Review Article

Cytogenetic and Genome Research

Cytogenet Genome Res 2015;145:243–252
DOI: 10.1159/000431108

Published online: June 9, 2015

Molecular Polymorphism and Divergence of Duplicated Genes in Tetraploid African Clawed Frogs *(Xenopus)*

Ben J. Evans[a] Taejoon Kwon[b]

[a]Department of Biology, McMaster University, Hamilton, Ont., Canada; [b]Department of Molecular Biosciences, Institute for Cellular and Molecular Biology, Center for Systems and Synthetic Biology, University of Texas, Austin, Tex., USA

Key Words

Dosage balance hypothesis · Gene duplication · Genome duplication · Mutation rate · Purifying selection

Abstract

Genome duplication creates redundancy in proteins and their interaction networks, and subsequent smaller-scale gene duplication can further amplify genetic redundancy. Mutations then lead to the loss, maintenance or functional divergence of duplicated genes. Genome duplication occurred many times in African clawed frogs (genus *Xenopus*), and almost all extant species in this group evolved from a polyploid ancestor. To better understand the nature of selective constraints in a polyploid genome, we examined molecular polymorphism and divergence of duplicates and single-copy genes in 2 tetraploid African clawed frog species, *Xenopus laevis* and *X. victorianus*. We found that molecular polymorphism in the coding regions of putative duplicated genes was higher than in singletons, but not significantly so. Our findings also suggest that transcriptome evolution in polyploids is influenced by variation in the genome-wide mutation rate, and do not reject the hypothesis that gene dosage balance is also important.

Over evolutionary time, gene duplication leads either to loss of one copy, continued expression of both copies with each being functionally interchangeable, or continued expression of both copies with functional divergence from one another as a consequence of changes in regulation or amino acid sequences. The genomic consequences of gene duplication by genome duplication versus by smaller-scale duplication differ in that the former scenario does not alter the relative dosages of interacting proteins, whereas the latter does [Lynch and Conery, 2000]. A key question in studies of genome duplication asks what factors contribute to the preservation of two functional duplicates versus the alternative scenario of one becoming a (nonfunctional) pseudogene while the other persists as a functional 'singleton'. Because mutations are generally deleterious, one copy of a pair of duplicated genes is expected to eventually become a pseudogene unless natural selection favors the persistence of both [Ohno, 1970]. It is possible, therefore, that the overall rate of mutation is of primary importance in governing the probability of pseudogenization, and that pseudogenization is more likely in rapidly evolving regions of the genome, irrespective of gene function or interactions with other proteins. The rates at which mutations that either improve or degrade protein function, respectively, govern the probabilities of subfunctionalization and neofunctionalization, and thus influence the probabilities that duplicate genes, are preserved by these mechanisms [Force et al., 1999; Lynch and Conery, 2000; Lynch et al., 2001; Lynch, 2007]. Alternatively or in addition, 'gene dosage balance' – that is, natural selection to maintain the relative stoichiometry of interacting proteins [Papp et al., 2003] – may influence the fate of genes duplicated by

KARGER 125

1424–8581/15/1454–0243$39.50/0

E-Mail karger@karger.com
www.karger.com/cgr

Ben J. Evans
Department of Biology, McMaster University
1280 Main Street West
Hamilton, ON L8S 4K1 (Canada)
E-Mail evansb@mcmaster.ca

whole-genome duplication (polyploidization). This could be evidenced, for example, by overrepresentation of retained duplicates of proteins that form stable protein complexes compared to those that do not [Qian and Zhang, 2008]. Natural selection favoring increased expression [Seoighe and Wolfe, 1999; Kondrashov et al., 2002; Kondrashov and Kondrashov, 2006; Conant and Wolfe, 2008], or functional buffering against deleterious mutations [Gu et al., 2003; Chapman et al., 2006], could also promote the retained function of both duplicate copies. Furthermore, patterns of duplicate gene loss and persistence may differ between polyploid genomes generated by autopolyploidization and allopolyploidization; for example, gene loss may occur in a biased fashion with respect to subgenomes [Comai, 2005; Evans, 2007]. Together, these issues are relevant to our understanding of diverse lineages that experienced genome duplication [Van de Peer et al., 2009], including lineages that are ancestral or relatively closely related to humans.

African clawed frogs offer fascinating subjects for the study of genome duplication. For the sake of consistency with most scientific literature, here, we consider all African clawed frogs to be members of the genus *Xenopus*. This clade comprises 2 groups [Kobel et al., 1996] that differ in their ancestral chromosome number, with the 'Silurana' group being descended from a diploid ancestor with 20 chromosomes and the 'Xenopus' group being descended from a diploid ancestor with 18 chromosomes (fig. 1). Only one known extant species of African clawed frog is diploid – *X. tropicalis*; this species is a member of the 'Silurana' group and has 20 chromosomes. Another widely studied African clawed frog, *X. laevis*, has 36 chromosomes and is a member of the 'Xenopus' group. *X. laevis* is considered to be 'pseudotetraploid' because it has hallmarks of tetraploidization (such as 9 sets of 4 morphologically similar chromosomes), yet it undergoes cell division as if it were a diploid, with each chromosome aligning with only one partner [Tymowska, 1991]. Complete genome sequences are available for *X. tropicalis* [Hellsten et al., 2010], and a draft genome sequence has been released for *X. laevis* [the *Xenopus laevis* Genome Project Consortium; available at Xenbase, Bowes et al., 2009]. Several other polyploid species have been identified in *Xenopus* including other tetraploid species with 36 or 40 chromosomes, octoploid species with 72 chromosomes, and dodecaploid species with 108 chromosomes [reviewed in Evans, 2008]. The mechanism of genome duplication is demonstrably via allopolyploidization for the tetraploids with 40 chromosomes, and for the octoploid and dodecaploid species, based on molecular evolutionary relationships with respect to lower-ploidy level species [Evans et al., 2005; Evans, 2007, 2008].

