Copyright © 2003 by the Genetics Society of America

Selection on Rapidly Evolving Proteins in the Arabidopsis Genome

Marianne Barrier,* Carlos D. Bustamante,† Jiaye Yu‡ and Michael D. Purugganan*,[1]

**Department of Genetics and ‡Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina 27695 and †Department of Statistics, University of Oxford, Oxford OX1 3TG, United Kingdom*

Manuscript received June 10, 2002
Accepted for publication November 11, 2002

ABSTRACT

Genes that have undergone positive or diversifying selection are likely to be associated with adaptive divergence between species. One indicator of adaptive selection at the molecular level is an excess of amino acid replacement fixed differences per replacement site relative to the number of synonymous fixed differences per synonymous site ($\omega = K_a/K_s$). We used an evolutionary expressed sequence tag (EST) approach to estimate the distribution of ω among 304 orthologous loci between *Arabidopsis thaliana* and *A. lyrata* to identify genes potentially involved in the adaptive divergence between these two Brassicaceae species. We find that 14 of 304 genes (~5%) have an estimated $\omega > 1$ and are candidates for genes with increased selection intensities. Molecular population genetic analyses of 6 of these rapidly evolving protein loci indicate that, despite their high levels of between-species nonsynonymous divergence, these genes do not have elevated levels of intraspecific replacement polymorphisms compared to previously studied genes. A hierarchical Bayesian analysis of protein-coding region evolution within and between species also indicates that the selection intensities of these genes are elevated compared to previously studied *A. thaliana* nuclear loci.

THE genetic architecture of species differences has been the subject of intense study in the last few years (Orr and Coyne 1992; Haag and True 2001; Wu 2001). There has been a concerted effort to identify genes responsible for adaptive differences between species to examine the genetic mechanisms that accompany evolutionary diversification (Haag and True 2001) and even speciation (Wu 2001). Adaptive morphological and physiological differences between species should leave a signature of positive selection at the molecular level and permit an analysis of evolutionary divergence at both the molecular genetic and the phenotypic levels (Haag and True 2001). By investigating loci whose sequences have been shaped by positive selection, it may be possible to unravel the evolutionary genetic mechanisms that underlie adaptive divergence between species and the origins and evolution of species differences.

One indicator of adaptive selection at the molecular level is an excess of amino acid replacement fixed differences per replacement site relative to the number of synonymous fixed differences per synonymous site ($\omega = K_a/K_s$; Li *et al.* 1981, 1985; Hughes and Yeager 1998). Purifying selection on amino acid variation, for example, causes a decrease in the rate of amino acid fixation and thus an inferred $\omega < 1$. If most amino acid variation is neutral, such as in pseudogenes, $\omega \sim 1$ (Li *et al.* 1981). Strong diversifying or positive selection operating on amino acid variation is associated with $\omega > 1$ (Hughes and Yeager 1998; Anisimova *et al.* 2001). Empirically, the distribution of ω varies radically among different classes of genes: in plant Brassicaceae species, the mean ω is ~0.14 (Tiffin and Hahn 2002), while for many mammalian pseudogenes ω has been shown to cluster around 1.0 (Bustamante *et al.* 2002a). In contrast, values of $\omega > 1$ have been observed in gamete recognition protein-coding genes (Swanson and Vacquier 1995), loci associated with host-parasite interaction (Hughes 1991), and genes involved in adaptation to specific environments (Messier and Stewart 1997). The elevated values of ω in these latter genes, which encode rapidly evolving proteins, are believed to arise from selection for divergence in protein structure and function.

Identifying genes with increased values of ω will be facilitated by evolutionary genomic approaches that permit investigators to sample and compare large numbers of genes between species genomes to search for those loci characterized by rapid evolution (Endo *et al.* 1996; Swanson *et al.* 2001; Tiffin and Hahn 2002). Although whole-genome sequences are not generally available from two closely related species, rapidly evolving proteins can be identified using expressed sequence tags (ESTs). An evolutionary EST approach has been used, for example, in demonstrating that genes encoding accessory gland-specific proteins in Drosophila species evolve faster than other loci in the genome, possibly as a

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos. BQ834040–BQ834596, BQ839827–BQ839830, AY140430–AY140446, and AY140459–AY140531.

[1]*Corresponding author:* Department of Genetics, North Carolina State University, 3513 Gardner Hall, Box 7614, Raleigh, NC 27695. E-mail: michaelp@unity.ncsu.edu

result of selection pressures associated with mate choice and intersexual genomic conflict (Swanson *et al.* 2001).

Our objective is to examine whether an evolutionary EST approach can identify genes with increased selection intensities in the Arabidopsis genome. Expressed sequence tags from developing inflorescences of *Arabidopsis lyrata* were compared to the whole-genome sequence of the model plant *A. thaliana* to estimate the distribution of nonsynonymous and synonymous substitution differences between these two Brassicaceae species. Fourteen genes with $\omega > 1$, which encode rapidly evolving proteins, were identified between these two plant taxa. These genes have accelerated rates of nonsynonymous substitutions that may be associated with adaptive evolution since the divergence of these two species ~5.2 MYA (Koch *et al.* 2000). Molecular population genetic analysis of six of these rapidly evolving protein loci confirms that the selection intensities on protein sequence change in these genes are significantly higher than those in previously studied Arabidopsis nuclear loci.

MATERIALS AND METHODS

Isolation and sequencing of expressed sequence tags: Seeds from individuals of a population of *A. lyrata* in Karhumaki, Russia were obtained from Outi Savolainen and Helmi Kuittinen. Total RNA was extracted from the *A. lyrata* inflorescences using the RNeasy plant mini kit (QIAGEN, Valencia, CA) and a cDNA library constructed in the plasmid vector pCMV-PCR using the PCR cDNA library construction kit (Stratagene, La Jolla, CA). Plasmid DNA from cDNA clones was isolated using the REAL Prep96 BioRobot kit (QIAGEN) with the BioRobot 9600 (QIAGEN) and sequenced from the 5′ end using an ABI Prism 3700 96-capillary automated sequencer (Perkin-Elmer, Norwalk, CT). Sequences were edited on the basis of Phred (Ewing *et al.* 1998) quality scores, with a Phred scoring threshold of 20. Ambiguous base calls were visually confirmed against the chromatograms. These EST sequences are deposited in GenBank (accession nos. BQ834040–BQ834596 and BQ839827–BQ839830).

Analyses of ESTs: *A. thaliana* sequences homologous to the high-quality *A. lyrata* EST sequences were identified by BLAST analysis against the *A. thaliana* whole-genome coding database found at The Arabidopsis Information Resource database (http://www.arabidopsis.org), using a maximum expected value (E) of e^{-5}. The GenBank nonredundant nucleotide sequence database was also searched to find the closest matching *A. thaliana* genomic bacterial artificial chromosome (BAC) clone sequence. The top matches from each database were visually aligned with their matching *A. lyrata* EST sequence. Calculations were made on the basis of pairwise comparisons between the *A. lyrata* EST and *A. thaliana* coding region genomic DNA sequence. EST-based nonsynonymous and synonymous distances were calculated using the modified Nei and Gojobori (1986) method as implemented in MEGA 2.1 (Kumar *et al.* 1993). A software package was developed to carry out a permutation analysis of nonsynonymous/synonymous substitution ratios for genes incorporating a modified Nei-Gojobori method (Zhang *et al.* 1998). Since there are several possible sources of sequence error in the experimental acquisition of the EST sequence data, we refer to these estimates of nonsynonymous and synonymous substitutions as K_a^* and K_s^*, respectively. In general, $K_a^* = K_a + \varepsilon$ and $K_s^* = K_s + \varepsilon$, where ε is an error term that reflects the empirical error in sequence determination. Sequence comparisons using duplicate ESTs suggest that $\varepsilon \sim 0.2\%$ (M. Barrier, unpublished results), although this may be an overestimate since some of these differences may arise from allelic variation.

