is extremely conservative and there are very likely many true findings that are being missed, i.e., many type II errors. In fact, the method in (20) estimates that around 46% of the null hypotheses are false for a drought effect and around 16% for a rust effect. Many of the modern methods for HDE data seek a balance between controlling for a certain proportion of type I errors and detecting truly significant results out of thousands of possible tests. Many of these methods are focused on the false discovery rate (FDR) first discussed by Benjamini and Hochberg (2).

4. The False Discovery Rate and Related Quantities in High-Dimensional Experiments

As stated earlier, FDR is similar to one minus the true-positive probability discussed earlier. Much work in statistical methods development has focused on a mathematical definition of FDR and methods to either bound it or estimate it. Many of these

methods work on the distribution of *P*-values from multiple tests, so herein we discuss FDR in this context. Stated in words, FDR is an expected proportion of hypothesis tests that are declared statistically significant, but that are false discoveries, i.e., the null hypothesis is actually true. **Table 9.1** shows quantities of interest in an HDE where there are a total of *K* hypothesis tests.

Table 9.1
Quantities of interest in an HDE. The total number of tests is equal to *K*. The row totals are known but column totals are not, nor are the individual values *A, B, C, D*

	Ho true	Ho false	Total
Tests that are *not* declared significant	A	B	$K-R$
Test that are declared significant	C	D	R
Total	M	$K-M$	K

4.1. Definitions of the False Discovery Rate

There are two approaches to using FDR in an HDE. One is to specify a desired FDR (or an upper bound for it) and select a threshold for statistical significance based on this desired upper limit. Another is to specify a threshold (i.e., significance level for a *P*-value) at which a hypothesis test will be declared significant, and then estimate the FDR at that threshold. We discuss the latter and show some ways that FDR can be estimated at a given threshold for significance.

In **Table 9.1**, the row totals are known. Once a threshold, τ, is set by the investigator, R is the number of *P*-values below that threshold. The number C is unknown and this is the number of false discoveries out of the total R rejected null hypotheses. The quantity C/R is the proportion of false discoveries. We can also note other quantities such as $B/(K–R)$ which is the proportion of null hypotheses that are false but that were not detected in the test (i.e., the *P*-value was above τ).

The proportion C/R is an unknown quantity from an HDE. FDR is defined with respect to this proportion as a parameter in an HDE for which estimates can be derived. Benjamini and Hochberg (2) defined *FDR* as follows:

$$FDR = E\left[C/R \;\; I_{\{R>0\}}\right] = E\left[C/R \middle| R>0\right] P(R>0), \quad [2]$$

where $I_{\{R>0\}}$ is an indicator function equal to 1 if $R>0$ and zero otherwise, and where $E()$ is an expectation operator representing a population average. Storey (19) defined the positive *FDR* as

$$pFDR = E\left[C/R \middle| R>0\right]. \quad [3]$$

Since $P(R>0) \geq 1-(1-\tau)^K$, and since K is usually very large, $FDR \approx pFDR$. For example, for $K=10{,}000$, $P(R>0) \geq 0.99995$ when $\tau = 0.001$ so we do not distinguish between FDR and $pFDR$ as the parameter being estimated and simply refer to it as FDR with estimates denoted by $\widehat{FDR}$. In fact there are other versions of FDR that have been defined that differ in the way the expectation is taken on the ratio C/R. Other examples are the marginal FDR, the empirical FDR, and the conditional FDR, but in many cases these different versions of FDR are numerically close with some being equivalent under certain conditions (47).

4.2. Estimating the False Discovery Rate and Related Quantities

The proportion M/K is the proportion of true null hypotheses among all K tests, a quantity that is unknown in an HDE and must be estimated. An estimate of this proportion (or an upper bound for it) is needed in order to produce an estimate of FDR, and many methods have produced estimates for this proportion (e.g., (20, 48, 49)). Let $\pi_0 = M/K$ and an estimate of this as $\hat{\pi}_0$, and define $P_R = R/K$, the proportion of rejected null hypotheses at a threshold τ, and note that P_R is a known quantity in an HDE. There are two (at least) basic techniques that are used to estimate FDR. One set of techniques produce an estimate of π_0 and then estimate FDR at a selected threshold τ using,

$$\widehat{FDR} = \frac{\tau \hat{\pi}_0}{P_R} \qquad [4]$$

These methods differ in how $\hat{\pi}_0$ is obtained with many methods focused on producing a conservative estimate. Clearly, $\hat{\pi}_0 = 1$ would be the most conservative and, if the distribution of P-values from multiple tests looks like that shown in **Fig. 9.2(A)**, then perhaps π_0 is close to 1. However, if distributions look like those in **Fig. 9.1**, then π_0 should be less than 1. Many methods that estimate π_0 use algorithms that assess how much the distribution of P-values deviates from a uniform distribution like in **Fig. 9.2(A)**.

