

Coding Sequence Divergence Between Two Closely Related Plant Species: *Arabidopsis thaliana* and *Brassica rapa* ssp. *pekinensis*

Peter Tiffin,¹ Matthew W. Hahn²

¹ Department of Ecology and Evolutionary Biology, University of California at Irvine, Irvine, CA 92664, USA

² Evolution, Ecology, and Organismal Biology Group, Department of Biology, Box 90338, Duke University, Durham, NC 27708, USA

Received: 25 June 2001 / Accepted: 2 November 2001

Abstract. To characterize the coding-sequence divergence of closely related genomes, we compared DNA sequence divergence between sequences from a *Brassica rapa* ssp. *pekinensis* EST library isolated from flower buds and genomic sequences from *Arabidopsis thaliana*. The specific objectives were (i) to determine the distribution of and relationship between K_a and K_s , (ii) to identify genes with the lowest and highest $K_a:K_s$ values, and (iii) to evaluate how codon usage has diverged between two closely related species. We found that the distribution of $K_a:K_s$ was unimodal, and that substitution rates were more variable at nonsynonymous than synonymous sites, and detected no evidence that K_a and K_s were positively correlated. Several genes had $K_a:K_s$ values equal to or near zero, as expected for genes that have evolved under strong selective constraint. In contrast, there were no genes with $K_a:K_s > 1$ and thus we found no strong evidence that any of the 218 sequences we analyzed have evolved in response to positive selection. We detected a stronger codon bias but a lower frequency of GC at synonymous sites in *A. thaliana* than *B. rapa*. Moreover, there has been a shift in the profile of most commonly used synonymous codons since these two species diverged from one another. This shift in codon usage may have been caused by stronger selection acting on codon usage or by a shift in the direction of mutational bias in the *B. rapa* phylogenetic lineage.

Key words: Synonymous substitution — Nonsynonymous substitution — $K_a:K_s$ — Codon bias — Mutation — Evolutionary rates — Genome

Introduction

There is a wide variation in the rates at which individual genes diverge between species—some genes are highly conserved and others diverge rapidly (Li 1997; Hughes 1999). However, little is known about the patterns of genomewide sequence divergence. One reason for this is that most analyses of sequence evolution have focused on individual genes or small gene families that have been investigated because of their biological functions, phenotypic effects, or expected evolutionary histories. Because investigations of individual genes have provided data on only a tiny fraction of the total number of genes within a genome, they may not provide an accurate description of genomewide DNA sequence divergence. A more robust picture of genomic divergence, one that may be useful for discerning the relative effects different evolutionary forces have had in causing that divergence, may be attained by examining large numbers of genes that have been selected without prior interest in their biological function or evolutionary histories (Charlesworth et al. 2001).

Differences in selection and mutation both contribute to the variation in rates of interspecific gene divergence. Analysis of individual genes or small groups of genes have identified loci that appear to have diverged in response to positive selection (Yang and Bielawski 2000)

Correspondence to: Peter Tiffin, Department of Plant Biology, University of Minnesota, 1445 Gortner Avenue, St. Paul, MN 55108-1095, USA; email: ptiffin@umn.edu.

and others that have evolved under varying levels of selective constraint (Rausher et al. 1999). The relative effects of positive selection and selective constraint (often referred to as “negative selection” or “purifying selection”) in determining the rate of divergence between gene sequences can be assessed by the ratio of substitutions at replacements sites (K_a) to substitutions at synonymous sites (K_s) (Hughes 1999; Yang and Bielawski 2000). A $K_a:K_s$ ratio <1 is consistent with a history of negative selection, although it does not rule out positive selection, and a $K_a:K_s$ ratio >1 indicates strong positive selection, although it does not mean that negative selection is not also acting (Yang and Bielawski 2000). Although $K_a:K_s$ values have the advantage of being easy to calculate, it is recognized that the sensitivity of $K_a:K_s$ ratios to detect positive selection is low and that the criterion of a $K_a:K_s$ ratio >1 as evidence for positive selection may be overly stringent (Hughes 1999). Nevertheless, the distribution of $K_a:K_s$ for a large number of genes may be useful for identifying the relative strengths of the evolutionary forces acting on individual genes—those with the lowest $K_a:K_s$ values are likely to be evolving under the strongest selective constraint, whereas those with the highest may be evolving in response to positive selection or relaxed constraint (Charlesworth et al. 2001). Although the distribution of $K_a:K_s$ values for large numbers of *Drosophila*, nematode, and mammalian genes has been examined (Wolfe and Sharp 1993; Makalowski and Boguski 1998; Wheelan et al. 1999), the distribution of $K_a:K_s$ values has not been identified for any pair of plant species.

