Original Paper

KNApSAcK gene classification system for *Arabidopsis thaliana*: Comparative genomic analysis of unicellular to seed plants

Hiroki Takahashi[1], Mai Kawazoe[1], Masayoshi Wada[1], Aki Hirai[1,3], Kensuke Nakamura[1], Md. Altaf-Ul-Amin[1], Yuji Sawada[2,3], Masami Yokota Hirai[2,3], Shigehiko Kanaya[1,3,*]

[1] Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Nara 630-0101, Japan; [2] RIKEN Plant Science Center, Yokohama, Kanagawa 230-0045, Japan; [3] JST, CREST, Kawaguchi, Saitama 332-0012, Japan

* E-mail: skanaya@gtc.naist.jp Tel: +81-743-72-5952 Fax: +81-743-72-5953

Received September 30, 2009; accepted October 21, 2009 (Edited by K. Saito)

Abstract Proper functional classification and statistical assessment of a set of genes is very important for the purpose of comparison of gene compositions in genomes between different plant species as well as for the post-genomic research such as assessment of tissue and cell conditions concerning gene expression and metabolite accumulation profiles in transcriptomics and metabolomics. So we defined five-level categories concerning *Arabidopsis thaliana* genes by surveying approximately 3,000 references and classified 14,525 of 27,677 genes into different categories. Based on the classification information accumulated from various sources, we have developed a software tool called 'Arabidopsis Gene Classifier'. By using this classifier system, we performed the comparative genomic analyses of five genome sequences of the plants, *Chlamydomonas reinhardtii*, *Physcomitrella patens*, *Selaginella moellendorffii*, *A. thaliana* and *Oryza sativa*, and extracted statistically significant differences in their gene compositions concerning metabolic pathways.

Key words: Gene function classifier, comparative genomics, metabolic pathway.

Biological research has transformed from a relatively poor discipline into one that now is data rich, mainly because of dramatic advances of high-throughput experimental technologies (Joyce and Palsson 2006). After determination of the *Arabidopsis thaliana* genome, in the post-genomic research such as transcriptomics and metabolomics, the systematic understanding of gene expression and metabolite accumulation in tissues and cells is an essential issue. To attain this purpose, several tools have been developed for function prediction based on gene co-expression such as ATTED-II (Obayashi et al. 2007), AthCoR@CSB.DB (Steinhauser et al. 2004), Genevestigator (Zimmermann et al. 2004) and KAGIANA (Aoki et al. 2007). Chervits et al. (Chervits et al. 1999) reported the first comparison between two complete eukaryotic genomes, i.e., budding yeast and worm. Biological roles of about 12% of the worm gene encoded proteins could be speculated from only sequence similarity to yeast genes. Identification of sequence and functional conservation across many organisms is useful for comprehensive understanding of cell biology. Availability of the genome sequences of more plants makes it possible to address some of the major questions regarding plants by comparative genomic analysis (Bowman et al. 2007; Merchant et al. 2007).

The Gene Ontology Consortium (Ashburner et al. 2000) has developed gene classification terms. The number of terms associated with GO in *A. thaliana* (TAIR. http://www.arabidopsis.ogr/) is, however, large (about 4,000). Out of those, some terms are obscure for biological researchers. So in order to understand gene functions more deeply and easily, we defined five-level categories concerning all *A. thaliana* genes by surveying approximately 3,000 references and classified 14,525 of 27,677 genes into different functional categories. Based on the classification information, we have developed a software tool called 'Arabidopsis Gene Classifier' for automatic functional classification.

A lot of genome sequencing projects are ongoing and one of the projects is for the lycophyte *Selaginella moellendorffii*, which has potential for unraveling several questions in plant evolution because this plant is considered to be ancestral or primitive for vascular plants. So by using the aforementioned classifier system, we performed the comparative genomic analyses of five genome sequences of the plants, *Chlamydomonas reinhardtii* (green algae), *Physcomitrella patens* (mosses), *S. moellendorffii* (club-mosses), *A. thaliana* (eudicots), and *Oryza sativa* (monocots).

