Patterns of Nucleotide Substitution Among Simultaneously Duplicated Gene Pairs in *Arabidopsis thaliana*

Liqing Zhang,* Todd J. Vision,† and Brandon S. Gaut*

*Department of Ecology and Evolutionary Biology, University of California; and †Department of Biology, University of North Carolina at Chapel Hill

We characterized rates and patterns of synonymous and nonsynonymous substitution in 242 duplicated gene pairs on chromosomes 2 and 4 of *Arabidopsis thaliana*. Based on their collinear order along the two chromosomes, the gene pairs were likely duplicated contemporaneously, and therefore comparison of genetic distances among gene pairs provides insights into the distribution of nucleotide substitution rates among plant nuclear genes. Rates of synonymous substitution varied 13.8-fold among the duplicated gene pairs, but 90% of gene pairs differed by less than 2.6-fold. Average nonsynonymous rates were fivefold lower than average synonymous rates; this rate difference is lower than that of previously studied nonplant lineages. The coefficient of variation of rates among genes was 0.65 for nonsynonymous rates and 0.44 for synonymous rates, indicating that synonymous and nonsynonymous rates vary among genes to roughly the same extent. The causes underlying rate variation were explored. Our analyses tentatively suggest an effect of physical location on synonymous substitution rates but no similar effect on nonsynonymous rates. Nonsynonymous substitution rates were negatively correlated with GC content at synonymous third codon positions, and synonymous substitution rates were negatively correlated with codon bias, as observed in other systems. Finally, the 242 gene pairs permitted investigation of the processes underlying divergence between paralogs. We found no evidence of positive selection, little evidence that paralogs evolve at different rates, and no evidence of differential codon usage or third position GC content.

Introduction

Synonymous and nonsynonymous substitution rates vary among genes, but the evolutionary forces underlying rate variation remain unclear. For nonsynonymous substitutions, substitution rates vary in part because proteins differ in their degree of selective constraint. For example, histone genes probably evolve slowly because most amino acid changes disrupt gene function and are therefore selectively deleterious (Nei 1987, pp. 47–52). Other factors also contribute to nonsynonymous rate variation, including the physical location of the gene in the genome (Wolfe, Sharp, and Li 1989a; Williams and Hurst 2000) and mutation biases (Bird 1980; Tsunoyama, Bellgard, and Gojobori 2001).

It is even more difficult to identify causes of synonymous rate variation among genes, but at least three molecular characteristics correlate with synonymous substitution rates. The first characteristic is GC content (Ticher and Graur 1989), which affects synonymous substitution rates because CpG dinucleotides are subject to high rates of mutation (Tsunoyama, Bellgard, and Gojobori 2001). The second characteristic is codon bias. Highly biased genes evolve more slowly, on average (Sharp and Li 1987; Akashi 1994a, 2001; Eyre-Walker and Bulmer 1995), presumably because selection for codon use limits the number of acceptable synonymous nucleotide substitutions in highly biased genes. Finally, mutation rates can differ among genomic regions, contributing to variation in synonymous substitution rates among genes (Wolfe, Sharp, and Li 1989a; Matassi, Sharp, and Gautier 1999). However, none of these three factors alone consistently explains variation in synonymous substitution rates (Ticher and Graur 1989; Wolfe, Sharp, and Li 1989a; Akashi 1994b, 1997; Matassi, Sharp, and Gautier 1999), and the interdependence between these factors is not always clear.

An important prerequisite for understanding causes of rate variation is to characterize the distribution of rates among genes. To date, most studies of the distribution of synonymous and nonsynonymous substitution rates have been restricted to animals. In plants, several studies have examined rate variation across evolutionary lineages of mitochondrial and chloroplast genes (e.g., Gaut et al. 1992; dePamphilis, Young, and Wolfe 1997; Laroche, Maggia, and Bousquet 1997), but few studies have either characterized rate variation among plant nuclear genes or discerned the factors contributing to rate variation (Wolfe, Sharp, and Li 1989b; Alvarez-Valin et al. 1999). One possible reason for the dearth of plant studies is that most plant nuclear genes are members of multigene families. Copy number within multigene families fluctuates (Clegg, Cummings, and Durbin 1997), and as a result it is often difficult to identify orthologs between species and hence to compare substitution rates among genes.

The recently sequenced *Arabidopsis thaliana* genome offers a unique opportunity to study substitution rate variation across plant nuclear genes. Many large arabidopsis chromosomal segments are duplicated (Mayer et al. 1999; AGI 2000), and these segments contain genes that were likely duplicated contemporaneously. Genetic distances can be compared among pairs of duplicated sequences (or gene pairs). For this special case in which the time of duplication is equivalent for all gene pairs, comparing genetic distances among gene pairs is equivalent to comparing nucleotide substitution rates among gene pairs.

