Research Report

Use of RNA and Genomic DNA References for Inferred Comparisons in DNA Microarray Analyses

BioTechniques 33:924-930 (October 2002)

H. Kim, B. Zhao, E.C. Snesrud, B.J. Haas, C.D. Town, and J. Quackenbush
The Institute for Genomic Research, Rockville, MD, USA

ABSTRACT

In most microarray assays, labeled cDNA molecules derived from reference and query RNA samples are co-hybridized to probes arrayed on a glass surface. Gene expression profiles are then calculated for each gene based on the relative hybridization intensities measured between the two samples. The most commonly used reference samples are typically isolates from a single representative RNA source (RNA-0) or pooled mixtures of RNA derived from a plurality of sources (RNA-p). Genomic DNA offers an alternative reference nucleic acid with a number of potential advantages, including stability, reproducibility, and a potentially uniform representation of all genes, as each unique gene should have equal representation in a haploid genome. Using hydrogen peroxide-treated Arabidopsis thaliana *plants as a model, we evaluated genomic DNA and RNA-p as reference samples and compared expression levels inferred through the reference relative to unexposed plants with expression levels measured directly using an RNA-0 reference. Our analysis demonstrates that while genomic DNA can serve as a reasonable reference source for microarray assays, a much greater correlation with direct measurements can be achieved using an RNA-based reference sample.*

INTRODUCTION

DNA microarray analysis has become the most widely used technique for global expression studies (7,9,10). Most approaches to microarray analysis use an experimental design in which cDNA molecules derived from query and reference RNA samples are labeled with distinguishable fluorescent dyes and co-hybridized to an array containing thousands of genes (3); both the query and the reference samples used in hybridization reactions are generally derived from RNA isolated from experimental tissue samples. One difficulty with microarray assays is the identification of an appropriate reference sample for comparison.

In studies evaluating expression differences in perturbed and non-perturbed systems that are otherwise identical, RNA isolated from the non-perturbed state generally serves as an excellent reference, as one can easily see changes in the perturbed state in pairwise comparisons. However, many microarray studies involve complex sets of comparisons between large numbers of samples. In these instances, identification of a single reference RNA source can be difficult. One solution that has been widely used is to create an RNA pool (RNA-p) derived from either all the samples under study or from independent collections of samples selected to achieve the widest possible representation of gene expression.

The construction of an appropriate reference sample in these cases presents a number of challenges. First, it is often difficult to identify samples that provide complete coverage for all of the genes represented on the array. This can be problematic if the query samples exhibit measurable expression for a significant number of genes that are not seen in the reference, as the hybridization intensities for these genes cannot be appropriately normalized or compared between arrays. Second, should the initial RNA reference sample be exhausted, it is often difficult to precisely reconstruct the original; any changes in the reference can potentially skew the resulting data sets, leading to problems during data analysis and interpretation.

Unlike RNA samples, which depend on tissue, state, and developmental stage, genomic DNA isolated from healthy organisms provides a stable reference nucleic acid, independent of variations in the growth, culture conditions, and subsequent purification processes. Further, genomic DNA should, in principal, provide nearly uniform representation of all of the genes on an array, as all genes should be equally represented in the genome. Finally, genomic DNA has the potential to be a truly "universal reference" that could be used in any experiment in a particular organism.

To test this hypothesis, to evaluate the relative performance of genomic DNA and RNA-p as a common reference for microarray gene expression profiling, and to demonstrate the feasibility of using a common reference to infer expression levels, we evaluated transcriptional response to oxidative stress in the model plant *Arabidopsis thaliana*. Specifically, we exposed plants to hydrogen peroxide and measured the expression at 3 and 6 h following exposure, as well as in unexposed plants. Using both genomic DNA and RNA pooled from all three time points (0, 3, and 6 h) as common reference

samples, we computed the expression relative to unexposed plants using a straightforward spot-by-spot normalization procedure. To evaluate the relative performance of these reference sources, we also made direct comparisons of expression using RNA from plants at the 3- and 6-h time points to RNA derived from unexposed plants (RNA-0).

