Genetics and Molecular Biology, 31, 2, 575-584 (2008)
Copyright © 2008, Sociedade Brasileira de Genética. Printed in Brazil
www.sbg.org.br

Research Article

Rapid sequence divergence rates in the 5 prime regulatory regions of young *Drosophila melanogaster* duplicate gene pairs

Michael H. Kohn

Ecology and Evolutionary Biology, Rice University, Houston, Texas, United States of America.

Abstract

While it remains a matter of some debate, rapid sequence evolution of the coding sequences of duplicate genes is characteristic for early phases past duplication, but long established duplicates generally evolve under constraint, much like the rest of the coding genome. As for coding sequences, it may be possible to infer evolutionary rate, selection, and constraint via contrasts between duplicate gene divergence in the 5 prime regions and in the corresponding synonymous site divergence in the coding regions. Finding elevated rates for the 5 prime regions of duplicated genes, in addition to the coding regions, would enable statements regarding the early processes of duplicate gene evolution. Here, 1 kb of each of the 5 prime regulatory regions of *Drosophila melanogaster* duplicate gene pairs were mapped onto one another to isolate shared sequence blocks. Genetic distances within shared sequence blocks ($d_{5'}$) were found to increase as a function of synonymous (d_S), and to a lesser extend, amino-acid (d_A) site divergence between duplicates. The rate $d_{5'}/d_S$ was found to rapidly decay from values > 1 in young duplicate pairs ($d_S < 0.3$) to 0.28 or less in older duplicates ($d_S > 0.8$). Such rapid rates of 5 prime evolution exceeding 1 (~neutral) predominantly were found to occur in duplicate pairs with low amino-acid site divergence and that tended to be co-regulated when assayed on microarrays. Conceivably, functional redundancy and relaxation of selective constraint facilitates subsequent positive selection on the 5 prime regions of young duplicate genes. This might promote the evolution of new functions (neofunctionalization) or division of labor among duplicate genes (subfunctionalization). In contrast, similar to the vast portion of the non-coding genome, the 5 prime regions of long-established gene duplicates appear to evolve under selective constraint, indicating that these long-established gene duplicates have assumed critical functions.

Key words: gene duplication, gene expression, selection, promoter evolution.

Received: October 23, 2007; Accepted: March 13, 2008.

Introduction

The alignment of orthologous sequences sampled from two or more related species can reveal evolutionarily conserved sequence blocks, an approach referred to as 'phylogenetic footprinting' (*e.g.* Fickett and Wasserman, 2000). The approach relies on the assumption that sequence blocks that contain functionally important motifs evolve under functional constraint (purifying selection), and thus, remain similar in their sequence over long periods of time (*e.g.* Koop, 1995). In contrast, alignments of non-functional sequences that evolve free of such constraint usually are less clear or not significant (Bergman and Kreitman, 2001). Overall, the footprint of varying degrees of selective constraint along alignments of orthologous, or homologous, sequences is manifest as a mosaic pattern of aligned and non-aligned sequence blocks (Bergman and Kreitman, 2001; Shabalina *et al.*, 2001; Bergman *et al.*, 2002; Castresana, 2002; Webb *et al.*, 2002). In non-coding sequences, such as enhancers and promoters, sequence blocks conserved between orthologs may be enriched for potential transcription factor binding sites (Fickett and Wasserman, 2000; Berman *et al.*, 2002). As more whole genome sequences begin to accumulate in the databases, comparative genomic approaches have become widely applied to aid with the annotation and evolutionary study of non-coding DNA (de Meaux, 2006; Haberer *et al.*, 2006; Li and Stephan, 2006; Hahn, 2007; Thomas *et al.*, 2007).

A wide range of evolutionary divergence times is captured within a single genome through the duplication of genes and their subsequent divergence (*e.g.* Ohno, 1970; Lynch and Conery, 2000; Conery and Lynch, 2001). Presumably, the extent to which gene duplicates, or paralagous genes, occur in the genome reflects their potential to provide a source for biological adaptation and diversification (*e.g.* Zhang *et al.*, 1998; Lynch and Conery, 2000; Conant

Send correspondence to Michael H. Kohn. Ecology and Evolutionary Biology, Rice University, MS 170, P.O. Box 1892, 77005-1892 Houston, Texas, United States of America. E-mail: hmkohn@rice.edu.

and Wagner, 2002; Gu *et al.*, 2002a; Hughes, 2002; Zhang, 2003). In recognition of the pivotal role gene duplication may play in evolution the mechanisms driving their origins and preservation have been a vibrant field of study that is experiencing a renaissance owing to the ever-growing number of genome sequencing projects (*e.g.* Ohno, 1970; Ohta, 1987; Clark, 1994; Hughes, 1994; Ohta, 1994; Walsh, 1995; King, 1998; Force *et al.*, 1999; Lynch and Force, 2000; Wagner, 2001; Hughes, 2002; Wagner, 2002a; Zhang, 2003; Taylor and Raes, 2004).

Whereas the origin and subsequent silencing of duplicate genes both appear to be frequent events, the evolutionary trajectories conducive to duplicate gene preservation may be restrictive (Force *et al.*, 1999; Lynch and Force, 2000). Importantly, the complement of functional duplicate genes that is sampled by genome sequencing projects and that can be studied for their molecular evolution should be comprised predominantly of those that have passed the 'selective sieve'. In other words, gene duplications detrimental to fitness have been removed by purifying selection and gene duplications free of selective constraint may have undergone mutations that rendered them non-functional pseudogenes whose evolution is governed by drift. Functional diversification of duplicates leading to the evolution of novel functions (neo-functionalization), or the partitioning of labor between them (sub-functionalization) could provide avenues for escape from non-functionalization and loss, because purifying selection would remove detrimental mutations from the functional duplicate genes once these have become indispensable (Ohta, 1988; Basten and Ohta, 1992; Hughes, 1994; Walsh, 1995; Force *et al.*, 1999; Lynch and Force, 2000; Wagner, 2002a,b).

Mutations in the 5 prime *cis*-regulatory regions of gene duplicates may promote functional diversification of duplicate genes (Wagner, 2000; Gu *et al.*, 2002b; Makova and Li, 2003; Papp *et al.*, 2003). To examine this possibility the 5 prime regulatory regions of gene duplicates could be searched for the footprint regulatory diversification, be it through positive selection or the loss of constraint (*i.e.* neutral processes), may have left. One such approach would be to compare the rate of divergence in the 5 prime regions relative to that at synonymous sites (Bird *et al.*, 2006; Eyre-Walker, 2006; Hahn 2007), as long as it is assumed synonymous sites follow neutral dynamics (see Akashi, 1995).

Here the evolution of 5 prime regulatory sequences of duplicate gene pairs in the *D. melanogaster* genome was studied. Specifically, (i) 1 kb of each of the 5 prime regions of the two members of a duplicate gene pairs identified previously (Lynch and Conery, 2000; Conery and Lynch 2001) were aligned. It was assumed that blocks of aligned sequence indicate regions of homology preserved owing to their recent divergence and/or by purifying selection. In analogy to phylogenetic footprinting this approach has been dubbed 'intragenomic footprinting' (Haberer *et al.* 2004; Haberer *et al.* 2006). (ii) Divergence of the 5 prime regions of duplicate gene pairs ($d_{5'}$) was expressed relative to divergence at synonymous sites (d_S) and amino-acid replacement sites (d_A) in these gene pairs. This is analogous to studies considering rates of coding sequence evolution of duplicate genes (*e.g.* Ohta, 1994; Lynch and Conery, 2000; Barrier *et al.*, 2001; Conery and Lynch, 2001; Thornton and Long, 2002; Kondrashov, 2005; Kondrashov and Kondrashov, 2006). (iii) Gene expression data from microarray experiments was compiled and related to *Drosophila* duplicate gene divergence (c.f. Wagner, 2000; Gu *et al.*, 2002b; Makova and Li, 2003; Castillo-Davis *et al.*, 2004; Haberer *et al.*, 2004; Casneuf *et al.*, 2006; Wang *et al.*, 2006; Tirosh and Barkai, 2007).

Methods

Collection and analysis of sequence data: The identification numbers for a set of 456 *D. melanogaster* duplicate gene pairs (Lynch and Conery, 2000, Conery and Lynch, 2001) were retrieved from (http://www.csi.uoregon.edu/projects/genetics/duplications/D.melanogaster.txt) and 1 kilobase (kb) of the nucleotide sequences annotated as the upstream 5 prime flanking regions and 5 prime untranslated regions (5' UTR) were retrieved for each gene via the Berkley Drosophila Genome Project (BDGP, Release 2) (http://www.fruitfly.org). Estimates of synonymous site divergence (d_S) and amino acid replacement site divergence (d_A) for the protein coding sequences of each duplicate gene pair were adopted from Lynch and Conery (2000) and Conery and Lynch (2001), who deduced them using PAML (Yang, 1997).

Duplicate gene pairs were grouped into divergence bins: $d_S < 0.1$ (N = 19), $0.1 < d_S < 0.25$ (N = 20), $0.25 < d_S < 0.5$ (N = 27), $0.5 < d_S < 0.75$ (N = 15), $0.75 < d_S < 1.0$ (N = 14), $1.0 < d_S < 1.25$ (N = 14), $1.25 < d_S < 1.5$ (N = 17) and $d_S > 1.5$ (N = 274). Young duplicated genes (*e.g.* $d_S < 1.0$) were comparatively scarce (N = 95 or ~22.5%) in this dataset, and in the Drosophila genome as a whole (Lynch and Conery, 2000; Conery and Lynch, 2001; Conant and Wagner, 2002; Gu *et al.*, 2002b; Thornton and Long, 2002). Similarly, d_A values were grouped into bins: $d_A < 0.1$ (N = 76), $0.1 < d_A < 0.2$ (N = 97), $0.2 < d_A < 0.3$ (N = 65), $0.3 < d_A < 0.4$ (57), $0.4 < d_A < 0.5$ (N = 43) and $d_A > 0.5$ (N = 86). It was assumed that gene conversion has not affected the estimation of genetic distances between gene duplicates.

A non-redundant set of 5 prime regions of *D. melanogaster* genes (set of single-copy genes) retrieved from http://www.fruitfly.org/seq_tools/datasets/Drosophila/promoter/ (Ohler *et al.*, 2002) was analyzed for comparison. The 5 prime regions of the duplicate genes and of the set of single-copy genes had similar GC contents (40.3 and 41.2%). Both datasets were screened for the presence of sequence elements known to occur in the *Drosophila* genome using the RepeatMasker software using the settings for insect genomes (http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker) (Thompson *et al.*, 1994), masked with "N", and excluded prior to alignment.

As done by Bergman and Kreitman (2001) the alignments of the 5 prime regions of each duplicate gene pair were done using the Dialign software (setting T = 1) (Morgenstern, 1999). For comparison, 5,000 alignments of randomly paired 5 prime regions drawn from the set of single-copy genes were done. Even if the Dialign alignment procedure may have its biases, as most procedures do, the comparison between the alignments of the 5 prime regions of duplicate genes and the alignments of randomly paired single copy genes should enable qualitative and quantitative statements regarding the significance of the sequence similarities observed in the 5 prime regions of the duplicate genes. Regions in the 5 prime regions that were aligned were converted as capital letters in the Fasta-formatted Dialign output. Aligned regions at least 10 nucleotides long were extracted and concatenated. The percentages of nucleotides that fell within such aligned regions was noted and referred to as 5 prime similarities (Table S1). Subsequently, for each alignment the number of perfectly matched base pairs within each aligned region was computed, leading to an estimate of sequence similarity within them (5 prime block similarity, Table S1). 5 prime block similarity values were transformed into a genetic distance ($d_{5'}$) using the HKY method (Hasegawa *et al.*, 1985) as implemented in PAML.

