Experimental design, analysis of variance and slide quality assessment in gene expression arrays

Sorin Draghici*, Alexander Kuklin, Bruce Hoff & Soheil Shams

Address
BioDiscovery Inc
11150 West Olympic Boulevard
Suite 1170
Los Angeles
CA 90064
USA
Email: sorin@biodiscovery.com

*To whom correspondence should be addressed

Current Opinion in Drug Discovery & Development 2001 **4**(3):332-337
 ISSN 1367-6733

A microarray experiment is a sequence of complicated molecular biology procedures relying on various laboratory tools, instrumentation and experimenter's skills. This paper discusses statistical models for distinguishing small changes in gene expression from the noise in the system. It describes methods for assigning statistical confidence to gene expression values derived from a single array slide. Some of the theory is discussed in the context of practical applications via software usage.

Keywords ANOVA, gene expression, microarray, statistical models

Introduction

Pharmaceutical research employs various high-throughput technologies to assess responses of biological subjects to drug candidates. Among these new technologies, DNA arrays are unsurpassed as a tool to monitor gene transcription for thousands of genes at a time. The first step of this technique involves spotting known sequences on a substrate, which in most cases are glass slides (microarrays) or nylon membranes (macroarrays). This is followed by reverse transcription of mRNA isolated from the biological subjects under study into cDNA. During the process of reverse transcription, the control and the experimental materials are differentially labeled (in most cases), pooled and hybridized to the arrays. cDNA strands in this pool hybridize to complementary sequences on the array by competing for them. The relative abundance of the corresponding mRNA from the two sources will be assessed by the measured signal.

Every procedure in the DNA array methodology is a potential source of fluctuation [1••,2••] leading to a lot of noise in the system as a whole (Table 1). Understanding these sources is of paramount importance to the correct interpretation of results. The experimental design can offer the possibility to control the various sources of variability and ultimately produce estimates of relative expressions and errors. The objective of this short article is to discuss some recent approaches in microarray experimental design and statistical confidence in array data analysis.

The objectives of microarray experiments are to reveal unknown genes and new gene functions as a result of experimental treatments, to find new gene expression patterns and use them as a basis for classification of physiological or pathological processes. Traditionally, DNA array experimenters have been reporting genes with large changes in expression (> 2 or 3 standard deviations). However, genes with small, but reproducible changes are also of considerable interest and these genes cannot be identified without appropriate statistical methods. To select these genes, researchers need to assign measures of statistical confidence to the results. The first question addressed is how to distinguish the small changes in gene expression from the noise in the system. A second important issue addressed here is how to assess the quality of a particular slide based solely on the information contained within. This is particularly important when repeating slides is not possible (eg, due to a limited quantity of mRNA) or is not desirable (eg, due to a high experimental cost).

Experimental design

Experimental design methods are used to develop processes affected minimally by external sources of variability [3••]. Such robust methods are important in DNA array technology in order to both minimize experimental costs and yield high-quality data.

Table 1. Sources of fluctuations in a microarray experiment [1••,2••].

Factor	Comments
mRNA preparation	Tissues, kits and procedures vary.
Transcription (RT-PCR)	Inherent variation in the reactions, type of enzymes used.
Labeling	Depends on the type of labeling and procedures, as well as the age of the labels.
Amplification (PCR protocol)	PCR is difficult to quantify.
Pin geometry variations	Different surfaces and properties due to production random errors.
Target volume	Fluctuates stochastically even for the same pin.
Target fixation	The fraction of target cDNA that is chemically linked to the slide surface from the droplet is unknown.
Hybridization parameters	Influenced by many factors such as the temperature of the laboratory, time, buffering conditions and others.
Slide inhomogeneities	Slide production parameters, batch-to-batch variations.
Non-specific hybridization	A typical source of errors.
Non-specific background	Bleeding from neighboring spots or artifacts.
Image analysis	Non-linear transmission effects and variations in spot shape; contaminants are not removed from the analysis.

