

Evolutionary rate variation in two conifer species, *Taxodium distichum* (L.) Rich. var. *distichum* (baldcypress) and *Cryptomeria japonica* (Thunb. ex L.f.) D. Don (Sugi, Japanese cedar)

Junko Kusumi^{1*}, Yoshihiko Tsumura² and Hidenori Tachida³

¹Department of Environmental Changes, Faculty of Social and Cultural Studies, Kyushu University, 744 Motoooka, Nishi-ku, Fukuoka 819-0395, Japan

²Forestry and Forest Products Research Institute, Matsunosato 1, Tsukuba, Ibaraki 305-8687, Japan

³Department of Biology, Faculty of Sciences, Kyushu University, 744 Motoooka, Nishi-ku, Fukuoka 819-0395, Japan

(Received 1 December 2014, accepted 1 May 2015; J-STAGE Advance published date: 18 December 2015)

With the advance of sequencing technologies, large-scale data of expressed sequence tags and full-length cDNA sequences have been reported for several conifer species. Comparative analyses of evolutionary rates among diverse taxa provide insights into taxon-specific molecular evolutionary features and into the origin of variation in evolutionary rates within genomes and between species. Here, we estimated evolutionary rates in two conifer species, *Taxodium distichum* and *Cryptomeria japonica*, to illuminate the molecular evolutionary features of these species, using hundreds of genes and employing *Chamaecyparis obtusa* as an outgroup. Our results show that the mutation rates based on synonymous substitution rates (*dS*) of *T. distichum* and *C. japonica* are approximately 0.67×10^{-9} and 0.59×10^{-9} /site/year, respectively, which are 15–25 times lower than those of annual angiosperms. We found a significant positive correlation between *dS* and GC3. This implies that a local mutation bias, such as context dependency of the mutation bias, exists within the genomes of *T. distichum* and *C. japonica*, and/or that selection acts on synonymous sites in these species. In addition, the means of the ratios of synonymous to nonsynonymous substitution rate in the two species are almost the same, suggesting that the average intensity of functional constraint is constant between the lineages. Finally, we tested the possibility of positive selection based on the site model, and detected one candidate gene for positive selection.

Key words: *Cryptomeria*, Cupressaceae, dispersion index, evolutionary rate, *Taxodium*

INTRODUCTION

Conifers are important tree species in forests and are distributed widely in the temperate, frigid and even subtropical zones, and some of them are also important for commercial purposes. There are 615 species (70 genera) in this group, which have diverse morphologies and ecological traits (Farjon, 2010). The genomes of conifers are characterized by large size, among the largest of any non-polyploid plant species, slow evolutionary rates in coding genes, and accumulations of a large amount of noncoding

DNA (Murray, 1998; Ahuja and Neale, 2005; Buschiazzi et al., 2012; Pavy et al., 2012; Ritland, 2012). Recent analyses have revealed that conifer genomes have huge amounts of repetitive sequences (probably non-functional) and low recombination rates in non-coding regions (Jaramillo-Correa et al., 2010; Kovach et al., 2010; Liu et al., 2011; Moritsuka et al., 2012). These observations contrast with those of angiosperm genomes reported so far. Unique genomic characteristics of conifers have lately attracted considerable attention as to what causes such differences in these genomic characteristics, and what kind of evolutionary features the conifer genome has: for example, how much intragenome and/or intergenome variation of evolutionary rates exists in conifer species?

In this study, we focused on one of the conifer families,

Edited by Yoko Satta

* Corresponding author. E-mail: jkusumi@scs.kyushu-u.ac.jp

DOI: <http://doi.org/10.1266/ggs.14-00079>

Cupressaceae *s. l.*, which contains economically important timber trees such as junipers, cypresses, redwoods and cedars. We previously estimated synonymous and non-synonymous substitution rates (dS and dN) at 11 nuclear genes in 10 conifer species belonging to three subfamilies (Taxodioideae, Cupressoideae and Sequoioideae) of Cupressaceae *s. l.* (Kusumi et al., 2002). We found a significant positive correlation between dS and dN , and between dS and GC content at the third codon position (GC3), but no correlation between dS and codon usage (measured by the effective number of codons, ENC). In addition, we found that the dispersion index (R , the variance to mean ratio of the evolutionary rates, a kind of indicator of neutrality) of nonsynonymous substitutions was close to 1. The dispersion index is expected to be close to 1 under various mutation-driven models of molecular evolution, unless underlying changes of parameters (e.g., population size) are very slow (Araki and Tachida, 1997; Cutler, 2000). Although Kusumi et al. (2002) suggested mutation-driven evolution in the Cupressaceae family, the number of analyzed genes was too limited to consider these trends as general evolutionary features of Cupressaceae. It is thus necessary to increase the number of genes, to examine to what extent the general features found in this study hold.

Taxodium distichum (L.) Rich var. *distichum* (baldcypress) is a keystone species of wetland forests in the southeastern United States, having unique morphological and ecological traits such as deciduousness, flood tolerance, salinity tolerance and knees (distinctive root morphology). *Cryptomeria* is a close relative of *Taxodium*, which consists of only one living species, *C. japonica* (Thunb. ex L.f.) D. Don (Sugi, Japanese cedar), which is endemic to Japan (Farjon, 2005). *Taxodium* and *Cryptomeria* belong to the same subfamily, Taxodioideae, and these genera diverged roughly 90 million years ago (MYA) (Aulenback and LePage, 1998; Leslie et al., 2012). Recently, partial sequences of more than thirty thousand complete cDNAs of *C. japonica* have been released (Futamura et al., 2008). To analyze the rate and pattern of nucleotide substitutions between the *Taxodium* and *Cryptomeria* lineages, we determined cDNA sequences from *T. distichum* using next-generation sequencing (NGS) technology. Advances in NGS techniques enable us to generate large-scale sequence data in a short time, even from non-model organisms. NGS offers us a resource for comparative analyses of the rate and pattern of nucleotide substitutions based on thousands of genes. Accumulating data from NGS platforms have been elucidating intragenome and interspecies variation in the rate and patterns of nucleotide substitution in coding and non-coding regions for plant species (Gaut et al., 2011). Such analyses also can elucidate the extent to which mutational and selective forces contribute to the variation in evolutionary rates between species; this has been one of

the outstanding problems in molecular evolution (Gaut et al., 2011; Hough et al., 2013).

