The African coelacanth genome provides insights into tetrapod evolution

The discovery of a living coelacanth specimen in 1938 was remarkable, as this lineage of lobe-finned fish was thought to have become 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 show the importance of the coelacanth genome as a blueprint for understanding tetrapod evolution.
Comoros Islands in the Indian Ocean, and only 309 individuals have been recorded in the past 75 years (R. Nulens, personal communication). The discovery in 1997 of a second coelacanth species in Indonesia, Latimeria menadoensis, was equally surprising, as it had been assumed that living coelacanths were confined to small populations off the East African coast. 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 Myr, leading to the supposition that, among vertebrates, this lineage is markedly slow to evolve. Latimeria has also been of particular interest to evolutionary biologists, owing to its hotly debated relationship to our last fish ancestor, the fish that first crawled onto land. In the past 15 years, targeted sequencing efforts have produced the sequences of the coelacanth mitochondrial genomes, HOX clusters and a few gene families. Nevertheless, coelacanth research has felt the lack of large-scale sequencing data. Here we describe the sequencing and comparative analysis of the genome of Latimeria chalumnae, the African coelacanth.

**Genome assembly and annotation**

The African coelacanth genome was sequenced and assembled using DNA from a Comoros Islands Latimeria chalumnae specimen (Supplementary Fig. 1). It was sequenced by Illumina sequencing technology and assembled using the short read genome assembler ALLPATHS-LG. The L. chalumnae genome has been reported previously to have a karyotype of 48 chromosomes. The draft assembly is 2.86 gigabases (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 (the contig size above which 50% of the total length of the sequence assembly can be found) of 12.7 kilobases (kb) and a scaffold N50 size of 924 kb, and quality metrics comparable to other Illumina genomes (Supplementary Note 1, and Supplementary Tables 1 and 2).

The genome assembly was annotated separately by both the Ensembl gene annotation pipeline (Ensembl release 66, February 2012) and by MAKER. The Ensembl gene annotation pipeline created gene models using protein alignments from the Universal Protein Resource (Uniprot) database, limited coelacanth complementary DNA data, RNA-seq data generated from L. chalumnae muscle (18 Gb of paired-end reads were assembled using Trinity software) 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) RNA-seq data to create gene models, and this produced 29,237 protein-coding gene annotations. In addition, 2,894 short non-coding RNAs, 1,214 long non-coding RNAs, and more than 24,000 conserved RNA secondary structures were identified (Supplementary Note 2, Supplementary Tables 3 and 4, Supplementary Data 1–3 and Supplementary Fig. 3). It was inferred that 336 genes underwent specific duplications in the coelacanth lineage (Supplementary Note 3, Supplementary Tables 5 and 6, and Supplementary Data 4).

**The closest living fish relative of tetrapods**

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

To seek a comprehensive answer we generated RNA-seq data from three samples (brain, gonad and kidney, and gut and 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 in which a 1:1 orthology ratio was clear and used CAT-GTR, a complex site-heterogeneous model of sequence evolution that is known to reduce tree-reconstruction artefacts (see Supplementary Methods). The resulting phylogeny, based on 100,583 concatenated amino acid positions (Fig. 1, posterior probability = 1.0 for the lungfish–tetrapod node) is maximally supported except for the relative positions of the armadillo and the 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 and Supplementary Fig. 4).

**The slowly evolving coelacanth**

The morphological resemblance of the modern coelacanth to its fossil ancestors has resulted in it being nicknamed ‘the living fossil’. This invites the question of whether the genome of the coelacanth is as slowly evolving as its outward appearance suggests. Earlier work showed that a few gene families, such as Hox and protocadherins, have comparatively slower protein-coding evolution in coelacanth than in other vertebrate lineages. To address the question, we compared several features of the coelacanth genome to those of other vertebrate genomes.

