Supporting Information

Sémon and Wolfe 10.1073/pnas.0708705105

SI Methods

EST Clustering. The aim of expressed sequence tag (EST) clustering is to group together sequences that are all transcripts of the same gene. The *X. laevis* WGD is relatively recent, so the sequences of the paralogs it created are still very similar (Modal nucleotide identity is 90%) as seen in Fig. S1. We used stringent criteria for clustering to ensure that we did not merge paralogous ESTs. We downloaded 547,704 *X. laevis* ESTs from dbEST and trimmed them to remove vector sequences. Repeats were masked by using RepeatMasker [Smit AFA, Hubley R, Green P (2004) *RepeatMasker Open-3.0*; http://www.repeatmasker.org]. Clustering was performed with TGICL (1), using complete mRNA and Refseq (2) sequences as seeds (10799 complete mRNAs for *X. laevis*). The first step consisted of a transitive clustering of pairs of sequences having ≥98% identity over at least 80 bp in a MEGABLAST alignment (3). Then, TGICL was used to make a multiple alignment between the sequences comprising each cluster and an assembly, using CAP3 (4). The same operation was performed in *S. tropicalis*, using 1,026,920 ESTs, 10,615 complete mRNAs, and 9,020 Refseq sequences. Coding regions in these contigs were predicted with ESTscan (5), trained with the *S. tropicalis* Refseq sequences.

Some of these predicted coding regions are very similar and most likely correspond to alternative splicing variants. To group the alternative transcripts of the same gene, we clustered the coding regions, using very stringent parameters (≥ 98% identity, over ≥100 bp or ≥80% of the smallest sequence's length), and then retained one sequence randomly from each of these sets. This procedure yielded 28,463 coding sequences for *X. laevis* and 28,860 for *S. tropicalis*.

Building Triplets of Homologous Genes. We searched for genes in *S. tropicalis* that have two coorthologs in *X. laevis*. We used TGICL to group the predicted protein sequences into gene families, by transitive clustering of pairs of genes with ≥60% protein identity over 70% of the sequence. We aligned the protein sequences in each family by using T-Coffee (6) and removed poorly aligned parts with Gblocks (7). The resulting alignments were back-translated into nucleotides and the corresponding trees were built by using PHYML (8). We then parsed the trees to retain 1,300 triplets where *S. tropicalis* was an outgroup to two *X. laevis* sequences. Our dataset of triplets is smaller than that recently used by Hellsten *et al.* (9) to study rate asymmetry in duplicated frog genes, because we did not use data from the unpublished *S. tropicalis* genome sequencing project.

ESTscan (5) was developed originally to predict coding sequences in individual ESTs. Because ESTs do not necessarily correspond to the sequence of full-length transcripts, ESTscan does not put much emphasis on predicting the translation start of genes. In some frog triplets, the coding sequence was predicted correctly to begin with a methionine codon in two sequences, but began a few nucleotides upstream in the last one. We considered this to be a misprediction if the third sequence coded for a methionine that aligned opposite the start codons of the other two. In that case, we trimmed the sequence so that all three coding regions begin with the common methionine.

Our triplets were identified based on phylogenetic analysis of gene families: we therefore ensure that the two paralogous copies in *X. laevis* are more similar to each other than either of the two *X. laevis*-*S. tropicalis* pairs from the same triplet. In vertebrates, synonymous substitutions are under weak constraint and *dS* values primarily reflect the age of the split between the sequences. Therefore, if the paralogous copies were created by WGD, we should observe that the ratio of the levels of synonymous substitution $dS_{(Xl1,\ Xl2)}/dS_{(St,\ Xl)}$ corresponds approximately to the ratio between the dates of the WGD and the speciation. We obtained a mean of 0.63 for the ratio of the values of *dS* (median: 0.66), which agrees with published estimates of the ratio of dates (0.50–0.67; refs. 10, 11).