MATERIALS AND METHODS

Microarray Preparation

DNA microarrays of chromosome 2 of *A. thaliana* were fabricated as described by Hegde et al. (3). Briefly, genomic fragments representing 4180 gene models on chromosome 2 were amplified by PCR, purified, and printed on SuperAmine™ aminosilane-coated microscope slides (Telechem, Sunnyvale, CA, USA) using an arraying robot built by Intelligent Automatic Systems (Cambridge, MA, USA). Printed slides were UV-irradiated at an integrated intensity of 350 mJ using a Stratalinker™ (Stratagene, La Jolla, CA, USA) to cross-link spotted DNA to the slides and were stored in a desiccated chamber until use.

Plant Materials and Isolation of RNA and Genomic DNA

Wild-type *A. thaliana* Columbia plants were germinated and grown in a 500-mL flask with 100 mL half-strength Murashige and Skoog medium (pH 5.7), vitamins, and sucrose without hormone (6) for 12 days with shaking at 100 rpm under constant light. Hydrogen peroxide (Sigma, St. Louis, MO, USA) was applied to the plants at a final concentration of 5 mM. Treated plants were harvested at 0, 3, and 6 h following exposure and were immediately frozen in liquid nitrogen for RNA extraction.

Total RNA was isolated using TRIZOL® (Invitrogen, Carlsbad, CA, USA) following the manufacturer's protocol. Pools of mRNA were enriched from total RNA preparations using the Oligotex™ mRNA purification kit (Qiagen, Valencia, CA, USA). Genomic DNA was prepared from untreated 12 day-old *Arabidopsis* plants using DNeasy® Plant Mini kit (Qiagen). Purified genomic DNA was digested with *Sau*3AI (New England Biolabs, Beverly, MA, USA) and purified with a QIAquick™ PCR purification kit (Qiagen) before labeling and hybridization.

Fluorescent Labeling and Hybridization

Labeling reactions with enriched mRNA were prepared as described in the TIGR standard operating procedures (http://atarrays.tigr.org). A 1-µg quantity of poly(A)-enriched mRNA was used as a template for random-primed first-strand cDNA synthesis in the presence of amino-allyl-dUTP (Sigma) during reverse transcription, followed by the conjugation of purified reaction products to the esters of Cy3™ and Cy5™ fluorescent dyes (Amersham Biosciences, Piscataway, NJ, USA). Digested genomic DNA samples (3 µg each) were labeled with amino-allyl-dUTP using random primers in the presence of Klenow enzyme (Invitrogen), followed by coupling to the Cy3 or Cy5 esters.

Hybridization reactions and microar-

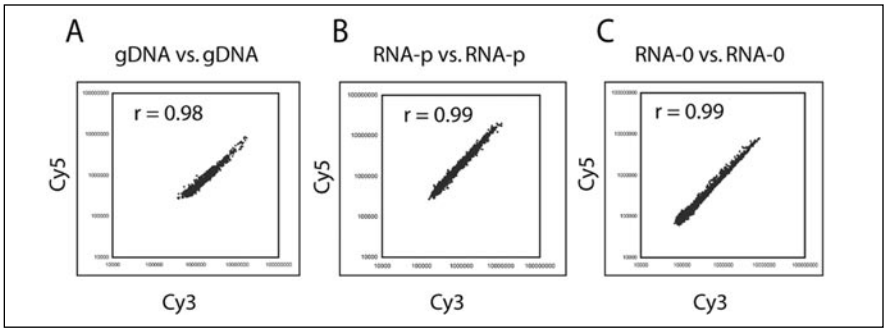


Figure 1. Reproducibility of labeling and hybridization for different nucleic acid samples. Scatter plots of self versus self hybridization intensities for (A) genomic DNA (gDNA), (B) RNA-p, and (C) RNA-0. Pearson correlation coefficients (r) for the two channels are shown in each plot.

ray image analyses were conducted as described in the TIGR microarray standard operating procedures (http://atarrays.tigr.org). Following hybridization, microarray slides were scanned using a laser scanner (GenePix® 4000, Axon Instrument, Foster City, CA, USA). Background-subtracted integrated intensities as well as background levels were measured for each spot using TIGR Spotfinder 1.0 (http://www.tigr.org/software; TIGR, Rockville, MD, USA). Spots that had lower intensities than local background were assigned 0 values. To remove unreliable low-intensity spots, those with intensities less than 1.5 times the corresponding local background level in either of the two channels were eliminated before data normalization.

