LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-JRNL-216808

Sequencing and analysis of 10967 full-length cDNA clones from Xenopus laevis and Xenopus tropicalis

R. D. Morin, E. Chang, A. Petrescu, N. Liao, R. Kirkpatrick, M. Griffith, Y. Butterfield, J. Stott, S. Barber, R. Babakaiff, C. Matsuo, D. Wong, G. Yang, D. Smailus, M. Brown-John, M. Mayo, J. Beland, S. Gibson, T. Olson, M. Tsai, R. Featherstone, S. Chand, A. Siddiqui, W. Jang, E. Lee, S. Klein, C. Pennacchio, R. M. Myers, E. D. Green, L. Wagner, D. Gerhard, M. Marra, S. J. M. Jones, R. Holt

November 3, 2005

Genome Research

Title: Sequencing and analysis of 10967 full-length cDNA clones from *Xenopus* laevis and *Xenopus* tropicalis

Ryan D. Morin, Elbert Chang, Anca Petrescu, Nancy Liao, Robert Kirkpatrick, Malachi Griffith, Yaron S. Butterfield, Jeffrey Stott, Sarah Barber, Ryan Babakaiff, Corey Matsuo, David Wong, George S. Yang, Duane E. Smailus, Mabel Brown-John, Michael Mayo, Jaclyn Beland, Susan Gibson, Teika Olson, Miranda Tsai, Ruth Featherstone, Steve Chand, Asim S. Siddiqui, Wonhee Jang, Ed Lee, Steven L. Klein, Christa Prange, Richard M. Myers, Eric D. Green, Lukas Wagner, Daniela S. Gerhard, Marco A. Marra, Steven J.M. Jones, Robert A. Holt (1).

(1) Communicating Author.

Keywords: Darwinian selection, molecular evolution, regulatory evolution, full ORF clone, full length cDNA, subfunctionalization, BC Genome Sciences Centre

Manuscript Type: RESOURCE

Abstract

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 an analysis of 10967 clones (8049 from *X. laevis* and 2918 from *X. tropicalis*). The clone set contains 2013 orthologs between *X. laevis* and *X. tropicalis* as well as 1795 paralog pairs within *X. laevis*. 1199 are in-paralogs, believed to have resulted from an allotetraploidization event approximately 30 million years ago, and the remaining 546 are likely out-paralogs that have resulted from more ancient gene duplications, prior to the divergence between the two species. We do not detect any evidence for positive selection by the Yang and Nielsen maximum likelihood method of approximating d_N/d_S. However, d_N/d_S for *X. laevis* in-paralogs is elevated relative to *X. tropicalis* orthologs. This difference is highly significant, and indicates an overall relaxation of selective pressures on duplicated gene pairs. Within both groups of paralogs, we found evidence of subfunctionalization, manifested as differential expression of paralogous genes among tissues, as measured by EST information from public resources. We have observed, as expected, a higher instance of subfunctionalization in out-paralogs relative to in-paralogs.

Introduction

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). Study of embryonic development in *Xenopus* has provided insights into many salient aspects of vertebrate development that would be difficult to study in vertebrate systems (Gilchrist et al. 2004). However, the study of *Xenopus* genetic material is difficult due to an allotetraploidization event in the *Xenopus* lineage estimated at 30 million years ago (MYA), which has generated a more complex genome in all extant *Xenopus* species except for *Xenopus tropicalis*, whose genome remains diploid (Hirsch et al. 2002; Graf and Kobel 1991). 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).

A number of groups have performed large-scale EST studies on libraries from various tissues of both *Xenopus laevis* and *Xenopus tropicalis* (Klein et al., 2002, Blackshear et al. 2004; Gilchrist et al. 2004). However, for analysis of transcripts and gene structure, the quality of data and coverage provided by EST reads can be limiting. Fully-sequenced 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 10967 full-length cDNA clones (8049 from *X. laevis* and 2918 from *X. tropicalis*) from cDNA libraries that were constructed using RNA from numerous tissues and whole animals in various developmental stages. In

addition to the various ongoing cDNA projects in Xenopus, the Joint Genome Institute (JGI) recently released a draft genome sequence for *X. tropicalis* (unpublished data, http://genome.jgi-psf.org/Xentr3/Xentr3.home.html). The EnsEMBL ab-initio gene prediction pipeline has been applied and the resulting annotations are publicly available online (www.ensembl.org/*Xenopus*_tropicalis; Birney et al. 2004; Hubbard et al. 2005). These resources are complementary in their nature, each providing information that provides a better understanding of the other.

The availability of full-ORF sequences for a large set of *Xenopus* clones provides a unique opportunity to study the mechanisms of molecular evolution within and between two closely related species. The *X. laevis* genome comprises 36 chromosomes (2n) whereas that of *X. tropicalis* comprises 10 chromosomes pairs in the diploid state (Hirsch et al. 2002). The duplicated genome of *X. laevis* is most likely the result of an allotetraploidization event, involving the combination of the entire chromosome set from two closely related ancestral *Xenopus* species (Kobel 1996; Evans et al. 2004). This process appears to be a common mode of speciation amongst the clawed (pipid) frogs (Evans et al. 2004). The putative *X. laevis* ancestral hybridization event created a full set of paralogs, each from one of the parent species involved in the mating. The redundancy of the tetraploid *X. laevis* genome is proposed to afford 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 we have undertaken a comparison of orthologs between the two *Xenopus* species and of the paralogs within *X. laevis* in an attempt to gain insight into the selective pressures on protein coding sequences that have occurred since the divergence of these two species from a common ancestor. We used the

EnsEMBL gene predictions for *X. tropicalis* to group *X. laevis* paralogs into in-paralogs and out-paralogs. In our analysis we determined that at the protein-coding level, all of these homologous genes appear to be evolving under purifying and not positive selection. In situations where two gene copies remain active in the genome, one would expect to find that they no longer fully overlap in function. The mechanism that would afford this circumstance, known as subfunctionalization, involves mutations that alter the expression profiles of one or both paralogs (Force et al. 1999). We have found significant evidence for the occurrence of this process between paralogs, with older paralogs having formed prior to the divergence of these two *Xenopus* species, demonstrating a higher frequency of subfunctionalization.

