

Evolutionary Aspects of Functional and Pseudogene Members of the Phytochrome Gene Family in Scots Pine

Maria Rosario García-Gil

Received: 14 December 2007 / Accepted: 9 June 2008 / Published online: 29 July 2008
© Springer Science+Business Media, LLC 2008

Abstract According to the neutral theory of evolution, mutation and genetic drift are the only forces that shape unconstrained, neutral, gene evolution. Thus, pseudogenes (which often evolve neutrally) provide opportunities to obtain direct estimates of mutation rates that are not biased by selection, and gene families comprising functional and pseudogene members provide useful material for both estimating neutral mutation rates and identifying sites that appear to be under positive or negative selection pressures. Conifers could be very useful for such analyses since they have large and complex genomes. There is evidence that pseudogenes make significant contributions to the size and complexity of gene families in pines, although few studies have examined the composition and evolution of gene families in conifers. In this work, I examine the complexity and rates of mutation of the phytochrome gene family in *Pinus sylvestris* and show that it includes not only functional genes but also pseudogenes. As expected, the functional *PHYO* does not appear to have evolved neutrally, while phytochrome pseudogenes show signs of unconstrained evolution.

Keywords Phytochrome · Pseudogenes · Gene family evolution · Scots pine

Introduction

The complexity of gene families in plant taxa is correlated with the size of their genomes (Kinlaw and Neale 1997; Ahuja and Neale 2005). Conifers have larger genomes than angiosperm trees, and pine genomes range in size from 21,500 Mb in *Pinus banksiana* to 37,000 Mb in *Pinus gerardiana* (Murray 1998). Their large genomes have been proposed to result from the accumulation of large amounts of transposable elements, large rDNA repeat units, large gene families, and pseudogenes. Examples of pseudogene members of larger gene families have been described in several conifers, including *Pinus taeda* (Skinner and Timko 1998), *Pinus thunbergii* (Wakasugi et al. 1994), *Larix* (Wei and Wang 2004), *Pinus monophylla* (Gernandt et al. 2001), and *Picea abies* (Kvarnheden et al. 1995). Two phytochrome pseudogenes have also been recorded in GenBank; one from *Ginkgo biloba* (X98699; Kolukisaoglu, unpublished data) and another from *Pinus sylvestris* (AJ286646; Schmidt and Schneider-Poetsch, unpublished data). These works report the presence of pseudogenes in conifer gene families, but there is still too little information to draw general conclusions regarding the contribution of pseudogenes to the size of gene families and their possible evolutionary roles, in conifers. Indeed, the cited study of *Picea abies* (Kvarnheden et al. 1995) is the only one in which the acquired data were analyzed from an evolutionary perspective.

Pseudogenes usually retain some degree of similarity with the genes from which they originated, but are non-functional due to the accumulation of mutations that have created premature stop codons, disruptive frame shift mutations, or alterations in their regulatory regions or splicing sites. Such mutations generally lead either to loss of the capacity to transcribe or translate the corresponding gene products or to the synthesis of nonfunctional proteins.

M. R. García-Gil (✉)
Department of Forest Genetics and Plant Physiology, SLU,
90183 Umea, Sweden
e-mail: M.Rosario.Garcia@genfys.slu.se

M. R. García-Gil
Department of Biology, University of Oulu, P.O. Box 3000,
90014 Oulu, Finland

Pseudogenes can originate from duplications of functional genes, from existing pseudogenes through mRNA retrotransposition (processed pseudogenes), or from genomic DNA duplication (duplicated pseudogenes [Mighell et al. 2000]).

Pseudogenes often evolve with little or no selective constraint. Therefore, according to the neutral theory of evolution, mutation and genetic drift should be the only forces shaping their evolution, and neutrally evolving pseudogenes provide opportunities to obtain direct estimates of mutation rates that are not biased by selection (Li et al. 1981). However, although pseudogenes are frequently nonfunctional, their formation does not always lead to either loss of gene function or unconstrained subsequent mutation. There are many examples of functional pseudogenes that play roles in gene expression, gene regulation, and generation of gene diversity (Balakirev and Ayala 2003). Functional pseudogenes may be under selective constraints, and thus their evolution will deviate from neutrality. Hence, testing for neutrality is crucial before drawing inferences regarding neutral mutation rates based on pseudogene sequences.

Several types of statistical methods can be used to detect signs of selection from DNA sequence data (Nielsen 2001; Swanson 2003), including methods based on analyses of allelic distributions and/or levels of variability, and others based on divergence/variability between different classes of mutations such as nonsynonymous and synonymous mutations. In analyses of allelic distributions, the effects of selection and demographic events can be confounded, so it is important to account for demographic patterns in the null models. The other type of methods are generally more robust since they are insensitive to demographic features and selection is tested by comparing variability in synonymous and nonsynonymous sites, or other categories of mutations, in the same genomic region. An example of this type of test is the McDonald and Kreitman (1991) test, in which the ratio of nonsynonymous-to-synonymous polymorphisms within species is compared to the ratio of nonsynonymous-to-synonymous fixed differences between species in a 2×2 contingency table. In another robust method for testing neutrality, proposed by Akashi (1999), frequency distributions in nonsynonymous and synonymous sites are compared using a test of homogeneity. However, the most direct method for detecting signs of selection is based on the ratio (ω) of replacement divergence between pairs of sequences (K_a) to synonymous divergence between pairs of sequences (K_s) in functional genes (Hughes 2000). When $\omega = 1$, $\omega < 1$, and $\omega > 1$, there is no evidence for selection, evidence for negative selection, and evidence for positive selection, respectively.

There are many statistical methods for testing departures from neutrality based on the ω parameter (Huelsenbeck and

Ronquist 2001; Nei and Kumar 2000; Suzuki and Gojobori 1999), most of which assume neutral evolution at synonymous sites. However, synonymous sites of functional genes may be under selective constraint due to differential codon usage (Akashi 1997; Li 1997), hitchhiking (Kaplan et al. 1989), or background selection (Charlesworth et al. 1993). Differential codon usage introduces bias from neutrality due to the preferential use of certain codons for specific amino acids, which may be related to the efficiency and/or accuracy of protein synthesis (Xia 1996), while hitchhiking and background selection change the frequency of synonymous sites in linkage disequilibrium with replacement sites under selection. Hence, more realistic methods that account for codon bias and transition/transversion rates provide better estimates of ω ratios (Yang and Bielawski 2000). Furthermore, the ω ratio is usually believed to reflect only the effect of selection, but this may not always be case. Recent studies have shown a strong correlation between the fixation probability of replacement mutations and the mutation rate, indicating that the ω ratio may not solely reflect the intensity of selection (Reed et al. 2007; Wyckoff et al. 2005).