Protein-coding gene evolution was examined using the phylogenomics data set described above (251 concatenated proteins) (Fig. 1). Pair-wise distances between taxa were calculated from the branch lengths of the tree using the two-cluster test proposed previously 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 the 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 that were tested were significantly more slowly evolving (0.890 substitutions per site) than the lungfish (1.05 substitutions per site), chicken (1.09 substitutions per site) and mammalian (1.21 substitutions per site) orthologues ($P < 10^{-6}$ in all cases) (Supplementary Data 5). In addition, as can be seen in Fig. 1, the substitution rate in coelacanth is approximately half that in tetrapods since the two lineages diverged. A Tajima’s relative rate test$^{21}$ confirmed the coelacanth’s significantly slower rate of protein evolution ($P < 10^{-20}$) (Supplementary Data 6).

We next examined the abundance of transposable elements in the coelacanth genome. Theoretically, transposable elements may make their greatest contribution 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 transposable-element superfamilies and has a relatively high transposable-element content (25%); this number is probably an underestimate as this is a draft assembly (Supplementary Note 5 and Supplementary Tables 7–10). Analysis of RNA-seq data and of the divergence of individual transposable-element copies from consensus sequences show that 14 coelacanth transposable-element superfamilies are currently active (Supplementary Note 6, Supplementary Table 10 and Supplementary Fig. 5). We conclude that the current coelacanth genome shows both an abundance and activity of transposable elements 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 fission events published previously$^{23}$ 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 per 1 Myr), compared to 20 events (0.054 fusions per 1 Myr) in the salmonander lineage and 21 events (0.057 fusions per 1 Myr) in the Xenopus lineage$^{23}$ (Supplementary Note 7 and Supplementary Fig. 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 Myr$^{25,26}$. When we compared the liver and testis transcriptomes of *L. menadoensis* to the *L. chalumnae* genome, we found an identity of 99.73% (Supplementary Note 8 and Supplementary Fig. 7), whereas alignments between 20 sequenced *L. menadoensis* bacterial artificial chromosomes (BACs) and the *L. chalumnae* genome showed an identity of 98.7% (Supplementary Table 11 and Supplementary Fig. 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 of 6 to 8 Myr ago)$^{28}$, whereas the rates of molecular evolution in *Latimeria* are probably affected by several factors, including the slower substitution rate seen in coelacanth. This suggests a slightly longer divergence time for the two coelacanth species.

### The adaptation of vertebrates to land

As the species with a 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 the land environment.

Over the 400 Myr that vertebrates have lived on land, some genes that are unnecessary for existence in their new environment 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 the coelacanth) but that are missing from tetrapod genomes. More than 50 such genes, including components of fibroblast growth factor (FGF) signalling, TGF-β and bone morphogenic protein (BMP) signalling, and WNT signalling pathways, as well as many transcription factors, were inferred to be lost based on the coelacanth data (Supplementary Data 7 and Supplementary Fig. 9). Previous studies of genes that were 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 that were shown previously to be absent in tetrapods (*And1* and *And2* (ref. 29), *Fgf24* (ref. 30) and *Asip2* (ref. 31)), were indeed present and intact in *Latimeria*, supporting the idea that they were lost in the tetrapod lineage.

We functionally annotated more than 50 genes lost in tetrapods using zebrafish data (gene expression, knock-downs and knockouts). Many genes were classified in important developmental categories (Supplementary Data 7): fin development (13 genes); otolith and ear development (8 genes); kidney development (7 genes); trunk, somite and 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 (for example, fin-to-limb transition and remodelling of the ear) 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 seem that the protein-coding portion of this gene family, along with several others (Supplementary Note 9, Supplementary Tables 12–16 and Supplementary Fig. 10), have remained largely conserved during the vertebrate land transition (Supplementary Fig. 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. They can act as promoters, enhancers, repressors and insulators$^{52,33}$, 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 (that is, 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, untranslated regions (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 (chromatin immunoprecipitation followed by sequencing) data obtained using antibodies against p300, a transcriptional coactivator. This resulted in a sevenfold 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 gene-ontology category enrichment was calculated for those genes. The most enriched categories were involved with smell perception (for example, sensory perception of smell, detection of chemical stimulus and olfactory receptor activity). 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 odors 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 and Supplementary Tables 17–24).

