nature

https://doi.org/10.1038/s41586-021-03198-8

Accelerated Article Preview

Giant lungfish genome elucidates the conquest of land by vertebrates

Received: 13 July 2020

Accepted: 6 January 2021

Accelerated Article Preview Published online 18 January 2021

Cite this article as: Meyer, A. et al. Giant lungfish genome elucidates the conquest of land by vertebrates. *Nature* https://doi.org/10.1038/s41586-021-03198-8 (2021).

Open access

Axel Meyer, Siegfried Schloissnig, Paolo Franchini, Kang Du, Joost Woltering, Iker Irisarri, Wai Yee Wong, Sergej Nowoshilow, Susanne Kneitz, Akane Kawaguchi, Andrej Fabrizius, Peiwen Xiong, Corentin Dechaud, Herman Spaink, Jean-Nicolas Volff, Oleg Simakov, Thorsten Burmester, Elly M. Tanaka & Manfred Schartl

Article

Giant lungfish genome elucidates the conquest of land by vertebrates

https://doi.org/10.1038/s41586-021-03198-8

Received: 13 July 2020

Accepted: 6 January 2021

Published online: 18 January 2021

Open access

Axel Meyer[1,12,13]✉, Siegfried Schloissnig[2,12], Paolo Franchini[1,12], Kang Du[3,4,12], Joost Woltering[1,12], Iker Irisarri[5,11], Wai Yee Wong[6], Sergej Nowoshilow[2], Susanne Kneitz[7], Akane Kawaguchi[2], Andrej Fabrizius[8], Peiwen Xiong[1], Corentin Dechaud[9], Herman Spaink[10], Jean-Nicolas Volff[9], Oleg Simakov[6,13]✉, Thorsten Burmester[8,13]✉, Elly M. Tanaka[2,13]✉ & Manfred Schartl[3,4,13]✉

Lungfishes belong to lobe-fined fish (Sarcopterygii) that in the Devonian 'conquered' land and gave rise to all land vertebrates, including humans[1–3]. We determined the largest chromosome-quality animal genome, the Australian lungfish, *Neoceratodus forsteri*. Its vast size (~14x of human) is attributable mostly to huge intergenic regions and introns with high repeat content (≈90%) whose components resemble tetrapods more (mostly LINE elements) than ray-finned fish. The lungfish genome continues to expand (its TEs are still active) independently and by different mechanisms than enormous salamander genomes. Synteny to other vertebrate chromosomes of 17 fully assembled macrochromosomes is maintained just as its conserved ancient homology of all microchromosomes to the ancestral vertebrate karyotype. Phylogenomic analyses ascertained that lungfish occupy an evolutionary key-position as closest living relatives to tetrapods, underscoring their importance for understanding innovations associated with terrestrialization[4,5]. Preadaptations to living on land include gaining of limb-like expression of developmental genes such as *hoxc13* and *sall1* in their lobed fins. Increased rates of evolution and duplication of genes associated with obligate air-breathing such as lung surfactants and the expansion of odorant receptor gene-families that detect airborne odours contribute to their tetrapod-like biology. These findings advance our understanding of this major transition during vertebrate evolution.

Lungfish (Dipnoi) share with the land-dwelling vertebrates the ability to breathe air though lungs that are homologous to ours. Since their discovery in the 19th century, lungfish attracted scientific interest and were initially thought to be amphibians[6,7]. We now know that they are more closely related to tetrapods than to ray-fined fish. Of the only six extant lungfish species four live in Africa, one in South America, and *Neoceratodus forsteri* in Australia. Lungfish appeared in the fossil record in the Devonian ~400 million years ago (Ma)[1]. Some consider lungfish as "living fossils" since their morphology barely changed over eons, for example >100 Ma fossils in Australia strongly resemble the surviving species – one of the oldest genera discovered exactly 150 years ago[2]. Due to its archaic characters, such as body shape, large scales, and paddle-shaped fins the Australian is the most archetypical extant lungfish. The South American and African lungfish lost their scales secondarily and simplified their fin morphology into thin filaments albeit showing the alternating gaits typical of terrestrial locomotion.

Together with the coelacanth and tetrapods lungfish are members of the Sarcopterygii (lobe-finned "fish") but due to the short branch that separates these three ancient lineages it remained difficult to resolve their relationships. Developments of powerful DNA-sequencing and computational methods allow now to revisit this long-standing evolutionary question using whole-genome-derived datasets with more robust orthology inferences. Recent analyses using large transcriptomic datasets tended to support the hypothesis that lungfish are the closest living relatives of tetrapods[4,5]. Lungfish hence are crucial for understanding the evolution and the preadaptations accompanying the transition of vertebrate life from water to land. This major evolutionary event required a number of evolutionary innovations including airbreathing, limbs, posture, prevention of desiccation, nitrogen excretion, reproduction, and olfaction. Lungfish are known to have the largest animal genome, but the mechanisms that led to and maintained their genome sizes are still poorly understood. Therefore,

[1]Department of Biology, University of Konstanz, Konstanz, Germany. [2]Research Institute of Molecular Pathology (IMP), Campus-Vienna-Biocenter, Vienna, Austria. [3]Developmental Biochemistry, Biocenter, University of Würzburg, Würzburg, Germany. [4]The Xiphophorus Genetic Stock Center, Texas State University, San Marcos, TX, USA. [5]Department of Biodiversity and Evolutionary Biology, Museo Nacional de Ciencias Naturales (MNCN-CSIC), Madrid, Spain. [6]Department of Neuroscience and Developmental Biology, University of Vienna, Vienna, Austria. [7]Biochemistry and Cell Biology, Biocenter, University of Würzburg, Würzburg, Germany. [8]Institut für Zoologie, Universität Hamburg, Hamburg, Germany. [9]Institut de Genomique Fonctionnelle, Ecole Normale Superieure, Universite Claude Bernard, Lyon, France. [10]Faculty of Science, Universiteit Leiden, Leiden, The Netherlands. [11]Present address: Department of Applied Bioinformatics, Institute for Microbiology and Genetics, University of Goettingen, and Campus Institute Data Science, Goettingen, Germany. [12]These authors contributed equally: Axel Meyer, Siegfried Schloissnig, Paolo Franchini, Kang Du, Joost Woltering. [13]These authors jointly supervised this work: Axel Meyer, Oleg Simakov, Thorsten Burmester, Elly M. Tanaka, Manfred Schartl.
✉e-mail: axel.meyer@uni-konstanz.de; oleg.simakov@univie.ac.at; thorsten.burmester@uni-hamburg.de; elly.tanaka@imp.ac.at; phch1@biozentrum.uni-wuerzburg.de

the Australian lungfish might provide insights both into tetrapod innovations and evolution and structure of giant genomes.

Genome sequencing, assembly and annotation

The largest animal genome sequenced so far is the 32Gb[8] genome of the axolotl salamander (*Ambystoma mexicanum*). To overcome the challenges to sequence and assemble the even larger genomes of lungfish we used long and ultra-long read Nanopore technology generating 1.2Tb in three batches: 601Gb with N50 read-length of 9kb, 532Gb with N50 of 27kb, and 1.5Gb with N50 of 46kb from a juvenile Australian lungfish. These were assembled into contigs using our further developed MARVEL assembler (Extended Data Fig. 1a, see Methods). This yielded a 37Gb assembly with N50 contig size 1.86Mb (Supplementary Table 1). To correct for indels, gaps, single-nucleotide polymorphisms (SNPs) and small local misalignments in the primary assembly we used 1.4Tb DNA and 499.8Gb RNA Illumina reads. The genome correction DNA data, sequenced at >30x coverage, were used to estimate genome size through frequencies of *k*-mers (Extended Data Fig. 2). The high completeness of the 37Gb assembly was ascertained in that 88.2% of the DNA and 84% of the RNA-seq reads aligned to the genome, estimating a total genome size of 43Gb (~30% larger than the axolotl[8]). This matches the *k*-mer value but is smaller than predicted by flow cytometry (52Gb[9]) and Feulgen photometry (75Gb[10]).

Next contigs were scaffolded using 271Gb Illumina Hi-C PE250 reads to a chromosome-scale assembly with N50 of 1.75Gb (Extended Data Fig. 1d, see Methods). HiC data were also used to detect misjoins by binning HiC contacts along the diagonal and identifying points depleted of contacts (Extended Data Fig. 1e). The largest scaffolds correspond to the 17 macrochromosomes or whole chromosome arms of the karyotype *of N. forsteri*. All 10 microchromosomes were also assembled into single scaffolds (see Supplementary Information).

A comprehensive multi-tissue *de novo* transcriptome assembly (BUSCO score >98% Core Vertebrate Genes, CVG) was constructed using RNA extracted from the same individual. For annotation of protein-coding genes evidence from transcript alignments and homology-based gene prediction were combined. This resulted in 31,120 high fidelity gene models. The genome assembly completeness was assessed using the predicted gene set and BUSCO pipeline detecting 91.4% of CVGs (233 genes) and 90.9% of vertebrate conserved genes (2586 genes) (Supplementary Table 2). Our analysis of non-coding RNAs predicted 17,095 ncRNA, including 1,042 tRNA, 1,771 rRNA, and 3,974 microRNAs (Supplementary Table 3, Supplementary Information).

Phylogeny of lungfish, coelacanth and tetrapods

Phylogenetic relationships among coelacanths, lungfishes and tetrapods have been debated[4,5,11]. We used Bayesian phylogenomics (Fig. 1) with 697 one-to-one orthologs for ten vertebrates, with a complex mixture model that can overcome long-branch attraction artefacts[4] and also used non-coding conserved genomic elements (96,601 aligned sites) (Extended Data Fig. 3a). Both data sets unequivocally support lungfish[4,5] as the closest living relatives of land vertebrates that last shared a common ancestor ~420 Ma (Extended Data Fig. 3b).

Synteny conserved of macro- and microchromosomes

Lineage-specific polyploidy events are important evolutionary forces[12] that can lead to genome expansions also in lungfish[9,13]. Despite the massive genome expansion the lungfish chromosomal scaffolds strongly resemble the ancestral chordate karyotype (Fig. 2a, Extended Data Fig. 4a,b). Based on 17 chordate linkage groups (CLGs)[14,15] and 6,337 markers mapped onto the lungfish genome we uncovered conserved syntenic correspondence between lungfish chromosomes and CLGs (Fig. 2a). The vertebrate ancestors underwent two whole genome duplications (2R). Also lungfish retain ancient 2R chromosomal fusions[15] where pre-2R CLG fusions are preserved intact, but substantially expanded (Fig. 2b). Almost all additional CLG fusions happened recently indicated by sharp syntenic boundaries (Fig. 2b). This and *N. forsteri*'s gene number confirms its diploidy.

All ten lungfish microchromosomes (inferred from karyotype[9] and our assembly - Extended Data Fig. 4) could be homologized to chicken and gar microchromosomes (Fig. 2c, Extended Data Fig. 4c,d) mostly even retained their co-linearity. This and conservation of some microchromosomes in gar, chicken and green anole[15,16] suggests that microchromosomes may date back to the earliest vertebrates. The complete retention of microchromosomes in the massively expanded lungfish suggests that stabilizing selection maintains these ancestral units. Supporting this, lungfish microchromosomes show on average higher gene densities and lower density of LINE (long interspersed nuclear elements) elements, the major contributors of genome size (Extended Data Fig. 4b) and also suggests different expansion dynamics of vertebrate micro- and macrochromosomes.

Hallmarks of the giant lungfish genome

A maximum likelihood reconstruction of ancestral vertebrate genome sizes shows two major independent genome expansion events in both lungfish and salamander lineages (Extended Data Fig. 3c) at initially similar (161-165 Mb myr^{-1}), but subsequently slower rates in lungfish (about 39 Mb myr^{-1}). The genome expansion happened in early lungfishes (~400-200 Ma) but slowed during Gondwanan break-up (~200Ma-present) (Extended Data Fig. 3c). Independently, genome size increased in salamanders in two independent DNA repeat expansion waves (Fig. 3b, Extended Data Fig. 3c, 5). LINE elements making up much of its recent genome growth (< 15% divergence, ~9% [4 Gb] also in an earlier burst in lungfish but not axolotl) (Extended Data Fig. 5a). Since mobilized TEs can interrupt gene function one might speculate that such bursts of TE activity might have caused novel gene functions.

Although syntenically highly conserved the lungfish genome has undergone extreme expansion through accumulation of transposable elements (TE). Standard repeat masking procedures of the 37Gb genome assembly identified 67.3% as repetitive (Fig. 3a, Supplementary Table 4). 24.65Gb is the highest repetitive DNA content in the animal kingdom. We tested whether the remaining 13Gb of the genome have signatures of repetitiveness that are obscured by genome size by applying a second round of repeat annotation on the hard-masked genome. This revealed an additional 23.92% of repetitive DNA (Fig. 3a) mostly classified as "Unknown" (adding 11% to the unknown portion of repetitive DNA) as well as "LINE" (+8.5%) (Supplementary Tables 5, 6). In total ~90% is repetitive and expanded in two waves (Fig. 3a, Extended Data Fig. 5).

Asking whether TE's are still active we analyzed polyA-RNA derived RNA-seq data, that likely encode proteins relevant for transposition activity. All major categories of TEs (1106/1821 = 60.7%) were expressed (Extended Data Fig. 6a). TE-families with higher copy numbers were also highly expressed in all three tissues tested. This and the finding of similar copies for many TE-families, suggests that several TE-types remain active contributing to the ongoing expansion of the lungfish genome. Identification of insertion polymorphisms between two lungfish species are necessary to confirm TE activity. Apparently, the transposon silencing machinery did not adapt to reduce overabundant TEs by copy number expansion or structural changes (Supplementary Table 7).

The repeat landscape (proportions of major TE-classes) of lungfish resembles tetrapods including axolotl, while the third extant sarcopterygian lineage, the coelacanth, is more "fish"-like (Fig. 3b). The two largest sequenced animal genomes expanded through different temporal dynamics. While long terminal repeat (LTR) elements are the most abundant TE-class (59%) in axolotl[8], LINEs (25.7%, mostly CR1 and L2 elements) dominate in lungfish (Extended Data Fig. 5, 6).

These two retrotransposon-classes belong to the same copy-and-paste (and not cut-and-paste) category propagate via different mechanisms[17]. Although global repeat compositions differ between lungfish and axolotl, the same LTR-class impacts their genic regions (Extended Data Fig. 6). (See Supplementary Information).

To further understand the enormous genome growth, we compared the genome structure of *N. forsteri* with other genomes (Extended Data Fig. 6c-d, 7). While compact genomes have small introns, intragenic non-coding regions usually increase with genome size[18]. The largest intron of the lungfish is 5.8Mb in the *DMBT1* gene and average intron size is 50Kb as in axolotl, compared to 1kb in fugu and 6kb in human. Introns in the *N. forsteri* genome comprise ~8Gb (21% of genome) – interestingly similar to human (21%), but only half of fugu (40%). This suggests that similar mechanisms affect the genic and intergenic compartments following expectations for genome size evolution[19].

In most genes the first intron typically is the largest. The biological relevance of this remains unclear. Surprisingly, also lungfish and axolotl's first introns are much larger than downstream introns (Extended Data Fig. 7) indicating that relatively larger first introns in smaller genomes are probably not due to space requirements of regulatory or structural motifs[20].

It has been suggested that the size of intragenic non-coding sequences and the extent of intron expansion is associated with organismal features such as metabolic rate[18] or functional categories of genes[8] e.g., developmental vs. non-developmental genes. Similar to axolotl[8] the introns in developmental genes are smaller also in lungfish than in non-developmental genes (p= 2.166e-08, Mann-Whitney U test) (Supplementary Table 8).

