Microarray expression profiling resources for plant genomics

Willem Albert Rensink and C. Robin Buell

The Institute for Genomic Research, 9712 Medical Center Dr., Rockville, MD 20850, USA

Large volumes of genomic data have been generated for several plant species over the past decade, including structural sequence data and functional annotation at the genome level. Various technologies such as expressed sequence tags (ESTs), massively parallel signature sequencing (MPSS) and microarrays have been used to study gene expression and to provide functional data for many genes simultaneously. This review focuses on recent advances in the application of microarrays in plant genomic research and in gene expression databases available for plants. Large sets of Arabidopsis microarray data are publicly available. Recently developed array platforms are currently being used to generate genome-wide expression profiles for several crop species. Coupled to these platforms are public databases that provide access to these large-scale expression data, which can be used to aid the functional discovery of gene function.

Functional discovery through microarray expression profiling

Access to a large collection of reference expression data from mutants, tissues or treatments can provide a tool for identifying the function of unknown genes (Figure 1). This concept of functional discovery via a compendium of expression profiles was first demonstrated in yeast [7]: the expression profile of uncharacterized yeast mutants was compared to the larger expression dataset of reference profiles and matched to similar expression profiles of other conditions that displayed similar global expression changes. An alternative approach is the analysis of co-expression. An analysis of the co-expression relationships among a large quantity of expression data from humans, flies, worms and yeast identified relationships that were conserved across evolution [8]. Although the yeast genome is well annotated and is a simpler eukaryote compared with that of higher plants, with the growing availability of microarray expression data, similar approaches could also be used to analyze plants. However, additional validation might be needed and full interpretation could be more complex because of the lower quality of annotation and the poorer understanding of plant genomes compared with those of yeast. For example, a recent study used publicly available Arabidopsis arrays to identify genes required for cellulose synthesis through the analysis of co-regulation with cellulose synthase genes [9], thereby illustrating the potential of mining many expression profiles for functional discovery via co-regulation. A co-response database (CSB.DB) based on publicly available expression data for Arabidopsis is available for the identification and mining of co-regulated genes [10]. Its usefulness was illustrated by the identification of brassinosteroid-related genes by co-regulation analysis of known signaling components [11]. Additional datasets such as protein–protein interaction...
Box 1. Microarray technology

**Definitions**

**Probe:** gene-specific DNA spotted on the array that will hybridize with the target.

**Target:** for an RNA sample, typically either total RNA or mRNA is isolated and labeled using a first strand cDNA synthesis step, either by direct incorporation of a fluorescent dye or by coupling the dyes to a modified nucleotide. For GeneChips, RNA is labeled by incorporating biotin into amplified cRNA, which is detected by fluorescently labeled streptavidin. For non-expression-based experiments, DNA rather than RNA can be labeled and hybridized to the array.

**Array platforms**

**On-slide synthesized arrays:** probes are synthesized on the array surface using DNA synthesis chemistry. The activation for oligonucleotide elongation is achieved using a mask (Affymetrix) or maskless (Nimbiogen) method. Alternatively, the reagents are delivered to each spot using ink-jet technology (Agilent).

**Spotted cDNA arrays:** clones from a cDNA library are amplified by the polymerase chain reaction (PCR) using generic primers for the vector. PCR products are purified and spotted on glass slides using a robotic arrayer.

**Spotted gene-specific sequence tag arrays:** unique segment(s) of the gene are amplified from genomic DNA or bacterial artificial chromosome clones using primers specific for each gene. The PCR products are purified and spotted on glass slides using a robotic arrayer.

**Spotted long oligonucleotide arrays:** oligonucleotides ranging from 50–70 bases are synthesized for a unique region of the genes of interest. Oligonucleotides are spotted on glass slides using a robotic arrayer.

Microarray expression profiling of crop species

Microarray platforms for crop species have been developed as well. Because the research communities for some of these species are smaller, several projects have been organized as consortia to provide a microarray expression platform for these species, either from Affymetrix or as synthesized long oligonucleotide sets (Table 1). Because most of these microarrays have become available only recently, little data are publicly available. However, studies using arrays with crop species has been reported for barley [28,29], grape [30], maize [31–34], pine [35], poplar [36], potato [37–39], tomato [40,41], soybean [42–44] and wheat [45].

Crop species-specific gene expression databases are publicly accessible. Several reference experiments using the barley 22K GeneChip [28] are available through barleyBase [46]. In addition to the barley expression data, barleyBase is expanding to PLEXdb, which is pursuing collection of expression data for several plant species. In addition, a tool is available for linking probes between the different species and platforms within PLEXdb (Table 1). Expression data for soybean generated using a spotted cDNA array are available through the Soybean Genomics and Microarray Database (SGMD) [47]. Large quantities of expression data for the Solanaceae, including potato and tomato, are available through the Solanaceae Gene Expression Database, and tomato gene expression data are available through a tomato expression database (Table 1). Maize gene expression data are available through the Zeamagine database [48].

