Exploring the Plant Transcriptome through Phylogenetic Profiling[1][w]

Klaas Vandepoele and Yves Van de Peer*

Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, B–9052 Ghent, Belgium

Publicly available protein sequences represent only a small fraction of the full catalog of genes encoded by the genomes of different plants, such as green algae, mosses, gymnosperms, and angiosperms. By contrast, an enormous amount of expressed sequence tags (ESTs) exists for a wide variety of plant species, representing a substantial part of all transcribed plant genes. Integrating protein and EST sequences in comparative and evolutionary analyses is not straightforward because of the heterogeneous nature of both types of sequence data. By combining information from publicly available EST and protein sequences for 32 different plant species, we identified more than 250,000 plant proteins organized in more than 12,000 gene families. Approximately 60% of the proteins are absent from current sequence databases but provide important new information about plant gene families. Analysis of the distribution of gene families over different plant species through phylogenetic profiling reveals interesting insights into plant gene evolution, and identifies species- and lineage-specific gene families, orphan genes, and conserved core genes across the green plant lineage. We counted a similar number of approximately 9,500 gene families in monocotyledonous and eudicotyledonous plants and found strong evidence for the existence of at least 33,700 genes in rice (*Oryza sativa*). Interestingly, the larger number of genes in rice compared to Arabidopsis (*Arabidopsis thaliana*) can partially be explained by a larger amount of species-specific single-copy genes and species-specific gene families. In addition, a majority of large gene families, typically containing more than 50 genes, are bigger in rice than Arabidopsis, whereas the opposite seems true for small gene families.

Comparative genomics provides a powerful means to study gene structure and the evolution of gene function and regulation. Analysis of genes or pathways in a broad phylogenetic context allows scientists to better understand how complex biological processes are regulated and evolve (Soltis and Soltis, 2003; Koonin et al., 2004). Although phylogenetic studies can provide important insights into gene and genome evolution (for examples, see Ermolaeva et al., 2003; Griffiths et al., 2003; Vandepoele et al., 2003), a dense taxonomical sampling is necessary to obtain a complete and accurate view of the evolutionary history of a biological process and its underlying genes. Similarly, to draw biologically relevant conclusions, the inference of orthology and paralogy between homologous genes requires a good phylogenetic sampling (for review, see Doyle and Gaut, 2000). Moreover, a coherent classification of homologous genes is essential for the high-throughput extraction of functional and evolutionary information from gene phylogenies. In this respect, the availability of numerous large-scale sequencing projects offers the opportunity to study homologous genes, typically gene families, from an evolutionary point of view. The construction of phylogenetic profiles, which reflect the presence or absence of a particular gene family in a biological species, is an effective method for the detection of conserved core genes, species-specific single-copy genes, species-specific gene family (SSGF) or lineage-specific gene family (LSGF) expansions, gene loss, and genes that have been transferred between nuclear and organellar genomes. Furthermore, analysis of the phylogenetic profiles of protein families and of domain fusion events helps to predict functional interactions and to deduce specific functions for numerous proteins (Kriventseva et al., 2001).

Perhaps the best known example of an integrated sequence-based system applying phylogenetic profiles is the clusters of orthologous groups (COG) database, which is a comprehensive repository of functionally annotated clusters of bacterial and eukaryotic orthologous genes (Tatusov et al., 2003). Although in bacteria, fungi, and animals various sequencing projects constantly enlarge the gene space (for an overview, see http://www.ncbi.nlm.nih.gov/genomes/static/EG_T.html), the situation is different for plants (Pryer et al., 2002). Apart from Arabidopsis (*Arabidopsis thaliana*) and rice (*Oryza sativa*), in which genome sequencing projects present a first overview of the eudicotyledonous and monocotyledonous gene repertoire, respectively (Arabidopsis Genome Initiative, 2000; Feng et al., 2002; Goff et al., 2002: Sasaki et al., 2002; Yu et al., 2002; Rice Chromosome 10 Sequencing Consortium, 2003), the majority of all other Viridiplantae, ranging from early land plants such as mosses and ferns to highly developed flowering plants, lack a comprehensive

[1] This work was supported by the Instituut voor de aanmoediging van Innovatie door Wetenschap en Technologie in Vlaanderen (predoctoral fellowship to K.V.).

* Corresponding author; e-mail yves.vandepeer@psb.ugent.be; fax 32–9–33–13809.

[w] The online version of this article contains Web-only data. www.plantphysiol.org/cgi/doi/10.1104/pp.104.054700.

overview of the proteins encoded by their genomes. On the other hand, an enormous amount of plant expressed sequence tags (ESTs)—single-pass sequence reads from reverse-transcribed mRNAs—is publicly available and provides a substantial representation of the plant transcriptome (Rudd, 2003). Because the overall number of plant ESTs is by far larger than that of plant proteins currently stored in public sequence repositories, the phylogenetic analysis of plant genes based on protein sequences is difficult and inefficient (Raes et al., 2003) and offers only a very limited view on the total amount of sequence information available.

Here, we present an integrated sequence repository (available as Sequence platform for the Phylogenetic analysis of Plant Genes [SPPG] in the section Databases at http://www.psb.ugent.be/bioinformatics/) that combines EST sequence data with protein information, providing an excellent starting point for plant comparative and evolutionary genomics. This is illustrated by the examination of several thousands of gene families distributed over a large number of different plant species, which reveals unique features about the evolution of plant gene families.

RESULTS AND DISCUSSION

EST Assembly, ORF Detection, Protein Clustering, and Functional Annotation

Initially, 106,174 proteins and 2,884,000 EST sequences from 32 different plant species were retrieved from EMBL and The Institute for Genomic Research (TIGR) to construct a nonredundant and high-quality data set of plant proteins. After the assembly of the EST sequences, annotation of open reading frames (ORFs) on EST clusters, and processing all currently available proteins for the plant species selected here (see ''Materials and Methods'' for technical details), a total of 86,077 nonredundant plant proteins from EMBL and TIGR were obtained, together with 253,857 EST clusters derived from more than 1.8 million clustered EST sequences (Table I; Fig. 1). Fifty-seven percent of all initial EST sequences could be assembled into an EST cluster comprising, on average, 6.16 ESTs. These results are very comparable with similar plant EST assembly initiatives (TIGR Plant Gene Indices, Quackenbush et al., 2001; PlantGDB, Dong et al., 2004). Nevertheless, because we applied more stringent assembly criteria to reduce the creation of chimeras and other artificial cDNAs (see ''Materials and Methods''), the overall number of EST assemblies per species is slightly smaller than that in PlantGDB and TIGR Gene Indices. For two-thirds of all EST clusters, an ORF longer than 50 codons could be determined, resulting in 166,306 protein sequences (Fig. 1). Thus, in total 252,383 nonredundant plant proteins were assigned to the final data set. Approximately 82% of all proteins (which corresponds to 207,023 proteins) could be assigned to 14,369 gene families, here defined as a set of two or more homologous gene sequences. Overall, a good correlation between the initial number of ESTs and the final number of clustered plant proteins was observed ($r^2 = 0.88$), which indicates that there is no significant bias in the EST assembly and ORF annotation routines applied for these different Viridiplantae species (see ''Materials and Methods'' for details). Whereas a minority of gene families (i.e. 4,275) contains only proteins derived from EST clusters, the majority (i.e. 10,094) consists of proteins from EMBL, TIGR, or both. In addition, 46% (6,664) of all gene families contain proteins derived from both EST clusters and EMBL or TIGR. Consequently, this subset corresponds to gene families with a dense sampling over the different plant species included in the data set, with an average total of 27 proteins per family from 9.7 different plant species. By contrast, the overall sampling density for all 14,369 gene families is 14.4 genes sampled over 5.5 different plants per family. Despite the fact that only 25% of all proteins derived from EST clusters are truly full length (i.e. the protein begins with a start codon and ends with a stop codon), the majority (86%) of all these proteins have significant homology with other proteins, offering additional information for the phylogenetic profiles (see below). Approximately 45,000 protein sequences were not clustered into gene families. Although 30% of these unclustered proteins represent single-copy species-specific genes (or orphan genes; see below), the majority corresponds to partial proteins, derived from incomplete ORFs annotated on non-full-length EST clusters, with sometimes only partial homology to other plant proteins. Indeed, one might expect that a number of gene families only comprising proteins derived from EST clusters will represent partial proteins. These proteins will not be clustered with the corresponding full-length proteins because they do not fulfill the global homology criterion required for being added to such a group of related proteins. We estimate that approximately 11% of all gene families form a group of related partial proteins, derived from EST clusters, for which a related full-length gene family exists (see ''Materials and Methods'').

