Factors that Contribute to Variation in Evolutionary Rate among *Arabidopsis* Genes

Liang Yang[1] and Brandon S. Gaut*,[1]

[1]Department of Ecology and Evolutionary Biology, University of California Irvine, Irvine

***Corresponding author**: E-mail: bgaut@uci.edu.

Associate editor: Aoife McLysaght

Research article

Abstract

Surprisingly, few studies have described evolutionary rate variation among plant nuclear genes, with little investigation of the causes of rate variation. Here, we describe evolutionary rates for 11,492 ortholog pairs between *Arabidopsis thaliana* and *A. lyrata* and investigate possible contributors to rate variation among these genes. Rates of evolution at synonymous sites vary along chromosomes, suggesting that mutation rates vary on genomic scales, perhaps as a function of recombination rate. Rates of evolution at nonsynonymous sites correlate most strongly with expression patterns, but they also vary as to whether a gene is duplicated and retained after a whole-genome duplication (WGD) event. WGD genes evolve more slowly, on average, than nonduplicated genes and non-WGD duplicates. We hypothesize that levels and patterns of expression are not only the major determinants that explain nonsynonymous rate variation among genes but also a critical determinant of gene retention after duplication.

Key words: nonsynonymous substitution, synonymous substitution, gene duplication, gene expression.

Introduction

Evolutionary rates vary among genes, but the underlying causes of rate variation remain obscure. To identify potential causes, researchers have investigated correlations between evolutionary rates and genomic parameters. For example, rates of both nonsynonymous and synonymous evolution correlate with gene characteristics like length (Marais and Duret 2001) and intron number (Larracuente et al. 2008), suggesting that rates are partly a property of gene organization. Rates also vary over chromosomal scales, suggesting that forces like recombination (Pál et al. 2001a) and mutation also contribute to rate variance among genes.

Historically, however, it has been thought that rate variation primarily reflects variation in functional importance (Zuckerkandl 1976). Many studies have tested this idea by assessing the correlation between evolutionary rates and measures of protein function, like "protein dispensability" (or overall importance) (Hirsh and Fraser 2001; Yang et al. 2003) and "protein stability" (Zeldovich et al. 2007; Lobkovsky et al. 2010). Although evolutionary rates do often correlate with these variables, rates seem to correlate most strongly with measures related to gene expression, (Pál et al. 2006) like codon bias (Sharp and Li 1987; Urrutia and Hurst 2001), expression level (Pál et al. 2001b; Subramanian and Kumar 2004; Drummond et al. 2006) and expression breadth (Duret and Mouchiroud 2000; Zhang and Li 2004). The reason for the correlation with gene expression is not entirely clear, but highly expressed genes may be under strong selective constraint for translation robustness (Drummond et al. 2005; Drummond and Wilke 2008) and broadly expressed genes may be constrained by the need to function in several biochemical environments (Duret and Mouchiroud 2000).

Recently, researchers have discovered another factor that correlates with evolutionary rate: duplication status. This correlation was discovered on the basis of ortholog comparisons, classifying ortholog pairs into those that have paralogs (i.e., duplicates) and into those that do not (singletons). This approach has revealed that slower evolution of duplicated genes is a common feature of eukaryotes (Yang et al. 2003; Davis and Petrov 2004; Jordan et al. 2004). Explanations for this observation include the ideas that conserved, functionally important genes are more likely to be retained as duplicates (Davis and Petrov 2004) and that retained duplicates are highly structurally constrained (Yang et al. 2003). In contrast to duplicates, singleton genes are more poorly annotated and may have less critical functions (Jordan et al. 2004).

Gene duplication is particularly common in plants due to the prevalence of whole-genome duplication (WGD) by polyploidy. WGD has occurred throughout the evolutionary history of flowering plants (Soltis PS and Soltis DE 2009), including a putative event at the base of the angiosperms (Jaillon et al. 2007; Edger and Pires 2009). In addition, many species, like *Arabidopsis* and maize, have experienced multiple WGD events over their evolutionary history (Vision et al. 2000; Gaut 2001). However, WGD is not the only mode of gene duplication. Genes may also be duplicated by segmental events that encompass large chromosomal regions, by dispersed duplication of single genes (Akhunov et al. 2003), and by tandem duplication. These alternate sources of gene duplication can be consequential; for example, in *Arabidopsis thaliana* tandem duplicates comprise almost as many genes (up to 18%) as those duplicated by WGD events (∼25%) (Lockton and Gaut 2005).

WGD and non-WGD genes differ in function. For example, genes retained as duplicates after WGD events are overrepresented for transcription factors and signal transducers (Blanc and Wolfe 2004; Maere et al. 2005). In contrast, tandemly duplicated genes, which represent one type of non-WGD duplicate, are biased for membranous proteins and genes involved in stress response (Rizzon et al. 2006). Given these functional differences, it seems plausible that evolutionary rates vary not only between singletons and duplicates but also between WGD and non-WGD duplicates.

To date, there have been no genome-wide studies of evolutionary rates among plant nuclear genes. As a result, there has been no accurate description of rate variation among genes, little investigation as to whether rates vary along chromosomes, and few attempts to correlate evolutionary rate with duplication status or other important evolutionary characteristics. The dearth of rate studies stems from a lack of sequenced closely related genomes that permit accurate ortholog identification (Gaut and Ross-Ibarra 2008). The *A. lyrata* genome sequence (Hu et al. 2011) removes this obstacle with respect to identifying orthologs in *A. thaliana*. *A. lyrata* and *A. thaliana* diverged ~13 million years ago (Ma) (Beilstein et al. 2010) and have ~80% sequence identity over whole-genome alignments (Hu et al. 2011). Moreover, the two genomes have shared WGD events, including one or two events near the base of the angiosperms (Simillion et al. 2002; Bowers et al. 2003) and a third event that occurred ~40 Ma, near the base of the Brassicaceae (Schranz and Mitchell-Olds 2006). As a result, the duplication status of individual genes should be well conserved between species.

In this study, we seek to characterize genome-wide patterns of rate variation among plant nuclear genes in the hope of inferring some of the evolutionary forces that shape this variation. We begin by estimating the number of synonymous substitutions per synonymous site (K_s), the number of nonsynonymous substitutions per nonsynonymous site (K_a) and their ratio K_a/K_s ($= \omega$) between *A. lyrata* and *A. thaliana* orthologs. We then use these rate estimates to address the following three questions. First, what is the distribution of K_s, K_a, and ω among genes and is this distribution clustered along the physical length of chromosomes? Second, what are the major correlates of evolutionary rates? Finally, do rates vary as a function of duplication status?

Materials and Methods

Orthologs, Duplicates, and Singletons

The orthologs and alignments for this study are the same as those used in Yang et al. (2011). K_s, K_a and ω were estimated in PAML (Yang 1997), using default parameters.

To identify duplicated and singleton genes within species, an all-against-all BlastP (Altschul et al. 1997) with default parameters was performed on all annotated protein sequences within *A. thaliana* and *A. lyrata*. Singleton genes were defined as those genes with no hit with an *e*-value $\leq$0.1 (Gu 2003). Duplicated genes were identified according to previous methods (Gu et al. 2002). Briefly, we first selected Blast alignments with *e*-values $\leq 10^{-10}$. Then two proteins were denoted as forming a link if 1) the alignable region length (L) was over 80% of the longer protein and 2) the identity (I) between them was $\geq$30% if the alignable region is longer than 150 amino acids or $I \geq 0.06 + 4.8L^{-0.32[1 + \exp(-L/1{,}000)]}$ (Rost 1999) if otherwise. Next, we submitted each protein as the query to search against the *A. thaliana* repetitive elements in Repbase (Jurka et al. 2005). Proteins were removed from further consideration if they formed a link due to their homology with the same repetitive element. Finally, a single-linkage algorithm was used to group proteins into families. If genes were not positively identified as either a "singleton" or as a "duplicate" according to the above definitions, then they were not assigned to groups based on duplication status.

