Sequencing and analysis of 10,967 full-length cDNA clones from *Xenopus laevis* and *Xenopus tropicalis* reveals post-tetraploidization transcriptome remodeling

Ryan D. Morin, Elbert Chang, Anca Petrescu, et al.

Genome Res. 2006 16: 796-803
Access the most recent version at doi:10.1101/gr.4871006

Resource

Sequencing and analysis of 10,967 full-length cDNA clones from *Xenopus laevis* and *Xenopus tropicalis* reveals post-tetraploidization transcriptome remodeling

Ryan D. Morin,[1] Elbert Chang,[1] Anca Petrescu,[1] Nancy Liao,[1] Malachi Griffith,[1] Robert Kirkpatrick,[1] Yaron S. Butterfield,[1] Alice C. Young,[3] Jeffrey Stott,[1] Sarah Barber,[1] Ryan Babakaiff,[1] Mark C. Dickson,[2] Corey Matsuo,[1] David Wong,[1] George S. Yang,[1] Duane E. Smailus,[1] Keith D. Wetherby,[3] Peggy N. Kwong,[3] Jane Grimwood,[2] Charles P. Brinkley III,[3] Mabel Brown-John,[1] Natalie D. Reddix-Dugue,[3] Michael Mayo,[1] Jeremy Schmutz,[2] Jaclyn Beland,[1] Morgan Park,[3] Susan Gibson,[1] Teika Olson,[1] Gerard G. Bouffard,[3] Miranda Tsai,[1] Ruth Featherstone,[1] Steve Chand,[1] Asim S. Siddiqui,[1] Wonhee Jang,[7] Ed Lee,[7] Steven L. Klein,[6] Robert W. Blakesley,[3] Barry R. Zeeberg,[4] Sudarshan Narasimhan,[9] John N. Weinstein,[4] Christa Prange Pennacchio,[8] Richard M. Myers,[2] Eric D. Green,[3] Lukas Wagner,[7] Daniela S. Gerhard,[5] Marco A. Marra,[1] Steven J.M. Jones,[1] and Robert A. Holt[1,10]

[1]British Columbia Genome Sciences Centre, BCCA, Vancouver, BC V5Z 1L3 Canada; [2]Stanford Human Genome Center and Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA; [3]NIH Intramural Sequencing Center, National Human Genome Research Institute, [4]Genomics and Bioinformatics Group, Laboratory of Molecular Pharmacology, [5]National Cancer Institute, and [6]National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland 20892, USA; [7]National Center for Biotechnology Information, National Library of Medicine, Bethesda, Maryland 20894, USA; [8]The I.M.A.G.E Consortium, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, California 94550, USA; [9]2SRA International, Fairfax, Virginia 22033, USA

Sequencing of full-insert clones from full-length cDNA libraries from both *Xenopus laevis* and *Xenopus tropicalis* has been ongoing as part of the *Xenopus* Gene Collection Initiative. Here we present 10,967 full ORF verified cDNA clones (8049 from *X. laevis* and 2918 from *X. tropicalis*) as a community resource. Because the genome of *X. laevis*, but not *X. tropicalis*, has undergone allotetraploidization, comparison of coding sequences from these two clawed (pipid) frogs provides a unique angle for exploring the molecular evolution of duplicate genes. Within our clone set, we have identified 445 gene trios, each comprised of an allotetraploidization-derived *X. laevis* gene pair and their shared *X. tropicalis* ortholog. Pairwise d_N/d_S comparisons within trios show strong evidence for purifying selection acting on all three members. However, d_N/d_S ratios between *X. laevis* gene pairs are elevated relative to their *X. tropicalis* ortholog. This difference is highly significant and indicates an overall relaxation of selective pressures on duplicated gene pairs. We have found that the paralogs that have been lost since the tetraploidization event are enriched for several molecular functions, but have found no such enrichment in the extant paralogs. Approximately 14% of the paralogous pairs analyzed here also show differential expression indicative of subfunctionalization.

[Supplemental material is available online at www.genome.org. The sequence data from this study have been submitted to GenBank under accession nos. BC040971–BC100665 (not exclusively). Keyword MGC and organism *Xenopus* will be required to get only the XGC sequences in the range.]

Xenopus laevis (the African claw-toed frog) has long been a preferred model organism among developmental biologists. Features such as ease of maintenance, oocyte size and number, and an easily manipulated reproductive system make it an ideal organism for the study of early embryonic development (De Sa and Hillis 1990). Studies of embryonic development in *Xenopus* have provided insights into many salient aspects of vertebrate development that would be difficult to study in other vertebrate systems (Gilchrist et al. 2004). However, the study of *Xenopus* genetic material is difficult because of an allotetraploidization event in the *Xenopus* lineage ~30 million years ago (Mya) (Bisbee et al. 1977; Evans et al. 2004) that generated a more complex genome in all extant *Xenopus* species except for *Xenopus tropica-*

The content of this publication does not necessarily reflect the views or policies of the U.S. Department of Health and Human Services, nor does mention of trade names, commercial products, or organization imply endorsement by the U.S. Government.

[10]Corresponding author.
E-mail rholt@bcgsc.ca; fax (604) 877-6085.

Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.4871006.

16:796–803 ©2006 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/06; www.genome.org

lis, whose genome remains diploid (Graf and Kobel 1991; Hirsch et al. 2002). With a less complex genome as well as a shorter generation time, *X. tropicalis* is more amenable to genetic manipulation and has become the preferred *Xenopus* species for genetic analyses (Hirsch et al. 2002).