Large data sets of coding sequences enable us to conduct a comprehensive survey of selective constraints throughout the genome. The ratio of nonsynonymous to synonymous substitution rate (dN/dS) is commonly used as an indicator for detecting the mode and strength of selective forces acting on protein-coding genes. An excess of nonsynonymous substitution ($dN/dS > 1$) suggests diversifying selection, while an excess of synonymous mutations ($dN/dS < 1$) indicates purifying selection, and no significant difference between synonymous and nonsynonymous mutation rates ($dN/dS = 1$) is taken as evidence for neutrality. Warren et al. (2010) performed a large-scale comparison of substitution rates and dN/dS values for genes duplicated in whole genome duplication in *Arabidopsis thaliana* to analyze the dependency of molecular evolutionary rates on biological function, which is characterized by Gene Ontology (GO) annotation. They categorized genes based on their role in the cell as described by GO and compared the dN/dS values among the functional groups. Their analysis revealed that there is a significant variation in dN/dS values between the gene groups; for example, defense response genes have significantly higher dN/dS values than the other gene groups, while genes involved in protein translation have significantly lower dN/dS values than the other gene groups.

Here, we analyzed variation of evolutionary rates in two conifer species, *T. distichum* and *C. japonica*, to investigate the molecular evolutionary features of Cupressaceae described above more extensively by examining hundreds of genes. We also compared the distribution of dN/dS values in lineages and functional categories of genes in *T. distichum* and *C. japonica*. In addition, we performed a maximum likelihood (ML) test developed by Yang et al. (2007) to detect sites undergoing diversifying selection.

MATERIALS AND METHODS

RNA-seq of *T. distichum* Young leaves of *T. distichum* were collected from Kyushu University forest in 2011. The samples were immediately frozen in liquid nitrogen and stored at -80°C until use. RNA was extracted using Concert Plant RNA Reagent (Invitrogen) according to the manufacturer's instruction. Contaminating genomic DNA was removed from the RNA sample by DNase treatment (Promega). Total RNA was sent to the Genome Analysis Consortium at Kyushu University for preparation and sequencing using the Roche GS-FLX (454) Titanium system. One microgram of total RNA was used to construct a normalized cDNA library. Amplified normalized cDNA was purified with a spin column, and then used for 454 sequencing. The cDNA library was

sequenced on one-eighth of a Roche GS-FLX (454) picotiter plate using Titanium chemistry. The sequence data from this study have been submitted to the DDBJ Sequence Read Archive (<http://trace.ddbj.nig.ac.jp/dra/index.html>) under accession no. DRA003169.

The raw sequence reads were assembled using MIRA software v. 3.4 (Chevreux et al., 2004) after removal of adaptor sequences and poly(A) tails. A total of 40,645 reads were used for assembly, but 14,475 reads were excluded from the assemblies because of their short length. Finally, we obtained 7,459 contig sequences whose average length was 465 bp and whose N90 was 340 bp. The average number of reads per contig was 3.65. We then searched for annotation of all obtained contig sequences using Blast2GO (Conesa et al., 2005). In total, 3,976 of the 7,459 sequences were mapped to GO terms, and we used them in the following analyses.

Protein-coding sequences for *T. distichum*, *C. japonica* and *Chamaecyparis obtusa* We used 3,976 contig sequences from *T. distichum* and 32,167 complete cDNA sequences from *C. japonica* obtained from a Sugi genome database (Futamura et al., 2008). We used *Chamaecyparis obtusa* (Siebold & Zucc.) Endl. as an out-group species, for which more than five thousand cDNA sequences have already been determined (Yamashita et al., 2008). *C. obtusa* belongs to Cupressoideae.

Open reading frame (ORF) search in conifer genes To identify protein-coding sequences, all possible ORFs found in the cDNA sequences of *T. distichum*, *C. japonica* and *C. obtusa* were queried against predicted protein sequences of *Picea sitchensis* (23,595 sequences distributed by TreeGenes, <http://dendrome.ucdavis.edu/>) using blastx. We selected ORFs with the same frame as the longest overlapping sequence with the best-scoring BLAST query against the protein sequence of *P. sitchensis*. Of the queried sequences, we found 1,640 ORFs of *T. distichum*, 3,992 ORFs of *C. japonica* and 1,399 ORFs of *C. obtusa* which matched the protein coding sequences of *P. sitchensis*.

We used the reciprocal best hit (RBH) approach (Moreno-Hagelsieb and Latimer, 2008) to infer putative orthologs between ORF sequences of *T. distichum* and *C. japonica* (the 2-species comparison), and between ORF sequences of *T. distichum*, *C. japonica* and *C. obtusa* (the 3-species comparison). From a BLAST search with -e threshold = 10^{-20} , we found a total of 610 RBH sequences with the 2-species comparison (2sp-RBH) and 202 RBH sequences with the 3-species comparison (3sp-RBH). Orthologous ORF sequences were aligned using MAFFT ver. 7 with the highest sensitivity (linsi) (Katoh and Toh, 2008; Katoh and Standley, 2013). Alignments shorter than 30 amino acids were discarded, and, finally, we obtained 600 2sp-RBH orthologous genes and 200 3sp-

RBH orthologous genes.

Data analysis Pairwise distances at non-synonymous sites (dN) and synonymous sites (dS) and dN/dS were estimated for individual 2sp-RBH orthologous genes using codeml (PAML 4.0) with the pairwise option (Yang, 2007). For the estimation of dS and dN , we used models allowing transition-transversion rate bias and unequal codon frequencies, which were determined using the empirical nucleotide frequencies at the three positions of the codon (F3X4 model). Lineage-specific estimates of dS , dN and dN/dS were calculated for individual 3sp-RBH orthologous genes. We used the model allowing independent dN/dS s for the three branches with star phylogeny.

Genes showing signs of saturated divergence and underestimation of dS were excluded from the following analyses. We discarded 12 genes with $dN/dS > 99$ and seven genes with $dS > 0.5$ from the 2sp-RBH orthologous gene set. For the 3sp-RBH orthologous gene set, we discarded 14 genes with high dS in one of the three lineages. Seventeen genes of the 3sp-RBH orthologous gene set showed high $dN/dS (> 100)$ in one of the three lineages because of suspected underestimation of dS , and we therefore discarded them when we calculated mean values of dN/dS . The final 2sp-RBH orthologous gene set contained 581 genes (average gap-free length = 305.1 bp), and the final 3sp-RBH orthologous gene set contained 186 genes (average gap-free length = 267.8 bp).

GC content at each codon position (GC1, GC2, GC3) was estimated by PAML 4.0. We estimated ENCprime (ENC'), a statistic to measure codon usage bias, using the ENCprime program (Novembre, 2002). ENC' can account for background base composition, and was developed by improving the effective number of codons (ENC; Wright, 1990).