Results

Clones for EST sequencing were selected at random from 30 cDNA libraries from various tissues and developmental stages of *X. laevis* and *X. tropicalis* (tables 1 and 2) and the ESTs were generated by the National Intramural Sequencing Center, Washington University Genome Sequencing Center and Agencourt Bioscience Corporation (Gerhard et al. 2004). The clones for full-insert sequencing were picked by an algorithm developed to identify clones will complete open reading frame (Klein et al., 2002, Gerhard et al., 2004). Each clone was fully sequenced to a consensus phred score of no less than 30 (Ewing and Green 1998) at each consensus position using previously described methods (Butterfield et al. 2002). Coding DNA segment (CDS) annotation of the clones was performed as previously described (Gerhard et al. 2004); only clones predicted to be complete were annotated with an MGC identifier, the other clones were identified with only an IMAGE identifier. All clone sequences were submitted to

Genbank and the clones are available through the IMAGE distribution network. In addition to the methods used to identify candidate clones from EST sequences for other organisms in MGC, we sought to identify candidate clones which 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 employed (see Methods) assumes that a stronger conservation will be observed in the CDS of paralogous genes than in the 5' untranslated region (UTR). 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 identified in Table 9. While there are 7 proteins with no significant hits to any mammalian protein, there are also 6 proteins with alignment scores of more than two standard deviations below the mean value to the most closely related human protein (Mean 73%, 13.8% standard deviation). To characterize the distance between each predicted protein and orthologous proteins outside pipidae, the percent identity with the most similar human protein aligning over at least 100 amino acids is shown.

We performed the following analysis on 10967 clones, the entirety of the *X. laevis* and *X. tropicalis* full-length cDNA sequences found on the XGC web page as of September 9, 2005 (XGC homepage, http://xgc.nci.nih.gov/). For each species, we performed a BLASTN search of each clone sequence against all sequences for that species to determine redundancy within the clone set. Any clone sequence that matched another with >98% identity in an alignment of at least 90% of its length was considered redundant. Clones resulting from alternative splicing were later removed as redundant during paralog assignment. This allowed us to classify 231 of the *X. laevis* clones and

148 of the *X. tropicalis* clones as redundant. We excluded all redundant clones and retained the longest representative clone for further analyses (Table 3). There were 10588 clones remaining after removal of redundant sequences. To assign homology among genes from the two *Xenopus* species we employed a modified reciprocal best hit (rbh) method (see Methods). We identified 2013 strict ortholog pairs between the two *Xenopus* species (Table 4). Our method allowed discernment between the likely tetraploidization-derived *X. laevis* paralogs (in-paralogs) as well as members of gene families that formed prior to the divergence of these two species (out-paralogs) (Remm et al. 2001). Of the 1745 paralog pairs we identified within the *X. laevis* clones, we found 546 to be out-paralogs (O'Brien et al. 2005), while the other 1199 we assumed to be in-paralogs, that is, two paralogous copies resulting from the tetraploidization event (see Methods). In other words, 546 of the *X. laevis* paralog pairs produce significant alignments with more than one gene in *X. tropicalis*. Of the in-paralogs, which are likely due to the doubling of the *X. laevis* ploidy, 437 have an ortholog represented in the *X. tropicalis* set of clones as well as a good ortholog in *Takifugu rubripes*. We assigned a *F. rubripes* homolog to each gene trio to provide an outgroup for phylogenetic-based estimation of d_N/d_S ratios between homologs. A *T. rubripes* ortholog was considered a good outgroup if it was identified by reciprocal blast (see Methods) hit with at least 2 of the 3 genes in the *Xenopus* gene trio. In addition, we assigned human orthologs in a similar fashion for the purpose of inferring function, as the Gene Ontology information for the EnsEMBL *X. tropicalis* genes was incomplete (see Methods),(Ashburner et al. 2000). We retrieved the GO category for biological process for every *X. laevis* and *X. tropicalis* gene; the result for the *X. laevis* clones is summarized in Figure 1.

To explore 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 (Yang et al. 1994; PAML software release 3.14). This method allows inference of evolutionary selection for mutations using the ratio of non-synonymous (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 (omega) > 1 is evidence for positive selection acting to modify the function of a gene (Thornton and Long, 2002; Zhang et al. 2002; Wong et al. 2004) whereas an omega significantly <1 suggests negative or purifying selection where functional constraint on the gene product has restricted non-synonymous mutation. None of the pairwise comparisons between gene homologs resulted in an omega significantly > 1 (Figure 2). The mean omega for *X. laevis* genes in our set is 0.1258, while the mean omega for *X. tropicalis* genes is 0.0918.