Distance estimation at high divergence levels can be associated with errors. Therefore, no attempt was made to resolve divergence times of $d_S > 1.5$. The estimation of very low synonymous divergence levels also can be associated with errors, in particular when the examined genes are short in length. To address this issue all 95 duplicate gene pairs with d_S up to 1 were re-analyzed to obtain estimates of $d_{5'}$ and $d_{5'}/ d_S$ that should be less likely to be affected by stochastic sampling. Specifically, first, sequences were extracted and aligned using the Dialign software. Second, Kimura's 2 parameter method was used to estimate $d_{5'}$, d_S, and d_A for each gene separately (Figure S1 and Table S2). Third, divergence times $d_{5'}$, d_S, and d_A were deduced from the concatenated sequences, the latter allows to obtain a weighted (by gene length) average of divergence times that should be less prone to stochastic sampling. For the concatenation process duplicate genes were grouped into the divergence bins $d_S < 0.1$ (N = 17), $0.1 < d_S < 0.2$ (N = 25), $0.2 < d_S < 0.3$ (N = 16), $0.3 < d_S < 0.4$ (N = 13), $0.4 < d_S < 0.5$ (N = 2), $0.5 < d_S < 0.6$ (N = 5), $0.6 < d_S < 0.7$ (N = 4), $0.7 < d_S < 0.8$ (N = 3), $0.8 < d_S < 0.9$ (N = 3), $0.9 < d_S < 1.0$ (N = 4).

Analysis of co-regulation of gene duplicates: Gene expression data from 267 Affymetrix GeneChips representing six independent investigations on *D. melanogaster* were retrieved from http://jbiol.com/content/supplementary/1475-4924-1-5-S1.txt (Spellman and Rubin, 2002). These dealt with embryo development, aging, DNA damage, immune response, and DDT resistance in adult flies and embryos subjected to 88 distinct conditions or experimental manipulations. For the description of the gene expression data and their analysis see Spellman and Rubin (2002). Here, Pearson's correlation coefficient (R) was computed across the expression levels provided by Spellman and Rubin (2002) to quantify the degree of co-regulation of duplicate genes. R was transformed using the expression ln((R+1)/R-1)) (Gu *et al.*, 2002b; Gu and Su 2007) and referred to as ln(R). The transformation of R into ln(R) enabled the analysis of sequence divergence and gene expression using linear regression (Gu *et al.*, 2002b). The expectation was that co-regulated duplicate genes would display high ln(R)-values when calculated over a series of conditions, because more similar regulatory regions should mediate more similar responses. For comparison, sampling with replacement from the expression profiles of the duplicate genes was done to yield 5,000 ln(R)-values computed between 10,000 randomly paired genes (Figure S2).

Results

Levels of 5 prime sequence similarities between duplicate genes: Alignments of the 5 prime non-coding regions of duplicate gene pairs resulted in a mosaic of aligned and non-aligned stretches of sequence. Only a percentage of sites in the 5 prime regions of duplicate genes fell within aligned stretches of sequence. Specifically, 5 prime similarities, a number that summarizes the percentage of nucleotide sites that fell within aligned stretches of sequence, were between 2 and 60% (median, 20.0%, mean 21.6%, 95% CI of mean, 20.4-22.1%) (Figure S3). 5 prime similarities were weakly correlated with synonymous and amino acid replacement site divergence between duplicate gene pairs (ANOVA, F_{ratio} 13.2, $R^2 = 0.06$, $p < 0.001$ and F_{ratio} 6.7, $R^2 = 0.03$, $p = 0.0014$, respectively).

The distribution of 5 prime block similarity values derived from the alignments of duplicate genes was compared to the distribution derived from 5,000 alignments of randomly paired genes (Figure 1). The expectation was that the 5 prime regions of randomly paired single-copy genes should reflect the degree to which DiAlign generated alignments between unrelated 5 prime regions of genes. For ~26% of random alignments no regions of any similarity were found that were 10 bp or longer. For a lower percentage 38/456 (~9%) of the duplicate gene pair dataset DiAlign could not identify such sequence blocks. These were excluded because they cannot be analyzed within a framework that considers per nucleotide site divergence rates. Their omission should have introduced a bias towards higher average levels of 5 prime block similarities among duplicate genes.

Levels of sequence similarity between the 5 prime regions of *Drosophila* duplicate genes exceeded those that

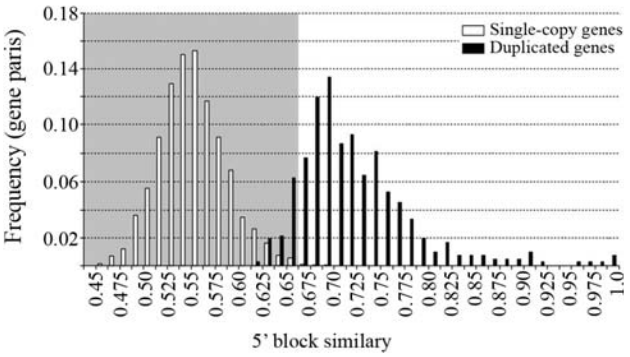

Figure 1 - The distribution of 5 prime block similarity values that were obtained from alignments of the 5 prime regions of duplicate gene pairs of the *D. melanogaster* genome (filled bars) and alignments of the 5 prime regions of randomly paired single copy *D. melanogaster* genes (open bars). The shaded area depicts the 99% range of 5 prime block similarity values obtained from alignments of the 5 prime region of randomly paired single copy genes.

were obtained from random alignments (Figure 1). Specifically, whereas the distribution of 5 prime block similarity scores that was based on alignments of randomly paired genes had a mean 5 prime block similarity of 0.553 (95% CI of mean: 0.552-0.555), mean 5 prime block similarity in alignments of the 5 prime regions of duplicate genes was 0.723 (95% CI: 0.717-0.729). The distribution of 5 prime block similarities between randomly paired single-copy genes was normally distributed, and was used to deduce the probability P to observe 5 prime block similarity values that were observed in the duplicate gene pairs after correction for multiple testing with the Bonferoni method. Forty-four duplicate gene pairs had average 5 prime block similarities that were not significant (< 0.67, n.s.), but 373 had average block similarities that exceeded random levels (= 0.67, p = 0.01). In total, average 5 prime block similarity between the 5 prime regions of duplicate genes was between 0.6 and 0.7 for 158 (34.6%) duplicate pairs, 0.7-0.8 for 216 (47.3%) pairs, 0.8-0.9 for 32 (7.0%) pairs, and 0.9-1.0 (2.4%) for 11 pairs. Thus, while DiAlign tended to find short sequence blocks even between 5 prime regions of random pairs of single copy genes, sequence similarity in the set of single-copy genes generally remained below those deduced from alignments of duplicated genes. For rate calculations below it was assumed that sequence similarities among the 5 prime regions of duplicated genes reflect sequence homology.

A contrast between $d_{5'}$ and d_S should enable inferences concerning the role of drift and selection on the evolution of the 5 prime regions of duplicated genes. Here it was found that $d_{5'}$ significantly increased with d_S (Figure 2A, $p < 0.0001$, $F_{Ratio} = 79.5$, $R^2 = 29\%$, ANOVA). This was less pronounced when $d_{5'}$ was related to d_A (Figure 2B, $p < 0.0001$, $F_{Ratio} = 24.6$, $R^2 = 11\%$, ANOVA). In addition, a decay of $d_{5'}/d_S$ as a function of d_S (Figure 2C, $p < 0.001$, $F_{Ratio} = 109.8$, $R^2 = 36$, ANOVA), and to a less systematic degree d_A (not shown, $p < 0.001$, $F_{Ratio} = 38.9$, $R^2 = 16\%$, ANOVA), was observed. Values for $d_{5'}/d_S$ larger than 1 were observed for nearly all, ~50%, and ~10% of duplicate pairs with $d_S < 0.1$, 0.1-0.25, and > 0.25-0.5, respectively. Duplicate gene pairs with $d_S < 0.25$ had mean and median $d_{5'}/d_S$ values exceeding 1. Thus, rapid rates of 5 prime block evolution close to 1, or exceeding 1, predominantly occurred in young duplicated genes, and these high rates were suggestive of relaxed constraint and/or positive selection. In contrast, the rate of 5 prime evolution observed slowed relative to that at synonymous sites, a pattern consistent with purifying selection and functional constraint. However, other homogenizing forces, such as gene conversion, should be considered as well.

To obtain $d_{5'}/d_S$ rates less likely affected by stochastic sampling of sites from individual gene pairs with low d_S, the sequences of duplicate gene pairs with $d_S < 1$ were concatenated in bins (c.f. Table 1). Bins with a weighted average of $d_S < 0.3$ displayed $d_{5'}/d_S$ ratios > 1 (Table 1). The corresponding average of d_A was 0.173 (Table 1). Thus, high rates of 5 prime sequence block evolution between young duplicate genes were not caused by the inclusion of a few genes with particularly high $d_{5'}/d_S$. The $d_{5'}/d_S$ ratios of

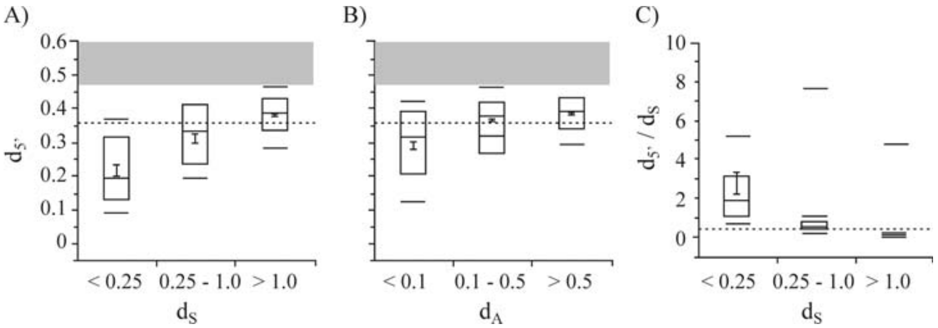

Figure 2 - The average genetic distance within aligned sequence blocks ($d_{5'}$) in relation to synonymous (A) and amino-acid site (B) divergence (d_S and d_A, respectively), and the evolutionary rate $d_{5'}/d_S$. The following quantiles are shown: 90%, 75%, mean (1 SE), median, 25%, and 10%. The means of the bins shown differ at $\alpha = 0.001$ (A) and $\alpha = 0.05$ (B) (Student's t-test). No further significant differences were found between any of the original bins given in the methods section. The grand mean is depicted by the dotted line. The shaded area represents the 99% range of values obtained from alignments of randomly paired single copy *D. melanogaster* genes (set of single-copy genes).

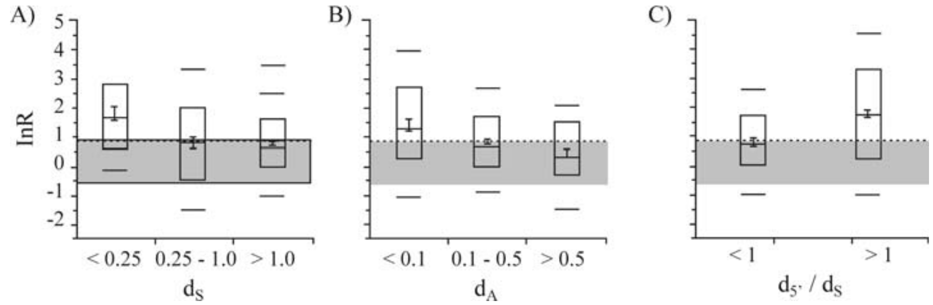

Figure 3 - The relationship between the correlation of gene expression, ln(R), between *D. melanogaster* duplicate gene pairs and their synonymous (A) and amino-acid site (B) divergence (d_S and d_A, respectively), and the evolutionary rate $d_{5'}/d_S$. Correlation of gene expression is expressed as the transformed Pearson's correlation coefficient over experimental conditions (see methods). The following quantiles are shown: 90%, 75%, mean (1 SE), median, 25%, and 10%. Only the means of the two bins $d_S < 0.25$ and $0.25 < d_S < 1$ differ at $\alpha = 0.0001$ (A) and $d_A < 0.1$ and $0.1 < d_A < 0.5$ differ at $\alpha = 0.0001$ (B) (Student's t-test). No further significant differences were found between any of the remaining bins given in the methods section. The grand mean is depicted by the dotted line. The shaded area represents the 99% range of ln(R) values obtained from a randomized dataset (c.f. Figure S2).

