COORDINACIÓN DE LA EXPRESIÓN GÉNICA DURANTE LA SÍNTESIS DE LIGNINA EN ÁLAMO (*Populus sp*)

VANESSA VIVIANA CASTRO RODRÍGUEZ

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE CIENCIAS
ESCUELA DE BIOLOGIA
BUCARAMANGA
2007

COORDINACIÓN DE LA EXPRESIÓN GÉNICA DURANTE LA SÍNTESIS DE LIGNINA EN ÁLAMO (*Populus sp*)

VANESSA VIVIANA CASTRO RODRÍGUEZ

Trabajo de grado presentado como requisito para el título de Biólogo

Dr. Ángel García-Gutiérrez Fernández
Director

Dr. Jorge Hernández Torres
Codirector

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE CIENCIAS
ESCUELA DE BIOLOGIA
BUCARAMANGA
2007

AGRADECIMIENTOS

A Ángel García-Gutiérrez Fernández por darme la oportunidad de trabajar a su lado, su apoyo incondicional y sus grandes enseñanzas hicieron que este gran trabajo fuese posible.

A Jorge Hernández Torres, por su apoyo y colaboración en la realización de este proyecto.

A los Drs. Francisco Cánovas Ramos, Francisco Ruiz Cantón y Juan Antonio Pérez Claros de la Universidad de Málaga, España, por su gran colaboración.

A mis padres Martha Cecilia Rodríguez Ramírez y Euripides Castro Rodríguez por su cariño, compresión y confianza.

A mis hermanas Mónika Castro y Lina Marcela Castro por su apoyo durante estos años de mi carrera.

A Isabel Natalia Sierra García, C. Julián Villabona Arenas y Kattia P. Palacio López por su valiosa amistad.

CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

RESUMEN

TITULO: Coordinación de la expresión génica durante la síntesis de lignina en álamo (*Populus sp*)
Autor: Vanessa Viviana Castro Rodríguez **
Palabras claves: *Populus sp*, Lignina, Bases de Datos, EST (Expressed Sequence Tags), Microarrays.

El advenimiento de proyectos genómicos/transcriptómicos estructurales y fisiológicos de plantas maderables ha dado lugar a nuevas posibilidades para estudiar la regulación génica de procesos metabólicos que, hasta el momento, ha sido difícil investigar. Una de las principales áreas de investigación es la biosíntesis de madera, debido a su interés ambiental e industrial. El cinamato, es el recurso de esqueleto carbonado de la ruta de los fenilpropanoides; es sintetizado por la enzima PAL a partir de la deaminación de Phe. El amonio liberado puede ser reciclado por enzimas involucradas en el metabolismo del nitrógeno, como son GS y GOGAT. Los residuos monolignoles de la lignina son sintetizados por actividad de CCoAOMT y AldOMT, enzimas metilantes. El carbono de estos grupos metilo procedería de la acción concertada de las enzimas GDC/SHMT y su transferencia a la ruta de los fenilpropanoides podría realizarse mediante las enzimas del ciclo C1. Este esquema ha sido parcialmente propuesto por varios autores. En el presente trabajo se analiza la estructura y duplicaciones de los genes que codifican estas enzimas mediante la utilización de información genómica del álamo proveniente de JGI. Tales resultados fueron ampliados y contrastados, a partir de información almacenada de PopulusDB para analizar la expresión génica órgano-dependiente y la co-expresión. Nuestros resultados están de acuerdo con un modelo que propone una fuerte coordinación de los genes mencionados en la formación de lignina.

---

[*] Proyecto de Grado

** Facultad de Ciencias. Escuela de Biología. Director: Ángel García-Gutiérrez Fernández. Codirector: Jorge Hernández Torres.

ABSTRACT

TITLE: Poplar genes involved in different metabolic process are expressed coordinately for lignin production
Author: Vanessa Viviana Castro Rodríguez **
Key words: *Populus sp*, Lignin, Database, EST (Expressed Sequence Tags), Microarrays.

The advent of structural and functional genomics/transcriptomics projects in woody plants is expediting new possibilities to study gene regulation of process until now difficult to be approached. One of the principal research subjects is the metabolism of woody formation due to the positive implications in ecological and industrial areas. Cinnamate, the source of carbon skeletons for the phenylpropanoid pathways, is synthetized by the PAL enzyme by deamination of Phe. The ammonium released could be recycled by nitrogen metabolism enzymes, as GS and GOGAT. The precursors of lignin are synthesized by the the CCoAOMT and AldOMT methylating enzymes. These methyl groups, provided by the S-adenosyl methionine (SAM) through the C1 cycle enzymes, would be provided by the concerted action of the GDC/SHMT enzymes. This scheme has already been partially proposed by several authors. In the present work we analyze the structure and duplications of the genes encoding these enzymes using *Populus* genomic information provided by the JGI. Such results were enlarged and contrasted, by gathering data from PopulusDB in order to analyze organ-dependent gene expression and coexpression. Our results agree with a model which proposes a tight coordination of the mentioned genes in the lignin formation.

* Proyecto de Grado.
** Facultad de Ciencias. Escuela de Biología. Director: Ángel García-Gutiérrez Fernández. Codirector: Jorge Hernández Torres.

---

* Graduation Project
** Sciences Faculty. Biology School. Project Director: Ángel García-Gutiérrez Fernández
Co-Director: Jorge Hernández Torres

INTRODUCTION

Woody plants are one of the most ecologic and economical important resources on the earth. The trees are the mainstay of a huge extension of ecosystems that support the natural living conditions of an enormous amount of species, the natural heritage of the planet. The trees have an environmental interest for forest repopulation programs and ground retention, and are essential components of the natural landscape. Furthermore, the forest ecosystems play a crucial paper in the production of the global C, are also responsible for responses to the climatic change, and for the maintenance of the biological diversity. Trees are also a valuable source of materials useful in industry of wood, pulp, paper and other articles of commercial importance (Gallardo et al. 2003). In the last years, a considerable number of tree species has been included in basic and applied research (Merkle and Dean 2000). Working with trees is especially difficult because of their great physical size, technical inconveniences in molecular and biochemical analysis, hardships for genetic transformation and regeneration in vitro, genomes sometimes extensive and long cycle of life.

The wood formation is one of the most characteristic processes that occur during the secondary development of trees. During more than ten years studies related with the structure, function, expression and regulation of genes, and the

corresponding proteins, involved in lignin synthesis have been carried out in woody plants. Lignin is one of the most abundant components of wood, and it is the second most abundant constituent in the biosphere, after cellulose. Lignin provides strength and support to the terrestrial plants, it is a structural part of the plant cell wall providing additional rigidity, it confers hydrophobic capacity to the vascular elements and it acts as a mechanical barrier against pathogen agents (Boerjan et al. 2003).

A characteristic of the lignin is the lack of nitrogen, what is especially important in woody plants. The lignin is a biopolymer composed of aromatic subunits derived of phenylpropanoid metabolites. The amino acid Phe proceeds from the photosynthetic erythrose through the shikimate pathway, and its nitrogen is supplied by glutamate transamination. Phe is deaminated by the enzyme phenylalanine ammonia-lyase (PAL) producing cinnamate, the initial metabolite of the phenylpropanoid pathway from which lignin residues are originated among other compounds (Whetten and Sederoff 1995). During the conversion of cinnamate to lignin, great quantities of methyl groups are introduced in specific positions of monolignols by two enzymes involved specifically in these methylations: Caffeoyl-CoA O-methyltransferase (CCoAOMT) and caffeic acid/5-hydroxyferulic acid O-methyltransferase (AldOMT). References about CCoAOMT and AldOMT in woody species as *Populus, Pinus, Tilia, Acacia, Betula, Acer, Fraxinus,* and *Ostrya* are reported by Li et al. (2006). The source of methyl groups that these enzymes incorporate in the monolignols is the S-adenosylmethionine

(SAM). According to Cantón et al. (2005) the C is transported to SAM through the C1 cycle enzymes 5,10-methylenetetrahydrofolate reductase (MTHFR), 5-Methyltetrahydrofolate-homocysteine methyltransferase (MTR), S-adenosylmethionine synthetase (MAT) and S-adenosylhomocysteine hydrolase (AHCY). The C for this C1 cycle is supplied by the pair glycine decarboxylase complex/serine hydroxymethyltransferase (GDC/SHMT). The GDC complex is composed by 4 different classes of enzymes named P, H, T and L (Douce et al. 2001). The importance of these pathways is not only qualitative, but also quantitative, because 30-40% of the photosynthetic carbon is extracted from the biosynthetic phenylpropanoid pathway (van Heerden et al. 1996).

Nitrogen is a restrictive element for the growth and development of terrestrial plants, the availability of nitrogen in plants constitutes a true challenge for their survival (Lam et al. 1996). The fixation of atmospheric nitrogen to generate nitrate or ammonium on the ground for plant assimilation is an expensive and slow process, and in addition the soil of woody plantations has low pH and is rich in lignin and other residual secondary metabolites that make the nitrification processes restrictive (Cánovas et al. 2006). Different plant metabolic processes liberate ammonium so that the plants have developed different metabolic routes for recycle this compound in order to guarantee the economy of this essential element during their whole life cycle (Suárez et al. 2002). The high quantities of ammonium liberated by PAL during the lignifying process could be reassimilated by the GS/GOGAT cycle (Cantón et al. 2005). The enzyme glutamine synthetase

(GS) incorporates ammonium to Glu synthesizing Gln, while glutamate synthase (GOGAT) transfers amidic nitrogen from the Gln to alpha keto glutarate generating Glu. The amino acids Gln and Glu are the main nitrogen donors to an enormous range of molecules.

One of the most useful resources to carry out molecular studies in woody plants is the information provided by genomic and transcriptomic projects. These sources are being used advantageously in many aspects of structural and physiology molecular research. Projects of this nature are being developed in woody plants like *Populus, Eucaliptus, Citrus, Prunus, Picea* and *Pinus* (Richmond and Somerville 2000; Dong et al. 2005; Li et al. 2006; González-Martínez et al. 2006). The genomic projects have served to enlarge plant genome information, as gene mapping, gene structure and evolutionary studies. In a similar way, transcriptomic studies have been used in species as *Arabidopsis, Oryza, Populus* and new species of commercial interest, advancing in the elaboration of cDNA libraries, and microarrays (Rensink and Buell 2005). These collections have thousands of transcripts of selected organs which facilitate the study of tissue and environmental dependent gene expression, and gene regulation and coexpression.

We are interested in study a possible coordinated system involving genes encoding enzymes directly or indirectly related with lignification, especially C1 cycle and monolignol methylating enzymes. In addition, since PAL is simultaneously the C source of the phenylpropanoid metabolism, and could be

related with the ammonium assimilation by the GS/GOGAT cycle, the study will include these genes. Partial models relating these pathways had already been proposed, so that positive results will aide to understand the regulation of the mentioned process.

