EST data suggest that poplar is an ancient polyploid

Lieven Sterck[1], Stephane Rombauts[1], Stefan Jansson[2], Fredrik Sterky[2], Pierre Rouzé[3] and Yves Van de Peer[1]

[1]Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Technologiepark 927, B-9052 Ghent, Belgium; [2]Department of Biotechnology, Kungliga Tekniska Hogskolan Royal Institute of Technology, AlbaNova University Center, SE-106 91 Stockholm, Sweden; [3]Laboratoire Associé de l'Institut National de la Recherche Agronomique (France), Ghent University, Technologiepark 927, B-9052 Ghent, Belgium

Author for correspondence:
Yves Van de Peer
Tel: +32 9 331 3807
Fax: +32 9 331 3809
Email: Yves.VandePeer@psb.ugent.be

Received: *28 October 2004*
Accepted: *4 January 2005*

Summary

• We analysed the publicly available expressed sequence tag (EST) collections for the genus *Populus* to examine whether evidence can be found for large-scale gene-duplication events in the evolutionary past of this genus.
• The ESTs were clustered into unigenes for each poplar species examined. Gene families were constructed for all proteins deduced from these unigenes, and K_S dating was performed on all paralogs within a gene family. The fraction of paralogs was then plotted against the K_S values, which resulted in a distribution reflecting the age of duplicated genes in poplar.
• Sufficient EST data were available for seven different poplar species spanning four of the six sections of the genus *Populus*. For all these species, there was evidence that a large-scale gene-duplication event had occurred.
• From our analysis it is clear that all poplar species have shared the same large-scale gene-duplication event, suggesting that this event must have occurred in the ancestor of poplar, or at least very early in the evolution of the *Populus* genus.

Key words: EST (expressed sequence tag) data, evolution, fossil record, genome duplication, K_S dating, polyploidy, *Populus* (poplar).

New Phytologist (2005) **167**: 165–170

© *New Phytologist* (2005) **doi**: 10.1111/j.1469-8137.2005.01378.x

Introduction

The genus *Populus* consists of some 30 species of woody plants including poplars, cottonwoods and aspens, all of which are found in the Northern hemisphere (Taylor, 2002). The genus is further subdivided into six distinct sections: Abaso, Turanga, Leucoides, Aigeiros, Tacamahaca and Populus (Eckenwalder, 1996 and references therein). The latter two sections each contain about a dozen species. Poplar trees have been used all over the world to produce a large variety of wood-based products including timber, pulp, and paper. In addition to their great economical value, poplars are also rapidly becoming the model organism for forest biology and tree biotechnology. They can be easily transformed and vegetatively propagated, and have rapid growth. Another major advantage of the poplars is that they have a modest genome size (≈ 500 Mbp; J. Tuskan, personal communication) organised in 19 chromosomes. It is therefore not surprising that, in 2001, poplar was selected as the first woody plant to have its genome sequenced. Because of its successful use as a model organism for trees, much genomic information and many resources are already available for poplar (for review see Brunner *et al.*, 2004), including several expressed sequence tag (EST) collections. Public EST libraries generally form an unbiased sampling of genes that are expressed in a wide variety of conditions and have proven to be an invaluable source of information (Rudd, 2003).

Blanc & Wolfe (2004) have proposed an elegant method to study the duplication past of genes in an organism's genome based on EST data when a complete genome annotation is not (yet) available. Within the EST data collections of different plant species, they identified paralogous sequences after which the time of duplication was inferred by estimating the number of synonymous substitutions (K_S) between two duplicates. Because synonymous substitutions do not alter the amino acid sequence, they are assumed to be under no or minimal selection, and to accumulate substitutions at a neutral and steady rate. When the number of duplicated pairs of genes is plotted against their age, inferred from the number of synonymous substitutions per synonymous site (K_S), the

