Epistemological issues in omics and high-dimensional biology: give the people what they want

Tapan S. Mehta, Stanislav O. Zakharkin, Gary L. Gadbury and David B. Allison

Physiol. Genomics 28:24-32, 2006. First published 12 September 2006;
doi:10.1152/physiolgenomics.00095.2006

Physiol Genomics 28: 24–32, 2006.
First published January 2007; doi:10.1152/physiolgenomics.00095.2006.

CALL FOR PAPERS | *2nd International Symposium on Animal Functional Genomics*

Epistemological issues in omics and high-dimensional biology: give the people what they want

Tapan S. Mehta,[1] Stanislav O. Zakharkin,[1] Gary L. Gadbury,[2] and David B. Allison[1,3,4]

[1]*Department of Biostatistics, Section on Statistical Genetics, University of Alabama at Birmingham, Birmingham, Alabama;* [2]*Department of Mathematics and Statistics, University of Missouri-Rolla, Rolla, Missouri;* [3]*Clinical Nutrition Research Center, University of Alabama at Birmingham; and* [4]*Department of Genetics, University of Alabama at Birmingham, Birmingham, Alabama*

Mehta T, Zakharkin SO, Gadbury GL, Allison DB. Epistemological issues in omics and high-dimensional biology: give the people what they want. *Physiol Genomics* 28: 24–32, 2006. First published September 12, 2006; doi:10.1152/physiolgenomics.00095.2006.—Gene expression microarrays have been the vanguard of new analytic approaches in high-dimensional biology. Draft sequences of several genomes coupled with new technologies allow study of the influences and responses of entire genomes rather than isolated genes. This has opened a new realm of highly dimensional biology where questions involve multiplicity at unprecedented scales: thousands of genetic polymorphisms, gene expression levels, protein measurements, genetic sequences, or any combination of these and their interactions. Such situations demand creative approaches to the processes of inference, estimation, prediction, classification, and study design. Although bench scientists intuitively grasp the need for flexibility in the inferential process, the elaboration of formal supporting statistical frameworks is just at the very start. Here, we will discuss some of the unique statistical challenges facing investigators studying high-dimensional biology, describe some approaches being developed by statistical scientists, and offer an epistemological framework for the validation of proffered statistical procedures. A key theme will be the challenge in providing methods that a statistician judges to be sound and a biologist finds informative. The shift from family-wise error rate control to false discovery rate estimation and to assessment of ranking and other forms of stability will be portrayed as illustrative of approaches to this challenge.

statistical genonics; proteomics; microarray experiments

HERE, WE WILL DISCUSS some of the unique statistical challenges facing investigators studying high-dimensional biology, describe some approaches being developed by statistical scientists, and offer an epistemological framework for the validation of proffered statistical procedures.

1: CHALLENGES OF HIGH-DIMENSIONAL BIOLOGY

Advances in modern technologies have led to the generation of tremendous amounts of genomic, proteomic, and other "omic" data. The term "high-dimensional biology" (HDB) has been proposed for investigations involving such high-throughput data. Types of HDB information are whole genome sequences and polymorphisms, expression levels of genes, protein abundance measurements, and combinations thereof. New tools address important biological questions such as the following. What is a gene (54)? How many genes do we have, and how can we determine their functions? What are the fundamental principles of metabolic processes? The identification of biomarkers, effects of mutations, and effects of drug treatments and the investigation of diseases as multifactorial phenomena can now be accomplished on an unprecedented scale.[1]

Our paper is targeted to statisticians, biologists, and those "hybrids" that increasingly bridge the interface. This is important because genomics is a rapidly evolving field, and we believe that progress can best be made when these mutually dependent disciplines work together. Therefore, we try to appeal to each of these groups, recognizing that, in some places, the level may be seen as somewhat esoteric for one and

Article published online before print. See web site for date of publication (http://physiolgenomics.physiology.org).

Address for reprint requests and other correspondence: D. B. Allison, Section on Statistical Genetics, Dept. of Biostatistics, and Clinical Nutrition Research Center, Dept. of Nutrition Sciences, Ryals Public Health Bldg., Suite 327, Univ. of Alabama at Birmingham, 1665 University Blvd., Birmingham, AL 35294 (e-mail: Dallison@UAB.edu).

[1] The 2nd International Symposium on Animal Functional Genomics was held May 16–19, 2006, at Michigan State University in East Lansing, MI, and was organized by Jeanne Burton of Michigan State University and Guilherme J. M. Rosa of University of Wisconsin, Madison, WI (see meeting report by Drs. Burton and Rosa, *Physiol Genomics* 28: 1-4, 2006).

somewhat basic for another. We discuss various issues related to drawing inferences from the HDB data. These include the increasing interest in questions involving the testing of multiple propositions simultaneously; appropriate inferential indicators for the types of questions biologists are interested in; and the need for replication of results across independent results, even (perhaps especially) in the face of extremely small *P* values.

The novel and challenging biological questions asked from HDB data have resulted in many specialized analytic techniques being developed. Often statistical methods applied to HDB data lack demonstration of their validity. It is important to evaluate the validity of the statistical methods applied (50, 86). In particular, we discuss resampling-based approaches for HDB data as a case study and illustrate some of the epistemological issues related to such inferential methods.

In summary, questions that biologists want to ask from HDB data require novel analytic approaches. Thus it is important that the statistical methods applied to HDB data are aimed at drawing inferences biologists are interested in and also that these analytic methods have sound epistemological foundations. While design issues are as important as inferential issues, we do not dwell on them here. Issues such as sample size and power estimation (24, 45, 62, 83) technical and biological replication (60), for definition of these terms see (11), and optimal pairing of samples (as well as dye labeling assignment) in two-channel microarray assays (76), and other topics are addressed elsewhere in the literature (6, 11, 17, 18, 39, 49, 52, 59, 65, 84).

2: INFERENCE IN HDB

We can address many questions via HDB data. Examples include the following. Which genes cause or are associated with a disease? Which genes are involved in a particular pathway? Which genes are differentially expressed under which condition? We will address some of the inferential procedures used to answer these questions.

