Experimental Design and Analysis of Antibody Microarrays: Applying Methods from cDNA Arrays

Jeanette E. Eckel-Passow,[1] Antje Hoering,[1] Terry M. Therneau,[1] and Irene Ghobrial[2]

Departments of [1]Health Sciences Research and [2]Hematology, Mayo Clinic, Rochester, Minnesota

Abstract

Protein expression microarrays, also called antibody arrays, represent a new technology that allows the expression level of proteins to be assessed directly. As is also the case with gene expression microarrays, it is hoped that protein expression microarrays will aid in biomarker discovery, predicting disease outcomes and response to treatments, and detecting molecular mechanisms and/or pathways associated with a particular disease state. However, accurately achieving these aims is dependent upon suitable experimental designs, normalization procedures that eliminate systematic bias, and appropriate statistical analyses to assess differential expression or expose expression patterns. In the last five years, a large amount of research has been devoted to two-color cDNA arrays to improve experimental design, normalization and statistical analyses to assess differential expression and classification. These methods are directly applicable to two-color antibody arrays. The objective of this article is to discuss statistical methods that have been developed for cDNA arrays and describe how the methods can be directly applied to antibody arrays. (Cancer Res 2005; 65(8): 2985-9)

Introduction

Gene expression microarray technology has exploded in the last five years with the completion of the Human Genome Project. Its use has been shown across a wide range of fields including but not limited to biomarker discovery, predicting disease outcomes and response to treatments, assessing coregulation via time course and/or dose-response experiments, and detecting molecular mechanisms and/or pathways associated with a particular disease state. Early on, it was thought that differential expression at the cDNA level would accurately predict expression at the protein level. However, correlation between DNA levels and protein levels can be modest at best (1, 2). For this reason, a new technology called antibody microarrays has been developed to assess differential expression directly at the protein level.

Antibody microarrays can be used for expression profiling of hundreds of thousands of proteins simultaneously, with the goal of identifying disease/protein or protein/protein relationships. Effectively, any biological sample can be used to evaluate protein expression including cells, whole tissues, and bodily fluids. The setup of two-color antibody arrays is virtually equivalent to cDNA arrays. Two biological samples are hybridized to a single array using two distinguishable fluorescent dyes. After hybridization, the array is scanned and independent images are produced for each biological sample. The images are then analyzed to determine the antibody feature location and to quantify the relative fluorescence intensity of each feature.

Consequently, there is a wealth of information contained within a protein expression microarray experiment, but several factors must be considered throughout the experimental process to insure that the correct information is extracted. From a statistical point of view, these consist of (i) choosing an appropriate experimental design to answer the question of interest (i.e., careful sample selection, correct allocation of samples to the array, and adequate sample size and replication), (ii) implementing an appropriate normalization procedure that adjusts for experimental effects (e.g., arrays, dyes, and feature location) so that expression levels can be effectively compared across biological samples, (iii) assessing differential expression via statistical methods that are capable of distinguishing meaningful biological changes in protein expression from random noise, and (iv) tools for clustering and/or classification. These factors have had a major amount of research devoted to them with regard to cDNA arrays. Users of antibody array technology are therefore at a great potential advantage as they can directly use the statistical methods that have been developed for cDNA arrays. This article will discuss sound statistical methods that have been developed for two-color cDNA arrays that are directly applicable to two-color antibody arrays, many of which are an improvement over some currently used methods (e.g., refs. 3, 4). The relevance of each of the aforementioned points (i-iv) is discussed and references are provided to direct the reader to more in-depth discussions.

Experimental Design

At the onset of the cDNA microarray era, the *reference design* was used in which samples of interest were labeled with one fluorescent dye (e.g., Cy3) and a single reference sample was labeled with another fluorescent dye (e.g., Cy5) on every array (Table 1). To quantify the expression between the samples of interest and the reference sample, the log ratio of the Cy3 intensity to the Cy5 intensity was computed for every feature on the array.

