MBE Advance Access published April 24, 2012

Pervasive Indels and Their Evolutionary Dynamics after the Fish-Specific Genome Duplication

Baocheng Guo,[1,2] Ming Zou,[3] and Andreas Wagner*,[1,2]

[1]Institute of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland

[2]The Swiss Institute of Bioinformatics, Quartier Sorge-Batiment Genopode, Lausanne, Switzerland

[3]Key Laboratory of Aquatic Biodiversity and Conservation, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan, People's Republic of China

***Corresponding author:** E-mail: andreas.wagner@ieu.uzh.ch.

Associate editor: Hervé Philippe

Abstract

Insertions and deletions (indels) in protein-coding genes are important sources of genetic variation. Their role in creating new proteins may be especially important after gene duplication. However, little is known about how indels affect the divergence of duplicate genes. We here study thousands of duplicate genes in five fish (teleost) species with completely sequenced genomes. The ancestor of these species has been subject to a fish-specific genome duplication (FSGD) event that occurred approximately 350 Ma. We find that duplicate genes contain at least 25% more indels than single-copy genes. These indels accumulated preferentially in the first 40 my after the FSGD. A lack of widespread asymmetric indel accumulation indicates that both members of a duplicate gene pair typically experience relaxed selection. Strikingly, we observe a 30–80% excess of deletions over insertions that is consistent for indels of various lengths and across the five genomes. We also find that indels preferentially accumulate inside loop regions of protein secondary structure and in regions where amino acids are exposed to solvent. We show that duplicate genes with high indel density also show high DNA sequence divergence. Indel density, but not amino acid divergence, can explain a large proportion of the tertiary structure divergence between proteins encoded by duplicate genes. Our observations are consistent across all five fish species. Taken together, they suggest a general pattern of duplicate gene evolution in which indels are important driving forces of evolutionary change.

Key words: indel, gene duplication, teleost, fish-specific genome duplication.

Introduction

Genetic changes that facilitate the evolution of new traits are key to adaptive evolution. Gene duplication is one such change because it provides redundant genetic materials and thus facilitates genetic innovation (Ohno 1970). Duplicate genes and their evolutionary fates have therefore attracted interest by numerous researchers, who have studied the evolution of duplicate genes across a broad range of organisms, including fungi (Kellis et al. 2004; Gu et al. 2005; VanderSluis et al. 2010), plants (Ha et al. 2009; Zhang, Huang, et al. 2010), fish (Brunet et al. 2006; Steinke et al. 2006; Semon and Wolfe 2007; Guo et al. 2009; Kassahn et al. 2009), and mammals (Robinson-Rechavi and Laudet 2001; Zhang et al. 2003; Farre and Alba 2010). Several scenarios have been proposed for the evolutionary fates of duplicate genes, including nonfunctionalization, where one of two copies becomes inactivated through mutation (Ohno 1970), subfunctionalization, where both copies assume more specialized functions than their ancestor (Hughes 1994; Force et al. 1999; Lynch and Force 2000), neofunctionalization, where one copy evolves a new function (Ohno 1970), and subneofunctionalization, a combination of the last two scenarios (He and Zhang 2005). Some analyses of duplicate genes focus on sequence divergence (Conant and Wagner 2003; Brunet et al. 2006; Steinke et al. 2006), whereas others focus on divergence in some aspect of gene function, such as gene expression (Gu et al. 2005), protein interaction (Wagner 2002; VanderSluis et al. 2010), localization (Kassahn et al. 2009), or protein structure (Kassahn et al. 2009). Even though gene duplication is important to understand genome evolution, genome-wide analyses of duplicate gene evolution are still limited, especially for vertebrate genomes that experienced one or more whole-genome duplication events.

Evolutionary divergence of genomes and genes—including duplicate genes—can be driven by nucleotide substitutions, insertions, and deletions. These events may be interdependent. For example, insertions and deletions—we refer to them as indels—are mutagenic and can increase the substitution rate in genomic regions flanking these indels (Tian et al. 2008). Conversely, genes with higher tolerance for nucleotide changes may also tolerate more indels. Although nucleotide substitution patterns in duplicate genes are well studied (Kellis et al. 2004; Brunet et al. 2006; Steinke et al. 2006), patterns of indel accumulation are poorly characterized, even though indels can lead to structural and functional divergence of homologous proteins (Reeves et al. 2006; Chan et al. 2007; Jiang and Blouin 2007; Chen et al. 2009) and thus play an important role in protein evolution (Grishin 2001; Hormozdiari et al. 2009). For example, Zhang, Wang, et al. (2010) showed that indels, as well as substitutions, are necessary to explain protein structure changes in homologous protein families. Zhang et al. (2011) demonstrated that indels can lead to a series of structural changes by analyzing protein structures in the SCOP structural classification database. Salari et al. (2008) found that indels could cause more severe functional changes than other types of sequence variation between paralogous protein pairs in *Escherichia coli*. Some indels can also induce functional divergence of proteins by changing the translational reading frame and thus causing frameshift mutations (Raes and Van de Peer 2005). Taken together, observations like these indicate that indels may be at least as important as nucleotide substitutions in the evolution of duplicate genes.

Teleost genomes are ideal to examine genome-wide duplicate gene evolution in vertebrates after whole-genome duplication. Extensive comparative genomics studies have demonstrated that an ancestor of teleosts experienced a whole-genome duplication about 350 Ma, the so-called fish-specific genome duplication (FSGD) (Amores et al. 1998; Taylor et al. 2003; Meyer and Van de Peer 2005; Guo et al. 2010). This event created thousands of duplicate genes in teleost genomes, many of which are retained to this day (Robinson-Rechavi and Laudet 2001; Taylor et al. 2003; Brunet et al. 2006; Steinke et al. 2006; Kassahn et al. 2009). We here studied the accumulation of indels and their effects on duplicate gene divergence in five available teleost genomes, those of the zebra fish *Danio rerio*, the stickleback *Gasterosteus aculeatus*, the medaka *Oryzias latipes*, the pufferfish *Takifugu rubripes*, and the pufferfish *Tetraodon nigroviridis*.

Materials and Methods

Candidate Gene Identification

We obtained 23,155 gene families from the database HOMOLENS version 4 (ftp://pbil.univ-lyon1.fr/databases/homolens4.php; Penel et al. 2009), which uses data from the Ensembl database release 49 (Hubbard et al. 2002). HOMOLENS allows us to retrieve reliably sets of orthologous genes for our study (Brunet et al. 2006; Studer et al. 2008). HOMOLENS has the same architecture as the related database HOVERGEN (Duret et al. 1994), and it stores genes from completely sequenced animal genomes. HOMOLENS organizes genes into families and includes precalculated alignments and phylogenic trees. In HOMOLENS 4, alignments are computed using MUSCLE (Edgar 2004) with default parameters; phylogenetic trees are constructed in PHYML (Guindon and Gascuel 2003) with conserved blocks of alignments selected with Gblocks (Castresana 2000) and the Jones, Taylor, and Thorton (JTT) amino acid substitution model (Jones et al. 1992). All individual phylogenetic trees are reconciled with a species tree using the program RAP (Dufayard et al. 2005), which can help detect ancient gene duplications and select orthologous genes (Penel et al. 2009).

For our analysis, we used genes from five completely sequenced teleost genomes, that is, the genomes of *D. rerio*, *G. aculeatus*, *O. latipes* (Kasahara et al. 2007), *Ta. rubripes* (Aparicio et al. 2002), and *Te. nigroviridis* (Jaillon et al. 2004), as well as genes from the genome of *Homo sapiens* (Venter et al. 2001) as an outgroup because the human gene complement is especially well annotated compared with other vertebrate genomes. We used the FamFetch client for HOMOLENS and its TreePattern functionality (Dufayard et al. 2005) to select genes that have only a single copy in *H. sapiens* and that fall into two major categories (fig. 1). The first category comprises single-copy genes with a phylogeny like that displayed in figure 1*A*. A gene with this phylogeny would have lost a duplicated copy after the FSGD event and would have undergone no further duplication in any of the five teleost genomes we study. The second category comprises genes whose duplicates have been retained in some species after the FSGD duplicate genes. If all five species had retained the gene and its duplicate, and if no further duplication of the gene occurred after the FSGD, the phylogeny of figure 1*B* would emerge. In our analysis, we focused on genes where at least two teleost species contained two paralogs of the gene, whereas loss or other duplications were allowed to occur in each teleost species. We did not require exactly two duplicates in all five species because doing so would have rendered our data set too small for meaningful statistical analysis. We required that duplicates must exist in at least two teleost species to increase the likelihood that the duplicates we study were truly part of the FSGD event.

Indel Characterization, Extraction, and Statistics in Singletons and Duplicate Genes

For the gene families that met the above criteria, we downloaded both their nucleic acid and amino acid sequences. We retained only the most closely related paralog pair for each species in each clade (fig. 1*B*) for further analysis and removed in each species more distant copies from gene families where further duplication events had occurred. See supplementary table S1 (Supplementary Material online) for sequences, we selected and their current Ensembl identifiers. We then aligned the amino acid sequences of each gene family from all five species and from the human genome using MUSCLE with default parameters, which is thought to give the best accurate alignments on average (Edgar 2004). We then inspected the alignments manually for possible artifacts. Subsequently, we calculated DNA alignments from protein alignments with a custom Perl script and analyzed these new alignments as described next.

We identified likely indels as gaps in our sequence alignments. Gaps starting at the N-terminus or ending at the C-terminus of the alignment were coded as one indel. Figure 2 illustrates how we characterized indels in singletons (fig. 2*A*) as well as in duplicates (fig. 2*B*). We note that indels cannot be uniquely polarized—identified as an insertion or a deletion—for singletons (fig. 2*A*) and for indels that occurred before the FSGD (pre-FSGD: Ins a&b and Del a&b; fig. 2*B*) in duplicates. We excluded all pre-FSGD indels (fig. 2*B*) from our analyses of association between indels

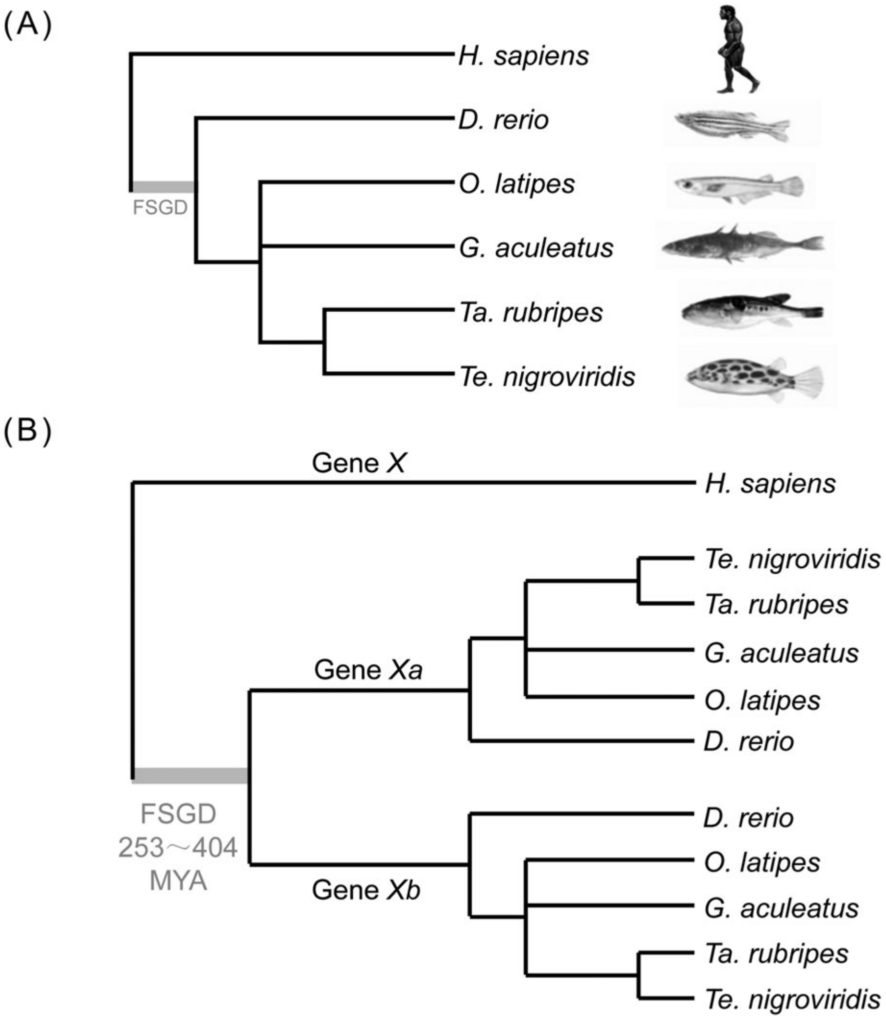

Fig. 1. Topologies used for gene family identification. (*A*) Singleton gene tree topology. Genes in this category have only a single copy in each teleost genome and in the human genome. They correspond to genes whose duplicate resulting from the FSGD was lost, and where no further duplication occurred after the FSGD. (*B*) Hypothetical "perfect" gene tree topology of a gene that was duplicated in the FSGD. Gene *X* stands for a hypothetical gene in our analysis and genes *Xa* and *Xb* stand for duplicates of this gene. The phylogenetic tree shown would be observed for a gene that has been duplicated in the FSGD, and where both duplicates have been preserved in each of the five study species without further duplication. The phylogenetic tree depicting the evolutionary relationships among the five fish species is modified from Miya et al. (2005), Nelson (2006), Li et al. (2008), and Negrisolo et al. (2010), and the time of the FSGD event is based on Taylor et al. (2003) and Meyer and Van de Peer (2005).

