Genome Evolution and Developmental Constraint in *Caenorhabditis elegans*

Cristian I. Castillo-Davis and Daniel L. Hartl

Department of Organismic and Evolutionary Biology, Harvard University

It has been hypothesized that evolutionary changes will be more frequent in later ontogeny than early ontogeny because of developmental constraint. To test this hypothesis, a genomewide examination of molecular evolution through ontogeny was carried out using comparative genomic data in *Caenorhabditis elegans* and *Caenorhabditis briggsae*. We found that the mean rate of amino acid replacement is not significantly different between genes expressed during and after embryogenesis. However, synonymous substitution rates differed significantly between these two classes. A genomewide survey of correlation between codon bias and expression level found codon bias to be significantly correlated with mRNA expression ($r_s = -0.30$ and $P < 10^{-131}$) but does not alone explain differences in *dS* between classes. Surprisingly, it was found that genes expressed after embryogenesis have a significantly greater number of duplicates in both the *C. elegans* and *C. briggsae* genomes ($P < 10^{-20}$ and $P < 10^{-13}$) when compared with early-expressed and nonmodulated genes. A similarity in the distribution of duplicates of nonmodulated and early-expressed genes, as well as a disproportionately higher number of early pseudogenes, lend support to the hypothesis that this difference in duplicate number is caused by selection against gene duplicates of early-expressed genes, reflecting developmental constraint. Developmental constraint at the level of gene duplication may have important implications for macroevolutionary change.

Introduction

Understanding interspecific morphological differences and patterns of diversity has long been the focus of both paleontologists and evolutionary biologists and has been the impetus for a prodigious amount of theoretical and speculative work. Much of this theory strives to establish general principles that are responsible for large-scale patterns of morphological diversity witnessed in extant and extinct taxa. Central among these is the concept of developmental constraint, the notion that the structure of animal development itself may place limits on the morphological space that organisms can explore over evolutionary time (Riedl 1978, pp. 102–114; Arthur 1988, pp. 13, 40, and 68; Raff 1996, pp. 292–320; Arthur 2000).

One rationale for developmental constraint follows from the belief that mutations that occur early in development are likely to be deleterious because later genetic and epigenetic events often depend on earlier ones (Riedl 1978). As a result, it has been hypothesized that, in general, evolutionary changes will be much more frequent in late development than in early development. Further, sensitivity to genetic perturbation (mutation) is thought to increase with increasing number of gene interactions, codependencies, and spatiotemporal precision in timing of gene expression (Arthur 1988, 1997; see also Goodwin, Kaufmann, and Murray 1993). Coupled with the observation that early development is plastic in many phyla, the expectation of constraint has been refined and theoretically localized to the so-called phylotypic stage, a stage in development—not necessarily the earliest stage—during which a maximal interaction of genetic modules occurs (Raff 1996, pp. 208–210).

Given that all developmental processes ultimately depend on the activity of specific sets of genes and their interactions, it should be expected that some amount of genetic difference underlies the differences in development between species. These differences may be manifested in proteins that are directly involved in developmental activities (e.g., morphogens) or in the regulatory sequences that control the interaction of these proteins (or both) (Sucena and Stern 2000). A genomic signature of developmental constraint may therefore be present in coding sequences or at the level of *cis*-acting regulatory sequences, or both.

Here, we undertake a test of the former hypothesis that proteins expressed early in ontogeny, specifically embryogenesis, evolve more slowly as a class than genes expressed later in ontogeny. Specifically, we test the null hypothesis that genes expressed during embryogenesis, which includes the hypothesized phylotypic stage, are no more constrained in their rate of molecular evolution than genes expressed later in development.

First, using newly available genome sequence data from *Caenorhabditis briggsae*, we estimate rates of amino acid substitution (*dN*) and synonymous substitution (*dS*) of genes expressed during and after embryogenesis in *Caenorhabditis elegans*. Second, we determine genomewide levels of gene duplication in each developmental class using both *C. elegans* and *C. briggsae* genomes. The null hypothesis of no developmental constraint predicts that genes expressed during and after embryogenesis will have similar rates of nonsynonymous and synonymous substitution and similar proportions of duplicate genes.

In *C. elegans*, embryogenesis takes approximately 12 h from fertilization (0 h) and is followed by four larval molts, with sexual maturity at 72 h at L_4 and death after approximately 2 weeks (Bird and Bird 1991, pp. 26, 77). We use existing *C. elegans* microarray expression data (Hill et al. 2000) which span eight time points through

Abbreviations: ENC, effective number of codons; ppm, parts per million.

Key words: comparative genomics, genome evolution, microarray analysis, developmental constraint, gene duplication, gene expression, molecular evolution.

Address for correspondence and reprints: Daniel L. Hartl, Biological Laboratories, Harvard University, 16 Divinity Avenue, Cambridge, Massachusetts 02138. E-mail: dhartl@oeb.harvard.edu.

Mol. Biol. Evol. 19(5):728–735. 2002
 ISSN: 0737-4038

development, including oocyte, 0, 12, 24, 36, 48, and 60 h, and 2 weeks, to identify genes that peak in expression during and after embryogenesis, referred to henceforth, for convenience, as early- and late-expressed. Genes with peak expression at the oocyte and 0-h stages (embryogenesis) were considered early-expressed, whereas genes with peak expression at the 12-, 24-, 36-, 48-, and 60-h stages (after embryogenesis) were considered late-expressed.