A strong codon bias can often reduce the overall synonymous substitution rate of a gene, resulting in an effective positive selection against some synonymous changes (Gu et al. 2002). We calculated the effective number of codons (ENC) for all clones in this study using the codonw program (http://codonw.sourceforge.net). The average ENC was 53.73 with a standard deviation of 3.75. A complete lack of codon bias is represented by an ENC of 61, wheras a strong bias will approach 20 (Fuglsang 2004). The average ENC observed suggests that codon bias has not been a major contributing factor to the synonymous substitution rates. Thus there is strong evidence for purifying selection acting on duplicated *X. laevis* genes, and that in general both copies of paralagous pairs have retained function.

A one-tailed t-test suggested that the mean omega observed for the set of in-paralogs is significantly higher than that obtained in comparing orthologous genes ($P = 3.58 \times 10^{-10}$). This indicates that although there is no evidence for positive selection on any homologous gene pair observed in this study, there appears to be an overall relaxation of selective pressures on duplicated gene pairs. It should be noted that the methods of homolog assignment (filtering out shorter alignments) and transcript sequencing in general creates a strong bias against pseudogenes that are no longer active. Nonetheless, there were 2376 *X. laevis* sequenced clones that are not included in the XGC collection of full-length transcripts due to mutations causing nonsense codons within the CDS. After removal of redundancy, we found 423 of these have functional paralogs within the *X. laevis* clones that were not assigned paralogs by comparison to the full-ORF set of transcripts. The average percent identity between these paralogs is 91.5%, the same as we observed for the in-paralogs. In addition to these clones, we detect some evidence of pseudogene-like selection on some of the genes without CDS disruptions. Some of the likelihood ratio tests failed to prove significant difference from 1 (see Methods). There are two possible explanations for this result. In some cases, it was obvious that, due to minimal sequence divergence between the homologs, the sequence data was not informative enough to provide an accurate estimate of d_N/d_S ;the result was usually extreme over- or underestimates of d_N/d_S. The other cause is that the real d_N/d_S value is near 1, indicating neutral selection. As expected, the genes of this type are dominated by *X. laevis* genes, as they are under relaxed selective pressure (Table 5).

To identify the function of genes for which both copies are preferentially retained, we explored the (GO) representation of the *X. laevis* in-paralog pairs. This approach

assumes that sampling bias did not influence our discovery of paralogous genes. We supported this assumption by tabulating the number of ESTs observed for each gene and comparing them between groups. The genes for which we found paralogous copies do not have significantly greater numbers of ESTs associated with them in UniGene as compared to the genes with no paralogs in our set of clones (t-test, $P = 2.75*10^{-5}$), indicating that their expression profiles are not inherently different than the average. We used the GoMiner tool to search for GO categories over-represented in the clones with active paralogs in *X. laevis* (Zeebert B et al. 2003). The genes for which both copies are preferentially retained appear to be dominated by those involved in protein degradation and intracellular signaling (Table 6).

We also devised a method using existing EST data to explore whether paralogous genes in *X. laevis* have begun to subfunctionalize at the level of gene expression (Force et al. 1999). Subfunctionalized genes may retain similar or identical CDS while obtaining tissue-specific functions due to mutations that alter their expression (Force et al. 1999). We addressed this possibility by evaluating the expression profiles of the paralogous *X. laevis* genes. The Unigene project contains approximately 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. 1997). The remainder of the ESTs in UniGene are derived from libraries constructed from whole bodies/embryos. Taking these tissue-specific ESTs, we matched each *X. laevis* and *X. tropicalis* FL-cDNA to ESTs where matches could be unambiguously assigned (BLASTN). For the 1199 *X. laevis* in-paralogs discussed previously, 839 had EST matches in tissue-specific libraries for both clones. Of these, 106 (12.6%) showed significant (alpha= 0.05) differential

tissue expression (Table 7). A higher proportion (20.6%) of the putative out-paralogs, which were duplicated prior to the speciation between *X. laevis* and *X. tropicalis*, demonstrate divergence between expressed tissues (Table 8).

A complementary approach to identifying functionally distinct paralogous genes is through aggregate rather than tissue-specific EST expression. Ranking paralogous genes by the number of ESTs from any tissue in which each member of the gene pair is expressed allows calculation of Spearman's rank correlation coefficient of overall expression between paralogous genes. This quantity is 0.49 when applied to in-paralogs and 0.35 for the out-paralogs. A correlation of 1 would indicate that all paralogous gene pairs were similarly expressed, whereas a correlation near 0 would indicate all gene pairs were expressed independently and randomly; the observed correlation of aggregate expression is intermediate to these extremes, providing a measure of how often paralogous gene pairs are similarly expressed. To assess how often paralogous genes have differential aggregate expression, we examined the highest and lowest deciles of aggregate expression. 68% of in-paralogs in the highest expression decile have a paralog which is also in the highest expression decile, and 38% of genes in the lowest expression decile have a paralog also in the lowest expression decile, suggesting 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

A large number of full-insert sequenced clones from these two closely related species, *X. laevis* and *X. tropicalis*, provides the community with a useful resource for functional genomic studies in these important model organisms. The clones presented

here were sequenced as part of the *Xenopus* Gene Collection project and are available from the website for that project (http://xgc.nci.nih.gov/). All clone sequences are available from Genbank. These sequences are an invaluable resource for the ongoing genome annotation effort. Further, the divergence of the parental *Xenopus* species 30 MYA in conjunction with the tetraploidization of the parental *X. laevis* genome provides a unique opportunity for molecular evolution studies and the sequences presented here provide the best resource to date for such a study.

