Genome-Wide Analysis of *LIM* Gene Family in *Populus trichocarpa, Arabidopsis thaliana*, and *Oryza sativa*

Dominique Arnaud, Annabelle Déjardin, Jean-Charles Leplé, Marie-Claude Lesage-Descauses, and Gilles Pilate*

Equipe 'Xylème' Unité Amélioration, Génétique et Physiologie Forestières, INRA-Orléans, Avenue de la Pomme de Pin, BP 20619 Ardon, F-45166 Olivet Cedex, France

Abstract

In Eukaryotes, LIM proteins act as developmental regulators in basic cellular processes such as regulating the transcription or organizing the cytoskeleton. The LIM domain protein family in plants has mainly been studied in sunflower and tobacco plants, where several of its members exhibit a specific pattern of expression in pollen. In this paper, we finely characterized in poplar six transcripts encoding these proteins. In *Populus trichocarpa* genome, the 12 *LIM* gene models identified all appear to be duplicated genes. In addition, we describe several new LIM domain proteins deduced from *Arabidopsis* and rice genomes, raising the number of *LIM* gene models to six for both species. Plant *LIM* genes have a core structure of four introns with highly conserved coding regions. We also identified new LIM domain proteins in several other species, and a phylogenetic analysis of plant LIM proteins reveals that they have undergone one or several duplication events during the evolution. We gathered several LIM protein members within new monophyletic groups. We propose to classify the plant LIM proteins into four groups: αLIM1, βLIM1, γLIM2, and δLIM2, subdivided according to their specificity to a taxonomic class and/or to their tissue-specific expression. Our investigation of the structure of the LIM domain proteins revealed that they contain many conserved motifs potentially involved in their function.

Key words: poplar; *Arabidopsis*; rice; LIM domain protein; tension wood

1. Introduction

LIM proteins have been named by the initials of the three first discovered LIM homeodomain proteins: LIN11, ISL1, and MEC3.[1–3] In Eukaryotes, LIM proteins contain one or more LIM domains, which are, in some cases, associated with a protein kinase domain or a homeodomain. The LIM domain is a cysteine–histidine-rich, zinc-coordinating domain consisting of two zinc fingers repeated in tandem.[4] It is conserved over a wide variety of species. The cysteine-rich protein (CRP) family in animals is a subclass of LIM proteins characterized by their two LIM domains, with the consensus sequence [C-X_2-C-X_{16-23}-H-X_2-C]-X_2-[C-X_2-C-X_{16-21}C-X_{2-3}-(C/D/H)], that are both followed by a short glycine-rich repeat. The CRP family in vertebrates involves four proteins: CRP1, CRP2, CRP3/MLP, and TLP, which act as molecular adapters.[5] Indeed, CRP1, CRP2, and CRP3 are able to bind α-actin and zyxin, two components of the cytoskeleton. Although the LIM domains from animal CRP proteins are structurally similar to the GATA-type zinc finger transcription factor,[6] their DNA-binding activity is yet unproven and CRP may rather be involved in protein–protein interactions.

The LIM protein family in plants is CRP-related proteins containing two LIM domains separated by a long inter-LIM domain. On the contrary to the animal CRP family, plant LIM proteins have a longer C-terminal domain and lack the glycine-rich region (GRR) following each LIM domain. For all plant LIM proteins, the two LIM domains of 52 residues have the following characteristic structure: [C-X_2-C-X_{17}-H-X_2-C]-X_2-[C-X_2-C-X_{17}-C-X_2-H].[7] The first gene encoding a LIM domain protein

---

in plants has been named *SF3*. Later renamed *HaPLIM1*, *SF3* expression was found specific to sunflower pollen.[8,9] The LIM proteins from sunflower, tobacco, and *Arabidopsis* have been classified into four groups: PLIM1 and PLIM2 specifically expressed in pollen and WLIM1 and WLIM2 expressed in the whole plant.[7] Like animal CRP, most plant LIM proteins are present in the cytoplasm and/or in the nucleus. This is the case for the sunflower protein HaWLIM1 that, for many different cell types, localizes either in the cytoplasm, in the nucleus, or in both.[10] Moreover, in protoplasts, HaWLIM1 seems to be associated with cortical microtubules, and it is also observed in the nucleus during the interphase.[11] As for CRP proteins, the tobacco NtWLIM1 binds F-actin and may be involved in actin cytoskeleton stability.[12] The sunflower protein HaPLIM1 has been detected both in small cytoplasmic structures located in the microspores and in the cortical region of mature pollen grains, where it concentrates in the actin-enriched germination cones, suggesting its interaction with the actin cytoskeleton.[13] Although HaPLIM1 was never found in the nucleus of vegetative cells, it exhibits a non-specific DNA- and RNA-binding activity.[14] Hence, the function of HaWLIM1 and HaPLIM1 in the transcriptional regulation remains unclear. In tobacco, NtLIM1 is clearly a transcription factor that binds to the PAL-box, a conserved motif present in the promoter of a number of genes from the phenylpropanoid pathway.[15] Transgenic tobacco plants with a reduced *NtLIM1* expression also present reduced lignin content and a decreased expression of PAL, 4CL, and CAD, three enzymes involved in the lignin biosynthesis. In poplar, the distribution of expressed sequence tags (ESTs) in different wood tissues indicates a rather high expression of an LIM protein homologue in tension wood.[16] Accordingly, microarray analyses also indicate a higher expression of some LIM transcription factor homologues in tension wood.[17] Tension wood, formed on the upper side of bent stems, is enriched in cellulose due to the formation of a supplementary gelatinous layer.

Our study focused on the plant LIM domain proteins containing only two LIM domains homologous to the animal CRP family. First, we finely describe, in this paper, the poplar *LIM* gene family. We determined the complete sequence of six cDNAs encoding LIM proteins and searched for LIM domain protein encoding genes in the *Populus trichocarpa* genome sequence. Secondly, we completed the inventory of the *Arabidopsis* and rice *LIM* gene family. To get a global overview of the plant LIM domain family, cDNAs and ESTs encoding LIM domain proteins have extensively been researched in plant sequence databases. Sequence analyses and phylogenetic studies revealed the structural diversity in plant LIM proteins. We named the genes coding for LIM domain proteins by following a nomenclature stemming from the phylogenetic analysis.

2. Materials and methods

2.1. Characterization of poplar LIM cDNA and gene sequences

The transcripts similar to LIM domain proteins were searched by basic local alignment search tool (BlastN) within a collection of 10 062 ESTs obtained from *Populus tremula* × *Populus alba* (clone INRA #717-1-B4, *Populus* section) wood cDNA libraries.[16] Among the *LIM* ESTs grouped in a same contig, we selected the longest cDNA clone and sequenced it. The entire cDNA was amplified by polymerase chain reaction (PCR) using the forward primer TriplexA and the reverse primer pTriplexB1[16] of the pTriplex vector (Clontech, Laboratories Inc., Mountain View, CA, USA). We further completed the forward and reverse sequencing with new primers specific to each *PtaLIM* gene (*Pta* stands for *P. tremula* × *P. alba*; Supplementary Table S1). The different *PtLIM* genes (*Pt* stands for *P. trichocarpa*, *Tacamahaca* section) were identified in the *Populus* genome database (http://www.genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html)[18] by keyword search, using the InterPro LIM domain annotation (IPR001781) as a query, and using the six cDNAs from *P. tremula* × *P. alba* as query for BlastN searches. We selected only the proteins containing two LIM domains. The prediction of exon/intron splicing in *P. trichocarpa LIM* genes was verified by sequencing reverse transcriptase (RT)–PCR fragments with specific primers against poplar *LIM* genes (see Supplementary Table S1). These fragments were produced after amplification of polyA RNA prepared from poplar wood samples.[19] Finally, the poplar ESTs homologous to the various *PtaLIM* transcripts identified were also searched against the GenBank database (http://www.ncbi.nlm.nih.gov/) and against the PopulusDB database (http://www.populus.db.umu.se), where cDNA libraries were built from *P. tremula* or *P. tremula* × *Populus tremuloides* (*Populus* section) samples.[20]

2.2. Database search for sequences coding for LIM domain protein

Until August 2006, we have used several ways to search in multiple databases all plant LIM proteins containing only two LIM domains. First, known genes and full-length cDNAs encoding LIM domain proteins from sunflower, tobacco, and *Arabidopsis* were collected from the literature and GenBank database.[7] Additional genes and full-length cDNA annotated as 'LIM domain protein', or 'LIM transcription factor' were found by keyword searches or using the InterPro LIM domain annotation (IPR001781) as a query. Alternatively, the poplar LIM protein sequences were used for BLAST (TBlastN and BlastP) searches at the GenBank non-redundant database. The genomic and cDNA sequences of *Arabidopsis thaliana* and *Oryza sativa* were obtained

from the GenBank, TIGR plant genomic group (http://plantgenomics.tigr.org/), and TAIR (http://arabidopsis.org) databases. Finally, to get more sequences encoding LIM domain proteins, ESTs homologous to plant LIM domain proteins were searched by BlastN at the GenBank EST database. This process was repeated with each newly identified set of plant *LIM* genes until no further sequences with significant similarity were identified. For each gene, the longest EST was translated in all reading frames using the EMBOSS Transeq program at the EMBL–EBI, and only those carrying the entire coding sequence (CDS) and some part of 3′ and 5′ untranslated region (UTR) were chosen for further phylogenetic analyses. When EST sequences were too short or did not contain the entire CDS, a consensus sequence was deduced using the Bioedit software. Finally, the deduced amino acid sequences were verified for those carrying the entire two LIM domains using the PROSITE database (http://www.expasy.org/prosite). The selected ESTs with their consensus sequences are listed in Supplementary Table S2.