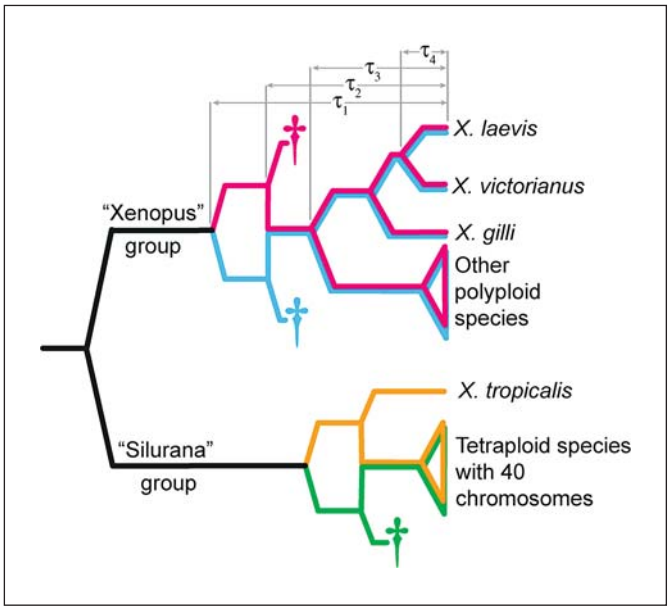

Fig. 1. Phylogenetic relationships among representative *Xenopus* species and lineages in this study. Merged evolutionary lineages represent genome duplication by allopolyploidization, and daggers indicate hypothetical diploid ancestors for whom an extant diploid descendant species is unknown. Divergence times are labeled for the diploid ancestors of tetraploids in the Xenopus group (τ_1), the putative allotetraploidization event in the Xenopus group (τ_2), the divergence of extant tetraploid species (τ_3), and the divergence of *X. laevis* and *X. victorianus* (τ_4).

In order to better understand the evolutionary forces that sculpt duplicate gene evolution in polyploid species, we examined molecular divergence and polymorphism of duplicates and singletons in the tetraploid African clawed frog species *X. laevis* and a closely related tetraploid species called *X. victorianus* that also has 36 chromosomes. Divergence of *X. victorianus* and *X. laevis* (indicated by τ_4 in fig. 1) occurred ~10 Mya based on multilocus data from the autosomes [Furman et al., 2014] and mitochondrial DNA [Evans et al., 2004]. An estimate of the time the tetraploid ancestor of these species originated (τ_2 in fig. 1) is unavailable because an extant descendant(s) of the ancestral diploid species is/are unknown (represented by the red and blue daggers in fig. 1). However, a divergence time estimate between other more distantly related tetraploid species that descended from this ancestor (τ_3 in fig. 1) [Evans et al., 2004; Chain and Evans, 2006] suggests that considerably more than the first half of the tetraploid evolutionary history of *X. victorianus* and *X. laevis* oc-

curred in their most recent common ancestor prior to their divergence as separate species (in other words that $\tau_3 > 2\tau_4$, and therefore that $\tau_2 >> 2\tau_4$; fig. 1).

If a genomic region is not subject to natural selection, it is said to evolve 'neutrally'. Under this circumstance, divergence between species and molecular polymorphism within a species are correlated: genomic regions with a high mutation rate are more diverged from an outgroup and also more polymorphic among individuals within a species than are genomic regions with a low mutation rate [Kimura, 1983]. However, natural selection can alter this correlation. A polymorphism-reducing effect of natural selection is seen, for example, if mutations in a genomic region were recently beneficial [selective sweep; Smith and Haigh, 1974] or deleterious [background selection; Charlesworth et al., 1993], or if a genomic region was not itself the target of natural selection, but genetically linked to another region that was [genetic hitchhiking; Smith and Haigh, 1974]. For these reasons, information on molecular divergence and polymorphism can offer insight into selective constraints on different genomic regions [Hudson et al., 1987], such as those that are present in one or two copies in a polyploid genome.

We used information from the UniGene database for *X. laevis* and the Ensembl database for *X. tropicalis* to identify duplicates – most of which are presumably homeologs, that is, paralogs generated by whole-genome duplication – and singletons in *X. laevis* and to evaluate levels of divergence of each of these gene categories. We then used a reduced representation genome sequencing approach called RADseq [Baird et al., 2008] to quantify molecular polymorphism in *X. victorianus*, and used information from *X. laevis* to identify putative duplicates and singletons in *X. victorianus* and to control for variation in the genome-wide rate of evolution.

We had 2 expectations related to divergence. First, we hypothesized that homeologs located in rapidly evolving genomic regions would undergo pseudogenization more rapidly or more frequently than those in slowly evolving regions. We expected this because pseudogenization of one homeolog is caused by mutations that occur after duplication. Thus, for the tetraploid *X. laevis*, we expected singletons – those genes whose extra copy underwent pseudogenization – to be more diverged than duplicates from their respective orthologous sequences in the outgroup species *X. tropicalis*. In a comparison between orthologous sequences in *X. victorianus* and *X. laevis*, we also expected putative singletons to be more diverged than putative duplicates. We note that these expectations assume that the rate of evolution remains similar in both homeologous regions after duplication as it was before duplication.

Furthermore, as a consequence of genetic redundancy, we also expected duplicated genes to be subject to a reduced level of purifying selection relative to singleton genes. This expectation is derived from the finding that purifying selection on duplicate genes in *Xenopus* is relaxed compared to a singleton ortholog in *X. tropicalis* based on the rate ratio of nonsynonymous to synonymous substitutions per site [Chain and Evans, 2006; Morin et al., 2006; Hellsten et al., 2007]. We therefore predicted that, after controlling for variation in mutation rate, duplicate genes would exhibit higher molecular polymorphism than singletons. As discussed above, the gene dosage balance hypothesis for duplicate gene retention posits that duplicated genes are retained by natural selection because of the important roles they play in duplicated regulatory networks [Papp et al., 2003]. Under this scenario, genes that become singletons presumably were less subject to natural selection favoring gene dosage balance of both homeologs. However, after one homeolog becomes a pseudogene, the remaining singleton would presumably still be subject to strong purifying selection to perform the ancestral function. Under this scenario, duplicates (which are under selection for gene dosage balance) and singletons might evolve, on average, under similar evolutionary constraints. In some duplicates, purifying selection might even be more extreme than some singletons, for example, if they are hubs in protein interaction networks [Hahn and Kern, 2005]. Overall, a similar level of molecular polymorphism in duplicates and singletons after controlling for variation in the mutation rate could be consistent with the dosage balance hypothesis.