As the cDNA microarray era progressed, it was noted that the reference design was not the optimal experimental design for all experiments. A commentary by Dobbin et al. (5) provides an in-depth discussion of experimental design considerations that have been proposed for two-color cDNA arrays that we suggest are also applicable to two-color antibody arrays. They state that the choice of experimental design should be dependent upon the overall objective of the experiment, the sources of variability in the system, and the labeling efficiencies between the two fluorescent dyes. With regard to a reference design, the log ratios eliminate variability at each feature that affects both dyes similarly. However,

Requests for reprints: Jeanette E. Eckel-Passow, Department of Health Sciences Research Mayo Clinic 200 First Street Southwest, Rochester, MN 55905. Phone: 507-538-6512; Fax: 507-266-2478; E-mail: eckel@mayo.edu.

Table 1. Experimental designs for two-color microarray experiments

Experimental design	Array 1	Array 2	Array 3	Array 4
Reference	A1/R	A2/R	B1/R	B2/R
Balanced block	A1/B1	B2/A2		
Incomplete block	A1/B1	B2/C1	C2/A2	
Loop	A1/B1	B1/A2	A2/B2	B2/A1

NOTE: Each cell contains the sample labeled with Cy3 followed by the sample labeled with Cy5.

it is now clear that Cy3 and Cy5 labeling differs in important ways resulting in intensity-dependent biases (6, 7). In addition to intensity-dependent biases, Kerr et al. (8) discovered a striking example of a gene-specific dye effect. That is, insulin-like growth factor II consistently produced larger expression values when hybridized with Cy3 regardless if it was associated with the control sample or the treated sample. At this point, it is unknown if protein-specific dye effects exist and thus an experimental design that allows for correct dye effect normalization is essential with the two-dye system.

Two other commonly used experimental designs in the cDNA array literature are the *balanced-block design* (5) and the *loop design* (9). The balanced-block design was inherited from agriculture experiments and in the current scenario each individual microarray is thought of as a "block." A balanced-block design for a two-class comparison would have samples from each of two classes hybridized to every array in such a way that the classes are balanced with respect to dyes. For experiments with more than two classes, the design is an *incomplete-block design* because not every class can be hybridized to every array because only two samples can be hybridized per array. However, even in an incomplete-block design, classes remain balanced with respect to dyes (Table 1).

As an alternative to the reference design and the block designs, Kerr and Churchill (9) proposed a *loop design* (Table 1). In the loop design, each biological sample is hybridized to two different arrays, each with a different dye. In comparison with the reference design, the block designs and the loop design allow for suitable dye effect normalization as a result of dye balance. Kerr and Churchill (9) show that the reference design is inefficient and can lead to confounding (e.g., class effects being confounded with a dye effect). To correct for the confounding of the class effect with the dye effect in the reference design, a dye swap can be done in which every array in the reference design is rerun with the dyes reversed. However, one obvious disadvantage of this approach is that it doubles the number of arrays and thus doubles the amount of resources needed. The reference design is also less efficient than the other designs because more information is collected from the reference group than each of the classes, although the reference group is not the primary class of interest.

Dobbin et al. (5) suggest that for class comparison studies, the balanced-block design is preferred, whereas the loop design may be superior for class discovery studies. A class comparison study compares two or more defined classes (e.g. tumor versus normal) to assess differential expression across classes. In a class discovery study, one seeks to discover new subgroups within a predefined class. In general, the two most important considerations in choosing an experimental design are the capability of answering the primary objective of the study as well as having desirable statistical qualities. Although important, factors such as efficiency should be judged on a case-by-case basis. As an example, consider a dose-response study that includes a vehicle group (dose = 0). In this scenario, it may be reasonable to collect more data on the vehicle group than each of the individual dose-concentrations because the vehicle group is the most appropriate group to compare each of the dose-concentrations. After doing so, the true effect of the dose-concentrations can be assessed independently of a vehicle effect. In summary, the choice of experimental design is not clear cut and largely depends on the primary objective of the study and deserves thorough consideration.

