The Structure and Early Evolution of Recently Arisen Gene Duplicates in the *Caenorhabditis elegans* Genome

Vaishali Katju[1] and Michael Lynch

Department of Biology, Indiana University, Bloomington, Indiana 47405

Manuscript received July 11, 2003
Accepted for publication September 8, 2003

ABSTRACT

The significance of gene duplication in provisioning raw materials for the evolution of genomic diversity is widely recognized, but the early evolutionary dynamics of duplicate genes remain obscure. To elucidate the structural characteristics of newly arisen gene duplicates at infancy and their subsequent evolutionary properties, we analyzed gene pairs with ≤10% divergence at synonymous sites within the genome of *Caenorhabditis elegans*. Structural heterogeneity between duplicate copies is present very early in their evolutionary history and is maintained over longer evolutionary timescales, suggesting that duplications across gene boundaries in conjunction with shuffling events have at least as much potential to contribute to long-term evolution as do fully redundant (complete) duplicates. The median duplication span of 1.4 kb falls short of the average gene length in *C. elegans* (2.5 kb), suggesting that partial gene duplications are frequent. Most gene duplicates reside close to the parent copy at inception, often as tandem inverted loci, and appear to disperse in the genome as they age, as a result of reduced survivorship of duplicates located in proximity to the ancestral copy. We propose that illegitimate recombination events leading to inverted duplications play a disproportionately large role in gene duplication within this genome in comparison with other mechanisms.

THE enormous disparity in the genome sizes of extant organisms is a striking reminder that genic and genome-wide duplications are a ubiquitous and evolutionarily important feature of genomes. Although the evolutionary significance of gene duplicates had been recognized by early geneticists and evolutionary biologists (Haldane 1933; Bridges 1935; Muller 1935, 1936), Ohno's 1970 treatise *Evolution by Gene Duplication* is largely credited with the empirical resurrection and theoretical development of the field. Ohno (1970) maintained that the evolution of new genes and novel biochemical processes could arise only via gene duplication. Although other mechanisms such as alternative splicing, post-transcriptional and post-translational modifications, and regulatory mutations among others can serve to increase the functional diversity of a gene without duplication, the pervasive role of gene duplication in the generation of genomic complexity cannot be denied. Gene duplication in conjunction with domain shuffling has frequently been suggested to play an important role in the origin of novel genes (Long and Langley 1993; Patthy 1994; Begun 1997; Chen *et al.* 1997; Nurminsky *et al.* 1998; Thomson *et al.* 2000). Furthermore, the origin of new gene families with disparate functions from ancestral genes is implicated in the evolution of organismal diversity (Patthy 1999).

[1]*Corresponding author:* Department of Biology, Jordan Hall 142, Indiana University, 1001 E. Third St., Bloomington, IN 47405.
E-mail: vkatju@bio.indiana.edu

Despite a widespread acceptance of the significance of gene duplication and extensive theoretical work relating to aspects of persistence and functionality of duplicated genes, empirical insight into the early evolution of newly arisen gene duplicates has been limited. Past studies of natural populations have revealed a handful of relatively young gene duplicates and intraspecific polymorphism for gene copy number (Maroni *et al.* 1987; Lyckegaard and Clark 1989; Theodore *et al.* 1991; Long and Langley 1993; Lootens *et al.* 1993; Lenormand *et al.* 1998). However, these cases are composed of either serendipitously discovered examples or genes known *a priori* to be of functional significance. The paucity of identifiable young duplicates and the potential bias involved in their identification has precluded statistically robust inferences with regard to the early evolution of gene duplicates. The advent of whole-genome sequencing vastly ameliorates the aforementioned constraints. The complete genomic sequence of an organism can be utilized to identify an unbiased sample of duplicates, thereby providing a large data set for addressing questions pertaining to functionality and persistence of gene duplicates. Two studies exemplify the advantages of such a genome-based approach. Lynch and Conery (2000) used synonymous-site divergence between gene-duplicate copies as a proxy for evolutionary age and arrived at some broad conclusions regarding the birth-death process of gene duplicates and the selective regimes faced by duplicated loci after conception. A more recent study (Gu *et al.* 2003) in *Saccharomyces cerevisiae* demonstrates that the knockout of single-copy

genes results in a greater fitness decline relative to the knockout of one copy of a duplicated pair, implying that gene duplication confers some degree of functional redundancy.

Population-genetic models have been employed to study the evolutionary dynamics of gene duplicates with regard to the probabilities of fixation (Spofford 1969; Ohta 1988b; Clark 1994; Lynch *et al.* 2001), gene silencing (Haldane 1933; Nei and Roychoudhury 1973; Bailey *et al.* 1978; Allendorf 1979; Kimura and King 1979; Takahata and Maruyama 1979; Li 1980; Kimura 1983; Watterson 1983), neofunctionalization (Ohta 1987, 1988a; Hughes 1994; Walsh 1995), subfunctionalization (Lynch and Force 2000; Lynch *et al.* 2001), and the evolution of redundancy (Cooke *et al.* 1997; Nowak *et al.* 1997; Krakauer and Nowak 1999; Wagner 1999). These theoretical models overwhelmingly assume that the process of gene duplication yields a gene copy that is functionally and structurally identical at birth to the progenitor copy. However, if structurally redundant copies comprise only a fraction of the entire set of gene duplicates, a singular focus on the evolutionary dynamics of one structural type of gene duplicate may fail to capture the complexity of the gene duplication process.

To test the common assumption of complete structural resemblance between gene duplicates at birth, we analyzed a population of recent gene duplicates (290 duplicate pairs with 10% or less divergence at synonymous sites) in the *Caenorhabditis elegans* genome. We posed several additional questions with respect to the features of gene duplicates at conception and early in their history. First, to what extent does a duplicated gene copy structurally resemble the progenitor copy at the nucleotide level, and how is this structural homology altered with time? Second, are the introns of the progenitor gene maintained in the duplicate copy and does reverse transcription of processed mRNA play a significant role in the gene duplication process? Third, where do duplicate copies tend to reside at or close to the time of conception, and does their location alter over evolutionary time? Finally, what is the approximate span (length) of the duplicated stretch of DNA?

MATERIALS AND METHODS

Identification of gene duplicates within the *C. elegans* genome: A total of 333 gene-duplicate pairs with K_S (number of substitutions per synonymous site) values ranging from 0.00 to 0.10 within the *C. elegans* genomic data set of Lynch and Conery (2000) were initially selected for inclusion in this analysis. This data set had excluded (i) duplicates belonging to multigene families (more than five family members), (ii) sequences showing similarity to known transposable elements, and (iii) potentially nonfunctional protein sequences that did not start with methionine. As the *C. elegans* genomic sequence had been revised substantially since the initial identification by Lynch and Conery (2000), the identity of each gene within the original data set was confirmed in WormBase (http://www.wormbase.org; Stein *et al.* 2001). Of the initial 333 pairs, 43 were excluded for any one of the following criteria: (i) the putative gene duplicates were found to be isoforms of the same gene, (ii) the sequence report with chromosomal location and other characteristics could not be located for one or both gene copies within WormBase, (iii) one or both of the gene copies were characterized as transposable elements, or (iv) no visible homology was apparent between the purported duplicates at the time of this analysis. Annotations were repeatedly confirmed for accuracy on WormBase during the analysis.

Grouping of gene-duplicate pairs into cohorts: To facilitate the comparison of putatively different cohorts of gene duplicates and discern evolutionary change with increasing sequence divergence, the 290 pairs of gene duplicates were originally classified into six cohorts on the basis of divergence at synonymous sites ($K_S = 0$, $0 < K_S \leq 0.01$, $0.01 < K_S \leq 0.03$, $0.03 < K_S \leq 0.05$, $0.05 < K_S \leq 0.07$, and $0.07 < K_S \leq 0.10$). However, statistical analyses revealed that the salient differences occurred between the putative youngest cohort ($K_S = 0$) and duplicate pairs with $0 < K_S \leq 0.10$. Hence, we present only results for the data set as classified into two cohorts, namely $K_S = 0$ and $0 < K_S \leq 0.10$ (68 and 222 gene-duplicate pairs, respectively).

