

## Large-scale computational analysis of poplar ESTs reveals the repertoire and unique features of expressed genes in the poplar genome

Sunchung Park, Sookyung Oh and Kyung-Hwan Han\*

*Department of Forestry Michigan State University 126 Natural Resources East Lansing MI 48824-1222 USA;*

*\*Author for correspondence (phone: (517) 353-4751 (office), (517) 432-6180 (lab); fax: (517) 432-1143;*

*e-mail: hanky@msu.edu)*

Received 11 December 2003; accepted in revised form 9 June 2004

*Key words:* Expressed sequence tags (ESTs), Comparative genomics, Poplar transcriptome

### Abstract

Perennial woody plants differ from annual herbaceous plants in several ways and are expected to have evolved to adopt a unique repertoire and expression profiles of functional genes. Poplar, a model tree species for which a large number of ESTs are publicly available, was used to carry out a large-scale comparative analysis with the expressed sequences of eight plant species. First, we obtained 105,831 poplar ESTs from public databases and identified a set of 25,282 unigenes (i.e., tentative non-redundant sequences). The majority of the unigenes (56%) had significant matches to *Arabidopsis* genes. We then estimated poplar multigene families by counting the tBLASTX matches of each unigene against the poplar unigene dataset itself. Forty-seven percent of the 25,282 unigenes were subsequently organized into 3,481 multigene families 89% of which had less than five copy members. In poplar, protein kinases represent the largest family followed by GTP-binding proteins and Myb transcription factors. Several multigene families had a higher copy number in poplar than in *Arabidopsis* hinting potential lineage-specific proliferation of poplar protein families. Such expansion may be related to the adaptation of perennial poplars for the high degree of environmental stresses that affects growth and survival. Comparison of poplar unigenes with the *Arabidopsis* transcriptome revealed that genes involved in transcriptional regulation are the most divergent while metabolism-related genes are the most conserved.

### Introduction

Tree growth is one of the most important biological processes on Earth. Its product wood is of primary importance to humans as timber for construction, fuelwood and wood-pulp for paper manufacturing. It is also the most environmentally cost-effective renewable source of energy. In order to survive various environmental changes over long periods of time, trees have evolved to adopt unique physiological properties. For example, bud dormancy enables the tree to overcome unfavorable growth conditions and avoid deep frost. We are interested in learning about the genetic regulation of the tree-specific developmental processes using *Populus* as a model.

The genus *Populus* belongs to the Salicaceae family and is comprised of about 30 species. Poplar has been widely accepted as a model system for tree biology by many researchers for several reasons: a) a small genome; b) easy clonal propagation which allows for replication of experiments; c) rapid growth; d) extensive genetic variation; e) high expected synteny with *Arabidopsis*; f) available high-throughput transgenic technology; and g) high quality genetic maps. Furthermore its complete genome sequencing project now has approximately 8× coverage of the genome (Tuskan et al. 2003). Therefore, it is no surprise that a large number of expressed sequence tags (ESTs) have been produced and analyzed from vari-

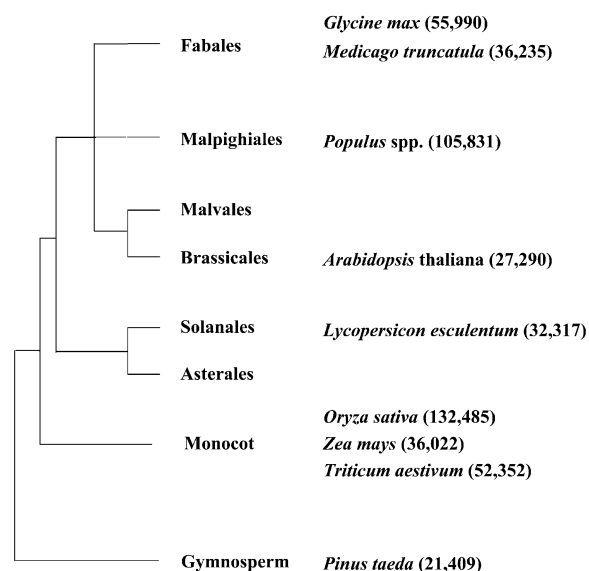


Figure 1. Phylogenetic relationships for the nine plant species used in the sequence analyses. The number of sequences used in the analyses for each species is given in parenthesis. The phylogenetic tree was adapted from Soltis et al. (1999) and abbreviated only to reflect the relative branching order of each species involved.

ous tissues of the *Populus* species (Sterky et al. 1998; Hertzberg et al. 2001; Kohler et al. 2003).

Single-pass sequencing of cDNA clones can reveal a substantial portion of the expressed genes of a genome. As a result large numbers of ESTs are being generated from a wide variety of plant species. Comparative analyses of the expressed sequence tags (ESTs) provide a powerful means for gene discovery as well as the study of molecular basis of plant diversity. First we retrieved a total of 105,831 poplar ESTs from NCBI (<http://www.ncbi.nih.gov>) and organized them into clusters to generate unigene sets using the StackPACK program. This program has been applied successfully to identify different genes from redundant ESTs of other species such as barley (Michalek et al. 2002) and fungus (Trail et al. 2003). With the identified unigene sets we conducted large-scale computational comparisons against ca. 394,100 expressed sequences of eight plant species that included four dicots, three monocots and one gymnosperm. These species have diverse phylogenetic relationships (Soltis et al. 1999) (Figure 1). Of these eight species, two complete plant genome sequences (*Arabidopsis* and rice) were included in the analyses. These analyses provided us with a unique opportunity to examine the repertoire and organization of expressed genes in

the poplar genome and to observe the differences in expression profiles between poplar and other plant species. The unigene set generated in this study is accessible on a publicly searchable website (<http://www.genomics.msu.edu/~han/poplar.html>).