While $K_a:K_s$ provides insight into the forces of selection acting on nonsynonymous sites, it implicitly assumes that selection is not acting on synonymous sites. However, selection acting to increase translational efficiency or translational accuracy may cause adaptive change at synonymous sites. Evidence for selection acting on synonymous mutations comes from the widespread nonrandom use of synonymous codons [“codon bias” (Ikemura 1981, 1985; Shields et al. 1988; Akashi 1994, 1995; Moriyama and Powell 1997)]. If the strength or effectiveness of selection acting on codon usage differs between closely related species, then selective forces acting on synonymous sites may also contribute to interspecific genome divergence. Moreover, interspecific differences in the strength or direction of mutational bias, which may also cause differences in codon usage (Sueoka 1988; Eyre-Walker 1991), may also contribute to divergence at synonymous sites. Differences in codon bias between closely related *Drosophila* species have been detected (Akashi 1996; Rodriguez-Trelles et al. 1999), but codon usage in closely related plant species has not been compared.

Genomewide sequence comparisons would ideally be examined using entire genome sequences; however, complete genome data are not yet available for closely

related eukaryotic species. In the absence of complete genome data, expressed sequence tag (EST) data offer an opportunity to examine large numbers of coding sequences that have been sequenced without bias to their function or expected evolutionary history. To characterize the pattern of DNA sequence divergence and codon bias between two closely related plant species, we compared sequences from 218 *Brassica rapa* ssp. *pekinensis* (Chinese cabbage) flower tissue expressed sequence tags to genome sequences from *A. thaliana*. These species were chosen because they are model dicot species and because preliminary investigations revealed that the sequences of these species have diverged considerably but are not so different as to cause difficulty with sequence alignment. The species themselves are estimated to have diverged from one another approximately 35 million years ago (Lagercrantz 1998). Our specific objectives were (i) to determine the distribution of $K_a:K_s$ and identify genes with the lowest and highest $K_a:K_s$ values and (ii) to determine how selective and mutational forces between these species have contributed to changes in codon usage and thus interspecific sequence divergence.

Materials and Methods

The data were obtained by searching the *Arabidopsis thaliana* sequences in GenBank for homologues of 310 sequences from a *Brassica rapa* subsp. *pekinensis* flower bud cDNA library [S.W. Ryu, C.O. Lim, and M.J. Cho, National University of Korea, 1999 (unpublished); accession numbers AT001683–AT002257]. The EST data set available from GenBank contained 395 sequences but 85 of these ESTs were not included in the analyses because BLASTN revealed that they had significant similarity to other ESTs within the same data set. When redundant ESTs were detected, only the longest sequence was included in analyses. The GenBank database was initially searched using BLASTN (Altschul et al. 1990), and only *B. rapa* sequences that exhibited significantly similarity (E values $<1 \times 10^{-7}$) to 150 bp or more of an *A. thaliana* sequence were analyzed further. *B. rapa* sequences with no significant matches in the genomic database were also used to search the database of *A. thaliana* ESTs. If more than one *A. thaliana* sequence exhibited significant similarity to the *B. rapa* sequence, then we assumed that the *A. thaliana* sequence with the greatest similarity is orthologous, and only this sequence was analyzed. BLASTN searches were initially conducted during July 2000. *B. rapa* sequences that did not show significant similarity to *A. thaliana* in this initial search were used to search GenBank again during April 2001. In addition, BLASTX and TBLASTX were used to search GenBank with all sequences for which a significant *Arabidopsis* homologue was not detected using BLASTN. These amino acid sequence-based algorithms detected homologues for 17 ESTs for which BLASTN detected no homologue.

For significant matches, open reading frames were identified using BioEdit (Hall 1999) and untranslated regions were removed. Because many of the sequences contained multiple open reading frames and because calculation of $K_a:K_s$ and measures of codon bias required accurate identification of coding regions, open reading frames were compared against the GenBank database using BLASTp (Altschul et al. 1990). None of the analyzed sequences contained complete protein coding sequences. Open reading frames that did not exhibit a significant match with previously identified putative or actual proteins were excluded from further analyses. EST sequences may contain sequencing errors; however, because errors should be distributed among syn-

onymous and nonsynonymous sites at equal frequencies and because K_a and K_s are measures of substitution per site, sequencing errors are not expected to bias strongly the results of our analyses.

To assess differences in mutational bias the sequences from 15 genes of full length that were available from both *B. rapa* and *A. thaliana* were examined. These genes were (*A. thaliana*/*B. rapa* accession numbers) 2-Cys peroxiredoxin (X97910/AF052202), amino alcohol phosphotransferase (AF091843/AF183933), acidic endochitinase (AB006069/AF207563), acyl-ACP thioesterase (Z36912/U17098), cinnamyl-alcohol dehydrogenase (AL161595/AF207555), 4-coumarate-CoA ligase (AL161549/AF207574), ω -3 fatty acid desaturase (D26508/AF056572), glutathione reductase (D89620/AF008441), oxoacyl-ACP reductase (X64464/AF179864), protease inhibitor II (AC005936/L31937), stearyl-ACP desaturase (AC002333/X60978), Cu/Zn superoxide dismutase (X60935/AF071112), terminal flower 1 (D87519/AB017529), thionin (L41245/AF090836), and thioredoxin (Z35474/AB010434). The untranslated regions (not including poly-A tails) and/or introns of these genes were used to analyze interspecific changes in GC content in regions thought to be under no selective pressure for nucleotide comparison.

