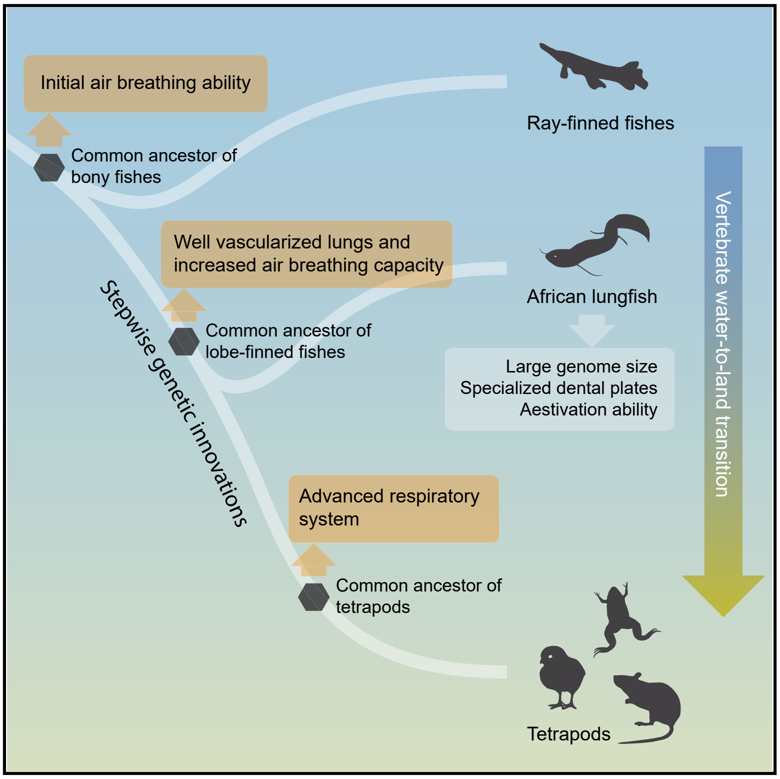
Cell

Resource

African lungfish genome sheds light on the vertebrate water-to-land transition

Graphical abstract


Initial air breathing ability
Common ancestor of bony fishes
Ray-finned fishes
Well vascularized lungs and increased air breathing capacity
Common ancestor of lobe-finned fishes
African lungfish
Large genome size
Specialized dental plates
Aestivation ability
Stepwise genetic innovations
Vertebrate water-to-land transition
Advanced respiratory system
Common ancestor of tetrapods
Tetrapods

Highlights

- Chromosome-level assembly of the largest vertebrate genome reported to date
- Continuous expansion of transposons contributed to the huge lungfish genome
- Genetic changes enhanced respiration, locomotion, and anxiolytic ability
- Three genetic innovation steps from bony fishes to lungfishes and then tetrapods

Authors

Kun Wang, Jun Wang, Chenglong Zhu, ..., Qiang Qiu, Shunping He, Wen Wang

Correspondence

zhaowm@big.ac.cn (W.Z.),
qiuqiang@lzu.edu.cn (Q.Q.),
clad@ihb.ac.cn (S.H.),
wwang@mail.kiz.ac.cn (W.W.)

In Brief

Genome assembly and analysis of the massive African lungfish genome has shed light on genetic innovations required for the water-to-land transition

Wang et al., 2021, Cell *184*, 1362–1376
March 4, 2021 © 2021 Elsevier Inc.
https://doi.org/10.1016/j.cell.2021.01.047

CellPress

CellPress

Cell

Resource

African lungfish genome sheds light on the vertebrate water-to-land transition

Kun Wang,[1,17] Jun Wang,[2,3,17] Chenglong Zhu,[1,4,17] Liandong Yang,[5,17] Yandong Ren,[1,4,17] Jue Ruan,[6,17] Guangyi Fan,[7,8,17] Jiang Hu,[9,17] Wenjie Xu,[1] Xupeng Bi,[8] Youan Zhu,[10] Yue Song,[7] Huatao Chen,[11] Tiantian Ma,[11] Ruoping Zhao,[4] Haifeng Jiang,[5] Bin Zhang,[12] Chenguang Feng,[1] Yuan Yuan,[1] Xiaoni Gan,[5] Yongxin Li,[1] Honghui Zeng,[5] Qun Liu,[7] Yaolei Zhang,[7] Feng Shao,[13] Shijie Hao,[7] He Zhang,[7] Xun Xu,[8] Xin Liu,[7] Depeng Wang,[9] Min Zhu,[10] Guojie Zhang,[8,4,14,15] Wenming Zhao,[12,*] Qiang Qiu,[1,*] Shunping He,[5,14,16,*] and Wen Wang[4,14,18,19,*]

[1]School of Ecology and Environment, Northwestern Polytechnical University, Xi'an 710072, China

[2]Joint Laboratory of Guangdong Province and Hong Kong Region on Marine Bioresource Conservation and Exploitation, College of Marine Sciences, South China Agricultural University, Guangzhou 510642, China

[3]School of Civil Engineering, Architecture and Environment, Hubei University of Technology, Wuhan 430068, China

[4]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China

[5]State Key Laboratory of Freshwater Ecology and Biotechnology, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan 430072, China

[6]Guangdong Laboratory for Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518120, China

[7]BGI-Qingdao, Qingdao 266555, China

[8]BGI-Shenzhen, Shenzhen 518083, China

[9]Grandomics Biosciences, Beijing 102200, China

[10]Institute of Vertebrate Paleontology and Paleoanthropology, China Academy of Sciences, Beijing 100044, China

[11]Key Laboratory of Animal Biotechnology of the Ministry of Agriculture, Department of Clinical Veterinary Medicine, College of Veterinary Medicine, Northwest A&F University, Yangling 712100, China

[12]Beijing Institute of Genomics, Chinese Academy of Sciences, and China National Center for Bioinformation, Beijing 100101, China

[13]Key Laboratory of Freshwater Fish Reproduction and Development, School of Life Sciences, Southwest University, Chongqing 400715, China

[14]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China

[15]Villum Center for Biodiversity Genomics, Section for Ecology and Evolution, Department of Biology, University of Copenhagen, Copenhagen, Denmark

[16]Institute of Deep-sea Science and Engineering, Chinese Academy of Sciences, Sanya 572000, China

[17]These authors contributed equally

[18]Present address: Northwestern Polytechnical University, Xi'an 710072, China

[19]Lead contact

*Correspondence: zhaowm@big.ac.cn (W.Z.), qiuqiang@lzu.edu.cn (Q.Q.), clad@ihb.ac.cn (S.H.), wwang@mail.kiz.ac.cn (W.W.)

https://doi.org/10.1016/j.cell.2021.01.047

SUMMARY

Lungfishes are the closest extant relatives of tetrapods and preserve ancestral traits linked with the water-to-land transition. However, their huge genome sizes have hindered understanding of this key transition in evolution. Here, we report a 40-Gb chromosome-level assembly of the African lungfish (*Protopterus annectens*) genome, which is the largest genome assembly ever reported and has a contig and chromosome N50 of 1.60 Mb and 2.81 Gb, respectively. The large size of the lungfish genome is due mainly to retrotransposons. Genes with ultra-long length show similar expression levels to other genes, indicating that lungfishes have evolved high transcription efficacy to keep gene expression balanced. Together with transcriptome and experimental data, we identified potential genes and regulatory elements related to such terrestrial adaptation traits as pulmonary surfactant, anxiolytic ability, pentadactyl limbs, and pharyngeal remodeling. Our results provide insights and key resources for understanding the evolutionary pathway leading from fishes to humans.

INTRODUCTION

The continuous increase in the oxygen content of the Earth's atmosphere during the Paleozoic era created conditions for the emergence of terrestrial animals (Hsia et al., 2013). After the ancestor of bony fishes first showed the ability to respire air (Bi et al., 2021; Liem, 1988), the ancestors of tetrapods successfully moved onto land. The water-to-land transition of vertebrates

Table 1. Assembly statistics and BUSCO assessment of the African lungfish genome

Genome assembly	Number of sequences	Total length (bp)	N50 (bp)	N90 (bp)	Longest (bp)	Percent of gaps
Contigs	74,217	39,061,217,646	1,603,990	329,778	17,525,427	-
Chromosomes	17	39,928,011,021	2,813,524,174	1,371,727,335	5,260,428,656	2.49%
Unplaced	12,640	126,313,591	10,894	5,905	346,105	0
Total	12,657	40,054,324,612	2,813,524,174	1,371,727,335	5,260,428,656	2.48%

BUSCO assessment	Total	Complete	Complete and single-copy	Complete and duplicated	Fragmented	Missing
tetrapoda_odb9	3,950	3,633 (92.0%)	3,514 (89.0%)	119 (3.0%)	154 (3.9%)	163 (4.1%)
vertebrata_odb9	2,586	2,468 (95.4%)	2,411 (93.2%)	57 (2.2%)	61 (2.4%)	57 (2.2%)

required the evolution of a series of body innovations (Ashley-Ross et al., 2013). The respiratory, sensory, locomotory, circulatory, and other systems had to be remodeled for terrestrial adaptation (Long and Gordon, 2004). In recent decades, paleontological studies have led to a progressive clarification of the process by which vertebrates emerged onto land (Clack, 2012). As an important complementary approach, a study of comparative genomics in tetrapods and their living sister lineages would provide pivotal perspectives to reveal the transition process and underlying molecular mechanisms. Tetrapods are nested within the tetrapodmorphs, one of the three living sarcopterygian (lobe-finned fish) lineages that also include coelacanths and lungfishes (Lu et al., 2012). Lungfishes, as the closest living relative of tetrapods, highlight ancestral state of the lobe-finned fishes that have remained in the water (Amemiya et al., 2013), representing a bridge to understanding the genetic basis and evolutionary process of these transitions. However, the large size of lungfishes' genomes, ranging from 40–130 gigabases (Gb) (Metcalfe et al., 2012), has posed a huge challenge in relevant studies. In this study, we successfully obtained a 40-Gb chromosome-level genome assembly with a high level of completeness and continuity for the African lungfish (*Protopterus annectens*) and conducted systematic analysis, from genomic sequences, expression profiles spanning most tissues, to experimental validation of some genes and regulatory elements, in order to explore the evolutionary genetic process from water to land ~420 million years ago (Ma) (Zhu et al., 2009).

RESULTS

Chromosome-level assembly

From Kmer analysis with short reads (Figure S1A) and flow cytometry (Figure S1B), the size of the African lungfish genome is estimated to be around 40 Gb. We sequenced long reads (~50× coverage, N50 read length 28.48Kb) using Oxford nanopore sequencing technology (ONT) (Figure S1C; Table S1). Combined with short-reads sequencing, optical maps, and Hi-C technology, we got a chromosome assembly with 17 chromosomes (Figure S1D), ranging from 862 Mb to 5.28 Gb, and 12,640 small unplaced scaffolds (totally, 126 Mb). The final length of the assembly is 40.05 Gb, with an N50 of 2.8 Gb (Table 1). The number and length of the assembled chromosomes are consistent with the lungfish karyotype (Figure S1E) (Suzuki and Yamanaka, 1988). More than 97.3% and 95.2% of the genomic regions could be covered 20-fold by long and short reads, respectively (Figure S1F). In addition, for the short pair reads, 96.6% of them could be properly paired, indicating a high level of continuity of the assembly. A total of 178 ultra-conserved elements (UCEs) across bony fishes were then used to assess the completeness of the assembly, and 171 (96.07%) of them could be aligned to the assembly. By comparison, 169 and 158 could be aligned to the coelacanth (Amemiya et al., 2013) and axolotl (Nowoshilow et al., 2018) genomes.

We further sequenced 120 Gb of PacBio full-length cDNA data and 227 Gb of short-reads RNA sequencing (RNA-seq) data from 14 samples of two individuals and assembled them into transcript datasets, including 150 kilo and 4.1 million transcripts from full-length and short reads RNA-seq, respectively (Table S1). Based on the transcript dataset and homologous proteins of vertebrates, we identified 19,457 protein-coding genes in the genome, containing 95.4% of complete conserved orthologs within vertebrates and 92.0% of complete conserved orthologs within tetrapods, based on BUSCO analysis (Table 1) (Seppey et al., 2019). The similar numbers of genes and *Ks* distribution between the lungfish and coelacanth/axolotl/frog (Figure S1G; Table S2) suggest that there has been no recent whole genome duplication (WGD) other than the shared two rounds of WGD from the ancestral vertebrates, and the expansion of transposable elements (TEs) is likely to have been the dominant force resulting in the large genome size in the African lungfish, as we shall show below. We then identified 11,837 1:1 orthologous genes between the lungfish and western clawed frog (Hellsten et al., 2010), which has high quality of chromosome-level genome assembly to examine the chromosomal relationship between the two species. A synteny plot reveals that the homologous chromosomal segments of the two species can be clearly identified (Figure 1A).

The phylogenetic relationship reconstructed from 5,149 1:1 orthologous genes in eight vertebrate species confirmed that the lungfishes are the closest sister lineage to tetrapods (Figure 1B). Although the bootstrap value of the concatenated gene trees receives 100% support, there exists extensive heterogeneity in the topologies of single gene trees (Figure S1H), indicating some extent of incomplete lineage sorting during early divergence of lobe-finned animals. The time of divergence of lungfishes and tetrapods is estimated to be 419 Ma (Figure 1B), about the beginning of the Devonian period. Through inference of recent demographic history, we found that the population of

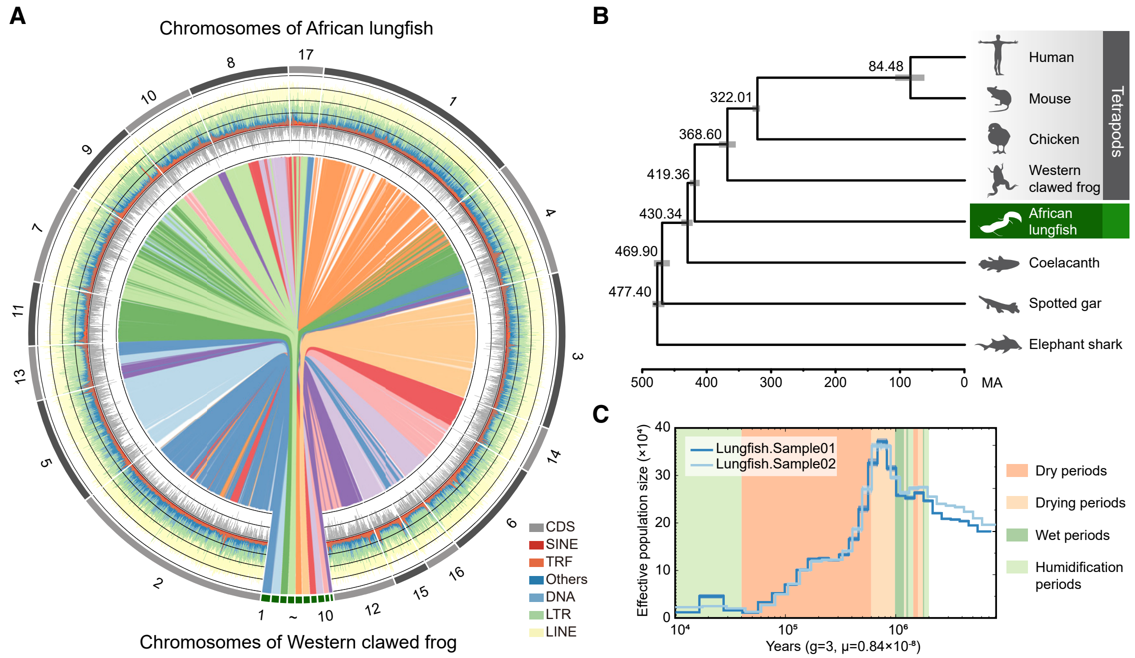


Figure 1. Chromosome-level genome assembly and evolutionary history of the African lungfish

(A) Synteny alignment of the African lungfish (*P. annectens*) and the western clawed frog. The numbers 1 to 17 refer to the chromosomes of the sequenced African lungfish and the bottom dark green circle refers to the chromosomes of western clawed frog. The densities of coding sequences (ranging between 0% to 0.3%) and different types of repeat sequences (with a range from 0% to 80%) in African lungfish are shown in the outer rims with a window size of 10 Mb.

(B) Phylogenetic relationships of the African lungfish and tetrapods. The numbers labeled on the tree refer to the estimated divergence time, and the gray rectangle on each node shows the 95% confidence interval.

(C) Recent population demographic history of the African lungfish and hydrological changes in the Limpopo catchment. The dark and light blue lines refer to the effective population sizes of two individuals, and the orange or green background refers to dry or wet periods of Limpopo catchment.

See also Figure S1.

this African lungfish experienced a rapid decline around 1 Ma (Figure 1C), consistent with a previous study that reported a long-term period of aridification in the Limpopo catchment of South Africa, which is an important habitat for the African lungfish, in the past 1 Ma (Caley et al., 2018). We also re-sequenced another lungfish individual and validated the demographic dynamics result.

Genome expansion

The expansion of the lungfish genome size was caused mainly by proliferation of TEs. A total of 61.7% of the lungfish genome was annotated as repeated sequences, representing 24.7 Gb (Figure 2A). Around 15 Gb of the entire genome was annotated as neither functional nor repeat sequences. These regions are most likely to represent an ancient burst of transposition followed by a long period of degeneration, resulting in very large numbers of unique sequences, a so-called "cemetery of TEs" (Sirijovski et al., 2005). The most abundant TE types are long interspersed nuclear elements (LINEs, 11.5 Gb), long terminal repeats (LTRs, 7.7 Gb), and DNA transposons (3.7 Gb). Among them, LINE/CR1 (7.8 Gb) and LTR/DIRS (6.1 Gb) are the two subclasses forming the highest proportion. We then estimated historical TE expansion activity by analyzing the Kimura distance (Chalopin et al., 2015), and the results suggest that TEs, especially retrotransposons, have been active within the last 70 million years (Myr) (Figure S2A). The recent activity of these TEs accounts for 5.9 Gb, with a peak expansion rate of 433 Mb per Myr, indicating that the genome size of the African lungfish has gradually increased over the past hundred Myr, consistent with previous work showing gradual increase in cell size over the Phanerozoic (Thomson, 1972).

The constant insertion of TEs may have impact on genome content and gene regulation. The mean and median intron lengths are very long in the African lungfish, and almost 16 Gb of the genome are intronic regions. The longest gene in the lungfish is 18 Mb, much longer than the ones in axolotl (6.7 Mb) and human (2.5 Mb). There are only 91 genes longer than 1 Mb in the human genome, but more than 5,000 in the lungfish. In contrast, the mean exon number and exon length are similar in the lungfish and other vertebrates, and species-specific gene family expansion in the African lungfish is at a similar level to other vertebrates (Table S2). By inspecting transcriptome data, we found that there is no obvious correlation between gene expression level and gene length; even genes with length greater than 1 Mb exhibit similar expression levels to those of other shorter genes (Figure S2B). We also found that changes in gene length did not have any great impact on gene expression in the African lungfish when compared to other species (Figure S2C). These results indicate that transcription efficacy for extra-long genes might have improved in the lungfish in order to keep gene expression balanced.

