Published in final edited form as:
*Nature*. 2013 April 18; 496(7445): 311–316. doi:10.1038/nature12027.

# Analysis of the African coelacanth genome sheds light on tetrapod evolution

**Chris T. Amemiya**[*,1,2], **Jessica Alföldi**[*,3], **Alison P. Lee**[4], **Shaohua Fan**[5], **Hervé Philippe**[6], **Iain MacCallum**[3], **Ingo Braasch**[7], **Tereza Manousaki**[5,8], **Igor Schneider**[9], **Nicolas Rohner**[10], **Chris Organ**[11], **Domitille Chalopin**[12], **Jeramiah J. Smith**[13], **Mark Robinson**[1], **Rosemary A. Dorrington**[14], **Marco Gerdol**[15], **Bronwen Aken**[16], **Maria Assunta Biscotti**[17], **Marco Barucca**[17], **Denis Baurain**[18], **Aaron M. Berlin**[3], **Gregory L. Blatch**[14,19], **Francesco Buonocore**[20], **Thorsten Burmester**[21], **Michael S. Campbell**[22], **Adriana Canapa**[17], **John P. Cannon**[23], **Alan Christoffels**[24], **Gianluca De Moro**[15], **Adrienne L. Edkins**[14], **Lin Fan**[3], **Anna Maria Fausto**[20], **Nathalie Feiner**[5,25], **Mariko Forconi**[17], **Junaid Gamieldien**[24], **Sante Gnerre**[3], **Andreas Gnirke**[3], **Jared V. Goldstone**[26], **Wilfried Haerty**[27], **Mark E. Hahn**[26], **Uljana Hesse**[24], **Steve Hoffmann**[28], **Jeremy Johnson**[3], **Sibel I. Karchner**[26], **Shigehiro Kuraku**[5,**], **Marcia Lara**[3], **Joshua Z. Levin**[3], **Gary W. Litman**[23], **Evan Mauceli**[3,***], **Tsutomu Miyake**[29], **M. Gail Mueller**[30], **David R. Nelson**[31], **Anne Nitsche**[32], **Ettore Olmo**[17], **Tatsuya Ota**[33], **Alberto Pallavicini**[15], **Sumir Panji**[24,****], **Barbara Picone**[24], **Chris P. Ponting**[27], **Sonja J. Prohaska**[34], **Dariusz Przybylski**[3], **Nil Ratan Saha**[1], **Vydianathan Ravi**[4], **Filipe J. Ribeiro**[3,*****], **Tatjana Sauka-Spengler**[35,37,38,39], **Giuseppe Scapigliati**[20], **Stephen M. J. Searle**[16], **Ted Sharpe**[3], **Oleg Simakov**[5,36], **Peter F. Stadler**[32], **John J. Stegeman**[26], **Kenta Sumiyama**[40], **Diana Tabbaa**[3], **Hakim Tafer**[32], **Jason Turner-Maier**[3], **Peter van Heusden**[24], **Simon White**[16], **Louise Williams**[3], **Mark Yandell**[22], **Henner Brinkmann**[6], **Jean-Nicolas Volff**[12], **Clifford J. Tabin**[10],

Correspondence and requests for materials should be addressed to: camemiya@benaroyaresearch.org, jalfoldi@broadinstitute.org, axel.meyer@unikonstanz.de, and kersli@broadinstitute.org.
[*]These authors contributed equally to this work
[**]Present address: Genome Resource and Analysis Unit, Center for Developmental Biology, RIKEN, Kobe, Japan
[***]Present address: Boston Children's Hospital, Boston, MA
[****]Present address: Computational Biology Unit, Institute of Infectious Disease and Molecular Medicine, University of Cape Town Health Sciences Campus, Anzio Road, Observatory 7925, South Africa
[*****]Present address: New York Genome Center, New York, NY

**Supplementary Information** is linked to the online version of the paper at www.nature.com/nature.

**Author contributions** JA, CTA, AM and KLT planned and oversaw the project. RD and CTA provided blood and tissues for sequencing. CTA and ML prepared the DNA for sequencing. IM, SG, DP, FJR, TS and DJ assembled the genome. NRS prepared RNA from *L. chalumnae* LF and JL made the *L. chalumnae* RNA-seq library. AC, MB, MAB, MF, FB, GS, AMF, AP, MG, GDM, JT-M and EO sequenced and analyzed the *L. menadoensis* RNA-seq library. BA, SMJS, SW, MC and MY annotated the genome. WH and CPP performed the lncRNAs annotation and analysis. PFS, SH, AN, HT, and SJP annotated ncRNAs. MG, GDM, AP, MR and CTA compared *L. chalumnae* and *L. menadoensis* sequence. HB, DB and HP performed the phylogenomic analysis. TMa and AM performed the gene relative rate analysis. AC, JG, SP, BP, PvH and UH performed the analysis, annotation and statistical enrichment of *L. chalumnae* specific gene duplications. NF and AM analyzed the homeobox gene repertoires. DC, SF, OS, J-NV, MS and AM analysed transposable elements. JJS analysed large scale rearrangements in vertebrate genomes. IB, JP, NF and SK analysed genes lost in tetrapods. TMi analyzed actinodin and pectoral fin musculature. CO and MS analysed selection in urea cycle genes. AL and BV performed the conserved non-coding element analysis. IS, NR, VR, NS and CT performed the analysis of autopodial CNEs. KS, TS-S and CTA examined the evolution of a placenta-related CNE. NRS, GWL, MGM, TO and CTA performed the IgM analysis. JA, CTA, AM and KLT wrote the paper with input from other authors. AG, DT and LW constructed fosmid libraries for the *L. chalumnae* genome assembly.

**Author Information**

The *L. chalumnae* genome assembly has been deposited in GenBank under the accession number AFYH00000000. The *L. chalumnae* transcriptome has been deposited under the accession number SRX117503 and the *P. annectans* transcriptomes have been deposited under the accession numbers SRX152529, SRX152530, and SRX152531. The *P. annectans* mitochondrial DNA sequence was deposited under the accession number JX568887. All animal experiments were approved by the MIT Committee for Animal Care.

The authors declare no competing financial interests.

**Neil Shubin**[41], **Manfred Schartl**[42], **David Jaffe**[3], **John H. Postlethwait**[7], **Byrappa Venkatesh**[4], **Federica Di Palma**[3], **Eric S. Lander**[3], **Axel Meyer**[5,8,25], and **Kerstin Lindblad-Toh**[3,43]

[1]Molecular Genetics Program, Benaroya Research Institute, Seattle, WA

[2]Department of Biology, University of Washington, Seattle, WA

[3]Broad Institute of MIT and Harvard, Cambridge, MA

[4]Comparative Genomics Laboratory, Institute of Molecular and Cell Biology, A*STAR, Biopolis, Singapore, Singapore

[5]Department of Biology, University of Konstanz, Konstanz, Germany

[6]Departement de Biochimie, Universite de Montreal, Centre Robert Cedergren, Montreal, Canada

[7]Institute of Neuroscience, University of Oregon, Eugene, OR

[8]Konstanz Research School of Chemical Biology, University of Konstanz, Konstanz, Germany

[9]Instituto de Ciencias Biologicas, Universidade Federal do Para, Belem, Brazil

[10]Department of Genetics, Harvard Medical School, Boston, MA

[11]Department of Anthropology, University of Utah, Salt Lake City, UT

[12]Institut de Genomique Fonctionnelle de Lyon, Ecole Normale Superieure de Lyon, Lyon, France

[13]Department of Biology, University of Kentucky, Lexington, KY

[14]Biomedical Biotechnology Research Unit (BioBRU), Department of Biochemistry, Microbiology & Biotechnology, Rhodes University, Grahamstown, South Africa