Determination of Orthology Relationships. To annotate the orthologs of the triplets in human and in zebrafish, we searched our frog sequences against human sequences from HOMOLENS (a database of homologous genes collated from Ensembl; Simon Penel and Laurent Duret, personal communication; http://pbil.univ-lyon1.fr/databases/HOMOLENS.html), using BLASTP (12). We retained the association between a frog triplet and a human sequence if the best matching human sequence was the same for all three sequences and if it was strong enough (*E* value $< 10^{-10}$) and specific enough (score of the second best hit <90% of the score of the best hit). This first filter associated 1,105 frog triplets to a unique HOMOLENS family.

We then aligned the protein sequences of the triplets with the sequences of the corresponding HOMOLENS family, using ClustalW (13) and Gblocks. A phylogenetic tree was drawn by using PHYML if the resulting alignment was ≥100 aa. We parsed the trees to retain topologies that corresponded to the species tree, that is with fishes being the outgroup to a clade composed of two monophyletic groups, the frog and the mammalian sequences. If the tree contained these three monophyletic groups but they did not branch in the expected order (for instance if frog and fish were grouped to the exclusion of human) we performed an SH test [Shimodaira-Hasegawa test, implemented in TREE-PUZZLE (14)] to check whether this topology was significantly more likely than the species tree. If not, we retained the assignment of the triplet to the corresponding family. By this method we associate 644 frog triplets with HOMOLENS families, containing at least one human sequence and one fish sequence. Among them, we obtained one or two orthologs in zebrafish for 529 triplets, and one ortholog in human for 570 triplets (after removing any family where a duplication occurred in the human lineage after the split between human and frog).

Alignment of the Triplets and Rates of Sequence Evolution. For each sequence triplet consisting of one gene (*St*) in *S. tropicalis* and its two coorthologs in *X. laevis* (*Xl1* and *Xl2*), we aligned the predicted proteins using T-Coffee, removed the gaps using Gblocks, and back-translated to obtain a codon alignment. These alignments were input to the program like-tri-test (15) to estimate branch-specific levels of nonsynonymous and synonymous divergence. We quantified the absolute level of asymmetry in nonsynonymous evolution between the duplicates in *X. laevis* as: $abs(dN1\text{-}dN2)/(dN1+dN2)$, where *dN1* and *dN2* are the nonsynonymous divergences on the *Xl1* and *Xl2* branches, respectively. Like-tri-test also allows the statistical significance of asymmetry to be estimated. For each pair of genes *Xl1* and *Xl2,* we tested whether a model where both paralogous copies are free to evolve at different nonsynonymous rates has a better fit than a null model where they are constrained to the same nonsynonymous rate (as in ref. 16). For this, we computed the likelihood of these two models and rejected the null model if twice the difference of the log-likelihood was >3.81.

For the comparisons between frog and human genes (Fig. 2)

we first identified human orthologs of the *S. tropicalis* genes in our triplets. After pairwise sequence alignment, using T-coffee (and Gblocks) as described above, we then used PAML (17) to compute *dN* and *dS* values between *St* and human, and for the corresponding *Xl1-Xl2* pair.

Estimating Expression Profiles in Zebrafish. We extracted 779,139 zebrafish ESTs from dbEST (March 2006) and classified the 125 libraries into 14 tissues (embryo, heart, eye, gill, olfactory, testis, digestive tract, brain, liver, skin, ovary, muscle, fin, kidney). These ESTs were mapped by using MEGABLAST to the 22,866 zebrafish CDS from Ensembl (November 2005 version; ref. 18) present in HOMOLENS. Only hits with high similarity (E value $< 10^{-10}$) and high specificity of mapping (the score of the second best hit is <95% of the score of the best hit) were retained, to prevent misassignment of ESTs to paralogs created by the WGD in teleosts.

