

Characterization of genes with tissue-specific differential expression patterns in *Populus*

Bo Segerman · Stefan Jansson · Jan Karlsson

Received: 5 December 2005 / Revised: 9 August 2006 / Accepted: 17 November 2006 / Published online: 10 January 2007
© Springer-Verlag 2007

Abstract Like many plants, *Populus* has an evolutionary history in which several, both recent and more ancient, genome duplication events have occurred and, therefore, constitutes an excellent model system for studying the functional evolution of genes. In the present study, we have focused on the properties of genes with tissue-specific differential expression patterns in poplar. We identified the genes by analyzing digital expression profiles derived by mapping 90,000+ expressed sequence tags (ESTs) from 18 sources to the predicted genes of *Populus*. Our sequence analysis suggests that tissue-specific differentially expressed genes have less diverged paralogs than average, indicating that gene duplication events is an important event in the pathway leading to this type of expression pattern. The functional analysis showed that genes coding for proteins involved in processes of functional importance for the specific tissue(s) in which they are expressed and genes coding for regulatory or responsive proteins are most common among the differentially expressed genes, demonstrating that the expression differentiation process is under strong selective pressure. Thus, our data supports a model where gene duplication followed by gene specialization or expansion of the regulatory and responsive networks leads to tissue-specific differential expression patterns. We have also searched for clustering of genes with similar expression pattern into gene-expression neighborhoods within the

Populus genome. However, we could not detect any major clustering among the analyzed genes with highly specific expression patterns.

Keywords Tissue-specific differential expression pattern · Gene duplication · Expression neighborhood

Introduction

Gene duplications have long been recognized as important events during the evolution of new genes and new gene functions (neofunctionalization; Ohno 1970). Plant genomes are often rich in duplicated genes originating from polyploidization events, with subsequent diploidization (Adams and Wendel 2005; Blanc et al. 2003; Simillion et al. 2002; Wolfe 2001; Sterck et al. 2005), and that of *Populus*, the first tree genome sequenced, is no exception (Tuskan et al. 2006). The *Populus* genome is especially interesting as its history contains, in addition to more ancient events, a fairly recent whole genome duplication. Newly duplicated genes are susceptible to degeneration (nonfunctionalization; Force et al. 1999), and a preservative mechanism can be subfunctionalization, i.e., division of the ancestral expression pattern or function(s) between the daughter genes (Force et al. 1999; Lynch and Force 2000; Prince and Pickett 2002). Subfunctionalization may thus lead to restriction of the duplicated genes' expression patterns. It should be noted that different degrees of neofunctionalization can occur at later stages after the subfunctionalization, sometimes referred to as subneofunctionalization (He and Zhang 2005). Alternatively, the copies can stay redundant (Gu et al. 2003).

In this study, we aimed to characterize the properties of genes for which tissue-specific differential expression patterns have evolved. As, e.g., subfunctionalization is a

Communicated by: S. Strauss

Electronic supplementary material Supplementary material is available in the online version of this article at <http://dx.doi.org/10.1007/s11295-006-0077-6> and is accessible for authorized users.

B. Segerman (✉) · S. Jansson · J. Karlsson
Umeå Plant Science Centre, Department of Plant Physiology,
Umeå University,
SE-90187 Umeå, Sweden
e-mail: bo.segerman@plantphys.umu.se

process restricting the expression pattern, we wanted to do our analysis in the context of duplicated genes, so it was important to distinguish in some cases between very similar transcripts. It is known that arrays, especially cDNA microarrays, may be influenced by cross-hybridization between closely related paralogous genes (Evertsz et al. 2001). A more robust measurement of transcript profiles, when comparing closely related genes, can be obtained by quantifying and comparing the frequencies of expressed sequenced tags (ESTs) picked from different nonnormalized cDNA libraries (Audic and Claverie 1997). However, this technique also has disadvantages, including the limitation that only genes with sufficiently high expression to be picked frequently from the cDNA libraries can be analyzed, and the questions that can be addressed are restricted to the conditions chosen in the available cDNA libraries. Important regulatory genes with low expression levels will not be picked up in this type of analysis.

We have previously reported the sequencing of 100,000+ ESTs from 18 nonnormalized cDNA libraries originating from different poplar tissues under different conditions (Sterky et al. 2004). This data set represents a comprehensive resource for deriving digital expression profiles, which was used here to search for genes showing tissue-specific differential expression patterns. The clustering of the ESTs into “clusters” and “contigs” (Sterky et al. 2004) gives an estimation of which poplar genes the ESTs are derived from. However, the availability of the genome sequence and the predicted gene models gives possibilities for a much more accurate annotation of the ESTs. We, therefore, decided to map the EST data set to the gene models/transcripts identified by the *Populus* genome project and generated digital expression profiles. Gene models/transcripts with specific enrichment in one of the 18 cDNA libraries were selected, and their relation to other poplar genes, functional profiles, and distribution in the genome were examined.

Thus, this study examines the properties of genes with tissue-specific differential expression patterns in *Populus*. Furthermore, as recent articles have suggested that clustering of similarly regulated genes into neighborhoods is a more general phenomenon than first believed (Oliver and Misteli 2005), we have analyzed the positioning of the identified genes within the genome. In addition, this study provides a catalog of interesting *poplar* genes for studying tissue- and condition-specific processes.

Results

Mapping of 90,000+ ESTs from 18 cDNA libraries to the *Populus* genes

The EST collection used in this study, derived from 18 nonnormalized cDNA libraries, has previously been de-