2.1: A Framework for One Gene

For ease of illustration, suppose there are two treatments, *treatment A* and *treatment B*, and gene expression is recorded for only one gene on each of *N* arrays. (We use the terms array and tissue/organism/etc. interchangeably, assuming that each biological specimen will have its own array.) Furthermore, suppose that *Y* is the variable denoting the outcome under *treatment A*, and *X* is the variable denoting the outcome under *treatment B*. We are assuming that each array presents a single measurement for each gene, such as occurs with single-channel arrays or two-color arrays where a ratio or log ratio represents "expression" for a gene. The *N* arrays are randomly divided into two treatment groups, *group A* and *group B*, of size *n* and *m*, respectively, where $n + m = N$. The standard random sampling framework is that $Y_1,\ldots,Y_n \sim^{iid} F_Y(y;\mu_Y,\sigma_Y)$ and $X_1,\ldots,X_m \sim^{iid} F_X(x;\mu_X,\sigma_X)$, where *iid* stands for identically and independently distributed, and the parameters μ_Y,σ_Y and μ_X,σ_X designate the mean and standard deviation of the two population models F_Y and F_X, respectively. When we are interested in determining whether the gene is differentially expressed because of treatment, the quantity representing this differential expression is frequently $\mu_Y - \mu_X$. A null hypothesis of no differential expression, $Ho{:}\mu_Y = \mu_X$, is evaluated using a test statistic computed on observed data and a *P* value computed under some assumed reference distribution (i.e., a distribution of a test statistic when *Ho* is true). The choice of an appropriate reference distribution is important for computing valid *P* values leading to meaningful tests of differential expression. Valid *P* values will be obtained if the required assumptions for the test are met, and they may not be valid of one or more assumptions are not met. In microarray experiments, sample sizes are often not large, and the expression levels may not be normally distributed. Alternative methods for obtaining a reference distribution fall under a category of methods sometimes called nonparametric or distribution free. These terms are slightly misleading in that parameters may still be estimated using these methods, and some assumptions regarding the distribution of data are required. The hypothesis-testing framework above refers to a classical frequentist approach. There are alternative approaches such as Bayesian. For more information about such approaches, please refer to *section 2.5*. We discuss two types of techniques that use the observed data and resampling strategies to produce a reference distribution for a test statistic.

2.2: Resampling-Based Procedures for HDB Data

Resampling-based inference (RBI) relies on resampling data instead of theoretical distributions to make inferences. RBI has the advantage of being robust and flexible enough to accommodate almost any novel statistic [e.g., the statistic after shrinkage of variance, from Cui et al. (15)] without the need for methodologists to perform analytic derivations but has the disadvantage of being computationally intensive. There are marked differences in how such approaches are implemented, and some confusion and uncertainty remain. Microarray investigators who use RBI often do not discuss these issues nor state why one RBI approach (e.g., bootstrap) is chosen over another (e.g., permutation testing)[2] or vice versa.

Frequently, unrecognized complexities arise when RBI is used for complex experiments, for example, when testing the difference between two groups (e.g., old and young mice) after controlling for some other factors (e.g., body fat). There are multiple ways to permute data in such circumstances, and only some will produce valid inferences (38). Another issue is the sampling unit, where a common error occurs in treating the gene rather than the case as the unit of analysis, such as in gene class testing (GCT) (40, 86). GCT attempts to conduct more powerful or informative analyses by testing for differential expression in entire predefined classes of genes rather than each gene singly. However, some GCT methods erroneously resample the genes rather than the cases. This type of resampling essentially ignores dependence among genes, ignores sample size, and can result in nonsensical results (e.g., tests whose power does not increase with sample size). Fortunately, several newer GCT methods are available that do not make this error (3, 27, 47). Some software use resampling procedures [e.g., the significance analysis of microarrays (SAM) algorithm] (70) but combine all resampled test statistics across all genes to obtain very small *P* values. This practice of combining

[2] It is common that the terms permutation test and randomization test are used interchangeably, and we do so here.

all resampled statistics is valid under the assumptions that *1*) the null distribution of the test statistic is the same for all transcripts, and *2*) all transcripts are independent. Unfortunately, there is no reason to assume either (40, 41). Therefore, some microarray software leaves the decision to the user and offers a choice of pooling or not pooling the resampled test statistics across genes (15, 78). Consequently, how we can obtain the benefit of pooling RBI statistics across transcripts without incurring the detriment of falsely assuming *1*) or *2*) is an important question meriting research. Two RBI techniques, randomization and the bootstrap, are discussed in more detail below.

2.2.1: Randomization. Continuing with the one-gene framework in *section 2.1*, where n arrays are randomly selected to receive *treatment A* (i.e., Y is observed) and the other m receive *treatment B* (X is observed), the null hypothesis (*Ho*) is described as a statement: the treatment assignment has no effect on the gene expression. A test statistic, t, is computed to represent a difference in gene expression due to treatment.

Suppose that r_j is the gene expression level of the jth array, which represents a value of Y_j or X_j depending on the treatment assignment for that array. The vector $(r_1,\ldots,r_N)$ represents the gene expression outcomes from the experiment for all N arrays. Under *Ho*, a randomization distribution of the test statistic is created by randomly choosing n values from $(r_1,\ldots,r_N)$ to represent outcomes to *treatment A* with the other m values representing outcomes to *treatment B*. Each time this is done, a test statistic, t^*, is computed, yielding $t_1^*,\ldots,t_C^*$, where $C = N!/(n!m!)$ is the number of unique treatment assignments, and each value of t^* is equally likely under *Ho*. The P value is then computed as the proportion of t^* that is as extreme or more extreme than the test statistic t observed from the data produced from the actual treatment assignment. The exactness of this P value depends on an assumption that is sometimes called unit-treatment additivity, meaning $Y = X + \tau$, where τ is a constant treatment effect; $\tau = 0$ under the null hypothesis of no differential expression. Note that unit-treatment additivity necessarily implies that the variances of the two outcome variables, Y and X, are the same and that the two distributions above, F_Y and F_X, differ only in their location (i.e., their mean). This assumption is sometimes referred to as exchangeability, meaning observations are exchangeable across treatment conditions under *Ho* or that the joint distribution of (X, Y) is the same as that of (Y, X).

As sample sizes become larger, the number C above becomes too large to allow for computation of exact P values. In this case a Monte Carlo approximation (i.e., choosing, for example, $C = 10{,}000$ random permutations) can give accurate results. If sample sizes are too small, these randomization tests produce P values that are too coarse, and it will often be algebraically impossible to obtain P values below some specified level (25). For instance, if $N = 6$ and $n = m = 3$, only 10 two-tailed P values are possible, the smallest being 0.1. In a microarray experiment, this makes using these P values for estimating false discovery rates and determining sets of genes that are most significant difficult at best. Note that the common rank-based tests (e.g., Mann-Whitney and Wilcoxon rank tests) are randomization tests after the appropriate transformation of data to ranks. Another limitation of randomization tests is that, in more complicated designs (e.g., multiple treatment groups and blocking variables, possibly with continuous covariates), implementing the permutations can be more complex and computationally time consuming. In some situations, the bootstrap can be useful.