## Results and discussion

### *Analysis of poplar ESTs and establishment of a unigene set*

We downloaded a total of 105,831 ESTs (as of January 2003) from public databases that were derived from various tissues of four poplar species (Table 1). A hybrid aspen species (*Populus tremula* × *P. tremuloides*) contributed the highest number of ESTs (56,147) followed by *P. balsamifera* ssp. *trichocarpa* (24,050) *P. tremula* (14,091) and *P. tremuloides* (11,543). The ESTs were organized into 25,282 unigenes, as described in the Methods section, each of which may represent a potentially distinct gene (i.e., unigene). While redundant transcripts are assembled into contigs, the low-frequency ESTs are not incorporated into contig assemblies and remain as singletons. The resulting 25,282-unigene set consists of 15,178 singletons and 10,104 contigs with an average length of 476-bp and 692-bp, respectively (Table 1). The 15,178 singletons consist of 9,005 from aspen, 2,892 from *P. trichocarpa*, 1,966 from *P. tremula* and 1,317 from *tremuloides*. The 25,282-unigenes account for about 72% of an estimated 35,000 poplar genes (Tuskan et al. 2003). A list of the entire poplar unigenes is provided on our project web site at <http://www.genomics.msu.edu/~han/poplar.html>. However, we should acknowledge that the number of unigenes might have been overestimated mostly due to the low quality and short-length attributes of ESTs that has been reported by other studies (Zhu et al. 2003; Rudd 2003). The attributes may be responsible for the singletons that failed to merge into contigs. Indeed, the singletons of the poplar unigene set had shorter sequence lengths than did the contigs (Figure 1). Fifty-five percent of singletons ranged less than 500-bp while 20% of contigs ranged less than 500-bp. The potential drawbacks (i.e., short length and singleton) were considered during interpretation of further analyses. For instance, the singletons could have been generated by orthologs less conserved among the four species. To assess the likelihood of orthologs appearing as unigenes, we compared the

Table 1. Poplar ESTs and unigene Set Statistics.

Species	Number of ESTs	Number of unigenes	Average Length (bp)
<i>P. tremula</i> × <i>P. tremuloides</i>	56147	20864	511
<i>P. trichocarpa</i>	24050	9113	542
<i>P. tremula</i>	14091	7295	461
<i>P. tremuloides</i>	11543	4683	507
Total	105831	41955	ND <sup>d</sup>
Poplar unigenes identified by clustering the combined 41955 sequences	41955 <sup>a</sup>	29818	ND
Filtering with BLASTCLUST		29644 <sup>b</sup>	ND
Final Poplar Unigene Set		25282 <sup>c</sup>	572

<sup>a</sup>The sequences consist of four species unigene sets; <sup>b</sup>BLASTCLUST algorithm was used to further remove any sequences with 95% nucleic acid identity covering 90% length region; <sup>c</sup>Any ESTs shorter than 250-bp sequence read were discarded in order to reduce sequence length bias in subsequent analyses. The poplar unigene set consists of 15,178 singletons and 10,104 contigs; <sup>d</sup>ND not determined.

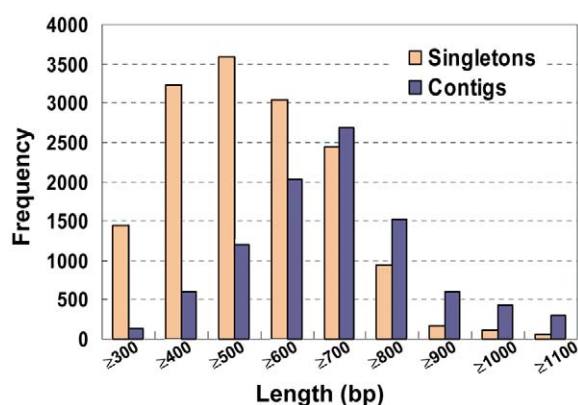


Figure 2. Distribution of sequence length (bp) of singletons and contigs that constitute the poplar unigene set.

singletons of three poplar species with the entire unigenes from aspen using tBLASTX algorithm. Any singletons that had higher similarity with aspen unigenes than with genes of any other eight non-poplar species used in this study were defined as potential orthologs although they could be closely related paralogs. Using the bits score as similarity criteria and less than  $1E-20$  as a threshold for significance of the similarity, we identified 571 singletons (275 from *P. trichocarpa*, 195 from *P. tremula* and 97 from *P. tremuloides*) as the potential orthologs. This number is relatively small compared to the 25,282 unigenes, suggesting that the less conserved orthologs are unlikely to affect the overall conclusions at the analysis stringency we used.

A surrogate annotation approach was used to annotate the poplar unigene set. The entire poplar unigene set was searched using the tBLASTX algorithm against the *Arabidopsis* transcriptome predicted from the complete genomic sequence (TIGR Genome Da-

tabases <http://www.tigr.org/tdb/>). Functional categories have been assigned using the Munich Information Center for Protein Sequences (MIPS) *Arabidopsis* database (MATDB [http://mips.gsf.de/proj/thal/db/search/search\\_frame.html](http://mips.gsf.de/proj/thal/db/search/search_frame.html)) search function. Poplar unigenes with an E-value of  $\leq 1.0E-20$  were assigned to the corresponding *Arabidopsis* annotation. This approach is based on the assumption that functionality is transferable based on sequence conservation. The majority of the unigenes (56%) had significant matches to *Arabidopsis* genes 75% of which matched to genes of known function but the remaining 25% were found to be unclassified or of unknown function. Of the assigned functions metabolism-related genes were the most numerous at 18% followed by cellular organization (9%) and signal transduction (8%) (Figure 3). Other significant functional categories include transcription regulation (7%), protein destination (7%), protein synthesis (4%), transport facilitation (5%) and cell division (3%). Kirst et al. (2003) recently reported that ~90% of pine ESTs had counterparts in *Arabidopsis* sequences (E value  $< 1.0E-10$ ). The 56% match in this study is substantially lower than their estimate. The difference might be explained by the facts that many of the poplar unigenes used in our analysis were not full-length (average length of 572-bp) and differing cut-off values ( $1.0E-10$  in pine study and  $1.0E-20$  in the current study) were used. Indeed, the unmatched unigenes had shorter sequence lengths (with an average of 490-bp) than did the matched ones (with an average of 645-bp). In addition, this functional classification analysis was performed solely based on ESTs, whose representation is highly dependent on the source libraries and expression level. Genes with low levels of expression might not

