The enormous disparity in the genome sizes of extant organisms is a striking reminder that genetic 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).

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 (Spoorford 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), neo-functionalization (Ohta 1987, 1988a; Hughes 1994; Walsh 1995), sub-functionalization (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 ($K_S = 0, 0 < K_S \leq 0.01, 0.01 < K_S \leq 0.05, 0.05 < K_S \leq 0.07$, 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 $0 < K_S \leq 0.10$.

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
Gene Duplicates in *C. elegans*

1. **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.

2. **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 or negative strand). Duplicates with inverse orientation 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 genomic 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.

3. **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 $0 < K_s \leq 0.10$).

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 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
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
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**

<table>
<thead>
<tr>
<th>Duplicate cohort</th>
<th>Tandem orientation</th>
<th>Nontandem orientation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>Inverse</td>
</tr>
<tr>
<td>( K_s = 0 )</td>
<td>13 (0.24)</td>
<td>25 (0.45)</td>
</tr>
<tr>
<td>( 0 &lt; K_s \leq 0.10 )</td>
<td>16 (0.13)</td>
<td>16 (0.13)</td>
</tr>
<tr>
<td>Total</td>
<td>29</td>
<td>41</td>
</tr>
</tbody>
</table>

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 $K_0 = 0$ 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 $K_0 = 0$ cohort and this excess alone might have caused the positive correlation between $K_0$ and physical distance between gene duplicates on the same chromosome. The difference in median distance between the $K_0 = 0$ and the $0 < K_0 \leq 0.10$ is quite dramatic. For gene duplicates with $K_0 = 0$, half of the duplicates are within 6.5 kb of each other, whereas half of the duplicates in the $0 < K_0 \leq 0.10$ group are within 32 kb of each other. Given that most gene duplicates in the $0 < K_0 \leq 0.10$ cohort are still relatively close to each other, the dispersion of duplicates uniformly across a chromosome does not explain the lack of relationship between $K_0$ and distance.

The majority of gene duplicates in the $K_0 = 0$ cohort occur on the same chromosome as tandem genes with inverse transcriptional orientation: As demonstrated earlier, the majority of gene-duplicate pairs in the $K_0 = 0$ 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 $K_0 = 0$ 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 $0 < K_0 \leq 0.10$ 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 $K_0 = 0$ cohort of gene duplicates had a twofold excess of duplicate copies in inverse orientation (69%) relative to those exhibiting direct orientation (31%). Within the $0 < K_0 \leq 0.10$ 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 $K_0 = 0$ cohort appear as tandemly organized loci. In contrast, the majority (74%) of the 125 gene-duplicate pairs on the same chromosome within the $0 < K_0 \leq 0.10$ 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 ($S$) is an appropriate indicator of the evolutionary age of a duplicate pair, at least for low estimates of $K_0$. 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 $K_0 = 0$ 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 movement, we would expect a progressive increase of inversely oriented gene duplicates is not increasing if gene duplicates are being moved apart predominantly by interchromosomal rearrangements (secondary movement). The observation that distant duplications in yeast are known to be extremely unstable and unless under selection from the one of the duplicated pair has been hit by a point mutation, given the high level of homologous recombination 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*) 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* = 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* = 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* 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* ≤ 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 other-
wise be prone to rapid loss by homologous recombi-
ation, is to insert a gene under selection (such as genes
for antibiotic resistance) between the duplicated re-
geions (Galitski and Roth 1997). The difference be-
tween the $K_i = 0$ 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 conver-
sion (Petes and Hill 1988; Semple and Wolfe 1999; 
Drouin 2002), they will appear young for their age and
give the impression that older (higher $K_i$) duplicates
have moved apart.

The nontandem duplications in our data set are more
likely to occur on the same chromosome than on differ-
ent 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 “transpo-
sition without transposiase” (Rappleye and Roth 1997)
or topoisomerase-II-mediated illegitimate recombi-
tion (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
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 re-
integrate into the genome, these cases may represent
gene duplication by reverse transcription of processed
or partially processed mRNA. An analysis of pseudo-
dgenes 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 $\sim 2.5$ kb
(Duret and Mouchiroud 1999; Vellai and Vida
1999). Within this data set of gene-duplicate pairs, the
median duplication span was $\sim 1.4$ 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 me-
dian duplication tract falls short of the average gene
length in C. elegans. Furthermore, our structural com-
parison 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 homol-
ogy 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 dupli-
cates within both the $K_i = 0$ and $0 < K_i \leq 0.10$ 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 duplica-
tion, 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 do-
 mains from constraints of universal coexpression if sepa-
rate domains of a protein are useful under different
conditions.

The creative potential of chimeric duplicates is well
appreciated in the context of evolution of organisms
diversity. For example, the demands imposed by a multi-
cellular existence in metazoans were met by an enor-
mous assemblage of novel animal-specific proteins that
arose as a result of partial/chimeric duplications in con-
junction with shuffling events (Doolittle 1985; Pat-
thy 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 re-
fect 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 $K_e = 0$ cohort vs. 21% partials and 43% chimerics in the $0 < K_e \leq 0.10$ 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 Berghorston 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 ROI-GM06827 to M.L.

**LITERATURE CITED**


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 Energet...