COMPARATIVE GENOMIC ANALYSIS OF ADAPTIVE AND ECONOMIC TRAITS RELATED GENES IN SOUTHERN PINES

A Dissertation

by

TOMASZ EDMUND KORALEWSKI

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2010

Major Subject: Genetics

COMPARATIVE GENOMIC ANALYSIS OF ADAPTIVE AND ECONOMIC TRAITS RELATED GENES IN SOUTHERN PINES

A Dissertation

by

TOMASZ EDMUND KORALEWSKI

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Konstantin V. Krutovsky
Committee Members,	Clare A. Gill
	Mariana Mateos
	Alan E. Pepper
Chair of Intercollegiate Faculty,	Craig J. Coates

August 2010

Major Subject: Genetics

ABSTRACT

Comparative Genomic Analysis of Adaptive and Economic Traits Related Genes in Southern Pines. (August 2010)

Tomasz Edmund Koralewski, B.S., Kazimierz Wielki University in Bydgoszcz, Poland;
M.S., University of Technology and Agriculture, Bydgoszcz, Poland
Chair of Advisory Committee: Dr. Konstantin V. Krutovsky

Four major Southern pines, *Pinus echinata* Mill., *P. elliottii* Engelm., *P. palustris* Mill. and *P. taeda* L. are evolutionarily young and closely related. They have not been intensely researched except *P. taeda*. In this study we addressed the questions of exon-intron structure, nucleotide variation and neutrality in adaptive and economic traits related genes, and phylogenetic relationships between these pines. Using publically available data in the NCBI databases, we first developed a series of statistical regression models. We defined functional relationships between the parameters that can be easily estimated from a small data sample (e.g. mean exon length and exon/gene ratio), and parameters that are difficult to assess (e.g. number of genes and exons). Second, we examined the effects of selection upon the set of studied genes in the four pines. We collected data from individuals representing all four Southern pines and merged them with previously published data, and applied four neutrality tests: Tajima's *D*, McDonald-Kreitman (MK), Hudson-Kreitman-Aguade (HKA), and synonymous-nonsynonymous nucleotide substitutions ratio. Finally, we analyzed phylogenetic relationships between the four Southern pines, and with respect to other selected pine species (*P. radiata*, *P. pinaster* and *P. sylvestris*), for which the nucleotide sequence data orthologous to the sequences newly generated in this study were available in the NCBI GenBank. We applied Maximum Parsimony, Maximum Likelihood and Bayesian Inference approaches.

Based on the statistical models we expect about 13-14 thousand genes in an organism with the mean exon length of 334.8 bp (like *P. taeda*). This number could be higher in plants (20-21 thousand). Furthermore, we identified signatures of selection in some of the studied genes, and demonstrated that different parts of a gene could be under different forms of selection. Therefore, the results of the neutrality tests performed at the entire gene level could be misleading. Finally, using twelve nuclear loci we confirmed very tight phylogenetic relationships within the subsection *Australes*, but the conclusions were not robust. Using two exclusive sets of three genes led to robust but conflicting results. Therefore, we demonstrated that conclusions about “species” trees based on “gene” trees may be misleading, especially for closely related species.

DEDICATION

To:

Mama, Tata, Ela and Hsiao-Hsuan

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Konstantin Krutovsky, for his continuing efforts in shaping me as a scientist, for caring about my and my wife's wellbeing, and for supporting me during the last four years of my life in multiple ways.

I would like to thank all my committee members, Dr. Clare Gill, Dr. Ruzong Fan, Dr. Mariana Mateos, and Dr. Alan Pepper, for their guidance and support throughout the course of my study.

I am also very grateful to Dr. Tom Byram for his support of my research.

I thank Dr. Judy Brooks, my everyday lab companion, who was not only my teacher and friend but also bottomless resource of knowledge about Texas and Texans.

Thanks also go to my friends, whom I met in College Station, and who made my life here easier, bringing optimism, support, advice and joy. Especially, to Agnieszka and Daniel Chmura for helping me find my way around once I arrived at College Station, Donita Bryan and Andrew Cartmill for standing beside me during my wedding, and Lihui Lee and Joseph Pollacco for bringing much color to our life over the last year.

I thank my mom and dad for being there for me always, and my sister for always being in the same team with me.

And lastly, I thank my wife, Hsiao-Hsuan Wang, for her encouragement, support and love. She was the reason to return home after work.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Four major Southern pines, *Pinus echinata* Mill. (shortleaf pine), *P. elliottii* Engelm. (slash pine), *P. palustris* Mill. (longleaf pine) and *P. taeda* L. (loblolly pine) grow in a diverse area that stretches from subtropical to warm temperate climate across thirteen Atlantic coast and Southern states. They provide enormous, multiple benefits to the ecosystem and human society.

Southern pine forests play a very important ecological role in carbon sequestration and climate change mitigation (Johnsen et al. 2001; van Minnen et al. 2008; Fahey et al. 2010; Malmsheimer et al. 2008). They provide habitat for many microbial, fungal, plant and animal species directly (e.g. for arthropods, birds and deer; Dickson and Segelquist 1979; Melchiors et al. 1985; Collins et al. 2002) and indirectly for species living in understory and litter layers (Carey and Johnson 1995; Holmes and Robinson 1988). Their seeds are an important food source for birds and rodents (Schultz 1999), and seedlings are browsed by larger animals (Michael 1985). They are keystone species and vital components of various management policies, e.g. soil erosion control, soil stabilization and watershed protection, and may play an important role in fire ecology through forest stand regeneration (Schultz 1999). They are also a recreational and ornamental component of the landscape.

This dissertation follows the style of Tree Genetics & Genomes.

Southern pines are among the most economically important crops cultivated in the USA (USDA Forest Service), providing lumber and pulp. They are also a great potential source for biofuel and alternative energy production (e.g. Frederick et al. 2008).

Despite the ecological and economic value, and with the exception of loblolly pine, the Southern pines are not as intensely researched as other crops. The loblolly pine is the most studied conifer species and has become a model species for conifers (Krutovsky et al. 2004). Although the genomic studies in pines are hindered by their large genome size (e.g. loblolly pine genome is ~24Gb, i.e. a few times larger than human genome; Grotkopp et al. 2004), significant progress has been made in comparative mapping and nucleotide polymorphism studies in recent years (e.g. Krutovsky et al. 2004; Krutovsky and Neale 2005; Neale and Ingvarsson 2008; Brown et al. 2004). The complete genome sequence for *P. taeda* is underway (USDA AFRI 2010). Because no other closely related coniferous species has been entirely sequenced, complete assembly and annotation is likely to be a very time-consuming process and pose multiple challenges.

In two previous large-scale studies on *P. taeda*, 34 drought-stress response, drought resistance and wood-quality related genes were analyzed for nucleotide diversity, linkage disequilibrium (LD) and signatures of selection (Gonzalez-Martinez et al. 2006a; Brown et al. 2004). Both samples included up to 32 individuals originating from various locations of the natural range of loblolly pine. In both studies the authors found low level of LD, moderate nucleotide diversity, and no population genetic structure. Despite some potential signatures of selection identified, the authors failed to reject neutrality in both studies. In addition, Gonzalez-Martinez et al. (2006a) did not find any evidence of selection acting upon an amino acid. However, only in the latter study were interspecific comparisons done, where *P. pinaster* was used as an outgroup. Interspecific tests, such as MK (McDonald and Kreitman 1991) and HKA (Hudson et al. 1987), are better suited for distinguishing the effects of demographic events from the effects of selection. We largely expanded this analysis via including multiple pine species. Applying these tests to the four Southern pines is particularly important because these species went through a severe bottleneck during the last glacial period that ended about 15,000 years ago. During this time their range is thought to have been constricted to two refugia, central Florida and the Caribbean, and Southern Texas and Mexico (Jackson et al. 2000; Wells et al. 1991; Schmidtling et al. 1999; Al-Rabab'ah and Williams 2002).

The recent common ancestry, common history and greatly overlapping habitat make relationships between the four Southern pines difficult to dissect. This problem has been addressed in larger phylogenetic studies on the genus. They have consequently been placed within subsection *Australes*; however, despite various methods applied and use of morphological and molecular data from both chloroplast and nuclear genomes, consensus about the ancestry within this group has not been reached (e.g. Grotkopp et al. 2004; Gernandt et al. 2005; Eckert and Hall 2006). We also largely expanded this phylogenetic analysis via including multiple additional genes.

In this study we focused on gene structure, nucleotide variation in wood quality and drought hardiness related genes, and phylogenetic relationships between the four Southern pines, in order to understand adaptive and evolutionary processes in these species and in the genus (Krutovsky and Neale 2005; Gonzalez-Martinez et al. 2007;

Gonzalez-Martinez et al. 2006a; Gonzalez-Martinez et al. 2006b). The main objectives of this study were to:

1) examine exon-intron structure and alternative splicing across evolutionarily diverse well-studied model organisms, and create regression models that could be used in pines for predicting genome-wide characteristics, such as number of all genes and number of all exons, based on parameters that are much easier to estimate, e.g. exon length and exon-gene ratio;
2) find selection signatures in wood quality and drought hardiness-related genes in the four Southern pines using neutrality tests;
3) refine the phylogenetic relationships between the four Southern pine species within the subsection *Australes*, and with other pine species in the family *Pinaceae* using genomic nucleotide sequence data.

The ability of predicting the genome-wide characteristics in the species, whose genomes have not been completely sequenced, including loblolly pine, will help guide the process of annotation and assembly of the genomic sequences. As more data become available, the regression models may be fine-tuned which will increase their precision.

The application of the neutrality tests to the expanded set of species and number of individuals will not only increase the statistical power of the data, but also allow for more thorough interspecific comparisons within the group of Southern pines. The data used for these analyses will also be used to investigate the phylogenetic relationship within this closely related group of Southern pines.

2. EVOLUTION OF INTRON-EXON STRUCTURE AND ALTERNATIVE SPLICING: WHAT WE LEARNED FROM COMPLETELY SEQUENCED GENOMES AND CAN PREDICT FOR NON-MODEL SPECIES

2.1. Overview

Despite significant advances in high-throughput DNA sequencing, many important species remain understudied at the genome level. Using NCBI GenBank database we performed a genome-wide analysis of such characteristics as alternative splicing, number of genes, gene products and exons in 36 completely sequenced model species. We created statistical regression models to fit these data and applied them to loblolly pine, an important species whose genome has not been completely sequenced yet. Using these models, the genome-wide characteristics, such as exon length and exon-gene ratio, can be predicted based on parameters estimated from available genomic data.

2.2. Introduction

Recent advances in high-throughput DNA sequencing led to significant progress in complete genome sequencing and opened unprecedented opportunities for comparative genome studies (Chi 2008; Mardis 2008; von Bubnoff 2008; Wheeler et al. 2008). The complete genome sequences are publicly available from constantly growing databases, such as the National Center for Biotechnology Information (NCBI) GenBank, and can be readily analyzed and compared for a number of evolutionarily distant species. The early comparisons revealed that the number of genes and metabolomic complexity progressively increase as species become more evolutionarily advanced (Adami et al. 2000; Lynch and Conery 2003; Graveley 2001; Valentine 2000), but their anatomical, morphological, physiological and behavioral complexity does not linearly correlate with the total number of genes discovered. For instance, whereas the number of protein coding genes in the human genome is only 14% greater than in the roundworm *Caenorhabditis elegans*, the evolutionary differences between these two species are immense. This suggests that regulatory and post-transcriptional processes might play an increasingly more important role throughout evolution. There are numerous mechanisms, processes and structures that affect gene regulation, such as methylation, chromatin structure, regulatory elements, transcription factors, polyadenylation, posttranslational modifications and compartmentalization of proteins, and others (for review see Orphanides and Reinberg 2002). However, alternative splicing (AS) is the only post-translational process that can increase proteomic complexity and number of various proteins without increasing the number of genes. Due to post-transcriptional modification and rearrangement of exons in the process of AS, additional variants are created among the mature mRNA transcripts (e.g. Black 2003; McKeown 1992). AS can promote adaptive and evolutionary potential of species without increasing the number of genes and maintenance cost that could be associated with it. For instance, the total hypothetical number of various proteins encoded by the *DSCAM* gene can reach 38,016 in *Drosophila melanogaster* (Black 2000). Therefore, one may expect that more evolutionarily advanced organisms have more elaborated and complex AS. We addressed this hypothesis in more detail in our study. Our objectives were to examine exon-intron structure in genomes of completely sequenced and fully annotated species, to infer AS data and to use this information for defining relationships between genes and proteomic complexity. We expect that these relationships can be used to predict the anticipated exon-intron structure and proteomic complexity in non-model species with large genomes, such as pines, that may remain unsequenced for a while. We applied our findings to loblolly pine (*Pinus taeda* L.), one of the most-studied coniferous species, which has a very large genome of 24.56 pg (~24 Gb; Grotkopp et al. 2004); complete genome sequencing for loblolly pine is underway (USDA AFRI 2010), but is still problematic and unavailable. The obtained knowledge is also essential for understanding the genetic control of the metabolomic complexity and functionality in the studied species and the evolutionary significance of AS in general.

2.3. Materials and methods

Selection of completely sequenced species for analysis. We selected the 36 most-annotated and featured species (Table 1) from eukaryotic genomic assemblies available in the NCBI GenBank (Benson et al. 2009). For 10 species in our set (*Arabidopsis thaliana, Caenorhabditis elegans, Drosophila melanogaster, Encephalitozoon cuniculi, Eremothecium gossypii*, *Homo sapiens, Mus musculus, Oryza sativa, Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*), data have been also entirely or almost entirely supported by records other than those provided by the NCBI GenBank. AS in rice was not documented in the present NCBI GenBank genome annotation, although it

Table 1 Exon-intron gene structure in completely sequenced genomes of 36 species

Taxonomic group	Species	Genes		Protein coding / Total gene ratio	CDSs	Exons	Exon length			CDS length			Exon / Gene ratio
		All	Protein coding				Mean	SD	Median	Mean	SD	Median	
Excavata	*Leishmania braziliensis*	7,898	7,897	0.9999	7,898	7,998	1,844.8	1,732.2	1,383.0	1,868.2	1,771.0	1,395.0	1.013
	Leishmania infantum	7,993	7,993	1.0000	7,993	8,069	1,839.9	1,660.6	1,401.0	1,857.6	1,720.0	1,401.0	1.010
	Trypanosoma brucei	9,336	8,772	0.9396	8,772	8,774	1,506.2	1,472.2	1,158.0	1,506.5	1,472.3	1,158.0	1.000
Chromalveolata	*Cryptosporidium parvum*	3,885	3,396	0.8741	3,396	3,440	1,821.1	1,942.4	1,321.5	1,844.7	1,945.5	1,341.0	1.013
	Guillardia theta	742	632	0.8518	632	648	851.1	747.6	625.5	872.6	745.2	633.0	1.025
	Hemiselmis andersenii	524	471	0.8989	471	471	1,018.3	808.6	774.0	1,018.3	808.6	774.0	1.000
	Plasmodium falciparum	5,300	5,263	0.9930	5,267	12,651	949.6	1,767.0	201.0	2,280.8	2,588.6	1,386.0	2.404
	Theileria parva	4,089	4,035	0.9868	4,035	14,447	393.2	680.8	159.0	1,408.0	1,230.2	1,080.0	3.580
Amoebozoa	*Dictyostelium discoideum*	13,322	13,322	1.0000	13,331	30,441	686.9	1,016.4	314.0	1,569.2	1,510.9	1,149.0	2.285
Fungi	*Aspergillus fumigatus*	9,859	9,630	0.9768	9,630	28,259	504.0	679.7	274.0	1,479.0	1,114.2	1,248.0	2.934
	Aspergillus niger	14,420	14,086	0.9768	14,086	50,371	370.4	565.1	176.0	1,324.4	1,103.9	1,089.0	3.576
	Candida glabrata	5,499	5,271	0.9585	5,272	5,356	1,485.5	1,104.1	1,239.0	1,509.2	1,097.6	1,260.0	1.016
	Cryptococcus neoformans	6,407	6,273	0.9791	6,475	39,350	257.2	327.8	150.0	1,608.8	1,105.8	1,371.0	6.273
	Debaryomyces hansenii	7,081	6,866	0.9696	6,872	7,227	1,274.3	1,036.1	1,059.0	1,340.5	1,026.7	1,113.0	1.053
	*Encephalitozoon cuniculi**	2,029	1,996	0.9837	1,996	2,011	1,072.3	812.2	846.0	1,080.4	810.2	852.0	1.008
	*Eremothecium gossypii**	4,971	4,714	0.9483	4,714	4,940	1,406.1	1,109.2	1,164.0	1,474.7	1,093.9	1,228.5	1.048
	Gibberella zeae	11,619	11,619	1.0000	11,619	37,454	477.2	699.6	242.0	1,538.2	1,233.6	1,272.0	3.224
	Kluyveromyces lactis	5,504	5,331	0.9686	5,331	5,461	1,377.1	1,062.9	1,146.0	1,410.7	1,054.8	1,173.0	1.024

Table 1 Continued

Taxonomic group	Species	Genes		Protein coding / Total gene ratio	CDSs	Exons	Exon length			CDS length			Exon / Gene ratio
		All	Protein coding				Mean	SD	Median	Mean	SD	Median	
Fungi	*Neurospora crassa*	10,093	9,699	0.9610	9,709	26,598	533.5	747.1	241.0	1,462.6	1,199.4	1,215.0	2.742
	Pichia stipitis	5,816	5,816	1.0000	5,816	8,383	1,025.5	954.8	816.0	1,478.1	1,043.1	1,251.0	1.441
	*Saccharomyces cerevisiae**	6,136	5,861	0.9552	5,861	6,185	1,412.6	1,132.8	1,179.0	1,490.7	1,149.2	1,224.0	1.055
	*Schizosaccharomyces pombe**	5,374	5,083	0.9459	5,084	9,844	722.9	964.7	338.0	1,400.4	1,101.6	1,140.0	1.937
	Ustilago maydis	6,604	6,495	0.9835	6,495	11,373	1,052.2	1,242.2	569.0	1,842.4	1,363.9	1,509.0	1.751
	Yarrowia lipolytica	7,180	6,660	0.9276	6,661	7,402	1,295.4	1,086.6	1,089.0	1,439.7	1,084.8	1,191.0	1.111
Viridiplantae	*Arabidopsis thaliana**	28,245	26,977	0.9551	30,705	138,876	236.8	316.5	134.0	1,208.0	883.6	1,041.0	5.148
	Oryza sativa	29,102	26,777	0.9201	26,777	128,267	250.2	353.9	132.0	1,198.3	868.4	1,023.0	4.790
	Ostreococcus 'lucimarinus'	7,603	7,603	1.0000	7,603	9,767	944.8	1,109.6	744.0	1,213.7	1,210.1	966.0	1.285
Metazoa (Nematoda)	*Caenorhabditis briggsae*	17,363	16,429	0.9462	16,429	98,457	209.8	228.2	149.0	1,257.1	1,104.2	996.0	5.993
	*Caenorhabditis elegans**	21,172	20,174	0.9529	23,759	124,949	203.1	227.9	146.0	1,322.0	1,423.6	1,029.0	6.194
Metazoa (Arthropoda)	*Anopheles gambiae*	12,423	11,971	0.9636	12,500	48,875	358.2	470.5	205.0	1,454.4	1,533.1	1,077.0	4.083
	*Drosophila melanogaster**	14,807	13,887	0.9379	17,837	56,580	401.0	560.7	216.0	1,719.8	1,868.3	1,266.0	4.074
	Drosophila pseudoobscura	11,875	9,606	0.8089	9,707	39,256	383.9	448.9	222.0	1,553.0	1,396.3	1,206.0	4.087
Metazoa (Mammalia)	*Canis lupus familiaris*	19,384	19,380	0.9998	31,837	194,624	169.2	243.8	124.0	1,748.3	1,599.6	1,311.0	10.043
	*Homo sapiens**	25,074	23,055	0.9195	27,904	201,083	174.5	277.6	124.0	1,548.4	1,852.9	1,128.0	8.722
	*Mus musculus**	26,314	25,533	0.9703	27,159	200,714	179.4	271.6	125.0	1,420.9	1,633.0	1,002.0	7.861
	Pan troglodytes	23,962	23,881	0.9966	40,767	177,922	170.2	240.3	122.0	1,440.1	1,342.4	1,095.0	7.450

* The most annotated species (see Materials and Methods for details)

has been reported previously (Campbell et al. 2006; Severing et al. 2009). Therefore, this species was excluded from the AS analysis.

Source of data. All genomic data were downloaded from the FTP directory of the NCBI GenBank (ftp://ftp.ncbi.nih.gov/genomes/MapView/). Sequences for *P. taeda* were downloaded from the Nucleotide database from the NCBI GenBank.

Genomic data analysis. Genomic data were analyzed using Perl scripts specifically written for this study. The downloaded files were screened, and chromosome ID, position and orientation of the exons on the chromosome, feature ID, AS type, transcript accession number, and group label were traced, partitioned and analyzed. Pseudogenes, mitochondrial, plastid and insufficiently annotated genes were excluded from further analysis. Total numbers of genes, protein coding genes and their coding sequences (CDSs) were calculated for each species. The number of exons and their boundaries were determined based on the coding structure of each protein coding gene ID recorded in their corresponding CDS section. For each gene supported by more than one CDS, the alternative coding sequences were compared with each other. Cases when corresponding exons had different boundaries or no matching counterpart were qualified as AS variants. Average and median lengths were calculated for both exons and CDSs. The exon estimates were computed based on all unique exons found in the genome. All CDSs, including alternatively spliced forms, were considered for estimation of the average and median CDS lengths. The exon/gene ratio was defined as the average number of exons per protein coding gene. AS ratio was defined as the ratio between the number of alternatively spliced and all protein coding genes. AS variants were categorized using the binary approach described by Nagasaki et al. (2006). Shannon's index *H* and equitability *E* were calculated for genes with AS to reflect richness and distribution evenness of AS forms (Shannon 1948).

Parameter estimation for Pinus taeda. In total, 99 complete CDS sequences representing protein coding genes in *P. taeda* were downloaded from NCBI GenBank, and their CDS structure and length were analyzed. The data were prescreened by a Perl script and rearranged manually. The majority of these sequences represented mRNA/cDNA. Only five CDSs represented genomic sequences and could provide complete information about exon-intron structure. Average and median CDS and exon lengths were calculated based on this information. Using the mean exon length and regression models developed based on the genomic data for other species (see below), we computed the expected total number of exons, total number of genes, exon/gene ratio, and total number of protein coding genes for *P. taeda*. Software package JMP version 5 was used for the statistical analysis.

2.4. Results

Analysis of complete genomes. The main results are summarized in Table 1. General trends demonstrated an increase in the number of genes, gene products and total number of exons as species advance evolutionarily. The exon/gene ratio also increases, but the average exon length becomes shorter whereas CDS length remains relatively constant. The results of regression analysis and estimates of the parameters are summarized in Table 2.

Average exon length as a predictor. A very strong negative correlation was observed between mean exon length and total number of exons (r^2 = 0.937, r^2_{adj} = 0.936; Fig. 1A). The negative correlation was weaker but statistically significant between mean exon length and either number of protein coding genes (r^2 = 0.712, r^2_{adj} = 0.703; Fig. 1B) or the total number of genes (r^2 = 0.706, r^2_{adj} = 0.697, Fig. A1, Appendix A). Mean exon length also strongly negatively correlated with exon/gene ratio (r^2 = 0.957, r^2 adjusted = 0.956; Fig. 1C). In all these cases the correlations were not linear. No statistically significant correlation was observed between mean exon length and mean CDS length (P = 0.132; Fig. A2, Appendix A).

Similar correlations were observed for median exon length, and the r^2 values were close to those obtained for mean exon length (see Table 2 for details).

Both average and median exon lengths become shorter in more advanced organisms (Fig. 2A).

Table 2 Predicted values for exon-intron gene structure and alternative splicing (AS) parameters for an organism with mean and median exon lengths of 334.8 and 198.0 bp, respectively, such as observed in *Pinus taeda*, based on results of regression analysis

Response (y)	**Factor (x)**	R^2	R^2_{adj}	***P*-value at 95% CI**	**Figure**	**Predicted**	**95% CI at population level**		**95% CI at individual level**	
							lower	**upper**	**lower**	**upper**
Number of exons	Mean exon length	0.937	0.936	<0.0001	1A	53,374	47,887	58,860	20,093	86,655
Number of protein coding genes	Mean exon length	0.712	0.703	<0.0001	1B	13,288	11,780	14,797	4,824	21,752
Exon / Gene ratio	Mean exon length	0.957	0.956	<0.0001	1C	4.245	4.049	4.441	3.146	5.344
Number of all genes	Mean exon length	0.706	0.697	<0.0001	A1	13,871	12,270	15,471	4,891	22,850
Mean CDS length	Mean exon length	0.065	0.038	0.1321	A2	-				
Number of protein coding genes	Number of exons	0.897	0.894	<0.0001	1D	-				
Number of protein coding genes	Number of all genes	0.996	0.996	<0.0001	1E	-				
Number of all genes	Number of exons	0.891	0.888	<0.0001	A3	-				
Number of protein coding genes	Exon / Gene ratio	0.648	0.638	<0.0001	A4	-				
Number of exons	Exon / Gene ratio	0.864	0.860	<0.0001	1F	-				
AS	Mean exon length	0.615	0.576	0.0025	A5	0.018	0	0.053	0	0.117
AS	Exon / Gene ratio	0.498	0.448	0.0103	A6	-				
AS	Number of CDSs	0.725	0.698	0.0004	A7	-				
Number of exons	Mean exon length	0.999	0.998	0.0175	-	71,010*	49,220	92,801	29,385	112,636
Number of protein coding genes	Mean exon length	0.997	0.994	0.0351	-	19,785*	13,411	26,159	7,063	32,508
Number of all genes	Mean exon length	0.990	0.980	0.0632	-	20,923*	8,371	33,474	0	45,975
Number of exons	Median exon length	0.852	0.848	<0.0001	-	59,904	51,343	68,465	8,737	111,070
Exon / Gene ratio	Median exon length	0.903	0.901	<0.0001	-	3.658	3.382	3.934	2.007	5.308
Number of all genes	Median exon length	0.710	0.701	<0.0001	-	12,342	10,853	13,831	3,441	21,243
Number of protein coding genes	Median exon length	0.715	0.707	<0.0001	-	11,827	10,422	13,231	3,432	20,221
Median CDS length	Median exon length	0.035	0.006	0.2767	-	-				

* Models constructed based on 3 plant species; see text for details

A Correlation of number of exons and mean exon length

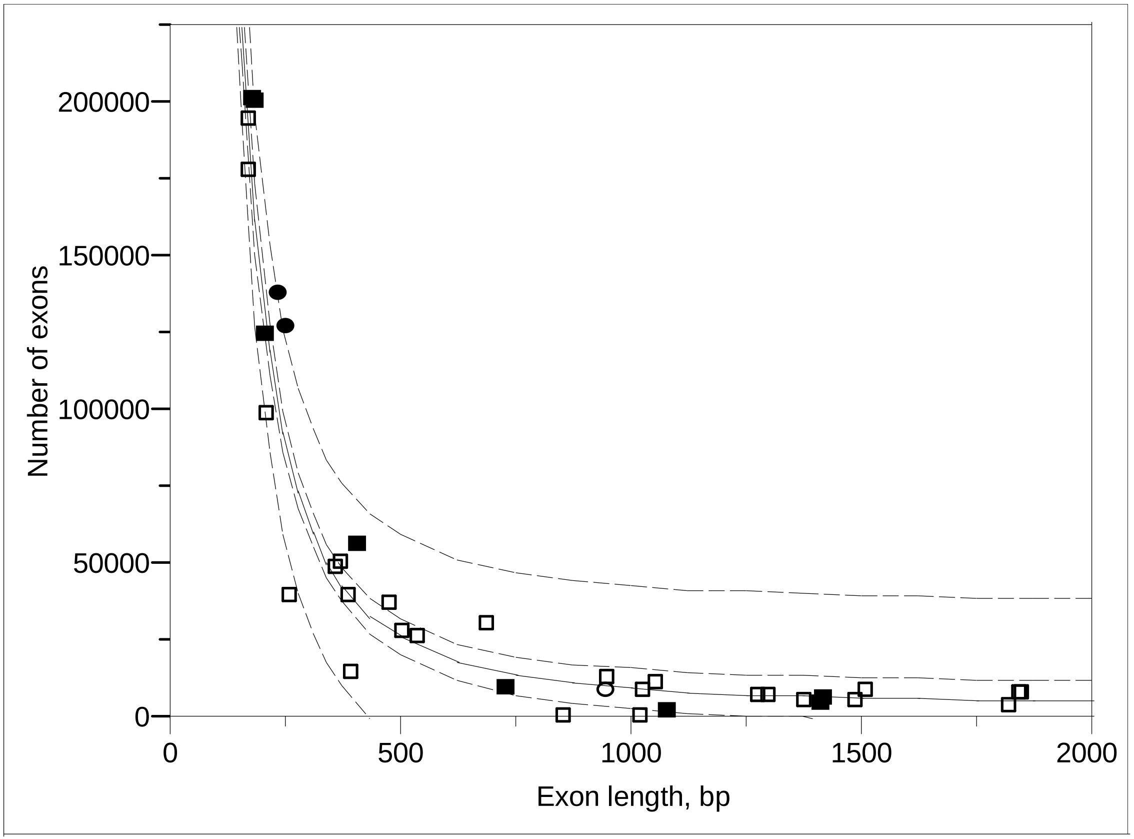


Fig. 1 Correlations of number of exons and mean exon length (A), number of protein coding genes and mean exon length (B), exon/gene ratio and mean exon length (C), number of protein coding genes and number of exons (D), number of protein coding genes and number of all genes (E), and number of exons and exon/gene ratio (F) based on 36 species studied. The most annotated species are represented by solid markers. Three plants in the dataset are marked by circle. 95% confidence intervals are presented for both population (internal dashed line) and individual (external dashed line) levels

