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

See next page for additional authors

Recommended Citation

Jiang, Hongmei; Wang, Jianlin; Tian, Lu; Chen, Z. Je rey; and Doerge, R. W. (2004). "STATISTICAL ANALYSIS OF 70-MER OLIGONUCLEOTIDE MICROARRAY DATA FROM POLYPLOID EXPERIMENTS USING REPEATED DYE-SWAPS," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1159

Author Information

Hongmei Jiang, Jianlin Wang, Lu Tian, Z. Je rey Chen, and R. W. Doerge

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

Hongmei Jiang[1], Jianlin Wang[2], Lu Tian[2],
Z. Je rey Chen[2], and R.W. Doerge[1]

[1]Department of Statistics, 1399 Math Building,
Purdue University, West Lafayette, IN 47907 USA
[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 *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 *Arabidopsis thaliana*. The experimental design included repeated dye-swaps, and analysis of variance (ANOVA) was employed to detect signi cant 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 ndings 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 owering plants are polyploids. The nature of polyploidy in *Arabidopsis, 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

speci city. 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 di erent uorescent 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 quanti cations of the two uorescent 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, modi ed 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 o springs 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 e ective 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 uorescent 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 di erent amounts, a dye-swap allows the assessment of this e ect. 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 in ates 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 e ect, MA plots (M versus A plot as de ned by Dudoit et al., 2002 [4]) were plotted, where M is the log-ratio of the background-corrected red and green intensities (i.e., $M = \log(\text{red}/\text{green})$); 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 di erentially 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: $M = \log(\text{red}/\text{green}) = \log(\text{At612}/\text{C.a.})$ and array 2: $M = \log(\text{red}/\text{green}) = \log(\text{C.a.}/\text{At612})$), it is obviously not due to di erences in the RNA samples, but instead the dye. That is, at the low gene expression levels, the dye e ect is biased and the green dye gives higher measurements than the red dye. We also noticed (Figure 2) that the unbalanced dye e ect 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 e ect, 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 di erentially expressed genes ([6]) or genes that change between treatments. Here, two ANOVA models were employed to detect di erentially expressed genes, and the di erences 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 rst 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

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

$$\log(y_{ijkgr}) = \quad + A_i + D_j + T_k + G_g + AG_{ig} + DG_{jg} + TG_{kg} + \quad _{ijkgr}. \tag{1}$$

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

$$z = \frac{|(\widehat{T_1} + \widehat{TG}_{1g}) - (\widehat{T_2} + \widehat{TG}_{2g})|}{\sqrt{\frac{1}{4n_g}\widehat{\ }^2}}, \tag{2}$$

where $\widehat{\ }^2 = \sum_{ijkgr} (y_{ijkgr} - \bar{y}_{i..g.} - \bar{y}_{.j.g.} - \bar{y}_{..kg.} + 2\bar{y}_{...g.})^2/(n - 10m)$, n is the total number of data points, m is the total number of genes, and $\bar{y}_{.j.g.}$ is the average intensity over the omitted indices i, k and r. Because there are 26,090 genes, the degrees of freedom for the error term is very large (approximately 180,000). The test statistic in (2) is a z-test statistic (i.e., normally distributed with mean 0 and variance 1 under the null hypothesis, $T_1 + TG_{1g} = T_2 + TG_{2g}$).

3.3.2 ANOVA: per-gene-variance approach

In order to acknowledge each gene's variation, we t the ANOVA model on a gene-by-gene basis. That is, for each gene g (g is xed in the following model), we have an ANOVA model as following,

$$\log(y_{ijkgr}) = \quad _g + A_{ig} + D_{jg} + T_{kg} + \quad _{ijkgr}. \tag{3}$$

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

$$t = \frac{|\widehat{T}_{1g} - \widehat{T}_{2g}|}{\sqrt{\frac{1}{4n_g}\widehat{\ }^2_g}}, \tag{4}$$

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

term is 6 (7 degrees freedom for the array e ect, and 1 degree of freedom for average gene intensity $_g$, treatment e ect, and dye e ect, 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 xed factor. The di erences 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 signi cant di erentially expressed genes. To address the multiple comparison problem, we simply employed Benjamini-Hochberg's FDR controlling procedure [7] at signi cance level 0.05. We identi ed 11,199 signi cantly di erentially 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 identi ed by both approaches. With respect to detecting di erentially expressed genes, the common-variance approach identi ed genes with large fold-changes, even some genes with large variations were found signi cant; the per-gene-variance approach detected both small and large fold-changes with small variations; and the signi cant genes identi ed 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 signi cant 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 di erentially 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 e ects as xed. 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 di erent lists of statistically signi cant differentially expressed genes due to di erent ways of modeling the errors. When the array e ect is treated as random (as well as its interaction with gene) in a mixed models approach, as proposed by Wol nger 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 di erent results.