2.3. Sequence and phylogenetic analysis of LIM domain proteins

The selected protein sequences were aligned using the ClustalW software package (http://www.ebi.ac.uk/clustalw/)[21] with minor adjustments. Phylogenetic analyses were carried out using the Phylogenetic Interference Package (PHYLIP) program, version 3.63 (http://evolution.genetics.washington.edu/phylip.html). Genetic distance matrices using protein polymorphism were calculated using PROTDIST software with the JTT amino acid substitution matrix as measure of distances.[22] A phylogenetic tree was then constructed using the neighbor-joining method thanks to NEIGHBOR software.[23] To estimate the statistical robustness of nodes, 1000 bootstrap samples were generated with SEQBOOT software, and the majority rule consensus tree was generated by CONSENSE software. The plant LIM family was also analyzed through a parsimonious method using the PROTPARS program with 1000 bootstrap replicates. Maximum likelihood analyses were performed using Phyml v2.4.1[24] with the JTT matrix and 100 bootstrap replicates. Maximum likelihood trees were generated with BIONJ, a modified neighbor-joining algorithm.[25] Trees were viewed and edited with Tree View,[26] and bootstrap values <50% were not reported.

Conserved motifs in LIM domain proteins were detected using the ClustalW alignment with few manual corrections, and the MEME program (http://www.meme.sdsc.edu/meme/meme.html).[27] The aligned protein sequences were shaded using the Bioedit software with a threshold of 90% for identical residues and a BLOSUM62 matrix for shading similar residues. Isoelectric point (p*I*) and molecular weight (Mw) were predicted using the p*I*/Mw tool at expasy (http://www.expasy.org/tools/pi_tool.html). PROSITE results were used to find putative ASN-glycosylation and phosphorylation sites.

3. Results and discussion

3.1. Survey and characterization of the LIM domain proteins in poplar, Arabidopsis *and rice*

3.1.1. Isolation of the cDNAs coding for poplar LIM domain proteins and identification of 12 gene models in the Populus trichocarpa *genome*

The distribution of 10 062 poplar ESTs in different wood cDNA libraries[16] revealed that a cDNA, named *PtaGLIM1a* (EF035035), homologous to the LIM protein SF3[8] was abundant in differentiating xylem from tension wood. Five other *PtaLIM* cDNAs present in this EST collection were found, and a complete sequence was determined for each. We named them *PtaGLIM1b* (EF035036), *PtaWLIM2a* (EF035040), *PtaPLIM2a* (EF035037), *PtaWLIM1a* (EF035038), and *PtaWLIM1b* (EF035039). All of the full-length cDNAs obtained contain an entire CDS, a 5′ and a 3′ UTR, and encode protein with two LIM domains (Fig. 1; Supplementary Fig. S1).

We searched the *P. trichocarpa* genome sequence for all the gene models coding for LIM domain proteins. We excluded several gene models coding for proteins containing only one LIM domain linked to either a cytochrome P450 domain or an ubiquitin interaction motif and focused on gene models with two LIM domains. Besides the gene models corresponding to the six cDNAs isolated in our laboratory, we identified six other gene models encoding LIM domain proteins, raising to twelve the number of *LIM* gene models in the *Populus* genome. In accordance with their phylogenetic relationship with other known plant LIM domain proteins, we named these genes *PtWLIM1a*, *PtWLIM1b*, *PtGLIM1a*, *PtGLIM1b*, *Ptβ LIM1a*, *Ptβ LIM1b*, *PtWLIM2a*, *PtWLIM2b*, *PtPLIM2a*, *PtPLIM2b*, *Ptδ LIM2a*, and *Ptδ LIM2b* (Fig. 2A; Supplementary Table S3). With
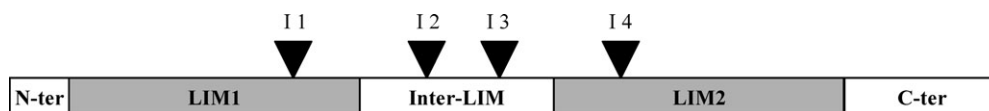

Figure 1. Schematic diagram of the structure of the poplar LIM domain proteins. The conserved LIM1 and LIM2 domains are indicated by gray boxes. Open boxes show the N-terminal, inter-LIM, and C-terminal domains. Arrows indicate the respective position of the four introns (I1–I4).
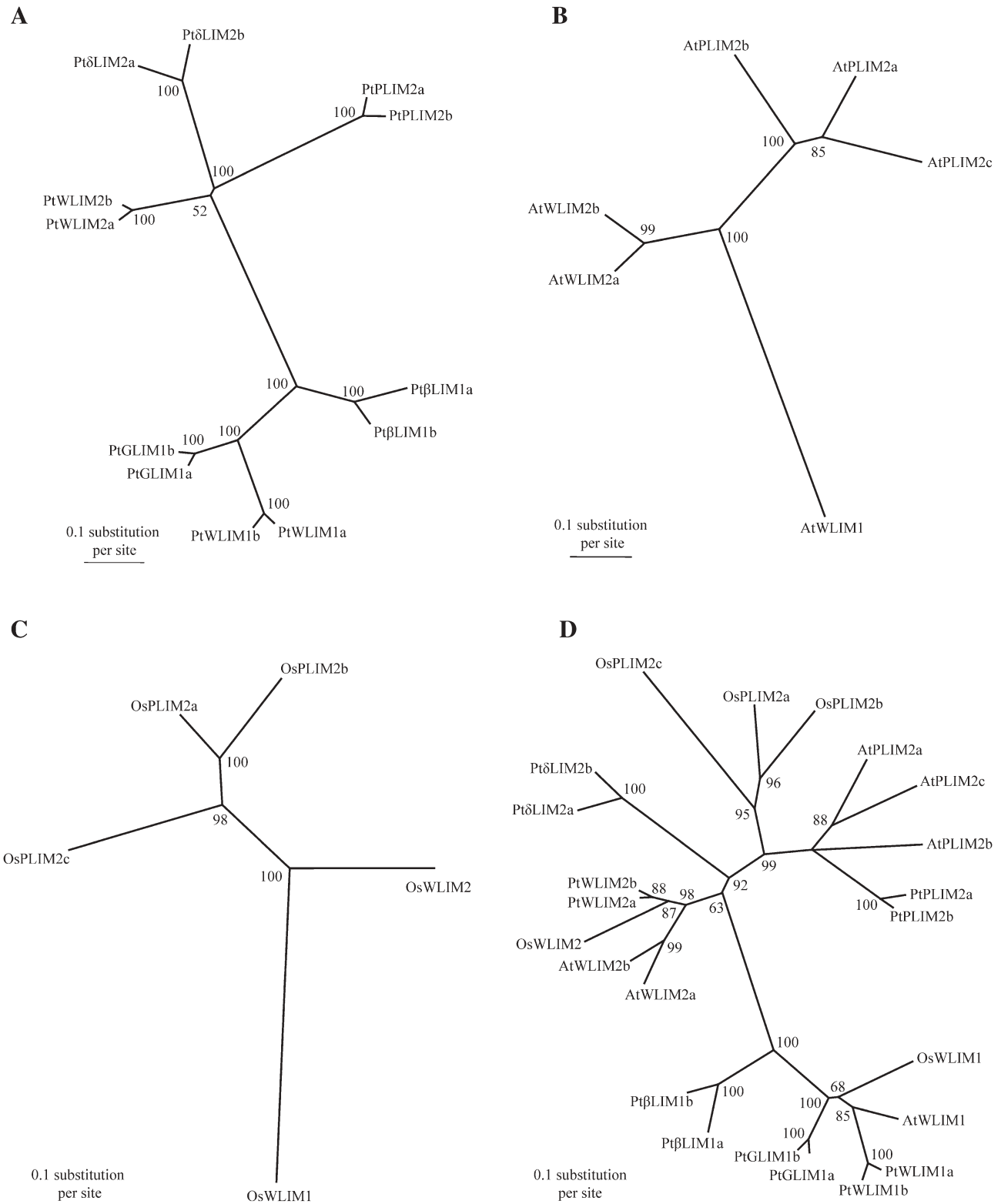
Figure 2. Phylogenetic trees of *P. trichocarpa*, *A. thaliana*, and *O. sativa* LIM domain proteins. Four phylogenetic trees of (**A**) 12 poplar, (**B**) six *Arabidopsis*, (**C**) six rice, and (**D**) all deduced LIM domain proteins are shown. Amino acid sequences of LIM domain proteins were analyzed by neighbor-joining method with genetic distance calculated by the JTT model of amino acids change. The numbers at the nodes represent percent of bootstraps values (≥50%) based on 1000 replications. The length of the branches is proportional to the expected numbers of amino acid substitutions per site with a scale provided at the bottom of the trees. A species acronym is added before each LIM protein name: At, *A. thaliana*; Os, *O. sativa*; Pt, *P. trichocarpa*.

new information from both full-length *PtaLIM* cDNAs and *P. trichocarpa* transcript sequences,[28] we have corrected and improved the annotated poplar genomic sequence for the *PtLIM* genes (Supplementary Data 1).