Key words: nucleotide substitution rates, positive selection, codon usage, regional mutation.

Address for correspondence and reprints: Brandon S. Gaut, Department of Ecology and Evolutionary Biology, 321 Steinhaus Hall, University of California, Irvine, California 92697-2525.

E-mail: bgaut@uci.edu.


© 2002 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038
In addition to providing insight into rate variation among genes, sequence data from duplicated chromosomal regions also facilitate characterization of patterns of sequence divergence between paralogs. Divergence patterns between paralogs have been of interest since Ohno (1970) suggested that most adaptive changes occur after gene duplication (see also Ohta and Kimura 1973). At the molecular level, adaptive changes can be detected by measuring the ratio of nonsynonymous ($K_a$) to synonymous substitution ($K_s$). A $K_a/K_s$ ratio >1 is strong evidence for positive selection having acted during sequence divergence, and a $K_a/K_s$ < 1 is consistent with purifying selection, although it does not rule out the possibility that positive selection acted. Although the sensitivity of $K_a/K_s$ to detect positive selection can be low (Hughes 1999), the distribution of $K_a/K_s$ among genes can be useful for characterizing the relative strengths of evolutionary forces acting on individual genes (Charlesworth, Charlesworth, and McVean 2001).

Sequence divergence can have other measurable effects. For example, shifts in selective constraint after duplication can lead to changes in rates of nucleotide substitution in one or both of the duplicated sequences (Goodman, Moore, and Matsuda 1975; Li and Gojobori 1983; Gonzalez and Jordan 2000). Similarly, paralogs often diverge in synonymous codon usage (Gaut et al. 1983; Gonzalez and Jordan 2000). Similarly, paralogs often diverge in synonymous codon usage (Gaut et al. 1983; Zhang, Kosakovsky-Pond, and Gaut 2001) either as a function of changes in mutation biases or as a function of shifts in gene expression (Fennoy and Bailey-Serres 1993; Duret and Mouchiroud 1999).

The large number of duplicated genes in *A. thaliana* permits unprecedented examination of rate variation among plant nuclear genes and also facilitates study of patterns of evolutionary divergence between paralogs. Here we characterize patterns of molecular evolution in a duplicated region between *arabidopsis* chromosomes 2 and 4. With data from 242 gene pairs, we address the following questions: (1) What is the distribution of synonymous and nonsynonymous substitution rates among gene pairs? (2) Are synonymous and nonsynonymous substitution rates a function of physical location, GC content, or codon usage? (3) Is there evidence for positive selection acting during the divergence of paralogs? (4) Is sequence divergence between paralogs accompanied by divergence in codon usage or evolutionary rate?

**Materials and Methods**

**Data set**

A duplicated region of 5.6 Mb on chromosome 2 and 4.6 Mb on chromosome 4 was initially identified by Terryn et al. (1999) and subsequently extended by Lin et al. (1999). We decided to study the genes of this region for two reasons. First, this is one of the largest duplicated segments in the *A. thaliana* genome. Second, this segment duplicated ~100 MYA (Vision, Brown, and Tanksley 2000), whereas other segments are more ancient and therefore too diverged for reasonable estimates of evolutionary rates.

The chromosome 2–4 duplication was further partitioned into five blocks (numbered 45, 49, 52, 54, and 56 by Vision, Brown, and Tanksley 2000) that differ in their order and orientation relative to one another on the two chromosomes but show only minor within-block re-arrangement (fig. 1). A minimum of four inversions is necessary to convert the order and orientation of blocks on one chromosome into that of the other. It is thought that these blocks arose from a single duplication event, both because of their suggestive spatial arrangement and based on a genome-wide analysis of the distribution of amino acid distances between duplicated genes (Vision, Brown, and Tanksley 2000). It should be emphasized, however, that our analyses do not assume that all blocks were duplicated at the same time, but we do assume that gene pairs within a single block were duplicated contemporaneously when they were in collinear order.

Our initial task was to identify gene pairs for further study. In the data set compiled by Vision, Brown, and Tanksley (2000), the five blocks contained 1183 genes (or tandem arrays of gene families) on chromosome 2, 1168 genes on chromosome 4, and 326 genes duplicated between chromosomes. In 52 of these 326 gene pairs, at least one copy belonged to a tandem array.

