Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Genome-wide identification of lineage-specific genes in *Arabidopsis*, *Oryza* and *Populus*

Xiaohan Yang, Sara Jawdy, Timothy J. Tschaplinski, Gerald A. Tuskan *

Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6422, USA

ARTICLE INFO

Article history:
Received 8 October 2008
Accepted 15 January 2009
Available online 29 January 2009

Keywords:
Lineage-specific genes
Gene expression
Arabidopsis
Rice
Poplar
Perennial plants
Woody plants

ABSTRACT

Protein sequences were compared among *Arabidopsis*, *Oryza* and *Populus* to identify differential gene (**DG**) sets that are in one but not the other two genomes. The **DG** sets were screened against a plant transcript database, the NR protein database and six newly-sequenced genomes (*Carica*, *Glycine*, *Medicago*, *Sorghum*, *Vitis* and *Zea*) to identify a set of species-specific genes (**SS**). Gene expression, protein motif and intron number were examined. 165, 638 and 109 **SS** genes were identified in *Arabidopsis*, *Oryza* and *Populus*, respectively. Some **SS** genes were preferentially expressed in flowers, roots, xylem and cambium or up-regulated by stress. Six conserved motifs in *Arabidopsis* and *Oryza* **SS** proteins were found in other distant lineages. The **SS** gene sets were enriched with intronless genes.

The results reflect functional and/or anatomical differences between monocots and eudicots or between herbaceous and woody plants. The *Populus*-specific genes are candidates for carbon sequestration and biofuel research.

The identification of taxon specific genes has both scientific and practical values [1]. Recently, computational approaches were used to identify genes unique to bacteria [1], virus [2] and Poaceae [3]. Three of the fully-sequenced plant species, *Arabidopsis* [4], *Oryza* [5–7] and *Populus* [8], represent three major types of higher plants: annual eudicots, annual monocots and perennial eudicots, respectively. Identification of species-specific genes in these taxa may provide insights into the molecular features distinguishing monocot from eudicot or herbaceous from woody plants. Additionally, transcript assemblies have been coalesced from expressed sequences collected from the NCBI GenBank Nucleotide database for more than 250 plant species representing a wide range of the evolutionary lineages [9]. Finally, genome sequences have been published for *Carica* [10] and *Vitis* [11] and draft/partial genome sequences are available in the public domain for *Glycine* (http://www.phytozome.net/soybean), *Medicago* (http://www.medicago.org/), *Sorghum* (http://genome.jgi-psf.org/) and *Zea* (http://www.maizesequence.org). These genomic data provide a broad and robust comparative resource for identification of species-specific genes in plants.

In this study, we initially identified three differential gene (**DG**) sets in the context of *Arabidopsis*, *Oryza* and *Populus*, i.e., 1) *Arabidopsis* genes without homologs in *Oryza* or *Populus*, 2) *Oryza* genes without homologs in *Arabidopsis* or *Populus*, and 3) *Populus* genes without homologs in *Arabidopsis* or *Oryza*. Then we used these three **DG** sets to query a customized database containing more than 250 plant transcript assemblies [9] followed by a query against the NR protein database and a query against a customized database containing annotated protein sequences from six recently-sequenced genomes (*Carica*, *Glycine*, *Medicago*, *Sorghum*, *Vitis* and *Zea*). The **DG** genes that have no homologs in the other species revealed three sets of species-specific (**SS**) genes in *Arabidopsis*, *Oryza* and *Populus*. To gain insights into the functions of the **SS** genes, we compared their expression pattern using microarray/digital northern data and identified conserved protein motifs that were over-represented in each taxon. The exon–intron structures were also examined to aid in the understanding of differential gene evolution.

Results

Differential genes in Arabidopsis, Oryza and Populus

Using a BLASTp search with an *e*-value cutoff of 0.1, we identified three differential gene (**DG**) sets that contained expression evidence from EST or full-length cDNA (FL-cDNA) in the context of *Arabidopsis*, *Oryza* and *Populus* (Fig. 1). The *Arabidopsis* **DG** set contained 917 genes without homologs in *Oryza* or *Populus*; the *Oryza* **DG** set contained 2781 genes without homologs in *Arabidopsis* or *Populus*; the *Populus* **DG** set contained 594 *Populus* genes without homologs in *Arabidopsis* or *Oryza* (Fig. 1).

To investigate the relationship between the three **DG** sets and genes in other plant species, we used a tBLASTn search (with an *e*-value cutoff of 0.1) to query a customized database containing the transcript assemblies [9]. The 250 plant species represented in the database were manually divided into 11 sub-datasets: algae, moss, fern, herbaceous gymnosperms (gymnosperm_herb), woody

* Corresponding author. Fax: +1 865 576 9939.
E-mail address: tuskanga@ornl.gov (G.A. Tuskan).

