Contrasting Rates of Molecular Evolution and Patterns of Selection among Gymnosperms and Flowering Plants

Amanda R. De La Torre,*[,1,2] Zhen Li,[3,4] Yves Van de Peer,[3,4,5] and Pär K. Ingvarsson[2,6]

[1]Department of Plant Sciences, University of California–Davis, Davis, CA

[2]Department of Ecology and Environmental Science, Umeå University, Umeå, Sweden

[3]Department of Plant Systems Biology, VIB, Ghent, Belgium

[4]Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium

[5]Genomics Research Institute, University of Pretoria, Hatfield Campus, Pretoria, South Africa

[6]Department of Plant Biology, Uppsala Biocenter, Swedish University of Agricultural Sciences, Uppsala, Sweden

***Corresponding author:** E-mail: ardelatorre@ucdavis.edu.

Associate editor: Stephen Wright

Abstract

The majority of variation in rates of molecular evolution among seed plants remains both unexplored and unexplained. Although some attention has been given to flowering plants, reports of molecular evolutionary rates for their sister plant clade (gymnosperms) are scarce, and to our knowledge differences in molecular evolution among seed plant clades have never been tested in a phylogenetic framework. Angiosperms and gymnosperms differ in a number of features, of which contrasting reproductive biology, life spans, and population sizes are the most prominent. The highly conserved morphology of gymnosperms evidenced by similarity of extant species to fossil records and the high levels of macrosynteny at the genomic level have led scientists to believe that gymnosperms are slow-evolving plants, although some studies have offered contradictory results. Here, we used 31,968 nucleotide sites obtained from orthologous genes across a wide taxonomic sampling that includes representatives of most conifers, cycads, ginkgo, and many angiosperms with a sequenced genome. Our results suggest that angiosperms and gymnosperms differ considerably in their rates of molecular evolution per unit time, with gymnosperm rates being, on average, seven times lower than angiosperm species. Longer generation times and larger genome sizes are some of the factors explaining the slow rates of molecular evolution found in gymnosperms. In contrast to their slow rates of molecular evolution, gymnosperms possess higher substitution rate ratios than angiosperm taxa. Finally, our study suggests stronger and more efficient purifying and diversifying selection in gymnosperm than in angiosperm species, probably in relation to larger effective population sizes.

Key words: **gymnosperms, angiosperms, substitution rates, selection, mutation, life-history traits.**

Article

Introduction

The study of the forces of mutation and selection and their effects at the molecular and phenotypic levels is crucial to understand how species have evolved over time (Lynch 2010). Under a strictly neutral model, the fate of substitutions at the molecular level is mainly determined by mutation and random genetic drift instead of by natural selection (Kimura 1968). Under the neutral theory, mutations are assumed to be selectively neutral, nearly neutral ($S = 0$) or strongly deleterious ($S = -\infty$) while advantageous mutations are assumed to be too rare to have a significant effect on sequence evolution. Later, Ohta (1992) proposed a modified version of the neutral theory in which a substantial fraction of substitutions are caused by the random fixation of slightly deleterious mutations, and that a small fraction of all new mutations may have positive selection coefficients (the so called "nearly neutral model"). Although the strictly neutral model remains the most commonly used null model in population genetics, the controversy regarding the different models of evolution often comes down to a discussion about the importance of positive selection (Nielsen and Yang 2003).

Selective pressures on amino acid mutations are often measured in comparative studies using the ratio of nonsynonymous to synonymous substitution rates (also called "omega" and denoted by ω or dN/dS). When an amino acid change is neutral, the rate of fixation will be the same as that of a synonymous mutation, and $\omega = 1$. Amino acid altering substitution rate ratios are denoted by $\omega < 1$ or $\omega > 1$, indicating negative selection, and positive selection, respectively (Yang and Nielsen 2002). Omega thus also gives an estimate of the rate of substitution at selected sites in comparison to the neutral substitution rate. Several codon-based likelihood models have been developed to study the distribution of ω among sites (site models) or among branches (branch models) using a phylogenetic framework (Yang et al. 1997; Yang and Nielsen 1998; Yang et al. 2000; Nielsen and Yang 2002). Estimating ω along particular lineages of a phylogeny allows the testing of hypotheses regarding

the relative effects of selection in certain lineages of interest (Nielsen 2005); and also for the estimation of the distribution of selection coefficients among taxa. Because positive selection most likely affects only a few sites at a few time points, the evaluation of individual sites (site models) and branches (branch models) has more power to detect adaptive evolution than the pairwise sequence comparison, in which substitution rates are averaged over all amino acid sites (Yang 2002).

Substitution rates have been reported to be variable across the tree of life, however the causes underlying this variation remain uncertain. In flowering plants, this variation has been explained by differences in life forms, height, generation times, genome size, environmental variables, and species richness (Gaut et al. 1992; Smith and Donoghue 2008; Lanfear et al. 2013; Bromham et al. 2015). Variable substitution rates have also been reported across nuclear and organelle genomes, and among genes with different functional categories (Bromham et al. 2015). Although considerable attention has been given to flowering plants (angiosperms), very few reports on the rate of molecular evolution exist for their sister seed plant clade, the gymnosperms, and to our knowledge differences in rates of molecular evolution among angiosperms and gymnosperms have never been tested within a phylogenetic framework.

Gymnosperms are an ancient and widespread plant clade that represent four of the five main lineages of seed plants, and includes cycads, ginkgos, gnetophytes, and conifers (Wang and Ran 2014). Gymnosperm lineages separated from each other during the Late Carboniferous to the Late Triassic (311–212 mya), earlier than the occurrence of the earliest extant angiosperms around 300 mya (Magallón et al. 2013). Despite being dominant through most of the Mesozoic, gymnosperms were severely affected by extreme climatic shifts especially during the late Neogene, which may have favored the disproportionate loss of ancient lineages and their replacement by younger lineages in the Northern Hemisphere (Won and Renner 2006; Crisp and Cook 2011; Leslie et al. 2012). Cenozoic extinctions may have contributed to the low diversity of extant gymnosperms (Wang and Ran 2014), however they are unlikely to explain the 30-fold difference in species diversity between gymnosperm and angiosperm species.

Besides differences in species diversity, gymnosperms and angiosperms differ in a number of features, of which contrasting reproductive biology (mating system, pollination type, and seed morphology), physiology (water-conducting systems), and life spans are the most prominent (Leitch and Leitch 2012; De La Torre et al. 2014a). Gymnosperms are typically outcrossing species, in which wind plays the main role in the pollination and dispersal of uncovered seeds. In addition, widespread gymnosperms are thought to have large effective population sizes and weak population structure (Neale and Kremer 2011). The recent genome sequencing of gymnosperm species has revealed that the enormous genomes of gymnosperms (20–40 Gb) are mainly composed of large and variable sets of transposable elements, and that they have similar numbers of protein-coding genes compared to other plant species (De La Torre et al. 2014a).

The highly conserved morphology of gymnosperms, evidenced by the similarity of extant species to the earliest fossil records, and the high levels of macrosynteny among conifers (Pavy et al. 2012), have led scientists to believe that gymnosperms are slow-evolving plants (Won and Renner 2006). However, this remains a controversial issue, with studies showing opposing results (Willyard et al. 2007; Palme et al. 2009; Buschiazzo et al. 2012; Chen et al. 2012). Limitations of previous studies include the use of a small number of genes and/or species, the use of highly diverged species compared using different gene sets, and the lack of a phylogenetic framework that includes species from different taxonomic families. Considering that the rate of molecular evolution strongly depends on the selective constraints of proteins or amino acids, and these constraints are variable among genes, it seems risky to conclude significant differences in rates of molecular evolution from the comparison among different gene sets. Moreover, a recent whole-genome study in *Picea* species has found contrasting rates of sequence divergence among genes in relation to their functional category, duplication status and gene family size (De La Torre et al. 2015a). Evolutionary and phylogenetic analyses among gymnosperms and angiosperms have also been limited by the absence of orthologous genes that allow for such comparisons.

In this study, we used a newly identified set of 42 single-copy genes obtained from whole genomic and/or transcriptomic data from a broad taxonomic sampling that includes all conifers (with the exception of Araucariaceae), cycads, ginkgo, and many angiosperms with sequenced genomes (Li Z, De La Torre AR, Sterck L, Canovas FM, Avila C, Von Arnold S, Ingvarsson PK, Van de Peer Y, in review). We aimed to test for differences in the rate of molecular evolution among gymnosperm and angiosperm species, and to understand the possible causes driving any such variation. With this, we hope we can contribute to a better understanding of the complex evolutionary relationships among major plant clades and to elucidate the main evolutionary processes that have shaped the seed plants we see today.

Results

Rates of Sequence Divergence

When evaluating pairwise estimates between species for each taxonomic family in the angiosperm and gymnosperm phylogeny (fig. 1), we found significant differences in the rates of sequence divergence between the two major plant clades. The results of the sign test indicated that dN, dS, and ω were significantly different between angiosperms and gymnosperms across the 42 genes evaluated. The number of synonymous substitutions per site (dS) was lower in gymnosperms than in angiosperms in all of the 42 genes studied ($P < 0.001$). Similarly, the number of non-synonymous substitutions per site (dN) was lower in gymnosperms for 86% of the studied genes ($P < 0.001$), whereas the ratio of non-synonymous/synonymous substitutions (ω) was higher in gymnosperms for 76.2% of the genes ($P < 0.001$) (see supplementary table S3, Supplementary Material Online). Consistent results were found when comparing branch estimations of dS

Table 1. Absolute Rates of Silent-Site Divergence (μ) for Each Taxonomic Family or Subfamily Based on 42 Single-Copy Nuclear Genes in 31 Gymnosperms and 34 Angiosperms Species.

Plant Clade	Taxa	Subtaxa	dS	*T* (years)	μ (site/year)
Gymnosperms	Pinaceae		0.102		7.28067E−10
		Pinus	0.070	84500000	4.16479E−10
		Cedrus_Abies	0.128	129700000	4.9387E−10
		Larix_Pseudotsuga	0.148	61600000	1.19939E−09
		Picea	0.061	38300000	8.02526E−10
	Cupressaceae		0.189	159200000	5.92216E−10
	Taxaceae		0.225	153000000	7.34573E−10
	Podocarpaceae		0.170	146100000	5.81248E−10
	Cycadales		0.333	248100000	6.70325E−10
	Gnetophytes		0.073	25000000	1.45547E−09
Angiosperms	Brassicaceae		0.352	27000000	6.51648E−09
	Malvaceae		0.373	40900000	4.56E−09
	Euphorbiaceae		0.463	51000000	4.54382E−09
	Fabaceae		0.502	53100000	4.72809E−09
	Cucurbitaceae		0.239	20000000	5.9795E−09
	Rosaceae		0.421	52300000	4.02113E−09
	Solanaceae		0.084	7300000	5.75068E−09
	Poaceae		0.372		6.0277E−09
		Poaceae I	0.457	39700000	5.76174E−09
		Poaceae II	0.287	22800000	6.29366E−09

Note.—Divergence times (*T*) are based on fossil calibration data and published studies (see supplementary table S2, Supplementary Material Online).