1. METHODOLOGY

The gene and protein sequences of the enzyme families GS, GOGAT, PAL, CCoAOMT, AldOMT, MTHFR, MTR, MAT, AHCY, and the H, P, T subunits of GDC were taken from the *Populus trichocarpa* genome database of the Joint Genome Institute, JGI (http://www.jgi.doe.gov/poplar). JGI protein identification numbers were adopted in this paper for nucleic and protein sequences adding “p” to the protein labels.

Protein alignments and trees were performed by using CLUSTALW. Nucleotide alignments among EST and CDS sequences were made by using CLUSTALW and MULTIALIN (Corpet 1998).

The p-distances among the exonic and intronic regions of the genes of each family were calculated by using the phylogenetic program MEGA 3.1 (Kumar et al. 2004). The introns of each gene were united, and the sewed intronic sequences were used to obtain the p-distances between the members of each family.

The *Populus* EST sequences of the enzyme families GS, GOGAT, PAL, CCoAOMT, AldOMT, MTHFR, MTR, MAT, AHCY, and the H, P, T subunits of GDC were obtained from the PopulusDB database of the Umea Plant Science

Centre (UPSC, http://www.*Populus*.db.umu.se). The clone quantity for different organs and tissues were equally obtained. These data were processed to obtain values representing the gene expression considering the library sizes.

Data for correlation were obtained from the experiment 7, “Global profile of wood-forming tissues” (Schrader et al. 2004; Hertzberg et al. 2001) as is presented in PopulusDB microarray database. This experiment analyze the gene expression of different progressive tissue cryosections covering from cambium to mature xylem cells of *P. tremula* X *tremuloides* stem, consisting the different samples in the transcripts prepared from the different sections. Expression values of EST clones were collected from all the samples and were organized in matrix of vertical columns each one with the data of each gene, so that the rows contained values corresponding to the same sample. Correlations analysis, p-values and principal component analysis (PCA) were calculated by introducing this matrix in the Minitab® 15.1.0.0 version of Minitab Inc, State College, Pennsylvania (USA).

The gene names used in the electronic northern and coexpression analysis were the corresponding protein identification numbers of the orthologous genes of the JGI *Populus* genome database, in order to facilitate the access to the gene structures. A correspondence among these numbers and the EST terminology utilized by PopulusDB is presented in the table 4. It is also showed the PU nomenclature used by PopulusDB for the EST clones of the microarrays experiments.

2. RESULTS AND DISCUSSION

The first tree genome completely sequenced has been poplar (*Populus trichocarpa*), a project of the Joint Genome Institute (JGI). This species has been selected by the reduced size of its genome, 4 times smaller than the human, and 50 times smaller than that of the pine, for which reason poplar is a perfect model for genomic studies in trees. On the other hand the PopulusDB database of the Umea Plant Science Centers (UPSC) stores EST collections from libraries constructed from different *Populus tremula* X *Populus tremuloides* tissues and organs (Sterky et al. 2004). The clones have been very useful to perform an excellent collection of microarray, which are stored in this database.

Poplars are paleopolyploid organisms, and recent genomic studies suggest that their genomes have suffered duplications along their evolution (Sterck et al. 2005). The members of the genus *Populus* are angiosperms of high growth rate, capable to utilize systems of vegetative propagation and also very valuable as transgene receptors (Gallardo et al. 2003). For these reasons poplar is considered an ideal model to carry out molecular studies in trees (Wullschleger et al. 2002). These plants produce a soft wood rich in cellulose, useful for pulp elaboration, paper, and wood derivatives as firewood and industrial wood because of its relative content in lignin. These characteristics

make *Populus* an excellent model for commercial perspectives. Poplar massive production would help to safeguard other forest species of ecological interest.

A general scheme associating the previous related metabolic process (see introduction) has been proposed by Cantón et al. (2005). We selected a work strategy based in obtaining structural genomic information about the referred genes and proteins before starting expression data processing. These data were carefully evaluated since the poplar genome is duplicated so that it was necessary to define faithfully the duplicated genes in order to handle this information in later studies, especially expression analysis. It is know that duplicated genes can diverge to acquire novel and independent functions, and it was necessary to have accurately information on duplication to assess the physiological diversity of the poplar genes. The hypothetically coexpressed genes would participate in a general cellular process geared by means of the combined action of interconnected metabolically pathways, and which main goal would be the lignin biosynthesis in woody living tissues (figure 1).

2.1 GENE DUPLICATIONS

In order to search candidate duplicated genes, the information of each family was independently analyzed as follows: Truncated or putative pseudogenes, and sequences that not clearly corresponded to the family were discarded. Protein selected sequences of each family were aligned obtaining the respective trees

(figure 2). The p-distance among the CDS sequences and intronic regions of the members of each family was calculated.

GS.- The GS family present 18 models, 9 are truncated sequences which then were not considered in the analysis, and 8 are entire. The last model p-589509 was not analyzed because its sequence scarcely seemed a typical GS but a considered nodulin protein found in fungi and plants and also present in archaea GS, although it had only regional and partial similitude. The alignment among the selected GS protein sequences produced 2 branches; the first had two GS2, and the other had 3 groups each one having two GS1 proteins (figure 2a). When the p-distances among the genes were calculated, the results were confirmed. The minimal distances corresponded with the pairs obtained by the protein alignment.

GOGAT.- The *Populus* genome have 10 GOGAT models, but only 4 are not truncated. These produced 2 branches one had a pair of Fd-GOGAT and the other a pair of NADH-GOGAT proteins (figure 2b). The sequence p-Fd-GOGAT-734780 does not present the 137 amino acids of the N amino terminal, because the corresponding nucleotidic sequence of the first exon has been considered part of the 3´-UTR region. A defective nucleotide in this region break the ORF, so that it has been selected a posterior first ATG. The Fd-GOGAT sequences presented more than 95% of similitude so that it is possible that the mutation is circumstantial and the protein is functional. In fact some *P. tremula* X *tremuloides* ESTs corresponding to this gene have been identified in the PopulusDB database. On the other hand 105 nucleotides of the NADH-GOGAT-700801 sequence have

been included in the 3´-UTR region lacking the corresponding 35 amino acids of the protein, which then were incorporated for the analysis. Once these corrections were made the tree agreed with the calculated p-distances.

PAL: The family have 6 genes one of them truncated. The alignment showed a region of 38 amino acids in p-PAL-82117 that not appeared in the plant PAL proteins. The PAL-82117 gene is probably originated by an internal duplication of PAL-725007 since the 2 sequences are 98.5% identical and the two genes are separated by an extension of about 13,000 bases. If PAL-82117 is not functional because of this insertion, it would explain that none corresponding EST was found in the PopulusDB database. The tree had 2 groups, with 2 and 3 sequences respectively (figure 2c), and according the p-values.

CCoAOMT: The alignment of the 5 protein sequences found in the genome database produced a tree with 3 branches, 2 of them had 2 paired sequences each one (figure 2d) agreeing the p-values.

AldOMT: When the AldOMT CDS sequences were previously aligned, AldOMT-412253 presented a defective base in position 20, although the amino acidic identity with the more proximal sequence was 97.2%. No ESTs corresponding to this gene were identified in the PopulusDB database, however it was provisionally added the corresponding A (in accordance with the paired sequence) in this position, so reconstructing the broken ORF. The alignment of the protein sequences produced 5 branches, 3 of them had only 1 sequence, and the other 2 branches had 2 sequences each (figure 2e). The p-distances confirmed the results.

MTHFR: Poplar have only two MTHFR genes, and the amino acid sequences aligned in high degree.

MTR: 5 models are present in the poplar genome, but one is a truncated gene. Protein alignment produced 2 branches, having 2 genes each one, and p-distances agreed these results (figure 2f).

AHCY: The *Populus* genome contain 4 AHCY models, but the AHCY-108722 and AHCY-108698 genes are identical to that of the *Ralstonia metallidurans* bacteria. When the protein sequences were examined in detail it was clear that corresponded respectively to the N-terminal and C-terminal regions of the bacterial protein, and in addition the two poplar proteins had large N-t and C-t extensions that do not exist in the bacteria. The poplar sequences were internally truncated, regarding the bacterial sequence, in an extension of more than 300 amino acids, presumably because they occupied almost the whole content of very short scaffolds. A possibility is the two sequences are united by unknown genomic information and consequently they are not really truncated, being the gene entire. The other 2 poplar AHCY proteins showed high identity with other plant AHCY (figure 2g).

MAT: There are 6 MAT genes in *Populus*, and the tree produced 2 groups each one with a pair of protein sequences, and two other sequences impaired. Nucleotidic alignment produced comparable results (figure 2h).

GDC: *Populus* have 2 GDC-T sequences and 2 GDC-P sequences, but for GDC-H there are 4 sequences, that the alignment grouped perfectly in two pairs and according the p-distances (figure 2i).

Once defined candidate duplicated genes, these were positioned on the corresponding poplar linkage groups (LG) or scaffolds. To check the duplications, the collinearity between huge segments surrounding each pair of hypothetic duplicated genes was analyzed. Localization of genes was carried out by tracking thousands of nucleotides upstream and downstream the candidate genes, and then the CDS sequences obtained from regions around the hypothetic duplicated genes were aligned between them. In this way it was made a general analysis by crossed alignments among the regional genes that allow us to find sequential identities and collinearity; the results confirmed all the hypothesized duplications. In many cases the selection of sequences to align was eased by indications provided by the genomic database about the character of the genes found. The sequence orientation of each pair of studied duplicated genes was conserved in all the cases analyzed. However some collinear duplicated genes were in inverted positions (figure 3 and table 1) and in addition other not duplicated coding sequences intercalated among the duplicated collinear genes suggest that chromosomal events had happened subsequently to the genome duplication.

2.2 *P. tremula* X *tremuloides* ESTs, AND DUPLICATED *P. trichocarpa* GENES

The PopulusDB database contains 102,019 ESTs compiled from 19 libraries of different *Populus* organs (Sterky et al. 2004). 17,727 ESTs are from *P. trichocarpa* and the rest from *P. tremula* and *P. tremula* X *tremuloides*. This collection is distributed in clusters, containing each one contigs which enclose ESTs used to

build consensus sequences. However the imbricated ESTs of each consensus could correspond to different genes, often duplications, so that the correspondence of each EST with the *P. trichocarpa* gene is also reported in the database. The existence of gene duplications in these 2 species generated interesting questions when the *P. tremula* X *tremuloides* ESTs and the *P. trichocarpa* genes were aligned, since it was obtained special information to treat questions about paralogous and orthologous genes in the context of the poplar genome duplication phenomenon.

