Large-Scale Profiling of the Arabidopsis Transcriptome

Tong Zhu and Xun Wang*

Novartis Agricultural Discovery Institute, Inc., 3115 Merryfield Row, San Diego, California 92121

DNA microarray is a powerful technology for parallel analysis of gene expression (Brown and Bostein, 1999). Since microarray technology emerged 5 years ago (Schena et al., 1995), the number of genes that can be monitored by this technology has increased from several hundreds (Yuan et al., 1998; Aharoni et al., 2000; Reymond et al., 2000) to several thousands (Arabidopsis Functional Genomics Consortium Microarray, http://afgc.stanford.edu/afgc_html). At Novartis Agricultural Discovery Institute, Inc. (NADII), microarray is an important component in our toolbox for transcription profiling. In addition, we have developed and are using other gene expression monitoring technology platforms for gene expression profiling (emphasizing coverage) and gene expression diagnostics (emphasizing throughput). These technologies include serial analysis of gene expression, cDNA fingerprinting, and microbead-based liquid microarray.

Here we focus on the microarray technologies used at NADII. It should be noted that there are at least two nomenclature systems that have been used to describe the hybridization partners in the microarray field. Consistent with the common nomenclature (Duggan et al., 1999; Lipshultz et al., 1999; Southern et al., 1999), we use "probe" to refer to the tethered nucleic acid molecules used to interrogate the experimental samples and "target" to refer to the free hybridization partner in the experimental sample.

Our gene expression microarray program has two technology platforms: oligonucleotide-based probe array (GeneChip) and cDNA-based array. For the purpose of gene discovery using Arabidopsis as a model system, a large-scale profile of its transcriptome is needed. We selected the oligonucleotide-based array (Lockhart et al., 1996) as our primary platform technology because of the following reasons. First, the transcript abundance of each gene can be accurately measured by multiple probe pairs. Second, the data can be produced with a moderate throughput and a large scale. The designed arrays are commercially manufactured using photolithography technology, therefore, the time-consuming and labor-intensive array fabrication process is eliminated. Moreover, human errors that often occur during the clone tracking process are also eliminated. Third, standardized data produced by the array can be easily normalized and interrogated. This is important when cross-project comparison is needed. And finally, the necessary genomic information for oligonucleotide probe selection is available from the Arabidopsis genome sequencing project (Lin et al., 1999; Mayer et al., 1999).

To profile the Arabidopsis transcriptome on a large scale, in addition to designing a high-density oligonucleotide probe array, we also tested and developed protocols for sample preparation; we developed the laboratory information management system (LIMS) for project, sample information, and data management; and we developed and integrated a number of analysis tools for data mining.

DESIGN AND CHARACTERIZATION OF THE ARABIDOPSIS GENOME ARRAY

To design an Arabidopsis oligonucleotide probe array, high-quality unique gene sequences must be obtained for probe selection. The quality of the sequences is critical because any mismatch introduced in the short oligonucleotide probes, in addition to the one in the mismatch probes, may significantly reduce the hybridization signal. For this reason Arabidopsis genomic sequences were used. Gene sequences were selected based on computational prediction and reference from matching expressed sequence tags (ESTs) and protein sequences. Predicted open reading frames in the bacterial artificial chromosomes were confirmed by blasting against the Arabidopsis EST database and SwissProt protein database. Sequences of known genes and approximately 100 high-quality EST clusters were also added to the collection. Redundant sequences and introns were then eliminated computationally. This approach ensured the sequence quality of the unigene set, although it may be biased toward abundantly expressed genes.

The final array contains probes from more than 8,000 Arabidopsis genes and 40 probes for spiking and negative controls. For each gene there are 16 probe pairs (probe sets) including perfect match probes and mismatch probes for cross-hybridization control. Among Arabidopsis genes presented in the array, about 70% are genes with known or predicted function and 30% are predicted genes with matching ESTs or proteins. There are approximately 700 genes with multiple probe sets because a single representative high quality probe set cannot be found (Table I).

The quality of the array was characterized by calculation of the rate of false changes (number of genes significantly changed over the total number of genes on the array; Lipshultz et al., 1999). Two cDNA and

* Corresponding author; e-mail xun.wang@nadii.novartis.com; fax 858–812–1097.