($t = 3$, df = 57.87, $P < 0.01$) and ω ($t = -4.0379$, df = 37.8, $P < 0.001$); however dN was not significantly different between angiosperms and gymnosperms when comparing branch estimations in all terminal branches ($t = -0.319$, df = 52.34, $P = 0.751$) (see supplementary table S4, Supplementary Material Online). The most contrasting differences were found in the absolute rates of silent-site divergence (μ), with angiosperms rates being, on average, seven times higher than that observed in gymnosperm species (5.35×10^{-9} vs. 7.71×10^{-10} synonymous substitutions/site/year, respectively). Families showing the highest rates were Brassicaceae and Poaceae, and the ones with the lowest rates were Cuppresaceae and Podocarpaceae. Within gymnosperms, Gnetophytes had the highest rates of sequence divergence (table 1).

Estimation of the Variation in Selective Pressures (ω) among Branches

We estimated the variation in selective pressures (ω) among stem branches in the gymnosperm and angiosperm phylogenies, using the branch models in PAML (fig. 1). The results of the model testing indicated that the most parameter-rich hypothesis (H4) fits the data best. This hypothesis assumes that all ancestral and terminal branches have different omegas. The second best hypothesis was H3, which suggests a long-term shift in selective pressure resulting in both lineages having different omegas ($\omega 0 \neq \omega A \neq \omega G$). Likelihood ratio tests were significant for H1–H3 ($2\Delta l = 3562$, $P < 0.001$) and H2–H3 ($2\Delta l = 129.31$, $P < 0.001$). Hypothesis H2 ranked third followed by H1 and, hypothesis H0, where there were no differences in selective pressures between lineages ($\omega 0 = \omega G = \omega A$) had the worst fit to the data (table 2). Omega ratios for gymnosperms were significantly higher than for angiosperms in all hypotheses tested with the exception of H0 (one-ratio model).

Estimation of Substitution Parameters Using Site Models

A discrete model (M3) that uses an unconstrained discrete distribution to model heterogeneous omega ratios among sites (Yang et al. 2000), showed the best fit to the data in all angiosperm taxonomic families with the exception of Brassicaceae (model 7: beta). The nearly neutral model (model 1a) that assumes a proportion of conserved sites with $\omega_0 < 1$ and a proportion of neutral sites with $\omega_1 = 1$, and the selection model (M2a) that allows an additional class of sites with $\omega_2 > 1$, had the worst fit in all angiosperm taxa. Excluding poorly fitted M1a and M2a, the average ω among all models ranged from 0.092 to 0.093 (Brassicaceae), 0.106–0.108 (Malvaceae), 0.117–0.125 (Fabaceae), 0.118–0.156 (Rosaceae), and 0.131–0.133 (Poaceae). Average ω ratios suggest a nonsynonymous mutation has only 9–13% as much chance as a synonymous mutation of being fixed, suggesting most sites are highly conserved in angiosperm taxa.

In gymnosperms, M8 (beta & ω) showed the best fit in Cupressaceae and Cycads, M10 (beta & gamma +1) in *Pinus*, and M2a (selection) in *Picea*. The beta model (M7) had the worst fit to the data in *Picea* and *Pinus*; and the gamma model (M5) in Cupressaceae and Cycads. Excluding poorly fitted M7 for *Picea* and *Pinus*, and M5 for Cuppressaceae and Cycads, the average ω among all models ranged from 0.66–0.67 (*Pinus*), 0.37–0.41 (*Picea*), 0.167–0.178 (Cuppresaceae), and 0.2–0.216 (Cycads). Overall, these estimates suggest a significantly higher chance of fixation of nonsynonymous mutations in gymnosperms (17–67%) than in angiosperm taxa. Parameters estimates for each of the six site models tested in each of the five angiosperm and four gymnosperm families

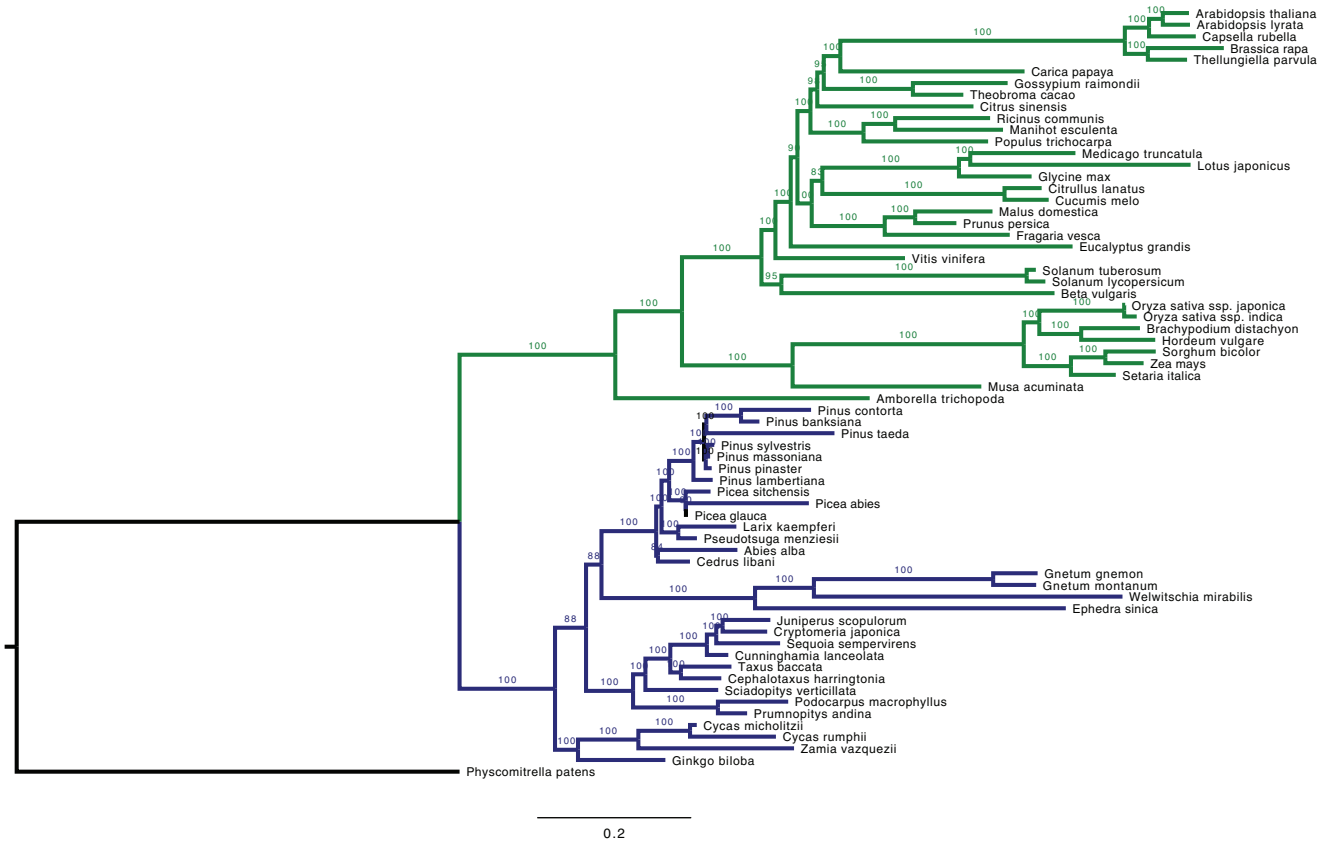

Fig. 1. Phylogenetic tree obtained from a concatenated alignment of 42 single-copy genes and 66 species distributed between two main seed plant lineages (angiosperms in green, gymnosperms in blue, and outgroup in black) inferred by RAxML. Stem branches for angiosperms and gymnosperms were used for model testing using the branch models implemented in codeml. Results of the model testing can be found in table 2.

Table 2. Parameter Estimates under Models of Variable Omega (ω) among Branches from the Gymnosperm and Angiosperm Lineages Based on a Concatenated Alignment of 31,737 Nucleotide Sites in 61 Species.

Hypothesis	Model	No. of Parameters	Background	Foreground	Parameter Estimates	ln *L*	AIC
H0	**M0**	**1**	$\omega 0 = \omega G = \omega A$		$\omega = 0.11356$	**−621777.74**	**1243557.48**
H1	**M2**	**2**	$\omega 0 = \omega G$	ωA	$\omega 0 = \omega G = 0.1238$; $\omega A = 0.0989$	**−621644.6499**	**1243293.3**
H2	**M2**	**2**	$\omega 0 = \omega A$	ωG	$\omega 0 = \omega A = 0.0935$; $\omega G = 0.2603$	**−619928.0209**	**1239860.042**
H3	**M2**	**3**	$\omega 0$, ωA, ωG		$\omega 0 = 0.0857$; $\omega A = 0.1024$; $\omega G = 0.2631$	**−619863.362**	**1239732.724**
H4	**M1**	**120**			ω **ranges between 0.05 to 0.76**	**−617598.8386**	**1235437.677**

Note.—We tested the following hypotheses: H0: Homogeneous selective pressure in both clades ($\omega 0 = \omega G = \omega A$); H1: Selective Pressure in the Angiosperm clade ($\omega 0 = \omega G$, ωA); H2: selective pressure in the gymnosperm clade ($\omega 0 = \omega A$, ωG); H3: long-term shift in selective pressure resulting in both clades having different omegas ($\omega 0$, ωG, ωA); and H4: all terminal branches in both clades have different omegas. For hypotheses H1 to H3, ωG and ωA identify estimated omegas on stem branches of the gymnosperm and angiosperm lineages, according to phylogenetic tree in figure 1; $\omega 0$ represents the estimated omega in unselected branches (background). Parameter estimates of model M1 (hypothesis H4) are detailed in supplementary table S7, Supplementary Material Online.

are included in supplementary table S5, Supplementary Material Online.

When comparing the proportion of conserved, neutral and positively selected sites under the discrete model (M3), we observed that all taxa had a higher proportion of conserved followed by neutral and then by positively selected sites. Conserved sites ($\omega < 1$) in angiosperms ranged from 55% to 77%, whereas gymnosperm sites ranged from 66% to 76%; differences between taxa were not significant ($P = 0.62$). The proportion of neutral sites ($\omega = 1$) ranged from 21% to 39% in angiosperms, and from 21% to 28% in gymnosperms; however the differences between medians were not significant ($P = 0.22$) (supplementary table S5, Supplementary Material Online).