and duplicate gene evolution. That is, we only considered indels that originated after the FSGD (post-FSGD indels: Ins a and Del a or Ins b and Del b; fig. 2*B*) for these analyses because such indels are specific to one of two paralogs and might thus contribute to the divergence of duplicate genes.

We counted all indels in singleton and duplicate genes, as defined in figure 2, with a custom Perl script. We then performed an exact binomial test to ask if the occurrence of post-FSGD insertions and deletions was equal between duplicate gene pairs and corrected the resulting *P* value for multiple testing with the Benjamini–Hochberg false discovery rate procedure (Benjamini and Hochberg 1995), as implemented in the R package ShotgunFunctionalizeR (Kristiansson et al. 2009).

To exclude the possibility that our results varied with alignment procedures, we also aligned the amino acid sequences of each gene family using MUSCLE with nondefault parameters (gapopen: −10, as opposed to the default gapopen = −2.9, distance1: Kmer6_6; and distance2: PctIdKimura), ClustalW (Thompson et al. 1994), MAFFT (Katoh and Toh 2008), and PRANK with default parameters (Loytynoja and Goldman 2005). We observed that different alignment procedures yielded only minor differences in indel incidence and length distribution that would not affect the major observations we report here (supplementary fig. S1, Supplementary Material online).

Gene prediction errors are common in current public databases, and they may affect our results (Nagy et al. 2008; Harrow et al. 2009; Hubisz et al. 2011; Nagy and Patthy 2011; Thompson et al. 2011; Prosdocimi et al. 2012). We thus carried out several additional analyses to ask whether our results are robust to such errors. First, most gene prediction errors (ca. 60%, according to an analysis by Prosdocimi et al [2012]) that affect indels occur in the regions of a gene that

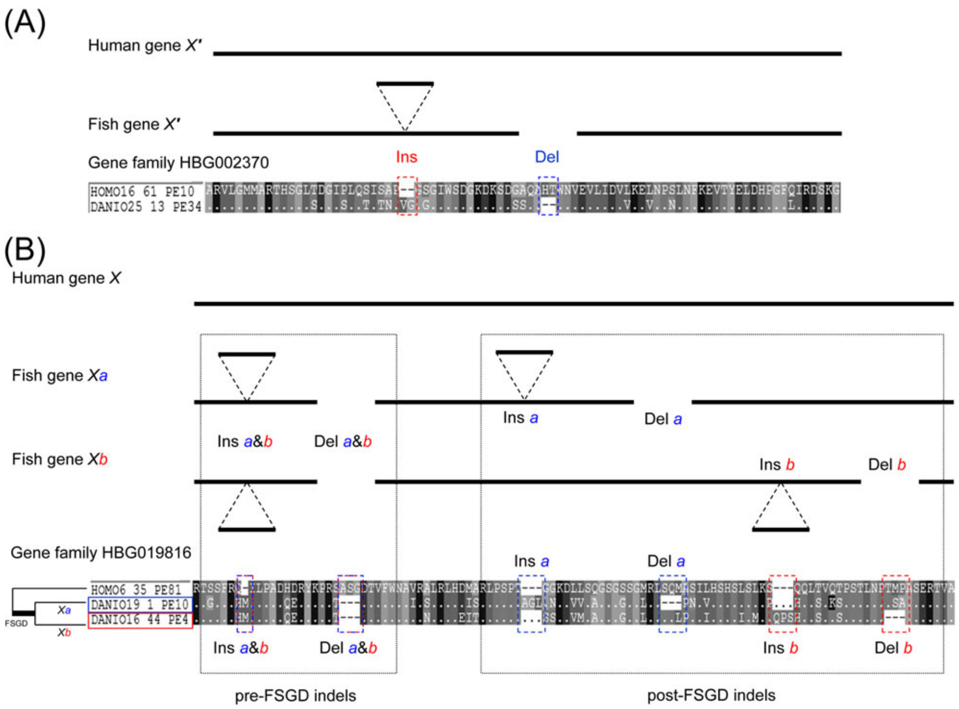

FIG. 2. Characterizing indels in singletons and duplicate genes. (*A*) Characterizing indels in singletons by comparing fish and human orthologs. Gene *X*′ stands for a hypothetical gene in our analysis. If the fish ortholog contains an additional stretch of DNA relative to humans, we refer to this stretch as an insertion (Ins); if it lacks a stretch of DNA, we refer to this absence as a deletion (Del). We use these designations for brevity, and note that an insertion (deletion) in a fish ortholog is not distinguishable from a deletion (insertion) in a human ortholog, based on comparing merely two genes. The alignment shows, as an example, an insertion and a deletion in the gene DANIO25_13_PE34 of *D. rerio* compared with its human ortholog HOMO16_61_PE10_form gene family HBG002370. (*B*) Characterizing indels in duplicate genes. Gene *X* stands for a hypothetical gene in our analysis. Genes *Xa* and *Xb* stand for duplicates of this gene. The left dashed rectangle indicates indels that occurred before the FSGD (pre-FSGD) and that were thus shared by both duplicated copies (Ins a&b and Del a&b). We follow the same convention as in panel (*A*) to label such indels as insertions or deletions, but note that they cannot be unambiguously identified as insertions or deletions. In contrast, indels that occurred after the FSGD (right dashed rectangle, post-FSGD indels) can be identified as either insertions (Ins a and Ins b) or deletions (Del a and Del b). Indels in a duplicate gene pair (DANIO19_1_PE10 and DANIO16_44_PE4) of *D. rerio* compared with its human ortholog HOMO6_35_PE81 from a partial alignment of the gene family HBG019816 are shown as an example. All gene identifiers (gene family ID and sequence name) used in this figure and figure 5A are based on database HOMOLENS version 4 (Penel et al. 2009) and can be used for retrieving information about these genes at http://pbil.univ-lyon1.fr/databases/homolens4.php.

encode the N-terminal or C-terminal part of a protein (Hubisz et al. 2011; Thompson et al. 2011; Prosdocimi et al. 2012). We thus repeated our analyses of indel accumulation and of indel density association with sequence divergence, with excluding such terminal indels. Second, we repeated these analyses for genes that contain no or only one intron because such genes are less likely to be subject to an important class of gene prediction errors, namely those affecting gene structure and alternative splicing variants. Finally, we repeated the analyses with recent Ensembl data (version 64, http://www.ensembl.org/index.html) to confirm that our findings are stable with respect to gene annotation updating. Supplementary figure S6 (Supplementary Material online) summarizes results from these analyses. Because they did not yield statistical patterns different from those reported in the main text, we do not discuss them in detail.

Evolutionary Distance Calculation and Phylogenetic Construction

We computed the number of nonsynonymous substitutions per nonsynonymous site (K_a) and the number of synonymous substitutions per synonymous site (K_s) between duplicate genes in each teleost using a maximum likelihood method (Yang and Nielsen 2000) with the YN00 program implemented in PAML version 4.4 (Yang 2007). We calculated the nucleic acid evolutionary distance between duplicate genes in each teleost genome using the Kimura 2-Parameter nucleotide substitution model implemented in PHYLIP-3.6b (Felsenstein 2004). We reconstructed the phylogenetic tree of each gene family using both the JTT (Jones et al. 1992) and the LG (Le and Gascuel 2008) amino acid substitution model in RAxML-7.0.4 (Stamatakis 2006) to confirm that their topologies are consistent with that in figure 1 when using different substitution models.

Indel Effects on Solvent Accessibility and Protein Structure

To analyze the effects of indels on protein structure, we took advantage of experimentally determined protein structures deposited in the Protein Data Bank (PDB) (Berman et al. 2000). Specifically, we first used BlastP (Altschul et al. 1997) with an *E* value threshold of 1.0×10^{-50} to identify proteins in PDB that matched duplicate genes in our data set. We only kept those gene families for further analysis where all family members matched a PDB protein with at least 50% amino acid identity and over at least 50 contiguous amino acids. We used the resulting data set to analyze how indel location depends on solvent accessibility of amino acids and on protein secondary structure and to study the effects of indels on tertiary structure, as described next.

We used ACCpro 4.01 with the most informative solvent accessibility threshold of 25% (Pollastri et al. 2002) to predict the relative solvent accessibility of each duplicate gene. The prediction of ACCpro uses multiple alignments of target proteins generated by PSI-BLAST as an input (Altschul et al. 1997). Each residue in a protein is predicted as buried (b) or exposed (e) by ACCpro. We predicted the secondary structures of duplicate genes using PSIPRED (Jones 1999), which uses a two-stage neural network for predicting protein secondary structure based on analyzing position-specific scoring matrices generated by PSI-BLAST (Altschul et al. 1997). PSIPRED is currently one of the most accurate protein secondary structure prediction methods available (Rost 2003). It assigns each amino acid in a protein to three states, helix (H), strand (S), and coil (C). PSIPRED requires a sequence database for analysis. We used the UniRef90 (Suzek et al. 2007) protein database for this purpose, where we masked low-complexity regions, transmembrane regions, and coiled-coil segments with the program PFILT implemented in PSIPRED before prediction.

After we had obtained secondary structure predictions, we used a custom Perl script to map the indels that we had identified in our duplicate genes onto the predicted secondary structure elements of their encoded proteins and onto the solvent accessibility of the amino acids. We considered that an indel in one duplicated copy occurred in a specific secondary structure type if both of the amino acids flanking the indel had this type of secondary structure in the other duplicate (which lacked the indel).

We used the "automated mode" of the SWISS-MODEL workspace (Arnold et al. 2006) for homology modeling of protein tertiary structures in our data set. This mode automatically selects suitable templates based on an input amino acid sequence, as well as information on the quality of the template structure, bound substrate molecules, and different conformational states of the template. The automated mode is suited for proteins where the target–template similarity is sufficiently high (usually exceeding 50%) to allow for fully automated modeling. Because of the high sequence divergence between our duplicate genes and known structure templates in PDB, modeling full structures with all loops and side chains would have been inaccurate and too time consuming, and we thus focused our analysis on the backbone structures of duplicate genes. We note that the accuracy of structural modeling could be influenced by low structural resolution of templates. To reduce the influence of this confounding factor, we used only duplicate genes with high-resolution structure templates (solved by X-ray crystallography to better than 2.2 Å resolution) and excluded all structures determined by nuclear magnetic resonance as well as structures that were derived from mutant proteins.