Table 1 - Divergence levels[1] at synonymous sites (S), in the 5 prime regions (5'), and at amino-acid replacement sites (A) sites, and the resulting rates $d_{5'}/d_S$ and d_A/d_S.

d_S bin[2]	S	5'	A	$d_S \pm$ SD	$d_{5'} \pm$ SD	$d_A \pm$ SD	$d_{5'}/d_S$	d_A/d_S
< 0.1	2252/122	5245/734	7371/384	0.057 ± 0.005	0.157 ± 0.006	0.054 ± 0.003	2.75	0.95
0.1-0.2	2972/397	6014/1067	9688/903	0.149 ± 0.008	0.207 ± 0.007	0.100 ± 0.003	1.39	0.67
0.2-0.3	2010/428	4459/1105	6402/974	0.258 ± 0.014	0.313 ± 0.011	0.173 ± 0.006	1.21	0.67
0.3-0.4	2974/782	2941/723	9965/1575	0.399 ± 0.014	0.310 ± 0.013	0.181 ± 0.005	0.78	0.45
0.4-0.5	143/46	513/68	346/90	0.452 ± 0.084	0.148 ± 0.019	0.334 ± 0.041	0.33	0.74
0.5-0.6	1134/417	1080/316	3648/698	0.562 ± 0.038	0.393 ± 0.026	0.227 ± 0.009	0.70	0.40
0.6-0.7	435/169	905/244	1321/320	0.622 ± 0.072	0.351 ± 0.026	0.304 ± 0.019	0.56	0.49
0.7-0.8	521/220	538/137	1598/589	0.739 ± 0.088	0.325 ± 0.032	0.564 ± 0.033	0.44	0.76
0.8-0.9	198/98	568/171	639/126	1.470 ± 0.659	0.409 ± 0.038	0.235 ± 0.023	0.28	0.16
0.9-1.0	328/158	845/201	911/185	1.152 ± 0.399	0.297 ± 0.024	0.244 ± 0.020	0.26	0.21

[1] The number of sites surveyed in base pairs (first number) and the number of divergent sites (second number).
[2] Divergence bins from Refs 1 and 2 and as described in methods.

more divergent duplicate gene pairs remained smaller than 1 in the concatenated data sets. The use of the concatenated sequences should provide conservative, *i.e.* lower, estimates for the rate $d_{5'}/d_S$. This was most notable in the divergence bin $d_S < 0.1$, where the average rate $d_{5'}/d_S$ computed as the mean over individual duplicate gene pairs yielded a value close to 7 (c.f. Table S2 for divergence estimates and rates derived from individual duplicate gene pairs). In contrast, when estimated from the concatenated sequences, a $d_{5'}/d_S$ value of 2.75 was obtained. Similarly, the rates $d_{5'}/d_S$ obtained for the remaining divergence bins were lower than the corresponding $d_{5'}/d_S$ values calculated as the mean over individual duplicate gene pairs. Taken together, the decay of $d_{5'}$ and $d_{5'}/d_S$ as function of d_S was suggestive of a phase of accelerated evolution in the 5 prime regions of young duplicated genes, *i.e.* those with d_S < 25%-30%. This was also true, qualitatively, when each individual duplicate gene pair was examined (Table S2).

Masked sequences in the 5 prime regions of duplicate genes: Besides nucleotide substitution, a range of possible other mutational events following gene duplication may alter the functionality of 5 prime regulatory sequences. These involve the insertion or deletion of various types of sequence elements (retro-elements and low-complexity/repeat sequences), or the insertion of the duplicate gene copies into regions that already were densely occupied by such sequence elements. As a proxy for the frequency of such events, the percentage of 5 prime sequence occupied by sequence elements that was recognized and masked by the RepeatMasker software was tabulated (Table 2). On average, only 4% of the total sequence data covering the 5 prime regions of duplicate genes were masked. A similar percentage (3.8%) was masked in the single-copy 5 prime regions, indicating that the majority of the duplicated 5 prime regions were not atypical with respect to such sequence elements when compared to 5 prime regions of single copy genes.

For about 10% of duplicate pairs masked sequences occupied as much as 18%-77% of the 5 prime region, indicating that larger-scale insertions or deletions of elements could affect the function of the 5 prime region. Simple repeats and low complexity-type sequences occupied the

Table 2 - The percentage of masked sequence occupied by various types of sequence elements and the percentages of the total sequence surveyed occupied by them in duplicate genes and the set of single-copy genes.

Element type*	Duplicate genes		Set of single-copy genes	
	Masked	Total	Masked	Total
LINEs	11.2%	0.4%	9.9%	0.4%
LTR elements	4.6%	0.2%	9.0%	0.3%
Gypsy-type:	4.2%	0.2%	2.0%	0.1%
PAO-type:	0.4%	-	7.0%	0.3%
DNA elements:	19.6%	0.8%	2.5%	0.1%
Tc1-type:	3.7%	0.1%	0.4%	-
Unclassified:	8.1%	0.3%	-	-
Satellites:	-	-	0.5%	-
Simple repeats:	19.4%	0.8%	21.6%	0.8%
Low complexity:	36.9%	1.5%	56.6%	2.1%
Total:		4.0%		3.8%

*SINEs, Copia, and small RNAs not found in either dataset.
- Not found in one of the two datasets.

largest percentage of the masked sequence (Table 2). There was a trend towards higher percentages of masked sequence (median > 10%) in comparatively young (d_S~25% or less) duplicate gene pairs when compared to the usually less than 5% masked sequence in duplicate gene pairs separated by d_S values > 25%. Perhaps, some of these repeats or low-complexity-type sequences are deleterious or form the basis for the evolution of motifs not recognized by RepeatMasker. However, the percentage of 5 prime sequence masked by RepeatMasker was not significantly related to d_S (not shown).

Evolution and expression of 5 prime sequences of duplicate genes: In yeast, young duplicate gene pairs tend to be more similar in their expression than are old duplicate pairs (Gu *et al.*, 2002a; Papp *et al.*, 2003; but see Wagner, 2000). Correlation of gene expression between duplicated genes might be a useful proxy for functional equivalence (Gu *et al.*, 2002c). Here, an analysis of the co-regulation of gene duplicates, as inferred from ln(R), showed that about 40% of *D. melanogaster* duplicate pairs were above the 99% range of randomly generated ln(R)-values (-0.79 and +0.99) and 10% were below that (c.f. Figure S1). Thus, half of the examined duplicate gene pairs conformed to random expectations. Co-regulation of duplicate genes may be 4 times more common than extreme divergence in regulation.

The correlation of expression of duplicate genes, expressed as ln(R), was found to decay as duplicate genes diverged at synonymous sites and at amino-acid replacement sites, but the relationships were weak. Specifically, a reduction of ln(R) between d_S < 0.25% (mean ln(R) = 1.8 ± 0.2, median 1.49) and d_S > 0.25 (mean ln(R) = 0.79 ± 0.1, median 0.64) was observed, but no further systematic trend was observed at higher divergence levels. Moreover, at high d_S the median and mean of ln(R) remained compatible with random expectations (Figure S1). However, expression was assayed over the whole fly, larvae, and embryos (Spellman and Rubin, 2002), such that only limited power would be expected to detect diversification of expression between gene duplicates *e.g.* at the level of tissues (c.f. Makova and Li, 2003). Overall, the *Drosophila* data fell in between the previously observed strong correlation between ln(R) and d_S (Gu *et al.*, 2002a) and a much weaker such relationship (Wagner, 2000), both observed in yeast. However, in this study emphasis was placed on the expression divergence after only 25 percent synonymous site divergence was observed, *i.e.* in young pairs of duplicated genes.

Duplicate gene pairs that diverged in their expression patterns displayed rapid rates of 5 prime sequence evolution. Specifically, gene pairs with $d_{5'}/d_S > 1$ displayed higher levels of correlation in gene expression than duplicate pairs with $d_{5'}/d_S < 1$ (Figure 3C, median ln(R) = 1.80, mean 1.64 ± 0.23 *vs.* median ln(R) = 0.69, mean 0.83 ± 0.07, respectively). More than 60% of the duplicate pairs with $d_{5'}/d_S > 1$ had ln(R) values that fell outside the random distribution of ln(R) values (Figure 3C). In contrast, none had ln(R) values that were below random levels.

5 prime block similarities between duplicate genes were not a good indicator for the co-regulation of expression. When $d_{5'}$ and ln(R) were grouped into those that were compatible with random expectations and those that were not, then one would have expected that random values of 5 prime block similarity values predominantly coincide with random ln(R) values (or *vice versa*). This was not the case. Only duplicate pairs with $d_{5'}$ higher than 0.8 differed from the remaining duplicate pairs in their correlation of expression (ln(R) > 1.4 *vs.* ln(R) < 0.9). However, for duplicate gene pairs with $d_{5'}$ exceeding 0.8 ln(R) values as low as -0.61 (c.f. Figure S1) were not uncommon, *e.g.* they were found in ~10% of the cases. Conversely, ln(R) values as high as 2.7 were found in ~10% of the gene pairs with $d_{5'}$ less than 0.8. Thus, while there was weak indication that $d_{5'}$ and ln(R) were dependent variables, the statistical resolution to document such a relationship was either limited or obscured by biological factors or the functional regulatory elements are located in regions that could not be aligned, and thus, $d_{5'}$ more closely approximates non-functional rates of evolution.

Discussion

The principle onto which ‘phylogenetic footprinting’ is based is that conservation between orthologous coding sequences reflects functional constraint (Fickett and Wasserman, 2000). Conservation between orthologous non-coding sequences also has been viewed as evidence for functional constraint (Tautz and Nigro, 1998; Bergman and Kreitman, 2001; Wasserman *et al.*, 2000; Bergman *et al.*, 2002; Webb *et al.*, 2002; Dermitzakis *et al.*, 2003; Haberer *et al.*, 2006, Thomas *et al.*, 2007). The possibility that negative selection on the 5 prime regions of genes may indeed be prevalent has been raised (Tautz and Nigro, 1998; Stone and Wray, 2001; Dermitzakis *et al.*, 2003; Hahn *et al.*, 2003; Kohn *et al.*, 2004; Andolfatto, 2005; Eyre-Walker, 2006; Hahn, 2007). More rapid rates of substitution take place in regions free of functional constraint (Andolfatto, 2005; Shapiro *et al.*, 2007). In the case of non-coding sequences rapid rates may be driven by nucleotide substitution, but also by mutational events (insertions, deletions, replication slippage) whose dynamics are not well understood (*e.g.* Comeron, 2001; Eyre-Walker, 2006). The dynamics of selective constraint on the 5 prime regions of *D. melanogaster* duplicate genes over time was manifest in the rate $d_{5'}/d_S$ (Figure 2A, and Table 1). Initially, for duplicate pairs separated by $d_S < 0.25$-0.3 $d_{5'}/d_S$ was larger than one. If it is assumed that d_S represents neutral divergence (Akashi, 1999), then $d_{5'}/d_S = 1$ indicate selective neutrality and $d_{5'}/d_S > 1$ positive selection. The majority of genes used here had low levels of codon usage bias (ENC 35 or more, Gu *et al.*, 2002b) and only 2% of genes had ENC levels between 32 and 35, suggesting that synonymous sites in this dataset should conform to neutrality reasonably well. Thus, as has been assumed by others here it was assumed that synonymous site divergence is useful measure for the relative ages of gene duplicate pairs (Kim and Yi, 2006; Wang *et al.*, 2006; Gu and Su, 2007; Guan *et al.*, 2007; Ha *et al.*, 2007; Jiang *et al.*, 2007; Johnston *et al.*, 2007; Roth *et al.*, 2007).

