Preferential subfunctionalization of slow-evolving genes after allopolyploidization in *Xenopus laevis*

Marie Sémon and Kenneth H. Wolfe*

Smurfit Institute of Genetics, Trinity College, Dublin 2, Ireland

Edited by Wen-Hsiung Li, University of Chicago, Chicago, IL, and approved April 11, 2008 (received for review September 13, 2007)

As paleopolyploid genomes evolve, the expression profiles of retained gene pairs are expected to diverge. To examine this divergence process on a large scale in a vertebrate system, we compare *Xenopus laevis*, which has retained ≈40% of loci in duplicate after a recent whole-genome duplication (WGD), with its unduplicated relative *Silurana (Xenopus) tropicalis*. This comparison of ingroup pairs to an outgroup allows the direction of change in expression profiles to be inferred for a set of 1,300 *X. laevis* pairs, relative to their single orthologs in *S. tropicalis*, across 11 tissues. We identify 68 pairs in which *X. laevis* is inferred to have undergone a significant reduction of expression in at least two tissues since WGD. Of these pairs, one-third show evidence of subfunctionalization, with decreases in the expression levels of different gene copies in two different tissues. Surprisingly, we find that genes with slow rates of evolution are particularly prone to subfunctionalization, even when the tendency for highly expressed genes to evolve slowly is controlled for. We interpret this result to be an effect of allopolyploidization. We then compare the outcomes of this WGD with an independent one that happened in the teleost fish lineage. We find that if a gene pair was retained in duplicate in *X. laevis*, the orthologous pair is more likely to have been retained in duplicate in zebrafish, suggesting that similar factors, among them subfunctionalization, determined which gene pairs survived in duplicate after the two WGDs.

rate of evolution | *Silurana tropicalis* | whole-genome duplication

Polyploidy, also termed whole-genome duplication (WGD) is a frequent phenomenon in eukaryotes (1). A WGD is followed by extensive and rapid genome restructuring involving many gene losses, so that only one of the two gene copies remains in most genomes that underwent ancient polyploidization [for example, fish and yeast (2, 3)]. Alterations in function are expected among genes retained in duplicate. In some cases, one copy may acquire a new function (neofunctionalization), while the other keeps the ancestral function. The models of Lynch and Force (4, 5) also propose the existence of subfunctionalization, in which each copy retains a subset of the functions of the ancestral gene. Sub and neofunctionalization models make different predictions about the rate and symmetry of sequence evolution in the duplicates.

Asymmetry in evolutionary rates between the protein sequences of the two copies is often interpreted as a footprint of neofunctionalization, especially if it is associated with evidence of positive selection in the accelerated copy (6). Several studies of paleopolyploid genomes have shown that rate asymmetry between the two copies can be widespread. For example, asymmetry was seen in 6% of retained gene pairs in *Xenopus laevis* and in 25–36% of pairs in teleost fishes (6–8). Relatively few examples of subfunctionalization of duplicated genes have been demonstrated so far, the best-known being those of fish *mitf* (9), *sox9* (10), *synapsin* (11), *POMC* (12), *mbx* (13), and the plant gene *RPL32-SODcp* (14). A few studies have attempted to detect subfunctionalization on a larger scale after WGD. Aury *et al.* (15) used successive rounds of WGD in *Paramecium* to test Force *et al.*'s (5) prediction that subfunctionalized gene pairs should be resistant to reduplication. Their results suggest that subfunctionalization has occurred, but only rarely, in *Paramecium* genes. Other studies of subfunctionalization after WGD have focused on complementary amino acid substitution in protein pairs (6) and on the differential loss of regulatory regions between duplicated copies of developmental genes (16).

The most powerful method currently available to study the divergence of function between duplicated genes on a large scale is the analysis of their transcription profiles. Many studies have shown expression divergence between WGD-duplicates (17–22). However, a major obstacle encountered in all these studies is that they could not differentiate between sub and neofunctionalization because the pattern of expression before duplication was unknown. This obstacle was overcome recently for gene pairs that were formed by WGD in *Saccharomyces cerevisiae* by comparing their pattern of expression to *Candida albicans*, an outgroup whose genome was not duplicated and therefore can be used to approximate the ancestral expression state (23).

Here, we apply a similar approach to search for evidence of gene subfunctionalization after WGD in a vertebrate system. We compare the expression profiles of gene pairs preserved in duplicate after WGD in *X. laevis* to the expression profiles of orthologous genes in the unduplicated clawed frog *S. tropicalis* (sometimes also called *X. tropicalis*). The WGD that has been proposed for *X. laevis* has not yet been validated by a complete genome sequence, but it is estimated to have occurred 21–54.6 Mya (6, 24, 25) and it is likely to have been an allopolyploidization because interspecies crosses in *Xenopus* often produce fertile polyploid offspring and phylogenetic studies have shown that other polyploid clawed frogs are ancient allopolyploids (24, 26, 27).

We used the extensive expressed sequence tag (EST) and cDNA sequence resources available for these species (20, 21, 25) to detect genes present in one copy in *S. tropicalis* and in two copies in *X. laevis*. We inferred the pattern of expression in these triplets and detected events of subfunctionalization. We then tested whether the subfunctionalized genes are a random subset of the genome.

Results

Construction of the Dataset. We clustered *Xenopus* expressed sequences (ESTs and full-length cDNAs) that are publicly available (558,503 sequences for *X. laevis* and 1,046,555 for *S. tropicalis*). We chose very stringent clustering parameters to avoid merging sequences expressed by paralogous genes [see *Methods*, supporting information (SI) *Methods*, and Fig. S1]. Using phylogenetic analysis, we built a dataset of 1,300 triplets, composed of one gene in *S. tropicalis* and its two coorthologs in *X. laevis*, whose duplication was most probably due to WGD.

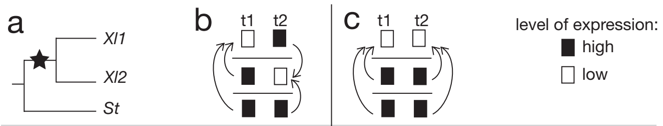
Author contributions: M.S. and K.H.W. designed research; M.S. performed research; M.S. analyzed data; and M.S. and K.H.W. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

*To whom correspondence should be addressed. E-mail: khwolfe@tcd.ie.