Inference of the Proportion of Sites under Selection

Models that allow for sites under positive selection such as M2a, M3, and M8, all suggested the presence of a very small number of weakly positively selected sites in angiosperm taxa. Models M3 and M8 identified between 0.7–1.3% sites under weak diversifying selection ($\omega = 1.65$ and $\omega = 1.38$, respectively) in Brassicaceae. However, likelihood ratio test statistics for comparing M1a and M2a, and M7 and M8 did not show significant results (table 3). In addition, the NEB analyses did not identify any sites under positive selection. For Malvaceae,

Table 3. Results of the Likelihood Ratio Tests for Evidence of Positive Selection in Each of the Taxonomic Families Studied.

Taxa	M1a–M2a (2ΔL)	*P* Value	M7–M8 (2ΔL)	*P* Value
Brassicaceae	0	NS	2.577926	0.2766
Malvaceae	0	NS	33.063102	0.00000007
Fabaceae	0	NS	22.744606	0.00001151
Rosaceae	0	NS	6.429714	0.04016
Poaceae	0	NS	15.428744	0.0004463
Pinus	511.4601	0	549.377428	0
Picea	116.785146	0	116.646654	0
Cupressaceae	0	NS	15.11157	0.000523
Cycadales	0	NS	6.535688	0.03808

NOTE.—Comparisons between site models M1a and M2a, and M7 and M8 for each taxonomic family were evaluated.

four sites under positive selection were identified by the BEB analysis ($P > 95\%$), and the likelihood ratio test for comparison between M7–M8 was significant ($2\Delta l = 33.063$, $P < 0.001$, $df = 2$). However, none of the other tested models identified sites under positive selection. Weakly to moderate diversifying selection ($\omega = 1.89$ and $\omega = 2.75$) was found in a very small proportion of sites (1.1% and 0.5%) in Fabaceae, based on the results of M3 and M8. In addition, the NEB analysis identified two sites under positive selection using M3, and the likelihood ratio test for comparison between M7–M8 was significant ($2\Delta l = 22.74$, $P < 0.001$, $df = 2$). In Rosaceae, M3 and M8 identified 0.2% of sites under strong diversifying selection. In addition, the BEB analysis identified two sites under positive selection using M3, and the likelihood ratio test for comparison between M7–M8 was significant ($2\Delta l = 6.42$, $P < 0.05$, $df = 2$). Finally, weakly diversifying selection ($\omega = 1.53$ and $\omega = 1.85$) was found in 0.7 and 1.3% sites in Poaceae. Four sites were identified by the BEB analyses in both M3 and M8, and the likelihood ratio test for M7–M8 was significant ($2\Delta l = 15.428$, $P < 0.001$ $df = 2$) (table 3, see supplementary table S5, Supplementary Material Online).

In contrast to angiosperm taxa, evidence for moderate to strong diversifying selection and a higher number of sites under positive selection was found in gymnosperm taxa. In *Pinus*, M2a, M3, and M8 identified 16.3%, 23.13%, and 22.21% of sites under moderate to strong diversifying selection ($\omega = 3.2$, $\omega = 2.73$, and $\omega = 2.8$, respectively). This taxonomic family had the highest number of sites under positive selection (240, 245, and 1070) evidenced by the results of the BEB analyses under M2a and M8, and the NEB analysis under M3. Likelihood ratio tests were significant for both the M1a–M2a ($2\Delta l = 511.46$, $P < 0.001$ $df = 2$) and the M7–M8 ($2\Delta l = 549.37$, $P < 0.001$ $df = 2$) comparisons. Similarly, in *Picea*, M2a identified 21.48% of weakly selected sites ($\omega = 1.19$), and 3.25% of sites under strong positive selection ($\omega = 3.29$). Models M3 and M8 identified 5.6% of sites under moderate positive selection ($\omega = 2.8$). The number of sites under positive selection was highly variable among models: 1, 301, and 885 based on the BEB analyses under M2a and M8, and by the NEB results based on M3. The LRT statistics were significant for both the M1a–M2a and M7–M8 comparisons (table 3). A small proportion (1.2–1.4%) of sites under moderate positive selection ($\omega = 2.21$ and $\omega = 2.32$) were evidenced by the results of M3 and M8 in Cuppresaceae. Three sites were found to be under positive selection based on the NEB analysis of M3. Finally, in Cycads, M3 identified 4.6% of weakly selected sites ($\omega = 1.71$); and M8 identified 1.9% of sited under moderate selection ($\omega = 2.38$). The LRT statistics for M7–M8 comparisons were significant for both Cuppressaceae and Cycads (table 3, see supplementary table S5, Supplementary Material Online).

Distribution of Selection Coefficients

Parameter estimates α and β from the gamma distribution of ω across sites in each taxonomic family were used to estimate the distribution of the selection coefficients of new mutations (Yang et al. 2000; Nielsen and Yang 2003). As expected, we found that α and β varied across taxonomic families, resulting in varying estimates of S and $f(S)$. For all taxonomic families, the distributions of selection coefficients for new mutations had a peak at $S = -5$, suggesting that the majority of sites are under purifying selection. In all angiosperms, except Brassicaceae, selection coefficients are mostly distributed with $S = [-10, 0]$; whereas in gymnosperms, the majority of selection coefficients are distributed within a wider range with $S = [-20, 10]$. Gymnosperm families had a greater proportion of sites under strong purifying selection ($S < -15$) than angiosperms ($t = -3.57$, $P < 0.05$); whereas angiosperm families had a greater proportion of neutral and mildly deleterious ($S \approx 0$ and $S > -15$) sites than gymnosperms ($t = 4.119$, $P < 0.05$) (fig. 2). The proportion of advantageous mutations (positive selection coefficients) also varied among families, with gymnosperm taxa having on average, more sites under positive selection than angiosperm taxa ($t = -5.74$, $P < 0.001$) (fig. 2).

Correlations with Life-History Traits

In order to explain the variation we observe in dN, dS, and ω between the major plant clades, we tested whether these variables showed any associations with life history traits. Our results suggest that generation time (time to reach reproductive maturity, measured in years) was strongly negatively correlated with μ (family level $r = -0.78$, $P < 0.001$) (table 4). Also, genome size (amount of DNA contained in a haploid nucleus and measured in picograms (1C)) was strongly negatively correlated with μ (family level $r = -0.67$) but positively correlated with ω (species level $r = 0.56$ $P < 0.001$, family level $r = 0.52$ $P < 0.05$) (table 4). Finally, species richness (number of extant species in each taxonomic family) was significantly correlated with dS ($r = 0.7$, $P < 0.001$) and dN ($r = 0.48$, $P < 0.05$) at the family level (not tested at species level).

We found significant differences ($P < 0.05$) in μ among groups with different life forms when using six groups (angiosperms dicots herbs, angiosperms dicots shrubs, angiosperms dicots trees, angiosperms monocots herbs, gymnosperms, and Gnetophytes). Differences in ω and μ were also significant among groups when comparing four groups (angiosperms herbs, angiosperms trees/shrubs, gymnosperms and Gnetophytes) and three groups (angiosperms herbs,

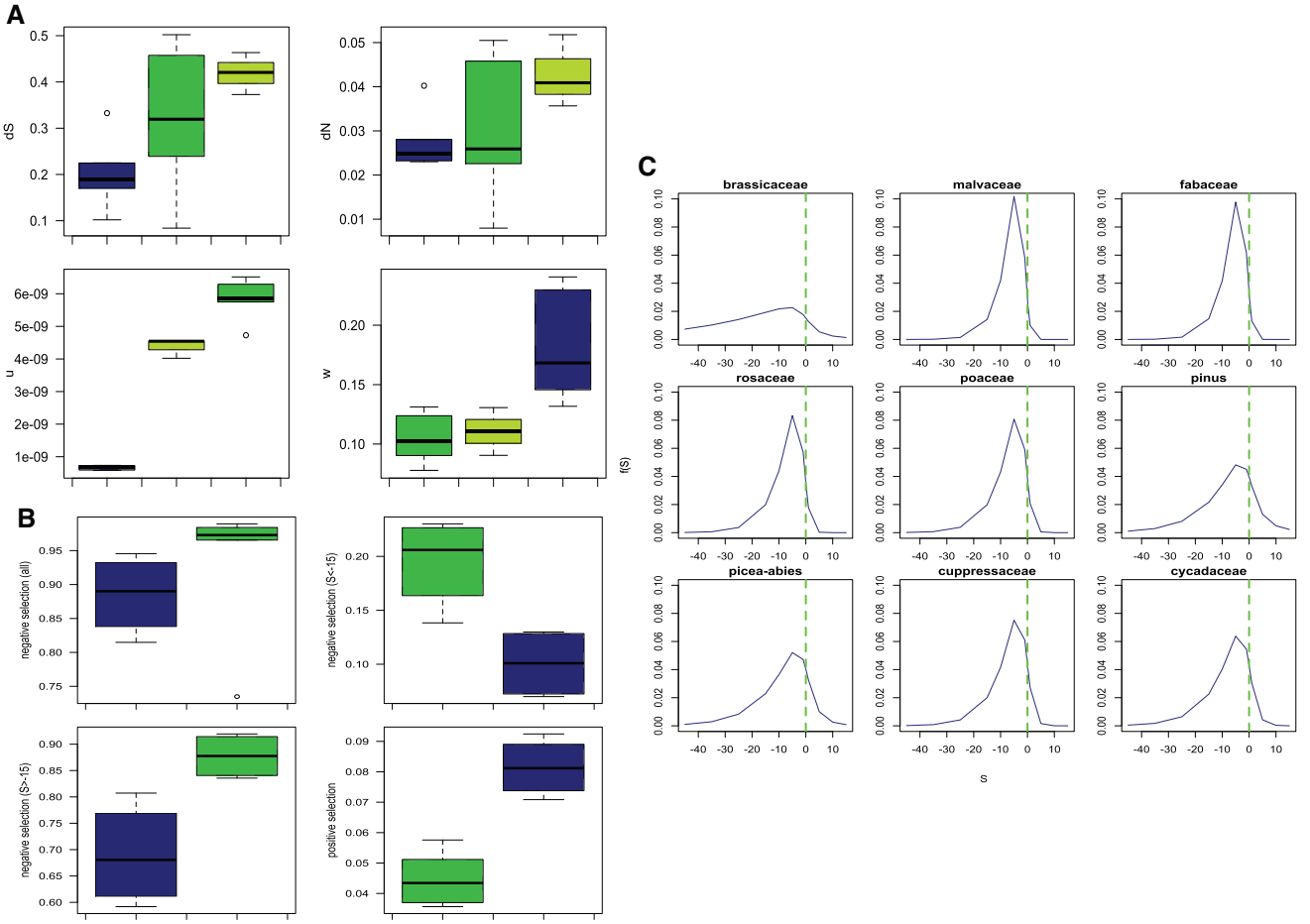

Fig. 2. (*A*) Boxplots showing differences in number of synonymous substitutions (dS), nonsynonymous substitutions (dN), absolute rate of silent-site divergence (μ), and substitution rate ratio (ω), among life forms defined as angiosperms herbs (green), angiosperms shrub/trees (light green), and gymnosperms (blue). Results of the statistical tests of comparisons among groups can be found in supplementary table S6, Supplementary Material Online. (*B*) Proportion of sites under negative selection, when all sites are evaluated (top left), when $S < -15$ (top right), and when $S > -15$ (bottom left); and proportion of sites under positive selection (bottom right). Green boxes represent angiosperm species, and blue boxes, gymnosperms. (*C*) Distribution of the selection coefficient (f(*S*)) of new mutations. It was calculated for all species in each taxonomic family studied, assuming a gamma distribution of ω among sites. Parameter estimates α and β were obtained from running the site model 5 in codeml (PAML) using a concatenated alignment of 29,000–31,000 sites per taxonomic family. Selection coefficients were then obtained replacing α and β in equation 1. Green dotted vertical lines are used to show the distribution of f(*S*) when $S \approx 0$ and $\omega = 1$ (neutrality). Taxonomic families containing less than 5 species were not analyzed due to difficulties in constructing the phylogenetic trees in RaxML.