0888-7543/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2009.01.002

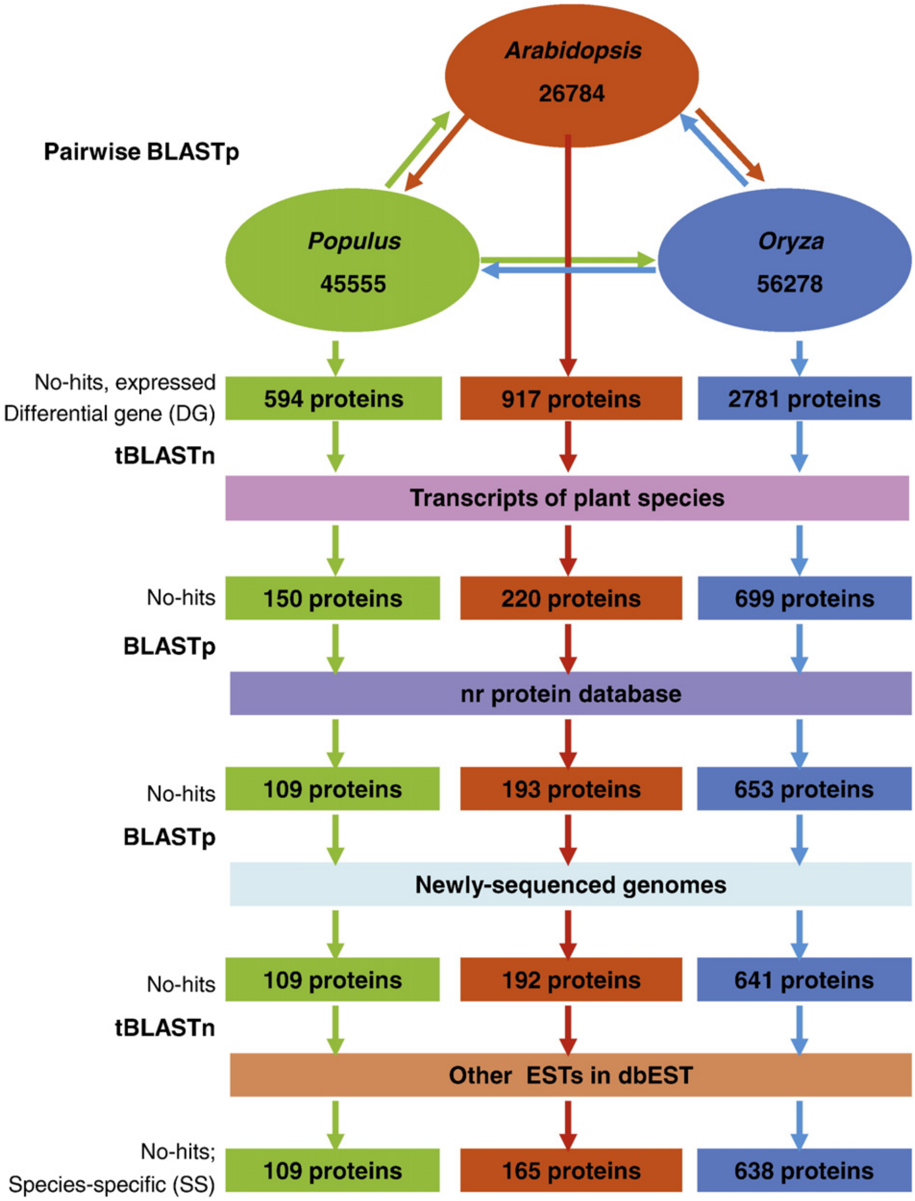

Fig. 1. Procedure for identifying species-specific genes in *Arabidopsis*, *Oryza* and *Populus*. The scoring matrix Blossum62, 80, Pam70, 30 were used for all the blast searches. The newly-sequenced genomes include *Carica*, *Glycine*, *Medicago*, *Sorghum*, *Vitis* and *Zea*.

gymnosperms (gymnosperm_woody), herbaceous basal angiosperms (angiosperm_basal_herb), woody basal angiosperms (angiosperm_basal_woody), herbaceous monocots (angiosperm_monocot_herb), woody monocots (angiosperm_monocot_woody), herbaceous eudicots (angiosperm_eudicot_herb) and woody eudicots (angiosperm_eudicot_woody). Results of the tBLASTn search revealed that the percentage of the *Arabidopsis* **DG** set with homology to herbaceous eudicots is higher than that of the *Oryza* **DG** set (90% vs. 26%, respectively; Fig. 2), whereas the percentage of the *Arabidopsis* **DG** set with homology to herbaceous monocots is lower than that of the *Oryza* **DG** set (21% vs. 92%, respectively; Fig. 2). These differences may be related to genes and processes that distinguish herbaceous monocots from herbaceous eudicots. The percentage of the *Populus* **DG** set with homology to woody eudicots is higher than that of the *Arabidopsis* **DG** set (77% vs. 26%, respectively; Fig. 2), whereas the percentage of the *Populus* **DG** set with homology to herbaceous eudicots is lower than that of the *Arabidopsis* **DG** set (71% vs. 90%, respectively; Fig. 2). These differences may be related to genes and processes that distinguish woody and herbaceous properties in eudicot plants.

Expression of genes in the differential gene sets

Gene expression in the *Arabidopsis* **DG** set that is associated with developmental and environmental responses was examined using a *K*-means clustering analysis of several *Arabidopsis* microarray datasets. For the developmental data set, using the whole seedling as the baseline reference, one cluster of 39 genes in the *Arabidopsis* **DG** set showed up-regulation in stamen (~8-fold) and pollen (~128-fold) (Supplementary Fig. S1 (Cluster 01)), one cluster of 27 genes showed up-regulated expression in root (~180-fold) (Supplementary Fig. S1 (Cluster 08)), one cluster of 49 genes showed up-regulated expression in flower tissues (~4000-fold) (Supplementary Fig. S1 (Cluster 10)) and one cluster of 88 genes showed up-regulated expression in mature pollen (~12-fold) (Supplementary Fig. S1 (Cluster 22)). Also, several clusters of genes exhibited down-regulated expression when

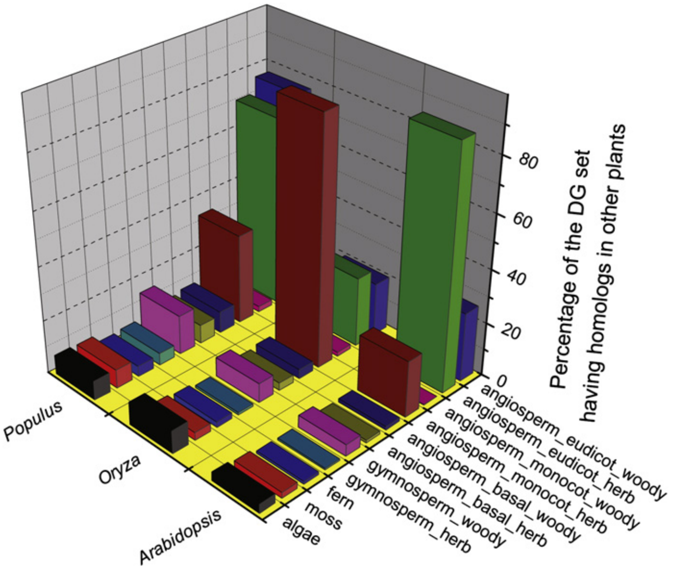

Fig. 2. The percentage of the differential gene sets in *Arabidopsis*, *Oryza* and *Populus* showing homology to other plant species as revealed by tBLASTn search against the transcript assemblies (TA) of more than 250 plant species (Childs et al., 2007), which were manually divided into 11 sub-datasets: algae, moss, fern, herbaceous gymnosperm (gymnosperm_herb), woody gymnosperm (gymnosperm_woody), herbaceous basal angiosperm (angiosperm_basal_herb), woody basal angiosperm (angiosperm_basal_woody), herbaceous monocot (angiosperm_monocot_herb), woody monocot (angiosperm_monocot_woody), herbaceous eudicot (angiosperm_eudicot_herb), and woody eudicot (angiosperm_eudicot_woody).

compared with the whole seedling, e.g. one cluster of 25 genes down-regulated in majority of the tissue types sampled (Supplementary Fig. S1 (Cluster 02)), one cluster of 6 genes down-regulated in shoot apex (~10-fold), carpel and pollen (~30-fold) (Supplementary Fig. S1 (Cluster 05)), and one cluster of 8 genes down-regulated in root (~70-fold), seed (~70-fold), stamen and pollen (~50-fold) (Supplementary Fig. S1 (Cluster 21)).