Sequence Analysis

K_a (the number of nonsynonymous differences divided by the number of nonsynonymous sites) and K_s (the number of synonymous differences divided by the number of synonymous sites) were calculated using DNAsp 3.50 (Rozas and Rozas 1999). DNAsp uses the method of Nei and Gojobori (1986) to identify synonymous and nonsynonymous sites and a Jukes-Cantor correction is applied to correct for multiple hits. The magnitude of codon bias in each sequence was estimated as the effective number of codons (ENC) (Wright 1990). ENC values range between 20 and 61 and are inversely related to codon bias. ENC values of 20 indicate that only a single codon is used for each amino acid, whereas an ENC of 61 indicates that all synonymous codons are used at equal frequencies. We used the ENC to estimate codon bias because it does not presuppose any knowledge of a set of optimal codons and because this metric is relatively unbiased when short sequences are analyzed (Wright 1990; Comeron and Aguade 1998). Nevertheless, calculating the ENC on short sequences can lead to estimates greater than 61. We set the ENC equal to 61 for all sequences with calculated ENC values greater than 61 (20 sequences had an ENC >61 in one or both species). For each sequence we also calculated the percentage of GC at synonymous sites (GC_{syn}), a measure thought to be positively correlated with codon bias in most species, including the plant species *Zea mays* and *A. thaliana* (Fennoy and Bailey-Serres 1993; Chiappello et al. 1998). Paired *t* tests in which the homologous sequences were paired were used to test whether ENC and GC_{syn} differed significantly between species. The *t* tests and correlations between K_a , K_s , ENC, and GC_{syn} were calculated using SAS (SAS Institute 1989).

Results and Discussion

Comparison of the Arabidopsis and Brassica Genomes

Two-hundred eighteen, or approximately 70% of the 310 unique *B. rapa* ESTs, had significant similarity to hypothetical or known open reading frames at least 150 nucleic acids long in the *A. thaliana* genome. The average length of these 218 sequences was 235 bp. Many of the remaining sequences had significant similarity to *A. thaliana* sequences but the regions of similarity were

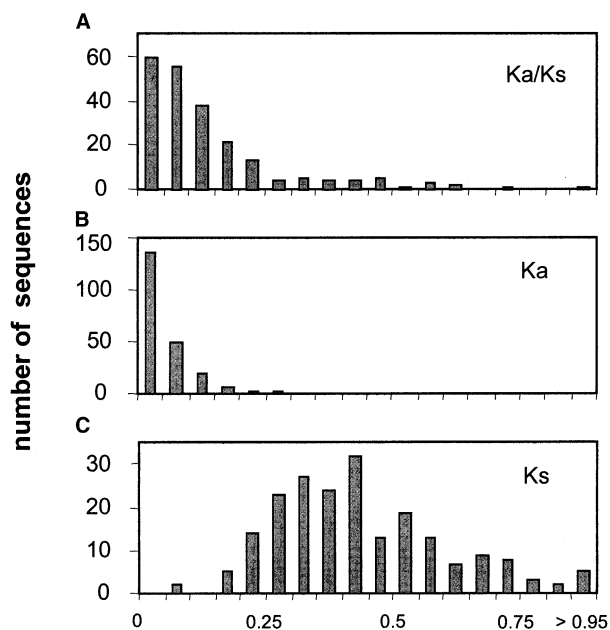


Fig. 1. Distributions of (A) the ratio of substitutions at nonsynonymous to those at synonymous sites ($K_a:K_s$), (B) K_a , and (C) K_s of 218 coding sequences from *Brassica rapa* ssp. *pekinensis* and *A. thaliana*.

shorter than 150 bp. Many of the sequences that were not significantly similar to *Arabidopsis* sequences were also not significantly similar to any other sequences in GenBank, including numerous *Brassica* EST sequences. For this reason we suspect that these ESTs contain 5' UTR regions of genes or noncoding contaminants. However, it is possible that significant homologues of these sequences were not detected because these sequences have been subject to strong positive selection and are no longer similar enough to *Arabidopsis* sequences for homologues to be detected using BLAST.