Gene families and individual genes have been functionally annotated based on the available gene descriptions and Gene Ontology (GO) annotations of protein sequences derived from EMBL and TIGR. Approximately 58,000 gene descriptions could be mapped on 11,938 different gene families, and 22,395 functional GO labels of Arabidopsis could be assigned to 4,099 gene families. When gene descriptions are transferred between different members of the same gene family, more than 80% of plant sequences can be labeled with functional information.

Gene Content in Chloroplast and Mitochondrial Genomes

In addition to assigning general gene descriptions to families or individual proteins, information about the nuclear or organellar origin of genes has also been

Table I. *Number of EST and protein sequences combined in EST assembly, CDS annotation, and protein clustering*

–, Not determined.

Species Name	FrameD IMM[a]	ESTs	EST Clusters	ESTs in Cluster	ESTs per Cluster	ORF from EST Clusters[b]	ePROT[b,c]	Total Proteins	Proteins in Gene Family	Orphan Proteins[d]
				%						
Arabidopsis	–	–	–	–	–	–	26,294	26,294	22,412	1,050 (253)
Beta vulgaris	Caryo	19,039	2,334	0.33	2.72	1,785	76	1,855	1,607	38
Brassica napus	Arabidopsis	37,896	4,913	0.58	4.49	3,676	813	4,350	3,996	95
Capsicum annuum	Aster	23,361	2,420	0.53	5.09	1,835	218	2,013	1,885	16
Chlamydomonas reinhardtii	Chlamy	141,100	9,528	0.72	10.62	2,480	738	3,015	1,889	131
Glycine max	Arabidopsis	342,149	28,726	0.78	9.32	19,189	1,332	20,180	17,726	584
Gossypium arboreum	Arabidopsis	38,967	3,577	0.43	4.67	2,613	38	2,648	2,397	57
Gossypium hirsutum	Arabidopsis	14,307	1,150	0.32	4.03	742	609	1,114	1,040	21
Helianthus annuus	Aster	60,785	5,024	0.60	7.30	2,589	306	2,827	2,496	70
Hordeum vulgare	Oryza	202,705	14,925	0.74	10.03	10,060	1,150	10,894	9,752	201
Lactuca sativa	Aster	69,319	6,698	0.66	6.87	4,679	83	4,755	4,150	90
Lotus corniculatus var. *Japonicus*	Arabidopsis	24,896	2,695	0.74	6.85	1,780	132	1,893	1,568	127
Lycopersicon esculentum	Aster	151,147	14,428	0.82	8.62	10,478	1,442	11,546	10,292	155
Medicago truncatula	Arabidopsis	188,367	16,867	0.83	9.22	12,567	233	12,772	11,143	191
Mesembryanthemum crystallinum	Caryo	26,563	2,471	0.70	7.50	1,673	207	1,857	1,674	34
Nicotiana tabacum	Aster	11,276	453	0.11	2.70	69	1,697	1,435	1,289	47
Oryza sativa	–	–	–	–	–	–	47,475	47,475	30,993	7,882 (704)
Physcomitrella patens	Physco	104,161	12,924	0.82	6.65	10,217	355	10,298	6,319	2,053
Pinus pinaster	Pinus	9,059	1,222	0.51	3.79	835	66	895	822	9
Pinus taeda	Pinus	73,349	5,325	0.40	5.53	2,895	154	3,004	2,466	97
Populus balsamifera subsp. *Trichocarpa*	Arabidopsis	24,579	2,320	0.58	6.18	1,725	49	1,758	1,621	28
Populus tremula	Arabidopsis	14,081	1,120	0.42	5.23	833	33	839	791	16
P. tremula × *Populus tremuloides*	Arabidopsis	56,048	4,757	0.55	6.50	3,401	68	3,467	3,197	81
Populus x canescens	Arabidopsis	10,499	754	0.26	3.60	526	20	546	513	13
Prunus persica	Arabidopsis	10,939	1,071	0.54	5.52	829	127	940	886	12
Solanum tuberosum	Aster	95,632	16,151	0.78	4.63	11,634	985	12,341	10,755	161
Sorghum bicolor	Oryza	134,740	15,303	0.80	7.06	10,925	426	11,278	9,771	219
Sorghum propinquum	Oryza	21,390	3,148	0.65	4.44	2,349	6	2,355	2,130	28
Triticum aestivum	Oryza	508,406	35,271	0.45	6.51	22,615	1,432	23,788	21,115	535
Vitis vinifera	Arabidopsis	111,849	11,163	0.82	8.17	6,674	233	6,861	5,952	260
Zea mays	Oryza	362,796	26,807	0.65	8.81	14,704	3,819	16,919	14,256	397
Zinnia elegans	Aster	9,836	312	0.07	2.15	119	57	171	120	4
Total/Average		2,899,241	253,857	0.57	6.16	166,496	96,219	252,383	207,023	14,457

[a]The IMM used to determine the ORF of an EST cluster (see "Materials and Methods" for more details). [b]The number of sequences before removing redundant entries. [c]The number of protein sequences available in EMBL/TIGR after removing transposon-like genes in Arabidopsis and rice. [d]The numbers in parentheses give the number of predicted orphan genes supported by EST/cDNA (match with >95% identity across >100 bp) for Arabidopsis and rice.

integrated, which allows us to determine the amount of chloroplast and mitochondrial DNA sequences that have been inserted into or transferred to the nucleus. In total, 704 chloroplast and 275 mitochondrial gene products were identified that could be clustered into 202 distinct gene families. Interestingly, in numerous gene families, genes from different origins were grouped. Sixty-six and 24 gene families were found uniquely for chloroplast or mitochondrial genomes, respectively, whereas 110 organelle families were identified for which homologs were also detected in the nuclear plant genome of Arabidopsis or rice. Two gene families were identified encoded by the chloroplast and mitochondrial genome (NADH dehydrogenase subunits 1 and 4). Gene families in both mitochondrial and nuclear genomes encode for cytochrome c subunits, ribosomal proteins, and tRNAs, whereas a wide variety of genes, covering 66 different gene families, was found in both chloroplast and nuclear genomes (for full list, see Supplemental Table I). In addition, 10 families were identified in the mitochondrial, chloroplast, and nuclear genomes of different species encoding ribosomal proteins, NADH dehydrogenase subunits, Fe-superoxide dismutase, ATP synthase