Table I. *Gene probes included in the Arabidopsis genome array*

Total Probe Sets	8875
Control probe sets	40
Arabidopsis probe sets	8,835
Arabidopsis genes	~8,100
cRNA sample quality control probe sets	9
Known function genes	~5,000
Hypothetical or unknown function genes	~3,100

subsequently cRNA (the antisense RNA synthesized by in vitro transcription using cDNA as templates in the presence of biotinylated ribonucleotides) samples were prepared in parallel from the same total RNA samples and hybridized to two different arrays manufactured in the same lot or different lots. Genes that showed changes of ≥2-fold and a signal threshold above the background (calculated according to the setting of the global scaling factor) were counted as false changes. Data from 15 pairs of array experiments indicated that false changes between two experiments using arrays of the same lot is 0.17% (based on eight pairs), whereas the false change using arrays of two different lots is 0.22% (based on seven pairs). In other words, approximately 16 to 20 genes among the 8,300 Arabidopsis genes may potentially show false change when an experiment is duplicated. Further analyses of these genes indicate that the fold change and expression levels are low and close to the threshold (Fig. 1).

The probe set quality was validated by hybridizing genomic DNA to the probe array. When a cRNA sample was hybridized to the array, gradient hybridization signals were observed. This gradient pattern could be due to the probe synthesis or the arrangement of the probes and expression level detected. To clarify this issue, fragmented and labeled genomic DNA (Winzeler et al., 1999) of Arabidopsis (Col-0) was hybridized to the array. As expected, an even hybridization signal was observed. This result indicated that the unevenness of the hybridization signal is indeed representing the relative amount of the transcripts. Further analysis indicated that 98% to 99% of the gene probe sets were hybridized by the genomic DNA and give a "present" call, suggesting a high affinity with the gene sequences.

The quality of the total RNA and subsequently synthesized cDNA and cRNA samples has direct impact on the array results. When we compared data generated from the same tissue samples with different total RNA extraction methods, a greater variation was observed. To control RNA quality, standard protocols were developed and quality control criteria for total RNA preparation and cRNA synthesis were established. In addition, selected housekeeping genes are used to ensure the quality of the array experiments. Probe sets were designed at 3′, middle, and 5′ end of the GAPDH and ubiquitin11 gene sequences. By comparing the ratio of the hybridization signal of 3′ and 5′ probe sets, one can deduce the quality of the

