J Mol Evol (2017) 84:253–258
DOI 10.1007/s00239-017-9794-8

LETTER TO THE EDITOR

Complex Genes Are Preferentially Retained After Whole-Genome Duplication in Teleost Fish

Baocheng Guo[1]

Received: 20 March 2017 / Accepted: 3 May 2017 / Published online: 10 May 2017

Abstract Gene duplication generates new genetic material which, if retained after duplication, may contribute to organismal evolution. A whole-genome duplication occurred in the ancestry of teleost fish and consequently there are many duplicated genes in teleost genomes. Indeed, it has been proposed that the evolutionary diversification of teleost fish may have been stimulated by the fish-specific genome duplication (FSGD). However, it is not clear which factors determine which genes are retained as duplicate copies and which return to a singleton state after duplication. In the present study, gene complexity, in terms of encoded protein length and functional domain number, is compared between duplicate and singleton genes for nine well-annotated teleost genomes. A total of 933 gene families with retained duplicates and 4590 singleton gene families are analysed. Genes with retained duplicates are found to be significantly longer (27.9–38.2%) and to have more functional domains (20.5–26.5%) than singleton genes in all the nine teleost genomes, suggesting that genes encoded longer proteins with and more functional domains were preferentially retained after whole-genome duplication in teleosts. This differential retention of duplicated genes will have increased the genomic complexity of teleost fish after FSGD which, together with differential duplicated gene retention as a lineage-splitting force, may have greatly contributed to the successful diversification of teleost fish.

Keywords Duplicated gene · Singleton gene · Protein length · Domain number

Introduction

Gene duplication is commonly believed to be of major evolutionary significance because it generates new genetic material giving opportunities for innovation (Ohno 1970; Stephens 1951). The fact that large numbers of genes belong to multigene families reveals that duplication has been very prevalent in eukaryotic evolution (Zhang 2003). However, in most cases, duplicated genes are expected to be functionally redundant immediately after duplication and thus prone to evolutionary loss; only a proportion of duplicated genes will be retained long term (Lynch and Conery 2000). It is important, therefore, to understand what are the determinants that predispose towards retention of duplicated genes. Gene complexity has been reported to be one determinant affecting retention of duplicated genes in yeast: retained duplicate genes in yeast tend to have longer protein sequences, more functional domains and more cis-regulatory motifs than singleton genes (He and Zhang 2005). However, this association needs further assessment in other organisms (Chain et al. 2011).

Gene duplication can involve single genes, arrays of genes or whole genomes. The last of these mechanisms provides the best opportunity for analysing post-duplication gene retention since paralogy relationships between chromosomes ensure that gene loss can be recognised unambiguously. Teleosts provided one of the first unambiguous examples of ancient whole-genome duplication in

Electronic supplementary material The online version of this article (doi:10.1007/s00239-017-9794-8) contains supplementary material, which is available to authorized users.

✉ Baocheng Guo
guobaocheng@ioz.ac.cn

[1] The Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing, China

Table 1 Number, mean length and mean domain number of duplicated and singleton genes in each of the nine teleost fish

	Duplicated genes			Singleton genes		
	Number (pairs)	Length (amino acid)	Domain number	Number	Length (amino acid)	Domain number
Cavefish	587	970.4 ± 22.3[a]	4.0 ± 0.1[a]	3805	740.8 ± 9.7	3.3 ± 0.04
Cod	468	950.6 ± 26.6	4.1 ± 0.1	3506	687.7 ± 9.8	3.4 ± 0.04
Fugu	560	1025.0 ± 33.6	4.4 ± 0.1	3600	743.1 ± 9.7	3.5 ± 0.05
Medaka	507	941.0 ± 22.3	4.2 ± 0.1	3415	706.4 ± 9.9	3.4 ± 0.05
Platyfish	560	927.5 ± 21.0	4.0 ± 0.1	3680	724.9 ± 9.5	3.3 ± 0.04
Stickleback	606	961.5 ± 23.5	4.2 ± 0.1	3878	718.9 ± 9.8	3.4 ± 0.05
Tetraodon	506	907.0 ± 20.4	4.3 ± 0.1	3397	695.3 ± 9.1	3.4 ± 0.05
Tilapia	657	993.5 ± 21.4	4.1 ± 0.1	3965	757.6 ± 9.8	3.3 ± 0.04
Zebrafish	598	1018.0 ± 41.2	4.8 ± 0.1	3858	737.5 ± 9.8	3.9 ± 0.05

