Rates and patterns of molecular evolution in inbred and outbred Arabidopsis

Stephen I. Wright, Béatrice Lauga, Deborah Charlesworth

▶ To cite this version:

Stephen I. Wright, Béatrice Lauga, Deborah Charlesworth. Rates and patterns of molecular evolution in inbred and outbred Arabidopsis. Molecular Biology and Evolution, 2002, 19 (9), pp.1407-1420. 10.1093/oxfordjournals.molbev.a004204 . hal-03518208

HAL Id: hal-03518208
https://univ-pau.hal.science/hal-03518208

Submitted on 9 Jan 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Rates and Patterns of Molecular Evolution in Inbred and Outbred *Arabidopsis*

Stephen I. Wright, Beatrice Lauga,[1] *and Deborah Charlesworth*

Institute of Cell, Animal, and Population Biology, Ashworth Laboratories, University of Edinburgh

The evolution of self-fertilization is associated with a large reduction in the effective rate of recombination and a corresponding decline in effective population size. If many spontaneous mutations are slightly deleterious, this shift in the breeding system is expected to lead to a reduced efficacy of natural selection and genome-wide changes in the rates of molecular evolution. Here, we investigate the effects of the breeding system on molecular evolution in the highly self-fertilizing plant *Arabidopsis thaliana* by comparing its coding and noncoding genomic regions with those of its close outcrossing relative, the self-incompatible *A. lyrata*. More distantly related species in the Brassicaceae are used as outgroups to polarize the substitutions along each lineage. In contrast to expectations, no significant difference in the rates of protein evolution is observed between selfing and outcrossing Arabidopsis species. Similarly, no consistent overall difference in codon bias is observed between the species, although for low-biased genes *A. lyrata* shows significantly higher major codon usage. There is also evidence of intron size evolution in *A. thaliana*, which has consistently smaller introns than its outcrossing congener, potentially reflecting directional selection on intron size. The results are discussed in the context of heterogeneity in selection coefficients across loci and the effects of life history and population structure on rates of molecular evolution. Using estimates of substitution rates in coding regions and approximate estimates of divergence and generation times, the genomic deleterious mutation rate (U) for amino acid substitutions in Arabidopsis is estimated to be approximately 0.2–0.6 per generation.

Introduction

The genomes of organisms vary in their rates and patterns of molecular evolution, including patterns of base composition (Tarrio, Rodriguez-Trelles, and Ayala 2001), rates of protein evolution (Keightley and Eyre-Walker 2000), and rates of insertion and deletion in noncoding DNA (Petrov, Lozovskaya, and Hartl 1996). Two primary explanations have been proposed to account for these patterns. First, mutation biases and rates may differ among species and genomic regions. Alternatively, there may be differences in the strength or efficacy of natural selection (or both) between genomic regions and species. The relative importance of the effects of mutation versus selection, and the strength and direction of any selective effects, are largely unclear (Rodriguez-Trelles, Tarrio, and Ayala 1999; Marais, Mouchiroud, and Duret 2001; Smith and Eyre-Walker 2001).

The rate of molecular evolution for mutations subjected to selection is not determined solely by mutation rates but by $N_e s$, the product of the effective population size and selection coefficient (s) (Kimura 1983, p. 45). If, as proposed by the nearly neutral theory, a large fraction of mutations are slightly deleterious (with selection coefficients of the order of $1/N_e$), differences in the effective population sizes can have substantial effects on their evolution (Ohta 1992). A lower effective size increases the role of drift relative to selection in determining the fate of mutations, increasing the frequencies and fixation rates of slightly deleterious mutations. Evidence for population size effects on rates of protein evolution in mammals (Keightley and Eyre-Walker 2000), Drosophila (DeSalle and Templeton 1988; Ohta 1993), and birds (Johnson and Seger 2001) are broadly consistent with the predictions of the nearly neutral model. In addition, the evidence for effects of population size on differences in codon usage bias between species (Akashi 1996, 1999; McVean and Vieira 2001) suggests that this evolutionary process is relevant to the evolution of codon bias.

The efficacy of selection can also differ among genomes and genome regions with different rates of crossing-over, through the effects of natural selection on the fate of linked segregating sites (Hill and Robertson 1966). These effects (hereafter ''hitchhiking'') are classified into three processes according to the direction and strength of the selection driving the process: directional selection on advantageous mutations (''selective sweeps,'' $N_e s \gg 1$; reviewed by Braverman et al. 1995; Barton 2000; Kim and Stephen 2002), strong selection against deleterious mutations (''background selection,'' $N_e s \ll -1$; Charlesworth B, Morgan, and Charlesworth D 1993), and weak Hill-Robertson interference between many weakly selected ($N_e s \sim 1$) sites segregating simultaneously (McVean and Charlesworth 2000; Tachida 2000). These processes lower N_e and thus reduce both polymorphism and efficacy of selection. The predicted correlations between crossing-over and neutral genetic variability are observed in several outbreeding species, including Drosophila (reviewed by Langley et al. 2000), plants (see Dvorak, Luo, and Yang 1998), and mammals (see Nachman 2001). Effects of recombination rates on the efficacy of selection, however, remain uncertain. Such effects have been proposed as the explanation for correlations between recombination and codon bias in

[1] Present address: Laboratoire d'Ecologie Moléculaire, Université de Pau et des Pays de l'Ardour, UFR Sciences et Techniques, Pau cedex, France.

Key words: inbreeding, *Arabidopsis*, deleterious mutation, codon bias, intron size.

Address for correspondence and reprints: Stephen Wright, ICAPB, Ashworth Laboratories, University of Edinburgh, King's Buildings, West Mains Road, Edinburgh EH9 3JT.
E-mail: stephen.wright@ed.ac.uk.

Mol. Biol. Evol. 19(9):1407-1420. 2002
© 2002 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

Drosophila (Kliman and Hey 1993; Comeron, Kreitman, and Aguade 1999), but it is difficult to exclude the possibility of regional differences in mutational biases or biased gene conversion (Marais, Mouchiroud, and Duret 2001).

The breeding systems of populations may also affect their molecular evolution (Charlesworth and Wright 2001). Mutational biases and selection pressures are likely to be similar at orthologous loci in close relatives, so related inbred and outbred plants may be useful for studying the effects of effective size and recombination. Under strict neutrality, the effective size of a population depends on the inbreeding coefficient (*F*); N_e is reduced by inbreeding and is halved with complete self-fertilization (Pollak 1987; Nordborg 2000). Inbreeding also reduces the effective recombination rates because crossover events rarely occur between mutations segregating in the population, since homozygosity is high. If natural selection acts at some sites, N_e, and thus the diversity, of inbred populations can be reduced by more than half (Charlesworth B, Morgan, and Charlesworth D 1993). With high levels of inbreeding, hitchhiking is expected to reduce the effective size not only of the nuclear genome, but also the organelle genomes, which become effectively linked to the nuclear genome (Charlesworth B, Morgan, and Charlesworth D 1993; Graustein et al. 2002). Such populations would also have increased expected rates of fixation of slightly deleterious mutations and reduced chances of fixation of beneficial mutations (Charlesworth 1994). For partially recessive mutations, high levels of homozygosity in inbred populations may partially oppose these effects (Charlesworth 1992). However, the effect of reduced N_e is likely to predominate because the fixation of deleterious mutations is not greatly affected by the dominance coefficient (Charlesworth 1994). With very high selfing rates, populations may also experience Muller's ratchet, the stochastic loss of individuals free from deleterious mutations (Heller and Maynard Smith 1979; Charlesworth D, Morgan, and Charlesworth B 1993), with concomitant fixation of deleterious mutations (Charlesworth B and Charlesworth D 1997; Gordo and Charlesworth 2000). Finally, the weedy life history of many inbreeding plant populations leads to strong population structure, highly variable population size, low pollen migration rates, and frequent extinction and recolonization. All these processes can further reduce the effective population size (Whitlock and Barton 1997) and hence the neutral variability (Pannell and Charlesworth 2000). Many studies on allozyme polymorphism (Hamrick and Godt 1990) and work on the nucleotide variability in species of the genera *Leavenworthia* (Liu, Zhang, and Charlesworth 1998; Liu, Charlesworth, and Kreitman 1999), *Lycopersicon* (Baudry et al. 2001), and *Arabidopsis* (Bergelson et al. 1998; Savolainen et al. 2000) have shown the predicted low diversity within selfing populations. Inbred populations thus seem to have highly reduced effective population sizes, and there is the potential for accumulation of weakly deleterious mutations.

Here, we compare the gene structure and substitution rates between the highly self-fertilizing *A. thaliana* (Abbott and Gomes 1989) and its self-incompatible outcrossing congener *A. lyrata* to investigate the effect of the breeding system on molecular evolution. Several studies have found an unexpectedly high level of amino acid polymorphism in species-wide samples of *A. thaliana* (Kawabe et al. 1997; Purugganan and Suddith 1998, 1999; Bustamante et al. 2002), which has been interpreted as reflecting reduced efficacy of purifying selection due to low effective population size. *Arabidopsis thaliana* also has low codon usage bias, again consistent with the reduced efficacy of selection. The weak correlation between the estimated gene expression levels and codon bias suggests that selection on synonymous sites may be weaker, and the average level of codon bias much smaller, in *A. thaliana* than that in *Drosophila melanogaster* (Duret and Mouchiroud 1999). Although these broadscale differences are consistent with the inbreeding mating system of this species causing a reduction in the efficacy of selection, comparisons of rates of amino acid substitution and levels of codon bias with related outcrossing species are important to test this hypothesis.