Replication and Sample Size

In general, the two types of replication commonly seen in gene expression or protein expression microarray experiments, are technical and biological replication. Technical replication refers to replicate measurements from the same biological sample, whereas biological replication refers to measurements from independent biological samples. It is imperative to understand the difference between technical and biological replication because with respect to generalizability, the utility of biological replication far exceeds that of technical replication. Appropriate biological replication helps to assure that the results can be generalized to the population from which the samples were obtained. If biological replication is replaced by technical replication then the corresponding results only apply to the samples at hand.

With regard to sample size, it is not sufficient to sample only one individual from each class nor is it sufficient to use only a single pooled sample from a set of subjects to represent a class of interest. Signal intensities across subjects from the same class are not expected to be equivalent because biological variability is inherent, as is measurement error. Thus, if differences exist when comparing only one subject from each class, expression differences due to class cannot be disentangled from simple biological variability or measurement error. Similarly, using a single pooled sample to represent a class is insufficient with regard to statistical models because biological and experimental variabilities are not separately estimable. In addition, it may be that an extreme outlier exists in the pool and this outlier can unduly influence the expression values obtained from the pool. If pooling is a necessity to obtain enough sample to hybridize to the array, then it is suggested that multiple independent pools be used instead of a single large pool. For example, if three subjects are needed to obtain enough sample to hybridize to a single array for a given class, then each array that contains this particular class should contain an independent set of three subjects that were pooled together specifically for that array. The take home point is that estimates of biological variability are necessary to determine whether differences between classes are actually biologically driven or spurious, and are thus crucial to array experiments.

How many biological samples are needed to accurately assess differential expression? The most general answer is that more is always better. For specific sample size calculations for various experimental designs, see Dobbin et al. (10). To obtain generalizable results, it is critical that the sample size consists of a set of samples that represents the population to which you want to

generalize the results. Likewise, in a class comparison study, a representative comparison group is also crucial. Thus, as is critical with any experiment, stratification with respect to all factors (biological and experimental) that could affect overall expression levels is highly recommended.

Normalization

To correctly assess differential expression across different classes of samples, it is important that systematic biases in the measured expression levels are eliminated. With regard to either cDNA or antibody arrays, systematic differences can arise due to experimental factors such as sample preparation, feature location, arrays, and fluorescent labeling efficiencies as well as various interactions of these effects. Thus, normalization is a critical initial step to balance the individual signal intensities across the experimental factors while maintaining the class effect of interest.

Recent articles using antibody array technology have used a normalization procedure referred to as an internally normalized ratio (INR; refs. 3, 4, 11). This normalization procedure was proposed for dye swap designs in which each sample of interest is represented on two arrays, each using a different dye. Thus, the sample of interest is labeled with Cy3 and a reference sample is labeled with Cy5 on one array, and on a second array, the same sample of interest is labeled with Cy5 and the reference sample is labeled with Cy3. For both arrays, the ratio of Cy3 to Cy5 is computed ($A_1 = \frac{S_{Cy3}}{R_{Cy5}}$ and $A_2 = \frac{R_{Cy3}}{S_{Cy5}}$) and the INR is the geometric mean of the two ratios ($INR = \sqrt{\frac{S_{Cy3}}{R_{Cy5}} \times \frac{S_{Cy5}}{R_{Cy3}}}$). Although the INR is capable of adjusting for a constant dye effect within each dye swap, it does not adjust for array effects, intensity-dependent effects that are often a result of the two-dye system (Fig. 1), or heterogeneous spreads in the data. Normalization procedures that account for these systematic effects should be considered.