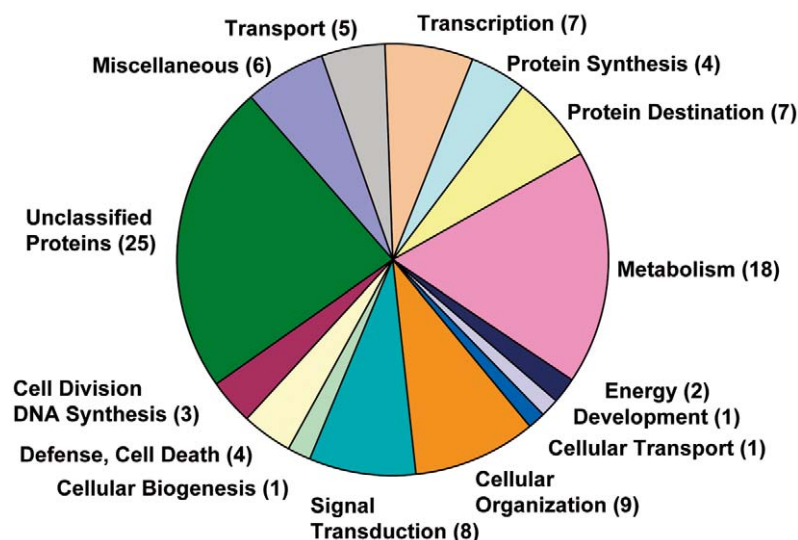


Figure 3. Functional categorization of the 14,291 poplar unigenes that had tBLASTX matches to *Arabidopsis* transcriptome with E-value threshold of  $\leq 1.0E-20$ . The 10,991 unigenes that did not have matches were excluded from the chart. Functional categories are according to the Munich Information Center For Protein Sequence (MIPS <http://mips.gsf.de>). Percentage of the unigenes in each functional category is given in parenthesis.

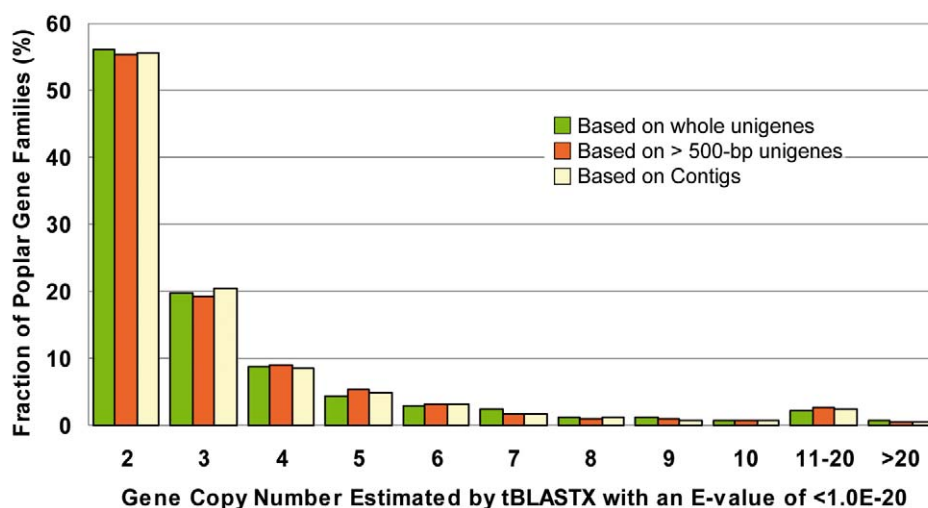


Figure 4. Gene copy number distribution of putative poplar gene families estimated by tBLASTX search against the poplar unigene data set itself. A total of 3,481, 2,303 or 1,995 multigene families were identified based on different criteria, and about 56% of which had two copy members.

have been represented in this analysis. This notion is supported by the fact that a large number of *Arabidopsis* ESTs (178,000 ESTs) from various sources representing over 60 distinct tissues, biotic and abiotic challenges, and developmental stages matched to only 16,115 unigenes (reviewed in Rudd 2003). Likewise, in the current classification analysis the presence of an EST is a reliable attribute while the

absence of an EST does not necessarily mean that the gene is not present in the genome and not expressed. Completion of the poplar genome sequencing will provide a better opportunity to examine the differences between poplar and other plant species.

*Poplar multigene families estimated by tBLASTX analysis of unigenes*

One characteristic feature of eukaryotic genomes is that a significant fraction of protein-coding genes belong to multigene families that are likely derived from gene duplication. For example, up to 80% of the *Arabidopsis* genes are members of multigene families (Lespinet et al. 2002). The main contributing forces for organizational and regulatory diversity in eukaryotes include the changes in domain architectures *via* domain accretion and shuffling gene loss in a particular lineage and lineage-specific proliferation of protein families (Aravind et al. 2000; Lander et al. 2001; Lespinet et al. 2002). Recent advances in structural and comparative genomics have provided valuable information about gene family organization in diverse plant species.

In order to describe the gene family organization in poplar we first computationally determined the number and size of poplar gene families by counting the tBLASTX matches of each unigene against the poplar unigene set itself with an E-value threshold of  $\leq 1.0E-20$ . It is important to note that these gene families do not necessarily indicate functional groupings. Of the 25,282 poplar unigenes, 11,909 (47%) had significant matches with other poplar unigenes at this threshold and were subsequently organized into 3,481 gene families. To determine the potential bias caused by short sequence length or singleton presence, we performed the same analysis using unigenes greater than 500-bp or excluding singletons (i.e., contigs). Of 15,088 poplar unigenes greater than 500-bp, 9,448 (63%) were grouped into 2,033 gene families. There were 10,104 poplar contigs (25,282 minus singletons), 5,401 (53%) of which had significant matches with other poplar unigenes, resulting in 1,995 gene families. The distribution of the gene family sizes is shown in Figure 4. Overall, the majority (56-55%) had two copy members. Gene families with fewer than five copy members accounted for around 90% of the entire gene families estimated in this study. In order to assess how the copy number of the gene families was conserved between the two species, we have compared the copy numbers of the 14 largest poplar gene families with those of corresponding *Arabidopsis* gene families. The *Arabidopsis* gene copy number was obtained by subjecting the *Arabidopsis* gene set (TIGR <http://www.tigr.org/tdb/>) to the same analysis as the poplar tBLASTX. In order to increase the stringency and therefore prevent

an overestimation of the poplar gene copy number in the analysis, the initial tBLASTX-determined copy numbers of the 14 largest poplar gene families were subjected to several refinement steps. First, we removed the sequences that had no matches to the known multigene family proteins in public databases. Then, the sequences that had less than 90% identity with any member of the multigene family with over 90% length region were removed. Finally, we removed the sequences that either did not share the conserved region of the protein family due to its short length or had long gap regions probably resulting from chimera when they were aligned using ClustalW software (Thompson et al. 1994). The resulting alignments are provided as supplemental materials ([http://forestry.msu.edu/biotech/Projects/Projects\\_Poplar-.htm](http://forestry.msu.edu/biotech/Projects/Projects_Poplar-.htm)). Table 2 lists the 14 highest copy number poplar gene families along with their corresponding *Arabidopsis* gene families. In poplar, protein kinases represent the largest family followed by Ubiquitin-domain proteins, Ras-related GTP-binding proteins, and Myb transcription factors. Several multigene families had lower copy numbers in poplar than in *Arabidopsis* (Table 2). However caution should be exercised in interpreting these results because two factors exist that may lead to a substantial underestimate of copy numbers for each gene family. First, many of the 25,282 poplar unigene sequences used in the analysis are partial sequences that generally result in higher E-values (i.e., lower similarity) compared to full-length sequences. It is quite possible that the tBLASTX search with a short coding region of partial EST sequences might miss the consensus motif(s) characteristics of the gene family if they are located in the middle of the protein. Second, the 25,282 poplar unigenes do not represent the entire poplar transcriptome while *Arabidopsis* sequences represent whole transcriptome on its genome. Possibly, the genes with low expression levels were not picked up during the EST sequencing resulting in a lower copy number estimation than the actual number. This may account for the lower copy number of some multigene families in poplar when compared to *Arabidopsis*.

