On the Formation of Novel Genes by Duplication in the *Caenorhabditis elegans* Genome

*Vaishali Katju**† and Michael Lynch**

*Department of Biology, Indiana University; and †Department of Biology, University of New Mexico

Gene duplication is thought to play the singular most important role in the formation of novel genes. The canonical model of gene duplication postulates that novel genes arise in a two-step fashion, namely, (1) the complete duplication of a gene followed by (2) the gradual accumulation of mutations in one or both copies leading to an altered function. It was previously demonstrated that more than 50% of newborn duplicates in *Caenorhabditis elegans* had unique exons in one or both members of a duplicate pair, indicating that many duplicates are not functionally identical to the progenitor copy at birth. Both partial and chimeric gene duplications contribute to the formation of novel genes. For chimeric duplications, the genomic sources of unique exons are diverse, including genic and intergenic regions, as well as repetitive elements. These novel genes derived from partial and chimeric duplications are equally likely to be transcriptionally active as copies derived from complete duplications of the ancestral gene. Duplication breakpoints in the ancestral copies are uniformly distributed in the genome, ruling out the role of any mechanism that restricts them to a particular type of sequence such as introns. Finally, both intron loss and gain contribute to the differential distribution of introns between two copies.

Introduction

The origin of novel genes is of inherent interest to evolutionary biologists, given their significance to adaptive evolution and organismal diversity. Multiple avenues leading to the creation of novel genes are now known, and these include shuffling events and gene duplication. Among these diverse mechanisms, gene duplication is thought to play the singular most important role in the formation of novel genes (Ohno 1970, 1973; Kimura and Ohta 1974; Li and Gojobori 1983; Hughes 1994).

How do duplicated genes proceed to evolve a new function? The canonical model of gene duplication postulates that a complete duplication of a preexisting gene yields a gene copy that is completely redundant to the ancestral copy with respect to sequence and functionality (Ohno 1973). While one of the two copies is thought to be under selective constraints to maintain the ancestral function, the other copy is often assumed to be free to neutrally accumulate additional mutations. The most common fate of a redundant gene duplicate is presumably nonfunctionalization (gene silencing), given that the majority of newly occurring mutations are deleterious by nature. Alternatively, both copies may independently accumulate deleterious mutations to the extent that they partition the ancestral function, commonly referred to as subfunctionalization (Force et al. 1999; Lynch and Force 2000). A third less probable fate may be the acquisition of new mutations that lead to neofunctionalization. Irrespective of the ultimate fate of the redundant copy, classical models of gene duplication largely assume that (1) duplication yields a completely redundant gene and (2) a divergence in the functionality of the progenitor and duplicated locus is achieved by gradual sequence divergence of one copy due to minor tinkering by point mutations or indels. Indeed, there are examples of gene duplicates diverging in function in this manner, such as visual pigment proteins in primates (S. Yokoyama and R. Yokoyama 1990; Radlwimmer and Yokoyama 1998) and pancreatic ribonuclease genes in colobine primates (Zhang, Rosenberg, and Nei 1998).

However, gene duplication can spawn several categories of duplicate loci with varying degrees of structural resemblance to the ancestral locus at conception. Partial gene duplications alone or in conjunction with shuffling events, internal duplications, gene fusion, indels, and integration of retroposed sequences have the potential to refashion a novel gene with a widely divergent intron-exon structure from the ancestral copy (Long and Langley 1993; Chen, DeVries, and Cheng 1997; Nurminsky et al. 1998; Thomson et al. 2000; Courseaux and Nahon 2001; Wang et al. 2002; Hirotsune et al. 2003). It has been argued that gradual divergence between complete duplicates can only lead to a minor change in function, whereas partial duplications in conjunction with gene fusion and shuffling events can lead to an immediate acquisition of a novel function conferring a great selective advantage (Patthy 1999). Duplications of this nature have largely been ignored by the body of population genetic theory dealing with the persistence and functionality of gene duplicates. Additionally, the majority of studies focusing on the early evolution of gene duplicates seldom provide a detailed characterization of the extent of structural homology between the two duplicate copies. A recent structural analysis of evolutionarily young gene duplicates in the *Caenorhabditis elegans* genome found that in more than 50% of the cases examined, the two paralogs exhibited structural heterogeneity with respect to their open reading frames (ORFs) (Katju and Lynch 2003). Moreover, there was no significant difference in the relative frequencies of these heterogeneous duplicates between the newborn ($K_S = 0$) and older cohort ($0 < K_S \leq 0.10$) of gene duplicates, suggesting that they may have as much potential to contribute to long-term evolution as do fully redundant (complete) duplicates.

This study focused on a subset of *C. elegans* gene duplicates with $0 \leq K_S \leq 0.10$, wherein the two paralogs exhibited structural differences in their ORFs. We posed five questions to further elucidate the processes involved in the

Key words: *Caenorhabditis*, chimeric duplication, gene duplication, partial duplication, partial duplication with recruitment.

E-mail: vkatju@unm.edu

Mol. Biol. Evol. 23(5):1056–1067. 2006
doi:10.1093/molbev/msj114
Advance Access publication February 24, 2006

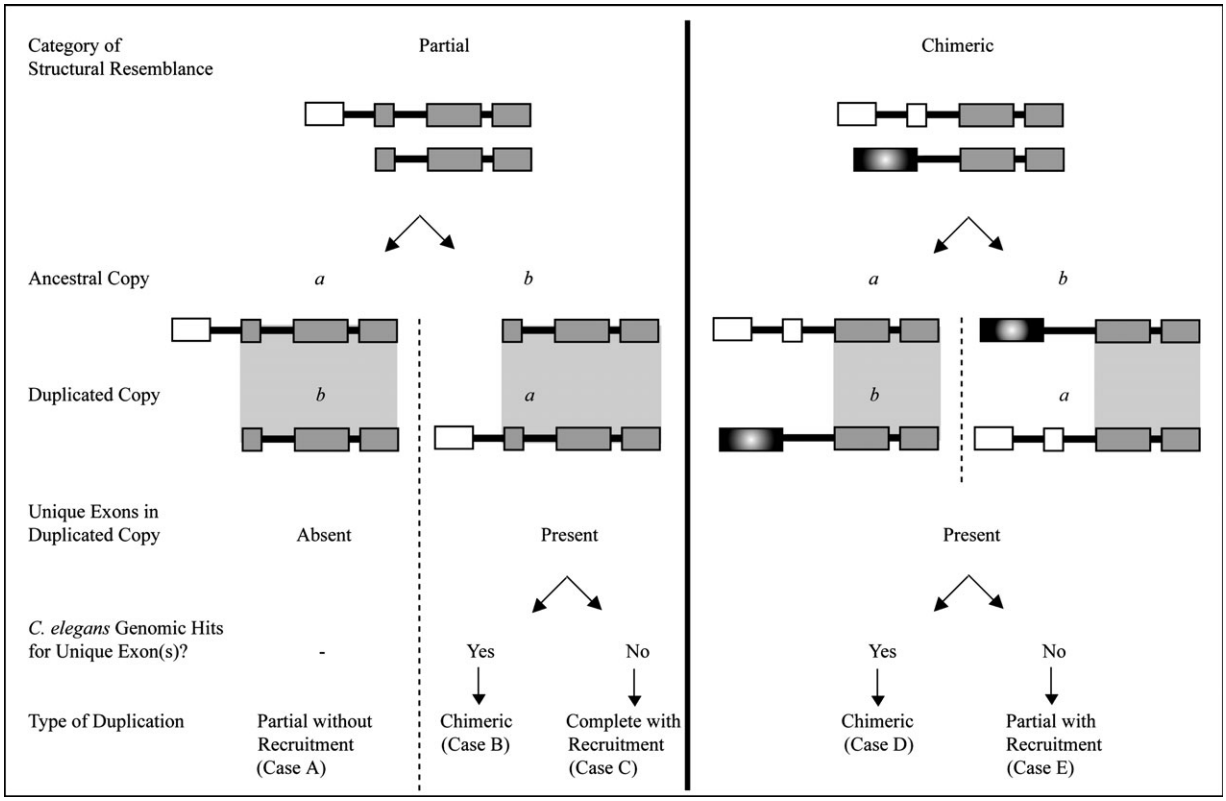

FIG. 1.—Schematic outlining the rationale and methods employed for the identification of the type of duplication event contributing to the formation of a novel duplicate gene. Large rectangles denote exons; narrow rectangles indicate introns. Correspondence of exons with identical color and pattern between the two copies denotes sequence homology. Duplicate pairs are first classified into two categories, namely, those with partial and chimeric structure. The two duplicate copies are labeled *a* and *b*. For each structural category, two scenarios are considered, namely, (1) copy *a* is ancestral or (2) copy *b* is ancestral. Identification of the ancestral copy in a duplicate pair after comparison with a *Caenorhabditis briggsae* ortholog and subsequent characterization of the duplication breakpoints helps determine the portion of the ancestral gene that was duplicated (area shaded gray and encompassing the two copies [ancestral and duplicated]). The next two steps determine if the duplicated copy contains any unique exons to the exclusion of the ancestral copy and when present, if these yield a hit in the *Caenorhabditis elegans* genome. Finally, based on these results, the type of duplication event is characterized as (1) partial without recruitment (Case A), (2) chimeric (Cases B and D), (3) complete with recruitment (Case C), or (4) partial with recruitment (Case E).

