Annotating Genes of Known and Unknown Function by Large-Scale Coexpression Analysis[1][W][OA]

Kevin Horan, Charles Jang, Julia Bailey-Serres, Ron Mittler, Christian Shelton, Jeff F. Harper, Jian-Kang Zhu, John C. Cushman, Martin Gollery, and Thomas Girke*

Department of Botany and Plant Sciences (K.H., C.J., J.B.-S., J.-K.Z., T.G.), and Department of Computer Science and Engineering (C.S.), University of California, Riverside, California 92521; Department of Biochemistry and Molecular Biology, University of Nevada, Reno, Nevada 89557 (R.M., J.F.H., J.C.C.); Department of Plant Science, Hebrew University of Jerusalem, Givat Ram Jerusalem 91904, Israel (R.M.); and TimeLogic, Division of Active Motif, Incline Village, Nevada 89451 (M.G.)

About 40% of the proteins encoded in eukaryotic genomes are proteins of unknown function (PUFs). Their functional characterization remains one of the main challenges in modern biology. In this study we identified the PUF encoding genes from Arabidopsis (*Arabidopsis thaliana*) using a combination of sequence similarity, domain-based, and empirical approaches. Large-scale gene expression analyses of 1,310 publicly available Affymetrix chips were performed to associate the identified PUF genes with regulatory networks and biological processes of known function. To generate quality results, the study was restricted to expression sets with replicated samples. First, genome-wide clustering and gene function enrichment analysis of clusters allowed us to associate 1,541 PUF genes with tightly coexpressed genes for proteins of known function (PKFs). Over 70% of them could be assigned to more specific biological process annotations than the ones available in the current Gene Ontology release. The most highly overrepresented functional categories in the obtained clusters were ribosome assembly, photosynthesis, and cell wall pathways. Interestingly, the majority of the PUF genes appeared to be controlled by the same regulatory networks as most PKF genes, because clusters enriched in PUF genes were extremely rare. Second, large-scale analysis of differentially expressed genes was applied to identify a comprehensive set of abiotic stress-response genes. This analysis resulted in the identification of 269 PKF and 104 PUF genes that responded to a wide variety of abiotic stresses, whereas 608 PKF and 206 PUF genes responded predominantly to specific stress treatments. The provided coexpression and differentially expressed gene data represent an important resource for guiding future functional characterization experiments of PUF and PKF genes. Finally, the public Plant Gene Expression Database (http://bioweb.ucr.edu/PED) was developed as part of this project to provide efficient access and mining tools for the vast gene expression data of this study.

Only a small percentage of the proteins encoded in animal or plant genomes are sufficiently characterized with regard to their cellular functions. The functions for the majority of these proteins remain either completely unknown (40%) or only partially understood (Gollery et al., 2006, 2007). In light of this significant knowledge deficit, our understanding about existing molecular functions (MFs) appears to be fundamentally incomplete. This is even more evident when we assume that the vast space of unexplored molecular and biological functions is composed of proteins with at least comparable or even greater diversity and importance for cellular processes than the known space. Efforts to narrow this knowledge gap will provide a wide spectrum of opportunities for advancing our understanding about plant and nonplant systems.

Two major methods are in use for defining proteins of unknown functions (PUFs) in model organisms. The widely used similarity approach considers all proteins as PUFs that show no detectable sequence or structural similarities to functionally characterized proteins in reference databases (Boeckmann et al., 2003; Leinonen et al., 2004). In contrast to this, the more conservative empirical approach defines as PUFs all proteins that lack direct experimental evidence as support for a specific function. Conceptually, the empirical approach incorporates most PUFs identified by the similarity approach, as well as functionally uncharacterized sequences that share sequence similarities with proteins of known function (PKFs). Sequence families and ortholog clusters are particularly affected by this fundamental difference between the two unknown definitions. For instance, when a group of related sequences contains one or more members of known function, then the similarity approach tends to assign all of them to the known space, whereas the empirical approach distinguishes between functionally characterized and uncharacterized candidates within groups of related

[1] This work was supported by National Science Foundation grants 2010–0420033, 2010–0420152, and IGERT–0504249, and National Institutes of Health grant P20–RR–016464.

* Corresponding author; e-mail tgirke@citrus.ucr.edu.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Thomas Girke (tgirke@citrus.ucr.edu).

[W] The online version of this article contains Web-only data.

[OA] Open Access articles can be viewed online without a subscription.

www.plantphysiol.org/cgi/doi/10.1104/pp.108.117366

sequences. As a result of this difference, most similarity-based PUFs of a given genome are either singletons or members of families that consist exclusively of uncharacterized sequences. These performance characteristics of the similarity concept result in an underestimation of the number of PUFs, because many genes in eukaryotic organisms are members of poorly characterized gene families (Horan et al., 2005). To illustrate this, all members of large families, like protein kinases or cytochrome P450s, will be assigned by the similarity approach to the known protein space, even though most of their members remain functionally uncharacterized (Wang et al., 2003; Nelson et al., 2004; Horan et al., 2005).

Dividing gene products into only two categories of known and unknown sequences is an oversimplification of a complex knowledge system with incremental and multifaceted differences. Consequently, every definition for drawing a strict separation line remains artificial and controversial. While acknowledging these difficulties, this study will adopt this two-class system mainly for practical reasons.

To advance our knowledge beyond a roadmap of knowing what we do not know, it is important to develop and apply approaches for predicting putative functions for PUFs. Bioinformatic techniques provide here a wide spectrum of opportunities. For instance, PUFs can be associated with remotely related PKFs by using sensitive sequence and structure similarity search strategies (Eddy, 1996; Altschul et al., 1997). The detected similarities can reveal important clues for testing their functions experimentally. Additionally, one can predict functional features from their sequences, such as subcellular targeting signals, secondary structures, and membrane domains (Schwacke et al., 2003; Gollery et al., 2006). Proteomics and protein interaction technologies provide additional important functional links (Johnson and Liu, 2006). However, for plants the required proteome resources are not yet available on a genome-wide level. One of the most promising and readily available information resources for systematic functional assignment studies of PUF genes represent large-scale gene expression data from public microarray databases. These data sets offer vast opportunities for associating PUF genes with MFs and cellular processes of coregulated PKF genes.

In this study we identified and analyzed the genome-wide PUF encoding genes from Arabidopsis (*Arabidopsis thaliana*) using both empirical and similarity strategies. Large-scale analysis of publicly available gene expression array data allowed us to associate PUF with PKF genes based on similarities of their expression and treatment response profiles. For this, cluster analysis was used to identify groups of coregulated PUF and PKF genes based on the similarity of their expression profiles across a wide range of tissue and treatment samples. Subsequently, enrichment analysis of Gene Ontology (GO) terms was applied to annotate the obtained clusters by overrepresented gene functions. Second, statistical analysis of differentially expressed genes (DEGs) allowed us to identify PUFs that exhibit generic and specific expression changes in response to a large number of different abiotic stress treatments. Finally, the Plant Gene Expression Database (PED) was developed to provide to the public efficient data mining utilities for the complex differential expression and clustering data of this project.

RESULTS AND DISCUSSION

Identification of PUFs

To obtain for this study a comprehensive set of PUFs from Arabidopsis, we compared three profoundly different PUF identification methods. The three approaches are based on GO annotations, sequence similarities, and protein domain searches.

First, we mined the GO annotations to estimate the number of PKFs and PUFs from a manually curated knowledge system that combines empirical and computational methods for assigning gene functions (Berardini et al., 2004; Falcon and Gentleman, 2007). Alternative pathway annotation systems from KEGG and AraCyc could have been used for the same purpose (Mueller et al., 2003; Kanehisa et al., 2006). However, due to the limited number of Arabidopsis genes (<40%) assigned to pathways, the GO system, with close to 95% genome coverage, appears to be currently the more efficient resource for identifying nearly complete PUF sets. This number includes the direct assignments to the root term of each ontology, which are the new GO annotations for sequences of unknown function (see ''Material and Methods'' for more details).

The evidence codes of the GO annotations specify which functional assignments are supported by experimental evidence data from the public domain and which annotations are solely based on computational prediction methods (Ashburner et al., 2000). To gain insight into the nature of the annotations with regard to the evidence type for assigning members to the known and unknown space, we combined in Table I the current set of 13 evidence codes into four custom categories. The category with the highest level of functional support (empirical) is based on direct evidence from traditional single sample experiments, the second one is based on large-scale screening data (large scale), the third one on computational predictions (sequence), and the fourth one on the GO-based PUF entries that lack functional support from experiments or in silico analyses. The detailed assignment schema of the evidence codes to the four categories is provided in the legend of Table I.

According to the above strategy, 32% to 38% of the Arabidopsis genes are currently annotated by the GO system as PUF encoding genes (Table I). This is largely in agreement with the estimates from previous studies (Wortman et al., 2003; Gollery et al., 2006). Interestingly,

Table I. *Functional classification by gene ontologies*

The numbers of protein coding loci from Arabidopsis are given for custom categories of evidence codes of the three gene ontologies: MF, BP, and CC. A description of the evidence codes is available on the GO project site (http://www.geneontology.org/GO.evidence.shtml). The number of loci with annotations in any of the three ontologies (Any) are given in the last two rows. The percentage values are calculated relative to the total number of protein coding genes represented in the three ontologies. The evidence codes are grouped into the following custom categories of functional assignments: empirical data (IC, IDA, IGI, IMP, IPI, TAS), large-scale experiments (IEP, RCA, NAS, NR), sequence similarity or feature predictions (IEA, ISS), and PUFs lacking functional data (ND). The last column (Missing) accounts for genes that lack annotations within the listed ontologies.

	Empirical	Large Scale	Sequence	PUFs	Missing
MF	1,918	9,061	4,677	8,665	2,228
%	7	34	18	33	8
BP	3,731	4,777	4,462	10,194	3,385
%	14	18	17	38	13
CC	3,333	1,661	8,527	8,426	4,602
%	13	6	32	32	17
Any	5,837	11,005	12,851	14,071	0
%	22	42	48	53	0

only 7% of all entries are functionally characterized by traditional one-gene-at-a-time experiments in the MF ontology and 14% in the biological process (BP) ontology, whereas 34% and 18% have functional support from high-throughput experiments, respectively. This means that 93% of the genes from Arabidopsis code for poorly characterized proteins or PUFs when the most conservative empirical criteria are applied within the MF ontology. The relative amount of PUFs for the combined empirical and large-scale categories is 59% in the MF ontology and 68% in the BP ontology. The cellular component (CC) ontology contains by far the largest number of entries with sequence-based annotations and the lowest for the empirical categories. This trend is due to the majority of the CC annotations presently being based on computational ab initio predictions of subcellular localizations, whereas annotations with experimental support are much less frequent in this category than in the other two ontologies. The subsequent analysis steps of this study utilize the standard PUF set from the MF ontology containing 8,665 members. These genes are exclusively assigned to the root term of the MF ontology (GO:0003674) and they carry the evidence code ND (no [biological] data available). The MF category was selected here because protein functions are most profoundly described at the mechanistic molecular level, whereas the other two ontologies, BP and CC, provide rather indirect information in this regard.