For duplicated genes, we classified them as to whether or not they were derived from WGD events according to the assignments of Blanc et al. (2003). The Blanc et al. (2003) data set contained 1,372 and 2,584 gene pairs representing early and a more recent WGD event, respectively. Because some genes were found in both age classes, we restricted our data set to genes with only one age annotation. We also repeated all analyses using the WGD duplication definitions of Bowers et al. (2003). All the results were qualitatively identical with the Blanc et al. (2003) definitions, and so we report only the Blanc et al. (2003) results here.

Principal Component Regression Analysis

We utilized principal component regression (PCR) analysis to explore the potential contribution of evolutionary parameters to the total variance in evolutionary rate among genes (Jolliffe 2002; Drummond et al. 2006). The package "pls" from *R* language (Ihaka and Gentleman 1996) was used to perform PCR, with K_a and K_s as response variables and 14 possible determinants of protein evolutionary rates as predictor variables (see below). We log transformed the predictor variables if log transformation led to a higher correlation coefficient, added a constant (0.001) before log transformation if the variables included zero values, and scaled the predictors by dividing each variable by its sample standard deviation before the PCR.

Potential Determinants of Protein Evolutionary Rates

We incorporated 14 gene characteristics into our PCR model based on availability and precedence in the literature. For each *A. thaliana* gene with an ortholog in *A. lyrata*, we calculated the following.

Gene Structure

We calculated statistics such as gene length, GC content, 5′ and 3′ UTR length, intron number, average intron length, and the frequency of optimal codons (F_{op}) (Ikemura 1985), as estimated by CodonW (http://codonw.sourceforge.net/) with preferred codons in *A. thaliana* from Wright et al. (2004).

Local Recombination Rate

We obtained *A. thaliana* genetic markers and genetic map positions from Singer et al. (2006). Given these markers, local recombination rates were estimated by using MareyMap (Rezvoy et al. 2007) with *LOESS* interpolation. The *LOESS* procedure depends on two parameters: the polynomial degree and the span, which describes the number of points used to calculate the local polynomial around a marker. Here, we employed *LOESS* with second degree polynomial fitting and a span consisting of 25% of the total number of points.

Levels and Patterns of Gene Expression

The expression data were obtained from the Arabidopsis Development Atlas (ADA, available at ftp://ftp.arabidopsis.org/, ExpressionSet ME00319), which contains triplicate expression estimates for ~80% of known *Arabidopsis* genes across 79 different tissues and developmental time points, using the Affymetrix ATH1 chip (Schmid et al. 2005). In order to minimize the effects of cross-hybridization, we matched each Affymetrix probe to the genome annotation and excluded any probe that matched multiple genes. The mean value of each triplicate was calculated for each probe under each condition. We used ADA data to estimate gene expression level and tissue specificity. The expression level of a gene was estimated by the average value of all the 79 samples. The tissue specificity was measured with the index τ (Yanai et al. 2005):

$$\tau = \frac{\sum_{j=1}^{n}[1 - \log_2 S(i,j)/\log S_2(i,\max)]}{n - 1},$$

where $n = 79$ is the number of tissues and conditions, and $S(i,\max)$ is the highest expression of gene i across the n tissues and conditions. The index τ ranges from 0 to 1, with higher value indicating higher specificity (or, synonymously, higher variation in expression across libraries). If a gene is expressed in only one library, τ approaches 1. In contrast, if a gene is expressed equally in all libraries, $\tau = 0$. The advantage of using τ rather than expression breadth as a measure of specificity has been documented previously (Liao and Zhang 2006).

We also examined Massively Parallel Signature Sequencing (MPSS) expression data for *A. thaliana* (http://mpss.udel.edu/at/) (Meyers et al. 2004), using similar methods. The MPSS data yielded qualitatively similar results, and we thus focused on the ADA data throughout the manuscript.

Function

We assessed the multifunctionality of a gene by counting the number of biological processes in which a gene is involved (Salathe et al. 2006), according to Gene Ontology (GO) Slim annotations that classify proteins to gain a high-level view of the functions (Prachumwat and Li 2006). GO annotations were obtained from The Arabidopsis Information Resource (http://www.arabidopsis.org).

Promoter Divergence

Yang et al. (2011) measured divergence between the upstream sequences of each orthologous gene pair by the shared motif method (Castillo-Davis et al. 2004). For each orthologous gene pair, we obtained the divergence score d_{SM}, defined as the fraction of both sequences that does not contain a region of significant local similarity (Castillo-Davis et al. 2004). A d_{SM} value of 0 indicates complete sharing of motifs between sequences, whereas a d_{SM} value of 1 indicates an absence of shared motifs. Yang et al. (2011) analyzed sequences encompassing 500 bp upstream from the translation start site, but results were qualitatively similar with longer upstream sequences (data not shown).

Duplication Mode

For the purposes of the PCR, genes were given discrete values to reflect duplication class: "1" for early WGD duplicates, "2" for recent WGD duplicates, "3" for non-WGD duplicates, and "4" for singletons.

Chromosomal Position

To include information about chromosomal location, we scaled the distance of each gene from the centromere. On each chromosomal arm, values ranged from 0 to 1, with higher values indicating greater physical distance from the centromere.

Results

The Distribution of Rate Variation among Orthologs

We began with 19,119 orthologous pairs (Yang et al. 2011) and then culled the data. First, we retained only the orthologous pairs that were defined as duplicated in both species or deemed as singletons in both species. Second, we retained only duplicated genes that had a single unambiguous assignment with regard to early or recent WGD events. Finally, we discarded 191 genes at the extreme tail of the K_s distribution ($K_s > 0.3$), which could denote either misalignment or potential sequence saturation. Our final data set consisted of 11,492 orthologous pairs, including 9,995 duplicated genes and 1,497 singletons.

K_a, K_s, and ω were calculated for these remaining 11,492 orthologs; their frequency distributions are provided in figure 1. The K_s distribution had a mean of 0.147 (table 1) and ranged from complete sequence identity ($K_s = 0.0$) for two genes (At2g07669 and At2g07772) to $K_s = 0.3$. The coefficient of variation (CV) of K_s was 0.30. The K_a estimates had a lower mean, at 0.028, and ranged from $K_a = 0.0$ to a high value of 0.29 (table 1). On average, however, K_s and K_a differed ~5-fold as reflected in average ω estimates of 0.203 (table 1). The ω distribution also lacked a prominent tail of genes with values >1.0 that could be indicative of positive selection; only 0.7% (90 of 11,492) of orthologous pairs yielded ω estimates >1.0 (see supplementary table S1, Supplementary Material online for details). Overall, K_a and K_s values were highly positively correlated across genes (Spearman's rank correlation $\rho = 0.21$, $P < 10^{-16}$), suggesting the possibility that common evolutionary mechanisms affect both synonymous and nonsynonymous sites.

