Methods for Transcriptional Profiling in Plants. Be Fruitful and Replicate

Blake C. Meyers*, **David W. Galbraith**, **Timothy Nelson**, **and Vikas Agrawal**

Department of Plant and Soil Sciences and Delaware Biotechnology Institute, University of Delaware, Newark, Delaware 19711 (B.C.M., V.A.); Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, Connecticut 06511 (T.N.); and Department of Plant Sciences, University of Arizona, Tucson, Arizona 85721 (D.W.G.)

Because of the tractability of large-scale RNA measurements compared with protein studies, the first application of genomics in many organisms is to catalog and then measure transcriptional activity. Substantial investment in the US and abroad has led to dramatic growth in the availability of gene sequences for many plant species. With these sequences in hand, many molecular biologists are building the resources and technologies to enable large-scale transcriptional analyses for different plant species. The availability of the complete genome sequence of Arabidopsis made this the first plant for which transcriptional profiling platforms were developed. The experience gained from the applications of these technologies in Arabidopsis will shape the direction of similar experiments performed in other plant species.

The ability to simultaneously measure the expression of thousands of genes is a powerful analytical system, and the availability of technologies for this has presented scientists with many new opportunities. In most plant species, these experiments are being conducted largely with microarrays, although there are a growing number of alternative technologies. Some of these alternative technologies generate data that are distinct from and complementary to microarray data. The massive datasets generated by gene expression technologies present novel statistical and analytical problems, resulting in a convergence of biology, mathematics, and computer science. Users have developed a broad range of applications for the platforms, so that the use of microarrays has gone beyond simple measurements of relative transcript abundance to include genotyping, tissue classification, and pathway studies. Competition is intense among commercial microarray vendors vying in the plant market, and new companies join the fray on a regular basis. For laboratories working in plant species other than Arabidopsis, or for students and teachers of plant molecular biology, the question arises of what lessons to take away from the experience of this model plant, and how to best apply these technologies and approaches without squandering limited resources.

TECHNOLOGIES FOR MEASURING GENE EXPRESSION

The last decade has seen major advances in technologies for measuring gene expression. However, no method is without serious limitations, so many more advances will be required before we have achieved the necessary sensitivity and scope. The forerunner of many of the current methods is the RNA gel blot (northern), in which a labeled probe is hybridized to an RNA target, and the resulting band size and signal intensity is used to confirm and quantify expression. Advances in genomic technologies now permit the simultaneous analysis of thousands of genes, although many are based on the same concept of specific probe-target hybridization. These methods, described in more detail in this section, most prominently include DNA microarrays. However, sequencing-based methods are an alternative; these methods started with the use of expressed sequence tags (ESTs), and now include methods based on short tags, such as serial analysis of gene expression (SAGE) and massively parallel signature sequencing (MPSS). Differential display techniques provide yet another means of analyzing gene expression; this family of techniques is based on random amplification of cDNA fragments generated by restriction digestion, and bands that differ between two tissues identify cDNAs of interest. With a well-characterized genome, it is possible to match fragments to specific genes (Shimkets et al., 1999). Most differential display techniques require a large number of reactions to achieve maximal coverage of all active transcripts, and it is difficult to sample every transcript. Differential display-like approaches have been reviewed elsewhere (Green et al., 2001) and will not be discussed in detail in this review. All of these transcriptional profiling technologies permit the analysis of complex mRNA populations from selected cells or tissues, producing large-scale measurements of gene expression, but different technologies provide data with different uses. In fact, none of the existing technologies address all experimental

* Corresponding author; e-mail meyers@dbi.udel.edu; fax 302–831–4841.

www.plantphysiol.org/cgi/doi/10.1104/pp.104.040840.

needs, and there are advantages and disadvantages to each. These differences make the technologies complementary; in addition to good experimental design and analysis, the validation of apparent quantitative differences in mRNA levels by using several of these complementary approaches is critically important.

Single Gene Measurements

Although measurements of single genes have advanced well beyond northern blots, northern blot data are still considered to be the gold standard. The basis for this confidence may be based more on historical reasons than on any data that indicate northerns are more reliable than other methods. In situ hybridizations can provide both a qualitative and quantitative assessment of gene expression in specific tissues. In recent years, quantitative real-time PCR (QRT-PCR) has been demonstrated to generate robust, quantitative expression data for a single gene; this method also offers rapid and reproducible results and a large dynamic range (Hayward-Lester et al., 1995; Bustin, 2002; Ginzinger, 2002; Klein, 2002). Fluorescence signals are generated by dyes that are specific to double-stranded DNA (dsDNA) or by sequence-specific fluorescently-labeled oligonucleotide primers. The signal is proportional to the amount of PCR product, and special PCR machines are designed to monitor the process of amplification in real time. The amplification curve is used to quantify the initial concentration of a specific transcript in a template mixture. One of the major advantages of QRT-PCR is a broad dynamic range that can precisely quantify transcript concentrations over more than eight orders of magnitude (Heid et al., 1996). QRT-PCR can be performed using a dye like SYBR Green and unlabeled primers, with one amplification target per tube and control reactions performed in parallel. Alternatively, a pair of gene-specific primers is synthesized, one of which is fluorescently labeled; several pairs of control primers are added to the sample and each primer pair labeled with a different fluorochrome to allow specific detection. The former method using SYBR Green is less expensive than the latter; in both cases, reactions are replicated and the results are averaged.

One of the more intriguing new methods for the measurement of single genes uses so-called polonies, which stands for polymerase-colonies (Mitra and Church, 1999; Mitra et al., 2003). While still in its infancy, this intriguing technology is based on the in situ amplification of DNA or cDNA in a thin-layer acrylamide gel on a microscope slide. Because the PCR products are essentially immobilized, the result of the amplification is large numbers of polonies distributed across the slide that are spherical colonies of DNA. Each polony is derived from a single template molecule, and specific genes or transcripts can be detected by hybridizing labeled probes, similar to a classic colony lift blot. By counting the proportion of polonies derived from a specific transcript compared to the total (detected by a nonspecific stain, for example), a quantitative estimate of gene expression can be obtained (Mitra and Church, 1999; Mikkilineni et al., 2004). Modifications of this technology may go beyond expression analysis to monitor RNA splicing (Zhu et al., 2003) or other applications.

The analysis of expression of single genes or small sets of genes will further advance with the increased availability of well-curated expression data in public repositories. Using these preexisting data sets, it may be possible to measure gene expression using only a computer and internet access. Such analyses constitute electronic or virtual northern blots. Several groups, including our own, have made plant gene expression data accessible from easy-to-use Web interfaces (see http://mpss.udel.edu or the gene expression section of http://www.arabidopsis.org). A more limited set of plant data are available as part of the Gene Expression Omnibus section of GenBank (http://www.ncbi.nlm.nih.gov/geo/); their SAGEmap Web page performs differential expression analyses and provides a limited ability to measure single genes (Lash et al., 2000). However, this site is primarily a repository for published SAGE data (described below), and by design it is not optimized for any particular organism. These resources provide starting points for researchers interested in specific genes or gene families.

DNA Microarrays

The DNA microarray has produced a revolution in expression analysis. These chips simultaneously determine expression levels for thousands of genes. Data are then analyzed for patterns of expression that change over various treatments or time points. Microarrays may be comprised of short oligonucleotides or complete cDNA clones and provide a rapid and relatively inexpensive way to monitor in parallel the expression of thousands of transcripts. Because microarrays have now been used in hundreds of publications and the technology has been discussed in scores of review articles, the reader is directed elsewhere for in-depth discussions and technical details.

Early microarrays were built of cDNA fragments robotically gridded and immobilized on microscope slides (Schena et al., 1995), much as if the probes for a northern blot were laid down in a dense pattern. This approach, though still widely used, requires the maintenance and handling of microtiter dishes, validation of clones, and large scale PCR reactions. A competing approach that has become the dominant system is based on short DNA oligonucleotides that serve as probes. There are several reasons for the dominance of these oligo arrays; one reason is that oligos can be synthesized either in plates or directly on solid surfaces (in situ synthesis), making it easier to obtain reliable amounts of material than for cDNA clones. In addition, even for a well-characterized plant like

Arabidopsis, cDNA clones may represent less than 60% of the predicted genes (Wortman et al., 2003). Oligo-based approaches can effectively target selected regions starting from only the DNA sequence, such as anonymous open reading frames found in genomic sequence. With any of these microarray technologies, one of the most serious problems is ensuring that cDNA or oligonucleotide sequences are correctly assigned to their source. This is a particular problem if any sort of spotting or gridding is used to build the microarray, because a small proportion of microtiter dishes and tubes inevitably are mishandled. A different concern for commercially manufactured arrays can be validating the identity or genomic location of a specific probe, as these probe sequences are often not available. The assumption that microarrays are manufactured without errors can lead to misinterpretations or delays in understanding data that result from poor sample tracking, informatics errors, or contamination.

For plant research, the tractability and genomic resources of Arabidopsis have made it an attractive system in which to develop or commercialize microarrays. Because development costs were high in the early days of microarrays, and because resources for plant research are limited, several academic groups formed a consortium (the Arabidopsis Functional Genomics Consortium, or AFGC) to produce and make publicly available the first Arabidopsis arrays (Wisman and Ohlrogge, 2000). While these arrays were used by many academic laboratories, commercial arrays such as those produced by Affymetrix (Santa Clara, CA) were quickly adopted as well. The AFGC ended on December 31, 2002 and its public microarray project was discontinued; some public groups still produce Arabidopsis arrays, representing the model that was an impetus for the development of core microarray facilities at many institutions. However, some of these core facilities are now gathering dust due to the centralization of microarray production and competition from commercial operations. In general, this has proven to be a positive step because it relieves research scientists of relatively mundane manufacturing responsibilities. For example, one of the most critical steps in array construction is quality control to ensure minimal variation from array to array. Companies or public groups focused solely on array production can afford to spend considerable effort to ensure quality control, and a competitive pressure for quality works to the benefit of the researcher. Companies were quick to recognize the commercial potential for Arabidopsis arrays and have aggressively pursued the production of Arabidopsis microarrays. The drawback of commercial production is that the high costs of overhead, labor, and development are included in the arrays, whereas these costs are often absorbed in academically-produced arrays. Another drawback to removing microarray production from the hands of researchers can be the loss of control over the content and format.