To test the presence of codon sites with $dN/dS > 1$, which can be considered as candidate sites undergoing diversifying selection, the ML methods with variable dN/dS among sites (the site model) of Yang (2007) were used. We analyzed 13 genes from the 3sp-RBH orthologous gene set that showed high dN/dS in one of the three lineages in the analysis described above (Supplementary Table S1). We used the following four models for the dN/dS distribution, implemented in codeml (Yang, 2007). The neutral model (M1) assumes conserved sites with $dN/dS < 1$ and neutral sites with $dN/dS = 1$ with a proportion p_0 and a proportion $p_1 = 1 - p_0$, respectively. The selection model (M2) adds an additional dN/dS class with frequency $p_2 = 1 - p_1 - p_0$, with dN/dS estimated from the data. The beta model (M7) assumes that dN/dS varies according to a beta distribution $B(p, q)$, whose domain is bounded within the interval (0, 1). Thus, this model does not allow for codon sites with $dN/dS > 1$. The beta & ω model (M8) adds a discrete dN/dS class to the beta (M7)

model to account for codons with $dN/dS > 1$. Sites with dN/dS drawn from the beta distribution $B(p, q)$ occur with a proportion p_0 , and the rest belong to a discrete dN/dS class (ω_1) and occur in a proportion $p_1 = 1 - p_0$. We compared two pairs of models (M1 vs. M2 and M7 vs. M8) by likelihood-ratio tests (LRTs) to examine the statistical significance of the fit of the model (Yang, 2007).

RESULTS AND DISCUSSION

Distribution of dN and dS for Cupressaceae We obtained 581 2sp-RBH orthologous genes (*Taxodium-Cryptomeria*), and then estimated pairwise dS and dN (Table 1). The mean value of dS averaged over the 581 genes was 0.102 and individual values ranged from complete identity ($dS = 0.0$) to 0.417. The mean value of dN was 0.0106, which was one-tenth that of dS . The distribution of dN values ranged from 0.0 to 0.104. The mean of dS averaged over the 581 genes was slightly higher than that estimated from seven genes in a previous study (0.0847; see Kado et al., 2003), while the mean of dN for the 581 genes was lower than that for the seven genes (0.0166), although the differences were not significant in either case (Welch two sample t -test, two-sided; dS , $p = 0.19$ and dN , $p = 0.36$). The coefficient of variation (CV) of dN was 1.08 and larger than that of dS (0.72). This tendency is also found in *Arabidopsis* species, in which the CV for dN is 0.93 and that for dS is 0.30 (Yang and

Gaut, 2011). Therefore, it seems likely that differences in functional constraint among the genes cause variation in nonsynonymous rates among the genes in both taxa.

Next, we estimated lineage-specific dS and dN using 186 3sp-RBH orthologous genes (*T. distichum*, *C. japonica*, *C. obtusa*) with the model assuming star phylogeny and different dN/dS among branches. Although the sum of substitution rates of *T. distichum* and *C. japonica* lineages was expected to be equal to the substitution rate of pairwise estimation, it did not exactly correspond to the pairwise estimation averaged over 581 genes for all categories of substitution rates. For example, the sum of lineage-specific dNs of *T. distichum* and that of *C. japonica* was estimated to be 0.0081, but it was lower than the mean dN of pairwise estimation based on the 581 genes (0.0106). Because we suspected that this disparity might be caused by differences in the analyzed genes or assuming different models for the ML estimations, we recalculated the mean of substitution rates of pairwise estimation using the identical gene set as those used for the lineage-specific estimation. The mean dN of the 186 genes for pairwise estimation was estimated to be 0.0087, which was closer to the sum of lineage-specific mean dNs of the two lineages (0.0081) than that of the 581 genes. Therefore, the main cause of the difference between the lineage-specific estimation and the pairwise estimation is likely to be the difference in the analyzed gene sets, and not the difference in assuming different models for the

Table 1. Substitution rates in protein-coding regions in Cupressaceae

	Number of genes	Mean length	dS	dN	dN/dS
Pairwise substitution rate ¹					
<i>Cryptomeria-Taxodium</i>	581	305.1	0.1021 (0.0734)	0.0106 (0.0116)	0.160
CV ³			0.72	1.08	
<i>Cryptomeria-Taxodium</i> ⁴	186	322.7	0.1079 (0.0758)	0.0087 (0.0105)	0.095
CV			0.7	1.19	
Lineage-specific substitution rates ²					
<i>Taxodium distichum</i>	186	267.8	0.0606 (0.0575)	0.0044 (0.0066)	0.106 ⁵
CV			0.95	1.48	
<i>Cryptomeria japonica</i>	186	267.8	0.0534 (0.0536)	0.0037 (0.0064)	0.107 ⁵
CV			1.00	1.93	

Mean substitution rates at synonymous sites (dS) and nonsynonymous sites (dN) are expressed as the number of substitutions per site. Standard deviations are shown in brackets.

¹ Estimates based on the 2sp-RBH orthologous genes.

² Estimates based on the 3sp-RBH orthologous genes assuming the star phylogeny.

³ Coefficient of variation.

⁴ Mean values averaged over the identical set of genes used in the lineage-specific estimation.

⁵ Mean dN/dS values were calculated from 169 genes (see details in MATERIALS AND METHODS).

ML estimations.

Although the difference was not significant, dS and dN of *T. distichum* were both slightly higher than those of *C. japonica*. Our previous study reported that the nucleotide diversity of *T. distichum* was significantly higher than that of *C. japonica* (Kusumi et al., 2010), and we concluded that this might mainly be due to the larger population size of *T. distichum*. The present results suggest that the difference in mutation rates of the two species may also contribute to the higher nucleotide diversity in *T. distichum* than in *C. japonica*.

Fossil evidence indicates that *Taxodium* and *Cryptomeria* diverged prior to the Maastrichtian, around 90 million years ago (MYA) (Aulenback and LePage, 1998; Leslie et al., 2012). Assuming that the two genera diverged 90 MYA, that synonymous sites are neutral and that the synonymous substitution rate is constant along the lineages, mutation rates of the *Taxodium* and *Cryptomeria* lineages were estimated to be 0.67×10^{-9} and 0.59×10^{-9} per site per year, respectively. These estimates are close to those of other conifers (e.g., 0.68×10^{-9} estimated for the *Pinus-Picea* divergence (Ritland, 2012), $0.70 - 1.31 \times 10^{-9}$ estimated for the *Pinus* subgenus divergence (Willyard et al., 2007)). Known mutation rates of angiosperms are 1.5×10^{-8} substitutions per site per year in *Arabidopsis* (Koch et al., 2006), 1.0×10^{-8} in *Oryza* (Swigoňová et al., 2004) and 0.25×10^{-8} in *Populus* species (Tuskan et al., 2006; Ingvarsson, 2008). The per year mutation rates of angiosperms species are 15 to 25 times higher than those of conifer species. However, conifer species generally have a long generation time (~25 years or more). If we consider per generation mutation rate assuming a generation time of 25 years, the rate is estimated to be 1.675×10^{-8} per site per generation in *Taxodium* and 1.475×10^{-8} in *Cryptomeria*, which are comparable to the mutation rates of annual angiosperms.

