Using ANOVA to Analyze Microarray Data

Gary A. Churchill

The Jackson Laboratory, Bar Harbor, ME, USA

Microarray data can be interrogated using analysis of variance (ANOVA), a powerful and general method of data analysis that has been extensively developed and studied for more than 75 years (1). ANOVA provides an integrated approach to normalization, estimation of expression levels, and testing for differential expression (2). This purpose of this article is to introduce some of the essential elements that one should be aware of when applying ANOVA to microarray data.

Consider a simple experiment in which RNA samples from twelve mice, four individual mice from each of three different strains, are assayed using microarrays. The purpose of the experiment is to identify genes that differ in expression levels among these strains. One could compare the expression level of each gene using Student's *t*-tests to make three comparisons, one for each different pair of strains. However, this is not an efficient approach, because it does not fully utilize all of the information available in the data. An alternative strategy considers the variability of the expression levels within and among strains. If the variability in the expression levels of a gene among strains is substantially greater than the variability within strains, this indicates that the gene is differentially expressed. The comparison of variation at different levels in a set of measurements is the essence of ANOVA.

Experimental Design

ANOVA provides a method of data analysis that is motivated by consideration of the experimental design. The design of an experiment should be determined by the scientific question that is being addressed and be balanced by the practical constraints of the experimental system. The method of analysis of the data should follow naturally from the design and should directly address the question that motivated the experiment.

Replication is an essential feature of experimental design that allows us to draw statistically valid conclusions. In our simple example, Mouse is replicated within each Strain to obtain an estimate of mouse-to-mouse variability. The type of replication used in an experiment has important implications for the types of inferences that one can make. If we had measured only a single mouse per strain and compared strain variances to the variance between replicated spots within each microarray, we could dramatically overstate the significance of any differences among strains because we fail to account for variation between individual mice within a strain. The design of microarray experiments has been the subject of several review papers (3,4).

The real power of ANOVA is most apparent in the analysis of multiple factor experiments. A factor is a group of treatments or conditions. In our simple example, Strain is a factor with three levels. Multiple factor experiments, in which the effects of several variables are interrogated simultaneously, are often more efficient and more comprehensive than a set of single factor experiments. We will consider an extension of our simple example that adds a factor Diet with two levels, high-fat and low-fat, as shown in Figure 1.

Data Preprocessing

Some preprocessing of the raw intensity data is required prior to ANOVA analysis. It is natural to think of the effects in a microarray experiment as being multiplicative. For example, doubling the amount of labeled cDNA in a hybridization reaction will double the signal intensity over a wide range of absolute amounts of cDNA. A logarithmic transform will convert these multiplicative effects into additive effects. But additive effects (prior to taking the log) are also present in raw microarray data (5), usually in the form of background fluorescence. Local background adjustments often introduce bigger problems than the ones they correct. Transformations that are routinely applied to microarray data are logarithm-like, but they also remove some of the systematic effects in the raw intensity values including those due to additive background (6–8).

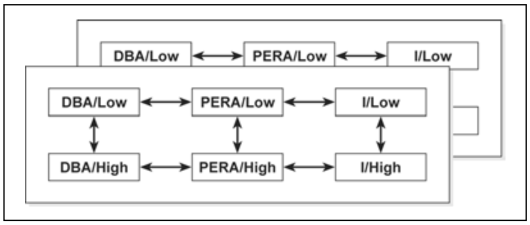

Figure 1. A microarray experiment design for a two-factor experiment. The first factor is Strain with levels DBA, PERA, and I. The second factor is Diet with levels Low and High. Liver RNA samples from individual mice were assayed on a two-color microarray platform using direct "dye-swap" comparisons as indicated by the double-headed arrows. The entire experiment was run twice using samples from a total of 12 individual mice and 28 microarrays. Each microarray has two color channels, thus each gene generates a total of 56 data points.

Microarray Analysis of Variance Models

An ANOVA model expresses the intensity value as a sum of components. Following Wolfinger et al. (9) we will define the ANOVA model in two stages: (*i*) the normalization model and (*ii*) the gene-specific model.

The normalization model removes effects due to overall variations in brightness between different arrays and dyes. ANOVA normalization is trivial. One simply subtracts the mean of the log-transformed intensity from each color channel on each array. We refer to transformed and normalized intensity values as Y.

In the second stage, we consider the data one gene at a time. For the experiment in Figure 1, Y represents the 56 data points associated with a given gene. The ANOVA model decomposes each element of Y into a sum of terms representing the contributions (effects) of different factors to the observed intensity. The terms in an ANOVA model include design factors and treatment factors. Design factors are related to the structure of the experiment and, for two-color arrays, should include the gene-specific Array (1,…, 28) and Dye (R, G) effects. Treatment variables define the conditions that apply to the individual RNA samples. In our example, these include Strain (DBA, PERA, I), Diet (High, Low), and Mouse (1,…,12). Levels of each factor are listed in parentheses.

