Received Date : 11-May-2012

Revised Date : 16-Oct-2012

Accepted Date : 07-Nov-2012

Article type : Original Article

Evolutionary analyses of non-family genes in plants

Chu-Yu Ye[1,2,3,4], Ting Li[1,3,4], Hengfu Yin[1], David J. Weston[1], Gerald A. Tuskan[1,2], Timothy J. Tschaplinski[1,2], Xiaohan Yang[1,2,*]

[1]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

[2]BioEnergy Science Center, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

[3]Present addresses: Beijing Forestry University, Beijing 100083, China (C.-Y.Y.); Pioneer Hi-Bred International, Johnston, IA 50131, USA (T.L.)

[4]These authors contributed equally to this work.

[*]Corresponding author:

Xiaohan Yang
Biosciences Division
Oak Ridge National Laboratory
P.O. Box 2008 MS-6422
Oak Ridge TN 37831-6422
Phone: (865)-241-6895; Fax: (865) 576-9939
E-mail: yangx@ornl.gov

Running title: Comparative analysis of non-family genes

Keywords: Non-family genes, plant genome, comparative genomics, evolution, functional genomics

SUMMARY

There are a large number of “non-family” (NF) genes that do not cluster into families with three or more members per genome. While gene families have been extensively studied, a systematic analysis of NF genes has not been reported. We performed comparative studies on NF genes in 14 plant species. Based on the clustering of protein sequences, we identified ~94,000 NF genes across these species that were divided into five evolutionary groups: Viridiplantae-wide, angiosperm-specific, monocot-specific, dicot-specific, and those that were species-specific. Our analysis revealed that the NF genes resulted largely from less frequent gene duplications and/or a higher rate of gene loss after segmental duplication relative to genes in both low-copy-number families (LF; 3 – 10 copies per genome) and high-copy-number families (HF; >10 copies). Furthermore, we identified functions enriched in the NF gene set as compared with the HF genes. We found that NF genes were involved in essential biological processes shared by all plant lineages (e.g., photosynthesis and translation), as well as gene regulation and stress responses associated with phylogenetic diversification. In particular, our analysis of an *Arabidopsis* protein-protein interaction network revealed that hub proteins with the top 10% most connections were over-represented in the NF set relative to the HF set. This research highlights the roles that NF genes may play in evolutionary and functional genomics research.

INTRODUCTION

A gene family can be defined as a group of genes with similar sequences, resulting from various gene duplication events and often sharing similar or partially redundant functions (De Grassi *et al.*, 2008; Demuth and Hahn, 2009). In the most inclusive manner, the majority of genes in plant genomes belong to gene families that contain three or more gene members per genome. Many gene families in plants have been extensively studied, e.g., MADS-box (Nam *et al.*, 2004; Parenicova *et al.*, 2003), ABC protein superfamily (Verrier *et al.*, 2008), NAC transcription factor (Ooka *et al.*, 2003), Glycosyltransferase superfamily (Ye *et al.*, 2011; Yin *et al.*, 2010; Yonekura-Sakakibara and Hanada, 2011), F-box (Yang *et al.*, 2008), AP2/ERF superfamily (Nakano *et al.*, 2006), and many protein kinase families (Hrabak *et al.*, 2003). Still, a large number of genes, with 1 – 2 gene copies per genome, do not belong to such gene families; we term these genes as non-family (NF) genes. Only a few studies have addressed the role of NF genes in sequenced plant species (Duarte *et al.*, 2010; Guo *et al.*, 2007), and as such, the function and evolution of NF genes remains unclear.

There is considerable variation in the number of genes within gene families. The variation in numbers of gene family members is partly due to the extent and retention of gene duplication (Chauve *et al.*, 2008; Ohno, 1970). Recent research indicates that the ancestral angiosperm genome contained 12,000 - 14,000 genes (Proost *et al.*, 2011), far less than the number of genes in extant angiosperm genomes (e.g., ~27,000 in *Arabidopsis*, ~26,000 in *Vitis*, ~40,000 in *Populus*, and ~40,000 in *Oryza*) (Goodstein *et al.*, 2012). The gene number expansion in extant angiosperm genomes appear to be created through various gene duplication events (Van de Peer *et al.*, 2009). For example, the F-box gene family is expanded in herbaceous plants relative to woody plants specifically due to a high rate of tandem duplications in the former (Yang *et al.*, 2008). Several models (e.g., neofunctionalization, subfunctionalization, balanced gene drive) have been proposed to explain the evolutionary fate of duplicated genes (Freeling, 2009; Hughes, 1994; Ohno, 1970; Yang *et al.*, 2006). Gene loss after duplication is one of the evolutionary modes that contribute to the contraction of gene families in *Arabidopsis thaliana* that experienced a loss of around 5,700 genes after divergence between *A. thaliana* and *A. lyrata* (Proost *et al.*, 2011).

To understand the evolution and function of NF genes in plants, we performed a large-scale comparative analysis of NF genes in diverse plant species ranging from algae to moss to angiosperm. In a comparative framework, we analyzed multiple aspects of the NF genes (1 – 2 copies per genome) in comparison with low-copy-number gene families (LF; 3 – 10 copies) and high-copy-number gene families (HF; >10 copies). Our analysis revealed differences in the frequency of gene duplication, evolutionary rate, gene ontology, subcellular localization and gene expression between NF genes and genes in LF and HF categories, and among NF genes in alternate evolutionary groups (i.e., Viridiplantae-wide, angiosperm-specific, monocot-specific, dicot-specific, and species-specific). We found that some plant functions involved more NF genes, relative to HF genes. We identified NF genes that were involved in stress responses as well as cell wall biosynthesis. Furthermore, our analysis of an *Arabidopsis* protein-protein interaction network revealed that a higher proportion of the NF gene set were classified as hubs (i.e., highly connected nodes) as compared with the HF gene set. We present this study to highlight the roles of NF genes in evolutionary and functional genomics research.

RESULTS

Non-family (NF) genes in plant genomes

Based on clustering analysis of protein sequences, we divided the protein-encoding genes from 14 diverse plant species, including algae (*Chlamydomonas reinhardtii* and *Volvox carteri*), moss (*Physcomitrella patens*), lycophyte (*Selaginella moellendorffii*), monocots (*Brachypodium distachyon*, *Oryza sativa*, *Sorghum bicolor*, and *Zea mays*), and dicots (*Arabidopsis thaliana*, *Carica papaya, Glycine max*, *Populus trichocarpa, Solanum tuberosum*, and *Vitis vinifera*), into three categories: 1) genes with only 1 – 2 copies per genome, i.e., "non-family" (NF) genes, 2) genes in low-copy-number families (LF), containing 3 - 10 copies per genome, and 3) genes in high-copy-number families (HF), containing more than 10 copies per genome (Figure 1). Non-vascular plants contained a significantly (P<0.01, Chi-square) higher proportion of NF genes and significantly (P<0.01, Chi-square) fewer HF genes relative to vascular plants (Figure 1). Furthermore, NF contained significantly (P<0.01, paired t-test) more single-copy genes than two-copy genes (Figure 2).

Among all NF genes, we identified five evolutionary groups: 1) NF1 -- genes having homologs in all the 14 species studied (i.e., found Viridiplantae-wide), 2) NF2 -- genes having homologs in all and only the 10 angiosperm species studied (i.e., angiosperm-specific), 3) NF3 -- genes having homologs in all and only the four monocot species studied (i.e., monocot-specific), 4) NF4 -- genes having homologs in all and only the six dicot species studied (i.e., dicot-specific), and 5) NF5 -- genes not conserved among the species studied (i.e., species-specific) (Figure 3). The number of genes in NF5 was significantly (P<0.01, ANOVA+LSD test) higher than that in the other groups (i.e., NF1, NF2, NF3, NF4). Interestingly, compared to the five NF groups (i.e., NF1-NF5), only four corresponding groups could be identified among the LF or HF genes and the dicot-specific group (LF4 or HF4) that contains LF or HF genes having homologs in all and only the six dicot species is absent (Figures S1 and S2), suggesting that the shared evolutionary history of the LF and HF gene families in dicots can be dated before the divergence between monocots and dicots.

Duplication and loss of NF genes

Using syntenic gene order, our analysis of duplication events in *A. thaliana*, *B. distachyon*, *G. max*, *O. sativa*, *P. trichocarpa*, *S. bicolor*, *V. vinifera* and *Z. mays* revealed that the frequency of tandem duplication in the NF genes is significantly (*P*<0.001, paired t-test) lower than that in both LF and HF categories, and the frequency of tandem duplication in the LF genes is significantly (*P*<0.001, paired t-test) lower than that in HF genes (Figure 4A). The frequency of segmental duplication in NF genes was (*P*<0.01, paired t-test) significantly lower than that in both LF and HF categories and there was no significant difference in frequency of segmental duplication between LF and HF categories (Figure 4B). Furthermore, NF genes had a significantly (*P*<0.01, paired t-test) higher frequency of gene loss after syntenic segmental duplication than that in both LF and HF categories, and genes in the LF category had a higher frequency (*P*<0.001, paired t-test) of gene loss than those in HF category (Figure 4C). These data suggest that the NF genes resulted mainly from a higher rate of gene loss after segmental duplication as well as a lower rate of tandem gene duplication.