Phytochromes are the most thoroughly characterized photoreceptors in plants; they have been found in both lower and higher plants (Mathews and Donoghue 1999) and they are involved in diverse developmental processes, including, among other things, seed germination and de-etiolation, cotyledon movements, shade avoidance, and flowering (Neff et al. 2000). In the present work, a novel phytochrome subfamily composed of several pseudogenes (*psPHY*) was characterized in Scots pine and compared to functional sequences of phytochrome genes, in particular, *PHYO* and *PHYP*. The main goals were to study the evolution and complexity of the phytochrome gene family, the mechanisms of pseudogene creation, and their contribution to the complexity of the pine genome.

Materials and Methods

Plant Material and DNA Extraction

The plant materials used in the study were seeds collected from five unrelated trees (spaced at least 500 m apart) from each of two natural populations of Scots pine in Finland (Lapinjärvi and Kolari, 66° and 60°N, respectively) and one in Spain (a small, relic population at Puebla de Lillo; 38°N). In a previous analysis of seeds collected from the same sources, García-Gil et al. (2003) found no association between variations in functional members of the *PHYO* and *PHYP* gene families, and a steep latitudinal cline in the timing of bud set. However, these findings should be treated with caution since the design of the experiment was not optimal for studying large-scale geographic population

structure. In the study reported here a single seed from each of the sampled trees was randomly chosen and DNA was extracted from the nutritive megagametophyte tissue surrounding the embryo using a FastDNA Kit, following the protocol recommended by the manufacturer (BIO 101). The megagametophytic tissue is haploid, allowing different members of gene families and different alleles of a gene to be readily distinguished after amplification, and only one megagametophyte was used per tree to amplify *PHYO* and phytochrome pseudogene (*psPHY*) sequences. Therefore, all *psPHY* and *PHYO* sequences from each tree can be treated as haplotypic data.

Sequencing

The *PHYO* (U60264) sequence from GenBank was used to design PCR primers (28) for the partial amplification of *PHYO* genomic sequences. The primers used for PCR amplification of *PHYO* were F771 (5'-GAT-GAACACGGTGAGGTAGTTGCA-3') and R2390 (5'-CTTTTTCACGATTTCAACCGAGTC-3'), which amplified the functional locus and also several phytochrome pseudogene sequences (*psPHY*). Standard PCR amplification procedures were used (DyNAzyme EXT; Finnzyme). The PCR products were resolved electrophoretically in agarose gels, purified from the gels using a QIAEX II kit (Qiagen), and then cloned using a TOPO PCR Cloning Kit (Invitrogen). Clones were sequenced using an automatic sequencer (ABI 377) and a BigDye kit (both from PE Applied Biosystems).

Pseudogene Identification

Several clones from each of five individual trees per population were sequenced, and, in total, 36, 62, and 41 sequences obtained from seeds originating from the Kolari, Lapinjärvi, and Puebla de Lillo populations, respectively, were identified as putative pseudogenes, all of which contained indels and/or premature stop codons. The error rate of Taq polymerase, determined by amplifying and sequencing the same clone three times, was estimated to be 10^{-3} (one mutation error every 1000 bp). All the *psPHY* sequences amplified showed higher frequencies of point mutations than expected from random Taq polymerase errors. Furthermore, the same stop codon mutations and indels were found in different trees from both the same populations and other populations, supporting this finding.

Phylogenetic Analysis

The highest similarity to the phytochrome gene family members in Scots pine was obtained by aligning *psPHYs* versus *PHYP* (EU203157), *PHYO* (EU203180), and *PHYN*

(AJ271627) sequences. The phylogenetic relationships between the *psPHY* sequences and the phytochrome gene family members in Scots pine were explored by two methods, as described below.

Distance-Matrix Method

Twelve hundred six-base-pair fragments from exon I of three phytochrome genes—L1_*PHYO* (where L1 indicates the origin of the sequence; the seed designated 1 from the Lapinjärvi population, EU203180), L1_*PHYP* (EU203157), and *PHYN* (AJ271627)—were aligned with *psPHY* sequences using ClustalX (Higgins and Sharp 1988), and the alignment was edited using GeneDoc Version 2.5.000. Phylogenetic analysis of the sequences was then carried out using neighbor joining and UPGMA distance algorithms based on the Kimura two-parameter method using Mega 2.1 software. Bootstrapping (10,000 replications) was used to estimate the branch reliability of the resulting trees. Indels were included in the phylogenetic reconstruction, but excluding them from the analysis did not change either the topology of the tree or the bootstrap values for branch support.

Tree-Searching Method (Bayesian Method)

MrBayes 3.1 software (Huelsenbeck and Ronquist 2001) was used to infer a phylogenetic tree in a Bayesian framework, with default settings except that the general structure of the substitution model was set as General Time Reversible, and the rates of substitution were set as gamma-distributed, with a proportion of invariable sites.

Neutrality Tests

Analysis of Sequence Variation

Single-nucleotide polymorphism among the sequences was quantified by the scaled mutation parameter θ ($\theta = 4N_e\mu$, where N_e is the effective population size and μ the mutation rate), based on the observed nucleotide diversity between two random sequences parameter, π (Nei 1987). In addition to the nucleotide polymorphisms, 1- to 2-bp indels (insertion/deletions) were observed. In subsequent phylogenetic analyses the indels were either treated as single-nucleotide-site polymorphisms or excluded.

F_{ST} , the proportion of nucleotide diversity due to differences among populations (Wright 1969), was estimated by analysis of molecular variance (AMOVA [Excoffier et al. 1992]) using Arlequin 2001 software, but only for *psPHY1* and *psPHY4*, since these were the only loci for which there was sufficient information.

Genetic mixture analysis with linked loci was performed using BAPS4 (Corander et al. 2007) software to investigate

possible clustering of the 15 individuals. If $F_{ST} = 0$ and no structure is found, the individual data from all of the examined populations can be merged into a single data set. This can be justified in cases where there is little or no differentiation between the populations due to long-distance pollen flow and random mating. Previous studies on among-population differentiation at other loci in material from the three populations sampled here support this scenario (Dvornyk et al. 2002; García-Gil et al. 2003).