An important distinction is drawn between model terms that are fixed and those that are random. To determine whether a term should be fixed or random, imagine a repetition of this experiment. If the effects would be the same in the repeated experiment, the term is fixed. In our example experiment, Strain, Diet, and Dye are treated as fixed effects. If the effect of the high-fat diet is to increase expression of a particular gene, we should see a similar increase in a repetition of the experiment. Random terms, on the other hand, would have different effects in the repeated experiment. A new set of arrays and new mice will vary in ways that we cannot predict from the original experiment. Random effects are sources of variation and should be considered as such when constructing test statistics. This helps to ensure that the conclusions we draw are repeatable in other contexts. Failure to recognize random effects in an experimental design can lead to misleading test results.

Now we can write an ANOVA model for our example as:

$$Y = \mu + \sim\text{Array} + \text{Dye} + \text{Strain} + \text{Diet} + \sim\text{Mouse} + \varepsilon \qquad \text{[Eq. 1]}$$

where Y, representing the log-transformed intensity, is decomposed into a sum of effects. The symbol μ represents the overall average intensity of all spots associated with the gene. The symbol ~ is used to denote the random terms. The random Array effect captures the variations in spot size for a given gene. The Dye term accounts for systematic gene-specific effects of each dye (10). The terms Strain and Diet capture gene expression changes that are attributable to the experimental conditions. The random term Mouse captures variability between individual mice within the same condition. The final term in the model, ε, represents measurement error. It is always included in an ANOVA model and is implicitly a random effect. These last two terms, Mouse and error, represent the variance to which the effects of experimental conditions are compared in the ANOVA F-test.

Different ANOVA models may be considered for the same set of experimental data. For example, if we suspect that the effects of Diet on gene expression may be Strain-dependent, we can include an interaction term:

$$Y = \mu + \sim\text{Array} + \text{Dye} + \text{Strain} + \text{Diet} + \text{Strain} * \text{Diet} + \sim\text{Mouse} + \varepsilon$$

where interaction between two factors is denoted by the * symbol. Much of the power of ANOVA analysis derives from the flexibility of fitting and comparing different models to the data.

Relative Expression Values

One useful feature of ANOVA is that it provides estimates of the effects on expression associated with each term in the model. The matrix algebra of ANOVA neatly extracts

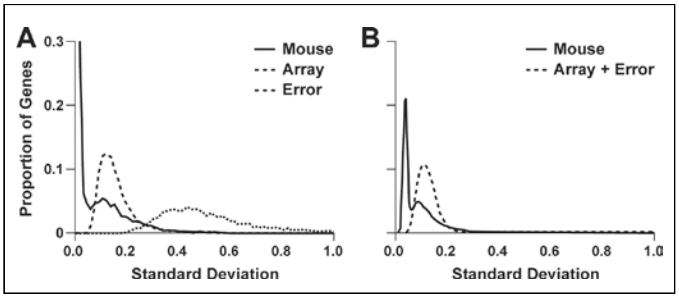

Figure 2. Variance components estimated using mixed model ANOVA. (A) Smoothed histograms of three variance components estimated for each of approximately 15,000 genes on a two-color microarray platform. The x-axis shows the square root of the estimated variances, and the y-axis indicates the proportion of genes with estimated variance within a sliding window. Although Array variance is the largest component, it does not substantially impact the mixed model F-test due the pairing of samples on each array. (B) The estimated variance components from the same samples assayed on a one-color array platform. With the one-color system, these two variance components (Array + error) are combined and both are included in the denominator of the F statistic. This underscores the importance of tight control in the production of one-color array platforms. It is of interest to note the consistent bimodal shape of the Mouse variance component. The majority of genes show little variation from mouse to mouse, yet a small proportion of all genes (approximately 10%–15%) show high levels of variation between individual but genetically identical mice.

quantities, such as the strain averaged effect of diet on the expression level of a gene. The effects of Diet are represented by two values, one for the high-fat level and one for the low-fat level of the factor. These numbers have little meaning by themselves, but the difference between relative expression values can be interpreted as the logarithm of the ratio of expression levels in the two conditions. In our second ANOVA model, we allowed for strain-specific Diet effects by including an interaction term. Now we obtain six relative expression values, one for the high-fat condition and another for the low-fat condition separately for each of the three strains.