We found that genes expressed early and late in development do not show significantly different rates of amino acid replacement, but they do show significantly different rates of synonymous substitution. This difference in synonymous substitution rates is most likely caused by significant variation in levels of codon-usage bias between the two classes of genes, which in turn reflects differences in expression level between early and late genes.

A highly significant correlation was found between number of gene duplicates per gene and developmental class, with early-expressed genes presenting far fewer paralogs per gene in both the *C. elegans* and *C. briggsae* genomes. This paucity of duplicates may involve developmental constraint at the level of gene duplication in embryogenesis or the selective retention and divergence of postembryonic gene duplicates.

The hypothesis of developmental constraint is supported by a similar distribution of nonmodulated gene duplicates and early gene duplicates as well as by an analysis of the distribution of class-related pseudogenes in the genome. More than twice as many pseudogenes as expected are derived from early-expressed genes, implying selective retention of nonfunctional duplicates in this class.

Methods

Expression Data

Expression data were obtained from Hill et al. (2000) in which DNA microarrays were used to examine mRNA expression at eight different stages of *C. elegans* development: oocyte, 0, 12, 24, 36, 48, and 60 h, and 2 weeks. Of the 18,908 genes on the array, 11,257 were called absent or absent at least once, according to the criteria of Hill et al. (2000). This left 7,651 genes for which reliable expression values were obtained. Of these, 6,235 were retained in the present analysis after excluding genes for which there was duplicate or conflicting expression information in the data of Hill et al. (2000) and genes whose sequences could not be easily retrieved from GenBank.

Genes with peak expression at the oocyte and 0-h 12-h stages were considered early-expressed, whereas genes with peak expression at the 12-, 24-, 36-, 48-, and 60-h stages were considered late expressed (clusters [1,1]–[2,6] and [5,2]–[6,6], respectively, as designated by Hill et al. [2000]), yielding 1,328 and 1,074 early and late genes, respectively. Genes with strongly bimodal or multimodal expression were excluded from the analysis [clusters (4,1)–(5,1)]. Genes that were not significantly modulated through ontogeny were obtained by subtracting all significantly modulated genes from the total list of genes analyzed, yielding 3,860 nonmodulated genes.

According to Hill et al. (2000), genes with low transcript abundance early in development may cluster with early-expressed genes because of limitations of array detection at later developmental stages. To identify any results caused solely by the inclusion of such rare transcripts, we determined how many genes in our subsample of early genes had an expression level less than 30 ppm at the 0-h stage, the approximate limit of detection of rare transcripts (supplementary material, Hill et al. 2000). Close to 50% of early genes were represented by rare transcripts, according to this criterion. However, exclusion of genes with rare transcripts from the analysis did not significantly affect the results; data utilizing the full set of genes are therefore reported. Minor differences between early genes with rare transcripts and other early genes, where they occur, are also reported for completeness.

Retrieval and Analysis of *C. briggsae* Orthologs and Paralogs

Sequences homologous to *C. elegans* genes were retrieved computationally from *C. briggsae* genomic sequence with GeneSeqer (Usuka, Zhu, and Brendel 2001) using nematode-specific splice-site settings. Finished, nonredundant genomic sequence of 12 Mb, approximately 25% of the *C. briggsae* genome, was obtained from the Genome Sequencing Center at Washington University, St. Louis (WUSTL) and used to probe each full-length *C. elegans* coding sequence. Introns were removed and exons concatenated computationally from the resulting set of 1,585 *C. elegans*-*C. briggsae* alignments comprising 909 unique *C. elegans* genes.

Because multiple alignments for each gene were often retrieved and because the *C. briggsae* genome sequence is not complete, special care was taken to establish orthology between sequences using the method of reciprocal best hits (Tatusov, Koonin, and Lipman 1997). First, the highest scoring *C. briggsae* sequence was identified for each gene based on a normalized similarity score from *C. elegans*-*C. briggsae* alignments, yielding 909 best-hit alignments. Second, a BLASTN (v. 2.1.2) search (Altschul et al. 1997) of each putative *C. briggsae* ortholog was performed against the *C. elegans* genome. If the initial *C. elegans* gene was retrieved as the best hit, the pair was accepted as orthologous; otherwise, the pair was rejected. This process resulted in a set of 492 valid alignments. Finally, sequences with stop codons were eliminated, yielding a final set of 201 genes.

Genes in each developmental class and their molecular functions, if known, are listed in the supplementary information and can be accessed through the Molecular Biology and Evolution web site (http://www.molbiolevol.org). Genes in each class correspond very well with expected class functions. For example, early-expressed genes in the sample include many transcription factors, a homeobox protein, and a regulator

of G-protein signaling. Late-expressed genes are, as expected, of more diverse functions and include various metabolic and structural genes, including synthases, transferases, collagen, and proteases.

Maximum likelihood estimates of nonsynonymous substitutions (*dN*) and synonymous substitutions (*dS*) between pairwise alignments were obtained with PAML (Yang 2000) using a codon-based model of sequence evolution with *dN*, *dS*, and transition-tranversion bias as free parameters and codon frequencies estimated from the data at each codon position (F3 × 4 model; Goldman and Yang 1994; Yang 2000).