Effective number of codons (ENC) values, which quantify the effective number of codons used in a gene (Wright 1986), were computed as a measure of codon bias in DnaSP 3.5. ENC values can range from 20 (when only one codon is used for each amino acid and thus the codon bias is maximal) to 61 (when all synonymous codons for each amino acid are used equally often).

Point Mutation Distribution

Synonymous and replacement sites in the psPHY sequences were identified by aligning them against PHYO from Scots pine. Then the point mutation distributions of synonymous vs replacement substitutions were analyzed, the codon positions classified accordingly, and chi-square tests for departure from neutrality applied using DnaSP 3.5 software.

Gratham's distances (which provide indications of degrees of amino acid dissimilarity) between amino acids encoded by the reference sequence and replacement amino acids encoded by the psPHY and PHYO sequences were estimated using DAMBE 4.1.19 software (Xia 2001).

Phylogeny-Based Method for Detecting Selection

Sites under positive selection were detected using MrBayes v3.1 software (Huelsenbeck and Ronquist 2001), which applies Markov chain Monte Carlo (MCMC) methods to generate draws from the posterior distribution of the parameters of the chosen model, set by choosing Nucmodel=codon (which analyses DNA sequences in terms of their codons) and Omegavar=M3. In M3 models the proportions of nonsynonymous-to-synonymous substitutions at all of the codons are considered and each site is assigned a so-called omega parameter (ω). If $\omega = 1$ there is deemed to be no selection pressure and the site is neutral, if $\omega < 1$ the site is under negative selection, and if $\omega > 1$ the site is under positive or diversifying selection. Distributions of the ω parameter were obtained from analyses of psPHY sequences versus Scots pine PHYO (EU203180), *Picea abies* PHYO (U60264), and *Arabidopsis thaliana* PhyA (homologue to PHYO; NM100818). The prior distributions of the parameters describing M3 models are the transition-to-transversion rates ratio $k = dT/dR$; the codon frequencies

π_1, \dots, π_{61} ; the three nonsynonymous-to-synonymous rates ratios ω_1, ω_2 , and ω_3 , with their corresponding frequencies p_1, p_2 , and p_3 ; the parameters associated with the tree topology, denoted T; and a parameter related to the branch lengths of the tree denoted B. The following prior distributions were considered: $dT/(dT + dR) \sim \beta(1,1)$; $\pi_1, \dots, \pi_{61} \sim \text{Dirichlet}(1, \dots, 1)$; $B \sim \text{exponential}(10)$; a uniform prior distribution on all the bipartite trees was assumed for T; and an exponential prior distribution for dN_1, dN_2, dN_3 , and dS (the rates of nonsynonymous substitutions for the three ω categories and rate of synonymous substitutions, respectively). It was assumed that $dN_1 < dN_2 < dN_3$ and $\omega_j = dN_j/dS$ for $j = 1, 2, 3$. One run and three chains with 25,000 burn-in cycles and 250,000 iterations of the algorithm were used to obtain the posterior distribution. During each iteration of the MCMC algorithm, MrBayes computes the probability that each codon belongs to each selection category and for each codon the sum of these probabilities for each selection category with $\omega > 1$ is reported. A site is deemed to be under positive selection if the posterior mean of the probabilities of positive selection is > 0.95 .

Among-Species Divergence

Replacement divergence between species (K_a) and synonymous divergence between species (K_s) at functional genes were estimated by the Kimura two-parameter method using DnaSP 3.5 software by aligning *Picea abies* PHYO against each Scots pine PHYO and each psPHY subfamily sequence, but for simplicity only the highest and lowest values within each locus are shown here (see Table 4).

Results

Pseudogene Identification

I amplified and sequenced the functional PHYO locus (accession numbers EU203172 to EU203183) and pseudogene (psPHY) (accession numbers EU203184 to EU203199 and EU203201 to EU203215) sequences. The pseudogene sequences showed several stop codons and indels when aligned with the L1_PHYO sequence (Table 1). Stop codons may prevent correct transcription and/or translation, leading to a nonfunctional gene (pseudogene), while indels found in the psPHY sequences changed the reading frame, at least in some parts of the sequence. Most of the indels were single-nucleotide deletions; two-nucleotide deletions were only detected at position 1716 bp (position based on the total length of the exonic region of PHYO). The same indels were found in more than one sequence, supporting the hypothesis that they were not due to random Taq polymerase errors. Four

Table 1 Deletions and insertions at each *psPHY* relative to the functional *PHYO* locus

Locus	875	1077	1137	1220	1360	1374	1440	1716	1776	1933	1946
<i>psPHY1</i>	i	—	—	—	—	—	d	—	—	—	d ^a
<i>psPHY2</i>	i	d ^b	d ^b	D	—	—	d	—	d ^c	—	d
<i>psPHY3</i>	—	—	—	—	d	—	—	d ^d	—	d	—
<i>psPHY4</i>	—	—	—	—	d ^e	i ^f	—	d	—	d	—

Note: The number assigned to each *psPHY* sequence (from 1 to 4) indicates its *psPHY* subgroup based on sequence similarity

^a Puebla de Lillo4 lacks the deletion

^b Deletion present only in Läpinjärvi1

^c Deletion present only in Läpinjärvi5

^d Two-base pair deletion

^e Puebla de Lillo1, Puebla de Lillo3, and Puebla de Lillo4 lack the deletion

^f Puebla de Lillo4 lacks the indel

Table 2 Number of clones sequenced from each individual representing each population

Sample	Kolari					Läpinjärvi					Puebla de Lillo					Σ
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	
<i>PHYO</i>	3	3	—	3	3	3	3	3	3	3	3	—	—	3	3	36
<i>psPHY1</i>	2	—	5	4	—	1	3	—	4	—	2	7	3	1	—	32
<i>psPHY2</i>	—	—	—	1	—	3	1	—	—	12	4	—	—	1	6	28
<i>psPHY3</i>	—	1	—	—	—	3	—	2	4	—	—	—	—	—	—	10
<i>psPHY4</i>	3	4	1	2	1	2	—	2	—	10	1	—	6	1	—	33
Σ	8	8	6	10	4	12	7	7	11	25	10	7	9	6	9	139

pseudogene subfamilies were detected and designated *psPHY1*, *psPHY2*, *psPHY3*, and *psPHY4*.