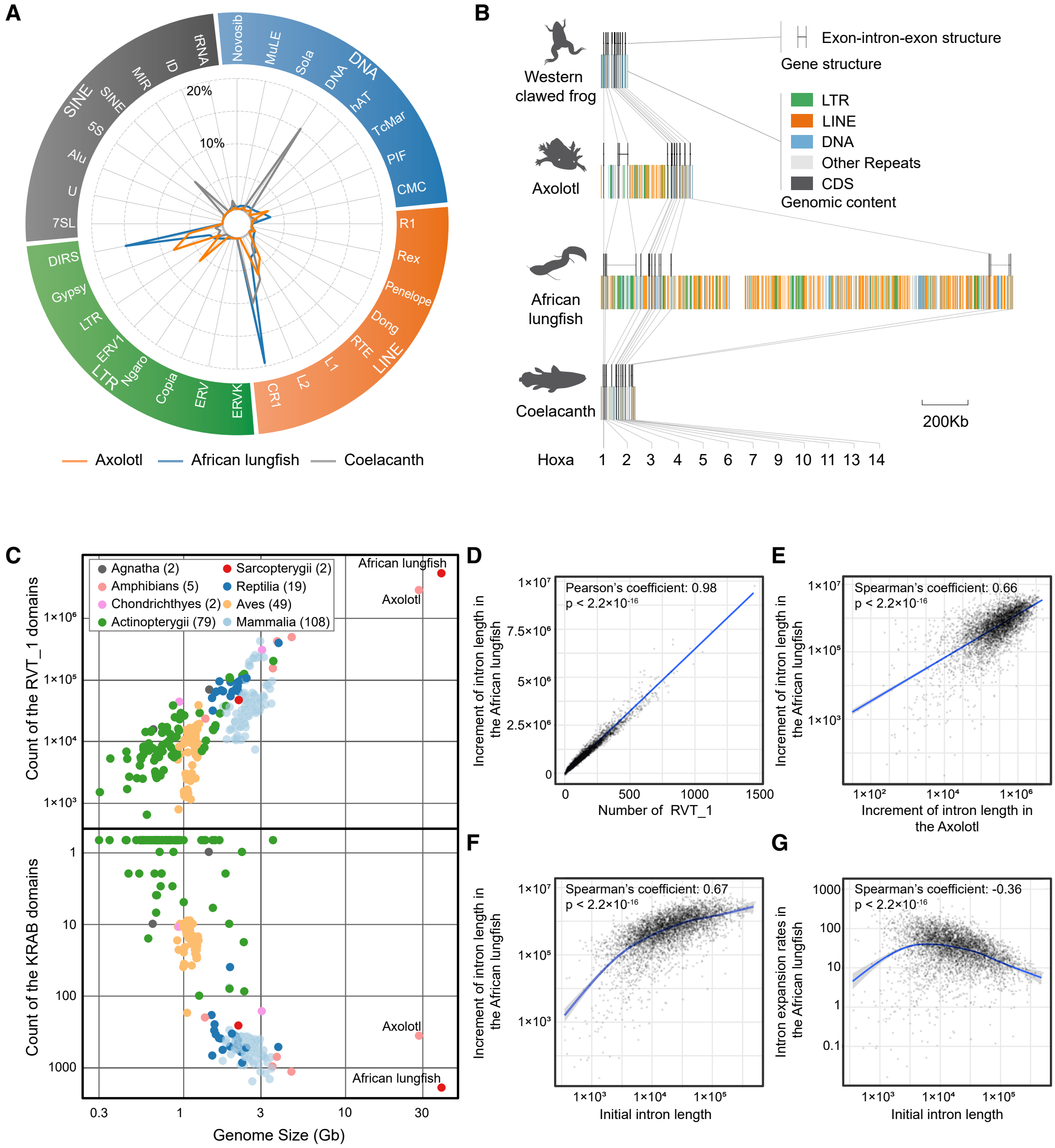


Figure 2. Independent genome expansion and enhanced TE repression in the African lungfish

(A) The African lungfish and axolotl show great differences in the composition of repetitive sequences. The coordinates represent the ratio of each category to the total length of the corresponding genome assembly.

(B) The lengths of gene intervals and introns in the *Hoxa* cluster increase independently in different species.

(C) The numbers of the domain RVT_1 and KRAB are positively correlated with genome sizes. The number at the right of the lineage name refers to the number of species in this lineage.

(D) The relationship between the increment of intron length and the number of the domain RVT_1 in the African lungfish.

(E) The increments of intron sizes are positively correlated in the African lungfish and the axolotl.

(F) The increments of intron sizes of the African lungfish are positively correlated with initial intron length.

(G) The expansion rate of intron sizes of the African lungfish is negatively correlated with initial intron length.

See also Figure S2.

The TE content of the African lungfish is very different from the case of the axolotl assembly, whose genome size is 32 Gb (Figure 2A). Consistently, their domain contents are also different (Figure S2D). When we take a closer look at the distances between different genes of the *Hoxa* gene cluster, it is clear that intergenic/intronic elongation occurred independently in the

African lungfish and the axolotl because they were caused by proliferation of different types of TEs (Figure 2B). These observations are consistent with previous comparative phylogenetic analysis showing that their genome sizes were increased independently (Organ et al., 2016). Despite their independent genome expansion, both the African lungfish and the axolotl have an enormous number of the domain RVT_1 (Figure 2C), encoding reverse transcriptase, which is always carried by retrotransposons. The number of RVT-1 has a positive correlation with the genome size in 266 vertebrates (Table S3), with more than 5.2 million in the African lungfish and only 61,631 in the mouse, implying that the retrotransposon is a main driver of genome amplification. These RVT_1 domains were found to be broadly distributed throughout the African lungfish genome (Figure S2E), implying that the accumulation process had continued for a long time.

We then investigated genome expansion across species by inspecting the intron length of orthologous genes. First, we observed that the increment of intron length of lungfish genes is linearly correlated with the number of RVT_1 domain (Figure 2D), reflecting the important role of retrotransposons in the genome expansion. Second, we observed that the increment of intron length in the African lungfish is positively correlated with that of the orthologous genes of axolotl (Figure 2E), indicating that longer introns have accumulated more TEs. Consistently, we observed a positive correlation between the increment of intron length and the initial intron length (Figure 2F). However, the genes with longer initial intron length tend to have smaller expansion rate (Figure 2G), although their absolute increments are larger (Figure 2F), indicating the selection against extreme extension of gene length.

Previous studies have proposed that the red queen hypothesis may explain the relationship between TE activity and repression strategies (Bruno et al., 2019; McLaughlin and Malik, 2017; Rogers et al., 2018). We detected 16,826 domains in 40 vertebrate species based on PFAM-A database (El-Gebali et al., 2019). Among them, only 54 domains have significant positive correlation between the domain counts and genome sizes (Table S4). Twelve of the 54 domains have count larger than 1,000 in the African lungfish. Seven of the 12 domains are associated with TEs, three are zinc-finger protein domains, one is the Filament domain, and one is the KRAB domain. Interestingly, the KRAB domain and one of the zinc-finger protein domains (zf-C2H2) together form the Kruppel-associated box zinc-finger proteins, which had been proved to play an important role in the silencing of transposable elements in embryonic stem cells (Imbeault et al., 2017). Moreover, the lungfish has the highest number of KRAB domains in 266 vertebrates analyzed (Figure 2C; Table S3), and the genes containing KRAB domains (KZFPs) also tend to be highly expressed in the ovary of the lungfish (Figure S2F). This is largely different from the expanded gene families containing RVT_1 that tend to have a low expression level in all tissues including ovary (Figure S2F). This suggests that the red queen hypothesis might hold true for the case of lungfishes because more KRAB domain and high expression of KRAB-domain-containing genes have evolved in the lungfish, possibly to repress TE activity. Taken together, the above results indicate that the African lungfish genome has a sloppy efficacy of dismissing junk DNAs while it has evolved higher TE repression and long gene transcription ability.

Ancestral karyotype of bony fishes

By using chromosome information from six species, including white-spotted bamboo shark (Zhang et al., 2019), bichir (Bi et al., 2021), spotted gar (Braasch et al., 2016), African lungfish (*P. annectens*), western clawed frog (Hellsten et al., 2010), and chicken (Warren et al., 2017), representing major lineages of vertebrates, a total of 32 proto-chromosomes were reconstructed (Figure 3A; Table S5) for the last common ancestor (LCA) of bony fishes, which is consistent with a previous estimation (Nakatani et al., 2007). While two fission events were observed from the LCA of bony fishes to the LCA of ray-finned fishes (Table S5), no major chromosomal fusion/fission event was observed during the evolution from the LCA of bony fishes to the LCA of tetrapods (Figure 3A).

In addition, we noticed that the proto-chromosomes 11-20 and 13-28 should have been fused independently in the lungfish and spotted gar because they are not fused in chicken, frog, bichir, or shark (Figures 3B and 3C). Similar independent fusions are also observed between the spotted gar and the western clawed frog, and between the lungfish and the bichir (Figures S3A and S3B; Table S5). The presence of shared fusion events in different lineages suggested that the chromosome fusion events might be correlated with certain characteristics of the proto-chromosomes themselves.

Evolutionary analyses on highly conserved elements, genes, and specifically expressed genes from bony fishes to tetrapods

We identified highly conserved elements (HCEs) in different lineages of bony fishes by using 12 species' genomes, including cartilaginous fishes, ray-finned fishes, and sarcopterygians. A total of 2,157, 1,191, and 4,916 HCEs were found to be gained by the CA of sarcopterygians, lungfish-tetrapods, and tetrapods, respectively. Besides, 388, 559, and 394 HCEs were lost in the CA of sarcopterygians, lungfish-tetrapods, and tetrapods, respectively, but are present in cartilaginous fishes and ray-finned fishes. These HCEs contain both conserved non-coding elements (CNEs) and coding elements. Moreover, from the transcriptomes of 220 samples of nine tissues of five species, including alligator gar, bichir, lungfish, frog, and mouse (Bi et al., 2021), we observed that 40 genes experienced expression pattern alternation in tissue specificity among tetrapods, lungfishes, and ray-finned fishes (Table S6). The most significant cases of these HCEs and specifically expressed genes will be discussed in the following sections.

Evolution of the respiratory system

Pulmonary surfactants are considered critical for lung evolution (Liem, 1988). They can reduce the surface tension of pulmonary alveoli and facilitate the expansion and contraction of the lungs during respiration. A previous study suggested that the cholesterol/phospholipid ratio (μg/μg) in surfactants is much higher in ray-finned fishes (~0.2–0.27) than in African lungfishes and tetrapods (~0.05–0.075) (Daniels and Orgeig, 2003). The role of cholesterol is subtle, in that it can confer the correct fluidity

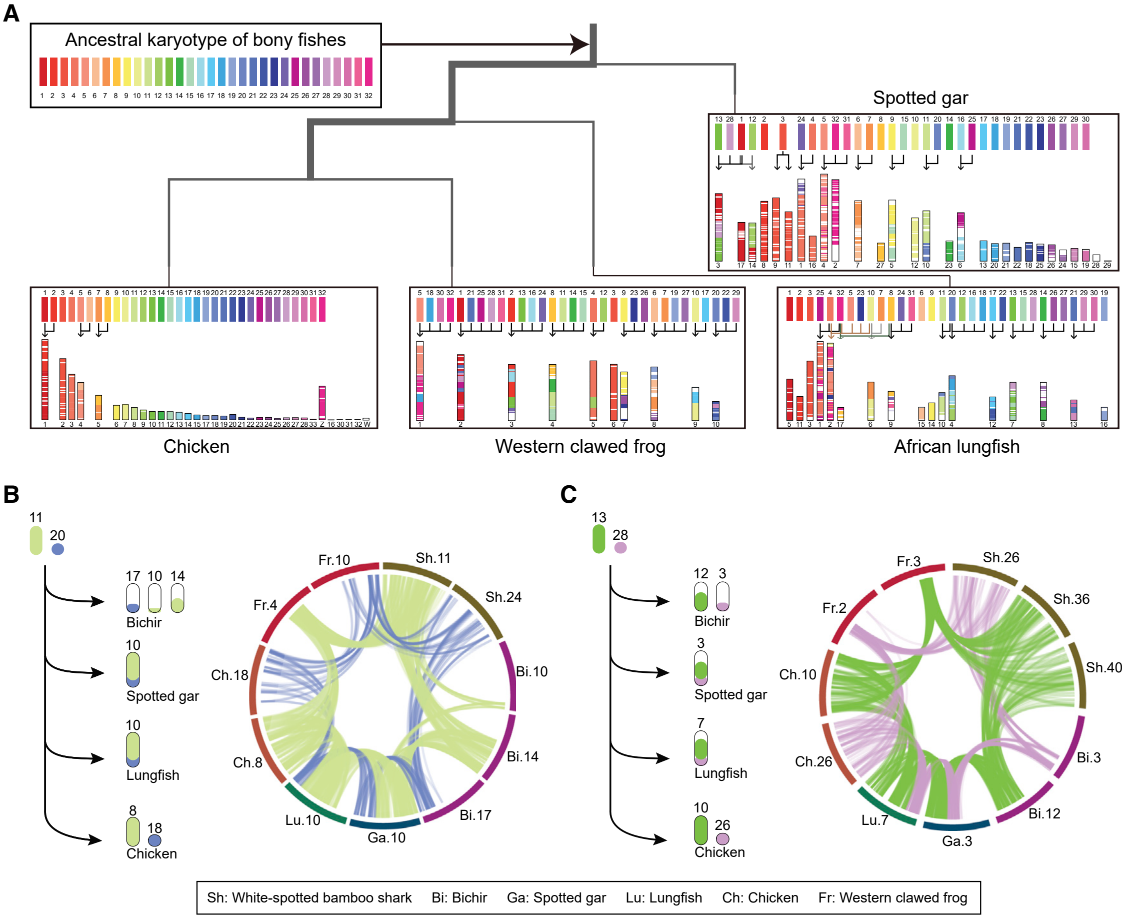


Figure 3. Conservation of proto-chromosome segments during the vertebrate water-to-land transition process

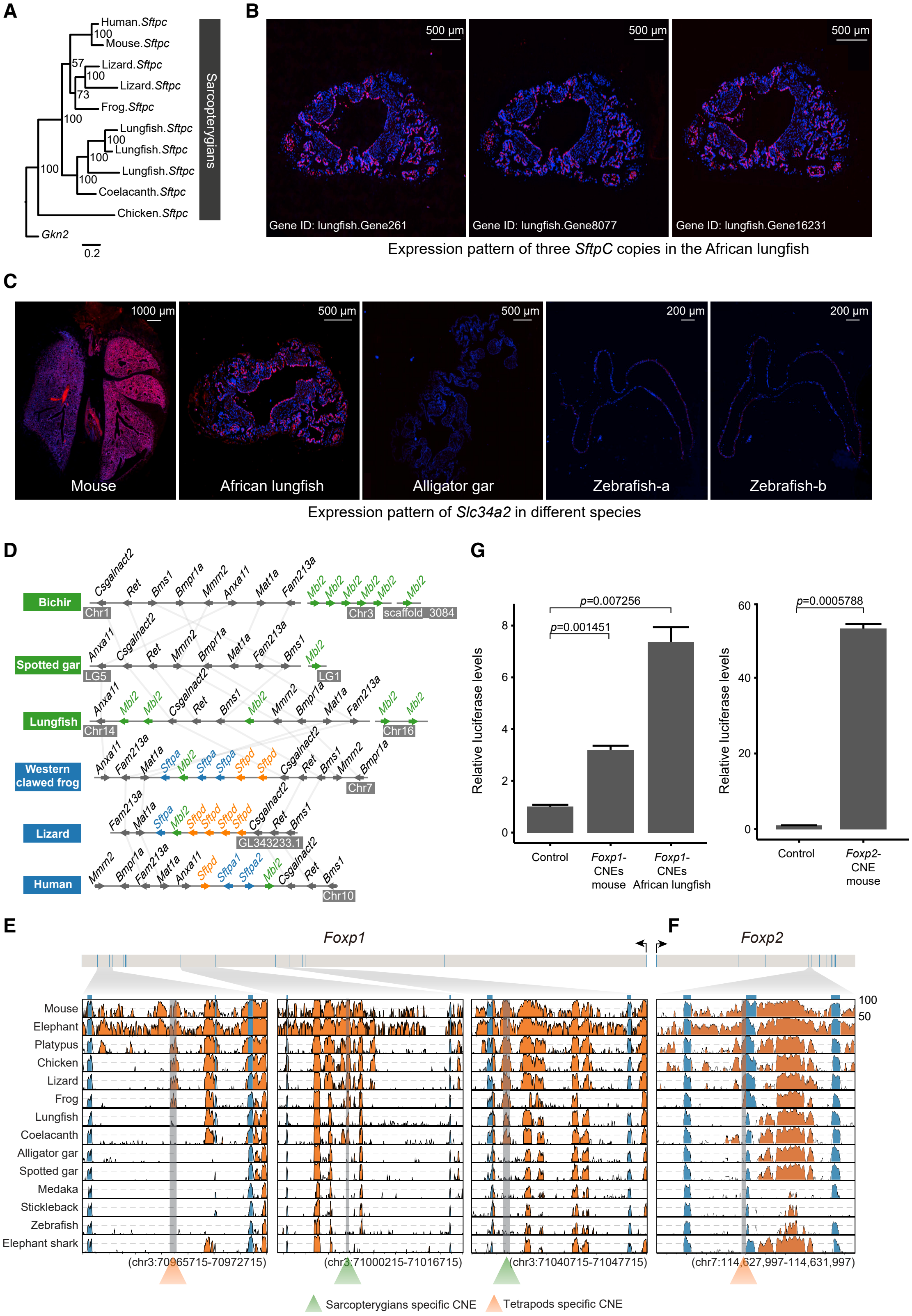
(A) The figure depicts a model for the distribution of the ancestral proto-chromosome segments in the genomes of spotted gar, chicken, western clawed frog, and the African lungfish.

(B–C) Two chromosome fusion events shared between spotted gar and African lungfish.

See also Figure S3 and Table S4.

and viscosity upon pulmonary surfactant (Orgeig and Daniels, 2001), while an excess of cholesterol inhibits surfactant function (Leonenko et al., 2007). To alleviate this conflict, the surfactant protein C (SP-C) is helpful in sustaining the presence of cholesterol in surfactant without functional impairment (Gómez-Gil et al., 2009; Roldan et al., 2016). The gene *Sftpc*, which encodes SP-C, was found to be originated via duplication, since the MRCA of sarcopterygians and then further duplications resulted in three copies in the African lungfish (Figure 4A). Our *in situ* hybridization results show clear evidence that they are indeed expressed in lungs of lungfishes (Figure 4B) and thus may function in regulating pulmonary surfactant.

