Evolutionary Patterns of Recently Emerged Animal Duplogs

Kiyoshi Ezawa[1,2,4], Kazuho Ikeo[2,3], Takashi Gojobori[2,3], and Naruya Saitou[1,*]

[1]Division of Population Genetics, National Institute of Genetics, Mishima, Japan

[2]Human Genome Network Project, National Institute of Genetics, Mishima, Japan

[3]DNA Data Analysis Laboratory, Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Mishima, Japan

[4]Present address: School of Advanced Sciences, Graduate University of Advanced Studies, Hayama, Japan

*Corresponding author: E-mail: saitounr@lab.nig.ac.jp.

Accepted: 19 July 2011

Abstract

Duplogs, or intraspecies paralogs, constitute the important portion of eukaryote genomes and serve as a major source of functional innovation. We conducted detailed analyses of recently emerged animal duplogs. Genome data of three vertebrate species (*Homo sapiens*, *Mus musculus*, and *Danio rerio*), *Caenorhabditis elegans*, and two *Drosophila* species (*Drosophila melanogaster* and *D. pseudoobscura*) were used. Duplication events were divided into six age-groups according to the synonymous distance (dS) up to 0.6. Duplogs were classified into four equal-sized classes on physical distances and into three classes on relative orientations. We observed the following shared characteristics among intrachromosomal multiexon duplogs: 1) inverted duplogs account for 20–50%, and about a half of the physically most distant 25%; 2) except for *C. elegans*, the composition of physical distances, that of relative orientations, and the proportion of inverted duplogs in each physical distance category are more or less uniform; 3) except for *C. elegans*, the characteristics of the youngest ($dS < 0.01$) duplogs are similar to the overall characteristics of the entire set. These results suggest that intrachromosomal duplogs with fairly long physical distances were generated at once, rather than resulting from tandem duplications and subsequent genomic rearrangements. This is different from the three well-known modes of gene duplication: tandem duplication, retrotransposition, and genome duplication. We termed this new mode as "drift" duplication. The drift duplication has been producing duplicate copies at paces comparable with tandem duplications since the common ancestor of vertebrates, and it may have already operated in the common ancestor of bilateral animals.

Key words: duplog, paralog, gene duplication, physical distance, transcriptional orientation, animals, genome-wide analysis, cross-sectional analysis.

Introduction

Gene duplication has long been one of the major subjects of evolutionary studies because it is considered as one of the major sources of genomic innovations (e.g., Haldane 1932; Muller 1935; Nei 1969; Ohno 1970; Lynch 2007). Genome sequence data revealed that eukaryotic genomes are fairly rich in duplicated genes (e.g., Lynch and Conery 2000; Rubin et al. 2000; Wapinski et al. 2007), and thus supporting the above concept. Fitch (1970) proposed to call duplicated genes as paralogs. Paralogous genes may exist either in the same species or in different species. Because duplicated gene pairs existing in one species are the main focus of this study, we would like to propose to call them "duplogs," as a subclass of paralogous sequences. Duplog is a synonym of "intraspecies paralogs" but is much shorter and easy to use. Wolfe (2000) proposed to call duplicated genes created through genome duplications as "ohnologs" after Susumu Ohno. All ohnologs in one species genome are duplogs. Duplogs are somewhat related to "inparalogs," paralogs in a given lineage that all evolved by gene duplications that happened after the radiation (speciation) event that separated the given lineage from the other lineage under consideration (Sonnhammer and Koonin 2002). Duplogs are, however, simply any kind of paralogs found in one species genome.

In evolutionary genomics, three duplication mechanisms have been well known to create duplogs of different positional relationships: 1) tandem duplication mostly creates

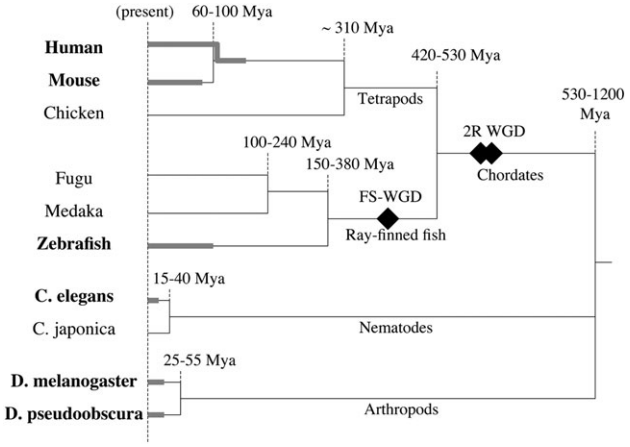

Fig. 1.—Six species used and the time frame of this study. The figure shows the six animal species used in this study (in boldface), which have finished or nearly finished genome assemblies, and some closely related outgroup species. Thick gray lines along the subject species lineages roughly span the time intervals equivalent to the neutral nucleotide divergence of 0.6 between duplogs.

physically close head-to-tail duplogs, 2) retrotransposition mostly creates unlinked or very distant intronless duplogs (reviewed, e.g., in Babushok et al. 2007), and 3) whole genome duplication (WGD) doubles the gene complement at least immediately after that (Ohno 1970). It is now believed that the common ancestor of vertebrates underwent two rounds of WGD (fig. 1; see also Dehal and Boore 2005) and that the common ancestor of teleost fish experienced another round of fish-specific WGD (fig. 1; see also Jallion et al. 2004; Woods et al. 2005). At the same time, we also know that the WGD events account for only a fraction of the duplicate genes of extant vertebrates.

Studying the long-term evolution of human and mouse duplogs, Friedman and Hughes (2003, 2004) found negative correlations between the proportion of linked duplog pairs and the sequence divergence, and they concluded that duplicate genes have been generated mainly via tandem duplication and have been physically separated via genome rearrangements. Since then, genome-wide studies have shown that tandem duplicate genes account for a considerable fraction (18–34%) of all duplogs in vertebrate genomes (Shoja and Zhang 2006; Pan and Zhang 2008), but the remaining fraction was not characterized except retrotransposed genes (Pan and Zhang 2007). Shoja and Zhang (2006) also found that head-to-tail duplog pairs in their set of "tandemly arrayed genes" tend to have smaller physical distances than inverted pairs. Combined with the arguments by Friedman and Hughes (2003, 2004), this may indicate that inverted and physically relatively distant duplogs have resulted from chromosomal rearrangements. To confirm that this is the case, however, detailed analyses on recently created duplogs are indispensable.

Recent duplications with 90% or more nucleotide identity, excluding retrotranspositions, have been actively studied as segmental duplications (SDs) since the advent of the human genome assemblies (International Human Genome Sequencing Consortium 2001; Bailey et al. 2002; Bailey and Eichler 2006). It was revealed that the human genome is abundant in interspersed SDs (Bailey et al. 2002, 2003; Bailey and Eichler 2006). A considerable fraction of interspersed SDs was explained by direct creations via mechanisms other than tandem duplications, possibly mediated by interspersed elements (Bailey et al. 2003). The abundance of SDs was often considered as specific to human or hominoids (Bailey and Eichler 2006; Marques-Bonet et al. 2009) because early analyses based on the whole genome shotgun (WGS) assemblies of mammalian genomes, such as the mouse genome, showed a paucity of SDs (e.g., Cheung et al. 2003). Recently, more careful analyses using the finished assembly of the mouse genome (She et al. 2008; Church et al. 2009) concluded that the SD content in the mouse genome is ca. 5%, which is comparable with that of the human genome, and that the mouse SDs are richer in tandem duplications. In these studies on primate and rodent SDs, however, tandem and interspersed SDs were roughly distinguished based solely on physical distances (with low resolutions of at best 1 Mb) taking no account of relative orientations. Besides, they did not examine the dependence of relative positions on the duplication age, leaving it unclear whether the interspersed SDs were indeed created directly or resulted from chromosomal rearrangements. As for recent duplogs in nonmammalian vertebrates, there are virtually no studies so far on the evolution of positional relationships. Regarding invertebrates, Katju and Lynch (2003) studied quite recent (synonymous distance [dS] < 0.1) duplogs in *Caenorhabditis elegans*. They found that inverted pairs account for a majority of *C. elegans* duplog pairs physically close to each other and suggested a possible duplication mechanism different from tandem duplication via unequal crossing-over. But a question remains as to whether such a mechanism is shared by other species or not.

To step up our understanding on duplog evolution, it is undoubtedly necessary to clarify whether physically relatively distant and/or inverted duplogs were created by one-step mechanisms or by tandem duplications and subsequent genomic rearrangements, and whether such mechanisms are shared across animals or specific to lineages. For this purpose, it is crucial to study the evolution of physical relationships between recently created duplogs. Theoretically, this could be achieved by comparing the chromosomal positions of orthologs of duplicate genes in closely related species. The problem is that species with finished-quality genome assemblies are still sparse in animal phylogeny (boldface species in fig. 1). It is now well appreciated that the draft genome assemblies based on the WGS technique is poor at locating duplicated DNA sequences, especially recently diverged ones (She et al. 2004; Church et al.

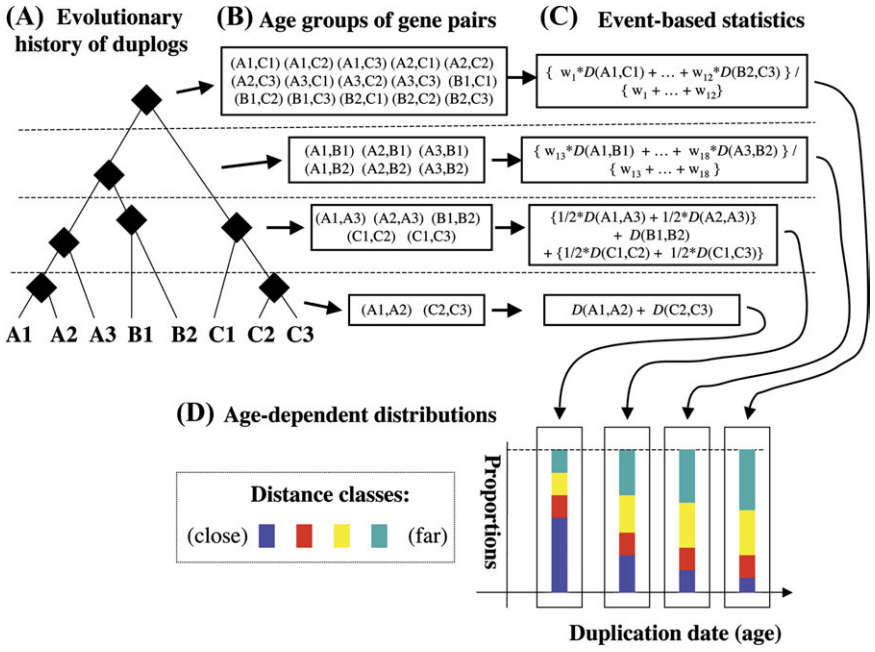

Fig. 2.—Cross-sectional analyses based on duplog pairs and those based on duplication events. For the illustration purpose, we showed the methods applied to a fictitious gene family, whose phylogenetic tree is given on the panel *A*. We distributed gene pairs into "age-groups" according to their duplication dates and then took statistics of physical properties on each age-group, as shown in panel *B*. Then, weighted statistics are assigned so that the weight factors for each duplication event add up to one, as shown in panel *C*. For details on the assignment of weight factors, see supplementary materials and methods (Supplementary Material online). After that, the statistics for different age-groups are juxtaposed for comparison, as shown in panel *D*.