Several groups have performed large-scale EST studies on libraries from various tissues from both *X. laevis* and *X. tropicalis* (Klein et al. 2002; Blackshear et al. 2001; Gilchrist et al. 2004). However, for analysis of transcripts and gene structures, the quality of data and coverage provided by EST reads can be limiting. Sequence-verified full-length cDNA clones are more informative and have a higher sequence quality standard. Here we report the full open reading frame (ORF) sequencing and coding DNA segment (CDS) analysis of 10,967 *Xenopus* full-length cDNA clones (8049 from *X. laevis* and 2918 from *X. tropicalis*). These clones are from libraries that were constructed using RNA from numerous tissues and whole animals in various developmental stages. We expect that these clones and their verified full ORF sequences will be a valuable resource for the community. Furthermore, the availability of full ORF sequences for a large set of *Xenopus* clones provides a unique opportunity to study molecular evolution in the context of allotetraploidization. The putative *X. laevis* ancestral allotetraploidization event created a full set of paralogs, each from one of the parent species involved in the mating (Graf and Kobel 1991). The redundancy of the resultant tetraploid *X. laevis* genome has, in theory, afforded this species greater freedom to accumulate mutations that may otherwise be deleterious in a diploid genome, such as that of *X. tropicalis*. In the present study, carefully defining gene trios (gene sets comprised of the two allotetraploidization-derived *X. laevis* paralogs and their shared *X. tropicalis* ortholog) has allowed us to distinguish paralogs arising from genome duplication from paralogs arising by ordinary within-species gene family expansion. We focus our initial analysis, presented here, on detecting signatures of purifying and positive selection, and on exploring the evolution of tissue-specific gene expression.

Results

X. laevis and *X. tropicalis* cDNA libraries (Supplemental Tables S1 and S2) were end-sequenced by the National Intramural Sequencing Center, Washington University Genome Sequencing Center, and Agencourt Bioscience Corporation (Gerhard et al. 2004). Candidate full ORF clones were selected as previously described (Klein et al. 2002; Gerhard et al. 2004). Each candidate full ORF clone was fully sequenced to a consensus phred score of no less than 30 (Ewing and Green 1998) at each consensus position by either transposon insertion or primer walking as previously described (Wilson and Mardis 1997; Gordon et al. 1998; Butterfield et al. 2002; Strausberg et al. 2002; Yang et al. 2005). Coding DNA segment (CDS) annotation of the full insert sequences was performed as previously described (Gerhard et al. 2004). Distinct from previous MGC projects, however, we took a second approach to clone selection in an attempt to identify clones that might encode either amphibian-specific proteins or proteins too weakly conserved at the N terminus to be identified by comparison with proteins from other organisms. The technique we used (see Methods) assumes that a stronger conservation will be observed in the CDS than in the 5′-untranslated region (UTR) of paralogous genes. Most of the clones identified by this technique (~80%) were also identified by comparison with proteins from other species. The full-length sequences identified by this method are listed in Table 1. While there are seven proteins with no significant hits ($E < 10^{-10}$) in any mammalian protein, there are also six proteins with alignment scores more than two standard deviations below the mean value to the most closely related human protein (mean 73%, 13.8% standard deviation). The small number of novel proteins identified as well as the overlap with proteins identified from protein or mRNA comparison suggest that there are few proteins present in this cDNA collection that are structurally distinct from previously identified proteins.

Following full insert sequencing of candidate clones, only those clones with verified complete ORFs were given an XGC (*Xenopus* Gene Collection) identifier. These 10,967 clones (8049 from *X. laevis* and 2918 from *X. tropicalis*) are considered the core group of XGC clones and are the basis of the analysis presented here. All clone sequences have been submitted to GenBank, and the physical clones are available through the IMAGE distribution network.

From the set of 10,967 clones, we identified 445 distinct gene trios for analysis (see Methods). Again, a trio is a gene set comprised of the two allotetraploidization-derived *X. laevis* paralogs and their shared *X. tropicalis* ortholog. To explore the signature of selection between *X. laevis* and *X. tropicalis*, we applied the d_N/d_S test using the maximum likelihood method of Yang and Nielsen available as the codeml component of the PAML software package (PAML software release 3.14) (Yang et al. 1997). This method allows inference of evolutionary selection for mutations using the ratio of nonsynonymous (d_N) to synonymous (d_S) mutations in the coding DNA sequence of a phylogeny of homologous genes. In general, a d_N/d_S ratio (ω) >1 is evidence for positive selection acting to modify the function of a gene (Thornton and Long 2002; Zhang et al. 2002), whereas an ω significantly <1 suggests negative or purifying selection where functional constraint on the gene product has restricted the amount of nonsynonymous mutation. None of the pairwise comparisons between *X. laevis* paralogs or between each *X. laevis* gene and its *X. tropicalis* ortholog resulted in an ω significantly >1 (Fig. 1), suggesting that purifying selection has continued to act on *X. laevis* genes duplicated by allotetraploidization, and that in general

Table 1. *X. laevis* genes identified without protein comparison

Xenopus laevis accession	Similarity to closest *Homo sapiens* protein
BC081278	42%
BC080430	40%
BC079813	25%
BC084980	—
BC079815	35%
BC082713	35%
BC079817	44%
BC086299	35%
BC079818	—
BC079819	—
BC097879	—
BC081276	—
BC097923	—
BC077645	—

Summary of the *X. laevis* genes with no clear ortholog in *H. sapiens*. These clones were selected for sequencing by the method described, which does not rely on sequence similarity but, rather, conservation in the 5′-UTR.