To investigate the distribution of K_s, K_a, and ω along *A. thaliana* chromosomes, we plotted mean values for nonoverlapping windows of 0.5 Mb, corresponding to an average of 48.9 genes in each window (fig. 2 for chromosome 1;

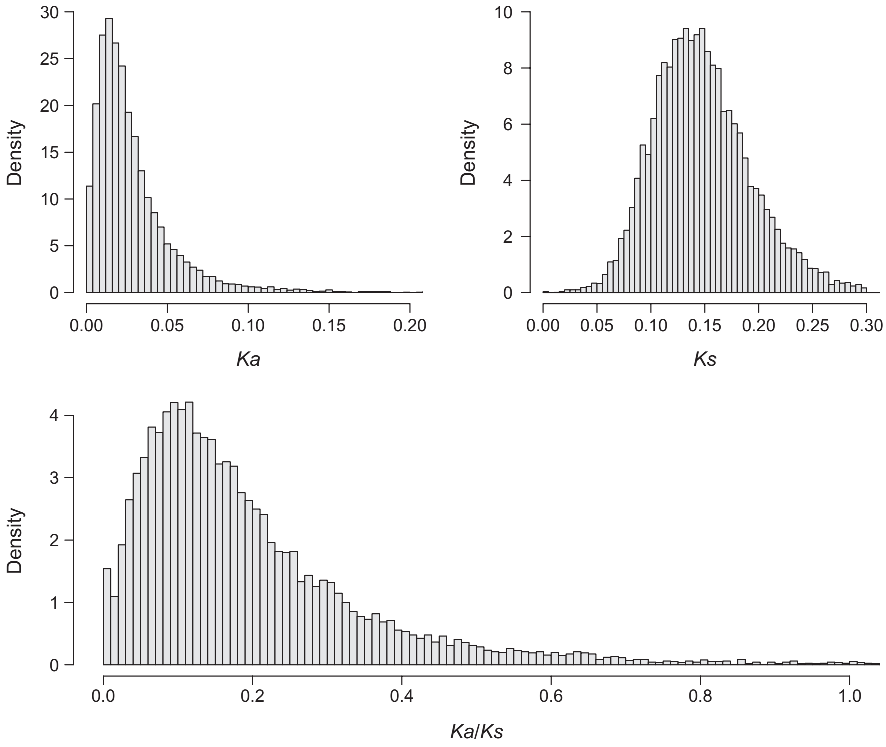

FIG. 1. The frequency distributions of K_a, K_s, and K_a/K_s (ω).

supplementary fig. S1–S4, Supplementary Material online, for chromosome 2–5, respectively). In general, there were few marked peaks for K_s (fig. 2). To test whether divergence values within windows were higher than expected, we randomly permuted K_s values among genes, holding gene location (and window definitions) constant. Over 10,000 permutations, we determined whether an observed K_s value for a window was extreme. Figure 2 provides an example whereby K_s values in a window are elevated for some regions near the centromeres and in the region spanning 24–27 Mb on chromosome 1. Generally, when K_s values were extreme, they tended to be elevated in arm regions proximal to centromeres and reduced near telomeres (fig. 2; supplementary figs. S1–S4, Supplementary Material online). We also performed permutation analyses for K_a and ω for which there were generally fewer regions of significantly high and low rates compared with K_s (fig. 2; supplementary figs. S1–S4, Supplementary Material online).

Determinants of Evolutionary Rates

To perform a general analysis of the factors that contribute to evolutionary rates, we selected 14 variables that might correlate with (or contribute to) evolutionary rates (table 2). The 14 variables were available for 5439 orthologs. Most of these

Table 1. Evolutionary Rates for Duplicates and Singletons.

		Early WGDs	Recent WGDs	Non-WGDs	Singletons	Total
K_a	Mean (SD)	0.021 (0.017)	0.024 (0.018)	0.031 (0.029)	0.032 (0.027)	0.028 (0.026)
	CV	0.81	0.75	0.94	0.84	0.93
	Range	0–0.16	0–0.15	0–0.28	0–0.29	0–0.29
	Range (90%)	0.0035–0.050	0.0039–0.060	0.0048–0.090	0.0076–0.063	0.0044–0.075
K_s	Mean (SD)	0.147 (0.041)	0.145 (0.042)	0.150 (0.045)	0.144 (0.049)	0.147 (0.044)
	CV	0.28	0.29	0.30	0.34	0.30
	Range	0.039–0.29	0.021–0.30	0.015–0.30	0–0.29	0–0.30
	Range (90%)	0.084–0.22	0.083–0.22	0.084–0.23	0.087–0.21	0.082–0.23
K_a/K_s	Mean (SD)	0.147 (0.114)	0.178 (0.136)	0.216 (0.208)	0.244 (0.260)	0.203 (0.194)
	CV	0.78	0.76	0.96	1.07	0.96
	Range	0.001–1.21	0.001–1.52	0.001–4.21	0.001–4.35	0.001–4.35
	Range (90%)	0.024–0.35	0.030–0.43	0.035–0.59	0.057–0.46	0.032–0.54
d_{SM}	Mean (SD)	0.183 (0.194)	0.192 (0.206)	0.244 (0.240)	0.299 (0.267)	0.229 (0.233)
	CV	1.06	1.07	0.98	0.89	1.02
	Range	0–1.00	0–1.00	0–1.00	0–1.00	0–1.00
	Range (90%)	0.008–0.616	0.008–0.628	0.010–0.75	0.012–0.819	0.009–0.715

NOTE.—SD, standard deviation; CV, coefficient of variation.

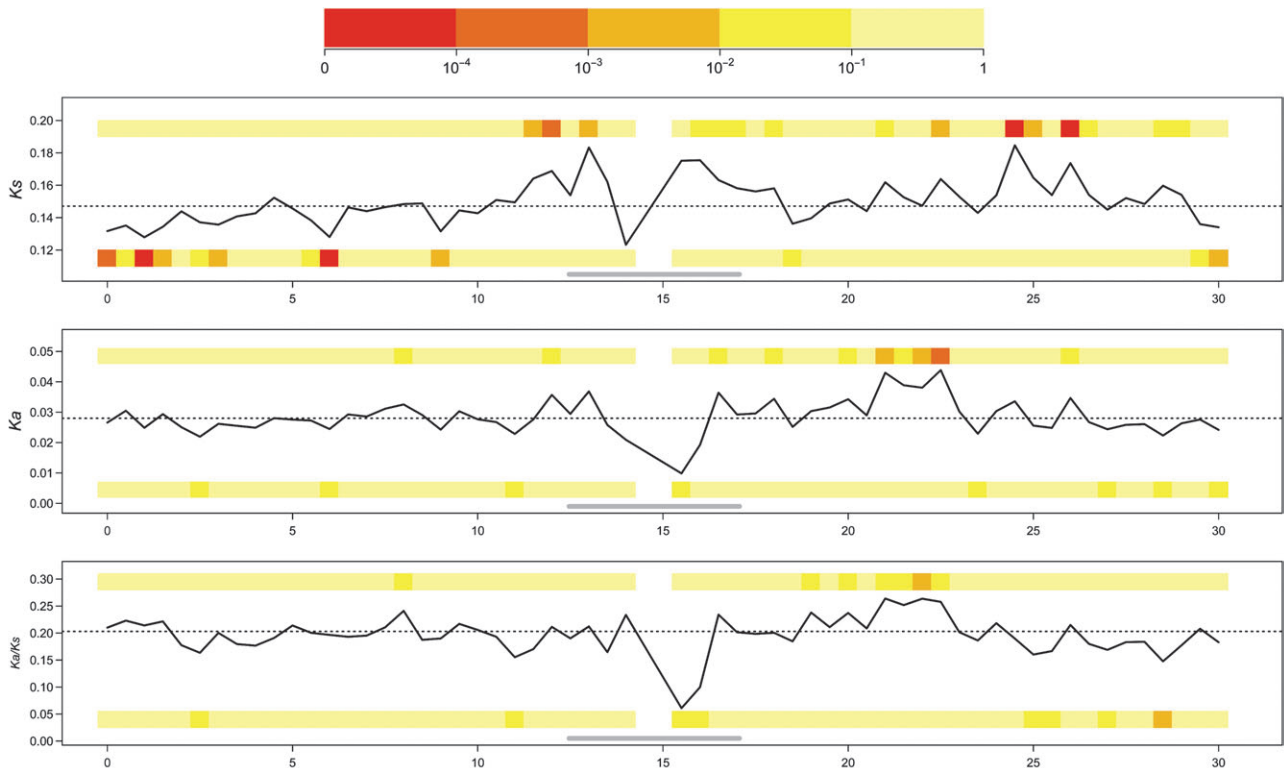

Fig. 2. The distributions of mean values of K_s, K_a, and K_a/K_s (ω) for 0.5 Mb nonoverlapping windows along chromosome 1 in *Arabidopsis thaliana*. To test whether divergence values within windows were higher or lower than expected, we randomly permuted the *K* values among genes, holding gene location (and window definitions) constant. Over 10,000 permutations, we determined whether the observed value for a window was extreme. The top bar in each plot shows the *P* values that the observed value is higher than expected; the bottom bar in each plot shows the *P* values that the observed value is lower than expected. The dotted lines indicate the mean values of evolutionary rates for all genes on chromosome 1.

variables were correlated with evolutionary rates in pairwise fashion (table 2). For example, all 14 variables except recombination rate were significantly correlated with K_a and ω (Spearman's rank correlation, $P < 10^{-9}$). Most variables were correlated with K_s as well, but the pattern differed slightly from K_a (table 2); for example, duplication mode was correlated with K_a but not K_s.