formation of these structurally novel ORFs by duplication. First, which of the two copies is the ancestral locus? Second, what kind of duplication event led to the formation of the derived novel copy? Third, if the derived copy contains unique exonic regions to the exclusion of the ancestral locus, what is the genomic source of these novel exons? Fourth, are the duplication termination points in the ancestral copies uniformly distributed in the genome or are they disproportionately restricted to a particular type of sequence (e.g., introns)? Finally, are these structurally novel duplicate copies evolutionary dead ends or are they transcriptionally active and perhaps functional?

Materials and Methods

Identification of Young Gene Duplicates Exhibiting Partial or Chimeric Structure in the *C. elegans* Genome

In a previous study of recently duplicated genes in *C. elegans*, we characterized gene-duplicate pairs on the basis of structural resemblance between the two paralogs (Katju and Lynch 2003). The data set comprised 290 gene-duplicate pairs with K_S values ranging from 0% to 10% and excluded duplicates belonging to multigene families (more than five family members) and sequences showing similarity to known transposable elements. Gene duplicates were classified as having (1) complete, (2) partial, or (3) chimeric structure (figure 1 in Katju and Lynch 2003). Complete gene duplicates had sequence homology extending between the initiation codon and the termination codon. If sequence homology was disrupted by indels, the gene duplicates were categorized as complete if sequence homology was resumed within the boundaries of the ORF. Partial gene duplicates comprised genes in which the longer paralog contained unique sequences that were absent from the shorter paralog. In chimeric gene duplicates, both paralogs contained unique sequences.

In order to elucidate how duplication can yield a structurally novel gene copy, we initially analyzed 172 gene-duplicate pairs with $0 < K_S \leq 0.10$ that were identified as exhibiting partial or chimeric structural resemblance (Katju and Lynch 2003). One of these duplicate pairs comprised paralogs with chimeric structure in addition to differential intron presence. Additionally, we further analyzed two duplicate pairs wherein introns were absent in one copy relative to the other despite complete homology in their coding sequence (complete structural resemblance),

to distinguish between "intron loss" versus "gain" in the duplicated copy.

Use of the *Caenorhabditis briggsae* Genome Sequence as an Outgroup to Determine the Ancestral Copy

For each of the 172 *C. elegans* duplicate pairs exhibiting partial or chimeric structural resemblance as per our delineation (Katju and Lynch 2003), sequence reports for both *C. elegans* paralogs were accessed from WormBase (http://www.wormbase.org; Stein et al. 2001). We further determined if either one or both the copies had a listed ortholog in the *C. briggsae* genome (Stein et al. 2003). Sixty-nine of the 172 duplicate pairs (40%) comprising the original data set had one or more *C. briggsae* orthologs listed on WormBase for one or both *C. elegans* paralogs. Subsequently, we accessed the amino acid and nucleotide sequence (spliced and unspliced versions) of the *C. briggsae* orthologs from WormBase and directly compared these to their putative *C. elegans* paralogs. Sequence analysis was implemented in the Se-Al Sequence Alignment Editor (Rambaut 1996). Initial sequence alignments were performed using Clustal software (Higgins, Bleasby, and Fuchs 1992) and completed visually.

It should be reiterated that this structural heterogeneity between the two *C. elegans* paralogs is not a result of minor indels or differential gene annotation assigning divergent exon-intron structure to the two paralogs. Rather, the structural differences are due to a break in homology between the two *C. elegans* paralogs within their ORFs. In order to unambiguously identify the progenitor copy in *C. elegans*, we required that the *C. briggsae* ortholog exhibit homology to one *C. elegans* paralog throughout the entire length of the latter, especially in the region of heterogeneity between the two *C. elegans* paralogs. Alignments based on amino acid sequences were preferable to nucleotide sequences, given the large estimated divergence time between these two congeneric species (Stein et al. 2003). Using this outgroup analysis, we were reliably able to determine the ancestral *C. elegans* paralog for 37 of the 69 (54%) structurally heterogeneous duplicate pairs with a listed *C. briggsae* ortholog. The remaining 32 *C. elegans* duplicate pairs were excluded because the *C. briggsae* ortholog exhibited equal similarity to both *C. elegans* paralogs (i.e., it failed to preferentially align with either one of the two *C. elegans* paralogs in the region of structural heterogeneity). Likewise, we were able to determine the ancestral *C. elegans* paralog for two additional complete duplicate pairs comprising paralogs with intron differences.

K_S values for each *C. elegans* duplicate pair in this data set were initially calculated by Lynch and Conery (2000) by employing a maximum-likelihood procedure in the PAML package (Yang 1997). To verify their results and given that we had new information on the location of the duplication breakpoints and the extent of sequence similarity between the two paralogs, we recalculated the K_S values using the Nei and Gojobori (1986) method. While there was some variation in the K_S values generated by the two techniques, the degree of synonymous divergence between the two paralogs was calculated to be less than 10% ($K_S \leq 0.10$) for 38 of the 39 duplicate pairs comprising this study.

Determination of the Genomic Source of Unique Exons in the Duplicated Copy

Once we were able to ascertain the direction of duplication by identifying the ancestral locus within a *C. elegans* duplicate pair, we sought to determine (1) if the derived duplicate copy possessed any unique ORF sequence to the exclusion of the ancestral copy, and if so, (2) the genomic source of these unique ORF exons. We performed a BlastN search of the nucleotides comprising the unique ORF sequence in the duplicated locus, if present, against the *C. elegans* genomic sequence within WormBase, using the Blast software package (Altschul et al. 1997) and only accepted matches with a conservative *E* value of 10^{-20} or lower. In cases where the query sequence generated multiple hits, we selected the genomic hit with the lowest *E* value as the probable genomic source of the unique ORF sequence in the duplicated copy. In some cases, the unique coding sequence in the derived copy was relatively short in length. Under these circumstances, it is possible that our Blast search would fail to yield *E* values matching our criterion of 10^{-20} or lower, leading us to underestimate the number of chimeric duplications and overestimate the frequency of partial duplications with recruitment.

In some cases, the query sequence generated multiple hits in the *C. elegans* genome, with similar *E* values and approximately equal representation on all six *C. elegans* chromosomes. Under these circumstances, this unique ORF sequence of the duplicate copy was additionally subjected to a global BlastN search against all sequenced genomes. If all of the resultant hits were still restricted to the *C. elegans* genome, we surmised that the source of the unique ORF in the derived copy was a family of *C. elegans*–specific repetitive elements.

To investigate if certain sequence types contribute differentially toward the recruitment of novel exons in chimeric duplicates, the genomic source of new exons was further classified as (1) genic (G), (2) intergenic (IG), (3) both genic and intergenic (G + IG), or (4) a repetitive element (RE).

Characterizing the Type of Duplication Event

We sought to determine the type of duplication event, once the identity of the ancestral gene and the genomic sources of new exons in the duplicated copy (if any) had been ascertained. Figure 1 provides a schematic representation of the rationale and methods employed in the determination of the type of duplication event leading to the formation of a novel gene. The data set of gene duplicates exhibiting structural heterogeneity comprised two categories of structural resemblance: partial and chimeric with one or both paralogs possessing unique ORF sequence to the exclusion of the other copy, respectively.