Isolation and sequencing of alleles: Six genes with K_a^*/K_s^* values >1 were chosen for molecular population genetic analysis. Primers were designed to amplify 1- to 2-kb regions of these genes on the basis of the *A. thaliana* sequences in the selected comparisons (see Table S1 at http://www.genetics.org/supplemental/). Leaf tissue from 10–13 *A. thaliana* ecotypes was obtained from single-seed propagated material provided by the Arabidopsis Biological Resource Center (see Table S2 at http://www.genetics.org/supplemental/). DNA was isolated from these *A. thaliana* ecotypes as well as 2–5 *A. lyrata* individuals (see above), using the DNeasy plant mini kit (QIAGEN). PCR of *A. thaliana* samples was performed with *Taq* DNA polymerase (Eppendorf, Madison, WI), using standard protocols. *A. thaliana* samples were sequenced directly via cycle sequencing with primers in both directions. PCR of *A. lyrata* samples was performed with the error-correcting *Tgo* polymerase (Roche, Indianapolis), using the manufacturer's amplification protocol. The error rate of error-correcting polymerases is <1 in 7000 bp (Olsen *et al.* 2002). Amplified *A. lyrata* products were cloned into pCR4Blunt-TOPO vector using the Zero Blunt TOPO PCR cloning kit (Invitrogen, San Diego). Plasmid miniprep DNA was isolated using the QIAprep miniprep kit (QIAGEN) and sequenced via cycle sequencing. DNA sequencing was conducted with a Prism 3700 96-capillary automated sequencer (Perkin-Elmer). GenBank accession numbers for these population data are AY140430–AY140446 and AY140459–AY140531.

Molecular population genetic data analysis: Sequences of *A. thaliana* and *A. lyrata* populations were aligned and visually corrected. All polymorphisms were visually checked against chromatographs or via resequencing. Analysis of polymorphism and divergence was carried out using DnaSP 3.5 (Rozas and Rozas 1999). *A. thaliana* species-wide silent site nucleotide diversity, π (Nei 1987), and θ (Watterson 1975) were estimated. The McDonald-Kreitman test (McDonald and Kreitman 1991) was performed to test for neutral evolution in the protein-coding region. A hierarchical Bayesian method is utilized to analyze McDonald-Kreitman-type tables for 12 previously studied (Bustamante *et al.* 2002b) and 6 rapidly evolving *A. thaliana* protein genes to estimate selection coefficients for replacement changes under a Poisson random field model (Sawyer and Hartl 1992). Details of the analytical approach utilized here are described in the accompanying appendix.

RESULTS AND DISCUSSION

Expressed sequence tags in *A. lyrata*: A small collection of ESTs were isolated and sequenced to obtain genes expressed in developing inflorescences in *A. lyrata*. From a cDNA library of ~2800 colonies, 768 clones were sequenced. From this sequence collection, 561 good-quality sequences of at least 200 bp in length were subjected to further analysis. *A. thaliana* orthologs to these *A. lyrata* ESTs were identified by BLAST searches of the whole-genome *A. thaliana* coding sequence database; 78 of the sequences did not match a clear *A. thaliana* ortholog and were not considered further. Ninety-five duplicate EST matches for *A. thaliana* coding

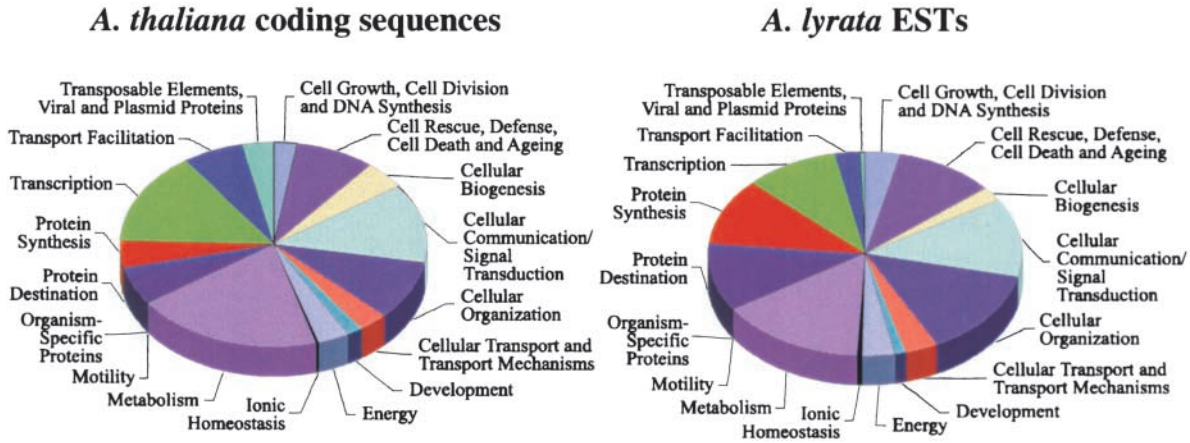

FIGURE 1.—Comparison of functional classifications of genes in the *A. thaliana* genome and the *A. lyrata* evolutionary ESTs.

sequences were also eliminated. The GenBank nonredundant nucleotide database was also searched for *A. thaliana* BAC clone sequences containing genes homologous to the *A. lyrata* EST sequences. By aligning the genomic BAC clone sequence with the *A. lyrata* and *A. thaliana* coding sequences, the boundaries of noncoding regions were located. After eliminating 84 sequence alignments with <150 bp of coding sequence, 304 unique ESTs remained for further analysis (see Table S3 at http://www.genetics.org/supplemental/). Although these ESTs represent coding region fragments, we subsequently refer to these as genes.

The *A. lyrata* EST sequences were assigned to different functional categories using the classifications of the orthologous *A. thaliana* sequences from the TAIR database (ARABIDOPSIS GENOME INITIATIVE 2000). Unclassified genes and those whose classifications were ambiguous were not included in this comparison. Of the >25,000 sequences in the *A. thaliana* coding sequence database, only slightly >4000 have thus far been unambiguously classified (ARABIDOPSIS GENOME INITIATIVE 2000). Eighty-seven of the 304 unique *A. lyrata* ESTs matched an *A. thaliana* coding sequence that has yet to be classified. In determining the count for each functional category, multiple categories listed for a single sequence were each counted as an equal fraction of the sample count. On the basis of this analysis, the range of functional categories represented by the 304 unique ESTs appears to be representative of those observed for the entire *A. thaliana* gene set (see Figure 1).