Historically, gene expression researchers have employed clustering methods as visualization tools to identify interesting gene expression patterns, while traditional statistical analysis methods were initially left in oblivion. Kerr and Churchill [4] observed that the classical statistical analysis approaches of the grandfather of statistics, Fisher, to analyze data from agricultural experiments in the field is very much applicable to array experiments. Crop scientists study yields of different crop varieties, but they can only directly compare data for varieties grown in the same field. When the fields are inadequate or insufficient, varieties have to be grown in different lots and subsequently compared. Additionally, conditions for different crops cultivated even in the same field are not the same. This situation resembles the typical array experiments. Gene expression values are compared across different slides and sections of a slide / array can be different due to hybridization problems.

This is where the experimental design comes into play. The three basic principles of experimental design are replication, randomization and blocking. Replication has two very important properties. Firstly, replication allows the researcher to calculate an estimate of the experimental error, which is necessary to determine if differences in the data are statistically different. Secondly, when the sample mean is used to estimate the effect of a factor in the experiment (see Glossary) replication provides better accuracy in estimating this effect.

Replication is a widely misunderstood term. In a strict etymological sense, to replicate means to perform the same task more than once. Often, the misunderstanding is related to the definition of the task to be performed. Thus, if the purpose is to understand and control the noise introduced by the location of the spot on the slide, one should replicate spots by printing exactly the same DNA at different locations on the same slide. If the purpose is to understand and control the noise introduced by the hybridization stage for instance, one should print several exact copies of a given slide (with all other parameters and DNA sources exactly the same) and hybridize several times with exactly the same mRNA in exactly the same conditions. Finally, if the purpose is to control the biological variability, different mRNA samples should be collected from similar specimens and the microarray should be used in exactly the same conditions from all other points of view. The common misunderstanding is related to the fact that often researchers refer to replicates without specifying which one factor was varied while keeping everything else constant. Even more misleading, sometimes the term 'replicates' is used to describe results obtained by varying several factors at the same time. For instance, combining two different expression values of the same gene obtained on different arrays, in different hybridization conditions and perhaps with different mRNA, will certainly contain more information than a single expression value, but can hardly be treated as a replicated experiment. The argument presented here refers to replicate spots on the same array. In other words, the design of this type of experiment requires that the same DNA be printed at different locations on the same array.

Replication of DNA sequences by printing them adjacent to one another has been used in the field, but is not always the best choice because of overshining (bleeding) occurrences that may eliminate data from two neighboring spots if their corresponding gene is highly expressed. Furthermore, printing all replicates of a gene in a limited area on the array means that if that particular area is affected by a local problem (eg, washing), all spots corresponding to the given gene will be affected and no reliable information about this gene will be available. It is better to distribute the spots randomly on the entire surface of the array. When spots are replicated on the same array, the source of variability is only the location of the spots on the slide and not the variability among slides. Lee and colleagues [5] demonstrated that by replicating the array three times on the same slide and combining the data, false positives and false negatives can be reduced considerably. Their work shows that even a single source of variability such as spot location can lead to considerable variation of the expression levels extracted from the array.

We have developed a software program that allows the user to: (i) design *in silico* arrays including spot replicates; (ii) keep track of the clones; and (iii) store the information associated with the experimental design in a database [6•]. Clones from all PCR plates used in the experiment that are printed (spotted) by one pin form a block called a 'sub-grid'. Clones can be randomly printed in replicates in different sub-grids. Alternatively, as Lee *et al* show, the whole array can be printed repeatedly on the same slide (if space permits) thus forming a super-grid [5]. The array by itself is a super-grid with 1 row and 1 column. A 3-fold replicated array will have 3 rows and 1 column. Advanced image analysis software packages (eg, ImaGene™ [7]) can perform the image analysis of arrays with a complex super-grid structure. For statistical analysis of data with replicates, the same gene ID or gene IDs with similar substrings in the name should be used for replicated spots. This enables the data analysis software packages to identify replicates and automatically perform the computational procedures involved in combining the replicates and calculating confidence measures based on such replicates.