The *P. tremula* X *tremuloides* sequences are not identical to the correspondent of *P. trichocarpa*, so that when an alignment among two *P. trichocarpa* duplicated genes and a *P. tremula* X *tremuloides* EST is made, this last present more identity with one of the genes, indicating this is the orthologous with the *P. tremula* X *tremuloides* gene. However the other duplicated gene is also very similar, having often nucleotides that appear specifically only in the EST. Furthermore is also possible to find specific common nucleotides between the duplicated genes that do not appear in the EST. Poplar genome duplication has been examined by Sterck et al. (2005), whom concluded that the event happened in an original species, which explain that the duplication is present in all the actual *Populus* species. In fact, when the previous described alignment is made, the distance between the two *P. trichocarpa* duplicated genes of an organism is greater than the distance between the *P. tremula X tremuloides* EST with the *P. trichocarpa* orthologous gene. A probable explanation is presented in the figure 4. After the

duplication, the two genes would paralogously diverge by accumulation of differences in the sequence. This paralogy can seem indeterminate, at least in comparison with others really paralogous and more distant members of the family, however the term is technically used in this case (the subject is discussed by Koonin 2001, 2005). Once the speciation befall, the two species, that inherited the same genomic content i.e. the two paralogous genes, start independent evolutionary ways, so that the genes continue undergoing changes in different manner. A comparison between nucleotidic sequences will prove a greater identity between orthologous when are compared with its duplicated paralogous due to the changes occurred on the two duplicated paralogous before the speciation, and moreover, changes after the speciation also explain common specific nucleotides between the ortologues when comparing with the paralogous sequences if the changes occur on these lasts. Nevertheless, changes on a gene after duplication could explain the corresponding specific nucleotidic identities between the two paralogous of the other organism, as well as corresponding specific identities between its paralogue and its ortologue. In these two cases it is necessary to suppose that these changes happen on nucleotides that had not previously suffered any mutation since the original state in the primitive species.

2.3 ORGAN-DEPENDENT EXPRESSION AND FUNCTIONAL DIVERGENCE

PopulusDB reports how many ESTs, i.e. clones corresponding to a determined gene appear in each organ or tissue, so that a scheme grouping all the genes under study was constructed in order to compare the representation of different

genes in each organ and keeping in mind the sizes of the libraries (table 2). This diagram allows extracting several conclusions concerning the functional specialization of divergent duplicated genes.

Organs containing a great deal of autotrophic tissues as young leaves presented an expected profile, a typical GS/GOGAT cycle is detected. GS proteins have two models in plants, chloroplastidic (GS2) and cytosolic enzymes (GS1) (Lam et al. 1996). The GS2-725770 presented an expression 5 times more than GS2-565302 in young leaves, however the expression of this last gene was about 3.5 fives greater in virus/fungal-infected leaves than in young leaves noticing a distribution of roles (see Rapp et al. 2005 about acquisition of novel functions in duplicated genes). The *P. trichocarpa* genes were structurally different, at least in the 3'UTR region, where the GS2-565302 gene presented an intercalated fragment bigger than 2,000 bases. Some trails indicate that the orthologous EST have also this region, suggesting that the supposed insertion happened before the speciation. Differential majority expression of the GS2-725770 gene was also present in tension wood, floral buds and female catkins. If this gene is preferentially used, being GS2-565302 reserved for emergency situations it will be a matter of research. The duplicated GS1-710427, GS1-648144 and the GS1-562842 cytosolic genes were also found in comparable levels in infected leaves, suggesting a distributive specialization of several members of the family for pathogen attacks. A glance to table 2 let discover other genes involved specifically in this task, and equally pointing alternative functions of duplicated genes.

GS1-648144 was the only GS1 gene expressed in leaves, but its duplicated GS1-710427 appeared specifically in roots inverting the situation and also representing a possible divergence of functions; however the expression is comparable in others organs. When the expression in leaves of these two last genes was compared according the results of the experiment 13, available in PopulusDB "Optimization of experimental design in microarray analysis" (Bylesjö et al. 2005), was verified that the gene GS1-648144 have expression levels 10 times higher than GS1-710427. There are not ESTs corresponding to the GS1-685196 gene in young leaves, and its expression was also very low in the microarray agreeing the electronic northern. On the other hand the other two pairs of duplicated genes (GS1-685196, GS1-562842 and GS1-569354, GS1-580657) also presented functional divergence.

There are two forms of GOGAT, Ferredoxin and NADH dependent. The first one is located in chloroplasts (García-Gutiérrez et al. 1995) working in photosynthetic tissues meanwhile NADH-GOGAT has been located in heterotrophic tissues as flower buds and nodules of alfalfa (Schoenbeck et al. 2000). Fd-GOGAT-734780 was also present in poplar young leaves, although in lower levels (this could be due to the lability of these transcripts) and interestingly the FGOGAT-695596, that did not appear in young leaves, was much more expressed in infected leaves instead of its duplicated gene. The two Fd-GOGAT duplicated genes showed an organ-dependent distribution of roles. A coordinated program against pathogen

attacks could be launched using selected GS and Fd-GOGAT isogenes, presumably for N recycling from infected cells (Pérez-García et al. 1998a, 1998b). NADH-GOGAT gene is not expressed in photosynthetic organs but in wood cell death and roots where Fd-GOGAT transcripts are not detected (a revision of GS and GOGAT in woody plants is reported by Suárez et al. 2002).

Tissue-dependent expression of PAL (Osakabe et al. 1995) and variety of PAL roles have already been reported by Kao et al. (2002) whom suggest that the expression of a gene in woody tissues does not exclude different metabolic roles in other tissues. According the electronic northern the 4 poplar PAL isogenes seem to be involved in lignification, especially PAL-696959 but the differential apparition in other organs agree with the idea of an amply distribution of functions. PAL-725007 seems to be poorly used, contrary to its duplicated PAL-696959. Differential expression was also notorious between the pair AldOMT-345776 and AldOMT-57441 being this last scarcely expressed. The generation of S lignin in woody poplar is carried out by AldOMT (Jouanin et al. 2000) so that the expression of this gene in tension wood correspond the highly specific ratio of syringyl (S) respect guaiacyl (G) residues typical of Fraxinus tension wood G layers (Prodhan et al.,1995). *Populus kitakamiensis* have 2 AldOMT isogenes differentially expressed in stem and leaves (Hayakawa et al. 1996). We prove they are duplicated genes, corresponding respectively with their orthologous AldOMT-345776 and AldOMT-57441 and according the expression of AldOMT-345776 in tension wood. The isogenes of the other methylating enzyme involved in lignin biosynthesis,

CCoAOMT also showed a differential organ-specific profile, as it is evident when the stronger transcript representation of CCoAOMT-691730 was compared with its duplicated CCoAOMT-722225 in the lignified organs tension wood and wood cell death, in female catkins and even petioles. CCoAOMT would be less implicated in methylation conducting to S-lignin, working preferentially for the production of G-lignin in poplar (Meng et al. 1998) and alfalfa (Guo et al. 2001; Chen et al. 2006).

There seems to be a different physiological evolution for the duplicated GDC-H-570626 and GDC-H-88775 genes because this last one was preferably expressed. A similar situation was observed for the duplicated GDC-T-833954 and GDC-T-828778. The differential levels of MTR, MTHFR, AHCY and MAT genes in tension wood was patent, suggesting that selected C1 cycle genes are expressed in this organ. A coordinated expression of these genes with the above analyzed CCoAOMT and AldOMT would support the proposed supply of methyl groups to these two enzymes through the C1 pathway (Cantón et al. 2005). Regarding AHCY, the isogene AHCY-597072 was in whole much minor represented than its duplicated AHCY-549785 in many organs, suggesting this last gene is preferentially used. This result is similar to the one described by Pereira et al. (2007) in *Arabidopsis thaliana* where there are two isogenes, one of them expressing itself notably in many organs, whereas the other is very low represented. MTR-594861 was also much more differently and preferably expressed in tension wood and cambial zone than its duplicated MTR-649045 isogene. However in wood cell death the levels of this last increased and

surpassed the expression of the first one. The transcripts of MTHFR-654300 are the most abundant in tension wood, also suggesting divergent physiological roles. Finally, different functionality for one MAT gene with respect to other members of the family has been reported by Shen et al. (2002) in *Arabidopsis thaliana*; similarly, a differential expression organ-dependent in this species is cited by Peleman et al. (1998), who found that the expression of a MAT isogene in roots and stems was stronger than in leaves. References about tissue-dependence expression are reported by Schröder et al. (1997) observing tissue-dependence and functional specialization for MAT isogenes of *Catharanthus roseus*. In a similar way, the duplicated poplar MAT-552593, and MAT-572548 pair also presented a differential expression in organs, and even also the MAT-566909 and MAT-564333 duplicated genes.

2.4 GENE COEXPRESSION

PopulusDB have a collection of microarrays prepared with an extensive and representative selection of clones from its proper database. The results as normalized data set of gene expression are accessible for the public, so that they can be directly used for different purposes. We decided to examine the coexpression of the studied *Populus* genes utilizing data from the experiment 7 because it contained information enough about these genes, and in addition the samples proceeded from poplar tissues involved in lignification, concretely sections from cambium to xylem regions of poplar stems. It was decided to use a

biostatistical strategy to examine the correlation among them, and although this class of calculation does not allow to know the expressions level of the genes in each tissue, it is expected to access to other class of valuable information, the coexpression through the different cellular types. Although the lignin production rate could be different in each tissue, all of them are manufacturing this compound so that it is expected to find some degree of correlation of the involved genes.

The correlation coefficients were calculated among all the genes listed in the table 3. In many cases different cDNA clones representing the same *P. trichocarpa* gene had been used, obtaining very coincident values. This allow us to check that the correlations calculated were correct in practically all the cases, and so positively assessing the viability of the original experiment for our aim. In all these cases a statistical mean among the microarray expression data corresponding to different ESTs representing the same gene were calculated; the new values were introduced substituting the previous ones to make a new general correlation calculation and so summarizing favorably the matrix (table 3). The explicitness of the results and the general viability of the p-values allowed us to perform a principal component analysis from the correlation matrix in order to reduce the dimensionality of the data set and to present a simplified scheme of the gene coexpression (figure 5). According the PCA, practically all the genes appeared tightly close in the first component suggesting that these genes as a whole are coordinately involved in the common process of lignification.