This article contains supporting information online at www.pnas.org/cgi/content/full/0708705105/DCSupplemental.



		subfunctionalization	asymmetric	Fisher test p-value
d	divergence in at least 2 tissues	19 (61)	49 (109)	0.01 (0.01)
e	divergence in exactly 2 tissues	15 (44)	41 (93)	0.02 (0.003)

Fig. 1. Principles of expression evolution in Xenopus sequence triplets. (*a*) Each triplet includes one gene in *S. tropicalis* (*St*) and its two coorthologs in *X. laevis* (*Xl1* and *Xl2*) created by WGD. (*b*) A case of subfunctionalization. Columns t1 and t2 represent two tissues. Arrows represent the results of statistical tests and point to the gene with the significantly lower expression level. Here, expression of gene *Xl1* in tissue t1 is significantly lower than expression of both *St* and *Xl2* in the same tissue. In tissue t2, gene *Xl2* shows lower expression. We infer that the gene was expressed in both tissues t1 and t2 before WGD and that subsequently the expression of *Xl1* decreased in t1, whereas expression of its paralog *Xl2* decreased in t2; this corresponds to a subfunctionalization pattern. (*c*) A case of asymmetric evolution of expression. Significant decreases in expression are inferred in *X. laevis* in two tissues, but they both involve the same *X. laevis* gene (here, *Xl1*). (*d*) Numbers of gene triplets showing subfunctionalization and asymmetric patterns of expression evolution as defined above. Numbers in parentheses do not include correction for multiple testing. *P* values by Fisher's test show that asymmetric evolution of expression is more frequent than subfunctionalization. (e) Numbers of triplets with subfunctionalized and asymmetric patterns of expression, defined as cases where there is a significant decrease of expression in exactly two tissues, and no significant decrease in the other nine tissues.

An early study based on a very small dataset proposed that 77% of genes were retained in duplicate in *X. laevis* (28). By using highly expressed genes to minimize errors associated with EST sampling, we estimate that ≈32–47% of genes were retained in double-copy in *X. laevis* after WGD (*SI Methods* and Figs. S2 and S3). Our figure is similar to Hellsten *et al.*'s (21) estimate of ≈25–50% retention. Gene loss has been less extensive after the relatively recent WGD in *X. laevis* than after the teleost-specific WGD, which is 10 times older (29, 30): In *Tetraodon nigroviridis* for instance, only 15% of genes were retained in duplicate (7).

We estimated the gene expression profiles of each triplet based on the tissue from which the ESTs were extracted. More precisely, we obtained for each gene in each triplet a measure of its level of expression in each of 11 tissues that had been used for library construction in both species (see *Methods* for a list). We measured the conservation of these expression patterns between the two *X. laevis* copies since WGD by a Spearman correlation coefficient. We find that the majority of duplicate pairs do not show much divergence in expression since WGD (median correlation rho = 0.64; Fig. S4), a result similar to that of Chain *et al.* (22).

Detection of Changes in Expression Profile: Subfunctionalization and Asymmetric Changes. We used parsimony to estimate the pattern of evolution of expression in each triplet. The principle of our analysis is shown in Fig. 1. We say that a pair of duplicates in *X. laevis* has become subfunctionalized if we infer that one gene copy 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 modified slightly a statistical test developed by Audic and Claverie (31) to take into account the effect of WGD and subsequent gene losses on the relative contribution of each gene to the total pool of mRNA in the cell (see *Methods*, *SI Methods*, and Figs. S5 and S6). We performed this test on the 1,300 triplets and found 61 examples of subfunctionalization (4%). This number drops to 19 triplets (1.2%) if we correct for multiple testing [false discovery rate (FDR) < 0.05] (32). These triplets are loci in which expression has been significantly decreased in one *X. laevis* copy in one tissue and in the other *X. laevis* copy in another tissue, whatever other changes happened (significant or not) in the nine remaining tissues. Among these 19 triplets, 15 have undergone significant changes in exactly two tissues and no significant changes in the other tissues (44 without correction for multiple testing).

We implemented another method to identify subfunctionalization between the two *X. laevis* copies. We constructed for each triplet the pattern Xl_{sum} by merging the patterns of expression of the two *X. laevis* copies (summing the number of ESTs per million for each tissue). Subfunctionalizations are cases where each of the copies in *X. laevis* has retained part of the ancestral function; therefore, the Spearman correlation of the patterns of expression between *S. tropicalis* and Xl_{sum} should be higher than both of the correlations between *S. tropicalis* and the individual *X. laevis* genes. This pattern was found in 11% of the triplets (144 triplets).

Cases of subfunctionalization, therefore, represent only a small proportion (1.2–11%) of the WGD-duplicates considered here; however, we have seen that most pairs have not diverged in expression since the WGD (Fig. S4). We tested whether, among the minority of genes that do show significant changes in expression in our dataset, the pattern of changes frequently corresponds to a subfunctionalization pattern. We searched in particular for two patterns of expression profile change, which we refer to as subfunctionalization and asymmetric change (Fig. 1 *b* and *c*). Both of these patterns involve decreases of expression in *X. laevis* compared with *S. tropicalis*. We detected 49 cases of asymmetric partitioning of expression (109 without multiple testing), defined as triplets where expression has decreased significantly in at least two tissues since WGD, in the same *X. laevis* copy (Fig. 1*c*). Therefore, we estimate that, among *X. laevis* gene pairs whose expression has diverged significantly, one-third (19 of 68) are subfunctionalized, which is less than the 50% expected by chance ($P = 0.01$ by Fisher's test). This ratio remains constant if we only consider genes with significant changes in exactly two tissues (Fig. 1*e*).

Relationship Between Rate of Sequence Evolution and Pattern of Expression Divergence. We examined whether the rate of evolution of a gene influences the evolution of its expression patterns after WGD. Because duplication tends to increase the rate of nonsynonymous sequence evolution [for instance, in *Xenopus* (21)], we instead measured this rate (*dN*) between two species whose genomes have not been duplicated: *S. tropicalis* and human. This *dN* value should be indicative of the gene's evolutionary rate before WGD. We find that genes that became subfunctionalized were more slowly evolving before WGD than the genes with no particular pattern of expression evolution (median *dN* values 0.154 and 0.214 respectively; $P = 0.018$ with two repetitions is significant at a 3.6% level; Fig. 2*a*). We consider this difference as biologically meaningful, even though its significance is marginal after Bonferroni correction, because the median *dN* is 40% lower in subfunctionalized genes than in the other genes, and because the power of the test is not very high given the small size of the datasets. In contrast, there is no significant rate difference between genes that later underwent an asymmetric pattern of expression evolution in *X. laevis* and those with no particular pattern of expression change ($P = 0.97$; Fig. 2*a*).

