Chasing the dream: plant EST microarrays
Todd Richmond* and Shauna Somerville†

DNA microarray technology is poised to make an important contribution to the field of plant biology. Stimulated by recent funding programs, expressed sequence tag sequencing and microarray production either has begun or is being contemplated for most economically important plant species. Although the DNA microarray technology is still being refined, the basic methods are well established. The real challenges lie in data analysis and data management. To fully realize the value of this technology, centralized databases that are capable of storing microarray expression data and managing information from a variety of sources will be needed. These information resources are under development and will help usher in a new era in plant functional genomics.

Introduction
Microarrays are widely recognized as a significant technological advance providing genome-scale information about gene expression patterns [1,2,3,4,5]. Two types of microarrays currently exist: the DNA-fragment-based and the oligonucleotide (oligo)-based microarrays. These differ primarily in the nature of the DNA fixed to the solid support. In this review, we focus on the DNA-fragment-based arrays, in which PCR-amplified inserts of partially sequenced cDNA clones are spotted onto glass microscope slides [7].

A number of reviews describe the methods and outline the advantages of DNA-fragment microarrays [2,3,5,6]. The major advantages of microarrays derive from the fact that data can be collected for large numbers of genes in one experiment permitting true genome-scale sampling of gene expression patterns. It is important to note that microarrays, like northern blots, provide a measure of steady-state RNA levels. Although such information can be very valuable, not all biological processes will be regulated at this level of gene expression and other methods are needed to characterize processes regulated at other levels.

Three general types of experiments can be performed with microarrays. In ‘marker discovery’ experiments, the goal is to discover a limited number of highly specific marker genes for a cell type, a developmental stage or an environmental treatment. In such experiments, the researcher is often interested only in genes that show a dramatic and selective induction or repression of expression. The second type of experiment, an exploratory or ‘biology discovery’ experiment, is one in which expression information about all genes is used to provide an integrative view of a plant’s response to a treatment. Such experiments have the potential to provide new insights into biological processes. In an example drawn from human biology, Iyer et al. [10] demonstrated that serum-starved cells respond to the addition of fresh serum in part by inducing a suite of genes involved in wounding. The wound response had not previously been implicated in this widely studied biological model. The third type of experiments is the ‘gene-function discovery’ experiments. As microarray data accumulates in centralized public databases, hypotheses as to the function of novel, uncharacterized genes can be generated by in silico analyses of when and where such genes exhibit altered expression. These kinds of experiments will be very beneficial to the field of plant biology in terms of both developing a broader view of how plants function and rapidly identifying novel genes most worthy of detailed characterization. In this review, we highlight some of our observations on the use of expressed sequence tag (EST)-based microarrays, and, in particular, on some of the plant-specific concerns, on the basis of our experience with Arabidopsis microarrays.

Construction of microarrays
The advantage of constructing microarrays from EST clones rather than from anonymous clones is that the subsequent analysis of gene expression patterns is much more informative. EST clones are the obvious choice for Arabidopsis because of the availability of a large collection from a public stock center — the Arabidopsis Biological Resource Center. There are a few limitations associated with this collection that should be noted. Most of the Arabidopsis ESTs consist of only 300-400 basepairs sequenced from the 5’ end of a longer cDNA clone [11]. Thus, the occurrence of chimeric clones consisting of sequences from unrelated genes will go undetected. Genes specifically induced under specialized environmental conditions or in specialized cell types will be under-represented in the EST collection, which was prepared from RNA from only four plant tissues. Recent estimates suggest that only a third of Arabidopsis genes [12,13] are represented among the current set of ~45,000 Arabidopsis ESTs.

The complete sequence of the Arabidopsis genome is scheduled to be available in 2000 [14], and the first comprehensive annotation to identify candidate genes will be completed shortly afterwards [15,16]. Thus, for
Arabidopsis, genes not represented by an EST clone can be recovered from genomic sequence information. It should be noted that the computer programs used to identify candidate genes are imperfect, and it may take some time to fully identify and describe all Arabidopsis genes properly. Therefore, Arabidopsis microarrays constructed in the near future may include sequences that are not expressed and there may be missing genes that are unusual enough to be missed by gene-finding software programs.