We excluded these genes from further analyses because of uncertainty as to the relative timing of tandem and chromosomal duplication events. Of the remaining 274 gene pairs, three were excluded from subsequent analyses because they proved to be too diverged to allow reasonable evolutionary inference. In addition, we excluded 29 genes that were suspect either because they were not in the optimal collinear order of gene pairs between the two chromosome copies or because they were too distant from other gene pairs. Collinearity was assessed by computing the longest common subsequence of gene pairs (Gusfield 1997, pp. 287–293). The distance between gene pairs was measured by the number of intervening genes on each chromosome. There are two such distances for a terminal gene pair in a
block, whereas there are four for an internal gene pair. We eliminated gene pairs from analysis that were overly distant from neighboring genes by only allowing one (in the case of terminal gene pairs) or two (in the case of internal gene pairs) distances to exceed a value of twelve. This cutoff was chosen to coincide with two standard deviations (SDs) about the mean distance calculated using all gene pairs.

The accession numbers of the 484 sequences used in the analyses are available at bgbox.bio.uci.edu/data/lq2acn.html.

Alignment

Amino acid alignments were obtained with ClustalW (Thompson, Higgins, and Gibson 1994), using default parameters. Visual inspection revealed that the alignments around gaps were sometimes ambiguous, and we therefore analyzed two alignment data sets. In the first data set, the residues around gaps were included. In the second data set, five amino acids were removed on either side of each gap. The two sets of amino acid alignments were then back translated using the known coding sequences to produce DNA sequence alignments. Gap treatment did not qualitatively affect our results, and we therefore report only the results for the data set with the original alignments. The DNA sequence alignments are available at bgbox.bio.uci.edu/.

Sequence Analyses

We calculated the percent identity of DNA sequences and protein sequences for all gene pairs. The expected relationship between DNA and amino acid identity under strictly neutral evolution was obtained by simulation, using Evolver (Yang 1997). We simulated DNA sequences with distances ranging from 0.1 to 2.0 along distance intervals of 0.05 units, corresponding to the range of distances observed between duplicated genes in our data set. For each distance, 20 pairs of DNA sequences of 500 codons were simulated. For all simulated data, percent identity was calculated from DNA sequences and translated protein sequences.

$K_s$ and $K_a$ between duplicated sequences were estimated by the maximum likelihood (ML) method implemented in PAML (Yang 1997), using the codon model of Goldman and Yang (1994). We used a likelihood ratio (LR) comparison to test for a $K_a/K_s$ ratio different from 1.0. To do this, two models were applied to the data: model 0 constrains the $K_a/K_s$ ratio to 1.0, and model 1 estimates the $K_a/K_s$ ratio as a free parameter. The LR of model 0 and model 1 was compared with the $\chi^2$ distribution with one degree of freedom, as detailed by Yang (1998).

Relative rate tests were performed with HYPHY (http://peppercat.statgen.ncsu.edu/~hyphy/), for both protein sequences and DNA sequences. In order to apply the relative rate test, we obtained outgroup sequences for a large number of the gene pairs. Ku et al. (2000) proposed that the duplication between chromosomes 2 and 4 occurred after the split of the tomato and A. thaliana lineages. If this is true, then plants as divergent, or more divergent, from A. thaliana than tomato can be used as the source of outgroup sequences. The Rosidae are the largest phylogenetic grouping in the NCBI taxonomy that contains A. thaliana and excludes tomato. Accordingly, we searched for candidate outgroup sequences among all GenBank records derived from angiosperms, excluding the Rosidae. TBLASTX (Altschul et al. 1997) was used to find matches between the 242 genes and the DNA sequences in the database. We required that both arabidopsis genes showed a strong match (expected value $<1 \times 10^{-10}$) to at least one of the sequences in the database for which a predicted or experimentally determined protein translation was available. In cases where multiple GenBank records met these criteria, the outgroup was selected from among the following NCBI-derived taxonomic groupings, in descending order of preference (and increasing taxonomic breadth): Asteridae (which includes tomato), core eudicots, eudicotyledons, and Magnoliophyta.

Each relative rate test was based on one gene pair and its outgroup. For DNA sequences, synonymous and nonsynonymous substitution rates were tested separately using the codon substitution model of Muse and Gaut (1994). For amino acid sequences, we used the substitution model of Dayhoff, Schwartz, and Orcutt (1978). The relative rate test uses an LR statistic to test the null hypothesis that two sequences evolve at equal rates (Muse and Weir 1992).

The effective number of codons (ENC) and GC content at synonymous third codon positions ($GC_s$) were calculated using codon W (http://www.molbiol.ox.ac.uk/cuf/). We used the ENC as a measure of codon usage because it is not biased by gene length, given a certain minimum length, or by amino acid composition (Wright 1990; Comeron and Aguade 1998). ENC values range from 20 to 61; a value of 20 represents extreme bias, and a value of 61 indicates random codon use. We tested for homogeneity of ENC and $GC_s$ between duplicated sequences by a permutation procedure described previously (Zhang, Kosakovsky-Pond, and Gaut 2001b), with the slight modification that the test statistic was the difference in ENC ($GC_s$) between sequences rather than the variance in ENC ($GC_s$) among sequences. Test statistics were based on 1000 permutations.