To compare protein structures between duplicate gene pairs, we employed the structural alignment program Secondary Structure Matching (SSM, Krissinel and Henrick 2004) through an online service (http://www.ebi.ac.uk/msd-srv/ssm/). We used this program only for protein pairs where at least 70% of amino acids in each protein were alignable in the structure. SSM computes a P-score, which reflects the statistical significance of an alignment, and is based on the root mean square deviation between the aligned regions, the length of the aligned region, and the number of indels in the aligned regions. Nontrivial alignments that reflect genuine structural similarity generally have a P-score of greater than 3 (Krissinel and Henrick 2004), and we analyzed only such alignments.

Indel-Induced Frameshift Mutation Identification in Duplicate Genes

The procedures we described thus far are based on amino acid sequences and are thus suitable for in-frame indels that affect multiples of 3 nt. They cannot be used for studying duplicate genes with indel-induced frameshift mutation (Raes and Van de Peer 2005). To study frameshifting indels, we first compared the nucleotide sequences and amino acid sequences among all members of a gene family. For this purpose, we used GeneWise (Birney et al. 2004) with default parameter settings. We then isolated those regions within the resulting multiple sequence alignments that included the frameshift, together with 25 flanking amino acids (75 flanking nucleotides). To avoid misalignments, we analyzed only those putative frameshifted regions further, where the DNA sequence identity in the frameshifted region was no lower than 70%, and where the amino acid sequence identity was no higher than 20%. We manually inspected alignments of the remaining candidates to avoid obvious further alignment artifacts.

Results

The Data Set

To investigate the impact of indels on the evolution of duplicate genes, we analyzed 4,571 gene families containing at least one gene for all six of our species (five teleost fish species and a human outgroup) in the database HOMOLENS 4. We employed a method based on gene tree topology to identify gene families for analysis in this study (see Materials and Methods). We identified 609 genes that had only a single copy in all six species (singletons) and 1,500 gene families likely to have been duplicated in the FSGD (FSGD families, see Materials and Methods). In this data set, the FSGD families comprise 731 duplicate gene pairs in *D. rerio*, 681 pairs in *G. aculeatus*, 1,162 pairs in *O. latipes*, 1,340 pairs in *Ta. rubripes*, and 1,047 pairs in *Te. nigroviridis* (table 1). A detailed gene family list is provided in supplementary table S2 (Supplementary Material online) for the singleton gene families and in supplementary table S3 (Supplementary Material online) for the FSGD families.

More Indels Accumulate in Duplicate Genes Than in Singletons

We first identified indels that occurred in both singletons and duplicate genes since teleosts diverged from their common ancestor with humans (fig. 2A, see Materials and Methods). We found that indels are pervasive in both singleton and duplicate genes. For example, 96.9% (590 of 609) singletons and 97.7% (1,429 of 1,462) duplicate genes in *D. rerio* showed indels. The corresponding numbers are similarly high in other species and exceed 95% for both singletons and duplicates in each species (table 1).

We next asked whether duplicate genes contain different numbers of indels than singletons. The answer was yes in each teleost species (Wilcoxon rank-sum tests, $P < 1.0 \times 10^{-6}$ for each species). In order to exclude alignment artifacts as confounding factors for genes with many indels, we excluded genes with more than 30 indels in both the singletons and duplicate genes from the analyses we describe next. After removal of these genes, the average indel number per singleton gene ranged from 6.8 (*Ta. rubripes*) to 8.0 (*Te. nigroviridis*). For duplicate genes, it ranged from 9.0 (*D. rerio* and *G. aculeatus*) to 10.8 (*Te. nigroviridis*). The difference in indel number between singleton and duplicate genes was still statistically significant for this reduced data set in each teleost (Wilcoxon rank-sum tests, $P < 1.2 \times 10^{-5}$ for each species). We next extended our analysis from the number of indels to their density, that is, the number of indels per kilobase (kbp). The average density of indels ranged from 3.0 (*Ta. rubripes*) to 3.5 (*Te. nigroviridis*) per kbp in singletons and from 4.0 (*D. rerio*) to 4.5 (*Te. nigroviridis*) per kbp in duplicate genes. Again, we found that duplicates have significantly higher indel density than singletons in each of the five teleost species (Wilcoxon rank-sum tests, $P < 1.0 \times 10^{-3}$ for each species). Based on the estimated divergence times of teleosts and humans of approximately 454 Ma (Peng et al. 2009), we can crudely estimate the rate of indel accumulation as approximately 6.5–7.6 $\times 10^{-4}$/kbp/my for singletons, and as 0.9–1.0 $\times 10^{-3}$/kbp/my for duplicate genes in the five teleosts. This means that indels accumulated in duplicate genes at least 25% faster than in singletons (*D. rerio*) since teleosts diverged from their common ancestor with humans.

Previous studies have suggested that gene duplicates often diverge asymmetrically in both sequence and function (Wagner 2002; Conant and Wagner 2003). This observation raises the possibility that among two duplicates, only one may accumulate indels faster than singletons do. To find out whether this is the case, we classified the member genes of a duplicate gene pair into two categories, the gene that contains more indels in a pair ("High" in fig. 3), and the gene that contains fewer indels ("Low" in fig. 3). We then compared the numbers of indels in these groups of genes to those of singletons and found that both categories of duplicates have more indels than singletons in each teleost (Wilcoxon rank-sum tests, $P < 1.0 \times 10^{-5}$). Figure 3 shows the results of this analysis separately for each species. The average indel densities for both categories of duplicates are significantly higher than those for singletons in each teleost

Table 1. Indels in Duplicates and Singletons.

	Duplicates							Singletons		
Species	Total Number of Duplicate Gene Pairs	Number of Duplicate Genes with Indels[a]	Number of Duplicated Pairs with Post-FSGD Indels	Number of Pre-FSGD Indels	Number of Post-FSGD Indels	Number of Post-FSGD Insertions	Number of Post-FSGD Deletions	Total Number of Singletons	Number of Singleton Genes with Indels[a]	Total Number of Indels
Danio rerio	731	1,429 (97.7%)	694 (94.9%)	4,520	7,412	2,621	4,791	609	590 (96.9%)	4,205
Gasterosteus aculeatus	681	1,326 (97.4%)	655 (96.2%)	3,422	7,502	3,061	4,441	609	587 (96.4%)	4,171
Oryzias latipes	1,162	2,276 (97.9%)	1,134 (97.6%)	6,312	13,764	5,811	7,953	609	591 (97.0%)	4,158
Takifugu rubripes	1,340	2,647 (98.8%)	1,312 (97.9%)	7,330	15,860	6,809	9,051	609	587 (96.4%)	4,069
Tetraodon nigroviridis	1,047	2,075 (99.1%)	1,035 (98.9%)	4,632	14,755	6,139	8,616	609	601 (96.9%)	4,794
Total	4,961	9,753 (98.3%)	4,830 (97.4%)	26,216	59,293	24,441	34,852	3,045	2,956 (97.1%)	21,397

[a] Indels that occurred between fish genes and their human orthologs since teleosts diverged from their common ancestor with human.

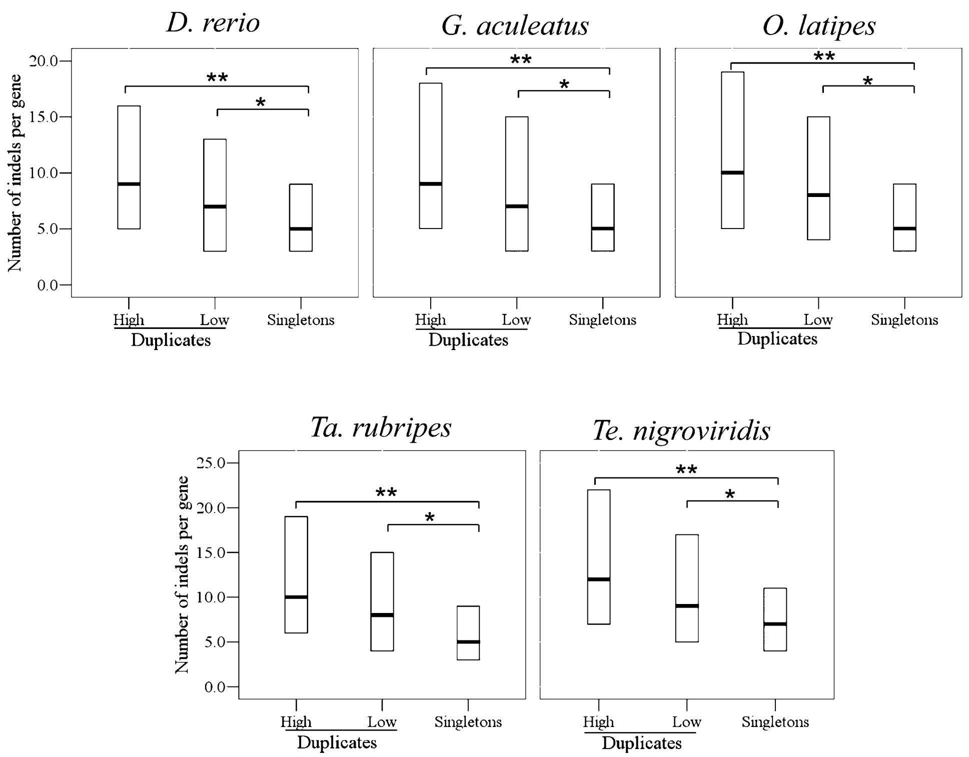

Fig. 3. Indels accumulate faster in duplicate genes. For the analysis shown here, we classified the member genes of a duplicate gene pair into two categories, the gene that contains more indels (High) and the gene that contains fewer indels in the pair (Low). Numbers of indels are shown for genes in each of these categories, as well as for singletons. Each panel shows data from one fish species, as indicated. Each vertically oriented rectangle shows the median (bar) and the interquartile range (box) of the number of indels per gene. Significant differences are indicated by asterisks, based on a Wilcoxon rank-sum test, *$P < 1.0 \times 10^{-5}$; **$P < 1.0 \times 10^{-22}$.

(Wilcoxon rank-sum tests, $P < 1.0 \times 10^{-3}$). In sum, the higher rate at which indels accumulate in duplicates is not just caused by faster evolution of one of the two duplicates. Instead, both duplicates typically accumulate more indels than singletons.

A Significant Excess of Deletions over Insertions

Our next analyses focused on post-FSGD indels, where our data allow us to distinguish between insertions and deletions (fig. 2*B*) as well as to study their profiles and effects on duplicate gene evolution. Post-FSGD indels are pervasive and occur in between 94.0% to more than 98.0% of duplicate genes, depending on the species (table 1). They account for approximately 62.1% (*D. rerio*) to 76.9% (*Te. nigroviridis*) of all indels that have been preserved in teleosts (the remainder being pre-FSGD indels). The average number of post-FSGD indels per duplicate gene ranges from 5.5 (*D. rerio*) to 8.0 (*Te. nigroviridis*), and their average density ranges from 2.5 (*D. rerio*) to 3.4 (*Te. nigroviridis*) per kbp. Given that the FSGD occurred approximately 350 Ma (Meyer and Van de Peer 2005), the average indel accumulation rate in duplicate genes since the FSGD event is approximately 7.0–9.8 × 10^{-4}/kbp/my.

Strikingly, duplicate genes usually contain more deletions than insertions (table 1). For example, duplicate genes in *D. rerio* contain 2,621 insertions and 4,791 deletions overall. The insertion to deletion ratio ranges from 1:1.3 (*Ta. rubripes*) to 1:1.8 (*D. rerio*). Each duplicate gene contains on average 1.9 insertions and 3.7 deletions in *D. rerio*. The corresponding numbers are 2.4 insertions and 3.8 deletions in *G. aculeatus*, 2.8 and. 3.9 in *O. latipes*, 2.8 and 3.9 in *Ta. rubripes*, as well as 3.3 and 5.0 in *Te. nigroviridis*, respectively. These differences between numbers of insertions and deletions are statistically significant for each teleost (Wilcoxon signed-ranks tests, $P < 1.0 \times 10^{-5}$ for all species).