Various normalization methods have been developed for cDNA arrays, including fixed effects ANOVA models (9), mixed ANOVA models (12), and more recently within-print tip local regression smoothing methods for the removal of systematic effects (6, 13). In general, if intensity-dependent biases do not exist, either the fixed effect linear model or the mixed model will suffice assuming that appropriate experimental effects are included in the model. However, if intensity-dependent biases do exist, which is often the case with the two-color system, the approach of Dudoit et al. (6) or Eckel et al. (13) is recommended. The modified MA plot in Fig. 1 depicts the systematic bias associated with estimating the true signal intensities of each feature in an unpublished antibody array experiment in which Ab Microarray (BD Clontech, Palo Alto, CA) chips were used, specifically chips from lots 3120314 and 4020056. Information regarding the chips is available at http://www.Bdbiosciences.com. Although the bias associated with the Cy5 dye seems to be constant and roughly centered at zero in Fig. 1, the bias associated with the Cy3 dye is noticeably nonlinear. If normalization were not needed for these data, the bias functions would be a horizontal line at $M = 0$.

One way to circumvent the nonlinear bias is via nonparametric local regression procedures (6, 13). That is, using a local regression procedure to estimate the bias functions as shown by the lines in Fig. 1 and subtracting it from the data. The important difference between the Dudoit et al. (6) and Eckel et al. (13) local regression approaches is that the Dudoit approach corrects for within-array intensity-dependent dye biases only (MA plots), whereas the Eckel approach corrects for within-array as well as across-array biases (modified MA plots). Regarding application, the Dudoit approach was developed with respect to a reference design and thus normalizes log ratios, whereas the Eckel approach is applicable to a more general class of experimental designs.

One last note with regard to normalization involves the concept of spatial or print tip effects. Each microarray is typically made up of a $p_1 \times p_2$ grid of print tips in which each grid contains $s_1 \times s_2$ features (often called spots) that are printed with the same group of print tips. It has been noted with cDNA arrays that spatial and/or print tip effects can exist as a result of spatial variation on the array, inconsistencies among the pens that make the array as well as inconsistencies in hybridization conditions across the array (7). Thus, it is recommended that normalization at the print tip level should also be considered, and this level of normalization can be added to either the ANOVA model or the local regression methods. With respect to an ANOVA model normalization procedure, a print tip effect is simply added to the model. In the local regression procedures, a local regression line is fit separately to each print tip on every array (6, 13). As an example, Fig. 2*A* displays a spatial effect. As a proxy for a spatial effect, each array was split into quarters and the Eckel normalization approach was implemented in which a local regression bias function was fit to each quarter of the array separately (upper left, upper right, lower left, and lower right quarter, respectively). Figure 2*A* displays an example where the bias associated with the lower half of the array (*dashed lines*) is systematically different

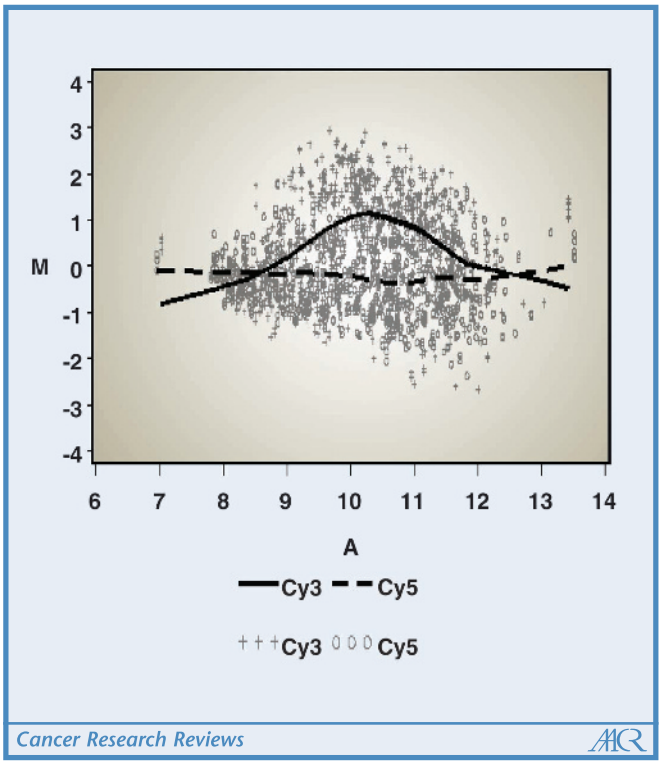

Figure 1. Modified MA plot as defined in the Eckel et al. (13) approach. If the data were perfectly normalized, the Cy3 and Cy5 bias functions would be at $M = 0$. Horizontal lines would suggest that a simple ANOVA normalization method would suffice; however, linear or nonlinear lines suggest intensity-dependent bias exists and a local regression procedure is the best normalization method.