An advantage of assaying polymorphism in *X. victorianus* is that sequences from *X. laevis* can be used to quantify variation in the mutation rate among genes (based on variation among genes in divergence). A drawback is that some of the putatively functional duplicates and singletons identified in *X. laevis* do not correspond respectively to functional duplicates and singletons in *X. victorianus* as a result of independent evolution that occurred in each species after their divergence. In other words, some functional duplicates in the most recent common ancestor of *X. laevis* and *X. victorianus* might still be functional duplicates in *X. laevis*, but became singletons in *X. victorianus* or vice versa. However, there is some reason to suspect that the rate of pseudogenization may be higher immediately following genome duplication as compared to later on [Maere et al., 2005; Scannell and Wolfe, 2008], which would make the patterns of pseudogenization more simi-

lar between *X. laevis* and *X. victorianus* than expected under a constant rate of pseudogenization. Furthermore, the proportion of retained duplicated genes is relatively high in *X. laevis*. Previous estimates of this proportion are ~30–50% [Hellsten et al., 2007; Sémon and Wolfe, 2008; Chain et al., 2011], and these may be underestimates due to missing data from expressed duplicates. Additionally, our previous work on duplicates and singletons from *X. laevis* and *X. gilli*, a more distantly related species pair, indicates that the duplicate/singleton status of many genes is conserved [Bewick et al., 2011; Furman et al., 2014]. As such, we anticipate that most duplicates in *X. laevis* are also duplicates in *X. victorianus* and likewise for singletons.

Methods

Identification of Duplicates and Singletons in X. laevis: Analysis of Divergence

We used the *X. laevis* UniGene Build #94 (Feb 16, 2013) and the *X. tropicalis* Ensembl (EnsEMBL v76) databases to identify functional duplicates and singletons in *X. laevis*. We selected the Ensembl release instead of the UniGene database for *X. tropicalis* because the Ensembl database had more annotations for coding regions. An Ensembl database for *X. laevis* is not currently available.

To identify putative functional duplicates and singletons in *X. laevis*, we used a modified reciprocal best BLAST [Altschul et al., 1997] hit approach using an expect value for each search of $1e^{-20}$. Functional duplicates in *X. laevis* sequences were defined as a pair of nucleotide sequences from the *X. laevis* UniGene database that both had as a top BLAST hit the same *X. tropicalis* nucleotide sequence in the Ensembl database, and where this *X. tropicalis* sequence also had as its top 2 best BLAST hits the same pair of *X. laevis* sequences. We imposed as an additional criterion that the alignment length identified from the *X. tropicalis* query be at least 200 bp long for both *X. laevis* sequences, although we did not constrain these alignments to overlap. We defined a singleton in *X. laevis* as a sequence that was the reciprocal best BLAST hit with a sequence from *X. tropicalis* with at least 200 bp of homologous sequence, and where other *X. laevis* sequences did not fulfill the above criteria for a functional duplicate. Some *X. laevis* sequences failed to meet the requirements for being a putative singleton or a duplicate and were discarded from analysis. This approach is conservative with respect to identification of functional duplicates because some putative singletons may actually be duplicates for which sufficient data were lacking from one homeolog or for which one homeolog was not identified due to divergence. For example, when no minimum homology length criterion was applied, there were 204 more pairs of functional duplicates identified.

We also calculated pairwise divergence between *X. laevis* duplicates. This calculation was restricted to the homologous regions identified by BLAST or, if more than one homologous region was identified, the region spanning multiple homologous regions. We imposed an additional criterion that this region be at least 200 bp long in order for the divergence to be calculated. Sequences were aligned with the program MAAFT, version 7.164b [Katoh and Standley, 2013] using the 'adjustdirectionaccurately' and 'genafpairtriads' options. In order to explore variation in the genome-wide rate of evolution of putative singletons and duplicates in *X. laevis*, we quantified divergence from the orthologous *X. tropicalis* sequences. A permutation test (described below) was used to test for a difference between the divergences of putative singletons and duplicates in *X. laevis*. We additionally performed these analyses for the entirety of the open reading frame (ORF) and the untranslated regions (UTRs) identified as described below.

Reduced Representation Genome Sequencing, Read Mapping and Genotyping

We used a reduced representation genome sequencing approach called restriction enzyme associated DNA [RADseq; Baird et al., 2008] to obtain sequences from homologous genomic regions of 8 samples of *X. victorianus*. We then inferred genotypes and levels of molecular polymorphism at genes identified as singletons or duplicates in *X. laevis* as discussed above. RADseq libraries were prepared by Floragenix (Portland, Oreg., USA) using the restriction enzyme *Sbf*I, and multiplexed, single-end sequencing was performed on one Illumina flow cell lane. These data have been deposited in the NCBI short read archive (accession number SRP050568). Seven of the *X. victorianus* samples (Field IDs: BJE00263-7, Museum IDs: MCZ A-138180-4, and Field IDs: BJE01488-9, no museum ID) were collected in the town of Lwiro, Democratic Republic of the Congo (GPS coordinates: –2.2455, 28.81325), and one sample (field ID: BJE00261, Museum ID: MCZ A-138178) was collected about 45 km south in Bukavu, Democratic Republic of the Congo (GPS coordinates: –2.5028, 28.87554).

We used the software package Stacks version 1.21 [Catchen et al., 2011] to preprocess these reads before alignment. This included de-multiplexing the 8 individuals based on barcodes and postprocessing of sequence reads after alignment. The Stacks program 'process_radtags' was used to stringently filter the data by (1) removing reads that did not have a correctly called barcode, (2) trimming the barcode and truncating the final read length to 75 bp, (3) removing reads with an uncalled base, (4) discarding reads in which the average sequence quality score within any sliding window of 11 bp was less than a Phred-scaled value of 10 (which corresponds to a 90% base call accuracy). The sequences of the adaptors used to make the RADseq Illumina libraries are not released by Floragenix, but (5) we also removed reads with adapter sequences matching ′GAT CGG AAG AGC GGT TCA GCA GGA ATG CCG AGA CCG ATC TCG TAT GCC GTC TTC TGC TTG′ and ′AGA TCG GAA GAG CGT CGT GTA GGG AAA GAG TGT AGA TCT CGG TGG TCG CCG TAT CAT T′ or sequences that were divergent from these by up to 2 nucleotides. Additionally, (6) we discarded reads that were marked by Illumina's chastity/purity filter as failing.