Five of the 14 largest poplar multigene families had a relatively higher copy number in poplar than in *Arabidopsis* (Table 2). These families include ubiquitin-domain protein, ubiquitin-conjugating enzyme, tubulin, peptidylprolyl isomerase and ABA-inducible protein. The proteins of these families had predicted biochemical characteristics that suggest their roles in protein degradation energy stress response and cellu-

Table 2. The 14 highest copy number poplar unigene families determined by tBLASTX with an E-Value threshold of  $< 1.0E-20$ .

Poplar Contig (Gene) Family	GenBank Accession Number	Copy No.	Putative Function of the Gene Family	Representative Gene of Corresponding <i>Arabidopsis</i> Family	<i>Arabidopsis</i> Gene Copy Number
cn726Aspn	BU896558	53 <sup>a</sup> (73 <sup>b</sup> )	Protein kinase	At1g56140	501
cn8047Aspn	BU894782	33 (53)	Ubiquitin-domain protein	At4g02890	19
cn16304Aspn	BU810510	41 (48)	GTP-binding protein	At3g46830	60
24067636Tcc	BU876112	31 (44)	Myb transcription factor	At4g18770	133
cn15003Aspn	BU894972	33 (37)	Ubiquitin-conjugating enzyme	At3g08690	30
cn933Aspn	BU895108	25 (34)	Peroxidase	At2g18150	73
cn6614Aspn	BU831012	24 (33)	Cytochrome P450	At4g37400	152
cn15387Aspn	BU831392	24 (31)	ADP-ribosylation factor	At2g47170	25
cn11888Aspn	BI128682	19 (30)	Tubulin	At1g50010	17
cn14980Aspn	BU831411	20 (29)	Peptidylprolyl isomerase <sup>c</sup>	At4g34870	20
24074413Tcc	BU882889	21 (28)	MADS-box transcription factor	At4g18960	41 <sup>d</sup>
cn753Tcc	BU882276	18 (28)	Water channel protein	At3g16240	24
cn51Tcc	BU874797	7 (27)	ABA-inducible protein	At5g38760	3
23534478Tmo	AY095297	18 (26)	Cellulose synthase	At4g39350	26

<sup>a</sup>The copy number estimation was further refined by excluding the genes that 1) had no matches to the known multigene family proteins present in NCBI protein database, 2) were shorter than 500-bp, 3) shared  $< 90\%$  nucleic acid sequence identity with any member of the multigene family over 90% length region, 4) did not share the conserved region mostly due to its short length, or 5) had long gap regions probably resulting from chimera when they were aligned using ClustalW software (Thompson et al. 1994); <sup>b</sup>The copy number determined using tBLASTX with an E-value threshold of  $\leq 1.0E-20$ ; <sup>c</sup>To support the gene duplication within the families, we performed phylogenetic analysis using peptidylprolyl isomerase protein family as an example, which was provided as supplemental figure; <sup>d</sup>Actual number of *Arabidopsis* MADS-box genes that can be identified by MADS-box domain search is  $\sim 80$  (Riechmann et al. 2000). The number 41 reflects the outcome of tBLASTX search using the entire cDNA sequence.

lar structure. The higher copy number of ubiquitin gene families in poplar may represent an evolutionary adaptation important for high protein turnover throughout different seasons. Interestingly, functionally related two multigene families such as ubiquitin-conjugating enzyme and peptidylprolyl isomerase which are involved in protein degradation and folding respectively (Hershko and Ciechanover 1998; Schiene and Fischer 2000) have a slightly higher copy number in poplar than in *Arabidopsis* (Table 2) and indeed these genes were more abundantly expressed in senescent autumn leaves than in the young summer leaves of poplar (Bhalerao et al. 2003). Another example of potential adaptational changes in poplar is the tubulin gene family that has 19 members in poplar and 17 in the *Arabidopsis* genome. The perennial nature of poplar growth and development requires adaptive evolution for cold hardiness. The microtubules of higher plants have been suggested to participate in low temperature stress response and adaptation (Nick 1998; Nyporko et al. 2003). In fact, those factors known to increase plant cold hardiness were also involved in the induction of elevated cold-stability in microtubules (Pihakaskimaunsbach and Puhakainen 1995). Also it has been shown that depo-

lymerization of microtubules is related to freezing injury in cotton (*Gossypium hirsutum* L.) (Rikin et al. 1980). The cold-resistance of microtubules appears to be correlated with their cold stability (Detrich 1997; Gupta et al. 2001). Parker and Detrich (1998) reported that an Antarctic fish (*Notothenia coriiceps*) had a substantial number ( $\sim 15$ ) of  $\alpha$ -tubulin subunits encoded by a multigene family suggesting that the expansion of tubulin gene families might have occurred as a result of adaptive evolution for low temperatures. Considering these observations in the literature, it is not surprising that poplar had an increase in copy number for this gene family. Another multigene family with a higher copy number in poplar encodes ABA-inducible protein, which is a subfamily of late embryogenesis abundant (LEA) protein family (Wise 2003). These hydrophilic proteins have been suggested to be involved in the protection of cellular structures from dehydration or hyper-osmotic stress by acting as a hydration buffer, by sequestration of ions, by direct protection of other proteins or membranes or through the re-naturation of unfolded proteins (Bray et al. 1993; Soulages et al. 2002; Koag et al. 2003).