For the stress dataset, using untreated plants as the baseline reference, one cluster of 39 genes showed up-regulation under UV-B and biotic stress (~32-fold) (Supplementary Fig. S2 (Cluster 08)) and one cluster of 21 genes showed up-regulated expression under light treatments (~8-fold) (Supplementary Fig. S2 (Cluster 15)).

One cluster of 212 genes in the *Oryza* **DG** set showed relatively high levels of expression in the leaf tissue (Supplementary Fig. S3 (Cluster 20)) and one cluster of 239 genes showed relatively high levels of expression in panicle tissue (Supplementary Fig. S3 (Cluster 21)). Likewise, one cluster of 9 genes in the *Populus* **DG** set showed relatively high levels of expression in the male flowers (Supplementary Figs. S4 (Cluster 07) and S5A), one cluster of 20 genes showed relatively high levels of expression in the female flowers (Supplementary Figs. S4 (Cluster 10) and S5B), one cluster of 31 genes showed relatively high levels of expression in the xylem tissue (Supplementary Figs. S4 (Cluster 18) and S6) and one cluster of 65 genes showed relatively high levels of expression in the flower buds (Supplementary Fig. S4 (Cluster 21)). There is an overlap between the *Populus* differential genes up-regulated in woody tissues and the *Populus* differential genes having homology to other woody plants. Specifically, 23 of the 31 genes that were preferentially expressed in xylem have blast hits in other woody plants (Supplementary Table S4).

Species-specific genes in Arabidopsis, Oryza and Populus

Using the three **DG** sets to query the transcript assemblies [9] by tBLASTn (with an *e*-value cutoff of 0.1) followed by querying the NR protein database and the six newly-sequenced genomes (*Carica*, *Glycine*, *Medicago*, *Sorghum*, *Vitis* and *Zea*) using BLASTp (with an *e*-value cutoff of 0.1), we identified 192 *Arabidopsis*-, 651 *Oryza*- and 145 *Populus*-specific genes that had no homologs in other species. Of the 145 *Populus*-specific genes, 36 were incorrectly annotated, i.e., truncated (without start codon or stop codon) or interrupted by internal stop codon and consequently they were excluded from the final *Populus*-specific gene list. Of the 651 *Oryza* **SS** genes, 10 were transposable elements and they were excluded from the final *Oryza*-specific gene list. Further search (tblastn) against the ESTs from other species in the dbEST database revealed that 27 *Arabidopsis*-specific proteins had hits in other lineages (mostly in *Brassica*), and 3 *Oryza*-specific proteins had hits in other lineages in Poacea. Therefore, the final species-specific gene list includes 165 *Arabidopsis*-, 638 *Oryza*- and 109 *Populus*-specific genes (Supplementary Tables S1–S3).

Expression of the species-specific genes

Expression of the *Arabidopsis*-specific genes associated with developmental and environmental responses was investigated using *K*-means clustering analysis of *Arabidopsis* microarray data.

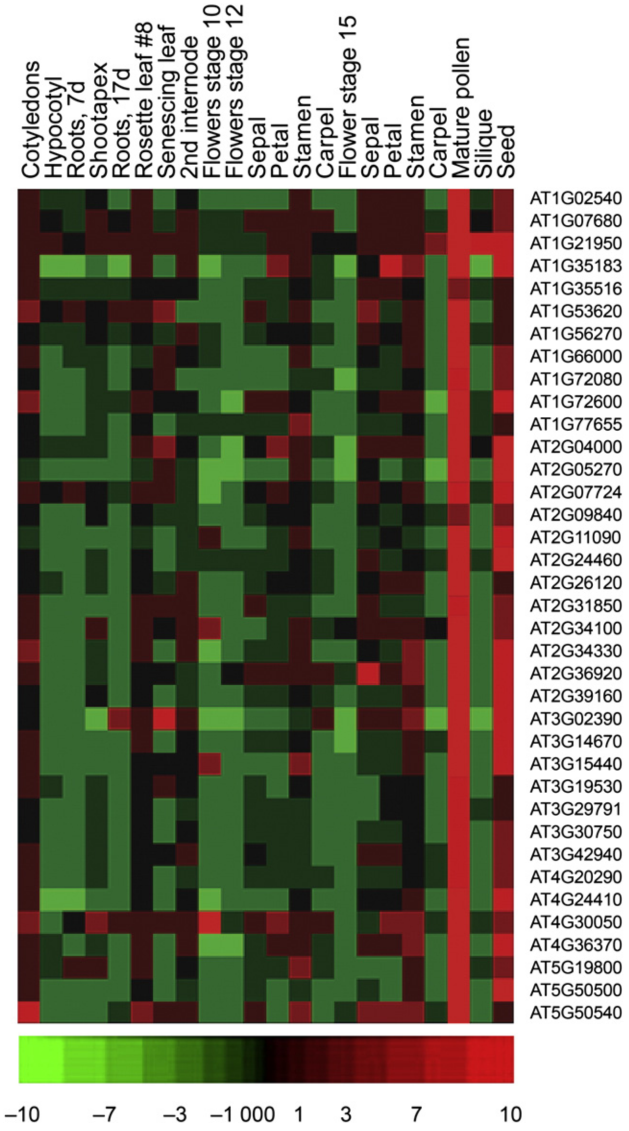

Fig. 3. Expression pattern of Cluster 01 (see Supplementary Fig. S7) of the *Arabidopsis* species-specific set as revealed by clustering analysis of the developmental microarray data. The whole seedling was used as a reference for comparison. The color scheme represents log ratio, with red indicating up-regulated and green down-regulated expression.

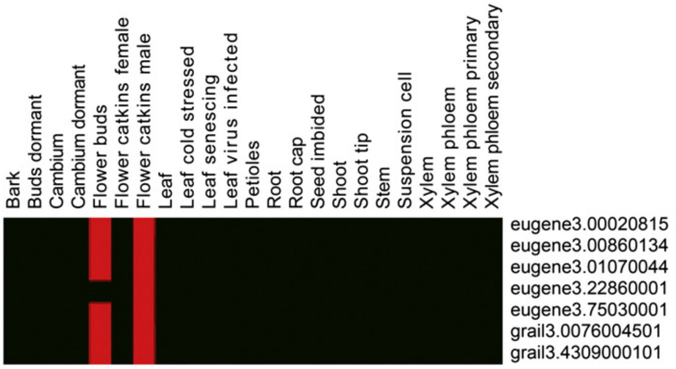

Fig. 4. Expression pattern of Cluster 06 (see Supplementary Fig. S10) of the *Populus* species-specific genes as revealed by clustering analysis of the digital northern data. Black represents no EST counts. Red indicates that EST counts are ≥1, with the brighter color the more EST counts.