Divergence at Synonymous and Nonsynonymous Sites

The 218 sequence pairs had mean values of K_a , K_s , and $K_a:K_s$ of 0.058, 0.474, and 0.14 and ranged from 0 to 0.48, 0.076 to 1.88 (K_s was saturated in one sequence; this sequence was removed from the analyses), and 0 to 0.87, respectively (Fig. 1). The coefficient of variation (CV), a measure of the variability of a sample relative to the sample mean, of K_a and K_s was 1.06 and 0.49, respectively. A Z test (Zar 1996) indicated that the variability in K_a was significantly higher than in K_s ($Z = 5.6$, $p < 0.01$; to meet assumptions of normality samples were square root transformed prior to conducting the Z test). This pattern is similar to what has been found in mammalian systems (Wolfe and Sharp 1993). Although the CV at synonymous sites was significantly less than the CV at nonsynonymous sites, there was still a wide range in the rates at which synonymous substitutions accumulate, perhaps reflecting intergene differences in mutation rates or the relative strength of selection acting on codon

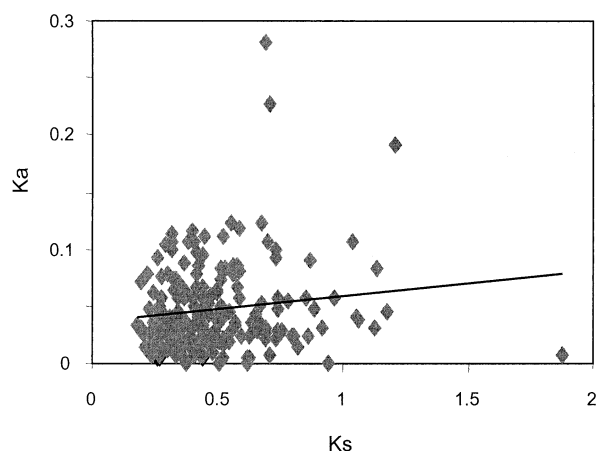


Fig. 2. Relationship between substitutions at nonsynonymous and those at synonymous sites. The estimated correlation between K_a and K_s was not significantly different from zero.

usage (Sharp and Li 1989). Regardless of the mechanisms, the large variation in K_s suggests that estimated times of species divergence or gene duplication events that are based on one or few genes may produce misleading results and these estimates should be viewed with caution.

In contrast to the significant positive correlations between K_a and K_s found between some *Drosophila* (Akashi 1994), bacterium (Sharp and Li 1987), mammal (Wolfe and Sharp 1993), and plant species (Alvarez-Valin et al. 1999), and between duplicated *A. thaliana* genes (Lynch and Conery 2000), we detected no evidence that the correlation between K_a and K_s from *A. thaliana* and *B. rapa* ssp. *pekinensis* was significantly different from zero ($r = 0.097$, $p = 0.14$) (Fig. 2). A positive correlation between K_a and K_s may result from intergenic differences in mutation rates (Ohta and Ina 1995), tandem mutations (mutations changing adjacent nucleotides simultaneously) (Wolfe and Sharp 1993), or correlated selective constraints on synonymous and nonsynonymous sites (Mouchiroud et al. 1995). The lack of correlation between K_a and K_s suggests either that these mechanisms are not operating in either *Arabidopsis* or *Brassica* or that other forces, such as mutational bias (see below), are strong enough to mask their effects.

Several sequences had $K_a:K_s$ values equal to zero, or only slightly greater than zero, suggesting that these genes have evolved under high selective constraint. The five sequences with the lowest $K_a:K_s$ values (all <0.005) showed significant similarity to 40S ribosomal protein S10, histone H3, an unidentified 40S ribosomal protein, sec61 translocation protein, and a heat-shock protein; all are housekeeping genes that typically evolve slowly (Li 1997).

There were no sequences with $K_a:K_s > 1$, and thus we found no strong evidence that positive selection has contributed to the interspecific sequence divergence of any of these genes. Nevertheless, because $K_a:K_s > 1$ is rec-

ognized as a very stringent criterion (Hughes 1999), identifying genes with the highest $K_a:K_s$ values may be useful for identifying genes that may have evolved in part in response to positive selection (Charlesworth et al. 2001). The sequences with the highest $K_a:K_s$ values had significant similarity to a cold-regulated protein ($K_a:K_s = 0.59$), a guanine-binding protein ($K_a:K_s = 0.61$), lipid-transfer proteins (two sequences: $K_a:K_s = 0.60$ and 0.75), and an anther-specific protein with a high similarity to putative cysteine-rich antifungal proteins ($K_a:K_s = 0.87$). Although these high $K_a:K_s$ values may reflect relaxed selective constraint, both lipid transfer proteins (Garcia-Olmedo et al. 1987) and antifungal proteins may be involved in plant defense against pathogens and may therefore be expected to evolve in response to positive selection.