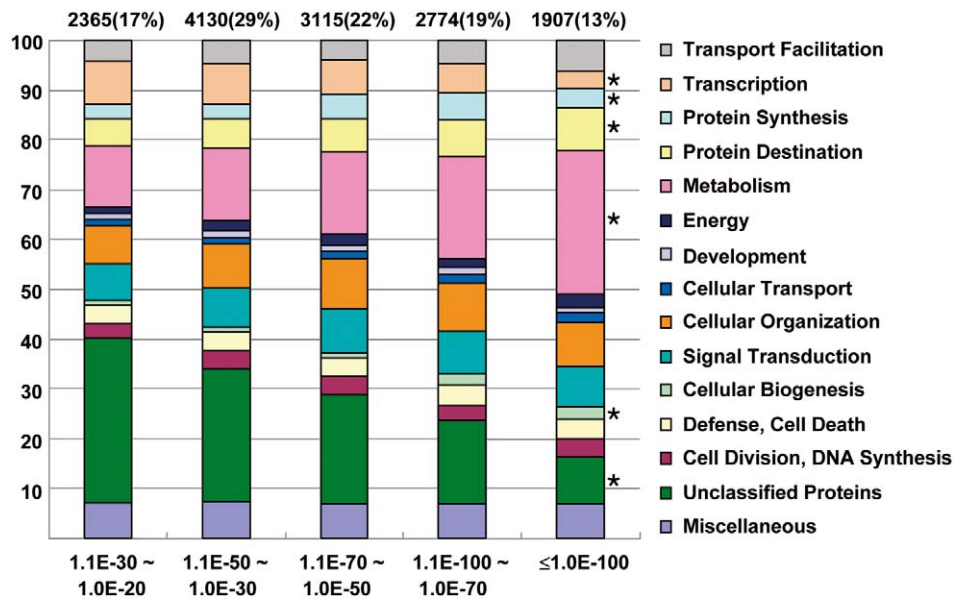


Figure 5. Percentage of poplar unigenes belonging to each functional category with different E-value thresholds. A total of 14,291 sequences had significant tBLASTX matches to the genes of known function according to MIPS (<http://mips.gsf.de>) functional categorization. The actual number and percentage of sequences matched with ranges of 1.0E-20~1.1E-30, 1.1E-30~1.0E-50, 1.1E-50~1.0E-70, 1.1E-70~1.0E-100 and < 1.0E-100 respectively are indicated on top of the graph. Asterisks indicate that the differed distribution of the categories is supported by contingency test ( $p < 1E-5$ ).

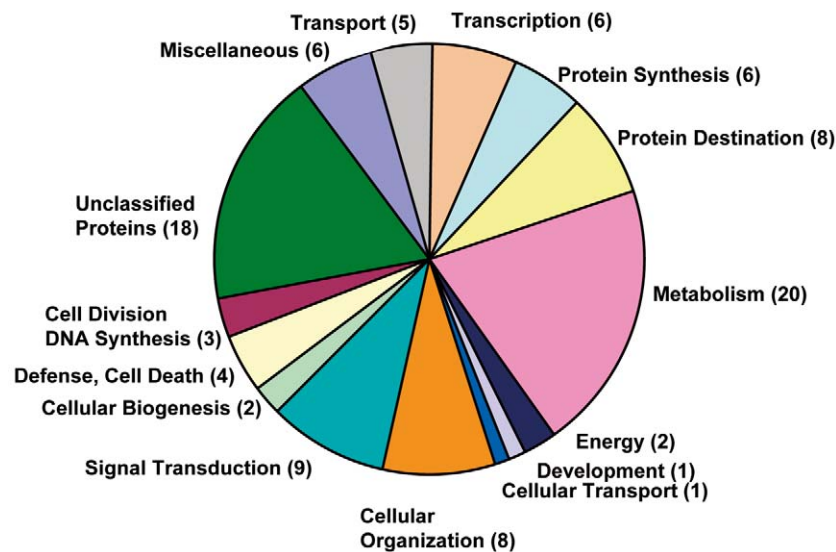


Figure 6. Functional categorization of the 12,064 poplar unigenes that had counterparts in three major vascular plant groups (dicots, monocots, and pine). Percentage of the unigenes in each functional category is given in parenthesis. For this analysis the plant gene sequences were pooled into three groups. All of the sequences from *Glycine max*, *Medicago truncatula*, *Arabidopsis thaliana* and *Lycopersicon esculentum* were pooled into the 'dicots' group which had a total of 151,832 sequences. Likewise the sequences from *Oryza sativa*, *Zea mays*, and *Triticum aestivum* into the 'monocots' group with a total of 220,859 sequences.

Multigene families with higher copy numbers could imply greater functional diversity providing resources for specific adaptations and evolution of new functional systems. Considering that the poplar genome is approximately four times larger than the *Arabidopsis* genome and only about 72% of the transcriptome is represented in this analysis it is reasonable to assume that the copy number of each poplar family should be less than or at most equal to that of corresponding *Arabidopsis* families if there is no poplar specific gene family expansion since divergence from the last common ancestor. Nonetheless we noted several multigene families had relatively higher copy numbers in poplar than in *Arabidopsis*. Potential physiological roles of the protein families (as described above) suggest that this kind of multigene family expansion might be common among the perennial temperate tree species. However it should be noted that the copy numbers of multigene families could also be overestimated considering the high rate of error in EST datasets (Ewing and Green 1998; Hillier et al. 1996), and highly expressed gene families could be over-represented during clustering analysis.

#### *Comparison of poplar with other plant species*

*Arabidopsis* is a herbaceous non-woody plant that does not undergo secondary growth under normal conditions. However it has recently been shown that *Arabidopsis* can be induced to express all of the major components of secondary growth (Lev-Yadun 1994; Zhao et al. 2000; Chaffey et al. 2002). It is therefore of great scientific interest to examine the degree of sequence similarity in expressed gene sequences between *Arabidopsis* a herbaceous plant and poplar a perennial woody species. In order to determine whether certain functional classes of genes represent characteristic differences between the perennial tree species (i.e., poplar) and an annual herbaceous plant (i.e., *Arabidopsis*) we carried out computational comparisons of the 25,282 poplar unigenes in all translated frames (tBLASTX algorithm) with *Arabidopsis* transcriptome predicted from the complete genomic sequence using the E-values of tBLASTX searches as an estimate of sequence conservation. Figure 5 displays the distribution of E-value matches with regard to functional categories for the poplar unigene set. Of the 25,282 poplar unigenes, 58% show strong similarity ( $\leq 1.0E-20$ ) with their *Arabidopsis* counterparts. Even at the strongest match