Methods

Sequence Information

Table 1 shows the genes studied, the species from which they were available, and the source of the sequences. Most sequences were partial coding regions, including multiple exons and introns. The sample includes 23 nuclear loci distributed across all five *A. thaliana* chromosomes and a single locus (*matK*) from the chloroplast genome. Sequences from *A. thaliana* were extracted from GenBank (National Center for Biotechnology Information, NCBI, http://www.ncbi.nlm.nih.gov) using either the complete genome sequence from the Columbia ecotype or a sequence from a published population survey. All nuclear genes previously sequenced in *A. lyrata* were extracted from GenBank, along with their orthologs in *A. thaliana*, with the exception of the putative self-incompatibility locus *SRK*, which is likely to be a pseudogene in *A. thaliana* (Kusaba et al. 2001). The *A. lyrata* sequences are derived from different source populations, including representatives from the European subspecies *petraea*, the Japanese subspecies *kawasakiana*, and the North American subspecies *lyrata*. Seven additional single-copy loci were selected for sequencing in *A. lyrata* (see subsequently). All *A. lyrata* loci were submitted to a BLAST search (Altschul et al. 1990, 1997; http://www.ncbi.nlm.nih.gov/BLAST) to confirm that the locus has a single clear ortholog in *A. thaliana* and to identify sequences available from the closest outgroup species in the Brassicaceae (Koch, Haubold, and Mitchell-Olds 2000, 2001). Two of the loci sequenced for this study (*HAT4* and *EnCoH1*) were selected from a 28-kb genomic region of the *A. thaliana* chromosome 4 that has been sequenced in the outgroup species *Capsella rubella* (Acarkan et al. 2000), which belongs to the sister group to *A. thaliana* and *A. lyrata* (Koch, Haubold, and Mitchell-Olds 2000, 2001). For an additional locus, *ABC1At*, the orthologous region was sequenced in *C. rubella* using DNA provided by H. Hurka. In total, seven loci in the analysis have an outgroup sequence from this sister group, including *C. rubella* and *A. glabra*. For an additional eight genes, sequences from more diverged outgroup species from the genera *Brassica*, *Leavenworthia*, *Matthiola*, or *Raphanus* were used. Because single sequences from each species are used in the comparisons, estimates of substitution rates along each lineage will include segregating polymorphisms as well as fixed differences between species. However, this should not bias the results, provided the species-wide coalescence times are similar in the two species. Given that the species-wide estimates of silent polymorphism do not appear to be very different in *A. thaliana* and *A. lyrata* (Savolainen et al. 2000) (despite highly reduced within-population variability in *A. thaliana*), this is reasonable in the present case.

Table 1 also shows for each locus the maximum number of matches with expressed sequence tags (ESTs) from the *A. thaliana* EST projects available in GenBank (obtained from G. Marais, personal communication). This index has been normalized for differences in total EST number across the libraries by dividing each value by the total number of ESTs in the source library. The number of EST matches is often used as a measure of the expression level and correlates with levels of codon bias in *A. thaliana* (Duret and Mouchiroud 1999).

DNA Extraction, Gene Amplification and Sequencing

DNA was extracted from *A. lyrata* individuals using the CTAB protocol (Junghans and Metzlaff 1990). PCR primers were designed using the *A. thaliana* sequence information and amplified in *A. lyrata* for 30 cycles consisting of 1 min denaturing at 95°C, 30 s annealing at 55°C, and 2 min extension at 72°C. PCR primer sequences are: Sc-ADH-F 5′-GGCATTCCTCCAGCGAC-3′, Sc-ADH-R 5′-CTTCCGTCGTCGTCTCTTC-3′; EnCoA-F 5′-CTGGTCGGTTACTTTTGTCG-3′, EnCoA-R 5′-CCTGTCACCAAAAATGCTATT-3′; HAT4-F 5′-CGTGAACAGACCACCGTC-3′, HAT4-R 5′-AGCGTCAAAAGTCAAGCCGT-3′; ABAp1-F 5′-CAAGCACAAAACCAACAGCC-3′, ABAp1-R 5′-CAAACCCATCTCGTGTCACC-3′; ABC1At-F 5′-CTTTACCAGG CTCGTCAATG-3′, ABC1At-R 5′-CATCACATCAGCACCTTGAC-3′; ETR1-F 5′-AGACCAAAGCTCATGCATTTCT-3′, ETR1-R 5′-TGTTGACTCATGAGATTAGAAGCA-3′; FKA1-F 5′-CCTGGATTCCTCAAAGCTCC-3′, FKA1-R 5′-TCCCAAATGCTCATGATCTG-3′. PCR fragments were cloned into the PCR 2.1 vector using the TA cloning kit (Invitrogen Life Technologies), and at least five clones were sequenced using the ABI automated sequencing facilities at the Institute of Cell, Animal and Population Biology, University of Edinburgh. PCR primers, the Universal M13 primers, as well as several internal sequencing primers were used in sequencing. In all cases, sequence analysis of clones indicated that the genes were single copy in *A. lyrata*, with one or two haplotypes identified per individual, with the exception of several clear PCR recombinants.

Rates of Protein Evolution

Single sequences from *A. thaliana*, *A. lyrata* and, where available, an outgroup species, were aligned using the CLUSTALW computer package (Thompson, Higgins, and Gibson 1994), and alignments were subsequently corrected by eye using the sequence editor GENEDOC (Nicholas, Nicholas, and Deerfield 1997). Pairwise estimates of K_a, the number of nonsynonymous substitutions per site, and K_s, the number of synonymous substitutions per site, were calculated for each gene using the program K-estimator version 5.5 (Comeron 1999), which uses the method of Comeron (1995) to estimate substitution rates. K-estimator was also used to obtain 95% confidence intervals for these estimates by Monte Carlo simulation. The total pairwise substitution rate was estimated by combining the sequences of all nuclear loci.

When an outgroup sequence was available, two approaches were utilized to estimate rates of synonymous and nonsynonymous substitutions in *A. thaliana* and *A. lyrata* independently. First, parsimony was used to estimate directly the numbers of nonsynonymous and synonymous substitutions in both lineages (Akashi 1996). For some sites, multiple substitutions precluded inference on the basis of parsimony, and these were excluded. For parsimony analysis, synonymous substitutions were counted only for codons that did not have a nonsynonymous difference. Differences in the total substitution rates between the lineages were assessed using Tajima's relative rate test (Tajima 1993). Second, rates of synonymous and nonsynonymous substitutions per site were estimated using the maximum likelihood method of the CODEML program in the PAML computer package (Yang 1997). This program estimates substitution rates, taking into account multiple substitutions per site, different rates of transitions and transversions, and effects of codon usage. Two models of sequence evolution were considered: (1) a model with a fixed K_a/K_s ratio across lineages, and (2) a model that allowed this ratio to differ for each species. Significance was assessed using the chi-square test with two degrees of freedom, where the chi-square statistic is $2(L_2 - L_1)$; L_2 is the log likelihood for the second (free ratios) model, whereas L_1 is the log likelihood for the model with fixed ratios (see Yang 1998). Standard errors (SE) were also estimated for the K_a/K_s ratios along each lineage using the PAML program, although these estimates provide only an approximate description of the likelihood surface (Yang 1997).

Comparisons of Codon Bias

Levels of codon bias and patterns of codon usage were examined for each locus in both Arabidopsis species. The GC content at third codon positions (GC_3) was used to measure codon usage bias. Chiapello et al. (1998) have shown that in *A. thaliana* GC_3 is highly correlated with the degree of biased codon usage and with gene expression levels. This measure is preferable to another standard measure of codon bias, ENC, because it measures more directly the frequency of pre-

Table 1
Genes Surveyed in An Analysis of Molecular Evolution in *Arabidopsis thaliana* and *A. lyrata*. Locus Names are Based on the *A. thaliana* Genome Project

Locus	Description	Species	Source[a] and Accession Number	Relative EST Abundance
ABAp1	Putative ABA-binding protein	*Arabidopsis thaliana*	AT4G01600	0
		Arabidopsis lyrata ssp. *petraea*	This study, AF494370	
ABC1At	ABC transporter	*Arabidopsis thaliana*	AT4G01660	51.41
		Arabidopsis lyrata ssp. *petraea*	This study, AF494371 (5′), AF494372 (3′)	
		Capsella rubella	This study, AF494373	
ADH1	Alcohol dehydrogenase 1	*Arabidopsis thaliana*	AT1G77120	33.33
		Arabidopsis ssp. *lyrata*	(Koch Haubold, and Mitchell-Olds 2000), AF110453	
		Capsella rubella	(Koch, Haubold, and Mitchell-Olds 2000), AF110435	
AOP1	2-Oxoglutarate–dependent dioxygenase 1	*Arabidopsis thaliana*	AT4G03070	0
		Arabidopsis lyrata ssp. *lyrata*	(Kliebenstein et al. 2001), AF417857	
AOP2	2-Oxoglutarate–dependent dioxygenase 2	*Arabidopsis thaliana*	(Kliebenstein et al. 2001), AF417858 (AT4G03060)	0
		Arabidopsis lyrata ssp. *lyrata*	(Kliebenstein et al. 2001), AF418239	
		Brassica oleraceaea	AY044425	
AOP3	2-Oxoglutarate–dependent dioxygenase 3	*Arabidopsis thaliana*	(Kliebenstein et al. 2001), AF417859 (AT4G03050)	2.55
		Arabidopsis lyrata ssp. *petraea*	(Kliebenstein et al. 2001), AF418280	
AP1	Apetala 1	*Arabidopsis thaliana*	AT1G69120	7.65
		Arabidopsis lyrata	(Lawton-Rauh, Buckler, and Purugganan 1999), AF143379	
		Brassica oleraceaea	(Anthony, James, and Jordan 1995), Z37968	
AP3	Apetala 3	*Arabidopsis thaliana*	AT3G54340	14.07
		Arabidopsis lyrata ssp. *petraea*	(Lawton-Rauh, Buckler, and Purugganan 1999, AF143380	
		Brassica oleraceaea	(Carr and Irish 1997), U67456	
CAUL	Cauliflower	*Arabidopsis thaliana*	AT1G26310	2.55
		Arabidopsis lyrata	(Purugganan and Suddith 1998), AF143381	
		Brassica rapa	(Li et al. 2000), AJ251300	
CHI	Chalcone flavone isomerase	*Arabidopsis thaliana*	AT3G55120	5.10
		Arabidopsis lyrata ssp. *petraea*	(Kuittinen and Aguadé 2000), AJ287322	
		Raphanus sativus	AF031921	
ChiA	Acidic endochitinase	*Arabidopsis thaliana*	AT5G24090	7.65
		Arabidopsis lyrata ssp. *kawasakiana*	(Kawabe et al. 1997), AB006072	
		Arabidopsis glabra	(Kawabe et al. 1997), AB006071	
CHS	Chalcone synthase	*Arabidopsis thaliana*	AT5G13930	98.49
		Arabidopsis lyrata ssp. lyrata	(Koch, Haubold, and Mitchell-Olds 2000), AF112100	
		Capsella rubella	(Koch, Haubold, and Mitchell-Olds 2000), AF112106	
EnCoH1	Enoyl-CoA hydratase	*Arabidopsis thaliana*	AT4G16800	0
		Arabidopsis lyrata ssp. *petraea*	This study, AF494369	
		Capsella rubella	(Acarkan et al. 2000), AJ400821	
ETR1	Ethylene receptor 1	*Arabidopsis thaliana*	AT1G66340	5.10
		Arabidopsis lyrata ssp. *lyrata*	This study, AF494374	
		Brassica oleraceaea	(Chen et al. 1998), AF047476	
F3H	Flavanone-3-hydroxylase	*Arabidopsis thaliana*	AT3G51240	16.67
		Arabidopsis lyrata ssp. *petraea*	(Aguadé 2001), AJ295607	
		Matthiola incana	(Britsch et al. 1993), X72594	
FAH1	Ferulate-5-hydroxylase	*Arabidopsis thaliana*	AT4G36220	5.10
		Arabidopsis lyrata ssp. *petraea*	(Aguadé 2001), AJ295586	
		Brassica napus	(Nair et al. 2000), AF214007	
FKA1	Fructokinase 1	*Arabidopsis thaliana*	AT2G31390	32.27
		Arabidopsis lyrata ssp. *lyrata*	This study, AF494374	
GL1	Glabrous 1	*Arabidopsis thaliana*	AT3G27920	0
		Arabidopsis lyrata ssp. *petraea*	(Hauser, Harr, and Schlotterer 2001), AF263720	
HAT4	Homeobox protein 4	*Arabidopsis thaliana*	AT4G16780	12.76
		Arabidopsis lyrata ssp. *petraea*	This study, AF494367	
		Capsella rubella	(Acarkan et al. 2000), AJ400821	