B Correlation of number of protein coding genes and mean exon length

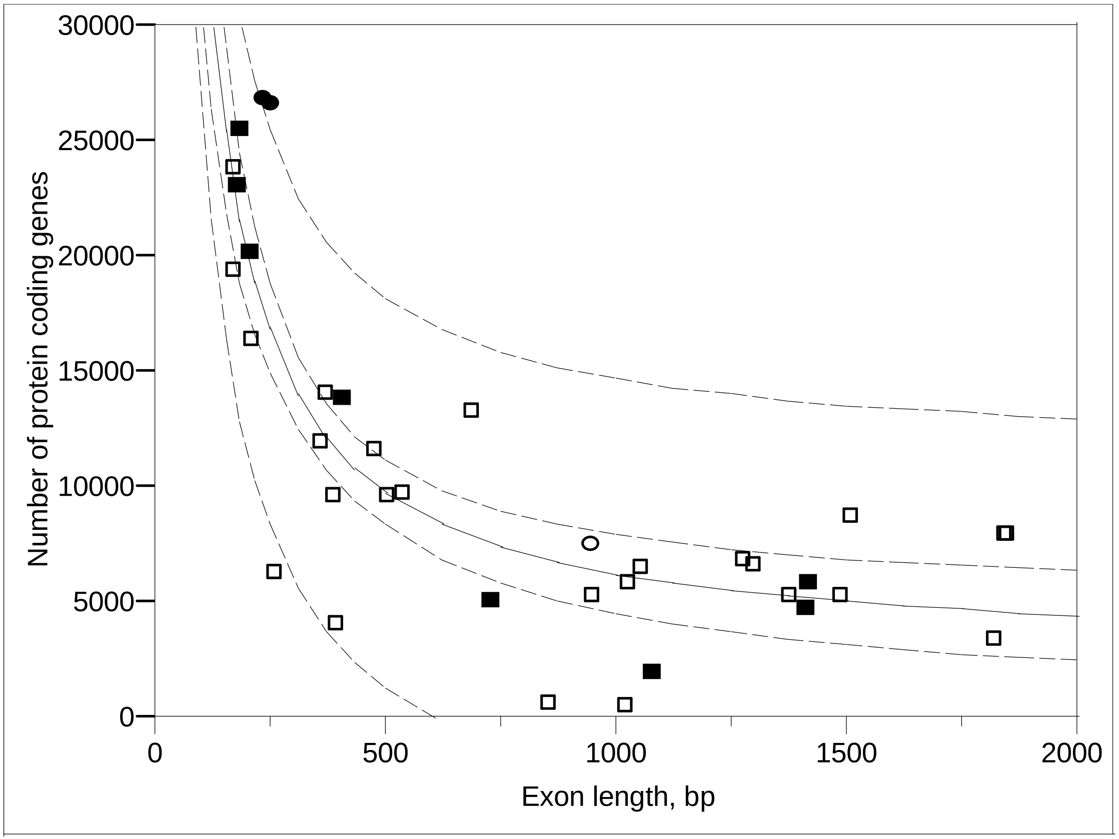


Fig. 1 Continued

C Correlation of exon/gene ratio and mean exon length

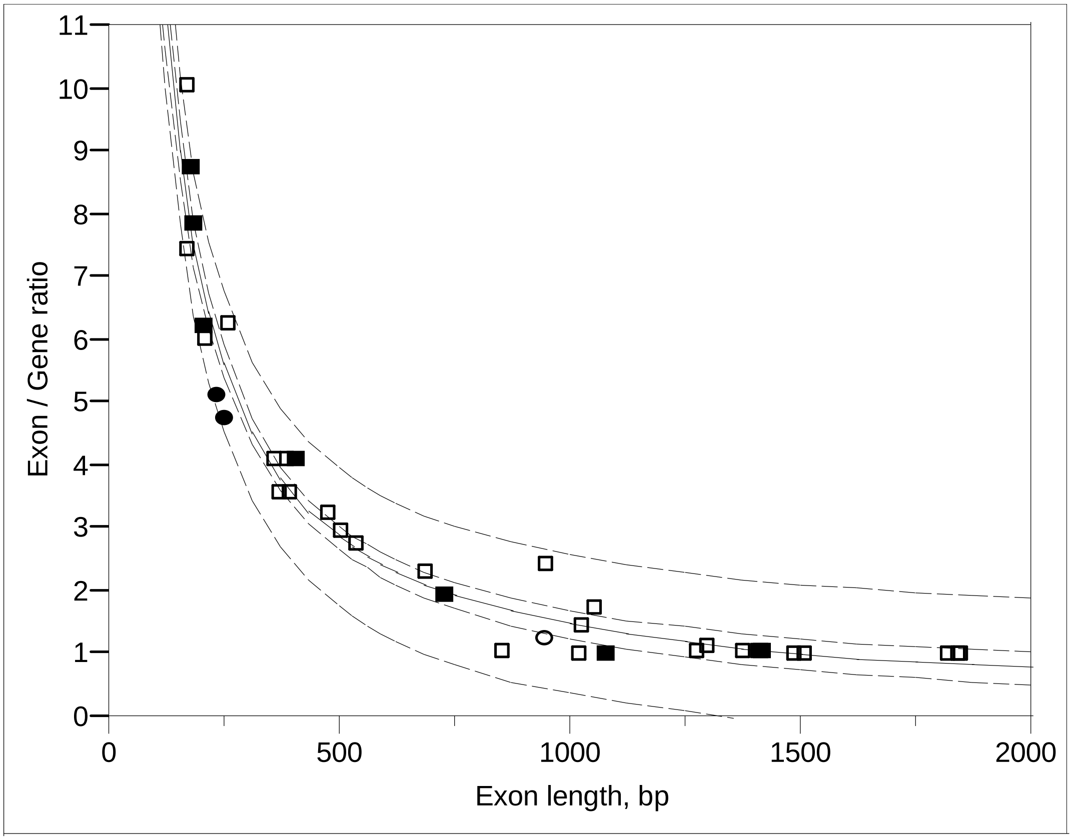


Fig. 1 Continued

D Correlation of number of protein coding genes and number of exons

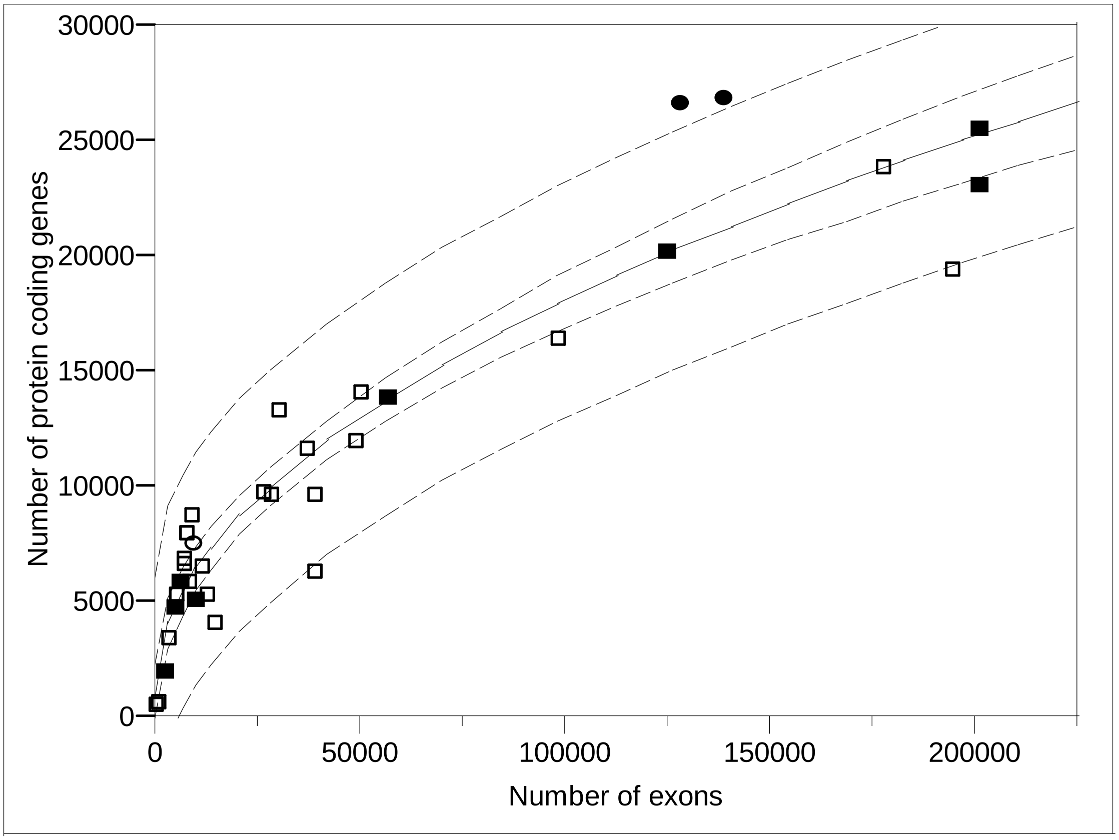


Fig. 1 Continued

E Correlation of number of protein coding genes and number of all genes

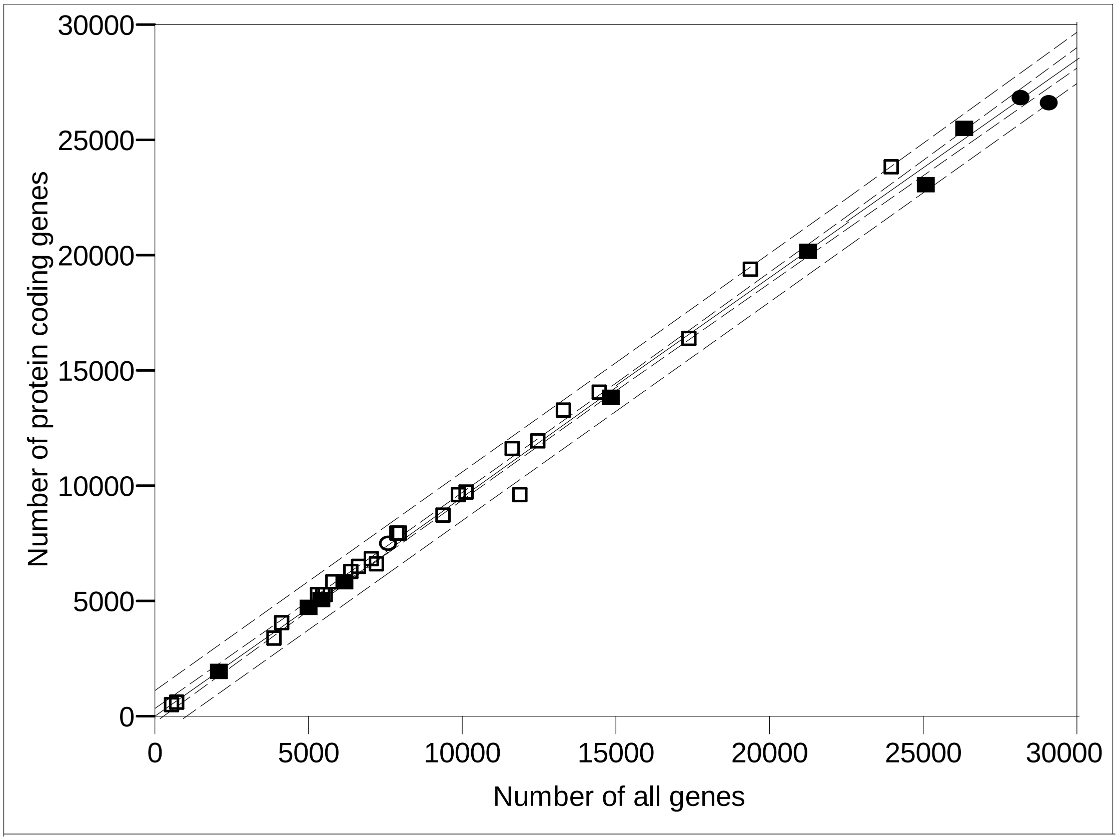


Fig. 1 Continued

F Correlation of number of exons and exon/gene ratio

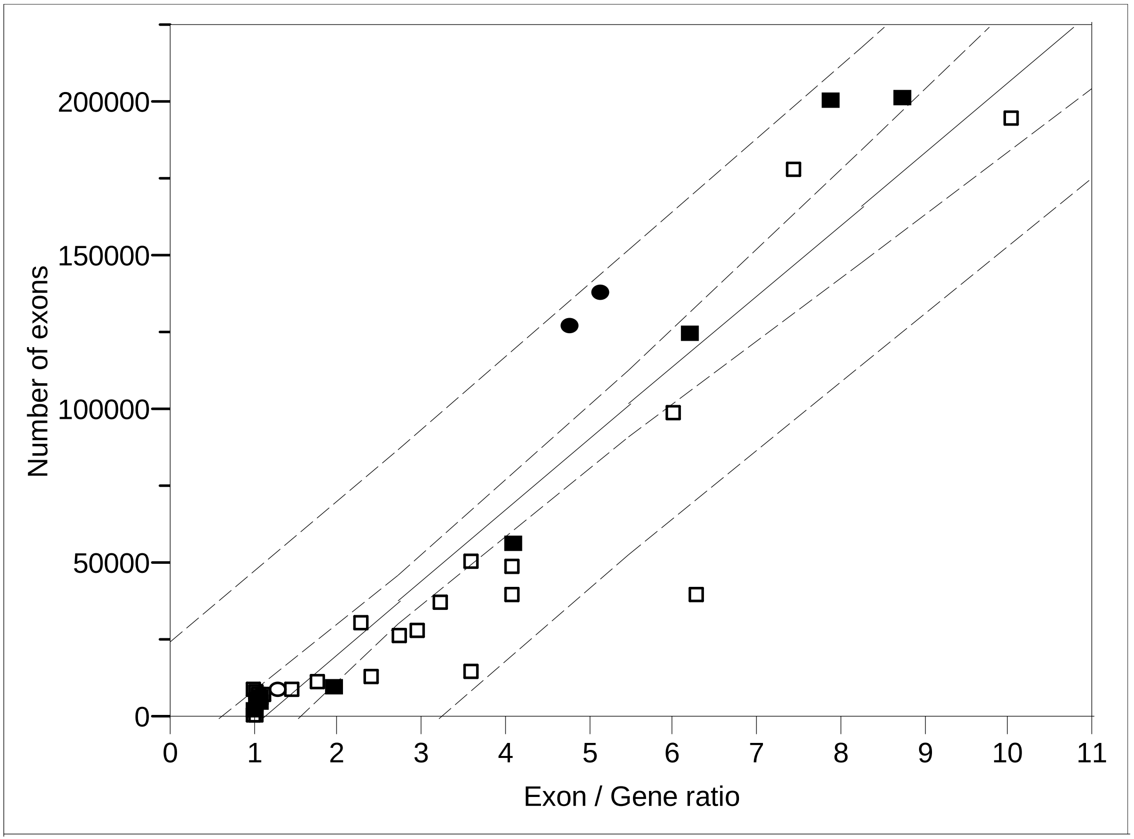


Fig. 1 Continued

A Mean (black) and median (gray) exon lengths averaged over taxonomic groups

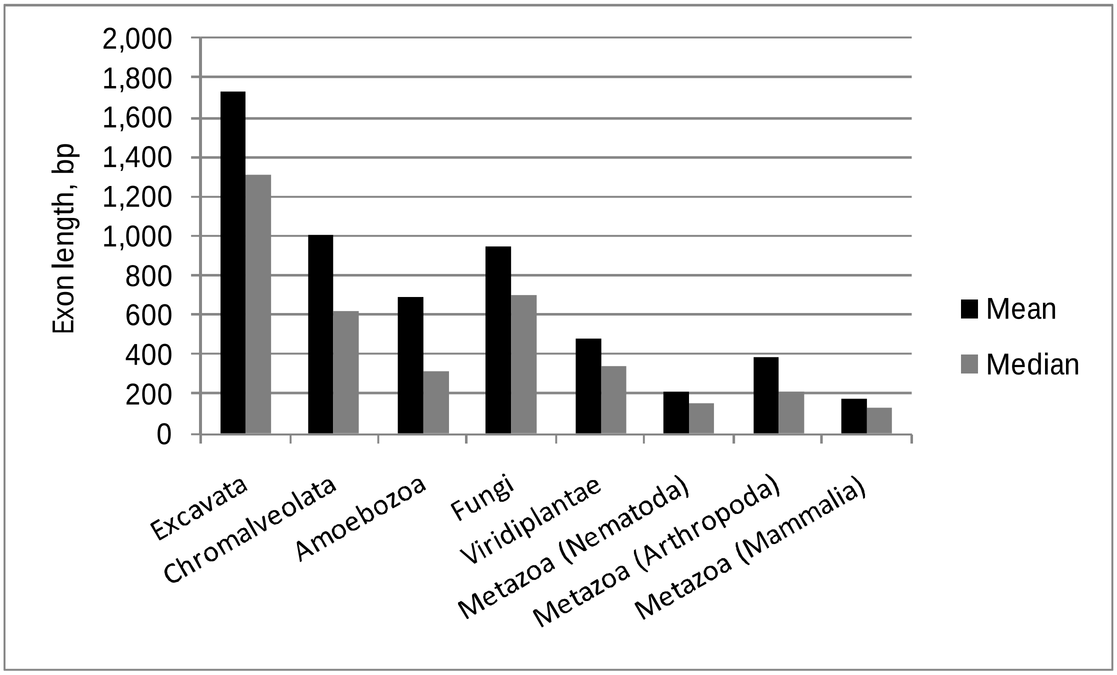


Fig. 2 Mean (black) and median (gray) exon lengths (A), number of all (black) and protein coding (gray) genes (B), exon/gene ratio (C) averaged over taxonomic groups, and ratio of alternatively spliced genes in five species (D)

B Number of all (black) and protein coding (gray) genes averaged over taxonomic groups

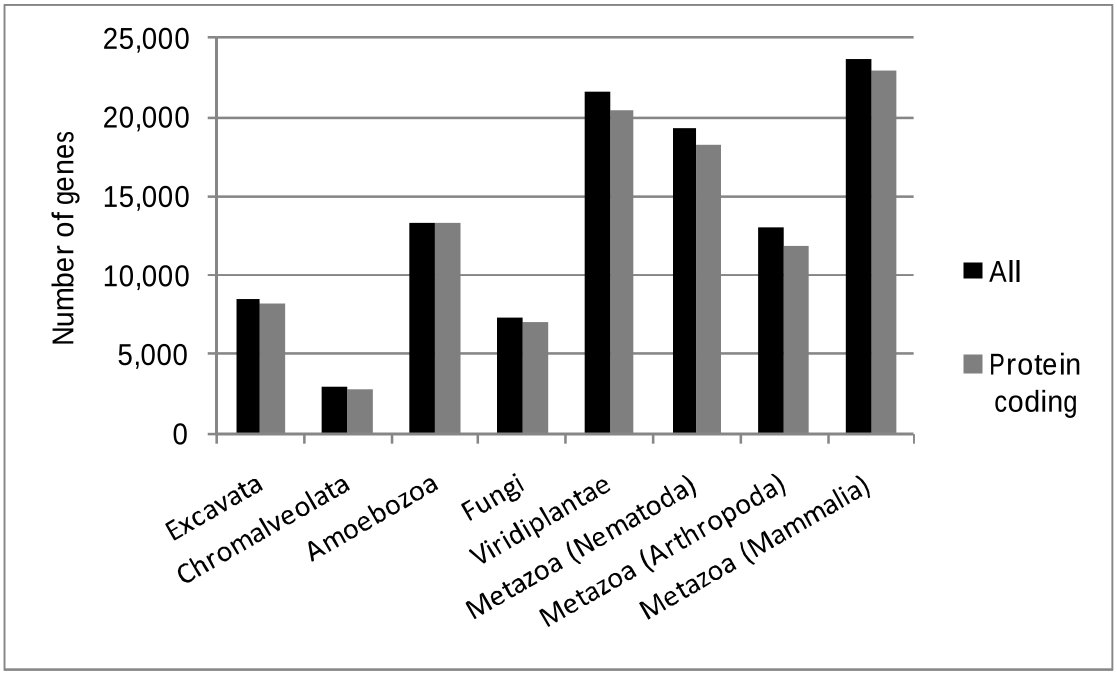


Fig. 2 Continued

C Exon/gene ratio averaged over taxonomic groups

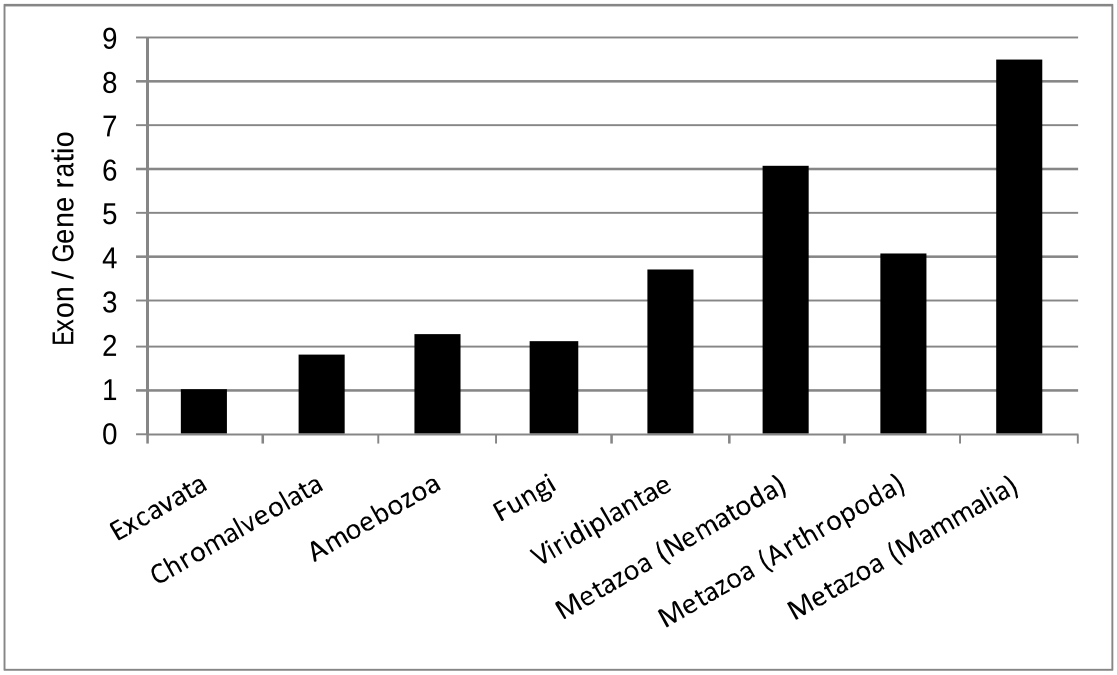


Fig. 2 Continued

D Ratio of alternatively spliced genes in five species

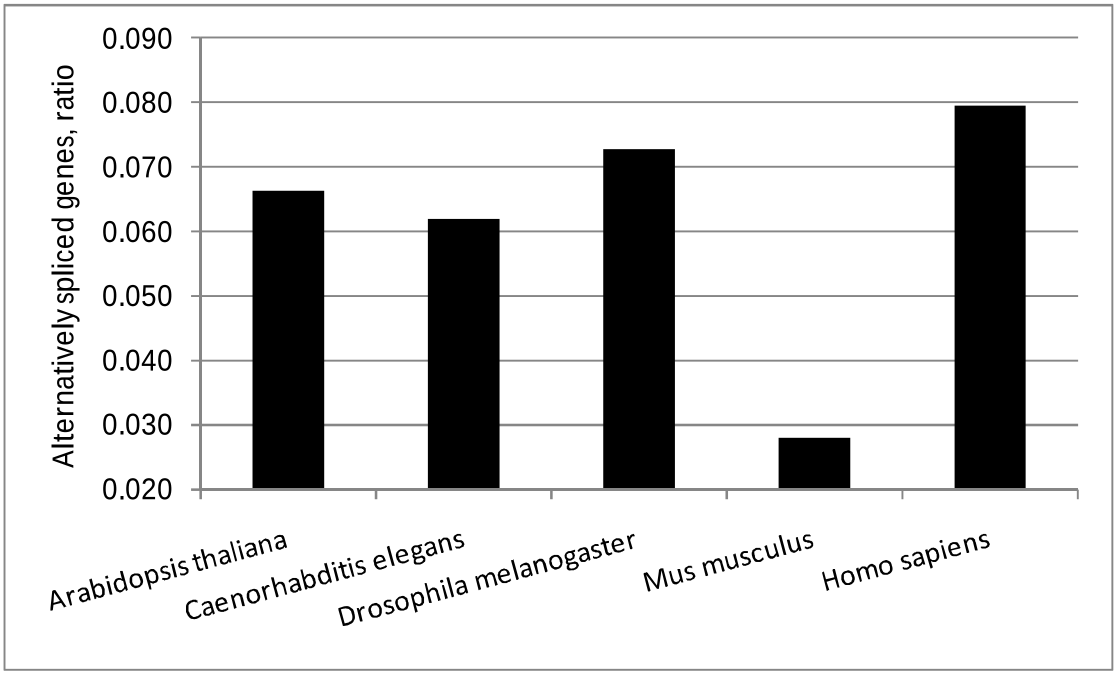


Fig. 2 Continued

Number of genes and exons in genomes. Also strong and nonlinear but positive correlations were found between the number of genes and number of exons ($r^2 = 0.897$, $r^2_{\mathrm{adj}} = 0.894$ for protein coding genes, Fig. 1D; $r^2 = 0.891$, $r^2_{\mathrm{adj}} = 0.888$ for the total number of genes, Fig. A3, Appendix A).

A very strong positive correlation was observed between the total number of genes and number of protein coding genes ($r^2 = 0.996$, $r^2_{\mathrm{adj}} = 0.996$; Fig. 1E). The total numbers of genes and protein coding genes were higher in more complex organisms (Fig. 2B).

Exon/gene ratio as a predictor. Strong positive linear correlation was observed between exon/gene ratio and the total number of exons in the genome ($r^2 = 0.864$, $r^2_{\mathrm{adj}} = 0.860$; Fig. 1F). In general, exon/gene ratio increases with the evolutionary progress of a taxonomic group (Fig. 2C). The relationship between exon/gene ratio and the number of protein coding genes was also strongly positive but rather nonlinear ($r^2 = 0.648$, $r^2_{\mathrm{adj}} = 0.638$; Fig. A4, Appendix A).

Alternative splicing. AS data were inferred from the annotated genomic data for 12 analyzed species. The five most annotated species (*A. thaliana, C. elegans, D. melanogaster, H. sapiens* and *M. musculus*) were analyzed in more detail (Table 3).