To compare the results obtained from the MF ontology with alternative PUF identification methods, we also used one sequence similarity and one domain-based approach using hidden Markov models. First, all predicted Arabidopsis proteins were searched against the Swiss-Prot database with the BLASTP program (Altschul et al., 1990; Wu et al., 2006). Protein sequences that showed no similarities to functionally characterized proteins in the Swiss-Prot database were classified as PUFs using an expectation value (E value) of 10^{-6} as the cutoff. Second, the same protein set was used to search the Pfam database with the HMMPFAM program (Eddy, 1996; Bateman et al., 2004). Likewise, sequences without similarities to protein domains of known function (E value $\geq 10^{-2}$) or those matching exclusively domains of unknown function were considered PUFs. Due to different calculation methods, the E values of the two search algorithms are not directly comparable. Therefore, we chose for both methods conservative cutoff values that are commonly used for sensitive sequence similarity searching with low false positive detection rates (e.g. Girke et al., 2004; Horan et al., 2005; Gollery et al., 2006). Table II provides a comparison of the results from the three different PUF identification approaches. Based on the chosen confidence thresholds, all three approaches identified PUF sets of comparable sizes with 8,272 to 8,681 members, whereas 5,456 to 6,260 PUFs are common among two, and 4,667 among all three methods. The corresponding gene lists for the three methods are provided in Supplemental Data S1.

To simplify the description of the subsequent functional analysis steps of this study, the remaining text is restricted to the PUF set obtained from the MF ontology, whereas the data for the remaining PUF identification methods are included in the corresponding Supplemental Data S1, S3, S5, and S7. The GO PUF set was given preference because of the high quality of the manually curated GO annotation system and its broad acceptance in the scientific community.

Relative Amount of Expressed Genes

To functionally associate PUF with PKF encoding genes based on the similarity of their mRNA expression profiles, large-scale gene expression analysis of publicly

Table II. *PUF identification by different methods*

The table provides a matrix representation of the number of PUFs determined by the three different identification methods: BLASTP searches against the Swiss-Prot (SWP) database, HMMpfam searches against Pfam, and the GOMF approach from Table I. The amount of PUFs common between pairwise comparisons of methods is provided in the corresponding row and column intersects of the matrix. The numbers of PUFs identified by all three methods (All) or by at least one of them (Any) are given in the last two rows, respectively. The percentage values are calculated relative to the total number of protein coding genes. The complete gene lists for the PUF sets are available in Supplemental Data S1.

Method	SWP	Pfam	GOMF
SWP	8,681 (32%)	6,260 (23%)	5,456 (20%)
Pfam	–	8,272 (31%)	5,788 (21%)
GOMF	–	–	8,665 (32%)
All	4,667 (17%)	–	–
Any	12,781 (47%)	–	–

available Affymetrix GeneChip microarrays was performed. Only experiment sets containing at least two replicate samples were used for this analysis to enable statistical analysis of DEGs and to increase the confidence of the obtained results. In total, the study included the raw expression data from 1,310 Affymetrix chips from the AtGenExpress and Gene Expression Omnibus (GEO) sites (Schmid et al., 2005; Barrett et al., 2006). Table III provides a summary of the chosen experiment sets that covers a wide spectrum of treatment series and tissue samples. The complete list of the analyzed data is available in Supplemental Data S2.

The relative amount of expressed genes can be expected to be lower in the PUF than in the PKF category because many predicted PUF genes may be the result of genome annotation artifacts or may represent untranscribed pseudogenes. In addition, a certain fraction of PUF genes may be expressed below the detection limit of the GeneChip microarray technology. To estimate the extent of these limitations, the amount of detectable genes across all experiment categories was compared between the PUF and PKF sets. The present call information of the nonparametric Wilcoxon signed rank test of the MAS5 algorithm provides for this purpose relatively reliable estimates (Liu et al., 2002; Schmid et al., 2005; McClintick and Edenberg, 2006). According to this test, the amount of detectable genes between the PUF and PKF sets differs 0.5% to 8% within the five frequency intervals plotted in Figure 1. The detailed data set of this analysis is available in Supplemental Data S3. Based on these rather small relative differences, it is likely that the majority of the PUF genes are expressed at high enough levels to obtain for them meaningful data in the downstream cluster and differential gene expression analyses of this study.

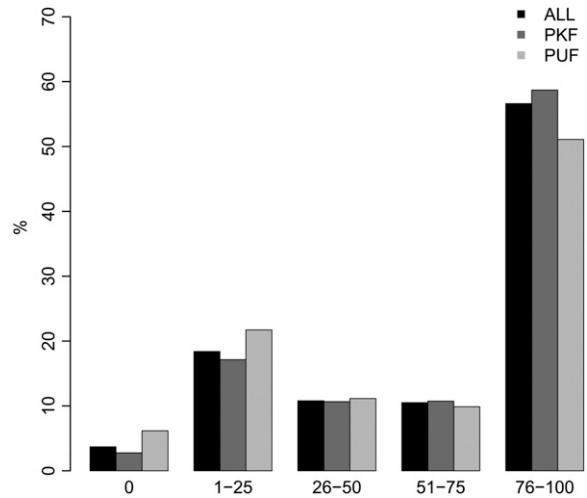

Figure 1. Relative amount of detectable genes. The relative amount of present calls is plotted for all genes (ALL), the PKF set and the PUF set using the five frequency intervals (bins): 0, 1 to 25, 26 to 50, 51 to 75, and 76% to 100% present calls. All experiment sets of this study were used for generating this plot. The complete present call data set for the individual experiment categories is available in Supplemental Data S3.

Cluster Analysis

Because many dynamic cellular processes are tightly associated with coordinated transcriptional changes, cluster analysis of gene expression profiles can be used to identify candidate sets of coregulated genes that are directly or indirectly involved in related processes (Steinhauser et al., 2004a; Gachon et al., 2005; Toufighi et al., 2005; Haberer et al., 2006; Jen et al., 2006; Vandepoele et al., 2006; Wei et al., 2006; Gutierrez et al., 2007). For instance, if a group of genes exhibits correlated expression profiles and it is significantly enriched in genes involved in a specific process then it is reasonable to assume that some of the PUF members of this cluster may share overlapping functions with its functionally characterized members. This association-based approach was applied here on a genome-wide level to systematically assign PUF to PKF genes based on the similarity of their expression profiles. Despite the great potential of this approach, it is important to keep in mind that correlation does not prove causal relationships. It only provides useful leads for establishing hypotheses and causal links in downstream investigations. Accordingly, the results of this study need to be interpreted as preliminary computer predictions that offer useful information for guiding future gene characterization experiments. Final evidence about gene and protein functions cannot be inferred directly from this data. Alternative network modeling approaches were not considered for this study because of the lack of efficient statistical methods to efficiently represent, score, and interpret the resulting network architectures on a genome-wide scale (e.g. Wolfe et al., 2005; Gutierrez et al., 2007; Ma et al., 2007). At this point, the traditional clustering approach appears to be more practical for the goals of this study.

To generate reliable and biologically relevant gene clusters form expression data, we evaluated several

Table III. *Analyzed gene expression arrays*

The table provides an overview of the different categories of GeneChip microarray experiments (first column) that were analyzed in this study. The following numeric columns contain the number of raw data (Cel) files, the amount of the corresponding biosamples (Samples), the number of performed comparisons in the DEG analysis (Comp), and the number of experiment sets (ExpSet) the raw data are derived from. A more detailed list of this data is available in Supplemental Data S2.

Category	Cel	Samples	Comp	ExpSet
Abiotic stress	524	254	129	10
Biotic stress	200	72	55	6
Chemical treatment	99	46	35	9
Tissue and development	237	79	40	1
Genotype	86	29	28	4
Hormone treatment	164	80	46	11
Sum	1310	560	333	41

available clustering algorithms (e.g. K-means, self-organizing maps) and selected agglomerative hierarchical clustering as the method of choice (Murtagh, 1985; Eisen et al., 1998; de Hoon et al., 2004; R Development Core Team, 2006). The hierarchical clustering method was chosen because of three main advantages: (1) the method requires no prior knowledge about the optimum number of the final clusters, (2) it is extremely robust in joining highly similar items into proper similarity groups, and (3) it provides an information-rich data output that represents the relative distances between all clustered items in a dendrogram (Becker et al., 1988). The main disadvantages of the approach are the complexity of its data output, the lack of predefined boundaries between clusters, and its weaker performance in identifying local expression similarities in a small subset of the samples (Prelic et al., 2006). However, most of these challenges can be overcome by applying efficient postprocessing methods of the obtained dendrograms, such as tree cutting methods (Gutierrez et al., 2007). Popular fuzzy clustering approaches (Krishnapuram et al., 2001) that allow memberships in several clusters—as opposed to strict clustering with unique memberships—were not considered for this study because it is difficult to efficiently prioritize and mine the complex cluster memberships from these methods in the downstream functional analysis steps. As an implementation of the hierarchical clustering algorithm, we used the hclust function (Murtagh, 1985) from the statistical programming environment R (R Development Core Team, 2006). As distance measurement we used correlation coefficients and as the cluster joining method complete linkage (see ''Material and Methods'' for more details). To obtain discrete clusters from the resulting dendrograms, we developed for this study a novel, to our knowledge, hierarchical threshold clustering (HTC) method. The corresponding R script is available in Supplemental Data S10. This method selects clusters in hierarchical clustering dendrograms based on a maximum tolerable distance between cluster members by applying an all-against-all distance test on all possible subtrees, while maintaining unique cluster memberships. As the threshold we chose for this step a minimum correlation coefficient of 0.6. This relatively conservative HTC setting ensures that all members of any given cluster share with all other members of the same cluster correlation coefficients between the selected cutoff of 0.6 and the highest possible value of 1.0. The exact cutoff value of 0.6 was chosen because it resulted in the highest enrichment of functionally related genes compared to alternative cutoff settings (Supplemental Data S4). Additionally, other gene expression correlation studies have used the same or very similar cutoff values (Haberer et al., 2006; Wei et al., 2006).

Applying the above strategy, we calculated four separate clustering data sets using both the Pearson correlation coefficients (PCC) and the Spearman correlation coefficients (SCC), in their signed and absolute forms as distance measures. The following text will refer to the four methods as PCC, SCC, PCCa, and SCCa, respectively (Supplemental Data S5). All four data sets were generated because of their complementary performance characteristics. The clustering with absolute correlation values allows the identification of positively and negatively correlated gene expressions, whereas the sign-specific approach joins only positively correlated items into similarity groups. The rank-based Spearman approach is limited to identifying global similarities in expression profiles, whereas the Pearson approach is very sensitive in detecting both global and local similarities. In particular, the latter detects local similarities with wide amplitude changes relative to the background, which can result in extreme cases in coclustering of outliers. A consensus approach between several or all methods was not considered because such a strategy would artificially deflate the cluster sizes and compromise the transparency of the results.