[a] Standard error of the mean

an animal lineage, an event in the common ancestor of all teleosts referred to as the fish-specific genome duplication (FSGD), teleost genome duplication (TSD) or the 3R. Teleost fish are thus ideal models to study the fate of duplicated genes. There is much interest in the genomic consequences of the FSGD and especially whether it influenced or promoted the diversification of teleosts (Amores et al. 1998; Taylor et al. 2003). Indeed, the FSGD gave rise to thousands of retained duplicated genes in extant teleost genomes (Guo et al. 2011, 2012; Inoue et al. 2015). To understand why some genes have been retained in teleost genomes after hundreds of million years while other duplicates were lost, here I compare gene complexity, assessed by protein length and number of protein domains, between FSGD-derived duplicated genes and singleton genes across nine well-annotated teleost genomes. Complex genes are found to be preferentially retained after whole-genome duplication in teleost fish.

Materials and Methods

Nine well-annotated teleost genomes were analysed: cavefish *Astyanax mexicanus*, cod *Gadus morhua*, fugu pufferfish *Takifugu rubripes*, medaka *Oryzias latipes*, platyfish *Xiphophorus maculatus*, three-spined stickleback *Gasterosteus aculeatus*, tetraodon pufferfish *Tetraodon nigroviridis*, tilapia *Oreochromis niloticus* and zebrafish *Danio rerio*. Species that were known to have undergone additional genome duplication after the FSGD were avoided.

The nine teleost genome sequences were retrieved from Ensembl release 76. Retained duplicated genes resultant from the FSGD event, and singleton genes whose duplicate copies have been lost, were retrieved from Inoue et al. (2015) in which orthologous and paralogous relationships of all teleost genes were resolved using tree-based methods with the human homolog as the outgroup. Specifically, to avoid false positive singleton and duplicated gene identification, only orthologous groups with singleton human genes and exact '1to1' or '1to2' (without further gene duplication after the FSGD event) teleost genes were used. The predicted number of functional domains for proteins encoded by each gene in each of the nine species was obtained from Ensembl using BioMart (Kasprzyk 2011). The statistics and figures were obtained with R version 3.3.2.

Results

Singleton genes, and genes with retained duplicates following the FSGD event, were retrieved from Inoue et al. (2015). In total, 933 gene families with retained duplicates and 4590 singleton gene families were used in the present study; the number of duplicated and singleton genes in each of the nine teleost fish is given in Table 1. The identities of each gene, and their human orthologs, are listed in Additional file Table S1.

The first measure of gene complexity examined was the length of encoded protein. Protein length was compared between post-FSGD-retained duplicates and singleton genes for each of the nine teleost genomes. The post-FSGD-retained duplicates encode deduced proteins with a mean length of 907.0 ± 20.4 (standard error of the mean) to 1025 ± 33.6 amino acids, and the mean length for singleton genes is 687.7 ± 9.8 to 757.6 ± 9.8 amino acids, depending on the species (Table 1). The mean length of post-FSGD retained duplicates is 27.9% (platyfish) to 38.2% (cod) greater than the mean length of singleton genes; the difference is significant in each of the nine teleost genomes (two-tailed Wilcoxon rank-sum tests, $P < 2.2 \times 10^{-6}$ for each species; Fig. 1a). The distribution of protein length shows that the significant

