STATISTICAL ANALYSIS OF 70-MER OLIGONUCLEOTIDE MICROARRAY DATA FROM POLYPLOID EXPERIMENTS USING REPEATED DYE-SWAPS

Hongmei Jiang

Jianlin Wang

Lu Tian

Z. Je rey Chen

R. W. Doerge

Follow this and additional works at: http://newprairiepress.org/agstatconference

Part of the Agriculture Commons, and the Applied Statistics Commons

This work is licensed under a Creative Commons Attribution-Noncommercial-No Derivative Works 4.0 License.

Recommended Citation


This is brought to you for free and open access by the Conferences at New Prairie Press. It has been accepted for inclusion in Conference on Applied Statistics in Agriculture by an authorized administrator of New Prairie Press. For more information, please contact cads@k-state.edu.
STaTISTICAL ANALYSIS OF 70-MER OLIGONUCLEOTIDE 
MICROARRAY DATA FROM POLYPLOID EXPERIMENTS 
USING REPEATED DYE-SWAPS

Hongmei Jiang\textsuperscript{1}, Jianlin Wang\textsuperscript{2}, Lu Tian\textsuperscript{2},
Z. Jefrey Chen\textsuperscript{2}, and R.W. Doerge\textsuperscript{1}

\textsuperscript{1}Department of Statistics, 1399 Math Building,
Purdue University, West Lafayette, IN 47907 USA
\textsuperscript{2}Genetics Program and Department of Soil and Crop Sciences,
Texas A&M University, College Station, TX 77843-2474 USA

Abstract

Polyploidy plays an important role in plant evolution. A series of \textit{Arabidopsis} autopolyploids and allopolyploids have been developed, and their transcript abundance compared using a 70-mer oligonucleotide microarray consisting of 26,090 annotated genes in \textit{Arabidopsis thaliana}. The experimental design included repeated dye-swaps, and analysis of variance (ANOVA) was employed to detect significant gene expression changes among and between the diploid, autopolyploid, and allopolyploid populations. Here, we discuss the statistical issues (replication, normalization, transformation, per-gene variance estimate, and the pooled estimate of variation) involved in analyzing these data, as well as the statistical findings of these analyses.

1 Introduction

A polyploid refers to cells or organisms that contain more than two complete sets of chromosomes. Polyploidy, the process of genome doubling that gives rise to organisms with multiple sets of chromosomes, has been very successful in nature and agriculture. In fact, more than 70\% of flowering plants are polyploids. The nature of polyploidy in \textit{Arabidopsis}, \textit{Brassica}, cotton, maize and wheat, has been studied intensely, however the cause of novel variation in polyploids is still not fully understood. For a recent review about polyploidy and research approaches, see Osborn et al. 2003 [1].

Microarray technology enables researchers to monitor tens of thousands of genes, or a whole genome, in a single experiment. This provides a powerful new approach to studying gene expression changes in the polyploids. For this study, a 70-mer oligonucleotide (oligos) microarray was employed [2]. The probes (features representing a gene on the array) are 70-mer oligos which are designed in the laboratory to ensure both high sensitivity and...
specificity. Like a cDNA spotted microarray, a 70-mer oligonucleotide microarray is also a two-channel or two-color array. The two total RNA samples of interest are labeled with different fluorescent dyes (Cy5 (red) and Cy3 (green)) during reverse transcription, then in equal amounts combined into one sample and applied to the array. After competitive hybridization, the array is scanned to produce images and separate quantifications of the two fluorescent dyes. For each spot (or gene feature) on the array, the red and green intensities represent the transcript abundance of the corresponding probe (or gene) in the two RNA samples. See [2] for more details about 70-mer oligonucleotide microarray technology.

A series of *Arabidopsis* polyploids (Figure 1, modified from Figure 1 of Comai et al. [9]) have been developed by doubling a single genome (autotetraploid), or combining two distinct but related genomes (allotetraploid) to study genetic and genomic consequences of genome duplication. The transcriptome variation among two autotetraploid parents and three independently derived allotetraploid offsprings were compared using a 70-mer oligonucleotide microarray consisting of 26,090 annotated genes in *Arabidopsis*. Here, we limit our discuss to the experimental design and data analysis details for comparing the gene expression changes between the two autotetraploids, *A. thaliana* 612 (At612) and *C. arenosa* (C.a.).

