Large-Scale Statistical Analyses of Rice ESTs Reveal Correlated Patterns of Gene Expression

Rob M. Ewing,[1] Alia Ben Kahla, Olivier Poirot, Fabrice Lopez, Stéphane Audic, and Jean-Michel Claverie

Structural and Genetic Information (SGI) Laboratory, Centre National de la Recherche Scientifique–Unité mixte de Recherche (CNRS–UMR) 1889, Marseille 13402 cedex 20, France

Large, publicly available collections of expressed sequence tags (ESTs) have been generated from *Arabidopsis thaliana* and rice (*Oryza sativa*). A potential, but relatively unexplored application of this data is in the study of plant gene expression. Other EST data, mainly from human and mouse, have been successfully used to point out genes exhibiting tissue- or disease-specific expression, as well as for identification of alternative transcripts. In this report, we go a step further in showing that computer analyses of plant EST data can be used to generate evidence of correlated expression patterns of genes across various tissues. Furthermore, tissue types and organs can be classified with respect to one another on the basis of their global gene expression patterns. As in previous studies, expression profiles are first estimated from EST counts. By clustering gene expression profiles or whole cDNA library profiles, we show that genes with similar functions, or cDNA libraries expected to share patterns of gene expression, are grouped together. Promising uses of this technique include functional genomics, in which evidence of correlated expression might complement (or substitute for) those of sequence similarity in the annotation of anonymous genes and identification of surrogate markers. The analysis presented here combines the application of a correlation-based clustering method with a graphical color map allowing intuitive visualization of patterns within a large table of expression measurements.

The development of distinct tissues and cell-types is a fundamental characteristic of the growth of higher organisms. Tissue and cellular differentiation, in turn, is highly dependent on specific patterns of gene expression and transcript accumulation.

In higher plants, a large volume of literature exists documenting spatial and temporal regulation of gene expression. It is increasingly clear that developmental pathways can be considered as modular, and that developmental transitions are accompanied by global changes in the expression of specific complements of genes (Doebley and Lukens 1998). For example, the intensively studied transition from etiolated to greening seedling involves coordinate regulation of many light-regulated genes (von Arnim and Deng 1996). Also increasingly clear, is the notion that complements of genes are best studied in parallel, which has become feasible with the development of new technologies (Schena et al. 1996, 1998; Wen et al. 1998).

Traditional approaches to the analysis of mRNA abundance, such as Northern blotting, tend to be limited by the number of transcripts that can be simultaneously analyzed. More recent innovations, such as hybridization to arrayed cDNA libraries or oligonucleotide chips permit simultaneous analysis of the abundance of thousands of transcripts (for review, see Brown and Botstein 1999). These latter approaches can be thought of as analog, because hybridization signal intensity reflects transcript abundance. In plants, the use of arrays of partially sequenced cDNAs has been successfully applied to the analysis of gene expression in light- and dark-grown seedlings of *Arabidopsis* (Desprez et al. 1998).

Digital analysis of gene expression can be achieved by generation of tags to expressed genes and transcript abundance inferred from the frequency of tags. This approach has been used with both conventional ESTs (Okubo et al. 1992; Lee et al. 1995; Takenaka et al. 1998) and in the SAGE technique with much shorter (10 bp) tags (Velculescu et al. 1995, 1997; Zhang et al. 1997). The availability of significant collections of expressed sequence tags from plant genomes presents an opportunity to analyze digital expression profiles for plant tissues and genes. Several studies have observed that the abundance of EST tags for many genes varies according to the tissue of origin of the cDNA library. (Uchimiya et al. 1992; Hofte et al. 1993; Umeda et al. 1994; Cooke et al. 1996; Yamamoto and Sasaki 1997). Because EST data is inherently noisy (Aaronson et al. 1996; Hillier et al. 1996; Wolfsberg and Landsman 1997), a rigorous statistical test was derived to assess the reliability of the identification of differentially expressed genes from EST counts sampled from different libraries (Audic and Claverie 1997). EST data has also been used to reveal alternative transcripts of the same

[1]Corresponding author.
E-MAIL ewing@igs.cnrs-mrs.fr; FAX 33 (0)4 91 16 45 49.

9:950–959 ©1999 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/99 $5.00; www.genome.org

gene, as well as their eventual library-specific distribution (Burke et al. 1998; Gautheret et al. 1998).

As of October 1998, there are ~37,000 *Arabidopsis* and 27,000 rice publicly available EST sequences, as well as smaller collections from other plant species (http://www.ncbi.nlm.nih.gov/dbEST). An important difference between the *Arabidopsis* and rice ESTs (at least for the purposes described in this report) is that a large proportion of the *Arabidopsis* ESTs were generated from a single cDNA library, prepared from a mixture of tissues (Newman et al. 1994; Delseny et al. 1997), whereas the rice ESTs are more evenly derived from a set of tissue and organ-specific cDNA libraries, therefore making them a more suitable starting point for gene expression studies (Yamamoto and Sasaki 1997).

A significant proportion of ESTs show no similarity to sequences in existing databases (Adams et al. 1992; Claverie 1996). Ascribing functions to those anonymous sequences has therefore become one of the major bottlenecks in plant and animal genomics. One way of gaining functional information on anonymous genes is by use of the two-hybrid system (for review, see Brent and Finley 1997). According to this approach, direct physical interactions of the product of an unknown gene are used to reveal its relationships with the product of (hopefully) better-characterized ones. Using publicly-available rice ESTs as a test set, we show that a multidimensional analysis of EST data can provide similar types of information, albeit based on the concept of statistical rather than physical interactions. Functional relationships between genes may then be inferred from the mathematical identification of significant similarities between their expression patterns.

Using the rice ESTs available in dbEST (Boguski et al.1993) we have computed an expression profile for each gene represented by at least 5 ESTs in 10 different cDNA libraries. For each of those genes, the expression profile is therefore derived from 10 expression measurements (EST counts). Correlation analysis was then used to point out significant similarities in the expression profiles of genes as well as to generate a graphical representation of gene clusters exhibiting related expression patterns. Our results indicate that genes with similar functions, or tissues expected to share patterns of gene expression, can be recognized by use of this type of analysis. The multidimensional analysis of EST data, in a way quite parallel to microarray experiments (DeRisi et al. 1996; Eisen et al. 1998), may thus constitute a new approach to the functional annotation of anonymous genes and to a more global understanding of plant physiology.

RESULTS

EST Database and Contigs

A breakdown of the rice cDNA libraries represented in dbEST (as of 10/98) is shown in Table 1. Preliminary investigations in which expression profiles were generated from all libraries with >100 ESTs showed that the smaller libraries gave misleading results (data not shown). Therefore, of the 27 cDNA libraries that contribute to the EST set, only the 10 largest (representing 95% of the total ESTs) were used in the analysis presented here. These 10 cDNA libraries contribute varying numbers of ESTs to the dataset used; the difference between the largest and smallest rice cDNA libraries used here is approximately fivefold (library 1073 has 5094 ESTs, library 1009 has 890 ESTs).