Duplicate pairs with partial similarity can arise in three different ways. (1) If "copy *a*" possessing the unique exons is ancestral (Case A in fig. 1), the shorter "copy *b*" represents an abbreviated version of the ancestral locus and lacks any nonhomologous sequence relative to its progenitor. The partial copy was able to complete its ORF by incorporating a novel start and/or stop codon from the fragment that was partially duplicated, by point mutations, small indels

or a refashioning of its exon-intron structure. This would represent a straightforward case of partial duplication of the ancestral gene, and the structural category would accurately represent the type of duplication event. This type of duplication was formally defined as a "partial duplication without recruitment" (henceforth referred to as "partial duplication"). (2) Alternatively, the shorter copy *b* lacking any unique ORF sequence is the ancestral locus. The next step was to determine if the unique exons in the longer, derived copy had any hits in the *C. elegans* genome. If genomic hits were identified, this would represent a case of "chimeric duplication" (Case B, fig. 1). Conversely, if the unique ORF sequence of the duplicate copy failed to yield any *C. elegans* genomic hits, it was taken to represent a complete duplication of the ancestral gene and the unique ORF sequence resulted from the subsequent recruitment of the neighborhood sequence from the duplicated copy's new genomic location, resulting in a novel reading frame. These cases were classified as "complete duplications with recruitment" (Case C, fig. 1). (3) Lastly, for duplicate pairs with chimeric structure (both paralogs have unique ORF sequences to the exclusion of the other copy), the type of duplication event depends on whether the unique ORF sequence in the duplicate yields any genomic hits for potential donors. If genomic hits were obtained, this would represent a "chimeric duplication" (Case D, fig. 1). Alternatively, the lack of any genomic hits to the duplicate copy's unique ORF sequence implies that the novel gene was created by a partial duplication of the ancestral copy followed by recruitment of the neighborhood sequence from its new genomic location. These were classified as "partial duplications with recruitment" (Case E, fig. 1).

There are multiple avenues for the formation of chimeric duplicate genes. Partial or complete duplications of two or more ancestral genes and the amalgamation of these sequences, either due to shuffling events or duplication across adjacent genes, can yield a chimeric gene. Likewise, duplicated intergenic region sequences can fuse with partially or wholly duplicated genes and form a novel chimeric gene. However, all these above-mentioned cases share one common thread, namely, that the duplicate copy resulted from the duplication of two or more ancestral genomic sources, be they genic or intergenic. Rather than split these categories, it was preferable to cluster them together as chimeric duplications, their trademark being an origin from two or more duplicated genomic sources. Alternatively, novel genes can also be created by partial duplication of an ancestral gene. Partial duplications create novel genes in themselves, but they differ from chimeric duplications in that there is only one duplicated ancestral source and the derived copy is able to complete its ORF either (1) by refashioning the exon-intron structure of the partially duplicated fragment to yield a novel ORF (partial duplications without recruitment) or (2) by incorporating the sequence from its new genomic neighborhood (partial duplications with recruitment). In the latter case, even though the duplicate appears to be chimeric in nature in that it contains a unique ORF sequence to the exclusion of the ancestral copy, the unique sequence is not derived from a duplication event. Hence, under our scheme, these cases are not considered to be true chimeric duplications.

Identification of Duplication Breakpoints in the Ancestral Copy

Having identified the ancestral copy within a structurally heterogeneous duplicate pair, we sought to determine the location of the two duplication breakpoints in the 5′ and 3′ directions beyond which homology between the two copies was terminated. For each ancestral source within this data set, a duplication breakpoint was assigned to one of three locations, namely, intergenic, exonic, or intronic. A chi-square test was used to compare the observed frequencies of exonic and intronic duplication breakpoints with the composition frequencies of exonic and intronic DNA in the *C. elegans* genome (The *C. elegans* Sequencing Consortium 1998).

Determining Functionality of Novel Genes Formed via Duplication

Complementary DNA (cDNA) information was collected from WormBase for all ancestral and duplicated copies to indirectly assess the potential functionality of these newly formed novel genes. To facilitate a more robust conclusion, cDNA information was also gathered for duplicate pairs exhibiting complete structural resemblance (Katju and Lynch 2003) and with a known single-copy ortholog in *C. briggsae*. The observed frequencies of duplicated copies with cDNA confirmation derived from (1) complete, (2) partial, versus (3) chimeric duplications were compared using a *G*-test (likelihood-ratio test) for goodness of fit (Sokal and Rohlf 1997) to determine whether the frequency of expressed duplicate copies depends on the type of duplication event.

Distinguishing Intron Loss Versus Gain in Gene-Duplicate Pairs with Differential Presence of Introns

In a preceding study, three gene-duplicate pairs with differential presence of introns within the two member copies had been identified (Katju and Lynch 2003). Of these, two pairs (W01D2.1/C54C6.1 and C03A7.14/C03A7.7) had homologous coding sequences but exhibited differential retention of introns. A third pair (B0035.2/C47A4.1) exhibited a chimeric structure (both member copies had unique exons) as well as differential presence of introns. We were able to identify a single-copy ortholog in *C. briggsae* for all three cases and determined the ancestral copy within *C. elegans* by directly comparing intron distributions across the three sequences (two *C. elegans* paralogs and one *C. briggsae* ortholog). This comparative approach helped determine the relative frequencies of intron loss and gain in the *C. elegans* gene duplicates.

Results

Sequence comparisons of two *C. elegans* paralogs with a single-copy ortholog in *C. briggsae* enabled the identification of the ancestral and derived copy in 37 pairs of evolutionarily young, but structurally heterogeneous, *C. elegans* gene duplicates (with one pair comprising paralogs exhibiting intron differences in addition to chimeric structure) as well as two gene-duplicate pairs with differential presence