Another set of techniques uses a mixture model framework to produce estimates of FDR. The mixture model (usually a two-component mixture) approach on a distribution of P-values uses a model of the form

$$F(p;\pi_0,\theta) = \pi_0 F_0(p) + (1-\pi_0)F_1(p), \qquad [5]$$

where F is a cumulative distribution function (CDF), p a P-value, F_0 a distribution of a P-value under the null hypothesis, F_1 a distribution of a P-value under the alternative hypothesis, π_0 is interpreted as before, and θ a (possibly vector) parameter of the distribution. Since valid P-values are assumed, F_0 is a uniform distribution. Estimating the components of the model in [5] yields estimates of FDR. The equation for FDR in a mixture model framework is

$$FDR = \frac{\pi_0 \tau}{\pi_0 \tau + (1 - \pi_0) F_1(\tau)} \quad [6]$$

Equation [6] has been defined as the positive FDR (19) but, as stated earlier, the different versions of FDR are close when K is large. Methods based on the mixture model framework differ in how the components of equation [6] are computed. Note that the only difference between equations [4] and [6] is the denominator. In [6], the denominator is the distribution function of a *P*-value and some have used a parametric form. Allison et al. (20) used a mixture of a uniform distribution and a beta distribution. The denominator in equation [4] is a version of the empirical distribution function which is a step function with increments of $1/K$ at each observed *P*-value. Another quantity called the local FDR (LFDR, (50)) can be directly defined from the mixture model in [5]. The definition is similar to [6] except that CDFs are replaced by the corresponding probability density function (pdf):

$$LFDR = \frac{\pi_0}{\pi_0 + (1 - \pi_0) f_1(\tau)}. \quad [7]$$

The interpretation of LFDR is the posterior probability that a test with a *P*-value equal to the threshold τ is a test for which the null hypothesis is true. As with FDR, LFDR will be smaller at smaller values of τ. Also, FDR can be thought of as a type of averaging of LFDR over all tests with a *P*-value $\leq \tau$ so, as a result, values of LFDR will be greater than FDR at a given τ. Estimates of FDR and LFDR are obtained in statistical methods by estimating the components in equations [6] and [7], respectively. Computing an estimate at thresholds equal to each observed *P*-value gives an FDR (LFDR) curve that is seen to be an increasing function of the *P*-values. **Figure 9.3** shows FDR and LFDR curves for the distribution of *P*-values shown for the drought effect in **Fig. 9.1**. The estimates were obtained using the mixture model method of Allison et al. (20). One can see that LFDR values are greater than FDR values. From the plot one can see (roughly) that for tests with a *P*-value smaller than 0.05, one would expect a proportion of around 0.10 false discoveries. For a test with a *P*-value equal to 0.05, one might estimate the posterior probability (LFDR) that the null hypothesis is actually true to be around 0.20. User-friendly software for fitting the mixture model of Allison et al. (20) and computing quantities based on the model was reported in Trivedi et al. (51) and is available at http://www.ssg.uab.edu/hdbstat.

In **Fig. 9.3** one can see that the FDR curve is a monotonically increasing function of the *P*-values. That is, FDR is smaller at smaller *P*-values. This does not necessarily happen when FDR is computed using equation [4] because the denominator is not a continuous function of the observed *P*-values. The FDR "curve" for the 100 smallest *P*-values obtained from the same data set is

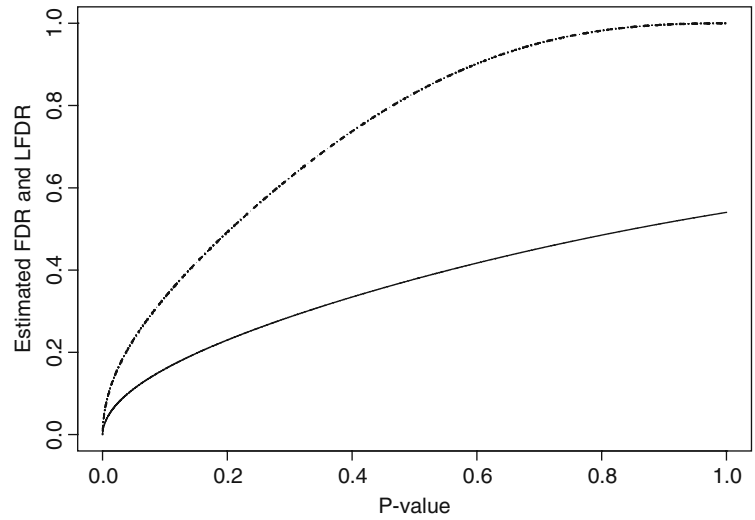

Fig. 9.3. FDR *(solid line)* and LFDR *(dashed line)* for the distribution of *P*-values in **Fig. 9.1** for the drought effect. Estimated quantities for the plots were obtained using the method in Allison et al. (20).

shown in the left panel of **Fig. 9.4**. There are some cases where a smaller *P*-value yields an increased estimate of FDR. Storey (52) defined the *Q*-value and interpreted it as a Bayesian posterior *P*-value, that is, it is a measure of the strength of an observed statistic (or *P*-value) with respect to the positive FDR. Estimates