Rice ESTs were organized into clusters and contig (consensus) sequences derived by a protocol adapted from Gautheret et al. (1998) (see Methods). Selected

Table 1. Breakdown of Rice cDNA Libraries Represented in dbEST

dbEST library identifier	Library description	Number of ESTs
1073	immature seed (5 days after pollination)	5094
307	green shoot (8 days old)	3790
961	callus	3542
75	callus	3229
193	etiolated shoot (8 days old)	3148
499	panicle (at flowering stage)	2025
101	seedling root	1884
535	panicle (at ripening stage)	1478
1275	[a]callus (H. Uchimiya)	1431
1009	panicle (shorter than 3 cm)	890
967	panicle (longer than 10 cm)	365
1010	immature leaf including apical meristem	358
322	FDRSC	310
621	etiolated leaf tissue	132
966	panicle (between 3 and 10 cm)	74
968	etiolated shoot	43
1404	late flower (panicle size 1–2 cm)	42
1396	etiolated leaf	32
1281	suspension culture	29
466	FDRRC	13
1137	[b]seed (A. Suzuki)	6
969	leaf (photoperiod insensitive)	2
1395	early flower (panicle size <1 cm)	2
1321	early embryogenesis	2
1243	—	2
1176	salt stressed	2
1266	[c]etiolated shoot (Y. Jiang)	1

Data as of October 1998. For each library, the dbEST identifier is shown with a short description if available (taken from dbEST). Only the 10 cDNA libraries contributing significant numbers of ESTs (libraries 1073 to 1009) were further used in this study.
[a]H. Uchimiya, University of Tokyo, Japan.
[b]A. Suzuki, National Institute of Agrobiological Resources, Tsukuba, Japan.
[c]Y. Jiang, Fudan University, Shanghai, China.

statistics of the clustered set of rice ESTs are shown in Table 2.

Derivation of Expression Profiles

Expression profiles were derived for each of the 707 contigs with 5 or more constituent ESTs. The cDNA library of origin was scored for each constituent EST of each contig, producing a two-way, contig versus library, table of raw EST counts (see Fig. 1). The content of this table was the primary data for all subsequent computations.

In a preliminary investigation, an alternative protocol was explored in which the raw EST counts were further reduced to a binary scale, such that simply the presence or absence of a given gene in a given library was recorded (one or more EST=1, none=0). The subsequent statistical analysis of this binary data (with, e.g., Fisher's 2×2 exact test) was found to be much less sensitive and meaningful than the analyses performed with the raw EST counts. The remainder of this report, therefore, focuses on the identification of correlated expression patterns with a statistical analysis of actual EST counts.

Assessing the Pairwise Similarity Between Expression Profiles

The first aim of our analysis is to identify pairs of genes (represented here by EST contigs) exhibiting a similar, multicondition (i.e., cDNA library) expression pattern. For each gene, the data consists of 10 numbers (EST counts) defining an expression profile (see Fig. 1 for overview of entire procedure). If two genes are expressed in a coordinated manner, we expect their expression profiles to have similar shapes, that is, the two series of EST counts to follow the same up or down trend. Given that the absolute EST counts vary widely

Table 2. Statistics of EST Clustering and Contiging

(*a*) *Preparation of rice EST clusters*	
ESTs analyzed	27877
ESTs remaining after quality control	27710
Clusters	3400
Singletons	11233
(*b*) *Breakdown of clusters with 5 or more ESTs*	
With 5 or more ESTs	707
With 100 or more ESTs	9
With between 10 and 99 ESTs	220
With between 5 and 9 ESTs	478
(*c*) *Contig sequence matches*	
Contig sequences finding homolog	602
Contig sequences with no homolog	105

[a]Using all available ESTs as of October 1998.
[b]ESTs from smaller cDNA libraries excluded; see Table 1.
[c]Matches to SWISS-PROT/TrEMBL, at scores >100 (default scoring matrix).

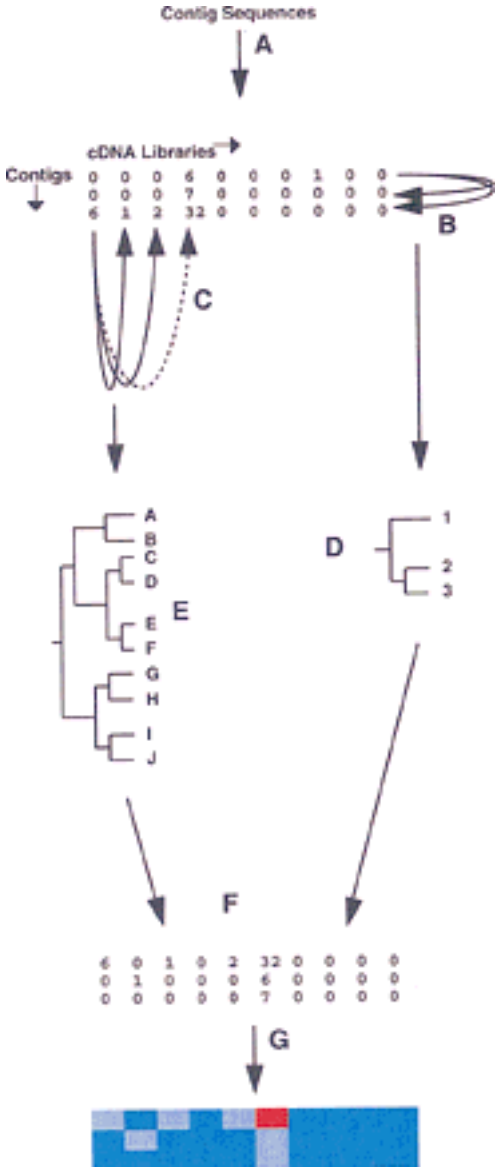

Figure 1 Overview of the procedure. (*A*)Derivation of expression profiles from each valid contig sequence (those representing five or more ESTs from the 10 cDNA libraries used) to form the primary data table in which contig sequence are arrayed in rows and cDNA libraries in columns. In the illustration, a hypothetical matrix of 3 contigs (1–3) by 10 libraries (*A–J*) is shown. (*B,C*) Pairwise calculation of similarity (Pearson correlation coefficient) between contig (row) and library (column) profiles, respectively. (*D,E*) Calculation of Euclidean distances for each contig/contig or library/library pair from Pearson coefficient matrices, and generation of contig (*D*) and library (*E*) dendrograms. (*F*) Separate reordering of rows (contigs) and columns (libraries) in the original data table, according to the hierarchical order in the contig and library dendrograms. In the reordered data table, similar contigs or similar libraries are adjacent. (*G*) Representation of the reordered data table as a clustered correlation map, in which the color in an individual cell reflects the underlying EST count.

between contigs (the number of constituent ESTs per contig ranges between 308 and 5) and libraries (there is a fivefold difference between the number of ESTs contributed by the largest and by the smallest cDNA libraries—see Table 1), a meaningful measure of expression profile similarity had to be independent of those absolute numbers. Within these constraints, the Pearson linear correlation coefficient (see Methods) represents a natural, easy to compute, similarity measure. The value of this coefficient varies from -1 to 1; a value close to 1 indicates a high similarity of the compared expression profiles (i.e., proportionality between the EST counts of two genes), whereas a value close to zero indicates no coordinated expression. A useful property of this coefficient is its capacity to also point out pairs of genes exhibiting opposite expression behavior (anti-correlated profiles, for example, sequences expressed in mutually exclusive sets of libraries), potentially another form of biologically interesting gene coupling. In this latter case, the Pearson coefficient value approaches -1.