Table 1
(*Continued*)

Locus	Description	Species	Source[a] and Accession Number	Relative EST Abundance
matK	Maturase K	*Arabidopsis thaliana*	Chlor.-Pos. 2056	—
		Arabidopsis lyrata ssp. *lyrata*	(Koch, Houbold, and Mitchell-Olds 2001), AF144342	
		Arabidopsis glabra	(Koch, Houbold, and Mitchell-Olds 2001), AF144333	
PGIC	Cytosolic phosphoglucose isomerase	*Arabidopsis thaliana*	AT5G42740	8.33
		Arabidopsis lyrata ssp. *petraea*	(Kawabe, Yamane, and Miyashita 2000), AB044969	
		Leavenworthia crassa	(Liu, Charlesworth, and Kreitman 1999), AF054455	
PIST	Pistillata	*Arabidopsis thaliana*	AT5G20240	42.21
		Arabidopsis lyrata	(Lawton-Rauh, Buckler, and Purugganan 1999), AF143382	
RPM1	Disease resistance protein	*Arabidopsis thaliana*	AT3G07040	5.10
		Arabidopsis lyrata ssp. *lyrata*	(Stahl et al. 1999), AF122982	
		Brassica napus	(Grant et al. 1998), AF105139	
Sc-ADH	Short-chain alcohol dehydrogenase	*Arabidopsis thaliana*	AT4G05530	15.31
		Arabidopsis lyrata ssp. *petraea*	This study, AF494368	

[a] For *A. thaliana*, locus names from the complete genome sequence are given. In cases where the sequence was derived from a population study, references and accession numbers are given with locus names in parentheses.

ferred codons. Pairwise comparisons were also made between *A. lyrata* and *A. thaliana* for the presence of major versus minor codons, as defined by multivariate analysis of codon usage in *A. thaliana* (Chiapello et al. 1998). In particular, for all codons which have the same amino acid, the number of cases where *A. lyrata* has a major codon and *A. thaliana* a nonmajor codon, and vice versa, were recorded (Akashi 1996). With the outgroup sequences, the rates of unpreferred relative to preferred synonymous substitutions were also estimated for each lineage independently, using the assumptions of parsimony (Akashi 1995, 1996; Takano-Shimizu 1999). Only codons that encode the same amino acid in all three lineages were used in this analysis. Because of the small number of *A. lyrata* genes currently available, this analysis assumes that codon preferences are the same as those in *A. thaliana*.

Estimating the Genomic Deleterious Mutation Rate (U)

The combined sequence data set of all nuclear loci in *A. thaliana* and *A. lyrata* was also used to estimate U, the genomic deleterious mutation rate for Arabidopsis, using the method of Keightley and Eyre-Walker (2000), which considers only the contribution of amino acid changes to deleterious mutation. Because this method uses estimates of substitution rates between a pair of species, it measures the average deleterious mutation rate for both species. The per-site estimate of the number of deleterious mutations between species (u) was calculated from the following equation using a program provided by P. Keightley (personal communication):

$$u = (\overline{K_{ts}N_{ts}} + \overline{K_{tv}N_{tv}}) - \frac{\overline{K_n}}{3}$$

Where K_{ts} is the per-codon number of synonymous transitions, K_{tv} is the per-codon number of synonymous transversions, K_n is the per-codon nonsynonymous substitution rate, and N_{ts} and N_{tv} are estimates of the proportion of the transitions and transversions in the sequence that would cause an amino acid substitution. To convert this into an estimate of the genome-wide deleterious mutation rate, the per-site value needs to be multiplied by the quantity Z:

$$Z = \frac{2 \times S \times I}{T}$$

Where S is the total number of base pairs of exon sequence in the genome, I the generation time, and T the number of years of divergence or twice the divergence time between the species. S was estimated using information from the *A. thaliana* genome sequence project (Arabidopsis Genome Consortium 2000: S = 33,249,250). I is thought to be 1 generation/year or less in natural populations of *A. thaliana*, whereas it varies from 1 to 2 generations/year for *A. lyrata*. For the calculation of U, I was assumed to be 1. The divergence time between *A. thaliana* and *A. lyrata* is unknown, and few estimates of silent substitution rates per unit time exist for dicotyledonous plants. Koch, Haubold, and Mitchell-Olds (2000) give an estimate of the silent substitution rates as 1.5×10^{-8} per year for the Brassicaceae, citing a study of fossil pollen deposits. This estimate of the nuclear substitution rate is at the high end of the estimates made across diverse plants, so we also use a lower-bound estimate of the per year substitution rate of 5.8×10^{-9} by Wolfe, Li, and Sharp (1987), which uses the divergence between monocots and dicots. Using these two substitution rates, we estimated T from the total number of synonymous substitutions per synonymous site between the two species from our complete data set of nuclear loci.

Results

Rates of Protein Evolution

Table 2 summarizes the pairwise estimates of synonymous and nonsynonymous divergence between *A.*

Table 2
Synonymous and Nonsynonymous Substitution Rates Between *A. thaliana* and *A. lyrata*. Rates Were Estimated Using the Method of Comeron (1995)

Locus	S[a]	R[b]	K_s (95% CI)[c]	K_a (95% CI)[c]	K_a/K_s	Syn_{thal}[d]	Rep_{thal}[d]	Syn_{lyr}[d]	Rep_{lyr}[d]
ABAp1	102.3	323.3	0.051 (0.007−0.107)	0.029 (0.010−0.051)	0.569	—	—	—	—
ABC1At (5′)	67.1	178.4	0.168 (0.062−0.304)	0.09671 (0.0482−0.152)	0.575	3	0	7	14
ABC1At (3′)	188.8	497.8	0.107 (0.058−0.166)	0.00201 (0−0.006)	0.019	—	—	—	—
ADH	317.6	859.4	0.147 (0.099−0.199)	0.0195 (0.0098−0.0288)	0.133	19	10	18	4
AOP1	223.0	737.3	0.229 (0.157−0.317)	0.0916 (0.0648−0.1281)	0.400	—	—	—	—
AOP2	303	892	0.0944 (0.054−0.139)	0.03792 (0.0417−0.0763)	0.402	10	10	2	10
AOP3	289.5	869.5	0.0953 (0.045−0.123)	0.05832 (0.0379−0.0728)	0.568	—	—	—	—
AP1	220.8	605.4	0.0730 (0.31−0.128)	0.0133 (0.0038−0.0245)	0.182	6	4	7	3
AP3	173.3	575.0	0.0854 (0.041−0.149)	0.01345 (0.0043−0.0271)	0.158	5	3	5	3
CAUL	196.3	583.5	0.0987 (0.048−0.156)	0.03150 (0.0170−0.0523)	0.319	7	7	8	8
CHI	179.8	534.8	0.210 (0.139−0.297)	0.04607 (0.0211−0.0772)	0.219	13	6	9	6
CHIA	226.0	678.0	0.114 (0.069−0.168)	0.03476 (0.0211−0.0502)	0.304	15	10	7	10
CHS	318.0	872.0	0.149 (0.098−0.205)	0.00749 (0.0022−0.0141)	0.050	19	2	19	2
EnCoH1	181.7	532.0	0.139 (0.080−0.211)	0.00756 (0.0018−0.0173)	0.054	14	2	9	1
ETR1	434.0	1,150.1	0.150 (0.103−0.198)	0.00524 (0.0009−0.0096)	0.035	17	2	27	3
F3H	271.5	858.7	0.188 (0.129−0.257)	0.00704 (0.0013−0.0142)	0.037	17	3	19	3
FAH1	362.1	1,047.0	0.110 (0.069−0.156)	0.008624 (0.0030−0.0147)	0.078	15	4	15	3
FKA1	113.6	319.7	0.1631	0	0	—	—	—	—
GL1	168.2	484.3	0.0561 (0.015−0.106)	0.01458 (0.0046−0.0286)	0.260	—	—	—	—
HAT4	106.5	274.3	0.142 (0.062−0.254)	0.00366 (0−0.0112)	0.026	9	0	5	1
matK	362.8	1,166.1	0.0267 (0.010−0.049)	0.01607 (0.0078−0.0237)	0.602	5	7	4	9
PGIC	397.0	1,316.0	0.1570 (0.116−0.202)	0.00840 (0.0035−0.0146)	0.054	23	5	19	3
PIST	186.1	504.8	0.0978 (0.041−0.173)	0.02109 (0.0078−0.0347)	0.216	—	—	—	—
RPM1	761.1	2,117.5	0.1011 (0.073−0.131)	0.02059 (0.0145−0.0276)	0.161	23	17	20	15
Sc-ADH	140.4	393.1	0.1501 (0.079−0.235)	0.00511 (0−0.015)	0.034	—	—	—	—
Total	5,775.8[e]	16,949.6[e]	0.126[e]	0.0211[e]	0.167[e]	217[f]	92[f]	193[f]	84[f]