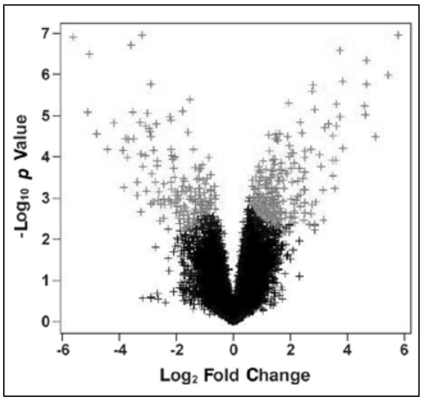

Figure 3. A volcano plot for the test of diet effect. The x-axis of the volcano plot shows the relative expression difference between low-fat and high-fat diets averaged across the three strains of mice. The y-axis shows the (unadjusted) significance of the gene-specific F-test. Each gene is represented as an individual point (+) on the graph. The red color indicates genes that were declared to be differentially expressed at the 0.05 FDR-adjusted significance level using the F-test with shrunken variance components (F_s in Reference 16).

Variance Components

Quantifying variation that arises from multiple sources is essential to the proper interpretation of microarray data. This is especially true of two- and multicolor systems but can also be relevant to one-color experiments. Variations can arise due to effects that are specific to spot-level data, whole arrays, and to the biological units (individual mice) in an experiment. In a simple ANOVA model, there is only one source of variation, the error term. For many experimental designs this "fixed effect" model is adequate. If there are multiple sources of variation, the ANOVA model is called a "mixed model" because it contains a mixture of fixed and random effects. The construction of F statistics using mixed model ANOVA properly accounts for the contributions of each of these sources of variation.

Variance components are estimated as part of the mixed model analysis, with one component per random effect per gene. Knowledge of the variance components can contribute to quality control and process improvement and can provide guidelines for sample size determination (11,12). The observation that some genes are more variable from animal to animal than others is a common feature seen in gene expression data, and it underscores the fact that each gene may have its own individual variance (Figure 2). Allowance for individual variance components for each gene is one advantage of the two-stage formulation of ANOVA.

Detecting Differential Expression

In ANOVA, the significance of the individual model terms can be tested. Thus we can determine, for any given gene, if the level of expression is altered by Diet or if it varies among Strains. Conducting a test for a term in an ANOVA model will generate a list of F statistics, one for each gene on the array. The significance (*p* value) of an F statistic can be assessed by reference to the standard F-distribution tables or by the analysis of many permuted versions of the original data. It is useful to compare the statistical significance of a gene to the magnitude of the expression change. The volcano plot (Figure 3), a scatter plot of the relative expression values against the *p* value for each gene, is one of the most useful tools for interpreting the results of ANOVA tests. Points in the extreme upper left and right corners of the volcano plot show the largest and statistically most significant changes in expression.

The most unique and challenging feature of microarray data is the multiplicity of probes. A typical microarray may contain more than 20,000 unique probes. The challenge here is to balance the two types of errors that occur when classifying genes as differentially expressed or not (13). We want to detect as many real differences as possible while maintaining control over the rate of false detection. We may choose a stringent but arbitrary level, such as 0.001, for the per-gene error rate in testing and acknowledge that this proportion of the nondifferentially expressed genes may be falsely declared significant. At the other extreme, we can control the probability of making any false detection across the entire set of genes by using a familywise error rate method such as the Bonferroni correction. However, the most popular and appealing approaches control the false discovery rate (14) to generate a list of differentially expressed genes such that the expected proportion of false detections on the list is bounded at a user-defined level.

Combining Data Across Genes

Why does the simple approach of gene-by-gene testing often fail? This problem arises because most microarray experiments assay only a small number of RNA samples. This translates to low power to detect differential expression using F- or Student's *t*-tests. One remedy for this problem is to combine information across genes to obtain improved estimates of the variance components before computing the gene-specific test statistics. One way to combine information is by pooling (averaging) the variance components estimates across all of the genes. This can improve power dramatically but at the risk of biasing results when some genes are truly more variable than others. A middle ground approach that achieves good power and yet allows for gene-to-gene variance heterogeneity is to "shrink" the variance component estimates by pulling each individual estimate closer, but not all the way, to the mean value across genes. F-tests based on shrunken estimates of variance components are powerful and robust (15,16).

Many gene expression changes are modest but highly coordinated across groups of genes. Thus, another potential way to take advantage of multiplicity would be to construct simultaneous tests for the differential expression of groups of genes. Information from all of the genes in a preselected group could be combined into a single test of significance by constructing an ANOVA F statistic. Alternatively, one could combine the data from multiple genes using a principle components summary and then test for the differential expression of this composite variable. Finding new ways to exploit multiplicity within the ANOVA framework is an aspect of microarray analysis that deserves further attention.