Finally, a significance level (*P* value) is associated with the computation of this correlation coefficient, allowing the evidence of pairwise coordinated gene expression to be ranked according to reliability [as with BLAST (Altschul et al. 1990) for sequence similarity].

To first confirm that computing Pearson's correlation coefficient is an appropriate way of identifying correlated expression profiles, groups of contigs with highly correlated profiles were analyzed. First, pairs of contigs with high correlation coefficients (in this case, $r > 0.94$), were identified within the 707×707 (symmetrical) matrix of pairwise gene expression profile correlation coefficients. These pairs of contigs were then organized into mutually matching clusters, whereby each profile in a cluster matches all of the others in the same cluster at the required stringency ($r > 0.94$). Table 3 shows two such clusters of contigs.

Table 3. Two Clusters of Correlated Contig Sequences *a* and *b*, with Highly Correlated ($r > 0.94$) Expression Profiles

		Library EST counts									
Contig	Putative identity	1073	535	307	499	961	75	1275	193	101	1009
(*a*)											
1	PRO2_ORYSA 13-kD Prolamin	293	15	0	0	0	0	0	0	0	0
4	GLU5_ORYSA glutelin	183	9	0	0	0	0	0	0	0	0
5	GLU2_ORYSA glutelin type II	145	28	0	0	0	0	0	0	0	0
7	PRO1_ORYSA 10-kD prolamin	145	10	0	0	0	0	0	0	0	0
12	—	74	0	0	0	0	0	0	0	0	0
15	—	63	0	0	0	0	0	0	0	0	0
34	PRO7_ORYSA prolamin	29	4	0	0	0	0	0	0	0	0
35	GLU3_ORYSA glutelin type-A III	27	3	0	0	0	0	0	0	0	0
41	GLGB_ORYSA 1,4-α-glucan branching enzyme	26	1	0	0	0	0	0	0	0	0
53	GLGB_MAIZE 1,4-α-glucan branching enzyme IIB	21	0	0	0	0	0	0	0	0	0
83	—	11	0	0	0	0	1	0	0	0	0
125	PODK_MAIZE pyruvate phosphate dikinase	12	0	0	0	0	0	0	0	0	0
148	PHSL_VICFA α-1,4 glucan phosphorylase	10	0	0	0	0	1	0	0	0	0
171	ASPR_HORVU phytepsin	10	0	0	0	0	0	0	0	0	0
209	UGST_ORYSA granule-bound starch synthase	8	1	0	0	0	0	0	0	0	0
304	SUS2_HORVU sucrose synthase 2	6	1	0	0	0	0	0	0	0	0
366	—	6	0	0	0	0	0	0	0	0	0
367	—	6	0	0	0	0	0	0	0	0	0
378	—	5	1	0	0	0	0	0	0	0	0
(*b*)											
3	RA05_ORYSA seed allergenic protein RA5	59	170	0	0	0	0	0	0	0	0
20	—	9	57	0	0	0	0	0	0	0	0
52	RA17_ORYSA seed allergenic protein RA17	5	21	0	0	0	0	0	0	0	0
95	IAA1_HORVU α-amylase inhibitor BMAI-1	2	16	0	0	0	0	0	0	0	0
114	P93615 ABA-induced plasma membrane protein	2	14	0	0	0	0	0	0	0	0
282	CDP1_ORYSA calcium-dependent protein kinase	0	8	0	0	0	0	0	0	0	0
308	P322_SOLTU probable protease inhibitor P322	1	7	0	0	0	0	0	0	0	0
481	OLE1_ORYSA oleosin 16 kD	0	6	0	0	0	0	0	0	0	0
488	—	0	6	0	0	0	0	0	0	0	0

Contigs are listed with a putative identification, if available, corresponding to the best match in the SWISS-PROT/TrEMBL databases (score > 100; default scoring matrix). The expression profile for each contig is shown for the 10 libraries used (identified by dbEST library identifiers), arranged in the same order as in the library dendrogram (see Fig. 2).

The expression profile and putative identity is shown for each contig. Profiles in both groups of contigs are characterized principally by expression in libraries 1073 and 535 (immature seed and panicle at ripening stage). However, the contigs form two discrete clusters on the basis of linear correlation. Thus, for the group of contigs in Table 3a, expression is several fold higher in library 1073 than in library 535, whereas the converse is true for the group of contigs in Table 3b. Most of the contigs in the two clusters encode proteins with seed-related functions, in particular storage proteins, concurring with previous observations of over-representation of prolamin and glutelin transcripts in rice seed cDNA libraries (Liu et al. 1995; Yamamoto and Sasaki 1997).

The Pearson correlation coefficient therefore permits fine-scale identification of sequences with correlated expression profiles.

Assessing the Pairwise Similarity Between cDNA Libraries

The degree of pairwise similarity between whole cDNA libraries can be similarly assessed with the Pearson correlation coefficient. The same table of multi-condition expression data is used, although with rows and columns exchanging roles. For each of the 10 sampled libraries, the profiles now consist of the 707 numbers (EST counts) characterizing the level of expression of each gene. If two tissues express a similar complement of genes, we expect the EST sampling of the corresponding cDNA libraries to exhibit similar profiles, hence, to be characterized by a high pairwise correlation coefficient. The computation of Pearson's coefficient between all cDNA libraries results in a 10×10 (symmetrical) matrix that will be used in building the graphical representation of the expression data.

A Two-Dimensional Graphical Representation Revealing Gene Clusters

The second aim of our study is to build a graphical representation of the whole table of multi-condition expression measurements, as a way to visualize clusters of genes obeying similar expression patterns.

To combine the library and contig data into a single representation, we adapted the clustered correlation approach pioneered by Weinstein et al. (1997) (Fig. 1). This technique involves reordering the results of multidimensional assays (in the latter study, N compounds vs. M tumors) so as to reveal discrete islands of regularities (e.g., different compounds affecting a similar subset of tumors, or different tumors affected by a common subset of compounds). This is performed by reordering the rows (and columns) of the data table so that the most similar ones are adjacent to each other. In our case, the data table consists of the expression measurements of 707 genes (rows) in 10 cDNA libraries (columns).

