Selection on Nuclear Genes in a *Pinus* Phylogeny

AE Palmé,†§ T Pyhäjärvi,†§ W Wachowiak,†‡§ and O Savolainen†§*

*Department of Evolutionary Functional Genomics, Uppsala University, Uppsala, Sweden; †Department of Biology, University of Oulu, Oulu, Finland; §Biocenter Oulu, University of Oulu, Oulu, Finland; and ‡Institute of Dendrology, Polish Academy of Sciences, Kórnik, Poland

In this study, we investigate natural selection in a pine phylogeny. DNA sequences from 18 nuclear genes were used to construct a very well-supported species tree including 10 pine species. This tree is in complete agreement with a previously reported supertree constructed from morphological and molecular data, but there are discrepancies with previous chloroplast phylogenies within the section *Pinus*. A significant difference in evolutionary rate between *Picea* and *Pinus* was found, which could potentially indicate a lower mutation rate in *Picea*, but other scenarios are also possible. Several approaches were used to study selection patterns in a set of 21 nuclear genes in pines and in some cases in *Picea* and *Pseudotsuga*. The overall pattern suggests efficient purifying selection resulting in low branch-specific d_n/d_s ratios with an average of 0.22, which is similar to other higher plants. Evidence for purifying selection was common and found on at least 55% of the branches. Evidence of positive selection at several sites was found in a *phytocyanin* homolog and significant differences in d_n/d_s among the branches in the gene tree in *dehydrin 1*. Several genes suitable for further phylogenetic analysis at various levels of divergence were identified.

Introduction

Selection has had a strong impact on the morphology and phenology of plants, but the genes involved in these adaptations are still largely unknown. Even in well-studied species such as *Arabidopsis thaliana* and its relatives, the genes responsible for adaptation in natural populations are generally unidentified, even though there are some important exceptions (Le Corre et al. 2002; Caicedo et al. 2004; Stinchcombe et al. 2004; Werner et al. 2005; Kivimäki et al. 2007; Filiault et al. 2008). In conifers, the information is even more limited, and only a few candidate genes for positive directional or balancing selection have been identified so far (Kusumi et al. 2002; Gonzalez-Martinez et al. 2006; Savolainen and Pyhäjärvi 2007; Eveno et al. 2008), even though conifers are important components in many northern ecosystems and valuable forestry trees. Conifer life history may give good opportunities for strong selection. Many species have large distributions and high levels of gene flow among regions resulting in large effective population sizes, which should make selection efficient (Kimura and Ohta 1969; Kimura 1983), even though variation in selection intensity in a heterogeneous environment might make it less so for new, nearly neutral mutations (Ohta 1972).

There is now a large array of tests available for examining nucleotide data for signs of selection (Biswas and Akey 2006). Importantly, different tests can detect selection on different timescales (Garrigan and Hedrick 2003), and thus by using different data sets selection both in the recent and distant past can be identified. In this paper, we will concentrate on tests for selection in the distant past using multispecies data sets. By studying evolution in a phylogenetic framework, selection events can potentially be located to a particular branch of the gene tree (e.g., Yang 1998; Creevey and McInerney 2002) and therefore be roughly timed to a period in the past of the species. In this way, selective events can be put into a larger context and be correlated with important evolutionary events. Multispecies data sets are also useful to identify particular sites under selection (e.g., Huelsenbeck and Dyer 2004; Yang et al. 2005), making it possible not only to identify the gene region under selection but also potentially the actual amino acid change causing the selective advantage. Functional analyses can then be used to verify and further study the effects of these candidate substitutions. For commercially important species such as pines, the genes identified as being influenced by selection may also have importance in tree breeding.

The basis of both branch-specific and site-specific searches for selection is generally the ratio between nonsynonymous and synonymous substitution rate (d_n/d_s) (Yang 1998; Huelsenbeck and Dyer 2004; Yang et al. 2005). Under completely neutral evolution, d_n and d_s are expected to be equal. Purifying selection decreases d_n and therefore also d_n/d_s, whereas positive selection has the potential to produce d_n/d_s ratios above one. However, a test for d_n/d_s above one across a whole gene and long evolutionary time is very conservative, as positive selection is not expected to be constantly acting on all sites of a gene (e.g., Nielsen and Yang 1998; Yang 1998; Liu and Zhu 2008). Therefore, methods that search for selection on specific sites or branches can improve the power to detect selection.

The genus *Pinus* consists of some 110 species and is divided into two subgenera, *Pinus* and *Strobus*, which are in turn each divided into two sections according to the most recent classification (Gernandt et al. 2005). *Pinus* has a rich fossil record, but the interpretation is not always straightforward, leading to diverse hypotheses on when important evolutionary events took place. Most evidence suggests that pines were present at least from the Middle Cretaceous (Millar 1998) and that the split between pines and spruce (genus *Picea*) occurred during the Cretaceous or Jurassic (Magallon and Sanderson 2005; Willyard et al. 2007). Some evolutionary hypotheses suggest that the subgenera *Pinus* and *Strobus* were already present during the Cretaceous (Millar 1998; Eckert and Hall 2006; Willyard et al. 2007), whereas others indicate an Eocene origin (Miller 1973; Willyard et al. 2007). During their evolution, pines have experienced large-scale environmental as well as distributional changes, for example, moving several times between Eurasia and America (Eckert and Hall 2006). Thus, we expect pine species to have been under continuing

Key words: selection, phylogeny, pine, nuclear genes, sequence polymorphisms, gene trees.

E-mail: anna.palme@ebc.uu.se.

Mol. Biol. Evol. 26(4):893–905. 2009
doi:10.1093/molbev/msp010
Advance Access publication January 23, 2009

selection for adaptation to different conditions and to be affected by recurrent demographic effects.

Molecular phylogenetic studies in pines have so far largely relied on chloroplast markers (Wang, Tsumura, et al. 1999; López et al. 2002; Eckert and Hall 2006). However, such markers are linked and therefore cannot provide independent information on the species phylogeny as compared with unlinked nuclear genes. To our knowledge, the only phylogenetic study of pine based on multiple nuclear genes is by Syring et al. (2005), where four low copy nuclear loci were analyzed in 12 pine species and combined with chloroplast and internal transcribed spacer (ITS) data. Several studies indicate that the inclusion of more genes increases the chance of retrieving the correct species tree (Parkinson et al. 1999; Soltis et al. 1999; Rokas et al. 2003) even though it does not resolve all phylogenetic problems (Delsuc et al. 2005; Jeffroy et al. 2006). In addition, gaps introduced in the data set by including genes that are not sequenced in all species or species where all genes are not sequenced should not be problematic, as long as there are enough informative characters in each species and not too limited taxon sampling within each gene (Wiens 2003, 2006). Unfortunately, the large and repetitive genomes of conifers (Kinlaw and Neale 1997) make finding orthologs in different species a demanding task and finding orthologs is crucial not only for correct phylogenetic inference but also in evolutionary analysis.

Here, we investigate the molecular evolution of 21 genes, some of which are candidates for the adaptively important traits cold tolerance and timing of bud set. By comparing orthologs from several *Pinus* and in some cases *Picea* and *Pseudotsuga* species, we examined the long-term evolutionary patterns of these genes and species. Specifically, we wanted to 1) construct a phylogeny based on many nuclear gene sequences, and compare it with earlier phylogenies, 2) compare evolutionary rates leading to the branches of *Pinus* and *Picea*, and 3) investigate the presence of negative and/or positive selection acting on these genes.

Materials and Methods

Genes and Species

Thirteen pine species from different parts of the *Pinus* phylogenetic tree were chosen for this study (table 1). They are divided into systematic groups as follows: 1) Subgenus *Pinus*, section Pinus, Subsection Pinus: *Pinus sylvestris*, *Pinus densiflora*, *Pinus nigra*, *Pinus resinosa*, and *Pinus thunbergii*; Subsection Pinaster: *Pinus pinaster*; section Trifoliae, Subsection Contortae: *Pinus contorta*, *Pinus banksiana*; Subsection Ponderosae: *Pinus ponderosa*; 2) Subgenus *Strobus*, section Quinquefoliae, subsection Strobus: *Pinus peuce*, *Pinus strobus*, *Pinus strobiformis*, and *Pinus lambertiana*. The classification is according to Gernandt et al. (2005), which is similar to that of Price et al. (1998). However, in Price et al. (1998), *P. pinaster* was placed in subsection Pinus, and section Trifoliae, section Quinquefoliae, and subsection Strobus mentioned above are replaced by "new world diploxylon pines," section Strobus, and subsection Strobi, respectively.