For the developmental data set using the whole seedling as the baseline reference, one cluster of 37 *Arabidopsis*-specific genes showed up-regulated expression in mature pollen (~10-fold) (Fig. 3; Supplementary Fig. S7 (Cluster 01)), one cluster of 7 genes showed up-regulated expression in root (~500-fold) (Supplementary Fig. S7 (Cluster 05)) and one cluster of 5 genes showed up-regulated expression in stamen (~16-fold) and pollen (~256-fold) (Supplementary Fig. S7 (Cluster 07)). For the stress dataset using untreated plants as references, only a single cluster of 12 genes showed up-regulated expression under heat stress (~64-fold) (Supplementary Fig. S8 (Cluster 19)).

Expression of the *Oryza*-specific genes was evaluated using *K*-means clustering analysis of the *Oryza* digital northern data. One cluster of 39 *Oryza*-specific genes showed relatively high levels of expression in the flower tissue (Supplementary Fig. S9 (Cluster 14)), one cluster of 47 genes showed relatively high levels of expression in the pistil tissue (Supplementary Fig. S9 (Cluster 16)) and one cluster of 39 genes showed relatively high levels of expression in the root tissue (Supplementary Fig. S9 (Cluster 17)). Similarly, one cluster of 7 *Populus*-specific genes showed relatively high levels of expression in the female flowers (Fig. 4, Supplementary Fig. S10 (Cluster 06)), one cluster of 12 genes showed relatively high levels of expression in the xylem tissue (Supplementary Fig. S10 (Cluster 14)), one cluster of 15 genes showed relatively high levels of expression in the cambium tissue (Supplementary Fig. S10 (Cluster 15)) and one cluster of 5 genes showed relatively high levels of expression in the leaf tissue (Supplementary Fig. S10 (Cluster 18)).

Table 1
Conserved motifs identified by MEME using amino acid sequences of specific genes from *Arabidopsis* (Supplementary Table S1) and *Oryza* (Supplementary Table S2)

Motif ID	Motif consensus	Genes
Motif 1	D[E]-E-E-E-E-E-E[D]-E[R]-D[E]-D [E]-D[ELR]-E[R]	6 genes from *Arabidopsis* specific group 1
Motif 2	E-K[R]-x-E[R]-E[R]-G[R]-E-E[K]-E-K [E]-E-E-E-x-D[E]	22 genes from *Oryza* specific group 1
Motif 3	L[V]-R-F-N-D-E[K]-P[L]-R-G-D[NL]-L [PQS]-L[S]-L[W]-S[K]-P[QS]-V[G]-M [ET]-A[FHLP]-T[HP]-P[QT]-K[IN]	4 genes from *Oryza* specific group 1
Motif 4	H[D]-H-H-H-R[CG]-H-H-H	8 genes from *Oryza* specific group 1
Motif 5	K[ERS]-G[KN]-E[GIKL]-D[NRI]-S [KNQ]-D[KMT]-D[KE]-D[LS]-D[KLY]-D [EGQ]-S[TV]-D[FVY]-G[HSTW]-S[DK]-S [DNW]-N[GTV]-D[EKNV]-D[NG]-L [NSY]-D[EFG]-S[GT]-D[Q]-D[HY]-D[FL]-D [EKMS]-D[T]-A[DKSV]-M[DNS]-K[HLM]-D [CHK]-K[DH]-I[LD]-S[FGT]-D[ES]-L[SQ]-F [DHY]-K[HI]-D[EHLY]-K[CGS]-D[I]	4 genes from *Oryza* specific group 1
Motif 6	E[K]-E[R]-R[EG]-E[GL]-E[DR]-D[E]-G[S]-R [EK]-K-K[E]-K[L]-E[DK]-E[R]-E-K-E[EKRT]-K	4 genes from *Oryza* specific gene expression Cluster 18

Table 2
Other species containing motifs listed in Table 1, as revealed by MAST search of the NR protein database

Motif	Organism		Protein	
	Taxonomy	Count	Name	Count
Motif 1				
	Eukaryota; Metazoa	6	Hypothetical protein	5
	Eukaryota; Alveolata	3	Unknown protein	1
	Eukaryota; Fungi	1	Hmgb1-prov protein	1
			Casein kinase II beta subunit CKB1	1
			alpha 2B adrenergic receptor	1
			MGC85274 protein	1
Motif 2				
	Eukaryota; Metazoa	9	Hypothetical protein	9
	Archaea; Euryarchaeota	1	Similar to melanoma antigen family A	1
Motif 3				
	Bacteria; Proteobacteria	1	Hypothetical protein	1
	Viruses; dsDNA viruses	9	lef-8	9
Motif 4				
	Bacteria; Firmicutes	1	Hypothetical protein	5
	Eukaryota; Alveolata	2	Histidine-rich protein	2
	Eukaryota; Metazoa	4	Unknown protein	1
	Bacteria; Proteobacteria	2	SJCHGC08151 protein	1
	Eukaryota; Mycetozoa	1	Proline-rich region	1
Motif 5				
	Eukaryota; Alveolata	5	Hypothetical protein	8
	Eukaryota; Mycetozoa	4	MIF4G domain protein	1
	Eukaryota; Fungi	1	Adenylate and guanylate cyclase catalytic Domain containing protein	1
Motif 6				
	Eukaryota; Metazoa	6	Hypothetical protein	6
	Bacteria; Firmicutes	1	LOC566027 protein	1
	Archaea; Crenarchaeota	1	LRRGT00096	1
	Eukaryota; Viridiplantae;	2	Unknown protein	1
			lti45 (low-temperature-induced protein)	1
			LTI45	

Only the top ten hits are listed.

Over-represented protein motifs in the species-specific genes

To identify the over-represented protein motifs, the **SS** gene set in each species were divided into groups based on protein sequences using the CLUSS program [12]. The **SS** genes in *Arabidopsis*, *Oryza* and *Populus* were divided into 6 (Supplementary Table S1), 5 (Supplementary Table S2) and 4 groups (Supplementary Table S3), respectively. Five protein motifs (Motifs 1–5) were identified by the MEME program from the group 1 sequences in *Arabidopsis* and *Oryza* (Table 1). In addition, one motif was identified by MEME from the *Oryza* **SS** gene expression Cluster 18 (Table 1; Supplementary Fig. S9). To investigate the evolutionary history of the protein motifs, the motif scoring matrix were used to query the NR protein database using the MAST program. The over-represented motifs identified from the **SS** sets in *Arabidopsis* and *Oryza* were also found in unknown or hypothetical proteins of other distantly-related organisms, most of which are in the early branches of the three of life, such as Archaea, Bacteria, Fungi and Metazoan (Table 2).