2009). At present, therefore, such "longitudinal" analyses cannot be used to track the evolution of duplog positions. Nevertheless, "cross-sectional" analyses may be used to infer the mechanisms forming relatively physically distant duplog pairs in the species with finished genome assemblies. A cross-sectional analysis examines the distributions of physical relationships between duplogs at different ages (fig. 2). If relatively distant duplogs resulted from tandem duplications and subsequent rearrangements, they will account for only a small portion of the youngest age-group and their proportion will increase with age (fig. 2*D*). If, in contrast, most of them were created de novo via one-step mechanisms, their proportion will be more or less uniform across ages, and the youngest age-group will contain a similar proportion of distant duplogs as the whole set of duplogs does.

Motivated by the above consideration, we conducted cross-sectional analyses. Because this study requires high-quality genome assemblies, we restricted our analyses to six animal species with finished or nearly finished genome assemblies (fig. 1): human (*Homo sapiens*), mouse (*Mus musculus*), zebrafish (*Danio rerio*), nematode worm (*C. elegans*), and two fruit flies (*Drosophila melanogaster* and *D. pseudoobscura*). Our cross-sectional studies revealed several trends shared by most of the species studied, which point to the one-step creation of randomly oriented duplogs at relatively large physical distances, as well as behaviors specific to one species, especially *C. elegans*.

Materials and Methods

Selection of Subject Animal Genomes

As of 12 May 2011, there are 128 animal species whose genome sequences are assembled (see National Center for Biotechnology Information Genome Project Statistics at http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html). Unfortunately, most of them are draft assemblies based on the WGS sequencing. Recent studies showed that WGS assemblies grossly underrepresent duplicated regions of the genomes, especially recent duplications (She et al. 2004; Church et al. 2009).

Because this study critically depends on the exact compositions and positions of duplogs, we restricted our analyses to the animals that have high-quality genome assemblies. So far, only four animal species have genome assemblies of finished qualities based on clone-by-clone sequencing: human (*H. sapiens*; International Human Genome Sequencing Consortium 2004), mouse (*M. musculus*; Church et al. 2009), nematode worm (*C. elegans*; The *C. elegans* Sequencing Consortium 1998; Hillier et al. 2005), and fruit fly (*D. melanogaster*; Celniker et al. 2002; Ashburner and Bergman 2005). In addition to these

four animals, we also included zebrafish (*D. rerio*) and another fruit fly (*D. pseudoobscura*) in our subject species.

Zebrafish was added with the hope that we could uncover duplication mechanisms operated in the genome of the vertebrate ancestor. Although it is still a draft assembly, the zebrafish genome sequence (the Sanger Institute genome build version Zv7) is built by tiling bacterial artificial chromosome clone sequences and by supplementing it with the WGS assembly (Danio rerio Sequencing Project 2007). We thus considered the zebrafish genome as more suitable for the study of recent duplogs than the genome of any other fish such as medaka (Kasahara et al. 2007) or tetraodon (Jaillon et al. 2004), whose genome assemblies are based mostly on WGS. After preliminary analyses, we found that the number of recent duplication events in the *D. melanogaster* genome is about an order of magnitude smaller than in other animals. So we tried to enhance the statistics on fruit fly duplication events by analyzing another species. *Drosophila pseudoobscura* was chosen because its genome assembly is more than just a WGS assembly, as it is augmented by separate sequences of repeat regions (Richards et al. 2005).

Peptide and cDNA Sequence Information Used in This Study

We downloaded files of the gene transcript (cDNA) sequences and the peptide sequences predicted in the human (*H. sapiens*), mouse (*M. musculus*), and zebrafish (*D. rerio*) genomes from the FTP site of the Ensembl database (Hubbard et al. 2009) version 52. The cDNA and peptide sequence files for two fruit fly species (*D. melanogaster* and *D. pseudoobscura*) were downloaded from the October 2008 version of FlyBase (Tweedle et al. 2009). Those for nematode worm *C. elegans* were downloaded from WormBase (Harris et al. 2010) version WS200.

We only used cDNA sequences with peptide counterparts, excluding cDNA products of the mitochondrial genes. The genomic map of exons, exon–transcript relationship, transcript–gene relationship, and translation starts and ends of the gene transcripts (cDNAs) were extracted from the mysql dumps provided at the Ensembl FTP site for vertebrates. The corresponding information for fruit flies and that for *C. elegans* were extracted from the genome feature tables in the GFF format provided at the FTP sites of FlyBase and WormBase, respectively.

Duplogs from the Six Animal Species

We conducted a series of screenings to retrieve pairs of duplogs, or intraspecies duplicated sequences, that have duplicated relatively recently in the genomes of the six animal species. We conducted BlastP (Altschul et al. 1990) homology searches against the set of translated cDNA sequences from the six species and their respective outgroup species that are supposed to have diverged from the subject species much earlier than the period we studied. Five vertebrate species (chicken, *Xenopus tropicalis*, zebrafish, *Tetraodon nigroviridis*, medaka) were used as outgroup for human and mouse, and eight outgroup species (human, mouse, chicken, *X. tropicalis*, *T. nigroviridis*, fugu, medaka, and stickleback) were used for zebrafish. *Caenorhabditis japonica* and *Pristionchus pacificus* were two outgroup species used for *C. elegans*, while five *Drosophila* species (*D. persimilis*, *D. willistoni*, *D. virilis*, *D. mojavensis*, and *D. grimshawi*) as well as its counterpart species were used as outgroup for *D. melanogaster* and *D. pseudoobscura*.

We then screened the resulting homologs using the numerical cutoff of 35% peptide identity, as well as a "natural cutoff" determined from the best outgroup homologs if available (Ezawa et al. 2006). Average dSs between the queries and the orthologs from the outgroup species are much larger than 0.6 (and typically larger than 2). Therefore, the introduction of such a natural cutoff will never miss a substantial fraction of duplogs with $dS < 0.6$. The surviving duplogs were aligned at the peptide level with the query sequence via Smith and Waterman's (1981) algorithm, which is implemented in the "ssearch" program of the FASTA package (available at http://fasta.bioch.virginia.edu/fasta_www2/fasta_down.shtml). The resulting pairwise alignments were transformed into their cDNA counterparts. We then masked the CpG dinucleotides for human and mouse because these sites are known to be hypermutable in mammals (Ehrlich and Wang 1981). For zebrafish, *C. elegans*, *D. melanogaster*, and *D. pseudoobscura*, we masked the repeat regions to avoid the misassignment of duplogs due to repeat sequences. We discarded the alignments containing less than 150 unmasked nucleotide sites excluding gapped sites. Next, we estimated the counts of synonymous sites and dS between the query and intraspecies paralogs via the "yn00" program of the PAML package (Yang 1997). Finally, we only kept duplog pairs whose dSs are less than the threshold value of 0.6. Among these recently generated duplog pairs, we only kept for our analyses those pairs 1) that have 100 or more synonymous sites and 2) that consist only of genes mapped onto chromosomes.

Cross-Sectional Analyses Based on Duplog Gene Pairs

Our main results were obtained through event-based cross-sectional analyses explained in the next subsection. In this subsection, however, we will explain pair-based analyses, because they provide a foundation of the event-based analyses. First, to get a broad sense of the age dependence, we subdivided the evolutionary period from the upper bound of $dS = 0.6$ to the present ($dS = 0$) into three time intervals using the boundaries at $dS = 0.2$ and 0.4. After confirming that duplication events in the youngest interval of $0 \leq dS < 0.2$ are much more abundant than those in the other two intervals ($0.2 \leq dS < 0.4$ and $0.4 \leq dS < 0.6$) for most of the

species, we further subdivided this youngest class with the boundaries dSs = 0.01, 0.03, and 0.1 to get a finer view of the age dependence. Especially, dS = 0.01 was chosen as it is the smallest measurable distance under our condition of 100 or more synonymous sites. Then, we classified each duplog pair into an age-group according to the interval the dS falls into (fig. 2*B*). The age-groups were labeled as follows: C1, $0 \leq dS < 0.01$; C2, $0.01 \leq dS < 0.03$; C3, $0.03 \leq dS < 0.1$; C4, $0.1 \leq dS < 0.2$; C5, $0.2 \leq dS < 0.4$; and C6, $0.4 \leq dS < 0.6$.

After that, we obtained statistics on the physical proximity and relative transcriptional orientations for each age-group. The "physical proximity" of a gene pair is a combination of the linkage and the physical distance. The pair is called "linked" when the genes are on the same chromosome and "unlinked" when they are on different chromosomes. The physical distance of a pair is defined by the length (in base pairs) of the sequence between the coding regions of the member genes. For each species, three boundaries were chosen to divide the entire set of linked duplog pairs from each species into four subsets of almost equal sizes. Then, to facilitate the comparison, we arranged the distributions for different age-groups in order of age (fig. 2*D*).

Cross-Sectional Analyses Based on Duplication Events

The cross-sectional analyses based on duplog gene pairs could potentially result in biased distributions depending on the histories and physical properties of large families. To alleviate such bias, we conducted cross-sectional analyses based on duplication events, by counting each duplication event, instead of each duplog pair, as a unit. We first merged duplog pairs that share the member genes into clusters (or families) of duplogs by using a single-linkage algorithm. Each of the resulting clusters should consist of duplogs that originated from a common ancestor gene since the time measured by dS = 0.6. The dS has been used to approximate the nucleotide substitution rate under neutral evolution (e.g., Lynch and Conery 2000) because synonymous substitutions by definition do not change amino acids and therefore are under weak, if any, selective pressure. Then, we constructed a rooted phylogenetic tree for each cluster via UPGMA (Sneath and Sokal 1973). Because the tree is rooted, we can identify the duplication event each duplog pair diverged from. By collecting the duplog pairs diverged from each duplication event, and by assigning an appropriate weight factor to each of the duplog pairs, we converted statistics on duplog pairs into those on duplication events (fig. 2*C*). The weight factors must add up to unity across the duplog pairs diverged from each duplication event, for we count each duplication event as a unit. How we assigned a weight factor to each gene pair is described in the supplementary materials and methods (Supplementary Material online). Then, we subdivided the duplication events into six age-groups using the same set of boundaries for dS values as in the above subsection. Finally, we added together the weight factors for duplog pairs belonging to each age-group and each class of physical properties (fig. 2*D*). Note that, in the event-based analyses, the three boundaries for the physical distance were chosen to divide the entire set of linked duplication events into four subsets of almost equal sizes. Note also that, in our method, the number of duplication events is defined as the number of duplog copies generated during the time period in question.