Fish-tetrapod transition: Genomic preadaptations

Positive selection analysis uncovered 259 genes, many of which are related to estrogen and female reproduction related categories (Supplementary Information, Supplementary Table 9). We compared these rate dynamics (16,471 gene families) (Supplementary Tables 10,11). And found in the lungfish lineage 24 families contracted and 107 expanded, possibly related to evolutionary innovations.

Air breathing: Evolution of lungs

All land-living vertebrates and adult lungfish are obligate air breathers. Interestingly, the pulmonary surfactant-protein-B-family expanded considerably. Surfactants are necessary components of the lipoprotein mixture that covers the lung surface ensuring proper pulmonary function. In lungfish the number of surfactants genes increased to a tetrapod-typical number (2-3x larger compared to fish) (Supplementary Table 12). This may indicate an adaptation to air-breathing in lungfish. We further investigated the expression of *shh*, an important regulator of lung development[21] during lungfish embryogenesis (Extended Data Fig. 8a). *Shh* is strongly expressed in the developing lungs (stage 43-48 embryos) visualizing the development of the right-sided lung (*Neoceratodus* has a unilateral lung). It develops strikingly similar to those of amphibians[22]. Altogether this highlights molecular signatures of lungs necessary for the conquest of land by sarcopterygians.

Olfaction: Evolution of the vomeronasal organ

Expansions were also noted for genes involved in olfaction. The gene complement of receptors for air-borne odorants, which is large and complex in tetrapods and small in fish is considerably expanded in lungfish, while several receptor classes for waterborne odors shrank, in particular zeta and eta receptors, which abound in teleosts (Supplementary Table 13). The vomeronasal organ (VNO) present in most tetrapods[23,24] is linked to pheromone reception and expresses a big repertoire of vomeronasal receptor (VR) genes that is particularly large in amphibians. In *N. forsteri* the VR gene family, known from fish and even lampreys, although their function in these species is unknown, strongly expanded. Lungfish possess a "VNO primordium"[25]. The notable expansion of the VR-gene family (especially V2Rs) in *N. forsteri* (Supplementary Table 14) shows that the VNO is a tetrapod innovation, which emerged at the water-to-land transition.

Evolution of terrestrial locomotion: Lobed fins

Sarcopterygians have elaborated endochondral skeletons hence the term: lobed-fins that are distally branched forming digits suitable for substrate-based locomotion. Our analysis of conserved tetrapod limb enhancer elements[26] indicates sarcopterygian origins for 31 (Fig. 4a, Extended Data Fig. 8b). Of these the *hs72* enhancer, related to *sall1*, drives autopodal expression (Fig. 4b). We find *sall1* strongly expressed in lungfish embryos similar to expression patterns reported for tetrapods[27] (Fig. 4b) but absent during zebrafish fin development[28]. Similar functions of *sall1* during mouse limb development[27] suggest that it contributed to the acquisition of sarcopterygian lobed-fins already in lungfish.

Hox clusters and fin-to-limb transition

The four *Neoceratodus hox* clusters (*hoxa-hoxd*) comprise 43 genes (Extended Data Fig. 9) and presence of *hoxb10* and *hoxa14* confirms their loss at the fish-to-tetrapod transition[11]. RNA-seq analysis of *hox* genes expression in *Neoceratodus* larval fins (Extended Data Fig. 8c) showed an unexpected expression of *hoxc* genes. To date *hoxc* gene expression in paired fins and limbs was only reported for mammals[29] related to the nail bed. We observed *hoxc13* expression in axolotl limbs (Fig. 4c) but it was absent in ray-finned fish pectoral fins (Extended Data Fig. 8d). Transcript localization in *Neoceratodus* embryos showed expression of *hoxc13* in the distal fin (Fig. 4c). This indicates an early gain of *hoxc13* expression in sarcopterygians suggesting cooption of this domain in tetrapods to pattern dermal limb elements such as nails, hoofs and claws. Together with *sall1* this demonstrates an early sarcopterygian origin of limb-like gene expression ready for tetrapod cooption facilitating the fin-to-limb transition and colonization of land.

Hox cluster expansion versus regulation

In line with the overall genome expansion, the *Neoceratodus hox* clusters are larger than in mouse, chicken and *Xenopus*, but with an uneven pattern of expansion (Extended Data Fig. 9). The clustering of *hoxd* genes results in their co-regulation by enhancers 3'and 5' of the cluster leading to co-expression of *hoxd9-d13* in the distal appendages[30–33]. During *Neoceratodus* fin development *hoxd11* expression is nearly absent from the *hoxd13* territory[34] (Fig. 4d) while in axolotl *hoxd9-d11* are excluded from the *hoxd13* digit domain[35] (Extended Data Fig. 8e). Such apparent loss of co-regulation between *hoxd9-d11* and *hoxd13* is similar to that caused by experimentally increased distances in the *hoxd* cluster[30] and suggests a disruption of enhancer sharing caused by the expansion of the *hoxd11-d13* intergenic region (Fig. 4e). Additional analyses in mouse, *Xenopus*, lungfish and axolotl shows that despite 5-10x differences in *hoxd* cluster size, the region comprising *hoxd8-hoxd11* remained fixed at ~25kb (Fig. 4e). This apparent constraint is likely due to sharing of enhancers located at the 3' end of the cluster[31]. Altogether this indicates that *hoxd* expansion has partially disrupted long-range enhancer sharing but that, conversely, such mechanisms have locally also constrained intergenic distances.

We sequenced and chromosome-level assembled (Supplementary Table 15) the largest animal genome and substantiate that lungfish are the closest living relatives of tetrapods. Despite the lungfish's unique genome expansion history the genic organization and chromosomal homology is maintained even including its microchromosomes.

Genomic preadaptations for the water-to-land transition of vertebrates include a larger complement of lung-expressed surfactant genes that might have facilitated the evolution of air-breathing through a lung. Moreover, the number of vomeronasal organ olfactory receptors as well as other receptor gene families that permit detection of air-born odors increased in the lineage leading to airbreathing lungfish. The uneven expansion of *hox* clusters demonstrates regulatory consequences of - and constraints on - genome expansion. The evolutionary trajectory of limb enhancers shows an early fish origin of the limb regulatory program with important changes towards preadaptations for terrestrialisation preceding the fin-to-limb transition. Genes that pattern the tetrapod limb but previously presumed absent from fins, such as *sall1* and *hoxc13*, gained new expression domains in the lobe-finned lineage. Such novelties might have predisposed the sarcopterygians to conquer land demonstrating how the lungfish genome can contribute to better understanding of this major transition during vertebrate evolution.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-021-03198-8.

1. Clack, J., Sharp, E. & Long, J. The fossil record of lungfishes. In: *The Biology of Lungfishes* (eds. Jorgensen, J. M. & Joss, J.) 1–42 (CRC Press, 2011).
2. Kemp, A. The biology of the Australian lungfish, *Neoceratodus forsteri* (Krefft 1870). *J. Morphol*. **190**, 181–198 (1986).
3. Carroll, R. L. *Vertebrate Paleontology and Evolution*. (W.H. Freeman & Co. Ltd., 1988).
4. Irisarri, I. & Meyer, A. The identification of the closest living relative(s) of tetrapods: Phylogenomic lessons for resolving short ancient internodes. *Syst. Biol*. **65**, 1057–1075 (2016).
5. Irisarri, I. *et al*. Phylotranscriptomic consolidation of the jawed vertebrate timetree. *Nat. Ecol. Evol* **1**, 1370–1378 (2017).
6. Krefft, G. Description of a giant amphibian allied to the genus *Lepidosiren* from the Wide Bay district, Queensland. *Proc. Zool. Soc. Long*. **1870**, 221–224 (1870).
7. Gunther, A. XIX. Description of *Ceratodus*, a genus of ganoid fishes, recently discovered in rivers of Queensland, Australia. *Phil. Trans. R. Soc. B Biol. Sci*. **161**, 511-571 (1871).
8. Nowoshilow, S. *et al*. The axolotl genome and the evolution of key tissue formation regulators. *Nature* **554**, 50–55 (2018).
9. Rock, J., Eldridge, M., Champion, A., Johnston, P. & Joss, J. Karyotype and nuclear DNA content of the Australian lungfish, *Neoceratodus forsteri* (Ceratodidae: Dipnoi). *Cytogenet. Cell Genet*. **73**, 187–189 (1996).
10. Pedersen, R. A. DNA content, ribosomal gene multiplicity, and cell size in fish. *J. Exp. Zool*. **177**, 65–78 (1971).
11. Amemiya, C. T. *et al*. The African coelacanth genome provides insights into tetrapod evolution. *Nature* **496**, 311–316 (2013).
12. Fox, D. T., Soltis, D. E., Soltis, P. S., Ashman, T.-L. & Peer, Y. V. de. Polyploidy: A Biological Force From Cells to Ecosystems. *Trends in Cell Biology* **30**, 688–694 (2020).
13. Vervoort, A. Tetraploidy in *Protopterus* (Dipnoi). *Experientia* **36**, 294–296 (1980).
14. Putnam, N. H. *et al*. The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**, 1064–1071 (2008).
15. Simakov, O. *et al*. Deeply conserved synteny resolves early events in vertebrate evolution. *Nat. Ecol. Evol*. **4**, 820–830 (2020).
16. Braasch, I. *et al*. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nat. Genet*. **48**, 427–437 (2016).
17. Jurka, J., Kapitonov, V. V., Kohany, O. & Jurka, M. V. Repetitive sequences in complex genomes: Structure and evolution. *Annu. Rev. Genom. Hum. Genet*. **8**, 241–259 (2007).
18. Zhang, Q. & Edwards, S. V. The evolution of intron size in amniotes: A role for powered flight? *Genome Biol Evol* **4**, 1033–1043 (2012).
19. Lynch, M. & Conery, J. S. The origins of genome complexity. *Science* **302**, 1401–1404 (2003).
20. Bradnam, K. R. & Korf, I. Longer first introns are a general property of eukaryotic gene structure. *PLOS ONE* **3**, e3093 (2008).
21. Kugler, M. C., Joyner, A. L., Loomis, C. A. & Munger, J. S. Sonic Hedgehog Signaling in the Lung. From Development to Disease. *Am. J. Respir. Cell Mol. Biol*. **52**, 1–13 (2015).
22. Rankin, S. A. *et al*. A molecular atlas of *Xenopus* respiratory system development. *Dev. Dyn*. **244**, 69–85 (2015).
23. Døving, K. B. & Trotier, D. Structure and function of the vomeronasal organ. *J. Exp. Biol*. **201**, 2913–2925 (1998).
24. Syed, A. S., Sansone, A., Hassenklöver, T., Manzini, I. & Korsching, S. I. Coordinated shift of olfactory amino acid responses and V2R expression to an amphibian water nose during metamorphosis. *Cell. Mol. Life Sci*. **74**, 1711–1719 (2017).
25. Nakamuta, S., Nakamuta, N., Taniguchi, K. & Taniguchi, K. Histological and ultrastructural characteristics of the primordial vomeronasal organ in lungfish. *Anat. Rec. (Hoboken)* **295**, 481–491 (2012).
26. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser–a database of tissue-specific human enhancers. *Nucleic Acids Res*. **35**, D88-92 (2007).
27. Kawakami, Y. *et al*. Sall genes regulate region-specific morphogenesis in the mouse limb by modulating Hox activities. *Development* **136**, 585 (2009).
28. Camp, E., Hope, R., Kortschak, R. D., Cox, T. C. & Lardelli, M. Expression of three spalt (sal) gene homologues in zebrafish embryos. *Dev Genes Evol* **213**, 35–43 (2003).
29. Fernandez-Guerrero, M. *et al*. Mammalian-specific ectodermal enhancers control the expression of *Hoxc* genes in developing nails and hair follicles. *Proc. Natl. Acad. Sci. USA* in press (2020) https://doi.org/10.1073/pnas.2011078117.
30. Spitz, F., Herkenne, C., Morris, M. A. & Duboule, D. Inversion-induced disruption of the Hoxd cluster leads to the partition of regulatory landscapes. *Nat. Genet*. **37**, 889–893 (2005).
31. Andrey, G. *et al*. A switch between topological domains underlies HoxD genes collinearity in mouse limbs. *Science* **340**, 1234167 (2013).
32. Montavon, T. & Duboule, D. Chromatin organization and global regulation of Hox gene clusters. *Phil. Trans. R. Soc. B Biol. Sci*. **368**, 20120367 (2013).
33. Woltering, J. M., Noordermeer, D., Leleu, M. & Duboule, D. Conservation and divergence of regulatory strategies at *Hox* loci and the origin of tetrapod digits. *PLoS Biol*. **12**, e1001773 (2014).
34. Woltering, J. M. *et al*. Sarcopterygian fin ontogeny elucidates the origin of hands with digits. *Sci. Adv* **6**, eabc3510 (2020).
35. Woltering, J. M., Holzem, M. & Meyer, A. Lissamphibian limbs and the origins of tetrapod hox domains. *Dev. Biol*. **456**, 138–144 (2019).

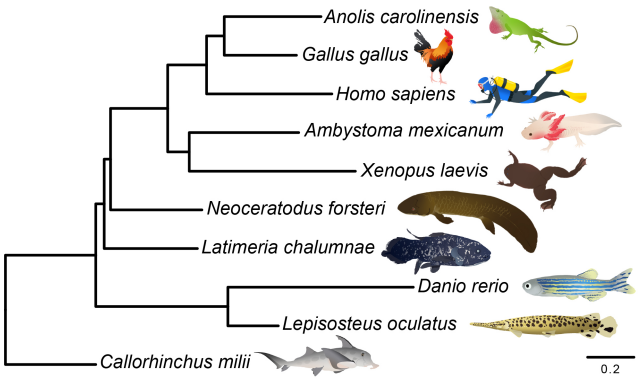


Fig. 1 | Bayesian phylogeny based on 697 orthologs (using PhyloBayes MPI; CATGTR). All branches were supported by posterior probabilities of 1. The protein and a non-coding conserved genomic elements data set (Extended data Fig. 3a) recovered identical and highly supported vertebrate relationships (posterior probability=1.0 and 100% bootstrap for all branches). Scale bar is expected amino acid replacements per site.