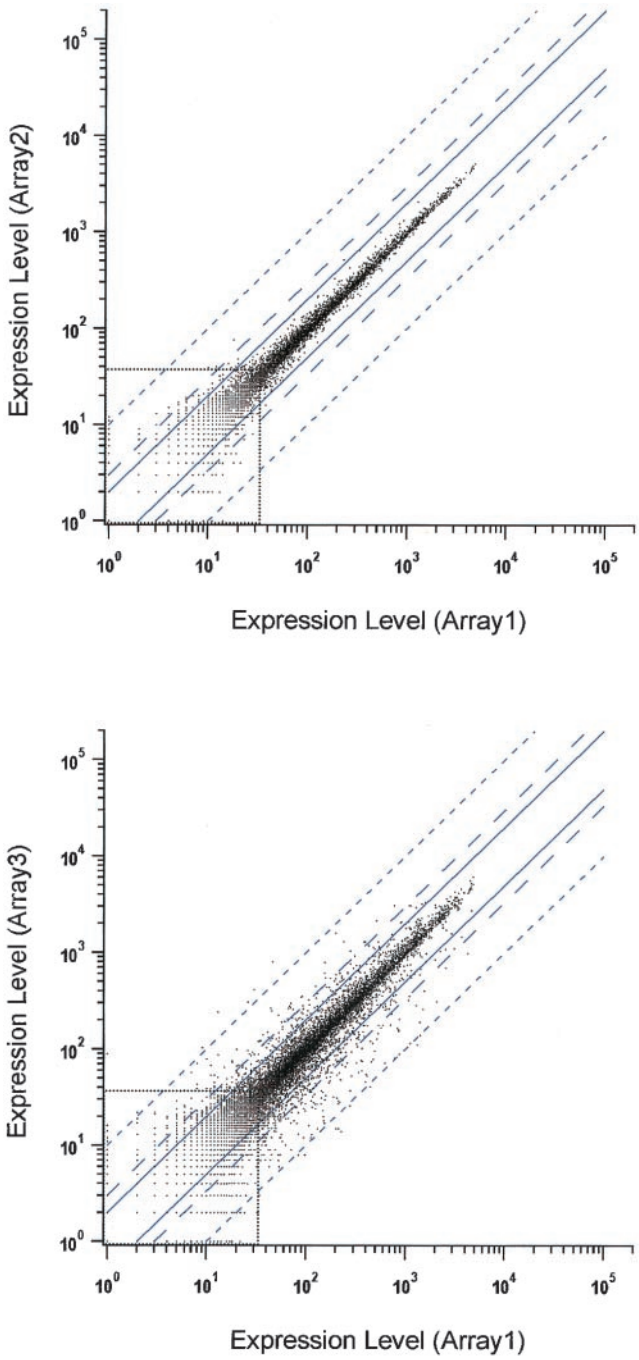

Figure 1. Two scatter plots showing the representative reproducibility of the microarray experiments (A) and the detection of the differentially expressed genes (B). Expression level of each gene, measured by average difference of hybridization signal intensity between perfect match and mismatch probes from duplicate arrays, was plotted. A, Two cRNA samples were independently prepared from the same total RNA sample and hybridized to different arrays. B, Two cRNA samples were prepared from two biological samples with two different treatments and hybridized to different arrays. Two-, 3-, and 10-fold changes in expression level between samples were indicated by the solid, long, and short dash lines. The dotted lines indicated the noise level.

labeled cRNA. Based on the data collected from 75 experiments, a consistent 3′/5′ ratio was obtained (Table II). These data validate our sample preparation procedure. Depending on the biological sample, approximately 60% to 68% of the total probe sets usually hybridize to the gene transcripts and are therefore called "present." A series of spiking experiments was conducted to determine the working dynamic range and sensitivity of the detection. The linear dynamic range is determined as 500-fold. Within this range the sensitivity of the Arabidopsis genome array is 1:100,000 to 300,000 (E. Tanimoto, personal communication).

With stringent quality controls the detected biological variations usually are greater than the technical variations appeared during the microarray experiments. To minimize the biological variations, pooling samples from individual plants is always necessary. Adding biological replications is also recommended in some experiments when large biological variations are expected.

IMPROVEMENT OF SAMPLE PREPARATIONS

A weakness of current microarray technologies is the inability to detect low abundant transcripts because of the limited dynamic range of detection and sensitivity (Bertucci et al., 1999; Lipshultz et al., 1999). This is in part due to a major obstacle in preparing high quality targets for microarray detection, which is the low efficiency of reverse transcriptase (RT) in synthesizing full-length cDNAs from the transcripts. Secondary structure elements intrinsic to many RNA transcripts impede access or terminate synthesis by RT altogether, which leads to a RT bias and reduces microarray sensitivity. To alleviate problems associated with RT bias and the efficiency of target cDNA synthesis we used a thermostable RT for first strand cDNA synthesis and demonstrated reproducible microarray detection with enhanced sensitivity and specificity. A reproducible 25% increase in the overall signal intensity and a 5% increase in the genes called present was evidenced when a thermostable RT was employed to generate a mRNA profile (H. Chang, B. Read, T. Yen, H. Dong, X. Wang, and T. Zhu, unpublished data).

Another challenge of the DNA microarray is the sample preparation process. It is time consuming and labor intensive. We use parallel sample preparation approaches to improve the throughput. By modifying the standard protocol recommended by Affymetrix, a number of samples were prepared in parallel in 96-well plates from total RNA to cRNA. Comparable results with those of standard methods were obtained. The false change is approximately 0.25%, slightly higher than the false change rate of the standard method (0.2%).

Table II. *Gene probes designed for quality control of prepared cRNA samples*

Data of the 3′/5′ ratio were collected from 75 independent array experiments.

Gene	Length	Ratio (3′/5′)
	bp	
GAPDH	1,295	1.39 ± 0.09
UBQ4	1,149	1.37 ± 0.17
UBQ11	1,140	1.82 ± 0.06

INFORMATION MANAGEMENT AND DATA ANALYSIS

Experiments using high-density microarrays produce large amounts of gene expression information regarding the biology of the sample. To manage and analyze the massive information and data generated from the microarray experiments, an integrated LIMS is needed.