Results

Statistics on the Six Animal Genomes

Basic statistics on the genomes of six animals used in this study are summarized in supplementary table S1 (Supplementary Material online). Broadly speaking, the genomes of invertebrates (*C. elegans*, *D. melanogaster*, and *D. pseudoobscura*) are an order of magnitude smaller than those of vertebrates (human, mouse, and zebrafish), whereas the numbers of genes are almost the same between invertebrates and vertebrates, with the former more than a half of the latter. So, we expect that the average physical distance between neighboring duplog genes should be an order of magnitude smaller in the invertebrates than in the vertebrates. This turned out to be roughly true, as we can see from the three boundaries of the inter-duplog physical distance at 25%, 50%, and 75% of all linked duplication events for each species (supplementary table S1, Supplementary Material online). For example, the median physical distances between linked vertebrate duplogs are 155, 95, and 78 kb for human, mouse, and zebrafish, respectively, whereas those between linked invertebrate duplogs are 7.4, 7.1, and 4.0 kb for *C. elegans*, *D. melanogaster*, and *D. pseudoobscura*, respectively.

Sets of Duplogs in the Six Animal Genomes

For each of the six animal genomes, with homology searches via Blast and a series of screening described in the Materials and Methods, we gathered a set of duplogs, or intraspecies paralogs, whose dSs are less than the threshold of 0.6 synonymous substitutions per synonymous site. We are interested in recent duplication events because only the short-term age dependence of duplog positions can unravel the mechanisms that form relatively physically distant duplogs, as explained in the Introduction. The threshold value of dS = 0.6 was chosen to avoid the historical correlation (or redundancy) between the human and mouse duplogs (fig. 1). The overall statistics on the set of duplog pairs is given in table 1. Broadly speaking, vertebrate genomes have an order of magnitude more duplog pairs than invertebrate genomes.

Figure 1 displays the time intervals used in this study as thick gray lines from the exterior nodes on the species

Table 1
Overall Counts of Duplogs and Duplication Events

	Duplog Pair Type[a]			
	Multiexon[a]	**Single Exon[a]**	**Mixed[a]**	**Total**
No. duplog pairs				
Human	4,524	789	823	6,136
Mouse	5,048	5,880	2,582	13,510
Zebrafish[b]	21,258 (3,071)	1,063 (277)	1,411 (187)	23,732 (3,535)
Caenorhabditis elegans	1,950	412	39	2,401
Drosophila Melanogaster	213	763	19	995
D. pseudoobscura	408	162	96	666
No. duplication events				
Human	1,215.3 [66.8]	263.2 [14.5]	340.8 [18.7]	1,819.2
Mouse	1,467.0 [46.0]	1,084.7 [34.0]	635.0 [19.9]	3,186.6
Zebrafish[b]	2,448.8 (871.1) [86.5 (83.4)]	188.3 (105.2) [6.7 (10.1)]	194.0 (67.7) [6.9 (6.5)]	2,831.0 (1,044.0)
C. elegans	1,340.0 [92.0]	103.2 [7.1]	13.8 [1.0]	1,457.0
D. melanogaster	115.5 [50.7]	98.1 [43.0]	14.3 [6.3]	228.0
D. pseudoobscura	324.5 [66.1]	106.5 [21.7]	60.0 [12.2]	491.0
Percentages of linked duplog generations among duplication events[c]				
Human	78.7	67.9	28.3	67.7
Mouse	67.9	76.8	25.4	62.5
Zebrafish[b]	67.9 (74.5)	73.2 (93.3)	70.3 (65.1)	68.4 (75.7)
C. elegans	82.3	71.9	42.8	81.2
D. melanogaster	97.4	99.0	62.9	95.9
D. pseudoobscura	69.0	66.2	52.0	66.3

NOTE.—We counted only those gene pairs mapped on chromosomes. The "number of duplication events" actually means the summation of weight factors from the gene pairs belonging to respective subsets. For details on the weight factors, see supplementary materials and methods (Supplementary Material online). The numbers therefore can be fractional and were rounded off to the nearest tenth. Numbers in brackets in "no. duplication events" denote the percentages of duplication events of each type (column) accounting for each species (row).

[a] Types of contributing gene pairs. "Multiexon" denotes a subset of gene pairs each consisting only of multiexon genes. "Single exon" represents a subset of pairs consisting only of single-exon genes. And "mixed" is designated for a subset of pairs each consisting of a single-exon and multiexon genes.

[b] Numbers in parentheses in this row are for a set of "stable" zebrafish duplogs that are consistently annotated on both the Zv7 and Zv8 assemblies (for details, see supplementary materials and methods, Supplementary Material online).

[c] "Linked" denotes a set of gene pairs each of which consists of genes on the same chromosome.

phylogeny. Time intervals corresponding to $0 \leq dS < 0.6$ considerably vary among the species because of the variation of synonymous substitution rate, measured per year, among the species. Broadly speaking, the neutral substitution rate is an order of magnitude higher in the invertebrates than in the vertebrates. Neutral substitution rate (per site/year) was estimated to be $(5.4 – 23) \times 10^{-8}$ for *C. elegans* and $(2.9 – 12) \times 10^{-8}$ for *D. melanogaster* (Cutter 2008), and those for human and mouse were estimated as $(0.8 – 1.2) \times 10^{-9}$ and $(2.5 – 5.0) \times 10^{-9}$, respectively (Yi et al. 2002). Consequently, the time intervals corresponding to $0 \leq dS < 0.6$ are an order of magnitude shorter for invertebrates than for vertebrates. As figure 1 indicates, the duplication events studied here are quite recent, hence the duplog pairs generated by the ancient WGD events (Ohno 1970; Jallion et al. 2004; Dehal and Boore 2005; Woods et al. 2005) should be negligibly few, if any, in our set of duplogs. We also know that animal genomes, especially mammalian genomes, have been bombarded with retrotransposition, and intronless retrotransposed duplicates of genes are abundant in these genomes (e.g., Babushok et al. 2007). To mitigate their influences, we divided duplog pairs into three categories: "multiexon" pairs consisting solely of genes with more than one exon, "single-exon" pairs consisting solely of genes with one exon, and "mixed" pairs consisting of both single-exon and multiexon genes (table 1). Because most retrotransposed duplicates are expected to be in the sets of single-exon and mixed pairs, we mainly used multiexon pairs to examine the patterns of duplication events due to mechanisms other than retrotransposition. In most cases, the multiexon duplog set behaved similarly to the whole duplog set. Because it is multiexon duplogs that are essential to this study, we will hereafter show the results on multiexon duplogs unless explicitly stated otherwise.

The total numbers of duplication events (with $dS < 0.6$) are more or less similar among the studied animals except two *Drosophila* species, which have experienced about an order of magnitude fewer duplication events compared with mouse (table 1), consistent with previous studies (Hedger and Ponting 2007; Zhou et al. 2008). Figure 3 shows the sizes of age-groups of all duplogs. We see that each of

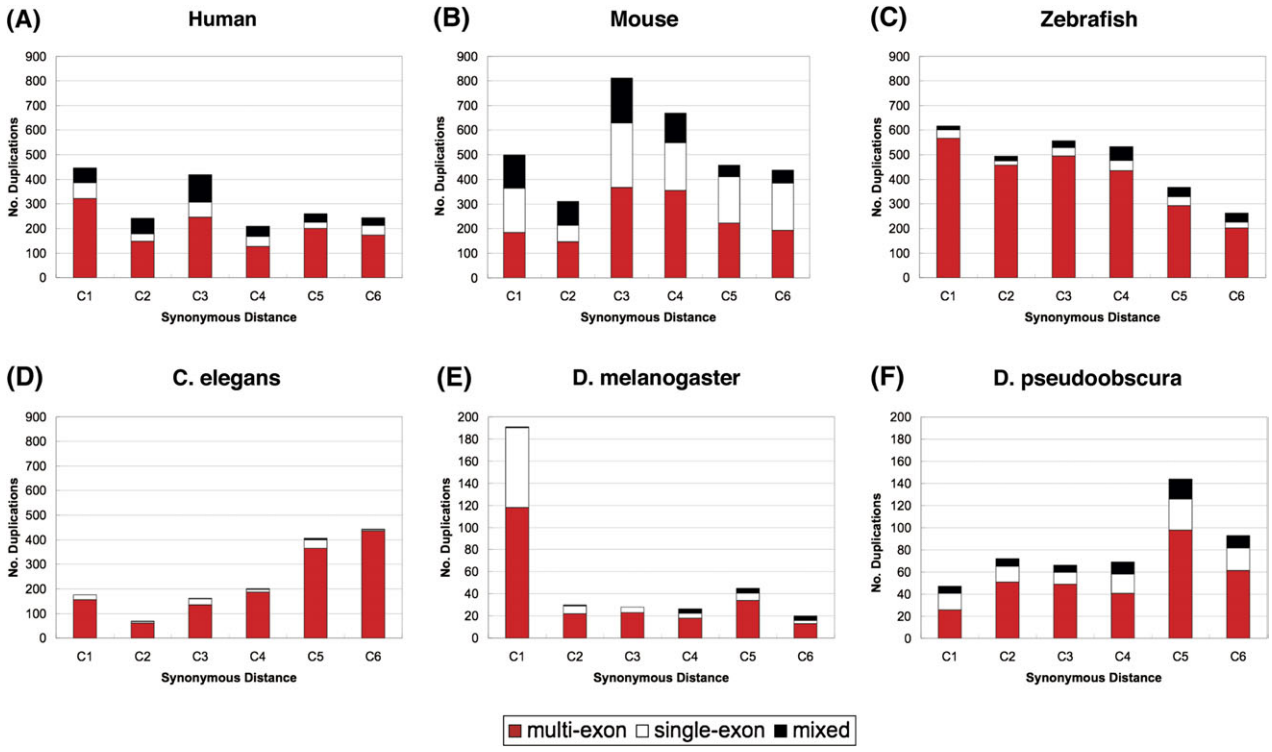

Fig. 3.—Sizes of age-groups of duplication events classified into contributions from multiexon, single-exon, and mixed pairs of duplogs. The red, white, and black bars stacked in each age-group represent the numbers of duplication events attributed to multiexon, single-exon, and mixed duplog pairs, respectively. Here, a "mixed" duplog pair consists of two duplogs, one multiexon and the other single exon. Only those pairs mapped on chromosomes are counted. Age-groups are defined by the time intervals measured in terms of the dS between duplogs. The age-groups used are as follows: C1 ($0 \leq dS < 0.01$), C2 ($0.01 \leq dS < 0.03$), C3 ($0.03 \leq dS < 0.1$), C4 ($0.1 \leq dS < 0.2$), C5 ($0.2 \leq dS < 0.4$), and C6 ($0.4 \leq dS < 0.6$). The panels *A*, *B*, *C*, *D*, *E*, and *F* show the graphs for human, mouse, zebrafish, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *D. pseudoobscura*, respectively. Caution must be exercised when comparing the age dependence of (observed) duplication events per unit time because the time interval (in dS) varies across the age-groups.

the age-groups consist of more than 100 duplication events in general, except those for fruit flies and the age-group with $0.01 \leq dS < 0.03$ for *C. elegans*. These sample sizes are appropriate for statistical analyses for most of the cases. Taking account of the approximately logarithmic scaling of the time intervals for the age-groups, the number of observed duplication events per unit time seems negatively correlated with the age of the events in all species. This was commonly observed in the previous analyses (e.g., Lynch and Conery 2000; International Human Genome Sequencing Consortium 2004) and is probably due to the loss of (functional) duplicate copies over time. Especially, the youngest age-group is expected to contain a substantial number of duplicates that are not fixed yet, as confirmed in recent studies (e.g., She et al. 2008; Zhou et al. 2008). Regarding the type of duplication events, multiexon duplogs account for 67%, 46%, 86%, 92%, 51%, and 66% of all duplogs in human, mouse, zebrafish, *C. elegans*, *D. melanogaster*, and *D. pseudoobscura*, respectively (table 1 and fig. 3). In mouse and *D. melanogaster*, single-exon duplogs give fairly large contributions. Functional analyses revealed that G-protein-coupled receptors including olfactory receptors and histones give major contributions to mouse and *D. melanogaster* duplogs, respectively.