Although we detected no evidence that positive selection has been an important force in causing divergence between the *A. thaliana* and *B. rapa* genomes, there are several reasons why the finding should be interpreted with caution. First, we examined only a small fraction of the nearly 26,000 open reading frames in *A. thaliana* (Arabidopsis Genome Initiative 2000), and if positive selection is rare, there is a high probability that we did not analyze enough genes to identify any that have evolved in response to strong positive selection. Based on samples from a wide range of taxa it has been estimated that less than 0.5% of genes have experienced positive selection strong enough to result in $K_a:K_s$ values >1 (Endo et al. 1996; Liberles et al. 2001). If the frequency of positively selected genes in *B. rapa* and *A. thaliana* is similar, then it is likely our sampling was not extensive enough to detect genes that have diverged in response to selection. Second, as discussed above, $K_a:K_s$ ratios, although an easy-to-apply screen, are an insensitive method for detecting positive selection (Hughes 1999; Yang and Bielawski 2000). Third, we analyzed only genes expressed in flower buds and expressed at levels high enough to be sampled from the EST library. As such, some genes, including regulatory genes that may be expressed at low levels but may be important for phenotypically important evolutionary changes (Doebly and Lukens 1998; Davidson 2001; Barrier et al. 2001), may not have been included in our analyses. Finally, identifying homologous sequences using BLAST searches may have resulted in ascertainment bias—those genes that have diversified in response to selection may no longer exhibit similarity significant enough to be identified with BLAST. This ascertainment bias would have biased our data set toward more conserved sequences.

Codon Bias and Shift in Codon Usage

Our analyses produced surprising results regarding the relative strength of codon bias in the two species, with

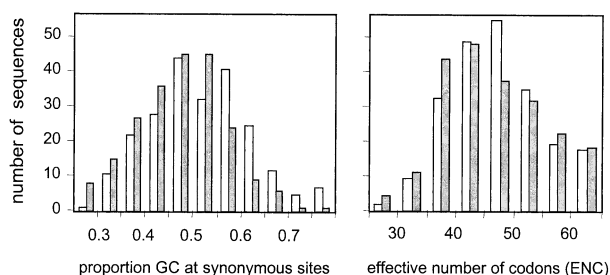


Fig. 3. Distribution of GC content at synonymous sites and effective number of codons (ENC) of 218 coding sequences from *Brassica rapa* ssp. *pekinensis* and *A. thaliana*. The white bars represent *B. rapa* data; the gray bars, *A. thaliana* data.

the ENC and GC content giving conflicting pictures of the strength of codon bias. The ENC, which is inversely related to the codon bias, was lower in *A. thaliana* than *B. rapa* (Fig. 3b). This difference was marginally significant when analyses were conducted on all data ($p = 0.075$), but the ENC was significantly lower in *A. thaliana* (ENC_{Arab}) than in *B. rapa* (ENC_{Brass}) when only sequences exhibiting some codon bias ($ENC < 61$) were analyzed ($ENC_{Arab} < ENC_{Brass}$; $t = 1.98$ $r < 0.05$), indicating stronger codon bias in *A. thaliana*. Although codon bias, as measured by ENC, was on average higher in *A. thaliana*, GC content at synonymous sites (GC_{syn}), which is often thought to be positively correlated with codon bias (Fennoy and Bailey-Serres 1993; Chiapello et al. 1998), was significantly higher in *B. rapa* (Fig. 3a) ($t = 6.93$, $p < 0.0001$). This difference was also highly significant when only sequences with an $ENC < 61$ were analyzed ($p < 0.0001$). As expected, ENC was negatively and significantly correlated with GC_{syn} in *B. rapa* ($r = -0.29$, $p < 0.0001$). However, ENC and GC_{syn} were not significantly correlated in *A. thaliana* ($r = -0.006$, $p > 0.90$).

The significant correlation between codon bias and GC_{syn} in *B. rapa*, and the lack of a significant relationship in *A. thaliana* can be explained if the most commonly used codons in *A. thaliana* end in A or T nucleotides and those from *B. rapa* end in G or C. To explore this hypothesis we used data from the Codon Usage Tabulated from GenBank database (Nakamura et al. 2000; www.kazusa.or.jp/codon) to compare synonymous codon usage among 97 coding sequences from *B. rapa*, 29,765 coding sequences from *A. thaliana*, and 666 coding sequences from *Lycopersicon esculentum* (the number of sequences represents the number available during October 2000). *Lycopersicon esculentum* is more distantly related to *B. rapa* and *A. thaliana* than either of these species are to one another and thus can be used as an outgroup to infer whether a shift in codon usage has occurred along the evolutionary lineage leading to *B. rapa*, the lineage leading to *A. thaliana*, or both. This analysis revealed that the most commonly used codons in *A. thaliana* and *L. esculentum* ended in A or T for all amino acids but lysine (Table 1). In contrast, in *B. rapa*

eight of the most commonly used codons ended in C or G, including six of nine codons that are twofold degenerate. Moreover, for all amino acids, synonymous codons that ended in C were used more frequently in *B. rapa* than in *A. thaliana* or *L. esculentum* and those ending in G were used more frequently in *B. rapa* for all amino acids but leucine (Table 1). Using parsimony criteria for ancestral state reconstruction, the differences between codon usage in *A. thaliana* and that in *B. rapa*, as detected in the Codon Usage Tabulated from GenBank database (Nakamura et al. 2000), appear to have evolved along the lineage leading to *B. rapa*.