Of course, many of these factors are intercorrelated, making it difficult to identify the contribution of individual factors to evolutionary rates. Although not without limitations (see Discussion), PCR is one approach to begin to tease apart the separate contribution of each predictor to the total variation in evolutionary rate among genes (Drummond et al. 2006). In this method, the 14 total predictor variables are scaled and then transformed orthogonally. In theory, the greatest variance explained by any projection of the data lies on the first principal component, and the factors with the greatest contribution to rate variation can be inferred.

We applied the PCR method separately to K_s and K_a variation. With regard to K_s, the key results are provided in figure 3, with variance estimates in supplementary table S2, Supplementary Material online. The first principal component represented 4.4% of rate variation, and in total 11.1% of the variation among genes was explained across significant principal components (supplementary table S2, Supplementary Material online). The first two ranked factors in the first principal component were intron number (2.0% of variation) and gene length (1.9%), both of which clearly superseded the remaining 12 factors in the percent of variance explained (fig. 3). It also should be noted that 1.7% of rate variation was explained by the combination of "chromosomal position" and "recombination rate" across principal components, suggesting that gene location is an important factor for describing $K_{s.}$ In contrast,

Table 2. Pairwise Correlations of Evolutionary Rates with Potentially Contributing Factors.

Variable	K_a	K_s	K_a/K_s
Duplication mode	0.13***	0.01	0.13***
Chromosomal position	−0.06***	−0.17***	0.001
Recombination rate	0.04*	0.12***	−0.002
Expression level	−0.42***	−0.09***	−0.39***
Tissue specificity (τ)	0.28***	0.12***	0.24***
d_{SM}	0.18***	0.11***	0.14***
F_{op}	−0.12***	0.04*	−0.14***
Multifunctionality	−0.19***	$-0.03^{\#}$	−0.18***
Gene length	−0.25***	−0.20***	−0.18***
5′ UTR length	−0.20***	−0.13***	−0.15***
3′ UTR length	−0.20***	−0.10***	−0.16***
Intron number	−0.19***	−0.26***	−0.10***
Average intron length	−0.11***	−0.03*	−0.09***
G + C content	−0.20***	−0.01	−0.20***

Note.—The coefficients were calculated based on Spearman rank correlation. $^{\#}P < 0.05$, $*P < 10^{-3}$, $**P < 10^{-6}$, $***P < 10^{-9}$.

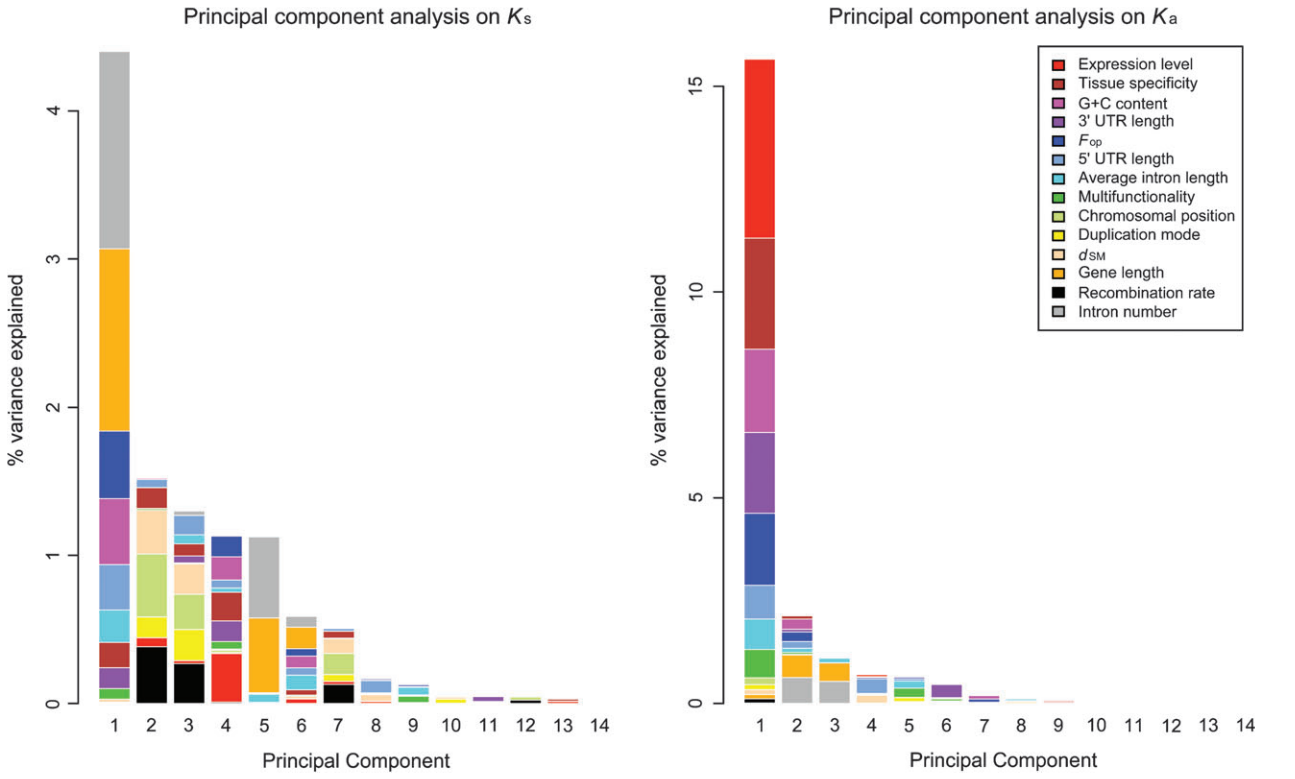

FIG. 3. Principal components regression on K_s and K_a. See supplementary tables S2 and S3, Supplementary Material online, for numerical data.

duplication status contributed very little (0.44%) to variation in K_s among genes.

For K_a, the first component explained 15.7% of rate variation, and all principle components captured a total of 21.4% of variation (fig. 3; supplementary table S3, Supplementary Material online). Among the 14 predictors, those that contributed most heavily to K_a variation were expression level and tissue specificity (τ), indicating gene expression best explains K_a variation among genes (see Discussion). Additional parameters of interest include G + C content, codon usage (F_{op}) and 3′ UTR lengths.

Evolutionary Rates between Duplicates and Singletons

Duplication status is not correlated with K_s variation among genes (table 2) and is at best a minor contributor to K_a variation based on PCR analysis (fig. 3). Yet, the dynamics of duplication and evolutionary rate are potentially interesting in their own right (Yang et al. 2003; Davis and Petrov 2004; Jordan et al. 2004). Accordingly, we contrasted the evolutionary rates of four groups that were categorized on the basis of their mode of duplication (see Materials and Methods): singletons (1,497 genes), early WGD (960 genes), recent-WGD (3,351 genes) and non-WGD duplicates (5,684 genes). This last category (non-WGD duplicates) may be best described as a "catch-all" category of duplicates of uncertain origin, including duplicates that are due to WGD events but not detected as such, tandem events, and duplications by transposition (Freeling et al. 2008).