Cross-Sectional Analyses

To figure out dominant mechanisms of physically relatively distant and/or inverted duplog formation, we examined the evolutionary patterns of positional relationships between duplogs, via cross-sectional analyses based on duplication events (fig. 2; see Materials and Methods for details). We first examined the composition of physical proximity, namely the linkage and the physical distance between the duplogs (fig. 4 and supplementary fig. S2, Supplementary Material online). In the set of all duplication events (supplementary fig. S2, Supplementary Material online), the proportion of unlinked events seems to show different age dependence from species to species. For events resulting in multiexon duplog pairs (fig. 4), the proportion of unlinked events seems to either remain almost constant (zebrafish, *D. melanogaster*) or gradually increase with the age for a while (human, mouse, *D. pseudoobscura*), except *C. elegans*. The latter behavior appears consistent with the production of unlinked duplogs from linked ones via genomic rearrangement

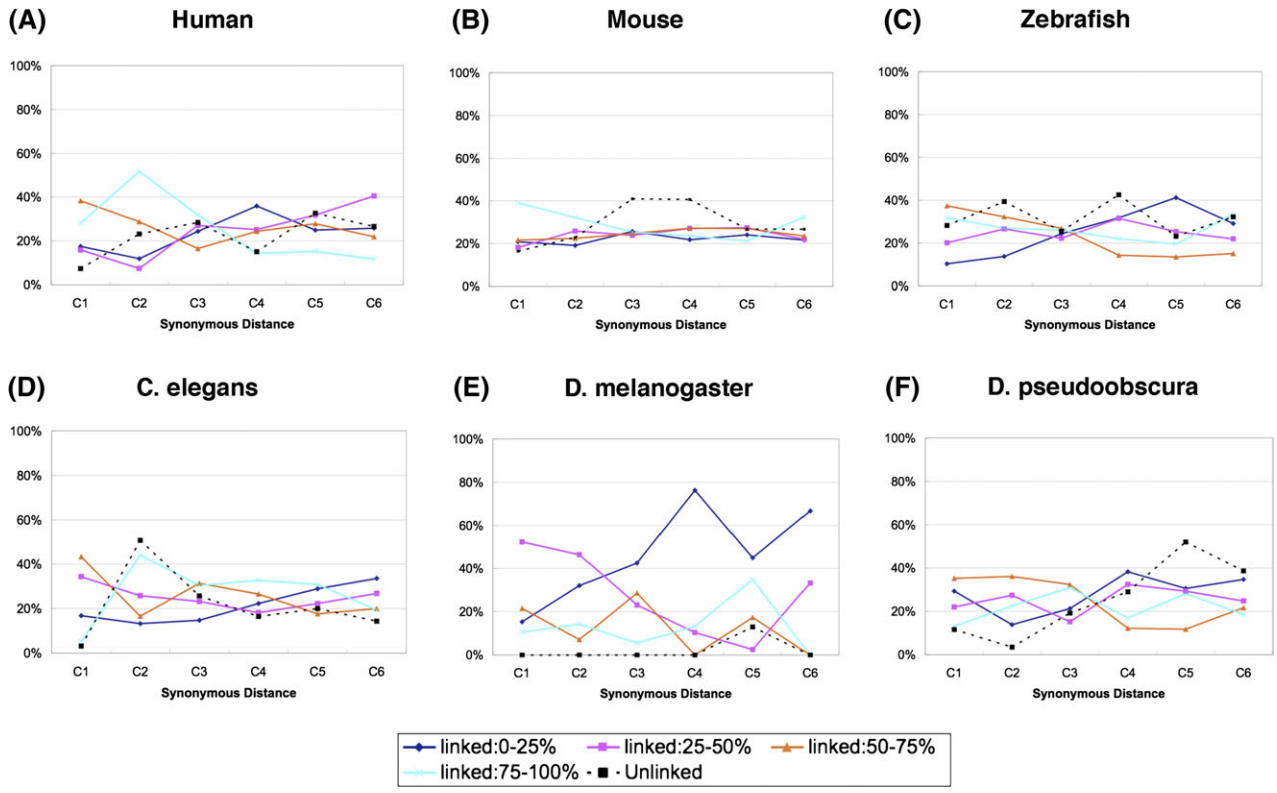

Fig. 4.—Age dependence of the composition of physical proximities between multiexon duplogs. In each panel, the black dashed line shows the age (in dS) dependence of the proportion of unlinked duplication events out of all multiexon duplication events mapped on chromosomes. The solid lines colored blue, pink, orange, and light blue, respectively, show the age (in dS) dependence of the proportions of the first, second, third, and fourth quartiles of the physical distance in the set of linked multiexon duplogs.

(e.g., Friedman and Hughes 2003, 2004; Conceição and Aguadé 2008). In some species, for example, mouse and zebrafish, however, the unlinked multiexon events account for a considerable portion of the youngest age-group, suggesting their de novo generation via mechanisms other than retrotransposition. Within the linked multiexon duplication events of each studied animal, the proportions of the third and fourth quartiles of the physical distance either remain almost the same through $0 \leq dS < 0.6$ or gradually decrease with age except for the fourth quartile in *C. elegans* (fig. 4; for statistical test results, see supplementary table S2, Supplementary Material online). This indicates that most of linked duplogs belonging to the most distant class (top 25% of events in the physical distance) were created de novo via mechanisms different from retrotransposition. *Caenorhabditis elegans* is unusual in that the proportion of the fourth quartile (top 25% in the physical distance) is fairly small (ca. 5%) in the youngest age-group.

Next, we examined relative orientations between linked multiexon duplogs. Overall, the proportion of inverted events (tail-to-tail and head-to-head events) varies across species from ca. 20% in *D. melanogaster* to over 40% in human (table 2). To get a clue about whether this is the nature of duplication events themselves or due to the secondary changes, we examined the age dependence of the composition of relative orientations (fig. 5) within linked multiexon duplogs. The proportion of inverted events either remains nearly constant throughout $0 \leq dS < 0.6$ or gradually decreases as the age increases (fig. 5; for statistical test results, see supplementary table S2, Supplementary Material online). And the proportion of tail-to-tail events and that of head-to-head events are approximately equal, aside from some fluctuations. This indicates that most of the inverted pairs in the genome are created from the beginning via mechanisms other than retrotransposition and did not result from the inversion of tandemly duplicated pairs. *Caenorhabditis elegans* seems anomalous: The proportion of inverted events in the age-group $0.01 \leq dS < 0.03$ (77%) is significantly larger than in the age-group $dS < 0.01$ (49%; P value = 0.0056 in Fisher's exact test).

The age dependences of the compositions of physical distance and of relative orientations observed above (figs. 4 and 5) seem to indicate one-step creation of duplogs with relatively large physical distance and/or inverted orientation. To further clarify the nature of such duplication mechanisms, we need to examine the physical distance dependence of the relative orientation composition, as well as the age dependence of such dependence. For this purpose, we first examined how the composition of relative orientations

Table 2
Compositions of Relative Orientations

Relative Orientation[a]	Head-to-Tail[a]	Tail-to-Tail[a]	Head-to-Head[a]	Total
All linked duplication events[b]				
Human	716.4 (58.2)	255.6 (20.8)	259.6 (21.1)	1,231.6
Mouse	1,374.7 (69.0)	301.3 (15.1)	315.4 (15.8)	1,991.5
Zebrafish	1,443.3 (74.6)	236.6 (12.2)	256.1 (13.2)	1,935.9
Zebrafish (stbl)[c]	623.7 (78.9)	87.5 (11.1)	79.5 (10.1)	790.7
Caenorhabditis elegans	745.2 (63.0)	190.1 (16.1)	247.1 (20.9)	1,182.4
Drosophila melanogaster	176.0 (80.5)	17.3 (7.9)	25.4 (11.6)	218.7
D. pseudoobscura	219.8 (67.5)	50.8 (15.6)	55.0 (16.9)	325.6
Linked duplication events (multiexon duplog pairs only[d]**)**				
Human	539.5 (56.4)	203.6 (21.3)	213.3 (22.3)	956.4
Mouse	694.4 (69.7)	145.0 (14.5)	157.4 (15.8)	996.8
Zebrafish	1,220.5 (73.4)	222.6 (13.4)	218.7 (13.2)	1,661.8
Zebrafish (stbl)[c]	496.8 (76.6)	78.7 (12.1)	73.1 (11.3)	648.6
C. elegans	709.7 (64.4)	170.6 (15.5)	222.1 (20.1)	1,102.4
D. melanogaster	93.0 (82.6)	10.5 (9.4)	9.0 (8.0)	112.5
D. pseudoobscura	150.8 (67.3)	38.5 (17.2)	34.6 (15.4)	224.0

NOTE.—The "number of duplication events" actually means the summation of weight factors from the gene pairs belonging to respective subsets. For details on the weight factors, see supplementary materials and methods (Supplementary Material online). The numbers therefore can be fractional and were rounded off to the nearest tenth. The percentage (in parentheses) in each cell is the proportion that the relative orientation in question (column) accounts for in the species in question (row).

[a] Relative transcriptional orientation of the contributing pairs. "Head-to-tail," "tail-to-tail," and "head-to-head" denote, respectively, the gene pairs of 5′-3′ 5′-3′, 5′-3′ 3′-5′, and 3′-5′ 5′-3′ orientations.

[b] A set of all duplication events whose resulting genes are both mapped on the same chromosomes.

[c] A set of "stable" zebrafish duplogs that are consistently annotated on both the Zv7 and Zv8 assemblies (for details, see supplementary materials and methods, Supplementary Material online).

[d] Only contributions from the multiexon gene pairs are counted.

depends on the physical distance (fig. 6). The proportion of inverted pairs in multiexon duplogs increases as the duplogs get more separated, reaching approximately half for the most distant class (top 25% in the physical distance), and the ratio between tail-to-tail and head-to-head events was nearly 1:1 for most classes of the physical distance. This pattern was also observed in previous studies (e.g., Shoja and Zhang 2006). By itself, the pattern is consistent both with tandem duplications followed by genomic rearrangements such as inversions and with the de novo creation of relatively distant duplogs with almost random orientations. When the pattern is combined with the age dependence of compositions of physical distance (fig. 4) and relative orientation (fig. 5), however, the latter scenario seems more plausible.