[15]Department of Life Sciences, University of Trieste, Trieste, Italy

[16]Department of Informatics, Wellcome Trust Sanger Institute, Hinxton, UK

[17]Department of Life and Environmental Sciences, Polytechnic University of the Marche, Ancona, Italy

[18]Department of Life Sciences, University of Liege, Liege, Belgium

[19]College of Health and Biomedicine, Victoria University, Melbourne, Australia

[20]Department for Innovation in Biological, Agro-food and Forest Systems, University of Tuscia, Viterbo, Italy

[21]Department of Biology, University of Hamburg, Hamburg, Germany

[22]Eccles Institute of Human Genetics, University of Utah, Salt Lake City, UT

[23]Department of Pediatrics, University of South Florida Morsani College of Medicine, Children's Research Institute, St. Petersburg, FL

[24]South African National Bioinformatics Institute, University of the Western Cape, Bellville, South Africa

[25]International Max-Planck Research School for Organismal Biology, University of Konstanz, Konstanz, Germany

[26]Biology Department, Woods Hole Oceanographic Institution, Woods Hole, MA

[27]MRC Functional Genomics Unit, Oxford University, Oxford, UK

[28]Transcriptome Bioinformatics Group, LIFE Research Center for Civilization Diseases, Universität Leipzig, Leipzig, Germany

[29]Graduate School of Science and Technology, Keio University, Yokohama, Japan

[30]Department of Molecular Genetics, All Children's Hospital, St. Petersburg, FL

[31]Department of Microbiology, Immunology and Biochemistry, University of Tennessee Health Science Center, Memphis, TN

[32]Bioinformatics Group, Department of Computer Science, Universität Leipzig, Leipzig, Germany

[33]Department of Evolutionary Studies of Biosystems, The Graduate University for Advanced Studies, Hayama, Japan

[34]Computational EvoDevo Group, Department of Computer Science, Universität Leipzig, Leipzig, Germany

[35]Institute of Molecular Medicine, Oxford University, Oxford, UK

[36]European Molecular Biology Laboratory, Heidelberg, Germany

[37]Max-Planck-Institute for Mathematics in the Sciences, Leipzig, Germany

[38]Fraunhofer Institute for Cell Therapy and Immunology, Leipzig, Germany

[39]The Santa Fe Institute, Santa Fe, NM

[40]Division of Population Genetics, National Institute of Genetics, Mishima, Japan

[41]University of Chicago, Chicago, IL

[42]Department Physiological Chemistry, Biocenter, University of Wuerzburg, Wuerzburg Germany

[43]Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden

## Abstract

It was a zoological sensation when a living specimen of the coelacanth was first discovered in 1938, as this lineage of lobe-finned fish was thought to have gone extinct 70 million years ago. The modern coelacanth looks remarkably similar to many of its ancient relatives, and its evolutionary proximity to our own fish ancestors provides a glimpse of the fish that first walked on land. Here we report the genome sequence of the African coelacanth, *Latimeria chalumnae*. Through a phylogenomic analysis, we conclude that the lungfish, and not the coelacanth, is the closest living relative of tetrapods. Coelacanth protein-coding genes are significantly more slowly evolving than those of tetrapods, unlike other genomic features . Analyses of changes in genes and regulatory elements during the vertebrate adaptation to land highlight genes involved in immunity, nitrogen excretion and the development of fins, tail, ear, eye, brain, and olfaction. Functional assays of enhancers involved in the fin-to-limb transition and in the emergence of extra-embryonic tissues demonstrate the importance of the coelacanth genome as a blueprint for understanding tetrapod evolution.

## Introduction

It was 1938 when Ms. Marjorie Courtenay-Latimer, the curator of a small natural history museum in East London, South Africa, discovered a large, peculiar looking fish among the myriad specimens delivered to her by a local fish trawler. *Latimeria chalumnae*, named after its discoverer[1], was over one meter long, bluish in coloration, and had conspicuously fleshy fins that resembled the limbs of terrestrial vertebrates. This discovery turned out to be a biological sensation and is considered one of the greatest zoological finds of the 20th century. *Latimeria* is the only living member of an ancient group of lobe-finned fishes

previously known only from fossils and believed to have been extinct since the Late Cretaceous period, about 70 million years ago (MYA)[1]. It took almost 15 years before a second specimen of this elusive species was discovered in the Comoros Islands in the Indian Ocean, and only a total of 309 individuals that are known to science, have been found in the past 75 years (Rik Nulens, personal communication)[2]. The discovery in 1997 of a second coelacanth species in Indonesia, *L. menadoensis*, was equally surprising, as it had been assumed that living coelacanths were confined to small populations off the East African coast[3–4]. Fascination with these fish is partly due to their prehistoric appearance – remarkably, their morphology is similar to that of fossils that date back at least 300 million years, leading to the supposition that this lineage is especially slow-evolving among vertebrates[1,5]. *Latimeria* has also been of particular interest to evolutionary biologists due to its hotly debated relationship to our last fish ancestor – the fish that first crawled up on land[6]. In the past 15 years, targeted sequencing efforts have yielded the sequences of the coelacanth mitochondrial genomes[7], HOX clusters[8], and a few gene families[9–10], but still, coelacanth research has felt the lack of large-scale sequencing data.

Here we describe the sequencing and comparative analysis of the genome of *L. chalumnae*, the African coelacanth.

## Genome assembly and annotation

The African coelacanth genome was sequenced and assembled (LatCha1.0) using DNA from a Comoros Islands *Latimeria chalumnae* specimen (Supplementary Figure 1). It was sequenced by Illumina sequencing technology and assembled via ALLPATHS-LG[11]. The *L. chalumnae* genome has previously been reported to have a karyotype of 48 chromosomes[12]. The draft assembly is 2.86 Gb in size and is composed of 2.18 Gb of sequence plus gaps between contigs. The coelacanth genome assembly has a contig N50 size of 12.7 kb and a scaffold N50 size of 924 kb, and quality metrics comparable to other Illumina genomes (See Supplementary Note 1, Supplementary Tables 1,2).

The genome assembly was annotated separately by both the Ensembl gene annotation pipeline (Ensembl release 66, February 2012) and by MAKER[13]. The Ensembl gene annotation pipeline created gene models using Uniprot protein alignments, limited coelacanth cDNA data, RNA-seq data generated from *L. chalumnae* muscle (18 Gb of paired end reads were assembled by Trinity[14], Supplementary Figure 2) as well as orthology with other vertebrates. This pipeline produced 19,033 protein coding genes containing 21,817 transcripts. The MAKER pipeline used the *L. chalumnae* Ensembl gene set, Uniprot protein alignments, and *L. chalumnae* (muscle) and *L. menadoensis* (liver and testis)[15] RNA-seq to create gene models, yielding 29,237 protein coding gene annotations. In addition, 2,894 short non-coding RNAs, 1,214 lncRNAs and more than 24,000 conserved RNA secondary structures were identified (Supplementary Note 2, Supplementary Tables 3–4, Supplementary Dataset 1–3, Supplementary Figure 3). 336 genes were inferred to have undergone specific duplications in the coelacanth lineage (Supplementary Note 3, Supplementary Tables 5–6, Supplementary Dataset 4).

## Closest living fish relative of tetrapods

The question of which living fish is the closest relative to ‘the fish that first crawled up on land’ has long captured our imagination: among scientists the odds have been placed on either the lungfish or the coelacanth[16]. Analyses of small to moderate amounts of sequence data for this important phylogenetic question (ranging from 1 to 43 genes) has tended to favor the lungfishes as the extant sister group to the land vertebrates[17], however, the alternative hypothesis that lungfish and coelacanth are equally closely related to the tetrapods could not be rejected with previous data sets[18].