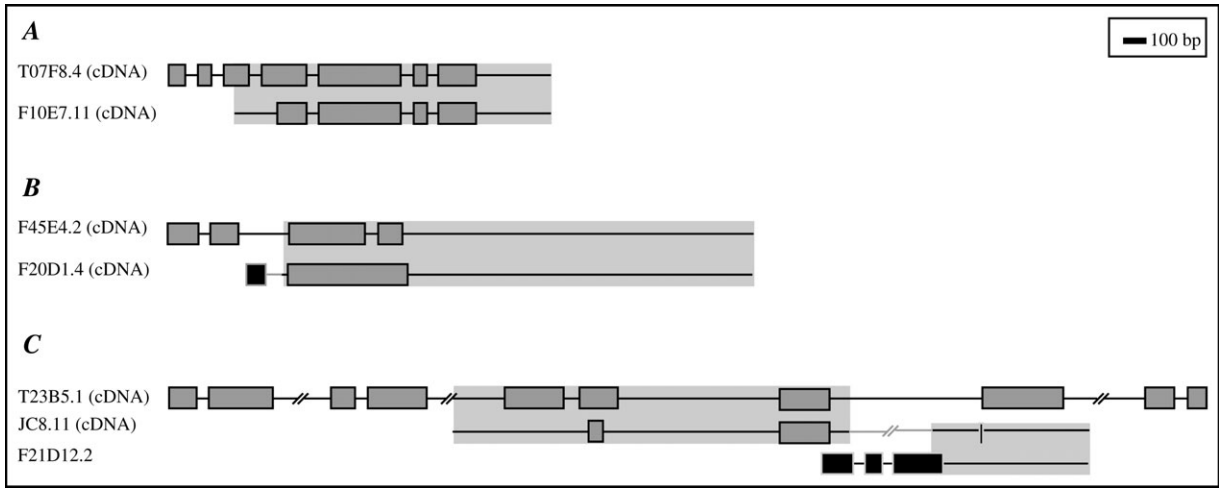

FIG. 2.—Representative examples of the three types of duplication events leading to the formation of novel genes. Shaded narrow rectangles denote exons, horizontal lines linking exons represent introns and duplicated UTR regions where applicable. The regions that were duplicated are highlighted in gray. Duplicated segments as determined by shared sequence homology between the ancestral and derived locus are also depicted by correspondence of regions with identical color and pattern. All figures are drawn to scale. Gene names followed by cDNA in parentheses denote paralogs that are either partially or fully confirmed by cDNA data. (*A*) Partial duplication (without recruitment) of ancestral locus T07F8.4 comprising seven exons resulted in the partial duplicate copy F10E7.11 ($K_S = 0.00$ over region of homology). The duplication was initiated at nucleotide position 54 of T07F8.4's exon 3, spanning the remainder downstream exons and introns of T07F8.4 as well as the first 326 nt of its 3′ UTR. Both the ancestral and derived copies are wholly confirmed by cDNA data. (*B*) Partial duplication with recruitment. The ancestral copy F45E4.2 comprises four exons. The duplication was initiated at nucleotide position 230 of F45E4.2's intron 2 and spans exons 3 through 4 and the first 1530 nt of the 3′ UTR. The derived copy F20D1.4's unique first exon and partial first intron contain a nonhomologous region of 156 nt that failed to yield any hits in the *Caenorhabditis elegans* genome, suggesting the recruitment of the additional sequence from its new genomic neighborhood. The ancestral and derived copies exhibit 2% divergence at synonymous sites over the region of homology. Ancestral copy F45E4.2 and derived copy F20D1.4 are wholly and partially confirmed by cDNA data, respectively. (*C*) Chimeric duplication. JC8.11 is a chimeric gene derived from two ancestral genes, T23B5.1 and F21D12.2. JC8.11's 5′ UTR (581 nt) and first two exons are derived from a region encompassing T23B5.1's intron 4 through partial intron 7 (79 nt). The next 687 nt of JC8.11's intron 2 have no corresponding hits in the *C. elegans* genome. JC8.11's remainder intron 2 (last 172 nt), terminal exon (8 nt), and 466 nt of 3′ UTR are homologous to a region encompassing F21D12.2's partial terminal exon and 3′ UTR. Ancestral copy T23B5.1 and derived copy JC8.11 are wholly and partially confirmed by cDNA data, respectively.

of introns within the two member copies despite complete structural homology in coding regions. Subsequently, it was first determined if the derived copy had any unique ORF sequence to the exclusion of the ancestral copy and second if these were derived from other sources in the *C. elegans* genome (detailed in fig. 1). This enabled the identification of the type of duplication event giving rise to the derived copy. A schematic example of each of the three observed types of duplication events is provided in figure 2.

Both Partial and Chimeric Gene Duplications Contribute to the Formation of Novel Genes

Twenty-three of 37 (62%) of the structurally heterogeneous duplicates were derived from partial duplications (either with or without recruitment of neighborhood sequence; Case A or E in fig. 1) compared to 14/37 (38%) from chimeric duplications (χ^2 test statistic = 2.19, df = 1, $0.5 > P > 0.1$). We found no cases representative of a complete duplication with recruitment (Case C, fig. 1).

Partial duplications can be further classified into two subcategories, namely, (1) those that were able to complete their ORF within the partial fragment that was duplicated and (2) those that recruited an additional neighborhood sequence to complete their ORF. These are defined as partial duplications with and without recruitment, respectively. Partial duplications were represented approximately equally by these two subcategories (10 partial duplications with recruitment and 13 partial duplications without recruitment).

Degree of Concordance Between Structural Category and Type of Duplication Event

How accurately does the structural resemblance between two genes resulting from gene duplication reflect the type of duplication event that occurred? We partitioned the two structural categories with respect to the type of duplication event. Of the 37 gene-duplicate pairs analyzed, 14 and 23 cases were designated as exhibiting partial and chimeric structure, respectively, as per the scheme followed in a previous study (Katju and Lynch 2003). Of the 14 gene-duplicate pairs with partial structure (only one copy possesses unique exons to the exclusion of the other copy), 13 (93%) did indeed result from a partial gene duplication as judged by the presence of a *C. briggsae* ortholog to the lengthier *C. elegans* copy. Only in one of these 14 cases was the shorter *C. elegans* copy represented in the *C. briggsae* genome. Thus, the lengthier *C. elegans* copy represents a novel gene derived from multiple ancestral sources (genic and intergenic) and was formed via a chimeric duplication.

The 23 gene-duplicate pairs with chimeric structure were characterized by both copies possessing a unique coding sequence to the exclusion of the other copy, in addition to their region of homology. Only 13/23 (56.5%) of these represented cases of true chimeric duplications wherein the duplicate copy was derived from duplications of two or more ancestral sources, be they genic or intergenic. For 11 of these 13 cases (~85%), the duplicate copy was

derived from two ancestral sources. The remaining two cases (15%) represent duplicate copies derived from three different ancestral sources. For the remaining 10 of the 23 duplicate pairs with chimeric structure (43.5%), the duplicate copy resulted from a partial duplication of the ancestral gene and subsequent recruitment of the neighboring sequence to the ORF (partial duplication with recruitment).

Diverse Genomic Sources for New Exons in Chimeric Duplicates

The genomic sources of unique coding regions in chimeric duplicates were identified by Blast searches against the *C. elegans* genome. The unique exons in 14 novel genes formed by chimeric duplications were derived from 17 ancestral sources. As mentioned earlier, in 11 cases, the unique coding sequence was derived from a single genomic source, whereas in the remaining three cases, two ancestral sources were identified. All four sequence categories (genic, intergenic, genic-intergenic, and repetitive elements) contributed to the formation of new exons in chimeric duplicates (see fig. 3). Two sequence types (intergenic and repetitive elements) contributed equally to the formation of unique exons, each with a relative frequency of ~29% (5/17). Genic-intergenic and purely genic contributions were marginally lower with relative frequencies of ~24% (4/17) and 18% (3/17), respectively.

Locations of Duplication Breakpoints are Strongly Correlated with the Frequency of Exonic and Intronic Sequences in the Genome

Intergenic, exonic, and intronic sequences comprise 47%, 27%, and 26% of the *C. elegans* genome, respectively (The *C. elegans* Sequencing Consortium 1998). The location of 100 duplication breakpoints in the ancestral sources (excluding repetitive elements) were identified. Of these breakpoints 43, 31, and 26 occurred in intergenic, exonic, and intronic sequences, respectively. We conducted a chi-square test to determine whether duplication breakpoints are more likely to occur in introns versus exons by comparing the observed frequencies of these duplication breakpoint locations relative to their expected values based on the genomic composition of exonic and intronic sequences within the *C. elegans* genome (fig. 4). We were unable to reject the null hypothesis that duplication breakpoints are randomly distributed with respect to exons and introns (χ^2 test statistic = 0.592, df = 1, $0.5 > P > 0.1$).

No Significant Difference in Transcriptional Activity of Duplicated Genes Arising from Three Diverse Types of Duplication Events

Functionality of new genes arising from complete, partial, and chimeric duplications was estimated by collecting cDNA confirmation data for each duplicate copy from WormBase. Approximately 41%, 23%, and 21% of new genes resulting from complete, partial, and chimeric duplications, respectively, were partially or wholly confirmed by cDNA data.

Although complete duplicates appear twice more likely to be transcribed relative to those formed by partial and chimeric duplications, this former category might be overestimated given that both copies are expected to yield the same cDNA sequence in the event of zero sequence divergence. Despite this potential for overestimation of cDNA data for complete gene duplicates, a *G*-test for goodness of fit revealed no significant association between the frequency of expression and the type of duplication event ($G_{adj} = 4.21$, df = 2, $0.5 > P > 0.1$).

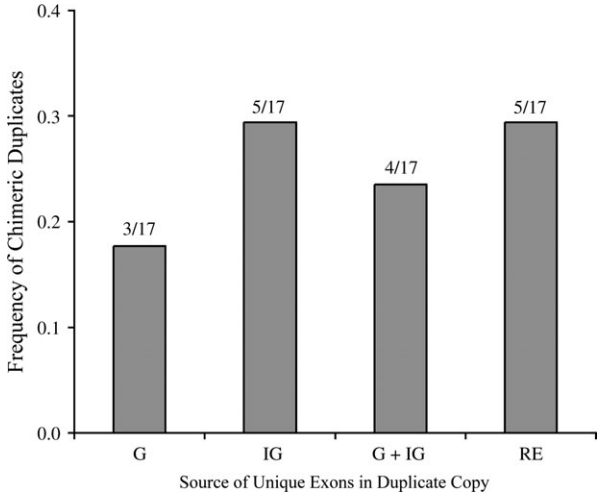

FIG. 3.—Relative contributions of four sequence types to the formation of unique exons in the duplicated copy ($n = 17$). The four sequence types are denoted as genic (G), intergenic (IG), both genic and intergenic (G + IG), and repetitive element (RE).