Each pair of genes (*a* and *b*) exhibits a high sequence similarity from 85% amino acids identity between PtδLIM2a and b to 95% amino acids identity between PtaGLIM1a and b (Fig. 2A; Supplementary Table S4). This high similarity strongly suggests a gene duplication, as previously observed in a recent study for the poplar cellulose synthase *CesA* gene family.[29] These observations are in accordance with the hypothesis that poplar is an ancient polyploid and that a large-scale duplication event has occurred in the ancestor of poplar.[18,30] The duplicated LIM proteins may have kept similar functions, but the distribution of ESTs as well as the expression pattern between duplicated genes (data not shown) differs slightly, suggesting some other differences within their regulatory regions. All the deduced proteins have the features of the plant LIM domain protein family as described previously,[7] with two LIM domains separated by a long-spacer named the inter-LIM region, a short N-ter domain, and a C-ter domain variable in length (Supplementary Fig. S1). The length of PtLIM proteins is rather constant, between 194 and 216 amino acids, and their Mw varies between 20.9 and 24.1 kDa (data not shown). Poplar LIM proteins share a level of homology ranging from 43 to 95% of amino acid identity (Supplementary Table S4). The highest divergence between poplar LIM protein sequences is mainly localized to the inter-LIM and C-ter domains. From the single analysis of amino acid identity between the different poplar LIM domain proteins, we differentiated these proteins into two major groups: LIM1 and LIM2. Within each group, the percentage of identity at the amino acid level is a lot higher (from 58 to 95%) than between the two groups (from 43 to 55%).

3.1.2.   The Arabidopsis *and rice genome contain six LIM gene models*

A previous study reported the identification of three *LIM* genes in the *A. thaliana* genome: *AtWLIM1* (At1g10200), *AtPLIM2* (At2g45800), and *AtWLIM2* (At2g39900).[7] Because the sequences of both the *Arabidopsis* and rice genomes were publicly available, we had the opportunity to identify extensively all the genes coding for LIM domain proteins. Bioinformatics analyses performed against GenBank and TIGR databases show that the *Arabidopsis* genome contains three other *AtLIM* genes. Because these genes seem to be duplicated, we named them *AtPLIM2b* (At1g01780), *AtPLIM2c* (At3g61230), and *AtWLIM2b* (At3g55770, also named *AtL2*[31]), whereas the previous *AtWLIM2* and *AtPLIM2* genes have been renamed *AtWLIM2a* and *AtPLIM2a*, respectively (Fig. 2B). The Supplementary Table S5 lists genomic, cDNA, and EST accessions for the six

Arabidopsis LIM genes, as well as the cDNA and genomic clones that contain errors or encode partial LIM domain proteins. It should be pointed out that the three related genes *AtPLIM2a*, *b*, and *c* overlap, respectively, with a gene encoding a phosphomannomutase (At2g45790), a gene encoding an unknown protein (At1g01770), and a gene encoding an oxydoreductase (At3g61220) that may affect the transcription of these *AtPLIM2* genes.

As in *Arabidopsis*, the rice (*Oryza sativa*) genome contains six genes coding for protein with two LIM domains: *OsWLIM1* (LOC_Os12g32620), *OsWLIM2* (LOC_Os03g15940), *OsPLIM2a* (LOC_Os02g42820), *OsPLIM2b* (LOC_Os04g45010), *OsPLIM2c* (LOC_Os10g35930), and *OsLIM* (LOC_Os06g13030) (Fig. 2C; Supplementary Table S5). *OsPLIM2a*, *b*, and *c* are very similar in their sequences and, therefore, may be considered in-paralogs genes. Unlike poplar, duplication is not the rule in *Arabidopsis* and rice (Fig. 2D). The genes *OsWLIM2*, *OsPLIM2b*, and *OsPLIM2c* are well supported by ESTs and are represented by full-length cDNAs.[32] For *OsWLIM1*, the identified cDNA (AK058220) is truncated. *OsPLIM2a* is also poorly represented at the mRNA level, with only two ESTs found and no published full-length cDNA. Therefore, for these two genes, we used their genomic sequences for the sequence alignment. In the case of *OsLIM*, we identified a very long transcript (AK102383) that encodes an unusual LIM protein of 1303 amino acids with two classical LIM domains and a very long C-ter domain that has no homology for any known protein. Only one (CI584223) of the 22 ESTs found by a BlastN search localizes in the 5′ part of the transcript at the level of the first LIM domain. Because of its unusually long C-ter domain, we did not include this LIM sequence in the phylogenetic analysis.

3.1.3.   *Genomic organization of poplar,* Arabidopsis, *and rice genes*

From the genomic analysis, we can infer that all plant *LIM* genes have a core gene structure with four introns within the coding sequence with the exception of *AtPLIM2a* and *c* genes that have two and three introns, respectively (Fig. 1; Supplementary Fig. S2). The position of the first and last introns is strictly conserved in the first and second LIM domain, respectively. In poplar and *Arabidopsis*, the *WLIM2* genes diverge from the other *LIM* genes by the occurrence of one (*AtWLIM2a*, *PtWLIM2a*, and *b*) or two (*AtWLIM2b*) supplementary introns in the 5′ UTR, before the ATG initiation codon. For these *WLIM2* genes in eudicots species, a mechanism of alternative splicing of the first intron may be involved in post-translational regulation. This is supported by northern-blot experiments revealing two hybridizing bands for *AtWLIM2b* only in the shoot and not in the root.[31]

The length of coding regions CR1, CR2, and CR4 is highly conserved between plant *LIM* genes, indicative of

a strict conservation in the length of the LIM domains during evolution (Supplementary Fig. S2). The first coding region is 135–138 bp long and only *OsWLIM1*, *OsWLIM2*, and *OsLIM* have a variable CR1 length. The second coding region is 97 bp long for genes belonging to the LIM1 group and 100 bp long for genes from the LIM2 group. The fourth exon, localized within the second LIM domain, is the most conserved exon, with a length of 90 bp for all *LIM* genes. The CR3 and CR5 are highly variable in length resulting in the differences observed at the amino acid level, respectively, in the interLIM region and the C-ter domain. Finally, the *PLIM2* and *δLIM2* genes contain the longest fifth coding region that reflects the extensive length of the C-ter region of the deduced proteins.

3.1.4. Identification of ESTs homologous to LIM domain protein

In a previous study, LIM proteins from sunflower, tobacco, and *Arabidopsis* have been classified into two groups, LIM1 and LIM2, and subdivided into four subgroups: PLIM1 and PLIM2, specifically expressed in pollen, and WLIM1 and WLIM2, widely expressed in plant.[7] Because of the availability of the genome sequence for *P. trichocarpa*, *A. thaliana*, and *O. sativa*, we found an increased number of genes belonging to the plant LIM domain family. The newly discovered LIM proteins may define new LIM subgroups or be related to the previously identified subgroups. To approach the diversity of the *LIM* gene family, an extensive BlastN search of cDNAs and ESTs encoding proteins with two LIM domains has been performed in NCBI plant sequence databases. In plants, we found 165 unigenes homologous to LIM domain protein, but we did not include 49 of them in the phylogenetic analysis because they contained only partial CDS (data not shown). Within the 116 unigenes coding for an entire LIM domain protein, 90 have a representative EST containing a full length CDS, whereas for each of the remaining 26 unigenes, a consensus was built to determine the complete CDS (Supplementary Table S2). We identified ESTs from a wide range of species in the different groups of

plants: bryophytes, conifers, piperales, monocotyledons, and eudicotyledons including rosids, asterids, caryophyllales, and banunculales subclasses. In tobacco (*Nicotiana tabacum*) and in the sunflower (*Helianthus annuus*), the number of different *LIM* transcripts has been raised to eight with the finding of novel ESTs, notably the *HpWLIM2* EST from *Helianthus petiolaris*, whose existence was previously suspected in this genus by others.[7]

3.2. Phylogenetic analysis of plant LIM proteins with regard to expression data

3.2.1. Four different groups

Phylogenetic trees have been constructed with the deduced amino acid sequence of the genes encoding LIM domain proteins (Fig. 3). We renamed the plant LIM proteins according to their phylogenetic relationship. The LIM1 and LIM2 groups identified previously[7] are clearly separated and supported by a high bootstrap value at the level of the TrLIM and PpLIM proteins from the mosses *Physcomitrella patens* and *Tortula ruralis* (Fig. 3). We were unable to place these two LIM proteins within either group because their sequence shares similarities with both LIM groups. Therefore, the phylogenetic trees have been rooted using sequences of these two Bryophyte LIM proteins. The plant LIM family can be divided into four groups, *α*LIM1, *β*LIM1, *γ*LIM2, and *δ*LIM2 resulting from the division of the LIM1 and LIM2 groups. These four groups are supported by high bootstrap values. This phylogenetic analysis confirms the existence of the PLIM1, WLIM1, PLIM2, and WLIM2 subgroups as described previously.[7] *β*LIM1 is a new group that has not been identified before, whereas the *α*LIM1 group includes PLIM1 and WLIM1 subgroups. The WLIM2 and PLIM2 subgroups belong, respectively, to the *γ*LIM2 and *δ*LIM2 groups. In addition, each group contains new subgroups, which are described below.