Accession and analysis of sequences: For each gene within a duplicate pair, two nucleotide sequence files were obtained from WormBase: (i) the unspliced version of the gene containing all exons and introns as well as 1 kb of flanking region in both the 5′ and 3′ directions and (ii) the predicted spliced version lacking introns. In addition, information about genomic location, strand orientation, cDNAs, and potential function was collected from the gene sequence report. All sequence analysis was implemented in the BioEdit Sequence Alignment Editor, Version 5.0.9 (Hall 1999). Initial sequence alignments were performed using CLUSTAL software (Higgins *et al.* 1992) and completed visually. A direct comparison of unspliced and spliced sequences of a gene yielded intron locations. Finally, both gene-duplicate copies were aligned to determine the extent of nucleotide sequence homology throughout the open reading frames (ORFs) and flanking regions. For cases where the homology between duplicates extended beyond 1 kb of the flanking region(s), an additional 1 kb of flanking region(s) was accessed from the database and subsequently aligned. The step of adding and aligning the flanking region(s) nucleotide sequence was iterated until no homology was apparent between the two duplicates for a continuous stretch of 1 kb on either end.

Classification of duplicate pairs into structural categories: On the basis of a direct comparison of the ORF nucleotide sequence of the two copies within a gene-duplicate pair, we classified duplicates as exhibiting one of three categories of structural resemblance, namely (i) *complete*, (ii) *partial*, or (iii) *chimeric* (Figure 1). Gene duplicates exhibiting nucleotide sequence homology between the initiation and the termination codons were categorized as having a complete structure. An assignment to this category is straightforward for duplicates with amino acid sequences of identical length. In the case of gene duplicates with amino acid sequences of differing length, duplicate pairs were designated as complete if the shorter copy exhibited nucleotide sequence homology to the lengthier copy throughout the latter's ORF, irrespective of differently demarcated exon-intron and flanking region(s) boundaries. The disruption of sequence homology between the two copies as a result of indels (including intron loss in one copy) was ignored as long as nucleotide sequence homology between the two copies was resumed within the ORF boundaries of the lengthier reference sequence, before the start of flanking region(s). Another class was composed of duplicate pairs with gene copies of differing amino acid sequence length wherein

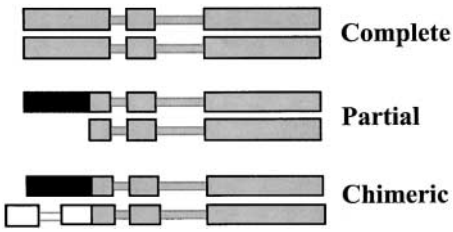

FIGURE 1.—A schematic of three different categories of gene duplicates based on the degree of structural resemblance. Long rectangles denote exons; short rectangles denote introns; correspondence of regions with identical color and pattern between the two duplicates copies reflects sequence homology. Gene duplicates with complete structure share sequence homology throughout their open reading frames from the start to the stop codon and possibly extending into flanking regions. Gene duplicates with partial structure comprise one duplicate copy with unique exons and/or introns that are absent in the other copy. Chimeric structural resemblance requires that both duplicate copies contain unique exons and/or introns to the exclusion of the other gene copy.

the entire ORF of the shorter gene was homologous to the longer gene's ORF, but the longer gene had a unique ORF sequence absent in the shorter gene. These duplicate pairs were classified as exhibiting a partial structure. A third class was composed of duplicate pairs with gene copies of differing amino acid sequence length wherein sequence homology between the two copies was disrupted within the ORFs of both genes, such that both had some unique ORF sequence to the exclusion of the other copy. These were classified as exhibiting a chimeric structure. Simply put, gene-duplicate copies with complete resemblance were homologous over their entire ORFs; those with partial resemblance had one copy with a unique ORF sequence that was absent in the other copy; and those with chimeric resemblance comprised pairs in which both copies had a unique ORF sequence to the exclusion of the other copy. The observed frequencies of the three structural categories of gene duplicates within the two cohorts ($K_S = 0$ and $0 < K_S \leq 0.10$) were compared using a *G*-test (likelihood-ratio test) for goodness of fit (SOKAL and ROHLF 1997).

Two issues deserve mention with respect to our structural classification scheme. First, our nucleotide sequence analysis revealed a frequent occurrence of small insertion/deletions (indels) in one copy relative to the other for duplicate pairs with increasing divergence at synonymous sites. These indels ranged from a few to several hundred base pairs and were located in both the coding and noncoding regions. Insofar as sequence homology between the two duplicate copies was resumed on both sides of the indel within the ORF and flanking regions, it was ignored under the parsimonious assumption that it occurred in the postduplication period as a mutation event and does not accurately reflect structural resemblance at origin. The second issue relates to gene annotation. The exon-intron organization and therefore the structure of many annotated genes within a genome are essentially predicted by computer programs such as Genefinder. Despite sequence homology, two duplicate copies can be assigned different exon-intron boundaries due to either inaccurate predictions by such programs or a disruption of the reading frame in one of the duplicate copies as a result of mutation(s). We frequently encountered cases wherein homologous nucleotide sequences are alternatively depicted as an exon and a flanking region in the two duplicate copies. Similarly, a genic region can be depicted alternatively as an exon and an intron in the two duplicate copies. To avoid the influence of erroneous annotations by gene-predicting programs, our method of structural classification is directly based on comparisons of nucleotide sequences between the start and termination codons of the two duplicates, irrespective of exon-intron predictions. In addition, we collected cDNA information from WormBase for all putative gene duplicates in our data set to indirectly verify the accuracy of gene predictions.

Span of duplication: Another aspect of gene duplication is concerned with understanding the frequency distribution of the span of duplication. In this study, this measure is restricted to duplicated nucleotide stretches containing identifiable open reading frames. The length of sequence homology (in kilobases) between two duplicate copies was taken to be the span of duplication. Of the 290 pairs of gene duplicates, the majority of the cases (276 of 290) involved the duplication of a single gene. However, 7 cases involved the duplication of multiple loci with intervening flanking regions. These linked sets of duplicated loci were treated as a single duplication event and assigned a single value for duplication span. Therefore, the $K_S = 0$ and $0 < K_S \leq 0.10$ cohorts comprise 62 and 221 duplication events, respectively.

As mentioned earlier, duplicate genes with increasing synonymous-site divergence often have numerous indels within their homologous regions, ranging in length from a few to several hundred base pairs. Under these circumstances, we generated two values for duplication span by separately considering each of the two duplicate loci in turn as the ancestral copy. The lower of the two duplication span values was included in the analysis. This may lead to a slight deflation in our estimate of the average length of duplication. For logistic purposes, we had also assumed that the duplication was terminated at the points beyond which no homology was apparent between the two gene copies for a continuous stretch of 1 kb on either end. This methodology is therefore biased against the detection of large indels. In other words, if an insertion/deletion of >1 kb occurred in one copy, we would fail to detect the resumption of homology between the two copies beyond the indel location. This too would lead to a deflation in our estimate of the length of duplication. Therefore, the values reported here are minimal estimates of duplication span.

Physical organization of duplicates residing on the same chromosome: We determined the relative strand orientation and physical organization of gene-duplicate copies located on the same chromosome. Duplicates on the same chromosome were categorized as having direct orientation if the direction of transcriptional orientation was preserved in both copies (*i.e.*, both duplicates were located on the positive strand or on the negative strand). Duplicates with inverse orientation on the same chromosome had one copy each on the positive and negative strands. Additionally, with respect to physical organization on the same chromosome, duplicates were classified as tandem if there were no intervening genes between the two copies or nontandem if intervening gene(s) were present. The observed frequencies of (i) direct *vs.* inverse duplicates and (ii) tandem *vs.* nontandem duplicates across the two cohorts ($K_S = 0$ and $0 < K_S \leq 0.10$) were compared using a *G*-test (likelihood-ratio test) for goodness of fit (SOKAL and ROHLF 1997).