To corroborate this idea, we conducted a cross-sectional analysis of the proportion of inverted pairs, separately for each of the four physical distance categories (fig. 7). (The sample sizes of the subsets, each specified by an age-group and a physical distance class, are available on request to the first or last author.) We should note that most of the subsets for *D. melanogaster* and *D. pseudoobscura* and some of the subsets for other animals have sizes that are too small (often less than 10) to give statistical significance. Taking account of fluctuations due to sampling errors, the proportion of inverted pairs in each quartile appears either nearly constant throughout the time interval $0 \leq dS < 0.6$ or gradually decreasing with age with a few exceptions (fig. 7; for statistical test results, see supplementary table S3, Supplementary Material online). This is true for three vertebrates and possibly for *Drosophila*, except for the second quartile in human, which exhibited a U-shaped age dependence. Especially, the proportion of inverted events in the youngest age-group ($dS < 0.01$) did not differ so much from that in the whole set ($0 \leq dS < 0.6$).

To see whether this pattern can be explained solely via tandem duplications and subsequent rearrangements, we compared the observed proportions of inverted duplogs in the youngest age-group ($dS < 0.01$) with theoretical expectations. We restricted our analysis mainly to the intersection of the youngest age-group ($dS < 0.01$) and the third quartile (50–75% from the bottom) of the physical distance. We avoided using the fourth quartile (top 25%) because the duplog pairs in this class could have indefinite probabilities of rearrangements due to unbound physical distance. For details on this analysis, see "Theoretical Estimation of the Proportion of Inverted Duplogs" subsection of supplementary materials and methods (Supplementary Material online). Here, we only note that we took account of the inversion rate disparity between duplication-rich regions and the remaining genomic regions. Based on the recent genome-wide analyses (Newman et al. 2005; Ranz et al. 2007), we estimated that recently duplicated regions on average underwent inversions ca. 67 times more frequently than the remaining regions for human, and that the rate disparity

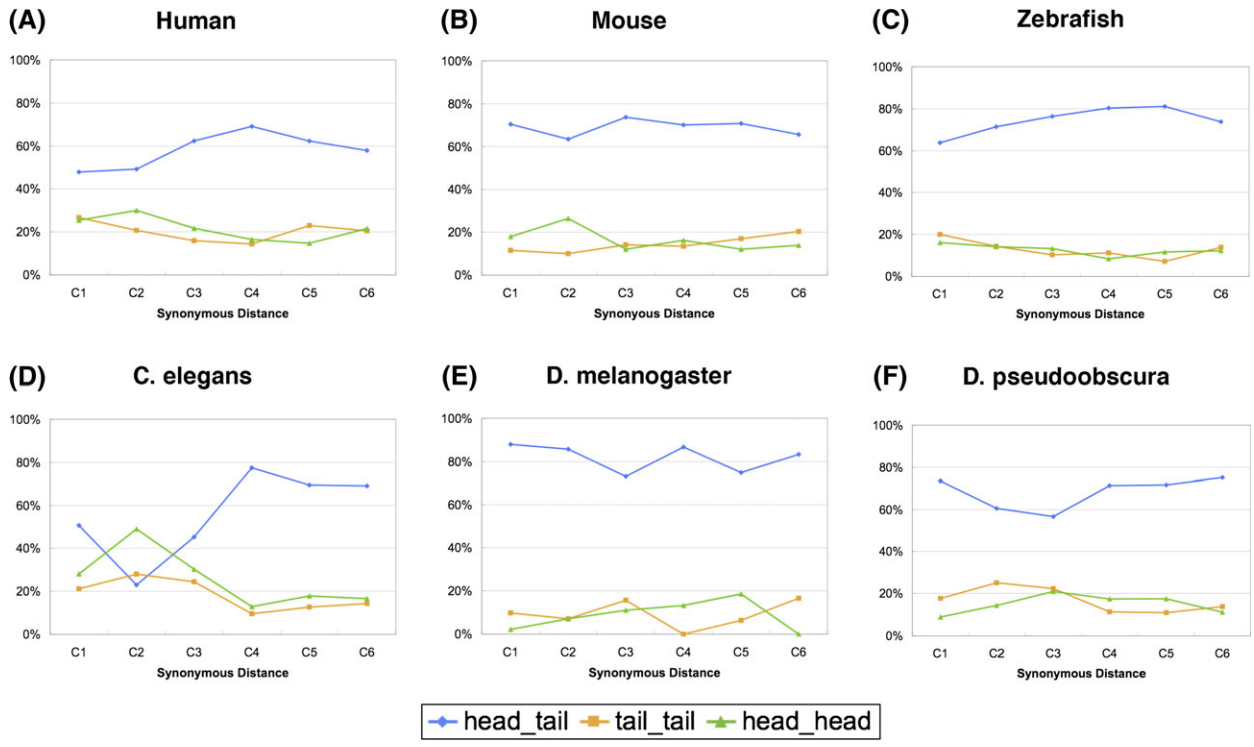

Fig. 5.—Age dependence of the composition of transcriptional relative orientations between linked multiexon duplogs. Each line graph shows the age dependence of the proportions of relative orientations in the set of linked multiexon events in each species. The levels of the light blue, light orange, and light green lines at each category represent the proportions of duplication events attributed to "head-to-tail," "tail-to-tail," and "head-to-head" gene pairs, respectively, which in turn mean gene pairs with the (5′-3′ 5′-3′), (5′-3′ 3′-5′), and (3′-5′ 5′-3′) configurations, respectively.

was ca. 162 times for *D. melanogaster* (supplementary materials and methods, Supplementary Material online). As the small *P* values in supplementary table S4 (Supplementary Material online) show, the theoretical estimation via the model of tandem duplications and subsequent chromosomal rearrangements alone fails to explain the observed proportion of inverted duplogs, which is about 4–15 times larger than the expectation, for any of the six animals tested.

It must be noted that the pattern of *C. elegans* duplogs in figure 7 totally differed from that of vertebrate duplogs. For this species, we observed a high proportion (>80%) of inverted duplogs in the bottom 25% of physical distance in dS < 0.03 followed by a plunge in the proportion (down to <20%) during $0.03 \leq$ dS < 0.10 (panel *D* in fig. 7). The second class (bottom 25–50% in the physical distance) also seems to show a similar but weak trend.

Reanalyses on Zebrafish Using a Stable Set of Duplogs

Our cross-sectional analyses showed that zebrafish duplogs behave similarly to mammalian duplogs. One caveat to these results is that the zebrafish genome assembly we used is still a draft assembly, and, in the worst case, all our findings on zebrafish duplogs could be artifacts stemming from the poor-quality portion of the assembly. After the first round of our analysis, an improved version of the draft zebrafish genome assembly, build Zv8, came out (The Danio rerio Sequencing Project 2008). So, we constructed a "stable" set of zebrafish duplogs consisting only of those duplogs that are mapped on both assemblies and whose annotations remain unchanged in the two assemblies (supplementary materials and methods, Supplementary Material online). This stable set should mostly consist of duplogs mapped on the clone-based portion of the genome assembly, and should therefore represent an almost random sampling of the duplogs from the finished assembly. Although the total number of duplication events reduced to approximately one-third (2,831 for Zv7 vs. 1,044 for "stable"; see table 1), the observed patterns remained almost unchanged (supplementary fig. S3, Supplementary Material online). We therefore believe that our observation was not an artifact and that zebrafish duplogs also share the patterns displayed by mammalian duplogs.

Cross-sectional Analyses Based on Mid-Intron Sequence Divergence

In this study, we used the dS between duplogs as a proxy to the duplication date. Previous studies revealed that

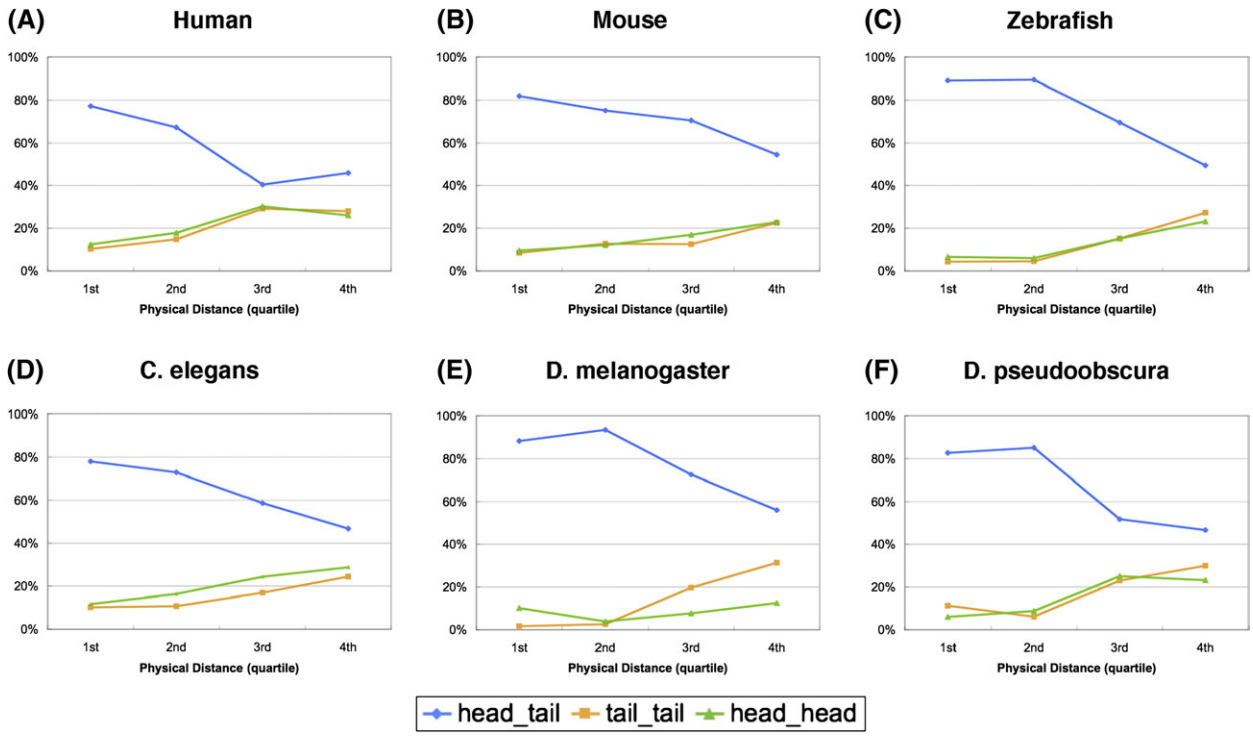

Fig. 6.—Physical distance dependence of the composition of transcriptional relative orientations between linked multiexon duplogs. Each line graph shows the physical distance dependence of the proportions of relative orientations in the set of linked multiexon events in each species. The meanings of the three colors are the same as in figure 5.

synonymous substitutions are also under weak selection over translation efficiency (Stenico et al. 1994; Akashi 1995), splicing efficiency (Parmley et al. 2006; Warnecke and Hurst 2007), and so on (e.g., Chuang and Li 2007). For human, mouse, and *Drosophila*, effects of selections on synonymous substitutions were shown to be so weak that dS can well approximate the neutral nucleotide divergence (e.g., Parmley et al. 2006; Cutter 2008). For *C. elegans*, in contrast, dS strongly correlates with the codon usage bias (Cutter 2008), questioning the use of raw dS values as proxies for the divergence dates. Regarding zebrafish, we do not know any such studies on dS. Because data on these two species contribute important conclusions in this study, we reconducted the cross-sectional analyses using the mid-intron sequence divergence (dI) as an alternative proxy to the divergence date (for details, see supplementary materials and methods for details, Supplementary Material online). Although dI is not completely free from selection either, the nature of (weak) selection on dI is different from that on dS. Consistent features between the two cross-sectional analyses, one based on dS and the other on dI, will therefore suggest the authenticity of the features. Supplementary figures S4 and S5 (Supplementary Material online) show that the main features for zebrafish remain valid and so do those for *C. elegans*, indicating that our conclusions are biologically significant. It should be noted, however, that the sample size for *C. elegans* is quite small (supplementary fig. S5*A*, Supplementary Material online). This may have obscured the timing of the switching from the inverted-predominance to the direct-predominance in the first quartile of the physical distance (supplementary fig. S5*E*, Supplementary Material online).