The lignin content of tobacco and poplar (Zhong et al. 1998, 2000) decreased notably when these plants were transformed with antisense constructions of AldOMT and CCoAOMT, confirming the transcendent participation of these genes. AldOMT and all the CCoAOMT poplar genes showed very high correlation values meaning a strong coexpression during the lignin production (These enzymes are reviewed by Boerjan et al. 2003 and Li et al. 2006). The signal intensity is increasing from the phloem through cambial tissues to the xylematic cells, where the expression is maximal for these 2 genes in the original microarray from which these data were obtained (Hertzberg et al. 2001). Sterky et al. (1998) had already obtained very similar results by scrutinizing clones from poplar EST libraries: The ratio of gene expression is maintained meanwhile the signals increase concomitantly from cambial cells through xylem. These conclusions combined with the correlation values suggest that the coexpression of the 3 CCoAOMT genes, and the AldOMT gene, occur by intensifying coordinately their expression along these tissues, irrespective their specific roles in the lignin biosynthesis. Poplar CCoAOMT is mainly involved in the synthesis of wood G units (Meng et al. 1998) whereas AldOMT would be responsible for the production of the S residues. A review about lignin composition is reported by Baucher et al. (1998). The similar high correlation values of the different poplar CCoAOMT isogenes hinder to specify the role of each one in the formation of lignin through the woody region.

All the C1 cycle genes, MTR, MAT, AHCY, and MTHFR showed very high correlation values among them. It is notice that although some genes as AHCY-540785 and MTR-73850 had low p-values, it is not easily disposable the

participation of anyone. The C1 cycle could supply methyl groups to the lignifying enzymes since many members of each family presented high correlation values and reliable p-values with CCoAOMT and AldOMT. The association of C1 with the methylating enzymes seems to occur as much in herbaceous as in woody plants. A proteomic study of the AldOMT, CCoAOMT and C1 cycle enzymes also report this class of correlation in lignifying cells of maize leaves (Vincent et al. 2005), specifying the implication of several isoforms of MAT and MTR. In tomato, MAT accumulates mainly in lignified organs, especially under saline stress conditions contributing to create rigidity by lignification of vessels (Sánchez-Aguayo et al. 2004). Shen et al. (2002) report decreasing in the lignin content in *Arabidopsis* plants with inactive mutated MAT, and in this species Peleman et al. (1998) have also cited a correlation between lignification and MAT expression. The provision of methyl groups in lignin formation in woody angiosperms and gymnosperms has been reported by Sterky et al. (1998) who found high abundance of MAT ESTs in poplar libraries of cambial and xylematic cells. GUS activity conducted by a MAT promoter was found around cambial zone of poplar, and MAT expression in these tissues was determined by using southern blot (Vander Mijnsbrugge et al. 1996). MAT, MTR, and AHCY transcripts were also present in pine secondary xylem libraries (Whetten et al. 2001). Based in the results of an eucalyptus microarray Kirst et al. (2004) present a scheme explaining the coordinated expression among AldOMT, CCoAOMT and the C1 cycle MTR, MAT and AHCY enzymes.

Glycine decarboxylase GDC-P-685558 and GDC-T-828778 genes presented high positive correlation values between them, but GDC-H-570626 had a value of -0.71 and -0.84 with each one indicating negative or inverse correlation although the p-values not showed confidence; indeed the PCA placed the GDC-H-570626 gene in the principal component region but in an inverted orientation; in this way other GDC-H isogene had to be involved in the complex with GDC-P-685558 and GDC-T-828778 genes, however no other GDC-H EST were used in the microarray. Rajinikanth et al. (2007) detected transcripts of PtgdcH1 (which correspond to GDC-H-570626) in xylematic tissues of poplar, and the authors also report the appearance of PtgdcH2 (GDC-H-88775) gene in this tissue. Although PtgdcH1 had higher expression than PtgdcH2, this last transcript could be the candidate encoding the H protein for the GDC complex in association with p-GDC-P-685558 and p-GDC-T-828778. These genes showed high correlation with the C1 cycle genes, MTR, MTHFR, MAT and AHCY, and with AldOMT, and CCoAOMT genes suggesting that the GDC complex could supply C atoms to an active C1 cycle implicated in lignification. The enzyme that would connect GDC with the cycle C1 is SHMT (Douce et al. 2001) which will be coupled by means THF cofactors (Mouillon et al. 1999), but unfortunately clones of this gene were not used in this experiment. The GDC complex is provided with metabolites derived from photorespiration in photosynthetic cells, where the C1 cycle seems to be linked to this activity by the GDC/SHMT system (Hanson and Roje 2001). However in non-photosynthetic organs the C metabolic source it is not known. A diagram showing a possible coordination between methyl apportion from SHMT and the C1 pathway

is presented by Vander Mijnsbrugge et al. (2000) who found by means of bidimensional electrophoresis that GDC is one of the most abundant proteins in differential poplar xylematic cells, where SHMT, AldOMT, and CCoAOMT are also preferentially present. SHMT transcripts have also been identified in pine secondary xylem (Whetten et al. 2001).

There are several reports about functional diversity of PAL isoenzymes in non woody organs. PAL protein and transcripts are located in different plant organs as leaves, stems, flowers and roots of *Nicotiana* (Fukasawa-Akada et al. 1996), and in young leaves, stems and petals of *Populus* (Subrarnaniam et al. 1993). Differential expression of PAL isogenes involved respectively in fruit ripening and flower development, has been also detected in Raspberry (Kumar and Ellis 2001). On the other hand a summary about PAL in woody forming tissues is reported by Li et al. (2006). PAL supply the carbon skeletons to the phenylpropanoid pathway from which are produced biomolecules with a wide range of functions; for example in aspen has been cited lignins, tannins and salicylate-derived phenolic glycosides (Kao et al. 2002). PAL isogenes show differential functionality in photosynthetic and woody tissues of aspen (Kao et al. 2002), and poplar tissue-dependent expression is also cited by Osakabe et al. (1995). The coexpression of PAL with AldOMT, and CCoAOMT was observed by detecting their concomitant augment from poplar cambial zone along the xylematic tissues (Hertzberg et al. 2001). PAL protein has been immunolocated in xylematic cells of poplar (Osakabe et al. 1996) and even similar methodology allows detecting its subcellular emplacement in

these tissues (Sato et al. 2004). Our correlation values agree this observation: The two PAL-725007 and PAL-696959 genes showed high correlation values with all the AldOMT and CCoAOMT, however only PAL-696959 had significant p-values and similar results were obtained regarding GDC and C1 cycle genes. This suggests that PAL-696959 is especially associated with the lignification supporting the proposed role of PAL providing material to the monolignol pathway and a differential involvement of each PAL isoenzyme in distinct process. The differential expression of PAL isogenes is expectable since lignin is not the only objective of the phenylpropanoid metabolism. In this respect, PAL promoter constructions direct GUS activity in distinctive expression models suggesting temporal and tissue distribution of PAL isoenzymes during the lignifying development of different poplar cells (Gray-Mitsumune et al. 1999). A complete vision of the process commented is reported by Prassinos et al. (2005) who amplified by AFLP technique the cDNA of 6 poplar tissues taken from different stems regions, and sequencing then the genes preferentially expressed. The resulting profile was a very similar pattern of progressive abundance of COMT, AldOMT, PAL, C1 cycle, and SHMT genes from cambial to xylem tissues, relating the activity of PAL supplying carbon skeletons for phenylpropanoid metabolites with the entry of C through C1 cycle for the methylation of monolignols.

The ammonium released by the PAL activity must be metabolically reassimilated since the plant economy does not permit wasting N. Ammonium recycling pathways are the GS/GOGAT cycle in angiosperms (Lea et al. 1990; Lam et al.

1996), and gymnosperms (Cánovas et al. 1998), and even the proposal GS/AS cycle (Ávila et al. 2001; Suárez et al. 2002). Razal et al. (1996) confirmed by spectroscopic analysis the sequential incorporation of N released by the PAL activity on Gln and GLu in potato discs. Moreover, these authors refer that the reassimilated ammonium was found in Phe in lignifying cells, indicating not only an active GS/GOGAT cycle, but a concatenated coordination with prephenate aminotransferase to reincorporate N from Glu to arogenate and then to Phe. Very similar model has been proposed by van Heerden et al. (1996) when they measured, by using a HPLC technique, the incorporation of N in Phe in cultured cells of Pinus. Not any GOGAT was used in the microarray of the experiment 7 nevertheless were used 2 GS1 clones. When correlations were calculated the GS1-580657 gene had good values with CCoAOMT, AldOMT-345776 and with some C1 cycle genes, being reliable in many cases according the p-values. However the correlation with PAL and GDC was moderate and the p-values gave low confidence. In fact this gene appears proximal to the whole of the lignifying genes in the PCA scheme although not so closed. These results suggest that although GS1-580657 could participate in lignification, its possible simultaneous participation in other different processes would disturb a correlation and PCA profile pointing exclusively a lignification role. Other GS1 genes not used in the microarray could be also partial or completely expressed, in which case the ammonium recycling would be a shared work which distribution is a matter of research. The GS1-710427 gene had much minor correlation values and a low confidence with the other genes than GS1-580657, and the PCA emplaced it to

the second component. This is also the case for GS2-725770 which presented a clear orientation in the second component but inverted with respect to GS1-710427, which was expected since GS2 encode a chloroplastidic enzyme.

3. CONCLUSIONS

PopulusDB and *Populus* genome databases are extraordinary resources to study molecular physiology of woody plants. The strategy to investigate the coordination of the process involved in lignin formation was to analyze the expression of genes participating in different pathways, which could be implicated in metabolic process related with this task. The selected genes were GDC complex protein, C1 cycle enzymes, GS and GOGAT, PAL, and CCoAOMT and AldOMT. Structural information extracted from *Populus trichocarpa* genomic database proved previous references reporting genome polyploidization, since the majority of the genes studied were duplicated. Gene duplications was recognized when screening ample genomic regions where the candidate duplicated genes were emplaced, other variable number of collinear duplicated genes were found; it is noticed that this could be an ancient event because several not duplicated genes were also found in these regions, presumably because chromosomal reorganizations occurred after the duplication took place. The opportunity to have also transcripts from *P. tremula* X *tremuloides* genes allowed to find the most probable *P. trichocarpa* orthologous genes with those from the hybrid. An investigation about the aligned regions found sequential traits understandable as a sequential order of mutation events through the duplication and speciation events, implying an evolutionary model in which the orthologous gene variability could be compared

with the structural paralogous gene divergence. The functional divergence was examined by means of an electronic northern presenting the organ-expression of the duplicated genes, concluding in numerous cases of function distribution. The statistical analyses of microarray expression data showed a strong correlation of a great majority of these genes, suggesting a coordinated contribution for the lignin synthesis, and a possible expression model in which several related metabolic pathways would be subordinated to this finality.