3.2.2. The αLIM1 group

The previous WLIM1 and PLIM1 subgroups are gathered within the *α*LIM1 group. With regards to the low bootstrap value

Figure 3. Phylogenetic tree of plant LIM domain proteins. Amino acid sequences of 149 LIM domain proteins were analyzed by neighbor-joining method with genetic distance calculated by the JTT model of amino acids change. The numbers at the nodes represent bootstraps values ($\geq$500) based on 1000 replications. The lengths of the branches are proportional to the expected numbers of amino acid substitutions per site with a scale provided at the bottom of the tree. LIM proteins have been renamed following their belonging to groups or subgroups WLIM1, PLIM1, FLIM1, *α*LIM1, *β*LIM1, WLIM2, PLIM2, or *δ*LIM2 that are in boxes. Poplar LIM proteins are in blue color, and plant LIM proteins whose expression pattern has been determined are in red color. A species acronym is added before each LIM protein name: Afp, *Aquilegia formosa* × *A. pubescens*; Am, *Antirrhinum majus*; At, *A. thaliana*; Bd, *Brachypodium distachyon*; Bn, *Brassica napus*; Bv, *Beta vulgaris*; Cc, *Coffea canephora*; Cl, *Curcuma longa*; Cs, *Citrus sinensis*; Cr, *Ceratopteris richardii*; Ec, *Eleusine coracana*; Ee, *Euphorbia esula*; Et, *Euphorbia tirucalli*; Eg, *Eucalyptus globulus*; Fv, *Fragaria vesca*; Geh, *Gerbera hybrida*; Gm, *Glycine max*; Gh, *Gossypium hirsutum*; Gr, *Gossypium raimondii*; Ga, *Gossypium arboretum*; Ha, *Helianthus annuus*; Hp, *Helianthus petiolaris*; Hv, *Hordeum vulgare*; In, *Ipomoea nil*; Le, *Lycopersicon esculentum*; Lp, *Lycopersicon pennellii*; Ls, *Lactuca sativa*; Md, *Malus* × *domestica*; Mt, *Medicago truncatula*; Mc, *Mesembryanthemum crystallinum*; Nt, *Nicotiana tabacum*; Ob, *Ocimum basilicum*; Pa, *Prunus armeniaca*; Pg, *Picea glauca*; Ph, *Petunia* × *hybrida*; Pit, *Pinus taeda*; Pp, *Physcomitrella patens*; Pt, *Populus trichocarpa*; Pv, *Panicum virgatum*; Sea, *Senecio aethnensis*; Sec, *Senecio chrysanthemifolius*; Sb, *Sorghum bicolor*; Sh, *Saruma henryi*; So, *Saccharum officinarum*; St, *Solanum tuberosum*; Sc, *Solanum chacoense*; Ta, *Triticum aestivum*; To, *Taraxacum officinale*; Tk, *Taraxacum kok-saghyz*; Tr, *Tortula ruralis*; Vv, *Vitis vinifera*; Zo, *Zingiber officinale*; Zm, *Zea mays*. For additional details on each gene, see Supplementary Table S2.
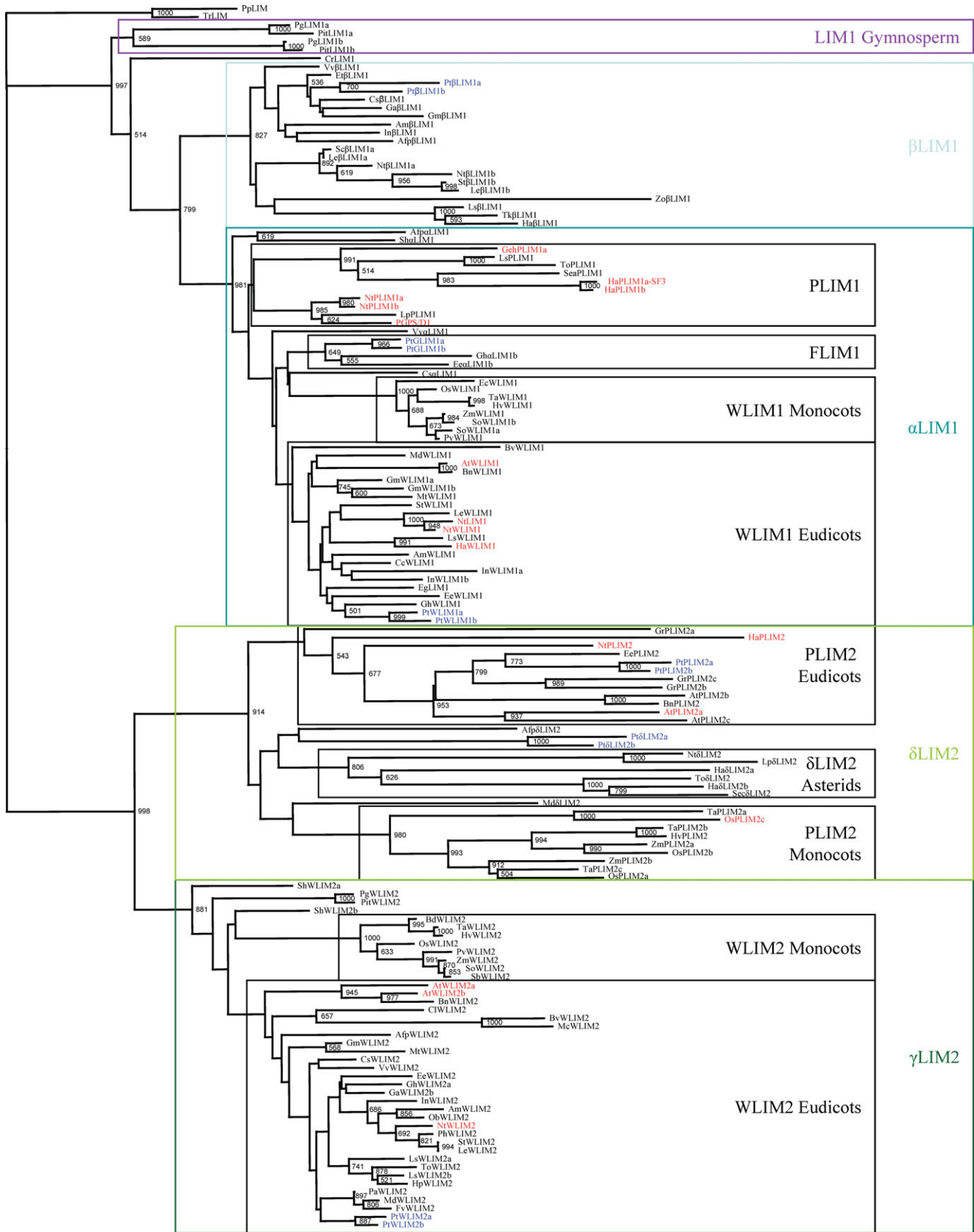
0.1 substitutions per site

supporting the node, the definition of these two sub-groups remains questionable (Fig. 3). In contrast, the monocots WLIM1 subgroup clearly forms a statistically significant new monophyletic clade distinct from the eudicots WLIM1 subgroup. Additionally, a fourth subgroup FLIM1 could also be assigned to the αLIM1 group. However, the respective position of the PLIM1, FLIM1, and monocots and eudicots WLIM1 subgroups within the αLIM1 group needs to be clarified. Indeed, the neighbor-joining trees generated from a matrix of distances calculated by the JTT method or from the maximum likelihood (PhyML) method gave us an FLIM1 subgroup close to the eudicots and monocots WLIM1 subgroups and clearly separated from the PLIM1 subgroup (Fig. 3; Supplementary Fig. S3). However, when using the parsimonious method, the FLIM1 subgroup takes place within the PLIM1 subgroup, between the PLIM1 proteins from the Solanaceae and the Asteraceae families (Supplementary Fig. S3). Furthermore, this last method favors the hypothesis of a common ancestor for the monocots WLIM1 and PLIM1 subgroups. Presently, the FLIM1 subgroup includes only a few proteins, namely PtGLIM1a and b, GhαLIM1, and EeαLIM1. Although additional sequences are definitely needed to strengthen this classification, there are strong arguments for the creation of a FLIM1 subgroup: (i) PtGLIM1a and b share 81% of amino acid identity with GhαLIM1 and 79% with EeαLIM1; (ii) RT–PCR analyses indicate that the *PtGLIM1a* and *b* genes are not expressed in pollen (data not shown), which could be an argument against grouping *PtGLIM1a* and *b* within the PLIM1 subgroup. With 58 ESTs coming mostly from different xylem cDNA libraries (Fig. 4), *PtGLIM1a* is probably highly expressed in xylem tissues and to a lesser extent in other organs with vascular tissues, such as leaves. In contrast to *PtGLIM1a*, *PtGLIM1b* ESTs are preferentially found in cambial zone and phloem. The *PtGLIM1a* and *b* names include a G, indicative of their abundance in tension wood with G-fibers.[16,17] Similarly, all of the 31 *GhαLIM1* ESTs were found in mature cotton fibers and a majority (29 ESTs) came from fibers harvested during secondary cell wall formation (Supplementary Table S2). Accordingly, we named this subgroup FLIM1 with the F, indicative of the high expression of its members in fibers. We can speculate that during the evolution of poplar and cotton species, these three proteins may have gained a novel function important for fiber differentiation or maturation and more particularly in fibers containing a cellulose-enriched secondary cell wall.