In the first step, a $N \times N$ row pairwise metric distance matrix is computed (see Methods) and then used to build a dendrogram that assembles all rows into a single tree. The rows are then reordered according to their hierarchical position in this tree. In our case, the contig/gene pairwise distance matrix is derived (see Methods) from the matrix of pairwise correlation coefficients described above. Adjacent genes then have similar expression profiles (Fig. 1). Given the large number of contigs (707), the complete contig dendrogram has not been reproduced here, although fragments are shown in Fig. 3, below. (The complete dendrogram and other data is available from the authors).

The same procedure is used to assign pairwise distances to cDNA libraries (Fig. 1) and reorder them in the table of EST counts. Adjacent libraries are then those apparently expressing the most similar subsets of genes (Fig. 1). The tree derived from the library correlation analysis is shown in Figure 2. As would be predicted, libraries derived from similar tissue types (callus libraries 1275, 961, and 75) or libraries derived from overlapping tissues (library 535 from panicle at ripening stage and library 1073 from immature seed) cluster together. This validates our method, and suggests that the cDNA libraries analyzed are reliable sources of expression data.

Other nearest neighbours on the tree include libraries 499 and 307 (panicle at flowering and green shoot at 8-days old, respectively). Interestingly, library 193 (etiolated shoot 8-days old) and library 307 (green shoot at 8 days old), are not paired, suggesting significant differences in expression patterns between these tissues. This is explained by the massive induction of light-regulated transcripts that occurs during the greening process, which are present in green but not

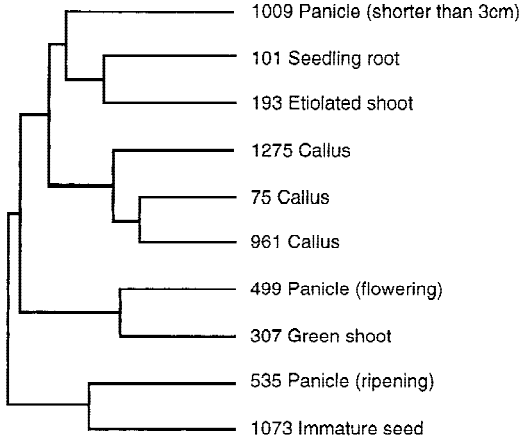

Figure 2 Dendrogram showing cDNA library similarities. Each library is identified by the dbEST library identifier and a short description.

etiolated tissue. These differences are also illustrated in the clustered correlation map shown in Figure 3.

Once optimally reordered according to both contig and to library similarity, the expression measurement table can be graphically represented as a map, with the color in a given cell reflecting the underlying EST count (Fig. 1). Following the reordering of rows and columns, clusters of genes exhibiting coordinated expression appear as blocks of similar color, and are readily identified either by visual inspection, or automatically via the use of classical image-processing techniques.

Figure 3 shows the complete clustered correlation map generated from the rice data. To illustrate ways in which the data may be explored, two fractions of the map have been expanded and annotated with contig numbers and putative identities (Fig. 3B, C). These show that contigs in close proximity on the map may represent genes with related functions. In addition, regions indicated by arrows on the map correspond to clusters of contigs expressed more or less specifically in a particular library; for example the green arrow indicates contig sequences expressed at high levels in library 307 (green shoot, 8-days old), many of which encode chloroplast component precursors.

The clustered correlation map therefore enables expression patterns of interest to be selected prior to identification of specific sequences. The clustered correlation map and associated results are available from the authors.

DISCUSSION

This report presents a new protocol for the analysis of EST data aimed at discovering correlated patterns of