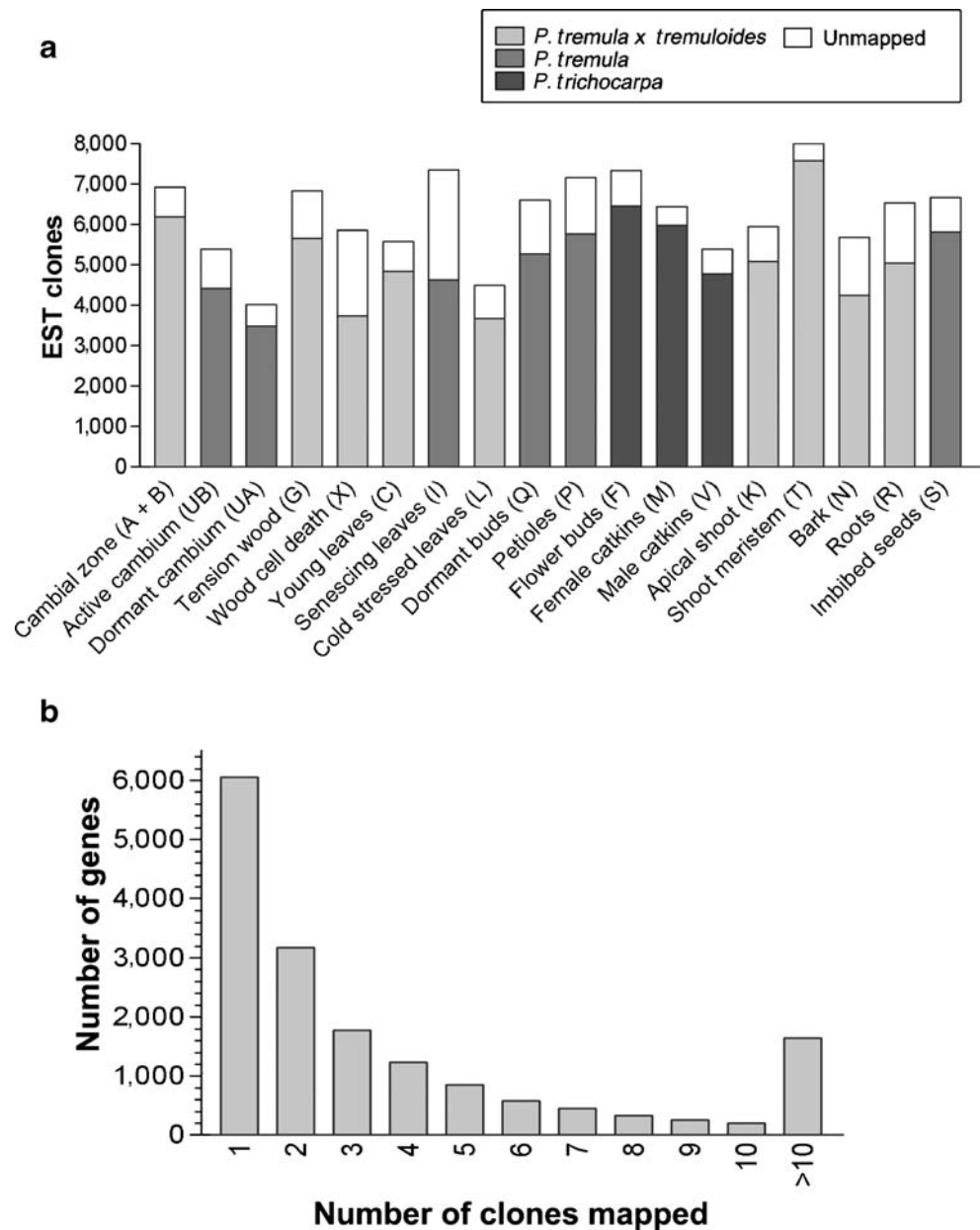
scribed (Sterky et al. 2004). To get a more accurate coupling between the ESTs (or rather the cDNA clones picked for sequencing) and the actual genes, we decided to map the ESTs directly to the gene models rather than use the clustration. We used a basic local alignment search tool (BLAST)-based mapping protocol to assign the ESTs to *Populus* gene models (hereafter referred to as genes) and accepted assignments when the best BLASTN high-scoring segment pair showed at least 90% identity over at least 30% of the sequence, or at least 80% identity over at least 50% of the sequence, or at least 75% identity over at least 70% of the sequence. Using this approach, we mapped 119,860 ESTs corresponding to 92,674 cDNA clones (some clones are sequenced more than once). All further analysis was done at the clone level. Approximately 91% of our clones were assigned to a gene, and typically between 3,500 and 7,500 clones were mapped per cDNA library (Fig. 1a). We had EST support for 16,435 of the 45,555 predicted genes, and there were 7,313 genes with at least 3 and 5,538 genes with at least 4 clones mapped (Fig. 1b).

Our cDNA libraries came from three different poplar species: *P. tremula* (European aspen), *P. tremula* x *tremuloides* (hybrid aspen), and *P. trichocarpa* (black cottonwood), whereas the genome is sequenced from *P. trichocarpa*. However, we did not observe any major differences between the species in mapping efficiency (Fig. 1a). The mean sequence identities among the mapped *P. trichocarpa*, *P. tremula*, and *P. tremula* x *tremuloides* sequences were 96.3, 93.6, and 94.3%, respectively. Thus, the sequence distance between these species seems to be generally around 2–3%, which should be sufficiently close for this analysis, although certain genes may have diverged more.

To clarify the source of the unmapped ESTs, we conducted a series of BLAST searches. We queried the unmapped sequences against the whole set of 337,000+ gene predictions (including alternative gene models predicted by different algorithms from the same loci), the *Populus* genome assembly, the *Populus* mitochondrion, the *Populus* chloroplast, the *Escherichia coli* genome, and the European Molecular Biology Laboratory fungi sequence collection. We assigned each EST to one of these groups or to the “Unmapped” group if the best *e* value was worse than $1e-10$. The results showed that most unmapped ESTs had a hit either to an alternative gene prediction that was not present in the filtered 45,555 set or to the genome (Fig. 2a). There were virtually no *E. coli* contaminants, but we found a number of ESTs with fungal origins, primarily in the senescing leaf (I) and bark (N) libraries.

To estimate the extent to which the sequences that were not mapped to any gene prediction but mapped to the genome represent undiscovered genes, we calculated the distance to the nearest exon in the nonredundant 45,555

Fig. 1 a Number of mapped ESTs for each of the cDNA libraries. ESTs were mapped to the *Populus* gene models with a BLAST-based protocol. Numbers of unmapped ESTs are represented in white. Species that provided the cDNA source tissue are indicated by color. **b** Number of genes (gene models) to which one or more EST clones mapped. 16,435 genes had EST support, and at least three, four, five, and six ESTs mapped to 7,313; 5,538; 4,313; and 3,458 genes, respectively