Among these four categories, there was no statistical difference in the K_s distributions (Mann–Whitney U test, $P >$ 0.01) (fig. 4A and supplementary fig. S5, Supplementary Material online). This result is consistent with pairwise correlations (table 2) and is expected if synonymous substitutions are approximately neutral and if genes have a similar divergence time. However, there were clear differences in the distributions of K_a and ω among categories. As a group, duplicated genes had significantly lower K_a and ω values than those of singleton genes (Mann–Whitney U test, $P < 10^{-10}$ for both; fig. 4B–C and supplementary figs. S6–S7, Supplementary Material online). Within classes of duplicated genes, early WGD duplicates had lower K_a and ω values than recent-WGD duplicates and non-WGD duplicates (Mann–Whitney U test, $P < 10^{-9}$ for both). There is thus evidence for K_a differences among all four duplication categories.

To attempt to verify these inferences, we analyzed upstream sequences. We found that d_{SM} for duplicated genes was significantly lower than those of singletons (Mann–Whitney U test, $P < 10^{-10}$; fig. 4D), mirroring the K_a results (fig. 4B). The d_{SM} results also corroborated the difference in K_a between WGD and non-WGD duplicates but, importantly, yielded no significant difference between early- and recent WGD classes (Mann–Whitney U test, $P =$ 0.56; fig. 4D). Overall, these d_{SM} results: i) suggest that protein divergence is correlated with divergence in upstream regions, and indeed K_a and d_{SM} were correlated over the entire data set (supplementary table S1, Supplementary Material online); ii) imply that the distinction between

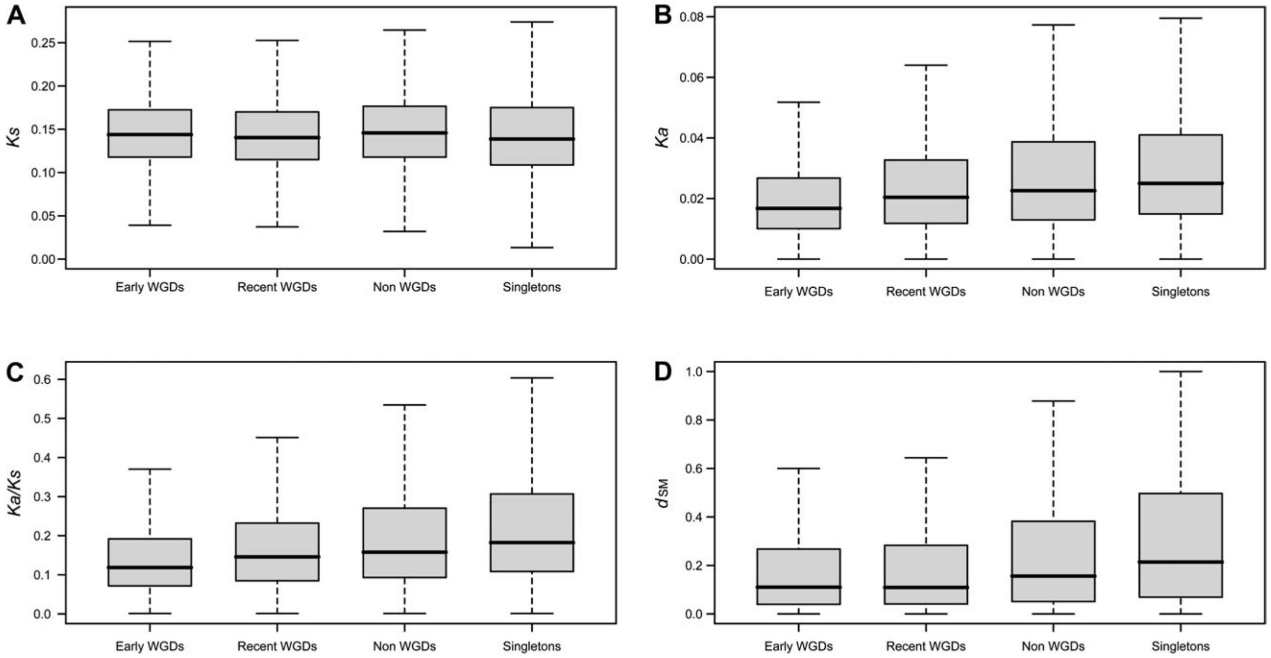

Fig. 4. The comparisons of K_s (*A*), K_a (*B*), K_a/K_s (*C*), and d_{SM} (*D*) among different types of duplicated and singleton genes. The bottom and top of each box are the first (lower) and third (higher) quartiles, and the band in the box is the median value. The ends of the whiskers represent 1.5 interquartile range of lower and higher quartiles, respectively.

duplicates and singletons is neither limited to amino acid replacements nor an artifact of coding region alignments but iii) do not provide independent confirmation that early- and recent WGD duplicates evolve at different rates.

Contrasts between Duplication Status and Gene Expression

Our preceding results reveal a potential inconsistency. On the one hand, PCR analyses attribute only a small proportion of rate variance to duplication status (fig. 3). On the other hand, direct contrasts reveal compelling K_a differences among some duplication categories (fig. 4*B*). Although there could be several reasons for this inconsistency, including shortcomings of the PCR method (see Discussion), one potential interpretation is that the orthogonal transformation in PCR removes an underlying correlation between gene expression and duplication. To try to tease apart the potential relationship between expression and duplication, we contrasted levels and patterns of *A. thaliana* gene expression among duplication categories for the 10,021 genes with expression data.

For average expression level, non-WGD genes are the outlier relative to the other three duplication categories and expressed at lower levels than the other gene classes (Mann–Whitney *U* test, $P < 10^{-9}$ for all three comparisons; fig. 5*A*). The other three categories did not differ statistically for expression level, but singletons were expressed at slightly higher levels (fig. 5*A*). Analysis of MPSS data yielded similar results but with singletons having significantly elevated expression over duplicates (data not shown).

Tissue specificity, τ, differed more widely among groups, with singletons expressed more broadly than duplicates (Mann–Whitney *U* test, $P < 10^{-15}$ for all three comparisons; fig. 5*B*) and early WGD genes expressed more specifically than either the recent WGD (Mann–Whitney *U* test, $P < 10^{-5}$) or the non-WGD (Mann–Whitney *U* test, $P < 10^{-6}$) genes. Analysis of MPSS data confirmed that singletons are expressed more broadly than duplicates (data not shown).

Discussion

Surprisingly, few studies have described evolutionary rate variation among plant nuclear genes, and the exceptions have been based on relatively small sample sizes. For example, Zhang et al. (2002) examined rate variation among a group of 242 genes, using contemporaneously duplicated *A. thaliana* paralogs to measure divergence; Tiffin and Hahn (2002) studied 218 putative orthologs between *A. thaliana* and *Brassica rapa*; and Wright et al. (2004) analyzed 83 orthologs between *A. thaliana* and *A. lyrata*, which represented the largest ortholog data set between these two species that were available at the time. Although there have been additional multigene studies of evolutionary rates among plant taxa (e.g., Wang et al. 2008), to our knowledge, none have approached the genome-wide scale of the analyses reported here, with >11,000 orthologous pairs.