The distributions of the obtained numbers of clusters including their sizes from the four clustering methods are summarized in Figure 2. Because the sign removal increases the potential pool sizes of gene pairs with correlation values above a given cutoff, one would expect larger cluster sizes for the data sets with absolute correlation values compared to their signed counterparts. This trend can be observed in the many individual clusters in Supplemental Data S5, but the effect is not very pronounced in the global representation of Figure 2. These relative increases in cluster sizes are not as frequent as expected because of two main reasons. First, the number of highly negatively correlated gene

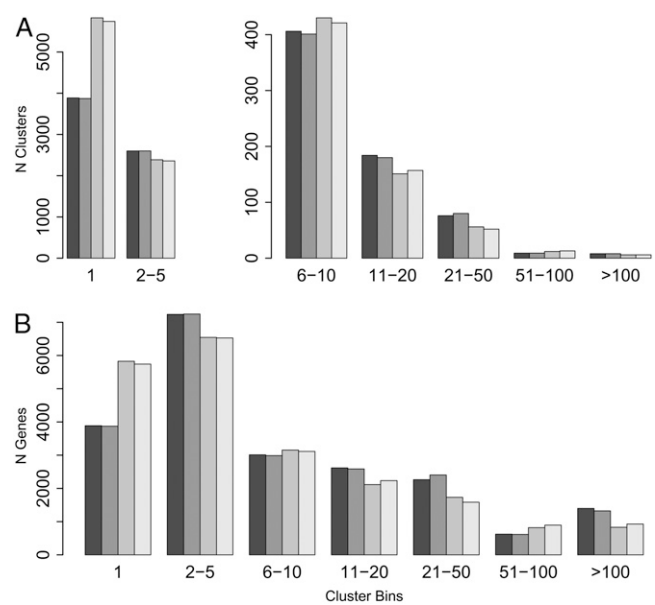

Figure 2. Cluster distributions. The numbers of clusters (A) and genes (B) are plotted for the cluster size intervals (bins) that are given along the abscissa. Each set of four bars, from left to right, contains the data for the clustering results using PCC, absolute PCC, SCC, and absolute SCC values as distance measures, respectively.

pairs is much smaller than the number of positively correlated gene pairs (data not shown; compare Haberer et al. [2006]). Second, the assignment of a negatively correlated gene to a cluster at an earlier stage of the hierarchical clustering process can prevent other potential members from joining the same cluster at a given cutoff level, if they do not share the required degree of correlation with the existing members. This is particularly the case in combination with a complete linkage joining method that was chosen for this study to minimize the number of false positive members in the generated clusters.

The most obvious differences among the four clustering data sets in Figure 2 are the numbers of singlet genes that do not join any clusters in the different methods. There are about 2,000 fewer singlet genes in the Pearson than in the Spearman data sets. This is expected because the latter method tends to generate slightly lower correlation values on gene expression data. Due to space restrictions, the subsequent text focuses on the clustering results from the distance method with the signed PCC, whereas the results for the other three methods are included in Supplemental Data S5. In addition, the clustering data for individual genes are available in the associated public database of this study (see below).

Functional Categorization of Gene Expression Clusters

Gene expression clusters with highly enriched functions provide more conclusive information about the potential roles of their PUF encoding members than clusters with very heterogeneous compositions. To functionally annotate the obtained clusters and to select the most informative gene sets with overrepresented gene functions, we performed enrichment analysis of GO terms using the hypergeometric distribution as a statistical test (Falcon and Gentleman, 2007). This method computes the enrichment test for all approximately 18,000 GO nodes of the three ontology networks and ranks the results by *P* values (see "Material and Methods"; Supplemental Data S9). The results of this method are more comprehensive and informative than generalized functional categorization systems, like GO slim or high-level pathway classification systems. Clusters with fewer than five members were excluded from this analysis because the predictive value of extremely small clusters is rather limited. The complete result set of this enrichment analysis is available in Supplemental Data S6. It contains the data for 916 clusters composed of a total of 11,077 genes. To prioritize the clusters based on the obtained enrichment data, we applied two selection filters. First, each cluster of interest needed to contain at least one overrepresented GO term in one of three ontologies (enrichment filter). Second, at least 20% of the cluster members had to be associated with this GO term to select clusters with relatively homogeneous compositions (uniformity filter). An overview of the number of clusters that meet these filter criteria is provided in Table IV. It contains the results for four different *P*-value cutoffs of the GO term enrichment filter ranging from 0.05 to 10^{-6}. The corresponding GO annotations for the prioritized cluster set that passed the most stringent selection criteria of 10^{-6} are listed in Table V. For space and readability reasons, the table presents only the highest ranking GO term for each of the three ontologies. The full set of GO annotations can be found in Supplemental Data S6. The following discussion of selected clusters is restricted to this most conservative data set (Table V). It contains 66 clusters with a total of 1,279 genes that include 277 PUF genes derived from 53 clusters (see Table IV). Our focus on these clusters does not indicate that the other clusters of this study are biologically less important. This selection is mainly based on the assumption that clusters with uniform GO annotations are particularly informative for functionally associating PUF with PKF genes.

Depending on the stringency of the applied prioritization filters listed in Table IV, our combined clustering and GO term enrichment strategy associated 277 to 1,541 PUF genes to overrepresented GO annotations. In comparison to the GO annotations currently available for these PUF genes, our method associated 216 to 1,050 of them to more specific GO terms in the MF category, 225 to 1,089 in the BP category, and 239 to 1,096 in the CC category (Supplemental Data S6). The large number of PUF genes associated with functionally informative annotations demonstrates the great potential of our approach for guiding future experimental studies on these genes.

Based on enrichment *P* values, the most highly overrepresented functional categories in the obtained cluster set are the BPs: ribosome assembly, photosynthesis pathways, and cell wall metabolism (Table V). This finding is largely in agreement with related gene

Table IV. *Overview of GO term enrichment analysis*

The amounts of clusters and genes are provided for different cluster prioritization filters that were applied to the GO term enrichment data of Supplemental Data S6. The values in parentheses represent the corresponding number of clusters containing PUF genes and the number of PUF genes in these clusters, respectively. The first row contains the counts for the unfiltered data set that considered only clusters with five or more members. The subsequent rows refer to the counts after applying the following two-component filter with four different stringency settings. (1) To select clusters with enriched GO terms, the clusters had to contain one or more overrepresented GO terms in at least one of the three ontologies based on the Bonferroni corrected *P* values of the enrichment analysis. The four different *P*-value cutoffs used for this filter are given in the first column. (2) In addition, ≥20% of the cluster members needed to be associated with the selected GO term to favor functionally homogeneous clusters.

Filter	Clusters	Genes
None	916 (794)	11,077 (2,884)
0.05	519 (429)	6,262 (1,541)
0.01	373 (301)	4,893 (1,126)
0.001	212 (170)	3,315 (744)
1E-06	66 (53)	1,279 (277)

Table V. *GO term enrichment data for prioritized clusters*

The GO annotations for the most conservative cluster prioritization filter from Table IV are provided. The three filtering criteria for selecting the presented clusters are described in the previous legend. Based on space and readability considerations, only the highest ranking GO term within each ontology is included here. As a result of our prioritization criteria, every cluster listed has at least one GO term assigned that meets both the enrichment (P value $\leq 10^{-6}$) and uniformity ($\geq$20%) criteria. If an ontology did not contain a GO term passing these filters then the candidate with the lowest P value was chosen. GO slim terms are used as table subtitles to organize the clusters based on a general BP classification schema. The different columns provide the identifiers of each cluster (CLID), the number of genes (CLSZ), the number of PUF genes, the number of genes matching a given GO term (Sample), the Bonferroni corrected P value of the hypergeometric distribution test (P value), the ontology type (Ont), and the corresponding GO term, respectively. The complete list of enriched GO terms and the associated gene identifiers for these clusters are available in Supplemental Data S6.

CLID	CLSZ	PUF	Sample	*P* Value	Ont	GO Term
Reproduction						
115	20	5	2	3.00E-06	BP	GO:0010344, seed oil body biogenesis
115	20	5	11	0.014	CC	GO:0016020, membrane
115		5	4	4.30E-07	MF	GO:0045735, nutrient reservoir activity
Carbohydrate metabolism						
95	21	4	4	5.90E-06	BP	GO:0006073, glucan metabolic process
95	21	4	8	1.80E-10	CC	GO:0005618, cell wall
95	21	4	4	1.90E-07	MF	GO:0005199, structural constituent of cell wall
131	18	1	6	1.10E-10	BP	GO:0006007, Glc catabolic process
131	18	1	7	1.80E-05	CC	GO:0005739, mitochondrion
131	18	1	2	1.30E-05	MF	GO:0004738, pyruvate dehydrogenase activity
248	11	2	3	9.30E-07	BP	GO:0005982, starch metabolic process
248	11	2	9	1.70E-07	CC	GO:0009507, chloroplast
248	11	2	5	0.003	MF	GO:0016740, transferase activity
300	11	3	3	1.70E-08	BP	GO:0005983, starch catabolic process
300	11	3	5	0.011	CC	GO:0044444, cytoplasmic part
300	11	3	2	0.0025	MF	GO:0016758, transferring hexosyl groups
548	7	0	3	2.30E-08	BP	GO:0006084, acetyl-CoA metabolic process
548	7	0	2	1.70E-06	CC	GO:0009346, citrate lyase complex
548	7	0	3	7.30E-09	MF	GO:0046912, transferring acyl groups
686	6	1	3	5.60E-08	BP	GO:0005982, starch metabolic process
686	6	1	2	0.0018	CC	GO:0005829, cytosol
686	6	1	2	3.10E-06	MF	GO:0001871, pattern binding
599	5	0	2	7.30E-06	BP	GO:0016138, glycoside biosynthetic process
599	5	0	2	0.19	CC	GO:0043231, intracellular membrane organelle
599	5	0	3	5.00E-07	MF	GO:0004497, monooxygenase activity
Nucleotide metabolism						
25	39	10	8	2.60E-07	BP	GO:0006259, DNA metabolic process
25	39	10	3	0.0045	CC	GO:0044427, chromosomal part
25	39	10	2	0.033	MF	GO:0003777, microtubule motor activity
29	37	5	13	2.10E-14	BP	GO:0006259, DNA metabolic process
29	37	5	6	5.30E-07	CC	GO:0005694, chromosome
29	37	5	15	1.20E-06	MF	GO:0003677, DNA binding
41	33	5	4	1.30E-05	BP	GO:0006399, tRNA metabolic process
41	33	5	21	1.80E-15	CC	GO:0009536, plastid
41	33	5	2	0.019	MF	GO:0004812, aminoacyl-tRNA ligase activity
Translation						
23	37	0	36	9.80E-45	BP	GO:0006412, translation
23	37	0	37	1.20E-64	CC	GO:0005840, ribosome
23	37	0	36	5.30E-65	MF	GO:0003735, structural constituent of ribosome
32	35	3	31	2.70E-35	BP	GO:0006412, translation
32	35	3	33	2.00E-50	CC	GO:0030529, ribonucleoprotein complex
32	35	3	31	2.40E-52	MF	GO:0003735, structural constituent of ribosome
37	36	11	8	0.00097	BP	GO:0006412, translation
37	36	11	11	5.50E-08	CC	GO:0005739, mitochondrion
37	36	11	4	0.00026	MF	GO:0008135, translation factor activity
39	34	1	29	7.70E-32	BP	GO:0006412, translation
39	34	1	29	4.60E-46	CC	GO:0005840, ribosome
39	34	1	29	3.10E-48	MF	GO:0003735, structural constituent of ribosome
182	11	0	9	4.50E-10	BP	GO:0006412, translation
182	11	0	10	1.80E-16	CC	GO:0005840, ribosome
182	11	0	10	7.80E-18	MF	GO:0003735, structural constituent of ribosome
239	13	0	8	2.50E-08	BP	GO:0006412, translation