The shift in codon usage in the *B. rapa* lineage is consistent with the effects of both weak selection acting on synonymous sites and mutational bias. Evidence for weak selection comes from the fact that the profile of most commonly used codons more closely matches the profile of optimal codons in *B. rapa* than in *A. thaliana* [assuming that the optimal codons that have been identified in *A. thaliana* (Miyashita et al. 1998) are also optimal in *B. rapa*]. Although there is no standard definition of “optimal” codons (sometimes called “favored” or “preferred” codons) for eukaryotic organisms, optimal codons are often identified as those most commonly used in a small subset of the most highly biased genes (Akashi 1995; Miyashita et al. 1998). Because highly expressed genes and highly conserved genes show the greatest use of optimal codons, these codons are presumably the ones that are favored by weak selection for translational efficiency and translational accuracy (Sharp and Li 1987). In *A. thaliana*, the most common codons always end in A or T, while the optimal codons (defined as the most over-represented codons in a sample of 268 high-biased genes) often end in G or C (Miyashita et al. 1998). Only 5 of 17 codons (one amino acid has no defined optimal codon) are both the most common and optimal in *A. thaliana* (Table 1). In contrast, 13 of the 17 defined optimal codons for *A. thaliana* are the most commonly used codons in *B. rapa*. Although in some unicellular organisms (Ikemura 1985) and *Drosophila* (Akashi 1995), the optimal codons are identical to the most commonly used codons (unpublished data), a lack of correspondence between the most common and optimal codons may not be unusual. For example, in *C. elegans*, only four optimal codons are also the most commonly used (Stenico et al. 1994; unpublished data).

More effective selection in *B. rapa* is consistent with expectations based on these plants’ mating systems. *Brassica rapa* reproduces largely through outcrossing, while *A. thaliana* reproduces through selfing. Selfing is expected to reduce recombination rates and weak selection acting on codon usage should be less effective with low recombination (Hill and Robertson 1966; Kliman and Hey 1993). If selection is more effectively shaping codon usage in *B. rapa*, then it is somewhat surprising that codon bias is stronger in *A. thaliana*. However, be-

Table 1. Synonymous codon usage in *Arabidopsis thaliana*, *Brassica rapa*, and *Lycopersicon esculentum*^a

Amino acid	Codon	<i>Arabidopsis thaliana</i>	<i>Brassica rapa</i>	<i>Lycopersicon esculentum</i>
Gly	GGG	0.16 ^a	0.18	0.14
	GGA	0.37^b	0.32	0.36
	GGT	0.34 ^{*,c}	0.33	0.35
	GGC	0.14	0.17	0.14
Glu	GAG	0.48 [*]	0.58	0.44
	GAA	0.52	0.42	0.56
Asp	GAT	0.68	0.60	0.71
	GAC	0.32 [*]	0.40	0.29
Val	GTG	0.26	0.29	0.25
	GTA	0.15	0.10	0.16
	GTT	0.40	0.34	0.43
	GTC	0.19 [*]	0.28	0.16
Ala	GCG	0.14	0.15	0.07
	GCA	0.27	0.24	0.32
	GCT	0.43[*]	0.41	0.46
	GCC	0.16	0.19	0.15
Arg	AGG	0.20	0.23	0.25
	AGA	0.35	0.34	0.35
	CGG	0.09	0.11	0.06
	CGA	0.12	0.11	0.11
	CGT	0.16 [*]	0.14	0.16
	CGC	0.07	0.07	0.07
Ser	AGT	0.16	0.16	0.18
	AGC	0.13	0.15	0.12
	TCG	0.10	0.12	0.07
	TCA	0.21	0.18	0.26
	TCT	0.28[*]	0.23	0.26
	TCC	0.12	0.16	0.12
Lys	AAG	0.51[*]	0.57	0.50
	AAA	0.49	0.43	0.50
Asn	AAT	0.53	0.43	0.62
	AAC	0.47 [*]	0.57	0.38
Ile	ATA	0.24	0.22	0.23
	ATT	0.41[*]	0.35	0.51
	ATC	0.34 [*]	0.43	0.26
Thr	ACG	0.15	0.16	0.09
	ACA	0.31	0.29	0.35
	ACT	0.34	0.26	0.40
	ACC	0.20 [*]	0.29	0.16
Cys	TGT	0.60	0.53	0.61
	TGC	0.40	0.44	0.39
Tyr	TAT	0.53	0.40	0.58
	TAC	0.47 [*]	0.60	0.42
Leu	TTG	0.22	0.21	0.26
	TTA	0.14	0.10	0.15
	CTG	0.11	0.10	0.10
	CTA	0.11	0.09	0.11
	CTT	0.26[*]	0.27	0.27
	CTC	0.17	0.23	0.12
Phe	TTT	0.53	0.42	0.59
	TTC	0.47 [*]	0.58	0.41
Gln	CAG	0.43 [*]	0.48	0.39
	CAA	0.57	0.52	0.61

Table 1. Synonymous codon usage in *Arabidopsis thaliana*, *Brassica rapa*, and *Lycopersicon esculentum*^a

Amino acid	Codon	<i>Arabidopsis thaliana</i>	<i>Brassica rapa</i>	<i>Lycopersicon esculentum</i>
His	CAT	0.62	0.50	0.66
	CAC	0.38 [*]	0.50	0.34
Pro	CCG	0.17	0.21	0.09
	CCA	0.34	0.32	0.41
	CCT	0.38[*]	0.33	0.39
	CCC	0.11	0.14	0.11

^a Frequencies are based on all sequences present in GenBank in October 2000.