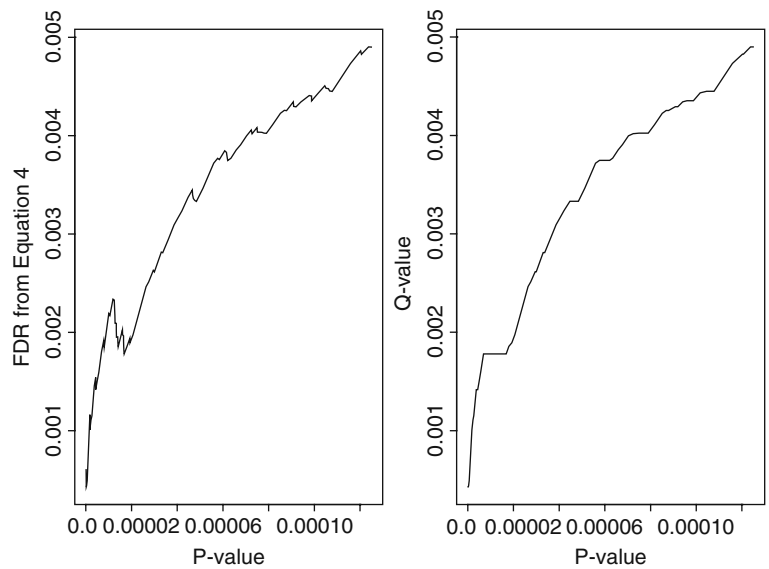

Fig. 9.4. The FDR for the smallest 100 *P*-values *(left panel)* using equation [4] and the *Q*-value *(right panel)* for the same *P*-values representing a drought effect (**Fig. 9.1**). The values for the plots were obtained using the smoothing method in Storey and Tibshirani (53).

of the *Q*-value from observed data should be monotonically increasing with the *P*-value. Storey (19) showed the algorithm to compute a *Q*-value from observed data and a plot of this *Q*-value is shown in the right panel of **Fig. 9.4**. A *Q*-value in **Fig. 9.4** computed at a given *P*-value is never larger at a smaller *P*-value. When there is no "signal" in a distribution of *P*-values (as seen in those in **Fig. 9.2**), the *Q*-value may remain large for all *P*-values, that is, for any list of tests that are rejected at a particular threshold, the proportion of false discoveries may be high. The software for computing *Q*-values is available as an R library called qvalue, available at www.r-project.org.

4.3. Sample Size Considerations for the False Discovery Rate and Related Quantities

Sometimes computed values of FDR can be large at even very small thresholds, and these large values may be due to the small sample sizes that are often common in HDEs. Gadbury et al. (34) presented a method to evaluate the role of sample size in bringing quantities like FDR to desired levels when the design is a comparison of two treatments. They also defined the expected discovery rate (EDR) which was the expected proportion of true alternative hypotheses that will be discovered in an HDE, i.e., the expected proportion $D/(K - M)$. Larger sample sizes yielded smaller values of FDR and larger values of EDR.

Recall from **Fig. 9.1** that the signal for a rust effect was not strong. Suppose that a two-treatment comparison study for differential expression due to a rust effect was being planned, and the signal present in **Fig. 9.1** for a rust effect was to be used as a pilot data set for planning sample size requirements for the new study. The *P*-values in **Fig. 9.1** were actually obtained from a two-factorial design structure, but for convenience and for purposes of illustration, we use this distribution as if it was obtained from a simple two treatment comparison study. **Figure 9.5** shows the technique reported in (34) that uses the distribution of *P*-values for a rust effect as a template but extrapolates EDR for various sample sizes and reports it for three different thresholds at which a null hypothesis is rejected. A smaller threshold yields a smaller EDR since fewer null hypotheses will be rejected; however, a smaller threshold yields a lower FDR because one is more certain that those null hypotheses that are rejected are true discoveries. One might notice that the EDR values in **Fig. 9.5** do not rise to the level of traditional power analyses in planning experiments. In HDEs there are many thousands of hypotheses being tested and an investigator might be content of discovering a smaller fraction of truly expressed genes for the purpose of follow-up research. A tool for implementing the method in (34) was reported in Page et al. (54) and is available online at www.poweratlas.org.

There are other results in the literature regarding sample size requirements in HDEs. Lee and Whitmore (55) investigated sample size requirements on type I and type II error probabilities.

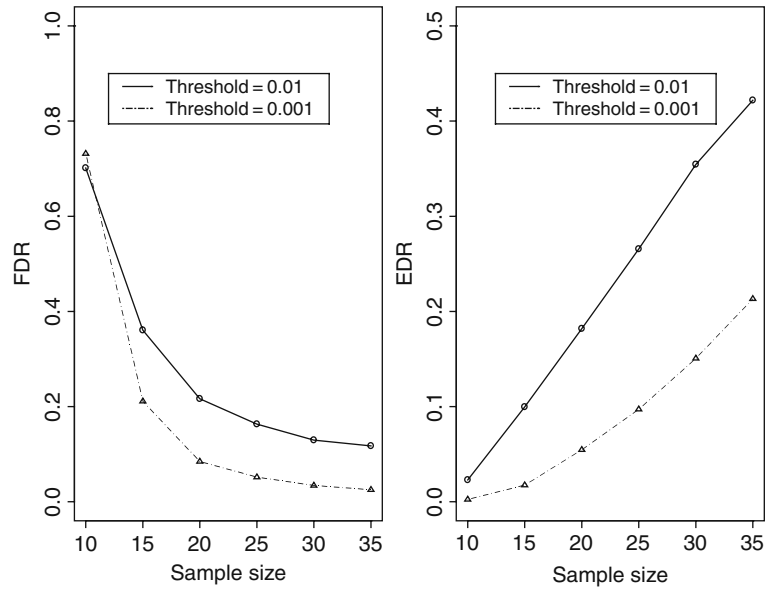

Fig. 9.5. Sample size analysis illustrated using *P*-values in **Fig. 9.1** for a rust effect as a pilot data set. Sample sizes reflect the number of microarrays in each of the two treatment groups in a two-treatment comparison study.