Evolutionary fate of NF genes

To investigate the evolutionary fate of NF genes after duplication, we examined the nonsynonymous/synonymous (*Ka/Ks*) ratio for gene pairs resulting from segmental duplication. Our analysis showed that NF genes had a significantly (*P*<0.01, ANOVA+LSD test) higher *Ka/Ks* ratio than genes in the LF and HF categories (Figure S3A), indicating that NF genes may have been under more relaxed selection than LF or HF genes. Furthermore, the species-specific NF genes (i.e., category NF5) had a significantly (*P*<0.01, ANOVA+LSD test) higher *Ka/Ks* ratio than the NF genes conserved among all plant species investigated (i.e., NF1) (Figure S3B), suggesting that the species-specific NF genes may have been under more relaxed selection than the conserved NF genes.

Functions of NF genes

Gene ontology (GO) analysis revealed that NF genes were disproportionately involved in various biological processes, with the most over-represented processes including nitrogen metabolism, biosynthetic process, biological regulation and response to stimulus (Figure S4). GO enrichment analysis revealed that many biological processes (e.g., translation, nucleoside metabolism, cofactor metabolism and photosynthesis) were significantly (*P*<0.05, Fisher) over-represented in conserved NF genes (i.e., shared by at least two species) relative to conserved HF genes (Figure 5, Table S1). Several biological processes (e.g., gene expression) were significantly (*P*<0.05, Fisher) enriched in species-specific NF genes relative to species-specific HF genes (NF5 vs. HF5; Table S2). Some biological processes (e.g., heterocycle biosynthetic process) were enriched (*P*<0.05, Fisher) in conserved NF genes in comparison with conserved LF genes (Tables S3). We also found some enriched (*P*<0.05, Fisher) biological processes in conserved and species-specific LF genes in comparison with conserved and species-specific HF genes, respectively (Tables S4 and S5). In addition, some biological processes were enriched (*P*<0.05, Fisher) in species-specific NF genes relative to conserved NF genes (Table S6).

To determine if there were differences in gene functions at the subcellular level among different gene-copy categories, we predicted the protein subcellular localization of genes in the NF, LF and HF categories. The results showed that genes found in 1) plasma membranes were over-represented in conserved (i.e., shared by at least two species) NF, LF and HF categories, 2) chloroplasts in conserved NF and LF categories, and 3) extracellular space and nucleus in species-specific NF and LF categories (*P*<0.01, the cumulative Poisson distribution; Table 1). These over-represented subcellular localizations were in concordance with the over-represented biological processes, such that chloroplast localization and photosynthesis genes were in conserved NF and LF genes and nucleus localization and regulation processes were in species-specific NF and LF genes (Table 1; Tables S1, S2, S4, S5). Furthermore, we found that chloroplast and mitochondrion localization was over-represented in NF1 (Viridiplantae-wide NF genes), nucleus localization over-represented in NF2 (angiosperm-specific NF genes) and NF5 (species-specific NF genes), and extracellular localization over-represented in NF3 and NF4 (*P*<0.01, the cumulative Poisson distribution; Table 2).

To obtain experimental support for our computational prediction of NF gene functions, we interrogated the AtGenExpress array data (Kilian *et al.*, 2007) for *Arabidopsis* genes in the NF1 category (Viridiplantae-wide NF genes) versus NF5 (species-specific NF genes). Our analysis revealed eight and 15 expression clusters for NF1 and NF5, respectively (Figures S5 and S6). This suggests that species-specific genes are involved in more diverse stress responses than the genes conserved among large evolutionary space.

Gene co-expression

Genes that are highly co-expressed may be involved in similar biological processes (Usadel *et al.*, 2009). Our previous co-expression analysis identified 692 genes (plus two obsolete gene models) associated with cell wall biosynthesis in *Arabidopsis* (Yang *et al.*, 2011a). Among these 692 genes, there were 70 NF genes, 177 LF genes and 445 HF genes (Tables S7 and S8). For example, two NF genes, AT2G31930 and AT2G41610, were co-expressed with known cell wall biosynthesis genes and preferentially expressed in stem tissue enriched with secondary cell wall materials (Figure 6). GO enrichment analysis showed that the secondary cell wall biosynthesis was significantly ($P<0.001$, Fisher) over-represented in this co-expressed gene subnetwork as compared with the whole *Arabidopsis* genome (Tables S9).

To explore the roles of NF genes in stress responses, we investigated the association of NF genes with known genes responsive to abiotic (i.e., cold, heat, salt, water deprivation) and biotic (i.e., bacteria and fungi) stresses in *Arabidopsis*. We identified 713 NF genes, 1,268 LF genes and 2,538 HF genes associated with stress responses (Tables S7 and S8). For example, the NF gene AT5G35320 was co-expressed with several heat shock protein genes (Figure S7A), and highly expressed under heat stress (Figure S7B). Genes related to heat response were significantly ($P<0.001$, Fisher) enriched in this co-expression subnetwork, as compared with the whole *Arabidopsis* genome (Tables S10). NF gene AT3G01420 was co-expressed with several biotic stress responsive genes (Figure 7A) and highly expressed in *Arabidopsis* leaves infected with *Pseudomonas* (Figure 7B). Genes related to defense response were significantly ($P<0.05$, Fisher) enriched in this co-expression gene subnetwork, as compared with the whole *Arabidopsis* genome (Tables S11).

Previous analysis revealed that 77% of one-to-one orthologous gene pairs between *A. thaliana* and *O. sativa* showed conserved co-expression (Movahedi *et al.*, 2011). We examined the co-expression conservation of one-to-one orthologous gene pairs in NF, LF and HF between *A. thaliana* and *O. sativa*. Our analysis revealed that NF genes had a significantly ($P<0.01$, the cumulative Poisson distribution) higher rate of co-expression conservation than HF genes (Figure S8).

Protein-protein interaction

Protein-protein interactions (PPIs) are crucial for a large number of cellular processes in living organisms (He and Zhang, 2006; Vallabhajosyula *et al.*, 2009). To determine the contribution of NF proteins in protein-protein interactions, we investigated an *Arabidopsis* PPI network, AtPIN (Brandao *et al.*, 2009). The results showed that a significantly (P<0.01, the cumulative Poisson distribution) higher proportion of the NF set could be classified as hubs, the top 10% most connected nodes (with a minimum degree of 36) in the PPI network, relative to the LF and HF sets (Figure 8; Table S12). In addition, all of the hubs in the PPI network were found in the conserved (i.e., shared by at least two species) NF, LF and HF categories, consistent with the view that hub proteins are generally evolutionary conserved (Vallabhajosyula *et al.*, 2009).

DISCUSSION

With the rapid accumulation of genome sequencing data in public databases, our knowledge about genes found in expanded families (i.e., three or more gene copies per genome) has greatly increased. However, questions related to the function and evolution of non-family (NF) genes (i.e., one or two copies per genome) remain unanswered. If the NF genes are involved in principal biological processes, why have they not expanded as much as the family genes have? To begin to address these questions, we performed a large-scale systematic analysis of NF genes in 14 plant genomes across a large evolutionary space. Our study provides initial evidence that NF genes play an important role in plant functions and their evolutionary dynamics are different from that of family genes. Key functions (e.g., cellular nitrogen compound metabolic process, translation, cofactor metabolic process) were enriched in NF genes (Figure 5; Table S1). In particular, some essential biological processes, such as photosynthesis in chloroplasts and respiration in mitochondria that are shared by all plant lineages, were significantly (P<1E^{-77}, Fisher) over-represented in the NF genes (Figure 5; Table S1). Our comparative analysis indicates that NF genes resulted largely from a higher rate of gene loss after segmental duplications and/or lower frequency of tandem duplications (Figure 4). This is consistent with previous reports showing that tandem duplication is one of the main factors contributing to the expansion of gene families (Cannon *et al.*, 2004; Yang *et al.*, 2008), and gene loss is a common aspect of evolutionary dynamics after gene duplication (Freeling, 2009; Yang *et al.*, 2006; Yang *et al.*, 2011b).