Real-time PCR confirmation of Populus gene expression

Because the annotation of the newly-sequenced *Populus* genome is in draft form and may contain annotation errors, we performed real-time RT-PCR analysis of 15 **SS** genes and 38 **DG** genes in five-tissues (i.e., bark, leaf, root, shoot tip and stem) from *Populus*. Fifty of the 53 *Populus* genes (94.3%) showed detectable expression in the tissues

Table 3
Multiple-tissue real-time RT-PCR analysis of expression of *Populus* specific (**SS**) and differential (**DG**) genes

Category	Gene name	Bark	SEM	Leaf	SEM	Root	SEM	Shoot	SEM	Stem	SEM
SS	estExt_fgenesh4_kg.C_LG_I0055[a]	130.6	39.9	14.2	3.8	140.0	38.6	1.0	0.6	91.8	29.8
SS	estExt_fgenesh4_pg.C_2630012	1.1	0.0	2.4	0.4	1.5	0.2	1.1	0.1	1.0	0.2
SS	eugene3.00020815	1.0	0.3	12.2	2.7	1.5	0.2	1.6	0.4	7.6	1.6
SS	eugene3.00021257	1.0	0.1	4.3	0.4	1.6	0.2	1.2	0.1	2.5	0.3
SS	eugene3.00091626	1.4	0.2	3.5	0.3	1.2	0.1	2.5	0.4	1.0	0.1
SS	eugene3.00102359	1.7	0.4	78.7	15.4	6.8	0.6	1.0	0.2	31.8	5.7
SS	eugene3.00120934	1.0	0.4	27.5	4.5	11.7	1.8	7.6	0.0	10.8	1.9
SS	eugene3.00280166	2.2	0.1	9.1	1.2	9.4	0.9	1.0	0.1	14.6	2.1
SS	eugene3.00400046[a]	7.4	0.9	2.8	0.4	19.8	4.9	12.5	1.8	1.0	0.2
SS	eugene3.01070044	1.6	0.4	11.2	2.7	1.0	0.2	5.1	1.1	0.0	0.0
SS	eugene3.02230021	18.2	3.3	4.8	0.3	2.1	0.5	1.0	0.2	0.0	0.0
SS	eugene3.04600002	1.2	0.1	5.3	0.5	1.0	0.1	2.4	0.2	1.8	0.2
SS	eugene3.14950001	1.4	0.1	5.5	0.5	1.9	0.2	1.0	0.1	2.0	0.2
SS	grail3.0043007301	3.1	0.8	11.2	1.7	1.0	0.2	2.0	0.4	2.8	0.6
SS	grail3.0169000301	1.4	0.2	2.8	0.2	1.6	0.2	1.3	0.1	1.0	0.1
DG	estExt_fgenesh4_kg.C_280013	6.0	1.1	95.7	31.9	1.0	0.3	6.2	0.7	16.8	2.6
DG	estExt_fgenesh4_pg.C_14770002	1.0	0.2	18.3	2.3	11.3	1.3	6.6	0.0	10.0	1.5
DG	estExt_fgenesh4_pg.C_290045	16.8	2.2	228.0	63.2	1.0	0.4	47.2	7.3	13.1	2.1
DG	estExt_fgenesh4_pg.C_LG_II1111[a]	1.0	0.3	292.4	66.6	29.6	3.8	2.7	0.5	103.8	27.4
DG	estExt_fgenesh4_pg.C_LG_VIII0722[a]	1.0	0.1	18.3	2.4	22.7	3.3	6.1	1.0	7.8	1.2
DG	estExt_fgenesh4_pg.C_LG_X0606	16.5	2.1	11.8	1.9	1.0	0.2	1.4	0.4	16.5	3.2
DG	estExt_fgenesh4_pg.C_LG_XVIII0908	40.1	6.5	29.8	5.6	4.1	0.5	1.0	0.3	5.3	0.8
DG	eugene3.00011774[a]	1.5	0.2	34.0	4.5	14.9	1.6	1.0	0.1	41.0	7.5
DG	eugene3.00051028[a]	1.0	0.2	22.0	2.5	11.5	1.4	15.2	0.9	6.0	0.9
DG	eugene3.00051604	0.0	0.0	1.5	0.4	1.1	0.2	1.0	0.0	1.1	0.1
DG	eugene3.00060029	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.1
DG	eugene3.00060132	1.0	0.3	21.0	2.6	2.3	0.4	7.1	8.2	6.7	1.0
DG	eugene3.00060513	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DG	eugene3.00061307	1.3	0.1	1.0	0.2	1.9	0.2	1.1	0.3	1.0	0.1
DG	eugene3.00070428	7.7	0.7	36.1	7.1	1.0	0.2	1.3	0.3	19.6	3.7
DG	eugene3.00080951	31.8	4.0	1.0	0.2	7.2	0.6	19.9	1.5	1.8	0.3
DG	eugene3.00081749[a]	47.1	4.4	92.0	12.9	160.6	31.0	1.0	0.3	194.0	35.0
DG	eugene3.00090211	11.9	1.6	14.6	1.2	1.0	0.1	2.8	0.3	7.8	1.0
DG	eugene3.00101292	1.9	0.2	3.7	0.3	1.0	0.1	2.1	0.1	1.7	0.2
DG	eugene3.00120867	6.0	0.6	7.5	0.7	1.0	0.1	2.3	0.3	5.3	0.6
DG	eugene3.00121045	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DG	eugene3.00180855	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DG	eugene3.00880016	2.2	0.2	4.5	0.5	1.2	0.1	2.4	0.1	1.0	0.1
DG	eugene3.02650003	11.7	1.8	3.9	0.5	1.9	0.3	1.7	0.0	1.0	0.2
DG	eugene3.08380002	1.6	0.2	4.4	0.6	4.1	0.4	1.0	0.0	1.8	0.3

For each gene, the gene expression level in the lowest expressed tissue was set to 1.0 as a reference for comparison and the gene expression levels in the other tissues were ratios relative to the reference.

[a] These genes showing relatively high expression in the root were placed into our *Populus* transformation pipeline for constitutive over-expression and RNAi knockdown.

sampled, indicating that the *Populus* **SS** and **DG** gene sets are functional (Table 3).

Exon–intron structure

To study gene structure, we examined the intron composition of the **SS** genes by dividing gene structures into 4 types: intronless, 1 intron, 2 introns and 3-or-more introns per gene. The **SS** sets in all three studied species (*Arabidopsis*, *Oryza* and *Populus*) contain more intronless genes than expected by chance alone when compared all other genes in each genomes ($P<1\times10^{-5}$) (Fig. 5). We performed GC content analysis for lineage-specific genes and found that the GC content in the coding region of the intronless genes (57%) is significantly higher than that of intron-containing genes (53%) ($P<1\times10^{-6}$). There are intronless genes that are also found in the protein motif groups. Specifically, 44% (=23/48) of the genes in the conserved protein motif groups lack introns. In particular, 59% (=13/22) of the genes in the conserved protein motif #2 (Table 1) group in *Oryza* lack introns.