"Power" was equal to 1 – *P*(type II error) which is analogous to the *EDR* in Gadbury et al. (34). Lee and Whitmore (55) also extended their results to situations where there may be more than two treatment groups where interest is in determining differential expression among several treatment groups. Pan et al. (56) used a *t*-type statistic to quantify differential expression and presented a model that, when fitted to a "pilot" data set, could be used to assess the number of replicates required to achieve desired power at a given threshold. The fitted model was considered fixed, a type I error was specified, and power computed for any specified effect size, e.g., standardized difference in mean expression levels between two groups.

Zein et al. (57) considered sample size effects on pairwise comparisons of different groups and discussed the role of both technical and biological variabilities. Actual data sets were used to develop parameter specifications for simulated data sets. They used the term sensitivity as analogous to *EDR*, and specificity that is analogous to 1 – *FDR*. They evaluated the effect of varying sample sizes on these two quantities for various simulated data sets and using different types of statistical tests for differential expression, e.g., *t*-tests and a rank-based test. More recently Shao and Tseng (58) presented a sample size calculation with an adjustment for dependence among tests in microarray studies. More discussion of power and sample size in HDEs is in (29).

5. Classification and Validation Strategies and Some Remarks on Future Developments

Thus far we have discussed design and inference issues in HDEs and focused on the multiple testing issues when many thousands of tests are being conducted simultaneously. Here we conclude this chapter by discussing some other topics and techniques.

5.1. Clustering High-Dimensional Data

Clustering is one of the earlier techniques and has been and is still popular (33). Cluster analysis attempts to group data into classes based on some similarity metric. An early illustration of cluster analysis on data from an HDE is Eisen et al. (59). They used clustering on two time course gene expression data sets and showed that some genes of similar functions would cluster together. They also applied a type of randomization procedure to assess whether the clusters were real or whether they were an artifact of the clustering procedure. Even in HDE data that are completely random (i.e., data are generated so that there are not real clusters), a clustering routine will find clusters, so one cannot always be sure that a cluster is real without some technique to assess its repeatability in similar experiments. Some have referred to this as stability of clustering and have compared the stability of different cluster routines under different conditions (60).

Attempts to evaluate the stability of clustering techniques have generally used resampling techniques such as the bootstrap. Kerr and Churchill (61) assessed the reliability of conclusions obtained using clustering on data from microarray experiments. They used a clustering technique on a data set and assessed the stability of clusters on simulated data sets. The simulated data sets were created by fitting an ANOVA model to data and resampling residuals from the model in a bootstrap routine. Another resampling approach was used to evaluate the number of clusters present in an HDE data set where mixture models were used as a basis for clustering (62). Kapp and Tibshirani (63) assessed the reproducibility of clusters by defining a "cluster quality measure" that is related to prediction accuracy, that is, the ability of a new datum to be classified to a previously defined cluster.

To reflect the variability in an experiment due to biological samples, the resampling unit should be at the level of the biological specimen, i.e., a microarray in a microarray experiment (11). However, sometimes sample sizes are too small to use resampling at this level to evaluate the reliability of clusters. Garge et al. (60) conducted a simulation study of four clustering techniques and found that all four techniques produced low stability scores when evaluated on microarray types of data sets. Although obtaining reproducible clusters in an HDE with relatively small samples may have challenges, clustering methods can still be useful as an

exploratory method for obtaining a general description of how genes covary with respect to their gene expression levels (33). One key advantage to a reliable clustering of data in a gene expression experiment is a reduction of dimension from one of many thousands of genes to a dimension of a smaller number of clusters, providing the clusters contain meaningful information about the function or classification of certain sets of genes.

5.2. Gene Class Testing

One challenge when analyzing data from an HDE such as a gene expression experiment is finding ways to successfully interpret the enormous number of results that are obtained (64). A type of analysis has emerged that appears designed to help address this challenge. Such analyses recognize that genes can be and have been placed into a priori categories and they use this categorical information in analytic strategies that can reduce the number of results about individual genes to a smaller number of more interpretable findings concerning classes or families. Gene class testing is a relatively recent technique, with some methods for implementing it still in development (45).