2 Experimental design

A simple and effective design for direct comparison of two treatment conditions or two types of samples (here we call them control and treatment samples) is a dye-swap, which is also known as a Latin-square design in classical statistics. This design uses two arrays but switches the color of the fluorescent dyes for the control and treatment samples when their mixed sample is hybridized to the array (for a review of microarray technology see [3]). On array 1, a control sample is labeled with red dye (Cy5) and treatment sample is labeled with green dye (Cy3); and on array 2, the treatment sample is labeled with red dye (Cy5) and control sample is labeled with green dye (Cy3). Because some genes incorporate each dye in different amounts, a dye-swap allows the assessment of this effect. In our experience, for genes with low intensities, the red dye (Cy5) typically yields a smaller (lower) intensity measurement than the green dye (Cy3).

There are approximately 26,000 genes in *Arabidopsis*, and all of these genes, except for some, were spotted only once on the 70-mer oligonucleotide array that we employed. For the genes that have replicated spots on the array, four of them were spotted 6 times and twelve of them were spotted 49 times. There were also some controls on the array that were spotted multiple times. Additionally, there were two biological replicates of each sample, and for each replicate, we used two repeated dye-swaps, which resulted in eight arrays (Table 1).

3 Methods

3.1 Background correction and log-transformation

Once each gene feature is assessed for both dyes (see [2] for technical details), we subtracted the background median intensity from the foreground median intensity for both red
and green intensities. If a background intensity is larger than the foreground intensity, the background-corrected gene intensity measurement is set to 1. Upon evaluation we found the variation of gene expression to increase as the mean intensity increased, so we employed the logarithm-transformation of the background-corrected intensities to stabilize the variance. Unfortunately, the log-transformation also inates the variation of genes with low intensity measurements, therefore we must be cautious when interpreting the results.

3.2 Normalization

As noted previously, there is a long history of one dye incorporating more or less than the other dye. To check for the dye effect, MA plots (M versus A plot as defined by Dudoit et al., 2002 [4]) were plotted, where M is the log-ratio of the background-corrected red and green intensities (i.e., $M = \log(\text{red/green})$); and A is the average log-intensity of the background-corrected red and green intensities (i.e., $A = (\log(\text{red}) + \log(\text{green}))/2$). If a gene is not differentially expressed, the red and green intensities are close to each other, and the log-ratio is close to 0. However, we noticed that when the average intensity is small, a lot of points fall below the horizontal (A) axis (i.e., there are a lot of gene features with the red intensity smaller than the green intensity) (Figure 2). Since this pattern was observed on both arrays in the dye-swap experiment where the dyes are exchanged between At612 and C.a. on array 1 and 2 (i.e., array 1: $M = \log(\text{red/green}) = \log(\text{At612/C.a.})$ and array 2: $M = \log(\text{red/green}) = \log(C.a./\text{At612})$), it is obviously not due to differences in the RNA samples, but instead the dye. That is, at the low gene expression levels, the dye effect is biased and the green dye gives higher measurements than the red dye. We also noticed (Figure 2) that the unbalanced dye effect for the controls do not have the same pattern as that for the non-control genes, and that the gene expression levels for the controls do not cover the whole range of the gene expression, which leads us to recommend not using the controls for normalization. Instead, we used a robust local regression [5] (loess function in the software package R) and the majority genes to remove the intensity-dependent dye effect, that is, the non-linear dependence of the log-ratio $M$ on the average log-intensity $A$ within each of the eight arrays. The distance between any one point and the loess smoothing line then becomes the new log-ratio. After normalization, the mean of log-ratio is roughly 0 and all data points scatter around the horizontal (A) axis (Figure 3). Linear transformations were then applied to the normalized log-ratio to get the normalized red and green intensities separately.

3.3 ANOVA models

Analysis of variance (ANOVA) models have been used to identify differentially expressed genes ([6]) or genes that change between treatments. Here, two ANOVA models were employed to detect differentially expressed genes, and the differences and similarities between them investigated for the purpose of understanding the result of the common variance assumption for each gene versus a per-gene variance assumption. For the first ANOVA model (1), we assume all 26,090 genes have the same variation; while the gene-based ANOVA model (3) is based on individual gene variation and requires thousands of ANOVA models.
3.3.1 ANOVA: common-variance approach

Different sources of variations in a microarray experiment include array, dye, treatment and gene [3]. The ANOVA model is:

\[
\log(y_{ijkgr}) = A_i + D_j + T_k + G_g + AG_{ig} + DG_{jg} + TG_{kg} + \epsilon_{ijkgr}.
\]  