Estimating the Fraction of Genes Retained in Duplicate after WGD. Because the whole genome sequence is not available, the frequency of genes retained in duplicate after WGD in *X. laevis* is unknown. An optimistic hypothesis is that the set of 1,300 triplets we detected represents all of the genes where two copies have been retained since WGD. We detected 8,116 sets of homologous genes with one gene in *S. tropicalis* and at least one ortholog in *X. laevis*. The number of genes in *X. laevis* before duplication is likely to lie between this value and the number of genes observed in the human genome (22,000 in Ensembl version August 2006; ref. 18). This suggests a lower limit estimate that 6–16% (1,300/22,000 or 1,300/8,116) of the loci were retained in duplicate since WGD. However, the frequency of duplicate gene retention is certainly much higher: Because it consists of sequencing only a subset of the mRNAs produced in a subset of all possible physiological conditions, EST analysis will not detect every gene encoded by the frog genomes. This detection problem is less important for highly expressed genes, especially given the large size of our EST datasets: simulations have shown that in a dataset containing 500,000 ESTs, nearly all highly expressed genes (producing >100 ESTs per million ESTs) are detected (19).

Under the hypothesis that the expression level has not changed between the three genes in a triplet, the expression level measured in *S. tropicalis* should be correlated with the probability that all three members of the triplet are detected. In other words, the frequency of genes retained in duplicate in *X. laevis* is estimated more accurately among genes that are highly expressed in *S. tropicalis*. As expected, the observed frequency of retention of genes in two-copies in *X. laevis* increases from 10% to 35% with increasing expression of the *S. tropicalis* ortholog (Fig. S2). Because the frequency does not appear to reach a plateau (Fig. S2), we conclude that the sensitivity of gene detection is still increasing even for highly expressed genes in *S. tropicalis*. It therefore seems likely that even the 35% retention level we see in highly expressed ESTs is an underestimate.

We developed a method to estimate the true level of duplicate gene retention in the *X. laevis* genome. Our data consists of triplet and doublet gene sets: a triplet has one *S. tropicalis* and two coorthologous *X. laevis* sequences, and a doublet has one *S. tropicalis* and one *X. laevis* sequence. The retention frequency, R, of genomic loci in duplicate is given by $R = t_r/(t_r + d_r)$, where t_r and d_r are (respectively) the real numbers of triplet and doublet loci that exist between the *X. laevis* and *S. tropicalis* genomes. The problem is that, when we use EST data to classify loci as triplets or doublets, some genes that were actually retained in duplicate in the *X. laevis* genome will be incorrectly scored as doublets instead of triplets if one of the *X. laevis* copies was not represented in the ESTs sequenced. Thus, the observed retention frequency $R_o = t_o/(t_o + d_o)$, where t_o and d_o are the observed numbers of triplets and doublets respectively, is an underestimate of R.

The observed number of triplets (t_o) is smaller than the real number (t_r) so that:

$$t_o = t_r f^2 g \quad [1]$$

where f is the probability that a gene that exists in the *X. laevis* genome is detected in the *X. laevis* EST data, and g is the probability that a gene that exists in the *S. tropicalis* genome is detected in the *S. tropicalis* EST data.

The observed number of doublets (d_o) depends on the detection of real doublets (d_r) but also on the misinterpretation of triplets (t_r) for doublets:

$$d_o = g[d_r f + 2t_r f(1 - f)] \quad [2]$$

Equations [1] and [2] allow the true retention frequency R to be expressed simply as a function of R_o and f:

$$R = R_o/[f + (f - 1)\ R_o], \quad [3]$$

which is defined if $R < 1$; that is, if $f > 2t_o/(d_o + 2\ t_o)$.

This model is valid under the simplifying assumption that f is the same for all genes. We estimate R using datasets composed of only the most highly expressed genes, either the top 10% or the top 20% of genes by expression in *S. tropicalis* (those with >196 ESTs or >95 ESTs, respectively; Fig. S2). We can assume in this dataset that f is high (highly expressed genes are easier to detect) and homogeneous. If we assume that $f = 1$ (all genes were detected), this equation yields an estimate of $R = 0.32$–0.35 depending on which threshold EST count we use to define highly expressed genes (Fig. S3; curves $R_o = 0.32$ and $R_o = 0.35$). If we assume that 20% of real *X. laevis* genes were not detected as ESTs ($f = 0.8$), the estimate of R rises only slightly, to 0.43–0.47. To obtain a value of $R = 0.75$ as proposed by Hughes and Hughes (20), it is necessary to hypothesize that we have missed 40% of the genes ($f = 0.6$), which is unrealistic given that we base our computation on the most expressed genes. The value of $R = 0.47$ is likely to be an overestimate, because the computation is based on the frequency of double-copy retention in highly expressed genes, which, as we show in the main text, have a higher retention frequency than the rest of the genome after a WGD. We conclude that the true value of R for *X. laevis* is $\approx 0.40 \pm 0.07$.