TABLE 1. Duplicated genes

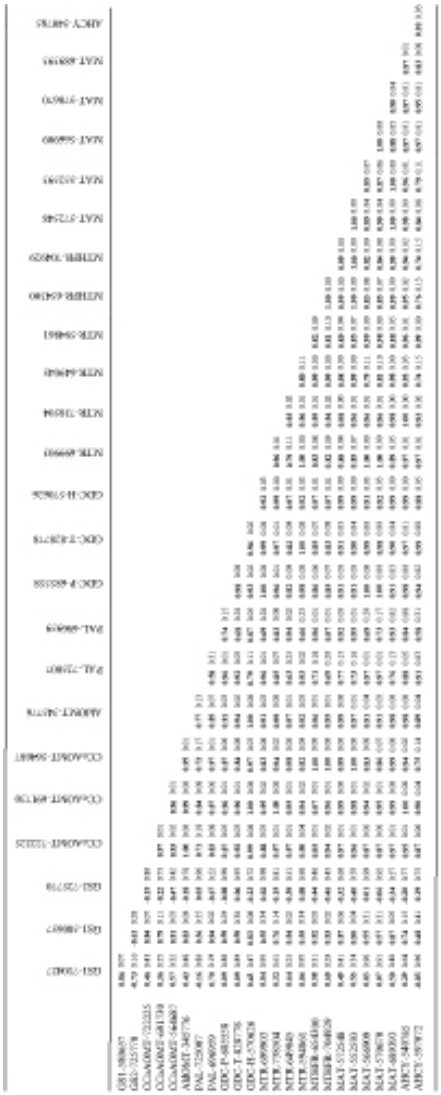
Duplications	Location	Bases screened		Collinear genes	
		D	U	D	U
GS1-685196	LG V	60,000	60,000	4	0
GS1-562842	LG VII	65,000	65,000		
GS1-710427	LG II	120,000	60,000	0	4
GS1-648144	LG IV	60,000	160,000		
GS1-569351	LG XII	140,000	60,000	0	2
GS1-580657	Scaffold 122	20,000	17,000		
GS2-565302	LG VIII	70,000	40,000	0	2
GS2-725770	LG X	40,000	110,000		
F8-GOGAT-695596	LG VI	200,225	210,808	7	2
F8-GOGAT-734780	LG XVI	190,076	171,918		
NADH-GOGAT-700801	LG XV	46,154	76,392	1	0
NADH-GOGAT-594762	Scaffold 662	17,328	111,453		
PAL-696959	LG VIII	187,589	92,562	10	12
PAL-725007	LG X	111,118	135,846		13
PAL-667815	LG XVI	217,849	244,952	5	2
PAL-739482	Scaffold 28	171,822	207,555		
CCoAOMT-564687	LG VIII	140,728	168,881	9	11
CCoAOMT-55191	LG X	168,096	125,304		
CCoAOMT-691730	LG I	175,689	180,418	11	2
CCoAOMT-722225	LG IX	107,508	135,813		
AMOMT-58062	LG XI	175,828	150,011	3	0
AMOMT-257389	LG XVI	155,565	128,544	4	
AMOMT-345776	LG XII	185,174	101,792	3	8
AMOMT-57441	LG XV	80,414	133,884		
MTHFR-654300	LG VII	1,763	122,425	0	12
MTHFR-704929	Scaffold 170	9,352	110,306		
MTR-649045	LG IX	103,161	84,203	3	3
MTR-594861	Scaffold 66	100,310	155,751		
MTR-699903	LG XIII	150,400	151,692	2	3
MTR-738504	LG XIX	230,497	186,265		
AHCY-549785	LG I	222,160	144,621	3	9
AHCY-597072	Scaffold 88	150,242	261,001		
MAT-552593	LG II	148,740	214,693	8	7
MAT-572548	LG XVI	110,434	117,259		
MAT-564333	LG VIII	138,381	132,917	2	6
MAT-566909	LG X	171,082	113,789	3	
GDC-H-570626	LG XII	88,552	136,783	7	2
GDC-H-88775	LG XV	142,334	110,945		
GDC-H-676524	Scaffold 29	149,030	178,199	7	7
GDC-H-580100	Scaffold 118	183,276	150,735		
GDC-P-685558	LG VI	169,615	158,250	7	3
GDC-P-577945	LG XVIII	183,401	233,536		
GDC-T-833954	LG XI	28,201	113,608	1	2
GDC-T-828778	Scaffold 252	30,193	82,337		

Columns from left to right: Duplicated genes, LG or scaffolds where the studied genes were emplaced, genome extensions inspected, and number of collinear duplicated genes found out. Double numeration means internal duplications. D: Downstream; U: Upstream.

TABLE 2. Genes expression

The left column indicate organs used to build the PopulusDB EST libraries. Protein identification numbers used by the corresponding orthologous genes of the JGI poplar genome were used to label the PopulusDB genes (top row). Table values were taken from the PopulusDB and represent the relative EST amounts according the library sizes. The correspondence of *P. trichocarpa* genes with PopulusDB EST clones is presented in the table 4.

TABLE 3. Correlation matrix

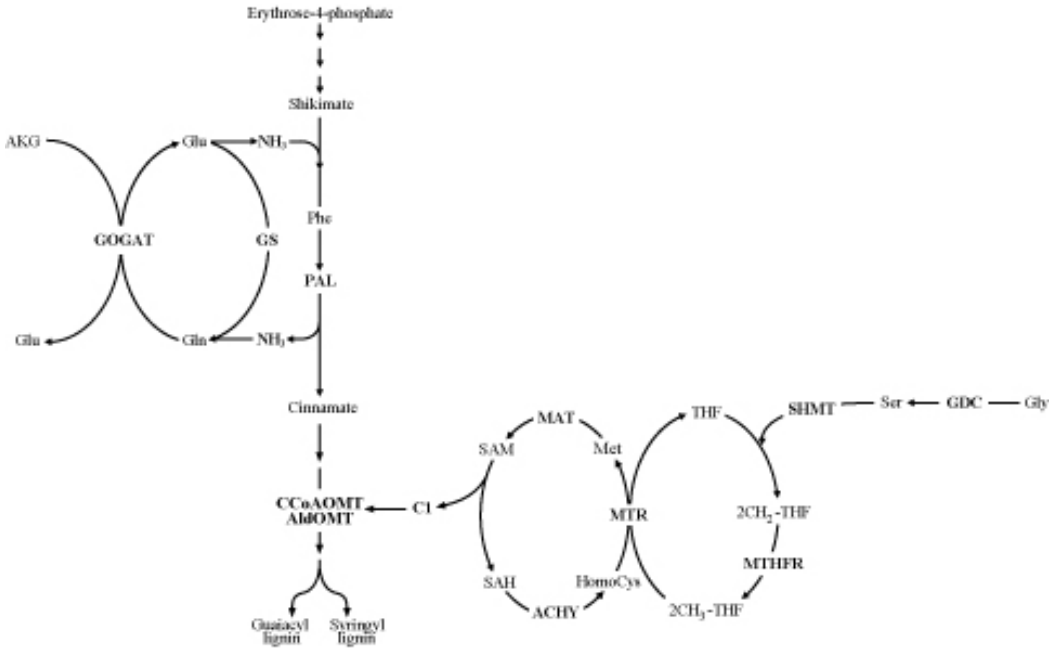


Bold characters represent correlation values. Regular characters are the p-values; p-values < 0.009 are written as 0.00. Protein identification numbers used by the corresponding orthologous genes of the JGI poplar genome were used to label the PopulusDB genes. The correspondence of *P. trichocarpa* genes with PopulusDB EST clones is presented in the table 4.

TABLE 4. Correspondence among JGI and PopulusDB genes.

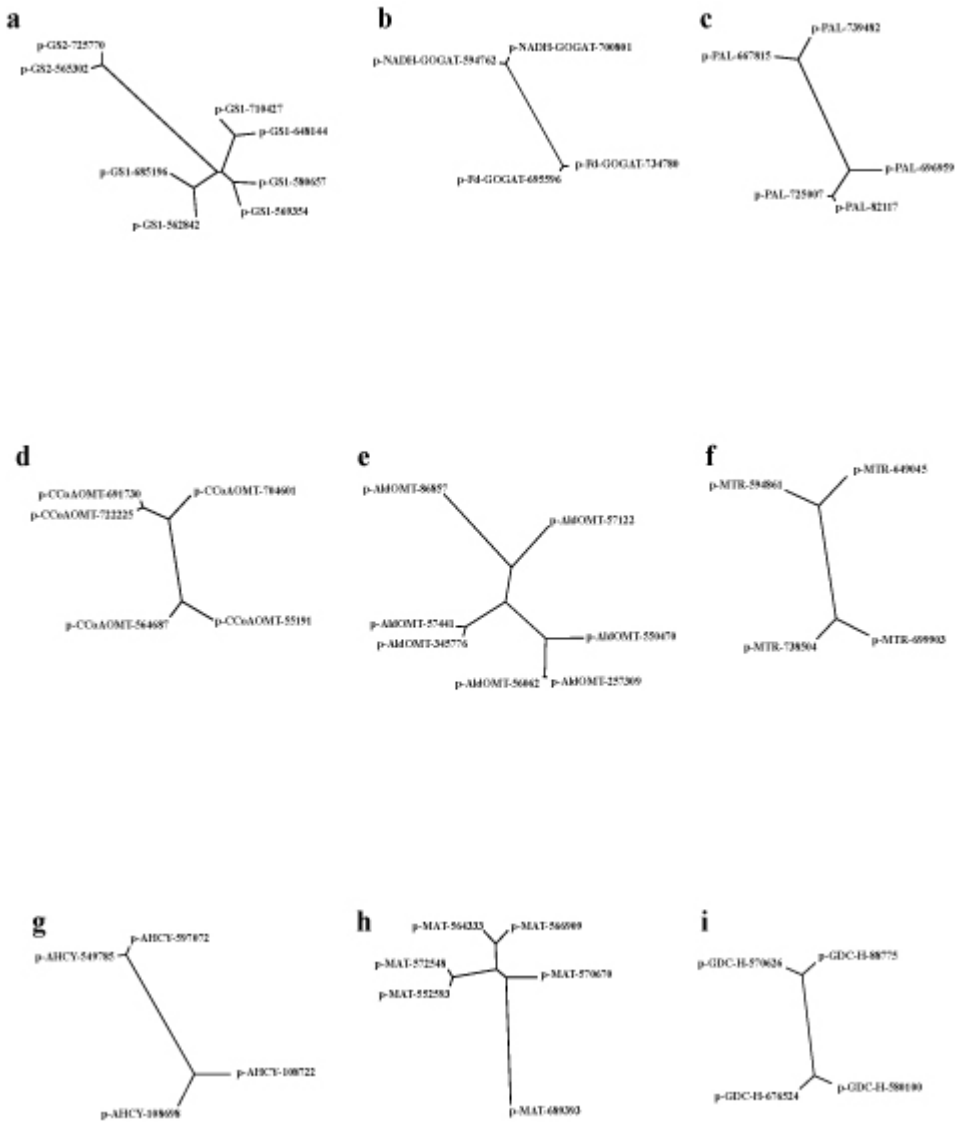
JGI	Populus DB	
Protein identification	EST code	PU code
GS1-685196	G089P77	PU06292
	G133P82	PU08292
GS1-562842	G068P15	PU06292
	A050P6	PU01775
GS1-710427	A073P68	PU02286
GS1-648144	UB12CPA11	PU04179
GS1-569354	A060P71	PU00168
	A083P21	PU02327
GS1-580657	A004P03	PU00010
	A089P51	PU02929
GS2-565302	Y010F07	-
	C011P07	-
GS2-725770	G065P67	PU06395
	F092P27	PU12880
Fd-GOGAT-695596	UA520PB04	PU04046
Fd-GOGAT-734780	C041P20	PU09069
NADH-GOGAT-700801	X076G05	PU28873
PAL-696959	A017P40	-
	F24P64	PU12211
	F50P34	PU12787
	G089P92	-
	B006P35	PU02719
PAL-725007	G119P27	PU07826
PAL-667815	M113A06	-
PAL-739482	P083H04	-
CCoAOMT-691730	A078P58	PU02461
CCoAOMT-722225	A057P63	PU01811
CCoAOMT-564687	A073P02	PU02140
	G101P19	-
AldOMT-345776	A015P67	PU00567
	A071P77	PU01944
AldOMT-57441	G120P55	PU07978
MTHFR-644300	A013P60	PU00592
	A081P19	PU02648
	A084P39	PU02625
	G091P18	PU07165
MTHFR-704929	G108P66	PU07366
MTR-649045	G092P14	PU06933
	G093P14	PU07062
MTR-594861	A039P12	PU01315
MTR-699903	A034P50	PU01393
MTR-738504	A042P20	PU01218
	A067P49	PU02268
AHCY-549785	A056P30	PU01655
	A033P80	-
	G063P01	-
	G066P52	-
AHCY-597072	A061P69	PU01624
	A007P33	-
	A039P60	-
	A084P37	-
	A085P34	-
MAT-552593	A051P67	PU01693
	A083P41	PU02615
	A065P30	PU02081
MAT-572548	A007P49	PU00333
	A012P68	PU00390
	G126P72	PU08231
MAT-564333	G064P93	-
	G080P67	-
MAT-566909	F016P07	PU11823
	F017P71	-
	G089P58	-
GDC-H-570626	A055P18	PU01744
	R025F05	PU22552
GDC-H-88775	UB56BP011	PU05496
GDC-P-685558	C053P66	PU09327
	C059P65	PU09278
	PF012P54	PU11896
	G066P51	PU06257
	I054P03	-
GDC-T-833954	C031P03	-
GDC-T-828778	G101P10	PU07242