Table 3 Alternative splicing types observed in five most studied species

Species	ES	IR	A3	A5	ME	A5A3	ESA3	A5ES	MEA3	A5ME	ESES	A5ESA3	Other	N	A	R	*H*	*H*$_{max}$	*E*
Arabidopsis thaliana	106	428	960	396	4	195	11	9	0	0	15	10	34	1,787	0.066	1.138	7.38	7.49	0.985
	0.049	*0.197*	*0.443*	*0.183*	*0.002*	*0.090*	*0.005*	*0.004*	*0.000*	*0.000*	*0.007*	*0.005*	*0.016*						
Caenorhabditis elegans	407	194	527	318	34	59	9	12	1	0	104	18	86	1,251	0.062	1.177	6.95	7.13	0.975
	0.230	*0.110*	*0.298*	*0.180*	*0.019*	*0.033*	*0.005*	*0.007*	*0.001*	*0.000*	*0.059*	*0.010*	*0.049*						
Drosophila melanogaster	464	292	297	165	102	62	31	16	2	0	79	9	67	1,008	0.073	1.284	6.66	6.92	0.964
	0.293	*0.184*	*0.187*	*0.104*	*0.064*	*0.039*	*0.020*	*0.010*	*0.001*	*0.000*	*0.050*	*0.006*	*0.042*						
Mus musculus	423	128	133	72	35	32	18	5	5	1	126	5	41	714	0.028	1.064	6.40	6.57	0.974
	0.413	*0.125*	*0.130*	*0.070*	*0.034*	*0.031*	*0.018*	*0.005*	*0.005*	*0.001*	*0.123*	*0.005*	*0.040*						
Homo sapiens	1,519	92	468	262	124	41	25	19	2	0	294	20	114	1,834	0.080	1.210	7.26	7.51	0.966
	0.510	*0.031*	*0.157*	*0.088*	*0.042*	*0.014*	*0.008*	*0.006*	*0.001*	*0.000*	*0.099*	*0.007*	*0.038*						

ES – exon skipping; IR – intron retention; A3 – alternative 3' splice site; A5 – alternative 5' splice site; ME – mutually exclusive exons; N – number of alternatively spliced genes; A – alternative splicing ratio (proportion of alternatively spliced genes); R – ratio of the total number of protein products to the total number of protein genes; *H* – Shannon's index; *H*$_{max}$ – maximum possible value of Shannon's index, where for a given *n*, *H* is a maximum and equal to log*n*, when all the *P*i are equal (i.e., 1/*n*); *E* – Shannon's equitability (*E*=*H*/*H*$_{max}$). Complex cases are denoted as combinations of these abbreviations. The numbers in italics are proportions of the type relative to all types.

Among the five most common AS types, alternative 3' splice sites type (A3) was the most frequent in *A. thaliana* and *C. elegans*, but exon skipping (ES) was the most frequent in *D. melanogaster*, *M. musculus* and *H. sapiens*.

The frequency of the ES type increases following organism complexity and reaches more than 51% of all AS types in human (Fig. 3). It is accompanied by a decrease in the intron retention (IR), A3 and alternative 5' splice sites (A5) types. The most common type in *A. thaliana* is A3 (44.3%), followed by IR (19.7%).

Alternative splicing ratio. The AS ratio, that is the ratio of the alternatively spliced genes and the total number of protein coding genes was highest (0.186) in *Pan troglodytes* among all analyzed species. When only the most-annotated species were considered, the highest AS ratio was observed in human (0.080), followed by the one in *D. melanogaster* (0.073) and *A. thaliana* (0.066) (Table 3). In general, the ratio increases with evolutionary progress (Fig. 2D).

AS negatively correlated with exon length and occurred more frequently in organisms with shorter exons (Fig. A5, Appendix A; $r^2 = 0.615$). Similarly, AS increases as exon/gene ratio ($r^2 = 0.498$; Fig. A6, Appendix A) and the total number of CDSs ($r^2 = 0.725$; Fig. A7, Appendix A) increase.

Among the five most-annotated species, both Shannon’s diversity index and equitability were highest in *A. thaliana* ($H = 7.38$, $E = 0.985$; Table 3) showing high richness and evenness of distribution. Second high value ($H = 7.26$) was observed in human, but the evenness was lower ($E = 0.966$).

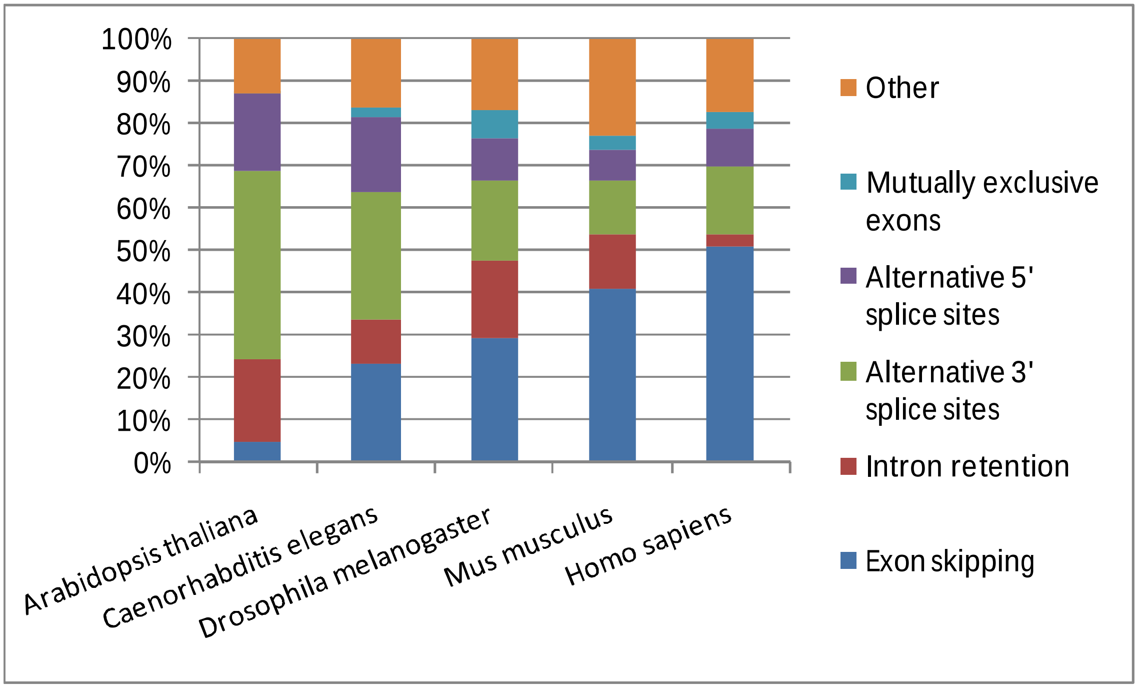


Fig. 3 Relative frequency of alternative splicing types in five species

Predictions for other species with large genomes such as Pinus taeda. The mean and median transcript lengths were practically the same (1278 bp) in *P. taeda*, based on available 99 complete CDSs (Table 4). However, the mean and median exon lengths were very different – 334.8 and 198 bp, respectively. These estimates are very preliminary and based only on 21 exons. An additional 43 complete exons were identified in partial CDSs. Their length was shorter – 166.2 bp on average, but these estimates could be biased toward shorter exons due to the PCR-biased amplicon resequencing. Based on the regression models created for the 36 complete genomes, we computed estimates for a hypothetical species with an average exon length such as the one observed in *P. taeda* (Table 2). The predicted exon/gene ratio was 4.245, very close to the observed 4.000 in *P. taeda* (95% CI on individual level: 3.146 to 5.344). The predicted total number of exons, genes and number of protein coding genes were 53,374 (95% CI on individual level: 20,093 to 86,655), 13,871 (95% CI on individual level: 4,891 to 22,850) and 13,288 (95% CI on individual level: 4,824 to 21,752), respectively. The estimates differed slightly when median exon length was used (see Table 2 for details).

2.5. Discussion

Despite significant progress in sequencing technologies, complete genomic data are still limited for eukaryotic organisms; and, more importantly, only a few extensively studied model species have been well annotated and featured. The most abundant data have been collected for microbial, fungal and some animal genomes, while vascular plants

Table 4 Exon and CDS lengths in *Pinus taeda* based on complete CDS sequences

Feature	CDS	Exon
Number	99	21
Mean length, bp	1278.3	334.8
Median length, bp	1278.0	198.0
Standard deviation	667.3	296.7

have been understudied. Only a few species have been completely sequenced and annotated in this underrepresented group so far, such as *Arabidopsis thaliana, Oryza sativa, Vitis vinifera, Physcomitrella patens,* and, recently, *Populus trichocarpa*, which remains relatively poorly annotated and featured. Therefore, due to insufficient experimental data, it is very likely that not all gene transcripts and AS products have been recorded in GenBank even for the best-studied species; this can cause underestimation of AS ratios in our study. As more experimental data are collected, the situation gradually improves with every new genome build that updates the number of genes, exons, and their locations on the chromosome. AS has different types, occurs at different developmental stages and tissues, and can be affected by environmental factors (Mano et al. 1999; Iida et al. 2004). AS is still insufficiently studied, and, therefore, not all AS events and types are well documented in the databases. Moreover, precise inference is made more difficult due to incomplete annotation, along with different stringency criteria, customary thresholds to classify true and erroneous AS events, and various gene models used in different species. To avoid these complications, we limited our analyses only to extensively studied completely sequenced genomes. However, we hope that the results obtained can be used for predictions in insufficiently studied and incompletely sequenced organisms, such as pine in our study.

Alternative splicing ratio. In general, both AS and number of genes are higher in more evolved organisms (Figs. 2B, D). However, surprisingly the AS rate was not always as high in more evolutionarily advanced species as expected and did not correlate linearly with their evolutionary progress. For instance, the ratio observed in *D. melanogaster* (0.073) was higher than in *A. thaliana* (0.066), but very close to the one in human (0.080). This could suggest that a relatively small number of genes in *D. melanogaster* compared to human (14,807 vs. 25,074, respectively) is compensated by a higher AS rate that increases proteomic and metabolomic complexity. We cannot completely exclude that the discrepancies are explained by insufficient AS data, but we can conclude in general that both the number of genes and the AS rate increase in more evolutionarily advanced species. Certainly, more experimental data are needed to increase the precision of estimates and predictions of AS ratios. For instance, a number of studies demonstrated AS in rice (McGuire et al. 2008; Campbell et al. 2006; Severing et al. 2009), but it is not documented in rice genomic data from the NCBI GenBank database, a likely indication that annotation of the rice genome is still in progress.

A substantially higher AS rate for shorter exons found in the study (Fig. A5) is consistent with previous studies that suggested that both tandem exon duplication (Kondrashov and Koonin 2001) and insertion of noncoding intron sequences (Kondrashov and Koonin 2003) could promote AS. Both within-gene duplication and AS would have less drastic effect on functionality of final proteins when they both deal with mutually exclusive exons (ME) that have shorter lengths. In addition, converting a part of an intron into an exon via AS has the risk of including a stop codon. This risk is higher when alternative exon sequences are longer.

Although not as drastic, more exons per gene is also associated with a higher AS rate (Fig. A6). This can be observed from the above described correlation of exon length and

AS because more exons per gene mean both shorter exons (assuming a constraint on the final gene product length) and more options for AS. Exon/gene ratio, similarly to AS, increases in evolutionarily advanced species (Figs. 2C, D). More advanced species also show higher numbers of CDSs. AS increases as number of CDSs increases (Fig. A7), playing an important role in creating higher proteomic complexity.

Previous studies reported the ES type of AS as the most frequent in mammals (Nagasaki et al. 2005; Sammeth et al. 2008). Our results are consistent with these findings. ES accounted for 51.0% of AS in human and 41.3% in mouse. Moreover, frequency of ES increases with complexity. Kim et al. (2007) used a modified approach that required the final number of ESTs in the compared organisms' genes to be the same to mitigate the bias in data availability for the studied species. They also found a high frequency of ES in mammals (~ 40%) and low IR (~10%).

It is the opposite in plants, where IR type was the most frequent (over 50%) in both rice and thale cress (Wang and Brendel 2006; Severing et al. 2009). The A3 was the second most frequent type, while ME was the least frequent type. Nagasaki et al. (2005) reported that IR accounted for over 42% of AS events in thale cress and 55% in rice. Although our analysis also found a very low level of ME in thale cress (0.2%), the most abundant type was A3 (44.3%) followed by IR (19.7%). Assuming that none of the classes is underrepresented in the dataset we used, this could indicate that IR tends to be overestimated in the EST/cDNA based studies, possibly due to the highest incidence of nonsense-mediated mRNA decay (NMD)-targeted products in this class. Wang and Brendel (2006) estimated that ~43% of AS events in *Arabidopsis* are potential NMD candidates, with IR showing the highest incidence of 40-48%. Conversely, Kim et al. (2007) found the rate of IR in *Arabidopsis* (~30%) less than A3 (~40%). In their study, ES accounted for approximately 5% of all types. McGuire et al. (2008) found that in *A. thaliana* IR accounted for 38.7% of splice variants, only slightly more than A3 (36.8%) and ES (7.7%) events. These results show great sensitivity to the methods and assumptions used. McGuire et al. (2008) discussed how including unspliced alignments may affect the outcomes.

Alternative evolutionary scenarios for plants. This study demonstrates that plants and animals may have used different mechanisms and strategies for developing proteomic and metabolomic complexity. The AS rate is low in plants compared to animals, whereas the number of genes is high (Figs. 2B, D). This could indicate that animals have evolved a more efficient system of managing the genomic information that allows them to increase proteomic complexity with the same or smaller number of genes. Flowering plants could have relied primarily on duplications (from exon shifting to entire chromosome or genome duplications that are common in flowering plants; Cui et al. 2006), duplication modifications and divergence. In contrast, large genomes in genus *Pinus* (class Coniferopsida) might be a result of retrotransposon expansion rather than polyploidy (Morse et al. 2009). Cui et al. (2006) found no evidence of recent genome duplications in *P. taeda* nor *P. pinaster*, and Grotkopp et al.(2004) estimated that the genome sizes varied from 22.10 pg to 36.89 pg in pines, with the putative common ancestor’s genome of 32.09 pg. Nevertheless, sporadic polyploidy has been observed in gymnosperms (for review see Ahuja 2005).

In order to check how well our regression models fit the real data, we compared the values observed in the three plants in our dataset with those predicted by the models. The total numbers of exons predicted based on the observed average exon length in *Oryza sativa* and *Arabidopsis thaliana* were underestimated in both cases (predicted 92,581 and 102,910 vs. observed 128,267 and 138,876, respectively), but the observed values were only slightly greater than the 95% confidence interval upper limit at the individual level (126,092 and 136,543, respectively; Fig. 1A). The observed number of exons in the primitive alga *Ostreococcus* ‘*lucimarinus*’ (9,767) was close to the predicted number (10,021) and fell within 95% CI.

Similarly, the observed numbers of protein coding genes (Fig. 1B) in both higher plants (26,977 in thale cress and 26,777 in rice) were only slightly higher than the upper 95% CI limit at individual level (26,282 and 25,445, respectively), and predicted values were underestimated (17,694 and 16,888, respectively). The total observed number of genes (Fig. A1) for the two species, again, fell only slightly above the 95% individual level CI (28,245 > 27,591 for thale cress and 29,102 > 26,715 for rice; predicted values: 18,480 and 17,637, respectively). In both cases the numbers observed in the alga fell within the 95% CI.

The discrepancy between Viridiplantae and other kingdoms can also be observed in the relationship between exon count and number of genes (Fig. 1D, Fig. A3), as well as exon/gene ratio and number of protein coding genes (Fig. A4). In our models the values for the two higher plants fall outside the upper 95% CI at the individual level in all these cases. Interestingly, comparison of the two evolutionarily youngest groups in kingdoms Metazoa and Viridiplantae reveals much longer exons in plants (236.8 bp in *A. thaliana* and 250.2 bp in *O. sativa*) than in mammals (ranging from 169.2 bp in *Canis lupus familiaris* to 179.4 bp in *M. musculus*), demonstrating that the processes that reduce exon length have been slower in plants. Body plan complexity is much greater in mammals than in angiosperms, and shorter mammalian exons coupled with lower number of genes could indicate greater pressure towards efficient use of gene space. The highest values of Shannon's index and equitability observed in *A. thaliana* (Table 3) indicated more even AS distribution than in four animal species, despite their higher evolutionary position. Perhaps AS is not the main mechanism in achieving the observed complexity level in plants. If AS is correlated with exon length, then longer exons in plants can imply that the pressure for greater AS rates is not as strong as in the case of higher animals. Other mechanisms, such as more frequent duplications and elevated retrotransposon activity in plants could be responsible for the high number of genes, and also greater exon lengths, through intron loss. Indeed, studies on animal species have shown a negative correlation between gene family size and AS frequency (Su et al. 2006; Hughes and Friedman 2008; Kopelman et al. 2005). This could explain not only lower rates of AS observed in plants but also the different patterns of AS forms, potentially increasing chances of NMD, a phenomenon not very well studied in plants (Campbell et al. 2006). Conversely, building upon the theory proposed by Lynch and Conery (2003), Babenko et al. (2004) suggested that intron gain/loss is not a commonly ongoing process, but rather may be triggered by certain dramatic evolutionary events that lead to long-term bottlenecks. Therefore the observed differences in exon lengths could be merely due to chance of the ancestors being affected by drastic events in the past. These conclusions seem to be supported by Sammeth et al. (2008), showing rather abrupt differences between invertebrates and vertebrates.

Current genomic data are insufficient to build separate robust regression models for plants. The conclusions about the total number of exons, the total number of genes and the total number of protein coding genes in *P. taeda* may therefore be biased; and the true values may be close to the upper 95% CI limit on the individual level, higher than the ones predicted by the proposed models (see below).

Short exons promote genomic complexity. The strong relationship between exon length and total number of exons (Fig. 1A) as well as exon length and exon/gene ratio (Fig. 1C) suggest that shorter exons increase potential for AS. Indeed, a much higher ratio of AS was observed in organisms with shorter exons (Fig. A5). However, more evolutionarily advanced organisms not only have shorter exons, but also more genes (Fig. 1B, Figs. 2A, B, and Fig. A1). The presence of shorter exons increases the potential for exon shuffling along with exon duplications; and, as a complement of AS, both increase proteomic and metabolomic complexity. It is likely that both evolved simultaneously and synergistically to amplify their effects on increasing physiological, behavioral and morphological complexity of the organisms through positive feedback loop-like mechanisms.

No statistically significant correlation was found between exon length and CDS length (Fig. A2). Since 3-dimensional protein structure and binding sites determine protein functionality, the length of the coding sequence seems to be of primary importance, and therefore the variation in the transcript length may be constrained. This could suggest that in the process of evolution, partitioning of the ancestral coding sequences has been occurring rather than extension through e.g. hypothetical stacking of coding blocks together. Such a process could have stimulated splicing out duplicated exons, eventually leading to alternatively spliced forms.

At the genome level, most of the species with less than 10,000 genes had a very small number of exons (Fig. 1D and Fig. A3). Consequently, the number of exons per gene was low in these species (Fig. 1F, Figs. 2B, C and Fig. A4). These observations show a general trend of genomic complexity increasing in evolutionarily advanced species.

The shortest exons identified in some of the analyzed species (including three plants) were only 1 bp long. We did not find any peer-reviewed publications experimentally confirming this observation. In previous studies Long et al. (1995) identified single base pair exons, and Deutsch and Long (1999) identified exons as short as 1 amino acid in a number of species including *A. thaliana* and human, although the exact length in bp is not clear. An experimental approach is necessary to find support for these structures and to verify that it is not an artifact resulting from exon/intron model assumptions. For instance, Kondrashov and Koonin (2001) used 9 bp as threshold.

Implications for Pinus taeda. Due to the small sample size, the *P. taeda* exon length estimate may be significantly biased. An alternative would be to include in the study completely sequenced exons from only partially sequenced genes. However, in this scenario shorter exons would be overrepresented due to the PCR amplicon length bias (typically a few hundred bp), making the mean and median underestimated. The observed exon/gene ratio was 4.000 based on 5 CDSs and 20 exons. This estimate is very close to the predicted exon/gene ratio of 4.245, based on the regression model, when the average exon length was the predictor or 3.658 when the median exon length was used (Table 2). The predicted values based on the average exon length for the other two higher plants analyzed were also close to the observed values (5.970 vs. 5.148 in thale cress and 5.654 vs. 4.790 in rice). The observed values for all three species fell inside the 95% CI at the individual level.

The number of protein coding genes and total number of genes expected in an organism with the average exon length of 334.8 bp (such as in *P. taeda*) is 13,871 and 13,288, respectively. These values seem to be underestimated as far as *P. taeda* is concerned, especially when compared with the other analyzed plants. Moreover, there are currently about nineteen thousand unique sequences in the NCBI UniGene database for *P. taeda*. The model severely underestimates the number of genes in the other two vascular plants described above as well. The number of protein coding genes in *A. thaliana* is underestimated by about 34.4% and *O. sativa* by about 36.9%. If *P. taeda* followed this bias, and the expected number of protein coding genes was also underestimated by approximately 35%; that would mean about 7,155 underestimated genes, which would raise the predicted number of protein coding genes to about 20,443 in this species, making this number more realistic.

Regression models that are based exclusively on the three examined plants and that follow the same logic as in the case of the 36 studied genomes also demonstrated that higher numbers are expected for loblolly pine (Table 2). The total number of genes expected would be 20,923 (upper limit is 45,975 at 95% confidence level; model significant at 93.7% confidence level) and the number of protein coding genes 19,785 (upper limit is 32,508 at 95% confidence level). These numbers seem to be more realistic when compared to the observed values in other higher plants, especially considering broad confidence intervals.

2.6. Conclusions

This study confirmed the general trend of increasing number of genes, gene products, and exons in the genome, along with higher exon/gene ratio and AS ratio as species become more complex. We demonstrated that parameters easily computable from small data samples (e.g. exon length or exon/gene ratio) are relatively good predictors of characteristics that are difficult to assess, such as total number of genes, gene products and exons. We also showed that taxonomic kingdoms may require different model calibration as their strategies to increase complexity throughout evolution have been different. As more genomic data become available and more species representing various taxonomic groups are annotated, these models can be tuned or applied to specific monophyletic groups, which will improve precision of the predictions.

3. MOLECULAR EVOLUTION OF ADAPTIVE AND DROUGHT RESISTANCE RELATED GENES IN FOUR SOUTHERN PINES FROM SUBSECTION *AUSTRALES*

3.1. Overview

Four major Southern pines, *Pinus echinata* Mill., *P. elliottii* Engelm., *P. palustris* Mill. and *P. taeda* L. are evolutionarily relatively young and closely related species. Due to their diverse habitat that covers the area from coastal plains to southern uplands of 13 southeastern states they have likely accumulated substantial variation in drought resistance loci. We searched for signatures of selection in 33 drought resistance related genes using neutrality tests such as Tajima's *D*, HKA, MK and synonymous-nonsynonymous substitutions ratio. Our study revealed statistically significant patterns of nucleotide variation that are consistent with balancing selection (e.g. in *cinnamyl alcohol dehydrogenase* and *caffeoyl CoA O-methyltransferase 1* genes) and purifying selection (e.g. in *early response to drought 3* gene), as well as combinations of different forms of selection (e.g. in *cinnamate 4-hydroxylase 1* and *putative wall-associated protein kinase* genes) and demographic events (such as bottleneck and population expansion) that possibly affected some loci.

3.2. Introduction

Loblolly (*Pinus taeda* L.), slash (*P. elliottii* Engelm.), shortleaf (*P. echinata* Mill.), and longleaf (*P. palustris* Mill.) pines are four major closely related Southern pines of the subsection *Australes* (section *Trifoliis*, genus *Pinus*). Their current natural area stretches from warm-temperate to subtropical climate of 13 southeastern states. *P. echinata* is the northernmost, and *P. elliottii* is the southernmost species, although their areas greatly overlap. Loblolly pine populations are mostly continuous. Extensive out-crossing and wind pollination result in high gene flow and low population structure at the neutral markers (Al-Rabab'ah and Williams 2002; Schmidtling et al. 1999; Gonzalez-Martinez et al. 2007; Eckert et al. 2010).

The natural range of the Southern pine populations is very broad, and they have likely accumulated much variation in adaptive trait-related genes (Schmidtling 2001) that allowed them to successfully adapt to different habitats and to expand into areas of diverse temperatures and rainfall. Specifically, they are relatively well adapted to drought conditions. Considering furthermore that the Southern pines are evolutionarily young species, we have a rare opportunity to study adaptive and evolutionary processes “in progress” via comparing their nucleotide variation.

To detect signatures of selection in nucleotide variation, a number of neutrality tests have been developed (for review see Kreitman 2000). The null hypothesis of no selection is based on the neutral theory of molecular evolution proposed by Kimura (1968), which considers mutation and genetic drift as major factors that affect nucleotide genetic variation and population genetic structure. We used the HKA (Hudson et al. 1987), Tajima’s *D* (Tajima 1989), MK (McDonald and Kreitman 1991) and synonymous and nonsynonymous nucleotide substitutions ratio (Li et al. 1985; Nei and Gojobori 1986) tests in our study; these are among the most common tests.

Tajima’s *D* statistic compares two within-species estimates of nucleotide diversity: one based on the number of segregating sites in the sample and predictions of neutral theory and the other based on the average number of pairwise nucleotide substitutions observed in the dataset. Positive values of the statistic are an indication of the excess of polymorphic alleles and may be a result of balancing or positive selection or a recent bottleneck. Negative values are an indication of the excess of rare alleles and could be a signature of a selective sweep or a recent population expansion.

The HKA test is a conservative approach based on the prediction that under neutrality the interspecific divergence would positively correlate with within-species polymorphism. The dataset should contain interspecific nucleotide sequence data from two or more genomic regions and intraspecific data from the same regions examined in both species, including population data for at least one of the species. The test examines whether the neutral mutation rates in the two loci vary significantly. The method is best suited for detecting balancing selection and processes that reduce variation, such as recent selective sweeps.

The MK test also uses interspecific data. It compares nonsynonymous nucleotide substitutions causing amino acid replacements with synonymous substitutions within the coding regions of the same locus. Under neutrality, the ratio of the fixed nonsynonymous and synonymous substitutions between species should be equal to the ratio of nonsynonymous to synonymous polymorphic substitutions within species. Although statistically simpler than HKA, the MK test is considered more powerful.

Synonymous and nonsynonymous nucleotide substitutions ratio is an alternative approach based on intraspecific variation. Typically, two equivalent approaches have been used. Li et al. (1985) proposed comparing the number of nonsynonymous substitutions per nonsynonymous site (K_A) and the number of synonymous substitutions per synonymous site (K_S), or K_A/K_S ratio. Nei and Gojobori (1986) proposed using the rate of nonsynonymous (d_N) and synonymous (d_S) substitutions, or d_N/d_S ratio. Both approaches indicate deviation from neutrality when the ratio is significantly greater than one ($K_A/K_S > 1$ or $d_N/d_S > 1$ – signature of positive selection) or smaller than one ($K_A/K_S < 1$ and $d_N/d_S < 1$ – signature of negative or purifying selection). The significance can be tested through the Z-test score, where $Z = (d_N – d_S)/(Var(d_S)+Var(d_N))^{½}$, generally, a one-tailed test; positive values indicate excess of nonsynonymous substitutions (signature of positive selection), while negative values indicate excess of synonymous substitutions (signature of negative or purifying selection).

In a previous study on loblolly pine, Brown et al. (2004) analyzed nucleotide diversity and linkage disequilibrium (LD) in 19 adaptive trait related genes in 32 individuals sampled from various locations within the species' natural range except Florida. Among them, 14 trees were first-generation selections from natural stands for the breeding program started in the mid 1950s, and 18 were second-generation breeding material. Their results demonstrated high absolute values of Tajima's D statistic in a few genes. However, the authors failed to reject neutrality.

Gonzalez-Martinez et al. (2006a) studied 18 drought-stress response related genes in loblolly pine. The sample included 32 megagametophytes. Out of a total of 31 trees, 22 were unrelated first-generation selections from the southeastern range of the species (including Florida) and 9 were second-generation selections from the parents from the Atlantic Coastal Plain provenance. Using Tajima's *D* they identified a possible selective sweep at *early response to drought 3* (*erd3*) gene, but pointed out that genetic hitchhiking or a recent population expansion could produce a similar effect. This result was not confirmed by the MK test, which is more robust to a potential bias due to demographic processes. Similarly, no robust conclusion was reached for *caffeoyl CoA O-methyltransferase 1* (*ccoaomt-1*) despite significant and positive Tajima's *D* statistic. A sliding window approach allowed for identification of a few regions in genes *putative wall-associated protein kinase* (*ppap12*) and *ug-2_498* with statistically significant Tajima's *D*. No selection acting upon amino acid sequences was identified from the ratio of nonsynonymous and synonymous substitutions.

These two studies examined in total 34 various drought tolerance, drought-stress response and wood-quality related genes in loblolly pine. Both confirmed a low level of LD, moderate nucleotide diversity, and failed to identify significant population genetic structure. We used loblolly pine primers designed in these two studies to amplify and sequence the same orthologous genes in the four Southern pine species and to use their sequences in a comparative genomic study. Our objective was to identify signatures of selection in the studied genes, given the extended data set for loblolly pine and new sequence data obtained for the other three species. The readily available genomic resources for loblolly pine, one of the most studied coniferous species, greatly helped us examine the other three Southern pines.