Detection of Changes in Expression Profile. We test whether one gene copy in *X. laevis* shows a significant decrease in expression level in one tissue, whereas the other copy shows a significant decrease in a different tissue (Fig. 1*b*). We use a statistical test developed by Audic and Claverie (30) with a slight modification to correct for a bias due to the effects of gene loss after WGD.

To explain this bias let us consider a simplified system (Fig. S5). Suppose that *S. tropicalis* has only 10 genes, each transcribed into 10 mRNAs per cell. So there is a total of 100 mRNAs per cell, and each gene makes 10% of the transcripts. Suppose also that there are 15 genes in *X. laevis*, including five pairs of duplicates. Each of these genes is transcribed at 10 mRNAs per cell, so there are a total of 150 mRNAs per cell. A gene whose expression has not changed produces 10 transcripts per cell in both species, but this represents 10% of the cellular mRNA in *S. tropicalis* and only 6.6% of cellular mRNA in *X. laevis* (Fig. S5). If no other evolution of expression happened, we would therefore expect the counts of ESTs per million to be lower for *X. laevis* genes than for *S. tropicalis* genes. The combined mRNA output of a retained pair of genes in *X. laevis* will be 13.3% of cellular mRNA, but each of the *X. laevis* genes alone produces a lower fraction of cellular mRNA than its *S. tropicalis* ortholog. The null hypothesis is that the ratio of expression levels between *S. tropicalis* and individual *X. laevis* genes should be approxi-

mately equal to the excess of genes in *X. laevis* due to WGD (32–47% according to our estimations; see above).

We estimated the levels of expression in both species as the number of ESTs observed in the 11 tissues divided by the total number of ESTs sequenced in the 11 tissues (Fig. S6). Expression levels of individual *X. laevis* genes, measured as EST counts per million, are significantly lower than expression levels in *S. tropicalis*. Genes that are single-copy in both species are more likely to follow the null expectation (no evolution happened to the pattern of expression since speciation). For these genes we observe a median of expression level in *X. laevis* = $1.85 \times 10^{-4,}$ lower than in *S. tropicalis* (2.29×10^{-4}; $n = 1382$, Wilcoxon P value $< 10^{-16}$). Fig. S6*c* shows the distribution of the ratio of expression levels for genes that are single-copy in both species. The median of this distribution is -0.26, which corresponds to a ratio of $e^{0.26} = 1.30$ *S. tropicalis* transcripts per *X. laevis* transcript. In other words, we observe that expression level is ≈30% greater in *S. tropicalis* than in *X. laevis*, which is in reasonable agreement with our estimates of the level of duplicate gene retention after WGD in *X. laevis*.

We need to take this effect into account in our definition of "significantly changed" expression, because the null expectation (if no other evolution happened to the pattern of expression) is that the observed number of ESTs in *S. tropicalis* should be $e^{0.26}$ times the observed number of ESTs in *X. laevis.* We incorporated this new threshold in Audic and Claverie's test to detect significant decreases in expression level in *X. laevis*.