angiosperms shrubs/trees, and gymnosperms) (see supplementary table S6, Supplementary Material Online). Number of synonymous (dS) and non-synonymous substitutions (dN) were not significantly different in any of the groups when assessed by $P < 0.05$. Boxplots showing differences among groups when groups are divided into three (angiosperms herbs, angiosperms trees/shrubs, and gymnosperms) are plotted in figure 2. In the Gnetophyte group, only *Gnetum* species were included and not *Welwitschia* and *Ephedra* (see Discussion part).

Discussion

Slower Rates of Molecular Evolution in Gymnosperms

Our study clearly suggests slower rates of molecular evolution in gymnosperm than in angiosperm protein-coding genes, evidenced by a lower number of synonymous substitutions (dS) and lower rates of silent-site divergence (μ).

Based on the evolutionary and phylogenetic analyses of 31,968 nucleotide sites from a wide taxonomic sampling, our results suggest μ values of $4.1–14 \times 10^{-10}$ in gymnosperm taxa, and $4–6.5 \times 10^{-9}$ in angiosperm taxa (table 1). Taken together, this corresponds to a 7-fold average variation among gymnosperms and angiosperms (7.71×10^{-10} vs. 5.35×10^{-9} synonymous substitutions/site/year, respectively). Our results lay within the ranges obtained by previous studies comparing a few Pinaceae species, suggesting the variation found in this taxonomic family is a relatively good representation of the variation found in all the gymnosperm taxa, with the exception of Gnetophytes. For example, Willyard et al. (2007), while comparing different species of *Pinus* found absolute rates of silent-site divergence of $7–13.1 \times 10^{-10}$. Similarly, Chen et al. (2012) estimated pairwise

Table 4. Correlation among Number of Synonymous (dS) and Nonsynonymous Substitutions (dN), Nonsynonymous/synonymous Rate Ratio (ω), and the Absolute Rate of Silent-Site Divergence (μ), with Life-History Traits Using the Phylogenetically Independent Contrast Method (PIC).

		Species Correlations		Taxonomic Family Correlations		
	Parameters	Generation Time (years)	Genome Size (1C)	Generation Time (years)	Genome Size (1C)	Species Richness
dS	*r*	−0.172	−0.135	−0.153	−0.311	0.703
	P-value	0.223	0.341	0.558	0.225	0.002
	df	50	50	15	15	15
dN	*r*	0.005	0.041	0.189	−0.026	0.485
	P-value	0.973	0.771	0.467	0.922	0.048
	df	50	50	15	15	15
M	*r*	−0.183	−0.037	−0.781	−0.670	0.057
	P-value	0.194	0.797	0.000	0.003	0.829
	df	50	50	15	15	15
Ω	*r*	0.212	0.567	0.482	0.527	−0.224
	P-value	0.132	0.000	0.050	0.030	0.387
	df	50	50	15	15	15

NOTE.—Substitution rates and life-history traits were estimated by species pairs (species correlations table); and averaged within taxonomic family (taxonomic family correlations table). We reject a null hypothesis of correlation equal to zero when $P < 0.05$ (shaded areas).

sequence divergence between three *Picea* and one *Taxus* species and obtained μ ranging from 5.5–12.4 $\times$ 10^{-10}. Finally, Buschiazzo et al. (2012), obtained a 25-fold difference when comparing substitution rates between orthologs of *Picea sitchensis* and *Pinus taeda* with *Arabidopsis thaliana*, and a 4-fold difference when comparing the same *Picea-Pinus* orthologs with *Populus trichocarpa*.

The pattern of slower rates of molecular evolution is consistent when grouping species according to their life forms, with gymnosperms showing lower dS and μ than angiosperms herbs and angiosperms shrubs and trees. This level of evolutionary conservation (which probably help explain the high levels of macrosynteny previously observed in some Pinaceae species) is surprising considering the ancient nature of the plant clade, which appeared on Earth much earlier than flowering plants.

Higher Substitution Rate Ratio (ω) in Gymnosperms than in Angiosperms

Equally surprising are the high substitution rate ratios (ω) found in gymnosperms despite their slow rates of evolution. Parameter estimates under models of variable ω between stem branches of the angiosperm and gymnosperms lineages based on a concatenated alignment of 31,737 nucleotide sites supported a model of different and higher ω in gymnosperms than in angiosperms (table 2). In addition, substitution parameters based on seven different site models, found angiosperms rate ratios varied from 0.09–0.13, whereas gymnosperms ratios varied from 0.17–0.67. Overall, these estimates suggest a significantly higher chance of fixation of nonsynonymous mutations than synonymous mutations in gymnosperms than in angiosperm taxa (supplementary table S5, Supplementary Material Online).

High ω ratios could result from a high dN over a low dS, however this does not seem to be the case for gymnosperms. Instead, higher ω ratios result from a low to moderate dN over a very low dS in gymnosperm genes. Our results are in contrast with previous smaller scale studies either showing significantly lower dS and dN as a cause of a 4-fold higher ω in gymnosperms than in angiosperms (Buschiazzo et al. 2012), or the ones that did not find significant differences in ω ratios among gymnosperms and angiosperms (Chen et al. 2012). These studies were limited by a small number of species that would not allow for evolutionary analyses within a phylogenetic context. When using a small number of species, pairwise estimates of dN, dS, and ω will strongly depend on the selection of species. In addition, the evaluation of individual sites (site models) and the branch models have more power to test variable ω ratios than the pairwise sequence comparisons, in which ω rates are averaged over all amino acid sites (Yang 2002). In the case of the Buschiazzo's study, highly diverged sets of species were used for comparisons (*P. sitchensis*–*P.taeda*, and *A.thaliana*–*P.trichocarpa*). High sequence divergence between species is often associated with difficulties in the alignment, different codon usage biases and nucleotide compositions. In addition, saturation of substitutions may be particularly problematic when using pairwise methods in comparison with branch methods.

Positive Selection and Adaptive Evolution in Gymnosperms

Under the Nearly Neutral Theory, the strength and efficacy of selection depends on the long-term effective population size (Ne). Theoretical approaches predict that as population sizes increase, the power of natural selection increases faster than the influx of new mutations (Akashi et al. 2012). Natural selection thus becomes more effective in removing deleterious mutations and in fixing advantageous mutations, which results in a lower substitution rate of deleterious mutations and also a higher substitution rate of advantageous mutations in large populations (Lanfear et al. 2014).

Current estimates suggest that species with large population sizes such as Drosophila, mice, bacteria, and rabbits show signs of adaptive evolution (Bierne and Eyre-Walker 2004; Carneiro et al. 2012; Phifer-Rixey et al. 2012). In contrast, little evidence of adaptive amino acid substitutions has been found in flowering plants (mostly with modest Ne $<$ 100,000) (Gossmann et al. 2010; Hough et al. 2013). Few exceptions

have been found in angiosperms with Ne >100,000 such as *Capsella grandiflora* (Slotte et al. 2010; Williamson et al. 2014); *Helianthus spp*. (Strasburg et al. 2011); and *Populus spp* (Ingvarsson 2010; Wang et al. 2016a, 2016b). In gymnosperms, previous reports of Ne range from 120,000–560,000 for *Picea* and *Pinus* species (Brown et al. 2004; Bouille and Bosuquet 2005; Syring et al. 2007).

In our study, we found a higher proportion of sites with positive selection coefficients and evidence of stronger diversifying selection in gymnosperms than in angiosperms ($P = 0.033$). Models that allow for sites under positive selection such as M2a, M3 and M8, all suggested the presence of a very small number (0.2–1.3%) of weakly positively selected sites ($\omega = 1.53–1.89$) in angiosperm taxa, in contrast to a higher number ($\leq$23.13%) of sites with moderate to strong diversifying selection ($\omega = 2.21–3.29$) in gymnosperm taxa. In addition, results of the BEB analysis using site model 2a identified 241 (0.759%) sites under selection in gymnosperms and none in angiosperms. The same analysis using site model M8 identified 546 (1.72%) sites under selection in gymnosperms and 10 (0.03%) in angiosperms. Finally, the NEB analysis using site model M3 identified 1960 (6.175%) sites in gymnosperms and only six (0.018%) in angiosperms. Likelihood ratio tests for the presence of diversifying selection were significant for *Picea* and *Pinus* (M1a–M2a comparison) and for all taxa with the exception of Brassicaceae under M7 and M8 comparisons. Under M3, *Picea* and *Pinus* showed the highest proportions of sites under selection, whereas Rosaceae showed the lowest proportion. Our results are consistent with previous estimates of low levels of adaptive evolution in eleven angiosperm species (Gossmann et al. 2010); and with estimates in *Pinus contorta* and *Pinus taeda*, in which the proportions of sites fixed by positive selection were 13–52% and 22–37%, respectively (Eckert et al. 2013; Hodgins et al. 2016).

