# The Genome of the Western Clawed Frog *Xenopus tropicalis*

Uffe Hellsten,[1] Richard M. Harland,[2] Michael J. Gilchrist,[3] David Hendrix,[2] Jerzy Jurka,[4] Vladimir Kapitonov,[4] Ivan Ovcharenko,[5] Nicholas H. Putnam,[6] Shengqiang Shu,[1] Leila Taher,[5] Ira L. Blitz,[7] Bruce Blumberg,[7] Darwin S. Dichmann,[2] Inna Dubchak,[1] Enrique Amaya,[8] John C. Detter,[9] Russell Fletcher, [2] Daniela S. Gerhard,[10] David Goodstein,[1] Tina Graves,[11] Igor V. Grigoriev,[1] Jane Grimwood,[1,12] Takeshi Kawashima,[2,13] Erika Lindquist,[1] Susan M. Lucas,[1] Paul E. Mead,[14] Therese Mitros,[2] Hajime Ogino,[15] Yuko Ohta,[16] Alexander V. Poliakov,[1] Nicolas Pollet,[17] Jacques Robert,[18] Asaf Salamov,[1] Amy K. Sater,[19] Jeremy Schmutz,[1,12] Astrid Terry,[1] Peter D. Vize,[20] Wesley C. Warren,[11] Dan Wells,[19] Andrea Wills,[2] Richard K. Wilson,[11] Lyle B. Zimmerman,[21] Aaron M. Zorn,[22] Robert Grainger,[23] Timothy Grammer,[2] Mustafa K. Khokha,[24] Paul M. Richardson,[1] and Daniel S. Rokhsar[1,2]

[1] Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA

[2] Center for Integrative Genomics, University of California Berkeley 94720, USA

[3] Division of Systems Biology, MRC National Institute for Medical Research, The Ridgeway, London, NW7 1AA, UK

[4] Genetic Information Research Institute, Mountain View, CA 94043, USA

[5] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA.

[6] Department of Ecology and Evolutionary Biology, Rice University, Houston, TX 77005, USA

[7] Dept of Developmental and Cell Biology, 4410 Natural Sciences Building 2, University of California

Irvine, CA 92697-2300, USA

[8] The Healing Foundation Centre, University of Manchester, Oxford Road, Manchester M13 9PT, UK

[9] DOE Joint Genome Institute, Los Alamos National Laboratory, Los Alamos NM 87545, USA

[10] Office of Cancer Genomics, NCI , NIH, DHHS Bethesda, Maryland, USA.

[11] Genome Sequencing Center, Washington University School of Medicine, St Louis, MO 63108, USA

[12] JGI HudsonAlpha Institute for Biotechnology, 601 Genome Way Huntsville, AL 35806, USA

[13] Okinawa Institute of Science and Technology, 12-22, Suzaki, Uruma, Okinawa 904-2234, Japan

[14] Department of Pathology, St Jude Children's Research Hospital, 262 Danny Thomas Place, D4047C, Mailstop 342, Memphis, TN 38105, USA

[15] Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara, Japan

[16] Department of Microbiology and Immunology, University of Maryland School of Medicine, Baltimore, MD 21201 USA

[17] Programme d'Epigénomique, CNRS, Genopole, Université d'Evry Val d'Essonne, F-91058 Evry, France

[18] Department of Microbiology & Immunology, Box 672, University of Rochester , Medical Center, Rochester, NY 14642 USA

[19] Dept. of Biology and Biochemistry, University of Houston, Houston TX 77204-5001, USA

[20] Department of Biological Sciences, University of Calgary, 2500 University Drive NW, Calgary, Alberta T2N 1N4, Canada

[21] MRC National Institute for Medical Research, London, UK

[22] Division of Developmental Biology, Cincinnati Children's Hospital Medical Center, Cincinnati OH 45229, USA

[23] Department of Biology, Gilmer Hall, PO Box 400328, Charlottesville, VA. 22904-4328, USA

[24] Department of Pediatrics and Genetics, Yale University School of Medicine, P.O. Box 208064, New Haven, CT 06520-8064, USA

## Abstract

**The western clawed frog *Xenopus tropicalis* is an important model for vertebrate development that combines experimental advantages of the African clawed frog *Xenopus laevis* with more tractable genetics. Here we present a draft genome sequence assembly of *X. tropicalis*. This genome encodes over 20,000 protein-coding genes, including orthologs of at least 1,700 human disease genes. Over a million expressed sequence tags validated the annotation. More than one-third of the genome consists of transposable elements, with unusually prevalent DNA transposons. Like other tetrapods, the genome contains gene deserts enriched for conserved non-coding elements. The genome exhibits remarkable shared synteny with human and chicken over major parts of large chromosomes, broken by lineage-specific chromosome fusions and fissions, mainly in the mammalian lineage.**

African clawed frogs (the genus *Xenopus*, meaning "strange foot") comprise more than twenty species of frogs native to Sub-Saharan Africa. The species *Xenopus laevis* was first introduced to the U.S. in the nineteen forties where a low-cost pregnancy test took advantage of the responsiveness of frogs to human chorionic gonadotropin(*1*). Since the frogs were easy to raise and had other desirable properties such as large eggs, external development, easily manipulated embryos and transparent tadpoles, *X. laevis* gradually developed into one of the most productive model systems for vertebrate experimental embryology(*2*).

However, *X. laevis* has a large paleotetraploid genome with an estimated size of 3.1 billion bases (Gbp) on 18 chromosomes and a generation time of 1-2 years. In contrast, the much smaller diploid western clawed frog, *X. tropicalis,* has a small genome, about 1.7 Gbp on 10 chromosomes (*3*), matures in only 4 months and requires less space than its larger cousin.  It is thus readily adopted as an alternative experimental subject for developmental and cell biology (Fig. 1).

As a group, amphibians are phylogenetically well-positioned for comparisons to other vertebrates, having diverged from the amniote lineage (mammals, birds, reptiles) some 360 million years ago. The comparison with mammalian and bird genomes also provides opportunity to examine the dynamics of tetrapod chromosomal evolution.

The *X. tropicalis* draft genome sequence described here was produced from ~7.6-fold redundant random shotgun sampling of genomic DNA from a seventh generation inbred Nigerian female. The assembly ((*4*), Tables S1-S3 and accession AAMC00000000) spans about 1.51 Gbp of scaffolds, with half of the assembled sequence contained in 272 scaffolds ranging in size from 1.56 to 7.82 Mb. Of known genes, 97.6% are present in the assembly, attesting to its near completeness in genic regions (*4*). Nearly two million *Xenopus* ESTs from diverse developmental stages and adult tissues complement the genome and enable studies of alternative splicing and identification of developmental stage- and tissue-specific genes (*4*).

Over a third of the frog genome consists of transposable elements (TEs),  (Table S7), higher than the 9% TE density in the chicken genome (*5*) but comparable to the 40-50% density in mammalian genomes(*6-7*). Many families of frog TEs are more than 25% divergent from their consensus sequence, so like mammalian and bird TEs they have persisted for as long as 20-200 million years (*5-6*). This contrasts with the faster turnover observed in insects, nematodes, fungi, and plants (*6, 8-9*). Recently active TEs (1-5 Mya) are more common in frogs than in mammals or birds and are comparable with prevalence in fish, insects, nematodes, and plants. Among these is an unusually high diversity of very young families of L1 non-LTR retrotransposons, Penelope, and DIRS retrotransposons. In contrast to other vertebrates, most recognizable transposable elements (72%) are DNA transposons, rather than the retrotransposons that dominate other genomes (*5-8, 10*).  Among these families(*11-12*), we identify *Kolobok* is a novel superfamily of DNA transposons. The genome also contains LTR retrotransposons of all major superfamilies, with higher diversity than in all other studied eukaryotes (Table S8). While most are ubiquitous, *Copia, BEL, and Gypsy* elements are not found in birds and mammals, suggesting that this subset became immobile after divergence from the amphibian lineage.

We estimate that the *X. tropicalis* genome contains 20,000 to 21,000 protein-coding genes using homology-based gene prediction methods and deep *Xenopus* EST and cDNA resources. These include orthologs of 79% of identified human disease genes (*4*). The genome contains 1,850 tandem expanded gene families with between 2 and 160 copies, accounting for nearly 24% of protein-coding loci. The largest expansion comprises tetrapod specific olfactory receptors (class II) occupying the first 1.7 Mb on scaffold_24. Other large expansions include protocadherins, bitter-taste receptors, and vomeronasal (pheromone) receptors (Table S9).

The *X. tropicalis* genome displays long stretches of gene colinearity with human and chicken (Fig. 2). Of the 272 largest scaffolds (totaling half the assembly) 267 show such colinearity (*4*). 60% of all gene models on these scaffolds can be directly associated with a human and/or chicken ortholog by conserved synteny. Patches of strict conserved colinearity are interrupted by large-scale inversions within the same linkage groups, and more rarely by chromosome breakage and fusion events, similar to the findings reported for human and chicken (Fig. 2, (*5*)) and in agreement with persistent conservation of linkage groups across chordates (*13*).

We uniquely placed 1,696 markers from the existing genetic map of *X. tropicalis* (http://tropmap.biology.uh.edu/map.html) onto a total of 691 scaffolds constituting more than 764 Mb of genomic sequence (*4, 14*). To identify lineage-specific fusion- and breakage-events within the mammals and sauropsids we identified blocks of conserved synteny between frog, human, and chicken. These blocks were detected using genomic probes comprising three-way orthologs between these tetrapods. 5,642 of these probes define conserved linkage blocks containing at least 15 genes and at least 2Mb of sequence (*4, 14*). The tetrapod ancestry of human and chicken chromosome 1 is outlined in Fig. 2. Remarkably, a core of more than 150 Mb of sequence spanning the centromere of human chr 1 (chicken chr 8, frog LG VII) has remained largely intact during ~360 million years of evolution since the tetrapod ancestor (Fig. 2A). Detailed shared synteny is interrupted by large-scale inversions, but gene order is frequently conserved over stretches of tens of Mb. Human chromosome 1 is seen to have grown by three lineage-specific mammalian fusions. In contrast, there are several mammalian-specific breakpoints (Fig. 2B). The genomic material on the entire q arm of chicken shows linkage conservation to frog LG VI while the human counterparts are scattered over regions of chromosomes 2, 3, 11, 13, 21, and X. The p arm indicates two mammalian breaks, suggesting that regions of chromosomes 7, 12, and 22 were once part of the same chromosome.

By extending this analysis to all human and chicken chromosomes we identified 22 human fusion and 21 fission events, versus only four fusions and one break in chicken. Clearly, the mammalian lineage has undergone considerably more rearrangement than the sauropsids, although the total chromosome count appears to have remained fairly constant. The segments analyzed here are distributed on 23 human and 22 chicken chromosomes, consistent with a derivation from 24 or 25 ancestral amniote chromosomes. Note that the chicken microchromosomes are unresolved by this analysis, preventing determination of the exact ancestral chromosome number. Both the vertebrate and eumetazoan ancestors have been suggested to have had about a dozen large chromosomes (*13, 15*). The current analysis indicates that the amniote ancestor had twice as many, suggesting substantial chromosome breakage on the amniotic stem.

The extensive conserved synteny among tetrapods allows us to provisionally place frog scaffolds without genetic markers onto the linkage map.  These are shown in Fig 2 as black bars within the blocks of conserved linkage with frog. A total of 170 large scaffolds containing about 200 Mb of sequence were assigned a linkage group in this manner. Such *in silico* inferred linkages will ultimately need to be verified experimentally, but have already proven useful in the positional identification and cloning of the gene responsible for the *muzak* mutation, which affects heart function (*16*).

The *X. tropicalis* genome exhibits extensive sequence conservation with other vertebrates, with the amphibian sequence filling a phylogenetic gap. Recognizable non-coding sequence conservation diminishes steadily with increasing evolutionary distance (Fig. S6). Frog genes adjacent to conserved non-coding sequences (CNS) are enriched or depleted in several gene ontology categories, including sensory perception of smell, response to stimulus, and regulation of transcription, among others (Table S16).

Gene deserts (defined as the top 3 percent of the longest intergenic regions) cover 17% of the genome and vary between 201 kbp and 1.2 Mbp. The 683 gene deserts contain almost 25% of CNSs. In mammalian genomes, these gene deserts have been found to harbor cis-regulatory elements(*17*).

The power of genome comparison and high-throughput transgenesis in *Xenopus* is illustrated in Fig. S7, where several mammalian-*Xenopus* CNS at the *Six3* locus were assayed for enhancers regulating its eye- and forebrain-specific expression. The analysis suggests that frog-mammal comparisons may be more suitable than fish-mammal comparisons for identifying conserved cis-regulatory elements (see, e.g., CNS5 in Fig. S7).

Developmental pathways controlling early vertebrate axis specification were first implicated by work in *Xenopus* (*2*) but some interesting amphibian modifications can be found.  For example, a *Wnt* ligand required for dorsal development, named *Wnt11b* in *X.*

*tropicalis*, has been lost from mammals, but is found in the chick and zebrafish (as *silberblick*) (*18*). Despite its retention in these vertebrates, there is no evidence to support a maternal role in axis formation similar to *Xenopus*. Similarly a *tbx16* homolog, *vegT,* is retained in frog, fish and chick, but is uniquely used in *Xenopus* for the establishment of the endoderm and mesoderm (*19*).

*X. tropicalis* also shows multiplications of genes deployed at the blastula and gastrula stages. For example, mammals have a single *nodal* gene, while *X. tropicalis* has more than 6. Synteny relationships reveal that *nodal4* on scaffold 204 is orthologous to the single human *nodal*, while a cluster of more than 6 *nodals* on scaffold 34 is orthologous to the chicken *nodal*. Further analysis suggests that these two *nodal* loci arose in one of the whole-genome duplications at the base of vertebrate evolution and that the birds and mammals subsequently lost different *nodal* genes, while the lizard *Anolis carolinensis* has retained both copies (*4*).

The theme of duplication is reiterated by several transcription factors that act during gastrulation (*4*). The transcriptional activator *siamois*, expressed in the organizer, is triplicated locally in the genome; so far this gene is unique to the frog. The *ventx* genes are expressed at the same time, but opposite the organizer, and are present in six linked copies.

Conservation of the vertebrate immune system is highlighted by mammalian and Xenopus genome comparisons (*20-21*). While orthology is usually obvious, synteny has been an important tool to identify diverged genes. For example, a diverged *CD8 beta* retains proximity to *CD8 alpha*, and *CD4* neighbors *Lag3* and *B* protein. Similarly, an Interleukin2/21-like sequence was identified in a syntenic region between the *tenr* and *centrin4* genes. The immunoglobulin repertoire provides further links between vertebrate immune systems. The *IgW* immunoglobulin was thought to be unique to shark/lungfish, but an orthologous *IgD* isotype in frog provides a connection between the fish and amniote gene families (*22-23*).

Unique antimicrobial peptides play an important role in skin secretions that are absent in birds, reptiles and mammals. Antimicrobial peptides (caerulein, levitide, magainin, PGLa/PYLa, PGQ, xenopsin), neuromuscular toxins (*e.g*. xenoxins) and neuropeptides (*e.g*. thyrotropin releasing hormone, TRH) (*24*) are secreted by granular glands and the first group represents an important defense against pathogens (*25*). Antimicrobial peptides are clustered in at least seven transcription units over 350 kbp on scaffold 811, with no intervening genes.

*X. tropicalis* occupies a key phylogenetic position among previously sequenced vertebrate genomes, namely amniotes and teleost fish. Given the utility of the frog as a genetic and developmental biology system and the large and increasing amounts of cDNA sequence from the pseudo-tetraploid *X. laevis*, the *X. tropicalis* reference sequence is well poised to advance our understanding of genome and proteome evolution in general, and vertebrate evolution in particular.

## References and Notes

1. L. Hogben, C. Gordon, *J. Exp. Biol.* **7**, (1930).
2. D. D. Brown, *J Biol Chem* **279**, 45291 (Oct 29, 2004).

3. J. Tymowska, *Cytogenet Cell Genet* **12**, 297 (1973).
4. SOM. (2010).
5. I. C. G. S. Consortium, *Nature* **432**, 695 (Dec 9, 2004).
6. E. S. Lander *et al.*, *Nature* **409**, 860 (Feb 15, 2001).
7. R. H. Waterston *et al.*, *Nature* **420**, 520 (Dec 5, 2002).
8. V. V. Kapitonov, J. Jurka, *Genetica* **107**, 27 (1999).
9. V. V. Kapitonov, J. Jurka, *Proc Natl Acad Sci U S A* **100**, 6569 (May 27, 2003).
10. RiceConsortium, *Nature* **436**, 793 (Aug 11, 2005).
11. N. L. Craig, R. Craigie, M. Gellert, A. M. Lambowitz, Eds., *Mobile DNA II*, (ASM Press, Washington, DC, 2002).
12. V. V. Kapitonov, J. Jurka, *DNA Cell Biol* **23**, 311 (May, 2004).
13. N. H. Putnam *et al.*, *Science* **317**, 86 (Jul 6, 2007).
14. SOD, (2010).
15. I. G. Woods *et al.*, *Genome Res* **15**, 1307 (Sep, 2005).
16. A. Abu-Daya, A. K. Sater, D. E. Wells, T. J. Mohun, L. B. Zimmerman, *Dev Biol* **336**, 20 (Dec 1, 2009).
17. M. A. Nobrega, I. Ovcharenko, V. Afzal, E. M. Rubin, *Science* **302**, 413 (Oct 17, 2003).
18. R. J. Garriock, A. S. Warkman, S. M. Meadows, S. D'Agostino, P. A. Krieg, *Dev Dyn* **236**, 1249 (May, 2007).
19. M. Kofron *et al.*, *Development* **126**, 5759 (Dec, 1999).
20. L. Du Pasquier, J. Schwager, M. F. Flajnik, *Annu Rev Immunol* **7**, 251 (1989).
21. J. Robert, Y. Ohta, *Dev Dyn* **238**, 1249 (Jun, 2009).
22. Y. Ohta, M. Flajnik, *Proc Natl Acad Sci U S A* **103**, 10723 (Jul 11, 2006).
23. Y. Zhao *et al.*, *Proc Natl Acad Sci U S A* **103**, 12087 (Aug 8, 2006).
24. G. Kreil, *Skin Secretions of Xenopus Laevis*. H. R. Tinsley, Ed., The Biology of Xenopus (The Zoological Society of London, Oxford, 1996).
25. L. A. R. Rollins-Smith, L.K.; Houston C.J.L.E.; Woodhams, D.C., *Antimicrobial peptide defenses in amphibian skin* Integrative and Comparative Biology (2005), vol. 45.