The number of clones sequenced for each *psPHY* locus from each individual (two to four in most cases) is listed in Table 2. The reliability of data obtained for loci with only one clone per individual was supported by amplifying the same loci in other individuals from the same or different populations. The only locus that was amplified from DNA obtained from all five sampled individuals of a population was *psPHY4*, which was found in all five sampled individuals of the Kolari population. The other *psPHY* loci were amplified from DNA obtained from only one to four individuals per population. However, all of the *psPHY* loci were detected in at least one representative of each population, except *psPHY3*, which was not amplified from any of the Puebla de Lillo samples. It should be noted that a conservative criterion was applied, whereby 21 sequences showing mutations that were amplified in only one clone and only one individual from all populations were excluded (data not shown), since they could have arisen from random errors, but they could of course be indicative of additional members of a *psPHY* gene subfamily. Based on this criterion, 139 sequences were included in further analysis, of a total of 160 sequenced clones.

Phylogenetic Tree and Sequence Similarity

All pseudogenes showed a higher similarity to the functional *PHYO* gene (ranging from 91.8% for *psPHY3* to 95.0% for *psPHY1*) than to the other members of the *PHY* gene family (*PHYP* and *PHYN* genes). Nucleotide divergence (K) values between the *psPHY* and the *PHYO* sequences indicate that *psPHY3* and *psPHY4* are more divergent from *PHYO* than *psPHY1* and *psPHY2* (Table 3). The NJ and UPGMA tree topologies also indicate that the *psPHY* sequences are most similar to the *PHYO* gene (Fig. 2a and b). *psPHY* sequences clustered in four groups, each of which corresponds to one of the pseudogene subfamilies I identified based on their similarity to *PHYO* (*psPHY1* to *psPHY4*). The NJ algorithm can incorporate different rates of evolution in different branches. Phylogenetic analysis based on the NJ model supported the hypothesis that *psPHYs* comprise a monophyletic group that has diverged from the *PHYO* subfamily (Fig. 2a), and the phylogenetic tree inferred using MrBayes (not shown) supports the NJ tree topology. In contrast, the UPGMA algorithm assumes that evolutionary rates have been constant over time and at each branch of the tree, and the UPGMA algorithm tree obtained in this study indicates that

Table 3 Nucleotide divergence at synonymous sites (K_s) among psPHY and PHYO sequences amplified from individual Läpinjärvi (standard errors in parentheses)

	PHYO	psPHY1	psPHY2	psPHY3	psPHY4
PHYO	—	0.041 (0.012)	0.047 (0.013)	0.079 (0.018)	0.069 (0.018)
psPHY1		—	0.024 (0.008)	0.098 (0.020)	0.088 (0.019)
psPHY2			—	0.094 (0.020)	0.096 (0.020)
psPHY3				—	0.015 (0.006)
psPHY4					—

the psPHY3/4 branch diverged from the PHYO/psPHY1/2 precursor (Fig. 2b). Both trees indicate that four duplication events occurred after the divergence of PHYN, an early subtype of the PHYO subfamily.

Neutrality Tests

Nucleotide Diversity

The alignment of psPHY sequences within each pseudogene subfamily is shown in Fig. 1. The populations were not divergent ($F_{ST} = 0$), therefore synonymous and replacement nucleotide diversity per locus parameters (π_s and π_a , respectively) were computed using the pooled sequence data from all of the individuals (Table 4). Searches for clustering by population using BAPS4 software also supported the absence of substructure for the studied

Table 4 Nucleotide diversity at the synonymous (π_s) and replacement sites (π_a) for each locus (standard errors in parentheses)

Locus	Synonymous, π_s	Replacement, π_a	π_s/π_a
PHYO ($n = 12$)	0.001 (0.0001)	0.0002 (0.0003)	0.2
psPHY1 ($n = 10$)	0.010 (0.003)	0.008 (0.002)	0.8
psPHY2 ($n = 7$)	0.015 (0.006)	0.011 (0.002)	0.7
psPHY3 ($n = 5$)	0.012 (0.005)	0.010 (0.002)	0.8
psPHY4 ($n = 11$)	0.007 (0.002)	0.008 (0.002)	1.1

Note: n = number of sequences analyzed

loci. π values were higher at each psPHY locus than at the PHYO locus, but none of them were significantly different. π_a/π_s ratios at psPHY were 3.5- to 5.5-fold higher than at PHYO and closer to the expected value (unity) under neutral evolution (in such cases the nucleotide diversity at synonymous and replacement sites is expected to be equal).

Mutation Distribution

The observed distributions of synonymous and replacement mutations are shown in Fig. 3a. The only locus at which more synonymous than replacement substitutions were found is the PHYO locus. Departure from neutrality was significant ($p < 0.05$) at this locus. At the psPHY loci the distributions of synonymous vs replacement mutation were consistent with expectations for unconstrained sequences, and there was no evidence of departure from neutrality. The pattern of point mutation distributions based on codon position classifications is shown in Fig. 3b. None of the loci (including PHYO) showed departure from neutrality.