A web-based LIMS has been developed for project management, sample submission, sample processing, sample tracking, data retrieving, sorting, visualization, and clustering. The sample information including genotype, treatments, and detailed growth conditions is standardized for comparison within our expression database and with data in the public domain. Data archived in the database are globally scaled to the same level for direct comparison and vertical search across many experiments. Data can be further normalized and selected based on their expression level and fold change. Internally developed tools, academic software, and commercial analysis software, such as Cluster and TreeView (Stanford University, CA), Cluster Analysis (Whitehead Institute, Massachusetts Institute of Technology, Cambridge, MA), GeneChip Suite (Affymetrix, Inc., Santa Clara, CA), Spotfire (Spotfire, Inc., Cambridge, MA), and GeneSpring (SiliconGenetics, Redwood City, CA) are used for pattern recognition and motif search during the data mining process.

ARABIDOPSIS TRANSCRIPTION PROFILES

With the high-density oligonucleotide probe array and improved sample preparation methods, more than 500 Arabidopsis transcription profiles were produced. These profiles consist of over four million data points and describe the gene expression patterns in different organs or tissues of Arabidopsis in various genetic conditions and growth environments.

The gene expression patterns in normal cells and tissues provide useful information about function. Tissue- and development-specific expressed genes, and constitutive expressed genes could be easily identified (Yuan et al., 1998). In a study conducted recently we analyzed the global expression pattern of over 8,000 genes in six major organs at different representative stages. The cluster analysis of the data indicated that because of the organ-specific gene expression, the organ samples are organized into clusters according to their functions. By such cluster anal-

yses, organ-specific expressed genes are easily identified.

One of the most attractive applications of microarrays is characterization of plant-microbial pathogen interaction and the subsequent seeking of effective means for plant disease control. Gene expression patterns of resistant and susceptible plants, mutants, or transgenic plants, with or without pathogen inoculation, can be compared to identify genes involving common stress, pathogenesis, and resistance. For example, clusters of genes involved in systemic acquired resistance and disease resistance were recently identified. These genes share common regulation patterns, or regulons, which contains PR-1, a reliable marker gene for systemic acquired resistance in Arabidopsis (Maleck et al., 2000).

By monitoring global gene expression changes between control and chemical treated Arabidopsis plants, metabolic pathways affected by the treatments, such as herbicides, could be identified and dissected. Mechanism or toxicity potentials of agricultural chemicals, including hormones, herbicides, fungicides, and insecticides could be characterized. Such an approach has been successfully applied in examining drug effects on gene expression in yeast (Gray et al., 1998).

The Arabidopsis transcription profiles have demonstrated their great value for gene discovery and regulatory pathways characterization. Pair-wise comparisons from individual projects are certainly a powerful way for gene discovery. However, a large expression database with normalized expression data from samples collected under different experimental conditions will be even more valuable for pattern identification and target search. At NADII the expression database of transcription profiles, as well as proteomic and metabolite profiles will be used in combination of reverse genetics tools for gene discovery.

RESOURCES AND ACADEMIC ACCESS

At NADII the high-density oligonucleotide probe array is the primary technology platform for transcription profiling. The Arabidopsis genome array, as the first high-density oligonucleotide probe array designed for plants, has already demonstrated its power in gene discovery. However, it only covers approximately one-third of the genome. To interrogate the gene expression pattern on a true genome scale, ideally, the second generation of the Arabidopsis oligonucleotide probe array should host probes for approximately 25,000 genes in the limited space (1.28×1.28 cm). This will require a reduction of the feature size, reduction of the probe pairs per genes, or other modifications of the current GeneChip technology. In collaboration with Affymetrix, this new array is currently under development.

Although Arabidopsis serves as an excellent model organism for dicotyledonous plants, rice has been proven as an ideal model system for cereal crops. Rice is economically important. It has the smallest genome (400 Mb) and shares a high degree of conservation of gene content and order with major cereal crops (Devos and Gale, 2000). Using the sequences from the NADII's Cereal Genomics program a high-density rice oligonucleotide probe array has been designed and it will be available for transcription profiling experiments in 2001.

In addition to the high-density oligonucleotide probe arrays, a number of complementary transcription profiling technologies including spotted DNA microarray and cDNA fingerprinting are also in place at NADII. Because the spotted DNA arrays can be fabricated to meet special needs, they provide alternative means to monitor the gene expression in a large scale for profiling purposes or for small scale diagnostic purposes. Plant species that lack extensive gene sequences or pre-made high-density oligonucleotide probe arrays could especially benefit by this approach (Aharoni et al., 2000).