Many methods use Gene Ontology (GO) terms for assignment of genes to classes, though other knowledge bases are available (*see* (65, 66) for discussion and illustration). The idea of gene class testing in an HDE is to identify classes or sets of genes that are differentially expressed across one or more treatment conditions, or that are associated with some phenotype. Pavlidis et al. (67) compared two computational methods for associating gene expression changes with age for selected sets of genes using GO classes. The two methods used different techniques to evaluate what GO classes are most associated with aging. Mootha et al. (68) presented a gene set enrichment analysis (GSEA) and illustrated its use on a gene expression study using human diabetic muscle. The technique used an enrichment score to quantify association of a gene set to a phenotypic class. The method was also illustrated on some cancer-related data sets that included leukemia and lung cancer (69). Goeman et al. (70) proposed their global test procedure and illustrated it on two examples. Their test produces one *P*-value for a group that is being tested. These are just a few of the methods that have been proposed for testing classes of genes for association with a phenotype or phenotypic class, e.g., "treatment" condition. Pan (71) proposed fitting mixture models to classes of genes and using these sub-mixture models within classes to determine differential expression, somewhat similar in concept to other approaches using mixture models that were fit to all genes, for example (20). A limitation of mixture models is that many measurements are usually needed to obtain a good fit (72) and, in this case, gene classes would need to be relatively large.

The fact that much recent development of statistical methods has occurred in gene class testing suggest its potential usefulness and promise as a tool for the analysis of HDE data. Many of the current methods have prompted some concerns (73) and others suffer from at least one flaw (33). Goeman and Buhlmann (45) review assumptions and limitations of some of the recent methods for gene class testing. Undoubtedly this area of research will continue as one of active interest.

5.3. Validating Methods Using Simulations

A general discussion of validity of findings in HDEs was given in Mehta et al. (74) and in Allison et al. (33). Here we discuss validity in the context of statistical methods and the results that they are designed to produce as was also done in Mehta et al. (11). Validity of results from an HDE data analysis depends on many of the same assumptions that are required for a valid analysis of data from a traditional experiment. Examples are assumptions about distributions of data (or residuals), the choice of the model used, and assumptions about random sampling and/or treatment assignment.

Many statistical methods that analyze data from HDEs produce conservative estimates of π_0 and FDR (i.e., estimates tend to be biased high). The properties of certain methods and the estimates that they produce can sometimes be evaluated using mathematical derivations and proofs. One example is Genovese and Wasserman (75) who looked at the FDR controlling procedure of Benjamini and Hochberg (2). The performance of many methods and comparisons of methods have been evaluated using computer simulation experiments.

One technique for simulating microarray data considered sources of variability in such data and created simulated data sets based on some knowledge of this variability gleaned from real data sets (27). Many methods simulate data sets using statistical distributions, often normal distributions (e.g., (49, 76)). Correlation structure, if considered at all, has been implemented in simulated data using a block-diagonal correlation matrix as was done, for example, in (20, 43).

Concern about how well-simulated data correspond to reality has generated interest in simulated data that are derived from actual data sets. Cattell and Jaspars (77) used the term plasmode to describe data that are constructed to reflect some aspect of reality. Mehta et al. (11) described a plasmode as a real (i.e., not computer-simulated but from actual biological specimens) data set for which some aspect of the truth is known. Plasmodes can be used to learn about the validity and lack of validity of certain statistical methods for microarray analysis. The great advantage of plasmodes is that, unlike with computer simulations, one need not question whether the particular distributions or correlations are realistic because they are taken directly from real data. Plasmodes are beginning to show promise as a valuable resource for the scientific community.

One example of a plasmode is a real microarray data set with specific mRNAs spiked-in (cf., (37, 78)). Evaluating whether a particular method can correctly detect the spiked mRNAs gives information about the method's ability to detect gene expression. Affycomp (37) is a set of tools and plasmode (spike-in) data sets on an integrated web site that allows investigators to analyze the same benchmark data sets using a new method.

Plasmodes could also be derived from a real data set in a manner for which some truth is known. Gadbury et al. (79) have explored techniques to do this in the context of a microarray experiment for a two-treatment comparison study. A distribution of realistic "effect sizes" in an HDE can be obtained by analyzing a real data set. A data set for which the global null hypothesis is true (the null hypothesis is true for all tests) may be obtained by dividing the data for one treatment group into two pseudo-treatment groups. Differentially expressed genes can be created by sampling effect sizes from the experiment and incorporating them into the data for one of the pseudo-treatment groups for a proportion $1 - \pi_0$ of genes. In the resulting plasmode data set a true value of π_0 and a true value of FDR at a particular threshold can be known. Methods can then be evaluated on their ability to estimate these quantities.

5.4. Future Developments Related to High-Dimensional Experiments

Experiments investigating genome-wide gene expression may shift to greater use of sequencing in place of microarrays as sequencing becomes less expensive. In this case, rather than evaluating expression for the set of genes represented on a microarray, any genes expressed may be analyzed by determining the frequency of occurrence of corresponding RNA in samples. This may make it easier to discover new genes that are differentially expressed, but it may also make it more difficult to study genes with low levels of expression. One difference for statistical analyses will be that only a certain (large) number of sequences can be evaluated from each sample, so a higher frequency of sequences corresponding to one gene will be associated with a lower frequency of sequences corresponding to other genes.

Proteomics, lipidomics, and metabolomics are becoming more approachable for more plant systems. The data sets generated in these new "Omics" fields may often be modeled using approaches similar to those for studies of genome-wide gene expression (transcriptomics). Ultimately a new challenge for statistics will be the development of good comparisons of responses to treatments across these different types of data sets. Biological questions may be answered using different sets of findings, possibly from different Omics experiments. As noted in Allison et al. (33), how best to examine intersections between sets of findings is a needed area of research as is how to evaluate complex multi-component hypotheses. Bayesian approaches might be helpful in these areas.