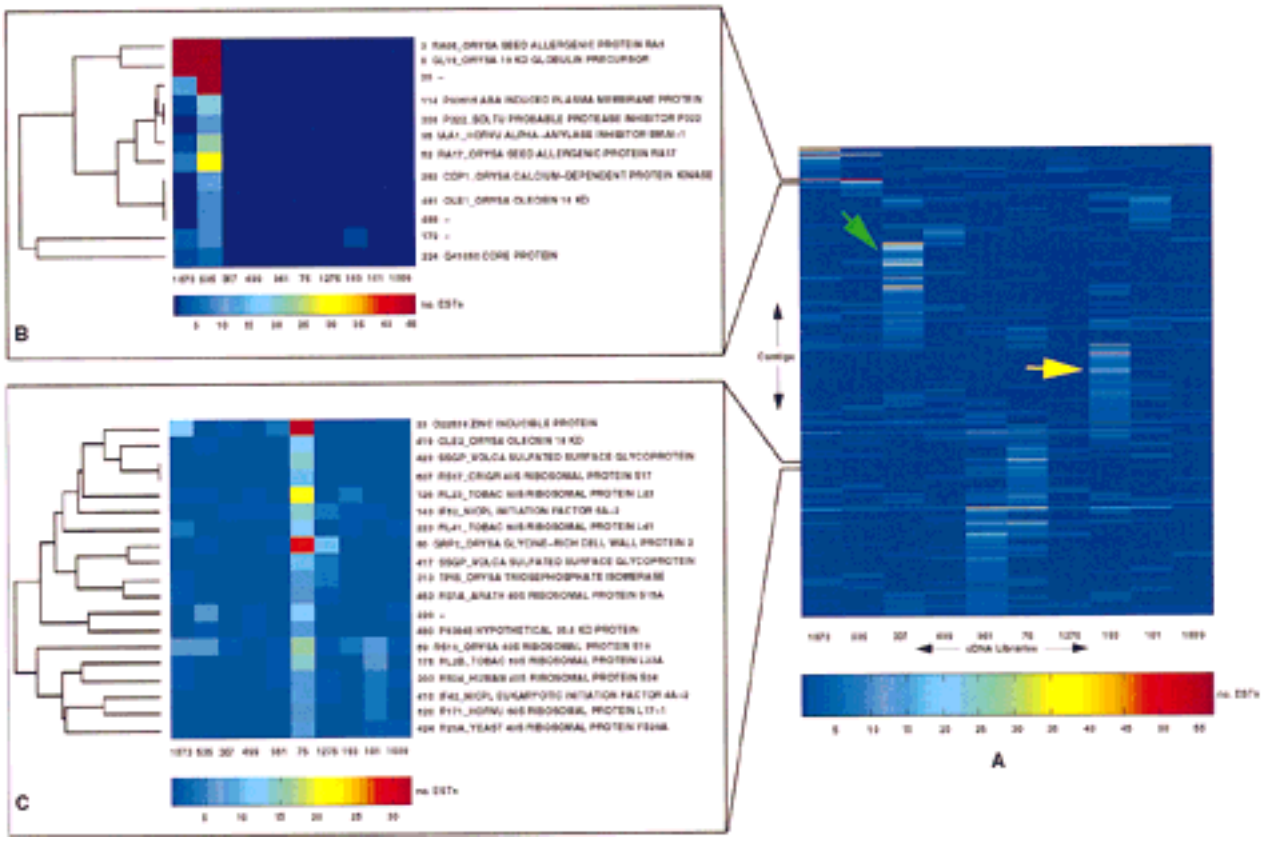

Figure 3 Clustered correlation map of the rice EST data. (*A*) The complete clustered correlation map of 707 contig sequences vs. 10 cDNA libraries is shown. Library identifiers (dbEST library identifiers, see Table 1) identify each library; contig identifiers are not shown in full, but are shown for each of the expanded regions in *B* and *C*. Absolute numbers of ESTs are represented according to the color scale shown below the map. Note that because of restrictions on the number of distinguishable colors, the color scale has been chosen so as to optimally represent a portion of the data (those cells with EST counts between 5 and 55; all cells with a value >55 are assigned the color red). The green and yellow arrows have been placed to show how the map might be used to identify groups of genes with particular expression patterns. In this case, the green arrow indicates a block of genes principally expressed in library 307 (green shoot, 8 days old) and the yellow arrow a block of genes principally expressed in library 193 (etiolated shoot, 8-days old). (*B,C*) Expanded regions of the clustered correlation map. To the *right* of each region, contigs are identified by contig number and putative identification (SWISS-PROT/TrEMBL identifier and description), if available. To the *left* of each region, the relevant portion of the dendrogram used to reorder the original data table is shown. Note the differences in color scale for *A*, *B*, and *C*, as the color scale was chosen in each case to optimally represent the required interval of EST counts.

gene expression between different tissues, with the rice EST database as a test set. Despite the inherent noise of EST data, and the relatively small size of the data set analyzed, our results show that coherent patterns of gene expression can be revealed. The approach permits both the association of tissues via their common patterns of gene expression and the association of genes via their tissue-dependent expression patterns.

The set of cDNA libraries used to generate the rice ESTs are sufficiently varied to cover each of the principal tissues in the plant life cycle (Yamamoto and Sasaki 1997). In addition, groups of libraries representing the same tissues at different developmental stages (e.g., libraries from panicle tissues) or the same tissue type under different growth conditions (e.g. libraries from callus tissue) are present within the 10 libraries analyzed in our study. By use of this data set, our methods show how whole transcriptomes from different tissues can be compared in a statistical manner. Tissues for which gene expression profiles would be expected to overlap, such as 1073 and 535 (immature seed and panicle at ripening stage, respectively), or 961, 75, and 1275 (all from callus), are found to have overlapping profiles. Similarly, genes with inter-related functions, such as those involved in seedling physiology shown in Table 3, are found to have correlated expression profiles. The strength of the method lies in the fact that clustering is based on expression profiles; prior knowledge of sequence identity is not required. Furthermore, the anonymous sequences in Table 3 (i.e., contigs 12, 15, 83, 366, 367, and 378 in the first cluster) illustrate how expression profile clustering might aid candidate gene selection; in this particular example, the anonymous sequences in Table 3 would be good candidates for identification of novel genes involved in seed metabolism.

Clustering genes by expression profile may also enable identification of novel regulatory elements, as genes with correlated profiles might be expected to have regulatory elements in common (DeRisi et al. 1997; Brazma et al. 1998). Other possible uses include the identification of surrogate markers (e.g., Figueroa et al. 1998; Johnson et al. 1998), whereby a conveniently assayed biomarker allows monitoring or prediction of a particular condition (e.g., a gene or cluster of genes whose expression profiles consistently correlate with an agricultural trait of interest).

Overall similarities between tissues are clearly revealed by the dendrogram or the two-dimensional clustered correlation map representation of expression profiles. These types of observations may contribute to a new understanding of the interrelationships between different tissues and developmental pathways. For example, it has long been hypothesized that leaves and certain floral organs derive from a common ancestral organ, an idea supported by documented instances of common regulatory processes during leaf and floral morphogenesis (Arber and Parkin 1907; Satina and Blakeslee 1941; Steeves and Sussex 1989; Bowman et al. 1993; Hofer et al. 1997). Large-scale studies of gene expression may support these hypotheses by identifying tissues with similar or overlapping patterns of gene expression.

The value of expression profiles from EST collections, and the potential for functional prediction are entirely dependent on the available data. In addition, certain assumptions are implicit when using EST collections for transcript profiling. First, to ensure that tag frequency correlates with the actual transcript abundance in a given tissue, the cDNA libraries should have been prepared in a comparable manner. For example, normalized cDNA libraries (e.g., Patanjali et al.1991), in which the frequencies of clones representing abundant and rare transcripts are normalized with respect to one another, are not suitable for a study of this type (although some large effect might still be detectable by a binary presence/absence coding of the original multicondition EST counts). In addition, ESTs should be contributed to the databases without prior selection for novel sequences (in some cases redundancy within EST sets is reduced by first screening the existing EST set and then only submitting sequences not already present). Potential errors may also originate from the EST-clustering procedure. For instance, ESTs derived from the 5′ and 3′ ends of a long transcript may constitute discrete contigs. However, this is not anticipated to be a major problem in the technique presented here.

The potential of large-scale gene expression analysis is most often discussed in the context of hybridization techniques such as cDNA microarrays (see Duggan et al. 1999) or synthetic oligonucleotide arrays (for a recent review, see Lipshutz et al. 1999). These technologies have been applied in several systems including two independent studies of the yeast transcriptome (Wodicka et al. 1997; Eisen et al. 1998), the monitoring of 1000 human genes in activated human T-cells (Schena et al. 1996), and the analysis of the fibroblast transcriptional response to serum (Iyer et al. 1999). Studies have also been performed on subsets of *Arabidopsis* cDNAs (Schena et al. 1995; Desprez et al. 1998) and on a subset of human genes related to inflammation (Heller et al. 1997). These accomplishments should not hide the fact that the high-density microarray technology is still only marginally accessible to academic laboratories (Cheung et al. 1999). On the other hand, the established EST (Adams et al. 1992; Okubo et al. 1992) or SAGE (Velculescu et al. 1995) approaches have proven their capacity in monitoring gene expression in a large variety of experimental systems (Lee et al. 1995; Anderson and Seilhamer 1997; Madden et al. 1997; Velculescu et al. 1997; Zhang et al. 1997; He et al. 1998; Hibi et al. 1998; Takenaka et al.