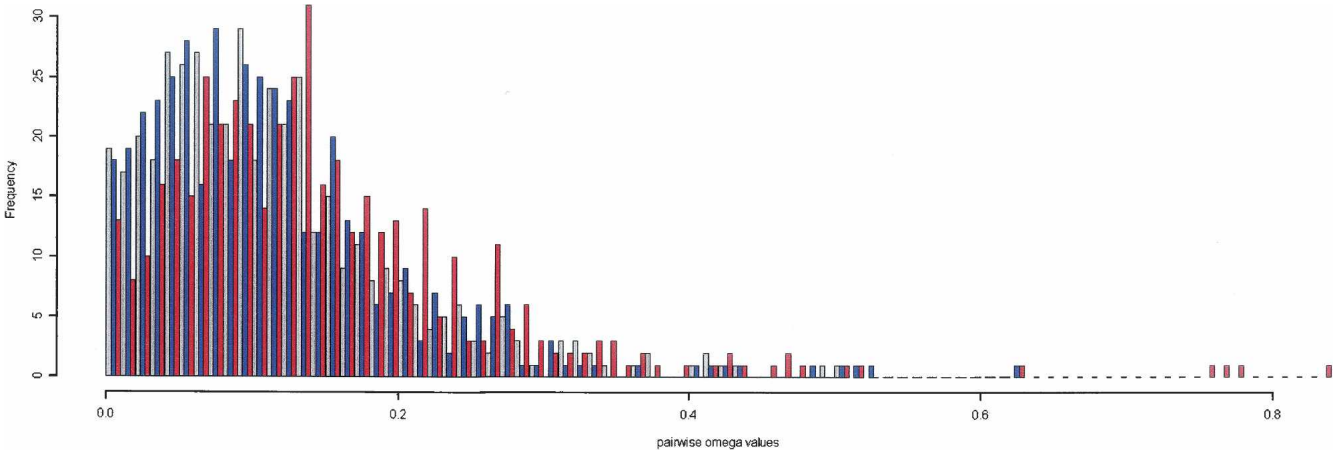

Figure 1. Frequencies of d_N/d_S ratio (ω) for pairwise comparisons between *X. laevis* and *X. tropicalis* genes. The distribution of d_N/d_S from pairwise comparisons of genes within gene trios is shown. ωs from *X. laevis* paralog pairs (shown in red) indicate a weaker selective constraint than the ω obtained from the comparisons of *X. laevis* paralogs with their *X. tropicalis* ortholog (shown in gray and blue). The ωs from both paralog–ortholog pairs follow a similar distribution with a lower median than the ω obtained from the paralogs in each trio ($P = 2.184 \times 10^{-7}$).

both copies have retained function. Interestingly, taken together, ωs from pairwise comparisons between *X. laevis* paralogs are significantly larger (Kruskal-Wallis rank sum test, $P = 2.184 \times 10^{-7}$) than either of the ωs from pairwise comparisons between each *X. laevis* gene and *X. tropicalis* paralog. These observations suggest that overall there has been relaxation of selective pressures on *X. laevis* duplicated gene pairs, allowing them more freedom to accumulate nonsynonymous substitutions.

Pairwise d_S estimates provide a relative measure of time since divergence of homologous genes and give an independent estimate of the topology of our defined trios. As expected, in all of the gene trios, the d_S between the *X. laevis* paralogs is lower than both d_Ss between each *X. laevis* paralog and its *X. tropicalis* ortholog. The trend (Fig. 2) shows a peak representing the divergence between the two species (blue and gray curves, median = 0.2915) and a peak representing the tetraploidization within the *X. laevis* genome (red curve, median = 0.2039). We calculated the effective number of codons (ENC) for all clones in this study using the codonW program (http://codonw.sourceforge.net) and did not see any evidence for an effect of codon bias on synonymous substitution rates.

To identify the function of genes for which duplicate copies are preferentially retained or lost after tetraploidization, we explored the Gene Ontology (GO) representation of *X. laevis* genes with and without paralogs. We used the High-Throughput GoMiner tool (http://discover.nci.nih.gov/gominer/htgm.jsp) (Zeeberg et al. 2005) to search for GO categories overrepresented in the clones with either active paralogs in *X. laevis* or no evidence for extant paralogs (see Methods). Several categories are enriched in the set of genes with no evidence for an extant para-