Twenty-three genes were selected for this study on the basis of their successful amplification in *P. sylvestris* and the amplification of a single gene product in the other species included here. When the latter was the case, we initially assumed orthology, but this was further investigated with phylogenetic analysis (see below). Gene *eph*, *151*, *207*, *175*, and *phy* have unusually high d_n/d_s ratios in a conifer expressed sequence tag data set (average 0.35), whereas *rpS10* and *rpS4* have a low d_n/d_s ratio (0.01 and 0.04) (Palmé et al. 2008). Dehydrin genes (*dhn1*, *dhn2*, *dhn3*, *dhn7*, and *dhn9*), *abaR*, and *gst2* are candidate loci for cold tolerance (Close 1997; Seppänen et al. 2000; Kalberer et al. 2006) and *a3ip2* and *gi* for timing of bud set as homologues of genes of the *A. thaliana* flowering time pathway have been demonstrated to act on growth cessation in trees (Böhlenius et al. 2006; Gyllenstrand et al. 2007; Ingvarsson et al. 2008). Many of these genes can be considered to have increased probability to be under selection as high d_n/d_s is suggestive of selection (Palmé et al. 2008), and both cold tolerance and growth cessation are important components of local adaptation (Mikola 1982; Morgenstern 1996; Hurme et al. 1997; Savolainen et al. 2007). For information on gene function, see supplementary table S1, Supplementary Material online. Due to varying amplification and sequencing success, not all genes were analyzed in all species (see table 1 for details), but we have chosen to include as many genes as possible, only excluding genes and species with minimal data or genes where we suspect orthology problems or balancing selection (large data set, see table 1). *Picea abies* or *Pseudotsuga menziesii* were mainly used as outgroups.

Most of the genes included in this study have been sequenced in population samples of *P. sylvestris* (see supplementary table S1, Supplementary Material online). In the genes *eph*, *151*, *207*, *175*, *phy*, *rpS10*, *rpS4*, *prof h2b*, and *rpL34* orthology-verified EST contig data from *P. pinaster*, *Pinus taeda*, *Picea glauca*, and *Pseudotsuga menziesii* are also available (Palmé et al. 2008), but for two loci (*phy* and *rpS4b*), the EST contig from *Pseudotsuga* was much shorter than the others and was therefore excluded from the analysis. Because initial analysis indicated that *abaR* was of special interest, a search for additional sequence information was conducted. The *P. sylvestris* sequence was used as a query to search the GenBank EST database with TblastX (http://www.ncbi.nlm.nih.gov/BLAST/). We retrieved the best hit for each of the following species: *P. pinaster* (e-value = $3e^{-60}$), *P. taeda* ($1e^{-57}$), *Picea abies* ($3e^{-31}$), and *Picea glauca* ($4e^{-37}$). The EST sequences are included in the selection analysis but not in the phylogenetic analysis.

Molecular Methods and Sequence Analysis

DNA was extracted from mega-gametophytes (haploid tissue) with a FastDNA Kit (QBiogene, Irvine, CA). As the DNA samples are haploid, it is possible to determine the haplotypes (multilocus combination of polymorphism) by direct sequencing. Polymerase chain reaction and sequencing reactions were performed according to the protocols for *P. sylvestris* described in Pyhäjärvi et al. (2007), Wachowiak et al. (2009), and Palmé et al. (2008). DNA sequences were edited and assembled in *Sequencher* (Gene Codes Corporation, Ann Arbor, MI). Multiple sequence

Table 1
Genes and Species Analyzed

	Genes																						
Species	*eph*	*151*	*207*	*175*	*phy*	*rpS10*	*rpS4a*[a]	*rpS4b*[a]	*rpL34*	*a3ip2*	*gi*	*laccase*	*gic*[b]	*myb-like*[b]	*abaR*	*gst2*	*phyP*	*dhn1*	*dhn2*	*dhn3*	*dhn7*	*dhn9*
Pinus sylvestris	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Pinus nigra	x	x	x	x	x	x	x	x	x		x	x	x	x	x		x	x		x		x
Pinus resinosa	x	x	x	x	x	x		x	x	x	x	x	x	x	x		x	x		x	x	x
Pinus pinaster		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
Pinus contorta	x		x	x		x		x	x		x	x	x	x			x	x	x	x	x	x
Pinus banksiana			x			x					x	x	x	x			x	x	x		x	x
Pinus ponderosa		x	x	x	x	x		x	x	x	x	x	x	x	x		x	x	x		x	x
Pinus peuce	x		x	x		x				x	x	x		x		x	x			x		
Pinus strobiformis	x		x	x		x	x		x	x	x	x		x		x	x					
Pinus lambertiana	x		x	x		x	x			x	x	x		x		x	x			x		
Picea abies			x	x		x		x			x	x	x				x					
Pseudotsuga menziesii			x			x				x	x											
Small data set[c]			Y			Y					Y	Y		Y			Y					
Large data set[c]	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y	
EST seq.[d]	x	x	x	x	x	x	x	x	x						x							
Length (bp)[e]	612	515	332	470	1,067	296	2,151	277	305	787	389	280	586	293	413	324	885	483	364	347	341	814

NOTE.—In a few cases, additional species or genes were analyzed: *abaR* and *phyP* were sequenced in *Pinus densiflora*, *dhn3* in *Pinus thunbergii*, *gst2*, *phyP*, and *dhn3* in *Pinus strobus*, *rpS4b* in *Picea glauca* and *rpS10* in *Larix sibirica*. Genes *h2b* and *prof* were analyzed in *P. sylvestris* and EST sequences are available for these genes.

[a] These sequences represent different parts of the same gene.

[b] Reading frame could not be assigned with confidence, so this gene was therefore not included in the selection analysis.

[c] Y, yes. Indicates that this gene is included in the data set.

[d] EST sequences available for this gene. In all cases except *abaR*, the EST data set was constructed as described in Palmé et al. (2008) and includes EST contigs from *P. pinaster, Pinus taeda, Picea glauca*, and *Pseudotsuga menziesii*, except for gene *phy* and *rpS4b* where the *Pseudotsuga* EST contig was excluded due to short length.

[e] Length of the DNA sequence analyzed for each gene. Coding and noncoding regions included but gaps excluded.

alignment was conducted in ClustalX 1.83 (Thompson et al. 1997) and if necessary edited manually in *BioEdit* 7.0.5.2 (Hall 1999) or *GeneDoc* (2.6.002) (Nicholas et al. 1997). Sequences have been deposited in GenBank, and accession numbers are given in supplementary table S1, Supplementary Material online.

Phylogenetic Analysis

Before including a gene in the phylogenetic analysis, a gene tree was constructed including all sequences from *P. sylvestris*, usually about 40 sequences (Pyhäjärvi et al. 2007; Palmé et al. 2008; Wachowiak et al. 2009), as well as the sequences produced in this study and in some cases groups of orthologous EST contigs (identified by a reciprocal best match strategy, Palmé et al. 2008). Genes that show a pattern suggestive of balancing selection or of the existence of paralogues (as indicated by gene tree structure and in some cases positive Tajima's D in *P. sylvestris*) were excluded from the phylogenetic analysis (genes *rpL34*, *dhn3*, and *dhn9*) because both can produce incorrect species phylogenies. In addition, genes with very low species coverage (*h2b*, *prof*) and species with very low gene coverage were excluded (*P. densiflora*, *P. thunbergii*, *P. strobus*, *Picea glauca*, and *Larix sibirica)*. Genes for which positive selection was inferred (see below) were included in the analysis as this process is not expected to cause errors in the phylogeny (Hang et al. 2003; Hagstrom et al. 2004).

Two data sets were analyzed: One large, including all genes, except those excluded as described above, and one small data set, which is a subset of the former with only genes with good species coverage (see table 1). The large and small data sets include 18 genes (10,865 bp) and 6 genes (2,475 bp), respectively. In the data sets for the individual genes, each species is represented by a single haplotype. For *P. sylvestris*, only the most common allele in the population data was included. Gaps were excluded in the alignment of each individual gene.

The individual genes were concatenated by modifying fasta files (Fasta data set splitter and Fasta data set joiner in FaBox, http://www.daimi.au.dk/~biopv/php/fabox/). *Pseudotsuga* was used as an outgroup in all analyses on the concatenated genes. Neighbor-Joining analysis was conducted in MEGA 3.1 (Kumar et al. 2004) with the Kimura 2-parameter model and bootstrapping with 1,000 replicates. Analyses were run with pairwise deletion of gaps (gaps here representing missing genes, see table 1).

Bayesian analysis was conducted in MrBayes 3.1.2 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003). The input file was created from a fasta file (with a FaBox tool, http://www.daimi.au.dk/~biopv/php/fabox/) and then modified to fit appropriate models and conditions. MrModeltest (Nylander 2004) was implemented to determine which model was appropriate for our data. The GTR + G model (general time reversible with gamma shaped rate variation across sites) was suggested for the large data set and HKY + G (Hasegwa, Kishino, Yano 85 model with gamma-shaped rate variation across sites) for the small data set. Each analysis was run for 1,000,000 generations and sampled every 1,000th generation, and the first 250 samples were discarded ("burn-in") before any inferences were made. Analyses were made both under the assumption of a molecular clock and without.