There is a rich literature reporting the influence of positive selection in Pinaceae species, which in absence of reports in other gymnosperm families, has frequently been used as an example of all gymnosperm taxa. The recent sequencing of some Pinaceae genomes have paved the way to test the influence of positive selection at the genome level (De La Torre et al. 2015a; Hodgins et al. 2016), which have confirmed what was previously found at the genetic (small number of genes) and population levels (Eckert et al. 2010; Buschiazzo et al. 2012; Pavy et al. 2012; De La Torre et al. 2015b). Furthermore, studies on quantitative trait variation suggest natural selection is highly efficient in producing a relatively fast evolutionary response in contrast to the slow evolutionary rate (Savolainen and Pyhäjärvi 2007). Examples of this include the evolution of cold adapted genotypes in several long-lived tree species in the northern hemisphere after the Last Glaciation (Mimura and Aitken 2007; Wachowiak et al. 2009; Holliday et al. 2010; Kujala and Savolainen 2012; De La Torre et al. 2014b). In addition, because of their long generation times and low mutation rates, gymnosperm genomes may retain the consequences of demographic events for a long time, suggesting that even species that have contracted their ranges and currently have small distributions may resemble those with long-term large effective population sizes.

Although our results suggest an important role of positive selection in the evolution of gymnosperm taxa, we do not believe that Pinaceae is a good representative of other gymnosperm taxa. *Pinus* and *Picea* have indeed significantly higher number of sites under diversifying selection than other gymnosperm taxa. Recent evolutionary radiations may have resulted in lower dS and higher ω in comparisons within these taxonomic families (Palme et al. 2009).

We expect our estimation of the proportion of sites under selection to be conservative because of the use of single copy genes. Single copy genes are usually ubiquitously expressed genes that encode for basic cellular functions that are preserved across taxa. For this reason, they are likely to be under stronger purifying selection due to functional and structural constraints than paralogous genes in gene families in both angiosperms (De Smet et al. 2013; Li et al. 2016) and gymnosperms (De La Torre et al. 2015a). These genes also seem to experience less frequent positive selection than paralogous genes, which tend to be more narrowly expressed (Larracuente et al. 2007). For this reason, we expect a lower proportion of conserved sites and a higher proportion of selected sites in paralogous genes (which account for the majority of genes in plant genomes). In fact, it has been suggested that many of the genes under diversifying selection in conifers belong to large multi-copy gene families such as Leucine Rich Repeats, Cytochrome P450, among others (Pavy et al. 2012; Neale et al. 2014; De La Torre et al. 2015a). In addition, some sites under positive selection may go undetected because structural constraints may induce purifying selection to push ω to values lower than 1 in highly conserved genes (Echave et al. 2016). However, the introduction of paralogous genes in a phylogeny has confounding effects for both the phylogenetic reconstruction and also for the estimation of sequence divergence among taxa, therefore, we believe than the use of single-copy genes is probably the best method for comparisons among diverse taxa. In our efforts to identify a highly confident set of very well aligned orthologs across a broad taxonomic sampling, the number of genes used in this study was significantly reduced raising concerns about the representativeness of these genes in a genome-wide context. However, comparisons with previous studies, suggest our estimates of rates of divergence are quite comparable with estimates of rates with much larger data sets in specific taxa (Gossmann et al. 2010; Eckert et al. 2013; Hodgins et al. 2016).

Stronger and More Efficient Purifying Selection in Gymnosperms

When estimating the distribution of selection coefficients, we found that the majority of sites in both angiosperms and gymnosperms were highly conserved, suggesting an important role of purifying selection in the evolution of plant genomes. Previous studies in angiosperm species have suggested that the largest proportion of mutations is strongly deleterious, and that the number of slightly and mildly deleterious mutations seems to be relatively conserved among species (Gossmann et al. 2010). In contrast with these studies, our study has found differences in the proportion of sites

under mildly and strong purifying selection. Gymnosperm families had a greater proportion of sites under strong purifying selection ($S < -15$) than angiosperms ($t = -3.57$, $P < 0.05$); whereas angiosperm families had a greater proportion of neutral and mildly deleterious ($S \approx 0$ and $S > -15$) sites than gymnosperms ($t = 4.119$, $P < 0.05$) (fig. 2).

This could be explained by the expectation that species with larger Ne generally experience stronger purifying selection (Slotte et al. 2010; Gossmann et al. 2010; Hough et al. 2013; Williamson et al. 2014). Slightly deleterious mutations will be less likely to segregate at higher frequencies and fix in large populations (Akashi et al. 2012). Such mutations are also selected against by the partial low degree of selfing and high early inbreeding depression in conifers (Williams and Savolainen 1996; Remington and O'Malley 2000). Larger population sizes may also explain the lower number of neutral sites in gymnosperms than in angiosperms found in this study, as it has been shown that the number of effectively neutral mutations is negatively correlated with Ne in many species (Piganeau and Eyre-Walker 2009; Gossmann et al. 2010).

Life Form Does Not Explain Differences in Molecular Evolution among Woody Angiosperms and Woody Gymnosperms, or within Gymnosperms

In flowering plants, it has been recently suggested that life form (also measured as plant height) is correlated with rates of molecular evolution, with shrubs and trees showing lower rates of molecular evolution than herbaceous plants (Gaut et al. 1992; Smith and Donoghue 2008; Lanfear et al. 2013). Similarly, in our study, we found significantly lower μ, on average, in angiosperm shrubs/trees compared to angiosperm herbs (see supplementary table S6, Supplementary Material Online, and, fig. 2). Both dS and dN showed large variation among angiosperm herbs, and differences in dS and dN among angiosperm shrubs/trees and herbs were not significant (see supplementary table S6, Supplementary Material Online). When incorporating gymnosperms into the analysis, life form on its own does not explain the differences in rates of molecular evolution among seed plants, as we observe significant differences in dN, dS, ω, and μ also among woody angiosperms and woody gymnosperms (fig. 2).

Our analyses showed that Gnetophytes, the only gymnosperm division that contains non-woody genera (*Welwitschia* and *Ephedra*), shows higher dN and μ than what we observe in other gymnosperm families (table 1). Therefore, differences in life form may drive differences in the rate of molecular evolution found in Gnetophytes. However, we still observe these differences when only including the woody genus *Gnetum* (and excluding *Ephedra* and *Welwitschia*), suggesting other factors may influence the differences in rate of sequence evolution (table 1). In fact, Gnetophytes are different from the rest of gymnosperms in a number of features that include but are not restricted to their morphology, ecology, and the presence of angiosperm-like characteristics such as special water-conducting wood vessels, and reproductive structures organized in compound strobili (Doyle and Donoghue 1986; Friedman 1998). For these reasons, the position of Gnetophytes within gymnosperms is contentious (Doyle 1998; Braukmann et al. 2009; Cibrian-Jaramillo et al. 2010; Ran et al. 2010; Xi et al. 2013). In any case, the lack of close, extant relatives to Gnetophytes suggests that this problem might never really be solved satisfactorily, regardless of the amount of sequence data we have access to. Our results suggest that estimations of rates of sequence divergence among a few species from very diverged Gnetophyte taxa (such as *Gnetum*, *Welwitschia* and *Ephedra*) may lead to inaccurate estimates of dN and dS. For this reason, we decided to exclude the more distant *Welwitschia* and *Ephedra* from all correlations with life history traits.

Generation Times and Their Effect on Low Substitution Rates

The differences in rates of molecular evolution between angiosperms with different life forms have been presumed to reflect differences in generation times between herbaceous and woody plants (Gaut et al. 1992; Smith and Donoghue 2008; Lanfear et al. 2013). The generation time hypothesis suggests that species with shorter generation times (e.g. herbaceous plants) accumulate more replication errors per unit time because they copy their genomes more often, which results in higher mutation rates (Li et al. 1996). Although generation time may not be a good indicator of the overall rate of genome replication because the number of mitotic cell divisions can vary substantially between generations and among plant species; it remains strongly associated with the long-term rates of meiosis in plants (Petit and Hampe 2006; Lanfear et al. 2013). Therefore, if a significant proportion of heritable mutations occur during meiosis, plants with longer generation times (e.g. gymnosperms) would have lower mutation and substitution rates per unit of time (Lanfear et al. 2013). Alternatively, it is been proposed that differences in the rates of mitosis (mitotic cell divisions that occur in the apical meristem before gametogenesis) can account for the observed differences in rates of molecular evolution among plants of different height (Lanfear et al. 2013). Lower absolute growth rates in long-lived woody perennials would translate in fewer cell divisions and less opportunities for DNA replications errors than short-lived plants (Bromham et al. 2015). This would result in a lower mutation rate per unit of time in long-lived species. Our results are consistent with this expectation, as we found a strong negative correlation between generation times and μ ($r = -0.78$, $P < 0.001$) (table 4). Long generation times, large effective population sizes and low recombination rates may also help explain the low synonymous polymorphism in gymnosperms (Savolainen and Pyhäjärvi 2007; Jaramillo-Correa et al. 2010).

In plants, in the absence of a segregated germline, gametes arise from the apical meristem late in development following periods of vegetative growth. Because of this, somatic mutations acquired during vegetative growth can be transmitted to the next generation (Watson et al. 2016). Longer growth periods are thought to result in a larger number of cell divisions, increasing the opportunities for mutations to occur per generation (Schultz and Scofield 2009). Our results suggest,

than on average, for the species under study, long-lived gymnosperms accumulate four times more mutations per generation than short-lived angiosperms (assuming an average generation time of 20 years in gymnosperms, and an average μ equal to 1.57×10^{-8} in gymnosperms, and 3.92×10^{-9} in angiosperms). This is coincident with previous reports of increased number of somatic mutations in gymnosperms species (Cloutier et al. 2003; O'Connell and Ritland 2004) and other woody perennials (Ally et al. 2010; Bobiwash et al. 2013). However, because the per-generation increase in mutation rate in woody perennials may be less than predicted from their differences in generation time (Petit and Hampe 2006), we would expect our reported values to be an over-estimation of the actual difference between plant seed clades. Differences in generation times would predict a mutation rate >100 times larger in mangrove trees than in annuals, when estimates were only 25 times larger (Klekowski and Godfrey 1989). Alternatively, there may be selection to reduce DNA replication-dependent errors through minimizing the number of cell divisions required during development, as recently suggested in *Arabidopsis* (Watson et al. 2016).

Larger Genomes Are Correlated with Larger Omegas and Slower Rates of Sequence Divergence

In our study, we also explored the relationship between genome size and rates of molecular evolution. Gymnosperm genomes are characterized by their enormous genome sizes (20–40 Gb), unique genome silencing mechanisms and low unequal recombination (Leitch and Leitch 2012, Nystedt et al. 2013). While angiosperm genomes are highly dynamic and have efficient mechanisms to counteract the increase in DNA amount stemming from WGDs or transposable elements (e.g. replication or recombination-based errors generating indels, unequal recombination between sister chromosomes, Grovel and Wendel 2010), gymnosperm genomes seem to be less dynamic and may have evolved their own epigenetic mechanisms to silence retrotransposons (Leitch and Leitch 2012). Our results indicate that genome size is positively correlated with ω ($r = 0.56$, $P < 0.001$ at the species level, and $r = 0.52$, $P < 0.05$ at the family level). In addition, genome size is negatively correlated with μ ($r = -0.67$, $P < 0.01$), suggesting species with large genomes such as gymnosperms have slower rates of molecular evolution than species with small genomes, confirming previous studies on angiosperms species (Bromham et al. 2015). A possible explanation for this is that if large genome sizes correlate with larger cells and a reduction in growth, then plants with larger genomes might have fewer replications and therefore less accumulated mutations per unit of time (Bromham et al. 2015).