Newer EST projects have noted the problems with the Arabidopsis EST collection and will sequence a larger number of clones from larger collections of diverse libraries. For example, The Institute for Genomic Research will sequence 150,000 tomato clones [17], and 300,000 soybean clones will be sequenced as part of the Soybean EST project [18]. Targets that exhibit expression patterns of interest to the user can be sequenced at the end of the expression experiment. One additional consideration that is unique to DNA-fragment-based microarrays is that handling mistakes during the production phase can lead to DNAs being mislabeled or contaminated before spotting on the slides. In most commercial microarray facilities, the DNAs are re-sequenced as part of the quality assurance process. In the absence of such quality control, microarray users should verify key microarray results (see section on Data verification).

Cross-hybridization between related sequences is another limitation of using EST clones to construct microarrays. Cross-hybridization to a weakly expressed gene by a probe for a highly expressed, related gene will confound data collected for the weakly expressed gene. Current estimates are that sequences with >70% sequence identity over >200 nucleotides are likely to exhibit some degree of cross-hybridization under standard conditions [19•]. As estimates of Arabidopsis genes occurring in gene families range from >12% [13] to 50% [12], this problem is of substantial concern. Thus, in future projects, there is some advantage to developing a non-redundant set of completely sequenced cDNA clones. For Arabidopsis, the availability of genomic sequence will help address some of the limitations associated with the EST clones. Primer pairs can be designed to amplify gene-specific fragments for each member of a gene family for spotting on microarrays.

**Sampling and experimental design**

The composite nature of plant organs adds additional constraints to the interpretation of microarray data. Gene expression in plant leaves, for example, is a composite of the expression patterns of various cell types (e.g. mesophyll, epidermal, vascular) and of various spatially distinct parts (e.g. upper cell layers versus lower cell layers). At present, our ability to sample single cells or a small number of uniform cells is restricted. Also, it is difficult or impossible
to obtain highly synchronized populations of *Arabidopsis* cells. These limitations should be considered when interpreting microarray expression data.

As with any experiment, experimental design and sample replication should be considered when designing microarray experiments. One of the strengths of the microarray technique is that relatively minor changes in gene expression can be measured. However, the resolving power of each experiment will be determined by the biological variation in the plant samples and the technical variation associated with the microarray technology (see Probe preparation and hybridization).

In designing an experiment, researchers should consider whether they wish to use the microarrays for ‘marker discovery’ or whether they wish to gain genome-scale knowledge of a biological process in ‘biology discovery’ experiments. In the first case, a limited number of genes are identified as of interest from a microarray experiment and then used in subsequent experiments to characterize a specific treatment or developmental stage. Such an experiment might consist of several replicates of two samples harvested at one time point under a single condition. A limited number of microarrays are required and the microarray data serves as the starting point for a project rather than the endpoint. In the second case, in which a detailed profile of gene expression is desired, a large number of samples are generated during the course of analyzing a variety of time points, treatment levels, developmental stages or genetic lines. Such experiments are extremely rich in information, although costly to perform.

### Probe preparation and hybridization

The design of experiments and collection of sample materials is only the first step in microarray experimentation. There is a whole range of technical issues involved with microarray hybridization and the collection of hybridization signals [7••]. Only a few of these points are specific to plant EST microarrays, and thus we will address them only briefly here.

Typically, two samples (e.g. from treated and control plants) are labeled with two distinct fluorochromes, mixed and hybridized simultaneously to the DNA-fragment based microarrays. Thus, DNA-fragment microarrays are best suited to side-by-side comparisons of the two samples hybridized to the same slide (Type I experiments [7••] or ‘marker discovery’ experiments). The data values are often expressed as a ratio of the signals collected for the two samples. Comparisons of multiple samples can be accomplished by using one sample as a common reference (Type II experiments [7••] or ‘biology discovery’ experiments). Thus, for each given gene, the ratio of the fluorescent signals derived from the two samples provides a relative measure of gene expression (i.e. whether mRNA accumulation is higher, lower or unchanged in one sample relative to the second sample).