[a] Number of synonymous sites.
[b] Number of replacement sites.
[c] 95% confidence intervals.
[d] Parsimony-based estimates of number of synonymous (Syn) and replacement (Rep) changes, in *A. thaliana* (*thal*) and *A. lyrata* (*lyr*) (see *Methods* for details).
[e] Total values excluding the *matK* chloroplast locus and the putative *ABC1At* pseudogene.
[f] Excludes only the putative *ABC1At* pseudogene.

thaliana and *A. lyrata* for the 24 loci. Levels of selective constraint, as measured by K_a/K_s ratios, are highly variable across loci, although the total ratio for nuclear loci is 0.17, indicating significant purifying selection on amino acid substitutions. However, five loci show K_a/K_s ratios greater than 0.4, suggesting that their protein sequences are evolving rapidly. One locus, *ABC1At,* has accumulated numerous deletions and frameshifts in its 5′ end in *A. lyrata* and is thus likely to be a pseudogene in this species. Within the region surveyed, there are at least five large deletions in *A. lyrata*, including three within exons. This is surprising, given the evidence for the essential function and high expression of this protein in *A. thaliana* (Cardazzo et al. 1998). The sequencing of this region in *C. rubella* confirmed that these deletions are specific to *A. lyrata*, and after the elimination of the deleted regions, parsimony analysis suggests a large excess of replacement substitutions in *A. lyrata* (table 2), although frameshifts preclude an accurate estimate. To examine the evolution of this gene in more detail, we subsequently amplified and sequenced the 3′ portion of this locus. In sharp contrast to the 5′ end, this region had no deletions, and the relative rates of replacement and synonymous substitution suggest strong selective constraints (table 1). This indicates that the gene has either become truncated or, more likely, it has been duplicated, with the second copy having degenerated. Because the sequenced 5′ end of this locus appears to be evolving neutrally in *A. lyrata*, this region of the gene was excluded from global comparisons of relative rates of substitution between the species.

Two additional nuclear loci, *AOP3* and *ABAP1*, show very high K_a/K_s ratios. Although this appears to be caused, in part, by the low estimates of K_s for these loci, *AOP3* has a relatively high K_a along with the other members of the *AOP* gene family. Relaxed constraint on these genes is perhaps not unexpected, given the evidence that these loci, involved in glucosinolate production, are expressed only in some populations of *A. thaliana* (Kliebenstein et al. 2001).

The single locus sequenced from the chloroplast genome, *matK*, also shows an unusually high K_a/K_s ratio, which is consistent with previous studies suggesting that it has one of the lowest levels of constraint among chloroplast genes, despite being widely distributed among plants (Young and dePamphilis 2000). *MatK* also shows low synonymous divergence compared with the nuclear loci, which is in accordance with the previously estimated greater than twofold reduction in synonymous substitution rates in chloroplast genes (Wolfe, Li, and Sharp 1987).

Despite a generally high level of selective constraint (Stahl et al. 1999), the coding region of the disease resistance locus *RPM1* contains a region with a deletion in *A. lyrata* compared with both *A. thaliana* and the outgroup sequence, leading to a large number of amino acid replacements in this region, which was therefore excluded from the estimates of the substitution rate. The *GLB1* locus, which is well characterized in *A. thaliana*, differs in *A. lyrata* by a frameshift in the last exon, leading to a smaller protein (Hauser, Harr, and Schlotterer 2001), and this region was also excluded from subsequent analysis.

With the exception of the putative *ABC1At* pseudogene, the genes with the highest K_a/K_s ratios tend to be those that have few or no matches with ESTs, suggesting that they are expressed at low levels (table 1). Indeed, a strong negative correlation is observed between the maximum number of EST matches and K_a/K_s (Spearman's $r = -0.671$, $P < 0.01$). The correlation appears to primarily reflect fewer nonsynonymous substitutions in highly expressed genes, rather than excess synonymous substitutions (K_a: Spearman's $r = -0.568$, $P < 0.05$; K_s: Spearman's $r = -0.303$, $P > 0.05$). This effect might reflect a greater selective constraint on genes that are more broadly or highly expressed. Alternatively, it may be caused by a higher level of annotation error for low-expression genes, which are generally less well characterized. For example, if the prediction of exon positions is generally poorer for genes that are less expressed, estimates of the K_a/K_s ratio could be inflated. However, this does not appear to be the cause of the effect; using only genes that have a complete cDNA sequenced in *A. thaliana*, the effect of EST number on K_a/K_s remains highly significant (Spearman's $r = -0.668$, $P < 0.01$).

Table 2 also shows parsimony-based estimates of synonymous and nonsynonymous substitutions in *A. thaliana* and *A. lyrata* for each locus with an available outgroup sequence. These analyses show that substitutions occur equally along both branches; there is no significant difference between the species in either synonymous (Tajima's test $\chi^2 = 1.40$, $P > 0.05$) or nonsynonymous (Tajima's test $\chi^2 = 0.36$, $P > 0.05$) substitution rates. Similarly, there is no consistent difference in the level of constraint between the species; the ratio of replacement to synonymous substitutions is not significantly different between species ($G = 0.021$, $P > 0.05$). Analyzing genes with low ($K_a/K_s > 0.1$, $N = 7$) and high ($K_a/K_s > 0.1$, $N = 9$) constraint separately, no significant difference in the ratio of nonsynonymous to synonymous substitutions is observed between the species for either class of genes (low, $G = 0.292$, $P > 0.05$; high, $G = 0.088$, $P > 0.05$).

Some of the outgroup species have high levels of silent divergence from *A. thaliana* and *A. lyrata*, making it difficult to infer the lineage in which substitutions have occurred (Tajima 1993; Bromham et al. 2000). As the relative distance of ingroup to outgroup increases, the power of the relative rate test is further decreased (Bromham et al. 2000). If the analysis is restricted to the six loci with outgroup sequences from *C. rubella* or *A. glabra*, which belong to the sister group to *A. thaliana* and *A. lyrata*, we still observe no lineage effects for either synonymous or nonsynonymous rates, or for K_a/K_s ($P > 0.05$). However, the number of synonymous substitutions in this sample is consistently greater in *A. thaliana*; five genes have higher numbers in *A. thaliana*, whereas no genes have more substitutions in *A. lyrata* (Wilcoxon signed ranks $Z = -2.03$, $P < 0.05$).

Maximum likelihood estimates of substitution rates, which take substitution biases and multiple substitutions per site into account, are broadly consistent with the parsimony-based analysis. Table 3 shows that the estimated K_a/K_s ratios for each locus are similar in both the *A. thaliana* and *A. lyrata* lineages, with no consistent difference in the level of selective constraint. For the vast majority of the genes, there was no evidence for a departure from a fixed level of selective constraint across all lineages, providing no evidence for a consistent change in the selective constraint since the divergence of these two lineages from the outgroup species. Only one locus, *AOP2*, shows a significant lineage effect on K_a/K_s, although this would not be significant after correcting for multiple tests. In this case, the estimated K_a/K_s ratio in *A. lyrata* is strikingly higher than 1, but this appears to be largely caused by an unusually low estimate of the synonymous substitution rate (table 2).

Table 3
Maximum Likelihood Estimates of the Ratio of Nonsynonymous to Synonymous Substitutions in *A. thaliana* and *A. lyrata*

	K_a/K_s (SE)		
Locus	*A. thaliana*	*A. lyrata*	$2(L_1 - L_2)$[a]
ADH	0.1592 (0.0635)	0.0802 (0.0419)	4.28
AOP2	0.15 (0)	89.0 (31.46)	10.46*
AP1	0.1368 (0)	0.1544 (0.0679)	0.67
AP3	0.1462 (0.0963)	0.1486 (0.104)	0.012
CAUL	0.2962 (0.107)	0.2393 (0.435)	0.64
CHI	0.2574 (0.00524)	0.1973 (0.0836)	2.89
CHIA	0.2154 (0.0945)	0.3915 (0.128)	4.84
CHS	0.0278 (0.0211)	0.0545 (0.0316)	2.64
EnCoH1	0.0376 (0.0289)	0.0612 (0.0491)	0.38
ETR1	0.0342 (0.0272)	0.0364 (0.0200)	0.28
F3H	0.0408 (0.0270)	0.0377 (0.0246)	0.072
FAH1	0.1005 (0.0505)	0.0513 (0.0339)	0.96
HAT4	0.001 (0)	0.0552 (0.063)	3.40
matK	0.356 (0.280)	0.506 (0.317)	1.65
PGIC	0.0556 (0.0254)	0.0393 (0.0235)	3.16
RPM1	0.2096 (0.0235)	0.1678 (0.0476)	0.28

NOTE.—SE refers to standard error.