resulting age distributions exhibit a typical L-shape, with many recently duplicated genes and much fewer older duplicates. Based on these age distributions, Lynch & Conery (2000, 2003) suggested a steady-state stochastic birth–death model for the dynamics of duplicated gene populations, from which they inferred the overall rate of both gene duplication and gene loss. However, large-scale gene or entire genome duplication events cause a sharp increase in the number of paralogs over a short period, and are visible as peaks in the L-shaped age distributions (Vandepoele *et al.*, 2003; Blanc & Wolfe, 2004; Van de Peer, 2004). Here we have analysed age distributions based on the EST data collections publicly available for poplar, to investigate whether evidence can be found for large-scale gene duplications in its evolutionary past.

Materials and Methods

Construction of the unigene sets

As no 'unigene' data sets are available for poplar (a unigene is the consensus sequence of a cluster of transcripts that represents a unique gene in a genome), we have assembled our own as follows. First, we selected all poplar species for which >10 000 entries exist in the dbEST database (Boguski *et al.*, 1993). Seven different *Populus* species meet this criterion: *Populus alba* × *Populus tremula*; *Populus balsamifera* ssp. *trichocarpa*; *P. balsamifera* ssp. *trichocarpa* × *Populus deltoides*; *P. tremula*; *P. tremula* × *Populus tremuloides*; *P. tremuloides*; and *Populus euphratica*. ESTs for these species were downloaded from the dbEST database. At a later stage we also included nonpublic ESTs from *P. balsamifera* ssp. *trichocarpa*, *P. tremula* and *P. tremula* × *P. tremuloides* to confirm our results (Sterky *et al.*, 2004). Next, the different data sets were screened for low complexity and vector contamination with the program SEQCLEAN (using the UniVec-Core database, ftp://ftp.ncbi.nih.gov/pub/UniVec), after which the cleaned ESTs are assembled into unigenes using the program TGICL. The program was run with default parameters except for the minimal overlap, which we set at 40 bases instead of the default 30. Both programs are available at http://www.tigr.org/tdb/tgi/software.

Defining paralogous relationships

The coding frame and putative coding sequence on our unigenes were determined with FRAMED (Schiex *et al.*, 2003), which uses a hidden Markov model (HMM) to search for the coding potential and has the ability to correct frameshifts. Because no HMM exists to recognize coding sequences in poplar, we used the *Arabidopsis* model provided by the FRAMED program. From all proteins predicted by FRAMED, we discarded all proteins that were shorter than 150 amino acids. Next, for each set of unigenes we performed an all-against-all BLASTP (Altschul *et al.*, 1997). Based on the BLAST results, we defined gene families for each *Populus* species. Two sequences were regarded as paralogs if the aligned region is longer then 150 amino acids and when the sequences showed more than 30% similarity.

K_S-based dating

For each gene family, all members were aligned with each other at the protein level with CLUSTALW (Thompson *et al.*, 1994), after which these alignments were used as a guide to align the corresponding nucleotide sequences. Then all N-containing codons and gaps were removed. Starting from these cleaned alignments, K_S was estimated using a maximum-likelihood approach as implemented in the program CODEML (Goldman & Yang, 1994) which is part of the PAML package (Yang, 1997). Because the program can become trapped in suboptimal optima and therefore produce incorrect K_S estimations, it was run five times and only the K_S with the best likelihood was used for each pair of sequences.

Corrections on number of K_S values

In order to exclude redundant sequences in our initial data sets, we identified all paralogous gene pairs that had both a synonymous and a nonsynonymous substitution rate equal to zero. Next, one of both sequences, preferably a singleton EST, was chosen from each of these pairs and all K_S estimations involving this sequence were discarded for further analysis. The elimination of these pairs will not influence our results concerning the large-scale duplication event, because the fraction of pairs discarded this way is very small, and because such pairs are randomly spread over all K_S bins.