Gene Duplications

Relative proportions of paralogs per gene in each developmental class were determined by two methods. First, the number of paralogs per gene in the *C. elegans* genome was estimated by counting the number of significant hits returned by BLAST searches of each *C. elegans* gene against the complete coding sequence of the *C. elegans* genome. E-values less than 1×10^{-10} were considered significant matches. Second, an estimate of the number of paralogs in each developmental class was carried out using 12 Mb of *C. briggsae* genomic sequence, approximately 25% of the genome (WUSTL, unpublished data). The number of paralogs per *C. elegans* gene in this random sample of the *C. briggsae* genome (WUSTL, unpublished data) was estimated by counting the number of significant alignments returned by GeneSeqer after correcting for alignments caused by alternative splicing predictions by GeneSeqer.

Codon Bias

The average mRNA expression in transcripts per million (ppm) from Hill et al. (2000) was calculated for each gene by taking mean expression across all life stages. Only transcripts that were detected in all repeated hybridizations for a particular life stage were used to calculate mean expression for a particular gene. Codon bias for the resulting 6,235 genes was measured in effective number of codons (ENC) (Wright 1990) calculated with the molecular evolutionary program MEA (E. N. Moriyama, personal communication). ENC measures deviation from expected random codon usage and is independent of hypotheses involving natural selection. ENC ranges from 20.0 (highest possible bias) to 61.0 (no bias). Because of the scope of this study, this measure of codon bias was used, instead of alternatives such as the codon adaptation index of Sharp and Li (1987) that rely on sets of preferred codons based on a small sample of genes (Wright 1990).

Pseudogene Analysis

A list of 305 known or suspected pseudogenes was obtained from Wormbase (Stein et al. 2001). A BLAST search of each pseudogene sequence against all genes in the early, late, and nonmodulated categories was performed. Pseudogenes with significant matches were then classified as early, late, or both if a pseudogene matched a gene in both the early and late expression classes. E values less than 1×10^{-10} were considered significant matches.

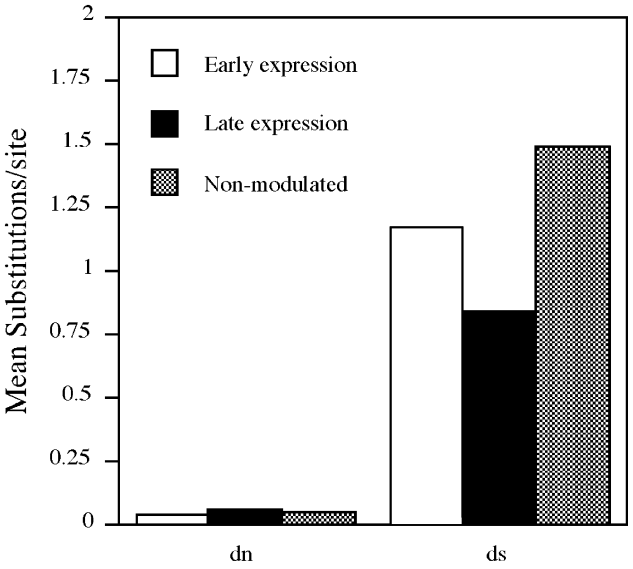

FIG. 1.—Mean rates of nonsynonymous (*dN*) and synonymous (*dS*) substitution in genes in early, late, and nonmodulated expression classes. *dN* is not significantly different between genes in different expression classes. *dS* is significantly different between early- and late-expressed genes as well as between nonmodulated and late-expressed genes.

Statistics

Tests of significant deviation from the null expectation of equal rates of molecular evolution (*dN* and *dS*) were carried out using nonparametric bootstrapping with replacement (10,000 replicates) in Mathematica (Wolfram 1999). Differences in numbers of duplicate genes in each class were tested using 2 × 2 contingency tables and the χ^2 statistic. Student's *t*-test (two-tailed) was used to test differences in mean expression and codon bias between classes. Spearman rank correlation coefficients (r_s) and associated *P*-values were calculated in Mathematica (Wolfram 1999).

Results

dN, *dS*, and Codon Bias

The null hypothesis of equal rates of amino acid replacement in genes expressed early and late in development could not be rejected. Rates of amino acid replacement (*dN*) were not significantly different between early-expressed ($n = 29$), late expressed ($n = 90$), and nonmodulated genes ($n = 105$): *dN* early = 0.0456, *dN* late = 0.0512, *dN* nonmodulated = 0.047, with 95% confidence intervals (CI) (0.0342–0.0585), (0.0400–0.0666), and (0.041–0.054), respectively (fig. 1). This result was not affected by focusing on early genes with transcript abundance above 30 ppm at the 0-h stage: *dN* early = 0.0447, 95% CI (0.0247–0.0696).