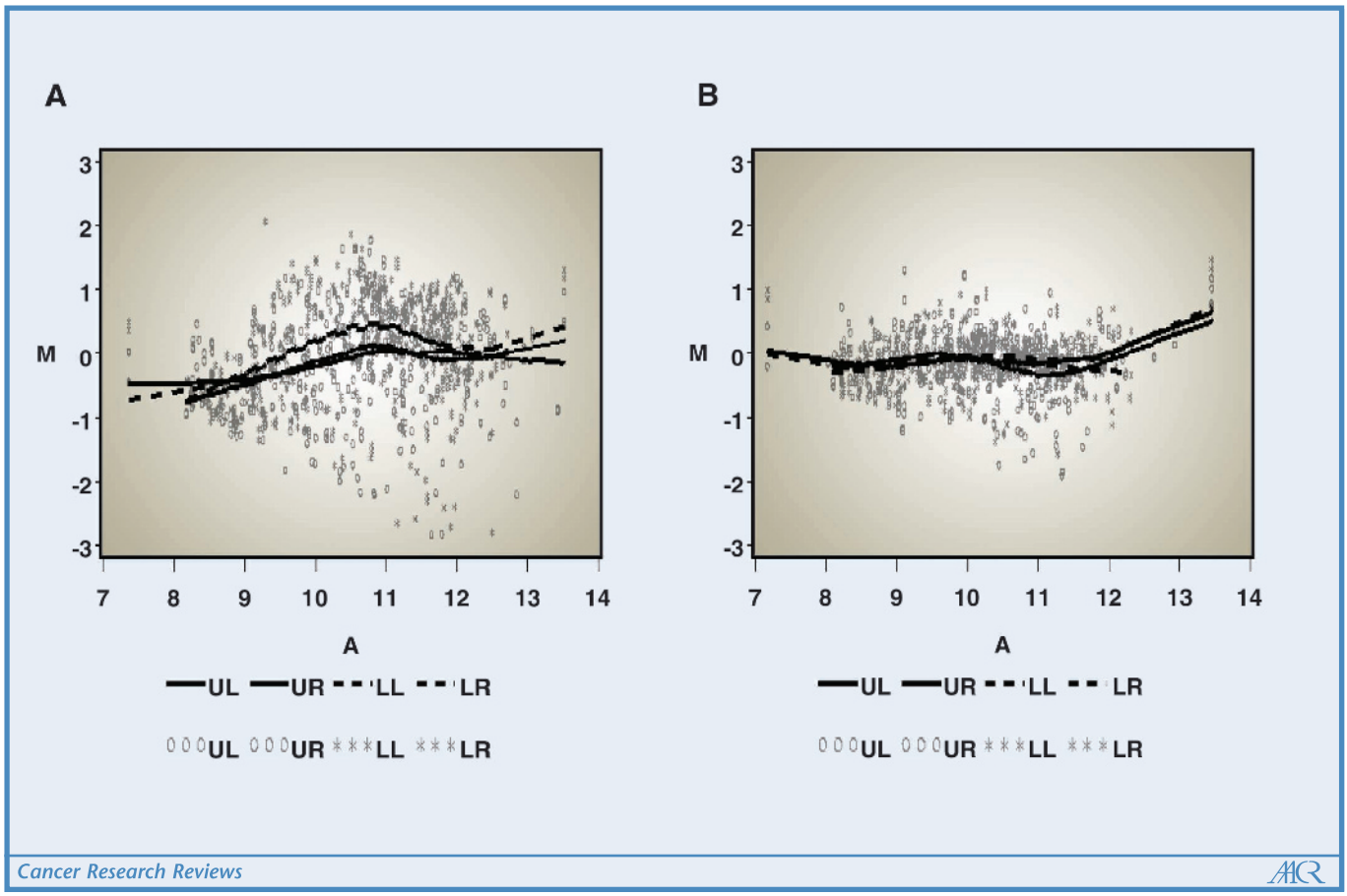

Figure 2. Modified MA plots of (*A*) an array that has a spatial effect and of (*B*) an array that does not have a spatial effect. Upper left (*UL*), upper right (*UR*), lower left (*LL*), and lower right (*LR*) quarter of the array.