Spatial correlation of $K_a$ and $K_s$

To examine the relationship between evolutionary distance and physical position, we performed a spatial correlation analysis. The following statistic (Chatfield 1999, p. 20) was calculated for a range of distances between gene pairs, where distance was defined as the number of intervening genes between gene pairs:

$$C_d = \frac{\sum_{i=1}^{n-d} (x_i - \bar{x})(x_{i+d} - \bar{x})}{ns_x^2}$$

where $x_i$ is the value of $K_a$ (or $K_s$) for the gene pair at position $i$, $x_{i+d}$ is the value of $K_a$ (or $K_s$) for the gene pair separated by $d = 1$ intervening genes, $\bar{x}$ and $s_x^2$ are
Table 1

$K_a$, $K_s$ and $K_a/K_s$ Statistics

<table>
<thead>
<tr>
<th>Block</th>
<th>Genes</th>
<th>Mean $K_a$ (SD)</th>
<th>CV</th>
<th>$K_a$ Range</th>
<th>Mean $K_s$ (SD)</th>
<th>CV</th>
<th>$K_s$ Range</th>
<th>Mean $K_a/K_s$ (SD)</th>
<th>$K_a/K_s$ Range</th>
<th>$\tau$ for $K_a$ and $K_s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>45 . . .</td>
<td>76</td>
<td>0.17 (0.12)</td>
<td>0.71</td>
<td>0.0005–0.78</td>
<td>0.90 (0.54)</td>
<td>0.60</td>
<td>0.42–4.27</td>
<td>0.19 (0.10)</td>
<td>0.001–0.51</td>
<td>0.60***</td>
</tr>
<tr>
<td>49 . . .</td>
<td>50</td>
<td>0.17 (0.13)</td>
<td>0.76</td>
<td>0.0006–0.59</td>
<td>0.75 (0.26)</td>
<td>0.35</td>
<td>0.41–1.79</td>
<td>0.21 (0.14)</td>
<td>0.001–0.68</td>
<td>0.77***</td>
</tr>
<tr>
<td>52 . . .</td>
<td>37</td>
<td>0.14 (0.09)</td>
<td>0.64</td>
<td>0.01–0.33</td>
<td>0.77 (0.25)</td>
<td>0.32</td>
<td>0.31–1.62</td>
<td>0.18 (0.11)</td>
<td>0.04–0.56</td>
<td>0.74***</td>
</tr>
<tr>
<td>54 . . .</td>
<td>23</td>
<td>0.19 (0.09)</td>
<td>0.47</td>
<td>0.04–0.42</td>
<td>0.92 (0.22)</td>
<td>0.24</td>
<td>0.69–1.49</td>
<td>0.21 (0.08)</td>
<td>0.05–0.33</td>
<td>0.63***</td>
</tr>
<tr>
<td>56 . . .</td>
<td>56</td>
<td>0.17 (0.10)</td>
<td>0.59</td>
<td>0.0026–0.49</td>
<td>0.79 (0.21)</td>
<td>0.27</td>
<td>0.51–1.62</td>
<td>0.22 (0.12)</td>
<td>0.004–0.48</td>
<td>0.77***</td>
</tr>
<tr>
<td>Total . .</td>
<td>242</td>
<td>0.17 (0.11)</td>
<td>0.65</td>
<td>0.0005–0.78</td>
<td>0.82 (0.36)</td>
<td>0.44</td>
<td>0.31–4.27</td>
<td>0.20 (0.11)</td>
<td>0.001–0.68</td>
<td>0.70***</td>
</tr>
</tbody>
</table>

Note: SD = standard deviation; CV = coefficient of variation.

* Significance of Kendall’s correlation, $P < 0.001$.

Results

The Distribution of Substitution Rates Among Gene Pairs

We collected 242 gene pairs duplicated between chromosome 2 and 4 that were in compact collinear order and not duplicated in tandem. Their chromosomal locations are shown in figure 1. Among all gene pairs, DNA sequence identity ranged from 46.8% to 92.8%, and protein sequence identity ranged from 33.2% to 100%. Over the entire region, $K_a$ had a range of 0.0005–0.78 nonsynonymous substitutions per nonsynonymous site and a mean of 0.17. The range of $K_s$ was 0.31–4.27 synonymous substitutions per synonymous site, with a mean of 0.82 (table 1 and fig. 2). Despite the larger range of $K_s$, coefficients of variation (CV) of $K_a$ and $K_s$ were similar, with a 1.5-fold difference between them (table 1). Furthermore, $K_a$ and $K_s$ were positively and significantly correlated among genes within each block, just as they were correlated over the entire five-block region (Kendall’s rank correlation $\tau = 0.70$, $P < 0.001$; table 1).