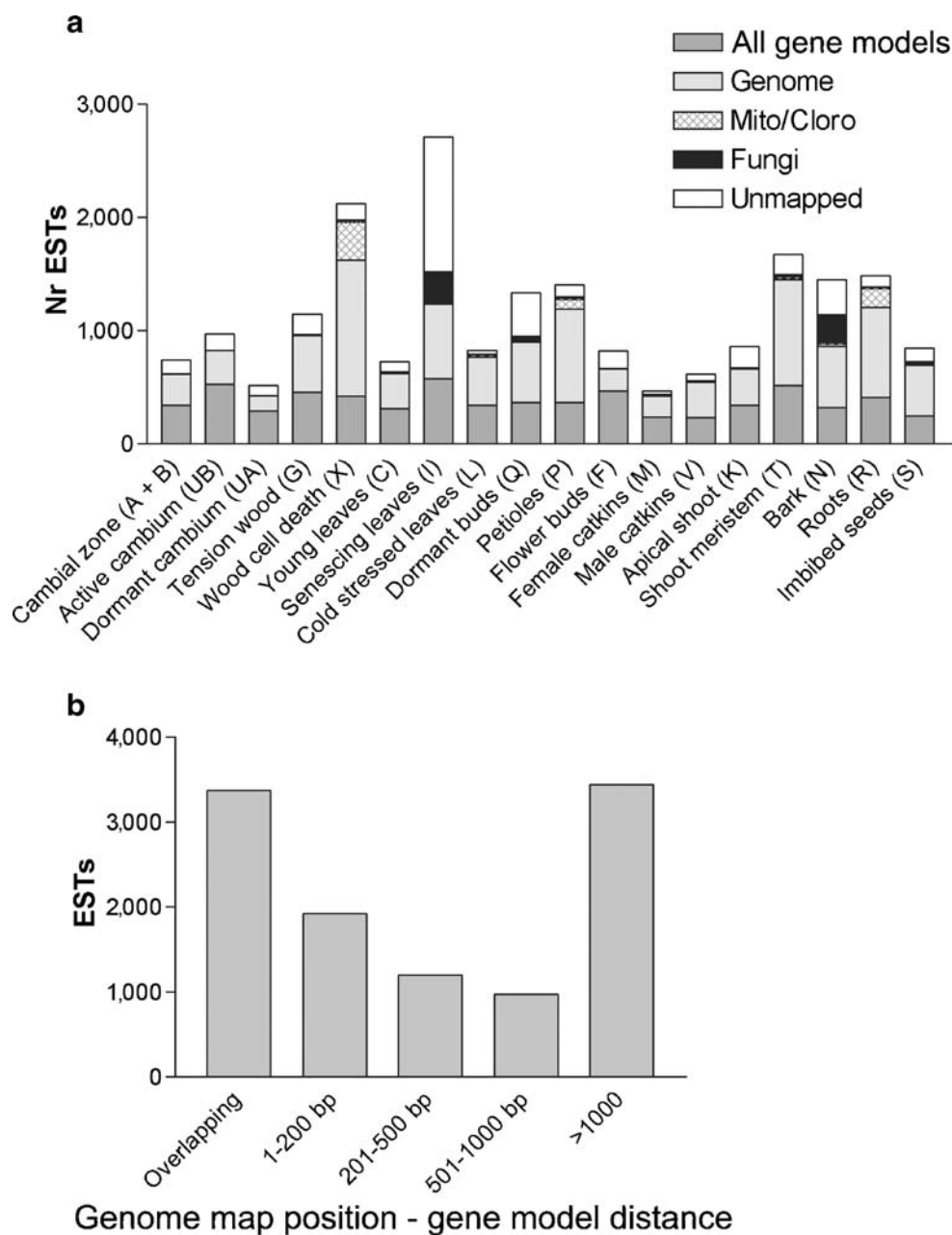


gene set. As shown in Fig. 2b, approximately half of these ESTs mapped within 200 bp of a gene. Furthermore, the orientation of 91% of the sequences that mapped within 200 bp of a gene matched the orientation of the respective gene. These findings indicate that the algorithms sometimes predicted too short UTR regions, which is consistent with knowledge that current gene prediction programs make mistakes, especially in the analyses of large genomes (Mathe et al. 2002). It is also possible that some of the ESTs may have originated from alternative splice forms. The EST mapping protocol used in the *Populus* genome project (<http://www.jgi.doe.gov/poplar/>) was slightly more stringent than ours, so the EST set used by the prediction programs excluded many of the ESTs that we have mapped.

Thus, our data suggest that many gene predictions may be improved by utilizing more of the available EST data. We also compared the different gene-calling algorithms. All of the algorithms used (grail, genewise, Eugene, fgenesh) appeared to be prone to predict too short genes, but genewise appeared somewhat worse and Eugene somewhat better than average (data not shown).

The fungal contaminants came primarily from the senescing leaf (I) and bark (N) libraries. The root library was derived from plants grown in agar under sterile conditions, which explains the absence of arbuscular mycorrhizal fungi sequences. The fungal sequences that scored originated from a diverse group of fungi, but some of the larger groups were from *Botrytis cinerea* (gray old, N

Fig. 2 a Origins of the unmapped ESTs. Unmapped ESTs were queried against fungal sequences, the *Populus* mitochondrion, the *Populus* chloroplast, the *Populus* genome, and all available gene model predictions (including alternative gene models predicted by different algorithms from the same loci) using BLASTN. **b** Distance in base pairs to the closest gene model for ESTs that scored well to the genome but not to a gene model in the filtered (nonredundant) set [light gray in (a)]



and I), *Yarrowia lipolytica* (yeast, N and I), *Schizosaccharomyces pombe* (yeast, primarily I), *Candida glabrata* (yeast, primarily I), *Debaryomyces hansenii* (yeast, primarily I), and *Neurospora crassa* (red bread mold, primarily N). In addition, of the sequences that remained unmapped (as described above), approximately a third were from the senescing leaves library, indicating that many of them may originate from fungal, or other, contaminants that are not yet represented in the sequence database. The diverse fungal flora of bark and wood from *Populus* grown in Alberta and British Columbia has been analyzed (Hutchison 1994, 1999), but with some exceptions (e.g. *Candida*); the fungal families isolated in Canada were quite different than

those on the Swedish trees. Several of the fungal sequences have been spotted on the Umeå Plant Science Centre cDNA microarray, providing scope for monitoring fungal infections during experiments.

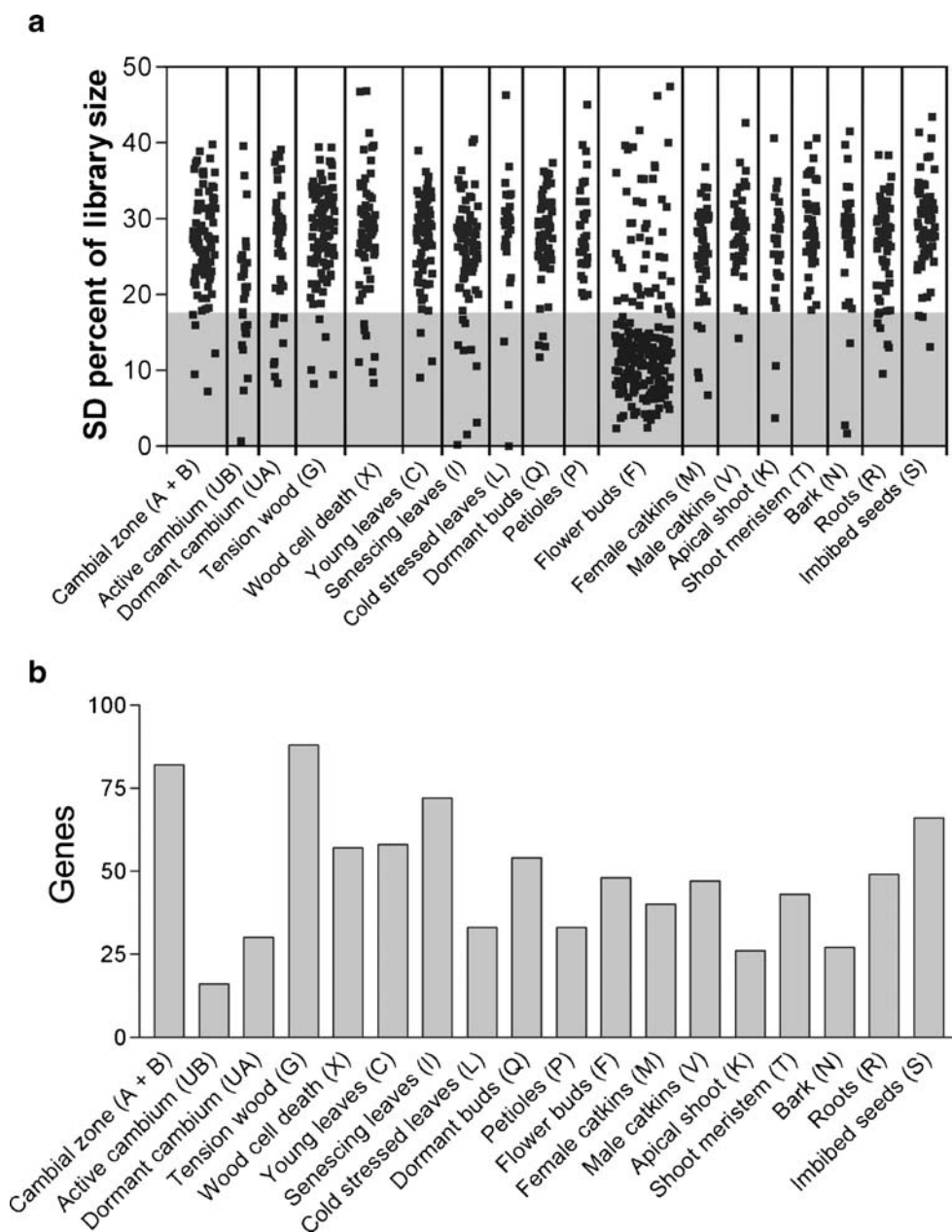
Identification of genes with tissue-specific differential expression patterns