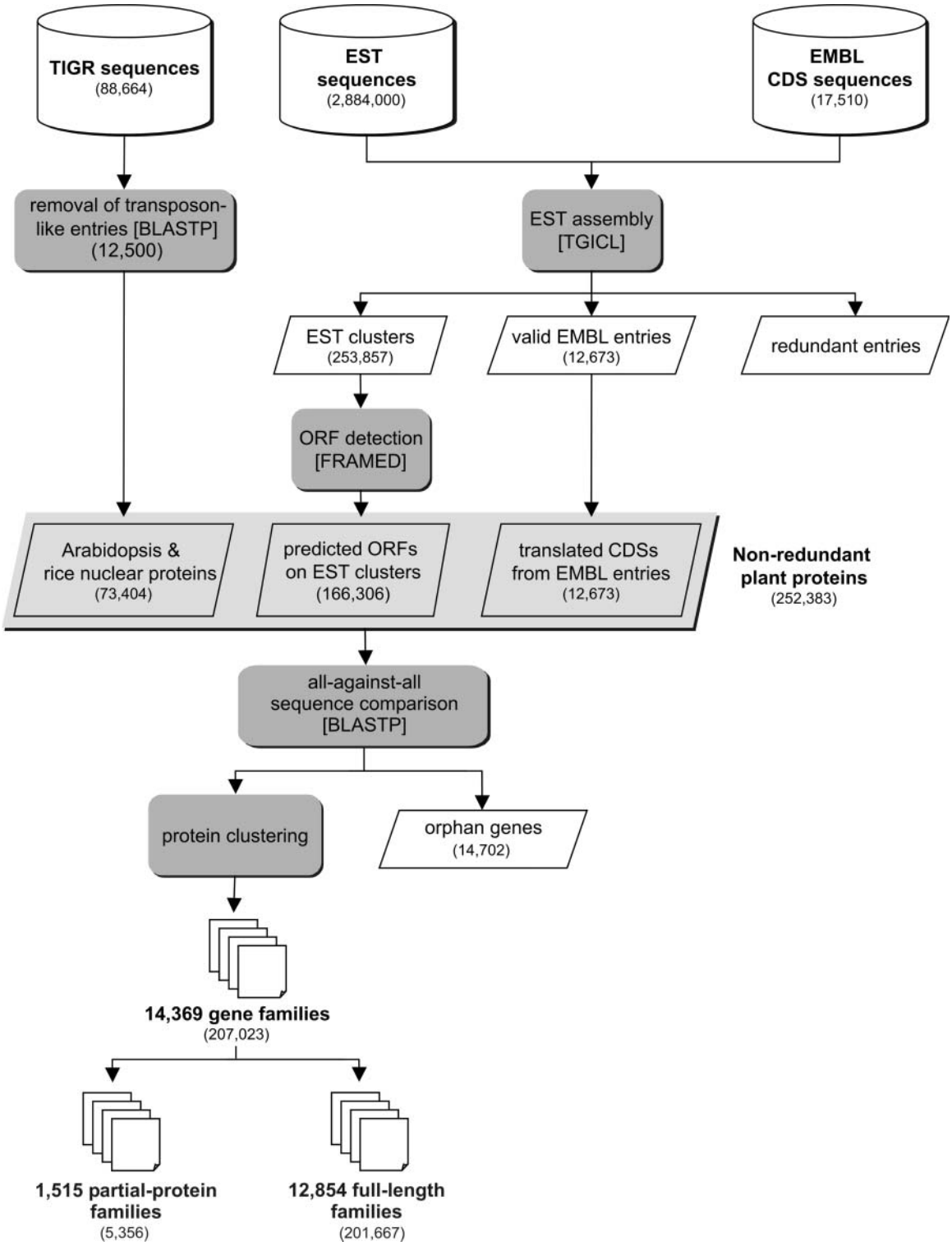

Figure 1. Schematic overview of the construction of the data set. The white barrels represent the initial sequence data retrieved from TIGR and EMBL, the dark gray boxes routines applied to manipulate and organize the data, whereas the light gray box describes the final amount of sequence data derived from the different sources (see text for details). Except for Arabidopsis and rice, whose nuclear protein-encoding genes were retrieved from TIGR, all other sequence data for the 32 species were obtained through EMBL. The numbers of sequences are indicated in parentheses.

subunit 1, and an Asn tRNA. This confirms previous findings that genes frequently are transferred from the chloroplast or mitochondrial genome to the nucleus, where they acquire new expression control and targeting signals for the correct expression, translation, and reimport into the organelle (Martin, 2003).

Strikingly, whereas 19% (15 out of 76 gene families) of all chloroplast gene functions in Arabidopsis are also present in the nuclear genome, in rice 37% (30 out of 81 gene families) of all chloroplast gene functions are found in the nuclear genome. This difference confirms previous findings that the rice nuclear genome is significantly more enriched with plastid genome sequences than that of Arabidopsis (Shahmuradov et al., 2003). Although recent gene transfers from the chloroplast to the nuclear genome might be associated with chloroplast genome reduction due to subsequent gene loss, the overall number of distinct gene functions in the rice chloroplast genome is not significantly different from that of the Arabidopsis chloroplast genome (81 and 76 gene families, respectively). Therefore, it is currently unclear whether this current redundancy represents the first step of the transfer of chloroplast gene functions to the rice nucleus and has any evolutionary consequences (Timmis et al., 2004).

Application of Phylogenetic Profiles for the Evolutionary Classification of Plant Genes

An overview of the number of proteins ascribed to gene families is shown in Table I. As expected, the largest numbers of proteins that can be assigned to gene families are derived from Arabidopsis and rice (22,412 and 30,993 genes, respectively), for which nearly complete nuclear genome sequences have been determined. Monocotyledonous plants, such as *Triticum aestivum*, *Zea mays*, *Sorghum bicolor*, and *Hordeum vulgare*, are also well represented, as well as the eudicotyledonous plants *Glycine max*, *Medicago truncatula*, *Solanum tuberosum*, *Lycopersicon esculentum*, and *Vitis vinifera*. For the moss *Physcomitrella patens*, more than 6,300 proteins are clustered into gene families, which can be explained by the exhaustive EST-sequencing efforts lately (Nishiyama et al., 2003). By contrast, for other plants only a limited number of protein sequences are available.

In addition to defining sensu stricto phylogenetic profiles at the species level, we also determined the overall presence of each gene family over distinct taxa of the Viridiplantae. The different taxa scored were, at lower taxonomic levels, Chlorophyta, Bryophyta, gymnosperms, and angiosperms, the latter being further subdivided in monocots and eudicots. At a higher taxonomic level, *Eurosids I*, *Eurosids II*, *Rosids*, *Asterids*, and Caryophyllales were discerned. Given the still very incomplete nature of most available plant gene sequences, these high-level phylogenetic profiles offer an alternative representation of the distribution of gene families within the green lineage (Fig. 2). Moreover, these alternative profiles provide a valuable tool for the extraction of information about the evolution of gene functions.

Core Plant Genes, Species- and Lineage-Specific Gene Families, and Orphans

Examination of the high-level phylogenetic profiles revealed that a total of 397 gene families covering

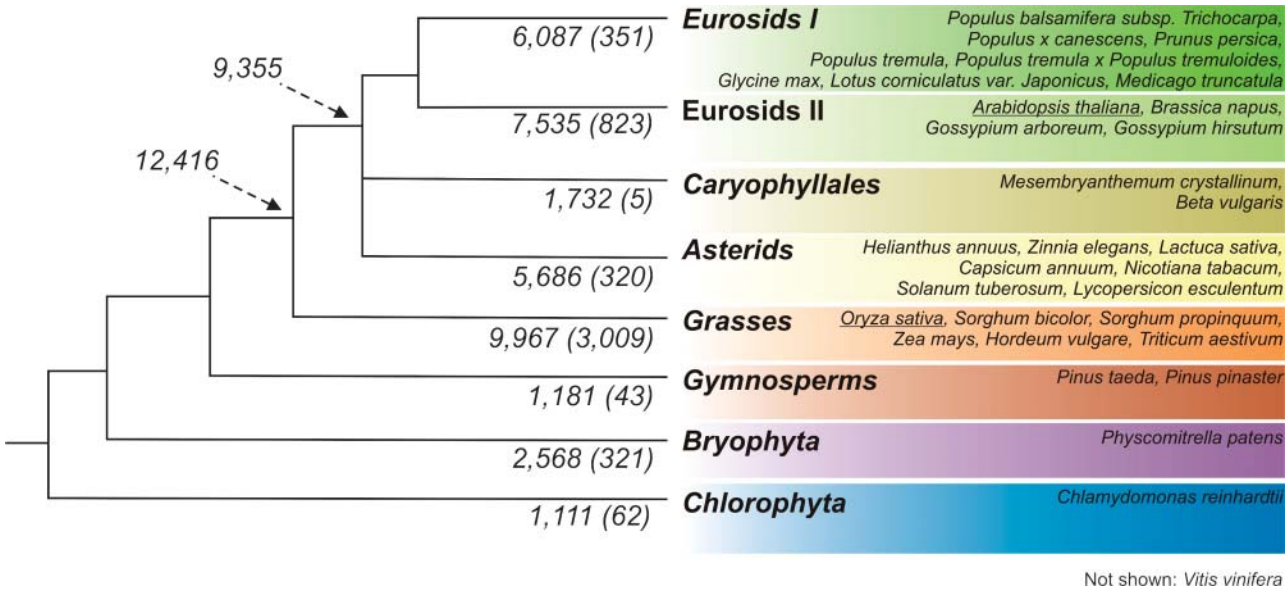

Figure 2. Phylogenetic distribution of all gene families over different taxa of the Viridiplantae. The number of gene families in one or more species belonging to a particular taxon is shown beneath the branches. The number of gene families exclusively found for a particular taxon is shown between parentheses. Families grouping partial protein sequences were discarded (see text for details). Arrows indicate the number of gene families found in the eudicots (9,355) and angiosperms (12,416).