The small number of novel proteins identified as well as the overlap with proteins identified from protein or mRNA comparison suggests that there are few proteins present in the XGC cDNA collection that are structurally distinct from previously identified proteins. Novel proteins present in either *X. laevis* or *X. tropicalis* cDNA libraries (but not both) could not be identified by this method. This method would also not allow detection of transcripts with very short 5' UTR, or very well conserved 5' UTR. Extrapolating from the 13 novel proteins identified from the 2013 orthologous pairs available, and noting that 77% of genes have differential conservation of 5' UTR and CDS, we estimate that 200 of this type of protein would be detected with completely-sampled transcriptomes.

Duplications are the most common precursor of new genes and their neofunctionalization or subfunctionalization into gene families (Force et al. 1999; Ohno 1970). It is known that whole genome duplications are common amongst 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). The events following gene duplication over an unknown (and likely variable) timescale determine the fate of one of the gene

copies. As a duplicate gene pair, is subjected to reduced selective pressure, it is thought that one version is likely to quickly become a pseudogene or completely inactivated after fixation of a null mutation within the population (Ohno 1970). We have found evidence for recent pseudogenization of 423 genes in *X. laevis*, as determined by expression of one functional copy and a highly similar paralog with one or more CDS interuptions.

The results of this study indicate that many of these redundant genes have remained active in *X. laevis* over a large time period. This result supports previous observations in *Xenopus*, fish and various plants (Hughes et al. 1993;Taylor et al. 2001; Adams and Wendel 2005). For the most part, the paralogs represented in this set appear to be undergoing purifying selection. The genes seen in Figure 2 (right tail) with a higher d_N/d_S ratio are possibly less functionally constrained and may eventually gain novel functions (or become pseudogenes). Calculation of d_N/d_S for *X. laevis* genes and the orthologous *X. tropicalis* gene by the Yang and Nielsen maximum likelihood method gave different distributions of omega, the means of which are significantly different. This result appears to indicate that over the timescale observed here the large majority of genes are retaining the same function in both *X. laevis* and *X. tropicalis* but divergence of protein sequence is more relaxed in paralogous gene pairs in the *X. laevis* genome than between the *Xenopus* orthologs we compared in this study.

Construction of cDNA libraries is biased towards capturing functionally-conserved genes as only actively transcribed genes (and transcribed pseudogenes) would have been captured in our libraries. Any unconstrained gene copies could already have become inactive pseudogenes since the tetraploidization event; supporting this, we encountered pseudogenes that are still actively transcribed. We have observed an

apparent overrepresentation of genes involved in protein degradation and intracellular signaling maintained as active paralog pairs. As a large proportion of paralogs have shown evidence of subfunctionalization in this study, it is possible that genes involved in these processes provide a selective advantage when maintained in duplicate. It may also be that mutation in these genes is more likely to act in a dominant negative manner.

None of the paralog pairs observed in this study show evidence of positive Darwinian selection. Because there is strong evidence for net purifying selection, existing in multiple copies has not freed the coding sequence of these genes from functional constraints. Regulatory regions, however, may be evolving, and supporting subfunctionalization at the expression level. The set of in-paralogs believed to have formed from tetraploidization show strong evidence for subfunctionalization, as indicated by the differential expression profiles observed in 12.6% of the paralog pairs analyzed. Paralogs that formed prior to the speciation between *X. laevis* and *X. tropicalis* (out-paralogs) show greater differential expression than the more recent in-paralogs. With the application of Spearman's rank to paralog pairs, a similar trend is observed, with the in-paralogs showing more concordance in expression patterns than the older out-paralogs. Further analyses using more high-throughput techniques, such as cDNA or affymetrix microarrays, which will be greatly facilitated by the cDNA sequences produced in this project, should provide more insights into this and other evolutionary processes. The high-quality resource of *Xenopus* FL-cDNA sequences and clones, available on-demand from the IMAGE Consortium distributors, will enable future investigations by both evolutionary and developmental biologists.

Methods

Sequencing of cDNAs

Clones were sequenced by transposon-mediated sequencing and finished with primer walking to close gaps and improve low quality regions. Methods for sequencing performed at the BC Genome Sciences Centre are as previously described (Butterfield et al. 2002; Yang et al. 2005) though specific sequencing methods used by the other sequencing centres was variable.

Ortholog assignment

We first compared the clones from each species using a reciprocal pairwise BLASTP analysis using the predicted protein sequences. This established, for a subset of *X. tropicalis* genes, the member of the *X. laevis* paralogous gene pair that was the best match. Because neither set of clones approaches a full representation of the transcriptome for the organism from which it was derived, we employed a more conservative orthology assignment by performing a BLASTP comparison of each putative ortholog pair defined by reciprocal BLAST analysis, above, against a non-redundant set of predicted *X. tropicalis* proteins from the EnsEMBL gene predictions of the JGI *X. tropicalis* genome (release 31.1a). If both clones shared a best hit in the EnsEMBL set, we retained them as strict orthologs. Otherwise, we assumed the hits were likely members of a gene family but not the representative ortholog between the two genomes.

Paralog assignment

A self-BLASTP search of all *X. laevis* clones provided a set of putative paralogs. We only considered BLASTP alignments with percent similarity above 30% for alignments > 150 amino acids, or using a length-dependent threshold for shorter alignments (Rost 1999; Gu et al. 2002). We also required that the alignment length was at least 80% of the

total length of the query sequence. Genes identified in this manner as having a paralogous copy in the clone set were compared to the EnsEMBL non-redundant representation of all predicted tropicalis genes with a pairwise BLASTP alignment. If any putative *X. laevis* paralog matched to more than one predicted *X. tropicalis* gene with the same requirements as described above, the pair was classified as an out-paralog.