2.2.2: Bootstrap. The bootstrap is another resampling approach, and it can produce results similar to those of randomization tests in some applications although it is fundamentally different (20). Consider the one-gene framework in *section 2.1* where $Y_1,\ldots,Y_n \sim^{iid} F_Y(y;\mu_Y,\sigma_Y)$ and $X_1,\ldots,X_m \sim^{iid} F_X(x;\mu_X,\sigma_X)$. The bootstrap works because the empirical distribution function of the data is a nonparametric maximum likelihood estimate of the population distributions. The empirical distribution function for Y, for example, is a discrete distribution with a mass of $1/n$ on each observed Y value. Repeated sampling from a population is approximated by resampling with replacement from the original sample. This is the usual nonparametric implementation of the bootstrap, but there are others. One is the parametric bootstrap, where the form of the distribution (say normal) is assumed for F_Y and F_X but the parameters are estimated from the data. Bootstrap samples are then drawn from the estimated normal distribution.

In hypothesis testing, care must be taken when implementing the resampling procedure to ensure that the bootstrap samples are drawn under the case where the null hypothesis is true. Hall and Wilson (29) provide some guidelines for appropriate resampling and for obtaining more accurate P values using the bootstrap. There are advantages and disadvantages of bootstrap tests relative to other resampling-based tests. See Ref. 28 for more details.

2.3: Differential Expression for a Study of K Genes

We now consider resampling strategies for simultaneously testing hypotheses for differential expression on K genes. A clear understanding of the sampling unit is critical for the proper implementation of the resampling techniques described earlier. The distribution models become multivariate, and the first of the two samples is now denoted as follows: $Y_1, Y_2,\ldots, Y_n \sim^{iid} F_Y(y,\mu_Y,\Sigma_Y)$, where $Y_j = (Y_{1j},\ldots,Y_{Kj})$, is a K-dimensional vector of gene expression values for the jth array, μ_Y is a K-dimensional vector of the mean expression for each gene represented on the arrays, and Σ_Y is a $K \times K$ variance-covariance matrix. Similarly, the second sample, $X_1,\ldots,X_m \sim^{iid} F_X(x,\mu_X,\Sigma_X)$, is now a multivariate random sample from a multivariate population model. The earlier discussions for testing a single gene for differential expression across the two 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. Gene-specific tests for differential expression seem to be the preferred method for analyzing microarray data, although methods that test for classes of genes have been proposed [e.g., Barry et al.(3)]. The result from gene-specific tests is a distribution of K test statistics or P values. The most differentially expressed genes are determined by a ranking procedure or assignment of a posterior probability (refer to *section 2.5* for discussion on the Bayesian approaches) of being differentially expressed. As mentioned earlier, bootstrap and 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.

The variance-covariance matrix for each distribution can be written as, for example, $\Sigma_Y = diag(\sigma_{Y1},\ldots,\sigma_{YK})P_Y diag(\sigma_{Y1},\ldots,\sigma_{YK})$, where $diag(\sigma_{Y1},\ldots,\sigma_{YK})$ is a diagonal matrix with entries σ_{Yi}, which are the standard deviations of gene expressions for the ith gene under *treatment A*, and P_Y is a $K \times K$ correlation matrix with 1 on the diagonal and off-diagonal entries $\rho_{Y,i,i'}$, which is the correlation between gene expression levels for gene i and gene i' under *treatment A*. Similar notation holds for random variables under *treatment B*, except that the variable Y is replaced by X. Resampling or permuting at the level of the gene will not preserve the structure of the correlation matrix. The appropriate procedure to mimic sampling from the multivariate distributions, F_Y and F_X, will require sampling at the level of the array. To overcome coarseness in the distribution of test statistics, some have fit a model to the gene expression data that includes gene effects and resampling or permuting the residuals from the model (e.g., Ref. 14). Whether the ANOVA model produced independent residuals with equal variance (across genes) is not clear. 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 (2, 23, 33, 77).

If one develops a method that requires exchangeability across genes (or sets of genes), then the sensitivity to violations of this assumption could be evaluated using simulation procedures, as in Allison et al. (1) or Gadbury et al. (25). Others have considered the issues of obtaining a correct reference distribution using appropriate resampling procedures (34, 80). Directly related to the issues of appropriate resampling procedures for valid inference are the issues of how to simulate data appropriately, that is, to mimic a sample produced from F_Y and/or F_X. More discussion of this appears in *section 3.2*.

2.4: Intersection-Union Testing

Biologists often wish to address questions involving multiple propositions simultaneously. For example, they may wish to ask whether a particular gene is differentially expressed in both of two (or more) tissues in response to some common stimulus or whether a particular gene is linked to a particular phenotype in both of two (or more) species. These questions about multiple propositions involve "and" rather than "or" questions. Investigators are asking whether each of several propositions is true, not whether any of several propositions is true. This is an important goal for biologists, and the distinction is important.

Traditionally, statisticians have devoted much attention and effort to testing multiple hypotheses or propositions simultaneously. Much of the multiple-testing literature has involved union-intersection tests (UIT), where the goal is to test the intersection of all null hypotheses against the union of all alternative hypotheses. All of the classic corrections for multiple testing are UITs. Examples can be found in the traditional literature on this topic (69) and even in the more modern literature that uses resampling-based methods [e.g., Westfall and Young (75)]. UITs test whether any of the null hypotheses is false, not whether all are false. In many cases, this is wholly appropriate, but this does not fully address the questions that biologists have when they ask about multiple propositions. To address some questions posed by biologists, such as in the examples offered above, intersection-union testing (IUT) is required (5, 64).

Most investigators using applications are unfamiliar with the formal aspects of IUTs. In trying to accomplish their goals, they often use homegrown approaches. For example, Kyng et al. (42) investigated differences in gene expression among cell lines of normal young individuals, elderly individuals, and individuals with Werner syndrome (WS; a disease of premature aging). They compared the cell lines of the old with those of the young and identified a number of transcripts that appeared to have significant differences. They then compared the cell lines from the WS individuals with those of the young normal individuals and also obtained a list of transcripts that appeared to have significant differences. The authors then noted that a great deal of overlap existed between the lists of genes obtained in these two analyses. They interpreted this to be evidence supporting the conjecture that WS is, at the transcript level, a process highly similar to accelerated normal aging. Although the conjecture may be correct, this particular line of evidence in support of it is highly questionable. First, by using a common control group, the authors introduced a dependency into the sample mean differences for old vs. young and WS vs. young. A model proper for examining whether the degree to which such overlap in the lists of differentially expressed genes is above that expected by chance would need to take this into account. Second, the authors do not consider that multiple transcripts may be correlated, that is, that not all gene transcription levels are dependent. This further complicates matters and renders highly questionable simply looking at the proportion as, for example, a two-by-two table of differentially expressed and not differentially expressed in each of the two conditions. Finally, in treating genes as simply having or not having evidence for differential expression, one loses the continuity of the data and the evidence they have to offer. Although statistical methodologists have spent a great deal of effort on multiple-testing corrections, investigators have been given little help with the IUT framework.