RESULTS

Self versus Self Hybridizations of Genomic DNA and cDNA Samples

As there have not been any published studies using genomic DNA as a reference for microarray expression analysis, we first investigated the reproducibility of genomic DNA labeling and hybridization to validate its suitability for this purpose. Equal quantities of the same isolate of genomic DNA were independently labeled with Cy3 and Cy5 and co-hybridized to a single *Arabidopsis* chromosome 2 microarray containing PCR products representing 4180 genes and pseudogenes present on the chromosome (5) (Figure 1). Fluorescence intensities were measured in both channels, and signals were normalized within each array using a total intensity approach (8). Normalized signal intensities correlated extremely between channels, with a Pearson correlation coefficient (r) of 0.98 (Figure 1A), and compared favorably to results obtained for pooled RNA and RNA derived from a single source (Figure 1, B and C). This suggests that genomic DNA can be reproducibly labeled and that there are few, if any, dye-specific labeling and hybridization artifacts.

Table 1. Hybridizations Performed in This Study

	Cy5-Labeled Sample				
Cy3-Labeled Sample	**RNA-p**	**Genomic DNA**	**RNA-0**	**RNA-3**	**RNA-6**
RNA-0			×	×	×
RNA-p	×		×	×	×
Genomic DNA		×	×	×	×

Use of Hydrogen Peroxide-Treated *Arabidopsis* Plants to Test the Reliability of References in Gene Expression Profiling

To evaluate the relative utility of RNA-p and genomic DNA reference sources, we examined their ability to assess temporal patterns of expression in *Arabidopsis* plants treated with hydrogen peroxide. We chose this system because RNA from unexposed plants (RNA-0) provides a natural basis for comparison with later time points. In addition, there are a number of well-known hydrogen peroxide-responsive genes located on chromosome 2 that could be used as markers to evaluate the results. These include genes coding for glutathione S-transferase 6 (GST6), phenylalanine ammonia lyase (PAL1), and the 70-kDa heat shock protein, which are involved in detoxification or defense processes (1,2,4). For both the RNA-p and genomic DNA references, we used the measured hybridization intensities to calculate gene expression 3 and 6 h after exposure relative to the zero time point; these results were evaluated using direct comparison of these time points to RNA from unexposed plants (RNA-0). In total, 11 hybridizations were performed, as shown in Table 1, including nine that were used to evaluate hydrogen peroxide response.

Data Normalization and Comparison

The first step in our analysis was to calculate inferred levels at the 3- and 6-h time points relative to unexposed plants using the comparisons between these and the reference nucleic acids (RNA-p or genomic DNA). These inferred measurements would then be