Discussion

"Fourth Mode" of Gene Duplication

In this study, we characterized the short-term evolution of duplogs using cross-sectional analyses. We analyzed six animals with high-quality genome assemblies: human, mouse, zebrafish, *C. elegans*, *D. melanogaster*, and *D. pseudoobscura*. Albeit with one or two exceptions, the duplog sets in the studied six animals shared the following evolutionary patterns, mainly among multiexon duplogs and mostly among all duplogs as well:

(i) Except for *C. elegans*, the proportions of the third and fourth quartiles (50–100%) of physical distance in the linked duplogs are almost unchanged across age-groups or decrease gradually as the age increases (fig. 4);
(ii) The proportion of inverted duplication events is almost unchanged through the age interval of $0 \leq dS < 0.6$ or decreases gradually as the events age (fig. 5);
(iii) The proportion of inverted events, with all ages mixed up, increases as the physical distance increases. Inverted events account for ca. 10–20% and approximately half in the bottom 25% and the top 25% of the physical distance, respectively (fig. 6); and

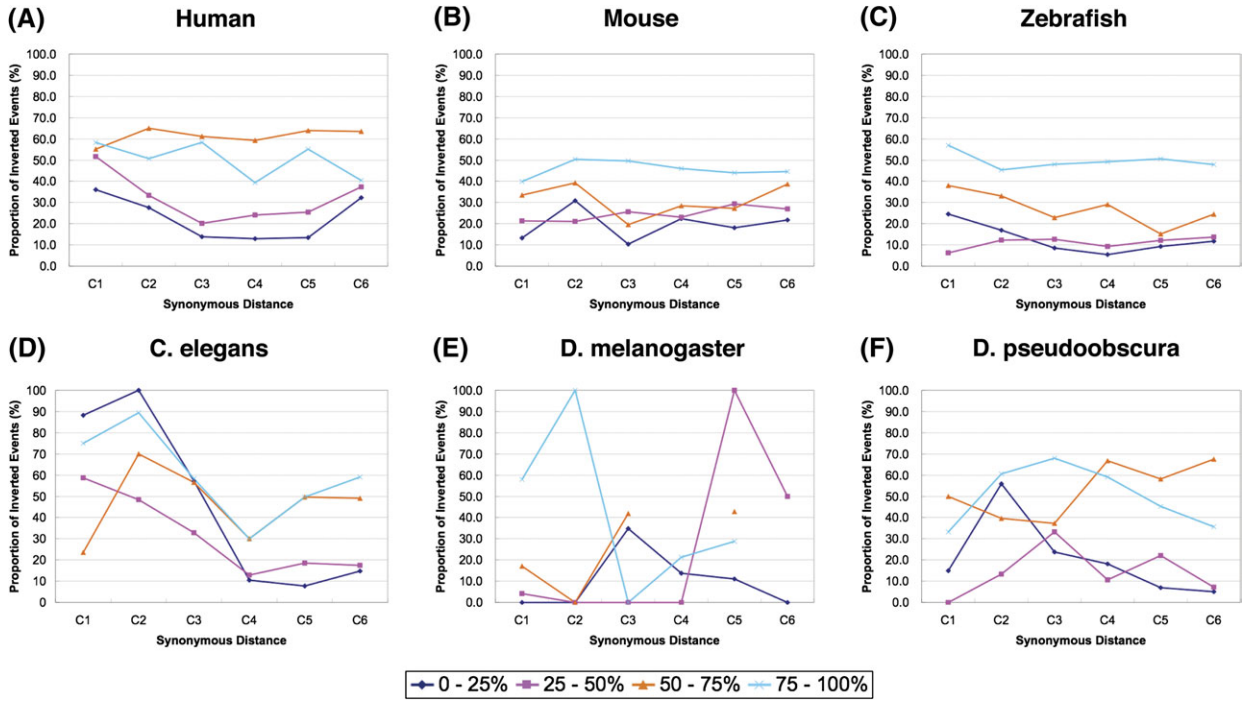

Fig. 7.—Proportions of inverted duplication events as functions of duplication date (in dS) in different classes of the physical distance (for linked multiexon duplog pairs). Each line graph shows the age (in dS) dependence of the proportion of inverted duplication events in a quartile classified by the physical distance between duplogs, contributed from linked multiexon duplog pairs. The blue, magenta, orange, and cyan lines represent, respectively, the first, second, third, and fourth quartiles of linked duplication events. Because *Drosophila melanogaster* and *D. pseudoobscura* experienced only small numbers of duplication events, stochastic fluctuations are so large that the proportions in panels *E* and *F* are not statistically meaningful in many data points (see also supplementary table S3, Supplementary Material online).

(iv) For the three vertebrates, the proportion of inverted events in each class of the physical distance remains almost constant across the age-groups or gradually decreases with age. The proportion hovers around 10–20% and around half in the bottom 25% and the top 25% of the physical distance, respectively (fig. 7).

Caenorhabditis elegans posed an exception against the patterns (i), (ii), and (iv), which will be discussed later. For the two *Drosophila* species, the patterns corresponding to (iv) were unclear, likely due to small sample sizes (fig. 7*E* and *F*).

The "static" pattern (iii) has been repeatedly observed also in the previous genome-wide analyses (e.g., Shoja and Zhang 2006). As far as we know, however, this study is the first to report patterns (i), (ii), and (iv) of the age dependence for animal duplogs. We emphasize that patterns (i)–(iv) were observed even for multiexon duplogs. This precludes the explanation via retrotransposition because such mechanism mostly creates single-exon duplogs (see, e.g., Babushok et al. 2007), unless premature long RNA transcripts were reverse-transcribed. Pattern (iii) is consistent with the classical view of tandem duplication followed by genomic rearrangements (e.g., Friedman and Hughes 2003, 2004; Conceição and Aguadé 2008; Hu et al. 2008). In contrast, it is difficult to explain patterns (i), (ii), and (iv) with this classical view. Patterns (i) and (ii) show that the most distant linked class (top 25% in the physical distance) and inverted pairs were already present in considerable proportions almost as soon as the duplications occurred. And pattern (iv) indicates that inverted pairs remain accounting for around half of the most distant linked class (top 25% in the physical distance) throughout the time interval of $0 \leq dS < 0.6$ we studied. Our statistical test showed that pattern (iv) cannot be explained by tandem duplication and subsequent chromosomal rearrangements alone (supplementary table S4, Supplementary Material online; see also Results). By analogy, we expect that such a mechanism alone cannot explain the patterns (i) and (ii), either.

The most natural interpretation for these observations would be "yet another mode of gene duplication" that is different from the three well-known duplication mechanisms: tandem duplication, retrotransposition, and whole-genome duplication. Here, we term this duplication mode as "drift" duplication. Its physical distance distribution appears to peak around a few hundred kilobase pairs for vertebrates and a few dozen kilobase pairs for invertebrates, which is in between those of tandem duplication (short

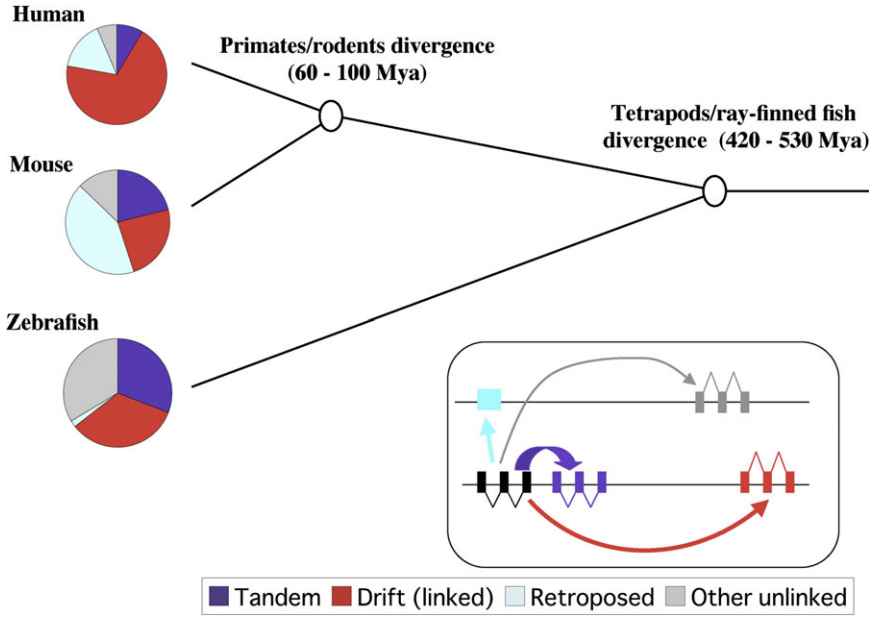

Fig. 8.—Evolution of the duplication mode composition in vertebrates. The pie charts at the exterior nodes of the phylogenetic tree represent the estimated compositions of the duplication modes in extant vertebrate species: human, mouse, and zebrafish. The inset above the key (on the bottom right) illustrates the color code of the duplication modes. A thin horizontal line represents a chromosome, and solid short rectangles (exons) connected by sharply bent lines represent genes, with black ones original and four colored ones duplicate copies.

range) and retrotransposition (long range, i.e., mostly unlinked). Drift duplications are almost randomly oriented, with the frequency ratio of head-to-tail : tail-to-tail : head-to-head ≈ 2:1:1, as opposed to tandem duplications due to unequal crossing-overs, which are mostly head-to-tail. A drift duplication can also create multiexon duplogs, as opposed to retrotransposition, whose products are mostly intronless. Retrotransposition is also drifting in a sense; however, it always passes through the RNA stage. This is the clear difference from drift duplication. With this name, "drift," we also implied that even some interchromosomal duplications may be attributed from drift duplication, though RNA-mediated duplications may be more frequent among interchromosomal duplications. DNA molecules for drift duplications are usually much larger than those for RNA-mediated duplications and may not be able to move to different chromosomes easily. This conjecture should be examined in future studies.