For the full set of early genes, rates of synonymous substitution (*dS*) were found to be significantly different

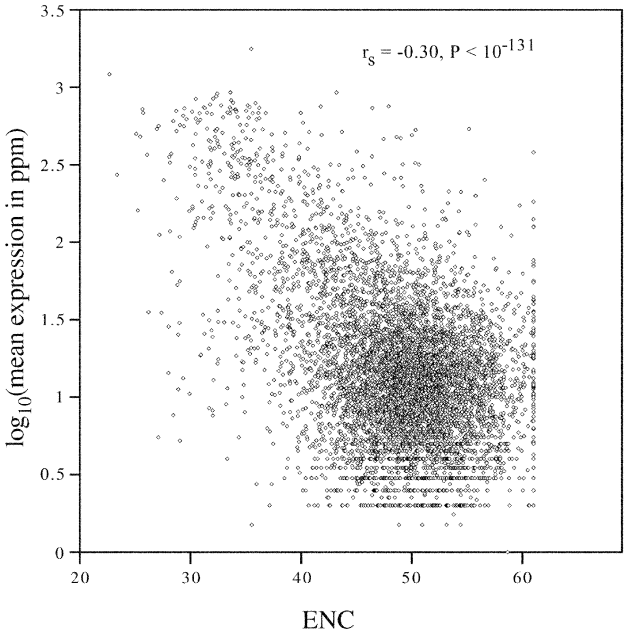

FIG. 2.—The relationship between mean mRNA expression and codon bias (ENC) in *C. elegans*, $r_s = -0.30$ and $P < 10^{-131}$.

between early-expressed and late-expressed genes: *dS* early = 1.355, *dS* late = 0.8678, with 95% CI (1.093–1.628) and (0.760–0.982), respectively, as well as between late-expressed and nonmodulated genes: *dS* nonmodulated = 1.487, with 95% CI (1.331–1.645) (fig. 1).

When genes with rare transcript abundance (<30 ppm) were excluded from this analysis, a mean difference in *dS* between early and late classes, although still apparent, was no longer statistically significant: *dS* early = 0.92131 and 95% CI (0.6588–1.2254). This may be because of a smaller sample size, ($n = 13$ vs. $n = 29$) or the relationship between transcript abundance, codon-usage bias, and *dS* described subsequently.

The differences we detected in *dS* are most simply explained to be the result of a mean difference in the codon-usage bias in each set of genes. For the early genes the mean ENC = 45.7, whereas for the late genes the mean ENC = 41.39 ($P < 0.05$). This pattern also holds when all genes in each expression class are analyzed: mean ENC early = 49.84 ($n = 1{,}269$), ENC late = 44.7 ($n = 1{,}043$), $P < 10^{-51}$ (data for nonmodulated genes are not shown).

To explore whether the difference in codon bias could be related to differences in levels of expression between the expression classes, the mean level of mRNA expression was calculated for all available genes in each class. Mean expression was indeed found to be significantly different for early and late classes, with early genes showing a mean of 27.18 ppm ($n = 1{,}269$) and late genes showing a mean of 90.45 ppm ($n = 1{,}043$), $P < 10^{-81}$.

To further explore the relationship between codon bias and expression, a genomewide survey of correlation between codon bias and expression was carried out. Codon bias was found to be significantly correlated with mRNA expression level with $r_s = -0.30$ ($n = 6{,}235$) and $P < 10^{-131}$ (fig. 2), consistent with the results found in yeast using a similar method (Coghlan and Wolfe 2000). Although such a correlation has been inferred for *C. elegans* previously using approximate methods (Duret and Mouchiroud 1999), the data here represent the first demonstration of a continuous relationship between mRNA expression and codon bias across the genome for a multicellular animal. A similar pattern of bias and expression held for modulated genes when examined independently as a class, $r_s = -0.47$ ($n = 3{,}147$) with $P < 10^{-174}$, but was weaker for nonmodulated genes as a class: $r_s = -0.15$, $n = 3{,}088$, $P < 10^{-16}$.

Duplicate Genes

The number of paralogs per gene was found to differ significantly between developmental classes. On an average, genes expressed during late development had significantly more paralogs, as well as more paralogs per gene, than genes expressed during early development. Only 18.36% of early-expressed genes ($n = 1{,}280$) had detectable paralogs in the *C. elegans* genome versus 35.31% of late-expressed genes ($n = 1{,}014$) (fig. 3). This pattern was even stronger in the *C. briggsae* genome: only 6.70% of early-expressed genes ($n = 165$) had detectable paralogs in the *C. briggsae* genome versus 38.8% of late-expressed genes ($n = 237$) (fig. 4).

Genes that were not significantly modulated through development had a distribution of duplicates more similar to that of early-expressed genes than that of late-expressed genes (figs. 3 and 4). In general, early expressed and nonmodulated genes had distributions of duplicates only marginally different from each other ($P = 0.26$ in *C. elegans* genome; $P = 0.05$ in *C. briggsae* genome) although both classes differed markedly in their distributions of duplicates from that of late-expressed genes—in the *C. elegans* genome: $P < 10^{-20}$ early versus late, $P < 10^{-23}$ nonmodulated versus late; in the *C. briggsae* genome: $P < 10^{-13}$ early versus late, $P < 10^{-12}$ nonmodulated versus late (figs. 3 and 4).

These results are not compromised by exclusion of early-expressed genes with low transcript abundance. For the reduced data set in *C. briggsae*, the fraction of early-expressed genes with detectable paralogs in the genome increases to 12.10%, whereas in *C. elegans*, this figure decreases to 12.74%. Differences in the number of paralogs remain statistically significant between classes (data not shown).

Pseudogene Analysis

Of the 305 annotated pseudogenes in the *C. elegans* genome, 48 pseudogenes showed significant similarity with genes in early or late expression classes. Twenty-seven pseudogenes matched early-expressed genes exclusively, and 13 pseudogenes matched late-expressed genes exclusively; eight pseudogenes showed significant similarity with one or more genes in each class and were excluded from further analysis. No processed pseudogenes, as identified by the presence of a poly-A tail, showed significant similarity with genes in early-expressed, late-expressed, or nonmodulated genes.