(*Table continues on following page.*)

Table V. (*Continued from previous page.*)

CLID	CLSZ	PUF	Sample	*P* Value	Ont	GO Term
239	13	0	6	3.30E-08	CC	GO:0005840, ribosome
239	13	0	6	8.40E-08	MF	GO:0003735, structural constituent of ribosome
299	11	1	7	3.40E-07	BP	GO:0006412, translation
299	11	1	7	2.40E-11	CC	GO:0005840, ribosome
299	11	1	7	2.40E-11	MF	GO:0003735, structural constituent of ribosome
Lipid metabolism						
73	26	8	6	1.80E-14	BP	GO:0019915, sequestering of lipid
73	26	8	7	6.50E-09	CC	GO:0005576, extracellular region
73	26	8	4	1.60E-06	MF	GO:0045735, nutrient reservoir activity
279	10	2	2	9.30E-06	BP	GO:0019374, galactolipid metabolic process
279	10	2	2	0.01	CC	GO:0031967, organelle envelope
279	10	2	5	1.90E-07	MF	GO:0042578, phosphoric ester hydrolase activity
Transport						
47	34	8	2	0.00042	BP	GO:0045036, protein targeting to chloroplast
47	34	8	23	2.20E-17	CC	GO:0009536, plastid
47	34	8	8	1	MF	GO:0003674, MF (PUF term)
288	11	4	3	2.40E-07	BP	GO:0045036, protein targeting to chloroplast
288	11	4	4	1.10E-07	CC	GO:0009941, chloroplast envelope
288	11	4	2	0.016	MF	GO:0022804, transmembrane transporter activity
536	7	0	5	9.20E-06	BP	GO:0006810, transport
536	7	0	5	3.60E-09	CC	GO:0005794, Golgi apparatus
536	7	0	4	7.20E-05	MF	GO:0005215, transporter activity
708	6	1	3	4.70E-08	BP	GO:0006606, protein import into nucleus
708	6	1	3	6.40E-07	CC	GO:0005635, nuclear envelope
708	6	1	3	2.30E-06	MF	GO:0008565, protein transporter activity
765	5	1	2	3.20E-05	BP	GO:0006820, anion transport
765	5	1	2	2.10E-05	CC	GO:0005741, mitochondrial outer membrane
765	5	1	2	8.80E-07	MF	GO:0008308, voltage-gated ion channel activity
Biological process						
17	43	26	28	1.10E-08	BP	GO:0008150, BP (PUF term)
17	43	26	23	1.50E-05	CC	GO:0005575, CC (PUF term)
17	43	26	26	2.70E-09	MF	GO:0003674, MF (PUF term)
Photosynthesis						
4	134	43	28	2.20E-37	BP	GO:0015979, photosynthesis
4	134	43	67	1.30E-89	CC	GO:0044436, thylakoid part
4	134	43	2	0.00058	MF	GO:0010242, oxygen evolving activity
9	88	18	6	2.00E-05	BP	GO:0015979, photosynthesis
9	88	18	47	4.20E-30	CC	GO:0009507, chloroplast
9	88	18	2	7.00E-04	MF	GO:0004045, aminoacyl-tRNA hydrolase activity
45	32	5	7	2.90E-10	BP	GO:0015979, photosynthesis
45	32	5	21	3.10E-16	CC	GO:0009507, chloroplast
45	32	5	5	1	MF	GO:0003674, MF (PUF term)
110	20	6	3	8.00E-05	BP	GO:0015979, photosynthesis
110	20	6	12	9.00E-10	CC	GO:0009507, chloroplast
110	20	6	2	0.00093	MF	GO:0004176, ATP-dependent peptidase activity
304	9	2	7	1.60E-15	BP	GO:0015979, photosynthesis
304	9	2	5	1.90E-11	CC	GO:0009523, PSII
304	9	2	3	5.10E-07	MF	GO:0046906, tetrapyrrole binding
428	8	2	2	0.024	BP	GO:0006091, generation of metabolites and energy
428	8	2	6	5.10E-07	CC	GO:0005739, mitochondrion
428	8	2	2	1.80E-06	MF	GO:0004449, isocitrate dehydrogenase activity
555	5	1	3	1.20E-06	BP	GO:0015979, photosynthesis
555	5	1	3	3.80E-10	CC	GO:0009502, photosynthetic electron transport chain
555	5	1	3	3.30E-06	MF	GO:0009055, electron carrier activity
923	5	2	3	5.20E-08	BP	GO:0009853, photorespiration
923	5	2	3	3.90E-08	CC	GO:0030964, NADH dehydrogenase complex
923	5	2	2	0.0047	MF	GO:0003735, structural constituent of ribosome
Cell organization and biogenesis						
77	24	5	6	4.20E-15	BP	GO:0009834, secondary cell wall biogenesis
77	24	5	3	0.011	CC	GO:0031225, anchored to membrane
77	24	5	6	1.70E-05	MF	GO:0016757, transferring glycosyl groups

(*Table continues on following page.*)

Table V. (*Continued from previous page.*)

CLID	CLSZ	PUF	Sample	*P* Value	Ont	GO Term
108	18	7	2	0.00084	BP	GO:0009831, cellulose and pectin modification
108	18	7	18	1.30E-09	CC	GO:0016020, membrane
108	18	7	2	0.0076	MF	GO:0008289, lipid binding
349	9	0	7	1.00E-15	BP	GO:0009664, cellulose and pectin biogenesis
349	9	0	6	0.00028	CC	GO:0012505, endomembrane system
349	9	0	7	7.60E-19	MF	GO:0005199, structural constituent of cell wall
953	5	1	2	9.70E-07	BP	GO:0010020, chloroplast fission
953	5	1	2	0.029	CC	GO:0009507, chloroplast
953	5	1	4	0.044	MF	GO:0005488, binding
Secondary metabolism						
12	73	13	3	0.0076	BP	GO:0046148, pigment biosynthetic process
12	73	13	34	1.30E-18	CC	GO:0009536, plastid
12	73	13	3	0.00059	MF	GO:0003746, translation elongation factor activity
143	17	2	4	8.70E-08	BP	GO:0046148, pigment biosynthetic process
143	17	2	8	1.40E-05	CC	GO:0009536, plastid
143	17	2	5	0.0023	MF	GO:0016491, oxidoreductase activity
347	10	2	3	1.20E-07	BP	GO:0009686, GA biosynthetic process
347	10	2	4	0.95	CC	GO:0005575, CC (PUF term)
347	10	2	5	1.80E-10	MF	GO:0016706, oxidoreductase activity
432	8	1	5	6.00E-13	BP	GO:0009813, flavonoid biosynthetic process
432	8	1	2	0.00017	CC	GO:0009705, membrane of vacuole
432	8	1	2	0.00023	MF	GO:0016706, oxidoreductase activity
600	5	0	2	9.30E-07	BP	GO:0009718, anthocyanin biosynthetic process
600	5	0	2	0.63	CC	GO:0005575, CC (PUF term)
600	5	0	4	0.00049	MF	GO:0016740, transferase activity
Response to stimulus						
68	22	3	7	5.40E-07	BP	GO:0006952, defense response
68	22	3	11	0.02	CC	GO:0016020, membrane
68	22	3	5	1.10E-06	MF	GO:0004888, transmembrane receptor activity
85	23	9	10	6.70E-18	BP	GO:0009408, response to heat
85	23	9	10	0.21	CC	GO:0005575, CC (PUF term)
85	23	9	2	0.043	MF	GO:0005516, calmodulin binding
90	22	5	6	1.60E-09	BP	GO:0009408, response to heat
90	22	5	5	0.04	CC	GO:0005634, nucleus
90	22	5	3	0.00052	MF	GO:0051082, unfolded protein binding
346	10	3	2	0.03	BP	GO:0009628, response to abiotic stimulus
346	10	3	9	6.20E-08	CC	GO:0009536, plastid
346	10	3	3	0.57	MF	GO:0003674, MF (PUF term)
356	9	9	8	1.10E-15	BP	GO:0009733, response to auxin stimulus
356	9	9	3	0.021	CC	GO:0043231, intracellular membrane-bound organelle
356	9	9	9	7.70E-05	MF	GO:0003674, MF (PUF term)
480	8	1	3	3.80E-05	BP	GO:0006979, response to oxidative stress
480	8	1	7	1.10E-08	CC	GO:0005739, mitochondrion
480	8	1	2	7.20E-05	MF	GO:0046933, hydrogen ion transporting ATP synthase
586	7	1	3	4.20E-05	BP	GO:0009737, response to abscisic acid stimulus
586	7	1	2	0.00013	CC	GO:0008287, Ser/Thr phosphatase complex
586	7	1	3	5.50E-07	MF	GO:0015071, protein phosphatase type 2C activity
748	5	0	3	6.50E-08	BP	GO:0009404, toxin metabolic process
748	5	0	4	0.01	CC	GO:0005737, cytoplasm
748	5	0	3	6.90E-08	MF	GO:0004364, glutathione transferase activity
912	5	0	4	7.30E-08	BP	GO:0006457, protein folding
912	5	0	3	1.80E-06	CC	GO:0009532, plastid stroma
912	5	0	3	1.20E-06	MF	GO:0051082, unfolded protein binding
Physiological process						
36	34	6	15	0.0011	BP	GO:0043170, macromolecule metabolic process
36	34	6	11	2.00E-08	CC	GO:0043228, nonmembrane-bound organelle
36	34	6	7	3.50E-06	MF	GO:0003735, structural constituent of ribosome
81	24	8	3	0.0011	BP	GO:0051188, cofactor biosynthetic process
81	24	8	14	6.70E-08	CC	GO:0009536, plastid
81	24	8	8	1	MF	GO:0003674, MF (PUF term)
130	15	1	3	3.80E-07	BP	GO:0010119, regulation of stomatal movement

(*Table continues on following page.*)

Table V. (*Continued from previous page.*)