3.3. Materials and methods

Source of data. The DNA was extracted from megagametophytes of four pine species: loblolly (*Pinus taeda* L.), slash (*P. elliottii* Engelm.), shortleaf (*P. echinata* Mill.), and longleaf (*P. palustris* Mill.) pines. PCR primers previously developed by Brown et al. (2004) and Gonzalez-Martinez et al. (2006a) were used to amplify and resequence the total of 51 amplicons (33 genes; Table 5) following standard PCR procedures. One or two unrelated megagametophytes of each species were sequenced. To minimize sequencing errors, both forward and reverse strands were sequenced, and a consensus sequence was obtained for each individual megagametophyte using the Sequencher computer program (ver. 4.2, Gene Codes Corporation, Ann Arbor, Michigan, USA, http://www.genecodes.com/). Data representing population sets for loblolly pine (Gonzalez-Martinez et al. 2006a; Brown et al. 2004) were downloaded from the PopSet database of GenBank (http://www.ncbi.nlm.nih.gov/). These two population sets were based on different individual trees, except two trees that were sequenced in both studies. The sequences for these two trees were pruned to make a non-redundant set. This set was included in the analysis together with newly-generated sequences of other species. Three loci were sequenced in both Brown et al. (2004) and Gonzalez-Martinez et al. (2006a) studies: *ccoaomt-1*, *phenylalanine ammonia-lyase 1* (*pal-1*) and *s-adenosyl methionine synthetase 2* (*sams-2*), and their sequences were combined to expand the population sets and respective multiple sequence alignments. For *4-coumarate:CoA ligase* (*4cl*; 4th segment of the gene, *4cl-4*), *coumarate 3-hydroxylase*

Table 5 Pine species, genes and number of individual nucleotide sequences studied

Gene	Abbreviations	***P. echinata***	***P. elliottii***	***P. palustris***	***P. taeda***	***P. radiata***	***P. sylvestris***	Reference
4-coumarate:CoA ligase	*4cl*	1	1	1	1			This study
					32			Brown et al. 2004
4-coumarate:CoA ligase (amplicon 4)	*4cl-4*	1	2	2	2			This study
					32			Brown et al. 2004
			2		32	2	2	Ersoz 2006
arabinogalactan 4	*agp-4*	2	2	2	2			This study
					32			Brown et al. 2004
arabinogalactan 6	*agp-6*	2	2	1	1			This study
					32			Brown et al. 2004
arabinogalactan-like (amplicon 2)	*agp-like*	2	2	2	2			This study
					32			Brown et al. 2004
alpha tubulin	*α-tubulin*	2	2	2	2			This study
					32			Brown et al. 2004
aquaporin, membrane intrinsic protein	*aqua-MIP*	2	2	2	2			This study
					32			Gonzalez-Martinez et al. 2006a
coumarate 3-hydroxylase	*c3h*	2	2	2	2			This study
					28			Brown et al. 2004
coumarate 3-hydroxylase (amplicon 1)	*c3h-1*	2	2	2	2			This study
					28			Brown et al. 2004
					32			Ersoz 2006
cinnamate 4-hydroxylase 1	*c4h-1*	1	1	1	1			This study
					32			Brown et al. 2004
cinnamate 4-hydroxylase 1 (amplicons 1, 4, and 5)	*c4h-1*	2	2	2	2			This study
					32			Brown et al. 2004

Table 5 Continued

Gene	Abbreviations	*P. echinata*	*P. elliottii*	*P. palustris*	*P. taeda*	*P. radiata*	*P. sylvestris*	Reference
cinnamate 4-hydroxylase 2	*c4h-2*	2	2	2	2			This study
					32			Brown et al. 2004
cinnamyl alcohol dehydrogenase	*cad*	1	1	1	1			This study
					28			Brown et al. 2004
caffeoyl CoA O-methyltransferase 1	*ccoaomt-1* or *ccoaomt*	0	1	1	1			This study
					32			Brown et al. 2004
					32			Gonzalez-Martinez et al. 2006a
cinnamoyl CoA reductase	*ccr* or *ccr-1*	2	2	2	2			This study
					32			Brown et al. 2004
cellulose synthase A3	*cesA3*	1	2	2	1			This study
					32			Brown et al. 2004
caffeate O-methyltransferase 2	*comt-2*	1	1	1	1			This study
					32			Brown et al. 2004
calcium-dependent protein kinase	*cpk3*	2	2	2	2			This study
					32			Gonzalez-Martinez et al. 2006a
dehydrin 1	*dhn-1*	1	0	0	1			This study
					32			Gonzalez-Martinez et al. 2006a
dehydrin 2	*dhn-2*	2	2	2	2			This study
					32			Gonzalez-Martinez et al. 2006a
			2			2	2	Ersoz 2006
early response to drought 3	*erd3*	1	1	1	1			This study
					32			Gonzalez-Martinez et al. 2006a
			2			2	1	Ersoz 2006
glycine hydroxymethyltransferase	*glyhmt*	1	1	1	1			This study
					32			Brown et al. 2004

Table 5 Continued

Gene	Abbreviations	*P. echinata*	*P. elliottii*	*P. palustris*	*P. taeda*	*P. radiata*	*P. sylvestris*	Reference
water-stress inducible protein 3	*lp3-3*	1	1	1	2			This study
					32			Gonzalez-Martinez et al. 2006a
putative cell-wall protein	*lp5-like* or *lp5*	1	0	0	0			This study
					32			Gonzalez-Martinez et al. 2006a
metallothionein-like	*mt-like*	1	1	1	1			This study
					32			Gonzalez-Martinez et al. 2006a
			2			2		Ersoz 2006
phenylalanine ammonia-lyase 1	*pal-1* or *pal*	0	1	1	1			This study
					32			Brown et al. 2004
					32			Gonzalez-Martinez et al. 2006a
protein phosphatase 2C-like	*pp2c*	1	1	1	1			This study
					32			Gonzalez-Martinez et al. 2006a
putative wall-associated protein kinase	*ppap12*	1	1	1	1			This study
					32			Gonzalez-Martinez et al. 2006a
LIM domain protein 1 (LIM transcription factor)	*ptlim1*	1	1	1	1			This study
					32			Brown et al. 2004
LIM domain protein 2 (LIM transcription factor)	*ptlim2*	0	1	1	1			This study
					32			Brown et al. 2004
cysteine protease	*rd21A-like*	0	1	1	1			This study
					32			Gonzalez-Martinez et al. 2006a
					1	2	1	Ersoz 2006
s-adenosyl methionine synthetase 1	*sam-1*	1	1	1	1			This study
					32			Brown et al. 2004

Table 5 Continued

Gene	Abbreviations	*P. echinata*	*P. elliottii*	*P. palustris*	*P. taeda*	*P. radiata*	*P. sylvestris*	Reference
s-adenosyl methionine synthetase 2	*sams-2* or *sam-2*	1	1	1	1			This study
					32			Brown et al. 2004
					32			Gonzalez-Martinez et al. 2006a
chloroplast Cu/Zn superoxide dismutase	*sod-chl*	1	1	1	1			This study
					32			Gonzalez-Martinez et al. 2006a
unknown; drought stress responsive (University of Georgia uniscript sequence #2_498)	*ug-2_498*	1	1	0	0			This study
					32			Gonzalez-Martinez et al. 2006a

(*c3h*; 1st segment, *c3h-1*), *dehydrin 2* (*dhn-2*), *erd3*, *metallothionein-like* (*mt-like*), and *cysteine protease* (*rd21A-like*) genes additional sequences from GenBank (Ersoz 2006) were included for *P. taeda*, *P. elliottii*, *P. radiata* and *P. sylvestris* (Table 5).

Multiple alignments and coding regions assignment. Multiple nucleotide alignments were done using BioEdit (ver. 7.0.9.0; Hall 1999) and SeaView (ver. 4.0; Galtier et al. 1996) software that implements the MUSCLE algorithm (Edgar 2004). Genes that were sequenced in multiple segments using several (but usually overlapping) amplicons generated by different primer pairs were concatenated and analyzed as a single sequence with the assumption that these segments represent the same gene. Sequences with extended gaps due to missing data (i.e. resulting from poor quality of the sequencing output) were excluded if an alternative sequence from the same species was available. The coding regions were assigned following available genomic data submitted to GenBank, primarily using the population sets submitted by Brown et al. (2004) and Gonzalez-Martinez et al. (2006a). Additional genomic and EST sequences from other pines and conifers, such as *Picea sitchensis* and *Pseudotsuga menziesii* were used to define exon-intron structure, if needed.

Neutrality tests. The loblolly pine nucleotide sequences newly generated in this study were merged into a single population set for each gene together with sequences downloaded from GenBank. Neutrality tests were performed using the DnaSP software (ver. 5.00.07; Librado and Rozas 2009). Tajima’s *D* (Tajima 1989) test was run with the sliding window option, where window length and step were 100 bp and 25 bp, respectively. Indels were excluded from the analysis and were not counted in the sliding window length. The HKA test (Hudson et al. 1987) was run for all nucleotide substitutions, while substitutions only in coding regions were considered in the MK test (McDonald and Kreitman 1991; Kreitman 2000; for review see Wray et al. 2003) . These two tests were used to compare the loblolly pine set with three other Southern pine species (*P.echinata*, *P. elliottii* and *P. palustris*), and, in a few cases, also with *P. radiata* and *P. sylvestris* (Table 6). Additionally, neutrality was tested through analysis of synonymous/nonsynonymous substitutions ratio as implemented in the MEGA software (ver. 4.1; Tamura et al. 2007). *Z*-test was run to test the null hypothesis. Sites with gaps or missing data were deleted in pairwise comparisons. Standard error was computed through bootstrap with 1,000 replicates. The Nei-Gojobori's nucleotide substitution model was used to calculate d_S and d_N (Jukes and Cantor 1969).

3.4. Results

Data for 33 genes were collected, and coding regions in 32 genes were assigned (there was no available data to assign exon-intron structure for *ug-2_498*). Due to poor quality of the sequence reads *water-stress inducible protein 1* (*lp3-1*) gene was dropped from the final analysis. For the same reason some of the sequences were significantly trimmed.

Table 6 Interspecific HKA and MK neutrality tests

Gene	Amp	Number of sequences						HKA, *P*					MK, *P*				
		Piec	*Piel*	*Pipa*	*Pita*	*Pira*	*Pisy*	*Piec-Pita*	*Piel-Pita*	*Pipa-Pita*	*Pira-Pita*	*Pisy-Pita*	*Piec-Pita*	*Piel-Pita*	*Pipa-Pita*	*Pira-Pita*	*Pisy-Pita*
4cl	5	1	1	1	33	0	0	0.732	0.198	0.998	-	-	0.200	0.465	1.000	-	-
4cl-4	1	1	4	2	66	2	2	0.602	0.599	0.566	0.304	0.707	-	-	-	-	-
agp-4	1	2	2	2	34	0	0	0.804	0.562	0.801	-	-	-	-	-	-	-
agp-6	2	2	2	1	33	0	0	0.791	0.867	0.984	-	-	-	-	-	-	-
agp-like	1	2	2	2	34	0	0	-	-	0.326	-	-	-	-	-	-	-
aqua-MIP	1	2	2	2	34	0	0	-	0.090	-	-	-	-	-	-	-	-
α-tubulin	2	2	2	2	34	0	0	0.937	0.505	0.631	-	-	-	-	-	-	-
c3h	5	2	2	2	30	0	0	0.835	0.479	0.485	-	-	-	1.000	1.000	-	-
c3h-1	1	2	2	2	62	0	0	0.963	1.000	0.453	-	-	-	-	-	-	-
c4h-1	5	1	1	1	33	0	0	0.510	0.791	0.873	-	-	1.000	-	0.533	-	-
c4h-1	3	2	2	2	34	0	0	0.686	0.524	0.881	-	-	-	-	-	-	-
c4h-2	1	2	2	2	34	0	0	0.866	0.866	0.866	-	-	-	-	0.182	-	-
cad	1	1	1	1	29	0	0	0.846	0.846	0.582	-	-	-	-	1.000	-	-
ccoaomt-1 or *ccoaomt*	1	0	1	1	65	0	0	-	0.982	0.847	-	-	-	-	-	-	-
ccr or *ccr-1*	2	2	2	2	34	0	0	0.691	0.902	0.984	-	-	-	-	-	-	-
cesA3	2	1	2	2	33	0	0	0.762	0.862	0.855	-	-	-	-	-	-	-
comt-2	2	1	1	1	33	0	0	0.976	0.874	0.349	-	-	1.000	1.000	0.545	-	-
cpk3	1	2	2	2	34	0	0	0.764	0.509	0.764	-	-	-	-	-	-	-
dhn-1	1	1	0	0	33	0	0	0.868	-	-	-	-	-	-	-	-	-
dhn-2	1	2	4	2	34	2	2	0.437	0.842	0.754	0.901	0.932	0.455	-	-	0.444	0.684
erd3	1	1	3	1	33	2	1	0.649	0.855	0.790	0.649	0.870	0.333	-	0.333	0.333	0.500
glyhmt	1	1	1	1	33	0	0	0.953	0.953	0.953	-	-	-	-	-	-	-
lp3-3	1	1	1	1	34	0	0	1.000	1.000	1.000	-	-	-	-	-	-	-
lp5-like or *lp5*	1	1	0	0	32	0	0	0.665	-	-	-	-	-	-	-	-	-
mt-like	1	1	3	1	33	2	0	0.767	0.556	0.870	0.871	-	-	-	-	-	-

Table 6 Continued

Gene	**Amp**	**Number of sequences**						**HKA, *P***					**MK, *P***				
		Piec	***Piel***	***Pipa***	***Pita***	***Pira***	***Pisy***	***Piec-Pita***	***Piel-Pita***	***Pipa-Pita***	***Pira-Pita***	***Pisy-Pita***	***Piec-Pita***	***Piel-Pita***	***Pipa-Pita***	***Pira-Pita***	***Pisy-Pita***
pal-1 or *pal*	1	0	1	1	65	0	0	-	0.659	0.883	-	-	-	-	-	-	-
pp2c	1	1	1	1	33	0	0	0.304	-	-	-	-	0.286	0.333	0.250	-	-
ppap12	1	1	1	1	33	0	0	0.728	0.496	0.489	-	-	1.000	-	-	-	-
ptlim1	1	1	1	1	33	0	0	-	0.080	-	-	-	-	-	-	-	-
ptlim2	1	0	1	1	33	0	0	-	0.960	0.960	-	-	-	-	-	-	-
rd21A-like	1	0	1	1	34	2	1	-	-	-	-	-	-	-	0.400	-	1.000
sam-1	1	1	1	1	33	0	0	0.924	0.785	0.197	-	-	-	-	-	-	-
sams-2 or *sam-2*	1	1	1	1	65	0	0	0.961	0.549	0.373	-	-	-	1.000	-	-	-
sod-chl	1	1	1	1	33	0	0	0.971	0.751	0.987	-	-	-	-	-	-	-
ug-2_498	1	1	1	0	32	0	0	-	-	-	-	-	-	-	-	-	-

Note: Amp – number of amplicons; *Piec* - *Pinus echinata* Mill., *Piel* - *P. elliottii* Engelm., *Pipa* - *P. palustris* Mill., *Pita* - *P. taeda* L., *Pira* - *P. radiata* D. Don, and *Pisy* - *P. sylvestris* L.

Tajima's D. Two genes showed statistically significant positive values (Table 7). The *D* statistic was 2.08 in *cinnamyl alcohol dehydrogenase* (*cad*; P < 0.05) and 3.18 in *ccoaomt-1* (P < 0.01), mainly due to silent mutations in both cases (*P* < 0.10 in *cad* and *P* < 0.01 in *ccoaomt-1*). The result was also confirmed by the sliding window; in *ccoaomt-1 P* values in both introns were < 0.01, and in the first two exons < 0.05. In *cad*, the strongest evidence came from the first two exons and the intron (*P* < 0.05).

The sliding window approach allowed us to identify nearly significant positive *D* value in the first exon of *arabinogalactan 4* (*agp-4*; *P* < 0.10). Although overall *D* was not significant, *D* based on nonsynonymous substitutions was significantly positive (*P* < 0.05) and almost significant based on coding regions *P* < 0.10.

Highly significant positive Tajima's *D* values were observed for all synonymous substitutions in *ppap12* (*P* < 0.05) and at the 3' end of the coding region (*P* < 0.001).

Weakly positive values (*P* < 0.10) were also observed in *sams-2* at the 3' end of the coding region and the beginning of 3' untranslated region (UTR).

Similarly, in *cinnamate 4-hydroxylase 1* (*c4h-1*) the sliding window identified significantly positive *D* values for the beginning of the first exon (*P* < 0.05) despite insignificant overall *D*.

Positive but insignificant values (*P* < 0.10) were observed in certain regions of *4cl*. After adding the set of 32 *P. taeda* sequences studied by Ersoz (2006) for the fourth amplicon of the gene (*4cl-4*), the region corresponding to the end of the second intron and beginning of the third exon demonstrated significant *D* (*P* < 0.01).

Table 7 Tajima's *D* neutrality test (*Pinus taeda*)

Gene	Amp	Seq	*D*	*P*	D_{coding}	*P*	D_{syn}	*P*	D_{nonsyn}	*P*	D_{silent}	*P*	Sliding window, bp (*D*)		
													P<0.10	*P<0.05*	*P<0.01*
4cl	5	33	1.184	>0.10	0.152	>0.10	0.514	>0.10	-1.272	>0.10	1.373	>0.10	1185-1284 (1.866); 1210-1309 (1.866); 1235-1334 (1.866); 1260-1454 (1.866); 1505-1604 (1.971); 1530-1629 (1.971); 1555-1654 (1.971); 1805-2071 (1.971)		
4cl-4	1	66	1.668	>0.10	1.315	>0.10	1.315	>0.10	n.a.	n.a.	1.668	>0.10		93-192 (2.160); 118-217 (2.445); 168-267 (2.402)	143-242 (2.684)
agp-4	1	34	0.304	>0.10	1.773	<0.10	-0.240	>0.10	2.212	<0.05	-0.805	>0.10	144-243 (1.718)		
agp-6	2	33	0.222	>0.10	-0.534	>0.10	-0.551	>0.10	-0.413	>0.10	0.465	>0.10	72-171 (-1.684); 147-246 (-1.610)	122-221 (-1.895)	
agp-like	1	34	0.566	>0.10	-	-	-	-	-	-	0.566	>0.10			
α-tubulin	2	34	-1.614	<0.10	-0.799	>0.10	-0.799	>0.10	n.a.	n.a.	-1.614	<0.10	402-501 (-1.650); 427-526 (-1.686); 452-569 (-1.703); 477-594 (-1.592)		
aqua-MIP	1	34	-1.434	>0.10	-0.188	>0.10	-0.188	>0.10	n.a.	n.a.	-1.434	>0.10			
c3h	5	30	-1.641	<0.10	-1.256	>0.10	n.a.	n.a.	-1.256	>0.10	-1.428	>0.10			
c3h-1	1	62	-1.525	>0.10	-1.387	>0.10	-1.080	>0.10	-1.071	>0.10	-1.315	>0.10			
c4h-1	5	33	-0.583	>0.10	-0.236	>0.10	-0.805	>0.10	0.668	>0.10	-0.801	>0.10	1840-2043 (-1.728)	1-117 (2.126); 43-142 (2.126)	
c4h-1	3	34	-0.670	>0.10	-0.241	>0.10	-0.798	>0.10	0.648	>0.10	-1.241	>0.10		1-117 (2.149); 43-142 (2.149)	

Table 7 Continued

Gene	**Amp**	**Seq**	***D***	***P***	D_{coding}	***P***	D_{syn}	***P***	D_{nonsyn}	***P***	D_{silent}	***P***	**Sliding window, bp (*D*)**		
													P<0.10	***P<0.05***	***P<0.01***
c4h-2	1	34	-0.177	>0.10	-0.177	>0.10	0.155	>0.10	-1.068	>0.10	0.155	>0.10			
cad	1	29	2.077	<0.05	1.026	>0.10	0.293	>0.10	1.595	>0.10	1.823	<0.10		1-144 (2.094); 245-344 (2.405); 270-369 (2.094); 295-394 (2.094)	
ccoaomt-1 or *ccoaomt*	1	65	3.177	<0.01	1.565	>0.10	1.565	>0.10	n.a.	n.a.	3.177	<0.01		1-149 (2.120); 200-299 (2.089); 225-324 (2.427); 250-349 (2.504); 275-374 (2.504); 325-440 (2.504); 350-465 (2.089)	75-174 (2.692); 100-199 (2.927); 125-224 (2.927); 150-249 (2.927); 175-274 (2.750); 300-399 (2.736)
ccr or *ccr-1*	2	34	-0.535	>0.10	-0.475	>0.10	-0.475	>0.10	n.a.	n.a.	-0.535	>0.10			
cesA3	2	33	-1.699	<0.10	-1.744	<0.10	-1.744	<0.10	n.a.	n.a.	-1.699	<0.10	866-965 (-1.744)		
comt-2	2	33	0.363	>0.10	0.022	>0.10	0.006	>0.10	0.031	>0.10	0.494	>0.10			
cpk3	1	34	0.216	>0.10	0.297	>0.10	0.549	>0.10	-0.483	>0.10	0.356	>0.10			
dhn-1	1	33	-0.389	>0.10	-0.110	>0.10	-0.050	>0.10	-0.143	>0.10	-0.427	>0.10			
dhn-2	1	34	0.639	>0.10	0.186	>0.10	0.376	>0.10	-0.130	>0.10	0.912	>0.10			
erd3	1	33	-2.103	<0.05	-1.502	>0.10	n.a.	n.a.	-1.502	>0.10	-1.888	<0.05			
glyhmt	1	33	0.434	>0.10	-0.472	>0.10	-0.472	>0.10	n.a.	n.a.	0.434	>0.10			
lp3-3	1	34	-0.694	>0.10	-0.694	>0.10	n.a.	n.a.	-0.694	>0.10	n.a.	n.a.			
lp5-like or *lp5*	1	32	-0.292	>0.10	-0.256	>0.10	-0.192	>0.10	-0.323	>0.10	-0.243	>0.10			
mt-like	1	33	-0.283	>0.10	-1.272	>0.10	n.a.	n.a.	-1.272	>0.10	0.182	>0.10			
pal-1 or *pal*	1	65	-1.139	>0.10	-1.089	>0.10	-0.558	>0.10	-1.075	>0.10	-0.953	>0.10	345-497 (-1.586)		

Table 7 Continued

Gene	**Amp**	**Seq**	***D***	***P***	D_{coding}	***P***	D_{syn}	***P***	D_{nonsyn}	***P***	D_{silent}	***P***	**Sliding window, bp (*D*)**		
													P<0.10	***P<0.05***	***P<0.01***
pp2c	1	33	-1.140	>0.10	-1.140	>0.10	-1.140	>0.10	n.a.	n.a.	-1.140	>0.10			
ppap12	1	33	0.723	>0.10	0.723	>0.10	2.500	<0.05	-0.389	>0.10	2.500	<0.05	89-188 (-1.728); 114-213 (-1.728)	239-338 (2.172); 289-388 (2.500); 314-393 (2.500)	264-363 (2.731)
ptlim1	1	33	-1.552	>0.10	-	-	-	-	-	-	-1.552	>0.10			
ptlim2	1	33	-1.609	<0.10	-1.058	>0.10	-1.140	>0.10	-0.466	>0.10	-1.728	<0.10			
rd21A-like	1	34	-1.087	>0.10	-1.475	>0.10	-1.224	>0.10	-1.138	>0.10	-0.914	>0.10			
sam-1	1	33	-1.431	>0.10	-1.388	>0.10	-1.388	>0.10	n.a.	n.a.	-1.431	>0.10			
sams-2 or *sam-2*	1	65	0.904	>0.10	0.948	>0.10	0.948	>0.10	n.a.	n.a.	0.904	>0.10	293-392 (1.774); 318-417 (1.774); 343-442 (1.774)		
sod-chl	1	33	0.382	>0.10	-0.454	>0.10	-0.828	>0.10	0.106	>0.10	0.401	>0.10			
ug-2_498	1	32	-1.224	>0.10	-	-	-	-	-	-	-	-		315-437 (-2.008)	

Note: Amp – number of amplicons; Seq – number of sequences.

Negative overall value of *D* was observed in *erd3* (*P* < 0.05), based primarily on the silent substitutions.

In other genes overall *D* was not statistically significant. However, the sliding window approach revealed negative *D* for the first exon in *arabinogalactan 6* (*agp-6*; *P* < 0.05) and for the middle region of *ug-2_498* (*P* < 0.05).

Nearly significant overall *D* was found in *c3h* (*D* = -1.64; *P* < 0.10). However, in a set with additional sequences studied by Ersoz (2006), the overall *D* was insignificant in the first segment (*c3h-1*). In *alpha tubulin* (*α-tubulin*) both overall *D* and *D* based only on silent mutations were -1.61 (*P* < 0.10), where most substitutions were in the first intron. Weakly negative *D* was observed in *pal-1* (*P* < 0.10) in the 3' UTR. In *LIM domain protein 2* (*ptlim2*) the overall *D* was -1.61 (*P* < 0.10), not much different from *D* based only on silent mutations (*D* = -1.73, *P* < 0.10) which became significant (*D* = -1.89, *P* < 0.05) for sequences studied by Brown et al. (2004).

Despite strong positive *D* observed in *c4h-1* and *ppap12*, the sliding window approach revealed a weakly negative *D* (*P* < 0.10) at the end of the second intron in *c4h-1* and near the beginning of the sequence in *ppap12*.

Weakly negative values for overall *D*, as well as for *D* based on coding, synonymous and silent substitutions (*P* < 0.10), were found in *cellulose synthase A3* (*cesA3*). Based on sliding windows, this was mostly attributed to the region directly upstream of the 3' UTR (*P* < 0.10).

HKA and MK tests. None of the results produced by the HKA test were statistically significant (Table 6). The *P*-value in the comparison *P. taeda* vs. *P. elliottii* was almost significant in *aquaporin membrane intrinsic protein* (*aqua-MIP*) and *LIM domain protein 1* (*ptlim1*; *P* = 0.09 and 0.08, respectively). Similarly, no significant results were observed in the MK test.

Synonymous and nonsynonymous substitutions ratio. For the loblolly pine sets, synonymous and nonsynonymous ratio test showed highly significant values in *4cl* (*P* = 0.001; although the *P*-value changed to 0.057, when only the fourth segment with the expanded population set was analyzed), *cinnamate 4-hydroxylase 2* (*c4h-2*; *P* = 0.019), and *putative cell-wall protein* (*lp5-like*; *P* = 0.004; Table 8). In several other genes the *P*-values were between 0.050 and 0.100, i.e. in *ccoaomt-1* (*P* = 0.087), *cesA3* (*P* = 0.051), and *dhn-2* (*P* = 0.080). In other species the sample size was very limited and consisted of sets from 2 to 4 individuals. In *P. palustris*, the *P*-value was 0.046 for *c4h-1* (amplicons 1, 4 and 5), and 0.086 in *4cl-4*. In *P. echinata* 0.075 and 0.095 for *c4h-2* and *cinnamoyl CoA reductase* (*ccr*), respectively. In *P. elliottii*, the *P*-value was also relatively low, but still insignificant for *dhn-2* (*P* = 0.088) and *erd3* (*P* = 0.076). In many cases the values for species other than loblolly pine could not be computed due to an insufficient number of individuals.