This article can be found at http://www.jspcmb.jp/

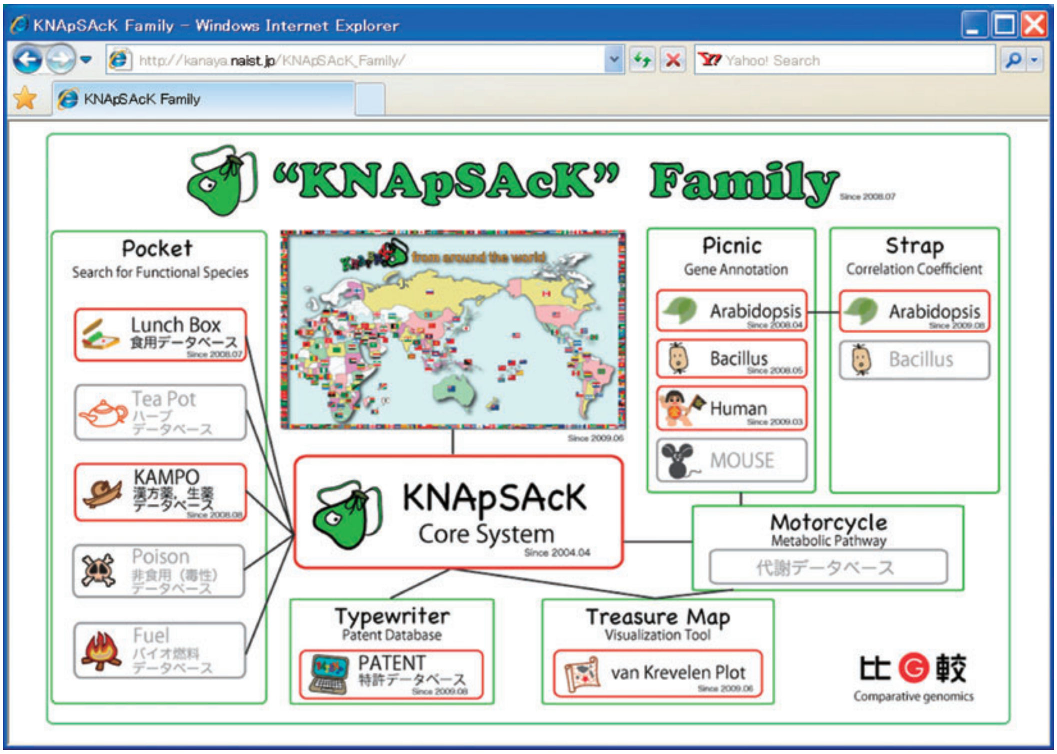

Figure 1. The main window of KNApSAcK family (http://kanaya.naist.jp/KNApSAcK_Family/).

