tion (*21*) and protonation state (*22*) of TyrL162 may be important for the mechanism of electron transfer, and the transient creation of a negatively charged tyrosyl group between heme c_{559} and P$^+$ could stabilize the oxidized form of the cytochrome subunit (*12*).

From the temperature dependence of the P$^+$:Q$_A^-$ → P:Q$_A$ charge recombination reaction, it was proposed that structural changes occurring in response to electron transfer decrease the free energy gap between P$^+$ and Q$_A^-$ of RC$_{sph}$ by about 12 kJ mol^{-1} (*23*). Similar measurements of the pH dependence of the RC$_{vir}$ charge recombination reaction showed the P$^+$:Q$_A^-$ state to be stabilized by a chemical group with an approximate pK_a of 9 (*24*). Prolonged exposure to bright light is also known to reversibly stabilize the P$^+$:Q$_A^-$ state of RC$_{sph}$, slowing the charge recombination reaction by up to three orders of magnitude (*3*, *25*, *26*). Light-induced deprotonation and conformational switching of TyrL162 could contribute to all three stabilization effects, because the creation of a phenolate anion in immediate proximity to the special pair effectively neutralizes the energetic penalty associated with the buried positive charge on P$^+$.

Electron paramagnetic resonance spectroscopy studies of RC$_{sph}$ mutants that increase the midpoint potential of the special pair have shown that TyrL162 can form a (deprotonated) tyrosyl radical in response to charge separation (*27*), confirming that TyrL162 can be deprotonated by photooxidation of the special pair. Photooxidation of photosystem II also creates a tyrosyl radical (Tyr$_Z^{\bullet}$) (*28*, *29*), which oxidizes the manganese cluster and leads to the synthesis of molecular oxygen from water. Crystal structures of photosystem II (*30*) reveal that Tyr$_Z$ lies between the oxygen-evolving center and the special pair in a position functionally analogous to that occupied by TyrL162 of RC$_{vir}$ (Fig. 4). Our findings suggest that the deprotonation and coupled conformational switching of TyrL162 may have aided the spontaneous formation of tyrosine radicals in an ancient reaction center, thereby creating the chemical potential to extract electrons from clusters of manganese atoms (*31*) and ultimately oxidize water to oxygen.