Acknowledgments

D. Allison and G. Gadbury acknowledge the support from NIH Grant U54 CA100949. K. Garrett acknowledges the support from NSF Grants DBI-0421427, DEB-0516046, and EF-0525712, and DOE Grant DE-FG02-04ER63892. Gadbury and Garrett are grateful to programs supporting research from the Ecological Genomics Institute, Kansas State University, and the NSF Long Term Ecological Research Program at Konza Prairie, Kansas, and the Kansas Agricultural Experiment Station (Contribution No. 09-118-B).

References

1. Wolfsberg, T.G., Wetterstrand, K.A., Guyer, M.S., Collins, F.S., and Baxevanis, A.D. (2002) A user's guide to the human genome. *Nature Genetics Supplement* **32**, 1–79.
2. Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**, 289–300.
3. Schadt, E.E., Li, C., Ellis, B., and Wong, W.H. (2001) Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry, Supplement* **37**, 120–125.
4. Quackenbush, J. (2002) Microarray data normalization and transformation. *Nature Genetics* **32**, 496–501.
5. Smyth, G.K. and Speed, T. (2003) Normalization of cDNA microarray data. *Methods* **31**, 265–273.
6. Ekstrom, C.T., Bak, S., Kristensen, C., and Rudemo, M. (2004) Spot shape modelling and data transformations for microarrays. *Bioinformatics* **20**, 2270–2278.
7. Travers, S.E., Smith, M.D., Bai, J.F., Hulbert, S.H., Leach, J.E., Schnable, P.S., Knapp, A.K., Milliken, G.A., Fay, P.A., Saleh, A., and Garrett, K.A. (2007) Ecological genomics: making the leap from model systems in the lab to native populations in the field. *Frontiers in Ecology and the Environment* **5**, 19–24.
8. Milliken, G.A., Garrett, K.A., and Travers, S.E. (2007) Experimental design for two-color microarrays applied in a pre-existing split-plot experiment. *Statistical Applications in Genetics and Molecular Biology* **6**, Article 20.
9. Kerr, M.K. (2003) Design considerations for efficient and effective microarray studies. *Biometrics* **59**, 822–828.
10. Fisher, R.A. (1966) The Design of Experiments, 8th edition. Hafner Publishing Company: New York.
11. Mehta, T.S., Zakharkin, S.O., Gadbury, G.L., and Allison, D.B. (2006) Epistemological issues in omics and high-dimensional biology: give the people what they want. *Physiological Genomics* **28**, 24–32.
12. Cui, X. and Churchill, G.A. (2003) Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology* **4**, 21.
13. Pepe, M.S., Longton, G., Anderson, G.L., and Schummer, M. (2003) Selecting differentially expressed genes from microarray experiments. *Biometrics* **59**, 133–142.
14. Gadbury, G.L., Page, G.P., Heo, M., Mountz, J.D., and Allison, D.B. (2003) Randomization tests for small samples: an application for genetic expression data. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* **52**, 365–76.
15. Xu, R. and Li, X. (2003) A comparison of parametric versus permutation methods with applications to general and temporal microarray gene expression data. *Bioinformatics* **19**, 1284–1289.
16. Mielke, P.W. and Berry, K.J. (2007) Permutation Methods: A Distance Function Approach. Springer: New York.
17. 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. *Journal of Computational Biology* **8**, 625–663.