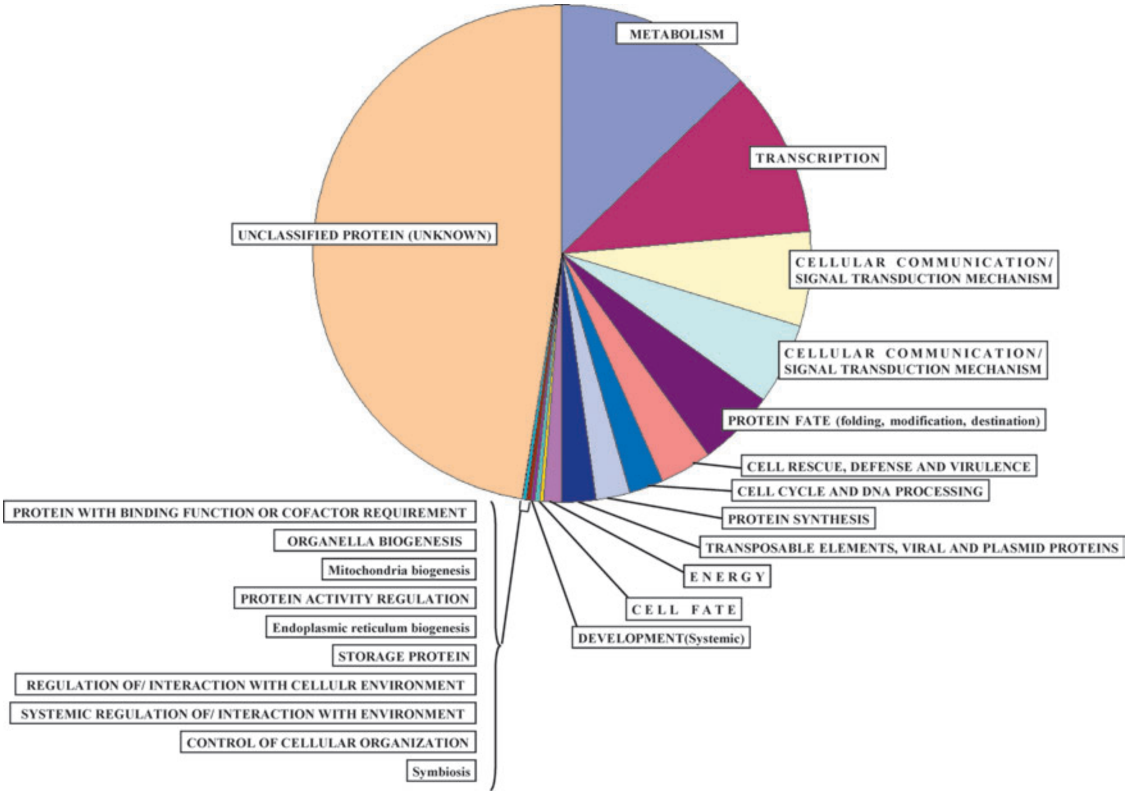

Figure 2. The pie chart showing the proportions of all genes corresponding to 23 Gene Classifier terms. 'METABOLISM' (12.7%), 'TRANSCRIPTION' (10.8%), 'CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM' (6.2%), 'TRANSPORT FACILITATION' (5.3%), 'PROTEIN FATE (folding, modification, destination)' (5.0%), 'CELL RESCUE, DEFENSE AND VIRULENCE' (3.3%), 'CELL CYCLE AND DNA PROCESSING' (2.3%), 'PROTEIN SYNTHESIS' (2.2%), 'TRANSPOSABLE ELEMENTS, VIRAL AND PLASMID PROTEINS' (2.2%), 'ENERGY' (1.2%), 'CELL FATE' (0.4%), 'DEVELOPMENT (Systemic)' (0.3%), 'Symbiosis' (0.2%), 'CONTROL OF CELLULAR ORGANIZATION' (0.2%), 'SYSTEMIC REGULATION OF/INTERACTION WITH ENVIRONMENT' (0.1%), 'REGULATION OF/INTERACTION WITH CELLULAR ENVIRONMENT' (0.1%), 'STORAGE PROTEIN' (0.1%), 'Endoplasmic reticulum biogenesis' (0.1%), 'PROTEIN ACTIVITY REGULATION' (<0.1%), 'Mitochondria biogenesis' (<0.1%), 'ORGANELLA BIOGENESIS' (<0.1%), 'PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic)' (<0.1%), 'UNCLASSIFIED PROTEINS' (47.5%).

Materials and methods

Data set

Five genome sequences, i.e., *C. reinhardtii* (Merchant et al. 2007), *P. patens* (Rensing et al. 2008), *S. moellendorffii* (http://www.jgi.doe.gov/), *A. thaliana* (TAIR), and *O. sativa* (IRGSP 2005), were used for the comparative genomic analyses in the present work.

Homology search

BLASTP (Altschul et al. 1997) analyses of all protein sequences for four organisms except *A. thaliana* were performed against the reference data set, i.e. the proteins of *A. thaliana* and *vice versa* to extract best-hit gene pairs. For first screening, all BLASTP results were filtered by e-value ≦1.0E-8, hit length coverage ≧60% of both a query and a reference sequence, and identity ≧60% of hit length. Finally, only reciprocal best-hit gene pairs were extracted. The best-hit genes of four organisms were assigned to gene function categories using Arabidopsis Gene Classifier (http://kanaya.naist.jp/GeneClassifier/top.jsp?fn=arabidopsis).

KNApSAcK family

Figure 1 shows the main window of KNApSAcK family (http://kanaya.naist.jp/KNApSAcK_Family/), which consists of eight parts. Species-metabolite relations can be retrieved by 'KNApSAcK Core System' which accumulated 80,029 species-metabolite relations comprised of 40,087 metabolites (6th Oct., 2009). 'Pocket' includes search systems for relationships between species and metabolites related to human life, such as 'Lunch Box' (edible plants in Japan), 'Tea Pot' (herb teas, in progress), 'KAMPO' (traditional Japanese medicines), 'Poison' (poisonous plants, in progress), and 'Fuel' (bio-fuel resources, in progress). Relationships between medicinal/edible plants and countries that utilize those plants are available in 'KNApSAcK from around the world' (at the top center, with the world map), which has accumulated 7,356 pairwise relationships between 119 countries and 4,538 medicinal/edible plants from records of scientific literatures (16th Sep., 2009). 'Typewriter' includes patent information concerning plants. 'Picnic' includes gene classification system for four species, i.e., *A. thaliana*, *Bacillus subtilis*, *Homo sapiens*, and *Mus musculus* (in progress). Gene function categories with five-level hierarchical structure based on references have been manually curated. Co-expressed genes for a target gene set can be retrieved by 'Strap' for two species, i.e., *A. thaliana* and *B. subtilis* (in progress). 'Treasure Map' is a visualization tool of metabolites based on van Krevelen Plot. Metabolic reactions are planned to be arranged in 'Motorcycle'.