Measures of genomic movement of one duplicate relative to the other: To determine if the genomic location of duplicate copies is altered over evolutionary time, two measures of location and dispersion were calculated as a function of divergence at synonymous sites: (i) the relative frequencies of duplicate pairs residing on the same *vs.* a different chromosome and (ii) the physical distance separating duplicate copies residing on the same chromosome.

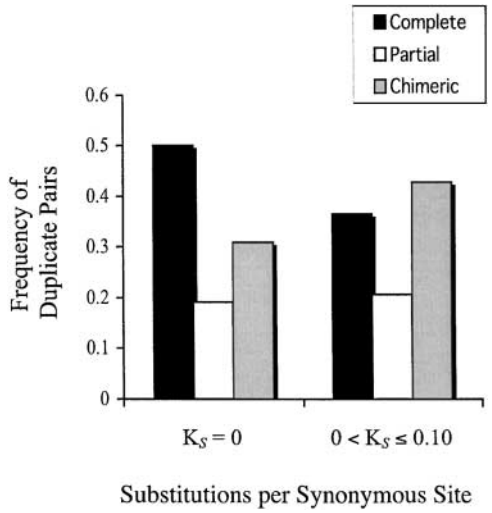

FIGURE 2.—Composition frequencies of three structural categories of gene duplicates within the two cohorts with different divergence at synonymous sites ($K_S = 0$ and $0 < K_S \leq 0.10$).

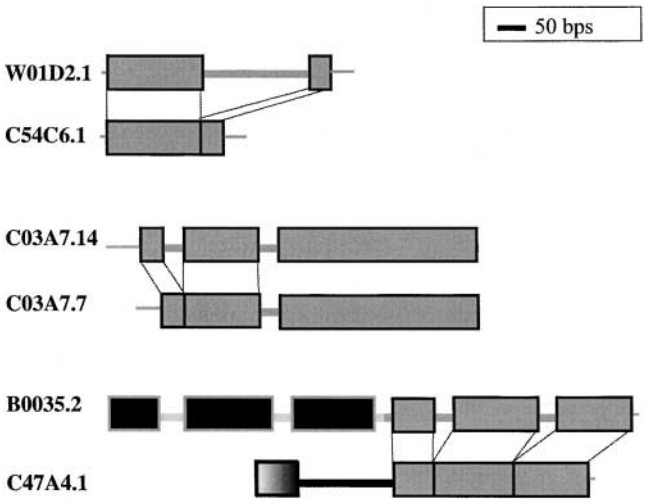

FIGURE 3.—Intron-exon organization of duplicate copies representing three potential cases of duplication by reverse transcription. Designated gene names appear on the left side of the schematic. Long rectangles denote exons; thick lines joining adjacent exons denote introns; thin lines denote the homologous flanking region between the two duplicate copies. Correspondence of regions with identical color and pattern between the two duplicate copies reflects sequence homology. Within each of the three gene-duplicate pairs, the gene copy on the top containing the intron(s) in question is taken as the reference for comparison. Dashed lines joining the two gene duplicates indicate the potential intron loss in the bottom copy relative to the top copy.

With respect to chromosomal location, we calculated the relative frequencies of gene-duplicate pairs with both member copies residing on the same chromosome *vs.* different chromosomes across both cohorts of gene duplicates ($K_S = 0$ and $0 < K_S \leq 0.10$). The observed frequencies of the two categories of chromosomal location across the two cohorts were compared using a *G*-test (likelihood-ratio test) for goodness of fit (SOKAL and ROHLF 1997). In addition, we used a simple logistic regression model (SPSS Version 10) to determine if there is a gradual secondary movement of duplicates to new locations in the genome with increased divergence at synonymous sites. Chromosomal location of both copies within a gene-duplicate pair was coded in a binary fashion: $Y = 0$ if both copies were located on the same chromosome and $Y = 1$ if the two copies resided on different chromosomes. Chromosomal location ($Y = 0$ or 1) was then plotted as a function of synonymous-site divergence between the two duplicate copies.

The physical distance (in base pairs) between duplicate copies on the same chromosome was plotted against synonymous-site divergence between gene duplicates to determine if duplicates on the same chromosome increasingly disperse with evolutionary time, which would be suggestive of intrachromosomal secondary movement by one copy or differential loss in the postduplication period. We independently calculated the correlation coefficient between physical distance and synonymous-site divergence for (i) all 180 gene-duplicate pairs on the same chromosome across both cohorts ($0 \leq K_S \leq 0.10$) and (ii) 125 gene-duplicate pairs within the $0 < K_S \leq 0.10$ cohort only. We tested for a significant sample correlation coefficient by employing (i) the nonparametric Kendall's coefficient of rank correlation test and (ii) the product-moment correlation coefficient (under the assumption of normality).

RESULTS

Early presence of structural heterogeneity between gene duplicates: Structural comparisons revealed that duplicates with partial and chimeric structural resemblance are present in high frequency even within the cohort with no synonymous-site divergence in homologous regions. Together, gene-duplicate pairs exhibiting partial or chimeric structure between the two copies comprise 50 and 64% of all duplicate pairs in the $K_S = 0$ and $0 < K_S \leq 0.10$ cohorts, respectively (Figure 2). A *G*-test for goodness of fit revealed no significant difference in the frequencies of the three structural categories between the two cohorts of gene duplicates ($G_{adj} = 4.24$, d.f. = 2, $0.1 < P < 0.5$).

Currently, 70% (203/290) of all predicted gene-duplicate pairs in our data set have cDNA sequence identified for at least one copy of a gene-duplicate pair. There were no significant differences among structural categories or K_S classes in the frequency of genes for which cDNA has been identified.

Minor role of reverse transcription in the origin of gene duplicates: The structural comparisons of gene duplicates also addressed the extent to which reverse transcription of processed mRNA contributes to gene duplication. Of the 290 gene-duplicate pairs analyzed, 278 were gene duplicates with introns in at least one gene copy. Intron(s) are preserved along the region of homology between the two copies in all but 3 of these 278 cases (~99%; Figure 3).

The first such case involves the duplicate pair C54C6.1/W01D2.1 wherein the two copies have different chromosomal locations and differ by one nonsynonymous substitution and a 3-bp indel (Figure 3). Both genes are members of the ribosomal protein L37 protein family. Gene locus W01D2.1, the lengthier copy, is

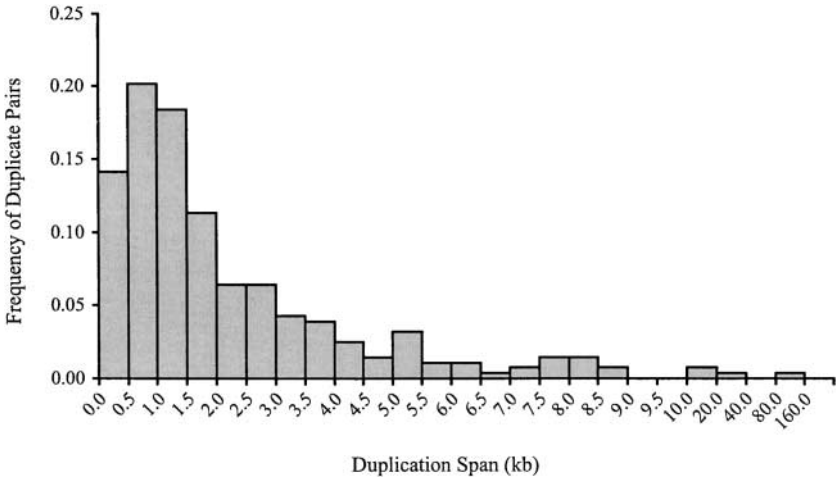

FIGURE 4.—Distribution of duplication spans (in kilobases) for 283 pairs of gene duplicates with 0–10% synonymous-site divergence.

composed of two exons separated by an intron of 300 bp. Locus C54C6.1 is composed of a single exon that is homologous to the two exons of locus W01D2.1, with the precise deletion of the intron.