Duplicate gene pairs separated by $d_S > 0.25$-0.3 displayed lower $d_{5'}$ than d_S values (Figure 2, Table 1), *i.e.* d_A/d_S 1 (Table 2). Thus, levels of constraint on the 5 prime regions of duplicate genes were found to be comparable to those at amino-acid replacement sites once substantial coding sequence divergence levels have been reached. In contrast, young duplicate pairs may experience reduced levels of constraint on their amino-acid changes (Figure 2B and Table 1) (Clark, 1994; Lynch and Conery, 2000; Kondrashov *et al.*, 2002). The degree to which the 5 prime regions of ancient duplicate pairs, which are fully saturated at synonymous sites, still can be aligned is remarkable. In the absence of constraint neutral sites should be entirely diverged after a few million years, or at $d_S \sim 1$.

However, the constraint imposed on 5 prime regions that can directly be attributed to transcription control may be less than intuition would suggest. During a previous study this conclusion was based on the similar levels of sequence similarity that can be detected from alignments of 5 prime regions of orthologous *Drosophila* genes as well as alignments of introns of orthologous genes (Bergman and Kreitman, 2001). Here, the weak relationship between 5 prime block similarities between duplicates and their weak correlations with expression (Figure 3) indicated that the constraint detected here at best was in part a direct result of transcription requirements. This could reflect a limited resolution of this study. However, biological implications of this finding are plausible, as much remains to be learned about regulatory non-coding sequences (*e.g.* Comeron, 2001; Fessele *et al.*, 2002; Ludwig, 2002; Hahn *et al.*, 2003; Bird *et al.*, 2006). Additional forces, such as gene conversion tracts spanning regions that are not involved in regulation can maintain sequence similarity in the 5 prime regions of duplicate genes (Ohta, 1985; Basten and Ohta, 1992; King, 1998; Maside *et al.*, 2003).

It is noteworthy that various other types of sequence elements (retro-elements and low-complexity/repeat sequences) located in the 5 prime regions of *D. melanogaster* duplicates became increasingly rare as duplicate genes diverged. Even though this was not further investigated here, the pattern pointed to their reduction over time. In human, repeat sequences occasionally have been linked to deleterious effects when located in the regulatory region of genes (Usdin and Grabczyk, 2000). Many types of low-complexity/repeat sequences may act as spurious transcription factor binding sites that are slightly deleterious (Stone and Wray, 2001).

The important assertion made in this report refers to the accelerated evolution in the 5 prime regions of young duplicates. The interpretation of the $d_{5'}/d_S$ rates relies on the premise that $d_{5'}$ and d_S of duplicate genes may be directly compared to one another, which may be questioned on a number of grounds. Most importantly, while it is quite certain that homologous sites in the coding regions of duplicate genes were compared, the possibility remains that non-homologous sites in the 5 prime regions of duplicate genes were compared. However, both the alignment and divergence estimation generally should be less problematic in young duplicate pairs compared to the alignment of old duplicate gene pairs. In fact, accelerated evolution in the 5 prime regions of young duplicate gene pairs was deduced from generally longer and more reliable alignments than those alignments of ancient duplicate pairs from which constraint was inferred.

The rapid divergence in the 5 prime regions of young *D. melanogaster* duplicates was found to coincide with their divergence at amino-acid replacement sites and low correlations of expression, as was expressed as ln(R) (Figure 3). To some degree this may reflect the functional diversification of duplicates with time. Data from yeast indicate that d_A and correlation of gene expression reflect functional equivalency of duplicates *in vivo* (Gu *et al.*, 2002b; Gu *et al.*, 2002c). In humans (Makova and Li, 2003) and in yeast duplicate gene expression patterns diverge rapidly (Gu *et al.*, 2002a; Papp *et al.*, 2003 but see Wagner, 2000). Data from orthologous genes now available from the newly released multiple Drosophila genome projects could be used next to assess whether the 5 prime regions of one or both copies of duplicate genes display accelerated evolution. This could help distinguishing between neo-functionalization (one copy accelerated) and sub-functionalization models (both copies accelerated), and to polarize the direction of change.

The possibility that advantageous mutations occur and positive selection acts on duplicate gene promoters has been raised before (Papp *et al.*, 2003; Seoighe *et al.*, 2003; Castillo-Davis *et al.*, 2004; Huminiecki and Wolfe, 2004; Jordan *et al.*, 2004; Lynch and Katju, 2004; He and Zhang, 2005; Crow and Wagner, 2006; Kim and Yi, 2006; Kondrashov and Kondrashov, 2006; Gu and Su, 2007; Jiang *et al.*, 2007; Johnston *et al.*, 2007). Here a pattern consistent with selection in Drosophila was observed. Complex selection patterns (Ohta, 1988; Basten and Ohta, 1992; Force *et al.*, 1999; Ludwig *et al.*, 2000; Lynch and Force, 2000; Tautz, 2000; Ludwig, 2002; Wagner, 2002a) and the diffuse link between sequence context and regulatory function (Carroll *et al.*, 2001; Fessele *et al.*, 2002) pose considerable challenges to the conclusive documentation of selection. However, the results presented here suggest that rapid evolution in the 5 prime regulatory regions of young duplicate genes, which tend to be rather equivalent in their function, appears to be a part of the footprint left by functional diversification. That positive selection is driving rates $d_{5'}/d_S$ in excess of 1 is conceivable when assuming that single nucleotide substitutions within 5 prime blocks are their major mode of change.

In sum, the 5 prime regulatory regions of very young *Drosophila* duplicate gene pairs diverge at rates faster than at synonymous sites. If the latter are viewed as a proxy for neutral divergence rates, then we can infer that the evolution of 5 prime sequences in young duplicate genes is driven by positive selection. Conceivably the process is facilitated by initial relaxation of selective constraint due to the overlapping functions of young duplicate pairs. Low levels of nonsynonymous site divergence and an analysis of *Drosophila* duplicate gene expression data presented supported functional redundancy of young gene duplicates. In contrast, as duplicate genes diverge over time in their coding sequences and expression patterns the 5 prime regulatory regions of them were found to display divergence rates as low as those at amino-acid replacement sites, suggesting that they evolve under selective constraint. An important next step in the analysis of duplicated gene evolution in *Drosophila* would be concerned with the symmetric, or asymmetric divergence of duplicate genes, which appears to be commonly seen in other organisms (Casneuf *et al.*, 2006; Chung *et al.*, 2006; Kim and Yi,2006; Tirosh and Barkai, 2007).

Acknowledgments

I thank Chung-I Wu for stimulating discussions and financial support, and Kevin Thornton, Casey Bergman, Matthias Gerberding, Sebastian Zöllner, Bettina Harr, and Jian Lu and two anonymous reviewers for important feedback.

References

Akashi H (1995) Inferring weak selection from patterns of polymorphism and divergence at 'silent' sites in *Drosophila* DNA. Genetics 139:1067-1076.

Akashi H (1999) Within- and between-species DNA sequence variation and the 'footprint' of natural selection. Gene 238:39-51.

Andolfatto P (2005) Adaptive evolution of non-coding DNA in *Drosophila*. Nature 437:1149-1152.

Barrier M, Robichaux RH and Purugganan MD (2001) Accelerated regulatory gene evolution in an adaptive radiation. Proc Natl Acad Sci USA 98:10208-10213.

Basten CJ and Ohta T (1992) Simulation study of a multigene family, with special reference to the evolution of compensatory advantageous mutations. Genetics 132:247-252.

Bergman CM and Kreitman M (2001) Analysis of conserved noncoding DNA in Drosophila reveals similar constraints in intergenic and intronic sequences. Genome Res 11:1335-1345.

Bergman CM, Pfeiffer BD, Rincon-Limas DE, Hoskins RA, Gnirke A, Mungall CJ, Wang AM, Kronmiller B, Pacleb J, Park S, *et al.* (2002) Assessing the impact of comparative genomic sequence data on the functional annotation of the *Drosophila* genome. Genome Biol 3:0086.1-0086.20.

Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM and Eisen, MB (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. Proc Natl Acad Sci USA 99:757-762.

Bird CP, Stranger BE and Dermitzakis ET (2006) Functional variation and evolution of non-coding DNA. Curr Opin Genet Dev 16:559-564.

Carroll SB, Grenier JK and Weatherbee SD (2001) From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design. Blackwell Science, Malden, 192 pp.

Casneuf T, De Bodt S, Raes J, Maere S and Van de Peer Y (2006) Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant *Arabidopsis thaliana*. Genome Biol 7:R13.

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

Castresana J (2002) Estimation of genetic distances from human and mouse introns. Genome Biol 3:0028.1-0028.7.

Chung WY, Albert R, Albert I, Nekrutenko A and Makova KD (2006) Rapid and asymmetric divergence of duplicate genes in the human gene coexpression network. Bmc Bioinfo 7:46.

Clark AG (1994) Invasion and maintenance of a gene duplication. Proc Natl Acad Sci USA 91:2950-2954.

Comeron JM (2001) What controls the length of noncoding DNA. Curr Opin Genet Dev 11:652-659.

Conant GC and Wagner A (2002) Genome History - A software tool and its application to fully sequenced genomes. Nucleic Acids Res 30:3378-3386.

Conery JS and Lynch M (2001) Nucleotide substitutions and the evolution of duplicate genes. Pacific Symp Biocomp 6:167-178.

Crow KD and Wagner GP (2006) What is the role of genome duplication in the evolution of complexity and diversity? Mol Biol Evol 23:887-892.

de Meaux J (2006) An adaptive path through jungle DNA. Nat Genet 38:506-507.

Dermitzakis ET, Bergman CM and Clark AG (2003) Tracing the evolutionary history of *Drosophila* regulatory regions with models that identify transcription factor binding sites. Mol Biol Evol 20:703-714.

Eyre-Walker A (2006) The genomic rate of adaptive evolution. Trends Ecol Evol 21:569-575.

Fessele S, Maier H, Zischek C, Nelson PJ and Werner T (2002) Regulatory context is crucial part of gene function. Trends Genet 18:60-63.

Fickett JW and Wasserman WW (2000) Discovery and modeling of transcriptional regulatory regions. Curr Opin Biotechnol 11:19-24.

Force A, Lynch M, Pickett FB, Amores A, Yan Y-L and Postlethwait J (1999) Preservation of duplicate genes by complementary degenerative mutations. Genetics 151:1531-1545.

Gu Z, Nicolae D, Henry H-S and Li W-H (2002a) Rapid divergence in expression between duplicate genes inferred from microarray data. Trends Genet 18:609-613.

Gu Z, Cavalcanti A, Chen F-C, Bouman P and Li W-H (2002b) Extend of gene duplication in the genomes of *Drosophila*, nematode and yeast. Mol Biol Evol 19:256-262.

Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW and Li W-H (2002c) Role of duplicate genes in genetic robustness against null mutations. Nature 421:63-66.

Gu X and Su ZX (2007) Tissue-driven hypothesis of genomic evolution and sequence-expression correlations. Proc Natl Acad Sci USA 104:2779-2784.

Guan YF, Dunham MJ and Troyanskaya OG (2007) Functional analysis of gene duplications in *Saccharomyces cerevisiae*. Genetics 175:933-943.

Ha M, Li WH and Chen ZJ (2007) External factors accelerate expression divergence between duplicate genes. Trends Genet 23:162-166.