The pie chart of Figure 2 shows the proportions of all *A. thaliana* genes corresponding to 23 main functional categories defined in our classification system, called 'Arabidopsis Gene Classifier' (there are some genes belonging to multi categories), i.e., 'UNCLASSIFIED PROTEINS' (47.5%), 'METABOLISM' (12.7%), 'TRANSCRIPTION' (10.8%), 'CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM' (6.2%), 'TRANSPORT FACILITATION' (5.3%), 'PROTEIN FATE (folding, modification, destination)' (5.0%) and etc. All gene function categories defined in the present study are listed in Supplemental Table 1.

Statistical analysis

Overrepresented and underrepresented gene function categories associated with a target gene set can be identified by Fisher's exact test. The one-tailed Fisher's exact *p*-values corresponding to overrepresentation of categories were calculated based on counts in 2×2 contingency tables. Counts n_{11}, n_{12}, n_{21}, and n_{22} of the contingency table are as follows: n_{11}, number of observations of a particular category in the target gene set; n_{12}, number of other categories in the target gene set; n_{21}, number of observations of the particular category in background gene set; and n_{22}, number of observations of other categories in the background gene set. The sums of n_{11}, n_{12}, and n_{21}, n_{22} are equal to number of top hit genes against *A. thaliana*, and number of query sequences of *A. thaliana* except top hit genes, respectively. Fisher's exact *p*-values were corrected to FDR *p*-values (Benjamini and Hochberg 1995).

Results and discussion

Comparative genomic analyses

To compare the overall sequence similarities among the genes of five plants, i.e., *C. reinhardtii*, *P. patens*, *S. moellendorffii*, *A. thaliana*, and *O. sativa*, we performed reciprocal BLASTP analyses and extracted top-hit genes against *A. thaliana* as discussed in detail in the Materials and methods section. Table 1 shows the summary of the results. The numbers of top-hit genes in *C. reinhardtii*, *P. patens*, *S. moellendorffii*, and *O. sativa* against *A. thaliana* were 585 (3.5%), 2,431 (6.8%), 2,428 (7.0%), 4,478 (16.6%), respectively. According to the fraction of top-hit genes out of total genes (query sequences) of each organism, *C. reinhardtii* is the farthest from *A. thaliana* among four plants and *O. sativa* is the closest to *A. thaliana*. *P. patens* and *S. moellendorffii* have similar percentage of top-hit genes and are located at the middle of *C. reinhardtii* and *O. sativa*. These results are consistent with the phylogenetic relationships among plants (Bowman et al. 2007).

Despite millions of years of evolution, the lycophytes (including *S. moellendorffii*) have retained many developmental features thought to be ancestral or primitive for vascular plants (Bowman et al. 2007). So, we compared top-hit genes in *S. moellendorffii* to those of *P. patens*. We found 1,659 common genes, 772 *P. patens* specific genes, and 769 *S. moellendorffii* specific genes (in Figure 3), suggesting that 772 genes of *P. patens* and *A. thaliana* might have derived from the common ancestor and needed only for the bryophytes and euphyllophytes but not for the lycophytes. In a similar way, 769 genes of *S. moellendorffii* and *A. thaliana* might be needed only for the lycophytes and euphyllophytes but not for the bryophytes. Out of 1,659 common genes, 429 genes were annotated to 'UNCLASSIFIED PROTEIN' category, which has no

Table 1. Summary of the BLASTP analyses.

Organism	Size (Mb)	# of query sequences	# of hit genes against *A. thaliana*
Chlamydomonas reinhardtii	121	16,888	585
Physcomitrella patens	480	35,938	2,431
Selaginella moellendorffii	~100	34,697	2,428
Oryza sativa	371.4	26,938	4,478
Arabidopsis thaliana	119.2	33,410	—

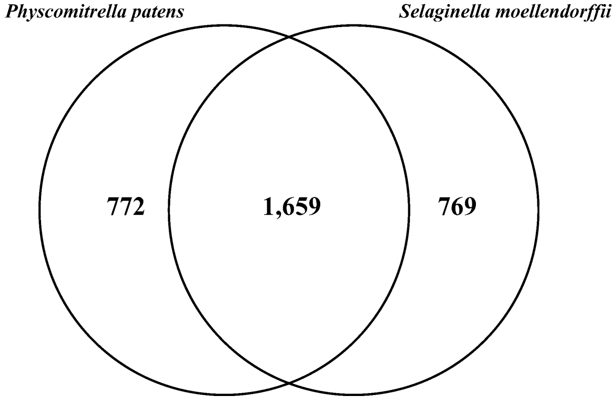

Figure 3. The Venn diagram of top-hit genes in *P. patens* and *S. moellendorffii* against *A. thaliana*. 1,659, 772 and 769 genes are common, specific for *P. patens* and *S. moellendorffii*, respectively.

homologous genes associated with known functions to date, indicating that these might play important roles across the bryophytes, lycophytes and euphyllophytes.