The maximum likelihood analysis was conducted in PHYLIP 3.65 (Felsenstein 2004). We specified the

transition–transversion ratio to 2.0, base frequencies to be estimated from the data and one category of mutation rates. The input order of the sequences was randomized. Bootstrap value was estimated from 1,000 resampled data sets.

Relative Rate Test

To test the hypothesis that *Picea* and *Pinus* have evolved at different rates, Tajima's relative rate test (Tajima 1993) was applied as implemented in MEGA 3.1 (Kumar et al. 2004). *Pseudotsuga menziesii* was used as an outgroup in all cases, but to avoid species specific effects, three different pine species were used as representatives of *Pinus*: *P. sylvestris*, *P. ponderosa*, and *P. lambertiana*. Two different data sets were analyzed. The first one includes both coding- and noncoding regions in genes *207*, *rpS10*, and *gi* but excludes all EST data. The second data set contains EST data when available and consists of coding regions from genes *eph*, *151*, *207*, *175*, *rpS10*, *rpS4*, *h2b*, *prof*, and *gi*. *Picea abies* is used as a representative of *Picea* when analyzing the first data set and *Picea glauca* in the second, based on the availability of outgroup sequences. All gaps were completely deleted. A *z*-test was used to test whether there was a difference between third-codon position and first + second position in the proportion of substitutions on the *Picea* branch.

Detecting Selection in Gene Trees

Data Sets for Selection Analysis

All genes where the coding regions could be reliably assigned by comparison to annotated genes in GenBank (all except *myb-like* and *gic*), were analyzed individually to detect selection patterns. In the cases where EST data are available, the analysis was conducted both with and without ESTs. If we had both an EST and new sequences from a certain species, only the new sequence was included. The input phylogenetic trees were constructed with Neighbor-Joining (see phylogenetic analysis above) and based on all available sites, coding as well as noncoding but excluding gaps. However, when including EST contigs, the phylogenetic trees were based only on overlapping sequences and thus excluded most noncoding regions. For the analysis conducted in CRANN (Creevey and McInerney 2003), the gene trees were rooted in accordance with the species tree, except for the *dhn3* tree, which was rooted according to the combined tree of dehydrins 3, 5, and 7. The tests for selection were conducted only on the coding regions, and all stop codons and codons with gaps with high frequency were excluded.

Neutral Substitution Test

The neutral substitution test was used with the main purpose to detect negative selection, but this test would also identify very strong positive selection. It tests if there is a difference between the observed ratio of replacement and silent substitution and the expected ratio. A significant result could indicate either positive or negative selection and is the equivalent to testing if d_n and d_s are significantly different. We used this test as it is implemented in CRANN 1.04 (Creevey and McInerney 2003). To get a picture of the overall pattern of selection, the analysis was also done on the concatenated sequences from all genes. The best tree (see below) was used as an input tree.

Relative Rate Ratio Test

To test for adaptive evolution, we used the relative rate ratio (RRR) test of Creevey and McInerney (2002) as it is implemented in the program CRANN 1.04 (Creevey and McInerney 2003). This method is an extension of the McDonald–Kreitman test adapted to phylogenetic trees. It is based on the comparison of replacement-invariable (RI) and replacement-variable (RV) sites to silent-invariable sites (SI) and silent variable (SV) at each internal branch. Under neutrality, the ratio of RI and RV should be the same as SI/SV, but selection can change this relationship. One of the advantages of this method is that directional and nondirectional positive selection can be separated from each other. Directional selection refers to the situation when a new mutation replaces the previous one, with no further substitutions, whereas under nondirectional selection an amino acid site change repeatedly during the evolution of a protein, for example, in response to a changing environment (Creevey and McInerney 2002). High levels of RI compared with expectations would indicate directional selection, whereas high levels of RV would indicate nondirectional selection. The significance is estimated using a Fishers exact test or a *G*-test when Fishers exact test becomes too computationally intensive.

PAML Analysis

The codon-based models (Goldman and Yang 1994) implemented in "codeml" of PAML (Yang 1997) were used to examine what model of evolution best explains our data and to study the possible effects of positive natural selection on these genes. Equilibrium codon frequencies were estimated from the average nucleotide frequencies at three codon positions (F3X4). Parameters (e.g., d_n/d_s, κ, and sequence divergence) were estimated by maximum likelihood and likelihood ratio tests (LRTs) were used to compare models. Gene trees constructed from the sequences in each gene were used as input tree topology.

Branch Models. For each gene, two branch models were fitted with maximum likelihood: a model with one d_n/d_s ratio, ω_0, for the whole tree (null model) and a model with free d_n/d_s for each branch (free-ratio model). The LRT was conducted to evaluate if the null model can be rejected. Strictly speaking, this is not a test for selection but for investigating whether d_n/d_s varies significantly among the branches. The branch-specific d_n/d_s values estimated in the free-ratio model were also used to investigate the frequency distribution of d_n/d_s in our data set. This analysis was conducted both on the full data set and on a subset where all ratios with d_s below 0.05 were excluded as low divergence leads to uncertain estimates. This particular cutoff (0.05) was chosen to make our results comparable with those of Roth and Liberles (2006).

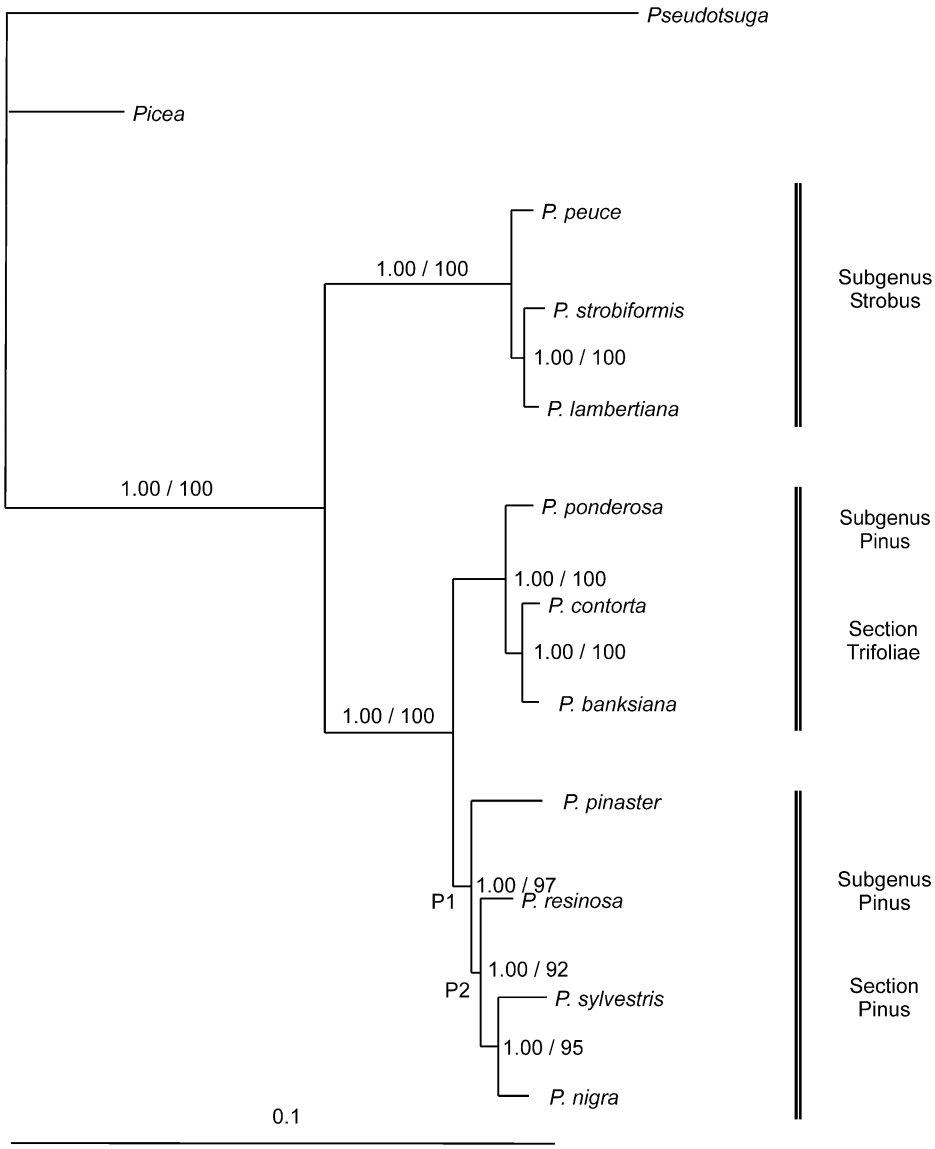

FIG. 1.—Bayesian estimate of the phylogeny based on the large data set (18 loci) without the assumption of a molecular clock. Numbers by the branches represent posterior probabilities of each clade/bootstrap values (%) from the maximum likelihood analysis on the large data set.

In addition to single gene analysis, three members of the dehydrin gene family were analyzed jointly to study if the lineage of *dhn5* is evolving differently from *dhn3* and *dhn7*. The genes *dhn3* and *dhn7* did not form separate monophyletic groups, and therefore, only the *dhn5* lineage was investigated. The null model was compared with a two-ratio model where *dhn3* and *dhn7* had the same d_n/d_s (ω_0) and *dhn5* had a different one (ω_5).