Competition is heating up among companies that can or do produce Arabidopsis microarrays. The popular Affymetrix GeneChip arrays are comprised of sets of 25-base oligonucleotides synthesized in situ via a photolithographic process (Lockhart et al., 1996); the original array design that included more than 8,000 genes was the first commercial Arabidopsis array on the market (Zhu and Wang, 2000). The most recent design that is often called the whole genome array (WGA) includes more than 24,000 genes (http://www.affymetrix.com). Rosetta Inpharmatics (Kirkland, WA) developed the process of ink-jet "printing" of 60-base probes (Hughes et al., 2001). The Arabidopsis array based on this technology is produced by Agilent Technologies (Palo Alto, CA) and includes 21,500 genes; later in 2004, this array will contain approximately 44,000 features. In addition to arrays produced by Agilent, other companies are now marketing so-called long oligo microarrays. These arrays typically are comprised of a single oligonucleotide primer of 50 to 70 nucleotides for each gene, and the oligos are synthesized in situ or synthesized using conventional methods and then spotted on the arrays (Barczak et al., 2003). Spotted oligo arrays offer several advantages, such as a low manufacturing cost and flexibility, but usually require a substantial commitment by a company to presynthesize the 20,000+ long oligos that are spotted on these arrays. However, once the oligos have been synthesized, the materials can be distributed to individual labs for use with conventional gridding robots. For example, Operon (a subsidiary of Qiagen) produces oligo sets for three plant species (http://oligos.qiagen.com/), and at least one academic group grids and distributes arrays based on these oligos (http://www.ag.arizona.edu/microarray/). Customized or whole-genome Arabidopsis arrays may potentially be made using any of the platforms based on rapid and flexible in situ synthesis. This includes platforms developed by NimbleGen Systems (Madison, WI; Nuwaysir et al., 2002) and febit ag (Mannheim, Germany; Baum et al., 2003). Nimblegen uses a flexible photolithographic process capable of synthesizing high-density arrays with oligos of 24 to 90 bases; febit produces a benchtop machine capable of producing arrays of up to 48,000 features per slide with an oligo length of approximately 30 nucleotides. Because of ongoing changes in the technologies and commercial competitors, it is impossible to provide a comprehensive list of microarray platforms. However, there are now many commercial microarray options now available to Arabidopsis researchers.

Microarrays are now becoming available for additional plant species. Rice (*Oryza sativa*) is a widely-studied organism for which the complete genome sequence is anticipated by end of 2004. As with Arabidopsis, early rice microarray experiments were based on limited sets of ESTs (Kawasaki et al., 2001). With more sequence data now available, Agilent has announced the release of a rice long-oligo microarray that includes approximately 60% of the estimated 50,000 rice transcripts (http://www.chem.agilent.com/). As with Arabidopsis, other companies are entering the business (for example, GreenGene Biotech; http://www.ggbio.com), heating up competition with a recently funded public rice array project (http://www.ricearray.org/). Despite a lack of genomic sequence data, other plant species have not been left without microarray resources. Academic collaborations have led to the development of microarrays for barley (*Hordeum vulgare*), cotton (*Gossypium hirsutum*), cabbage (*Brassica capitata*), maize (Zea mays), potato (*Solanum tuberosum*), tomato (*Lycopersicon esculentum*), and wheat (*Triticum aestivum*); commercial interest in developing arrays for these and other plant species is growing. As in the case of Arabidopsis, the release of commercial microarray products can drive some academics out of the array manufacturing business. However, because the primary motivation for some academic labs to fabricate microarrays is to generate the resources they need for experimentation, the entrance of a commercial competitor may be welcomed.

Despite the broad adoption of microarrays as a research tool, there are several technical issues with the technology, some of which are better understood than others. Most of these limitations result from the principle of hybridization that is at the core of the technology. For example, cross-hybridization, the hybridization of multiple targets to single probes, remains poorly characterized. Genome duplications impede the design of oligos that distinguish between closely related sequences (Ishii et al., 2000). In many plant species, genome duplications resulting in cross-hybridization may be a limitation for determining the expression of any single gene; in Arabidopsis, one of the most simple genomes, approximately 60% of the genome is duplicated and 17% of the genes are present in tandem arrays (Blanc et al., 2000; Grant et al., 2000; Vision et al., 2000; Simillion et al., 2002). The general migration from cDNA to oligo arrays means that probes can be selected based on regions of dissimilarity among generally similar genes, improving specificity (Talla et al., 2003). Hybridization and washing conditions are a critical issue for any array platform; these conditions are influenced by variations in temperature, ionic strength, or pH. The limit of detection for Affymetrix chips is approximately 1/100,000 transcripts (http://www.affymetrix.com); changes in genes expressed near this level are difficult to detect with statistical significance (Ishii et al., 2000). Background signal intensities at this level are similar to signals of many weakly expressed transcripts (Duggan et al., 1999). Spotted microarrays built from presynthesized components have several potential sources of variation that differ from those of arrays manufactured by in situ synthesis. Spotted microarrays are subject to variation in the pin geometry, variations in spot geometry, and differences in the amount of material deposited onto and subsequently bound to the slide surface. The method of preparation of the RNA and labeled cDNA targets used in any microarray experiment can also introduce variation, as there are many methods for the processing, isolation, and labeling of RNA samples, and factors such as the degradation rate of transcripts may also affect the final data (Auer et al., 2003). Sequence-specific differences in the efficiency of dye incorporation may also produce variation for biologically-irrelevant reasons. In the use of microarrays, the source of variation, whether technical or biological, should be identified and quantitatively estimated by replicating experiments at two levels—technical replications that are separate preparations and arrays run for the same RNA sample, and biological replications that are RNA samples extracted from separate but identically treated biological materials (Lee et al., 2000). It is important to note that technical variation appears much lower for in situ synthesized and spotted oligo arrays than for those produced from PCR amplicons, and this consistency decreases the relative importance of technical replicates to the point at which these may be eliminated while retaining biological replications (Zhu and Wang, 2000).

An involvement of statistics is inevitable given the large numbers of simultaneous measurements that can be made using microarrays, and these large numbers raise problems that are not normally encountered in molecular biology. For example, an alpha value of 0.05 would be viewed as highly satisfactory for most biological measurements, where the α value is the accepted probability of detecting a false positive for a single event (a Type I error). However, when making independent measurements of 26,000 genes (events) on a typical Arabidopsis whole-genome microarray, this α value allows 1,300 false positives for the experiment. Since downstream procedures, which are more labor intensive, less high throughput, and more expensive per unit of information, cannot reasonably accommodate this proportion of false leads, the importance of achieving more restrictive α values is readily apparent. This is possible through replication of the microarray experiments and requires greater numbers of microarrays as well as an appropriate statistical design. A particularly accessible review of this area has been provided by Draghici (2002). Among statistical treatments, the application of mixed model ANOVA methods to microarray data has considerable promise for both spotted and in situ synthesized microarrays (Kerr et al., 2000; Wolfinger et al., 2001). General agreement has not yet been reached on the optimal statistical treatment for the sets of 10 or more probes designed for each gene represented on the Affymetrix microarrays (probe level expression data). There are advantages to using existing statistical methodologies instead of the standard Affymetrix software; better accuracy and sensitivity are provided by the use of various types of models or probe level data (Li and Wong, 2001a, 2001b), ANOVA analyses (Chu et al., 2002), or analyses of inherent noise (Naef et al., 2002; Draghici et al., 2003). Identification of the sources of variance in expression data is essential to enable the detection of small but biologically relevant differences in transcriptional profiles (Jin et al., 2001). It has been clearly demonstrated that the failure to apply appropriate statistical analyses to microarray data can result in misleading conclusions (Hsieh et al., 2003).

Tag-Based Methods

Exhaustive sequencing of ESTs is a common method for gene expression profiling, although the primary purpose of EST sequencing is usually to generate genic sequence data. EST data are generated by large-scale, single-pass, partial sequencing of cDNA clones (approximately 500 bp), usually from a large number of libraries representing diverse tissues (Adams et al., 1995). Comparisons of EST frequencies in different libraries can expose differential gene expression on a broad basis (Okubo et al., 1992, 1995; Matsubara and Okubo, 1993; Ewing et al., 1999). In theory, the abundance of an EST is an exact digital representation of the number of copies of a transcript in the tissue. Large numbers of ESTs derived from diverse tissues produce quantitative estimates of gene expression, but ESTs are relatively slow and costly to generate, making it difficult to achieve saturation of a library. Theoretically, expression profiles could be derived for very weakly expressed genes if ESTs were sequenced in sufficient number. This has been performed with human EST libraries that contain tens of thousands of sequences (Okubo et al., 1992, 1995; Matsubara and Okubo, 1993; Adams et al., 1995; Okubo et al., 1995; Kawamoto et al., 2000). In plants, Ewing et al. (1999) compared and analyzed 10 rice libraries containing between 1,000 and 5,000 ESTs and were able to identify statistically significant patterns of gene expression among several rice tissues. However, public plant EST libraries are in general too small or from too many sources for accurate quantitative expression analyses, although private companies have amassed databases of more than a million plant ESTs (Mazur et al., 1999). For Arabidopsis, there are currently 196,988 ESTs or cDNAs in GenBank (as of January, 2004; http://www.ncbi.nlm.nih.gov/dbEST), but because most of these were generated either from a single library of mixed tissues or were selected from normalized libraries (Newman et al., 1994; Delseny et al., 1997), the Arabidopsis EST abundance does not accurately reflect expression levels. In general, the low total number of EST sequences for a given organism confounds accurate estimates of gene expression levels.