Transcription factors are important regulators of gene expression. A superfamily of transcription factors is MYB superfamily, which has the largest number of members compared to any *A. thaliana* gene family (Riechmann and Ratcliffe 2000; Iida et al. 2005). Out of top-hit genes in *P. patens*, three genes, i.e., gw1.65.181.1, e_gw1.61.201.1 and gw1.106.41.1, correspond to genes of MYB superfamily, At1g35515 (AtMYB8), At1g66380 (AtMYB114) and At2g31180 (AtMYB14), respectively. In *S. moellendorffii*, one gene, i.e., e_gw1.86.192.1, correspond to At5g52600 (AtMYB82). PlnTFDB (Riano-Pachon et al. 2007) indicates that *P. patens* an *S. moellendorffii* have 62 and 22 genes of MYB family, respectively. Our results suggest that those four MYB genes might have important roles in plants because of higher sequence similarities than other MYB genes.

The characteristics of four plants based on gene classifier system

To statistically estimate the functional differences of conserved genes among four plants, we used 720 third-level categories in our classification system. Fisher's exact test based on 2×2 contingency table was performed for four sets of top-hit genes. The threshold of FDR corrected *p*-value was set to 1.00E-2. Tables 2 (a)–(d) show all significant overrepresented categories. The numbers of significant categories in *C. reinhardtii*, *P. patens*, *S. moellendorffii*, and *O. sativa* against *A. thaliana* were 11, 36, 32, and 37, respectively. Eight categories were overrepresented in all results, e.g., 'ribosomal protein', 'intermediately carbon metabolism', 'small nuclear ribonucleoprotein (snRNP)', 'branched chain amino acids from aspartate', and 'isoprenoid biosynthesis', suggesting that genes associated with these categories are highly conserved among plants and probably play the important roles across all plants. Supplemental Table 2 shows GO terms of 'biological process' (TAIR) overrepresented in four species by the same way, indicating that the defined orders of hierarchical structure about terms are different, and some terms are obscure for understanding biological functions concerned with genes.

According to Rensing et al. (Rensing et al. 2004), genes associated with the biosynthetic pathways of carotenoids were conserved in *C. reinhardtii*, *P. patens* and *A. thaliana* and paralog frequencies of those genes in *P. patens* were the highest. On the other hand, our analyses detected only three genes (At3g04870 (ZDS), At4g14210 (PDS) and At5g17230 (PSY)) in 'Carotenoid biosynthesis' in *C. reinhardtii*, while in addition to those three genes, 8, 7 and 13 genes were detected in *P. patens*, *S. moellendorffii*, and *O. sativa*, respectively. These three genes are related to the pathway from geranylgeranyl diphophate (GGPP) to lycopene, implying that out of many reactions in the biosynthetic pathway of carotenoids, the reactions from GGPP to lycopene catalyzed by PSY, PDS and ZDS should be essential and necessary across all plants. In the context of evolutionary process of plants, it can be said that the biosynthetic pathway of carotenoids might have started from these reactions.

Conclusion

We defined five-level gene function categories concerning *A. thariana* genes by surveying approximately 3,000 references and classified 14,525 of 27,677 genes into different functional categories. Based on the classification information, we have developed a freely available software tool called 'Arabidopsis Gene Classifier' for automatic functional classification of a set of target genes. Comparative genomic analyses of five plants by using the classification system revealed that the reactions from GGPP to lycopene catalyzed by PSY, PDS and ZDS in the biosynthetic pathway of carotenoids are essential and necessary across all five plants. Taken together, the gene classifier system is useful for estimating gene functions not only for *A. thaliana* but also for other organisms based on sequence similarity. The more the genomes of different plants will be

Table 2. Gene function categories overrepresented in (a) *C. reinhardtii*, (b) *P. patens*, (c) *S. moellendorffii*, and (d) *O. sativa* (FDR *p*-value<0.01).

(a) *C. reinhardtii* (Overrepresentation)

FDR *p*-Value	1st-level category	2nd-level category	3rd-level category
6.21E-37	**PROTEIN SYNTHESIS**	**translation**	**ribosomal protein**
1.04E-24	**METABOLISM**	**C-compound and carbohydrate metabolism**	**Intermediary carbon metabolism**
8.50E-18	**TRANSCRIPTION**	**mRNA processing**	**small nuclear ribonucleoprotein (snRNP)**
4.23E-08	**METABOLISM**	**lipid, fatty-acid and isoprenoid metabolism**	**isoprenoid biosynthesis**
5.94E-08	**METABOLISM**	**amino acid metabolism**	**branched chain amino acids from aspartate**
9.82E-05	PROTEIN FATE (folding, modification, destination)	protein folding and stabilization	chaeronin
5.25E-04	PROTEIN FATE (folding, modification, destination)	Ubiquitin 26S Proteasome proteolytic pathway	20S core protease
7.50E-04	TRANSPORT FACILITATION	vesicular transport (Golgi network, etc.)	G-protein mediated signal transduction
2.32E-03	**METABOLISM**	**amino acid metabolism**	**—**
8.15E-03	**METABOLISM**	**amino acid metabolism**	**aromatic amino acids (Phe, Tyr, Trp) metabolism**
9.19E-03	**METABOLISM**	**nucleotide metabolism**	**pyrimidine biosynthesis**