Haberer G, Hindemitt T, Meyers BC and Mayer KFX (2004) Transcriptional similarities, dissimilarities, and conservation of *cis*-elements in duplicated genes of arabidopsis. Plant Physiol 136:3009-3022.

Haberer G, Mader MT, Kosarev P, Spannagl M, Yang L and Mayer KFX (2006) Large-scale *cis*-element detection by analysis of correlated expression and sequence conservation between arabidopsis and *Brassica oleracea*. Plant Physiol 142:1589-1602.

Hahn MW, Stajich JE and Wray GA (2003) The effects of selection against spurious transcription factor binding sites. Mol Biol Evol 20:901-906.

Hahn MW (2007) Detecting natural selection on *cis*-regulatory DNA. Genetica 129:7-18.

Hasegawa M, Kishino H and Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160-174.

He XL and Zhang JZ (2005) Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics 169:1157-1164.

Hughes AL (1994) The evolution of functionally novel proteins after gene duplication. Proc R Soc London Soc Ser B 256:119-124.

Hughes AL (2002) Adaptive evolution after gene duplication. Trends Ecol Evol 18:433-434.

Huminiecki L and Wolfe KH (2004) Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. Genome Res 14:1870-1879.

Jiang, HF, Liu DY, Gu ZL and Wang W (2007) Rapid evolution in a pair of recent duplicate segments of rice. J Exp Zool 308B:50-57.

Johnston CR, O'Dushlaine C, Fitzpatrick DA, Edwards RJ and Shields DC (2007) Evaluation of whether accelerated protein evolution in chordates has occurred before, after, or simultaneously with gene duplication. Mol Biol Evol 24:315-323.

Jordan IK, Marino-Ramirez L and Koonin EV (2005) Evolutionary significance of gene expression divergence. Gene 345:119-126.

Kim SH and Yi SV (2006) Correlated asymmetry of sequence and functional divergence between duplicate proteins of *Saccharomyces cerevisiae*. Mol Biol Evol 23:1068-1075.

King LM (1998) The role of gene conversion in determining sequence variation and divergence in the *Est-5* gene family in *Drosophila pseudoobscura*. Genetics 148:305-316.

Kohn MH, Fang S and Wu C-I (2004) Inference of positive and negative selection on the 5 regulatory regions of *Drosophila* genes. Mol Biol Evol 21:374-383.

Kondrashov FA, Rogozin IB, Wolf Y I and Koonin EV (2002) Selection in the evolution of gene duplications. Genome Biol 3:0008.1-0008.9.

Kondrashov AS (2005) Evolutionary biology - Fruitfly genome is not junk. Nature 437:1106.

Kondrashov FA and Kondrashov AS (2006) Role of selection in fixation of gene duplications. J Theor Biol 239:141-151.

Koop BF (1995) Human and rodent DNA sequence comparisons: A mosaic model of genomic evolution. Trends Genet 11:367-371.

Li HP and Stephan W (2006) Inferring the demographic history and rate of adaptive substitution in *Drosophila*. PLoS Genetics 2:1580-1589.

Ludwig MZ, Bergman C, Patel NH and Kreitman M (2000) Evidence for stabilizing selection in a eukaryotic enhancer element. Nature 403:564-567.

Ludwig MZ (2002) Functional evolution of noncoding DNA. Curr Opin Genet Dev 12:634-639.

Lynch M and Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151-1155.

Lynch M and Force A (2000) The probability of duplicate gene preservation by subfunctionalization. Genetics 154:459-473.

Lynch M and Katju V (2004) The altered evolutionary trajectories of gene duplicates. Trends Genet 20:544-549.

Makova KD and Li W-H (2003) Divergence in the spatial pattern of gene expression between human duplicate genes. Genome Res 13:1638-1645.

Maside X, Bartolome C and Charlesworth B (2003) Inferences on the evolutionary history of the S-element family of *Drosophila melanogaster*. Mol Biol Evol 20:1183-1187.

Morgenstern B (1999) DIALIGN 2: Improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics 15:211-218.

Ohler U, Liao G-C, Niemann H and Rubin GM (2002) Computational analysis of core promoters in the *Drosophila* genome. Genome Biol 3:0087.1-0087.12.

Ohno S (1970) Evolution by Gene Duplication. George Allen and Unwin, London, 160 pp.

Ohta T (1985) A model of duplicative transposition and gene conversion for repetitive DNA families. Genetics 110:513-524.

Ohta T (1987) Simulating the evolution of gene duplication. Genetics 115:207-213.

Ohta T (1988) Evolution by gene duplication and compensatory advantegous mutations. Genetics 120:841-847.

Ohta T (1994) Further examples of evolution by gene duplication revealed through DNA sequence comparisons. Genetics 138:1331-1337.

Papp B, Pal C and Hurst LD (2003) Evolution of *cis*-regulatory elements in duplicated genes of yeast. Trends Genet 19:417-422.

Roth C, Rastogi S, Arvestad L, Dittmar K, Light S, Ekman D and Liberles DA (2007) Evolution after gene duplication: Models, mechanisms, sequences, systems, and organisms. J Exp Zool Part 308B:58-73.

Seoighe C, Johnston CR and Shields DC (2003) Significantly different patterns of amino acid replacement after gene duplication as compared to after speciation. Mol Biol Evol 20:484-490.

Shabalina SA, Ogurtsov AY, Kondrashov VA an Kondrashov AS (2001) Selective constraint in intergenic regions of human and mouse genomes. Trends Genet 17:373-376.

Shapiro JA, Huang W, Zhang CH, Hubisz MJ, Lu J, Turissini DA, Fang S, Wang HY, Hudson RR, Nielsen R, *et al.* (2007) Adaptive genic evolution in the *Drosophila* genomes. Proc Natl Acad Sci USA 104:2271-2276.

Spellman PT and Rubin GM (2002) Evidence for large domains of similarly expressed genes in the *Drosophila* genome. J Biol 1:5.

Stone JR and Wray GA (2001) Rapid evolution of *cis*-regulatory sequences via local point mutations. Mol Biol Evol 18:1764-1770.

Tautz D (2000) Evolution of transcriptional regulation. Curr Opin Genet Dev 10:575-579.

Tautz D and Nigro L (1998) Microevolutionary divergence pattern of the segmentation gene hunchback in *Drosophila*. Mol Biol Evol 15:1403-1411.

Taylor JS and Raes J (2004) Duplication and divergence: The evolution of new genes and old ideas. Annu Rev Genet 38:615-643.

Thomas BC, Rapaka L, Lyons E, Pedersen B and Freeling M (2007) *Arabidopsis* intragenomic conserved noncoding sequences. Proc Natl Acad Sci USA 104:3348-3353.

Thompson JD, Higgins DG and Gibson TJ (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673-4680.

Thornton K and Long M (2002) Rapid divergence of gene duplicates on the *Drosophila melanogaster* X chromosome. Mol Biol Evol 19:918-925.

Usdin K and Grabczyk E (2000) DNA repeat expansions and human disease. Cell Mol Life Sci 57:914-931.

Wagner A (2000) Decoupled evolution of coding region and mRNA expression patterns after gene duplication: Implications for the neutralist-selectionist debate. Proc Natl Acad Sci USA 97:6579-6584.

Wagner A (2001) Birth and death of duplicated genes in completely sequenced eukaryotes. Trends Genet 17:237-239.

Wagner A (2002a) Selection and gene duplication: A view from the genome. Genome Biol 3:1012.1-1012.3.

Wagner A (2002b) Asymmetric functional divergence of duplicate genes in yeast. Mol Biol Evol 19:1760-1768.

Walsh JB (1995) How often do duplicated genes evolve new functions? Genetics 139:421-428.

Wang R, Chong K and Wang T (2006) Divergence in spatial expression patterns and in response to stimuli of tandem-repeat paralogues encoding a novel class of proline-rich proteins in *Oryza sativa*. J Exp Bot 57:2887-2897.

Wasserman WW, Palumbo M, Thompson W and Fickett JW (2000) Human-mouse genome comparisons to locate regulatory sites. Nat Genet 26:225-228.

Webb CT, Shabalina SA, Ogurtsov AY and Kondrashov AS (2002) Analysis of similarity within 142 pairs of orthologous intergenic regions of *Caenorhabditis elegans* and *Caenorhabditis briggsae*. Nucleic Acids Res 30:1223-1229.

Yang Z (1997) PAML: A program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555-556.

Zhang J, Rosenberg HF and Nei M (1998) Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc Natl Acad Sci USA 95:3708-3713.

Zhang J (2003) Evolution by gene duplication: An update. Trends Ecol Evol 18:292-298.

Internet Resources

Web page that contains the information and analysis methods used for the paper on the evolution of duplicate genes by Lynch and Conery (2000), http://www.csi.uoregon.edu/projects/genetics/duplications/D.melanogaster.txt (December 2002).

Berkley Drosophila Genome Project (BDGP, Release 2) (http://www.fruitfly.org) (December 2002).

A non-redundant set of the 5 prime regions of *D. melanogaster* genes, http://www.fruitfly.org/seq_tools/datasets/Drosophila/promoter/ (December 2002) (Ohler *et al.*, (2002).

RepeatMasker software, (http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker) (Thompson *et al.*, 1994).

Gene expression data from 267 Affymetrix GeneChips representing six independent investigations on *D. melanogaster* compiled by Spellman and Rubin (2002), http://jbiol.com/content/supplementary/1475-4924-1-5-S1.txt (December 2002).

Associate Editor: Louis Bernard Klaczko

Table S1. Listing of 418 gene duplicate pairs for which alignments of their 5 prime regions were analyzed. Given are, from left to right, Gene IDs #1 and 2, the length of the alignment in base pairs (bp), 5 prime similarity, the number of matched up nucleotides in the aligned segment, and resulting 5 prime block similarity (see methods).

ID1	ID2	Alignment length	5 prime similarity	Aligned nucleotides	5 prime block similarity
CG7027	CG17355	125	0.13	114	0.91
CG8825	CG8826	220	0.22	167	0.76
CG18259	CG6961	441	0.44	423	0.96
CG10102	CG12505	202	0.2	140	0.69
CG4137	CG15190	523	0.52	508	0.97
CG4216	CG7271	241	0.24	206	0.85
CG7952	CG4575	267	0.27	186	0.7
CG3678	CG17556	350	0.35	285	0.81
CG15797	CG15910	97	0.1	72	0.74
CG1742	CG12628	152	0.15	128	0.84
CG12700	CG11941	576	0.58	525	0.91
CG4479	CG4478	421	0.42	383	0.91
CG18606	CG10476	310	0.31	266	0.86
CG6343	CG5585	167	0.17	133	0.8
CG6891	CG6900	16	0.02	13	0.81
CG13060	CG13041	582	0.58	507	0.87
CG17438	CG17441	327	0.33	276	0.84
CG15600	CG8565	150	0.15	114	0.76
CG13402	CG18313	274	0.27	242	0.88
CG1524	CG1527	294	0.29	226	0.77
CG7045	CG7046	182	0.18	161	0.88
CG13402	CG18157	227	0.23	197	0.87
CG9111	CG9118	223	0.22	166	0.74
CG18078	CG3875	98	0.1	73	0.74
CG18281	CG17637	380	0.38	338	0.89
CG11719	CG18396	253	0.25	181	0.72
CG18157	CG9404	276	0.28	253	0.92
CG5334	CG5347	395	0.4	321	0.81
CG13402	CG9404	236	0.24	202	0.86
CG18372	CG10146	402	0.4	290	0.72
CG17597	CG17320	290	0.29	209	0.72
CG18157	CG18313	317	0.32	313	0.99
CG12493	CG10630	324	0.32	275	0.85
CG1740	CG10174	184	0.18	136	0.74
CG3895	CG18414	45	0.05	32	0.71
CG18078	CG3927	129	0.13	99	0.77
CG9404	CG18313	368	0.37	332	0.9
CG18078	CG3584	124	0.12	94	0.76
CG13069	CG13051	478	0.48	396	0.83
CG1365	CG1367	254	0.25	186	0.73
CG7370	CG17812	28	0.03	25	0.89
CG15577	CG15578	342	0.34	244	0.71