CLID	CLSZ	PUF	Sample	*P* Value	Ont	GO Term
130	15	1	2	1	CC	GO:0005575, CC (PUF term)
130	15	1	5	0.041	MF	GO:0016787, hydrolase activity
134	17	7	12	1.10E-20	BP	GO:0006511, ubiquitin-dependent catabolic process
134	17	7	12	1.70E-28	CC	GO:0000502, proteasome complex
134	17	7	7	6.50E-08	MF	GO:0008233, peptidase activity
199	13	1	9	3.90E-07	BP	GO:0009058, biosynthetic process
199	13	1	6	5.60E-12	CC	GO:0044445, cytosolic part
199	13	1	6	2.10E-08	MF	GO:0003735, structural constituent of ribosome
224	12	3	2	0.045	BP	GO:0044249, cellular biosynthetic process
224	12	3	8	1.00E-07	CC	GO:0009507, chloroplast
224	12	3	2	0.043	MF	GO:0003723, RNA binding
293	11	3	2	0.00012	BP	GO:0042775, ATP synthesis coupled electron transport
293	11	3	9	3.80E-11	CC	GO:0005739, mitochondrion
293	11	3	3	2.10E-05	MF	GO:0015078, hydrogen ion transmembrane transporter
366	8	1	3	7.90E-06	BP	GO:0006457, protein folding
366	8	1	6	4.50E-10	CC	GO:0005783, endoplasmic reticulum
366	8	1	2	0.0016	MF	GO:0031072, heat shock protein binding
406	9	1	4	1.00E-06	BP	GO:0006511, ubiquitin-dependent protein catabolic
406	9	1	4	6.20E-09	CC	GO:0000502, proteasome complex
406	9	1	3	0.00063	MF	GO:0008233, peptidase activity
520	7	0	2	3.00E-06	BP	GO:0006121, mitochondrial electron transport
520	7	0	2	3.40E-06	CC	GO:0045273, respiratory chain complex II
520	7	0	3	4.90E-07	MF	GO:0016627, oxidoreductase for CH-CH groups
728	6	0	6	5.50E-12	CC	GO:0005783, endoplasmic reticulum
728	6	0	2	0.0051	MF	GO:0008233, peptidase activity
790	5	1	3	8.60E-06	BP	GO:0006511, ubiquitin-dependent catabolic process
790	5	1	3	1.10E-08	CC	GO:0005839, proteasome core complex
790	5	1	3	0.00012	MF	GO:0008233, peptidase activity
895	5	0	5	0.0099	BP	GO:0008152, metabolic process
895	5	0	5	2.20E-07	CC	GO:0005739, mitochondrion
895	5	0	2	9.50E-08	MF	GO:0004774, succinate-CoA ligase activity
943	5	2	2	0.029	BP	GO:0009058, biosynthetic process
943	5	2	4	4.70E-07	CC	GO:0005783, endoplasmic reticulum
943	5	2	2	0.44	MF	GO:0003674, MF (PUF term)

coregulation studies in Arabidopsis (Haberer et al., 2006; Wei et al., 2006). With regard to ribosome assembly, 124 of the 410 GO annotated genes for cytosolic, plastidial, and mitochondrial ribosome components appear in seven clusters (see Table V; cluster identifiers 23, 32, 37, 39, 182, 239, and 299); 272 ribosomal genes appear in clusters with five or more members of the nonprioritized data set. Although cluster 23 consists exclusively of genes annotated as ribosomal genes (GO:0005840; *P* value 1.2×10^{-64}), the other six clusters are highly enriched in ribosomal genes and they contain among others 16 PUF genes. Equally interesting is the observation that photosynthesis-related annotations are highly overrepresented in five large clusters (cluster identifiers 4, 9, 45, 110, and 304). These clusters represent 51 of all the 121 genes that are currently annotated by the GO system as photosynthesis components (GO:0015979). Because both processes, photosynthesis as well as ribosomal activities, require the coordinated assembly of many proteins to large complexes and protein-protein interaction networks, it is not unexpected that their corresponding genes are tightly coregulated. In alignment with the association hypothesis of this study, several of the PUF members in these functionally extremely uniform clusters may be involved in processes that are connected to the enzymatic or regulatory networks of photosynthesis and ribosomal activities.

Interestingly, our method also identified a cluster (identifier 77) that is highly enriched in cell wall-related annotations (e.g. GO:0009834; *P* value 4.2×10^{-15}), such as cellulase synthase genes. A very similar cluster of genes was recently described and experimentally verified by two groups (Brown et al., 2005; Persson et al., 2005) who specifically mined public expression data for genes that are coregulated with the cellulose synthase genes *CESA4*, *CESA7*, and *CESA8*. In addition, comparable results were described by Jen et al. (2006). This example demonstrates that our genome-wide expression clustering approach generates biologically meaningful data. An additional interesting cell wall-related cluster is cluster 349 that contains eight genes for Pro-rich extensin domain proteins.

The majority of the clusters in our data set contain one or more PUF genes (Table IV), but only a few of the larger clusters consist predominantly of PUF genes.

Cluster 17 represents an exception to this rule. The 43 members of this cluster contain 26 PUF genes, and its characterized members show no clear enrichment of specific functions. Based on the high abundance of PUF genes in the entire data set (approximately 32%), PUF-gene-enriched clusters occur much less frequent than those enriched in PKF genes; clusters consisting exclusively of PUF genes are entirely absent (Table IV). One explanation for this difference could be that the expression of most PUF genes is controlled by the same regulatory networks as many PKF genes. If this is the case, PUF genes are more likely to appear in expression clusters together with PKF genes than without them.

Our method also identified clusters that are enriched in abiotic stress-response annotations. For instance, clusters 85 and 912 are highly enriched in heat stress-related genes (GO:0009408; P values 6.7×10^{-18} and 1.8×10^{-6}). Interestingly, 10 of the 23 members in the cluster 85 were identified by the subsequent DEG analysis of this study as genes that respond specifically to heat stress and to a much lesser extent to other types of abiotic stresses (see Supplemental Data S7). Based on the available coexpression data, the nine PUF genes of this cluster are now excellent candidates for discovering novel gene functions involved in heat stress-response pathways. Additionally, this example illustrates that the two chosen approaches of this study, expression clustering and DEG analysis, complement and confirm each other. The hypoxia cluster 203 is another interesting abiotic stress cluster (Supplemental Data S6; Fukao and Bailey-Serres, 2004). This cluster does not appear in the most stringently prioritized data set (Table V), because it did not pass the applied uniformity filter. Nevertheless, it is enriched in hypoxia-responsive genes (cluster identifier 203; GO:0001666; P value 2.0×10^{-5}), and it contains several members that are involved in cellular respiration processes, such as genes for the alcohol dehydrogenase ADH1 (AT1G77120), a pyruvate dehydrogenase (AT4G33070), and a hemoglobin-like oxygen binding protein that affects ATP levels under hypoxia (AT2G16060; Hebelstrup et al., 2007). Whether the five PUF genes of this cluster are also involved in hypoxia-response processes can be addressed in experimental studies.

In conclusion, the combined clustering and gene function enrichment strategy allowed us to associate a considerable fraction of the PUF encoding gene pool with tightly coexpressed gene sets of known function. Depending on the chosen stringency settings, the approach allowed us to assign 277 to 1,541 PUF genes (Table IV) to more specific GO terms than those available in the latest GO annotation release for Arabidopsis.

Analysis of DEGs

DEG analysis can identify groups of genes that exhibit expression changes in response to specific treatments or cellular changes. Because this information is not easily obtainable from clustering of global expression profiles, DEG analysis of publicly available expression data complements the previous approach by associating PUF with PKF encoding genes based on common differential expression responses to environmental changes, such as abiotic stresses. If a group of genes shares similar expression patterns across a wide spectrum of treatments then it is likely that certain members are involved in similar or connected response pathways to these perturbations. The association of genes with these response mechanisms can provide valuable information for future functional characterization experiments of PUF or PKF genes.

One of the main challenges of performing systematic DEG analyses on large and diverse gene expression data sets from public sources is the identification of the given design parameters to determine for each experiment set its biologically most meaningful analysis strategy. This step is extremely crucial because every analysis needs to focus on the specific treatment factors of an experiment. The alternative of performing simply all possible comparisons will provide meaningless results for many experimental designs because it would generate a large number of illegitimate contrasts between biologically incomparable samples. To define reasonable analysis strategies for public GeneChip microarray expression data sets, all their replicates and the most useful sample comparisons need to be determined manually to provide the proper experimental design parameters to the downstream statistical methods for identifying DEGs. The MIAME and MGED ontology annotations (Brazma et al., 2001; Whetzel et al., 2006) of the public microarray depositories provide the essential information about the experiments, but efficient facilities to completely automate the DEG analyses on a large scale are not available at this point.

To perform large-scale DEG analysis of public expression data, we chose for this study a human-supervised analysis strategy, in which we determined for each experiment set its optimum analysis parameters. The goal of this analysis was to identify all PUF and PKF genes that respond to specific or a wide range of conditions by enumerating their significant expression modulations in the corresponding experiment classes. For this, the available experiment annotations were manually evaluated and the most reasonable set of sample comparisons were recorded in an experiment definition table that contained all the required input parameters to control the downstream statistical DEG analysis in an automated manner (Supplemental Data S2). Typically, we chose for each experiment set a design strategy that focused the analysis on the primary treatment as the main experimental factor. Multifactorial analysis strategies were avoided as much as possible. For instance, when an experiment contained a stress treatment as the primary experimental factor and time or different tissue types as secondary factors, then we compared only samples from identical tissues that were collected at the same time points. Additionally, comparisons between different experiment sets

were not considered to exclude unknown variables, such as sample handling differences between laboratories (Hong et al., 2006). It is important to stress here that, depending on the design of a given experiment and its available annotations, it is often difficult to select a single most meaningful analysis strategy. Thus, our chosen strategy may not provide a perfect solution for every experiment set, but it represents a practical and reasonable compromise for performing systematic DEG analyses on large expression data sets from public databases.

In total our large-scale DEG analysis survey included 333 comparisons between samples with two to four technical or biological replicates from 41 experiment sets of six experiment categories. Table III provides an overview of the corresponding sample and experiment sets, and Supplemental Data S2 contains all detailed information including the chosen analysis strategies for these data sets. Because the abiotic stress category is by far the largest data set, containing 524 chip hybridizations of 254 biosamples (Kilian et al., 2007), the following description of our DEG results will be restricted to this most comprehensive treatment category (Table VI). The data for the other categories are provided in the online database of this project (see below). As the statistical method for identifying DEGs with the determined experiment analyses strategies, we used linear models for microarray data (LIMMA) from Smyth (2004, 2005), using in all cases as the confidence threshold a false discovery rate (FDR) of ≤ 0.01 in combination with a minimum fold-change filter of 2.