Interestingly, the PLIM1 subgroup contains only LIM proteins from Solanaceae and Asteraceae species from the asterids subclass (Fig. 3). Many expression studies have demonstrated the pollen-specific expression of these *PLIM1* genes. For example, the sunflower HaPLIM1a (SF3) is a protein exclusively and highly expressed in maturing pollen,[13] and its colocalization with actin in the germinating cone suggests that HaPLIM1a participates in pollen germination or pollen tube elongation. The tobacco *NtPLIM1a* gene has also been shown to be highly expressed in mature pollen,[33] whereas in *Gerberea hybrida*, the expression of *GhPLIM1a* gene is highly upregulated in stamens as evidenced in microarray studies.[34] Likewise, the petunia *PGPS/D1* gene encoding a PLIM1 protein has been found to be highly expressed in germinating pollen,[35] and the accumulation of the *PhPLIM1* transcript increases during anther maturation suggesting again an involvement of PLIM1 proteins in pollen tube elongation. PLIM1 proteins from Solanaceae species are clearly separated from those belonging to the Asteraceae species, suggesting that PLIM1 proteins have diverged during the speciation of these two families.

The monocots WLIM1 subgroup contains proteins with high homology (from 89 to 99% of amino acid identity). Monocots *WLIM1* ESTs originate from a large amount of tissue including inflorescences. Only the closely related *SoWLIM1a* and *b* genes from sugarcane seem to be duplicated, as the two maize *WLIM1* transcripts described are probably alleles of the same gene. Surprisingly, a high number of *WLIM1* ESTs has been sequenced from the maize (200) and sugarcane (128) cDNA libraries. In particular, *SoWLIM1a* and *b* with 76 ESTs is the biggest cluster of the 'shoot-root transition zone from adult plant' cDNA library.

All plant species from core eudicotyledons are represented within the large eudicots WLIM1 subgroup. Within this subgroup, only poplar, *Ipomoea*, and soybean species have in-paralogs WLIM1 proteins (Fig. 3). In general, each plant *WLIM1* gene is represented by a large number of ESTs originating from a wide variety of tissues, sometimes including inflorescence organs. The *WLIM1* genes have been the most studied *LIM* genes in plants, but their precise function is yet to be defined. In the sunflower, the HaWLIM1 protein is localized in the nucleus and/or the cytoplasm for different cell types,[10] and may be associated with the tubulin cytoskeleton in protoplasts.[11] A tobacco *WLIM1* gene, *NtLIM1*, highly expressed in the stem, acts as a transcription activator. Indeed, the binding of the NtLIM1 protein to the PAL-box motif, present on the promoter of several genes from the lignin biosynthesis pathway, leads to the activation of the transcription of these genes. Moreover, an important reduction in the lignin content has been observed in antisense *Ntlim1* transgenic tobacco.[15] NtLIM1 protein mostly differs from NtWLIM1 protein[7] by seven additional residues in the C-ter of the protein. The NtWLIM1 protein binds actin directly with a high affinity, enhances the stability of actin cytoskeleton, and promotes the bundling of actin filaments.[12] The poplar ESTs homologous to *PtWLIM1a* and *b* are more represented in vascular tissues with many ESTs from phloem, cambium, and xylem (Fig. 4). In *Eucalyptus globulus*, the homologous *EgLIM1* gene (AB208710) and
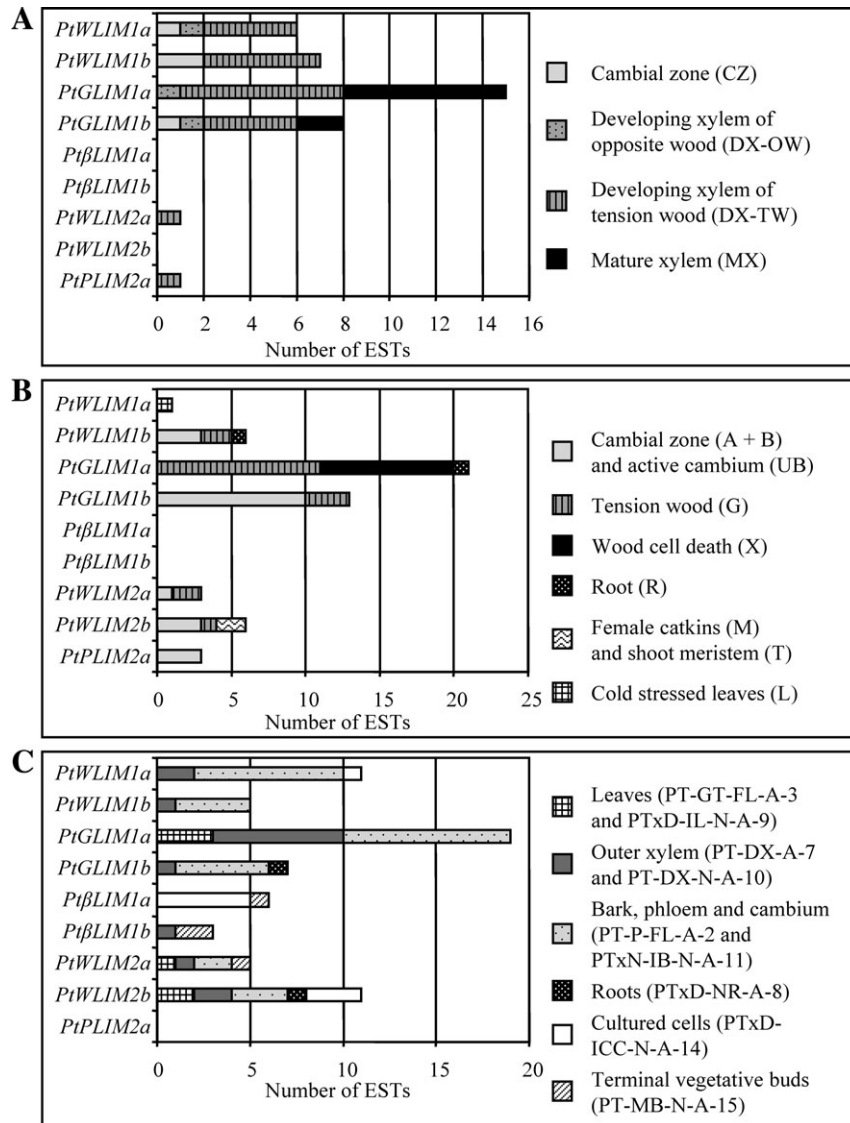
Figure 4. Distribution of poplar LIM EST in different poplar cDNA libraries. (**A**) Libraries from INRA-Orléans.[16] (**B**) Libraries from Umea Plant Science Centre (UPSC).[20] (**C**) Libraries from University of British Columbia (UBC):[28] cDNA libraries prepared from young and mature leaves, along with green shoot tips (PT-GT-FL-A-3), local mature leaves harvested after continuous feeding by *M. disstria* (PTxD-IL-N-A-9), outer xylem harvested biweekly between April and October 2002 (PT-DX-A-7 and PT-DX-N-A-10), phloem and cambium (PT-P-FL-A-2), bark (with phloem and cambium attached) harvested after continuous feeding by *C. lapathi* (PTxN-IB-N-A-11), roots harvested from 3-month-old trees grown in hydroponic media without nitrogen source for 24 and 48 h, as well as trees grown in regular media (PTxD-NR-A-8), cultured cells grown in media supplemented with salicylic acid, benzothiadazole, methyl jasmonate, chitosan, or *Pollacia radiosa* extract (PTxD-ICC-N-A-14); for more details, see Ralph et al.[28]

the corresponding cDNA (AB208709) have also been isolated, but there is no expression information available. We have been unable to position the ShαLIM1, AfpαLIM1, VvαLIM1, and CsαLIM1 proteins in either subgroup of the αLIM1 group, as illustrated by the low bootstrap values supporting the nodes.

In conifer trees, the LIM1 proteins are very similar in structure in white spruce (*Picea glauca*) and loblolly pine (*Pinus taeda*). For both species, the LIM1 proteins have undergone a probable duplication event, and it is difficult to class them into either the αLIM1 or the βLIM1 subgroup (Fig. 3). In *Pinus taeda*, the *PitLIM1b* (clone

6C12H) gene is more highly expressed in the xylem than in any other part of the tree.[36] Therefore, at least in tree species, the LIM domain proteins from the WLIM1 subgroup, as well as those from the FLIM1 subgroup, are preferentially expressed during vascular development.