1. Pertea G, *et al.* (2003) *Bioinformatics* 19:651–652.
2. Pruitt KD, Tatusova T, Maglott DR (2005) *Nucleic Acids Res* 33:D501–D504.
3. Zhang Z, Schwartz S, Wagner L, Miller W (2000) *J Comput Biol* 7:203–214.
4. Huang X, Madan A (1999) *Genome Res* 9:868–877.
5. Lottaz C, Iseli C, Jongeneel CV, Bucher P (2003) *Bioinformatics* 19:ii103–ii112.
6. Notredame C, Higgins DG, Heringa J (2000) *J Mol Biol* 302:205–217.
7. Castresana J (2000) *Mol Biol Evol* 17:540–552.
8. Guindon S, Gascuel O (2003) *Syst Biol* 52:696–704.
9. Hellsten U, *et al.* (2007) *BMC Biol* 5:31.
10. Evans BJ, Kelley DB, Melnick DJ, Cannatella DC (2005) *Mol Biol Evol* 22:1193–1207.
11. Chain FJ, Evans BJ (2006) *PLoS Genet* 2:e56.
12. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) *J Mol Biol* 215:403–410.
13. Thompson JD, Higgins DG, Gibson TJ (1994) *Nucleic Acids Res* 22:4673–4480.
14. Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) *Bioinformatics* 18:502–504.
15. Conant GC, Wagner A (2003) *Genome Res* 13:2052–2058.
16. Cusack BP, Wolfe KH (2007) *Trends Genet* 23:270–272.
17. Yang Z (2007) *Mol Biol Evol* 24:1586–1591.
18. Birney E, *et al.* (2004) *Nucleic Acids Res* 32:D468–D470.
19. Reverter A, McWilliam SM, Barris W, Dalrymple BP (2005) *Bioinformatics* 21:80–89.
20. Hughes MK, Hughes AL (1993) *Mol Biol Evol* 10:1360–1369.

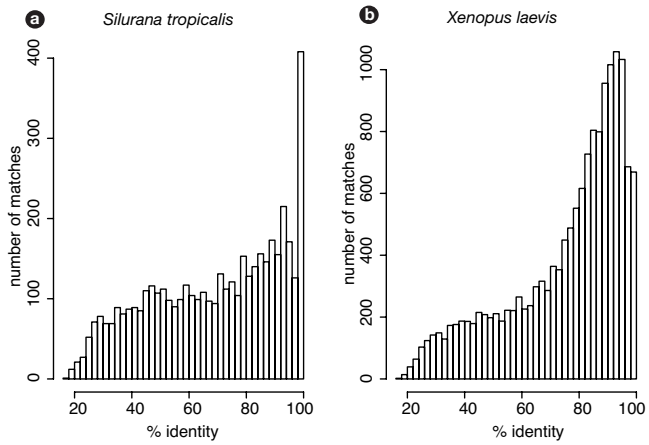

Fig. S1. The excess of recent duplicates in *X. laevis* by comparison to *S. tropicalis* is the hallmark of a recent WGD in *X. laevis.* We estimated the number of paralogs in each species by the number of pairs of coding sequences that align highly significantly (BLASTN *E* value $< 10^{-10}$; ref. 12), and for each species we show the relationship between the number of these matches and the percentage nucleotide identity, which can, to a first approximation, be considered as a proxy for the age of the duplicates. (*a*) The number of paralogs does not depend on nucleotide similarity in *S. tropicalis*, apart from a peak of very similar duplicates (>98% identity) that are probably attributable to alternative splicing variants that were separated during the assembly of the clusters. (*b*) The plot for *X. laevis* is very different, and the excess of duplicates centered on 90% DNA sequence identity is most likely due to WGD.

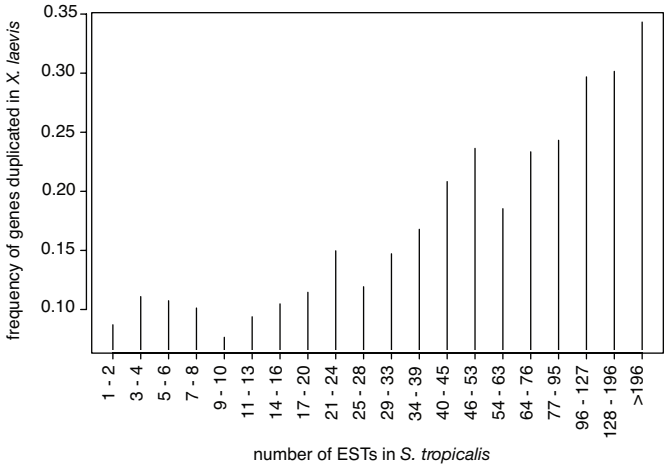

Fig. S2. The frequency of genes detected in two copies in *X. laevis* increases with the level of expression of their ortholog in *S. tropicalis*. The total set of 8,116 genes in *S. tropicalis* with at least one ortholog in *X. laevis* was divided into 20 bins of equal size according to expression level (number of ESTs) in *S. tropicalis*. The plot shows the frequency of genes retained in two copies in *X. laevis* for each of these 20 bins. The range of ESTs for each bin is indicated on the *x* axis.