1998; de Waard et al. 1999), including plants (Uchimiya et al. 1992; Hofte et al. 1993; Liu et al. 1995; Yamamoto et al. 1997). The EST approach is unique in allowing both expression measurements and the discovery of new genes at the same time, whereas microarray techniques are limited to a repertoire of previously identified sequences. Furthermore, ESTs have a wide range of applications including mapping and studies of colinearity (Sasaki 1996). Several studies have shown that EST/SAGE sampling experiments can reliably identify differentially expressed genes (Lee et al. 1995; Audic and Claverie 1997; He et al. 1998; Greller and Tobin 1999). In more recent work, it has been shown that the analysis of EST data can provide valuable insight into the existence and the expression patterns of alternative transcript forms (Burke et al. 1998; Gautheret et al. 1998). In the present article, we show that the analysis of this data can be extended beyond the simple recognition of differential expression to the identification of gene subsets exhibiting coordinated expression patterns.

From a statistical point of view, multicondition expression data obtained from hybridization arrays or cDNA tag sampling are quite similar. They both result in gene abundance estimates stored in a gene versus cDNA library table. Thus, it is expected that after a first step of signal processing (such as noise filtering, pixel detection, thresholding, and normalization) specific to the microarray technique involved, similar statistical treatment could be applied. In the case of EST or tag data, initial signal processing consists mainly of selecting genes and libraries for which total tag counts are large enough to eventually lead to statistically significant inferences (in our own study, selecting contigs representing five or more ESTs, and those cDNA libraries from which >800 ESTs have been generated). Our analysis is then quite similar to the approach independently followed by Eisen et al. (1998) to identify coordinated gene expression in yeast using cDNA microarrays. For instance, both use the Pearson correlation coefficient as the primary statistical parameter to quantify the similarity of expression profiles. However, slightly different metrics for the subsequent hierarchical clustering of genes were used; whereas Eisen et al. (1998) directly used the pairwise correlation coefficient between genes, we computed a true Euclidean distance from the whole gene versus gene correlation coefficient matrix. The distance between two genes is thus computed from the similarity of their expression with all other genes in the matrix, and not from a single pairwise correlation. This procedure, which minimizes the influence of random fluctuation in tag counting, might also serve in smoothing the noise of microarray pixel data. The sensitivity of expression analysis from EST data depends to an extent on the number of ESTs sequenced. Theoretically, expression profiles could be derived for even very weakly expressed genes if sufficient numbers of ESTs were generated. This contrasts with current limitations of microarray technology, in which sensitivity is limited by the quantity of RNA used per hybridization, making detection of very weakly expressed transcripts difficult (see Duggan et al. 1999).

The nature of our multicondition expression data also allowed us to perform hierarchical clustering of both rows (genes) and columns (cDNA libraries), resulting in a two-dimensional clustering (following Weinstein et al. 1997) indicative of both gene and library expression similarity. Similar genes are thus graphically clustered into islands of simple shape (Fig. 3). In a subsequent development of our display program, the visual recognition of these islands will be supplemented by standard image processing algorithms, an attractive alternative to the complexity of more abstract clustering algorithms.

With increased definition of EST collections, (e.g., cDNA libraries prepared from tighter developmental windows, or cDNA libraries prepared from specific cell types), digital expression profiles will become increasingly valuable sources of expression information. This information, alongside expression data from other large-scale approaches, has an important role to play in our efforts to assign function to anonymous sequences.

METHODS

EST Database and Contigs

Rice ESTs were extracted from GenBank version 107 with Batch Entrez at the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/Entrez/batch.html). dbEST (Boguski et al. 1993) reports were obtained with the Sequence Retrieval System (SRS) at the Human Genome Mapping Project (http://iron.hgmp.mrc.ac.uk/).

Rice ESTs were quality controlled and organized into contigs as described elsewhere (Ewing et al. 1999; http://igs-server.cnrs-mrs.fr/ewing). The protocol involved a classical preliminary cleaning of the EST data (vector removal, elimination of low quality sequences), a stringent pairwise comparison of all cleaned EST sequences, followed by the separate contiging of overlapping ESTs. Because our aim is a statistical analysis of gene expression profiles, contigs derived from fewer than five constituent ESTs were excluded from the study. Putative identities (Table 2) were assigned to every resulting contig sequence by querying them against the SWISS-PROT/TrEMBL (36.0) database (Bairoch and Apweiler 1998) with gapped BLASTx (Altschul et al. 1997).

Contig and Library Correlation Analysis

The similarity between contigs (genes) or cDNA library expression profiles was estimated by Pearson's r coefficient, quantifying the degree of linear correlation between two variables, $X = (x_1,x_2,...,x_N)$ and $Y = (y_1,y_2,...,y_N)$.

Given a sample of N pairs of score, r quantifies the extent to which we can make useful predictions on the value of Y from the knowledge of the corresponding X score. The measure of correlation, r, is computed as

$$r = \frac{\sum_{i=0}^{i=N} (\chi_i - \overline{X})(y_i - \overline{Y})}{\sqrt{\left(\sum_{i=0}^{i=N} (\chi_i - \overline{X})^2 \sum_{i=0}^{i=N} (y_i - \overline{Y})^2\right)}}$$

and has a value between −1 and +1. Values of r near to 0 indicate a low degree of correlation. Positive values of r indicate that high values of X are associated with high values of Y. Negative values of r indicate that low values of X are associated with high values of Y (anti-correlation) or vice versa.

The pairwise gene expression correlation coefficients were computed by the repetitive use of the above formula, in which X and Y are different genes associated with their corresponding EST counts $(x_1,x_2, \ldots ,x_N)$ and $(y_1,y_2, \ldots ,y_N)$ measured in cDNA libraries 1,2 ,N (with N = 10). The result of these computations constitutes a 707 × 707 symmetrical matrix of correlation values and a matrix of pairwise gene distances was subsequently derived from it as described below.

Alternatively, a table of the pairwise library correlation coefficient was computed, now taking X and Y as different libraries associated with the EST counts $(x_1,x_2, \ldots ,x_N)$ and $(y_1,y_2, \ldots ,y_N)$ corresponding to the various genes 1,2, . . . ,N (with N = 707). The result of these computations constitutes a 10 × 10 symmetrical matrix of correlation values. As for the gene distance values, a matrix of pairwise library distances was derived as described below.

Hierarchical Classification of Genes and Libraries

The hierarchical classification (dendrogram) of objects requires the calculation of the distance between all pairs of objects. From the gene correlation matrix constructed previously (the elements of which are r values ranging from −1 to 1), a pairwise Euclidean distance matrix was derived as follows. The Euclidean distance d, between two sets, $X = (x_1,x_2, \ldots ,x_N)$ and $Y = (y_1,y_2, \ldots ,y_N)$ is simply computed as

$$d(X,Y) = \sqrt{\sum_{i=0}^{i=N} (\chi_i - y_i)^2}$$

The above formula can then be used for all pairs of genes X and Y defined by their list of correlation coefficients $(x_1,x_2, \ldots ,x_N)$ and $(y_1,y_2, \ldots ,y_N)$.