Our approach to identifying genes with tissue-specific differential expression patterns was to search for genes that were specifically overrepresented in a particular cDNA library compared to the other 17 cDNA libraries. To do this, we first generated a master table with clone counts for each

gene and library. We then conducted a series of Fisher exact tests in which we compared the proportion of clones in each library (that had an EST) with the proportion of clones in all other libraries (fused) for each gene (7,313 genes with at least three ESTs mapped were analyzed, in total 35,531 tests). Then, for each library, we selected genes that (1) showed a *p* value lower than 0.001 in the Fisher exact test and (2) were up-regulated at least fivefold compared to the level expected from a random distribution of the clones. We found 1,120 cases where a gene was overrepresented in a library. However, we were concerned about the quality of the flower bud (F) library, as there were signs that its complexity had declined during the latter stages of its sequencing (clones from the same transcripts were picked

over and over again due to technical problems with the cDNA-library). We therefore calculated the standard deviation of the clone order in the pick list from the corresponding library for each case of identified overrepresentation (Fig. 3a). We then excluded all cases that showed a standard deviation of <18% of the library size (gray zone in Fig. 3a). Thus, a relatively large proportion of the library-specific clones of the excluded genes had been picked during a short time window. After this filtering, 869 cases of overrepresentation remained, representing 821 genes (Supplementary Table 1). The genes that were significant at 5% (545 genes), 1% (473 genes), and 0.1% (389 genes) according to the stringent Bonferroni correction for multiple tests are labeled with ‘*’, ‘**’, and ‘***’,

Fig. 3 (a) Variation in position within the ordered clone pick list for the identified up-regulated genes. A low value means that most of the clones were picked during a short time window of the analysis. The F library showed reductions in quality in the later stages of sequencing. Gene models within the gray zone were filtered away. (b) Number of identified tissue specifically expressed genes in each library, after filtering. The identification method was based on a Fisher exact test



respectively, in Supplementary Table 1. A single asterisk (*) corresponded to $p < 1.4 \times 10^{-6}$, double asterisks (**) to $p < 2.8 \times 10^{-7}$, and triple asterisks (***) to $p < 2.8 \times 10^{-8}$. Approximately 95% of the genes were up-regulated in a single cDNA library, and approximately 5% were up-regulated in two, often related, libraries. The distribution of genes with tissue-specific differential expression patterns between the different cDNA libraries is shown in Fig. 3b. Supplementary Table 1 is also available as an Excel file at http://www.populus.db.umu.se/proj_downl.php

Sequence distance to paralogous genes

If the differential expression patterns of these genes are consequences of gene duplication events, one would expect them to have a close paralog somewhere else in the genome. To test this hypothesis, we defined the sequence similarity of the closest other poplar gene. As shown in Fig. 4, for most of the tissue specifically expressed genes, another closely related gene was present. The sequence similarity distance to the closest presumed paralog was smaller in this group than the average distance for the whole gene set. Of the differentially expressed genes, 84% scored at least half of the self-score (the maximum score that the sequence can obtain, i.e., score against a 100% identical sequence) toward another gene as compared to 67.6% in the “all genes”-set. The presumed paralog was generally not in the identified list of genes with tissue-

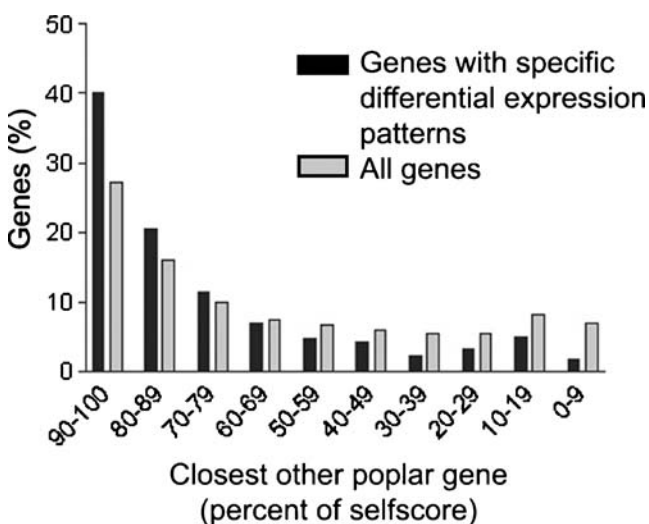


Fig. 4 Scoring distance to the closest other gene within the *Populus* genome for the tissue specifically expressed genes and all poplar genes. The distance to presumed paralogs was measured as a percentage of the self-scores (i.e., BLASTP score normalized for length by dividing with the score when the sequence is compared to itself). For the genes with tissue-specific differential expression patterns, 72% scored better than 70%, 84% scored better than 50%, and 10% scored worse than 30% of the self-score with the closest other gene. Corresponding numbers for all genes were 53, 68, and 21%, respectively

specific differential expression, indicating it generally had broader expression pattern (for less than 1% of the duplication pairs both sister genes were found to have a specific differential expression pattern).

Functions of the genes with tissue-specific differential expression patterns

A specific expression pattern is often considered to be indicative of the gene function, so we examined the functional profiles of the tissue-specific differentially expressed genes we had identified. The genes were annotated according to their closest *Arabidopsis* hits (The Institute for Genomic Research [TIGR]) and divided into broad functional categories based on a manual inspection of the annotation and the *Arabidopsis* gene ontology (Supplementary Table 1). If the BLASTP score of the closest *Arabidopsis* protein was less than 100, the function was classified as “unknown.” Thus, it should be noted that the functions assigned in this list are not confirmed in *Populus*. Nevertheless, most *poplar* genes appear to have a closely related *Arabidopsis* gene, implying a high degree of functional similarity. The functional profiles of the genes identified as having a tissue-specific differential expression pattern are summarized in Fig. 5 and Supplementary Fig. 1. In many cases, the functions of the genes appear to clearly reflect the functions of the source tissue of the cDNA library. In addition, regulatory and responsive genes were also common, especially in developing, dormant, and remodeling tissues. A short summary of the genes that were found to be differentially expressed in each library is presented in Table 1. For convenience, the genes included are sometimes described as proteins or processes rather than genes encoding proteins.