Both Intron Loss and Gain Contribute to Differential Intron Distributions Between *C. elegans* Paralogs

A direct comparison of the sequences of both *C. elegans* paralogs to their *C. briggsae* ortholog enabled the

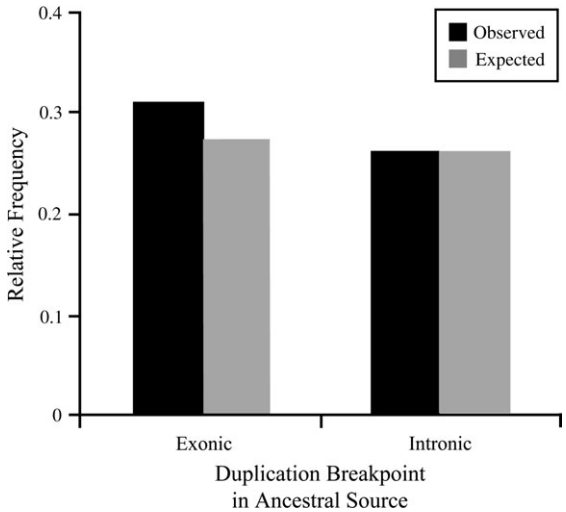

FIG. 4.—Distribution comparing observed and expected frequencies of duplication breakpoints in two types of sequences (exonic and intronic). Black shading denotes observed location of duplication breakpoints in ancestral copies. Light gray shading denotes expected frequencies if duplication breakpoints occurred uniformly in the genome.

identification of the ancestral copy within three *C. elegans* gene-duplicate pairs exhibiting differential intron distributions. Of these three cases, two represent intron loss from the duplicated copy, suggesting a role of reverse transcription in duplication. Interestingly, one case represents an intron gain by the duplicated copy.

The first case detailed here represents an intron loss in the duplicated copy. Gene duplicates W01D2.1 and C54C6.1 are mostly identical in coding sequence and belong to the L37 ribosomal protein (large subunit) family. Copy W01D2.1 is composed of two exons separated by a 300-bp intron. The *C. briggsae* ortholog CBG04239, too, retains the single intron separating the two exons. Given that these two species are estimated to have diverged 50–120 MYA (Coghlan and Wolfe 2002), it is no surprise that sequence similarity observed in introns is largely restricted to the area encompassing the exon-intron splice junctions. The single intron in the *C. elegans* ancestral copy has been lost in the duplicate C54C6.1, such that the two ancestral exons now comprise one exon. The location of the two *C. elegans* duplicates on different chromosomes, the precise deletion of the intron in the derived copy, and the lack of homology between the two copies in the 5′ untranslated region (UTR) strongly suggests that the duplication occurred via reverse transcription of a complete messenger RNA (mRNA) transcript. Alternatively, a simple DNA-level precise deletion of the intron could have resulted in the observed intron loss (Robertson 1998, 2000).

In contrast, the second case represents an intron gain by the duplicated copy. Gene duplicates C03A7.7 and C03A7.14 with $K_S = 0.17$ are homologous in their exonic regions but differ with respect to their intron distributions. Paralog C03A7.7 comprises two exons separated by a 54-bp intron, whereas paralog C03A7.14 is composed of three exons and two introns (44- and 54-bp). Outgroup analysis shows that the *C. elegans* paralog C03A7.7 and the *C. briggsae* ortholog CBG08690 have similar exon-intron structures comprising two exons and an intervening intron, suggesting that the duplicated copy C03A7.14 in *C. elegans* gained an intron. A Blast search within the *C. elegans* genome using the 44-bp intron sequence as a query generated five hits of similar identities (ranging from 79% to 84%). Three of the five hits were in other introns, whereas two hits matched the intergenic sequence.

Finally, the third case represents intron loss in the duplicated copy, with slight modifications relative to the first case discussed above. The two paralogs within gene-duplicate pair B0035.2/C47A4.1 with $K_S = 0.069$ display a chimeric structure, with each paralog possessing unique exons to the exclusion of the other copy. Locus B0035.2 with six exons was identified as the ancestral locus based on sequence homology to the *C. briggsae* ortholog CBG06069. C47A4.1, the putative duplicate copy, is composed of two exons. Homology between these two *C. elegans* paralogs is largely restricted toward the 3′ region of their ORFs. Ancestral copy B0035.2's three terminal exons are represented as a single exon in the duplicate C47A4.1 with the precise deletion of the two ancestral introns. C47A4.1's unique first exon and majority of its first intron failed to generate any hits in the *C. elegans* genome. It is possible that reverse transcription of a partially processed mRNA transcript roughly spanning the terminal half of the ancestral gene yielded a partially duplicated fragment lacking the associated introns. The resultant partially duplicated fragment in turn recruited the neighboring sequence from its new insertion site to fashion a novel first exon and intron and was hence classified as a partial duplication with recruitment.

Discussion

Gene duplication can spawn a diverse set of progeny loci with differing degrees of resemblance to their ancestors. Although this has been well appreciated in principle, most of the treatment of the early evolution of duplicated genes assumes that they are redundant and implicitly identical at birth. In a preceding study of evolutionarily young gene duplicates within the *C. elegans* genome (Katju and Lynch 2003), gene-duplicate pairs were classified as exhibiting complete, partial, or chimeric structure. The latter two structural categories consist of pairs with member copies exhibiting considerable structural heterogeneity despite their young evolutionary age and contributed to greater than 50% of the gene duplications in the data set. However, these categories of structural resemblance may not accurately reflect the types of duplication that created them (fig. 1). For example, a gene-duplicate pair in which one copy is lengthier than the other would have been characterized as a partial gene duplicate but might result from a true partial gene duplication or a complete duplication with recruitment of the new sequence. Likewise, gene-duplicate pairs with chimeric structure may result from partial duplications with recruitment, in addition to true chimeric duplications.

In this study, 172 evolutionarily young but structurally heterogeneous pairs were analyzed to elucidate how the process of gene duplication yields novel genes at birth or close to the time of conception. Taking advantage of the availability of the genome of a congeneric species, *C. briggsae*, the first aim was to identify the ancestral gene in a structurally heterogeneous gene-duplicate pair in *C. elegans*, based on unambiguous homology for both gene sequence and gene structure with a single-copy ortholog in *C. briggsae*. We were able to unambiguously identify a single-copy ortholog in *C. briggsae* for only 37 of the 172 (21.5%) structurally heterogeneous duplicate pairs analyzed. Although *C. elegans* and *C. briggsae* are classified as congeneric species and are morphologically very similar, the *briggsae-elegans* split is estimated to have occurred ~50–120 MYA, with a median speciation date estimated at ~88 MYA (Coghlan and Wolfe 2002). Employing a conservative estimate of the average rate of origin of new duplicates (Lynch and Conery 2000), 50% of all genes within the *C. elegans* genome would be expected to duplicate at least once on a timescale of 70 Myr (approximating the period since the *briggsae-elegans* split). This rapid rate of origin of new duplicates within a lineage coupled with the possibility that many of these duplications may be partial or chimeric in nature may preclude the unambiguous characterization of a single-copy ortholog in a moderately distant relative. Hence, each lineage can be expected to have a large number of supposed orphan genes with no clear orthologs in close relatives due to rapid gene turnover.

Approximately 60% (172 of 290) of gene-duplicate pairs with $0.0 \leq K_S \leq 0.10$ in *C. elegans* displayed structural heterogeneity between member copies. Therefore, only 40% of newborn gene duplicates in *C. elegans* arose from complete duplications of the ancestral ORF. However, in all probability, this is an inflated estimate of complete duplications given that an unknown fraction of these may further lack the full endowment of regulatory elements required for performing the ancestral function (Lynch and Katju 2004). In this study of 37 structurally heterogeneous duplicate pairs, 62% and 38% of novel genes arose by partial (with and without recruitment of the neighborhood sequence) and chimeric duplications, respectively. If these results are generalized to the entire data set of evolutionarily young duplicates in *C. elegans*, the relative frequencies of surviving copies from complete, partial, and chimeric duplications are 41%, 37%, and 22%, respectively. We conclude that gene duplication can create novel genes de novo and in conjunction with shuffling events. Structural differences between ancestral and duplicated copies at or close to conception may very well dictate divergent evolutionary trajectories for the two copies at the very onset. The main caveat to our conclusions is that they rest on the reliability of annotation programs to accurately identify gene structure. Although cDNA evidence exists for some of the ORFs in this analysis, it is not always complete. Sometimes ESTs are available for parts of a derived copy without extending to all the novel exons. Because the ESTs seldom provide confirmation for the entire gene, targeted RT-PCR may provide confirmation of the predicted gene structure for the novel copies.