Distribution of nucleotide substitution rates between *A. thaliana* and *A. lyrata*: Comparisons of the coding sequences of the 304 ESTs from *A. lyrata* with the whole genome sequence from *A. thaliana* allow us to compare the distribution of the rates of nucleotide substitution between these two species. The distributions of both synonymous and nonsynonymous substitutions are shown in Figure 2. The mean length of aligned coding sequences is 111 ± 2.19 codons. The distribution of the number of synonymous nucleotide substitutions ranges from 0.000 to 0.552 synonymous substitutions per synonymous site. The distribution of K_s^* has one mode between 0.10 and 0.15 and has a long tail. The mean synonymous substitution distance is 0.119 ± 0.004 (see Figure 2A), which is comparable to estimates observed in previous comparisons of *A. thaliana/A. lyrata* loci. If we assume a divergence time of 5.2 MYA for the two species (KOCH *et al.* 2000), this indicates that the average synonymous mutation rate for Arabidopsis nuclear genes is $\sim 1.1 \times 10^{-8}$ substitutions/site/year.

The distribution of the number of nonsynonymous substitutions between the two species differs from that observed for synonymous substitutions. The nonsynonymous distance distribution has a mode of <0.050 nonsynonymous substitutions/nonsynonymous site, and the frequency decreases with increasing nonsynonymous substitution distances until ~0.150 (see Figure 2B). The range of K_a^* is 0.000–0.159 nonsynonymous substitutions/nonsynonymous site, with a mean of 0.025 ± 0.002. As expected, the mean $K_a^* < K_s^*$, which reflects that action of purifying selection that prevents many nonsynonymous mutations from reaching fixation between species. The mean rate of nonsynonymous substitutions in nuclear genes between the two species is 0.24×10^{-8} substitutions/site/year, which is approximately fivefold lower than the average synonymous mutation rate.

The distribution in evolutionary rates between species assumes that the comparisons are for orthologs between the two species. Identifying the correct orthologs between species will be confounded by gene duplications immediately at or prior to the most recent common ancestor of *A. thaliana* and *A. lyrata*, followed by deletion of one of the duplicate gene copies in the *A. thaliana* genome. It is unclear what the rate of gene duplication/deletion is within these species genomes.

Variation in selective constraints among Arabidopsis genes: The historical action of selection on a gene can be inferred from the relative ratio of nonsynonymous to synonymous substitutions, $\omega = K_a/K_s$ (LI *et al.* 1981,

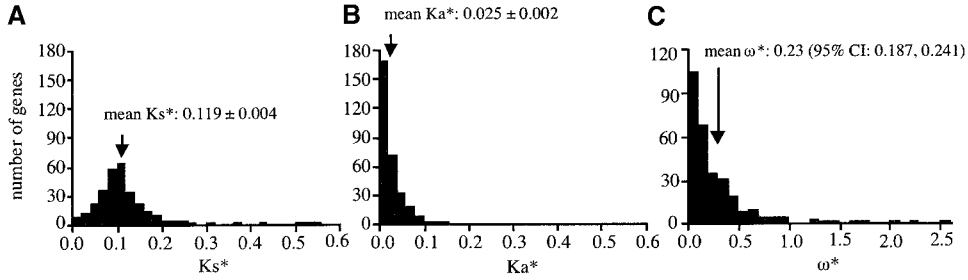

FIGURE 2.—Distributions of (A) synonymous (K_s^*) and (B) nonsynonymous (K_a^*) sequence distances, as well as (C) ω* estimates, among 304 orthologs between *A. thaliana* and *A. lyrata*.

1985; HUGHES and YEAGER 1998; ANISIMOVA *et al.* 2001). The evolutionary EST data can be used to estimate the distribution of the selection parameter ω* between *A. thaliana* and *A. lyrata*. The value of $\omega^* = K_a^*/K_s^*$ is ascertained for each of the 304 orthologous pairs between the two species, corrected for sequencing errors. The estimates of ω* range from 0.00 to 2.59; the distribution is similar to that of K_a^* in that the most genes have a low ω* value, and the distribution decreases with increasing ω* (see Figure 2C). The mean value of ω*, obtained by bootstrap resampling of K_a^* and K_s^* pair values 100,000 times (BUSTAMANTE *et al.* 2002a), is 0.213 [95% confidence interval (C.I.): $0.187 \leq \omega^* \leq 0.241$]. This is higher than previous estimates of $\omega \sim 0.14$ for a set of comparisons between *A. thaliana* and *Brassica rapa* (TIFFIN and HAHN 2002) and $\omega \sim 0.18$ for four gene comparisons between *A. thaliana* and *A. lyrata* (LAWTON-RAUH *et al.* 1999).

Of the 304 orthologous gene pairs in this evolutionary EST study, 37 genes (12%) have $\omega^* = 0$. These represent genes whose encoded proteins are under very strong selective constraint. At the other extreme, 14 genes (~5%) have ω* values >1, suggesting that these genes may have accelerated rates of nonsynonymous substitutions associated with higher selection intensities on amino acid replacement changes. These loci include genes that encode RNA and zinc-finger helicases, extensin-like proteins, and zinc-finger transcription factors. More than half of these rapidly evolving protein loci (8 out of 14 genes) encode hypothetical or putative proteins of unknown function. Genes with $\omega > 1$ have a higher mean K_a^* (0.075 ± 0.011) and a lower mean K_s^* (0.042 ± 0.009) than genes that comprise the entire EST data set. The increased ω estimates for these genes thus stem from having both elevated absolute levels of nonsynonymous substitutions and lower levels of synonymous substitutions.

It is possible that identifying 14 loci with $\omega^* > 1$ may not represent genes subject to gene-specific selection mechanisms, but may simply be due to stochastic sampling from a set of 304 loci. To test this possibility, we approximated the distribution of K_a^*/K_s^* under the null hypothesis that all genes evolved according to some common evolutionary process such as selection. To approximate this null distribution, 1000 data sets were simulated, each of which consisted of 304 simulated gene pairs. The 304 simulated gene pairs were generated by sampling the aligned *A. thaliana* and *A. lyrata* codons from the EST data set without replacement. The lengths of the 304 simulated genes matched the lengths of the 304 genes in the actual EST data set. For every pseudodata set, the ω* ratio was separately estimated for all 304 simulated genes, and the number of these 304 ω* estimates that exceeded 1 was then counted. None of the 1000 simulated data sets yielded as many as 14 genes with $\omega^* > 1$ ($P < 0.001$; see Figure 3). In fact, 10 was the highest number of genes with K_a^*/K_s^* estimates >1 and this occurred only once. The mean number of genes with $\omega^* > 1$ under this null hypothesis was 2.121 with a sample standard error of 0.047. This indicates that finding 14 genes with $\omega^* > 1$ by chance in a set of 304 loci is improbable under the null hypothesis that the action of evolutionary forces is homogenous across all loci.

Molecular population genetics of Arabidopsis loci that encode rapidly evolving proteins: One other approach to confirm the roles of positive or diversifying selection in the evolution of rapidly evolving protein genes is to examine the levels and patterns of nucleotide variation at these loci within and between species (MCDONALD and KREITMAN 1991; SCHMID and AQUADRO 2001). If the elevated levels of replacement differences between species arise from neutral processes, we should also observe a comparable increase in intraspecific replacement polymorphisms. Moreover, molecular population genetics also provides methods of selection analysis that determine whether genes are evolving according to the

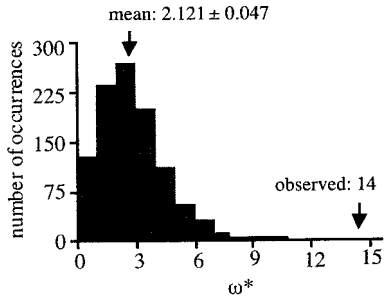

FIGURE 3.—Distribution of number of loci with $\omega^* > 1$ in permutation analysis. The mean for 1000 permuted gene data sets and the observed value from the actual observed distribution are indicated.

predictions of the neutral theory or have been subject to adaptive selection (Nielsen 2001).