NADII values global agricultural research efforts and academic collaborations. We are delighted to contribute our custom Arabidopsis genome array to the general public via Affymetrix. We actively seek opportunities for collaboration with academia. In fact, approximately 50% of the transcriptional profiling projects are academic collaboration. These collaborative projects were developed based on the mutual interests of NADII and our collaborators. To participate in the program, researchers are encouraged to submit a research proposal. The research proposal should include: (a) research background and objectives; (b) proposed experiments; (c) significance of the research and its potential impact in terms of agriculture; (d) research timeline; and (e) selected references. The proposals will be selected according to the scientific merits, importance, and potential applications in agricultural and related fields. Additional criteria will include the preliminary research conducted and the number of arrays required. For selected projects, NADII will bear the cost of array experiments and data analysis conducted at NADII. The collaborative research agreement grants the rights of accessing to raw data and publishing the results to the academic collaborators under certain terms. For more information, please visit our website under "academic collaboration" at www.nadii.com.

ACKNOWLEDGMENTS

We thank members of the NADII microarray group, Hur-Song Chang, Bin Han, Yen Kim Tran, Betsy Read, and James Schmeits for contributing data described in this paper. We also thank Guangzhou Zou for developing the microarray LIMS; Darrell Ricke, E. Li, Roman Rozenshteyn, Dana Alcivare of NADII, Gene Tanimoto, Mike Mittmann, Walt Short, Liz Kerr, Trace Lane, Tarif Awad, Mike Troutman, Helin Dong of Affymetrix, and Ron Davis and George

Karlin-Neumann of Stanford University for their efforts in array design; David Lockhart, Steve Whitham, Liang Shi, and Helin Dong for their suggestions.

Received September 1, 2000; accepted September 18, 2000.

LITERATURE CITED

Aharoni A, Keizer LC, Bouwmeester HJ, Sun Z, Alvarez-Huerta M, Verhoeven HA, Blaas J, van Houwelingen AM, De Vos RC, van der Voet H, Jansen RC, Guis M, Mol J, Davis RW, Schena M, van Tunen AJ, O'Connell AP (2000) Identification of the SAAT gene involved in strawberry flavor biogenesis by use of DNA microarrays. Plant Cell **12:** 647–662

Bertucci F, Bernard K, Loriod B, Chang YC, Granjeaud S, Birnbaum D, Nguyen C, Peck K, Jordan BR (1999) Sensitivity issues in DNA array-based expression measurements and performance of nylon microarrays for small samples. Hum Mol Genet **8:** 1715–1722

Brown PO, Bostein D (1999) Exploring the new world of the genome with DNA microarrays. Nat Genet **21:** 33–37

Devos KM, Gale MD (2000) Genome relationships: the grass model in current research. Plant Cell **12:** 637–646

Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM (1999) Expression profiling using cDNA microarrays. Nat Genet **21:** 10–14

Gray NS, Wodicka L, Thunnissen AM, Norman TC, Kwon S, Espinoza FH, Morgan DO, Barnes G, LeClerc S, Meijer L, Kim S-H, Lockhart DJ, Schultz PG (1998) Exploiting chemical libraries, structure, and genomics in search for kinase inhibitors. Science **281:** 533–538

Lin XY, Kaul S, Rounsley S, Shea TP, Benito MI, Town CD, Fujii CY, Mason T, Bowman CL, Barnstead M, Feldblyum TV, Buell CR, Ketchum KA, Lee J, Ronning CM, Koo HL, Moffat KS, Cronin LA, Shen M, Pai G, Van Aken S, Umayam L, Tallon LJ, Gill JE, Adams MD, Carrera AJ, Creasy TH, Goodman HM, Somerville CR, Copenhaver GP, Preuss D, Nierman WC, White O, Eisen JA, Salzberg SL, Fraser CM, Venter JC (1999) Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. Nature **402:** 761–768

Lipshultz RJ, Fodor SPA, Gingeras TR, Lockhart DJ (1999) High density synthetic oligonucleotide arrays. Nat Genet **21:** 20–24

Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittman M, Wang C, Kobayashi M, Horton H, Brown EL (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol **14:** 1675–1680

Maleck K, Levine A, Eulgem T, Morgan A, Schmid J, Lawton KA, Dangl JL, Dietrich RA (2000) The transcriptome of Arabidopsis during systemic acquired resistance. Nat Genet (in press)