In our previous study (Kusumi et al., 2010), we estimated the population growth rate of *T. distichum* using ML and Bayesian methods implemented in LAMARC 2.1 (Kuhner, 2009). The parameter g , the exponential population growth rate measured in units of $1/u$ (mutations per generation), ranged from 925 to 1,576, indicating that the population size of *T. distichum* has gradually increased. In that report, we estimated the rate of population growth per year using a synonymous substitution rate between Taxodioideae and Sequoioideae averaged over 10 genes (1.9×10^{-9} /site/year) as a neutral mutation rate of these species; however, this rate is three times higher than that estimated in the present study. Therefore, in our previous study, per year population growth rate was calculated to be $1.0 - 1.7 \times 10^{-6}$ per year in *T. distichum*. If we recalculate the per year growth rate using the newly estimated neutral mutation rate of *T. distichum*, the per year population growth rate is estimated to be approximately $0.36 - 0.60 \times 10^{-6}$ per year.

Although the onset of the population expansion is not known, the size of the population of *T. distichum* may have increased by approximately 1.4 – 1.8 times during the past 1 million years.

Correlation between dS , dN , GC12, GC3 and ENC'

In previous work we reported weak but significant positive correlations between dS vs dN and dS vs GC3s in *C. japonica* and *Thujopsis dolabrata* (Cupressoideae) (Kusumi et al., 2002). Moreover, Kado et al. (2008) estimated dS and dN between *C. japonica* and *C. obtusa* (pairwise estimation) for 10 nuclear genes, and reported a positive correlation between dS vs GC3 but no correlation between dS vs dN . Here, we investigated relationships between dS , dN , GC12 (average GC content at the first and second codon position), GC3 and ENC' using pairwise estimates between *T. distichum* and *C. japonica* (Supplementary Table S2). There is a significant positive correlation between dS and GC3 with the pairwise estimates ($p < 0.00001$, adjusting for multiple testing by Holm's methods), although the correlation between dS and dN is not significant ($p > 0.01$, adjusting for multiple testing by Holm's methods) (Fig. 1, A and B). We also investigated correlation between dS and GC3 and between dS and dN using lineage-specific estimates in *T. distichum*, *C. japonica* and *C. obtusa* (Supplementary Fig. S1). The lineage-specific estimates also show a significant correlation between dS and GC3. As shown in our previous study, there is no significant correlation between dS and ENC' (Fig. 1C).

The significant correlation between dS and GC3 may indicate translational selection or mutational bias. If translational selection exists in conifer species, we expect a correlation between dS and codon usage and/or between dS and the gene expression level. As noted above, we could not find a significant correlation between dS and ENC'. In addition, we investigated the correlation between dS and the number of reads for each gene, as an index of gene expression level. The result shows that there is no significant correlation between dS and the number of reads (Supplementary Fig. S2); therefore, translational selection seems unlikely to contribute to the positive correlation between dS and GC3. However, the total number of reads in our analysis may be too small to detect a difference in levels of expression between genes. Comprehensive investigation of gene expression using microarrays or NGS instruments will be necessary for a more precise measurement of levels of gene expression. Furthermore, we need to consider the mutation models used for ML estimation when we draw biological conclusions derived from genome-scale analyses involving estimation of synonymous substitution rates. Aris-Brosou and Bielawski (2006) conducted a simulation analysis to evaluate two different ways of modeling uneven codon frequencies. The results showed that evo-

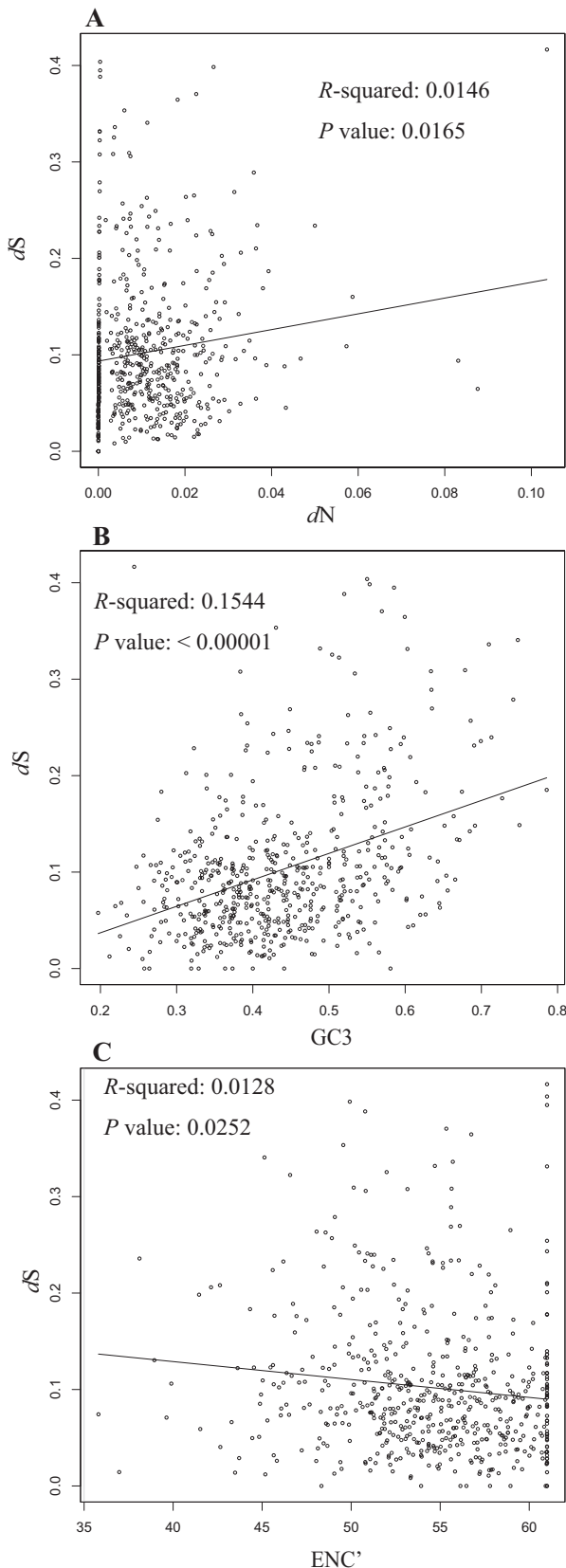


Fig. 1. The relationships between dS and dN (A), dS and GC3 (B) and dS and ENC' (C). Estimates based on the 2sp-RBH orthologous genes (581 genes) are shown.

dence for translational selection can be difficult to find when codon frequencies are uneven, and suggested that the difference in ways of modeling uneven codon frequencies can have a dramatic impact on rate estimates and affect biological conclusions about genome evolution (Aris-Brosou and Bielawski, 2006).

On the other hand, mutation bias across nuclear genes has been reported in maize and *A. thaliana* (Gaut et al., 2011). Both species have higher mutation rates in G and C residues than in A and T residues. Here, we obtained a significant correlation between GC3 and dS , suggesting mutation bias in conifers as found in angiosperms, but we will need additional data to construct an appropriate model for nucleotide substitutions in conifers. More specifically, analyses of the context dependency and methylation dependency of the mutation process (Morton, 1995; Morton et al., 2006) and local variation in GC content are necessary to determine the cause of the correlation.