The second case involves the gene-duplicate pair C03A7.14/C03A7.7 with 5.6% substitutions per synonymous site (Figure 3). The lengthier and shorter loci comprise three and two exons, respectively. Exons 1 and 2 of the lengthier copy (C03A7.14) are fused as one exon minus the intervening intron in the shorter copy (C03A7.7). The two duplicates are separated by nine intervening genes and a physical distance of ~31 kb on chromosome V.

The third case involves the gene-duplicate pair B0035.2/C47A4.1 with 6.9% substitutions per synonymous site (Figure 3). The two loci display a chimeric structure relative to one another, each having unique exons to the exclusion of the other locus. The region of homology toward the 3′ end is composed of three exons with intervening introns in one gene and a single exon minus both introns in the other gene. The two duplicates are separated by a physical distance of 2.4 Mb on chromosome IV.

Predominance of duplications involving short sequence tracts: Figure 4 displays the distribution of duplication spans for all 283 duplication events analyzed. With the exception of four cases involving duplicated clusters of genes spanning ~10.1, 15.8, 23.5, and 108.3 kb, respectively, all duplication span values were <8.7 kb. The L-shape of the distribution implies that duplications involving relatively short tracts of sequence are extremely frequent. In contrast, lengthier duplication events, including partial chromosomal duplications, are relatively rare. In this data set 70% (199/283) of all duplication events resulted in a duplication span of <2 kb. The >0.5- to 1-kb duplication span class has the highest frequency of duplicate pairs (57/283 = 20%), followed by the >1- to 1.5-kb class (52/283 = 18% of all duplicate pairs). The median value for duplication span within this data set was 1419 bp.

Increase in genomic distance between duplicates over evolutionary time: With respect to chromosomal location, we calculated the frequencies of gene-duplicate pairs with both member copies residing (i) on the same chromosome *vs.* (ii) on different chromosomes. Approximately 89% (55/62) of duplicate pairs comprising the $K_S = 0$ cohort had both copies residing on the same chromosome compared to only 56% (125/221) in the $0 < K_S \leq 0.10$ cohort (Figure 5). A *G*-test for goodness of fit revealed chromosomal location to be highly associated with the degree of divergence at synonymous sites ($G_{adj} = 24.6$, d.f. = 1, $P \ll 0.0001$). With respect to physical distance between duplicate copies residing on the same chromosome, Kendall's coefficient of rank correlation revealed a significant positive correlation between physical distance and sequence divergence at synonymous sites if all 180 gene-duplicate pairs are considered ($\tau = 0.317$; $P < 0.0001$; Figure 6). The median physical distance between duplicates residing on the

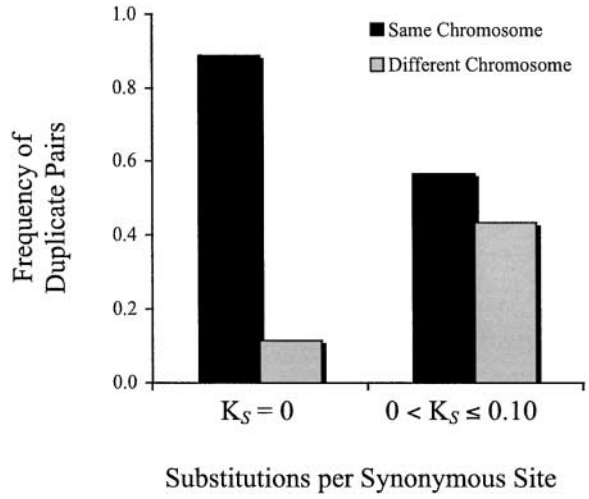

FIGURE 5.—Frequencies of gene-duplicate pairs with both copies residing on the same chromosome *vs.* different chromosomes within the two gene-duplicate cohorts with different synonymous-site divergence ($K_S = 0$ and $0 < K_S \leq 0.10$).

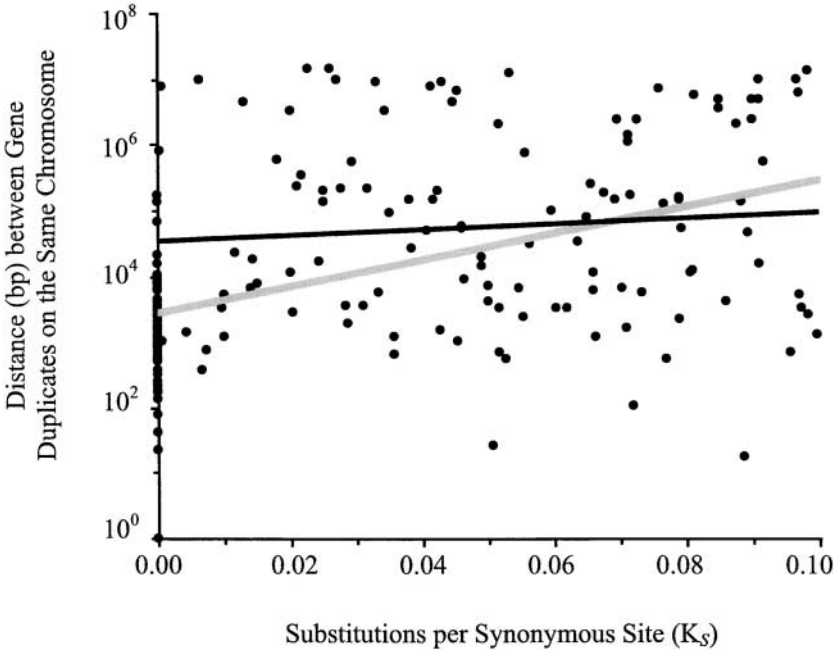

FIGURE 6.—Relationship between physical distance (in base pairs) separating two duplicate copies residing on the same chromosome and divergence at synonymous sites. The solid line was calculated for 125 gene-duplicate pairs residing on the same chromosome in the $0 < K_S \leq 0.10$ cohort ($r = 0.083$; d.f. = 123; $P > 0.05$). The shaded line was calculated for all 180 gene-duplicate pairs residing on the same chromosome, including 55 pairs within the $K_S = 0$ cohort ($r = 0.406$; d.f. = 178; $P < 0.01$).

same chromosome was 1138 and 8644 bp for the $K_S = 0$ and $0 < K_S \leq 0.10$ cohorts, respectively. Therefore, not only are gene duplicates within the $K_S = 0$ cohort more likely to occur on the same chromosome relative to older cohorts, but also they tend to be closely spaced together on the same chromosome (often as tandem loci; see Table 1). Hence, these distance measures are consistent with a pattern of increased genomic distance between gene duplicates over evolutionary time.

A gradual increase in genomic distance (greater likelihood of residence on different chromosomes and/or increased distance between gene copies on the same chromosome) between gene duplicates with increased synonymous-site divergence would support secondary movement by one or both copies in the postduplication period as the mechanism for genomic dispersal. Conversely, a lack of correlation between distance measures and synonymous-site divergence would argue against the hypothesis of secondary movement leading to genomic dispersal of gene duplicates.

Logistic regression analysis on the chromosomal location data found no significant effect of synonymous-site divergence on chromosomal location of gene duplicates (Wald test statistic = 0.181, d.f. = 1, $P = 0.67$). When gene duplicates are broken up into smaller cohorts ($0 < K_S \leq 0.01$, $0.01 < K_S \leq 0.03$, $0.03 < K_S \leq 0.05$, $0.05 < K_S \leq 0.07$, $0.07 < K_S \leq 0.10$), there is a large jump in frequency of duplicates residing on different chromosomes from the $K_S = 0$ to the next cohort ($0 < K_S \leq 0.01$) but no further increase in older cohorts. These results argue against the hypothesis of secondary movement by gene duplicates to different chromosomes with increasing evolutionary time.