than the upper half of the array (*solid lines*). As a point of reference, Fig. 2*B* displays an array that does not seem to have a spatial effect. Thus, MA plots at the print tip level are useful in determining which normalization procedure will best eliminate the bias associated the data at hand.

Assessing Differential Expression

The body of literature on statistical methods to determine differential expression in the gene expression microarray field is far too large to discuss in depth in this review. To mention a few, Kerr and Churchill (9) and Wolfinger et al. (12) discuss the use of fixed ANOVA and mixed effects models respectively, Efron et al. (14) and Newton et al. (15) developed novel empirical Bayes approaches, and Tusher et al. (16) developed a nonparametric approach. The main concept we want to address in this section is that looking at fold change between a sample of interest and a reference sample is not sufficient to determine significant differential expression (17). Fold change simply addresses how the mean is behaving and does not take into consideration the amount of variability in the measurements. Thus, a protein may exhibit a large fold change, but it may also be extremely unstable and thus unreliable. Therefore, statistical methods that take into account deviations in the mean as well as random variability in the system are significantly preferred (i.e., a signal-to-noise measure).

Classification and Clustering

Another application of gene or protein expression microarray technology is to sort the data by expression levels and to group together genes or proteins with similar regulation. The idea is that proteins/genes that group together might contribute to the same biological pathway. Clustering is an *exploratory* technique that can give valuable insight into underlying relationships that are otherwise not readily found in multidimensional data and thus can suggest interesting hypotheses concerning relationships (18). As an example, clustering of gene expression data has shown to be a valuable tool for classifying types and subtypes of cancer (19). Clustering methods can be used for class discovery or for the classification of known categories. A large amount of literature lists on the applications of clustering and classification methods to microarray data. Some nonmodel-based methods include hierarchical clustering, K-means clustering, self-organizing maps, and principal component analysis, whereas some model-based methods include classic linear and nonlinear models (20, 21). The literature is far too expansive for a comprehensive discussion here; however, different clustering techniques for microarray data are discussed in more detail by Azuaje et al. (22) and Stanford et al. (23) and a comparison of clustering techniques is discussed by Datta and Datta (24) and Huang and Pan (20). Clustering is an exploratory tool and thinking about what tool to use and particularly about the underlying distance metric is useful.

Summary

The objective of this review was to provide an overview of statistical concepts to consider when designing and analyzing protein expression microarray experiments. We discussed why some currently implemented techniques for normalization and analysis of differential expression are not recommended and in particular discussed the pitfalls associated with the INR and fold change analyses. In general, we suggest (a) implementing an experimental design that allows the primary question of interest to be answered as well as appropriate normalization of systematic effects; (b) implementing a normalization procedure that adjusts for array, dye, print tip, and intensity-dependent effects that are inherently included in the two-dye system to assure that the estimated class effects are biologically driven instead of reflecting systematic effects; (c) with regard to assessing differential expression, it is important to use a statistical model that is capable of distinguishing true signal from random noise; and (d) if class discovery is the goal, using several different clustering tools instead of a single cluster analysis to verify interesting findings.

Acknowledgments

Received 9/3/2004; revised 11/23/2004; accepted 12/23/2004.

Grant support: National Cancer Institute grant R25 CA92049 and the Multiple Myeloma Research Foundation (MMRF).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

The authors thank Lyle Burgoon and an anonymous reviewer for their informative comments that have improved this review.