A gene family of n members can be created by at most $n-1$ gene-duplication events. However, the number of possible pairwise comparisons within a gene family is $n(n-1)/2$ and, in particular for large gene families, can thus be considerably larger than the number of gene duplications. Therefore, in order to eliminate redundant K_S values when building the age distribution for duplicated genes, phylogenetic trees were constructed for each gene family using an average linkage clustering algorithm. Starting from each gene as a separate cluster, the clusters with the lowest mean intercluster K_S value were iteratively merged. The splits in the average linkage tree represent the $n-1$ retained duplication events. For each split, the m K_S measurements between the two merged gene clusters were added to the K_S distribution with a weight $1/m$ (Blanc & Wolfe, 2004).

A final cleaning step was performed by excluding all pairs that had a K_S larger than 3.5.

Results

For four out of the six sections of the genus *Populus* there is a representative in our data set (Table 1). The initial number of

Table 1 Numbers of sequences for all poplar species used in this analysis

Species	Section in genus *Populus*	No. initial ESTs	No. unigenes	No. proteins	No. paralogs	No. families	No. paralogs per family	No. used K_S values
Populus alba × *Populus tremula*	Populus	10 446	1 565	1 166	543	180	3.02	351
Populus balsamifera ssp. *trichocarpa*	Tacamahaca	26 825	8 306	4 722	1 701	565	3.01	1 069
Populus balsamifera ssp. *trichocarpa* × *Populus deltoides*	Tacamahaca × Aigeiros	16 480	2 453	1 738	731	245	2.98	463
Populus euphratica	Turanga	13 903	2 162	1 587	528	185	2.85	318
Populus tremula	Populus	31 288	15 737	6 630	2 536	830	3.06	1 546
Populus tremula × *Populus tremuloides*	Populus	65 981	24 159	10 732	4 822	1 479	3.26	3 155
Populus tremuloides	Populus	12 813	2 228	1 479	454	158	2.87	265

downloaded transcripts for each species ranged from ≈ 10 000 for the smallest set to 66 000 for the largest. These sequences were clustered into unigenes for each species separately. The final unigene set consists of the assembled transcripts (contigs) and the singleton transcripts (singlets). For a substantial part of each set we were able to extract a protein longer than 150 amino acids (Table 1). Gene families were defined at the protein level using BLASTP.

Figure 1 shows age distributions for ESTs for the seven different poplar species obtained by plotting the number of paralogs against their K_S values (see Materials and Methods). The high number of paralogs with very low K_S values refers to sequence pairs that result from recent small-scale gene duplications, an ongoing process in most species (Lynch & Conery, 2000; Blanc & Wolfe, 2004). However, as can be clearly observed, for all species there is a sharp increase in the number of sequence pairs with a K_S between 0.20 and 0.30, which means that a large number of gene duplicates must have been created at about the same time. The most plausible explanation for such an observation is a large-scale gene- or even entire genome-duplication event (Van de Peer, 2004). The sudden increase in the number of paralogs is even more evident when we plot the cumulative percentage of paralogs (grey line, Fig. 1). It is clearly seen that there is a sudden increase in the number of K_S values in the lowest K_S bins ($K_S < 0.45$) indicating that, for all *Populus* species, more than half the duplicated genes originated through very recent small-scale gene duplications and one relatively recent large-scale duplication event (Fig. 1). The K_S distribution for higher K_S values ($K_S > 0.90$) again shows a small increase in the number of duplicates. It should be noted that the error on the K_S estimations rises quickly for K_S values > 1 (Li, 1997), and that these values therefore should be interpreted with caution. Nevertheless, we believe that this increase probably reflects older large-scale gene-duplication events early in the evolution of the angiosperms (Simillion *et al.*, 2002; Bowers *et al.*, 2003).

It could be argued that the data sets of hybrid poplars should not be used for studying duplication history in the genus *Populus* by means of K_S dating, because the hybridization event could interfere with the duplication event. However, in general the difference between two alleles is much smaller than the differences between duplicates (not shown). Moreover, we believe our results show that this is not a concern, the strongest argument being the fact that we observe the large-scale duplication peak in all species examined at the same time, regardless of whether the poplar species are hybrids (Fig. 1a,d,f) or not (Fig. 1b,c,e,g).