Gene pairs within individual blocks exhibited a similar range of DNA and protein identities to the entire region, but there were apparent differences among blocks. For example, block 45 had a relatively high synonymous CV, and blocks 52 and 54 had a relatively narrow range of $K_a$ and $K_s$ (table 1). These observations could reflect sampling phenomena because some blocks contained small numbers of duplicated genes. To test for among-block rate variation, we performed an ANOVA. The ANOVA indicated that $K_a$ and $K_s$ are homogeneous.
among blocks ($K_s$: $P = 0.45$, $K_a$: $P = 0.09$), further suggesting that the blocks may have been duplicated contemporaneously.

Over the whole region, 58% (141/242) of gene pairs had higher DNA than protein identity, and DNA sequence identity was always higher than protein sequence identity when amino acid identity was greater than ~80% (fig. 3). To determine the relationship between DNA and protein identities relative to a model of strictly neutral evolution, we simulated pairs of gene sequences under neutrality. For simulations, $K_a/K_s$ was set to 1.0; κ, the transition and transversion parameter was set to 3.0; and the codon frequency matrix was based on tabulated frequencies from 14,647,315 Arabidopsis thaliana codons (http://www.kazusa.or.jp/codon/). The simulations indicated that DNA sequence identity was always higher than protein sequence identity under the neutral model (fig. 3), and the result was qualitatively similar with different transition:transversion ratios (κ = 2, 3, or 4; data not shown). In our Arabidopsis data set, amino acid sequence identity was always higher than that for simulated gene pairs with equivalent DNA sequence identity (fig. 3). Thus, selective constraint appears to be slowing the pace of nonsynonymous nucleotide substitution for all the gene pairs in our data set.

Examining Factors That Contribute to Rate Variation Among Gene Pairs

Four factors are commonly cited as contributing to rate variation among genes: (1) physical location within the genome, (2) GC content, (3) codon usage bias, and (4) differing levels of selective constraint on amino acid substitutions. In this section, we examine the relationship between rate variation and the first three factors.

We tested the relationship between substitution rate and physical distance with two methods. First, we applied a spatial autocorrelation test to both $K_a$ and $K_s$. There were no significant results with $K_a$, but autocorrelation for $K_s$ was greater than expected at the 5% significance level for some distances—for example, genes separated by 10 genes are more highly correlated than expected at random (fig. 4). We should note, however, that $K_s$ autocorrelation statistics do not demonstrate significantly high autocorrelation in neighboring genes, as expected under the hypothesis that physical location is a contributing factor to variation in synonymous substitution rates among genes. Thus, the $K_s$ results are difficult to interpret but suggest only weakly that physical location contributes to synonymous rate variation among genes. Second, we applied the permutation method of Williams and Hurst (2000) to test the null hypothesis that rates are random with respect to physical location. This test also failed to reject the null hypothesis for $K_a$ ($P = 0.66$), but the test was borderline significant for $K_s$ ($P = 0.067$). Thus, both tests provide suggestive, but inconclusive, evidence that there could be a relationship between $K_s$ and physical distance.

GC content and codon bias have also been shown to correlate with evolutionary rate (Ticher and Graur 1989; Tsunoyama, Bellgard, and Gojobori 2001). We calculated ENC and GCₐ for each gene pair. Among all genes, ENC ranged from 35.4 to 61.0, and GCₐ varied from 0.27 to 0.68. We examined correlations among GCₐ, codon bias, and evolutionary rates with Kendall’s nonparametric test of correlation (table 2). The results are inconsistent among blocks, but two correlations stand out. The first is the correlation between $K_a$ and GCₐ, which is negative in four of five blocks and significantly negative over the entire region. The second correlation is between $K_s$ and ENC, which is positive in four of five blocks and significantly positive over the entire region.