References and Notes

1. J. Deisenhofer, O. Epp, K. Miki, R. Huber, H. Michel, *Nature* **318**, 618 (1985).
2. M. H. Stowell *et al.*, *Science* **276**, 812 (1997).
3. G. Katona *et al.*, *Nat. Struct. Mol. Biol.* **12**, 630 (2005).
4. V. Srajer *et al.*, *Science* **274**, 1726 (1996).
5. F. Schotte *et al.*, *Science* **300**, 1944 (2003).
6. U. K. Genick *et al.*, *Science* **275**, 1471 (1997).
7. B. Perman *et al.*, *Science* **279**, 1946 (1998).
8. R. H. Baxter *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 5982 (2004).
9. Materials and methods are available as supporting material on *Science* Online.
10. A. B. Wöhri *et al.*, *Biochemistry* **48**, 9831 (2009).
11. B. Dohse *et al.*, *Biochemistry* **34**, 11335 (1995).
12. J. M. Ortega, B. Dohse, D. Oesterhelt, P. Mathis, *Biophys. J.* **74**, 1135 (1998).
13. M. Leonhard, W. Mäntele, *Biochemistry* **32**, 4532 (1993).
14. S. Buchanan, H. Michel, K. Gerwert, *Biochemistry* **31**, 1314 (1992).
15. R. L. Thurlkill, G. R. Grimsley, J. M. Scholtz, C. N. Pace, *Protein Sci.* **15**, 1214 (2006).
16. N. S. Lewis, D. G. Nocera, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 15729 (2006).
17. L. Hammarström, S. Styring, *Philos. Trans. R. Soc. London Ser. B* **363**, 1283 (2008).
18. J. Wachtveitl, J. W. Farchaus, P. Mathis, D. Oesterhelt, *Biochemistry* **32**, 10894 (1993).
19. X. M. Gong, M. L. Paddock, M. Y. Okamura, *Biochemistry* **42**, 14492 (2003).
20. Y. Ohtsuka, K. Ohkawa, H. Nakatsuji, *J. Comput. Chem.* **22**, 521 (2001).
21. B. Cartling, *J. Chem. Phys.* **95**, 317 (1991).
22. B. Cartling, *Chem. Phys. Lett.* **196**, 128 (1992).
23. B. H. McMahon, J. D. Müller, C. A. Wraight, G. U. Nienhaus, *Biophys. J.* **74**, 2567 (1998).
24. P. Sebban, C. A. Wraight, *Biochim. Biophys. Acta* **974**, 54 (1989).
25. F. van Mourik, M. Reus, A. R. Holzwarth, *Biochim. Biophys. Acta* **1504**, 311 (2001).
26. U. Andréasson, L. E. Andréasson, *Photosynth. Res.* **75**, 223 (2003).
27. A. J. Narváez, R. LoBrutto, J. P. Allen, J. C. Williams, *Biochemistry* **43**, 14379 (2004).
28. B. A. Barry, G. T. Babcock, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 7099 (1987).
29. R. J. Debus, B. A. Barry, I. Sithole, G. T. Babcock, L. McIntosh, *Biochemistry* **27**, 9071 (1988).
30. K. N. Ferreira, T. M. Iverson, K. Maghlaoui, J. Barber, S. Iwata, *Science* **303**, 1831 (2004); published online 5 February 2004 (10.1126/science.1093087).
31. J. F. Allen, W. Martin, *Nature* **445**, 610 (2007).
32. Refined crystallographic coordinates and structure factor amplitudes are deposited within the Protein Data Bank (PDB) with codes 2x5u (dark state) and 2x5v (photoactivated state). We thank L. Hammarström for valuable discussions and acknowledge financial support from the Swedish Science Research Council (Vetenskapsrådet), European Commission [Research Training Network (RTN) "Fast light-actuated structural changes" (FLASH), and Integrated Project, the European Membrane Protein Consortium (EMEP)], European Molecular Biology Organization (EMBO), Human Frontier Science Program (HFSP), and the University of Gothenburg Quantitative Biology Platform.

Supporting Online Material

www.sciencemag.org/cgi/content/full/328/5978/630/DC1
Materials and Methods
Figs. S1 to S10
Tables S1 to S3
Scheme S1
References

21 December 2009; accepted 17 March 2010
10.1126/science.1186159

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] Daniel S. Rokhsar[1,2]

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 more than 20,000 protein-coding genes, including orthologs of at least 1700 human disease genes. Over 1 million expressed sequence tags validated the annotation. More than one-third of the genome consists of transposable elements, with unusually prevalent DNA transposons. Like that of other tetrapods, the genome of *X. tropicalis* contains gene deserts enriched for conserved noncoding elements. The genome exhibits substantial 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 20 species of frogs native to Sub-Saharan Africa. The species *Xenopus laevis* was first introduced to the United States in the 1940s 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 to 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 an 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 to S3 and accession number 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 2 million *Xenopus* expressed sequence tags (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*).

More than one-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 to 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 to 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 to 5 million years ago) are more common in frogs than in mammals or birds, and their prevalence is comparable to that in fish, insects, nematodes, and plants. Among these is an unusually high diversity of very young families of L1 non-LTR (long terminal repeat) retrotransposons, Penelope, and DIRS retrotransposons. In contrast to those of other vertebrates, most recognizable frog TEs (72%) are DNA transposons, rather than the retrotransposons that dominate other genomes (*5*–*8*, *10*). Among these families (*11*, *12*), we identified *Kolobok* as a previously uncharacterized 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). Although 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.

Using homology-based gene prediction methods and deep *Xenopus* EST and cDNA resources, we estimated that the *X. tropicalis* genome contains 20,000 to 21,000 protein-coding genes. These include orthologs of 79% of identified human disease genes (*4*). The genome contains 1850 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*). Sixty percent 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 the human and chicken genome (Fig. 2) (*5*) and in agreement with persistent conservation of linkage groups across chordates (*13*).