In line with the newly evolved *Sftpc* genes, there are three genes that are highly expressed in the lungs of African lungfishes, African clawed frog, and mouse, but not the swim bladders of bichir and alligator gar, and are all functionally related to pulmonary surfactants. One of them (Figure S4A), *Slc34a2*, has a key role in transporting phosphate released from phospholipids during pulmonary surfactant recycling (Izumi et al., 2017). Our *in situ* hybridization results confirm that *Slc34a2* is barely expressed in the swim bladders of ray-finned fishes but is highly expressed in the lungs of lungfish and tetrapods (Figure 4C). Previous studies suggested that *Slc34a2* was expressed in the intestine and kidney of teleosts to increase phosphorus utilization efficiency (Chen et al., 2016, 2017). The recruitment of *Slc34a2* into the lungs of sarcopterygians could meet the greater demand for pulmonary surfactants. The other two lung highly expressed genes (Figure S4A), *Nkx2-1* and *Nrp1*, are associated with the upregulation of pulmonary surfactants (Attarian et al., 2018; Joza et al., 2012). We identified a sarcopterygians-specific CNE located upstream of *Nrp1* (Figure S4B). Our *in vitro* reporter assay experiment indicates that the CNE has enhancer activity (Figure S4C), which might have changed the expression pattern of the *Nrp1* gene from the actinopterygians to sarcopterygians. Taken together, the emergence of *Sftpc* and enhanced ability for pulmonary surfactant recycling might have contributed to the comparable pulmonary oxygen diffusing capacity between the African lungfishes and amphibians (Jorgensen and Joss, 2011).



(legend on next page)

Further alternations of pulmonary surfactant proteins are found in the lineage leading to the LCA of tetrapods. In the rest members of the surfactant proteins (SPs), *Sftpb*, encoded SP-B, is the most ancient member of the family; it is present in all bony fishes but absent in cartilaginous fishes (Figure S4D). We observed that while the *Sftpb* gene is specifically expressed in lungs of tetrapods, it tends to be more broadly expressed in multiple tissues of ray-finned fishes and lungfishes (Figure S4E). The hydrophilic surfactant proteins SP-A and SP-D are mainly involved in the pulmonary innate host defense system (Haagsman and Diemel, 2001). We observed that the *SP-A* and *SP-D* encoded genes originated from duplications of the *Mbl2* gene since the LCA of tetrapods (Figures 4D, S4F, and S4G). In addition, previous studies have indicated that the Foxp1/2 transcription factors are essential in lung development (Shu et al., 2007). We observed two sarcopterygian-specific CNEs and a tetrapod-specific CNE located in the intron region of the gene *Foxp1* (Figure 4E), and a tetrapod-specific CNE proximal to the *Foxp2* (Figure 4F). Results from our reporter assay suggest that these CNEs have potential enhancer activity (Figure 4G).

All these results suggest a conjecture that the evolution of lung respiratory capacity might have taken place in three steps. The first step was that the common ancestors of bony fishes already had an initial air-breathing ability as revealed by Liem (1988) and (Bi et al. 2021), which is also corroborated by the presence of *Sftpb* in all bony fishes as observed in this study. The second step was an increase in air breathing capacity through such genetic innovations as the emergence of *Sftpc* and CNEs proximal to *Foxp1* in the LCA of sarcopterygians. The third step might have resulted from further genetic innovation, including the appearance of *Sftpa*, *Sftpd*, and CNE proximal to *Foxp2*, providing the last critical basis for the advanced respiratory system in the tetrapod lineage.

Origin of pentadactyl limb and terrestrial locomotion

The appearance of five digits is the hallmark event in vertebrates' water-to-land transition (Clack, 2009). Previous studies indicated that *Hoxa13* and *Hoxd13* are essential for digit morphogenesis (Zákány and Duboule, 1999). The exclusive expression of *Hoxa11* and *Hoxa13* in tetrapod limbs (Davis et al., 2007), while they have largely overlapping expression in fins (Metscher et al., 2005), may play a pivotal role in this process (Kherdjemil et al., 2016). Interestingly, we observed a tetrapod-specific CNE with a length of 67 bp, located 200 bp upstream of *Hoxa11* (Figure 5A), which overlaps with the antisense transcripts *Hoxa11as204* and *Hoxa11as205*. A previous study suggested that the expression pattern of *Hoxa11as205* is responsible for the exclusive expression of *Hoxa11* and *Hoxa13* and contributes to the transition from polydactyl limbs to a pentadactyl limb (Kherdjemil et al., 2016). Our results support this inference, and this tetrapod-specific CNE may be a key genetic innovation for the origin of the pentadactyl limb. We then investigated this CNE in different tetrapod lineages (Table S7). While it is highly conserved across amphibians, crocodiles, turtles, and mammals, it is considerably altered in the genome of snakes and birds (Figure 5B), which further suggests that it might be related to the origin of digits.

The fin-to-limb transition resulted in the three parts of tetrapod limbs, which, from proximal to distal, are the stylopod, zeugopod, and autopod. Two ancient actinotrichia proteins (encoded by *and1/2* and *and3*) were found to be crucial for the fin-to-limb transition by determining the fate of apical ectodermal ridge (AER) (Zhang et al., 2010), and interestingly, while the coelacanths have both copies, the lungfish has only one (*and1/2*), and tetrapods have lost both (Figure S5A). The progressive loss of actinotrichia proteins had been hinted by transcriptome data (Biscotti et al., 2016) and is now validated by our genome data. This finding is consistent with previous observations that radius is also present in modern and fossil lungfish (Johanson et al., 2007; Jude et al., 2014). In addition, we found that about 40 continuous amino acids (AAs) at the beginning of *Hoxb13* are highly conserved in non-tetrapod vertebrates, while they have been lost in tetrapods (Figure 5C). A previous study showed that this gene is expressed in the distal mesenchyme of developing hind limbs (Carlson et al., 2001). The alteration of *Hoxb13* at the protein level might be related to the fin-to-limb transition.

Aside from changes in morphogenesis, the tetrapods require their axons, especially motoneurons, to leave the cord to innervate the muscles of the limbs. The axons to and from the forelimbs produce cervical enlargement, while the axons to and from the hind limbs/tails provide lumbar enlargement (Butler and Hodos, 2005). *Hoxc10* and *Hoxd10* have been proved to play key roles in establishing the lumbar motoneuron columnar, divisional, and motor pool identity in mouse (Wu et al., 2008). The two tetrapod-specific CNEs (Figure 5D) located upstream (3 Kb and 2.5 Kb) of *Hoxc10* may be related to terrestrial locomotion of tetrapods, and our reporter assay shows that they are candidate enhancers (Figure S5B).

Figure 4. Evolution of the lung respiratory functions in sarcopterygians

(A) Bayesian tree of *Sftpc* genes in Sarcopterygians. The numbers at each node refer to the corresponding supporting probability in percentage.

(B) Results of *in situ* hybridization analysis of the *Sftpc* genes in the lung of the African lungfish. There are three copies in the lungfish *Sftpc* gene family, and we designed specific probe for each copy to detect the expression signals (red) of the three genes.

(C) Results of *in situ* hybridization analysis of *Slc34a2* in the lungs of mouse and African lungfish and swim bladders of alligator gar and zebrafish. The zebrafish has two copies of Slc34a2, and we designed two specific probes for them. The red regions indicate the expression of these genes.

(D) Origin of the *Sftpa* and *Sftpd* genes from tandem duplication of *Mbl2* in the LCA of tetrapods. The relative positions of genes on chromosomes are indicated by arrows, with arrows to the right representing the forward strand and arrows to the left representing the reverse strand.

(E–F) Four CNEs in the intronic regions of *Foxp1* and *Foxp2*, which are indicated by different colored triangles, using human (GRCh38) as references.

(G) Results of reporter assays suggest that the CNEs in the intronic regions of *Foxp1* and *Foxp2* could significantly improve the expression level of luciferases. In the left panel, the three CNEs of mouse and the two CNEs of the African lungfish were connected, respectively, for testing. Significance was tested by Student's t test and data are represented as mean ± SEM.

See also Figure S4.

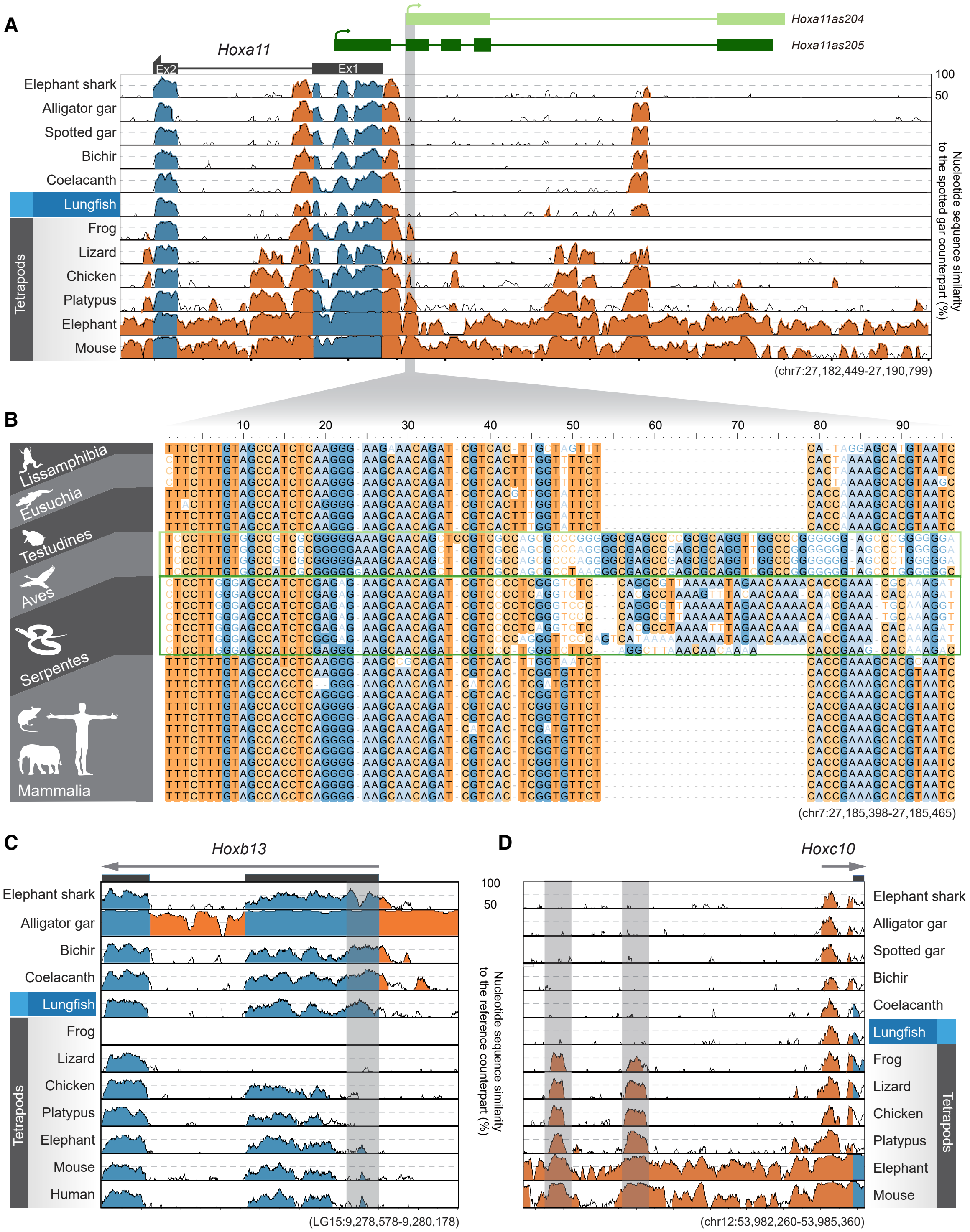


Figure 5. The CNEs potentially related to the origin of digits and other changes related to terrestrial locomotion

(A) VISTA sequence conservation plot of the tetrapod-specific CNE around *Hoxa11*, using human (GRCh38) as a reference. Ex1: exon1 of *Hoxa11*; Ex2: exon2 of *Hoxa11*. The two antisense RNAs that overlap with this CNE are indicated with green lines.

(B) Sequence alignment shows that the tetrapod-specific CNE was changed specifically in Serpentes and Aves. The list of species used here is in Table S7.

(legend continued on next page)

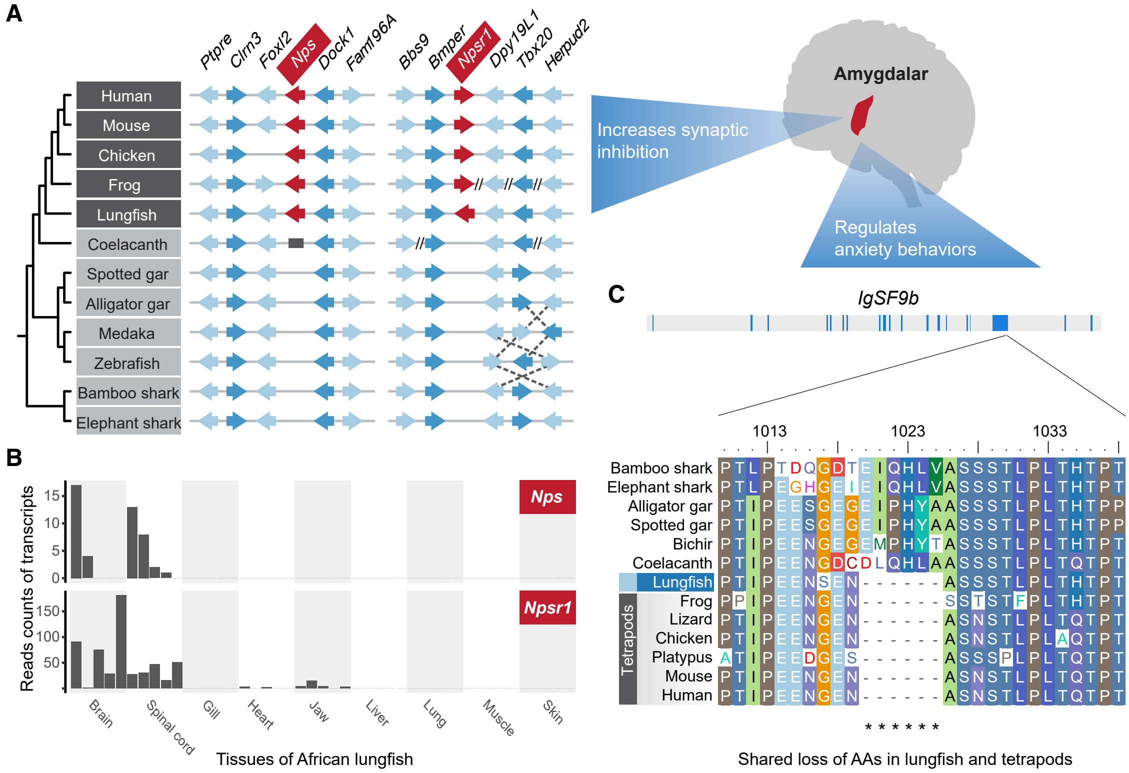


Figure 6. Alterations in genes associated with the amygdala and anxiolytic ability

(A) Origin of *Nps* and *Npsr1* since the appearance of lungfishes. The black rectangle in the *Nps* of coelacanth indicates a precursor sequence.

(B) Expression of *Nps* and *Npsr1* in nine tissues of the African lungfish.

(C) AA alignment of *IgSF9b*. The gray bar with blue stripes shows the exon-intron structure of *IgSF9b*.

See also Figure S5.

Potentially enhanced anxiolytic ability

One of the well-known changes that occurred within the brain during vertebrate water-to-land transition is in the limbic system, in which the basic subdivisions and connections of the amygdala are established in tetrapods (Bruce and Neary, 1995) and perhaps also in lungfishes (González and Northcutt, 2009; Northcutt, 2009). Through a genome-wide scan, we identified two new genes, *Npsr1* and *Nps*, that appeared in the lineage of the LCA of lungfishes and tetrapods (Figure 6A). These two genes encode Neuropeptide S receptor (NPSR) and Neuropeptide S (NPS), respectively, which can promote arousal and anxiolytic-like effects (Xu et al., 2004), moderate stress processing, and increase synaptic inhibition in the amygdala (Dannlowski et al., 2011; Medina et al., 2014; Streit et al., 2014). They had been previously supposed to be only tetrapod-specific genes (Reinscheid, 2007). Our results show that the *Npsr1* gene originated from the duplication of the mesotocin receptor gene in sarcopterygians and is expressed mainly in the brain and spinal cord (Figure 6B). *Nps* originated *de novo* from precursor sequences in sarcopterygians (Figure S5C). We also observed evidence of their expression in brain and spinal cord samples in this species (Figure 6B), indicating that the lungfish has a primitive NPS/NPSR system. Evidence also showed that the Nps is produced by the amygdala (Xu et al., 2007). Together with these results, this pair of newly originated genes may imply enhanced anxiolytic ability in the lineage of lungfishes and tetrapods.

In addition to NPS/NPSR, gamma-aminobutyric acid (GABA) within the amygdala is another critical inhibitory neurotransmitter used to control feelings of fear and anxiety (Babaev et al., 2018b). Two HCEs across bony fishes located in the coding regions of GABA-related genes were found to have deletion or insertion in the lineage leading to the LCA of tetrapods and lungfishes. The first gene, *IgSF9b*, whose product has a six-AA deletion (Figure 6C), was proved to be strongly expressed in GABAergic interneurons and to form a complex with *Nlgn2* as an inhibitory synaptic organization (Babaev et al., 2018a; Woo et al., 2013). The second gene, *Arfgef1*, which could maintain membrane surface postsynaptic GABA$_A$ receptors (Teoh et al., 2020), has a

(C) VISTA sequence conservation plot of the specific loss of 5′ AAs of *Hoxb13* in tetrapods, using the spotted gar (LepOcu1) as a reference. The gray shadowed column refers to the location of lost sequence in tetrapods.

(D) The tetrapod-specific CNEs close to *Hoxc10*, using human (GRCh38) as a reference. The gray shadowed columns refer to the locations of newly gained elements in tetrapods.

See also Figure S5.

product containing a two-AA insertion (Figure S5D). Moreover, *Gh*, the only gene that was observed to be expressed specifically in the brain of tetrapods and the African lungfish (Table S6), encodes growth hormone. The upregulation of *Gh* in the amygdala has been found to be related to the establishment of fear memory (Gisabella et al., 2016). Overall, these genetic innovations shared by lungfishes and tetrapods are consistent with previous observations that the tetrapod-like amygdaloid complex originated in the lineage leading to the CA of lungfishes and tetrapods (González et al., 2010; Maximino et al., 2013).

Pharyngeal remodeling

Along with the shift of the primary respiratory function from gills to lungs in terrestrial vertebrates, the branchial arches no longer generated gills. The branchial arch reduced in number, from five in sarcopterygian fishes to four or three in tetrapods (Graham and Shone, 2019). The gene *Hoxb3* is thought to be important in regulating the development of pharyngeal arches (Tomotsune et al., 2000). A CNE across most vertebrates located 4 Kb upstream of *Hoxb3* was found to be no longer conserved in tetrapods (Figure S5E), which might be a consequence of the relaxation of selection caused by the loss of two pharyngeal arches.