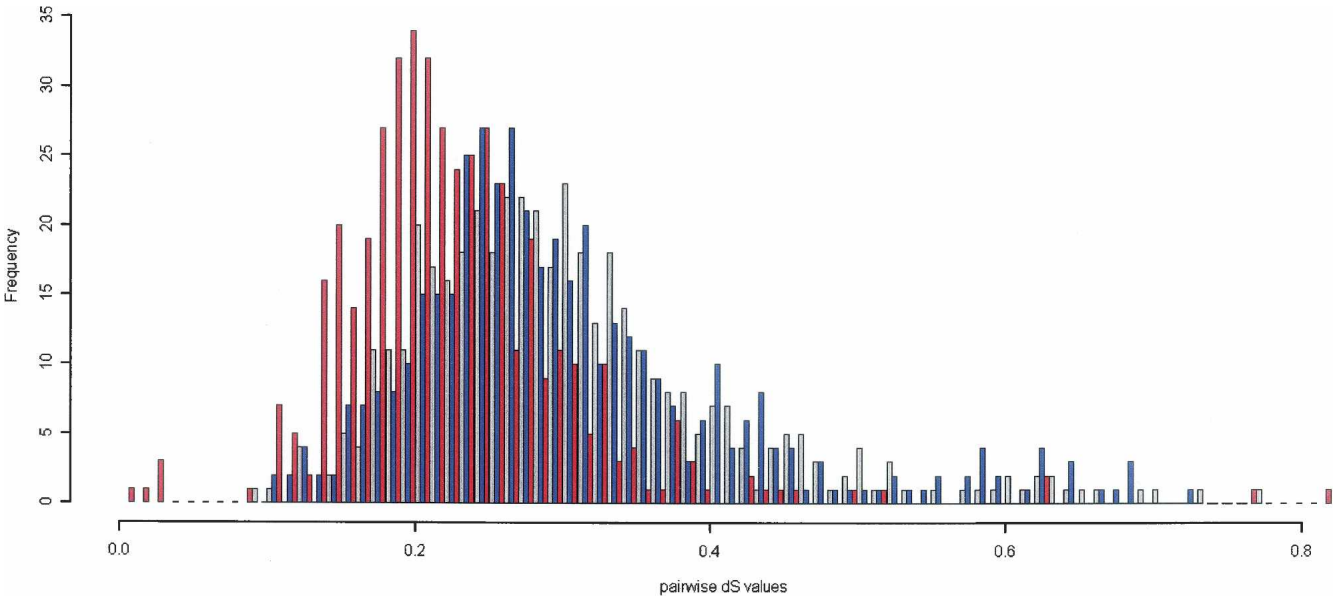

Figure 2. Distribution of d_S for pairwise comparisons between paralogs and orthologs. Distribution of d_S from pairwise comparison between *X. laevis* paralogs (red) and from pairwise comparisons between each *X. laevis* paralog from a trio with its *X. tropicalis* ortholog (blue and gray). The small number (eight in total) of d_S values that were >1 were eliminated to provide an appropriate scale.

log (Table 2; Fig. 3; Supplemental Fig. S6), but we saw no significant enrichment of GO categories in the set of genes with retained paralogs. GO categories enriched in the single-copy genes, which have lost the tetraploidization-derived paralog, can be grouped into three main clusters related to respiration (cluster 1), nucleic acid processing (cluster 2), and nucleoside metabolism (cluster 3) (Table 2; Supplemental Fig. S5).

Next, we explored whether paralogous genes in *X. laevis* have begun to subfunctionalize at the level of gene expression. Subfunctionalized genes may retain similar or identical CDS while obtaining tissue-specific functions through mutations that alter their expression (Force et al. 1999). The UniGene project contains ~27,000 expressed sequence tags each from *X. laevis* and *X. tropicalis* that are derived exclusively from tissue-specific libraries (builds 63 and 24 respectively) (Pontius et al. 2003). Taking these tissue-specific ESTs, we matched each to its corresponding *X. laevis* and *X. tropicalis* gene where matches could be unambiguously assigned (BLASTN) (Altschul et al. 1997). For the 1039 *X. laevis* allotetraploidization-derived paralogs in our gene set (see Methods), 842 had EST matches in tissue-specific libraries for both clones. Of these, 118 (14.0%) showed significant ($P < 0.05$, corrected for multiple testing) differential tissue expression (Table 3; Supplemental Table S3) consistent with subfunctionalization. Next, we ranked *X. laevis* paralogs by the number of ESTs from any tissue in which each member of the pair is expressed. This approach gave a Spearman's rank correlation coefficient of 0.49. A correlation of 1 would indicate that all paralogous gene pairs were similarly expressed, whereas a correlation near 0 would indicate that all gene pairs were expressed independently and randomly; the observed correlation of aggregate expression is intermediate to these extremes, which again is consistent with a degree of subfunctionalization in *X. laevis*. To assess how often paralogous genes have differential aggregate expression, we examined the highest and lowest deciles of aggregate expression. In the highest expression decile, 68% of genes have a paralog that is also in the highest expression decile, and in the lowest expression decile, 38% of genes have a paralog also in the lowest expression decile. This observation suggests that the function of both copies of highly expressed genes in *X. laevis* is more often conserved than is the function of sparsely expressed gene pairs.

Discussion

We report 10,967 full insert verified clones and sequences from *X. laevis* and *X. tropicalis* that have been generated by the XGC project. This clone set provides the community with a useful resource for functional genomic studies in these important model organisms. Furthermore, we have gained insight into the evolution of protein-coding sequences in a genome duplicated by species hybridization. It is well established that new genes typically arise through duplication of existing genes (Ohno 1970; Force et al. 1999). It is also known that whole genome duplications are common among the pipid frogs (Evans et al. 2004), and it is thought that these events are the major method of speciation within this genus (Kobel 1996; Evans et al. 2004). Because one member of a new duplicate gene pair is initially redundant, it is likely to quickly become a pseudogene after fixation of a null mutation within the population (Ohno 1970). The present study focuses on cDNA sequences; thus only actively transcribed genes (and potentially, rarely transcribed pseudogenes) contribute to our data set. For these expressed genes, the results of the present study indicate that many of the redundant copies have remained active in *X. laevis* over tens of millions of years. This result is consistent with previous observations in *Xenopus*, fish, and vari-

Table 2. GO Categories enriched in genes with no paralogs in *X. laevis*

GO ID	Description	Total genes in category	Genes in category with no paralog	Enrichment	False discovery rate
GO:0006139	Nucleobase, nucleoside, nucleotide, and nucleic acid metabolism	446	125	1.450582	0
GO:0046112	Nucleobase biosynthesis	6	5	4.313063	0.0520
GO:0016070	RNA metabolism	108	34	1.629379	0.0533
GO:0006350	Transcription	85	28	1.704928	0.0571
GO:0006259	DNA metabolism	153	44	1.48843	0.0650
GO:0006082	Organic acid metabolism	116	35	1.561626	0.0600
GO:0019752	Carboxylic acid metabolism	116	35	1.561626	0.0600
GO:0009112	Nucleobase metabolism	7	5	3.696911	0.0717
GO:0006260	DNA replication	43	16	1.925833	0.0692
GO:0006099	Tricarboxylic acid cycle	10	6	3.105405	0.0600
GO:0009060	Aerobic respiration	10	6	3.105405	0.0600
GO:0045333	Cellular respiration	10	6	3.105405	0.0600
GO:0046356	Acetyl-CoA catabolism	10	6	3.105405	0.0600
GO:0051187	Cofactor catabolism	13	7	2.786902	0.0589
GO:0006206	Pyrimidine base metabolism	5	4	4.140541	0.0548
GO:0006207	De novo pyrimidine base biosynthesis	5	4	4.140541	0.0548
GO:0006544	Glycine metabolism	5	4	4.140541	0.0548
GO:0007223	Frizzled-2 signaling pathway	5	4	4.140541	0.0548
GO:0019856	Pyrimidine base biosynthesis	5	4	4.140541	0.0548
GO:0006396	RNA processing	68	22	1.674483	0.0579
GO:0006412	Protein biosynthesis	256	65	1.314136	0.0784