[a] Two times the difference in likelihoods between a model with fixed K_a/K_s ratios among branches compared with a model with free ratios.

* $P < 0.05$.

For this locus, a model which allows the *A. thaliana* ratio to vary, while the outgroup and *A. lyrata* have the same ratio, differs significantly from a fixed ratio model ($\chi^2 = 7.24$, 1 df, $P < 0.05$), whereas a model allowing a free ratio in *A. lyrata* does not significantly improve the likelihood ($\chi^2 = 1.11$, $P > 0.05$). This suggests that the lineage difference at this locus largely represents reduced K_a/K_s in *A. thaliana*.

Evolution of Codon Usage Bias

Pairwise differences between the species in the presence of major versus nonmajor codons, as well as overall GC_3 are shown in table 4. In total, there are 32 more cases of *A. lyrata* having a major codon when *A. thaliana* has a nonmajor codon than the reciprocal case, but out of the total of 392 codons which differ, this is not significant (Wilcoxon ranked sign test $P > 0.05$). There is also no significant difference in GC_3 between the two species (Wilcoxon ranked sign test $P > 0.05$). Similarly, numbers of unpreferred relative to preferred substitutions do not differ significantly between species, using either the total numbers of preferred and unpreferred substitutions ($G = 0.02$, $P > 0.05$) or considering the subset of loci with sequences from the least diverged outgroups ($G = 0.14$, $P > 0.05$, $N = 5$).

If codon bias is at equilibrium with respect to mutation and selection, an equal number of unpreferred and preferred codons is expected within each lineage. Conversely, if there has been a recent change in the mutational biases or selective pressure, the numbers of preferred and unpreferred substitutions will be different (Akashi 1995, 1996). Both species show a similar indication of a slight excess of unpreferred over preferred substitutions; this difference is marginally significant for *A. thaliana* (Tajima's test $\chi^2 = 4.4$, $P < 0.05$) but not for *A. lyrata* ($\chi^2 = 3.47$, $P > 0.05$). When considering the loci sequenced in close outgroup species, however,

Table 4
Patterns of Codon Usage Bias in *A. thaliana* and *A. lyrata*

			GC_3			Preferred		Unpreferred	
Locus	$Major_{thal}$, Non_{lyr}[a]	$Major_{lyr}$, Non_{thal}[a]	*thal*	*lyr*	*Outgroup*	*thal*	*lyr*	*thal*	*lyr*
ABApl	2	2	0.390	0.379	—	—	—	—	—
ABC1At	5	8	0.381	0.393	—	—	—	—	—
ADH	14	11	0.464	0.434	0.449	7	4	6	7
AOP1	6	14	0.328	0.379	—	—	—	—	—
AOP2	6	9	0.317	0.305	0.336	3	0	3	1
AOP3	5	8	0.318	0.331	—	—	—	—	—
AP1	6	4	0.496	0.496	0.513	2	2	2	4
AP3	4	3	0.498	0.481	0.498	2	3	0	2
CAUL	6	4	0.527	0.533	0.492	3	2	1	3
CHI	7	10	0.528	0.564	0.504	4	3	4	2
CHIA	7	11	0.426	0.445	0.462	3	2	9	3
CHS	16	13	0.597	0.627	0.646	4	3	10	11
EnCoH1	6	5	0.324	0.309	0.355	2	1	4	4
ETR1	7	19	0.382	0.389	0.467	4	8	4	4
F3H	18	14	0.518	0.509	0.610	6	5	6	9
FAH1	14	8	0.499	0.483	0.472	6	5	4	8
FKA1	2	8	0.394	0.433	—	—	—	—	—
GL1	2	1	0.401	0.401	—	—	—	—	—
HAT4	2	5	0.415	0.407	0.459	0	1	4	2
matK	—	—	0.254	0.256	0.252	—	—	—	—
PgiC	16	20	0.382	0.397	0.353	4	9	11	5
PIST	5	7	0.490	0.505	—	—	—	—	—
RPM1	19	22	0.403	0.411	0.468	4	5	10	9
Sc-ADH	5	6	0.364	0.376	—	—	—	—	—
Total	180	212				54	53	78	74

[a] Homologous codons where a major codon is present in one species and a nonmajor codon is present in the other.
thal, *A. thaliana*; lyr, *A. lyrata*.

the difference is significant for both species (*A. thaliana*, 16 preferred, 33 unpreferred, $\chi^2 = 5.9$, $P < 0.05$; *A. lyrata*, 11 preferred, 27 unpreferred, $\chi^2 = 6.74$, $P < 0.05$).

Given the variation in codon bias across loci, some genes may be under weak or no selection on codon usage, whereas stronger purifying selection may be acting on others. It is thus worthwhile to analyze the differences in codon bias among the species across different categories of overall levels of codon bias. Dividing the genes into high-bias ($GC_3 > 0.4$ for both species) and low-bias ($GC_3 < 0.4$ for both species), a significantly higher number of major codon occurrences are observed in *A. lyrata* for low-biased genes (*A. lyrata* total, 90; *A. thaliana* total, 59; Wilcoxon ranked sign test $Z = -2.375$, $P < 0.05$) but there is no significant difference for high-biased genes (*A. lyrata* total, 112; *A. thaliana* total, 120; $P > 0.05$). Both species show a marginally significant correlation between GC_3 and maximum EST matches, and this correlation is stronger in *A. lyrata* than in *A. thaliana* (*A. thaliana*, Spearman's $r = 0.402$, one-tailed $P < 0.05$, *A. lyrata*, Spearman's $r = 0.484$, one-tailed $P < 0.05$). However, this correlation is largely caused by the *CHS* gene, which has both an exceptionally high GC_3 and number of EST matches (tables 1, 2). Excluding this locus, the correlation becomes nonsignificant for *A. thaliana* ($r = 0.316$, $P > 0.05$) and marginally significant for *A. lyrata* ($r = 0.409$, $P < 0.05$). No significant correlation is observed between K_s and the GC_3 values of either species (*A. thaliana* $r = 0.046$, $P > 0.05$; *A. lyrata* $r = 0.13$, $P > 0.05$), suggesting that codon usage is not an important determinant of synonymous substitution rates in this sample of genes.

Evolution of Intron Size

There is evidence for substantial intron size evolution between *A. thaliana* and *A. lyrata*. Noncoding regions have accumulated a large number of insertion-deletion differences between species, and the cumulative result is a 5% reduction in the amount of DNA derived from intron sequence in *A. thaliana* compared with *A. lyrata* in the 19 genes sampled (16,883 base pairs in *A. thaliana*, 17,846 in *A. lyrata*). Figure 1 shows the difference in intron size at each locus between the species. Comparing all 87 introns in the data set, intron size is consistently smaller in *A. thaliana* (Wilcoxon ranked sign test $Z = -2.864$, $P < 0.01$). Comparing the total intron size per gene, however, the difference is not significant in this sample ($N = 19$, $Z = -1.932$, $P = 0.053$). This difference between the total length per gene versus lengths of individual introns probably reflects the fact that several regions sequenced have only a few small introns that show little difference between species (fig. 1), rather than suggesting that the effect is restricted to a small number of loci. An analysis of the alignments suggests that the intron size difference primarily reflects differences in the accumulation of small insertions and deletions and simple sequence repeats; no insertion-deletion event between the species from this sample appeared to be the result of a large insertion such as a transposable element. Because many of the outgroup gene sequences are from cDNA, there is little information on the intron size in these species, and only a sample of six genes could be analyzed. There is no significant difference in intron size between both species and the outgroup sequence (Wilcoxon signed rank test, $P > 0.05$), but the total intron size from this sample of genes was larger than that in either Arabidopsis species (3,287 bp compared with 3,257 bp in *A. lyrata* and 3,091 bp in *A. thaliana*).

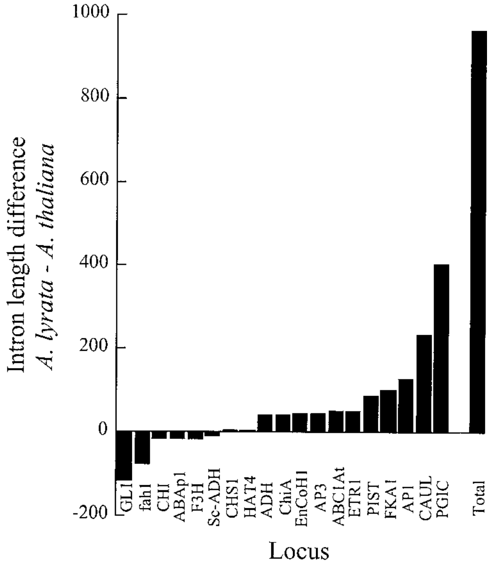

FIG. 1.—Evolution of intron size in Arabidopsis. The total difference in intron size between *A. thaliana* and *A. lyrata* is shown for each locus.

Estimate of the Genomic Deleterious Mutation Rate

Using our pooled estimates of synonymous substitution rates, the divergence time between *A. thaliana* and *A. lyrata* is estimated as between 4.2 and 10.9 MYA, using the estimates of substitution rate of 1.5×10^{-8} and 5.8×10^{-9}, respectively. The combined data generate a per-site estimate of genomic deleterious mutations between the two species of 0.077. Using these divergence time estimates, U is estimated to be between 0.22 and 0.58.

Discussion

Effects of Mating System on Molecular Evolution

From the sample of genes examined in this study, there is no evidence for an elevated rate of replacement substitution relative to synonymous substitution in *A. thaliana*, in comparison with its outcrossing congener *A. lyrata*. Similarly, the overall levels of codon usage bias do not differ substantially between the species. Both these results are in conflict with the theoretical prediction of a reduced efficacy of natural selection in self-fertilizing plants. The lack of evidence for an effect of inbreeding is particularly surprising, given our high estimate of U, which should generate high levels of background selection in a selfing population (Charlesworth

B, Morgan, and Charlesworth D 1993). Although the uncertainty of the divergence time means that this mutation rate estimate should be treated with caution, it is within the range of the estimates in *A. thaliana* based on inbreeding depression ($U = 0.5$; Charlesworth B, Charlesworth D, and Morgan 1990) and a mutation accumulation experiment ($U = 0.1$; Schultz, Lynch, and Willis 1999). In what follows, we discuss several possible explanations for our results and suggest methods to distinguish between these possibilities.