It is not certain at this point whether this mode, namely drift duplication, is due to a single duplication mechanism or not. We should note that this might also be explained via a considerable proportion of duplication events in extremely unstable regions undergoing tremendously frequent rearrangements (Eichler and Sankoff 2003; Pevzner and Tesler 2003; Murphy et al. 2005). When estimating the theoretical proportions of inverted duplogs in supplementary table S4 (Supplementary Material online), we took account of the rearrangement rate disparity between recently duplicated regions and the remaining genomic regions. Still, it is possible that the duplicated regions in fact consist of relatively stable regions and extremely unstable regions. In the future, analyses of finished genome assemblies in closely related species or analyses of structural variations within species with finished reference genomes will reveal whether our observation is due to rearrangements in extremely unstable regions or to genuine creation mechanisms of relatively distant and randomly oriented duplogs. Here, we will simply assume the latter and continue our discussion.

Recent analyses of the human genome revealed rich instances of recent SDs nonrandomly distributed across the human genome (; Bailey et al. 2002; Bailer and Eichler 2006). A substantial proportion of such SDs was found to be of interspersed type (Bailey et al. 2002, 2003; Bailey and Eichler 2006). The finished mouse genome assembly analysis also revealed rich instances of recent SDs (She et al. 2008; Church et al. 2009). Although these authors emphasized the differences of mouse SDs from human SDs such as the enrichment of tandem duplications, their classification of linked SDs was quite coarse grained at low physical distance resolutions of 1 Mb. In contrast, our cross-sectional analysis strongly indicates that a considerable fraction of mouse SDs classified as "tandem" so far, as well as a majority of interspersed SDs, are created de novo via the drift duplication, and that de novo creation of duplogs via the drift duplication is a common nature of mammalian genomes. This in turn suggests that the drift duplication has been actively operating in the genome since the common ancestor of placental mammals (fig. 8).

Actually, we could further extend the period during which the drift duplication has been active. We observed

that zebrafish duplogs also exhibit the patterns similar to mouse (rather than human) duplogs (figs. 4–7), indicating a substantial contribution of the drift duplication in this species. Hence, we can infer that the drift duplication has been generating duplicate genes in rates comparable with those of tandem duplications at least since the common ancestor of tetrapods/ray-finned fish (fig. 8). So far, tandem duplications due to unequal crossing-over have often been regarded as a dominant mechanism to produce recent duplicate genes (e.g., Shoja and Zhang 2006; Pan and Zhang 2007, 2008). To the best of our knowledge, the present study is the first to point out that the drift duplication also has produced as many duplicate genes as, or even more than, tandem duplications at least in vertebrate genomes.

Various modes of duplications among three vertebrate species are compared in figure 8. The drift duplication, shown in red, constitutes a major part in all three species, especially in human. While retroposed duplication is most frequent in mouse, other unlinked duplication is more than one-third of the total in zebrafish. Although we restricted the drift duplications to intrachromosomes, it may be possible that the same molecular mechanisms are also involved in interchromosomal duplications classified as "other unlinked" in this figure. If so, this type may be the major mechanism for duplog generation.

Let us now discuss the mechanism causing the drift duplication. In the human genome, over one-fourth of the recent interspersed SDs seem to be explained by Alu-mediated mechanisms (Bailey et al. 2003), but about a half of the events seem unaccounted for. For mouse, many but not most of the SDs are bounded by LINEs (long interspersed elements) or LTRs (long terminal repeats), suggesting mechanisms mediated by such repeats (She et al. 2008). As for zebrafish or *C. elegans*, we do not know previous studies on the mechanisms potentially causing the drift duplication. Regarding fruit flies, a recent large-scale experimental screening of eight genomes in the *D. melanogaster* subgroup identified 17 duplicates generated with the mediation of repetitive elements (Yang et al. 2008). It would be interesting to carefully examine the boundaries of the SDs resulting from the drift duplication in zebrafish and *C. elegans*, and if many of them turn out to be mediated by species-specific repeat sequences. We have to note, however, that a majority of the SDs seem to have been caused by repeat-independent mechanisms (Zhou and Mishra 2005). Determining the specific mechanisms responsible for the drift duplication would require correlating features in the flanking sequences of duplicated regions (Bailey et al. 2003; Zhou and Mishra 2005) with positional relationships between the duplicated regions.

Recent "Expansion" of Histone Tandem Array in *Drosophila melanogaster*

We observed that, regarding all duplogs in the two fruit flies, whereas *D. pseudoobscura* duplogs seem to conform to the general patterns (i)–(iv) discussed above, *D. melanogaster* duplogs were eccentric in the sense that they did not follow any of patterns (i)–(iv). Especially, the proportion of inverted events was quite small (ca. 9%) in $dS < 0.01$ for the top 25% of the physical distance, and this seemed to be causing most of the eccentric patterns. When restricted to multiexon events, however, the proportion of inverted events was around half in the subclass in question (fig. 7*E*). So, we carefully examined the single-exon pairs that have $dS < 0.01$ and belong to the top 25% in the physical distance. We found that tandem repeats of the histone gene cluster (Celniker et al. 2002) make an overwhelming contribution of 63 events to this subclass. When we reconducted the cross-sectional analyses after removing these histone tandem repeats, the remaining *D. melanogaster* duplogs largely followed the patterns commonly observed in other animals (except *C. elegans*) (data not shown). This implies that, aside from the huge tandem array of the histone gene cluster, duplogs of the two fruit flies also follow the general rules (i)–(iv). We are not saying that the huge tandem array of histones is specific to *D. melanogaster*. Such repeat might be missing in the *D. pseudoobscura* genome because it is still a WGS-based draft assembly. Whether this is true or not will be revealed if a clone-based analysis is conducted. Rather, we can say that the effect of this tandem repeat stood out because the whole set of *D. melanogaster* duplication events is pretty small (table 1 and fig. 3*E*), which made the patterns of *D. melanogaster* duplogs "appear" eccentric. Tandem duplication may be a predominant duplication mechanism in the fruitfly genomes. Still, it seems that these genomes are also undergoing the drift duplication to some degree. This raises the possibility that the drift duplication have been operating since the common ancestor of bilateral multicellular animals. It would be premature, however, to conclude this conjecture now. The sample size to support this is too small, totaling ca. 11 multiexon events with $dS < 0.01$ belonging to the top 25% of the physical distance in the *D. melanogaster* and the *D. pseudoobscura* genomes.

At face value, the large tandem array of *D. melanogaster* histone gene clusters appears to indicate that a burst of tandem duplications occurred very recently to amplify the histone cluster. This apparent evolutionary pattern is, however, also consistent with the strong concerted evolution, as observed for the tandem clusters of ribosomal RNA genes (for review, see Eickbush and Eickbush 2007). The pattern could also be explained with strong selective pressures on synonymous substitutions, maybe due to the requirements to conserve mechanisms for histone gene regulation. To determine which mechanism actually generated the observed pattern of the histone gene clusters in *D. melanogaster*, it would be inevitable to conduct evolutionary analyses of the histone clusters based on clone-based sequences from closely related *Drosophila* species.

Unique Features in *Caenorhabditis elegans* Duplogs

Caenorhabditis elegans duplogs displayed quite different patterns than the duplogs of other five animals studied, posing exceptions against (i), (ii), and (iv) (figs. 4, 5, and 7). Especially against (iv), the closest class (bottom 25% in the physical distance) of linked duplogs switched from inverted-rich states in $0 \leq dS < 0.03$ to direct (head-to-tail)-rich ones in $dS \geq 0.1$ (panel *D* of fig. 7). The predominance of inverted pairs among recent ($0 \leq dS < 0.03$) physically close duplogs is consistent with the finding by Katju and Lynch (2003) and, at the same time, specific to *C. elegans* among the six animals studied. This suggests some mechanisms specific to the *C. elegans* lineage. Katju and Lynch (2003) mainly proposed two mechanisms that may have caused this pattern: illegitimate recombination and frequent elimination of head-to-tail duplog pairs. Here, we would like to point out another possible mechanism: preferential homogenization between closely located inverted duplogs. Instances of intense homogenization between inverted duplicons are known, for example, on the male-specific region (MSR) of human Y chromosome (Rozen et al. 2003). And it should be noted that the homologous recombination rate of *C. elegans* is very low because it mostly self-fertilizes. Whether or how *C. elegans* has evaded Muller's ratchet has therefore been discussed (e.g., Loewe and Cutter 2008). It is possible that intense homogenization between inverted duplogs have liberated the recombination-poor *C. elegans* genome at least partially from Muller's ratchet, as was proposed for the recombination-free MSR of human Y chromosome (Rozen et al. 2003).

In contrast, it also merits a mention that the patterns for *C. elegans* duplogs are actually quite similar to those for vertebrate duplogs if we focus on the age-groups with $0.1 \leq dS < 0.6$, which were not studied by Katju and Lynch (2003). To see whether this is just a coincidence or due to shared underlying mechanisms, especially the drift duplication, across bilateral animals would require more detailed analyses.

Summary and Future Tasks

Clone-based genome assemblies of finished quality are crucial to the genome-wide study of recent duplogs, or intraspecies duplicate genes (She et al. 2004; Church et al. 2009). To elucidate genome-wide trends of recent evolution of positional relationships between duplogs, an ideal way would be to compare the chromosomal positions of orthologs of duplicate genes using finished-quality genome assemblies of closely related species. Unfortunately, at present, finished genome assemblies are too sparse in the animal kingdom to conduct such an ideal analysis. As a practical substitute, we applied cross-sectional analyses to the sets of recent duplogs from the six animal genomes of finished or nearly finished qualities. The analyses uncovered a common but unexpected feature, namely substantial contributions from drift duplication. The analyses also illustrated idiosyncrasies of duplogs in the animal species studied, especially the sharp drop with age of the proportion of inverted duplogs during $0.03 \leq dS < 0.1$ in the closest quartile of *C. elegans* duplogs (fig. 7*D*), as well as the "burst" of tandem duplications in the most distant quartile of *D. melanogaster* duplogs. So far, tandem duplications due to unequal crossing-over have often been invoked as a dominant mechanism to generate duplicate genes (see, e.g., Friedman and Hughes 2003, 2004; Shoja and Zhang 2006; Pan and Zhang 2008). For the first time, our cross-sectional analyses revealed that the drift duplication has also been generating duplicate genes in rates comparable with or even higher than the rates of tandem duplications at least since the vertebrate ancestor. It would be premature, however, to conclude that the patterns we observed are universal even across animals because the number of sampled genomes (i.e., 6) is far from large enough to make a general conclusion, although this restriction was inevitable to keep our analyses reliable.