We used the 2014-10-09 release of the program GSNAP [Wu and Watanabe, 2005; Wu and Nacu, 2010] to map the remaining high-quality reads to release 94 of the UniGene *X. laevis* database. For GSNAP, (a) we mapped reads with only one path (alignment) and suppressed reads with more than one path, (b) allowed up to 5 mismatches for mapped reads, and (c) set an insertion/deletion penalty of 2. We then used samtools version 0.1.19 [Li et al., 2009] to convert '.sam'-formatted files from GSNAP to '.bam' format. We used the 'pstacks' and 'cstacks' programs from Stacks to collate aligned reads from each individual, with 'pstacks' reporting, for each individual, only those genomic regions with a minimum depth of 3 reads (these 3 reads forming what is called a 'stack'), and with 'cstacks'-clustering reads based on genomic position. In 'pstacks',

we conservatively allowed no more than 2 mismatches between reads that were combined to form a 'stack', and in 'cstacks', we conservatively allowed no more than 2 mismatches between stacks (i.e. 2.7% of the 75-bp read) from different samples in order for them to be combined into the catalog of alleles. We selected these low values in hopes of reducing the chances that sequences from homeologous genomic regions would be combined into a single genotype. We infer that these values are conservative based on an analysis of divergence between expressed homeologs of *X. laevis* detailed above. A drawback of these conservative settings is that we may underestimate polymorphism at some genes because some diverged alleles would be excluded from analysis. For this reason, we also explored how alternative settings of 1, 3 or 4 mismatches for both programs affected the results. We used the program 'sstacks' to match data across the 8 samples to the catalog generated by 'pstacks'.

We used a 'bounded' model for genotype calling [Catchen et al., 2013] with an upper bound for the error rate set to 0.05, with χ^2 significance level to compare the likelihoods of a homozygous versus heterozygous genotype call of 0.05. A lower upper bound of the error rate has a higher chance that a heterozygous genotype would be called at a polymorphic position, as opposed to inferring a homozygous genotype in which a sequencing error occurred. To evaluate whether the error rate setting affected our inferences, we also used, for each of the 4 mismatch parameter values detailed above, settings of 0.01 or 0.001, and a χ^2 significance level of 0.01.

For analyses with each combination of parameter settings, the 'rxstacks' program was used to make corrections to genotype and haplotype calls based on information from all individuals, as detailed in the Stacks manual. For 'rxstacks', any catalog entry was removed that had 2 or more individuals (25%) with a confounded match where multiple loci map to the same region. We also removed haplotypes in excess of the biologically expected maximum of 2, and we re-called SNPs after removing sequencing errors using a bounded SNP model with the same settings as described earlier for 'pstacks'. We additionally set a lower limit for the mean natural logarithm of the likelihood of each catalog locus at –8. Following this, 'pstacks', 'cstacks' and 'sstacks' were rerun on the corrected data to infer polymorphism statistics.

Using scripts written in Perl, we then calculated molecular polymorphism and divergence statistics for *X. victorianus* orthologs of the putative singletons and duplicates of *X. laevis* identified as described above. These statistics included pairwise nucleotide diversity (π) [Tajima, 1983] and Watterson's [1975] θ, and percent sequence divergence between the *X. laevis* reference sequence and a randomly selected allele from the 8 *X. victorianus* individuals. The pairwise divergences were then corrected for multiple substitutions using the method of Jukes and Cantor [1969] and corrected for ancestral polymorphism by subtracting π [Charlesworth and Charlesworth, 2010, pp 258–259], and the resulting corrected divergence (d_{JC_AP}) was used to standardize π. For these statistics, 95% CIs were estimated by bootstrapping by site as described in Evans et al. [2014]. In order to test for a difference in the molecular polymorphism of putative singletons and duplicates, departure of the null hypothesis that the ratio of π/d_{JC_AP} for singletons:duplicates is equal to 1 was assessed with a z-test as described previously [Evans and Charlesworth, 2013; Evans et al., 2014].

Identification of ORFs and UTRs in X. laevis UniGenes

In order to predict coding regions, the *X. laevis* UniGene sequences were translated to all-6-frame protein sequences and then mapped to reference proteomes of human, mouse, chicken, zebrafish, and *X. tropicalis* (EnsEMBL v72). We considered an ORF to be validated if it uniquely mapped to the ORFs of 5 reference species. For situations where, for a given nucleotide sequence, more than one possible reading frame matched the reference proteomes, we developed 2 rules to potentially validate one of them. The first rule is that if the number of reference genomes that matched a particular frame was greater than twice the number of reference genomes that matched an alternative frame, the former frame was considered validated. If this criterion was not met, the second rule was that if the length of the matched frame was twice as long as the length of the second best matching frame, then the former frame was considered validated.

Results

Putative Singletons Are More Diverged from an Ortholog than Duplicates

We used BLAST [Altschul et al., 1997] to identify putative duplicates and singletons in the tetraploid species *X. laevis* based on comparisons to sequences from the closely related diploid species *X. tropicalis*. We identified 11,150 *X. tropicalis* sequences with 1 or 2 orthologs in *X. laevis*. About 34% of these *X. tropicalis* sequences were inferred to be orthologous to expressed duplicates in *X. laevis* (3,757 pairs of functional duplicates; 7,514 *X. laevis* genes in total) and the remainder were inferred to be orthologous to a singleton in *X. laevis* (7,393 putative *X. laevis* singletons). Our estimated proportion of functional duplicates matches other recent estimates [Hellsten et al., 2007; Sémon and Wolfe, 2008; Chain et al., 2011], but we suspect that it is an underestimate as a consequence of missing data and the challenge of identifying highly diverged duplicates.

We calculated the average pairwise divergence between each homeolog for *X. laevis* and the *X. tropicalis* ortholog and for *X. laevis* singletons, the pairwise divergence to the *X. tropicalis* ortholog. Across duplicates and singletons, the average of these divergences was 10.2% (n = 11,150 comparisons). When the average divergence was calculated after weighting by the alignment lengths, it was 10.0%. When these gene categories were considered separately, *X. laevis* singletons were slightly more diverged from *X. tropicalis* (10.37%) than *X. laevis* duplicates (9.96%). We also calculated the average pairwise divergence between BLAST-identified homeologous regions within *X. laevis* duplicates with at least 200 bp of overlapping sequence. The average divergence between *X. laevis* homeologs was 7.2% (weighted average was 7.1%, n = 3,334 comparisons).