Discussion

Although the number of EST sequences and paralogs is quite small for some species (Table 1), some general trends can be observed for all collections. For example, for all EST sets, even the smaller ones, the large-scale duplication event can clearly be uncovered. This proves that the approach used is quite robust and suggests that general duplication trends in a genome can already be recognized, even in species for which there is a limited amount of EST data available. For all seven species, a sharp increase in the number of paralogs is observed at a K_S of 0.20–0.30, which points to a large-scale duplication event that has occurred at the same time in all species. As it is most unlikely that there was a duplication event independently in all these species at the same time, we conclude that the duplication event must have occurred very early in, if not before, the evolution of the genus *Populus*.

When K_S values are converted to time ($T = K_S/2\lambda$), using a rate (λ) of 1.5×10^{-8} synonymous substitutions per synonymous site per year, as proposed by Koch *et al.* (2000), the polyploidy event in poplar is estimated to have occurred some 8 myr ago. Using a different rate of 9.1×10^{-9} synonymous mutations per synonymous site per year, as suggested elsewhere (Lynch & Conery, 2000), the large-scale gene-duplication event in poplar is dated at ≈ 13 myr ago. Although

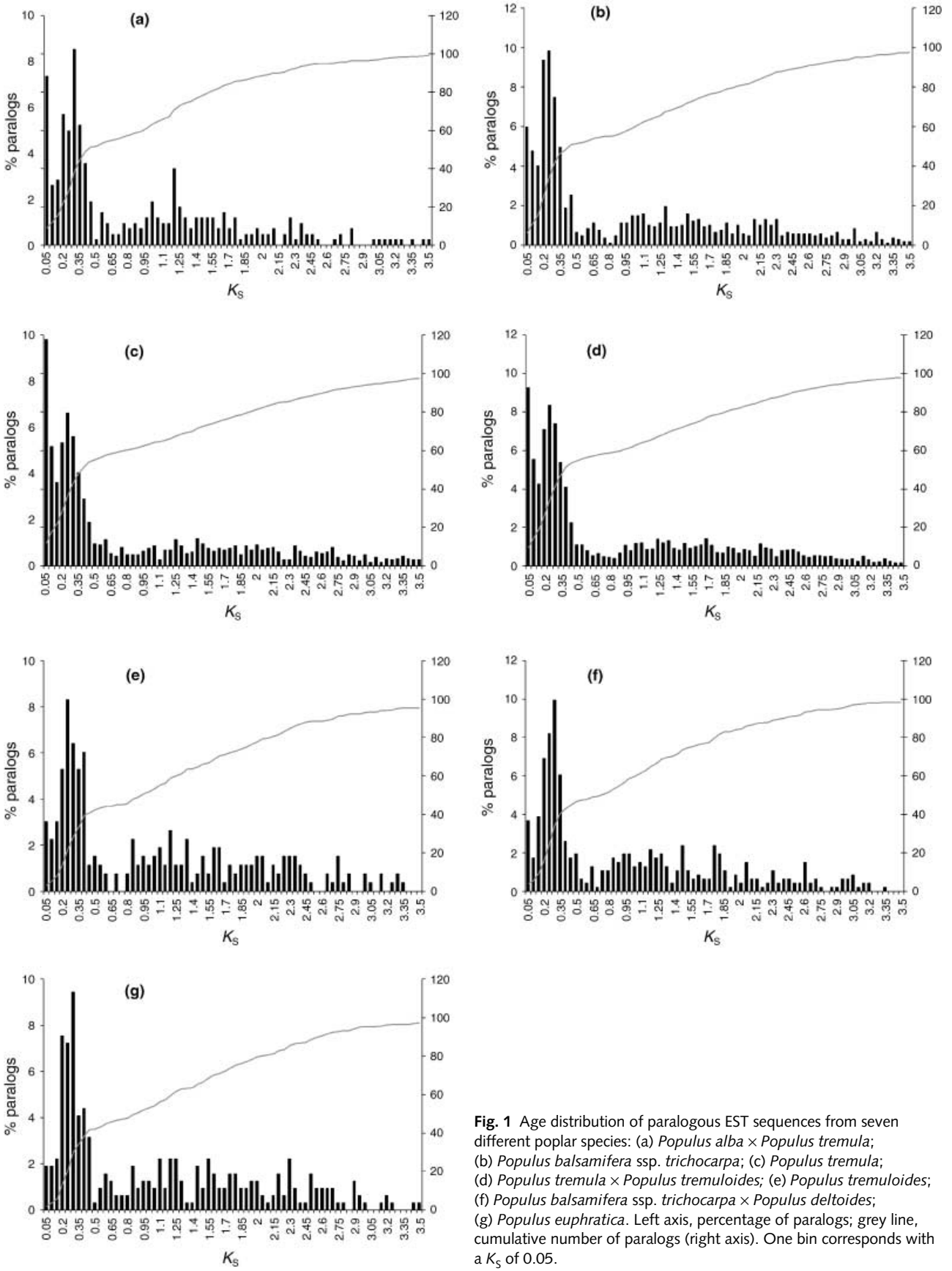

Fig. 1 Age distribution of paralogous EST sequences from seven different poplar species: (a) *Populus alba* × *Populus tremula*; (b) *Populus balsamifera* ssp. *trichocarpa*; (c) *Populus tremula*; (d) *Populus tremula* × *Populus tremuloides;* (e) *Populus tremuloides*; (f) *Populus balsamifera* ssp. *trichocarpa* × *Populus deltoides*; (g) *Populus euphratica*. Left axis, percentage of paralogs; grey line, cumulative number of paralogs (right axis). One bin corresponds with a K_S of 0.05.