Classical IUT entails the use of the MIN-Test (44). In the MIN-Test, one rejects the union of null hypotheses in favor of the intersection of alternative hypotheses if and only if the largest P value for the statistical significance test on each of the individual component hypotheses is less than or equal to some preset alpha (α) level. This approach has several problems, the first of which is that the IUT will not have a predefined size. That is, using this approach, one can only state that one's type I error rate for the IUTs will be less than or equal to α, not equal to α. Because of this, the tests will not be very powerful under many circumstances. The second problem is that this method again does not fully utilize the continuous information available in the data about the strength of evidence. For example, if *investigator A* conducted two-component hypothesis tests (e.g., gene X is differentially expressed under *condition 1*, and gene X is differentially expressed under *condition 2*) and obtained P values of 0.049 and 0.048 for the two component hypotheses, *investigator A* would be able to state that the IUT was significant at the 0.049 level. Similarly, if *investigator B* conducted the same two experiments and obtained P values of 0.049 and 10^{-20}, *investigator B* would still only be able to say that the IUT null hypothesis was rejected at the α level of

0.049. This seems incongruous given the clearly differential evidence that the two investigators obtained. Unfortunately, getting out of this difficulty may not be possible in a strict frequentist paradigm.

As we begin to integrate more information across more experiments, conditions, methods, tissues, and species and move into the age of "integromics" (37, 74), omic investigators will need IUTs more frequently. It is incumbent on us to give the people what they want and begin developing better IUTs. The Bayesian framework in which one fits models to large collections of data by modeling the *P* values (1), test statistics (35), or effect-size indicators (51) may offer an excellent approach in this regard. Additional research is warranted to extend the methods developed from simply looking at a single vector of results from one microarray experiment to examining multiple vectors of results from multiple microarray or other omic-level experiments.

2.5: Which Inferential Indicator Do Biologists Want and Understand?

The staple of inference in modern statistics is the frequentist *P* value. As mentioned in *section 2.1*, this value indicates the probability of obtaining data that depart as much or more from the expectation under the null hypothesis as the data that were actually observed if, in fact, the null hypothesis were true. Perhaps not surprisingly, this somewhat cumbersome description is not easily grasped by many nonstatisticians. Indeed, survey research has shown that many nonstatisticians believe that a *P* value indicates the probability that the null hypothesis is false (79). A smaller but still nontrivial proportion of nonstatisticians believe that a *P* value indicates the probability that a result will not replicate if the experiment is repeated. The fact that most nonstatisticians misunderstand *P* value suggests two things. First, if we are to continue using *P* values, we need to work more diligently to help our colleagues understand what they are. Second, classic frequentist thinking may not effectively capture how most nonstatisticians think, and they may be more amenable to other indicators of evidential strength. Most physicians and physiologists seem intuitively to be Bayesians and comfortable talking about the probability that a null hypothesis is true or false. That is, they are interested in a probability that the null hypothesis is true given the observed data rather than probabilities associated with observed data given a true null hypothesis, which is the basis for the definition of a *P* value. Bayesian techniques that facilitate these interpretations are available and being used more frequently in HDB investigations (81).

It is ironic then that, as the current age of discovery-based HDB bloomed, the first reaction of many statisticians was to offer family-wise error rate (FWER) methods (19, 72, 73). FWER is the probability of making one or more type I errors in a family or set of comparisons or inferences (21). For example, in a microarray study for detecting differentially expressed genes between two conditions, FWER could be defined as the probability of incorrectly identifying at least one differentially expressed gene among those genes that are not truly differentially expressed (48). Although this has some appeal and some correction for multiple testing does seem necessary, when we communicate with our colleagues who use these applications and ask "Do you wish us to be certain, within some very small probability such as 5%, that we never tell you that a particular gene is differentially expressed (or linked, or associated, etc.) in this study of tens of thousands of genes when in fact it is not differentially expressed (or linked, or associated, etc.)?", the response is invariably "No." Our colleagues in the biology labs tell us that they do not mind if we give them a few false positives as long as a substantial proportion of the results offered are not false positives. This essentially describes the false discovery rate (FDR), and it appears that biologists are far more interested in maintaining a reasonably low FDR than a FWER. Loosely defined, FDR is the expected proportion of "findings" that are in fact erroneous (for further discussion and some useful techniques, see Refs. 67, 68). How low an FDR must be may vary from investigator to investigator and context to context, but it suggests that, as statisticians, we may have overemphasized FWER control. Statistical geneticists have embraced this, and an enormous amount of research over the past 5 years has gone into building newer and better FDR procedures.

A detailed review of specific FDR methods and related approaches is beyond the scope of this paper. To place FDR and related techniques in historical and conceptual context, we note that the concept of posterior probabilities having a relation to *P* values goes back for decades (58). The idea of controlling FDR by examining distributions of *P* values goes back at least to 1982 (61), and, to our knowledge, the term FDR was coined and the concept first thoroughly elucidated in 1995 (4). Calculations of posterior probabilities (a very close relation to FDR estimation techniques) in microarray work appeared in the published literature at least as early as 2000 (45). Modeling the *P* values in microarray studies to produce FDR and posterior probability estimates appeared in the published literature by 2002 (1). Several papers reviewing and exploring the connections among FDR and related techniques have now appeared (57), and papers comparing the empirical performance of various FDR methods are starting to appear. Interestingly, such empirical comparisons do not show any one method to be consistently better than all others (7, 53, 56, 82).

This work on FDR procedures has largely been spurred by microarray research but is now being embraced in many other areas. However, one must wonder whether even FDR is what we are really interested in. Consider the case of a very complex trait, obesity. Given the number of genes that have been knocked out and shown to cause an obesity phenotype, ethylnitrosourea (ENU) mutagenesis deletions that have been created, and resulting obesity phenotypes, Jurgen Naggert (personal communication, 2004) estimated that there could easily be on the order of several thousand genes contributing to obesity. If we allow ourselves a rough assumption that these genes are approximately evenly distributed throughout the genome, no region of the genome cannot be linked to obesity. If every region of the genome is linked to obesity, then in what sense can we actually even talk about false discoveries in a genome scan for obesity? There can be no false discoveries because every area of the genome is linked to obesity. The question of interest may instead be about which other regions have the strongest linkage with obesity.