SAGE, like EST sequencing, is a quantitative or digital method of gene expression analysis. Unlike EST sequencing, SAGE extracts only a 10- to 14-base tag from a unique position within each species of mRNA (Velculescu et al., 1995; Zhang et al., 1997). These short SAGE tags are derived from a position directly 3′-adjacent to the 3′-most recognition site for a particular restriction enzyme, such as *Nla*III. The tag sequence and position are important for the identification of the gene from which the tag was derived. Whereas ESTs each require a single sequencing read, SAGE tags are released from cDNAs by restriction enzymes, ligated together, amplified by PCR, and sequenced as concatamers. This results in a higher throughput and lower cost for SAGE than ESTs. A number of modifications to the original protocol have been reported. Modifications that increase the length of the tag include the LongSAGE method (Saha et al., 2002) that produces 21- or 22-base tags, and the SuperSAGE method that produces 26-base tags (Matsumura et al., 2003); a recent report describes modifications that dramatically improve the efficiency of LongSAGE library construction (Gowda et al., 2004). The primary limitation of SAGE or its variants is the cost of sequencing reactions; even at $1 per read, SAGE tags cost roughly $0.04 each and a library of 100,000 tags would cost $4,000. Sampling error has also been a source of bias in SAGE (Stollberg et al., 2000), although increasing the number of available tags addresses this problem.

A recent advance in tag-based gene expression analysis is MPSS, developed and commercialized by Lynx Therapeutics (Hayward, CA). MPSS is based on methods to clone individual cDNA molecules on microbeads and sequence, in parallel, short tags or signatures from these cDNAs (Brenner et al., 2000a, 2000b). A complex mix of cDNAs, such as those derived from a particular plant tissue, is cloned onto microbeads, with the representation of molecules on the beads identical to that in the original sample (e.g. one cDNA per bead). Using an unconventional but ingenious method of sequencing, large numbers of beads are sequenced in parallel. A series of digestion, ligation, and hybridization reactions are performed in consecutive steps while the beads are immobilized in a flow-cell underneath a high-power microscope so that the reagents flow over and around the beads, and there are no gels or capillaries (Brenner et al., 2000a). The final output of MPSS is a set of abundances for thousands of distinct 17- or 20-base signatures, most of which uniquely identify a particular transcript. The parallel sequencing method produces millions of MPSS signatures in only a few weeks; however, the technology is sufficiently complex that unlike SAGE, it cannot be performed in individual laboratories. On a per-tag basis, MPSS is currently less than half the cost of SAGE.

The sequence-based expression data from ESTs, SAGE, or MPSS experiments have many uses. The availability of complete genome sequences permits the direct comparison of tags to genomic sequence and further extends the utility of the data (Meyers et al., 2004b). The identification of transcribed regions is performed by aligning the signatures to genomic sequence. The expression levels of nearly all polyadenylated transcripts can be quantitatively determined, and the abundance of a given tag for a specific library is representative of the expression level of the corresponding gene. The approximate location of the polyadenylation site for each transcript is known because both SAGE and MPSS tags are derived from defined restriction sites in the 3′ end of a transcript. Several distinct SAGE or MPSS tags matching different sites within a single gene indicate alternative polyadenylation or 3′ splicing. Expressed tags that uniquely match to unannotated regions of the genome provide experimental evidence for novel transcripts (Meyers et al., 2004c). Quantitative methods for the analysis of tag frequencies and detection of differences among libraries have been published (Audic and Claverie, 1997; Greller and Tobin, 1999; Lash et al., 2000; Stekel et al., 2000).

Genome duplications complicate the unique assignment of short tags to specific genes, particularly when members of a gene family have a high degree of similarity. Issues of genome duplications are likely to be particularly relevant to many plant species that have polyploid origins and show evidence of large-scale segmental duplications. The short length of SAGE tags (usually 14 bases) complicates the assignment of tags to distinct genes in even minimally complex genomes (a tag-to-gene ambiguity; Lash et al., 2000; Stollberg et al., 2000). Tag-to-gene ambiguities may be avoided by using longer tag sequences, such as 20-base MPSS signatures, 21-base LongSAGE tags (Saha et al., 2002), or the 26-base SuperSAGE tags (Matsumura et al., 2003). An analysis of potential MPSS signatures in the Arabidopsis genome demonstrates that 18.1% of 17-base tags and 12.5% of 20-base tags are duplicated (Meyers et al., 2004b). Analyses using the Arabidopsis genome indicate that there is a diminishing return for tag lengths beyond 20 bases, such that it may be more economical to sacrifice some specificity to obtain a greater number of tags of approximately 20 bases and sort out differential expression among nearly identical gene family members using different techniques (C.D. Haudenschild and B.C. Meyers, unpublished data). A gene may also have more than one unique tag as a result of alternative termination of some transcripts, creating a gene-to-tag ambiguity (Lash et al., 2000; Meyers et al., 2004b).

Methods like SAGE have not been applied extensively to plant species, but more and more examples can be found in the literature (Matsumura et al., 1999; Chakravarthy et al., 2003; Jung et al., 2003; Lee and Lee, 2003; Fizames et al., 2004). Early applications in nonplant species used SAGE to characterize transcriptomes (Velculescu et al., 1995, 1997), to study the differences between them (Zhang et al., 1997), to annotate genomic sequences (Saha et al., 2002), and for whole-genome studies of transcriptional activity (Caron et al., 2001). In our laboratory, we have been using MPSS to analyze gene expression in Arabidopsis, and we have developed a Web site for public access to these data (Meyers et al., 2004a, 2004b). For reasons that are not entirely clear, MPSS has been more rapidly adopted in the plant community than in animal species, although there are only a few published studies outside of our own laboratory (e.g. Hoth et al., 2002, 2003; Christensen et al., 2003). One limitation for all of the tag-based methods compared to microarrays is that the cost of biological or technical replications is prohibitive, so estimates of variance for the tag-based methods are incomplete or poorly characterized.

THE DANGERS OF PROLIFERATING TECHNOLOGIES

There are both advantages and disadvantages to the growing number of competing technologies and technology platforms for the measurement of gene expression. Some comparisons are not entirely fair; for example, the two broad categories that we describe above, tag-based systems and microarrays, have different and complementary uses (see below), so these are not directly competing technologies. Competition among microarray platforms has led to lower costs, improved quality control, and increased numbers of genes per array, at least in the case of Arabidopsis. The disadvantage of having a proliferation of array platforms is that it can create orphan data. In other words, experiments performed with an older generation or different type of a microarray may be difficult to compare to data derived from the latest microarray format. This may necessitate the repetition of experiments to directly confirm other laboratory's findings.

The prospect of comparing data across experiments raises the question of whether the measurements from gene expression technologies are directly comparable and how good the correlations are. While no definitive answer yet exists, several groups have or currently are addressing this question. In a comparison of SAGE with the Affymetrix oligonucleotide microarrays, the two approaches correlate for genes expressed at high levels, and SAGE is more accurate than for genes expressed at low levels (Ishii et al., 2000). We are currently conducting comparisons of MPSS and microarray analyses. Among microarray platforms, several comparisons have been published. Tan et al. (2003) compared gene expression measurements generated from identical human RNA samples using the Affymetrix (25-mer), Agilent (60-mer), and Amersham (30-mer; Piscataway, NJ) microarray platforms. A total of five arrays were used for each time point in their analysis, including technical and biological replicates. Their results demonstrated considerable variation for comparisons of significant gene expression changes, and correlations in gene expression levels across the different platforms were modest (Pearson's correlation coefficient average of 0.53, range of 0.48–0.60). In addition, although many of the genes present on each microarray platforms were the same, the differentially expressed genes identified by each technology were not substantially overlapping. Other studies have compared spotted cDNA microarrays with Affymetrix GeneChip arrays and found a poor correlation between these disparate array types (Kuo et al., 2002; Yuen et al., 2002), although the level of experimental replication in these studies was not clear. Poor statistical designs or a lack of replications could also generate low correlations. In general, published cross-platform analyses suggest that the conclusions derived from a microarray analysis may be largely dependent upon the type of platform used in the experiment. This is not encouraging news, and suggests that a great deal remains to be learned about factors intrinsic to different microarray platforms that can affect the data.

Incongruous data or conclusions from gene expression measurements performed using different technology platforms may result from several sources of variation. A very simple example is that the set of genes represented in the arrays may not be identical; the Agilent, Affymetrix, and Qiagen/Operon probe sets for Arabidopsis microarrays each represent 21,500 to 24,197 genes, but only 17,149 genes are shared among the three platforms. However, there are additional issues in such a comparison, because oligo lengths, positions, and numbers per gene vary among manufacturers. It is possible that some genes are better measured by the probes on different microarray platforms, and no single type of array accurately measures every gene. It may take many years of empirical studies before we achieve optimal designs and understand the impact of the sequence and position of the oligo on the signal strength. The process of correlating design features with expression data would be facilitated if all manufacturers released the sequence of the probes on their arrays. Probe sequences are considered proprietary information by some companies because of a fear that competitors will market arrays based on identical probes or use the information to decipher design algorithms. With some exceptions, complete sets of probe sequences can be hard to obtain except via nondisclosure agreements with manufacturers. In fact, oligo design software is still rapidly developing (e.g. Mei et al., 2003; Nielsen et al., 2003; Talla et al., 2003), so it is highly unlikely that any existing microarrays contain a complete set of optimally-designed probes. It may also be desirable (although perhaps not plausible) for array manufacturers to agree on a set of standard template sequences; if different splice variants or models of a gene are used for probe design, it is possible that probes with the same gene identifier may be measuring different transcripts. Standardization of experimental design and methods would also facilitate comparisons of array data produced by different labs. One of the first steps in this direction was the development of a standard set of technical details that should be reported for every microarray experiment. The minimal information about a microarray experiment (MIAME) protocol requires the reporting of enough details to ensure that the results of a microarray experiment could be interpreted or repeated independently (Brazma et al., 2001). These basic data should be sufficient to store the data in public repositories such as GenBank and enable the use of standardized data analysis tools.

In the coming years and as sequence databases are populated with ESTs and genomic data for diverse plant species, the research community working in each of these organisms may face the question of which gene expression platform to choose. This may be an issue if it comes down to a choice among commercial platforms, because several of the major microarray production companies charge significant set-up fees (although for a large-enough market, these fees may be waived and absorbed into the sales of the arrays). The barley GeneChip microarray is an example of an organized and united approach taken by a consortium of plant researchers to build resources for expression profiling in a crop species that had not attracted the interest of commercial microarray manufacturers (Close et al., 2004). An international group of laboratories focused and coordinated their efforts to develop a single microarray platform for transcriptional profiling. A public data storage site, BarleyBase (http://barleybase.org/), was constructed as part of this project to integrate expression profiling data from all researchers using the platform. This creates a synergistic effect because all array data generated for barley will be directly and easily comparable. BarleyBase is also incorporating controlled vocabularies to facilitate cross-species comparisons (Close et al., 2004). The coordinated development of the barley microarray may represent a paradigm for other plant species in which too many technology platforms could diminish the utility of individual data sets and fragment the research community.