threshold ( $\leq 1.0E-100$ ) eight percent (1,907) of the 25,282 poplar unigenes matched with *Arabidopsis* sequences showing a very high level of conservation. While the frequencies of genes in most functional categories were similar among the five different match stringency (E-value) categories those of genes involved in a few cellular functions changed significantly as the match stringency increased. For example, the proportion of genes belonging to the metabolism category decreased from 29% at the strongest match stringency ( $\leq 1.0E-100$ ) to 12% when the E-value cutoff was reduced to  $1.0E-20 \sim 1.1E-30$  suggesting that metabolic functions are more constant across plant species (i.e., more ancestral gene functions) (Figure 5). On the other hand transcription regulation-related genes showed opposite trends. Their frequencies decreased from 9 to 3% as the E-value stringency changed from  $1.0E-20$  to  $1.0E-100$ . This along with the fact that *Arabidopsis* can be induced to undergo secondary growth suggests that the major differences between poplar and *Arabidopsis* (e.g., secondary growth perennial growth habit) may be in transcriptional control rather than in structural genes. Genes of unknown function are among the 'less-conserved.' For example 33% of the poplar unigenes having relatively weak similarity ( $\leq 1.0E-20$ ) with their *Arabidopsis* counterparts were categorized as 'unclassified' function while only 9% of those with the strongest similarity ( $\leq 1.0E-100$ ) matched to *Arabidopsis* genes for which no putative functions have been assigned. Genes involved in most other cellular functions had similar frequencies among the five different match stringency categories suggesting that they are relatively fast evolving. The differential distribution of the functional categories based on the match stringencies was confirmed by a contingency test ( $P < 1E-5$ ).

In order to investigate whether the highly conserved or divergent genes between poplar and *Arabidopsis* are also conserved or divergent in other vascular plants we computationally compared the poplar unigenes with ca. 394,100 expressed sequences including putative ORFs from the genome sequences of four dicots (*Arabidopsis*, barrel medic, soybean, tomato) three monocots (maize, rice, wheat) and one gymnosperm (pine) (Figure 1). For this analysis the 394,100 plant gene sequences were divided into three groups ('dicots', 'monocots' and 'pine'). All sequences from *Glycine max*, *Medicago truncatula*, *Arabidopsis thaliana*, and *Lycopersicon esculentum* were pooled into the 'dicots' group which had a total of 151,832 sequences. Likewise the sequences from *Oryza sativa*, *Zea mays*, *Triticum aestivum* went into the 'monocots' group making a total of 220,859 sequences. The 'pine' group repre-

sents only one species (*Pinus taeda*) with 21,409 ESTs available. A large number of poplar unigenes did not have any detectable homologues to other plant species, ranging from 6,835 (27%) (an E-value threshold of  $\leq 1.0E-10$ ) to 9,398 (37%) ( $\leq 1.0E-20$ ). However, about 30% (7,581) of the 25,282 poplar unigenes had significant matches ( $\leq 1.0E-20$ ) in all of the three groups suggesting their involvement with common biological functions in plant species. The proportions in each functional category of the 12,064 unigenes that had counterparts in all of the three groups were very similar to those of the entire poplar unigene set. The largest functional category was 'metabolism' (20%) followed by 'unclassified proteins' (18%) and 'signal transduction' (9%) (Figure 6). A total of 7,760 poplar unigenes matched to pine ESTs with E-value cutoff of  $\leq 1.0E-20$  while 13,422 and 15,257 poplar unigenes had significant matches to monocot and dicot sequences respectively ( $\leq 1.0E-20$ ). The number of poplar unigenes that matched to monocot sequences represents a majority (53%) of the 25,282 poplar unigenes.

## Conclusion

About 72% of the 35,000 estimated poplar genes were identified in the unigene set described in this report. High proportions of the unigenes had counterparts in the genomes of *Arabidopsis* and other plant species supporting the hypothesis that many of the functional genes are conserved among different plant species. We estimate that at least 47% of poplar genes belong to multigene families based on the observation that 11,909 of the 25,282 unigenes had significant tBLASTX matches ( $\leq 1.0E-20$ ) to other poplar unigenes. Comparison of poplar multigene families with those of *Arabidopsis* suggests that the expansion of poplar multigene families occurred in support of adaptation to various stresses affecting perennial growth. Another striking feature of the difference between the poplar and *Arabidopsis* transcriptome is that the genes involved in transcriptional regulation are the most divergent while metabolism-related genes are highly conserved. While the findings described in this report increase our understanding of the poplar transcriptome it should be noted that completion of the on-going poplar genome sequencing would provide an invaluable opportunity to verify these findings and advance our knowledge of tree growth.

## Materials and methods

### *Assembly and clustering of poplar ESTs*

As of January 2003 a total of 105,831 poplar ESTs were available from four *Populus* species. Those ESTs were retrieved using a custom Perl script from dbEST and the non-redundant nucleotide database available at the NCBI (<http://www.ncbi.nih.gov>). The ESTs were separately assembled by individual species using the StackPACK (version 2.2) clustering system (<http://www.e genetics.com>) using default settings for EST assembly (Miller et al. 1999). The process includes sequential steps of masking, clustering assembly, alignment, analysis and consensus partitioning. The masking step employed CrossMatch (<http://www.genome.washington.edu/uwgc/analysis-tools/phrap.htm>) to mask vector artifacts and repetitive sequences as simple repeats. The masked sequences were clustered based on their relative similarity (the default is greater than 96% identical over a window of 150 bases) determined by a d2 cluster (Hide et al. 1994) that is word-based greedy clustering algorithm. The related but loose clusters were further aligned and assembled using PHRAP (<http://www.genome.washington.edu/uwgc/analysis-tools/phrap.htm>) to improve alignment quality by generating particularly distinct sequences as singletons and highly related sequences as sub-contigs. The aligned sub-contigs were further analyzed using the CRAW alignment analysis tool (Chou and Burke 1999). CRAW is used to analyze sub-contigs for error and alternative expression forms, partition the sub-contigs, maximize consensus sequence length, create final alignments and select the best consensus sequence. For each species the ESTs were organized into contiguous overlapping consensus ('contigs') while the low-frequency ESTs were not incorporated into contig assemblies and remain as singletons. The resulting singletons and consensus sequences were used as tentative unigene sets in this study. Then the four unigene sets (a total of 41,955 sequences) were assembled again using the same StackPACK clustering system to produce a 29,818 poplar unigene set. In order to eliminate any transcripts that potentially resulted from the same locus, this poplar unigene set was re-assembled using the BLASTCLUST program (<ftp://ftp.ncbi.nlm.nih.gov/blast/documentations/RE-ADME.bcl>) with a threshold of 95% sequence identity covering a 90% length region. To avoid potential mosaic sequences of different species, the sequence

with the longest sequence read in each contig was selected instead of the resulting consensus sequences, and used as a representative of the contig regardless of the species origin. The resulting contigs and singletons were further screened to eliminate any sequences shorter than 250-bp. The final 25,282 unigenes that may represent distinct genes were used for further analysis.