We uniquely placed 1696 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 analyzed blocks of conserved synteny between frog, human, and chicken. These blocks were detected with genomic probes comprising three-way orthologs between these tetrapods. Of these probes, 5642 define conserved linkage blocks containing at least 15 genes and at least 2 Mb of sequence (*4*, *14*). The tetrapod ancestry of human and chicken chromosome 1 is outlined in Fig. 2. Notably, a core of more than 150 Mb of sequence spanning the centromere of human chromosome 1 [chicken chromosome 8, frog linkage group (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 megabases. Human chromosome 1 is seen to have grown by three lineage-specific

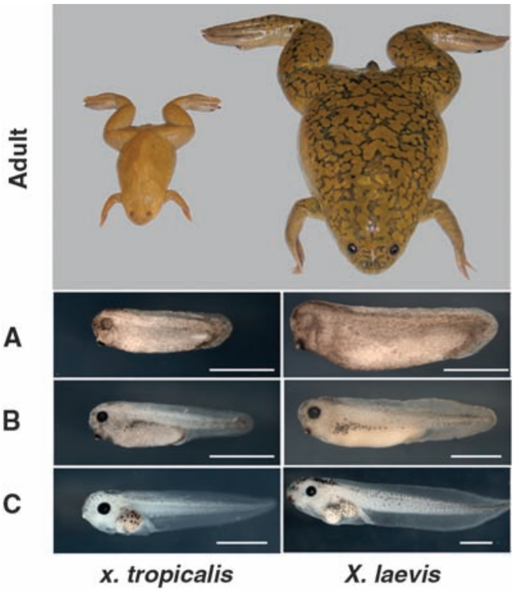

Fig. 1. Comparison of adults and tadpoles of *X. tropicalis* and *X. laevis*. Adult body length is 5 and 10 cm, respectively. (**A**) Tailbud, (**B**) swimming tadpole, and (**C**) feeding tadpole. Bar, 1 mm.

[1]Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA. [2]Center for Integrative Genomics, University of California Berkeley, Berkeley, CA 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]Department of Developmental and Cell Biology, 4410 Natural Sciences Building 2, University of California Irvine, 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, National Cancer Institute, NIH, DHHS Bethesda, MD 20892, USA. [11]Genome Sequencing Center, Washington University School of Medicine, St. Louis, MO 63108, USA. [12]Joint Genome Institute 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 and Immunology, Box 672, University of Rochester, Medical Center, Rochester, NY 14642, USA. [19]Department 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 NW7 1AA, UK. [22]Division of Developmental Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, USA. [23]Department of Biology, Gilmer Hall, Post Office Box 400328, Charlottesville, VA 22904–4328, USA. [24]Department of Pediatrics and Genetics, Yale University School of Medicine, Post Office Box 208064, New Haven, CT 06520–8064, USA.

*To whom correspondence should be addressed. E-mail: uhellsten@lbl.gov

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, whereas 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 that of 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. The chicken microchromosomes are unresolved by this analysis, however, 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 noncoding sequence conservation diminishes steadily with increasing evolutionary distance (fig. S6). Frog genes adjacent to conserved noncoding 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% 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* CNSs 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 that in *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, whereas *X. tropicalis* has more than six. Synteny relationships reveal that *nodal4* on scaffold 204 is orthologous to the single human *nodal*, whereas a cluster of more than six *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, whereas 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*). Although 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 interleukin-2/interleukin-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),