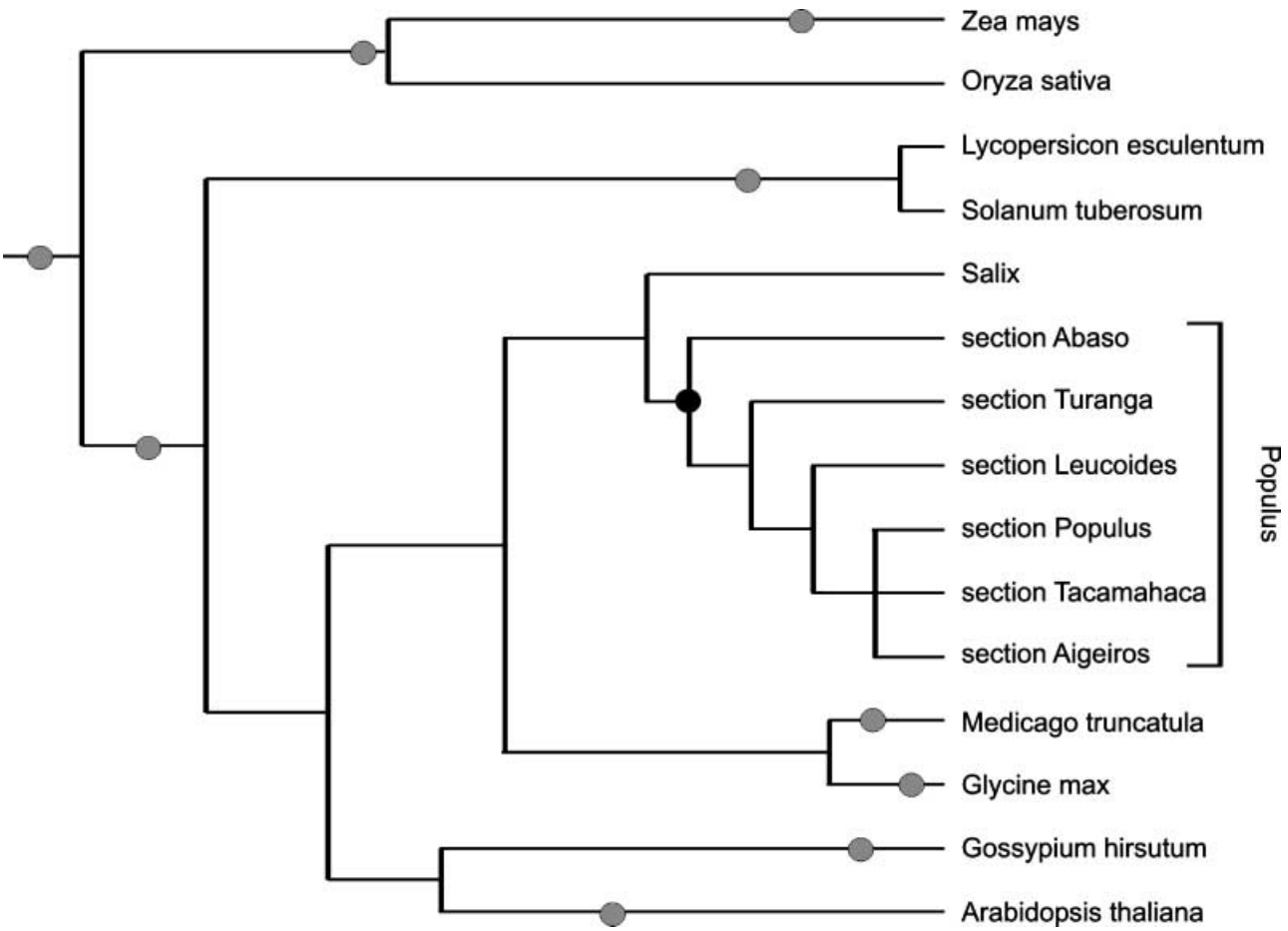

Fig. 2 Schematic representation of the phylogeny of the genus *Populus* and some relevant plant species, based on Blanc & Wolfe (2004); Eckenwalder (1996); Wikström *et al*. (2001). Grey dots indicate large-scale duplication events proposed by Blanc & Wolfe (2004; see also Van de Peer, 2004); a black dot denotes the large-scale duplication event in poplar proposed in the current study.

these dates differ considerably and should be interpreted with caution, it is clear that all the poplar species examined share the polyploidy event. This can be explained only if the genome duplication occurred in the ancestor of all these species, somewhere between 8 and 13 myr ago. This also means that the divergence of the different poplar sections must be more recent than the polyploidy event. The earliest fossils claimed as being of poplar are 58-myr-old leaves ascribed to the section Abaso, which is probably one of the earliest diverging poplar species (Fig. 2; Eckenwalder, 1996), but for which unfortunately no EST data exist. Therefore we cannot conclude for sure whether the Abaso section shares the duplication event. In this respect it would also be interesting to examine the duplication past of other members of the family Salicaceae, such as the sister genus of *Populus*, *Salix*, which is closely related (Leskinen & Alström-Rapaport, 1999; Wikström *et al*., 2001), to see whether they share the same duplication event. Unfortunately too few EST data are available for the other Salicaceae, so this question remains unanswered. The earliest fossil evidence ascribed to the other poplar sections is claimed to be between 18 and 40 myr old (Eckenwalder, 1996 and references therein), which predates the polyploidy event, and is thus clearly in disagreement with our data. There are two possible explanations for this incongruence. The first is that the poplar fossils are not correctly ascribed to the different poplar sections. Alternatively, it is possible that the rate of synonymous substitutions (λ) for poplar is somewhat different than the value generally used for dicots (see above). This is not unlikely considering the fact that the generation time of a species is known to affect its nucleotide-substitution rate (Gaut, 1998) and that poplar has a much longer generation time than most other plant species used in molecular research. Careful calibration of some poplar molecular markers in the future may shed further light on this.