References

1. Svingen PA, Loegering D, Rodriquez J, et al. Components of the cell death machine and drug sensitivity of the national cancer institute cell line panel. Clin Cancer Res 2004;10:6807–20.
2. Chen G, Gharib TG, Huang CC, et al. Discordant protein and mRNA expression in lung adenocarcinomas. Mol Cell Proteomics 2002;1:304–13.
3. Anderson K, Potter A, Baban D, Davies KE. Protein expression changes in spinal muscular atrophy revealed with a novel protein microarray technology. Brain 2003;126:2052–64.
4. Yamagiwa Y, Marienfeld C, Meng F, Holcik M, Patel T. Translational regulation of X-linked inhibitor of apoptosis protein by interleukin-6: a novel mechanism of tumor cell survival. Cancer Res 2004; 64:1293–8.
5. Dobbin K, Shih JH, Simon R. Questions and answers on design of dual-labeled microarrays for identifying differentially expressed genes. J Natl Cancer Inst 2003;95:1362–9.
6. Dudoit S, Yang YH, Callow MJ, Speed TP. Statistical methods for identifying differentially expressed gene in replicated cDNA microarray experiments. Statistica Sinica 2002;12:111–39.
7. Quackenbush J. Microarray data normalization and transformation. Nature Genetics Supplement 2002;32: 496–501.
8. Kerr MK, Leiter EH, Churchill GA. Analysis of a designed microarray experiment. Proceedings of the IEEE-Eurasip Nonlinear Signal and Image Processing Workshop, June 3-6, 2001, (http://www.jax.org/staff/Churchill/labsite/research/expression/leiter.pdf).
9. Kerr MK, Churchill GA. Experimental design for gene expression microarrays. Biostatistics 2001;2:183–201.
10. Dobbin K, Shih JH, Simon R. Statistical design of reverse dye microarrays. Bioinformatics 2003;19: 803–10.
11. Gosmanov AR, Umpierrez GE, Carabel AH, Cuervo R, Thomason DB. Impaired expression and insulin-simulated phosphorylation of Akt-2 in muscle of obese patients with atypical diabetes. Am J Physiol Endocrinol Metab 2004;287:E8–15.
12. Wolfinger RD, Gibson G, Wolfinger ED, et al. Assessing gene significance from cDNA microarray expression data via mixed models. J Comput Biol 2001;8:625–37.
13. Eckel JE, Gennings C, Therneau TM, Burgoon LD, Boverhof DR, Zacharewski TR. Normalization of two-channel microarray experiments: a semiparametric approach. Bioinformatics (October 28, 2004), doi:10.1093/bioinformatics/bti105.
14. Efron B, Tibshirani R, Storey JD, Tusher V. Empirical Bayes analysis of a microarray experiment. Journal of American Statistical Association 2001;96: 1151–60.
15. Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. J Comput Biol 2001;8:37–52.
16. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 2001; 98:5116–21.
17. Kutalik Z, Inwald J, Gordon SV, et al. Advanced significance analysis of microarray data based on weighted resampling: a comparative study and application to gene deletions in *Mycobacterium bovis*. Bioinformatics 2004;20:357–63.
18. Johnson RA, Wichern DW. Applied multivariate statistical analysis. Chapter 12. Clustering, distance methods, and ordination. New Jersey: Prentice Hall; 1998.
19. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531–7.
20. Huang X, Pan W. Linear regression and two-class classification with gene expression data. Bioinformatics 2003;19:2072–8.
21. Luan Y, Li H. Clustering of time-course gene expression data using a mixed-effects model with B-splines. Bioinformatics 2003;19:474–82.
22. Azuaje F, Bolshakova N. A practical approach to microarray data analysis. Chapter 13. Clustering genomic expression data: design and evaluation principles. London: Kluwer; 2003.
23. Stanford DC, Clarkson DB, Hoering A. A practical approach to microarray data analysis. Chapter 14. Clustering or automatic class discovery: hierarchical methods. London: Kluwer; 2003.
24. Datta S, Datta S. Comparisons and validation of statistical clustering techniques for microarray gene expression data. Bioinformatics 2003;19:459–66.