Applying the above DEG analysis strategy, we were able to identify 269 PKF and 104 PUF genes that showed expression changes in the majority of the 10 considered abiotic stress categories (Fig. 3; Supplemental Data S7). This set of a total of 373 generic stress DEGs was determined by filtering the generated DEG data set for members that showed one or more significant expression changes in at least 80% of all stress categories. Interestingly, 95% of these DEGs also appear in the generated gene expression clusters of the previous analysis (Supplemental Data S5). The subsequent GO term enrichment analysis revealed that stress-related annotations are highly overrepresented in this group of DEGs (see Supplemental Data S8). About 48 of its members (13%) are associated with the GO term "response to stress" from the BP ontology (GO:0006950; P value 2.0×10^{-13}). This enrichment indicates that our strategy has a high selectivity for identifying stress-response genes. Therefore, many PUF encoding genes in this data set may be directly or indirectly involved in generic stress-response pathways. Among the different groups of identified stress responsive genes (see below; Fig. 3), the generic stress DEG set represents by far the largest group.

Similarly, other studies have shown that stress-regulated genes frequently exhibit expression changes to a wide range of different abiotic stress treatments rather than a refined subset of stresses (Rodriguez and Redman, 2005; Kilian et al., 2007). The group of generic stress DEGs contains 48 genes that are annotated as transcription regulators in the MF ontology (GO:0030528; P value 2.5×10^{-3}; Supplemental Data S8). This enrichment emphasizes the central role of transcription factors for the control of many stress-response pathways. Moreover, it opens the possibility that several of the 104 PUF genes of this data set may be involved in similar transcription control processes.

We also used the generated abiotic stress DEG data set for identifying genes that respond predominantly to a specific type of stress. These specific stress DEGs were defined as follows. Firstly, they had to show in 25% of all comparisons of a given stress type significant changes. Secondly, they had to exhibit at the same time at least four times as many changes than in the other nine stresses (Fig. 3; Supplemental Data S7). This frequency-based filtering approach appeared to be more efficient for associating DEGs with specific stresses than overly strict filtering methods. This is the case because most stress-response genes are not highly specific for a single type of stress (Kilian et al., 2007). As a result, strict filtering for genes responding only to a single stress will fail to identify any candidate genes in our comprehensive data sets. It is important to emphasize here that the chosen filtering approach is a practical compromise, but not a perfect solution to the problem of assigning DEGs reliably to different stress types. Therefore, the complete DEG results are provided in Supplemental Data S7 where users can apply their own custom filters and prioritize strategies.

With the chosen frequency filter we were able to identify specific stress DEG sets within six of the 10 treatment types (Table VI; Fig. 3). The data sets for the stress treatments—light, oxidative, and wounding stress—did not contain any genes that meet our filtering criteria, and the drought data set contained only a single member. The lack of specific stress DEGs in these data sets indicates that the genome-wide expression response patterns to these four stresses widely overlap

Table VI. *Abiotic stress treatments*

The table provides an overview of the different types of abiotic stress experiment sets (Stress) that were used in the DEG analysis of this study. The numeric columns contain the number of the analyzed GeneChip microarrays (Chips), the number of the corresponding biosamples (Samples), and the number of the performed comparisons (Comp). A more detailed list of this data is available in Supplemental Data S2.

Stress	Chips	Samples	Comp
Heat	68	34	17
Cold	48	24	12
Osmotic	48	24	12
Salt	48	24	12
Drought	56	28	14
Oxidative	48	24	12
Wounding	56	28	14
UV-B	56	28	14
Light	48	16	10
Genotoxic	48	24	12

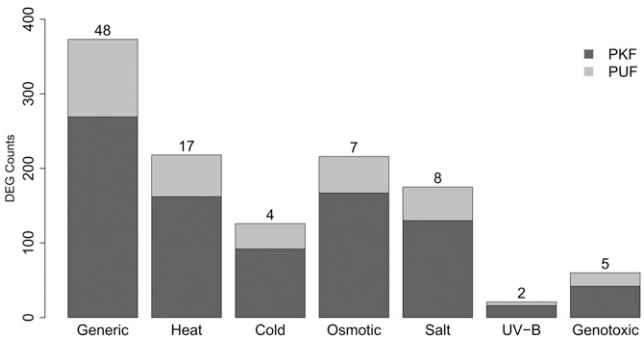

Figure 3. Generic and specific stress DEGs. The number of PUF and PKF encoding genes are plotted that were identified as generic and specific stress DEGs. The values above the bars provide the corresponding numbers of genes that are currently annotated with the GO term "response to stress" (GO:0006950 in BP ontology). The different stress types are given along the abscissa. Genes responding to the majority of the 10 abiotic stresses were considered as generic stress DEGs (Generic), whereas those responding predominantly to a specific type of stress were classified as specific stress DEGs. The following filters were used for assigning genes to the two stress categories. (1) Generic stress-response genes are those that showed in at least 80% of all stress treatments one or more significant changes. (2) Specific stress-response genes are those that showed in ≥25% of all comparisons of a given stress significant changes, and exhibited there ≥4 times as many changes than in the other nine stresses. For both filters, the observed expression changes were only counted when they met our confidence criteria of a FDR ≤0.01 and a fold change ≥2. The specific stress data for the four treatment sets—light, oxidative, drought, and wounding—are not plotted here because their data sets did not contain any DEGs that met our specific stress criteria.

with those from other stresses. For the remaining six treatment categories we identified in total 608 PKF and 206 PUF genes that responded predominantly to single stresses. The functional analysis of these specific stress DEG sets with our GO term enrichment approach showed no outstanding enrichment of specific gene functions. Instead, the results contained mostly moderately enriched GO annotations from a wide spectrum of molecular processes and BPs (Supplemental Data S8). Similar to the generic stress data, the different groups of specific stress DEGs included various marker genes that are characteristic for stress-related gene sets. For instance, they contained many genes that are annotated with the GO term "response to stress" (see Fig. 3). This term is significantly enriched in the heat stress data set (P value 1.3×10^{-2}), whereas the other five treatment sets contain it with considerable, but not significantly enriched frequencies (P values of $\geq 5 \times 10^{-2}$). In addition, the heat stress and genotoxic stress data sets showed the expected enrichment of genes that are associated with heat response and DNA repair processes, respectively (GO:0009408; P value 4.9×10^{-3} and GO:0006281; P value 6.1×10^{-5}).

In summary, the above large-scale DEG study identified a comprehensive set of candidate PKF and PUF genes that are involved in generic and specific stress-response pathways. These results suggest the existence of one or more abiotic stress-response regulons in Arabidopsis similar to the environmental stress regulon described in yeast (*Saccharomyces cerevisiae*; Gasch et al., 2000; Gasch, 2002). Furthermore, the generated data sets represent an important resource for other scientists who are interested in addressing more specific questions relevant to abiotic stress research by querying the generated DEG information in alternative ways (see Supplemental Data S7; online database).

Plant Unknown-eome and Gene Expression Databases

To provide efficient access to the extensive data sets of this study, we have developed two publicly available online portals: the Plant Unknown-eome Database (POND; http://bioweb.ucr.edu/scripts/unknownsDisplay.pl) and the PED (http://bioweb.ucr.edu/PED). The POND interface provides query and download options for the latest PUF sets from Arabidopsis. Their predictions are based on the three search methods used for this study: (1) BLASTP searches against the PKFs from Swiss-Prot, (2) HMM searches against the Pfam domain database, and (3) retrieval of the unknown annotations from the GO system (MF).

The PED integrates our diverse coexpression data with a variety of online tools for user-friendly DEG analysis, cluster visualization, and data mining (Fig. 4). The aim of this service is not to duplicate or compete with the excellent Web resources that are already available for array-based expression data from plants, such as GEO, Genevestigator, BAR, AtGenExpress, ATC, Page-Man, CSB.DB, and MetNet (Steinhauser et al., 2004b; Zimmermann et al., 2004, 2005; Schmid et al., 2005; Toufighi et al., 2005; Yang et al., 2005; Barrett et al., 2006; Grennan, 2006; Jen et al., 2006; Usadel et al., 2006). Instead PED complements the available resources by providing a subset of the publicly available Affymetrix expression data from Arabidopsis in preanalyzed form using various statistical methods for DEG identification combined with expression cluster information for coregulation analysis. To provide high-confidence data, the database is restricted to data sets with two or more replicates. The following text provides a brief overview of the most interesting features of the database.

All expression data in PED were normalized with the RMA and MAS5 algorithms (Irizarry et al., 2003; Qin et al., 2006). The incorporation of the expression values from both normalization methods increases the utility spectrum of the provided data sets. The quantile-based RMA method generates more accurate expression measures for weakly expressed genes, whereas the MAS5 scaling approach is more appropriate for comparisons between expression studies (Lim et al., 2007). The option to identify DEGs by statistical modeling is a very unique feature of this online service. For this, PED provides the results of experiment design-based expression changes from several statistical methods, such as LIMMA (Smyth, 2004, 2005). The corresponding

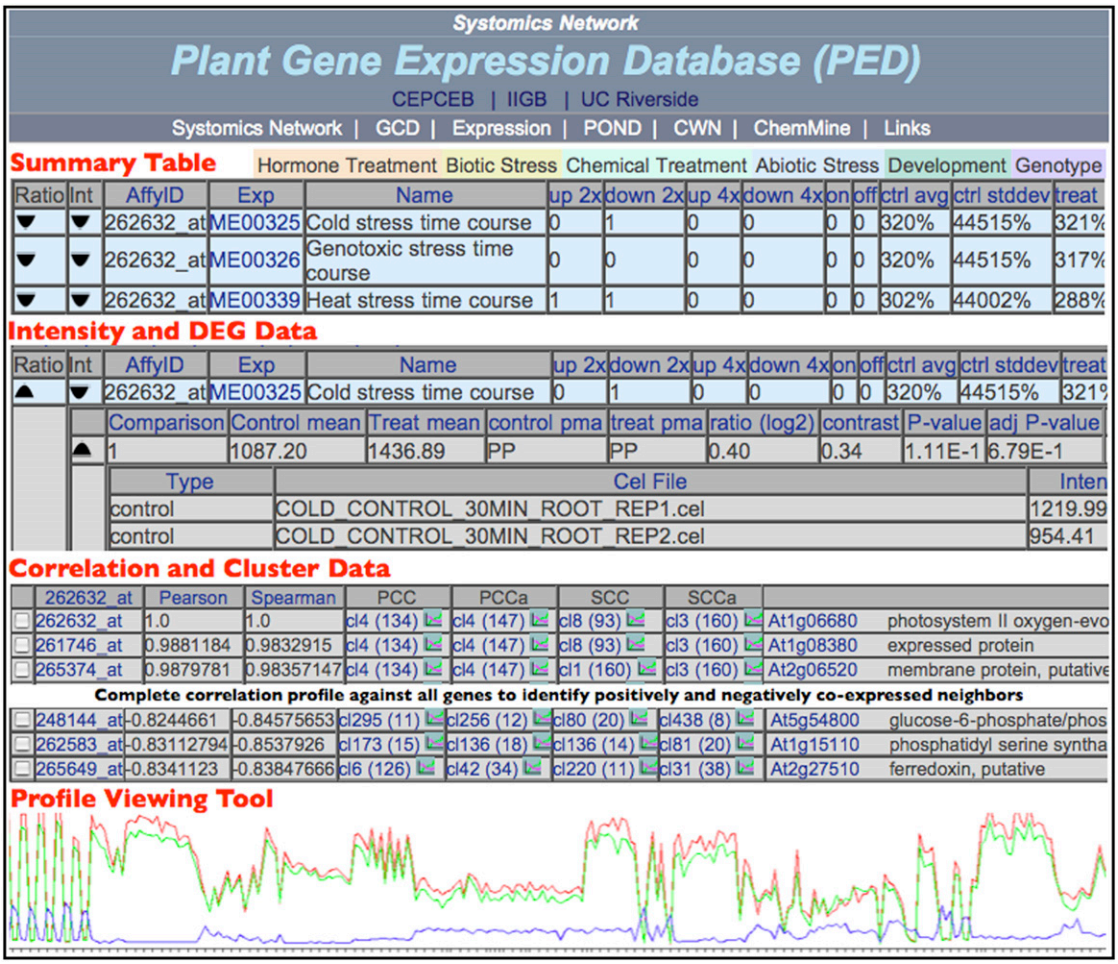

Figure 4. The PED database. The outline illustrates important utilities of the database (http://bioweb.ucr.edu/PED).