This preferential subfunctionalization of slowly evolving genes was unexpected. It is not a bias because of differences in mutation rate or in the age of the duplicates, because the levels of synonymous substitution are not significantly different among the three categories of genes (Fig. 2*b*). There is, however, another possible bias. By construction, the triplets with either subfunctionalization or asymmetric change of expression are expressed in more tissues and at a higher level than triplets that do not show these patterns (Fig. 2*c*), and it is known that genes expressed in many tissues evolve more slowly than other genes (33, 34). However, after correcting for this bias we still find that genes that became subfunctionalized are descended from exceptionally slowly evolving ancestors (Fig. 3*a*; $P = 0.009$ with four

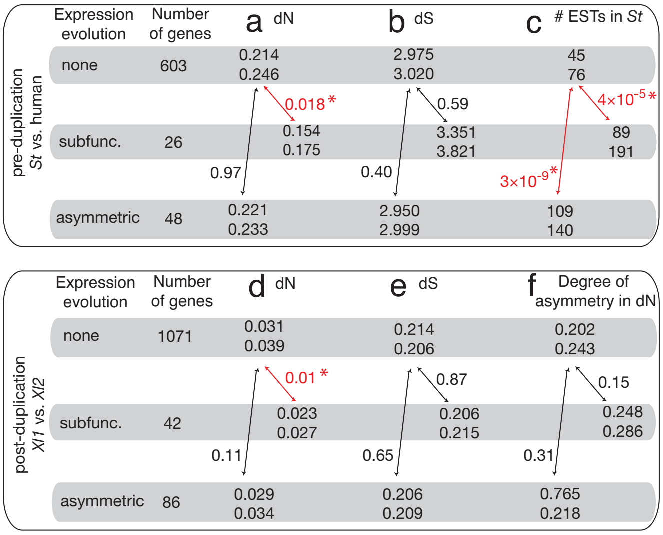


Fig. 2. Rates of sequence evolution are correlated with expression evolution. Patterns of expression evolution are classified as subfunctionalized or asymmetric as in Fig. 1e; ''none'' refers to triplets that show neither of these patterns. *dN*, *dS*, and #ESTs are, respectively, the levels of nonsynonymous and synonymous sequence evolution and the number of ESTs in *S. tropicalis (St)*. The median and mean of these variables are indicated (above and below, respectively), and *P* values for Wilcoxon tests for the pairwise comparison between the set ''none'' and each of the other sets of genes are shown near the arrows. Significant *P* values after Bonferroni correction (two test repetitions per panel) are indicated in red and marked with an asterisk. (*a–c*) Preduplication rates of sequence evolution are estimated between *S. tropicalis* and human. (*d–f*) Postduplication rates are computed between the two copies in *X. laevis (Xl1* and *Xl2*).

repetitions is significant at a 3.6% level). Moreover, this method of analysis also shows an opposite pattern in genes that developed asymmetric expression patterns after WGD: The ancestors of these genes were unusually quickly evolving (Fig. 3*b*; $P = 0.001$ with four repetitions is significant at a 0.4% level). These results are very surprising because they show that the rate of sequence evolution before the duplication influences the pattern of evolution of expression after the duplication.

We then asked whether the triplets of genes with subfunctionalization or asymmetric patterns of expression evolution have a particular rate of evolution after WGD. We computed the nonsynonymous divergence between the two paralogous copies in *X. laevis* and compared the mean values among groups with different patterns of expression evolution (Fig. 2*d*). Nonsynonymous divergence between *X. laevis* copies is significantly smaller for subfunctionalized genes (median *dN*: 0.023) than for genes with no particular pattern of expression evolution (median *dN*: 0.031; $P = 0.01$ by Wilcoxon test with two repetitions is significant at a 2% level). This effect is only marginally significant after correction for expression bias (Fig. 3*c*, $P = 0.03$ with four repetitions is marginally significant at a 12% level). Genes with an asymmetric pattern of evolution do not have a particular rate of evolution after duplication (Fig. 3*d*; $P = 0.08$ is not significant after Bonferroni correction). We also evaluated the asymmetry in the rates of nonsynonymous substitution between the two copies in *X. laevis*, using like-tri-test (35), but we found no link between sequence asymmetry and the pattern of expression evolution (Fig. 2*f*). A recent study showed that rates of nonsynonymous sequence evolution increased after WGD in *X. laevis* sequences, but this was a small effect only visible after concatenation of the sequences (21). Therefore, it is possible that nonsynonymous rates of evolution are asymmetric after duplication, but this effect is not visible on a gene-by-gene basis.

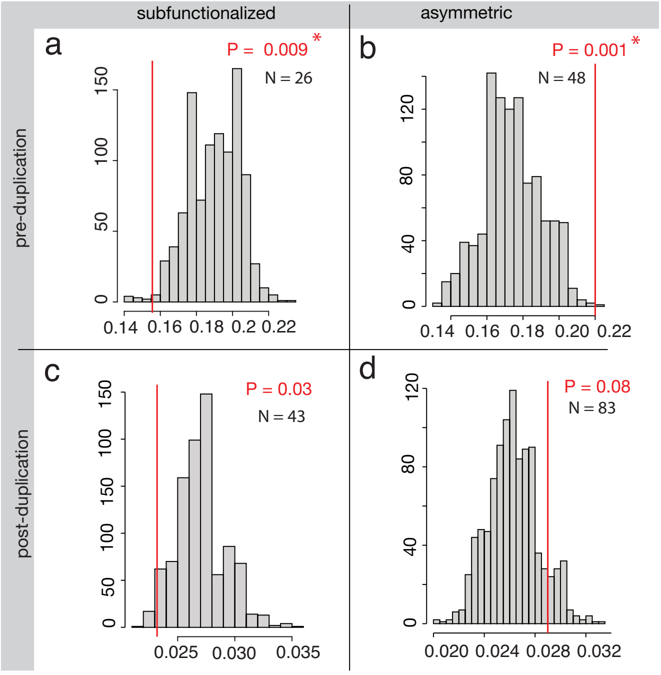


Fig. 3. The rate of nonsynonymous sequence evolution (*dN*) before the duplication influences the pattern of evolution of expression after the duplication. Shown are comparisons of an observed median *dN* value (red line) with a histogram of the distribution of expected values; the *x* axis in each panel is in *dN* units, and the *y* axis shows the number of genes in the histogram. *P* values for the comparisons between the observed values and the distributions are shown. Asterisks mark tests that are significant at a 5% level after Bonferroni correction for the four test repetitions. (*a*) The median value (0.154) of $dN_{\text{(human, }S.\ tropicalis\text{)}}$ for the set of 26 genes that show subfunctionalization in *X. laevis* and have an annotated ortholog in human is superimposed on a histogram showing the distribution of median values of $dN_{\text{(human, }S.\ tropicalis\text{)}}$ obtained from 1,000 samples. Each of these samples contains 100 genes chosen randomly from triplets with neither subfunctionalized nor asymmetric pattern of evolution of expression, such that the distribution of the levels of expression (measured in *S. tropicalis*) in each sample is the same as in the distribution observed in the set of 26 subfunctionalized genes. (*b–d*) Comparisons of observed median *dN* values with histograms of the distributions of median *dN* values obtained from same-size samples of genes showing no significant expression evolution and after correction for expression level. The four panels compare the observed median *dN* values (red line) of loci whose *X. laevis* coorthologs show either subfunctionalization (*a* and *c*) or asymmetric (*b* and *d*) patterns of expression evolution, to distributions sampled from triplets that do not show such patterns and after correction for expression level. *a* and *b* show *dN* values calculated between *S. tropicalis* and human, representing the preduplication rate of evolution, and *c* and *d* show *dN* values calculated from comparisons between the *X. laevis* coorthologs, representing the postduplication rate.