OPEN VERSUS CLOSED TECHNOLOGIES AND THE IDENTIFICATION OF NOVEL TRANSCRIPTS

Technologies such as ESTs, SAGE, or MPSS require no prior knowledge of the sequences of the transcripts and can discover previously unknown transcripts. This feature defines an open architecture for these expression technologies. In contrast, closed architectures, like most microarrays, are based on existing knowledge of genes, with probe sets designed to match known or predicted transcripts. The data derived from the open technologies can be used to annotate genomic sequence, whereas data from closed technologies is often cheaper to obtain and can more easily be used for focused experiments. However, one of the more interesting applications of the microarray is the development of a hybrid approach. In different organisms, several groups have constructed true WGAs containing tiled probe sets that include nearly every nucleotide in the genome (Kapranov et al., 2002; Yamada et al., 2003). WGAs have extended the potential of microarrays by creating an open system on a platform generally characterized as closed. Such arrays have recently been applied to Arabidopsis and led to the identification of transcription from unannotated regions of the genome (Yamada et al., 2003). In addition, these tiled arrays uniquely offer the ability to characterize, at the whole-genome level, transcriptional variants that differ in the use of splice sites and exons and to describe previously uncharacterized 5′ or 3′ untranslated regions.

In fact, transcriptional data from open technologies suggest that automated annotations of genomic sequence fail to identify many transcripts. Through the application of WGAs, MPSS, and targeted RACE experiments, the Arabidopsis genome is still yielding previously unknown transcripts, although the genome was mostly completed and first annotated more than 3 years ago (Arabidopsis Genome Initiative, 2000; Xiao et al., 2002; Yamada et al., 2003; Meyers et al., 2004b). The WGA and MPSS data of Yamada et al. (2003) and Meyers et al. (2004c) suggest that a comprehensive annotation of transcripts encoded in a genome requires significant experimental data beyond the complete sequencing of chromosomal DNA. Many of these RNA molecules may not encode proteins, but could have independent functions as regulatory molecules. Transcripts that do not encode proteins but can function directly as RNA molecules are called noncoding RNAs (ncRNAs; Eddy, 2001). With the exception of housekeeping RNAs, like tRNAs or small nucleolar RNAs, relatively few potential regulatory ncRNAs have been characterized from plants; those that have been identified appear to be plant-specific (MacIntosh et al., 2001). Nearly all of the 29,000+ predicted genes in Arabidopsis encode proteins; very few ncRNAs are annotated (MacIntosh et al., 2001; Wortman et al., 2003). Natural anti-sense transcripts (NATs) overlap with transcribed coding regions and may be involved in the regulation of gene expression (Vanhee-Brossollet and Vaquero, 1998). These NATs and other ncRNAs are a major component of the diversity of transcripts produced in higher eukaryotes (Eddy, 2001; Numata et al., 2003; Yelin et al., 2003). Some of the first experiments using SAGE and MPSS in plant genomes have identified a number of anti-sense transcripts (Gibbings et al., 2003; Meyers et al., 2004b). Therefore, the comprehensive use of open transcriptional profiling approaches will add significant new information to any sequenced genome by identification of ncRNAs, NATs, or other transcripts that are poorly predicted. Because the transcriptional complexity of sequenced genomes has yet to be fully explored, microarray designs should be flexible and facilitate the addition of newly discovered transcripts.

There are additional transcripts missing from or insufficiently measured by current technology platforms. Methods also need to be developed for high-throughput quantification of splice variants. Simultaneous quantification of all splice variants of a single gene is presently done on a gene-by-gene basis using QRT-PCR (Renner and Pilger, 1999; Goel et al., 2001). Large numbers of variants of known transcripts have been found in Arabidopsis, generated by alternative splicing or polyadenylation (Haas et al., 2003; Meyers et al., 2004b). These variants may have novel functions. Additionally, there are no systematic processes for identification and quantification of microRNAs (miRNAs), which have important biological roles in plants and animals (Carrington and Ambros, 2003). These small RNA molecules (approximately 21 nucleotides) play regulatory roles in plant development and are processed from longer noncoding transcripts (Aukerman and Sakai, 2003; Palatnik et al., 2003). However, it is not yet clear that all possible miRNAs have been characterized from Arabidopsis. A technology to measure these on a global scale would contribute greatly to our understanding and open the door to novel experiments.

Future genomics projects will take advantage of the advances in techniques and technologies to deliver genomes at a fraction of previous costs. We anticipate that high-throughput open technologies, such as MPSS, will be important because the data can be used to annotate genomic sequence. Ultimately, it may be possible to estimate the extent of gaps in the genomic sequence based on the percentage of unmatched MPSS signatures. Statistical approaches to estimating the complete size and complexity of the human transcriptome based on limited SAGE data were unsuccessful (Stern et al., 2003), but it may be possible to estimate the complexity of the Arabidopsis transcriptomes using more extensive sets of MPSS data.

TISSUE ISSUES: MEASUREMENTS OF GENE EXPRESSION IN SPECIFIC CELL TYPES

Multicellular eukaryotic organisms comprise complex interspersions of different cell types. Higher plants are no exception, and it is increasingly apparent that methods are required to isolate specific cell types when considering gene expression in the whole organ. Typical experiments may utilize intact leaves, flowers, or other organs that comprise multiple cell types and utilize RNA that is isolated essentially from a population or mixture of cells. For certain studies, this homogenization of a heterogeneous starting material may dilute, alter, or mask the true biological state of individual cells. The averaging of a response across millions of cells may produce a signal that is artificial and accurately reflects none of the varied transcriptional states found in individual cells. Signals that emanate from a single plant cell (perhaps one under attack from a pathogen) may be found in a gradient that decreases with distance from the source, such that the timing and magnitude of the transcriptional response varies dramatically in cells that are further from the source. However, until technologies are better able to precisely measure the state of single cells, this will remain speculation.

Several methods are being employed to allow subsets of cells to be isolated and analyzed for gene expression with the techniques described above. These methods are described in more detail below, but one limitation that still exists is the large amount of RNA required for an experiment. Standard microarray experiments utilize fluorescent dyes that necessitate microgram quantities; SAGE and MPSS library construction requires similar quantities of starting material. The use of radioactively-labeled targets requires only nanogram quantities for accurate detection and measurement, but methods employing radiation, such as macroarrays (the larger cousin of microarrays with probes gridded on nylon membranes), have been predominantly supplanted due to relatively low throughput. Amplification of small quantities of RNA may provide a way around this requirement. Methods and products for RNA amplification are available, but amplification could bias the representation in the sample due to variation in the length or sequence of the transcripts. Amplification methods are complicated slightly for oligo-based microarrays; the immobilized probe on these arrays consists of a single strand of DNA, and to ensure strand specificity for the RNA target, amplification methods must ensure production of the complementary target. We have developed accurate methods based on in vitro transcription for the linear amplification of plant total RNA that start from as little as 50 ng of material; we have also developed methods for exponential amplification of picogram quantities of RNA (F.-C. Gong and D. W. Galbraith, unpublished data).

Isolation of Cell-Specific RNA and Other Macromolecules by Laser-Capture Microdissection

Several methods have been developed for the isolation of macromolecules such as DNA, RNA, and protein from selected cells. Some schemes rely on tissue dissociation (e.g. tissue digestion and cell sorting) and thus rely on the prior identification of cell-specific markers (see below). Other techniques, such as direct micropipetting of cell contents, are highly labor-intensive or have limited access to internal tissues (Karrer et al., 1995; Brandt et al., 2002). In contrast, laser-capture microdissection (LCM) provides a rapid means of isolating pure cellular preparations directly from heterogeneous tissues, based on conventional histological identification (Emmert-Buck et al., 1996). Specific markers can assist with the identification of the desired cells, including prestaining with β-glucuronidase reporters (N. Gandotra and T. Nelson, unpublished data) but this is not a requirement. The LCM system can also incorporate immunological identification of specific cells to assist the laser-harvest step. Two studies to date have reported the use of laser microdissected cells from plant tissues as the source of RNA for profiling on microarrays (Asano et al., 2002; Nakazono et al., 2003).

In the LCM version developed at the National Institutes of Health (Emmert-Buck et al., 1996) and commercially available as the Arcturus Pix-Cell system (http://www.arctur.com), a HeNe laser beam is used to fasten selected cells to a thermoplastic film suspended above a tissue slice while it is viewed on an inverted microscope. Cells harvested onto the film can be subjected to high efficiency procedures for the isolation and analysis of DNA, RNA, and protein. The advantage of this version of LCM method is that the low-power infrared laser dimples the adhesive film onto individual cells (for review, see Roberts, 2002); the cells are not struck by the laser beam. Images are obtained of samples before and after cell harvest, as well as of the harvested cells. The harvest of hundreds or thousands of individual cells is feasible, using either a manual aim-and-fire method or a fully automated method in which the desired cells are marked on a screen for robotic harvest from the slide. A variety of proof-of-concept and analytical studies have demonstrated that the DNA, RNA, and protein obtained from LCM-harvested cells can be suitable for microarray-based RNA expression profiling, proteomic protein profiling and genomic mutational analysis (Banks et al., 1999; Jin et al., 1999; Luo et al., 1999; Simone et al., 2000; Wong et al., 2000; Craven et al., 2002; Ohyama et al., 2002; Nakazono et al., 2003).

Kerk et al. (2003) optimized LCM for use with tissues from a variety of plants, including rice, maize, Arabidopsis, radish (*Raphanus sativus*), and other species. Their approach used conventional histological methods, including paraffin-embedding; this method provides high-resolution access to cells of all ages and types, and is stable enough to permit archiving and resampling of the tissue. RNA can be isolated from paraffin-archived materials for at least several months without degradation in quality. In addition, samples can be taken from multiple sections onto the same collecting film to pool cells that are rare, such as single cells from a particular location. Using the paraffin methodology, recoveries of 10 ng of RNA/50 LCM-harvested cells are possible, sufficient for a strong signal by single-round RT-PCR from a moderately expressed gene or to serve as a template for linear amplification into probes for microarrays (N. Gandotra, T. Ceserani, S.L. Tausta, and T. Nelson, unpublished data).