CG9906	CG11958	62	0.06	45	0.73
CG11941	CG11942	446	0.45	361	0.81
CG9902	CG7692	312	0.31	257	0.82
CG6999	CG10993	362	0.36	282	0.78
CG12700	CG11942	450	0.45	357	0.79
CG6997	CG10993	318	0.32	260	0.82
CG12359	CG11023	191	0.19	128	0.67
CG1924	CG11958	133	0.13	92	0.69
CG12224	CG3397	386	0.39	293	0.76
CG6997	CG6999	299	0.3	238	0.8
CG18078	CG4021	153	0.15	113	0.74
CG15645	CG13732	336	0.34	250	0.74
CG14850	CG14851	328	0.33	227	0.69
CG1924	CG9906	191	0.19	137	0.72
CG9111	CG1180	172	0.17	137	0.8
CG14213	CG9573	270	0.27	198	0.73
CG14500	CG14499	311	0.31	237	0.76
CG18106	CG18108	279	0.28	212	0.76
CG9111	CG1179	166	0.17	136	0.82
CG8066	CG8050	108	0.11	108	1
CG15646	CG12708	263	0.26	171	0.65
CG4960	CG8331	185	0.19	139	0.75
CG11386	CG10944	223	0.22	146	0.65
CG9111	CG1165	122	0.12	83	0.68
CG13794	CG13795	265	0.27	182	0.69
CG9046	CG9271	240	0.24	164	0.68
CG13324	CG13323	210	0.21	155	0.74
CG15332	CG15333	82	0.08	62	0.76
CG11520	CG10810	288	0.29	192	0.67
CG8219	CG7398	290	0.29	198	0.68
CG12477	CG7184	272	0.27	196	0.72
CG12699	CG18469	40	0.04	40	1
CG8628	CG15829	161	0.16	117	0.73
CG8628	CG8629	220	0.22	158	0.72
CG18107	CG18108	302	0.3	302	1
CG1365	CG1373	217	0.22	162	0.75
CG2694	CG11322	285	0.29	190	0.67
CG9906	CG11235	147	0.15	100	0.68
CG9046	CG9048	211	0.21	143	0.68
CG1365	CG1878	195	0.2	156	0.8
CG10811	CG10810	205	0.21	140	0.68
CG13617	CG12842	248	0.25	178	0.72
CG15332	CG18293	154	0.15	106	0.69
CG10813	CG10812	240	0.24	165	0.69
CG13063	CG13043	191	0.19	154	0.81
CG14850	CG8087	298	0.3	217	0.73
CG13589	CG13590	192	0.19	157	0.82
CG11444	CG4438	238	0.24	177	0.74
CG6857	CG6881	164	0.16	112	0.68

CG6246	CG12287	62	0.06	48	0.77
CG13793	CG13794	603	0.6	472	0.78
CG1367	CG1878	250	0.25	191	0.76
CG17760	CG17759	116	0.12	89	0.77
CG17734	CG11825	259	0.26	176	0.68
CG13793	CG13795	256	0.26	170	0.66
CG10813	CG10810	293	0.29	206	0.7
CG2043	CG2044	284	0.28	218	0.77
CG1319	CG4205	153	0.15	105	0.69
CG8989	CG5825	138	0.14	99	0.72
CG5767	CG5770	350	0.35	255	0.73
CG18372	CG4740	191	0.19	138	0.72
CG14534	CG14254	214	0.21	153	0.71
CG2555	CG6956	151	0.15	111	0.74
CG8893	CG12055	187	0.19	127	0.68
CG10812	CG11520	224	0.22	175	0.78
CG15375	CG14708	223	0.22	157	0.7
CG4094	CG4095	158	0.16	107	0.68
CG11350	CG13705	174	0.17	131	0.75
CG4099	CG8856	99	0.1	74	0.75
CG13706	CG13705	195	0.2	133	0.68
CG11314	CG11315	313	0.31	211	0.67
CG4787	CG5422	214	0.21	151	0.71
CG7350	CG16931	401	0.4	313	0.78
CG6152	CG6145	307	0.31	210	0.68
CG7224	CG15283	179	0.18	134	0.75
CG4717	CG4761	307	0.31	203	0.66
CG4717	CG18455	188	0.19	129	0.69
CG1984	CG1980	275	0.28	192	0.7
CG4099	CG3212	169	0.17	107	0.63
CG18087	CG6132	432	0.43	340	0.79
CG3367	CG9194	174	0.17	117	0.67
CG16827	CG8095	202	0.2	129	0.64
CG14666	CG10090	84	0.08	60	0.71
CG1946	CG1942	192	0.19	132	0.69
CG6045	CG18516	138	0.14	101	0.73
CG1701	CG11113	177	0.18	134	0.76
CG7994	CG1056	178	0.18	128	0.72
CG6131	CG15884	58	0.06	44	0.76
CG3212	CG8856	138	0.14	100	0.72
CG18324	CG18327	352	0.35	248	0.7
CG13079	CG10363	135	0.14	105	0.78
CG12254	CG13609	191	0.19	136	0.71
CG9223	CG9908	198	0.2	143	0.72
CG8091	CG5370	166	0.17	121	0.73
CG18538	CG18537	367	0.37	281	0.77
CG6733	CG6738	349	0.35	271	0.78
CG1635	CG1774	240	0.24	160	0.67
CG4035	CG1442	169	0.17	120	0.71

CG14851	CG8087	212	0.21	158	0.75
CG2984	CG1906	138	0.14	103	0.75
CG10254	CG2013	276	0.28	190	0.69
CG17300	CG8189	277	0.28	186	0.67
CG10813	CG11520	224	0.22	165	0.74
CG1878	CG1373	292	0.29	202	0.69
CG14304	CG14301	180	0.18	120	0.67
CG7311	CG8256	268	0.27	181	0.68
CG2952	CG8193	196	0.2	138	0.7
CG15587	CG4791	122	0.12	88	0.72
CG9847	CG14715	83	0.08	64	0.77
CG6062	CG13289	208	0.21	154	0.74
CG14910	CG13300	178	0.18	125	0.7
CG9656	CG3978	219	0.22	158	0.72
CG8922	CG7014	285	0.29	196	0.69
CG7594	CG7599	315	0.32	218	0.69
CG11267	CG9920	157	0.16	105	0.67
CG16983	CG12700	168	0.17	113	0.67
CG4095	CG6140	137	0.14	88	0.64
CG5097	CG4312	196	0.2	145	0.74
CG12562	CG4988	160	0.16	115	0.72
CG1695	CG1702	191	0.19	147	0.77
CG8168	CG5661	159	0.16	115	0.72
CG5372	CG8095	82	0.08	68	0.83
CG3396	CG8337	156	0.16	121	0.78
CG11392	CG1442	110	0.11	76	0.69
CG7198	CG6154	170	0.17	130	0.76
CG12994	CG17193	318	0.32	215	0.68
CG8541	CG8543	206	0.21	132	0.64
CG17109	CG6733	312	0.31	242	0.78
CG16983	CG11941	152	0.15	96	0.63
CG9336	CG9338	161	0.16	120	0.75
CG1942	CG1941	259	0.26	180	0.69
CG1689	CG1379	227	0.23	167	0.74
CG9555	CG17906	323	0.32	238	0.74
CG17860	CG12348	215	0.22	141	0.66
CG5984	CG3937	208	0.21	144	0.69
CG4686	CG2996	57	0.06	43	0.75
CG7486	CG7788	161	0.16	115	0.71
CG5925	CG5887	160	0.16	118	0.74
CG3739	CG11626	244	0.24	180	0.74
CG15397	CG4341	204	0.2	148	0.73
CG2043	CG11650	197	0.2	125	0.63
CG3719	CG3315	76	0.08	59	0.78
CG1065	CG6255	205	0.21	135	0.66
CG6742	CG16728	222	0.22	158	0.71
CG15354	CG15630	176	0.18	130	0.74
CG9090	CG4994	197	0.2	152	0.77
CG14355	CG3199	291	0.29	197	0.68

CG16983	CG11942	165	0.17	115	0.7
CG5836	CG10384	219	0.22	155	0.71
CG5334	CG7184	262	0.26	180	0.69
CG1262	CG7489	223	0.22	154	0.69
CG13539	CG12637	267	0.27	185	0.69
CG6640	CG6645	182	0.18	138	0.76
CG9396	CG9399	194	0.19	144	0.74
CG7465	CG11352	232	0.23	172	0.74
CG9032	CG12810	161	0.16	118	0.73
CG9855	CG17991	157	0.16	107	0.68
CG9129	CG9130	291	0.29	212	0.73
CG14570	CG14568	187	0.19	134	0.72
CG15286	CG11984	242	0.24	173	0.71
CG5822	CG6268	159	0.16	107	0.67
CG10146	CG4740	185	0.19	132	0.71
CG7362	CG7069	222	0.22	153	0.69
CG6737	CG9013	105	0.11	81	0.77
CG5095	CG11584	209	0.21	145	0.69
CG18540	CG18539	350	0.35	271	0.77
CG5347	CG12477	200	0.2	146	0.73
CG5372	CG16827	169	0.17	128	0.76
CG18174	CG14884	167	0.17	116	0.69
CG16775	CG5506	133	0.13	101	0.76
CG10037	CG12287	298	0.3	196	0.66
CG13547	CG13973	259	0.26	183	0.71
CG14156	CG11742	133	0.13	101	0.76
CG3367	CG8713	249	0.25	163	0.65
CG13651	CG11849	179	0.18	131	0.73
CG1534	CG12334	207	0.21	153	0.74
CG4094	CG6140	171	0.17	119	0.7
CG7216	CG7214	187	0.19	125	0.67
CG17916	CG11735	166	0.17	126	0.76
CG6627	CG5191	176	0.18	116	0.66
CG13803	CG8985	196	0.2	137	0.7
CG4766	CG4746	193	0.19	131	0.68
CG1058	CG8546	94	0.09	72	0.77
CG5334	CG12477	362	0.36	237	0.65
CG6347	CG11459	275	0.28	176	0.64
CG5804	CG15829	190	0.19	141	0.74
CG13309	CG13308	274	0.27	195	0.71
CG5304	CG9254	51	0.05	38	0.75
CG18662	CG7933	154	0.15	112	0.73
CG15861	CG6124	95	0.1	61	0.64
CG9354	CG6090	276	0.28	193	0.7
CG10249	CG5841	145	0.15	111	0.77
CG9361	CG9194	200	0.2	136	0.68
CG7486	CG8091	177	0.18	129	0.73
CG3874	CG9620	201	0.2	144	0.72
CG15171	CG10751	139	0.14	94	0.68