There are still a large number of NF genes annotated as unknown, putative or hypothetical, especially for the NF genes in the species-specific group (i.e. NF5). These NF genes should be considered as candidates for future functional studies. It was reported that the proportion of essential singleton genes was higher than that of essential duplicated genes in yeast and nematode, whereas the singletons and duplicates were determined to be equally essential in mouse (Liao and Zhang, 2007). In plants, there is currently very limited information on NF mutants relative to yeast, nematode and mouse. Future efforts are needed to create knockout mutants for majority if not all of the genes in several model plant species for estimating the proportion of essential genes among non-family and family genes. Hubs tend to be more essential than non-hub proteins (He and Zhang, 2006; Vallabhajosyula *et al.*, 2009). We found that hub proteins in the *Arabidopsis* protein-protein interaction network were over-represented in the NF set relative to the HF set (Figure 8), suggesting that NF genes may play more important roles than expected in plants.

MATERIALS AND METHODS

Clustering of protein sequences

The annotated non-TE (transposable element) protein sequences of 14 plant species, including *Chlamydomonas reinhardtii* (www.Phytozome.net; Phytozome), *Volvox carteri* (Phytozome), *Physcomitrella patens* (Phytozome), *Selaginella moellendorffii* (Phytozome), *Brachypodium distachyon* (Phytozome), *Sorghum bicolor* (Phytozome)*, Vitis vinifera* (Phytozome), *Carica papaya* (Phytozome), *Glycine max* (Phytozome), *Populus trichocarpa* (Phytozome), *Oryza sativa* (rice.plantbiology.msu.edu), *Zea mays* (www.maizesequence.org), *Solanum tuberosum* (potatogenomics.plantbiology.msu.edu), and *Arabidopsis thaliana* (www.*Arabidopsis*.org), were used to cluster protein sequences. The longest protein sequence was selected in case of multiple transcripts annotated for one gene locus. All-against-all BLASTP search of these protein sequences were performed using Blast+ (Camacho *et al.*, 2009) with an E-value cutoff of 1E^{-3} followed by clustering analysis using TRIBE-MCL with an inflation value of 1.2 (Enright *et al.*, 2002). Based on the protein clustering analysis, the genes were classified into three categories: Non-family (NF) that was defined as the gene clusters containing 1 - 2 genes in one genome, low-copy-number family (LF) defined as the gene clusters containing 3 - 10 genes in one genome, and high-copy-number family (HF) defined as the gene clusters containing more than ten genes in one genome.

Identification of gene duplication and gene loss

The information for segmental duplication was obtained from SynMap (synteny.cnr.berkeley.edu). The tandem duplicated genes were identified and defined as an array of two or more genes that were in the same protein cluster and were found within a 100-kb genomic window (Yang *et al.*, 2008). Gene loss after segmental duplications was defined as the absence of homologous genes within one copy of the duplicated blocks, as compared with the other copy of the duplicated segments in the syntenic regions. For example, if a syntenic block contains duplicated fragments of "gene A --- gene B --- gene C --- gene D --- gene E --- gene F --- gene G --- gene H (fragment 1; gene F is not generated by tandem duplication)" and "gene A* --- gene B* --- gene C* --- gene D* --- gene E* --- gene G* --- gene H* (fragment 2; genes A*, B*, C*, D*, E*, G*, and H* are duplicated copies of genes A, B, C, D, E, G, and H, respectively)", we considered that there was a gene loss (i.e., gene F*) on fragment 2.

Evolutionary analysis of non-family genes

Nonsynonymous (*Ka*) and synonymous (*Ks*) substitution rates of full-length coding sequences were calculated using the KaKs_Calculator with the GMYN method (Wang *et al.*, 2010). The same numbers (3679) of gene pairs generated from segmental duplication were randomly selected from NF, LF, and HF categories for *Ka/Ks* analysis. More than 70% of coding regions in the shorter sequences of duplicated gene pairs were aligned for *Ka*/*Ks* analysis, as determined by analysis of 200 randomly selected gene pairs (Figure S9).

Gene ontology (GO) analysis

We obtained whole genome GO term annotation for the 14 species investigated in this study using Blast2GO with a BlastP e-value hit filter of 1E^{-6}, an annotation cutoff value of 55, and GO weight of 5 (Conesa *et al.*, 2005). GO enrichment analysis was performed using Blast2GO or agriGO (bioinfo.cau.edu.cn/agriGO/) with a Fisher exact test (Du *et al.*, 2010), and GO terms were summarized using REVIGO (Supek *et al.*, 2011). For pair-wise enrichment comparison between the NF, LF and HF sets, all the genes in each set were used. For example, for calculating enrichment of GO terms in the conserved NF set relative to the conserved HF set, all of the conserved NF genes were studied with all of the conserved HF genes used as a reference.

Prediction of protein subcellular localization

The prediction of protein subcellular localization was performed using the Yloc (Briesemeister *et al.*, 2010) with the model set as "YLoc+" and version set as "Plants", WoLF PSORT (Horton *et al.*, 2007) with the organism set as "Plant", and CELLO (Yu *et al.*, 2006) with the organism set as "Eukaryotes". The consensus results predicted by these three methods (i.e. the same results obtained by the three different methods) were adopted. 6,000 randomly selected, conserved (i.e., shared by at least by two species) NF/LF/HF genes, 6,000 randomly selected, species-specific NF genes, all of the species-specific LF genes (6,225 genes), and all of the species-specific HF genes (7,164 genes) were used for subcellular localization analysis. The random selection of gene sets was repeated three times and results from analysis of protein subcellular localization were consistent among the repeated random gene sets.

Analysis of gene expression

Expression data for *Arabidopsis* genes were obtained from AtGenExpress (Kilian *et al.*, 2007; Schmid *et al.*, 2005). K-means clustering of the gene expression pattern was performed using SC(2)ATmd (Olex and Fetrow, 2011) with optimal number of clusters determined by Figure of Merit. 263 species-specific NF genes (NF5), the same number as Viridiplantae-wide NF genes (NF1), were randomly selected for clustering analysis. The random selection of gene sets was repeated three times and results from clustering analysis were consistent among the repeated random gene sets. *Arabidopsis* co-expression data were obtained from ATTED-II (Obayashi *et al.*, 2007). The co-expression conservation data for one-to-one orthologs in *A. thaliana* and *O. sativa* were obtained from Movahedi *et al*. (2011).

Analysis of protein-protein interaction

The *Arabidopsis* protein-protein interaction (PPI) network data was obtained from AtPIN (Brandao *et al.*, 2009). The hubs were defined as the top 10% most connected nodes (with a minimum degree of 36).

Statistical analysis

Statistical analyses, including Chi-square tests, paired t-tests and ANOVA with LSD tests, were performed using R (www.r-project.org/).

ACKNOWLEDGEMENTS

We would like to thank S. D. Wullschleger for thoughtful and insightful comments on the manuscript. This work was supported by the U.S. DOE Office of Biological and Environmental Research, Genomic Science Program, the U.S. DOE BioEnergy Science Center and the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory. The BioEnergy Science Center is a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science. Oak Ridge National Laboratory is managed by UT-Battelle, LLC for the U.S. Department of Energy under Contract Number DE–AC05–00OR22725.

CONFLICT OF INTEREST

The authors have no conflict of interest to declare.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Table S1. Biological processes over-represented in conserved non-family genes as compared with conserved high-copy-number family genes.

Table S2. Biological processes over-represented in species-specific non-family genes as compared with species-specific high-copy-number family genes.

Table S3. Biological processes over-represented in conserved non-family genes as compared with conserved low-copy-number family genes.

Table S4. Biological processes over-represented in conserved low-copy-number family genes as compared with conserved high-copy-number family genes.

Table S5. Biological processes over-represented in species-specific low-copy-number family genes as compared with species-specific high-copy-number family genes.

Table S6. Biological processes over-represented in species-specific non-family genes as compared with conserved non-family genes.

Table S7. Number of non-family, low-copy-number family and high-copy-number family genes related to stress responses and cell wall biosynthesis in *Arabidopsis*.

Table S8. List of *Arabidopsis* non-family genes related to stress responses and cell wall biosynthesis based on GO annotation and NF genes that were co-expressed with genes involved in stress responses and cell wall biosynthesis.

Table S9. Biological processes over-represented in the gene cluster in Figure 6 as compared with the *Arabidopsis* genome.

Table S10. Biological processes over-represented in the gene cluster in Supplemental Figure S7A as compared with the *Arabidopsis* genome.

Table S11. Biological processes over-represented in the gene cluster in Figure 7A as compared with the *Arabidopsis* genome.

Table S12. Non-family hub genes in an *Arabidopsis* protein-protein interaction network.

Figure S1. Numbers of genes in the low-copy-number families in different phylogenic groups.

Figure S2. Numbers of genes in the high-copy-number families among different phylogenetic groups.

Figure S3. The non-synonymous to synonymous substitution (*Ka/Ks*) ratio for the coding region of paralogous genes generated from segmental duplications.