Simultaneously, the second arch in the embryonic development of human and other tetrapods was remodeled to cover more caudal arches, which resulted in the internalization of the posterior arches and the loss of an external opening at the posterior end (Graham and Shone, 2019). The parathyroid glands have then appeared from the posterior pharyngeal pouches since tetrapods emerged; these glands are responsible for regulating the release of calcium from internal stores (Graham and Shone, 2019), and this function is achieved by gills in ray-finned fishes (Okabe and Graham, 2004). The gene *Pax1* was found to be directly related to the development of pharyngeal pouches posterior to the second arch (Okada et al., 2016) and the parathyroid gland (Su et al., 2001). We identified a tetrapod-specific CNE upstream of *Pax1*, which could act as an enhancer (Figures S5F and S5G). The *Gcm2* gene is essential for the differentiation and survival of parathyroid glands (Liu et al., 2007). We noticed that this gene was highly expressed in the gills of ray-finned fishes and the African lungfish, but expressed at a low level in the gill of clawed frog (Figure S5H), suggesting that the corresponding functions had been shifted from the gills to the parathyroid gland in tetrapods.

Specialized characteristics in lungfishes

The African lungfishes can aestivate in the mud for several months or even years to get through the dry and hot season (Filogonio et al., 2017). We conducted aestivation treatment for six African lungfish (*P. annectens*) individuals, collected and analyzed transcriptome data from 11 tissues of the six aestivated individuals and three non-aestivated individuals. We found that the overall expression profiles are highly similar between aestivated and non-aestivated individuals (Figures S6A and S6B). The downregulated genes are enriched in hormone activity in brain (hypergeometric test, adjusted p value = 8.4×10^{-13}) and collagen-containing extracellular matrix in skin (adjusted p value = 8.0×10^{-3}) (Figure S6B). Among the ten genes that are upregulated in ten or more tissues, eight of them are heat shock protein genes (Figure S6C). The globally increased expression of heat shock proteins may protect lungfishes in the dry and hot seasons.

The dentition pattern is another unique feature of lungfishes that was considered to be highly derived in sarcopterygian fishes (Qiao and Zhu, 2009). Their individual teeth, arranged in a radial pattern, are integrated into the dental plate and will not be lost through shedding. To a certain extent, they are similar to the statodont dentition of placoderms that also never sheds (Zerina and Smith, 2005). We examined 51 genes that have been reported to be associated with tooth development (Bei, 2009) in 18 vertebrate species. None of the genes that are present in most vertebrates was found to be absent in the lungfishes. And only one gene, *Pax9*, was found to be with marked specific mutations in the regions that are conserved in vertebrates other than lungfishes (Figure S6D). In addition, we also detected positive selection signal in the *Pax9* gene of lungfishes (Figure S6E). *Pax9* determines the exact location of tooth germ appearances, and mutation in this gene is related to oligodontia in human beings (Suda et al., 2011). The specific mutations in lungfishes may lead to their specific dentitions.

DISCUSSION

The large genome size of lungfishes has resulted in the long-standing absence of their genome sequences, which has been a formidable obstacle to study the evolutionary mechanism of vertebrates' water-to-land transition. Combining ONT single molecular sequencing technology, optical mapping, and Hi-C technology, we successfully obtained a high-quality chromosome-level genome assembly for the African lungfish, which is the largest sequenced genome reported so far. Compared with the recently released Australian lungfish genome (Meyer et al., 2021), which missed 21%–35% of the genome sequences, our assembly is nearly complete and much more continuous with doubled chromosome N50 size. Our assembly also resulted in much better gene annotation for African lungfish and was over seven times higher in number of 1:1 orthologs across vertebrates than that in Australian lungfish. Nevertheless, these two available lungfish genomes provide important resources for understanding land-to-water transition of vertebrates and evolution of the mysterious lungfish species.

Our comparative genomic analysis revealed that the vertebrates' transition to land involved different kinds of genetic innovations. It is intriguing that none of the vertebrate lineages successfully emerged onto the land except the monophyletic group of tetrapods. Given the fact that the primitive respiration system is still preserved in non-teleost ray-finned fishes (Liem, 1988) and that they already had most of the genes essential for life on the land (Bi et al., 2021), we propose a three-step scenario for the water-to-land evolution: first, the common ancestor of bony fishes evolved the initial air-breathing ability; then, the genetic innovations associated with air-breathing ability (e.g., the new gene *Sftpc* and the recruitment of *Slc34a2*), nervous system (e.g., the new gene *Nps* and *Npsr*), and other improvements had rendered the CA of lungfishes and tetrapods the ability to leave the water temporarily; and finally, the CA of tetrapods acquired the respiration and locomotion system adapting to the terrestrial living. It is noteworthy that vomeronasal receptor (VR) gene expansion in lungfishes may not be a unique innovation in tetrapods as suggested by Meyer et al. (2021), because hagfish, some ray-finned fishes, and some tetrapods also have significantly expanded VRs (Bi et al., 2021), and the *hs72* enhancer nearby the *Sall1* gene is not Sarcopterygian in origin because it is present in ray-finned fishes except the zebrafish. It is clear that more studies are needed in order to reveal the concrete functional roles of those genetic changes observed in these two initial lungfish genomics studies.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - ○ Source organisms
  - ○ Cell lines
- METHOD DETAILS
  - ○ DNA isolation and genomic sequencing
  - ○ Estimation of genome size
  - ○ Genome assembly and chromosome anchoring
  - ○ Genome quality evaluation
  - ○ RNA preparation and sequencing
  - ○ Annotation of genome sequences
  - ○ Assessment of whole genome duplication (WGD)
  - ○ Phylogenetic relationship
  - ○ Estimation of increment of intron sizes
  - ○ Reconstruction of ancestral karyotype
  - ○ Analysis of conserved elements
  - ○ Identification of new genes
  - ○ Comparative transcriptome analysis
  - ○ *In situ* hybridization
  - ○ Dual-luciferase assay
  - ○ Evolution of tooth related genes
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cell.2021.01.047.

ACKNOWLEDGMENTS

We are very grateful to Drs. Erwei Zuo and Zhifang Li for their great help in our functional analysis. This work was supported by the Strategic Priority Research Program of CAS (XDB13000000) to W.W., S. He, and W.Z.; the 1000 Talent Project of Shaanxi Province to Q.Q. and K.W.; Research Funds for Interdisciplinary Subject, NWPU (19SH030408) to Q.Q.; grants from the State Key Laboratory of Freshwater Ecology and Biotechnology, the Youth Innovation Promotion Association of the CAS, and the NSFC of China (41876179 and 31972866) to S. He and L.Y.; and Guangdong Province Universities and Colleges Pearl River Scholar Funded Scheme (2018) and the National Key Research and Development Program of China (2018YFD0900604) to J.W. G.Z. is supported by the Villum Foundation (grant no. 25900). We also thank the supports from Fish10k project and Frasergen Bioinformatics Co., Ltd.

AUTHOR CONTRIBUTIONS

W.W., S. He, Q.Q., W.Z., G.Z., K.W., and J.W. designed and managed the project. L.Y. and H.J. collected the samples and performed the flow cytometry experiments and *in situ* hybridization experiments. R.Z., X.G., and H. Zeng generated the sequencing libraries. J.H. and D.W. developed the NextDenovo software for read correction, and J.R. developed wtdbg2 for initial contig assembly. K.W. and C.Z performed genome assembly and gene annotation. G.F., Y.S., Q.L., Y. Zhang, S. Hao, H. Zhang, X.X., and X.L. performed the optical mapping, Hi-C chromosomal assembly, and repeats annotation. K.W., C.Z., W.X., X.B., Y.Y., C.F., and Y.L. performed the evolutionary analysis. Y. Zhu and M.Z. provided fossil records and revised the sections related to the water-to-land transition. F.S. estimated the divergence time of repeat sequences. H.C. and T.M. performed the *in vitro* reporter assay experiments. K.W., W.W., C.Z., J.W., Q.Q., and G.Z. wrote and revised the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 17, 2020
Revised: November 9, 2020
Accepted: January 27, 2021
Published: February 4, 2021

REFERENCES

Amemiya, C.T., Alföldi, J., Lee, A.P., Fan, S., Philippe, H., Maccallum, I., Braasch, I., Manousaki, T., Schneider, I., Rohner, N., et al. (2013). The African coelacanth genome provides insights into tetrapod evolution. Nature *496*, 311–316.

Ashley-Ross, M.A., Hsieh, S.T., Gibb, A.C., and Blob, R.W. (2013). Vertebrate land invasions-past, present, and future: an introduction to the symposium. Integr. Comp. Biol. *53*, 192–196.

Attarian, S.J., Leibel, S.L., Yang, P., Alfano, D.N., Hackett, B.P., Cole, F.S., and Hamvas, A. (2018). Mutations in the thyroid transcription factor gene NKX2-1 result in decreased expression of SFTPB and SFTPC. Pediatr. Res. *84*, 419–425.

Babaev, O., Cruces-Solis, H., Piletti Chatain, C., Hammer, M., Wenger, S., Ali, H., Karalis, N., de Hoz, L., Schlüter, O.M., Yanagawa, Y., et al. (2018a). IgSF9b regulates anxiety behaviors through effects on centromedial amygdala inhibitory synapses. Nat. Commun. *9*, 5400.

Babaev, O., Piletti Chatain, C., and Krueger-Burg, D. (2018b). Inhibition in the amygdala anxiety circuitry. Exp. Mol. Med. *50*, 18.

Bei, M. (2009). Molecular genetics of tooth development. Curr. Opin. Genet. Dev. *19*, 504–510.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., and Haussler, D. (2004). Ultraconserved elements in the human genome. Science *304*, 1321–1325.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. *27*, 573–580.

Benton, M., Donoghue, P.C.J., Asher, R., Friedman, M., Near, T., and Vinther, J. (2015). Constraints on the timescale of animal evolutionary history. Palaeontologia Electronica *18*, 1–107.

Bi, X., Wang, K., Yang, L., Pan, H., Jiang, H., Wei, Q., Fang, M., Yu, H., Zhu, C., Cai, Y., et al. (2021). Tracing the genetic footprints of vertebrate landing in non-teleost ray-finned fishes. Cell *184*, this issue, 1377–1391.e14.

Biscotti, M.A., Gerdol, M., Canapa, A., Forconi, M., Olmo, E., Pallavicini, A., Barucca, M., and Schartl, M. (2016). The Lungfish Transcriptome: A Glimpse into Molecular Evolution Events at the Transition from Water to Land. Sci. Rep. *6*, 21571.

Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., et al. (2004). Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res. *14*, 708–715.

Braasch, I., Gehrke, A.R., Smith, J.J., Kawasaki, K., Manousaki, T., Pasquier, J., Amores, A., Desvignes, T., Batzel, P., Catchen, J., et al. (2016). The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. Nat. Genet. *48*, 427–437.

Bruce, L.L., and Neary, T.J. (1995). The limbic system of tetrapods: a comparative analysis of cortical and amygdalar populations. Brain Behav. Evol. *46*, 224–234.

Bruno, M., Mahgoub, M., and Macfarlan, T.S. (2019). The Arms Race Between KRAB-Zinc Finger Proteins and Endogenous Retroelements and Its Impact on Mammals. Annu. Rev. Genet. *53*, 393–416.

Butler, A.B., and Hodos, W. (2005). Comparative Vertebrate Neuroanatomy: Evolution and Adaptation, Second Edition (Wiley).

Caley, T., Extier, T., Collins, J.A., Schefuß, E., Dupont, L., Malaizé, B., Rossignol, L., Souron, A., McClymont, E.L., Jimenez-Espejo, F.J., et al. (2018). A two-million-year-long hydroclimatic context for hominin evolution in southeastern Africa. Nature *560*, 76–79.

Carlson, M.R., Komine, Y., Bryant, S.V., and Gardiner, D.M. (2001). Expression of Hoxb13 and Hoxc10 in developing and regenerating Axolotl limbs and tails. Dev. Biol. *229*, 396–406.

Chalopin, D., Naville, M., Plard, F., Galiana, D., and Volff, J.-N. (2015). Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. Genome Biol. Evol. *7*, 567–580.

Chen, P., Tang, Q., and Wang, C. (2016). Characterizing and evaluating the expression of the type IIb sodium-dependent phosphate cotransporter (slc34a2) gene and its potential influence on phosphorus utilization efficiency in yellow catfish (Pelteobagrus fulvidraco). Fish Physiol. Biochem. *42*, 51–64.

Chen, P., Huang, Y., Bayir, A., and Wang, C. (2017). Characterization of the isoforms of type IIb sodium-dependent phosphate cotransporter (Slc34a2) in yellow catfish, Pelteobagrus fulvidraco, and their vitamin D_3-regulated expression under low-phosphate conditions. Fish Physiol. Biochem. *43*, 229–244.

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890.

Clack, J.A. (2009). The Fin to Limb Transition: New Data, Interpretations, and Hypotheses from Paleontology and Developmental Biology. Annu. Rev. Earth Planet. Sci. *37*, 163–179.

Clack, J.A. (2012). Gaining Ground, The Origin and Evolution of Tetrapods, Second Edition (Indiana University Press).

Coates, M.I., Finarelli, J.A., Sansom, I.J., Andreev, P.S., Criswell, K.E., Tietjen, K., Rivers, M.L., and La Riviere, P.J. (2018). An early chondrichthyan and the evolutionary assembly of a shark body plan. Proc. Biol. Sci. *285* https://doi.org/10.1098/rspb.2017.2418.

Daniels, C.B., and Orgeig, S. (2003). Pulmonary surfactant: the key to the evolution of air breathing. News Physiol. Sci. *18*, 151–157.

Dannlowski, U., Kugel, H., Franke, F., Stuhrmann, A., Hohoff, C., Zwanzger, P., Lenzen, T., Grotegerd, D., Suslow, T., Arolt, V., et al. (2011). Neuropeptide-S (NPS) receptor genotype modulates basolateral amygdala responsiveness to aversive stimuli. Neuropsychopharmacology *36*, 1879–1885.

Davis, M.C., Dahn, R.D., and Shubin, N.H. (2007). An autopodial-like pattern of Hox expression in the fins of a basal actinopterygian fish. Nature *447*, 473–476.

De Bie, T., Cristianini, N., Demuth, J.P., and Hahn, M.W. (2006). CAFE: a computational tool for the study of gene family evolution. Bioinformatics *22*, 1269–1271.

Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C., Shamim, M.S., Machol, I., Lander, E.S., Aiden, A.P., and Aiden, E.L. (2017). De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science *356*, 92–95.

Dudchenko, O., Shamim, M., Batra, S., Durand, N., Musial, N., Mostofa, R., Pham, M., Glenn St Hilaire, B., Yao, W., Stamenova, E., et al. (2018). The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. bioRxiv. https://doi.org/10.1101/254797.

Dunbrack, R., Green, J.M., and Mlewa, C.M. (2006). Marbled lungfish growth rates in Lake Baringo, Kenya, estimated by mark-recapture. J. Fish Biol. *68*, 443–449.

Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S.P., Huntley, M.H., Lander, E.S., and Aiden, E.L. (2016). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell Syst. *3*, 95–98.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A., et al. (2019). The Pfam protein families database in 2019. Nucleic Acids Res. *47* (D1), D427–D432.

Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. *20*, 238.

Filogonio, R., Joyce, W., and Wang, T. (2017). Nitrergic cardiovascular regulation in the African lungfish, Protopterus aethiopicus. Comp. Biochem. Physiol. A Mol. Integr. Physiol. *207*, 52–56.

Frazer, K.A., Pachter, L., Poliakov, A., Rubin, E.M., and Dubchak, I. (2004). VISTA: computational tools for comparative genomics. Nucleic Acids Res. *32*, W273-W279.

Gisabella, B., Farah, S., Peng, X., Burgos-Robles, A., Lim, S.H., and Goosens, K.A. (2016). Growth hormone biases amygdala network activation after fear learning. Transl. Psychiatry *6*, e960–e960.

Gómez-Gil, L., Schürch, D., Goormaghtigh, E., and Pérez-Gil, J. (2009). Pulmonary surfactant protein SP-C counteracts the deleterious effects of cholesterol on the activity of surfactant films under physiologically relevant compression-expansion dynamics. Biophys. J. *97*, 2736–2745.

González, A., and Northcutt, R.G. (2009). An immunohistochemical approach to lungfish telencephalic organization. Brain Behav Evol *74*, 43–55.

González, A., Morona, R., López, J., Moreno, N., and Northcutt, G. (2010). Lungfishes, Like Tetrapods, Possess a Vomeronasal System. Frontiers in Neuroanatomy *4*. https://doi.org/10.3389/fnana.2010.00130.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. *29*, 644–652.

Graham, A., and Shone, V. (2019). Evolution and Development of Fishes (Cambridge University Press).

Haagsman, H.P., and Diemel, R.V. (2001). Surfactant-associated proteins: functions and structural variation. Comp. Biochem. Physiol. A Mol. Integr. Physiol. *129*, 91–108.

Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., and Wortman, J.R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. *9*, R7.

Harris, R.S. (2007). Improved pairwise alignment of genomic DNA (The Pennsylvania State University).

Hellsten, U., Harland, R.M., Gilchrist, M.J., Hendrix, D., Jurka, J., Kapitonov, V., Ovcharenko, I., Putnam, N.H., Shu, S., Taher, L., et al. (2010). The genome of the Western clawed frog Xenopus tropicalis. Science *328*, 633–636.

Hsia, C.C., Schmitz, A., Lambertz, M., Perry, S.F., and Maina, J.N. (2013). Evolution of air breathing: oxygen homeostasis and the transitions from water to land and sky. Compr. Physiol. *3*, 849–915.

Hu, J., Fan, J., Sun, Z., and Liu, S. (2020). NextPolish: a fast and efficient genome polishing tool for long-read assembly. Bioinformatics *36*, 2253–2255.

Imbeault, M., Helleboid, P.Y., and Trono, D. (2017). KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. Nature *543*, 550–554.

Izumi, H., Kurai, J., Kodani, M., Watanabe, M., Yamamoto, A., Nanba, E., Adachi, K., Igishi, T., and Shimizu, E. (2017). A novel *SLC34A2* mutation in a patient with pulmonary alveolar microlithiasis. Hum. Genome Var. *4*, 16047.

Johanson, Z., Joss, J., Boisvert, C.A., Ericsson, R., Sutija, M., and Ahlberg, P.E. (2007). Fish fingers: digit homologues in sarcopterygian fish fins. J. Exp. Zoolog. B Mol. Dev. Evol. *308*, 757–768.

Jones, B.R., Rajaraman, A., Tannier, E., and Chauve, C. (2012). ANGES: reconstructing ANcestral GEnomeS maps. Bioinformatics *28*, 2388–2390.

Jorgensen, J.E., and Joss, J.E. (2011). The Biology of Lungfishes (Taylor & Francis Ltd).

Joza, S., Wang, J., Fox, E., Hillman, V., Ackerley, C., and Post, M. (2012). Loss of semaphorin-neuropilin-1 signaling causes dysmorphic vascularization reminiscent of alveolar capillary dysplasia. Am. J. Pathol. *181*, 2003–2017.

Jude, E., Johanson, Z., Kearsley, A., and Friedman, M. (2014). Early evolution of the lungfish pectoral-fin endoskeleton: evidence from the Middle Devonian (Givetian) Pentlandia macroptera. Frontiers in Earth Science *2*. https://doi.org/10.3389/feart.2014.00018.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. *110*, 462–467.

Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.

Kent, W.J. (2002). BLAT–the BLAST-like alignment tool. Genome Res. *12*, 656–664.