To seek a comprehensive answer we generated RNA-seq data from three samples (brain, gonad/kidney, gut/liver) from the West African lungfish, *Protopterus annectens*, and compared it to gene sets from 21 strategically chosen jawed vertebrate species. To perform a reliable analysis we selected 251 genes where 1–1 orthology was clear and used CAT-GTR, a complex site-heterogeneous model of sequence evolution known to reduce tree reconstruction artefacts[19] (see Methods). The resulting phylogeny, based on 100,583 concatenated amino acid positions, (Figure 1, PP=1.0 for the lungfish-tetrapod node) is fully resolved except for the relative positions of armadillo and elephant. It corroborates known vertebrate phylogenetic relationships and strongly supports the conclusion that tetrapods are more closely related to lungfish than to the coelacanth (Supplementary Note 4, Supplementary Figure 4).

## How slowly evolving is the coelacanth?

The morphological resemblance of the modern coelacanth to its fossil ancestors has resulted in it being nicknamed 'the living fossil'[1]. This invites the question: Is the genome of the coelacanth as slowly evolving as its outward appearance suggests? Earlier work found that a few gene families, such as Hox and protocadherins, showed comparatively slower protein-coding evolution in coelacanth than in other vertebrate lineages[8,10]. To address this question, we examined several types of genomic changes in the coelacanth compared to other vertebrates.

Protein-coding gene evolution was examined using the 251 concatenated protein phylogenomics dataset (Figure 1). Pair-wise distances between taxa were calculated from the branch lengths of the tree using the Two-Cluster test proposed by Takezaki *et al.*[20] to test for equality of average substitution rates. Then, for each of the following species and species clusters (coelacanth, lungfish, chicken and mammals), we ascertained their respective mean distance to an outgroup consisting of three cartilaginous fishes (elephant shark, little skate and spotted catshark). Finally, we tested whether there was any significant difference in distance to the outgroup of cartilaginous fish for every pair of species and species clusters, using a Z-statistic. When these distances to the outgroup of cartilaginous fish were compared, we found that the coelacanth proteins tested were significantly more slowly evolving (0.890 substitutions/site) than the lungfish (1.05 substitutions/site), chicken (1.09 substitutions/site) and mammalian (1.21 substitutions/site) orthologues (Supplementary Dataset 5), in all cases with p-values $<10^{-6}$. Additionally, as can be seen in Figure 1, the substitution rate in coelacanth is approximately half that in tetrapods since the two lineages diverged. A Tajima relative rate test[21] confirmed the coelacanth's significantly slower rate of protein evolution (Supplementary Dataset 6).

Secondly, we examined the abundance of transposable elements (TEs) in the coelacanth genome. Theoretically, TEs might contribute most significantly to the evolution of a species by generating templates for exaptation to form novel regulatory elements and exons, and by acting as substrates for genomic rearrangement[22]. We found that the coelacanth genome contains a wide variety of TE superfamilies and has a relatively high TE content (25%); this number is likely an underestimate due to the draft nature of the assembly (Supplementary Note 5, Supplementary Tables 7–10). Analysis of RNA-seq data and of the divergence of individual TE copies from consensus sequences show that 14 coelacanth TE super-families are currently active (Supplementary Note 6, Supplementary Table 10, Supplementary Figure 5). We conclude that the current coelacanth genome shows both an abundance and activity of TEs similar to many other genomes. This contrasts with the slow protein evolution observed.

Analyses of chromosomal breakpoints in the coelacanth genome and tetrapod genomes reveal extensive conservation of synteny and indicate that large-scale rearrangements have occurred at a generally low rate in the coelacanth lineage. Analyses of these rearrangement classes detected several previously published fission events that are known to have occurred in tetrapod lineages and at least 31 interchromosomal rearrangements that occurred in the coelacanth lineage or the early tetrapod lineage (0.063 fusions/million years), compared to 20 events (0.054 fusions/million years) in the salamander lineage and 21 events (0.057 fusions/million years) in the *Xenopus* lineage[23] (Supplementary Note 7, Supplementary Figure 6). Overall, these analyses indicate that karyotypic evolution in the coelacanth lineage has occurred at a relatively slow rate, similar to that of non-mammalian tetrapods[24].

In a separate analysis we also examined the evolutionary divergence between the two species of coelacanth, *L. chalumnae* and *L. menadoensis*, found in African and Indonesian waters respectively. Previous analysis of mitochondrial DNA showed a sequence identity of 96%, but estimated divergence times range widely from 6 to 40 million years[25–26]. When we compared the liver and testis transcriptomes of *L. menadoensis*[27] to the *L. chalumnae* genome, we found an identity of 99.73% (Supplementary Note 8, Supplementary Figure 7), whereas alignments between 20 sequenced *L. menadoensis* BACs and the *L. chalumnae* genome showed an identity of 98.7% (Supplementary Table 11, Supplementary Figure 8). Both the genic and genomic divergence rates are similar to those seen between the human and chimpanzee genomes (99.5% and 98.8% respectively, divergence time 6–8 million years ago)[28], while the rates of molecular evolution in *Latimeria* are likely affected by multiple factors including the slower substitution rate seen in coelacanth, thereby suggesting a slightly larger divergence time for the two coelacanth species.

## Vertebrate adaptation to land

As the sequenced genome closest to our most recent aquatic ancestor, the coelacanth provides a unique opportunity to identify genomic changes that were associated with the successful adaptation of vertebrates to an important new environment – land.

Over the 400 MY interval that vertebrates have lived on land, genes that are unnecessary for existence in their new environment would have been eliminated. To understand this aspect of the water-to-land transition, we surveyed the *Latimeria* genome annotations to identify genes that were present in the last common ancestor of all bony fish (including coelacanth) but that are missing from tetrapod genomes. More than 50 such genes including components of the Fgf signaling, TGF-beta/Bmp signaling, and Wnt signaling pathways, as well as many transcription factor genes, were inferred to be lost based on the coelacanth data (Supplementary Dataset 7, Supplementary Figure 9). Previous studies of genes lost in this transition could only compare teleost fish to tetrapods, meaning that differences in gene content could have been due to loss in the tetrapod or in the lobe-finned fish lineages. We were able to confirm that four genes previously shown to be absent in tetrapods (*Actinodin* genes[29], *Fgf24*[30], *Asip2*[31]), were indeed present and intact in *Latimeria*, supporting their loss in the tetrapod lineage.

We functionally annotated the >50 genes lost in tetrapods using zebrafish data (gene expression, knock-downs and knock-outs). Many genes were classified in important developmental categories (Supplementary Dataset 7): Fin development (13 genes), otolith and ear development (8 genes), kidney development (7 genes), trunk/somite/tail development (11 genes), eye (13 genes), and brain development (23 genes). This implies that critical characters in the morphological transition from water to land (fin-to-limb transition, remodelling of the ear, etc.) are reflected in the loss of specific genes along the phylogenetic branch leading to tetrapods. However, homeobox genes, which are responsible

for the development of an organism's basic body plan, show only slight differences between *Latimeria*, ray-finned fish and tetrapods; it would appear that the protein-coding portion of this gene family, along with several others (Supplementary Note 9, Supplementary Tables 12–16, Supplementary Figure 10), have remained largely conserved during the vertebrate land transition. (Supplementary Figure 11).