(1)

where \(i = 1, \ldots, 8; j = 1, 2; k = 1, 2; g = 1, \ldots, 26090; \) and \(r = 1, \ldots, n_g (n_g \text{ is the number of replicated spots of gene } g)\); \(\bar{y}\) is the average gene intensity over all arrays, dyes, treatments and genes, and \(A, D, T \text{ and } G\) are the array, dye, treatment and gene main effects, while \(AG, DG \text{ and } TG\) are the interactions between array and gene, dye and gene, and treatment and gene, respectively. For this polyploid experiment the treatment effect is the parent (diploid, autopolyploid, or allopolyploid) effect. The error terms \(\epsilon_{ijkgr}\) are independent normal with mean 0 and variance \(\sigma^2\). When we test for differential expression of gene \(g\) between two treatments, we use \(T_k + TG_{kg}\) ([3]) and the test statistic is

\[
z = \frac{|(\hat{T}_1 + \hat{T}G_{1g}) - (\hat{T}_2 + \hat{T}G_{2g})|}{\sqrt{\frac{1}{4n_g} \sum_{ijkgr} (y_{ijkgr} - \bar{y}_{..g} - \bar{y}_{..j.g} - \bar{y}_{..kg} + 2\bar{y}_{..g})^2}}.
\]

(2)

where \(\sum_{ijkgr} (y_{ijkgr} - \bar{y}_{..g} - \bar{y}_{..j.g} - \bar{y}_{..kg} + 2\bar{y}_{..g})^2 = (16n_g - 10)\). For most genes spotted only once on the array \((n_g = 1)\), there are 16 observations, so the degrees of freedom for the error

3.3.2 ANOVA: per-gene-variance approach

In order to acknowledge each gene’s variation, we fit the ANOVA model on a gene-by-gene basis. That is, for each gene \(g\) (\(g\) is fixed in the following model), we have an ANOVA model as following,

\[
\log(y_{ijkgr}) = G_g + A_{ig} + D_{jg} + T_{kg} + \epsilon_{ijkgr}.
\]

(3)

Notice that here \(G_g\) is the average gene intensity for gene \(g\), while \(A, D \text{ and } T\) are the gene-specific array, dye and treatment effects, respectively. We assume the error terms \(\epsilon_{ijkgr}\) are independent normal with mean 0 and variance \(\sigma^2_g\). In fact when we assume the array is a fixed factor, \(G_g = +G_g\) in (1), \(A_{ig} = A_i + AG_{ig}\), and so on. We use \(T_{kg}\) to test the treatment effect for each gene and the test statistic is

\[
t = \frac{|T_{1g} - T_{2g}|}{\sqrt{\frac{1}{4n_g} \sum_{ijkgr} (y_{ijkgr} - \bar{y}_{..g} - \bar{y}_{..j.g} - \bar{y}_{..kg} + 2\bar{y}_{..g})^2}}.
\]

(4)
term is 6 (7 degrees freedom for the array effect, and 1 degree of freedom for average gene intensity \( g \), treatment effect, and dye effect, respectively). The test statistic in (4) has a t-distribution with 6 degrees of freedom.

### 3.3.3 Comparisons between two ANOVA models

Both the common-variance approach and per-gene-variance approach give the same estimates of the gene expression changes (i.e., the numerators in (2) and (4)) when array is a fixed factor. The differences in the two test statistics lies in how to estimate the variances of the gene expression changes, and the relative degrees of freedom of the test statistics. For the common-variance approach, we assume the residuals have a constant variance, and use all the observations to estimate the variation. For the per-gene-variance approach, only the observations related to a given gene are used. If in fact, all genes have the same variation, the common-variance approach is more powerful than the per-gene-variance approach, since the degrees of freedom is larger. However, it is well known that the residuals are non-normally distributed with non-constant variance (discussed later). When the common-variance approach is used, the genes with small fold-change AND small variation may not be detected, but the disadvantage in using the gene-basis ANOVA is that the degrees of freedom are small due to the limited number of replicate of spots and arrays.

### 4 Results and summary

For the polyploid experiment at hand, we applied ANOVA models (1) and (3) to detect statistically significant differentially expressed genes. To address the multiple comparison problem, we simply employed Benjamini-Hochberg’s FDR controlling procedure [7] at significance level 0.05. We identified 11,199 significantly differentially expressed genes between At612 and C.a. using the per-gene-variance approach, 4,363 genes using the common-variance approach, and 3,923 genes were identified by both approaches. With respect to detecting differentially expressed genes, the common-variance approach identified genes with large fold-changes, even some genes with large variations were found significant; the per-gene-variance approach detected both small and large fold-changes with small variations; and the significant genes identified by both approaches have large fold-changes and small variation. Interestingly, there are also several genes with small fold-changes and large variation that are statistically significant by either one or both approaches. In fact these same genes, as mentioned earlier, were spotted more than one time on the array. This illustrates the well known principle, “the bigger the sample size, the more powerful the test.” This being said it is not a fair assessment when there are more replications for one gene than another in the same microarray experiment, because genes with more replications at the spot level will have a bigger chance to be detected as differentially expressed.