FIGURE 1. Hypothesis of pathways coordinated for lignin production



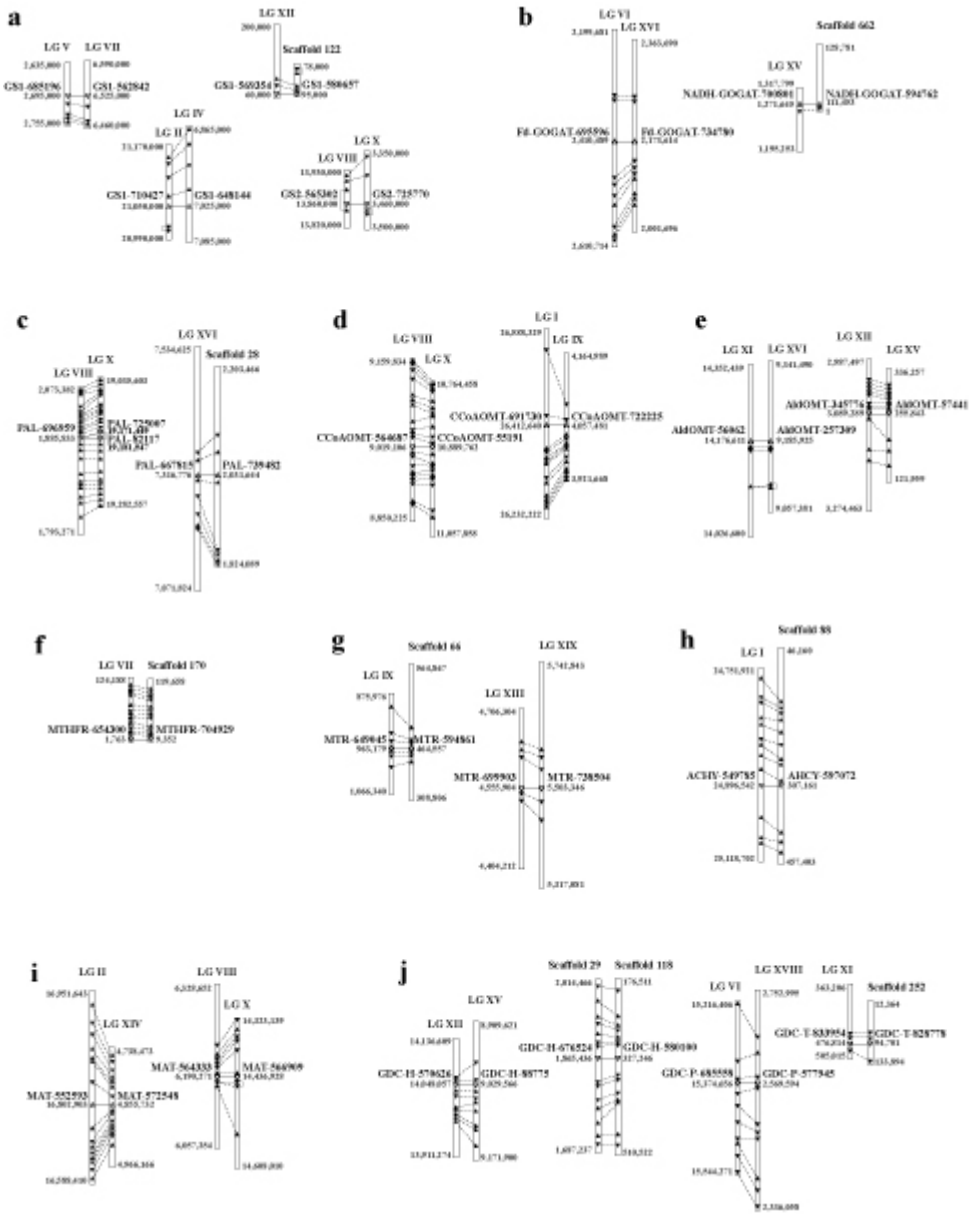
GS: Glutamine synthetase. GOGAT: Glutamate synthase. AKG: Alpha keto glutarate. PAL: Phenylalanine ammonia-lyase. CCoAOMT: Caffeoyl-CoA O-methyltransferase. AldOMT: Caffeic acid/5-hydroxyferulic acid O-methyltransferase. MTHFR: 5,10-Methylenetetrahydrofolate reductase. THF: Tetrahydrofolate. MTR: 5-Methyltetrahydrofolate- homocysteine methyltransferase (methionine synthase). MAT: S-adenosylmethionine synthetase. SAM: S-adenosylmethionine. AHCY: S-adenosyl homocysteine hydrolase. GDC: Glycine decarboxylase. SHMT: Serine hydroxymethyltransferase. SAH: S-adenosyl-homocysteine.

FIGURE 2. Protein alignment of *Populus* families.



a: GS. b: GOGAT. c: PAL. d: CCoAOMT. e: AldOMT. f: MTR. g: AHCY. h: MAT. i: GDC-H. The closest branches suggest candidate duplicated sequences.

FIGURE 3. Gene duplications



Arrows indicate the 5'-3' gene orientation. White arrows are the duplicated studied genes united by horizontal solid line. Black arrows are the duplicated collinear genes specified by dotted lines. Internal duplications are also marked by grey lines. LG or scaffolds numbers are written at the top of each. Grey numbers indicate the genomic location.

a: Downstream the GS1-685196 and GS1-562842 genes and upstream GS1-710427 and GS1-648144 genes were located 4 collinear duplicated genes. The pair GS1-569354 and GS1-580657

presented 2 collinear duplications in the upstream region, but the downstream region could not be examined because the scaffold finished just where the GS gene was located. 2 collinear duplications were placed upstream the pair GS2-565302 and GS2-725770.

b: When the region surrounding the duplicated Fd-GOGAT-695596 and Fd-GOGAT-734780 genes was examined, 9 duplicated genes appeared collinearly disposed. The gene NADH-GOGAT-594762 was in a scaffold of only 128,781 bases, consequently an extensive examination was not possible. However 1 collinear ORF was found downstream NADH-GOGAT positions.

c: The duplicated genes of PAL-696959 were PAL-725007 and PAL-82117 which were closely placed. A total of 22 collinear genes confirmed the duplication of the genomic region. PAL-667815 and PAL-739482 were also duplicated genes, as it was showed when in the region were found 7 other collinear genes.

d: The duplication of CCoAOMT-564687 was CCoAOMT-55191: 11 and 9 collinear duplications were found respectively upstream and downstream. Similar regions were examined for the duplicated CCoAOMT-691730 and CCoAOMT-722225, finding 13 collinear duplications.

e: The downstream region AldOMT-56062 and AldOMT-257309 had 3 collinear genes. By contrast, 8 (upstream) and 3 (downstream) collinear duplicated genes were detected for AldOMT-345776 and AldOMT-57441.

f: It was not possible to examine the downstream region of the MTHFR-654300 and MTHFR-704929 because this last was placed close to the end of the scaffold 170. However the upstream region contained 12 collinear duplicated genes.