As vertebrates transitioned to a new land environment, changes occurred not only in gene content, but also in the regulation of existing genes. Conserved non-coding elements (CNEs) are strong candidates for gene regulatory elements and can act as promoters, enhancers, repressors and insulators[32–33], and have been implicated as major facilitators of evolutionary change[34]. To identify CNEs that originated in the most recent common ancestor of tetrapods, we predicted CNEs that evolved in various bony vertebrate (i.e., ray-finned fish, coelacanth and tetrapod) lineages and assigned them to their likely branch points of origin. To detect CNEs, conserved sequences in the human genome were identified using MULTIZ alignments of bony vertebrate genomes, and then known protein-coding sequences, UTRs and known RNA genes were excluded. Our analysis identified 44,200 ancestral tetrapod CNEs that originated after the divergence of the coelacanth lineage. They represent 6% of the 739,597 CNEs that are under constraint in the bony vertebrate lineage. We compared the ancestral tetrapod CNEs to mouse embryo ChIP-seq data obtained using antibodies against p300, a transcriptional co-activator. This resulted in a 7-fold enrichment in the p300 binding sites for our candidate CNEs and confirmed that these CNEs are indeed enriched for gene regulatory elements.

Each tetrapod CNE was assigned to the gene whose transcription start site was closest, and GO category enrichment was calculated for those genes. The most enriched categories were involved with smell perception (sensory perception of smell, detection of chemical stimulus, olfactory receptor activity etc.). This is consistent with the notable expansion of olfactory receptor family genes in tetrapods compared with teleosts, and may reflect the necessity of a more tightly regulated, larger and more diverse repertoire of olfactory receptors for detecting airborne odorants as part of the terrestrial lifestyle. Other significant categories include morphogenesis (radial pattern formation, hind limb morphogenesis, kidney morphogenesis) and cell differentiation (endothelial cell fate commitment, epithelial cell fate commitment), which is consistent with the body plan changes required for land transition, as well as immunoglobulin VDJ recombination, which reflects the presumed response differences required to address the novel pathogens that vertebrates would encounter on land (Supplementary Note 10, Supplementary Tables 17–24).

A major innovation of tetrapods is the evolution of limbs characterised by digits. The limb skeleton consists of a stylopod (humerus or femur), the zeugopod (radius/ulna and tibia/fibula), and an autopod (wrist/ankle and digits). There are two major hypotheses about the origins of the autopod – either it was a novel feature of tetrapods, or it has antecedents in the fins of fish[35] (Supplementary Note 11, Supplementary Figure 12). We examine here the Hox regulation of limb development in ray-finned fish, coelacanth, and tetrapods to address these hypotheses.

In mouse, late phase digit enhancers are located in a gene desert located proximal to the HOX-D cluster[36]. Here we provide an alignment of the HoxD centromeric gene desert of coelacanth with tetrapods and ray-finned fishes (Figure 2a). Among the six cis-regulatory sequences previously identified in this gene desert[36], three sequences show sequence conservation restricted to tetrapods (Supplementary Figure 13). However, one regulatory sequence (Island 1) is shared between tetrapods and coelacanth, but not with ray-finned fish (Figure 2b, Supplementary Figure 14). When tested in a transient transgenic assay in mouse, the coelacanth sequence of Island 1 was able to drive reporter expression in a limb specific

pattern (Figure 2c), making it likely that Island 1 was a lobe-fin developmental enhancer in the fish ancestor of tetrapods that was then coopted into the autopod enhancer of modern tetrapods. In this case, the autopod developmental regulation was derived from an ancestral lobe-finned fish regulatory element.

Changes in the urea cycle provide an illuminating example of the adaptations associated with transition to land. Excretion of nitrogen is a major physiological challenge for terrestrial vertebrates. In aquatic environments, the primary nitrogenous waste product is ammonia, which is readily diluted by surrounding water before it reaches toxic levels, but on land, less toxic substances such as urea or uric acid must be produced instead (Supplementary Figure 15). The widespread and almost exclusive occurrence of urea excretion in amphibians, some turtles and mammals has led to the hypothesis that the use of urea as the main nitrogenous waste product was a key innovation in the vertebrate transition from water to land[37].

With the availability of gene sequences from coelacanth and lungfish, it became possible to test this hypothesis. We used a branch-site model in the HYPHY package[38], which estimates dN/dS (ω) values among different branches and among different sites (codons) across a multiple species sequence alignment. For the rate-limiting enzyme of the hepatic urea cycle, carbamoyl phosphate synthase I (CPS1), only one branch of the tree shows a strong signature of selection ($p = 0.02$), namely the branch leading to tetrapods and the branch leading to amniotes (Figure 3); no other enzymes in this cycle showed a signature of selection. Conversely, mitochondrial arginase (ARG2), which produces extrahepatic urea as a byproduct of arginine metabolism but which is not involved in the production of urea for nitrogenous waste disposal, did not show any evidence of selection in vertebrates (Supplementary Figure 16). This leads us to conclude that adaptive evolution occurred in the hepatic urea cycle during the vertebrate land transition. In addition, it is interesting to note that of the five amino acids of CPS1 that changed between coelacanth and tetrapods, three are in important domains (ATP-A site, ATP-B site, subunit interaction domain) and a fourth is known to cause a malfunctioning enzyme in human patients if mutated[39].

The adaptation to a terrestrial lifestyle necessitated major changes in the physiological milieu of the developing embryo and fetus, resulting in the evolution and specialization of extraembryonic membranes of the amniote mammals[40]. The placenta, in particular, is a complex structure that is critical for providing gas and nutrient exchange between mother and fetus and is also a major site of hematopoiesis[41].

We have identified a region of the coelacanth HOX-A cluster that may have been involved in the evolution of extraembryonic structures in tetrapods, including the eutherian placenta. Global alignment of the coelacanth *Hoxa14-a13* region with the homologous regions of the horn shark, chicken, human and mouse yielded a CNE just upstream of the coelacanth *Hoxa14* gene (Supplementary Figure 17a, arrow). This conserved stretch is not found in teleost fishes but is highly conserved among horn shark, chicken, human and mouse despite the fact that the latter three have no *Hoxa14* orthologues, and that the horn shark *Hoxa14* gene has become a pseudogene. This CNE, HA14E1, corresponds to the proximal promoter-enhancer region of the *Hoxa14* gene in *Latimeria*. HA14E1 is >99% identical between mouse, human and all other sequenced mammals, and would thus be considered an ultraconserved element[42]. The high level of conservation suggests that this element, which already possessed promoter activity, may have been coopted for other functions despite the loss of the *Hoxa14* gene in amniotes (Supplementary Figure 17bc). Expression of human HA14E1 in a mouse transient transgenic assay did not give notable expression in the embryo *proper* at day 11.5[43], which was unexpected since its location would predict that it would regulate axial structures caudally[44]. A similar experiment in chick embryos using the

chicken HA14E1 also showed no activity in the AP-axis. However, stunning expression was observed in the extraembryonic *area vasculosa* of the chick embryo (Figure 4a). Examination of a *Latimeria* BAC *Hoxa14*-reporter transgene in mouse embryos showed that the Hoxa14 gene is specifically expressed in a subset of cells in an extraembryonic region at E8.5 (Figure 4b).

These findings suggest that the HA14E1 region may have been evolutionarily recruited to coordinate regulation of posterior HoxA genes (*Hoxa13*, *Hoxa11* and *Hoxa10*), which are known to be expressed in the mouse allantois and are critical for early formation of the mammalian placenta[45]. Although *Latimeria* does not possess a placenta, it is a livebearer and has very large, vascularised eggs, but the relationship of *Hoxa14*, the HA14E1 enhancer, and blood island formation in the coelacanth remains unknown.