Partial duplications, by definition, represent an abbreviated version of the ancestral gene. Half of these partial duplicates are able to complete their ORF without recruitment of additional sequence from their new genomic neighborhood, via point mutations that introduce a novel initiation or termination codon or slight frameshifts that refashion the ancestral exon-intron structure. Hence, partial duplications without recruitment are likely to code for a protein product similar to that of the ancestral locus, albeit a truncated version. The generation of a partial duplicate may accelerate the partitioning of ancestral function between the two copies given that one copy may already lack a domain or key structural elements at birth (Katju and Lynch 2003; Lynch and Katju 2004). On the other hand, both partial duplications with recruitment and chimeric duplications may bear greater potential for the formation of a gene product that is radically different from the ancestral protein. How alterations to gene structure via incomplete gene duplications (alone or in conjunction with shuffling events) alter the functionality of a gene still remains a central question that needs to be addressed empirically.

It should be pointed out that the observed structural differences between the two paralogs may have arisen either (1) at the time of duplication or (2) in the postduplication period and that we cannot distinguish between these two possibilities. As an example of the latter scenario, an initial partial duplication of an ancestral gene may have been followed by the recruitment of an adjacent sequence or an independently duplicated segment from elsewhere in the genome to yield a derived copy with unique sequence. Likewise, an initial complete duplication of an ancestral locus with subsequent deletion would result in what appears to be a partial gene duplication. However, a substantial portion (50%) of gene duplicates that have not had time to diverge at synonymous sites ($K_S = 0$) already displays structural differences with respect to one another. Moreover, there is a small but insignificant increase in the frequency of structurally heterogeneous (partial and chimeric) gene-duplicate pairs with increased K_S (Katju and Lynch 2003). If postduplication structural changes account for a significant fraction of the gene-duplicate pairs analyzed here, they occurred soon after the original duplication, before base-substitution changes had time to appear.

To what extent does the degree of structural resemblance between two young duplicates faithfully represent the type of duplication? Gene duplicates exhibiting partial structural resemblance are in all probability derived from partial duplications. In other words, given two gene duplicates of differing lengths, wherein only the longer copy has unique exons to the exclusion of the shorter copy, the longer copy is most likely to be the ancestral copy. For 93% of the cases analyzed, gene-duplicate pairs exhibiting partial structure indeed represented partial duplications with the lengthier copy being the ancestral locus. This prediction is supported by findings in both the worm (Katju and Lynch 2003) and mouse genomes (Mallon et al. 2004) that, on average, duplicated genes tend to be shorter than their ancestral counterparts. In contrast, for duplicate pairs exhibiting a chimeric structure (wherein both copies possess unique exons to the exclusion of the other copy), the duplicate copy is equally likely to have arisen from a chimeric or partial duplication with recruitment. A duplicated copy can gain new exons by refashioning a new sequence recruited from its genomic neighborhood or via shuffling events that mediate fusion with fragments duplicated elsewhere in the genome.

Chimeric duplicates derived from the fusion of multiple duplicated fragments can play a significant role in the origin of evolutionary novelties. Shuffling of fragments or domains can dramatically alter the regulation and functionality of the recipient, as well as increase its probability of survivorship if the new function is viewed favorably by natural selection. The significance of shuffling events as an evolutionary mechanism is well established (Patthy 1985), although the extent to which exons truly represent functional domains remains a point of contention (Stoltzfus et al. 1994; Doolittle 1995). This study enables the determination of the types of genomic sequence that partake in such shuffling events. Chimeric duplicates derived from multiple ancestral sources appear to gain unique exons from a diversity of genomic sources, namely, genic and intergenic regions as well as repetitive elements (fig. 3). Furthermore, the relative contributions of these different sequences toward the formation of unique exons are approximately equal, given that no one particular sequence category appears to be significantly overrepresented. The diversity of sequences contributing to the formation of novel exons suggests enormous genomic fluidity and exchange.

There has been a great deal of interest in the evolutionary role of introns, and it is widely accepted that introns can promote recombination between parts of genes resulting in

chimeric genes, a process usually referred to as exon shuffling (Gilbert 1978). Moreover, it has been argued that introns are located between functional domains and that there is a correspondence between functional domains and exons (Go 1981, 1983; de Souza et al. 1996, 1998; Fedorov et al. 2003), although this notion has been strongly contended (Traut 1988; Patthy 1991; Doolittle 1995; Wolf, Kondrashov, and Koonin 2000, 2001). The reason for this, the argument goes, is that joining parts of different genes together inside functional domains is far less likely to result in functional products than joining genes between functional domains. If the intron-exon structure corresponds in some way to functional domains, and if chimeric duplicate genes that join gene segments in their introns are more likely to be functional than chimeric genes joined inside their exons, this would be reflected in the locations of the duplication breakpoints: chimeric duplicate genes joining gene segments in their introns would be more common than genes joined inside their exons after accounting for the relative lengths of introns and exons. The observed numbers of duplication breakpoints within exonic and intronic regions are very similar to the expected values based on the fractions of exonic and intronic regions in the *C. elegans* genome (fig. 4). This fails to support the hypothesis that breakpoints occur preferentially in one particular type of sequence (such as introns). The duplication breakpoints appear to be effectively random with respect to these two types of sequences. These conclusions rest on the assumption that the ORFs analyzed here are functional genes. If they are not, the apparent random distribution of duplication breakpoints would simply result from lack of natural selection against inappropriate joining of coding fragments. Interestingly, Conant and Wagner (2005) analyzed shuffling events in conserved genes within the *Drosophila* and *Caenorhabditis* genomes and similarly conclude a lack of association between shuffling boundaries and exon-intron junctions.

Are these novel genes functionally active or are they evolutionary dead ends? The frequency of complete duplicates in this sample for which cDNA information exists in WormBase is twice that of partial and chimeric duplicates; however, this difference is not statistically significant and duplicates derived from partial and chimeric duplications may be equally likely to be transcriptionally active as those resulting from complete duplications.

However, approximately 67% of duplicate copies lack expression based on cDNA data. At first glance, this may suggest nonfunctionalization as the predominant fate for most duplicate copies. However, the compilation of the *C. elegans* EST database is still very much a work in progress. Furthermore, a lack of detectable transcriptional activity is not evidence of pseudogenization. Duplicates may adopt regulatory roles as has been demonstrated in recent studies of the mouse *Makorin1-p1* gene (Hirotsune et al. 2003; Podlaha and Zhang 2004), wherein the presumptive pseudogene regulates the expression of its functional homolog. Genes with regulatory roles may have lower levels of transcriptional activity than housekeeping genes and are therefore less likely to be detected in EST studies. Additionally, gene copies originating from duplicative retrotransposition and assumed to be nonfunctional at birth have been demonstrated to evolve novel functions (Long et al. 2003). A large number of pseudogenes exhibit the standard trademarks of functionality (an excess of synonymous over nonsynonymous substitutions, preservation of the ORF, and evolutionary conservation) to the extent that several authors have questioned the legitimacy of their current designation as pseudogenes (Balakirev and Ayala 2003; Hirotsune et al. 2003; Podlaha and Zhang 2004). Finally, a lack of functionality now does not preclude future exaptation. An intriguing case of pseudogene resurrection involves bovine seminal ribonuclease, derived from a duplication of pancreatic ribonuclease approximately 35 MYA and rendered functional 5–8 MYA (Trabesinger-Ruef et al. 1996). Functionally inactive duplicates do have an adaptive potential if they can persist in the genome. Future insertion of transposable elements in their genomic vicinity may provide novel promoters to drive their expression (Brosius 1999; Nekrutenko and Li 2001; Ganko et al. 2003), and fortuitous shuffling events with other fragments may create new chimeric genes encoding novel protein products.