^b For each amino acid, the most frequently used codon is in *boldface* (due to rounding error, the reported frequencies may be equal).

^c Asterisks designate those codons that exhibit the greatest increase in frequency of use with increased codon bias in *A. thaliana* (Miyashita et al. 1998).

cause we estimated codon bias using the ENC, a measure of codon usage that does not take into account optimal codons [e.g., F_{opt} , the frequency of optimal codons (Ikemura 1985)], the difference in codon bias may be strongly influenced by mutation bias. In particular, the greater codon bias we detected in *A. thaliana* may reflect mutational bias toward suboptimal codons ending in A or T rather than weak selection acting in *B. rapa*.

To investigate the possibility that mutational bias contributes to the difference in codon usage, we examined GC content in the intron and untranslated regions of 15 homologous genes from *A. thaliana* and *B. rapa*. Because these regions are not expected to experience selection for nucleic acid use, differences in GC content should reflect differences in mutational bias. In 12 of these genes, the GC content in noncoding regions is higher in *B. rapa* than *A. thaliana* and a binomial test rejects the null hypothesis that noncoding regions in *B. rapa* have a GC content equal to that in *A. thaliana* ($p < 0.01$). These results suggest that a change in mutation bias is at least partially responsible for the shift in codon usage that has occurred since the divergence of these two species.

Conclusion

To gain insight into the forces responsible for differences in DNA coding sequences in closely related plant species, we examined the divergence and codon bias of 218 homologous sequences from *Arabidopsis thaliana* and *Brassica rapa* ssp. *pekinensis*. Although there was wide variation in the rate at which these sequences have evolved, we detected no strong evidence that positive selection has contributed to divergence of the genes we examined. In contrast, it is possible that weak selection in *B. rapa* has caused codon use to evolve to correspond more closely to the profile of optimal codons. Although

the shift in codon use is consistent with the expectation that weak selection is more effective in the outcrossing *B. rapa* than the predominantly selfing *A. thaliana*, the shift appears to have been at least partially the result of an interspecific difference in the direction or strength of mutational bias. Shifts in synonymous codon usage also have been detected between species in the *Drosophila saltans* species group (Rodriguez-Trelles et al. 1999), suggesting that shifts in the profile of the most commonly used codons may not be uncommon. Moreover, if the shift in synonymous codon usage has been caused by changes in mutation pressure, then our results suggest that mutational bias may be an important factor in causing intergenome sequence divergence.

Acknowledgments. We thank Mark D. Rausher, Liquing Zhang, and Brandon S. Gaut for discussion and comments that improved this work. The idea to compare ESTs to the *A. thaliana* genome data was suggested by A.D. Long. P.T. was supported by a National Research Initiative Competitive Grants Program/United States Department of Agriculture award (99-35301-8076); M.W.H. was supported by a National Institutes of Health training grant administered through the Duke University Program in Genetics.