Figure S4. The top 15 biological processes of genes classified as non-family genes in *Arabidopsis*.

Figure S5. K-means clustering of expression pattern of *Arabidopsis* for genes classified as non-family genes in NF1 group (i.e., Viridiplantae-wide; defined in Fig. 3).

Figure S6. K-means clustering of expression pattern of *Arabidopsis* for genes classified as non-family genes in NF5 group (i.e., species-specific; defined in Fig. 3).

Figure S7. A network analysis of an *Arabidopsis* non-family gene involved in heat response.

Figure S8. Co-expression conservation of NF, LF and HF genes between *A. thaliana* and *O. sativa*.

Figure S9. Proportion of coding regions in the shorter sequences of duplicated gene pairs were aligned for *Ka*/*Ks* analysis in 200 randomly selected gene pairs.

REFERENCES

Brandao, M.M., Dantas, L.L. and Silva-Filho, M.C. (2009) AtPIN: *Arabidopsis thaliana* protein interaction network. *Bmc Bioinformatics*, **10**, 454.

Briesemeister, S., Rahnenfuhrer, J. and Kohlbacher, O. (2010) YLoc-an interpretable web server for predicting subcellular localization. *Nucleic Acids Research*, **38**, W497-W502.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. (2009) BLAST+: architecture and applications. *Bmc Bioinformatics*, **10**, 421.

Cannon, S.B., Mitra, A., Baumgarten, A., Young, N.D. and May, G. (2004) The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC Plant Biol*, **4**, 1-21.

Chauve, C., Doyon, J.P. and El-Mabrouk, N. (2008) Gene family evolution by duplication, speciation, and loss. *J Comput Biol*, **15**, 1043-1062.

Conesa, A., Gotz, S., Garcia-Gomez, J.M., Terol, J., Talon, M. and Robles, M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674-3676.

De Grassi, A., Lanave, C. and Saccone, C. (2008) Genome duplication and gene-family evolution: the case of three OXPHOS gene families. *Gene*, **421**, 1-6.

Demuth, J.P. and Hahn, M.W. (2009) The life and death of gene families. *Bioessays*, **31**, 29-39.

Du, Z., Zhou, X., Ling, Y., Zhang, Z. and Su, Z. (2010) agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Res*, **38**, W64-70.

Duarte, J.M., Wall, P.K., Edger, P.P., Landherr, L.L., Ma, H., Pires, J.C., Leebens-Mack, J. and dePamphilis, C.W. (2010) Identification of shared single copy nuclear genes in *Arabidopsis*, *Populus*, *Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. *Bmc Evol Biol*, **10**, 61.

Enright, A.J., Van Dongen, S. and Ouzounis, C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*, **30**, 1575-1584.

Freeling, M. (2009) Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annu Rev Plant Biol*, **60**, 433-453.

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N. and Rokhsar, D.S. (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res*, **40**, D1178-1186.

Guo, W.J., Li, P., Ling, J. and Ye, S.P. (2007) Significant comparative characteristics between orphan and nonorphan genes in the rice (*Oryza sativa* L.) genome. *Comp Funct Genom*, 21676.

He, X. and Zhang, J. (2006) Why do hubs tend to be essential in protein networks? *PLoS Genet*, **2**, e88.

Horton, P., Park, K.J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C.J. and Nakai, K. (2007) WoLF PSORT: protein localization predictor. *Nucleic Acids Res*, **35**, W585-587.

Hrabak, E.M., Chan, C.W., Gribskov, M., Harper, J.F., Choi, J.H., Halford, N., Kudla, J., Luan, S., Nimmo, H.G., Sussman, M.R., Thomas, M., Walker-Simmons, K., Zhu, J.K. and Harmon, A.C. (2003) The *Arabidopsis* CDPK-SnRK superfamily of protein kinases. *Plant Physiology*, **132**, 666-680.

Hughes, A.L. (1994) The evolution of functionally novel proteins after gene duplication. *Proc Biol Sci*, **256**, 119-124.

Kilian, J., Whitehead, D., Horak, J., Wanke, D., Weinl, S., Batistic, O., D'Angelo, C., Bornberg-Bauer, E., Kudla, J. and 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.

Liao, B.Y. and Zhang, J.Z. (2007) Mouse duplicate genes are as essential as singletons. *Trends Genet*, **23**, 378-381.

Movahedi, S., Van de Peer, Y. and Vandepoele, K. (2011) Comparative network analysis reveals that tissue specificity and gene function are important factors influencing the mode of expression evolution in *Arabidopsis* and rice. *Plant Physiology*, **156**, 1316-1330.

Nakano, T., Suzuki, K., Fujimura, T. and Shinshi, H. (2006) Genome-wide analysis of the ERF gene family in *Arabidopsis* and rice. *Plant Physiology*, **140**, 411-432.

Nam, J., Kim, J., Lee, S., An, G.H., Ma, H. and Nei, M.S. (2004) Type I MADS-box genes have experienced faster birth-and-death evolution than type II MADS-box genes in angiosperms. *P Natl Acad Sci USA*, **101**, 1910-1915.

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

Ohno, S. (1970) *Evolution by gene duplication*: Springer-Verlag.

Olex, A.L. and Fetrow, J.S. (2011) SC(2)ATmd: a tool for integration of the figure of merit with cluster analysis for gene expression data. *Bioinformatics*, **27**, 1330-1331.

Ooka, H., Satoh, K., Doi, K., Nagata, T., Otomo, Y., Murakami, K., Matsubara, K., Osato, N., Kawai, J., Carninci, P., Hayashizaki, Y., Suzuki, K., Kojima, K., Takahara, Y., Yamamoto, K. and Kikuchi, S. (2003) Comprehensive analysis of NAC family genes in *Oryza sativa* and *Arabidopsis thaliana*. *DNA Res*, **10**, 239-247.

Parenicova, L., de Folter, S., Kieffer, M., Horner, D.S., Favalli, C., Busscher, J., Cook, H.E., Ingram, R.M., Kater, M.M., Davies, B., Angenent, G.C. and Colombo, L. (2003) Molecular and phylogenetic analyses of the complete MADS-box transcription factor family in *Arabidopsis*: new openings to the MADS world. *Plant Cell*, **15**, 1538-1551.

Proost, S., Pattyn, P., Gerats, T. and Van de Peer, Y. (2011) Journey through the past: 150 million years of plant genome evolution. *Plant J*, **66**, 58-65.

Schmid, M., Davison, T.S., Henz, S.R., Pape, U.J., Demar, M., Vingron, M., Scholkopf, B., Weigel, D. and Lohmann, J.U. (2005) A gene expression map of *Arabidopsis thaliana* development. *Nat Genet*, **37**, 501-506.

Supek, F., Bosnjak, M., Skunca, N. and Smuc, T. (2011) REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One*, **6**, e21800.

Usadel, B., Obayashi, T., Mutwil, M., Giorgi, F.M., Bassel, G.W., Tanimoto, M., Chow, A., Steinhauser, D., Persson, S. and Provart, N.J. (2009) Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. *Plant Cell Environ*, **32**, 1633-1651.

Vallabhajosyula, R.R., Chakravarti, D., Lutfeali, S., Ray, A. and Raval, A. (2009) Identifying hubs in protein interaction networks. *PLoS One*, **4**, e5344.

Van de Peer, Y., Fawcett, J.A., Proost, S., Sterck, L. and Vandepoele, K. (2009) The flowering world: a tale of duplications. *Trends in Plant Science*, **14**, 680-688.

Verrier, P.J., Bird, D., Burla, B., Dassa, E., Forestier, C., Geisler, M., Klein, M., Kolukisaoglu, U., Lee, Y., Martinoia, E., Murphy, A., Rea, P.A., Samuels, L., Schulz, B., Spalding, E.J., Yazaki, K. and Theodoulou, F.L. (2008) Plant ABC proteins--a unified nomenclature and updated inventory. *Trends in Plant Science*, **13**, 151-159.

Wang, D., Zhang, Y., Zhang, Z., Zhu, J. and Yu, J. (2010) KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinformatics*, **8**, 77-80.

Yang, X., Kalluri, U.C., Jawdy, S., Gunter, L.E., Yin, T., Tschaplinski, T.J., Weston, D.J., Ranjan, P. and Tuskan, G.A. (2008) The F-box gene family is expanded in herbaceous annual plants relative to woody perennial plants. *Plant Physiology*, **148**, 1189-1200.

Yang, X., Tuskan, G.A. and Cheng, M.Z. (2006) Divergence of the Dof gene families in poplar, Arabidopsis, and rice suggests multiple modes of gene evolution after duplication. *Plant Physiology*, **142**, 820-830.