Arabidopsis and rice: reference species and reference microarray expression data

The *Arabidopsis* community adopted microarrays as a tool when they were first developed. The initial AG Affymetrix array represented only ~8000 unique genes [14], whereas the more recent ATH1 array represents ~23 750 unique genes [15]. It has been shown that expression data obtained with both array versions are largely similar and comparable [16]. In addition to the Affymetrix platform, long oligonucleotide arrays for *Arabidopsis* are available as well as spotted arrays with gene-specific sequence tags [17]. However, based on the predominance of literature reports and available datasets, the ATH1 array now appears to be the platform of choice for the *Arabidopsis* community. Together with ArrayExpress [18], The Nottingham Arabidopsis Stock Center (NASC) now provides the largest resource for *Arabidopsis* expression data because the *Arabidopsis* Information Resource [19] recently stopped collecting expression data. Large quantities of expression data have been generated using the ATH1 array in individual studies and as part of the expression profiling service at NASC [20] (Table 1). All hybridizations and data processing are performed at NASC using user-supplied RNA. Another large *Arabidopsis* expression dataset is provided through the AtgenExpress project, which is also available through NASC. This dataset includes a gene expression map of *Arabidopsis* covering samples of many stages of development [21]. A detailed gene expression map of the *Arabidopsis* root is also available [22].

Currently, only limited amounts of expression data are available for rice. Several published studies have used relatively small-scale spotted cDNA microarrays addressing rice pollination and fertilization [23], responses to elicitors [24], and salt stress [25]. Two studies have been published that used a proprietary rice Affymetrix array to study grain filling of rice [26] and drought tolerance [27]. No whole genome expression data providing a comprehensive expression map for rice are available to date but work is in progress at the Virtual Center for Cellular Expression Profiling of Rice in which laser capture microdissection is being used for whole genome expression profiling of rice cell types (Table 1). Several public array platforms are now available for rice (Table 1). Coupled with the public availability of genome-scale rice arrays, expression databases dedicated to housing rice expression data have also been developed (Table 1).
near future, it can be assumed that large amounts of expression data will be generated; for data integration and cross-species comparisons, special care should be given to the identification of homologous genes and probes representing these genes.

On-line analysis tools for microarray expression data
To navigate the large quantity of data now present in expression databases, tools are being developed to identify experiments of interest as well as expression profiles for genes of interest. On-line expression analysis tools provide some capability for expression data analysis. Expression Profiler is a platform for web-based data analysis [49] in which analysis tools for gene expression data preprocessing, filtering, clustering and visualization methods are integrated with the ArrayExpress repository of expression data. Analysis tools for the Arabidopsis expression data are available at NASC with a more comprehensive set of analysis tools available through Genevestigator [50], which is a web-based platform for the selection and visualization of Arabidopsis expression data. Mapman was developed as a tool to display genomics data onto diagrams of biological processes and metabolic pathways [51]. Although the tool can be downloaded and installed locally to analyze Arabidopsis expression data, some sample data are available for online analysis and display. Barleybase [46] has implemented several on-line visualization and analysis tools for statistical analysis of the data present in the database. In addition to on-line analysis tools, several open-source software packages have been developed for microarray data analysis and interpretation, such as the Bioconductor project, TM4 software and the web-based and BASE system (reviewed in [52]).

Microarray expression data integration and exchange
Even though early expression profiling studies relied mainly on publishing lists of up- or down-regulated genes, it is now well recognized that all experimental information and data associated with expression studies should be made available upon publication. Thus, any published expression study should provide enough information for other researchers to reproduce the experiment and to validate the findings. However, the complex nature of microarray expression data, such as experimental conditions, target and probe generation and relative expression ratios, prevent simple data exchange mechanisms. The MGED (Microarray Gene Expression Data) society has taken the lead in the development of standards accompanying microarray expression data [53].

Figure 1. Functional discovery through expression profiling. Expression data from various sources and plant species provide expression data using a range of platforms. Expression databases provide access to the data. Through data analysis and integration with available genome sequences and annotation data, putative gene functions can be revealed. GeneChip is a registered trademark of Affymetrix.
 importância, o mínimo informação sobre microarrays (MIAF) diretrizes descrevem requisitos para um acurado descreve do ensaios de plantas microarrays, bem como a descrição do microarray e os procedimentos experimentais usados para gerar os dados de expressão. Atualmente, a maioria das MIAF diretrizes têm sido adotadas por muitos métodos de expressão plantas.