Novel cDNA candidate identification

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. 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 UTR. Sequences showing any increase in conservation (cooresponding to sequence a conservation of 1.6 standard deviations below the mean) and with a 5' UTR of at least 25 nt were selected.

Paralog Retention Analysis

For each *X. laevis* paralog pair, we found the best representative protein in Uniprot by a BLASTP search against a recent Uniprot release downloaded October, 2005 from the Uniprot download centre (http://www.pir.uniprot.org/database/download.shtml; Uniprot Download Centre). As we wanted to ensure that a gene with the same function was used, hits with an e value < 10^{-40} were not used in this analysis. We generated the Gominer input file to reflect the presence or absence of each *X. laevis* gene sequenced in this project in the set of paralogs sequenced here by indication of a 1 for genes present and a -1 for genes absent. We ran GoMiner on the command line and used alpha= 0.05 to identify significant findings.

Multiple Sequence Alignments

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

d_N/d_S Estimation

We used the PAML package to estimate d_N/d_S using the maximum likelihood method of the codeml program (Yang and Nielsen 1998). We ran codeml with the three-ratio model allowing the outgroup and one branch to vary in d_N/d_S ratio (ran once for each branch). We then ran the program again, successively fixing omega at 1 at each branch in the tree while letting the *T. rubripes* outgroup vary in d_N/d_S. We performed the likelihood ratio test for significance on each branch with an omega >1 (alpha = 0.05). We also concluded that all genes with an omega larger than the largest significant omega by this test (0.5356) and less than two show evidence of evolution under neutral selection.

Tissue Expression Profiles

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 (parsed from the FASTA header), 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 (alpha= 0.05).

Acknowledgements:

We would like to thank the National Institutes of Health, who kindly provided the funding for this research. This work would not have been possible without the cDNA libraries and EST sequences provided by various groups and we want to specifically thank Igor Dawid and Thomas Sargent (NICHD), Donald Brown (Carnegie Institute), Bruce Blumberg (UC at Irvine) and Robert Grainger (UVA). We would also like to thank Diana Palmquist and Elizabeth Chun who assisted with finishing some of our clones. Work by CPP 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.

Figure Legends:

Figure 2
A histogram comparing the distribution of d_N/d_S for all *X. laevis* genes analyzed to those of *X. tropicalis* orthologs. The shift of the *X. laevis* genes to the right suggests a relaxed constraint on these genes.

Figures
Figure 1: Gene Ontology (GO) Biological Processes associated with *X. laevis* genes (assigned by human ortholog)