Convergent Outcomes of Two Independent WGDs in Teleost Fish and *X. laevis*. We have seen that subfunctionalized genes in *X. laevis* are distinctive because they were slowly evolving before WGD. Therefore, it is possible that some genes are more prone to subfunctionalization than others. To test this hypothesis, we compared the outcomes of two WGDs that occurred independently in vertebrates: one in *X. laevis* and one at the base of teleost fish lineage. First, we tested the null hypothesis that the two WGDs should have independent results in terms of double-copy retentions. In other words, whether a gene pair was or was not retained in duplicate in *X. laevis* should have no bearing on whether or not its orthologous pair was retained in fish. Because the genome of zebrafish has been completely sequenced, we can assess with certainty whether a gene pair was retained in two copies in this species after its WGD, whereas this is not feasible

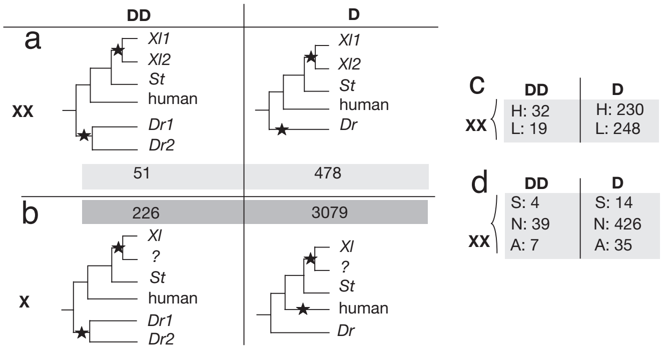


Fig. 4. Comparison of genes retained in double-copy in *X. laevis* and in zebrafish shows that the two independent WGDs do not have independent outcomes. (*a*) 529 triplets with two copies in *X. laevis* (designated XX) were sorted into groups whose orthologs in zebrafish are double-copy or single-copy (designated DD and D, respectively). Zebrafish gene pairs (*Dr1* and *Dr2*) in the same Homolens family were attributed to the teleost WGD if their duplication is older than the speciation of zebrafish and pufferfish. (*b*) 3305 doublets (designated X) for which we could not find a second copy in *X. laevis* were sorted similarly according to their duplication status in zebrafish. (*c*) Distribution of the 529 XX triplets by their expression level in *S. tropicalis* (two classes: H, high; L, low) and their duplication status in zebrafish. (*d*) Distribution of the 529 XX triplets by their pattern of expression evolution (three classes: S, subfunctionalization; A, asymmetric partitioning; N, neither) and their duplication status in zebrafish.

in *X. laevis*. We identified reliable orthologs in zebrafish for half of our triplets (529 genes; Fig. 4*a*). The fraction of these families that contain two WGD coorthologs in zebrafish is 9.5% (51 of 529). This fraction is significantly higher than the corresponding fraction for gene families in which only one *X. laevis* copy was found (6.8%; 226 of 3,305; Fig. 4*b*; $P = 0.036$ by Fisher's test). The latter set corresponds to families for which we did not detect a second gene in *X. laevis*, which in some cases might have been because no EST was sampled rather than because the gene is truly single-copy in *X. laevis*. For this reason, our test is a conservative one, and we conclude that genes retained in duplicate after one vertebrate WGD event have increased probability of also having been retained after the other.

What is the reason for this convergence? It has been shown that highly expressed genes are overretained after WGD (15, 36). If expression level differences are responsible for the nonindependence of the two WGDs, we would expect that highly expressed genes should have higher frequencies of retention in duplicate than weakly expressed genes, after both WGDs. We divided our dataset of *Xenopus*-fish orthologs into two classes depending on their expression level in *S. tropicalis* (low or high; Fig. 4*c*) and observed that the proportion of genes retained in duplicate in zebrafish is higher (12%; 32 of 262) for genes highly expressed in *S. tropicalis* than for low-expression genes (7%; 19 of 267; $P = 0.04$ by Fisher's test). The genes responsible for the nonindependence of the two duplications are therefore highly expressed.

We can ask whether these highly expressed genes have been retained for the same reason after the two WGDs. By definition, subfunctionalized genes are expressed in several tissues and they are also highly expressed (Fig. 2*c*), so it is possible that the genes convergently retained in duplicate after the two WGDs are enriched in subfunctionalized genes. Indeed, we find that genes that have been subfunctionalized in *X. laevis* have a higher frequency of parallel retention in zebrafish (22%; 4 of 18; Fig. 4*d*) than do those that show no pattern of expression divergence or an asymmetric divergence pattern (8 and 16%, respectively; data from Fig. 4*d*; $P = 0.037$ by χ^2 test of homogeneity among these three categories). These results suggest that gene pairs retained by subfunctionalization in *X. laevis* also tended to be retained by subfunctionalization in zebrafish. Unfortunately, we cannot directly test whether these pairs have been subfunctionalized in zebrafish, because no expression data are available for any outgroup species that diverged shortly before the teleost WGD. However, our hypothesis of convergent subfunctionalization receives some support from a comparison of the divergence of expression profiles between the pairs in zebrafish and in *X. laevis* (Fig. S7).

Discussion

In their pioneering study of 17 duplicated genes in *X. laevis*, Hughes and Hughes (28) already noticed that in four cases the two copies were expressed in different tissues or at different developmental times, and this trend was confirmed recently in larger datasets (20, 21). We detect relatively little subfunctionalization in our dataset (1.2–11% of the WGD-duplicates considered). This may be because most pairs have not diverged in expression since the WGD and subfunctionalization actually only happened in a small percentage of pairs. Alternatively, our ability to detect subfunctionalization is perhaps limited. If we had complete information about transcription in every tissue, we could more accurately detect significant expression divergence between gene pairs in some particular tissues and hence obtain a reliable estimate of the fraction of genes undergoing subfunctionalization. The frequency of subfunctionalization we estimate here is a lower limit, because we examined a limited number of tissues, but we cannot propose an upper bound for this figure. Even though we were not able to estimate the frequency of subfunctionalization, we could still examine the characteristics of genes that became subfunctionalized.

We find that some genes are predisposed to subfunctionalization. Genes that underwent subfunctionalization in *X. laevis* tend to be slowly evolving in other species, and conversely genes with an asymmetric pattern of expression evolution in *X. laevis* tend to evolve faster than expected in these outgroups. These results are only of medium statistical significance, partly because the limited size of the datasets weakens the power of the tests, and partly because of the necessity to correct for multiple testing. Nonetheless, we observed that the rate of sequence evolution influences the retention of some genes after WGD, and we propose a model to explain our observations.