CG8931	CG5755	204	0.2	142	0.7
CG17027	CG17026	59	0.06	47	0.8
CG13112	CG14741	209	0.21	148	0.71
CG8261	CG15844	182	0.18	129	0.71
CG6342	CG4900	272	0.27	193	0.71
CG9059	CG11319	223	0.22	139	0.62
CG12526	CG11735	200	0.2	138	0.69
CG13706	CG13703	145	0.15	102	0.7
CG10142	CG17988	196	0.2	134	0.68
CG3734	CG18493	331	0.33	221	0.67
CG9953	CG3739	286	0.29	200	0.7
CG14777	CG11077	222	0.22	141	0.64
CG12379	CG8206	157	0.16	118	0.75
CG9240	CG11110	254	0.25	168	0.66
CG8056	CG10334	215	0.22	143	0.67
CG14206	CG12275	137	0.14	107	0.78
CG3765	CG14940	192	0.19	143	0.74
CG7054	CG5430	178	0.18	128	0.72
CG15380	CG5308	169	0.17	118	0.7
CG14840	CG6301	94	0.09	63	0.67
CG12895	CG14757	289	0.29	188	0.65
CG3344	CG4572	236	0.24	155	0.66
CG6726	CG17109	206	0.21	143	0.69
CG5917	CG7198	255	0.26	171	0.67
CG6871	CG9314	253	0.25	168	0.66
CG1635	CG1638	265	0.27	183	0.69
CG1982	CG4649	194	0.19	142	0.73
CG1946	CG1941	185	0.19	128	0.69
CG8719	CG8721	174	0.17	123	0.71
CG9331	CG9332	358	0.36	273	0.76
CG10476	CG10659	61	0.06	44	0.72
CG2206	CG15622	163	0.16	113	0.69
CG7547	CG6737	194	0.19	135	0.7
CG6943	CG1410	163	0.16	109	0.67
CG7890	CG8339	187	0.19	136	0.73
CG6217	CG14682	181	0.18	123	0.68
CG6186	CG3666	220	0.22	149	0.68
CG8664	CG8661	228	0.23	153	0.67
CG13042	CG13043	133	0.13	100	0.75
CG10140	CG10154	238	0.24	164	0.69
CG1499	CG12063	129	0.13	89	0.69
CG5112	CG8839	192	0.19	136	0.71
CG9338	CG14401	238	0.24	156	0.66
CG7547	CG9013	185	0.19	127	0.69
CG7291	CG3153	230	0.23	158	0.69
CG5346	CG5340	204	0.2	138	0.68
CG9330	CG10181	134	0.13	104	0.78
CG9427	CG9796	174	0.17	130	0.75
CG16914	CG8510	144	0.14	99	0.69

CG15143	CG15144	267	0.27	174	0.65
CG11316	CG1099	224	0.22	162	0.72
CG13029	CG17197	123	0.12	86	0.7
CG1221	CG18321	161	0.16	115	0.71
CG6946	CG8205	223	0.22	153	0.69
CG11207	CG1655	125	0.13	101	0.81
CG3025	CG1894	220	0.22	149	0.68
CG10140	CG10725	206	0.21	145	0.7
CG3812	CG17608	210	0.21	144	0.69
CG7788	CG5370	161	0.16	120	0.75
CG10474	CG1827	184	0.18	132	0.72
CG15257	CG10090	251	0.25	172	0.69
CG3415	CG3699	235	0.24	153	0.65
CG6790	CG4907	170	0.17	111	0.65
CG15144	CG15145	161	0.16	102	0.63
CG1571	CG10859	175	0.18	126	0.72
CG14781	CG15395	121	0.12	84	0.69
CG15177	CG15178	272	0.27	177	0.65
CG5008	CG13429	298	0.3	200	0.67
CG4465	CG4459	197	0.2	142	0.72
CG12684	CG15708	230	0.23	163	0.71
CG14736	CG17424	213	0.21	159	0.75
CG4769	CG14508	210	0.21	144	0.69
CG3568	CG2909	221	0.22	147	0.67
CG4549	CG4252	173	0.17	121	0.7
CG6376	CG2161	195	0.2	139	0.71
CG1049	CG18330	150	0.15	106	0.71
CG10797	CG5411	231	0.23	153	0.66
CG15379	CG14068	167	0.17	118	0.71
CG14782	CG6051	232	0.23	159	0.69
CG13547	CG12484	212	0.21	157	0.74
CG7860	CG10474	230	0.23	158	0.69
CG14610	CG12883	279	0.28	182	0.65
CG13941	CG10102	402	0.4	290	0.72
CG1705	CG6211	174	0.17	117	0.67
CG17031	CG1101	192	0.19	134	0.7
CG1966	CG2252	171	0.17	115	0.67
CG10605	CG10571	140	0.14	108	0.77
CG13133	CG4463	171	0.17	124	0.73
CG4714	CG4712	127	0.13	96	0.76
CG12255	CG4818	298	0.3	200	0.67
CG7465	CG13705	232	0.23	169	0.73
CG4465	CG6006	193	0.19	131	0.68
CG17725	CG10924	197	0.2	134	0.68
CG10037	CG6246	176	0.18	123	0.7
CG12676	CG3393	317	0.32	221	0.7
CG15583	CG2297	180	0.18	128	0.71
CG10916	CG3884	273	0.27	194	0.71
CG17109	CG6738	368	0.37	267	0.73

CG9656	CG10278	186	0.19	137	0.74
CG11560	CG17153	382	0.38	251	0.66
CG6612	CG6092	188	0.19	136	0.72
CG6726	CG6733	168	0.17	126	0.75
CG3819	CG14062	242	0.24	170	0.7
CG14214	CG8860	220	0.22	153	0.7
CG17028	CG17029	286	0.29	199	0.7
CG3253	CG15483	249	0.25	162	0.65
CG11928	CG14340	135	0.14	105	0.78
CG6901	CG17930	255	0.26	182	0.71
CG8142	CG5313	165	0.17	109	0.66
CG9906	CG9429	213	0.21	140	0.66
CG9707	CG9709	209	0.21	145	0.69
CG15813	CG9488	216	0.22	151	0.7
CG5988	CG5993	108	0.11	90	0.83
CG10752	CG12857	354	0.35	239	0.68
CG4837	CG4827	61	0.06	43	0.7
CG8865	CG5522	141	0.14	105	0.74
CG12700	CG8881	221	0.22	157	0.71
CG11941	CG8881	135	0.14	90	0.67
CG10173	CG6264	175	0.18	137	0.78
CG3612	CG17369	181	0.18	127	0.7
CG18537	CG18536	516	0.52	367	0.71
CG11392	CG4035	225	0.23	162	0.72
CG5191	CG5112	214	0.21	151	0.71
CG12743	CG3251	185	0.19	130	0.7
CG9427	CG13822	300	0.3	208	0.69
CG7349	CG3283	60	0.06	45	0.75
CG16874	CG9271	204	0.2	144	0.71
CG4152	CG6019	115	0.12	94	0.82
CG1617	CG13412	93	0.09	65	0.7
CG2262	CG12399	134	0.13	92	0.69
CG5515	CG15605	278	0.28	192	0.69
CG5398	CG17664	74	0.07	53	0.72
CG2043	CG8697	244	0.24	164	0.67
CG14999	CG8142	141	0.14	92	0.65
CG12206	CG11461	224	0.22	154	0.69
CG3210	CG8479	104	0.1	71	0.68
CG4237	CG6742	189	0.19	126	0.67
CG3988	CG6208	184	0.18	140	0.76
CG1349	CG6646	189	0.19	146	0.77
CG7054	CG17919	206	0.21	130	0.63
CG10160	CG13334	149	0.15	95	0.64
CG15107	CG18608	213	0.21	154	0.72
CG6894	CG7431	176	0.18	132	0.75
CG6617	CG18467	195	0.2	129	0.66
CG14411	CG5026	177	0.18	129	0.73
CG18589	CG10363	219	0.22	152	0.69
CG13597	CG2252	196	0.2	136	0.69

CG10260	CG5373	151	0.15	105	0.7
CG5472	CG12130	188	0.19	124	0.66
CG4700	CG4383	196	0.2	135	0.69
CG13941	CG12505	240	0.24	168	0.7
CG10966	CG8657	93	0.09	64	0.69
CG5355	CG2528	251	0.25	172	0.69
CG7334	CG15706	225	0.23	141	0.63
CG17140	CG17139	247	0.25	165	0.67
CG9076	CG9077	210	0.21	149	0.71
CG9709	CG5009	218	0.22	150	0.69
CG12498	CG10955	34	0.03	27	0.79
CG4855	CG4838	200	0.2	142	0.71
CG10124	CG1442	201	0.2	127	0.63
CG5804	CG8498	267	0.27	171	0.64
CG5373	CG4141	155	0.16	106	0.68
CG5978	CG8449	255	0.26	173	0.68
CG9046	CG16874	97	0.1	68	0.7
CG6998	CG5450	255	0.26	184	0.72
CG10478	CG9423	105	0.11	70	0.67
CG1724	CG15257	122	0.12	94	0.77
CG4104	CG5177	83	0.08	56	0.67
CG4706	CG4900	201	0.2	153	0.76
CG6775	CG6734	184	0.18	128	0.7
CG1550	CG11323	148	0.15	95	0.64
CG4907	CG13978	121	0.12	90	0.74
CG3992	CG10278	211	0.21	143	0.68
CG6874	CG14251	190	0.19	131	0.69
CG12824	CG12825	211	0.21	137	0.65
CG8117	CG3710	158	0.16	108	0.68
CG11942	CG8881	243	0.24	164	0.67
CG16954	CG7235	146	0.15	109	0.75
CG3788	CG6330	109	0.11	76	0.7
CG7533	CG7590	158	0.16	126	0.8

Table S2. Listing of the duplicate gene pairs with dS-values of less than one that were re-analyzed (see methods). Given are, from left to right, gene IDs #1 and 2, bin assigned based on Lynch and Conery (2000) $d_{S(L\&C)}$, re-calculated $d_{5'}$, d_S, and d_A and their respective 1 standard deviations (SD), as well as the resulting rates $d_{5'}/d_S$ and d_A/d_S. Furthermore, given are the number sites examined in the 5 prime region, synonymous sites and non-synonymous sites in base pairs ($bp_{5'}$, bp_S, bp_A) and the number of sites divergent between gene pairs ($n_{5'}$, n_S, n_A).