#### *Other plant sequence data sets used in the analyses*

Computational analyses were performed on several sets of plant gene sequences obtained from public databases. The *Arabidopsis* and rice gene sets as predicted from the entire genomic sequence contain 27,290 and 80,916 sequences respectively (as of March 2003) and are available through TIGR (<http://www.tigr.org/tdb/>). The minimally-redundant EST sets of three dicots, three monocots and one gymnosperm species were obtained from the Gene Indices of TIGR (<http://www.tigr.org/tdb/tgi/plant.shtml>) which were constructed by first clustering and then assembling ESTs and annotated gene sequences from GenBank to produce a set of unique high-fidelity virtual transcripts or tentative consensus (TC) sequences (Quackenbush et al. 2000).

#### *Similarity searches*

Similarity searches were carried out with the Stand-alone BLAST programs (Altschul et al. 1997) using executable copies obtained from the NCBI (v 2.2.2). The sequence data sets used here were formatted first as a BLAST searchable database file using Formatdb executable, one of the BLAST tools available from NCBI (for a manual see <ftp://ftp.ncbi.nih.gov/blast/documents/README.formatdb>). Searches were performed through comparisons of protein sequences with translation of nucleotide query or database sequences using gapped BLASTX or tBLASTX (Altschul et al. 1997). Nucleotide queries were pre-processed with DUST to mask low-complexity regions and protein query sequences (including six-frame translations of ESTs) were filtered with SEG (Wootton and Federhen 1996). Custom Perl scripts (for scripts visit our laboratory web site at [http://forestry.msu.edu/biotech/Projects/Projects\\_Poplar.htm](http://forestry.msu.edu/biotech/Projects/Projects_Poplar.htm)) and relational databases (Microsoft Access) were used to automate searches on large sets of query sequences and to extract summary information (e.g. score and E-value of best hit). Queries were consid-

ered to have a significant similarity for E-value cutoff of  $\leq 1.0E-20$  a potential similarity for E-value  $\leq 1.0E-10$  and no significant similarity for an E-value cutoff of  $> 1.0E-5$ . These user-defined E-values were based on the criteria used in other comparative genomics studies (Rubin et al. 2000; Van der Hoeven et al. 2002).

#### **Web Site References**

Documentation for the Formatdb program:  
<ftp://ftp.ncbi.nih.gov/blast/documents/README.formatdb>;  
 Documentation for the BLASTCLUST program:  
<ftp://ftp.ncbi.nih.gov/blast/documents/README.formatdb>;  
 Electric Genetics Home Page: <http://www.e genetics.com>;  
 MATDB Search Page: [http://mips.gsf.de/proj/thal/db/search/search\\_frame.html](http://mips.gsf.de/proj/thal/db/search/search_frame.html);  
 NCBI Public Databases: <ftp://ftp.ncbi.nih.gov/blast/db>;  
 Searchable web site for the data presented:  
<http://www.genomics.msu.edu/~han/poplar.html>  
 Stand-alone BLAST program: <ftp://ftp.ncbi.nih.gov/blast/executables>;  
 Supplemental data: [http://forestry.msu.edu/biotech/Projects/Projects\\_Poplar.htm](http://forestry.msu.edu/biotech/Projects/Projects_Poplar.htm);  
 TIGR Gene Index Databases: <http://www.tigr.org/tbd/plant.shtml>;  
 TIGR Genome Databases: <http://www.tigr.org/tdb>;

#### **Acknowledgement**

We would like to thank Marilyn Ruthig for editorial assistance and the Genomics Technology Support Facility staff at Michigan State University for their help and consultation with the bioinformatics analysis. This work was supported by USDA Grants to the Eastern Hardwood Utilization Program at Michigan State University (No. 98-34158-5995 00-34158-9236 and 01-34158-11222).

#### **References**

Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W. and Lipman D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389-3402.