The original methods for preparing a probe required >1 µg of poly(A)+ RNA [7••]; however, it can be difficult to obtain this much mRNA from specialized tissues or cell types. In the future, methods of probe preparation that are based on total RNA or adapted for use with very small amounts of RNA may be more practical.
amounts of mRNA, and that use more efficient methods of label incorporation, will extend the utility of microarray analysis to difficult samples.

One of the chief technical specifications of microarrays is sensitivity. Sensitivity, determined by the minimal signal that can be reliably detected above background, is dependent in part on background levels and on the specific incorporation of label into the probes. Claims have been made that EST microarrays can detect 1 part in 100,000 [20•] to 500,000 [21]. Incyte Pharmaceuticals indicates that with their GEM microarrays [22•], 2 pg out of 600 ng of poly(A) RNA (1:300,000) can be detected [23•]. These levels are generally considered sufficient to detect a mRNA present at a few copies per cell [20•].

Another measure of sensitivity is the ability to reliably determine minor changes in mRNA levels between two samples. Incyte suggests the technical variation associated with their GEM arrays, as measured by the coefficient of variation, is 15% [23•]. Our own experiments, using replicate 2500-element Arabidopsis arrays and experimental samples, show an average coefficient of variation that is closer to 37%. Incyte GEM arrays can be used to determine a differential expression greater than 1.75-fold. In our laboratory, we have established standards for minimal expression levels of more than three standard deviations above the background of the slide and for differential expression of more than three-fold when analyzing three independent replicates.

Data collection

Microarray hybridization signals are collected by specialized scanners, which are essentially computer-controlled inverted scanning fluorescent microscopes, with most models providing confocal imaging. The output of these readers is typically a 16-bit TIFF image file, limiting the dynamic range of the data to less than five orders of magnitude. The adoption of charge-coupled device (CCD)-based detectors [24] may increase the range of signal detection as improved methods and slide coatings [25] reduce background noise. The TIFF files are processed with image analysis software (often supplied with the microarray reader) to produce a measure of the hybridization signal for each DNA sample on the array. The ideal image analysis software package would be able to detect automatically each valid spot on the array, flag bad spots, subtract the background, and determine the expression levels with high accuracy and precision.

### Table 3

<table>
<thead>
<tr>
<th>Microarray data analysis software.</th>
<th>Name</th>
<th>Type</th>
<th>Reference/URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>QuantArray</td>
<td>Data visualization</td>
<td>GSI Luminica</td>
<td><a href="http://www.genscan.com/lifescience_frame/software.htm">http://www.genscan.com/lifescience_frame/software.htm</a></td>
</tr>
<tr>
<td>Stingray</td>
<td>Data visualization, clustering</td>
<td>Molecular Applications Group</td>
<td><a href="http://www.mag.com">http://www.mag.com</a></td>
</tr>
<tr>
<td>Array Vision</td>
<td>Data extraction, visualization</td>
<td>Imaging Research, Inc.</td>
<td><a href="http://www.imagingresearch.com">http://www.imagingresearch.com</a></td>
</tr>
<tr>
<td>Spotfire</td>
<td>Data visualization, clustering</td>
<td><a href="http://www.spotfire.com">http://www.spotfire.com</a></td>
<td></td>
</tr>
<tr>
<td>GeneCluster</td>
<td>Clustering using self-organizing maps</td>
<td><a href="http://www.genesys.mit.edu/MPL/GeneCluster/GeneCluster.html">http://www.genesys.mit.edu/MPL/GeneCluster/GeneCluster.html</a> [34]</td>
<td></td>
</tr>
<tr>
<td>GeneSpring</td>
<td>Data visualization; clustering; database links</td>
<td>Silicon Genetics</td>
<td><a href="http://www.silicongenetics.com/GeneSight/">http://www.silicongenetics.com/GeneSight/</a></td>
</tr>
<tr>
<td>GenEx</td>
<td>Public web database; data storage; data visualization</td>
<td>National Center for Genome Resources; Silicon Genetics</td>
<td><a href="http://www.ncgr.org/research/genex/">http://www.ncgr.org/research/genex/</a>; <a href="http://genes.silicongenetics.com/">http://genes.silicongenetics.com/</a></td>
</tr>
<tr>
<td>arraySCOUT</td>
<td>Data visualization; clustering, database links</td>
<td>Lion Bioscience</td>
<td><a href="http://www.lionbioscience.com">http://www.lionbioscience.com</a></td>
</tr>
<tr>
<td>ArrayDB</td>
<td>LIMS</td>
<td>National Human Genome Research Institute</td>
<td><a href="http://www.nhgri.nih.gov/DIR/LCG/15K/HTML/">http://www.nhgri.nih.gov/DIR/LCG/15K/HTML/</a></td>
</tr>
<tr>
<td>LifeTools</td>
<td>LIMS</td>
<td>Incyte</td>
<td><a href="http://www.incyte.com/products/Liets/Loets.html">http://www.incyte.com/products/Liets/Loets.html</a></td>
</tr>
<tr>
<td>Microarray Database</td>
<td>LIMS</td>
<td>Stanford University</td>
<td><a href="http://genome-www.stanford.edu/8100/">http://genome-www.stanford.edu/8100/</a></td>
</tr>
</tbody>
</table>
local background, and output a mean signal-intensity above background for each spot [26,27].