Likewise, we find no evidence for a gradual increase in distance between duplicate copies residing on the same chromosome with evolutionary time. As men-

TABLE 1

Total number and frequencies (in parentheses) of gene-duplicate pairs with direct *vs.* inverse orientation within two age cohorts with different synonymous-site divergence ($K_S = 0$ and $0 < K_S \leq 0.10$)

	Tandem orientation		Nontandem orientation		
Duplicate cohort	Direct	Inverse	Direct	Inverse	Total
$K_S = 0$	13 (0.24)	25 (0.45)	4 (0.07)	13 (0.24)	**55**
$0 < K_S \leq 0.10$	16 (0.13)	16 (0.13)	43 (0.34)	50 (0.40)	**125**
Total	**29**	**41**	**47**	**63**	**180**

Data are composed of 180 pairs of gene duplicates with both copies residing on the same chromosome. Within each orientation category, data are further classified as (i) tandem (an absence of intervening genes between the two gene-duplicate copies) *vs.* (ii) nontandem (the presence of intervening genes between the two gene-duplicate copies).

tioned earlier, we found a significant positive relationship between synonymous-site divergence and physical distance between gene duplicates residing on the same chromosome. However, if gene-duplicate pairs within the $K_S = 0$ cohort (55 pairs) are removed from the data set, the correlation between the two variables is no longer evident ($\tau = 0.055$; $P = 0.366$; Figure 6). Significance tests of the sample correlation coefficient under the assumption of normality yielded P values similar to the nonparametric Kendall's coefficient of rank correlation test (see Figure 6). Thus, there is a significant excess of closely spaced gene duplicates in the $K_S = 0$ cohort and this excess alone might have caused the positive correlation between K_S and physical distance between gene duplicates on the same chromosome. The difference in median distance between the $K_S = 0$ and the $0 < K_S \leq 0.10$ is quite dramatic. For gene duplicates with $K_S = 0$, half of the duplicates are within 6.5 kb of each other, whereas half of the duplicates in the $0 < K_S \leq 0.10$ group are within 32 kb of each other. Given that most gene duplicates in the $0 < K_S \leq 0.10$ cohort are still relatively close to each other, the dispersion of duplicates uniformly across a chromosome does not explain the lack of relationship between K_S and distance.

The majority of gene duplicates in the $K_S = 0$ cohort occur on the same chromosome as tandem genes with inverse transcriptional orientation: As demonstrated earlier, the majority of gene-duplicate pairs in the $K_S = 0$ cohort have both gene copies located on the same chromosome (Figure 5). Table 1 represents the relative strand orientation and physical organization of 180 gene-duplicate pairs with both copies on the same chromosome. Within the $K_S = 0$ cohort, we observe the following frequencies: inverse tandem (45%) > direct tandem and inverse nontandem (24% each) > direct nontandem (7%). The following pattern is observed for the $0 < K_S \leq 0.10$ cohort: inverse nontandem (40%) > direct nontandem (34%) > direct tandem and inverse tandem (13% each).

We observed a striking difference between the two cohorts of gene duplicates with respect to strand orientation of the two copies. The $K_S = 0$ cohort of gene duplicates had a twofold excess of duplicate copies in inverse orientation (69%) relative to those exhibiting direct orientation (31%). Within the $0 < K_S \leq 0.10$ cohort, gene duplicates are equally likely to occur in direct *vs.* inverse orientation (47 and 53%, respectively). A *G*-test for goodness of fit comparing the two duplicate cohorts rejects the null hypothesis that the frequencies of strand orientation are independent of sequence divergence at synonymous sites ($G_{adj} = 4.199$, d.f. = 1, $P < 0.05$).

Likewise, the two gene-duplicate cohorts also exhibit differences with respect to physical organization when both copies are present on the same chromosome. The majority (69%) of the 55 gene-duplicate pairs on the same chromosome within the $K_S = 0$ cohort appear as tandemly organized loci. In contrast, the majority (74%) of the 125 gene-duplicate pairs on the same chromosome within the $0 < K_S \leq 0.10$ cohort exhibit a nontandem organization. A *G*-test for goodness of fit comparing the two cohorts rejects the null hypothesis that physical organization on the same chromosome (tandem *vs.* nontandem) is independent of sequence divergence at synonymous sites ($G_{adj} = 30.341$, d.f. = 1, $P \ll 0.001$).

DISCUSSION

We have focused on 290 gene-duplicate pairs in the *C. elegans* genome with <10% sequence divergence at synonymous sites to address questions relating to the structure and genomic location of presumably young gene duplicates and the possible mechanisms of gene duplication. We conducted our analysis under the initial assumption that the number of substitutions per synonymous site (K_S) is an appropriate indicator of the evolutionary age of a duplicate pair, at least for low estimates of K_S. However, concerted evolution, particularly gene conversion, has the potential to homogenize the sequence of previously diverged duplicate copies in relation to one another, so that they appear evolutionarily young. Unfortunately, the methods to detect and test for gene conversion in the absence of a close outgroup sequence do not work when there is high sequence identity between the copies (Sawyer 1989; Maynard-Smith 1992). A partial-genome analysis of duplicate genes within *C. elegans* detected gene conversion events in only 2% of the duplicate pairs, with the majority (85%) of these cases restricted to members of gene families (Semple and Wolfe 1999). If these estimates fairly reflect the frequency of gene conversion in *C. elegans* and the fact that multigene families (more than five gene family members) were excluded in this particular data set of duplicates (Lynch and Conery 2000), there is perhaps not much reason for concern.

Slippage and unequal exchange are expected to result in tandem gene duplicates with direct orientation and these mechanisms are often invoked as an explanation for closely spaced gene duplicates. On the basis of an apparent excess of tandem duplicates in a partial-genome analysis of gene duplicates in *C. elegans*, it was concluded that slippage or unequal crossing over rather than transposition was the primary gene duplication mechanism within this genome (Semple and Wolfe 1999). However, our analysis shows that gene duplicates on the same chromosome across both cohorts are frequently in inverse orientation with respect to one another (58%; Table 1). Furthermore, within the $K_S = 0$ cohort, 69 and 66% of the total and tandem gene duplicates, respectively, are in inverse orientation. Inversion of repeats has been explained by secondary chromosomal rearrangements after duplication (*e.g.*, Achaz *et al.* 2000). Indeed, comparisons of gene order among

genomes have implicated a major role for local-scale gene inversion events in genome evolution (Gilley and Fried 1999; Llorente *et al.* 2000; Seoighe *et al.* 2000; Fischer *et al.* 2001). Nonetheless, secondary rearrangements are unlikely to account for the majority of inverse orientation gene duplicates in the *C. elegans* genome, considering that (i) they are already in high frequency in the youngest cohort and (ii) the frequency of inversely oriented gene duplicates is not increasing with increased synonymous-site divergence. This suggests that inversions are part and parcel of the original duplication event. Inverse orientation gene duplication has also been suggested to be common and to play a role in generating local inversions in Saccharomyces species (Fischer *et al.* 2001) and bacteria (Eisen *et al.* 2000).

Several models of inverted duplications have been proposed, especially in conjunction with the phenomenon of gene amplification in mammalian cells (Passananti *et al.* 1987; Hyrien *et al.* 1988). A structural analysis of inverted duplications in mammalian cells led Passananti *et al.* (1987) to conclude that these did not appear to involve any transposable elements but were instead generated by an illegitimate recombination event. During DNA replication, strand switching by the DNA polymerase can lead to the formation of inverted duplicates (Cohen *et al.* 1994; Bi and Liu 1996; Lin *et al.* 2001). Gordon and Halliday's (1995) simple model of strand misalignment-realignment may also explain the mechanism of formation of inverted duplicates. Under their scenario, sequence complementarity at inverted-repeat sites facilitates the misalignment of the nascent leading strand onto the lagging-strand template and its eventual realignment back onto the leading-strand template, thereby leading to the duplication and inversion of the replicated sequence with respect to its original orientation.