The Distribution of $K_a/K_s$ Across Gene Pairs

To examine whether positive selection commonly contributed to sequence divergence between paralogs, we estimated $K_a/K_s$ for all gene pairs and tested whether $K_a/K_s$ exceeded 1.0. Estimated $K_a/K_s$ values ranged from
Table 2
Kendall’s Correlation (τ) Among GC_s, ENC and Substitution Rate, with Probability Value P

<table>
<thead>
<tr>
<th>BLOCK</th>
<th>K_a and GC_s</th>
<th>K_a and ENC</th>
<th>K_s and GC_s</th>
<th>K_s and ENC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>τ  P</td>
<td>τ  P</td>
<td>τ  P</td>
<td>τ  P</td>
</tr>
<tr>
<td>45 . . . .</td>
<td>0.015 0.85</td>
<td>0.018 0.82</td>
<td>0.213 0.006</td>
<td>0.217 0.006</td>
</tr>
<tr>
<td>49 . . . .</td>
<td>−0.285 0.003</td>
<td>0.014 0.89</td>
<td>−0.138 0.16</td>
<td>0.081 0.41</td>
</tr>
<tr>
<td>52 . . . .</td>
<td>−0.224 0.05</td>
<td>0.006 0.97</td>
<td>−0.056 0.63</td>
<td>−0.030 0.81</td>
</tr>
<tr>
<td>54 . . . .</td>
<td>−0.249 0.10</td>
<td>−0.194 0.21</td>
<td>−0.032 0.83</td>
<td>0.190 0.20</td>
</tr>
<tr>
<td>56 . . . .</td>
<td>−0.046 0.62</td>
<td>−0.007 0.93</td>
<td>0.118 0.20</td>
<td>0.157 0.09</td>
</tr>
<tr>
<td>Total . . .</td>
<td>−0.111 0.01</td>
<td>−0.001 0.98</td>
<td>0.058 0.18</td>
<td>0.134 0.002</td>
</tr>
</tbody>
</table>

Ten of the 14 gene pairs that were significant for protein sequences were also significant for K_a-based tests, showing that results were reasonably consistent between methods. For K_s, six of 105 gene pairs were significant at P < 0.05, but only one remained significant after Bonferroni correction.

To further characterize patterns of sequence divergence between duplicated genes, we measured both GC_s and ENC for all gene pairs. Table 3 provides average ENC and GC_s for both chromosomes and all blocks. We applied paired t-tests to determine whether the two chromosomes differ significantly for either measure; no significant difference was detected between chromosomes for either ENC (t = 0.478, P = 0.633) or GC_s (t = −1.591, P = 0.112). This result did not preclude the possibility that any two paralogous sequences had diverged significantly in codon usage. To test this possibility, we applied a permutation test of homogeneous ENC (or GC_s) to each of the 242 gene pairs (Zhang, Kosakovsky-Pond, and Gaut 2001b). No paralogs exhibited significant difference in either measure after Bonferroni correction (data not shown), and thus there has been no detectable divergence in codon usage or GC content after duplication.

Discussion

In this study, we examined sequence divergence among gene pairs located in the duplicated region between chromosomes 2 and 4 of A. thaliana. Gene pairs were studied under the assumption that they were duplicated contemporaneously. We took three precautions to ensure the accuracy of this assumption. First, the chromosome 2–4 duplication consists of five rearranged blocks that presumably owe their duplication to the same event, but we studied each block separately as well as in combination. Second, we did not consider arrayed genes in tandem because it is difficult to determine the relative timing of tandem and chromosomal duplication events. Finally, we studied only genes that occur in compact collinear arrangement between the two chromosomes; collinearity reinforces the notion that the gene pairs originated during a single duplication event.

Our study differs from previous studies of rate variation among genes in that we used paralogous, rather than orthologous, sequences to estimate substitution rates. One potential complication is that gene conversion can occur between paralogs. If gene conversion has oc-
curred in some gene pairs but not others, the net effect is to increase the range and variance of $K_a$ and $K_s$ among gene pairs. We tested for the presence of gene conversion in all gene pairs with Sawyer’s (1989) method but found no evidence for conversion between paralogs (data not shown). We should note, however, that most tests for gene conversion detect events that affect a portion of a gene rather than a complete gene.

Despite the tendency of gene conversion to inflate the range of substitution rates, we found $K_a$ ranges similar to previous studies. For example, $K_a$ ranged 15-fold in a study of 24 Drosophila orthologs (Zeng et al. 1998) and 20-fold in a study of 363 rat-mouse orthologs (Wolfe and Sharp 1993). In our study, synonymous substitution rates varied up to 10.4-fold within an individual block and 13.8-fold over the entire data set (table 1 and fig. 2).