All nonredundant Gene Ontology (Ashburner et al. 2000) terms enriched in the set of clones suspected to have no expressed paralogs in *X. laevis*. Categories with a false discovery rate (FDR) ≤0.10 are shown. Several very large generic categories were removed for clarity. No statistically significant categories were found for genes for which a tetraploidization-derived paralog was found.

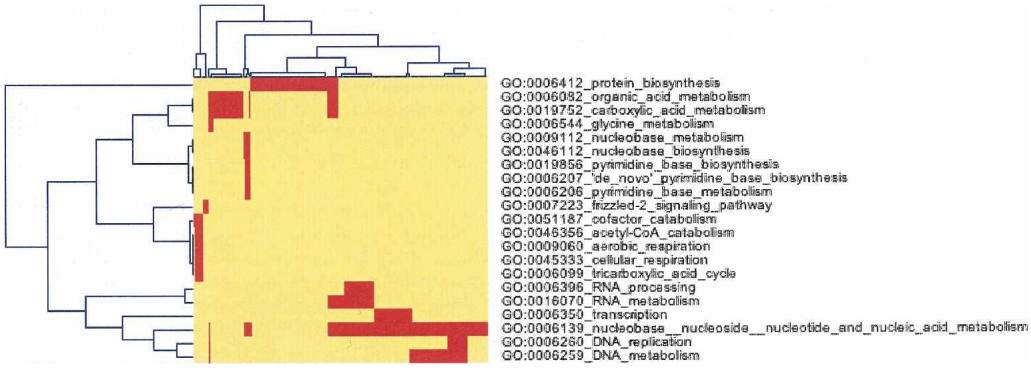

Figure 3. Clustered Image Map of genes with no paralog versus GO categories for categories with significant enrichment. Thumbnail clustered image map (CIM) of genes (*top*) versus categories (*right*) for categories with a false discovery rate (FDR) ≤0.10. Very large generic categories have been removed to improve visualization. Clustering was performed with the Genesis Client (Sturn et al. 2002; http://genome.tugraz.at/Software/GenesisCenter.html). Three major clusters can be seen. Processes involved in general metabolism (far *left*) include "cofactor catabolism," "acetyl-CoA catabolism," "aerobic respiration," "cellular respiration," and "tricarboxylic acid cycle." Processes involved in nucleic acid processing (*bottom right*) include "RNA metabolism," "transcription," "nucleobase metabolism," "DNA replication," and "DNA metabolism." The third cluster contains GO categories involved in nucleoside metabolism such as "nucleobase metabolism," "pyrimidine base biosynthesis," and "nucleobase biosynthesis." The full-size CIM in which all genes are displayed is available as Supplemental Figure S6.

ous plants (Hughes and Hughes 1993; Taylor et al. 2001; Adams and Wendel 2005).

Given that we observe 1039 paralog pairs after sampling only 8049 genes and using a conservative paralog assignment method, it is possible that retention of both copies within *X. laevis* has benefited the species. It has previously been estimated that *X. laevis* has retained about half of its duplicate genes (Bisbee et al. 1977). While there is no evidence for selective retention by functional category, we do provide evidence consistent with a model of selective loss of certain duplicate genes after genome duplication. Interestingly, the functional groups selectively lost in *Xenopus*, namely, respiration, nucleic acid processing, and nucleoside metabolism (Table 2; Supplemental Fig. S5), are included in those known to have been selectively lost after whole genome duplication in *Arabidopsis thaliana* (Maere et al. 2005). Genes in these categories, therefore, appear more likely than other genes to confer dosage sensitivity.

We searched for evidence of selection on the remaining tetraploidization-derived paralogs in *X. laevis*. Using the Yang and Nielsen maximum likelihood method (Yang 1997, 1998), we calculated pairwise d_N/d_S ratios between tetraploidization-derived *X. laevis* paralogs and between each of these *X. laevis* paralogs and

Table 3. Differential tissue expression for *X. laevis* paralogs from EST information (top 20 genes)