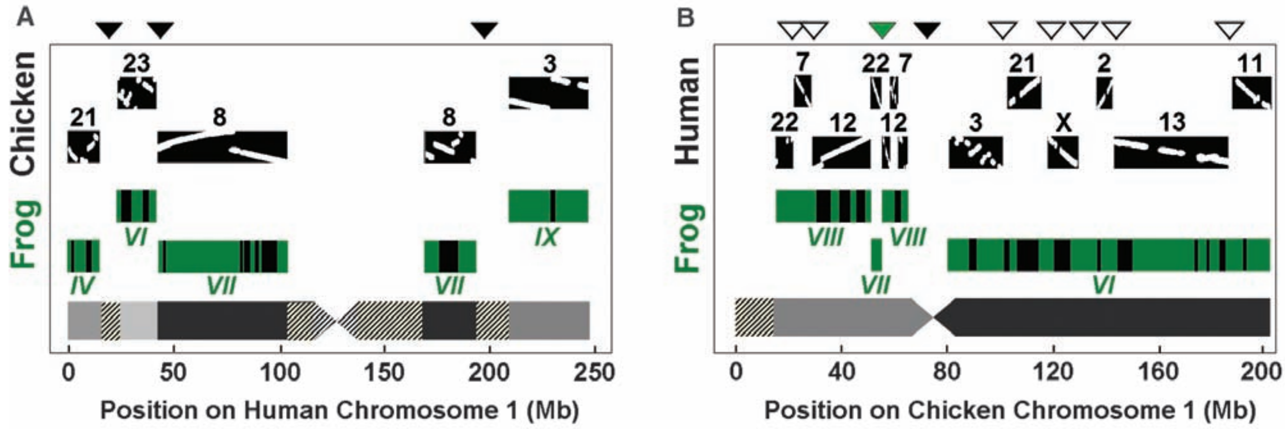

Fig. 2. Blocks of conserved tetrapod linkage for human (**A**) and chicken (**B**) chromosome 1 reveal fusions (solid black triangles) and break points (unfilled triangles) in amniotes. A total of three human fusions (A), seven human breaks (B), and one chicken fusion (B) are observed. The green triangle in (B) indicates the position of an apparent frog-specific break or ancestral amniote fusion. Gray 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 (A) and 97 in (B). Bars on the frog blocks show the location of scaffolds that do not contain markers from the linkage map, but have been predicted to associate with the linkage group by conserved synteny.

and neuropeptides (e.g., thyrotropin-releasing hormone) (*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 >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**, 286 (1930).
2. D. D. Brown, *J. Biol. Chem.* **279**, 45291 (2004).
3. J. Tymowska, *Cytogenet. Cell Genet.* **12**, 297 (1973).
4. Supporting material is available on *Science* Online.
5. International Chicken Genome Sequencing Consortium, *Nature* **432**, 695 (2004).
6. International Human Genome Sequencing Consortium, *Nature* **409**, 860 (2001).
7. Mouse Genome Sequencing Consortium, *Nature* **420**, 520 (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 (2003).
10. International Rice Genome Sequencing Project, *Nature* **436**, 793 (2005).
11. N. L. Craig, R. Craigie, M. Gellert, A. M. Lambowitz, Eds., *Mobile DNA II* (American Society for Microbiology, Washington, DC, 2002).
12. V. V. Kapitonov, J. Jurka, *DNA Cell Biol.* **23**, 311 (2004).
13. N. H. Putnam *et al.*, *Science* **317**, 86 (2007).
14. Dataset S1 is available on *Science* Online.
15. I. G. Woods *et al.*, *Genome Res.* **15**, 1307 (2005).
16. A. Abu-Daya, A. K. Sater, D. E. Wells, T. J. Mohun, L. B. Zimmerman, *Dev. Biol.* **336**, 20 (2009).
17. M. A. Nobrega, I. Ovcharenko, V. Afzal, E. M. Rubin, *Science* **302**, 413 (2003).
18. R. J. Garriock, A. S. Warkman, S. M. Meadows, S. D'Agostino, P. A. Krieg, *Dev. Dyn.* **236**, 1249 (2007).
19. M. Kofron *et al.*, *Development* **126**, 5759 (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 (2009).
22. Y. Ohta, M. Flajnik, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 10723 (2006).
23. Y. Zhao *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 12087 (2006).
24. G. Kreil, in *The Biology of Xenopus*, R. C. Tinsley, H. R. Kobels, Eds. (The Zoological Society of London, Oxford, 1996), pp. 263–277.
25. L. A. Rollins-Smith, *Integr. Comp. Biol.* **45**, 137 (2005).

26. This work was performed under the auspices of the U.S. Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory, under contract DE-AC02-05CH11231, Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344, and Los Alamos National Laboratory under contract 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 R. Gibbs and S. Scherer of the Human Genome Sequencing Center, Baylor College of Medicine, for their contributions to identification and mapping of simple sequence length polymorphisms.