Power to Detect Substitution Differences

One explanation for the similar levels of selective constraint in both lineages is simply a lack in the power to detect significant differences among lineages. Our analyses are currently restricted to a small fraction of the genome; the detection of an effect of the breeding system may require substantially more sequence information. However, given that studies using similar samples of genes in other systems have detected significant lineage effects with apparently small differences in the effective population size (e.g., Akashi 1996), any effects of the breeding system in Arabidopsis must be very weak, which is surprising, given the potential for large differences in the efficacy of natural selection between self- and cross-fertilizing populations.

Another possibility is that the species used are too divergent to accurately infer substitution rates, and the signal of significant differences in substitution rates is not detected. An analysis of the subset of genes from the closest outgroup species does suggest a lineage effect on synonymous substitution rates but does not change our conclusions that selective constraint on amino acid substitutions is the same in the two lineages. Furthermore, all species divergences are well below saturation at replacement sites, and we find no evidence for significant differences in the numbers of amino acid substitutions between species (table 2). Provided that the mutation rate per unit time does not differ greatly between the species, this suggests that there is no major difference in the rate of fixation of amino acids. The conclusions remain unchanged when the maximum likelihood method is used, although approximate estimates of the SE of K_a/K_s were often quite large. In the case of codon bias, our pairwise comparisons of codon usage for a larger sample of genes generated similar conclusions to those based on parsimony, again suggesting that our conclusions are robust for this set of loci.

Nearly Neutral versus Neutral Models

The lack of effect of the breeding system on amino acid substitution and codon bias could be accounted for if there is no large class of slightly deleterious mutations in Arabidopsis, so that most synonymous and amino acid substitutions that have fixed since the divergence of the species studied were effectively neutral with respect to fitness in both species. Alternatively, mutations with small deleterious effects in heterozygotes may occur, but they may be nearly recessive, and experience strong selection in homozygotes, preventing their spread in selfing populations. Although the reduction in the effective size caused by background selection is thought to outweigh the purging effects of high levels of homozygosity (Charlesworth 1994), a theoretical investigation of the interaction of these effects remains preliminary, and the quantitative importance for the fixation of deleterious mutations in selfers versus outcrossers will depend on the distributions of selective effects and dominance coefficients. The strength of background selection also depends on the deleterious mutation rate, so another possibility is that this may be too low to substantially reduce N_e. However, our high estimate of U based on the sequence data, and comparable estimates using levels of inbreeding depression, makes this explanation seem unlikely.

Conclusions based on a ‘‘random’’ sample of loci are also complicated by the presence of differences in the strength and direction of selection among genes and individual sites. Our analysis relied on pooling a heterogeneous set of loci, which clearly vary in their levels of selective constraint, and some of these loci may also have been subject to positive selection on amino acid mutations. Because of the substantial differences among these loci, the effects of the breeding system and recombination rate are very likely to have differential effects on these different classes of genes. In the case of codon bias, we observed higher levels of major codon usage in *A. lyrata* when only the low-biased genes are examined. This may reflect strong selection ($N_e s \gg 1$) on highly biased genes in both species, whereas low-biased genes are probably under weaker selection, allowing more deleterious mutations to accumulate in selfing lineages. The effects of variation in selection coefficients should be investigated in more detail by sampling a larger number of weakly and highly expressed genes in *A. lyrata* and an outgroup species and comparing codon bias and amino acid substitution in the two lineages for these different classes of loci.

Population Size Changes

Models predicting an accumulation of slightly deleterious mutations in selfing populations assume that population size has remained constant in both species. However, if *A. lyrata* has undergone a population bottleneck, the rates of amino acid substitution and unpreferred codon substitution could have been elevated, obscuring any differences when compared with the inbreeding species *A. thaliana*. Such a population bottleneck could have been associated with postglacial recolonization of northern Europe and North America (Comes and Kadereit 1998). Consistent with a bottleneck hypothesis, our preliminary evidence suggests a departure from equilibrium codon usage in both species, with an excess of unpreferred substitutions. This suggests a reduction in codon usage bias in both lineages since divergence from their ancestor, similar to recent conclusions from comparisons of codon usage in *D. melanogaster* and *D. simulans* (McVean and Vieira 2001). However, a shift in the mutational bias or codon preference in *A. lyrata* remains a possibility.

Studies on polymorphism at the *ADH* locus in *A. lyrata* (Savolainen et al. 2000) found an excess of intermediate frequency variants in North American populations, which is consistent with a bottleneck hypothesis, but this was not evident from samples of European populations at this locus, although sample sizes were small. Our own data on polymorphism in European populations at several other loci has also not found this (B. Lauga, S. I. Wright, and D. Charlesworth, unpublished data). The possibility that the effective population size has been reduced in both species can be tested by gathering more data on polymorphism in *A. lyrata* and by examining a larger sample of outgroup species to test for elevated substitution rates in both species.

Conversely, if the absolute population size has increased in *A. thaliana* since the evolution of selfing, the efficacy of selection may not be low. Species-wide samples of polymorphism in *A. thaliana* often indicate a frequent skew in the frequency spectrum toward rare variants, a unimodal distribution of pairwise differences (Kuittinen and Aguade 2000), and a lack of genetic isolation by distance (Bergelson et al. 1998; but see Sharbel, Haubold, and Mitchell-Olds 2000), as expected under a recent population expansion. However, none of these features is consistent across all loci, and it is unclear to what degree the frequency of rare variants simply reflects the presence of strong population structure rather than global population size changes. It is also likely that such recent historical expansion events occurred too recently to affect patterns of molecular evolution, and a historical signature of higher rates of slightly deleterious fixation would therefore still be expected.

A problem with explanations based on population size change is the evidence for highly reduced levels of polymorphism within populations of *A. thaliana* and the evidence for strongly subdivided populations (Abbott and Gomes 1989; Berge, Nordal, and Hestmark 1998; Bergelson et al. 1998). The effects of the breeding system on the efficacy of selection in strongly subdivided populations remain poorly understood, and it is unclear what forms of migration and selection would allow selection to be effective species-wide in *A. thaliana*. Given the evidence for similar levels of species-wide polymorphism in both inbreeding and outbreeding Arabidopsis species (Savolainen et al. 2000), it is possible that strong population structure allows locally high frequencies of deleterious mutations, while preventing species-wide fixation. One might then frequently find an excess of replacement polymorphism in species-wide samples, as observed in *A. thaliana* (Kawabe et al. 1997; Purugganan and Suddith 1998, 1999), in contrast to the pattern in Drosophila nuclear genes (Weinreich and Rand 2000; Bustamante et al. 2002). However, the effect of high levels of homozygosity on the purging of deleterious mutations from local populations is uncertain. Clearly, more theoretical investigations of these effects, and more detailed analyses of polymorphism and population subdivision in both species, are necessary to evaluate the interaction between the breeding system and population structure in influencing the efficacy of natural selection.

How Long has A. thaliana Been Self-Fertilizing?

It is often suggested that selfing species persist only for short evolutionary times (reviewed in Takebayashi and Morrell 2001) and rapidly become extinct. If most of the selfing lineages are very recently derived, it may be difficult to detect deleterious mutation accumulation. However, if self-fertilizing populations become extinct before substantial mutation accumulation has occurred, the genetic explanation for their short evolutionary life spans (Takebayashi and Morrell 2001) cannot be correct. The amount of time during which *A. thaliana* has been self-fertilizing is unknown, but as a maximum estimate, the divergence between *A. thaliana* and its close relatives has been estimated as between 3.1 and 9 MYA (Koch, Haubold, and Mitchell-Olds 2000). However, it is much more difficult to assess the minimum time during which the species has been self-fertilizing. Comparisons of the putative *A. lyrata* self-incompatibility locus (*SRK*) indicate that the most similar *A. thaliana* sequence encodes a truncated kinase domain, and the locus is probably a pseudogene (Kusaba et al. 2001). Loss of function mutations at this locus may have caused *A. thaliana*'s self-fertility or could have occurred after self-fertilization evolved. However, it is impossible to infer the precise date of this event because of the high silent and replacement polymorphisms among *A. lyrata* alleles (Schierup et al. 2001). Nevertheless, the observation of high silent and amino acid divergence between species using the most similar *A. lyrata SRK* allele (C. Bartholomé and D. Charlesworth, unpublished data) suggests that this locus may have been nonfunctional for a long time; therefore, *A. thaliana* has probably been self-fertile for most of the time since its separation from *A. lyrata*. As further information becomes known about the genome-wide patterns of linkage disequilibrium in *A. thaliana* (Nordborg et al. 2002), the data might be used to estimate the historical rate of self-fertilization (Nordborg 2000), although this is also complicated by the presence of population structure.

Effects of Gene Expression on Patterns on Molecular Evolution

A surprising result from our study is that the gene expression level explains a large proportion of the observed variance in selective constraint on amino acids among loci. Using the number of EST matches as a crude estimate of gene expression, we observe a strong correlation with amounts of amino acid, but not silent, substitutions, even for loci with a complete cDNA sequence. This suggests that less-expressed genes either have low selective constraints on amino acid substitutions or else that they are more likely to be subject to positive selection in Arabidopsis. In mammals, there is a correlation between the breadth of expression and K_a/K_s, which has been interpreted as evidence that a higher proportion of replacement changes affect function in genes expressed in many tissues (Duret and Mouchiroud 2000). Although the EST database for *A. thaliana* does not include sufficient sampling across tissues to investigate this in detail, the breadth of expression may generally be associated with the overall level of gene expression (Akashi 2001). As more quantitative information on gene expression becomes available, it will be important to investigate in more detail the effects of expression levels on substitution rates and gene structure.