Despite the limited number of genes in previous studies, they revealed similar patterns of rate variation among plant nuclear genes. For example, Zhang et al. (2002) documented ~14-fold range of synonymous rate variation among genes, with 90% of genes represented in a more narrow window of 2.6-fold rate variation. Our results also indicate that 90% of genes fall within a window of 2.6-fold rate variation (table 1). The consistency between studies is remarkable, especially given that Zhang et al. (2002)

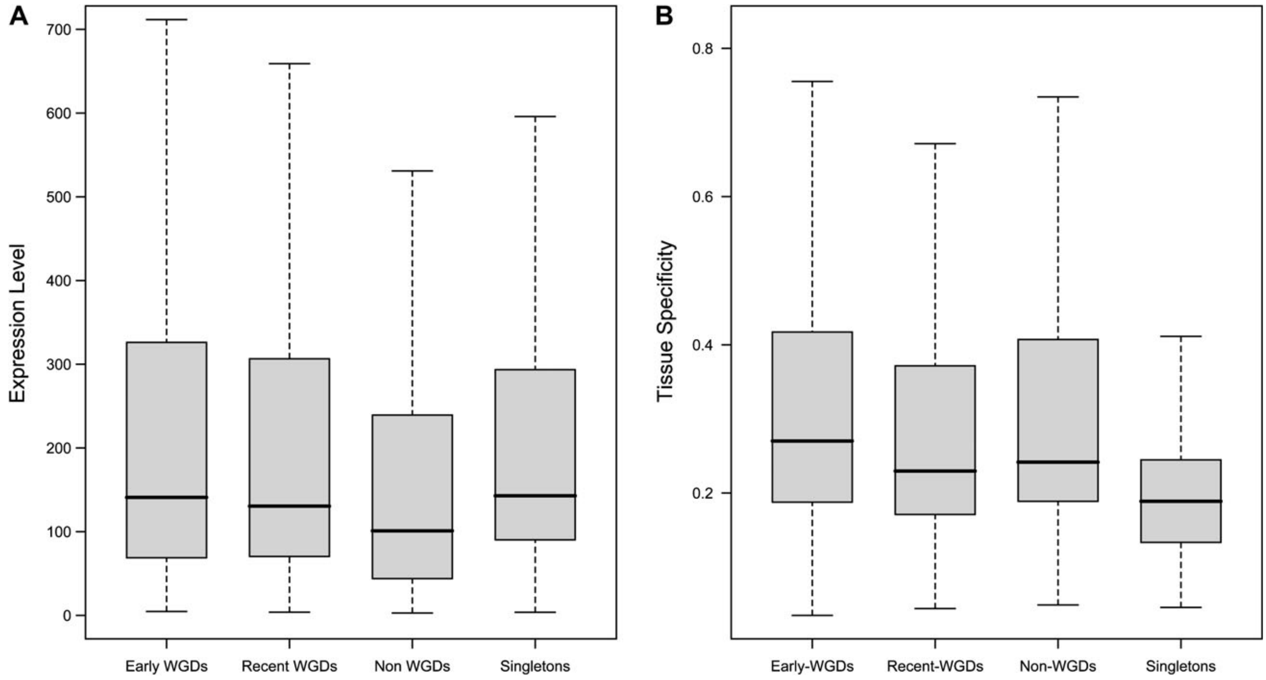

Fig. 5. The comparisons of total expression level (*A*) and specificity (τ) (*B*) among different types of duplicated and singleton genes. The bottom and top of each box are the first (lower) and third (higher) quartiles, and the band in the box is the median value. The ends of the whiskers represent 1.5 interquartile range.

studied paralogs potentially subjected to gene conversion that also diverged on a time frame roughly an order of magnitude higher than that of the orthologs studied here.

Our results are also consistent with these previous papers both in estimating an average ω of ~0.2, signaling strong constraint on amino acid replacements, and in identifying very few genes with ω values >1.0. It is worth noting that the small number (90) of genes with $\omega > 1.0$ have no obvious functional biases as measured by GO analyses (data not shown). Although it may be possible that ortholog contrasts lack statistical power to detect $\omega > 1.0$, the overarching impression is that the type of positive selection detected by ω analyses has not been a common feature of species divergence.

K_a and K_s Are Correlated across Genes

Like many previous studies (Alvarez-Valin et al. 1998; Makalowski and Boguski 1998; Smith and Hurst 1999; Zhang et al. 2002; Castillo-Davis et al. 2004), we document a positive correlation between K_a and K_s across genes. The reason for this correlation is not well established, but it could be caused by at least three nonexclusive phenomena. First, selection for translational speed, efficiency, and accuracy may affect both K_a and K_s (Wright et al. 2004; Drummond and Wilke 2008) and thus drive the correlation. Second, variation in mutation rates along chromosomes may lead to genomic regions that covary in K_a and K_s. Third, even if mutation rates are uniform across chromosomes, deleterious mutations may be culled less effectively from regions of low recombination, again potentially leading to genomic regions that covary in K_a and K_s.

If the latter two phenomena contribute, we expect clustering of evolutionary rates along chromosomes. Indeed, we have identified several 0.5 Mb windows of enhanced or decreased substitution rates (fig. 2; supplementary figs. S1–S4, Supplementary Material online). Perhaps, the most interesting of these are on chromosome 1, which has a marked K_a peak at ~21–24 Mb and a K_s peak from ~24 to 27 Mb (fig. 2). This entire region coincides to a peak of high synonymous nucleotide diversity among *A. thaliana* accessions (Clark et al. 2007). Clark et al. (2007) noticed that this region on chromosome 1 contained several disease resistance genes and hypothesized that this was a region particularly prone to balancing selection. However, the correspondence of high divergence and high polymorphism in the same genomic window strongly implies that there is variation in mutation rates along the chromosome, as postulated for mammalian genomes (Lercher et al. 2001). We thus hypothesize that at least one of the causes of K_a and K_s rate correlation is shared mutation rates and also that variation in mutation rates contributions to rate variation among genes. Interestingly, the region of high rates on chromosome 1 may be syntenous to a centromeric region in *A. lyrata* (Hu et al. 2011), suggesting that chromosomal rearrangements may affect mutation rates.

Gene Expression Is the Major Predictor of Rate Variation among Genes

It is unlikely that mutation rates alone predict rate variation among genes. We thus examined 14 variables to assess their contributions to evolutionary rates. Typically, the correlations between evolutionary rates and predictor variables have been measured by partial correlation or multiple regression. Drummond et al. (2006) demonstrated that these approaches can generate spurious but highly significant results when the predictor variables are measured with error (noise), and they introduced the PCR approach as an alternative. In turn, Plotkin and Fraser (2007) showed that PCR is itself neither robust to measurement noise nor to variation in the predictor variables. Unfortunately, then, there is as yet no ideal method to partition the factors that contribute to evolutionary rates. Some of our predictor variables are virtually free of noise, such as gene length and F_{op}, but other variables—such as gene expression data—do contain noise. Here, we have employed PCR as an exploratory tool, recognizing that the approach can suffer from the same weaknesses as other approaches but has the advantage of easy interpretation.

One outcome of our PCR analysis is that the first principal component explains a small proportion of variation in K_s (4.4%) and K_a (15.7%) among genes. In contrast, expression level alone accounts for >25% of rate variation among genes in *Drosophila* (Lemos et al. 2005), bacteria and *Chlamydomonas* (Rocha and Danchin 2004; Rocha 2006), and ~50% of K_a variation among yeast genes (Drummond et al. 2006).

There may be at least two reasons why our PCR captures comparatively little variation. The first is that our model lacks predictor variables that assess functional importance directly largely because such information is unavailable. Although there is some information about gene essentiality for a subset of *A. thaliana* genes (Hanada et al. 2009), to our knowledge, there is no genome-wide information on (for example) protein dispensability or abundance. To circumvent the lack of functional data, we included "multifunctionality" as a predictor based on GO annotations. Associations between GO functions and evolutionary rates have been detected for *A. thaliana* early WGD genes (Warren et al. 2010), but multifunctionality explains little of the variation across all genes (fig. 3; supplementary tables S2 and S3, Supplementary Material online). Hopefully, future work will be able to incorporate critical functional parameters as they become available. The second reason is that PCR cannot capture all sources of variation. For example, in *A. lyrata* and *A. thaliana*, which diverged recently on an evolutionary timescale (~13 Ma; Beilstein et al. 2010), polymorphisms that segregated in the ancestral species may contribute substantial stochastic variation in divergence among genes. Unfortunately, it is unclear how to incorporate such information into the PCR.