Analysis of within-species nucleotide diversity in *A. thaliana* was undertaken for six genes identified from the evolutionary EST analysis as having $\omega^* > 1$ (see Table 1). These genes have a range of ω^* from 1.28 to 2.03 and were chosen to encompass the range of high ω^* values. These sampled genes include the Arabidopsis *NAC2* transcriptional activator (Xie *et al.* 2000) and a locus homologous to the human *p55.11* protein-coding gene (Boldin *et al.* 1995). Four others are hypothetical or putative proteins predicted in the *A. thaliana* genome annotation and represent genes of unknown function. All these genes are found in the *A. thaliana* EST sequence database, indicating that they are expressed in the developing plant. The genome annotation of the correct reading frame for one gene (AT2G04410) is ambiguous; we have relied on the original genome annotation to identify the reading frame for this locus and this choice does not significantly affect our results.

Alleles of each of these genes were isolated and sequenced from 10–13 ecotypes in *A. thaliana* and 2–5 individuals from a Russian population of *A lyrata*. The latter sample was included to provide an interspecific comparison; the sample sizes from the *A. lyrata* population were too small to permit a meaningful assessment of diversity for these genes in this species. The sequenced portions range in size from ~0.4 to 1.6 kb and include exon sequences that encompass the protein-coding regions that display the high ω^* values observed in the evolutionary EST analysis. The number of codons analyzed in these molecular population genetic data sets ranges from 84 to 341 codons. For four of the six genes, the amount of coding sequence assayed in the molecular population genetic analysis was ~1.5–6 times greater than the size of the sequenced ESTs. The other two had coding sequence lengths nearly equal to the length in the evolutionary EST analysis. Levels of within-species nucleotide diversity at silent sites, π, for these six rapidly evolving protein genes ranged from 0.0003 to 0.0140 in *A. thaliana*, with a mean of 0.0050 ± 0.0022 (see Table 1). This is slightly lower but comparable to the mean of ~0.007 observed for other previously studied *A. thaliana* nuclear genes (Miyashita *et al.* 1998; Purugganan and Suddith 1998; Aguade 2001; Olsen *et al.* 2002). The mean value for silent-site θ is 0.0057 ± 0.0023.

The mean ω estimates for the regions sequenced in this population genetic survey are all, except for the unknown gene AT2G04410, <1 (see Table 1). This reduction in ω values compared to those obtained in the EST study may arise from the longer lengths of coding sequences analyzed for most of the genes in the molecular population genetic study and underlying heterogeneities in selective constraint across these loci. All of the ω values, however, are greater than the mean for *A. thaliana* genes. Moreover, the mean K_s from this ex-

TABLE 1

Nucleotide variation levels of *Arabidopsis thaliana* genes encoding rapidly evolving proteins

Gene	Gene ID[a]	Function[b]	EST length (codons)	EST ω	n	K_a[c]	K_s[c]	ω[c]	Length (bp)	Length (codons)	π (silent)	θ (silent)
1	AT3G22060	Hypothetical	59	1.33	13	0.053	0.177	0.30	917	200	0.0067	0.0086
2	AT5G04410	*NAC2*	56	2.03	10	0.042	0.097	0.43	1574	341	0.0021	0.0024
3	AT4G28470	*p55.11*-like	100	1.50	10	0.014	0.048	0.30	363	117	0.0038	0.0047
4	AT1G67140	Hypothetical	128	1.28	11	0.052	0.087	0.60	1470	319	0.0003	0.0005
5	AT4G15950	Putative	140	1.63	12	0.045	0.110	0.41	991	142	0.0140	0.0146
6	AT2G04410	Unknown	60	1.49	13	0.038	0.015	2.62	1011	84	0.0029	0.0036

[a] Arabidopsis genome sequence gene reference number.
[b] Homologies or functional classifications.
[c] K_a, K_s, and K_a/K_s values estimated for the entire sequenced region in the sequence variation analysis.

TABLE 2

Replacement and synonymous changes at rapidly evolving *A. thaliana* genes

Gene	Gene ID[a]	Length (codons)	Polymorphisms R	Polymorphisms S	Fixed differences R	Fixed differences S
1	AT3G22060	200	6	4	19	24
2	AT5G04410	341	1	0	27	22
3	AT4G28470	117	0	2	3	5
4	AT1G67140	319	8	0	23	18
5	AT4G15950	142	7	8	12	8
6	AT2G04410	84	2	0	5	2

[a] Arabidopsis genome sequence gene reference number.

panded sequence region is 0.09 ± 0.02 substitutions/site, which is close to the mean for the entire data set (mean $K_s^* = 0.119 \pm 0.004$). Permutation analysis indicates that the mean K_s of the genes in the molecular population genetic study is not significantly different from the mean K_s of the EST data set ($P < 0.22$). Thus, the molecular population genetic analysis is based on sequence information whose synonymous substitution rate is comparable to the mean for *A. thaliana* genes.

Elevated levels of fixed replacement differences among rapidly evolving Arabidopsis protein genes: The relative levels of within- to between-species polymorphisms in nucleotide sites that encode a gene's products provide information on the selective forces that act in protein-coding regions (McDonald and Kreitman 1991; Bustamante *et al.* 2002b). Levels of within-species replacement and synonymous polymorphisms as well as fixed differences between *A. thaliana* and *A. lyrata* in six rapidly evolving protein genes are shown in Table 2. The levels of evolutionary change observed for these six rapidly evolving protein loci can be compared with the levels and patterns of nucleotide variation observed among 12 other previously studied *A. thaliana* genes (Bustamante *et al.* 2002b). In this study, the latter genes represent a set of Arabidopsis nuclear loci chosen without regard to their rates of protein evolution. The previously studied genes have a mean K_a of 0.020 ± 0.014 and a mean K_s of 0.148 ± 0.014. These genes have ω values ranging from 0.03 to 0.30.

The posterior distributions of the interspecies divergence time, *t*, between *A. thaliana* and *A. lyrata* are comparable between the two gene classes (for the method, see the appendix). For the previously studied gene class, the mean of the posterior distribution for t_1 equaled 8.6 in multiples of twice the effective population size with 95% highest posterior credibility interval (CI) of $[6.9 \leq t_1 \leq 11.2]$. Using data for the rapidly evolving protein genes only, the posterior distribution of t_2 has mean 9.5 and 95% highest posterior (HP) CI of $[7.0 \leq t_2 \leq 13.9]$. Using all of the data, we get 95% HPCI of $[7.41 \leq t \leq 11.22]$ for *t* with the posterior distribution having a mean of 9.16. The similarity in interspecies divergence time estimates suggests that the data from both gene classes compare loci of similar divergence times and are thus likely orthologous, and not paralogous, between *A. thaliana* and *A. lyrata*. This also indicates that the levels of synonymous substitutions between the two gene classes give comparable estimates of divergence time, indicating that the levels of interspecific synonymous divergence are comparable between both gene classes.