It is quite possible that slippage or unequal exchange does indeed lead to a large number of tandem duplications. Direct tandem repeats, however, are expected to be very unstable and unless under selection from the outset, are easily lost by the very same mechanisms that created them in the first place (Anderson and Roth 1977; Olson 1991; Lovett *et al.* 1994; Galitski and Roth 1997).

Gene-duplicate copies typically reside close to one another in the genome, most often as tandem and inversely oriented genes on the same chromosome. The observation that gene duplicates often reside on the same chromosome has been noted in other eukaryotic genomes as well (Rubin *et al.* 2000). With increasing sequence divergence at synonymous sites, surviving gene-duplicate copies tend either to be farther apart from each other on the same chromosome or to appear on different chromosomes. The observation that distant intrachromosomal repeats tend to be more diverged in sequence led Achaz *et al.* (2000, 2001) to propose a two-phase model wherein intrachromosomal repeats are mostly created in tandem by unequal crossing over or slippage and subsequently made distant by chromosomal rearrangements. For the *C. elegans* genome, Lercher *et al.* (2003) have also described a correlation between sequence similarity of gene duplicates and the distance between them and explained the relationship between the two by secondary movement.

If gene duplicates are being moved apart predominantly by interchromosomal rearrangements (secondary movement), we would expect a progressive increase with age (K_S) in the frequency of duplicate pairs with the two copies located on different chromosomes. The chromosomal location data analyzed here do indeed show a significant enrichment with time of duplicate pairs with the two copies located on different chromosomes. However, the increase in the frequency of gene duplicates on separate chromosomes with increasing sequence divergence is primarily due to the fact that gene duplicates within the $K_S = 0$ cohort are overwhelmingly located on the same chromosome. When these are excluded from the data, no further increase in frequency of gene duplicates occurs on separate chromosomes with increasing synonymous-site divergence. Similar results emerge when the distance between duplicates residing on the same chromosome is analyzed, in that there is no relationship between synonymous-site divergence and distance when the $K_S = 0$ class is excluded. There is no doubt that chromosomal rearrangements (inversions and translocations) occur frequently in the *C. elegans* genome (Coghlan and Wolfe 2002). If later rearrangements are the primary reason for the relationship between K_S and physical distance in the genome, these rearrangements appear to preferentially recognize and translocate duplicate copies onto a different chromosome, or far apart on the same chromosome, in a very narrow evolutionary window ($0 < K_S \leq 0.01$). However, the mechanisms responsible for moving gene duplicates apart presumably cannot distinguish between young and old gene duplicates and stop operating once one of the duplicated pair has been hit by a point mutation.

There are two alternatives to the secondary movement explanation for the relationship between synonymous-site divergence and distance between gene duplicates, namely (i) differential retention of gene duplicates and (ii) gene conversion; the frequency of both may depend on the distance between duplicate copies. First, gene duplicates in genomic proximity to the cognate copy are probably less stable than duplicates far apart. Although all gene duplicates can be lost by a simple deletion, closely spaced gene duplicates can also be lost by slippage or recombination with the cognate partner resulting in unequal exchange. For example, tandem duplications in yeast are known to be extremely unstable, given the high level of homologous recombination within this genome (Olson 1991). For duplicates

spaced farther apart, such homologous exchange would result in the loss of intervening genes and likely would be selected against. In fact, a common way to stabilize duplications in microbial genomes, which would otherwise be prone to rapid loss by homologous recombination, is to insert a gene under selection (such as genes for antibiotic resistance) between the duplicated regions (Galitski and Roth 1997). The difference between the $K_S = 0$ cohort and older duplications could then be primarily because closely spaced duplicates are highly unstable and get lost relatively rapidly unless there is selection to maintain copies immediately or shortly after birth. Second, because closely spaced gene duplicates are more likely to be subject to gene conversion (Petes and Hill 1988; Semple and Wolfe 1999; Drouin 2002), they will appear young for their age and give the impression that older (higher K_S) duplicates have moved apart.

The nontandem duplications in our data set are more likely to occur on the same chromosome than on different chromosomes. This suggests that the duplication mechanisms involve interaction between sites that are in physical proximity in the nucleus. Such mechanisms could involve replicative processes such as "transposition without transposase" (Rappleye and Roth 1997) or topoisomerase-II-mediated illegitimate recombination (Bae *et al.* 1988; Holt *et al.* 2002), both of which can lead to inverted orientation of gene duplicates, spaced at a distance on the same chromosome.

We found only three pairs of gene duplicates for which intron(s) were missing in one member relative to the other copy. Such a condition could result from either the insertion of intron(s) in one copy or their precise deletion in the other duplicate. In each of these three cases, the duplicate copies were located either on different chromosomes or at a considerable distance away from each other on the same chromosome. Since reverse-transcribed genes are expected to randomly reintegrate into the genome, these cases may represent gene duplication by reverse transcription of processed or partially processed mRNA. An analysis of pseudogenes in the *C. elegans* genome (which were excluded from this study) found that only a small fraction (10%) of these appear to be processed (Harrison *et al.* 2001). In contrast, processed pseudogenes comprise 80% of all pseudogenes within the human genome (Dunham *et al.* 1999). Given the concordance of our results with those from an independent study (Harrison *et al.* 2001), we conclude that RNA-mediated transposition is unlikely to play a significant role in gene duplication within the *C. elegans* genome.

The average gene length in *C. elegans* is ~2.5 kb (Duret and Mouchiroud 1999; Vellai and Vida 1999). Within this data set of gene-duplicate pairs, the median duplication span was ~1.4 kb, and 70% (199/283) of all duplication events resulted in a duplication span of <2.5 kb (Figure 4). The L-shaped frequency distribution of duplication spans indicates that, aside from a few lengthy regional duplications, the average duplication event within this genome is fairly localized and spawns relatively short tracts of duplicate sequence that may not encompass entire genes. These results lend credence to the idea that partial gene duplications are to be expected (Averof *et al.* 1996).

The mechanisms responsible for gene duplication (except for reverse transcription) are unlikely to respect gene boundaries. Many, if not most, gene duplications should therefore include gene fractions rather than complete copies, resulting in either a partial copy of the original or a chimeric gene fusion of a partial copy to another gene. This hypothesis is bolstered by our duplication span analysis demonstrating that the median duplication tract falls short of the average gene length in *C. elegans*. Furthermore, our structural comparison results (see below) are consistent with the idea that incomplete gene duplications are common.

We compared the ORF nucleotide sequences of both duplicates to determine the extent of sequence homology between them. Our results indicate that mosaicism or structural heterogeneity between duplicate copies is visible very early in their evolutionary history, if not at birth. Approximately half of the *C. elegans* gene duplicates within both the $K_S = 0$ and $0 < K_S \leq 0.10$ cohorts have unique coding region sequence to the exclusion of the other copy, in addition to the region of homology. To what degree such partial or chimeric gene duplicates contribute to the creative process in evolution by gene duplication is an important question. Some partial gene duplications could be maintained by a process of duplication, degeneration, and conservation (Force *et al.* 1999; Lynch and Force 2000). Under this scenario, a deleterious mutation in the parental gene of a duplicate pair can be compensated for by its partial cognate copy and this in turn would lead to conservation of both copies. Partial duplication may also free different domains from constraints of universal coexpression if separate domains of a protein are useful under different conditions.