Mayer K, Schuller C, Wambutt R, Murphy G, Volckaert G, Pohl T, Dusterhoft A, Stiekema W, Entian KD, Terryn N, Harris B, Ansorge W, Brandt P, Grivell L, Rieger M, Weichselgartner M, de Simone V, Obermaier B, Mache R, Muller M, Kreis M, Delseny M, Puigdomenech P, Watson M, Schmidtheini T, Reichert B, Portatelle D, Perez-Alonso M, Boutry M, Bancroft I, Vos P, Hoheisel J, Zimmerman W, Wedler H, Ridley P, Langham S-A, McCullagh B, Bilham L, Robben J, Van Der Schueren J, Grymonprez B, Chuang Y-J, Vandenbussche F, Braeken M, Weltjens I, Voet M, Bastiaens I, Aert R, DeFoor E, Weitzenegger T, Bothe G, Ramsperger U, Hilbert H, Braun M, Holzer E, Brandt A, Peters S, Van Staveren M, Dirkse W, Mooijman P, Klein Lankhorst R, Rose M, Hauf J, Kötter P, Berneiser S, Hempel S, Feldpausch M, Lamberth S, Van Den Daele H, De Keyser A, Buysshaert C, Gielen J, Villarroel R, De Clercq R, Van Montagu M, Rogers J, Cronin A, Quail M, Bray-Allen S, Clark L, Doggett J, Hall S, Kay M, Lennard N, Mclay K, Mayes R, Pettett A, Rajandream M-A, Lyne M, Benes V, Rechmann S, Borkova D, Blöcker H, Scharfe M, Grimm M, Löhnert T-H, Dose S, De Haan M, Maarse A, Schäfer M, Müller-Auer S, Gabel C, Fuchs M, Fartmann B, Granderath K, Dauner D, Herzl A, Neumann S, Argiriou A, Vitale D, Liguori R, Piravandi E, Massenet O, Quigley F, Clabauld G, Mündlein A, Felber R, Schnabl S, Hiller R, Schmidt W, Lecharny A, Aubourg S, Chefdor F, Cooke R, Berger C, Montfort A, Casacubert AE, Gibbons T, Weber N, Vandenbol M, Bargues M, Terol J, Torres A, Perez-Perez A, Purnelle B, Bent E, Johnson S, Tacon D, Jesse T, Heijnen L, Schwarz S, Scholler P, Heber S, Francs P, Bielke C, Frishman D, Haase D, Lemcke K, Mewes HW, Stocker S, Zaccaria P, Bevan M, Wilson RK, De La Bastide M, Habermann K, Parnell L, Dedhia N, Gnoj L, Schutz K, Huang E, Spiegel L, Sehkon M, Murray J, Sheet P, Cordes M, Abu-Threideh J, Stoneking T, Kalicki J, Graves T, Harmon G, Edwards J, Latreille P, Courtney L, Cloud J, Abbott A, Scott K, Johnson D, Minx P, Bentley D, Fulton B, Miller N, Greco T, Kemp K, Kramer J, Fulton L, Mardis E, Dante M, Pepin K, Hillier L, Nelson J, Spieth J, Ryan E, Andrews S, Geisel C, Layman D, Du H, Ali J, Berghoff A, Jones K, Drone K, Cotton M, Joshu C, Antonoiu B, Zidanic M, Strong C, Sun H, Lamar B, Yordan C, Ma P, Zhong J, Preston R, Vil D, Shekher M, Matero A, Shah R, Swaby IK, O'Shaughnessy A, Rodriguez M, Hoffman J, Till S, Granat S, Shohdy N, Hasegawa A, Hameed A, Lodhi M, Johnson A, Chen E, Marra M, Martienssen R, McCombie WR (1999) Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. Nature **402:** 769–777

Reymond P, Weber H, Damond M, Farmer EE (2000) Differential gene expression in response to mechanical wounding and insect feeding in Arabidopsis. Plant Cell **12:** 707–719

Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science **270:** 460–470

Southern E, Mir K, Shchepinov M (1999) Molecular interactions on microarrays. Nat Genet **21:** 5–9

Winzeler EA, Richards DR, Conway AR, Goldstein AL, Kalman S, McCullough MJ, McCusker JH, Stevens DA, Wodicka L, Lockhart DJ, Davis RW (1998) Direct allelic variation scanning of the yeast genome. Science **281:** 1194–1197

Yuan Y, Gilmore J, Conner T (1998) Towards Arabidopsis genome analysis: monitoring expression profiles of 1400 genes using cDNA microarrays. Plant J **15:** 821–833