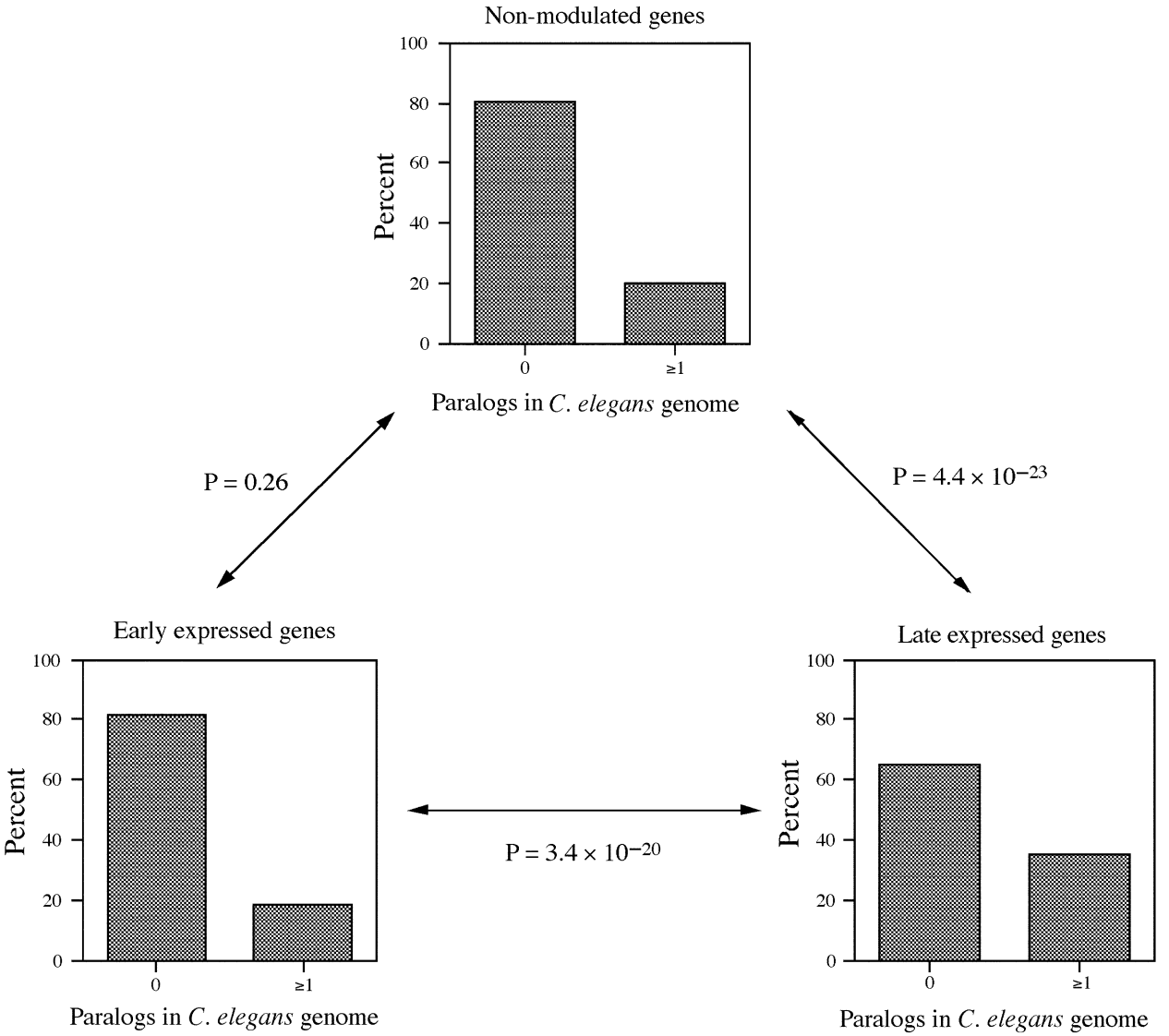

FIG. 3.—Proportion of paralogs showing significant similarity with genes in early, late, and nonmodulated expression classes in the *C. elegans* genome. Early-expressed genes have significantly fewer paralogs in the genome than late-expressed genes. Nonmodulated and early-expressed genes have a statistically similar distribution of paralogs.

The expected number of pseudogenes in early and late classes was calculated by adjusting for the number of genes in each expression class and for the fact that late-expressed genes have on an average 2.22 times more paralogs per gene than early-expressed genes. The observed 27 and 13 pseudogene matches for early and late expression classes, respectively, was significantly different from the null expectation of 14.3 and 25.7 pseudogenes matches per class ($P < 10^{-5}$ by 2×2 contingency table). Similar results were obtained using E values progressively greater than 1×10^{-10} (data not shown).

Discussion

Numerous studies have identified genes that have undergone rapid and identifiable evolution because of positive selection, e.g., male reproductive proteins (Civetta and Singh 1998; Wyckoff et al. 2000; Parsch et al. 2001); other genes have been found to evolve extremely slowly or not at all because of strong purifying selection, e.g., some homeobox genes (Kumar and Hedges 1998) and histones (Li 1997, p. 189). At least one study in mammals has identified a pattern of selection (differences in mean amino acid substitution rates) among genes that vary in their breadth of tissue expression (Duret and Mouchiroud 2001). Thus, it is interesting that our examination of over 200 genes across development yielded no statistical differences in rates of amino acid evolution among genes expressed during and after embryogenesis and genes not significantly modulated through development (fig. 1).