Even though PCR explains only a low percentage of variation, it provides insights into the forces that affect K_s and K_a. For K_s, some contributors are factors, like recombination rate, which vary on chromosomal scales, but intron number and gene length are also major contributors. For K_a, the results are consistent with previous results from plants, yeast, and mammals in revealing a substantial relationship with both the level and the specificity of expression (Duret and Mouchiroud 2000; Pál et al. 2001b; Zhang et al. 2002; Wright et al. 2004; Drummond et al. 2005, 2006; Wall et al. 2005; Drummond and Wilke 2008). The favored explanation for a positive correlation between K_a and expression specificity is that widely expressed genes are constrained either due to multiple selective environments in different cells due to multiple biochemical contexts (Duret and Mouchiroud 2000) or due to high functional densities. Similarly, the positive correlation between K_a and expression level may reflect high purifying selection on abundant proteins, selection on the speed and accuracy of translation, or selection for robustness against mistranslations (Drummond et al. 2005; Drummond and Wilke 2008). We cannot differentiate among those explanations here, but our study establishes that gene expression is a major predictor of K_a variation on a genomic scale in a plant system.

What Drives the Evolutionary Rate of Duplicated Genes?

Previous studies have shown that singleton genes evolve more rapidly than duplicated genes (Nembaware et al. 2002; Yang et al. 2003; Davis and Petrov 2004; Jordan et al. 2004). By considering different classes of duplicated genes, we have uncovered a somewhat more complex pattern of nonsynonymous evolution. On average, singletons evolve more rapidly than duplicates; non-WGD genes evolve more rapidly than WGD genes; and recent WGD genes evolve more rapidly than early WGD genes (fig. 4*B*). The hierarchy of K_a divergence is corroborated by d_{SM} (fig. 4*D*). The lone exception to this corroboration, for which we do not have an explanation, is that the recent WGD and early WGD promoters evolve at the same rate but the early WGD genes have statistically lower K_a values. Nonetheless, similar patterns of K_a and d_{SM} indicate that protein divergence is correlated with divergence in upstream regions (Castillo-Davis et al. 2004; Chin et al. 2005), suggesting that some aspects of selective constraint are shared between protein and upstream regions.

The four groups also vary with respect to levels and patterns of gene expression (fig. 5). Singletons are expressed more broadly than duplicated genes and tend to be expressed at higher levels; non-WGD genes are lowly expressed, on average; and the slowly evolving early WGD genes are expressed with high specificity. Interestingly, these patterns of gene expression do not follow the expected correlation with evolutionary rate. Previous studies have found that genes with low K_a are expressed in more tissues at "higher" levels (Duret and Mouchiroud 2000; Pál et al. 2001b; Subramanian and Kumar 2004; Zhang and Li 2004), and our groups do not meet this expectation. For example, the slowest evolving group (early WGD duplicates) has neither the highest expression level nor the broadest expression specificity. However, the expected patterns do hold within groups. For example, we compared K_a and expression level within the early WGD duplicates and found the expected negative correlation (supplementary table S4, Supplementary Material online). Overall, these observations suggest that the groups are distinct not only for evolutionary rates but also for some aspect of their expression dynamics.

What, then, drives theses differences among groups? The short answer is that we do not know, but one possibility is that dosage balance plays a critical role. Dosage balance is thought to be a factor in the retention of WGD duplicates (Edger and Pires 2009) because the loss of an individual duplicate destroys stoichiometric relationships in macromolecular complexes and signaling pathways. Gout et al. (2010) have extended this idea by postulating that patterns of gene expression drives the retention and relatively slow evolution of duplicated genes due to an associated cost of perturbing gene expression. In other words, there may be strong selection to retain WGD duplicates, largely through constraint of gene expression. Compared with WGD duplicates, both singletons and non-WGD duplicates may need less adherence to dosage balance and are thus relatively free to diverge both functionally (Rizzon et al. 2006) and in gene expression (Ganko et al. 2007). As a result, one would predict different evolutionary rates among groups, partially as a function of gene expression, as demonstrated here.

Supplementary Material

Supplementary figures S1–S5 and tables S1–S4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work has benefited from the comments of O. Tenaillon, M. Tenaillon, S. Takuno, H. Sakai, B. von Holdt, and two anonymous reviewers. This work was supported by National Science Foundation grant DEB-0723860 to B.S.G.

References

Akhunov ED, Goodyear AW, Geng S, et al. (33 co-authors). 2003. The organization and rate of evolution of wheat genomes are correlated with recombination rates along chromosome arms. *Genome Res*. 13:753–763.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 25:3389–3402.

Alvarez-Valin F, Jabbari K, Bernardi G. 1998. Synonymous and nonsynonymous substitutions in mammalian genes: intragenic correlations. *J Mol Evol*. 46:37–44.

Beilstein MA, Nagalingum NS, Clements MD, Manchester SR, Mathews S. 2010. Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. *Proc Natl Acad Sci U S A*. 107:18724–18728.

Blanc G, Hokamp K, Wolfe KH. 2003. A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Res*. 13:137–144.

Blanc G, Wolfe KH. 2004. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell*. 16:1679–1691.

Bowers JE, Chapman BA, Rong J, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422:433–438.

Castillo-Davis CI, Hartl DL, Achaz G. 2004. *cis*-Regulatory and protein evolution in orthologous and duplicate genes. *Genome Res*. 14:1530–1536.

Chin CS, Chuang JH, Li H. 2005. Genome-wide regulatory complexity in yeast promoters: separation of functionally conserved and neutral sequence. *Genome Res*. 15:205–213.

Clark RM, Schweikert G, Toomajian C, et al. (18 co-authors). 2007. Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science* 317:338–342.

Davis JC, Petrov DA. 2004. Preferential duplication of conserved proteins in eukaryotic genomes. *PLoS Biol*. 2:E55.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci U S A*. 102:14338–14343.

Drummond DA, Raval A, Wilke CO. 2006. A single determinant dominates the rate of yeast protein evolution. *Mol Biol Evol*. 23:327–337.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant contraint on coding-sequence evolution. *Cell*. 134:341–352.

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

Edger PP, Pires JC. 2009. Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Res*. 17:699–717.

Freeling M, Lyons E, Pedersen B, Alam M, Ming R, Lisch D. 2008. Many or most genes in *Arabidopsis* transposed after the origin of the order Brassicales. *Genome Res*. 18:1924–1937.

Ganko EW, Meyers BC, Vision TJ. 2007. Divergence in expression between duplicated genes in *Arabidopsis*. *Mol Biol Evol*. 24:2298–2309.

Gaut BS. 2001. Patterns of chromosomal duplication in maize and their implications for comparative maps of the grasses. *Genome Res*. 11:55–66.

Gaut BS, Ross-Ibarra J. 2008. Selection on major components of angiosperm genomes. *Science* 320:484–486.

Gout JF, Kahn D, Duret L. 2010. The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. *PLoS Genet*. 6:e1000944.

Gu X. 2003. Evolution of duplicate genes versus genetic robustness against null mutations. *Trends Genet*. 19:354–356.

Gu Z, Cavalcanti A, Chen FC, Bouman P, Li WH. 2002. Extent of gene duplication in the genomes of *Drosophila*, nematode, and yeast. *Mol Biol Evol*. 19:256–262.

Hanada K, Kuromori T, Myouga F, Toyoda T, Li WH, Shinozaki K. 2009. Evolutionary persistence of functional compensation by duplicate genes in *Arabidopsis*. *Genome Biol Evol*. 1:409–414.

Hirsh AE, Fraser HB. 2001. Protein dispensability and rate of evolution. *Nature* 411:1046–1049.

Hu TT, Pattyn P, Bakker EG, et al. (30 co-authors). 2011. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat Genet*. (in press).