Search for gene-expression neighborhoods

Similarly expressed genes sometimes tend to cluster into “gene-expression neighborhoods” within the genome. As 18 sets of genes with similar (tissue-specific differential) expression patterns were available, we examined possible tendencies toward such clustering among them. For this purpose, we plotted the localization of the identified genes for each library on the genome assembly (Fig. 6). Each “window” in the graph represents 100,000 bases. Some similarly regulated tandem duplications resulting in two (and in one case three) genes were located in the same window. Apart from the tandem duplications, there was a tendency towards gene-clustering in the seed library (S) at LG_X:5,100,000–5,800,000 where five unrelated seed-specific genes were located. Another interesting observation was that at LG_V:16,450,000–16,700,000 there

Fig. 5 **a** Functions of the tissue-specific differentially expressed genes. The functions of the identified genes were divided into broad categories, and their distribution in each library were plotted. **b** Energy related functions. **c** Cell wall-related functions. **d** Signal transduction related functions. **e** Hormone signaling related functions. More graphs are available in Supplementary Fig. 1

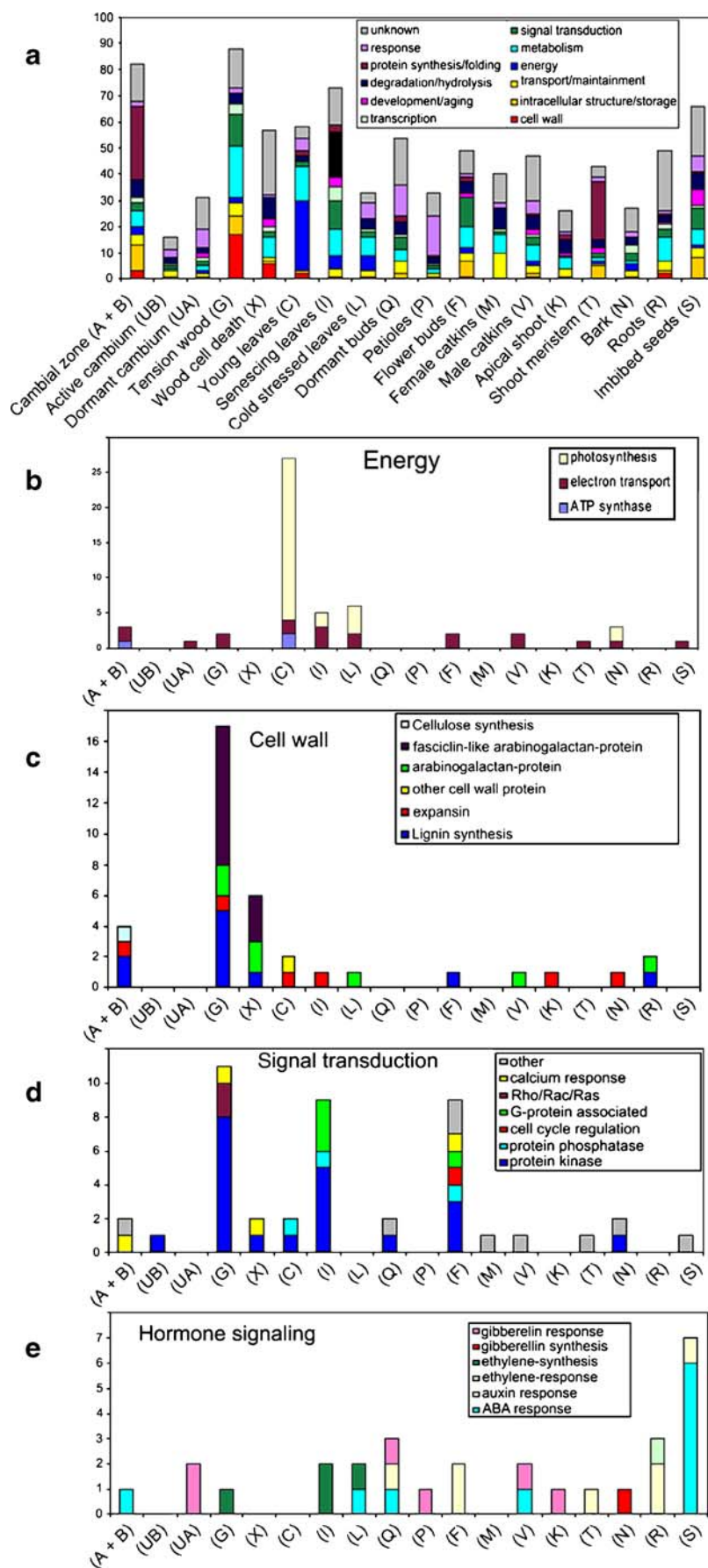


Table 1 Summary of the functions for the genes with specific differential expression patterns

Libraries	
Cambial zone library (A+B)	Ribosomal proteins, cell wall synthesis/remodeling, cytoskeleton proteins (primarily actin related), transcription factors, ABA-responsive protein
Active cambium library (UB)(a thinner cambial section)	Transcription factor, putative kinase, phosphate starvation-induced proteins, pectinase
Dormant cambium library (UA)	Low temperature-responsive/cold-acclimation proteins, water deprivation-response proteins, and defense-response proteins, responses to gibberellin, late embryogenesis-abundant protein
Tension wood library (G) ^a	Cell wall-related proteins, carbohydrate, glycine and methionine biosynthesis, tubulin isoforms, kinases, ethylene biosynthesis, transcription factors
Wood cell death library (X) ^b	Cell wall-related proteins, proteases
Young leaves library (C) ^c	Photosynthesis, carbohydrate metabolism cell wall proteins, ubiquitin ligase, protease, a protein kinase and a phosphatase
Senescing leaves library (I) ^c	Proteases, phosphoesterases, an exonuclease, chlorophyll breakdown, ethylene-synthesis, transcription factors, protein kinases, SPX domain proteins
Cold-stressed leaves library (L)	Photosynthesis, carbohydrate metabolism, polysaccharide catabolism, stress responses, amylases, dehydrin, low temperature/salt-responsive protein, putative ABA-inducible protein, ethylene-forming enzyme
Dormant bud library (Q)	Dehydrins, low temperature/salt response protein, putative Pi starvation-induced proteins, pathogenesis-related proteins, protease inhibitors, an ABA-, an auxin-, and a gibberellin-responsive protein
Petioles library (P)(subjected to various stress conditions)	Class I heat shock proteins, a putative gibberellin-regulated protein, a tubulin isoform, and a myb family transcription factor
Flower Buds library (F)	Signal transduction proteins, protein kinases flower development proteins, auxin responsive proteins, histones, intracellular vesicle transport, flavonoid synthesis
Female Catkins library (M)	Kunitz family proteins, inositol-3-phosphate synthase, WRKY family transcription factor, a pectinesterase, an endochitinase, and two osmotin-like proteins
Male catkins library (V)	Anther development protein, a floral homeotic protein, a gibberellin-regulated protein, a putative ABA-inducible protein, a late embryogenesis abundant protein, flavonoid biosynthesis, proteases, a cellulase, a pectinesterase, a lipase, plant defensin proteins, and an uclacyanin
Apical shoot library (K)	Two lipases, a chitinase, a protein involved in lignin biosynthesis, a histone, a ribosomal protein, and two defense response proteins. A gibberellin-regulated protein
Shoot meristem library (T)	Ribosomal proteins, histones, an auxin-responsive protein, a rapid alkalinization factor (RALF), an axial regulator—YABBY1/abnormal floral organs protein (AFO)/filamentous flower protein (FIL), and an embryo-specific protein
Bark library (N)	WRKY family transcription factors, photosynthesis proteins, disease resistance proteins, a phytochrome A, signal transduction protein. An enzyme involved in gibberellin biosynthesis
Root library (R)	Glutathione <i>S</i> -transferases, a metallothionein, transcription factors, a peroxidase, and a haloacid dehalogenase-like hydrolase. Auxin response proteins and an ethylene-response protein
Imbibed seeds library (S)	Storage proteins, developmental proteins, and hormone response proteins. Dehydrins, one low temperature- and salt- responsive protein, a Bax inhibitor, ABA-responsive proteins, and an auxin down-regulated protein

^a See also (Andersson-Gunneras et al. 2006)

^b See also (Moreau et al. 2005)

^c See also (Bhalerao et al. 2003)

seemed to be an enrichment of hormone-responsive proteins. Taken all genes in account, there were three gibberellin-regulated, one ABA-responsive, and four auxin-responsive genes located in this region. However, in general, the genes were essentially randomly spread. Thus, we could not detect any obvious accumulation into gene-expression neighborhoods among the identified tissue specifically expressed genes. However, as this data set only covers a portion of the differentially expressed genes, we cannot exclude that such neighborhoods may exist.