Figure 4 and supplementary figure S2 (Supplementary Material online) display the length distribution of insertions and deletions. Shorter indels are much more frequent than longer indels, with the relative majority of indels affecting few codons, and one-codon indels being the most frequent. For example, 19.2% (in *Te. nigroviridis*) to 27.1% (in *Ta. rubripes*) of insertions and 22.1% (in *Te. nigroviridis*) to 32.0% (in *D. rerio*) of deletions affect only a single codon. Importantly, in most or all length categories, deletions are in excess over insertions, and the number of insertions decreases more quickly with increasing length than the number of deletions does. For example, insertions longer than 14 codons account for 7.5% of insertions, whereas deletions longer than 14 codons account for 24.3% of all deletions in *D. rerio*. The only exception to this pattern is in *Te. nigroviridis*, in which insertions 22–31 codons in length are in excess over deletions of the same length. This exception aside, the general bias toward deletions is further underscored by the observation that very long deletions (exceeding 50 codons) are significantly more abundant

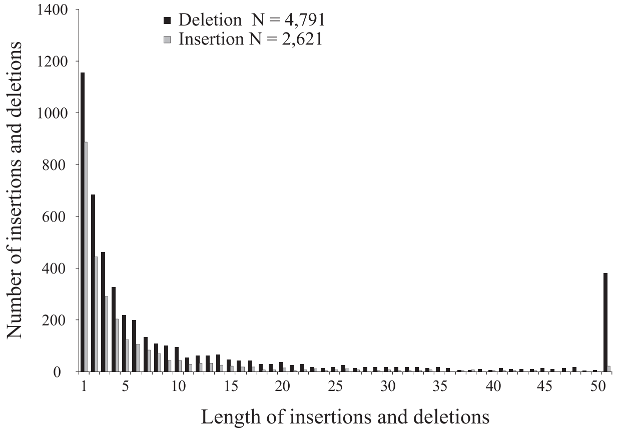

FIG. 4. Length distributions of post-FSGD insertions and deletions in duplicate genes in *Danio rerio*. The lengths of insertion and deletion are shown in numbers of codons. Total numbers of insertions and deletions are shown in the legend. Note the small but consistent excess of insertions over deletions in many length categories.

than very long insertions in all five species (*t*-test, $P < 0.01$) (fig. 4 and supplementary fig. S2, Supplementary Material online). Almost 50% of very long insertions and up to 80% of very long deletions occurred near the N-terminus or the C-terminus of the encoded protein (data not shown).

Only a Minority of Duplicates Accumulates Indels Asymmetrically

The ability to polarize indels for post-FSGD duplicates allows us to revisit our analysis of asymmetric divergence for insertions and deletions separately. Many duplicate genes contain unequal numbers of insertions, deletions, or both in all five fish species (supplementary table S4, Supplementary Material online). For example, 81.4% of duplicate gene pairs in *D. rerio* showed different numbers of indels, with similarly high percentages in the other four species (supplementary table S4, Supplementary Material online, column 9). We asked whether the accumulation of insertions, deletions, or indels was significantly different between the members of a duplicate gene pair, indicating asymmetric indel accumulation in the two duplicated copies. To this end, we first performed multiple binomial tests of the null hypothesis that indel numbers are equal between two members of a gene pair and used the Benjamini–Hochberg procedure to correct for multiple testing at a false discovery rate of $\alpha < 0.05$ (Benjamini and Hochberg 1995). This analysis showed that only a minority of duplicates accumulates indels asymmetrically (supplementary table S4, Supplementary Material online). Specifically, between 4.1% and 5.1% of duplicates accumulate insertions asymmetrically (depending on the species), and between 4.8% and 8.3% of genes accumulate deletions asymmetrically.

Fewer Indels Arise long after the FSGD

We next asked whether indels occurred preferentially soon or long after the FSGD. To this end, we reconstructed the phylogeny of each gene family from its aligned amino acid sequences and mapped indels onto this phylogeny, which allowed us to subdivide indels into three age categories (fig. 5A). The first category comprises old indels, which occurred after the FSGD event but before teleost diversification. An indel in this category would be shared by all five teleosts (fig. 5A, "old"). The second category comprises intermediate indels that occurred during the diversification of teleosts. These indels would be shared by at least two Euteleostei species (*G. aculeatus*, *O. latipes*, *Ta. rubripes*, and *Te. nigroviridis*) (fig. 5A, "intermediate"). The third category comprises young indels, which occur in only one teleost species late and are thus specific to only one copy of a duplicate gene (fig. 5A, "young"). We note that each indel in any one of the three categories would occur in only one of two paralogs within a genome because we are only concerned with post-FSGD indels.

Figure 5B shows the results of our analysis, which we restricted to genes with no more than 30 indels to avoid alignment artifacts (694 gene families). All five species show more old and intermediate indels than young indels. Specifically, we observed that 23.6% (*Te. nigroviridis*) to 38.7% (*D. rerio*) of indels are old indels, 38.0% (*D. rerio*) to 51.8% (*G. aculeatus*) are intermediate indels, and 13.5% (*G. aculeatus*) to 31.4% (*Te. nigroviridis*) are young indels. We next carried out chi-square tests to ask whether indels are equally likely to have occurred in each of the three stages (fig. 5A), taking into consideration the different approximate length of each stages (Peng et al. 2009). The test rejects this null hypothesis at $P < 1.2 \times 10^{-8}$ for all five species. Taken together, these observations show that indels accumulated preferentially soon after the FSGD but that indels continue to accumulate, albeit at a lower rate, long thereafter. The data in figure 5B also underscore a pattern that we showed in the previous section, namely an excess of deletions over insertions in multiple age categories. We asked whether this excess was significant using a χ^2 test of the null hypothesis that deletions are as abundant as insertions and did so for each indel age category in each species. We found that deletions are in excess over insertions for old and intermediate indels ($P < 2.2 \times 10^{-16}$) but not for young indels. These observations show that the excess of deletions was particularly pronounced soon after the FSGD and becomes less pronounced later on.

Indels Occur Preferentially in Loop Regions and near Solvent-Exposed Amino Acids

Indels can potentially disrupt protein structure, and this disruptive effect might leave traces in the evolutionary record. In other words, the extent to which proteins can tolerate indels might depend on their structure. To find out whether this is the case in our data set, we first focused on the predicted secondary structures of the protein products of duplicated genes. In a protein's secondary structure, each amino acid can be part of a helix (H), a β-sheet (S), or a loop or coil region (C). We mapped indels onto these secondary structure elements for duplicated gene families where such an analysis was possible (see Materials and Methods, pertinent gene families are provided in supplementary table S5, Supplementary Material online).

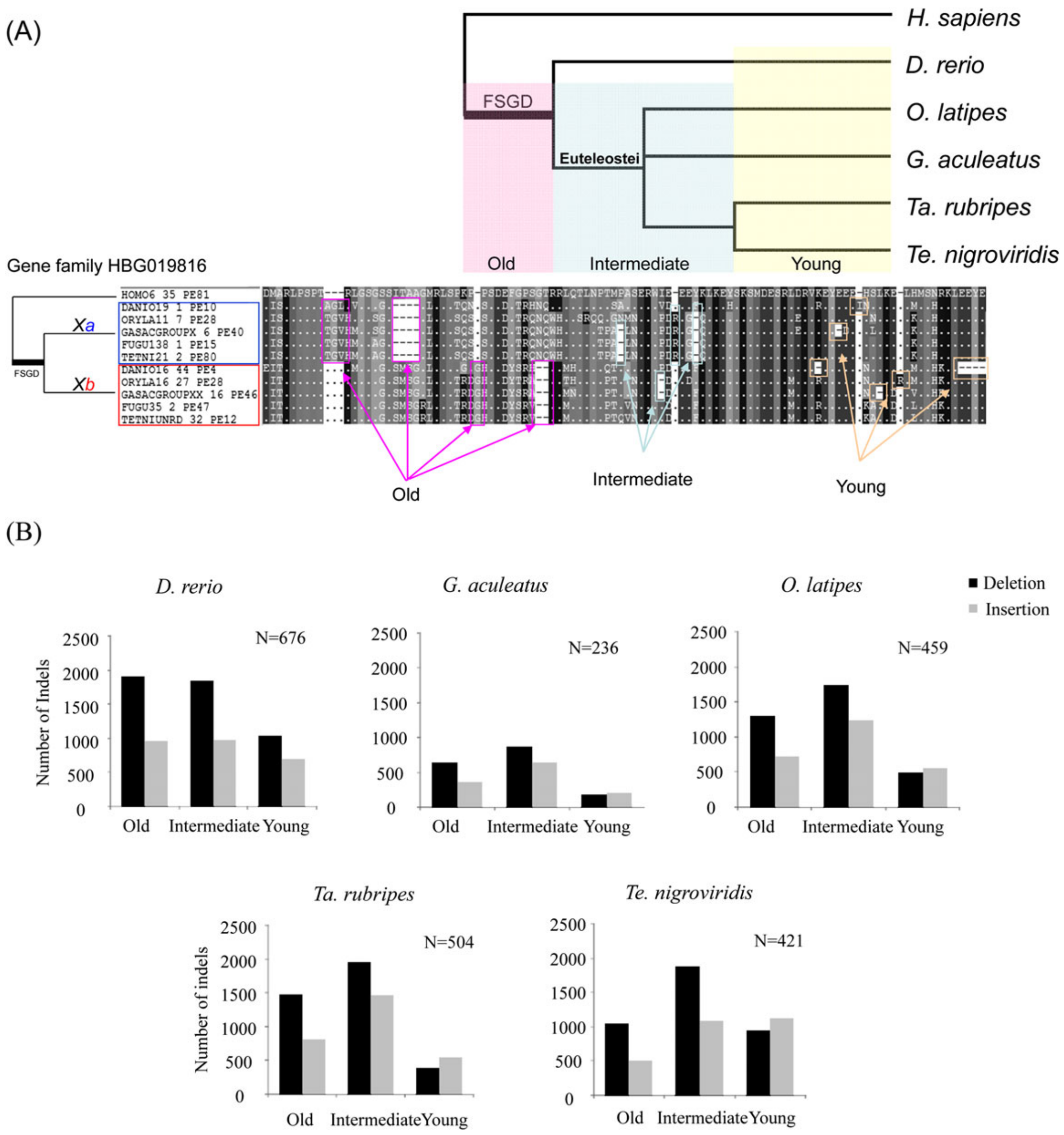

Fig. 5. Indels of different ages differ in their abundance. (*A*) Classification of post-FSGD indels into three age categories based on the phylogeny of the five model teleosts: early (purple), intermediate (blue), and late (yellow). See text for details. The example shows indels in a partial alignment of a duplicate gene family HBG019816. Candidate examples for old, intermediate, and late indels are framed with purple, blue, and yellow rectangles in the alignment. (*B*) Number of insertions and deletions in the three age categories in each teleost species. The numbers *N* in the inset indicate the number of duplicate gene pairs in each teleost.

The second column from the left in table 2 (background residues) shows the total percentage of amino acids that fall into each of the three secondary structure categories. Columns 3, 4, and 5 indicate the percentage of insertions, deletions, and indels in each of these three categories. It is evident from table 2 that both insertions and deletions occur preferentially in loop regions and that they are underrepresented both in helical and β-sheet regions. For example, whereas 27% of amino acids in a protein are contained in helical regions, fewer than 6% of indels occur in helical regions. This difference is just as dramatic for amino acids in β-sheets. They cover more than 13% of amino acids in a protein, but only 2.2% of indels occur in them. Conversely, on average, ≈59% of a protein's amino acids are parts of a loop region, whereas more than 78% of indels occur in loop regions. A χ^2 test shows that indels are significantly more abundant in loop regions than in helical or β-sheet regions ($P < 2.0 \times 10^{-16}$). When insertions and deletions are considered separately, the same excess of indels in loop regions emerges and is also statistically significant ($P < 2.5 \times 10^{-16}$). Supplementary tables S6 and S7 (Supplementary Material online) contain the same data but for each of the five fish genomes separately. Here also, one observes an excess of indels in loop regions that is significant ($P < 2.2 \times 10^{-16}$ for each species). In a related analysis, we asked whether the three different elements of secondary structure showed differences in the length of indels that occur in them but found no such differences

Table 2. Secondary Structure Context of Indel Events.