## Acknowledgments

## Acknowledgments

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396. This research was supported in part by the Intramural Research Program of the NIH, National Library of Medicine, and by a grant to R.K.W. from the National Human Genome Research Institute (NHGRI U01 HG02155) with supplemental funds provided by the National Institute of Child Health and Human Development. We thank Richard Gibbs and Steve Scherer of the Human Genome Sequencing Center, Baylor College of Medicine, for their contributions to SSLP identification and mapping.

## Figure Legends

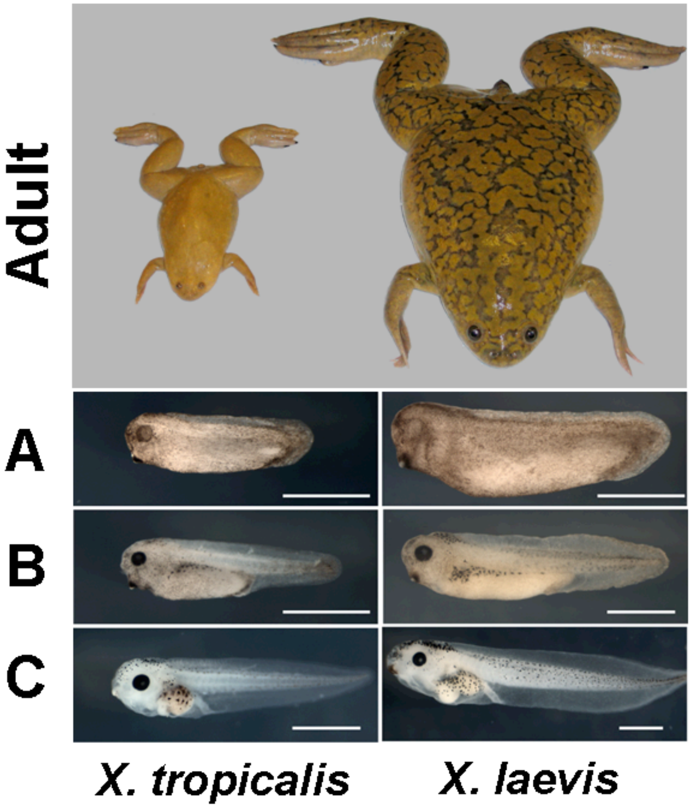


**Fig. 1: Comparison of adults and tadpoles of *X.tropicalis* and *X. laevis*. Adult body length is 5 and 10 cm respectively. (A) tailbud (B) swimming tadpole (C) feeding tadpole. White bar indicates 1 mm.**

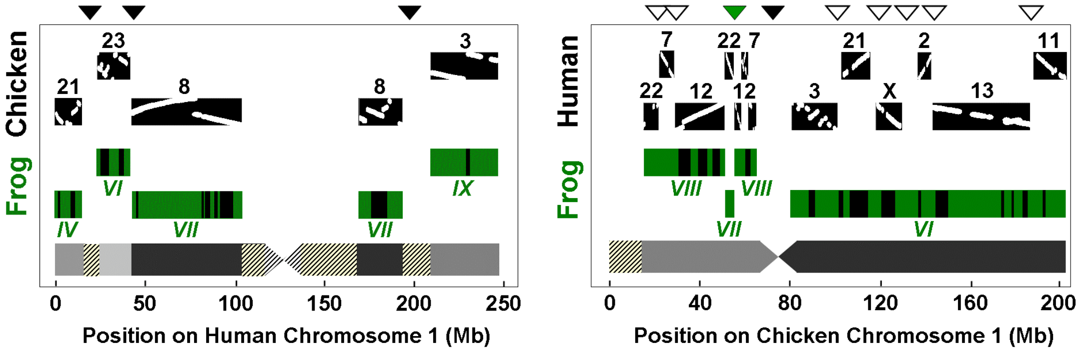


**Figure 2: Blocks of conserved tetrapod linkage for human (panel A) and chicken (panel B) chromosome 1 reveal fusions (solid black triangles) and break points (unfilled triangles) in amniotes. A total of three human fusions (panel A), seven human breaks (panel B), and one chicken fusion (panel B) is observed. The green triangle in panel B indicates the position of an apparent frog-specific break or ancestral amniote fusion. Grey areas indicate origin in different ancestral chromosomes. Shaded areas show larger regions with insufficient three-way synteny information. Detailed comparison of gene order in human and chicken reveals multiple large-scale inversions (dot plots on the black blocks). The green frog blocks consist of multiple scaffolds, 55 in panel A and 97 in panel B. Bars on the frog blocks show the location of scaffolds which do not contain markers from the linkage map, but have been predicted to associate with the linkage group by conserved synteny.**

# Supporting Online Material

# The Genome of the Western Clawed Frog *Xenopus tropicalis*

Uffe Hellsten,[1] Richard M. Harland,[2] Michael J. Gilchrist,[3] David Hendrix,[2] Jerzy Jurka,[4] Vladimir Kapitonov,[4] Ivan Ovcharenko,[5] Nicholas H. Putnam,[6] Shengqiang Shu,[1] Leila Taher,[5] Ira L. Blitz,[7] Bruce Blumberg,[7] Darwin S. Dichmann,[2] Inna Dubchak,[1] Enrique Amaya,[8] John C. Detter,[9] Russell Fletcher, [2] Daniela Gerhard,[10] David Goodstein,[1] Tina Graves,[11] Igor V. Grigoriev,[1] Jane Grimwood,[1,12] Takeshi Kawashima,[2,13] Erika Lindquist,[1] Susan M. Lucas[1], Paul E. Mead,[14] Therese Mitros,[2] Hajime Ogino,[15] Yuko Ohta,[16] Alexander V. Poliakov,[1] Nicolas Pollet,[17] Jacques Robert,[18] Asaf Salamov,[1] Amy K. Sater,[19] Jeremy Schmutz,[1,12] Astrid Terry,[1] Peter D. Vize,[20] Wesley C. Warren,[18] Dan Wells,[19] Andrea Wills,[2] Lyle B. Zimmerman,[21] Aaron M. Zorn,[22] Robert Grainger,[23] Timothy Grammer,[2] Mustafa K. Khokha,[24] Paul M. Richardson,[1] and Daniel S. Rokhsar[1,2]

[1] Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA

[2] Center for Integrative Genomics, University of California Berkeley 94720, USA

[3] Division of Systems Biology, MRC National Institute for Medical Research, The Ridgeway, London, NW7 1AA, UK

[4] Genetic Information Research Institute, Mountain View, CA 94043, USA

[5] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA.

[6] Department of Ecology and Evolutionary Biology, Rice University, Houston, TX 77005, USA

[7] Dept of Developmental and Cell Biology, 4410 Natural Sciences Building 2, University of California Irvine, CA 92697-2300, USA

[8] The Healing Foundation Centre, University of Manchester, Oxford Road, Manchester M13 9PT, UK

[9] DOE Joint Genome Institute, Los Alamos National Laboratory, Los Alamos NM 87545, USA

[10] Office of Cancer Genomics, NCI , NIH, DHHS Bethesda, Maryland, USA.

[11] Genome Sequencing Center, Washington University School of Medicine, St Louis, MO 63108, USA

[12] JGI HudsonAlpha Institute for Biotechnology, 601 Genome Way Huntsville, AL 35806, USA

[13] Okinawa Institute of Science and Technology, 12-22, Suzaki, Uruma, Okinawa 904-2234, Japan

[14] Department of Pathology, St Jude Children's Research Hospital, 262 Danny Thomas Place, D4047C, Mailstop 342, Memphis, TN 38105, USA

[15] Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara, Japan

[16] Department of Microbiology and Immunology, University of Maryland School of Medicine, Baltimore, MD 21201 USA

[17] Programme d’Epigénomique, CNRS, Genopole, Université d'Evry Val d'Essonne, F-91058 Evry, France

[18] Department of Microbiology & Immunology, Box 672, University of Rochester , Medical Center, Rochester, NY 14642 USA

[19] Dept. of Biology and Biochemistry, University of Houston, Houston TX 77204-5001, USA

[20] Department of Biological Sciences, University of Calgary, 2500 University Drive NW, Calgary, Alberta T2N 1N4, Canada

[21] MRC National Institute for Medical Research, London, UK

[22] Division of Developmental Biology, Cincinnati Children's Hospital Medical Center, Cincinnati OH 45229, USA

[23] Department of Biology, Gilmer Hall, PO Box 400328, Charlottesville, VA. 22904-4328, USA

[24] Department of Pediatrics and Genetics, Yale University School of Medicine, P.O. Box 208064, New Haven, CT 06520-8064, USA

## Supplementary Note 1. Shotgun sequencing and genome assembly

### DNA source and material preparation

Genomic DNA was prepared from a 7th generation inbred Nigerian frog derived by brother-sister mating from an original mating pair. Nuclear DNA was isolated from erythrocytes and other tissues by standard methods.

### Shotgun library preparation and sequencing (plasmid and fosmid)

Plasmid- and fosmid-end sequencing was performed using standard library protocols and Sanger dye-terminator chemistries on the ABI-3730 and MegaBACE 4000 sequencing instruments. Sequencing totals are shown in Table S1, with insert sizes estimated self-consistently from the shotgun assembly. Sequence coverage from high quality reads is computed assuming a nominal genome size of 1.7 Gb. High quality reads are longer than 200 bp free of vector sequence and with Phred $Q > 20$. All traces for this project were deposited in the NCBI Trace Archive with species_code="XENOPUS TROPICALIS", CENTER_NAME = "JGI" and SOURCE_TYPE=”GENOMIC”.

### BAC libraries and sequencing

Two BAC libraries were used, and described here briefly for completeness. More information can be found at the Children’s Hospital of Oakland (CHORI) BAC resources website (http://bacpac.chori.org), from which libraries and filters can be obtained.

- The CHORI-216 Nigerian frog *Xenopus tropicalis* BAC library (segment 1) was constructed by Dr. Michael Nefedov in Pieter de Jong's laboratory at BACPAC Resources, Children's Hospital Oakland Research Institute. Genomic DNA was isolated at Virginia Mason Research Laboratory in collaboration with Chris Amemiya, from a Nigerian male frog in the 7th generation of inbreeding (N7).
- The ISB-1 Xenopus tropicalis BAC library was constructed by Shizhen Qin, Monica Dors, Brian Birditt, and Jeremy Burke in Leroy Hood's Laboratory at Institute for System Biology. High-molecular-weight DNA was isolated from blood which was obtained from a female *Xenopus tropicalis* through Dr. Robert Grainger (University of Virginia).

BAC-end sequencing was performed at Washington University Genome Sequencing Center.

**Assembly**

The *X. tropicalis* assembly 4.1 (summarized in Tables S2-S3) was produced at the DOE Joint Genome Institute with JAZZ, described in (*1*). Assembly 4 contains 19,501 scaffolds with an average coverage of 7.65X. Roughly half of the genome is contained in 272 scaffolds, all at least 1.56 Mb in length. Ubiquitous long tandem arrays of ~30-200 bp repetitive elements and incomplete coverage from the partial-digest BAC libraries limited the range of the sequence assembly. Scaffolds showing homology to a known prokaryotic contaminant as well as non-cellular or vector contamination have been removed. This Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the project accession AAMC00000000.

To assess the completeness of the Xentr4.1 assembly, we aligned 5,417 full length insert *X. tropicalis* cDNAs obtained from NCBI using GMAP(*2*). 5,394 (99.6%) have some alignment to the genome assembly, and 5,288 (97.6%) aligned with better than 95% identity over more than 50% of their length. Thus the draft assembly captures the vast majority of the expressed genome. Alignment with PASA (*3*) identifies 132 "incontiguous alignments" indicating cDNAs with hits to two or more scaffolds, which could be chimeric cDNAs, genes split across scaffold boundaries, or misassemblies within genes.

**Table S1: Sequencing summary**

| Insert size | number of reads | Estimated sequence depth | estimated clone depth |
|---|---|---|---|
| 2.95 kb plasmid | 9.4M | 3.30x | 6.9x |
| 8.30 kb plasmid | 11.0M | 3.75x | 19.6x |
| 38.5 kb fosmid | 1.9M | 0.55x | 14.5x |
| 57.4 kb BAC (ISB1) | 64.8K | 0.02x | 1x |
| 140 kb BAC (CH216) | 192.5K | 0.06x | 7.1x |
| Total | 22.55M | 7.68x | 49.1x |

**Table S2: Xentr4.1 assembly summary**

| | |
|---|---|
| Main genome scaffold total: | 19,759 scaffolds |
| Main genome contig total: | 191,450 contigs |
| Main genome scaffold sequence total: | 1,513.9 MB |
| Main genome contig sequence total: | 1359.4 MB (10.2% gap) |
| Main genome scaffold N/L50: | 272 scaffolds > 1.6 MB |
| Main genome contig N/L50: | 22,312 contigs > 17.0 KB |
| Number of scaffolds > 50 KB: | 1,683 |
| % main genome in scaffolds > 50 KB: | 89.1% |

**Table S3: Xentr4.1 assembly cumulative statistics**

| Scaffolds longer than ... | Number of scaffolds | Number of contigs in these scaffolds | Net scaffold length | Net contig length | % contig coverage |
|---|---|---|---|---|---|
| 5 mb | 19 | 9,408 | 114,541,727 | 111,088,087 | 96.98% |
| 2.5 mb | 126 | 41,071 | 475,010,582 | 459,219,917 | 96.68% |
| 1 mb | 447 | 89,121 | 973,414,426 | 935,280,328 | 96.08% |
| 500 kb | 728 | 112,024 | 1,174,400,214 | 1,123,342,508 | 95.65% |
| 250 kb | 975 | 124,872 | 1,263,955,455 | 1,203,305,907 | 95.20% |
| 100 kb | 1,366 | 135,500 | 1,326,565,464 | 1,251,671,939 | 94.35% |
| 50 kb | 1,683 | 140,070 | 1,348,839,729 | 1,266,924,112 | 93.93% |
| 25 kb | 2,546 | 145,659 | 1,380,320,822 | 1,280,387,063 | 92.76% |
| 10 kb | 6,309 | 161,846 | 1,435,979,832 | 1,313,483,945 | 91.47% |
| 5 kb | 13,574 | 179,867 | 1,492,946,094 | 1,341,754,220 | 89.87% |
| 2.5 kb | 19,188 | 190,811 | 1,512,650,785 | 1,358,153,127 | 89.79% |
| 1 kb | 19,759 | 191,450 | 1,513,925,492 | 1,359,399,966 | 89.79% |
| All | 19,759 | 191,450 | 1,513,925,492 | 1,359,399,966 | 89.79% |

## Supplementary Note 2. cDNA resources and EST sequencing

### Role of EST data in genome assembly and genomic analysis

One of the goals of genome assembly is the systematic elucidiation of gene models for the species. Expressed sequence data is an integral part of this process, allowing confirmation of exons, and being the most effective method for delineation of the UTRs. The other primary factor affecting the generation of reliable gene models is the integrity of the genome assembly, and the fragmented nature of the current assembly does create some difficulties. In this assembly we estimate that the mean distance between in-scaffold gaps is 8.6k, which, in combination with the large number of smaller scaffolds, suggests that significant numbers of genes will be poorly modeled, or not modeled at all. An analysis of the 28,704 Ensembl transcripts generated by gene modeling on this assembly suggests that 14,417 (50%) have one or both ends of the open reading frame truncated or ill-defined, and a further 8,926 (31%) have a complete open reading frame but are missing one or both UTRs.

These data suggest that the combination of the limitations of a partially assembled genome, compounded by an insufficency of expressed sequence data, make gene modeling for *Xenopus tropicalis* a significant challenge. There is however a large amount of EST data (see below), and there is also some possibility that EST or cDNA data could assist the genome assembly process itself, by enabling scaffold joining, if assembly programs were capable of using this as input. These factors, and the more straightforward requirement for well defined gene mRNA sequence data, have led the *Xenopus* community to adopt an active EST sequencing strategy to maximize gene coverage and diversity.

## Current EST and cDNA library resources

The genome sequence has been complemented by over 1.2 million *Xenopus tropicalis* EST sequences from 65 cDNA libraries that sample a useful range of developmental stages and adult organs and tissues, summarized in Table S4 (NCBI UniGene EST data resource (*4*)). In addition there are ~678,000 ESTs from diverse *Xenopus laevis* libraries. The ESTs provide a rich resource for the characterization of *Xenopus tropicalis* genes, and since many libraries were constructed in expression-ready vectors, they also provide an excellent resource for functional experiments with individual clones, or for screening by expression cloning (*5-6*).

Most cDNA libraries were made from the Nigerian strain, but some were also generated from an outbred strain from Ivory Coast (TGA). The degree of polymorphism between these libraries is low, with an estimated rate of ~1/300 in 3’ untranslated regions, based on manual count in TGA libraries aligned with Nigerian 3' UTRs of combined length of about 10,000 nucleotides.