The distribution of $K_a$ and $K_s$ from arabidopsis gene pairs differs from previous studies of non-plant taxa in two noteworthy ways. First, the ratio of mean synonymous to mean nonsynonymous substitution rates is relatively low. The ratio is ~5 for our arabidopsis data (table 1), but it is as high as 24 for bacterial genes (Sharp 1991) and ~7 for rat-mouse comparisons (Wolfe and Sharp 1993), based on comparisons between orthologous sequences. Differences in this ratio may reflect different population sizes and life histories (Sharp 1991), but may also reflect differences between ortholog and paralog comparisons. For example, Kondrashov et al. (2002) recently found that the ratio of synonymous to nonsynonymous substitution is ~twofold lower between paralogs than orthologs. Thus, the forces affecting this ratio remain unclear at present. Second, the ratio of CVs between nonsynonymous and synonymous rates is lower for this study than previous studies. In rat-mouse comparisons, for example, the CV for nonsynonymous rates was fourfold higher than the CV for synonymous rates (Wolfe and Sharp 1993). Among arabidopsis gene pairs, there is only a 1.5-fold difference between nonsynonymous and synonymous CVs (table 1), indicating that synonymous and nonsynonymous rates vary among genes to roughly the same extent.

Our analyses represent the most extensive comparison of rates among plant nuclear genes to date. Some previous studies have been based on a much smaller number of sequence comparisons. For example, Wolfe, Sharp, and Li (1989b) and Gaut (1998) found that synonymous rates vary up to 2.5-fold among genes, but these studies were based on only 11 and nine nuclear genes, respectively. A more recent study made 212 orthologous sequence comparisons between arabidopsis and Brassica rapis (Tiffin and Hahn 2002), but relied on EST data, which could bias results depending on the accuracy of EST data and the gene regions sequenced. Nonetheless, Tiffin and Hahn (2002) described two rate characteristics that are similar to our results: (1) an ~24-fold range of synonymous rate variation among genes and (2) a twofold difference in CV between nonsynonymous and synonymous rates. Together these two plant studies indicate that synonymous rates vary at least an order of magnitude among plant nuclear genes and also that variation in nonsynonymous and synonymous rates are relatively similar among genes, based on the CV. It is clear, however, that a general picture of plant nuclear gene evolution requires additional studies, particularly given the fact that different plant lineages evolve with different mutational patterns (Tiffin and Hahn 2002) and different rates (e.g., Eyre-Walker and Gaut 1997).

With the increasing availability of genome sequences, several studies have used genetic distances (either $K_a$ or $K_s$) to make genome-wide inferences about the timing and frequency of gene and genome duplication (Gaut and Doebley 1997; Lynch and Conery 2000; Vision, Brown, and Tanksley 2000; Friedman and Hughes 2001). For example, Lynch and Conery (2000) used $K_s$ values between duplicated arabidopsis sequenc- es as a proxy for divergence time; in essence they assumed that substitution rates were homogeneous among gene pairs. Their use of $K_s$ has come under criticism (Long and Thornton 2001; Zhang, Gaut, and Vision 2001a), but this study provides quantitative insights about the degree to which the assumption can be misleading. In a worst-case scenario, $K_s$ provides time estimates that differ up to 14-fold (table 1). However, the average effect of the homogeneous rate assumption is not nearly as dramatic because ~90% of the gene pairs in this study fall within a $K_s$ range of 0.462–1.188 (fig. 2), and thus most genes (~90%) vary less than 2.57-fold in rate. Overall, however, our study indicates that $K_s$ variation among genes is generally higher than previously reported for plants (Wolfe, Sharp, and Li 1989b; Gaut 1998), and thus one should be cautious when equating $K_s$ with time.

Factors Contributing to Rate Variation Among Gene Pairs

What are the forces that contribute to variation in evolutionary rates among gene pairs? The first possibility, discussed earlier, is gene conversion. Although we did not detect gene conversion in any of our gene pairs, the possibility of gene conversion should not be ignored. However, the fact that the range of variation found in our 242 gene pairs is similar to that of orthologous gene pairs from other studies (rat-mouse, drosophilids and Arabidopsis-Brassica; see earlier) suggests that the effects of gene conversion have been negligible. In addition, a similar study of yeast paralogs found no evidence of gene conversion, suggesting gene conversion may not be widespread (Pál, Papp, and Hurst 2001).

The origin of duplicated chromosomes can also affect the variance in rates across gene pairs. If chromosomal duplication originated via a polyploid event, the original gene pairs contained varied levels of residual standing variation (polymorphism) at the onset of duplication. Standing variation contributes to variation in genetic distances among gene pairs (discussed in Gaut and Doebley 1997). However, for highly diverged gene pairs like those we have studied here, the level of standing variation at the time of origin should be very low relative to the total divergence between paralogs. Hence the origin effect probably contributes little to the ob-
served variance in substitution rates among the 242 arabidopsis gene pairs.