Status	Total Residues (%)	Insertion (%)	Deletion (%)	Total Indels (%)
Secondary structure[a]				
H	27.27	5.15	6.54	5.95
S	13.64	1.48	2.79	2.24
C	59.09	75.12	81.46	78.79
Solvent accessibility[b]				
b	53.60	20.08	16.43	17.97
e	46.40	39.28	52.41	46.88

[a] Secondary structure predicted by PSIPRED 3.21, with H = helix, S = β-sheet, and C = coil.

[b] Relative solvent accessibility predicted by ACCpro 4.01, with b = buried and e = exposed. The percentages in columns 2–5 do not add up to 100 because a small number of indels occur at the boundary between two regions.

(Mann–Whitney *U*-test, $P > 0.1$). All three secondary structure elements contain indels with an average of approximately seven codons in length (data not shown).

An important distinction among amino acids is whether an amino acid is buried inside a protein or exposed to solvent. Globular proteins, for example, have a surface of solvent-exposed amino acids, and a core of amino acids that are buried. Disrupting this core is more likely to perturb protein structure than disrupting solvent-exposed regions. We asked whether indels preferentially affect buried or solvent-exposed amino acids. Column 2 from the left of table 2 shows the overall percentage of amino acids in our study proteins that are buried ("b") and solvent exposed ("e"). Columns 3–5 show the incidence of insertions, deletions, and indels for these two categories of amino acids. For instance, only 18% of indels occur at buried amino acids, whereas 47% occur at exposed amino acids. (The remaining 35% occur between buried and exposed amino acids.) The excess of indels at exposed amino acids stands in contrast to the total percentage of exposed amino acids, which comprise a minority (46.4%) of all amino acids. The excess is statistically significant for both kinds of indels (χ^2 test, $P < 2.0 \times 10^{-16}$), a pattern that also holds for each of the five fish species separately (χ^2 test, $P < 2.2 \times 10^{-16}$).

Indels Are Associated with High Sequence Divergence

On the one hand, genes that evolve faster and that are thus under lower evolutionary constraints, might tolerate indels more readily. On the other hand, indels themselves can be mutagenic and increase sequence divergence in a genomic region (Tian et al. 2008). In either case, one would predict that a larger numbers of indels in a gene should be associated with greater sequence divergence between two duplicate genes. To test this prediction, we examined the statistical association between the divergence of duplicate genes and the abundance of indels in these genes. Specifically, we used the number of nonsynonymous substitutions per nonsynonymous site (K_a) as a measure of divergence between two duplicate genes. K_a ranged from 0.0032 to 1.2644 for duplicate genes in our data set. We used indel density, which ranged from 0.27 to 19.7 indels/kbp, as a measure of indel abundance.

Figure 6 shows the relationship between K_a and indel density for each of the five teleosts. K_a increases with indel density. For example, the average K_a in *D. rerio* between duplicate genes without indels is 0.14, which increases to 0.38 for duplicate genes with more than 10 indels/kbp. The statistical association between K_a and indel density is significantly different from zero in each of the five species (fig. 6; Spearman's rho ≥ 0.48, $P < 1.0 \times 10^{-6}$). Supplementary figure S3 (Supplementary Material online) shows the relationship between indel density on the one hand, and the fraction of synonymous substitutions per synonymous site K_s, as well as the ratio K_a/K_s—a measure of evolutionary constraint on a gene, and the pairwise nucleotide distance between duplicate genes on the other hand. Synonymous divergence K_s exceeds 1.0 for 99.6% of duplicates in each species. This means that synonymous substitutions are near saturation in many genes and can thus not be estimated accurately (Li 1997). It is therefore not surprising that K_s and indel density do not show a significant association in any of the five species (Spearman's rho ≤ 0.32, $P > 0.1$). In contrast, both K_a/K_s and overall nucleotide diversity are positively associated with indel density in all five species (Sperman's rho $> 0 .32$, $P < 1.0 \times 10^{-5}$). Overall, the associations between overall and nonsynonymous divergence, on the one hand, and indel density, on the other hand, are consistent both with the postulated mutagenic nature of indels as well as with lower evolutionary constraints on genes with many indels (Tian et al. 2008; Hollister et al. 2010).

Effects of Indels on Tertiary Structure

To study the effects of indels on the tertiary structure of proteins encoded in our duplicate genes, we estimated the structural divergence of those proteins for which structural information is available (see Materials and Methods). Specifically, we focused on the structure's backbone defined by the spatial coordinates of its alpha carbon atoms and estimated structural divergence as the root mean square deviation of these coordinates between two duplicate proteins. We determined indel density (number of indels per 100 amino acids) and amino acid divergence (percentage of amino acids that differed between two proteins) from those regions whose structure could be aligned (see Materials and Methods). We merged data from all five fish species for this analysis. See supplementary table S8 (Supplementary Material online) for the duplicate gene pairs, we used in this analysis. Supplementary table S9 (Supplementary Material online) summarizes the results of this analysis.

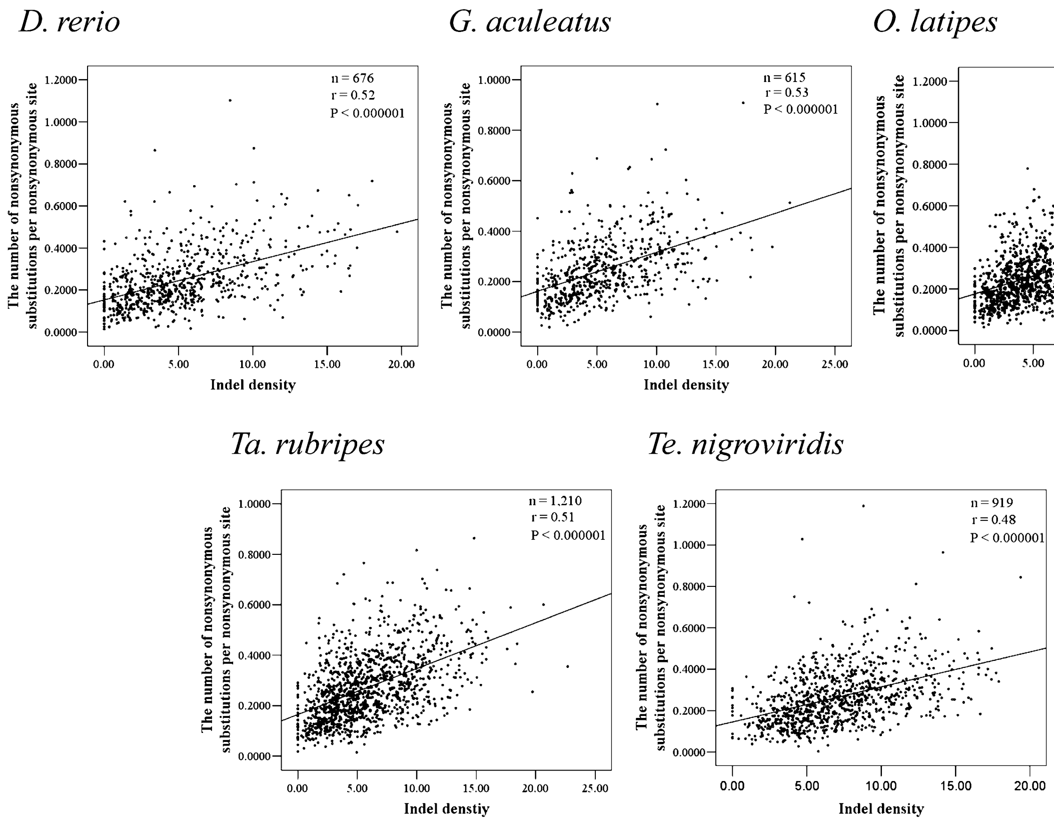

FIG. 6. Association between K_a and indel density in duplicate genes for each teleost. K_a is the number of nonsynonymous substitutions per nonsynonymous site that occurred between two duplicate genes. Indel density is the number of indels in a duplicate gene pair per kbp. The insets show the numbers n of duplicate gene pairs, as well as the Spearman's rank correlation coefficient r, and the P value for a test of the null hypothesis that r is identical to zero.

We found that a majority of duplicate protein pairs with sufficient structural data (67.1%, 897 of 1,337 gene pairs) contained indels in at least one of the two members of a pair. To exclude possible structural divergence between duplicates due to the existence of different homology modeling templates, we only focused on those 88.6% (1,184 of 1,337) of duplicate gene pairs with identical templates. The mean structural divergence of duplicate protein pairs with indels (0.363 ± 0.175 Å) was ten times larger than that of pairs without indels (0.031 ± 0.015 Å), a difference that is significant (Mann–Whitney U-test, $P < 1.0 \times 10^{-4}$). Because amino acid changes can also cause structural divergence (Grishin 2001), the question arises whether most structural divergence occurs due to the presence of indels or due to amino acid divergence. We found that structural divergence increases only weakly with amino acid divergence (Spearman's rho = 0.24, $P < 1.0 \times 10^{-5}$, n = 781; fig. 7A) but depends more strongly on the presence of indels (Spearman's rho = 0.74, $P < 1.0 \times 10^{-6}$, n = 781; fig. 7B). In addition, we observed a significant linear correlation between amino acid divergence and indel density (Spearman's rho = 0.29, $P < 1.0 \times 10^{-5}$, n = 781; fig. 7C) in duplicate genes with indels, just as we had in our closely related previous analysis on K_a (fig. 6). Zhang, Wang, et al. (2010) showed that structural divergence within gene families can be explained by a bilinear model combining indels and substitutions. However, for our data, a bilinear regression analysis of structural divergence as a dependent variable against indel density and amino acid divergence as independent variables shows that amino acid divergence is not significantly associated with structural divergence (P = 0.48). Overall, our results indicate that indels influence structural divergence to a much greater extent than amino acid changes do.

Indel-Induced Frameshift Mutations in Duplicate Genes

Thus far, all of our analyses were focused on in-frame indels, which affect multiples of 3 nt. Indels without this property cause frameshift mutations and thus are much more disruptive to protein sequence and structure. However, because such mutations can also be preserved and contribute to functional divergence of duplicate genes (Raes and Van de Peer 2005), we inquired about the incidence of frameshifting indels in our data set. Briefly, they are few comprising less than 2.0% (55 of 3,914) of all duplicates in four of the five species.