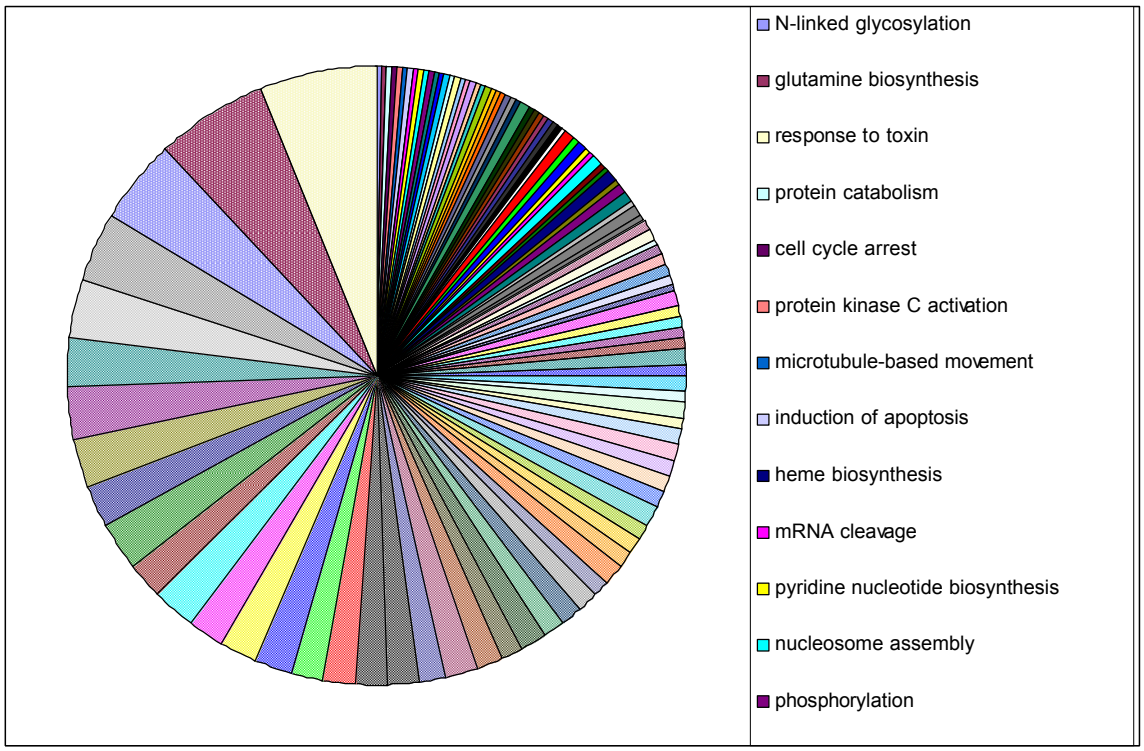
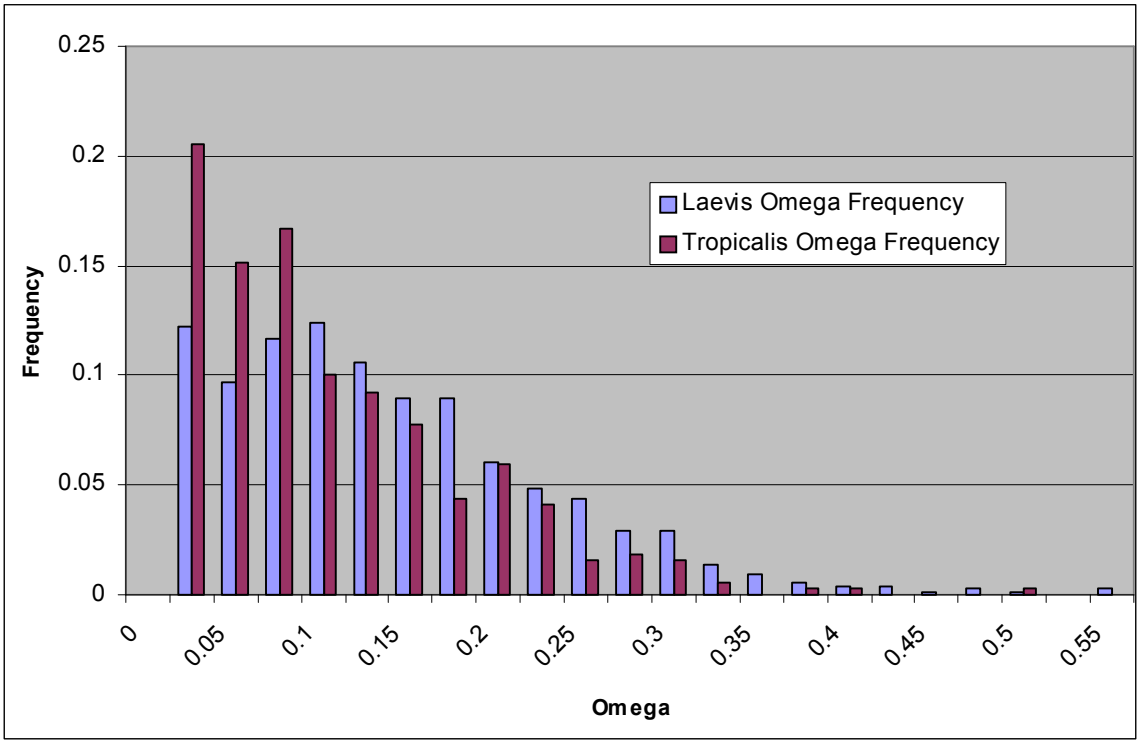


Figure 2: Frequencies of d_N/d_S ratio (omega) for *X. laevis* and *X. tropicalis* genes



Tables:

Table 1: Library names and descriptions for *Xenopus laevis* clones

Library name	Tissue	Vector	Clones
NICHD_XGC_Brn1	adult *Xenopus* brain	pCMV-SPORT6.ccdb	355
Xenopus laevis unfertilized egg cDNA library	unfertilized egg	pBluescript SK-	14
Wellcome CRC pSK egg	Egg	pBluescript SK-	403
Xenopus laevis gastrula non normalized	gastrula (stages 10.5-11.5)	pBluescript SK-	18
NICHD_XGC_Emb10	neurula	pExpress-1	46
NICHD_XGC_Emb9	neurula	pExpress-1	22
NICHD_XGC_Emb1	Embryo (stage 10)	pCMV-SPORT6	1010
NICHD_XGC_Emb2	Embryo (stage 17/19)	pCMV-SPORT6	224
NICHD_XGC_Emb3	Embryo (stage 24-25)	pCMV-SPORT6	68
NICHD_XGC_Emb4	Embryo (stage 31/32)	pCMV-SPORT6	1619
NICHD_XGC_Eye1	adult *Xenopus* eye	pCMV-SPORT6.ccdb	452
NICHD_XGC_He1	adult *Xenopus* heart	pCMV-SPORT6	52
NICHD_XGC_Kid1	adult *Xenopus* kidney	pCMV-SPORT6	778
NICHD_XGC_Li1	Liver	pCMV-SPORT6	80
NICHD_XGC_Lu1	adult *Xenopus* lung	pCMV-SPORT6	114
Xenopus laevis oocyte non normalized	Oocyte (stages 5-6)	pBluescript SK-	29
NICHD_XGC_OO1	Oocyte	pCMV-SPORT6.ccdb	595
NICHD_XGC_Ov1	Ovary	pCMV-SPORT6	640
NICHD_XGC_Sp1	adult *Xenopus* spleen	pCMV-SPORT6	595
NICHD_XGC_Te2	adult testis	pExpress-1	163
NICHD_XGC_Te2N	adult testis	pExpress-1	251
NICHD_XGC_Tad1	whole tadpole (stage 53)	pDNR-LIB	270
NICHD_XGC_Tad2	whole tadpole (stage 62)	pDNR-LIB	251
Total			8049