Discussion

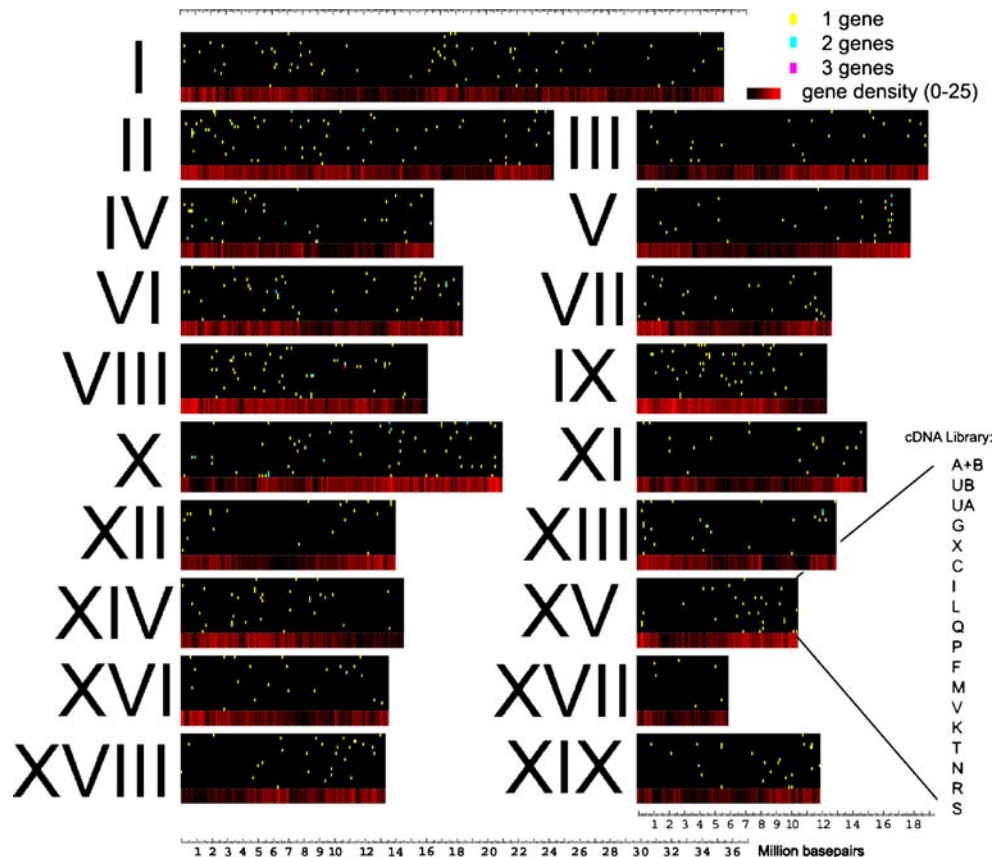
The genomes of higher organisms contain a large number of genes compared to unicellular organisms, but many of the genes are only expressed in certain places and under certain conditions. Our aim in this study was to learn more about genes with tissue-specific differential expression patterns in *Populus* and how they evolve. The recently sequenced *Populus* genome (Tuskan et al. 2006) is an interesting model system for addressing this kind of

question, as it has undergone several large-scale duplication events, both recent and ancient, and is therefore, probably, in the middle of an active gene-evolution phase. After a duplication event, the selective pressure acting on the sister genes is reduced, and four possible gene-fates have been discussed: (1) one of the sister genes may degenerate (nonfunctionalization), (2) the ancestral gene-expression pattern or, possibly, function(s) may be divided between the daughter genes (subfunctionalization), (3) one of the sister genes may obtain a new function (neofunctionalization; Force et al. 1999; Lynch and Force 2000; Prince and Pickett 2002), and (4) the copies may stay redundant (Gu et al. 2003). In practice, subfunctionalization may be accompanied by different degrees of neofunctionalization, and recent results suggest that rapid subfunctionalization followed by prolonged neofunctionalization is a major pathway during gene evolution (He and Zhang 2005). Our data provide strong support for the hypothesis that gene duplication and probably also subfunctionalization are important factors driving the evolution of tissue-specific gene-expression patterns in *Populus*. The genes with tissue-specific differential expression patterns appear to have more closely related paralogs than average, indicating that most of the tissue-specific differentially expressed genes originated from a duplication event. Also supporting the

importance of duplication events in the evolution of tissue specificity in expression patterns is the interesting observation that the “old” (100–150 Myr—fourfold synonymous third-codon transversion position—4DTV=0.31+0.02) duplication pairs, as defined in the *Populus* genome project, are clearly more rich in genes with tissue-specific expression patterns than the “recent” (10–50 Myr 4DTV=0.07+0.001) pairs, which in turn are richer in tissue-specific genes than random pairs (Tuskan et al. 2006).

Subfunctionalization has been suggested to be a rather rapid process that serves to preserve the sister genes after a duplication event (He and Zhang 2005). In theory, preservation is ensured when expression of one of the genes is lost, or becomes too weak to meet essential needs, in one temporal or spatial location and the other gene in another. None of the sister genes can then fulfill the ancestral function(s) alone, but together they can. However, there may still be a large overlap in the expression patterns. It has been shown that certain functional classes are preferentially retained after whole genome duplication events (Blanc and Wolfe 2004). For instance, in *Arabidopsis*, families of transcription factors and signal transducers are strongly expanded during whole-genome duplication events (Maere et al. 2005). Our data suggest that the process leading to tissue-specific differential expression

Fig. 6 Position on the chromosomes of the tissue specifically expressed genes. The linkage groups (chromosomes) are divided into windows of 100,000 bases. The tissue specifically expressed genes were assigned to the windows based on the coordinates of the center of the gene. Windows with no, one, two, and three tissue specifically expressed genes are colored black, yellow, blue, and violet, respectively. The total gene density is also plotted in red. The libraries are denoted as follows: cambial zone (AB); active cambium (UB); dormant cambium (UA); tension wood (G); wood cell death (X); young leaves (C); senescing leaves (I); cold-stressed leaves (L); dormant buds (Q); petioles (P); flower buds (F); female catkins (M); male catkins (V)



patterns is under functional selection pressure and is accompanied by gene specialization or rearrangement of the regulatory or responsive network. Thus, our data could be explained if the function of one of the sister genes tends to shift after a duplication event so that it is better adapted to a certain tissue, which may concurrently have negative effects in other locations where it is expressed. This results in a skewed selection pressure that may restrict the expression pattern of the gene. The initial broad subfunctionalization process may be rapid, but the following reduction in the expression pattern into tissue specificity is probably a slower process accompanied and driven by gene specialization.