Kherdjemil, Y., Lalonde, R.L., Sheth, R., Dumouchel, A., de Martino, G., Pineault, K.M., Wellik, D.M., Stadler, H.S., Akimenko, M.A., and Kmita, M. (2016). Evolution of Hoxa11 regulation in vertebrates is linked to the pentadactyl state. Nature *539*, 89–92.

Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. *37*, 907–915.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res. *19*, 1639–1645.

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol. Biol. Evol. *33*, 1870–1874.

Leonenko, Z., Gill, S., Baoukina, S., Monticelli, L., Doehner, J., Gunasekara, L., Felderer, F., Rodenstein, M., Eng, L.M., and Amrein, M. (2007). An elevated level of cholesterol impairs self-assembly of pulmonary surfactant into a functional film. Biophys. J. *93*, 674–683.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595.

Li, H., and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. Nature *475*, 493–496.

Liem, K.F. (1988). Form and Function of Lungs: The Evolution of Air Breathing Mechanisms. Am. Zool. *28*, 739–759.

Liu, Z., Yu, S., and Manley, N.R. (2007). Gcm2 is required for the differentiation and survival of parathyroid precursor cells in the parathyroid/thymus primordia. Dev. Biol. *305*, 333–346.

Liu, L., Yu, L., and Edwards, S.V. (2010). A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol. Biol. *10*, 302.

Long, J.A., and Gordon, M.S. (2004). The greatest step in vertebrate history: a paleobiological review of the fish-tetrapod transition. Physiol. Biochem. Zool. *77*, 700–719.

Loong, A.M., Pang, C.Y., Hiong, K.C., Wong, W.P., Chew, S.F., and Ip, Y.K. (2008). Increased urea synthesis and/or suppressed ammonia production in the African lungfish, Protopterus annectens, during aestivation in air or mud. J. Comp. Physiol. B *178*, 351–363.

Löytynoja, A. (2014). Phylogeny-aware alignment with PRANK. Methods Mol. Biol. *1079*, 155–170.

Lu, J., Zhu, M., Long, J.A., Zhao, W., Senden, T.J., Jia, L., and Qiao, T. (2012). The earliest known stem-tetrapod from the Lower Devonian of China. Nat. Commun. *3*, 1160.

Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics *27*, 764–770.

Maximino, C., Lima, M.G., Oliveira, K.R., Batista, Ede.J., and Herculano, A.M. (2013). ‘‘Limbic associative’’ and ‘‘autonomic’’ amygdala in teleosts: a review of the evidence. J. Chem. Neuroanat. *48-49*, 1–13.

McLaughlin, R.N., Jr., and Malik, H.S. (2017). Genetic conflicts: the usual suspects and beyond. J. Exp. Biol. *220*, 6–17.

Medina, G., Ji, G., Grégoire, S., and Neugebauer, V. (2014). Nasal application of neuropeptide S inhibits arthritis pain-related behaviors through an action in the amygdala. Mol. Pain *10*. https://doi.org/10.1186/1744-8069-10-32.

Metcalfe, C.J., Filée, J., Germon, I., Joss, J., and Casane, D. (2012). Evolution of the Australian lungfish (Neoceratodus forsteri) genome: a major role for CR1 and L2 LINE elements. Mol. Biol. Evol. *29*, 3529–3539.

Metscher, B.D., Takahashi, K., Crow, K., Amemiya, C., Nonaka, D.F., and Wagner, G.P. (2005). Expression of Hoxa-11 and Hoxa-13 in the pectoral fin of a basal ray-finned fish, Polyodon spathula: implications for the origin of tetrapod limbs. Evol. Dev. *7*, 186–195.

Meyer, A., Schloissnig, S., Franchini, P., Du, K., Woltering, J., Irisarri, I., Wong, W.Y., Nowoshilow, S., Kneitz, S., Kawaguchi, A., et al. (2021). Giant lungfish genome elucidates the conquest of land by vertebrates. Nature. https://doi.org/10.1038/s41586-021-03198-8.

Mistry, J., Finn, R.D., Eddy, S.R., Bateman, A., and Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. *41*, e121.

Nakatani, Y., Takeda, H., Kohara, Y., and Morishita, S. (2007). Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. Genome Res. *17*, 1254–1265.

Northcutt, R.G. (2009). Telencephalic organization in the spotted African Lungfish, Protopterus dolloi: a new cytological model. Brain Behav Evol *73*, 59–80.

Nowoshilow, S., Schloissnig, S., Fei, J.F., Dahl, A., Pang, A.W.C., Pippel, M., Winkler, S., Hastie, A.R., Young, G., Roscito, J.G., et al. (2018). The axolotl genome and the evolution of key tissue formation regulators. Nature *554*, 50–55.

Okabe, M., and Graham, A. (2004). The origin of the parathyroid gland. Proc. Natl. Acad. Sci. USA *101*, 17716–17719.

Okada, K., Inohaya, K., Mise, T., Kudo, A., Takada, S., and Wada, H. (2016). Reiterative expression of pax1 directs pharyngeal pouch segmentation in medaka. Development *143*, 1800–1810.

Organ, C., Struble, M., Canoville, A., de Buffrénil, V., and Laurin, M. (2016). Macroevolution of genome size in sarcopterygians during the water–land transition. Comptes Rendus Palevol *15*, 65–73.

Orgeig, S., and Daniels, C.B. (2001). The roles of cholesterol in pulmonary surfactant: insights from comparative and evolutionary studies. Comp. Biochem. Physiol. A Mol. Integr. Physiol. *129*, 75–89.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. *33*, 290–295.

Pond, S.L., Frost, S.D., and Muse, S.V. (2005). HyPhy: hypothesis testing using phylogenies. Bioinformatics *21*, 676–679.

Qiao, T., and Zhu, M. (2009). A new tooth-plated lungfish from the Middle Devonian of Yunnan, China, and its phylogenetic relationships. Acta Zoologica *90*, 236–252.

Reinscheid, R.K. (2007). Phylogenetic appearance of neuropeptide S precursor proteins in tetrapods. Peptides *28*, 830–837.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. *16*, 276–277.

Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. *11*, R25.

Rogers, R.L., Zhou, L., Chu, C., Márquez, R., Corl, A., Linderoth, T., Freeborn, L., MacManes, M.D., Xiong, Z., Zheng, J., et al. (2018). Genomic Takeover by Transposable Elements in the Strawberry Poison Frog. Mol. Biol. Evol. *35*, 2913–2927.

Roldan, N., Nyholm, T.K.M., Slotte, J.P., Pérez-Gil, J., and García-Álvarez, B. (2016). Effect of Lung Surfactant Protein SP-C and SP-C-Promoted Membrane Fragmentation on Cholesterol Dynamics. Biophys. J. *111*, 1703–1713.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A., and Huelsenbeck, J.P. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. *61*, 539–542.

Ruan, J., and Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. Nat. Methods *17*, 155–158.

Sanderson, M.J. (2003). r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics *19*, 301–302.

Sansom, J.I., Davies, S.N., Coates, I.M., Nicoll, S.R., and Ritchie, A. (2012). Chondrichthyan-like scales from the Middle Ordovician of Australia. Palaeontology *55*, 243–247.

Seppey, M., Manni, M., and Zdobnov, E.M. (2019). BUSCO: Assessing Genome Assembly and Annotation Completeness. Methods Mol. Biol. *1962*, 227–245.

Shu, W., Lu, M.M., Zhang, Y., Tucker, P.W., Zhou, D., and Morrisey, E.E. (2007). Foxp2 and Foxp1 cooperatively regulate lung and esophagus development. Development *134*, 1991–2000.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. *15*, 1034–1050.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics *31*, 3210–3212.

Sirijovski, N., Woolnough, C., Rock, J., and Joss, J.M. (2005). NfCR1, the first non-LTR retrotransposon characterized in the Australian lungfish genome, Neoceratodus forsteri, shows similarities to CR1-like elements. J. Exp. Zoolog. B Mol. Dev. Evol. *304*, 40–49.

Smith, M.D., Wertheim, J.O., Weaver, S., Murrell, B., Scheffler, K., and Kosakovsky Pond, S.L. (2015). Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. Mol. Biol. Evol. *32*, 1342–1353.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. *34*, W435-W439.

Streit, F., Haddad, L., Paul, T., Frank, J., Schäfer, A., Nikitopoulos, J., Akdeniz, C., Lederbogen, F., Treutlein, J., Witt, S., et al. (2014). A functional variant in the neuropeptide S receptor 1 gene moderates the influence of urban upbringing on stress processing in the amygdala. Stress *17*, 352–361.

Su, D., Ellis, S., Napier, A., Lee, K., and Manley, N.R. (2001). Hoxa3 and pax1 regulate epithelial cell death and proliferation during thymus and parathyroid organogenesis. Dev. Biol. *236*, 316–329.

Suda, N., Ogawa, T., Kojima, T., Saito, C., and Moriyama, K. (2011). Non-syndromic oligodontia with a novel mutation of PAX9. J. Dent. Res. *90*, 382–386.

Suzuki, A., and Yamanaka, K. (1988). Chromosomes of an African Lungfish, Protopterus annectens. Proceedings of the Japan Academy, Series B *64*, 119–121.

Teoh, J., Subramanian, N., Pero, M.E., Bartolini, F., Amador, A., Kanber, A., Williams, D., Petri, S., Yang, M., Allen, A.S., et al. (2020). Arfgef1 haploinsufficiency in mice alters neuronal endosome composition and decreases membrane surface postsynaptic GABA$_A$ receptors. Neurobiol. Dis. *134*, 104632.

Thomson, K.S. (1972). An attempt to reconstruct evolutionary changes in the cellular DNA content of lungfish. J. Exp. Zool. *180*, 363–371.

Tiersch, T.R., and Chandler, R.W. (1989). Chicken Erythrocytes as an Internal Reference for Analysis of DNA Content by Flow Cytometry in Grass Carp. Trans. Am. Fish. Soc. *118*, 713–717.

Tippmann, H.F. (2004). Analysis for free: comparing programs for sequence analysis. Brief. Bioinform. *5*, 82–87.

Tomotsune, D., Shirai, M., Takihara, Y., and Shimada, K. (2000). Regulation of Hoxb3 expression in the hindbrain and pharyngeal arches by rae28, a member of the mammalian Polycomb group of genes. Mech. Dev. *98*, 165–169.

Vurture, G.W., Sedlazeck, F.J., Nattestad, M., Underwood, C.J., Fang, H., Gurtowski, J., and Schatz, M.C. (2017). GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics *33*, 2202–2204.

Wang, Y., Tang, H., Debarry, J.D., Tan, X., Li, J., Wang, X., Lee, T.H., Jin, H., Marler, B., Guo, H., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. *40* https://doi.org/10.1093/nar/gkr1293.

Wang, Y., Lu, Y., Zhang, Y., Ning, Z., Li, Y., Zhao, Q., Lu, H., Huang, R., Xia, X., Feng, Q., et al. (2015). The draft genome of the grass carp (Ctenopharyngodon idellus) provides insights into its evolution and vegetarian adaptation. Nat. Genet. *47*, 625–631.

Warren, W.C., Hillier, L.W., Tomlinson, C., Minx, P., Kremitzki, M., Graves, T., Markovic, C., Bouk, N., Pruitt, K.D., Thibaud-Nissen, F., et al. (2017). A New Chicken Genome Assembly Provides Insight into Avian Genome Structure. G3 (Bethesda) *7*, 109–117.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (New York: Springer-Verlag).

Woo, J., Kwon, S.K., Nam, J., Choi, S., Takahashi, H., Krueger, D., Park, J., Lee, Y., Bae, J.Y., Lee, D., et al. (2013). The adhesion protein IgSF9b is coupled to neuroligin 2 via S-SCAM to promote inhibitory synapse development. J. Cell Biol. *201*, 929–944.

Wu, Y., Wang, G., Scott, S.A., and Capecchi, M.R. (2008). Hoxc10 and Hoxd10 regulate mouse columnar, divisional and motor pool identity of lumbar motoneurons. Development *135*, 171–182.

Xu, Y.L., Reinscheid, R.K., Huitron-Resendiz, S., Clark, S.D., Wang, Z., Lin, S.H., Brucher, F.A., Zeng, J., Ly, N.K., Henriksen, S.J., et al. (2004). Neuropeptide S: a neuropeptide promoting arousal and anxiolytic-like effects. Neuron *43*, 487–497.

Xu, Y.L., Gall, C.M., Jackson, V.R., Civelli, O., and Reinscheid, R.K. (2007). Distribution of neuropeptide S receptor mRNA and neurochemical characteristics of neuropeptide S-expressing neurons in the rat brain. J. Comp. Neurol. *500*, 84–102.

Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-Even, A., Horn-Saban, S., Safran, M., Domany, E., et al. (2005). Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics *21*, 650–659.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24*, 1586–1591.

Zákány, J., and Duboule, D. (1999). Hox genes in digit development and evolution. Cell Tissue Res. *296*, 19–25.

Zerina, J., and Smith, M.M. (2005). Origin and evolution of gnathostome dentitions: a question of teeth and pharyngeal denticles in placoderms. Biol. Rev. Camb. Philos. Soc. *80*, 303–345.

Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. (2000). A greedy algorithm for aligning DNA sequences. J. Comput. Biol. *7*, 203–214.

Zhang, Z., Li, J., Zhao, X.Q., Wang, J., Wong, G.K., and Yu, J. (2006). KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genomics Proteomics Bioinformatics *4*, 259–263.

Zhang, J., Wagh, P., Guay, D., Sanchez-Pulido, L., Padhi, B.K., Korzh, V., Andrade-Navarro, M.A., and Akimenko, M.A. (2010). Loss of fish actinotrichia proteins and the fin-to-limb transition. Nature *466*, 234–237.

Zhang, Y., Gao, H., Li, H., Guo, J., Wang, M., Xu, Q., Wang, J., Lv, M., Guo, X., Liu, Q., et al. (2019). Dynamic chromosome rearrangements of the white-spotted bamboo shark shed light on cartilaginous fish diversification mechanisms. bioRxiv, 602136. https://doi.org/10.1101/602136.

Zhu, M., and Fan, J. (1995). Youngolepis from the Xishancun Formation (Early Lochkovian) of Qujing; China. Geobios *28*, 293–299.

Zhu, M., Zhao, W., Jia, L., Lu, J., Qiao, T., and Qu, Q. (2009). The oldest articulated osteichthyan reveals mosaic gnathostome characters. Nature *458*, 469–474.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
African lungfish	This study	N/A
Critical commercial assays		
RNeasy Mini kit	QIAGEN	N/A
Anti-DIG-AP antibody	Jackson	N/A
DMEM	GIBCO	N/A
FBS	GIBCO	N/A
Antibiotic-Antimycotic	GIBCO	N/A
pGL4.23 vector	Promega	N/A
Turbofect reagent	ThermoFisher Scientific	N/A
Dual-Luciferase Reporter Assay System	Promega	N/A
Multimode microplate reader	Spark Tecan	N/A
Deposited data		
Sequencing data for African lungfish	This study	National Genomic Data Center (https://bigd.big.ac.cn/bioproject/) under the accession number PRJCA002950.
Genome assembly for African lungfish	This study	National Genomic Data Center (https://bigd.big.ac.cn/gwh/) under the accession number GWHANVS00000000.
Ensembl database release 96	European Bioinformatics Institute	http://apr2019.archive.ensembl.org/
Experimental models: cell lines		
HEK293T	DMEM	N/A
Software and algorithms		
fastp v0.19.4	Chen et al., 2018	https://github.com/OpenGene/fastp
Jellyfish v2.2.10	Marçais and Kingsford, 2011	https://www.cbcb.umd.edu/software/jellyfish/
GenomeScope	Vurture et al., 2017	http://qb.cshl.edu/genomescope/
Wtdbg2.huge	Ruan and Li, 2020	https://github.com/ruanjue/wtdbg-2.huge
NextDenovo v1.0	Hu Jiang	https://github.com/Nextomics/NextDenovo
NextPolish v1.01	Hu et al., 2020	https://github.com/Nextomics/NextPolish
Bionano Solve hybrid scaffold pipeline	Pacific Biosciences	https://bionanogenomics.com/wp-content/uploads/2018/04/30073-Bionano-Solve-Theory-of-Operation-Hybrid-Scaffold.pdf
3D-DNA v180419	Dudchenko et al., 2017	https://github.com/theaidenlab/3d-dna
Juicebox Assembly Tools v1.9.9	Dudchenko et al., 2018	https://github.com/aidenlab/Juicebox/wiki/Juicebox-Assembly-Tools
Trinity v2.8.5	Grabherr et al., 2011	https://github.com/trinityrnaseq/trinityrnaseq
Iso-Seq2 from SMRT Link v5.1.0	Pacific Biosciences	https://www.pacb.com/wp-content/uploads/2018-10-NA-UGM-Iso-Seq-Method.pdf
BLAT v36x2	Bejerano et al., 2004	https://genome.ucsc.edu/cgi-bin/hgBlat?command=start
minimap2 v2.17	Li, 2018	https://github.com/lh3/minimap2
Burrows-Wheeler Aligner (bwa mem v0.7.17)	Li and Durbin, 2010	https://github.com/lh3/bwa
RepeatMasker v4.0.7	Arian Smit & Robert Hubley	http://www.repeatmasker.org/
RepeatModeler v1.0.11	Arian Smit & Robert Hubley	http://www.repeatmasker.org/RepeatModeler/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
parseRM	Aurelie Kapusta	https://github.com/4ureliek/Parsing-RepeatMasker-Outputs
TRF v4.09	Benson, 1999	https://tandem.bu.edu/trf/trf.html
Exonerate v2.4.0	European Bioinformatics Institute	https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate
Augustus v3.3.2	Stanke et al., 2006	http://bioinf.uni-greifswald.de/augustus/
EVM v1.1.1	Haas et al., 2008	https://evidencemodeler.github.io/
BUSCO V3	Simão et al., 2015	https://busco.ezlab.org/v3
BLAST v2.9.0	Zhang et al., 2000	https://blast.ncbi.nlm.nih.gov/Blast.cgi
EMBOSS v6.6.0	Rice et al., 2000	http://emboss.sourceforge.net/
HMMER2 v3.3.1	Mistry et al., 2013	http://hmmer.org/
KaKs_calculator v2.0	Zhang et al., 2006	https://bigd.big.ac.cn/tools/kaks
MAFFT v7.407	Katoh and Standley, 2013	https://mafft.cbrc.jp/alignment/software/
RAxML v8.2.12	Stamatakis, 2014	https://cme.h-its.org/exelixis/web/software/raxml/index.html
MP-EST v2.0	Liu et al., 2010	http://faculty.franklin.uga.edu/lliu/mp-est
PAML v4.9e	Yang, 2007	http://abacus.gene.ucl.ac.uk/software/paml.html
r8s v1.8.1	Sanderson, 2003	https://sourceforge.net/projects/r8s/
PSMC v0.6.5	Li and Durbin, 2011	https://github.com/lh3/psmc
OrthoFinder v2.3.1	Emms and Kelly, 2019	https://github.com/davidemms/OrthoFinder
CAFE v3.1	De Bie et al., 2006	https://github.com/hahnlab/CAFE
PRANK v.170427	Löytynoja, 2014	http://wasabiapp.org/software/prank/
MrBayes v3.2.7a	Ronquist et al., 2012	https://nbisweden.github.io/MrBayes/download.html
ggplot2 v3.3.2	Wickham, 2016	https://ggplot2.tidyverse.org/
MCScanX	Wang et al., 2012	https://github.com/wyp1125/MCScanX
ANGeS v1.01	Jones et al., 2012	http://paleogenomics.irmacs.sfu.ca/ANGES/
Circos v0.69-9	Krzywinski et al., 2009	http://circos.ca/
LASTZ v1.04.03	Harris, 2007	http://www.bx.psu.edu/~rsharris/lastz/
UCSC tools	The University of California	https://genome.ucsc.edu/util.html
MULTIZ v012109	Blanchette et al., 2004	https://github.com/multiz/multiz
PHAST v1.5	Siepel et al., 2005	http://compgen.cshl.edu/phast/
VISTA v1.4.26	Frazer et al., 2004	http://genome.lbl.gov/vista/
bioedit v7.2	Tippmann, 2004	https://softfamous.com/bioedit/
HISAT2 v2.1.0	Kim et al., 2019	http://daehwankimlab.github.io/hisat2/
StringTie v2.0.6	Pertea et al., 2015	https://ccb.jhu.edu/software/stringtie/
preprocessCore v1.44.0	Ben Bolstad	http://bioconductor.org/packages/release/bioc/html/preprocessCore.html
edgeR v3.32.0	Robinson and Oshlack, 2010	https://bioconductor.org/packages/release/bioc/html/edgeR.html
MEGA v7.0	Kumar et al., 2016	https://www.megasoftware.net/
HyPhy v2.5.15	Pond et al., 2005	http://www.hyphy.org/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Wen Wang (wwang@mail.kiz.ac.cn).