Table 8 Ratio of nonsynonymous (d_N) and synonymous (d_S) nucleotide substitutions

Gene	**Amp**	***P. echinata***		***P. elliottii***		***P. palustris***		***P. taeda***		***P. radiata***		***P. sylvestris***	
		P	***Z***	***P***	***Z***	***P***	***Z***	***P***	***Z***	***P***	***Z***	***P***	***Z***
4cl	5	n/c	n/c	n/c	n/c	n/c	n/c	0.001	-3.286	-	-	-	-
4cl-4	1	n/c	n/c	1.000	0.000	0.086	-1.730	0.057	-1.924	1.000	0.000	0.154	-1.436
agp-4	1	1.000	0.000	0.461	-0.739	0.311	1.018	0.343	0.951	-	-	-	-
agp-6	2	0.560	-0.584	1.000	0.000	n/c	n/c	0.285	-1.073	-	-	-	-
agp-like	1	1.000	0.000	1.000	0.000	0.336	0.967	1.000	0.000	-	-	-	-
α-tubulin	2	0.311	-1.017	0.348	-0.942	1.000	0.000	0.349	-0.940	-	-	-	-
aqua-MIP	1	1.000	0.000	1.000	0.000	1.000	0.000	0.210	-1.261	-	-	-	-
c3h	5	0.291	-1.062	0.149	-1.453	0.307	-1.026	0.162	1.406	-	-	-	-
c3h-1	1	1.000	0.000	0.305	-1.030	1.000	0.000	0.744	0.328	-	-	-	-
c4h-1	5	n/c	n/c	n/c	n/c	n/c	n/c	0.122	-1.559	-	-	-	-
c4h-1	3	0.125	-1.547	0.289	-1.064	0.046	-2.018	0.331	-0.975	-	-	-	-
c4h-2	1	0.075	-1.796	1.000	0.000	1.000	0.000	0.019	-2.386	-	-	-	-
cad	1	n/c	n/c	n/c	n/c	n/c	n/c	0.425	-0.801	-	-	-	-
ccoaomt-1 or *ccoaomt*	1	-	-	n/c	n/c	n/c	n/c	0.087	-1.724	-	-	-	-
ccr or *ccr-1*	2	0.095	-1.681	0.146	-1.465	0.307	-1.027	0.133	-1.513	-	-	-	-
cesA3	2	n/c	n/c	0.319	1.000	0.312	-1.016	0.051	-1.972	-	-	-	-
comt-2	2	n/c	n/c	n/c	n/c	-	-	0.230	-1.206	-	-	-	-
cpk3	1	1.000	0.000	0.231	-1.203	1.000	0.000	0.107	-1.626	-	-	-	-
dhn-1	1	n/c	n/c	-	-	-	-	0.131	-1.519	-	-	-	-
dhn-2	1	1.000	0.000	0.088	-1.721	1.000	0.000	0.080	-1.768	0.298	1.045	0.509	-0.663
erd3	1	n/c	n/c	0.076	-1.789	n/c	n/c	0.152	1.443	1.000	0.000	n/c	n/c

Table 8 Continued

Gene	**Amp**	***P. echinata***		***P. elliottii***		***P. palustris***		***P. taeda***		***P. radiata***		***P. sylvestris***	
		P	***Z***	***P***	***Z***	***P***	***Z***	***P***	***Z***	***P***	***Z***	***P***	***Z***
glyhmt	1	n/c	n/c	n/c	n/c	n/c	n/c	0.181	-1.345	-	-	-	-
lp3-3	1	n/c	n/c	n/c	n/c	n/c	n/c	0.233	1.200	-	-	-	-
lp5-like or *lp5*	1	n/c	n/c	n/c	n/c	n/c	n/c	0.004	-2.973	-	-	-	-
mt-like	1	n/c	n/c	1.000	0.000	n/c	n/c	0.159	1.418	1.000	0.000	-	-
pal-1 or *pal*	1	-	-	n/c	n/c	n/c	n/c	0.383	-0.875	-	-	-	-
pp2c	1	n/c	n/c	n/c	n/c	n/c	n/c	0.293	-1.056	-	-	-	-
ppap12	1	n/c	n/c	n/c	n/c	n/c	n/c	0.231	-1.204	-	-	-	-
ptlim1	1	n/c	n/c	n/c	n/c	n/c	n/c	1.000	0.000	-	-	-	-
ptlim2	1	-	-	n/c	n/c	n/c	n/c	0.876	-0.157	-	-	-	-
rd21A-like	1	-	-	n/c	n/c	n/c	n/c	0.143	-1.473	1.000	0.000	n/c	n/c
sam-1	1	n/c	n/c	n/c	n/c	n/c	n/c	0.136	-1.502	-	-	-	-
sams-2 or *sam-2*	1	n/c	n/c	n/c	n/c	n/c	n/c	0.186	-1.330	-	-	-	-
sod-chl	1	n/c	n/c	n/c	n/c	n/c	n/c	0.617	-0.501	-	-	-	-
ug-2_498	1	-	-	-	-	-	-	-	-	-	-	-	-

Note: Amp – number of amplicons; n/c – non-computable.

3.5. Discussion

Southern pines are an evolutionarily relatively young and closely related group. It has been hypothesized that during the last glacial period that ended about 15,000 years ago (equivalent to about 500 generations of loblolly pine), their range was within the regions of central Florida and the Caribbean (Wells et al. 1991; Schmidtling et al. 1999; Jackson et al. 2000). In addition, Mexico and Southern Texas (including Lost Pines) have been proposed as a western refuge of *Pinus taeda* (Al-Rabab'ah and Williams 2004; Wells et al. 1991; Schmidtling et al. 1999). It implies two historically separated refuges of loblolly pine: east and west of the Mississippi river, respectively (Al-Rabab'ah and Williams 2002).

Therefore, in the case of Southern pines, recent demographic expansion could seriously affect nucleotide variation and complicate the search for signatures of selection caused by adaptive processes. Both demographic events and selection can leave similar signatures in the genomes that are often very difficult to dissect. In this study multiple tests were applied to examine the data from multiple perspectives. None of them, however, allowed for robust conclusions regarding adaptive (positive) selection.

Although dissection of selection from recent demographic events could be facilitated through interspecific comparisons (Kreitman 2000; Hudson et al. 1987; McDonald and Kreitman 1991), our dataset was still too limited. Only the *P. taeda* population set had sufficient representation. In the other three species, more than two samples per gene were available only in *P. elliottii*. Both MK and HKA tests failed to reject neutrality in all loci. The evidence of selection was the most pronounced in *aqua-MIP* and *ptlim1*, but still insignificant (HKA test, P = 0.09 and 0.08, respectively).

The ranges of the four Southern pines are largely overlapping. Therefore, selection pressure and molecular evolution in adaptive traits genes could be similar. An outgroup species which faces adaptive challenges different than the four studied pines could be very informative in the case of these interspecific tests.

Evidence of balancing selection. Several genes showed positive values of Tajima's *D*, a result that is consistent with balancing or positive selection. The overall Tajima's *D* was positive and significant in *ccoaomt-1* and *cad* (P < 0.01 and < 0.05, respectively). The *Z*-test score was negative in both genes and nearly significant in *ccoaomt-1* (P = 0.087), but insignificant in *cad*. These results are inconsistent and difficult to interpret because they may reflect the combination of both factors – demographic and selective. This pattern could indicate balancing selection acting in coding (*cad*) and both coding and noncoding (*ccoaomt-1*) regions, as well as slightly negative selection in combination with fast recent expansion following a bottleneck.

Brown et al. (2004) attributed the high positive value of *D* in *ccoaomt-1* and *cad* to the predominant presence of two haplotypes with multiple fixed differences. Gonzalez-Martinez et al. (2006a) presented a similar conclusion regarding *ccoaomt-1*. In the case of *cad*, the expanded set with an additional sequence obtained in this study only slightly affected the *D* value. In the case of *ccoaomt-1*, both population sets from the abovementioned studies were merged with the newly sequenced data in this study and slightly increased *D* (*D* = 3.18 vs. 2.81 and 2.49 in previous studies, respectively). Along with the *Z*-test outcomes, this result leads to a more elaborate and complex conclusion.

The sliding window approach was useful in cases where overall Tajima's *D* was not significant. *Agp-4* showed statistically significant Tajima's *D* due to nonsynonymous substitutions (P < 0.05) in the last exon right before the 3' UTR. In *ppap12* Tajima's *D* was significant for the synonymous substitutions (*P* < 0.05), and sliding window detected a region with *P* < 0.01. In *sams-2* the sliding window showed *P* < 0.10 mostly over the 3' UTR. However, other tests failed to reject neutrality in these three genes. Tajima's *D* was significantly positive (*P* < 0.05) in a short region of the first exon of *c4h-1*. The *Z*-test also rejected neutrality at this locus in *P. palustris* (*P* = 0.046); however, the result is not very reliable because the set consisted of only two individuals. This pattern indicates that some regions of the locus can be under selection, while other regions are selectively neutral, and that different forms of selection can affect different regions of the same gene. Therefore, overall *D* statistic and lack of significance can be misleading sometimes due to averaging opposite effects.

In *4cl* overall Tajima's *D* was positive but not significant despite highly significant negative *Z*-test score (*P* = 0.001). The sliding window approach attributed much of the positive Tajima's *D* signal to sites within introns (*P* < 0.10). After including an additional 32 *P. taeda* sequences from population sets for the *4cl-4* segment studied by Ersoz (2006), the Tajima's *D* raised (*P* < 0.01 around intron-exon junction) and *Z*-test became less significant (*P* = 0.057). This pattern could indicate weak purifying selection acting upon the introns and reducing variability in closely linked regions of the gene. This can indicate that introns are not entirely neutral and can be under balancing or purifying selection (Collins 1988; Castillo-Davis et al. 2002; Carvalho and Clark 1999), acting to lower the cost of transcription, ensure correct splicing and appropriate level of affinity for regulatory factors.

In general, despite some significant positive Tajima's *D* values, the *Z*-test failed to find any significant signs of strong positive selection. This is consistent with the previous studies on loblolly pine (Brown et al. 2004; Gonzalez-Martinez et al. 2006a).

Evidence of purifying selection. Apart from *erd3*, the only gene with statistically significant overall Tajima's *D* ($P < 0.05$), signatures of possible purifying or negative selection were found also in a few other loci. In *cesA3* the Tajima's *D* score was almost significantly negative, which was also in agreement with the nearly significant outcome of the *Z*-test.

When five amplicons in *c4h-1* were considered, apart from regions showing positive *D* values, an area in an intron showed negative *D* with $P < 0.10$. Similarly, in *ppap12* a part of a coding region was identified as under possible selection with $P < 0.10$ for Tajima's *D* sliding window test. These are very interesting examples of how different sections of one gene can be affected by various potentially opposite factors.

Two loci (*ug_2-498*, and coding region in *agp-6*) demonstrated negative Tajima's *D* values when sliding window was applied, although other tests did not support these findings. Nearly significant negative values were identified in *α-tubulin*, *c3h*, *pal-1*, and in *ptlim2*. These results might show the possibility of a weak purifying selective pressure that was not detectable by other tests.

In addition, purifying selection was indicated by the *Z*-test at 95% confidence level in *c4h-2* and *lp5-like* genes in *P. taeda*. Two other genes demonstrated nearly significant values ($P < 0.10$; *dhn-2* in *P. taeda* and *ccr* in *P. echinata*). Although these results were not supported by significant results from other tests, the *Z*-test is considered as very robust and may indicate purifying selection in these cases.

Recent population expansion vs. selection. Perhaps the biggest challenge in studies on selection is discriminating between natural selection and demographic events. In this study some genes demonstrated negative Tajima's *D* values due to silent substitutions, while *Z*-test values were not statistically significant. In others, neutrality was not rejected. Although negative Tajima's *D* values could indicate also a recent population expansion, when coupled with significant or nearly significant values of the *Z*-test they could be a strong signature of purifying selection. Similarly, positive Tajima's *D* values alone could indicate a fast expansion following a bottleneck, but when accompanied by significant or nearly significant negative *Z*-test values they can be a strong signature of balancing selection. In many cases the strongest evidence for Tajima's *D* values came either from synonymous/silent substitutions or was strongly localized and detected via sliding window.

Highly localized significant Tajima's *D* values could be an indication of selection rather than a result of recent population expansion. Gonzalez-Martinez et al. (2006a) found no or little evidence for population substructure or recent expansion in the studied area using 21 unlinked nuclear microsatellite markers representing most of loblolly pine linkage groups. F_{ST} was also low among the three studied regions. Al-Rabab'ah and Williams (2002) found only slight population differentiation across the loblolly pine range, but identified genetic differentiation between the two main parts of the loblolly pine range, i.e. east and west of the Mississippi river, respectively. If we assume, however, that the observed pattern is due to recent post-glacial expansion (demographic event), then this genome-wide effect would be observed in most if not all studied loci, while selection usually affects only a few genes, and its effect could be very different depending on the form of selection.

The wide spectrum of results from these tests is an indication of various factors influencing pine genomes. It is often very difficult, if not impossible, to discriminate the demographic effects from signatures of selection. It is evident that the processes shaping variation on the molecular level are not homogeneous. Highly significant positive Tajima's D scores in certain loci and highly significant negative D in others cannot be easily explained by the population expansion alone. Very likely, both population expansion and selective pressure for more efficient water use have been acting simultaneously. Further studies focusing not only on *P. taeda* but also on other Southern pines and outgroup species that face different environmental challenges may help to resolve this problem.

3.6. Conclusions

Despite extensive sequencing efforts, the amount of data available for studies of evolution at the molecular level in pines is still limited. In this study the data for *P. taeda* came from an expanded data set, but is still based mostly on a little more than 30 individuals for most loci. However, despite a relatively small sample size, we demonstrated that various functional parts of a gene can have different signatures of selection that can be opposite and cancel each other at the locus level. To make better use of the available tests, a pine species that faces other adaptive challenges should be included for interspecific comparisons.

4. PHYLOGENETIC RELATIONSHIPS BETWEEN FOUR MAJOR SOUTHERN PINE SPECIES FROM SUBSECTION *AUSTRALES,* GENUS *PINUS*

4.1. Overview

The phylogenetic relationships between four closely related Southern pines, *Pinus echinata* Mill., *P. elliottii* Engelm., *P. palustris* Mill., and *P. taeda* L. from subsection *Australes* have not been unambiguously classified. These evolutionarily young species share the habitat, face similar adaptive challenges, and were similarly affected by the recent ice age. In addition, similar phenology promotes hybridization documented within this group. Using Maximum Parsimony, Maximum Likelihood, and Bayesian Inference methods, 12 nuclear loci (*4-coumarate:CoA ligase*, *arabinogalactan 6*, *cinnamyl alcohol dehydrogenase*, *caffeoyl CoA O-methyltransferase 1*, *cellulose synthase A3*, *dehydrin 2*, *early response to drought 3*, *putative cell-wall protein*, *metallothionein-like*, *phenylalanine ammonia-lyase 1*, *cysteine protease*, and *s-adenosyl methionine synthetase 2*) were examined in these and three other pine species. The results demonstrated that interpretation of phylogenetic relationships between the Southern pines depends heavily on the subsets of genes selected. Alternative and in some cases conflicting phylogenies were reconstructed.

4.2. Introduction

Pines represent genus *Pinus* (order Coniferales, family Pinaceae) that consists of 110-120 species and constitutes a nearly ubiquitous group across the Northern Hemisphere (Eckert and Hall 2006). They belong to the most important crops in the USA (USDA Forest Service) and worldwide. Pines are keystone species; pine forests play a very important ecological role and provide important habitat for numerous species. Recent estimates suggest that *Pinus* diverged from their most recent common ancestor approximately 123-156 MYA (Gernandt et al. 2008) or maximum 225 MYA (Eckert and Hall 2006).

Four major Southern pines investigated in this study, loblolly (*Pinus taeda* L.), slash (*P. elliottii* Engelm.), shortleaf (*P. echinata* Mill.), and longleaf (*P. palustris* Mill.) belong to subsection *Australes* (section *Trifoliis*, genus *Pinus*). Pines from this subsection are thought to have begun to diverge 5-18 MYA (Willyard et al. 2007; see Axelrod 1986 for a discussion on fossil records), and, therefore, are a relatively young group with eleven species according to the traditional classification (Little and Critchfield 1969). Although their current habitats greatly overlap, stretching across 13 Southern states, *P. palustris* and *P. elliottii* could have been separated during the Pleistocene, and *P. taeda* was constricted to two refugia, while *P. echinata*'s range was probably continuous (Schmidtling 2003). The close relationship between these species is supported by natural hybridization that occurs between *P. taeda* and *P. echinata* (Smouse and Saylor 1973; Mergen et al. 1965) and between *P. palustris* and *P. elliottii* (Mergen 1958; see Price 1989 for examples of other natural hybrids in pines). Recent phylogenetic studies demonstrated that dissecting the ancestry within this group is problematic (Grotkopp et al. 2004; Eckert and Hall 2006; Gernandt et al. 2005).

Modern studies on classification of pines began in the early twentieth century. Pioneering work by Shaw (1914) laid foundations for the later studies by Little and Critchfield (1969), which became a classic reference in pine phylogeny. Subsequent development of molecular techniques provided researchers with genetic markers, such as allozymes (e.g. Shurkhal et al. 1992; Wu et al. 1999; Krutovskii et al. 1994), random amplified polymorphic DNA (e.g. RAPD; Wu et al. 1999), and later more informative DNA sequence markers, such as SNPs (Gernandt et al. 2005; Gernandt et al. 2001).

The taxonomic classification of *Pinus* has been fine-tuned multiple times. New molecular data have helped to verify the existing views on the relationships within the genus. Regarding subsection *Australes*, twenty-one morphological characters were used in the study by Adams and Jackson (1997). They found a very close relationship between *P. taeda*, *P. pungens* and *P. palustris* but failed to infer which one was more ancestral. Grotkopp et al. (2004) studied relationships between genome sizes and phylogeny, environmental factors, and biological traits. They used the supertree approach and confirmed the tight relationship between *P. palustris* and *P. taeda*. They suggested that *P. pungens, P. echinata, P. elliottii*, and *P. radiata* are more distant to this clade. This hypothesis was challenged by Gernandt et al. (2005). The strict consensus tree for 101 species based on 2 chloroplast genes demonstrated closer relationship between *P. taeda* and *P. pungens*, placing *P. echinata*, *P. elliottii* and *P. palustris* as sister taxa to this clade. Eckert and Hall (2006) used 4 chloroplast loci to study phylogeny of 83 pines. They confirmed tight relationships between Southern pines, but *P. taeda*, *P. elliottii* and *P. pungens* were grouped in one clade, while *P. radiata*, *P. palustris* and *P. echinata* were in another. Apparently, more analyses with additional nuclear loci need to be done to resolve these controversies. In our study, 12 nuclear loci were used including newly sequenced; that is 3-4 times more than the number of genes used in other studies.

The objective of this study was to refine the phylogenetic relationships between *P. echinata*, *P. elliottii*, *P. palustris* and *P. taeda*, employing twelve nuclear protein coding genes. *P. pinaster* (subsection *Pinaster* or *Pinus*, depending on different classifications, respectively), *P. sylvestris* (subsection *Pinus*) and *P. radiata* (section Pinus, classification to a subsection is controversial; Millar 1999) were used as outgroups.

4.3. Materials and methods

Source of data, species selection and outgroup species identification. The NCBI GenBank was scanned for nucleotide sequences available in other pines for the genes studied in Chapter 2 of this dissertation. Three pine species, *P. pinaster*, *P. sylvestris*, and *P. radiata*, were identified with 5, 9 and 12 common genes, respectively (Table 9). Nucleotide sequence data for *P. radiata* were the most complete, and, therefore, it was considered the best candidate for an outgroup for the Southern pines in the study. Analysis based on all 12 genes could be performed for this group of five species. Therefore, in the first stage of the analysis we examined the feasibility of using *P. radiata* as an outgroup. We identified eight shared genes that could be used for all seven

Table 9 Pine species, genes, and 4 combinations of data studied

Gene	Abbreviation	*P. echinata*	*P. elliottii*	*P. palustris*	*P. taeda*	*P. pinaster*	*P. radiata*	*P. sylvestris*
4-coumarate:CoA ligase (amplicons 1-4)	*4cl*	+	+	+	+	-	AY634350.1	EU392780.1
arabinogalactan 6 (amplicon 1)	*agp-6-1*	+	+	+	+	-	AY634318.1	-
cinnamyl alcohol dehydrogenase	*cad*	+	+	+	+	-	AF060491.1	-
caffeoyl CoA O-methyltransferase 1	*ccoaomt-1* or *ccoaomt*	-	+	+	+	AM502291.1	EU394088.1	EU394089.1
cellulose synthase A3 (amplicon 1)	*cesA3-1*	+	+	+	+	-	EU392879.1	EU392880.1
dehydrin 2	*dhn-2*	+	+	+	+	EU020010.1	EU394115.1	EU394116.1
early response to drought 3	*erd3*	+	+	+	AY874639.1*	EU020011.1	EU394093.1	EU394094.1
putative cell-wall protein	*lp5-like*	+	-	-	AY867662.1	-	EU394124.1	EU394125.1
metallothionein-like	*mt-like*	+	+	+	+	-	EU394130.1	-
phenylalanine ammonia-lyase 1	*pal-1* or *pal*	-	+	+	+	EU120508.1	EU394102.1	EU394103.1
cysteine protease	*rd21A-like*	-	+	+	AY867788.1*	EU020015.1	EU394108.1	EU394109.1
s-adenosyl methionine synthetase 2	*sams-2* or *sam-2*	+	+	+	+	-	EU394098.1	EU394099.1

Note: "+" – sequences generated in this study; "-" – data not available; * – sequences generated in this study were used for 8 genes and 7 species combination.

species (combination 8-7; Table 10 and Fig. 4) and could be a good compromise between the number of species and the number of nucleotide sites.

The second stage involved three configurations of the dataset, representing various combinations of number of genes and species. All 12 genes were studied in six species (excluding *P. pinaster*; combination 12-6), 9 genes in 5 species (excluding *P. pinaster* and *P. radiata*; combination 9-5), and 5 genes in all 7 species (combination 5-7; Table 10 and Figs. 5-7). In all analyses *P. sylvestris* and *P. pinaster* (if included) were considered as outgroups. At this stage, two new sequences of *P. taeda* sequenced in this study and described in Chapter 2 of this dissertation were replaced with longer sequences downloaded from the NCBI GenBank, i.e. *early response to drought 3* (*erd3*) and *cysteine protease* (*rd21A-like*), and an additional sequence was included for *putative cell-wall protein* (*lp5-like*) gene.

Multiple alignments. The DNA sequences were aligned using BioEdit ver. 7.0.9.0 (Hall 1999) and ClustalW multiple alignment algorithm (Thompson et al. 1994), and fine-tuned manually. Four amplicons for *4-coumarate:CoA ligase* (*4cl*) gene sequenced in this study were concatenated. Conversions from FASTA to PHYLIP and NEXUS formats were done using SeaView ver. 4.0 (Galtier et al. 1996). Following the alignment, the sequences for all genes for each species were concatenated.

Table 10 Six combinations of data studied

Gene	**Combination of number of genes (first number) and species (second number)**					
	8-7	**12-6**	**9-5**	**5-7**	**3-6 a**	**3-6 b**
4cl	+*	+	+	-	-	+
agp-6-1	-	+	-	-	-	-
cad	-	+	-	-	-	-
ccoaomt-1 or *ccoaomt*	+	+	+	+	-	+
cesA3-1	+	+	+	-	-	+
dhn-2	+	+	+	+	+	-
erd3	+	+	+	+	+	-
lp5-like	-	+	+	-	-	-
mt-like	-	+	-	-	-	-
pal-1 or *pal*	+	+	+	+	-	-
rd21A-like	+	+	+	+	+	-
sams-2 or *sam-2*	+	+	+	-	-	-

Note: "+" and "-" mean that gene was either included or not in the combination, respectively; * – only amplicon 4 was used (*4cl-4*).

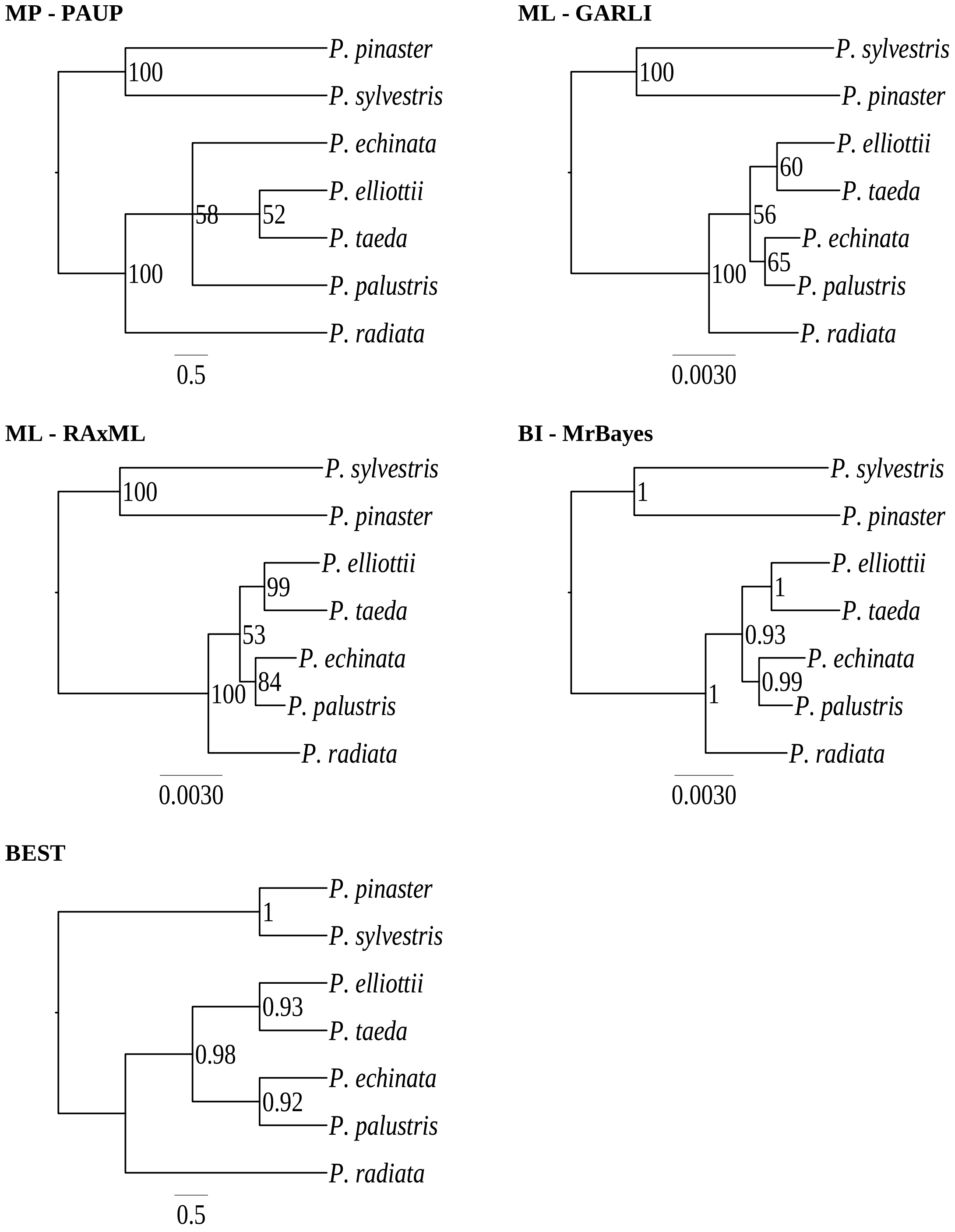


Fig. 4 Cladograms (MP - PAUP and BEST) and phylograms (ML - GARLI, ML - RAxML, and BI - MrBayes) for the dataset of 8 genes and 7 species. Bootstrap values (MP and ML) and posterior probability values (BI and BEST) are shown at the nodes

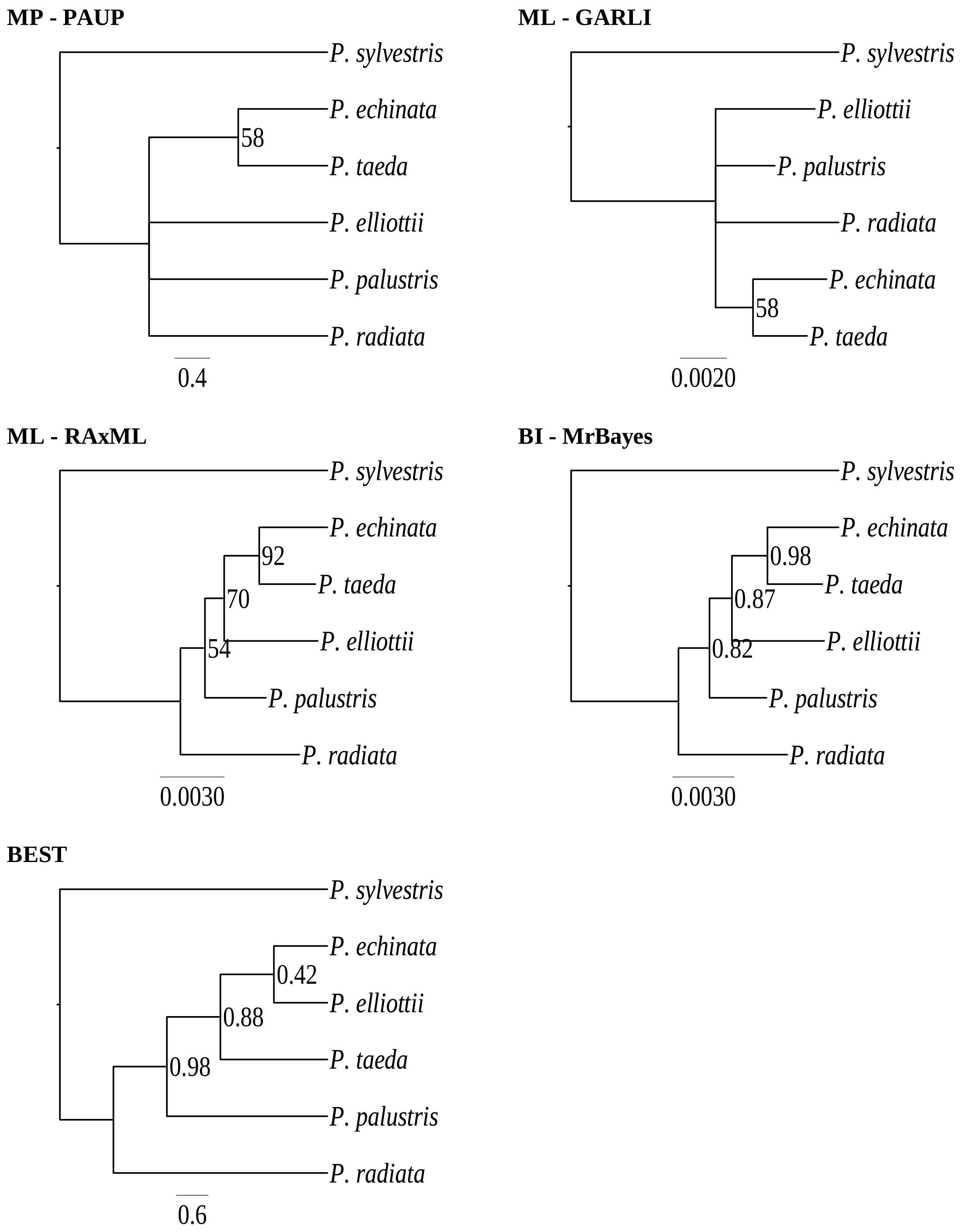


Fig. 5 Cladograms (MP - PAUP and BEST) and phylograms (ML - GARLI, ML - RAxML, and BI - MrBayes) for the dataset of 12 genes and 6 species. Bootstrap values (MP and ML) and posterior probability values (BI and BEST) are shown at the nodes