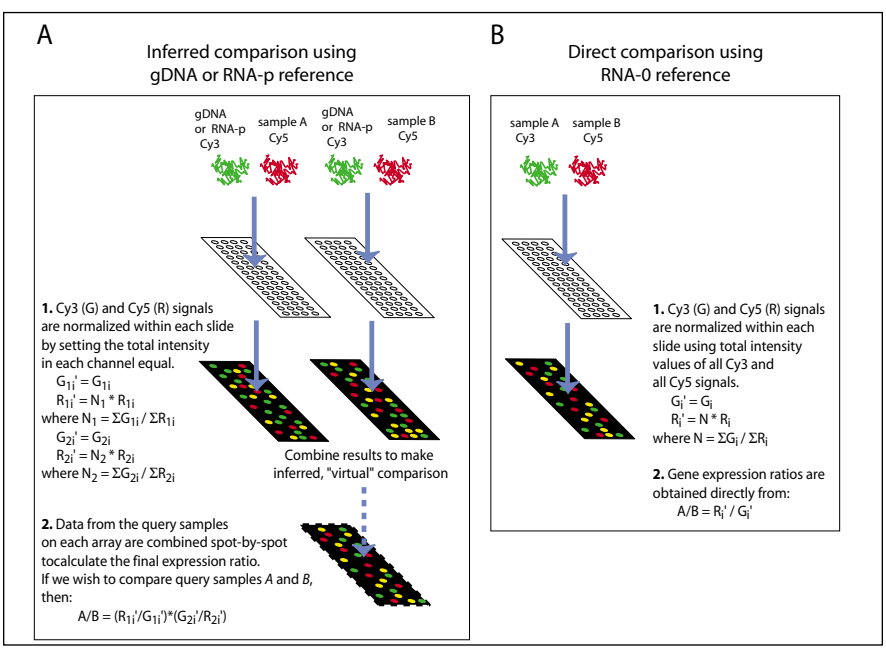
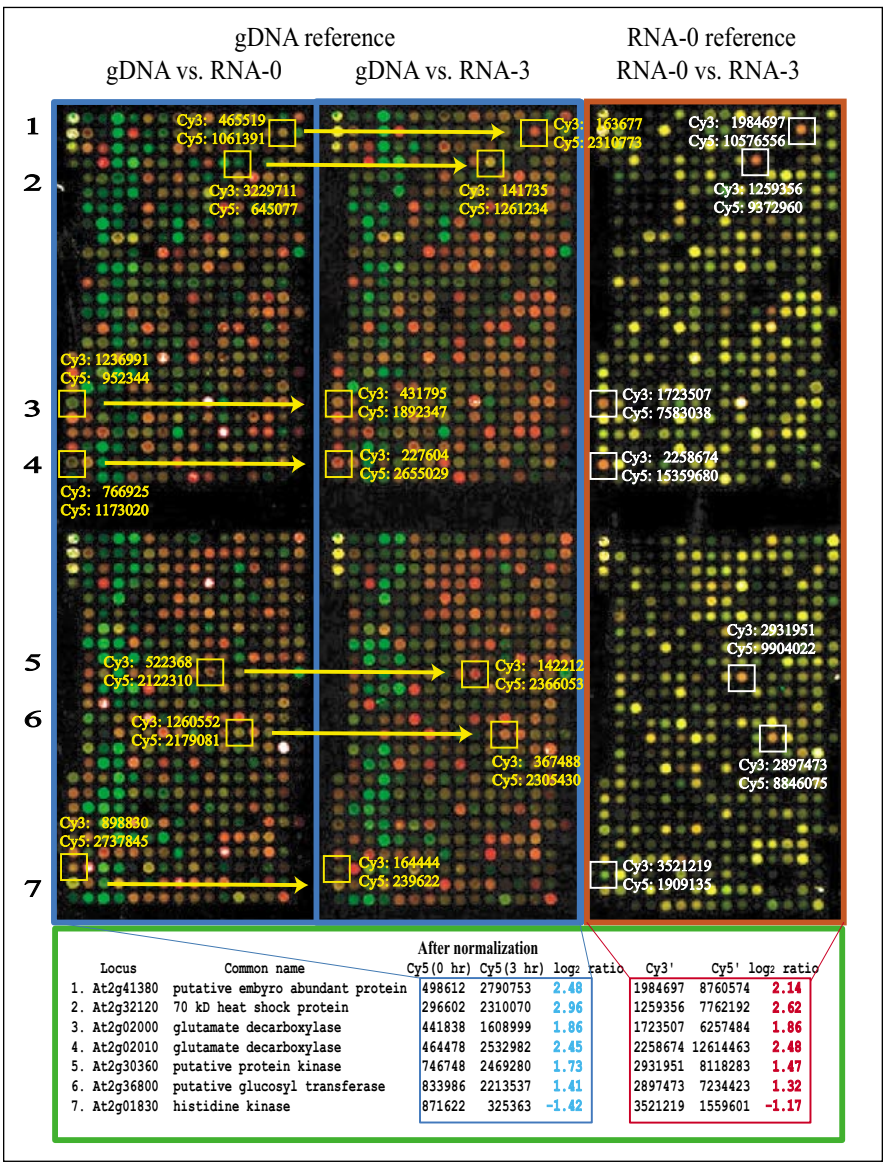


Figure 2. Overview of data collection and analysis. This figure illustrates how one can compare two experimental states (A to B) using either common reference samples or direct comparisons. (A) For common references, measured hybridization intensities are normalized within the slide by scaling the intensities for each spot such that the total hybridization intensity in each channel is equal. Ratios between states are then computed spot-by-spot using the common reference to balance the signals. (B) For direct comparisons, total intensity normalization is used. Gene expression ratios between conditions are then computed directly for each gene.

evaluated relative to direct comparisons between the 3- and 6-h time points with the unexposed plants. Hybridization intensities for the elements on each slide were first adjusted using total intensity normalization.