* The terms with bold font were common through (a) to (d).

(b) *P. patens* (Overrepresentation)

FDR *p*-Value	1st-level category	2nd-level category	3rd-level category
3.64E-36	**METABOLISM**	**C-compound and carbohydrate metabolism**	**Intermediary carbon metabolism**
1.37E-26	**PROTEIN SYNTHESIS**	**translation**	**ribosomal protein**
7.21E-20	**TRANSCRIPTION**	**mRNA processing**	**small nuclear ribonucleoprotein (snRNP)**
1.09E-12	**METABOLISM**	**lipid, fatty-acid and isoprenoid metabolism**	**isoprenoid biosynthesis**
1.74E-08	TRANSCRIPTION	mRNA processing	splicing-related
2.34E-08	PROTEIN FATE (folding, modification, destination)	proteolytic degradation	Clp Protease complexes
4.29E-08	**METABOLISM**	**amino acid metabolism**	**branched chain amino acids from aspartate**
5.03E-08	TRANSPORT FACILITATION	ABC Superfamily	Soluble ABC protein
2.60E-07	**METABOLISM**	**amino acid metabolism**	**—**
2.26E-06	PROTEIN SYNTHESIS	translation	Eukaryotic initiation factor 1
2.30E-06	PROTEIN FATE (folding, modification, destination)	Ubiquitin 26S Proteasome proteolytic pathway	20S proteasome subunit
3.43E-06	**METABOLISM**	**amino acid metabolism**	**aromatic amino acids (Phe, Tyr, Trp) metabolism**
4.96E-06	METABOLISM	nucleotide metabolism	purine biosynthesis
1.84E-05	TRANSPORT FACILITATION	vesicular transport (Golgi network, etc.)	—
1.93E-05	**METABOLISM**	**nucleotide metabolism**	**pyrimidine biosynthesis**
2.38E-05	METABOLISM	secondary metabolism	Carotenoid biosynthesis
2.50E-05	PROTEIN FATE (folding, modification, destination)	protein modification	FK506-binding protein
3.10E-05	ENERGY	photosynthesis	Photosystem II
3.31E-05	TRANSPORT FACILITATION	NST-TPT family	—
1.80E-04	PROTEIN FATE (folding, modification, destination)	protein folding and stabilization	—
2.60E-04	TRANSPORT FACILITATION	MC family	—
2.76E-04	ENERGY	—	—
6.46E-04	METABOLISM	lipid, fatty-acid and isoprenoid metabolism	lipid and fatty-acid metabolism
6.46E-04	METABOLISM	biosynthesis of prenyl diphosphates	—
9.13E-04	ENERGY	photosynthesis	Photosystem I
9.54E-04	TRANSPORT FACILITATION	V-Type ATPase	transport mechanism
1.48E-03	TRANSPORT FACILITATION	vesicular transport (Golgi network, etc.)	clathrin adaptor complex
1.82E-03	ENERGY	Mitochondrial complex I	—
1.87E-03	PROTEIN FATE (folding, modification, destination)	Ubiquitin 26S Proteasome proteolytic pathway	20S core protease
2.53E-03	PROTEIN SYNTHESIS	translation	—
3.93E-03	METABOLISM	amino acid metabolism	aspartate/ornithine cycle/nitric oxide from glutamate
4.28E-03	METABOLISM	C-compound and carbohydrate metabolism	Ascorbic acid biosynthesis
4.53E-03	PROTEIN FATE (folding, modification, destination)	Ubiquitin 26S Proteasome proteolytic pathway	19 S regulatory particle
4.57E-03	METABOLISM	C-compound and carbohydrate metabolism	Starch
4.67E-03	TRANSCRIPTION	mRNA processing (splicing)	spliceosome
9.92E-03	CELL RESCUE, DEFENSE AND VIRULENCE	stress response	antioxidant

Table 2. Continued.