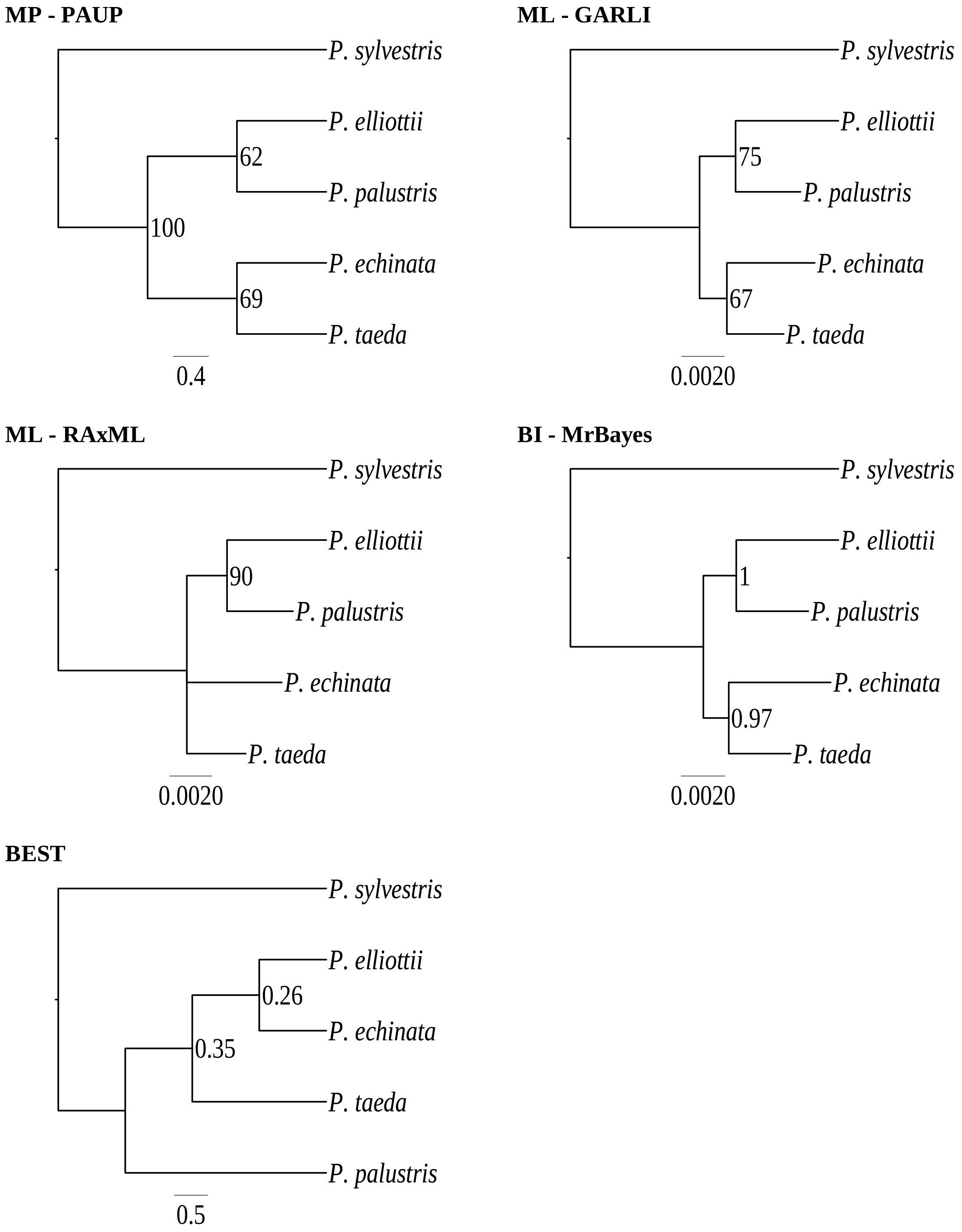


Fig. 6 Cladograms (MP - PAUP and BEST) and phylograms (ML - GARLI, ML - RAxML, and BI - MrBayes) for the dataset of 9 genes and 5 species. Bootstrap values (MP and ML) and posterior probability values (BI and BEST) are shown at the nodes

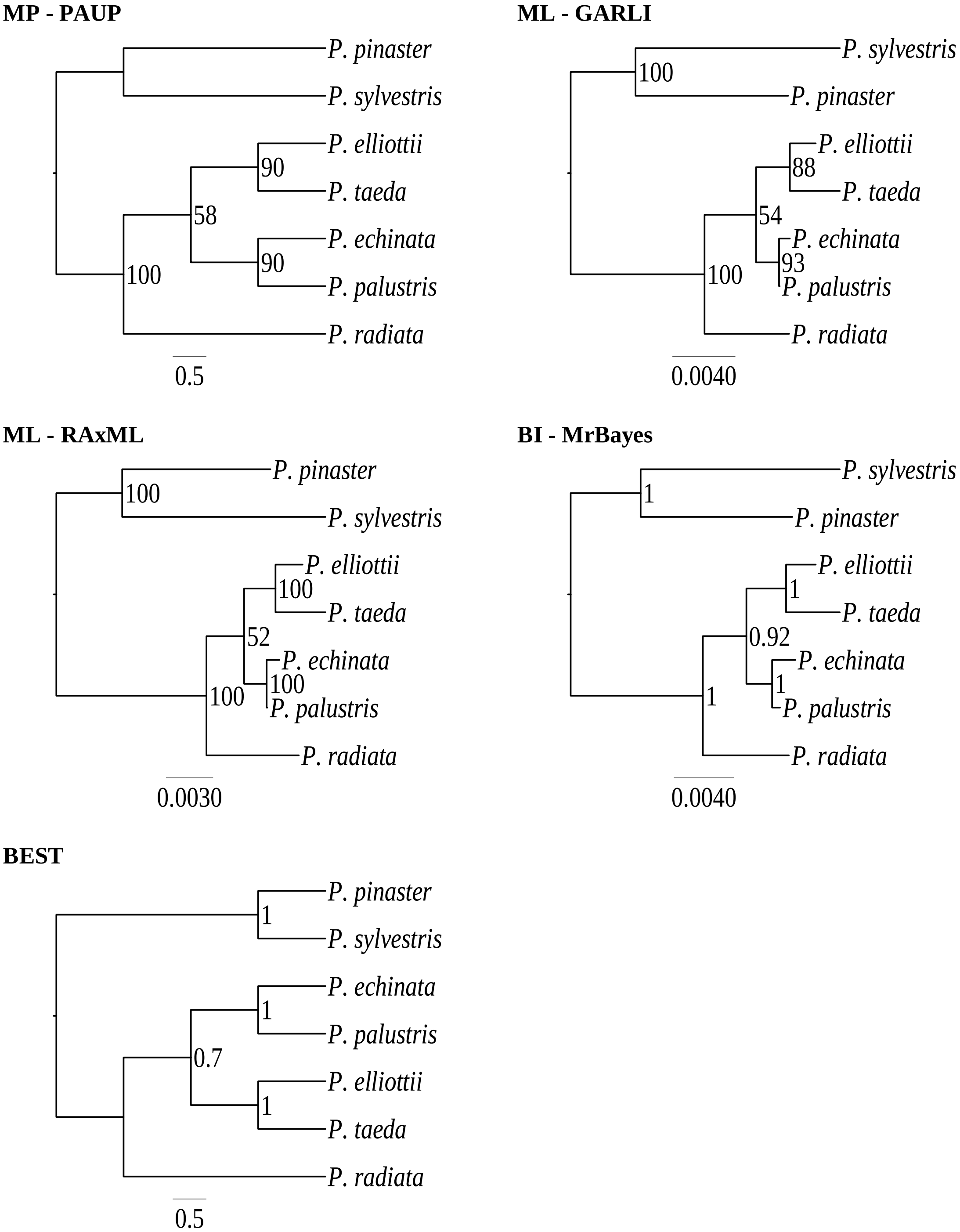


Fig. 7 Cladograms (MP - PAUP and BEST) and phylograms (ML - GARLI, ML - RAxML, and BI - MrBayes) for the dataset of 5 genes and 7 species. Bootstrap values (MP and ML) and posterior probability values (BI and BEST) are shown at the nodes

Phylogenetic analysis. MP analysis was performed using PAUP* 4.0 b10 (Swofford 2003). Most parsimonious trees were found through the heuristic search with 200 random addition sequences followed by tree bisection-reconnection (TBR) branch swapping. Alignment gaps were treated as missing data and multiple states as uncertainty. Bootstrap analysis involved 500 replicates, and majority consensus rule was applied to generate a consensus phylogenetic tree.

ML analysis was done using RAxML ver. 7.2.6 (Stamatakis et al. 2008; Stamatakis et al. 2005; Stamatakis 2006) on the CIPRES cluster (http://www.phylo.org/), and GARLI ver. 1.0 (Zwickl 2006), both with 100 bootstrap replicates. RAxML, one of the fastest implementations of ML inference (Stamatakis et al. 2007), was run under the assumptions of the general time-reversible model (Tavaré 1986), GTR+Γ, where Γ is the shape parameter of the gamma distribution of the substitution rates over sites (Yang 1993). GARLI employs genetic algorithms for the parameter optimization and was run under HKY+Γ without partitioning of the data. The bootstrap analyses were summarized using SumTrees ver. 2.0.2 (Sukumaran and Holder 2010).

BI was performed using MrBayes ver. 3.1.2 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003) on the Brazos Cluster at Texas A&M University (http://brazos.tamu.edu/). The analyses ran for 100,000,000 generations. Sampling frequency was set to 5,000, and number of runs and number of chains were both set to 4. The outputs were inspected for stationarity using three criteria: the plot of the log likelihood values, the standard deviation of split frequencies, and the potential scale reduction factor (PSFR) values. The burn-in was determined individually for each analysis.

BEST ver. 2.3 (Liu 2008) was used for Bayesian estimation of species trees (Table 11). The analyses ran for approximately 74, 63, 80, and 55 million generations for combinations 8-7, 12-6, 9-5, and 5-7, respectively, with 2 runs, 2 chains each, and with sampling frequency of 5,000.

Data partitioning and model selection. For the purpose of these tests, data partitions were extracted out of each of the four datasets (i.e. containing 8, 12, 9, and 5 genes) using SeaView ver. 4.0, and saved in separate files. For the ML analysis, jModelTest ver. 0.1.1 (Posada 2008; Guindon and Gascuel 2003) was run for non-partitioned data. Akaike information criterion (AIC) was used to infer the model for GARLI, which does not consider data partitions. RAxML allows for data partitioning and uses much faster algorithms, therefore was run for four configurations of partitions (Table 12): no partitioning, two partitions (codon positions 1 and 2 jointly, and codon positions 3 and noncoding sites jointly), three partitions (codon positions 1 and 2 jointly, codon positions 3, and noncoding sites), and four partitions (codon positions 1, 2 and 3, and noncoding sites). Based on the likelihood scores, AIC was calculated and the best configuration selected.

For the purpose of BI, seven configurations were tested: no partitions, codon positions 1 and 2 jointly, codon positions 3 and noncoding sites jointly, and four separate partitions for codon positions 1, 2 and 3, and noncoding sites. jModelTest was run for

Table 11 jModelTest results for BEST analysis

Gene	Combination of number of genes (first number) and species (second number)							
	8-7		12-6		9-5		5-7	
	n	**model**	***n***	**model**	***n***	**model**	***n***	**model**
4cl	519	JC	1914	K80+G	1914	K80	-	-
agp-6-1	-	-	596	F81	-	-	-	-
cad	-	-	483	JC	-	-	-	-
ccoaomt-1 or *ccoaomt*	593	JC	593	JC	593	JC	593	JC
cesA3-1	668	F81	668	F81	668	F81	-	-
dhn-2	590	K80	586	K80	586	K80	590	K80
erd3	885	F81	883	F81	883	F81	883	F81
lp5-like	-	-	513	F81	513	F81	-	-
mt-like	-	-	456	JC	-	-	-	-
pal-1 or *pal*	497	JC+G	497	JC	497	JC	497	JC+G
rd21A-like	983	TrN+G	983	HKY	983	HKY	983	HKY+G
sams-2 or *sam-2*	544	JC	544	JC	544	JC	-	-

Note: "-" – gene was not used in this combination; *n* – number of sites.

Table 12 Partitioning configurations for each dataset according to AIC (based on RAxML results) and BIC (based on jModelTest)

Combination*	**Criterion**	**Partitioning**				
		No partitions	**1-2, 3-*N***	**1-2, 3, *N***	**1, 2, 3-*N***	**1, 2, 3, *N***
8-7	AIC	17,262.5	17,123.6	17,072.8	17060.8	*17,013.3*
	BIC	17,401.3	17,336.6	17,359.3	*17,319.0*	17,341.7
12-6	AIC	27,640.9	27,401.7	27,257.3	27244.0	*27,100.0*
	BIC	27,796.6	27,596.4	27,490.6	27,483.7	*27,377.9*
9-5	AIC	22,528.2	22,321.7	22,195.8	22226.2	*22,100.3*
	BIC	22,669.9	22,499.5	22,392.2	22,448.1	*22,340.8*
5-7	AIC	11,990.6	11,881.2	11,849.7	11842.1	*11,810.5*
	BIC	12,109.2	*12,064.8*	12,103.0	12,065.6	12,103.8
3-6 a	AIC	7,913.6	7,833.6	7,821.2	7818.8	*7,806.3*
	BIC	8,004.7	*7,970.1*	8,014.4	7,997.0	8,041.3
3-6 b	AIC	9,921.6	9,874.8	9,774.6	9799.8	*9,699.6*
	BIC	10,029.6	10,007.3	9,953.2	9,997.9	*9,943.8*

Note: * Combination of number of genes (first number) and species (second number);
N – noncoding sites; 1-2 – codon positions 1 and 2 jointly; 3-*N* – codon positions 3 and noncoding sites jointly; 1, 2, 3, *N* – four separate partitions for codon positions 1, 2 and 3, and noncoding sites, respectively. In italics are the minimum values for each combination for both AIC and BIC.

each of these sets and the Bayesian Information Criterion (BIC) scores were recorded (Table 13). BIC scores were calculated for all five possible combinations of these seven partitions (e.g. codon partitions 1 and 2 jointly, and codon positions 3, and noncoding sites), such that each combination ensured complete coverage of the dataset. Following this algorithm, models were selected for ML and BI.

Partitioned Bremer Support (PBS). TreeRot ver. 3 (Sorenson and Franzosa 2007) and PAUP* were used to infer support for the nodes provided by each gene in the dataset containing 12 genes (Table 14).

Phylogenetic trees visualization. The trees figures were prepared using FigTree ver. 1.2.2 (http://tree.bio.ed.ac.uk/software/figtree/).

4.4. Results

The results of maximum parsimony (MP), maximum likelihood analysis (ML) and Bayesian inference (BI) did not allow for unambiguous placement of *P. radiata* as a sister species to the clade comprising the four Southern pines (Fig. 4), showing fairly low bootstrap support. The Approximately Unbiased (AU) test (Shimodaira 2002) implemented in CONSEL (Shimodaira and Hasegawa 2001) was used to assess topologies with seven alternative positions of *P. radiata* with regards to the four Southern pines. The test's result failed to reject only one of the alternatives, that placed *P. radiata* as a sister species to *P. echinata.* This topology was absent from the results produced

Table 13 Model selection (jModelTest results) for partitions in each dataset

Combination*	**Partition**	***n***	**AIC**				**BIC**			
			-*ln*(*L*)	**AIC**	***k***	**model**	**-*ln*(*L*)**	**BIC**	***k***	**model**
8-7	no partitions	5,279	8,632.1	17,296.3	16	HKY+G	8,632.1	17,401.3	16	HKY+G
	codon pos. 1 & 2	2,208	3,341.5	6,713.0	15	HKY	3,336.5	6,796.2	16	HKY+G
	codon pos. 3 & *N*	3,071	5,205.9	10,443.9	16	HKY+G	*5,205.9*	*10,540.4*	*16*	*HKY+G*
	codon pos. 1	1,105	1,646.1	3,322.1	15	HKY	*1,646.1*	*3,397.3*	*15*	*HKY*
	codon pos. 2	1,103	1,644.1	3,316.2	14	F81	*1,638.1*	*3,381.4*	*15*	*HKY*
	codon pos. 3	1,103	1,881.9	3,789.7	13	K80+G	1,883.9	3,851.9	12	K80
	N	1,968	3,294.9	6,621.8	16	HKY+G	3,294.9	6,711.2	16	HKY+G
12-6	no partitions	8,716	13,834.8	27,697.6	14	HKY+G	13,834.8	27,796.6	14	HKY+G
	codon pos. 1 & 3	3,855	5,645.1	11,314.1	12	F81	5,639.3	11,386.0	13	HKY
	codon pos. 3 & *N*	4,861	8,045.8	16,119.6	14	HKY+G	8,045.8	16,210.4	14	HKY+G
	codon pos. 1	1,929	2,730.0	5,484.0	12	F81	*2,730.0*	*5,550.7*	*12*	*F81*
	codon pos. 2	1,926	2,815.9	5,655.8	12	F81	*2,815.9*	*5,722.6*	*12*	*F81*
	codon pos. 3	1,926	3,209.9	6,447.8	14	HKY+G	*3,209.9*	*6,525.7*	*14*	*HKY+G*
	N	2,935	4,733.6	9,495.2	14	HKY+G	*4,733.6*	*9,578.9*	*14*	*HKY+G*
9-5	no partitions	7,181	11,281.7	22,587.4	12	HKY+G	11,281.7	22,669.9	12	HKY+G
	codon pos. 1 & 2	3,156	4,576.1	9,174.2	11	HKY	4,576.1	9,240.8	11	HKY
	codon pos. 3 & *N*	4,025	6,579.6	13,183.1	12	HKY+G	6,579.6	13,258.7	12	HKY+G
	codon pos. 1	1,579	2,247.7	4,515.4	10	F81	*2,247.7*	*4,569.0*	*10*	*F81*
	codon pos. 2	1,577	2,273.3	4,566.7	10	F81	*2,273.3*	*4,620.3*	*10*	*F81*
	codon pos. 3	1,577	2,571.0	5,165.9	12	HKY+G	*2,571.0*	*5,230.3*	*12*	*HKY+G*
	N	2,448	3,913.7	7,851.5	12	HKY+G	*3,913.7*	*7,921.1*	*12*	*HKY+G*

Table 13 Continued

Combination*	**Partition**	***n***	**AIC**				**BIC**			
			-*ln*(*L*)	**AIC**	***k***	**model**	**-*ln*(*L*)**	**BIC**	***k***	**model**
5-7	no partitions	3,546	5,989.2	12,010.4	16	HKY+G	5,989.2	12,109.2	16	HKY+G
	codon pos. 1 & 2	1,528	2,371.1	4,772.3	15	HKY	*2,366.2*	*4,849.7*	*16*	*HKY+G*
	codon pos. 3 & *N*	2,018	3,546.7	7,125.4	16	HKY+G	*3,546.7*	*7,215.1*	*16*	*HKY+G*
	codon pos. 1	765	1,175.8	2,379.6	14	F81	1,175.8	2,444.5	14	F81
	codon pos. 2	763	1,158.4	2,344.8	14	F81	1,153.2	2,406.0	15	HKY
	codon pos. 3	764	1,337.8	2,705.5	15	HKY	1,337.8	2,775.1	15	HKY
	N	1,254	2,182.2	4,396.4	16	HKY+G	2,178.5	4,478.2	17	TrN+G
3-6 a	no partitions	2,452	3,951.6	7,929.3	13	HKY	3,951.6	8,004.7	13	HKY
	codon pos. 1 & 2	1,110	1,668.6	3,361.2	12	F81	*1,663.5*	*3,418.1*	*13*	*HKY*
	codon pos. 3 & *N*	1,342	2,229.2	4,484.3	13	HKY	*2,229.2*	*4,551.9*	*13*	*HKY*
	codon pos. 1	556	828.8	1,681.5	12	F81	828.8	1,733.4	12	F81
	codon pos. 2	554	817.9	1,659.9	12	F81	817.9	1,711.7	12	F81
	codon pos. 3	555	928.8	1,881.6	12	F81	931.3	1,925.8	10	K80
	N	787	1,292.5	2,610.9	13	F81+G	1,295.2	2,670.5	12	F81
3-6 b	no partitions	3,175	4,958.4	9,944.8	14	HKY+G	4,958.4	10,029.6	14	HKY+G
	codon pos. 1 & 2	1,321	1,896.5	3,811.0	9	JC	1,896.5	3,857.7	9	JC
	codon pos. 3 & *N*	1,854	3,022.2	6,072.3	14	HKY+G	3,022.2	6,149.7	14	HKY+G
	codon pos. 1	661	914.4	1,852.8	12	F81	*914.4*	*1,906.7*	*12*	*F81*
	codon pos. 2	660	931.8	1,887.6	12	F81	*931.8*	*1,941.5*	*12*	*F81*
	codon pos. 3	659	993.9	2,013.7	13	HKY	*993.9*	*2,072.1*	*13*	*HKY*
	N	1,195	1,962.1	3,952.3	14	HKY+G	*1,962.1*	*4,023.5*	*14*	*HKY+G*

Note: * – Combination of number of genes (first number) and species (second number); ***N*** – noncoding sites; ***n*** – sample size (number of characters); ***k*** – number of parameters. In italics are the partitions included in BI analysis.

Table 14 Partitioned Bremer Support (PBS) values for the genes in each dataset

Combination **	Node	*ccoaomt-1* or *ccoaomt*	*dhn-2*	*erd3*	*cesA3-1*	*pal-1* or *pal*	*rd21A-like*	*sams-2* or *sam-2*	*agp-6-1*	*lp5-like*	*mt-like*	*cad*	*4cl*	Total
12-6	whole tree	24	29	33	8	14	35	12	6	27	8	6	48	250
	1 (*P. echinata*, *P. taeda*)	3	-1	-2	0	1	0	0	0	-2	0	0	2	
	2 (*P. echinata*, *P. taeda*, *P. elliottii*)	3	-1	-2	0	1	0	0	0	-2	0	0	2	
	3 (*P. echinata*, *P. taeda*, *P. elliottii*, *P. palustris*)	3	-0.50	-2	0	1	-0.50	0	0	-1	0	0	1	
9-5	whole tree	17	26	30	6	12	31	11	-	20	-	-	42	195
	1 (*P. elliottii*, *P. palustris)*	1	0	0	1	0	-1	0	-	0	-	-	0	
	2 (*P. echinata*, *P. taeda*)	0	-1	0	0	0	0	0	-	0	-	-	2	
8-7	whole tree	36	33	30	8	20	43	12	-	-	-	-	14*	196
	1 (*P. echinata*, *P. palustris*, *P. elliottii*, *P. taeda*, *P. radiata*)	0	11	7	0	0	0	0	-	-	-	-	-1*	
	2 (*P. elliottii*, *P. taeda*)	-1	2	4	-1	0	0	0	-	-	-	-	-3*	
	3 (*P. echinata*, *P. palustris*)	-0.33	2	1.33	-0.33	0	0	0	-	-	-	-	-1.67*	
	4 (*P. echinata*, *P. palustris*, *P. elliottii*, *P. taeda*)	1.50	1	-1	0	0.50	-0.50	0	-	-	-	-	-0.50*	

Table 14 Continued

Combination **	**Node**	***ccoaomt-1*** **or** ***ccoaomt***	***dhn-2***	***erd3***	***cesA3-1***	***pal-1*** **or** ***pal***	***rd21A-like***	***sams-2*** **or** ***sam-2***	***agp-6-1***	***lp5-like***	***mt-like***	***cad***	***4cl***	**Total**
5-7	whole tree	36	33	31	-	20	44	-	-	-	-	-	-	164
	1 (*P. echinata*, *P. palustris*, *P. elliottii*, *P. taeda*, *P. radiata*)	0	11	7	-	0	0	-	-	-	-	-	-	
	2 (*P. elliottii*, *P. taeda*)	1	-0.50	2	-	0.50	1	-	-	-	-	-	-	
	3 (*P. echinata*, *P. palustris*)	0	2	0	-	0	0	-	-	-	-	-	-	
	4 (*P. echinata*, *P. palustris*, *P. elliottii*, *P. taeda*)	3	0	-2	-	1	-1	-	-	-	-	-	-	

Note: * – only amplicon 4 was used (*4cl-4*); ** – Combination of number of genes (first number) and species (second number).

by MP, ML or BI. Therefore, the hypothesis that *P. radiata* is not a sister lineage (and thus, appropriate outgroup) for the four Southern pines could not be rejected with this dataset.

None of the four dataset combinations unambiguously resolved the phylogenetic relationships between the studied species. For the combinations 8-7 (Fig. 4) and 5-7 (Fig. 7), most analyses grouped *P. echinata* with *P. palustris,* and *P. elliottii* with *P. taeda*. Although the bootstrap support varied between the methods, the topology of these two clades was consistent except for MP in combination 8-7. Interestingly, the AU test rejected all alternative hypotheses to the ML tree except for *P. radiata* being a sister species to *P. echinata.*

On the contrary, the combinations 12-6 (Fig. 5) and 9-5 (Fig. 6) indicate the closest relationship between *P. echinata* and *P. taeda*. Although the bootstrap support for this clade was as low as 58 in MP and ML in combination 12-6, this topology was supported across all methods. The clade *P. elliottii – P. palustris* was supported by all methods only in the combination 9-5 and was not resolved when 12 genes were analyzed. *P. radiata*, analyzed in combination 12-6, was placed as an ancestral taxon to the four Southern pines by ML analysis conducted by RAxML and BI using MrBayes, while GARLI (ML) and PAUP* (MP) failed to resolve these relationships.

Trees produced by MrBayes (BI) topologically matched the trees produced by RAxML (ML) in all the cases except for combination 9-5, where RAxML failed to resolve relationship between *P. taeda* and *P. echinata,* and MrBayes showed high posterior probability for this clade (0.97). Although bootstrap values used in ML and

MP methods cannot be directly compared with the posterior probability values produced by BI (Cummings et al. 2003), both are in consensus for the clade support in general. Specifically, the posterior probability values tend to be excessively higher than the expectations in the range 0.85-1 (Cummings et al. 2003), and, therefore, are more prone to lead to erroneous conclusions (Erixon et al. 2003). This seems to be the case also in this study.

The BEST approach produced trees that were in agreement with MrBayes and RAxML except for the combination 9-5, where posterior probabilities were very low (Fig. 6).

The model selected for GARLI was HKY+Γ (no partitioning of the data is currently available in this software; Table13). Only in the case of combination 3-6a, this model assumed no Γ parameter. In the RAxML analyses GTR+Γ model was used. The tested partitioning configurations included 4, 3, 2 and no partitions. For all datasets, however, the optimal AIC value under GTR+Γ assumptions was achieved for 4 partitions (Table 12).