We are no longer interested in strictly testing the truth or falsity of null hypotheses but, rather, in ranking the alternative hypotheses with respect to the strength of evidence in favor of them and/or the strength of the parameter quantifying the

alternative hypothesis. How shall we best give investigators information about this ranking, and how shall we quantify our confidence in any given ranking? This is an area that people are beginning to investigate (55). More research is clearly needed in this area, and we look forward to having this research come to fruition.

2.6: Need for Replication vs. a Single Small P Value

In this section, we refer to replication in its traditional usage as attempting to reproduce entire studies through obtaining new data and not to the design issue of obtaining multiple data points within a study [with respect to the latter, see Churchill (11)]. Many investigators have long held the need for replication before a finding is accepted. Such calls for replication have been especially strong within the community of investigators studying complex genetics. Recently, a method was proposed in which a very large number of hypothesis tests were conducted in a single sample, and then a few of the most significant results were confirmed or "replicated" in the same data by using a different test that was independent of the first test (43). This creative approach is likely to be much used but may not give applied investigators what they want and need.

Years ago, Lykken (46) described different types of replication. He offered that perhaps we should be most confident in a finding when we can replicate it using methods most different from those that resulted in the original finding. This call for "constructive replication" seems to be nowhere more appropriate than in the field of complex trait genetics. For example, Zhang et al. (87) reported a finding that the gene encoding the enzyme responsible for the synthesis of brain serotonin, tryptophan hydroxylase-2 (hTPH2), exhibited a functional variant associated with major depression. This variant, GA1463A, attenuated the capacity of hTPH2 in serotonin production by 80%. This study also observed a strikingly high frequency of this variant in a cohort of unipolar depressed subjects. However, independent large studies later by several other groups were not able to reproduce these results or even find that the polymorphisms existed (26, 71, 90). These puzzling findings suggest that no amount of replication within a single sample, however statistically sophisticated, can substitute for true constructive replication. This problem is not unique to the field of genomics. In 2000, Siefe (63) investigated what might be called the "5 sigma problem," whereby certain findings from the field of physics, although seemingly incontrovertible from the basis of their vanishingly small P values, nevertheless turned out to be incorrect. Apparently, this occurrence is not all that uncommon. According to Siefe, this may result from an overly optimized experiment to detect an event or systematic technical errors (e.g., incorrect settings in computers or a faulty lens used for recording observations) getting introduced in an experiment, leading to 5 sigma results.

Examples of 5 sigma results that fail to replicate abound in the field of complex trait genomics. For example, an exciting paper on obesity was published recently (30). This paper found that a polymorphism in the gene INSIG2 was highly significantly associated with obesity in multiple samples. The initial finding came by use of the self-replication method mentioned earlier and held up for multiple separate samples. In this same paper, however, a large sample of 2,700 subjects failed to confirm the finding. The failure to find such a result seems most unlikely given the sample size of 2,700 and given its relative strength in the other samples. Moreover, in commentary in the journal, P. Froguel (13) said that he also failed to observe the result in a separate sample of 10,000 subjects. This clearly seems like a 5 sigma anomaly. It underscores the need for true constructive replications and perhaps the need for our scientific community to begin more formal investigations of how to conduct and consider replication findings.

3: EPISTEMOLOGICAL FOUNDATIONS OF RESAMPLING-BASED INFERENTIAL METHODS

Epistemology is a branch of philosophy that deals with the nature, origin, and scope of knowledge. Mehta et al. (50) and Zakharkin et al. (86) offer guidelines about evaluating the validity of statistical methods in HDB and illustrate some of the seemingly obvious but not universally appreciated statistical issues. In an experiment, we observe measurements of variables from samples drawn from populations, and these observations form our base data. Using these data, we wish to make inferences from imperfect measurements about the real variables they represent and then from the sample cases to the whole population. Some authors resample across genes when trying to assess the stability of certain microarray results, whereas the samples-to-population perspective holds that the sampling units should be cases (e.g., mice) and not genes (22). A very recent paper from Klebanov and Yakovlev (40) explains the rationale of not drawing inference by using genes as sampling units and the risk involved in doing so. Similarly, methods were proposed where inferences about gene expression differences between populations are made by comparing observed sample differences with an estimated null distribution of differences based on technical rather than biological replicates. This conflates the standard error of measurement with the standard error of the sample statistic. We offer a framework for evaluating the validity of inferential methods for HDB.

3.1: Mathematical Proofs

Methods should be evaluated with respect to well-defined objective criteria such as sensitivity, specificity, type I error rate control, power, unbiasedness, reproducibility, etc. The classical way to characterize the operating characteristics of methods is via mathematical proofs. Mathematical proofs are well accepted but can be difficult, if not impossible, in many situations, particularly when typical assumptions (e.g., independence and distribution assumptions) are not met. In such situations, the use of simulations becomes key.

3.2: Simulations

The importance of simulations is well known to many statisticians, and such readers may choose to skip this section. Simulations are valuable, because some truth about F_Y and F_X is known to the investigator, who can then determine how well a statistical method reveals this truth. Initial procedures for simulating microarray data assumed F_Y and F_X to have a parametric form (often a normal distribution). Methods for determining the mean vectors for these distributions were usually reasonable, determining the variances were more difficult, and determining the correlation matrices, P, were nearly impossible. Many simulations used the identity matrix for P and simulated genes as independent variables. Allison et al. (1)

and Gadbury et al. (25) used a block diagonal structure for *P*. For example, if 10,000 genes are being studied, then *P* would be block diagonal with 20 blocks of dimension 500. Each block would be an equicorrelation matrix with 1 on the diagonal and ρ everywhere else, and ρ would vary over different simulation runs to simulate weak, medium, or strong dependence among groups of genes. Although some argument could be made that this is reasonable, it is likely far from ideal, in that the correlation structure of an actual data set is unknown and negative values for ρ produce negative definite matrices. As larger studies have become available, it has made sense to borrow the concept of the bootstrap to simulate data with the same variance-covariance structure as an actual data set. This has been termed a plasmode simulation and is described later.

3.3: Plasmodes

3.3.1: What are plasmodes? Concerns about how well simulated data correspond to reality can partially be addressed by using plasmodes. To our knowledge, the term "plasmode" was introduced as early as 1967 by Cattell and Jaspars (8). A plasmode is 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. The availability of large plasmode databases of microarray data allows a method to be evaluated with hundreds if not thousands of data sets.