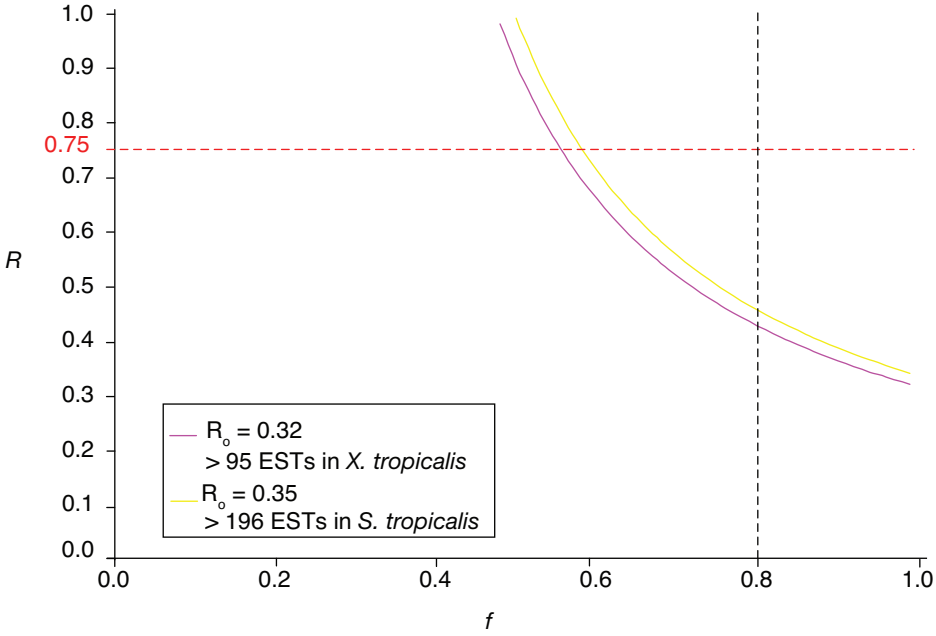

Fig. S3. Computation of the double-copy retention frequency (*R*) in *X. laevis* for different values of the probability that an extant gene is detected in the *X. laevis* EST data (*f*). The computation is based on the observed double-copy retention in the most highly expressed genes (pink for the 20% most highly expressed genes, yellow for the 10% most highly expressed genes). The red dashed line represents the double-copy retention estimated by Hughes and Hughes (20). The black dashed line shows the values of *R* obtained for $f = 0.8$, that is when 80% of the genes are detected.

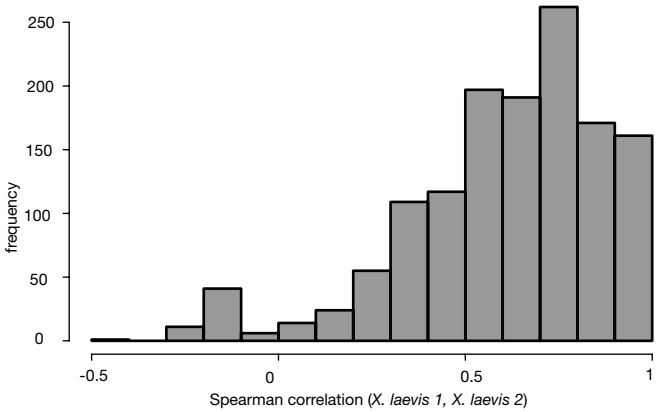

Fig. S4. Distribution of levels of conservation of expression patterns in 1,300 pairs of paralogous genes in *X. laevis* created by the WGD, measured as a Spearman correlation coefficient across 11 tissues.