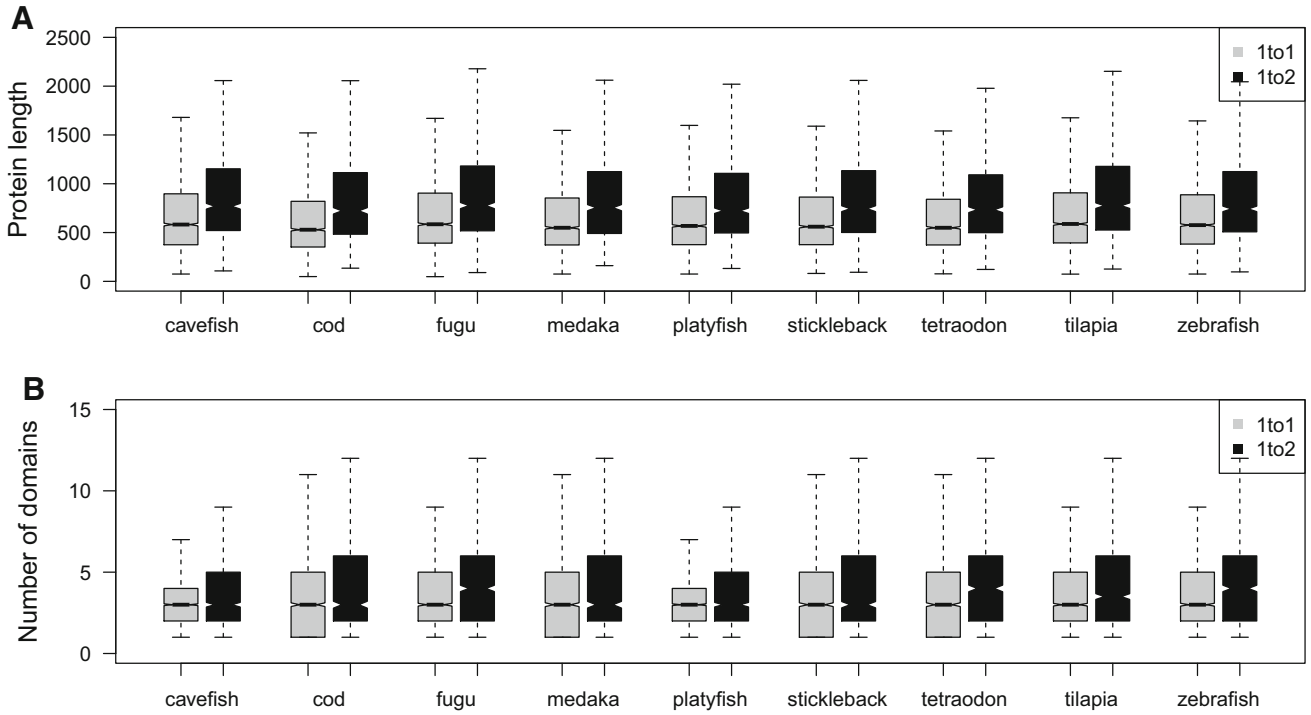

Fig. 1 Comparisons of **a** protein length in amino acids and **b** functional domain number between retained duplicates and singleton genes in nine teleost genomes. *1to1* singleton genes, *1to2* retained duplicate genes

difference in protein length between duplicated and singleton genes is not the result of outliers, and that long proteins (length >600, 700 or 800 amino acids) are frequent among retained duplicates than among singleton genes (Fig. 2a). Similar results were obtained when one random copy of a pair of duplicated genes was used in statistical comparison for each species, or when comparisons were restricted to the 101 duplicated genes and 1702 singleton genes found in all of the nine teleost genomes.

The second measure of complexity examined was the number of functional protein domains. The mean number of functional domains in proteins encoded by post-FSGD-retained duplicate genes ranged from 4.0 ± 0.1 (cavefish and platyfish) to 4.8 ± 0.1 (zebrafish), and the mean for singleton genes ranged from 3.3 ± 0.04 (cavefish, platyfish, and tilapia) to 3.9 ± 0.05 (zebrafish) (Table 1). Hence, the mean domain number is 20.5% (cod) to 26.5% (tetraodon) greater for retained duplicates that for singleton genes; the difference is significant in each of the nine teleost genomes (two-tailed Wilcoxon rank-sum tests, $P < 3.0 \times 10^{-13}$ for each species; Fig. 1b). The distribution of functional domain number shows that difference in domain number does not result from outliers (Fig. 2b). Similar results were obtained when comparisons were performed with different datasets, as outlined above for protein length comparisons.

Discussion

The most salient finding of this study is that duplicated genes with longer protein length and more functional domains are preferentially retained after whole-genome duplication in teleost fish. This finding is consistent with the analyses in yeast which concluded that the higher the complexity of a gene, the higher its probability of retention after duplication (He and Zhang 2005). Together, these studies suggest that the complexity of a gene, or its encoded protein product, is a determinant of retention probability after duplication across widely divergent organisms.