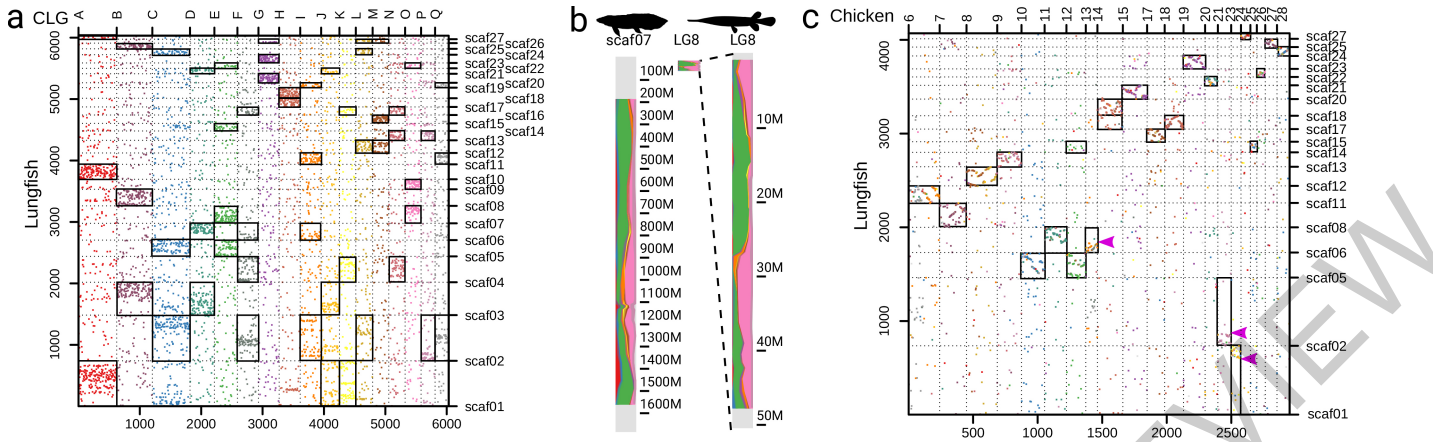


Fig. 2 | Conserved synteny and chromosomal expansion in lungfish. **a**, Mapping of chordate linkage groups (CLGs) onto lungfish chromosomes. Shown are orthologous gene family numbers. Each dot represents an orthologous gene family, CLGs as defined in ref. [15]. Scaffolds 01-17 represent lungfish macrochromosomes and 18-27 microchromosomes. Significantly enriched CLGs on lungfish chromosomes indicated by rectangles (for raw data see Extended Data Fig. 4f). **b**, Expansion of homologous chromosomes in lungfish, compared to spotted gar (upper, here only LG8 shown, the others are in Extended Data Fig. 4a). Chromosomes are partitioned into bins and CLG content is profiled, chromosomal position is plotted next to each chromosome. LG8 in gar has a prominent jawed vertebrate-specific fusion of CLGE+O, which is retained throughout the whole chromosome in lungfish, despite being >30-fold larger. The small box in the middle is the unexpanded LG8 of spotted gar. **c**, Preservation of microchromosomes. Chicken microchromosomes are plotted (for gar Extended Data Fig. 4d) along with their lungfish homologs with >50 orthologs. Scaffolds 01-17 represent lungfish macrochromosomes and 18-27 microchromosomes. For chicken only microchromosomes shown. Significantly enriched chicken microchromosomes on lungfish chromosomes indicated by rectangles (for raw data: Fig. 4e). Most chicken microchromosomes are in one-to-one correspondence between lungfish and chicken, but some were recently incorporated into macrochromosomes. Those lungfish macrochromosomes, e.g., scaffold01 or scaffold02, have significant association with both chicken macro- and microchromosomes. However, those fusions are recent in lungfish, because the positions of chicken orthologs is restricted to specific areas of lungfish chromosomes. Silhouettes are from[34].

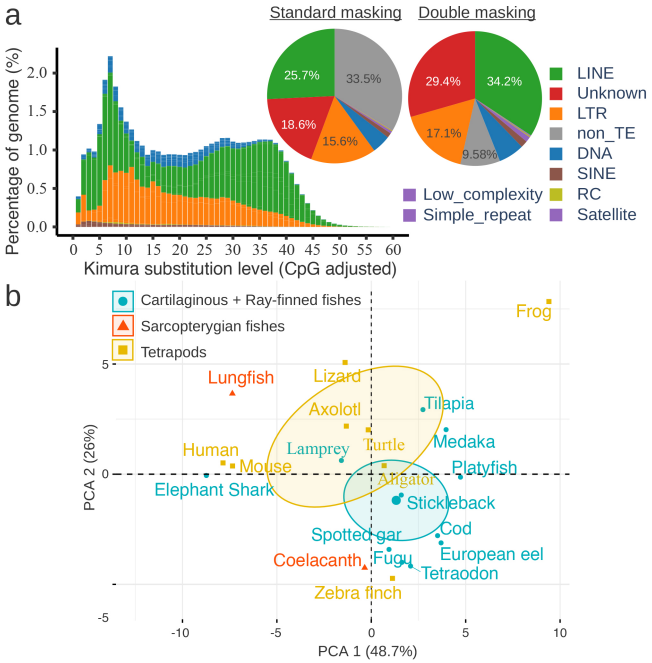


Fig. 3 | Composition of repetitive elements in the lungfish genome. **a**, (Pie chart) Overall composition of repetitive elements from unmasked assembly (first TE annotation, left), together with the annotation from the hard masked genome (second TE annotation, right). (Bar chart) Repeat landscape of major classes of transposable elements. Kimura substitution level (%) for each copy against its consensus sequence used as proxy for expansion history of the transposable elements. Older copies (old expansion) accumulated more mutations and show higher divergence from the consensus sequences. **b**, Principle component analysis of composition of repetitive elements (LTR, LINE, SINE, DNA and Unknown, filtered by 80/80 rule) of vertebrates.

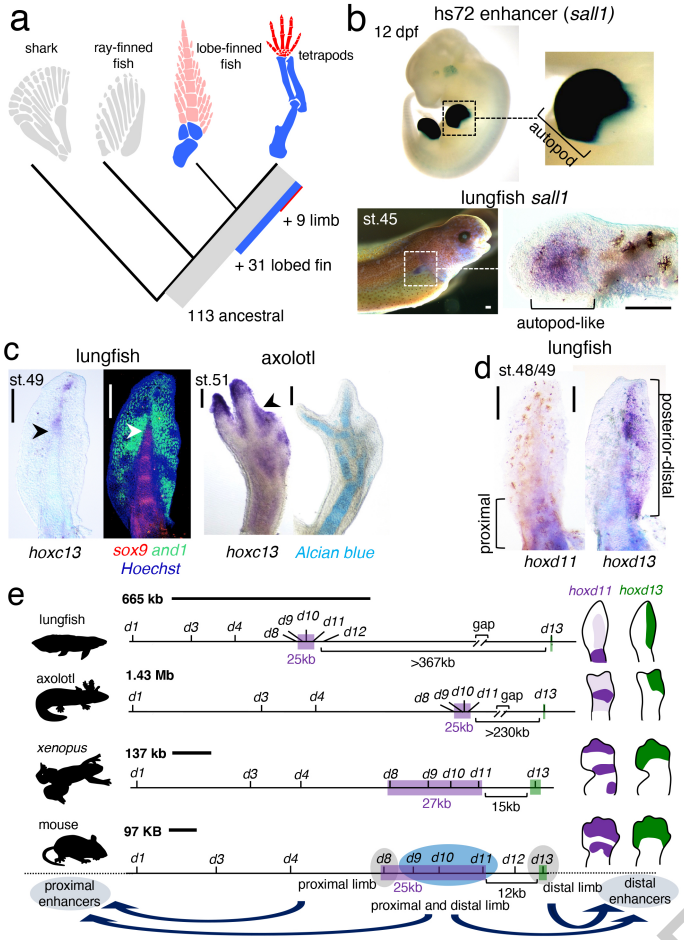


Fig. 4 | Regulatory preadaptation of fleshly-lobed fin and *hoxd* gene regulation. a, Analysis of 330 validated mouse and human limb enhancers shows deep evolutionary origin of limb regulatory program and 31 enhancers are associated with the emergence of the lobed fin. **b**, Of these, the hs72 enhancer located near the *sall1* gene drives strong LacZ in mouse autopods (N = 3/3 embryos, lacZ stained embryos courtesy of VISTA enhancer[26]). *Sall1* is expressed in a similar autopodial-like domain in lungfish pectoral fins (N = 2/2 fins). **c**, *Hoxc13* is expressed in a distal lungfish area overlapping with the central metapterygial axis (*sox9*) and fin fold (*and1*) (arrowheads) (N = 2/2 fins). Similar expression present in axolotl limbs (arrowhead) (N = 4/4 limbs) indicating deep sarcopterygian origin for this expression domain. **d**, During lungfish fin development *hoxd11* and *hoxd13* are expressed in mostly non-overlapping proximal and posterior-distal fin domains (N= 4/4 fins each). **e**. The lungfish *hoxd* cluster has increased in size compared to mouse and *Xenopus* but may be smaller than the axolotl *hoxd* cluster. In lungfish and axolotl expansion has occurred in the 3' and 5' regions of the cluster, whereas the central *hoxd8-hoxd11* region (lilac box) remained stable ~25 kb, forming a separate "mini-cluster". The *hoxd* cluster is regulated by 3' and long-range enhancers. *Hoxd9-hoxd11* (lilac) and *hoxd13* (green) are subject to enhancer sharing[31] and co-expressed in the distal limb in mouse and *Xenopus*[31,35], whereas the increased genomic distance between *hoxd13* and *hoxd9-hoxd11* has disrupted their co-expression in lungfish and axolotl distal appendages. The preserved clustering of *hoxd8-hoxd11* can be explained by enhancer sharing 3` of the cluster[31], which likely places constraints on their intergenic distances. Axolotl and *Xenopus hoxd11/13*[35], lungfish *hoxd11/13* domains after ref. [34] and panel c. (Supplementary Table 16 lists primers for probes). Scale bars in panel a-d 0.2 mm. Abbreviations: st.; stage. Silhouettes are from[34].

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Biological Materials

Biopsy material for DNA and RNA isolation was obtained from a juvenile Australian lungfish (*Neoceratodus fosteri*) imported from Australia (CITES Permit No.: PWS 2017-AU-000242). Due to the immature status of the gonad the sex could not be determined. The same specimen was used for genome sequencing (muscle), construction of the Hi-C library (spleen) and transcriptome sequencing of brain, gonad and liver. The second set of reads was generated from lungfish embryos (embryonic stage 52, GenBank accession numbers SRR6297462-6297470)[34]. Embryos were bred and collected under permit ARA 2009.039 at Macquarie University, Australia.

DNA extractions, genome sequencing and assembly

Genome sequencing. HMW and ultra HMW DNA was prepared by FutureGenomics (Leiden, The Netherlands) and Nextomics (Wuhan, China) and sequenced using Nanopore technology (for statistics see Supplementary Table 1).

gDNA for genome correction from snap frozen lungfish muscle tissue (0.3 g) was isolated by standard gDNA isolation protocol. Library preparation was performed using the Westburg NGS DNA library kit. The final library was excised by Pippin prep with 400bp DNA size and sequenced (Illumina Nova-seq S2; PE150) at Vienna Bio Center NGS facility.

Hi-C library was generated as described[36,37], with modifications detailed in Supplementary Methods. Final Hi-C libraries were sequenced (Illumina Nova-seq SP; PE150) at Vienna Bio Center NGS facility.

Genome assembly. 96 M reads comprising 1.2 Tbp were assembled using the MARVEL genome assembler[8]. We first aligned 1% of the reads against all other reads. From these 1%-against-all alignments we derived information on the repetitive elements present in the reads and used transitive transfer to repeat annotate all reads used in the assembly. Regions were deemed repetitive when the depth of the alignments for a given read exceeded the expected depth fourfold. Given the alignment of the 1% against every other read in the assembly we then transferred the repeat annotation of the 1% using the alignments to the respective position in the aligned reads. Here the assumption is that when region (a,b) in read A aligns to (c,d) in read B and for a <= rb <= re <= b with (rb, re) being a repetitive element, this than can be mapped using the alignment to a corresponding region in B, which then can be tagged as repetitive as well. The final repeat masking track covered 28,7% of the 1.2 Tbp.

We then processed with an all-against-all alignment, with repeat masking in place, yielding 5 billion alignments. Based on these alignments read qualities were derived at 100 bp resolution, highlighting low sequencing quality regions in the reads. Using the alignments and the read qualities structural weaknesses (chimeric breaks, high-noise regions and other sequencing artifacts) in the reads were repaired (see Supplementary Methods, See Extended Data Fig. 10).

Repaired reads were then used for a new round of alignments, again with repeat masking as described above, in place. After alignment the default MARVEL assembly pipeline proceeded as shown in the included examples of the source distribution (Extended Data Fig. 1).

For the current MARVEL source code repository see https://github.com/schloi/MARVEL. For sample execution scripts see: https://github.com/schloi/MARVEL/examples/.

Scaffolding

We used an agglomerative hierarchical clustering based scaffolding approach utilizing various normalizations (Extended Data Fig. 1). For details see Supplementary Methods.

We created initial clusters by selecting the largest contigs with the fewest contacts between them, each contig serving as a single cluster. We then added contigs based on unique assignability to clusters. This was followed by scaffolding the cluster separately, visual inspection of an approximate contact map derived during the scaffolding process and return of wrongly assigned contigs to the set of unassigned contigs. We created contact maps for all clusters and merged or split clusters based on the signal within those. The process of assigning contigs, scaffolding, merging and splitting clusters was repeated until no more useful changes could be made to the clusters (Supplementary Table 15 for comparison of chromosome and scaffold DNA content).

For the public source code repository see: https://github.com/schloi/MARVEL/hic

The MARVEL assembler and scaffolder has been used to obtain a chromosome-scale axolotl genome assembly that has been validated in comparison to the chromosome-scale meiotic scaffolding from ref. [38] and is available as described in ref. [39].

Genome assembly correction

For correction of errors (indels, base substitutions and small gaps) remaining after the genome assembly, we applied a two-step procedure using DNAseq and RNAseq reads separately. In brief, we sequenced the same genomic DNA sample and generated 4,693,324,032 high-quality read pairs (2x150bp) (30x coverage). Additionally, we used the RNAseq reads from the *de novo* transcriptome assembly to correct indels, but not base substitutions in transcribed regions (see Supplementary Methods, Supplementary Results and See Extended Data Fig. 10 for details).

Transcriptome assembly

RNA was isolated from brain, spinal cord, eyes, gut, gonad, liver, jaw, gills, pectoral fin, caudal fin, trunk muscles and larval fin. Libraries were constructed using NEBNext® Ultra™ II Directional RNA library preparation kit (New England Biolabs, Ipswich, USA), Illumina TruSeq RNA sample preparation kit (Illumina, San Diego, USA) or Lexogen Total RNA-seq Library Prep Kit V2 (Lexogen, Vienna, Austria). Paired-end sequencing, performed with Illumina platforms, yielded approximately 1,150 million (M) raw reads.

Raw reads, filtered and corrected using Trimmomatic v0.36[40] and RCorrector v1.0.2[41], were assembled using *de novo* and reference-guided approaches. For *de novo* assembly, only reads derived from poly-A selected RNA were processed using the Oyster River Protocol (ORP) v2.2.8[42]. Briefly, reads were assembled using Trinity v2.8.4 (k-mer=25), SPAdes v3.13.3[43] (k-mer=55), SPAdes (k-mer=75) and Trans-Abyss v2.0.1[44] (k-mer=32), respectively. The four different assemblies were then merged using the OrthoFuser module[45,46] implemented in ORP. Completeness of the *de novo* assembled transcriptome was assessed with BUSCO v3[47] using Core Vertebrate Genes (CVG) and Vertebrata genes (vertebrata_odb9 database) in the gVolante webserver[48]. For reference-guided assembly, all reads were aligned to the *N. forsteri* genome, each sample independently, using the program HISAT2 v2.1.0[49] (maximum intron length set to 3 Mbp). The resulting mapping files were parsed by StringTie v1.3.6[50] and transcripts reconstructed from each aligned sample were merged in a single consensus "gtf" file.