Clustering analysis has enabled the prediction of full-length cDNA clones, their reorganization into non-redundant collections (*7*), and their input into various large scale full-insert sequencing programs. These sequencing programs, as well as many smaller efforts, have resulted in the deposition of 26,194 mRNA sequences in GenBank, representing 11,421 genes (data from NCBI-UniGene, *Xenopus tropicalis* build 47; assuming one UniGene cluster equals one gene). Although what proportion of these full-insert mRNA sequences contain the full open reading frame is not clear. EST data and full-length sequences are also available in the Xenopus Gene Collection (*8*).

**Table S4: List of EST libraries, from the UniGene database (build 47, Jan. 2009), with the number of ESTs currently in GenBank from each library, also the source of the mRNA/tissue, and the cDNA library preparation, where available (names generally refer to labs or organisations; name followed by / is an individual in a lab).**

| UniGene ID | Library Title | ESTs Submitted | mRNA:cDNA Source | Stage:Tissue |
|---|---|---|---|---|
| 8701 | Wellcome CRC pCS107 tropicalis St10-12 | 3474 | Zorn | gastrula:whole body |
| 8773 | Wellcome CRC pCS107 tropicalis egg | 2721 | Zorn | egg:whole body |
| 9665 | XGC-gastrula | 59853 | Amaya:Zorn | gastrula:whole body |
| 9908 | XGC-neurula | 60504 | Amaya:Zorn | neurula:whole body |
| 9909 | XGC-egg | 62459 | Amaya:Zorn | egg:whole body |
| 10829 | NICHD_XGC_Emb5 | 10096 | Strausberg | mixed:whole body |
| 10830 | NICHD_XGC_Emb6 | 9367 | Strausberg | neurula:whole body |
| 10895 | NICHD_XGC_Emb7 | 8507 | Strausberg | neurula:whole body |
| 10896 | NICHD_XGC_Emb8 | 8778 | Strausberg | tadpole:whole body |
| 14247 | NICHD_XGC_Swb1 | 3882 | OpenBiosystems | adult:whole body |
| 14248 | NICHD_XGC_Swb1N | 5817 | OpenBiosystems | adult:whole body |
| 14469 | XtSt10-30 | 9077 | Niehrs | mixed:whole body |
| 14603 | XGC-tadpole | 40462 | Amaya:Gurdon | tadpole:whole body |
| 15539 | Xenopus tropicalis xtbs plasmid library | 21784 | Pollet | mixed:mixed |
| 15540 | Xenopus tropicalis xthr plasmid library | 25962 | Pollet | tailbud embryo:head |
| 15887 | XGC-tailbud | 39479 | Amaya:Gurdon | tailbud embryo:whole body |
| 16078 | XGC-tailbud-head | 27776 | Amaya:Gurdon | tailbud embryo:head |
| 16801 | NIH_XGC_tropTad5 | 105127 | Harland:Fletcher/Harland | tadpole:whole body |

| | | | | |
|---|---|---|---|---|
| 16852 | NIH_XGC_tropGas5 | 841 | Grainger:Peng/Blumberg | mixed:whole body |
| 16853 | NIH_XGC_tropGas7 | 96383 | Harland:Fletcher/Harland | gastrula:whole body |
| 16854 | NIH_XGC_tropGas6 | 5248 | Grainger:Peng/Blumberg | gastrula:whole body |
| 16855 | NIH_XGC_tropNeu5 | 3952 | Grainger:Peng/Blumberg | neurula:whole body |
| 16856 | NIH_XGC_tropBrn2 | 19429 | Grammer/Harland:JGI | adult:brain |
| 16857 | NIH_XGC_tropBrn3 | 20708 | Grammer/Harland:JGI | adult:brain |
| 16858 | NIH_XGC_tropBrn4 | 19686 | Grammer/Harland:JGI | adult:brain |
| 16859 | NIH_XGC_tropTe3 | 19918 | Grammer/Harland:JGI | adult:testis |
| 16860 | NIH_XGC_tropTe4 | 20971 | Grammer/Harland:JGI | adult:testis |
| 16861 | NIH_XGC_tropTe5 | 20476 | Grammer/Harland:JGI | adult:testis |
| 16862 | NIH_XGC_tropInt1 | 20920 | Grainger:Tabb/Blumberg | adult:intestine |
| 16863 | NIH_XGC_tropHrt1 | 19364 | Grainger:Tabb/Blumberg | adult:heart |
| 16864 | NIH_XGC_tropLiv1 | 22125 | Grainger:Tabb/Blumberg | adult:liver |
| 16865 | NIH_XGC_tropMet5 | 1356 | Buchholz/Shi:JGI | metamorphosis:whole body |
| 16866 | NIH_XGC_tropMet6 | 917 | Buchholz/Shi:JGI | metamorphosis:whole body |
| 16867 | NIH_XGC_tropMet4 | 250 | Buchholz/Shi:JGI | metamorphosis:whole body |
| 16868 | NIH_XGC_tropMet2 | 645 | Buchholz/Shi:JGI | metamorphosis:whole body |
| 16869 | NIH_XGC_tropMet3 | 249 | Buchholz/Shi:JGI | metamorphosis:whole body |
| 16870 | NIH_XGC_tropKid1 | 11852 | Grainger:Tabb/Blumberg | adult:kidney |
| 16871 | NIH_XGC_tropFat1 | 11645 | Grainger:Tabb/Blumberg | adult:adipose tissue |
| 16872 | NIH_XGC_tropLun1 | 22250 | Grainger:Tabb/Blumberg | adult:lung |
| 16873 | NIH_XGC_tropOva1 | 23702 | Grainger:Tabb/Blumberg | adult:ovary |

| | | | | |
|---|---|---|---|---|
| 16874 | NIH_XGC_tropSto1 | 20022 | Grainger:Tabb/Blumberg | adult:stomach |
| 16875 | NIH_XGC_tropSkeMus1 | 21364 | Grainger:Tabb/Blumberg | adult:skeletal muscle |
| 16876 | NIH_XGC_tropOvi1 | 22486 | Grainger:Tabb/Blumberg | adult:oviduct |
| 16877 | NIH_XGC_tropSki1 | 20143 | Grainger:Tabb/Blumberg | adult:skin |
| 16878 | NIH_XGC_tropSpl1 | 16993 | Grainger:Tabb/Blumberg | adult:spleen |
| 16879 | NIH_XGC_tropBrn1 | 0 | Grainger:Tabb/Blumberg | adult:brain |
| 16880 | NIH_XGC_tropTe6 | 633 | Grainger:Tabb/Blumberg | adult:testis |
| 17804 | NIH_XGC_tropMet7 | 10 | Buchholz/Shi:JGI | metamorphosis:whole body |
| 17805 | NIH_XGC_tropMet8 | 246 | Buchholz/Shi:JGI | metamorphosis:whole body |
| 17806 | NIH_XGC_tropMet9 | 4 | Buchholz/Shi:JGI | metamorphosis:whole body |
| 17807 | NIH_XGC_tropMet10 | 271 | Buchholz/Shi:JGI | metamorphosis:whole body |
| 20560 | NICHD_XGC_tropInt_54 | 2835 | Gerhard | adult:intestine |
| 20561 | NICHD_XGC_tropInt_60 | 3009 | Gerhard | adult:intestine |
| 20562 | NICHD_XGC_tropInt_62 | 3466 | Gerhard | adult:intestine |
| 20682 | Xenopus tropicalis embryo gastrula | 56305 | Ueno | gastrula:whole body |
| 20886 | NICHD_XGC_tropBone1 | 13517 | Grammer/Harland:JGI | adult:bone |
| 20891 | NICHD_XGC_tropInt_66 | 3642 | Gerhard | unknown:intestine |
| 20892 | NICHD_XGC_tropInt_63 | 3545 | Gerhard | unknown:intestine |
| 20901 | NICHD_XGC_tropThy1 | 15267 | Flajnik:JGI | adult:thymus |
| 20911 | NICHD_XGC_tropLimb_m | 15360 | Pollet:JGI | metamorphosis:limb |
| 20912 | NICHD_XGC_tropSp1 | 16700 | Flajnik:JGI | adult:spleen |
| 20931 | NICHD_XGC_tropPanc1 | 10331 | Blitz/Cho:JGI | adult:pancreas |
| 20947 | NICHD_XGC_tropTail_m | 16139 | Pollet:JGI | metamorphosis:tail |

| | | | | |
|---|---|---|---|---|
| 20953 | NICHD_XGC_tropTe1 | 22892 | Grammer/Harland:JGI | adult:testis |
| 20954 | NICHD_XGC_trop_25 | 30387 | Grammer/Harland:JGI | tailbud embryo:whole body |
| 21298 | NICHD_XGC_tropEye1 | 14675 | Grainger:Express Genomics | adult:head |

**cDNA library strategy**

To augment the amount and diversity of the initial collection of *Xenopus tropicalis* EST data, new libraries were constructed from stages or organs where considerable additional diversity of cNDAs was expected; these included stage 25 whole embryo, limb and tail tissues from metamorphic stage organisms, eye, bone, and immune system from adult tissues. In addition, further sequencing was carried out from two existing libraries: adult brain and testis, where the EST coverage indicated that there was considerable additional diversity to be explored. In addition, new libraries gave us an opportunity to harmonize library production in the vector pCS107/8, which can be used for direct expression of mRNA. Figure S1 illustrates the success of this approach, with the wide range of embryonic stages and adult tissues sampled at a (relatively) uniform depth.

Subsequent analysis of the EST data from this part of the project, with clone selection being internationally coordinated to minimize redundancy, fed into the final phase of full-insert sequencing.

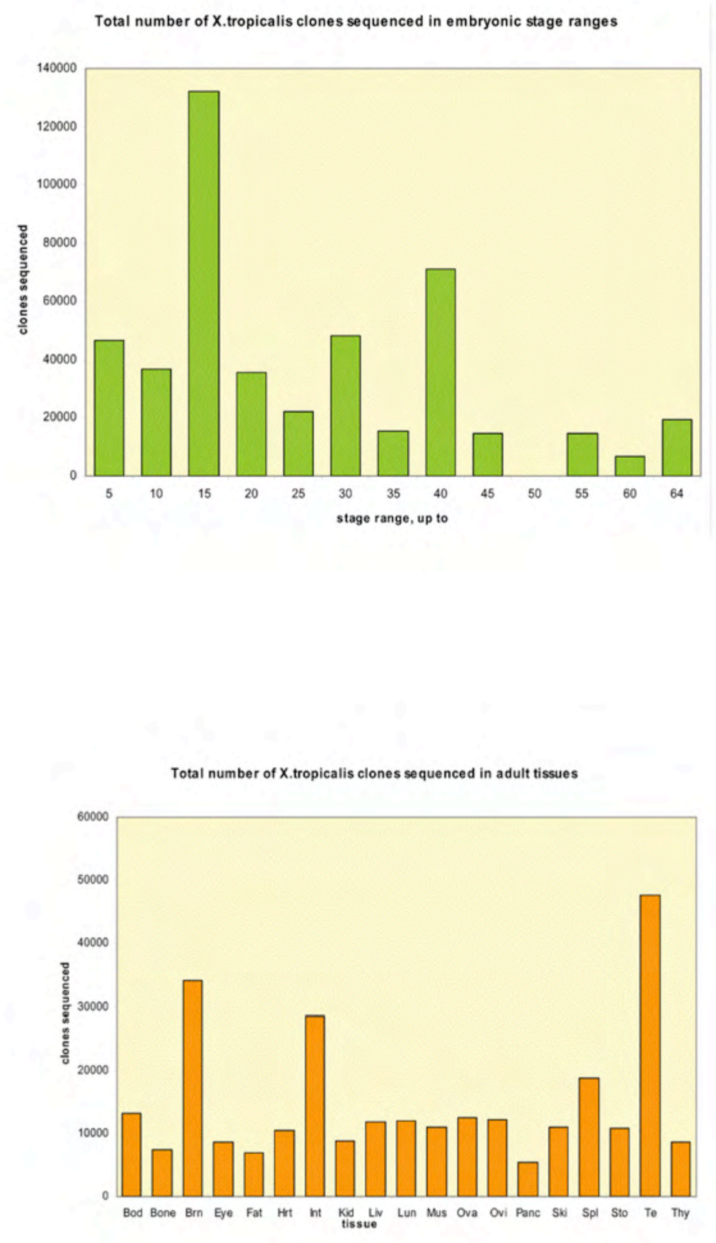


**Figure S1: Distribution of numbers of EST sequenced clones over the range of embryonic development stages and adult tissues showing effective exploration of gene 'expression space'. (a)**

**distribution of number of clones sequenced by approximate developmental stage. For libraries generated from a range of stages, clone counts have been distributed equally over the range eg. Gastrula = stages 10-13. Egg is counted as stage 1. Stage count are then grouped into bins of five for the figure. (b) distribution of number of clones by tissue type; all tissue types are for adult tissues.**

## Using EST data from adult tissues to predict embryonic gene function

The large collection of EST data from well-defined and diverse libraries can be used to make functional (although somewhat general) predictions about gene behavior, or, for example, to help focus searches for genes expressed in embryonic tissues, of which we present two examples here. As the ESTs are derived from staged embryos and specific adult tissues, and because the libraries are generally not normalized, the EST counts provide a representative sampling of transcript populations for different tissues or stages. This allows us to easily distinguish between genes expressed at low and high levels, and between ubiquitous and tissue specific genes. A list of identified tissue-specific genes is provided as an excel sheet (Supplementary data 1).

Embryonic libraries (especially from earlier stages) are generally made from the whole body, so genes involved (for example) in specific embryogenis programs are hard to discern from EST data alone. However, genes with restricted expression in specific adult tissues are also likely be expressed in the equivalent developing embryonic or tadpole tissue, and may be important in the development of that tissue. To test this, we selected for *in situ* hybridization a small set of clones from genes with predominant expression in adult brain or liver. Many of these did show embryonic tissue-specific expression, though in the case of the liver-specific ESTs, their expression was restricted to the endodermal germ layer, but was not necessarily liver-specific. See Figure S2 and Tables S5 and S6 for the *in situ* expression images, and the genes and EST clusters used to drive the experiment.

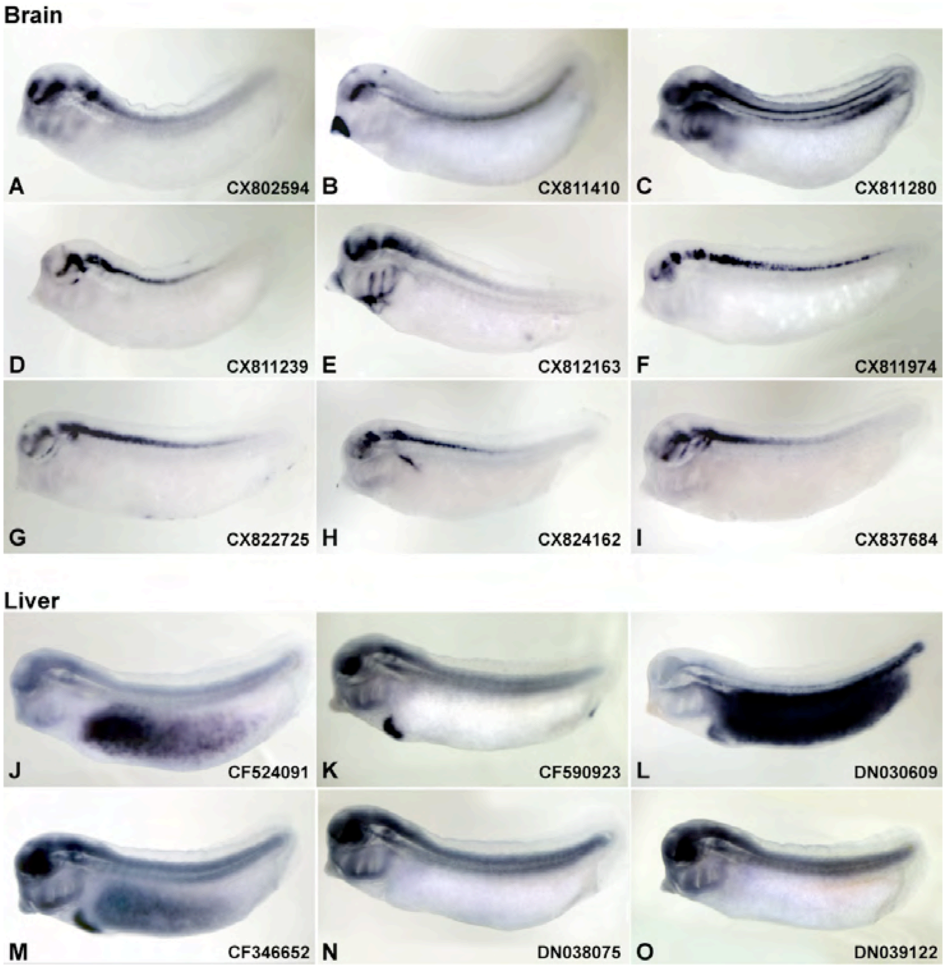


**Figure S2. Expression of EST clusters with high representation in adult brain or liver libraries at the tailbud tadpole stage. A-I. EST clusters in which 100% of the constituent ESTs were present in an adult brain library were analyzed by in situ hybridization for their expression at several stages during development. Examples of ESTs with strong expression in the nervous system at tailbud stages are shown here; of 14 ESTs tested, 7 showed expression only in the central nervous system, 3 showed expression in the CNS as well as in some other organ (such as the cement gland, B, or**

**branchial arches, C and E), 3 showed weak or background staining, and one showed ubiquitous expression. J-L. ESTs from clusters in which 95% or more of the constituents were expressed in the adult liver. Examples of ESTs with endodermal expression are shown. Of 9 ESTs tested, 5 showed strong endodermal expression, including intestine (J,M), liver (K,M), or pan-endodermal (L) expression. Three ESTs showed apparent neural expression which may have been nonspecific (N,O) and one had weak or background expression. Plasmid DNA from candidate ESTs in vector pCMVsport6 or CS107 was isolated by alkaline lysis, linearized with Sal1 or EcoR1 respectively, and purified on a Qiagen QIAquick column. Digoxigenin-labelled probes were prepared from these linear templates with T7 RNA polymerase and used to stain embryos by in situ hybridization . Tailbud stage embryos are shown with anterior to the left, genbank accession numbers are given in the bottom right of each image, additional identifying information is given in Table S5 and S6. For annotation, the cluster was blasted against the genome and the identity of the gene determined through sequence similarity and synteny with the mammals; where no synteny is apparent, the blast similarity is noted.**