We will have to analyze more animals when their finished genome assemblies come out. It will also be interesting to apply our cross-sectional analyses to the finished genomes of other eukaryotes (e.g., Dolinski and Botstein 2005; Haas et al. 2005), or even to the finished genomes of procaryotes (see, e.g., Walter et al. 2009). The general features we observed were largely attributed to the creation mechanisms of duplicated genes. But mechanisms of post-duplication evolution, such as preferential loss or homogenization of particular classes of duplogs (Rodin and Parkhomchuk 2004; Ezawa et al. 2006, 2010; Xu et al. 2008), might make considerable contributions especially to the species-specific features. Finally, although this study focused mainly on the positional relationships between duplogs, it may also be interesting to examine the dependence of functional differences between duplogs on their positional relationships. Such analyses are left for future studies.

Supplementary Material

Supplementary materials and methods, figures S1–S5, and tables S1–S4 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This paper is dedicated to the memory of late Dr. Walter M. Fitch, who passed away in March 2011. The authors are grateful to Dr Kenta Sumiyama for discussions on this study. This study was conducted as a part of the Genome Network Project as well as Grant-in-Aid for Scientific Research presided by the Ministry of Education, Culture, Sports, Science and Technology of Japan. This work was also supported in part by the US National Library of Medicine (grant LM010009-01 to Dan Graur and Giddy Landan at the University of Houston).

Literature Cited

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

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

Ashburner M, Bergman CM. 20005. *Drosophila melanogaster*: a case study of a model genomic sequence and its consequences. Genome Res. 15:1661–1667.

Babushok DV, Ostertag EM, Kazazian HH Jr. 2007. Current topics in genome evolution: molecular mechanisms of new gene formation. Cell Mol Life Sci. 64:542–554.

Bailey JA, Eichler EE. 2006. Primate segmental duplications: crucibles of evolution, diversity and disease. Nat Rev Genet. 7:552–564.

Bailey JA, Liu G, Eichler EE. 2003. An Alu transposition model for the origin and expansion of human segmental duplications. Am J Hum Genet. 73:823–834.

Bailey JA, et al. 2002. Recent segmental duplications in the human genome. Science 297:1003–1007.

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

Celniker SE, et al. 2002. Finishing a whole-genome shotgun: release 3 of the *Drosophila melanogaster* euchromatic genome sequence. Genome Biol. 3:1–14.

Cheung J, et al. 2003. Recent segmental and gene duplications in the mouse genome. Genome Biol. 4:R47.

Chuang JH, Li H. 2007. Similarity of synonymous substitution rates across mammalian genomes. J Mol Biol. 65:236–248.

Church DM, et al. 2009. Lineage-specific biology revealed by a finished genome assembly of the mouse. PLoS Biol. 5:e1000112.

Conceição IC, Aguadé M. 2008. High incidence of interchromosomal transpositions in the evolutionary history of a subset of *Or* genes in *Drosophila*. J Mol Evol. 66:325–332.

Cutter AD. 2008. Divergence times in *Caenorhabditis* and *Drosophila* inferred from direct estimation of the neutral mutation rate. Mol Biol Evol. 25:778–786.

Danio rerio Sequencing Project. 2007. Seventh Assembly, Zv7, of the zebrafish genome released. Hinxton (UK): Wellcome Trust Sanger Institute. [updated 2007 Nov 2; cited 2011 Aug 29]. Available from: http://www.sanger.ac.uk/Projects/D_rerio/Zv7_assembly_information.shtml

Danio rerio Sequencing Project. 2008. Zv8, the 8th integrated whole genome assembly of the zebrafish genome has been released. Hinxton (UK): Wellcome Trust Sanger Institute. [updated 2008 Dec 17; cited 2011 Aug 29]. Available from: http://www.sanger.ac.uk/Projects/D_rerio/Zv8_assembly_information.shtml

Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol. 3:e314.

Dolinski K, Botstein D. 2005. Changing perspectives in yeast research nearly a decade after the genome sequence. Genome Res. 15:1611–1619.

Ehrlich M, Wang RY. 1981. 5-Methylcytosine in eukaryotic DNA. Science 212:1350–1357.

Eichler E, Sankoff D. 2003. Structural dynamics of Eukaryotic chromosomal evolution. Science 301:793–797.

Eickbush TH, Eickbush DG. 2007. Finely orchestrated movements: evolution of the ribosomal RNA genes. Genetics 175:477–485.

Ezawa K, Ikeo K, Gojobori T, Saitou N. 2010. Evolutionary pattern of gene homogenization between primate-specific paralogs after human and macaque speciation using the 4-2-4 method. Mol Biol Evol. 27:2152–2171.

Ezawa K, OOta S, Saitou N. 2006. Proceedings of the SMBE tri-national young investigator's workshop 2005. Genome-wide search of gene conversions in duplicated genes of mouse and rat. Mol Biol Evol. 23:927–940.

Fitch WM. 1970. Distinguishing homologous from analogous proteins. Syst Zool. 19:99–113.

Friedman R, Hughes AL. 2003. The temporal distribution of gene duplication events in a set of highly conserved human gene families. Mol Biol Evol. 20:154–161.

Friedman R, Hughes AL. 2004. Two patterns of genome organization in mammals: the chromosomal distribution of duplicate genes in human and mouse. Mol Biol Evol. 21:1008–1013.

Haas BJ, et al. 2005. Complete reannotation of the *Arabidopsis* genome: methods, tools, protocols, and the final release. BMC Biol. 3:7.

Haldane JBS. 1932. The causes of evolution. London: Longmans and Green.

Harris TW, et al. 2010. WormBase: a comprehensive resource for nematode research. Nucleic Acids Res. 38:D463–D467. Available from: http://www.wormbase.org and FTP site: ftp.sanger.ac.uk/pub2/wormbase.

Hedger A, Ponting CP. 2007. Evolutionary rate analyses of orthologs and paralogs from 12 *Drosophila* genomes. Genome Res. 17:1837–1849.

Hillier LW, et al. 2005. Genomics in *C. elegans*: so many genes, such a little worm. Genome Res. 15:1651–1660.

Hu S, et al. 2008. Evolution of the *CYP2ABFGST* gene cluster in rat, and a fine-scale comparison among rodent and primate species. Genetica 133:215–226.

Hubbard TJP, et al. 2009. Ensembl 2009. Nucleic Acids Res. 37:D690–D697. Available from: http://www.ensembl.org and ftp.ensembl.org/pub.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. Nature 431:931–945.

Jaillon O, et al. 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature 431:946–957.

Kasahara M, et al. 2007. The medaka draft genome and insights into vertebrate genome evolution. Nature 447:714–719.

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

Loewe L, Cutter AD. 2008. On the potential for extinction by Muller's Ratchet in *Caenorhabditis elegans*. BMC Evol Biol. 8:125.

Lynch M. 2007. The origins of genome architecture. Sunderland (MA): Sinauer Associates.

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

Marques-Bonet T, et al. 2009. A burst of segmental duplications in the genome of the African great ape ancestor. Nature 457:877–881.

Muller HJ. 1935. The origin of chromosomal deficiencies as minute deletions subject to insertion elsewhere. Genetics 17:237–252.

Murphy WJ, et al. 2005. Dynamics of mammalian chromosome evolution inferred from multispecies chromosomal maps. Science 309:613–617.

Nei M. 1969. Gene duplication and nucleotide substitution in evolution. Nature 221:40–42.

Newman TL, et al. 2005. A genome-wide survey of structural variation between human and chimpanzee. Genome Res. 15:1344–1356.

Ohno S. 1970. Evolution by gene duplication. New York: Springer.
Pan D, Zhang L. 2007. Quantifying the major mechanisms of recent gene duplications in the human and mouse genomes: a novel strategy to estimate gene duplication rates. Genome Biol. 8:R158.
Pan D, Zhang L. 2008. Tandemly arrayed genes in vertebrate genomes. Comp Funct Genomics. 2008:545269.
Parmley JL, et al. 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. Mol Biol Evol. 23:301–309.
Pevzner P, Tesler G. 2003. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. Proc Natl Acad Sci U S A. 100:7672–7677.
Ranz JM, et al. 2007. Principles of genome evolution in the Drosophila melanogaster species group. PLoS Biol. 5:e152.
Richards S, et al. 2005. Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution. Genome Res. 15:1–18.
Rodin SN, Parkhomchuk DV. 2004. Position-associated GC asymmetry of gene duplicates. J Mol Evol. 59:372–384.
Rozen S, et al. 2003. Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. Nature 423:873–876.
Rubin GM, et al. 2000. Comparative genomics of the eukaryotes. Science 287:2204–2215.
She X, Cheng Z, Zöllner S, Church DM, Eichler EE. 2008. Mouse segmental duplication and copy number variation. Nat Genet. 40:909–914.
She X, et al. 2004. Shotgun sequence assembly and recent segmental duplications within the human genome. Nature 431:927–930.
Shoja V, Zhang L. 2006. A roadmap of tandemly arrayed genes in the genomes of human, mouse, and rat. Mol Biol Evol. 23:2134–2141.
Smith TF, Waterman MS. 1981. Identification of common molecular sequences. J Mol Biol. 147:195–197.
Sneath PHA, Sokal RR. 1973. Numerical taxonomy. San Francisco (CA): Freeman.
Sonnhammer ELL, Koonin EV. 2002. Orthology, paralogy, and proposed classification for paralog subtypes. Trends Genet. 18:619–620.
Stenico M, Lloyd AT, Sharp PM. 1994. Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. Nucleic Acids Res. 22:2437–2446.
Tweedle S, et al. 2009. FlyBase: enhancing *Drosophila* Gene Ontology annotations. Nucleic Acids Res. 37:D555–D559. Available from: http://flybase.org.
Walter MC, et al. 2009. PEDANT covers all complete RefSeq genomes. Nucleic Acids Res. 37:D408–D411.
Wapinski I, Pfeffer A, Friedman N, Regev A. 2007. Natural history and evolutionary principles of gene duplication in fungi. Nature 449:54–61.
Warnecke T, Hurst LD. 2007. Evidence for a trade-off between translational efficiency and splicing regulation in determining synonymous codon usage in *Drosophila melanogaster*. Mol Biol Evol. 24:2755–2762.
Wolfe K. 2000. Robustness—it's not where you think it is. Nat Genet. 25:3–4.
Woods IG, et al. 2005. The zebrafish gene map defines ancestral vertebrate chromosomes. Genome Res. 15:1307–1314.
Xu S, et al. 2008. Gene conversion in the rice genome. BMC Genomics. 9:93.
Yang S, et al. 2008. Repetitive element-mediated recombination as a mechanism for new gene origination in Drosophila. PLoS Genetics. 4:e3.
Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Compt Appl Biosci. 13:555–556.
Yi S, Ellsworth DL, Li W- H. 2002. Slow molecular clocks in Old World Monkeys, apes, and humans. Mol Biol Evol. 19:2191–2198.
Zhou Q, et al. 2008. On the origin of new genes in *Drosophila*. Genome Res. 18:1446–1455.
Zhou Y, Mishra B. 2005. Quantifying the mechanisms for segmental duplications in mammalian genomes by statistical analysis and modeling. Proc Natl Acad Sci U S A. 102:4051–4056.

Associate editor: Yoshihito Niimura