3.3.2: Different ways of creating plasmodes. 3.3.2.1: IN WET LAB. Plasmode data sets can be created in wet labs by directly manipulating biological samples. A simple example is a real microarray data set with specific mRNAs spiked in (12, 32). Evaluating whether a particular method can correctly detect the spiked mRNAs gives information about the method's ability to detect gene expression. An outstanding resource for the field is Affycomp (32), 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 and then compare results with those of many other methods that have been applied to the same data (31, 89). The nonspecific binding data set in Zhijin et al. (88) and the "controlled" data set in Choe et al. (9) can also be considered to be plasmode data sets created in wet labs. With respect to the latter data set (9), there has been recent debate as to whether it was constructed in such a way that allows it to be legitimately useful for method evaluation (10, 16). Nevertheless, plasmodes are the start of a valuable resource for the scientific community. Certain data sets are commonly used (e.g., yeast cell cycle data) and could become de facto standards for methods evaluation (66).

3.3.2.2: APPLYING RESAMPLING-BASED METHODS ON REAL DATA SETS WITH UNKNOWN TRUTH. Plasmodes could also be created from a real data set where the truth is unknown by applying resampling-based methods. The randomization technique can be used to create plasmode data sets where the null hypothesis is true. Null plasmode data sets can be used in at least two ways. First, the type I error rate of the method can be estimated by applying a method to each data set. Any significant results obtained are, by definition, type I errors (false positives). Second, anyone wishing to evaluate the reproducibility (stability) of results produced by one or more methods can apply each method separately to two groups within each data set and examine the extent to which similar results are obtained. Methodologists can evaluate how successful any particular method of analyzing data is in detecting the true effects and not detecting the null effects and then be able to have common, identifiable, benchmark data sets for comparing analytic approaches.

Plasmode data sets with modest or large effects can also be created. A real template data set can be selected that involves two groups in which there appear to be real but small differences in gene expression (i.e., only a modest portion of genes differentially expressed with only moderate difference in expression across the two groups). The mixture-modeling approach used for power projections (1, 24) is used to fit a model to the data; the proportion of differentially expressed genes and the distribution of standardized mean differences in expression for those differentially expressed genes can then be estimated. The estimated effects from the template data set (scaled to the study-specific variances) can be added to a particular group (of the two groups) of the null data sets. This will create a data set with a realistic distribution of null and nonnull effects and, again, preserve the marginal distributions and covariances among gene expression levels. This process can be repeated to create plasmode data sets with different template data sets in which there appear to be moderate-to-large real differences in gene expression. This method of plasmode generation differs from a standard simulation study, in that data are generated in such a manner that the covariance structure among all genes and the marginal residual distributions for all genes are preserved and generated from real data and therefore, by definition, known to be realistic. In contrast, it is recognized (36) that no one knows how to simulate data that have such characteristics.

Methodologists can evaluate how successful any particular method of analyzing data is in detecting the true effects and in not detecting the null effects. The result will be common, identifiable, benchmark data sets against which they can judge the comparative performance of analytic approaches.

4: DISCUSSION

In conclusion, the fields of complex trait genetics and HDB have led to new challenges to our ability to offer veridical findings and provide statistical methods that provide information that biologists find to be germane and are epistemologically sound. Statisticians have begun to offer new approaches to this and to exciting developments that have broken us out of our traditional approaches in favor of new epistemological criteria.

ACKNOWLEDGMENTS

Present address of S. O. Zakharkin: Solae, LLC, St. Louis, MO 63188.

GRANTS

This work was supported by the University of Alabama at Birmingham Statistical Genetics Postdoctoral Training Program Grant No. T32-HL-072757-04 and the Center for Nutrient-Gene Interaction in Cancer Prevention Grant No. 5U54-CA-100949-04 from the National Institutes of Health and

grant no. 0217651 from the National Science Foundation Plant Genome Research Program (Design and Analysis of Microarray Gene Expression Studies in Plants: Toward Sound Statistical Procedures).