53,796 proteins were present in *Chlorophytes*, *Bryophytes*, gymnosperms, and angiosperms. These conserved gene families thus represent a set of core genes found in all major divisions of the Viridiplantae. As expected, the functional classification of these gene families shows that they encode basic components of the plant cell machinery, such as genes involved in translation, ribosomal structure, posttranslational modifications, energy production, secretion, amino acid transport, and metabolism (see Supplemental Fig. 1). The number of core proteins in Arabidopsis identified here (4,177) is larger than the 1,152 Arabidopsis proteins conserved in all eukaryotes (Gutierrez et al., 2004), which can be explained by the presence of a large number of gene functions specific to the green lineage but absent from other eukaryotic kingdoms. Indeed, we find that only 10% of the Arabidopsis plant core genes is part of the eukaryotic core as defined by Gutierrez et al. (2004), suggesting a large number of plant-specific core gene functions. As expected, a large number of these plant-specific core genes are involved in photosynthesis. Surprisingly, when combining the set of 3,848 plant-specific Arabidopsis proteins identified by Gutierrez et al. (2004) with the phylogenetic profiles computed here, only 3% of these proteins belong to the set of core gene families. This indicates that a large fraction of these putative plant-specific genes are part of SSGFs or LSGFs and do not belong to the set of plant core genes, as it is defined now by including more plant species. It should be noted that in our data set only 8 gene families were found in all 32 plant species, which is very illustrative for the current poor status of gene sampling in plants. Surprisingly, 26 core gene families (i.e. 7% of all core families) correspond to genes with unknown function, which suggests that they represent essential, albeit unexplored, gene functions in plants. This number is significantly higher than the 2% of uncharacterized core gene families in pan-eukaryotic orthologous groups (KOGs; 18/860; http://www.ncbi.nlm.nih.gov/COG/new/; Koonin et al., 2004). Genes typically used for reconstructing the phylogenetic relationships between different plant phyla were found in a majority, if not all, species (e.g. tubulin, actin, Rubisco subunits, heat shock protein hsp70, and elongation factor 1α).

In contrast with the set of core genes, a large number of gene families are specific to one particular plant. Initially, 3,337 SSGFs were identified when querying the profiles of all gene families. Because the general gene family delineation was performed with rather conservative criteria, less stringent protein clustering parameters were applied in order to determine the real number of SSGFs, LSGFs, and orphan genes (see "Materials and Methods"). In total, 1,116 SSGFs containing 5,180 proteins were detected, with the largest number in rice, Arabidopsis, and Physcomitrella, covering 637 (approximately 4,258 proteins), 187 (approximately 1,241 proteins), and 164 (approximately 408 proteins) gene families, respectively. The availability of a complete genome sequence for Arabidopsis and rice may be the reason for the larger number of SSGF proteins, whereas for Physcomitrella the absence of sequence data from closely related species in combination with the large number of available EST/cDNA sequences explains the high amount of SSGF proteins. Approximately 82% of all SSGF proteins lack a functional annotation, which indicates that they play a role in unknown or poorly characterized biological processes. Although one might expect that LSGFs will be hard to detect in an incomplete and fragmented plant data set (Jabbari et al., 2004), several examples were obtained by querying the phylogenetic profiles. An overview of some SSGFs and LSGFs for which functional information is available is given in Table II. The largest SSGF was found in Arabidopsis and codes for Ulp1 proteases, a eukaryotic class of Cys proteases. Examples of genes driving unique taxa-specific biological processes are matrix metalloproteases, lytic enzymes digesting the cell walls of mating-type gametes during mating in *Chlamydomonas reinhardtii* (Kinoshita et al., 1992), specific nodulin genes participating in nodule formation and function in legume plants (Kevei et al., 2002; Mergaert et al., 2003), and zeins, a class of seed storage proteins typically found in panicoid cereals (Shewry and Halford, 2002).

To estimate the real number of orphan genes for a particular organism, we compared these proteins with the total data set by using less strict sequence similarity criteria than those used for the construction of the gene families. Still more than 14,000 orphan genes were detected, the largest number being found in rice and the lowest in *Zinnia elegans* (Table I). Interestingly, the number of expressed orphan genes is only 6,482 because almost one-half of all putative orphans are predicted genes of rice and Arabidopsis lacking proof of expression (no EST- or cDNA-supported gene model). *P. patens* seems to be the organism with the highest number of expressed orphan genes (2,053) in the full data set, which can be explained by its unique taxonomic position and current EST/cDNA sequencing status. Indeed, *P. patens* is the only moss representative in the data set and has a high number of ESTs yielding more than 10,000 different moss proteins. Overall, disregarding *P. patens*, the observed correlation between the number of initial EST sequences and the final number of orphan genes for all plant species is linear (r^2 = 0,83; y = 0.0011x + 25.482). Hence, within these plant species, the chance of detecting new orphan genes only increases with one new orphan per approximately 900 additional ESTs. In this respect, the 131 orphan genes for *C. reinhardtii*, which also lacks closely related species in this data set and has a high number of ESTs (>140,000), seems unexpectedly low. Most probably, the fact that only 26% of all *C. reinhardtii* EST clusters yielded a protein sequence of more than 50 amino acids compared to 79% for *P. patens*, for which overall longer cDNA sequences could be obtained, reduces the number of detectable Chlamydomonas orphan genes. The current

Table II. *Examples of SSGFs and LSGFs*

Phylogenetic Profile	Taxon[a]	Family ID	No. of Homologous Proteins	Function	Comments
SSGF	Arabidopsis	10207	102	Ulp1 protease family	
SSGF	Arabidopsis	1607	72	F-box protein family	Confirmed by EST/cDNA
SSGF	Arabidopsis	2397	21	Cytochrome P-450 aromatase-related	
SSGF	*Chlamydomonas reinhardtii*	12880	10	Matrix metalloprotease	Homologs found in Volvox and vertebrates
SSGF	*C. reinhardtii*	6240	2	Hyp-rich glycoprotein	
SSGF	*C. reinhardtii*	7610	4	Sulfur deprivation response regulator	Homologs found in bacteria
SSGF	*C. reinhardtii*	4173	4	Nitrite transporter NAR1	Homologs found in bacteria
SSGF	*C. reinhardtii*	287	6	Perphorin	Homologs found in Volvox
SSGF	*Glycine max*	6927	6	Nodulin 22	homologs found in Phaseolus
SSGF	*Medicago truncatula*	5884	3	Nodule-specific Gly-rich protein	
SSGF	*Nicotiana tabacum*	4926	5	Putative translation transactivator	
SSGF	*Pinus taeda*	4061	6	Nonspecific lipid transfer protein	
SSGF	*Vitis vinifera*	11153	7	Putative Pro-rich cell wall protein	
SSGF	*Zea mays*	10123	7	Basal layer antifungal peptide	
SSGF	*Z. mays*	12935	6	MURA-like protein	
SSGF	*Mesembryanthemum crystallinum*	5565	3	Antimicrobial peptide 1 precursor	Homologs found in related Caryophyllales
LSGF	Gymnosperms (2 species)	7133	3	Gly-rich protein	
LSGF	Asterids (4 species)	9108	22	Proteinase inhibitor II protein	
LSGF	Asterids (3 species)	1844	12	γ-Thionin 1 precursor	
LSGF	Asterids (2 species)	5981	10	Probable metallocarboxypeptidase inhibitor	
LSGF	Asterids (3 species)	1996	8	Cys-rich extensin-like protein	
LSGF	Asterids (4 species)	7470	5	Hypothetical protein SENU1, senescence up-regulated	
LSGF	Eurosids I (2 Fabaceae species)	3404	18	Albumin 1	
LSGF	Eurosids I (2 Fabaceae species)	2428	2	Nodulin 6	
LSGF	Monocots (3 species)	4462	20	γ-Gliadin	
LSGF	Monocots (3 species)	5153	80	Zein-α precursor	
LSGF	Monocots (5 species)	6364	39	Pollen allergen	
LSGF	Monocots (4 species)	10095	39	α-Amylase inhibitor	
LSGF	Monocots (5 species)	9761	12	Protein synthesis inhibitor	