The hypothesis that proteins expressed early in ontogeny, specifically embryogenesis, evolve more slowly as a class than genes expressed later in ontogeny is not supported by the data. According to the results presented here, if timing of developmental deployment impacts rates of protein evolution, it does not seem to do so in

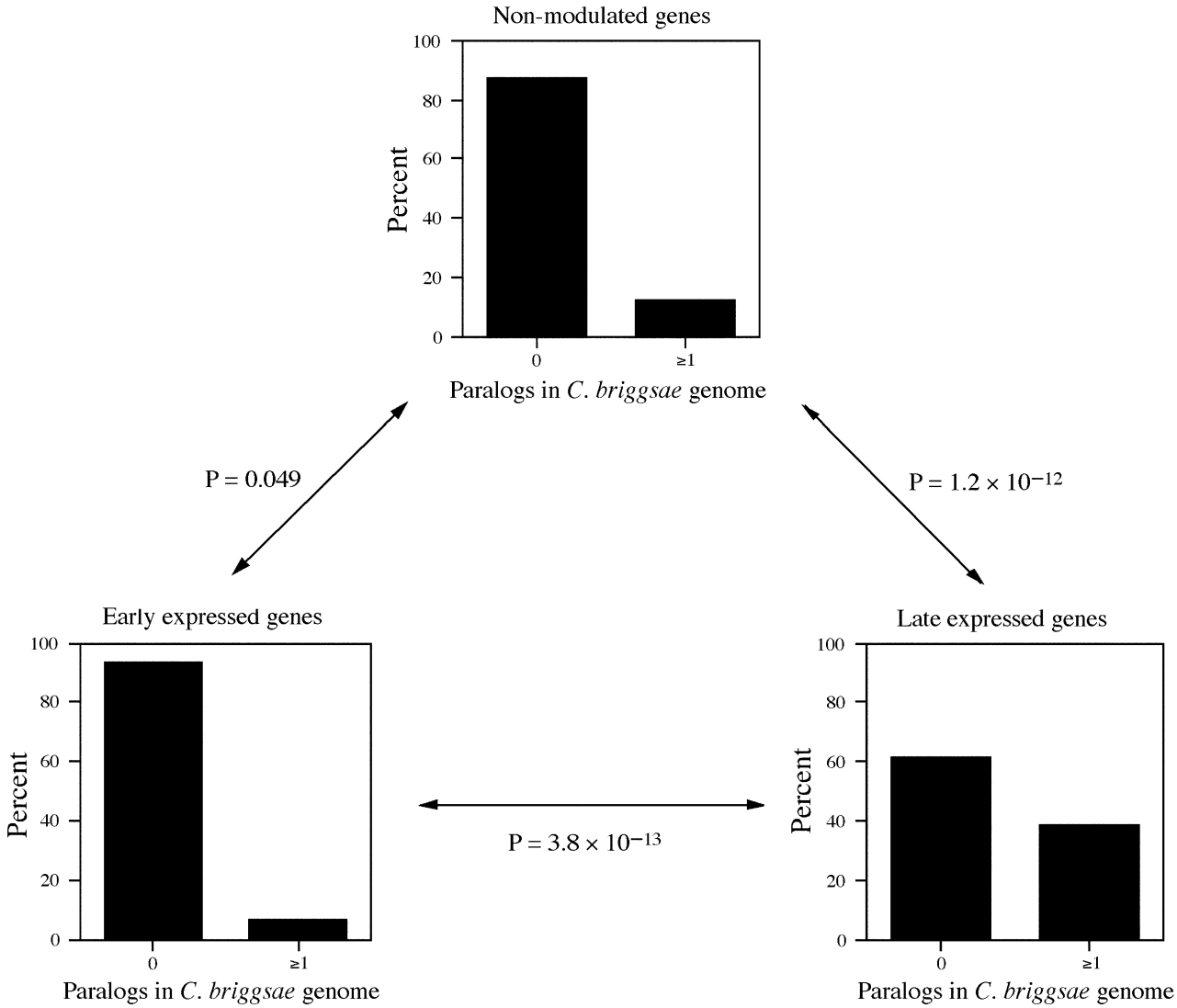

FIG. 4.—Proportion of paralogs showing significant similarity with genes in early, late, and nonmodulated expression classes in the *C. briggsae* genome. Early-expressed *C. elegans* genes have significantly fewer paralogs in this genome than late-expressed genes. Nonmodulated and early-expressed genes have a similar but marginally different distribution of paralogs in *C. briggsae*.

a dramatic manner. Although it is still possible that the evolution of key amino acid residues is constrained in early-expressed genes, no evidence of a large-scale effect on protein evolution because of development constraint is supported.

Surprisingly, rates of synonymous substitution (*dS*) were found to be significantly different between early-expressed and late-expressed genes. This difference is likely a consequence of differences in codon bias between the two classes; in fact, *dS* is strongly negatively correlated with codon bias, with $r_s = 0.74$ ($P < 10^{-20}$, data not shown). In contrast, it has been shown for a much smaller data set in *Drosophila* that synonymous substitution rates are independent of codon bias when maximum likelihood methods, identical to those used here, are used to estimate *dS* (Dunn et al. 2001). Although likelihood methods which incorporate codon-usage patterns in estimates of *dS* are superior to approximate methods, these methods are still prone to underestimation of *dS* if the degree of divergence between sequences is high and codon bias is strong (Dunn et al. 2001). However, in the data analyzed here, *dS* is moderate (mean *dS* = 1.18) and codon bias is not extreme (mean ENC late = 41.39)—conditions under which maximum likelihood methods perform very well (Dunn, Bielawski, and Yang 2001). Here, the observed difference in *dS* between early and late expression classes is unlikely to be an artifact of underestimation of *dS* in the late expression class.