Table 2: Library names and descriptions for *Xenopus tropicalis* clones

Library name	Tissue	Vector	Clones
NICHD_XGC_Emb5	gastrula (stages 10-13)	pCMV-SPORT6.ccdb	513
NICHD_XGC_Emb6	neurula (stages 14-19)	pCMV-SPORT6.ccdb	430
NICHD_XGC_Emb8	tadpole (stages 40-45)	pCMV-SPORT6.ccdb	229
NICHD_XGC_Emb7	tailbud (stages 20-27)	pCMV-SPORT6.ccdb	688
XtSt10-30	Embryo, mix of stage 10, 20 and 30, mixed sex.	pRKW2	231
NICHD_XGC_Swb1	Whole body, male, 10 months old, strain F6, normal	pExpress-1	216
NICHD_XGC_Swb1N	Whole body, male, 10 months old, strain F6, normal	pExpress-1	611
Total			2918

Table 3: Clones sequenced from Each *Xenopus* species after filtering for redundancies

Species	Total Sequenced	non-redundant
X. tropicalis	2918	2770
X. laevis	8049	7818

Total	10967	10588

Table 4: Homology within *Xenopus* cDNA clones sequenced

Total ortholog pairs between *X. laevis* and *X. tropicalis*	2013
Total *X. laevis* paralog pairs	1745
X. laevis in-paralog pairs	1199
X. laevis out-paralog pairs	546
X. laevis in-paralogs with *X. tropicalis* and *T. rubripes* ortholog (used in dN/dS analysis)	437

Table 5: Accessions and d_N/d_S estimates for genes indicating no selective constraint

X. laevis Genes		*X. tropicalis* Genes	
Accession	omega (d_N/d_S)	Accession	omega (d_N/d_S)
BC087451	0.5453	BC087992	0.6146
BC060483	0.5464	BC061315	0.6789
BC077935	0.5500	BC075513	0.6826
BC054174	0.5665	BC080500	0.7280
BC071003	0.5729	BC090370	0.7730
BC099298	0.5755	BC067322	0.9236
BC072984	0.5828		
BC070676	0.5892		
BC099246	0.6021		
BC073251	0.6450		
BC078076	0.6472		
BC072765	0.6657		
BC099280	0.6767		
BC083036	0.6796		
BC047976	0.6865		

BC072854	0.6970
BC082908	0.7055
BC086274	0.7645
BC072891	0.7913
BC068638	0.8368
BC084386	0.8400
BC077496	0.8473
BC085024	0.8827
BC088964	0.9652
BC080014	1.0918
BC060356	1.1261
BC084787	1.1963
BC077290	1.3437
BC077201	1.7795

Table 6: Over-represented Gene Ontology categories in *X. laevis* genes with actively-transcribed paralogs

GO Term	Description	P (using best homolog)	P (using human ortholog)
GO:0000188	*inactivation of MAPK*	0.0618	0.0200
GO:0004299	**proteasome endopeptidase activity**	0.0439	0.0877
GO:0004556	alpha amylase activity	0.0316	0.0330
GO:0004725	*protein tyrosine phosphatase activity*	0.0312	0.0573
GO:0004726	*non-membrane spanning protein tyrosine phosphatase activity*	0.0587	0.0613
GO:0004840	**ubiquitin conjugating enzyme activity**	0.0044	0.0200
GO:0004842	**ubiquitin-protein ligase activity**	0.0446	0.0862
GO:0005839	**proteasome core complex**	0.0533	0.0717
GO:0005901	Caveola	0.0316	0.0613
GO:0006800	oxygen and reactive oxygen species metabolism	0.0819	0.0337
GO:0006979	response to oxidative stress	0.0414	0.0665
GO:0007043	intercellular junction assembly	0.0044	0.0613
GO:0007582	physiological process	0.0782	0.0522
GO:0008639	**small protein conjugating enzyme activity**	0.0044	0.0200
GO:0016160	amylase activity	0.0316	0.0330
GO:0016599	caveolar membrane	0.0316	0.0613
GO:0017017	*MAP kinase phosphatase activity*	0.0316	0.0613
GO:0030151	molybdenum ion binding	0.0316	0.0119
GO:0045216	intercellular junction assembly and/or maintenance	0.0102	0.0949
GO:0045934	negative regulation of nucleobase, nucleoside, nucleotide or nucleic acid metabolism	0.0159	0.0560
GO:0048468	cell development	0.0587	0.0949

Table 7: Differential tissue expression for *X. laevis* in-paralogs from EST information (top ten genes)

P	accession 1	accesion 2	EST count 1	EST count 2	Tissue
1.67E-14	BC054950	BC056840	78	10	Liver
4.28E-14	BC054151	BC041281	53	2	Testes
1.15E-11	BC072139	BC068905	19	86	Testes
8.37E-08	BC054976	BC045004	52	11	Testes
4.77E-07	BC079705	BC059977	0	21	Kidney
7.63E-06	BC082829	BC053760	17	0	Heart
0.000745	BC072304	BC056053	3	18	Testes
0.000829	BC046664	BC041210	9	29	Testes
0.000977	BC081147	BC072297	10	0	Testes
0.000977	BC061650	BC084848	10	0	Testes

Table 8: Differential tissue expression for *X. laevis* out-paralogs from EST information (top ten genes)

P	accession 1	accesion 2	EST count 1	EST count 2	Tissue
5.89E-13	BC071136	BC077532	2	49	Testes
5.96E-08	BC056039	BC080076	0	24	Testes
2.63E-06	BC084109	BC084348	17	56	Brain
5.65E-06	BC060381	BC044961	61	21	Testes
7.63E-06	BC100214	BC060415	0	17	Kidney
1.53E-05	BC053786	BC074339	16	0	Lung
1.93E-05	BC099349	BC054964	5	29	Kidney
3.05E-05	BC074315	BC082868	15	0	Brain
0.000122	BC042249	BC042282	0	13	Ovary
0.000122	BC044063	BC098958	13	0	Brain

Table 9: *Xenopus leavis* genes identified without protein comparison

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

References:

Adams, Keith L. and Jonothan F Wendel. 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 Research 29(1)37-40

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 25(1):25-9.