## Coelacanth lacks IgM

Immunoglobulin M (IgM), a class of antibodies, has been reported in all vertebrate species thus far characterised and is considered to be indispensable for adaptive immunity[46]. Interestingly, IgM genes cannot be found in coelacanth despite an exhaustive search of the coelacanth sequence data, and even though all other major components of the immune system are present (Supplementary Note 12, Supplementary Figure 18). Instead, we found two IgW genes (Supplementary Figures 19–21), immunoglobulin genes only found in lungfish and cartilaginous fish and which are believed to have originated in the ancestor of jawed vertebrates[47] and to have been subsequently lost in teleosts and tetrapods. IgM was similarly absent from the *Latimeria* RNA-seq data, although both IgW genes were found as transcripts. To further characterise the apparent absence of IgM, we exhaustively screened large genomic *L. menadoensis* libraries using numerous strategies and probes and also performed PCR with degenerate primers that should universally amplify IgM sequences. The lack of IgM in *Latimeria* raises questions as to how coelacanth B cells respond to microbial pathogens and whether the IgW molecules can serve a compensatory function, even though there is no indication that the coelacanth IgW was derived from vertebrate IgM genes.

## Discussion

Ever since its discovery, the coelacanth has been referred to as a ‘living fossil’ due to its morphological similarities to its fossil ancestors[1]. However, questions have remained as to whether it truly is slowly evolving, as morphological stasis does not necessarily imply genomic stasis. In this study, we determined that *L. chalumnae*’s protein-coding genes show a decreased substitution rate compared to those of other sequenced vertebrates, even though its genome as a whole does not show evidence of low genome plasticity. The reason for this lower substitution rate is still unknown, although a static habitat and a lack of predation over evolutionary timescales could be contributing factors to a lower need for adaptation. A closer examination of gene families that show either unusually high or low levels of directional selection indicative of adaptation in the coelacanth, could tell us a great deal about which selective pressures, or lack thereof, shaped this evolutionary relict (Supplementary Note 13, Supplementary Figure 22).

The vertebrate land transition is one of the most important steps in our evolutionary history. We conclude that the closest living fish to the tetrapod ancestor is the lungfish, not the coelacanth. However, the coelacanth is critical for our understanding of this transition, as the lungfish have intractable genome sizes (estimated at 50–100 Gb)[48]. We have already learned a great deal about our adaptation to land through coelacanth whole genome analysis, and we have shown the promise of focused analysis of specific gene families involved in this

process. Still, further study of these changes between tetrapods and the coelacanth will undoubtedly yield important insights as to how a complex organism like a vertebrate can so drastically change its way of life.

**Methods:** Appear in the online supplement.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

Acquisition and storage of *Latimeria chalumnae* samples was supported by grants from the African Coelacanth Ecosystem Programme of the South African National Department of Science and Technology. Generation of the *Latimeria chalumnae* and *Protopterus annectens* sequence by Broad Institute of MIT and Harvard was supported by grants from the National Human Genome Research Institute (NHGRI). KLT is the recipient of a EURYI award from the ESF. We would also like to thank the Genomics Sequencing Platform of the Broad Institute for sequencing the *L. chalumnae* genome and *L. chalumnae* and *P. annectens* transcriptomes, Said Ahamada, Robin Stobbs and the Association pour le Protection de Gombesa (APG) for their help in obtaining coelacanth samples, Yu Zhao for the use of data from *Rana chensinensis,* and Leslie Gaffney, Catherine Hamilton and John Westlund for assistance with figure preparation.

## References

1. Smith JLB. A Living Fish of Mesozoic Type. Nature. 1939; 143:455–456. doi:10.1038/143455a0.
2. Nulens, R.; Scott, L.; Herbin, M. An Updated Inventory of All Known Specimens of the Coelacanth, Latimeria Spp: By Rik Nulens, Lucy Scott and Marc Herbin. 2010.
3. Erdmann M, Caldwell R, Kasim Moosa M. Indonesian 'king of the sea' discovered. Nature. 1998; 395:335.
4. Smith, JL. Old Fourlegs: The story of the coelacanth. Longmans, Green; 1956.
5. Zhu M, et al. Earliest known coelacanth skull extends the range of anatomically modern coelacanths to the Early Devonian. Nat Commun. 2012; 3:772. doi:ncomms1764 [pii] 10.1038/ncomms1764. [PubMed: 22491320]
6. Zimmer, C. At the Water's Edge: Fish with Fingers, Whales with Legs, and How Life Came Ashore but Then Went Back to Sea. Free Press; 1999.
7. Zardoya R, Meyer A. The complete DNA sequence of the mitochondrial genome of a "living fossil," the coelacanth (Latimeria chalumnae). Genetics. 1997; 146:995–1010. [PubMed: 9215903]
8. Amemiya CT, et al. Complete HOX cluster characterization of the coelacanth provides further evidence for slow evolution of its genome. Proc Natl Acad Sci U S A. 2010; 107:3622–3627. doi: 0914312107 [pii] 10.1073/pnas.0914312107. [PubMed: 20139301]
9. Larsson TA, Larson ET, Larhammar D. Cloning and sequence analysis of the neuropeptide Y receptors Y5 and Y6 in the coelacanth Latimeria chalumnae. Gen Comp Endocrinol. 2007; 150:337–342. doi:S0016-6480(06)00296-6 [pii] 10.1016/j.ygcen.2006.09.002. [PubMed: 17070811]
10. Noonan JP, et al. Coelacanth genome sequence reveals the evolutionary history of vertebrate genes. Genome Res. 2004; 14:2397–2405. doi:gr.2972804 [pii] 10.1101/gr.2972804. [PubMed: 15545497]
11. Gnerre S, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A. 2011; 108:1513–1518. doi:1017351108 [pii] 10.1073/pnas.1017351108. [PubMed: 21187386]
12. Bogart JP, Balon EK, Bruton MN. The chromosomes of the living coelacanth and their remarkable similarity to those of one of the most ancient frogs. J Hered. 1994; 85:322–325. [PubMed: 7930502]
13. Cantarel BL, et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 2008; 18:188–196. doi:gr.6743907 [pii] 10.1101/gr.6743907. [PubMed: 18025269]