Genome annotation

Repeats and transposable elements annotation. *N. forsteri* repeat sequences were predicted using RepeatMasker (v.4.0.7) with default TE Dfam database and a *de novo* repeat library constructed using RepeatModeler (v.1.0.10), including the RECON (v.1.0.8), RepeatScout (v.1.0.5) and rmblast (.2.6.0), with default parameters. TEs not classified by RepeatModeler were analyzed using PASTEC (https://urgi.versailles.inra.fr/Tools/) and DeepTE[51]. Repeat sequences of *Ambystoma mexicanum* [AmexG_v3.0.0, www.axolotl-omics.org] were predicted using the same approach. Repetitive sequences of *Anolis carolinensis* [GenBank accession GCA_000090745.2], *Xenopus tropi-*

calis [GCA_000004195.4], *Rhinatrema bivittatum* [GCA_901001135.1], *Latimeria chalumnae* [GCA_000325985.2], *Lepisosteus oculatus* [GCA_000242695.1], *Danio rerio* [GCA_000002035.4] and *Amblyraja radiata* [GCF_010909815.1]) were identified using Dfam TE Tools Container (github.com/Dfam-consortium/TETools) including RepeatModeler (v.2.0.1) and RepeatMasker (v.4.1.0). To further examine remaining intergenic sequences, we predicted repetitive sequences again using the same workflow on the genome hard-masked with repeats already predicted by RepeatMasker.

Kimura distance-based distribution analysis, TE composition PCA analysis

Kimura substitution levels between the repeat consensus to its copies were calculated using a utility script calcDivergenceFromAlign.pl bundled in RepeatMasker. Repeat landscape plots were produced with the R script nf_all_age_plot.R and nf_am_rb_age_plots.R, using the divsum output from calcDivergenceFromAlign.pl. Principal component analysis on repetitive element composition was performed on R (v.3.6) using factoextra package (v.1.0.6). Repetitive element compositions (SINE, LINE, DNA, LTR, Unknown) were calculated from the predicted libraries. Repetitive element copies were filtered by the 80/80 rule (equal or longer than 80bp, equal or more than 80 per cent identity compared with the consensus sequence). Repetitive element composition of other vertebrates was obtained from ref. [52].

TE composition by gene length, LTR family analysis

Repetitive sequence composition within genes (grouped by length) was examined by calculating the coverage (bp) of each class of repetitive element, normalized by gene length. We examined LTR family enrichment in genic regions. All calculations and visualizations are summarized in the jupyter notebook file te_general_analysis.ipynb. All python scripts ran on Python >= 3.7 and used the package gffutils (v.0.10.1) (https://github.com/daler/gffutils) to operate large gene and repetitive element annotation files from big genomes. Plots were generated using Plotly python API (https://plot.ly).

TE content in genic regions

Intron position was calculated by GenomeTools (v.1.5.9). The sum of the coverage of the repetitive element (e.g. LINE/CR1) was normalized by the length of the genic feature considered (Supplementary Table 17) (e.g. intron 8) using python script te_cnt_class.py.

TE expression

TE expression was assessed with TEtools[53] on gonad, brain and liver polyA-RNA data. Because of the large size of lungfish genome, a random subset of 10% of all TE copies was used. TE family counts were normalized by TE family consensus length (count * $1e^6$/ consensus_length) and library size. Normalized counts were plotted against TE family copy numbers.

Annotation of protein-coding genes

Protein coding genes were predicted by combining transcript and homology-based evidence. For transcript evidence, assembled transcripts (see above) were mapped to the assembly using Gmapl v2019-05-12[54] and the gene structure was inferred using the PASA pipeline v2.2.3[55]. Expression of each transcript was measured using the whole RNA-seq dataset (see above, section "Transcriptome assembly") and the pseudoalignment algorithm implemented in Kallisto v0.46.1[56]. For homology evidence, we collected manually curated proteins from UniProtKB/SWISSPROT database (UniProtKB/Swiss-Prot 2020_03)[57] and protein sequences of *Callorhinchus milii*, *Latimeria chalumnae*, *Lepisosteus oculatus* and *Xenopus tropicalis* from Ensembl (www.ensembl.org) and NCBI (www.ncbi.nlm.nih.gov/genome), and aligned them to the repeat masked assembly using Exonerate v2.2[58]. Transcript and homology-based evidence were then combined by prioritizing the former (homology-based predicted genes were removed when intersecting a gene predicted using the reconstructed transcripts). The combined gene set was then processed by two rounds of "PASA compare" in order to add untranslated region (UTR) annotations and models for alternatively spliced isoforms. Low-quality gene models were removed by applying three further quality-filtering steps in an iterative fashion: 1) single-exon genes were retained only when no similarity with exons of multi-exonic genes was found (similarity was identified with the *glsearch36* module implemented in the FASTA v36.3.8g package[59] with e-value cutoffs of $1e^{-10}$ and identity cutoffs of 80); 2) genes intersecting repeat elements were removed when >50% (single-exonic genes) and >90% (multi-exonic genes) were covered by repeats; 3) genes with internal stop codon(s) were removed. The completeness of the predicted protein-coding gene set was assessed with BUSCO using the Core Vertebrate Genes (CVG) and the Vertebrata genes (vertebrata_odb9 database) in the gVolante webserver.

To annotate the lungfish Hox clusters, *Hox* genes were first identified using BLAST with vertebrate orthologs as query (see Supplementary Methods).

Annotation of non-coding RNA genes

Non-coding RNA genes were annotated using tRNAscan-SE v.2.0.3[60] and Infernal v.1.1.2[61]. The same procedure was applied to the genomes of the nine other focal species. For each of the ten species, the corresponding miRNA sets (obtained from miRBase v.22[62] database) were used to predict miRNA target sites on 3′ UTRs of canonical mRNAs using miRanda v.3.3[63]. Further details are provided in Supplementary Information.

Annotation of conserved non-coding elements (CNEs)

Whole genome alignments. The masked versions of the genome assemblies of the 10 species used for the phylogenetic tree (Fig. 1) were used to build a whole-genome alignment with the human genome as reference (10-way WGA). Briefly, each pairwise alignment was constructed using Lastz v.1.03.73[64] and further processed using UCSC Genome Browser tools[65]. Multiple alignments were generated using as input the nine pairwise alignments in "maf" format with the programs Multiz v.11.2 and Roast.v.3.0[66].

Detection of conserved elements. The phylogenetic hidden Markov model (phylo-HMM) implemented in phastCons[67] (run in rho-estimation mode) was used to predict a consistent set of conserved genomic elements in the 10-species whole genome alignment (10-way WGA). A neutral model of substitutions was calculated using phyloFit[67] with the general reversible substitution model (REV) from four-fold degenerate (4d) sites. Raw CNEs detected by phastCons were merged when their distance was < 10 bp, and subsequently CNEs < 50 bp were removed. Protein-coding CNEs and those intersecting non-coding RNA genes, pseudogenes, retrotransposed elements and antisense genes (annotated in the human genome) were removed.

Expansion of the genome in intergenic regions

The final filtered set of CNEs was used to investigate expansion of intergenic spaces. We compared the distance of non-exonic elements that are conserved in lungfish and three tetrapods (human, chicken and axolotl). To obtain informative CNE pairs, we selected those CNEs that: 1) were present in all four genomes; 2) were located in intergenic space; 3) were located in the same contig/chromosome in each species; 4) did not have a gene in between them. The remaining set of 223 CNE pairs were used to calculate intergenic distance and region-specific expansion of the lungfish genome (Supplementary Table 18).

Lineage-specific acceleration of CNEs

The program phyloP was used to test each CNE for lineage-specific accelerated evolution[67,68] in the lungfish branch. A likelihood ratio test to compute p-value of acceleration with respect to a neutral model of

evolution for each of the conserved elements in the alignment was used. CNEs showing FDR adjusted p-values < 0.05 were considered significantly accelerated (ACC-CNEs). The ACC-CNEs were checked for overlap with a set of 1,978 experimentally validated human and mouse noncoding fragments with gene enhancer activity (data from "VISTA Enhancer Browser"[26]) (Supplementary Table 19).

Macro-synteny analysis

Amphioxus annotation[15] was mapped onto the lungfish assembly using TBLASTN. The chordate linkage group (CLG) identity of amphioxus genes was used to determine CLG composition of lungfish chromosomal scaffolds. Dot plots were done using scripts available at https://bitbucket.org/viemet/public/src/master/CLG/.

Comparison of intron size

Intron size was compared between lungfish, axolotl, human and fugu for one-to-one orthologs. Intron sizes of each gene were calculated from the gff files of each genome. Genes without start codon were removed to avoid the pseudo-intron order. The intron size was compared first in absolute bp, then in the value normalized by each genome size (lungfish: 44032 Mb; axolotl 32768 Mb; human 3000 Mb and fugu 400 Mb).

Orthology assignment

Protein sequences of *Anolis carolinensis*, *Callorhinchus milii*, *Danio rerio*, *Gallus gallus*, *Homo sapiens*, *Latimeria chalumnae* and *Lepisosteus oculatus* were downloaded from Ensembl (Lepisosteus_oculatus), and of *Xenopus laevis* from NCBI (www.ncbi.nlm.nih.gov/genome). Sequences of *Ambystoma mexicanum* were taken from ref. [39]. In case of alternative splicing, we kept the longest sequence for the gene. All proteins were pooled together as the query and database for an all vs. all BLASTP. From the result we determined a H-score between each two proteins as representative of the distance for sequence similarity[69], and launched a clustering using Hcluster_sg[70]. Finally, for each cluster, a gene tree was built using TreeBeST and orthology between genes was assigned.

Phylogeny inference

The phylogeny was inferred using the set of 697 orthologous proteins. Individual loci were filtered with PREQUAL[71], aligned with MAFFT ginsi[72] and highly incomplete positions (>80%) trimmed with BMGE[73]. Orthology was ensured by manual inspection of ML gene trees (IQ-TREE) and alignments (MAFFT ginsi) for loci showing high branch-length disparity and five individual sequences were removed. Loci were concatenated into a final matrix containing 10 taxa and 697 loci, totaling 383,894 aligned amino acid positions, of which 208,588 (54%) were variable. Phylogeny was inferred using PhyloBayes MPI v.1.7[74] under the site-heterogeneous CAT-GTR model, shown to avoid phylogenetic artifacts when reconstructing basal sarcopterygian relationships[4]. Two independent MCMC chains were run until convergence (>4,000 cycles), assessed a posteriori using PhyloBayes' built-in functions (maxdiff = 0, meandiff = 0, ESS>100 for all parameters after discarding the first 25% cycles as burnin). Post-burnin trees were summarized into a fully resolved consensus tree with posterior probabilities of 1 for all bipartitions.

Whole-genome alignment-based phylogeny

The 10-species whole genome alignment (10-way WGA) was processed by MafFilter v.1.3.0[75] to keep only alignment blocks > 300 bp that were present in all species. Filtered non-coding blocks were then concatenated and exported in phylip format. Poorly aligned regions were removed using trimAl v1.2 with option "-automated1". The final data set (99,601 aligned nucleotides) that were used to reconstruct the phylogeny with RAxML v.8.2.4 under the GTRGAMMA model and 1,000 bootstrap replicates.

Genome size evolution

Genome size evolution was modelled by ML using the 'fastAnc' function in the phytools R package[76]. We used a time-calibrated tree representing all major jawed vertebrate lineages obtained from the phylotranscriptomic tree of ref. [5]; ages are a genome-wide estimates across 100 time-calibrated trees inferred from 100 independent gene jackknife replicates inferred in PhyloBayes v.4.1[77] under a log-normal autocorrelated clock model with 16 cross-validated fossils as uniform calibrations with soft bounds, the CAT-GTR substitution model and a birth-death tree prior. Genome size data (haploid DNA content or c-value) were obtained from ref. [78]. Genome size estimates were averaged per species (if multiple were available) and in six species genome size was approximated as the average of closely related species within the same genera. For *Neoceratodus*, the k-mer based estimation was used (43 Gb; c-value = 43.97 pg). Ancestral genome sizes were used to calculate the rates of genome evolution for selected branches.

Molecular clock analyses

Divergence times were inferred with a relaxed molecular clock with autocorrelated rates, as implemented in MCMCTree within the PAML package v.4.9h[79]. A total of 6 fossil calibrations were used as uniform priors[80]. For further details see Supplementary Methods.

Dynamics of gene family size

CAFE[81] was used to infer gene birth/death rates (lambda) and retrieve gene families under significant dynamics. As input we took the species tree with divergence time from the output of MCMCTREE, and the results of gene clusters from Hcluster_sg. Each gene cluster was deemed as gene family. We run CAFE under the model where a global lambda was set across the whole tree. To symbolize each gene family, we took the longest member as representative and BLAST-searched with diamond[82] against SWISSPROT and NR databases. From both the best hit was retained.

To compare the repertoire of olfactory receptors, taste receptors and pulmonary surfactant proteins across all studied species, we followed the same procedure for each species. First, we collected sequences of olfactory receptors, taste receptors and pulmonary surfactant proteins from Swiss-Prot and NR database as query. For sequences from NR database, we only kept those with ID starting with "NP_", which are supported by the RefSeq eukaryotic curation group. Second, we mapped the query set to each genome using Exonerate in server model (maxintron set to 6M for lungfish and axolotl). The alignment was extended to start/stop codon when possible. Third, we BLAST-searched all retrieved sequences to NR database and removed those with best hit not an olfactory receptor, taste receptor or pulmonary surfactant. The final result sequences had alignment coverage ranging from 32% to 100% (first quartile 95%), and percentage of identity from 17% to 100 (first quartile 62%) to its query.

Following a previous study[83] we defined the final sequences into three categories based on their alignment to its query: 1) pseudogene, sequences with premature stop codon or frameshift; 2) truncated gene, sequences without premature stop codon and frameshift but broken ORF (start or stop codon missing); 3) intact gene, sequences with intact ORF.

Positive selection analysis

Two different models were calculated. Model 1 to find genes positively selected in lungfish and model 2 for genes commonly positively selected in tetrapods and lungfish. Genomes included were *Neoceratodus forsteri*, *Ambystoma mexicanum* (this study), Ensembl genomes: *Danio rerio* (Danio_rerio.GRCz11), *Anolis carolinensis* (Anolis_carolinensis.AnoCar2.0), *Lepisosteus oculatus* (Lepisosteus_oculatus.LepOcu1), *Latimeria chalumnae* (Latimeria_chalumnae.LatCha1), *Callorhinchus milii* (Callorhinchus_milii.Callorhinchus_milii-6.1.3), *Xenopus tropicalis*

(GCF_001663975.1_Xenopus_laevis_v2), *Gallus gallus* (Gallus_gallus.GRCg6a) and *Homo sapiens* (Homo_sapiens.GRCh38). The *Xenopus tropicalis* genome (GCF_001663975.1_Xenopus_laevis_v2) was downloaded from NCBI. Protein and cDNA files from all species were downloaded. To identify orthologous proteins, all protein sequences were compared to lungfish using Inparanoid[84] (default settings). To match protein and cDNA, sequences were searched by TBLASTN and only 100% hits were kept. Codon alignments for the protein/cDNA sequence pairs were constructed using pal2nal v.14[85]. Resulting sequences were aligned by MUSCLE[86] (option: -fastaout) and poorly aligned positions and divergent regions of cDNA were eliminated by Gblocks v.0.91b[87] (options: -b4 10 -b5 n –b3 5 –t=c). An inhouse script was used to convert the Gblocks output to PAML format.

As phylogenetic tree we took the species tree with divergence times from MCMCTREE as input for detection of positive selection with *Callorhinchus milii* as outgroup. For the phylogenetic analyses by maximum likelihood the 'Environment for Tree Exploration' (ETE3) toolkit[88], which automates CodeML and Slr analyses by using pre-configured evolutionary models, was used. For detection of genes under positive selection in lungfish, we compared the branch-specific model bsA1 (neutral) with model bsA (positive selection) using a likelihood ratio test (FDR <=0.05). To detect sites under positive selection Naive Empirical Bayes (NEB) probabilities for all 4 classes were calculated for each site. Sites with a probability > 0.95 for either site class 2a (positive selection in marked branch and conserved in rest) or 2b (positive selection in marked branch and relaxed in rest) were considered. Two models were calculated. In model 1 only the branch for lungfish was marked, in model 2 all tetrapods and lungfish were marked for positive selection.