Genes retained in duplicate after WGD are more likely to belong to gene families with slow rates of sequence evolution (7, 37, 38) or high expression levels (15, 36). Slow sequence evolution is correlated with a high level (or wide breadth) of expression in both yeast and vertebrates (33, 39), and both observations may be due to the same phenomenon. Davis and Petrov (37) were unable to find an obvious explanation for their discovery that slowly evolving genes are preferentially retained in duplicate, but they suggested that the bias may be an indirect correlation due to a third variable that is responsible for the retention and is correlated with the other two. Candidates for this third variable include the presence of many *cis*-regulatory regions (5, 40), of genes coding for multidomain proteins (41), and pleiotropic genes (model 3 in ref. 42). Other models predict that genes with a particular function, such as regulatory genes (43), should be retained in duplicate more often than expected after WGD. Alternatively there may be a direct relationship between the expression level (or the rate of sequence evolution) and the propensity to be retained in duplicate. Highly expressed genes may be retained in duplicate after WGD simply because they are beneficial for gene dosage (15). We discuss below that the rate of evolution seems to be directly responsible for double-copy retention in *Xenopus*, at least for the subset of gene pairs whose expression is divergent.

We have shown that slowly evolving genes are more subject to subfunctionalization. Theoretical studies of subfunctionalization

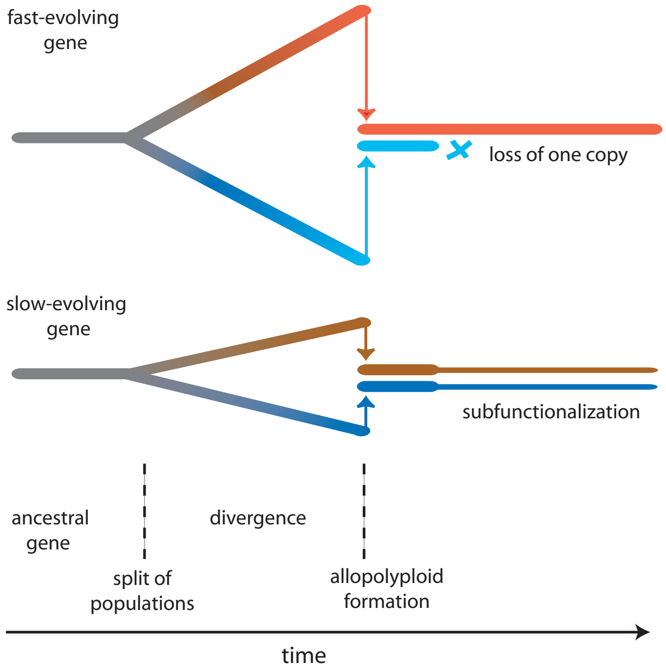


Fig. 5. Model to explain why slowly evolving genes are preferentially subfunctionalized after allopolyploidization. The horizontal axis represents time. Initially two populations diverge and their genes begin to accumulate sequence differences, represented by color changes and vertical separation. They subsequently hybridize to form an allopolyploid. For a fast-evolving gene (*Upper*), the two copies are very different at the time of allopolyploid formation. They are unlikely to be functionally indistinguishable, and it is probable that only the better-functioning one of them will be retained. For a slow-evolving gene (*Lower*), the two copies are less different and more likely to be able to replace one another functionally. Subfunctionalization (represented by thinning of the lines) may result.

do not predict that genes becoming subfunctionalized should evolve more slowly than others before duplication (5, 40). On the contrary, subfunctionalization is supposed to be a neutral event that occurs because of neutral mutations impairing different subfunctions in the duplicates. Because we correct for expression bias (Fig. 3), we show more exactly that among a pool of genes with the same level of expression, slowly evolving genes are more likely to be subfunctionalized, and fast evolving genes are more likely to have an asymmetric pattern of expression evolution. We can extrapolate that genes that are subfunctionalized are going to be retained in two copies in the genome (both copies are necessary to perform the ancestral function), but, in contrast, it is likely that many genes with an asymmetric pattern of evolution of expression will eventually return to single-copy state. This is confirmed by our comparison between WGDs in *Xenopus* and zebrafish: subfunctionalized genes in *Xenopus*, but not genes with an asymmetric pattern of expression evolution, are retained in two copies more than expected in zebrafish.

These observations lead us to propose that slowly evolving genes were more easily subfunctionalized in *X. laevis* and therefore more easily retained long after WGD. Our model of gene evolution after WGD in *X. laevis* is illustrated in Fig. 5, which is based on the assumption that the WGD was an allopolyploidization, as is most likely (26, 27). Most models to explain gene retention after WGD postulate that the two copies are equal at birth (e.g., refs. 5, 43) but this is not true in the case of allopolyploidization. In our model, two diverging populations accumulate sequence differences, but to a greater extent in faster-evolving genes than in slower-evolving genes. When their two genomes are merged by allopolyploidization, the slower-evolving loci have accumulated fewer substitutions so the two copies may still be interchangeable and subfunctionalization can occur. In contrast, in the faster-evolving genes it is more likely that there are functional differences between the two copies, and one of them functions better than the other. If so, it may be deleterious for the better-functioning gene copy to lose any of its subfunctions. Such a situation will prevent subfunctionalization from happening; instead, the worse-functioning gene copy will be lost completely. In *X. laevis*, we observe the consequences of an allopolyploidization of medium age, where nearly half the genes are still in duplicate. For the fast-evolving genes in this genome we tend to see an asymmetry in the expression patterns and we anticipate that the worse-functioning copy will be lost eventually. Our observations and our model contradict a previous hypothesis by Spring (44), who suggested that slower-evolving genes would be more redundant at the time of allopolyploidization and therefore easier to lose.

Note that our model is also valid for an autopolyploidization or any other kind of gene duplication, if the genes can survive in duplicate long enough to attain sequence divergence. In each case the genes with slower rates of nonsynonymous substitution are expected to remain equivalent (and therefore prone to subfunctionalization) for a longer time. This hypothesis is supported by our laboratory's previous work on WGD in yeast, where we found that slow-evolving genes retained their interchangeability for a longer time period after WGD than fast-evolving genes (45, 46). If this idea is correct it can account for the preferential retention of slow-evolving genes after any kind of duplication, as seen by Davis and Petrov (37). Thus, subfunctionalization may be a force in the long-term evolution of duplicated genes, in addition to its originally postulated role (5) in their initial preservation.

Methods

The methods we used for stringent EST clustering, building triplets of homologous *Xenopus* genes, establishing orthology relationships, and estimating rates of sequence evolution are described in *SI Methods*. The derivation of our estimate that 32–47% of genes were retained in double-copy in *X. laevis* after WGD is also given in *SI Methods*.