The creative potential of chimeric duplicates is well appreciated in the context of evolution of organismal diversity. For example, the demands imposed by a multicellular existence in metazoans were met by an enormous assemblage of novel animal-specific proteins that arose as a result of partial/chimeric duplications in conjunction with shuffling events (Doolittle 1985; Patthy 1985). However, the relative role of complete gene duplicates followed by gradual accumulation of point mutations *vs.* partial or chimeric gene duplications is not well understood, the latter having the potential to create genes with radically different functions from their predecessors. Although most of the theoretical work has been directed at complete gene duplicates that are essentially redundant at birth, it may not accurately reflect on the relative importance of different types of

duplications. The frequency of different structural categories of *C. elegans* gene duplicates that were the subject of this study did not change radically with increasing synonymous divergence (19% partials and 31% chimerics in the $K_S = 0$ cohort *vs.* 21% partials and 43% chimerics in the $0 < K_S \leq 0.10$ cohort). Taken at face value, this means that partial and chimeric gene duplicates not only are present at "birth," but also have at least as much potential to contribute to long-term evolution as do complete gene duplicates. Of course, it is also possible that gene duplicates do not stay within their structural category. For example, if many partial and chimeric gene duplicates were nonfunctional and had no initial evolutionary potential, their loss could have been countered by the creation of new partial and chimeric genes from complete duplicates. The relationship between a gene duplicate's structure at birth and its future evolutionary potential remains to be determined.

We thank Ulfar Bergthorsson for critical reading of the manuscript and are grateful to two anonymous reviewers for helpful comments on the manuscript. We give a special thanks to Sarah Otto, the communicating editor, for extremely insightful and constructive suggestions. This research has been supported by a National Science Foundation Integrative Graduate Education and Research Traineeship Program in Evolution, Development, and Genomics graduate fellowship to V.K. and a National Institutes of Health grant RO1-GM36827 to M.L.

LITERATURE CITED

Achaz, G., E. Coissac, A. Viari and P. Netter, 2000 Analysis of intrachromosomal duplications in yeast *Saccharomyces cerevisiae*: a possible model for their origin. Mol. Biol. Evol. **17:** 1268–1275.

Achaz, G., P. Netter and E. Coissac, 2001 Study of intrachromosomal duplications among the eukaryote genomes. Mol. Biol. Evol. **18:** 2280–2288.

Allendorf, F., 1979 Rapid loss of duplicate gene expression by natural selection. Heredity **43:** 247–258.

Anderson, R. P., and J. R. Roth, 1977 Tandem genetic duplications in phage and bacteria. Annu. Rev. Microbiol. **31:** 473–505.

Averof, M., R. Dawes and D. Ferrier, 1996 Diversification of arthropod Hox genes as a paradigm for the evolution of gene functions. Semin. Cell Dev. Biol. **7:** 539–551.

Bae, Y.-S., I. Kawasaki, H. Ikeda and L. F. Liu, 1988 Illegitimate recombination mediated calf thymus DNA topoisomerase II *in vitro*. Proc. Natl. Acad. Sci. USA **85:** 2076–2080.

Bailey, G. S., R. T. Poulter and P. A. Stockwell, 1978 Gene duplication in tetraploid fish: model for gene silencing at unlinked duplicated loci. Proc. Natl. Acad. Sci. USA **75:** 5575–5579.

Begun, D. J., 1997 Origin and evolution of a new gene descended from alcohol dehydrogenase in Drosophila. Genetics **145:** 375–382.

Bi, X., and L. F. Liu, 1996 DNA rearrangement mediated by inverted repeats. Proc. Natl. Acad. Sci. USA **93:** 819–823.

Bridges, C. B., 1935 Salivary chromosome maps. J. Hered. **26:** 60–64.

Chen, L., A. L. DeVries and C. H. Cheng, 1997 Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. Proc. Natl. Acad. Sci. USA **94:** 3811–3816.

Clark, A. G., 1994 Invasion and maintenance of a gene duplication. Proc. Natl. Acad. Sci. USA **91:** 2950–2954.

Coghlan, A., and K. H. Wolfe, 2002 Fourfold faster rate of genome rearrangement in nematodes than in *Drosophila*. Genome Res. **16:** 857–867.

Cohen, A., D. Hassin, S. Karby and S. Lavi, 1994 Hairpin structures are the primary amplification products: a novel mechanism for generation of inverted repeats during gene amplification. Mol. Cell. Biol. **14:** 7782–7791.

Cooke, J., M. A. Nowak, M. Boerlijst and J. Maynard-Smith, 1997 Evolutionary origins and maintenance of redundant gene expression during metazoan development. Trends Genet. **13:** 360–364.

Doolittle, R. F., 1985 The geneaology of some recently evolved vertebrate proteins. Trends Biochem. Sci. **10:** 233–237.

Drouin, G., 2002 Characterization of the gene conversions between the multigene family members of the yeast genome. J. Mol. Evol. **55:** 14–23.

Dunham, I., N. Shimizu, B. A. Roe, S. Chissoe, A. R. Hunt *et al.*, 1999 The DNA sequence of human chromosome 22. Nature **402:** 489–495.

Duret, L., and D. Mouchiroud, 1999 Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila* and *Arabidopsis*. Proc. Natl. Acad. Sci. USA **96:** 4482–4487.

Eisen, J. A., J. F. Heidelberg, O. White and S. L. Salzberg, 2000 Evidence for symmetric chromosomal inversions around the replication origin in bacteria. Genome Biol. **1:** 0011.0011–0011.0019.

Fischer, G., C. Neuveglise, P. Durrens, C. Gaillardin and B. Dujon, 2001 Evolution of gene order in the genomes of two related yeast species. Genome Res. **11:** 2009–2019.

Force, A., M. Lynch, F. Bryan Pickett, A. Amores, Y. Yan *et al.*, 1999 Preservation of duplicate genes by complementary degenerative mutations. Genetics **151:** 1531–1545.

Galitski, T., and J. R. Roth, 1997 Pathways of homologous recombination between chromosomal direct repeats in *Salmonella typhimurium*. Genetics **146:** 751–767.

Gilley, J., and M. Fried, 1999 Extensive gene order differences within regions of conserved synteny between the *Fugu* and human genomes: implications for chromosomal evolution and the cloning of disease genes. Hum. Mol. Genet. **8:** 1313–1320.

Gordon, A. J., and J. A. Halliday, 1995 Inversions with deletions and duplications. Genetics **140:** 411–414.

Gu, Z., L. M. Steinmetz, X. Gu, C. Scharfe, R. W. Davis *et al.*, 2003 Role of duplicate genes in genetic robustness against null mutations. Nature **421:** 63–66.

Haldane, J. B. S., 1933 The part played by recurrent mutation in evolution. Am. Nat. **67:** 5–19.

Hall, T. A., 1999 BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/99/NT. Nucleic Acids Symp. Ser. **41:** 95–98.

Harrison, P. M., N. Echols and M. B. Gerstein, 2001 Digging for dead genes: an analysis of the characteristics of the pseudogene population in the *Caenorhabditis elegans* genome. Nucleic Acids Res. **29:** 818–830.

Higgins, D. G., A. J. Bleasby and R. Fuchs, 1992 CLUSTAL V: improved software for multiple sequence alignment. Comput. Appl. Biosci. **8:** 189–191.

Holt, S. J., W. A. Cress and J. Van Staden, 2002 Evidence for dynamic alteration in histone gene clusters of *Caenorhabditis elegans*: A topoisomerase II connection? Genet. Res. **79:** 11–22.

Hughes, A. L., 1994 The evolution of functionally novel proteins after gene duplication. Proc. R. Soc. Lond. Ser. B Biol. Sci. **256:** 119–124.

Hyrien, O., M. Debatisse, G. Buttin and B. Robert de Saint Vincent, 1988 The multicopy appearance of a large inverted duplication and the sequence at the inversion joint suggest a new model for gene duplication. EMBO J. **7:** 407–417.

Kimura, M., 1983 *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge, UK.

Kimura, M., and J. L. King, 1979 Fixation of a deleterious allele at one of two "duplicate" loci by mutation pressure and random drift. Proc. Natl. Acad. Sci. USA **76:** 2858–2861.