Functional clustering was done with IPA (QIAGEN Inc., qiagenbioinformatics.com/products/ingenuity-pathway-analysis) and DAVID (https://david.ncifcrf.gov/home.jsp) using human homologs with default settings.

In situ hybridization

In situ hybridisation was performed as described[34,89] with modifications (see Supplementary Methods).

Hox gene RNAseq analysis

Hox gene RNAseq analysis was performed on a st. 52 lungfish larva RNAseq dataset (SRR6297462-SRR6297470)[37] (see Supplementary Methods).

Limb enhancer analysis

330 non-redundant VISTA enhancer elements[26] were searched by BLASTN against *Xenopus laevis*, *Xenopus tropicalis*, *Nanorana parkeri*, axolotl, reedfish, sterlet, gar, elephant shark, coelacanth (LatCha1), and *Neoceratodus* genomes to determine conservation (see Supplementary Methods).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Data and all codes will be publicly available at the time of publication at Github and NCBI Bioproject (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA644903)

Code availability

Custom code has been deposited at https://github.com/labtanaka/meyer_lungfish For the current MARVEL source code repository see https://github.com/schloi/MARVEL. For sample execution scripts csee: https://github.com/schloi/MARVEL/examples/.

36. Nagano, T. *et al*. Comparison of Hi-C results using in-solution versus in-nucleus ligation. *Genome Biol*. **16**, 175 (2015).
37. Wutz, G. *et al*. Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *EMBO J*. **36**, 3573–3599 (2017).
38. Smith, J. J. *et al*. A chromosome-scale assembly of the axolotl genome. *Genome Res*. **29**, 317–324 (2019).
39. Nowoshilow, S. & Tanaka, E. M. Introducing www.axolotl-omics.org – an integrated -omics data portal for the axolotl research community. *Exp. Cell Res*. **394**, 112143 (2020).
40. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
41. Song, L. & Florea, L. Rcorrector: Efficient and accurate error correction for Illumina RNA-seq reads. *GigaScience* **4**, 48 (2015).
42. MacManes, M. D. The Oyster River Protocol: a multi-assembler and kmer approach for de novo transcriptome assembly. *PeerJ* **6**, e5428 (2018).
43. Chikhi, R. & Medvedev, P. Informed and automated k-mer size selection for genome assembly. *Bioinformatics* **30**, 31–37 (2014).
44. Robertson, G. *et al*. De novo assembly and analysis of RNA-seq data. *Nat. Meth*. **7**, 909–912 (2010).
45. Emms, D. M. & Kelly, S. OrthoFinder: Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol*. **16**, 157 (2015).
46. Li, W., Jaroszewski, L. & Godzik, A. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* **17**, 282–283 (2001).
47. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
48. Nishimura, O., Hara, Y. & Kuraku, S. gVolante for standardizing completeness assessment of genome and transcriptome assemblies. *Bioinformatics* **33**, (2017).
49. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: A fast spliced aligner with low memory requirements. *Nat. Meth*. **12**, 357–360 (2015).
50. Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc*. **11**, 1650–1667 (2016).
51. Yan, H., Bombarely, A. & Li, S. DeepTE: A computational method for de novo classification of transposons with convolutional neural network. *Bioinformatics* **36**, 4269–4275 (2020).
52. Chalopin, D. & Volff, J.-N. Analysis of the spotted gar genome suggests absence of causative link between ancestral genome duplication and transposable element diversification in teleost fish. *J. Exp. Zool. Part B Mol. Dev. Evol*. **328**, 629–637 (2017).
53. Lerat, E., Fablet, M., Modolo, L., Lopez-Maestre, H. & Vieira, C. TEtools facilitates big data expression analysis of transposable elements and reveals an antagonism between their activity and that of piRNA genes. *Nucleic Acids Res*. **45**, e17 (2017).
54. Wu, T. D. & Watanabe, C. K. GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
55. Haas, B. J. *et al*. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res*. **31**, 5654–5666 (2003).
56. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol*. **34**, 525–527 (2016).
57. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res*. **47**, D506–D515 (2019).
58. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
59. Pearson, W. R. Finding protein and nucleotide similarities with FASTA. *Curr. Protoc. Bioinformatics* **53**, 3.9.1-3.9.25 (2016).
60. Chan, P. P. & Lowe, T. M. tRNAscan-SE: Searching for tRNA genes in genomic sequences. *Methods Mol. Biol*. **1962**, 1–14 (2019).
61. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
62. Kozomara, A., Birgaoanu, M. & Griffiths-Jones, S. miRBase: from microRNA sequences to function. *Nucleic Acids Res* **47**, D155–D162 (2019).
63. Enright, A. J. *et al*. MicroRNA targets in *Drosophila*. *Genome Biol*. **5**, R1 (2003).
64. Harris, R. Improved pairwise alignment of genomic DNA. (Pennsylvania State University, 2007).
65. Kent, W. J. *et al*. The Human Genome Browser at UCSC. *Genome Res*. **12**, 996–1006 (2002).
66. Blanchette, M. *et al*. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res*. **14**, 708–715 (2004).
67. Siepel, A. *et al*. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*. **15**, 1034–1050 (2005).
68. Cooper, G. M. *et al*. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res*. **15**, 901–913 (2005).
69. Cho, Y. S. *et al*. The tiger genome and comparative analysis with lion and snow leopard genomes. *Nat. Commun*. **4**, 2433 (2013).
70. Ruan, J. *et al*. TreeFam: 2008 Update. *Nucleic Acids Res*. **36**, D735-740 (2008).
71. Whelan, S., Irisarri, I. & Burki, F. PREQUAL: detecting non-homologous characters in sets of unaligned homologous sequences. *Bioinformatics* **34**, 3929–3930 (2018).
72. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol*. **30**, 772–780 (2013).
73. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): A new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol*. **10**, 210 (2010).
74. Lartillot, N., Rodrigue, N., Stubbs, D. & Richer, J. PhyloBayes MPI: Phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst. Biol*. **62**, 611–615 (2013).
75. Dutheil, J. Y., Gaillard, S. & Stukenbrock, E. H. MafFilter: A highly flexible and extensible multiple genome alignment files processor. *BMC Genomics* **15**, 53 (2014).
76. Revell, L. J. phytools: An R package for phylogenetic comparative biology (and other things). *Meth. Ecol. Evol*. **3**, 217–223 (2012).

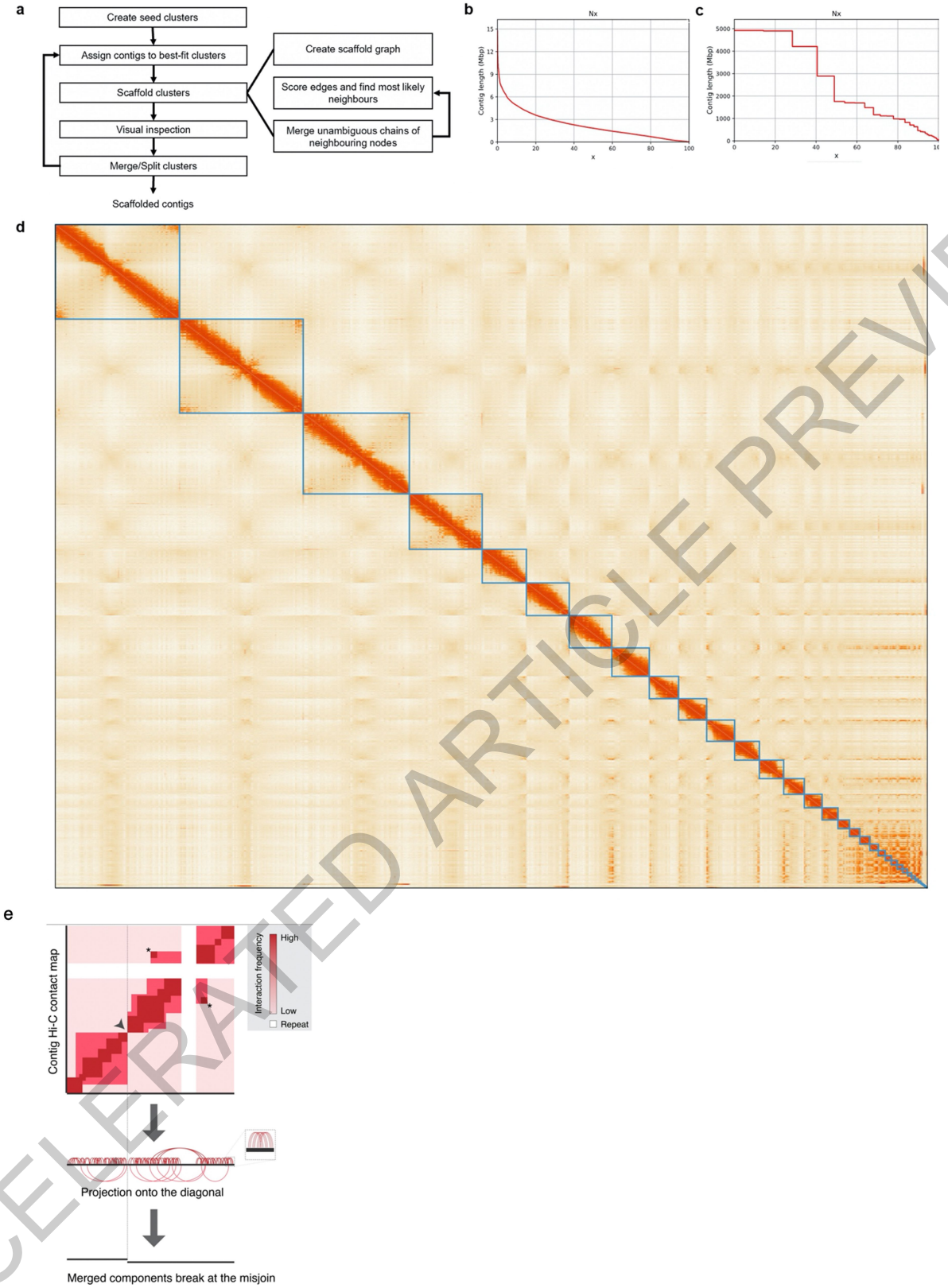
77. Lartillot, N., Lepage, T. & Blanquart, S. PhyloBayes 3: A Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**, 2286–2288 (2009).
78. Gregory, T. R. Animal Genome Size Database. http://www.genomesize.com. (2020).
79. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
80. Marjanović, D. Recalibrating the transcriptomic timetree of jawed vertebrates *bioRxiv* (2019) https://www.biorxiv.org/content/10.1101/2019.12.19.882829v1.
81. De Bie, T., Cristianini, N., Demuth, J. P. & Hahn, M. W. CAFE: A computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269–1271 (2006).
82. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat Meth* **12**, 59–60 (2015).
83. Niimura, Y. Olfactory receptor multigene family in vertebrates: from the viewpoint of evolutionary genomics. *Curr. Genomics* **13**, 103–114 (2012).
84. O'Brien, K. P., Remm, M. & Sonnhammer, E. L. L. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.* **33**, D476-480 (2005).
85. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nuleic Acids Res.* **34**, W609–W612 (2006).
86. Edgar, R. C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
87. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540–552 (2000).
88. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol* **33**, 1635–1638 (2016).
89. Woltering, J. M. *et al*. Axial patterning in snakes and caecilians: evidence for an alternative interpretation of the Hox code. *Dev. Biol.* **332**, 82–89 (2009).
90. Bickelmann, C. *et al*. Noncanonical Hox, Etv4, and Gli3 gene activities give insight into unique limb patterning in salamanders. *J. Exp. Zool. B Mol. Dev. Evol.* **330**, 138–147 (2018).

Acknowledgements We dedicate this paper to the memories of Jenny A. Clack from Cambridge University and Robert L. Carroll from the Redpath Museum at McGill University whose groundbreaking work on the paleontology of sarcopterygian fishes contributed to our understanding of the water-land transition of vertebrates. This work was supported by the German Science Foundation (DFG) through a grant to AM, TB and MS (Me1725/24-1, Bu956/23-1, Scha408/16-1), to JMW (Wo2165/2-1) and core funding from the IMP to EMT. JNV and MS were supported by a joint grant of the French Research Agency (ANR Evobooster) and DFG (SCHA408/13-1). II was supported by the Spanish Ministry of Economy and Competitiveness (MINECO) (Juan de la Cierva-Incorporación fellowship IJCI-2016- 29566) and the European Research Council (Grant Agreement No. 852725; ERC-StG 'TerreStriAL' to Jan de Vries, University of Goettingen). WYW and OS were supported by the Austrian Science Fund grants P3219 and I 4353. WYW is supported by Croucher Scholarships for Doctoral Study. AK was supported by a fellowship from the Japanese Society for the Promotion of Science (JSPS) postdoctoral fellowship for Overseas Researchers Program. We thank Daniel Ocampo Daza (www.egosumdaniel.se) for generously sharing his vertebrate illustrations, J. Joss and P. Sordino for gift of lungfish embryos, and L. Pennacchio for Vista enhancer images.

Author contributions A.M., T.B., M.S. conceived the study and coordinated the work. A.M., and MS wrote the manuscript with contributions from all other authors. S.S.: genome assembly into contigs and HiC scaffolding. P.F.: transcriptome analysis, annotation and CNE analyses. K.D.: genome annotation, analysis of gene family dynamics, genome expansion. JMW: analysis and annotated *hox* clusters, embryonal RNA-seq, *in situ* hybridization. I.I.: phylogenetic analyses, molecular clock, ancestral character state reconstruction. W.Y.W.: repeat and syntenic analysis. S.N.: genome correction and initial transcript alignment. S.K.: Positive selection analysis. A.K.: HiC library preparation, library prep for genome correction. A.F.: transcriptome generation. P.X.: annotation of ncRNAs. C.D., J.N.V.: transposon and repeat analysis. H.S.: contributed resources. O.S.: syntenic analyses. E.T.: supervised HiC, genomic sequencing, analyzed data.