As expected, there is a significant elevation in the levels of fixed replacement differences between *A. thaliana* and *A. lyrata* in these six rapidly evolving protein genes. Among these loci, a total of 89 of the 168 coding region differences between these two species (~53%) result in amino acid replacements in the encoded proteins. In contrast, only 123 of 373 fixed differences (~33%) in previously studied Arabidopsis nuclear genes are replacement differences. The contrast in relative levels of replacement to synonymous fixed differences between these two gene classes is significant (Fisher's exact test, $P < 2 \times 10^{-5}$).

By comparison, the relative levels of within-species replacement to synonymous nucleotide polymorphisms within *A. thaliana* do not differ significantly between the rapidly evolving protein loci and previously studied nuclear genes. Among the genes in this study, 24 of the 38 intraspecific coding region polymorphisms (~63%) are replacement polymorphisms. Among 12 previously studied *A. thaliana* nuclear genes, 108 of 212 polymorphisms (~51%) are replacement changes. The relative levels of within-species replacement to synonymous site polymorphisms are not significantly different between the two gene classes (Fisher's exact test, $P < 0.22$). These results indicate that while the rapidly evolving protein loci have increased levels of fixed replacement differences this is not accompanied by a significant increase in relative levels of intraspecific replacement polymorphisms.

Rapidly evolving protein genes display elevated selection intensities in protein-coding regions: Selection in a specific protein-coding gene is conventionally detected in a test of homogeneity (the McDonald-Kreitman test) that examines within- and between-species replacement to synonymous nucleotide changes (McDonald and Kreitman 1991). Despite the overall increase in between-species fixed replacement differences between *A. thaliana* and *A. lyrata* in these six rapidly evolving protein genes, none of these individual genes show evidence of positive selection (Fisher's exact tests, $P <$ 0.10–1.00).

Although none of these individual genes show evidence of positive selection, each gene contains information regarding the selective forces that act on amino acid replacements (Bustamante *et al.* 2002b). Using the cell entries from a conventional McDonald-Kreitman

contingency table it is possible to estimate the four parameters in a Poisson random field model (θ^S for synonymous sites, θ^R for replacement sites, t for interspecies divergence time, and γ for replacement sites selection coefficient; Bustamante *et al.* 2002b). For a set of McDonald-Kreitman-type tables from the same species pairs, we can also model variation in selection among genes. This information can be analyzed using a hierarchical Bayesian framework to describe the probability distribution of the selection intensity, γ, for each individual Arabidopsis gene (Bustamante *et al.* 2002b; see appendix). These selection intensities can be considered as the relative levels of selection on amino acid replacements with respect to synonymous site changes. Variation in selection among genes is modeled as normally distributed with unknown mean, μ, and variance, σ^2, for both the rapidly evolving protein and the previously studied gene classes.

The Markov chain Monte Carlo (MCMC) sampling scheme described in the appendix was used to draw from the joint posterior probability distribution of several parameters in the model given the data in the McDonald-Kreitman tables for all 18 genes. These include the mean and variance parameters for previously studied (μ_1, σ_1) and rapidly evolving protein loci (μ_2, σ_2), the scaled species divergence parameter (t), the vectors of mutation rates at synonymous sites (θ^S) and replacement sites (θ^R), and the vector of selection coefficients (γ). At completion of the sampling scheme, we have 10,000 quasi-independent vectors for each parameter in the model drawn from the joint probability distribution of the parameters given the data.

The distribution of the sampled values of γ for each locus [$\gamma_i^{(1)}, \gamma_i^{(2)}, \ldots, \gamma_i^{(10,000)}$ for $1 \leq i \leq 18$] within and between the two classes of Arabidopsis genes provides several striking results. The means of the γ draws for each gene in the class of previously studied Arabidopsis loci ranged from -2.285 to $+0.941$. Six of these loci have 95% HPCIs that are entirely <0 (see Figure 4). These negative selection intensities suggest that most amino acid replacements are slightly deleterious and persist due to the inbreeding associated with the predominant selfing observed in this species (Bustamante *et al.* 2002b). In contrast, the means of the γ samples for the rapidly evolving protein loci ranged from -0.823 to 0.856. Three of the six rapidly evolving protein genes (50%) have $\gamma < 0$. Only one of these genes, which encodes a protein of unknown function, has a 95% HPCI <0. Three genes have $\gamma > 0$, although the 95% HPCIs for all of these also encompass 0.

The posterior distribution of μ, the average selective effect of amino acid replacement changes for rapidly evolving protein genes, shows a shift in the positive direction (see Figure 5). The previously studied Arabidopsis loci have a posterior mean for the average selective effect of amino acid replacements (μ_1) of -0.9622. We find that the posterior probability that $\mu_1 > 0$, $P(\mu_1 \geq 0)$, is 0.02. In contrast, the posterior distribution for μ_2 (the average selective effect of amino acid replacement in the rapidly evolving protein genes) has a mean of $+0.1201$ and $P(\mu_2 \geq 0) \sim 0.56$. The posterior mean of the difference between the average selective effects ($\mu_1 - \mu_2$) is -1.0823. This analysis indicates that amino acid replacement changes in these rapidly evolving protein genes are more beneficial (or less deleterious) than those found in previously studied nuclear loci.

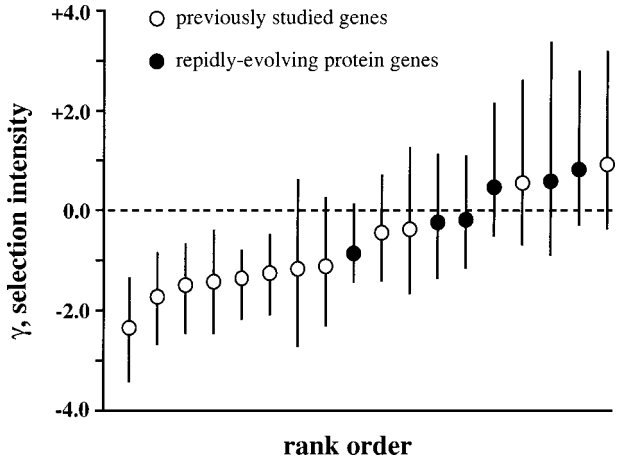

Figure 4.—Selection intensities γ of *A. thaliana* nuclear genes. The open and solid circles are γ values for previously studied and rapidly evolving *A. thaliana* protein genes, respectively. The bars indicate the 95% highest posterior credible intervals of γ estimates. The genes from left to right are *AP3*, *PgiC*, *PI*, *AP1*, *ChiA*, *CAL*, *TFL1*, *FAH1*, hypothetical gene (AT1G67140), *Adh1*, *F3H*, unknown gene (AT2G04410), putative gene (AT4G15950), hypothetical gene (AT3G22060), *LFY*, *p55.11*-like gene (AT4G28470), *NAC2*, and *CHI*.