[a]The number of species covered by the gene family for specific taxa is indicated in parentheses.

sequencing and gene annotation of the Chlamydomonas genome will probably reveal additional information about the amount of Chlorophyta-specific and orphan genes (Grossman et al., 2003).

Gene Loss in Arabidopsis and Rice

To determine specific gene-loss events in Arabidopsis and rice, we searched the phylogenetic profiles for conserved gene functions present in numerous eudicots and grasses but absent in Arabidopsis and rice, respectively. Subsequently, we used less stringent sequence similarity criteria (see ''Materials and Methods'') to validate whether a particular gene family indeed was absent in the full proteome of Arabidopsis or rice. We identified seven gene families that were present in five or more plant species, including related Eurosid II species, but were absent from Arabidopsis. A detailed search with protein sequences of related plants for the missing genes against the raw genomic Arabidopsis bacterial artificial chromosome (BAC) sequences yielded three loci with significant similarity (Table III; Supplemental Table II). This indicates that these loci may represent active genes missed by the current gene annotation efforts, whereas the absence of the other four gene families could point to gene loss in Arabidopsis. An alternative explanation is that these four gene functions do exist in Arabidopsis but are located in currently unsequenced chromosomal regions, such as centromeres (Yamada et al., 2003; Nagaki et al., 2004). In rice, 62 gene-loss events were detected for gene families with homologs in five or more other species, including other cereals. For more than 70% (45/62) of the missing gene families, a homologous rice locus could be identified on the raw BAC sequences. Although this number might reflect a higher degree of gene loss in rice than in Arabidopsis, this observation is most probably biased due to the

Table III. *Potential gene-loss events in Arabidopsis and rice*

Loss in	Family ID	Species[a]	Family Size[b]	Function	Probe[c]	Evolutionary Conservation[d]
Arabidopsis	152	6	11	Unknown	29729_1996.1	*Gossypium arboreum*, grasses
Arabidopsis	7748	8	13	Unknown	3635_1091.1	*Gossypium hirsutum, Glycine max, Solanum tuberosum*
Arabidopsis	10777	11	13	Unknown	29729_1339.1	*G. arboreum, G. max, Medicago truncatula*
Arabidopsis	11343	7	8	Unknown	29729_2160.1	*G. arboreum, G. max, M. truncatula*
Rice	2118	6	7	Unknown	4513_12561.1	*Hordeum vulgare, Lycopersicon esculentum*
Rice	2431	8	18	Unknown	4513_2513.1	*Triticum aestivum, H. vulgare*
Rice	3448	5	5	Oxidoreductase, 2OG-Fe(II) oxygenase family	4577_13672.1	*Zea mays*, Arabidopsis, *Beta vulgaris*
Rice	3563	6	6	Unknown	132711_1331.1	*H. vulgare, Sorghum propinquum*
Rice	3705	7	8	Unknown	4565_16492.1	*T. aestivum, Z. mays*
Rice	5304	6	6	Unknown	4577_22894.1	*Z. mays, L. esculentum, G. max*
Rice	5357	6	7	Ubiquitin family protein	4513_2676.1	*H. vulgare*, Arabidopsis, *M. truncatula*
Rice	6190	7	8	Unknown	4577_1043.1	*Z. mays*, Arabidopsis, *G. max*
Rice	6248	7	7	Unknown	4513_12850.1	*T. aestivum, H. vulgare, Z. mays*
Rice	6840	5	6	Brix domain-containing protein	4513_1319.1	*T. aestivum, H. vulgare, Z. mays*
Rice	8373	11	14	Quinolinate phosphoribosyltransferase	4558_2586.1	*T. aestivum, Z. mays, Sorghum bicolor*
Rice	9577	5	5	MA3 domain-containing protein	4513_5042.1	*T. aestivum, Z. mays, H. vulgare*
Rice	9601	8	13	Unknown	4577_2849.1	*Z. mays*, Arabidopsis, *G. max*
Rice	10255	5	6	Temperature sensing protein related	4558_5791.1	*T. aestivum, S. bicolor*, Arabidopsis
Rice	10829	11	12	Unknown	4513_7620.1	*T. aestivum, H. vulgare, Z. mays*
Rice	10975	5	5	Unknown	4513_5641.1	*T. aestivum, H. vulgare, Z. mays*
Rice	13242	11	12	Unknown	4513_2178.1	*T. aestivum, H. vulgare, Z. mays*

[a]The number of species in which the gene family was found. [b]The total number of proteins assigned to this gene family. [c]The protein ID of the probe that was used to search against the raw genomic BAC sequences (see text for details). [d]A subset of taxa where the gene family was found.

current incomplete status of the rice sequencing project. The gene families that are currently untraceable in Arabidopsis and rice are shown in Table III.

Despite the high number of publicly available protein and EST sequences for monocots that are extremely valuable for extrinsic gene prediction approaches (Mathé et al., 2002; Allen et al., 2004), these observations indicate that the current gene annotation in rice still suffers from a number of missed genes. In addition, the high number of unclustered rice genes (approximately 8,600 genes) and putative orphans currently lacking any evidence of expression (approximately 7,000 genes) indicate that further improvement and retraining of gene prediction programs, together with newly developed extrinsic gene prediction methods, seems inevitable for fully exploiting the rice genome sequence (Rouzé et al., 1999; Bennetzen et al., 2004). When compiling all results, our data provides strong evidence for the existence of 33,708 rice genes (30,993 genes organized in gene families + 704 expressed orphan genes + 2,011 unclustered genes with EST/cDNA support) when excluding 12,398 proteins resembling transposable elements (see "Materials and Methods"). Note that this is a very conservative estimation, since it has been shown that a considerable amount, up to 37% in Arabidopsis, of genes lacking EST/cDNA support do represent active genes (Yamada et al., 2003). When taking into account the large number of unclustered rice proteins that are partially homologous with other plant proteins (6,252 proteins matched other rice or plant proteins with a BLASTP E-value $<1e-05$), the estimated number of rice genes increases to 39,960. Whether this set of proteins corresponds to genuine genes or pseudogenes, as observed in other eukaryotic genomes (Mounsey et al., 2002; Torrents et al., 2003), remains to be determined.