The conflicting results regarding relationships between the four Southern pines were consistent because the discrepancies depended on the dataset and not on the method used. The partitioned Bremer support (PBS) test (Table 14) identified those genes that support a particular clade (positive values), or an alternative one (negative values). The distinction is specifically noticeable for the dataset 12-6, where genes *4cl*, *caffeoyl CoA O-methyltransferase 1* (*ccoaomt-1*) and *phenylalanine ammonia-lyase 1* (*pal-1*) supported all three nodes, while genes *dehydrin 2* (*dhn-2), erd3* and *lp5-like* supported an alternative (Table 14 and Fig. 5). The analysis of these results led to selection of two triplets of genes: *dhn-2, erd3* and *rd21A-like* (support for the node *P. taeda* and *P. elliottii*), and *4cl, ccoaomt-1* and *cellulose synthase A3* (*cesA3-1*; support for the node *P. taeda* and *P. echinata*), combinations 3-6a and 3-6b, respectively. This approach allowed for significant improvement in bootstrap and posterior probability scores in all four methods used (except BEST, that was not used here; Figs. 8 and 9). As expected, the topologies between these two combinations varied by the way the four Southern pines were coupled, and also by the position of *P. radiata*.

4.5. Discussion

As the nucleotide sequence data are becoming more abundant, data processing is becoming more challenging. Although the dataset in this study included only 12 loci and 7 species, the analysis of the total evidence might be misleading. Low bootstrap support or posterior probability could be interpreted as insufficient data coverage. In this case joint analysis of the whole dataset introduced additional uncertainty as far as the final phylogeny is concerned.

The outcome of the PBS test shows that the genes that provided support for the cluster of *P. taeda* and *P. elliottii* (*dhn-2, erd3* and *rd21A-like*) are involved in drought and dehydration recognition and response (see Ersoz 2006 for the summary on selected genes function; Table 15). These genes strongly supported two clades: *P. elliottii – P. taeda* and *P. echinata – P. palustris* in MP, ML (GARLI), and BI analyses. In addition,

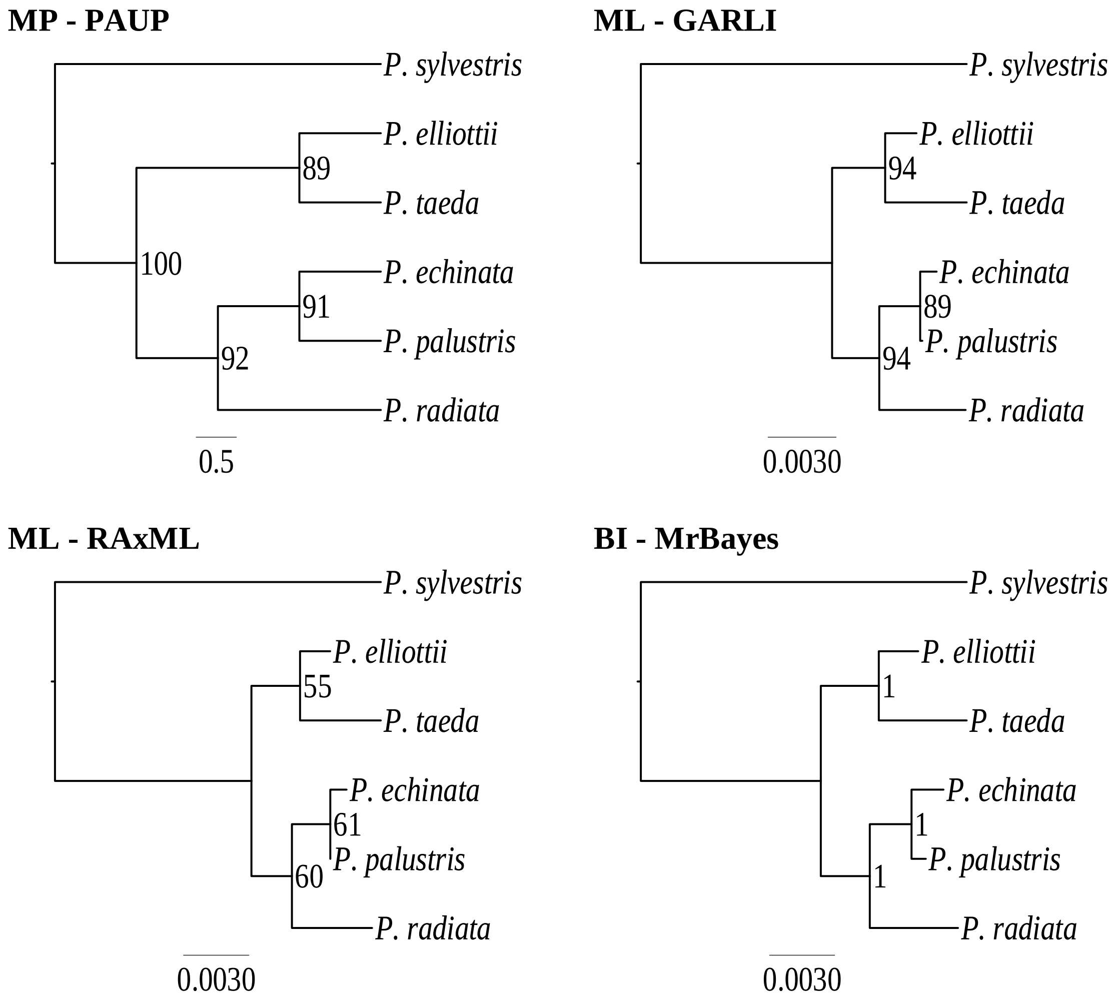
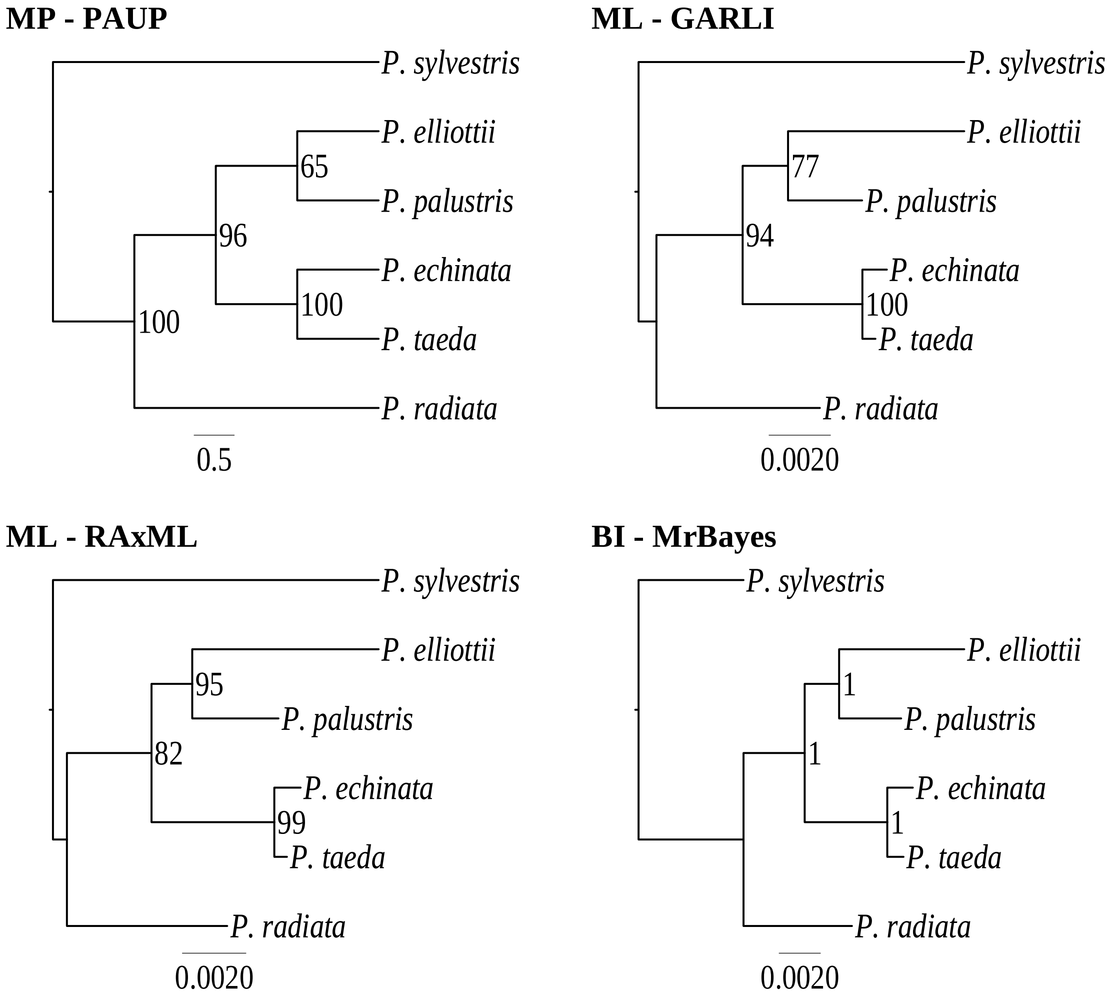


Fig. 8 Cladograms (MP - PAUP and BEST) and phylograms (ML - GARLI, ML - RAxML, and BI - MrBayes) for the dataset of 3 genes (*dhn-2*, *erd3* and *rd21A-like*) and 6 species. Bootstrap values (MP and ML) and posterior probability values (BI) are shown at the nodes



MP - PAUP

ML - GARLI

ML - RAxML

BI - MrBayes

Fig. 9 Cladograms (MP - PAUP and BEST) and phylograms (ML - GARLI, ML - RAxML, and BI - MrBayes) for the dataset of 3 genes (*4cl*, *ccoaomt-1* and *cesA3-1*) and 6 species. Bootstrap values (MP and ML) and posterior probability values (BI) are shown at the nodes

Table 15 Function of selected genes (Ersoz 2006)

Gene	Function
4cl	enzyme; mono-lignol biosynthesis
ccoaomt-1 or *ccoaomt*	enzyme; mono-lignol biosynthesis
cesA3-1	enzyme; cell wall biosynthesis
dhn-2	enzyme; drought response
erd3	transcription factor; dehydration recognition and response
rd21A-like	enzyme; drought response

the position of *P. radiata* was very well-supported as a sister taxon to *P. echinata – P. palustris* clade, proving its place within subsection *Australes*. Only RAxML-based ML analysis showed low bootstrap support, although the topology did not vary from the other three approaches.

The dataset that included the other gene triplet (*4cl, ccoaomt-1* and *cesA3-1*), associated with wood quality, very strongly supported another clade (*P. echinata – P. taeda*). The clade *P. elliottii – P. palustris* was also well defined, but with a considerably lower bootstrap support in MP and ML (GARLI) analyses. In this case, *P. radiata* was also very well supported, however, positioned as an ancestral species in relation to all four Southern pines. Interestingly, these three genes play a role in cell-wall biosynthesis (Ersoz 2006). This dichotomy clearly shows that various evolutionary adaptations may affect differently phylogenetic relationships within the subsection.

An additional important factor is hybridization occurring between the Southern pines, as described in both natural and artificial settings (Mergen 1958; Smouse and Saylor 1973; Mergen et al. 1965; Price 1989). Gene flow that occurred between the species at different times in their history can also complicate their relationships via introducing different selective alleles at different times. Therefore, these similarities may be not only characteristic of a close relationship between two species, but also reflect parallel adaptation and opportunistic gene exchange.

A recent study on multiple species from subgenus *Strobus* (Syring et al. 2007) using an intron from IFG8612 revealed high levels of polymorphisms shared among the species. The authors concluded that the likely reason for this is incomplete lineage sorting. It is very likely that in the case of *Australes* we deal with a similar phenomenon. In this study only one sequence per gene per species was used. Data from multiple alleles per species could help investigate this phenomenon in *Australes*.

4.6. Conclusions

This study confirmed very tight relationships within the subsection *Australes*. The dataset consisting of 12 loci was sufficient to show that in the case of *Australes* not only the amount of genetic data but also their partitioning and configuration may severely impact the conclusions. Using only three genes proved to be sufficient to achieve very high bootstrap support (MP and ML) and posterior probability values (BI), but still can be misleading in identifying true phylogenetic relationships. This study clearly demonstrated that the phylogenetic trees that are based on a limited number of genes could be very different, and very likely represent “gene” trees, rather than “species” trees. Therefore, trees obtained in the study based on the small number of genes should be considered very cautiously and critically. Moreover, proper partitioning may help to understand the ancestral relationships between the species more than the amount of data used.

5. CONCLUSIONS

The four major Southern pines, *Pinus echinata* Mill. (shortleaf pine), *P. elliottii* Engelm. (slash pine), *P. palustris* Mill. (longleaf pine) and *P. taeda* L. (loblolly pine), are an important component of the Southeastern ecosystem and the landscape of the thirteen states. They provide multiple benefits to human society, playing an important role in landscaping, erosion control and watershed management, as well as being an important source of timber and pulp for industry. They are the keystone species that provide habitat and protection for numerous species of microorganisms, fungi, plants and animals.

These four pine species share common evolutionary history, ancestry, and vastly sympatric or overlapping area. Affected by the harsh conditions of the recent glacial period that ended about 15 thousand years ago, and being well adapted to climatic conditions of their current range that stretches from subtropical to warm temperate, they have likely accumulated significant variation in the adaptive trait loci.

Despite the multiple benefits that they provide, they have not been as intensely studied as loblolly pine, and knowledge about the organization of their genomes is limited. In addition, their recent dynamic evolutionary history creates additional opportunities to study evolutionary processes in progress. Although the complete sequencing of the loblolly pine genome is on the way, due to its large size, complete assembly and annotation will require much time and effort.

We used already available data for completely sequenced organisms, data newly collected in this study, and publically available data for selected wood-quality and water-stress related genes in loblolly pine. We proposed novel comparative genomic approaches to further our understanding of genome-wide characteristics in incompletely sequenced non-model species, such as the four Southern pines. We investigated the effect of selection pressures in these selected genes, and analyzed phylogenetic relationships between these pine species. We answered three specific questions: (1) What can be predicted about the genome-wide characteristics of loblolly pine, based on the genomic data of completely sequenced species? (2) What effects has selection had upon the set of studied genes in the four Southern pines? (3) What are the phylogenetic relationships between the four Southern pines, and with respect to other selected pine species?

To predict the genome-wide characteristics of loblolly pine, we developed a series of statistical regression models. Using data publically available in NCBI GenBank, we inferred relationships between the parameters that can be relatively easily estimated from available data (such as mean exon length and exon/gene ratio), and parameters that are difficult to assess (e.g. number of protein coding genes, number of all genes and exons). We confirmed the general trend of increasing number of genes, gene products, and exons in the genome, along with higher exon/gene ratio and alternative splicing (AS) ratio as species become more evolutionarily advanced. Although our results indicate that different taxonomic kingdoms may have followed various evolutionary paths and may require different calibration of the model parameters, the number of completely sequenced evolutionarily distant plant species should be extended to allow further conclusions.

To elucidate the effects of selection on drought tolerance, drought-stress response and wood-quality related genes, we used previously published data (Ersoz 2006; Gonzalez-Martinez et al. 2006a; Brown et al. 2004) and expanded the dataset by newly acquired sequences in this study for the four Southern pines. Despite the relatively small dataset of a little over 30 *P. taeda* individuals in most loci, and no more than 4 individuals from the other species, we found signatures of selection in some of these genes studied. In addition, we demonstrated that different parts of a gene could be under different forms of selection and could mislead neutrality tests performed at the entire gene level. To better discriminate between the effects caused by selection and those caused by recent demographic events, a more distant pine species that faces other adaptive challenges should be included for interspecific comparisons; more complex models that include both selection and demographic events should be tested using coalescence approaches.

To resolve the relationships between the four Southern pines, we studied three additional pine species, i.e. *P. radiata*, *P. pinaster* and *P. sylvestris*, for which the nucleotide sequence data orthologous to the sequences newly generated in this study are available in the NCBI GenBank. We confirmed very tight phylogenetic relationships within the subsection *Australes*. The study of 12 genes demonstrated that both the number of genes and their partitioning can greatly affect the conclusions for this group of Southern pines. Within this dataset we identified two triplets of genes that supported alternative topologies for the four Southern pines, each with high bootstrap support and posterior probability values, depending on the method applied. We demonstrated that drawing conclusions about species trees should be done with caution because the phylogenetic trees based on a limited number of genes could be very different and likely represent “gene” trees rather than “species” trees. In addition, appropriately applied data partitioning may be a useful tool in more detailed understanding of the ancestral relationships between the species.

As the most studied coniferous species, loblolly pine has become a model species for conifers (Krutovsky et al. 2004). Before the complete genomic data are available, the regression models developed in this study will help bridge the gap in understanding the structure of the pine genomes and genomes of other incompletely sequenced non-model species, until more data are collected. New genomic data for pines and other evolutionarily distant plants will help fine-tune the proposed models. Moreover, we demonstrated that newly obtained nucleotide sequence data can be combined with already publically available data and used to expand the sample size needed for study. To better elucidate the molecular evolution in the studied genes, interspecific tests should be used, as they are better suited for dissecting the demographic effects from selection. For this purpose incorporating data from more distant pine species that face other adaptive challenges should be considered. Finally, we advanced knowledge about the relationships between the four Southern pines, which will certainly continue to grow as more data are collected.

REFERENCES

Adami C, Ofria C, Collier TC (2000) Evolution of biological complexity. Proceedings of the National Academy of Sciences of the United States of America 97 (9):4463-4468

Adams DC, Jackson JF (1997) A phylogenetic analysis of the southern pines (*Pinus* subsect *Australes* Loudon): biogeographical and ecological implications. Proceedings of the Biological Society of Washington 110 (4):681-692

Ahuja MR (2005) Polyploidy in gymnosperms: revisited. Silvae Genetica 54 (2):59-69

Al-Rabab'ah MA, Williams CG (2002) Population dynamics of *Pinus taeda* L. based on nuclear microsatellites. Forest Ecology and Management 163 (1-3):263-271

Al-Rabab'ah MA, Williams CG (2004) An ancient bottleneck in the Lost Pines of central Texas. Molecular Ecology 13 (5):1075-1084

Axelrod D (1986) Cenozoic history of some western American pines. Annals of the Missouri Botanical Garden 73 (3):565-641

Babenko VN, Rogozin IB, Mekhedov SL, Koonin EV (2004) Prevalence of intron gain over intron loss in the evolution of paralogous gene families. Nucleic Acids Research 32 (12):3724-3733

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2009) GenBank. Nucleic Acids Research 37:D26-D31

Black DL (2000) Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. Cell 103 (3):367-370

Black DL (2003) Mechanisms of alternative pre-messenger RNA splicing. Annual Review of Biochemistry 72:291-336

Brown GR, Gill GP, Kuntz RJ, Langley CH, Neale DB (2004) Nucleotide diversity and linkage disequilibrium in loblolly pine. Proceedings of the National Academy of Sciences of the United States of America 101 (42):15255-15260

Campbell MA, Haas BJ, Hamilton JP, Mount SM, Buell CR (2006) Comprehensive analysis of alternative splicing in rice and comparative analyses with *Arabidopsis*. BMC Genomics 7:327

Carey AB, Johnson ML (1995) Small mammals in managed, naturally young, and old-growth forests. Ecological Applications 5 (2):336-352

Carvalho AB, Clark AG (1999) Intron size and natural selection. Nature 401 (6751):344

Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA (2002) Selection for short introns in highly expressed genes. Nature Genetics 31 (4):415-418

Chi KR (2008) The year of sequencing. Nature Methods 5 (1):11-14

Collins C, Conner R, Saenz D (2002) Influence of hardwood midstory and pine species on pine bole arthropods. Forest Ecology and Management 164 (1-3):211-220

Collins RA (1988) Evidence of natural selection to maintain a functional domain outside of the 'core' in a large subclass of Group I introns. Nucleic Acids Research 16 (6):2705-2715

Cui LY, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A, Albert VA, Ma H, dePamphilis CW (2006) Widespread genome duplications throughout the history of flowering plants. Genome Research 16 (6):738-749

Cummings MP, Handley SA, Myers DS, Reed DL, Rokas A, Winka K (2003) Comparing bootstrap and posterior probability values in the four-taxon case. Systematic Biology 52 (4):477-487

Deutsch M, Long M (1999) Intron-exon structures of eukaryotic model organisms. Nucleic Acids Research 27 (15):3219-3228

Dickson JG, Segelquist CA (1979) Breeding bird populations in pine and pine-hardwood forests in Texas. Journal of Wildlife Management 43 (2):549-555

Eckert A, van Heervaarden J, Wegrzyn J, Nelson C, Ross-Ibarra J, Gonzalez-Martinez S, Neale D (2010) Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda* L., Pinaceae). Genetics (in press)

Eckert AJ, Hall BD (2006) Phylogeny, historical biogeography, and patterns of diversification for *Pinus* (Pinaceae): phylogenetic tests of fossil-based hypotheses. Molecular Phylogenetics and Evolution 40 (1):166-182

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32 (5):1792-1797

Erixon P, Svennblad B, Britton T, Oxelman B (2003) Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. Systematic Biology 52 (5):665-673

Ersoz E (2006) Candidate gene-association mapping for dissecting fungal disease resistance in loblolly pine. Ph.D. dissertation, University of California, Davis, CA

Fahey T, Woodbury P, Battles J, Goodale C, Hamburg S, Ollinger S, Woodall C (2010) Forest carbon storage: ecology, management, and policy. Frontiers in Ecology and the Environment 8 (5):245-252

Frederick WJ, Lien SJ, Courchene CE, DeMartini NA, Ragauskas AJ, Iisa K (2008) Production of ethanol from carbohydrates from loblolly pine: a technical and economic assessment. Bioresource Technology 99 (11):5051-5057

Galtier N, Gouy M, Gautier C (1996) SEAVIEW and PHYLO_WIN: Two graphic tools for sequence alignment and molecular phylogeny. Computer Applications in the Biosciences 12 (6):543-548

Gernandt DS, Liston A, Pinero D (2001) Variation in the nrDNA ITS of *Pinus* subsection *Cembroides*: implications for molecular systematic studies of pine species complexes. Molecular Phylogenetics and Evolution 21 (3):449-467

Gernandt DS, Lopez GG, Garcia SO, Liston A (2005) Phylogeny and classification of *Pinus*. Taxon 54 (1):29-42

Gernandt DS, Magallon S, Lopez GG, Flores OZ, Willyard A, Liston A (2008) Use of simultaneous analyses to guide fossil-based calibrations of Pinaceae phylogeny. International Journal of Plant Sciences 169 (8):1086-1099

Gonzalez-Martinez SC, Ersoz E, Brown GR, Wheeler NC, Neale DB (2006a) DNA sequence variation and selection of tag single-nucleotide polymorphisms at candidate genes for drought-stress response in *Pinus taeda* L. Genetics 172 (3):1915-1926

Gonzalez-Martinez SC, Krutovsky KV, Neale DB (2006b) Forest-tree population genomics and adaptive evolution. New Phytologist 170 (2):227-238

Gonzalez-Martinez SC, Wheeler NC, Ersoz E, Nelson CD, Neale DB (2007) Association genetics in *Pinus taeda* L. I. Wood property traits. Genetics 175 (1):399-409

Graveley BR (2001) Alternative splicing: increasing diversity in the proteomic world. Trends in Genetics 17 (2):100-107

Grotkopp E, Rejmanek M, Sanderson MJ, Rost TL (2004) Evolution of genome size in pines (*Pinus*) and its life-history correlates: supertree analyses. Evolution 58 (8):1705-1729

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology 52 (5):696-704

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series 41:95-98

Holmes RT, Robinson SK (1988) Spatial patterns, foraging tactics, and diets of ground-foraging birds in a northern hardwoods forest. Wilson Bulletin 100 (3):377-394

Hudson RR, Kreitman M, Aguade M (1987) A test of neutral molecular evolution based on nucleotide data. Genetics 116 (1):153-159

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17 (8):754-755

Hughes AL, Friedman R (2008) Alternative splicing, gene duplication and connectivity in the genetic interaction network of the nematode worm *Caenorhabditis elegans*. Genetica 134 (2):181-186

Iida K, Seki M, Sakurai T, Satou M, Akiyama K, Toyoda T, Konagaya A, Shinozaki K (2004) Genome-wide analysis of alternative pre-mRNA splicing in *Arabidopsis thaliana* based on full-length cDNA sequences. Nucleic Acids Research 32 (17):5096-5103

Jackson ST, Webb RS, Anderson KH, Overpeck JT, Webb T, Williams JW, Hansen BCS (2000) Vegetation and environment in Eastern North America during the Last Glacial Maximum. Quaternary Science Reviews 19 (6):489-508

Johnsen K, Wear D, Oren R, Teskey R, Sanchez F, Will R, Butnor J, Markewitz D, Richter D, Rials T (2001) Meeting global policy commitments: carbon sequestration and southern pine forests. Journal of Forestry 99 (4):14-21

Jukes T, Cantor C (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian protein metabolism. Academic Press, New York, pp 21-132

Kim E, Magen A, Ast G (2007) Different levels of alternative splicing among eukaryotes. Nucleic Acids Research 35 (1):125-131

Kimura M (1968) Evolutionary rate at molecular level. Nature 217 (5129):624-626

Kondrashov FA, Koonin EV (2001) Origin of alternative splicing by tandem exon duplication. Human Molecular Genetics 10 (23):2661-2669

Kondrashov FA, Koonin EV (2003) Evolution of alternative splicing: deletions, insertions and origin of functional parts of proteins from intron sequences. Trends in Genetics 19 (3):115-119

Kopelman NM, Lancet D, Yanai I (2005) Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. Nature Genetics 37 (6):588-589

Kreitman M (2000) Methods to detect selection in populations with applications to the human. Annual Review of Genomics and Human Genetics 1:539-559

Krutovskii KV, Politov DV, Altukhov YP (1994) Genetic differentiation and phylogeny of stone pine species based on isozyme loci. Proceedings of the International Workshop on Subalpine Stone Pines and Their Environment: The Status of Our Knowledge 309:19-30

Krutovsky KV, Neale DB (2005) Nucleotide diversity and linkage disequilibrium in cold-hardiness- and wood quality-related candidate genes in Douglas fir. Genetics 171 (4):2029-2041

Krutovsky KV, Troggio M, Brown GR, Jermstad KD, Neale DB (2004) Comparative mapping in the Pinaceae. Genetics 168 (1):447-461

Li WH, Wu CI, Luo CC (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Molecular Biology and Evolution 2 (2):150-174

Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25 (11):1451-1452

Little EL, Critchfield WB (1969) Subdivisions of the genus *Pinus* (pines). USDA Forest Service, Washington, DC, Miscellaneous Publication Number 1144

Liu L (2008) BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics 24 (21):2542

Long MY, Rosenberg C, Gilbert W (1995) Intron phase correlations and the evolution of the intron/exon structure of genes. Proceedings of the National Academy of Sciences of the United States of America 92 (26):12495-12499

Lynch M, Conery JS (2003) The origins of genome complexity. Science 302 (5649):1401-1404

Malmsheimer RW, Heffernan P, Brink S, Crandall D, Deneke F, Galik C, Gee E, Helms JA, McClure N, Mortimer M, Ruddell S, Smith M, Stewart J (2008) Forest management solutions for mitigating climate change in the United States. Journal of Forestry 106 (3):115-117

Mano S, Hayashi M, Nishimura M (1999) Light regulates alternative splicing of hydroxypyruvate reductase in pumpkin. Plant Journal 17 (3):309-320

Mardis ER (2008) Next-generation DNA sequencing methods. Annual Review of Genomics and Human Genetics 9:387-402

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in *Drosophila*. Nature 351 (6328):652-654

McGuire AM, Pearson MD, Neafsey DE, Galagan JE (2008) Cross-kingdom patterns of alternative splicing and splice recognition. Genome Biology 9 (3):R50

McKeown M (1992) Alternative messenger-RNA splicing. Annual Review of Cell Biology 8:133-155

Melchiors M, Silker T, Reeb J (1985) Deer use of young pine plantations in southeastern Oklahoma. The Journal of Wildlife Management 49 (4):958-962

Mergen F (1958) Genetic variation in needle characteristics of slash pine and in some of its hybrids. Silvae Genetica 7:1-9

Mergen F, Stairs G, Snyder E (1965) Natural and controlled loblolly x shortleaf pine hybrids in Mississippi. Forest Science 11 (3):306-314

Michael J (1985) Growth of loblolly pine treated with hexazinone, sulfometuron methyl, and metsulfuron methyl for herbaceous weed control. Southern Journal of Applied Forestry 9 (1):20-26

Millar C (1999) Evolution and biogeography of Pinus radiata, with a proposed revision of its Quaternary history. New Zealand Journal of Forestry Science 29 (3):335-365

Morse AM, Peterson DG, Islam-Faridi MN, Smith KE, Magbanua Z, Garcia SA, Kubisiak TL, Amerson HV, Carlson JE, Nelson CD, Davis JM (2009) Evolution of genome size and complexity in *Pinus*. PLoS ONE 4 (2):e4332