Krakauer, D. C., and M. A. Nowak, 1999 Evolutionary preservation of redundant duplicated genes. Semin. Cell Dev. Biol. **10:** 555–559.

Lenormand, T., T. Guillemaud, D. Bourguet and M. Raymond, 1998 Appearance and sweep of a gene duplication: adaptive response and potential for new functions in the mosquito *Culex pipiens*. Evolution **52:** 1705–1712.

Lercher, M. J., T. Blumenthal and L. D. Hurst, 2003 Coexpression of neighboring genes in *Caenorhabditis elegans* is mostly due to operons and duplicate genes. Genome Res. **13:** 238–243.

Li, W. H., 1980 Rate of gene silencing at duplicate loci: a theoretical study and interpretation of data from tetraploid fishes. Genetics **95:** 237–258.

Lin, C. T., W. H. Lin, Y. L. Lyu and J. Whang-Peng, 2001 Inverted

repeats as genetic elements for promoting DNA inverted duplication: implications in gene amplification. Nucleic Acids Res. **29:** 3529–3538.

Llorente, B., A. Malpertuy, C. Neuveglise, J. de Montigny, M. Aigle *et al.*, 2000 Genomic exploration of the hemiascomycetous yeasts: 18. Comparative analysis of chromosome maps and synteny with *Saccharomyces cerevisiae*. FEBS Lett. **487:** 101–112.

Long, M., and C. H. Langley, 1993 Natural selection and the origin of *jingwei*, a chimeric processed functional gene in *Drosophila*. Science **260:** 91–95.

Lootens, S., J. Burnett and T. B. Friedman, 1993 An intraspecific gene duplication polymorphism of the urate oxidase gene of *Drosophila virilis*: a genetic and molecular analysis. Mol. Biol. Evol. **10:** 635–646.

Lovett, S. T., T. J. Gluckman, P. J. Simon, V. J. Sutera and P. T. Drapkin, 1994 Recombination between repeats in *Escherichia coli* by a recA-independent, proximity-sensitive mechanism. Mol. Gen. Genet. **245:** 294–300.

Lyckegaard, E. M., and A. G. Clark, 1989 Ribosomal DNA and *Stellate* gene copy number variation on the Y chromosome of *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **86:** 1944–1948.

Lynch, M., and J. Conery, 2000 The evolutionary fate and consequences of duplicate genes. Science **290:** 1151–1155.

Lynch, M., and A. Force, 2000 The probability of duplicate gene preservation by subfunctionalization. Genetics **154:** 459–473.

Lynch, M., M. O'Hely, B. Walsh and A. Force, 2001 The probability of preservation of a newly arisen gene duplicate. Genetics **159:** 1789–1804.

Maroni, G., J. Wise, J. E. Young and E. Otto, 1987 Metallothionein gene duplications and metal tolerance in natural populations of *Drosophila melanogaster*. Genetics **117:** 739–744.

Maynard-Smith, J., 1992 Analyzing the mosaic structure of genes. J. Mol. Evol. **34:** 126–129.

Muller, H. J., 1935 The origination of chromatin deficiencies as minute deletions subject to insertion elsewhere. Genetica **17:** 237–252.

Muller, H. J., 1936 Bar duplication. Science **83:** 528–530.

Nei, M., and A. K. Roychoudhury, 1973 Probability of fixation of nonfunctional genes at duplicate loci. Am. Nat. **107:** 362–372.

Nowak, M. A., M. C. Boerlijst, J. Cooke and J. M. Smith, 1997 Evolution of genetic redundancy. Nature **388:** 167–171.

Nurminsky, D. I., M. V. Nurminskaya, D. De Aguiar and D. L. Hartl, 1998 Selective sweep of a newly evolved sperm-specific gene in Drosophila. Nature **396:** 572–575.

Ohno, S., 1970 *Evolution by Gene Duplication*. Springer-Verlag, New York.

Ohta, T., 1987 Simulating evolution by gene duplication. Genetics **115:** 207–213.

Ohta, T., 1988a Further simulation studies on evolution by gene duplication. Evolution **42:** 375–386.

Ohta, T., 1988b Time for acquiring a new gene by duplication. Proc. Natl. Acad. Sci. USA **85:** 3509–3512.

Olson, M. V., 1991 Genome structure and organization in *Saccharomyces cerevisiae*, pp. 1–39 in *The Molecular and Cellular Biology of the Yeast Saccharomyces: Genome Dynamics, Protein Synthesis, and Energetics*, edited by J. R. Broach, J. R. Pringle and E. W. Jones. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Passananti, C., B. Davies, M. Ford and M. Fried, 1987 Structure of an inverted duplication formed as a first step in a gene amplification event: implications for a model of gene amplification. EMBO J. **6:** 1697–1703.

Patthy, L., 1985 Evolution of the proteases of blood coagulation and fibrinolysis by assembly from modules. Cell **41:** 657–663.

Patthy, L., 1994 Introns and exons. Curr. Opin. Struct. Biol. **4:** 383–392.

Patthy, L., 1999 Genome evolution and the evolution of exon-shuffling: a review. Gene **238:** 103–114.

Petes, T. D., and C. W. Hill, 1988 Recombination between repeated genes in microorganisms. Annu. Rev. Genet. **22:** 147–168.

Rappleye, C. A., and J. R. Roth, 1997 Transposition without transposase: a spontaneous mutation in bacteria. J. Bacteriol. **179:** 2047–2052.

Rubin, G. M., M. D. Yandell, J. R. Wortman, G. L. Gabor Miklos, C. R. Nelson *et al.*, 2000 Comparative genomics of the eukaryotes. Science **287:** 2204–2215.

Sawyer, S., 1989 Statistical tests for detecting gene conversion. Mol. Biol. Evol. **6:** 526–538.

Semple, C., and K. H. Wolfe, 1999 Gene duplication and gene conversion in the *Caenorhabditis elegans* genome. J. Mol. Evol. **48:** 555–564.

Seoighe, C., N. Federspiel, T. Jones, N. Hansen, V. Bivolarovic *et al.*, 2000 Prevalence of small inversions in yeast gene order evolution. Proc. Natl. Acad. Sci. USA **97:** 14433–14437.

Sokal, R. R., and F. J. Rohlf, 1997 *Biometry: The Principles and Practice of Statistics in Biological Research*. W. H. Freeman, San Francisco.

Spofford, J. B., 1969 Heterosis and the evolution of duplications. Am. Nat. **103:** 407–432.

Stein, L., P. Sternberg, R. Durbin, J. Thierry-Mieg and J. Spieth, 2001 WormBase: network access to the genome and biology of *Caenorhabditis elegans*. Nucleic Acids Res. **29:** 82–86.

Takahata, N., and T. Maruyama, 1979 Polymorphism and loss of duplicate gene expression: a theoretical study with application of tetraploid fish. Proc. Natl. Acad. Sci. USA **76:** 4521–4525.

Theodore, L., A. S. Ho and G. Maroni, 1991 Recent evolutionary history of the metallothionein gene Mtn in *Drosophila*. Genet. Res. **58:** 203–210.

Thomson, T. M., J. J. Lozano, N. Loukili, R. Carrio, F. Serras *et al.*, 2000 Fusion of the human gene for the polyubiquitination coeffector UEV1 with Kua, a newly identified gene. Genome Res. **10:** 1743–1756.

Vellai, T., and G. Vida, 1999 The origin of eukaryotes: the difference between prokaryotic and eukaryotic cells. Proc. R. Soc. Lond. Ser. B Biol. Sci. **266:** 1571–1577.

Wagner, A., 1999 Redundant gene functions and natural selection. J. Evol. Biol. **12:** 1–16.

Walsh, J. B., 1995 How often do duplicated genes evolve new functions? Genetics **139:** 421–428.

Watterson, G. A., 1983 On the time for gene silencing at duplicate loci. Genetics **105:** 745–766.

Communicating editor: S. P. Otto