Evolution of Intron Size in *Arabidopsis*

In our sample of loci, intron size was consistently smaller in *A. thaliana* than in *A. lyrata*. This contrasts with the patterns observed in *D. melanogaster*, where regions of low recombination have larger introns on an average (Carvalho and Clark 1999; Comeron and Kreitman 2000). However, it is consistent with measurements of DNA content in the two species studied; *A. lyrata* is estimated to have an approximately fourfold greater genome size than *A. thaliana* (O. Savolainen, personal communication). Comparative mapping of a bacterial artificial chromosome clone containing the putative self-incompatibility locus in *A. lyrata* has also provided evidence that intergenic sequences are consistently larger in *A. lyrata* in comparison with those of *A. thaliana* (Kusaba et al. 2001), so a general decrease in the sizes of noncoding regions is possible in *A. thaliana*. The contrast may reflect directional selection on intron size in a fast-growing annual, given the observed negative correlation in plants between genome size and weediness (Bennett, Leitch, and Hanson 1998). However, evidence for high rates of deletion in the *ABC1At* pseudogene in *A. lyrata* indicates that mutation may rapidly eliminate ''junk'' DNA, suggesting that selection may be maintaining large intron sizes in these species. Comparisons of segregating insertions and deletions with those that are fixed between species should be helpful in distinguishing between effects of mutational biases and selection (Comeron and Kreitman 2000), although the estimation of numbers of insertion and deletion events is a challenge even for modestly diverged species.

Acknowledgments

We thank B. Charlesworth and D. Schoen for comments on the manuscript, P. Keightley for helpful discussion and for providing his program for estimating *U*, G. Marais for providing EST data, and H. Hurka for providing genomic DNA. This work was supported by a NERC senior fellowship to D.C. and by a Commonwealth fellowship and NSERC PGSB scholarship to S.I.W.

LITERATURE CITED

ABBOTT, R. J., and M. F. GOMES. 1989. Population genetic structure and outcrossing rate of *Arabidopsis thaliana* (L.) Heynh. Heredity **62**:411–418.

ACARKAN, A., M. ROSSBERG, M. KOCH, and R. SCHMIDT. 2000. Comparative genome analysis reveals extensive conservation of genome organisation for *Arabidopsis thaliana* and *Capsella rubella*. Plant J. **23**:55–62.

AGUADÉ, M. 2001. Nucleotide sequence variation at two genes of the phenylpropanoid pathway, the *FAH1* and *F3H* genes, in *Arabidopsis thaliana*. Mol. Biol. Evol. **18**:1–9.

AKASHI, H. 1995. Inferring weak selection from patterns of polymorphism and divergence at ''silent'' sites in Drosophila DNA. Genetics **139**:1067–1076.

———. 1996. Molecular evolution between *Drosophila melanogaster* and *D. simulans*: reduced codon bias, faster rates of amino acid substitution, and larger proteins in *D. melanogaster*. Genetics **144**:1297–1307.

———. 1999. Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination. Genetics **151**:221–238.

———. 2001. Gene expression and molecular evolution. Curr. Opin. Genet. Dev. **11**:660–666.

ALTSCHUL, S. F., W. GISH, W. MILLER, E. W. MYERS, and D. J. LIPMAN. 1990. Basic local alignment search tool. J. Mol. Biol. **215**:403–410.

ALTSCHUL, S. F., T. L. MADDEN, A. A. SCHAFFER, J. ZHANG, Z. ZHANG, W. MILLER, and D. J. LIPMAN. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389–3402.

ANTHONY, R. G., P. E. JAMES, and B. R. JORDAN. 1995. The cDNA sequence of a cauliflower apetala-1/squamosa homolog. Plant Physiol. **108**:441–442.

BARTON, N. H. 2000. Genetic hitchhiking. Philos. Trans. R. Soc. Lond. B: Biol. Sci. **355**:1553–1562.

BAUDRY, E., C. KERDELHUE, H. INNAN, and W. STEPHAN. 2001. Species and recombination effects on DNA variability in the tomato genus. Genetics **158**:1725–1735.

BENNETT, M. D., I. J. LEITCH, and L. HANSON. 1998. DNA amounts in two samples of angiosperm weeds. Ann. Bot. Suppl. A: 121–134.

BERGE, G., I. NORDAL, and G. HESTMARK. 1998. The effect of breeding systems and pollination vectors on the genetic variation of small plant populations within an agricultural landscape. Oikos **81**:17–29.

BERGELSON, J., E. STAHL, S. DUDEK, and M. KREITMAN. 1998. Genetic variation within and among populations of *Arabidopsis thaliana*. Genetics **148**:1311–1323.

BRAVERMAN, J. M., R. R. HUDSON, N. L. KAPLAN, C. H. LANGLEY, and W. STEPHAN. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics **140**:783–796.

BRITSCH, L., J. DEDIO, H. SAEDLER, and G. FORKMANN. 1993. Molecular characterization of flavanone 3 beta-hydroxylases. Consensus sequence, comparison with related enzymes and the role of conserved histidine residues. Eur. J. Biochem. **217**:745–754.

BROMHAM, L., D. PENNY, A. RAMBAUT, and M. D. HENDY. 2000. The power of relative rates tests depends on the data. J. Mol. Evol. **50**:296–301.

BUSTAMANTE, C. D., R. NIELSEN, S. A. SAWYER, K. M. OLSEN, M. D. PURUGGANAN, and D. L. HARTL. 2002. The cost of inbreeding in Arabidopsis. Nature **416**:531–534.

CARDAZZO, B., P. HAMEL, W. SAKAMOTO, H. WINTZ, and G. DUJARDIN. 1998. Isolation of an *Arabidopsis thaliana* cDNA by complementation of a yeast abc1 deletion mutant deficient in complex III respiratory activity. Gene **221**:117–125.

CARR, S. M., and V. F. IRISH. 1997. Floral homeotic gene expression defines developmental arrest stages in *Brassica oleracea* L. vars. botrytis and italica. Planta **201**:179–188.

CARVALHO, A. B., and A. G. CLARK. 1999. Intron size and natural selection. Nature **401**:344.

CHARLESWORTH, B. 1992. Evolutionary rates in partionally self-fertilizing species. Am. Nat. **140**:126–148.

———. 1994. The effect of background selection against deleterious mutations on weakly selected, linked variants. Genet. Res. **63**:213–227.

Charlesworth, B., and D. Charlesworth. 1997. Rapid fixation of deleterious alleles can be caused by Muller's ratchet. Genet. Res. **70**:63–73.

Charlesworth, B., D. Charlesworth, and M. T. Morgan. 1990. Genetic loads and estimates of mutation rates in highly inbred plant populations. Nature **347**:380–382.

Charlesworth, B., M. T. Morgan, and D. Charlesworth. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics **134**:1289–1303.

Charlesworth, D., M. T. Morgan, and B. Charlesworth. 1993. Mutation accumulation in finite outbreeding and inbreeding populations. Genet. Res. **61**:39–56.

Charlesworth, D., and S. I. Wright. 2001. Breeding systems and genome evolution. Curr. Opin. Genet. Dev. **11**: 685–690.

Chen, H.-H., Y.-Y. Charng, F. Y. Shang, and J.-F. Shaw. 1998. Molecular cloning and sequencing of a broccoli cDNA (Accession No. AF047476) encoding an *ETR*-type ethylene receptor. (PGR98-088). Plant Physiol. **117**:717.

Chiapello, H., F. Lisacek, M. Caboche, and A. Henaut. 1998. Codon usage and gene function are related in sequences of *Arabidopsis thaliana*. Gene **209**:GC1–GC38.

Comeron, J. M. 1995. A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. J. Mol. Evol. **41**:1152–1159.

———. 1999. K-Estimator: calculation of the number of nucleotide substitutions per site and the confidence intervals. Bioinformatics **15**:763–764.

Comeron, J. M., and M. Kreitman. 2000. The correlation between intron length and recombination in Drosophila. Dynamic equilibrium between mutational and selective forces. Genetics **156**:1175–1190.

Comeron, J. M., M. Kreitman, and M. Aguade. 1999. Natural selection on synonymous sites is correlated with gene length and recombination in Drosophila. Genetics **151**:239–249.

Comes, H. P., and J. W. Kadereit. 1998. The effect of Quaternary climatic changes on plant distribution and evolution. Trends Plant Sci. **3**:432–438.

DeSalle, R., and A. R. Templeton. 1988. Founder effects and the rate of mitochondrial DNA evolution in Hawaiian Drosophila. Evolution **42**:1076–1084.

Duret, L., and D. Mouchiroud. 1999. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc. Natl. Acad. Sci. USA **96**:4482–4487.

———. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol. Biol. Evol. **17**:68–74.

Dvorak, J., M. C. Luo, and Z. L. Yang. 1998. Restriction fragment length polymorphism and divergence in the genomic regions of high and low recombination in self-fertilizing and cross-fertilizing aegilops species. Genetics **148**: 423–434.

Gordo, I., and B. Charlesworth. 2000. The degeneration of asexual haploid populations and the speed of Muller's ratchet. Genetics **154**:1379–1387.

Grant, M. R., J. M. McDowell, A. G. Sharpe, M. de Torres Zabala, D. J. Lydiate, and J. L. Dangl. 1998. Independent deletions of a pathogen-resistance gene in Brassica and Arabidopsis. Proc. Natl. Acad. Sci. USA **95**: 15843–15848.

Graustein, A., J. M. Gaspar, J. R. Walters, and M. F. Palopoli. 2002. Levels of DNA polymorphism vary with mating system in the nematode genus Caenorhabditis. Genetics **161**:99–107.