Finally, we investigated the e ect 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 in ates 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 di erentially 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 di erentially 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 signi cant di erentially expressed genes between two autotetraploids using repeated dye-swaps experimental design. When the array e ect is treated as xed, these two approaches give the same estimates of gene expression changes, but yield di erent lists of di erentially expressed genes. Being aware of the assumptions behind these two models, and the similarities and di erences 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

[1] Osborn, T. C., Pires, J. C., Birchler, J. A., Auger, D. L., Chen, Z. J., Lee, H.-S., Comai, L., Madlung, A., Doerge, R., Colot, V., and Martienssen, R. A. Understanding mechanisms of novel gene expression in polyploids. *Trends in Genetics* **19**, 141–147 (2003).

[2] Lee, H.-S., Wang, J., Tian, L., Jiang, H., Black, M., Madlung, A., Watson, B., Lukens, L., Pires, J., Wang, J. J., Comai, L., Osborn, T., Doerge, R. W., and Chen, Z. J. Sensitivity of 70-mer oligonucleotides and cDNAs for microarray analysis of gene expression in Arabidopsis and its related species. *Plant Biotechnology Journal* **2**, 45–57 (2004).

[3] Craig, B. A., Black, M. A., and Doerge, R. W. Gene expression data: The technology and statistical analysis. *Journal of Agricultural, Biological, and Environmental Statistics (JABES)* **8**(1), 1–28 (2003).

[4] Dudoit, S., Yang, Y. H., Speed, T. P., and Callow, M. J. Statistical methods for identifying di erentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* **12**(1), 111–139 (2002).

[5] Cleveland, W. S., Grosse, E., and Shyu, W. M. Local regression models. In Statistical Models in S, Chambers, J. M. and Hastie, T. J., editors, chapter 8, 309:376. CRC Press, Inc. (1991).

[6] Kerr, M. K., Martin, M., and Churchill, G. A. Analysis of variance for gene expression microarray data. *Journal of Computational Biology* **7**, 819–837 (2000).

[7] Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**, 289:300 (1995).

[8] Wol nger, R. D., Gibson, G., Wol nger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., and Paules, R. S. Assessing gene signi cance from cDNA microarray expression data via mixed models. *Journal of Computational Biology* **8**(6), 625–637 (2001).

[9] Comai, L., Tyagi, A. P., Winter, K., Holmes-Davis, R., Reynolds, S. H., Stevens, Y., and Byersb, B. Phenotypic instability and rapid gene silencing in newly formed Arabidopsis allotetraploids. *The Plant Cell* **12**(9), 1551–1568 (2000).

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.

	Biological replicate 1				Biological replicate 2			
	Dye-Swap 1		Dye-Swap 2		Dye-Swap 3		Dye-Swap 4	
Array	1	2	3	4	5	6	7	8
Red dye (Cy5)	At612	C.a.	At612	C.a.	At612	C.a.	At612	C.a.
Green dye (Cy3)	C.a.	At612	C.a.	At612	C.a.	At612	C.a.	At612

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 o spring.

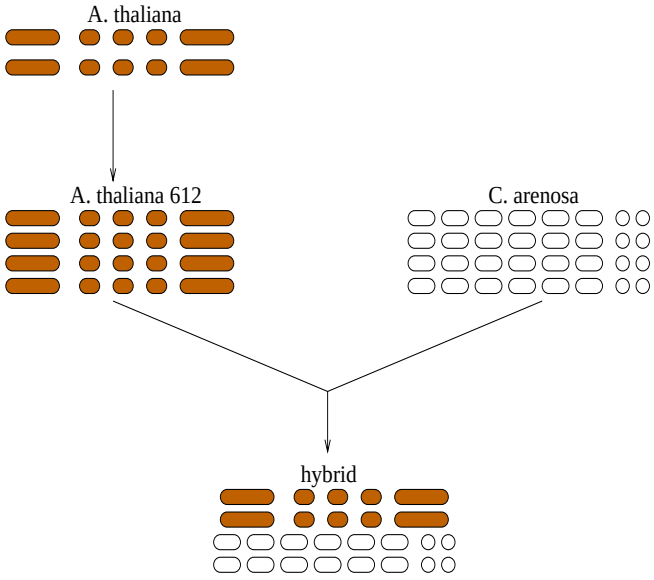

Figure 2: MA plots for the two arrays of one dye-swap experiment. The left plot is array 1 with At612 labeled red and C.a. labeled green; the right plot is array 2 with At612 labeled green and C.a. labeled red. The loess smoothing lines for all genes (both black and purple points) and the controls (purple points only) are in blue and, red respectively