There are two distinct implications of this conclusion. First, the finding is relevant for understanding the evolutionary mechanisms at play after gene or genome duplication. It is logical to suppose that one of a pair of genes would be soon lost after duplication except under particular circumstances, such as dosage advantage, subfunctionalization, or neofunctionalization. The preferential retention of duplicated genes encoding more complex proteins is compatible with subfunctionalization being a prevalent force, since functional divergence could occur more rapidly in more complex genes after duplication (He and Zhang 2005). Second, the finding is relevant to discussions concerning the relation between genomic and organismal complexity (He and Zhang 2005; Yang et al. 2003). The evolutionary significance of gene duplication has long been

conjectured (Ohno 1970; Stephens 1951), but the mechanisms by which duplicate genes contribute to phenotypic evolution or diversification are not fully resolved. The whole-genome duplication event shared by all teleost fish preceded teleost diversification, and hence it has been suggested this event contributed to the great diversification of teleosts (Amores et al. 1998; Taylor et al. 2003), either through increased genomic complexity and/or by contributing to reproductive isolation through differential loss of duplicated genes (Lynch and Conery 2000; Semon and Wolfe 2007). This proposed relation between FSGD and diversification has been challenged on the grounds that much of teleost diversity is found in clades that radiated long after the genome duplication, and because palaeontological evidence does not show a rapid increase in disparity post-FSGD (Clarke et al. 2016); however, these criticisms are weakened by the finding that gene divergence can be delayed until long after genome duplication (Macqueen and Johnston 2014; Martin and Holland 2014, 2017). The preferential retention of complex genes after the FSGD, reported here, is particularly relevant to these discussions, since it indicates that eventual loss of