REFERENCES

1. **Allison DB, Gadbury G, Heo M, Fernandez J, Lee CK, Prolla TA, Weindruch R.** A mixture model approach for the analysis of microarray gene expression data. *Comput Stat Data Anal* 39: 1–20, 2002.
2. **Baldi P, Long A.** A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* 17: 509–519, 2001.
3. **Barry WT, Nobel AB, Wright FA.** Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics* 21: 1943–1949, 2005.
4. **Benjamini Y, Hochberg Y.** Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B* 57: 289–300, 1995.
5. **Berger RL, Hsu JC.** Bioequivalence trials, intersection-union tests, and equivalence confidence sets. *Stat Sci* 11: 283–319, 1996.
6. **Bolstad BM, Collin F, Simpson KM, Irizarry RA, Speed TP.** Experimental design and low-level analysis of microarray data. *Int Rev Neurobiol* 60: 25–58, 2004.
7. **Broberg P.** A comparative review of estimates of the proportion unchanged genes and the false discovery rate. *BMC Bioinformatics* 6: 199, 2005.
8. **Cattell RB, Jaspars J.** A general plasmode (No. 30-10-5-2) for factor analytic exercises and research. *Multivariate Behav Res Monographs* 67: 1–212, 1967.
9. **Choe SE, Boutros M, Michelson AM, Church GM, Halfon MS.** Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biol* 6: R16, 2005.
10. **Choe SE, Boutros M, Michelson AM, Church GM, Halfon MS.** Response to 'A reanalysis of a published Affymetrix GeneChip control dataset' by Dabney and Storey in Genome Biology 7: 401, 2006. *Genome Biol* 7: 401.3–401.6, 2006.
11. **Churchill GA.** Fundamentals of experimental design for cDNA microarrays. *Nat Genet* 32: 490–495, 2002.
12. **Cope LM, Irizarry RA, Jaffee HA, Wu Z, Speed TP.** A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics* 20: 323–331, 2004.
13. **Couzin J.** Gene variant may boost obesity risk. *ScienceNOW Daily News*, 13 April 2006.
14. **Cui X, Churchill GA.** Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol* 4: 210, 2003.
15. **Cui XQ, Hwang TG, Qui J, Blades J, Churchill GA.** Improved statistical tests for differential gene expression by shrinking variance component estimates. *Biostatistics* 6: 59–75, 2005.
16. **Dabney AR, Storey JD.** A reanalysis of a published Affymetrix GeneChip control dataset. *Genome Biol* 7: 401, 2006.
17. **Dhiman N, Bonilla R, O'Kane DJ, Poland GA.** Gene expression microarrays: a 21st century tool for directed vaccine design. *Vaccine* 20: 22–30, 2001.
18. **Dobbin K, Shih JH, Simon R.** Questions and answers on design of dual-label microarrays for identifying differentially expressed genes. *J Natl Cancer Inst* 95: 1362–1369, 2003.
19. **Dudoit S, van der Laan MJ, Pollard KS.** Multiple testing. Part I. Single-step procedures for control of general type I error rates. *Stat Appl Genet Mol Biol* 3: 13, 2004.
20. **Efron B, Tibshirani RJ.** *An Introduction to the Bootstrap*. Boca Raton, FL: CRC, 1993.
21. **Everitt BS.** *The Cambridge Dictionary of Statistics*. Cambridge, UK: Cambridge Univ. Press, 1998, p.124.
22. **Famili AF, Liu G, Liu Z.** Evaluation and optimization of clustering in gene expression data analysis. *Bioinformatics* 20: 1535–1545, 2004.
23. **Fox RJ, Dimmic MD.** A two-sample Bayesian t-test for microarray data. *BMC Bioinformatics* 7: 126, 2006.
24. **Gadbury GL, Page GP, Edwards JW, Kayo T, Prolla TA, Weindruch R, Permana PA, Mountz J, Allison DB.** Power analysis and sample size estimation in the age of high dimensional biology: a parametric bootstrap approach illustrated via microarray research. *Stat Methods Med Res* 13: 325–338, 2004.
25. **Gadbury GL, Page GP, Heo M, Mountz JD, Allison DB.** Randomization tests for small samples: an application for genetic expression data. *J R Stat Soc C* 52: 365–376, 2003.
26. **Glatt CE, Carlson E, Taylor TR, Risch N, Reus VI, Schaefer CA.** Response to Zhang et al. (2005): Loss-of-function mutation in tryptophan hydroxylase-2 identified in unipolar major depression. Neuron 45, 11–16. *Neuron* 48: 704–705, 2005.
27. **Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC.** A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 20: 93–99, 2004.
28. **Good P.** *Permutation Tests. A Practical Guide to Resampling Methods for Testing Hypotheses*. New York: Springer, 1994.
29. **Hall P, Wilson SR.** Two guidelines for bootstrap hypothesis testing. *Biometrics* 47: 757–762, 1991.
30. **Herbert A, Gerry NP, McQueen MB, Heid IM, Pfeufer A, Illig T, Wichmann HE, Meitinger T, Hunter D, Hu FB, Colditz G, Hinney A, Hebebrand J, Koberwitz K, Zhu X, Cooper R, Ardlie K, Lyon H, Hirschhorn JN, Laird NM, Lenburg ME, Lange C, Christman MF.** A common genetic variant is associated with adult and childhood obesity. *Science* 312: 279–283, 2006.
31. **Hochreiter S, Clevert DA, Obermayer K.** A new summarization method for Affymetrix probe level data. *Bioinformatics* 22: 943–949, 2006.
32. **Irizarry RA, Wu Z, Jaffee HA.** Comparison of Affymetrix GeneChip expression measures. *Bioinformatics* 22: 789–794, 2006.
33. **Ishwaran H, Rao JS, Kogalur UB.** BAMarray: Java software for Bayesian analysis of variance for microarray data. *BMC Bioinformatics* 7: 59, 2006.
34. **Jianqing F, Chen Y, Chan HM, Tam Paul KH, Yi R.** Removing intensity effects and identifying significant genes for Affymetrix arrays in macrophage migration inhibitory factor-suppressed neuroblastoma cells. *Proc Natl Acad Sci USA* 102: 17751–17756, 2005.
35. **Johnson VE.** A Bayesian test for goodness-of-fit. *Annals Stat* 32: 2361–2384, 2004.
36. **Jung SH, Jang W.** How accurately can we control the FDR in analyzing microarray data? *Bioinformatics* 22: 1730–1736, 2006.
37. **Katoh M.** WNT2B: comparative integromics and clinical applications. *Int J Mol Med* 16: 1103–1108, 2005.
38. **Kennedy PE, Cade BS.** Randomization tests for multiple regression. *Comm Stat Simul Comput* 25: 923–936, 1996.
39. **Kerr MK, Churchill GA.** Related articles, statistical design and the analysis of gene expression microarray data. *Genet Res* 77: 123–128, 2001.
40. **Klebanov L, Yakovlev A.** Treating expression levels of different genes as a sample in microarray data analysis: is it worth a risk? *Stat Appl Genet Mol Biol* 5: 9, 2006.
41. **Kowalski J, Drake C, Schwartz RH, Powell J.** Non-parametric, hypothesis-based analysis of microarrays for comparison of several phenotypes. *Bioinformatics* 20: 364–373, 2004.
42. **Kyng KJ, May A, Kolvraa S, Bohr VA.** Gene expression profiling in Werner syndrome closely resembles that of normal aging. *Proc Natl Acad Sci USA* 100: 12259–12264, 2003.
43. **Laird NM, Lange C.** Family-based designs in the age of large-scale gene-association studies. *Nat Rev Genet* 7: 385–394, 2006.
44. **Laska EM, Meisner MJ.** Testing whether an identified treatment is best. *Biometrics* 45: 1139–1151, 1989.
45. **Lee ML, Whitmore GA.** Power and sample size for DNA microarray studies. *Stat Med* 21: 3543–3570, 2002.
46. **Lykken DT.** Statistical significance in psychological research. *Psychol Bull* 70: 51–159, 1968.
47. **Mansmann U, Meister R.** Testing differential gene expression in functional groups. Goeman's global test versus an ANCOVA approach. *Methods Inf Med* 44: 449–453, 2005.
48. **McClure J, Wit E.** *Statistics for Microarrays: Design, Analysis and Inference*. New York: Wiley, 2004, p.181.
49. **McShane LM, Shih JH, Michalowska AM.** Statistical issues in the design and analysis of gene expression microarray studies of animal models. *J Mammary Gland Biol Neoplasia* 8: 359–374, 2003.
50. **Mehta T, Tanik M, Allison DB.** Towards sound epistemological foundation of statistical methods for high-dimensional biology. *Nat Genet* 36: 943–947, 2004.
51. **Newton MA, Noueiry A, Sarkar D, Ahlquist P.** Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* 5: 155–176, 2004.
52. **Page GP, Edwards JW, Barnes S, Weindruch R, Allison DB.** A design and statistical perspective on microarray gene expression studies in nutrition: the need for playful creativity and scientific hard-mindedness. *Nutrition* 19: 997–1000, 2003.