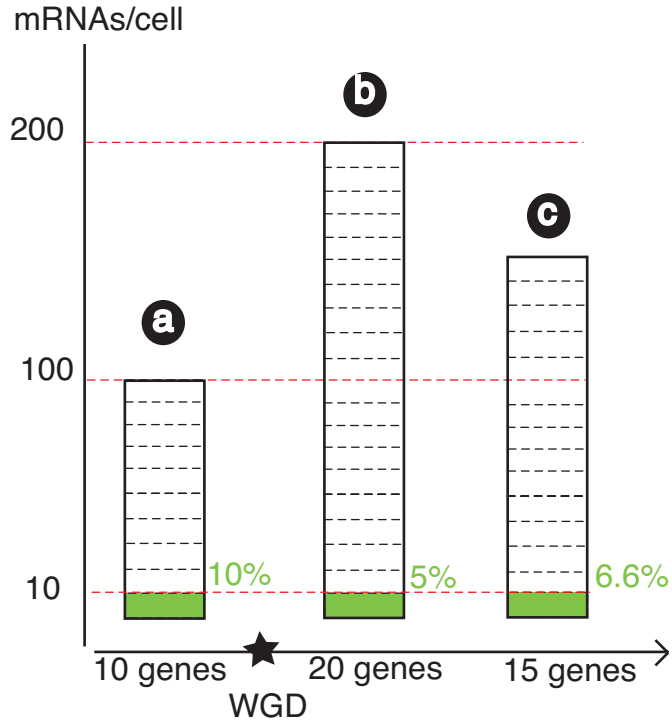

Fig. S5. Simplified system explaining why observed levels of expression should be lower in *X. laevis* than in *S. tropicalis*. (*a*) For illustration, we imagine that *S. tropicalis* has only 10 genes, each transcribed into 10 mRNAs per cell. (*b* and *c*) After WGD (*b*) and gene loss, five pairs of genes are retained in duplicate in *X. laevis* (*c*). Each of the 15 genes is transcribed at 10 mRNAs per cell. Any given gene produces 10 transcripts in both species but this represents 10% of the cellular mRNA in *S. tropicalis* and only 6.6% of cellular mRNA in *X. laevis*.