One practical disadvantage of the common-variance approach is that one can not easily apply general statistical software to estimate the large number of parameters in model (1). However this can be easily addressed using a programming language, such as R and MATLAB. On the other hand, one practical advantage of the common-variance approach is model validation. Standard statistical QQ plots and residual plots can be used to check
the model assumptions, such as constant variance and normal distribution. Unfortunately, for the per-gene-variance approach (3), with tens of thousands of genes, it is not feasible to produce and assess this many visual plots.

In this paper, we treat all effects as fixed. The two approaches, the common-variance approach and the per-gene-variance approach, give the same estimates of gene expression changes between two treatment conditions, but different lists of statistically significant differentially expressed genes due to different ways of modeling the errors. When the array effect is treated as random (as well as its interaction with gene) in a mixed models approach, as proposed by Wolflnger et al., 2001 [8], the previous conclusions still hold when variance components are computed using traditional ANOVA estimates. However when the method of restricted maximum likelihood(REML) is used, the two ANOVA models produce substantially different results.

Finally, we investigated the effect of the log-transformation. The standard deviation which was computed using the per-gene-variance approach for each individual gene was plotted against its corresponding average gene intensity for the log-transformed data (Figure 4). It can be seen that the log-transformation works well for the genes with large intensities, however it isates the variation for genes with low intensities. The per-gene-variance approach can only detect genes with large intensity levels; while the common-variance approach is able to detect some differentially genes at the low intensity level, since it looks for large fold-changes. Therefore, when applying the per-gene-variance approach to log-transformed data, it will be di–cult to identify differentially expressed genes with low intensity levels, unless this is compensated by increasing the number of replicates.

To summarize, two ANOVA models, common-variance approach (1) and per-gene-variance approach (3), were compared in the context of identifying statistically significant differentially expressed genes between two autotetraploids using repeated dye-swaps experimental design. When the array effect is treated as fixed, these two approaches give the same estimates of gene expression changes, but yield different lists of differentially expressed genes. Being aware of the assumptions behind these two models, and the similarities and differences between them allows researchers to interpret their results appropriately.

Acknowledgments

We thank other members of the National Science Foundation Plant Genome funded Polyploid Project: Drs Tom Osborn (University of Wisconsin, Madison), Jim Birchler (University of Missouri), Luca Comai (University of Washington), Rob Martienssen (Cold Spring Harbor Laboratory).

References


Table 1: Experimental design for comparing two autotetraploids *A. thaliana* 612 (At612) and *C. arenosa* (C.a.). There are two biological replicates for each of the two autotetraploids, and for each replicate, two repeated dye-swaps are used.

<table>
<thead>
<tr>
<th></th>
<th>Biological replicate 1</th>
<th>Biological replicate 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dye-Swap 1</td>
<td>Dye-Swap 2</td>
</tr>
<tr>
<td><strong>Array</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Red dye (Cy5)</td>
<td>At612</td>
<td>C.a.</td>
</tr>
<tr>
<td>Green dye (Cy3)</td>
<td>C.a.</td>
<td>At612</td>
</tr>
</tbody>
</table>
Figure 1: *A. thaliana* (A.t) is a diploid with 5 chromosomes. *A. thaliana* 612 and *C. arenosa* (C.a.) are autotetraploids each having four complete sets of the same chromosomes. The cross of the two autotetraploids produces allotetraploid offspring.
Figure 2: MA plots for the two arrays of one dye-swap experiment. The left plot is array 1 with At612 labeled red and C.a. labeled green; the right plot is array 2 with At612 labeled green and C.a. labeled red. The loess smoothing lines for all genes (both black and purple points) and the controls (purple points only) are in blue and, red respectively.
Figure 3: MA plots for one array before and after loess normalization. The left plot represents the before normalization scenario, while the plot on the right represents the after normalization scenario.
Figure 4: The effect of log-transformation. The black points represent significant genes detected by common-variance approach; green points by per-gene-variance approach; and red points by both approaches.