3.2.3. The new βLIM1 group   The new βLIM1 group contains LIM proteins from species mostly belonging to the asterids and rosids subclasses, including the poplar duplicated PtβLIM1a and b proteins. Interestingly, there are no *Arabidopsis* or rice proteins into this group (Fig. 3). The proteins from Solanaceae and Asteraceae

form two distinct smaller clades within this group. Within the asterids subclass, only the Solanaceae species (tobacco, potato, and tomato) have in-paralogs βLIM1 proteins. The expression or function of the *βLIM1* genes is yet unknown. Like the genes from the WLIM1 and WLIM2 subgroups, *βLIM1* ESTs originate from a variety of tissues including flower and fruit (data not shown), but they are preferentially found in roots, in cells that undergo differentiation like immature cotton fibers, and in undifferentiated cells like callus. In poplar, the *PtβLIM1a* ESTs almost exclusively originate from cell suspension cultures and the *PtβLIM1b* ESTs are more widely represented in xylem, vegetative buds, and roots (Fig. 4). Although *βLIM1* ESTs originate from all parts of the plant, a significant number has been sequenced from undifferentiated cell samples.

3.2.4.   The γLIM2 group   The γLIM2 group is clearly separated from the δLIM2 group with a high bootstrap value (Fig. 3). The γLIM2 group contains the monocots and eudicots WLIM2 subgroups that are clearly separated. As in the WLIM1 subgroup, the short branches are indicative of a high conservation between WLIM2 protein sequences within each subgroup. In several eudicots such as *Arabidopsis*, poplar, cotton, and lettuce, *WLIM2* genes are duplicated, whereas this is never the case in monocot species. For both monocots and eudicots WLIM2 subgroups, a high number of ESTs from various tissues, including inflorescences, have been sequenced. This is the case in *Arabidopsis* for the *AtWLIM2a* gene[7] as well as in poplar, for the *PtWLIM2a* and *b* genes (Fig. 4). The wide range of expression and the high conservation of *WLIM2* genes suggest that they may be involved in basic cellular processes. This is substantiated by the difficulty to produce homozygous mutants for *AtWLIM2a* gene.[37] In a recent microarray study, *AtWLIM2a* appeared highly expressed in siliques, and its expression was induced in the *Arabidopsis pkl* mutant.[37] The PKL protein is a negative regulator of transcription that acts as a repressor of embryonic identity during seed germination, and *AtWLIM2a* gene may be involved in the regulation of embryo development. In another study, *AtWLIM2a* expression is also gradually induced during leaf senescence.[38] Taken together, the *AtWLIM2a* gene may be more precisely involved in the seed maturation. The duplicate of *AtWLIM2a*, *AtWLIM2b* (*AtL2*), is more highly expressed in roots than in shoots. In addition, this expression seems to be affected by nitrogen availability.[31]

3.2.5.   The δLIM2 group   Because of the increased number of LIM sequences available, the δLIM2 group is considerably enlarged. The δLIM2 group is divided into three monophyletic subgroups: the eudicots PLIM2, monocots PLIM2, and Asterids δLIM2 that contains the sequences from the new LIM domain proteins identified in sunflower and tobacco plants (Fig. 3). Several duplication events have occurred within the *δLIM2* genes, indicative of an important diversification within this group. Because of the long length of branches, the δLIM2 proteins are on average more divergent each other than proteins from any other groups. Most eudicots *PLIM2* genes appear preferentially expressed during pollen development. This is the case for *AtPLIM2a* expressed in hydrated pollen grains,[39] for the tobacco *NtPLIM2*, and the sunflower *HaPLIM2* genes.[7] Likewise, in *Gerbera hybrida*, the ESTs encoding an homologue to *HaPLIM2* also appeared exclusively expressed in stamens [34] and in tomato, all the EST sequences orthologous to the *NtPLIM2* gene originated from a pollen cDNA library (data not shown). Surprisingly, a poplar cDNA library prepared from male catkins[20] does not contain any *PtPLIM2a* and *b* EST, but we observed for the *PtPLIM2a* gene a strong expression in mature anthers (data not shown). It should be noted that a few *PtPLIM2a* ESTs originate from cambial zone, tension wood, and root poplar libraries (Fig. 4), but no *PtPLIM2b* EST has been sequenced yet.

Monocots PLIM2 proteins are phylogenetically separated from eudicots PLIM2 proteins supported by a high bootstrap value. In monocots, the PLIM2 proteins are generally found in triplicate, resulting from at least two duplication events. Contrary to the highly conserved monocots WLIM2 proteins, monocots PLIM2 proteins are more divergent. As for eudicots, the monocots *PLIM2* genes are generally highly expressed in pollen, but not exclusively. Of the 200 *ZmPLIM2a* ESTs found in public databases, 156 come from flower cDNA libraries, and 66 from these ESTs originate from mature pollen. The majority of *OsPLIM2c* and *OsPLIM2b* ESTs also come from flower cDNA libraries. *OsPLIM2c* (AK072520) expression has been studied using microarray analysis during pollination and early embryogenesis in the rice pistil.[40] Their maximal accumulation in the pistil occurs at anthesis and decreases gradually during the following 24 h, when the pollen tube reaches the micropyle. However, the same study reveals a wider expression of the *OsPLIM2c* gene, suggesting an involvement in some other processes than pollen tube elongation. We speculate that within monocots and rosids species, the extensive duplication of their *PLIM2* genes may be a counterpart to the lack of specialization of the genes from the αLIM1 group toward pollen-specific expression.

In asterids species, *LIM* genes also seem duplicated within the δLIM2 group. Contrary to the *Arabidopsis PLIM2* genes found in triplicate within the PLIM2 subgroup, the newly discovered tobacco and sunflower δLIM2 proteins are more distantly related from the previously identified HaPLIM2 and NtPLIM2 and therefore belong to a new Asterids δLIM2 subgroup (Fig. 3). The Asterids δLIM2 subgroup is surrounded by the two monocots and eudicots PLIM2 subgroups. Because dicots and

monocots *PLIM2* genes are strongly expressed in pollen, it is probable that these asterids *δLIM2* genes are also strongly expressed in pollen, but it is not clear if these duplication events have brought any new functions in pollen development or in other processes. Poplar PtδLIM2a and b protein sequences deduced from the *P. trichocarpa* genome are phylogenetically distant from the PLIM2 subgroups. So far there is no expression data for their potential pollen-specific expression.