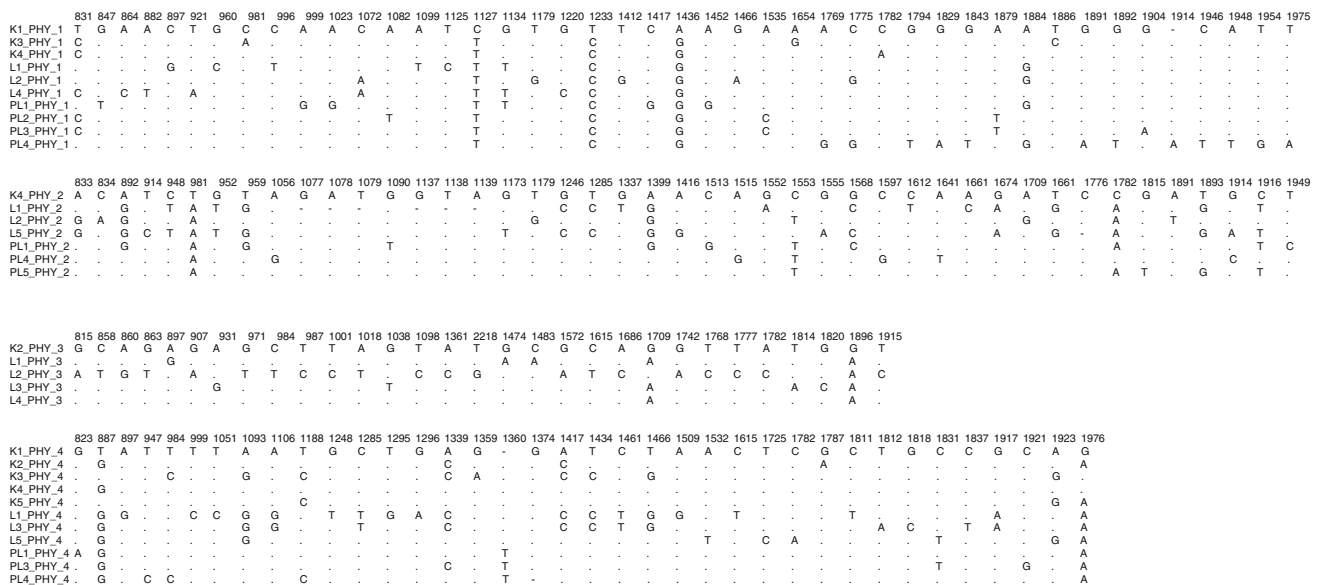


Fig. 1 Nucleotide polymorphism within each psPHY locus in Scots pine

The most frequent replacement substitutions at psPHY loci did not result in replacement by similar amino acids; instead, they generally had high Grantham's distances (48.7 for psPHY1 to 80.5 for psPHY4). In contrast, the Grantham's distance between substituted amino acids at the PHYO loci was moderate (30).

Codon Model

The hierarchical Bayesian method of Huelsenbeck and Ronquist implemented in MrBayes was applied to search for codons under positive selection. MCMC convergence was established informally, through visual inspection of the MCMC traces from the model parameters and functions of the parameters such as the likelihood function. Convergence seemed to be reached after the first 30,000 iterations. The values sampled during the iterations prior to convergence (burn-in values) were discarded. My inferences were based on a sample of posterior draws constructed by subsampling the remaining postconvergence MCMC samples every 100th iteration. For the PHYO, no codon was found to have evolved under positive selection; instead p_1 (the frequency of ω_1) was 0.99, supporting negative selection.

Among-Species Divergence (K)

Adaptive evolution at the protein level is widely inferred from the ratio (ω) of nonsynonymous divergence between species (K_a) to synonymous divergence between species (K_s) at functional gene loci. Under the neutral evolution model, the amount of divergence between sequences from two species is given by

$$k = 2\mu t + 4N_e\mu$$

where μ is the mutation rate, t is the time since the species diverged, and N_e is the ancestral effective population size. If $\omega = 1$, there is no evidence for selection; if $\omega < 1$, there is evidence for negative selection; and if $\omega > 1$, there is evidence for positive selection. The K_a/K_s ratio (ω) at PHYO is <1 (0.17), indicating negative selection (Table 5). K_a/K_s ratios are also <1 at psPHYs, but the values are still at least 2.5 times higher than the value obtained for PHYO.

Discussion

There is published evidence indicating that multigene families, composed of both functional genes and pseudogenes, account for significant proportions of the large, complex genomes of conifers (Ahuja and Neale 2005). However, most of these indications have been obtained using cDNA probes in Southern (DNA) hybridizations, and

Table 5 Rates of synonymous and replacement substitutions per year/per generation at each locus vs *Picea* PHYO (U60264) (standard errors in parentheses)

		K_s	μ_s^a	K_a	K_a/K_s
PHYO	L	0.121 (0.024)	0.43	0.021 (0.005)	0.17
	H	0.128 (0.025)	0.46	0.022 (0.005)	0.17
psPHY1	L	0.126 (0.024)	0.45	0.054 (0.008)	0.42
	H	0.144 (0.026)	0.51	0.065 (0.009)	0.46
psPHY2	L	0.138 (0.025)	0.49	0.066 (0.009)	0.49
	H	0.158 (0.027)	0.56	0.077 (0.010)	0.49
psPHY3	L	0.163 (0.029)	0.58	0.098 (0.012)	0.60
	H	0.167 (0.029)	0.59	0.102 (0.012)	0.62
psPHY4	L	0.159 (0.028)	0.57	0.087 (0.011)	0.54
	H	0.164 (0.029)	0.58	0.094 (0.012)	0.57

Note: The time of divergence between *Pinus sylvestris* and *Picea abies* based on the fossil record (140 MYA [Miller 1977]) was used to estimate the rate of substitutions per site and per year. L and H represent the lowest and highest rate of substitution within each locus, respectively

^a Substitutions per site per 10^{-9} year ($\mu = K/2T$)

studies on the nature and evolution of the gene families involved are still scarce.

In this work, I compare sequences of four phytochrome pseudogenes (psPHY) and the functional PHYO, PHYN, and PHYP genes in Scots pine to assess the evolutionary implications of differences among them. Each of the examined psPHY sequences is 1716 bp long and shows deletions which probably alter or completely prevent their expression. All of the psPHY sequences show higher homology with the functional PHYO gene than with the other sequences. NJ and UPGMA trees indicate that psPHY subfamilies cluster into two subfamilies (psPHY1/2 and psPHY3/4). As expected, the nucleotide divergence at synonymous sites (K_s) is lower within each subfamily than between the subfamilies (Table 3). The NJ algorithm can incorporate different rates of evolution, which are graphically represented as differences in branch lengths. The NJ tree indicates that the psPHY3/4 subfamily has evolved more rapidly than the psPHY1/2 subfamily, since it has longer branch lengths (Fig. 2a). Although evolutionary constraints may differ among genes (McInerney 2006), pseudogenes are expected a priori to be neutral and, thus, should evolve without constraint. Therefore, the evolutionary rate of both psPHY subfamilies is expected to be the same, in the absence of other evolutionary events, like gene conversion or hitchhiking events. Other possible explanations for the apparent difference in their evolutionary rates are that the ancestor of the psPHY3/4 subfamily may have become a pseudogene earlier than the ancestor of the psPHY1/2 subfamily, or psPHY3/4 may have diverged from a functional ancestor with a higher