Lack of constraint at the level of protein evolution does not preclude constraint at the level of regulatory sequence evolution or at specific functional domains within proteins, two possibilities not tested here. A test of this hypothesis awaits (1) comparison of expression profiles through ontogeny across multiple species, with tests of selection on developmentally important *cis*-acting regulatory sequences, and (2) assessment of all functionally relevant domains in the proteome.

Timing in developmental deployment is, however, significantly correlated with levels of gene duplication in *C. elegans*. Strikingly, more than 81% of early-expressed *C. elegans* genes have no paralogs in the *C. elegans* genome, and more than 90% have only single matches in the *C. briggsae* genome. In contrast, more than a third of all late-expressed genes have one or more paralogs in the *C. elegans* genome, and close to 40% have one or more paralogs in the *C. briggsae* genome. What is responsible for this dramatic difference in the number of paralogs between early- and late-expressed genes?

Two possibilities present themselves: a biased origin of duplicates in late-expressed genes or a biased loss of duplicates in early-expressed genes. The mechanisms by which gene duplications are thought to occur—homologous recombination, replication slippage, and transposition (Li 1997)—are general molecular phenomena and thus make the biased origin of paralogs in one class over another unlikely. Instead, the lower than expected number of duplicates in the early-expressed class is most likely caused by a biased loss of duplicates in this class over evolutionary time.

According to one study, the origin of duplicate genes in *C. elegans* is of the order of 0.0208/gene/Myr, giving an expected 383 dup/genome/Myr (Lynch and Conery 2000). If we assume that many genes involved in embryogenesis in both *C. elegans* and *C. briggsae* are inherited from a common ancestor approximately 20–50 MYA, the expected number of gene duplicates in the early development class is of the order of 551–1,379 (1,326 genes $\times$ 0.0208 dup/gene/Myr $\times$ 20–50 Myr) or at least 0.42–1 duplication per gene in each the *C. elegans* and *C. briggsae* genomes.

Thus, more than 40% of all early- (and late-) expressed genes are expected to have duplicated at least once in both the *C. elegans* and *C. briggsae* lineages since their divergence. This estimate is a minimum estimate, as many genes active during and after embryogenesis are likely to be much older than 50 Myr. The proportion of genes with duplicates in the late-expression class (0.35–0.39) falls close to the above estimate. In contrast, the proportion of genes with duplicates in the early-expression class (0.07–0.18) falls well below the estimated 0.42–1.0 duplicates per gene. Given the prodigious rate of gene duplication in *C. elegans*, coupled with the nonspecific mechanisms by which gene duplications are thought to occur, the lower than expected number of duplicates in the early-expression class is likely explained by a biased loss of duplicates of early-expressed genes.

Two alternative hypotheses may explain this biased loss: a duplicate gene may simply not be needed during embryogenesis, whereas late-expressed duplicate genes experience selective divergence for postembryonic roles (for example, Force et al. 1999). Alternatively, a duplicate gene may be actively selected against because of harmful effects caused by disruption of embryogenesis. In the latter case, one expects the number of nonprocessed pseudogenes derived from early-expressed genes to be disproportionately enriched because of the retention of the products of primarily those duplication events that result in nonfunctional duplicates, i.e., those involving partial duplication, frameshifts, and stop codons. Such an enrichment of pseudogenes among early-expressed genes in the genome of *C. elegans* is indeed found. Almost twice as many pseudogenes as expected are found among early-expressed genes ($P < 10^{-5}$), consistent with a selective hypothesis of duplicate loss. Unfortunately, this result is also consistent with the neutral hypothesis under different scenarios, for example, if the rate of selective divergence is extremely high.

Better support for a hypothesis of selective loss of early duplicates is found in the distribution of duplicates of nonmodulated genes. Because nonmodulated genes are expressed early (as well as late in development), these genes are putatively exposed to the same selection pressure as genes expressed uniquely during embryogenesis. Under the hypothesis of developmental constraint, passage through the hypothesized selective bottleneck of embryogenesis, in effect, marks these genes as early, despite their continued presence later in ontogeny. This scenario of purifying selection predicts that the nonmodulated class will have a distribution of duplicates more similar to early-expressed genes than late-expressed genes. The alternative hypothesis, early duplicate neutrality and selective divergence of late duplicates, makes the opposite prediction.

We found that early-expressed and nonmodulated genes have a strikingly similar distribution of duplicates; these distributions are only marginally or nonsignificantly different from each other in the *C. elegans* and *C. briggsae* genomes, respectively (figs. 3 and 4). This distribution of duplicates lends support to the hypothesis of selective loss of duplicates of early-expressed genes over the neutral hypothesis of early duplicate degeneration and late duplicate divergence.