For direct comparison measurements of later time points with the untreated RNA-0 reference, these normalized values were then used to compute an inferred expression ratio for each gene, comparing later times to the unexposed reference (time 0): T/T_0, where T represents the expression level at either the 3- or 6-h time point. For both the genomic DNA and RNA-p reference samples, the total intensity was also adjusted to be equal between slides, and expression ratios measured for each gene at each time point relative to the reference were used to infer the T/T_0 ratio:

$$\frac{T}{T_0} = \left[\frac{T}{R_1}\right] * \left[\frac{R_0}{T_0}\right]$$



After normalization

	Locus	Common name	Cy5(0 hr)	Cy5(3 hr)	log2 ratio	Cy3'	Cy5'	log2 ratio
1.	At2g41380	putative embyro abundant protein	498612	2790753	2.48	1984697	8760574	2.14
2.	At2g32120	70 kD heat shock protein	296602	2310070	2.96	1259356	7762192	2.62
3.	At2g02000	glutamate decarboxylase	441838	1608999	1.86	1723507	6257484	1.86
4.	At2g02010	glutamate decarboxylase	464478	2532982	2.45	2258674	12614463	2.48
5.	At2g30360	putative protein kinase	746748	2469280	1.73	2931951	8118283	1.47
6.	At2g36800	putative glucosyl transferase	833986	2213537	1.41	2897473	7234423	1.32
7.	At2g01830	histidine kinase	871622	325363	-1.42	3521219	1559601	-1.17

Figure 3. Scanned images of genomic DNA (gDNA) reference and RNA-0 sets and comparisons of the normalized ratios for sample genes. Seven spots that showed significant up- (red) or down-regulation (green) in the RNA-0 dataset were chosen for comparisons and are indicated by white open squares in both reference set images. Cy3 and Cy5 raw intensity values generated with TIGR Spotfinder 1.0 for the given spots are shown as yellow (genomic DNA set) or white text (RNA-0 set) near the associated spots. For the genomic DNA sample, corresponding spots in images from both channels are shown. Final normalized values and calculated log2 ratios of the corresponding seven genes are shown in the box below the images.

where R_1 and R_0 are the measured hybridization intensities for the reference sample (either genomic DNA or RNA-p) for each gene on the time T (3 or 6 h after treatment) or time T_0 slide, respectively. The underlying hypothesis here is that the representation of the reference should be equivalent between hybridizations and, consequently, that the measured fluorescence intensity for each gene measured in the reference channel should be equal across all slides. This scaling preserves the ratio between the query and reference sample for each gene on each slide and allows meaningful comparisons between genes across slides, as it provides a normalization of the data to account for spot variability. This procedure is shown schematically in Figure 2, where samples A and B represent the T and T_0 samples, respectively.

All calculated gene expression ratios were $\log_2$-transformed before comparing measurements. Differentially expressed genes at the 95% confidence level for each reference set were determined by assuming the $\log_2$ ratio values for each data set are normally distributed and selecting genes with values more than 1.96 standard deviations from the mean of all $\log_2$ ratio values. Figure 3 shows hybridization images using both RNA-0 and genomic DNA reference sources, along with ratios calculated for seven significantly regulated genes, comparing expression at the 3-h time point relative to time zero.