A Closer Look at Arabidopsis and Rice

Comparing all conserved gene families between Arabidopsis and rice makes it possible to verify whether the larger number of genes in rice, as suggested in the past (Goff et al., 2002; Yu et al., 2002) and partially confirmed here, can be the consequence of gene amplification in specific families. A detailed comparison of all 5,910 gene families containing 18,461 and 22,149 genes in Arabidopsis and rice, respectively, is given in Figure 3. We found that 51% of these gene families have the same copy number in both model plants, whereas 10% of all gene families have a more than 2-fold size difference. Interestingly, the best-fit line shows that in general large gene families, containing more than 50 genes, are larger in rice

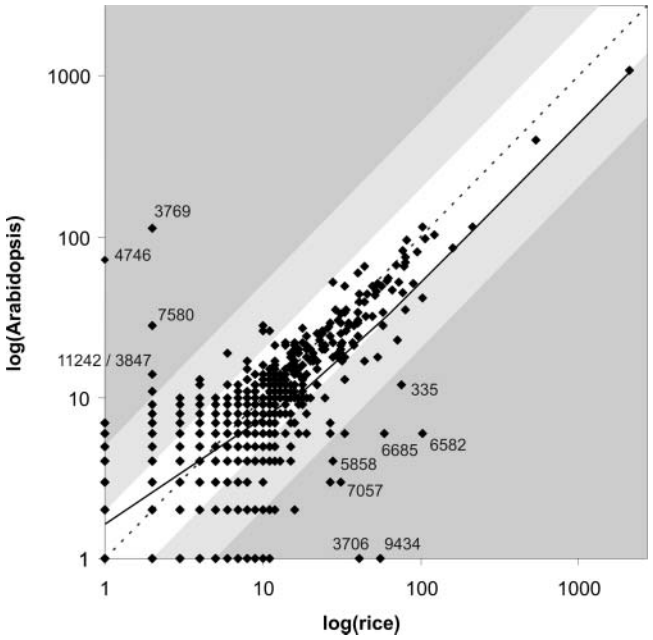

Figure 3. Size variations of all 5,910 gene families shared by Arabidopsis and rice. The position of each dot representing a gene family describes the number of genes identified in Arabidopsis and rice (abscissa and ordinate, respectively). The dotted line shows the 1:1 ratio and the black line the best-fit line ($y = 0.5399x + 1.1002$; $r^2 = 0.95$). The dark gray and light gray areas indicate a >5-fold and >2-fold size difference, respectively, whereas the white area indicates a <2-fold size difference. The gene families indicated by their family ID are: 335, F-box domain-containing protein; 3706, NB-ARC domain/disease resistance protein (CC-NBS-LRR class); 3769, disease resistance protein (TIR-NBS-LRR class); 3847, EXS family protein; 4746, kelch repeat-containing F-box family protein; 5858, chalcone synthase; 6582, putative speckle-type protein/BTB/POZ domain; 6685, unknown; 7057, glycosyl hydrolase family 18; 7580, pentatricopeptide repeat-containing protein; 9434, disease resistance protein (NBS-LRR class); and 11242, F-box family protein.

than in Arabidopsis, whereas the opposite, slightly counterbalancing pattern is observed for small gene families containing less than 5 genes (Fig. 3). Moreover, 76% of all gene families with a >5-fold size difference are bigger in rice compared with Arabidopsis. Examples of gene families that strongly vary in copy number are coding for Toll/interleukin receptor (TIR) and non-TIR nucleotide-binding site (NBS)-Leu-rich repeat (LRR) disease resistance genes (Zhou et al., 2004), Kelch repeat-containing F-box proteins, BTB/POZ domain-containing proteins, glycosyl hydrolases, and F-box family proteins (Fig. 3). Phylogenetic analysis on a subset of gene families with a higher copy number in rice than in Arabidopsis indicates that they have expanded significantly in rice, after the divergence of monocots and eudicots from their last common ancestor (Supplemental Fig. 2). The expansion of the chalcone synthase family in rice, a catalyst in the first steps of flavonoid biosynthesis, might reflect an adaptive strategy in its evolution because previous analyses have reported the extensive differentiation in gene expression among duplicate copies of chalcone synthase genes (Durbin et al., 2000). Likewise, the expansion of receptor-like kinases involved in defense and disease control in rice, for which we observe a >1.9 size difference, offers advanced sensing toward diverse extracellular signals (Shiu et al., 2004). Similar patterns of gene family expansion were also observed in gene families that are larger in Arabidopsis than in rice, which suggests that the extension of gene families through gene duplication is a more common phenomenon in higher plants than massive reduction through gene loss. The presence of a number of large gene families with similar copy numbers in both plant model systems, such as gene families covering transcription factors, transporter proteins, cytochrome P450s, and phosphatases, corresponds with previously reported findings (Goff et al., 2002).

Apart from analyzing the conserved gene families between Arabidopsis and rice, we also examined the distribution of gene families containing Arabidopsis or rice genes over a wider range of plant species using the high-level phylogenetic profiles (see above). Although 69% of the gene families in grasses is also present in eudicots, 3,006 gene families are unique to the grasses, of which 42% represent grass-specific families found in multiple cereals. These results correspond with previous estimates of putative monocot-specific genes using sugarcane (*Saccharum officinarum*) ESTs (Vincentz et al., 2004). In addition, we found that 11% of all families present in the grasses with homologs in eudicots were absent in Arabidopsis, which confirms our findings that gene loss in specific lineages or species is common. This number is considerably higher than the 2% of sugarcane sequences that matched homologous non-Arabidopsis eudicot sequences and is most probably caused by the higher number of eudicotyledonous species used here, compared with the analysis of Vincentz et al. (2004). The reverse query indicates that also 11% of all families conserved between monocots and eudicots are absent in rice, suggesting that the amount species-specific gene loss in monocots and eudicots is very similar. Although the overall evolutionary distributions of gene families are very similar for Arabidopsis and rice (Fig. 3; Supplemental Fig. 3), the number of rice-specific gene families (and genes) is approximately 2- to 3-fold larger for rice than for Arabidopsis (see above). Thus, this set of genes, together with the set of orphan genes, also accounts for the larger number of genes currently found in rice than in Arabidopsis. Finally, the fact that 914 gene families are detected solely in the fully sequenced genomes of Arabidopsis and rice indicates that a fraction of plant gene functions is currently not covered by gene discovery efforts through EST sequencing.

Conclusion

Recent estimates show that approximately 43,000 plant protein sequences are known, which can be

classified into approximately 4,053 gene families (Mohseni-Zadeh et al., 2004). Although an enormous amount of ESTs are publicly available for a variety of plant species, these sequences only represent partial information about transcribed genes and lack annotated coding sequence information. Consequently, phylogenetic analysis of plant genes and gene families based on protein information combined with manual addition of homologous plant ESTs is very time consuming and has an overall low success rate. Analysis of the data set described here suggests that approximately 19,300 different gene functions (i.e. 12,854 full-length gene families + 6,482 expressed orphans) exist in the green plant lineage. When all gene families covering partial proteins are discarded, 9,355 gene families are found in the eudicots, of which 89% are found in multiple species, with an additional 2,353 expressed orphan genes. Similarly, 9,967 gene families have been detected in the grasses, of which 82% are found in multiple species, together with 2,084 expressed orphan genes for specific cereals. These numbers suggest that the total number of gene functions in monocots and eudicots is comparable and seems to indicate that a substantial portion of the recently described rice genes are anomalous sequences representing incorrect gene predictions or pseudogenes (Goff et al., 2002; Yu et al., 2002; Bennetzen et al., 2004; Jabbari et al., 2004). Nevertheless, a significant difference in copy number between Arabidopsis and rice was uncovered for a subset of large gene families, SSGFs, and orphan genes, confirming the larger number of genes in rice compared to Arabidopsis. Clearly, the large number of expressed orphans, together with numerous examples of SSGFs and LSGFs, complemented with the observations of gene loss in Arabidopsis and rice, illustrates the high plasticity of plant genomes.

MATERIALS AND METHODS

Construction of the Data Set

The data set consists of two subsets, one including publicly available plant proteins and the other containing EST sequences. The protein data set covers data extracted from EMBL (Kulikova et al., 2004) for 30 different plant species, whereas the EST set contains data of more than 2.8 million ESTs for these plant species. Sequence information for Arabidopsis (*Arabidopsis thaliana*) and rice (*Oryza sativa*), for which a nuclear genome sequence is available, was obtained from TIGR (Arabidopsis release 5 from January 2004; Wortman et al., 2003; rice release 2 April 2004; Yuan et al., 2003). If multiple protein sequences were available for the same locus, the protein of the first gene model was retained. Altogether 102 Arabidopsis proteins with similarity to known plant transposable elements (BLASTP E-value $<1e-05$ with Swiss-Prot transposable elements) were not retained for further analysis. For rice, all 12,398 proteins with the gene description "transposon" or "retrotransposon" were discarded.