A maioria das instituições envolvidas em ensaios de escala grande expressão emprega bancos de dados expressão para análises de dados internamente salvo e/ou uso ou para exportar os dados para o público. Infelizmente, apesar de expressão bibliotecas ou e MIAF o grupo têm adotado seu próprio esquema de banco de dados, que complica os dados de troca. Como consequência, os MGED diretizes têm desenvolvido padrões para armazenar e microarrays de troca chamado MAGE-ML [54].

O requisito de disseminar microarrays de forma pública, preferivelmente fora do site dos autores, resultou na criação de bancos de dados da expressão como (Tabela 1). Atualmente, há dois maiores bancos de dados: o Gene Expression Omnibus (GEO) no National Center for Biotechnology Information (NCBI) [55] e ArrayExpress no European Bioinformatics Institute (EBI) [18]. Ambos os bancos são organizados de maneira similar: primeiro, um microarray é definido com um layout e identificação para os genes em um microarray; seguido, dados de expressão e de ensaios de expressão com diferentes bancos de dados para a integração, requerendo que os dados sejam confiáveis e reproduzíveis não apenas entre plataformas, mas também entre laboratórios. A maioria das plataformas de microarrays proporciona um sistema de busca de dados para GEO, que permite a pesquisa de perfiles de gênero em várias experimentos. Assim, estes bancos de dados fornecem um repositório e um arquivo de dados de expressão, além de um recurso para análise de dados de expressão.

**Técnicas com microarrays**

Para fazer uso dos vários bancos de dados expressão, os dados de função de sequência ou tipo de genoma são utilizados, mas a maioria dos bancos de dados de sequência geralmente fornece informações não suficientes para a integração entre plataformas. No entanto, não todas as plataformas compartilham a mesma base de dados de informações, e a GEO é o principal repositório de dados de expressão. No entanto, GEO não fornece dados de expressão para o desenvolvimento de comunidades de expressão, embora ambos os repositórios forneçam funções de busca de dados e GEO permite a pesquisa de perfis de gênero em vários experimentos. Assim, estes bancos de dados fornecem um repositório e um arquivo de dados de expressão, além de um recurso para análise de dados de expressão.
result of probe-specific effects can be observed [56]. Platforms that provide a high level of technical reproducibility result in higher quality data, an increased capability of detecting differentially expressed genes and more reproducible results among different microarray platforms [57]. An important consideration for cross-platform comparisons is the method used to ‘link’ the probes between the platforms because this will influence the correlation between the platforms. Sequence-based linking of gene identifiers rather than text-based matching identifiers increases cross-platform comparisons [58]. Another level of variability can be derived from between laboratory comparisons. Expression data generated with the same RNA on the same platform resulted in large differences between laboratories, but the best-performing laboratories coupled with the most robust platforms resulted in expression data with good correlation [59]. The application of standard protocols and methods among laboratories using the same platform will increase the reproducibility of expression data among laboratories [60]. Cross-platform reproducibility is best achieved when the sample labeling, hybridization and data processing is technically well performed. This was illustrated by the comparison of various microarray platforms for Arabidopsis that showed good correspondence among platforms of gene-specific PCR products, as well as long and short oligonucleotides [17]. If available, high-quality data generated using robust platforms with alternative probes and technology can be integrated. One complication will be the selection of datasets that meet quality standards. Thus, it can be preferable to use either datasets generated with the same platform or by the same laboratory for functional discovery of gene function. But in general, conclusions derived from expression data using alternative platforms should be reproducible with an alternative microarray expression platform. Overlying all these considerations is the need for replicate data, technical and biological, coupled with appropriate statistical analyses, something that until recently has typically been lacking in plant gene expression studies.

Tiling arrays to identify all transcripts

Another approach to study genome-wide expression is the detection of all transcriptional activity on the chromosomes using tiling arrays. Rather than using gene-specific probes to detect the expression of genes, the complete genome, including the intergenic regions is represented by probes on the array. In addition to detecting transcripts, other applications for tiling arrays include comparative genome hybridizations to detect deletions and polymorphisms, methylation profiling and the analysis of chromatin immuno-precipitation samples [61] because these types of experiments cannot be performed as well with the common expression arrays. The first plant species for which tiling expression data are available are rice and Arabidopsis. The Arabidopsis genome sequence was represented on high-density Affymetrix arrays. RNA samples were hybridized to these arrays to identify genes and gene structures experimentally [62]. For Arabidopsis, this resulted in the adjustment of gene models as well as the identification of novel transcripts.