Combining data from all five species, we found very few frameshifting indels in four of the five species. Specifically, we found only 55 candidate examples (2% of 3,914 indels) that are not obvious alignment artifacts in the genomes of *D. rerio*, *G. aculeatus*, *O. latipes*, and *Ta. rubripes*. Supplementary table S10 (Supplementary Material online) lists the gene families in which these indels occurred, and supplementary figure S4

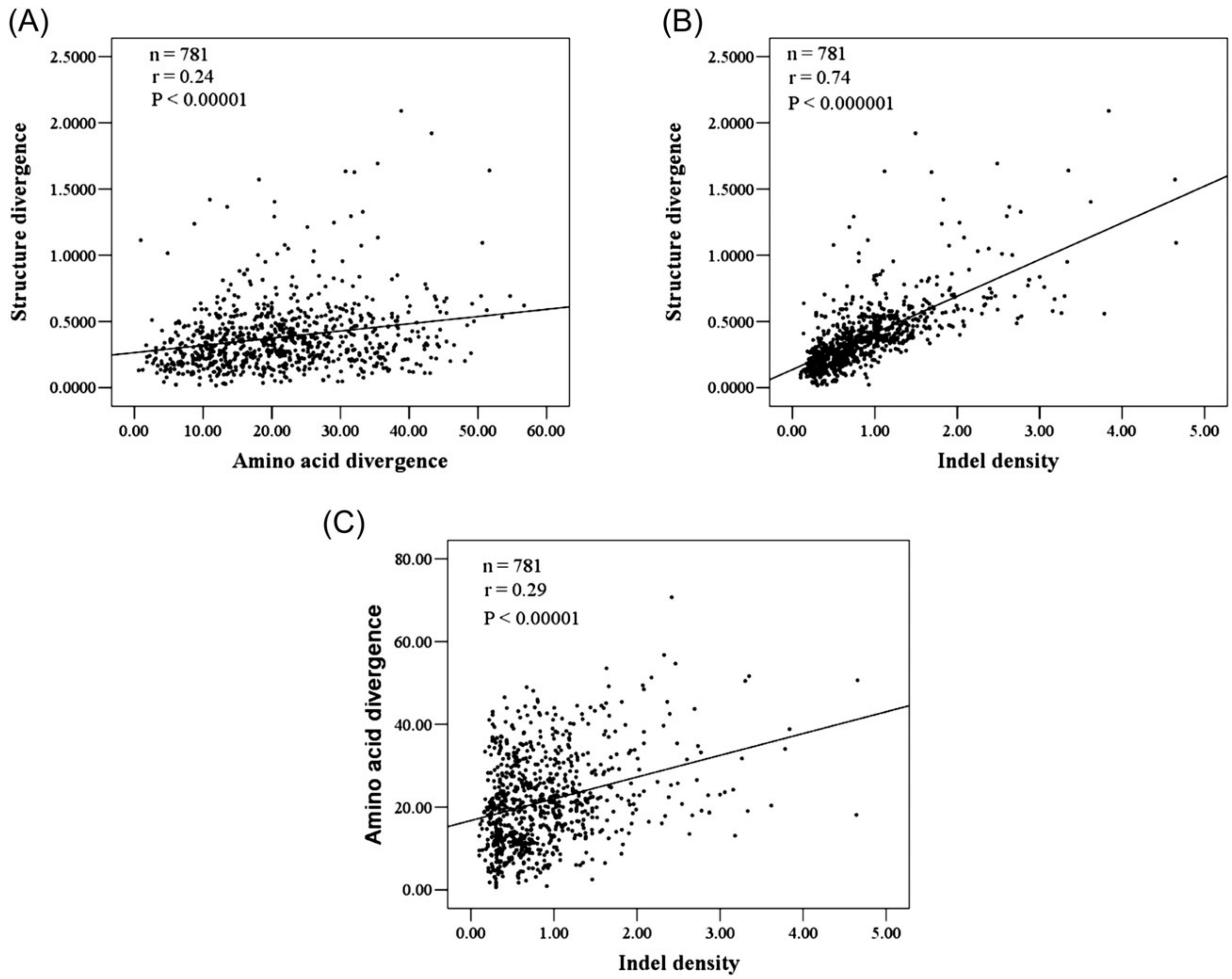

FIG. 7. Linear correlations of (*A*) structure divergence versus amino acid divergence, (*B*) structure divergence versus indel density, and (*C*) amino acid divergence versus indel density for duplicate genes with indels. The insets show the numbers *n* of duplicate gene pairs with indels, Spearman's rank correlation coefficient *r*, and the *P* value for a test of the null hypothesis that *r* is identical to zero.

(Supplementary Material online) shows typical examples. The remaining species, *Te. nigroviridis*, contained 350 such candidate frameshift mutations, which are too many to manually inspect all alignments. We inspected the alignments of 30 gene families in this species and found that 25 of the 30 candidates are not likely to be alignment artifacts. However, we cannot rule out that these indels in *Te. nigroviridis* might result from artifacts in its genome annotation, and thus, we excluded these indels from further analysis.

We next asked where, how, and when the remaining 55 frameshift mutations occurred. We observed no frameshift mutation at the N terminus, 47 such mutations in the middle of a protein, and 8 at the C-terminus. The preference for middle regions, and the avoidance of terminal regions was highly significant ($P < 1.05 \times 10^{-15}$, two-tailed χ^2 test, based on a null hypothesis of a 1:1:1 distribution of N-terminal, middle, and C-terminal indels). Nine of the frameshift mutations were caused by insertions and 18 by deletions. There is no significant association between the region in a protein where a frameshift mutation occurred and whether it occurred through an insertion or through a deletion (Fisher's exact test, $P > 0.05$). We add a note of caution to the interpretation of this data: Even though we inspected alignments carefully for potential artifacts, it is difficult to exclude with certainty that some apparent N- and C-terminal indels are due to genome sequencing or gene prediction errors. Finally, frameshift mutations could occur at different times after the FSGD event (fig. 5A). We observed an excess of frameshift mutations that are young, that is, 44 of 55 frameshift mutations occur in only one duplicate and in only one teleost.

Discussion

Past studies on primate and bacterial genomes showed that indels are ten times less abundant than nucleotide substitutions (Nachman and Crowell 2000; Chen et al. 2009). Nonetheless, many indels survive in genomes to this day, implying that they were either not very deleterious, perhaps neutral, or even beneficial when they first arose (Pascarella and Argos 1992; Grishin 2001; Denver et al. 2004; Taylor et al. 2004; Tian et al. 2008; Yang et al. 2010). In the five teleost genomes we study, more than 95% of singletons or duplicate genes carry indels compared with their human orthologs. Figures like these demonstrate that indels are important sources of genetic variation in fish genomes. In this discussion, we focus on their role in the evolution of duplicate genes.

Models of duplicated gene evolution differ in their predictions of how gene duplications affect sequence evolution. On the one hand, Ohno's classical model of neofunctionalization holds that one duplicated copy continues to function, whereas the other is no longer under purifying selection and can diverge freely (Ohno 1970). In this scenario, the second gene would evolve much faster than the first. This prediction is consistent with the observation that many duplicate genes diverge asymmetrically, that is, at significantly different rates, in both sequence (Conant and Wagner 2003) and function (Wagner 2002). On the other hand, both duplicates might experience relaxed selection and diverge after the duplication (Force et al. 1999; Lynch and Conery 2000; Lynch and Force 2000), which suggests that they both would evolve faster than single-copy genes. Our data, albeit from old duplicates, largely support the second scenario. Specifically, we observe that the incidence of indels is at least 25% higher for both members of a duplicate gene pair than in single-copy genes. (Note that the gene duplicates we analyze are old, with a synonymous divergence $K_s > 1$ for 99.6% of duplicate pairs, supplementary fig. S5, Supplementary Material online, such that much of the increased incidence of indels in duplicate genes may reflect relaxed selection in the distant past.) Also relevant in this regard is our observation that only 1.5–8% of duplicate genes accumulate indels asymmetrically, that is, one duplicate shows significant more or fewer indels than the other. These observations are not consistent with the first scenario. They suggest that both duplicates typically experience relaxed purifying selection.

The most severe relaxation of selection after gene duplication is expected to occur within a few million years (Force et al. 1999; Lynch and Conery 2000; Lynch and Force 2000). Our analysis is fully consistent with this notion because we found significantly fewer young indels than old indels (which originated close to the FSGD). Specifically, up to 40% of all indels occurred after the FSGD but before teleost diversification. The FSGD event is likely to have occurred approximately 350 Ma (Meyer and Van de Peer 2005), and the divergence time between *D. rerio* and other euteleosts (e.g., *G. aculeatus*, *O. latipes*, *Ta. rubripes*, and *Te. nigroviridis*) has been estimated at 307 my (Peng et al. 2009). Thus, ≈40% of our indels occurred within approximately the first 10% of the time interval since the FSGD or in the first 40 my after the FSGD. Furthermore, when we compared the incidence of indels in three age categories between duplicated genes and singletons, we found that indels in the youngest age category are no more abundant than in single-copy genes. This implies that relaxed selection on duplicate genes does not continue to this day.

Our duplicate genes contain 30–80% more deletions than insertions, depending on the genome, a difference that is statistically significant, and that exists for indels in most length categories. In principle, this excess of deletions can have two possible causes. First, it could be caused by a bias in the mutational mechanisms causing indels, such that deletions occur more frequently than insertions. This would be expected on the basis of the thermodynamics of replication slippage, in which an insertion requires the melting and rereplication of a segment of previously duplicated DNA, whereas in a deletion, a polymerase only needs to skip unreplicated bases (Petrov 2002). Second, deletions may be no more frequent than insertions, but selection may drive them to high frequency or to fixation more often.

We note that deletion bias we see is consistent with previous studies in vertebrates. For example, deletions are more abundant than insertions in rodent protein-coding genes (Taylor et al. 2004) and in mammalian genomes (Fan et al. 2007). Human pseudogenes (Zhang and Gerstein 2003) and noncoding nucleotide sequences of primates (Saitou and Ueda 1994) contain more small deletions than insertions. Because insertions and deletions within processed pseudogenes and in noncoding sequences may be selectively neutral, we speculate that the deletion bias we see does not reflect the action of natural selection but a general property of DNA mutations in genes and genomes.

In addition to a prevalence of deletions, our data also show another bias, namely a prevalence of short over long indels. Specifically, single-codon insertions and deletions are the most frequent indels, and the frequency of indels decreases approximately exponentially with their length. This distribution of indel lengths may again be a combination of two causes that are indistinguishable from our data. The first cause is a bias in mutations, which may preferentially create short indels. One candidate process that may account for some of the shortest indels is alternative splicing at NAGNAG motifs (N stands for any nucleotide), which has been implicated in an increased incidence of single-codon indels at 3′ splice sites in both mammals and Drosophila (Bradley et al. 2012). The second cause is natural selection, which may preferentially eliminate longer indels, because they may disrupt protein structure and function to a greater extent. A prevalence of short indels has also been observed in other organisms, for example, in a comparison of 19 mammalian genomes (Fan et al. 2007) as well as in orthologous genes in the mouse and rat genomes (Taylor et al. 2004). Previous authors (Wolf et al. 2007) suggested that indel length may increase with indel age. This suggestion is consistent with the observation that short indels are much less abundant in our study genomes, where 20% of all indels affect only one codon, than in rodents, where 60% affect only one codon (Taylor et al. 2004). We note that mice and rats diverged approximately 16 Ma (Springer et al. 2003), whereas the indels we study are up to 350 my old (Meyer and Van de Peer 2005).

Our data also show that duplicate genes with many indels are highly diverged in their amino acid sequences (fig. 6). These observations are consistent with observations from interspecies genome comparisons between primates, rodents, fruit flies, rice, yeast, and bacteria (Tian et al. 2008; Chen et al. 2009; Zhu et al. 2009); with observations from intraspecies genomic comparisons within yeast, *Arabidopsis*, and *Oryza* (Tian et al. 2008; Zhang et al. 2008; Hollister et al. 2010); and with protein evolution data (Yang et al. 2009; Zhang et al. 2011). The association between indel number and amino acid divergence can have two principal causes. On the one hand, genes that are subject to weak evolutionary constraints can tolerate many genetic changes, including nucleotide changes and indels. In other words, the association may exist because selection constrains sequence divergence in different genes to a different extent. On the other hand, indels themselves could be mutagenic, and they may cause an increased incidence in point mutations in nearby nucleotides (Tian et al. 2008). A candidate reason is that indels can cause meiotically paired chromosomes to form heteroduplex DNA in heterozygotes, which may attract DNA repair systems to indel-containing regions, which may lead to a higher mutation rate in these regions (Tian et al. 2008). Although the two scenarios are difficult to distinguish, the second scenario and its likely importance (Hodgkinson et al. 2009; Conrad, Bird, et al. 2010; Conrad, Pinto, et al. 2010; Hollister et al. 2010) suggest that indels may have at least contributed to the sequence divergence of duplicate genes.