Data analysis
The real heart of microarray experimentation and, ironically, the most underdeveloped part, is data analysis. When an experiment is finished, the researcher is left with a series of signal intensities from thousands of clones. The true potential of microarrays as tools for gaining a global, highly integrated view of plant biology cannot be realized without large-scale relational databases to house data from multiple experiments and sophisticated data mining and pattern-recognition software tools for extracting multidimensional information from microarray data sets [26].

For simple comparisons, as is typical for marker-discovery experiments, there are a variety of data analysis methods. For small data sets, spreadsheets and statistical software can be used to find differentially expressed genes. Our laboratory uses a collection of custom Perl scripts [28] to process and analyze microarray data (T Richmond, unpublished data). Companies that provide microarrays or microarray services often provide the necessary software and information to examine the experimental data (Table 3). For large-scale microarray projects, laboratory information management systems (LIMS) are essential. There are a number of microarray projects developing database tools for storing, processing and retrieving microarray data [27,29–33]. Although often freely available, the software tools being developed by these groups are not amenable to inexperienced researchers, requiring computer skills for their installation and maintenance that are beyond those of the typical biologist.

Regardless of the software program used, the first step in data analysis involves the normalization of data. In order to...
compare separate fluorescent channels, or the results from separate experiments, the raw data must be transformed into normalized values. Several different methods of data normalization can be used. One method identifies a standard set of ‘housekeeping’ genes with expression levels that remain relatively constant under a variety of different conditions and uses their aggregate expression level to adjust the fluorescent signal of each target to a normalized value. The challenge is to identify a significant number (20–200) of medium to highly expressed genes with constant expression over a wide range of experimental conditions for each plant species. This requires a relatively large initial set of data in order to choose such genes.

Another method of normalization, which is not dependent on a predetermined set of housekeeping genes, makes the assumption that under the conditions being tested, most genes will not change in expression, and uses the mean, mode or median fluorescence for each channel as a mean of normalization. Although this may work in most cases, there are instances where this assumption can break down. For specialized cDNA microarrays with a limited number of genes, many of the genes may change in expression and thus the addition of a large number of controls may be necessary. Figure 1 gives an example of the data collected from three different experiments, showing the range of different expression that can be seen.

Another simple method of normalization is to convert the fluorescence of each target spot into a normalized value that is based on the standard deviation (Student normalization) or the mean and standard deviation (Z-score normalization) of the fluorescence of each channel as a whole. (For Student normalization, $S = \chi_{\sigma}$; for Z-score normalization, $Z = \chi_{\sigma}$, where $\chi$ = sample value, $\mu$ = mean, and $\sigma$ = standard deviation of the channel data.) We carried out a series of experiments in triplicate to compare the methods of normalization described above. The results from the comparison of methods for three experiments, demonstrating that there is little difference among the normalization methods. The chief differences appear at the cutoff thresholds for differential expression (typically three-fold); small changes in the threshold (0.1–0.5-fold) can result in large changes in the number of differentially expressed genes. We have found that Student normalization to be the simplest to use and typically use it for our experiments.