In a multi-channel experiment, microarrays are probed with differently labeled cDNA samples. The most frequently used experiment involves two channels, which are usually labeled with Cy3 and Cy5 dyes. The same spot is used to assess the expression of a gene in both the control and the test samples. In this way, homogeneous material is employed, which is a prerequisite for blocking experimental designs. Kerr and Churchill [8] note that arrays can be considered as experimental blocks with two dye colors. The paired comparison design may be used to compare two categories of a factor by using spot replicates on the same array [2••]. This design is a subclass of a more general type of design called randomized block design. The block represents a restriction on complete randomization because the treatment combinations are randomized within the block. However, when there are more than two categories of the factor of interest, not every category (variety) can be present on each array. Then the design is termed incomplete block design.

In each step of a microarray experiment, researchers encounter various factors contributing to variability of the data (Table 1). These factors, however, may interact with

each other. The objective is to find the combination of factors that will lead to an optimal system with minimal variability and maximal undistorted responses. Factorial designs have been very popular in studies where it is necessary to find the joint effect of the factors on a response. Wildsmith *et al* [2] reported that in their microarray system, the probe preparation step is a major contributor to the overall experimental variability. This step depends on several factors, such as: (i) the type of fluorescent label; (ii) the age of the fluorescent label; (iii) the type of enzyme; (iv) the age of the deoxynucleotide triphosphates (dNTPs); (v) the incubation time; and (vi) the type of RNA, ie, total or poly(A) mRNA. In this case, there are six factors (k = 6) and two levels. This experimental design is known as a 2^k factorial design and it has always been useful in early stages of experimental work where there are many factors involved in the process. An experiment using such an approach requires 2^k observations in general. In the example above, the experiment will require $2^6 = 64$ runs. Wildsmith *et al* decided to study the array-to-array variability within each factor combination. In order to achieve this they used 32 runs instead of 64, with two microarrays for each run. This design allowed them to estimate all factor effects and all two-factor interactions. The enzyme type, label type and RNA type were found to be the most important factors among the six considered.

ANOVA and microarrays

Kerr and Churchill [4] discuss an approach for gene expression data analogous to the approach used by Fisher to analyze the crop yield data. In order to compare several crop varieties grown on different blocks, Fisher used V_j to designate the effect of variety j, B_i to designate the effect of block i, and μ to designate the overall mean. According to Fisher, one can simultaneously estimate the relative yield of crop varieties and the relative effects of the blocks of land using the following linear model:

$$y_{ij} = \mu + B_i + V_j + \varepsilon_{ij} \quad (1)$$

Here, ε is the random error in the experiment. The objectives are to find the variety with the highest yield by testing appropriate hypotheses and accept or reject them on a statistical basis. For hypothesis testing, we assume that the model errors are independently distributed random variables with a mean of zero. Their variance σ^2 is assumed to be constant for all levels of the factors.

Microarray experiments may involve multiple arrays to compare multiple samples. In order to account for the multiple sources of variation in a microarray experiment, Kerr and Churchill [3••] used the analysis of variance (ANOVA) approach and proposed the following model:

$$\log(y_{ijkg}) = \mu + A_i + D_j + V_k + G_g + (AG)_{ig} + (VG)_{kg} + \varepsilon_{ijkg} \quad (2)$$

In this model, μ is the overall mean signal of the array, A_i is the effect of the i^{th} array, D_j represents the effect of the j^{th} dye, V_k represents the effect of the k^{th} variety, G_g represents the variation of the g^{th} gene, $(AG)_{ig}$ is the effect of a particular gene on a given array, $(VG)_{kg}$ represents the interaction between the k^{th} variety and the g^{th} gene and ε_{ijkg} represents the error term for array i, dye j, variety k and gene g. The errors are assumed to be independent and of zero mean.