Indels contribute more to tertiary structure divergence between duplicate genes than amino acid changes because indel density, but not amino acid divergence, shows a highly positive association with structure divergence in our data. These observations are consistent with analyses based on proteins in structure databases (Zhang, Wang, et al. 2010; Zhang et al. 2011). By affecting protein structure, indels can also contribute to the functional divergence between homologous proteins (Pascarella and Argos 1992; Grishin 2001; Reeves et al. 2006; Chan et al. 2007; Jiang and Blouin 2007; Wolf et al. 2007; Zhang, Wang, et al. 2010; Zhang et al. 2011). Extreme examples involve the loss of entire protein domains, such as a complete deletion of the N-terminal domain (five exons) of the duplicated sialic acid synthase in Antarctic zoarcid fish. This deletion helped turn the protein into an antifreeze protein (Deng et al. 2010). Similarly, domain losses may have facilitated accelerated evolution and neofunctionalization of duplicate snake venom metalloproteinase toxin genes (Casewell et al. 2011).

Because protein structure is usually a prerequisite for function, one would predict that changes in this structure may be subject to strong selection. Our observations support this expectation. First, we showed that indels in our duplicate genes preferentially occur in loop regions of a protein's secondary structure, where they do not disrupt helices or pleated β-sheets. In the latter two secondary structure elements, we find significantly fewer indels than expected from the total percentage of a protein's structure contained in helices and β-sheets. Our observations are consistent with other work, demonstrating that indels are underrepresented in proteins with well-defined folds and that they accumulate in loops (Pascarella and Argos 1992; Taylor et al. 2004). However, some indels outside loop regions may well be functionally important. For example, small secondary structure insertions may help proteins evolve new functions because they can modify a protein's active site geometry, its surface conformation, and its molecular interaction partners (Reeves et al. 2006; Chan et al. 2007). Second, we observed that indels occur preferentially in solvent-exposed regions of a protein and not in the buried hydrophobic core, where they may disrupt protein structure. This is consistent with the previous observation that buried amino acids evolve more slowly than exposed amino acids (Goldman et al. 1998). Third, indels are especially prevalent in regions encoding the N- or C-terminus of a protein, where 98% of our duplicated pairs contain indels. This prevalence can be explained by a "permutation by duplication" model (Peisajovich et al. 2006) that is supported by fruit fly data (Yang et al. 2008). Indels could thus be especially important in generating new protein structures in those regions.

Finally, we also identified a small percentage of candidate indels that cause a frameshift mutation, although we note that it is difficult to exclude alignment artifacts in the identification of such indels. Frameshift mutations are often considered to be deleterious and of little importance for the evolution of novel gene functions, but their preservation may be facilitated by the redundancy that gene duplication creates (Ohno 1970). In this case, they may even attain important roles in the functional divergence of genes and proteins. For instance, Vandenbussche et al. (2003) observed three instances of functional divergence through frameshift indels after duplication in the MADS-box gene family in plants. Relatedly, Janssens et al. (2008) proposed that frameshift mutations may help explain the divergence of C-terminal sequences in the MIKC gene subfamilies and the retention of duplicated MIKC genes.

Several caveats of this study are noteworthy. First, duplicate genes in fish genomes could have originated before the FSGD, during the FSGD, or after the FSGD. A potential limitation of our work is that we cannot be absolutely certain to have identified only duplicates that originated in the FSGD. However, by enforcing that duplicate genes must exist in at least two species and that their phylogeny has the structure shown in figure 1*B*, we can exclude most of genes that underwent duplication long after the FSGD. We also note that the age of our duplicates is consistent with the notion that originated in the FSGD. Specifically, none of our duplicate genes are young, as indicated by the observation that their synonymous divergence at synonymous sites (K_s) exceeds one for 99.6% of genes (Li 1997). An independent benchmark of our analysis is provided by Kassahn et al. (2009), who studied 615 FSGD duplicated gene pairs in *D. rerio*. Our analysis independently identified 614 of these 615 genes, suggesting that our procedure misses only a small fraction of duplicate genes. Taken together, these observations suggest that most of our identified duplicates stem from the FSGD.

Second, we are aware that conclusions of studies like ours critically depend on alignment quality (Tian et al. 2008). To identify possible artifacts that stem from specific alignment algorithms, we aligned the amino acid sequences of each gene family from all five species and from the human genome with ClustalW (Thompson et al. 1994), with MAFFT (Katoh and Toh 2008), and with PRANK (Loytynoja and Goldman 2005). Those algorithms yielded consistent results, for example, with respect to the patterns depicted in figure 4 (supplementary fig. S1, Supplementary Material online), which suggests that our observations are not sensitive to alignment procedures.

Third, sequencing and gene prediction errors may affect studies like ours (Nagy et al. 2008; Harrow et al. 2009; Hubisz et al. 2011; Nagy and Patthy 2011; Thompson et al. 2011; Prosdocimi et al. 2012). We thus pursued three independent approaches to verify that our findings are robust to such errors. The first uses the observation that most gene prediction errors that affect indels occur in the regions of a gene that encode the N-terminal or C-terminal part of a protein (Hubisz et al. 2011; Thompson et al. 2011; Prosdocimi et al. 2012). We thus repeated our analyses after excluding all such terminal indels. This procedure decreased indel numbers (e.g., by 17% in *D. rerio*), but it did not affect our results materially. Post-FSGD indels are still pervasive between duplicates and the length distribution of indels (fig. 4), as well as the association between sequence divergence and indel density (fig. 6) persists (supplementary fig. S6*A* and *B*, Supplementary Material online). Second, we showed that our findings also hold for gene pairs that contain no intron (26 pairs) or only one intron (155 pairs) because such gene pairs are less likely to be subject to an important class of gene prediction errors, namely those affecting gene structure and alternative splicing variants (supplementary fig. S6*C* and *D*, Supplementary Material online). Furthermore, we also repeated our analyses with recently updated Ensembl gene annotation data (version 64) and used data from *D. rerio* and *Te. nigroviridis* to show that the patterns we detect persist after this update (supplementary fig. S6*E* and *F*, Supplementary Material online). The likely reason why our observations are robust is that our analysis focuses on global genome-wide patterns of indel accumulation, and not on case studies of individual genes, which would be much more sensitive to gene prediction errors.

Fourth, our tertiary structure analyses depend on the accuracy of homology modeling, which is influenced by the evolutionary distance between the target and the template proteins (Bordoli et al. 2008). To exclude possible modeling artifacts, we focused on duplicate genes that are sufficiently similar that the same template protein can be used in homology modeling. We did not model side chains but focused on the protein backbone because backbone predictions are more reliable for divergent sequences (Bordoli et al. 2008). In addition, we compared predicted structures only in regions of duplicate genes that have high sequence identities of greater than 70%. Finally, we note that our observations are consistent with earlier observations based on experimentally determined protein structures (Zhang et al. 2011).

Fifth, we only used human genes as outgroups in our indel characterization because the human genome is well-annotated compared with other genomes. However, we note that a more complex analysis might use outgroup genes from additional species. They might help polarize pre-FSGD and post-FSGD indels as insertions or deletions with greater accuracy. We leave such an analysis for future work.

In conclusion, our results show that indels have played important roles during the evolution of duplicate genes. Both members of a duplicate gene pair contain significantly more indels than singletons. Indels pervasively occurred between duplicates, and these indels accumulated preferentially early after the genome duplication. Indels are associated with high amino acid divergence, and they lead to the structural divergence of proteins. The consistency of our observations across five completely sequenced genomes suggests their generality. Indels are an important driving force of duplicate gene and genome divergence.

Supplementary Material

Supplementary tables S1–S10 and figures S1–S6 are available at *Molecular Biology and Evolution* online (http://mbe.oxfordjournals.org/).

Acknowledgments

We wish to thank Prof. Peter Holland (University of Oxford) for insightful discussions during the conception of this work. We are grateful to Prof. Torsten Schwede (University of Basel), Marco Biasini (University of Basel), and Dr Junfeng Huang (Sun Yat-sen University) for technical advice on protein structure analysis and to Dr Niv Sabath (University of Zurich) for statistical assistance. We also thank Dr Eric Hayden (University of Zurich) for reading an early version of the manuscript as well as the editor and three anonymous reviewers for constructive comments. We acknowledge support through Swiss National Science Foundation grant 315230-129708 as well as through the YeastX project of SystemsX.ch and the University Priority Research Program in Systems Biology at the University of Zurich.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389–3402.

Amores A, Force A, Yan YL, et al. (13 co-authors). 1998. Zebrafish hox clusters and vertebrate genome evolution. *Science* 282:1711–1714.

Aparicio S, Chapman J, Stupka E, et al. (41 co-authors). 2002. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* 297:1301–1310.

Arnold K, Bordoli L, Kopp J, Schwede T. 2006. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* 22:195–201.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B.* 57:289–300.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235–242.

Birney E, Clamp M, Durbin R. 2004. GeneWise and Genomewise. *Genome Res.* 14:988–995.

Bordoli L, Kiefer F, Arnold K, Benkert P, Battey J, Schwede T. 2008. Protein structure homology modeling using SWISS-MODEL workspace. *Nat Protoc.* 4:1–13.

Bradley RK, Merkin J, Lambert NJ, Burge CB. 2012. Alternative splicing of RNA triplets is often regulated and accelerates proteome evolution. *PLoS Biol.* 10:e1001229.

Brunet FG, Crollius HR, Paris M, Aury JM, Gibert P, Jaillon O, Laudet V, Robinson-Rechavi M. 2006. Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol Biol Evol.* 23:1808–1816.

Casewell NR, Wagstaff SC, Harrison RA, Renjifo C, Wuster W. 2011. Domain loss facilitates accelerated evolution and neofunctionalization of duplicate snake venom metalloproteinase toxin genes. *Mol Biol Evol.* 28:2637–2649.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol.* 17:540–552.

Chan SK, Hsing M, Hormozdiari F, Cherkasov A. 2007. Relationship between insertion/deletion (indel) frequency of proteins and essentiality. *BMC Bioinformatics* 8:227.

Chen JQ, Wu Y, Yang H, Bergelson J, Kreitman M, Tian D. 2009. Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria. *Mol Biol Evol.* 26:1523–1531.

Conant GC, Wagner A. 2003. Asymmetric sequence divergence of duplicate genes. *Genome Res.* 13:2052–2058.

Conrad DF, Bird C, Blackburne B, Lindsay S, Mamanova L, Lee C, Turner DJ, Hurles ME. 2010. Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. *Nat Genet.* 42:385–391.

Conrad DF, Pinto D, Redon R, et al. (28 co-authors). 2010. Origins and functional impact of copy number variation in the human genome. *Nature* 464:704–712.

Deng C, Cheng CH, Ye H, He X, Chen L. 2010. Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict. *Proc Natl Acad Sci U S A.* 107:21593–21598.

Denver DR, Morris K, Lynch M, Thomas WK. 2004. High mutation rate and predominance of insertions in the Caenorhabditis elegans nuclear genome. *Nature* 430:679–682.

Dufayard JF, Duret L, Penel S, Gouy M, Rechenmann F, Perriere G. 2005. Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics* 21:2596–2603.

Duret L, Mouchiroud D, Gouy M. 1994. HOVERGEN: a database of homologous vertebrate genes. *Nucleic Acids Res.* 22:2360–2365.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.

Fan Y, Wang W, Ma G, Liang L, Shi Q, Tao S. 2007. Patterns of insertion and deletion in Mammalian genomes. *Curr Genomics.* 8:370–378.

Farre D, Alba MM. 2010. Heterogeneous patterns of gene-expression diversification in mammalian gene duplicates. *Mol Biol Evol.* 27:325–335.

Felsenstein J. 2004. PHYLIP (phylogeny inference package) version 3.6. Distributed by the author. Seattle (WA): Department of Genome Sciences, University of Washington.

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

Goldman N, Thorne JL, Jones DT. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* 149:445–458.