Site-Specific Models. Site-specific models were fitted to the data from all genes to test whether positive selection has acted on specific codons. Two pairs of models were tested: M1a against M2a and M7 against M8. In the nearly neutral M1a, a proportion p_0 of sites have $d_n/d_s < 1$ and $1 - p_0$ have $d_n/d_s = 1$. In M2a, there is an additional site class p_2 where $d_n/d_s > 1$. In M7, d_n/d_s follows a beta distribution with parameters p and q and varies between 0 and 1. M8 is similar to M7 but has additional class where $d_n/d_s > 1$. Posterior probabilities of the mean d_n/d_s and for a specific site to have $d_n/d_s > 1$ were calculated by Bayes empirical Bayes procedure implemented in PAML (Yang et al. 2005).

Multiple Testing Correction

It is not clear how multiple testing within a tree should be corrected for in the neutral substitution and RRR tests, and these methods have so far largely been applied without correction (Carginale et al. 2004; Guillet-Claude et al. 2004; Mes and Stal 2005; Borrelli et al. 2006). We have not only the issue of multiple testing within a tree but also have to consider testing of many gene trees. To correct for multiple testing problems, a false discovery rate (FDR) (Benjamini and Hochberg 1995) approach was chosen. The method has been applied earlier on site-specific tests, for example, in a study on phylogeny of 12 *Drosophila* genomes

Table 2
Summary Table of Gene Function, Divergence, d_n/d_s (ω_0), and Selection Tests for Each Gene

Gene	Hypothetical Function	K (JC)[a]	No. Subst.[b]	ω_0[c]	Tests for Selection[d]
Eph	Epoxide hydrolase	0.02/0.05	68	0.34	Neg*
151	Unnamed gene	0.01/—	44	0.23	Neg*
207	Unnamed gene	0.02/0.06	48	0.13	Neg*
175	Unnamed gene	0.01/0.04	94	0.25	Neg*, Dir
Phy	Phytocyanin	0.03/—	148	0.80	Site*
rpS10	Ribosomal protein S10	0.02/0.03	46	0.02	Neg*
rpS4a	Ribosomal protein S4	—/0.11	32	0.05	Neg*
rpS4b	Ribosomal protein S4	0.00/—	18	0.13	Neg*
Prof	Profilin	—/—	21	0.00	Neg*
h2b	Histone	—/—	56	0.25	Neg*
rpL34	Ribosomal protein L34	0.02/0.10	20	0.00	Neg*
a3ip2	ABI3-interacting protein 2	0.02/0.07	26	0.17	Neg*
Gi	Gigantea	0.01/0.06	76	0.64	—
Laccase	Laccase	0.03/0.09	60	0.33	Neg*
gic[e]	Not known	0.04/—	n	n	n
Myb-like[e]	Myb transcription factor -like	0.02/0.05	n	n	n
abaR	Abscisic acid—responsive protein	0.05/—	69	0.93	Site
gst2	Glutathione-S-transferase	—/0.08	15	0.94	—
phyP	Phytochrome P	0.01/0.05	62	0.32	Neg*
dhn1	Dehydrin	0.05/—	49	0.38	Neg* (Branch)*
dhn2	Dehydrin	0.05/—	17	0.54	—
dhn3	Dehydrin	0.01/0.01	10	0.30	—
dhn7	Dehydrin	0.07/—	23	0.51	—
dhn9	Dehydrin	0.04/—	41	0.43	Neg*, Non-dir

NOTE.—All values are given for the data sets including ESTs, when those were available (table 1). An "n" indicates that a test or a calculation has not been conducted.

[*] Indicates significance after multiple testing correction ($Q < 0.05$). For details see supplementary tables S4, S5, and S6, Supplementary Material online.

[a] Total divergence with Jukes and Cantor correction between *P. sylvestris* and Section Trifoliae/Subgenus *Strobus*.

[b] The number of substitutions above the most basal branch estimated with CRANN.

[c] Total d_n/d_s for the entire gene tree (ω_0) estimated with PAML according to the null model

[d] The results of the selection tests. "Neg" indicates significant negative selection according to the neutral substitution test, "Dir" and "Nondir" indicate significant positive directional and nondirectional selection according to the relative rate ratio test. "Site" suggests significant selection on some sites and "Branch" significant differences in d_n/d_s among branches according to the analysis in PAML. Observe that the branch test is not strictly speaking a test for selection.

[e] Reading frame could not be assigned with confidence so this gene was therefore not included in the selection analysis.

(Drosophila 12 Genomes Consortium 2007). In FDR analysis, q-values corresponding to each p-value are calculated. The q-value of a test is the minimum FDR when that particular test is considered significant. Also p_0, which is an overall proportion of true null hypotheses among all cases, is reported. To calculate q-values and p_0, the software QVALUE (Storey and Tibshirani 2003) was used. All nonobserved p-values were set to 1. FDR analysis was conducted separately for the p-values obtained from neutral substitution, RRR, site-specific and branch-specific tests.

Results

Species Trees

The topologies generated by different data sets and phylogenetic methods are identical or very similar. The same typology is retrieved from the small data set independently of the phylogenetic inference method and the large data set supports this topology both when the Bayesian method (both with and without a molecular clock) or the maximum likelihood method is used. This will be referred to as the best topology (fig. 1). The Neighbor-Joining tree based on the large data set is different from the best topology within Section Pinus, where *P. pinaster* and *P. resinosa* were clustered with *P. sylvestris* as a sister group and the bootstrap values within this group were low (data not shown).

Overall, most branches in the best topology are strongly supported, independently of the phylogenetic inference method or data set: In the Neighbor-Joining tree, based on the small data set all branches are supported by bootstrap values above 98%, except for the branch leading to the group containing *P. resinosa*, *P. nigra*, and *P. sylvestris* (branch P2 in fig. 1), which has a bootstrap value of 74%. With the Bayesian method, the analysis of the large data set resulted in a tree where all clades had a posterior probability of 1.0 (fig. 1), whereas the small data set tree had two branches with lower support (branch P1 0.98 and branch P2 0.92). With the maximum likelihood method, both the small and the large data set produced a tree where all branches except one had bootstrap values above 95%. The deviating branch (P2 in fig. 1) had a bootstrap value of 92% in the large data set tree (fig. 1) and only 61% in the small data set tree.

Relative Rate Test

Independently of which data set is used and what species is chosen to represent the genera, there is a significant difference in the rate of nucleotide substitution between the branch leading to *Picea* and the branch leading to *Pinus* ($p < 0.01$ in all cases). (See supplementary table S2, Supplementary Material online, for net genetic distances between *Pseudotsuga*, *Picea*, and *Pinus*.) In each analysis, there are fewer mutations along the *Picea* branch than on the *Pinus* branch (supplementary table S3, Supplementary Material online). The data set without ESTs suggests

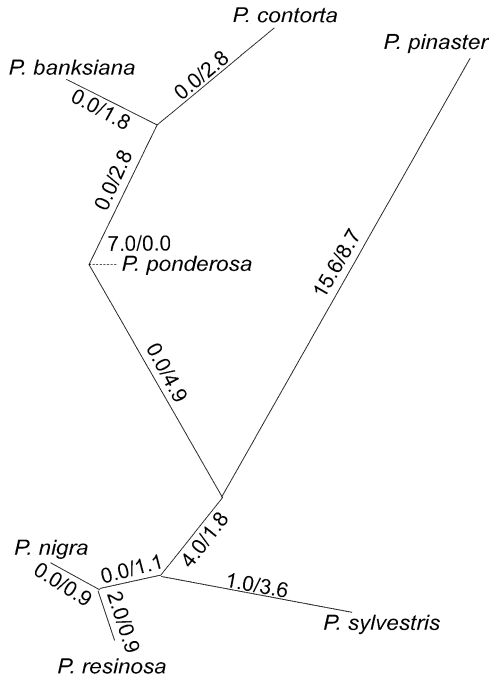

FIG. 2.—Unrooted neighbor-joining tree for *dhn1* based on synonymous sites divergence (numbers are predicted non-synonymous/synonymous changes in each branch). Branch leading to *P. ponderosa* is stretched for visualization; the length of the external branch is zero.

that the rate on the former branch is around 43–45% of the rate of the latter, whereas the data set including ESTs indicates 51–65%. A separate analysis was conducted on third codon position and jointly on first and second codon positions. The same trend of slow evolution along the *Picea* branch was evident both in third and first + second codon positions and the rate difference was frequently significant (*P* for third/first + second codon position, respectively: *Picea glauca* (EST contig) − *P. sylvestris* 0.041/0.064,− *P. ponderosa* 0.033/0.034, − *P. lambertiana* 0.002/0.369; *Picea abies* − *P. sylvestris* 0.086/0.041, − *P. ponderosa* 0.086/0.023, − *P. lambertiana* 0.008/0.117). There was no significant difference between the third codon position and first + second position in the proportion of substitutions on the *Picea* branch compared with the *Pinus* branch ($p > 0.05$ in all cases).