To estimate expression profiles of frog genes, we classified the available *Xenopus* EST libraries into tissues (104 libraries in *X. laevis*, 51 in *S. tropicalis*) and identified the following 11 tissues (or developmental stages) as being common between the two species: brain, embryo, heart, kidney, liver, lung, ovary, skin, spleen, tadpole, and testis. By construction, each contig in a triplet is composed of ESTs that were used to infer its pattern of expression. Zebrafish EST analysis is described in *SI Methods*.

To detect differences in expression level between the two copies in *X. laevis* (denoted *Xl1* and *Xl2*) and *S. tropicalis* (*St*) in one tissue, we used Audic and Claverie's Bayesian test (31), which takes the total number of ESTs sequenced in each tissue from each species into account. We modified the test slightly because the null expectation is that the EST count of gene *St* should be ≈1.3 times greater (exact value: $e^{0.26}$; see *SI Methods*) than the individual EST counts of its orthologs *Xl1* and *Xl2*. To detect a significant decrease in the expression of gene *Xl1* in a particular tissue we tested whether, for this tissue in the two species, (*i*) the EST count of *Xl1* × $e^{0.26}$ is significantly lower than the count of *St*, and (*ii*) the EST count of *Xl2* × $e^{0.26}$ is not significantly lower than the count of *St.*

ACKNOWLEDGMENTS. We thank G. Conant for help with computing of asymmetry of sequence evolution; L. Guéguen for help with FDR and tree parsing; and K. Byrne, G. Conant, B. Cusack, C. Frank, J. Gordon, N. Khaldi, J. Mower, D. Scannell, M. Webster, M. Woolfit, and two anonymous referees for helpful comments. This work was supported by Irish Research Council for Science, Engineering and Technology and the Science Foundation Ireland.

1. Otto SP, Whitton J (2000) Polyploid incidence and evolution. *Annu Rev Genet* 34:401–437.
2. Jaillon O, *et al.* (2004) Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature* 431:946–957.
3. Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH (2006) Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* 440:341–345.

4. Lynch M, Force A (2000) The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154:459–473.
5. Force A, *et al.* (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151:1531–1545.
6. Chain FJ, Evans BJ (2006) Multiple mechanisms promote the retained expression of gene duplicates in the tetraploid frog Xenopus laevis. PLoS *Genet* 2:e56.
7. Brunet FG, *et al.* (2006) Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol Biol Evol* 23:1808–1816.
8. Steinke D, Salzburger W, Braasch I, Meyer A (2006) Many genes in fish have species-specific asymmetric rates of molecular evolution. *BMC Genomics* 7:20.
9. Altschmied J, *et al.* (2002) Subfunctionalization of duplicate mitf genes associated with differential degeneration of alternative exons in fish. *Genetics* 161:259–267.
10. Cresko WA, *et al.* (2003) Genome duplication, subfunction partitioning, and lineage divergence: Sox9 in stickleback and zebrafish. *Dev Dyn* 228:480–489.
11. Yu WP, Brenner S, Venkatesh B (2003) Duplication, degeneration and subfunctionalization of the nested synapsin-Timp genes in Fugu. *Trends Genet* 19:180–183.
12. de Souza FS, Bumaschny VF, Low MJ, Rubinstein M (2005) Subfunctionalization of expression and peptide domains following the ancient duplication of the proopiomelanocortin gene in teleost fishes. *Mol Biol Evol* 22:2417–2427.
13. Chang L, Khoo B, Wong L, Tropepe V (2006) Genomic sequence and spatiotemporal expression comparison of zebrafish mbx1 and its paralog, mbx2. *Dev Genes Evol* 216:647–654.
14. Cusack BP, Wolfe KH (2007) When gene marriages don't work out: Divorce by subfunctionalization. *Trends Genet* 23:270–272.
15. Aury JM, *et al.* (2006) Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. *Nature* 444:171–178.
16. Woolfe A, Elgar G (2007) Comparative genomics using Fugu reveals insights into regulatory subfunctionalization. *Genome Biol* 8:R53.
17. Casneuf T, De Bodt S, Raes J, Maere S, Van de Peer Y (2006) Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant Arabidopsis thaliana. *Genome Biol* 7:R13.
18. Blanc G, Wolfe KH (2004) Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *Plant Cell* 16:1679–1691.
19. Duarte JM, *et al.* (2006) Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of Arabidopsis. *Mol Biol Evol* 23:469–478.
20. Morin RD, *et al.* (2006) 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.
21. Hellsten U, *et al.* (2007) Accelerated gene evolution and subfunctionalization in the pseudotetraploid frog Xenopus laevis. *BMC Biol* 5:31.
22. Chain FJ, Ilieva D, Evans BJ (2008) Duplicate gene evolution and expression in the wake of vertebrate allopolyploidization. *BMC Evol Biol* 8:43.
23. Tirosh I, Barkai N (2007) Comparative analysis indicates regulatory neofunctionalization of yeast duplicates. *Genome Biol* 8:R50.
24. Evans BJ, Kelley DB, Melnick DJ, Cannatella DC (2005) Evolution of RAG-1 in polyploid clawed frogs *Mol Biol Evol* 22:1193–1207.
25. Pollet N, Mazabraud A (2006) in *Genome Dynamics, Volume 2, Vertebrate Genomes*, ed Volff J-N (Karger, Basel, Switzerland), pp 138–153.
26. Kobel HR (1996) in *The Biology of Xenopus*, eds Tinsley RC, Kobel HR (Clarendon, Oxford), pp 391–401.
27. Evans BJ (2007) Ancestry influences the fate of duplicated genes millions of years after polyploidization of clawed frogs (Xenopus). *Genetics* 176:1119–1130.
28. Hughes MK, Hughes AL (1993) Evolution of duplicate genes in a tetraploid animal, Xenopus laevis *Mol Biol Evol* 10:1360–1369.
29. Vandepoele K, De Vos W, Taylor JS, Meyer A, Van de Peer Y (2004) Major events in the genome evolution of vertebrates: Paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proc Natl Acad Sci USA* 101:1638–1643.
30. Hoegg S, Brinkmann H, Taylor JS, Meyer A (2004) Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. *J Mol Evol* 59:190–203.
31. Audic S, Claverie JM (1997) The significance of digital gene expression profiles. *Genome Res* 7:986–995.
32. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A Practical and powerful approach to multiple testing. *J Roy Stat Soc B* 57:289–300.
33. Duret L, Mouchiroud D (2000) Determinants of substitution rates in mammalian genes: Expression pattern affects selection intensity but not mutation rate. *Mol Biol Evol* 17:68–74.
34. Zhang L, Li WH (2004) Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Mol Biol Evol* 21:236–239.
35. Conant GC, Wagner A (2003) Asymmetric sequence divergence of duplicate genes. *Genome Res* 13:2052–2058.
36. Seoighe C, Wolfe KH (1999) Yeast genome evolution in the post-genome era. *Curr Opin Microbiol* 2:548–554.
37. Davis JC, Petrov DA (2004) Preferential duplication of conserved proteins in eukaryotic genomes. *PLoS Biol* 2:E55.
38. Davis JC, Petrov DA (2005) Do disparate mechanisms of duplication add similar genes to the genome? *Trends Genet* 21:548–551.
39. Drummond DA, Raval A, Wilke CO (2006) A single determinant dominates the rate of yeast protein evolution. *Mol Biol Evol* 23:327–337.
40. Lynch M, O'Hely M, Walsh B, Force A (2001) The probability of preservation of a newly arisen gene duplicate. *Genetics* 159:1789–1804.
41. Gibson TJ, Spring J (1998) Genetic redundancy in vertebrates: Polyploidy and persistence of genes encoding multidomain proteins. *Trends Genet* 14:46–49.
42. Nowak MA, Boerlijst MC, Cooke J, Smith JM (1997) Evolution of genetic redundancy. *Nature* 388:167–171.
43. Birchler JA, Pal-Bhadra M, Bhadra U (2003) Dosage dependent gene regulation and the compensation of the X chromosome in Drosophila males. *Genetica* 117:179–190.
44. Spring J (1997) Vertebrate evolution by interspecific hybridisation—are we polyploid? *FEBS Lett* 400:2–8.
45. Scannell DR, *et al.* (2007) Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. *Proc Natl Acad Sci USA* 104:8397–8402.
46. Byrne KP, Wolfe KH (2007) Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication. *Genetics* 175:1341–1350.