A major innovation of tetrapods is the evolution of limbs characterized by digits. The limb skeleton consists of a stylopod (humerus or femur), the zeugopod (radius and ulna, or tibia and fibula), and an autopod (wrist or ankle, and digits). There are two major hypotheses about the origins of the autopod; that it was a novel feature of tetrapods, and that it has antecedents in the fins of fish (Supplementary Note 11 and Supplementary Fig. 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 that is proximal to the HOX-D cluster. Here we provide an alignment of the HOX-D centromeric gene desert of coelacanth with those of tetrapods and ray-finned fishes (Fig. 2a). Among the six cis-regulatory sequences previously identified in this gene desert, three sequences show sequence conservation restricted to tetrapods (Supplementary Fig. 13). However, one regulatory sequence (island 1) is shared by tetrapods and coelacanth, but not by ray-finned fish (Fig. 2b and Supplementary Fig. 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 (Fig. 2c). This suggests 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 Fig. 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.

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, which estimates the ratio of synonymous (ds) to non-synonymous (dn) substitutions (o 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 (Fig. 3); no other enzymes in this cycle showed a signature of selection. Conversely, mitochondrial arginase (ARG2), which produces extrarenal urea as a byproduct of arginine metabolism but is not involved in the production of urea for nitrogenous waste disposal, did not show any evidence of selection in vertebrates (Supplementary Fig. 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 (the two ATP-binding sites of the five amino acids of CPS1 that changed between coelacanth and tetrapods, three are in important domains (the two ATP-binding sites and the subunit interaction domain) and a fourth is known to cause a malfunctioning enzyme in human patients if mutated.

The adaptation to a terrestrial lifestyle necessitated major changes in the physiological environment of the developing embryo and fetus, resulting in the evolution and specialization of extra-embryonic membranes of the amniote mammals. In particular, the placenta is a complex structure that is critical for providing gas and nutrient exchange between mother and fetus, and is also a major site of haematopoiesis.

We have identified a region of the coelacanth HOX-A cluster that may have been involved in the evolution of extra-embryonic structures in tetrapods, including the eutherian placenta. Global alignment of the coelacanth Hoxa14–Hoxa13 region with the homologous regions of the horn shark, chicken, human and mouse revealed a CNE just upstream of the coelacanth Hoxa14 gene (Supplementary Fig. 17a). 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 chicken, human and mouse 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 more than 99% identical between mouse, human and all other sequenced mammals, and would therefore be considered to be an ultra-conserved element. 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 Fig. 17bc). Expression of human HA14E1 in a mouse...
The coelacanth lacks immunoglobulin-M

Immunoglobulin-M (IgM), a class of antibodies, has been reported in all vertebrate species that have been characterized so far, and is considered to be indispensable for adaptive immunity. 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 and Supplementary Fig. 22). Further study of these changes between tetrapods and the coelacanth may provide important insights into how a complex organism like a vertebrate can markedly change its way of life.

**Methods Summary**

A full description of methods, including information on sample collection, sequencing, assembly, annotation, all sequence analysis and functional validation, can be found in the Supplementary Information.

Received 5 September 2012; accepted 20 February 2013.