Software

Robust and reliable software is essential due to the numerical complexity of mixed model ANOVA computations. However, once the basic features of interfacing with a software package have been mastered, the data analyst can focus on the models and corresponding interpretations of the data with little concern for the details of the calculations. Among commercially available software packages for microarray analysis, the SAS microarray analysis system stands out because it is built upon the gold standard for mixed model ANOVA software, PROC MIXED (17). We have written a free, open source package for microarray data analysis in the R computing environment. R/maanova (18) is a powerful command-line analysis system that can interface with the many other statistical and microarray-specific packages written for the R environment. The newest release of R/maanova is available at http://www.jax.org/staff/churchill/labsite. A graphical user interface is currently under development.

Summary

ANOVA provides a general approach to the analysis of single and multiple factor experiments on both one- and two-color microarray platforms. Mixed model ANOVA is important because in many microarray experiments there are multiple sources of variation that must be taken into consideration when constructing tests for differential expression of a gene.

The genome is large, and the signals of expression change can be small, so we must rely on rigorous statistical methods to distinguish signal from noise. We apply statistical tests to ensure that we are not just making up stories based on seeing patterns where there may be none.

References

1. **Fisher, R.A.** 1925. Statistical Methods for Research Workers. Oliver and Boyd, Edinburg.
2. **Kerr, M.K., M. Martin, and G.A. Churchill.** 2000. Analysis of variance for gene expression microarray data. J. Comput. Biol. *7*:819-837.
3. **Yang, Y.H. and T.P. Speed.** 2002. Design issues for cDNA microarray experiments. Nat. Rev. Genet. *3*:579-583.
4. **Churchill, G.A.** 2002. Fundamentals of experimental design for cDNA microarrays. Nat. Genet. *32(Suppl)*:490-495.
5. **Rocke, D.M. and B. Durbin.** 2001. A model for measurement error for gene expression arrays. J. Comput. Biol. *8*:557-569.
6. **Yang, Y.H., S. Dudoit, P. Luu, D.M. Lin, V. Peng, J. Ngai, and T.P. Speed.** 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. *30*:e15.
7. **Cui, X.Q., M.K. Kerr, and G.A.Churchill.** 2003. Transformations for cDNA microarray data. Stat. Appl. Genet. Mol. Biol. *2*:article 4.
8. **Irizarry, R.A., B.M. Bolstad, F. Collin, L.M. Cope, B. Hobbs, and T.P. Speed.** 2003. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. *31*:e15.
9. **Wolfinger, R.D., G. Gibson, E.D. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari, and R.S. Paules.** 2001. Assessing gene significance from cDNA microarray expression data via mixed models. J. Comput. Biol. *8*:625-637.
10. **Jin, W., R.M. Riley, R.D. Wolfinger, K. White, G. Passador-Gurgel, and G. Gibson.** 2001. The contributions of sex, genotype and age to transcriptional variance in *Drosophila Melanogaster.* Nat. Genet. *29*:389-395.
11. **Cui, X.Q. and G.A. Churchill.** 2003. How many mice and how many arrays? Replication in mouse cDNA microarray experiments, p. 139-154. *In* K.F. Johnson and S.M. Lin (Eds.), Methods of Microarray Data Analysis III. Kluwer Academic, Norwell, MA.
12. **Chen, J.J., R.R. Delongchamp, C.-A. Tsai, H. Hsueh, F. Sistare, K.L. Thompson, V.G. Desai, and J.C. Fuscoe.** 2004. Analysis of variance components in gene expression data. Bioinformatics *20*:1436-1446.
13. **Dudoit, S., J.P. Shaffer, and J.C. Boldrick.** 2003. Multiple hypothesis testing in microarray experiments. Stat. Sci. *18*:71-103.
14. **Storey, J.D. and R. Tibshirani.** 2003. Statistical significance for genome-wide studies. Proc. Natl. Acad. Sci. USA *100*:9440-9445.
15. **Cui, X.Q. and G.A. Churchill.** 2003. Statistical tests for differential expression in cDNA microarray experiments. Genome Biol. *4*:210.
16. **Cui, X.Q., G. Hwang, J. Qiu, N. Blades, and G.A. Churchill.** Improved statistical tests for differential gene expression by shrinking variance components. Biostatistics (In press).
17. **Littell, R.C., G. Milliken, W.W. Stroup, and R.D. Wolfinger.** 1996. SAS System for Mixed Models. SAS Institute, Cary, NC.
18. **Wu, H., M.K. Kerr, and G.A. Churchill.** 2002. MAANOVA: a software package for the analysis of spotted cDNA microarray experiments. *In* The Analysis of Gene Expression Data: Methods and Software. Springer, New York.