EST sequences were transformed into EST clusters (also called unigene or tentative consensus) and a set of singleton ESTs with the EST clustering software developed by TIGR (Pertea et al., 2003). ESTs were clustered and assembled in such a way that paralogous gene sequences should be maintained as such and not merged into a single chimeric EST cluster. To this end, conservative parameters were applied (minimum percent identity 99%, minimum length of overlap 50 bp, and a maximum mismatched overhang of 20 bp), which are more stringent than those of TIGR Gene Indices or National Center for Biotechnology Information Unigenes (for a detailed comparison of these and others EST assembly efforts, see Parkinson et al., 2002). All mRNA sequences of all genes in the protein data set were also incorporated during the EST clustering. As a consequence, all ESTs perfectly matching an existing plant mRNA were remapped to one gene sequence, avoiding inclusion of redundancy in the data set. Similarly, redundant genes in the protein layer were removed because identical mRNAs were merged into a single gene sequence.

Next, putative ORFs were delineated for all EST clusters. For these EST clusters containing experimentally derived mRNAs, the corresponding coding sequence (CDS) information was retained. For all other sequences, the coding frame and putative CDS were determined with the FrameD software tool (Schiex et al., 2003). When validating the FrameD software against a subset of mRNAs from the protein set from different species, its overall sensitivity was good (85% for mRNAs with an EMBL CDS annotation using the Arabidopsis Interpolated Markov Model [IMM]), but rather low for species without a specific IMM, such as Chlamydomonas and Pinus (2% and 63%, respectively). Because different plants have different codon usages and only a limited number of plant IMMs is available in FrameD, additional IMMs were required to have a good overall ORF detection sensitivity not biased toward particular plant species. Therefore, we first created training sets for each plant species for which no IMM was available, based on the annotated CDS of mRNAs present in the EMBL database. After a careful evaluation of the available FrameD IMMs on the training sequences from different plants and a detailed comparison of the codon usage in the 30 plant species under investigation (data not shown), we constructed five new IMMs (one for Chlamydomonas, Physcomitrella, the Pinaceae, the Asterids, and the Caryophyllales). Note that not all new models are species specific because some models were built with sequences from several closely related plant species (see Supplemental Table III). Finally, for each plant species, ORFs were determined on the EST clusters with FrameD using a specific IMM (Table I; additional parameters $-E$ for eukaryotic EST analysis and $-C$ for correcting frameshifts; Schiex et al., 2003). Only putative ORFs with a minimal length of 50 codons were retained.

All translated coding sequences of the EST clusters and all sequences from the protein data set were used to construct gene families by applying sequence-based protein clustering (Li et al., 2001). First, an all-against-all sequence comparison was performed using BLASTP (Altschul et al., 1997), and relevant hits were retained (Li et al., 2001). Briefly, this method considers two proteins as being homologous only when they share a substantially conserved region on both molecules with a minimum amount of sequence identity. In this manner, homology based on the partial overlap of single protein domains between two multidomain proteins, which occasionally leads to significant E-values in BLAST, is not retained. The proportion of identical amino acids in the aligned region between the query and target sequence is recalculated to $I' = I \times \min(n_1/L_1, n_2/L_2)$, where L_i is the length of sequence i and n_i is the number of amino acids in the aligned region of sequence i. This value I' is then used in the empirical formula for protein clustering proposed by Rost (1999). These additional criteria prevent that partial ORFs derived from two EST clusters, which in reality originated from the same gene, were counted as two distinct family members. Finally, all valid homologous protein pairs (e.g. protein A is homologous to protein B, protein B is homologous to protein C) were subject to a simple-linkage clustering routine to delineate protein gene families (for example, family with proteins A, B, and C). In total, more than 39 million BLAST hits were evaluated, and more than 6.4 million valid homologous protein pairs were used for delineating the gene families. An evaluation of Li's method (Li et al., 2001) applied on yeast sequences showed that it behaves equally well compared to other automatic protein clustering algorithms (Yang et al., 2003). Although one might argue that by using this method partial proteins will be split from their complete homologous counterparts (see below), we prefer this conservative clustering approach because a less stringent protein clustering would lead to the creation of superfamilies, obscuring every pattern of evolutionary conservation for a specific gene function. Additional information about different protein clustering strategies that were evaluated can be found at http://www.psb.ugent.be/bioinformatics/.

GO Functional Annotation

GO gene associations for Arabidopsis proteins were retrieved from TIGR (ftp.tigr.org/pub/data/a_thaliana/ath1/DATA_RELEASE_SUPPLEMENT/) and remapped to the generic GO Slim classification scheme (ftp.geneontology.org/pub/go/GO_slims/goslim_generic.go) with the Perl script map2slim.pl (available at www.geneontology.org).

Analysis of Gene Families Consisting of Partial Proteins

Throughout this analysis, we assumed that Arabidopsis and rice genes derived from the genome sequencing projects represented full-length proteins. Given the fact that the family delineation algorithm does not create family relationships between homologous proteins that vary extremely in length (i.e. that lack global homology), we believe that gene families including Arabidopsis and rice proteins will generally not contain clustered partial proteins. These full-length families represent the majority of all gene families (i.e. 68% of all 14,639 gene families). We obtained 4,341 gene families without Arabidopsis and/or rice homologs that might contain partial proteins (designated partial protein families [PPFs]). For each of the 14,369 gene families, a random gene representative was selected and compared with all other gene representatives. Subsequently, all significant similarities (BLASTP E-value <1e−15) between genes representing full-length families and PPFs were scored. Finally, we identified these PPFs that were significantly shorter than the homologous full-length family. We found 1,415 and 1,515 PPFs that were more than 50% and more than 30% shorter than the homologous full-length family, respectively. To reduce the chance of overpredicting the final number of gene families, we selected the 1,515 gene families that were at least 30% shorter than their full-length counterpart as gene families consisting of partial proteins. These families were discarded when the number of gene families in the different lineages is discussed (Fig. 2). Applying other E-value similarity and length-difference cutoffs yielded similar results (data not shown).

Analysis of Orphans, SSGFs, and LSGFs

All orphan proteins or proteins of gene families specific for one plant species or lineage were compared against the full set of proteins using less stringent criteria (BLASTP E-value <1e−05) compared to the criteria applied by the protein clustering algorithm for delineating gene families (see above). These proteins without non-self BLAST hits (i.e. only hitting themselves) were designated orphans, whereas only those genes uniquely matching proteins of the same species or lineage were retained as species or lineage specific, respectively.

Distribution of Materials

Upon request, all novel materials described in this publication will be made available in a timely manner for noncommercial research purposes, subject to the requisite permission from any third-party owners of all or parts of the material. Obtaining any permission will be the responsibility of the requestor.

ACKNOWLEDGMENTS

We thank M. De Cock for help with the manuscript and F. Dierick for technical assistance.

Received October 11, 2004; returned for revision November 10, 2004; accepted November 10, 2004.