3.3. Sequence and structural analysis of plant LIM domain protein

3.3.1. Toward a new characterization of the two LIM domains consensus sequences in plant

The plant LIM domain proteins present a lot of common features with the CRP proteins, but also several specificities. This is the case for the GRR following each LIM domain found in the animal CRP proteins that lacks in all the plant LIM domain proteins.[7] However, for all plant LIM domain proteins, a glycine residue is strictly conserved nine amino acids after the last zinc ligand for each LIM domain (Fig. 5B). The potential nuclear targeting signal (KKYGPK) present in the CRP family of animal LIM proteins is also missing in the plant LIM domain family. The LIM domains contain two zinc finger repeated in tandem with the characteristic structure $[C-X_2-C-X_{17}-H-X_2-C]-X_2-[C-X_2-C-X_{17}-C-X_2-H]$ (Fig. 5).[7] Compared with the CRP family in animals, the second LIM domain in plants is atypical with the second ligand of the second zinc finger replaced by a conserved glycine. It has been proposed that closer histidine or cysteine makes the formation of two alternative structures $-C-X-H-X_{18}-C-X_2-H-$ or $-C-X_4-C-X_{15}-C-X_2-H-$ possible for the zinc coordination.[10] However, the increased number of protein sequences analyzed reveals that this histidine residue is sometimes lacking in the LIM proteins from the βLIM1 and δLIM2 groups (Supplementary Fig. S4). Only the cysteine residue is conserved in all plant LIM domain proteins, strongly suggesting that the $-C-X_4-C-X_{15}-C-X_2-H-$ is the most probable structure for the
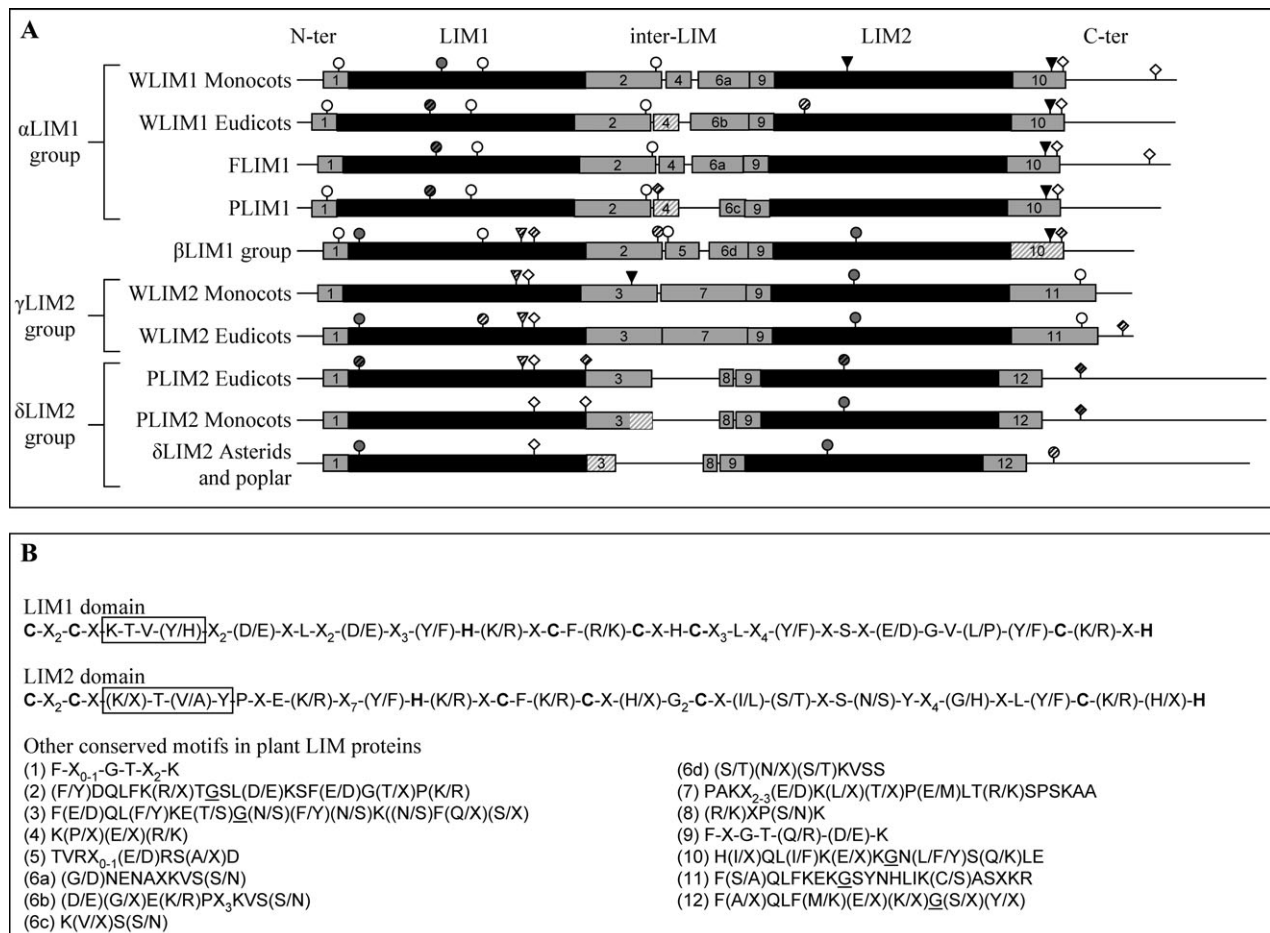

Figure 5. (**A**) Diagrammatic representation of motifs found in the sequences of the plant LIM family proteins. The LIM1 and LIM2 domains are represented by black boxes. ASN glycosylation sites are shown in triangles. cAMP- and cGMP-dependent protein kinase (cAMP), CKII, Tyr, and PKC phosphorylation sites are, respectively, shown in grey and white diamonds, grey and white circles. Non-conserved motifs are striped. Conserved motifs surrounding the LIM domains and identified with ClustalW[21] and MEME[27] programs are in gray boxes. (**B**) Consensus sequences of the two plant LIM domains and the 15 motifs identified. Cysteine and histidine coordinating zinc finger are highlighted in bold. The last conserved glycine after each LIM domain is underlined and the conserved K[T/A]VY motif is boxed. X represents any amino acid residue.

second zinc finger. Interestingly, the first cysteine of the first LIM domain is missing only for the sugarcane-duplicated SoWLIM1b protein and in maize for the second allelic version of ZmWLIM1 protein supported by the cDNA AY112454 (data not shown). If this cysteine is compulsory for the formation of the zinc finger, the functional significance of this absence remains to be elucidated. In both animal CRP and plant LIM families, with the exception of the monocots PLIM2 protein, a highly conserved K[T/A]VY motif is found in the first LIM domain close to the second cysteine of the first zinc finger (Fig. 5B). This common motif may be important in the correct folding of the LIM1 domain, and is also found, but conserved to a lesser extent, at the same position in the second LIM domain. In general, the amino acids surrounding the LIM domains are also highly conserved between LIM proteins of each subgroup with the noticeable exception of the δLIM2 proteins (Fig. 5A and B). It should be pointed out that the LIM1 domain is generally more conserved than the LIM2 domain. However, in WLIM2 proteins, both LIM domains are highly conserved.

3.3.2. Structural variation may reflect functional differences

As for animals, plant LIM proteins may have a function in the regulation of cell differentiation or development. LIM proteins could have multiple partners in the cell, such as actin or tubulin cytoskeleton, DNA and other potential proteins. It has been proposed that plant LIM proteins may serve as a shuttle between the cytoplasm and the nucleus for yet unknown functions. In this respect, we searched for putative phosphorylation and glycosylation sites in the LIM protein sequences using the Prosite database. Thanks to the high number of sequences collected, we have been able to perform ortholog sequence comparisons within subgroups to find conserved motifs using the MEME program. Most of the differences observed between these LIM proteins are concentrated either in the inter-LIM region or in the C-ter domain (Fig. 5). Within the αLIM1 group, PLIM1 proteins are more divergent each other than WLIM1 and FLIM1 proteins. The C-ter domain of PLIM1 proteins is more different than in WLIM1 proteins and contains additional proline residues (Supplementary Fig. S4). FLIM1 proteins mostly resemble WLIM1 proteins, particularly at the level of the LIM domains. Their inter-LIM domain is more similar to those of monocots WLIM1 proteins, whereas their C-ter domain differs clearly from that of the other subgroups. Proteins within the βLIM1 group carry on their LIM domains, the structural properties characteristic of both WLIM2 and αLIM1 groups, but the inter-LIM and C-ter domains are specific to this group. WLIM2 proteins have highly conserved sequences, even at the level of the inter-LIM region, except for WLIM2 proteins from conifers and caryophyllales. The δLIM2 group is the most heterogeneous

LIM group even at the level of LIM domains. The δLIM2 proteins are characterized by a long C-ter domain highly variable in amino acid composition, and composed of many acidic amino acids, particularly glutamic acid (Supplementary Fig. S4). It has been suggested for tobacco NtLIM1 protein that this acidic domain may function as a transcription activator.[15] Interestingly, on the contrary to PLIM2 and αLIM1 proteins, the C-ter domain of WLIM2 proteins is rather basic and contains non-polar amino acids except one conserved acidic amino acid at the end of the protein. The C-ter domain of βLIM1 proteins has a variable p*I*. The p*I* is rather basic in the Solanaceae-duplicated proteins, and acid to neutral for the Asteraceae proteins (data not shown). Different kinds of motifs like the ASN glycosylation site, the casein kinase II (CKII), the tyrosine (Tyr), and the protein kinase C (PKC) phosphorylation sites are localized inside and outside the LIM domains (Fig. 5). For example, in αLIM1 and βLIM1 proteins and only in these proteins, a putative PKC phosphorylation site is found just before the first zinc ligand of the LIM1 domain. However, the FLIM1 proteins (from the αLIM1 group) do not possess this phosphorylation site. Finally, another PKC phosphorylation site is exclusively found at the C-ter domain in WLIM2 proteins from angiosperm and also gymnosperm species, suggesting a key function for this site.

3.4. Concluding remarks

In this work, we report on the first analysis of poplar LIM domain protein family with the characterization of six *PtaLIM* cDNAs. Moreover, we updated the *LIM* genomic sequences in *P. trichocarpa*, *A. thaliana*, and *O. sativa*. The set of ESTs collected and the subsequent phylogenetic classification of LIM proteins provide a useful database for future researches on this family. We have defined four phylogenetic groups, αLIM1, βLIM1, γLIM2, and δLIM2, and positioned into these groups, the previously characterized PLIM1, WLIM1, PLIM2 and WLIM2 subgroups. The βLIM1 is a newly identified group, whose genes have an apparent preferential expression in undifferentiated cells. However, more detailed expression studies are needed for the functional analysis of these βLIM1 proteins. Besides, the pollen-specific expression of *PLIM1* and *PLIM2* genes, several subgroups seem to be specific to a plant class or subclass. This is the case for the PLIM1 subgroup that is represented only by LIM protein from the asterids subclass. Additionally, it appears that LIM proteins from monocots and eudicots plant are phylogenetically distant, suggesting specific functions within each taxonomic class. Finally, a new FLIM1 subgroup was created within the αLIM1 group, and the corresponding genes appear to be highly expressed in poplar G-fibers or cotton fibers. In poplar, the distribution of ESTs

corresponding to *WLIM1* and *GLIM1* genes suggests the involvement of LIM proteins in wood fibers formation and/or vascular development. Despite the fact that plant LIM proteins could both bind to actin and act as a transcription factor, the function of LIM proteins remains largely unknown. Structural analysis is the first step to the functional characterization of LIM proteins in plants. In the future, we will need to determine whether all plant LIM proteins are able to bind actin cytoskeleton and/or DNA, and to examine the function of PLIM1 and PLIM2 proteins in pollen development or pollen tube elongation.