Flow Sorting of Cell-Type Specific Nuclei or Protoplasts

Specific cell types can be labeled with fluorescent proteins and protoplasts prepared and purified using flow cytometry and cell sorting. The sorted protoplasts can then be subjected to gene expression analyses. The green fluorescent protein (GFP) of *Aequorea victoria* is the prototypic label; specific cell types can be tagged by driving expression of such proteins with highly specific promoters. This approach was used by Birnbaum et al. (2003) to create a gene expression map of the developing Arabidopsis root. Groups of genes with coordinated expression, as determined using Affymetrix GeneChips, defined local expression domains. Statistically significant overrepresentation of genes of known functions within the local expression domains provided testable hypotheses about root development. These hypotheses concerned the influences and involvement of signal transduction, hormone responses, gene organization, and other regulatory mechanisms. The map also provides a useful resource for the design of further experimental and computational strategies to explore gene regulation in roots. One caveat is the possible influence of the process of protoplast production on gene expression patterns. For Arabidopsis roots, this influence appears minor (Birnbaum et al., 2003), although subtle changes in genes expressed at low levels may not have been detected by the expression platform. For organ systems, the question also exists as to whether protoplasts can be successfully isolated from all cell types that are present within that organ.

The approach of GFP-based cell type-specific labeling can also be applied to subcellular organelles such as nuclei (Galbraith, 2003). Flow sorting of GFP-tagged nuclei from homogenates of transgenic plants allows rapid purification of sources of primary transcripts. Given that polyadenylation is essentially cotranscriptional (Orphanides and Reinberg, 2002), this approach should provide information about transcriptional regulation that is unaffected by the types of perturbation of gene expression associated with protoplast production. A further advantage is that plant homogenization can be adapted more readily for high throughput handling than can protoplast production.

APPLICATIONS OF TRANSCRIPTIONAL PROFILING: AN EXPANDING RANGE OF POSSIBILITIES

Dissection of Changes in Gene Expression Levels

One of the temptations of whole-genome expression platforms is to simply generate data for discovery purposes. While this may be a valid approach for open technologies in which the data can be used for genome annotation, it is harder to justify for microarrays and other closed technology platforms. Despite the ease of producing reams of data, it will be meaningless unless experiments are properly designed with the appropriate biological materials and replicates. The extraction of meaningful data requires analytical strategies and the interpretation depends on close interactions among biologists, computer scientists, and statisticians.

The detection of differential expression among two types of tissues differing by some experimental variable is one of the most basic questions addressed with transcriptional analysis. Typically, a user-defined cutoff or threshold for the ratio of expression levels in the two tissues is used to identify differentially expressed genes. The underlying assumption is that genes with differential expression are somehow involved in the condition that distinguished the tissues. The statistical methods for identifying such genes have been much better developed in recent years (for review, see Slonim, 2002), and are able now to identify up- or down-regulated genes with statistical significance. The end product is a list of candidate genes believed to be involved in the phenotype of interest; these genes must then be validated using much more time-consuming functional studies. The integration of pathway information could lead to the association of pathways with a process when genes in that pathway are overrepresented in the differentially expressed genes. Although for most organisms few data are available describing the pathways and related genes, such data may be generated empirically by the application of pattern discovery methods. These methods include the numerous clustering techniques designed to construct groups of genes with related patterns within the dataset. This simplifies and structures the data based on inherent patterns rather than imposing assumptions made a priori. Ultimately, it may be possible to reconstruct or model complex signaling pathways by combining interferences made from transcriptional profiling data with biochemical and metabolic data.

Categorization of Tissues Based on Expression Patterns

Expression profiling provides a comprehensive approach for the molecular characterization of tissues, treatments, or cell types. The state of the transcriptome represents a phenotype that provides a clear physiological picture of cellular activity (Hughes et al., 2000). Class prediction methods are statistical techniques that can be used to classify expression profiles from different samples into known groups (for review, see Slonim, 2002). The use of microarray phenotypes for tissue classification is most widely and successfully used in cancer research; the molecular data can distinguish tumors more reliably than other approaches, resulting in more accurate disease diagnoses (Russo et al., 2003). Comparisons among different samples of the same cancer type reveal distinct subgroups, provide a molecular classification of the cancer type, and can determine the stages of progression of the disease (Russo et al., 2003). Hierarchical clustering analysis of the array data is used to sort specimens. These studies have defined candidate marker genes that can discriminate between normal and diseased tissues. The combined sets of diagnostic marker genes may be used to develop specialized or customized arrays that contain only the diagnostic genes of specific interest. However, while the idea of customized arrays was pertinent when array densities were low and most arrays were homemade, this strategy may be less important as costs decrease for high-density commercial arrays for which uninformative genes can be ignored.

In plants, this type of classification based on transcriptional profiles could be applied to the sorting of mutants based on perturbations in distinct signaling pathways. This strategy does not require optimal microarray probe design or even that the probes identify known genes. The microarray elements must serve as molecular markers, providing detectable signals and behaving independently. Moreover, complete coverage of all genes by the technology is not critical, as long as the genes that are represented provide enough resolution for diagnosis or identification. Every informative array element or probe will provide an additional dimension for the analysis and for maximum resolution and significance; these probes should outnumber the distinct pathways or mutants under analysis.

Application of Technologies to Diverse Genotypes

Natural variation in gene expression levels between closely related plant varieties can be treated as a genetic polymorphism. Microarrays or other methods can be used to describe patterns of gene expression among individuals in a mapping population. Each pattern constitutes a molecular phenotype. Transcript abundance levels differing in the parents of a mapping population and segregating among the progeny can be mapped and characterized as quantitative traits (for review, see Cheung and Spielman, 2002). These expression profiles may be more easily interpreted or quantified than some visible phenotypes. Differences in expression of a given gene may result either from allelic differences in its promoter or from effects of distal regulatory loci. In both cases, the variation is due to genetic differences that can be subjected to genetic analysis. In parallel, the individuals in the population can be genotyped using standard molecular techniques. With molecular phenotypic and genotypic data, expression level differences can be mapped using approaches based on quantitative traits, and with these data, quantitative phenotypic measurements may be associated with genetic markers (Jansen and Nap, 2001). Accessions of Arabidopsis are rich in genetic variation for many traits (Alonso-Blanco and Koornneef, 2000), and the analysis of this natural variation using quantitative methods may provide more insight into plant signaling and gene function than classical mutagenesis studies. This is because of the complexity of variation found between ecotypes and because variation in the genetic background may increase the penetrance of certain weak alleles or promote novel phenotypes resulting from gene interactions. Another important point is that alterations in the transcriptional activity of a gene may have more significant effects than polymorphisms that alter the protein sequence. Substantial variation in gene expression has been demonstrated between primate species and among fish populations (Enard et al., 2002; Oleksiak et al., 2002), suggesting that natural selection may act as, or more, effectively on transcriptional than translational differences. In plants, most such studies will first be carried out in Arabidopsis due to the experimental advantages of this model plant; there is little doubt that gene expression analysis ultimately will be used to characterize and to map complex phenotypes in many plant species.

Which technology platforms will be used for studies of natural variation in gene expression? All of the platforms described above will measure variation in expression, but some will also be sensitive to genotypic differences that could interfere with measurements of expression. For example, the oligos used in some microarray platforms are short enough to be sensitive to sequence polymorphisms within the homologous region of the transcript. The short probes (25-base oligos) used on Affymetrix arrays will be most sensitive to single nucleotide polymorphisms (SNPs); one base difference in the length of the oligo is enough to substantially diminish hybridization. Because Affymetrix uses 10 or more probes for each gene, differences in hybridization intensity among the probes may be attributed to genomic polymorphisms. In fact, some research groups have exploited this property using labeled genomic DNA to identify SNPs or insertion/deletion events. An early and elegant study demonstrated polymorphic hybridization to Affymetrix microarrays due to strain-specific differences in yeast (*Saccharomyces cerevisiae*; Winzeler et al., 1998). Borevitz et al. (2003) used Affymetrix arrays to assess the polymorphisms in the Landsberg ecotype of Arabidopsis by hybridization of genomic DNA to the array designed from the Columbia genome. In contrast to the 25-mer oligos, long oligos (70-mers) are more tolerant to polymorphisms, presumably because the additional nucleotides provide greater stability. This has been demonstrated in experiments using RNA from *Arabidopsis thaliana*, *Arabidopsis arenosa*, and *Brassica oleracea* (Lee et al., 2004). Whole-genome long-oligo arrays could be used to analyze gene expression in a wide variety of related species with smaller genotypic effects on hybridization. This reduced sensitivity to SNPs means that long-oligo microarrays will not be useful for distinguishing expression levels of alleles or closely related gene families.

Measurement of Allele-Specific Differences

Beyond simply measuring expression level differences among homozygous inbred lines, an additional challenge for gene expression technologies will be to characterize and quantify subtle allele-specific differences in expression at heterozygous loci. Hybrid vigor is a well-characterized but poorly understood trait that is important to modern agriculture; one possible explanation for hybrid vigor is transgressive variation in expression. Expression differences for a particular allele in a hybrid compared with the parental lines result either from imprinting (Oakey and Beechey, 2002; a cis effect) or trans-acting regulatory elements encoded in the two genomes. Imprinting is generally associated with monoallelic expression (Oakey and Beechey, 2002), so biallelic nonparental expression is indicative of trans-acting regulation of expression. To put it differently, the promoter and other adjacent regulatory elements for a given allele are identical in the F_1 hybrid and parental lines, so any differences in expression for a specific allele between an inbred parent and the F_1 hybrid must result from the interchromosomal effects in the hybrid. Similar intergenome effects may alter gene expression patterns in polyploids (Osborn et al., 2003). Draft sequences of rice indica and japonica varieties have been published (Goff et al., 2002; Yu et al., 2002), and these data create a unique opportunity for large-scale measurements of differential expression in closely related varieties and hybrids, because the sequence of alleles from each variety will be known and may be used for measurements of allele-specific expression levels.