The approach presented in this paper is different and has a 2-fold purpose. Firstly, we would like to estimate the amount of noise on a single slide and be able to use this knowledge in order to distinguish between genes that are truly upregulated and genes that may only appear to be so due to the noise on the slide. The currently-used approach to identifying up- or downregulated genes is to choose *a priori* a threshold on the expression values distribution. For instance, upregulated genes are often considered to be those genes that have an expression level that is beyond two standard deviations from the sample mean. The implicit assumptions made here are that: (i) the expression levels are normally distributed; and (ii) the noise variance is much lower than the variance of the expression level distribution. However, these assumptions often do not hold. For instance, the distribution of the expression values presented in Figure 1 is bimodal, asymmetric and can hardly be approximated by a normal distribution. Furthermore, if the noise variance happens, as it may, to be much larger than the variance due to differential regulation, then taking two standard deviations as the definition of differentially regulated genes does not ensure that the genes are truly differentially regulated as opposed to simply being affected by the noise. The approach presented here is aimed at being able to make exactly this crucial distinction using information on the noise distribution collected through the use of replicates.

The second purpose of our approach is to give the laboratory researcher some feedback about the quality of the slide. Such feedback is essential in improving the laboratory protocols towards obtaining an overall good quality for the whole process. An accurate estimation of the noise on one particular slide may serve as such feedback, allowing the researcher to understand important factors in the overall process. In order to do this, we have modified the Kerr-Churchill statistical model as follows:

$$\log R(g,s) = \mu + G(g) + \varepsilon(g,s) \quad (3)$$

where logR(g,s) is the average log ratio over the whole array, G(g) is a term for the differential regulation of gene g and ε(g,s) is a zero-mean noise term. In this model, one can calculate the following estimates:

$$\hat{\mu} = \frac{1}{n \cdot m} \sum_{g,s} \log(R(g,s)) \quad (4)$$

$$\hat{G}(g) = \frac{1}{m} \sum_{g} \log(R(g,s)) - \hat{\mu} \quad (5)$$

where $\hat{\mu}$ is the estimate of the average log ratio, $\hat{G}(g)$ is the estimate of the effect of gene g, m is the number of replicates and n is the number of genes.

Using the estimate above, one can now calculate an estimate of the noise as follows:

$$\hat{\varepsilon}(g,s) = \log(R(g,s)) - \hat{\mu} - \hat{G}(g) \quad (6)$$

Note that the only assumption made here is that the noise has zero mean. This is a very reasonable assumption since we are referring to the noise that affects various spots within the same slide. No particular shape (such as Gaussian) is assumed for the noise distribution, which makes this approach very general and very powerful.

The proposed approach allows us to calculate an empirical distribution of the noise. This distribution can be used to achieve the two purposes mentioned above. In order to find the genes that are truly differentially regulated for a given confidence level, it is sufficient to calculate the deviation (from the mean) of the empirical distribution that corresponds to the given confidence level. In order to avoid the use of any particular statistical model, this is done by numerically integrating the area under the graph of the empirical distribution. The deviation thus calculated is then applied to the histogram of the gene expression values distribution (see Figure 1). In mapping the confidence interval to the gene expression levels, there is a scaling factor of:

$$\frac{1}{\sqrt{m - 1}}$$

where m is the number of replicates.

Let us assume that the deviation from the mean of the empirical distribution corresponding to the 5% confidence level is x. Let G be the mean of the expression values distribution. The genes that have expression values that fall above

$$G + \frac{x}{\sqrt{m - 1}}$$

have less than 5% probability of being there because of the noise (only 5% of the noise values fall more than x from the noise mean). In other words, such genes are upregulated with a confidence level of 95% (see Figure 1). A similar reasoning can be applied when looking for downregulated genes or for genes that are up- or downregulated with at least a specific-fold difference (see Figure 2). The second purpose mentioned above is also achieved because the width of the empirical noise distribution (eg, the area corresponding to 90% of the noise samples in Figure 1 and Figure 2) is inversely proportional to the quality of the slide (a cleaner slide means less noise, which means a narrower distribution).