14. Grabherr MG, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011; 29:644–652. doi:nbt.1883 [pii] 10.1038/nbt.1883. [PubMed: 21572440]
15. Pallavicini A, et al. Analysis of the transcriptome of the Indonesian coelacanth. Latimeria menadoensis. 2012 submitted.
16. Schultze, HP.; Trueb, L. Origins of the Higher Groups of Tetrapods: Controversy and Consensus. Comstock Pub. Associates; 1991.
17. Meyer A, Dolven SI. Molecules, fossils, and the origin of tetrapods. J Mol Evol. 1992; 35:102–113. [PubMed: 1501250]
18. Brinkmann H, Venkatesh B, Brenner S, Meyer A. Nuclear protein-coding genes support lungfish and not the coelacanth as the closest living relatives of land vertebrates. Proc Natl Acad Sci U S A. 2004; 101:4900–4905. doi:10.1073/pnas.0400609101 0400609101 [pii]. [PubMed: 15037746]
19. Lartillot N, Philippe H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol. 2004; 21:1095–1109. doi:10.1093/molbev/msh112msh112 [pii]. [PubMed: 15014145]
20. Takezaki N, Rzhetsky A, Nei M. Phylogenetic test of the molecular clock and linearized trees. Mol Biol Evol. 1995; 12:823–833. [PubMed: 7476128]
21. Tajima F. Simple methods for testing the molecular evolutionary clock hypothesis. Genetics. 1993; 135:599–607. [PubMed: 8244016]
22. Bejerano G, et al. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature. 2006; 441:87–90. doi:nature04696 [pii] 10.1038/nature04696. [PubMed: 16625209]
23. Voss SR, et al. Origin of amphibian and avian chromosomes by fission, fusion, and retention of ancestral chromosomes. Genome Res. 2011; 21:1306–1312. doi:gr.116491.110 [pii] 10.1101/gr.116491.110. [PubMed: 21482624]
24. Smith JJ, Voss SR. Gene order data from a model amphibian (Ambystoma): new perspectives on vertebrate genome structure and evolution. BMC Genomics. 2006; 7:219. doi:1471-2164-7-219 [pii] 10.1186/1471-2164-7-219. [PubMed: 16939647]
25. Inoue JG, Miya M, Venkatesh B, Nishida M. The mitochondrial genome of Indonesian coelacanth Latimeria menadoensis (Sarcopterygii: Coelacanthiformes) and divergence time estimation between the two coelacanths. Gene. 2005; 349:227–235. doi:S0378-1119(05)00017-X [pii] 10.1016/j.gene.2005.01.008. [PubMed: 15777665]
26. Holder MT, Erdmann MV, Wilcox TP, Caldwell RL, Hillis DM. Two living species of coelacanths? Proc Natl Acad Sci U S A. 1999; 96:12616–12620. [PubMed: 10535971]
27. Canapa A, et al. Composition and Phylogenetic Analysis of Vitellogenin Coding Sequences in the Indonesian Coelacanth Latimeria menadoensis. J Exp Zool B Mol Dev Evol. 2012; 318:404–416. doi:10.1002/jez.b.22455. [PubMed: 22711571]
28. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature. 2005; 437:69–87. doi:nature04072 [pii] 10.1038/nature04072. [PubMed: 16136131]
29. Zhang J, et al. Loss of fish actinotrichia proteins and the fin-to-limb transition. Nature. 2010; 466:234–237. doi:10.1038/nature09137. [PubMed: 20574421]
30. Jovelin R, et al. Evolution of developmental regulation in the vertebrate FgfD subfamily. Journal of experimental zoology. Part B, Molecular and developmental evolution. 2010; 314:33–56. doi:10.1002/jez.b.21307.
31. Braasch I, Postlethwait JH. The teleost agouti-related protein 2 gene is an ohnolog gone missing from the tetrapod genome. Proceedings of the National Academy of Sciences of the United States of America. 2011; 108:E47–E48. doi:10.1073/pnas.1101594108. [PubMed: 21406593]
32. Navratilova P, et al. Systematic human/zebrafish comparative identification of cis-regulatory activity around vertebrate developmental transcription factor genes. Dev Biol. 2009; 327:526–540. doi:S0012-1606(08)01320-1 [pii] 10.1016/j.ydbio.2008.10.044. [PubMed: 19073165]
33. Xie X, et al. Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. Proc Natl Acad Sci U S A. 2007; 104:7145–7150. doi:0701811104 [pii] 10.1073/pnas.0701811104. [PubMed: 17442748]

34. Jones FC, et al. The genomic basis of adaptive evolution in threespine sticklebacks. Nature. 2012; 484:55–61. doi:nature10944 [pii] 10.1038/nature10944. [PubMed: 22481358]
35. Shubin N, Tabin C, Carroll S. Deep homology and the origins of evolutionary novelty. Nature. 2009; 457:818–823. doi:nature07891 [pii] 10.1038/nature07891. [PubMed: 19212399]
36. Montavon T, et al. A regulatory archipelago controls Hox genes transcription in digits. Cell. 2011; 147:1132–1145. doi:S0092-8674(11)01273-6 [pii] 10.1016/j.cell.2011.10.023. [PubMed: 22118467]
37. Wright PA. Nitrogen excretion: three end products, many physiological roles. J Exp Biol. 1995; 198:273–281. [PubMed: 7699310]
38. Kosakovsky Pond SL, et al. A random effects branch-site model for detecting episodic diversifying selection. Mol Biol Evol. 2011; 28:3033–3043. doi:msr125 [pii] 10.1093/molbev/msr125. [PubMed: 21670087]
39. Haberle J, et al. Molecular defects in human carbamoy phosphate synthetase I: mutational spectrum, diagnostic and protein structure considerations. Hum Mutat. 2011; 32:579–589. doi: 10.1002/humu.21406. [PubMed: 21120950]
40. Carroll, RL. Vertebrate Paleontology and Evolution. W.H. Freeman and Company; 1988.
41. Gekas C, et al. Hematopoietic stem cell development in the placenta. Int J Dev Biol. 2010; 54:1089–1098. doi:103070cg [pii] 10.1387/ijdb.103070cg. [PubMed: 20711986]
42. Bejerano G, et al. Ultraconserved elements in the human genome. Science. 2004; 304:1321–1325. doi:10.1126/science.1098119 1098119 [pii]. [PubMed: 15131266]
43. Vista Enhancer Browser. <http://enhancer.lbl.gov/cgi-bin/imagedb3.pl?form=presentation&show=1&experiment_id=501&organism_id=1> (
44. Wellik DM. Hox patterning of the vertebrate axial skeleton. Dev Dyn. 2007; 236:2454–2463. doi: 10.1002/dvdy.21286. [PubMed: 17685480]
45. Scotti M, Kmita M. Recruitment of 5' Hoxa genes in the allantois is essential for proper extra-embryonic function in placental mammals. Development. 2012; 139:731–739. doi:dev.075408 [pii] 10.1242/dev.075408. [PubMed: 22219351]
46. Bengten E, et al. Immunoglobulin isotypes: structure, function, and genetics. Curr Top Microbiol Immunol. 2000; 248:189–219. [PubMed: 10793479]
47. Ota T, Rast JP, Litman GW, Amemiya CT. Lineage-restricted retention of a primitive immunoglobulin heavy chain isotype within the Dipnoi reveals an evolutionary paradox. Proc Natl Acad Sci U S A. 2003; 100:2501–2506. doi:10.1073/pnas.0538029100 0538029100 [pii]. [PubMed: 12606718]
48. Gregory, TR. The Evolution of the Genome. Elsevier Academic Press, Inc.; 2004.
49. Stamatakis A, Ludwig T, Meier H. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics. 2005; 21:456–463. doi:bti191 [pii] 10.1093/bioinformatics/bti191. [PubMed: 15608047]
50. Smith JJ, Sumiyama K, Amemiya CT. A living fossil in the genome of a living fossil: Harbinger transposons in the coelacanth genome. Mol Biol Evol. 2012; 29:985–993. doi:msr267 [pii] 10.1093/molbev/msr267. [PubMed: 22045999]