Sequence based-measurements of gene expression such as LongSAGE or MPSS are sensitive to single nucleotide polymorphisms and therefore could be used to globally quantify allele-specific expression. However, the sequence of both alleles must be known to ensure a specific match for the tag. For microarrays, a priori knowledge of SNP locations enables the use of short oligonucleotides, such as those present on the Affymetrix arrays, to measure differential expression between alleles. This type of analysis was performed using human genes and demonstrated that a significant proportion of the alleles that were examined were differentially expressed (Lo et al., 2003). Differential display-type methods, which distinguish genes based on restriction site polymorphisms, can be used to screen for allele-specific expression (Hagiwara et al., 1997); differential-display approaches are advantageous when the sequence of one or both alleles is unknown. This approach has been used to identify allele-specific differences in expression for small numbers of maize genes (Guo et al., 2003). Whole-genome analyses of allele-specific expression in plants will require gene sequences from multiple varieties and may require specialized microarrays that detect SNPs to distinguish alleles.

FUTURE DIRECTIONS

Eventually it may be possible to perform global expression profiling experiments on single plant cells. Attempts have been made to do this for human cancer cells (Klein et al., 2002). LCM can isolate RNA from a single cell, which can then be amplified by a linear method into sufficient probe for an array experiment. Few technologies exist to precisely measure single cell transcription without amplification. One recent report used oligomer DNA probes tagged with fluorophores to detect RNAs by fluorescence in situ hybridization (Levsky et al., 2002). However, this analysis was limited to 11 genes. These experiments suggested that gene expression is stochastic and that a single sampled cell may have properties highly divergent from the average (Levsky et al., 2002; Levsky and Singer, 2003). Studies of bacterial colonies also show substantial stochasticity in gene expression, suggesting that for biological reasons, substantial noise will be inherent in any measure of gene expression (Elowitz et al., 2002). This makes it important to pool cells of a type and to compare multiple samples to understand their average or typical behavior. This could be done using LCM or flow sorting, as described above; it is relatively easy to collect samples of hundreds of cells of one type by LCM, as long as the histological preparation makes them visible and accessible.

Because existing transcriptional profiling methods require the physical disruption of tissues and cells, gene expression is measured only in discrete time points. Ideally, future technologies should monitor transcripts in situ and in real time for the duration of a treatment or developmental phase. The technique mentioned above using labeled DNA probes and FISH permits this type of analysis (Levsky et al., 2002). However, significant advances will be required to make this more practical and to enable large-scale measurements of transcriptional activity.

Intriguing advances in DNA and protein detection are being made with nanoparticles. The laboratory of Chad Mirkin (Northwestern University) has developed methods based on metal nanoparticles coated with oligonucleotides and Raman-active dyes (Cao et al., 2002; Nam et al., 2003). These nanoparticles serve as probes that detect RNA, DNA, or protein targets; these nanoparticles incorporate Raman dyes that can have a wide range of nonoverlapping spectra properties. For the detection of nucleotides, these targets may have first been captured on an underlying chip in microarray format. The use of the Raman dyes permits a large number of multiplexed reactions. The method is also extremely sensitive, with a current unoptimized detection limit of 20 femtomolar (10^{-15}; Cao et al., 2002). A related approach applied to protein detection has a sensitivity approximately a million times higher than standard techniques (Nam et al., 2003). While these types of advances may not generate practical, high-throughput applications for plant biologists for many years, they typify the technologies that will be needed.

Integration of gene expression data with other data sources will, in the future, become a more standard way of molecular experimentation. However, a fundamental challenge remains the development of technologies and mathematical approaches to incorporate disparate and complex data sets. As described above, the full transcriptional complexity of plant genomes is still being described, and it would be a big step forward to measure all functional RNA transcripts, including miRNAs, ncRNAs, and products of alternative splicing and polyadenylation. Such a step would approach a truly global analysis of gene expression. In addition, the methods that we have reviewed above are nearly all based on poly(A) RNA. The concentration of cellular poly(A) RNA is a function of complex processes of transcription, modification, nuclear export, and degradation. Future progress will require devising novel methods and technologies to measure and dissect posttranscriptional processes. Gene expression is also inextricably linked to translation, and measuring proteins and metabolites from the same sample as transcriptional analyses will pose additional challenges. The ability to integrate all of these data over real time and for single cells will require technologies well beyond those that currently exist. The noise that results from the stochastic nature of gene expression will require substantial replication, and the source and amount of variation in measurements due to the technologies will need to be elucidated.

It has been proposed, due to similarities to the semiconductor industry Moore's law, it should be possible in the not too distant future to sequence a human genome for $1000 (http://www.venterscience.org/news.html). Assuming that these technologies are equally applicable to any genome, this would have tremendous implications for plant genetics. However, if the $1000 genome is a goal for the future, we should concurrently aim for a $10 global gene expression measurement. Reduced prices would facilitate better experimental design by eliminating financial restrictions on replication and would open the door to novel types of experiments.

Received February 11, 2004; returned for revision March 19, 2004; accepted March 19, 2004.

LITERATURE CITED

Adams MD, Kerlavage AR, Fleischmann RD, Fuldner RA, Bult CJ, Lee NH, Kirkness EF, Weinstock KG, Gocayne JD, White O, et al (1995) Initial assessment of human gene diversity and expression patterns based on 83 million nucleotides of cDNA sequence. Nature **377:** 3–174

Alonso-Blanco C, Koornneef M (2000) Naturally occurring variation in Arabidopsis: an underexploited resource for plant genetics. Trends Plant Sci **5:** 22–29

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana.* Nature **408:** 796–815

Asano T, Masumura T, Kusano H, Kikuchi S, Kurita A, Shimada H, Kadowaki K (2002) Construction of a specialized cDNA library from plant cells isolated by laser capture microdissection: toward comprehensive analysis of the genes expressed in the rice phloem. Plant J **32:** 401–408

Audic S, Claverie JM (1997) The significance of digital gene expression profiles. Genome Res **7:** 986–995

Auer H, Lyianarachchi S, Newsom D, Klisovic MI, Marcucci U, Kornacker K (2003) Chipping away at the chip bias: RNA degradation in microarray analysis. Nat Genet **35:** 292–293

Aukerman MJ, Sakai H (2003) Regulation of flowering time and floral organ identity by a MicroRNA and its APETALA2-like target genes. Plant Cell **15:** 2730–2741

Banks RE, Dunn MJ, Forbes MA, Stanley A, Pappin D, Naven T, Gough M, Harnden P, Selby PJ (1999) The potential use of laser capture microdissection to selectively obtain distinct populations of cells for proteomic analysis–preliminary findings. Electrophoresis **20:** 689–700

Barczak A, Rodriguez MW, Hanspers K, Koth LL, Tai YC, Bolstad BM, Speed TP, Erle DJ (2003) Spotted long oligonucleotide arrays for human gene expression analysis. Genome Res **13:** 1775–1785

Baum M, Bielau S, Rittner N, Schmid K, Eggelbusch K, Dahms M, Schlauersbach A, Tahedl H, Beier M, Guimil R, et al (2003) Validation of a novel, fully integrated and flexible microarray benchtop facility for gene expression profiling. Nucleic Acids Res **31:** e151

Birnbaum K, Shasha DE, Wang JY, Jung JW, Lambert GM, Galbraith DW, Benfey PN (2003) A gene expression map of the Arabidopsis root. Science **302:** 1956–1960

Blanc G, Barakat A, Guyot R, Cooke R, Delseny M (2000) Extensive duplication and reshuffling in the Arabidopsis genome. Plant Cell **12:** 1093–1101

Borevitz JO, Liang D, Plouffe D, Chang HS, Zhu T, Weigel D, Berry CC, Winzeler E, Chory J (2003) Large-scale identification of single-feature polymorphisms in complex genomes. Genome Res **13:** 513–523

Brandt S, Kloska S, Altmann T, Kehr J (2002) Using array hybridization to monitor gene expression at the single cell level. J Exp Bot **53:** 2315–2323

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, et al (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet **29:** 365–371

Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, et al (2000a) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol **18:** 630–634

Brenner S, Williams SR, Vermaas EH, Storck T, Moon K, McCollum C, Mao JI, Luo S, Kirchner JJ, Eletr S, et al (2000b) In vitro cloning of complex mixtures of DNA on microbeads: physical separation of differentially expressed cDNAs. Proc Natl Acad Sci USA **97:** 1665–1670

Bustin SA (2002) Quantification of mRNA using real-time reverse transcription PCR (RT-PCR): trends and problems. J Mol Endocrinol **29:** 23–39

Cao YC, Jin R, Mirkin CA (2002) Nanoparticles with Raman spectroscopic fingerprints for DNA and RNA detection. Science **297:** 1536–1540

Caron H, van Schaik B, van der Mee M, Baas F, Riggins G, van Sluis P, Hermus MC, van Asperen R, Boon K, Voute PA, et al (2001) The human transcriptome map: clustering of highly expressed genes in chromosomal domains. Science **291:** 1289–1292

Carrington JC, Ambros V (2003) Role of microRNAs in plant and animal development. Science **301:** 336–338

Chakravarthy S, Tuori RP, D'Ascenzo MD, Fobert PR, Despres C, Martin GB (2003) The tomato transcription factor Pti4 regulates defense-related gene expression via GCC box and non-GCC box cis elements. Plant Cell **15:** 3033–3050

Cheung VG, Spielman RS (2002) The genetics of variation in gene expression. Nat Genet **32** (suppl.): 522–525

Christensen TM, Vejlupkova Z, Sharma YK, Arthur KM, Spatafora JW, Albright CA, Meeley RB, Duvick JP, Quatrano RS, Fowler JE (2003) Conserved subgroups and developmental regulation in the monocot rop gene family. Plant Physiol **133:** 1791–1808

Chu TM, Weir B, Wolfinger R (2002) A systematic statistical linear modeling approach to oligonucleotide array experiments. Math Biosci **176:** 35–51

Close TJ, Wanamaker S, Caldo RA, Turner SM, Ashlock DA, Dickerson JA, Wing RA, Muehlbauer GJ, Kleinhofs A, Wise RP (2004) A new resource for cereal genomics: 22K barley GeneChip comes of age. Plant Physiol **134:** 960–968

Craven RA, Totty N, Harnden P, Selby PJ, Banks RE (2002) Laser capture microdissection and two-dimensional polyacrylamide gel electrophoresis: evaluation of tissue preparation and sample limitations. Am J Pathol **160:** 815–822