Audic, S. and Claverie, J.M. 1997. The Significance of Digital Gene Expression Profiles. Genome Research 7(10):986-95.

Birney, E., Andrews, D., Bevan, P., Caccamo, M., Cameron, G., Chen, Y., Clarke, L., Coates, G., Cox, T., Cuff, J. *et al*. 2004. EnsEMBL 2004. Nucleic Acids Res 32:D468–D470.

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. and Soares, M. B. 2001. The NIEHS *Xenopus* maternal EST project: interim analysis of the first 13,879 ESTs from unfertilized eggs. Gene 267: 71-87

Rost, B. 1999. Twilight zone of protein sequence alignments. Protein Engineering 12(2) 85-94.

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. Jun 1;30(11):2460-8.

Ewing B and Green P. 1998. Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. Genome Res. Mar;8(3):186-94.

De Sa, R. O., and Hillis, D.M. 1990. Phylogenetic Relationships of the Pipid Fogs *Xenopus* and Silurana: An Integration of Ribosomal DNA Morphology. Mol Biol Evol 7(4)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: phyleogeogroaphy and implications for polyploidy evolution. Molecular Phylogenetics and Evolution 33: 197-213.

Force, A., M. Lynch, F.B. Pickett, A. Amores, Y. Yan and J. Postlethwait. 1999. Preservation of Duplicate Genes by Complementary, Degenerative Mutations. Genetics 151: 1531-1545.

Fuglsang, A. 2004. The 'effective number of codons' revisited. Biochemical and Biophysical Research Communications 317:957-964.

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. Devel Biol 271 498-516.

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

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. Oct;14(10B):2121-7.

Gibson, T. J. and Spring, J. 2002. Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins. Trends Genet. 14:46-49.

Gu, Z., Cavalcanti A., Chen F. C., Bouman P., and Li, W. H. 2002. Extent of gene duplication in the genomes of Drosophila, nematode, and yeast. Mol Biol Evol. 19:256-262.

Hirsch, N., Zimmerman, L. and Grainger, R. 2002. *Xenopus*, the Next Generation: *X. tropicalis* Genetics and Genomics. Devel Dynamics 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(6):1360-1369.

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: Tinsley, R.C., Kobel, H.R. (Eds.), The Biology of *Xenopus*. Clarendon Press, Oxford, pp. 391-401.

Maido R., Storm, C. E. V. and Sonnhammer, E. L. L. 2001. Automatic Clustering of Orthologs and In-paralogs from Pairwise Species Comparisons. Journal of Molecular Biology 314, 1041-1052.

Ohno, S. Evolution by Gene Duplication (Allen and Unwin, London, 1970).

Pontius JU, Wagner L, Schuler GD. 2003. UniGene: a unified view of the transcriptome. In: The NCBI Handbook. Bethesda (MD): National Center for Biotechnology Information.

Rost, B. 1999. Twilight zone of protein sequence alignments. Protein Engineering vol.12 no.2 pp 85-94.

Strausberg, R. L., Feingold, E. A., Grouse, L. H., Derge, J. G., Klausner, R. D., Collins, F. S., Wagner, L., Shenmen, C.M., Schuler, GD, Altschul, SF. 2002. Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. PNAS. 99(26):16899-903. Epub 2002 Dec 11.

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. Phil Trans R Soc Lond B 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. 22:4673-80.

Thornton, K. and Long, M. 2002. Rapid Divergence of Gene Duplicates on the Drosophila melanogaster X Chromosome. Mol Biol Evol 19(6):918-925.

Wernersson, R. and Pederson, A. G. 2003. RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. Nuc A Res 31(13):3537-3539.

Wong, W. S. W., Yang, Z., Goldman, N. and Nielsen, R. 2004. Accuracy and Power of Statistical Methods for Detecting Adaptive Evolution in Protein Coding Sequences and for Identifying Positively Selected Sites. Genetics 168:1041-1051.

Yang Z et al. Phylogenetic analysis by maximum likelihood (PAML), Version 1.3. University of California, Berkely, California, USA

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.

Zhang, L., Vision, T. and Gaut, B. 2002. Patterns of Nucleotide Substitution Among Simultaneously Duplicated Gene Pairs in Arabidopsis thaliana. Mol Biol Evol 19(9):1464-1473.

Zeeberg, B. R., Feng, W., Wang, G., Wang, M. D., Fojo, A. T., Sunshine, M., Narasimhan, S., Kane, D. W., Reinhold, W. C., Lababidi, S. et al. 2003. GoMiner: A

Resource for Biological Interpretation of Genomic and Proteomic Data. Genome Biology. 4(4):R28.

Web Site References

http://xgc.nci.nih.gov/; *Xenopus* Gene Collection Home page
http://www.ensembl.org/*Xenopus*_tropicalis; EnsEMBL Home page
http://genome.jgi-psf.org/Xentr3/Xentr3.home.html; DOE JGI *Xenopus* Home page
http://codonw.sourceforge.net; codonW software Home page
http://www.pir.uniprot.org/database/download.shtml; Uniprot Download Centre
ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/ , NR protein database, release date: March 7, 2005.