- Aravind L., Watanabe H., Lipman D.J. and Koonin E.V. 2000. Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proc. Natl. Acad. Sci. USA* 97: 11319–11324.
- Bhalerao R., Keskitalo J., Sterky F., Erlandsson R., Bjorkbacka H., Birve S.J., Karlsson J., Gardstrom P., Gustafsson P., Lundeberg J. and Jansson S. 2003. Gene expression in autumn leaves. *Plant Physiol.* 131: 430–442.
- Bray E.A. 1993. Molecular Responses to Water Deficit. *Plant Physiol.* 103: 1035–1040.
- Chaffey N., Cholewa E., Regan S. and Sundberg B. 2002. Secondary xylem development in *Arabidopsis*: a model for wood formation. *Physiol. Plant* 114: 594–600.
- Chou A. and Burke J. 1999. CRAWview: for viewing splicing variation gene families and polymorphism in clusters of ESTs and full-length sequences. *Bioinformatics* 15: 376–381.
- Detrich H.W. 3rd. 1997. Microtubule assembly in cold-adapted organisms: functional properties and structural adaptations of tubulins from antarctic fishes. *Comp. Biochem. Physiol. A Physiol.* 118: 501–513.
- Ewing B. and Green P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* 8: 186–194.
- Gupta M.L. Jr., Bode C.J., Dougherty C.A., Marquez R.T. and Himes R.H. 2001. Mutagenesis of beta-tubulin cysteine residues in *Saccharomyces cerevisiae*: mutation of cysteine 354 results in cold-stable microtubules. *Cell Motil. Cytoskeleton* 49: 67–77.
- Hershko A. and Ciechanover A. 1998. The ubiquitin system. *Annu. Rev. Biochem.* 67: 425–479.
- Hertzberg M., Aspeborg H., Schrader J., Andersson A., Erlandsson R., Blomqvist K., Bhalerao R., Uhlen M., Teeri T.T., Lundeberg J., Sundberg B., Nilsson P. and Sandberg G. 2001. A transcriptional roadmap to wood formation. *Proc. Natl. Acad. Sci. USA* 98: 14732–14737.
- Hide W., Burke J. and Davison D.B. 1994. Biological evaluation of d2 an algorithm for high-performance sequence comparison. *J. Comput. Biol.* 1: 199–215.
- Hillier L.D., Lennon G., Becker M., Bonaldo M.F., Chiapelli B., Chisoe S., Dietrich N., DuBuque T., Favello A., Gish W., Hawkins M., Hultman M., Kucaba T., Lacy M., Le M., Le N., Mardis E., Moore B., Morris M., Parsons J., Prange C., Rifkin L., Rohlfing T., Schellenberg K., Marra M. and et al. 1996. Generation and analysis of 280000 human expressed sequence tags. *Genome Res.* 6: 807–828.
- Kirst M., Johnson A.F., Baucom C., Ulrich E., Hubbard K., Staggs R., Paule C., Retzel E., Whetten R. and Sederoff R. 2003. Apparent homology of expressed genes from wood-forming tissues of loblolly pine (*Pinus taeda* L.) with *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA* 100: 7383–7388.
- Koag M.C., Fenton R.D., Wilkens S. and Close T.J. 2003. The binding of maize DHN1 to lipid vesicles. Gain of structure and lipid specificity. *Plant Physiol.* 131: 309–316.
- Kohler A., Delaruelle C., Martin D., Encelot N. and Martin F. 2003. The poplar root transcriptome: analysis of 7000 expressed sequence tags. *FEBS Lett.* 542: 37–41.
- Lander E.S., Linton L.M., Birren B., Nusbaum C., Zody M.C., Baldwin J., Devon K., Dewar K., Doyle M., FitzHugh W. and et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409: 860–921.
- Lespinet O., Wolf Y.I., Koonin E.V. and Aravind L. 2002. The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res.* 12: 1048–1059.
- Lev-Yadun S. 1994. Induction of sclereid differentiation in the pith of *Arabidopsis thaliana* (L) Heynh. *J. Exp. Bot.* 45: 1845–1849.
- Michalek W., Weschke W., Pleissner K.P. and Graner A. 2002. EST analysis in barley defines a unigene set comprising 4000 genes. *Theor. Appl. Genet.* 104: 97–103.
- Miller R.T., Christoffels A.G., Gopalakrishnan C., Burke J., Ptitsyn A.A., Broveak T.R. and Hide W.A. 1999. A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. *Genome Res.* 9: 1143–1155.
- Nick P. 1998. Signaling to the microtubular cytoskeleton in plants. In *International Review of Cytology – a Survey of Cell Biology* 184: 33–80.
- Nyorko A.Y., Demchuk O.N. and Blume Y.B. 2003. Cold adaptation of plant microtubules: structural interpretation of primary sequence changes in a highly conserved region of alpha-tubulin. *Cell Biol. Intl.* 27: 241–243.
- Parker S.K. and Detrich H.W. 1998. Evolution organization and expression of alpha-tubulin genes in the antarctic fish *Notothenia coriiceps*. Adaptive expansion of a gene family by recent gene duplication inversion and divergence. *J. Biol. Chem.* 273: 34358–34369.
- Pihakaskimaunsbach K. and Puhakainen T. 1995. Effect of cold-exposure on cortical microtubules of rye (*Secale cereale*) as observed by immunocytochemistry. *Physiologia Plantarum* 93: 563–571.
- Quackenbush J., Liang F., Holt I., Pertea G. and Upton J. 2000. The TIGR gene indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Res.* 28: 141–145.
- Rikin A., Atsmon D. and Gitler C. 1980. Chilling injury in cotton (*Gossypium hirsutum* L.) – Effects of anti-microtubular drugs. *Plant and Cell Physiology* 21: 829–837.
- Rubin G.M., Yandell M.D., Wortman J.R., Gabor Miklos G.L., Nelson C.R., Hariharan I.K., Fortini M.E., Li P.W., Apweiler R., Fleischmann W. and et al. 2000. Comparative genomics of the Eukaryotes. *Science* 287: 2204–2215.
- Rudd S. 2003. Expressed sequence tags: alternative or complement to whole genome sequences? *Trends in Plant Science* 8: 321–329.
- Schiene C. and Fischer G. 2000. Enzymes that catalyse the restructuring of proteins. *Curr. Opin. Struct. Biol.* 10: 40–45.
- Soltis P.S., Soltis D.E. and Chase M.W. 1999. Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. *Nature* 402: 402–404.
- Soulages J.L., Kim K., Walters C. and Cushman J.C. 2002. Temperature-induced extended helix/random coil transitions in a group 1 late embryogenesis-abundant protein from soybean. *Plant Physiol.* 128: 822–832.
- Sterky F., Regan S., Karlsson J., Hertzberg M., Rohde A., Holmberg A., Amini B., Bhalerao R., Larsson M., Villarroel R., Van Montagu M., Sandberg G., Olsson O., Teeri T.T., Boerjan W., Gustafsson P., Uhlen M., Sundberg B. and Lundeberg J. 1998. Gene discovery in the wood-forming tissues of poplar: analysis of 5 692 expressed sequence tags. *Proc. Natl. Acad. Sci. USA* 95: 13330–13335.
- Thompson J.D., Higgins D.G. and Gibson T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting position-specific gap

- penalties and weight matrix choice. *Nucleic Acids Res.* 22: 4673–4680.
- Trail F., Xu J.R., San Miguel P., Halgren R.G. and Kistler H.C. 2003. Analysis of expressed sequence tags from *Gibberella zeae* (anamorph *Fusarium graminearum*). *Fungal Genet. Biol.* 38: 187–197.
- Tuskan G.A., Wulschleger S.D., Difazio S.P., Gunter L.E., Schuster M.E., Land M.L., Larimer F.W., Ritland K., Boore J.L. and Rokhsar D.S. 2003. The *Populus* chloroplast genome: a comparison of genome structure and organization. *Plant and Animal Genome XI*. January 11 – 15 2003 San Diego California, USA.
- Van der Hoeven R., Ronning C., Giovannoni J., Martin G. and Tanksley S. 2002. Deductions about the number organization and evolution of genes in the tomato genome based on analysis of a large expressed sequence tag collection and selective genomic sequencing. *Plant Cell* 14: 1441–1456.
- Wise M.J. 2003. LEAping to conclusions: A computational reanalysis of late embryogenesis abundant proteins and their possible roles. *BMC Bioinformatics* 4: 52.
- Wootton J.C. and Federhen S. 1996. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* 266: 554–571.
- Zhao C., Johnson B.J., Kositsup B. and Beers E.P. 2000. Exploiting secondary growth in *Arabidopsis*. Construction of xylem and bark cDNA libraries and cloning of three xylem endopeptidases. *Plant Physiol.* 123: 1185–1196.
- Zhu W., Schlueter S.D. and Brendel V. 2003. Refined annotation of the *Arabidopsis* genome by complete expressed sequence tag mapping. *Plant Physiology* 132: 469–484.