1. Smith, J. L. B. A living fish of mesozoic type. Nature 143, 455–456 (1939).
2. Nuijens, R., Scott, L. & Herbin, M. An Updated Inventory of All Known Specimens of the Coelacanth, Latimeria Spp. Smithiana Vol. 3 (South African Institute for Aquatic Biodiversity, 2010).
3. Erdmann, M. V., Calder, R. L. & Kasim Moosa, M. Indonesian 'king of the sea' discovered. Nature 395, 335 (1998).
4. Smith, J. L. B. 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. Nature Commun. 3, 772 (2012).
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. and Meyer, A. Complete DNA sequence of the mitochondrial genome of a "living fossil," the coelacanth (Latimeria chalumnae). Genes 146, 995–1010 (1997).
8. Amemiya, C. T. et al. Complete Hox cluster characterization of the coelacanth provides further evidence for slow evolution of its genome. Proc. Natl Acad. Sci. USA 107, 3622–3627 (2010).
9. Larsson, T. A., Larson, E. T. & Larhammar, D. Cloning and sequence analysis of the novel retroposon. J. Exp. Zool. 287, 227–237 (2005).
10. Tajima, F. Simple methods for testing the molecular evolutionary clock hypothesis. Mol. Biol. Evol. 21, 158–161 (2004).
11. Gnerre, S. et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl Acad. Sci. USA 108, 1513–1518 (2011).
12. Bogart, J. P., Balon, E. K. & Bruton, M. N. The chromosomes of the living coelacanth and their remarkable similarity to those of one of the most ancient frogs, J. Hered. 85, 322–325 (1994).
13. Cantarel, B. L. et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18, 188–196 (2008).
14. Scotti, M. & Kmita, M. Recruitment of 5 cis-regulatory elements to control the expression of a novel coelacanth Hoxa gene in the allantois is essential for proper extra-embryonic function in placental mammals. Development 137, 729–730 (2010).
15. Scotti, M. & Kmita, M. Recruitment of 5’ Hoxa genes in the allantochord is essential for proper extra-embryonic function in placental mammals. Development 135, 327–338 (2008).
16. Smith, J. J., Sumiyama, K. & Amemiya, C. T. Lineage-restricted retention of a primitive immunoglobulin heavy chain isotype within the Dnjo reveals an evolutionary paradox. Proc. Natl Acad. Sci. USA 100, 2501–2506 (2003).
17. Gregory, T. R. The Evolution of the Genome 1–71 (Elsevier Academic, 2004).
18. Scotti, M., Atkins, J. E. & Kmita, M. A fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21, 465–463 (2005).
19. Smith, J. J., Sumiyama, K. & Amemiya, C. T. A living fossil in the genome of a living fish: Harbinger transposons in the coelacanth genome. Mol. Biol. Evol. 29, 985–993 (2012).

Supplementary information is available in the online version of this paper.

Acknowledgements Acquisition and storage of Latimeria chalumnae samples was supported by grants from the Natural History Museum, the African Coelacanth Ecosystem Programme of the South African National Department of Science and Technology, Generation of the Latimeria chalumnae and Proteus angilis sequences were performed at the Broad Institute of the Massachusetts Institute of Technology (MIT) and Harvard University supported by grants from the National Human Genome Research Institute (NHGRI). CLT is the recipient of a EURYI award from the European Science Foundation. 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. annectans transcriptomes. S. Hamada, R. Stobbs and the Association for the Protection of Gombessa (APG) for their help in obtaining coelacanth samples. The authors would like to thank T. Ma. and A.M. for performing the gene relative-rate analysis. A.C., N.R.S. and C.T.A. prepared RNA from Rana cinnamomea and L. Gaffney, C. Hamilton and J. Westlund for assistance with figure preparation.

Author Contributions J.A., C.T.A., A.M. and K.L.T. planned and oversaw the project. R.A.D. and C.T.A. provided blood and tissues for sequencing. C.T.A. and M.L. prepared the DNA for sequencing. I.M., S.G., D.P., F.J.R., T.S. and D.B.J. assembled the genome. N.R.S. and C.T.A. prepared RNA from L. chalumnae and P. annectans, and J.L. and M.L. prepared RNA from L. menadoensis RNA-seq data. A.M. and K.L.T. planned and oversaw the project.

The complete DNA sequence of the mitochondrial genome of a "living fossil," the coelacanth (Latimeria chalumnae). Genes 146, 995–1010 (1997).

The complete DNA sequence of the mitochondrial genome of a "living fossil," the coelacanth (Latimeria chalumnae). Genes 146, 995–1010 (1997).