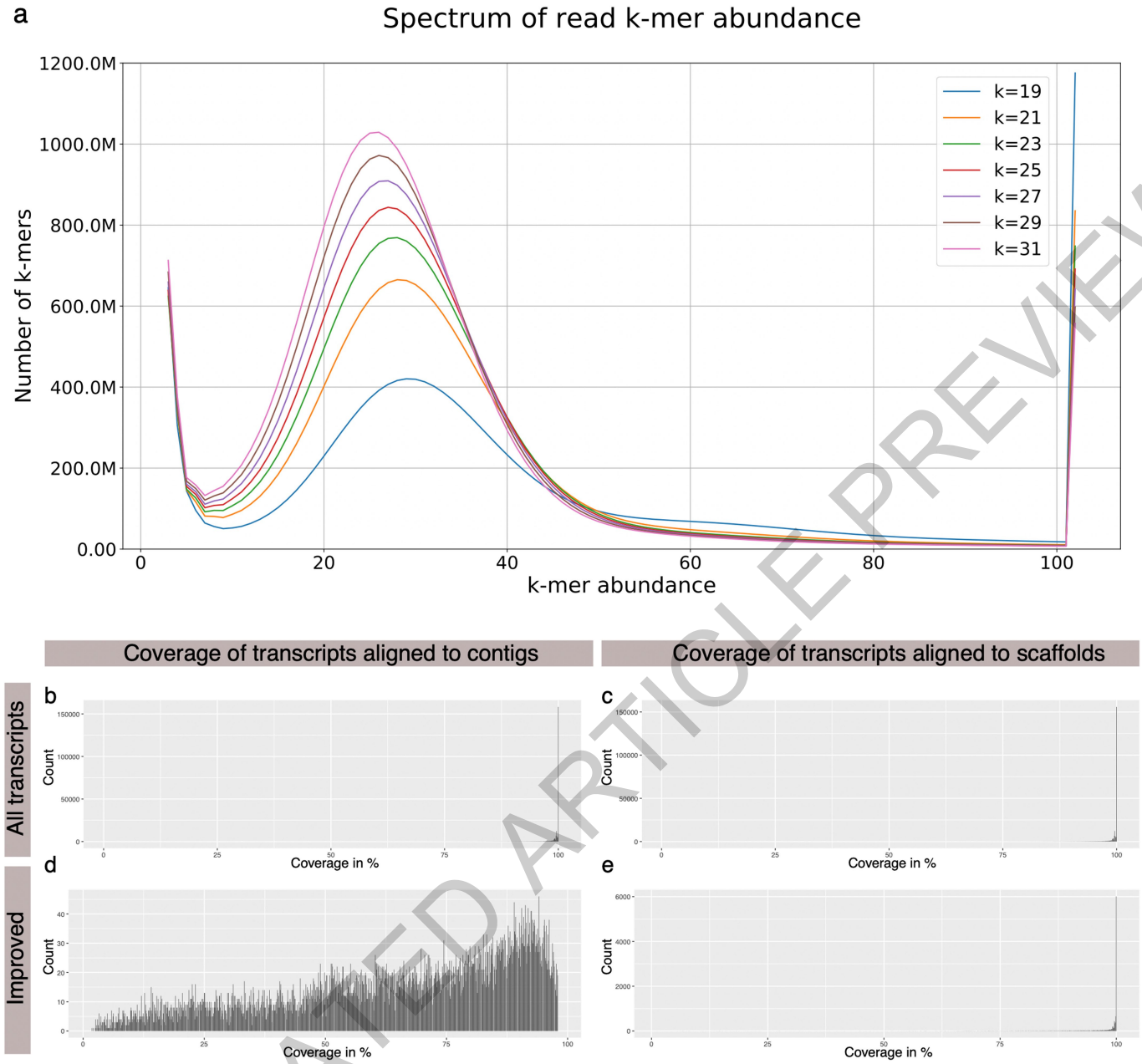
Competing interests The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-03198-8.
Correspondence and requests for materials should be addressed to A.M., O.S., T.B., E.M.T. or M.S.
Peer review information *Nature* thanks Benedict Paten, Igor Schneider, Ryan Lorig-Roach, Marina Haukness and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer review reports are available.
Reprints and permissions information is available at http://www.nature.com/reprints.

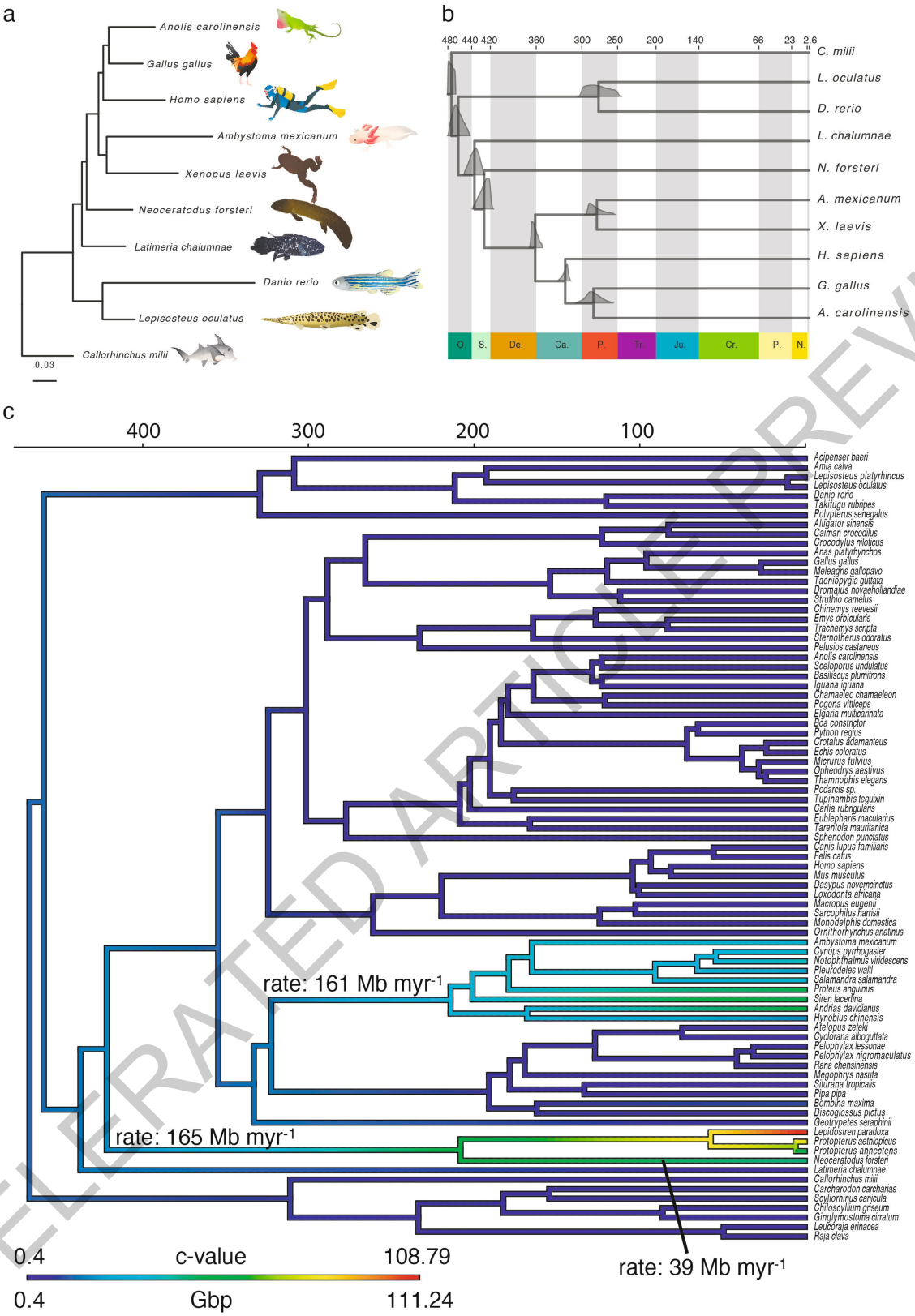


Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Schematic overview of the scaffolding procedure. **a**, Scaffolding consists conceptually of two nested loops. The inner loop, depicted on the right takes a list of contigs, their contact information and iteratively performs a global agglomerative clustering until convergence or until no more contigs can be joined. This loop is nested in the main procedure, which takes as input a list of seed contigs, assigns contigs these initial clusters, scaffolds these and allows for visual inspection and merging/splitting of the clusters. **b**, N(x) plot of the assembled contigs. On the Y-axis the contig length is shown for which the collection of all contigs of that length or longer covers at least x% (X-axis) of the assembly. **c**, N(x) plot of the scaffolded genome. On the Y-axis the contig length is shown for which the collection of all scaffolds of that length or longer covers at least x% (X-axis) of the assembly. **d**, HiC contact heat map of the scaffolded portion of the Lungfish genome assembly, ordered by scaffold length. Blue boxes indicate the scaffold boundaries. The four largest scaffolds represent both chromosome arms on a single scaffold. Remaining scaffolds are split into chromosome arms or represent microchromosomes. **e**, Schema illustrating the contig misjoin detection process. Hi-C contacts are binned along the diagonal (inserts a,b). Points that not crossed by a sufficient number of contacts are deemed potential misjoins and are thus separated (insert c, dotted line).

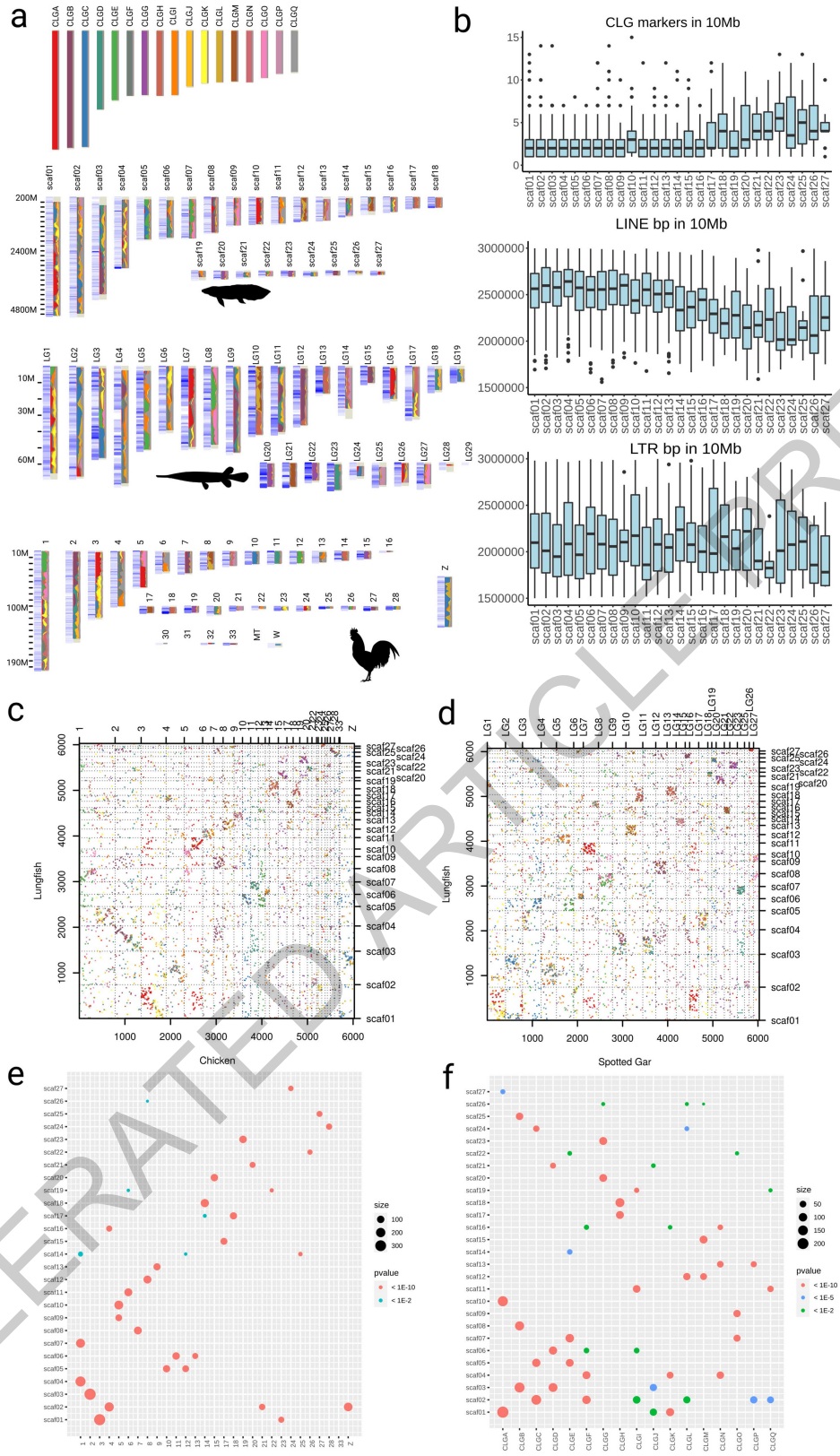


Extended Data Fig. 2 | k-mer frequency analysis and transcript coverage by genomic sequences. a, The Illumina dataset was used to generate the spectra of k-mer abundances using 7 different k-mer sizes. **b-e**, Transcript coverage by genomic sequences. **b**. Histogram of the proportion of all transcript lengths covered by the alignment to contigs. **c**, Histogram of the proportion of all transcript lengths covered by the alignment to scaffolds. **d**, Histogram of the proportion of the transcript lengths covered by the alignment to contigs or **e**, to scaffolds of those transcripts whose alignment was improved after scaffolding.



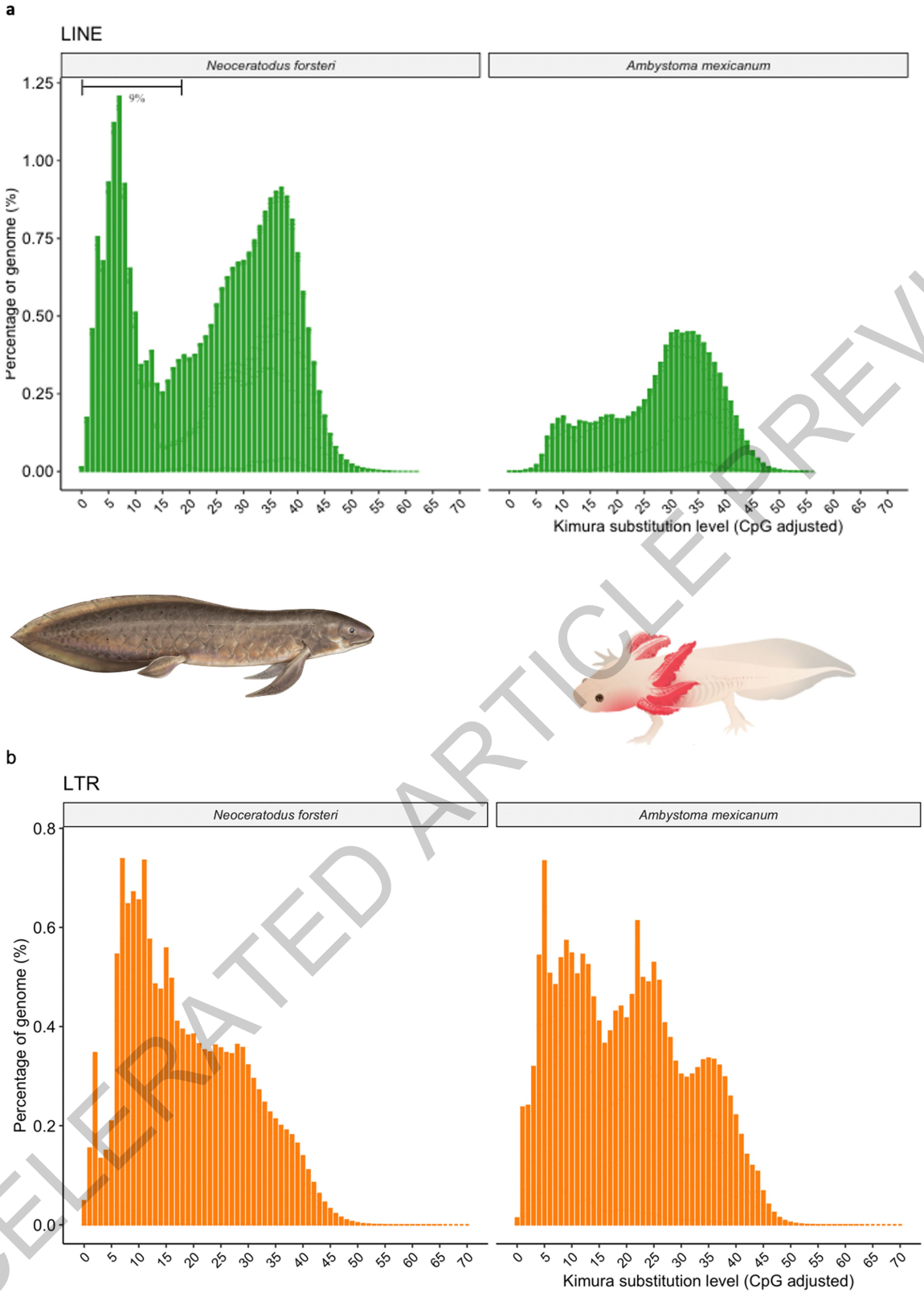
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | CNE-based phylogeny, divergence times and rates of genome evolution. a, Maximum likelihood phylogeny from non-coding conserved alignment blocks totaling 99,601 informative sites (using RAxML; GTRGAMMA). All branches were supported by 100% bootstrap value and scale bar is in expected nucleotide replacements per site. Branch lengths of the trees obtained by the CNE method or from the protein sequences show a high correlation (R2=0.84; p<0.05). **b**, Relaxed clock time-calibrated phylogeny (MCMCTree). Plots at nodes correspond to full posterior distribution of inferred ages. Scale is in million years ago and main geologic periods are highlighted. Plot generated with MCMCTreeR (https://github.com/PuttickMacroevolution/MCMCtreeR). **c**, Evolution of genome size in jawed vertebrates. Maximum likelihood reconstruction of ancestral genome sizes using a time-calibrated phylotranscriptomic tree[8] and genome size values obtained from ref. [78]. Branch lengths are in millions years ago and colors denote genome size (c-value in pg or Gbp). Rates of genome expansion are given for the ancestral branches of lungfishes and salamanders, as well as for the *Neoceratodus* terminal branch.