Table 1. Divergence between *X. victorianus* and *X. laevis* and *X. victorianus* polymorphism statistics for combined ORFs and UTRs, OFSs only, and UTRs only

	Duplicates							Singletons						
	genes	bp	*S*	π	θ_W	d_{JC_AP}	π/d_{JC_AP}	genes	bp	*S*	π	θ_W	d_{JC_AP}	π/d_{JC_AP}
ORFS and UTRs	149	12,128	17 (10–26)	0.00039 (0.00020–0.00063)	0.00042 (0.00025–0.00065)	0.01780 (0.01553–0.02015)	0.02188 (0.01123–0.03609)	151	11,722	33 (22–44)	0.00078 (0.00049–0.00109)	0.00085 (0.00057–0.00113)	0.02244 (0.01973–0.02541)	0.03485 (0.02175–0.04998)
ORFs only	101	7,616	12 (6–19)	0.00051 (0.00022–0.00086)	0.00047 (0.00024–0.00075)	0.01380 (0.01119–0.01662)	0.03710 (0.01545–0.06457)	107	7,687	15 (8–24)	0.00042 (0.00020–0.00068)	0.00059 (0.00031–0.00094)	0.02243 (0.01904–0.02616)	0.01865 (0.00878–0.03107)
UTRs only*	55	4,081	5 (1–10)	0.00020 (0.00003–0.00043)	0.00037 (0.00007–0.00074)	0.02420 (0.01968–0.02973)	0.00835 (0.00151–0.01834)	45	3,470	15 (8–23)	0.00150 (0.00074–0.00237)	0.00130 (0.00069–0.00200)	0.2132 (0.01640–0.02677)	0.07028 (0.03469–0.12274)

ORFs and UTRs of putative duplicates and singletons using settings 2, 0.05, and 0.05 for mismatch, the upper bound of the error parameter, and alpha level for χ^2 significance for Stacks (see Methods). Other information includes the number of genes (genes), sites genotyped in all individuals (bp), segregating sites (*S*), the pairwise nucleotide diversity (π), Watterson's theta (θ_W), divergence with Jukes-Cantor correction for multiple substitutions and correction for ancestral polymorphism (d_{JC_AP}; see Methods), and the standardized diversity, π/d_{JC_AP}. 95% CIs from bootstrapping are in parentheses. An asterisk indicates that the standardized diversity estimate of the UTRs of singletons is significantly higher than that of the UTRs of duplicates according to a z-test ($p < 0.05$). For this combination of Stacks parameters, the standardized diversity in the ORFs of singletons was significantly lower than that of the UTRs of singletons. In contrast, the standardized diversity in the ORFs of duplicates was significantly higher than that of the UTRs of duplicates.

In order to test whether this disparity in average outgroup divergence of *X. laevis* duplicates and singletons was greater than expected by chance, we compared the observed difference to permutated differences calculated by randomly shuffling the observed outgroup divergence among the duplicates and singletons. The observed difference was small (0.41%), but significantly higher than the random expectation ($p < 0.001$). This indicates that singletons in *X. laevis* are more diverged than duplicates from their respective orthologs in *X. tropicalis* than expected by chance.

To explore whether these results were consistent when the entirety of the protein coding and UTRs were considered separately, we used stringent criteria (see Methods) to identify putative ORFs in the *X. laevis* UniGene database. This exercise resulted in annotations for 16,224 of the 31,306 *X. laevis* UniGenes (51.8%). Thus, about half of these sequences were not annotated because the homology of the *X. laevis* reading frames did not pass our stringent criteria. The average pairwise divergence of the ORFs of *X. laevis* singletons and duplicates was 10.2 and 9.3%, respectively, and that for the UTRs was 21.5 and 19.8%, respectively. Consistent with the results discussed above based only on the homologous regions identified by BLAST, permutation tests indicate that divergence of *X. laevis* duplicates in each of these transcribed regions was significantly lower than that for singletons. Divergence between *X. laevis* homeologs was 7.3 and 16.1% in the ORFs and the untranscribed regions, respectively.

The genetic divergence between orthologs of *X. victorianus* and *X. laevis* was ~2% (table 1; online suppl. table 1; see www.karger.com/doi/10.1159/000431108 for all suppl. material), which is ~30% of the inferred level of divergence between the BLAST identified *X. laevis* homeologous regions described above. Similar to the comparisons between *X. laevis* and *X. tropicalis*, when sequences from the combined UTRs and the ORFs (including genes with ambiguous ORFs) were compared, we observed higher divergence between orthologs of *X. laevis* and *X. victorianus* for putative singletons than for putative duplicates for all combinations of parameter settings used in Stacks (table 1; online suppl. table 1). This was also the case when the comparison was restricted to only the ORFs (table 1; online suppl. table 2). However, the opposite was observed when the comparison between *X. laevis* and *X. victorianus* was restricted to only the UTRs (table 1; online suppl. table 3).

Molecular Polymorphism of Putative Duplicates in X. victorianus Is Not Significantly Different from Singletons

After read filtering (see Methods), an average of ~2.5 million reads were obtained for each of the 8 *X. victorianus* individuals (range 1.7–2.9 million) of which 5.6% mapped to the *X. laevis* UniGene database (range 5.3–5.9%) for an average of 142,076 mapped reads per individual (range 90,881–161,965). Average coverage across individuals was 29 reads per base (range 20.0–32.2). Indi-

vidual BJE261 had the lowest coverage, whereas the coverage for the other 7 individuals was similar.

Our results indicated that the standardized molecular polymorphism (π/d_{JC_AP}) of putative singleton genes was higher than that of putative duplicates in *X. victorianus* when ORFs and UTRs are analyzed together (table 1; online suppl. table 1). However, this difference was not significant irrespective of the settings we used in Stacks ($p > 0.05$, z-tests). This nonsignificant result was unexpected under the hypothesis that duplicates are subject to relaxed purifying selection relative to singleton genes [Ohno, 1970; Kimura and Ohta, 1974].

We therefore quantified molecular variation separately in the ORFs and in the UTRs of genes whose ORFs could be unambiguously identified using conservative criteria detailed above. When only the ORFs were considered, the standardized molecular polymorphism was higher in putative duplicates than in putative singletons (table 1; online suppl. table 2), but again this difference was not statistically significant ($p > 0.05$, z-test). Similar to the analysis of the entirety of the sequence, the standardized molecular polymorphism was higher in the UTRs of singletons. This difference was significant in several of the combinations of parameters for Stacks that we attempted (table 1; online suppl. table 3).

We also compared the standardized molecular polymorphism in the UTR to the ORFs for both singletons and duplicates. For singletons, the UTRs were more diverse than ORFs for all Stacks parameter settings, and this difference was significant for about half of them ($p > 0.05$, z-test, data not shown). Unexpectedly, however, the standardized molecular polymorphism was lower in the UTR than the ORFs for the putative duplicates, and this difference was significant for several Stacks parameter combinations ($p < 0.05$, z-test, data not shown). The lower (but not significantly different) level of polymorphism across the entirety of the transcript in putative duplicates compared to putative singletons thus appears to be due to a dearth of observed polymorphism in the UTRs of putative duplicates.