Nagasaki H, Arita M, Nishizawa T, Suwa M, Gotoh O (2005) Species-specific variation of alternative splicing and transcriptional initiation in six eukaryotes. Gene 364:53-62

Nagasaki H, Arita M, Nishizawa T, Suwa M, Gotoh O (2006) Automated classification of alternative splicing and transcriptional initiation and construction of visual database of classified patterns. Bioinformatics 22 (10):1211-1216

Neale DB, Ingvarsson PK (2008) Population, quantitative and comparative genomics of adaptation in forest trees. Current Opinion in Plant Biology 11 (2):149-155

Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Molecular Biology and Evolution 3 (5):418-426

Orphanides G, Reinberg D (2002) A unified theory of gene expression. Cell 108 (4):439-451

Posada D (2008) jModelTest: Phylogenetic model averaging. Molecular Biology and Evolution 25 (7):1253-1256

Price R (1989) The genera of Pinaceae in the southeastern United States. Journal of the Arnold Arboretum 70 (2):247-305

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19 (12):1572-1574

Sammeth M, Foissac S, Guigo R (2008) A general definition and nomenclature for alternative splicing events. PLoS Computational Biology 4 (8):e1000147

Schmidtling RC (2001) Southern pine seed sources. USDA Forest Service, Southern Research Station, General Technical Report SRS-44

Schmidtling R (2003) The southern pines during the Pleistocene. ISHS Acta Horticulturae 615:203-209

Schmidtling RC, Carroll E, LaFarge T (1999) Allozyme diversity of selected and natural loblolly pine populations. Silvae Genetica 48 (1):35-45

Schultz RP (1999) Loblolly - the pine for the twenty-first century. New Forests 17 (1-3):71-88

Severing EI, van Dijk ADJ, Stiekema WJ, van Ham RCHJ (2009) Comparative analysis indicates that alternative splicing in plants has a limited role in functional expansion of the proteome. BMC Genomics 10:154

Shannon CE (1948) A mathematical theory of communication. Bell System Technical Journal 27 (3):379-423

Shaw GR (1914) The genus *Pinus*. Riverside Press, Cambridge, MA

Shimodaira H (2002) An approximately unbiased test of phylogenetic tree selection. Systematic Biology 51 (3):492-508

Shimodaira H, Hasegawa M (2001) CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics 17 (12):1246-1247

Shurkhal A, Podogas A, Zhivotovsky L (1992) Allozyme differentiation in the genus *Pinus*. Silvae Genetica 41 (2):105-109

Smouse P, Saylor L (1973) Studies of the *Pinus rigida-serotina* complex II. Natural hybridization among the *Pinus rigida-serotina* complex, *P. taeda* and *P. echinata*. Annals of the Missouri Botanical Garden 60:192-203

Sorenson MD, Franzosa EA (2007) TreeRot, version 3. Boston University, Boston, MA

Stamatakis A (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22 (21):2688-2690

Stamatakis A, Blagojevic F, Nikolopoulos DS, Antonopoulos CD (2007) Exploring new search algorithms and hardware for phylogenetics: RAxML meets the IBM cell. Journal of VLSI Signal Processing Systems for Signal Image and Video Technology 48 (3):271-286

Stamatakis A, Hoover P, Rougemont J (2008) A rapid bootstrap algorithm for the RAxML web servers. Systematic Biology 57 (5):758-771

Stamatakis A, Ludwig T, Meier H (2005) RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21 (4):456-463

Su ZX, Wa JM, Yu J, Huang XQ, Gu X (2006) Evolution of alternative splicing after gene duplication. Genome Research 16 (2):182-189

Sukumaran J, Holder M (2010) DendroPy: a Python library for phylogenetic computing. Bioinformatics 26 (12):1569-1571

Swofford D (2003) PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4. Sinauer Associates, Sunderland, MA

Syring J, Farrell K, Businsky R, Cronn R, Liston A (2007) Widespread genealogical nonmonophyly in species of *Pinus* subgenus *Strobus*. Systematic Biology 56 (2):163-181

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123 (3):585-595

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Molecular Biology and Evolution 24 (8):1596-1599

Tavaré S (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. Lectures on Mathematics in the Life Sciences 17:57–86

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

USDA Agriculture and Food Research Initiative (AFRI) (2010) Competitive Grants Program SB, FY 2010 Request for Applications. National Loblolly Pine Genome Sequencing, Program Area Code: A6141, pp 10-11

USDA Forest Service. Southern Forest Resource Assessment. http://www.srs.fs.usda.gov/sustain/. Accessed May 2010

Valentine JW (2000) Two genomic paths to the evolution of complexity in bodyplans. Paleobiology 26 (3):513-519

van Minnen J, Strengers B, Eickhout B, Swart R, Leemans R (2008) Quantifying the effectiveness of climate change mitigation through forest plantations and carbon sequestration with an integrated land-use model. Carbon Balance and Management 3:3

von Bubnoff A (2008) Next-generation sequencing: the race is on. Cell 132 (5):721-723

Wang BB, Brendel V (2006) Genomewide comparative analysis of alternative splicing in plants. Proceedings of the National Academy of Sciences of the United States of America 103 (18):7175-7180

Wells OO, Switzer GL, Schmidtling RC (1991) Geographic variation in Mississippi loblolly pine and sweetgum. Silvae Genetica 40 (3-4):105-119

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature 452 (7189):872-876

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*. Molecular Biology and Evolution 24 (1):90-101

Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA (2003) The evolution of transcriptional regulation in eukaryotes. Molecular Biology and Evolution 20 (9):1377-1419

Wu J, Krutovskii KV, Strauss SH (1999) Nuclear DNA diversity, population differentiation, and phylogenetic relationships in the California closed-cone pines based on RAPD and allozyme markers. Genome 42 (5):893-908

Yang ZH (1993) Maximum-likelihood-estimation of phylogeny from DNA-sequences when substitution rates differ over sites. Molecular Biology and Evolution 10 (6):1396-1401

Zwickl D (2006) Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Ph.D. dissertation, The University of Texas, Austin, TX

APPENDIX A

SUPPLEMENTARY FIGURES FOR SECTION 2

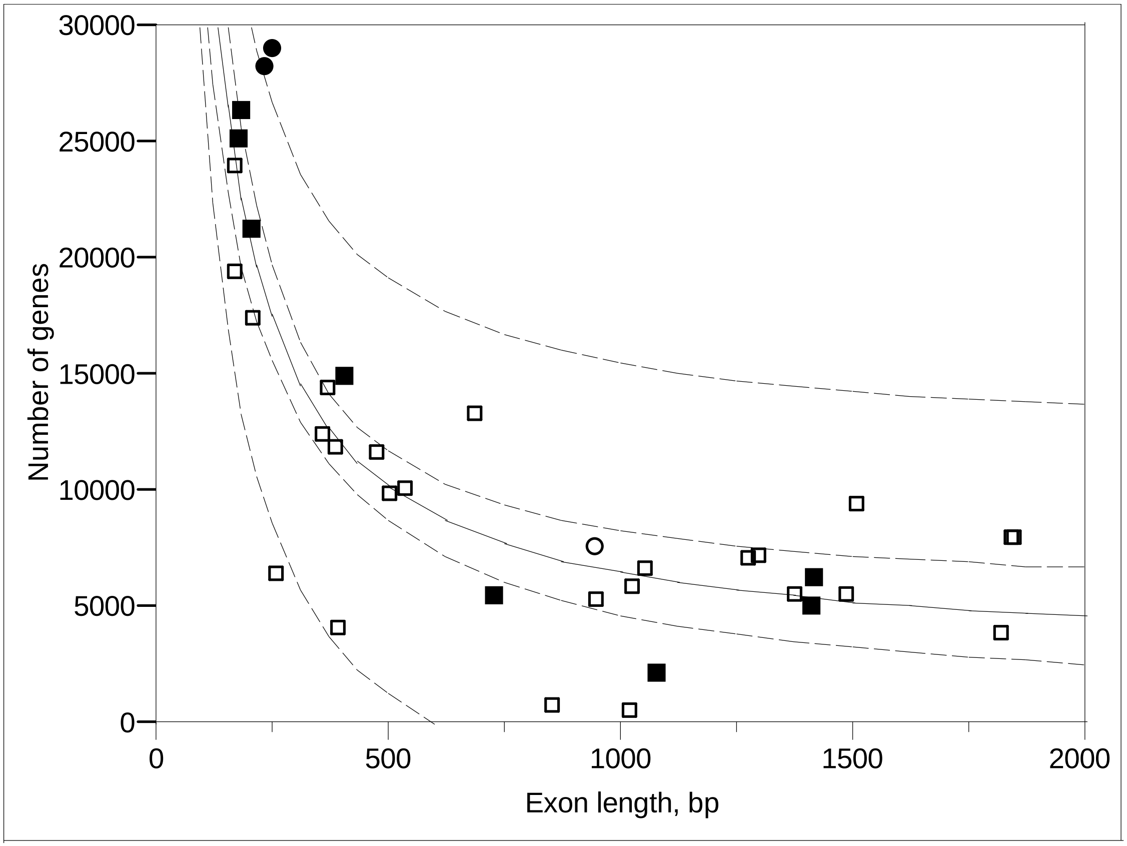


Fig. A1 Correlation of number of all genes and mean exon length

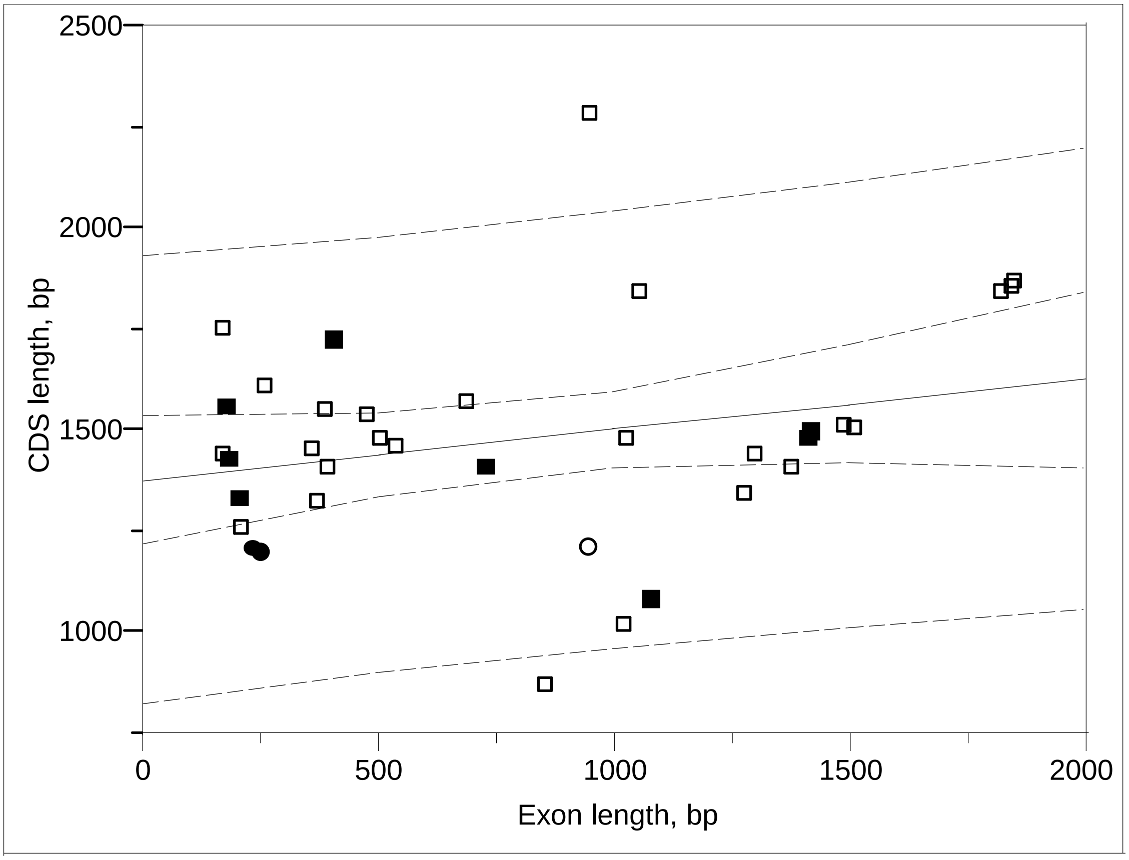


Fig. A2 Correlation of mean CDS length and mean exon length

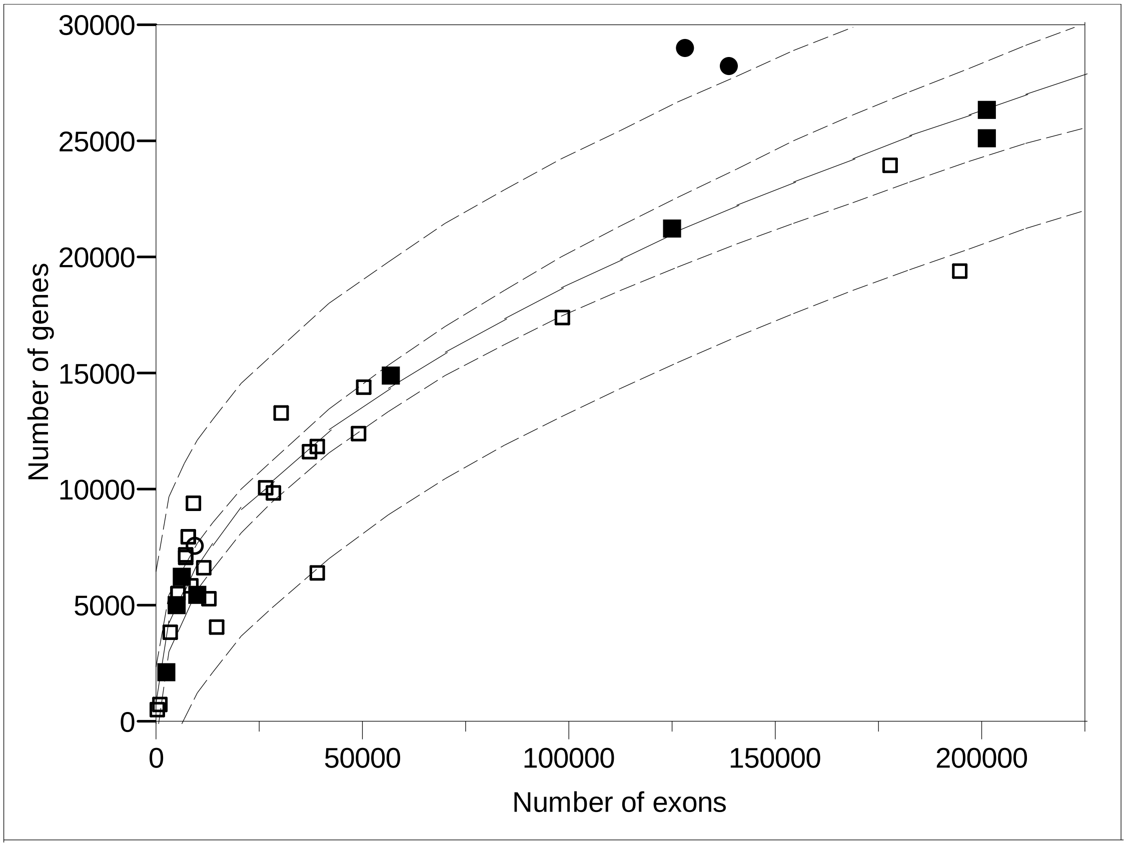


Fig. A3 Correlation of number of all genes and number of all exons

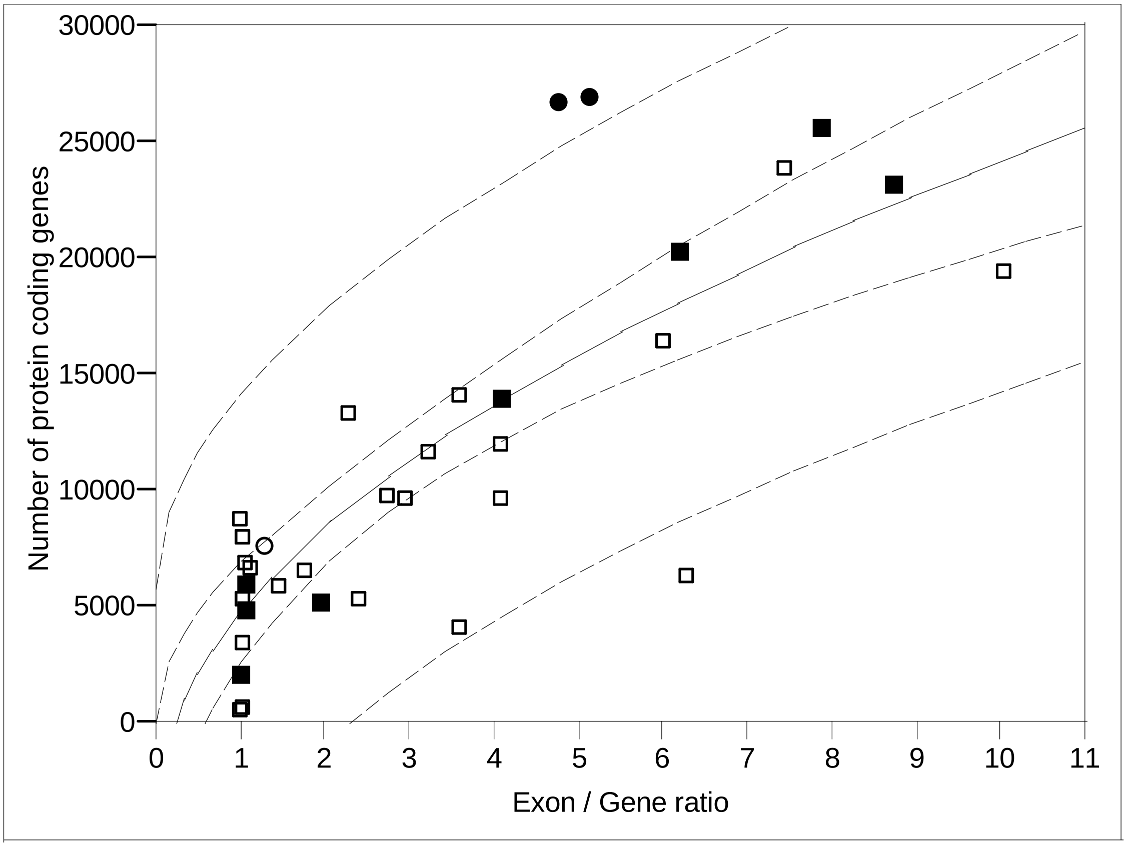


Fig. A4 Correlation of number of protein coding genes and exon/gene ratio

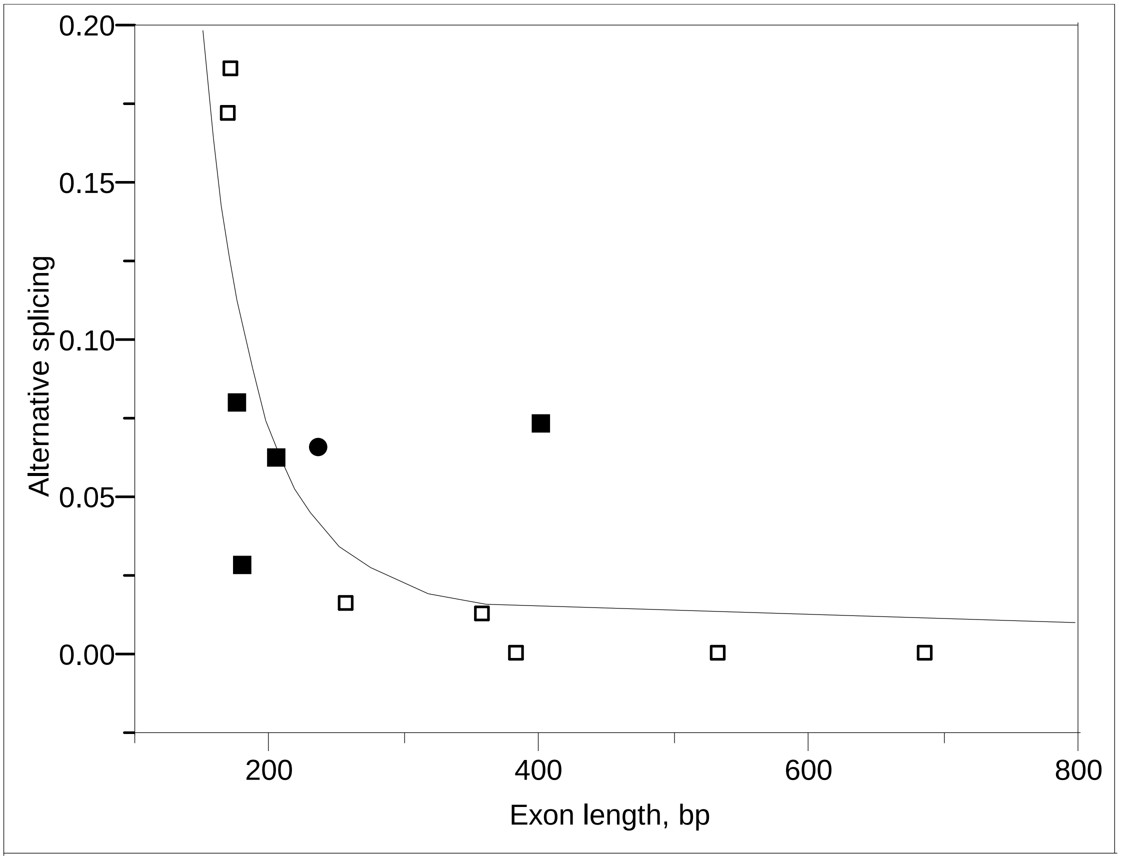


Fig. A5 Correlation of alternative splicing ratio and mean exon length. Only species with alternative splicing were considered

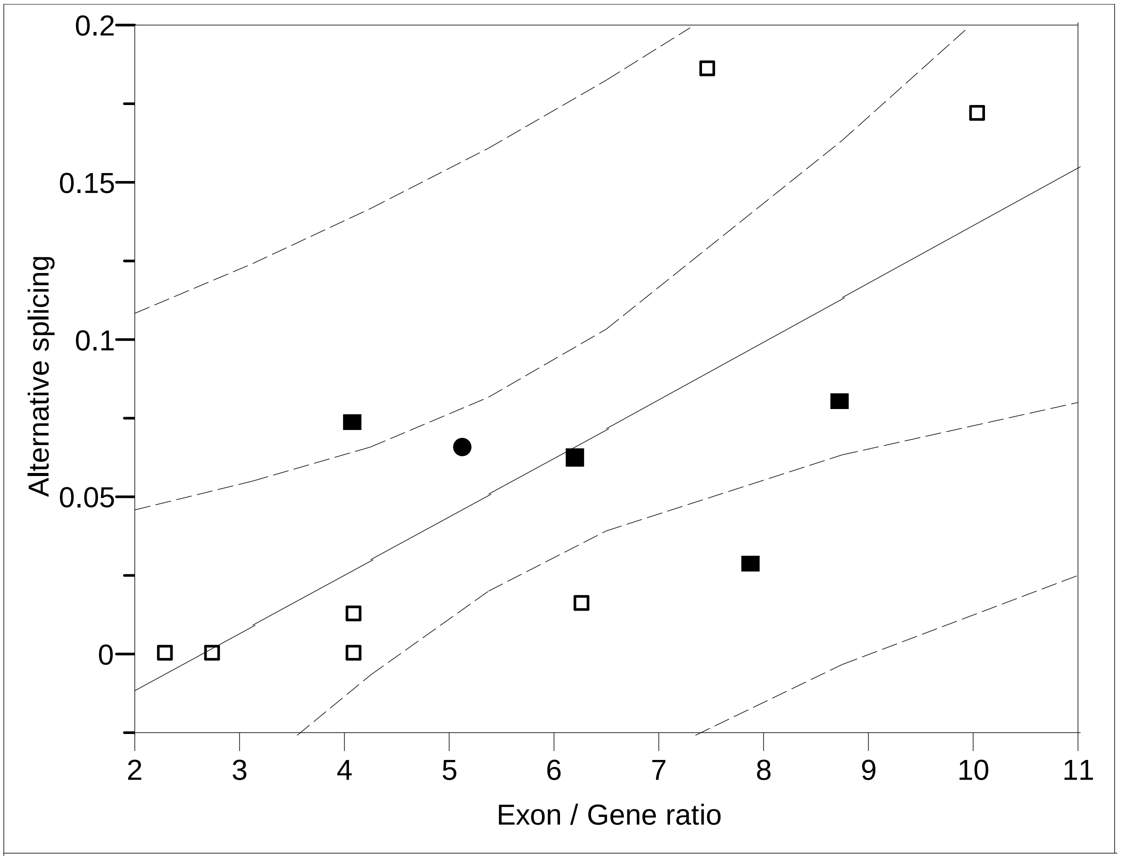


Fig. A6 Correlation of alternative splicing ratio and exon/gene ratio. Only species with alternative splicing were considered

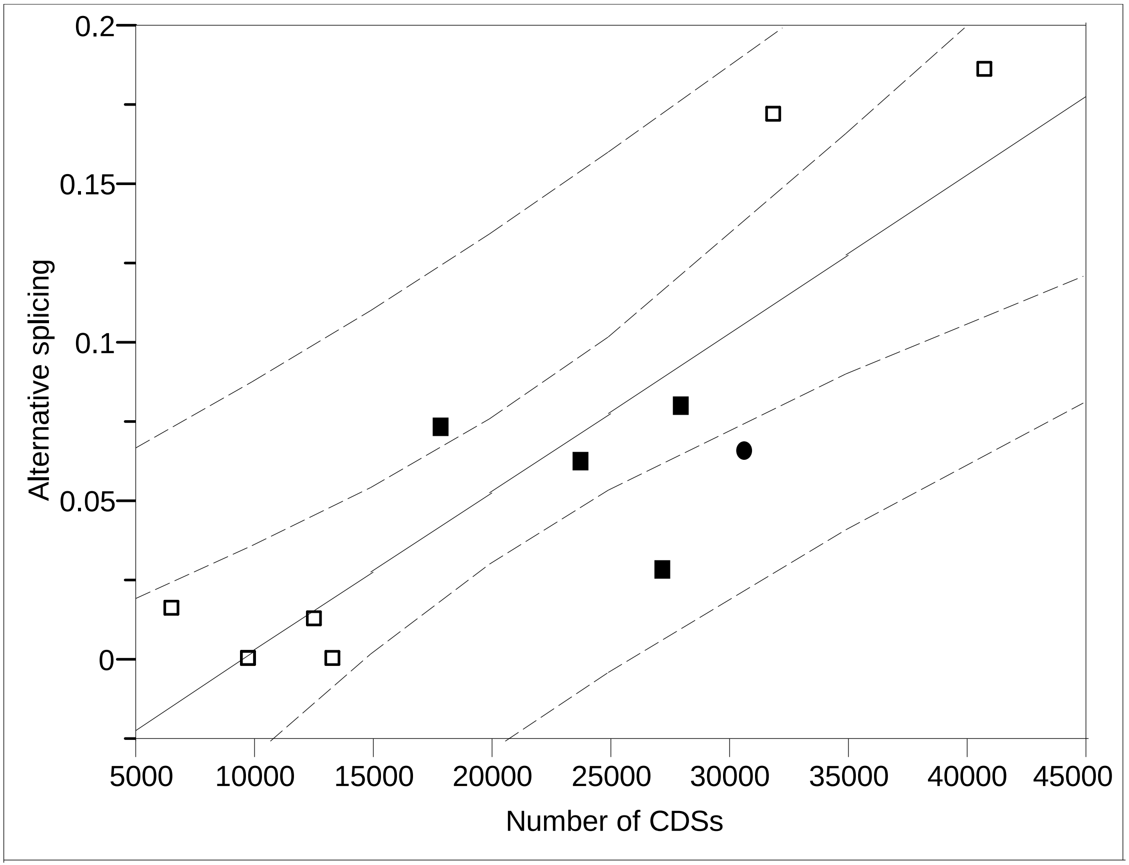


Fig. A7 Correlation of alternative splicing ratio and number of all CDSs. Only species with alternative splicing were considered

VITA

Name: Tomasz Edmund Koralewski

Address: 2138 TAMU
College Station, Texas 77843-2138
Phone: (979) 458 0471

Email address: tkoral@tamu.edu

Education: Ph.D., Genetics
Texas A&M University, College Station, Texas
2010

B.S., Biology
Kazimierz Wielki University in Bydgoszcz, Poland
2004

M.S., Electronics and Telecommunications Engineering
University of Technology and Agriculture, Bydgoszcz, Poland
2001