**Clusters from http://genomics.nimr.mrc.ac.uk/online/xt-fl-db.html**

**Table S5: Developmental expression of ESTs highly represented in adult brain libraries.**

| Genbank ID | Image clone | Cluster | Annotation | Expression at tailbud stages | Figure letter |
|---|---|---|---|---|---|
| CX802594 | 7645482 | Xt7.1-CAAJ16299.3 | none | Neural | A |
| CX805047 | 7647121 | Xt7.1-CAAJ16990.5 | syngr3 | Weak/background | -- |
| CX805307 | 7646920 | Xt7.1-CAAJ17138.3 | gprin1 | Weak/background | -- |
| CX805074 | 7647173 | Xt7.1-CAAJ17007.5 | none | Neural | -- |
| CX811410 | 7650370 | Xt7.1-CAAJ14326.5 | map6 | Brain, cement gland | B |
| CX811280 | 7650447 | Xt7.1-CAAJ12186.5 | slit1 | Neural, branchial arches, eye, posterior mesoderm | C |
| CX811239 | 7650489 | Xt7.1-CAAK1859.5 | islr2 | Neural | D |
| CX812163 | 7651056 | Xt7.1-CAAJ23320.5 | pea15 | Neural, branchial arches, heart | E |
| CX811974 | 7650933 | Xt7.1-CAAJ23213.5 | gad1 | Neural | F |
| CX812278 | 7650930 | Xt7.1-CAAJ23389.5 | C16orf45 | Ubiquitous | -- |
| CX812864 | 7651300 | Xt7.1-CAAJ23717.5 | gria2 | Weak/background | -- |
| CX822725 | 7653983 | Xt7.1-CAAK2773.3 | sv2a | Neural | G |
| CX824162 | 7654657 | Xt7.1-CAAK3547.5 | sez6L | Neural | H |
| CX837684 | 7658446 | Xt7.1-CAAK7946.3 | fbxl16 | Neural | I |

**Table S6: Developmental Expression of ESTs highly represented in adult liver libraries.**

| | | | | | |
|---|---|---|---|---|---|
| CF346652 | 6997954 | Xt7.1-XZT53274 | itih2 | Liver, intestine, neural | M |
| CF524143 | 7017636 | Xt7.1-IMAGE:7017636.5 | mgc89221 | Ventral endoderm | -- |
| CF524091 | 7017690 | Xt7.1-CAAR7812.3 | fga | intestine | J |
| CF590923 | 7023535 | Xt7.1-CABC7493.3 | tdo2 | Liver, neural | K |
| DN030609 | 7740202 | Xt7.1-CAAR7993.3.5 | a2ml1 | All endoderm | L |
| DN038075 | 7744611 | Xt7.1-CAAR2659.5 | collagen IV | Neural/nonspecific | N |

## Supplementary Note 3. Identification and characterization of transposable/repetitive elements

**Table S7: Content of TEs in the frog genome**

| Classes of TEs | Percent of the genome<br>% |
|---|---|
| **Total DNA transposons** | **25** |
| “cut and paste”: | |
| *hAT* | 6.1 |
| ***Kolobok*** | 5.8 |
| ***Harbinger*** | 4.7 |
| ***Mariner (Tc1, Pogo groups)*** | 4.7 |
| ***PiggyBac*** | 1.3 |
| ***Merlin*** | < |
| Unclassified | 1.9 |
| “rolling circle” *Helitrons* | 0.6 |
| “self-synthesizing” *Polintons* | 0.01 |
| **Total retrotransposons** | **9** |
| LTR retrotransposons: | |
| Gypsy | |
| BEL | 1.3 |
| Copia | 0.3 |
| ERV I | 0.02 |
| ERV III | 0.1 |

| | |
|---|---|
| Unclassified | 0.03 |
| DIRS | 0. |
| Non-LTR retrotransposons: | 0.6 |
| *CR1, L2, REX1* clades | |
| SINEs | 3.8 |
| L1 (L1, Tx1 clades) | 0.4 |
| Penelope | 1.2 |
| | 0.9 |
| **Unclassified TEs** | **0.5** |
| **Total TEs** | **34.5** |

**Table S8: Diversity of LTR retrotransposons in eukaryotes**

| Species | Copia | BEL | Gypsy | ERV | | |
|---|---|---|---|---|---|---|
| | | | | I | II | III |
| Mammals | - | - | - | + | + | + |
| Chicken | - | - | - | + | + | + |
| Frog | + | + | + | + | - | + |
| Zebrafish | + | + | + | + | - | - |
| Insects | + | + | + | - | - | - |
| Nematodes | + | + | + | - | - | - |
| Fungi | + | - | + | - | - | - |
| Plants | + | - | + | - | - | - |

## Most of the transposable elements are DNA transposons

In bulk, copies of DNA transposons comprise 72% of all TEs, making the frog genome unique among all other studied animals, where the major TE-derived portion of the genome is composed of retrotransposons (*9-13*). All known classes of DNA transposons are present in the frog genome. These include "cut and paste" transposons (*14*), rolling-circle *Helitrons* (*15*), and self-synthesizing *Polintons* (*16*).

Five superfamilies of “cut and paste” DNA transposons (*hAT*, *Harbinger*, *Mariner, piggyBac,* and *Kolobok;* Table S7) are most prolific in terms of a proportion of the frog genome made up of their copies. While the first four superfamilies are well established (*14, 17*), Kolobok is a novel superfamily of DNA transposons. We identified an autonomous *Kolobok-1_XT* transposon that codes for a *Kolobok* transposase (TPase), which is not similar to known proteins, excluding those encoded by *Kolobok* transposons spread in genomes of vertebrates, insects, and nematodes. In addition to the TPase, the precise 4-bp TTAA target site duplications and terminal inverted repeats with the 5’-AG termini form distinctive hallmarks of this superfamily. The same TTAA target site duplication is also a characteristic of the *piggyBac* superfamily; however *piggyBac* transposons have the 5’-CC termini that bind to the *piggyBac* TPase, which is not similar to the *Kolobok* TPase.

### ***Helitron* transposons**

After discovery of *Helitrons* in plants and nematodes (*15*), a few ORFs coding for proteins similar to the *Helitron* replicase/helicase have been identified in the fish genomes (*18*). In vertebrates, the Helitron replicase/helicase contains an additional domain derived from an endonuclease encoded by CR1 non-LTR retrotransposons. At the same time, no single full-size autonomous or non-autonomous *Helitron* has been identified so far. Importantly, terminal sequences of vertebrate *Helitrons* have not been reported. As a result, given the conserved CR1-like endonuclease, it was not clear how much the recruited endonuclease might have changed the rolling-circle transposition mechanism typical for standard *Helitrons* (*15*). We derived consensus sequences of two families of non-autonomous frog *Helitrons* (*Helitron-N1_XT*, *Helitron-N2_XT*; elements from each family are less than 10% divergent from their consensus sequences). Based on identification of numerous insertions of these elements into copies of other TEs, we found that vertebrate *Helitrons* preserve main features of standard *Helitrons*, including duplication-free insertions into ApT target sites and structural hallmarks of the termini (*15*). This finding implies that recruitment of the endonuclease by the vertebrate *Helitrons* did not change the mechanism of transposition drastically.

While above 0.6% of the frog genome is made up of copies of non-autonomous *Helitrons*, the genome contains only a few copies of autonomous *Helitrons* that belong to two families (*Helitron-1_XT* and *Helitron-2_XT*).

### LTR retrotransposons

TEs from all major superfamilies of LTR retrotransposons, including *Copia*, *BEL*, *Gypsy,* and endogenous retroviruses (classes I and III), populate the frog genome.
In terms of such diversity, frog is a champion among all other eukaryotes (Table S8). It is possible, given that the class III endogenous retroviruses form the oldest retroviral fossils preserved in the mammalian genomes, present in the chicken genome (*9*) and absent in the fish genome, that they have been evolved endogenously from a *Gypsy*-like retrotransposon in a common ancestor of mammals, birds, and amphibians some 400-500 million years ago.

Since *Copia, BEL, and Gypsy* elements are not present in birds and mammals, excluding some genes derived from *Gypsy*-encoded proteins in common ancestors of mammals or mammals and birds, it is likely that all these LTR retrotransposons have become immobile in common ancestors of mammals and reptiles after their split from amphibians.

## ***DIRS*** **and** ***Penelope*** **retrotransposons**

About 1% of the frog genome is made of recently transposed copies of *DIRS* retrotransposons. *DIRS* form a separate class of retrotransposons that have probably evolved from an ancestral *Gypsy* LTR retrotransposon after its DDE integrase was replaced by a tyrosine recombinase (*19*). The frog genome contains over 100 highly diverse families of DIRS (each family is composed of a few copies that are >90% identical to each other). We derived consensus sequences of 52 DIRS families. Some of these families can still be active (less than 2% divergence of copies from their consensuses, and all three ORFs coding for *DIRS*-specific proteins are free of stop-codons).

Another 1% of the frog genome is composed of *Penelope* retrotransposons that form the most ancient class of eukaryotic retrotransposons (*20*). Frog *Penelope* elements are characterized by frequent 5' truncations; they usually form inverted structures, similar to those observed in fish *Penelopes* (*21*). We reconstructed consensus sequences for 20 families of frog Penelopes. Some of these families (*Penelope-5_XT*, *Penelope-6_XT*) have been very active in the last few million years (>1000 copies, 3% divergence from the consensus sequences), another families (*Penelope-7_XT*) are quite old (10% divergence from the consensus), and families like Penelope-10_XT are very young (<1% divergence) and are composed of only a few copies, which may still be mobile.

Since *DIRS* and *Penelope* elements are not present in the genomes of birds and mammals, they likely were immobilized in common ancestors of birds and mammals, after their split from amphibians, and later became extinct due to random mutations and deletions of their immobile genomic copies.

**Non-LTR retrotransposons**

All non-LTR retrotransposons identified in frog constitute above 5.4% of the genome (Table S7). They can be classified into five clades (CR1, L2, Rex1, L1, and Tx1) which are wide spread in eukaryotes, including birds and mammals (*9-11*). However, in contrast to the genomes of birds (contain only CR1s) and mammals (L1s and very old fossilized L2s), the frog genome contains young families of CR1, L2, Rex1, L1 and Tx1 elements.

The frog genome harbors more than 100 families of L1 and Tx1 retrotransposons that are usually composed of a small number of copies (sometimes just one). We derived consensus sequences of 68 young L1 and Tx1 families (1% to 5% intra-family divergence; inter-family divergence >30%). The evolution of L1/Tx1 in the frog genome differs from that of non-LTR retrotransposons in mammals and birds (*9-10*) and is similar to the evolution of non-LTR retrotransposons in plants and insects (*12, 22*). While ~100 of highly divergent families of frog retrotransposons were active approximately at the same time, only a few lineages of non-LTR retrotransposons were active simultaneously in mammals and birds.

Several families of the frog Tx1 retrotransposons are characterized by remarkable target-site specificity. For instance, retrotransposons from four different young families, *L1-52_XT* to *L1-56_XT*, are inserted at the same target site in different copies of U2 small RNA. *L1-60_XT* and *L1-61_XT* elements are inserted at the same target sites in copies of the *MSAT2_XT* and *Sat2_XT* satellites, respectively. Copies of *Tx1_XT*, which is 85% identical to the Tx1 from *Xenopus laevis,* are inserted into the same site in copies of the *piggyBac-N1_XT* transposon.

**SINE elements**

Despite presence of diverse young L1, Tx1, CR1, L2 and Rex1 non-LTR retrotransposons there are no active SINE elements in the frog genome. However it contains two families of SINEs that became immobile millions of years ago: SINE2-1_XT (constitute 0.33% of the genome) and MIR_XT (0.01% of the genome). Both SINEs contain tRNA-derived pol III internal promoters and have been retrotransposed using reverse transcriptase and endonuclease activities provided by L2 retrotransposons.

## Methods

New families of transposable elements were identified using CENSOR (*23*). First, we detected all fragments of the frog genome coding for proteins similar to catalytic cores of transposases, reverse transcriptases, and DNA polymerases representing all known classes of TEs collected in Repbase (*24*). The detected DNA sequences have been clustered based on their pairwise identities by using BLASTclust from the standalone NCBI BLAST package (the pairwise DNA identity threshold was equal to 80%). Each cluster has been treated as a potential family of TEs described by its consensus sequence. The consensus sequences were built automatically based on multiple alignments of the cluster sequences expanded in both directions and manually modified based on structural characteristics of known TEs. A library of TEs was produced by merging the identified consensus sequences with DNA sequences of *X. tropicalis* TEs reported previously in literature and collected in Repbase. Using CENSOR, we identified genomic copies of TEs similar to the library sequences. Second, given known consensus sequences of the library TEs, we detected automatically all putative insertions longer than 50-bp present in the identified genomic copies of the library TEs. The identified insertions have been treated as putative novel TEs not similar to the library TEs. They have been clustered based on their pairwise DNA identities using BLASTclust. In each cluster, a consensus sequence was derived based on multiple alignment of the cluster sequences. After manual refinements of the consensus sequences, the identified families of TEs were classified based on their structural hallmarks, including target site duplications, terminal repeats, encoded proteins and similarities to TEs classified previously (*25*). Identified TEs are deposited in Repbase.

## Supplementary Note 4. Identification and characterization of protein-coding genes

Using homology-based gene prediction methods and the deep *Xenopus* ESTs and cDNAs resources we identified 27,415 candidate protein-coding loci and 35,996 transcripts. This overestimates the actual gene count, partly due to genes extending over multiple small scaffolds, and partly due to our generous inclusion of single-exon gene candidates. Transcript assemblies were made by PASA (*3*) from *Xenopus tropicalis* ESTs/cDNAs using *X. tropicalis* genome assembly Xentr4 as reference and criteria of 95% identity and 50% coverage (*X. tropicalis* PASA), and transcript assemblies from *X. laevis* ESTs/cDNAs using *X. tropicalis* genome assembly Xentr4 as reference and criteria of 90% identity and 70% coverage (*X. laevis* PASA). ESTs/cDNAs were downloaded from NCBI. *X. tropicalis* genome sequences were repeat-masked by RepeatMasker (*26*). Both sets of transcript assemblies were aligned to *X. tropicalis* repeat-masked genome using blat, and human and chicken (ENSEMBL release 55) peptides were aligned using NCBI BLASTX. Putative gene loci were determined based on blat alignments and BLASTX alignments with possible extension of 500 BP at either end. Best ORFs for transcript assemblies was obtained by studying 3-frame translation homology to human peptides (-e 1E-5) or longest ORFs were kept if no homology was found and if the ORF is at least 150 BP long. Human and chicken peptides, and transcript assembly ORFs at a given locus were used as protein templates for both GenomeScan (*27*) and Fgenesh+ (*28*) gene predications along with locus location as range constraint. Gene predictions were fed into *X. laevis* PASA for 2 rounds of annotation comparison and update. Gene models from *X. laevis* PASA were fed into *X. tropicalis* PASA for another 2 rounds of annotation comparison and update. Gene model transcripts have a valid flag if PASA has improved and validated transcripts based on ESTs/cDNA alignments.

Peptides of gene models from *X. tropicalis* PASA were aligned to human and chicken peptides for homology and synteny analysis. Gene models were discarded if their CDS overlap with repeats exceeds 20%. After filtering for repeats, all transcripts in a locus were kept if they were validated by PASA runs while only one transcript (longest CDS length) was kept if it has ESTs/cDNA, homology support, or synteny to human or chicken. Transcripts from *X. tropicalis* annotation version 4.1 that have synteny to those in the human genome but are not represented by any gene model in the current annotation were promoted (507 gene models). All candidate loci are supported by EST evidence or peptide homology to human or chicken, with 86% being supported over at least 80% of the CDS length by either ESTs and/or sequence homology and 55% being supported over at least 80% of the CDS length by ESTs/cDNAs alone.

We believe the inferred 27,415 candidate loci to be an overestimate of the true gene count for two main reasons. First, there are only 12,015 gene models on the 272 largest scaffolds, which contain half the total assembled genomic sequence, which would suggest a total gene count of ~24,000, with the extra 3000+ gene models being due to exons from single genes covering multiple small scaffolds. Second, of the 7145 genes on the largest scaffolds which have confirmed human or chicken orthologs, only 786 (11%) are one- or two-exon genes. In comparison, 2,267 of the remaining 4,870 genes without human or chicken orthologs (47%) have one or two exons. This suggests that as many as 36% or 1750 genes of the set without orthologs could be annotated single-exon genes in pseudogenic regions. If this interpretation is correct, the total number of genes could be as low as 2 x (12,000 - 1,750) ~ 20,500.