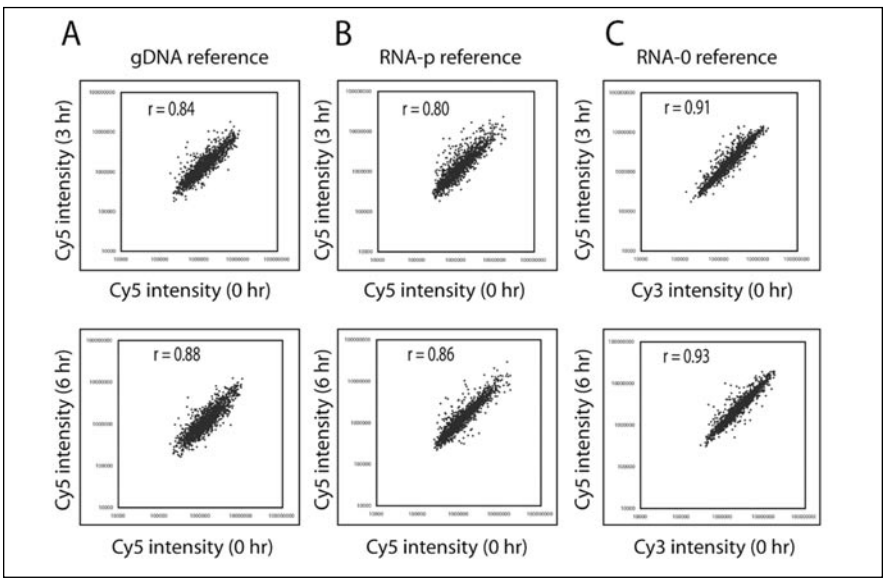


Figure 4. Scatter plots of normalized gene expression levels at 3 and 6 h following hydrogen peroxide exposure relative to unexposed plants (0 h). The data here are shown for each of the three reference samples used (A) genomic DNA (gDNA), (B) RNA-p, and (C) RNA-0. Only a minority of genes assayed shows noticeable changes in expression, and, consequently, measurements are highly correlated. As expected, inferred measurements using common reference nucleic acids show more variability than direct measurement between time points.

Comparison of Reference Samples

To facilitate our comparison, we limited analysis to those genes with hybridization intensities detectable above background on all arrays at a particular time point following exposure (1608 genes for 0 to 3 h comparisons, and 1669 genes for 0 to 6 h); scatter plots of the

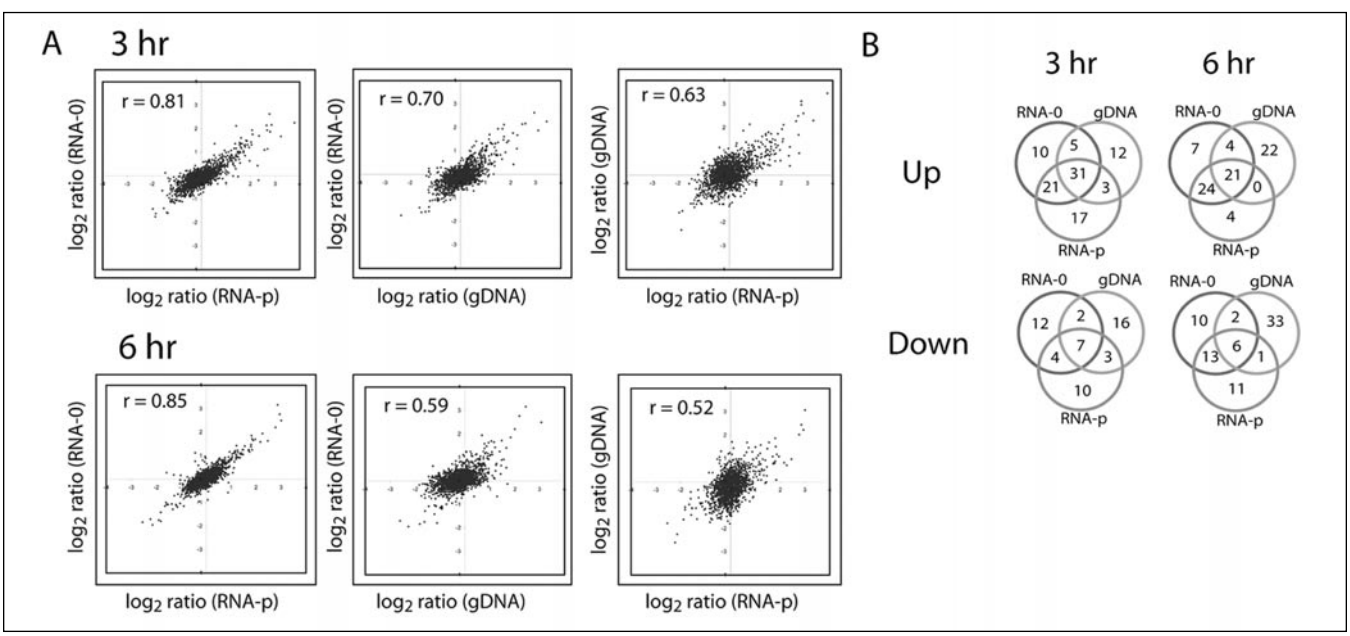


Figure 5. Comparison of differential expression in hydrogen peroxide-exposed plants calculated using each of three reference sources. (A) Direct comparison of log2 ratios for each pair of reference samples at both the 3- and 6-h time points. (B) Comparison of the genes identified as being differentially up- or down-regulated at the 95% confidence level (see Materials and Methods) using each of the three reference samples. A complete list of genes is available at http://atarrays.tigr.org.