Evolution of rapidly evolving protein genes in the genome: Approximately 5% of the inflorescence-expressed genes examined in this evolutionary study have values of $\omega^* > 1$ between *A. thaliana* and *A. lyrata* orthologs and are potential candidates for genes associated with adaptive divergence between these two species. The high proportion of genes with $\omega^* > 1$ suggests that rapidly evolving protein-coding loci may represent a significant portion of genes in eukaryotic genomes. This proportion, however, is higher than that observed in a similar analysis between *A. thaliana* and *B. rapa*, two species that last shared a common ancestor ~35 MYA (Tiffin and Hahn 2002). On the basis of 218 coding sequences from a floral EST-based *B. rapa* data set, no gene was identified with $\omega > 1$. A larger study using 3595 gene sequences across all species represented in DNA databases also indicates that the number of genes that have $\omega > 1$ is <0.5%, suggesting the number of genes in this class may be very low (Endo *et al.* 1996). In these two studies, however, the comparisons were between distantly related taxa. It is likely that the ability to identify genes with $\omega > 1$ may be facilitated by using more closely related species, where the signature for

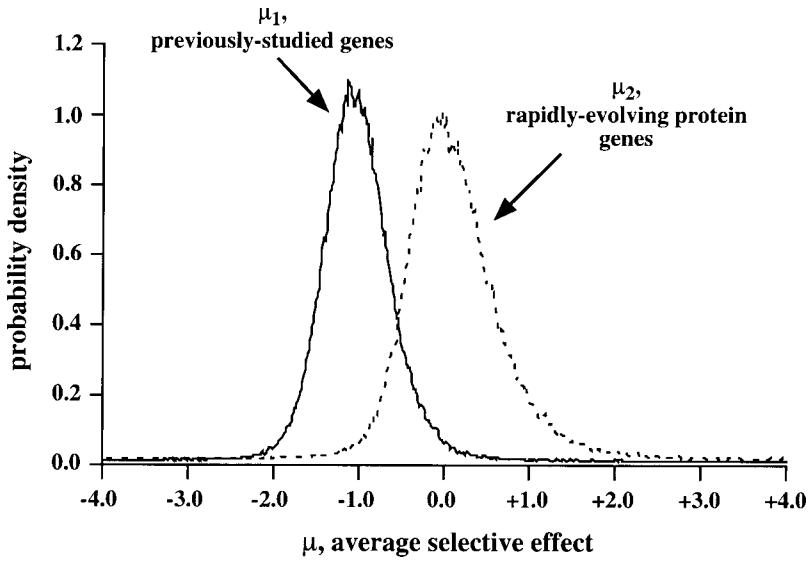

FIGURE 5.—Distribution of selection coefficients for *A. thaliana* nuclear genes. μ_1 (solid line) and μ_2 (dotted line) are the selection coefficient distributions for previously studied and rapidly evolving *A. thaliana* protein loci.

accelerated nonsynonymous substitution, possibly arising from positive selection, may be more readily apparent. Indeed, a study of male accessory gland ESTs from closely related Drosophila species has identified 11% of genes with $\omega > 1$ (SWANSON *et al.* 2001).

Molecular population genetic analysis confirms the increased selection intensities associated with genes that display accelerated rates of nonsynonymous evolution. In previously studied *A. thaliana* nuclear genes, most replacement changes are slightly deleterious and their estimated selection intensities are generally negative (BUSTAMANTE *et al.* 2002b). Many *A. thaliana* nuclear loci studied to date possess high levels of within-species replacement polymorphisms (PURUGGANAN and SUDDITH 1998), few of which go to fixation and contribute to differences between *A. thaliana* and *A. lyrata*. In contrast, the class of rapidly evolving protein genes that were identified in this evolutionary EST study as having accelerated rates of nonsynonymous evolution generally possesses higher selection intensities on amino acid replacements. This is evident in the shift of the distribution of selection coefficients, μ, toward positive values compared to the distribution of previously studied *A. thaliana* genes (see Figure 5).

Positive selection associated with the fixation of protein sequence variants may explain the increased selection intensities on these genes. The increase in selection intensities for rapidly evolving protein genes may also arise, however, from neutral evolutionary forces on replacement polymorphisms, leading to increased fixation of amino acid changes. This is underscored by the distribution of selection coefficients, μ_2, for the rapidly evolving protein gene class, which while shifted to the positive direction is nevertheless centered near $\mu = 0$ (see Figure 5). All the rapidly evolving protein genes used in the molecular population genetic study, however, are expressed in both species and there are no premature stop codons in these loci. This suggests that if neutral evolution underlies the increased selection intensities in these rapidly evolving protein loci, they do not appear to be associated with pseudogene formation.

It is likely that both neutral evolution and positive selection may be responsible for the rapidly evolving protein genes identified in this evolutionary EST study. Nevertheless, evolutionary ESTs do appear to provide a general genomic approach to identify loci associated with increased selection intensities on protein sequence, some of which may underlie adaptive evolution between these species. It should be noted that while some of the loci with $\omega^* > 1$ identified in this evolutionary EST study may be associated with adaptive divergence, this may underestimate the role of positive or diversifying selection in shaping gene structure and function in the genome. The criteria of $\omega^* > 1$ as an indicator of selection can be overly stringent, as it requires that amino acid fixations occur throughout the gene and does not recognize adaptive fixation of small numbers of replacement changes. Moreover, it does not identify genes in which the selective force acts on regulatory regions of the gene, which is believed to be a major factor in adaptive divergence between species (DOEBLEY and LUKENS 1998; SUCENA and STERN 2000).

The function of the genes that encode rapidly evolving proteins remains largely unknown. Of the three genes in the molecular population genetic analysis that have selection intensities >0, one encodes a previously unknown protein while the other two are homologous to known genes in *A. thaliana* or other eukaryotic organisms. One gene is *NAC2*, which belongs to a family of transcription factors required in Arabidopsis development (XIE *et al.* 2000). The other gene encodes a protein homologous to the human p55.11 protein that binds to the tumor necrosis factor p55 receptor (BOLDIN *et al.* 1995). The precise functions of these genes in Arabidopsis remain to be elucidated. Expression studies as well as phenotypic analysis of T-DNA insertion mutants

for these genes may permit an assessment of their functions and may also provide clues as to the traits that are the targets of selection in the divergence between *A. thaliana* and *A. lyrata*.

The authors thank the NCSU Genome Research Laboratory for use of its facilities and the Purugganan laboratory and three anonymous reviewers for helpful suggestions that improved the article. We are also grateful to Jeff Thorne for suggesting the permutation analysis. M.B. was funded by a National Institutes of Health training grant on the genetics of complex traits and C.D.B. by a Marshall-Sherfield Fellowship from the Marshall Commemoration Commission. This work was funded in part by grants from the National Science Foundation Integrated Research Challenges in Environmental Biology program to M.D.P, J. I. Schmitt, and T. F. C. Mackay; and from the Alfred P. Sloan Foundation to M.D.P.

LITERATURE CITED

Aguade, 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.

Anisimova, M., J. P. Bielawsi and Z. H. Yang, 2001 Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol. Biol. Evol. **18:** 1585–1591.

Arabidopsis Genome Initiative, 2000 Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408:** 796–815.

Boldin, M. P., I. L. Mett and D. Wallach, 1995 A protein related to a proteasomal subunit binds to the intracellular domain of the p55 TNF receptor upstream to its 'death domain'. FEBS Lett. **367:** 39–44.

Bustamante, C. D., R. Nielsen and D. L. Hartl, 2002a A maximum likelihood method for analyzing pseudogene evolution: implications for silent site evolution in humans and rodents. Mol. Biol. Evol. **19:** 110–117.