The involvement of functional selection in the evolution of tissue-specific differential expression patterns preferentially favors genes coding for proteins involved in processes of functional importance for the tissues where they are expressed. Examples of functional processes in which tissue-specific differential expressed genes tend to become involved include: photosynthesis (leaves), protein synthesis (cambium, shoot meristem), cell wall formation (wood), flavonoid synthesis (flowers), and degradation (senescing leaves, wood cell death). Tissue specifically expressed responsive proteins were found primarily in dormant and stressed tissues, although various defense-responsive proteins with tissue-specific differential expression patterns were found in many tissue types. Thus, defense responses seem to have different requirements in different tissues. Tissue specifically expressed signal transduction proteins were mostly found in tissues undergoing some kind of morphological change (senescing leaves, flower buds, and tension wood-forming tissues). Hormone synthesis/response was also a common function among the genes with tissue-specific differential expression patterns. This is not surprising, as an essential feature of hormone regulation is spatial specificity.

Some of the tissue specifically expressed genes we identified did not have obvious paralogs. These genes may have obtained their specificity by an alternative process, other than gene duplication. Subfunctionalization is believed to arise from degenerative mutations in the regulatory *cis*-elements (Force et al. 1999), but it is important to remember that specific up-regulation of a transcription factor can affect the expression of many genes in certain tissues in *trans*. This may be an elegant way for evolution to specifically enhance the expression of a whole regulon that is important for specific tissues. An interesting observation is that the differentially expressed genes that had no obvious paralog were generally poorly annotated (72% of the group with scores less than 20% of self-score), and many had no Arabidopsis homologue (39% of the group with scores less than 20% of self-score). As the annotations were derived from Arabidopsis (the most

thoroughly studied plant), these sets may represent tree-specific genes that have been lost in the Arabidopsis evolutionary line or possibly gained in the *Populus* evolutionary line. The wood cell death and root libraries were especially rich in these “orphan” genes, suggesting that wood formation and root function may have genetically diverged particularly strongly between these species. However, the differences between roots could reflect differences in habitat rather than tree-herb-specific differences.

Prokaryotes have long been known to cluster related genes into operons, but eukaryotic gene expression has traditionally been considered to be randomly distributed within the genome. However, recent studies have shown that co-expression of neighboring genes occurs in yeast (Cohen et al. 2000), *Drosophila* (Cohen et al. 2000), *Caenorhabditis elegans* (Lercher et al. 2003), *Homo sapiens* (Lercher et al. 2002), and *Arabidopsis* (Williams and Bowles 2004). Higher level co-regulation is speculated to involve phenomena such as the specific organization of chromatin structure, shared regulatory elements, and epigenetic mechanisms. Tandem duplication also results in pairs of genes that are located close to each other and have similar expression patterns. As we had identified several data sets with genes that have similar, tissue-specific, expression patterns, we examined whether they showed any tendency to cluster on the chromosomes. To do this, we plotted the locations of the genes on bands representing the chromosomes. However, we could not detect any major co-localization and, therefore, concluded that the regulation of these genes seem to be largely independent, originating from their own regulatory elements. This suggests that gene-expression neighborhoods are not widespread among genes with tissue-specific expression patterns, although we cannot exclude that they exist, as we have only analyzed a portion of the tissue specifically expressed genes.

In conclusion, we have found evidence in the *Populus* model system of gene duplication dependent process, directed by functional selection and specialization, leading to tissue-specific differential expression patterns. We have also shown that the process in general appears to be local to the duplicated gene, i.e., it does not involve gene-expression neighborhoods. Thus, this study sheds light on gene with tissue-specific differential expression patterns and the processes involved in the evolution of them.

Materials and methods

Data sources

The ESTs can be downloaded at <http://www.populus.db.umu.se> and are all archived in GenBank. The cDNA

libraries have been described previously (Sterky et al. 2004). However, it should be noted that the petiole library (P) also contains sequences from heat-treated petioles. The species from which the libraries originated are indicated in Fig. 1, and the samples were collected, briefly, as follows: the cambial zone library (AB)—bark was peeled and tissue was scraped from both exposed surfaces with a scalpel; active cambium (UB)—30- μ m sections were obtained by cryosectioning (July 10th 2001); dormant cambium (UA)—30- μ m sections were obtained by cryosectioning (October 5th 2001); tension wood (G)—wood scrapings were taken from a tree inclined for 3 weeks in the greenhouse; wood cell death (X)—tissues collected consisted mainly of xylem cells in which the cell wall was fully developed, but included some in which secondary cell wall formation was not yet complete; young leaves (C)—trees cultured in a greenhouse in fertilized peat under natural light supplemented with metal halogen lamps; senescing leaves (I)—leaves were collected from a single wild tree on the Umeå University campus a few days before visible leaf senescence was observed; cold-stressed leaves (L)—greenhouse-grown plants were transferred to 5°C, then fully developed leaves were sampled 3 and 4 days after transfer and pooled; dormant buds (Q)—dormant buds were collected in February; petioles (P)—petioles were collected from several individuals, growing in long-day conditions and stressed in different ways were pooled; flower buds (F)—immature female inflorescence tissue was collected; female catkins (M)—flushing catkins were collected in early spring; male catkins (V)—flushing catkins were collected in early spring; apical shoot (K)—150 apical shoots (top 3 mm, biggest leaf approximately 5 mm, weight approximately 4 mg) from 3-month-old greenhouse-grown plants were collected and pooled; shoot meristem (T)—shoot apices were dissected under a microscope; bark (N)—bark was sampled from under the “crown” and 75 cm downwards, the sample was peeled with a potato peeler, avoiding buds, and the cells were inspected under a microscope; roots (R)—roots were collected from plants grown in agar under sterile conditions, the whole root system (primary roots, still white) up to 0.5–1 cm from the stem was used; imbibed seeds (S)—seeds from a seed lot were imbibed, and samples were taken both immediately after imbibition and after 24 h and pooled. The *Populus* gene models, genome sequence, mitochondrion sequence and chloroplast sequence were provided by the *P. trichocarpa* genome project. The *E. coli* genome was downloaded from GenBank (accession U00096). The fungi sequences were downloaded (04-02-2005) from EBI. (<ftp://ftp.ebi.ac.uk/pub/databases/embl/release/fun.dat.gz>). The coordinates of the gene models were provided by the *P. trichocarpa* genome project. *Arabidopsis* annotations were downloaded from TIGR (<http://www.tigr.org>) and the gene ontology

classifications from TAIR (<http://www.arabidopsis.org>). Data were handled using MySQL and Filemaker.