Table 2. Expected Hydrogen Peroxide-Responsive Genes and the $\log_2$ Ratio of Gene Expression in Exposed Plants[a]

		Genomic DNA		RNA-p		RNA-0	
Locus	**Common Name**	**3 h**	**6 h**	**3 h**	**6 h**	**3 h**	**6 h**
At2g32120	70-kDa heat shock protein	**2.97**	**2.21**	**2.79**	**2.96**	**2.62**	**2.75**
At2g19310	Putative small heat shock protein	**1.38**	**1.48**	**1.23**	**1.02**	**0.87**	**0.89**
At2g37040	*PAL*1	**1.21**	-0.48	**1.62**	0.42	**1.36**	0.28
At2g47730	*GST*6	**1.50**	**0.96**	**2.08**	**2.48**	**1.93**	**1.61**
At2g29450	*GST*	**1.86**	0.78	0.55	-0.44	**1.36**	0.76

[a]These are measured relative to unexposed plants, for each of the three reference sources. Genes showing significant differential expression (>95% confidence) are bolded. Note that values greater than 0 imply induction, while those less than 0 imply repression.

normalized $\log_2$ ratios obtained using our three reference sources are shown in Figure 4. As can be most clearly seen in the direct RNA-0 comparisons, the majority of genes assayed did not exhibit a significant change in expression following exposure to hydrogen peroxide. Further, as might be expected, inferred measurements based on independent comparison to a reference sample are more variable than direct measurements. However, it is noteworthy that both genomic DNA- and RNA-p-based comparisons exhibit similar levels of correlation, suggesting that the accuracy of genomic DNA-based measurement are comparable to those obtained using other reference nucleic acids.

To test this observation, we examined correlations between the $\log_2$ ratio expression measurements with each of the three reference sources, calculated relative to unexposed plants, at the 3- and 6-h time points. As can be seen in Figure 5A, at both 3 and 6 h, the calculated expression ratios correlate much better between the RNA-based measurements than between genomic DNA and either RNA-based measure. We then compared the genes identified as significantly up- or down-regulated at the 95% confidence level in each direct or inferred comparison to unexposed plants (Figure 5B). Again, the genomic DNA reference measurements were found to be the most divergent, sharing

only 46% with the direct RNA-0 measurements. In contrast, the RNA-p measurements had 72% of its differentially expressed genes validated by the RNA-0 measurements. This clearly suggests that pooled RNA is superior to genomic DNA as a reference sample for array hybridizations. This may be because hybridization kinetics in genomic DNA-based samples is more complex than that involving RNA-based reference samples. When only RNA is used as the source nucleic acid, labeling involves creating first-strand cDNA products for both the query and the reference sample. However, genomic DNA references contain representatives of both the coding and noncoding strands. As such, solution hybridization interactions may interfere with hybridization to the arrayed probes, particularly for genes expressed at low to moderate levels.

However, all three sets of experiments, including those obtained using genomic DNA, successfully reported significant induction of the three marker genes, 70-kDa heat shock protein (At2g32120), *PAL*1 (At2g37040), and *GST*6 (At2g47730) (Table 2). We also examined the expression of two additional genes that code for a putative small heat shock protein (At2g19310) and another *GST* (At2g29450) that are closely related to the genes previously identified as being regulated by hydrogen peroxide (Table 2). The RNA-0 and genomic DNA reference samples identified these genes as significantly differentially expressed, while the RNA-p set failed to report differential expression for *GST*. In more extensive time course experiments examining expression at 0, 1, 3, 6, and 12 h after exposure and using an RNA-0 reference, *GST* clearly showed induction in response to hydrogen peroxide, peaking 1 h after exposure (unpublished data). The consistency of these with published results suggests that both direct and inferred comparisons can yield results that are biologically meaningful.