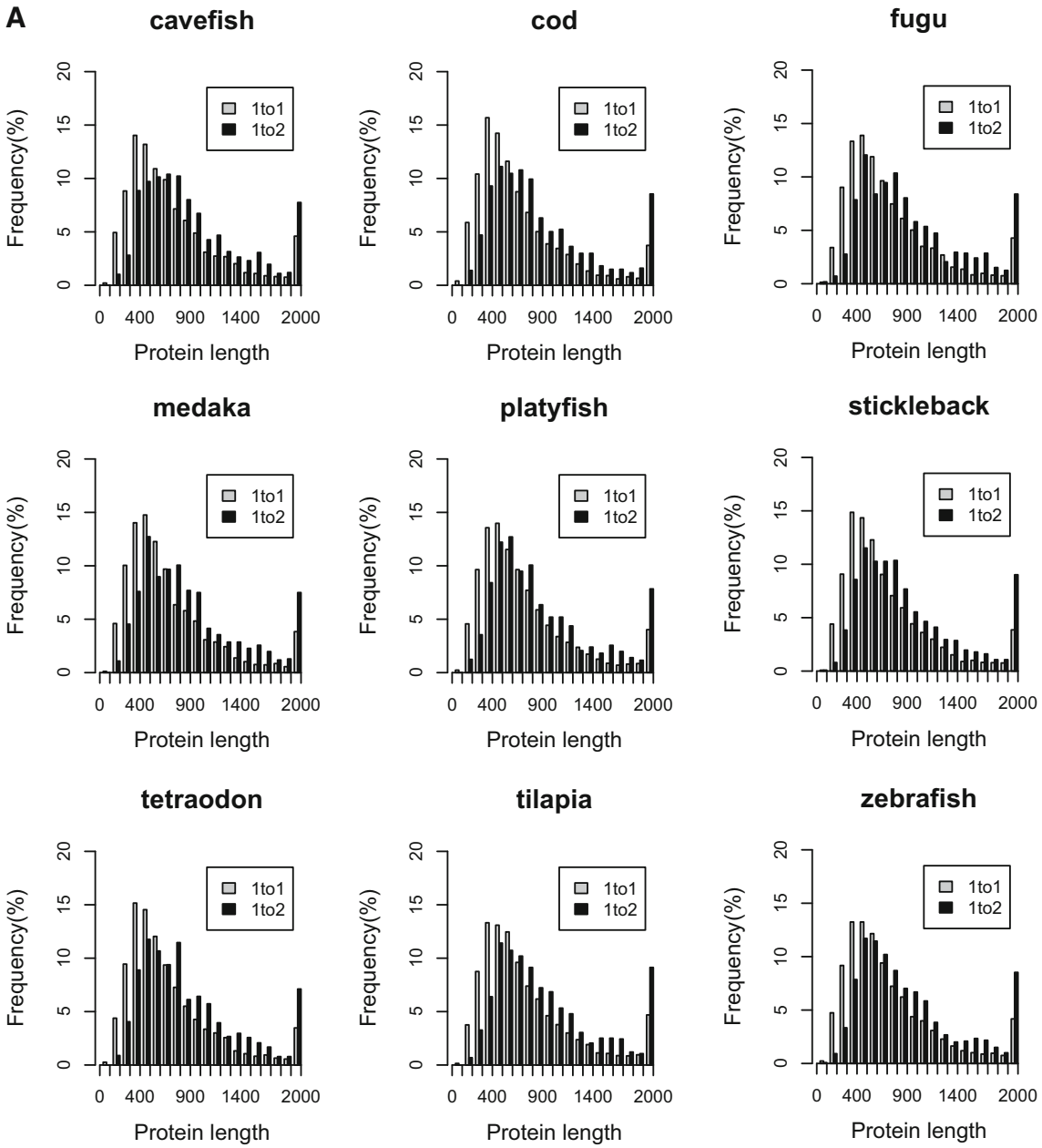

Fig. 2 Distribution of **a** protein length in amino acids and **b** functional domain number between retained duplicate and singleton genes in nine teleost genomes. *1to1* singleton genes, *1to2* retained duplicate genes