Discussion

While flowering plants share many common aspects in growth and development, they have distinguishing features at the taxonomic level. Phylogenetic analyses based on both morphological and molecular data have obviously separated *Oryza* (monocot) from the eudicot species *Arabidopsis* and *Populus* [13]. Similarly, our results demonstrate that the percentage of the *Arabidopsis* or *Populus* **DG** showing homology to eudicots is higher than that of the *Oryza* **DG** set, whereas the percentage of the *Arabidopsis* or *Populus* **DG** showing homology to monocots is lower than that of the *Oryza* **DG** set. Although both *Arabidopsis* and *Populus* are eudicots, they also have distinct growth and developmental habits, an herbaceous annual vs. a woody perennial, respectively. Some of the molecular elements responsible for the differences between *Arabidopsis* and *Populus* are revealed by our results where the percentage of the *Populus* **DG** set with homology to woody eudicots is higher than that of the *Arabidopsis* **DG** set and the percentage of the *Populus* **DG** set with homology to herbaceous eudicots is lower than that of the *Arabidopsis* **DG** set. In combination these results demonstrated that some genes in either **DG** set are preferentially expressed in root, flower or xylem tissue. The differences in gene expression may be reflective of the differences in reproductive or stem anatomy characteristics in *Arabidopsis* and *Populus*.

Although the species-specific genes in *Arabidopsis* and *Oryza* have no homologs in other species, they do share some motifs with other organisms, mostly in the early branches of the tree-of-life. It is interesting that the proteins in other organisms sharing these motifs are largely functionally-unknown or hypothetical. Future molecular and biochemical experiments will be needed to investigate the functions of the unknown protein motifs. It is possible that the

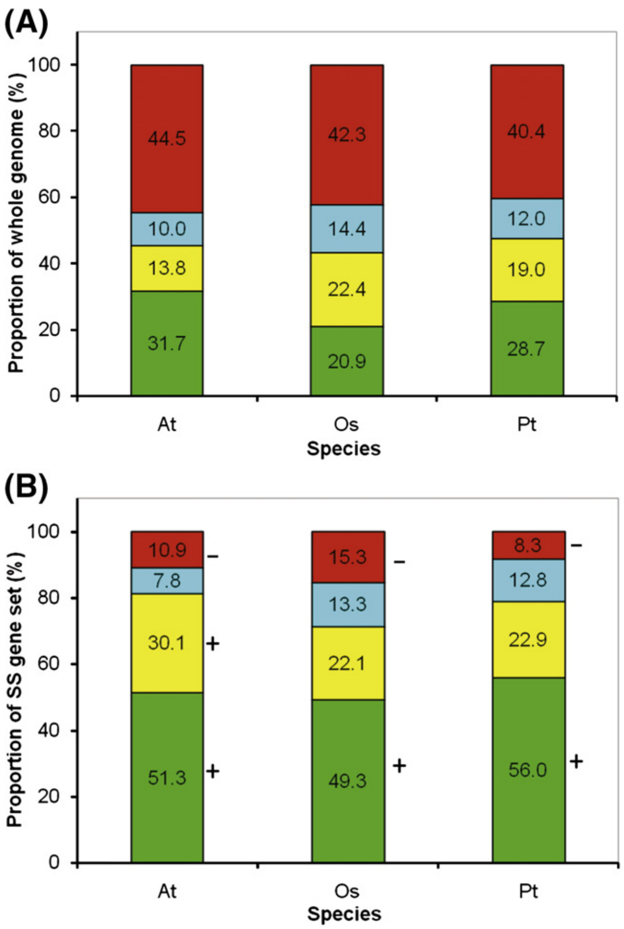

Fig. 5. Number of introns per gene in *Arabidopsis* (At), *Oryza* (Os) and *Populus* (Pt). (A) All genes in the whole genome; (B) Species-specific (**SS**) genes. "+" indicates that species-specific genes are over-represented and "−" under-represented at $P<1\times10^{-5}$, as compared with all genes in the whole genome. Gene structures were divided into 4 types: 0 (green), 1 (yellow), 2 (blue), and 3-or-more introns (red) per gene.

conserved motifs in the species-specific proteins of *Arabidopsis* and *Oryza* may have resulted from domain co-option from other organisms. Intraspecific protein domain co-option has been reported in *Populus* gene evolution [8,14]. It is also possible that the shared motifs are contained in ancestral genes that occurred early in gene evolution and have been subsequently lost in most higher plants.

Carmel et al. [15] have inferred that high intron density was reached in the early evolutionary history of plants and the last common ancestor of multicellular life forms harbored approximately 3.4 introns per kb, a greater intron density than in most of the extant fungi and in some animals. A recent report also implies that rates of intron creation were higher during earlier periods of plant evolution [16]. We recently reported that lineage-specific F-box gene is over-represented by intronless gene structure [17]. In this study, we also found that intronless genes were enriched in the species-specific gene sets as compared with the whole-genome annotation gene set. It is tempting to hypothesize that these species-specific sets resulted from recent lineage-specific expansion. Further studies are needed to test this hypothesis. We found that the GC content in the coding region of the intronless lineage-specific genes is significantly higher than that of intron-containing genes. This is consistent with previous reports revealing that high GC content class was enriched with intronless genes in plants [18,19].

The expression pattern of the species-specific genes indicates that some of these genes are associated with flower, root, leaf, or xylem development as well as stress response. We suggest that these tissue-specific or stress-responsive species-specific genes are involved in the molecular processes underlying the taxa phenotypic features in these plant species. As such, the species-specific genes will be valuable source of information for understanding the molecular mechanism underlying the distinguishing features in growth and development in the three model plant species. Future experiments involving over-expression and/or RNAi knockdown are needed to decipher the functions of these specific genes. Genes preferentially expressed in the *Populus* root tissue are potential candidate genes for carbon sequestration research. Eight *Populus* **DG**/**SS** genes showing relatively high expression in the root tissue have been placed in our *Populus* transformation pipeline for functional studies using the over-expression and RNAi knockdown strategy.