Variability of the numbers of synonymous and non-synonymous substitutions among lineages To examine variability of the numbers of synonymous and nonsynonymous substitutions among lineages, we calculated the dispersion index (R) with the weighting factor of Gillespie (1989). The dispersion index is defined as the ratio of the variance to the mean, and is a normalized measure of the dispersion of a probability distribution. Simple models of molecular evolution assume that sequences evolve according to a Poisson process in which nucleotide or amino acid substitutions occur as rare independent events with a constant rate. In these models, the expected value of R is 1. We estimated R with weighting factors (see below) using the number of synonymous and nonsynonymous substitutions (S and N) estimated from the lineage-specific estimation. As in our previous study (Kusumi et al., 2002), the weighting factors are proportional to the total number of substitutions across all 186 genes along the lineage in the respective substitution categories (synonymous and nonsynonymous substitutions). Estimated weighting factors are 0.615, 0.567 and 1.818 for nonsynonymous substitutions and 0.625, 0.571 and 1.805 for synonymous substitutions, for *T. distichum*, *C. japonica* and *C. obtusa*, respectively. Distributions of R for nonsynonymous and synonymous substitutions are shown in Fig. 2. For the nonsynonymous substitutions, the weighted R ranged from 0.003 to 7.815, with a mean value of 1.674. For the synonymous substitutions, weighted R ranged from 0.015 to 18.228, with the mean value of 2.813. Our previous study based on 11 genes of Cupressaceae species (Kusumi et al., 2002) reported that the weighted R of nonsynonymous substitution rate was close to 1, which is expected under a simple Poisson process with the same rate among lineages (Kimura, 1983). In contrast, the estimate of R at nonsynonymous sites in the present study was slightly higher than 1, which means

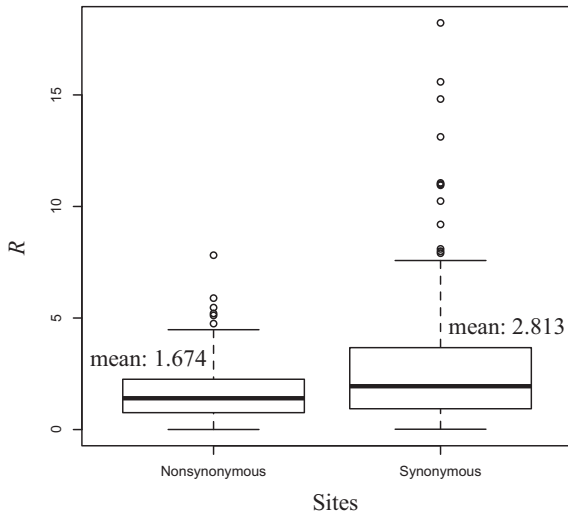


Fig. 2. Distribution of R for nonsynonymous and synonymous substitutions.

overdispersion of the molecular clocks in the Cupressaceae lineages. This difference may be caused by the difference in analyzed genes. Kim and Yi (2008) reported that proteins belonging to different GO terms exhibit different degrees of dispersion in nonsynonymous sites. For example, genes involved in regulation of transcription tend to show lower indices of dispersion, whereas protein hormone genes tend to have large indices of dispersion. As our previous estimation was based on only 11 nuclear genes, the index may have been somewhat biased. The mean of R for nonsynonymous substitutions based on 186 genes in the Cupressaceae lineages, 1.674, was almost at the same level as that in the *Drosophila* lineages (mean R (t) for amino acid substitutions = 1.836 (Bedford and Hartl, 2008)). Because the estimate of R for nonsynonymous sites in the Cupressaceae lineages was still larger than 1 after adjusting lineage effects by weighting factors, the contribution of selective forces on each gene to the variation of nonsynonymous substitution could not be ignored. In addition, Bedford and Hartl (2008) found that the observed deviation from the Poisson expectation in the *Drosophila* lineages is a linear function of the rate at which substitutions occur in a phylogeny, and they therefore concluded that deviations from the Poisson expectation arise from gene-specific temporal variation in substitution rates. For the Cupressaceae lineages, temporal variation in the mutation rate may contribute to the overdispersion of the nonsynonymous substitution; however, we need data from additional related taxa to confirm this. On the other hand, the mean R for the synonymous sites in the Cupressaceae lineages, 2.813, was much greater than those for both the nonsynonymous sites and synonymous sites in the *Drosophila* lineages (mean R (t) for the four-fold synonymous sites = 1.707 (Bedford and Hartl, 2008)).

The synonymous substitution rate could be affected by a variety of factors in lineage- and gene-specific manners. It should be noted that synonymous sites may be subject to selection through mechanisms such as translational efficiency, RNA editing, microRNA binding and conservation of splice signals in animals and plants (Chamary et al., 2006), although evidence for selection acting on synonymous sites is limited in the conifer species. On the other hand, context dependency of mutation rates is known in angiosperms (G and C residues have higher mutation rate than A and T residues) (Morton et al., 2006; Ossowski et al., 2010; Gaut et al., 2011). The Cupressaceae species showed significant positive correlations between dS and GC3, and it is thus possible that the context-dependent mutation rates affect the variation of the synonymous substitution rates in the Cupressaceae species.

Variability of dS/dN among the genes and lineages

We estimated dN/dS values based on pairwise comparison using 581 2sp-RBH orthologous genes. The mean of dN/dS averaged over the 581 genes was 0.160 (95% CI = 0.142, 0.179) (Table 1). This value was higher than the mean of the lineage-specific dN/dS s of *T. distichum* (0.106, 95% CI = 0.077, 0.134) and *C. japonica* (0.107, 95% CI = 0.077, 0.138) lineages, but if we averaged over dN/dS of the pairwise estimation using the same set of genes as those of the lineage-specific estimation, it became 0.095 (CI = 0.074, 0.115), which was similar to the lineage-specific estimation. Therefore, as shown in dN , the difference between the pairwise estimation and lineage-specific estimation in dN/dS is likely to be caused by the difference in the analyzed genes. In addition, the means of dN/dS of *T. distichum* and *C. japonica* lineages were almost the same. (We used a model with different dN/dS for each branch for the ML estimation.) The mean dN/dS of the *C. obtusa* lineage (from the common ancestor of *T. distichum* and *C. japonica* to *C. obtusa*) was 0.095 (95% CI = 0.074, 0.116), which was again similar to those of the *T. distichum* and *C. japonica* lineages. On the other hand, there was no correlation of dN/dS values at orthologous genes between any combinations of the lineages (data not shown). This suggests that the intensity of functional constraint is constant among the three lineages on average, but it may be variable among the lineages in their respective orthologous genes. In the Pinaceae family, the means of branch-specific dN/dS were estimated to be 0.12, 0.14 and 0.15 for the *Pinus*, *Picea* and *Pseudotsuga* branches (Palme et al., 2008) based on 109, 71 and 128 orthologous genes, respectively. Even though different genes were used for the estimation, the means for dN/dS of Pinaceae are similar to those of Cupressaceae, suggesting strong functional constraints in conifer protein evolution throughout the genome. In *Arabidopsis*, the averaged value of dN/dS was estimated

to be 0.203, which is higher than in the conifer species (Yang and Gaut, 2011). These results suggest that most genes in conifer species are under strong purifying selection and that diversifying selection is rare. On the other hand, Buschiazzi et al. (2012) estimated dN/dS using the divergence between Sitka spruce (*P. sitchensis*) and loblolly pine (*Pinus taeda*), and reported that dN/dS , averaged over 3,723 genes, was 0.3137. This value is more than twice that of Palme et al. (2008) and also that between *T. distichum* and *C. japonica* estimated in this study. It is unclear what causes this difference, but differences in the analyzed genes, the divergence time between species and the models for substitutions employed in the analyses may explain it.