CG1	CG2	$d_{S(L\&C)}$	$d_{5'}$	SD	d_S	SD	d_A	SD	$d_{5'}/d_S$	d_A/d_S	$bp_{5'}$	$n_{5'}$	bp_S	n_S	bp_A	n_A
CG9111	CG1179	<0.1	0.212	0.042	0.000	0.000	0.125	0.041			166	30	37	0	88	10
CG8825	CG8826	<0.1	0.302	0.047	0.008	0.008	0.021	0.007	37.75	2.63	220	53	131	1	439	9
CG4137	CG15190	<0.1	0.027	0.007	0.010	0.010	0.020	0.007	2.70	2.00	522	14	101	1	400	8
CG18259	CG6961	<0.1	0.042	0.010	0.012	0.006	0.002	0.001	3.50	0.17	441	18	343	4	1080	2
CG4216	CG7271	<0.1	0.164	0.029	0.014	0.007	0.002	0.001	11.71	0.14	241	35	292	4	970	2
CG6997	CG10993	<0.1	0.214	0.030	0.024	0.050	0.190	0.023	9.07	8.05	318	58	132	26	467	77
CG3678	CG17556	<0.1	0.219	0.029	0.031	0.018	0.003	0.003	7.06	0.10	350	65	98	3	346	1
CG1742	CG12628	<0.1	0.181	0.039	0.033	0.019	0.016	0.008	5.48	0.48	152	24	94	3	260	4
CG12700	CG11941	<0.1	0.095	0.014	0.043	0.022	0.088	0.017	2.21	2.05	576	51	97	4	339	28
CG9111	CG1165	<0.1	0.448	0.090	0.057	0.041	0.228	0.060	7.86	4.00	122	39	36	2	89	17
CG18606	CG10476	<0.1	0.160	0.025	0.059	0.021	0.017	0.006	2.71	0.29	310	44	142	8	486	8
CG4479	CG4478	<0.1	0.097	0.016	0.069	0.029	0.036	0.011	1.41	0.52	421	38	92	6	340	12
CG10102	CG12505	<0.1	0.421	0.065	0.072	0.026	0.075	0.014	5.85	1.04	202	62	117	8	423	30
CG7952	CG4575	<0.1	0.414	0.056	0.076	0.035	0.102	0.023	5.45	1.34	267	81	70	5	213	20
CG1524	CG1527	<0.1	0.287	0.039	0.093	0.030	0.000	0.000	3.09	0.00	294	68	116	10	338	0
CG13402	CG18157	<0.1	0.148	0.028	0.094	0.048	0.052	0.018	1.57	0.55	227	30	46	4	161	8
CG9906	CG11958	<0.1	0.359	0.102	0.095	0.016	0.142	0.012	3.78	1.49	62	17	407	36	1199	153
CG7045	CG7046	0.1-0.2	0.127	0.029	0.104	0.040	0.142	0.024	1.22	1.37	182	21	73	7	305	39
CG17438	CG17441	0.1-0.2	0.178	0.026	0.113	0.025	0.130	0.014	1.58	1.15	327	51	230	21	754	89
CG13060	CG13041	0.1-0.2	0.143	0.017	0.118	0.037	0.011	0.006	1.21	0.09	582	75	102	11	270	3
CG18078	CG3927	0.1-0.2	0.289	0.059	0.118	0.043	0.214	0.037	2.45	1.81	129	30	74	8	214	39
CG18157	CG9404	0.1-0.2	0.089	0.019	0.124	0.033	0.098	0.016	0.72	0.79	276	23	133	15	452	41
CG15600	CG8565	0.1-0.2	0.301	0.056	0.126	0.066	0.096	0.031	2.39	0.76	150	36	35	4	112	10
CG18078	CG3875	0.1-0.2	0.326	0.075	0.129	0.048	0.177	0.032	2.53	1.37	98	25	68	8	226	35
CG9111	CG9118	0.1-0.2	0.327	0.050	0.133	0.069	0.011	0.011	2.46	0.08	223	57	33	4	93	1
CG15797	CG15910	0.1-0.2	0.330	0.076	0.134	0.031	0.203	0.024	2.46	1.51	97	25	173	21	465	81
CG13402	CG18313	0.1-0.2	0.129	0.024	0.136	0.054	0.065	0.018	0.95	0.48	274	32	57	7	210	13
CG18281	CG17637	0.1-0.2	0.121	0.019	0.143	0.022	0.050	0.007	0.85	0.35	380	42	350	45	1142	55
CG7027	CG17355	0.1-0.2	0.094	0.029	0.145	0.054	0.088	0.022	0.65	0.61	125	11	62	8	201	17
CG18157	CG18313	0.1-0.2	0.013	0.006	0.150	0.032	0.128	0.018	0.09	0.85	317	4	134	18	457	53
CG9404	CG18313	0.1-0.2	0.106	0.018	0.157	0.035	0.093	0.014	0.68	0.59	368	36	164	23	565	49
CG17597	CG17320	0.1-0.2	0.368	0.048	0.158	0.027	0.017	0.005	2.33	0.11	290	81	263	37	871	15
CG6891	CG6900	0.1-0.2	0.221	0.138	0.167	0.062	0.080	0.023	1.32	0.48	16	3	54	8	173	13
CG7370	CG17812	0.1-0.2	0.117	0.070	0.167	0.072	0.080	0.025	0.70	0.48	28	3	41	6	145	11
CG1740	CG10174	0.1-0.2	0.336	0.056	0.171	0.050	0.055	0.014	1.96	0.32	184	48	86	13	304	16

CG5334	CG5347	0.1-0.2	0.221	0.028	0.172	0.041	0.175	0.022	1.28	1.02	395	74	132	20	482	74
CG18372	CG10146	0.1-0.2	0.367	0.041	0.176	0.037	0.016	0.006	2.09	0.09	402	112	162	25	492	8
CG15646	CG12708	0.1-0.2	0.516	0.072	0.177	0.038	0.067	0.012	2.92	0.38	263	92	155	24	472	30
CG12493	CG10630	0.1-0.2	0.172	0.026	0.180	0.038	0.186	0.021	0.96	1.03	324	49	159	25	561	91
CG13402	CG9404	0.1-0.2	0.163	0.029	0.182	0.058	0.067	0.017	0.90	0.37	236	34	69	11	252	16
CG18078	CG3584	0.1-0.2	0.304	0.063	0.188	0.060	0.142	0.028	1.62	0.76	124	30	68	11	227	29
CG9046	CG9271	0.1-0.2	0.441	0.063	0.198	0.049	0.421	0.060	2.23	2.13	240	76	112	19	238	73
CG18078	CG4021	0.2-0.3	0.337	0.061	0.207	0.064	0.200	0.034	1.63	0.97	153	40	68	12	226	39
CG11941	CG11942	0.2-0.3	0.226	0.027	0.217	0.054	0.261	0.034	1.04	1.20	446	85	103	19	331	71
CG11386	CG10944	0.2-0.3	0.505	0.076	0.222	0.055	0.235	0.033	2.27	1.06	223	77	101	19	304	60
CG14500	CG14499	0.2-0.3	0.297	0.039	0.223	0.059	0.199	0.032	1.33	0.89	311	74	90	47	657	45
CG15577	CG15578	0.2-0.3	0.382	0.046	0.233	0.070	0.206	0.038	1.64	0.88	342	98	66	13	193	34
CG15645	CG13732	0.2-0.3	0.327	0.040	0.242	0.058	0.148	0.021	1.35	0.61	336	86	104	21	415	55
CG6997	CG6999	0.2-0.3	0.245	0.034	0.243	0.050	0.273	0.029	1.01	1.12	299	61	138	28	491	109
CG12700	CG11942	0.2-0.3	0.249	0.028	0.260	0.060	0.250	0.031	0.96	0.96	450	93	108	23	337	76
CG1365	CG1367	0.2-0.3	0.347	0.049	0.262	0.091	0.000	0.000	1.32	0.00	254	68	47	10	143	0
CG13617	CG12842	0.2-0.3	0.374	0.053	0.264	0.069	0.113	0.023	1.42	0.43	248	70	83	18	240	25
CG13324	CG13323	0.2-0.3	0.363	0.056	0.271	0.071	0.032	0.011	1.34	0.12	210	58	81	18	255	8
CG3895	CG18414	0.2-0.3	0.386	0.128	0.275	0.033	0.065	0.008	1.40	0.24	45	13	389	87	1147	71
CG14213	CG9573	0.2-0.3	0.346	0.047	0.275	0.045	0.204	0.020	1.26	0.24	270	72	210	17	262	115
CG12359	CG11023	0.2-0.3	0.470	0.075	0.279	0.038	0.197	0.016	1.68	0.71	191	63	296	67	1069	182
CG13069	CG13051	0.2-0.3	0.199	0.024	0.287	0.083	0.146	0.033	0.69	0.51	478	82	65	15	160	21
CG10811	CG10810	0.2-0.3	0.442	0.068	0.296	0.086	0.813	0.205	1.49	2.75	205	65	63	15	144	63
CG9902	CG7692	0.3-0.4	0.206	0.030	0.300	0.028	0.119	0.008	0.69	0.40	312	55	626	150	254	245
CG11719	CG18396	0.3-0.4	0.378	0.053	0.303	0.050	0.025	0.007	1.25	0.08	253	72	190	46	605	15
CG1924	CG11958	0.3-0.4	0.424	0.081	0.306	0.039	0.113	0.010	1.39	0.37	133	41	329	80	1201	125
CG12224	CG3397	0.3-0.4	0.302	0.035	0.323	0.050	0.046	0.009	0.93	0.14	386	93	213	54	669	30
CG9111	CG1180	0.3-0.4	0.244	0.045	0.347	0.134	0.068	0.028	0.70	0.26	172	35	34	9	92	6
CG8628	CG8629	0.3-0.4	0.373	0.056	0.347	0.108	0.091	0.023	1.07	0.26	220	62	52	14	200	17
CG1924	CG9906	0.3-0.4	0.375	0.060	0.351	0.042	0.163	0.013	1.07	0.46	191	54	356	96	1278	184
CG2694	CG11322	0.3-0.4	0.477	0.063	0.367	0.046	0.269	0.020	1.30	0.73	285	95	319	89	1020	224
CG6246	CG12287	0.3-0.4	0.278	0.083	0.371	0.044	0.540	0.039	0.75	1.46	62	14	349	98	990	356
CG4960	CG8331	0.3-0.4	0.315	0.053	0.373	0.077	0.113	0.018	0.84	0.30	185	46	117	33	396	41
CG12477	CG7184	0.3-0.4	0.368	0.050	0.381	0.061	0.368	0.034	0.97	0.97	272	76	196	56	581	162
CG6999	CG10993	0.3-0.4	0.271	0.034	0.387	0.077	0.464	0.049	0.70	1.20	362	80	124	36	440	144
CG8066	CG8050	0.3-0.4	0.000	0.000	0.394	0.130	0.123	0.025	0.00	0.31	108	0	72	21	240	27
CG9046	CG9048	0.4-0.5	0.453	0.069	0.439	0.094	0.459	0.064	1.03	1.05	211	68	107	34	249	81
CG18107	CG18108	0.4-0.5	0.000	0.000	0.492	0.185	0.100	0.034	0.00	0.20	302	0	35	12	97	9
CG6857	CG6881	0.5-0.6	0.442	0.076	0.511	0.066	0.387	0.030	0.86	0.76	164	52	305	106	801	232
CG18106	CG18108	0.5-0.6	0.301	0.041	0.519	0.200	0.062	0.026	0.58	0.12	279	67	34	12	101	6
CG13794	CG13795	0.5-0.6	0.434	0.059	0.556	0.187	0.106	0.027	0.78	0.19	265	83	47	17	164	16

CG8219	CG7398	0.5-0.6	0.442	0.057	0.583	0.056	0.173	0.010	0.76	0.30	290	92	583	219	2039	310
CG15332	CG15333	0.5-0.6	0.348	0.086	0.599	0.110	0.311	0.030	0.58	0.52	82	22	165	63	544	134
CG11444	CG4438	0.6-0.7	0.328	0.048	0.606	0.122	0.194	0.024	0.54	0.32	238	61	138	53	429	72
CG14850	CG14851	0.6-0.7	0.423	0.052	0.609	0.137	0.355	0.048	0.69	0.58	328	101	112	43	279	76
CG9906	CG11235	0.6-0.7	0.448	0.082	0.622	0.136	0.735	0.104	0.72	1.18	147	47	489	121	368	155
CG13589	CG13590	0.6-0.7	0.214	0.039	0.679	0.215	0.073	0.018	0.32	0.11	192	35	64	26	245	17
CG15332	CG18293	0.7-0.8	0.431	0.077	0.706	0.134	0.350	0.030	0.61	0.50	154	48	191	79	673	181
CG6343	CG5585	0.7-0.8	0.245	0.046	0.751	0.123	0.871	0.057	0.33	1.16	167	34	285	121	782	403
CG1365	CG1373	0.7-0.8	0.323	0.050	0.787	0.337	0.036	0.016	0.41	0.05	217	55	46	20	143	5
CG11520	CG10810	0.8-0.9	0.477	0.063	0.826	0.379	0.079	0.023	0.58	0.10	288	96	45	20	162	12
CG10813	CG10812	0.8-0.9	0.433	0.062	0.846	0.230	0.107	0.028	0.51	0.13	240	75	45	23	161	16
CG12699	CG18469	0.8-0.9	0.000	0.000	0.860	0.152	0.429	0.053	0.00	0.50	40	0	108	55	316	98
CG14850	CG8087	0.9-1.0	0.355	0.046	0.946	0.326	0.445	0.057	0.38	0.47	298	81	126	58	292	93
CG8628	CG15829	0.9-1.0	0.357	0.063	0.962	0.545	0.429	0.068	0.37	0.45	161	44	50	23	196	61
CG1365	CG1878	0.9-1.0	0.239	0.042	1.034	0.290	0.082	0.025	0.23	0.08	195	39	46	26	143	11
CG13063	CG13043	0.9-1.0	0.230	0.041	1.127	0.641	0.076	0.017	0.20	0.07	191	37	106	51	280	20