**Table S9: Large tandem expanded gene families in the frog genome. Position on scaffolds and number of members in the clusters are shown, in addition to the most frequent PFAM domain in each cluster, with description.**

| Scaffold | Position | Count | PFAM | Description |
|---|---|---|---|---|
| scaffold_24 | 4504-1748538 | 160 | PF00001 | 7-transmembrane receptor |
| scaffold_546 | 18708-430931 | 43 | PF00028 | Cadherin domain |
| scaffold_442 | 683044-1065308 | 38 | PF00001 | 7-transmembrane receptor |
| scaffold_63 | 316148-1065922 | 33 | PF00067 | Cytochrome P450 |
| scaffold_676 | 161319-498850 | 32 | PF00096 | Zinc finger, C2H2 |
| scaffold_190 | 1279831-1972459 | 31 | PF01094 | Receptor family ligand binding |
| scaffold_899 | 5599-310571 | 30 | PF00001 | 7-transmembrane receptor |
| scaffold_677 | 45143-552454 | 30 | PF00096 | Zinc finger, C2H2 |
| scaffold_680 | 15280-519617 | 29 | PF07562 | Nine Cysteines Domain of family 3 GPCR |
| scaffold_91 | 1870444-2180530 | 28 | PF00001 | 7-transmembrane receptor |
| scaffold_325 | 275380-731628 | 28 | PF00001 | 7-transmembrane receptor |
| scaffold_535 | 290585-683826 | 28 | PF00096 | Zinc finger, C2H2 |
| scaffold_290 | 18832-462495 | 27 | PF00001 | 7-transmembrane receptor |
| scaffold_315 | 943681-1254697 | 27 | PF00001 | 7-transmembrane receptor |
| scaffold_657 | 420760-603725 | 27 | PF00001 | 7-transmembrane receptor |
| scaffold_882 | 2844-321753 | 27 | PF00001 | 7-transmembrane receptor |

| | | | | |
|---|---|---|---|---|
| scaffold_942 | 5207-272885 | 27 | PF00001 | 7-transmembrane receptor |
| scaffold_675 | 38100-445169 | 26 | PF01094 | Receptor family ligand binding |
| scaffold_150 | 2011882-2270784 | 26 | PF00001 | 7-transmembrane receptor |
| scaffold_532 | 24438-229690 | 26 | PF00001 | 7-transmembrane receptor |
| scaffold_406 | 739581-1069127 | 26 | PF00001 | 7-transmembrane receptor |
| scaffold_34 | 2576260-2850813 | 26 | PF00001 | 7-transmembrane receptor |
| scaffold_913 | 4514-291939 | 25 | PF00001 | 7-transmembrane receptor |

Regions of tandem-expanded gene families were identified clusters of genes along a scaffold whose corresponding peptides were showing similarity at a BLAST expectation value smaller than 0.001, and with a maximum allowed number of 2 non-similar genes on any strand betweem any two members of the cluster. The largest clusters are shown in Table S9.

## Supplementary Note 5. Human disease gene orthologs in frog

We evaluated *X. tropicalis* genes that are orthologous to human disease related genes using online databases and a recently published disease classification system. Online Mendelian Inheritance in Man (*29*) is a comprehensive and continually updated database of human genes and genetic phenotypes (*30*). OMIM contains information on all types of heritable traits, not just diseases. The OMIM Morbid Map is a catalog diseases described in OMIM. Allelic variants for a disease are not always included in OMIM, and the criterion for inclusion are distinctive phenotype, high population frequency, historic significance, and unusual mechanism of mutation. Therefore, diseases with annotated allelic variants have a more established relationship between genotype and phenotype.

Many of these disease related OMIM terms have been mapped to human genes (Entrez gene IDs) and are available for download at NCBI (ftp://ftp.ncbi.nlm.nih.gov/gene/). Because our proteome data for other organisms was downloaded from Ensembl, we mapped these to ensembl gene IDs. Results are shown in Tables S10-S13. Orthologous genes were determined by performing a blastp alignment of the proteomes of each organism under consideration against the human proteome using a BLOSUM45 matrix, and subsequently defining orthologs as reciprocal best hits from this alignment.

Using a human disease classification for disease-associated OMIM terms defined by Goh et. al, we examined the distribution of *tropicalis* orthologs for various disease classes as compared to other model organisms(*31*). Most disease classes show a comparable fraction of *tropicalis* orthologs present, and some such as metabolic, renal, and muscular show a higher percentage than other model organisms. When only OMIM terms that have confirmed allelic variants are used, opthamological, metabolic, and nutritional categories show a greater percentage of orthologs than other model organisms, including mammals.

While *X. tropicalis* is diploid, *X. laevis* is tetraploid, and orthologs of disease genes might be expected to have been subject to selection, perhaps by subfunctionaliztion. Based on an EST-based collection of 20,223 genes in *X. laevis*, two *X. laevis* co-orthologs dating back to the tetraploidization event have been found for least 14% of the *X. tropicalis* genes with human orthologs (*32*). This is also true specifically for the disease-related genes, and since genes retained in multiple copies are often subfunctionalized(*33*), further analysis of *X. laevis* duplicates may provide finer resolution in studying the effects of such genes.

**Table S10: All OMIM Terms and Human disease-related genes.**

| | |
|---:|---|
| 20605 | OMIM Terms |
| 3605 | OMIM Terms in Morbid Map |
| 1697 | OMIM Terms in Morbid Map that are in a defined disease group(*31*). |
| 3552 | Human Genes associated with OMIM terms in Morbid Map |
| 2460 | Human Ensembl Ids associated with Gene Ids associated with Morbid Map |
| 873 | Orthologous genes in *D.melanogaster* |
| 1905 | Orthologous genes in *D.rerio* |
| 1924 | Orthologous genes in *X.tropicalis* |
| 1836 | Orthologous genes in *G.gallus* |
| 2335 | Orthologous genes in *M. musculus* |

**Table S11: OMIM Terms with at least one identified allelic variant**

| | |
|---:|---|
| 20605 | OMIM Terms |
| 2251 | OMIM Terms in Morbid Map |
| 1617 | OMIM Terms in Morbid Map that are in a defined disease group (31). |
| 2249 | Human Genes associated with OMIM terms in Morbid Map |
| 2229 | Human Ensembl Ids associated with Gene Ids associated with Morbid Map |
| 801 | Orthologous genes in *D.melanogaster* |
| 1747 | Orthologous genes in *D.rerio* |
| 1761 | Orthologous genes in *X.tropicalis* |
| 1668 | Orthologous genes in *G.gallus* |
| 2131 | Orthologous genes in *M. musculus* |

**Table S12: Orthologous genes associated with OMIM diseases and the distribution amongst different categories.**

| | Dme | Dre | Xtr | Gga | Mmu | Hsa |
|---|---:|---:|---:|---:|---:|---:|
| #total orthologs | 6078 | 11467 | 12884 | 11898 | 16884 | 23517 |
| #total in disease class | 705 | 1622 | 1600 | 1555 | 1959 | 2052 |
| Cancer | 82 | 168 | 165 | 172 | 198 | 207 |
| Cancer (%) | 1.349 | 1.465 | 1.281 | 1.446 | 1.173 | 0.88 |
| Renal | 12 | 35 | 42 | 43 | 53 | 57 |
| Renal (%) | 0.197 | 0.305 | 0.326 | 0.361 | 0.314 | 0.242 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Muscular | 20 | 62 | 54 | 44 | 63 | 65 |
| Muscular (%) | 0.329 | 0.541 | 0.419 | 0.37 | 0.373 | 0.276 |
| Gastrointestinal | 10 | 31 | 25 | 24 | 33 | 34 |
| Gastrointestinal (%) | 0.165 | 0.27 | 0.194 | 0.202 | 0.195 | 0.145 |
| Nutritional | 1 | 15 | 17 | 15 | 21 | 22 |
| Nutritional (%) | 0.016 | 0.131 | 0.132 | 0.126 | 0.124 | 0.094 |
| Skeletal | 14 | 48 | 43 | 46 | 55 | 56 |
| Skeletal (%) | 0.23 | 0.419 | 0.334 | 0.387 | 0.326 | 0.238 |
| Endocrine | 22 | 69 | 68 | 69 | 91 | 96 |
| Endocrine (%) | 0.362 | 0.602 | 0.528 | 0.58 | 0.539 | 0.408 |
| Ophthamological | 27 | 102 | 97 | 85 | 113 | 118 |
| Ophthamological (%) | 0.444 | 0.89 | 0.753 | 0.714 | 0.669 | 0.502 |
| Respiratory | 7 | 20 | 23 | 24 | 33 | 33 |
| Respiratory (%) | 0.115 | 0.174 | 0.179 | 0.202 | 0.195 | 0.14 |
| Dermatological | 17 | 50 | 52 | 47 | 73 | 77 |
| Dermatological (%) | 0.28 | 0.436 | 0.404 | 0.395 | 0.432 | 0.327 |
| Metabolic | 151 | 226 | 229 | 198 | 254 | 260 |
| Metabolic (%) | 2.484 | 1.971 | 1.777 | 1.664 | 1.504 | 1.106 |
| Neurological | 109 | 222 | 213 | 199 | 239 | 250 |
| Neurological (%) | 1.793 | 1.936 | 1.653 | 1.673 | 1.416 | 1.063 |
| Psychiatric | 10 | 24 | 21 | 25 | 27 | 30 |
| Psychiatric (%) | 0.165 | 0.209 | 0.163 | 0.21 | 0.16 | 0.128 |
| Bone | 10 | 31 | 30 | 32 | 43 | 44 |
| Bone (%) | 0.165 | 0.27 | 0.233 | 0.269 | 0.255 | 0.187 |
| Developmental | 21 | 49 | 46 | 45 | 50 | 53 |
| Developmental (%) | 0.346 | 0.427 | 0.357 | 0.378 | 0.296 | 0.225 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Immunological | 9 | 70 | 72 | 83 | 105 | 116 |
| Immunological (%) | 0.148 | 0.61 | 0.559 | 0.698 | 0.622 | 0.493 |
| Connective tissue | 9 | 25 | 29 | 34 | 44 | 51 |
| Connective tissue (%) | 0.148 | 0.218 | 0.225 | 0.286 | 0.261 | 0.217 |
| Hematological | 35 | 92 | 99 | 92 | 125 | 135 |
| Hematological (%) | 0.576 | 0.802 | 0.768 | 0.773 | 0.74 | 0.574 |
| Ear,Nose,Throat | 13 | 31 | 31 | 32 | 44 | 43 |
| Ear,Nose,Throat (%) | 0.214 | 0.27 | 0.241 | 0.269 | 0.261 | 0.183 |
| Multiple | 100 | 185 | 177 | 172 | 204 | 210 |
| Multiple (%) | 1.645 | 1.613 | 1.374 | 1.446 | 1.208 | 0.893 |
| Cardiovascular | 26 | 67 | 67 | 74 | 91 | 95 |
| Cardiovascular (%) | 0.428 | 0.584 | 0.52 | 0.622 | 0.539 | 0.404 |

**Table S13: Othologous genes associated with OMIM diseases that have at least one identified allelic variant, and the distribution amongst different categories.**

| | Dme | Dre | Xtr | Gga | Mmu | Hsa |
|---|---|---|---|---|---|---|
| #total orthologs | 6078 | 11467 | 12884 | 11898 | 16884 | 23517 |
| #total in disease class | 681 | 1555 | 1537 | 1489 | 1881 | 1967 |
| Cancer | 65 | 126 | 123 | 131 | 150 | 156 |
| Cancer (%) | 1.069 | 1.099 | 0.955 | 1.101 | 0.888 | 0.663 |
| Renal | 12 | 35 | 42 | 43 | 53 | 56 |
| Renal (%) | 0.197 | 0.305 | 0.326 | 0.361 | 0.314 | 0.238 |
| Muscular | 20 | 59 | 52 | 43 | 61 | 63 |
| Muscular (%) | 0.329 | 0.515 | 0.404 | 0.361 | 0.361 | 0.268 |
| Gastrointestinal | 9 | 30 | 24 | 22 | 31 | 32 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Gastrointestinal (%) | 0.148 | 0.262 | 0.186 | 0.185 | 0.184 | 0.136 |
| Nutritional | 0 | 13 | 15 | 13 | 19 | 19 |
| Nutritional (%) | 0 | 0.113 | 0.116 | 0.109 | 0.113 | 0.081 |
| Skeletal | 14 | 47 | 41 | 44 | 53 | 54 |
| Skeletal (%) | 0.23 | 0.41 | 0.318 | 0.37 | 0.314 | 0.23 |
| Endocrine | 22 | 69 | 68 | 69 | 91 | 96 |
| Endocrine (%) | 0.362 | 0.602 | 0.528 | 0.58 | 0.539 | 0.408 |
| Ophthamological | 27 | 102 | 97 | 85 | 113 | 118 |
| Ophthamological (%) | 0.444 | 0.89 | 0.753 | 0.714 | 0.669 | 0.502 |
| Respiratory | 7 | 19 | 22 | 22 | 31 | 31 |
| Respiratory (%) | 0.115 | 0.166 | 0.171 | 0.185 | 0.184 | 0.132 |
| Dermatological | 17 | 50 | 52 | 47 | 73 | 77 |
| Dermatological (%) | 0.28 | 0.436 | 0.404 | 0.395 | 0.432 | 0.327 |
| Metabolic | 151 | 225 | 227 | 196 | 252 | 258 |
| Metabolic (%) | 2.484 | 1.962 | 1.762 | 1.647 | 1.493 | 1.097 |
| Neurological | 108 | 219 | 210 | 195 | 235 | 245 |
| Neurological (%) | 1.777 | 1.91 | 1.63 | 1.639 | 1.392 | 1.042 |
| Psychiatric | 8 | 21 | 19 | 23 | 23 | 26 |
| Psychiatric (%) | 0.132 | 0.183 | 0.147 | 0.193 | 0.136 | 0.111 |
| Bone | 10 | 31 | 30 | 32 | 43 | 44 |
| Bone (%) | 0.165 | 0.27 | 0.233 | 0.269 | 0.255 | 0.187 |
| Developmental | 21 | 49 | 46 | 45 | 50 | 53 |
| Developmental (%) | 0.346 | 0.427 | 0.357 | 0.378 | 0.296 | 0.225 |
| Immunological | 9 | 69 | 72 | 82 | 104 | 115 |
| Immunological (%) | 0.148 | 0.602 | 0.559 | 0.689 | 0.616 | 0.489 |
| Connective tissue | 9 | 25 | 29 | 34 | 43 | 50 |

|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
| Connective tissue (%) | 0.148 | 0.218 | 0.225 | 0.286 | 0.255 | 0.213 |
| Hematological | 35 | 92 | 99 | 92 | 125 | 134 |
| Hematological (%) | 0.576 | 0.802 | 0.768 | 0.773 | 0.74 | 0.57 |
| Ear,Nose,Throat | 13 | 29 | 30 | 31 | 43 | 42 |
| Ear,Nose,Throat (%) | 0.214 | 0.253 | 0.233 | 0.261 | 0.255 | 0.179 |
| Multiple | 98 | 181 | 174 | 169 | 200 | 206 |
| Multiple (%) | 1.612 | 1.578 | 1.351 | 1.42 | 1.185 | 0.876 |
| Cardiovascular | 26 | 64 | 65 | 71 | 88 | 92 |
| Cardiovascular (%) | 0.428 | 0.558 | 0.505 | 0.597 | 0.521 | 0.391 |

## Supplementary Note 6. Conserved synteny among frog, human and chicken

To compare the frog proteome to the human and chicken proteomes, we first performed all-against-all sequence alignments of predicted peptides, longest per locus, within all three species. For human and chicken we used ENSEMBL models versions 55. The peptides were aligned using BLASTp (*34*) with an e-value cut-off of $10^{-3}$. We assigned unique position IDs to all loci by numbering them in the order in which they occur on the chromosomes and scaffolds. Next, we scanned all genomes for putative tandem expanded families, here defined as clusters of peptides showing sequence similarity to neighboring genes, allowing a maximum of two intervening, non-participating genes on any strand. Such clusters were replaced with a single gene, the longest member of the cluster.

We then identified reciprocal highest scoring hits between the remaining genes in each pair of genomes (human-frog, frog-chicken, chicken-human). The overwhelming majority of hits in the all-against-all sequence comparison are due to weak, super-family level sequence similarity. We eliminated such hits by restricting further analysis to pairwise alignments of genes from two species in which the score of the alignment is at least 40% of that of the maximum of each of the members reciprocal best hits scores. This approach has the advantage of retaining hits of low scores that may be due to rapid sequence evolution between orthologs, while reducing hits between low-scoring distant paralogs. After filtering of the shorter members of tandem regions and genes without any hits satisfying the 40% criterion above, we were left with 14,334 human, 12,575 chicken, and 17,880 frog genes. These genes were re-numbered in strict consecutive orders in preparation for the study of detailed conserved synteny.

To identify regions of conserved synteny we implemented an algorithm similar to that described by Blanc. *et. al.*(*35*) in which the genomes are scanned for clusters of genes from a region in one genome, where each member shows sequence similarity to a member of a similar localized cluster within the other genome. The mapping of each such gene to its counterpart in the other genome forms can be visualized as rungs in a ladder, defining a block of conserved synteny. We allowed blocks to be as small as two rungs. Furthermore we allowed up to two intervening genes between any two rungs, to account for the possibility that orthologs may have been lost or gene models may be wrong. This resulted in 2,089 human-chicken blocks containing 12,712 rungs, 2,867 human-frog blocks with 12,953 rungs, and 2,396 chicken-frog blocks with 10,655 rungs. The largest human-chicken block contains 267 loci, while the largest chicken-frog block contains 74 loci. However, block sizes in human-frog and chicken-frog comparisons are limited by the size of assembled frog scaffolds.

To assess the number of false positive (FP) predictions of conserved synteny blocks, we applied the above algorithm to data where the gene order had been randomly scrambled by re-assigning new gene IDs to all existing gene positions. Such simulations showed that at most 3% of predicted rungs are FP. These are nearly all 2-rung blocks and will be eliminated by the 3-way conserved synteny requirements later in the analysis.