Acknowledgements

We would like to thank Stefanie De Bodt, Steven Maere and Klaas Vandepoele for their help in producing and interpreting

the K_S estimations. We also acknowledge the support of our institution and thank all members of the biocomputing group for helpful discussions. This work was supported by a grant from the European Community, FOOD-CT-2004-506223-GRAIN LEGUMES.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**: 3389–3402.

Blanc G, Wolfe KH. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* **16**: 1667–1678.

Boguski MS, Lowe TM, Tolstoshev CM. 1993. dbEST – database for 'expressed sequence tags'. *Nature Genetics* **4**: 332–333.

Bowers JE, Chapman BA, Rong J, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**: 433–438.

Brunner AM, Busov VB, Strauss SH. 2004. Poplar genome sequence: functional genomics in an ecologically dominant plant species. *Trends in Plant Science* **9**: 49–56.

Eckenwalder JE. 1996. Systematics and evolution of *Populus*. In: Stettler RF, Bradshaw HD Jr, Heilman PE, Hinckley TM, eds. *Biology of Populus and its Implications for Management and Conservation*. Ottawa, Canada: NRC Research Press, National Research Council of Canada, 7–32.

Gaut BS. 1998. Molecular clocks and nucleotide substitution rates in higher plants. In: Hecht MK, Macintyre RJ, Clegg MT, eds. *Evolutionary Biology*, **Vol. 30**. New York, USA: Plenum Press, 93–120.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* **11**: 725–736.

Koch MA, Haubold B, Mitchell-Olds T. 2000. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae). *Molecular Biology and Evolution* **17**: 1483–1498.

Leskinen E, Alström-Rapaport C. 1999. Molecular phylogeny of Salicaceae and closely related Flacourtiaceae: evidence from 5.8 S, ITS 1 and ITS 2 of the rDNA. *Plant Systematic Evolution* **215**: 209–227.

Li WH. 1997. *Molecular Evolution*. Sunderland, MA, USA: Sinauer Associates.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290**: 1151–1155.

Lynch M, Conery JS. 2003. The evolutionary demography of duplicate genes. *Journal of Structural and Functional Genomics* **3**: 35–44.

Rudd S. 2003. Expressed sequence tags: alternative or complement to whole genome sequences? *Trends in Plant Science* **8**: 321–329.

Schiex T, Gouzy J, Moisan A, de Oliveira Y. 2003. FRAMED: a flexible program for quality check and gene prediction in prokaryotic genomes and noisy matured eukaryotic sequences. *Nucleic Acids Research* **31**: 3738–3741.

Simillion C, Vandepoele K, Van Montagu M, Zabeau M, Van de Peer Y. 2002. The hidden duplication past of *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences, 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, Sundberg B, Gustafsson P, Uhlen M, Bhalerao RP, Nilsson O, Sandberg G, Karlsson J, Lundeberg J, Jansson S. 2004. A Populus EST resource for plant functional genomics. *Proceedings of the National Academy of Sciences, USA* **101**: 13951–13956.

Taylor G. 2002. *Populus*: arabidopsis for forestry. Do we need a model tree? *Annals of Botany (London)* **90**: 681–689.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**: 4673–4680.

Van de Peer Y. 2004. Computational approaches to unveil ancient genome duplications. *Nature Reviews Genetics* **5**: 752–763.

Vandepoele K, Simillion C, Van de Peer Y. 2003. Evidence that rice and other cereals are ancient aneuploids. *Plant Cell* **15**: 2192–2202.

Wikström N, Savolainen V, Chase MW. 2001. Evolution of the angiosperms: calibrating the family tree. *Proceedings of the Royal Society of London B* **268**: 2211–2220.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences* **13**: 555–556.