The proposed model has been implemented in the data mining tool GeneSight [10] (Figure 3). This tool allows the user to select which type of regulation is of interest (up or down), as well as to continuously vary the confidence level while monitoring the number and distribution of the genes that are differentially regulated at the current confidence level.

Figure 1. Using the empirically calculated noise distribution to calculate the confidence level for up- and downregulated genes.

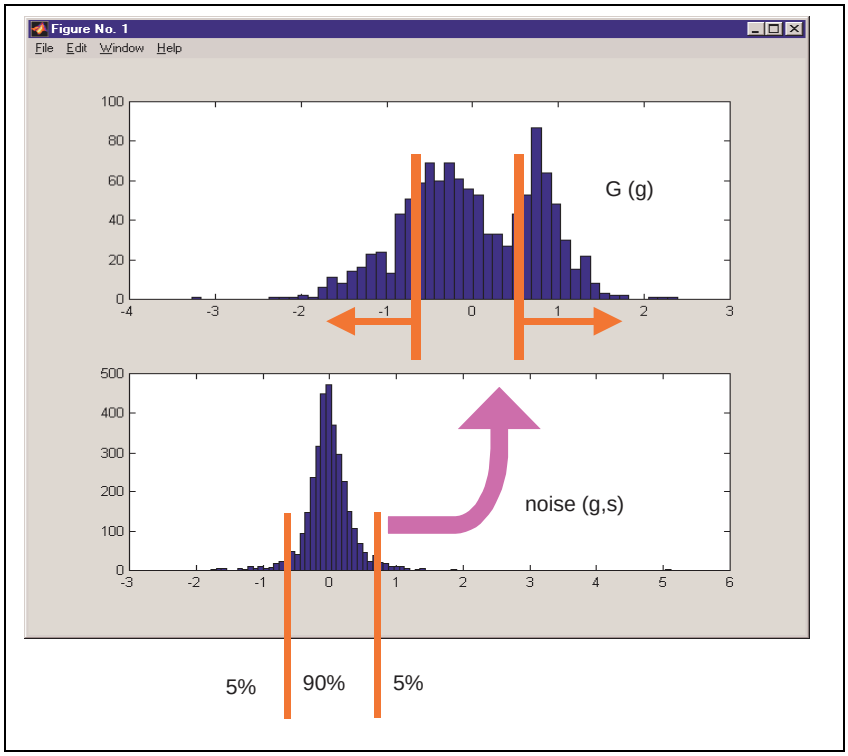


Figure 2. Using the empirically calculated noise distribution to calculate the confidence level for upregulated genes by a given fold in difference: The confidence interval is centred around the required fold difference and gene to the right of the interval selected.