Yang, X., Ye, C.Y., Bisaria, A., Tuskan, G.A. and Kalluri, U.C. (2011a) Identification of candidate genes in *Arabidopsis* and *Populus* cell wall biosynthesis using text-mining, co-expression network analysis and comparative genomics. *Plant Sci*, **181**, 675-687.

Yang, X., Ye, C.Y., Cheng, Z.M., Tschaplinski, T.J., Wullschleger, S.D., Yin, W.L., Xia, X.L. and Tuskan, G.A. (2011b) Genomic aspects of research involving polyploid plants. *Plant Cell Tiss Org*, **104**, 387-397.

Ye, C.Y., Li, T., Tuskan, G.A., Tschaplinski, T.J. and Yang, X. (2011) Comparative analysis of GT14/GT14-like gene family in *Arabidopsis*, *Oryza*, *Populus*, *Sorghum* and *Vitis*. *Plant Sci*, **181**, 688-695.

Yin, Y., Chen, H., Hahn, M.G., Mohnen, D. and Xu, Y. (2010) Evolution and function of the plant cell wall synthesis-related glycosyltransferase family 8. *Plant Physiology*, **153**, 1729-1746.

Yonekura-Sakakibara, K. and Hanada, K. (2011) An evolutionary view of functional diversity in family 1 glycosyltransferases. *Plant J*, **66**, 182-193.

Yu, C.S., Chen, Y.C., Lu, C.H. and Hwang, J.K. (2006) Prediction of protein subcellular localization. *Proteins*, **64**, 643-651.

Table 1. Subcellular localization of non-family (NF; 1 - 2 copies per genome) genes, low-copy-number family (LF; 3 - 10 copies) genes and high-copy-number family (HF; >10 copies) genes. "Conserved" indicates that the genes are shared by at least two species.

Subcellular localization	**NF conserved** (%)	**NF species-specific** (%)	**LF conserved** (%)	**LF species-specific** (%)	**HF conserved** (%)	**HF species-specific** (%)
chloroplast	13.6*	1.6	13.4*	1.6	4.9	0.8
cytoplasm	20.7	9.0	24.0*	12.3	27.7*	29.2*
extracellular space	2.5	12.6*	5.0	20.4*	5.1	9.5
mitochondrion	3.8*	3.4	2.7	2.2	1.7	1.6
nucleus	49.4	71.0*	41.1	60.3*	44.8	57.1
plasma membrane	9.7*	2.4	13.7*	3.2	15.7*	1.8

*Overrepresentation (P<0.01) of gene groups in each subcellular localization; P value was calculated using the cumulative Poisson distribution.

Table 2. The relative abundance (i.e., proportion of genes in each gene group) of subcellular localization in different non-family (NF; 1 - 2 copies per genome) genes. NF1 (Viridiplantae-wide), NF2 (angiosperm-specific), NF3 (monocot-specific), NF4 (dicot-specific), and NF5 (species-specific) as defined in Figure 3.

Subcellular localization	**NF1** (%)	**NF2** (%)	**NF3 & NF4** (%)	**NF5** (%)
chloroplast	24.5*	1.9	4.3	1.6
cytoplasm	29.9*	1.9	8.5	9.0
extracellular space	0.7	3.8	26.6*	12.6
mitochondrion	3.9*	0.6	1.6	3.4
nucleus	35.1	86.0*	55.9	71.0*
plasma membrane	5.8*	5.7*	3.2	2.4

*Overrepresentation (P<0.01) of gene groups in each subcellular localization; P value was calculated using the cumulative Poisson distribution.

Figure Legends

Figure 1. Distribution of genes classified as non-family (**NF**; 1 - 2 copies per genome) genes, low-copy-number (**LF**; 3 - 10 copies) genes and high-copy-number (**HF**; >10 copies) genes in 14 plant species.

Figure 2. Number of single- and two-copy genes classified here as non-family genes in each of the 14 plant species. The number of single-copy genes is significantly (*P*<0.01, paired t-test) higher than that of two-copy genes.

Figure 3. Distribution of non-family (**NF**; 1 - 2 copies) genes among different phylogenetic groups. NF1 gene group contains NF genes having homologs in all the 14 species; NF2 contains NF genes having homologs in all and only the 10 angiosperm species; NF3 contains NF genes having homologs in all and only the four monocot species; NF4 contains NF genes having homologs in all and only the six dicot species; and NF5 contains species-specific NF genes. The numbers of genes in each species were normalized based on *Glycine*, which has the highest number of genes among the 14 species studied.

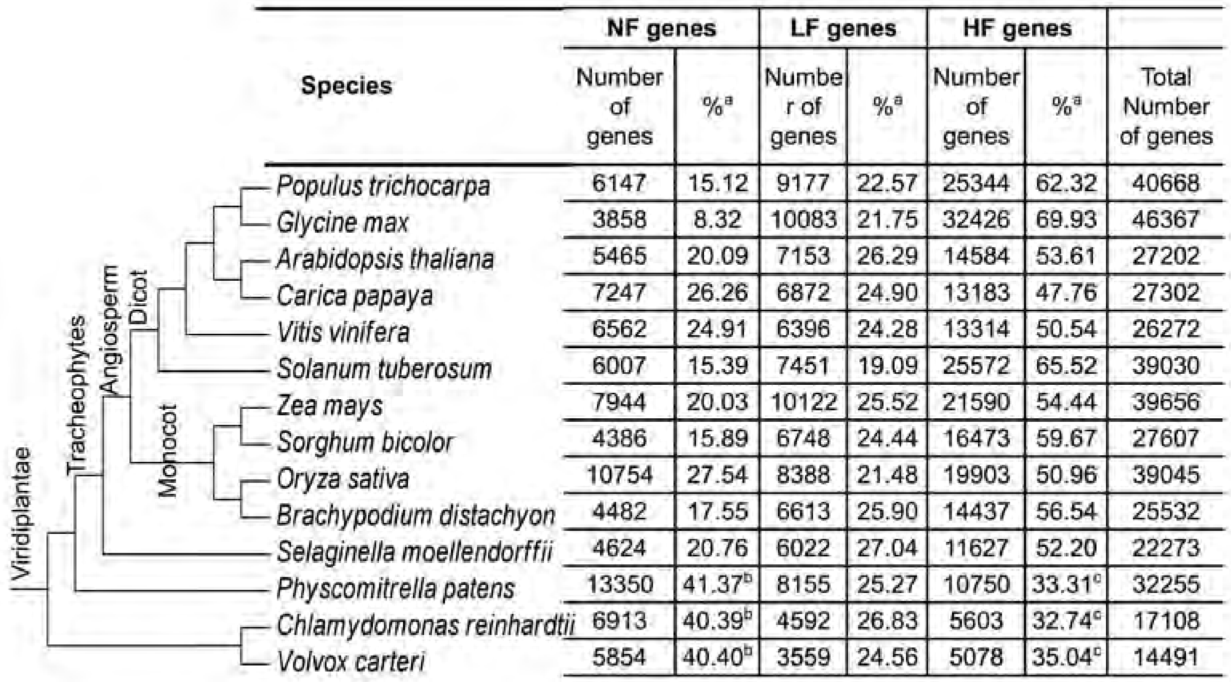
Figure 4. Duplication and loss of genes classified as non-family (**NF**; 1 - 2 copies), low-copy-number (**LF**; 3 - 10 copies) and high-copy-number (**HF**; >10 copies) genes in *Arabidopsis*, *Brachypodium*, *Glycine*, *Oryza*, *Populus*, *Sorghum*, *Vitis* and *Zea*. (A) The frequency of tandem duplication in the NF genes is significantly lower than that in both LF (*P*=0.00003) and HF (*P*=0.0001) categories, and the frequency of tandem duplication in the LF genes is significantly (*P*=0.0006) lower than that in HF genes; (B) the frequency of segmental duplication in NF genes was significantly lower than that in both LF (*P*=0.0015) and HF (*P*=0.0077) categories; and (C) NF genes had a significantly higher frequency of gene loss after syntenic segmental duplication than that in both LF (*P*=0.0028) and HF (*P*=0.0006) categories, and genes in the LF category had a higher frequency (P=0.0001) of gene loss than those in HF category. The significance (P values) of pairwise comparisons between NF, LF and HF are estimated by paired t-tests.

Figure 5. Biological processes over-represented (P<1E^{-77}, Fisher) in conserved (i.e., shared by at least two species) non-family (**NF**; 1 - 2 copies) genes as compared with conserved high-copy-number family (**HF**; >10 copies) genes. The GO terms were summarized using REVIGO (http://revigo.irb.hr/) and the enriched biological processes were listed in Table S1.