18. Sackrowitz, H. and Samuel-Cahn, E.P. (1999) P values as random variables—expected P values. *The American Statistician* **53**, 326–331.
19. Story, J.D. (2002) A direct approach to false discovery rates, *Journal of the Royal Statistical Society, Series B* **64**, 479–498.
20. Allison, D.B., Gadbury, G.L., Heo, M., Fernandez, J.R., Lee, C., Prolla, T.A., and Weindruch, R.A. (2002) Mixture model approach for the analysis of microarray gene expression data. *Computational Statistics and Data Analysis* **39**, 1–20.
21. Ruppert, D., Nettleton, D., and Hwang, J.T.G. (2007) Exploring the information in P-values for the analysis and planning of multiple-test experiments. *Biometrics* **63**, 487–495.
22. Schweder, T. and Spjøtvoll, E. (1982) Plots of P-values to evaluate many tests simultaneously. *Biometrika* **69**, 493–502.
23. Berger, J.O. and Sellke, T. (1987) Testing a point null hypothesis: The irreconcilability of P values and evidence. *Journal of the American Statistical Association* **82**, 112–122.
24. Broberg, P. (2004) A new estimate of the proportion unchanged genes in a microarray experiment. *Genome Biology* **5**, P10.
25. Langaas,M., Lindqvist, B.H., and Ferkingstad, E. (2005) Estimating the proportion of true null hypotheses, with application to DNA microarray data. *Journal of the Royal Statistical Society, Series B* **67**, 555–572.
26. Frank, E.E. (2007) The effects of drought and pathogen stress on gene expression and phytohormone concentrations in *Andropogon gerardii*. M.S. Thesis; Kansas State University: Manhattan, KS.
27. Singhal, S., Kyvernitis, C.G., Johnson, S.W., Kaisera, L.R., Leibman, M.N., and Albelda, S.M. (2003) Microarray data simulator for improved selection of differentially expressed genes. *Cancer Biology and Therapy* **2**, 383–391.
28. Zakharkin, S.O., Kim, K., Mehta, T., Chen, L., Barnes, S., Scheirer, K.E., Parrish, R.S., Allison, D.B., and Page, G.P. (2005) Sources of variation in Affymetrix microarray experiments. *BMC Bioinformatics* **29**, 214.
29. Gadbury, G.L., Xiang, Q., Edwards, J.W., Page, G.P., and Allison, D.B. (2006) The role of sample size on measures of uncertainty and power. In: Allison, D.B., Page, G.P., Beasley, T.M., Edwards, J.W., ed. DNA Microarrays and Related Genomics Techniques. Boca Raton: Chapman & Hall/CRC: 77–94.
30. Brody, J.P., Williams, B.A., Wold, B.J., and Quake, S.R. (2002) Significance and statistical errors in the analysis of DNA microarray data. *Proceedings of the National Academy of Sciences of the United States of America* **99**(20), 12975–12978.
31. Nguyen, D.V., Arpat, A.B., Wang, N., and Caroll, R.G. (2002) DNA microarray experiments: biological and technical aspects. *Biometrics* **58**, 701–717.
32. Rosa Guilherme, J.M., Steibel, J.P., and Tempelman, R.J. (2005) Reassessing design and analysis of two-colour microarray experiments using mixed effects models. *Comparative and Functional Genomics* **6**(3), 123–131.
33. Allison, D.B., Cui, X., Page, G.P., and Sabripour, M.(2006) Microarray data analysis: From disarray to consolidation and consensus. *Nature Review Genetics* **7**, 55–65.
34. Gadbury, G.L., Page, G.P., Edwards, J.W., Kayo, T., Prolla, T.A., Weindruch, R., Permana, P.A., Mountz, J., and Allison, D.B. (2004) Power analysis and sample size estimation in the age of high dimensional biology: a parametric bootstrap approach illustrated via microarray research. *Statistical Methods in Medical Research* **13**, 325–38.
35. Hurlbert, S.H. (1984) Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* **54**, 187–211.
36. Efron, B. and Tibshirani, R.J. (1993) An Introduction to the Bootstrap. Boca Raton, FL: CRC Press.
37. Irizarry, R.A., Wu, Z., and Jaffee, H.A. (2006) Comparison of Affymetrix GeneChip expression measures. *Bioinformatics* **22**, 789–794.
38. Ishwaran, H., Rao, J.S., and Kogalur, U.B. (2006) BAMarray: Java software for Bayesian analysis of variance for microarray data. *BMC Bioinformatics* **7**(1), 59.
39. Qiu, X., Klebanov, L., and Yakovlev, A. (2005) Correlation between gene expression levels and limitations of the empirical Bayes methodology for finding differentially expressed genes. *Statistical Applications in Genetics and Molecular Biology* **4**, Article 34.
40. Qiu, X., Xiao, Y., Gordon, A., and Yakovlev, A. (2006) Assessing stability of gene selection in microarray data analysis. *BMC Bioinformatics* **7**, 50.