(c) *S. moellendorffii* (Overrepresentation)

FDR *p*-Value	1st-level category	2nd-level category	3rd-level category
3.31E-36	**METABOLISM**	**C-compound and carbohydrate metabolism**	**Intermediary carbon metabolism**
2.59E-16	**PROTEIN SYNTHESIS**	**translation**	**ribosomal protein**
1.75E-14	**METABOLISM**	**lipid, fatty-acid and isoprenoid metabolism**	**isoprenoid biosynthesis**
5.83E-12	**TRANSCRIPTION**	**mRNA processing**	**small nuclear ribonucleoprotein (snRNP)**
6.84E-11	ENERGY	photosynthesis	Photosystem II
4.69E-08	**METABOLISM**	**amino acid metabolism**	**branched chain amino acids from aspartate**
4.05E-07	**METABOLISM**	**nucleotide metabolism**	**pyrimidine biosynthesis**
4.07E-07	PROTEIN FATE (folding, modification, destination)	proteolytic degradation	Clp Protease complexes
1.07E-06	TRANSCRIPTION	mRNA processing	splicing-related
1.18E-06	**METABOLISM**	**amino acid metabolism**	—
3.15E-06	ENERGY	photosynthesis	Photosystem I
5.21E-06	METABOLISM	nucleotide metabolism	purine biosynthesis
8.25E-06	TRANSPORT FACILITATION	ABC Superfamily	Soluble ABC protein
1.50E-05	METABOLISM	C-compound and carbohydrate metabolism	Ascorbic acid biosynthesis
1.57E-05	**METABOLISM**	**amino acid metabolism**	**aromatic amino acids (Phe, Tyr, Trp) metabolism**
2.47E-05	PROTEIN FATE (folding, modification, destination)	protein modification	FK506-binding protein
3.60E-05	TRANSPORT FACILITATION	NST-TPT family	—
3.71E-05	PROTEIN SYNTHESIS	translation	Eukaryotic initiation factor 1
4.05E-05	TRANSPORT FACILITATION	MC family	—
1.12E-04	TRANSPORT FACILITATION	vesicular transport (Golgi network, etc.)	clathrin adaptor complex
1.76E-04	METABOLISM	secondary metabolism	Carotenoid biosynthesis
1.88E-04	PROTEIN FATE (folding, modification, destination)	Ubiquitin 26S Proteasome proteolytic pathway	20S proteasome subunit
5.76E-04	METABOLISM	biosynthesis of prenyl diphosphates	—
5.80E-04	METABOLISM	amino acid metabolism	aspartate/ornithine cycle/nitric oxide from glutamate
5.89E-04	TRANSPORT FACILITATION	vesicular transport (Golgi network, etc.)	—
1.43E-03	PROTEIN FATE (folding, modification, destination)	protein folding and stabilization	chaeronin
1.47E-03	METABOLISM	C-compound and carbohydrate metabolism	Starch
1.81E-03	PROTEIN FATE (folding, modification, destination)	proteolytic degradation	FtsH protease
1.86E-03	ENERGY	Mitochondrial complex I	—
5.54E-03	METABOLISM	lipid, fatty-acid and isoprenoid metabolism	lipid and fatty-acid metabolism
6.12E-03	TRANSPORT FACILITATION	V-Type ATPase	transport mechanism
8.05E-03	TRANSPORT FACILITATION	V-Type ATPase	—

(d) *O. sativa* (Overrepresentation)

FDR *p*-Value	1st-level category	2nd-level category	3rd-level category
2.21E-26	**METABOLISM**	**C-compound and carbohydrate metabolism**	**Intermediary carbon metabolism**
8.16E-18	**PROTEIN SYNTHESIS**	**translation**	**ribosomal protein**
7.74E-11	ENERGY	photosynthesis	Photosystem II
1.30E-10	METABOLISM	lipid, fatty-acid and isoprenoid metabolism	lipid and fatty-acid metabolism
1.26E-09	**TRANSCRIPTION**	**mRNA processing**	**small nuclear ribonucleoprotein (snRNP)**
1.35E-08	**METABOLISM**	**lipid, fatty-acid and isoprenoid metabolism**	**isoprenoid biosynthesis**
2.93E-07	TRANSPORT FACILITATION	MC family	—
3.59E-07	METABOLISM	secondary metabolism	Carotenoid biosynthesis
9.88E-07	TRANSCRIPTION	mRNA processing	splicing-related
1.22E-06	METABOLISM	C-compound and carbohydrate metabolism	Ascorbic acid biosynthesis
2.67E-06	TRANSPORT FACILITATION	vesicular transport (Golgi network, etc.)	—
6.16E-06	METABOLISM	C-compound and carbohydrate metabolism	Starch
6.70E-06	**METABOLISM**	**amino acid metabolism**	**branched chain amino acids from aspartate**
1.39E-05	METABOLISM	—	—
1.58E-05	METABOLISM	nucleotide metabolism	purine biosynthesis
5.13E-05	ENERGY	photosynthesis	Photosystem I
8.71E-05	ENERGY	Mitochondrial complex I	Carbonic anhydrase subunts
9.27E-05	METABOLISM	amino acid metabolism	aspartate/ornithine cycle/nitric oxide from glutamate

Table 2. Continued.