Permutation tests discussed above indicate that singletons evolve more quickly than duplicates in *X. laevis*, and singletons, therefore, are expected to be more polymorphic by virtue of their higher mutation rate. For this reason, we controlled for differences in the mutation rate using corrected divergence from *X. laevis* (see Methods), and therefore, our results cannot be solely explained by variation in mutation rates among these gene categories.

In this study, it is possible that *X. victorianus* sequences from pseudogenes or from an expressed gene that was missing from the UniGene database might have inappropriately mapped to a homeologous sequence that was present in the UniGene database. Mismapped reads would be expected to increase polymorphism, and this might inflate polymorphism estimates from putative singletons to a greater degree than duplicates, depending on the extent of missing data. We explored this possibility in multiple ways. As detailed above, we used a range of values for the mismatch parameter, which controls the allowable level of allelic divergence. We also disallowed information from genomic regions with more than the biological expectation of 2 diverged alleles, a scenario that would be expected with mismapped reads. In general, similar results were recovered across combinations of Stacks settings (online suppl. tables 1–3).

Another indication of potentially mismapped reads is substantially higher coverage (approx. by a factor of 2) compared to loci without mismapped reads. To explore this possibility, we tested for genes that contained sites that were either polymorphic or diverged from the *X. laevis* reference sequence, whether there was a correlation between the number of heterozygous genotypes and the average coverage per individual. In putative singletons, there was no significant correlation between coverage and the number of heterozygous sites ($p > 0.05$ for all settings attempted in Stacks, ANOVA test), which would be expected if highly polymorphic loci were polymorphic due to mismapped reads.

Although peripheral to the question of whether polymorphism in singletons is augmented by mismapped homeologous reads, we also tested whether this correlation was present in duplicates. Mismapping of reads could inflate polymorphism in duplicates if one homeolog was partially sequenced while the other was completely sequenced, and reads mismapped to the more completely sequenced homeolog. Polymorphism in duplicates could also be inflated by mismapped reads from segmental duplicates that originated after polyploidization. Consistent with these possibilities, in duplicates we found a significant positive correlation between the number of heterozygous sites and the average coverage per individual at some Stacks settings (mismatch penalty equal to 1, 2 or 3, and error rate and χ^2 significance level equal to 0.001 and 0.01, respectively, or a mismatch penalty equal to 4, and error rate and χ^2 significance level equal to 0.01 and 0.01, respectively). These significant correlations in duplicates were not strong, but could reflect mismapping of reads in some duplicates.

Another possible indication of mismapped reads is sites that are inferred to be heterozygous in all individuals as a result of divergence between homeologs. We there-

fore evaluated polymorphism after excluding positions that were heterozygous in all individuals from the analysis. After excluding these sites (which comprised a small proportion of the variable positions due to the conservative genotype filter we applied, see Methods), levels of polymorphism were essentially identical to the analyses presented in table 1 and the online supplementary tables.

Overall, these results are inconsistent with the possibility that molecular polymorphism of singletons was inappropriately overestimated due to mismapped reads from homeologs or pseudogenes. An additional concern that we do not address here is that the *X. laevis* UniGene database is incomplete and that many putative *X. laevis* singletons are in fact functional duplicates for which data is missing for one homeolog. To explore this possibility, these analyses could be repeated on more comprehensive compilations of unique *X. laevis* transcripts, or on the draft assembly of the *X. laevis* genome (the *Xenopus laevis* Genome Project Consortium), both of which will soon be available for genomic analyses.

Discussion

Divergence and Pseudogenization

It is possible that variation among genomic regions in the mutation rate could influence the probability of pseudogenization, with pseudogenization occurring more frequently in rapidly evolving regions. Consistent with this possibility, divergence between orthologs of *X. laevis* and *X. tropicalis* was slightly higher for *X. laevis* singletons than for duplicates, and a permutation test suggested that this difference was significant. A qualitatively similar pattern was observed in comparisons between orthologs of putative singletons and duplicates in *X. laevis* and *X. victorianus* when data from the ORFs and UTRs were combined and when the ORFs were analyzed separately, although not when the UTRs were analyzed separately. An alternative explanation is that other factors, such as natural selection, act to slightly accelerate divergence of singletons after pseudogenization of a homeolog or to decelerate divergence of functional duplicates relative to singletons. It is also possible that the divergence estimates of duplicates and singletons were influenced by misclassification of genes. For example, gene conversion or missing data could have caused some duplicates to be categorized as singletons; the influence of misclassification in our analysis is unclear.

Previous studies suggest that slowly evolving genes tend to be preferentially retained as duplicates [Davis and Petrov, 2004; Chain et al., 2011], or undergo subfunctionalization [Sémon and Wolfe, 2008], and our permutation tests are consistent with these studies. If variation in rates of evolution of the silenced homeologous partner of *X. laevis* singletons matches that of the singleton, these results suggest that pseudogenization is more likely in rapidly evolving portions of the genome. Studies of yeast homeologs in which one copy was lost suggest that the lost copy tends to be the faster evolving of the 2 homeologs [Byrne and Wolfe, 2007], which also suggests an important role of mutation rate variation in the probability of pseudogenization. An alternative explanation is that singletons evolve more quickly after a homeologous gene becomes pseudogenized, although this latter explanation lacks a biological rationale.

Molecular Polymorphism on Putative Singletons and Duplicates

There exists variation among genomic regions in the rate of mutation and in the nature of natural selection, both of which influence molecular polymorphism. After controlling for genome-wide variation in the mutation rate, the inferred levels of molecular polymorphism in putative singletons and duplicates of *X. victorianus* were not significantly different, with polymorphism over the entire transcripts being higher for putative singletons, but within the ORF being higher for putative duplicates. There are at least 2 explanations for these results. The first is that we lacked statistical power to detect a significant difference between putative singletons and duplicates. To further contextualize this possibility, we performed a power analysis in which we fixed the observed levels of polymorphism for the ORFs of singletons and duplicates and scaled the amount of data until a significant difference was detected. This analysis indicated that ~2.1 times as much data would be required in order to detect a significant difference between the observed levels of polymorphism in the ORFs for the Stacks parameter settings presented in table 1.