experiment analysis strategies are available for online viewing and download. A combinatorial query page allows searching for DEGs by specific treatments and filtering by various quantitative values to obtain candidate gene lists with strategies that resemble typical microarray analysis routines. Furthermore, the expression intensity and DEG data in PED are fully integrated with a comprehensive set of gene coexpression data from correlation and cluster analyses. To identify for a gene of interest its most positively or negatively coregulated neighbors, the interface contains a correlation tool that provides for every gene on the arrays the Pearson and Spearman correlation profiles against all other genes. Information on discrete expression clusters is combined with the correlation data. It contains the four separate HTC cluster data sets that were generated by this study using as distance measures the two correlation coefficients in their signed and absolute forms (see previous section). An expression profile plotting tool is available for evaluating the quality of expression clusters or visualizing the expression patterns for custom gene sets across all samples in the database. This utility offers convenient options for inspecting the vast number of expression clusters of this study efficiently. Extensive download options for imports into local spreadsheet programs are available on all query levels for intensity, DEG, correlation, and cluster data.

While the backend of the database is based on PostgreSQL and the web interface is implemented in Java, the framework of data analysis and online tools is largely designed around R and BioConductor utilities (Gentleman et al., 2005; R Development Core Team, 2006). The latter design feature will allow us to routinely add to PED's online services in the future additional useful tools from the wide spectrum of statistical data analysis packages that are provided by the R open source community.

CONCLUSION

We present here one of the most comprehensive gene coregulation studies that are currently available for Arabidopsis. Our study is unique, to our knowledge, by focusing on the analysis on PUF genes and their

systematic association with functional annotations of PKF genes. By applying a combination of genome-wide cluster and DEG analysis methods, we identified many interesting groups of potentially coregulated genes from a wide range of BPs and stress-response pathways. This approach allowed us to assign 1,541 PUF genes to relative specific and functionally informative GO terms. These gene associations provide a valuable resource for guiding future functional characterization experiments of PUF and PKF genes. In addition, the developed large-scale expression data analysis methods and the associated database represent important components of a future open-source framework for other scientists who are interested in performing similar studies or utilizing public gene expression resources more efficiently. Finally, users of the provided data sets should keep two limitations in mind. First, the generated associations are hypotheses and not final proofs of gene functions. Second, even the most careful statistical approaches for large-scale data can only reduce, but not fully eliminate, errors in the decision-making processes associated with the interpretation of microarray data.

MATERIALS AND METHODS

Sequence Similarity and Domain Searches

Sequence similarity searches of the Arabidopsis (*Arabidopsis thaliana*) proteome against the SwissProt database were performed with the BLASTP program (Altschul et al., 1997) using an E value of 1×10^{-6} as the cutoff and the default settings for the remaining parameters. The Arabidopsis protein sequences were obtained from The Arabidopsis Information Resource (TAIR) site (version 7 release; ftp://ftp.arabidopsis.org/home/tair/Sequences), and the SwissProt sequences (Wu et al., 2006) were downloaded from the ExPASy site (release 54.4; ftp://ftp.expasy.org/databases/uniprot). To query only the functionally characterized protein space, all entries annotated as sequences of unknown function were removed from the SwissProt data set.

To identify protein domains of known function in the above Arabidopsis proteins, domain searches against the hidden Markov models of the Pfam database (Bateman et al., 2004) were performed with the HMMPFAM program (Eddy, 1996) using an E value of 1×10^{-2} as the cutoff. The global models of the Pfam release 22 were used for these searches (ftp://ftp.sanger.ac.uk/pub/databases/Pfam/). Matches against domains of unknown function were ignored in the postprocessing of the search results to identify only candidate sequences with domains of known functions.

GO Analysis

The Arabidopsis gene-to-GO mappings from TAIR/The Institute for Genomic Research were used for all GO analysis steps of this study. They were downloaded from the GO site (10/12/2007 release; http://geneontology.org). Direct assignments to the root node of each ontology were considered as unknown function annotations. These root assignments, in combination with the evidence code ND, are the new official GO terms for sequences of unknown function. The former terms, MF unknown (GO:0005554), BP unknown (GO:0000004), and CC unknown (GO:0008372), were discontinued by the consortium on 10/17/2006. In the subsequent GO term enrichment analysis steps, the new unknown annotations to the root were considered as artificial terminal annotations. This was necessary because the root node is connected with all other genes in the GO network, which makes it impossible to obtain for the new unknown annotations meaningful enrichment data with most GO analysis approaches. This modification does not affect the results for any of the other GO nodes.

The hypergeometric distribution was used to test gene sets for the overrepresentation of GO terms. To perform this test, we developed a set of modular functions using the R language for statistical computing for their implementation (R Development Core Team, 2006). The corresponding GOHyperGAll script computes for a given sample population of genes the enrichment test for all nodes in the GO network, and returns raw and adjusted P values. As an adjustment method for multiple testing, it uses the Bonferroni method according to Boyle et al. (2004). GOHyperGAll is based on the GOstats package (Falcon and Gentleman, 2007) from the BioConductor project (Gentleman et al., 2005), and it provides similar utilities as the hyperGTest function included in this package. The main differences of our method are that it simplifies the usage of custom gene-to-GO mappings, and it contains various utilities for efficiently analyzing large numbers of gene sets from cluster analyses in batch mode. All functions of the GOHyperGAll script are available in Supplemental Data S9.

Microarray Analysis

A total of 1,310 Affymetrix raw data Cel files were downloaded from the AtGenExpress and GEO sites (Schmid et al., 2005; Barrett et al., 2006; Kilian et al., 2007). All of them are derived from the Affymetrix ATH1 gene GeneChip microarray for Arabidopsis, and the corresponding samples contained at least two replicate samples. A summary of the utilized experiment sets is provided in Table III, whereas a detailed description of the analyzed data with their experimental design parameters is provided in Supplemental Data S2. The required probe set-to-locus mappings for the ATH1 chip were obtained from TAIR (ftp://ftp.arabidopsis.org/home/tair/Microarrays/Affymetrix; version 2/5/2007). All ambiguous probe sets on this chip were treated in the gene enumeration steps of this study in the following manner: controls and probe sets matching no or several loci in the Arabidopsis genome were ignored in the downstream analysis steps. In addition, redundant probe sets that represent the same locus several times were counted only once.

The normalization of the raw data Cel files was performed in R using the MAS5 and RMA algorithms that are implemented in the affy package form the BioConductor project (Irizarry et al., 2003, 2006; Qin et al., 2006). To allow in the DEG analysis comparisons between the different samples of an experiment set, the RMA normalization was performed in batches for entire experiment sets (Table III). This batch normalization is only required for the quantile-based RMA approach, but not for the MAS5 scaling approach. The present call information of the nonparametric Wilcoxon signed rank test was computed with the affy package to estimate the amount of unexpressed genes (Liu et al., 2002; McClintick and Edenberg, 2006). The obtained expression values from both normalization methods were uploaded to the PED database.

For the DEG analysis, the replicates and the most appropriate sample comparisons were determined manually for each experiment set. The generated analysis strategies were recorded in experiment definition tables (Supplemental Data S2). These tables were used to control the downstream DEG analysis steps in an automated manner by providing all information on replicates and sample comparisons to the statistical test methods. The actual analysis of DEGs was performed with the LIMMA package from Smyth (2004, 2005). The Benjamini and Hochberg method was selected to adjust P values for multiple testing and to determine FDRs (Benjamini and Hochberg, 1995). As confidence threshold we used an adjusted P value of ≤ 0.01 in combination with a minimum fold-change filter of 2. All DEG analyses were performed on both the MAS5 and RMA normalized data sets. Although both DEG analysis results were uploaded to the PED database, only the RMA set is discussed in this study because the RMA algorithm provides more accurate measurements on weaker expressed genes (Qin et al., 2006).

Cluster Analysis

The correlation and cluster analysis steps were performed in R on the MAS5 normalized expression data set. For this, the mean values from replicated biological measurements were combined in one large expression matrix. The RMA data were not used for cluster analysis, because they are less reliable for correlation studies than MAS5 data (Lim et al., 2007). The Pearson and Spearman correlation coefficients were calculated with the cor function in R. The obtained correlation coefficients were transformed into a correlation-based distance matrix after subtracting their values from 1. Four separate distance matrices were calculated for the Pearson and Spearman correlation coefficients in their signed and absolute forms. The matrices were passed on to the hclust function (Murtagh, 1985; R Development Core Team, 2006) that performs agglomerative hierarchical clustering. Complete linkage was used as cluster joining method.

To obtain from hierarchical dendrograms discrete clusters, we developed, to our knowledge, a new HTC method for this project. This method identifies subclusters in dendrograms based on a minimum tolerable similarity cutoff between all cluster members. This is achieved by applying an all-against-all similarity test for the clusters from all possible subtrees. At the same time, unique cluster memberships are maintained and all items in the processed dendrogram are assigned to clusters with one or more members. The corresponding HTC R script is available in Supplemental Data S10. As the cutoff we used for this cluster selection procedure a correlation coefficient of $\geq$0.6. This cutoff was chosen because it resulted in the highest enrichment of functionally related genes compared to alternative cutoffs settings (Supplemental Data S4). As a result of this method, the members of every identified cluster shared with all other members of the same cluster correlation coefficients between 0.6 and 1.0.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Data S1. PUF sets.

Supplemental Data S2. GeneChip microarray experiments and analysis strategies.

Supplemental Data S3. PMA data.

Supplemental Data S4. HTC cutoff selection.

Supplemental Data S5. Cluster data.

Supplemental Data S6. GO analysis of clusters.

Supplemental Data S7. DEG analysis.

Supplemental Data S8. GO analysis of DEGs.

Supplemental Data S9. R script for GO term enrichment analysis.

Supplemental Data S10. R script for HTC clustering.