Not all blocks of conserved synteny are orthologous. In many cases, blocks are paralogous, originating in one of the two rounds of whole-genome duplication in the vertebrate ancestor (*36*). Such blocks tend to be much smaller than orthologous blocks due to more extensive scrambling of gene order by inversions since the last common ancestor of the region. To define putative synteny-confirmed orthologs for each pair of species, we sorted the segments in order from longest to shortest, then went through this list of segments and assigned the rungs in the segment as orthologs unless the corresponding genomic segments had already been masked by a longer segment. In ambiguous cases with two segments of the same length covering the same region we refrained from any ortholog calling. This procedure resulted in 9,759 human-chicken, 9,651 human-frog, and 7,885 chicken-frog unique orthologous pairs. Such synteny-confirmed orthologs are ubiquitous throughout the genomes. For example, half of all 45 kb segments and 98% of all 500 kb segments in the *X. tropicalis* genome overlap blocks of conserved synteny to human.

To further eliminate FP predictions and paralogous segments we restricted ourselves to three-way clusters of orthologous genes in which all pairs of genes are synteny-confirmed and consistent. We found 6,265 such clusters which act as genomic probes, associating locations on all three genomes with a single position on an ancestral tetrapod chromosome. Of these clusters, 5,645 were in blocks consisting of at least 15 genes scattered over an area of at least 2 Mb. We settled on this level of resolution for our genome comparison studies, and the corresponding data are included as Supplementary data 2.

## Association of regions with the meiotic map to predict syntenic Superscaffolds

**Table S14: Mapping of *X. tropicalis* scaffolds to linkage groups and chromosomes. Nearly two-thirds of the assembled sequence has been associated with linkage groups by markers and conserved synteny. Linkage groups have been mapped to chromosomes by FISH (*37*).**

| Chromosome | Linkage Group | # Scaffolds | Mb | # Genes |
|---|---|---|---|---|
| 1 | I | 134 | 151.0 | 2,582 |
| 2 | VI | 117 | 128.8 | 2,194 |
| 3 | VIII | 77 | 76.9 | 1,294 |
| 4 | VII | 76 | 100.4 | 1,836 |
| 5 | IX | 76 | 94.5 | 1,282 |
| 6 | II | 85 | 116.3 | 1,544 |
| 7 | IV | 99 | 87.5 | 1,406 |
| 8 | V | 96 | 101.1 | 2,045 |
| 9 | III | 58 | 73.9 | 1,456 |
| 10 | X | 43 | 33.6 | 805 |
| ***Total mapped*** | | 861 | 963.8 | 16,444 |
| ***Unmapped*** | | 18,898 | 396.2 | 10,971 |

Scaffolds were associated with linkage groups by means of 2,204 microsatellite markers from the existing linkage map at http://tropmap.biology.uh.edu/map.html. Microsatellite primer sequences were mapped onto the assembly by the short-read aligner BWA(*38*), essentially treating the forward and reverse primers like paired-end reads. 2,088 (95%) of the markers were successfully mapped. This approach revealed that some of the large scaffolds are likely misassembled hybrids of more than one linkage group. For example, the first 2.7 Mb of scaffold_2 contain five markers from LG3, whereas the remaining 4.7 Mb has 11 markers from LG9. Other large scaffolds with at least two markers from each of two linkage group are scaffolds 5, 112, 241, 252, 266 and 332.

A total of 1,696 markers from one of the major ten linkage groups (i.e. excluding unresolved clusters on the linkage map) could be used to uniquely assign a linkage group to 691 scaffolds. The results are included in (*39*). This allows conserved linkage to be studied on a much larger scale than the individual scaffold lengths would otherwise permit. Furthermore, blocks of conserved synteny to human and chicken are typically a few to tens of megabases long, which is much larger than the size of most scaffolds. Hence, in many cases where scaffolds with unassigned linkage group have conserved synteny to a region in human or chicken embedded within regions with conserved synteny to scaffolds which *do* have a consistently assigned linkage group, we will assign such scaffolds to the same linkage group by means of association. Such an assignment is only to be taken as a prdiction, not proof of linkage, though many such predictions have been verified in meiotic mapping, e.g., (*40*). Table S14 summarizes the assignments of scaffolds to linkage groups.

## Supplementary Note 7. Conserved sequence elements between frog and other vertebrates

**Whole-genome DNA Alignments.**

The selection of vertebrate species for whole genome sequencing was largely based on varying evolutionary distance with the central goal to better annotate and understand the genome of Homo sapiens. This resulted in the selection of various mammals (mouse), a bird (chicken), an amphibian (frog), and two fish species (fugu and zebrafish) to capture a diversity of Classes within the Subphylum Vertebrata. Each of these species has provided differing windows into the evolutionary history and constraint of the human genome.

To align genomes we have used the VISTA framework(*41*) with algorithms that combine both global and local alignment methods. First, we obtained a map of large blocks of conserved synteny between the two species by applying Shuffle-LAGAN global chaining algorithm(*42-43*) to local alignments produced by translated BLAT(*44*). After that we applied Supermap, the fully symmetric whole-genome extension to the Shuffle-LAGAN algorithm(*45*). Then, in each syntenic block we applied Shuffle-LAGAN a second time to obtain a more fine-grained map of small-scale rearrangements such as inversions.

**Comparative genomic analysis.** To explore the evolutionary history of the frog genome, we performed nucleotide alignments to the genomes of human (hg18), mouse (mm9), chicken (galGal3), fugu (fr2), and zebrafish (danRer5). Overall, we found only small differences in the total number of conserved elements in pair-wise comparisons between frog-human and frog-chicken. However, we observed approximately 30 to 40%

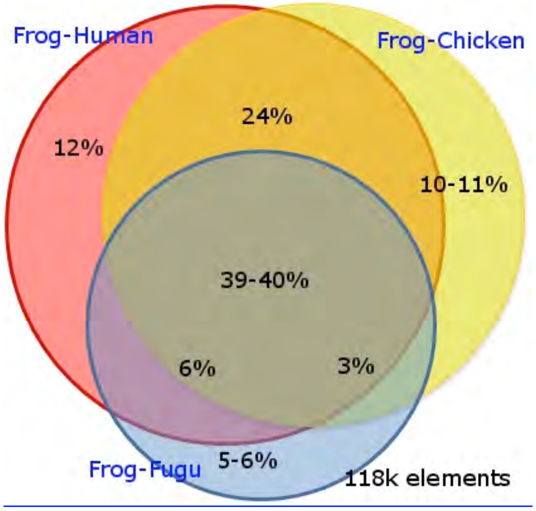
fewer conserved elements when comparing the frog and fish genomes. For instance, while we identified 127k conserved regions in frog-human and 126k in frog-chicken genome comparisons, we found 92k in frog-zebrafish and only 75k in frog-fugu comparisons (Table S15). This is in contrast to traditional comparative genomic views where the human genome has served as the reference for comparisons. In such studies, the number of conserved regions significantly decreases between human-mouse, human-chicken, human-frog, and human-zebrafish with 1.5M, 217k, 142k, and 92k conserved elements, respectively. (Table S15). This altered perspective from a frog-centric viewpoint reflects a distinct and almost equidistant position of the frog genome in the phylogeny of currently sequenced tetrapod genomes (with fish-frog displaying slightly more sequence divergence).

We next binned the frog conserved regions from each pair-wise genome comparison into coding and noncoding fractions to infer their functional nature. In frog-human comparisons, we found that 70% of frog conserved regions overlapped annotated gene exons, mRNA, ESTs, and/or gene predictions (Table S15). Thus, the remaining frog-human conserved regions appear noncoding and this fraction remained approximately the same in comparisons with other vertebrate genomes (the exception being frog-fugu where the noncoding fraction was ~16%) (Table S15). In total, we identified 35k frog-chicken, 32k frog-human, 23k frog-zebrafish and 12k frog-fugu conserved noncoding sequences (CNSs). Again, this is in contrast to human centric studies where human-mouse or human-chicken yield 1.2M and 87k CNSs, respectively. The equidistant nature of the frog genome to other tetrapods is further reflected in the observation that a minimum of 11% of frog CNSs was conserved in two or more vertebrate species (Figure S3 B). This overlap was the highest for comparisons among frog-human and frog-chicken CNSs (~80%). These results are consistent with the large evolutionary distance separating teleost fish (fugu and zebrafish) compared to amniote species (chicken and human). Thus, the existing sequenced vertebrate genomes provide limited additional value in functionally annotating the frog genome based on evolutionary constraint and suggests that additional amphibians genomes would be required to accomplish such a goal.

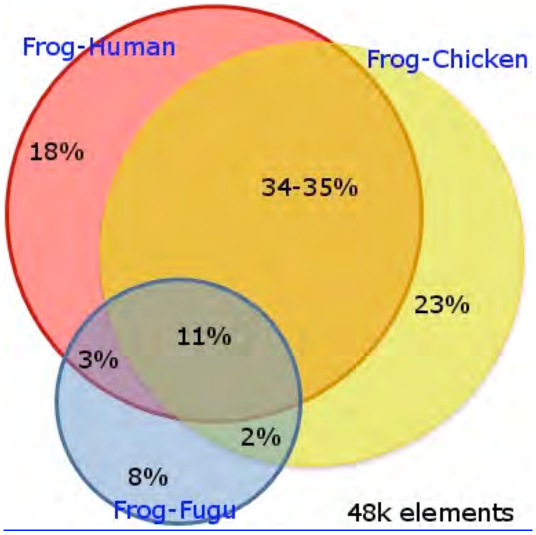
To characterize the biological function of genes associated with noncoding conservation in the frog genome, we annotated genes flanking human-frog CNSs using Gene Ontology analysis(*46*). Based on its maturity, we employed gene annotation corresponding to the human genome. Gene ontology analysis of genes being flanked by CNSs identifies multiple enriched and depleted gene categories (Table S16).

Finally, we analyzed the fraction of human exons that is conserved in different species, including frog. These results are represented in Figure S4. This fraction was computed using exon annotation included in RefSeq(*47*) or UCSC Known Genes(*48*). A given exon was considered conserved if 50% of its sequence overlapped with conserved sequences for the indicated species.

**Genome architecture.** Regarding some architectural features of the frog genome, genes comprise 25%, while regular intergenic regions (defined as having lengths between the 25th and 75th pecentile) range from 5 to 35kb and correspond to 11% of the genome. We also identified approximately 680 gene deserts, defined as the top 3% of the longest intergenic intervals in the frog genome, spanning 231Mb or about 17% of the genome. The remaining 44% of the genome is in intergenic regions of uncategorized lengths. The CNS density tends to be higher in gene deserts. For instance, on average, we found that 83% (569/683) of frog gene deserts contain at least one frog-human CNS with their average being 14 frog-human CNS per CNS-containing gene desert. This is in contrast to intragenic regions, where only 13% (1547/11390) contain at least one frog-human CNS with their average being 2 frog-human CNSs per CNS-containing gene desert. The large number of CNSs within gene deserts further supports the existence of functional elements within these gene-void regions.



A



B

**Figure S3: Conservation among tetrapod genomes. A) Number of coding sequences. B) Number of noncoding sequences.**

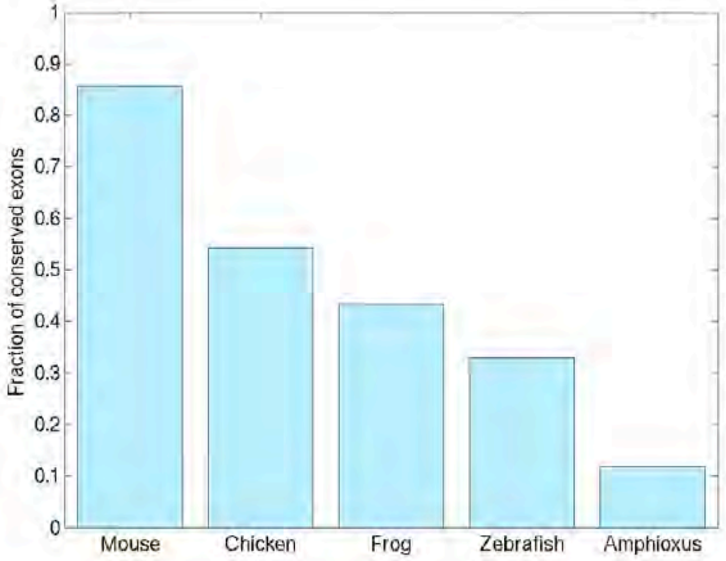


**Figure S4: Fraction of human exons that is conserved in different species, as determined by nucleotide alignments.**

**Table S15: Evolutionary conservation of the frog and the human genomes across sequenced vertebrate species. Number of coding sequences is based on overlap with annotated exons (GenBank EST, GenBank mRNA, exons of mapped RefSeq genes, exons of MGC genes, exons of Genscan gene predictions, and exons of mapped human proteins).**

| | **Coding** | **Noncoding** | **Total** |
|---|---|---|---|
| Frog-Chicken | 91k (72%) | 35k (28%) | 126 (100%) |
| Frog-Human | 95 (75%) | 32 (25%) | 127 (100%) |
| Frog-Zebrafish | 69k (75%) | 23k (25%) | 92k (100%) |
| Frog-Fugu | 63k (84%) | 12k (16%) | 75k (100%) |
| Human-Zebrafish | 78k (85%) | 14k (15%) | 92k (100%) |
| Human-Frog | 100k (70%) | 42k (30%) | 142k (100%) |
| Human-Chicken | 130k (60%) | 87k (40%) | 217k (100%) |
| Human-Mouse | 301k (20%) | 1200k (80%) | 1501k (100%) |

**Table S16: Functional analysis. Table A shows over-represented GO categories, whileTable B lists under-represented GO categories.**

**A**

| GO ID | Description | p-value | Actual Number | Expected Number | Enrichment |
|---|---|---|---|---|---|
| GO:0004984 | Olfactory receptor activity (molecular function) | 3E-94 | 229 | 38 | 5.9 |
| GO:0007608 | sensory perception of smell (biological process) | 2E-57 | 206 | 50 | 4.1 |
| GO:0000786 | nucleosome (cellular component) | 23E-05 | 28 | 7 | 3.9 |
| GO:0006334 | nucleosome assembly (biological process) | 9E-05 | 40 | 14 | 2.8 |
| GO:0050896 | Response to stimulus (biological process) | 24E-35 | 251 | 97 | 2.6 |
| GO:0006511 | ubiquitin-dependent protein catabolic process (biological process) | 1E-05 | 80 | 38 | 2.1 |
| GO:0007186 | G-protein coupled receptor protein signaling pathway (biological process) | 8E-29 | 353 | 177 | 2.0 |
| GO:0003723 | RNA binding (molecular function) | 0.01 | 235 | 172 | 1.4 |
| GO:0003677 | DNA binding (molecular function) | 5E-05 | 463 | 355 | 1.3 |
| GO:0006355 | regulation of transcription, DNA-dependent (biological process) | 4E-4 | 821 | 686 | 1.2 |
| GO:0005634 | nucleus (cellular component) | 4E-11 | 1718 | 1453 | 1.2 |

**B**

| GO_ID | Description | p-value | Actual Number | Expected Number | Depletion |
|---|---|---|---|---|---|
| GO:0007156 | homophilic cell adhesion (biological process) | 7E-26 | 47 | 171 | 0.3 |
| GO:0007218 | neuropeptide signaling pathway (biological process) | 4E-05 | 25 | 66 | 0.4 |
| GO:0007155 | Cell adhesion (biological process) | 5E-36 | 170 | 397 | 0.4 |
| GO:0007169 | transmembrane receptor protein tyrosine kinase signaling pathway (biological process) | 0.01 | 30 | 64 | 0.5 |
| GO:0031225 | anchored to membrane (cellular component) | 0.006 | 42 | 82 | 0.5 |
| GO:0005509 | calcium ion binding (molecular function) | 5E-25 | 302 | 529 | 0.6 |
| GO:0005887 | integral to plasma membrane (cellular component) | 6E-15 | 314 | 490 | 0.6 |
| GO:0030054 | Cell junction (cellular component) | 0.001 | 133 | 201 | 0.7 |
| GO:0006811 | ion transport (biological process) | 3E-4 | 167 | 245 | 0.7 |

| | | | | | |
|---|---|---|---|---|---|
| GO:0016020 | membrane (cellular component) | 8E-30 | 1236 | 1645 | 0.8 |
| GO:0016021 | integral to membrane (cellular component) | 2E-13 | 1246 | 1526 | 0.8 |

**B**

## Evolutionary distances

The molecular (4DTv) distances between vertebrates shown in Fig. S6 represent *the expected number of transversions between two orthologous four-fold degenerate codon sites since the last common ancestor of the species*. To evaluate these distances, we performed all-against all pairwise BLASTp (*34*) of the Human proteome to each of mouse, chicken, frog, zebrafish, and fugu, using Ensembl models. We next combined reciprocal highest scoring hits as candidate orthologs. In 4,549 cases, a human gene had a putative ortholog in all five other species, defining a candidate cluster of orthologs. These clusters were aligned using clustalW(*49*), and gap-free blocks of at least 20 amino residues flanked by fully conserved amino residues (using in-house scripts) were extracted and concatenated. Next, columns of fully conserved four-fold degenerate amino acids (P, T, V, A or G) were selected, and for each species it was noted whether a pyrine (A or G) or a pyrimidine (C or T) nucleotide was present at the 3rd codon position. This was encoded as 0 and 1 and concatenated into six “state strings” each of length 207,604 representing the pyrimidine-purine states at all conserved 4D sites. From these characters, a Bayesian phylogenetic tree was created using mrBayes(*50*) with a binary model (Fig. S5). From this tree we infer the following 4DTv distances: Human-Mouse: 0.18, Human-Chicken: 0.57, Human-Frog:  0.91 and Human-Zebrafish: 1.22. These distances are used in Fig. S6. Comparison beyond human-fish does not yield reliable results due to the large degree of saturation (4DTv >> 1).