Next, we compared dN/dS among functional categories of genes. We used the web-based program CateGORizer to analyze Gene Ontology classification categories (Hu et al., 2008). We classified GO terms of the 2sp-RBH orthologous genes into subclasses of the Molecular function (MF) and Biological process (BP) classes using Plant GO slim categories. We allowed redundancy in classification, which means that one 2sp-RBH orthologous gene can be classified into multiple subclasses. We show results for subclasses each having more than five classified genes (Fig. 3). In both the MF and BP class, we could not find significant differences in dN/dS among the subclasses. This result is in contrast to that of *A. thaliana*, which showed significant differences in dN/dS between functionally categorized gene groups (Warren et al., 2010). For example, dN/dS s of defense response genes were significantly higher than those of the other genes in *A. thaliana*, suggesting that the defense response genes are under either weaker purifying selection or stronger diversifying selection than the other genes. This suggests that the dependency of evolutionary rate on biological function is small or that genes under diversifying selection are rare in Cupressaceae. However, since we used RNA-seq data from young leaves, the majority of the genes used in the analysis may be housekeeping genes and leaf-specific genes. Such limitation of the data may account for the rarity of genes with high dN/dS . Whole-genome sequencing or targeted sequencing with NGS instruments will be necessary to confirm the frequency of diversifying selection in Cupressaceae.

Finally, we tested the presence of codon sites with $dN/dS > 1$ that can be considered as candidate sites undergoing diversifying selection, using the genes that showed high dN/dS in the branch model estimation. We chose 13 3sp-RBH orthologous genes with high dN/dS in at least one of the three lineages, and analyzed them using the ML models with variable dN/dS among sites. We compared the alternative models, M2a and M8, with their corresponding null models, M1a and M7 (M1a vs. M2a and M7 vs. M8), by likelihood-ratio tests (LRTs) to exam-

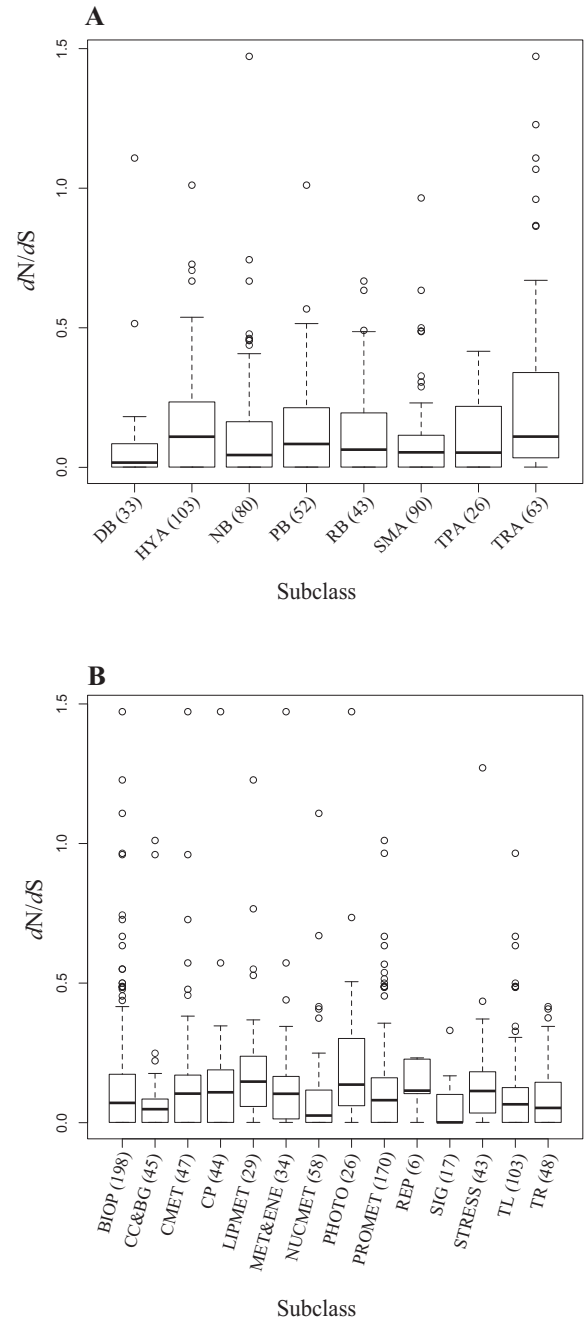


Fig. 3. Variation of dN/dS among functional categories, (A) the Molecular function class and (B) the Biological process class. The number of genes categorized into each subclass is indicated in parentheses. DB: DNA binding; HYA: hydrolase activity; NB: nucleotide binding; PB: protein binding; RB: RNA binding; SMA: structural molecule activity; TPA: transporter activity; TRA: transferase activity; BIOP: biosynthetic process; CC&BP: cellular component organization and biogenesis; CMET: carbohydrate metabolic process; CP: catabolic process; LIPMET: lipid metabolic process; MET&ENE: generation of precursor metabolites and energy; NUCMET: nucleobase, nucleoside, nucleotide and nucleic acid metabolic process; PHOTO: photosynthesis; PROMET: protein metabolic process; REP: reproduction; SIG: signal transduction; STRESS: response to stress; TL: translation; TR: transport.