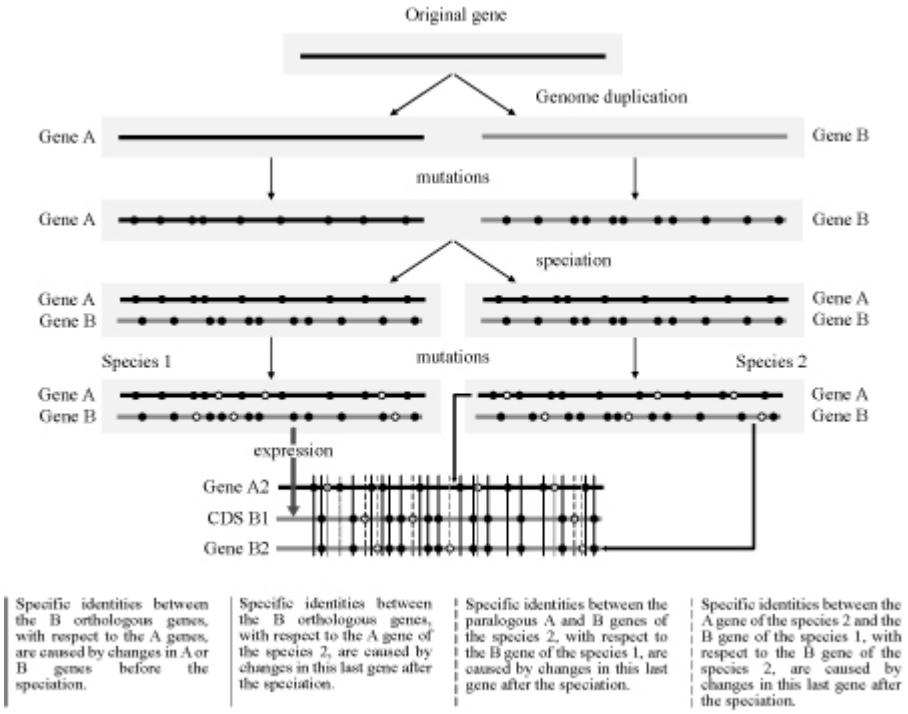
g: MTR-649045 and MTR-594861 genes had 6 collinear duplications and for MTR-699903 and MTR-738504 the screening counted 5.

h: ACHY-549785 and ACHY-597072 were also duplicated genes. The screening of the genomic region showed 12 duplications collinearly disposed.

i: The genomic region enclosing the duplicated MAT-552593 and MAT-572548 had 15 collinear duplicated genes; 2 duplications were located downstream and 6 upstream for MAT-564333 and MAT-566909.

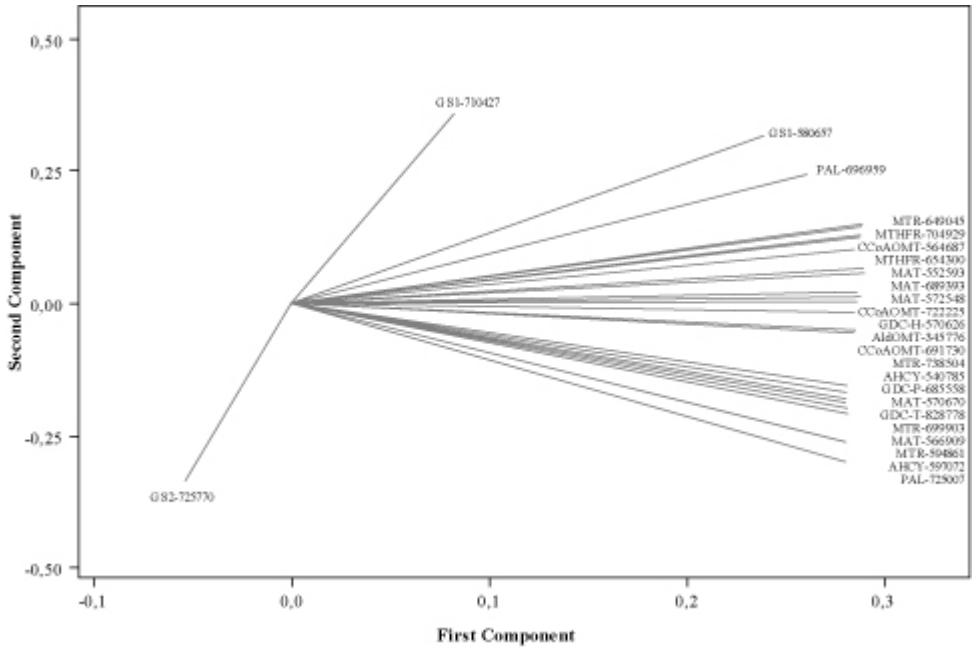
j: The genomic region of the GDC-H-570626 and GDC-H-88775 genes contained 9 collinear duplicated genes. 14 duplications were found for the GDC-H-676524 and GDC-H-580100, 10 for GDC-P-685558 and GDC-P-577945, and finally only 3 for GDC-T-833954 and GDC-T-828778.

FIGURE 4. Genome polyploidization



Genome polyploidization produce the duplication of a particular original gene generating copies A and B. Progressive independent changes on the A and B duplicated genes are inherited when a speciation event happens. New punctual different changes on the A and B genes produce even more divergence between A and B genes of the species 1 and 2. Orthologous genes are A1 with A2, and B1 with B2. Paralogous genes are A1 with B1, and A2 with B2. Comparison among duplicated A2 and B2 genes with the B1 transcript: Solid lines indicate mutations respectively before (dark) or after (grey) speciation, that represent conserved similitude between B1 and B2 orthologous. However changes happened on B1 after the speciation (dark dashed lines), represent conserved residues between the paralogous A2 and B2. Moreover, changes on B2 afters speciation (grey dashed lines) represent common elements between A2 and B1, i.e. between a gene and the paralogous of its orthologous.

FIGURE 5. Principal Component Analysis (PCA).



Plot of the analyzed variables (gene expressions) on the two first principal components: 90.782% and 3.283% of the variance respectively. Practically, all the gene coexpression values were positively correlated with the first principal component, with the exception of GDC-H-570625 which shows an inverse relationship. The second principal component is mainly characterized by the mutually exclusive expression of GS1-10427 and GS2-725770 respectively.

REFERENCES

ÁVILA C, SUÁREZ MF, GÓMEZ-MALDONADO J, CÁNOVAS FM. Spatial and temporal expression of two cytosolic glutamine synthetase genes in Scots pine: Functional implications on nitrogen metabolism during early stages of conifer development. The Plant Journal, 2001; 25(1): 1-11.

BAUCHER M, MONTIES B, MONTAGU MV, BOERJAN W. Biosynthesis and genetic engineering of lignin. Critical Reviews in Plant Sciences, 1998; 17(2):125-197.

BOERJAN W, RALPH J, BAUCHER, M. Lignin biosynthesis. Annu. Rev. Plant Biol, 2003; 54: 519-546.

BYLESJÖ M, ERIKSSON D, SJÖDIN A, SJÖSTRÖM M, JANSSON S, ANTTI H, TRYGG J. MASQOT: A method for cDNA microarray spot quality control. BMC Bioinformatics, 2005; 6:250.

CÁNOVAS FM, CANTÓN FR, GARCÍA-GUTIÉRREZ A, GALLARDO F, CRESPILLO R. Molecular physiology of glutamine and glutamate biosynthesis in developing conifer seedlings. Physiologia Plantarum, 1998; 103: 287–294.

CÁNOVAS FM, GALLARDO F, PING ZP, PASCUAL MB. Transgenic approaches to engineer nitrogen metabolism. Tree Transgenesis, 2006; 157-178.

CANTÓN FR, SUÁREZ MF, CÁNOVAS FM. Molecular aspects of nitrogen mobilization and recycling in trees. Photosynthesis Research, 2005; 83: 265-278.

CHEN F, SRINIVASA REDDY MS, TEMPLE S, JACKSON L, SHADLE G, DIXON RA. Multi-site genetic modulation of monolignol biosynthesis suggests new routes

for formation of syringyl lignin and wallbound ferulic acid in alfalfa (Medicago sativa L.), The Plant Journal, 2006; 48: 113–124.

CORPET F. Multiple sequence alignment with hierarchical clustering. Nucl. Acids Res, 1988; 16 (22): 10881-10890.

DONG Q, KROISS L, OAKLEY FD, WANG BB, BRENDEL V. Comparative EST analyses in plant systems. Methods in enzymology, 2005; 395: 400-418.

DOUCE R, BOURGUIGNON J, NEUBURGER M, RÉBEILLÉ F. The glycine decarboxylase system: A fascinating complex. Trends in Plant Science, 2001; 6 (4): 167-176.

FUKASAWA-AKADA T, KUNG SD, WATSON JC. Phenylalanine ammonia-lyase gene structure, expression, and evolution in nicotiana. Plant Mol Biol, 1996; 30(4): 711-722.

GALLARDO F, FU J, JING ZP, KIRBY EG, CÁNOVAS FM. Genetic modification of amino acid metabolism in woody plants. Plant Physiology and Biochemistry, 2006; 41: 587-594.

GARCÍA-GUTIÉRREZ A, CANTÓN FR, GALLARDO F, SÁNCHEZ-JIMÉNEZ F, CÁNOVAS FM. Expression of ferredoxin-dependent glutamate synthase in dark-grown pine seedlings. Plant molecular biology, 1995; 27: 115-128.

GONZÁLEZ-MARTÍNEZ SG, KRUTOVSKY KV, NEALE DB. Forest-tree population genomics and adaptive evolution. New Phytologist, 2006; 170: 227-238.

GUO D, CHEN F, INOUE K, BLOUNT JW, DIXON RA. Downregulation of caffeic Acid 3-O-methyltransferase and caffeoyl CoA 3-O-methyltransferase in transgenic

alfalfa: Impacts on lignin structure and implications for the biosynthesis of G and S Lignin. The Plant Cell, 2001; 13: 73-88.

GRAY-MITSUMUNE M, MOLITOR EK, CUKOVIC D, CARLSON JE, DOUGLAS CJ. Developmentally regulated patterns of expression directed by poplar PAL promoters in transgenic tobacco and poplar. Plant Molecular Biology, 1999; 39: 657-669.

HANSON AD, ROJE S. One-carbon metabolism in higher plants. Annu. Rev. Plant Physiol. Plant Mol. Biol, 2001, 52:119-137.

HAYAKAWA T, NANTO K, KAWAI S, KATAYAMA Y, MOROHOSHI N. Molecular cloning and tissue-specific expression of two genes that encode caffeic acid O-methyltransferases from *Populus kitakamiensis*. Plant Science, 1996; 113(2): 157-165.

HERTZBERG M, ASPEBORG H, SCHRADER J, ANDERSSON A, ERLANDSSON R, BLOMQVIST K, BHALERAO R, UHLÉN M, TEERI TT, LUNDEBERG J, SUNDBERG B, NILSSON P, SANDBERG G. A transcriptional roadmap to wood formation. PNAS, 2001; 98 (25): 14732-14737.

JOUANIN L, GOUJON T, DE NADAÏ V, MARTIN MT, MILA I, VALLET C, POLLET B, YOSHINAGA A, CHABBERT B, PETIT-CONIL M, LAPIERRE C. Lignification in transgenic poplars with extremely reduced caffeic acid O-methyltransferase activity. Plant Physiology, 2000; 123: 1363-1373.

KAO YY, HARDING SC, TSAI CJ. Differential expression of two distinct phenylalanine ammonia-lyase genes in condensed tannin-accumulating and lignifying cells of quaking aspen. Plant Physiology, 2002; 130: 796-807.

KIRST M, MYBURG AA, DE LEÓN JPG, KIRST ME, SCOTT J. Coordinated genetic regulation of growth and lignin revealed by quantitative trait locus analysis of cDNA microarray data in an interspecific backcross of Eucalyptus. Plant Physiology, 2004; 135: 2368-2378.

KOONIN EV. An apology for orthologs -or brave new memes. Genome Biology, 2001; 2(4): 1005.1–1005.2.