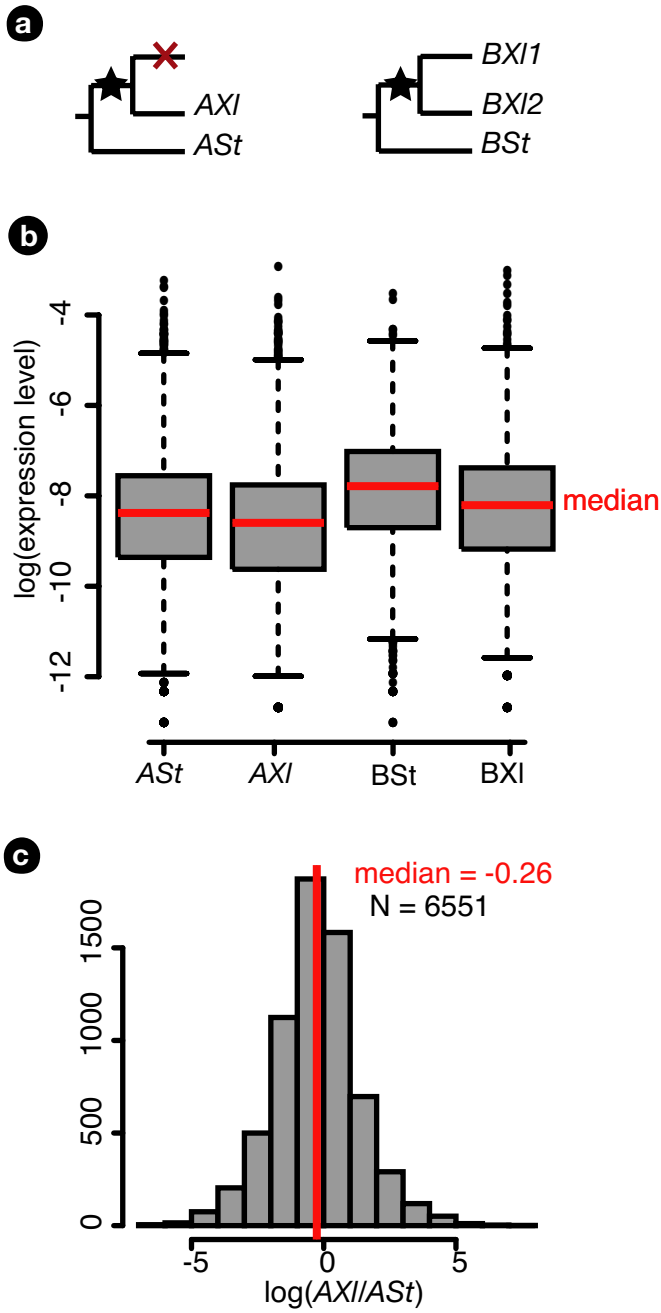

Fig. S6. The null expectation is that observed expression levels should be ≈30% greater in *S. tropicalis* (*St*) than in *X. laevis (Xl).* (*a*) We designate genes that are single-copy in both species as ''A,'' and genes that are members of a retained duplicate pair in *X. laevis* as ''B.'' (*b*) Box-plots of the observed expression levels in *S. tropicalis* and *X. laevis* for genes of types A and B, measured as the number of ESTs observed in the 11 tissues divided by the total number of ESTs sequenced in the 11 tissues. As expected, expression levels are significantly higher in *S. tropicalis* than in *X. laevis*, for genes of both types A and B. (*c*) Distribution of the ratio of expression levels in *X. laevis* and *S. tropicalis* for genes that are single-copy (type A) in both species.

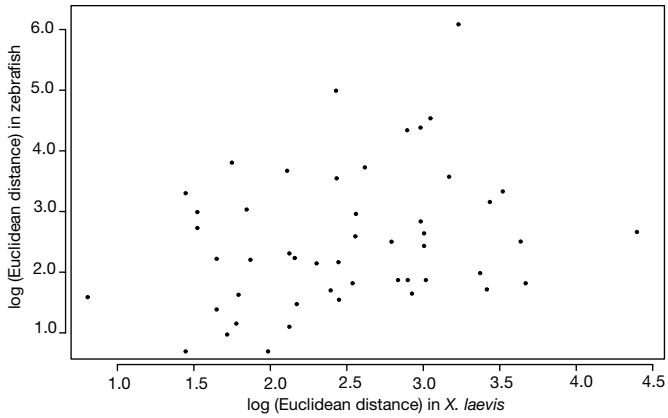

Fig. S7. Comparison of the expression divergences observed after WGD in zebrafish and after WGD in *X. laevis*. If orthologous pairs were retained for the same reasons after the two duplications, we should observe a correlation between the levels of within-pair expression divergence in the two species, because genes retained for dosage should have a low divergence in both cases and genes retained by subfunctionalization a higher divergence. For each of the 49 orthologous families that were retained in duplicate in both zebrafish and *X. laevis*, we measured the divergence of expression profiles between the two copies within each species, using Euclidian distances. The plot shows a moderate correlation between these Euclidean distances ($R = 0.28$; $P = 0.04$; $n = 49$). Note there is a possible bias in this analysis, because Euclidean distances and the total number of ESTs are correlated, and the level of expression is conserved across species, which may cause an indirect correlation between the Euclidean distances in different species. For instance, the number of EST in *S. tropicalis* is correlated with the Euclidian distance between the two copies in *X. laevis* ($R = 0.64$; $P < 10^{-5}$; $n = 49$) and the numbers of ESTs are correlated between orthologs in *X. laevis* and zebrafish ($R = 0.53$; $P < 10^{-5}$; $n = 49$). To correct for this bias, we verified that the Euclidian distances divided by the number of ESTs are still moderately correlated between zebrafish and *X. laevis* ($R = 0.28$; $P = 0.05$; $n = 49$).