Table 2. Log-likelihood values and parameter estimates under the models of variable dN/dS (ω) among sites

Sequence ID	Number of codon sites	Number of species	Model	$2\Delta l$	d.f.	P value	Estimates of parameters of selection model			
							Selected sites*	Pr [$\omega > 1$]	Post mean $\pm$ SE for ω	
Tax-rep-c611 (TIP1;1, aquaporin)	224	5	M1a vs. M2a	9.5	2	< 0.01	$p_0 = 0.90021$,	15 V	0.952*	7.577 $\pm$ 2.477
							$p_1 = 0.08104$,	213 A	0.952*	7.574 $\pm$ 2.477
							$p_2 = 0.01875$,			
							$\omega_0 = 0.09660$,			
							$\omega_1 = 1.000$,			
							$\omega_2 = 12.28297$			
			M7 vs. M8	16.91	2	< 0.001	$p_0 = 0.97924$,	15 V	0.975*	7.344 $\pm$ 2.395
							$p_1 = 0.02076$,	141 Y	0.967*	7.294 $\pm$ 2.460
								211 T	0.967*	7.288 $\pm$ 2.463
							$p = 0.51131$	213 A	0.978*	7.359 $\pm$ 2.372
							$q = 2.89520$			
							$\omega = 11.31830$			

* Positively selected sites based on Bayes Empirical Bayes analysis (Yang et al., 2005)

ine the fit of the model (Yang, 2007). To enhance the power of the ML test we added homologous genes of related species to our alignments. Parameter estimates and LRTs suggested the presence of positively selected sites in a gene homologous to TIP1;1 (γ TIP) of *A. thaliana*, a member of the tonoplast family of aquaporins (Table 2). This gene showed high dN/dS in the *T. distichum* branch ($dN/dS = 0.925$) but moderate dN/dS in the *C. japonica* and *C. obtusa* branches (0.339 and 0.232, respectively) in the branch model estimation. We used homologous genes from three *Picea* species, *P. sitchensis* (gb|EF083674.1), *P. glauca* (gb|BT102589.1) and *P. abies* (emb|AJ005078.2), for the ML test of this gene. In the LRT tests comparing M1a vs. M2a and M7 vs. M8 using the χ^2 distribution with d. f. = 2, the P values were 0.0087 and 0.00021, respectively. The neutral model (M1a or M7) was therefore rejected in favor of a selection model (M2a or M8, respectively) in each comparison. The selection model M2a indicated that two sites were positively selected with a posterior mean of $dN/dS = 7.577$ and 7.574, respectively, and both sites were also detected by the M8 (beta& ω) model. Aquaporin is a water-specific membrane protein channel that constitutes a large gene family in animals and plants (Törnroth-Horsefield et al., 2006). Plants regulate the flux of water into and out of the cell by means of aquaporin channels. *T. distichum* is characterized by its unique habitat of wetland forests, where annual flooding occurs and the salinity is high, especially in estuaries. Diversification of aquaporin may therefore be responsible for adaptation to water stress and salinity. We found only one significant gene among 13 candidate genes which showed high dN/dS in at least one of the lineages. This may be because high dN/dS

found in the remaining 12 genes is due to sampling errors caused by their short sequence lengths; alternatively, the limitation of statistical power in a test based on the site model may have caused negative results.

Conclusion Our results showed that species of Cupressaceae had very low mutation rates and strong functional constraint in their coding genes. We found a significant positive correlation between dS and GC3 but no correlation between dS and dN or between dS and ENC'. The correlation between dS and GC3 could be explained by mutation bias, but a more precise estimation of dS is necessary to understand the biological implications of these observations. Overdispersion of the molecular clocks in the Cupressaceae lineage was observed in both nonsynonymous and synonymous substitutions. The mean R at nonsynonymous sites was 1.674, which was almost the same as that in the *Drosophila* lineage. In contrast, the mean R at synonymous sites was much higher than that at nonsynonymous sites. If the overdispersion of the number of substitutions at synonymous sites is caused by temporal and local variation of mutation rates, low variance of the number of nonsynonymous substitutions may imply that Cupressaceae genomes are under strong purifying selection. Finally, we could find only one candidate gene for diversifying selection among hundreds of genes, and no significant difference in dN/dS among GO functional categories. These results suggest that diversifying selection is rare in the Cupressaceae species, but we cannot ignore the possibility that the negative results are caused by a limitation of the data used for the test. In addition to comparative analyses of evolutionary rates such as that presented

here, analyses of patterns of polymorphism in combination with divergence using population genetic approaches will be necessary to elucidate the role of selection on the whole genomes of Cupressaceae species.

We thank two anonymous reviewers for their thoughtful comments on an earlier version of this manuscript. We are also grateful to Dr. K. Tashiro for his help in 454 sequencing. This study was partially supported by Grants-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (no. 22370083, 26291082).