Materials availability

This study did not generate any new unique reagents.

Data and code availability

The genome sequences, raw genome and transcriptome sequencing data for *Protopterus annectens* have been deposited at National Genomic Data Center (https://bigd.big.ac.cn/bioproject/) under the accession number PRJCA002950.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Source organisms

All animal experiments were approved by the Animal Research and Ethics Committee of the Institute of Hydrobiology, Chinese Academy of Sciences. The samples were collected legally and in accordance with the Animal Care and Use Ethics policy of Chinese Academy of Sciences. The African lungfish (*Protopterus annectens*) samples were purchased from an ornamental fish market in Guangzhou, China and then identified in the Institute of Hydrobiology, Chinese Academy of Sciences. Two samples were used for genome and transcriptome sequencing. The first sample "Lungfish.Sample01" was a male individual of ~2 years old collected in 2018. The sample "Lungfish.Sample02" was another male individual of ~2 years old collected in 2015. For the aestivation treatment, a total of nine individuals of ~2 years old were used. Six samples were randomly selected and fasted for 96 h and individually induced to aestivate in muddy substrata in vitreous tanks (L 35 cm × W 21 cm × H 28 cm) under conditions of 27–29°C and 85%–90% humidity as described by Loong et al. (Loong et al., 2008). Among them, there were 2 males and 2 females, but the gender of the other two individuals were not recorded. The three remaining samples were maintained in freshwater served as controls, with two males and one female. Aestivating and control lungfishes were euthanized on day 36 (6 days for induction and 30 days for maintenance) by blowing the head. For the *in situ* hybridization on lung tissues, samples from four species were collected. Samples of lungfish and alligator gar (*Atractosteus spatula*) were about two years old. Samples of mouse (*Mus musculus*, strain: C57BL/6) and zebrafish (*Danio rerio*, strain: AB) were three and ten months, respectively. We didn't identify the gender for these samples.

Cell lines

The HEK293T cells were purchased from National Collection of Authenticated Cell Cultures (Shanghai, China). The cell line was authenticated by short tandem repeat analysis. The cells were maintained in DMEM (GIBCO, USA) supplemented with 10% FBS (GIBCO) and 1% Antibiotic-Antimycotic (GIBCO) at 37°C, 5% CO_2.

METHOD DETAILS

DNA isolation and genomic sequencing

For the sample "Lungfish.Sample01," the DNA for sequencing was extracted from its muscle tissue. For long read sequencing, the libraries constructed with high-quality genomic DNA were sequenced on both GridION and PromethION platforms, resulting a total of 337 lanes of data. For short read sequencing, a paired-end library with an insert size of ~300 bp and a 100 bp paired-end read length was constructed and sequenced with the MGISEQ-2000 platform. For optical mapping, the Bionano Saphyr^{TM} system based on NanoChannel Array Technology was used. For Hi-C sequencing, the libraries constructed were sequenced on the BGI-Shenzhen MGISEQ-2000 platform with 150 bp paired-end read length. For the sample "Lungfish.Sample02," genomic DNA was extracted from its muscle tissue and the Illumina HiSeq 2500 platform was used for 127 bp and 250 bp paired-end read length sequencing. The short reads were filtered using fastp v0.19.4 (Chen et al., 2018) with default parameters.

Estimation of genome size

We used K-mer frequency-based and flow cell-based methods for genome size estimation. First, the K-mers were counted using Jellyfish v2.2.10 (Marçais and Kingsford, 2011) with the parameter "-C -m 51 -s 10000000000 -t 50." The output file was then used as the input for GenomeScope (Vurture et al., 2017) to estimate the genome size. Second, one African lungfish individual was anesthetized with MS-222 at a concentration of 50 mg/L and about 0.1 mL of blood sample was collected from the caudal artery by an injector with 0.2% heparin. Two microliters of blood cells were diluted in 1 mL cooled 0.01M PBS and then centrifuged at 94 g for 5 min to isolate leucocytes and erythrocytes. The red blood cells (RBCs) were re-suspended in PBS and diluted to 1.0 × 106 cell/mL for each sample. The RBCs were then re-suspended in 70% ethanol, fixed at 4°C overnight, washed and suspended in PBS. Next, 1 mL of propidium iodide (PI) staining solution (0.1% Triton X-100, 10 μg/mL PI, and 100 μg/mL DNase - free RNase A in PBS) was added, and the mixture kept in the dark at room temperature for 30 min. The same method was applied to process blood from both chicken (*Gallus gallus domesticus*) (genome size 1.25 pg) (Tiersch and Chandler, 1989) and grass carp (*Ctenopharyngodon idella*) (genome size 1 pg) (Wang et al., 2015), which were used as standards. The genome size was calculated from to the fluorescence value by the formula: Fish genome size in picograms = reference genome size × (E1/E2), where E1 was the fluorescence in the channel used for the lungfish sample and E2 was the fluorescence in the channel used for the reference species.

Genome assembly and chromosome anchoring

The software NextDenovo v1.0 (https://github.com/Nextomics/NextDenovo) was used for the self-correction of long reads sequenced with ONT platforms. Using the default parameters, the reads were aligned against themselves and the regions of overlap were compared to generate consensus sequences. The corrected reads were assembled with wtdbg2.huge (commit a98e6dc) (Ruan and Li, 2020) with parameters "-t 96 -k 0 -p 21 -AS 4 -K 0.05 -s 0.5 -L 10000." Contigs were polished with NextPolish v1.01 (Hu et al., 2020) with three rounds of alignment with long reads and three rounds of alignment with short reads. Then, the Bionano *de novo* optical map and the polished contigs were input into the Bionano Solve hybrid scaffold pipeline (https://bionanogenomics.com/wp-content/uploads/2018/04/30073-Bionano-Solve-Theory-of-Operation-Hybrid-Scaffold.pdf) and an AGP file was generated to guide the scaffolding of the FASTA file. Finally, the Hi-C short reads were aligned to the scaffolds with Juicer (Durand et al., 2016), and the anchoring was performed with 3D-DNA v180419 (Dudchenko et al., 2017). The Juicebox Assembly Tools v1.9.9 (Dudchenko et al., 2018) were then applied for the manual correction of the connections.

Genome quality evaluation

The 481 ultra-conservative elements (UCEs) identified in humans, rats, and rabbits that were longer than 200 bp were used in assessing the quality of the lungfish genome (Bejerano et al., 2004). The BLAT v36x2 (Kent, 2002) software package was applied to the genome of bichir and spotted gar with the sensitive parameters, "-minIdentity=60 -minScore=30 -minMatch=1 stepSize=8 -mask=lower," filtering out those sequences that were less than 75% matches and less than 50 bp in length, yielding 178 ultraconservative elements conserved in the Osteichthyes. These Osteichthyes-conserved UCEs were used to access the integrity of the lungfish genome. We performed the same assessment for the genomes of coelacanth, axolotl, Western clawed frog, and chicken. The sequenced long- and short- reads were then aligned back to the assembled genome to check the completeness of the genome using, respectively, minimap2 v2.17 (Li, 2018) and Burrows-Wheeler Aligner v0.7.17 (Li and Durbin, 2010), with the default parameters. Finally, to check the reliability of chromosomal anchoring, the physical length of each chromosome was measured from a photograph in a previous publication (Suzuki and Yamanaka, 1988) and the lengths of chromosomes were used for correlation analysis.

RNA preparation and sequencing

Both the above two lungfish individuals were used for transcriptome sequencing. RNA was extracted for all the samples using TRIzol (Invitrogen) and subsequently purified using a RNeasy Mini Kit (QIAGEN). For "Lungfish.Sample01," a total of 6 samples, gill, heart, liver, lung, muscle and kidney, were used for RNA isolation and sequenced separately on the MGISEQ-2000 platform. The RNA of these 6 samples was divided into three groups, each of which was processed via full-length transcriptome sequencing using the PacBio RS II platform. For "Lungfish.Sample02," 8 samples, gut, lung, liver, muscle, brain, skin, kidney and heart, were used for sequencing with the Illumina Hiseq 2500 platform. These samples were also divided into three groups and sequenced on the PacBio RS II platform.

For studying the aestivation, a total of 11 tissues were collected in six aestivated and three control individuals, including brain, spinal cord, heart, liver, lung, muscle, skin, gill, kidney, gut and eye. The collected samples were excised, frozen in pre-cooled freezing tubes in liquid nitrogen, and stored at −80°C until use. For each sample, the RNA was isolated and sequenced on the MGISEQ-2000 platform. Besides, the ovary samples from four individuals were also collected for short reads transcriptome sequencing on the MGISEQ-2000 platform.

Annotation of genome sequences

The repetitive sequences were annotated using both homology-based and *de novo* predictions. To enable better parallel computation and accelerate the annotation of repetitive sequences, here we used the scaffold rather than the chromosome-level genome assembly. First, the transposable elements were identified using RepeatMasker v4.0.7 (http://www.repeatmasker.org) and RepeatProteinMask v1.36 with the Repbase transposable element library (Jurka et al., 2005). Second, RepeatModeler v1.0.11 (http://www.repeatmasker.org/RepeatModeler.html) was used to construct a *de novo* transposable element library, which was then used to predict repeats with RepeatMasker v4.0.7. The divergence time of transposons was estimated using parseRM (https://github.com/4ureliek/Parsing-RepeatMasker-Outputs). As a complement, tandem repeats were predicted using TRF v4.09 (Benson, 1999). We also applied the same protocol to annotate the genome sequences of axolotl, coelacanth and Western clawed frog for further comparison.

For the annotation of the protein domains across genome, the HMMs model files for all the domains from the Pfam-A database were downloaded and searched in the genome of 40 species genomes (Table S4). Each genome was translated with 6-phase model using EMBOSS v6.6.0 (Rice et al., 2000), and HMMER2 v3.3.1 (Mistry et al., 2013) was used for the domains search with the default parameters. We also downloaded 266 vertebrate genomes from the NCBI, ENSEMBL databases (Table S3) for searching the KRAB and RVT_1 domains using the same procedure.

For the prediction of protein coding genes, we first assembled the transcriptome from the 14 samples of the two lungfish individuals (Lungfish.Sample01 and Lungfish.Sample02). The short-read transcriptomes were assembled with Trinity v2.8.5 (Grabherr et al., 2011) and the full-length transcriptomes were assembled with Iso-Seq2 from SMRT Link v5.1.0 (https://www.pacb.com/wp-content/uploads/2018-10-NA-UGM-Iso-Seq-Method.pdf) using the default parameters. The assembled transcripts with an open reading frame of at least 30 amino acids were aligned to the genome using BLAT with default parameters. We kept only the best alignment for each transcript and filtered out transcripts with coverage below 70% of the query length. Second, the annotated proteins of six species, including elephant shark, spotted gar, zebrafish, coelacanth, chicken and human, were downloaded from ENSEMBL (v96). We retained only the longest transcript from each gene as a representative and used BLAT to align the genes to the genome assembly. Alignments with coverage below 70% of the query length were again filtered out. The alignments were then grouped based on position on the reference genome. Groups for which there was at least one item of homology-based evidence and one item of transcript-based evidence, and groups for which transcript-based evidence was available from more than three independent samples, were retained. Afterward, for each group, the transcript with the best alignment was used for initial gene prediction with the software Exonerate v2.4.0 (https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate). We used Augustus v3.3.2 (Stanke et al., 2006) for model training using the initial gene sets and performed gene prediction within each group. Finally, EVM v1.1.1 (Haas et al., 2008) was used to integrate all evidence to produce the final gene sets. The completeness of the gene sets was assessed against both the vertebrate and tetrapod lineages with Benchmarking Universal Single-Copy Orthologs (BUSCO) (Simão et al., 2015).

Assessment of whole genome duplication (WGD)

The distribution of synonymous substitutions per site (Ks) within paralogs was used to examine the most recent WGD event in African lungfish. The protein sequences of African lungfish were aligned against themselves with BLAST v2.9.0 (Zhang et al., 2000) (-evalue 1e-10). When one gene and another gene were mutual best hits (excluding hits to themselves), they were identified as a paralog pair. The Ks was calculated via KaKs_calculator v2.0 (Zhang et al., 2006) for each paralog pair. For comparison, we also plotted the Ks distribution for coelacanth, axolotl, Western clawed frog, and spotted gar, which are species that are known to have no additional WGD other than the two rounds of genome duplication common to all vertebrates, and African clawed frog, which had an independent WGD event.

Phylogenetic relationship

To clarify the phylogenetic relationships of the African lungfish, we selected seven additional species that have experienced no recent WGD event, elephant shark, spotted gar, coelacanth, Western clawed frog, chicken, mouse and human (ENSEMBL 96), that together represent the jawed vertebrates. Two different datasets were used to infer the phylogenetic tree. First, the reciprocal best hit (RBH) method was adopted to identify the genes that are 1:1 orthologous among all the 8 species, which resulted in 5,149 ortholog pairs. For the concatenated tree, the protein sequences of each orthologous pair were aligned by MAFFT v7.407 (Katoh and Standley, 2013). The aligned sequences were then joined and used as input to RAxML v8.2.12 (Stamatakis, 2014) to infer a maximum likelihood (ML) tree, with the parameter: -f a -m PROTGAMMAAUTO -p 15256 -T 50 -x 271828 -N 100. For the species tree, the ML tree of each ortholog pair was inferred with RAxML as above and the ML trees were collected to build the species tree with MP-EST v2.0 (Liu et al., 2010). To estimate the divergence time, we extracted the 4-fold sites from the 5149 ortholog pairs and estimated the divergence time using MCMCTree in the PAML v4.9e (Yang, 2007) package with the calibration based on five fossil records (> 460 Ma for jawed vertebrate ancestor (Coates et al., 2018; Sansom et al., 2012); > 425 Ma for bony fish ancestor (Zhu et al., 2009); > 419 Ma for the appearance of lungfishes (Benton et al., 2015; Zhu and Fan, 1995)); > 318 Ma for birds and mammals ancestor (Benton et al., 2015); > 61.6 Ma for the crown Euarchontoglires node (Benton et al., 2015)). We then estimated the mutation rate of each lineage with r8s v1.8.1 (Sanderson, 2003) using the same dataset and fossil record as above and reconstructed the dynamic effective population size of African lungfish with PSMC v0.6.5 (Li and Durbin, 2011), using a generation time of 3 years (Dunbrack et al., 2006).

In addition, the protein sequences of the eight species were also clustered with OrthoFinder v2.3.1 (Emms and Kelly, 2019). The gene family expansion and contraction was evaluated used CAFE v3.1 (De Bie et al., 2006) with results from the OrthoFinder pipeline. A conditional P value was calculated for each gene family, and families with conditional P values lower than 0.05 were considered to have had a significantly accelerated rate of expansion or contraction.

Estimation of increment of intron sizes

The gene sets of seven species, including the African lungfish, the axolotl (Nowoshilow et al., 2018), and human, chicken, coelacanth, spotted gar and elephant shark from ENSEMBL 96, were collected for generation 1:1 orthologous dataset using the reciprocal best hits methods. The genes without introns were excluded. For each ortholog gene pairs, the initial intron length was estimated by calculating the median intron length of the five species including human, chicken, coelacanth, spotted gar and elephant shark. The increment of intron sizes was estimated by subtracting the initial intron length from the intron length of the African lungfish or the axolotl. The expansion rate of intron size was estimated by dividing the intron length of the African lungfish or the axolotl by the initial intron length. In order to reveal the correlation between the different factors, the linear fitting and curve fitting were done by the function ‘geom_smooth’ in the R package ‘ggplot2’ v3.3.2 (Wickham, 2016).

Reconstruction of ancestral karyotype

A total of six species were used, including white-spotted bamboo shark (Zhang et al., 2019), bichir (Bi et al., 2021), spotted gar, African lungfish, Western clawed frog (Hellsten et al., 2010), and chicken (Warren et al., 2017) were used for reconstruction of the ancestral karyotype. Chicken was adopted as a reference genome and BLAST was used for two-by-two interspecies comparisons and reciprocal best hits to obtain a set of genes homologous between species. The corrected posterior binomial test (q-value < 0.05, number of homologous genes > = 20) was applied to identify chromosomes that were homologous among species. The default parameters of MCScanX (Wang et al., 2012) were used for the identification of inter-chromosomal co-linear blocks. Finally, ANGeS v1.01 (Jones et al., 2012) was used to construct the ancestral karyotype. Interspecies co-collinearity was displayed using Circos v0.69-9 (Krzywinski et al., 2009).