Side-by-side comparisons quickly inspire larger experiments that are based on time-courses, dose-response series or multiple treatments. For examining multiple data sets, cluster analysis is the most common method. Depending on the resolving power of the data set and software programs, the number of clusters and their associated patterns may reflect the number of distinct responses associated with a treatment or developmental sequence — responses that may not always be reflected in cytological or physiological parameters that can be readily measured. As an example of such methods, cluster analysis of the expression data for the cell cycle in yeast revealed the expected stages of G0, G1, G2, S and M in this well-characterized experimental system.

Clustering is also a powerful tool for predicting the function of previously uncharacterized genes. The utility of this approach has been clearly demonstrated in a number of systems. Iyer et al. identified over 200 previously unknown genes whose expression was regulated during the response of fibroblasts to serum. The coordinate expression of uncharacterized genes with genes of known function provides clues to the function of the unknown genes (the ‘guilt-by-association’ rule for gene-function discovery). Plant biologists will be able to conduct similar experiments, classifying genes on the basis of their coordinated expression in response to environmental stresses or developmental stages. For plants with fully sequenced genomes, another outcome of such cluster analysis is the discovery of candidate cis-acting regulatory elements from the identification of sequence elements shared in common among all genes of a cluster. Such discoveries are likely to accelerate future discoveries of transcription factors and other DNA-binding proteins.

**Microarray databases and information management**

The uniform goal of microarray data analysis software is to provide an easy way for the researcher to identify potentially interesting candidate genes and to provide convenient, detailed information about these genes. What readily becomes apparent when studying thousands of genes simultaneously is that one has personal knowledge of only a small fraction of genes. Information that is readily accessible at the point of data analysis is therefore an extremely valuable aid in interpreting results from microarray experiments. Towards this goal, microarray databases should ideally contain links to a multitude of external information sources, including citation databases, sequence databases, metabolic pathway databases and protein domain databases. Where possible, microarray databases should also contain links to information regarding expression data collected from other sources, map position, predicted protein motifs, functional classifications, related genes in both plants and other organisms, known alleles and mutations, interacting protein partners and availability of DNA stocks, seed stocks and so on. For most plant species, including Arabidopsis, such information is now being collected in a systematic fashion.

The eventual goal is to have a unified and comprehensive information resource, which will link microarray expression
data to all available gene-related data. There are several groups working on establishing public repositories for microarray expression data [45,46]. The recently funded TMR (The Arabidopsis Information Resource) project [47], with its aim of providing a link from Arabidopsis to all other plants and *cucurbit* (S Rhee, C Somerville, personal communication), is the first step in this direction for plant biologists.

**Data verification**

Once candidate gene(s) of interest have been identified, researchers must decide how to proceed. As mentioned above, the gene identities of key genes should be verified by sequencing the spotted EST. In addition, many reviewers will not yet accept microarray data on its own, requiring secondary confirmation of differential expression.

There are a number of standard and newly emerging techniques that can be used to obtain secondary confirmation: RNA gel blots, quantitative or competitive PCR, and quantitative ribonuclease protection assays. Each has its own advantages, depending on the tissue and/or conditions being evaluated. RNA gel blots are simple and straightforward if sufficient amounts of tissue are available. Quantitative PCR, especially using real-time instruments such as the LightCycler [48] or the TaqMan [49] systems, is the method of choice if source material is limiting. Difficulties with primer design, genomic DNA contamination, and choice of internal controls, however, can make setting up these systems tedious, especially if the expression pattern of a large number of genes must be evaluated.

**Future prospects**

For Arabidopsis and plants of significant commercial interest (e.g. maize, soybean, rice), commercial vendors will probably produce microarrays for sale. There are many advantages to commercializing the production of microarrays. As long as the product is commercially viable, the companies can invest significantly more resources in producing large numbers of high-quality microarrays than can an academic laboratory. Ideally, all but the most specialized microarrays will be produced commercially in the future.