P-value	Gene 1	Gene 2	EST counts	Tissue	GO term	GO function	Best human BLAST hit
5.35E-138	BC072139	BC088696	21 and 562	Testes	GO:0006414	Translational elongation	Eukaryotic translation elongation factor 1 δ
1.06E-54	BC060381	BC044961	263 and 21	Testes	GO:0006826	Iron ion transport	Ferritin, heavy polypeptide 1
5.22E-23	BC054950	BC056840	111 and 10	Liver	GO:0006826	Iron ion transport	Lactoferrin (copper form)
1.59E-15	BC054151	BC041281	58 and 2	Testes	GO:0003735	Structural constituent of ribosome	Small ribosomal subunit
1.42E-10	BC046680	BC057216	1 and 37	Lung	GO:0006826	Iron ion transport	Ferritin, heavy polypeptide 1
9.31E-10	BC082829	BC053760	30 and 0	Heart	GO:0005509	Calcium ion binding	Troponin C, slow
9.31E-10	BC077795	BC072841	30 and 0	Lung	GO:0005554	Molecular function unknown	Hypothetical protein
1.42E-09	BC043895	BC054956	2 and 37	Testes	GO:0006412	Protein biosynthesis	Ribosomal protein L4
3.73E-09	BC041282	BC041307	28 and 0	Testes	GO:0006412	Protein biosynthesis	Ribosomal protein S8
4.77E-07	BC073285	BC059975	21 and 0	Kidney	GO:0005554	Molecular function unknown	Latexin
1.37E-05	BC079989	BC054174	3 and 25	Brain	GO:0005554	Molecular function unknown	No human hits
7.83E-05	BC043906	BC077529	22 and 3	Ovary	GO:0006810	Transport	Migration-inducing protein MIG8
0.000122	BC070531	BC044682	13 and 0	Kidney	GO:0006839	Mitochondrial transport	Uncoupling protein homolog
0.000259	BC092102	BC072833	1 and 15	Testes	GO:0005554	Molecular function unknown	Thioredoxin-dependent peroxide reductase 2
0.000428	BC072304	BC056053	3 and 19	Heart	GO:0018149	Peptide cross-linking	Uncoupling protein-2
0.000462	BC041242	BC073375	31 and 64	Testes	GO:0006412	Protein biosynthesis	Ribosomal protein S9
0.000488	BC081147	BC072297	11 and 0	Testes	GO:0005554	Molecular function unknown	ADP-ribosylhydrolase like 1 isoform 1
0.000488	BC045043	BC073411	11 and 0	Testes	GO:0005554	Molecular function unknown	WD40 protein Ciao1 variant
0.000488	BC078599	BC074203	11 and 0	Testes	GO:0006334	Nucleosome assembly	H2A histone family, member Q
0.000488	BC059972	BC044112	11 and 0	Heart	GO:0006468	Protein amino acid phosphorylation	Pyruvate dehydrogenase kinase, isoenzyme 4

A summary of the top 20 paralog pairs with evidence for subfunctionalization (sorted with most significant *P*-value at the top). The EST count for gene 1 and gene 2 of each paralog pair in the tissue showing differential expression is shown. The Gene Ontology (GO) category (Ashburner et al. 2000) that best describes the putative function of each gene pair is included in addition to the name of the best *Homo sapiens* BLASTP hit. All 118 cases of potential subfunctionalization are supplied in Supplemental Table S3.

its *X. tropicalis* ortholog. While none of the paralog pairs observed in this study provides direct evidence of positive Darwinian selection, it must be noted that the evolutionary distance between *X. laevis* and *X. tropicalis* is approaching the practical limit of detecting positive selection by the d_N/d_S test (Hughes et al. 2000). Also, all d_N/d_S ratios are averages across the length of each gene, and may mask situations in which positive selection has only occurred in a site-specific manner. Interestingly, overall, d_N/d_S ratios between tetraploidization-derived *X. laevis* paralogs were significantly elevated relative to d_N/d_S ratios between *X. laevis/X. tropicalis* orthologs (Fig. 1). This result is consistent with the notion that gene duplication has, to a small but measurable degree, freed expressed *X. laevis* sequence from functional constraint.

A subset of *X. laevis* paralogs also shows strong evidence for subfunctionalization, as indicated by the differential expression observed between ~14% of the paralog pairs analyzed. The paralogs that show differential expression at the highest significance levels are overrepresented in a few molecular functions (Table 3). In the top 20 subfunctionalized pairs, five are involved in protein translation, and five are potentially involved in either iron or calcium transport. With the application of Spearman's rank to paralog pairs, we have detected a global trend of many of the paralog pairs toward subfunctionalization. The observation of numerous examples in which two allotetraploidization-derived gene copies show differential expression lends support to the theory of regulatory evolution (King and Wilson 1975), which holds that evolution is mediated by modification of gene expression patterns as well as by coding sequence changes.

Methods

The analyses described here were performed on the 10,967 *Xenopus* sequences available from the XGC Web site as of September 9, 2005 (XGC homepage, http://xgc.nci.nih.gov/).

Selection of clones for sequencing

Candidate clones for full insert sequencing were selected as previously described (Gerhard et al. 2004). Identification of additional putatively novel *Xenopus* clones was as follows. Pairs of orthologous (in the reciprocal best-match sense) *X. tropicalis* and *X. laevis* sequences were aligned, and the 5′-most ATG in the sequence was used to divide the alignment. Then 674 pairs of characterized genes were examined to determine mean conservation in CDS and 5′-UTR (92.9% and 89.7%, respectively), as well as to compute the variance of the conservation difference between CDS and UTRs. Sequences showing any increase in conservation (corresponding to a sequence conservation of 1.6 standard deviations below the mean) and with a 5′-UTR of at least 25 nt were selected for sequencing.

Paralog and ortholog assignment

For each species, an all-by-all BLASTN search was executed, and any sequences with >98% nucleic acid identity over 90% of their length were removed as redundant copies of the same clone, allelic copies, or recent gene duplicates. Next, unmatched sequences from the *X. laevis* reciprocal BLASTN and those with more than one significant ($e < 10^{-20}$) match were excluded, leaving 1354 putative allotetraploidization-derived *X. laevis* paralogs with 93.1% ± 2.72% nucleic acid identity (mean ± SD). Finally, a reciprocal BLASTP search of each *X. laevis* paralogous against the set of *X. tropicalis* predicted nonredundant proteins (EnsEMBL homepage, http://www.ensembl.org/*Xenopus*_tropicalis) identified 1039 paralog pairs for which both clones had a common best *X. tropicalis* hit. Note that the above approach automatically prevents us from considering gene families with more than 1:1 paralog mapping and prevents assignment of tetraploidization-derived paralogs to those that formed prior to the speciation between *X. laevis* and *X. tropicalis*. Also, by considering the best nonredundant match of each *X. laevis* clone sequence, we assigned the most recent paralog pairs, assuming that few paralog pairs have formed since the tetraploidization event. Of the 1039 gene trios, 445 have an ortholog that is represented in the set of 2918 *X. tropicalis* full ORF cDNAs. Our d_N/d_S analysis is based on these 445 trios, and we included the additional paralog pairs with no *X. tropicalis* XGC ortholog in the retained paralog and subfunctionalization analyses.