Analysis of conserved elements

First, the genome sequences from human (GRCh38) and spotted gar (LepOcu1) were used as references, to construct 12-way whole genome alignments (WGAs) of human (GRCh38), mouse (GRCm38), anole lizard (AnoCar2.0), chicken (Gallus_gallus-5.0), Western clawed frog (JGI_4.2), African lungfish, coelacanth (LatCha1), spotted gar (LepOcu1), alligator gar, bichir (Bi et al., 2021), elephant shark (Callorhinchus_milii-6.1.3) and brownbanded bamboo shark (GCA_003427335.1). The pairwise alignment was conducted using LASTZ v1.04.03 (Harris, 2007) with the parameters: H = 2000, K = 2200, L = 6000, Y = 3400. The UCSC tools were applied for 'chaining' and 'netting' to ensure that all the alignments were non-overlapped. The roast program in MULTIZ v012109 (Blanchette et al., 2004) was then used to construct WGAs. Second, to identify highly-conserved elements (HCEs) that originated in various lineages, PhyloFit from the PHAST v1.5 (Siepel et al., 2005) package was used to construct non-conserved and conserved models with the four-fold degenerate sites and the first codons from coding sequences. The conserved elements in each 12-way WGA were defined using the parameters: target-coverage = 0.3, expected-length = 45, most-conserved = true, score = true, along with the above models. Subsequently, in these conserved alignments, we defined two types of HCEs, the ancient HCEs with spotted gar as reference, which required a high level of similarity (70% for every species) within cartilaginous fish and ray-finned fish, and the tetrapods-HCEs using human as reference, which required a high similarity (70% for every species) within tetrapods. The HCEs that had been gained in tetrapods were extracted from the WGAs with human as reference. The HCEs that had been lost from the tetrapods were extracted from the WGAs with spotted gar as reference. In addition, for the HCEs that are presented in all the 12 species mentioned above, we specifically identified the HCEs with fixed insertion/deletions/linked substitutions (with more than 6bp) between tetrapods, the African lungfish, the coelacanth, ray-finned fishes and cartilaginous fishes. The alignments of HCEs referred to in the manuscript were manually checked and plotted with VISTA v1.4.26 (Frazer et al., 2004) and bioedit v7.2 (Tippmann, 2004).

Identification of new genes

The HCEs located in the coding regions that gained by the lineage leads to the LCA of lungfishes were used for searching possible new genes. For each HCE in the coding regions, the gene that containing this HCE was used for searching homologs with BLAST v2.9.0 in the gene sets of 12 bony fishes that are mentioned in the above section. We also performed manually gene annotation in those species to avoid omissions in their general feature format (GFF) files. If we were completely unable to find its homologous in coelacanth, ray-finned fishes and cartilaginous fishes, the gene was considered as a *de novo* originated gene. Otherwise, all these homology sequences were aligned together with the software PRANK v.170427 (Löytynoja, 2014), and a phylogeny trees was constructed with MrBayes v3.2.7a (Ronquist et al., 2012), using the parameter: "lset nst=6 rates=invgamma Ngen=1000000 Nruns=2." When the sequences from a certain of sarcopterygian lineage has paralog(s) but not in other outgroup species on the phylogenetic tree, the gene was considered as a new duplicated gene in that lineage, with the closet paralog identified as its parent gene. We also further manually checked the syntenic relationship of these genes in bony fishes to validate the duplication events.

Comparative transcriptome analysis

For identifying genes differentially expressed across species during water-to-land transition. The RNA-Seq reads from nine tissues of five species (Bi et al., 2021) (Table S1) were mapped to their own genomes with HISAT2 v2.1.0 (Kim et al., 2019). For each species, the expression level (RPKM) of genes was calculated with StringTie v2.0.6 (Pertea et al., 2015). To construct the inter-species expression matrix, we applied a relaxed RBH method. Using mouse as the reference species, we first identified the mouse genes that had 1:1 homologs in each of the other four species separately, using the RBH method. Then, if a mouse gene had a 1:1 homolog in all four other species, we treated the corresponding five genes as a homologous gene set. We eventually obtained an expression matrix with a total of 7,870 homologous gene pairs and 220 samples across five species. The expression matrix was quantile normalized using the "preprocessCore" v1.44.0 in R (https://github.com/bmbolstad/preprocessCore). For each species, genes showing tissue-specific expression were identified with the *Tau* method (Yanai et al., 2005). If a gene had a *Tau* value larger than 0.8 and the expression level for a tissue was more than half of that in the tissue with the highest expression, the gene was then identified as being tissue-specifically expressed in that tissue. For each tissue, we focused on three classes of genes that had undergone nodal alterations in tissue-specific expression (TSE): genes that showed TSE in ray-finned fish and the African lungfish; genes that had TSE in the African lungfish and tetrapods, and genes that showed TSE in tetrapods alone. For the identified genes, we manually checked their corresponding expression levels in zebrafish (data from Bi et al., 2021).

To investigate the expression pattern of aestivation state of the African lungfishes, we aligned the RNA-seq reads from the 11 tissues of 6 aestivated and 3 non-aestivated African lungfish individuals to the lungfish genome with HISAT2 v2.1.0 (Kim et al., 2019). The gene expression level of each gene was with StringTie v2.0.6. The differentially expressed genes for each tissue were identified using an exact test provided in the edgeR v3.32.0 software package, where genes with fold change greater than 4 and false discovery rate values less than 0.01 (Robinson and Oshlack, 2010).

In situ hybridization

The lungs from mouse and the African lungfish (*P. annectens*), and swim bladder from alligator gar and zebrafish were sampled and fixated with 4% paraformaldehyde (PFA) at 4°C for 2-12 h. The tissues were next dehydrated in gradient ethanol, dipped in paraffin and finally embedded in paraffin wax. Then the paraffin block sectioned (5 μm), and mounted on slides. After the sample preparation, the tissue slides first pass sequentially in xylene, ethanol and DEPC water, and then were boiled in repairing liquid for 5 min and cooled naturally. Followingly, the samples on the slides were circled by the Pap Pen and then digested with 20 μg/mL proteinase K at 37°C for 20 min. After a brief wash with purified water, the slides were washed in PBS 3 times for 5 min each. To block the endogenous peroxidase, 30% methanol-H_2O_2 was added to the sample and incubated for 15 min at room temperature in the dark and then washed in PBS 3 times for 5 min each. Subsequently, a total of 8 probes were designed for hybridization, including: three copies of the African lungfish *SftpC*: 5′-GATAGTATTCAGATCCTCTCTTCTTGCCTCCCCTT-3′, 5′-CACTATCCCACTTCATCACAAGGC ACAAATCTC-3′, and 5′-GCGACCATTACCGTTTCCTTGTTTAGCCC-3′; mouse *Slc34a2*: 5′-CAGTCTTGGCTACAGGAGTCCCGT TGTCATT-3′, African lungfish *Slc34a2*: 5′-CAGATGAA GTCGCCATTATCCCAGCAGA-3′, alligator gar *Slc34a2*, 5′-GCCAACAGCCA GATCCGACAAAGA AGAGT-3′, two copies of zebrafish *Slc34a2*: 5′-CTGGAGAGTCGTGTTTGAGGCTGTGTGG-3′ and 5′-CGTCG GGGTTTGGGGCTCATGTTTGTCC-3′. For each probe, samples from corresponding species were prehybridized in a prehybridization solution at 37°C for 1 h and then hybridized in a hybridization solution containing the digoxigenin (DIG)-labeled RNA probe at 37°C for 12-16 h. The hybridized samples were then washed in a solution of 2 × SSC at 37°C for 10 min, 1 × SSC at 37°C for 5 min twice, and 0.5 × SSC at room temperature for 10 min, respectively. After blocking in BSA for 30 min at room temperature, the slides were incubated with anti-DIG-AP antibody (Jackson) at 37°C for 50 min and then washed in PBS 3 times for 5 min each. The FISH hybridization signal was stained by the anti-DIG antibody-conjugated peroxidase and tyramide signal amplification (FITC-TSA)/(DAPI). The images were acquired on a fluorescence microscope (NIKON Eclipse ci, JAPAN).

Dual-luciferase assay

The chemically synthesized fragments of Mus-*Hoxc10* (Ensembl: Chr15: 102964014-102964265, Chr15: 10294797-10295065), Mus-*Pax1* (Ensembl: Chr2: 147340589-147341041), Mus-*Foxp1* (Ensembl: Chr6: 98,971,277-98,971,646, Chr6: 99,003,798-99,004,060, Chr6: 98,938,098-98,938,454), Lungfish-*Foxp1* (Chr7:1,386,992,832-1,386,994,080, Chr7:1,387,112,483-1,387,113,686), Mus-*Foxp2* (Ensembl: Chr6: 15,377,659-1,538,036), Mus-*Nrp1* (Ensembl: Chr8: 128424884-128425508), and Lungfish-*Nrp1* (Chr3:2182603878-2182605198) were inserted into the pGL4.23 vector (Promega, USA) and were sequenced (TsingKe Biotech, China). The sequence information of inserted fragments was listed in Table 1. The cells were seeded in 24-well plates. Twenty-four h later, 100 ng constructed pGL4.23 vectors were co-transfected with 5 ng pRL-CMV as the internal control into HEK293T cells using Turbofect reagent (ThermoFisher Scientific, USA). Cellular lysates were collected 36 h after transfection using passive lysis buffer. Luciferase activity was measured using the Dual-Luciferase Reporter Assay System (Promega) by a Multimode microplate reader (Spark Tecan, Switzerland). Light output from transcriptional activity was divided by the output from Renilla luciferase activity to normalize the samples.

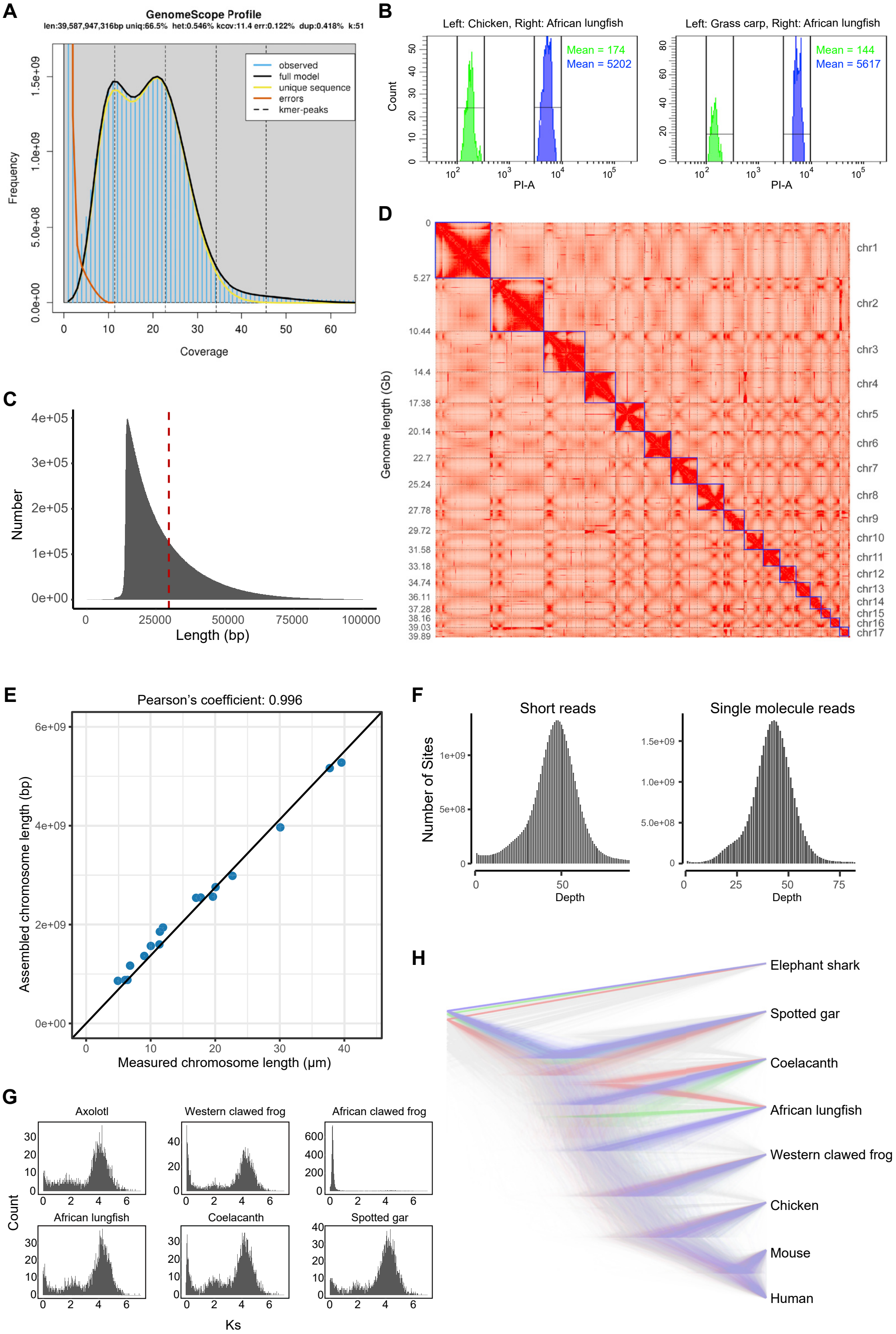
Evolution of tooth related genes

A total of 51 genes were collected from a previous paper (Bei, 2009) for examination on 18 species, including elephant shark, zebrafish, stickleback, medaka, spotted gar, coelacanth, Western clawed frog, chicken, platypus, elephant, mouse and human from Ensembl 96, brownbanded bamboo shark (GCA_003427335.1) and common lizard (GCA_011800845.1) from NCBI, bichir and alligator gar (Bi et al., 2020), the African lungfish and the Australian lungfish (*Neoceratodus forsteri*). The gene set of the Australian lungfish (*N. forsteri*) was assembled from the transcriptome data (SRR8131642, SRR3632078) downloaded from the NCBI SRA database using Trinity v2.8.5. For each gene, the protein sequences were collected, aligned by prank and a phylogenetic tree was reconstructed by RAxML v8.2.12 to ensure they are ortholog genes. The alignment of the protein sequences was visualized by the software MEGA v7.0 (Kumar et al., 2016) to manually check if there exist shared alternations in two lungfish species on the regions that are conserved in other species. Lastly, for the identified genes, we applied the aBSREL model (Smith et al., 2015) of the HyPhy v2.5.15 package (Pond et al., 2005) to estimate the selection pressure on the common ancestor of lungfishes.

QUANTIFICATION AND STATISTICAL ANALYSIS

Quantification approaches and statistical analyses used in the genome sequencing and assembly, genome quality assessment, evolutionary analysis and comparative transcriptome analysis can be found in the relevant sections of the Method Details.

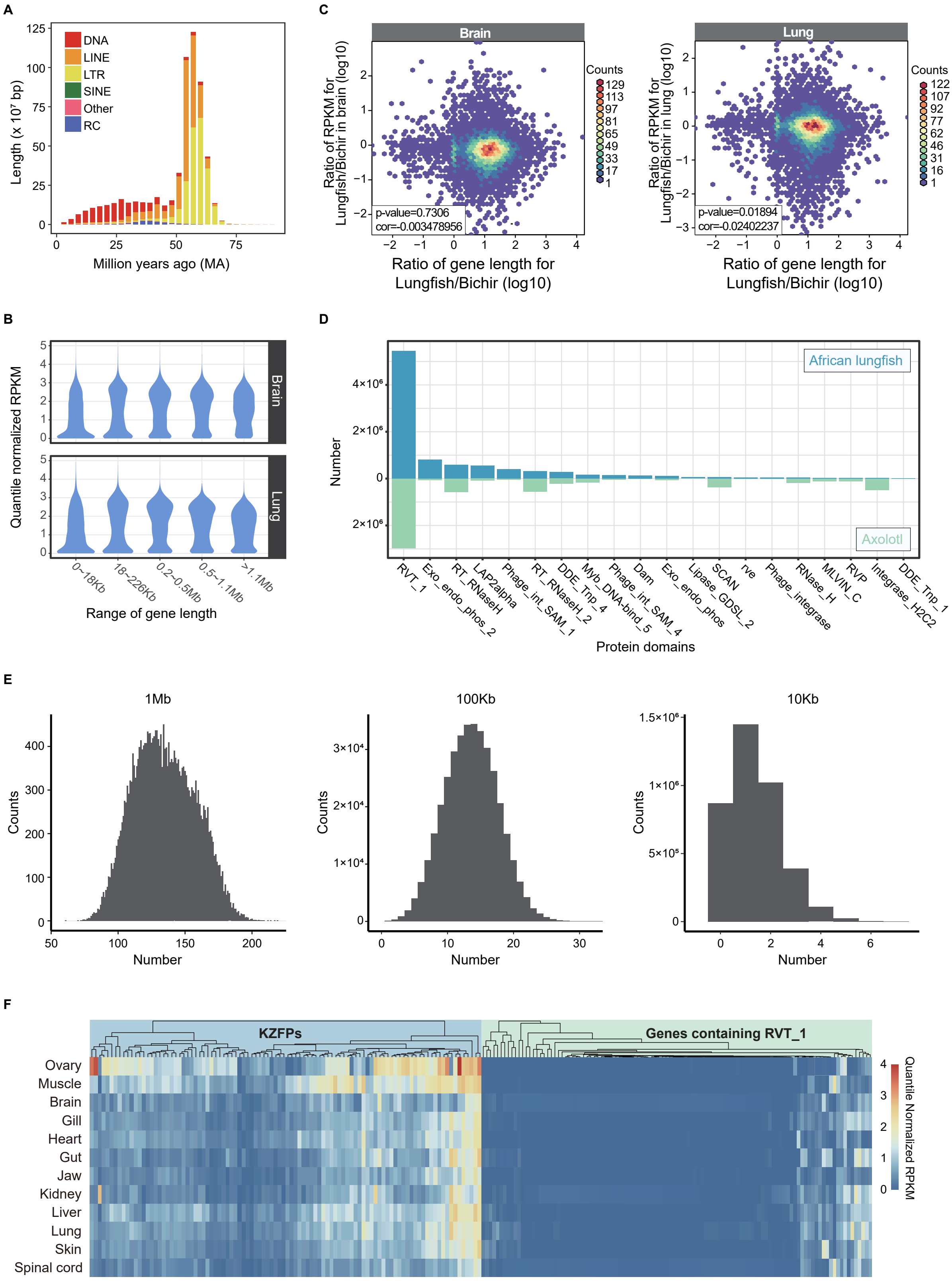
Supplemental Figures



(legend on next page)

Figure S1. Genome assembly and phylogenetic relationship of the African lungfish (*Protopterus annectens*), related to Figure 1