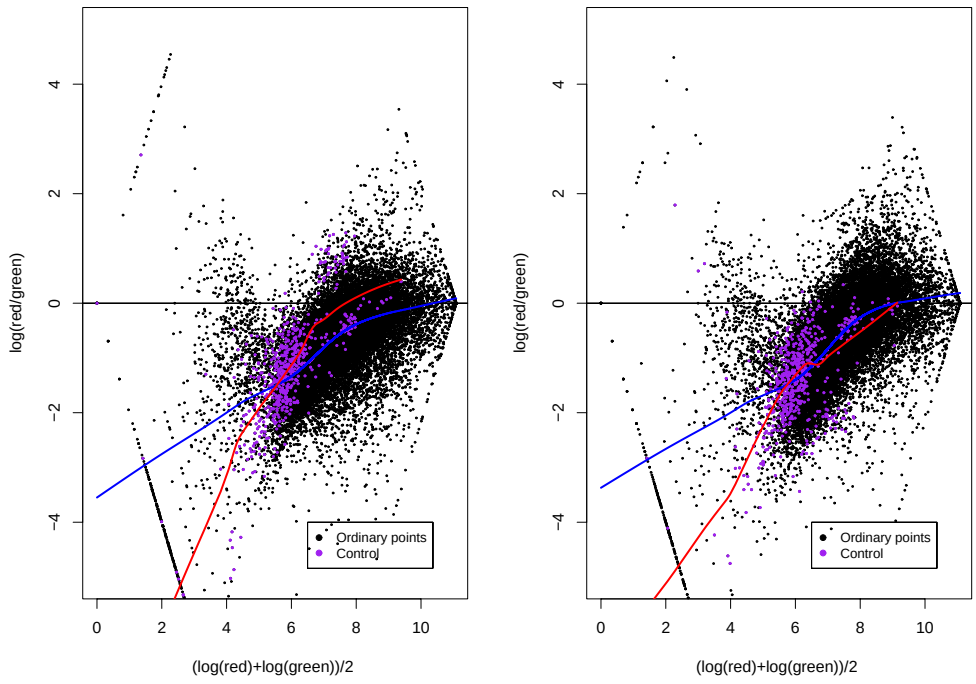

Figure 3: MA plots for one array before and after loess normalization. The left plot represents the before normalization scenario, while the plot on the right represents the after normalization scenario.

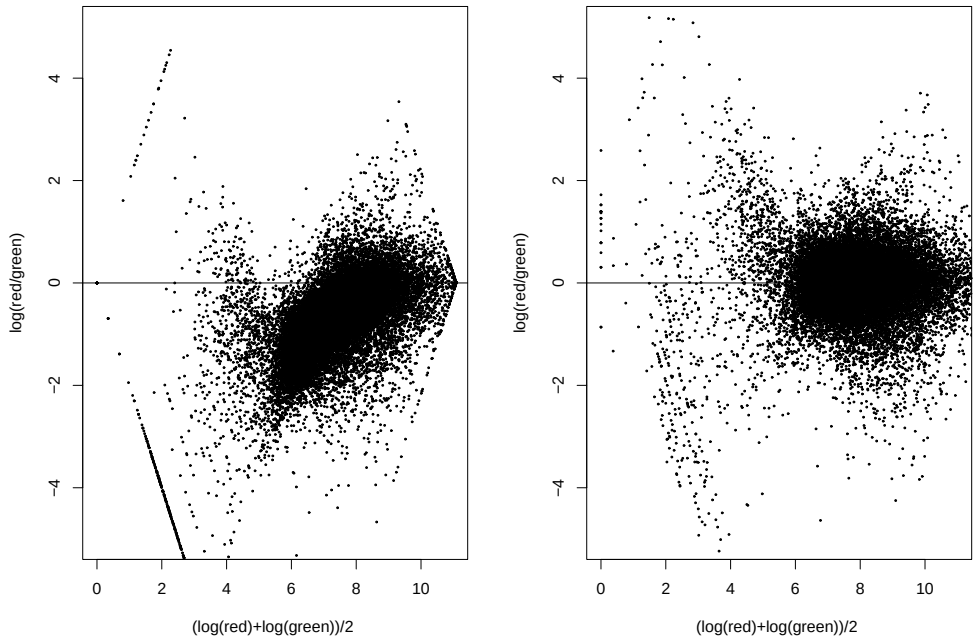

Figure 4: The e ect of log-transformation. The black points represent signi cant genes detected by common-variance approach; green points by per-gene-variance approach; and red points by both approaches.