Ihaka R, Gentleman R. 1996. R: a language for data analysis and graphics. *J Comput Graphical Stat*. 5:299–314.

Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol*. 2:13–34.

Jaillon O, Aury JM, Noel B, et al. (56 co-authors). 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449:463–467.

Jolliffe IT. 2002. Principal component analysis. New York: Springer.

Jordan IK, Wolf YI, Koonin EV. 2004. Duplicated genes evolve slower than singletons despite the initial rate increase. *BMC Evol Biol*. 4:22.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res*. 110:462–467.

Larracuente AM, Sackton TB, Greenberg AJ, Wong A, Singh ND, Sturgill D, Zhang Y, Oliver B, Clark AG. 2008. Evolution of protein-coding genes in *Drosophila*. *Trends Genet*. 24:114–123.

Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL. 2005. Evolution of proteins and gene expression levels are coupled in *Drosophila* and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. *Mol Biol Evol*. 22:1345–1354.

Lercher MJ, Williams EJ, Hurst, LD. 2001. Local similarity in evolutionary rates extends over whole chromosomes in human-rodent and mouse-rat comparisons: implications for understanding the mechanistic basis of the male mutation bias. *Mol Biol Evol.* 18:2023–2029.

Liao BY, Zhang J. 2006. Low rates of expression profile divergence in highly expressed genes and tissue-specific genes during mammalian evolution. *Mol Biol Evol.* 23: 1119–1128.

Lobkovsky AE, Wolf YI, Koonin EV. 2010. Universal distribution of protein evolution rates as a consequence of protein folding physics. *Proc Natl Acad Sci U S A.* 107: 2983–2988.

Lockton S, Gaut BS. 2005. Plant conserved non-coding sequences and paralogue evolution. *Trends Genet.* 21:60–65.

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y. 2005. Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci U S A.* 102:5454–5459.

Makalowski W, Boguski MS. 1998. Synonymous and nonsynonymous substitution distances are correlated in mouse and rat genes. *J Mol Evol.* 47:119–121.

Marais G, Duret L. 2001. Synonymous codon usage, accuracy of translation, and gene length in *Caenorhabditis elegans*. *J Mol Evol.* 52:275–280.

Meyers BC, Tej SS, Vu TH, Haudenschild CD, Agrawal V, Edberg SB, Ghazal H, Decola S. 2004. The use of MPSS for whole-genome transcriptional analysis in Arabidopsis. *Genome Res.* 14: 1641–1653.

Nembaware V, Crum K, Kelso J, Seoighe C. 2002. Impact of the presence of paralogs on sequence divergence in a set of mouse-human orthologs. *Genome Res.* 12:1370–1376.

Pál C, Papp B, Hurst LD. 2001a. Does the recombination rate affect the efficiency of purifying selection? The yeast genome provides a partial answer. *Mol Biol Evol.* 18:2323–2326.

Pál C, Papp B, Hurst LD. 2001b. Highly expressed genes in yeast evolve slowly. *Genetics* 158:927–931.

Pál C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. *Nat Rev Genet* 7:337–348.

Plotkin JB, Fraser HB. 2007. Assessing the determinants of evolutionary rates in the presence of noise. *Mol Biol Evol.* 24: 1113–1121.

Prachumwat A, Li WH. 2006. Protein function, connectivity, and duplicability in yeast. *Mol Biol Evol.* 23:30–39.

Rezvoy C, Charif D, Gueguen L, Marais GAB. 2007. MareyMap: an R-based tool with graphical interface for estimating recombination rates. *Bioinformatics* 23:2188–2189.

Rizzon C, Ponger L, Gaut BS. 2006. Striking similarities in the genomic distribution of tandemly arrayed genes in *Arabidopsis* and rice. *PLoS Comput Biol.* 2:e115.

Rocha EP. 2006. The quest for the universals of protein evolution. *Trends Genet.* 22:412–416.

Rocha EP, Danchin A. 2004. An analysis of determinants of amino acids substitution rates in bacterial proteins. *Mol Biol Evol.* 21:108–116.

Rost B. 1999. Twilight zone of protein sequence alignments. *Protein Eng.* 12:85–94.

Salathe M, Ackermann M, Bonhoeffer S. 2006. The effect of multifunctionality on the rate of evolution in yeast. *Mol Biol Evol.* 23:721–722.

Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Scholkopf B, Weigel D, Lohmann JU. 2005. A gene expression map of *Arabidopsis thaliana* development. *Nat Genet.* 37:501–506.

Schranz ME, Mitchell-Olds T. 2006. Independent ancient polyploidy events in the sister families *Brassicaceae* and *Cleomaceae*. *Plant Cell.* 18:1152–1165.

Sharp PM, Li WH. 1987. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Mol Biol Evol.* 4:222–230.

Simillion C, Vandepoele K, Van Montagu MC, Zabeau M, Van de Peer Y. 2002. The hidden duplication past of *Arabidopsis thaliana*. *Proc Natl Acad Sci U S A.* 99:13627–13632.

Singer T, Fan Y, Chang HS, Zhu T, Hazen SP, Briggs SP. 2006. A high-resolution map of *Arabidopsis* recombinant inbred lines by whole-genome exon array hybridization. *PLoS Genet.* 2(9):e144.

Smith NG, Hurst LD. 1999. The effect of tandem substitutions on the correlation between synonymous and nonsynonymous rates in rodents. *Genetics* 153:1395–1402.

Soltis PS, Soltis DE. 2009. The role of hybridization in plant speciation. *Annu Rev Plant Biol.* 60:561–588.

Subramanian S, Kumar S. 2004. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. *Genetics.* 168:373–381.

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

Urrutia AO, Hurst LD. 2001. Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. *Genetics* 159:1191–1199.

Vision TJ, Brown DG, Tanksley SD. 2000. The origins of genomic duplications in *Arabidopsis*. *Science* 290:2114–2117.

Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, Eisen MB, Feldman MW. 2005. Functional genomic analysis of the rates of protein evolution. *Proc Natl Acad Sci U S A.* 102:5483–5488.

Wang Y, Diehl A, Wu F, Vrebalov J, Giovannoni J, Siepel A, Tanksley SD. 2008. Sequencing and comparative analysis of a conserved syntenic segment in the *Solanaceae*. *Genetics* 180:391–408.

Warren AS, Anandakrishnan R, Zhang L. 2010. Functional bias in molecular evolution rate of *Arabidopsis thaliana*. *BMC Evol Biol.* 10:125.

Wright SI, Yau CB, Looseley M, Meyers BC. 2004. Effects of gene expression on molecular evolution in *Arabidopsis thaliana* and *Arabidopsis lyrata*. *Mol Biol Evol.* 21:1719–1726.

Yanai I, Benjamin H, Shmoish M, et al. (12 co-authors). 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics.* 21:650–659.

Yang J, Gu Z, Li WH. 2003. Rate of protein evolution versus fitness effect of gene deletion. *Mol Biol Evol.* 20:772–774.

Yang L, Takuno S, Waters ER, Gaut BS. 2011. Lowly-expressed genes in *Arabidopsis thaliana* bear the signature of possible pseudogenization by promoter degradation. *Mol Biol Evol.* 28:1193–1203.

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

Zeldovich KB, Chen P, Shakhnovich EI. 2007. Protein stability imposes limits on organism complexity and speed of molecular evolution. *Proc Natl Acad Sci U S A.* 104:16152–16157.

Zhang LQ, Li WH. 2004. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Mol Biol Evol.* 21:236–239.

Zhang LQ, Vision TJ, Gaut BS. 2002. Patterns of nucleotide substitution among simultaneously duplicated gene pairs in *Arabidopsis thaliana*. *Mol Biol Evol.* 19:1464–1473.

Zuckerkandl E. 1976. Evolutionary processes and evolutionary noise at the molecular level. I. Functional density in proteins. *J Mol Evol.* 7:167–183.