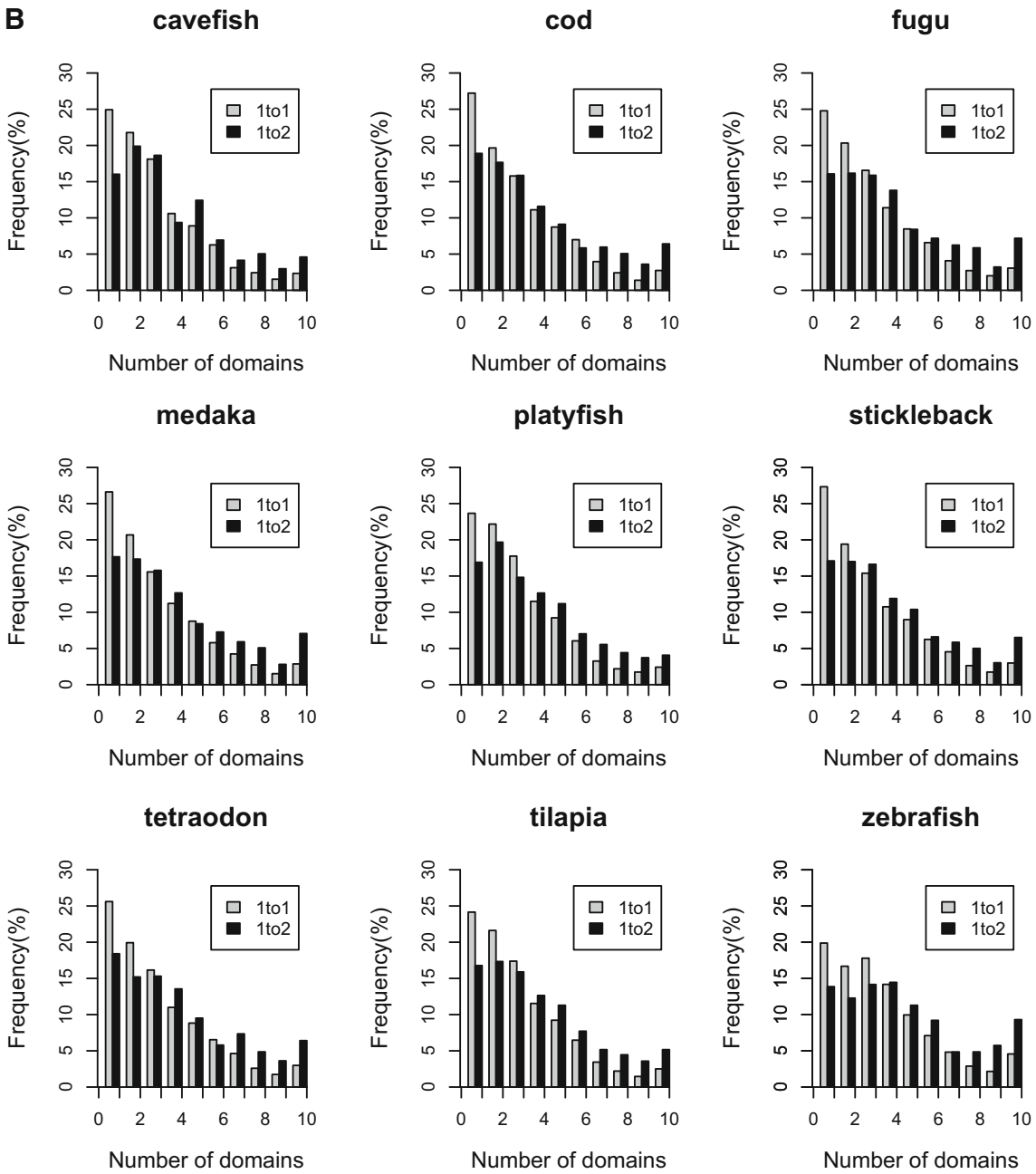

Fig. 2 continued

duplicate genes was not random. The preferential retention of duplicate genes encoding more complex proteins is consistent with a model in which the FSGD generated additional genomic complexity which may have been exploited during the successful diversification of teleost fish.

Acknowledgements This work was supported by CAS Pioneer Hundred Talents Program and the National Natural Science Foundation of China (31672273). I thank Prof. Peter Holland from University of Oxford for useful discussions and language revision.

Compliance with Ethical Standards

Conflict of interest Author declares that he has no conflict of interest.

References

Amores A, Force A, Yan YL, Joly L, Amemiya C, Fritz A, Ho RK, Langeland J, Prince V, Wang YL, Westerfield M, Ekker M,

Postlethwait JH (1998) Zebrafish hox clusters and vertebrate genome evolution. Science 282:1711–1714
Chain FJJ, Dushoff J, Evans BJ (2011) The odds of duplicate gene persistence after polyploidization. BMC Genom 12:599. doi:10.1186/1471-2164-12-599
Clarke JT, Lloyd GT, Friedman M (2016) Little evidence for enhanced phenotypic evolution in early teleosts relative to their living fossil sister group. Proc Natl Acad Sci USA 113:11531–11536
Guo B, Wagner A, He S (2011) Duplicated gene evolution following whole-genome duplication in teleost Fish. In: Friedberg F (ed) Gene duplication. InTech, Rijeka, pp 27–36
Guo B, Zou M, Wagner A (2012) Pervasive indels and their evolutionary dynamics after the fish-specific genome duplication. Mol Biol Evol 29:3005–3022
He X, Zhang J (2005) Gene complexity and gene duplicability. Curr Biol 15:1016–1021
Inoue J, Sato Y, Sinclair R, Tsukamoto K, Nishida M (2015) Rapid genome reshaping by multiple-gene loss after whole-genome duplication in teleost fish suggested by mathematical modeling. Proc Natl Acad Sci USA 112:14918–14923
Kasprzyk A (2011) BioMart: driving a paradigm change in biological data management. Database J Biol Databases Curation 2011:bar049. doi:10.1093/database/bar049
Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155
Macqueen DJ, Johnston IA (2014) A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. Proc R Soc B: Biol Sci 281:20132881
Martin KJ, Holland PWH (2014) Enigmatic orthology relationships between hox clusters of the African butterfly fish and other teleosts following ancient whole-genome duplication. Mol Biol Evol 31:2592–2611
Martin KJ, Holland PW (2017) Diversification of hox gene clusters in osteoglossomorph fish in comparison to other teleosts and the spotted gar outgroup. J Exp Zool B: Mol Dev Evol. doi:10.1002/jez.b.22726
Ohno S (1970) Evolution by gene duplication. Springer, New York
Semon M, Wolfe KH (2007) Reciprocal gene loss between Tetraodon and zebrafish after whole genome duplication in their ancestor. Trends Genet 23:108–112
Stephens SG (1951) Possible significance of duplication in evolution. Adv Genet Inc Mol Genet Med 4:247–265
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
Yang J, Lusk R, Li WH (2003) Organismal complexity, protein complexity, and gene duplicability. Proc Natl Acad Sci USA 100:15661–15665
Zhang J (2003) Evolution by gene duplication: an update. Trends Ecol Evol 18:292–298