Data analysis

All sequence comparisons were done with WU-BLAST 2.0 (<http://www.blast.wustl.edu/>) using standard parameters. The ESTs were queried against the *Populus* gene models, the *Populus* genome, the *E. coli* genome, the *Populus* mitochondrion, the *Populus* chloroplast, and the fungi sequence collection with BLASTN. The BLAST score, *e* value, identity, and coordinates of the hit were extracted from the result file. The gene models were compared to the *Arabidopsis* proteome (TIGR) using BLASTP to get an annotation. The coordinates of the genome hits were compared with the coordinates of the gene models provided by the *Populus* genome project. The distances from gene models were measured to the closest exon boundary. The tissue-specific differentially expressed genes were identified by Fisher tests (two tailed) performed using the R package (<http://www.r-project.org/>). The count of ESTs for a gene in a library vs the size of that library was compared with the sum of the ESTs for all other libraries vs the sum of the sizes of those libraries. Only genes with an EST count of at least three was considered, and for a particular gene, only libraries that had ESTs were tested. In total 35,531 tests were carried out. Bonferroni corrections of the cutoffs were carried out at 5, 1, and 0.1%, and genes were classified accordingly in the result table. Significant cases of down-regulation found were not analyzed, as they were generally consequences of strong up-regulation in another library. The identified genes were then filtered by calculating the standard deviation of the order in the clone-pick list using the R package. This step was carried out to ensure that the EST clones had been picked with an even spread-across-the-projects time course. We discarded tissue-specific genes that had a standard deviation within the clone-pick order of less than 18% of the number of clones in that library. Sequence distance to paralogs were calculated by BLASTN with the gene queried against all other genes. The genes were also queried against themselves to get the maximal achievable score (self-score) using BLASTN. The percentage of the paralog score of the self-score was then calculated. Functional classification was done by manual inspection of the TIGR annotations and GO data. C++ was used for the calculations and chromosome plots. The chromosome plots were generated using a stepping window of size 100,000 bp and compared to the gene coordinate lists provided by the *Populus* genome project. The count of tissue-specific genes in each window was calculated and visualized as bands. Also, the total gene frequency in each window was calculated.

Acknowledgment This work was supported by the Knut and Alice Wallenberg Foundation and the Swedish Foundation for Strategic Research.

References

- Adams KL, Wendel JF (2005) Polyploidy and genome evolution in plants. *Curr Opin Plant Biol* 8:135–141
- Andersson-Gunneras S, Mellerowicz EJ, Love J, Segerman B, Ohmiya Y, Coutinho PM, Nilsson P, Henrissat B, Moritz T, Sundberg B (2006) Making cellulose enriched gelatinous fibers in poplar: global analysis of transcripts and metabolites identifies biochemical and developmental regulators in secondary wall biosynthesis. *Plant J* 45:144–165
- Audic S, Claverie JM (1997) The significance of digital gene expression profiles. *Genome Res* 7:986–995
- Bhalerao R, Keskitalo J, Sterky F, Erlandsson R, Bjorkbacka H, Birve SJ, Karlsson J, Gardstrom P, Gustafsson P, Lundeberg J et al (2003) Gene expression in autumn leaves. *Plant Physiol* 131:430–442
- Blanc G, Hokamp K, Wolfe KH (2003) A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. *Genome Res* 13:137–144
- Blanc G, Wolfe KH (2004) Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *Plant Cell* 16:1679–1691
- Cohen BA, Mitra RD, Hughes JD, Church GM (2000) A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat Genet* 26:183–186
- Evertsz EM, Au-Young J, Ruvolo MV, Lim AC, Reynolds MA (2001) Hybridization cross-reactivity within homologous gene families on glass cDNA microarrays. *Biotechniques* 31:1182, 1184, 1186 passim
- Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151:1531–1545
- Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li WH (2003) Role of duplicate genes in genetic robustness against null mutations. *Nature* 421:63–66
- He X, Zhang J (2005) Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* 169:1157–1164
- Hutchison L (1994) Some wood-inhabiting yeasts of trembling aspen (*Populus tremuloides*) from Alberta and northeastern British Columbia. *Mycologia* 86:386–391
- Hutchison L (1999) Wood-inhabiting microfungi isolated from *Populus tremuloides* from Alberta and northeastern British Columbia. *Can J Bot* 77:898–905
- Lercher MJ, Urrutia AO, Hurst LD (2002) Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat Genet* 31:180–183
- Lercher MJ, Blumenthal T, Hurst LD (2003) Coexpression of neighboring genes in *Caenorhabditis elegans* is mostly due to operons and duplicate genes. *Genome Res* 13:238–243
- Lynch M, Force A (2000) The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154:459–473
- Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y (2005) Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci USA* 102:5454–5459
- Mathe C, Sagot MF, Schiex T, Rouze P (2002) Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res* 30:4103–4117
- Moreau C, Aksenov N, Lorenzo MG, Segerman B, Funk C, Nilsson P, Jansson S, Tuominen H (2005) A genomic approach to investigate developmental cell death in woody tissues of *Populus* trees. *Genome Biol* 6:R34
- Ohno S (1970) Evolution by gene duplication. Springer, Berlin Heidelberg New York
- Oliver B, Misteli T (2005) A non-random walk through the genome. *Genome Biol* 6:214
- Prince VE, Pickett FB (2002) Splitting pairs: the diverging fates of duplicated genes. *Nat Rev Genet* 3:827–837
- Simillion C, Vandepoele K, Van Montagu MC, Zabeau M, Van de Peer Y (2002) The hidden duplication past of *Arabidopsis thaliana*. *Proc Natl Acad Sci USA* 99:13627–13632
- Sterky F, Bhalerao RR, Unneberg P, Segerman B, Nilsson P, Brunner AM, Charbonnel-Campaa L, Lindvall JJ, Tandre K, Strauss SH et al (2004) A *Populus* EST resource for plant functional genomics. *Proc Natl Acad Sci USA* 101:13951–13956
- Sterck L, Rombauts S, Jansson S, Sterky F, Rouze P, Van de Peer Y (2005) EST data suggest that poplar is an ancient polyploid. *New Phytol* 167:165–170
- Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen GL, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, Cunningham R, Davis J, Degroev S, Dejardin A, Depamphilis C, Detter J, Dirks B, Dubchak I, Duplessis S, Ehrling J, Ellis B, Gendler K, Goodstein D, Gribskov M, Grimwood J, Groover A, Gunter L, Hamberger B, Heinze B, Helariutta Y, Henrissat B, Holligan D, Holt R, Huang W, Islam-Faridi N, Jones S, Jones-Rhoades M, Jorgensen R, Joshi C, Kangasjarvi J, Karlsson J, Kelleher C, Kirkpatrick R, Kirst M, Kohler A, Kalluri U, Larimer F, Leebens-Mack J, Leple JC, Locascio P, Lou Y, Lucas S, Martin F, Montanini B, Napoli C, Nelson DR, Nelson C, Nieminen K, Nilsson O, Pereda V, Peter G, Philippe R, Pilate G, Poliakov A, Razumovskaya J, Richardson P, Rinaldi C, Ritland K, Rouze P, Ryaboy D, Schmutz J, Schrader J, Segerman B, Shin H, Siddiqui A, Sterky F, Terry A, Tsai CJ, Uberbacher E, Unneberg P, Vahala J, Wall K, Wessler S, Yang G, Yin T, Douglas C, Marra M, Sandberg G, Van de Peer Y, Rokhsar D (2006) The genome of black cottonwood, *Populus trichocarpa*. *Science* 313 (5793):1596–1604
- Williams EJ, Bowles DJ (2004) Coexpression of neighboring genes in the genome of *Arabidopsis thaliana*. *Genome Res* 14:1060–1067
- Wolfe KH (2001) Yesterday's polyploids and the mystery of diploidization. *Nat Rev Genet* 2:333–341