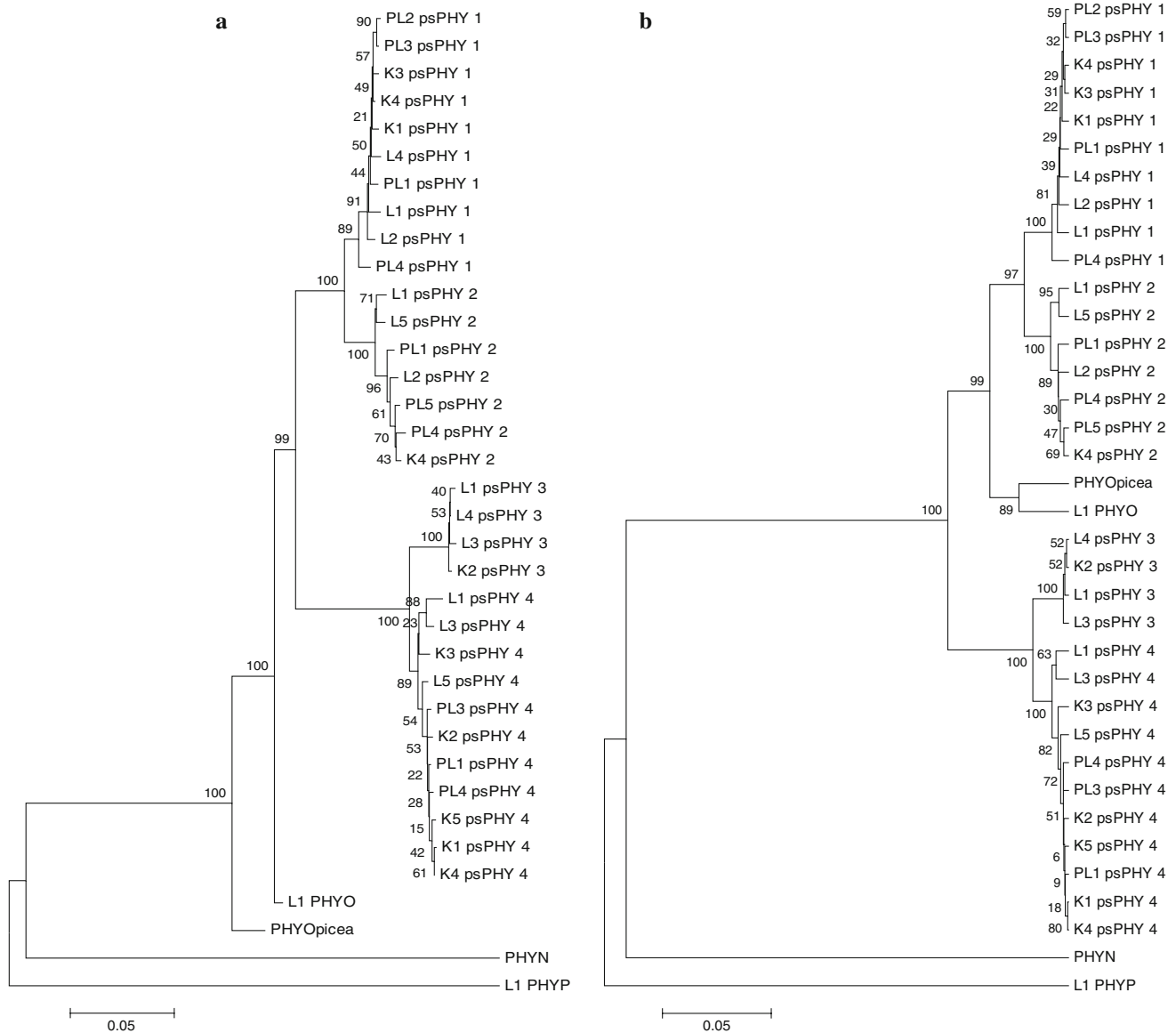


Fig. 2 (a) Neighbor-joining phylogenetic tree. (b) UPGMA phylogenetic tree. GenBank accession numbers for L1_PHYO, L1_PHP, and PHYO picea: EU203180, EU203157, and U60264, respectively.

The letters K, L, and PL refer to individuals from the Kolari, Läpinjärvi, and Puebla de Lillo populations, respectively. Bootstrap values are shown on the tree branches

evolutionary rate than the psPHY1/2 functional ancestor. In contrast to the NJ algorithm, UPGMA does not allow for differences in evolutionary rates among branches; instead it indicates simply that the psPHY3/4 subfamily is older than psPHY1/2 (Fig. 2b). Mr Bayes phylogenetic analysis is a tree-searching or discrete data method. This type of method is potentially more informative than distance-matrix methods, since evolutionary changes at each specific site are considered (Baldauf 2003). MrBayes phylogenetic analysis supports the same tree topology as NJ, indicating that the psPHY3/4 family has evolved more rapidly than the psPHY1/2 family. Again, there are two possible explanations for this observation: earlier

pseudogenization or departure from a more rapidly evolving functional ancestor.

The results provide several lines of evidence supporting the a priori expectations of unconstrained evolution of the psPHY sequences and constrained evolution of the functional PHYO sequence: first, in the lower nucleotide diversity at synonymous and replacement mutations at PHYO (Table 4). In addition, point mutation distributions among codons and synonymous/replacement sites indicate that the evolution of the psPHYs has not departed from neutrality (Fig. 3a and b). In contrast, constrained evolution at PHYO is supported by significant departure from neutrality of its synonymous vs replacement mutation