Bustamante, C. D., R. Nielsen, S. Sawyer, K. M. Olsen, M. D. Purugganan *et al.*, 2002b The cost of inbreeding in Arabidopsis. Nature **416:** 531–534.

Doebley, J., and L. Lukens, 1998 Transcriptional regulators and the evolution of plant form. Plant Cell **10:** 1075–1082.

Endo, T., K. Ikeo and T. Gojobori, 1996 Large-scale selection for genes on which positive selection may operate. Mol. Biol. Evol. **13:** 685–690.

Ewing, B., L. Hillier, M. C. Wendl and P. Green, 1998 Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. **8:** 175–185.

Gelman, A., J. S. Carlin, H. S. Stern and D. B. Rubin, 1997 *Bayesian Data Analysis*. Chapman & Hall, Boca Raton, FL.

Haag, E. S., and J. R. True, 2001 Perspective: From mutants to mechanisms? Assessing the candidate gene paradigm in evolutionary biology. Evolution **55:** 1077–1084.

Hughes, A. L., 1991 Circumsporozoite protein genes of malaria parasites (*Plasmodium* spp.): evidence for positive selection on immunogenic regions. Genetics **127:** 345–353.

Hughes, A. L., and M. Yeager, 1998 Natural selection at major histocompatibility complex loci of vertebrates. Annu. Rev. Genet. **32:** 415–435.

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

Kumar, S., K. Tamura and M. Nei, 1993 *MEGA: Molecular Evolutionary Genetics Analysis*, Version 1.0. Pennsylvania State University, University Park, PA.

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

Li, W.-H., T. Gojobori and M. Nei, 1981 Pseudogenes as a paradigm of neutral evolution. Nature **292:** 237–239.

Li, W.-H., C.-I Wu and C. C. Luo, 1985 A new method for estimating synonymous and nonsynonymous rates on nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. **2:** 150–174.

McDonald, J., and M. Kreitman, 1991 Adaptive protein evolution at the *Adh* locus in Drosophila. Nature **351:** 652–654.

Messier, W., and C. B. Stewart, 1997 Episodic adaptive evolution of primate lysozymes. Nature **385:** 151–154.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller and E. Teller, 1953 Equation of state calculations by fast computing machines. J. Chem. Phys. **21:** 1087–1092.

Miyashita, N. T., A. Kawabe, H. Innan and R. Terauchi, 1998 Intra- and interspecific DNA variation and codon bias of the *Alcohol Dehydrogenase* (*Adh*) locus in Arabis and Arabidopsis species. Mol. Biol. Evol. **15:** 1420–1429.

Nei, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Nei, M., and T. Gojobori, 1986 Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3:** 418–426.

Nielsen, R., 2001 Statistical tests of selective neutrality in the age of genomics. Heredity **86:** 641–647.

Olsen, K. M., A. Womack, A. R. Garrett, J. I. Suddith and M. D. Purugganan, 2002 Contrasting evolutionary forces in the *Arabidopsis thaliana* floral developmental pathway. Genetics **160:** 1641–1650.

Orr, H. A., and J. A. Coyne, 1992 The genetics of adaptation: a reassessment. Am. Nat. **140:** 725–742.

Purugganan, M. D., and J. I. Suddith, 1998 Molecular population genetics of floral homeotic loci: departures from the equilibrium-neutral model at the *AP3* and *PI* genes of *Arabidopsis thaliana*. Genetics **151:** 839–848.

Rozas, J., and R. Rozas, 1999 DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15:** 174–175.

Sawyer, S. A., and D. L. Hartl, 1992 Population genetics of polymorphism and divergence. Genetics **132:** 1161–1176.

Schmid, K. J., and C. F. Aquadro, 2001 The evolutionary analysis of "orphans" from the Drosophila genome identifies rapidly diverging and incorrectly annotated genes. Genetics **159:** 589–598.

Sucena, E., and D. L. Stern, 2000 Divergence of larval morphology between *Drosophila sechellia* and its sibling species caused by cis-regulatory evolution of *ovo/shaven-baby*. Proc. Natl. Acad. Sci. USA **97:** 4530–4534.

Swanson, W. J., and V. D. Vacquier, 1995 Extraordinary divergence and positive Darwinian selection in a fusagenic protein coating the acrosomal process of abalone spermatozoa. Proc. Natl. Acad. Sci. USA **92:** 4957–4961.

Swanson, W. J., A. G. Clark, H. M. Waldrip-Dail, M. F. Wolfner and C. F. Aquadro, 2001 Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in Drosophila. Proc. Natl. Acad. Sci. USA **98:** 7375–7379.

Tiffin, P., and M. Hahn, 2002 Coding sequence divergence between two closely-related plant species—*Arabidopsis thaliana* and *Brassica rapa* ssp. *pekinensis*. J. Mol. Evol. **54:** 746–753.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

Wu, C.-I, 2001 The genic view of the process of speciation. J. Evol. Biol. **14:** 851–865.

Xie, Q., G. Frugis, D. Colgan and N. H. Chua, 2000 *Arabidopsis NAC1* transduces auxin signal downstream of *TIR1* to promote lateral root development. Genes Dev. **14:** 3024–3036.

Zhang, J., H. F. Rosenberg and M. Nei, 1998 Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc. Natl. Acad. Sci. USA **95:** 3708–3713.

Communicating editor: M. K. Uyenoyama

APPENDIX

It has been shown (Sawyer and Hartl 1992) that under the assumptions of the Poisson random field setting, the sampling distributions for the number of polymorphic sites within species, S, and fixed differences

between two species, K, are independent Poisson-distributed random variables with rates

$$E(S) = \theta\frac{2\gamma}{1 - e^{-2\gamma}}(F(\gamma, m) + F(\gamma, n)) \quad \text{(A1)}$$

$$E(K) = \theta\frac{2\gamma}{1 - e^{-2\gamma}}(G(\gamma, m) + G(\gamma, n) + t), \quad \text{(A2)}$$

where γ is the scaled selection coefficient for new mutations ($2N_e s$), θ is the scaled mutation rate ($4N_e \cdot \mu$), t is the scaled time since *species* divergence (number of generations since divergence/$2N_e$), N_e is the effective population size, and n and m are the sample sizes from the two species. The functions $F(\gamma, n)$ and $G(\gamma, n)$ are as previously described (Sawyer and Hartl 1992; Bustamante *et al.* 2002b).

Using the cell entries from a conventional McDonald-Kreitman table (McDonald and Kreitman 1991), it is possible to estimate four parameters in such a model: θ^S (mutation rate for silent sites), θ^R (mutation rate for replacement sites), t, and γ (selection intensity for replacement sites) assuming silent sites are neutral (*i.e.*, $\gamma = 0$ for all silent sites). For a set of such tables from the same species pairs, it is also possible to model variation in selection among genes by specifying a distribution for γ and estimating the parameters of this hyperdistribution given the data in all of the tables. A convenient form to use is the normal distribution since selection coefficients can be either positive or negative. It should also be noted that the species divergence time is a shared parameter across all the tables in such an analysis.

Hierarchical model: The analysis we present is based on a description of the joint and marginal posterior probability distributions of the following model, described in three parts.