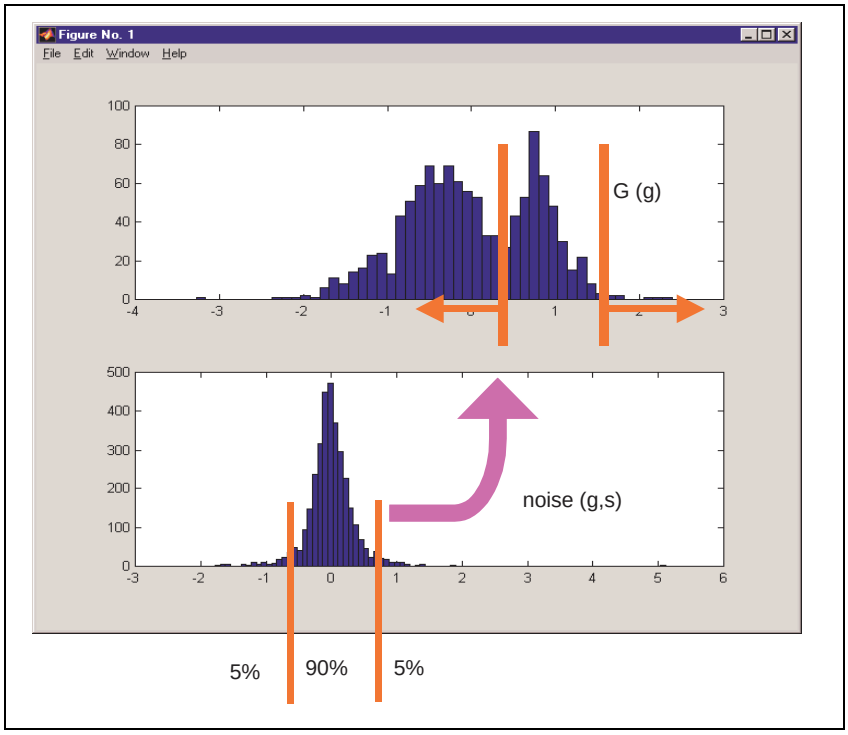
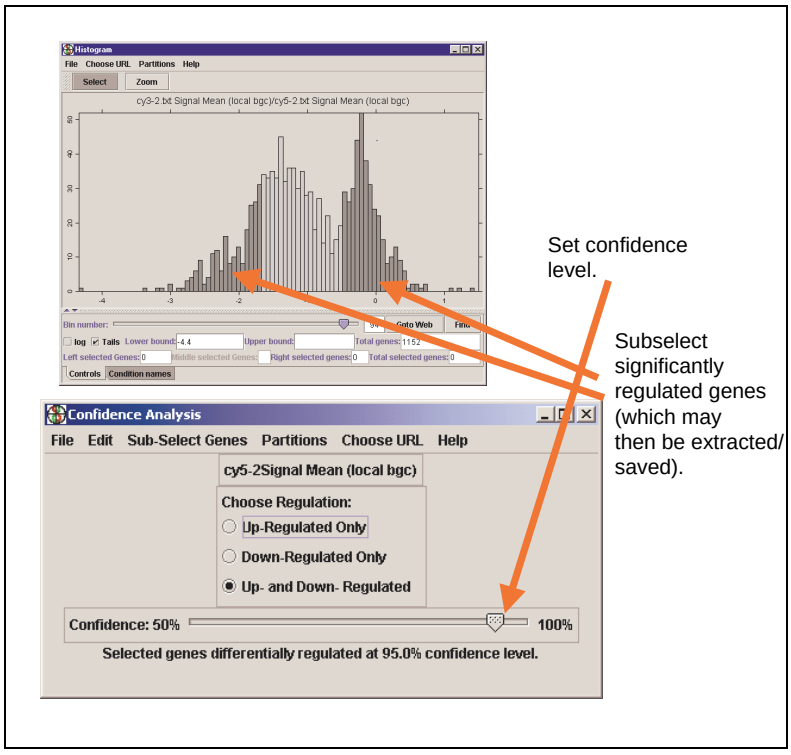


Figure 3. The implementation of the technique in GeneSight 2.0.



Conclusions

In this paper we have discussed issues related to the experimental design and interpretation of microarrays, focusing on the use of the ANOVA approach in order to achieve two objectives, namely: (i) the ability to differentiate between genes that have different expression values due to random fluctuations in the process and genes that are truly differentially regulated; and (ii) provide the experimental researcher with a feedback measure of the quality of the experimental process and the amount of noise introduced at the slide level. Being able to differentiate between noise and genuine differential regulation up to the resolution allowed

by the random factors involved in the process, is proposed as a better alternative to considering a pre-determined threshold on the expression values affected by the noise. The method proposed uses replicate spots on a given slide to estimate an empirical distribution of the noise at slide level. Given this distribution and a chosen confidence level, one can establish which genes are differentially regulated beyond the influence of the noise. Furthermore, the width of the noise distribution can be used as a quality measure for the noise introduced at the slide level. In turn, this can be used as a feedback to allow the fine-tuning of the laboratory protocols used to produce the slides.