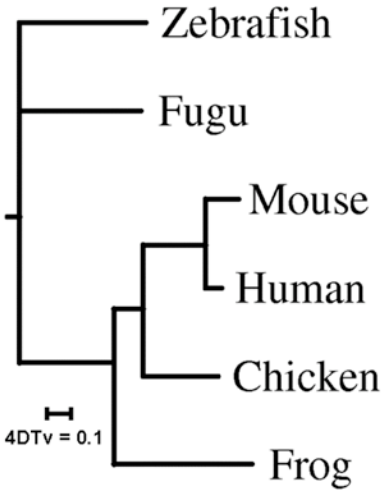


**Fig S5: Phylogenetic tree based on purine/pyrimidine content at the 3rd codon position in more than 200,000 fully conserved amino acids. The scale bar indicates 0.1 transversion per site.**

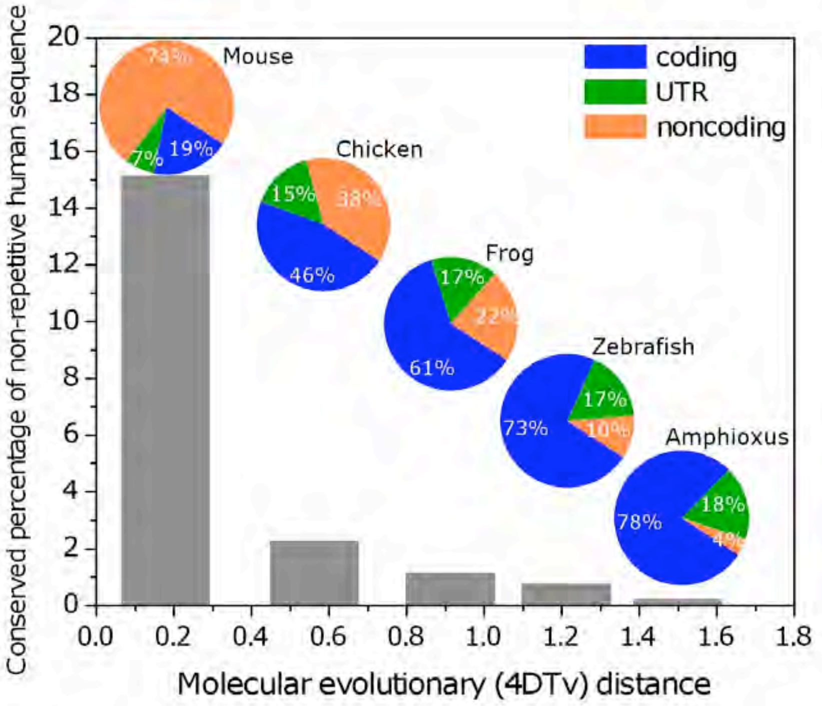


**Fig. S6: Percentage of conserved non-repetitive human sequence annotated as coding, noncoding or untranslated regions (UTR), shown for alignments with the sequences of each indicated species, as a function of the molecular evolutionary (4DTv) distance. The 4DTv distance is the expected number of transversions to have occurred at a four-fold degenerate codon site in a conserved amino acid since the divergence of the species. The 4DTv distance between human and Amphioxus is highly saturated, and the value of 1.5 represents a rough estimate.**

## Function of conserved non-coding regions

As stated in the main text, the dominant category of conserved non-coding sequence (CNS) is shared by tetrapods (38%) to the exclusion of fish, with only half as many (19%) also shared with fish. Many such CNSs are cis-regulatory elements(*51*), and the lower degree of conservation between tetrapods and fish may reflect their unique subfunctionalization and neofunctionalization that occurred after the whole genome duplication in the teleost lineage(*52*). Thus the comparison of mammals to amphibian or avian genomes may be most productive for predicting enhancer function.

An example of the utility of the *X. tropicalis* genome in identifying highly conserved regulatory elements is shown for the anterior neural homeobox gene *six3* (*53*) in Fig. S7. The alignment of mouse and human genomes shows peaks of conservation embedded in the high degree of overall similarity, but it is more difficult to pick out clear peaks of conservation than in the human to frog comparison. When mammalian and fish sequences are compared, the divergence causes many of these peaks to become indistinct. However, the real test of the predictive power is to test these elements in transgenic embryos.

In Fig. S7 a cloning-free co-transgenesis assay was used to identify enhancers responsible for *six3* eye- and forebrain-specific expression (*54*). The mammalian-*Fugu* genome comparison cannot identify one of these conserved enhancers, CNS5, showing that the mammalian-*Xenopus* comparison may be more suitable than the mammalian-fish comparison for identifying cis-regulatory elements with conserved functions.

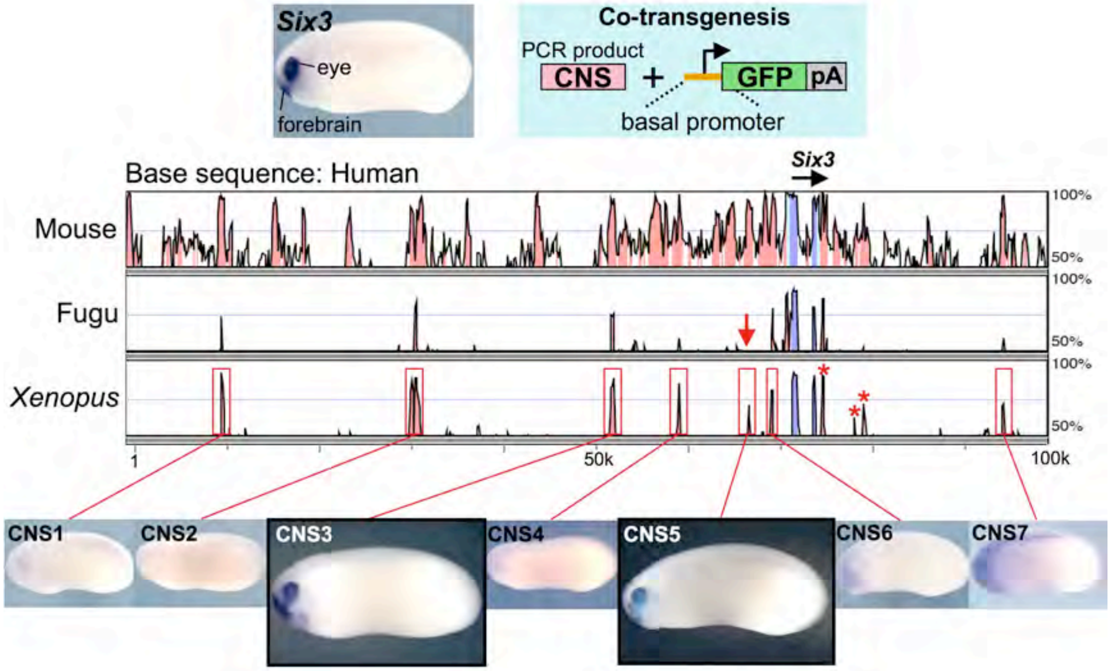


**Figure S7: Combination of the mammalian-*Xenopus* genome comparison and the co-transgenesis reporter assay identifies eye- and forebrain-specific enhancers of a homeobox gene, *Six3*. A 100 kb genomic sequence of the human *Six3* locus (hg18, chr2: 44950620-45050620) is aligned with its orthologous sequences of mouse (mm8, chr17: 85461507-85596280), Fugu (FUGU4, scaffold 124: 457398-501325) and *Xenopus* (Xentr4, scaffold 25: 1427798-1552971) using mVISTA(*41*), middle panel. Peaks shaded with light blue, light cyan and vermillion represent conserved regions in coding exons, untranslated exons and intergenic regions, respectively. The scale at the bottom of the alignment indicates relative positions in the human *Six3* locus. Red boxes indicate the Human-*Xenopus* CNSs (CNS1 - CNS7) that were amplified from *Xenopus* genomic DNA by PCR, and co-injected into *Xenopus* eggs along with a *β-actin* basal promoter-GFP cassette(*54*) (right upper panel)GFP expression in the resulting embryos was analyzed at early tailbud stages by in situ hybridization for maximum sensitivity (bottom panel). Among the seven CNSs tested, only CNS3 and CNS5 drove GFP expression in the eye and part of the ventral forebrain, which mostly recapitulated the expression of endogenous *Six3* (left upper panel). The red arrow in the Human-Fugu alignment**

**indicates the position of CNS5 that is conserved between human and *Xenopus* but not clearly between human and Fugu. The sequences indicated with asterisks in the Human-*Xenopus* alignment were not subjected to the enhancer assay, because they appear to be repetitive sequences rather than CNSs. Although mVISTA, used in this particular example for comparative analysis, is a valuable comparative tool, PipMaker(*55*) often reveals more subtle conservation of CNSs when comparing mammalian and *X. tropicalis* sequence because of its ability to detect small inversions and deletions(*54*).**

## Supplementary Note 8. Developmental genes in frog

The initial patterning of the amphibian embryo follows from the animal-vegetal polarity of the egg, and as outlined in the main text, *VegT* is the essential vegetally localized component that establishes the equatorial mesoderm, and vegetal endoderm. One of the main activities of *VegT* is to activate the numerous copies of *nodal*, and the duplication of *nodal* genes in the amphibian lineage illustrates how this gene has evolved in the vertebrate lineage from an ancient nodal gene that was present before the duplication of the protostomes and deuterostomes(*56*).

*Xenopus* possesses a remarkable multiplicity of *nodal* genes. While the mammals manage with a single *nodal* gene, the frog has expanded this family to include six *nodal* relatives (*Xnr*s) that were first characterized as cDNAs in *Xenopus laevis* (*57-60*). Different cDNAs also illustrate that the *Xnr3* and *Xnr5* genes must have duplicated further, and this has been verified through examination of the *Xnr5* locus in *Xenopus laevis*(*61*). All but *Xnr3* signal through the *Smad2* pathway, while *Xnr3* has diverged to function in a *Smad2*-independent pathway and act as a secreted inhibitor of BMP signaling(*62*).

Synteny relationships reveal the evolutionary history and dynamic chromosome rearrangements of two nodal loci (Figs. S8, S9). The *X. tropicalis* version of *Xnr4* shares synteny, both upstream and downstream, with the single nodal locus in mammals and therefore, because of this clear orthology, is termed *nodal.* The synteny is shared with one of the three *nodal*s in teleost genomes and hence this gene may reflect one ancestral configuration, as has been recognized previously(*63*). Interestingly, this *nodal* appears to have been cleanly deleted from the syntenic location in the chicken genome, between the *eif4ebp2* and *paladin* (*pald*) genes. Instead, the chicken *nodal* shows synteny to a separate cluster of *X. tropicalis nodal* genes. This chicken *nodal* gene lies between another copy of *eif4ebp* (orthologous to the mammalian *eif4bp1*) and the *ash2l* gene. The mammals have a syntenic stretch of genes, but lack *nodal* in this location. Thus, just as with the potential deletion of chicken *nodal* from between the *eif4ebp2* and *pald* genes the mammalian lineage may have deleted an ancestral *nodal* from between the *eif4ebp1* and *ash2l* genes. The presence of two evolutionarily distinct *nodal* loci in these amniote lineages prompted us to search a tetrapod outgroup to birds and mammals, in addition *to X. Tropicalis*, that may have retained both *nodal* loci. Indeed, inspection of the *Anolis carolinensis* (reptile) genome assembly shows *nodal* genes to have been retained in both syntenic locations.

In *X. tropicalis*, this locus adjacent to *ash2l, star*, and *lsm1* genes has a complex of at least 9 *nodal* genes. This contains the *nodal 1*, *2*, *3*, *5* and *6* genes (similar to *X. Laevis Xnr*s *1, 2,3 ,5* and *6*). The complex of these *nodal* genes is poorly represented in the whole genome assembly, since a separate analysis of two BAC sequences from independent libraries shows that the *nodal5* gene is present as five copies, consistent with the multitude of genes and cDNA isoforms found in *X. laevis*. In addition, *nodal3* is probably present as three copies(*64*). Yet other *nodal* genes are found on small scaffolds, but these may also represent the assembly of residual reads after misassembly of scaffold 34, since they are nearly identical to the *nodal 3* and *5* genes on this scaffold. We conclude from this analysis that at least two *nodal* loci were retained in vertebrates following the last genome duplication, and subsequently were retained and amplified, as in *X,* tropicalis, or were lost independently in the chicken and mammalian lineages. *Nodal* not only fulfills an essential role in mesoderm development, but also is the conserved mediator of left-right asymmetry, through its expression in the left lateral plate mesoderm. Both *nodal* loci show evidence of ancestral left-sided expression, since *nodal1* in *Xenopus* and the chick *nodal*, adjacent to *ash2l*, are left-sided, while the mammalian *nodal*, adjacent to *paladin*, is also left-sided. Most other Xenopus *nodal*s have lost this expression including *X.laevis Xnr4*, the locus orthologous to mammalian *nodal* (*65-66*).

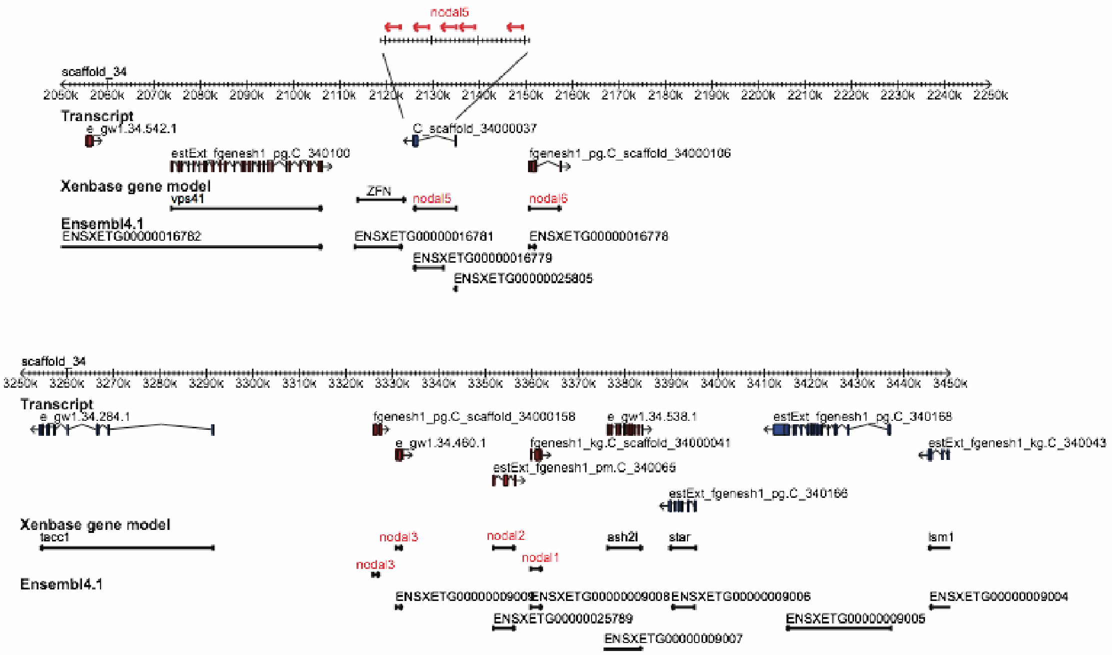

nodal5
scaffold_34
2050k 2060k 2070k 2080k 2090k 2100k 2110k 2120k 2130k 2140k 2150k 2160k 2170k 2180k 2190k 2200k 2210k 2220k 2230k 2240k 2250k
Transcript
e_gw1.34.542.1
C_scaffold_34000037
estExt_fgenesh1_pg.C_340100
fgenesh1_pg.C_scaffold_34000106
Xenbase gene model
vps41
ZFN
nodal5
nodal6
Ensembl4.1
ENSXETG00000016782
ENSXETG00000016781
ENSXETG00000016778
ENSXETG00000016779
ENSXETG00000025805
scaffold_34
3250k 3260k 3270k 3280k 3290k 3300k 3310k 3320k 3330k 3340k 3350k 3360k 3370k 3380k 3390k 3400k 3410k 3420k 3430k 3440k 3450k
Transcript
e_gw1.34.284.1
fgenesh1_pg.C_scaffold_34000158
e_gw1.34.538.1
estExt_fgenesh1_pg.C_340168
e_gw1.34.460.1
fgenesh1_kg.C_scaffold_34000041
estExt_fgenesh1_kg.C_340043
estExt_fgenesh1_pm.C_340065
estExt_fgenesh1_pg.C_340166
Xenbase gene model
tacc1
nodal3
nodal2
ash2l
star
ism1
nodal1
nodal3
Ensembl4.1
ENSXETG00000009009
ENSXETG00000009008
ENSXETG00000009006
ENSXETG00000009004
ENSXETG00000025789
ENSXETG00000009005
ENSXETG00000009007

**Figure S8: Structure of the second *nodal* locus in Xenopus tropicalis on scaffold 34. The *nodals* are present in two ~50kb segments on scaffold 34, from 2120K to 2160K, and 3,320K to 3,370K. However, due to misassembly of the reiterated *nodal5* gene, a 30kb segment should be inserted to replace the existing *nodal5* gene. Until these regions are resequenced and assembled, the possibility that additional expansions of the nodal locus have been concealed should also be entertained, and indeed there is evidence for three copies of *nodal3* from cDNA sequences(*64*).**