Physical location is a third potential contributor to rate variation among genes. For example, Williams and Hurst (2000) found that nonsynonymous substitution rates between mouse and rat vary as a function of genome location. Similarly, Matassi, Sharp, and Gautier (1999) documented location effects on synonymous substitution rates in human-rodent comparisons. Both of these effects likely reflect variable mutation rates along chromosomes (Casas et al. 1997; Lercher, Williams, and Hurst 2001). We examined the relationship between physical location and nucleotide substitution rates with two methods, and neither analysis provided evidence that physical location affects nonsynonymous substitution rates. In contrast, there is a hint that physical location and synonymous substitution rates are correlated, but the effect, if present, is weak. It should be noted that previous studies focused on whole genomes, with neighboring genes up to 5 cM apart (Matassi, Sharp, and Gautier 1999; Williams and Hurst 2000; Lercher, Williams, and Hurst 2001), which is roughly 10 Mb in humans (Williams and Hurst 2000). In contrast, our entire study focused on a region of 5 Mb, with neighboring genes separated by only a few kb. If the physical scale that affects substitution rates is large, our analyses could miss location effects.

Variation in selective constraint among proteins also contributes to rate variation among gene pairs. Deviation from a molecular clock is one potential measure of variation in selective constraint between paralogs, but on the whole our analyses uncovered little deviation from clock-like evolution. With some caveats, \( K_s/K_a \) can also be used for characterizing the relative strength of evolutionary forces acting on individual gene pairs. For the 242 gene pairs in our study, \( K_s/K_a \) ranged from 0.0 to 0.70, with a mean of 0.20, suggesting extensive variation in selective constraint among gene pairs. However, \( K_s/K_a \) never exceeded 1.0 (table 1 and fig. 3), and there is thus no evidence that positive selection has driven divergence between any of the paralogs on the duplicated regions of chromosomes 2 and 4. Altogether, variation in selective constraint, but not positive selection, likely contributes to variation in substitution rates among arabidopsis gene pairs.

Both GC content and codon bias have been shown to correlate with nucleotide substitution rates (Moriyama and Gojobori 1992; Alvarez-Valin et al. 1999), and hence GC content and codon bias may also contribute to nucleotide substitution rate variation among gene pairs. We detected no correlations between GC content and \( K_a \) in arabidopsis gene pairs, but did find a negative correlation between GC content and \( K_s \). The phenomena underlying the negative correlation are unclear, but the lack of correlation between GC content and \( K_s \) is not surprising given the relatively homogeneous base composition of the \( A. thaliana \) genome (Barakat, Matassi, and Bernardi 1998; AGI 2000). We also found a correlation between codon bias and \( K_a \) that is similar to that documented in other species (Sharp and Li 1987; Moriyama and Hartl 1993, pp. 847–858; Akashi 1994b). This correlation is consistent with strong codon bias (low ENC) limiting acceptable synonymous changes and thus retarding substitution rates.

One interesting feature of codon bias is that it often also correlates with gene expression level—i.e., highly expressed genes are more biased (Gouy and Gautier 1982; Duret and Mouchiroud 1999; Coghlan and Wolfe 2000). Surprisingly, highly expressed genes also evolve with relatively low \( K_s \) (Duret and Mouchiroud 2000). Both of these observations suggest that yet another factor—gene expression—contributes to variation in \( K_s \) and perhaps \( K_s/K_a \) among genes. To briefly explore this possibility with the arabidopsis gene pairs, we employed the database procedures of Duret and Mouchiroud (2000) and counted the number of BLAST hits for each gene pair to an arabidopsis EST database (http://www.kazusa.or.jp/en/plant/arabi/EST/). We then regarded the number of hits as a rough approximation of the expression level of a gene pair (this approximation is particularly rough given the diverse origin and preparation methods of cDNA libraries used for EST sequencing), and compared this expression level with evolutionary rates. Based on pairwise correlations among \( K_s/K_a \) and the number of EST hits, there is a significantly negative partial correlation between expression and \( K_s/K_a \) \((r = -0.172, P = 0.007)\) but no significant correlation between expression and \( K_a/K_s \) \((r = 0.037, P = 0.57)\). More detailed analyses require better expression data, but this significant negative correlation is consistent with recent ideas that evolutionary rates may be either proximally or secondarily a function of gene expression (reviewed in Akashi 2001).

Acknowledgments

We thank A. McLysaght, M. Tenaillon, M. Mondragon, M. Clegg, and P. Tiffin for valuable comments, and S. Muse provided valuable advice. We also thank S. Kosakovsky-Pond and Z. Yang for assistance with Hyphy and PAML, respectively. This work was funded by NSF grants DBI-9872631 and DEB-0104796. T. J. V. was supported by the USDA-ARS Center for Agricultural Bioinformatics and the Cornell Theory Center.

LITERATURE CITED


Substitution Between Duplicated Genes in *Arabidopsis thaliana*


Brian Golding, reviewing editor

Accepted April 29, 2002