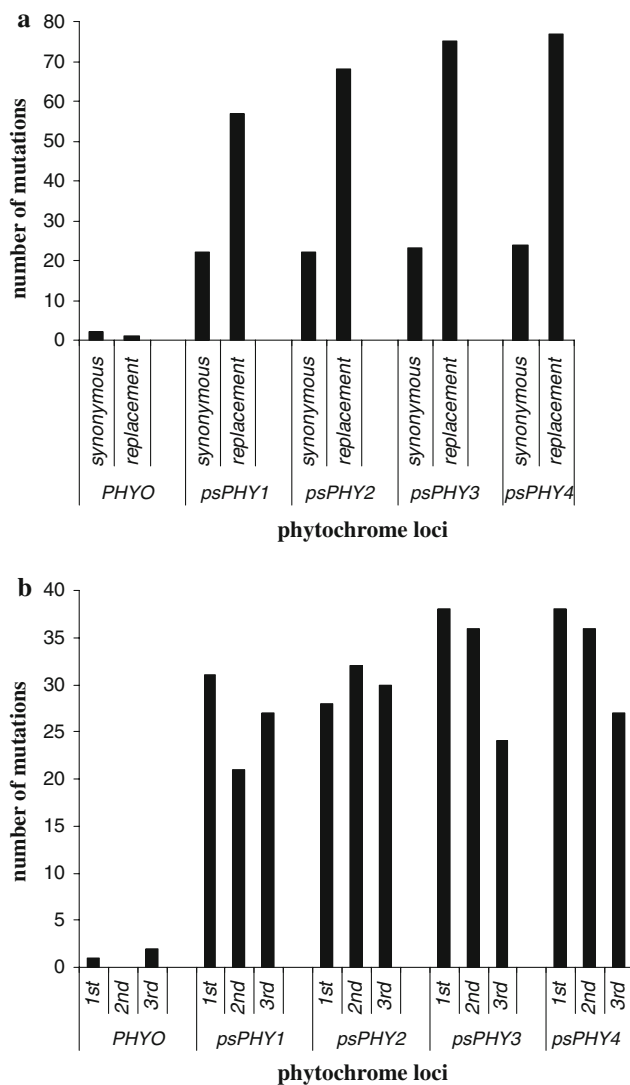


Fig. 3 Point mutation distributions based on synonymous vs replacement substitutions (a) and codon positions (b). Synonymous and replacement sites at each *psPHY* were identified from the alignment of each *psPHY* subfamily vs the PL1_ *PHYO* functional gene

distribution (Fig. 3b), although the absence of detected departure from neutrality has weak statistical support (only three mutations vs three codon categories). The Grantham's distance and AT-level data provide further evidence of the difference in evolutionary patterns between *PHYO* and the *psPHYs*. Grantham's distance between replacement amino acids encoded by the *PHYO* sequences and those in the reference sequence is small (30), indicating that all of the substitutions result in similar amino acids, as expected in a functional gene. In contrast, the most frequent substitutions at *psPHY* result in replacements with high Grantham's distances (48.7 at *psPHY1* to 80.5 at *psPHY4*), indicative of nonconservative changes. Randomly distributed mutations are predicted to produce higher proportions of nonconservative changes. This random pattern of mutations is also

indicated by an increase in AT levels, as found in all *psPHYs*. Increases in AT-levels have also been described in other unconstrained DNA sequences, e.g., *cdc2* pseudogenes in *Picea* (Kvarnheden et al. 1995) and the Helena retrotransposon in *Drosophila* (Petrov and Hartl 2000).

The hierarchical Bayesian analysis of the *PHYO* gene resulted in a p_1 value (frequency of ω_1) of 0.99, clearly supporting negative selection. Convergence was achieved in this study without problems after the first 30,000 iterations, as expected for well-diverged sequences. The K_a/K_s ratio (ω) data also highlight the differential evolution of *psPHYs* and *PHYO*, since *psPHYs* ω parameter values are at least 2.5 times higher than those of *PHYO*. An interesting aspect of the ω parameter values obtained for the *psPHY* subfamily is that they deviate significantly from the expected value of one (indicative of neutral evolution). Indeed, at *PHYO* and *psPHYs* the ω parameter is <1 , supporting negative selection. This result would be expected if the *psPHYs*' pseudogenization is still recent, in which case the sequences would still contain signs of their former functionality. In other words, signs of previously constrained evolution may still be being erased in the *psPHY* sequences. The signs of former functionality are clearest in the *psPHY1/2* subfamily, for which values of the ω parameter are lower than in the *psPHY3/4* subfamily. Lynch and Conery (2000) proposed that the fate of a large proportion of genes after duplication is to become silenced due to the accumulation of unconstrained mutations. This process of gene silencing due to degeneration is called nonfunctionalization, which is one possible pathway of gene evolution following duplication (Watterson 1983). The process of nonfunctionalization is gradual, and duplicated genes will still retain signs of their former functionality in their nucleotide sequence for a certain time. Various evolutionary factors, including, among other things, the mutation rate and gene conversion, will determine how long it takes for all signs of former constrained evolution to be erased. An example of nonfunctionalization of a phytochrome gene (encoding phyD) has been described in *Arabidopsis* (Aukerman et al. 1997). The cited authors proposed that loss of its function was nonlethal since phyD and phyB have overlapping functions.

Further interesting aspects of gene family evolution that are not fully understood are the processes affecting or regulating the size of gene families. In conifers, the evidence suggests that gene families are substantially larger than in angiosperms. Theoretically, a lack of constraint on genome sizes, long generation times, and low fixation rates (characteristic of species with large populations) are likely to lead to low rates of removal of nonsense sequences (i.e., pseudogenes). These theoretical expectations of unconstrained accumulation of pseudogenes appear to be met in Scots pine, and the *psPHYs* show several indications of a

lack of constraint on gene family size. First, although deletions are more frequent, insertions were also detected in the psPHY sequences (Table 1). Two pseudogenes of the sHSP gene family in *Picea abies* described by Schubert et al. (2002) also show insertions. Second, more substitutions than deletions were found in the psPHY sequences, in accordance with trends found in *cdc2* pseudogenes in *Picea abies* by Gernandt et al. (2001), which is a characteristic feature of unconstrained genomes (Petrov et al. 1996; Petrov et al. 2000).

In conclusion, this work presents novel findings regarding the complexity and evolution of the phytochrome gene family in Scots pine. The acquired data and the recent deposition of another phytochrome pseudogene sequence, homologous to *PHYN*, in GenBank (AJ286646) indicate that the composition and evolution of the phytochrome gene family are more complex than previously realized. Our knowledge of gene family evolution in conifers is still quite limited, but ongoing efforts to elucidate aspects of conifer genomics and genetics should provide interesting advances in our understanding of the processes involved, the ultimate constraints, and the evolutionary implications of changes in the sizes of gene families in both conifers specifically and organisms generally.

Acknowledgments I thank Professor Outi Savolainen for useful comments on the manuscript, and I gratefully acknowledge the financial support from the European Science Foundation grant and the Marie Curie fellowship (QLK5-CT-2000-51233) provided under the 5th Framework Programme and Bioscience and Environmental Research Council.