Figure 6. An example of *Arabidopsis* non-family (NF) genes involved in cell wall biosynthesis. (A) Gene co-expression network, with the yellow dots representing the non-family genes with unknown function, the blue dots representing known genes associated with cell wall biosynthesis, the red dots representing other co-expressed genes and the green lines connecting two co-expressed genes. The network was drawn using the NetworkDrawer function of ATTD-II (http://atted.jp/) with the NF gene AT2G31930 and its directly co-expressed genes as query; (B) expression pattern of the NF genes and associated known cell wall biosynthesis genes. Microarray expression data were obtained from AtGenExpress (http://www.weigelworld.org).

Figure 7. An *Arabidopsis* non-family (NF) gene involved in biotic stress response. (A) Gene co-expression network, with the yellow dots representing the non-family gene, the blue dots representing known biotic responsive genes, the red dots representing other co-expressed genes and the green lines connecting two co-expressed genes. The network was drawn using the NetworkDrawer function of ATTD-II (http://atted.jp/) with the NF gene AT3G01420 and its directly co-expressed genes as query; (B) expression pattern of the NF gene and associated biotic responsive genes in *Arabidopsis* under *Pseudomonas* treatment. Microarray expression data were obtained from AtGenExpress (http://www.weigelworld.org).

Figure 8. Hub genes in non-family (**NF**; 1 - 2 copies), low-copy-number family (**LF**; 3 - 10 copies) and high-copy-number family (**HF**; >10 copies) sets. Hubs were defined here as the top 10% most connected nodes in the *Arabidopsis* protein-protein interaction network (http://bioinfo.esalq.usp.br/atpin/atpin.pl).

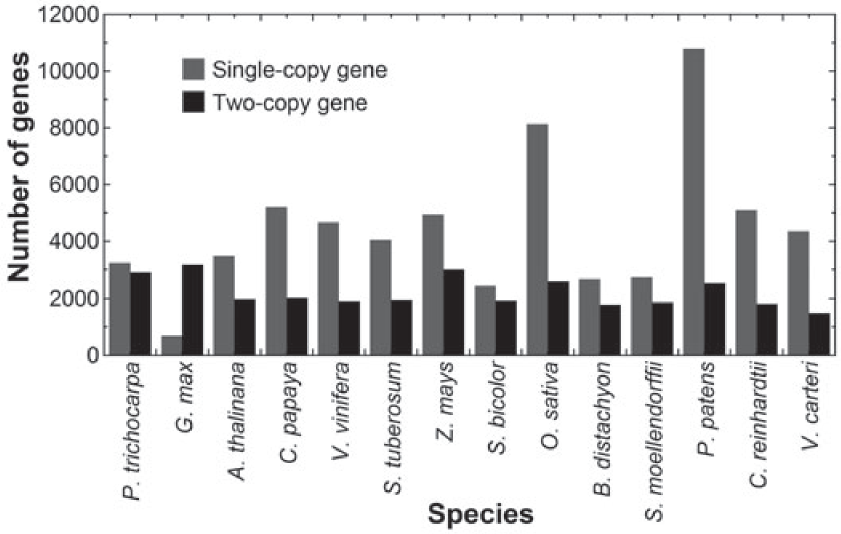

Viridiplantae
Tracheophytes
Angiosperm
Dicot
Monocot

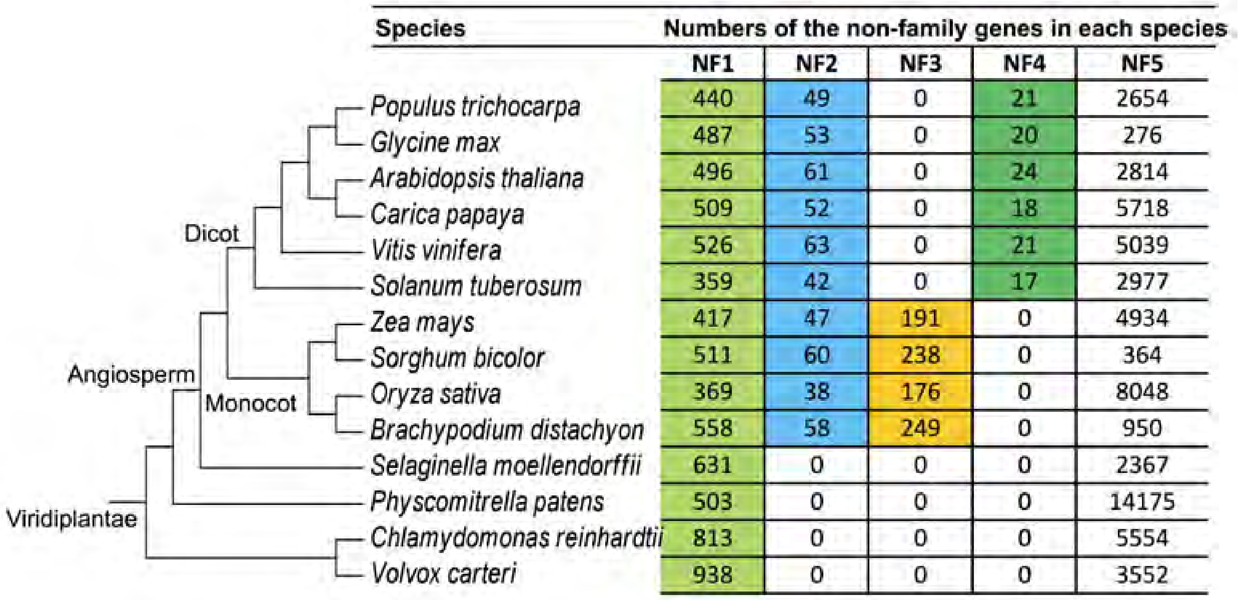
Species	NF genes		LF genes		HF genes		
	Number of genes	%[a]	Number of genes	%[a]	Number of genes	%[a]	Total Number of genes
Populus trichocarpa	6147	15.12	9177	22.57	25344	62.32	40668
Glycine max	3858	8.32	10083	21.75	32426	69.93	46367
Arabidopsis thaliana	5465	20.09	7153	26.29	14584	53.61	27202
Carica papaya	7247	26.26	6872	24.90	13183	47.76	27302
Vitis vinifera	6562	24.91	6396	24.28	13314	50.54	26272
Solanum tuberosum	6007	15.39	7451	19.09	25572	65.52	39030
Zea mays	7944	20.03	10122	25.52	21590	54.44	39656
Sorghum bicolor	4386	15.89	6748	24.44	16473	59.67	27607
Oryza sativa	10754	27.54	8388	21.48	19903	50.96	39045
Brachypodium distachyon	4482	17.55	6613	25.90	14437	56.54	25532
Selaginella moellendorffii	4624	20.76	6022	27.04	11627	52.20	22273
Physcomitrella patens	13350	41.37[b]	8155	25.27	10750	33.31[c]	32255
Chlamydomonas reinhardtii	6913	40.39[b]	4592	26.83	5603	32.74[c]	17108
Volvox carteri	5854	40.40[b]	3559	24.56	5078	35.04[c]	14491

[a] Percent of all the protein-encoding genes in each plant species.