LITERATURE CITED

Allen JE, Pertea M, Salzberg SL (2004) Computational gene prediction using multiple sources of evidence. Genome Res **14:** 142–148

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

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

Bennetzen JL, Coleman C, Liu R, Ma J, Ramakrishna W (2004) Consistent over-estimation of gene number in complex plant genomes. Curr Opin Plant Biol **7:** 732–736

Dong Q, Schlueter SD, Brendel V (2004) PlantGDB, plant genome database and analysis tools. Nucleic Acids Res **32** (Database issue)**:** D354–D359

Doyle JJ, Gaut BS (2000) Evolution of genes and taxa: a primer. Plant Mol Biol **42:** 1–23

Durbin ML, McCaig B, Clegg MT (2000) Molecular evolution of the chalcone synthase multigene family in the morning glory genome. Plant Mol Biol **42:** 79–92

Ermolaeva MD, Wu M, Eisen JA, Salzberg SL (2003) The age of the *Arabidopsis thaliana* genome duplication. Plant Mol Biol **51:** 859–866

Feng Q, Zhang Y, Hao P, Wang S, Fu G, Huang Y, Li Y, Zhu J, Liu Y, Hu X, et al (2002) Sequence and analysis of rice chromosome 4. Nature **420:** 316–320

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

Griffiths S, Dunford RP, Coupland G, Laurie DA (2003) The evolution of CONSTANS-like gene families in barley, rice, and Arabidopsis. Plant Physiol **131:** 1855–1867

Grossman AR, Harris EE, Hauser C, Lefebvre PA, Martinez D, Rokhsar D, Shrager J, Silflow CD, Stern D, Vallon O, et al (2003) *Chlamydomonas reinhardtii* at the crossroads of genomics. Eukaryot Cell **2:** 1137–1150

Gutierrez RA, Green PJ, Keegstra K, Ohlrogge JB (2004) Phylogenetic profiling of the *Arabidopsis thaliana* proteome: What proteins distinguish plants from other organisms? Genome Biol **5:** R53

Jabbari K, Cruveiller S, Clay O, Le Saux J, Bernardi G (2004) The new genes of rice: a closer look. Trends Plant Sci **9:** 281–285

Kevei Z, Vinardell JM, Kiss GB, Kondorosi A, Kondorosi E (2002) Glycine-rich proteins encoded by a nodule-specific gene family are implicated in different stages of symbiotic nodule development in *Medicago* spp. Mol Plant Microbe Interact **15:** 922–931

Kinoshita T, Fukuzawa H, Shimada T, Saito T, Matsuda Y (1992) Primary structure and expression of a gamete lytic enzyme in *Chlamydomonas reinhardtii*: similarity of functional domains to matrix metalloproteases. Proc Natl Acad Sci USA **89:** 4693–4697

Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, Makarova KS, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, et al (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol **5:** R7

Kriventseva EV, Biswas M, Apweiler R (2001) Clustering and analysis of protein families. Curr Opin Struct Biol **11:** 334–339

Kulikova T, Aldebert P, Althorpe N, Baker W, Bates K, Browne P, van den Broek A, Cochrane G, Duggan K, Eberhardt R, et al (2004) The EMBL Nucleotide Sequence Database. Nucleic Acids Res **32** (Database issue)**:** D27–D30

Li WH, Gu Z, Wang H, Nekrutenko A (2001) Evolutionary analyses of the human genome. Nature **409:** 847–849

Martin W (2003) Gene transfer from organelles to the nucleus: frequent and in big chunks. Proc Natl Acad Sci USA **100:** 8612–8614

Mathé C, Sagot MF, Schiex T, Rouzé P (2002) Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res **30:** 4103–4117

Mergaert P, Nikovics K, Kelemen Z, Maunoury N, Vaubert D, Kondorosi A, Kondorosi E (2003) A novel family in *Medicago truncatula* consisting of more than 300 nodule-specific genes coding for small, secreted polypeptides with conserved cysteine motifs. Plant Physiol **132:** 161–173

Mohseni-Zadeh S, Louis A, Brezellec P, Risler JL (2004) PHYTOPROT: a database of clusters of plant proteins. Nucleic Acids Res **32** (Database issue)**:** D351–D353

Mounsey A, Bauer P, Hope IA (2002) Evidence suggesting that a fifth of annotated *Caenorhabditis elegans* genes may be pseudogenes. Genome Res **12:** 770–775

Nagaki K, Cheng Z, Ouyang S, Talbert PB, Kim M, Jones KM, Henikoff S, Buell CR, Jiang J (2004) Sequencing of a rice centromere uncovers active genes. Nat Genet **36:** 138–145

Nishiyama T, Fujita T, Shin IT, Seki M, Nishide H, Uchiyama I, Kamiya A, Carninci P, Hayashizaki Y, Shinozaki K, et al (2003) Comparative genomics of *Physcomitrella patens* gametophytic transcriptome and *Arabidopsis thaliana*: implication for land plant evolution. Proc Natl Acad Sci USA **100:** 8007–8012

Parkinson J, Guiliano DB, Blaxter M (2002) Making sense of EST sequences by CLOBBing them. BMC Bioinformatics **3:** 31

Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B, et al (2003) TIGR Gene Indices

clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics **19:** 651–652

Pryer KM, Schneider H, Zimmer EA, Ann Banks J (2002) Deciding among green plants for whole genome studies. Trends Plant Sci **7:** 550–554

Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana R, White J (2001) The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. Nucleic Acids Res **29:** 159–164

Raes J, Vandepoele K, Simillion C, Saeys Y, Van de Peer Y (2003) Investigating ancient duplication events in the *Arabidopsis* genome. J Struct Funct Genomics **3:** 117–129

Rice Chromosome 10 Sequencing Consortium (2003) In-depth view of structure, activity, and evolution of rice chromosome 10. Science **300:** 1566–1569

Rost B (1999) Twilight zone of protein sequence alignments. Protein Eng **12:** 85–94

Rouzé P, Pavy N, Rombauts S (1999) Genome annotation: which tools do we have for it? Curr Opin Plant Biol **2:** 90–95

Rudd S (2003) Expressed sequence tags: alternative or complement to whole genome sequences? Trends Plant Sci **8:** 321–329

Sasaki T, Matsumoto T, Yamamoto K, Sakata K, Baba T, Katayose Y, Wu J, Niimura Y, Cheng Z, Nagamura Y, et al (2002) The genome sequence and structure of rice chromosome 1. Nature **420:** 312–316

Schiex T, Gouzy J, Moisan A, de Oliveira Y (2003) FrameD: a flexible program for quality check and gene prediction in prokaryotic genomes and noisy matured eukaryotic sequences. Nucleic Acids Res **31:** 3738–3741

Shahmuradov IA, Akbarova YY, Solovyev VV, Aliyev JA (2003) Abundance of plastid DNA insertions in nuclear genomes of rice and *Arabidopsis*. Plant Mol Biol **52:** 923–934

Shewry PR, Halford NG (2002) Cereal seed storage proteins: structures, properties and role in grain utilization. J Exp Bot **53:** 947–958

Shiu SH, Karlowski WM, Pan R, Tzeng YH, Mayer KF, Li WH (2004) Comparative analysis of the receptor-like kinase family in Arabidopsis and rice. Plant Cell **16:** 1220–1234

Soltis DE, Soltis PS (2003) The role of phylogenetics in comparative genetics. Plant Physiol **132:** 1790–1800

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, et al (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics **4:** 41

Timmis JN, Ayliffe MA, Huang CY, Martin W (2004) Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet **5:** 123–135

Torrents D, Suyama M, Zdobnov E, Bork P (2003) A genome-wide survey of human pseudogenes. Genome Res **13:** 2559–2567

Vandepoele K, Simillion C, Van de Peer Y (2003) Evidence that rice and other cereals are ancient aneuploids. Plant Cell **15:** 2192–2202

Vincentz M, Cara FA, Okura VK, da Silva FR, Pedrosa GL, Hemerly AS, Capella AN, Marins M, Ferreira PC, Franca SC, et al (2004) Evaluation of monocot and eudicot divergence using the sugarcane transcriptome. Plant Physiol **134:** 951–959

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

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

Yang J, Lusk R, Li WH (2003) Organismal complexity, protein complexity, and gene duplicability. Proc Natl Acad Sci USA **100:** 15661–15665

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

Yuan Q, Ouyang S, Liu J, Suh B, Cheung F, Sultana R, Lee D, Quackenbush J, Buell CR (2003) The TIGR rice genome annotation resource: annotating the rice genome and creating resources for plant biologists. Nucleic Acids Res **31:** 229–233

Zhou T, Wang Y, Chen JQ, Araki H, Jing Z, Jiang K, Shen J, Tian D (2004) Genome-wide identification of NBS genes in japonica rice reveals significant expansion of divergent non-TIR NBS-LRR genes. Mol Genet Genomics **271:** 402–415