Lastly, comparative analysis with *C. briggsae* also allowed us to determine the respective roles of intron loss versus gain in the differential distribution of introns between three gene-duplicate pairs. Two cases represent loss of introns in the derived copy, suggesting either reverse transcription of RNA as a mechanism of duplication or duplication followed by precise DNA-level intron deletion by nonhomologous recombination (Robertson 1998, 2000). One case of intron loss potentially represents reverse transcription of a partially processed pre-mRNA, leading to the formation of a semiprocessed gene wherein only some introns were eliminated, akin to the formation of the preproinsulin-I gene in rat and mice (Perler et al. 1980). In the third case, the derived copy actually gained an intron that cleaved the first ancestral exon into two exons. This 44-bp intron sequence matches both intergenic and other intronic sequences in the genome, suggesting that introns are derived from elements (Palmer and Logsdon 1991; Logsdon 1998) that may be present in intergenic regions or have come to exist as introns elsewhere in the genome (donor introns) (Coghlan and Wolfe 2004; Logsdon 2004).

Ohno (1970) stated that nothing in evolution is created de novo. In his view, every existing gene must have arisen from a preexisting gene. In a strict sense, this view is upheld with respect to gene duplication given that new genes arise from preexisting genetic material. However, Ohno (1970) also championed the notion (both explicitly and implicitly) that gene duplication yields copies structurally and functionally identical to the progenitor locus (via a complete gene duplication) and evolutionary innovation was facilitated by the accumulation of "forbidden" mutations in one copy in the period following gene duplication. In this study, we demonstrate how gene duplication can fashion structurally novel genes de novo from extant genomic material via partial duplications and often in conjunction with shuffling events. The novel structure of genes derived from partial and chimeric duplications predisposes them to divergent evolutionary trajectories at the very onset. These results demonstrating the widespread occurrence of partial and chimeric duplicates with limited semblance to their ancestral counterparts also have bearing

Table 1
List of 39 Gene-Duplicate Pairs in *Caenorhabditis elegans* with an Identified Single-Copy Ortholog in *Caenorhabditis briggsae*

Ancestral Copy	Duplicate Copy	K_S	Structural Category	Type of Duplication	*C. briggsae* Ortholog for Ancestral Copy	Source of Unique Exons	Duplication Breakpoints	cDNA
T05C3.5	C24G6.5	0.00	Chimeric	Partial with recruitment	CBG09381	—	UTR/E2	+/–
W03B1.9	W03B1.5	0.00	Partial	Partial	CBG13411	—	I3/UTR	+/–
D2045.2	H04D03.3	0.00	Partial	Partial	CBG18387	—	E13/Stop	+/–
F20B10.1	F20B10.2	0.00	Chimeric	Chimeric	CBG06244	SC-IG	UTR/E7	+/+
C09F9.2	C09F9.4	0.00	Partial	Partial	CBG04265	—	E5/UTR	+/–
C24A8.1	C24A8.4a	0.00	Chimeric	Chimeric	CBG14393	SC-G+IG	I6/UTR	–/–
F28H1.4a	F47B3.3	0.00	Partial	Partial	CBG12087	—	I3/UTR	+/–
C25G4.10	T04A11.3	0.00	Partial	Partial	CBG03342	—	E3/UTR	+/–
T07F8.4	F10E7.11	0.00	Partial	Partial	CBG13072	—	E3/UTR	+/+
Y40H7A.10	Y40H7A.9	0.00	Chimeric	Chimeric	CBG13740	DC-IG	I4/UTR	+/–
Y51H4A.17	Y51H4A.20	0.00	Chimeric	Chimeric	CBG13675	RE	I5/E8	+/–
W01D2.1	C54C6.1	0.00	Complete	Intron loss	CBG04239	—	—	+/+
ZK1127.9a	ZK1127.6	0.00	Chimeric	Partial with recruitment	CBG11210	—	E5/UTR	+/+
F44E7.2	K09H11.7	0.01	Chimeric	Partial with recruitment	CBG09323	—	E1/E3	+/–
Y51H4A.17	Y51H4A.19	0.01	Chimeric	Partial with recruitment	CBG13675	—	I2/I4	+/–
Y51H4A.17	Y51H4A.18	0.02	Partial	Partial	CBG13675	—	UTR/I1	+/+
C49C3.1	Y43D4A.1	0.02	Chimeric	Chimeric	CBG00451	SC-G	UTR/E3	+/–
W09C3.1	M04F3.3	0.02	Partial	Partial	CBG03328	—	I1/UTR	–/–
C27F2.2	F17C8.6	0.02	Partial	Partial	CBG18206	—	I17/UTRl	+/–
C14C11.1	ZC317.6	0.02	Partial	Partial	CBG09348	—	I3/UTR	+/–
F45E4.2	F20D1.4	0.02	Chimeric	Partial with recruitment	CBG05926	—	I2/UTR	+/+
C14C11.6	ZC317.1	0.02	Chimeric	Chimeric	CBG09254	DC-IG	UTR/E5	+/–
C36C9.4	T25D1.1	0.03	Chimeric	Chimeric	CBG16972	DC-IG	UTR/UTR	–/–
C27A7.6	Y43D4A.4	0.03	Chimeric	Chimeric	CBG23388	DC-G+IG; RE	E3/I11	+/–
M01G12.5	M01G12.2	0.03	Partial	Partial	CBG18680	—	E1/Stop	+/–
B0379.2	Y106G6E.3	0.04	Chimeric	Chimeric	CBG03810	RE	I1/UTR	+/–
B0250.1	B0250.7	0.04	Chimeric	Chimeric	CBG05588	SC-G; RE	I2/E3	+/–
C49A9.1	E02H9.4	0.05	Chimeric	Partial with recruitment	CBG05496	—	E7/UTR	+/–
T23B5.1	JC8.11	0.06	Chimeric	Chimeric	CBG03486	DC-G+IG	I4/I7	+/+
K08F11.5	C47C12.4	0.07	Chimeric	Chimeric	CBG01740	RE	UTR/E8	+/–
B0035.2	C47A4.1	0.07	Chimeric	Partial with recruitment (intron loss)	CBG06069	—	I3/UTR	+/–
ZK1127.11	T02G5.6	0.07	Partial	Partial	CBG24743	—	E11/I15	+/–
T25G12.2	C33E10.10	0.07	Chimeric	Partial with recruitment	CBG15942	—	E4/I5	–/+
C54E10.1	T08G5.9	0.07	Partial	Partial	CBG15607	—	I14/UTR	+/–
Y37A1B.5	F42G8.8	0.07	Chimeric	Chimeric	CBG18761	SC-G	I4/I5	+/–
C56C10.3	Y61A9LA.5	0.08	Partial	Chimeric	CBG04319	DC-G+IG; SC-IG	UTR/UTR	+/+
F55A3.3	F55A3.7	0.09	Chimeric	Partial with recruitment	CBG12204	—	I6/E6	+/–
Y48A6C.4	Y48A6C.1	0.09	Chimeric	Partial with recruitment	CBG13183	—	E3/UTR	+/–
C03A7.7	C03A7.14	0.17	Complete	Intron gain	CBG08689	—	—	+/+

NOTE.—The list includes two gene-duplicate pairs with complete homology in coding regions but differential intron distributions. Columns 1 and 2 identify the ancestral and derived locus, respectively. Column 3 presents the synonymous-site divergence (K_S) between the two paralogs in the region of homology. Columns 4 lists the assigned category of structural resemblance from a previous study conducted prior to the availability of the *C. briggsae* genome. Column 5 identifies the type of duplication event leading to the formation of the novel derived gene. Column 6 lists an identified single-copy ortholog in *C. briggsae*. Column 7 lists the source of unique exons in the derived copy, if present (DC denotes location on a different chromosome than the previously identified ancestral copy; SC denotes an alternate ancestral source residing on the same chromosome as the originally identified ancestral copy; IG denotes an intergenic region; G denotes a genic region; G+IG denotes both genic and intergenic regions; RE denotes repetitive element). Column 8 lists the two duplication breakpoints in the ancestral copy in the 5′ to 3′ direction (E denotes exon; I denotes intron; UTR denotes untranslated region; Stop denotes termination codon). Column 9 lists cDNA information with a ''+'' and ''–''denoting cDNA presence and absence, respectively.

for gene duplication theory. The actual rates of neofunctionalization, subfunctionalization, and nonfunctionalization may be substantially different from predictions made by models simulating the evolution of complete (redundant) duplicates.