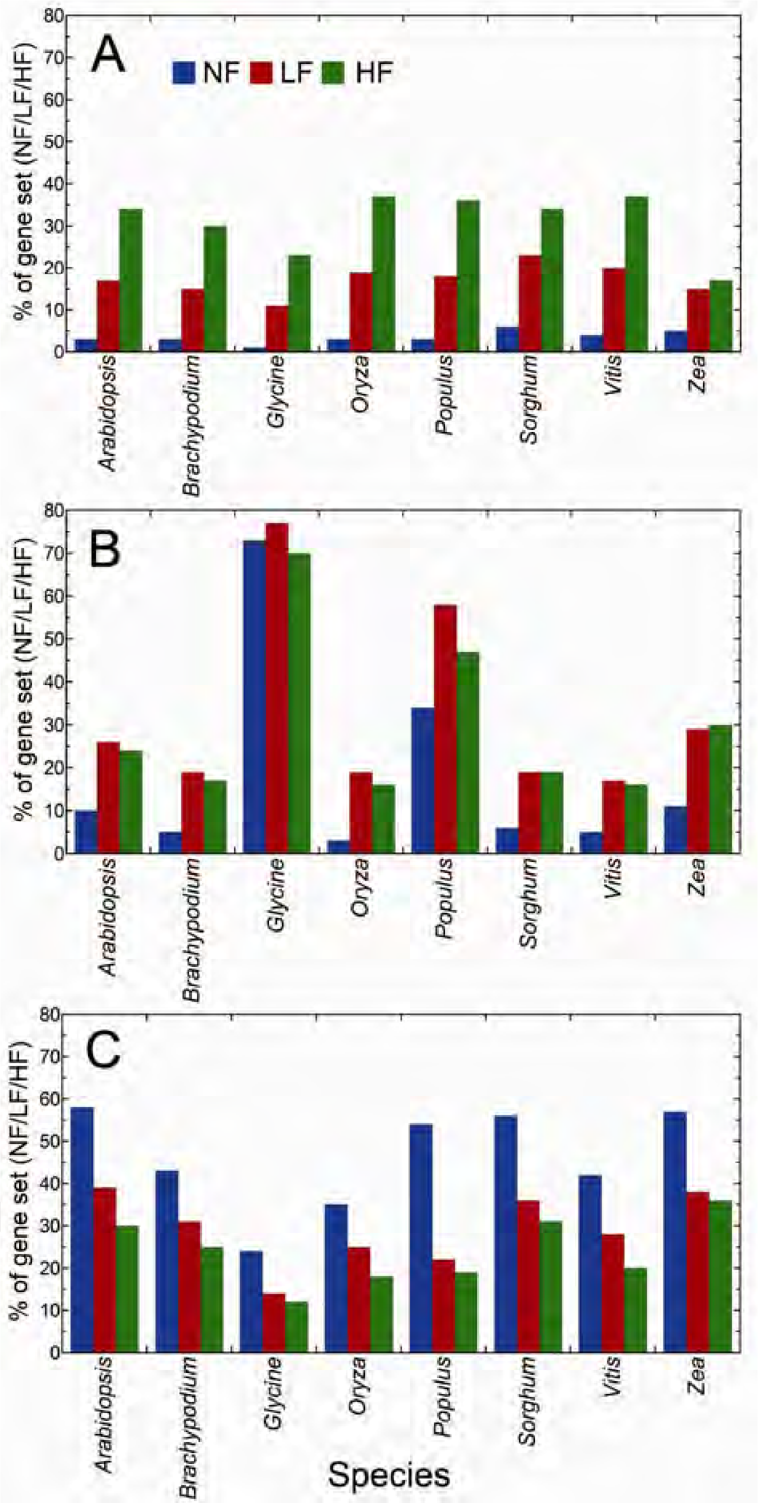
[b] Lower plants have significantly (*P*<0.01, Chi-square) higher proportion of NF genes than higher plants (i.e., Tracheophytes).

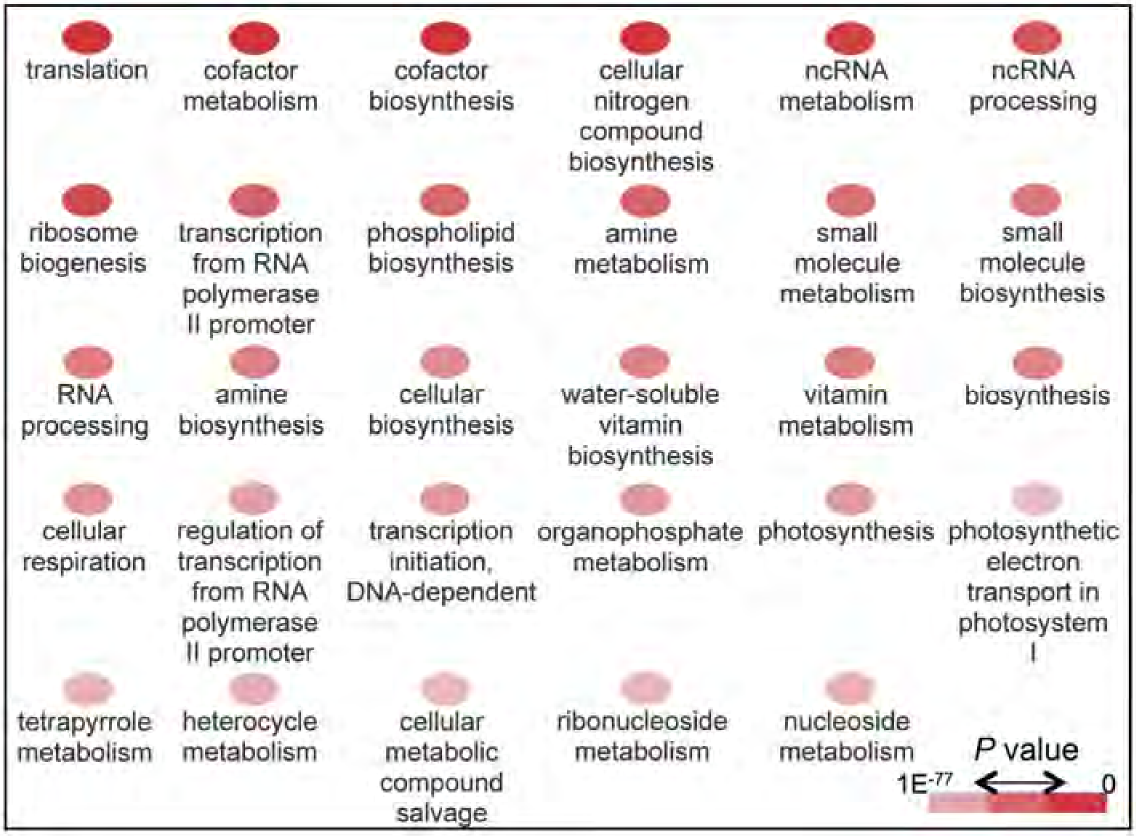
[c] Lower plants have significantly (*P*<0.01, Chi-square) lower proportion of HF genes than higher plants.