Detecting Selection in Gene Trees

Neutral Substitution Test

Negative selection was found in most of the genes on which the neutral substitution test was conducted (table 2) and in many cases on several branches (supplementary table S4, Supplementary Material online). Negative selection was detected most frequently on the most basal branch or branches of the trees. However, this pattern might very well be caused by lack of power as the outer branches have few substitutions and therefore low power to detect selection. The analysis of the concatenated data set indicates an overall pattern of negative selection on all the branches except the branch leading to *P. lambertiana* and *P. strobiformis*, which had few substitutions.

There was no evidence of negative selection in seven of the genes. For some genes such as *dhn2*, *dhn3* and *dhn7*, and *gst2* this could potentially be a problem of low power as they have a low number of substitutions occurring in the gene tree (table 2). However, genes *phy*, *gi*, and *abaR* have a large number of substitutions above the basal node (table 2).

RRR Test

In most of the genes, the RRR test indicated no significant pattern of positive selection on any branch of the gene tree (see supplementary table S4, Supplementary Material online, and table 2). In two cases, genes *175* and *dhn9*, the initial analysis suggested selection, but those were not significant after correction for multiple testing. In gene *175*, the ratio between replacement invariable and variable (RI/RV) was much higher than the ratio between silent invariable and variable (SI/SV) on the branch leading up to the *Pinus* subgenus, which could suggest positive directional selection. The opposite pattern was seen in *dhn9* on the branch leading to Subgenus *Pinus* Section Pinus, which could indicate nondirectional selection.

PAML Analysis

Branch Models. There was only one gene, *dhn1*, where d_n/d_s varied significantly among branches ($2\Delta l = 33.92$, df = 10, *p*-value = 0.0002, *q*-value = 0.004; fig. 2). In most genes, there was no significant difference in d_n/d_s among branches (table 2, supplementary table S5, Supplementary Material online). Similarly, in the combined analysis of *dhn*3, *dhn*5, and *dhn*7, we found no significant difference in d_n/d_s between the *dhn5* clade and the clade including both dehydrins 3 and 7. Some loci had surprisingly large ω_0 (d_n/d_s across the whole tree according to the null model), for example, 1.87 in *abaR* or 1.08 in *eph*. In both these cases, ω_0 was smaller (see table 2) when also ESTs were included. Average transition–transversion ratio was 2.4.

Site-Specific Models. Two pairs of site-specific models were tested: M1a against M2a and M7 against M8, but the comparisons gave similar results, so only the latter is reported here. In most genes, we found no evidence of positive selection (supplementary table S6, Supplementary Material online). The initial analysis indicates that in two genes (*abaR* and *phy*), models that allow some sites to be under positive selection fitted the data better than models that do not. However, only the latter was significant after correction for multiple testing (table 2, supplementary table S6, Supplementary Material online) when ESTs were included. In *phy*, we identified four sites that were positively selected with posterior probability higher than 0.95 and d_n/d_s values ranging from 6.1 to 6.3 (supplementary fig. S1, Supplementary Material online).

False Discovery Rate Analysis

Q-values for individual tests are reported in supplementary tables S4, S5, and S6, Supplementary Material

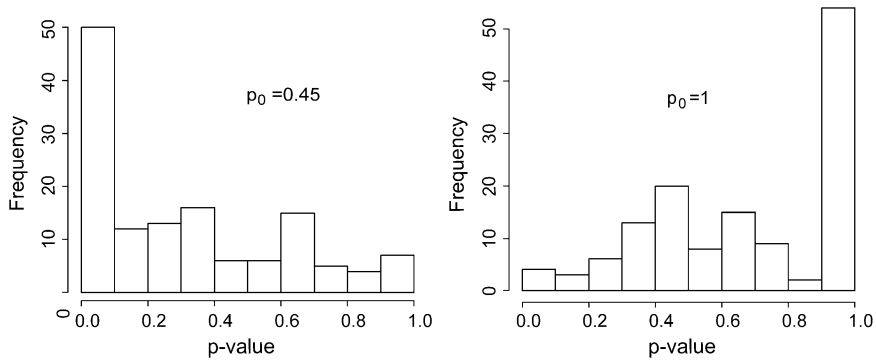

FIG. 3.—Distribution of *p*-values of 134 branches tested for negative (neutral substitution test, on the left) and positive selection (relative rate ratio test, on the right). p_0 is the proportion of true nulls among all tested cases.

online. For all tests for positive selection (PAML and RRR) p_0, the overall proportion of true null hypotheses, was 1, but for neutral substitution tests, p_0 was 0.45 (fig. 3). This suggests that there has been very little positive selection but that negative selection has been acting on at least 55% of the branches studied.

Frequency Distribution of d_n/d_s

To study the frequency distribution of branch-specific d_n/d_s in the whole data set, the free-ratio model was used to estimate d_n/d_s on each branch in each gene tree. Altogether 302 branches were analyzed. Most branches have a low d_n/d_s, and the class between 0 and 0.1 is the largest (fig. 4). Excluding branches with d_s below 0.05, the mean value of d_n/d_s over all branches and genes was 0.22, or 0.20 if the dehydrins were not included. Values ranged from 0 to 0.75. Without the 0.05 cutoff, the mean is 0.23 (with dehydrins), and values range from 0 to 1.6.

The distribution of site-specific d_n/d_s was estimated with the M7 model in PAML, where d_n/d_s was allowed to vary between 0 and 1 (supplementary fig. S2, Supplementary Material online). In eight of the 20 loci, the distribution was L-shaped (*rpS4a*, *rpS10*, *rpL34*, *h2b*, *prof*, *207*, *a3ip2*, and *phyP*), but three of these loci did not have any nonsynonymous changes, and the parameters of beta distribution reached their limiting values (*rpL34*, *h2b*, and *prof*). Eleven genes have a U-shaped distribution (*151*, *phy*, *gst2*, *abaR*, dehydrins 1, 2, 3, 7, 9, *gi*, and *laccase*), and three have intermediate distributions (*rpS4b, eph*, and *175*).

Discussion

Phylogenetic Analysis of Species Trees

The phylogeny on the large data set is based on 18 different genes and produces a single well-resolved tree (fig. 1) both with the Bayesian and the maximum likelihood method. Independent of phylogenetic method, the same tree is also produced with the small data set, which has no missing data within *Pinus*, even though some branches are less well supported. The correspondence between trees from the large and small data sets and with the tree of Grotkopp et al. (2004) (see below) suggests that the missing data have not affected the phylogeny negatively when using the Bayesian or the maximum likelihood methods. The inclusion of more genes has rather increased the support of the tree, even though there are quite many missing data. The Neighbor-Joining method is as expected more sensitive to missing data.

The best phylogeny (fig. 1) confirms many of the main features of published phylogenies and traditional taxonomic groups. The two main groups within *Pinus* correspond to the two subgenera, *Strobus* and *Pinus*. Monophyly of these groups have been reported for phylogenies based on chloroplast markers (Wang, Tsumura, et al. 1999; Eckert and Hall 2006), ITS (Liston et al. 1999) as well as for nuclear genes (Syring et al. 2005). *Pinus* is in turn divided into two well-supported monophyletic groups (fig. 1) that represent the sections Pinus and Trifoliae (Gernandt et al. 2005). The placement of *P. pinaster* in the same group as the others in section Pinus was uncertain in the ITS phylogeny (Liston et al. 2003) but strongly supported both by chloroplast phylogenies (López et al. 2002; Eckert and Hall 2006) and our tree. The relationships within the subgenus *Strobus* are in complete agreement with the chloroplast phylogeny (Eckert and Hall 2006) as is the structure within section Trifoliae but not within section Pinus. In the chloroplast phylogenies, *P. nigra* and *P. resinosa* form a monophyletic group to which *P. sylvestris* is a sister group (López et al. 2002; Eckert and Hall 2006) but Eckert and Hall (2006) point out that this structure results in an impossible dispersal scenario between North America and Eurasia. The best phylogeny suggests that *P. sylvestris* and *P. nigra* are the most closely related and that *P. resinosa* is their sister group (fig. 1). This relationship is supported by a supertree constructed from morphological and molecular data (Grotkopp et al. 2004), which is in fact identical to our tree. As the phylogeny of Grotkopp et al. (2004) and the current phylogeny are both based on multiple but independent data sets, this lends strong support to the results and resolves some incongruence among previous phylogenies based on fewer markers.