The number of **SS** genes in *Oryza* (638) is much higher than that in *Arabidopsis* (165) or *Populus* (109), even though the number of genes in the *Oryza* genome (42,653) is equivalent to that in the *Populus* genome (45,555) and 1.5 times that in the *Arabidopsis* genome (27,000) [8,20,21]. Even though the **SS** gene set in *Oryza* was identified by query against 1) the transcript assemblies from algae, moss, fern, gymnosperm to angiosperm including both monocot and eudicot, 2) the NR protein database which contains proteins identified in many diverse organisms, and 3) the *Carica*, *Glycine*, *Medicago*, *Sorghum*, *Vitis* and *Zea* genomes, we cannot conclusively determine that all of these genes arose from lineage-specific expansion after speciation. Annotation and assembly errors may account for some of the predicted gene models in the examined species. These issues will be resolved with additional genomic sequence, publicly available expression data sets, and functional characterization of the **SS** gene set.

Materials and methods

Genome sequences

The *Arabidopsis* protein sequences were downloaded from TAIR release 7 (http://www.arabidopsis.org/). The *Oryza* protein sequences were downloaded from TIGR release 5 (http://rice.plantbiology.msu.edu/). The *Populus* protein sequences were downloaded from JGI release 1.1 (http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html). The TIGR Plant Transcript Assemblies were downloaded from: http://plantta.tigr.org/index.shtml, which were built from the expressed transcripts in more than 250 plant species collected from dbEST (ESTs) and the NCBI GenBank nucleotide database (full-length and partial cDNAs) [9]. *Carica* protein sequences were downloaded from ftp://asgpb.mhpcc.hawaii.edu/papaya/annotation.genbank_submission/. *Glycine* protein sequences were downloaded from ftp://ftp.jgi-psf.org/pub/JGI_data/Glycine_max/. *Medicago* protein sequences were downloaded from http://www.medicago.org/genome/. *Sorghum* protein sequences were downloaded from http://genome.jgi-psf.org/Sorbi1/Sorbi1.home.html. *Vitis* protein sequences were downloaded from http://www.genoscope.cns.fr/spip/Vitis-vinifera-whole-genome.html. *Zea* protein sequences were downloaded from http://www.maizesequence.org/index.html.

Homolog search

The species-specific genes were identified in a pipeline (Fig. 1) based on a homolog search using BLASTp or BLASTn [22] with an *e*-value cutoff of 0.1 and scoring matrix of Blossum62, Blossum80, Pam70 and Pam30. Because the *e*-value of BLAST search is influenced by the database size, we use the standardized NCBI NR protein database size for all the BLAST searches in this study.

Expression evidence

The expression evidence from EST or full-length cDNA (FL-cDNA) for *Arabidopsis* genes were obtained from TAIR release 7 (http://www.arabidopsis.org/). The expression evidence from EST or FL-cDNA for *Oryza* genes were obtained from TIGR release 5 (http://rice.plantbiology.msu.edu/). The expression evidence from EST or FL-cDNA for *Populus* genes was determined by minimal 97% identity over an alignment of at least 100 bp and at least 80% length of the shorter sequences [17].

Analysis of gene expression

Two *Arabidopsis* microarray datasets were compiled from AtGen-Express [23,24]. The developmental data set contains cotyledons, hypocotyl, roots (7 or 17 d), shoot apex (vegetative), rosette leaf (# 8), senescing leaves, 2nd internode, flowers (stage 10/11, 12 or 15), sepals (flowers stage 12 or 15), petals (flowers stage 12 or 15), stamens (flowers stage 12 or 15), carpels (flowers stage 12 or 15), mature pollen), siliques (with seeds stage 5; late heart to mid torpedo embryos) and seeds (stage 10, without siliques; green cotyledons embryos). The gene expression levels are expressed as $\mathrm{LOG}_2(x/y)$, where x is the detection signal from the above tissue types and y is the detection signal from seedling (green parts).

The environmental data set contains cold (4 °C; 1 or 3 h), salt (150 mM NaCl; 1 or 3 h), drought treatments (1 or 3 h after 15 min dry air stream leading to 10% loss of fresh weight), oxidative treatments (10 μM methyl viologen; 1 h or 3 h), UV-B (1 or 3 h after 15 min exposure to ultraviolet-B light, 1.18 $\mathrm{W/m^2}$ Philips TL40W/12), heat (38 °C; 1 or 3 h), pathogen (*Phytophthora infestans*; 6, 12 or 24 h; in control treatments, H_2O was applied to leaves), blue light treatment (4 h), far-red light treatment (4 h), red light treatment (4 h) and white light treatment (4 h). Dark treatment (4 h) was used as a control for the light experiments. *K*-means clustering of the *Arabidopsis* microarray data was performed using EPCLUST (http://ep.ebi.ac.uk/EP/EPCLUST/) with correlation distance (uncentered).

For *Oryza* and *Populus* EST analysis, the coding sequences were used to search the dbEST using BLASTn. Expression evidence from EST sequences was determined by minimal 97% identity over an alignment of at least 100 bp and at least 80% length of the shorter sequences [17]. *K*-means clustering of the square-root transformed EST data (EST counts per tissue) was performed using EPCLUST (http://ep.ebi.ac.uk/EP/EPCLUST/) with Euclidean distance.

Protein classification

The specific protein sequences were clustered into groups using the CLUSS program [12].

Motif identification

Protein motifs of specific genes were identified statistically using MEME [25] with motif length set as 6 to 100, maximum motif number<100, and *e*-value<0.005. The MAST program [26] was used to search protein motifs.

Real-time PCR

P. trichocarpa ‘Nisqually-1’ stem and leaf tissues were taken from plants grown *in vitro* on media containing Murashige and Skoog salts [27], 3% sucrose and 0.25% Gelrite (PhytoTechnology Laboratories) at 23 °C ± 1 °C under cool-white fluorescent light (approximately 125 mmol $\mathrm{m^{-2}\ s^{-1}}$, 16-h photoperiod). Root, shoot tip, petiole and bark tissues were taken from plants grown in a greenhouse under natural lighting and temperatures ranging from 25 °C to 35 °C. Total RNA was extracted from root, stem, shoot tip, petiole, leaf and bark using the Spectrum Plant Total RNA kit (Sigma-Aldrich) and then treated with AMPD1 DNase I (Sigma-Aldrich) to eliminate DNA, according to the manufacturer's instructions. RNA purity was determined spectrophotometrically and quality was determined by examining rRNA bands on agarose gels. cDNA was synthesized from 2 μg of RNA using the PowerScript PrePrimed Single Shots with random hexamers as primer (CLONTECH Laboratories) in a 20 μl reaction.