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-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:

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

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:

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

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

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

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 ≈0.40 ± 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 approximately 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 $\approx$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.

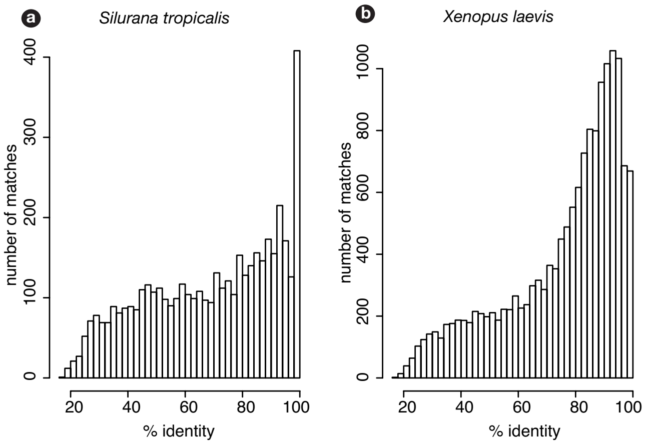


Fig. S1. The excess of recent duplicates in *X. laevis* by comparison to *S. tropicalis* is the hallmark of a recent WGD in *X. laevis.* We estimated the number of paralogs in each species by the number of pairs of coding sequences that align highly significantly (BLASTN *E* value $< 10^{-10}$; ref. 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.

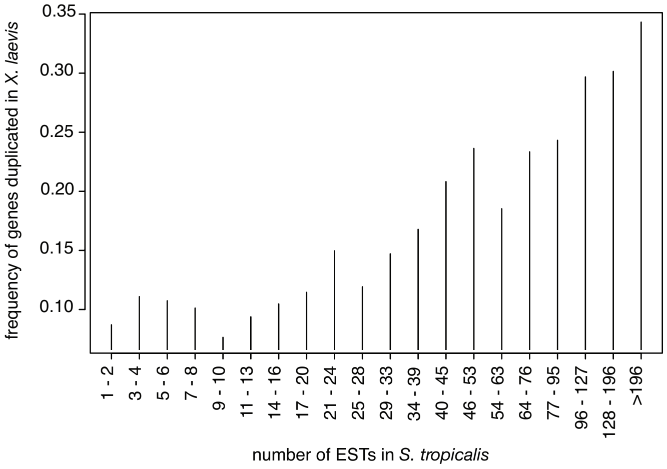


Fig. S2. The frequency of genes detected in two copies in *X. laevis* increases with the level of expression of their ortholog in *S. tropicalis*. The total set of 8,116 genes in *S. tropicalis* with at least one ortholog in *X. laevis* was divided into 20 bins of equal size according to expression level (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.

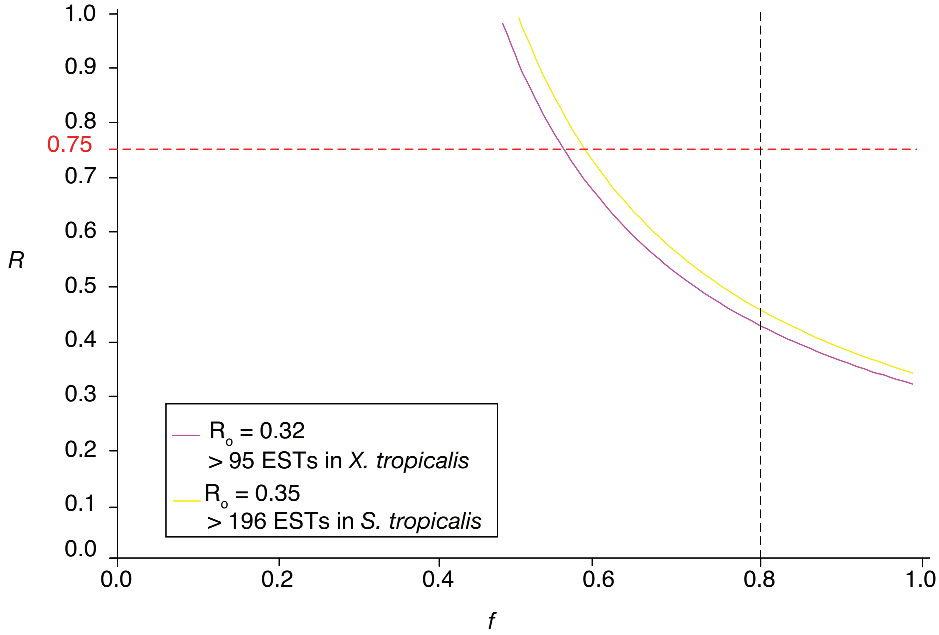


Fig. S3. Computation of the double-copy retention frequency (*R*) in *X. laevis* for different values of the probability that an extant gene is detected in the *X. laevis* EST data (*f*). The computation is based on the observed double-copy retention in the most highly expressed genes (pink for the 20% 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.