Acknowledgments

We are grateful to Ulfar Bergthorsson, Jason Bragg, Austin Hughes, and Jeffrey Palmer for constructive comments on earlier versions of the manuscript. We also thank

four anonymous reviewers for their extremely detailed and helpful 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 and a Indiana University Summer Dissertation fellowship to V.K. and a National Institutes of Health grant RO1-GM36827 to M.L.

Literature Cited

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389–3402.

Balakirev, E. S., and F. J. Ayala. 2003. Pseudogenes: are they "junk" or functional DNA? Annu. Rev. Genet. **37**:123–151.

Brosius, J. 1999. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene **238**:115–134.

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.

Coghlan, A., and K. H. Wolfe. 2002. Fourfold faster rate of genome rearrangement in nematodes than in *Drosophila*. Genome Res. **12**:857–867.

———. 2004. Origins of recently gained introns in *Caenorhabditis*. Proc. Natl. Acad. Sci. USA **101**:11362–11367.

Conant, G. C., and A. Wagner. 2005. The rarity of gene shuffling in conserved genes. Genome Biol. **6**:R50.1–R50.14.

Courseaux, A., and J. L. Nahon. 2001. Birth of two chimeric genes in the Hominidae lineage. Science **291**:1293–1297.

de Souza, S. J., M. Long, L. Schoenbach, S. W. Roy, and W. Gilbert. 1996. Intron positions correlate with module boundaries in ancient proteins. Proc. Natl. Acad. Sci. USA **93**:14632–14636.

de Souza, S. J., M. Long, R. J. Klein, S. Roy, S. Lin, and W. Gilbert. 1998. Toward a resolution of the introns early/late debate: only phase zero introns are correlated with the structure of ancient proteins. Proc. Natl. Acad. Sci. USA **95**:5094–5099.

Doolittle, R. F. 1995. The multiplicity of domains in proteins. Annu. Rev. Biochem. **64**:287–314.

Fedorov, A., S. Roy, X. Cao, and W. Gilbert. 2003. Phylogenetically older introns strongly correlate with module boundaries in ancient proteins. Genome Res. **13**:1155–1157.

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

Ganko, E. W., V. Bhattacharjee, P. Schliekelman, and J. F. McDonald. 2003. Evidence for the contribution of LTR retrotransposons to *C. elegans* gene evolution. Mol. Biol. Evol. **20**:1925–1931.

Gilbert, W. 1978. Why genes in pieces? Nature **271**:501.

Go, M. 1981. Correlation of DNA exonic regions with protein structural units in haemoglobin. Nature **291**:90–92.

———. 1983. Modular structural units, exons, and function in chicken lysozyme. Proc. Natl. Acad. Sci. USA **80**: 1964–1968.

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

Hirotsune, S., N. Yoshida, A. Chen, L. Garrett, F. Sugiyama, S. Takahashi, K. I. Yagami, A. Wynshaw-Boris, and A. Yoshiki. 2003. An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. Nature **423**:91–96.

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

Katju, V., and M. Lynch. 2003. The structure and early evolution of recently arisen gene duplicates in the *Caenorhabditis elegans* genome. Genetics **165**:1793–1803.

Kimura, M., and T. Ohta. 1974. Probability of gene fixation in an expanding finite population. Proc. Natl. Acad. Sci. USA **71**:3377–3379.

Li, W. H., and T. Gojobori. 1983. Rapid evolution of goat and sheep globin genes following gene duplication. Mol. Biol. Evol. **1**:94–108.

Logsdon, J. M. Jr. 1998. The recent origins of spliceosomal introns revisited. Curr. Opin. Genet. Dev. **8**:637–648.

———. 2004. Worm genomes hold the smoking guns of intron gain. Proc. Natl. Acad. Sci. USA **101**:11195–11196.

Long, M., E. Betran, K. Thornton, and W. Wang. 2003. Origin of new genes: glimpse from young and old. Nat. Rev. Genet. **4**:865–875.

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.

Lynch, M., and J. S. 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., and V. Katju. 2004. The altered evolutionary trajectories of gene duplicates. Trends Genet. **20**:544–549.

Mallon, A. M., L. Wilming, J. Weekes et al. (16 co-authors). 2004. Organization and evolution of a gene-rich region of the mouse genome: a 12.7-Mb region deleted in the Del(13)*Svea*36H mouse. Genome Res. **14**:1888–1901.

Nei, M., and T. Gojobori. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3**:418–426.

Nekrutenko, A., and W. H. Li. 2001. Transposable elements are found in a large number of human protein-coding genes. Trends Genet. **17**:619–621.

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.

———. 1973. Ancient linkage groups and frozen accidents. Nature **244**:259–262.

Palmer, J. D., and J. M. Logsdon Jr. 1991. The recent origin of introns. Curr. Opin. Genet. Dev. **1**:470–477.

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

———. 1991. Modular exchange principles in proteins. Curr. Opin. Struct. Biol. **1**:351–361.

———. 1999. Protein evolution. Blackwell Science Ltd., Oxford.

Perler, F., A. Efstratiadis, P. Lomedico, W. Gilbert, R. Kolodner, and J. Dodgson. 1980. The evolution of genes: the chicken preproinsulin gene. Cell **20**:555–566.

Podlaha, O., and J. Zhang. 2004. Nonneutral evolution of the transcribed pseudogene *Makorin1-p1* in mice. Mol. Biol. Evol. **21**:2202–2209.

Radlwimmer, F. B., and S. Yokoyama. 1998. Genetic analyses of the green visual pigments of rabbit (*Oryctolagus cuniculus*) and rat (*Rattus norvegicus*). Gene **218**:103–109.

Rambaut, A. 1996. Se-Al: sequence alignment editor. (http://evolve.zoo.ox.ac.uk/).

Robertson, H. M. 1998. Two large families of chemoreceptor genes in the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae* reveal extensive gene duplication, diversification, movement, and intron loss. Genome Res. **8**: 449–463.

———. 2000. The large *srh* family of chemoreceptor genes in *Caenorhabditis* nematodes reveals processes of genome evo-

lution involving large duplications and deletions and intron gains and losses. Genome Res. **10**:192–203.

Sokal, R. R., and F. J. Rohlf. 1997. Biometry: the principles and practice of statistics in biological research. W. H. Freeman and Company, New York.

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.

Stein, L. D., Z. Bao, D. Blasiar et al. (33 co-authors). 2003. The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. PLoS Biol. **1**:166–192.

Stoltzfus, A., D. F. Spencer, M. Zuker, J. M. Logsdon Jr., and W. F. Doolittle. 1994. Testing the exon theory of genes: the evidence from protein structure. Science **265**:202–207.

The *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science **282**:2012–2018.

Thomson, T. M., J. J. Lozano, N. Loukili et al. (11 co-authors). 2000. Fusion of the human gene for the polyubiquitination coeffector UEV1 with *Kua*, a newly identified gene. Genome Res. **10**:1743–1756.

Trabesinger-Ruef, N., T. Jermann, T. Zankel, B. Durrant, G. Frank, and S. A. Benner. 1996. Pseudogenes in ribonuclease evolution: a source of new biomacromolecular function? FEBS Lett. **382**:319–322.

Traut, T. W. 1988. Do exons code for structural and functional units in proteins? Proc. Natl. Acad. Sci. USA **85**:2944–2948.

Wang, W., F. G. Brunet, E. Nevo, and M. Long. 2002. Origin of *sphinx*, a young chimeric RNA gene in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **99**:4448–4453.

Wolf, Y. I., F. A. Kondrashov, and E. V. Koonin. 2000. No footprints of primordial introns in a eukaryotic genome. Trends Genet. **16**:333–334.

———. 2001. Footprints of priomordial introns on the eukaryotic genome: still no clear traces. Trends Genet. **17**:499–501.

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

Yokoyama, S., and R. Yokoyama. 1990. Molecular evolution of visual pigment genes and other G-protein coupled receptor genes. Pp. 307–322 *in* N. Takahata and J. F. Crow, eds. Population biology of genes. Baifukan, Tokyo.

Zhang, J., H. F. Rosenberg, and M. Nei. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc. Natl. Acad. Sci. USA **95**:3708–3713.

Naoko Takezaki, Associate Editor

Accepted February 20, 2006