ACKNOWLEDGMENTS

We thank the community projects—R, BioConductor, TAIR, The Institute for Genomic Research, GEO, and AtGenExpress—for providing the excellent software and data resources that were used by this project. T.G. acknowledges support from the Bioinformatics Core Facility, the Center for Plant Cell Biology, and the Institute for Integrative Genome Biology at University of California, Riverside.

Received February 3, 2008; accepted March 10, 2008; published March 19, 2008.

LITERATURE CITED

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol **215:** 403–410

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25:** 3389–3402

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet **25:** 25–29

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R (2006) NCBI GEO: mining tens of millions of expression profiles—database and tools update. Nucleic Acids Res **35:** D760–D765

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, et al (2004) The Pfam protein families database. Nucleic Acids Res **32:** D138–D141

Becker RA, Chambers JM, Wilks AR (1988) The New S Language. Wadsworth and Brooks/Cole, Pacific Grove, CA

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate—a practical and powerful approach to multiple testing. J Roy Statist Soc Ser B Methodological **57:** 289–300

Berardini TZ, Mundodi S, Reiser L, Huala E, Garcia-Hernandez M, Zhang P, Mueller LA, Yoon J, Doyle A, Lander G, et al (2004) Functional annotation of the Arabidopsis genome using controlled vocabularies. Plant Physiol **135:** 745–755

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, et al (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res **31:** 365–370

Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G (2004) GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics **20:** 3710–3715

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

Brown DM, Zeef LA, Ellis J, Goodacre R, Turner SR (2005) Identification of novel genes in *Arabidopsis* involved in secondary cell wall formation using expression profiling and reverse genetics. Plant Cell **17:** 2281–2295

de Hoon MJ, Imoto S, Nolan J, Miyano S (2004) Open source clustering software. Bioinformatics **20:** 1453–1454

Eddy SR (1996) Hidden Markov models. Curr Opin Struct Biol **6:** 361–365

Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA **95:** 14863–14868

Falcon S, Gentleman R (2007) Using GOstats to test gene lists for GO term association. Bioinformatics **23:** 257–258

Fukao T, Bailey-Serres J (2004) Plant responses to hypoxia—is survival a balancing act? Trends Plant Sci **9:** 449–456

Gachon CM, Langlois-Meurinne M, Henry Y, Saindrenan P (2005) Transcriptional co-regulation of secondary metabolism enzymes in Arabidopsis: functional and evolutionary implications. Plant Mol Biol **58:** 229–245

Gasch AP (2002) The environmental stress response: a common yeast response to environmental stresses. *In* S Hohmann, ed, Topics in Current Genetics, Vol. 1. Springer Verlag, Heidelberg, Germany, pp 11–70

Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO (2000) Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell **11:** 4241–4257

Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W (2005) Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer, New York

Girke T, Lauricha J, Tran H, Keegstra K, Raikhel N (2004) The Cell Wall Navigator database. A systems-based approach to organism-unrestricted mining of protein families involved in cell wall metabolism. Plant Physiol **136:** 3003–3008

Gollery M, Harper J, Cushman J, Mittler T, Girke T, Zhu JK, Bailey-Serres J, Mittler R (2006) What makes species unique? The contribution of proteins with obscure features. Genome Biol **7:** R57

Gollery M, Harper J, Cushman J, Mittler T, Mittler R (2007) POFs: what we don't know can hurt us. Trends Plant Sci (in press)

Grennan AK (2006) Genevestigator. Facilitating web-based gene-expression analysis. Plant Physiol **141:** 1164–1166

Gutierrez RA, Lejay LV, Dean A, Chiaromonte F, Shasha DE, Coruzzi GM (2007) Qualitative network models and genome-wide expression data define carbon/nitrogen-responsive molecular machines in Arabidopsis. Genome Biol **8:** R7

Haberer G, Mader MT, Kosarev P, Spannagl M, Yang L, Mayer KF (2006) Large-scale cis-element detection by analysis of correlated expression and sequence conservation between Arabidopsis and Brassica oleracea. Plant Physiol **142:** 1589–1602

Hebelstrup KH, Igamberdiev AU, Hill RD (2007) Metabolic effects of hemoglobin gene expression in plants. Gene **398:** 86–93

Hong F, Breitling R, McEntee CW, Wittner BS, Nemhauser JL, Chory J (2006) RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. Bioinformatics **22:** 2825–2827

Horan K, Lauricha J, Bailey-Serres J, Raikhel N, Girke T (2005) Genome cluster database. A sequence family analysis platform for Arabidopsis and rice. Plant Physiol **138:** 47–54

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP (2003) Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res **31:** e15

Irizarry RA, Gautier L, Bolstad BM, Magnus Astrand CM, Cope LM, Gentleman R, Gentry J, Halling C, Huber W, MacDonald J, et al (2006) affy: Methods for Affymetrix Oligonucleotide Arrays. R Package Version 1.12.1. http://www.bioconductor.org

Jen CH, Manfield IW, Michalopoulos I, Pinney JW, Willats WG, Gilmartin PM, Westhead DR (2006) The Arabidopsis co-expression tool (ACT): a WWW-based tool and database for microarray-based gene expression analysis. Plant J **46:** 336–348

Johnson O, Liu J (2006) A traveling salesman approach for predicting protein functions. Source Code Biol Med **1:** 3

Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M (2006) From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res **34:** D354–D357

Kilian J, Whitehead D, Horak J, Wanke D, Weinl S, Batistic O, D'Angelo C, Bornberg-Bauer E, Kudla J, Harter K (2007) The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. Plant J **50:** 347–363

Krishnapuram R, Joshi A, Nasraoui O, Yi L (2001) Low-complexity fuzzy relational clustering algorithms for Web mining. IEEE Trans Fuzzy Syst **9:** 595–607

Leinonen R, Diez FG, Binns D, Fleischmann W, Lopez R, Apweiler R (2004) UniProt archive. Bioinformatics **20:** 3236–3237

Lim WK, Wang K, Lefebvre C, Califano A (2007) Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. Bioinformatics **23:** 282–288

Liu WM, Mei R, Di X, Ryder TB, Hubbell E, Dee S, Webster TA, Harrington CA, Ho MH, Baid J, et al (2002) Analysis of high density expression microarrays with signed-rank call algorithms. Bioinformatics **18:** 1593–1599

Ma S, Gong Q, Bohnert HJ (2007) An Arabidopsis gene network based on the graphical Gaussian model. Genome Res **17:** 1614–1625

McClintick JN, Edenberg HJ (2006) Effects of filtering by Present call on analysis of microarray experiments. BMC Bioinformatics **7:** 49

Mueller LA, Zhang P, Rhee SY (2003) AraCyc: a biochemical pathway database for Arabidopsis. Plant Physiol **132:** 453–460

Murtagh F (1985) Multidimensional Clustering Algorithms. COMPSTAT Lectures 4. Physica-Verlag, Wuerzburg, Germany

Nelson DR, Schuler MA, Paquette SM, Werck-Reichhart D, Bak S (2004) Comparative genomics of rice and Arabidopsis. Analysis of 727 cytochrome P450 genes and pseudogenes from a monocot and a dicot. Plant Physiol **135:** 756–772

Persson S, Wei H, Milne J, Page GP, Somerville CR (2005) Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. Proc Natl Acad Sci USA **102:** 8633–8638

Prelic A, Bleuler S, Zimmermann P, Wille A, Bühlmann P, Gruissem W, Hennig L, Thiele L, Zitzler E (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics **22:** 1122–1129

Qin LX, Beyer RP, Hudson FN, Linford NJ, Morris DE, Kerr KF (2006) Evaluation of methods for oligonucleotide array data via quantitative real-time PCR. BMC Bioinformatics **7:** 23–23

R Development Core Team (2006) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna

Rodriguez R, Redman R (2005) Balancing the generation and elimination of reactive oxygen species. Proc Natl Acad Sci USA **102:** 3175–3176

Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Scholkopf B, Weigel D, Lohmann JU (2005) A gene expression map of Arabidopsis thaliana development. Nat Genet **37:** 501–506

Schwacke R, Schneider A, van der Graaff E, Fischer K, Catoni E, Desimone M, Frommer WB, Flugge UI, Kunze R (2003) ARAMEMNON, a novel database for Arabidopsis integral membrane proteins. Plant Physiol **131:** 16–26

Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol **3:** 3

Smyth GK (2005) Limma: linear models for microarray data. *In* R Gentleman, V Carey, S Dudoit, R Irizarry, W Huber, eds, Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer, New York, pp 397–420

Steinhauser D, Junker BH, Luedemann A, Selbig J, Kopka J (2004a) Hypothesis-driven approach to predict transcriptional units from gene expression data. Bioinformatics **20:** 1928–1939

Steinhauser D, Usadel B, Luedemann A, Thimm O, Kopka J (2004b) CSB.DB: a comprehensive systems-biology database. Bioinformatics **20:** 3647–3651

Toufighi K, Brady SM, Austin R, Ly E, Provart NJ (2005) The Botany Array Resource: e-Northerns, Expression Angling, and promoter analyses. Plant J **43:** 153–163

Usadel B, Nagel A, Steinhauser D, Gibon Y, Bläsing OE, Redestig H, Sreeniva-sulu N, Krall L, Hannah MA, Poree F, et al (2006) PageMan: an interactive ontology tool to generate, display, and annotate overview graphs for profiling experiments. BMC Bioinformatics **7:** 535

Vandepoele K, Casneuf T, Van de Peer Y (2006) Identification of novel regulatory modules in dicot plants using expression data and comparative genomics. Genome Biol **7:** R103

Wang D, Harper JF, Gribskov M (2003) Systematic trans-genomic comparison of protein kinases between Arabidopsis and *Saccharomyces cerevisiae*. Plant Physiol **132:** 2152–2165

Wei H, Persson S, Mehta T, Srinivasasainagendra V, Chen L, Page GP, Somerville C, Loraine A (2006) Transcriptional coordination of the metabolic network in Arabidopsis. Plant Physiol **142:** 762–774

Whetzel PL, Parkinson H, Causton HC, Fan L, Fostel J, Fragoso G, Game L, Heiskanen M, Morrison N, Rocca-Serra P, et al (2006) The MGED Ontology: a resource for semantics-based description of microarray experiments. Bioinformatics **22:** 866–873

Wolfe CJ, Kohane IS, Butte AJ (2005) Systematic survey reveals general applicability of ''guilt-by-association'' within gene coexpression networks. BMC Bioinformatics **6:** 227

Wortman JR, Haas BJ, Hannick LI, Smith RK, 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

Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, et al (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res **34:** D187–D191

Yang Y, Engin L, Wurtele ES, Cruz-Neira C, Dickerson JA (2005) Integration of metabolic networks and gene expression in virtual reality. Bioinformatics **21:** 3645–3650

Zimmermann P, Hennig L, Gruissem W (2005) Gene-expression analysis and network discovery using Genevestigator. Trends Plant Sci **10:** 407–409

Zimmermann P, Hirsch-Hoffmann M, Hennig L, Gruissem W (2004) GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. Plant Physiol **136:** 2621–2632