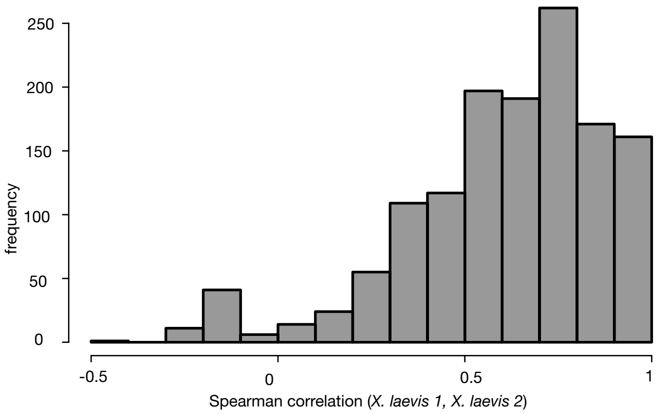


Fig. S4. Distribution of levels of conservation of expression patterns in 1,300 pairs of paralogous genes in *X. laevis* created by the WGD, measured as a Spearman correlation coefficient across 11 tissues.

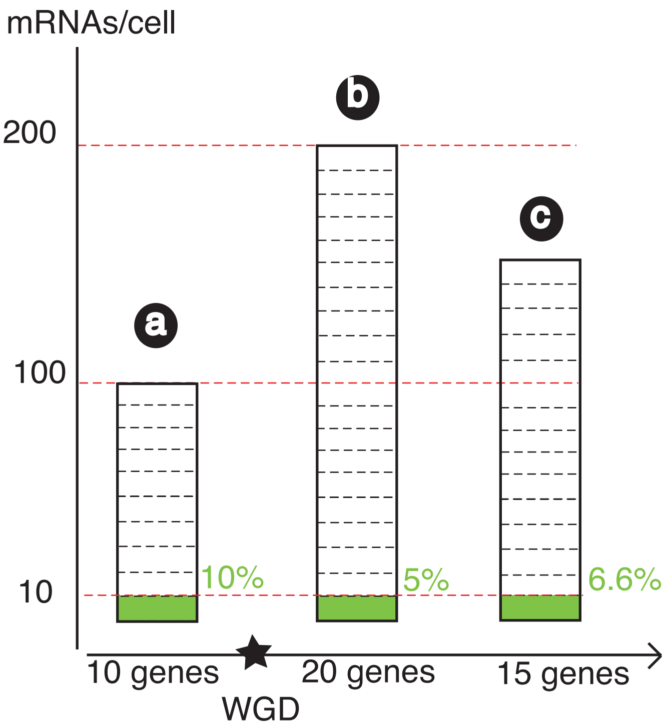


Fig. S5. Simplified system explaining why observed levels of expression should be lower in *X. laevis* than in *S. tropicalis*. (*a*) For illustration, we imagine that *S. tropicalis* has only 10 genes, each transcribed into 10 mRNAs per cell. (*b* and *c*) After WGD (*b*) and gene loss, five pairs of 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*.

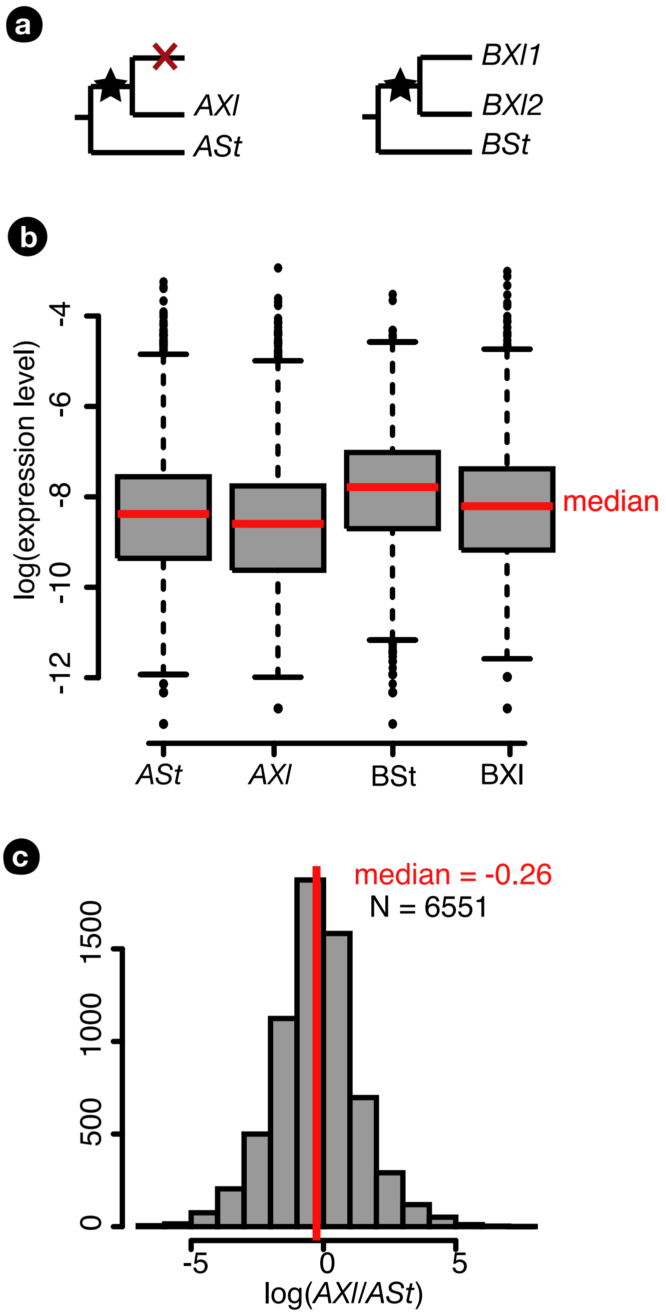


Fig. S6. The null expectation is that observed expression levels should be ≈30% greater in *S. tropicalis* (*St*) than in *X. laevis (Xl)*. (*a*) We designate genes that are single-copy in both species as ''A,'' and genes that are members of a retained duplicate pair in *X. laevis* as ''B.'' (*b*) Box-plots 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.

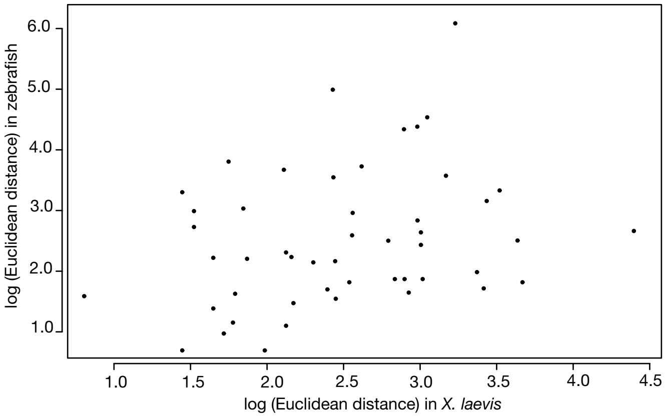


Fig. S7. Comparison of the expression divergences observed after WGD in zebrafish and after WGD in *X. laevis*. If orthologous pairs were retained for the same reasons after the two duplications, we should observe a correlation between the levels of within-pair expression divergence in the two species, 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; $P = 0.05$; $n = 49$).