Hamrick, J. L., and M. J. Godt. 1990. Allozyme diversity in plant species. Pp. 43–63 *in* A. H. D. Brown, M. T. Clegg, A. L. Kahler, and B. S. Weir, eds. Plant population genetics, breeding, and genetic resources. Sinauer, Sunderland, Mass.

Hauser, M. T., B. Harr, and C. Schlotterer. 2001. Trichome distribution in *Arabidopsis thaliana* and its close relative *Arabidopsis lyrata*: molecular analysis of the candidate gene GLABROUS1. Mol. Biol. Evol. **18**:1754–1763.

Heller, J., and J. Maynard Smith. 1979. Does Muller's ratchet work with selfing? Genet. Res. **32**:289–294.

Hill, W. G., and A. Robertson. 1966. The effect of linkage on limits to artificial selection. Genet. Res. **8**:269–294.

Johnson, K. P., and J. Seger. 2001. Elevated rates of nonsynonymous substitution in island birds. Mol. Biol. Evol. **18**: 874–881.

Junghans, H., and M. Metzlaff. 1990. A simple and rapid method for the preparation of total plant DNA. Biotechniques **8**:176.

Kawabe, A., H. Innan, R. Terauchi, and N. T. Miyashita. 1997. Nucleotide polymorphism in the acidic chitinase locus (*ChiA*) region of the wild plant *Arabidopsis thaliana*. Mol. Biol. Evol. **14**:1303–1315.

Kawabe, A., K. Yamane, and N. T. Miyashita. 2000. DNA polymorphism at the cytosolic phosphoglucose isomerase (*PgiC*) locus of the wild plant *Arabidopsis thaliana*. Genetics **156**:1339–1347.

Keightley, P. D., and A. Eyre-Walker. 2000. Deleterious mutations and the evolution of sex. Science **290**:331–333.

Kim, Y., and W. Stephan. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics **160**:765–777.

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

Kliebenstein, D. J., J. Kroymann, P. Brown, A. Figuth, D. Pedersen, J. Gershenzon, and T. Mitchell-Olds. 2001. Genetic control of natural variation in Arabidopsis glucosinolate accumulation. Plant Physiol. **126**:811–825.

Kliman, R. M., and J. Hey. 1993. Reduced natural selection associated with low recombination in *Drosophila melanogaster*. Mol. Biol. Evol. **10**:1239–1258.

Koch, M. A., B. Haubold, and T. Mitchell-Olds. 2000. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in Arabidopsis, Arabis, and related genera (Brassicaceae). Mol. Biol. Evol. **17**:1483–1498.

———. 2001. Molecular systematics of the Brassicaceae: evidence from coding plastidic *matK* and nuclear *Chs* sequences. Am. J. Bot. **88**:534–544.

Kuittinen, H., and M. Aguade. 2000. Nucleotide variation at the CHALCONE ISOMERASE locus in *Arabidopsis thaliana*. Genetics **155**:863–872.

Kusaba, M., K. Dwyer, J. Hendershot, J. Vrebalov, J. B. Nasrallah, and M. E. Nasrallah. 2001. Self-incompatibility in the genus Arabidopsis: characterization of the *S* locus in the outcrossing *A. lyrata* and its autogamous relative *A. thaliana*. Plant Cell **13**:627–643.

Langley, C. H., B. P. Lazzaro, W. Phillips, E. Heikkinen, and J. M. Braverman. 2000. Linkage disequilibria and the site frequency spectra in the *su*(*s*) and *su*(*w*(*a*)) regions of the *Drosophila melanogaster* X chromosome. Genetics **156**: 1837–1852.

Lawton-Rauh, A. L., E. S. Buckler 4th, and M. D. Purugganan. 1999. Patterns of molecular evolution among paralogous floral homeotic genes. Mol. Biol. Evol. **16**:1037–1045.

Li, X. F., R. J. Shen, P. L. Liu, Z. C. Tang, and Y. K. He. 2000. Molecular characters and morphological genetics of *CAL* gene in Chinese cabbage. Cell. Res. **10**:29–38.

Liu, F., D. Charlesworth, and M. Kreitman. 1999. The effect of mating system differences on nucleotide diversity at the phosphoglucose isomerase locus in the plant genus Leavenworthia. Genetics **151**:343–357.

Liu, F., L. Zhang, and D. Charlesworth. 1998. Genetic diversity in Leavenworthia populations with different inbreeding levels. Proc. R. Soc. Lond. B: Biol. Sci. **265**:293–301.

Marais, G., D. Mouchiroud, and L. Duret. 2001. Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. Proc. Natl. Acad. Sci. USA **98**:5688–5692.

McVean, G. A., and B. Charlesworth. 2000. The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. Genetics **155**:929–944.

McVean, G. A., and J. Vieira. 2001. Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in Drosophila. Genetics **157**:245–257.

Nachman, M. W. 2001. Single nucleotide polymorphisms and recombination rate in humans. Trends Genet. **17**:481–485.

Nair, R. B., R. W. Joy IV, E. Kurylo, X. Shi, J. Schnaider, R. S. Datla, W. A. Keller, and G. Selvaraj. 2000. Identification of a CYP84 family of cytochrome P450-dependent mono-oxygenase genes in brassica napus and perturbation of their expression for engineering sinapine reduction in the seeds. Plant Physiol. **123**:1623–1634.

Nicholas, K. B., H. B. Nicholas Jr., and D. W. Deerfield II. 1997. GeneDoc: analysis and visualization of genetic variation. EMBNEW.NEWS **4**:14.

Nordborg, M. 2000. Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. Genetics **154**:923–929.

Nordborg, M., J. O. Borevitz, J. Bergelson et al. (12 coauthors). 2002. The extent of linkage disequilibrium in *Arabidopsis thaliana*. Nat. Genet. **7**:7.

Ohta, T. 1992. The nearly neutral model of molecular evolution. Annu. Rev. Ecol. Syst. **23**:263–286.

———. 1993. Amino acid substitution at the *Adh* locus of Drosophila is facilitated by small population size. Proc. Natl. Acad. Sci. USA **90**:4548–4551.

Pannell, J. R., and B. Charlesworth. 2000. Effects of metapopulation processes on measures of genetic diversity. Philos. Trans. R. Soc. Lond. B: Biol. Sci. **355**:1851–1864.

Petrov, D. A., E. R. Lozovskaya, and D. L. Hartl. 1996. High intrinsic rate of DNA loss in Drosophila. Nature **384**:346–349.

Pollak, E. 1987. On the theory of partially inbreeding finite populations. I. Partial selfing. Genetics **117**:353–360.

Purugganan, M. D., and J. I. Suddith. 1998. Molecular population genetics of the Arabidopsis CAULIFLOWER regulatory gene: nonneutral evolution and naturally occurring variation in floral homeotic function. Proc. Natl. Acad. Sci. USA **95**:8130–8134.

———. 1999. Molecular population genetics of floral homeotic loci. Departures from the equilibrium-neutral model at the APETALA3 and PISTILLATA genes of *Arabidopsis thaliana*. Genetics **151**:839–848.

Rodriguez-Trelles, F., R. Tarrio, and F. J. Ayala. 1999. Switch in codon bias and increased rates of amino acid substitution in the *Drosophila saltans* species group. Genetics **153**:339–350.

Savolainen, O., C. H. Langley, B. P. Lazzaro, and H. Fréville 2000. Contrasting patterns of nucleotide polymorphism at the alcohol dehydrogenase locus in the outcrossing *Arabidopsis lyrata* and the selfing *Arabidopsis thaliana*. Mol. Biol. Evol. **17**:645–655.

Schierup, M. H., B. K. Mable, P. Awadalla, and D. Charlesworth. 2001. Identification and characterization of a polymorphic receptor kinase gene linked to the self-incompatibility locus of *Arabidopsis lyrata*. Genetics **158**:387–399.

Schultz, S. T., M. Lynch, and J. H. Willis. 1999. Spontaneous deleterious mutation in *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA **96**:11393–11398.

Sharbel, T. F., B. Haubold, and T. Mitchell-Olds. 2000. Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and postglacial colonization of Europe. Mol. Ecol. **9**:2109–2118.

Smith, N. G., and A. Eyre-Walker. 2001. Synonymous codon bias is not caused by mutation bias in G+C-rich genes in humans. Mol. Biol. Evol. **18**:982–986.

Stahl, E. A., G. Dwyer, R. Mauricio, M. Kreitman, and J. Bergelson. 1999. Dynamics of disease resistance polymorphism at the *Rpm1* locus of Arabidopsis. Nature **400**:667–671.

Tachida, H. 2000. Molecular evolution in a multisite nearly neutral mutation model. J. Mol. Evol. **50**:69–81.

Tajima, F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. Genetics **135**:599–607.

Takano-Shimizu, T. 1999. Local recombination and mutation effects on molecular evolution in Drosophila. Genetics **153**:1285–1296.

Takebayashi, N., and P. L. Morrell. 2001. Is self-fertilization an evolutionary dead end? Revisiting an old hypothesis with genetic theories and a macroevolutionary approach. Am. J. Bot. **88**:1143–1150.

Tarrio, R., F. Rodriguez-Trelles, and F. J. Ayala. 2001. Shared nucleotide composition biases among species and their impact on phylogenetic reconstructions of the Drosophilidae. Mol. Biol. Evol. **18**:1464–1473.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

Weinreich, D. M., and D. M. Rand. 2000. Contrasting patterns of nonneutral evolution in proteins encoded in nuclear and mitochondrial genomes. Genetics **156**:385–399.

Whitlock, M. C., and N. H. Barton. 1997. The effective size of a subdivided population. Genetics **146**:427–441.

Wolfe, K. H., W. H. Li, and P. M. Sharp. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc. Natl. Acad. Sci. USA **84**:9054–9058.

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

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

Young, N. D., and C. W. dePamphilis. 2000. Purifying selection detected in the plastid gene *matK* and flanking ribozyme regions within a group II intron of nonphotosynthetic plants. Mol. Biol. Evol. **17**:1933–1941.

Brandon Gaut, reviewing editor

Accepted April 5, 2002