41. Owen, A. (2005) Variance in the number of false discoveries. *Journal of the Royal Statistical Society, Series B* **67**, 411–426.
42. Hu, X. (2007) Distributional aspects of P-value and their use in multiple testing situations. Ph.D. Dissertation. University of Missouri – Rolla: Rolla, Missouri.
43. Nettleton, D., Hwang, G.J.T., Caldro, R.A., and Wise, R.P. (2006) Estimating the number of true null hypotheses from a histogram of p-values. *Journal of Agricultural, Biological, and Environmental Statistics* **11**, 337–356.
44. Efron, B. (2007) Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association* **102**, 93–103.
45. Goeman, J.J. and Buhlmann, P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* **23**, 980–987.
46. Hochberg, Y., and Tamhane, A.C. (1987) Multiple Comparisons Procedures. New York: John Wiley & Sons, Inc.
47. Tsai, C., Hsueh, H., and Chen, J.J. (2003) Estimation of false discovery rates in multiple testing: application to gene microarray data. *Biometrics* **59**, 1071–1081.
48. Pounds, S. and Morris, S.W. (2003) Estimating the occurrence of false positive and false negative in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics* **19**(10), 1236–1242.
49. Nguyen, D. (2004) On estimating the proportion of true null hypotheses for false discovery rate controlling procedures in exploratory DNA microarray studies. *Computational Statistics & Data Analysis* **47**, 611–637.
50. Efron, B. (2004) Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association* **99**, 96–104.
51. Trivedi, P., Edwards, J.W., Wang, J., Gadbury, G.L., Srinivasasainagendra, V., Zakharkin, S.O., Kim, K., Mehta, T., Brand, J.P.L., Patki, A., Page, G.P., and Allison, D.B. (2005) HDBStat!: A platform-independent software suite for statistical analysis of high dimensional biology data. *BMC Bioinformatics* **6**, 86.
52. Storey, J.D. (2003) The positive false discovery rate: A Bayesian interpretation and the q-value. *The Annals of Statistics* **31**, 2013–2035.
53. Storey, J.D. and Tibshirani, R. (2003) Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* **100**, 9440–9445.
54. Page, G.P., Edwards, J.W., Gadbury, G.L., Yelisetti, P., Wang, J., Trivedi, P., Allison, D.B. (2006) The PowerAtlas: a power and sample size atlas for microarray experimental design and research. *BMC Bioinformatics* **7**, 84.
55. Lee, M.L.T. and Whitmore, G.A. (2002) Power and sample size for DNA microarray studies. *Statistics in Medicine* **21**, 3543–3570.
56. Pan, W., Lin, J., and Le, C.T. (2002) How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. *Genome Biology* **3**(5), 1–10.
57. Zien, A., Fluck, J., Zimmer, R., and Lengauer, T. (2003) Microarrays: how many do you need? *Journal of Computational Biology* **10**, 653–667.
58. Shao, Y. and Tseng, C.-H. (2007) Sample size calculation with dependent adjustment for FDR-control in microarray studies. *Statistics in Medicine* **26**, 4219–4237.
59. Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Science* **95**, 14863–14868.
60. Garge, N.R., Page, G.P., Sprague, A.P., Gorman, B.S., and Allison, D.B. (2005) Reproducible clusters from microarray research: Wither? *BMC Bioinformatics* **6**(Suppl 2), S10.
61. Kerr, M.K. and Churchill, G.A. (2001) Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proceedings of the National Academy of Science* **98**, 8961–8965.
62. McLachlan, G.J. and Khan, N. (2004) On a resampling approach for tests on the number of clusters with mixture model-based clustering of tissue samples. *Journal of Multivariate Analysis* **90**, 90–105.
63. Kapp, A.V. and Tibshirani, R. (2007) Are clusters found in one dataset present in another dataset? *Biostatistics* **8**, 9–31.
64. Breitling, R., Amtmann, A., and Herzyk, P. (2004) Iterative Group Analysis (iGA): a simple tool to enhance sensitivity and facilitate interpretation of micro array experiments. *BMC Bioinformatics* **5**(1), 34.

65. Osier, M.V. (2006) Postanalysis interpretation: "What do I do with this gene list?" In: Allison DB, Page GP, Beasley TM, Edwards JW, ed. DNA Microarrays and Related Genomics Techniques. Chapman & Hall. CRC: Boca Raton, FL, 321–333.
66. Osier, M.V., Zhao, H., and Cheung, K.-H. (2004) Handling multiple testing while interpreting microarrays with the gene ontology database. *BMC Bioinformatics* **5**, 124.
67. Pavlidis, P., Qin, J., Arango, V., Mann, J.J., and Sibille, E. (2004) Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex. *Neurochemical Research* **29**(6), 1213–1222.
68. Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M.J., Patterson, N., Mesirov, J.P., Golub, T.R., Tamayo, P., Spiegelman, B., Lander, E.S., Hirschhorn, J.N., Altshuler, D., and Groop, L.C. (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately down-regulated in human diabetes. *Nature Genetics* **34**(3), 267–273.
69. 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. *Proceedings of the National Academy of Science* **43**, 15545–15550.
70. Goeman, J.J., van de Geer, S.A., de Kort, F., and van Houwelingen, H.C. (2004) A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* **20**(1), 93–99.
71. Pan, W. (2005) Incorporating gene functional annotations in detecting differential gene expression. *Journal of the Royal Statistical Society*, Series C-Applied Statistics **55**, 301–316.
72. Xiang, Q., Edwards, J.W., and Gadbury, G.L. (2006) Interval estimation in a finite mixture model: Modeling P-values in multiple testing applications. *Computational Statistics and Data Analysis* **51**, 570–586.
73. Damian, D. and Gorfine, M. (2004) Statistical concerns about the GSEA procedure. *Nature Genetics* **36**, 663.
74. Mehta, T., Tanik, M., and Allison, D.B. (2004) Towards sound epistemological foundation of statistical methods for high-dimensional biology. *Nature Genetics* **36**, 943–947.
75. Genovese, C. and Wasserman, L. (2002) Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society,*Series B **64**, 499–517.
76. Hsueh, H., Chen, J.J., and Kodell, R.L. (2003) Comparison of methods for estimating the number of true null hypotheses in multiplicity testing. *Journal of Biopharmaceutical Statistics* **13**(94), 675–689.
77. Cattell ,R.B. and Jaspars, J. (1967) A general plasmode (No. 30-10-5-2) for factor analytic exercises and research. *Multivariate Behavioral Research Monographs* **67**, 1–212.
78. Choe, S.E., Boutros, M., Michelson, A.M., Church, G.M., and Halfon, M.S. (2005) Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biology* **6**(2), R16.
79. Gadbury, G.L., Xiang, Q., Yang, L., Barnes, S., Page, G.P., Allison, D.B. (2007) Evaluating Statistical Methods Using Plasmode Data Sets in the Age of Massive Public Databases: An Illustration using False Discovery Rates. *Plos Genetics* **4(6)**, e1000098.