The future of *Arabidopsis* microarrays is likely to diverge from those to be developed for other plant species. With the completion of the genome sequence, full genome microarrays can be constructed. In addition, the availability of high-quality, error-free genomic sequence will permit the construction of oligo-based microarrays. The National Science Foundation has recently funded a group headed by J Ecker to produce an *Arabidopsis* gene expression oligo-array in collaboration with Affymetrix (Sunnyvale, California; Table 1). In addition to gene expression studies, oligo-based microarrays have been designed for use in scoring single nucleotide polymorphisms molecular markers in population surveys and in genetic crosses [50]. M Mindrinos in R Davis’ group (Stanford University) is currently developing such a ‘mapping chip’ for *Arabidopsis*.

One innovation is to use microarrays in ‘reverse Southern’ experiments to monitor changes in gene-copy number [51]. The application of these methods can potentially be very helpful in the analysis of the various kinds of somatic mutants that exist in plants. In addition, such an analysis could be extended to population genetic studies of gene evolution through duplication and deletions.

Further in the future, a method of generating ‘virtual masks’ using optical devices [52•,53•] and inkjet-based methods (Packard Instruments Company, Meriden, Connecticut) have been proposed as alternative methods to the established photolithography-based method for producing oligo-based microarrays or the spotting method for DNA-fragment microarrays. These technologies may lead to reductions in production costs and make the technology more accessible to all plant biologists. There are various other technical innovations being introduced that will enhance microarray experimentation, including multi-laser scanners that permit the use of more than two fluorescent dyes at a time (GSI Lumonics, Watertown, Massachusetts; Virtek, Woburn, Massachusetts), and new dyes with improved properties (Molecular Probes, Eugene, Oregon; New England Nuclear, Boston, Massachusetts).

**Conclusions**

DNA-fragment-based microarrays offer the promise of dramatically enhancing our understanding of all aspects of plant biology. Microarrays are proving to be very valuable for the initial functional characterization of unknown genes as well as enabling detailed analyses of already well-known processes. This technology will greatly facilitate the analysis of mutants and other genetic resources that are available for plants. The full potential of this technology depends on the establishment of public databases to house microarray expression data so that a maximum number of researchers can bring their expertise and intuition to bear on the interpretation of the expression data.

**Acknowledgements**

Supported in part by the US Department of Agriculture, the US Department of Energy and the Carnegie Institution of Washington. We also thank Robin Steavely (University of Illinois at Urbana-Champaign) for his helpful advice during data normalization. We thank Gérard de Boer, Greg Gallois, Wolf Schellhies and John Vogel for their comments and suggestions on the manuscript. TAR thanks C Somerville for his support of his research.

Publication # 1427 of the Carnegie Institution of Washington.

**References and recommended reading**

Papers of particular interest, published within the annual period of review, have been highlighted as:

* of outstanding interest
  • of special interest


A thoughtful commentary on the impact of DNA microarrays.

unknown genes by clustering with genes of known identity and function.

• Staudt LM, Hudson J Jnr, Boguski MS.

A collection of chapters covering all aspects of DNA microarray technology,

• 8. Kehoe D, Villand P, Somerville S.


22. Incyte Pharmaceuticals: Towards genomewide expression profiling: DNA microarrays.

Escherichia coli K-12.

A comprehensive list of brief descriptions and pointers to internet sites for various molecular biology databases.

The transcriptional control of gene expression on a genomic scale.

FTX (tagging and Genomes) Information Resource (TAIR).

Arabidopsis genome Annotation Database (AGAD).

Molecular biology database list.

A database for post-genome analysis.

Kyoto Encyclopedia of Genes and Genomes (KEGG).

A novel surface, attachment and gene expression arrays.


Both this paper and [53] outline an innovative use of new technology that may revolutionize oligonucleotide array construction.

53. Gardner S: Digital Optical Chemistry System. URL: http://pompous.swmed.edu/expbio/doc/ This website, like [52], describes a system that uses a unique UV light projector to manufacture biological/chemical arrays using UV photoreaction. This new technology may revolutionize oligonucleotide array construction.