References

- Ahuja MR, Neale DB (2005) Evolution of genome size in conifers. *Silvae Genetica* 54:126–137
- Akashi H (1997) Distinguishing the effects of mutational biases and natural selection on DNA sequence variation. *Genetics* 147:1989–1991
- Akashi H (1999) Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics* 136:927–935
- Aukerman MJ, Hirschfeld M, Wester L, Clack T, Amasino RM, Sharrock RA (1997) A deletion in the PHYD gene of the *Arabidopsis* Wassilewskija ecotype defines a role for phytochrome D in far red/red light sensing. *Plant Cell* 9:1317–1326
- Balakirev ES, Ayala FJ (2003) Pseudogenes: are they “junk” or functional DNA? *Annu Rev Genet* 37:123–151
- Baldauf SL (2003) Phylogeny for the faint of heart: a tutorial. *Trends Genet* 19:345–351
- Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics* 134:1289–1303
- Dvornyk V, Sirviö A, Mikkonen M, Savolainen O (2002) Low nucleotide diversity at the *pall* locus in the widely distributed *Pinus sylvestris*. *Mol Biol Evol* 19:179–188
- Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondria DNA restriction data. *Genetics* 131:479–491
- García-Gil MR, Mikkonen M, Savolainen O (2003) Nucleotide diversity at two phytochrome loci along a latitudinal cline in *Pinus sylvestris*. *Mol Ecol* 12:1195–1206
- Gernandt DS, Liston A, Pineiro D (2001) Variation in the nrDNA ITS of *Pinus cembra*: implications for molecular systematics studies of pine species complexes. *Mol Phylogenet Evol* 21:449–467
- Higgins DG, Sharp PM (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* 73:237–244
- Huelsenbeck JP, Ronquist F (2001) MRBAYES: bayesian inference of phylogenetic trees. *Bioinformatics* 17:754–755
- Hughes AL (2000) Adaptive evolution of genes and genomes. Oxford University Press, New York
- Kaplan NL, Hudson RR, Langley CH (1989) The hitchhiking effect revisited. *Genetics* 123:887–899
- Kinlaw CS, Neale DB (1997) Complex gene families in pine genomes. *Trends Plant Sci* 2:356–359
- Kvarnheden A, Albert VA, Engström P (1995) Molecular evolution of *cdc2* pseudogenes in spruce (*Picea*). *Plant Mol Biol* 36:767–774
- Li WH (1997) Molecular evolution. Sinauer Associates, Sunderland, MA
- Li WH, Gojobori T, Mei M (1981) Pseudogenes as a paradigm of neutral evolution. *Nature* 292:237–239
- Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290:1151–1155
- Mathews S, Donoghue MJ (1999) The root of angiosperm phylogeny inferred from duplicate phytochrome genes. *Science* 286:947–950
- McDonald JH, Kreitman M (1991) Adaptive evolution of the “Adh” locus in “*Drosophila*”. *Nature* 351:652–654
- McInerney JO (2006) The causes of protein evolutionary rate variation. *Trends Ecol Evol* 21:230–232
- Mighell AJ, Smith NR, Robinson PA, Markham AF (2000) Vertebrate pseudogenes. *FEBS Lett* 468:109–114
- Miller CN (1977) Mesozoic conifers. *Bot Rev* 43:217–280
- Murray BG (1998) Nuclear DNA amounts in gymnosperms. *Ann Bot* 82 (Suppl A):3–15
- Neff MM, Fankhauser C, Chory J (2000) Light an indicator of time and place. *Genes Dev* 14:257–271
- Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York
- Nei M, Kumar S (2000) Molecular evolution and phylogenetics. Oxford University Press, New York
- Nielsen R (2001) Statistical tests of selective neutrality in the age of genomics. *Heredity* 86:641–647
- Petrov DA, Hartl DL (1999) Patterns of nucleotide substitution in *Drosophila* and mammalian genomes. *Proc Natl Acad Sci USA* 96:1475–1479
- Petrov DA, Lozovskaya ER, Hartl DL (1996) High intrinsic rate of DNA loss in *Drosophila*. *Nature* 384:346–349
- Petrov DA, Sangster TA, Johnston S, Hartl DL, Shaw KL (2000) Evidence of DNA loss as a determinant of genome size. *Science* 287:1060–1062
- Reed DH, Nicholas AC, Stratton GE (2007) Genetic quality of individuals impacts population dynamics. *Anim Conserv* 10:275–283
- Schubert R, Manteuffel R, Eich J, Häger KP (2002) Molecular characterization and evolution of the cytosolic class II 17.0 kDa small heat-shock protein gene family from *Picea abies* (L.). *Plant Sci* 163:1–12
- Skinner JS, Timko MP (1998). Lobloily pine (*Pinus taeda* L) contains multiple expressed genes encoding light-dependent NADPH:protochlorophyllide oxidoreductase (POR). *Plant Cell Physiol* 39:795–806

- Suzuki Y, Gojobori T (1999) A method for detecting positive selection at single amino acid sites. *Mol Biol Evol* 16:1315–1328
- Swanson WJ (2003) Adaptive evolution of genes and gene families. *Curr Opin Genet Dev* 13:617–622
- Wakasugi T, Tsudzuki J, Nakashima K, Tsudzuki T, Sugiura M (1994) Loss of all Ndh genes as determined by sequencing the entire chloroplast genome of the black pine *Pinus thumbergii*. *Proc Natl Acad Sci USA* 91:9794–9798
- Watterson GA (1983) On the time for gene silencing at duplicated loci. *Genetics* 105:745–766
- Wei XX, Wang XQ (2004) Recolonization and radiation in *Larix* (Pinaceae): evidence from nuclear ribosomal DNA paralogues. *Mol Ecol* 13:3115–3123
- Wright S (1969) *Evolution and the genetics of populations, vol 2. The theory of gene frequencies*. University of Chicago Press, Chicago
- Wright F (1986) The “effective number of codons” used in a gene. *Gene* 87:23–29
- Wyckoff GJ, Malcom CM, Vallender EJ, Lahn BT (2005) A highly unexpected strong correlation between fixation probability of nonsynonymous mutations and mutation rate. *Trends Genet* 21:381–385
- Xia X (1996) Maximizing transcription efficiency causes codon usage bias. *Genetics* 144:1309–1320
- Xia X (2001) *Data analysis in molecular biology and evolution*. Kluwer Academic, New York
- Yang Z, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. *TREE* 15:496–503