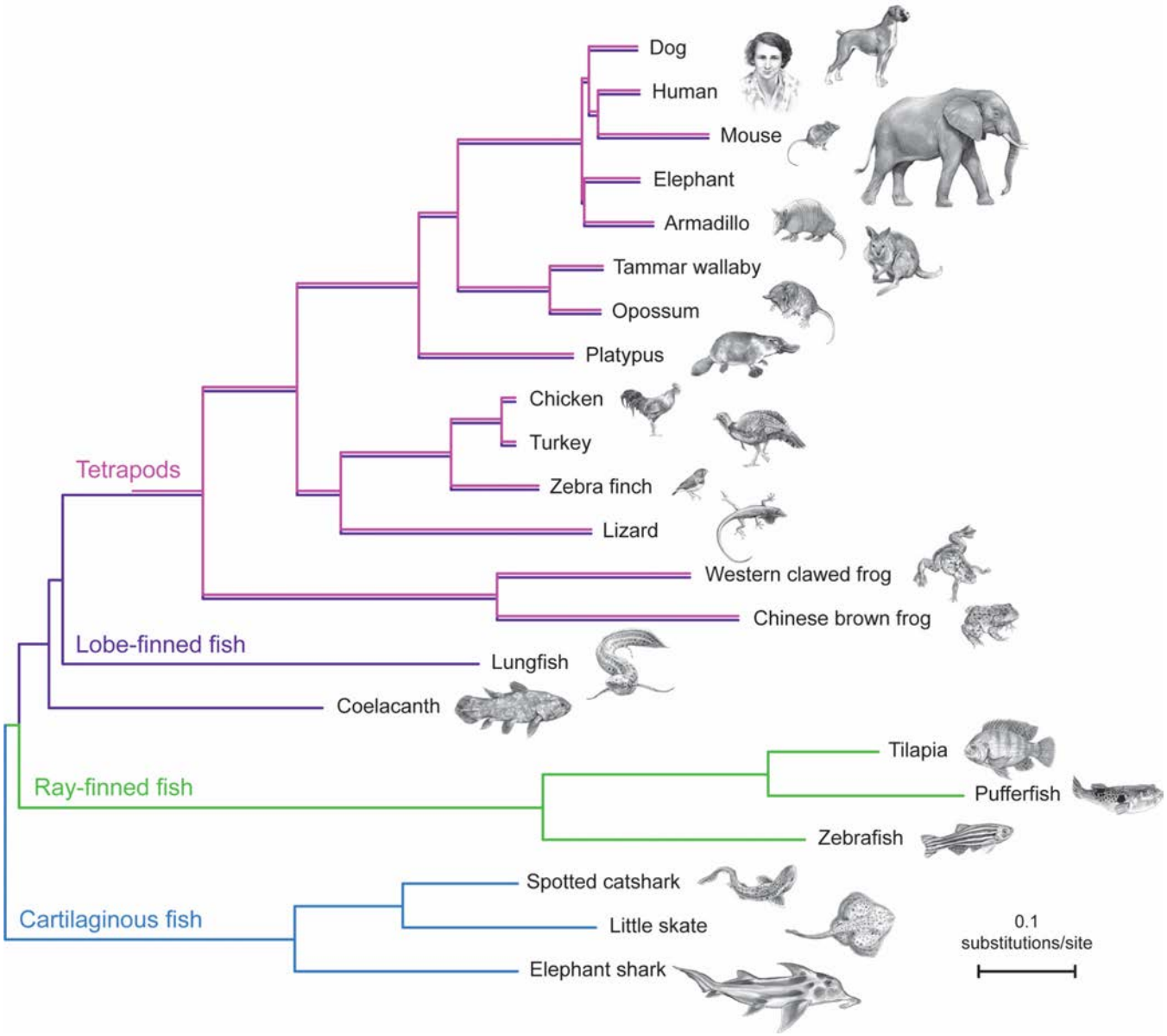


**Figure 1. A phylogenetic tree of a broad selection of jawed vertebrates shows that lungfish, not coelacanth, is the closest relative of tetrapods**

Multiple sequence alignments of 251 genes present as 1-to-1 orthologs in 22 vertebrates and with a full sequence coverage for both lungfish and coelacanth were used to generate a concatenated matrix of 100,583 unambiguously aligned amino acid positions. The Bayesian tree was inferred using PhyloBayes under the CAT+GTR+$\Gamma_4$ model with confidence estimates derived from 100 jackknife tests (1.0 posterior probability)[49]. The tree was rooted on cartilaginous fish. It shows both that lungfish is more closely related to tetrapods than coelacanth and that the protein sequence of coelacanth is slowly evolving.

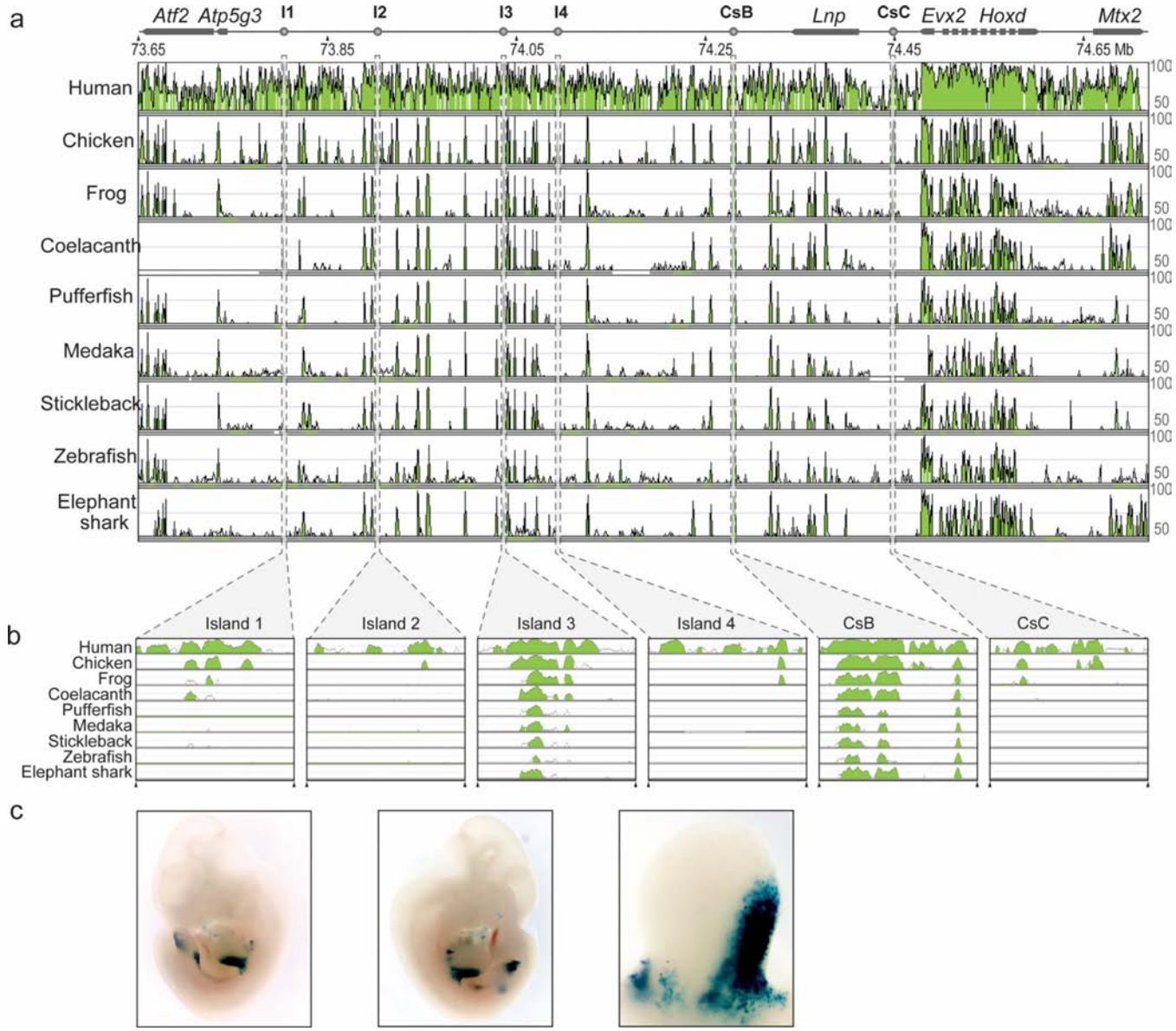


**Figure 2. Alignment of the HOX-D locus and upstream gene desert identifies conserved limb enhancers**

(a) Organization of the mouse HOX-D locus and centromeric gene desert, flanked by the ATF2 and MTX2 genes. Limb regulatory sequences (I1, I2, I3, I4, CsB and CsC) are noted. Using the mouse locus as a reference (NCBI37/mm9 assembly), corresponding sequences from human, chicken, frog, coelacanth, pufferfish, medaka, stickleback, zebrafish and elephant shark were aligned. Alignment shows regions of homology between tetrapod, coelacanth and ray-finned fishes. (b) Alignment of vertebrate cis-regulatory elements I1, I2, I3, I4, CsB and CsC. (c) Expression patterns of coelacanth Island I in a transgenic mouse. Limb buds indicated by arrowheads in the first two panels. The third panel shows a close-up of a limb bud.