The weakness of the phylogenetic analysis presented here is the low number of species. Even though this should not have a large effect on the accuracy of the tree (Rokas and Carroll 2005), it limits the information that can be gained from it. Wiens (2006) suggested a strategy using genes with low evolutionary rates to produce a

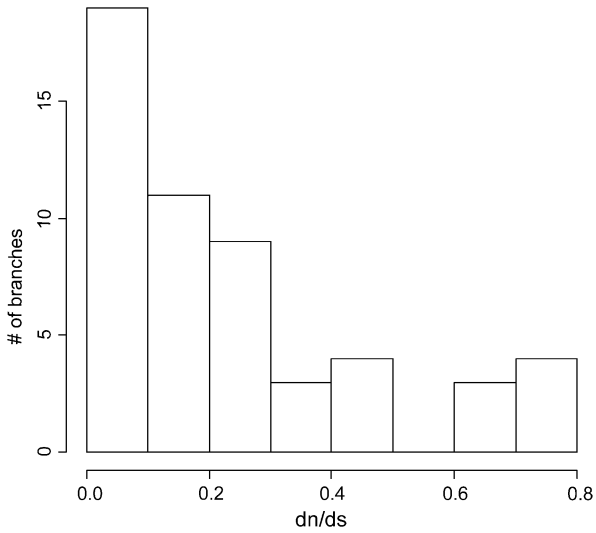

FIG. 4.—The frequency distribution of 53 branch specific nonsynonymous divergence to synonymous divergence ratios (d_n/d_s) for all branches with synonymous divergence (d_s) larger than 0.05. The free-ratio-model was used to estimate branch-specific d_n/d_s.

"phylogenetic scaffold" and then adding genes with higher evolutionary rates to resolve shorter branches. The genes used here were all first analyzed in *P. sylvestris*, and then sequencing in other species was attempted. This has resulted in a larger sequencing success in the species most closely related to *P. sylvestris*, but there are also a large number of genes that were analyzed in species from Subgenus *Strobus* (see table 1), making them useful as "scaffolding genes." The genes included in the small data set show potential for this purpose as they were easy to sequence in all the pine species. To add resolution within sections and subsections, additional genes with high divergence would be useful. The highest divergence levels within Subgenus *Pinus* are found in three dehydrins (table 2), but working with members of a gene family can be difficult. More promising candidates for phylogenetic studies would be *abaR*, *gic*, *laccase*, and *phy*, which show the highest levels of divergence excluding dehydrins. In addition *rpS4a*, *laccase*, or *gst2* show high levels of intersubgenus divergence (see table 2).

Evolutionary Rate Slower in *Picea* Than in *Pinus*

We found a significant difference in the rate of nucleotide evolution between the branch leading to *Picea* and the branches leading to *Pinus*. Earlier studies on a limited number of genes indicate similar differences, even though the differences in evolutionary rates were not statistically tested. Gene trees based on both chloroplast genes and on a nuclear gene have shorter branches leading to *Picea* than to *Pinus* (Wang, Tsumura, et al. 1999; Wang et al. 2000; Eckert and Hall 2006). It is difficult to determine conclusively from our data if it is *Picea* or *Pinus* that displays a divergent pattern, compared with other conifer genera. Judging from branch lengths (fig. 1), *Picea abies* seems to have a reduced evolutionary rate, but the opposite, with increased rates in *Pinus*, is indicated in some studies (Wang et al. 2000; Eckert and Hall 2006). Whether this pattern is general to *Picea* or if it is particular to *Picea abies* and *Picea glauca* is also difficult to determine, but the patterns found by Wang et al. (2000) for other *Picea* species suggest a general trend within this genus. The rate of substitution does not differ between first + second and third codon positions, which would be consistent with a neutral mutation hypothesis, rather than a change in the efficiency of selection. *Picea abies* (Heuertz et al. 2006) also seems to have lower synonymous nucleotide diversity than most pine species (Savolainen and Pyhäjärvi 2007), in accordance with the possibly lower rate mutation in the branch leading to *Picea*.

Selection Patterns

The most common type of selection found in our data set is negative selection (table 2 and supplementary table S4, Supplementary Material online). FDR analysis suggests that as much as 55% of all analyzed branches (fig. 3) have been under negative selection. Estimates of d_n/d_s for whole phylogenies (table 2) or branches (fig. 4) are generally well below 1. This is in accordance with our expectations as most mutations in functional genes are expected to be disadvantageous. The average branch-specific d_n/d_s was 0.22 if a cutoff value of 0.05 for d_s was used. This is very similar to the average found in a study of over 4,000 gene families in higher plants using the same cutoff value (0.21; Roth and Liberles 2006) and in a comparison between *A. thaliana* and *A. lyrata* (0.21; Barrier et al. 2003). However, it is slightly higher than in a comparison between *A. thaliana* and *Brassica rapa* (0.14; Tiffin and Hahn 2002) and a conifer EST data set where branch-specific estimates for an internal branch, the branch leading to *Picea glauca*, and the branch leading to *Pseudotsuga menziesii* were 0.12, 0.14, and 0.15, respectively, after applying a 0.05 d_s cutoff (Palmé et al. 2008). The active choice of genes with an increased probability to be under positive selection could have increased the d_n/d_s estimates in our data set compared with other randomly selected genes and in addition five genes were specifically chosen for their high d_n/d_s. On the other hand, the EST data set could potentially be enriched in genes with low d_n/d_s, because highly expressed and broadly expressed proteins, which should dominate the EST data set, tend to evolve more slowly than other genes (Zhang and Li 2004; Drummond et al. 2005; Popescu et al. 2006).

In the frequency distribution of branch-specific d_n/d_s, the size class with the lowest d_n/d_s is the largest (fig. 4), which is also found in the species pair comparisons in Brassicaceae (Tiffin and Hahn 2002; Barrier et al. 2003). This pattern is more pronounced when the d_s cutoff is not used (data not shown) because a low d_s is often accompanied with a zero d_n. The frequency distribution is similar to that of Roth and Liberles (2006), but the peak of d_n/d_s is further to the left, indicating a higher frequency of very low d_n/d_s values and a somewhat fatter tail in our data. This would be consistent with efficient selection in large conifer populations, but given the limited number and nonrandom choice of genes, it is difficult to draw definite conclusions.

The high d_n/d_s in some genes could indicate relaxed purifying selection or a combination of negative selection and positive selection acting during some time periods or on some sites. *Phy* has high d_n/d_s and significant positive selection at some sites (table 2), suggesting that positive selection has played a role in increasing d_n/d_s. In the other genes with high d_n/d_s or absence of significant negative selection, we find no evidence of positive selection, suggesting simply relaxed purifying selection. In accordance, the frequency distribution of site-specific d_n/d_s in many genes was U-shaped (supplementary fig. S2, Supplementary Material online), suggesting the presence of sites under neutral or nearly neutral evolution. However, the power of the tests varies greatly among the data sets, as it depends on the number of substitutions (see table 2 for the number of substitutions in each gene tree) making it difficult to compare genes.

Even if negative selection is the dominating force, there is some evidence for the presence of positive selection from tests on individual genes (table 2). Among the 21 genes analyzed, initial tests suggests that there are four genes with positive selection (19%), but after correction for multiple testing, there is only one significant gene left (5%). It is not straightforward to compare the proportion of genes under selection in different studies, as different tests and types of data sets are used, genes are generally not randomly chosen, and there is only a limited number of studies on conifers. Only tentative comparisons can therefore be made. The proportion of genes found to be under selection within species varies extensively among studies (0–38%), but the upper estimates are most likely overestimates (Wright and Gaut 2005). In a review on selection in forest tree species, 22 of 151 genes (15%) were found to be under positive or balancing selection, and the percentage in the conifer subgroup was even lower (11%) (Savolainen and Pyhäjärvi 2007). In Cupressaceae, site-specific models identified two candidate genes among 11 analyzed (18%) without correction for multiple testing (Kusumi et al. 2002), which is surprisingly similar to our estimate.

As we are investigating selection patterns during a long evolutionary time period, we would expect to find a higher proportion of genes under positive selection than studies concentrating on one species. However, the power to detect selection is very low at the tip of the branches due to low divergence between species, and in practice, we are only efficiently testing the most basal branches in each gene tree. On the other hand, our sample of genes is not a random one and should be enriched in genes with positive selection, which would increase the estimate. This would also be true for several of the studies reviewed in Savolainen and Pyhäjärvi (2007). It is clear that more extensive multigene studies with randomly chosen genes are needed both in conifers and other plant species to reliably quantify the frequency of positive selection.

The Phytocyanin Homolog: A Candidate for Positive Selection

For *phy*, the site models indicate that some sites have been under positive selection during its evolution (table 2, supplementary table S6, Supplementary Material online). This is true only when the data included ESTs, which could indicate either that the power was increased due to a larger number of polymorphisms or absence of positive selection in the timescale covered by the data from *Pinus*. The two hypotheses are difficult to separate because we did not get positive results in the RRR test, which could have given information on where in the phylogeny selection occurred.