For real-time RT-PCR analysis using gene-specific primers the cDNA was diluted 50-fold. Amplification reactions (25.0 μl) were carried out using iQ™ SYBR® Green Supermix according to the instructions provided by Bio-Rad Laboratories. Each reaction contained a cDNA template (1.0 μl), SYBR® Green supermix (12.5 μl), sterile water (8.5 μl) and the appropriate forward and reverse 5 μM primer pair (1.5 μl each). The gene used as a control to normalize the data for differences in input RNA and efficiency of reverse transcription between the samples was an actin gene expressed at a constant rate across tissue types. PCR amplification reactions were performed in triplicate. The thermal cycling conditions took place on an iCycler Real Time PCR detection system (Bio-Rad Laboratories 2005) and included 3 min at 95 °C, 40 cycles of 95 °C for 15 s, 55 °C for 20 s and 72 °C for 20 s, 1 min at 95 °C, 80 cycles at 55 °C for 10 s with the temperature increasing by 0.5 °C after each cycle and then held at 4 °C until plates were removed from the machine. Data analysis was carried out using DART-PCR version 1.0 [28] and qBASE [29]. DART-PCR version 1.0 was used to calculate primer efficiency. This information was then used in qBASE, along with cycle threshold values, to calculate fold change in expression of each gene as compared to its expression in the tissue where transcript levels were the lowest.

Intron analysis

Information about the number of introns per gene was obtained from *Arabidopsis* genome annotation release 7 (http://www.arabidopsis.org/), *Oryza* genome annotation release 5 (http://rice.plantbiology.msu.edu/) and *Populus* genome annotation release 1.1 (http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html).

GC content analysis

The GC content of coding sequences were performed using EMBOSS [30]. A two-tailed *T*-test was used to compare the mean GC content of coding sequence between the intronless and intron-containing genes.

Acknowledgments

We thank Stan Wullschleger and Udaya Kalluri for reviewing the manuscript and valuable comments, and J.C. Tuskan for input on the design of the study. Funding for this research was provided by the U.S. Department of Energy, Office of Science, Biological and Environmental Research Carbon Sequestration Program. ORNL is managed by UT-Battelle, LLC for the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2009.01.002.

References

[1] R. Mazumder, D.A. Natale, S. Murthy, R. Thiagarajan, C.H. Wu, Computational identification of strain-, species- and genus-specific proteins, BMC Bioinformatics 6 (2005) 279.

[2] H. Ogata, J.M. Claverie, Unique genes in giant viruses: regular substitution pattern and anomalously short size, Genome Res. 17 (2007) 1353–1361.

[3] M.A. Campbell, W. Zhu, N. Jiang, H. Lin, S. Ouyang, K.L. Childs, B.J. Haas, J.P. Hamilton, C.R. Buell, Identification and characterization of lineage-specific genes within the Poaceae, Plant Physiol. 145 (2007) 1311–1322.

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

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

[6] International Rice Genome Sequencing Project, The map-based sequence of the rice genome, Nature 436 (2005) 793–800.

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

[8] G.A. Tuskan, S. Difazio, S. Jansson, J. Bohlmann, I. Grigoriev, U. Hellsten, N. Putnam, S. Ralph, S. Rombauts, A. Salamov, et al., The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray), Science 313 (2006) 1596–1604.

[9] K.L. Childs, J.P. Hamilton, W. Zhu, E. Ly, F. Cheung, H. Wu, P.D. Rabinowicz, C.D. Town, C.R. Buell, A.P. Chan, The TIGR plant transcript assemblies database, Nucleic Acids Res. 35 (2007) D846–D851.

[10] R. Ming, S. Hou, Y. Feng, Q. Yu, A. Dionne-Laporte, J.H. Saw, P. Senin, W. Wang, B.V. Ly, K.L. Lewis, et al., The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus), Nature 452 (2008) 991–996.

[11] O. Jaillon, J.M. Aury, B. Noel, A. Policriti, C. Clepet, A. Casagrande, N. Choisne, S. Aubourg, N. Vitulo, C. Jubin, et al., The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla, Nature 449 (2007) 463–467.

[12] A. Kelil, S. Wang, R. Brzezinski, A. Fleury, CLUSS: clustering of protein sequences based on a new similarity measure, BMC. Bioinformatics. 8 (2007) 286.

[13] APG II, An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II, Bot. J. Linn. Soc. 141 (2003) 399–436.

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

[15] L. Carmel, Y.I. Wolf, I.B. Rogozin, E.V. Koonin, Three distinct modes of intron dynamics in the evolution of eukaryotes, Genome Res. 17 (2007) 1034–1044.

[16] S.W. Roy, D. Penny, Patterns of intron loss and gain in plants: intron loss-dominated evolution and genome-wide comparison of *O. sativa* and *A. thaliana*, Mol. Biol. Evol. 24 (2007) 171–181.

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

[18] N. Carels, G. Bernardi, Two classes of genes in plants, Genetics 154 (2000) 1819–1825.

[19] N.N. Alexandrov, V.V. Brover, S. Freidin, M.E. Troukhan, T.V. Tatarinova, H. Zhang, T.J. Swaller, Y.P. Lu, J. Bouck, R.B. Flavell, et al., Insights into corn genes derived from large-scale cDNA sequencing, Plant Mol. Biol. 69 (2009) 179–194.

[20] S. Ouyang, W. Zhu, J. Hamilton, H. Lin, M. Campbell, K. Childs, F. Thibaud-Nissen, R.L. Malek, Y. Lee, L. Zheng, et al., The TIGR rice genome annotation resource: improvements and new features, Nucleic Acids Res. 35 (2007) D883–D887.

[21] B.J. Haas, A.L. Delcher, S.M. Mount, J.R. Wortman, R.K. Smith Jr., L.I. Hannick, R. Maiti, C.M. Ronning, D.B. Rusch, C.D. Town, et al., Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies, Nucleic Acids Res. 31 (2003) 5654–5666.

[22] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.

[23] J. Kilian, D. Whitehead, J. Horak, D. Wanke, S. Weinl, O. Batistic, C. D'Angelo, E. Bornberg-Bauer, J. Kudla, K. Harter, 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 (2007) 347–363.

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

[25] T.L. Bailey, C. Elkan, Fitting a mixture model by expectation maximization to discover motifs in biopolymers, Proc. Int. Conf. Intell. Syst. Mol. Biol. 2 (1994) 28–36.

[26] T.L. Bailey, M. Gribskov, Combining evidence using p-values: application to sequence homology searches, Bioinformatics 14 (1998) 48–54.

[27] T. Murashige, F. Skoog, A revised medium for rapid growth and bioassay with tobacco tissue cultures, Physiol. Plant. 15 (1962) 473–497.

[28] S.N. Peirson, J.N. Butler, R.G. Foster, Experimental validation of novel and conventional approaches to quantitative real-time PCR data analysis, Nucleic Acids Res. 31 (2003) e73.

[29] J. Hellemans, G. Mortier, A. De Paepe, F. Speleman, J. Vandesompele, qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data, Genome Biol. 8 (2007) R19.

[30] P. Rice, I. Longden, A. Bleasby, EMBOSS: the European Molecular Biology Open Software Suite, Trends Genet. 16 (2000) 276–277.