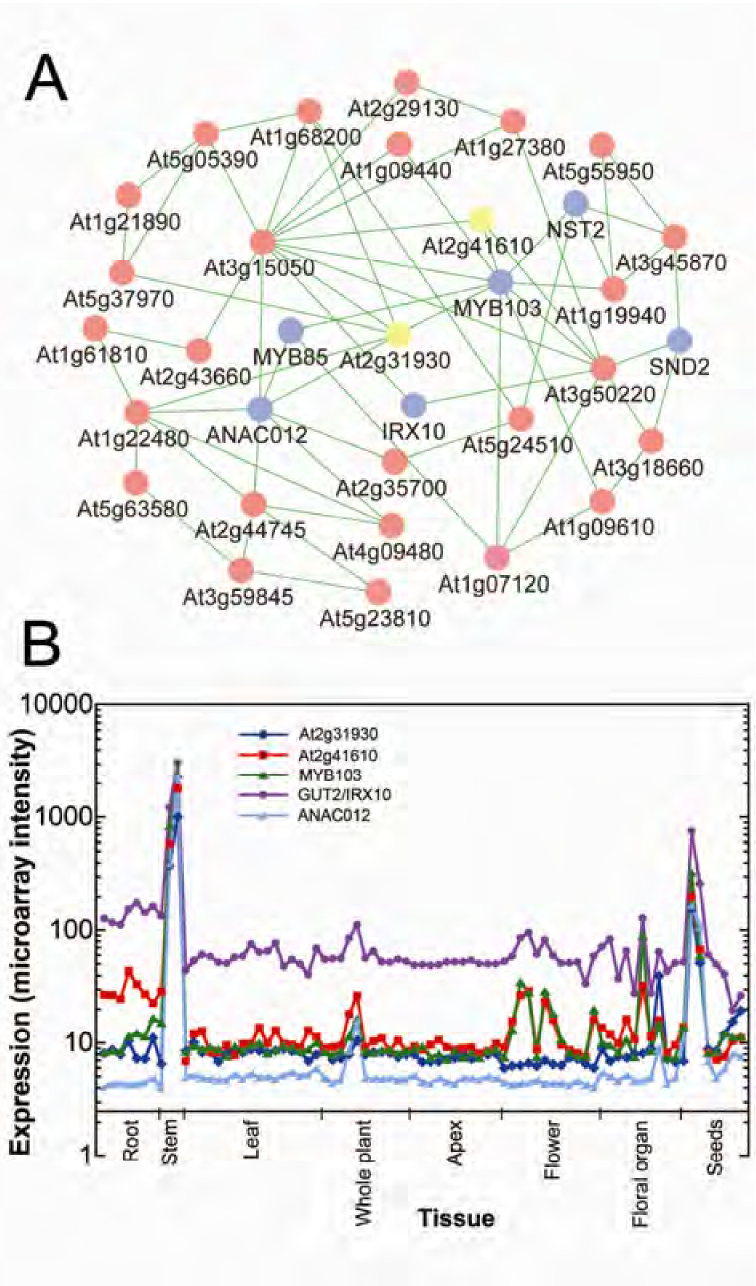

Viridiplantae
Angiosperm
Dicot
Monocot

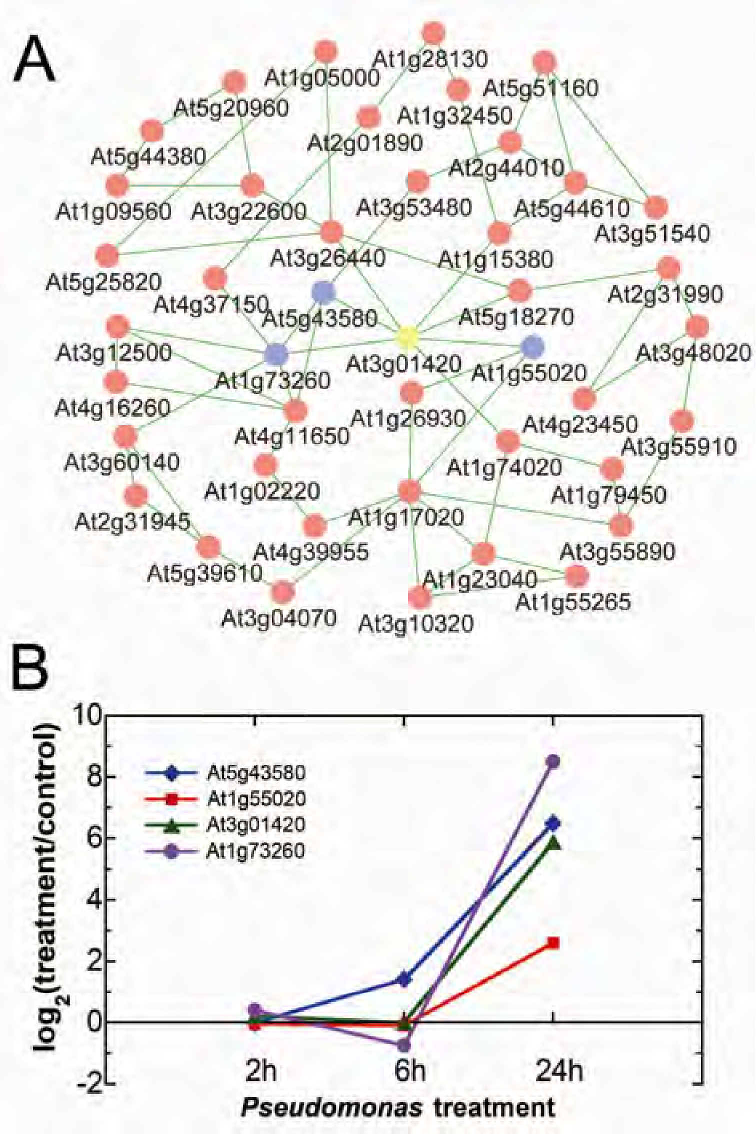
Species	Numbers of the non-family genes in each species				
	NF1	NF2	NF3	NF4	NF5
Populus trichocarpa	440	49	0	21	2654
Glycine max	487	53	0	20	276
Arabidopsis thaliana	496	61	0	24	2814
Carica papaya	509	52	0	18	5718
Vitis vinifera	526	63	0	21	5039
Solanum tuberosum	359	42	0	17	2977
Zea mays	417	47	191	0	4934
Sorghum bicolor	511	60	238	0	364
Oryza sativa	369	38	176	0	8048
Brachypodium distachyon	558	58	249	0	950
Selaginella moellendorffii	631	0	0	0	2367
Physcomitrella patens	503	0	0	0	14175
Chlamydomonas reinhardtii	813	0	0	0	5554
Volvox carteri	938	0	0	0	3552

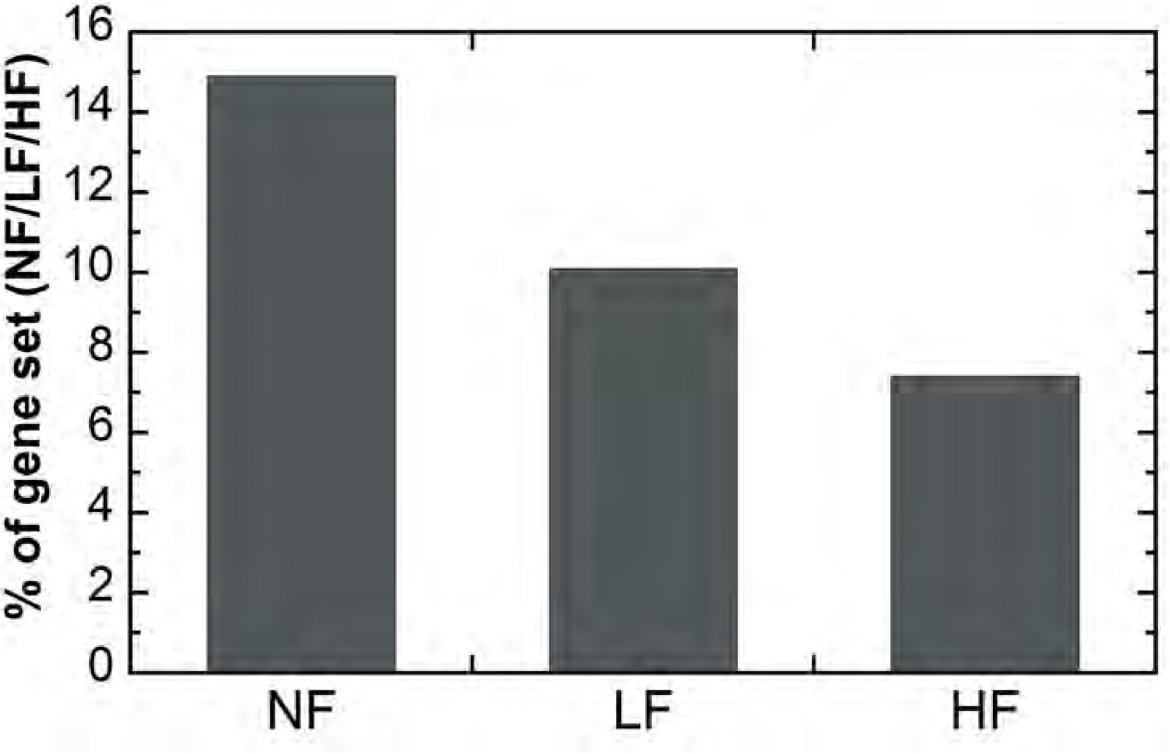

A
🟦 NF 🟥 LF 🟩 HF
B
C
% of gene set (NF/LF/HF)
0
10
20
30
40
50
60
70
80
Arabidopsis
Brachypodium
Glycine
Oryza
Populus
Sorghum
Vitis
Zea
Species



translation
cofactor metabolism
cofactor biosynthesis
cellular nitrogen compound biosynthesis
ncRNA metabolism
ncRNA processing
ribosome biogenesis
transcription from RNA polymerase II promoter
phospholipid biosynthesis
amine metabolism
small molecule metabolism
small molecule biosynthesis
RNA processing
amine biosynthesis
cellular biosynthesis
water-soluble vitamin biosynthesis
vitamin metabolism
biosynthesis
cellular respiration
regulation of transcription from RNA polymerase II promoter
transcription initiation, DNA-dependent
organophosphate metabolism
photosynthesis
photosynthetic electron transport in photosystem I
tetrapyrrole metabolism
heterocycle metabolism
cellular metabolic compound salvage
ribonucleoside metabolism
nucleoside metabolism
P value
1E-77
0


A
At2g29130
At1g68200
At5g05390
At1g27380
At1g09440
At5g55950
At1g21890
At2g41610
NST2
At3g15050
At3g45870
At5g37970
MYB103
At1g19940
At1g61810
MYB85
At2g31930
SND2
At2g43660
At3g50220
IRX10
At1g22480
ANAC012
At5g24510
At2g35700
At3g18660
At5g63580
At2g44745
At1g09610
At4g09480
At1g07120
At3g59845
At5g23810
B
10000
1000
100
10
1
Expression (microarray intensity)
At2g31930
At2g41610
MYB103
GUT2/IRX10
ANAC012
Root
Stem
Leaf
Whole plant
Apex
Flower
Floral organ
Seeds
Tissue


A
At1g28130
At1g05000
At5g51160
At5g20960
At1g32450
At2g01890
At5g44380
At2g44010
At1g09560
At3g22600
At3g53480
At5g44610
At3g51540
At3g26440
At1g15380
At5g25820
At2g31990
At4g37150
At5g43580
At5g18270
At3g12500
At3g48020
At3g01420
At1g55020
At1g73260
At4g16260
At1g26930
At4g23450
At4g11650
At3g55910
At3g60140
At1g74020
At1g02220
At1g79450
At2g31945
At1g17020
At4g39955
At3g55890
At5g39610
At1g23040
At1g55265
At3g04070
At3g10320
B
$\log_2$(treatment/control)
10
8
6
4
2
0
-2
At5g43580
At1g55020
At3g01420
At1g73260
2h
6h
24h
Pseudomonas treatment


% of gene set (NF/LF/HF)
16
14
12
10
8
6
4
2
0
NF
LF
HF