A second explanation is that functional duplicates and singletons do in fact evolve under similar evolutionary constraints in this relatively 'mature' polyploid genome (>>20 million generations old, assuming a generation time of 1 year). As discussed earlier, purifying selection on duplicates could be of similar magnitude to that on singletons if natural selection maintains dosage balance of interacting proteins encoded by duplicated genes [Papp et al., 2003; Qian and Zhang, 2008]. A weakness of this study is that the prediction for the dosage balance hypothesis (that there is no difference in the level of purifying selection on singletons and duplicates) is a null

hypothesis that we failed to reject. Additional information on protein interaction networks would permit a more comprehensive assessment of whether proteins with many interactions tend to be overrepresented as duplicates [Edger and Pires, 2009], as is the case in the polyploid plant *Arabidopsis thaliana* [Bekaert et al., 2011]. Another explanation for these results is that a similar level of purifying selection on duplicates and singletons exists as a consequence of functional divergence of the homeologous pairs. Moreover, Chain et al. [2011] did not recover a significant difference between singleton and duplicate genes in *X. laevis* in another index of purifying selection when the effect was evaluated jointly with other information on the pattern of expression and information content of genes. In that study, the strength of purifying selection was measured in terms of the rate ratio of nonsynonymous to synonymous substitutions per site in orthologs of *X. tropicalis* in order to estimate these rates without the influence of whole-genome duplication. Due to the limited data and associated dearth of statistical power, here, we did not separately analyze synonymous and nonsynonymous polymorphisms.

While we cannot reject the prediction associated with the dosage balance hypothesis, the (not significantly) higher level of polymorphism in ORFs of duplicates compared to singletons matches the proposal that duplicates are subject to relaxed purifying selection based on estimates of the rate ratio of nonsynonymous to synonymous substitution [Chain and Evans, 2006; Morin et al., 2006; Hellsten et al., 2007]. However, the paucity of polymorphism in the UTR of duplicates compared to the ORF of these genes was unexpected. This could reflect a polymorphism-reducing influence of purifying selection on the UTR of duplicates, a region that can be involved in the regulation of translation [Araujo et al., 2012]. Alternatively, it could be an unidentified technical artifact related to mismapped reads, sequencing errors or the sequence filtering. Future efforts that leverage the complete genome sequence from *X. laevis* and more complete sequence information from expressed transcripts undoubtedly will contribute statistical power to help address these possibilities.

Inferences made herein, of course, do not extend through time to earlier stages of polyploid genome evolution, when patterns of pseudogenization and selective constraints on singletons and duplicates may have been quite different. Importantly, the inference based on molecular polymorphism comes with the caveat that the genes from *X. victorianus* were categorized as putative duplicates and singletons based on information from another diverged species *(X. laevis)*. This was necessary because we lack comprehensive information from the *X. victorianus* transcriptome across developmental stages and tissue types (this information is available for *X. laevis*). The degree to which this caveat would affect our conclusions depends on (a) the rate of pseudogenization and the variation in this rate since tetraploidization, (b) the proportion of tetraploid evolution that occurred prior to versus after the speciation of *X. laevis* and *X. victorianus*, and (c) the degree to which missing information from the *X. laevis* UniGene database led us to incorrectly characterize *X. laevis* duplicates as singletons. A compelling direction for further exploration would leverage these data from *X. victorianus* to standardize diversity values from a natural population of *X. laevis* from Western Cape Province, South Africa, which is the probable source of many laboratory strains from which *X. laevis* sequence databases were generated. Another concern not addressed here is the possibility that gene conversion occurs between homeologs, and that this could lead to misinterpretations about their divergence, duplicate versus singleton status and other evolutionary characteristics [Katju and Bergthorsson, 2010]. Future work will also shed light on the possibility that more than one tetraploidization event generated extant *Xenopus* tetraploids with 36 chromosomes [Bewick et al., 2011], a scenario that would potentially have consequences for the duration of tetraploid ancestry prior to the speciation of *X. laevis* and *X. victorianus*.

Acknowledgements

B.E. thanks Frédéric Chain, Adam Bewick, and Ben Furman for numerous enlightening discussions about *Xenopus* genomics, and one reviewer for constructive comments on an earlier version of the manuscript.

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402 (1997).

Araujo PR, Yoon K, Ko D, Smith AD, Qiao M, et al: Before it gets started: regulating translation at the 5′ UTR. Comp Funct Genomics 2012: 475731 (2012).

Baird NS, Etter PD, Atwood TS, Currey MC, Schriver AL, et al: Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3:e3376 (2008).

Bekaert M, Edger PP, Pires JC, Conant GC: Two-phase resolution of polyploidy in the *Arabidopsis* metabolic network gives rise to relative and absolute dosage constraints. Plant Cell 23:1719–1728 (2011).