Active selection against duplicates of genes expressed during embryogenesis is compelling evidence for developmental constraint at the level of gene duplication. Because duplicate genes are often held to be the substrate of evolutionary novelty (Force et al. 1999; Lynch and Conery 2000), an inability to retain duplicates of genes expressed during embryogenesis may have important implications for macroevolutionary change. Further investigation into the distribution of duplicates through development in genomes of other phyla with different modes of development is necessary before the generality of this phenomenon and its importance in evolution can be assessed.

Acknowledgments

Special thanks to Justin Blumensteil for supplying the key reference in this work. Additional thanks to Gordon Kindlmann of the University of Utah Scientific Computing and Imaging Institute for computational resources, Etsuko Moriyama for the latest copy of MEA, and Kym Hallsworth-Pepin and WUSTL for *C. briggsae* sequence and information. We thank members of the Hartl lab, Josh Cherry, and Daniel Neafsey for suggestions. This work was supported by NIH grants GM0035

and GM58423 and a NIH Genetics Training Grant to C.I.C.-D.

LITERATURE CITED

ALTSCHUL, S. F., T. L. MADDEN, A. A. SCHAFFER, J. ZHANG, Z. ZHANG, W. MILLER, and D. J. LIPMAN. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389–3402.

ARTHUR, W. 1988. A theory of the evolution of development. John Wiley & Sons, Chichester.

———. 1997. The origin of animal body plans: a study in evolutionary developmental biology. Cambridge University Press, New York.

———. 2000. The concept of developmental reprogramming and the quest for an inclusive theory of evolutionary mechanisms. Evol. Dev. **2**(1):49–57.

BIRD, A. F., and J. BIRD. 1991. The structure of Nematodes. 2nd edition. Academic Press, Harcourt Brace & Jovanovich Publishers, San Diego, Calif.

CIVETTA, A., and R. S. SINGH. 1998. Sex-related genes, directional selection, and speciation. Mol. Biol. Evol. **15**:901–909.

COGHLAN, A., and K. H. WOLFE. 2000. Relationship of codon bias to mRNA concentration and protein length in *Saccromyces cerevisiae*. Yeast **16**:1131–1145.

DUNN, K. A., J. P. BIELAWSKI, and Z. YANG. 2001. Substitution rates in Drosophila nuclear genes: implications for translational selection. Genetics **157**:295–305.

DURET, L., and D. MOUCHIROUD. 1999. Expression pattern and, suprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabadopsis*. Proc. Natl. Acad. Sci. USA **96**:4482–4487.

———. 2001. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol. Biol. Evol. **17**(1):68–74.

FORCE, A., M. LYNCH, F. B. PICKETT, A. AMORES, Y. YAN, and J. POSTLETHWAIT. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics **151**: 1531–1545.

GOLDMAN, N., and Z. YANG. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11**(5):725–726.

GOODWIN, B. C., S. KAUFMANN, and J. D. MURRAY. 1993. Is morphogenesis an intrinsically robust process? J. Theor. Biol. **163**:135–144.

HILL, A. A., C. P. HUNTER, B. T. TSUNG, G. TUCKER-KELLOGG, and E. L. BROWN. 2000. Genomic analysis of gene expression in *C. elegans*. Science **290**:809–812.

KUMAR, S., and S. B. HEDGES. 1998. A molecular timescale for vertebrate evolution. Nature **392**:917–920.

LI, W. H. 1997. Molecular evolution. Sinauer, Sunderland, Mass.

LYNCH, M., and J. S. CONERY. 2000. The evolutionary fate and consequence of duplicate genes. Science **290**:1151–1154.

PARSCH, J., C. D. MEIKLEJOHN, E. HAUSCHTECK-JUNGEN, P. HUNZIKER, and D. L. HARTL. 2001. Molecular evolution of the oncus and janus genes in the *Drosophila melanogaster* species subgroup. Mol. Biol. Evol. **18**:801–811.

RAFF, R. A. 1996. The shape of life. The University of Chicago Press, Chicago, Ill.

RIEDL, R. 1978. Order in living organisms. Wiley, New York.

SHARP, P. M., and W. H. LI. 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. **15**: 1281–1295.

STEIN, L., P. SERNBERG, R. DURBIN, J. THIERRY-MIEG, and J. SPIETH. 2001. Wormbase: network access to the genome and biology of *Caenorhabditis elegans*. Nucleic Acids Res. **29**(1):82–86.

SUCENA, E., and D. L. STERN. 2000. Divergence of larval morphology between *Drosophila sechellia* and its sibling species caused by cis-regulatory evolution of *ovo/shaven-baby*. Proc. Natl. Acad. Sci. USA **97**(9):4530–4534.

TATUSOV, R. L., E. V. KOONIN, and D. J. LIPMAN. 1997. A genomic perspective on protein families. Science **278**:631–637.

USUKA, J., W. ZHU, and V. BRENDEL. 2000. Optimal spliced alignment of homologous cDNA to a genomic DNA template. Bioinformatics **16**:203–211.

WOLFRAM, S. 1999. Mathematica. Version 4.0.1.0. Wolfram Research, Champaign, Ill.

WRIGHT, F. 1990. The 'effective number of codons' used in a gene. Gene **87**:23–29.

WYCOFF, G., W. WANG, and C. I. WU. 2000. Rapid evolution of reproductive proteins in the descent of man. Nature **403**: 304–309.

YANG, Z. 2000. Phylogenetic analysis by maximum likelihood (PAML), Version 3.0. University College London, UK.

PIERRE CAPY, reviewing editor

Accepted January 15, 2002