Delseny M, Cooke R, Raynal M, Grellet F (1997) The Arabidopsis thaliana cDNA sequencing projects. FEBS Lett **405:** 129–132

Draghici S (2002) Statistical intelligence: effective analysis of high-density microarray data. Drug Discov Today **7:** S55–S63

Draghici S, Kulaeva O, Hoff B, Petrov A, Shams S, Tainsky MA (2003) Noise sampling method: an ANOVA approach allowing robust selection of differentially regulated genes measured by DNA microarrays. Bioinformatics **19:** 1348–1359

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

Eddy SR (2001) Non-coding RNA genes and the modern RNA world. Nat Rev Genet **2:** 919–929

Elowitz MB, Levine AJ, Siggia ED, Swain PS (2002) Stochastic gene expression in a single cell. Science **297:** 1183–1186

Emmert-Buck MR, Bonner RF, Smith PD, Chuaqui RF, Zhuang Z, Goldstein SR, Weiss RA, Liotta LA (1996) Laser capture microdissection. Science **274:** 998–1001

Enard W, Khaitovich P, Klose J, Zollner S, Heissig F, Giavalisco P, Nieselt-Struwe K, Muchmore E, Varki A, Ravid R, et al (2002) Intra- and interspecific variation in primate gene expression patterns. Science **296:** 340–343

Ewing RM, Kahla AB, Poirot O, Lopez F, Audic S, Claverie JM (1999) Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. Genome Res **9:** 950–959

Fizames C, Munos S, Cazettes C, Nacry P, Boucherez J, Gaymard F, Piquemal D, Delorme V, Commes T, Doumas P, et al (2004) The Arabidopsis root transcriptome by serial analysis of gene expression. Gene identification using the genome sequence. Plant Physiol **134:** 67–80

Galbraith DW (2003) Global analysis of cell type-specific gene expression. Comp Funct Genomics **4:** 208–215

Gibbings JG, Cook BP, Dufault MR, Madden SL, Khuri S, Turnbull CJ, Dunwell JM (2003) Global transcript analysis of rice leaf and seed using SAGE technology. Plant Biotechnol J **1:** 271–285

Ginzinger DG (2002) Gene quantification using real-time quantitative PCR: an emerging technology hits the mainstream. Exp Hematol **30:** 503–512

Goel A, Seth P, Chauhan SS (2001) Specific amplication of mRNA splice variants by RT-PCR. Biotechniques **30:** 944–949

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science **296:** 92–100

Gowda M, Jantasuriyarat C, Dean R, Wang G-L (2004) Robust-LongSAGE (RL-SAGE) for both gene discovery and transcriptome analysis. Plant Physiol **134:** 890–897

Grant D, Cregan P, Shoemaker RC (2000) Genome organization in dicots: genome duplication in Arabidopsis and synteny between soybean and Arabidopsis. Proc Natl Acad Sci USA **97:** 4168–4173

Green CD, Simons JF, Taillon BE, Lewin DA (2001) Open systems: panoramic views of gene expression. J Immunol Methods **250:** 67–79

Greller LD, Tobin FL (1999) Detecting selective expression of genes and proteins. Genome Res **9:** 282–296

Guo M, Rupe MA, Danilevskaya ON, Yang X, Hu Z (2003) Genome-wide mRNA profiling reveals heterochronic allelic variation and a new imprinted gene in hybrid maize endosperm. Plant J **36:** 30–44

Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, Salzberg SL, White O (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res **31:** 5654–5666

Hagiwara Y, Hirai M, Nishiyama K, Kanazawa I, Ueda T, Sakaki Y, Ito T (1997) Screening for imprinted genes by allelic message display: identification of a paternally expressed gene impact on mouse chromosome 18. Proc Natl Acad Sci USA **94:** 9249–9254

Hayward-Lester A, Oefner PJ, Sabatini S, Doris PA (1995) Accurate and absolute quantitative measurement of gene expression by single-tube RT-PCR and HPLC. Genome Res **5:** 494–499

Heid CA, Stevens J, Livak KJ, Williams PM (1996) Real time quantitative PCR. Genome Res **6:** 986–994

Hoth S, Ikeda Y, Morgante M, Wang X, Zuo J, Hanafey MK, Gaasterland T, Tingey SV, Chua NH (2003) Monitoring genome-wide changes in gene expression in response to endogenous cytokinin reveals targets in Arabidopsis thaliana. FEBS Lett **554:** 373–380

Hoth S, Morgante M, Sanchez JP, Hanafey MK, Tingey SV, Chua NH (2002) Genome-wide gene expression profiling in Arabidopsis thaliana reveals new targets of abscisic acid and largely impaired gene regulation in the abi1-1 mutant. J Cell Sci **115:** 4891–4900

Hsieh WP, Chu TM, Wolfinger RD, Gibson G (2003) Mixed-model reanalysis of primate data suggests tissue and species biases in oligonucleotide-based gene expression profiles. Genetics **165:** 747–757

Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, Lefkowitz SM, Ziman M, Schelter JM, Meyer MR, et al (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nat Biotechnol **19:** 342–347

Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, et al (2000) Functional discovery via a compendium of expression profiles. Cell **102:** 109–126

Ishii M, Hashimoto S, Tsutsumi S, Wada Y, Matsushima K, Kodama T, Aburatani H (2000) Direct comparison of GeneChip and SAGE on the quantitative accuracy in transcript profiling analysis. Genomics **68:** 136–143

Jansen RC, Nap JP (2001) Genetical genomics: the added value from segregation. Trends Genet **17:** 388–391

Jin L, Thompson CA, Qian X, Kuecker SJ, Kulig E, Lloyd RV (1999) Analysis of anterior pituitary hormone mRNA expression in immunophenotypically characterized single cells after laser capture microdissection. Lab Invest **79:** 511–512

Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G, Gibson G (2001) The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. Nat Genet **29:** 389–395

Jung SH, Lee JY, Lee DH (2003) Use of SAGE technology to reveal changes in gene expression in Arabidopsis leaves undergoing cold stress. Plant Mol Biol **52:** 553–567

Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP, Gingeras TR (2002) Large-scale transcriptional activity in chromosomes 21 and 22. Science **296:** 916–919

Karrer EE, Lincoln JE, Hogenhout S, Bennett AB, Bostock RM, Martineau B, Lucas WJ, Gilchrist DG, Alexander D (1995) In situ isolation of mRNA from individual plant cells: creation of cell-specific cDNA libraries. Proc Natl Acad Sci USA **92:** 3814–3818

Kawamoto S, Yoshii J, Mizuno K, Ito K, Miyamoto Y, Ohnishi T, Matoba R, Hori N, Matsumoto Y, Okumura T, et al (2000) BodyMap: a collection of 3′ ESTs for analysis of human gene expression information. Genome Res **10:** 1817–1827

Kawasaki S, Borchert C, Deyholos M, Wang H, Brazille S, Kawai K, Galbraith D, Bohnert HJ (2001) Gene expression profiles during the initial phase of salt stress in rice. Plant Cell **13:** 889–905

Kerk NM, Ceserani T, Tausta SL, Sussex IM, Nelson TM (2003) Laser capture microdissection of cells from plant tissues. Plant Physiol **132:** 27–35

Kerr MK, Martin M, Churchill GA (2000) Analysis of variance for gene expression microarray data. J Comput Biol **7:** 819–837

Klein CA, Seidl S, Petat-Dutter K, Offner S, Geigl JB, Schmidt-Kittler O, Wendler N, Passlick B, Huber RM, Schlimok G, Baeuerle PA, Riethmuller G (2002) Combined transcriptome and genome analysis of single micrometastatic cells. Nat Biotechnol **20:** 387–392

Klein D (2002) Quantification using real-time PCR technology: applications and limitations. Trends Mol Med **8:** 257–260

Kuo WP, Jenssen T-K, Butte AJ, Ohno-Machado L, Kohane IS (2002) Analysis of matched mRNA measurements from two different microarray technologies. Bioinformatics **18:** 405–412

Lash AE, Tolstoshev CM, Wagner L, Schuler GD, Strausberg RL, Riggins GJ, Altschul SF (2000) SAGEmap: a public gene expression resource. Genome Res **10:** 1051–1060

Lee H-S, Wang J, Tian L, Jiang H, Black MA, Madlung A, Watson B, Lukens L, Pires JC, Wang JJ, et al (2004) Sensitivity of 70-mer oligonucleotides and cDNAs for microarray analysis of gene expression in Arabidopsis and its related species. Plant Biotechnol J **2:** 45–57

Lee JY, Lee DH (2003) Use of serial analysis of gene expression technology to reveal changes in gene expression in Arabidopsis pollen undergoing cold stress. Plant Physiol **132:** 517–529

Lee ML, Kuo FC, Whitmore GA, Sklar J (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proc Natl Acad Sci USA **97:** 9834–9839

Levsky JM, Shenoy SM, Pezo RC, Singer RH (2002) Single-cell gene expression profiling. Science **297:** 836–840

Levsky JM, Singer RH (2003) Gene expression and the myth of the average cell. Trends Cell Biol **13:** 4–6

Li C, Wong WH (2001a) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci USA **98:** 31–36

Li C, Wong WH (2001b) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Genome Biol **2:** RESEARCH0032

Lo HS, Wang Z, Hu Y, Yang HH, Gere S, Buetow KH, Lee MP (2003) Allelic variation in gene expression is common in the human genome. Genome Res **13:** 1855–1862

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

Luo L, Salunga RC, Guo H, Bittner A, Joy KC, Galindo JE, Xiao H, Rogers KE, Wan JS, Jackson MR, et al (1999) Gene expression profiles of laser-captured adjacent neuronal subtypes. Nat Med **5:** 117–122

MacIntosh GC, Wilkerson C, Green PJ (2001) Identification and analysis of Arabidopsis expressed sequence tags characteristic of non-coding RNAs. Plant Physiol **127:** 765–776

Matsubara K, Okubo K (1993) cDNA analyses in the human genome project. Gene **135:** 265–274

Matsumura H, Nirasawa S, Terauchi R (1999) Technical advance: transcript profiling in rice (Oryza sativa L.) seedlings using serial analysis of gene expression (SAGE). Plant J **20:** 719–726