Our study aims to contribute to the unexplored field of molecular evolution in seed plants by investigating the differences in rates of molecular evolution among gymnosperms and angiosperms using a phylogenetic framework. The recent genomic and transcriptomic resources in gymnosperm species opened a window to understand the evolution of this ancient and important plant clade. However, gymnosperm resources are still limited in comparison to their sister clade of flowering plants. We hope that with the development of new genomic resources, studies in molecular evolution will include a broader taxonomic sampling that provides us with a more complete understanding of the evolution of seed plants.

Materials and Methods

Generation of Phylogenetic Markers

To develop our set of phylogenetic markers, we performed deep sequencing and assembly of whole transcriptomes from two conifer species (*Pinus pinaster* and *Pinus sylvestris*) (Li Z et al., (in review), transcriptome data can be found at: http://bioinformatics.psb.ugent.be/supplementary_data/zheli/phylo/). We integrated this data with whole-transcriptome data from another 29 gymnosperms, 34 angiosperms, and 1 outgroup (*Physcomitrella patens*) obtained from public databases PlantGDB, oneKP, TreeGenes and PLAZA v3.0 (Proost et al. 2015). We used OrthoMCL (Li et al. 2003) to build orthologous gene families across the species. To reduce the number of single-copy genes, we selected only the gene families that were conserved and had low copy number. Later we used hidden Markov probabilistic models implemented in HMMER to build an HMM profile for each gene family based on multiple sequence alignment (Eddy 2009). These HMM profiles were then used to assign additional proteins to the existing gene families. Then, we used HMMSEARCH to find the best protein hit of an HMM profile in each species and selected the markers with only reciprocal best hits for phylogenomic analysis. In order to increase the spectrum of the phylogenetic markers, we only selected single-copy markers that were present in a majority of species.

Multiple sequence alignments were carried out for each gene family based on amino acid sequences using Muscle v3.8.31 (Edgar et al. 2004). Trimal v1.4 (Capella-Gutierrez et al. 2009) was used to back translate the amino acid alignments into coding sequence alignments, and to remove low quality alignment regions and spurious sequences. We defined a spurious sequence as a sequence having less than 70% of the total alignment positions that were present in 75% of 66 species. Genes that lost their sequences in more than 10% of the species, after removing spurious sequences, were not used for further analyses. This way, we obtained 42 single-copy markers across 66 seed plant species that were used for evolutionary and phylogenetic analyses. Annotations of all genes are reported in supplementary table S1, Supplementary Material Online.

Rates of Sequence Divergence

Multiple sequence alignments for each of the 42 single-copy genes were divided in two groups containing either angiosperms (34 species) or gymnosperms (31 species), to allow further comparisons between major plant taxa. Alignment gaps and low quality regions were manually removed using Jalview version 2.8.1 (Waterhouse et al. 2009). Also, species containing more than 30% gaps or ambiguous sites were not kept for further analyses.

Pairwise estimates (runmode= −2, seqtype = 1, model = 0, NSsites = 0) of the number of synonymous substitutions per site (dS), nonsynonymous substitutions per site (dN) and nonsynonymous/synonymous rate ratio (also called "omega" and denoted by ω or dN/dS) were calculated between species in terminal branches for each of the taxonomic families of the angiosperm and gymnosperm phylogeny using the maximum likelihood method of Goldman and Yang (1994) in the codeml program from the PAML package (version 4.8, Yang 2007). The analysis was repeated twice for each of the 42 genes. For each sequence pair, only the results with the higher lnL (log likelihood) were retained. We discarded genes with dS values lower than 0.01, as these values may result in inaccurate estimates of ω; and also genes with dS or dN > 3 which suggest saturation of substitutions. Abnormally high omega rates ($\omega > 10$) were also discarded (Villanueva-Cañas et al. 2013). To evaluate the differences in synonymous (dS) and non-synonymous substitutions (dN) and their ratio (ω) between angiosperms and gymnosperms, we used a sign test in the R package PASWR (Ugarte et al. 2008) to test the null hypothesis that the median for the differences between the pairs equals zero. We tested the alternative hypotheses of a higher ω in gymnosperms than in angiosperms; and a lower dN and dS in gymnosperms than in angiosperm species for each of the genes. Significance values were calculated for all genes for each of the variables measured using a one-sided exact binomial test in R. In addition to the pairwise estimates, we also calculated dS, dN, and ω for each terminal branch of the phylogeny using the free-ratios branch model (Model 1) implemented in codeml. Differences in dS, dN, and ω between plant lineages were tested using a Welch two-sample *t*-test in R.

Absolute rates of silent-site divergence were estimated for each taxonomic family, assuming the "molecular clock" hypothesis (Kimura 1968, Kimura and Ohta 1971), in which the change at the molecular level occurs constantly through time across evolutionary lineages. Synonymous substitutions (dS) for each gene were averaged for all pairs of species in each of the six taxonomic families within gymnosperms (Pinaceae, Cuppresaceae, Taxaceae, Podocarpaceae, Cycadaceae, and Gnetophytes); and in eight taxonomic families in angiosperms (Brassicaceae, Malvaceae, Euphorbiaceae, Fabaceae, Cucurbitaceae, Rosaceae, Solanaceae, and Poaceae). Poaceae was further divided in two groups (each containing a different ancestral branch) as Poaceae I and Poaceae II (see supplementary figure S1, Supplementary Material Online). Absolute rates of silent-site divergence were calculated for each of the genes using the formula $\mu = dS/2T$ where μ is the synonymous divergence rate per site per year, dS is the mean of synonymous substitutions per site, and *T* is the time of divergence between two species in years (Gaut et al. 2011). Overall rates of silent-site divergence for each family were calculated using the same formula, where dS is the mean of synonymous substitutions per synonymous site across all 42 genes, and *T* is the time of divergence in years. Estimations of divergence times (*T*) were based on fossil records and previously published divergence times (see supplementary table S2, Supplementary Material Online). Whenever possible, we selected the median time of divergence between the species in each taxonomic group, making sure the estimates are consistent with previous estimates for each taxa.

Estimation of the Variation in Selective Pressures (ω) among Branches

To understand the variation in selective pressures among the gymnosperm and angiosperm plant lineages, we formulated the following hypotheses: H0: homogeneous selective pressure in both lineages ($\omega 0 = \omega G = \omega A$); H1: selective pressure in the angiosperm clade ($\omega 0 = \omega G$, ωA); H2: selective pressure in the gymnosperm clade ($\omega 0 = \omega A$, ωG); H3: long-term shift in selective pressure resulting in both clades having different omegas ($\omega 0$, ωG, ωA); and H4: all terminal branches in both clades have different omegas. All estimated omegas for hypotheses H1 to H3 were calculated on the stem branches of the angiosperm (ωA) and the gymnosperm (ωG) clades according to the phylogenetic tree in figure 1. The estimated omega in unselected branches (background) is represented by $\omega 0$.

We tested the different hypotheses using branch models (Yang 1998; Yang and Nielsen 1998) in PAML 4.8 (Yang 1997; Yang 2007) that allow ω to vary among branches in the tree. The one-ratio model (runmode = 0, seqtype = 1, model = 0), which assumes one ω ratio for all branches, was first used to estimate the branch lengths from a concatenated sequence alignment of all 42 single-copy genes (31,968 nucleotide sites), and a phylogenetic tree (see below for details on tree building). This model was also used to test the null hypothesis (H0). The resulting branch lengths were then used as initial values to run the program three times with model 2 (model = 2, NSSites = 0) to test hypotheses H1, H2, and H3. This model allows different branch groups to have different omegas, according to the different branch labels assigned. Omega in selected branches (foreground) was then compared with omega in unselected branches (background). Finally, we tested hypothesis 4 by fitting model 1 (model = 1, NSsites = 0), also called the free-ratios model. A likelihood ratio test was used to assess deviations from the null model (both angiosperms and gymnosperms had the same ω) for the gene set. Corrections for multiple testing were done using the Benjamini and Hochberg method (1995) with a false discovery rate threshold of 0.05.

Phylogenetic trees used in all branch models were constructed applying a posteriori partitioning, inferred from Bayesian searches of the substitution rate matrix using a mixture model approach, to the concatenated alignment by carrying out BayesPhylogenies (Pagel and Meade 2004) with 10 million generations of Markov chain Monte Carlo analysis. We finally obtained 12 partitions by the Perl script provided in Xi et al. (2012). The phylogeny of seed plants was inferred by RAxML (8.2) with the concatenated alignment and the partitions, and edited for publication with FigTree v.1.4.2 (http://tree.bio.ed.ac.uk).

Estimation of Substitution Parameters and Inference of the Proportion of Sites under Selection for Each Taxonomic Family

We tested different site models in codeml to estimate substitution parameters; infer omega distributions and assess the proportion of conserved ($\omega < 1$), neutral ($\omega = 1$), and selected ($\omega > 1$) nucleotide sites in each taxonomic family. We used the following parameters: runmode = 0 (user tree); model = 0; NSsites = 0,1,2,3,5,7,8,10; kappa (transition/transversion rate ratio) to be estimated from the data. Models M1a (nearly neutral), M2a (positive selection), M3 (discrete), M5 (gamma), M7 (beta), M8 (beta & ω), and M10 (beta & gamma + 1) detailed in Yang et al. (2000), were tested for each taxonomic family independently (all terminal branches in each clade), using a concatenated alignment of 42 genes. By testing all species that share a most recent common ancestor, we make sure that species had a similar amount of time to accrue genetic differences, and that therefore any difference in the amount of genetic change accumulated through time represents a difference in the evolutionary rate. For this analysis, three species were removed due to problems in the concatenated alignment (*Beta vulgaris*, *Cryptomerica japonica* and *Welwitschia mirabilis*). Input phylogenetic trees were constructed for each taxonomic family using the GTR + GAMMA model for maximum likelihood of phylogenetic inference in RAxML (Stamatakis 2014). We used the Akaike Information Criteria (AIC) to rank models and estimate the most probable distribution of omega for each taxonomic family. Runs were repeated to ensure the convergence of results.