By the same method, the 10 × 10 matrix of library correlation coefficients was used to derive pairwise distance values between libraries. The gene and library distance matrices were then used to build their associated dendrograms according to the UPGMA algorithm (Sokal and Michener 1958), implemented in the neighbor program (Kuhner and Felsenstein 1994). Dendrograms were plotted with the njplot program (Perriere and Gouy 1996). The order of contigs and libraries in their respective dendrograms were used to reorder the original data table. The reordered data table was then used as the basis for plotting the clustered correlation map, generated with Matlab 5.2 (MathWorks, Inc.).

ACKNOWLEDGMENTS

The financial support of Novartis Crop Protection, Inc. is gratefully acknowledged. We also thank Dr David Robertson for help with dendrograms and Suzanne Dixon for reading the manuscript.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Aaronson, J., B. Eckman, R. Blevins, J. Borkowski, J. Myerson, S. Imran, and K. Elliston. 1996. Toward the development of a gene index to the human genome: An assessment of the nature of high-throughput EST sequence data. *Genome Res.* **6:** 829–845.

Adams, M., M. Dubnick, A. Kerlavage, R. Moreno, J. Kelley, T. Utterback, J. Nagle, C. Fields, and J. Venter. 1992. Sequence identification of 2,375 human brain genes. *Nature* **355:** 632–634.

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Altschul, S., T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Anderson L. and J. Seilhamer. 1997. A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis* **18:** 533–537.

Arber, E. and J. Parkin. 1907. On the origin of angiosperms. *J. Linnean Soc.* **38:** 29–80.

Audic, S. and J.-M. Claverie. 1997. The significance of digital gene expression profiles. *Genome Res.* **7:** 986–995.

Bairoch, A. and R. Apweiler. 1998. *Nucleic Acids Res.* **26:** 38–42.

Boguski, M.S., T.M.J. Lowe, and C.M. Tolstoshev. 1993. DbEST—database for "expressed sequence tags". *Nat. Genet.* **4:** 332–333

Bowman, J., J. Alvarez, D. Weigel, E. Meyerowitz, and D. Smyth. 1993. Control of flower development in *Arabidopsis thaliana* by apetala1 and interacting genes. *Development* **119:** 721–743.

Brent, R. and R.L. Finley. 1997. Understanding gene and allele function with two-hybrid methods. *Annu. Rev. Genet.* **31:** 663–704.

Brown, P.O. and D. Botstein. 1999. Exploring the new world of the genome with DNA microarrays. *Nat. Genet.* **21:** 33–37.

Burke, J., H. Wang, W. Hide, and D.B. Davison. 1998. Alternative gene form discovery and candidate gene selection from gene indexing projects. *Genome Res.* **8:** 276–290.

Cheung, V.G., M. Morley, F. Aguilar, A. Massimi, R. Kucherlapati, and G. Childs. 1999. Making and reading microarrays. *Nat. Genet.* (Suppl.)**21:** 15–19.

Claverie, J.-M. 1996. Exploring the vast territory of uncharted ESTs. In *Genomes, molecular biology and drug discovery*. Chapter 4, pp. 56–71. Academic Press,

Cooke, R., M. Raynal, M. Laudie, F. Grellet, M. Delseny, P. Morris, D. Guerrier, J. Giraudat, F. Quigley, G. Clabault, et al. 1996. Further progress towards a catalogue of all *Arabidopsis* genes: Analysis of a set of 5000 non-redundant ESTs. *Plant J.* **9:** 101–124.

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

DeRisi, J.L., L. Penland, P.O. Brown, M.L. Bittner, P.S. Meltzer, M. Ray, Y. Chen, Y.A. Su, and J.M. Trent. 1996. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat. Genet.* **14:** 457–460.

DeRisi, J.L., V.R. Iyer, and P.O. Brown. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278:** 680–686.

Desprez, T., J. Amselem, M. Caboche, and H. Hofte. 1998. Differential gene expression in Arabidopsis monitored using cDNA arrays. *Plant J.* **14:** 643–652.

de Waard, V., B.M.M. Berg, J. Veken, R. Heienbrok, H. Pannekoek, and A.J. Zonneveld. 1999. Serial analysis of gene expression to assess the endothelial cell response to an atherogenic stimulus. *Gene* **226:** 1–8.

Doebley, J. and L. Lukens. 1998. Transcriptional regulators and the evolution of plant form. *Plant Cell* **10:** 1075–1082.

Duggan, D.J., M. Bittner, Y. Chen, P. Meltzer, and J.M. Trent. 1999.