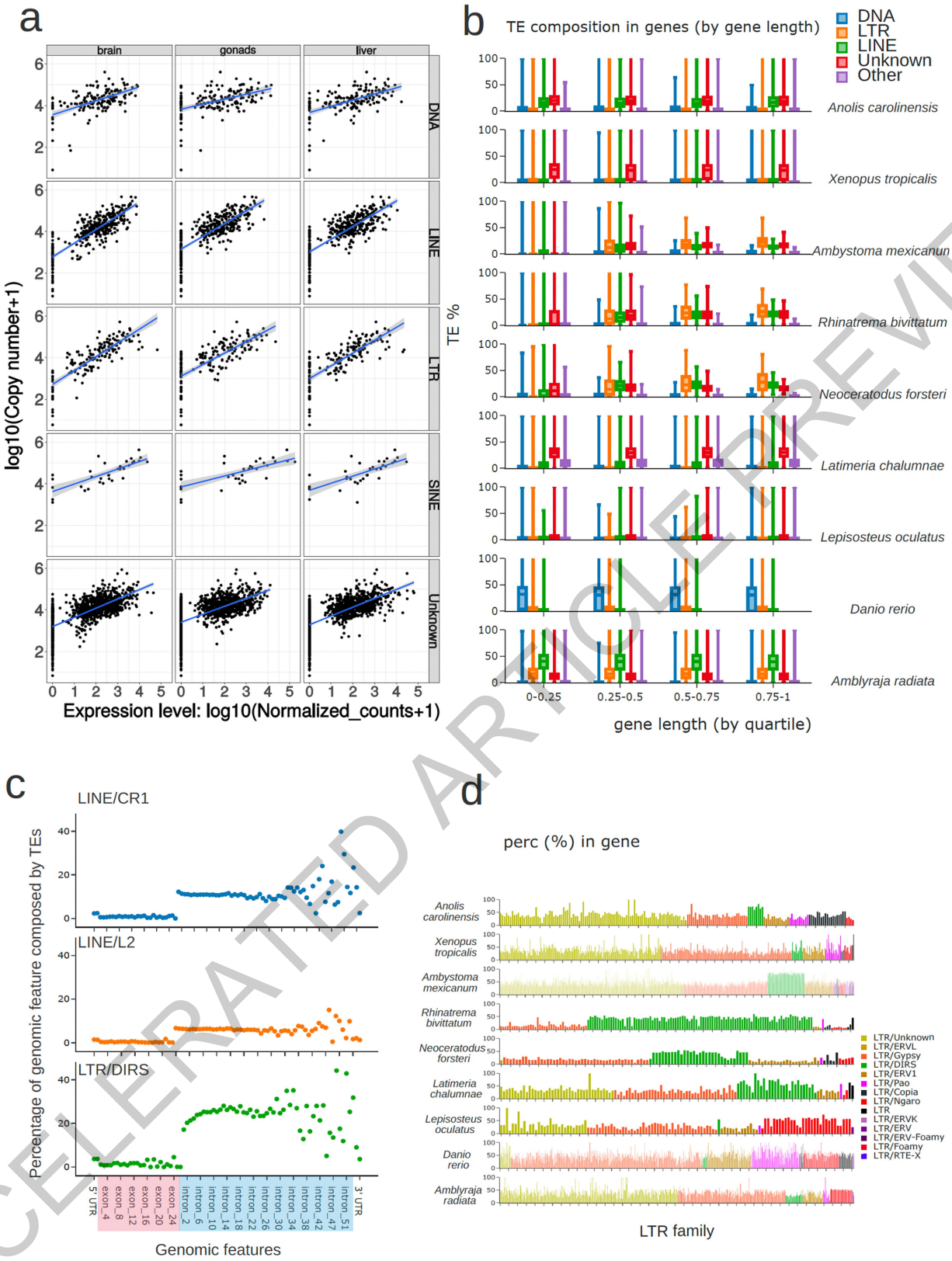


Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Chordate linkage group (CLG), gene and repeat density along lungfish chromosomes. a, CLG content profiled within windows of 20 genes with available orthology and CLG identity and using a 10 gene sliding window. If genes were more than 10Mb or 100Mb apart in gar or lungfish, respectively, breaking the 20 gene window, the area is highlighted as grey, indicating areas lacking sufficient amount of orthologous CLG markers. Blue bar indicates gene density (as measured by the 6,337 marker genes used in the CLG analysis) along 10Mb windows – white or grey indicates gene desert, blue indicates gene-rich areas. Upper row: previously reconstructed CLGs and their colour labels, followed by lungfish, spotted gar and chicken. **b**, Gene and repeat density along 10Mb windows on lungfish chromosomes. Y-axis shows count of CLG genes, LINE, and LTRs per 10Mb window, respectively. Microchromosomes show higher gene density and lower LINE density, while LTR density remains stable. **c**, Conserved macrosynteny between lungfish and **c**, chicken and **d**, spotted gar. Chromosomes of chicken (**c**) and gar (**d**) are plotted along with their homologous lungfish chromosomes. The majority of the chromosomes and co-linearity are retained one-to-one. Some recent incorporation of microchromosomes into lungfish macrochromosomes (scaf02) has occurred, as evident by sharp syntenic boundaries. **e**+**f**, Significance of the association (homology) between chicken and lungfish chromosomes. Colors correspond to the significance power of the association, or -log10 (adjusted Fisher's exact test p-values). Fisher test was run on the number of orthologous gene families shared between any given pair of chromosomes in chicken and lungfish, compared to the overall distribution of orthologous gene families on all other chromosomes. Most chicken microchromosomes (chromosome 6 onwards) are in one-to-one correspondence between lungfish and chicken, but some were recently incorporated into macrochromosomes. Those lungfish macrochromosomes, e.g., scaffold 01 or scafold 02, have significant association with both chicken macro and microchromosomes. However, those fusions are very recent in lungfish, because the positions of chicken orthologs is restricted to specific areas of the lungfish chromosome (also seen as a clear boundary in Fig. 2c). "Size" refers to the number of shared orthologous gene families between homologous chromosomes. **f**, Significance of the association (homology) between Chordate Linkage Groups and lungfish chromosomes. Fisher test was run on the number of orthologous gene families shared between any given pair of chromosomes in CLG and lungfish, compared to the overall distribution of orthologous gene families on all other chromosomes. Silhouette of the lungfish is from[34].

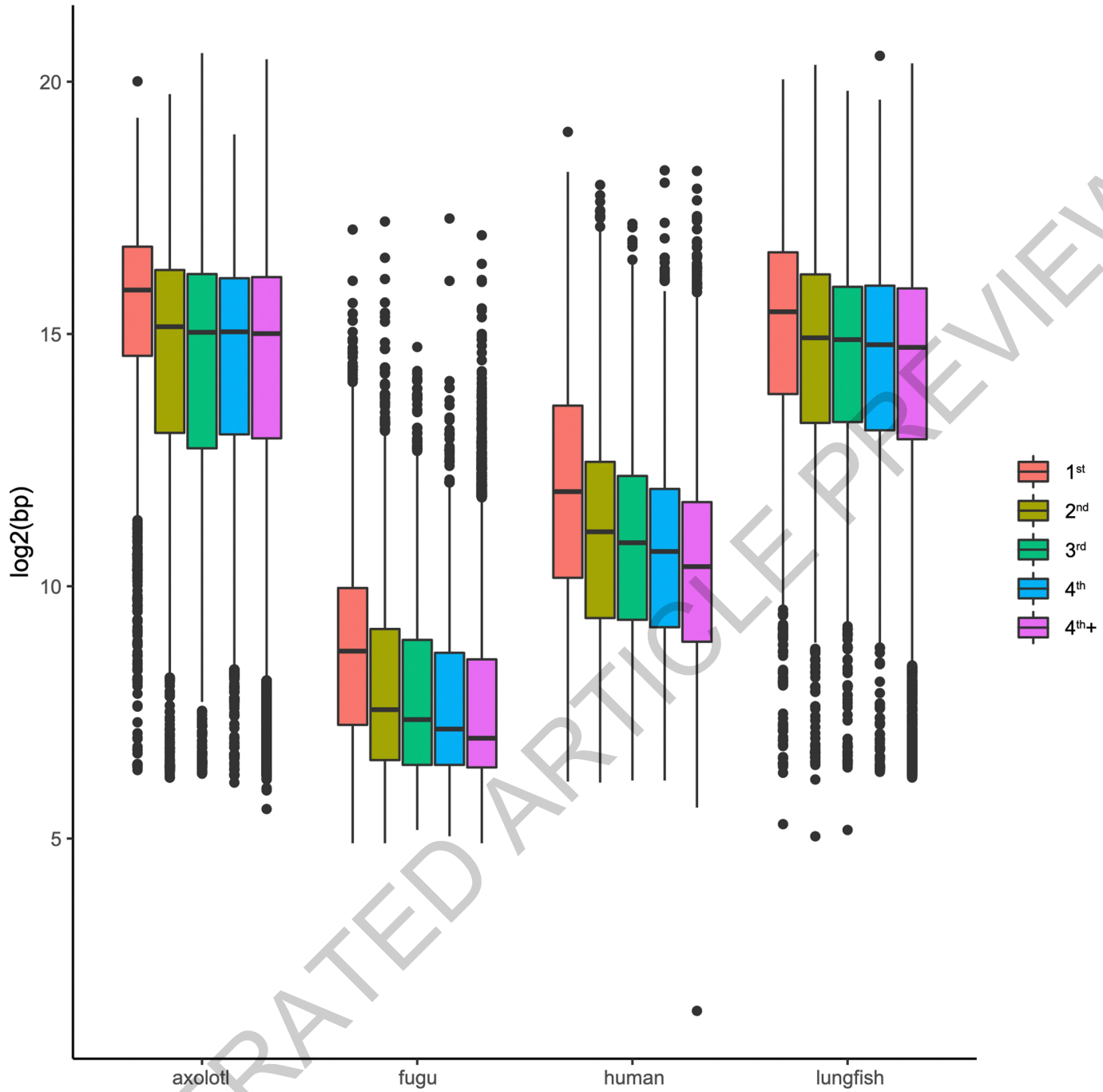


Extended Data Fig. 5 | Age estimation plots on LINE and LTR classes (Kimura plots). a, Repeat landscape of LINE and **b**, LTR of lungfish and axolotl. The two main peaks indicate there were two major LINE expansions in lungfish. The recent expansion (div <=15% from the consensus sequences) contributed to 9% of the lungfish genome. The LTR landscapes are similar in these two species.

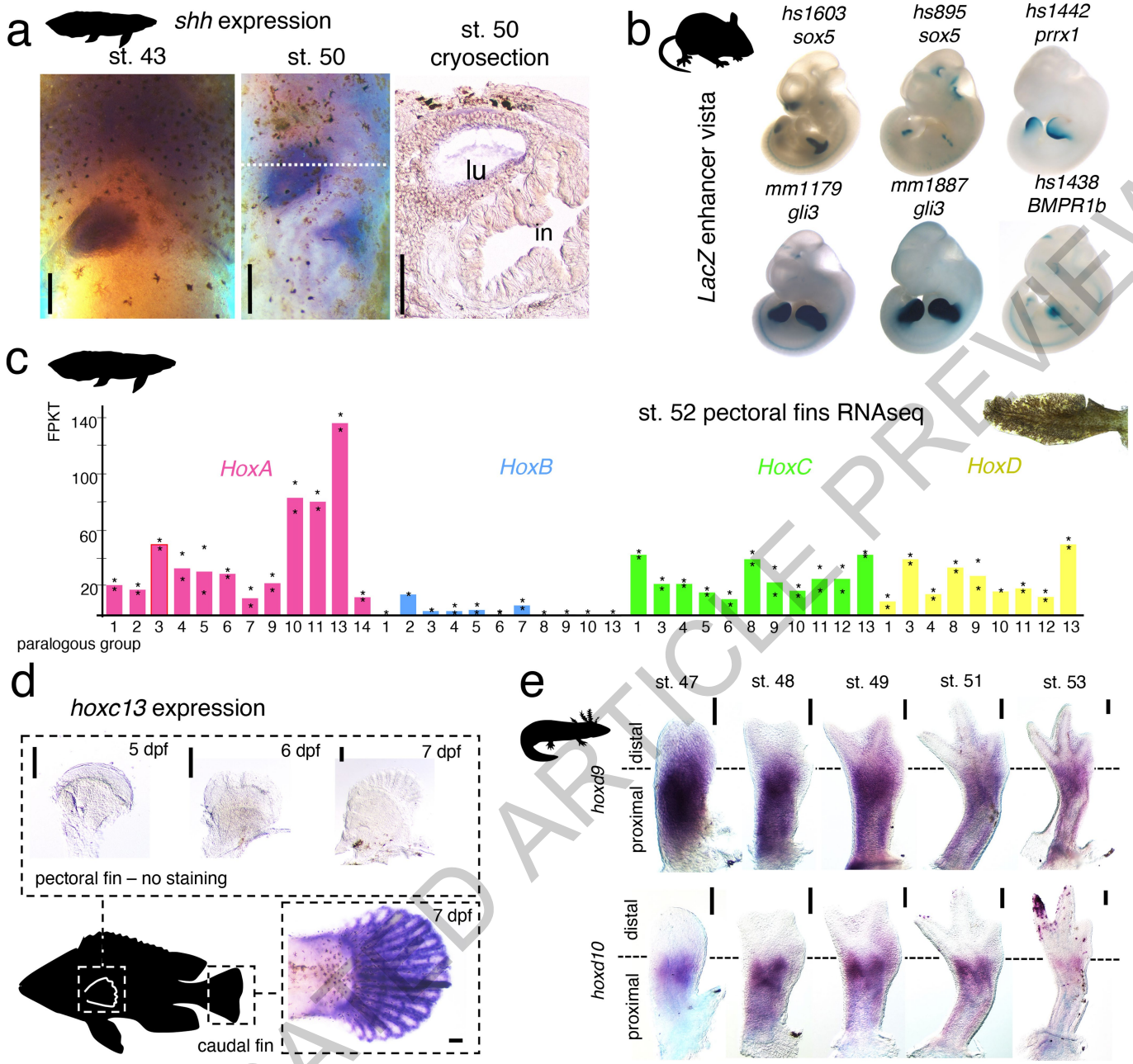


Extended Data Fig. 6 | See next page for caption.