In addition, we also estimated the number of sites under positive selection using two different analyses. In the first one, we compared models M1a–M2a and M7–M8 using likelihood ratio tests. In the second one, we used the Bayes Empirical Bayes (BEB) (Yang et al. 2005), implemented in site models M2a and M8, to calculate the posterior probabilities for site classes and to identify sites under positive selection when the likelihood ratio test was significant. We also used the results of the Naïve Empirical Bayes (NEB) but only when BEB results were not available (site model M3).

Distribution of Selection Coefficients

While omega (ω), as formulated in the codon models, is a property of a particular site or collection of sites in the DNA sequence; the selective coefficient (S) is a property of a particular allele or mutation (Nielsen and Yang 2003). Therefore, it is possible to infer the distribution of S from the distribution of ω if some assumptions are made in the mutation model. We assume no interference in the fixation process of multiple mutations at different sites, and further that there are no more than two alleles segregating at the same nucleotide site. We also assume that all non-synonymous mutations at the same amino acid site have equal selection coefficients S. If these assumptions are not met, we may under-estimate the selection coefficients. We use the inverse mapping from ω to S to obtain the distribution of the selective coefficient of new mutations from the distribution of ω. Model 5 (Yang et al. 2000) assumes that omega (ω) is gamma distributed among codon sites with parameters α and β, and therefore the probability density function for S is obtained by:

$$f(S) = (\beta e^{S} S/h(S))^{\alpha} e^{-\beta e S/h(S)} (h(S) - S)/Sh(S)\Gamma(\alpha), \quad -\infty < S < \infty, \tag{1}$$

where $h(S) = (e^{s} - 1)$ (Nielsen and Yang 2003). We assumed a constant population size among lineages within each taxonomic family.

Correlations with Life-History Traits

In order to explain the variation in substitution rates and ω between major plant clades, we evaluated the associations of these with several biological and life history traits. Generation time (time to reach maturity), genome size (amount of DNA contained in a haploid nucleus measured in picograms), and species richness (number of extant species in each taxonomic group or family) were obtained from public databases. Because phylogenetic non-independence of data points can lead to an inflation of the significance of observed relationships between substitution rates and traits (Lanfear et al. 2010), we used the phylogenetically independent contrasts (PIC) method (Felsenstein 1985) implemented in the R package APE to estimate the correlations between variables (Paradis et al. 2004). This analysis was done twice, either using all species in the phylogeny or using only taxonomic families. For the latter analysis, a phylogenetic tree with average branch lengths for each family, and family-average estimates of all variables were used as input files in APE. Phylogenetic trees were imported in Newick format and then saved with FigTree v1.4.2 (Rambaut 2006, http://tree.bio.ed.ac.uk) before PIC analysis.

We also tested for significant differences in dN, dS, ω, and μ among groups with different life forms using a Kruskal-Wallis one-way analysis of variance (rank sum test). The analysis was repeated three times taking in consideration different groups. In the first one there were six groups as follows: angiosperms dicots herbs, angiosperms dicots shrubs, angiosperms dicots trees, angiosperms monocots herbs, gymnosperms and Gnetophytes. In the second one, there were four groups: angiosperms herbs, angiosperms trees/shrubs, gymnosperms and Gnetophytes. And in the last test three groups were considered (angiosperms herbs, angiosperms shrubs/trees, and gymnosperms) while Gnetophytes were excluded.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank Joseph Bielawski for valuable suggestions on data analyses, and Outi Savolainen and Jeff Ross-Ibarra for comments on previous versions of the manuscript. This work was supported by the European 7th Framework Programme under the ProCoGen (Promoting Conifer Genomic Resources)

project. YVdP would like to acknowledge the Multidisciplinary Research Partnership "Bioinformatics: from nucleotides to networks" Project (no. 01MR0310W) of Ghent University and the European Union Seventh Framework Programme (FP7/2007-2013) under European Research Council Advanced Grant Agreement 322739–DOUBLEUP.

References

Akashi H, Osada N, Ohta T. 2012. Weak selection and protein evolution. *Genetics*. 192:15–31.

Ally D, Ritland K, Otto SP. 2010. Aging in a long-lived clonal tree. *Plos Biol*. 8(8):e1000454.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate-a practical and powerful approach to multiple testing. *J R Stat Soc Series B*. 57(1):289–300.

Bierne N, Eyre-Walker A. 2004. The genomic rate of adaptive amino acid substitutions in *Drosophila*. *Mol Biol Evol*. 21:1350–1360.

Bobiwash K, Schultz ST, Schoen DJ. 2013. Somatic deleterious mutation rate in a woody plant: estimation from phenotypic data. *Heredity*. 111(4):338–344.

Bouille M, Bousquet J. 2005. Trans-species shared polymorphisms at orthologous nuclear gene loci among distant species in the conifer Picea (Pinaceae): implications for the long-term maintenance of genetic diversity in trees. *Am J Bot*. 92:63–73.

Braukmann TW, Kuzmina M, Stefanovic S. 2009. Loss of all plastid ndh genes in Gnetales and conifers: extent and evolutionary significance for the seed plant phylogeny. *Curr Gene*. 55:323–337.

Bromham L, Hua X, Lanfear R, Cowman PF. 2015. Exploring the relationships between mutation rates, life history, genome size, environment, and species richness in flowering plants. *Am Nat*. 185(4):507–524.

Brown GR, Gill GP, Kuntz RJ, Langley CH, Neale DB. 2004. Nucleotide diversity and linkage disequilibrium in loblolly pine. *PNAS*. 101(42):15255–15260.

Buschiazzo E, Ritland C, Bohlmann J, Ritland K. 2012. Slow but not low: genomic comparisons reveal slower evolutionary rate and higher dN/dS in conifers compared to angiosperms. *BMC Evol Biol*. 12:8.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. TrimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. 25:1972–1973.

Carneiro M, Albert FW, Melo-Ferreira J, Glatier N, Gayral P, Blanco-Aguiar JA, Villafuerte R, Nachman MW, Ferrand N. 2012. Evidence for widespread positive and purifying selection across the European rabbit (*Oryctolagus cuniculus*) genome. *Mol Biol Evol*. 29(7):1837–1849.

Chen J, Uebbing S, Gyllenstrand N, Lagercrantz U, Lascoux M, Kallman T. 2012. Sequencing of the needle transcriptome from Norway spruce (*Picea abies* Karst L.) reveals lower substitution rates, but similar selective constraints in gymnosperms and angiosperms. *BMC Genom*. 13:589.

Cibrian-Jaramillo A, De La Torre-Barcena JE, Lee EK, Katari MS, Little DP, Stevenson DW, Martiensen R, Coruzzi GM, DeSalle R. 2010. Using phylogenomic patterns and gene ontology to identify proteins of importance in plant evolution. *Gen Biol Evol*. 2:225–239.

Cloutier D, Rioux D, Beaulieu J, Schoen DJ. 2003. Somatic stability of microsatellite loci in Eastern white pine, *Pinus strobus* L. *Heredity*. 90(3):247–252.

Crisp MD, Cook LG. 2011. Cenozoic extinctions account for the low diversity of extant gymnosperms compared with angiosperms. *New Phytol*. 192:997–1009.

De La Torre AR, Birol I, Bousquet J, Ingvarsson PK, Jansson S, Jones SJM, Keeling CI, MacKay J, Nilsson O, Ritland K, et al. 2014a. Insights into Conifer Giga-genomes. *Plant Physiol*. 166:1–9.

De La Torre AR, Roberts DR, Aitken SN. 2014b. Genome-wide admixture and ecological niche modeling reveal the maintenance of species boundaries despite long history of interspecific gene flow. *Mol Ecol*. 23(8):2046–2059.

De La Torre AR, Ling YC, Van de Peer Y, Ingvarsson PK. 2015a. Genome-wide analysis reveals diverged patterns of codon bias, gene expression, and rates of sequence evolution in *Picea* gene families. *Gen Biol Evol*. 7(4):1002–1015.

De La Torre AR, Ingvarsson PK, Aitken SN. 2015b. Genetic architecture and genomic patterns of gene flow between hybridizing species of *Picea*. *Heredity*. 00:1–12. doi: 10.1038/hdy.2015.19.

De Smet R, Adams KL, Vandepoele K, Van Montagu MCE, Maere S, Van de Peer Y. 2013. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *PNAS*. 110(8):2898–2903.

Doyle JA. 1998. Phylogeny of vascular plants. *Annu Rev Ecol Syst*. 20:567–599.

Doyle JA, Donoghue MJ 1986. Seed plant phylogeny and the origin of angiosperms: An experimental cladistic approach. *M J Bot Rev*. 52:321.

Echave J, Spielman SJ, Wike CO. 2016. Causes of evolutionary rate variation among protein sites. *Nat Rev Genet*. 17:109–121.

Eckert AJ, Bower AD, Gonzalez-Martinez SC, Wegrzyn JL, Coop G, Neale DB. 2010. Back to nature: ecological genomics of loblolly pine (*Pinus taeda*, Pinaceae). *Mol Ecol*. 19(17):3789–3805.

Eckert AJ, Wegrzyn JL, Liechty JD, Lee JM, Cumbie P, Davis JM, Goldfarb B, Loopstra CA, Palle SR, Quesada T, et al. 2013. The evolutionary genetics of the genes underlying phenotypic associations for Loblolly pine (*Pinus taeda*, Pinaceae). *Genetics*. 195:1353–1372.

Eddy SR. 2009. A new generation of homology search tools based on probabilistic inference. *Genome Inform*. 23(1):205–211.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nuc Acids Res*. 32(5):1792–1797.

Felsenstein J. 1985. Phylogenies and the comparative method. *Am Nat*. 125:1–15.

Gaut B, Muse SV, Clark WD, Clegg MT. 1992. Relative rates of nucleotide substitution at the *rbc*L locus of monocotyledonous plants. *J Mol Evol*. 35:292–303.

Gaut B, Yang L, Takuno S, Eguiarte LE. 2011. The patterns and causes of variation in plant nucleotide substitution rates. *Annu Rev Ecol Evol Syst*. 42:245–266.

Goldman N, Yang Z. 1994. A Codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol*. 11(5):725–736.

Gossmann TI, Song BH, Windsor AJ, Mitchell-Olds T, Dixon CJ, Kapralov MV, Filatov DA, Eyre-Walker A. 2010. Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Mol Biol Evol*. 27(8):1822–1832.

Grover C, Wendel JF. 2010. Recent insights into mechanisms of genome size change in plants. *J Botany*. 2010:article ID 382732, 1–8.

Hodgins KA, Yeaman S, Nurkowski K, Rieseberg LH, Aitken SN. 2016. Expression divergence is correlated with sequence evolution but not positive selection in conifers. *Mol Biol Evol*. 33(6):1502–1516.