A commercial version of the Arabidopsis tiling array will become available sometime next year. Because the genome size of rice is much larger than the genome of Arabidopsis, more arrays or higher density arrays are required for tiling the complete rice genome. The transcriptome of rice chromosome 10 has been analyzed using a short oligonucleotide tiling array [63] and chromosome 4 has been analyzed using a PCR-fragment-based tiling array [64]. Both studies provide evidence for expression of genes and insights into regulation of transcription along the chromosome. Owing to the lower density of these rice tiling arrays compared with those of Arabidopsis, the ability to refine gene structure using these arrays is limited.

An advantage of tiling arrays over expression arrays is that they are less biased for the expressed sequences because no assumption is made about the gene structure and which part of the genome sequence is coding. Another non-biased high-throughput method to determine expression levels for many genes simultaneously is provided by massively parallel signature sequencing (MPSS), which produces short sequence signatures and whose frequency is correlated to the expression level. To be maximally successful, both approaches require the availability of the genome sequence. These approaches seem to be well suited to generate a reference for gene expression in various tissues. MPSS was successfully applied to Arabidopsis to analyze the transcriptome in 14 libraries [65] and is currently being used to examine the rice transcriptome via 60 different libraries (Table 1).

Future perspectives

The plant community has embraced gene expression profiling technology, which is evident by the number of species for which one or more array platforms are available to date. As illustrated above, large expression databases have the potential to accelerate functional discoveries. Although these expression databases provide access to expression data and the experimental conditions used to generate the expression data, to complement genome annotation efforts effectively and to increase gene functional discovery, additional efforts are desirable. A first step would be the generation of reference expression datasets of cell, tissue or developmental-specific gene expression for each plant species. These gene expression compendia, such as AtgenExpress and the rice expression project, would provide a starting point to catalogue gene expression throughout the life cycle for each species. A second step would be the addition of gene expression data of mutants and treatments that will aid in the identification of expression and regulatory networks. A third step would be the integration of multiple data types because with the increased availability of expression platforms for multiple plant species, it is likely that the function of several unknown genes will be derived from coupling expression data with other functional and/or phenotypic data. Thus, to exploit these data fully and to embrace a complete systems approach, alternative data types such as knockout phenotypic data, metabolomics and proteomics data will need to be integrated with expression data and/or made readily available. A fourth step would be increased availability of pre-analyzed statistical and
clustering analyses because, owing to computational limitations, on-line analysis tools have limited capabilities to perform data integrations on a genome-scale. Improving large-scale databases could accelerate these efforts. Currently, plant gene expression data are scattered and stored in multiple databases, often separated between species, which inhibits cross-species comparisons. Special attention should be given to the identification of true orthologs among these plant species. These advances will be made available in the coming years not only for Arabidopsis and rice but also for other species. Coupled with these advances will be the improved structural annotation of the genomes and an increase in the number of genes with a known function. With the non-expression applications of arrays, even more functional information about the genome should be available, accelerating the discovery of the function of all the genes within a plant genome.

Acknowledgements
Work on plant gene expression databases at The Institute for Genomic Research is provided by funding to C.R.B. by the National Science Foundation Plant Genome Research Program (DBI 0113887 and DBI 021663).

References
6 Yuan, Q. et al. (2005) The institute for genomic research osa1 rice genome annotation database. Plant Physiol. 138, 18–26
18 Parkinson, H. et al. (2005) ArrayExpress – a public repository for microarray gene expression data at the EBI. Nucleic Acids Res. 33, D553–D555
36 Moreau, C. et al. (2005) A genomic approach to investigate developmental cell death in woody tissues of Populus trees. Genome Biol. 6, R34
40 Baxter, C.J. et al. (2005) Comparison of changes in fruit gene expression in tomato introgression lines provides evidence of genome-wide transcriptional changes and reveals links to mapped QTLs and described traits. J. Exp. Bot. 56, 1591–1604
43 Vodkin, L. et al. (2004) Microarrays for global expression constructed with a low redundancy set of 27,500 sequenced cDNAs representing an array of developmental stages and physiological conditions of the soybean plant. BMC Genomics 5, 73
50 Thimm, O. et al. (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J. 37, 914–939
58 Mecham, B.H. et al. (2004) Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. Nucleic Acids Res. 32, e74
63 Li, L. et al. (2005) Tiling microarray analysis of rice chromosome 10 to identify the transcriptome and relate its expression to chromosomal architecture. Genome Biol. 6, R52
64 Jiao, Y. et al. (2005) A tiling microarray expression analysis of rice chromosome 4 suggests a chromosome-level regulation of transcription. Plant Cell 17, 1641–1657