Part 1: Let γ be the vector of selection coefficients, with $\gamma_1, \ldots, \gamma_{12}$ being the set of previously studied loci (Bustamante *et al.* 2002b) and $\gamma_{13}, \ldots, \gamma_{18}$ representing rapidly evolving protein loci. θ^R and θ^S are the corresponding vectors of mutation rates at replacement and silent sites. Denote the mean and variance of the distribution of γ among previously studied genes as μ_1 and σ_1^2 and use μ_2 and σ_2^2 for the analogous quantities for the rapidly evolving protein genes.

Part 2: Set a truncated uniform prior distribution for t on $(0, T)$, where T is chosen on the basis of prior information on the upper bound for the species divergence time. We used $T = 100$, corresponding to an upper bound of between 20 and 200 million years ago.

Part 3: Assume a normal conjugate prior probability distribution for the mean and variance parameters for each of the two classes of genes so that

$$\mu_i|\sigma_i^2 \sim N\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right) \quad \text{(A3)}$$

$$\sigma_i^2 \sim \text{Inv} - \chi^2(\nu_0, \sigma_0^2), \quad \text{(A4)}$$

where μ_0, κ_0, ν_0, and σ_0^2 are parameters of the prior distributions for μ_i and σ_i^2, Inv $- \chi^2$ refers to an inverse χ^2 distribution, and $i \in \{1, 2\}$ indexes the two sets of hyperparameters for both classes of genes. The notation is borrowed from Gelman *et al.* (1997). Note that if κ_0 and ν_0 are chosen to be small and σ_0^2 to be large, the prior distribution will be uninformative. In our runs we used $\mu_0 = 0$, $\sigma_0^2 = 100$, $\kappa_0 = 0.001$, $\nu_0 = 0.001$.

Results for conditional posterior distributions: The joint posterior distribution of interest $p(\gamma, \theta^R, \theta^S, t, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2|\text{data})$ can be approximated using a Markov chain Monte Carlo sampling scheme similar to that implemented in Bustamante *et al.* (2002b) using the following results:

Result 1: The conditional posterior distribution of $\mu_1|\sigma_1^2, \gamma$ depends only on the entries in γ that are members of the class i and can be shown to be normally distributed as

$$\mu_i|\sigma_i^2, \gamma \sim N\left(\frac{(\kappa_0/\sigma_i^2)\mu_0 + (J_i/\sigma_i^2)\overline{\gamma}_i}{\kappa_0/\sigma_i^2 + J_i/\sigma^2}, \frac{1}{\kappa_0/\sigma_i^2 + J_i/\sigma_i^2}\right), \quad \text{(A5)}$$

where $\overline{\gamma}_i$ is the arithmetic average of the entries in γ for class i and J_i is the number of genes in the class.

Result 2: The marginal posterior distribution of σ_i^2 conditional on γ, which depends only on the sample variance of the entries in γ that are members of the class i and the parameters of the prior distribution, has an inverse χ^2 distribution with parameters ν_{J_i} and $\sigma_{J_i}^2$ as given in Gelman *et al.* (1997).

Result 3: Using independent Gamma prior distributions with parameters α_0 and β_0 for each of the mutation rates yields independent Gamma posterior distributions conditional on t and γ with parameters $\alpha_0 + K + S$ and

$$\beta_0 + \frac{2\gamma}{1 - e^{-2\gamma}}(F(\gamma, m) + F(\gamma, n) + t + G(\gamma, n) + G(\gamma, m)).$$

The mean of this distribution is α/β and the variance is α/β^2. As such, if α_0 and β_0 are chosen to be small, the prior will be uninformative. For all our analysis, we used $\alpha = \beta = 0.001$.

Result 4: The posterior distribution $p(t|\theta^R, \theta^S, \gamma, \text{data})$ is proportional to the likelihood function at the point $(t, \theta^R, \theta^S, \gamma)$ if $t < T$ and 0 otherwise.

Result 5: The joint conditional posterior distribution $p(\gamma|\theta^R, \theta^S, t, \mu, \sigma^2, \text{data})$ factors into the product of the individuals' conditional distributions $p(\gamma_j|\theta_j^R, \theta_j^S, t, \mu_i, \sigma_i^2, K_j^R, S_j^R, \text{data})$. Furthermore, the conditional posterior distribution $p(\gamma_j|\theta_j^R, \theta_j^S, t, \mu_i, \sigma_i^2, K_j^R, S_j^R)$ for a given gene j in class i is proportional to the product of the likelihood for the gene given θ_j^R, θ_j^S, γ_j, and t, and the probability density of a normal distribution with mean μ_i and variance σ_i^2, at the point γ_j.

Markov chain Monte Carlo algorithm: Given the model and results outlined above, it is then possible to sample from $p(\gamma, \theta^R, \theta^S, t, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2|\text{data})$ using the following algorithm (Metropolis *et al.* 1953) for each chain. The algorithm we employ in this analysis has the following steps.

Step 1: Initialize γ by drawing a value for γ_j for $1 \leq j \leq J_1 + J_2$ independently from a normal distribution with mean near 0 and a reasonably large variance. We used several starting values for the mean in the range $[-5, 5]$ and the variance in the range $[1, 100]$.

Step 2: Using the values in γ, update σ_i^2 for $i \in \{1, 2\}$ by sampling from the conditional distribution of $\sigma_i^2|\gamma$, which is inverse-χ^2 distributed as detailed above.

Step 3: Using the values in γ and σ_i^2 for $i \in \{1, 2\}$, update μ_i by sampling a new value from a normal distribution with the updated parameters in Result 1 above.

Step 4: Update t by using Metropolis sampling.

a. Sample a proposal value t' from a $U(t - \delta_t, t + \delta_t)$ distribution.

b. If $p(t'|\theta^R, \theta^S, \gamma \text{ data}) > p(t|\theta^R, \theta^S, \gamma|\text{data})$, set $t = t'$. Otherwise, set $t = t'$ with probability proportional to the ratio of these two quantities.

Step 5: Update each γ_j in γ by using $J_1 + J_2$ independent Metropolis steps as follows.

a. Sample a proposal value γ_j' from a $U(\gamma_j - \delta, \gamma_j + \delta)$.

b. If $p(\gamma_j'|\theta_j^R, \theta_j^S, t, \mu_i, \sigma_i^2, S_j^R, K_j^R) > p(\gamma_j|\theta_j^R, \theta_j^S, t, \mu_i, \sigma_i^2, S_j^R, K_j^R)$, set $\gamma_j = \gamma_j'$. Otherwise, set $\gamma_j = \gamma_j'$ with probability proportional to their ratio.

Step 6: For each gene, draw a value for θ_j^R and θ_j^S using the result that the posterior distribution for $\theta_j|\gamma_j, t$ is a Gamma distribution with parameters as described in Result 3 above.

Step 7: Repeat steps 2–6.

We used the above algorithm to approximate the joint posterior distributions using 10 different starting points (*i.e.*, 10 different chains) run for 10,000 steps each after an initial 2000-step burn-in and retention of draws every 10 steps (for a total of 10,000 draws for each parameter in the model). For the Metropolis step for updating t, we used a proposal distribution with $\delta_t = 1.0$, which gave a rejection rate of $\sim$26.36% for the 100,000 draws retained after the initial burn-in. To measure convergence we used a $\sqrt{\hat{R}}$ statistic that was below 1.01 for all parameters in the model before samples were retained (conventionally one retains after 1.2 or less), illustrating that the 10 chains had converged well before we retained our samples.