References

1. Freyd, G., Kim, S. K. and Horvitz, H. R. 1990, Novel cysteine-rich motif and homeodomain in the product of the *Caenorhabditis elegans* cell lineage gene *lin-II*, *Nature*, **344**, 876–879.

2. Karlsson, O., Thor, S., Norberg, T., Ohlsson, H. and Edlund, T. 1990, Insulin gene enhancer binding protein Isl-1 is a member of a novel class of proteins containing both a homeo- and a Cys–His domain, *Nature*, **344**, 879–882.

3. Way, J. C. and Chalfie, M. 1988, *mec-3*, a homeobox-containing gene that specifies differentiation of the touch receptor neurons in *C. elegans*, *Cell*, **54**, 5–16.

4. Dawid, I. B., Breen, J. J. and Toyama, R. 1998, LIM domains: multiple roles as adapters and functional modifiers in protein interactions, *Trends Genet.*, **14**, 156–162.

5. Weiskirchen, R. and Günther, K. 2003, The CRP/MLP/ TLP family of LIM domain proteins: acting by connecting, *BioEssays*, **25**, 152–162.

6. Perez-Alvarado, G. C., Miles, C., Michelsen, J. W., et al. 1994, Structure of the carboxy-terminal LIM domain from the cysteine rich protein CRP, *Nat. Struct. Biol.*, **1**, 388–398.

7. Eliasson, Å., Gass, N., Mundel, C., et al. 2000, Molecular and expression analysis of a LIM protein gene family from flowering plants, *Mol. Gen. Genet.*, **264**, 257–267.

8. Baltz, R., Domon, C., Pillay, D. and Steinmetz, A. 1992, Characterization of a pollen-specific cDNA from sunflower encoding a zinc finger protein, *Plant J.*, **2**, 713–721.

9. Baltz, R., Evrard, J. L., Domon, C. and Steinmetz, A. 1992, A LIM motif is present in a pollen-specific protein, *Plant Cell*, **4**, 1465–1466.

10. Mundel, C., Baltz, R., Eliasson, A., et al. 2000, A LIM-domain protein from sunflower is localized to the cytoplasm

11. Briere, C., Bordel, A.-C., Barthou, H., et al. 2003, Is the LIM-domain protein HaWLIM1 associated with cortical microtubules in sunflower protoplasts?, *Plant Cell Physiol.*, **44**, 1055–1063.

12. Thomas, C., Hoffmann, C., Dieterle, M., Van Troys, M., Ampe, C. and Steinmetz, A. 2006, Tobacco WLIM1 is a novel F-Actin binding protein involved in actin cytoskeleton remodeling, *Plant Cell*, **18**, 2194–2206.

13. Baltz, R., Schmit, A.-C., Kohnen, M., Hentges, F. and Steinmetz, A. 1999, Differential localization of the LIM domain protein PLIM-1 in microspores and mature pollen grains from sunflower, *Sex. Plant Reprod.*, **12**, 60–65.

14. Baltz, R., Evrard, J.-L., Bourdon, V. and Steinmetz, A. 1996, The pollen-specific LIM protein PLIM-1 from sunflower binds nucleic acids in vitro, *Sex. Plant Reprod.*, **9**, 264–268.

15. Kawaoka, A., Kaothien, P., Yoshida, K., Endo, S., Yamada, K. and Ebinuma, H. 2000, Functional analysis of tobacco LIM protein Ntlim1 involved in lignin biosynthesis, *Plant J.*, **22**, 289–301.

16. Déjardin, A., Leplé, J.-C., Lesage-Descauses, M.-C., Costa, G. and Pilate, G. 2004, Expressed sequence tags from poplar wood tissues—a comparative analysis from multiple libraries, *Plant Biol.*, **6**, 55–64.

17. Andersson-Gunneras, S., Mellerowicz, E. J., Love, J., et al. 2006, Biosynthesis of cellulose-enriched tension wood in *Populus*: global analysis of transcripts and metabolites identifies biochemical and developmental regulators in secondary wall biosynthesis, *Plant J.*, **45**, 144–165.

18. Tuskan, G. A., DiFazio, S., Jansson, S., et al. 2006, The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray), *Science*, **313**, 1596–1604.

19. Lafarguette, F., Leplé, J.-C., Déjardin, A., et al. 2004, Poplar genes encoding fasciclin-like arabinogalactan proteins are highly expressed in tension wood, *New Phytol.*, **164**, 107–121.

20. Sterky, F., Bhalerao, R. R., Unneberg, P., et al. 2004, A *Populus* EST resource for plant functional genomics, *Proc. Natl. Acad. Sci. USA*, **101**, 13951–13956.

21. Chenna, R., Sugawara, H., Koike, T., et al. 2003, Multiple sequence alignment with the Clustal series of programs, *Nucleic Acids Res.*, **31**, 3497–3500.

22. Jones, D. T., Taylor, W. R. and Thornton, J. M. 1992, The rapid generation of mutation data matrices from protein sequences, *Comput. Appl. Biosci.*, **8**, 275–282.

23. Saitou, N. and Nei, M. 1987, The neighbor-joining method: a new method for reconstructing phylogenetic trees, *Mol. Biol. Evol.*, **4**, 406–425.

24. Guindon, S. and Gascuel, O. 2003, A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood, *Syst. Biol.*, **52**, 696–704.

25. Gascuel, O. 1997, BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data, *Mol. Biol. Evol.*, **14**, 685–695.

26. Page, R. D. M. 1996, Tree View: an application to display phylogenetic trees on personal computers, *Comput. Appl. Biosci.*, **12**, 357–358.

27. Bailey, T. L. and Elkan, C. 1994, Fitting a mixture model by expectation maximization to discover motifs in biopolymers, *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.

28. Ralph, S., Oddy, C., Cooper, D., et al. 2006, Genomics of hybrid poplar (*Populus trichocarpa × deltoides*) interacting with forest tent caterpillars (*Malacosoma disstria*): normalized and full-length cDNA libraries, expressed sequence tags, and a cDNA microarray for the study of insect-induced defences in poplar, *Mol. Ecol.*, **15**, 1275–1297.

29. Djerbi, S., Lindskog, M., Arvestad, L., Sterky, F. and Teeri, T. 2005, The genome sequence of black cottonwood (*Populus trichocarpa*) reveals 18 conserved cellulose synthase (*CesA*) genes, *Planta*, **221**, 739–746.

30. Sterck, L., Rombauts, S., Jansson, S., Sterky, F., Rouze, P. and Van de Peer, Y. 2005, EST data suggest that poplar is an ancient polyploid, *New Phytol.*, **167**, 165–170.

31. Tranbarger, T. J., Al-Ghazi, Y., Muller, B., Teyssendier de la Serve, B., Doumas, P. and Touraine, B. 2003, Transcription factor genes with expression correlated to nitrate-related root plasticity of *Arabidopsis thaliana*, *Plant Cell Environ.*, **26**, 459–469.

32. Kikuchi, S., Satoh, K., Nagata, T., et al. 2003, Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice, *Science*, **301**, 376–379.

33. Sweetman, J., Spurr, C., Eliasson, A., Gass, N., Steinmetz, A. and Twell, D. 2000, Isolation and characterisation of two pollen-specific LIM domain protein cDNAs from Nicotiana tabacum, *Sex. Plant Reprod.*, **12**, 339–345.

34. Laitinen, R. A. E., Immanen, J., Auvinen, P., et al. 2005, Analysis of the floral transcriptome uncovers new regulators of organ determination and gene families related to flower organ differentiation in *Gerbera hybrida* (Asteraceae), *Genome Res.*, **15**, 475–486.

35. Guyon, V. N., Astwood, J. D., Garner, E. C., Dunker, A. K. and Taylor, L. P. 2000, Isolation and characterization of cDNAs expressed in the early stages of flavonol-induced pollen germination in Petunia, *Plant Physiol.*, **123**, 699–710.

36. Allona, I., Quinn, M., Shoop, E., et al. 1998, Analysis of xylem formation in pine by cDNA sequencing, *Proc. Natl. Acad. Sci. USA*, **95**, 9693–9698.

37. Dean Rider, S., Henderson, J. T., Jerome, R. E., Edenberg, H. J., Romero-Severson, J. and Ogas, J. 2003, Coordinate repression of regulators of embryonic identity by *PICKLE* during germination in *Arabidopsis*, *Plant J.*, **35**, 33–43.

38. Gepstein, S., Sabehi, G., Carp, M.-J., et al. 2003, Large-scale identification of leaf senescence-associated genes, *Plant J.*, **36**, 629–642.

39. Becker, J. D., Boavida, L. C., Carneiro, J., Haury, M. and Feijo, J. A. 2003, Transcriptional profiling of *Arabidopsis* tissues reveals the unique characteristics of the pollen transcriptome, *Plant Physiol.*, **133**, 713–725.

40. Yoshida, K., Endo, M., Nakazono, M., et al. 2005, cDNA microarray analysis of gene expression changes during pollination, pollen-tube elongation, fertilization, and early embryogenesis in rice pistils, *Sex. Plant Reprod.*, **17**, 269–275.