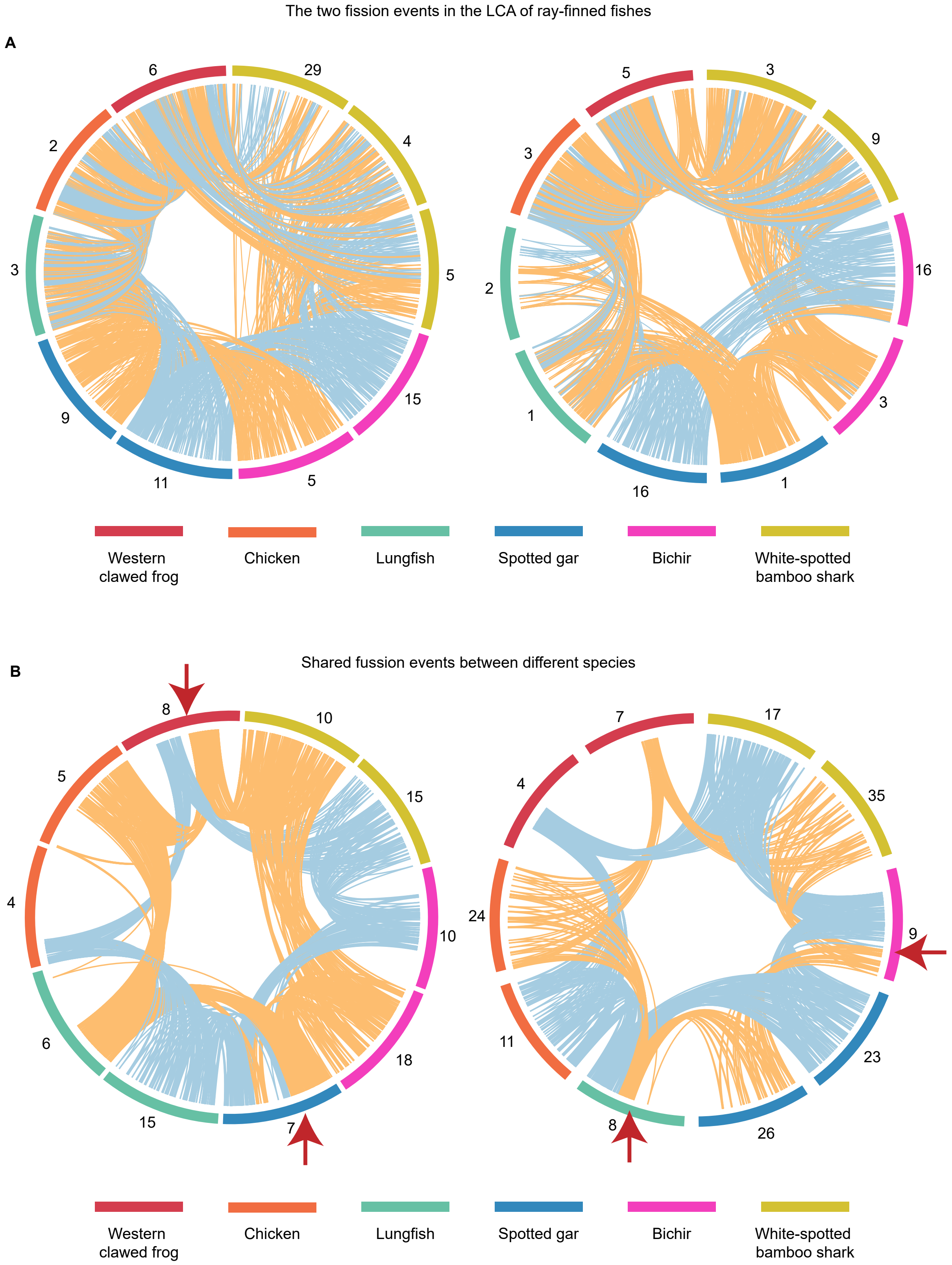
(A) Estimated genome size of African lungfish based on K-mer analysis. The result of Genome Scope can be viewed at the following URL: (http://qb.cshl.edu/genomescope/analysis.php?code=nxoFwyt06pxboISs2lSC). (B) Flow cytometry histograms for chicken, grass carp, and African lungfish erythrocytes on the basis of PI fluorescence dye. Left: Chicken and African lungfish contrast. Right: Grass carp and African lungfish contrast. (C) Distribution of ONT reads length. The red-dotted line indicates the N50, 29,717. (D) The interaction map of Hi-C data. (E) Chromosomes lengths consistent with observed physical lengths. (F) The depth of long and short reads mapped to scaffolds. Left: The depth distribution of short reads generated by MGISEQ 2000. Right: The depth distribution of single molecule reads generated by ONT platform. (G) Ks distribution between paralogous of the axolotl, coelacanth, African lungfish, western clawed frog, African clawed frog, and spotted gar. (H) The heterogeneity of gene trees. The blue, red, and green trees represent the most, second most, and third most topological structures, respectively.



(legend on next page)

Figure S2. Genome expansion and its effects on the African lungfish, related to Figure 2

(A) The estimated amplification time of repeat sequences. (B) Violin map of the relationship between gene length and expression level. The gene set of the African lungfish was divided into five groups according to length from smallest to longest. Here the brain and lung were selected as representative tissues. (C) The relationship between gene length and expression level. The organ brain and lung were selected as representatives. (D) The most abundant protein domains in the African lungfish and the axolotl. (E) The distribution of RVT_1 domains in different sizes (left: 1Mb, middle: 100Kb, right: 10Kb) of non-overlap sliding window in the African lungfish genome assembly. (F) The expression pattern of genes containing KRAB domain and the genes containing RVT_1 domain in different tissues.



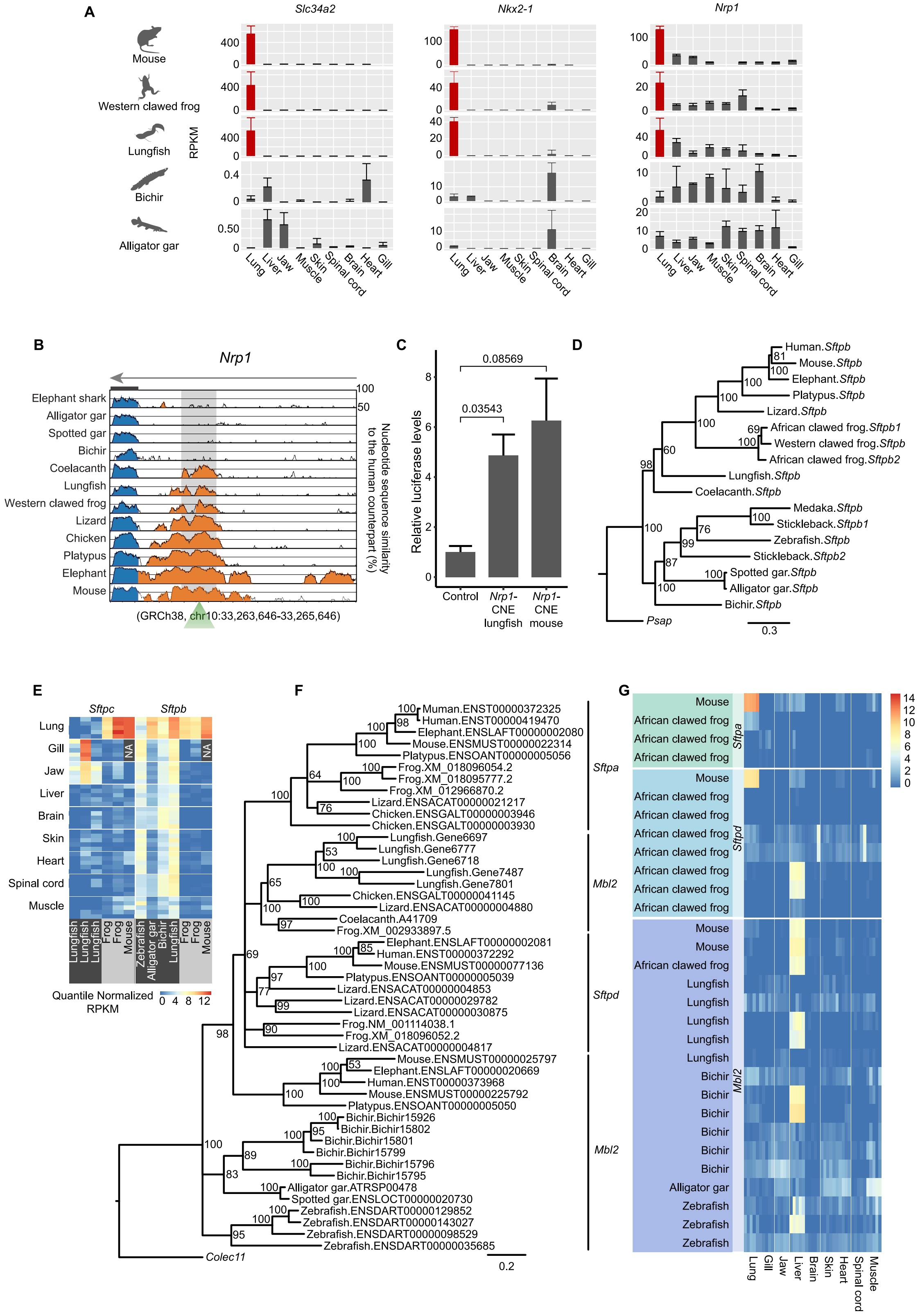
The two fission events in the LCA of ray-finned fishes

Shared fussion events between different species

(legend on next page)

Figure S3. Chromosomal fission and fusion events during the evolution of bony fishes, related to Figure 3

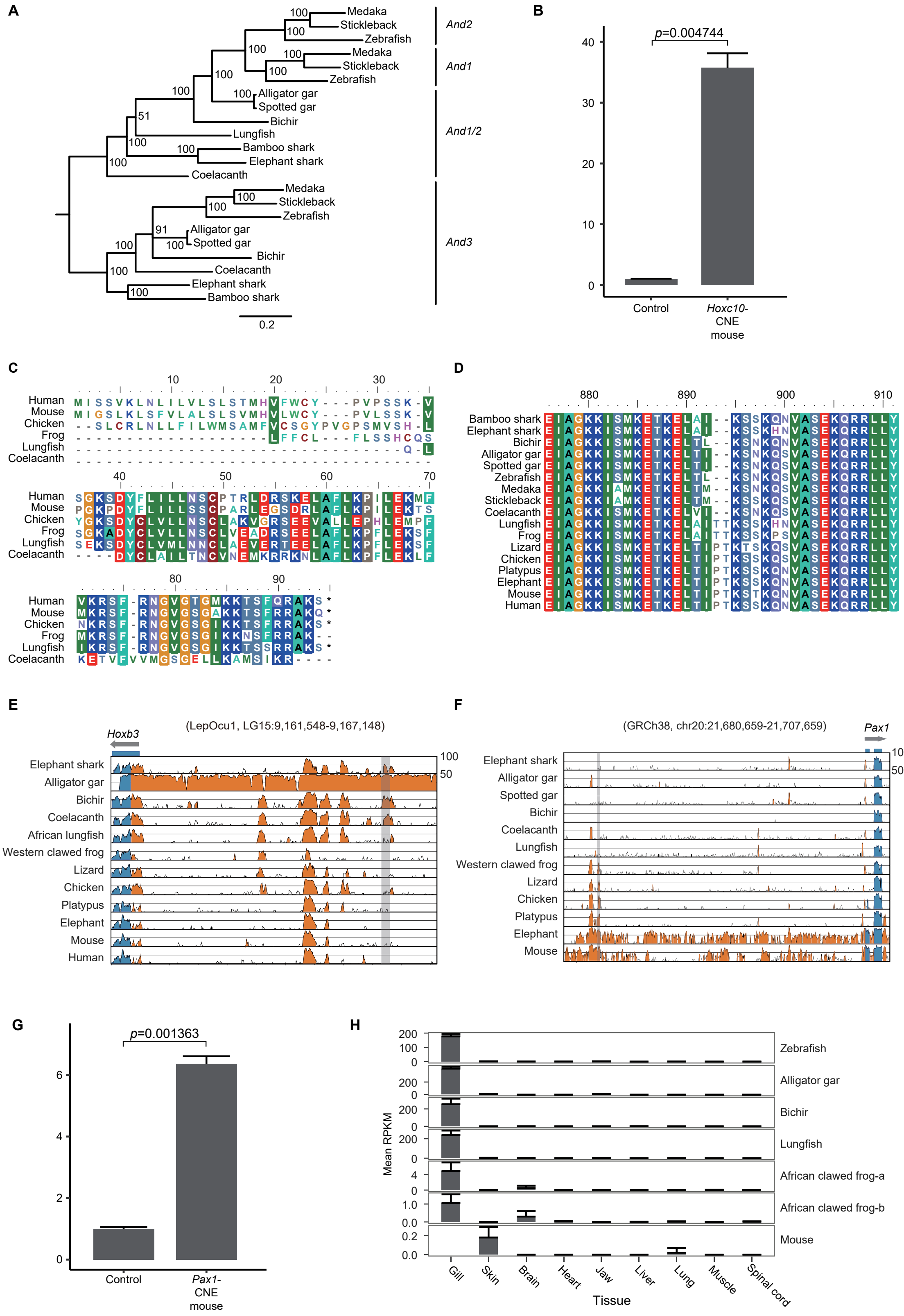
(A) The two fission events for the common ancestor of ray-finned fish. (B) The shared fusion events between western clawed frog and spotted gar; and between lungfish and bichir. The red arrows indicate the fusion points in corresponding species.



(legend on next page)

Figure S4. Genomic evolution of the respiration system during the water-to-land transition, related to Figure 4

(A) The expression Patterns of *Slc34a2, Nkx2-1* and *Nrp1* in five species. The red bar shows the genes that were tissue-specifically expressed in the lung. (B) Appearance of a CNE in the intron region of *Nrp1*. (C) Results of *in vitro* reporter assays suggest the CNE in the upstream of *Nrp1* could significantly improve the expression level of luciferases. The *Nrp1*-CNE lungfish refers to the sequence of that CNE in the African lungfish genome, while the *Nrp1*-CNE mouse refers to that in the mouse genome. Significance was tested by Student's t test and data are represented as mean ± SEM. (D) The Bayesian tree of the gene *Sftpb* across vertebrates. The numbers in each label indicate the probabilities in percent. (E) The expression pattern of *Sftpc* in three species and *Sftpb* in five species. There are three copies of *Sftpc* in the African lungfish, and two copies of *Sftpc* and *Sftpb* in the African clawed frog. "NA" in the heatmap indicates that data are not available. (F) The Bayesian tree of *Sftpa*, *Sftpd*. The numbers in each label indicate the probabilities in percent. (G) The expression pattern of *Sftpa*/*Sftpd*/*Mbl2* in different species.



(legend on next page)

Figure S5. Genetic changes related to locomotion, anxiolytic ability, and Pharyngeal remodeling, related to Figures 5 and 6

(A) The progressive loss of actinotrichia proteins in the African lungfish. The numbers in each label indicate the probabilities in percent. (B) Results of *in vitro* reporter assays suggest the CNEs in the upstream of *Hoxc10* could significantly improve the expression level of luciferases. Significance were tested by Student's t test and data are represented as mean ± SEM. (C) The protein alignments of *Nps*. The amino acids of coelacanth were translated from its corresponding precursor sequence. (D) The 2 AAs insertion shared by lungfish and tetrapod in the gene *Arfgef1*. (E) The Vista plot of the lost CNE upstream of *Hoxb3* that are conserved in cartilaginous fishes, ray-finned fishes are lost in tetrapods. (F) The Vista plot exhibit a CNE upstream of *Pax1* that gained by tetrapods. (G) Results of *in vitro* reporter assays suggest the CNEs in the upstream of *Pax1* could significantly improve the expression level of luciferases. Significance was tested by Student's t test and data are represented as mean ± SEM. (H) The expression of *Gcm2* in different bony fish lineages. The expression level of the gene *Gcm2* is much lower in the gill of African clawed frog than that in ray-finned fish and lungfishes. Data are represented as mean ± SEM.

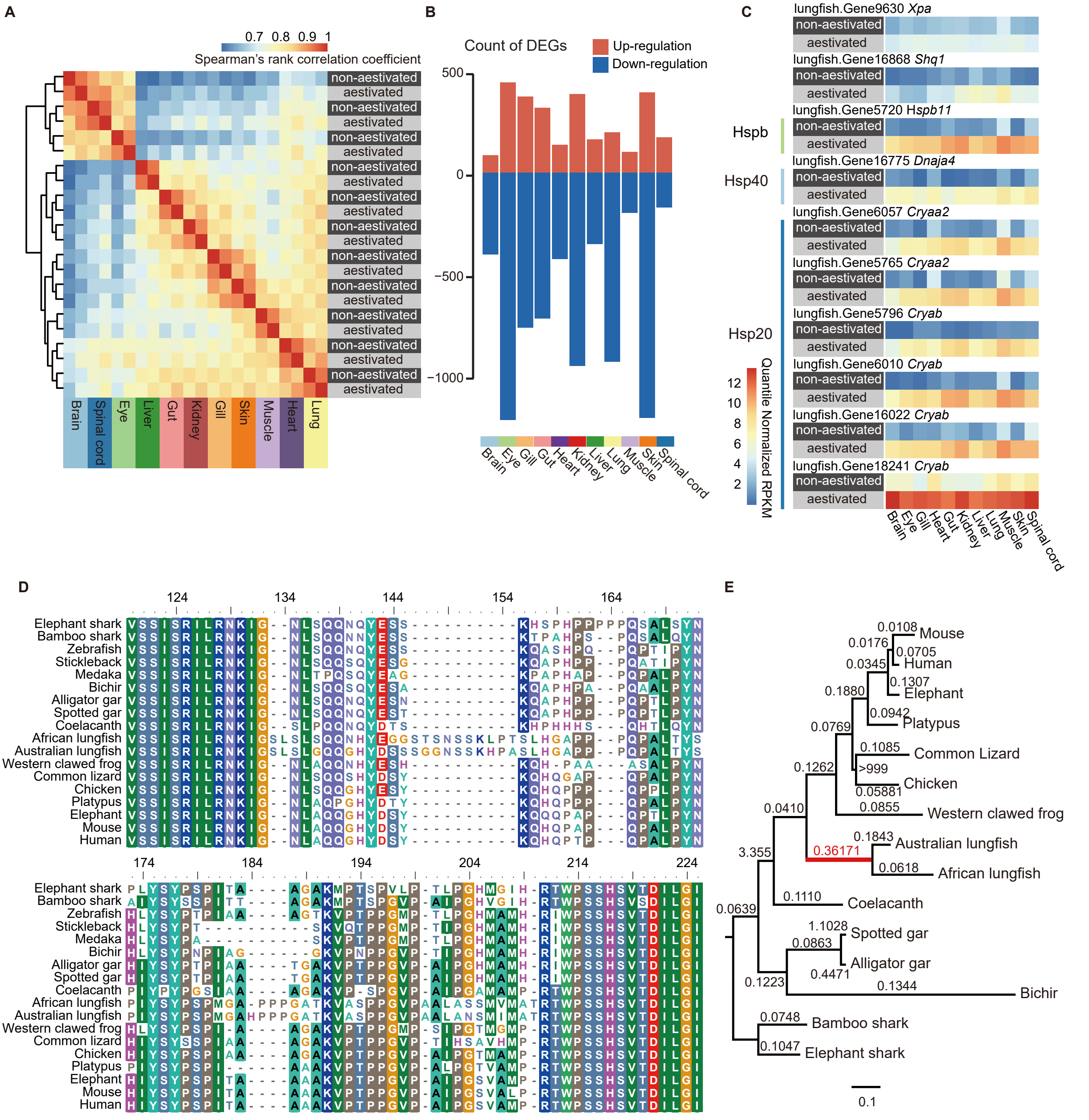


Figure S6. The aestivation transcriptome and specific alternations in the Pax9 gene of the African lungfish, related to STAR methods

(A) The Spearman's correlation between non-aestivated and aestivated samples in 11 tissues. (B) The number of Differentially expressed genes (DEGs). The red bars indicate genes with upregulation in aestivated samples, while the blue bars indicate genes with downregulation in aestivated samples. (C) The shared upregulation genes in multiple tissues of aestivated samples. (D) The specific mutations in the *Pax9* gene in lungfishes, including the African lungfish and the Australian lungfish. (E) The distribution of dN/dS (nonsynonymous substitution rate/synonymous substitution rate) in different nodes of the *Pax9* gene. The common ancestor of lungfishes (colored in red) were set as the foreground, and was significantly positively selected (p = 0.038).