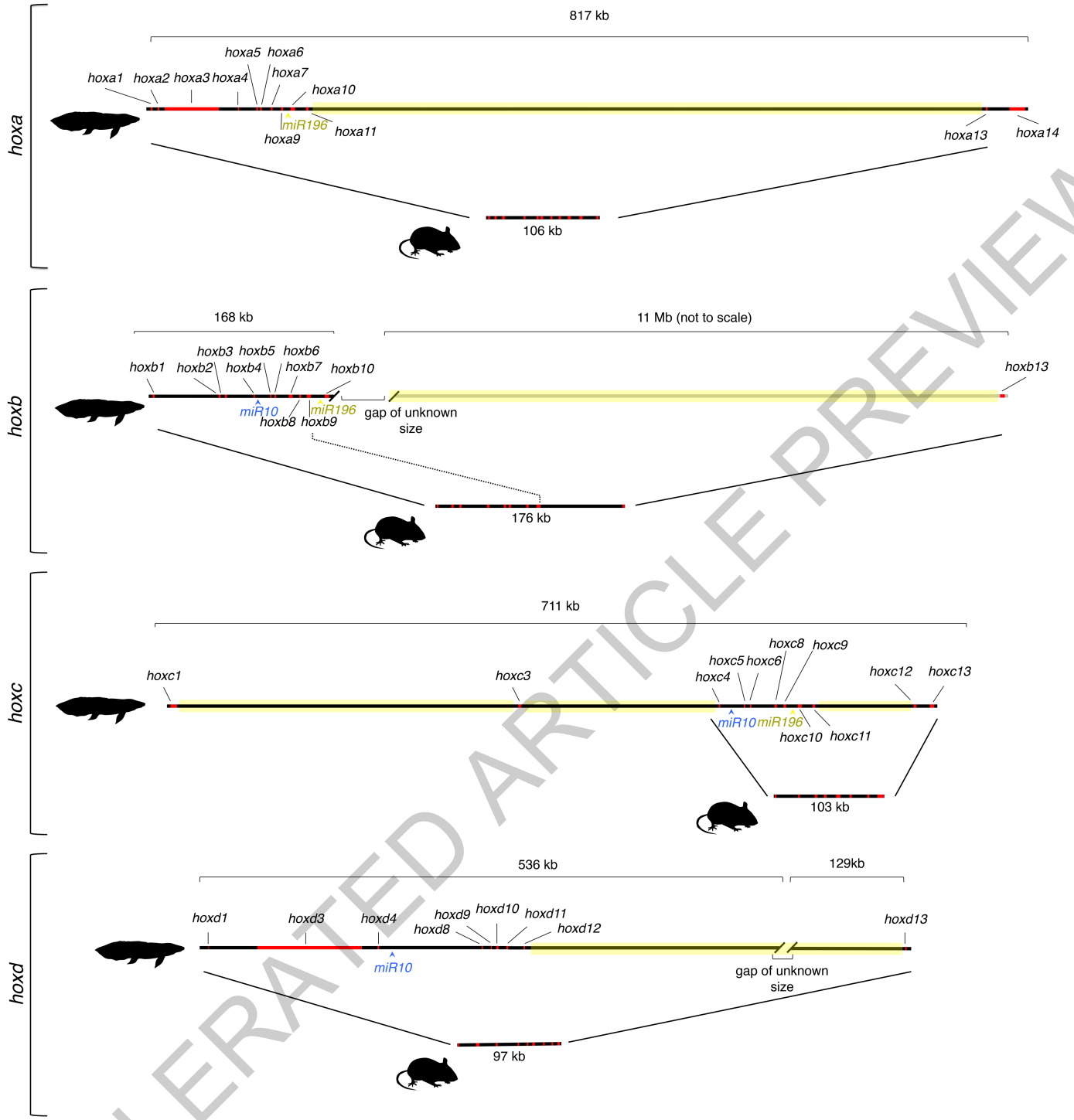
Extended Data Fig. 6 | Correlation between TE family expression and copy number in the genome. a, Expression was estimated for each TE family using polyA-enriched RNAseq data from gonad, brain and liver. For all tissues and TE classes a positive correlation is observed between expression level and copy number: when a TE family is highly expressed, this family tend to have more copies. However, some families are distant from the correlation line, with a high expression and low copy number or vice versa. The expression levels of TE families are globally correlated in the three tissues. **b**. Composition of different classes of repetitive elements in genic regions. Gene and repetitive element annotations were obtained from published reference genomes (see method Repeats and transposable elements annotation section). The percentage of different classes of repetitive elements in genic region (including UTRs, exons and introns) were calculated as percentage of the number of basepair (bp) covered by the repetitive element, normalised by the size of the genes. Genes are grouped by length. As the size of genes varies across species, we grouped them by quartile division per species. The genic LTR% (orange) increases in longer genes in lungfish, axolotl and caecilian (vertical lines show the minimum and maximum of the percentage of TE's in genes). The boxplot shows the median, and the 25% and the 75% quartiles; whiskers show 1.5 times the interquartile range. Outliers extend beyond 1.5× interquartile ranges from either hinge. **c**, Percentage of the genic regions that are occupied by different classes of transposable elements. (Top and middle) LINE/CR1 and LINE/L2, which are classified in the same clade of LINE and are closely related, compose ~5.1% and 2.9% of the lungfish genome, respectively. (Bottom) On average, introns (blue) harbor a high number of LTR/DIRS (~20 to 30%) than exons (red). **d**, Percentage of LTR families in genic regions (including UTRs). The LTR/DIRS is enriched in genic regions in lungfish and axolotl.



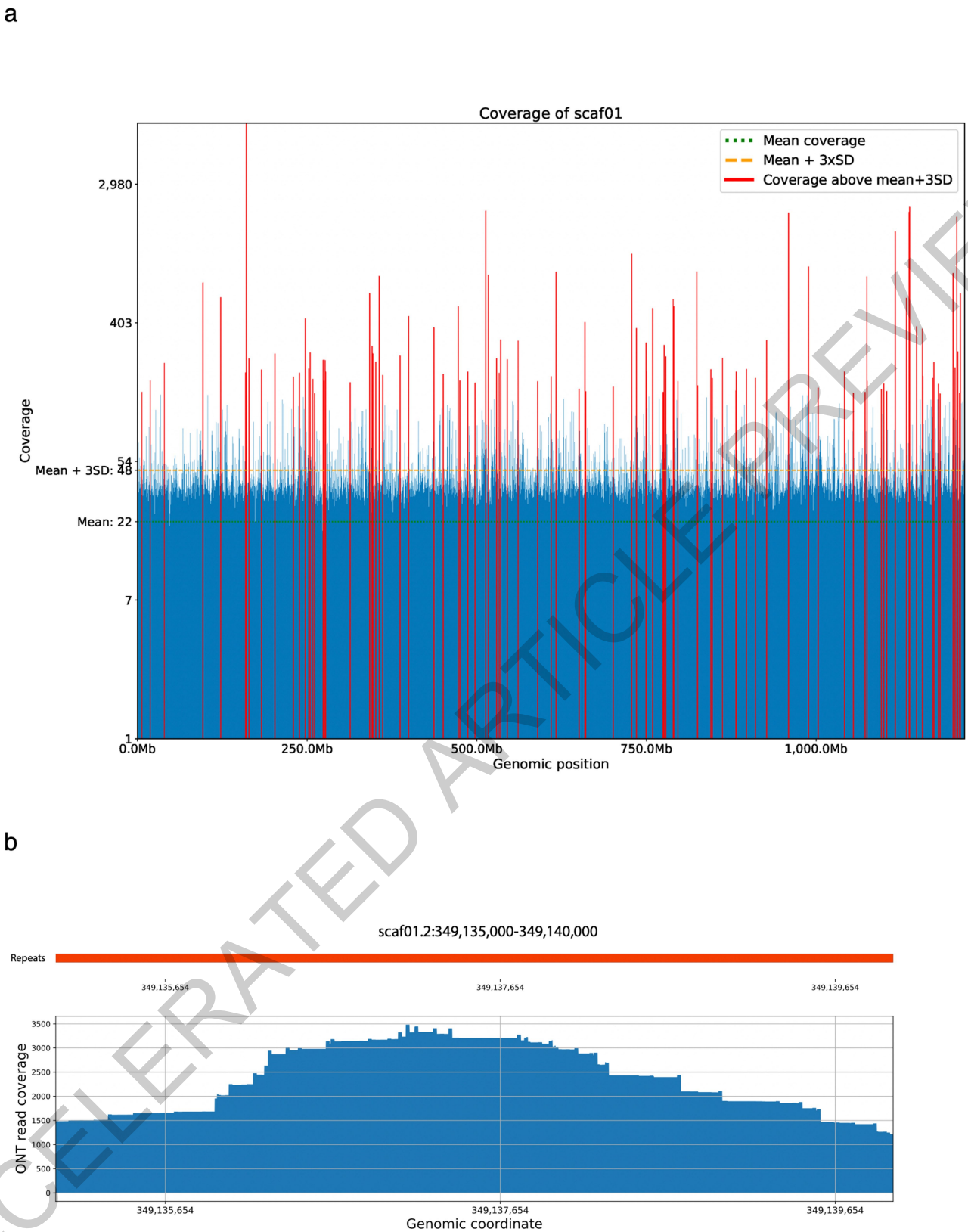
Extended Data Fig. 7 | Boxplot of intron sizes in axolotl, fugu, human and lungfish. For axolotl (axo), fugu (fug), human(hum) and lungfish (lung) the length (Y-axis is log2 scale of base pairs) of the 1st, 2nd, 3rd, 4th, 5th+ introns show a consistent pattern with the 1st intron always being the longest intron, both in the giant lungfish and axolotl genomes as well as in the tiny fugu (400Mb) genome.



Extended Data Fig. 8 | Gene expression data in Australian lungfish, ray-finned fish and axolotl salamader. a, *Neoceratodus* only has a single right lung. *Shh* expression in the *Neoceratodus* lung Anlage. st. 43 ventral view (N = 1/1), anterior up. st. 48 ventral view, anterior up (N = 1/1). The appearance of the lung *Anlage* and *shh* expression is similar to that in *Xenopus*. Transverse section across dotted line. Abbreviations: lu; lung, in: intestine. Scale bars 0.2 mm. **b**, LacZ enhancer assays in mouse 12 dpf embryos show the regulatory activity of several ultra-conserved enhancers that emerged in association with the evolution of the lobed fin. These include elements located near important limb developmental genes that contribute to the sturdy sarcopterygian fin archetype (also see main text and supplementary results). Reported LacZ limb expression: *hs1603* N = 7/7, hs895 N = 5/8, *hs1442* N = 10/11, *mm1179* N = 7/7, *mm1887* N = 6/6, *hs1438* N = 5/11. **c**, Hox gene expression from RNA-seq analysis of stage 52 pectoral fins (N = 2). Individual datapoints shown with asterisks, the hight of the bar indicates average expression. Overlapping datapoints indicated with a single asterisk. High expression of posterior *hoxa* and *hoxd* genes (except for *hoxa14*), low expression of *hoxb* genes and unexpectedly high expression of *hoxc* genes. **d**, Absence of *hoxc13* expression from pectoral, but not caudal fins in the ray finned cichlid *Astatotilapia burtoni*. A staging series of cichlid pectoral fins (5-7 days post fertilization) does not show expression of *hoxc13*, whereas this gene stains strongly in the caudal fin (N = 4/4 embryos per stage). This result is consistent with a sarcopterygian origin of *hoxc13* expression in the distal paired fins and limbs. Scale bars 0.1 mm. **e**, Non canonical patterns of *hoxd9* and *hoxd10* expression in axolotl limbs (N = 2/2 limbs per stage). Expression of *hoxd9* and *hoxd10* during Axolotl limb development shows strong expression in a proximal limb domain but absence or low expression in the distal limb/digit domain. This non-canonical expression is similar to that previously reported for *hoxd11*[35,90] and suggest a loss of contact with the distal limb enhancers located 5′ of the *hoxd* cluster, caused by the expansion of the posterior *hoxd* cluster (see main text). Scale bars 0.2 mm. Silhouettes are from[34].



Extended Data Fig. 9 | Comparison of *Neoceratodus* and mouse *hox* clusters. Four *hox* clusters are present in the *Neoceratodus* genome (*hoxa-hoxd*), comprising 43 genes and 6 conserved microRNA genes (*miR10* and *miR196*). *Neoceratodus* preserves a copy of *hoxb10* and *hoxa14*, which are lost in tetrapods. The 3' *hoxc* cluster contains the *hoxc1* and *hoxc3* genes, which are lost in several tetrapod lineages but have been shown to be part of the original tetrapod *hox* complement. In line with the overall expansion of the *Neoceratodus* genome its *hox* clusters are larger than their mouse counterparts. Expansion has occurred unevenly across the clusters and intergenic regions of highest expansion are indicated with yellow markup (*hoxa11-hoxa13*, *hoxb10-hoxb13*, *hoxc1-hoxc3-hoxc4*, *hoxc11-hoxc12*, and *hoxd12-hoxd13)*. Furthermore, the introns of *hoxa3* and *hoxd3* are enlarged. All clusters shown (both mouse and *Neoceratodus*) are drawn to scale with the respective sizes indicated, except for the 11Mb between *hoxb10* and *hoxb13*, which is drawn about 20 fold reduced. The *Neoceratodus hoxb13* and *hoxd13* are present on separate contigs and the exact genomic distance to their nearest neighbouring *hox* gene has not been determined. The sizes for the *hoxb* and *hoxd* clusters therefore represent a lower limit. The mouse has lost *hoxa14* and the indicated synteny for *hoxa* runs from *hoxa1* through *hoxa13*. Similarly, the mouse *hoxc* cluster lacks *hoxc1* and *hoxc3* and the comparative *hoxc* synteny runs from *hoxc4* through *hoxc13*. Gene labels are included for the *Neoceratodus* cluster whereas in the mouse clusters genes are only indicated using red boxes. MicroRNAs are only indicated for the *Neoceratodus* clusters. Silhouettes are from[34].



Extended Data Fig. 10 | Validation of the assembly of the *Neoceratodus* genome. a, Read coverage along the assembly showing a portion of scaffold 01. Red lines mark regions exhibiting a coverage >3 standard deviations from mean. Overall, these regions represent 0.09% of the genome. **b**, Representative region showing read pile-up with coverage in excess of 3 standard deviations from the mean. The entire region is contained within a region annotated as repetitive by RepeatMasker (red interval).

nature research

Corresponding author(s): Axel Meyer

Last updated by author(s): Dec 18, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☒	☐	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used to collect the data. I don't know if we should list NanoPore software here. Probably not, as it was done by a company. For RNAseq data, I also don't think we need to list the sequencer software, it doesn't make sense. Therefore, I think the statement in the first sentence should be enough
Data analysis	MARVEL https://github.com/schloi/MARVEL

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

PRJNA645042

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	n/a
Data exclusions	No data were excluded
Replication	n/a
Randomization	The data were not randomized
Blinding	No blinding was necessary as we used all data for the genome assembly and analyses thereof

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	*For laboratory animals, report species, strain, sex and age OR state that the study did not involve laboratory animals.*
Wild animals	*Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.*
Field-collected samples	*For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*
Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.