Glossary

The main terms used throughout are briefly described in the following:

Blocking A technique in experimental design used to increase the precision of an experiment. A block is a portion of the experimental material that is more homogenous than the whole material. A microarray slide probed with two dyes is considered as a block.

Categories of factors These are also known as varieties, levels or treatments of the factor. For example, different concentrations of a drug used to treat tissue cultures, different time points, etc.

Factor effects The difference between the observed signals at the two factor settings in a 2^k factorial design.

Factorial design 2^k factorial A design that uses 2 x 2 x 2 x 2 ·· ··x 2 = 2^k observations on k factors, each one at two levels.

Factors Groups of experimental entities. Factors can be strains of living organisms (mice, yeast, etc), viruses, tissue types or lines of tissue cultures, drug treatments and others.

Replication Repetition of the basic experiment in identical conditions. In microarray technology, printing of the same array design several times on the slide yields spot replicates. Printing the same configuration on different slides yields slide replicates.

Statistical design of experiments The process of planning experiments such that they yield suitable data for analysis with statistical methods (models).

Variety A category of the factor under study.

References

•• of outstanding interest
• of special interest

1. Schuchhardt J, Beule D, Malik A, Wolski E, Eickhoff H, Lehrach H, Herzel H: **Normalization strategies for cDNA microarrays.** *Nucleic Acids Res* (2000) **28**(10):e47.

•• *This paper focuses on sources of noise in microarray experiments. The authors propose theoretical and experimental strategies for improving quantative evaluation of cDNA microarrays.*

2. Wildsmith SE, Archer GE, Winkley AJ, Lane PW, Bugelski PJ: **Maximization of signal derived from cDNA microarrays.** *BioTechniques* (2000) **30**:202-208.

•• *This article describes the use of factorial design of experiments to assess the effects of different experimental factors in the system under study. It is of particular interest to researchers trying to identify sources of variation in their experimental set-up before they commit to a large experiment.*

3. Kerr MK, Martin M, Churchill GA: **Analysis of variance for gene expression microarray data.** *J Comput Biol* (2001) in press.
http:// www.jax.org/research/churchill/pubs/index.html.

•• *Presents applications of ANOVA in microarray data analysis. The authors discuss statistical models for array data analysis.*

4. Kerr MK, Churchill GA: **Statistical design and the analysis of gene expression.** *Genet Res* (2001) in press. *http://www.jax.org/research/churchill/pubs/index.html* 2000.

5. Lee M-LT, Kuo FC, Whitmore GA, Sklar J: **Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations.** *Proc Natl Acad Sci USA* (2000) **97**(18):9834-9839.

6. Kuklin A, Smith J, Chen J-Y: **DNA experimental design and databasing with CloneTracker.** *BioDiscovery Inc Manual* (2001).

• *This booklet gives examples of how to achieve replicates of spots and arrays with the software CloneTracker designed to achieve virtual design, clone tracking and databasing of array experiments.*

7. **ImaGene™**: *BioDiscovery Inc*, Los Angeles, CA, USA. *http://www.biodiscovery.com/products/imagene/imagene.html*

8. Kerr MK, Churchill GA: **Experimental design for gene expression analysis.** *Biostatistics* (2001) in press. *http://www.jax.org/research/churchill/pubs/index.html.*

9. Montgomery DC (Ed): *Design and Analysis of Experiments*. John Wiley & Sons, New York, NY, USA.

10. **GeneSight™**: *BioDiscovery Inc*.
http://www.biodiscovery.com/ products/genesight/genesight.html