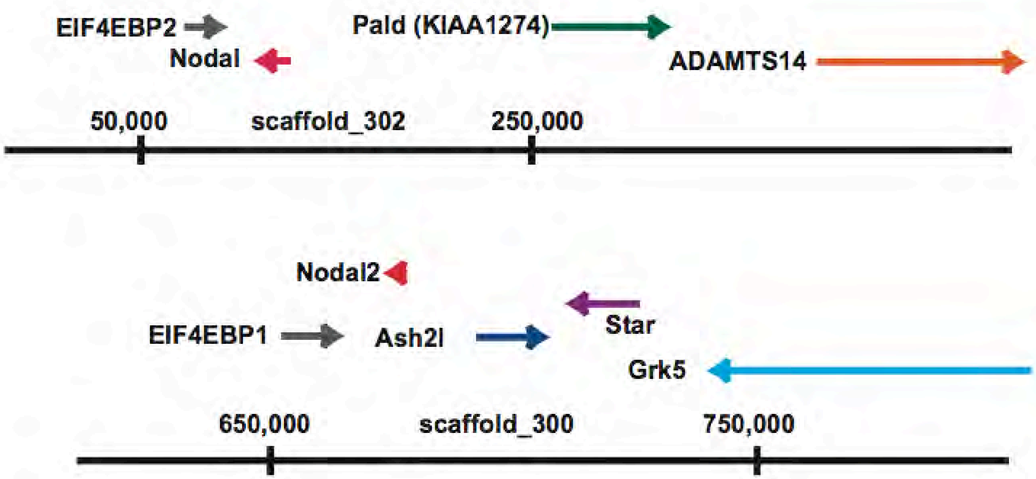


**Figure S9: The Lizard *Anolis carolinensis* has retained the ancestral configuration of nodal genes. We searched in the lizard genome for nodal genes, both by blast similarity, and by searching the neighborhood of the two *eif4ebp* genes and the neighboring *ash2l,* and *pald* genes. The figure summarizes the arrangement of the two loci, where nodal has been retained in A. carolinensis. In contrast, the chicken has deleted the nodal gene adjacent to *pald,* while the mammals have deleted the gene adjacent to *ash2l***

The *nodal*s are expressed in overlapping domains, so it is not clear why there has been selection for multiple copies, but the theme of duplication of early signaling activities is also displayed by a number of transcription factors that operate early in patterning the embryo. Many organizer signaling components were first identified in *Xenopus laevis* by their function or expression, and their homologues are present in *Xenopus tropicalis*, often in multiple copies. This includes the early organizer transcripts from *siamois*, a paired homeodomain transcriptional activator, whose genes are triplicated locally in the genome. This gene is unusual in appearing to be frog specific, with no similar gene found in fish or amniotes, nor in the other deuterostome lineages (*Ciona*, *Branchiostoma*, *Saccoglossus*, nor *Strongylocentrotus*). Other overlapping and duplicated activities present in Spemann's organizer, are found in other vertebrates, as expected from the functional similarity to the embryonic shield of the fish and Hensen's node of the amniotes. In the case of the BMP antagonists *chordin*, *follistatin*, *noggin*, *nodal-related3* and *cerberus*, these are multiple activities that share function but no obvious primary structure. Among these all are encoded by unique genes with the exception of the triplicated *nodal3*. Likewise, the *wnt* antagonists, *dkk1*, *frzb*, *crescent* encode distinct proteins which nonetheless cooperate to mediate head formation; these are all present in single copies

Other transcription factors that set up the mesoderm and endoderm are present in multiple copies, with *ventx* relatives (ventrally expressed paired- family homeodomain activities) locally duplicated to six copies. *Bix* genes (encoding paired family homeodomains), *brachyury* (early mesodermally expressed T-box genes), and *sox17* (SRY-box transcription factors) are also found in multiple copies. Interestingly, there is also a local triplication of the gene encoding the stem cell pluripotency facter *pou5f1* (*67*) (also known in mammals as *oct3/4*). The tropicalis genes are linked to the nearby gene *fut7* as are fish *pou5f1* genes, and *fut7* is found near a *pou5f1* gene on *Monodelphis* chromosome

1, implying that the anamniote *pou5f1* genes are indeed orthologs to mammalian Oct4s. Interestingly, another stem cell pluripotency gene, *nanog*, while present in amniotes, appears lacking in the genomes of frog and teleost fishes. The amplification of these genes or the multiple unique genes encoding overlapping functions may represent the selection for a robust and rapid induction of pattern formation, which imposes a considerable transcriptional burden on the embryo.

As expected from general conservation of signaling, most components and genes can be found in the *Xenopus* genome; most are readily identified through their syntenic relationship to the genes of mammals, and are present as unique copies.

## Supplementary Note 9. Immune system genes in frog

The immune system of the frog is similar to that of other vertebrates, with components of the adaptive and innate immune system. In particular, it suggests a co-evolution of gene members of both the adaptive and innate immune system, though before genomic information was available, several components had not been identified unambiguously. As might be expected from the different ecological niche occupied by Xenopus, many cell surface receptor families have expanded, including Non-classical class *Ib*, *Fc* receptor-like, *CD2*, *NKp30* gene families. One can speculate that this may be related to the need for a functional immune system very early in ontogeny and metamorphosis. Proof that the Xenopus tropicalis genome provides a useful intermediate between amniotes and fish comes from the analysis of the immunoglobulin isotypes. The IgW shark/lungfish immunoglobulin was thought to have been lost subsequently in evolution. However, the frog sequence shows a related *IgD* isotype that makes a connection between the fish and amniotes IgD.

However, the immune responses mounted by frogs are somewhat attenuated compared to that of mammals; thus, repeated immunization results in lower titer and affinity of antibodies, T cells expand their population lesser than those of mammals in vitro and in vivo, and Lipopolysaccharide only elicits poor inflammatory responses. One hypothesis for the difference might be a less expanded set of immune regulators. A difficulty in addressing this question is that many immune regulators are not easily recognized in sequence similarity searches. However, the high level of conserved synteny between frog and mammalian genomes has enabled the unambiguous identification of a number of regulators, such as *CD8 beta*, whose proximity to *CD8 alpha* is conserved, and *CD4*. Subsequent work has confirmed that these markers identify CD8 positive T cells (including *Natural Killer* and *CD8 alpha* expressing cells) as well as *CD4* positive T cells (likely T- helper cells). Similarly, an Interleukin2/21 like sequence was identified in a syntenic region between the *tenr* and *centrin4* genes, though no EST support is yet available for the expression.

In mammals, the high affinity *LPS* receptor system (*TLR-4* + *CD14*) plays an important role in activation/maturation of antigen presenting cells (e.g., dendritic cells) that up-regulate co-stimulatory molecules and release cytokines needed for an optimal and robust T cell activation(*68*). Interestingly, while all the human *TLR* orthologs are present and well conserved in *Xenopus*, *Xenopus TLR-4* seems to be divergent(*69-70*) and *CD14* appears to be missing. This might account for the poor responses of *Xenopus* to LPS exposure.

In summary, information on genes involved in immunity so far extracted for *the X. tropicalis* genome reveal a remarkable overall conservation with mammals with further specialization (expansion or contraction of certain gene families). Future data-mining of the *Xenopus* genome will provide important insights into the evolution and development of the vertebrate immune system.

**Skin peptides**

*Xenopus* has become well known for its production of antimicrobial peptides in skin secretions. Remarkably, similar peptides have not been found in birds, reptiles and mammals. Both antimicrobial peptides (caerulein, levitide, magainin, PGLa/PYLa, PGQ, xenopsin), neuromuscular toxins (e.g. xenoxins) and neuropeptides (e.g. thyrotropin releasing hormone, TRH (*71*) are secreted by granular glands and constitute an important defense against pathogens(*72*)). The sequence and activity of amphibian antimicrobial peptides is well described (see (*73*) for a review) so that the corresponding genes can be annotated. Interestingly, antimicrobial peptides are clustered in at least seven transcription units encoding the antimicrobial peptides spread over 350 kbp on scaffold_811, with no intervening genes. Each transcription unit is composed of four to five exons as deduced from EST alignments, and expressed from the onset of metamorphosis, in skin, bone and thymus. The recent expansion of the gene set is also suggested by their amino-terminal portion which is highly similar to cholecystokinin and gastrin neuropeptides, encoded by a single exon. However, the genes encoding cholecystokinin (scaffold_1166), xenoxin (scaffold_521) and TRH (scaffold_353) are located on separate scaffolds, and in syntenic regions compared to mammalian genomes. Thus the structure of the antimicrobial peptide locus reflects the combinatorial rearrangements of exons that occurred during evolution. It is tempting to speculate that splicing events may produce different mRNA encoding different preproproteins; these might encode a variety of peptides that mediate aspects of innate immunity.

**Table S17: antimicrobial genes**

| CDNA | (Xentr4) |
|---|---|
| Pgla_xentr | scaffold_811:21284-21414,24320-24350,25963-26022,27347-27509 |
| Levitide_xentr | scaffold_811:234823-234876,238354-238517,242742-242789,244266-244346,249729 |
| magainin_xentr | scaffold_811:272044-272092,275080-275231,276929-276991,278721-278779 |
| caerulein_xentr | scaffold_811:300765-300847,302701-302849,304911-304964,306529-306698 |
| caerulein2_xentr | scaffold_811:343808-343954,345109-345194,347389-347537,348657-348694 |
| prepropgq_xentr | scaffold_811:47814-47950,50296-50387,52003-52071,53604-53761,54914-54957 |
| prepropgq2_xentr | scaffold_811:68508-68616,71069-71153,71843-71911,75139-75296,81538-81572 |
| cholecystokinin | scaffold_1166:19699-35312 |
| Xenoxin | scaffold_521:624513-626954 |
| TRH | scaffold_353:194552-202299 |

# References

1. S. Aparicio *et al.*, *Science* 297, 1301 (Aug 23, 2002).

2. T. D. Wu, C. K. Watanabe, *Bioinformatics* 21, 1859 (May 1, 2005).

3. B. J. Haas *et al.*, *Nucleic Acids Res* 31, 5654 (Oct 1, 2003).

4. D. L. Wheeler *et al.*, *Nucleic Acids Res* 31, 28 (Jan 1, 2003).

5. J. Voigt, J. A. Chen, M. Gilchrist, E. Amaya, N. Papalopulu, *Mech Dev* 122, 289 (Mar, 2005).

6. J. A. Chen, J. Voigt, M. Gilchrist, N. Papalopulu, E. Amaya, *Mech Dev* 122, 307 (Mar, 2005).

7. M. J. Gilchrist *et al.*, *Dev Biol* 271, 498 (Jul 15, 2004).

8. S. L. Klein *et al.*, *Dev Dyn* 225, 384 (Dec, 2002).

9. L. W. Hillier *et al.*, *Nature* 432, 695 (Dec 9, 2004).

10. E. S. Lander *et al.*, *Nature* 409, 860 (Feb 15, 2001).

11. R. H. Waterston *et al.*, *Nature* 420, 520 (Dec 5, 2002).

12. V. V. Kapitonov, J. Jurka, *Genetica* 107, 27 (1999).

13. I. R. G. S. Project, *Nature* 436, 793 (2005/08/11/print, 2005).

14. N. L. Craig, R. Craigie, M. Gellert, A. M. Lambowitz, Eds., *Mobile DNA II*, (ASM Press, Washington, DC, 2002).

15. V. V. Kapitonov, J. Jurka, *Proc Natl Acad Sci U S A* 98, 8714 (Jul 17, 2001).

16. V. V. Kapitonov, J. Jurka, *Proc Natl Acad Sci U S A* 103, 4540 (Mar 21, 2006).

17. V. V. Kapitonov, J. Jurka, *DNA Cell Biol* 23, 311 (May, 2004).

18. R. T. Poulter, T. J. Goodwin, M. I. Butler, *Gene* 313, 201 (Aug 14, 2003).

19. R. T. Poulter, T. J. Goodwin, *Cytogenet Genome Res* 110, 575 (2005).

20. M. B. Evgen'ev, I. R. Arkhipova, *Cytogenet Genome Res* 110, 510 (2005).

**21. D. E. Dalle Nogare, M. S. Clark, G. Elgar, I. G. Frame, R. T. Poulter, *Mol Biol Evol* 19, 247 (Mar, 2002).**

**22. V. V. Kapitonov, J. Jurka, *Proc Natl Acad Sci U S A* 100, 6569 (May 27, 2003).**

**23. O. Kohany, A. J. Gentles, L. Hankus, J. Jurka, *BMC Bioinformatics* 7, 474 (2006).**

**24. J. Jurka *et al.*, *Cytogenet Genome Res* 110, 462 (2005).**

**25. V. V. Kapitonov, J. Jurka, *Nat Rev Genet* 9, 411 (May, 2008).**

**26. A. H. Smit, R; Green, P. (1996-2004).**

**27. R. F. Yeh, L. P. Lim, C. B. Burge, *Genome Res* 11, 803 (May, 2001).**

**28. A. A. Salamov, V. V. Solovyev, *Genome Res* 10, 516 (Apr, 2000).**

**29. J. Amberger, C. A. Bocchini, A. F. Scott, A. Hamosh, *Nucleic Acids Res* 37, D793 (Jan, 2009).**

**30. V. A. McKusick, *Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders*. (Johns Hopkins University Press, Baltimore, ed. 12, 1998).**

**31. K. I. Goh *et al.*, *Proc Natl Acad Sci U S A* 104, 8685 (May 22, 2007).**

**32. U. Hellsten *et al.*, *BMC Biol* 5, 31 (2007).**

**33. A. Force *et al.*, *Genetics* 151, 1531 (Apr, 1999).**

**34. S. F. Altschul *et al.*, *Nucleic Acids Res* 25, 3389 (Sep 1, 1997).**

**35. G. Blanc, K. Hokamp, K. H. Wolfe, *Genome Res* 13, 137 (Feb, 2003).**

**36. N. H. Putnam *et al.*, *Science* 317, 86 (Jul 6, 2007).**

**37. M. K. Khokha *et al.*, *Dev Dyn* 238, 1398 (Jun, 2009).**

**38. H. Li, R. Durbin, *Bioinformatics* 25, 1754 (Jul 15, 2009).**

**39. SOD, (2010).**

**40. A. Abu-Daya, A. K. Sater, D. E. Wells, T. J. Mohun, L. B. Zimmerman, *Dev Biol* 336, 20 (Dec 1, 2009).**

**41. K. A. Frazer, L. Pachter, A. Poliakov, E. M. Rubin, I. Dubchak, *Nucleic Acids Res* 32, W273 (Jul 1, 2004).**

42. **M. Brudno *et al.*, *Bioinformatics* 19 Suppl 1, i54 (2003).**

43. **M. B. Sundararajan, M. et. al., in *WABI 2004, 4th Workshop on Algorithms in Bioinformatics*. (Bergen, Norway, 2004).**

44. **W. J. Kent, *Genome Res* 12, 656 (Apr, 2002).**

45. **I. Dubchak, A. Poliakov, A. Kislyuk, M. Brudno, *Genome Res* 19, 682 (Apr, 2009).**

46. **L. Taher, I. Ovcharenko, *Bioinformatics* 25, 578 (Mar 1, 2009).**

47. **K. D. Pruitt, D. R. Maglott, *Nucleic Acids Res* 29, 137 (Jan 1, 2001).**

48. **F. Hsu *et al.*, *Bioinformatics* 22, 1036 (May 1, 2006).**

49. **J. D. Thompson, T. J. Gibson, D. G. Higgins, *Curr Protoc Bioinformatics* Chapter 2, Unit 2 3 (Aug, 2002).**

50. **F. Ronquist, J. P. Huelsenbeck, *Bioinformatics* 19, 1572 (Aug 12, 2003).**

51. **A. Visel, J. Bristow, L. A. Pennacchio, *Semin Cell Dev Biol* 18, 140 (Feb, 2007).**

52. **D. Kurokawa *et al.*, *Proc Natl Acad Sci U S A* 103, 19350 (Dec 19, 2006).**

53. **X. Zhou, T. Hollemann, T. Pieler, P. Gruss, *Mech Dev* 91, 327 (Mar 1, 2000).**

54. **H. Ogino, M. Fisher, R. M. Grainger, *Development* 135, 249 (Jan, 2008).**

55. **S. Schwartz *et al.*, *Genome Res* 10, 577 (Apr, 2000).**

56. **C. Grande, N. H. Patel, *Nature* 457, 1007 (Feb 19, 2009).**

57. **C. M. Jones, M. R. Kuehn, B. L. Hogan, J. C. Smith, C. V. Wright, *Development* 121, 3651 (Nov, 1995).**

58. **W. C. Smith, R. McKendry, S. Ribisi, Jr., R. M. Harland, *Cell* 82, 37 (Jul 14, 1995).**

59. **N. Qureshi, R. E. Dugan, W. W. *Cleland*, J. W. Porter, *Biochemistry* 15, 4191 (Sep 21, 1976).**

60. **S. Takahashi *et al.*, *Development* 127, 5319 (Dec, 2000).**

61. **S. Takahashi *et al.*, *Genesis* 44, 309 (Jul, 2006).**

62. **C. S. Hansen, C. D. Marion, K. Steele, S. George, W. C. Smith, *Development* 124, 483 (Jan, 1997).**

63. X. Fan, S. T. Dougan, *Dev Genes Evol* 217, 807 (Dec, 2007).

64. Y. Haramoto *et al.*, *Dev Biol* 265, 155 (Jan 1, 2004).

65. L. A. Lowe *et al.*, *Nature* 381, 158 (May 9, 1996).

66. E. M. Joseph, D. A. Melton, *Dev Biol* 184, 367 (Apr 15, 1997).

67. S. Frankenberg, A. Pask, M. B. Renfree, *Dev Biol* 337, 162 (Jan 1, 2010).

68. R. Medzhitov, P. Preston-Hurlburt, C. A. Janeway, Jr., *Nature* 388, 394 (Jul 24, 1997).

69. A. Ishii, M. Kawasaki, M. Matsumoto, S. Tochinai, T. Seya, *Immunogenetics* 59, 281 (Apr, 2007).

70. J. C. Roach *et al.*, *Proc Natl Acad Sci U S A* 102, 9577 (Jul 5, 2005).

71. G. Kreil, *Skin Secretions of Xenopus Laevis*. H. R. Tinsley, Ed., The Biology of Xenopus (The Zoological Society of London, Oxford, 1996).

72. L. A. R. Rollins-Smith, L.K.; Houston C.J.L.E.; Woodhams, D.C., *Antimicrobial peptide defenses in amphibian skin* Integrative and Comparative Biology (2005), vol. 45.

73. A. C. Rinaldi, *Curr Opin Chem Biol* 6, 799 (Dec, 2002).