KOONIN EV. Orthologs, Paralogs, and Evolutionary Genomics. Annu. Rev. Genet, 2005; 39: 309–338.

KUMAR A, ELLIS BE. The phenylalanine ammonia-lyase gene family in raspberry. Structure, expression, and evolution. Plant Physiology, 2001, 127: 230-239.

KUMAR S, TAMURA T, MASATOSHI N. MEGA3: Integrated software for Molecular Evolutionary Genetic Analysis and sequence alignment. Briefings in bioinformatics, 2004; 5(2): 150-163.

LAM HM, COSCHIGANO KT, OLIVEIRA IC, MELO-OLIVEIRA R, CORUZZI GM. The molecular-genetics of nitrogen assimilation into amino acids in higher plants. Annu. Rev. Plant Physiol. Plant Mol. Biol, 1996; 47: 569-593.

LEA PJ, ROBINSON SA, STEWART GR. The enzymology and metabolism of glutamine, glutamate, and asparagine. In The Biochemistry of Plants, 1997; 16: 121-159.

LI L, LU S, CHIANG V. A genomic and molecular view of wood formation. Critical reviews in Plant Sciences, 2006; 25: 215-233.

MENG H, CAMPBELL WH. Substrate profiles and expression of caffeoyl coenzyme A and caffeic acid O-methyltransferases in secondary xylem of aspen during seasonal development. Plant Molecular Biology, 1998; 38: 513-520.

MERKLE SA, DEAN JFD. Forest tree biotechnology. Current Opinion in Biotechnology, 2000; 11: 298-302.

MOUILLON JM, AUBERT S, BOURGUIGNON J, GOUT E, DOUCE R, RÉBEILLÉ F. Glycine and serine catabolism in non-photosynthetic higher plant cells: Their role in C1 metabolism. The Plant Journal, 1999; 20 (2): 197-205.

OSAKABE Y, NANTO K, KITAMURA H, KAWAI S, KONDO Y, FUJII T, TAKABE K, KATAYAMA Y, MOROHOSHI N. Immunocytochemical localization of phenylalanine ammonia-lyase in tissues of *Populus kitakamiensis*. Planta, 1996; 200(1): 13-19.

OSAKABE Y, OHTSUBO Y, KAWAI S, KATAYAMA Y, MOROHOSHI N Structure and tissue-specific expression of genes for phenylalanine ammonia-lyase from hybrid aspen, *Populus* kitakamiensis. Plan Science, 1995; 105: 217-226.

PELEMAN J, BOERJAN W, ENGLER G, SEURINCK J, BOTTERMAN J, ALLIOTTE T, MONTAGU MV, INZÉ D. Strong cellular preference in the expression of a housekeeping gene of Arabidopsis thaliana encoding S-adenosylmethionine synthetase. The Plant Cell, 1998; 1: 81-93.

PEREIRA L, TODOROVA M, CAI X, MAKAROFF C, EMERY R, MOFFATT B. Methyl recycling activities are co-ordinately regulated during plant development. Journal of Experimental Botany, 2007; 58 (5) 1083-1098.

PÉREZ GARCÍA, DE VICENTE A, CANTÓN FR, CAZORLA FM, CODINA JC, GARCÍA-GUTIÉRREZ A, CÁNOVAS FM. Light-dependent changes in the relative proportions of tomato glutamine synthetase polypeptides in response to Pseudomonas syringae infection or phosphinothricin treatment. Physiologia Plantarum, 1998a; 102: 377-384.

PÉREZ-GARCÍA A, PEREIRA S, PISSARRA J, GARCÍA-GUTIÉRREZ A, SALEMA R, DE VICENTE A, CÁNOVAS FM. Cytosolic localization in tomato mesophyll cells of a novel glutamine synthetase induced in response to a Pseudomonas syringae infection or phosphinothricin treatment. Planta, 1998b; 206: 426-434.

PRASSINOS C, KO JH, YANG J, HAN KH. Transcriptome profiling of vertical stem segments provides insights into the genetic regulation of secondary growth in hybrid aspen trees. Plant Cell Physiol, 2005; 46(8): 1213-1225.

PRODHAN AKMA, OHTANI J, FUNADA R, ABE H, FUKAZAWA K. Ultrastructural investigation of tension wood fibre in Fraxinus mandshurica Rupr. var. japonica Maxim. Annals of Botany, 1995; 75: 311-317.

RAJINIKANTH M, HARDING SA, TSAI CJ. The glycine decarboxylase complex multienzyme family in *Populus*. Journal of Experimental Botany, 2007; 1-10.

RAPP RA, WENDEL JF; Epigenetics and plant evolution. New Phytologist, 2005; 168: 81-91.

RAZAL R, ELLIS S, SINGH S, LEWIS NG, TOWERS GHN. Nitrogen recycling in phenylpropanoid metabolism. Phytochemistry, 1996; 41 (1): 31-35.

RENSINK WA, BUELL CR. Microarray expression profiling resources for plant genomics. Trends Plant Sci, 2005; 10(12): 603-609.

RICHMOND T, SOMERVILLE S; Chasing the dream: Plant EST microarrays. Current Opinion in Plant Biology, 2000; 3:108-116.

SÁNCHEZ-AGUAYO I, RODRÍGUEZ-GALÁN JM, GARCÍA R, TORREBLANCA J, PARDO JM. Salt stress enhances xylem development and expression of S-adenosyl-L-methionine synthase in lignifying tissues of tomato plants. Planta, 2004; 220(2): 278-285.

SATO T, TAKABE K, FUJITA M. Immunolocalization of phenylalanine ammonia-lyase and cinnamate-4-hydroxylase in differentiating xylem of poplar. C. R. Biologies, 2004; 327: 827-836.

SCHOENBECK MA, TEMPLE SJ, TREPP GB, BLUMENTHAL JM, SAMAC DA, GANTT SJ, HERNANDEZ G, VANCE CP. Decreased NADH-glutamate synthase activity in nodules and flowers of alfalfa (Medicago sativa L.) transformed with an antisense glutamate synthase transgene. Journal of Experimental Botany, 2004; 51: 29-39.

SCHRADER J, NILSSON J, MELLEROWICZ E, BERGLUND A, NILSSON P, HERTZBERG M, SANDBERG G. A High-Resolution transcript profile across the wood-forming meristem of Poplar identifies potential regulators of cambial stem cell identity. The Plant Cell, 2004; 16: 2278-2292.

SCHRÖDER G, EICHEL J, BREINIG S, SCHRÖDER J. Three differentially expressed S-adenosylmethionine synthetases from Catharanthus roseus: Molecular and functional characterization. Plant Molecular Biology, 1997; 33: 211–222.

SHEN B, LI C, TARCZYNSKI MC (2002) High free-methionine and decreased lignin content result from a mutation in the Arabidopsis S-adenosyl-Lmethionine synthetase 3 gene. The Plant Journal, 2002; 29(3): 371-380.

STERCK L, ROMBAUTS S, JANSSON S, STERKY F, ROUZE P, VAN DE PEER Y. EST data suggest that poplar is an ancient polyploidy. New Phytologist, 2005; 167: 165-170.

STERKY F, BHALERAO RR, UNNEBERG P, SEGERMAN B, NILSSON P, BRUNNER AM, CHARBONNEL-CAMPAA L, LINDVALL JJ, TANDRE K, STRAUSS SH, SUNDBERG B, GUSTAFSSON P, UHLÉN M, BHALERAO RP, NILSSON O, SANDBERG G, KARLSSON J, LUNDEBERG J, JANSSON S. A *Populus* EST resource for plant functional genomics. Plant biology, 2004; 101: 13951-13956.

STERKY F, REGAN S, KARLSSON J, HERTZBERG M, ROHDE A, ROLMBERG A, AMINI B, BHALERAO R, LARSSON M, VILLARROEL R, MONTAGU MV, SANDBERG G, OLSSONI O, TEERI TT, BOERJAN W, GUSTAFSSON P, UHLÉN M, SUNDBERG B, LUNDEBERG J. Gene discovery in the wood-forming tissues of poplar: Analysis of 5,692 expressed sequence tags. Proc. Natl. Acad. Sci. USA, 1998; 95: 13330-13335.

SUÁREZ MF, ÁVILA C, GALLARDO F, CANTÓN FR, GARCÍA-GUTIÉRREZ A, CLAROS MG, CÁNOVAS FM. Molecular and enzymatic analysis of ammonium assimilation in woody plants. Journal of Experimental Botany, 2002; 53: 891-904.

SUBRARNANIAM R, REINOLD S, MOLITOR EK, DOUGLAS CJ. Structure, inheritance, and expression of hybrid Poplar (*Populus* trichocarpa x *Populus* deltoides) phenylalanine ammonia-lyase genes. Plant Physiol, 1993; 102: 71-83.

VAN HEERDEN PS, NEIL TOWERS GS, LEWIS NG. Nitrogen Metabolism in Lignifying Pinus taeda Cell Cultures. The Journal of Biological Chemistry, 1996; 271(21): 12350–12355.

VANDER MIJNSBRUGGE K, MEYERMANS H, MONTAGU MV, BAUW G, BOERJAN W. Wood formation in poplar: identification, characterization, and seasonal variation of xylem proteins. Planta, 2002; 210: 589-598.

VANDER MIJNSBRUGGE K, MONTAGU MV, INZÉ D, BOERJAN W. Tissue-specific expression conferred by the S-adenosyl-L-methionine synthetase promoter of Arabidopsis thaliana in transgenic poplar. Plant Celt Physiol, 1996; 37(8): 1108-1115.

VINCENT D, LAPIERRE C, POLLET B, CORNIC G, NEGRONI L, ZIVY M. Water deficits affect caffeate O-methyltransferase, lignification, and related enzymes in maize leaves. A Proteomic Investigation. Plant Physiology, 2005; 137: 949-960.

WHETTEN R, SEDEROOF RR. Lignin biosynthesis. The Plant Cell, 1995; 7: 1001-1013.

WHETTEN R, SUN YH, ZHANG Y, SEDEROFF RR. Functional genomics and cell wall biosynthesis in loblolly pine. Plant Molecular Biology, 2001; 47: 275-291.

WULLSCHLEGER SD, JANSSON S. Genomics and forest biology: Populus emerges as the perennial favorite. The Plant Cell. 2002; 14: 2651-2655.

ZHONG R, MORRISON WH, NEGREL J, YE ZH. Dual methylation pathways in lignin biosynthesis. The Plant Cell, 1998; 10: 2033-2045.

ZHONG R, MORRISON WH, HIMMELSBACH DS, POOLE FL, YE ZH, RICHARD B. Essential role of caffeoyl coenzyme a O-methyltransferase in lignin biosynthesis in woody Poplar plants. Plant Physiology, 2000; 124: 563-577.