Multiple sequence alignments

We performed all protein alignments using CLUSTALW with default parameters (Thompson et al. 1994). We used the RevTrans program (Wernersson and Pedersen 2003) to produce codon-aware alignments for more accurate predictions of substitution rates. Files were formatted from RevTrans output (FASTA) to codeml input format with a Perl script.

d_N/d_S estimation

We performed pairwise d_N,d_S and ω calculations using the yn00 component of the PAML package. We also calculated d_N/d_S with the maximum likelihood method of the codeml program (Yang et al. 1997) in order to perform likelihood ratio tests. Each d_N/d_S calculation was performed twice. In one run, the d_N/d_S was fixed at 1, and in the other run it was free to vary. Twice the difference between the two log likelihood values was compared to the χ^2 distribution with 1 degree of freedom for rejection of the hypothesis that $d_N/d_S = 1$ ($P < 0.05$).

Gene Ontology analysis of retained paralogs

The aim of this analysis was to determine whether genes of specific functions have been selectively retained in multiple copies within the *X. laevis* genome or whether redundant copies have been selectively lost after duplication. Clones with no expressed paralog were defined by first assigning a UniGene cluster to each clone as described above. We limited our analysis to clones for which this corresponding UniGene cluster contained at least 20 ESTs to minimize the potential of assigning genes with low expression as single-copy genes. Single-copy genes were then assigned by the absence of a second significant ($e < 10^{-30}$) BLASTN hit in UniGene. We searched for GO categories enriched in either the 1527 single-copy genes or the 1039 dual-copy genes using High-Throughput GoMiner (http://discover.nci.nih.gov/gominer/htgm.jsp) (Zeeberg et al. 2005) and an in-house customized version of the GO Consortium (Ashburner et al. 2000) database. This database was populated with all GO annotations assigned by Interproscan (Apweller et al. 2001) to the *X. laevis* genes. The statistical significance of enrichment was based on a false discover rate (FDR) threshold value of 0.10, using High-Throughput GoMiner's correction for multiple testing.

Tissue expression comparisons

We matched every clone to its best UniGene cluster by a BLASTN of each clone against a representative EST from each UniGene cluster in the files Xl.seq.uniq and Str.seq.uniq downloaded from NCBI (UniGene Download page, ftp://ftp.ncbi.nih.gov/

repository/UniGene). We then performed a second BLASTN search of each clone against all ESTs in its representative cluster and only considered matches with alignment lengths >200 bp and percent identities >90%. As some (~20) of the genes shared a common UniGene cluster, we only used ESTs from those clusters if they could be unambiguously assigned to one copy by differential percent identity. In all other cases, all reads from the best UniGene cluster for each clone were used. Using the library name for each EST, we determined the source tissue of the EST as defined in the files Xl.lib.info and Str.lib.info. A Perl script counted the total number of ESTs of each gene found in each tissue. Using a previously described Bayesian method (Audic and Claverie 1997), we assigned *P*-values to each gene pair showing a significant difference in expression ($P < 0.05$). As the variability in expression patterns was strongest in tissues where the genes are highly expressed, we chose to only compare the expression of the two paralogs in the tissue in which they showed to be most highly expressed (highest EST count). A Bonferroni correction was used to compensate for multiple hypothesis testing where applicable, although a maximum of two tests were done for each paralog pair.

Acknowledgments

This project has been funded in part with Federal funds from the National Human Genome Research Institute, National Institutes of Health, under contract No. U01 HG002155-06S1. This research was supported (in part) by the Intramural Research Program of the National Cancer Institute (NCI) of the National Institutes of Health (NIH). This work would also not have been possible without the cDNA libraries and EST sequences provided by various groups and we specifically thank Igor Dawid and Thomas Sargent (NICHD), Donald Brown (Carnegie Institute), Bruce Blumberg (UC at Irvine), and Robert Grainger (UVA). Work by C.P.P. was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48. We thank Diana Palmquist and Elizabeth Chun, who assisted with finishing some of the clones. R.A.H., S.J.M.J., and M.A.M. are Michael Smith Foundation for Health Research Scholars.

References

Adams, K.L. and Wendel, J.F. 2005. Polyploidy and genome evolution in plants. *Curr. Opin. Plant Biol.* **8:** 135–141.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Apweller, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D.R., et al. 2001. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* **29:** 37–40.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. The Gene Ontology Consortium. 2000. Gene Ontology: Tool for the unification of biology. *Nat. Genet.* **25:** 25–29.

Audic, S. and Claverie, J.M. 1997. The significance of digital gene expression profiles. *Genome Res.* **7:** 986–995.

Bisbee, C.A., Baker, M.A., Wilson, A.C., Irandokht, H.A., and Fischberg, M. 1977. Albumin phylogeny for clawed frogs. *Science* **195:** 785–787.

Blackshear, P.J., Lai, W.S., Thorn, J.M., Kennington, E.A., Staffa, N.G., Moore, D.T., Bouffard, G.G., Beckstrom-Sterberg, S.M., Touchman, J.W., de Fatima Bonaldo, M., et al. 2001. The NIEHS *Xenopus* maternal EST project: Interim analysis of the first 13,879 ESTs from unfertilized eggs. *Gene* **267:** 71–87.