Bewick AJ, Anderson DW, Evans BJ: Evolution of the closely related, sex-related genes DM-W and *DMRT1* in African clawed frogs *(Xenopus)*. Evolution 65:698–712 (2011).
Bowes JB, Snyder KA, Segerdell E, Gibb R, Jarabek CJ, et al: Xenbase: a *Xenopus* biology and genomics resource. Nucleic Acids Res 36:D761–772 (2009).
Byrne KP, Wolfe KH: Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication. Genetics 175: 1341–1350 (2007).
Catchen JM, Amores A, Hohenlohe PA, Cresko W, Postlethwait JH: Stacks: building and genotyping loci de novo from short-read sequences. G3 (Bethesda) 1:171–182 (2011).
Catchen JM, Hohenlohe PA, Bassham S, Amores A, Cresko WA: Stacks: an analysis tool set for population genomics. Mol Ecol 22:3124–3140 (2013).
Chain FJJ, Evans BJ: Multiple mechanisms promote the retained expression of gene duplicates in the tetraploid frog *Xenopus laevis*. PLoS Genet 2:e56 (2006).
Chain FJJ, Dushoff J, Evans BJ: The odds of duplicate gene persistence after polyploidization. BMC Genomics 12:599 (2011).
Chapman BA, Bowers JE, Feltus FA, Paterson AH: Buffering of crucial functional by paleologous duplicated genes may contribute cyclicality to angiosperm genome duplication. Proc Natl Acad Sci USA 103:2730–2735 (2006).
Charlesworth B, Charlesworth D: Elements of Evolutionary Genetics. (Roberts & Co., Greenwood Village 2010).
Charlesworth B, Morgan MT, Charlesworth D: The effect of deleterious mutations on neutral molecular variation. Genetics 134:1289–1303 (1993).
Comai L: The advantages and disadvantages of being polyploid. Nat Rev Genet 6:836–846 (2005).
Conant GC, Wolfe KH: Turning a hobby into a job: how duplicated genes find new functions. Nat Rev Genet 9:938–950 (2008).
Davis JC, Petrov DA: Preferential duplication of conserved proteins in eukaryotic genomes. PLoS Biol 2:E55 (2004).
Edger P, Pires JC: Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. Chromosome Res 17:699–717 (2009).
Evans BJ: Ancestry influences the fate of duplicated genes millions of years after duplication in allopolyploid clawed frogs *(Xenopus)*. Genetics 176:1119–1130 (2007).
Evans BJ: Genome evolution and speciation genetics of allopolyploid clawed frogs *(Xenopus* and *Silurana)*. Front Biosci 13:4687–4706 (2008).
Evans BJ, Charlesworth B: The effect of nonindependent mate pairing on the effective population size. Genetics 193:545–556 (2013).
Evans BJ, Kelley DB, Tinsley RC, Melnick DJ, Cannatella DC: A mitochondrial DNA phylogeny of African clawed frogs: phylogeography and implications for polyploid evolution. Mol Phylogenet Evol 33:197–213 (2004).
Evans BJ, Kelley DB, Melnick DJ, Cannatella DC: Evolution of RAG-1 in polyploid clawed frogs. Mol Biol Evol 22:1193–1207 (2005).
Evans BJ, Zeng K, Esselstyn JA, Charlesworth B, Melnick DJ: Reduced representation genome sequencing suggests low diversity on the sex chromosomes of Tonkean macaque monkeys. Mol Biol Evol 31:2425–2440 (2014).
Force A, Lynch M, Pickett B, Amores A, Yan YL, Postlethwait JH: Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545 (1999).
Furman BLS, Bewick AJ, Harrison TL, Greenbaum E, Gvoždík V, et al: Pan-African phylogeography of a model organism, the African clawed frog *Xenopus laevis*. Mol Ecol 24:909–925 (2014).
Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li WH: Role of duplicate genes in genetic robustness against null mutations. Nature 421: 63–66 (2003).
Hahn MW, Kern AD: Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol Biol Evol 22:803–806 (2005).
Hellsten U, Khokha MK, Grammer TC, Harland RM, Richardson P, Rokhsar DS: Accelerated gene evolution and subfunctionalization in the pseudotetraploid frog *Xenopus laevis*. BMC Biol 5:31 (2007).
Hellsten U, Harland RM, Gilchrist MJ, Hendrix D, Jurka J, et al: The genome of the Western clawed frog *Xenopus tropicalis*. Science 328: 633–636 (2010).
Hudson RR, Kreitman M, Aguadé M: A test of neutral molecular evolution based on nucleotide data. Genetics 116:153–159 (1987).
Jukes TH, Cantor CR: Evolution of protein molecules, in Munro HN (ed): Mammalian Protein Metabolism, pp 21–132 (Academic Press, New York 1969).
Katju V, Bergthorsson U: Genomic and population-level effects of gene conversion in *Caenorhabditis* paralogs. Genes (Basel) 1:452–468 (2010).
Katoh K, Standley SM: MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780 (2013).
Kimura M: The neutral theory of molecular evolution, in Nei M, Koehn R (eds): Evolution of Genes and Proteins, pp 208–233 (Sinauer, Sunderland 1983).
Kimura M, Ohta T: On some principles governing molecular evolution. Proc Natl Acad Sci USA 71:2848–2852 (1974).
Kobel HR, Loumont C, Tinsley RC: The extant species, in Tinsley RC, Kobel HR (eds): The Biology of *Xenopus*, pp 9–33 (Clarendon Press, Oxford 1996).
Kondrashov FA, Kondrashov AS: Role of selection in fixation of gene duplicates. J Theor Biol 239:141–151 (2006).
Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV: Selection in the evolution of gene duplications. Genome Biol 3:RESEARCH0008 (2002).
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al: The sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25: 2078–2079 (2009).
Lynch M: The Origins of Genome Architecture. (Sinauer, Sunderland 2007).
Lynch M, Conery JS: The evolutionary fate and consequences of duplicate genes. Science 290: 1151–1155 (2000).
Lynch M, O'Hely M, Walsh B, Force A: The probability of preservation of a newly arisen gene duplicate. Genetics 159:1789–1804 (2001).
Maere S, DeBodt S, Raes J, Casneuf T, Van Montagu MCE, et al: Modeling gene and genome duplication in eukaryotes. Proc Natl Acad Sci USA 102:5454–5459 (2005).
Morin RD, Chang E, Petrescu A, Liao N, Griffith M, et al: Sequencing and analysis of 10,967 full-length cDNA clones from *Xenopus laevis* and *Xenopus tropicalis* reveals post-tetraploidization transcriptome remodeling. Genome Res 16:796–803 (2006).
Ohno S: Evolution by Gene Duplication. (Springer, Berlin 1970).
Papp B, Pál C, Hurst LD: Dosage sensitivity and the evolution of gene families in yeast. Nature 424:194–197 (2003).
Qian W, Zhang J: Gene dosage and gene duplicability. Genetics 179:2319–2324 (2008).
Scannell DR, Wolfe KH: A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. Genome Res 18:137–147 (2008).
Sémon M, Wolfe KH: Preferential subfunctionalization of slow-evolving genes after allopolyploidization in *Xenopus laevis*. Proc Natl Acad Sci USA 105:8333–8338 (2008).
Seoighe C, Wolfe KH: Yeast genome evolution in the post-genome era. Curr Opin Microbiol 2: 548–554 (1999).
Smith JM, Haigh J: The hitch-hiking effect of a favourable gene. Genet Res 23:23–25 (1974).
Tajima F: Evolutionary relationship of DNA sequences in finite populations. Genetics 105: 437–460 (1983).
Tymowska J: Polyploidy and cytogenetic variation in frogs of the genus *Xenopus*, in Green DS, Sessions SK (eds): Amphibian Cytogenetics and Evolution, pp 259–297 (Academic Press, San Diego 1991).
Van de Peer Y, Maere S, Meyer A: The evolutionary significance of ancient genome duplications. Nat Rev Genet 10:725–732 (2009).
Watterson GA: On the number of segregating sites in genetical models without recombination. Theor Popul Bio 7:256–276 (1975).
Wu TD, Nacu S: Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26:873–881 (2010).
Wu TD, Watanabe CK: GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21:1859–1875 (2005).