Holliday JA, Yuen M, Ritland K, Aitken SN. 2010. Postglacial history of a widespread conifer produces inverse clines in selective neutrality tests. *Mol Ecol*. 19(18):3857–3864.

Hough J, Williamson RJ, Wright SI. 2013. Patterns of selection in plant genomes. *Annu Rev Ecol Evol Syst*. 44:31–49.

Ingvarsson PK. 2010. Natural selection on synonymous and nonsynonymous mutations shapes patterns of polymorphism in *Populus tremula*. *Mol Biol Evol*. 27:650–660.

Jaramillo-Correa JP, Verdu M, Gonzalez-Martinez SC. 2010. The contribution of recombination to heterozygosity differs among plant evolutionary lineages and life-forms. *BMC Evol Biol*. 10:22.

Kimura M. 1968. Evolutionary rate at the molecular level. *Nature*. 217:624–626.

Kimura M, Ohta T. 1971. Protein polymorphism as a phase of molecular evolution. *Nature*. 229:467–469.

Klekowski EJ, Godfrey PJ. 1989. Aging and mutation in plants. *Nature*. 340(6232):389–391.

Kujala ST, Savolainen O. 2012. Sequence variation patterns along a latitudinal cline in Scots pine (*Pinus sylvestris*): signs of clinal adaptation? *Tree Genet Genomes*. 8:1451–1467.

Lanfear R, Welch JJ, Bromham L. 2010. Watching the clock: studying variation in rates of molecular evolution between species. *Tree*. 25:495–503.
Lanfear R, Ho SYW, Davies TJ, Moles AT, Aarssen L, Swenson NG, Warman L, Zanne AE, Allen AP. 2013. Taller plants have lower rates of molecular evolution. *Nature Commun*. 4:1879.
Lanfear R, Kokko H, Eyre-Walker A. 2014. Population size and the rate of evolution. *Tree*. 29(1):33–41.
Larracuente AM, Sackton TB, Greenberg AJ, Wong A, Singh ND, Sturgill D, Zhang Y, Oliver B, Clark AG. 2007. Evolution of protein-coding genes in Drosophila. *Trends Genet*. 24(3):114–123.
Leitch AR, Leitch IJ. 2012. Ecological and genetic factors linked to contrasting genome dynamics in seed plants. *New Phyto*. 194:629–646.
Leslie AB, Beaulieu JM, Rai HS, Crane PR, Donoghue MJ, Mathews S. 2012. Hemisphere-scale differences in conifer evolutionary dynamics. *PNAS*. 109(40):16217–16221.
Li WH, Ellsworth DL, Krushkal J, Chang BHJ, Hewett-Emmett D. 1996. Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. *Mol Phylo Evol*. 5(1):182–187.
Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*. 13:2178–2189.
Li Z, Defoort J, Tasdighian S, Maere S, Van de Peer Y, De Smet R. 2016. Gene duplicability of core genes is highly consistent across all angiosperms. *Plant Cell*. 28(2):326–344.
Lynch M. 2010. Evolution of the mutation rate. *Trends Genet*. 26(8):345–352.
Magallón S, Hilu KW, Quandt D. 2013. Land plant evolutionary timeline: gene effects are secondary to fossil constraints in relaxed clock estimation of age and substitution rates. *Am J Bot*. 100:556–573.
Mimura M, Aitken SN. 2007. Adaptive gradients and isolation-by-distance with postglacial migration in *Picea sitchensis*. *Heredity*. 99:224–232.
Neale DB, Kremer A. 2011. Forest tree genomics: growing resources and applications. *Nature Rev Genet*. 12:111.
Neale DB, Wegrzyn JL, Stevens KA, Zimin AV, Puiu D, Crepeau MW, Cardeno C, Koriabine M, Holtz-Morris AE, Liechty JD, et al. 2014. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biol*. 15:R59.
Nielsen R, Yang Z. 2003. Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. *Mol Biol Evol*. 20(8):1231–1239.
Nielsen R. 2005. Molecular signatures of natural selection. *Annu Rev Gen*. 39:197–218.
Nystedt B, et al. 2013. The Norway spruce genome sequence and conifer genome evolution. *Nature*. 497:579–584.
O'Connell LM, Ritland K. 2004. Somatic mutations at microsatellite loci in western Redcedar (*Thuja plicata*: Cupressaceae). *J Hered*. 95(2):172–176.
Ohta T. 1992. The nearly neutral theory of molecular evolution. *Annu Rev Ecol Syst*. 23:263–286.
Pagel M, Meade A. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst Biol*. 53:571–581.
Palme AE, Pyhäjärvi T, Wachowiak W, Savolainen O. 2009. Selection on nuclear genes in a *Pinus* phylogeny. *Mol Biol Evol*. 26(4):893–905.
Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*. 20(2):289–290.
Pavy N, Pelgas B, Laroche J, Rigault P, Isabel N, Bousquet J. 2012. A spruce gene map infers ancient plant genome reshuffling and subsequent slow evolution in the gymnosperm lineage leading to extant conifers. *BMC Biol*. 10:84.
Petit RJ, Hampe A. 2006. Some evolutionary consequences of being a tree. *Annu Rev Ecol Evol Syst*. 37:187–214.
Phifer-Rixey M, Bonhomme F, Boursot P, Churchill GA, Pialek J, Tucker PK, Nachman MW. 2012. Adaptive evolution and effective population size in wild house mice. *Mol Biol Evol*. 29(10):2949–2955.
Piganeau G, Eyre-Walker A. 2009. Evidence for variation in the effective population size of animal mitochondrial DNA. *PLoS One*. 4:e4396.
Proost S, Van Bel M, Vaneechoutte D, Van de Peer Y, Inze D, Mueller-Roeber B, Vanderpoele K. 2015. PLAZA3.0: an access point for plant comparative genomics. *Nucl Acid Res*. 43(D1):D974–D981.
Ran JH, Gao H, Wang XQ. 2010. Fast evolution of the retroprocessed mitochondrial rps3 gene in Conifer II and further evidence for the phylogeny of gymnosperms. *Mol Phylogenet Evol*. 54:136–149.
Remington DL, O'Malley DM. 2000. Whole-genome characterization of embryonic stage inbreeding depression in a selfed Loblolly Pine family. *Genetics*. 155(1):337–348.
Savolainen O, Pyhäjärvi T. 2007. Genomic diversity in forest trees. *Curr Opin Plant Biol*. 10:162–167.
Schultz ST, Scofield DG. 2009. Mutation accumulation in real branches: fitness assays for genomic deleterious mutation rate and effect in large-statured plants. *Am Nat*. 174:163–175.
Slotte T, Foxe JP, Hazzouri KM, Wright SI. 2010. Genome-wide evidence for efficient positive and purifying selection in *Capsella grandiflora*, a plant species with large effective population size. *Mol Biol Evol*. 27(8):1813–1821.
Smith SA, Donoghue MJ. 2008. Rates of molecular evolution are linked to life history traits in flowering plants. *Science*. 322(5898):86–89.
Stamatakis A. 2014. RAxML Version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 30(9):1312–1313.
Strasburg JL, Kane NC, Raduski AR, Bonin A, Michelmore R, Rieseberg LH. 2011. Effective population size is positively correlated with levels of adaptive divergence among annual sunflowers. *Mol Biol Evol*. 28(5):1569–1580.
Syring J, Farrell K, Businsky R, Cronn R, Liston A. 2007. Widespread genealogical nonmonophyly in species of *Pinus* subgenus *Strobus*. *Syst Biol*. 56:163–181.
Ugarte MD, Militino AF, Arnholt AT. 2008. Probability and statistics with R. Chaoman & Hall, CRC.
Villanueva-Cañas JL, Laurie S, Alba MM. 2013. Improving genome-wide scans of positive selection using protein isoforms of similar length. *Genome Biol Evol*. 52(2):457–467.
Wachowiak W, Balk PA, Savolainen O. 2009. Search for nucleotide diversity patterns of local adaptation in dehydrins and other cold-related candidate genes in Scots pine (*Pinus sylvestris* L.). *Tree Genet Genomes*. 5:117–132.
Wang J, Street NR, Scofield DG, Ingvarsson PK. 2016a. Natural selection and recombination rate variation shape nucleotide polymorphism across the genomes of three related *Populus* species. *Genetics*. 202(3):1185–1200.
Wang J, Street NR, Scofield DG, Ingvarsson PK. 2016b. Variation in linked selection and recombination drive genomic divergence during allopatric speciation of European and American aspens. *Mol Biol Evol*. 33(7):1754–1767.
Wang XQ, Ran JH. 2014. Evolution and biogeography of gymnosperms. *Mol Phylo Evol*. 75:24–40.
Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview version 2: a multiple sequence alignment and analysis workbench. *Bioinformatics*. 25(9):1189–1191.
Watson JM, Platzer A, Kazda A, Akimcheva S, Valuchova S, Nizhynska V, Nordborg M, Riha K. 2016. Germline replications and somatic mutation accumulation are independent of vegetative life span in Arabidopsis. *PNAS*. 113(43):12226–12231.
Williams CG, Savolainen O. 1996. Inbreeding depression in conifers: implications for breeding strategy. *Forest Sci*. 42(1):102–117.
Williamson RJ, Josephs EB, Platts AE, Hazzouri KM, Haudry A, Blanchette M, Wright SI. 2014. Evidence for widespread positive and negative selection in coding and conserved noncoding regions of *Capsella grandiflora*. *PLoS Genet*. 10(9):e1004622. DOI:10.1371/journal.pgen.1004622.
Willyard A, Syring J, Gernandt DS, Liston A, Cronn R. 2007. Fossil calibration of molecular divergence infers a moderate mutation rate and recent radiations in *Pinus*. *Mol Biol Evol*. 24(1):90–101.
Won H, Renner S. 2006. Dating dispersal and radiation in the gymnosperm *Gnetum* (Gnetales)-Clock calibration when outgroup relationships are uncertain. *Syst Biol*. 55:610–622.

Xi Z, Rest JS, Davis CC. 2013. Phylogenomics and coalescent analyses resolve extant seed plant relationships. *PLoS ONE*. 8(11):e80870.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl BioSci*. 13:555–556.

Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol*. 15:568–573.

Yang Z, Nielsen R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J Mol Evol*. 46:409–418.

Yang Z, Nielsen R, Goldman N, Krabbe-Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*. 155:431–449.

Yang Z. 2002. Inference of selection from multiple species alignments. *Curr Op Gen Dev*. 12:688–694.

Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol*. 19:908–917.

Yang Z, Wong WSW, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol*. 22:1107–1118.

Yang Z. 2007. PAML 4: a program package for phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 24:1586–1591.