Butterfield, Y.S., Marra, M.A., Asano, J.K., Chan, S.Y., Guin, R., Krzywinski, M.I., Lee, S.S., MacDonald, K.W., Mathewson, C.A., Olson, T.E., et al. 2002. An efficient strategy for large-scale high-throughput transposon-mediated sequencing of cDNA clones. *Nucleic Acids Res.* **30:** 2460–2468.

De Sa, R.O. and Hillis, D.M. 1990. Phylogenetic relationships of the pipid frogs *Xenopus* and *Silurana*: An integration of ribosomal DNA morphology. *Mol. Biol. Evol.* **7:** 365–376.

Evans, B.J., Kelley, D.B., Tinsley, R.C., Melnick, D.J., and Cannatella, D.C. 2004. A mitochondrial DNA phylogeny of African clawed frogs: Phylogeography and implications for polyploid evolution. *Mol. Phylogenet. Evol.* **33:** 197–213.

Ewing, B. and Green, P. 1998. Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res.* **8:** 186–194.

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

Gerhard, D.S., Wagner, L., Feingold, E.A., Shenmen, C.M., Grouse, L.H., Schuler, G., Klein, S.L., Old, S., Rasooly, R., Good, P., et al. 2004. The status, quality, and expansion of the NIH Full-Length cDNA Project: The Mammalian Gene Collection (MGC). *Genome Res.* **14:** 2121–2127.

Gilchrist, M.J., Zorn, A.M., Voigt, J., Smith, J.C., Papalopulu, N., and Amaya, E. 2004. Defining a large set of full-length clones from a *Xenopus tropicalis* EST project. *Dev. Biol.* **271:** 498–516.

Gordon, D., Abajian, C., and Green, P. 1998. Consed: A graphical tool for sequence finishing. *Genome Res.* **8:** 195–202.

Graf, J. and Kobel, H. 1991. Genetics of *Xenopus laevis*. *Methods Cell Biol.* **36:** 663–669.

Hirsch, N., Zimmerman, L., and Grainger, R. 2002. *Xenopus*, the next generation: *X. tropicalis* genetics and genomics. *Dev. Dyn.* **225:** 422–433.

Hughes, M.K. and Hughes, A.L. 1993. Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*. *Mol. Biol. Evol.* **10:** 1360–1369.

Hughes, M.K., Green, J.A., Garbayo, J.M., and Roberts, R.M. 2000. Adaptive diversification within a large family of recently duplicated, placentally expressed genes. *Proc. Natl. Acad. Sci.* **97:** 3319–3323.

King, M.C. and Wilson, A.C. 1975. Evolution at two levels in humans and chimpanzees. *Science* **188:** 107–116.

Klein, S.L., Strausberg, R.L., Wagner, L., Pontius, J., Clifton, S.W., and Richardson, P. 2002. Genetic and genomic tools for *Xenopus* research: The NIH *Xenopus* Initiative. *Dev. Dyn.* **225:** 384–391.

Kobel, H.R. 1996. Allopolyploid speciation. In *The biology of Xenopus* (eds. R.C. Tinsley and H.R. Kobel), pp. 391–401. Clarendon Press, Oxford.

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., and Van de Peer, Y. 2005. Modeling gene and genome duplication in eukaryotes. *Proc. Natl. Acad. Sci.* **102:** 5454–5459.

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

Pontius, J.U., Wagner, L., and Schuler, G.D. 2003. UniGene: A unified view of the transcriptome. In *The NCBI Handbook*, pp. 857–868. National Center for Biotechnology Information, Bethesda, MD.

Strausberg, R.L., Feingold, E.A., Grouse, L.H., Derge, J.G., Klausner, R.D., Collins, F.S., Wagner, L., Shenmen, C.M., Schuler, G.D., and Altschul, S.F. 2002. Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl. Acad. Sci.* **99:** 16899–16903.

Sturn, A., Quackenbush, J., and Trajanoski, Z. 2002. Genesis: Cluster analysis of microarray data. *Bioinformatics* **18:** 207–208.

Taylor, J.S., Van de Peer, Y., Braasch, I., and Myer, A. 2001. Comparative genomics provides evidence for an ancient genome duplication event in fish. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **356:** 1661–1679.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.

Thornton, K. and Long, M. 2002. Rapid divergence of gene duplicates on the *Drosophila melanogaster* X chromosome. *Mol. Biol. Evol.* **19:** 918–925.

Wernersson, R. and Pedersen, A.G. 2003. RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Res.* **31:** 3537–3539.

Wilson, R.K. and Mardis, E.R. 1997. Fluorescence-based DNA sequencing. In *Genome analysis: A laboratory manual: Analyzing DNA* (eds. B. Birren et al.), Vol. 1, pp. 301–395. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Yang, Z. 1997. PAML, a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13:** 555–556.

———. 1998. Likelihood ratio tests for detecting positive selection and

application to primate lysozyme evolution. *Mol. Biol. Evol.* **15:** 568–573.
Yang, G., Stott, J.M., Smailus, D.M., Barber, S.A., Balasundaram, M., Marra, M.A., and Holt, R.A. 2005. High-throughput sequencing: A failure mode analysis. *BMC Genomics* **6:** 2.
Zeeberg, B.R., Qin, H., Narasimhan, S., Sunshine, M., Cao, H., Kane, D.W., Reimers, M., Stephens, R.M., Bryant, D., Burt, S.K., et al. 2005. High-Throughput GoMiner, an 'industrial-strength' integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of Common Variable Immune Deficiency (CVID). *BMC Bioinformatics* **6:** 168.
Zhang, L., Vision, T., and Gaut, B. 2002. Patterns of nucleotide substitution among simultaneously duplicated gene pairs in *Arabidopsis thaliana*. *Mol. Biol. Evol.* **19:** 1464–1473.

Received October 28, 2005; accepted in revised form March 16, 2006.