Expression profiling using cDNA microarrays. *Nat. Genet.* (Suppl.) **21:** 10–14.
Eisen, M., P. Spellman, P. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.* **95:** 14863–14868.
Ewing, R., O. Poirot, and J.-M. Claverie. 1999. Comparative analysis of the Arabidopsis and rice expressed sequence tag (EST) sets. In *Silico Biol.* **1:** 18 [http://www.bioinfo.de/isb/1999/o1/0018]
Figueroa, J.A., S. Raad, L. Tadlock, V.O. Speights, and J.J. Rinehart. 1998. Differential expression of insulin-like growth factor binding proteins in high versus low Gleason score prostate cancer. *J. Urol.* **159:** 1379–1383.
Gautheret, D., O. Poirot, F. Lopez, S. Audic, and J.-M. Claverie. 1998. Alternative polyadenylation in human mRNAs: A large-scale analysis by EST clustering. *Genome Res.* **8:** 524–530.
Greller, L.D. and F.L. Tobin. 1999. Detecting selective expression of genes and proteins. *Genome Res.* **9:** 282–296.
He, T.C., A.B. Sparks, C. Rago, H. Hermeking, L. Zawel, L.T. Costa, P.J. Morin, B. Vogelstein, and K.W. Kinzler. 1998. Identification of c-MYC as a target of the APC pathway. *Science* **281:** 1509–1512.
Heller, R.A., M. Schena, A. Chai, D. Shalon, T. Bedilion, J. Gilmore, D.E. Woolley, and R.W. Davis. 1997. Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc. Natl. Acad. Sci.* **94:** 2150–2155.
Hibi, K., Q. Liu, G.A. Beaudry, S.L. Madden, W.H. Westra, S.L. Wehage, S.C. Yang, R.F. Heitmiller, A.H. Bertelsen, D. Sidransky, and J. Jen. 1998. Serial analysis of gene expression in non-small cell lung cancer. *Cancer Res.* **58:** 5690–5694.
Hillier, L., G. Lennon, M. Becker, M. Bonaldo, B. Chiapelli, N. Dietrich, T. DuBuque, A. Favello, W. Gish, M. Hawkins et al. 1996. Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.* **6:** 807–828.
Hofer, J., L. Turner, R. Hellens, M. Ambrose, P. Matthews, A. Michael, and N. Ellis. 1997. Unifoliata regulates leaf and flower morphogenesis in pea. *Curr. Biol.* **7:** 581–587.
Hofte, H., T. Desprez, J. Amselem, H. Chiapello, M. Caboche, A. Moisan, M. Jourjon, J. Charpenteau, P. Berthomieu, and D. Guerrier. 1993. An inventory of 1152 expressed sequence tags obtained by partial sequencing of cDNAs from *Arabidopsis thaliana*. *Plant J.* **4:** 1051–1061.
Iyer, V., M. Eisen, D. Ross, G. Schuler, T. Moore, J. Lee, J. Trent, L. Staudt, J. Hudson, M. Boguski, D. Lashkari, D. Shalon, D. Botstein, and P.O. Brown. 1999. The transcriptional program in the response of human fibroblasts to serum. *Science* **283:** 83–87.
Johnson, B.J., I. Estrada, Z. Shen, S. Ress, P. Willcox, M.J. Colston, and G. Kaplan. 1998. Differential gene expression in response to adjunctive recombinant human interleukin-2 immunotherapy in multidrug-resistant tuberculosis patients. *Infect. Immun.* **66:** 2426–2423.
Kuhner, M.K. and J. Felsenstein. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* **11:** 459–468.
Lee, N., K. Weinstock, E. Kirkness, J. Earlehughes, R.F.S. Marmaros, A.J. Glodek, M. Adams, A. Kerlavage, and C.F.J. Venter. 1995. Comparative expressed-sequence-tag analysis of differential gene-expression profiles in pc-12 cells before and after nerve growth- factor treatment. *Proc. Natl. Acad. Sci.* **92:** 8303–8307.
Lipshutz, R.J., S.P.A. Fodor, T.R. Gingeras, and D.J. Lockhart. 1999. High density synthetic oligonucleotide arrays. *Nat. Genet.* (Suppl.) **21:** 20–24.
Liu, L., C. Hara, M. Umeda, Y. Zhao, T. Okita, and H. Uchimiya. 1995. Analysis of randomly isolated cDNAs from developing endosperm of rice (*Oryza sativa*): Evaluation of expressed sequence tags, and expression levels of mRNAs. *Plant Mol. Biol.* **29:** 685–689.
Madden, S.L., E.A. Galella, J. Zhu, A.H. Bertelsen, and G.A. Beaudry. 1997. SAGE transcript profiles for p53-dependent growth regulation. *Oncogene* **15:** 1079–1085.
Newman, T., F. Bruijn, P. Green, K. Keegstra, H. Kende, L. McIntosh, J. Ohlrogge, N. Raikhel, S. Somerville, M. Thomashow 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.
Okubo, K., N. Hori, R. Matoba, T. Niiyama, A. Fukushima, Y. Kojima, and K. Matsubara. 1992. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nat. Genet.* **2:** 173–179.
Patanjali, S.R., S. Parimoo, and S.M. Weisman. 1991. Construction of a uniform-abundance (normalised) cDNA library. *Proc. Natl. Acad. Sci.* **85:** 2444–2448.
Perriere, G. and M. Gouy. 1996. WWW-query: An online retrieval system for biological sequence banks. *Biochimie* **78:** 364–369.
Sasaki, T. 1996. Rice cDNAs as a model for expressed genes of plants. *Symp. Soc. Exp. Biol.* **50:** 11–15.
Satina, S. and A. Blakeslee. 1941. Periclinal chimeras in Datura stramonium in relation to development of leaf and flower. *Am .J. Bot.* **28:** 862–871.
Schena, M., D. Shalon, R.W. Davis, and P.O. Brown. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270:** 467–470.
Schena, M., D. Shalon, R. Heller, A. Chai, P. Brown, and R. Davis. 1996. Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. *Proc. Natl. Acad. Sci.* **93:** 10614–10619.
Schena, M., T.P. Theriault, K. Konrad, E. Lachenmeier, and R.W. Davis. 1998. Microarrays: Biotechnology's discovery platform for functional genomics. *Trends Biotechnol.* **16:** 301–316.
Sokal, R. and C. Michener. 1958. A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.* **28:** 1409–1438.
Steeves, T. A. and I. M. Sussex. 1989. *Patterns in plant development.* Cambridge University Press, Cambridge, UK.
Takenaka, M., E. Imai, T. Kaneko, T. Ito, T. Moriyama, A. Yamauchi, M. Hori, S. Kawamoto, and K. Okubo. 1998. Isolation of genes identified in mouse renal proximal tubule by comparing different gene expression profiles. *Kidney Int.* **53:** 562–572.
Uchimiya, H., S. Kidou, T. Shimazaki, S. Aotsuka, S. Takamatsu, R. Nishi, H. Hashimoto, Y. Matsubayashi, N. Kidou, M. Umeda, and A. Kata. 1992. Random sequencing of cDNA libraries reveals a variety of expressed genes in cultured-cells of rice (*Oryza sativa*). *Plant J.* **2:** 1005–1009.
Umeda, M., C. Hara, Y. Matsubayashi, H. Li, Q. Liu, F. Tadokoro, S. Aotsuka, and H. Uchimiya. 1994. Expressed sequence tags from cultured-cells of rice (*Oryza sativa*) under stressed conditions—analysis of transcripts of genes engaged in ATP-generating pathways. *Plant Mol. Biol.* **25:** 469–478.
Velculescu, V., L. Zhang, B. Vogelstein, and K. Kinzler. 1995. Serial analysis of gene expression. *Science* **270:** 484–487.
Velculescu, V., L. Zhang, W. Zhou, B. Vogelstein, M. Basrai, D. E. Bassett, P. Hieter, B. Vogelstein, and K. Kinzler. 1997. Characterisation of the yeast transcriptome. *Cell* **88:** 243–251.
von Arnim, A. and X. Deng. 1996. Light control of seedling development. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **47:** 215–243.
Weinstein, J., T. Myers, P. Connor, S. Friend, K.W. Kohn, T. Fojo, S. Bates, L. Rubinstein, N. Anderson, J. Buolamwini et al. 1997. An information-intensive approach to the molecular pharmacology of cancer. *Science* **275:** 343–349.
Wen, X., S. Fuhrman, G. Michaels, D. Carr, S. Smith, J. Barker, and R. Somogyi. 1998. Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl. Acad. Sci.* **95:** 334–339.
Wodicka, L., H. Dong, M. Mittmann, M.H. Ho, and D.J. Lockhart. 1997. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nat. Biotechnol.* **15:** 1359–1367.
Wolfsberg, T. and D. Landsman. 1997. A comparison of expressed sequence tags (ESTs) to human genomic sequences. *Nucleic Acids Res.* **25:** 1626–1632.
Yamamoto, K. and T. Sasaki. 1997. Large scale EST sequencing in rice. *Plant Mol. Biol.* **35:** 135–144.
Zhang, L., W. Zhou, V. Velculescu, S. Kern, R. Hruban, S. Hamilton, B. Vogelstein, and K. Kinzler. 1997. Gene expression profiles in normal and cancer cells. *Science* **276:** 1268–1272.

Received April 26, 1999; accepted in revised form August 4, 1999.