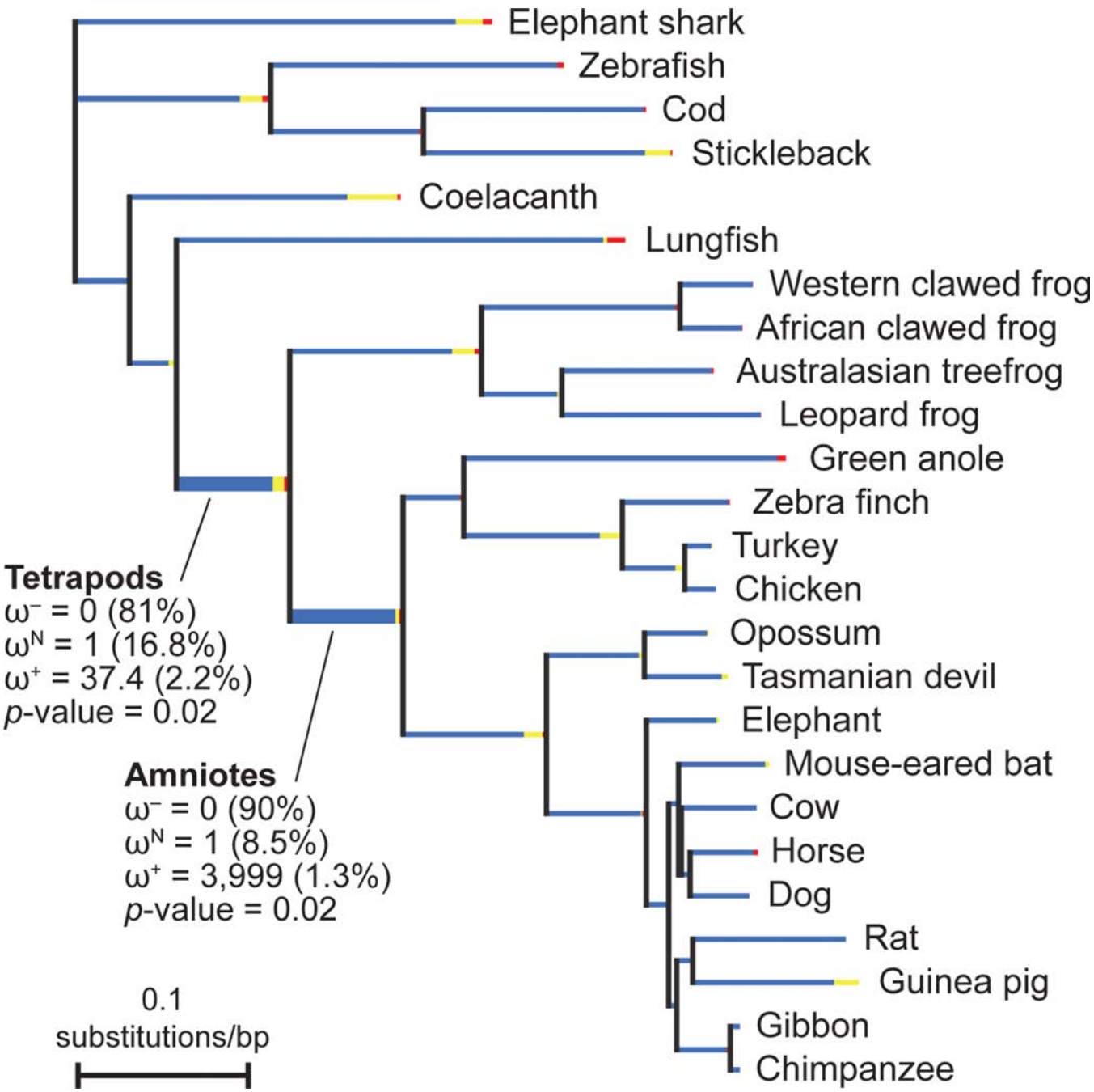


**Figure 3. Phylogeny of *CPS1* coding sequences used to determine positive selection within the urea cycle**

Branch lengths are scaled to the expected number of substitutions/nucleotide and branch color indicates the strength of selection (dN/dS or ω) with red corresponding to positive or diversifying selection (ω > 5), blue to purifying selection (ω = 0), and yellow to neutral evolution (ω = 1). Thick branches indicate statistical support for evolution under episodic diversifying selection. The proportion of each color represents the fraction of the sequence undergoing the corresponding class of selection.

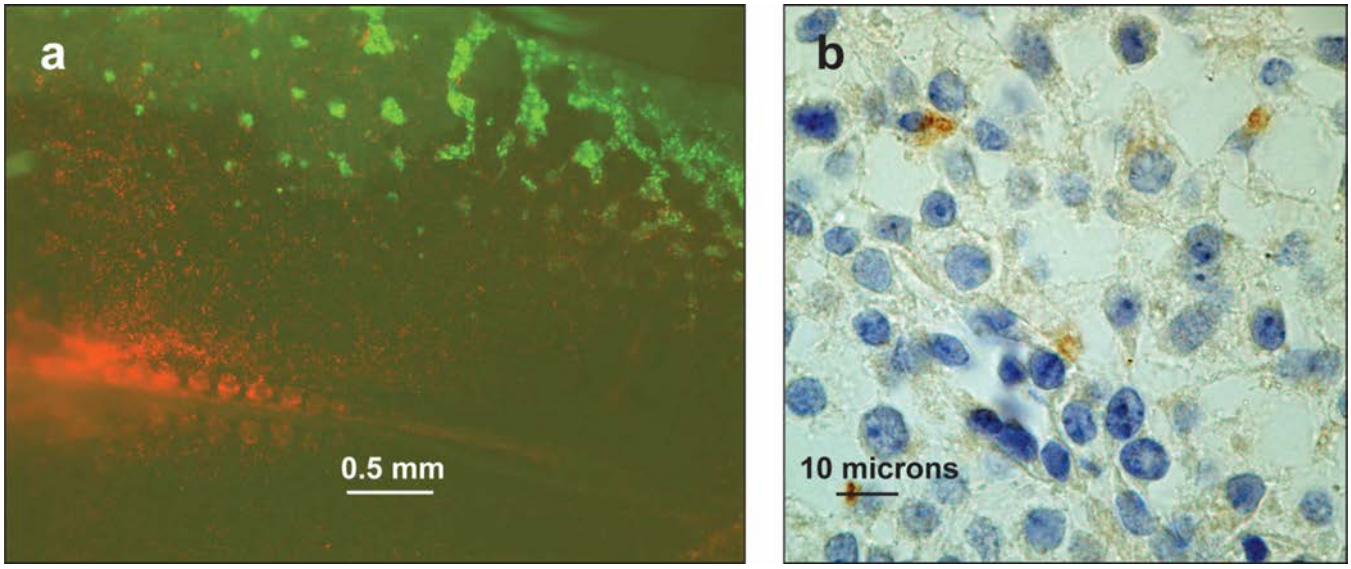


**Figure 4. Transgenic analysis implicates involvement of *Hox* CNE HA14E1 in extraembryonic activities in the chick and mouse**

(A) Chicken HA14E1 drives reporter expression in blood islands in chick embryos. A construct containing chicken HA14E1 upstream of a minimal (TK) promoter driving eGFP was electroporated in HH4 stage chick embryos together with a nuclear mCherry construct. GFP expression was analyzed at stage ~ HH11. The green aggregations and punctate staining are observed in the blood islands and developing vasculature. (B) Expression of *Latimeria Hoxa14* reporter transgene in the developing placental labyrinth of a mouse embryo. A field of cells from the labyrinth region of an E8.5 embryo from a BAC transgenic line containing coelacanth *Hoxa14-Hoxa9* [50] in which the *Hoxa14* gene had been supplanted with the gene for red fluorescence protein (RFP). Immunohistochemistry was used to detect RFP (brown staining in a small number of cells).