Matsumura H, Reich S, Ito A, Saitoh H, Kamoun S, Winter P, Kahl G, Reuter M, Kruger DH, Terauchi R (2003) Gene expression analysis of plant host-pathogen interactions by SuperSAGE. Proc Natl Acad Sci USA **100:** 15718–15723

Mazur B, Krebbers E, Tingey S (1999) Gene discovery and product development for grain quality traits. Science **285:** 372–375

Mei R, Hubbell E, Bekiranov S, Mittmann M, Christians FC, Shen MM, Lu G, Fang J, Liu WM, Ryder T, et al (2003) Probe selection for high-density oligonucleotide arrays. Proc Natl Acad Sci USA **100:** 11237–11242

Meyers BC, Lee DK, Vu TH, Tej SS, Edberg SB, Matvienko M, Tindell LD (2004a) Arabidopsis MPSS: an online resource for quantitative expression analysis. Plant Physiol **135:** 801–813

Meyers BC, Tej SS, Vu TH, Haudenschild C, Agrawal V, Edberg SB, Ghazal H, Decola S (2004b) The use of MPSS for whole-genome transcriptional analysis in Arabidopsis. Genome Res (in press)

Meyers BC, Vu TH, Tej SS, Ghazal H, Matvienko M, Agrawal V, Ning J, Haudenschild C (2004c) Analysis of the transcriptional complexity of Arabidopsis by massively parallel signature sequencing. Nat Biotechnol (in press)

Mikkilineni V, Mitra RD, Merritt J, DiTonno JR, Church GM, Ogunnaike B, Edwards JS (2004) Digital quantitative measurements of gene expression. Biotechnol Bioeng **86:** 117–124

Mitra RD, Church GM (1999) In situ localized amplification and contact replication of many individual DNA molecules. Nucleic Acids Res **27:** e34

Mitra RD, Shendure J, Olejnik J, Edyta Krzymanska O, Church GM (2003) Fluorescent in situ sequencing on polymerase colonies. Anal Biochem **320:** 55–65

Naef F, Hacker CR, Patil N, Magnasco M (2002) Empirical characterization of the expression ratio noise structure in high-density oligonucleotide arrays. Genome Biol **3:** RESEARCH0018

Nakazono M, Qiu F, Borsuk LA, Schnable PS (2003) Laser-capture microdissection, a tool for the global analysis of gene expression in specific plant cell types: identification of genes expressed differentially in epidermal cells or vascular tissues of maize. Plant Cell **15:** 583–596

Nam JM, Thaxton CS, Mirkin CA (2003) Nanoparticle-based bio-bar codes for the ultrasensitive detection of proteins. Science **301:** 1884–1886

Newman T, de Bruijn FJ, Green P, Keegstra K, Kende H, McIntosh L, Ohlrogge J, Raikhel N, Somerville S, Thomashow M, et al (1994) Genes galore: a summary of methods for accessing results from large-scale partial sequencing of anonymous Arabidopsis cDNA clones. Plant Physiol **106:** 1241–1255

Nielsen HB, Wernersson R, Knudsen S (2003) Design of oligonucleotides for microarrays and perspectives for design of multi-transcriptome arrays. Nucleic Acids Res **31:** 3491–3496

Numata K, Kanai A, Saito R, Kondo S, Adachi J, Wilming LG, Hume DA, Hayashizaki Y, Tomita M (2003) Identification of putative noncoding RNAs among the RIKEN mouse full-length cDNA collection. Genome Res **13:** 1301–1306

Nuwaysir EF, Huang W, Albert TJ, Singh J, Nuwaysir K, Pitas A, Richmond T, Gorski T, Berg JP, Ballin J, et al (2002) Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. Genome Res **12:** 1749–1755

Oakey RJ, Beechey CV (2002) Imprinted genes: identification by chromosome rearrangements and post-genomic strategies. Trends Genet **18:** 359–366

Ohyama H, Mahadevappa M, Luukkaa H, Todd R, Warrington JA, Wong DT (2002) Use of laser capture microdissection-generated targets for hybridization of high-density oligonucleotide arrays. Methods Enzymol **356:** 323–333

Okubo K, Hori N, Matoba R, Niiyama T, Fukushima A, Kojima Y, Matsubara K (1992) Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. Nat Genet **2:** 173–179

Okubo K, Itoh K, Fukushima A, Yoshii J, Matsubara K (1995) Monitoring cell physiology by expression profiles and discovering cell type-specific genes by compiled expression profiles. Genomics **30:** 178–186

Oleksiak MF, Churchill GA, Crawford DL (2002) Variation in gene expression within and among natural populations. Nat Genet **32:** 261–266

Orphanides G, Reinberg D (2002) A unified theory of gene expression. Cell **108:** 439–451

Osborn TC, Pires JC, Birchler JA, Auger DL, Chen ZJ, Lee HS, Comai L, Madlung A, Doerge RW, Colot V, et al (2003) Understanding mechanisms of novel gene expression in polyploids. Trends Genet **19:** 141–147

Palatnik JF, Allen E, Wu X, Schommer C, Schwab R, Carrington JC, Weigel D (2003) Control of leaf morphogenesis by microRNAs. Nature **425:** 257–263

Renner W, Pilger E (1999) Simultaneous in vivo quantitation of vascular endothelial growth factor mRNA splice variants. J Vasc Res **36:** 133–138

Roberts JP (2002) The cutting edge in laser microdissection. Biophotonics International **9:** 50–53

Russo G, Zegar C, Giordano A (2003) Advantages and limitations of microarray technology in human cancer. Oncogene **22:** 6497–6507

Saha S, Sparks AB, Rago C, Akmaev V, Wang CJ, Vogelstein B, Kinzler KW, Velculescu VE (2002) Using the transcriptome to annotate the genome. Nat Biotechnol **20:** 508–512

Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science **270:** 467–470

Shimkets RA, Lowe DG, Tai JT, Sehl P, Jin H, Yang R, Predki PF, Rothberg BE, Murtha MT, Roth ME, et al (1999) Gene expression analysis by transcript profiling coupled to a gene database query. Nat Biotechnol **17:** 798–803

Simillion C, Vandepoele K, Van Montagu MC, Zabeau M, Van De Peer Y (2002) The hidden duplication past of *Arabidopsis thaliana.* Proc Natl Acad Sci USA **99:** 13627–13632

Simone NL, Paweletz CP, Charboneau L, Petricoin EF III, Liotta LA (2000) Laser capture microdissection: beyond functional genomics to proteomics. Mol Diagn **5:** 301–307

Slonim DK (2002) From patterns to pathways: gene expression data analysis comes of age. Nat Genet **32** (suppl.): 502–508

Stekel DJ, Git Y, Falciani F (2000) The comparison of gene expression from multiple cDNA libraries. Genome Res **10:** 2055–2061

Stern MD, Anisimov SV, Boheler KR (2003) Can transcriptome size be estimated from SAGE catalogs? Bioinformatics **19:** 443–448

Stollberg J, Urschitz J, Urban Z, Boyd CD (2000) A quantitative evaluation of SAGE. Genome Res **10:** 1241–1248

Talla E, Tekaia F, Brino L, Dujon B (2003) A novel design of whole-genome microarray probes for Saccharomyces cerevisiae which minimizes cross-hybridization. BMC Genomics **4:** 38

Tan PK, Downey TJ, Spitznagel EL Jr, Xu P, Fu D, Dimitrov DS, Lempicki RA, Raaka BM, Cam MC (2003) Evaluation of gene expression measurements from commercial microarray platforms. Nucleic Acids Res **31:** 5676–5684

Vanhee-Brossollet C, Vaquero C (1998) Do natural antisense transcripts make sense in eukaryotes? Gene **211:** 1–9

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. Science **270:** 484–487

Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett DE Jr, Hieter P, Vogelstein B, Kinzler KW (1997) Characterization of the yeast transcriptome. Cell **88:** 243–251

Vision TJ, Brown DG, Tanksley SD (2000) The origins of genomic duplications in Arabidopsis. Science **290:** 2114–2117

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

Wisman E, Ohlrogge J (2000) Arabidopsis microarray service facilities. Plant Physiol **124:** 1468–1471

Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS (2001) Assessing gene significance from cDNA microarray expression data via mixed models. J Comput Biol **8:** 625–637

Wong MH, Saam JR, Stappenbeck TS, Rexer CH, Gordon JI (2000) Genetic mosaic analysis based on Cre recombinase and navigated laser capture microdissection. Proc Natl Acad Sci USA **97:** 12601–12606

Wortman JR, Haas BJ, Hannick LI, Smith RK Jr, Maiti R, Ronning CM, Chan AP, Yu C, Ayele M, Whitelaw CA, et al (2003) Annotation of the Arabidopsis genome. Plant Physiol **132:** 461–468

Xiao YL, Malik M, Whitelaw CA, Town CD (2002) Cloning and Sequencing of cDNAs for Hypothetical Genes from Chromosome 2 of Arabidopsis. Plant Physiol **130:** 2118–2128

Yamada K, Lim J, Dale JM, Chen H, Shinn P, Palm CJ, Southwick AM, Wu HC, Kim C, Nguyen M, et al (2003) Empirical analysis of transcriptional activity in the Arabidopsis genome. Science **302:** 842–846

Yelin R, Dahary D, Sorek R, Levanon EY, Goldstein O, Shoshan A, Diber A, Biton S, Tamir Y, Khosravi R, et al (2003) Widespread occurrence of antisense transcription in the human genome. Nat Biotechnol **21:** 379–386

Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science **296:** 79–92

Yuen T, Wurmbach E, Pfeffer RL, Ebersole BJ, Sealfon SC (2002) Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays. Nucleic Acids Res **30:** e48

Zhang L, Zhou W, Velculescu VE, Kern SE, Hruban RH, Hamilton SR, Vogelstein B, Kinzler KW (1997) Gene expression profiles in normal and cancer cells. Science **276:** 1268–1272

Zhu J, Shendure J, Mitra RD, Church GM (2003) Single molecule profiling of alternative pre-mRNA splicing. Science **301:** 836–838

Zhu T, Wang X (2000) Large-scale profiling of the Arabidopsis transcriptome. Plant Physiol **124:** 1472–1476