REFERENCES

- Ahuja, M. R., and Neale, D. B. (2005) Evolution of genome size in conifers. *Silvae Genet.* **54**, 126–137.
- Araki, H., and Tachida, H. (1997) Bottleneck effect on evolutionary rate in the nearly neutral mutation model. *Genetics* **147**, 907–914.
- Aris-Brosou, S., and Bielawski, J. P. (2006) Large-scale analyses of synonymous substitution rates can be sensitive to assumptions about the process of mutation. *Gene* **378**, 58–64.
- Aulenback, K. R., and LePage, B. A. (1998) *Taxodium wallisii* Sp. Nov.: first occurrence of *Taxodium* from the Upper Cretaceous. *Int. J. Plant Sci.* **159**, 367–390.
- Bedford, T., and Hartl, D. L. (2008) Overdispersion of the molecular clock: temporal variation of gene-specific substitution rates in *Drosophila*. *Mol. Biol. Evol.* **25**, 1631–1638.
- Buschiazio, E., Ritland, C., Bohlmann, J., and 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.
- Chamary, J. V., Parmley, J. L., and Hurst, L. D. (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat. Rev. Genet.* **7**, 98–108.
- Chevreur, B., Pfisterer, T., Drescher, B., Driesel, A. J., Müller, W. E. G., Wetter, T., and Suhai, S. (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res.* **14**, 1147–1159.
- Conesa, A., Gotz, S., Garcia-Gomez, J. M., Terol, J., Talon, M., and Robles, M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676.
- Cutler, D. J. (2000) Understanding the overdispersed molecular clock. *Genetics* **154**, 1403–1417.
- Farjon, A. (2005) *A Monograph of Cupressaceae and Sciadopitys*. Royal Botanical Gardens, Kew, UK.
- Farjon, A. (2010) *A Hand Book of The World's Conifers*. pp. 526, Brill, Leiden-Boston.
- Futamura, N., Totoki, Y., Toyoda, A., Igasaki, T., Nanjo, T., Seki, M., Sakaki, Y., Mari, A., Shinozaki, K., and Shinohara, K. (2008) Characterization of expressed sequence tags from a full-length enriched cDNA library of *Cryptomeria japonica* male strobili. *BMC Genomics* **9**, 383.
- Gaut, B., Yang, L., Takuno, S., and Eguiarte, L. E. (2011) The patterns and causes of variation in plant nucleotide substitution rates. *Annu. Rev. Ecol. Evol. Syst.* **42**, 245–266.
- Gillespie, J. H. (1989) Lineage effects and the index of dispersion of molecular evolution. *Mol. Biol. Evol.* **6**, 636–647.
- Hough, J., Williamson, R. J., and Wright, S. I. (2013) Patterns of selection in plant genomes. *Annu. Rev. Ecol. Evol. Syst.* **44**, 31–49.
- Hu, Z.-L., Bao, J., and Reecy, J. M. (2008) CateGORizer: a web-based program to batch analyze Gene Ontology classification categories. *Online J. Bioinformatics* **9**, 108–112.
- Ingvarsson, P. K. (2008) Multilocus patterns of nucleotide polymorphism and the demographic history of *Populus tremula*. *Genetics* **180**, 329–340.
- Jaramillo-Correa, J. P., Verdú, M., and González-Martínez, S. C. (2010) The contribution of recombination to heterozygosity differs among plant evolutionary lineages and life-forms. *BMC Evol. Biol.* **10**, 22.
- Kado, T., Yoshimaru, H., Tsumura, Y., and Tachida, H. (2003) DNA variation in a conifer, *Cryptomeria japonica* (Cupressaceae sensu lato). *Genetics* **164**, 1547–1559.
- Kado, T., Matsumoto, A., Ujino-Ihara, T., and Tsumura, Y. (2008) Amounts and patterns of nucleotide variation within and between two Japanese conifers, sugi (*Cryptomeria japonica*) and hinoki (*Chamaecyparis obtusa*) (Cupressaceae sensu lato). *Tree Genetics & Genomes* **4**, 133–141.
- Katoh, K., and Standley, D. M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780.
- Katoh, K., and Toh, H. (2008) Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics* **9**, 286–298.
- Kim, S.-H., and Yi, S. V. (2008) Mammalian nonsynonymous sites are not overdispersed: comparative genomic analysis of index of dispersion of mammalian proteins. *Mol. Biol. Evol.* **25**, 634–642.
- Kimura, M. (1983) *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK.
- Koch, M. A., Haubold, B., and Mitchell-Olds, T. (2000) Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae). *Mol. Biol. Evol.* **17**, 1483–1498.
- Kovach, A., Wegrzyn, J. L., Parra, G., Holt, C., Bruening, G. E., Loopstra, C. A., Hartigan, J., Yandell, M., Langley, C. H., Korf, I., et al. (2010) The *Pinus taeda* genome is characterized by diverse and highly diverged repetitive sequences. *BMC Genomics* **11**, 420.
- Kuhner, M. K. (2009) Coalescent genealogy samplers: windows into population history. *Trends in Ecol. & Evol.* **24**, 86–93.
- Kusumi, J., Tsumura, Y., Yoshimaru, H., and Tachida, H. (2002) Molecular evolution of nuclear genes in Cupressaceae, a group of conifer trees. *Mol. Biol. Evol.* **19**, 736–747.
- Kusumi, J., Zidong, L., Kado, T., Tsumura, Y., Middleton, B. A., and Tachida, H. (2010) Multilocus patterns of nucleotide polymorphism and demographic change in *Taxodium distichum* (Cupressaceae) in the lower Mississippi River alluvial valley. *Am. J. Bot.* **97**, 1848–1857.
- Leslie, A. B., Beaulieu, J. M., Rai, H. S., Crane, P. R., Donoghue, M. J., and Mathews, S. (2012) Hemisphere-scale differences in conifer evolutionary dynamics. *Proc. Natl. Acad. Sci. USA* **109**, 16217–16221.
- Liu, W., Thummasuwan, S., Sehgal, S. K., Chouvarine, P., and Peterson, D. G. (2011) Characterization of the genome of bald cypress. *BMC Genomics* **12**, 553.
- Moreno-Hagelsieb, G., and Latimer, K. (2008) Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinformatics* **24**, 319–324.
- Moritsuka, E., Hisataka, Y., Tamura, M., Uchiyama, K., Watanabe, A., Tsumura, Y., and Tachida, H. (2012) Extended linkage disequilibrium in noncoding regions in a conifer, *Cryptomeria japonica*. *Genetics* **190**, 1145–1148.
- Morton, B. R. (1995) Neighboring base composition and trans-

- version/transition bias in a comparison of rice and maize chloroplast noncoding regions. *Proc. Natl. Acad. Sci. USA* **92**, 9717–9721.
- Morton, B. R., Bi, I. V., McMullen, M. D., and Gaut, B. S. (2006) Variation in mutation dynamics across the maize genome as a function of regional and flanking base composition. *Genetics* **172**, 569–577.
- Murray, B. G. (1998) Nuclear DNA amounts in gymnosperms. *Ann. Bot.* **82** (suppl. 1), 3–15.
- Novembre, J. A. (2002) Accounting for background nucleotide composition when measuring codon usage bias. *Mol. Biol. Evol.* **19**, 1390–1394.
- Ossowski, S., Schneeberger, K., Lucas-Lledo, J. I., Warthmann, N., Clark, R. M., Shaw, R. G., Weigel, D., and Lynch, M. (2010) The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* **327**, 92–94.
- Palme, A. E., Wright, M., and Savolainen, O. (2008) Patterns of divergence among conifer ESTs and polymorphism in *Pinus sylvestris* identify putative selective sweeps. *Mol. Biol. Evol.* **25**, 2567–2577.
- Pavy, N., Pelgas, B., Laroche, J., Rigault, P., Isabel, N., and 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 Biology* **10**, 84.
- Ritland, K. (2012) Genomics of a phylum distant from flowering plants: conifers. *Tree Genetics & Genomes* **8**, 573–582.
- Swigoňová, Z., Lai, J., Ma, J., Ramakrishna, W., Llaca, V., Bennetzen, J. L., and Messing, J. (2004) Close split of sorghum and maize genome progenitors. *Genome Res.* **14**, 1916–1923.
- Törnroth-Horsefield, S., Wang, Y., Hedfalk, K., Johanson, U., Karlsson, M., Tajkhorshid, E., Neutze, R., and Kjellbom, P. (2006) Structural mechanism of plant aquaporin gating. *Nature* **439**, 688–694.
- Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., et al. (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**, 1596–1604.
- Warren, A., Anandakrishnan, R., and Zhang, L. (2010) Functional bias in molecular evolution rate of *Arabidopsis thaliana*. *BMC Evol. Biol.* **10**, 125.
- Willyard, A., Syring, J., Gernandt, D. S., Liston, A., and Cronn, R. (2007) Fossil calibration of molecular divergence infers a moderate mutation rate and recent radiations for *Pinus*. *Mol. Biol. Evol.* **24**, 90–101.
- Wright, F. (1990) The “effective number of codons” used in a gene. *Gene* **87**, 23–29.
- Yamashita, S., Yoshida, M., Yamamoto, H., and Okuyama, T. (2008) Screening genes that change expression during compression wood formation in *Chamaecyparis obtusa*. *Tree Physiol.* **28**, 1331–1340.
- Yang, L., and Gaut, B. S. (2011) Factors that contribute to variation in evolutionary rate among *Arabidopsis* genes. *Mol. Biol. Evol.* **28**, 2359–2369.
- Yang, Z. (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591.
- Yang, Z., Wong, W. S. W., and Nielsen, R. (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.* **22**, 1107–1118.