(d) *O. sativa* (Overrepresentation)

FDR *p*-Value	1st-level category	2nd-level category	3rd-level category
9.33E-05	**METABOLISM**	**amino acid metabolism**	**aromatic amino acids (Phe,Tyr,Trp) metabolism**
1.34E-04	PROTEIN FATE (folding, modification, destination)	protein modification	FK506-binding protein
2.34E-04	TRANSPORT FACILITATION	ABC Superfamily	Soluble ABC protein
4.32E-04	PROTEIN FATE (folding, modification, destination)	—	—
4.46E-04	PROTEIN SYNTHESIS	translation	Eukaryotic initiation factor 1
7.47E-04	PROTEIN FATE (folding, modification, destination)	proteolytic degradation	Clp Protease complexes
7.63E-04	**METABOLISM**	**amino acid metabolism**	**—**
9.76E-04	METABOLISM	nucleotide metabolism	—
1.00E-03	CELLULAR COMMUNICATION /SIGNAL TRANSDUCTION MECHANISM	intracellular signalling	Protein phosphatase
1.54E-03	PROTEIN FATE (folding, modification, destination)	protein folding and stabilization	—
1.72E-03	ENERGY	Mitochondrial complex I	—
3.51E-03	TRANSPORT FACILITATION	vesicular transport (Golgi network, etc.)	clathrin adaptor complex
7.95E-03	ENERGY	—	—
8.17E-03	**METABOLISM**	**nucleotide metabolism**	**pyrimidine biosynthesis**
8.33E-03	ENERGY	respiration	—
9.07E-03	METABOLISM	nucleotide metabolism	pyrimidine modification
9.28E-03	TRANSPORT FACILITATION	POT family(NRT1 family)	ion transporters
9.31E-03	CELLULAR COMMUNICATION /SIGNAL TRANSDUCTION MECHANISM	intracellular signalling	enzyme mediated signal transduction
9.97E-03	METABOLISM	Sulfur metabolism	sulfate assimilation

sequenced, the more evolutionary process of plants could be unraveled based on *A. thaliana* gene functions.

Acknowledgements

We would like to acknowledge *S. moellendorffii* genome sequence project. The sequence data of *S. moellendorffii* were produced by the US Department of Energy Joint Genome Institute http://www.jgi.doe.gov/ in collaboration with the user community. We also acknowledge The Japan Science and Technology Agency, CREST grant, for supporting this work.

References

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. *Nucl Acids Res* 25: 3389–3402

Aoki K, Ogata Y, Shibata D (2007) Approaches for extracting practical information from gene co-expression networks in plant biology. *Plant Cell Physiol* 48: 381–390

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25: 25–9

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol* 57: 289–300

Bowman JL, Floyd SK, Sakakibara K (2007) Green genes-comparative genomics of the green branch of life. *Cell* 129: 229–234

Chervitz SA, Hester ET, Ball CA, Dolinski K, Dwight SS, et al. (1999) Using the Saccharomyces Genome Database (SGD) for analysis of protein similarities and structure. *Nucl Acids Res* 27: 74–8

Iida K, Seki M, Sakurai T, Satou M, Akiyama K, Toyoda T, Konagaya A, Shinozaki K (2005) RARTF: database and tools for complete sets of Arabidopsis transcription factors. *DNA Res* 12: 247–256

IRGSP (2005) The map-based sequence of the rice genome. *Nature* 436: 793–800

Joyce AR, Palsson BO (2006) The model organism as a system: integrating 'omics' data sets. *Nat Rev Mol Cell Biol* 7: 198–210

Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, et al. (2007) The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science* 318: 245–250

Obayashi T, Kinoshita K, Nakai K, Shibaoka M, Hayashi S, Saeki M, Shibata D, Saito K, Ohta H (2007) ATTED II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in Arabidopsis. *Nucl Acids Res* 35: D863–D869

Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, et al. (2008) The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science* 319: 64–69

Riano-Pachon DM, Ruzicic S, Dreyer I, Mueller-Roeber B (2007) PlnTFDB: an integrative plant transcription factor database. *BMC Bioinformatics* 8: 42

Riechmann JL, Ratcliffe OJ (2000) A genomic perspective on plant transcription factors. *Curr Opin Plant Biol* 3: 423–434

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

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