References

- Akashi H (1994) Synonymous codon usage in *Drosophila melanogaster*: Natural selection and translational accuracy. *Genetics* 163:927–935
- Akashi H (1995) Inferring weak selection from patterns of polymorphism and divergence at “silent” sites in *Drosophila* DNA. *Genetics* 139:1067–1076
- Akashi H (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
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410
- Alvarez-Valin F, Jabbari K, Carels N, Bernardi G (1999) Synonymous and nonsynonymous substitutions in genes from *Gramineae*: Intra-genic correlations. *J Mol Evol* 49:330–342
- Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796–815
- Barrier M, Robichaux RH, Purugganan MD (2001) Accelerated regulatory gene evolution in an adaptive radiation. *Proc Natl Acad Sci USA* 98:10208–10213
- Charlesworth D, Charlesworth B, McVean GAT (2001) Genome sequences and evolutionary biology, a two-way interaction. *Trends Ecol Evol* 16:235–242
- Chiapello H, Lisacek F, Caboche M, Henaut A (1998) Codon usage and gene function are related in sequences of *Arabidopsis thaliana*. *Gene* 209:GC1–GC38
- Comeron JM, Aguade M (1998) An evaluation of measures of synonymous codon usage bias. *J Mol Evol* 47:268–274
- Davidson EH (2001) *Genomic regulatory systems: Development and evolution*. Academic Press, San Diego
- Doebley J, Lukens L (1998) Transcriptional regulators and the evolution of plant form. *Plant Cell* 10:1075–1082
- Endo T, Ikeo K, Gojobori T (1996) Large-scale search for genes on which positive selection may operate. *Mol Biol Evol* 13:685–690
- Eyre-Walker AC (1991) An analysis of codon usage in mammals: selection or mutation bias? *J Mol Evol* 33:442–449
- Fennoy SL, Bailey-Serres J (1993) Synonymous codon usage in *Zea mays* L. nuclear genes is varied by levels of C and G-ending codons. *Nucleic Acids Res* 21:5294–5300
- Garcia-Olmedo F, Salcedo G, Sanchez-Monge R, Gomez L, Royo J, Carbonero P (1987) Plant proteinaceous inhibitors of proteinases and α -amylases. *Oxford Surv Plant Mol Cell Biol* 4:275–334
- Hall TA (1999) BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp Ser* 41:95–98
- Hill WG, Robertson A (1966) The effect of linkage on limits to artificial selection. *Genet Res* 8:269–294
- Hughes AL (1999) *Adaptive evolution of genes and genomes*. Oxford University Press, New York
- Ikemura T (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J Mol Biol* 151:389–409
- Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol* 2:13–34
- Kliman RM, Hey J (1993) Reduced natural selection associated with low recombination in *Drosophila melanogaster*. *Mol Biol Evol* 10:1239–1258
- Lagercrantz U (1998) Comparative mapping between *Arabidopsis thaliana* and *Brassica nigra* indicates that Brassica genomes have evolved through extensive genome replication accompanied by chromosome fusions and frequent rearrangements. *Genetics* 150:1217–1228
- Li W-H (1997) *Molecular evolution*. Sinauer Associates, Sunderland, MA.
- Liberles DA, Schreiber DR, Govindarajan S, Chamberlin SG, Benner SA (2001) The adaptive evolution database (TAED). *Genome Biol* 2:0028.1–0028.6
- Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290:1151–1155
- Makalowski W, Boguski MS (1998) Evolutionary parameters of the transcribed mammalian genome: An analysis of 2,820 orthologous rodent and human sequences. *Proc Natl Acad Sci USA* 95:9407–9412
- Miyashita NT, Kawabe A, Innan H, Terauchi R (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
- Moriyama EN, Powell JR (1997) Codon usage bias and tRNA abundance in *Drosophila*. *J Mol Evol* 45:514–523
- Mouchiroud D, Gautier C, Bernardi G (1995) Frequencies of synonymous substitutions in mammals are gene-specific and correlated with frequencies of nonsynonymous substitutions. *J Mol Evol* 40:107–113
- Nakamura Y, Gojobori T, Ikemura T (2000) Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res* 28:292
- Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* 3:418–426
- Ohta T, Ina Y (1995) Variation in synonymous substitution rates among mammalian genes and the correlation between synonymous and nonsynonymous divergences. *J Mol Evol* 41:717–720
- Rausher MD, Miller RE, Tiffin P (1999) Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. *Mol Biol Evol* 16:266–274
- Rodriguez-Trelles F, Tarrío R, Ayala FJ (1999) Switch in codon bias and increased rates of amino acid substitution in the *Drosophila saltans* species group. *Genetics* 153:339–350
- Rozas J, Rozas R (1999) DnaSP version 3: An integrated program for

- molecular population genetics and molecular evolution analysis. *Bioinformatics* 15:174–175
- SAS Institute Inc. (1989) SAS/STAT user's guide. Version 6, 4th ed. SAS Institute, Cary, NC
- Sharp PM, Li WH (1987) The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Mol Biol Evol* 43:222–230
- Sharp PM, Li W-H (1989) On the rate of DNA sequence evolution in *Drosophila*. *J Mol Evol* 28:398–402
- Shields DC, Sharp PM, Higgins DG, Wright F (1988) 'Silent' sites in *Drosophila* genes are not neutral: Evidence of selection among synonymous codons. *Mol Biol Evol* 5:704–716
- Stenico M, Lloyd AT, Sharp PM (1994) Codon usage in *Caenorhabditis elegans*: Delineation of translational selection and mutational biases. *Nucleic Acids Res* 22:2437–2446
- Sueoka N (1988) Directional mutation pressure and neutral molecular evolution. *Proc Natl Acad Sci USA* 85:2653–2657
- Wheelan SJ, Boguski MS, Duret L, Makalowski W (1999) Human and nematode orthologs: Lessons from the analysis of 1800 human genes and the proteome of *Caenorhabditis elegans*. *Gene* 238:163–170
- Wolfe KH, Sharp PM (1993) Mammalian gene evolution: Nucleotide sequence divergence between mouse and rat. *J Mol Evol* 37:441–456
- Wright F (1990) The effective number of codons used in a gene. *Gene* 87:23–29
- Yang ZH, Bielawski JP (2000) Statistical methods for detecting molecular evolution. *Trends Ecol Evol* 15:496–503
- Zar JH (1996) *Biostatistical analysis*, 3rd ed. Prentice Hall, Englewood Cliffs, NJ