53. **Pawitan Y, Murthy KR, Michiels S, Ploner A.** Bias in the estimation of false discovery rate in microarray studies. *Bioinformatics* 21: 3865–3872, 2005.
54. **Pearson H.** What is a gene? *Nature* 441: 399–401, 2006.
55. **Pepe MS, Longton G, Anderson GL, Schummer M.** Selecting differentially expressed genes from microarray experiments. *Biometrics* 59: 133–142, 2003.
56. **Pounds S, Cheng C.** Improving false discovery rate estimation. *Bioinformatics* 20: 1737–1745, 2004.
57. **Pounds SB.** Estimation and control of multiple testing error rates for microarray studies. *Brief Bioinform* 7: 25–36, 2006.
58. **Pratt JW.** Bayesian interpretation of standard inference statements. *J R Stat Soc B* 27: 169–203, 1965.
59. **Pusztai L, Hess KR.** Clinical trial design for microarray predictive marker discovery and assessment. *Ann Oncol* 15: 1731–1737, 2004.
60. **Rosa GJM, Steibel JP, Tempelman RJ.** Reassessing design and analysis of two-colour microarray experiments using mixed effects models. *Comp Funct Genomics* 6: 123–131, 2005.
61. **Schweder T, Spjotvoll E.** Plots of P-values to evaluate many tests simultaneously. *Biometrika* 69: 493–502, 1982.
62. **Seo J, Gordish-Dressman H, Hoffman EP.** An interactive power analysis tool for microarray hypothesis testing and generation. *Bioinformatics* 22: 808–814, 2006.
63. **Siefe C.** CERN's gamble shows perils, rewards of playing the odds. *Science* 289: 2260–2262, 2000.
64. **Sierra-Cavazos JH, Berger RL.** Intersection-union tests in dissolution profile testing [Online]. NCSU Institute of Statistics Mimeo Series 2521, 1999 (http://www.west.asu.edu/rlberge1/papers.html).
65. **Simon R, Radmacher MD, Dobbin K.** Design of studies using DNA microarrays. *Genet Epidemiol* 23: 21–36, 2002.
66. **Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B.** Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell* 9: 3273–3297, 1998.
67. **Storey JD.** A direct approach to false discovery rates. *J R Stat Soc B* 64: 479–498, 2002.
68. **Storey JD.** The positive false discovery rate: a Bayesian interpretation and the q-value. *Annals Stat* 31: 2013–2035, 2003.
69. **Toothaker LE.** *Multiple Comparisons for Researchers*. Thousand Oaks, CA: SAGE, 1991.
70. **Tusher VG, Tibshirani R, Chu G.** Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* 98: 5116–5121, 2001.
71. **Van Den Bogaert A, De Zutter S, Heyrman L, Mendlewicz J, Adolfsson R, Van Broeckhoven C, Del-Favero J.** Response to Zhang et al. (2005): Loss-of-function mutation in tryptophan hydroxylase-2 identified in unipolar major depression. Neuron 45, 11–16. *Neuron* 48: 705–706, 2005.
72. **van der Laan MJ, Dudoit S, Pollard K.** Multiple testing. Part II. Step-down procedures for control of the family-wise error rate. *Stat Appl Genet Mol Biol* 3: 14, 2004.
73. **van der Laan MJ, Dudoit S, Pollard KS.** Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Stat Appl Genet Mol Biol* 3: 15, 2004.
74. **Venkatesh TV, Harlow HB.** Integromics: challenges in data integration. *Genome Biol* 3: REPORTS4027, 2002.
75. **Westfall P, Young S.** *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*. New York: Wiley, 1992.
76. **Wit E, Nobile A, Khanin R.** Near-optimal designs for dual channel microarray studies. *Appl Stat* 54: 817–830, 2005.
77. **Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS.** Assessing gene significance from cDNA microarray expression data via mixed models. *J Comput Biol* 8: 625–663, 2001.
78. **Wu H, Kerr K, Cui X, Churchill G.** MAANOVA: a software package for the analysis of spotted cDNA microarray experiments. In: *The Analysis of Gene Expression Data: Methods and Software*, edited by Parmigiani G, Garrett ES, Irizarry R, and Zeger S. New York: Springer, 2002, p. 313–341.
79. **Wulff HR, Andersen B, Brandenhoff P, Guttler F.** What do doctors know about statistics? *Stat Med* 6: 3–10, 1987.
80. **Xie Y, Pan W, Khodursky AB.** A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data. *Bioinformatics* 21: 4280–4288, 2005.
81. **Yang D, Zakharkin SO, Page GP, Brand JP, Edwards JW, Bartolucci AA, Allison DB.** Applications of Bayesian statistical methods in microarray data analysis. *Am J Pharmacogenomics* 4: 53–62, 2004.
82. **Yang JJ, Yang MC.** An improved procedure for gene selection from microarray experiments using false discovery rate criterion. *BMC Bioinformatics* 7: 15, 2006.
83. **Yang MC, Yang JJ, McIndoe RA, She JX.** Microarray experimental design: power and sample size considerations. *Physiol Genomics* 16: 24–28, 2003.
84. **Yang YH, Speed T.** Design issues for cDNA microarray experiments. *Nat Rev Genet* 3: 579–588, 2002.
85. **Zakharkin SO, Kim K, Mehta T, Chen L, Barnes S, Scheirer KE, Parrish RS, Allison DB, Page GP.** Sources of variation in Affymetrix microarray experiments. *BMC Bioinformatics* 29: 214, 2005.
86. **Zakharkin SO, Mehta T, Tanik M, Allison DB.** Epistemological foundations of statiscal methods for high-dimensional biology. In: *DNA Microarrays and Related Genomic Techniques: Design, Analysis, and Interpretation of Experiments*, edited by Allison DB, Page GP, Beasley MT, and Edwards JW. Boca Raton, FL: CRC, 2006, p. 55–75.
87. **Zhang X, Gainetdinov RR, Beaulieu JM, Sotnikova TD, Burch LH, Williams RB, Schwartz DA, Krishnan KR, Caron MG.** Loss-of-function mutation in tryptophan hydroxylase-2 identified in unipolar major depression. *Neuron* 45: 11–16, 2005.
88. **Zhijin W, Irizarry R, Gentleman R, Murillo FM, Spencer F.** A model based background adjustment for oligonucleotide expression arrays. *J Am Stat Assoc* 99: 468, 2004.
89. **Zhou L, Rocke DM.** An expression index for Affymetrix GeneChips based on the generalized logarithm. *Bioinformatics* 21: 3983–3989, 2005.
90. **Zhou Z, Peters EJ, Hamilton SP, McMahon F, Thomas C, McGrath PJ, Rush J, Trivedi MH, Charney DS, Roy A, Wisniewski S, Lipsky R, Goldman D.** Response to Zhang et al. (2005): Loss-of-function mutation in tryptophan hydroxylase-2 identified in unipolar major depression. Neuron 45, 11–16. *Neuron* 48: 702–703, 2005.