Supporting Online Material

www.sciencemag.org/cgi/content/full/328/5978/633/DC1
Materials and Methods
SOM Text
Figs. S1 to S9
Tables S1 to S17
References
Dataset S1

22 October 2009; accepted 16 February 2010
10.1126/science.1183670

Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing

Jared C. Roach,[1]* Gustavo Glusman,[1]* Arian F. A. Smit,[1]* Chad D. Huff,[1,2]* Robert Hubley,[1] Paul T. Shannon,[1] Lee Rowen,[1] Krishna P. Pant,[3] Nathan Goodman,[1] Michael Bamshad,[4] Jay Shendure,[5] Radoje Drmanac,[3] Lynn B. Jorde,[2] Leroy Hood,[1]† David J. Galas[1]†

We analyzed the whole-genome sequences of a family of four, consisting of two siblings and their parents. Family-based sequencing allowed us to delineate recombination sites precisely, identify 70% of the sequencing errors (resulting in >99.999% accuracy), and identify very rare single-nucleotide polymorphisms. We also directly estimated a human intergeneration mutation rate of ~1.1 × 10^{-8} per position per haploid genome. Both offspring in this family have two recessive disorders: Miller syndrome, for which the gene was concurrently identified, and primary ciliary dyskinesia, for which causative genes have been previously identified. Family-based genome analysis enabled us to narrow the candidate genes for both of these Mendelian disorders to only four. Our results demonstrate the value of complete genome sequencing in families.

Whole-genome sequences from four members of a family represent a qualitatively different type of genetic data than whole-genome sequences from individual or sets of unrelated genomes. They enable inheritance analyses that detect errors and permit the identification of precise locations of recombination events. This leads in turn to near-complete knowledge of inheritance states through the precise determination of the parental chromosomal origins of sequence blocks in offspring. Confident predictions of inheritance states and haplotypes power analyses that include the identification of genomic features with nonclassical inheritance patterns, such as hemizygous deletions or copy number variants (CNVs). Identification of inheritance patterns in the pedigree permits the detection of ~70% of sequencing errors and sharply reduces the search space for disease-causing variants. These analyses would be far less powerful in studies that had fewer markers (such as standard genotype or exome data sets) or that had sequences from fewer family members.

DNA from each family member was extracted from peripheral blood cells and sequenced at CGI (Mountain View, California) with a nanoarray-based short-read sequencing-by-ligation technology (*1*), including an adaptation of the pairwise end-sequencing strategy (*2*). Reads were mapped to the National Center for Biotechnology Information (NCBI) reference genome (fig. S1 and tables S1 and S2). Polymorphic markers used for this analysis were single-nucleotide polymorphisms (SNPs) with at least two variants among the four genotypes of the family, averaging 802 base pairs (bp) between markers. We observed 4,471,510 positions at which at least one family member had an allele that varied from the reference genome. This corresponds to a Watterson's theta (θ_W) of 9.5 × 10^{-4} per site for the two parents and the reference sequence (*3*), given the fraction of the genome successfully genotyped in each parent (fig. S1). This is a close match to the estimate of $\theta_W = 9.3 \times 10^{-4}$ that we obtained by combining two previously published European genomes and the reference sequence (*4*). Of the 4.5 million variant positions, 3,665,772 were variable within the family; the rest were homozygous and identical in all four members. Comparisons to known SNPs show that 323,255 of these 3.7 million SNPs are novel.

For each meiosis in a pedigree, each base position in a resulting gamete will have inherited one of two parental alleles. The number of inheritance patterns of the segregation of alleles in

[1]Institute for Systems Biology, Seattle, WA 98103, USA. [2]Department of Human Genetics, Eccles Institute of Human Genetics, University of Utah, Salt Lake City, UT 84109, USA. [3]Complete Genomics, Inc. (CGI), Mountain View, CA 94043, USA. [4]Department of Pediatrics, University of Washington, Seattle, WA 98195, USA. [5]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA.

*These authors contributed equally to this work.

†To whom correspondence should be addressed. E-mail: dgalas@systemsbiology.org (D.J.G.); lhood@systemsbiology.org (L.H.)