Grishin NV. 2001. Fold change in evolution of protein structures. *J Struct Biol.* 134:167–185.

Gu X, Zhang Z, Huang W. 2005. Rapid evolution of expression and regulatory divergences after yeast gene duplication. *Proc Natl Acad Sci U S A.* 102:707–712.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 52:696–704.

Guo B, Gan X, He S. 2010. Hox genes of the Japanese eel Anguilla japonica and Hox cluster evolution in teleosts. *J Exp Zool B Mol Dev Evol.* 314:135–147.

Guo B, Tong C, He S. 2009. Sox genes evolution in closely related young tetraploid cyprinid fishes and their diploid relative. *Gene* 439:102–112.

Ha M, Kim ED, Chen ZJ. 2009. Duplicate genes increase expression diversity in closely related species and allopolyploids. *Proc Natl Acad Sci U S A.* 106:2295–2300.

Harrow J, Nagy A, Reymond A, Alioto T, Patthy L, Antonarakis SE, Guigo R. 2009. Identifying protein-coding genes in genomic sequences. *Genome Biol.* 10:201.

He X, Zhang J. 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* 169:1157–1164.

Hodgkinson A, Ladoukakis E, Eyre-Walker A. 2009. Cryptic variation in the human mutation rate. *PLoS Biol.* 7:e1000027.

Hollister JD, Ross-Ibarra J, Gaut BS. 2010. Indel-associated mutation rate varies with mating system in flowering plants. *Mol Biol Evol.* 27:409–416.

Hormozdiari F, Salari R, Hsing M, Schonhuth A, Chan SK, Sahinalp SC, Cherkasov A. 2009. The effect of insertions and deletions on wirings in protein-protein interaction networks: a large-scale study. *J Comput Biol.* 16:159–167.

Hubbard T, Barker D, Birney E, et al. (35 co-authors). 2002. The Ensembl genome database project. *Nucleic Acids Res.* 30:38–41.

Hubisz MJ, Lin MF, Kellis M, Siepel A. 2011. Error and error mitigation in low-coverage genome assemblies. *PLoS One* 6:e17034.

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

Jaillon O, Aury JM, Brunet F, et al. (61 co-authors). 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature* 431:946–957.

Janssens SB, Viaene T, Huysmans S, Smets EF, Geuten KP. 2008. Selection on length mutations after frameshift can explain the origin and retention of the AP3/DEF-like paralogues in Impatiens. *J Mol Evol.* 66:424–435.

Jiang H, Blouin C. 2007. Insertions and the emergence of novel protein structure: a structure-based phylogenetic study of insertions. *BMC Bioinformatics* 8:444.

Jones DT. 1999. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol.* 292:195–202.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci.* 8:275–282.

Kasahara M, Naruse K, Sasaki S, et al. (38 co-authors). 2007. The medaka draft genome and insights into vertebrate genome evolution. *Nature* 447:714–719.

Kassahn KS, Dang VT, Wilkins SJ, Perkins AC, Ragan MA. 2009. Evolution of gene function and regulatory control after whole-genome duplication: comparative analyses in vertebrates. *Genome Res.* 19:1404–1418.

Katoh K, Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform.* 9:286–298.

Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature* 428:617–624.

Krissinel E, Henrick K. 2004. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr D Biol Crystallogr.* 60:2256–2268.

Kristiansson E, Hugenholtz P, Dalevi D. 2009. ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. *Bioinformatics* 25:2737–2738.

Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. *Mol Biol Evol.* 25:1307–1320.

Li C, Lu G, Ortí G. 2008. Optimal data partitioning and a test case for ray-finned fishes (Actinopterygii) based on ten nuclear loci. *Syst Biol.* 57:519–539.

Li W-H. 1997. Molecular evolution. Sunderland (MA): Sinauer Associates.

Loytynoja A, Goldman N. 2005. An algorithm for progressive multiple alignment of sequences with insertions. *Proc Natl Acad Sci U S A.* 102:10557–10562.

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

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

Miya M, Satoh TP, Nishida M. 2005. The phylogenetic position of toadfishes (order Batrachoidiformes) in the higher ray-finned fish as inferred from partitioned Bayesian analysis of 102 whole mitochondrial genome sequences. *Biol J Linn Soc Lond* 85:289–306.

Meyer A, Van de Peer Y. 2005. From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *Bioessays* 27:937–945.

Nachman MW, Crowell SL. 2000. Estimate of the mutation rate per nucleotide in humans. *Genetics* 156:297–304.

Nagy A, Hegyi H, Farkas K, Tordai H, Kozma E, Banyai L, Patthy L. 2008. Identification and correction of abnormal, incomplete and mis-predicted proteins in public databases. *BMC Bioinformatics* 9:353.

Nagy A, Patthy L. 2011. Reassessing domain architecture evolution of metazoan proteins: the contribution of different evolutionary mechanisms. *Genes* 2:578–598.

Negrisolo E, Kuhl H, Forcato C, Vitulo N, Reinhardt R, Patarnello T, Bargelloni L. 2010. Different phylogenomic approaches to resolve the evolutionary relationships among model fish species. *Mol Biol Evol.* 27:2757–2774.

Nelson JS. 2006. Fishes of the world. 4th ed. New York: John Wiley and Sons.

Ohno S. 1970. Evolution by gene duplication. New York: Springer-Verlag.

Pascarella S, Argos P. 1992. Analysis of insertions/deletions in protein structures. *J Mol Biol.* 224:461–471.

Peisajovich SG, Rockah L, Tawfik DS. 2006. Evolution of new protein topologies through multistep gene rearrangements. *Nat Genet.* 38:168–174.

Penel S, Arigon AM, Dufayard JF, Sertier AS, Daubin V, Duret L, Gouy M, Perriere G. 2009. Databases of homologous gene families for comparative genomics. *BMC Bioinformatics.* 10(Suppl 6):S3.

Peng Z, Diogo R, He S. 2009. Teleost fishes (Teleostei). In: Hedges SB, Kumar S, editors. The timetree of life. New York: Oxford University Press. p. 335–338.

Petrov DA. 2002. Mutational equilibrium model of genome size evolution. *Theor Popul Biol.* 61:531–544.

Pollastri G, Baldi P, Fariselli P, Casadio R. 2002. Prediction of coordination number and relative solvent accessibility in proteins. *Proteins* 47:142–153.

Prosdocimi F, Linard B, Pontarotti P, Poch O, Thompson JD. 2012. Controversies in modern evolutionary biology: the imperative for error detection and quality control. *BMC Genomics* 13:5.

Raes J, Van de Peer Y. 2005. Functional divergence of proteins through frameshift mutations. *Trends Genet.* 21:428–431.

Reeves GA, Dallman TJ, Redfern OC, Akpor A, Orengo CA. 2006. Structural diversity of domain superfamilies in the CATH database. *J Mol Biol.* 360:725–741.

Robinson-Rechavi M, Laudet V. 2001. Evolutionary rates of duplicate genes in fish and mammals. *Mol Biol Evol.* 18:681–683.

Rost B. 2003. Rising accuracy of protein secondary structure prediction. In: Chasman D, editor. Protein structure determination, analysis, and modeling for drug discovery. New York: Dekker. p. 207–249.

Saitou N, Ueda S. 1994. Evolutionary rates of insertion and deletion in noncoding nucleotide sequences of primates. *Mol Biol Evol.* 11:504–512.

Salari R, Schönhuth A, Hormozdiari F, Cherkasov A, Sahinalp SC. 2008. The relation between indel length and functional divergence: a formal study. Proceedings of the 8th International Workshop on Algorithms in Bioinformatics; 2008 Sep 15–19. Karlsruhe (Germany): Springer-Verlag. p. 330–341.

Semon M, Wolfe KH. 2007. Reciprocal gene loss between Tetraodon and zebrafish after whole genome duplication in their ancestor. *Trends Genet.* 23:108–112.

Springer MS, Murphy WJ, Eizirik E, O'Brien SJ. 2003. Placental mammal diversification and the Cretaceous-Tertiary boundary. *Proc Natl Acad Sci U S A.* 100:1056–1061.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.

Steinke D, Salzburger W, Braasch I, Meyer A. 2006. Many genes in fish have species-specific asymmetric rates of molecular evolution. *BMC Genomics* 7:20.

Studer RA, Penel S, Duret L, Robinson-Rechavi M. 2008. Pervasive positive selection on duplicated and nonduplicated vertebrate protein coding genes. *Genome Res.* 18:1393–1402.

Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* 23:1282–1288.

Taylor JS, Braasch I, Frickey T, Meyer A, Van de Peer Y. 2003. Genome duplication, a trait shared by 22000 species of ray-finned fish. *Genome Res.* 13:382–390.

Taylor MS, Ponting CP, Copley RR. 2004. Occurrence and consequences of coding sequence insertions and deletions in Mammalian genomes. *Genome Res.* 14:555–566.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22:4673–4680.

Thompson JD, Linard B, Lecompte O, Poch O. 2011. A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. *PLoS One* 6:e18093.

Tian D, Wang Q, Zhang P, Araki H, Yang S, Kreitman M, Nagylaki T, Hudson R, Bergelson J, Chen JQ. 2008. Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature* 455:105–108.

Vandenbussche M, Theissen G, Van de Peer Y, Gerats T. 2003. Structural diversification and neo-functionalization during floral MADS-box gene evolution by C-terminal frameshift mutations. *Nucleic Acids Res.* 31:4401–4409.

VanderSluis B, Bellay J, Musso G, Costanzo M, Papp B, Vizeacoumar FJ, Baryshnikova A, Andrews B, Boone C, Myers CL. 2010. Genetic interactions reveal the evolutionary trajectories of duplicate genes. *Mol Syst Biol.* 6:429.

Venter JC, Adams MD, Myers EW, et al. (274 co-authors). 2001. The sequence of the human genome. *Science* 291:1304–1351.

Wagner A. 2002. Asymmetric functional divergence of duplicate genes in yeast. *Mol Biol Evol.* 19:1760–1768.

Wolf Y, Madej T, Babenko V, Shoemaker B, Panchenko AR. 2007. Long-term trends in evolution of indels in protein sequences. *BMC Evol Biol.* 7:19.

Yang H, Wu Y, Feng J, Yang S, Tian D. 2009. Evolutionary pattern of protein architecture in mammal and fruit fly genomes. *Genomics* 93:90–97.

Yang H, Zhong Y, Peng C, Chen JQ, Tian D. 2010. Important role of indels in somatic mutations of human cancer genes. *BMC Med Genet.* 11:128.

Yang S, Arguello JR, Li X, et al. (12 co-authors). 2008. Repetitive element-mediated recombination as a mechanism for new gene origination in Drosophila. *PLoS Genet.* 4:e3.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol.* 17:32–43.

Zhang P, Gu Z, Li W-H. 2003. Different evolutionary patterns between young duplicate genes in the human genome. *Genome Biol.* 4:R56.

Zhang PG, Huang SZ, Pin AL, Adams KL. 2010. Extensive divergence in alternative splicing patterns after gene and genome duplication during the evolutionary history of Arabidopsis. *Mol Biol Evol.* 27:1686–1697.

Zhang W, Sun X, Yuan H, Araki H, Wang J, Tian D. 2008. The pattern of insertion/deletion polymorphism in Arabidopsis thaliana. *Mol Genet Genomics.* 280:351–361.

Zhang Z, Gerstein M. 2003. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Res.* 31:5338–5348.

Zhang Z, Huang J, Wang Z, Wang L, Gao P. 2011. Impact of indels on the flanking regions in structural domains. *Mol Biol Evol.* 28: 291–301.

Zhang Z, Wang Y, Wang L, Gao P. 2010. The combined effects of amino acid substitutions and indels on the evolution of structure within protein families. *PLoS One* 5:e14316.

Zhu L, Wang Q, Tang P, Araki H, Tian D. 2009. Genomewide association between insertions/deletions and the nucleotide diversity in bacteria. *Mol Biol Evol.* 26:2353–2361.