DISCUSSION

There is an increasing number of published studies showing that differentially expressed genes identified by direct comparison on DNA microarrays can be validated using other techniques, such as RT-PCR or Northern analysis. However, there is growing interest in using common reference samples for DNA microarray analyses, with the underlying assumption being that comparisons to the common reference can be used to infer any direct comparisons between samples. While most studies use pooled RNA as the common reference, RNA sources are inherently unstable, as each time the RNA is collected, differences in growth or collection conditions can change the composition of the final RNA sample. For this reason, there has been a growing interest in a "universal" reference that would not depend on the age, treatment, or tissue of the source. Genomic DNA is one potential universal reference.

Our results indicate that inferred comparisons using pooled RNA as a common reference can indeed compare favorably to direct comparisons in the sense that the gene-by-gene expression ratios correlate rather well and that the most significantly differentially expressed genes are in good concordance. Obviously, these correlations would improve with greater replication, but already our results suggest that even simple comparisons can be achieved using a common reference.

Although genomic DNA does have a number of properties that would make it a good potential reference source, our results show that it does not perform nearly as well in the estimation of inferred comparisons as does pooled RNA. It should be noted, however, that even genomic DNA was able to identify the highly and significantly differentially expressed genes. This suggests that one could use genomic DNA as an alternative in specific applications when sufficient RNA is unavailable or a suitable RNA reference does not exist. However, any results derived from these analyses must be carefully validated, as they deviate from those obtained using RNA-based reference sources.

ACKNOWLEDGMENTS

We wish to thank N.H. Lee, S. Wang, and H. Wang for helpful discussions and comments on the manuscript and J. White, V Sharov, A.I. Saeed, and J. Li for bioinformatics support for the microarray work. The authors also wish to thank M. Heaney and S. Lo for database support, and V. Sapiro, B. Lee, J. Shao, S. Gregory, C. Irwin, J. Neubrech, R. Kramchedu, M. Sengamalay, and E. Arnold for computer system support. This work was supported by a US National Science Foundation grant no. NSF 9975920 to J.Q.

REFERENCES

1. **Chen, W. and K.B. Singh.** 1999. The auxin, hydrogen peroxide, and salicylic acid induced expression of the *Arabidopsis* GST6 promoter is mediated in part by an ocs element. Plant J. *19*:667-677.
2. **Grant, J.J., B.W. Yun, and G.J. Loak.** 2000. Oxidative burst and cognate redox signalling reported by luciferase imaging: identification of a signal network that functions independently of ethylene, SA, and Me-JA but is dependent on MAPKK activity. Plant J. *24*:569-582.
3. **Hegde, P., R. Qi, K. Abernathy, C. Gay, S. Dharap, R. Gaspard, J. Earle-Hughes, E.C. Snesrud, et al.** 2000. A concise guide to cDNA microarray analysis. BioTechniques *29*:548-562.
4. **Lamb, C. and R.A. Dixon.** 1997. The oxidative burst in plant disease resistance. Annu. Rev. Plant Physiol. Plant Mol. Biol. *48*:251-275.
5. **Lin, X., S. Rounsley, T.P. Shea, M.I. Benito, C.D. Town, C.Y. Fujii, T. Mason, et al.** 1999. Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. Nature *402*:761-768.
6. **Murashige, T. and F. Skoog.** 1962. A revised medium for rapid growth and bioassays with tobacco tissue culture. Physiol. Plant *15*:473-497.
7. **Perou, C.M., T. Sorlie, M.B. Eisen, M. van de Rijn, S.S. Jeffrey, C.A. Rees, J.R. Pollack, D.T. Ross, et al.** 2000. Molecular portraits of human breast tumours. Nature *406*:747-752.
8. **Quackenbush, J.** 2001. Computational analysis of cDNA microarray data. Nat. Rev. Genet. *2*:418-427.
9. **Reymond, P., H. Weber, M. Damond, and E.E. Farmer.** 2000. Differential gene expression in response to mechanical wounding and insect feeding in *Arabidopsis*. Plant Cell *12*:707-719.
10. **Schena, M., D. Shalon, R.W. Davis, and P.O. Brown.** 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science *270*:467-470.

Received 26 March 2002; accepted 28 May 2002.

Address correspondence to:

Dr. John Quackenbush
The Institute for Genomic Research
9712 Medical Center Drive
Rockville, MD 20850, USA
e-mail: johnq@tigr.org