The best Blast hit for the *P. sylvestris* sequence is a *P. taeda* phytocyanin homolog (AAF75824), which has a DNA sequence identical to the *P. taeda* EST contig of this study. In *P. taeda*, the phytocyanin homolog is expressed in embryo, needles, roots as well as stem but not in cones (Expression profile, NCBI). The precise function of *phy* is unknown, but its close similarity to phytocyanin and other blue copper proteins indicates that it functions as an electron transporter. The four amino acids that are part of the copper-binding site (Zhang et al. 2000) are all conserved among the species studied here, further suggesting conservation of that function. The sites identified as being under positive selection by PAML are all located in the cell-wall structural domain as identified by Zhang et al. (2000) and within 26 amino acids of each other (supplementary fig. S1, Supplementary Material online). The identification of candidate sites for selection opens up for functional or association studies that could help to identify the phenotypic and fitness effects of different alleles in this locus.

Dehydrin 1: Variation in d_n/d_s among Branches

We identify significant differences in d_n/d_s among the branches in the *dhn1* gene tree (table 2, supplementary table S5, Supplementary Material online). This could either be caused by differences in the amount of selective constraint among branches or alternatively d_n/d_s could be increased in some cases due to positive selection. The site-specific models with positive selection did not fit the data better than neutral or nearly neutral models. We also found no support for positive selection with the RRR test (table 2), whereas purifying selection is clearly occurring at least on some branches (table 2, supplementary table S4, Supplementary Material online). However, the power to detect selection on the outer branches is low, so this cannot be used to rule out positive selection.

A branch that could potentially cause the deviation is the branch going to *P. ponderosa* as it has seven nonsynonymous changes and no synonymous changes (fig. 2). Interestingly, two of the replacements were found also in *P. pinaster* but not in the rest of the species (supplementary fig. S3, Supplementary Material online). This is surprising, because the two species are relatively distant in the gene tree. There are several possible explanations for the observed pattern, such as lineage sorting or recombination between members of different gene families, but the distribution of substitutions makes the latter unlikely. Another possibility is that positive selection has affected these particular substitutions and thus increasing the probability of fixation of independent mutations. The two nucleotide changes are the easiest way to change from Val to Ala (one transition) and the site model M8 indicates that they are among the four amino acid sites with the highest d_n/d_s,

making them potential candidates for positive selection. Two other dehydrin genes have been identified as under positive selection in *P. pinaster* (Eveno et al. 2008), suggesting that dehydrins might be frequently under selection.

Conclusion

The first stage of this study has been the identification and sequencing of several genes in many conifer species. Identifying such orthologous genes is not a trivial task in the large unsequenced genomes of conifers, and the genes identified here have the potential to be useful not only in this study but also in other conifer projects as orthologous genes are central both in the study of gene and species evolution. Despite the low level of nucleotide variation among conifer species, this data set enabled us to construct a very well-supported phylogeny of 10 pine species and also to identify a slowdown of evolutionary rate along in *Picea* compared with *Pinus*. Selection analysis demonstrates not only a clear impact of purifying selection on most genes but also signs of positive selection and at least one interesting candidate gene for further study (*phy*).

Supplementary Material

Supplementary figures S1–S3 as well as supplementary tables S1–S6 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The Academy of Finland, and EU projects TREES-NIPS and EVOLTREE, are acknowledged for funding postdoctoral stays of A.P. and W.W. in Oulu. We thank Soile Finne for laboratory assistance. T.P. would like to thank Finnish Graduate School in Population Genetics and Mobility Centre EVOLTREE for funding.

Literature Cited

Barrier M, Bustamante CD, Yu J, Purugganan MD. 2003. Selection on rapidly evolving proteins in the *Arabidopsis* genome. Genetics. 163:723–733.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate – a practical and powerful approach to multiple testing. J Roy Stat Soc B Met. 57:289–300.

Biswas S, Akey JM. 2006. Genomic insights into positive selection. Trends Genet. 22:437–446.

Borrelli L, De Stasio R, Filosa S, Parisi E, Riggio M, Scudiero R, Trinchella F. 2006. Evolutionary fate of duplicate genes encoding aspartic proteinases. Nothepsin case study. Gene. 368:101–109.

Böhlenius H, Huang T, Charbonnel-Campaa L, Brunner AM, Jansson S, Strauss SH, Nilsson O. 2006. CO/FT regulatory module controls timing of flowering and seasonal growth cessation in trees. Science. 312:1040–1043.

Caicedo AL, Stinchcombe JR, Olsen KM, Schmitt J, Purugganan MD. 2004. Epistatic interaction between *Arabidopsis fri* and *flc* flowering time genes generates a latitudinal cline in a life history trait. Proc Natl Acad Sci USA. 101: 15670–15675.

Carginale V, Trinchella F, Capasso C, Scudiero R, Riggio M, Parisi E. 2004. Adaptive evolution and functional divergence of pepsin gene family. Gene. 333:81–90.

Close TJ. 1997. Dehydrins: a commonality in the response of plants to dehydration and low temperature. Physiol Plantarum. 100:291–296.

Creevey CJ, McInerney JO. 2002. An algorithm for detecting directional and non-directional positive selection, neutrality and negative selection in protein coding DNA sequences. Gene. 300:43–51.

Creevey CJ, McInerney JO. 2003. Crann: detecting adaptive evolution in protein-coding DNA sequences. Bioinformatics. 19:1726.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet. 6:361–375.

Evolution of genes and genomes on the *Drosophila* phylogeny. Nature. 450:203–218.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci USA. 102:14338–14343.

Eckert AJ, Hall BD. 2006. Phylogeny, historical biogeography, and patterns of diversification for *Pinus* (Pinaceae): phylogenetic tests of fossil-based hypotheses. Mol Phylogenet Evol. 40:166–182.

Eveno E, Collada C, Guevara MA, et al. (11 co-authors). 2008. Contrasting patterns of selection at *Pinus pinaster* Ait. Drought stress candidate genes as revealed by genetic differentiation analyses. Mol Biol Evol. 25:417–437.

Felsenstein J. 2004. PHYLIP (phylogeny inference package) version 3.6 *Distributed by the author*. Seattle, USA: Department of Genome Sciences, University of Washington.

Filiault DL, Wessinger CA, Dinneny JR, Lutes J, Borevitz JO, Weigel D, Chory J, Maloof JN. 2008. Amino acid polymorphisms in *Arabidopsis* phytochrome B cause differential responses to light. Proc Natl Acad Sci USA. 105:3157–3162.

Garrigan D, Hedrick PW. 2003. Perspective: detecting adaptive molecular polymorphism: lessons from the MHC. Evolution. 57:1707–1722.

Gernandt DS, López GG, García SO, Liston A. 2005. Phylogeny and classification of *Pinus*. Taxon. 54:29–42.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 11:725–736.

Gonzalez-Martinez SC, Ersoz E, Brown GR, Wheeler NC, Neale DB. 2006. DNA sequence variation and selection of tag single-nucleotide polymorphisms at candidate genes for drought-stress response in *Pinus taeda* L. Genetics. 172: 1915–1926.

Grotkopp E, Rejmanek M, Sanderson M, Rost T. 2004. Evolution of genome size in pines (*Pinus*) and its life-history correlates: supertree analyses. Evolution. 58: 1705–1729.

Guillet-Claude C, Isabel N, Pelgas B, Bousquet J. 2004. The evolutionary implications of *knox*-I gene duplications in conifers: correlated evidence from phylogeny, gene mapping, and analysis of functional divergence. Mol Biol Evol. 21: 2232–2245.

Gyllenstrand N, Clapham D, Källman T, Lagercrantz U. 2007. A Norway spruce FLOWERING LOCUS T homolog is implicated in control of growth rhythm in conifers. Plant Physiol. 144:248–257.

Hagstrom GI, Hang DH, Ofria C, Torng E. 2004. Using avida to test the effects of natural selection on phylogenetic reconstruction methods. Artif Life. 10:156–166.

Hall T. 1999. Bioedit: a user-friendly biological sequence alignment editor and analysis program for windows 95/98/nt. Nucleic Acids Symp Ser. 41:95–98.

Hang D, Ofria C, Schmidt TM, Torng E. 2003. The effect of natural selection on phylogeny reconstruction algorithms. In: Genetic and evolutionary computation — GECCO 2003. Berlin/Heidelberg: Springer. p. 13–24.

Heuertz M, De Paoli E, Källman T, Larsson H, Jurman I, Morgante M, Lascoux M, Gyllenstrand N. 2006. Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of Norway spruce [*Picea abies* (L.) Karst]. Genetics. 174:2095–2105.

Huelsenbeck JP, Dyer KA. 2004. Bayesian estimation of positively selected sites. J Mol Evol. 58:661–672.

Huelsenbeck JP, Ronquist F. 2001. Mrbayes: bayesian inference of phylogenetic trees. Bioinformatics. 17:754–755.

Hurme P, Repo T, Savolainen O, Pääkkönen T. 1997. Climatic adaptation of bud set and frost hardiness in Scots pine (*Pinus sylvestris*). Can J For Res. 27:716–723.

Ingvarsson PK, Garcia MV, Luquez V, Hall D, Jansson S. 2008. Nucleotide polymorphism and phenotypic associations within and around the phytochrome B2 locus in European aspen (*Populus tremula*, Salicaceae). Genetics. 178:2217–2226.

Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: the beginning of incongruence? Trends Genet. 22:225–231.

Kalberer SR, Wisniewski M, Arora R. 2006. Deacclimation and reacclimation of cold-hardy plants: current understanding and emerging concepts. Plant Sci. 171:3–16.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press.

Kimura M, Ohta T. 1969. The average number of generations until fixation of a mutant gene in a finite population. Genetics. 61:763–771.

Kinlaw CS, Neale DB. 1997. Complex gene families in pine genomes. Trends Plant Sci. 2:356–359.

Kivimäki M, Kärkkäinen K, Gaudeul M, Loe G, Ågren J. 2007. Gene, phenotype and function: *glabrous*1 and resistance to herbivory in natural populations of *Arabidopsis lyrata*. Mol Ecol. 16:453–462.

Kumar S, Tamura K, Nei M. 2004. Mega3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform. 5:150–163.

Kusumi J, Tsumura Y, Yoshimaru H, Tachida H. 2002. Molecular evolution of nuclear genes in Cupressacea, a group of conifer trees. Mol Biol Evol. 19:736–747.

Le Corre V, Roux F, Reboud X. 2002. DNA polymorphism at the *frigida* gene in *Arabidopsis thaliana*: extensive non-synonymous variation is consistent with local selection for flowering time. Mol Biol Evol. 19:1261–1271.

Liston A, Gernandt DS, Vining TF, Campbell CS, Piñero D. 2003. Molecular phylogeny of *Pinaceae* and *Pinus*. Acta Hort. 615:107–114.

Liston A, Robinson WA, Pinero D, Alvarez-Buylla ER. 1999. Phylogenetics of *Pinus* (Pinaceae) based on nuclear ribosomal DNA internal transcribed spacer region sequences. Mol Phylogenet Evol. 11:95–109.

Liu Q, Zhu H. 2008. Molecular evolution of the *MLO* gene family in *Oryza sativa* and their functional divergence. Gene. 409:1–10.

López GG, Kamiya K, Harada K. 2002. Phylogenetic relationships of diploxylon pines (subgenus *Pinus*) based on plastid sequence data. Int J Plant Sci. 163:737–747.

Magallon SA, Sanderson MJ. 2005. Angiosperm divergence times: the effect of genes, codon positions, and time constraints. Evolution. 59:1653–1670.

Mes THM, Stal LJ. 2005. Variable selection pressures across lineages in *Trichodesmium* and related cyanobacteria based on the heterocyst differentiation protein gene *het*R. Gene. 346: 163–171.

Mikola J. 1982. Bud-set phenology as an indicator of climatic adaptation of Scots pine in Finland. Silva Fenn. 16: 221–228.

Millar CI. 1998. Early evolution of pines. In: Richardson DM, editor. Ecology and biogeography of *Pinus*. Cambridge: Cambridge University Press. p. 69–91.

Miller CN Jr. 1973. Silicified cones and vegetative remains of Pinus from Eocene of British Columbia. Contributions from the Museum of Paleontology, The University of Michigan. 24:101–118.

Morgenstern EK. 1996. Geographic variation in forest trees, genetic basis and application of knowledge in silviculture. Vancouver: UBC Press.

Nicholas KB Jr, Nicholas HB, Deerfield DWI. 1997. Genedoc: analysis and visualization of genetic variation. EMBNEW News. 4:14.

Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics. 148:929–936.

Nylander JAA. 2004. MrModeltest 2.2. Program distributed by the author. Evolutionary Biology Center, Uppsala University.

Ohta T. 1972. Population size and rate of evolution. J Mol Evol. 1:305–413.

Palmé AE, Wright M, Savolainen O. 2008. Patterns of divergence among conifer ESTs and polymorphism in *Pinus sylvestris* identify putative selective sweeps. Mol Biol Evol. 25:2567–2577.

Parkinson CL, Adams KL, Palmer JD. 1999. Multigene analyses identify the three earliest lineages of extant flowering plants. Curr Biol. 9:1485–1491.

Popescu CE, Borza T, Bielawski JP, Lee RW. 2006. Evolutionary rates and expression level in Chlamydomonas. Genetics. 172:1567–1576.

Price RA, Liston A, Strauss SH. 1998. Phylogeny and systematics of *Pinus*. In: Richardson DM, editor. In: Ecology and biogeography of *Pinus*. Cambridge: Cambridge University Press. p. 44–68.

Pyhäjärvi T, García-Gil MR, Knürr T, Mikkonen M, Wachowiak W, Savolainen O. 2007. Demographic history has influenced nucleotide diversity in European *Pinus sylvestris* populations. Genetics. 177:1713–1724.

Rokas A, Carroll SB. 2005. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. Mol Biol Evol. 22:1337–1344.

Rokas A, Williams BL, King N, Carroll SB. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature. 425:798–804.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: bayesian phylogenetic inference under mixed models. Bioinformatics. 19:1572–1574.

Roth C, Liberles DA. 2006. A systematic search for positive selection in higher plants (Embryophytes). BMC Plant Biol. 6:12.

Savolainen O, Pyhäjärvi T. 2007. Genomic diversity in forest trees. Curr Opin Plant Biol. 10:162–167.

Savolainen O, Pyhäjärvi T, Knürr T. 2007. Gene flow and local adaptation in trees. Annu Rev Ecol Evol Syst. 38:595–619.

Seppänen MM, Cardi T, Borg Hyökki M, Pehu E. 2000. Characterization and expression of cold-induced glutathione s-transferase in freezing tolerant *Solanum commersonii*, sensitive *S. tuberosum* and their interspecific somatic hybrids. Plant Sci. 153:125–133.

Soltis PS, Soltis DE, Chase MW. 1999. Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. Nature. 402:402–404.

Stinchcombe JR, Weinig C, Ungerer M, Olsen KM, Mays C, Halldorsdottir SS, Purugganan MD, Schmitt J. 2004. A

latitudinal cline in flowering time in *Arabidopsis thaliana* modulated by the flowering time gene *frigida*. Proc Natl Acad Sci USA. 101:4712–4717.
Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. Proc Natl Acad Sci USA. 100:9440–9445.
Syring J, Willyard A, Cronn R, Liston A. 2005. Evolutionary relationships among *Pinus* (Pinaceae) subsections inferred from multiple low-copy nuclear loci. Am J Bot. 92: 2086–2100.
Tajima F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. Genetics. 135:599–607.
Thompson J, Gibson T, Plewniak F, Jeanmougin F, Higgins D. 1997. The Clustal X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 24:4876–4882.
Tiffin P, Hahn MW. 2002. Coding sequence divergence between two closely related plant species: *arabidopsis thaliana* and *Brassica rapa* ssp. *Pekinensis*. J Mol Evol. 54:746–753.
Wachowiak W, Balk P, Savolainen O. 2009. Search for nucleotide patterns of local adaptation in dehydrins and other cold related candidate genes in Scots pine (*Pinus sylvestris* L.). Tree Genet Genomes. 5:117–132.
Wang X-Q, Tank DC, Sang T. 2000. Phylogeny and divergence times in Pinaceae: evidence from three genomes. Mol Biol Evol. 17:773–781.
Wang X-R, Tsumura Y, Yoshimaru H, Nagasaka K, Szmidt AE. 1999. Phylogenetic relationships of Eurasian pines (*Pinus*, Pinaceae) based on chloroplast *rbc*l, *mat*k, *rpl*20-*rps*18 spacer, and *trn*v intron sequences. Am J Bot. 86: 1742–1753.
Werner JD, Borevitz JO, Warthmann N, Trainer GT, Ecker JR, Chory J, Weigel D. 2005. Quantitative trait locus mapping and DNA array hybridization identify an *flm* deletion as a cause for natural flowering-time variation. Proc Natl Acad Sci USA. 102:2460–2465.
Wiens JJ. 2003. Incomplete taxa, incomplete characters, and phylogenetic accuracy: is there a missing data problem? J Vertebr Paleontol. 23:297–310.
Wiens JJ. 2006. Missing data and the design of phylogenetic analyses. J Biom Inform. 39:34–42.
Willyard A, Syring J, Gernandt DS, Liston A, Cronn R. 2007. Fossil calibration of molecular divergence infers a moderate mutation rate and recent radiations for *Pinus*. Mol Biol Evol. 24:90–101.
Wright SI, Gaut BS. 2005. Molecular population genetics and the search for adaptive evolution in plants. Mol Biol Evol. 22:506–519.
Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol. 15:568–573.
Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13:555–556.
Yang Z, Wong WSW, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol. 22:1107–1118.
Zhang L, Li W-H. 2004. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. Mol Biol Evol. 21:236–239.
Zhang Y, Sederoff RR, Allona I. 2000. Differential expression of genes encoding cell wall proteins in vascular tissues from vertical and bent loblolly pine trees. Tree Physiol. 20:457–466.

Koichiro Tamura, Associate Editor

Accepted January 15, 2009